Sampling Errors in the Estimation of Empirical Orthogonal Functions

GERALD R. NORTH, THOMAS L. BELL AND ROBERT F. CAHALAN

Goddard Laboratory for Atmospheric Sciences, NASA/Goddard Space Flight Center, Greenbelt, MD 20771

FANTHUNE J. MOENG

Applied Research Corporation, Landover, MD 20771 and Goddard Laboratory for Atmospheric Sciences, NASA/Goddard Space Flight Center, Greenbelt, MD 20771

(Manuscript received 23 November 1981, in final form 5 April 1982)

ABSTRACT

Empirical Orthogonal Functions (EOF's), eigenvectors of the spatial cross-covariance matrix of a meteorological field, are reviewed with special attention given to the necessary weighting factors for gridded data and the sampling errors incurred when too small a sample is available. The geographical shape of an EOF shows large intersample variability when its associated eigenvalue is "close" to a neighboring one. A rule of thumb indicating when an EOF is likely to be subject to large sampling fluctuations is presented. An explicit example, based on the statistics of the 500 mb geopotential height field, displays large intersample variability in the EOF's for sample sizes of a few hundred independent realizations, a size seldom exceeded by meteorological data sets.

1. Introduction

Climate may be defined as the multivariate, multiple-time probability distribution of states of the ocean–ice–atmosphere system. A primary goal in modern climatology is the measurement and understanding of the parameters describing the stationary probability distribution. Secular changes in these parameters are of interest since they constitute a measure of climatic change. The task of determining the parameters is not easy since the natural variability of the climate system can lead to large sampling errors in estimates of properties of the parent distribution. Many climate variables have long time scales so that even a very long record may contain only a few statistically independent samples. For example, while the time between independent samples for estimating the mean of the 500 mb geopotential height field at a point is between 6 and 10 days, for sea-surface temperatures this time may be up to several months. The independent sampling time depends upon the parameter being estimated and the statistical model used for the variable in the estimation procedure (Leith, 1973; Jones, 1975).

Since their introduction in meteorology (Obukhov, 1947; Lorenz, 1956; Kutzbach, 1967), empirical orthogonal functions (EOF's) or principal components have become very popular as a convenient means of representing climatological fields. The EOF's are defined as the eigenvectors of the cross-covariance matrix between grid points. The eigenvectors are linear combinations of the individual station data with weights chosen so that the sums are uncorrelated with each other. These weights for the various stations can be represented as a contour map. EOF's (the weights) are a property of the parent probability distribution of the climate; hence, their forms are of great interest. We cannot know the exact EOF's for climate but must be satisfied with estimates of them based upon a finite number of independent realizations of the instantaneous state of the field. The purpose of this paper is to provide an estimate of the sampling errors encountered in some common climatological applications. While much of the underlying theory exists in the statistical literature (*e.g.*, Girschick, 1939; Anderson, 1963), it is usually not in a form convenient to the climatologist. We were guided by some earlier applied studies by Preisendorfer and Barnett (1977) and Buell (1979).

The outline of the paper is as follows: Section 2 introduces EOF's for continuous stochastic fields and establishes notation. Section 3 makes use of linear algebra to estimate the first-order perturbation of eigenvalues and eigenvectors of the covariance matrix due to sampling errors. Section 4 provides a numerical example of how a homogeneous field with parameters comparable to those encountered in meteorology generates rather large intersample variability even with a population of a few hundred. Conclusions and a "rule of thumb" are presented in the last section.

0027-0644/82/070699-08$06.00

2. Empirical orthogonal functions

In this section we define empirical orthogonal functions (EOF's), establish notation, and recall some properties of EOF's. We shall be examining a sequence of fields, each example or realization of which is a continuous function in space. The sequence may be constructed from a single field which evolves in time. If the correlation of the field with itself at a later time becomes small for long enough time intervals, then independent realizations may be obtained by choosing them at sufficiently long intervals. Just how long these intervals must be is often difficult to determine in advance since different spatial scales frequently have different autocorrelation times. We assume here that a long enough interval between realizations is used and that all realizations are independent. Measurement and analysis errors are ignored here although in practice they can be significant, we intend to study this additional effect in a later paper. The fluctuations from one realization to the next are just the natural variations of the field. By examining enough realizations and assuming the time series to be stationary (*i.e.*, ignoring seasonal and secular trends) we can learn as much as we please about the probability distribution underlying the process. In this section we discuss the EOF's for such a process. By considering continuous fields first, a natural choice of the weighting of data at discrete grid points can be formulated.

Consider a meteorological field $T(\mathbf{r})$ defined at the point $\mathbf{r}$ on a two-dimensional domain (the arguments given here are easily generalized to 3 dimensions). It will be assumed that the mean of $T(\mathbf{r})$ has been subtracted out so that $\langle T(\mathbf{r})\rangle \equiv 0$, where the angular brackets denote ensemble average or expectation value. The covariance of the field between any two points $\mathbf{r}$ and $\mathbf{r}'$ may be defined

$$\gamma(\mathbf{r}, \mathbf{r}') = \langle T(\mathbf{r})T(\mathbf{r}')\rangle \tag{1}$$

Principle components or EOF's are ordinarily defined as the eigenvectors of a finite-dimensional covariance matrix, but since the covariance matrix here is a continuous function of two vector variables it is more natural to consider $\gamma(\mathbf{r}, \mathbf{r}')$ as the kernel of an integral equation eigenvalue problem. The formalism is often referred to as the Karhunen-Loève expansion of T, and the basis functions or EOF's are the eigenfunctions of the integral equation

$$\frac{1}{A}\int_A \gamma(\mathbf{r}, \mathbf{r}')\phi_\alpha(\mathbf{r}')d\mathbf{r}' = \lambda_\alpha\phi_\alpha(\mathbf{r}), \tag{2}$$

where A is the area of the domain, λ_α is the eigenvalue associated with the eigenfunction $\phi_\alpha(\mathbf{r})$, and α is an integer index. The integral extends over the finite domain of interest. It is obvious that EOF shapes will depend upon the geometry of the domain.

