A Significance Test for Principal Components Applied to a Cyclone Climatology

JAMES E. OVERLAND AND R. W. PREISENDORFER

Pacific Marine Environmental Laboratory/NOAA, Seattle, WA 98105

(Manuscript received 27 April 1981, in final form 4 December 1981)

ABSTRACT

A technique is presented for selection of principal components for which the geophysical signal is greater than the level of noise. The level of noise is simulated by repeated sampling of principal components computed from a spatially and temporally uncorrelated random process. By contrasting the application of principal components based upon the covariance matrix and correlation matrix for a given data set of cyclone frequencies, it is shown that the former is more suitable to fitting data and locating the individual variables that represent large variance in the record, while the latter is more suitable for resolving spatial oscillations such as the movement of primary storm tracks.

1. Introduction

During the development of a set of indices to relate cyclone climatology to interannual variability of sea-ice extent in the Bering Sea (Overland and Pease, 1982), an empirical orthogonal function (EOF) analysis was performed on the spatial correlation matrix of the number of cyclones transiting through 56 2° latitude × 4° longitude cells during the ice-growth season, October–February, for 23 years—1957/58 through 1979/80. The first three EOF's of the Bering Sea data set are shown in Fig. 1. The first EOF shows a northwest-southeast negative correlation of cyclone counts which is similar to variation in cyclone tracks implied by the North Pacific Oscillation (Walker and Bliss, 1932). Plausible physical interpretation can be suggested for the second and third EOF's, particularly the third, which shows a Siberian versus Alaskan preference for northward propagating cyclones.

The need for some form of EOF selection rule in geophysical studies has become apparent in recent years as an increasing number of researchers have used principal component analysis to study large data sets in meteorological and oceanographic settings. This method of analysis unfortunately is potentially dangerous in the sense that too much is often required of it or, worse yet, read into its results. The situation is similar to inferring geophysical conclusions from a correlation, regression analysis or Fourier spectral decomposition of a time series. The intent of this note is to demonstrate a selection rule as applied to the Bering Sea data set for determining if the eigenvalues of an EOF analysis of a geophysical data set can be distinguished from those produced from a spatially and temporally uncorrelated random process. We interpret the rule as indicating that physical interpretation of EOF's is suspect when the corresponding geophysical eigenvalue is less than one generated from the random data set, unless alternate significance measures are employed. The selection rule presented and applied below is but one of a set of rules recently devised and tested in Preisendorfer *et al.* (1981). In the latter study there was also a brief review of some other selection rules devised in the field of psychometry where similar problems have been encountered for at least 50 years.

2. Application of the selection rule

Preisendorfer and Barnett (1977) suggested a Monte Carlo technique for selecting eigenvalues in an EOF analysis for which the geophysical signal is above the level of noise. Let d_j, $j = 1, \ldots, p$, be the eigenvalues of the spatial correlation matrix $\Phi(x, x')$ computed from n data sets, such that $d_1 > d_2 \cdots > d_p$. We form the normalized eigenvalue statistic

$$T_j = d_j(\sum_{j=1}^{p} d_j)^{-1}, \quad j = 1, \ldots, p. \qquad (1)$$

We now form the null hypothesis that the geophysical data set is randomly drawn from a population of uncorrelated gaussian variables. Use a random number generator to generate independent sequences of length n for p independent gaussian variables of zero mean and unit variance and compute the correlation matrix. Compute the eigenvalues of the correlation matrix and repeat the experiment (say) one hundred times. If δ_j^r, $j = 1, \cdots p$, is the set of eigenvalues produced by the rth Monte Carlo experiment, the statistic analogous to (1) is

