Tellus (1996), 48A, 584–592
Printed in UK – all rights reserved

TELLUS
ISSN 0280–6495

Metric-based principal components:
data uncertainties

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

(Manuscript received 24 July 1995; in final form 15 December 1995)

ABSTRACT

Seeking an index characterizing the best-determined mode of variability leads to a natural generalization of principal-component analysis with an explicit metric characterizing the uncertainties of the data. This formalism, which distinguishes between state-space patterns and patterns of coefficients defining principal components, allows the more accurate data to exert a greater influence on the definition of the indices than they do in conventional principal-component analysis; in all other aspects, the new formalism is the same as the old. Within the context of the simple example of Bretherton and collaborators, metric-based principal-component analysis is shown to be capable of finding correlated patterns of variability in two different data sets.

1. Introduction

This paper presents a generalization of principal-component analysis (Pearson, 1901; Jolliffe, 1986) that accounts for uncertainties in the data. This is particularly important in oceanography and meteorology where the data characterize spatio-temporal averages. For example, COADS summaries (Woodruff et al., 1987) provide sea-surface temperatures in the form of monthly means for 2° × 2° latitude-by-longitude cells. The number of observations contributing to a particular mean reflects the number of ships that happened to be in the region that month. For a particular month, one cell's mean may be computed from hundreds of observations, while others may be based on only two or three, and there may be no observations at all for some cells. Another example might be provided by meteorological analyses of heights of 500 hPa pressure surfaces: the height is better estimated over populated regions than over the oceans where the observational network is sparse. Principal-component analysis involves estimating covariances from such data. Clearly, covariances between poorly sampled regions are less meaningful than those between regions that have been well observed throughout the analysis interval, so it is desirable to discount their influence.

Traditional principal components are linear combinations of the original variables, e.g., sea-surface temperatures and/or 500 hPa heights, which are uncorrelated within the sample and are ordered by their efficiency at explaining the total variance of the data. They can be determined variationally by seeking the linear combination of variables that has the greatest variance, subject to the constraint that the coefficient vector has unit length. Similarly, principal components that account for uncertainties in the data can be found by seeking the linear combination having the greatest variance relative to the variance of its error, i.e. by maximizing a signal-to-noise ratio. This variational problem leads to a generalized eigenproblem involving two matrices, the first being the usual sample covariance matrix of the data and the second being a covariance matrix

[*] thacker@aoml.erl.gov.

characterizing the uncertainties of the data. The second matrix plays the role of a metric defining what is meant by length of the coefficient vector. If all variables are estimated with equal accuracy, the second matrix becomes the identity, the length becomes the usual Pythagorean length, and this analysis reduces to the traditional principal-component analysis. However, in the general case where some variables are known with more accuracy than others, the resulting principal components are defined to emphasize the better-observed aspects of the data.

A distinguishing feature of metric-based principal-component analysis is that there are *two* sets of EOF-like patterns. 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. The second patterns are obtained from the first by multiplying by the metric, in this case by the error-covariance matrix. The decomposition of the data into principal components and EOFs 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, 1996). 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 analyze the natural variability exhibited by data of varying reliability.

The organization of the paper is the following: Section 2 defines a measure of uncertainty for an arbitrary linear combination of variables and derives metric-based principal components by seeking the linear combination with the greatest ratio of variability to uncertainty. Section 3 shows how the metric-based principal-component analysis is related to an underlying metric-based singular-value decomposition of the data, and Section 4 points out that the metric makes the analysis invariant under changes in the units of measurement. Section 5 re-examines the simple example of Bretherton et al. (1992) and shows that metric-based principal components are capable of finding correlated patterns in two sets of data. Section 6 discusses computational results for data from model-based analyses of temperature at 25 m depth in the tropical Pacific Ocean, and Section 7 lists the conclusions.

2. Metric-based principal components

A collection of time series of values of climatic variables such as temperature and pressure at various locations can be represented as a series of state vectors $x(t_j)$, one for each time t_j, $j = 1, \ldots, n$. Any linear combination

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

provides an index that characterizes some aspect of the temporally varying climatic state. Traditional principal components are indices that are constructed to be mutually uncorrelated over the analysis interval and to have the property that the first principal component is the index accounting for the greatest fraction of total variance, the second accounts for the greatest fraction of remaining variance, etc. When the data are subject to uncertainty, the principal components are also subject to uncertainty. The problem addressed here is that of defining a set of principal-component-like indices that are defined by their ability to account for the best-determined aspects of the variability.