Since the covariance γ is bounded, continuous and symmetric in its arguments, several properties of the system follow (Courant and Hilbert, 1953):

1) The eigenvalues are positive and discrete, although some eigenfunctions may have the same eigenvalue (degeneracy). In general there will be infinitely many.

2) The functions $\phi_\alpha(\mathbf{r})$ may be normalized so that they form an orthonormal set:

$$\frac{1}{A}\int_A \phi_\alpha(\mathbf{r})\phi_\beta(\mathbf{r})d\mathbf{r} = \delta_{\alpha\beta}\,. \tag{3}$$

3) The functions $\phi_\alpha(\mathbf{r})$ are a complete set:

$$\sum_{\alpha=1}^{\infty} \phi_\alpha(\mathbf{r})\phi_\alpha(\mathbf{r}') = \delta(\mathbf{r} - \mathbf{r}'). \tag{4}$$

It follows from (3) and (4) that each realization of the field $T(\mathbf{r})$ can be expanded into an infinite series of the $\phi_\alpha(\mathbf{r})$

$$T(\mathbf{r}) = \sum_{\alpha=1}^{\infty} T_\alpha\phi_\alpha(\mathbf{r}), \tag{5}$$

where

$$T_\alpha = \frac{1}{A}\int_A \phi_\alpha(\mathbf{r})T(\mathbf{r})d\mathbf{r}. \tag{6}$$

Each realization of the field $T(\mathbf{r})$ is represented via (6) by a set of mode amplitudes T_α, $\alpha = 1, 2, \ldots$. The statistics of $T(\mathbf{r})$ require that the T_α have a probability distribution satisfying

$$\langle T_\alpha\rangle = 0$$

and

$$\langle T_\alpha T_\beta\rangle = \lambda_\alpha\delta_{\alpha\beta}\,, \tag{7}$$

i.e., the different components of the series (5) are uncorrelated with each other and the variance of the mode α is just the eigenvalue λ_α associated with mode α. Generating random fields using Eqs. (5) and (7) can be useful for many different purposes such as that of stochastic modeling (see North and Cahalan, 1981) or that of estimating sampling errors as in Section 3 of this paper.

It can be shown directly from (5) and (7) that the variance of the field at a point is

$$\langle T(\mathbf{r})^2\rangle = \sum_{\alpha=1}^{\infty} \lambda_\alpha\phi_\alpha(\mathbf{r})^2 \tag{8}$$

and the average variance integrated over the domain is

$$\frac{1}{A}\int_A \langle T(\mathbf{r})^2\rangle d\mathbf{r} = \sum_{\alpha=1}^{\infty} \lambda_\alpha\,. \tag{9}$$

Because of Eq. (9) we may think of λ_α as the portion of total variance "explained" by the EOF $\phi_\alpha(\mathbf{r})$.

It is convenient to label the eigenfunctions so that the eigenvalues are in descending order, i.e.,

$$\lambda_1 > \lambda_2 > \lambda_3 \cdots\,. \tag{10}$$

We can then consider truncating the sequence at some level n by retaining only the first n terms in the

formulas (5), (8) and (9). By the ordering (10) we are assured of keeping the largest contributors to the total variance (9). The property of EOF's which makes them special is that they are the *optimum* set of basis functions for a given truncation n. That is, for a given n no other basis set can explain more of the average variance, i.e., generate a larger subsum approximating (9). The proof of this theorem can be found in various forms in the literature (e.g., Lorenz, 1956; Davis, 1976).

In practice we never have continuous measurements of $T(\mathbf{r})$ at each point $\mathbf{r}$ but must be satisfied with a finite and sometimes irregularly spaced discrete grid of m points in the domain A. If it is our aim to learn about the intrinsic EOF's $\phi_\alpha(\mathbf{r})$, independent of the grid, we must find an appropriate scaling which eliminates the geometrical effects of grid spacing. The integral equation (2) may be approximated by the finite sum

$$\frac{1}{A}\sum_{j=1}^{m}\gamma(\mathbf{r}_i, \mathbf{r}_j)(\Delta A)_j\phi_\alpha(\mathbf{r}_j) = \lambda_\alpha\phi_\alpha(\mathbf{r}_i), \qquad (11)$$

where $\mathbf{r}_i$, $i = 1, 2 \cdots, m$ are the sequence of m individual grid points and $(\Delta A)_i$ is an area element containing the point $\mathbf{r}_i$. The shapes of the elements might for example be chosen as polygons. Multiplying (11) through by $(\Delta A)_i^{1/2}$ allows us to define a new symmetric matrix and the associated finite-dimensional eigenvalue problem (Buell, 1971, 1978)

$$\sum_{j=1}^{m}\Gamma_{ij}\Phi_\alpha(j) = \lambda_\alpha\Phi_\alpha(i), \qquad (12)$$

where

$$\Gamma_{ij} = (\Delta A)_i^{1/2}\gamma(\mathbf{r}_i, \mathbf{r}_j)(\Delta A)_j^{1/2}/A, \qquad (13)$$

$$\Phi_\alpha(i) = \phi_\alpha(\mathbf{r}_i)(\Delta A)_i^{1/2}/A^{1/2}. \qquad (14)$$

