JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 90, NO. C5, PAGES 8995–9005, SEPTEMBER 20, 1985

Statistics for the Evaluation and Comparison of Models

CORT J. WILLMOTT,[1] STEVEN G. ACKLESON,[2] ROBERT E. DAVIS,[1] JOHANNES J. FEDDEMA,[1]
KATHERINE M. KLINK,[1] DAVID R. LEGATES,[1] JAMES O'DONNELL,[2] AND CLINTON M. ROWE[1]

Procedures that may be used to evaluate the operational performance of a wide spectrum of geophysical models are introduced. Primarily using a complementary set of difference measures, both model accuracy and precision can be meaningfully estimated, regardless of whether the model predictions are manifested as scalars, directions, or vectors. It is additionally suggested that the reliability of the accuracy and precision measures can be determined from bootstrap estimates of confidence and significance. Recommended procedures are illustrated with a comparative evaluation of two models that estimate wind velocity over the South Atlantic Bight.

1. INTRODUCTION

Over the past two decades, the development of numerical models that can simulate atmospheric, oceanic, or other geophysical processes has increasingly become a major focus of the physical science community. One important aspect of the model development process, the evaluation of model performance, has received relatively little attention in the geophysical literature. The problem is magnified by a lack of agreement among scientists regarding the most suitable measures and procedures for determining (1) model accuracy, i.e., the extent to which model-predicted events approach a corresponding set of independently obtained, reliable observations (usually measured), and precision, i.e., the degree to which model-predicted values approach a linear function of the reliable observations, and (2) the extent to which the model's behavior is consistent with prevailing scientific theory. Evaluations of the former kind are often referred to as "operational," whereas the latter variety frequently are termed "scientific." Successful model evaluations are clearly comprised of both operational and scientific examination, although the operational evaluation of precision and accuracy often provides the most tangible means of establishing model credibility. For this reason, the development, examination, and recommendation of methods that may be used to determine and compare the accuracy and precision of models are of primary concern. Guidelines for the objective scientific evaluation of model performance are also needed, but, since they are more intrinsically linked to a specific model or problem, fewer general recommendations can be made.

Within this paper, our goal is to modify and extend salient scalar-based evaluation measures to model-predicted and observed variables whose elements may be vectors or directions. Such general statistics are required for the operational evaluation of those simulation models that predict geophysical phenomena such as wind or ocean current direction or velocity. A computer-intensive method of estimating the reliability associated with various evaluation indices, "the bootstrap" [*Efron and Gong*, 1983], also is discussed as an alternative to standard parametric, statistical ways of determining confidence and "significance." In several respects, bootstrap methods appear to be superior to many of the other nonparametric techniques as well [cf. *Efron*, 1981*a*; *Efron and Gong*, 1983]. Methodological points are illustrated by an operational evaluation of two models that estimate wind velocity over the continental shelf.

Previous discussions [e.g., *Fox*, 1981; *Preisendorfer and Barnett*, 1983; *Willmott*, 1984] also have focused on operational evaluation, but in virtually all cases, the elements of the model-predicted and observed variables or fields were scalars. The relative abilities of Pearson's product-moment correlation coefficient and a variety of difference measures to compare one-dimensional variables, for instance, have been examined by *Fox* [1981, 1984], *Willmott* [1981, 1982, 1984], and others, while *Preisendorfer and Barnett* [1983] have extended a set of statistics to the comparison of model-predicted and observed fields. (For a review of scalar-based evaluation methods, we suggest the following sample papers: *Nash and Sutcliffe* [1970], *McCuen and Snyder* [1975], *Johnson and Bras* [1980], *Willmott* [1981, 1982, 1984], *Fox* [1981, 1984], *Won* [1981], *MacKay and Bornstein* [1981], *Rao and Visalli* [1981], *Gordon* [1982], *James and Burges* [1982], *Harr et al.* [1983].) Such studies typically reach different conclusions regarding the efficacy of correlation, tests of statistical significance (both parametric and nonparametric), and certain difference measures, which underscores the uncertainty that researchers face when testing a model, comparing two or more models, or selecting the most appropriate model from the literature. However, there is widespread agreement on the virtues of data-display graphics and difference measures in general.

Preisendorfer and Barnett's [1983] paper is of particular interest because they too develop measures for higher-dimensional problems, but their recommendations differ from ours in several important respects. Their "trinity [of] statistics" are all dimensionless and rest on a well-known decomposition of the difference variable into its first and second moments [cf. *Fox*, 1981], the second moment then being partitioned into correlated and uncorrelated terms [cf. *Bevington*, 1969]. Our measures, on the other hand, ascribe the "model-reality" differences ("error") to the model, and we explicitly treat vector or directional observations in addition to scalars. *Preisendorfer and Barnett* [1983] also discuss three nonparametric approaches to the estimation of significance, but they do not consider the bootstrap nor do they directly concern themselves with the estimation of a statistic's confidence or reliability.

2. MEASURES OF ACCURACY AND PRECISION

Difference or error measures, especially the root-mean-square error (RMSE), are increasingly being used in the comparison and evaluation of simulation models and are be-

[1] Center for Climatic Research, Department of Geography, University of Delaware, Newark.
[2] College of Marine Studies, University of Delaware, Newark.

Paper number 5C0173.
0148-0227/85/005C-0173$05.00

ginning to replace the correlation- and skill-based indices as the paramount measures of accuracy. This represents a desirable trend because the correlation- and skill-based measures are not consistently related to model accuracy, but since the difference indices have not been thoroughly investigated, their effective application to a wide variety of geophysical models and data requires further exploration. With this in mind, the array of difference measures previously discussed by *Willmott* [1981, 1982, 1984] is modified so that it may be used to evaluate and compare geophysical models that predict vector as well as scalar variables or fields.

Geophysical variables are often measured and modeled at discrete times and/or locations even though they may, in reality, be continuous. For this reason, geophysical variables are assumed to be discrete within this treatment. When the elements of a model-predicted (**P**) and an observed (**O**) variable are scalars, a difference variable **D**, where $\mathbf{D} = \mathbf{P} - \mathbf{O}$, can readily be defined. Indices that purport to describe **D** are called "difference measures." (Uppercase bold notation, i.e., **P**, **O** and **D**, is used to indicate variables or sets whose elements are scalars, directions, or vectors. The jth element of such a variable is given in lowercase bold, e.g., $\mathbf{d}_j$.) When the elements of **P** and **O** are vectors or directions, the calculation of **D** can be made by vector subtraction of the corresponding elements.

A set of difference measures that describe **D** when its elements are scalars [*Willmott*, 1982, 1984] subsequently can be generalized to problems in which the elements of **D** are directions or vectors. These measures include the root-mean-square error (RMSE), the systematic root-mean-square error (RMSE_s), the unsystematic root-mean-square error (RMSE_u), and the index of agreement (d_2). Additional indices such as the mean absolute error (MAE) and a modified version of the index of agreement (d_1) also are presented for reasons discussed below.

With the exception of RMSE, our measures differ from those recommended by *Fox* [1981, 1984], *Preisendorfer and Barnett* [1983] and most others primarily because we assume that all the error is contained within **P** and that **O** is error free. If **O** is known to contain nontrivial errors, they should be excised prior to the application of our statistics. We also consider the dimensions of **P** and **O** to be commensurate and important for interpretational purposes, and therefore virtually all of our statistics (save d_1 and d_2) preserve the metric. Where a statistic is dimensionless (d_1 and d_2) and consequently difficult to physically interpret [cf. *Preisendorfer and Barnett*, 1983], we assign the limits of 0.0 and 1.0 to facilitate understanding.

