Sampling Errors in Statistical Models of Tropical Cyclone Motion: A Comparison of Predictor Screening and EOF Techniques

LLOYD J. SHAPIRO

Atlantic Oceanographic and Meteorological Laboratory, NOAA, Hurricane Research Division, Miami, FL 33149

(Manuscript received 13 September 1983, in final form 29 February 1984)

ABSTRACT

Statistical model significance, sampling and forecast errors are compared between linear regression models developed from preselected and ordered Empirical Orthogonal Function (EOF) predictors and those selected by a forward stepwise screening technique. As a particular application, grid-point height predictors are used to forecast tropical storm displacements in a storm-heading oriented coordinate system.

Critical correlations for model significance and upper bounds on expected sampling errors are derived from a Monte Carlo method. It is found that dependence among predictors selected by screening reduces expected sampling errors below those for the same number of independent screened predictors. For the given application, expected forecast errors for screened predictors are only slightly greater than those for EOFs.

1. Introduction

Statistical models provide important objective guidance for operational forecasts of tropical cyclone motion. Neumann and Pelissier (1981) describe the models currently in use at the National Hurricane Center (NHC). These models incorporate geopotential height data defined on a grid surrounding the cyclone. Predictors statistically related to tropical cyclone motion are typically selected from the available grid-point data.

As with all statistical forecast models, care must be taken in the selection of predictors. Due to random errors, regression equations derived from the developmental data typically will not perform as well on an independent sample, giving larger forecast errors. The greater the number of predictors, the greater the potential errors. The addition of more predictors to a model, although always apparently increasing the skill, may in fact degrade the forecast.

a. Preselection and ordering of predictors

Davis (1976, 1977) has clearly discussed the effect of random errors on the significance of linear statistical forecast models. The "hindcast" skill is the reduction of variance from the developmental data. Due to imperfect sample statistics, the hindcast skill will be greater than the true model skill. The (positive) bias in the hindcast skill due purely to chance is called the "artificial" skill. The expected artificial skill increases with the number of predictors in the regression model and is inversely proportional to the correlation time scales in the observations. The expected forecast skill on an independent sample is diminished in proportion to the artificial skill (Davis, 1977; Lorenz, 1956, 1977). Significance levels of sample hindcast skill can be estimated in terms of artificial skill (Chelton, 1983).

Davis and Chelton emphasize the importance of selecting model predictors by *a priori* criteria in order to limit artificial skill. Too many predictors will decrease model significance and reduce the expected forecast skill. The choice of the method used to select the predictors made available to the model is ultimately a subjective decision. Empirical Orthogonal Functions (EOFs) provide a convenient means of preselecting and ordering the predictors. Shaffer and Elsberry (1982) used EOFs in the development of a statistical model of tropical cyclone motion. Any other method, if *a priori* in nature, could be used to order the predictors.

Explicit techniques such as that described by Overland and Preisendorfer (1982) have been used to reject EOFs that are below the noise level, thereby limiting the number of predictors made available to the forecast model. In this study, a more empirical method is used. Ordered sets of EOFs are entered in the model, based upon the hindcast skill significance level. EOFs that are truly noise will show little skill and will (almost always) be rejected. Since the EOFs are preordered irrespective of their predictive skills, the *a priori* nature of the selection process is retained.

b. Screening of predictors

The operational statistical and statistical-dynamical models at NHC use a forward stepwise screening technique (e.g., see Draper and Smith, 1981) to select individual grid-point predictors that contribute most to the reduction of variance of tropical cyclone motion. Neumann and Lawrence (1975) describe the predictors and grid systems used to develop the regression equa-

tions in a recent forecast model. Traditionally, the stepwise screening algorithm was continued until the added reduction of variance due to one additional predictor was less than 1%.

Neumann *et al.* (1977), following Lund (1970), used a Monte Carlo technique to determine the critical correlation coefficient (or *F*-ratio) required for a given level of model significance when all grid-point predictors were made available to the screening program. In this technique the test statistic is established by replacing the original predictand series with a randomly ordered set. The stepwise screening algorithm is then used many times to select a specified number of predictors, each time using a different set of randomly ordered predictands but keeping the same set of predictors. The distribution of correlation coefficients determined in this way provides critical values at any level of significance. The method retains the spatial dependence of the grid-point predictors. Since a large number of predictors were available, but only a few were selected, the critical value for a given level of significance was much greater than that derived from classical tests. The classical tests assume that all available predictors are included in the model. Neumann *et al.* (1977) found that the 1% added reduction of variance cutoff was too liberal, incorporated too many insignificant predictors and thus potentially degraded the forecast.

An empirical critical partial correlation coefficient can be derived from the Monte Carlo technique by selection of just one predictor from those available. Neumann *et al.* (1977) found that when a single predictor is selected the critical correlation determined empirically from dependent grid-point predictors is smaller than that derived from an analytical expression which assumes independent predictors. They used the empirical critical partial correlation coefficient to limit the number of predictors selected by screening.

c. Comparison of EOF and screening methods

Davis (1977) compared the sampling errors of a model with a preselected set of predictors to those of a model with predictors selected by stepwise screening. Using a Monte Carlo technique, the average artificial skill of a model developed by screening was found to be considerably greater than a model with an equal number of *a priori* chosen predictors. Davis only considered a model with independent predictors. Tropical cyclone forecast models, however, are developed from dependent grid-point predictors. Thus, the results of Davis (1977) are not directly applicable to the evaluation of artificial skill in these models.

Klein and Walsh (1983) compared errors from dependent (developmental) and independent data samples between models developed from ten grid-point predictors selected by screening and from ten EOFs. The models represented specifications of monthly surface temperatures at individual stations from concurrent 700 mb data. The independent sample comprised only 11 months. They found that the predictors selected by screening outperformed the EOFs in both the developmental and independent samples. In the latter case the selected grid point gave about 10% smaller errors. No consideration was given, however, to obtaining comparable model significance or optimizing error reduction in the models. The use of an equal predetermined number of screened and EOF predictors makes comparisons between the potential predictive abilities of the two methods difficult. Similar comments apply to Klein (1983).