Although the metric factors $(\Delta A)_i^{1/2}$ enter the definitions of Γ_{ij} and $\Phi_\alpha(i)$, they do not affect the scaling of the eigenvalues λ_α. Hence, the eigenvalues computed from (12) are approximations to the first few eigenvalues of the continuous field EOF's. Approximations to the true continuous EOF's can be obtained by solving (14) for $\phi_\alpha(\mathbf{r}_i)$. Note that the resulting EOF shapes and eigenvalues λ_α will be different from the true EOF shapes and eigenvalues unless the metric factor is included. In what follows we shall see that the sampling errors incurred in estimating EOF's depend strongly upon the spectrum of eigenvalues. Since we want to estimate the EOF's and eigenvalues characterizing the true field independent of the station or grid point locations it is important that we have an idea about the spacing of the true eigenvalues. The metric factor introduced here insures that we at least have an approximation to the true eigenvalues which in principle becomes exact for a large enough number of grid points. In an actual case one should be sure that the spacing of grid points is smaller than an autocorrelation length for the field; otherwise important contributions to the variance at smaller scales will be lost and the result will include a kind of aliasing error in the retained modes. For a further discussion of this point we recommend the article by Buell (1978).

Before turning to the sampling problem we wish to mention a few peculiarities connected with the physical interpretation of EOF's for real physical fields. An obvious difficulty arises in physically interpreting an EOF if it is not even well-defined intrinsically. This can happen for instance if two or more EOF's have the same eigenvalue. It is easily demonstrated that any linear combination of the members of the degenerate multiplet is also an EOF with the same eigenvalue. Hence, in the case of a degenerate multiplet one can choose a continuous range of linear combinations which will all satisfy the Karhunen-Loève Equation (2) and which are indistinguishable in terms of their contribution to the average variance (9). This ambiguity for degenerate multiplets is central to understanding the sampling theory in this paper. Such degeneracies often arise from a symmetry in the problem but they can be present for no apparent reason (accidental degeneracy).

There are a few cases in which the EOF's can be found analytically and since these are instructive as well as useful we mention a few here:

1) Rotational invariance of statistics on the sphere (Obukhov, 1947; North and Cahalan, 1981) leads to the spherical harmonics as EOF's. Similarly on the circle the EOF's are the sine and cosine Fourier basis.

2) Translational invariance on the infinite line leads to the Fourier integral representation and is the foundation of spectral analysis of stationary time series (Papoulis, 1965).

3) Linear mechanical systems such as a drum head observed at a fixed interval after evolving from random initial conditions lead to the normal modes as the EOF's (North, 1982).[1]

3. Perturbation of EOF's by sampling errors

The EOF's of a stochastic field are a property of the underlying multi-dimensional probability distribution. Any estimate of the EOF's from a finite number of realizations will be subject to sampling errors. It is obviously important to establish some estimate of the sampling errors involved before interpreting sample EOF's. In this section we present an estimate of the errors which is first order in $(N^{-1})^{1/2}$, where N is the number of realizations. From this error es-

[1] Preisendorfer has also found examples of linear mechanical systems whose EOF's can be found. The work is described in "Principal Components and the Motions of Simple Dynamical Systems," Scripps Institution of Oceanography. Reference Series, 79-11, April, 1979.

timate we derive a rule of thumb determining whether a sample EOF is expected to be a faithful representation of the true EOF.

The familiar methods of linear analysis can be used to estimate the shifts in eigenvectors and eigenvalues when the covariance matrix they are derived from has added to it a small symmetric perturbation matrix (Mathews and Walker, 1965; Anderson, 1963). Let the ij element of the exact m gridpoint covariance matrix be Γ_{ij}. (We use the discrete index notation in this section for convenience; we use Greek letters for population quantities and Latin letters for sample quantities.) We denote the exact eigenvectors of Γ by Φ_α, corresponding to eigenvalues $\lambda_\alpha \alpha = 1, 2 \cdots m$. The eigenvalue equation is

$$\sum_{j=1}^{m} \Gamma_{ij}\Phi_\alpha(j) = \lambda_\alpha \Phi_\alpha(i). \tag{15}$$

In practice one has only an estimate of Γ_{ij}. In fact we may write

$$S_{ij} = \Gamma_{ij} + \epsilon V_{ij}, \tag{16}$$

where for a particular sample of N realizations, S_{ij} is the (symmetric) sample covariance matrix, ϵV_{ij} represents the sampling error with ϵ a parameter of order $(2/N)^{1/2}$, and V_{ij} is of the order of

$$V_{ij} \approx [(\Gamma_{ii}\Gamma_{jj} + \Gamma_{ij}^2)/2]^{1/2}. \tag{17}$$

These last statements come from the standard error of a covariance between Gaussian random variables.