Consider the vector $\mathbf{d}_j$ (the jth element of **D**) as the resultant of the jth model-predicted ($\mathbf{p}_j$) minus the jth observed ($\mathbf{o}_j$) vector or direction, that is,

$$\mathbf{d}_j = \mathbf{p}_j - \mathbf{o}_j \tag{1}$$

where, for our purposes, j refers to a location in time or space. The subscript j could also refer to a location in time and space, or even to position in a multivariate hyperspace, but the difference measures described below are virtually meaningless if the dimensions of the hyperspace cannot be made satisfactorily commensurate. Average error subsequently may be described by

$$E^{1/\gamma} = \left[\left(\sum_{j=1}^{N} \omega_j |\mathbf{d}_j|^\gamma\right) \Big/ \left(\sum_{j=1}^{N} \omega_j\right)\right]^{1/\gamma} \qquad 0 < \gamma \tag{2a}$$

where N is the number of directions or vectors contained in **D** and ω_j is a scalar weight that corrects $|\mathbf{d}_j|$ when j is temporally or spatially overrepresentative or underrepresentative ($\omega_j \neq \bar{\omega}$). Commonly, it is assumed that $\omega_j = 1$ for all j, but when j refers to observations in an irregularly spaced or systematically varying time or space series, each ω_j must represent the relative size of the jth interval or cell size in order for $E^{1/\gamma}$ to be unbiased. When comparing general circulation model (GCM) predicted and observed fields at the nodes (j) of a latitude-longitude grid, for instance, each ω_j must correct for the latitudinal variation in the grid-cell size. The important special cases of $E^{1/\gamma}$ are

$$E^{1/\gamma} = \text{MAE} \qquad \gamma = 1 \tag{2b}$$

$$E^{1/\gamma} = \text{RMSE} \qquad \gamma = 2 \tag{2c}$$

A relative average error similarly can be defined as

$$d_\gamma = 1 - \left[\sum_{j=1}^{N} \omega_j |\mathbf{d}_j|^\gamma\right] \Big/ \left[\sum_{j=1}^{N} \omega_j (|\mathbf{p}_j - \bar{\mathbf{o}}| + |\mathbf{o}_j - \bar{\mathbf{o}}|)^\gamma\right] \tag{3}$$

$$0 < \gamma$$

where $\bar{\mathbf{o}}$ is the weighted mean direction or vector of the elements contained in **O**, i.e.,

$$\bar{\mathbf{o}} = \sum_{j=1}^{n} \omega_j \mathbf{o}_j \Big/ \sum_{j=1}^{n} \omega_j$$

Pertinent special cases of (3) are *Willmott*'s [1981] index of agreement (d_2) and a modified index of agreement (d_1).

Values of γ other than 1 or 2 may be appropriate in particular situations, but $\gamma = 1$ and $\gamma = 2$ have certain advantages. In the absence of any geophysical justification, values of γ other than 1 represent an arbitrary weighting of each $\mathbf{d}_j$, and therefore, with the exception of $\gamma = 2$, they are avoided. When $\gamma = 2$, $E^{1/\gamma}$ and d_γ are set in the familiar format of variance, which not only facilitates interpretation but also enhances further mathematical or statistical analysis. A useful decomposition of the average error into its systematic and unsystematic components, for instance, can be accomplished using ordinary least squares (OLS). *Willmott* [1981, 1984] describes this decomposition for scalar elements, and here we generalize it.

For direction or vector elements, the systematic portion of the error can be written

$$\text{RMSE}_s = \left[\sum_{j=1}^{N} \omega_j |\hat{\mathbf{p}}_j - \mathbf{o}_j|^2 \Big/ \sum_{j=1}^{N} \omega_j\right]^{0.5} \tag{4}$$

while the unsystematic part is

$$\text{RMSE}_u = \left[\sum_{j=1}^{N} \omega_j |\hat{\mathbf{p}}_j - \mathbf{p}_j|^2 \Big/ \sum_{j=1}^{N} \omega_j\right]^{0.5} \tag{5}$$

where $\hat{\mathbf{p}}_j$ is an OLS estimate of $\mathbf{p}_j$, which is derived from the temporally or spatially weighted (by ω) regression of **P** on **O**. When $\mathbf{p}_j$, $\mathbf{o}_j$, and $\hat{\mathbf{p}}_j$ are directions or vectors, the estimation of $\mathbf{p}_j$ requires a separate regression for each component. Equations (4) and (5) are a complete partitioning of the error, since $\text{RMSE}^2 = \text{RMSE}_s^2 + \text{RMSE}_u^2$. Linear bias produced by a model is described by RMSE_s, whereas RMSE_u may be interpreted as a measure of precision.

Difference measures are easily interpreted because they are scalars, that is, RMSE, RMSE_s, RMSE_u, and MAE all take on the units of $|\mathbf{d}_j|$, whereas d_1 and d_2 are dimensionless and bounded by 0 and 1. Once again, *Willmott* [1981, 1982, 1984] has discussed the interpretation of these indices with scalar data; therefore the following paragraph is limited to direction and vector applications.

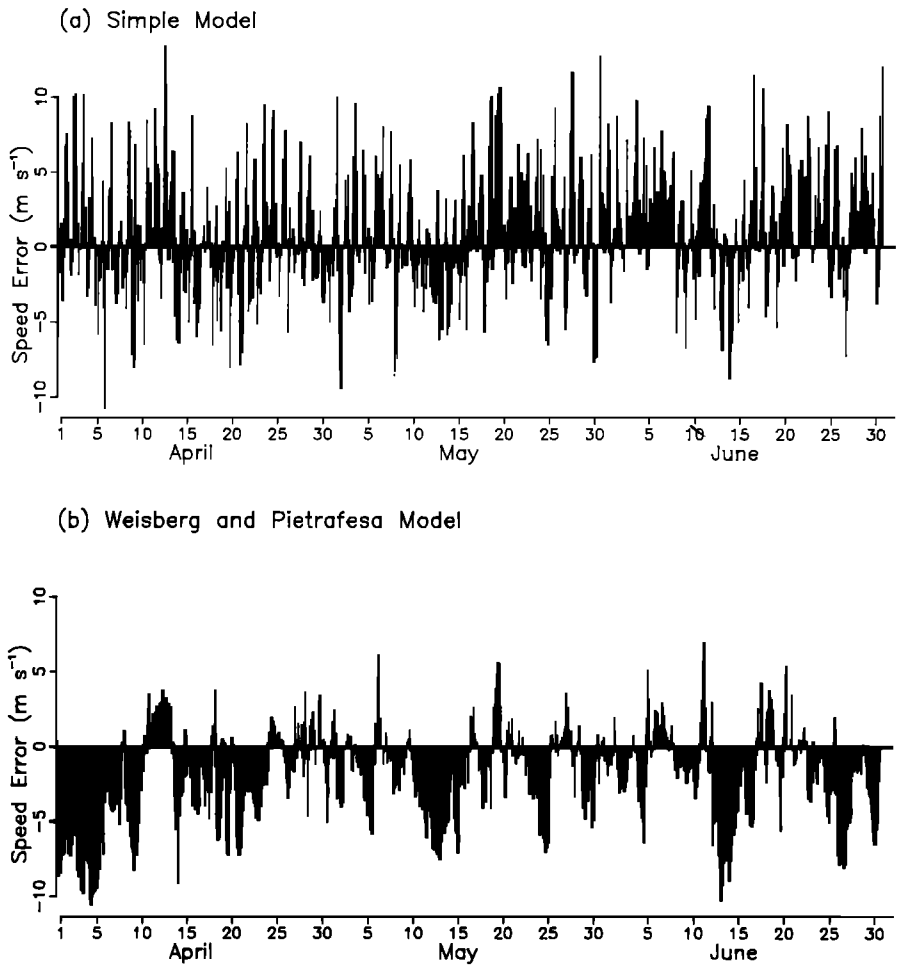

Fig. 1. Time series plots of the 3-hourly differences between (*a*) simple model-predicted and observed wind speeds and (*b*) W-P model-predicted and observed wind speeds for spring 1979 at NDBC data buoy 41002.

When used with directional or vector data, RMSE and MAE describe the average magnitude of **D** in N space, while the relative measures, d_1 and d_2, depict the degree to which **D** approaches the null set. For all γ, $d_\gamma = 1$ when **D** is the null set and d_γ approaches 0 as |**D**| approaches the combined variability in **P** and **O** about $\bar{o}$. It should be noted that RMSE $\geq$ MAE and $d_2 \geq d_1$ which makes both RMSE and d_1 conservative measures of average error. Specifically, the extent to which RMSE^2(MSE) exceeds MAE^2 is equal to the variability of the $|\mathbf{d}_j|$'s about MAE. A similar relationship exists between d_1 and d_2. Normally, RMSE, RMSE$_s$, RMSE$_u$, and MAE are unbounded on the upper end; however, when the elements of **D** are directions, their upper limit becomes 2. Interpretation of RMSE, RMSE$_s$, RMSE$_u$, and MAE when the $\mathbf{d}_j$'s are directions then may be enhanced by converting to angular measure. Illustrating for RMSE, the translation is

$$\varepsilon_2 = 2 \sin^{-1} (\text{RMSE}/2) \qquad 0 \leq \varepsilon_2 \leq \pi \tag{6}$$

When these difference measures are accompanied by data-display graphics, summary univariate statistics (e.g., the vector means of **P** and **O**), estimates of reliability, and salient sensitivity studies, they form the base of a comprehensive approach to the evaluation of model performance.

