NOTES AND CORRESPONDENCE

Interpretation of Extended Empirical Orthogonal Function (EEOF) Analysis

JENG-MING CHEN AND PATRICK A. HARR

Department of Meteorology, Naval Postgraduate School, Monterey, California

6 January 1993 and 29 January 1993

ABSTRACT

Application of an empirical orthogonal function (EOF) analysis to a data matrix that contains two or more variable fields has been referred to as extended EOF (EEOF) analysis. Coherence between individual features contained within one EEOF has been implied to represent interrelationships between the fields (in the case of a combination of different variables) or propagating features (in the case of the same field at different times). However, caution must be exercised in the interpretation of interrelationships within one EEOF because the derivation of the EEOFs is based on the optimization of the variance of every EEOF as an entity and may not indicate correlations among substructures within one EEOF. These types of problems associated with interpretation of EEOF analyses are highlighted through an analytic example and application to a dataset with known statistical properties.

Although other multivariate analysis techniques such as singular value decomposition and canonical correlation analysis are being used with more frequency, it is important to highlight potential difficulties associated with the EEOF technique that has been an integral analysis tool in meteorological research.

1. Introduction

Empirical orthogonal function (EOF) analysis, similar to principal component analysis (PCA) (Pearson 1901; Hotelling 1933), which is widely used in psychology and biostatistics, has become increasingly popular in meteorological research since its introduction by Lorenz (1956). Differences between EOF analysis and PCA are discussed by Richman (1986, 1987) and Jolliffe (1987). Primarily introduced as a data reduction technique (Lorenz 1956), EOF analysis has been employed in a variety of meteorological analyses (Richman 1986 and references therein). The primary reasons for the increased popularity of EOF analysis are the capability to reduce the dimensionality and to efficiently describe coherent variability in large datasets. These attributes make EOF analysis a particularly useful tool in climate research. Furthermore, EOF analysis can be applied to any combination of variables—observed or numerically generated, same field or not, regular grids or individual stations—and different time steps. The technique is also useful for analyzing vector quantities (Hardy and Walton 1978; Legler 1983; Klink and Willmott 1989) and complex valued time series composed of the scalar data (real part) and the data phase shifted one-quarter cycle (imaginary part) (Barnett 1983; Horel 1984; Davis et al. 1991).

Corresponding author address: Patrick A. Harr, Code MR/HP, Department of Meteorology, Naval Postgraduate School, Monterey, CA 93943-5000.

Physical interpretation of the spatial EOF patterns requires careful consideration of several factors, such as choice of dispersion matrices, sampling errors (North et al. 1982), and rotations of the EOF axes (Richman 1986, 1987). In particular, effects of the orthogonality constraints inherent in the EOF analysis and their relation to physical processes being inferred from the interpretation of the EOFs should be given special attention.

Physical interpretation of EOFs becomes even more difficult when the data matrix is composed of heterogeneous variables. Two types of heterogeneous datasets have been used in EOF analysis. The first type contains different field variables and has been labeled EOF complexes (Kutzbach 1967), composite EOF (Weare 1987), or multivariate EOF (MV-EOF) (Wang 1992). In MV-EOF analysis, patterns or substructures of different variables within one EOF are interpreted as the interrelated patterns between those variables. The second type contains the same variable at different times, and this has been referred to as extended EOF (EEOF) analysis (Weare and Nasstrom 1982; Lau and Chan 1985, 1986; Lau and Lau 1990). In EEOF analysis, successive patterns or substructures (Fig. 1) of the same variable within one EEOF are interpreted as the propagation or evolution in time of the first substructure of the function.