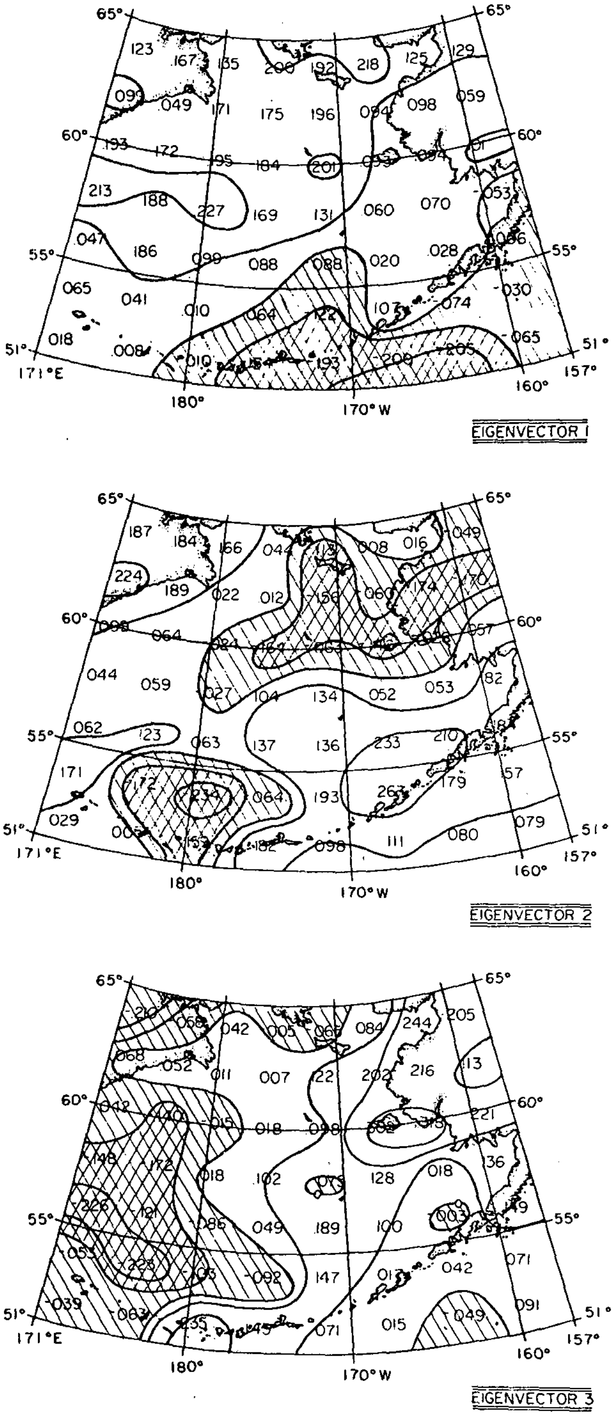

FIG. 1. Contours of the first three empirical orthogonal functions (EOF) of the correlation matrix of the Bering Sea cyclone data set.

$$U_j^r = \delta_j^r\left(\sum_{j=1}^{p} \delta_j^r\right)^{-1}, \quad \begin{array}{l} j = 1, \ldots, p, \\ r = 1, \ldots, 100. \end{array} \qquad (2)$$

For fixed j, order the U_j^r so that

$$U_j^1 \leqslant U_j^2 \leqslant \cdots \leqslant U_j^{100}.$$

TABLE 1. Values of U_j^{95} for selected values of p and n.

		n				
P		20	60	100	200	1000
9	$j = 1$	29.78	29.78	18.33	15.89	13.11
	$j = 2$	22.00	17.33	15.67	14.33	12.44
	$j = 3$	17.89	15.00	14.11	13.22	12.11
	$j = 4$	14.67	13.22	12.78	12.33	11.67
	$j = 5$	11.56	11.44	11.56	11.56	11.33
36	$j = 1$	15.00	8.69	6.94	5.64	3.94
	$j = 2$	12.67	7.58	6.47	5.14	3.78
	$j = 3$	10.94	7.03	5.92	4.89	3.69
	$j = 4$	9.83	6.47	5.56	4.69	3.58
	$j = 5$	8.72	6.03	5.25	4.47	3.50
64	$j = 1$	12.00	6.50	5.03	3.86	2.47
	$j = 2$	10.69	5.89	4.61	3.58	2.38
	$j = 3$	9.50	5.38	4.34	3.44	2.33
	$j = 4$	8.78	5.08	4.19	3.28	2.27
	$j = 5$	7.91	4.77	3.91	3.17	2.23
100	$j = 1$	10.45	5.31	3.98	2.91	1.74
	$j = 2$	9.29	4.81	3.72	2.75	1.69
	$j = 3$	8.57	4.55	3.55	2.65	1.66
	$j = 4$	7.95	4.30	3.39	2.56	1.62
	$j = 5$	7.39	4.14	3.23	2.47	1.59

Compare T_j with the distribution represented by the pair of values $[U_j^5, U_j^{95}]$ for each $j = 1, \ldots, p$. Rule N (Preisendorfer and Barnett, 1977) is given by the following: Terminate the sequence T_j at $j = p'$, where p' is the largest integer m such that T_m exceeds U_m^{95}. A table of U_j^{95}, $j = 1, 5$, is given in Table 1 for a range of values of n and p.