From the sample covariances S_{ij} one can determine sample eigenvectors f_α and eigenvalues l_α by solving

$$\sum_{j=1}^{m} S_{ij} f_\alpha(j) = l_\alpha f_\alpha(i). \tag{18}$$

We wish to know how the sample eigenvectors $f_\alpha(i)$ and eigenvalues l_α differ from the exact ones $\Phi_\alpha(i)$ and λ_α. We use the standard method of expanding all perturbed quantities into m power series in the small parameter ϵ:

$$f_\alpha(i) = \Phi_\alpha(i) + \epsilon f_\alpha^{(1)}(i) + \epsilon^2 f_\alpha^{(2)}(i) + \cdots, \tag{19}$$

$$l_\alpha = \lambda_\alpha + \epsilon l_\alpha^{(1)} + \epsilon^2 l_\alpha^{(2)} + \cdots. \tag{20}$$

We shall be content with the first-order corrections in this paper, although equations for higher orders can be developed (Anderson, 1963). Such a crude estimate should be sufficient in many climatological applications. After inserting the expansions (19) and (20) into (18) and collecting the coefficients of the various powers of ϵ, we may make use of the orthogonality of the $f_\alpha(i)$ to arrive at the first-order estimates

$$l_\alpha^{(1)} = \sum_{i,j} \Phi_\alpha(i) V_{ij} \Phi_\alpha(j), \tag{21}$$

$$\equiv (\Phi_\alpha, V\Phi_\alpha), \tag{22}$$

the latter in an inner product notation, and

$$f_\alpha^{(1)}(i) = \sum_{\beta \neq \alpha} \frac{(\Phi_\beta, V\Phi_\alpha)}{\lambda_\beta - \lambda_\alpha} \Phi_\beta(i), \tag{23}$$

where we have again used the inner product notation introduced in (22).

We proceed now to find simple practical approximations to these first-order quantities to facilitate their use. First note the qualitative difference between the two expressions: At this order the shift in eigenvector depends strongly upon the *spacing* of eigenvalues, whereas the shift of eigenvalues does not.

To first order the *shift* in eigenvalue λ_α is given by

$$\delta\lambda_\alpha \equiv \epsilon l_\alpha^{(1)} \approx \lambda_\alpha (2/N)^{1/2}, \tag{24}$$

where we have estimated the inner product (22) by λ_α and ϵ by $(2/N)^{1/2}$ based upon (17). The difference denominators in (23) largely determine whether $\delta\Phi_\alpha$ is large. If no other eigenvalue is close to λ_α the error is very small for f_α. On the other hand, if another eigenvalue $\lambda_{\alpha'}$ is very close to λ_α then we can expect the term $\beta = \alpha'$ to dominate the sum in (23) and we obtain the estimate

$$\delta\Phi_\alpha(i) = \epsilon f_\alpha^{(1)}(i) \approx (2/N)^{1/2} \frac{\lambda_\alpha}{\Delta\lambda_\alpha} \Phi_{\alpha'}(i), \tag{25}$$

where $\Delta\lambda_\alpha$ is the *spacing* $\lambda_\alpha - \lambda_{\alpha'}$, and we have estimated the inner product in (23) by λ_α. Making use of (24) we may write

$$\delta\Phi_\alpha(i) \approx \frac{\delta\lambda_\alpha}{\Delta\lambda_\alpha} \Phi_{\alpha'}(i). \tag{26}$$

In other words if the sampling error in the eigenvalue is comparable to the distance to a nearby eigenvalue, then the sampling errors in the EOF will be comparable to the "nearby" EOF. The instability due to sampling when the eigenvalues are closely spaced is, of course, implicit in the statistical literature cited earlier (Anderson, 1963).

A physical interpretation of this result is immediately apparent. If the sampling error in eigenvalue is comparable to the spacing, a kind of "effective degeneracy" occurs. We have already seen that degeneracy leads to an intrinsic ambiguity in defining the EOF, since any linear combination of the possible eigenvectors in the multiplet (subspace) is also an eigenvector. But even if no degeneracy actually exists, some eigenvalues may be close enough to each other that sampling errors lead to an effective degeneracy and mixing occurs. That is, a particular sample will lead to one linear combination and another sample may pick out a drastically different linear combination of the nearby eigenvectors. The result is wildly differing patterns from one sample to the next.

The derivation of formulas (22) and (23) does not hold if there is degeneracy in the unperturbed case. In this case it is clear that one must choose the un-

perturbed degenerate eigenvectors carefully to avoid the singularity in the sum (23). This can be done in practice but need not concern us here, since the qualitative effects are clear from the discussion above.

In the next section we present a numerical example that illustrates the preceding formal discussion.

4. Numerical example

In this section we illustrate the sampling error problem by considering a numerical example with a statistical structure reasonably close to a climatological field. Although we use statistics derived from the statistics of the 500 mb geopotential height field, the calculations are not to be construed as a theory or empirical study of the EOF's of the 500 mb height field.

We shall construct a Karhunen-Loève expansion to generate independent realizations of a stochastic field having homogeneous statistics on the sphere. As noted in Section 2, this can be accomplished by using the complex spherical harmonics as a basis set

$$F(\hat{\mathbf{r}}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} f_{lm} Y_l^m(\hat{\mathbf{r}}), \tag{27}$$

where the radial unit vector $\hat{\mathbf{r}}$ denotes position on the sphere. The spherical harmonics are standard $Y_l^m(\hat{\mathbf{r}}) = N_{lm} P_l^{|m|}(\sin\theta) e^{im\phi}$, $P_l^{|m|}$ are the associated Legendre polynomials, θ latitude, ϕ longitude and the $N_{lm} = N_{l,-m}$ are real constants chosen so that the spherical harmonics have unit normalization: $\int d\Omega |Y_l^m|^2 = 1$. The f_{lm} are random complex variables constrained by the realness of $F(\hat{\mathbf{r}})$ to satisfy $f_{lm}^* = f_{l,-m}$ (asterisks denote complex conjugation). It can be shown (for details see North and Cahalan, 1981; North *et al.*, 1982) that the f_{lm} are to be drawn from a probability distribution such that $\langle f_{lm} \rangle = 0$ and

$$\langle f_{lm}^* f_{l'm'} \rangle = \sigma_l^2 \delta_{ll'} \delta_{mm'} . \tag{28}$$