Recently, Shapiro and Neumann (1984; hereafter referred to as SN) developed models of tropical cyclone motion based upon synoptic deep-layer mean geopotential height predictors. A 113-point grid was used, oriented both geographically and with respect to current storm heading. A forward stepwise screening of predictors, with a 1% added reduction of variance cutoff, allowed up to eight predictors in the model. Based on the developmental data, SN found that the error was lowered by about 13% when the storm-heading-oriented system was used instead of the geographically-oriented system for a 24 h forecast. The ten EOFs that accounted for the greatest amount of variance in the height data were also utilized as predictors of 24 h motion. Similar reductions of variance based on the developmental data were obtained as from the screening technique. Neither the statistical significance of the model predictors nor the expected forecast skill were evaluated.

d. Outline of paper

The present paper makes a direct comparison of significance, and artificial and forecast skills between linear regression models developed from preselected and ordered predictors (e.g., EOFs) and from a forward stepwise screening technique. The effect of predictor dependence is explicitly evaluated. Although the model of tropical cyclone motion formulated by SN is used as a specific application, the methods and qualitative results are relevant to all statistical models in which predictors are not independent.

Section 2 presents the data and EOFs used in the analysis. The significance, and artificial and forecast skills are evaluated for a model with preordered EOFs as predictors, taking into account the serial correlation of the observations. The expected artificial skill of a model derived from forward stepwise screening of predictors is evaluated in Section 3, with an upper bound derived from a Monte Carlo method. The artificial skills are compared when independent predictors or dependent grid-point predictors are made available to the screening technique. The results are applied to the evaluation of forecast skill for the models formulated by SN. Section 4 derives critical hindcast skills for

model significance when independent and dependent predictors are selected by screening. The analysis allows a direct comparison of model forecast skills when EOFs of grid-point predictors are selected. The results of the analysis are summarized in Section 5.

2. Sampling errors

a. Description of data

The developmental data used in the present analysis comprises 24, 48 and 72 h storm displacement forecasts. The cases are selected from all available forecast situations at 0000 and 1200 GMT for tropical storms and hurricanes in the North Atlantic basin from 1965 through 1980. Pressure-weighted deep-layer mean geopotential heights, interpolated from National Meteorological Center operational analyses, are used as predictors of storm motion. The heights are defined on a storm-centered grid, which has a uniform 278 km spacing measured along latitude and longitude circles. The grid is oriented with respect to the current storm heading, with positive displacements in the direction of the storm heading (y') and to the right (x'). A complete list of symbols is given in the Appendix. Figure 1 shows the grid, which is oriented for illustration with respect to a storm heading 23° east of north. There are 113 grid points, all contained within a circle of 1700 km radius. Further details of the data and grid rotation are given in SN. For 24, 48 and 72 h displacement forecasts, there are N = 795, 621 and 477 cases available, respectively. Table 1 presents the average displacement $\mathbf{X}' = (X', Y')$ and standard deviation $s(\mathbf{X}')$ for each component in the storm-heading-oriented system. The total variance of displacement is measured by

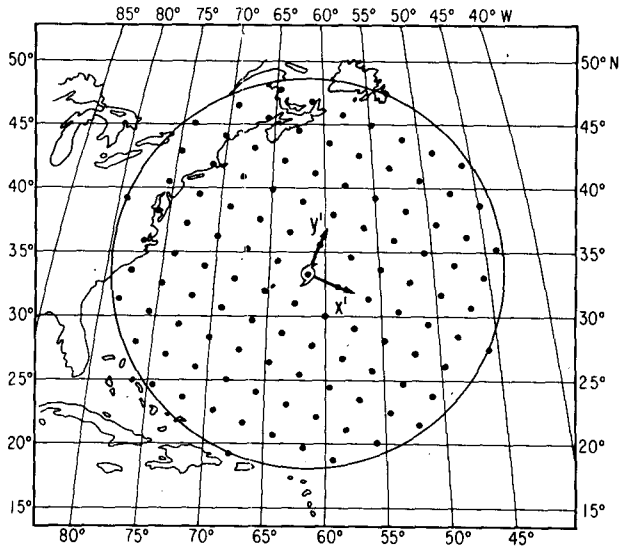

FIG. 1. Grid system, oriented with respect to storm heading. In this example, storm heading is 23° east of north, center of grid is at 33°N, 61°W. The orientation and position of the grid are the average for the sample used to predict 24 h motion.

TABLE 1. Distribution of displacements in storm-heading-oriented coordinate system for 24, 48 and 72 h forecasts.

		$\bar{X}'$ (km)	$s(X')$	$s(\text{tot})$
24 h	X'	39	163	443
	Y'	534	411	
48 h	X'	172	497	849
	Y'	793	688	
72 h	X'	317	849	1240
	Y'	819	906	

$$s^2(\text{tot}) \equiv s^2(X') + s^2(Y').$$

As discussed in SN, for 24 h displacements the variance in the rotated system is primarily in the along-track (y') direction. This has the effect of reducing the total variance of storm motion 40% below that in the geographically-oriented system. For 48 and 72 h this advantageous effect is reduced or eliminated.

b. EOFs

The heights of the 113 grid points provide a large number of available predictors for a statistical forecast model. The method of EOFs was used by Lorenz (1956) to reduce the number of predictors entered into such a model. The properties of EOFs are concisely discussed in Appendix B of Davis (1976). The modes are independent and are ordered with respect to their contribution to the total variance (or energy) in the height data. The first mode explains the greatest portion of the variance. The lowest-order modes, which are the most energetic, can be used as uncorrelated predictors of storm displacements.

Together, the first ten height EOFs in the rotated coordinate system explain 98.5% of the total height variance. The first four EOFs are shown in Fig. 2.

c. Artificial and forecast skills

The reduction of variance, or hindcast skill S_H, is derived from averages over a finite set of observations (e.g., see Davis, 1976). Due to sampling errors, S_H overestimates the true model skill, which is defined by (infinite) ensemble averages. Chelton (1983) distinguishes between the true model skill S, as defined by expected values (ensemble averages), and the true model skill applied to a particular data set, as defined by sample averages. He denotes the latter value by $\tilde{S}$. We retain his notation in the present paper. Hindcast skill S_H overestimates $\tilde{S}$ by an amount called the artificial skill S_A (Davis, 1976; Chelton, 1983). Thus by definition,