The purpose of this note is to demonstrate that these EEOF (or MV-EOF) interpretations may be misleading. Although recent presentation of the singular value decomposition (SVD) technique (Bretherton 1992)

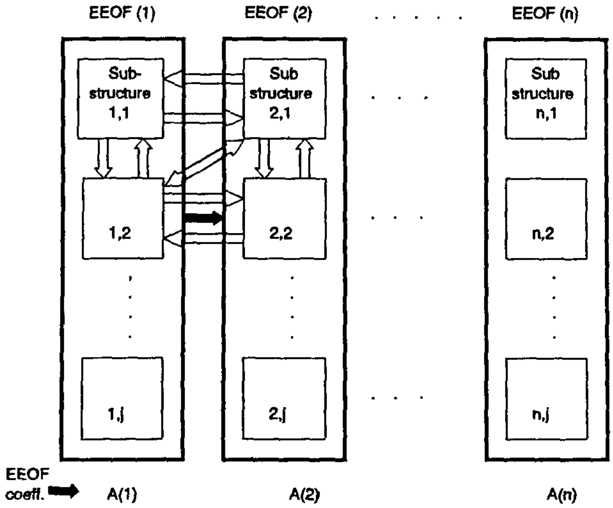

FIG. 1. Schematic diagram of EEOFs and their substructures. The solid arrow represents the orthogonal relationship between total functions. Open arrows emphasize the nonorthogonality between substructures within one function or between functions.

within the meteorological literature presents a viable alternative to EEOF analysis, it is important to analyze the potential problems with the EEOF technique that has been used for some time.

Since the mathematical derivation of EOF analysis is based on the optimization of the variance of each EOF as an entity, it has no reference to the correlations among partial structures within one EOF. These concepts are schematically represented in Fig. 1 as a set of n EEOFs with j substructures in each function that may represent time lags at, say, -1, 0, and 1 days. Substructures within one EEOF (or MV-EOF) are not orthogonal to either their counterparts or other substructures in other EEOFs. Furthermore, the EEOF coefficients [$A(k)$, $k = 1, \cdots, n$, in Fig. 1], which may be thought of as composed of a contribution from each substructure as

$$A(k) = a_1(k) + a_2(k) + \cdots + a_i(k) + \cdots \quad i = 1, \cdots, j; \quad k = 1, \cdots, n,$$

are derived from the total EEOF structure, so the correlation between substructures cannot be directly computed (i.e., it is not possible to compute the $a_i(k)$ terms directly). Therefore, interrelationships among substructures within one EEOF cannot be directly assessed (Chang and Chen 1992).

The difficulty in interpretation of EEOF analysis will be illustrated in an analytic example in section 2. A simple test dataset constructed with known statistical properties will be used in section 3 to further illustrate possible difficulties with EEOF interpretation.

2. Limits on EEOF interpretation

Typically, EEOF analysis is an EOF analysis applied on a data matrix that contains fields at different time lags. As a data reduction technique, inclusion of heterogenous data enhances the use of EOF analysis without major difficulties (Richman 1987). Since the derivation of EOF analysis is based on the optimization of total variance with little reference to correlation between substructures within one function, extra care must be taken when interpreting substructures within one EOF or EEOF as intercorrelated features. This is best illustrated by considering an EOF analysis of a two-variable dataset. Define X and Y as the mean removed variables, S and W as two principal components, and (a, b) as the principal component loading (EOF). Because of the orthonormality in space between EOFs, the relationships between S, W, X, and Y can be written as

$$S = aX + bY \qquad (1)$$

$$W = bX - aY, \qquad (2)$$

with $a^2 + b^2 = 1$. Since by definition there is no correlation in time between S and W,

$$a^2\overline{XY} - ab(\overline{X^2} - \overline{Y^2}) - b^2\overline{XY} = 0, \qquad (3)$$

where an overbar represents an average over time. Definition of a loading partition ratio, $r = a/b$, and a normalization factor, $m = 1 + r^2$, allows

$$S = m^{-1/2}(rX + Y) \qquad (4)$$

$$W = m^{-1/2}(X - rY) \qquad (5)$$

$$r^2(\overline{XY}) - r(\overline{X^2} - \overline{Y^2}) - (\overline{XY}) = 0, \qquad (6)$$

or

$$r^2cq - r(q^2 - 1) - cq = 0. \qquad (7)$$

Here,

$$q \equiv \left(\frac{\overline{X^2}}{\overline{Y^2}}\right)^{1/2} \qquad (8)$$

is a variance ratio of X to Y, and

$$c \equiv \frac{\overline{XY}}{[(\overline{X^2})(\overline{Y^2})]^{1/2}} \qquad (9)$$

is the time correlation coefficient between X and Y. The EOF solution r (loading partition ratio) versus the correlation c for different values of the variance ratio q is presented in Fig. 2, and r versus q for different values of c is presented in Fig. 3. Comparison of these figures indicates that the partition of the loadings is much more sensitive to the variance ratio of X to Y than to the correlation between X and Y for reasonable values of c (say, $c > 0.4$). The loading partition ratio shows only excessive sensitivity to c for values less than 20%. Unfortunately, these are cases that are not desirable when a researcher is investigating interrelationships (i.e., correlations) between fields.