We have normalized the spectrum σ_l^2 to give unit variance at each point while the wavenumber dependence is at our discretion. For the purposes of this example we have chosen the spherical wavenumber dependence shown in Fig. 1. The spectrum is similar to the one exhibited by the 500 mb geopotential height field in midlatitudes (North *et al.*, 1982). Although we could study sampling errors for the $Y_l^m(\hat{\mathbf{r}})$ by looking at sample estimates of the covariance of $F(\hat{\mathbf{r}})$ over the whole sphere, to create a situation more common to empirical studies we have restricted the data to a rectangular patch on the sphere of dimensions 20–60°N in latitude and 100° of width in longitude. Such a domain is large enough to be of interest climatologically and is typical of a regional study. We shall examine the EOF's of the field $F(\hat{\mathbf{r}})$ generated by (27) restricted to the patch. The EOF's in this case are not the spherical harmonics but an orthonormal set defined only on the patch and satisfying (2) with the integral so restricted.

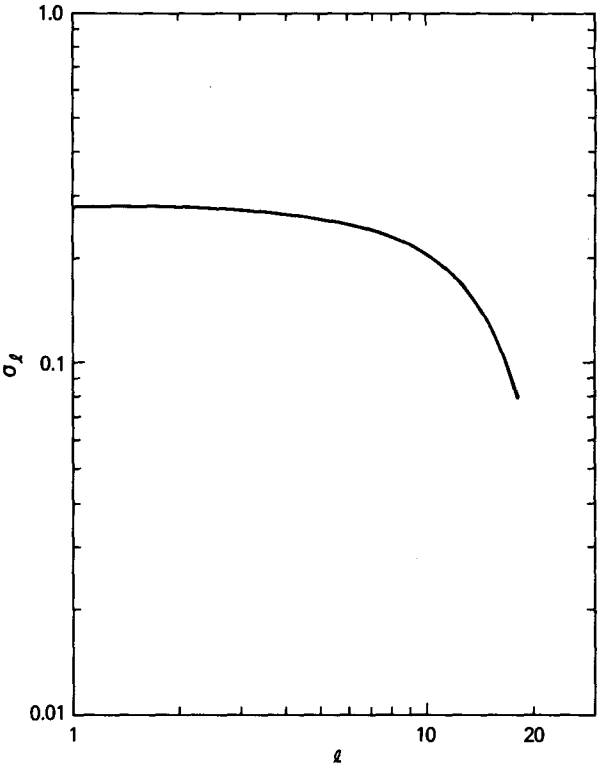

FIG. 1. Standard deviation spectrum of spherical harmonic components of the random field used in covering the sphere for the numerical example of Section 4. The field is defined by (27) and (28).

The patch (20–60°N and 100° longitude) is represented by a grid of 189 points. In our computations we have omitted the area metric factors in the covariance matrix (13), hence we must caution the reader again about relating the results directly to the 500 mb surface. The results we obtain satisfactorily illustrate the sampling theory.

The exact covariance matrix can be computed from the exact covariance statistics with the formula (27) and (28) for the grid points on the patch. After this is done the exact eigenvectors may be computed numerically. Fig. 2 shows coarse contour maps of the first four EOF's for this field. Note that the EOF's are left-right symmetric (because of the left-right symmetry of the patch) but not up-down symmetric because of the spherical geometry. Also shown above each pattern is the corresponding eigenvalue λ_α. It is especially noteworthy that $(\lambda_3 - \lambda_4)/\lambda_3 \approx 0.02$ is small compared to the other separations. A sample covariance matrix can be computed from a collection of N realizations of the field on the patch, leading to a set of sample EOF's. Figs. 3a–c show the first four sample EOF's from three experiments, each experiment having $N = 300$ realizations. Evidently the six sets of sample EOF's vary considerably from