3. Reliability and Significance Estimation

Dissatisfaction with the standard statistical approaches to the estimation of reliability and significance, principally through parametric confidence estimates and hypothesis tests, prompted *Willmott* [1981, 1982, 1984] to admonish their use in the area of model evaluation in favor of an informed scientific interpretation of the accuracy measures. Several other researchers [e.g., *Fox*, 1981] have also recognized the problems associated with testing an accuracy measure for significance or using postulated, underlying frequency distributions to establish confidence bounds; nevertheless, they cautiously recommend and use such procedures. Even though Fox and Willmott proposed quite dissimilar approaches to reliability evaluation, both recommendations stem from the frequently observed inadequacies associated with the traditional methods and the perception that no better statistical methods existed. It now appears, however, that certain nonparametric statistical procedures provide credible, quantitative estimates of reliability [*Efron*, 1981*a*, *b*; *Efron and Gong*, 1983] and perhaps significance.

Once an estimate of model accuracy ($\hat{\Theta}$) has been calculated from an N-element difference variable (**D**), we principally seek a range of values, a confidence interval, through which $\hat{\Theta}$ would be expected to vary. The magnitude of such an interval is then a measure of the reliability of $\hat{\Theta}$. A concomitant parametric test for the statistical significance of $\hat{\Theta}$ is still thought to be questionable [*Willmott*, 1981, 1982, 1984], but it may be useful to calculate and report the probability that $\hat{\Theta}$ is larger or smaller than some value. This significance-like calculation will be introduced following the discussion of confidence inter-

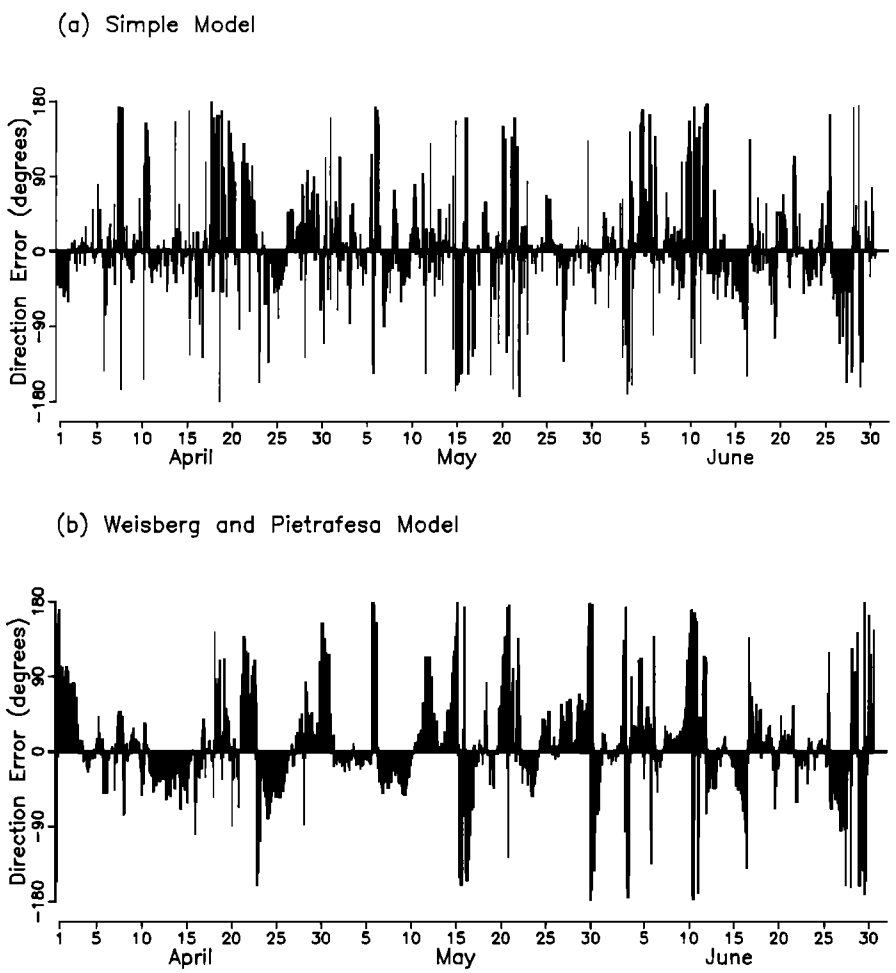

Fig. 2. Time series plots of the 3-hourly differences between (*a*) simple model-predicted and observed wind directions and (*b*) W-P model-predicted and observed wind directions for spring 1979 at NDBC data buoy 41002.

vals. Before describing nonparametric approaches to the estimation of confidence intervals, it is useful to review their conceptual, parametric roots.

A well-behaved probability distribution, such as the student's-*t* distribution, is traditionally postulated, and this nets a characteristic probability density function ($f(\Theta)$). The upper (b) and lower (a) bounds of a confidence interval can then be selected such that

$$\int_a^b f(\Theta)\, d\Theta = P\{a < \Theta < b\} = 1 - \alpha \tag{7}$$

where $P\{a < \Theta < b\}$ is the probability that Θ falls between a and b, and α is an a priori chosen probability (often 0.05). Recall also that $f(\Theta)$ has the properties $f(\Theta) \geq 0$ and

$$\int_{-\infty}^{\infty} f(\Theta)\, d\Theta = 1$$

and that $\hat{\Theta}$ is ordinarily centered between a and b. Usually, however, the true character of $f(\Theta)$ is unknown, which suggests that its a priori identification as student's-*t* or normal, for example, and the ensuing estimation of a and b, are somewhat speculative.

Efron [1981*a*, *b*] and *Efron and Gong* [1983] alternately suggest that a nonparametric approach called the "bootstrap" may be used to estimate the reliability of $\hat{\Theta}$. Like other nonparametric procedures, the bootstrap makes no a priori assumption about the shape of $f(\Theta)$, but it constructs an empirical distribution function $\hat{f}(\Theta)$ by resampling a set of N independent observations. The bootstrap is superior to other nonparametric methods for evaluating the confidence (reliability) of $\hat{\Theta}$, because the bootstrap estimate of the standard error is the nonparametric maximum likelihood estimate of the standard error [*Efron*, 1981*a*]. *Efron* [1981*a*] goes on to say that "if we want to do better, we have to use some form of estimation which is not truly nonparametric."

For our purposes, the N observations are the elements of **D**. According to Efron, a bootstrap sample (**D***) of size N is randomly chosen one element at a time from **D** with replacement. Once a **D*** has been selected, a bootstrap measure of accuracy ($\hat{\Theta}^*$) may be calculated. If this process is repeated B times, it yields an empirically derived frequency distribution ($\hat{f}(\Theta)$) which approaches the true $f(\Theta)$ as B becomes large. The degree to which $\hat{f}(\Theta)$ approaches $f(\Theta)$, of course, also depends on N or the extent to which **D** reflects its parent population. Both the central tendency ($\hat{\Theta}_\cdot^*$) and the dispersion ($\hat{\sigma}^*$) associated with $\hat{f}(\Theta)$ now can be expressed as

$$\hat{\Theta}_\cdot^* = B^{-1} \sum_{k=1}^{B} \hat{\Theta}_k^* \tag{8}$$

and

$$\hat{\sigma}^* = \left[(B-1)^{-1} \sum_{k=1}^{B} (\hat{\Theta}_k^* - \hat{\Theta}_\cdot^*)^2 \right]^{0.5} \tag{9}$$