$$S_A = S_H - \tilde{S}. \tag{1}$$

Denoting expected values by angular brackets,

$$\langle S_A \rangle = \langle S_H \rangle - S, \tag{2}$$

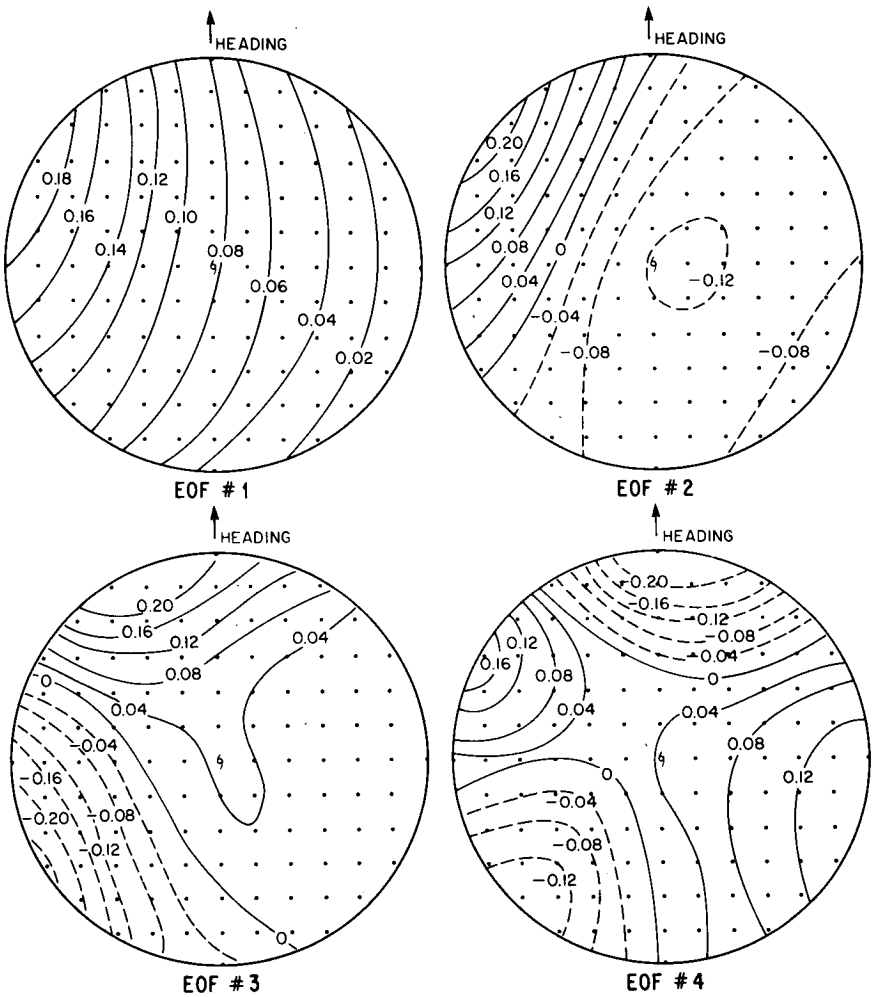

FIG. 2. Height EOFs in rotated grid system.

where $\langle \tilde{S} \rangle = S$ has been used. The expected forecast skill $\langle S_F \rangle$ for an independent sample is less than the true skill by approximately the same amount that the hindcast skill exceeds the true skill (Lorenz, 1956, 1977; Davis 1976, 1977), with

$$\langle S_F \rangle \approx S - \langle S_A \rangle. \tag{3}$$

Then, from (2) and (3),

$$\langle S_F \rangle \approx \langle S_H \rangle - 2\langle S_A \rangle. \tag{4}$$

Thus, as the artificial skill increases, the expected forecast skill decreases proportionally. Random errors degrade the forecast. Chelton (1983) has derived the approximate expression for the expected artificial skill when m preselected predictors are used:

$$\langle S_A \rangle \approx \frac{m}{N^*}(1 - S). \tag{5}$$

This relation is valid when the correlation time scales for each predictor are nearly equal. The effective number of cases N^*, which accounts for the serial correlation between observations, is assumed $\gg 1$. Limiting cases of (5) were derived by Lorenz (1956, 1977) for serially uncorrelated observations ($N^* = N$), and by Davis (1976, 1977) for $S \ll 1$. The value of $\langle S_A \rangle$ increases both as the number of predictors m increases, and as the effective number of observations N^* decreases. It decreases as the true skill S increases.

d. Effective number of cases

Davis (1976, 1977) discussed the effect of serial correlations among observations on hindcast skill significance. When long time scale variations are present, and correlations between observations are large at substantial lags, the effective number of degrees of freedom N^* is reduced. This reduction has the effect of increasing sampling errors and the artificial skill. Only when the observations are statistically independent does $N^* = N$.

Davis (1976, 1977) developed a formula based on an integral correlation time scale for computing N^*

from the sample data. His analysis was restricted to cases with small skill ($S \ll 1$). Chelton (1983) removed this restriction in his derivation.

An alternative simple method proposed by Davis (1977) for the estimation of N^* is used in the present analysis. Many long-lag hindcasts (estimates by the height predictors of storm displacements at a time far removed) are made for which the true skill is assumed to vanish ($S = 0$). The average hindcast skill obtained in this way is completely artificial and so, from (2) and (5), equals m/N^*. This gives an estimate of N^* when m predictors are included. This method (with $m = 1$) is used to estimate N/N^* for each of the first ten height EOFs. The average of the ten values of N/N^* is 2.12.[1] The variation of N/N^* from one predictor to another is relatively small, with a standard deviation of 0.37. Thus, the approximation implicit in the derivation of (5) that each predictor (or set of predictors) is associated with the same value of N^* can reasonably be made. For 24, 48 and 72 h forecasts, N = 795, 621 and 477, so that N^* = 375, 293 and 225, respectively.

e. *Model significance and forecast errors for EOFs*

The addition of predictors to a statistical model will always increase the hindcast skill (Davis, 1976). Classically, the significance of a second model derived from the first by the addition of predictors can be evaluated with the "F" statistic (e.g., Kleinbaum and Kupper, 1978). If the F-ratio is greater than a critical value derived from the F-distribution, the second model has significantly greater skill than the first. In the classical tests, the number of degrees of freedom is computed assuming N statistically independent observations. Serial correlations among observations may be accounted for by replacing N by N^* in the formulas (Chelton, 1983). (Discarding intermediate observations to obtain serial independence would reduce the number of observations to $\sim N^*$, thus losing valuable information. The present technique utilizes the entire sample, with N observations.) When N^* is large, the asymptotic form of the F-distribution can be used to derive the test for statistical significance of the additional predictors at the α level:

$$\frac{(S_H)_2 - (S_H)_1}{1 - (S_H)_1} \geqslant \frac{\chi_{m_2-m_1}(1-\alpha)}{N^*}, \qquad (6)$$

where $\chi_{m_2-m_1}(1-\alpha)$ is the 100 $(1-\alpha)$ percentage point of the chi-squared distribution with $m_2 - m_1$ degrees of freedom. Here the original m_1-predictor model has hindcast skill $(S_H)_1$, and the second m_2-predictor model has hindcast skill $(S_H)_2$. Thus, $m_2 - m_1$ predictors are added by the new model. The quantity on the left in (6) is simply the square of the multiple-partial correlation coefficient due to the additional predictors [cf. Eq. (11.6) of Kleinbaum and Kupper, 1978]. As a special case, where the original model has zero predictors, so that $m_1 = 0$ and $(S_H)_1 = 0$, (6) becomes

$$S_H \geqslant S_H^{\text{crit}}(1-\alpha) = \frac{\chi_m(1-\alpha)}{N^*}. \qquad (7)$$

Here $m \equiv m_2$ and $S_H \equiv (S_H)_2$. This relation, which is also derived directly by Chelton (1983), provides a quantitative measure of the α significance level for S_H for an m-predictor model.

