Climatic indices from meteorological and oceanographic reanalyses: Analysis error.

W. C. Thacker *

Atlantic Oceanographic and Meteorological Laboratory
4301 Rickenbacker Causeway
Miami FL 33149 US

May 8, 1995

Abstract

Estimates of meteorological and oceanographic fields obtained by assimilating forty to fifty years of data into forecast models can provide an excellent description of the evolving seasonal-to-interannual climate, but one which is subject to the uncertainty of estimating gridded fields from local observations. The purpose of this paper is to present a way to construct indices characterizing the variability exhibited by such reanalyses while accounting for analysis errors. Seeking indices with minimal uncertainty leads naturally to a metric-based principal-component analysis in which the metric is specified by the covariances of the analysis errors.

1 Introduction.

The National Weather Service is currently reanalyzing meteorological and oceanographic data using consistent models and assimilation procedures over the period extending from the 1950's to the present (Kalnay and Jenne, 1991). Such reanalyses provide time series of high-resolution fields characterizing

* thacker@aoml.erl.gov

seasonal-to-interannual variability, which can be used in constructing statistical forecast models. Statistical models are typically formulated in terms of a relatively small number of indices, *e.g.*, principal components (Pearson, 1901; Preisendorfer, 1988; Jolliffe, 1986), which account for the bulk of the spatial and temporal variability exhibited by the climatic fields (Graham *et al.*, 1987; Barnston and Ropelewski, 1992; Penland and Magorian, 1993). Future values of the indices are computed as linear combinations of present and past values, the coefficients having been determined by fitting to the available data. Forecasts for the climatic fields are then computed from the predicted values of the indices as superpositions of empirical orthogonal functions, or equivalently, as a regression of the local variables on the indices (Thacker, 1995a).

The number of indices required to reproduce faithfully all of the spatial details of every meteorological and oceanographic field for the entire reanalysis interval is limited by the length of the interval. For example, a reanalysis producing forty years of monthly-mean fields can be represented precisely using no more than 480 indices. Moreover, the predictable aspects of the variability might be accounted for using far fewer than 480. First, the indices necessarily reflect the particular 40-year interval and cannot be expected to characterize some other 40-year interval with the same precision; *i.e.*, they are subject to sampling error (Preisendorfer *et al.*, 1981; von Storch and Hannoschöck, 1985; Thacker, 1995a). Second, they reflect the uncertainty associated with estimating gridded fields from local observations; *i.e.*, they are subject to analysis error.

These two types of errors are conceptually different. Analysis errors would exist even in the limit of an infinitely long record, when sampling error should vanish. Conversely, sampling error is a concern, even if the monthly-mean fields were known precisely. The concept of sampling error is most easily understood within the context of prediction: because the coefficients of each predictor are determined from a limited sample, increasing the number of predictors can reduce the model's performance. Here, the predictors are chosen from a set of indices, and the performance is judged by how well the predictors can explain the spatial details of the climatic fields. Thus, two decisions are involved: the first involves the specification of potential predictors and the second involves the selection of which and how many to use.

This paper presents a method for defining a set of indices that measure the climatic variability with respect to the uncertainties associated with analysis errors. The first index should be as free from analysis error as possible while also accounting for as much of the variability as possible. The second should be uncorrelated with the first, so it might serve as an independent predictor, and it should also explain as much of the remaining well-determined variability as possible, and so on. The method that achieves this is based on the assumption that the uncertainties in the monthly-mean fields produced by the reanalysis are similar for each month of the record. Given the practical difficulty in assessing the actual uncertainties, this assumption is reasonable enough. It is needed so that the analysis errors can be characterized by a single error-covariance matrix. A starting point for estimating this matrix is the reanalysis scheme, which is based on optimal interpolation with a specified error-covariance matrix.

