Diagnostic Verification of Temperature Forecasts

ALLAN H. MURPHY, BARBARA G. BROWN* AND YIN-SHENG CHEN†

Department of Atmospheric Sciences, Oregon State University, Corvallis, Oregon

(Manuscript received 13 September 1988, in final form 16 June 1989)

ABSTRACT

A diagnostic approach to forecast verification is described and illustrated. This approach is based on a general framework for forecast verification. It is "diagnostic" in the sense that it focuses on the fundamental characteristics of the forecasts, the corresponding observations, and their relationship.

Three classes of diagnostic verification methods are identified: 1) the joint distribution of forecasts and observations and conditional and marginal distributions associated with factorizations of this joint distribution; 2) summary measures of these joint, conditional, and marginal distributions; and 3) performance measures and their decompositions. Linear regression models that can be used to describe the relationship between forecasts and observations are also presented. Graphical displays are advanced as a means of enhancing the utility of this body of diagnostic verification methodology.

A sample of National Weather Service maximum temperature forecasts (and observations) for Minneapolis, Minnesota, is analyzed to illustrate the use of this methodology. Graphical displays of the basic distributions and various summary measures are employed to obtain insights into distributional characteristics such as central tendency, variability, and asymmetry. The displays also facilitate the comparison of these characteristics among distributions—for example, between distributions involving forecasts and observations, among distributions involving different types of forecasts, and among distributions involving forecasts for different seasons or lead times. Performance measures and their decompositions are shown to provide quantitative information regarding basic dimensions of forecast quality such as bias, accuracy, calibration (or reliability), discrimination, and skill. Information regarding both distributional and performance characteristics is needed by modelers and forecasters concerned with improving forecast quality. Some implications of these diagnostic methods for verification procedures and practices are discussed.

1. Introduction

Verification of temperature forecasts has traditionally consisted of the computation of a few overall measures of performance such as the mean absolute error or mean square error (e.g., Carter and Polger 1986; Murphy and Daan 1985). Such practices may be adequate to describe the general state of the art of temperature forecasting or to assess overall trends in the quality of temperature forecasts. However, current verification procedures and practices are inadequate when the objective is either to identify the fundamental strengths and weaknesses in temperature forecasts or to provide modelers and forecasters with feedback as a basis for improving the quality of such forecasts. Moreover, the needs of users for information regarding the basic characteristics of temperature forecasts are not adequately met by overall performance measures.

A general framework for forecast verification was recently described by Murphy and Winkler (1987) (hereafter referred to as MW87). This framework is based on the joint (probability) distribution of forecasts and observations and on the conditional and marginal distributions associated with factorizations of the joint distribution. Since these distributions describe the fundamental statistical characteristics of the forecasts and observations and their relationship, they appear to represent a sound basis for a logically coherent and genuinely insightful approach to forecast verification—an approach that has been lacking heretofore. Although the outlines of such an approach—referred to here as *diagnostic verification*—were implicit in MW87, the approach itself was not described explicitly in the earlier paper. Moreover, this approach—and its associated methodology—have not as yet been applied to samples of real forecasts and observations. Clearly, such an application constitutes the ultimate test of the utility of the diagnostic approach to verification.

Diagnostic verification not only represents a sound approach to forecast verification, it also contains a useful set of verification methods. The fundamental elements in this set of methods are the joint, conditional,

* Current address and joint affiliation: Environmental and Societal Impacts Group, National Center for Atmospheric Research, Boulder, CO.

† Permanent address: Department of Meteorology, Nanjing Institute of Meteorology, Nanjing, Jiangsu, 210035, P.R.C.