It is interesting to note from Fig. 2 that for a variance ratio near 1, the loading partition ratio exhibits large

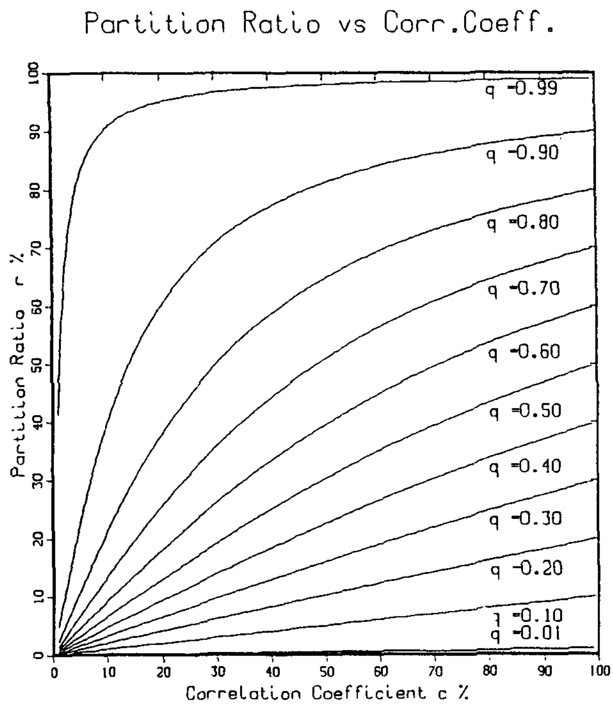

FIG. 2. Loading partition ratio r versus the correlation coefficient c for different values of the variation ratio q.

variation when the correlation coefficient is small (say <0.1). This represents the problems caused by sampling errors in EOF analysis as defined by North et al. (1982). Ideally, EOFs are obtained with no correlation between any two principal components. However, sampling errors may induce small correlations between principal components. When the variances of the principal components are similar, new components will be formed as linear combinations of the original.

Although the analytic derivation for the two field cases (heterogeneous data matrix) is extremely complicated, the mathematics are analogous to the two-variable case. Equation (3) is then written as

$$\overline{(\mathbf{AX})(\mathbf{AY})} - [\overline{(\mathbf{AX})(\mathbf{BX})} - \overline{(\mathbf{AY})(\mathbf{BY})}] - \overline{(\mathbf{BX})(\mathbf{BY})} = 0. \quad (10)$$

In this case, the correlation coefficient c, the variance ratio q, and the loading partition ratio r can vary upon different substructures within the loadings **A**, **B**. As in the two-variable case, the solution for the loading partition ratio r for varying q and c exhibits greater sensitivity to the variance ratio. Therefore, the resulting EEOFs represent maximum variance solutions with little regard to correlations between substructures within individual functions.

In summary, the constraint on EOF analysis is the optimization of variance over the entire function. For substructures within one function, there is no optimization constraint on either their individual variance or correlation. Since the loading partition ratio is more sensitive to the variance ratio than to correlations, substructures within one EEOF are locally optimal patterns rather than functionwide interrelated patterns. It is possible that significant correlation does exist between substructures if the most significant spatial features are also the most persistent in time. However, the same conclusions could also be drawn from examining correlations between principal components obtained from a regular EOF analysis.

Finally, a substructure in one EEOF may be correlated with its counterpart or with other substructures in other EEOFs. No unique solution exists for projecting a variable field onto the substructure hyperspace. Correlations between substructures and variance partitions of each substructure cannot be computed directly. Therefore, inclusion of heterogeneous data in EOF analysis can make physical interpretation of the resulting functions very difficult.