Seeking the index of greatest variance and least uncertainty leads to a generalized eigenproblem involving both the sample-covariance matrix and the error-covariance matrix. The role of the error-covariance matrix is that of a metric: it allows the variability of the monthly-mean fields to be compared to the variability allowed by the uncertainty associated with the interpolation of the data. If the error-covariance matrix were replaced with the identity, the eigenproblem would reduce to that of standard principal-component analysis and variability would be measured with respect to the units in which the variables are expressed. Thus, the method might be referred to as metric-based principal component analysis with the error-covariance matrix serving as the metric.

Conventional principal-component analysis of the sample correlation matrix provides another example of metric-based principal-component analysis (Thacker, 1995a). In that case the metric is determined by the diagonal of the sample covariance matrix; every variable is measured with respect to its own standard deviation. When the error-covariance matrix is the metric, it also provides a scale for measuring variability, even though it is not diagonal; the original variables can be linearly transformed to a new set of variables whose errors are uncorrelated, and these variables are measured with respect to the standard deviations of their errors.

Metric-based principal-component analysis can be regarded as a consequence of metric-based singular-value decomposition, which requires two metrics, one for rows and the other for columns (Greenacre, 1984; Escoufier, 1987). For example, canonical correlation analysis follows from the singular-value decomposition of a matrix of covariance between variables in two different sets with the two metrics taken to be the covariance matrices of the variables within each of the two sets. When a data matrix is decomposed, one metric typically represents the distance (dissimilarity) between multivariate samples and the other represents the relative importance (weight) of a sample. Here, each sample represents the climatic state at a particular time, and the metric specifies how distances between two different climatic states (two different times) are to be measured.

One advantage that metric-based principal-component analysis has over conventional principal-component analysis is that the metric forces us to state explicitly how the variables are to be scaled. Consequently, the results are dimensionally homogeneous, even when the analysis involves variables of different type, such as temperature and pressure. The conventional approaches, on the other hand, can give different results when the variables are expressed in different units. (Escoufier (1987) points out that, to avoid having the procedure dictate the the course of the analysis, care should be taken in specifying, not only the scale with which each variable is to be measured, but also the zero point for each variable.)

Because metric-based principal-component analysis accommodates data having different units, it is not surprising that it can be used to find correlated patterns in two sets of data. In section 4 below, the example from (Bretherton *et al.* (1992)) is reexamined, and prescribed patterns in noisy artificial data are shown to be recoverable when the known noise-covariance matrix is the metric.

A distinguishing feature of metric-based principal-component analysis is that there are *two* sets of EOF-like patterns (Thacker, 1995a). The first are the patterns computed as metric-based eigenvectors of the sample covariance matrix. These are patterns of coefficients used in defining indices as linear combinations of the original variables, *i.e.*, of each variable at each grid point; the units of the coefficients are reciprocally related to those of the corresponding variables, so that the indices are dimensionally homogeneous. The second patterns are obtained from the first by multiplying by the metric, in this case by the error-covariance matrix; they have the same units as the original variables. These are patterns of regression coefficients, which indicate what values would be predicted for the original variables based on using a single index as a predictor. Because the indices are uncorrelated, predictions based on multiple indices can be added; using all indices as predictors reproduces all of the variability of the *sample* exactly. Thus, the decomposition of the data into principal components and EOF's becomes a decomposition into products of indices and patterns of the second type. Thus, it is appropriate to refer to the patterns of the second type as state-space patterns and to those of the first type as dual or coefficient patterns.

The distinction between state-space patterns and their duals is also encountered in the problem of detecting global warming (Hasselmann, 1979; Hasselmann, 1993; Thacker, 1995b). For that problem the state-space pattern is suggested by the difference between two model simulations (the signal) and the adjoint pattern (the fingerprint) is used to find a signal in observations. Because the objective is to determine whether the signal is sufficiently strong to reject a hypothesis of no climatic change, the metric for that problem is determined by natural variability. This should be contrasted with the objective of this paper, which is to characterize the natural variability in the presence of analysis errors.