Just as the covariance between pairs of time series can be characterized by the sample covariance matrix

$$\boldsymbol{A} = \frac{1}{n-1} \sum_{j=1}^{n} x(t_j) x(t_j)^{\mathrm{T}}, \tag{2}$$

where, for simplicity, the variables are assumed to be in the form of departures from mean values, their uncertainties can be characterized by an error-covariance matrix

$$\boldsymbol{B} = \langle \delta x(t_j) \delta x(t_j)^{\mathrm{T}} \rangle. \tag{3}$$

For example, if the data are COADS monthly-mean sea-surface temperatures, $\boldsymbol{B}$ might be estimated from a Monte Carlo computation of the distribution of covariances resulting from randomly generated monthly-means consistent with the standard errors of the data. If the data were in the form of meteorological analyses, $\boldsymbol{B}$ could be taken to be the same error-covariance matrix as was used in optimally interpolating the data; in that case, the errors can be expected to be correlated, and $\boldsymbol{B}$ will not be diagonal. Because no aspects of the data can be expected to be free of uncertainty, $\boldsymbol{B}$ should not be singular.

Just as the sample variance of the index $\gamma(t_j) = \alpha^{\mathrm{T}} x(t_j)$ is given by:

$$\langle \gamma^2 \rangle = \alpha^{\mathrm{T}} \boldsymbol{A} \alpha, \tag{4}$$

its uncertainty is measured by

$$\langle (\delta\gamma)^2 \rangle = \alpha^{\mathrm{T}} \boldsymbol{B} \alpha. \tag{5}$$

Thus, the ratio

$$\lambda^2 = \frac{\alpha^{\mathrm{T}} \boldsymbol{A} \alpha}{\alpha^{\mathrm{T}} \boldsymbol{B} \alpha} \tag{6}$$

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,

$$\boldsymbol{A} \alpha_k = \lambda_k^2 \boldsymbol{B} \alpha_k, \tag{7}$$

for eigenvalues λ_k^2 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 data. The eigenvalue is the ratio of the variance of the index to its error variance.

If $\boldsymbol{B}$ were the identity matrix, (7) 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. In general, all variables are not equally well determined, so $\boldsymbol{B}$ is not the identity. Even then, the data can be represented by variables that are equally well determined, e.g., by the principal components of $\boldsymbol{B}$ normalized to have unit error variance. In terms of these variables the generalized eigenproblem (7) becomes an ordinary, single-matrix, principal-component eigenproblem with total variance referenced to a common level of uncertainty. Thus, the first metric-based principal component is the index that explains the largest fraction of the uncertainty-referenced variance; the second is uncorrelated with the first and explains the largest fraction of the remainder, and so on.

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

$$\alpha_k^{\mathrm{T}} \boldsymbol{B} \alpha_l = \delta_{k,l}, \tag{8}$$

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

$$\alpha_k^{\mathrm{T}} \boldsymbol{A} \alpha_k = \lambda_k^2. \tag{9}$$

Note that (8) implies that indices have uncorrelated errors and that each index has unit error variance. Using (7), 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 two roles. First, they are vectors of coefficients defining the principal components as linear combinations of the state variables, and second, they represent the patterns of variability associated with the temporally varying principal components. 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

$$\beta_k = \boldsymbol{B} \alpha_k. \tag{10}$$

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

$$\beta_k^{\mathrm{T}} \alpha_l = \delta_{k,l}; \tag{11}$$

their orthonormality condition is

$$\beta_k^{\mathrm{T}} \boldsymbol{B}^{-1} \beta_l = \delta_{k,l}. \tag{12}$$

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

$$x(t_j) = \sum_k \gamma_k(t_j) \beta_k. \tag{13}$$

This decomposition is a consequence of the metric-based singular-value decomposition discussed in Section 3 below. Another way of looking at (13) is as a linear statistical model for the state variables; the indices are uncorrelated predictors and the patterns are Gauss–Markov weights (Thacker and Lewandowicz, 1996).