3. Example

The difficulty of EEOF interpretation is investigated by constructing a test dataset from a set of EOFs (Fig. 4) and PCs obtained from operational analyses of 700-mb zonal wind anomalies in the western Pacific between June and October 1979. Normally, an EEOF analysis is applied using an augmented dispersion matrix with one component, **X**, representing the data field at a base time, and a second component, **Y**, representing the data field lagged in time with respect to the base time. The EOFs and their PCs (P_i) derived from

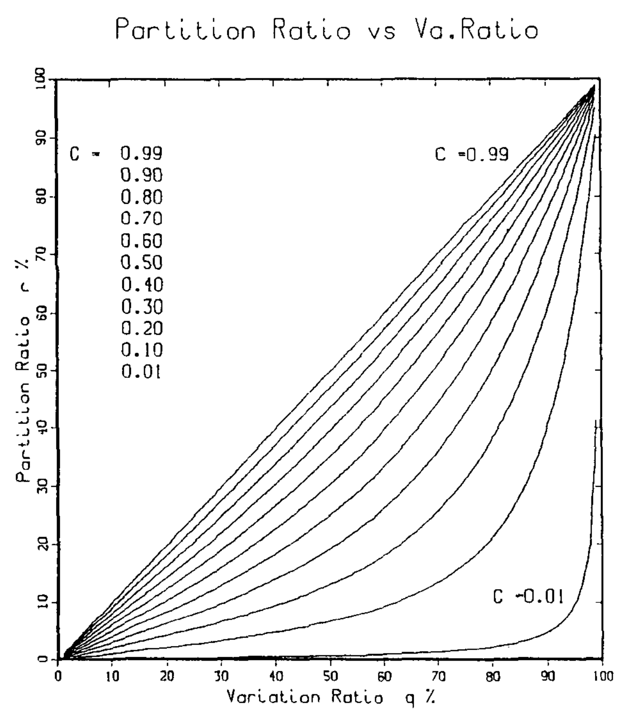

FIG. 3. Loading partition ratio r versus the variation ratio q for different values of the correlation c.