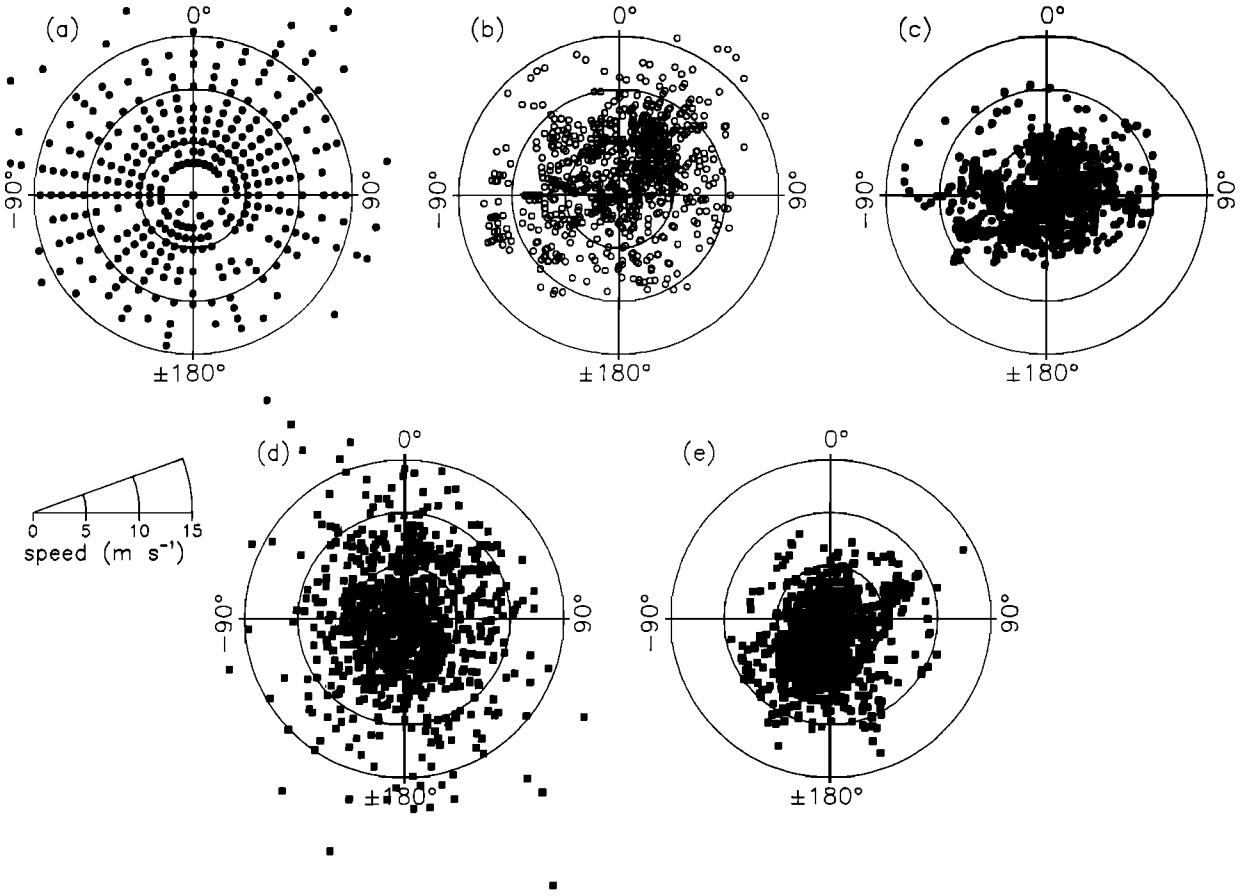

Fig. 3. Polar scatterplots of (*a*) simple model-predicted (solid circles), (*b*) observed (open circles), and (*c*) W-P model-predicted (solid circles) 3-hourly wind velocities at NDBC data buoy 41002 during spring 1979. Polar scatterplots of the differences between (*d*) simple model-predicted and observed (solid squares) and (*e*) W-P model-predicted and observed wind velocities (solid squares) are also given.

where k refers to the kth bootstrap sample. A confidence interval subsequently can be defined such that

$$P\{(\hat{\Theta}_{\cdot}^{*} - \zeta_1\hat{\sigma}^{*}) < \Theta < (\hat{\Theta}_{\cdot}^{*} + \zeta_2\hat{\sigma}^{*})\} = 1 - \alpha \qquad (10)$$

where ζ_1 and ζ_2 are the magnitudes of the confidence bounds in $\hat{\sigma}^{*}$ units and, once again, α has been set by the researcher. Since $\hat{f}(\Theta)$ is discrete, the confidence limits for $\hat{\Theta}$ are taken in percentile form as

$$\#\,[\hat{\Theta}_k^{*} \le (\hat{\Theta}_{\cdot}^{*} - \zeta_1\hat{\sigma}^{*})]B^{-1} = \alpha/2 \qquad (11a)$$

and

$$\#\,[\hat{\Theta}_k^{*} \ge (\hat{\Theta}_{\cdot}^{*} + \zeta_2\hat{\sigma}^{*})]B^{-1} = \alpha/2 \qquad (11b)$$

where # denotes the number of $\hat{\Theta}_k^{*}$'s that satisfy the inequality. *Efron and Gong* [1983] further describe a means to correct a confidence interval for statistical bias or when $\hat{f}(\Theta)$ exhibits an asymmetry, although we do not discuss or use this refinement.

Bootstrap statistics are not as efficient as their parametric counterparts in that they substitute "raw computing power for theoretical analysis"; however, they provide "crude but trustworthy nonparametric answers" when "parametric assumptions are difficult to justify" [*Efron and Gong*, 1983]. Some principal advantages of nonparametric estimates and bootstrap methods in particular, over their parametric counterparts, include (1) assumptions about the underlying but unknown frequency distribution of Θ do not effect the methods' validity or cloud interpretation, and (2) confidence can readily be established for any accuracy measure of interest even if its theoretical distributional characteristics previously have not been derived and cataloged.

When a histogram of the $\hat{\Theta}_k^{*}$'s is plotted, it can enhance the interpretation of confidence as well as bias and it also represents a comprehensive expression of the reliability of $\hat{\Theta}$.

Even though a determination of $(\hat{\Theta}_{\cdot}^{*} - \zeta_1\hat{\sigma}^{*})$ and $(\hat{\Theta}_{\cdot}^{*} + \zeta_2\hat{\sigma}^{*})$ is sufficient for most evaluation problems, it is also possible to make interpretations of significance, that is, whether or not $\hat{\Theta}$ is meaningful. Take, for instance, the generic null (H_0) and alternative (H_a) hypotheses:

$$H_0\colon \hat{\Theta} = c$$

$$H_a\colon \hat{\Theta} > c$$

where c can be any value, although the test is often performed with $c = 0$. With the bootstrap estimation of $\hat{f}(\Theta)$, the pertinent probability can be cast as

$$P\{\hat{\Theta} > c\} = 1 - \alpha_0 \approx \#\,[\hat{\Theta}_k^{*} > c]B^{-1} \qquad (12)$$

where α_0 is the observed probability level. This calculation is essentially of the same form as (11) and, for this reason, provides little new information. Nonetheless, a value of α_0 which is smaller than the probability of making a type 1 error (α), set previously by the researcher, may be interpreted as significant, and thus H_0 may be rejected and H_a accepted. In a similar