The results of metric-based principal-component analysis of a 126-month record of the temperature field at 25-meters depth in the tropical Pacific Ocean were qualitatively similar to those of a conventional principal-component analysis of the same data. It would be difficult to distinguish the first few state-space patterns from the first few conventional EOF's. However, the first few conventional principal components explain a greater percentage of the total variability than the percentage of variability to uncertainty explained by the same number of indices based on a metric corresponding to the error-covariance matrix for the reanalyses. This fact is best understood by considering the scale dependence of analysis errors. Because the greatest uncertainties are at the largest scales, which correspond to the greatest variability, the smaller scales become more important. Thus, more indices, *i.e.*, smaller scales, are needed to account for the same percentage of uncertainty-corrected variability.

2 Metric-based principal components.

The results of a multi-decadal reanalysis can be represented as a series of state vectors $x(t_j)$, one for each time t_j, $j = 1, \ldots, n$. For example, $x(t_j)$ might be a column vector containing the monthly mean values of every variable at every grid point for the jth month of the analysis period, or it might contain only selected variables at selected points that are to characterize the state of the climate. Because the reanalysis produces estimates on high-resolution grids used for numerical modelling, the dimensionality of x can be expected to be quite high and not all of the variables will be statistically independent. If the results of the reanalysis are to be used to build a statistical model describing seasonal-to-interannual variability, it will be necessary to use a much smaller set of variables. One possibility is to base the statistical model on a relatively few selected variables. Another is to reduce the resolution by averaging neighboring values to produce variables on a coarser grid; for example, the Niño 3 Index might be one of the variables of the statistical model. Still another be to use differences of widely separated values; the Southern Oscillation Index, for example. More generally, the variables used for the statistical model might be any linear combinations of the variables provided by the reanalysis. Any linear combination

$$\gamma(t_j) = \alpha^T x(t_j) \tag{1}$$

provides an index that characterizes some aspect of the climatic state. The problem addressed here is that of determining coefficient vectors α_k defining indices γ_k that collectively summarize the variability exhibited by the reanalysis, which can be used as predictors in a statistical model.

It is important to recognize that the reanalysis provides only an estimate of the evolving climate subject to unknown errors $\delta x(t_j)$. Nevertheless, the uncertainty of the state can be characterized by the error-covariance matrix

$$B =< \delta x(t_j) \delta x(t_j)^T > , \tag{2}$$

which is assumed here to be the same for all t_j. For example, B might be taken to be the same error-covariance matrix as was used in optimally interpolating the data. Because the errors can be expected to be correlated, B cannot be expected to be diagonal; and because no aspects of the analysis can be expected to be free of uncertainty, B should not be singular. Using B, it is possible to estimate the uncertainty of any index;

$$\alpha^T B \alpha =< \alpha^T \delta x \delta x^T \alpha >=< \delta\gamma^2 > \tag{3}$$

is the variance of the error of the index $\gamma = \alpha^T x$.

The variability exhibited by the reanalysis is characterized by the sample covariance matrix

$$A = \frac{1}{n} \sum_{j=1}^{n} x(t_j) x(t_j)^T , \tag{4}$$

where, for simplicity, the variables are assumed to be in the form of departures from mean values or from the mean seasonal cycle. Because the variance of the index $\gamma(t_j) = \alpha^T x(t_j)$ is given by $\alpha^T A \alpha$, the ratio

$$\lambda^2 = \frac{\alpha^T A \alpha}{\alpha^T B \alpha} \tag{5}$$

represents the variability of the index relative to its uncertainty. The best-determined aspect of the climatic state corresponds to the index γ defined by the coefficient vector α for which λ^2 is maximum. Requiring the derivative of λ^2 with respect to α to vanish results in the generalized eigenproblem,

$$A\alpha_k = \lambda_k^2 B\alpha_k \ , \tag{6}$$

for eigenvalues λ_k and eigenvectors α_k. The eigenvector α_1 corresponding to the largest eigenvalue λ_1 is the coefficient vector that defines the index γ_1 associated with the aspect of the climatic state best determined by the reanalysis.