Corresponding author address: Allan H. Murphy, Department of Atmospheric Sciences, Oregon State University, Corvallis, OR 97331.

and marginal distributions of the forecasts and observations. In addition, the body of diagnostic verification methodology includes summary measures of these distributions as well as overall performance measures and their decompositions. In contrast to traditional verification methods, these diagnostic methods focus on basic statistical characteristics of the forecasts and observations and their relationship. Particular attention is devoted to assessing the conditional characteristics of the observations given the forecasts and vice versa. To facilitate insights into basic characteristics of forecasting performance, diagnostic verification also makes extensive use of graphical displays in describing and summarizing various results. Information forthcoming from this approach and the associated body of methodology is more likely to be useful to modelers and forecasters in the process of identifying deficiencies in forecasts—and of obtaining insights into ways in which forecasts might be improved—than information forthcoming from traditional verification approaches and methods.

The primary purposes of the present paper are to describe a diagnostic approach to forecast verification—and its attendant methodology—and to illustrate the use of this methodology by presenting some results of a diagnostic analysis of short-range temperature forecasts. The diagnostic approach to forecast verification is outlined in section 2. This section focuses on the identification of a potentially useful body of diagnostic verification methodology and briefly contrasts these methods with current procedures. Mathematical details related to various aspects of this methodology are incorporated into several appendices. The insights provided by this approach are illustrated in section 3 by applying these diagnostic verification methods to a sample of U.S. National Weather Service (NWS) maximum temperature forecasts. Section 4 reviews the important features of these methods, summarizes the insights provided by their use, and discusses their implications for verification procedures and practices. In addition, some possible extensions of the diagnostic approach to forecast verification presented here are outlined in this section.

2. Diagnostic verification: general approach and methods

a. General approach

As indicated in MW87, the joint distribution of forecasts and observations provides a general framework for forecast verification. If the forecasts and observations are denoted by f and x, respectively, then this joint distribution can be denoted by $p(f, x)$. For a sample of data (forecasts *and* observations), $p(f, x)$ specifies the relative frequency of occurrence of particular combinations of values of f and x. With the application to temperature forecasts in mind, both f and x are assumed to be defined on the set of integer values (i.e., the forecasts are nonprobabilistic). The joint distribution $p(f, x)$ contains information about the forecasts, the observations, and the relationship between the forecasts and observations. In fact, $p(f, x)$ contains *all* of the nontime-dependent information relevant to forecast verification.

The information contained in $p(f, x)$ is more accessible when this distribution is factored into conditional and marginal distributions. Two such factorizations are possible; namely, 1) $p(f, x) = p(x|f)\,p(f)$ and 2) $p(f, x) = p(f|x)\,p(x)$. In the first factorization, $p(f, x)$ is factored into conditional distributions of the observations given the forecasts, $p(x|f)$, and the marginal distribution of the forecasts, $p(f)$. For a sample of data, the conditional distribution $p(x|f)$ specifies the relative frequency of occurrence of the various observations when a particular forecast is made, whereas the marginal distribution $p(f)$ specifies the relative frequency of use of the various possible forecasts.

The distributions $p(x|f)$ and $p(f)$ relate to two distinct characteristics of the forecasts, calibration and refinement, both of which are of interest for verification purposes. Specifically, forecasts are said to be perfectly calibrated (or completely reliable) if $E(x|f) = f$ for all f, where $E(x|f)$ is the expected (or mean) value of the conditional distribution $p(x|f)$. Thus, a temperature forecasting system is perfectly calibrated if, for each forecast value f, the mean observed temperature is equal to f. The marginal or predictive distribution of the forecasts, $p(f)$, relates to the refinement (or sharpness) of the forecasts. A temperature forecasting system that produces the same forecast on each occasion is completely unrefined. Complete refinement is difficult to define in the case of such nonprobabilistic forecasts. However, for perfectly accurate forecasts (which are also perfectly refined), $p(f)$ is necessarily identical to the marginal distribution of the observations, $p(x)$.