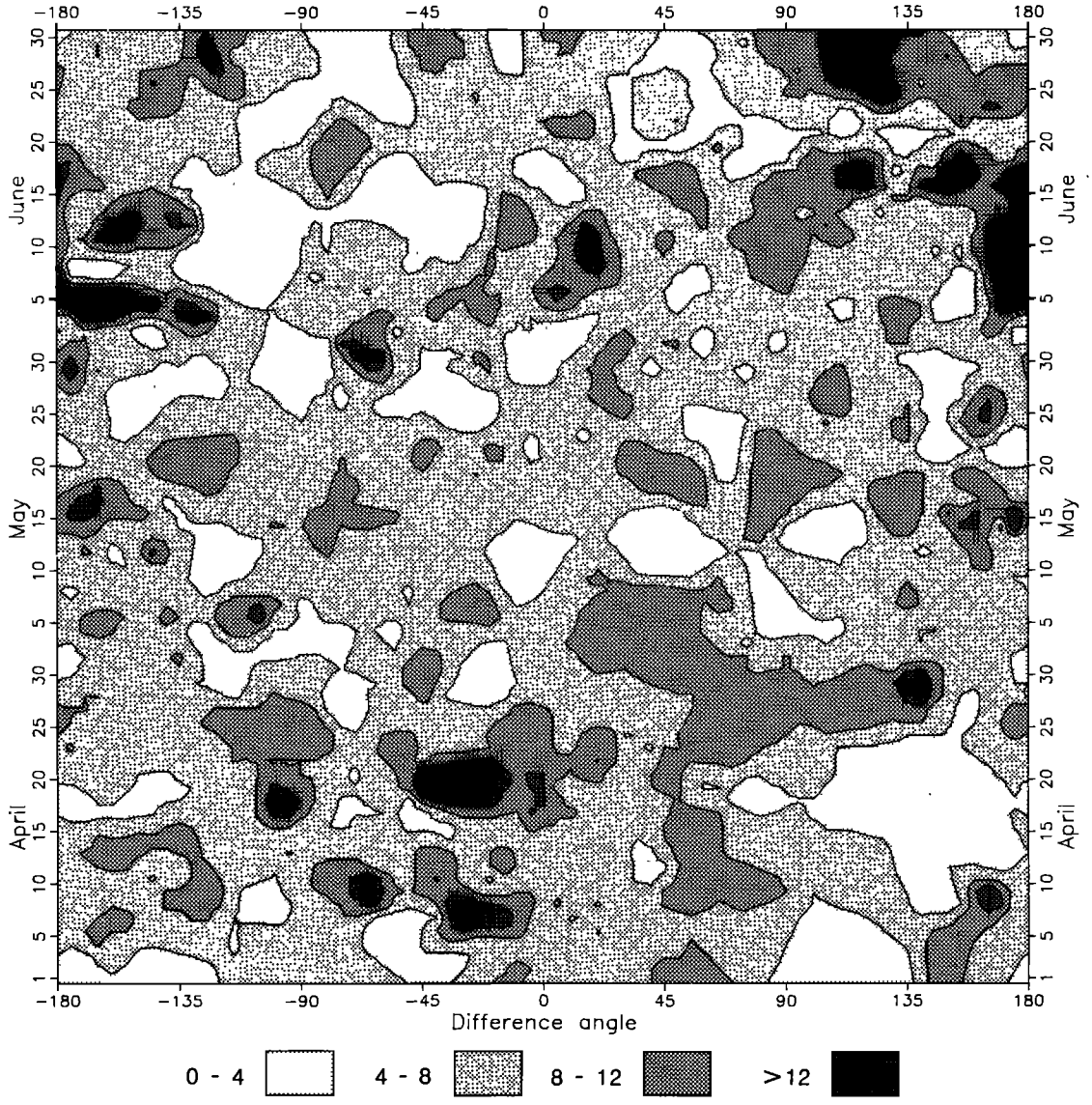

Fig. 4*a*

Fig. 4. Plots of the spring 1979 velocity error field (m s^{-1}) associated with (*a*) the simple model and (*b*) W-P model. They were mapped on a cylinder and then projected into a planar difference-angle, time space.

vein, it is possible to examine the difference between model accuracies by evaluating

$$H_0: \hat{\Theta}_1 = \hat{\Theta}_2$$

$$H_a: \hat{\Theta}_1 > \hat{\Theta}_2$$

on the basis of the bootstrapped frequency distribution associated with $(\hat{\Theta}_1 - \hat{\Theta}_2)$. While the bootstrap has certain advantages over parametric tests, we reiterate that any statistically derived interpretation of significance should be made with caution [*Willmott*, 1984].

4. Operational Evaluation and Comparison of Two Wind Velocity Models

A general problem in oceanography is the development of accurate models (e.g., transfer functions) that can be used to extrapolate readily available coastal meteorological observations to offshore environs of interest. Wind velocity is of considerable interest because of its overriding influence on the sea state and circulation and because its over-the-ocean measurements are few. To illustrate these model evaluation methods, two models that estimate near-surface wind velocities over the South Atlantic Bight (SAB) are compared.

One of these models represents a simple rule-of-thumb in which the wind speed over the SAB is assumed to be twice the observed coastal speed while the direction is taken as that observed at the coastal meteorological station. This transfer function is referred to as the "simple model." The second model, developed by *Weisberg and Pietrafesa* [1983], is a statistically derived transfer function between the measured wind at Charleston, South Carolina (CHS), and that observed at the NOAA Data Buoy Center (NDBC) data buoy 41002 (≈300 km east of CHS). This model, referred to as the W-P model, was specified with three-hourly wind velocities observed during the spring seasons of 1976–1978. Since our focus is on operational evaluation, no attempt is made to examine the scientific basis of either model.

Our evaluation of these models' performances is based upon CHS and buoy 41002 data observed in April, May, and June of 1979, the same data used by *Weisberg and Pietrafesa* [1983]

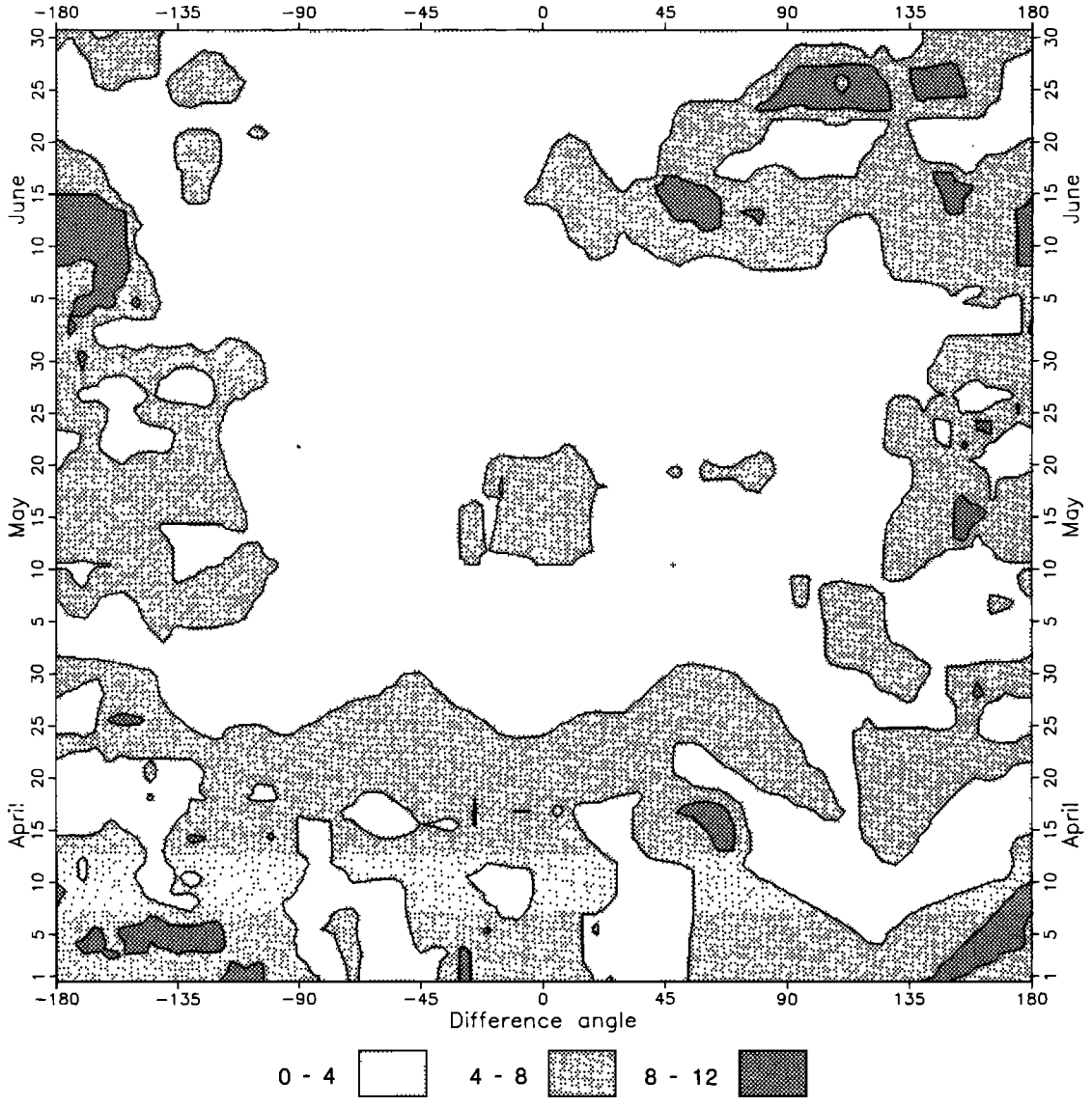

Fig. 4*b*

to examine the performance of their model for spring. The measured as well as the predicted time series ($\mathbf{P}^{wp}$) were provided by Weisberg and Pietrafesa. Using the CHS observed data, a second predicted time series at buoy 41002 was estimated according to the simple model ($\mathbf{P}^{s}$). The following discussion represents our evaluation of the ability of the W-P model as well as the simple model to estimate three-hourly wind velocities at buoy 41002 from CHS data during the spring. Since the time differential is constant at 3 hours, we assume that $\omega_j = \bar{\omega} = 1.0$. *Weisberg and Pietrafesa* [1983] also examined the performance of their model during the summer, fall, and winter of 1979, but we confine our evaluation to spring.