Application of these statistics allows evaluation of the significance of a preordered set of regression models, using the first 5, 10 or 15 EOFs as predictors of storm displacement. The addition of only one EOF at a time could lead to a model with too few predictors, omitting physically meaningful and useful higher-order EOF modes. The prespecified order of the individual EOFs may not correspond exactly to the order of their physical relationship to storm displacement. The use of sets of five smooths over these misorderings and sampling errors that contribute to a misordering within a given set. Application of (7) indicates that in all cases shown in Table 2 a model with $m = 5$ EOF predictors is significant at the 5% level. Then the next five EOFs are added to the model ($m_1 = 5$, $m_2 = 10$), and tested for significance. If the additional predictors add significant skill [by (6)], they are retained. The number of predictors $m = m_2$ retained in this analysis, together with $S_H(\mathbf{X}')$ for each displacement component, is given in Table 2. In all but one case, $m_2 = 10$ predictors are incorporated into the model.

The combined standard hindcast error e_H for both displacements, defined by

$$e_H^2 \equiv [1 - S_H(X')]s^2(X') + [1 - S_H(Y')]s^2(Y'), \qquad (8)$$

TABLE 2. Skills and errors using m height EOFs as predictors of storm motion.

		m	S_H	e_H (km)	$\langle S_A \rangle$	e (km)	e_F (km)
24 h	X'	10	0.236	270	0.020	274	277
	Y'	10	0.688		0.008		
48 h	X'	10	0.272	661	0.025	672	682
	Y'	10	0.458		0.018		
72 h	X'	10	0.283	1020	0.032	1037	1054
	Y'	5	0.361		0.014		

[1] Classical runs tests for serial correlation (e.g., Siegel, 1956) provide another estimate of N/N^*. These tests evaluate the expected number of runs that would be obtained from a series of length N if all observations were uncorrelated. Correlations among observations reduce the number runs, just as they reduce N^*. Thus, N/N^* may be estimated by the ratio between the expected and actual number of runs in the predictand series (displacements). In this way, N/N^* is estimated to be ≈ 2.65. This value is more conservative than that from the long-lag hindcasts, since the serial correlation of only the predictand is considered. Neumann *et al.* (1977) also noted that a reduction in sample size by a factor of ~ 2.5 is necessary to achieve effective serial independence for hurricane track forecasts.

is also given in Table 2. Incorporating $m = m_2$ predictors, and using the values of N^* derived in Section 2d, $\langle S_A \rangle$ can be evaluated from (2) and (5), while $\langle S_H \rangle$ is estimated by S_H for the sample data. Table 2 gives the values of $\langle S_A \rangle$ for each displacement component. When S_H is substantial, the $(1 - S)$ factor in (5) greatly reduces $\langle S_A \rangle$; S is derived from (2), and $\langle S_F \rangle$ from (4). The combined standard error e is defined in terms of the true model skill [cf. (8)]:

$$e^2 \equiv [1 - S(X')]s^2(X') + [1 - S(Y')]s^2(Y'). \quad (9)$$

Similarly, the combined standard expected forecast error e_F is defined by

$$e_F^2 \equiv [1 - \langle S_F(X') \rangle]s^2(X') + [1 - \langle S_F(Y') \rangle]s^2(Y'). \quad (10)$$

Here e_F is an estimate of the expected model forecast skill on an independent sample. When $\langle S_A \rangle = 0$ for each component, $e^2 = e_F^2 = \langle e_H^2 \rangle$. The larger the expected artificial skill, the larger are e and e_F.

Random errors degrade the forecast, so $e_F > e > e_H$ in all cases, as seen in Table 2. The random errors are small, however, so that forecast error is greater than the corresponding hindcast error only by about 3%.

3. Sampling errors from screening of predictors

An alternative to the preselection of the predictors that are included in a statistical model is the forward stepwise screening technique. In this method individual predictors are entered into the model in the order of their contribution to the reduction of predictand variance. In this section the expected artificial skill of a model developed by stepwise predictor screening is evaluated both for independent predictors and those selected from (dependent) grid points.

a. Independent predictors

Davis (1977, his Fig. 1) evaluated the average artificial skill of a model developed by screening of independent, normally distributed, predictors. A Monte Carlo method was used, assuming zero true model skill ($S = 0$). When only a few predictors were selected from those available, Davis found that the average artificial skill of a model developed by screening was very much greater than when an equal number of predictors was preselected.

Using a Monte Carlo simulation equivalent to that of Davis (1977), both $k = 10$ and $k = 113$ independent predictors are first made available to a screening program.[2] The average model hindcast skill, which is purely artificial, is computed from 100 reorderings of the predictand series, when $m = 1, 2, \ldots, k$ predictors are selected by the usual screening technique. For each m, the average of S_H gives an estimate of $\langle S_A \rangle$ when $S = 0$. For convenience, the predictand series is assumed random and normally distributed. As in the analysis by Neumann *et al.* (1977), the use of actual storm displacements instead would make very little difference in the quantitative results. A direct comparison with classical results is facilitated by the assumption of normality. The effective number of preselected predictors $\bar{m}$ from the Monte Carlo method is defined from the average artificial skill $\langle S_A \rangle$ by

$$\langle S_A \rangle_{S=0} \equiv \bar{m}/\mathcal{N}, \quad (11)$$

where $\mathcal{N}$ is the record length. For large $\mathcal{N}$, $\bar{m}$ is independent of the length of the record. If the series were random, $\mathcal{N} = N$; but since there is substantial serial correlation, in practice $\mathcal{N} = N^*$ is used, with N^* evaluated as in Section 2d.