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 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 $\boldsymbol{M}$ is defined by the pair of equations,

$$\boldsymbol{M}r_k = \sigma_k \boldsymbol{C}c_k, \tag{14}$$

$$\boldsymbol{M}^{\mathrm{T}} c_k = \sigma_k \boldsymbol{R}r_k, \tag{15}$$

where the vectors r_k span the row space of $\boldsymbol{M}$, c_k span the column space, and σ_k are the metric-based singular values of $\boldsymbol{M}$. The positive, symmetric matrices $\boldsymbol{C}$ and $\boldsymbol{R}$ are the metrics for the column and row spaces, respectively. For standard singular-value decomposition, $\boldsymbol{C}$ and $\boldsymbol{R}$ are taken to be identity matrices. Because $\boldsymbol{M}$ 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 eqs. (14) and (15) into each other yields a pair of generalized eigenproblems,

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

$$\boldsymbol{M}\boldsymbol{R}^{-1}\boldsymbol{M}^{\mathrm{T}}c_k = \sigma_k^2 \boldsymbol{C}c_k, \tag{17}$$

which share the same eigenvalues. The matrices $\boldsymbol{R}$ and $\boldsymbol{C}$ determine the orthonormality conditions for the row and column spaces:

$$r_k^{\mathrm{T}}\boldsymbol{R}r_l = \delta_{k,l}, \tag{18}$$

$$c_k^{\mathrm{T}}\boldsymbol{C}c_k = \delta_{k,l}. \tag{19}$$

The matrix $\boldsymbol{M}$ can be decomposed into a sum of rank-one matrices formed by outer products of the dual-basis vectors $\boldsymbol{R}r_k$ and $\boldsymbol{C}c_k$ of the row and column spaces with corresponding singular values as coefficients:

$$\boldsymbol{M} = \sum_k \sigma_k (\boldsymbol{C}c_k)(\boldsymbol{R}r_k)^{\mathrm{T}}. \tag{20}$$

Using the orthonormality conditions (18) and (19), it is easy to verify that this representation of $\boldsymbol{M}$ satisfies the defining eqs. (14) and (15).

Note the similarity of the eigenproblems (16) and (17) to those of canonical-correlation analysis (Hotelling, 1936; Kendall et al., 1983; Graham et al., 1987); $\boldsymbol{M}$ corresponds to the matrix of covariances between two sets of data, while $\boldsymbol{R}$ and $\boldsymbol{C}$ 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 $\boldsymbol{D}$, which 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 $\boldsymbol{B}$, while the column-space metric is the identity matrix multiplied by the adjusted length of the time series $\boldsymbol{C} = (n-1)\boldsymbol{I}$. Using these matrices for $\boldsymbol{M}$, $\boldsymbol{R}$, and $\boldsymbol{C}$, eqs. (14) and (15) become

$$\boldsymbol{D}\alpha_k = \lambda_k (n-1)\boldsymbol{I}c_k, \tag{21}$$

$$\boldsymbol{D}^{\mathrm{T}}c_k = \lambda_k \boldsymbol{B}\alpha_k. \tag{22}$$

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 $\boldsymbol{D}$ gives γ_k, the vector containing time series of the kth index, the column-space basis vector c_k in (21) should be proportional to γ_k. While the γ_k was normalized so that $\gamma_k^{\mathrm{T}}\gamma_k = (n-1)\lambda_k^2$, i.e., so that the variance of the index was the eigenvalue λ_k^2, $c_k = \gamma_k/(n-1)\lambda_k$ is normalized so that $c_k^{\mathrm{T}}(n-1)\boldsymbol{I}c_k = 1$ to be consistent with (19). Eq. (22) 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 eqs. (16) and (17) become

$$\frac{1}{n-1}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}\alpha_k=\lambda_k^2\boldsymbol{B}\alpha_k, \tag{23}$$

$$\frac{1}{n-1}\boldsymbol{D}\boldsymbol{B}^{-1}\boldsymbol{D}^{\mathrm{T}}c_k=\lambda_k^2c_k. \tag{24}$$

The matrix on the left-hand side of (23) is the sample covariance matrix $\boldsymbol{A}$, so (23) is the same as (7). 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 (22).