In the second factorization, $p(f, x)$ is factored into conditional distributions of the forecasts given the observations, $p(f|x)$, and the marginal distribution of the observations, $p(x)$. The conditional probabilities that constitute $p(f|x)$ are generally referred to as likelihoods, since they indicate the "likelihood" that a particular forecast is associated with a given observation. For a sample of data, these likelihoods are estimated by the corresponding conditional relative frequencies. Analogously, the marginal distribution $p(x)$ specifies the probability of occurrence of the respective observations. Thus, $p(x)$ consists of the sample climatological probabilities (or sample base rates).

The likelihoods, $p(f|x)$, indicate the extent to which a forecast discriminates among the various values of x. For a temperature forecast f, if the values of the likelihoods are very similar for different values of x, then the forecast is not very discriminatory—in the extreme, when $p(f|x)$ is the same for all x, the forecast is not at all discriminatory. When the likelihoods are

very different for different values of x, the temperature forecast is much more discriminatory; in particular, it is perfectly discriminatory (for each f) when $p(f|x)$ equals zero for all values of x except one. Note that the marginal distribution of the observations, $p(x)$, is the only component of either factorization that does not involve the forecast f in any way. Thus, it is a characteristic of the forecasting situation, rather than of the forecasting system or forecaster. The characteristic of concern in this case is uncertainty, which is related to the variability of the observations. Forecasting situations involving a narrow range of temperature values and/or a peaked distribution are characterized by relatively little uncertainty, and thus are relatively less difficult situations in which to forecast. In contrast, forecasting situations involving a wide range of temperature values and/or a fairly uniform distribution are characterized by relatively great uncertainty, and thus are relatively more difficult situations in which to forecast.

It is evident that the individual components in the two factorizations measure different characteristics of forecasting systems and/or forecasting situations. This fact implies that all four components will be of interest for verification purposes. For a discussion of the relationships between the factorizations—and their respective components—see MW87.

b. *Specific methods*

The framework described in section 2a provides a rational basis for a diagnostic approach to forecast verification. This approach is "diagnostic" in the sense that it is concerned primarily with identifying and describing, in a quantitative manner whenever possible, the basic characteristics of the forecasts, the observations, and their relationship. Here, a "verification method" is any mathematical or statistical measure or pictoral or graphical display that provides insight into or summarizes one or more of these basic characteristics. In assembling the verification methods to be described in this section, the application to nonprobabilistic forecasts of a continuous variable has been kept in mind.

Three specific classes of diagnostic verification methods are identified in this section: 1) the basic distributions themselves; 2) summary measures of these distributions; and 3) traditional performance measures and their decompositions. First, since the joint, conditional, and marginal distributions describe the fundamental statistical characteristics of the forecasts and/or observations, they necessarily constitute a potentially useful set of verification methods. Of course, recognition of the fact that these distributions represent important sources of information in the context of forecast verification is hardly a new concept. For example, the results of verification studies involving forecasts of discrete variables or events are often summarized in the form of unconditional or conditional contingency tables (e.g., Brier and Allen 1951). Moreover, indirect use is frequently made of (unconditional) contingency tables in computing performance measures for such forecasts (e.g., the fraction or percent correct). However, the fundamental role of these distributions in forecast verification has generally not been recognized explicitly, and they have seldom if ever been considered in the case of forecasts of continuous variables such as temperature. Nevertheless, we believe that the conditional distributions $p(x|f)$ and $p(f|x)$, since they necessarily describe the relationship between f and x, can provide especially valuable insights into forecasting performance. This class of methods will be referred to as the *basic distributions*.

Notwithstanding the fundamental information vis-à-vis the forecasts and observations—and their relationship—that is contained in the basic distributions, it is also desirable to summarize the most important features of these distributions—and thereby forecast quality—in terms of a few specific measures or parameters. The features of interest include such distributional characteristics as central tendency, variability, and asymmetry. In choosing such parameters, it seems reasonable to focus at least initially on measures directly related to the basic distributions themselves. Traditional choices for such summary measures include the mean as a measure of central tendency and the variance (or standard deviation) as a measure of variability. Moreover, the correlation coefficient represents a traditional measure of the overall association (or linear relationship) between the forecasts and observations.