Principally, we wish to compare $\mathbf{P}^{s}$ and $\mathbf{P}^{wp}$ to $\mathbf{O}$, the set of 728 spring 1979 wind velocities drawn from the record of buoy 41002. As suggested previously [*Willmott*, 1981, 1982, 1984], it is useful to begin such an evaluation with graphic representations of the relationships between the predicted and observed variables. In this vein, *Weisberg and Pietrafesa* [1983] present useful time series plots of the north and east components associated with $\mathbf{P}^{wp}$, $\mathbf{O}$, and the observed wind velocities at CHS. They also give frequency domain plots of the coherence, amplitude, and phase associated with the differences between $\mathbf{P}^{wp}$ and $\mathbf{O}$ in order to illustrate the nature of the difference variable ($\mathbf{D}^{wp}$), where $\mathbf{D}^{wp} = \mathbf{P}^{wp} - \mathbf{O}$. From these graphics alone, however, it is difficult to appreciate the nature of the errors described by $\mathbf{D}^{wp}$ because each error ($\mathbf{d}_j^{wp}$) is decomposed into components, cartesian or spectral, prior to graphic summary. Because of the added dimension, the graphic presentation of time or space series whose elements are vectors is more complex than the presentation of scalar series, and several additional plots are needed. It might be useful, for example, to compute and plot as time series the $\mathbf{d}_j^{wp}$'s and $\mathbf{d}_j^{s}$'s associated with the wind speeds or directions separately (Figures 1 and 2), assuming that one of these aspects of the wind was of interest independently of the other. If speeds were of primary concern, say, for the estimation of the sea state, comparable time series plots of $\mathbf{D}^{wp}$ and $\mathbf{D}^{s}$ computed for the speeds alone reveal considerable temporal autocorrelation within the error portion of $\mathbf{P}^{wp}$, which could possibly be damped. The average error in the W-P model estimate, however, is considerably less than the error associated with the simple model estimates (Figure 1). A less pronounced temporal autocorrelation is also evident within the error portion

TABLE 1. Statistical Measures of the Simple Model's Ability to Estimate Spring 1979 Wind Velocities Over the South Atlantic Bight

Statistic	Value	Confidence Limits (95%) Lower	Upper
$\lvert\bar{o}\rvert$	2.02	1.63	2.37
$\angle\bar{o}$	2.32	−9.78	12.09
s_o	6.97	6.74	7.23
$\lvert\bar{p}^s\rvert$	1.59	1.17	1.97
$\angle\bar{p}^s$	9.38	−7.62	25.25
s_p	8.62	8.26	8.92
MAE^s	6.54	6.26	6.80
$RMSE^s$	7.63	7.31	7.96
$RMSE_s{}^s$	2.57	*	
$RMSE_u{}^s$	7.18		
$d_1{}^s$	0.53	0.51	0.55
$d_2{}^s$	0.74	0.72	0.77

*Values not computed because $RMSE_s$ and $RMSE_u$ are not independent of one another.

TABLE 2. Statistical Measures of the W-P Model's Ability to Estimate Spring 1979 Wind Velocities Over the South Atlantic Bight

Statistic	Value	Confidence Limits (95%) Lower	Upper
$\lvert\bar{o}\rvert$	2.02	1.69	2.38
$\angle\bar{o}$	2.32	−9.55	12.15
s_o	6.97	6.73	7.18
$\lvert\bar{p}^{wp}\rvert$	0.14	0.02	0.50
$\angle\bar{p}^{wp}$	−100.19	94.93	60.40
$s_p{}^{wp}$	5.16	4.96	5.34
MAE^{wp}	4.79	4.60	4.97
$RMSE^{wp}$	5.45	5.25	5.63
$RMSE_s{}^{wp}$	4.06	*	
$RMSE_u{}^{wp}$	3.63		
$d_1{}^{wp}$	0.57	0.55	0.59
$d_2{}^{wp}$	0.80	0.78	0.81

*Values not computed because $RMSE_s$ and $RMSE_u$ are not independent of one another.

of the W-P model estimates of spring wind directions (Figure 2).

Even though the graphics presented above (Figures 1 and 2) and those given by *Weisberg and Pietrafesa* [1983] illustrate the temporal patterns within selected error components, it would additionally be useful to examine the direction and speed together. One way this may be accomplished is with polar scatterplots of each $\mathbf{d}_j^s$ and $\mathbf{d}_j^{wp}$ (Figures 3*d* and 3*e*). These plots suggest that the simple model produces a relatively large but unsystematic error in both the directional and the speed components, i.e., in velocity. The W-P model is much better at estimating the speed, but it does not reproduce direction well. This tendency is also apparent in the polar scatterplots of $\mathbf{P}^{wp}$ and $\mathbf{O}$, where the east-west portion of the variance is represented well by the W-P model but the north-south portion is not (Figures 3*b* and 3*c*). Direct translation of

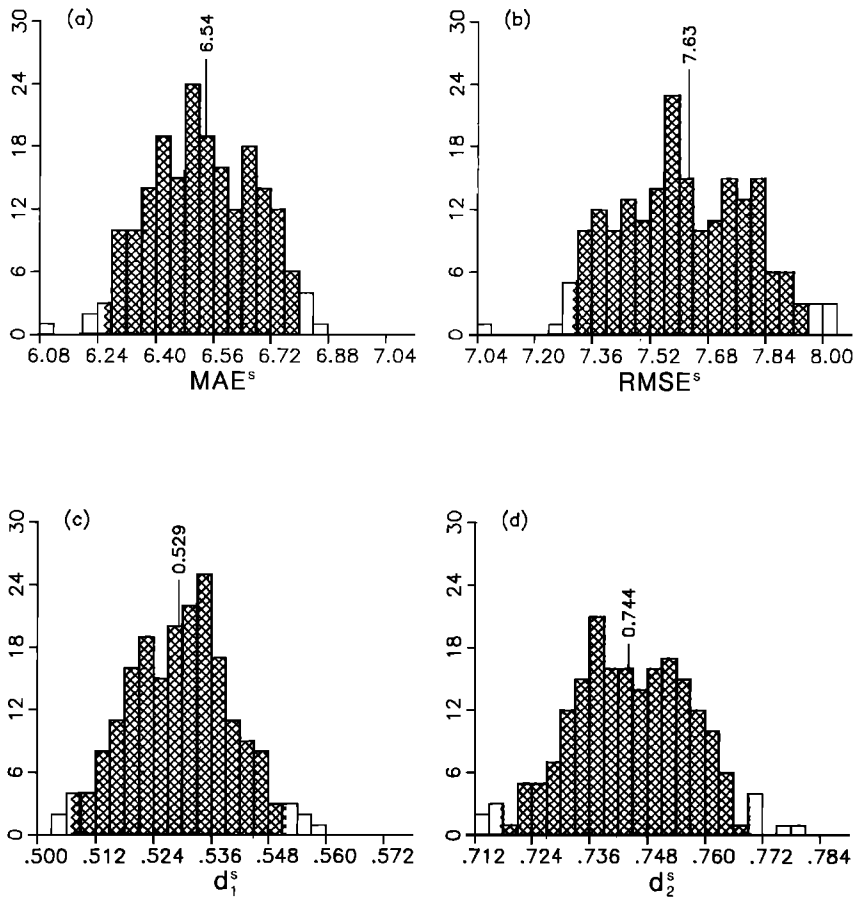

Fig. 5. Bootstrapped histograms of (*a*) the mean absolute error (MAE^s), (*b*) the root-mean-square error ($RMSE^s$), (*c*) the modified index of agreement ($d_1{}^s$) and (*d*) the index of agreement ($d_2{}^s$) associated with the simple model. The 95% bootstrap confidence interval is shaded.