If B were the identity matrix, (6) would be the standard eigenproblem defining principal components. In that case λ_k^2 would correspond to the amount of total variance explained by the principal component γ_k. For any positive, symmetric matrix B, maximizing the ratio λ^2 defined in (5) leads to the generalized eigenproblem (6) for indices, with the meaning of the eigenvalues depending on the meaning of B. Here, B is the error-covariance matrix, so each eigenvalues is the ratio of an index's variability to its uncertainty. Metric-based principal components, like standard principal components, are mutually uncorrelated, and just as standard principal components correspond to the indices explaining the most variability, next most variability, *etc.*, these correspond to the greatest variability relative to uncertainty, second greatest, and so on.

The error-covariance matrix B serves as a metric and enters into the normalization and orthogonality of the eigenvectors. When they are normalized so that

$$\alpha_k^T B\alpha_l = \delta_{k,l} \ , \tag{7}$$

the variance of the index γ_k is the corresponding eigenvalue, *i.e.*,

$$\alpha_k^T A\alpha_k = \lambda_k^2 \ . \tag{8}$$

Note that (7) implies that indices have uncorrelated errors and that each index has unit error variance. Using (6), it is easy to see that the indices themselves are mutually uncorrelated.

In standard principal-component analysis, the eigenvectors of the sample covariance matrix (the EOF's) play a dual role. First, they are vectors of coefficients defining the principal components as linear combinations of the state variables, and second, they represent the pattern of variability associated with the temporally varying principal component. For metric-based principal components, these two roles are played by two distinct sets of vectors. The eigenvectors α_k are the coefficient vectors defining the principal components γ_k, while the patterns of variability are the vectors

$$\beta_k = B\alpha_k \ . \tag{9}$$

The state-space patterns are bi-orthogonal to the coefficient vectors,

$$\beta_k^T \alpha_l = \delta_{k,l} \ ; \tag{10}$$

their orthonormality condition is

$$\beta_k^T B^{-1} \beta_l = \delta_{k,l} \ . \tag{11}$$

The reanalysis variables can be decomposed into a series of products of temporally varying indices and their corresponding state-space patterns:

$$x(t_j) = \sum_k \gamma_k(t_j)\beta_k \ . \tag{12}$$

This decomposition is a consequence of the metric-based singular-value decomposition discussed in section 3 below. Another way of looking at (12) is as a linear statistical model for the state variables; the indices are uncorrelated predictors and the patterns are Gauss-Markov weights (Thacker, 1995a).

3 Metric-based singular-value decomposition.

Standard singular-value decomposition, *i.e.*,the decomposition of a matrix into a sum of rank-one matrices formed as an outer product of basis vectors for the rows and columns, is based on the premise that the basis vectors are orthogonal with an identity matrix as metric. This premise can be generalized to allow for orthogonality relative to any specified metric. Because metric-based singular-value decomposition is not well-known to to meteorologists and oceanographers, it is presented here first within a general context and then specialized to the case of a data matrix and an error metric.

Metric-based singular-value decomposition of the matrix M is defined by the pair of equations,

$$Mr_k = \sigma_k C c_k \tag{13}$$

and

$$M^T c_k = \sigma_k R r_k \ , \tag{14}$$

where the vectors r_k span the row space of M, c_k span the column space, and σ_k are the metric-based singular values of M. The positive, symmetric matrices C and R are the metrics for the column and row spaces, respectively. For standard singular-value decomposition, C and R are taken to be identity matrices. Because M can be rectangular, the dimension of the row space might not be the same as that of the column space; every basis vector of the smaller space is paired with a basis vector of the larger space and the unpaired vectors from the larger space correspond to singular values that are zero.