In assessing the statistical characteristics of the basic distributions, it may also be desirable to consider measures that do not require assumptions regarding the shapes of such distributions (e.g., symmetry, normality). For example, it is frequently useful to determine various quantiles of the marginal and/or conditional distributions. Quantiles employed in this paper include the 0.50th quantile (median), the 0.75th and 0.25th quantiles (upper and lower quartiles, respectively), and the 0.90th and 0.10th quantiles. In situations in which the basic distributions are asymmetric or otherwise nonnormal, the median and interquartile range (the difference between the upper and lower quartiles) might be more appropriate measures of central tendency and variability, respectively, than the mean and standard deviation. For a relatively "well-behaved" variable such as temperature—for which the distributions of forecasts and observations might be reasonably symmetric and even approximately normal (jointly, conditionally, and/or marginally)—the difference between the mean and median generally would be quite small. To assess the asymmetry of distributions in this paper, we compute a measure that consists simply of the difference between the 0.90th quantile minus the median and the median minus the 0.10th quantile. This measure is zero for a symmetric distribution and (generally) non-

zero for an asymmetric distribution. We will refer to the class of (traditional and nontraditional) measures of the characteristics of the basic distributions as the *summary measures.*

In interpreting the summary measures of the conditional distributions, $p(x|f)$ and $p(f|x)$, we will make use of simple linear regression models in which the forecasts are regressed on the observations and vice versa. These regression models are described in detail in appendix A. The models provide "standards of reference," which can be used to evaluate the extent to which the observations given the forecasts and the forecasts given the observations are conditionally unbiased. As noted in appendix A, the phrase "conditionally unbiased" has the same meaning as perfectly calibrated (or completely reliable). It is demonstrated in appendix A that, from the perspective of the regression model associated with the distributions $p(x|f)$, conditionally unbiased forecasts are optimal. However, forecasts that are conditionally unbiased in this sense will necessarily (unless they are perfect) be conditionally biased from the perspective of the regression model associated with the distributions $p(f|x)$. Moreover, this latter model prescribes the amount of conditional bias that is consistent with the overall degree of association between the forecasts and observations (as measured by the correlation coefficient).

It is also desirable to identify some means of assessing the extent to which the forecasts discriminate among the observations, as characterized by the conditional distributions $p(f|x)$. Recall that a forecast is discriminatory if, for fixed f, $p(f|x)$ is different for different values of x, whereas it is not at all discriminatory if $p(f|x)$ is the same for all x (see section 2a). To evaluate qualitatively the amount of discrimination provided by a set of forecasts, we can examine and compare these conditional distributions for different values of x. If the forecasts are discriminatory, then the $p(f|x)$ should be quite well separated, whereas if the forecasts are not very discriminatory, then the $p(f|x)$ will overlap to a considerable degree. As a quantitative measure of the degree of discrimination between two observations (e.g., x_i and x_j) provided by the forecast (f), we can use the likelihood ratio (LR), where $\mathrm{LR}(f; x_i, x_j) = p(f|x_i)/p(f|x_j)$. This ratio is close to 1 in the case of very little discrimination, and it is exactly 1 in the case of no discrimination. Moreover, as discrimination increases, the value of LR will differ increasingly from 1. A measure of discrimination based on the likelihood ratio is defined in appendix B.

Calculation of certain familiar measures of performance can also be quite useful. The traditional performance measures considered here include the mean (algebraic) error (ME), the mean square error (MSE), and the skill score (SS). These measures are defined in appendix C. The ME is a measure of overall (or systematic; or unconditional) bias, the MSE is a measure of accuracy, and the SS is a measure of skill (i.e., relative accuracy). This class of methods will be referred to as the *performance measures.*

The above-mentioned performance measures are currently computed in conjunction with many verification programs. In addition, however, it is possible to decompose the MSE, as well as the SS based on the MSE, into terms that measure other characteristics of the forecasts and observations. Mathematical expressions for two such decompositions are given in appendix D. From these expressions and the interpretations of the respective terms, it is evident that the two decompositions provide additional insights into basic characteristics of forecasting performance.