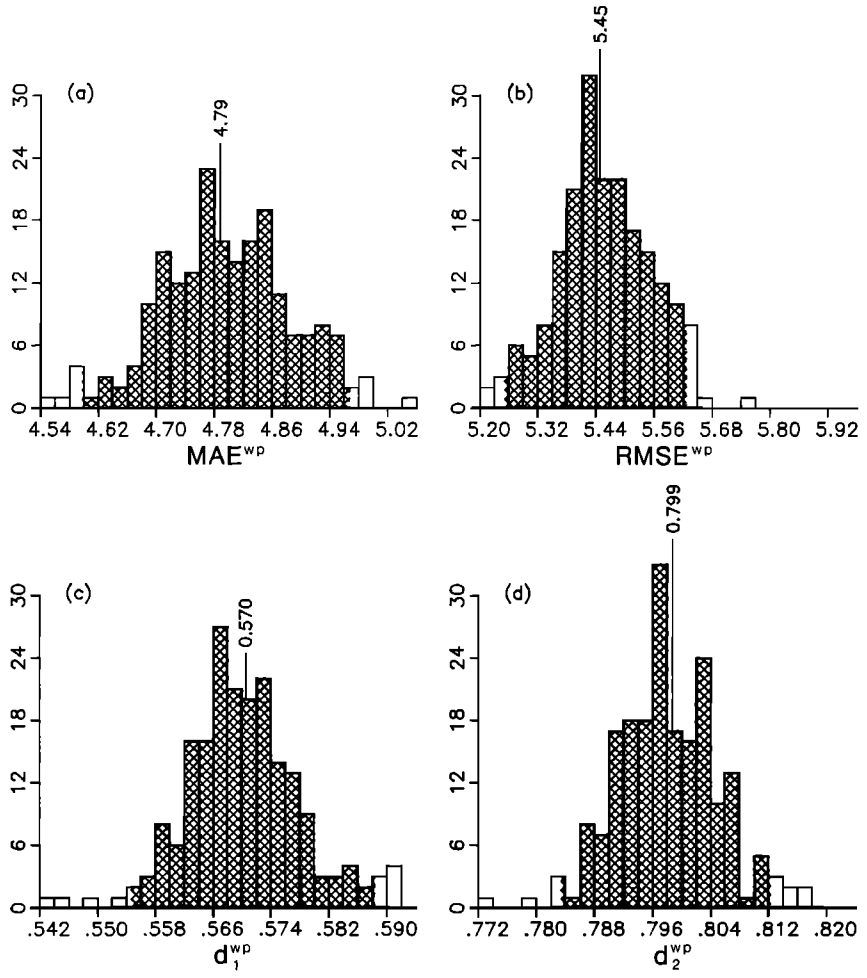

Fig. 6. Bootstrapped histograms of (*a*) the mean absolute error (MAE^{wp}), (*b*) the root-mean-square error ($RMSE^{wp}$), (*c*) the modified index of agreement ($d_1{}^{wp}$), and (*d*) the index of agreement ($d_2{}^{wp}$) associated with the W-P model. The 95% bootstrap confidence interval is shaded.

the CHS 36-point wind direction summaries in the simple model produces a crude 10° resolution in the predicted directions, and the multiplication by 2 overestimates the magnitude (Figure 3*a*).

All dimensions may be included in the graphic analysis by mapping the magnitudes associated with $\mathbf{D}^s$ and $\mathbf{D}^{wp}$ in direction-time space, where $x_j = f(\theta_j)$, $y_j = g(\text{time})$, $z_j = |\mathbf{d}_j|$ and θ_j is the direction of $\mathbf{d}_j$ (Figure 4). Consistent with the polar scatterplot (Figure 3*d*), the map of $\mathbf{D}^s$ indicates that (1) the error field produced by the simple model is highly variable and (2) the magnitude and direction of the $\mathbf{P}^s$ error field are virtually uncorrelated with each other or with time (Figure 4*a*). Within $\mathbf{D}^{wp}$, on the other hand, the error magnitudes increase slightly with the error direction, which also is apparent in the polar scatterplot (Figure 3*e*). Moreover, two time dependencies are apparent in the error field. During April, the correlation between $|\mathbf{d}_j{}^{wp}|$ and $\theta_j{}^{wp}$ is weakened, and higher $|\mathbf{d}_j{}^{wp}|$'s are coincident with both small and large $\theta_j{}^{wp}$'s. There is also an increase in $|\mathbf{d}_j{}^{wp}|$ during mid-June, which occurs when the $\theta_j{}^{wp}$'s are positive, i.e., when $\mathbf{p}_j{}^{wp}$ is erroneously rotated in a clockwise direction with respect to $\mathbf{o}_j$.

The quantitative indices recommended above concur with the more qualitative, graphic analysis (cf. Figures 3 and 4) and indicate that the W-P model is demonstrably more accurate at estimating SAB wind velocities during spring than is the simple model (Tables 1 and 2). Again, it may be useful to quantitatively evaluate the speed and direction separately, but in this illustration we treat only the velocity error field.

When evaluating vectors (velocities in this case), the interpretation of the simple summary statistics (i.e., $\bar{\mathbf{p}}$, $\bar{\mathbf{o}}$, s_p, and s_o) is not as straightforward as when the elements of $\mathbf{D}$ are scalars or directions. Since $\bar{\mathbf{p}}$ and $\bar{\mathbf{o}}$ are vector means, their magnitudes and directions cannot be equated with average speed or direction. Rather, $\bar{\mathbf{p}}$ and $\bar{\mathbf{o}}$ describe the average direction ($\angle\bar{\mathbf{p}}$ and $\angle\bar{\mathbf{o}}$) and magnitude ($|\bar{\mathbf{p}}|$ and $|\bar{\mathbf{o}}|$) associated with $\mathbf{P}$ and $\mathbf{O}$, respectively. The standard deviations, s_p and s_o, on the other hand, can be interpreted as the average distances between each $\mathbf{p}_j$ and $\bar{\mathbf{p}}$ and each $\mathbf{o}_j$ and $\bar{\mathbf{o}}$. Our summary univariate measures then indicate that the observed 1979 spring wind field at buoy 41002 has a tendency to be from the north (2.3°) at a strength of $\approx$2.0 ms^{-1}. Since $|\bar{\mathbf{o}}|$ is small relative to s_o (Tables 1 and 2), however, it can be concluded that this tendency is weak, which concurs with the impression obtained from the polar scatterplot of $\mathbf{O}$ (Figure 3*b*). The simple model statistics ($\bar{\mathbf{p}}^s$ and $s_p{}^s$) more nearly approximate $\bar{\mathbf{o}}$ and s_o than do the corresponding W-P model statistics ($\bar{\mathbf{p}}^{wp}$ and $s_p{}^{wp}$), but because the signal in $\mathbf{O}$ is weak and $s_p{}^s$ and $s_p{}^{wp}$ are so much larger than $|\bar{\mathbf{p}}^s|$ and $|\bar{\mathbf{p}}^{wp}|$, such an observation is inconclusive. A more appropriate interpretation is that the velocity distributions associated with $\mathbf{O}$, $\mathbf{P}^s$, and $\mathbf{P}^{wp}$ are relatively uniform in that no strong central tendency is evident. Note that as $|\bar{\mathbf{p}}|$ or $|\bar{\mathbf{o}}|$ approaches zero, their corresponding direction or sense

becomes meaningless (cf. the direction of $\bar{\mathbf{p}}^{wp}$ and its confidence bounds). In this example, the pairwise difference measures are much more revealing of model performance than the summary univariate indices.