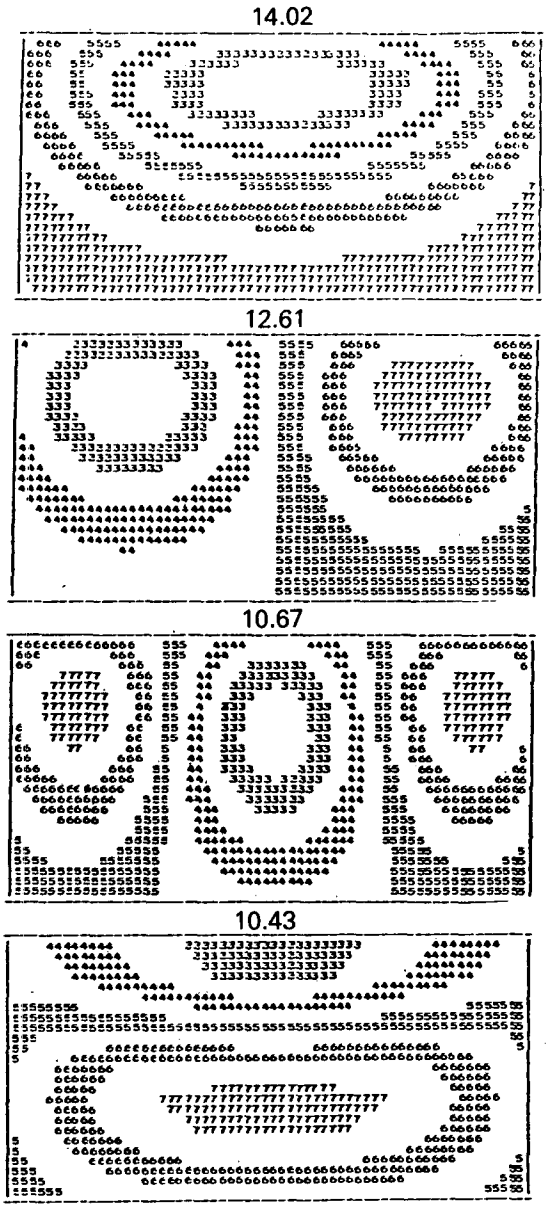

FIG. 2. The first four exact EOF patterns for the field (27) restricted to a patch on the sphere (20 to 60°N and 100° of width in longitude). The number above each contour map is the exact eigenvalue associated with the corresponding EOF pattern.

each other and from the exact EOF's in Fig. 2. For $N = 300$ we can estimate from (24) that the sampling error in the first four eigenvalues is typically about $12(2/300)^{1/2} = 0.98$ or

$$\delta\lambda \approx 1.0, \quad N = 300.$$

In Fig. 4 we illustrate schematically the spacing of eigenvalues with error bars indicative of the sampling error $\pm\delta\lambda$. To the extent that the large N treatment of the statistics is valid, sample eigenvalues should lie within the error bars ~68% of the time.

Fig. 4 also shows the error bars for $N = 1000$, and a corresponding group of $N = 1000$ sample EOF's are mapped in Fig. 5a, b, c. As can be seen in Fig. 4 the sampling errors in this case are smaller than the spacing for $\alpha = 1$ and 2 but still larger than the spacing for $\alpha = 3$ and 4. This is reflected in the sample EOF's of Fig. 5. In fact many thousands of realizations would be necessary to resolve the EOF's for $\alpha = 3$ and 4.

5. Conclusions

In this paper we have reviewed some properties of EOF's, shown how they may be computed from grid point information, and used standard linear analysis to estimate sampling errors. The discussion has focussed on the application to climatology where the EOF's are considered to be a manifestation of the parent probability distribution. Any secular changes in time of the EOF's would indicate climatic change. Unfortunately, the problem of detecting such a change is complicated by the sampling errors incurred in estimating the shapes of the EOF's. We suggest a rule of thumb for estimating the sampling errors. The rule is simply that if the sampling error of a particular eigenvalue $\lambda[\delta\lambda \sim \lambda(2/N)^{1/2}]$ is comparable to or larger than the spacing between λ and a neighboring eigenvalue, then the sampling errors for the EOF associated with λ will be comparable to the size of the neighboring EOF. The interpretation is that if a group of true eigenvalues lie within one or two $\delta\lambda$ of each other, then they form an "effectively degenerate multiplet," and sample eigenvectors are a random mixture of the true eigenvectors. The ambiguity in choosing the proper linear combination within the multiplet leads to enhanced sampling error.

The example shown in this paper suggests that in many cases of climatological interest the sampling errors are unacceptably large for samples of a few hundred independent realizations. In fact, depending upon the spectrum of eigenvalues thousands of realizations may be necessary even to resolve the correct number of maxima and minima for the EOF's. Each EOF will have different sampling requirements depending upon the nearness of neighboring eigenvalues. Wallace and Gutzler (1981) provide another example of the variability of EOF patterns from one sample to another in their investigation of EOF's for the Northern Hemisphere wintertime geopotential height field. The rule of thumb described here indicates that the first two pairs of EOF's derived from the record available to them are likely to be mixed by sampling fluctuations, and this is borne out in their analysis of a similar but independent record.

Of course, in many applications one EOF will stand out, accounting for a large part of the variance. Such "stand out" EOF's will have small sampling errors ($\Delta\lambda$ large, $\delta\lambda$ small) and often can be related to some physical inhomogeneity in the problem. We