Figure 3 displays $\bar{m}$ as a function of the number of predictors selected. Davis (1977, his Fig. 1) also evaluated the case when $k = 10$, with virtually the same result as that presented here. When the available predictors are all selected ($m = k$), then [from (5) with $S = 0$] $\bar{m} = m = k$. As noted by Davis (1977) and Lorenz (1977), when fewer predictors are selected, with $m < k$, screening reduces artificial skill below that of a model with all available predictors included. Thus, $\bar{m} < k$. The artificial skill is, however, greater than when the same number of predictors are preselected, so that $\bar{m} > m$. For few predictors selected (m small), Fig. 3 shows that $\bar{m} \gg m$, so screening gives much greater artificial skill. The more predictors available to the screening program, the greater the skill expected by chance. Thus, for a given number of independent predictors selected,

$$\bar{m}(k = 113) \gg \bar{m}(k = 10).$$

b. Dependent predictors

When the 113 grid-point height predictors are available to the screening algorithm, dependence among predictors from point to point affects the artificial skill. Since the Monte Carlo technique reorders only the predictand series, dependence among predictors is conserved. The same method was applied by Neumann *et al.* (1977). The open circles in Fig. 3 display the effective number of preselected predictors $\bar{m}$ when the dependent heights are screened. Predictor dependence is seen to reduce $\bar{m}$ below that for the same number of independent predictors (crosses). In the present example, for $m = 1$ predictor selected, dependence reduces $\bar{m}$ from 7.7 to 3.8, about 50%. We also note that for $m = 1$, $\bar{m}$ for the 113 dependent grid-point predictors $\approx \bar{m}$ for 10 independent predictors (solid circles). Thus, in the first selection, dependence among

[2] For convenience the independent predictors used are EOFs derived from the 113 grid-point height predictors. Any other set of uncorrelated predictors, not related to the height data, could also have been used without altering the results.

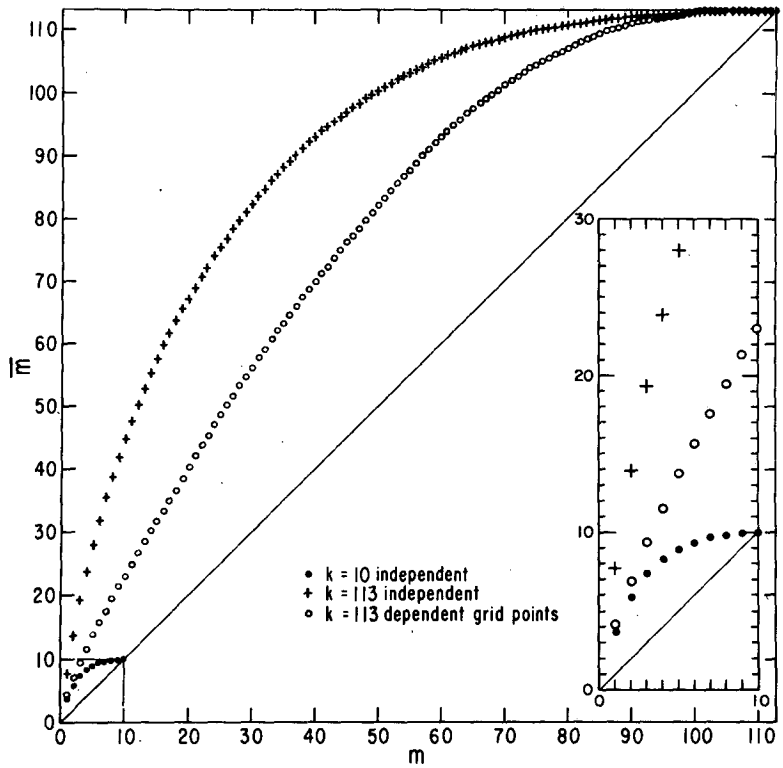

FIG. 3. Effective number of preselected predictors $\bar{m}$ as a function of the number of selected predictors m; k is the number of predictors available to a screening program. The inset expands the graph for $m \leq 10$.

the 113 grid-point predictors provides the screening program with a potential error equivalent to only about ten independent predictors. This observation, which will be expanded upon in the following section, highlights the substantial error-reducing ability of predictor dependence. Dependence among predictors reduces the effective number of predictors $\bar{m}$ in the same way that serial correlation reduces the effective number of cases N^*.

c. *Bounds on model skill*

The Monte Carlo technique used to derive Fig. 3, and Fig. 1 of Davis (1977), requires the assumption of zero true model skill ($S = 0$). In that case, when few predictors are employed, $\langle S_A \rangle$ for a model with predictors selected by screening is much greater than when an equal number of predictors are preselected. In practice, however, $S \neq 0$. Some predictors are, in fact, correlated with the predictand. When predictors are preselected, the $(1 - S)$ factor in (5) shows that $S > 0$ reduces artificial skill. When predictors are screened, $S > 0$ also tends to order and thus preselect the predictors; those predictors with greater skill contribution tend to be selected first. When $S = 0$, apparent model skill is due purely to random errors, so that all predictors are seen as equivalent (*a priori*) by the screening program. Preordering reduces artificial skill; thus, $\langle S_A \rangle$ derived from the Monte Carlo method ($S = 0$) provides an upper bound on the actual expected artificial skill. In the limit that the selected predictors are completely determined because of their substantial skill contribution, the expected artificial skill is given by (5).

Hindcast skills and errors were derived by Shapiro and Neumann (1984) for models incorporating grid-point predictors of tropical cyclone motion. A stepwise screening of predictors was continued until the additional reduction of variance in each displacement component was less than 1%. Table 3 shows the number of predictors m selected and hindcast skills S_H for each component, and the hindcast errors e_H. Estimates of true model errors e and forecast errors e_F are also shown, using two estimates of $\langle S_A \rangle$; e (lower) is determined from (2) and (9), with $\langle S_A \rangle$ derived from (5) for preselected predictors, while e (upper) uses $\langle S_A \rangle$ from (11), the Monte Carlo estimate ($S = 0$), with $\mathcal{N} = N^*$. When predictors are preselected, (4) [or (3)] and (5) together can be used in (10). The estimate of e_F (lower) is determined in this way. When predictors are selected by screening, the expression for $\langle S_F \rangle$ corresponding to (4) and (5) cannot be easily derived. For a given $\langle S_A \rangle$, however, a qualitatively correct estimate of e_F can be made from (4), assuming [in (3)] that random errors degrade the forecast by an amount close

TABLE 3. Skills and errors using m grid-point height predictors selected by screening, with 1% added reduction of variance cutoff.