Substituting equations (13) and (14) into each other yields a pair of generalized eigenproblems,

$$M^T C^{-1} M r_k = \sigma_k^2 R r_k \tag{15}$$

and

$$M R^{-1} M^T c_k = \sigma_k^2 C c_k \ , \tag{16}$$

which share the same eigenvalues. The matrices R and C determine the orthonormality conditions for the row and column spaces:

$$r_k^T R\, r_l = \delta_{k,l} \tag{17}$$

and

$$c_k^T C\, c_k = \delta_{k,l} \ . \tag{18}$$

The matrix M can be decomposed into a sum of rank-one matrices formed by outer products of the dual-basis vectors Rr_k and Cc_k of the row and column spaces with corresponding singular values as coefficients:

$$M = \sum_k \sigma_k (Cc_k)(Rr_k)^T \ . \tag{19}$$

Using the orthonormality conditions (17) and (18), it is easy to verify that this representation of M satisfies the defining equations (13) and (14).

Note the similarity of the eigenproblems (15) and (16) to those of canonical-correlation analysis (Hotelling, 1936; Kendall *et al.*, 1983; Graham *et al.*, 1987); M corresponds to the matrix of covariances between two sets of data, while R and C correspond to the within-set covariance matrices. Thus, canonical-correlation analysis provides an example of metric-based singular-value decomposition. This might be contrasted with the standard singular-value decomposition of the cross-covariance matrix (Bretherton *et al.*, 1992).

The metric-based principal-component analysis of the sample covariance matrix discussed above follows from the metric-based singular-value decomposition of the data matrix D. D is defined so that each row is the transpose of the state vector for some particular time and each column is a time series of one of the state variables. The row-space metric is the error-covariance matrix B, while the column-space metric is the identity matrix multiplied by the length of the time series $C = nI$. Using these matrices for M, R, and C, equations (13) and (14) become

$$D\alpha_k = \lambda_k nI c_k \tag{20}$$

and

$$D^T c_k = \lambda_k B\alpha_k \ . \tag{21}$$

Anticipating the fact that the row-space basis vectors are the generalized eigenvectors of the sample covariance matrix, r_k has been replaced with α_k. Because multiplying α_k by D gives γ_k, the vector containing time series of the kth index, the column-space basis vector c_k in (20) should be proportional to γ_k. While the γ_k was normalized so that $\gamma_k^T\gamma_k = n\lambda_k^2$, *i.e.*, so that the variance of the index was the eigenvalue λ_k^2, $c_k = \gamma_k/n\lambda_k$ is normalized so that $c_k^T nI\, c_k = 1$ to be consistent with (18). Equation (21) expresses the fact that the state-space pattern β_k is related to the correlation between the index c_k and the original variables.

The two eigenvalue equations (15) and (16) become

$$\frac{1}{n} D^T D\alpha_k = \lambda_k^2 B\alpha_k \tag{22}$$

and

$$\frac{1}{n} DB^{-1}D^T c_k = \lambda_k^2 c_k \ . \tag{23}$$

The matrix on the left-hand side of (22) is the sample covariance matrix A, so (22) is the same as (6). The second eigenproblem can be solved for the time series of the indices over the reanalysis interval. Because there generally will be far fewer months of available data than there will be variables characterizing the climatic state, this eigenproblem should be much smaller and easier to solve. Once it has been solved, the coefficients defining the indices, can be computed using (21).

The generalized singular-value decomposition (19) of the data matrix D is

$$D = \sum_k \lambda_k (nc_k)(B\alpha_k)^T = \sum_k \gamma_k \beta_k^T \ . \tag{24}$$

This is simply the matrix representation of the expansion in terms of patterns and time series given in equation (12).

4 An example from Bretherton, Smith, and Wallace.

In discussing several techniques for identifying correlated patterns in two sets of data, *e.g.*, two different fields, Bretherton *et al.* (1992) suggest a simple "toy" problem. The data in each set were composed of two parts, signal and noise. The signals were prescribed patterns of spatial variability ϕ_1 and ϕ_2 having the same temporal behavior $f(t)$, which was normalized to have unit variance. In particular, the number of data in each set was taken to be the same and ϕ_1 was taken to be equal in magnitude and opposite in sign to ϕ_2. The noise was characterized by the a time-independent covariance matrix V, which plays the same role as the covariance matrix of reanalysis errors B.