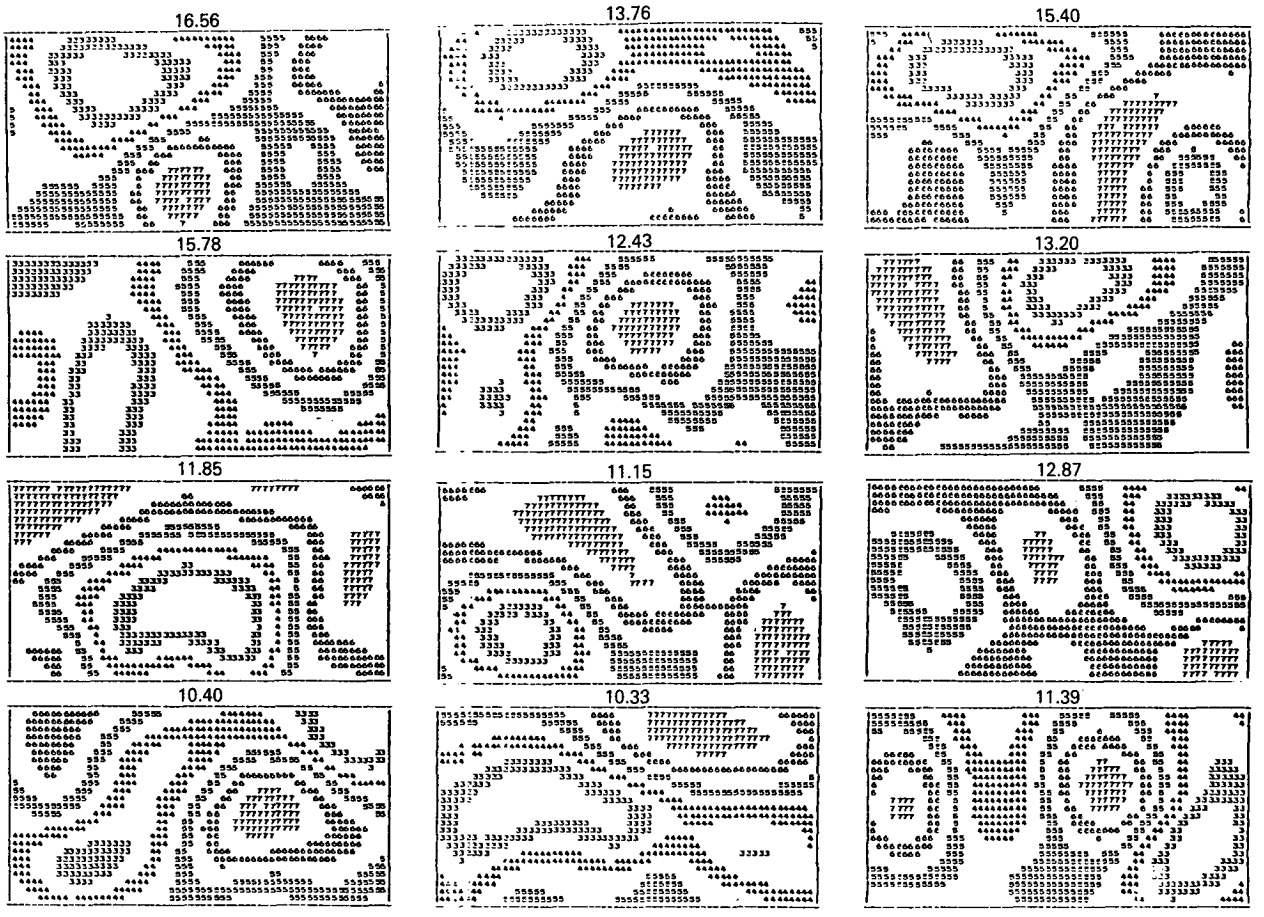

FIG. 3a, b, c. The columns are three separate $N = 300$ sample estimates of the exact EOF's shown in Fig. 2. Above each sample EOF is the corresponding sample eigenvalue. Note that the intersample variability is so large that the true patterns are scarcely recognizable.

have examined some cases (not discussed here) of this type with an "island" in the patch contributing enhanced variability within its borders. Our rule of thumb worked as expected with a stand-out EOF showing small sampling errors but the others showing errors in accord with the rule.

We should note that often one is not especially interested in the *shapes* of the EOF's per se. In many applications one may wish only to have a convenient basis set for representing data. The sample EOF's still form a complete basis set and a subsum may account for about as much variance as the corresponding subsum of true EOF's. The problem focussed upon in this paper occurs when near multiplets get mixed by sampling error. So long as all of the mixed multiplet members are included there is no special problem in representing data at the same level of fit. However, in choosing the point of truncation, one should take care that it does not fall in the middle of an "effective multiplet" created by the sampling problem, since there is no justification for choosing to keep part of the multiplet and discarding the rest. Other than this, the rule of thumb unfortunately provides no guidance in selecting a truncation point

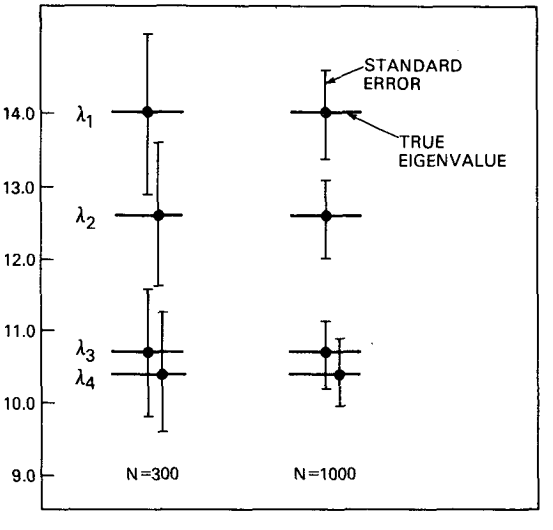

FIG. 4. Schematic diagram of the first four eigenvalues λ_α corresponding to the exact patterns of Fig. 2. The error bars represent the standard error (one standard deviation error due to sampling) for each eigenvalue. For $N = 300$ (left) the standard error is comparable to or larger than the eigenvalue spacing for all four eigenvalues, whereas for $N = 1000$ (right) λ_1 and λ_2 are resolved but λ_3 and λ_4 still are close compared to the sampling error.