		m	S_H	e_H (km)	e (km) Lower	e (km) Upper	e_F (km) Lower	e_F (km) Upper	$\bar{m}$
24 h	X'	4	0.229	271	272	284	274	298	11.5
	Y'	6	0.690						15.6
48 h	X'	3	0.238	652	659	681	665	706	9.3
	Y'	7	0.496						17.5
72 h	X'	7	0.318	993	1008	1046	1021	1096	17.5
	Y'	5	0.395						13.7

to that by which the hindcast is artifically enhanced. An equivalent assumption is implicit in the discussions of Lorenz (1977) and Klein and Walsh (1983). In this way e_F(upper) is derived with $\langle S_A \rangle$ from (11). In Table 3 $\bar{m}$ is also given; $\bar{m} \gg m$, but is about one-half the corresponding value for independent predictors (Fig. 3). The estimates of e and e_F provide approximate lower and upper bounds on the expected true model and forecast errors. The actual values of e and e_F depend upon the true skill of each individual predictor. The lower estimates of 48 and 72 h will not be realized in practice since, as determined in Section 4, statistically insignificant predictors are included in these cases. As determined from e_F(upper) in Table 3, each estimated forecast error is at most about 10% greater than the corresponding hindcast error. Thus, the models derived by Shapiro and Neumann (1983) are expected to have value in an actual forecast situation.

4. Critical hindcast skills and model significance

a. Critical hindcast skills

Classical screening techniques apply an F-test or partial correlation cutoff, accepting predictors into the model until the F-ratio or partial correlation coefficient due to one additional predictor falls below a critical value. The critical value at a given level of significance is given by (6), with $m_2 - m_1 = 1$, only when the order of the selected predictors is predetermined. When the predictors are selected in the order of their contribution to the reduction of variance, (6) is not valid. Neumann *et al.* (1977) used the Monte Carlo technique, as applied in Section 3, to determine empirically the critical values of variance reduction (S_H) when predictors were selected from a set of dependent grid-point predictors. Their analysis was for a different grid and data set than that used in the present analysis.

In the course of diagnosing the average artificial skills used to derive Fig. 3, we have determined critical values of S_H at the 5% significance level, based upon correlations with random, normally distributed predictands. For each m, the critical values were obtained from the cumulative percentage frequency distribution of the 100 values of S_H determined from the Monte Carlo simulations. The results are shown in Fig. 4a, for $m = 1$ to 10 predictors selected by screening. When N^* is large, the ordinate $N^* S_H^{\text{crit}}(1 - \alpha)$ is independent of N^* [cf. (7)]. If m predictors were preselected, (7) indicates that

$$N^* S_H^{\text{crit}}(1 - \alpha) = \chi_m(1 - \alpha).$$

The values of χ_m (0.95) are shown by the stars in Fig. 4a. In agreement with theory, the empirical value of

$$N^* S_H^{\text{crit}}(1 - \alpha) \approx \chi_m(0.95)$$

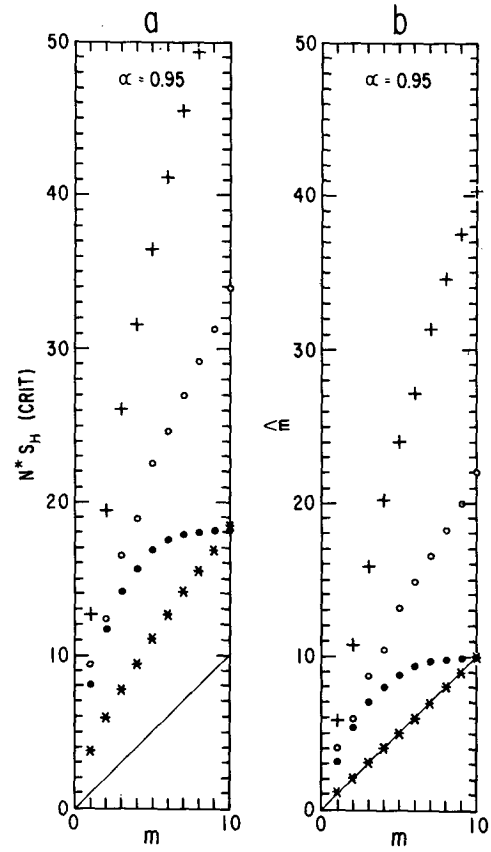

FIG. 4. (a) Critical hindcast skill at 5% significance level times the effective number of independent cases, $N^* S_H$(crit) as a function of m. Stars represent values for preselected predictors, discussed in text. Other symbols as in Fig. 3. (b) As in (a) for for $\hat{m}$, defined by $\chi_{\hat{m}}(\alpha = 0.05) = N^* S_H$(crit).

when $m = k = 10$, since then screening plays no role in predictor selection. When $m \ll k$, however, screening artificially greatly inflates hindcast skill, with

$$N^*S_H^{\text{crit}}(1 - \alpha) \gg \chi_m(1 - \alpha).$$

Qualitatively, $S_H^{\text{crit}}(1 - \alpha)$ depends on m in very much the same way as does $\bar{m}$ (Fig. 3). For a given m, dependent predictors give a smaller critical S_H than do the same number of independent ones. The results for $\alpha = 0.01$ (not shown) are very much the same as those in Fig. 4a, with $S_H^{\text{crit}}(1 - \alpha)$ of course larger in all cases.

b. Relation to expected artificial skill

In order to directly compare the results of Fig. 4a with those in Section 3, the empirical values of $N^*S_H^{\text{crit}}(1 - \alpha)$ from the Monte Carlo method can be related to an effective number of preselected predictors $\hat{m}$, defined by

$$S_H^{\text{crit}}(1 - \alpha) \equiv \chi_{\hat{m}}(1 - \alpha)/N^*. \tag{12}$$

This definition is analogous to (11) for $\bar{m}$ in terms of $\langle S_A \rangle$. Both $\bar{m}$ and $\hat{m}$ represent an equivalent number of preselected predictors; $\bar{m}$ is used in applications that involve average artificial skill, and $\hat{m}$ in those that involve critical hindcast skills. In general, $\hat{m}$ may depend upon α, as well as m and the set of available predictors; $\hat{m}$ is shown in Fig. 4b, derived from N^*S_H(crit) in Fig. 4a. When predictors are preselected, (12) reduces to the theoretical relation (7), with $\hat{m} = m$. Thus, the stars in Fig. 4b all lie on the diagonal. The empirical values of $\hat{m}$ in Fig. 4b are close to corresponding values of $\bar{m}$ in Fig. 3, but are not identical.