One technique they examined for finding the patterns ϕ_1 and ϕ_2 in the presence of noise was combined principal-component analysis; the two sets of data were combined and principal components were computed for the sample covariance matrix of the combined data. (Such computations are generally not dimensionally homogeneous.) They found that the first EOF differed from the prescribed pattern. However, if metric-based principal components were used, the state-space pattern of the first index would have corresponded exactly to the prescribed pattern.

When the record of observations is sufficiently long, the signal and noise are uncorrelated and the sample covariance matrix A can be separated into contributions from signal and noise,

$$A = S + V \; . \tag{25}$$

For this example, the contribution from the signal is the rank-one matrix,

$$S = \Phi\Phi^T \; , \tag{26}$$

where

$$\Phi = \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} \; . \tag{27}$$

So, the eigenproblem (6) becomes,

$$S\alpha_k = (\lambda_k - 1)V\alpha_k \; . \tag{28}$$

Because S is a rank-one matrix, it can have only one non-zero eigenvalue. Thus, $\lambda_k = 1$ for all k except $k = 1$; *i.e.*, there is only a single index for which the variability exceeds the uncertainty. The corresponding (un-normalized) eigenvector is the dual of the prescribed pattern,

$$\alpha_1 = V^{-1}\Phi \ . \tag{29}$$

By substituting (29) into (28) and using (26), it is easy to verify that the corresponding eigenvalue is

$$\lambda_1 = 1 + \Phi^T V^{-1}\Phi \ . \tag{30}$$

The corresponding state-space pattern is

$$\beta_1 = V\alpha_1 = \Phi \ . \tag{31}$$

Thus, the first state-space pattern is the same as the pattern of the prescribed signal.

5 Computational results for correlated, scale-dependent, analysis errors.

The 25-meters-depth, monthly-mean temperature field of the tropical Pacific Ocean for the 126-month period extending from July 1982 through December 1992, which was taken from a reanalysis produced by National Meteorological Center's Ocean Analysis System (Ji *et al.*, 1995), was arbitrarily selected for testing metric-based principal-component analysis. The object was to determine how principal components based on an analysis-error metric compare to conventional principal components.

The statistical interpolation scheme used in the reanalysis was much the same as that described by Derber and Rosati (1989), *i.e.*, the prior error-covariance matrix has the form L^p, where L is a matrix representation of a local five-point averaging operator and where the exponent p indicates the number of times the averaging is repeated. Such an error-covariance matrix is extremely convenient for data assimilation, but it presents computational difficulties here. Because the tropical 25-meter temperature fields are represented on a high-resolution grid, the matrices A and B in the eigenproblem (6) are 4919×4919. Not only is this eigenproblem large, most of the eigenvalues are identically zero. It is much more convenient to solve the much smaller eigenproblem (23), for which the matrix $DB^{-1}D^T$ is only 126×126. For computational convenience, a sparse approximation to the inverse of their error-covariance matrix was needed for computing the matrix $DB^{-1}D^T$. Therefors, B was defined implicitly via

$$B^{-1} = \epsilon^{-2}(I + (b^2\nabla^2)^2) \ , \tag{32}$$

where the finite-difference Laplacian operator ∇^2 is represented as 5-point stencils, which varies from point to point due to the irregularities of the spherical grid. It was constructed to be symmetric and to have Neumann boundary conditions. The identity matrix I, which serves to guarantee that B^{-1} is not singular, does not interfere with B being a smoothing operator. The coefficient b determines the extent of the smoothing. Although we have not established a precise correspondence between values of b and the e-folding scale of 4° used by Ji *et al.* (1995), we have chosen $b = 200\,\mathrm{km}$ based on computations indicating that this error model should resemble theirs. This level of smoothing would reduce all component wavelengths in the data that are shorter than 4°to less than 1/50 their original amplitude. The amount of smoothing is independent of the value of ϵ, which determines the overall level of errors. Although the appropriate value of ϵ is unknown, an arbitrary value can be prescribed; the eigenvalues representing variability with respect to uncertainty are determined up to an unknown scale factor, which is sufficient to study the relative importance of the indices, and the corresponding patterns are unaffected.