The generalized singular-value decomposition (20) of the data matrix $\boldsymbol{D}$ is

$$\boldsymbol{D}=\sum_k\lambda_k(n-1)c_k(\boldsymbol{B}\alpha_k)^{\mathrm{T}}=\sum_k\gamma_k\beta_k^{\mathrm{T}}. \tag{25}$$

This is simply the matrix representation of the expansion in terms of patterns and time series given in eq. (13).

4. The metric and units of measurement

Traditional principal components can depend on the units in which the data are expressed. For example, if some of the data are velocities and others are pressures, the principal components obtained when the data are expressed in meters per second and hectopascals are not the same as those for miles per hour and inches of mercury. Total variance is not dimensionally homogeneous; squared meters per second are added to squared hectopascals. This is an example of the well-known scaling problem of principal components. Another example is provided by the difference between principal components from the correlation matrix and those from covariance matrix; in computing the former, the former data are expressed in units of their standard deviations, while for the latter, the data are usually expressed in common units of measurement.

Metric-based principal components, on the other hand, are invariant under change of units. To achieve this invariance, the coefficients should transform covariantly, if the data transform contravariantly, i.e., the units of α should be reciprocally related to those of x, so that the expression (1) for the index is dimensionally homogeneous. Similarly, the generalized eigenvalue eq. (7) is dimensionally homogeneous, with the eigenvalue λ^2 being dimensionless. (Contrast this to the common convention of dimensionless coefficients and eigenvalues with units of variance, which only works when there is a single type of data.) State-space patterns have the same units as the data, the metric $\boldsymbol{B}$ guaranteeing dimensional homogeneity in eq. (10). In fact, all of the equations of Sections 2 and 3 are dimensionally homogeneous. Rescaling of the data necessarily results in a rescaling of both the sample covariance matrix $\boldsymbol{A}$ and the error covariance matrix $\boldsymbol{B}$, so the scaling enters the formalism explicitly via the metric rather than implicitly via the choice of units.

Although the focus here is on a metric that allows variability to be measured with respect to uncertainty, other metrics can be used to achieve other goals. For example, if the metric is a dimensional *identity matrix*, each element having the same units as the corresponding element of the covariance matrix, the results are numerically equivalent to those of conventional covariance-matrix principal-component analysis; however, the principal components will be dimensionless, and except for their units coefficient and state-space patterns will be indistinguishable. Correlation-matrix principal components, on the other hand, correspond to a metric $\boldsymbol{B}$ with principal diagonal the same as that of the sample covariance matrix $\boldsymbol{A}$ and all other entries zero; their state-space patterns are the eigenfunctions of the correla tion matrix scaled by each variable's standard-deviation.

5. An example from Bretherton et al.

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 to sign to ϕ_2. The noise was characterized by the time-independent covariance matrix $\boldsymbol{V}$, which plays the same role as the error-covariance matrix $\boldsymbol{B}$.

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 $\boldsymbol{A}$ can be separated into contributions from signal and noise,

$$\boldsymbol{A} = \boldsymbol{S} + \boldsymbol{V}. \tag{26}$$

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

$$\boldsymbol{S} = \Phi\Phi^{\mathrm{T}}, \tag{27}$$

where

$$\Phi = \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}. \tag{28}$$

So, the eigenproblem (7) becomes,

$$\boldsymbol{S}\alpha_k = (\lambda_k^2 - 1)\boldsymbol{V}\alpha_k. \tag{29}$$

Because $\boldsymbol{S}$ is a rank-one matrix, it can have only one non-zero eigenvalue. Thus, $\lambda_k^2 = 1$ 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,

$$\alpha_1 = \boldsymbol{V}^{-1}\Phi. \tag{30}$$

By substituting (30) into (29) and using (27), it is easy to verify that the corresponding eigenvalue is

$$\lambda_1^2 = 1 + \Phi^{\mathrm{T}}\boldsymbol{V}^{-1}\Phi. \tag{31}$$

The corresponding state-space pattern is

$$\beta_1 = \boldsymbol{V}\alpha_1 = \Phi. \tag{32}$$

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

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

The 25 m-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 is approximated as being proportional to $\boldsymbol{L}^p$, where $\boldsymbol{L}$ 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 m temperature fields are represented on a high-resolution grid, the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ in the eigenproblem (7) 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 (24), for which the matrix $\boldsymbol{DB}^{-1}\boldsymbol{D}^{\mathrm{T}}$ is only 126 × 126. For computational convenience, a sparse approximation to the inverse of their error-covariance matrix was needed for computing the matrix $\boldsymbol{DB}^{-1}\boldsymbol{D}^{\mathrm{T}}$. Here, $\boldsymbol{B}$ was defined implicitly via*

$$\boldsymbol{B}^{-1} = \varepsilon^{-2}(\boldsymbol{I} + (b^2\nabla^2)^2), \tag{33}$$