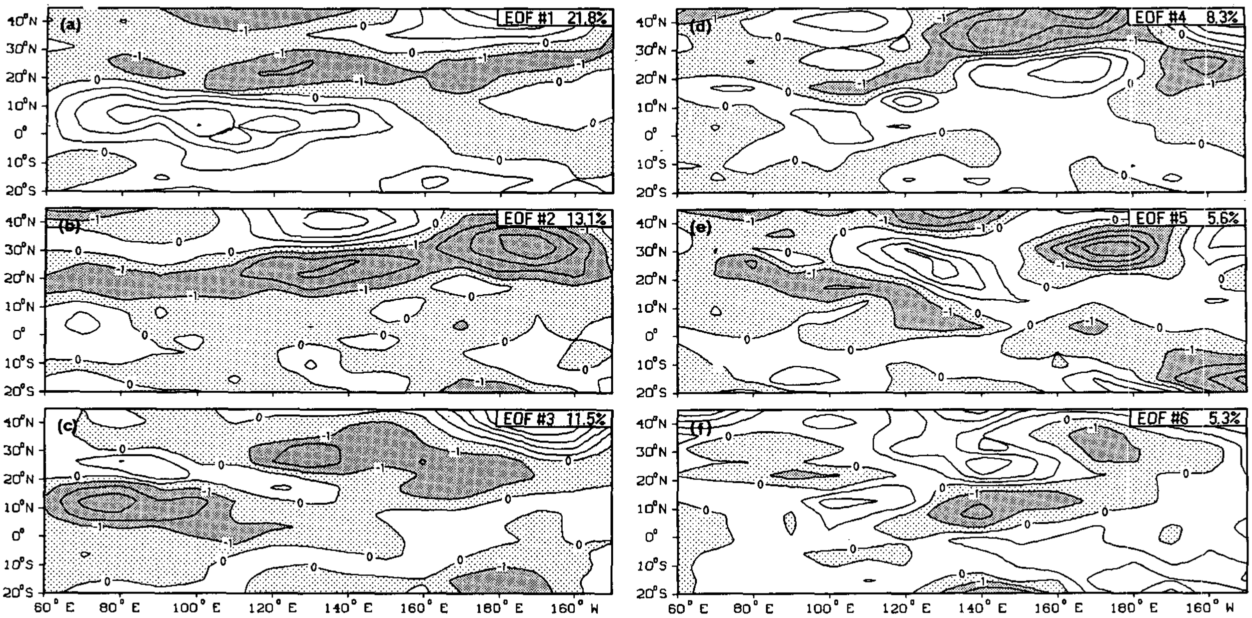

FIG. 4. Spatial EOF patterns of 700-mb anomalous zonal wind during June–October 1979. Negative values are shaded. Units are nondimensional, and the contour interval is 1.0. The percentage of variance explained by each function is indicated.

the zonal wind anomalies are combined to represent components of an augmented dispersion matrix. The **X** component is written

$$\mathbf{X} = (X_1, X_2, X_3, X_4, X_5, Y_6) \times (\mathrm{EOF}_1, \mathrm{EOF}_3, \mathrm{EOF}_5, \mathrm{EOF}_2, \mathrm{EOF}_4, \mathrm{EOF}_6)^{\mathrm{T}}.$$

The X_i are defined such that

$$\mathbf{X} = \begin{bmatrix} 1.0P_1 + 0.4P_2 \\ 0.5P_3 + 0.5P_4 \\ 0.2P_5 + 0.5P_6 \\ 0 \\ 0 \\ 0 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathrm{EOF}_1 \\ \mathrm{EOF}_3 \\ \mathrm{EOF}_5 \\ \mathrm{EOF}_2 \\ \mathrm{EOF}_4 \\ \mathrm{EOF}_6 \end{bmatrix}. \quad (11)$$

The **Y** component is defined as

$$\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4, Y_5, Y_6) \times (\mathrm{EOF}_1, \mathrm{EOF}_3, \mathrm{EOF}_5, \mathrm{EOF}_2, \mathrm{EOF}_4, \mathrm{EOF}_6)^{\mathrm{T}},$$

where the Y_i are defined such that

$$\mathbf{Y} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1.0P_2 \\ 0.5P_4 \\ 0.5P_6 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathrm{EOF}_1 \\ \mathrm{EOF}_3 \\ \mathrm{EOF}_5 \\ \mathrm{EOF}_2 \\ \mathrm{EOF}_4 \\ \mathrm{EOF}_6 \end{bmatrix}. \quad (12)$$

Constants were chosen so that the leading two EEOFs will be composed of the same EOFs and can easily demonstrate the possibility of misleading interpretations. Creation of the augmented matrix in this manner allows direct assessment of the interrelationships between EEOF substructures in terms of their correlations and variances. This is possible only because we know the exact structure of each dataset [defined by (11) and (12)], which is never known in practice. By definition, correlations between any two components i and j are

$$p_i p_j = \delta_{ij}.$$

Therefore,

$$X_i X_j = \delta_{ij}$$

$$Y_i Y_j = \delta_{ij}.$$

Finally, correlations between the X and Y components can be defined as

$$X_i Y_{j+3} = c_{i,j+3}\delta_{ij}.$$

Due to the unique construction of this test dataset, the correlations between substructures, $C_{i,j}$, may be computed (Table 1). By construction, the correlation between X_3 and Y_6 is largest, and correlation between X_1 and Y_4 is smallest.

The first substructure of EEOF_1 (Fig. 5a) is similar in sign and magnitude to EOF_1 (Fig. 4a), and the sec-

TABLE 1. Correlations between the substructures (X and Y) of the case 1 dataset constructed from principal components of 700-mb anomalous zonal wind, as defined by Eqs. (11) and (12).

i, j	corr (X_i, Y_j)
1, 4	0.2959
2, 5	0.6478
3, 6	0.9249

ond substructure of $EEOF_1$ (Fig. 5b) is similar to EOF_2 in sign but has smaller magnitude (Fig. 4b). These patterns suggest a positive correlation between the substructures. This interpretation may be extended to include features in one substructure as representing some alteration of the same feature in the other substructure (e.g., propagation of the feature in time and space).