The state-space patterns were compared with the EOF's computed as eigenvectors of the sample covariance matrix, and the two sets of patterns were found to be qualitatively similar. For both sets of patterns, the maxima were located in regions of greatest variability. This was also the case for state-space patterns based on the sample correlation matrix. In that case the eigenvectors of the correlation matrix had to be rescaled by the local standard deviation to convert them to state-space patterns with units of temperature. And for all three sets, the patterns corresponding to the largest eigenvalues varied on the largest spatial scales, while those corresponding to the smallest eigenvalues varied on the smallest scales. Given three sets of maps of the first dozen patterns for each of the three cases, it would be difficult to identify which maps corresponded to which case. They all resembled typical EOF maps, so there was no point to present them here.

The index time series for the three analyses were also compared. Index-by-index comparison showed somewhat larger differences than did the spatial patterns, but again, it would be difficult to examine three sets of indices and discern which came from which analysis. A more important comparison was the ability of the first few indices to explain the variability exhibited by the reanalyses. This comparison was made in two ways. First the eigenvalue spectra were compared to determine what percentage of the total variability was accounted for by the individual indices, and then maps of temperature fields based on the first few indices were examined to see how well the fields were approximated.

In comparing the eigenvalue spectra, it is important to keep in mind that the eigenvalues have different meanings for each of the three cases. For conventional covariance-matrix principal components, it is a percentage of the total variance; for correlation-matrix principal components, it is the predictive ability of the indices expressed as fraction of the total number of grid cells they can replace; and for metric-based principal components, it is total variance relative to error variance. The first several conventional principal components accounted for a greater fraction of their spectral energy than did the same number of components in the other two cases. To explain the same fraction, slightly more correlation-matrix principal components were needed, as might be expected, because they give equal emphasis to regions of low and high variability. Still more error-metric indices were required to account for the same amount of variability. This might be explained by the fact that the error-covariance matrix corresponds to the greatest uncertainties associated with the largest scales and the least with the smallest scales. Because there is more emphasis on the smaller scales, more indices are required. Of course, in all three cases, the complete set of indices explained all of the variability.

In comparisons of maps of the temperature fields reconstructed from the first few indices, the results were similar. The greatest level of detail attained with the fewest indices was attained using conventional principal components, and the least, for metric-based indices. It should be kept in mind, however, that some of this detail is due to analysis errors.

The more important question, "Which indices are best at explaining the variability of independent data?" was not addressed, because a much longer reanalysis interval would be needed to answer properly. Part of the record would be needed for determining the coefficient vectors and the state-space patterns, and another part would be required for verification. Indices would have to be computed from the verification data, using the coefficient vectors determined from the training set, and then products of indices and state-space patterns would have to be summed for comparison with the second part of the reanalysis. In other words, the indices would be used as predictors in linear statistical models for the original variables. Finally, the models based on error-metric principal components would have to be compared with models using predictors from the other sets of indices. Only then would it be possible to judge which indices really explain more of the "predictable" variability.

The sort of predictability discussed in the previous paragraph is *contemporaneous* predictability. It reflects the extent to which the values of a few indices at any given time reflect the all of the state variables, *e.g.*, the temperature at each grid point, for the same time. The ultimate goal is to predict the state variables at a *future* time. That goal might be achieved in two steps: first, future values of the indices would be forecast from their present and past values, and then, the future values of state variables would be predicted contemporaneously from the predicted indices. So an equally important question is, "Which indices are best forecast from their present and past values?" Again, no attempt has been made to answer this question, due to the lack of sufficient data, but it is reasonable to think that indices less subject to analysis error should make better predictors.