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 $\boldsymbol{I}$, which serves to guarantee that $\boldsymbol{B}^{-1}$ is not singular, does not interfere with $\boldsymbol{B}$ being a smoothing operator. The coefficient b determines the extent of the smoothing. Although

* This expression can be regarded as a two-term approximation to the power-series expansion of the inverse of $\boldsymbol{L}^p = (\boldsymbol{I} + \nabla^2/4)^p$.

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 $b = 200$ 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 3 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. In spite of this general similarity, it is important to understand that the sets of patterns did differ in details. A single pattern in one set generally had features found in several patterns in the other sets. Rather than regarding a state-space pattern as a mode of variability, it is better to think of it as the pattern of covariance between the index and the original variables, i.e., $\lambda_k \beta_k = A\alpha_k = \mathrm{cov}(x, \gamma_k)$.

Computations were also carried out using a hypothetical error model for which errors were much smaller along the well-sampled ship tracks than for the poorly-observed grid cells. Surprisingly, this 4th set of state-space patterns was qualitatively similar to the other three sets. Only when the error variance of the poorly-sampled cells was seven orders of magnitude greater than that of the well-sampled was there any indication of the ship tracks in the dominant patterns. However, the coefficient patterns did reveal the ship tracks clearly, i.e., the indices were constructed from the more reliable data. The absence of ship-track signatures in the state-space patterns is due to the correlations between temperatures of pairs of cells that are not too far apart: the index must characterize them similarly.

The index time series for the four analyses were also compared. Index-by-index comparison exhibited somewhat larger differences than did the state-space patterns, but again, it would be difficult to discern which came from which analysis. Although the seasonal cycle was clearly dominant in the first indices of all three sets, and the ENSO signal could be seen in the next few indices, the temporal behavior of a particular index from one set generally appeared to be a linear combination of several indices from the other sets. Clearly the indices from one set characterize different aspects of the variability than do those from another set.

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 can 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.

7. Conclusions

The first conclusion of this paper is that seeking indices characterizing the well-determined aspects of climatic data leads to a natural extension of the familiar principal-component analysis. The error-covariance matrix characterizing the uncertainty of the data 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 variables, which are defined by their ability to explain total variability relatve 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, e.g., 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. 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 on results for a 126-month reanalysis of oceanic thermal data, is that although metric-based principal components are qualitatively similar to conventional principal components, they place more emphasis on the well-determined aspects of the data. 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 well-observed variability.

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

* To see why, replace ∇ with wavenumber in eq. (3).

which conclusions are specific to the sample, it is necessary to have additional data that can be used for verification.

A final comment 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

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

Derber, J. and A. Rosati, 1989. A global oceanic data assimilation system. *J. Phys. Oceanogr.* **19**, 1333–1347.

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. *J. Geophys. Res.* **92**, 14251–14270.

Hasselmann, K., 1979. On the signal-to-noise problem in atmospheric response studies. In: *Meteorology of tropical oceans*, 251–259. Royal Meteorological Society.

Hasselmann, K., 1993. Optimal fingerprints for the detection of time-dependent climate change. *J. 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. *Mon. Wea. Rev.* **123**, 460–481.

Jolliffe, I. T., 1986. *Principal component analysis.* Springer-Verlag, New York.

Kendall, M. G., A. Stuart, and J. K. Ord, 1983. *The advanced theory of statistics*, vol. 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.

Thacker, W. C., and R. Lewandowicz, 1996. Climatic indices, principal components, and the Gauss–Markov theorem. *J. Climate.*, in press.

Thacker, W. C., 1996. Climatic fingerprints, patterns, and indices. *J. Climate*, in press.

Woodruff, S. D., R. J. Slutz, R. L. Jenne and P. M. Steurer, 1987. A comprehensive ocean-atmosphere data set. *Bull. Amer. Meteor. Soc.* **22**, 1239–1250.