Examination of $EEOF_2$ is necessary to further interpret the EEOF analysis. The first substructure of $EEOF_2$ (Fig. 6a) represents a negative amplitude EOF_1, but it also has much less amplitude (Fig. 4a). The second substructure of $EEOF_2$ (Fig. 6b) represents a positive amplitude EOF_2 with similar magnitude. Similar to $EEOF_1$, $EEOF_2$ also represents the contributions of EOF_1 and EOF_2 to the dataset. However, a negative correlation is implied by the $EEOF_2$ substructures. Therefore, the EEOF analysis correctly identified two modes of variability in the dataset, as defined by (11) and (12), but the correlations or relationships between these modes are contrary to each other.

The difficulty of interpretating the EEOFs demonstrated above was shown by subjectively analyzing the resultant patterns in light of the intrinsic modes known to exist in the dataset. In this case the modes are known exactly as defined by (11) and (12), but in practice they are not, and the subjective pattern analysis may be difficult.

This test case emphasizes the effect of the maximization of variance constraint on the EEOF analysis. By design, the substructure correlation $C_{1,4}$ is smallest (Table 1). The X_1 and Y_4 components are composed of the first and second PCs. Although there is little correlation between the substructures, the large variances associated with them dictate that they be represented by the leading EEOFs. The largest substructure correlation is defined by $C_{3,6}$ (Table 1), which is composed of the fourth and sixth PCs. Although there is extremely high correlation between the substructures,

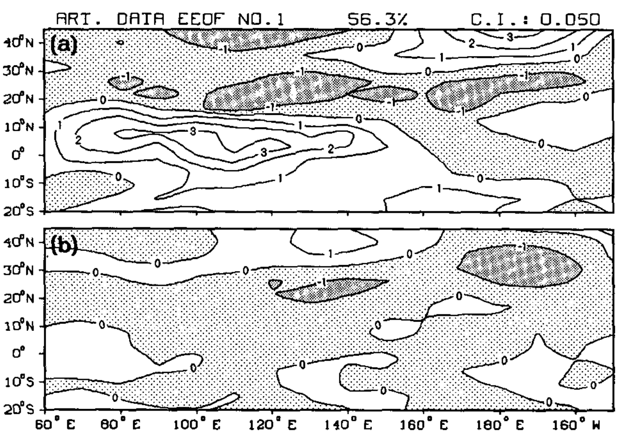

FIG. 5. (a) Substructure 1 of the spatial $EEOF_1$ pattern for the case 1 dataset. (b) As in (a) except for substructure 2. Negative values are shaded.

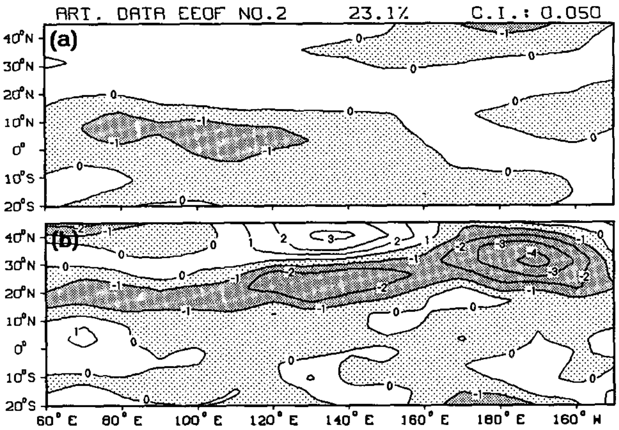

FIG. 6. (a) Substructure 1 of the spatial $EEOF_2$ pattern for the case 1 dataset. (b) As in (a) except for substructure 2. Negative values are shaded.

the maximization of variance criterion results in the pattern being placed in low-order EEOFs (not shown).

4. Summary

The purpose of this note was to examine several of the implicit aspects of EEOF analysis and show how potentially incomplete consideration of these components may result in misleading interpretations. It is possible that these problems could be masked when particular signals in the data are very strong (i.e., consistent in time). However, these signals could be examined by a regular EOF analysis, thereby avoiding the problems identified here.