Both average error measures (MAE and RMSE) suggest that the W-P model predictions are about 2 m s^{-1} more accurate than the corresponding simple model predictions. The relative error indices (d_1 and d_2) concur, but they additionally indicate that the relative magnitude of the average difference between models is small; for example, $|d_1{}^{wp} - d_1{}^{s}| \approx 0.04$. Whether an average difference between the errors of ≈ 2 m s^{-1} or 4% is important is a problem-specific scientific interpretation that we do not address; however, it can be concluded that this difference did not occur by chance. Inspection of the 95%, bootstrap confidence intervals indicates that the empirically derived frequency distributions associated with MAE^{wp}, $RMSE^{wp}$, $d_1{}^{wp}$, and $d_2{}^{wp}$ virtually do not overlap the corresponding distributions associated with MAE^{s}, $RMSE^{s}$, $d_1{}^{s}$, and $d_2{}^{s}$ (Figures 5 and 6). It is reasonable to conclude therefore that the W-P model is "significantly" more accurate than the simple model. One could also conclude that, since $d_1{}^{wp}$, $d_1{}^{s}$, $d_2{}^{wp}$, and $d_2{}^{s}$ are meaningfully larger than, say, 0.5, both models represent significant error-reducing descriptions of spring SAB wind velocities. While the null and alternative hypotheses $RMSE^{wp} = RMSE^{s}$ and $RMSE^{wp} < RMSE^{s}$, for example, could be more formally evaluated by setting α, bootstrapping the difference distribution and numerically integrating such additional analyses are unnecessary. It is clear that significance would be achieved for any typical value of α, and therefore we do not make and present such tests.

Consistent with the error patterns depicted in the polar scatterplots (Figure 3), the magnitudes and relative magnitudes of the systematic ($RMSE_s$) and unsystematic ($RMSE_u$) errors indicate that the W-P model produces a relatively large systematic error (Table 2). That is, 56% of the mean-square error is attributable to linear systematic causes, which most likely reside in the W-P model. The magnitude of the unsystematic error (≈ 3.6 m s^{-1}), on the other hand, suggests that the W-P model is quite precise. By contrast, the simple model is very imprecise ($RMSE_u = 7.18$ m s^{-1}), and since $(RMSE_u/RMSE)^2 = 0.89$, the possibility of meaningful improvement by simple refinement is remote (Table 1). Based upon the error measures, one may confidently conclude that the W-P model is both more accurate and more precise than the simple model. A respecification (with regard to spring) of the W-P model parameters should improve its accuracy.

5. SUMMARY AND CONCLUSIONS

With the development and use of simulation models becoming a major focus in the geophysical community, the need to evaluate a model's performance comprehensively and objectively or to compare competing models has become an important but underinvestigated aspect of modeling research. Not only is the model evaluation literature sparse, but the discussion is often specific to a small class of problems (e.g., air pollution or solar radiation models) and frequently the recommendations are contradictory. For the purpose of expanding the discussion of model evaluation, we have presented an array of measures and procedures that may be used to evaluate operationally (i.e., compare model-predicted events to reliable observations) a wide variety of geophysical models.

It has been recommended that a small set of complementary difference measures can represent an objective and meaningful description of a model's ability to reproduce reliable observations precisely or accurately, regardless of whether the events of interest are scalars, directions, or vectors. The core of this set of difference measures is made up of the root-mean-square error (RMSE), the systematic root-mean-square error ($RMSE_s$), the unsystematic root-mean-square error ($RMSE_u$), and the index of agreement (d_2), although the mean absolute error (MAE) and a modified index of agreement (d_1) supply related but useful information. It also has been argued that bootstrapping provides a general and reliable way to evaluate both the confidence and significance associated with each of the difference indices or, for that matter, any statistic of interest. When these difference measures are used in conjunction with the appropriate univariate statistics and data-display graphics, the operational evaluation of the performance of one or more models can be comprehensively accomplished.

While our points have been illustrated with the comparative evaluation of only two models that estimate wind velocity at a single location in the South Atlantic Bight, these methods may be extended to several other interesting problems, such as the comparison of model-predicted and observed flow fields. Model-predicted and observed wind velocity maps, for instance, could be quantitatively compared. If **P** and **O** are time series, on the other hand, time-dependent errors within the model could be detected by the calculation and interpretation of the difference measures at lags other than zero. To gain even further insight into the nature and sources of the error variable or field, it may also be useful to partition **D** into its spectral [cf. *Weisberg and Pietrafesa*, 1983] or eigenvector [cf. *Preisendorfer and Barnett*, 1983] components. Several other extensions also could be conceived, but even when the above-described evaluation is conducted in its most basic form, the ability of one or more models to reproduce nature accurately can be dependably assessed.

Acknowledgments. Part of this paper is based upon work supported by the National Science Foundation under grant ATM-8306946. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation. Professors Weisberg and Pietrafesa graciously supplied us with both observed and model-predicted data, and we gratefully acknowledge their assistance.

REFERENCES

Bevington, P. R., *Data Reduction and Error Analysis for the Physical Sciences*, McGraw-Hill, New York, 1969.

Efron, B., Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods, *Biometrika*, *68*, 589–599, 1981*a*.

Efron, B., Nonparametric standard errors and confidence intervals, *Can. J. Statist.*, *9*, 139–172, 1981*b*.

Efron, B., and G. Gong, A leisurely look at the bootstrap, the jackknife and cross-validation, *Am. Statist.*, *37*, 36–48, 1983.

Fox, D. G., Judging air quality model performance, *Bull. Am. Meteorol. Soc.*, *62*, 599–609, 1981.

Fox, D. G., Uncertainty in air quality modeling, *Bull. Am. Meteorol. Soc.*, *65*, 27–36, 1984.

Gordon, N. D., Evaluating the skill of categorical forecasts, *Mon. Weather Rev.*, *110*, 657–661, 1982.

Harr, P. A., T. L. Tsui, and L. R. Brody, Identification of systematic errors in a numerical weather forecast, *Mon. Weather Rev.*, *111*, 1219–1227, 1983.

James, L. D., and S. J. Burges, Selection, calibration and testing of hydrologic models, in *Hydrologic Modeling of Small Watersheds*, edited by C. T. Haan, H. P. Johnson, and D. L. Brakiensiek, American Society of Agricultural Engineers, St. Joseph, Mich., 1982.

Johnson, E. R., and R. L. Bras, Multivariate short-term rainfall prediction, *Water Resour. Res.*, *16*, 173–185, 1980.

MacKay, K. P., and R. D. Bornstein, Statistical evaluation of air

quality simulation models, Environmetics 81: Selected Papers, Alexandria, Va., 1981.

McCuen, R. H., and W. M. Snyder, A proposed index for comparing hydrographs, *Water Resour. Res.*, *11*, 1021–1024, 1975.

Nash, J. E., and J. V. Sutcliffe, River flow forecasting through conceptual models, 1, A discussion of principles, *J. Hydrol.*, *10*, 282–290, 1970.

Preisendorfer, R. W., and T. P. Barnett, Numerical model-reality intercomparison test using small-sample statistics, *J. Atmos. Sci.*, *40*, 1884–1896, 1983.

Rao, S. T., and J. R. Visalli, On the comparative assessment of the performance of air quality models, *J. Air Pollut. Contr. Assoc.*, *31*, 851–860, 1981.

Weisberg, R. H., and L. J. Pietrafesa, Kinematics and correlation of the surface wind field in the South Atlantic Bight, *J. Geophys. Res.*, *88*, 4593–4610, 1983.

Willmott, C. J., On the validation of models, *Phys. Geogr.*, *2*, 184–194, 1981.

Willmott, C. J., Some comments on the evaluation of model performance, *Bull. of Am. Meteorol. Soc.*, *63*, 1309–1313, 1982.

Willmott, C. J., On the evaluation of model performance in physical geography, in *Spatial Statistics and Models*, edited by G. L. Gaile and C. J. Willmott, D. Reidel, Hingham, Mass., 1984.

Won, T. K., Model validation methods, in *Handbook of Radiation Estimation Methods*, IEA Programme to Develop and Test Solar Heating and Cooling Systems, Canada, 1981.

S. G. Ackleson and J. O'Donnell, College of Marine Studies, University of Delaware, Newark, DE 19716.

R. E. Davis, J. J. Feddema, K. M. Klink, D. R. Legates, C. M. Rowe, and C. J. Willmott, Center for Climatic Research, Department of Geography, University of Delaware, Newark, DE 19716.

(Received February 25, 1985;
accepted February 25, 1985.)