6 Conclusions.

The first conclusion of this paper is that seeking indices characterizing the well-determined aspects of model-based analyses of climatic data leads to a natural extension of the familiar principal-component analysis. The error-covariance matrix characterizing the uncertainty of the analyses appears in the formalism as an explicit metric, which serves to measure variability with respect to uncertainty caused by analysis errors. The indices are mutually uncorrelated linear combinations of the analysis variables, which are defined by their ability to explain total variability relative to uncertainty. The EOF's of conventional principal-component analysis are differentiated into two different types of patterns, which are related via the metric: state-space patterns and patterns of coefficients that define the indices.

The second conclusion is that metric-based principal-component analysis follows from metric-based singular-value decomposition of a data matrix. Any matrix can be decomposed into a sum of outer products of column- and row-space basis vectors that are orthonormal with respect to specified column- and row-space metrics. When the rows and columns of the data matrix correspond to variables and instances, the row-space metric specifies how the different variables should be measured, *i.e.*, relative to their analysis errors , and the column-space metric specifies the relative importance of data at different times. Such a formulation not only makes explicit how data of different types should be compared, it also allows computations to be dimensionally homogeneous so that the results do not depend on the choice of units for the variables. A related result is that canonical-correlation analysis can be recognized to follow from the singular-value decomposition of the cross-covariance matrix for data in two different sets with the two within-set covariance matrices serving as metrics.

The third conclusion is that metric-based principal-component analysis is capable of finding correlated patterns of variability in two different data sets. This was demonstrated within the context of the simple example of Bretherton *et al.* (1992), where the patterns were prescribed, as were the statistics of the background noise, which plays the role of analysis error. Because there is only one prescribed pattern, there is only a single index for which the variability is greater than the uncertainty. The state-space pattern for that index was shown to be exactly the same as the specified pattern.

The fourth conclusion, which is based results for a 126-month reanalysis of oceanic thermal data, is that metric-based principal components differ little qualitatively from conventional principal components. However, because they account for variability with respect to uncertainty rather than variability with respect to the unit of measurement and because the uncertainty is greatest for the largest scales, the largest scales account for a somewhat greater percentage of the total variability than of the variability with respect to uncertainty.

The fifth conclusion is that sampling error affects metric-based principal components in exactly the same way as it affects conventional principal components. To gauge the extent to which conclusions are specific to the sample, it is necessary to have additional data that can be used for verification.

The final conclusion is that metric-based principal components might be preferred to conventional principal components as predictors for statistical models of seasonal-to-interannual change, because they should provide a characterization of the observations that is less subject to analysis error. For the potential of this method to be realized, care should be taken when reanalyzing data to provide accurate estimates of analysis errors.

References

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

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.

Derber, J., and A. Rosati (1989). A global oceanic data assimilation system. *Journal of Physical Oceanography*, **19**, 1333–1347.

Escoufier, Y. (1987). The duality diagram: A means for better practical applications. In *Developments in Numerical Ecology*, volume G14 of *NATO ASI Series*, pages 251–259. Springer-Verlag.

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.

Hasselmann, K. (1979). On the signal-to-noise problem in atmospheric response studies. In *Meteorology of Tropical Oceans*, pages 251–259. Royal Meteorological Society.

Hasselmann, K. (1993). Optimal fingerprints for the detection of time-dependent climate change. *Journal of Climate*, **6**, 1957–1971.

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

Ji, M., A. Leetmaa, and J. Derber (1995). An ocean analysis system for climate studies. *Monthly Weather Review*, pages 460–481.

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

Kalnay, E., and R. Jenne (1991). Summary of the nmc/ncar reanalysis workshop of April 1991. *Bulletin American Meteorological Society*, **72**, 1897–1904.

Kendall, M. G., A. Stuart, and J. K. Ord (1983). *The Advanced Theory of Statistics*, volume 3. Charles Griffin & Company Ltd., 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., F. W. Zwiers, and T. P. Barnett (1981). Foundations of principal component selection rules. SIO Ref. Ser. 81–4, Scripps Institution of Oceanography.

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

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

Thacker, W. C. (1995b). Indice characterizing patterns of climatic change. Submitted to *Journal of Climate*.

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.