Finally, the purpose of many applications of EEOF analysis has been to identify and interpret interrelationships between substructures of the EEOFs. It has been shown that the EEOFs are constructed according to the maximization of variance of the total function and not correlation. Therefore, application of EEOF analysis for the intent of analyzing correlations may lead to problems. Fortunately, several alternative multivariate analysis methods are available for analyzing interrelationships between two sets of variables. Canonical correlation analysis (CCA) is a well-known multivariate technique that seeks patterns with maximum correlation between two sets of variables regardless of their variance (Glahn 1968). The SVD technique that maximizes the covariance between two sets of variables represents a compromise between an EEOF analysis and CCA.

Acknowledgments. The authors would like to thank Prof. Russ Elsberry for carefully reviewing the manuscript and making valuable comments. This research was supported by the Naval Postgraduate School direct research funding and the National Science Foundation Grant ATM-9106495.

REFERENCES

Barnett, T. P., 1983: Interaction of the monsoon and Pacific trade wind systems at inter-annual time scales. I: The equatorial zone. *Mon. Wea. Rev.*, **111,** 756–773.

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

Chang, C.-P., and J.-M. Chen, 1992: A statistical study of winter monsoon cold surges over the South China Sea and the large-scale equatorial divergence. *J. Meteor. Soc. Japan,* **70,** 287–302.

Davis, J. M., F. L. Estis, P. Bloomfield, and J. F. Monahan, 1991: Complex principal component analysis of sea-level pressure over the eastern USA. *J. Climatol.,* **11,** 27–54.

Glahn, H. R., 1968: Canonical correlation and its relationship to discriminant analysis and multiple regression. *J. Atmos. Sci.,* **25,** 23–31.

Hardy, D. M., and J. J. Walton, 1978: Principal components analysis of vector wind measurements. *J. Appl. Meteor.,* **17,** 1153–1162.

Horel, J. D., 1984: Complex principal component analysis: Theory and examples. *J. Climate Appl. Meteor.,* **23,** 1660–1673.

Hotelling, H., 1933: Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.,* **24,** 417–441.

Jolliffee, I. T., 1987: Rotation of principal components: Some comments. *J. Climatol.,* **7,** 509–512.

Klink, K., and C. Willmott, 1989: Principal components of the surface wind field in the United States: A comparison of analyses based upon wind velocity, direction and speed. *J. Climatol.,* **9,** 293–308.

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

Lau, K.-H., and N.-C. Lau, 1990: Observed structure and propagation characteristics of tropical summertime synoptic scale disturbances. *Mon. Wea. Rev.,* **118,** 1888–1913.

Lau, K. M., and P. H. Chan, 1985: Aspects of the 40–50 day oscillation during the northern winter as inferred from outgoing longwave radiation. *Mon. Wea. Rev.,* **113,** 1889–1909.

——, and ——, 1986: Aspects of the 40–50 day oscillation during the northern summer as inferred from outgoing longwave radiation. *Mon. Wea. Rev.,* **114,** 1354–1367.

Legler, D. M., 1983: Empirical orthogonal function analysis of wind vectors over the tropical Pacific region. *Bull. Amer. Meteor. Soc.,* **64,** 234–241.

Lorenz, E. N., 1956: Empirical orthogonal functions and statistical weather prediction. Science Report 1, Statistical Forecasting Project, Department of Meteorology, Massachusetts Institute of Technology, 49 pp. [NTIS AD 110268.]

North, G. R., T. L. Bell, R. F. Cahalan, and F. J. Moeng, 1982: Sampling errors in the estimation of empirical orthogonal functions. *Mon. Wea. Rev.,* **110,** 699–706.

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

Richman, M. B., 1986: Rotation of principal components. *J. Climatol.,* **6,** 293–335.

——, 1987: Rotation of principal components: A reply. *J. Climatol.,* **7,** 511–520.

Wang, B., 1992: The vertical structure and development of the ENSO anomaly mode during 1979–1989. *J. Atmos. Sci.,* **49,** 698–712.

Weare, B. C., 1987: Relationships between monthly precipitation and SST variations in the tropical Pacific region. *J. Atmos. Sci.,* **115,** 2687–2698.

——, and J. S. Nasstrom, 1982: Examples of extended empirical orthogonal function analysis. *Mon. Wea. Rev.,* **110,** 481–485.