When a single predictor is selected ($m = 1$) from a set of k independent predictors, analytic formulas can be deduced for both $S_H^{\text{crit}}(1 - \alpha)$ and $\langle S_A \rangle_{S=0}$, and thus $\hat{m}$ and $\bar{m}$. If the probability that S_H for a single predictor is not greater than a given critical value is $(1 - \alpha)^{1/k}$, then the probability that none of k (independent) predictors has a value of S_H greater than the same critical value is $1 - \alpha$. Thus, when $m = 1$ predictor is chosen from k independent ones, the critical value of S_H at the α level can be derived from (7), with $(1 - \alpha)$ replaced by $(1 - \alpha)^{1/k}$, so that

$$S_H^{\text{crit}}(1 - \alpha) = \chi_1[(1 - \alpha)^{1/k}]/N^*. \tag{13}$$

If $\alpha = 0.05$ and $k = 10$, $(1 - \alpha)^{1/k} = 0.995$, and $N^*S_H^{\text{crit}}(1 - \alpha) = 7.9$. This value is confirmed empirically in Fig. 4a (and Davis, 1977). When $k = 113$, $N^*S_H^{\text{crit}}(1 - \alpha) = 12.2$, also found in Fig. 4a. For small α, the approximation $(1 - \alpha)^{1/k} \approx 1 - \alpha/k$ can be made. This approximation has been used to adjust critical values for the availability of a large number of predictors (e.g., Miller, 1962). The adjustment is only valid, however, when one predictor is selected from an independent set.

For independent predictors, (12) and (13) imply for $m = 1$,

$$\chi_{\hat{m}}(1 - \alpha) = \chi_1[(1 - \alpha)^{1/k}]. \tag{14}$$

As a special case, we choose α such that $\hat{m}$ [from (14)] equals $\bar{m}$ *and* $S_H^{\text{crit}}(1 - \alpha)$ [from (13)], equals $\langle S_A \rangle_{S=0}$. In that case, (11) and (12) together imply that

$$\chi_{\hat{m}}(1 - \alpha) = \hat{m}. \tag{15}$$

Then, (14) and (15) can be solved simultaneously for α and $\hat{m} = \bar{m}$ ($m = 1$). When $k = 10$, $\bar{m}$ ($m = 1$) ≈ 3.8, while for $k = 113$, $\bar{m}$ ($m = 1$) ≈ 7.9. These values are confirmed in Fig. 3a for independent predictors.

When predictors are dependent, (14) is not valid. In general, however, the significance level $\alpha = \alpha'$ at which

$$S_H^{\text{crit}}(1 - \alpha) = \langle S_A \rangle_{S=0}$$

for $m = 1$ can be found empirically. Then, the equivalent number of independent predictors available, k_e, is defined by

$$S_H^{\text{crit}}(1 - \alpha') \equiv \chi_1[(1 - \alpha')^{1/k_e}]/N^*. \tag{16}$$

If the predictors are independent, $k_e = k$ [from (13)]. For the 113 dependent grid-point predictors, $k_e \approx 11.6$. Thus, in terms of average sampling errors, when the first predictor is selected the 113 dependent predictors are equivalent to only $\sim$10 independent predictors. This point was alluded to in the discussion of Fig. 3. It provides an *a posteriori* reason for the selection of $\sim$10 EOFs in the prediction model of Section 2.

Livezey and Chen (1983) used a Monte Carlo simulation to evaluate the statistical significance of fields (sets) of correlations with individual grid points. As they noted, cross-correlations in the fields (dependence among grid points) reduces the effective number of degrees of "independent clusters" of grid points, corresponding to k_e in the present screening analysis.

c. Bounds on model skill

When $m = 1$, $S_H^{\text{crit}}(1 - \alpha)$ is an estimate of the square of the critical partial correlation coefficient to be used in the screening selection of predictors. Empirically (Fig. 4a), $N^*S_H^{\text{crit}}(1 - \alpha) \approx 9.4$ at the 5% level of significance for the 113 dependent grid-point predictors. When $\alpha = 0.01$, the value (not shown) is $\approx$12.7. The stepwise screening analysis of Section 2 was redone, entering individual predictors into the model, based upon their contribution to the reduction of variance. Predictors were accepted until the partial correlation coefficient fell below $[N^*S_H^{\text{crit}}(1 - \alpha)]^{1/2}$. The results for $\alpha = 0.05$ are summarized in Table 4. In general, fewer predictors are allowed into the model than when the 1% added reduction of variance cutoff was used (Table 3) without regard to statistical significance.

At the 5% level of significance, e (lower) and e_F

TABLE 4. As in Table 3, but for grid-point predictors selected with critical partial correlation coefficient cutoff at 5% significance level.

		m	S_H	e_H (km)	e (km)		e_F (km)	
					Lower	Upper	Lower	Upper
24 h	X'	3	0.214	271	272	285	274	298
	Y'	6	0.690					
48 h	X'	2	0.221	673	677	691	680	697
	Y'	4	0.449					
72 h	X'	3	0.211	1051	1054	1074	1063	1102
	Y'	2	0.346					

(lower) are both about equal to the corresponding estimated errors derived from the EOF analysis (Table 2). On average, e (upper) and e_F (upper) are about 3% and 5% greater, respectively. The estimated upper bound on e_F for a 24 h forecast is 298 km; with the EOF predictors, the estimate of e_F is 277 km. Thus, the use of EOFs reduces the 24 h forecast error on the rotated grid at most by ~8%. By comparison, grid rotation reduces the error by about 13% over that for the geographically oriented grid (SN).

Since a 5% level of significance is used, predictors with no true skill are rejected 95% of the time. Predictors with some true skill may also be rejected. Thus, the set of selected predictors tends to be limited to those with substantial skill. True model and forecast errors tend to approach their lower estimates, valid for preselected predictors. The use of a stricter 99% cutoff (not shown) reduces the number of predictors selected and increases the lower estimates to about 2% greater than the corresponding EOF estimates. The upper estimates are about 4% greater. When a 5% or 1% significance level is used, preselection and ordering of predictors with EOFs thus gives smaller expected true model and forecast errors than stepwise screening. This result supports the conclusion of Davis (1977) that preselection reduces errors. Even considering the uncertainties in the estimates of e and e_F, however, it can be concluded that on average the differences between the two methods in the present application are relatively small.