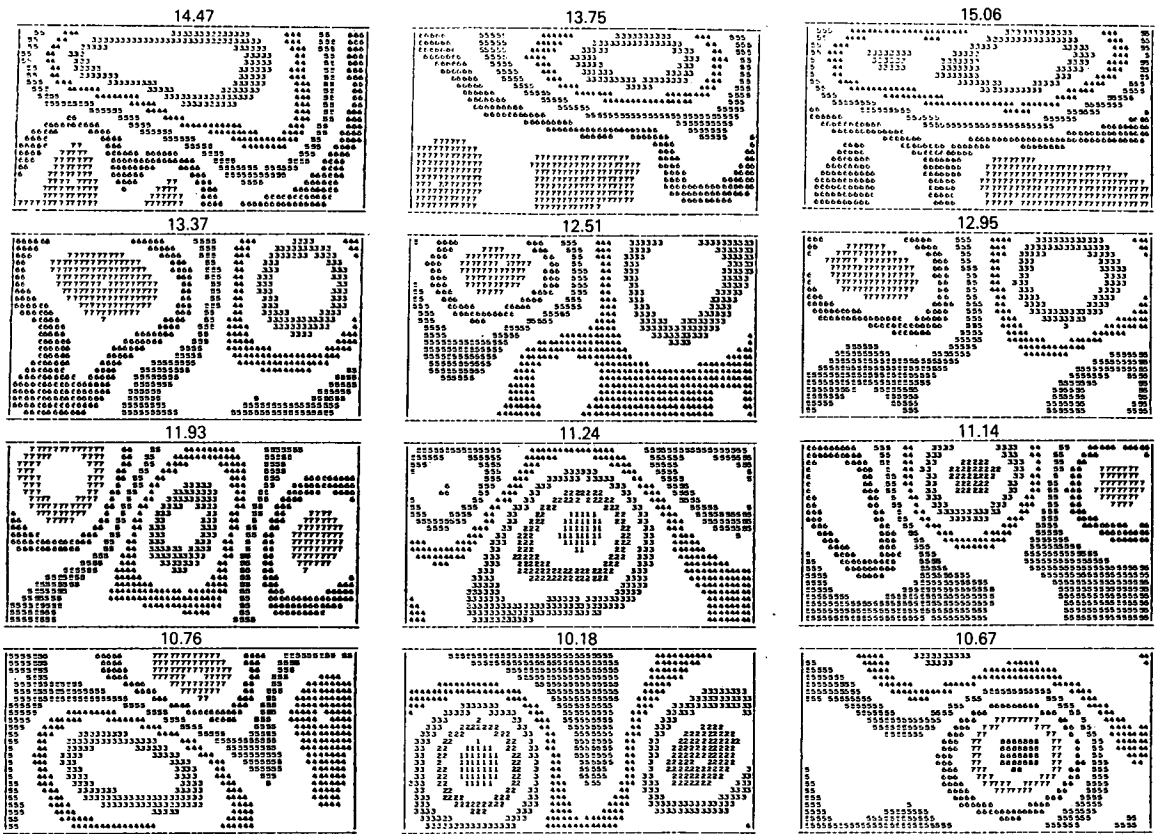

FIG. 5a, b, c. As in Fig. 3 except $N = 1000$. Note that in this case EOF's one and two look like their exact counterparts in Fig. 2, but EOF's three and four are still mixed.

for using a subset of EOF's to represent a large data set efficiently. Additional assumptions about the nature of the "noise" in the data must be made.

Acknowledgments. We are grateful to C. E. Buell and R. D. Preisendorfer for their helpful suggestions for improving the manuscript.

REFERENCES

Anderson, T. W., 1963: Asymptotic theory for principal component analysis. *Ann. Math. Stat.*, **34,** 122–148.

Buell, C. E., 1971: Integral equation representation for factor analysis. *J. Atmos. Sci.*, **28,** 1502–1505.

——, 1978: The number of significant proper functions of two-dimensional fields. *J. Appl. Meteor.*, **17,** 717–722.

——, 1979: On the physical interpretation of empirical orthogonal functions. *Preprints 6th Conf. on Probability and Statistics in Atmospheric Sciences*, Banff, Amer. Meteor. Soc., 112–117.

Courant, R., and D. Hilbert, 1953: *Methods of Mathematical Physics*, Vol. 1. Interscience, 560 pp.

Davis, R. E., 1976: Predictability of sea surface temperature and sea level pressure anomalies over the North Pacific Ocean. *J. Phys. Oceanogr.*, **6,** 249–266.

Girshick, M. A., 1939: On the sampling theory of roots of determinantal equations. *Ann. Math. Stat.*, **10,** 203–224.

Jones, R. H., 1975: Estimating the variance of time averages. *J. Appl. Meteor.*, **14,** 159–163.

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

Leith, C. E., 1973: The standard error of time-average estimates of climatic means. *J. Appl. Meteor.*, **12,** 1066–1069.

Lorenz, E. N., 1956: Empirical orthogonal functions and statistical weather prediction. Sci. Rep. No. 1, Statist. Forecasting Proj., Dept. Meteor., MIT, 49 pp.

Mathews, J. and R. L. Walker, 1965: *Mathematical Methods of Physics.* W. A. Benjamin, 475 pp.

North, G. R., 1982: Empirical orthogonal functions and normal modes. To appear in *Amer. J. Phys.*

——, and R. F. Cahalan, 1981: Predictability in a solvable stochastic climate model. *J. Atmos. Sci.*, **38,** 504–513.

——, F. J. Moeng, T. L. Bell and R. F. Cahalan, 1982: The latitude dependence of the variance of zonally averaged quantities. *Mon. Wea. Rev.*, **110,** 319–326.

Obukhov, A. M., 1947: Statistically homogeneous fields on a sphere. *Usp. Mat. Nauk*, **2,** 196–198.

Papoulis, A., 1965: *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, 583 pp.

Preisendorfer, R. W., and T. P. Barnett, 1977: Significance tests for empirical orthogonal functions. *Preprints 5th Conf. on Probability and Statistics in Atmospheric Sciences*, Las Vegas, Amer. Meteor. Soc., 169–172.

Wallace, J. M., and D. S. Gutzler, 1981: Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Mon. Wea. Rev.*, **109,** 784–812.