For the first five EOF's of the Bering Sea data set, Table 2 lists T_j, the normalized eigenvalues, and T_j/U_j^{95}, the ratio which determines the application of rule N. The first EOF represents 22.3% of the variance while the second, third and fourth represent 9.9, 9.1 and 8.1%. While EOF's 2–5 represent a significant percentage of the total variance of the original data set fit by these functions, they fail rule N. We conclude that except for the first, we cannot distinguish the present meteorological eigenvalues from ones generated by a spatially and temporally uncorrelated random process.

As an additional example we apply rule N to the data set generated by Hayden (1981) who expanded annual cyclone frequencies for 74 cells covering

TABLE 2. Summary of normalized eigenvalues of the correlation matrix of the Bering Sea cyclone data set; $p = 56$, $n = 23$. Rule N for selection of geophysical eigenvalues is satisfied for values of $T_j/U_j^{95} > 1.0$.

	j				
	1	2	3	4	5
T_j (×100)	22.3	9.9	9.1	8.1	6.5
T_j/U_j^{95}	1.91	0.97	0.99	0.98	0.86

TABLE 3. Summary of normalized eigenvalues of the correlation matrix from the data set of Hayden (1981); $p = 74$, $n = 96$.

	j			
	1	2	3	4
T_j (×100)	28.0	17.3	6.6	5.5
T_j/U_j^{95}	5.77	3.81	1.53	1.40

eastern North America and the western North Atlantic for 94 years, 1885–1978, in an EOF analysis also using the correlation matrix. We simulated a random process, uncorrelated over 74 spatial points and independently sampled 94 times, and repeated this simulation 100 times to compute the U_j^{95} values associated with Hayden's data set. Values of T_j and T_j/U_j^{95} for Hayden's computation are shown in Table 3. The percent variance accounted for by the largest eigenvalue for both studies is close: 22% versus 28%. The first four EOF's in Hayden's study appear to contain meteorological content distinguished from noise, based upon rule N. Hence, attempts at physical interpretation of these four EOF's are reasonable.

The forms of (1) and (2) also apply to principal component analysis via the covariance matrix. The geophysical variables are first centered about their mean value. The first EOF of the covariance matrix for the Bering Sea data set is shown in Fig. 2. The normalized eigenvalues and the results of applying rule N based upon a set of randomly generated covariance matrices are listed in Table 4. The primary difference between the first geophysical eigenvector of the covariance matrix and the first geophysical eigenvector of the correlation matrix is that the magnitude of relative maxima is greater with the covariance matrix. The first two principal components of the data set obtained by the covariance matrix satisfy rule N, although only the first one is much larger than that generated by noise.

3. Discussion

Comparison of results of the EOF analysis applied to the correlation and covariance matrix of the same data set illustrates an important point in choosing between one or the other approach for application to a geophysical problem. Since the sum of the eigenvalues equals the trace of the matrix, the principal components in the covariance approach are affected by the variance of each spatial variable as well as the covariance between variables. The covariance approach would therefore be particularly useful in locating specific regions with high variance relative to the rest of the field; an example would be in resolving the spatial distribution of sea-surface temperature anomalies. In an application of the correlation matrix the sum of the eigenvalues will again equal the trace of the matrix, but the contribution toward the vectorial direction represented by the EOF's is exclusively from the off-diagonal elements. The spatial-pattern-detection property of the correlation approach, as displayed in contour maps of the EOF modes, is advantageous in such applications as the cyclone climatology, in which one is specifically interested in spatial oscillations or variations of primary storm tracks.

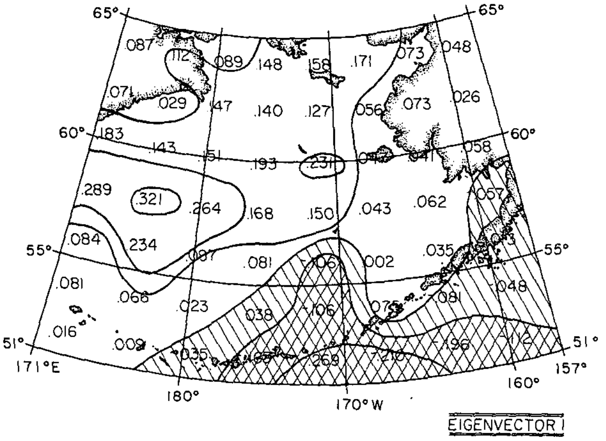

FIG. 2. Contours of the first empirical orthogonal function of the covariance matrix of the Bering Sea cyclone data set.