5. Summary

Preselection and ordering of predictors, such as EOFs, limits artificial skill and reduces sampling errors that degrade forecasts. Screening of predictors to select those that contribute most to the reduction of predictand variance allows greater opportunity for random errors and thus increases expected artificial skill over that for an equal number of preselected predictors. The greater the number of available predictors for screening, the greater the artificial skill. Dependence among predictors selected by screening, however, reduces sampling error below that for the same number of screened independent predictors. In the particular example discussed in Section 3, 113 grid-point height predictors were available to forecast storm tracks. For a single predictor selected by screening, dependence among the screened predictors reduced the expected artificial skill (when true skill is zero) by about 50%. As derived in Section 4, when the first selection is made the available grid-point predictors provide the screening program with a sampling error equivalent to that obtained from only about ten available preselected independent predictors. Moreover, the use of a proper critical partial correlation coefficient, determined from a Monte Carlo method in Section 4, restricts the number of predictors allowed in the forecast model. This restriction, in itself, limits artificial skill.

For the grid-point height predictors, on the average, the expected forecast errors derived from the screening of individual predictors were at worst about 5% greater than those from preselected and ordered EOFs. At best, the two methods of model development provided nearly equal expected forecast skills. The evaluation of estimated true model and expected forecast skills for preselected EOFs was straightforward. Estimates of bounds on these skills for screened predictors, with upper bounds from a Monte Carlo simulation, provide an alternative to estimates from an independent data sample. The method avoids removal of part of the developmental sample for the independent test. For the present model comparison, the bounds on skill are sufficiently precise to conclude that although the use of EOF techniques to preselect predictors may have small advantage, care in selection of individual predictors by screening can lead to comparable forecast ability.

Acknowledgments. Continuing discussions with Dudley Chelton have been of considerable value. His comments on an earlier version of this paper, and those of Bill Klein and John Walsh, have also helped the presentation. Dale Martin's drafting, and Gail Derr and Jorge Betancourt's typing skills greatly eased the preparation of the manuscript.

APPENDIX

List of Symbols

x', y' Coordinate axes in cross-track and along-track directions

$\mathbf{X}'$ = (X', Y'); storm displacements in (cross-, along-) track directions

$s(\mathbf{X}')$	Standard deviation of individual displacement components
$s(\text{tot})$	$s^2(\text{tot}) \equiv s^2(X') + s^2(Y')$
e_H	Combined standard hindcast error for both displacement components [defined in (8)]
e	Combined standard true model error [defined in (9)]
e_F	Combined standard expected forecast error [defined in (10)]
$e(\text{lower}), e(\text{upper})$	Bounds on combined error
S	True model skill
$\tilde{S}$	S, applied to a given data set
S_H	Hindcast skill (= reduction of variance)
S_A	Artificial skill
S_F	Forecast skill
$\langle(\ )\rangle$	Expected value
N	Number of cases (observations)
N^*	Effective number of independent observations
$\mathcal{N}$	Record length used in Monte Carlo method
m	Number of predictors selected
$\bar{m}$	Effective number of preselected predictors [determined from average artificial skill in (11)]
$\hat{m}$	Effective number of preselected predictors [determined from critical hindcast skill in (12)]
$m_i, (S_H)_i$	m, S_H for models $i = 1, 2$
k	Number of available predictors
k_e	Equivalent number of independent predictors available [defined in (16)]
$S_H^{\text{crit}}(1-\alpha)$	Critical value of hindcast skill
$F_{\text{crit}}(1-\alpha)$	Critical value of F-statistic
$\chi_m(1-\alpha)$	100 $(1-\alpha)$ percentage point of chi-squared distribution with m degrees of freedom
α	Significance level
α'	Empirical α from Monte Carlo technique

REFERENCES

Chelton, D. B., 1983: Effects of sampling errors in statistical estimation. *Deep-Sea Res.*, **30,** 1083–1103.

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.

——, 1977: Techniques for statistical analysis and prediction of geophysical fluid systems. *Geophys. Astrophys. Fluid Dyn.*, **8,** 245–277.

Draper, N. R., and H. Smith, 1981: *Applied Regression Analysis,* 2nd ed. Wiley, 709 pp.

Klein, W. H., 1983: Objective specification of monthly mean surface temperature from mean 700 mb heights in winter. *Mon. Wea. Rev.*, **111,** 674–691.

——, and J. E. Walsh, 1983: A comparison of pointwise screening and empirical orthogonal functions in specifying monthly surface temperature from 700 mb data. *Mon. Wea. Rev.*, **111,** 669–673.

Kleinbaum, B., and L. Kupper, 1978: *Applied Regression Analysis and Other Multivariate Methods.* Wadsworth, 556 pp.

Livezey, R. E., and W. Y. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.*, **111,** 46–59.

Lorenz, E. N., 1956: Empirical orthogonal functions and statistical weather prediction. Rep. No. 1, Statistical Forecasting Project, Dept. Meteor., Massachusetts Institute of Technology, 49 pp. [AFCRC-TN-57-256; NTIS AD-110-268.]

——, 1977: An experiment in nonlinear statistical weather prediction. *Mon. Wea. Rev.*, **105,** 590–602.

Lund, I. A., 1970: A Monte Carlo method for testing the statistical significance of a regression equation. *J. Appl. Meteor.*, **9,** 330–332.

Miller, R. G., 1962: *Statistical Prediction by Discriminant Analysis. Meteor. Monogr.,* No. 25, Amer. Meteor. Soc., 54 pp.

Neumann, C. J., and M. B. Lawrence, 1975: An operational experiment in the statistical-dynamical prediction of tropical cyclone motion. *Mon. Wea. Rev.*, **103,** 665–673.

——, and J. M. Pelissier, 1981: Models for the prediction of tropical cyclone motion over the North Atlantic: An operational evaluation. *Mon. Wea. Rev.*, **109,** 522–538.

——, M. B. Lawrence and E. L. Caso, 1977: Monte Carlo significance testing as applied to statistical tropical cyclone prediction models. *J. Appl. Meteor.*, **16,** 1165–1174.

Overland, J. E., and R. W. Preisendorfer, 1982: A significance test for principal components applied to a cyclone climatology. *Mon. Wea. Rev.*, **110,** 1–4.

Shaffer, A. R., and R. L. Elsberry, 1982: A statistical–climatological tropical cyclone track prediction technique using an EOF representation of the synoptic forcing. *Mon. Wea. Rev.*, **110,** 1945–1954.

Shapiro, L. J., and C. J. Neumann, 1984: On the orientation of grid systems for the statistical prediction of tropical cyclone motion. *Mon. Wea. Rev.*, **112,** 188–199.

Siegel, S., 1956: *Nonparametric Statistics.* McGraw-Hill, 312 pp.