The difference between using covariance and correlation may then have relevance in applications of rule N. The part of the rule based on the Monte Carlo experiment can be interpreted as finding the lengths of the axes of p-dimensional ellipsoid generated by spatially uncorrelated noise. This is compared with the ellipsoid generated from the geophysical data set with the length of the axes equal to twice the square root of variance accounted for by each principal component. Because variances of individual variables do not contribute to increasing the length of the axes in the correlation approach, rule N should be a good test for spatial correlation in the geophysical field. In comparing Table 2 with Table 4, the magnitudes T_1 and U_1^{95} are both greater for the covariance case than the correlation case in this sample. Thus, the rule seems as conservative when applied to the covariance matrix as when applied to the correlation matrix.

TABLE 4. Summary of normalized eigenvalues of the covariance matrix of the Bering Sea data set; $p = 56$, $n = 23$.

	j				
	1	2	3	4	5
T_j (×100)	23.6	10.6	8.7	8.2	6.3
T_j/U_j^{95}	1.88	1.01	0.91	0.96	0.82

4. Conclusion

We have presented a test for principal components which is designed to determine if the eigenvalues of a geophysical data set can be distinguished from one drawn at random. The test is based upon examination of the eigenvalues derived from prenormalized data (correlation matrix) or those based upon the covariance matrix compared to eigenvalues generated from gaussian noise and is referred to as a dominant-variance selection procedure in Preisendorfer *et al.* (1981). It should be noted, however, that even if a particular principal component fails rule N, this need not eliminate the possibility that other criteria, such as the ability of the particular principal component to track closely time evolution of a geophysical process in the presence of large noise variance, may still be physically relevant. This option, called a time-history selection procedure, is discussed in Preisendorfer *et al.* (1981).

When rule N is applied to the case of the covariance matrix, the magnitudes of the normalized eigenvalues are influenced by both spatial correlation and the variance of individual variables. With the correlation matrix only spatial correlation contributes to the magnitude of the geophysical signal relative to one sampled from a spatially and temporally uncorrelated random process. In representing the variance of large data sets the covariance matrix is preferred; application of rule N will lead to the reduction of the dimension of the representation. The correlation matrix has certain advantages in resolving relationships between variables.

We consider that the question of significance is important for geophysical interpretation of EOF's and note that several recent MWR articles using EOF analysis do not address this issue. We hope that our note will provide the forum for further discussion of candidate selection rules and their geophysical interpretation so that a set of potentially good selection rules can be established.

Acknowledgment. This research is a contribution to the Marine Services Project at PMEL. It was supported in part by the Bureau of Land Management through interagency agreement with the National Oceanic and Atmospheric Administration, under which a multiyear program responding to needs of petroleum development of the Alaskan continental shelf is being conducted and managed by the Outer Continental Shelf Environmental Assessment Program (OCSEAP). Discussions with C. H. Pease and assistance from J. Gray, V. May and L. Lu are gratefully acknowledged.

REFERENCES

Hayden, B. P., 1981: Secular variation in Atlantic coast extratropical cyclones. *Mon. Wea. Rev.*, **109**, 159–67.

Overland, J. E., and C. H. Pease, 1982: Cyclone climatology of the Bering Sea and its relation to sea ice extent. *Mon. Wea. Rev.*, **110**, 5–13.

Preisendorfer, R. W., and T. P. Barnett, 1977: Significance tests for empirical orthogonal functions. *Reprints Fifth Conf. Probability and Statistics in Atmospheric Sciences*, Las Vegas, Amer. Meteor. Soc., 169–72.

——, F. W. Zwiers and T. P. Barnett, 1981: Foundations of principal component selection rules. SIO Rep. 81-7, May 1981, Scripps Institution of Oceanography, 200 pp.

Walker, C. T., and E. W. Bliss, 1932: World Weather V. *Mem. Roy. Meteor. Soc.*, **4**, 53–84.