Finally, an important feature of the diagnostic approach to forecast verification—and its associated body of methodology—is the use of graphical displays of the results. As the results presented in section 3 demonstrate, displays such as bivariate histograms, box plots, likelihood diagrams, and even simple x–y plots provide a means of summarizing information concerning basic aspects of forecast quality in an efficient yet transparent manner. These displays can be particularly useful in conjunction with the evaluation of the basic distributions and the associated summary measures. Specifically, such displays can greatly facilitate the interpretation and enhance the usefulness of the results of verification studies.

The methods of diagnostic verification employed in this paper are summarized in Table 1. In this table the methods are classified first by probability distribution and then by distributional or performance characteristic. The types of displays used to depict these methods are also indicated.

c. *Computation of diagnostic measures*

Computation of the diagnostic verification measures described in section 2b—and to be employed and illustrated in section 3—is relatively easy. In fact, most of these measures can be computed by using common statistical software packages on a mainframe computer, minicomputer, or microcomputer. The graphical procedures used to prepare insightful displays of verification data rely for the most part on simple x–y plotting routines, which are included in most statistical or graphical software packages. More complex graphical displays such as bivariate histograms and box plots can also be prepared by selected software packages of this type. Thus, most of the computations illustrated here can be accomplished without requiring the creation of new software. Of course, in practice it may be necessary to modify existing software or to formulate new subroutines that will, for example, automatically perform diagnostic verification computations on a regular basis.

3. Application to NWS temperature forecasts

Some results of an application of the diagnostic approach to forecast verification—and its associated

TABLE 1. Methods of diagnostic verification, classified by probability distribution and distributional or performance characteristic. Summary or performance measures and displays for each characteristic are also indicated.

Characteristic	Measure	Display
(1) Joint Distribution: $p(f, x)$		
correspondence	joint distribution	bivariate histogram
association	correlation coefficient	line diagram
accuracy	(a) mean square error (MSE) (b) root mean square error (RMSE)	line diagram line diagram
skill	skill score (SS)	line diagram
(2) Marginal Distributions: $p(f)$ and $p(x)$		
central tendency	(a) mean (b) median	line diagram box plot
variability	(a) variance (b) interquartile range	line diagram box plot
asymmetry	quantiles	box plot
extremes	extreme quantiles	box plot
bias	(a) mean error (ME) (b) component of SS	line diagram line diagram
(3) Conditional Distributions: $p(x/f)$		
conditional central tendency	conditional median	conditional quantile diagram
conditional variability	conditional interquartile range	conditional quantile diagram
conditional asymmetry	conditional quantiles	conditional quantile diagram
conditional extremes	conditional extreme quantiles	conditional quantile diagram
conditional bias (reliability)	(a) conditional median (b) component of SS	conditional quantile diagram line diagram
(4) Conditional Distributions: $p(f/x)$		
conditional central tendency	conditional median	conditional quantile diagram
conditional variability	conditional interquartile range	conditional quantile diagram
conditional asymmetry	conditional quantile	conditional quantile diagram
conditional extremes	conditional extreme quantiles	conditional quantile diagram
discrimination	(a) conditional distributions (b) discrimination (DIS)	conditional distributions diagram line diagram

methodology—are presented in this section. These results are based on an analysis of a sample of temperature forecasts for Minneapolis, Minnesota. However, the purpose of describing this particular application is not to investigate the characteristics of temperature forecasts for a specific location. Instead, the objectives here are (a) to present some methods that can be used to identify and compare various basic dimensions of forecast quality within the overall context of diagnostic verification and (b) to demonstrate by example the kinds of information about forecasts, observations, and the relationship between forecasts and observations that can be obtained by adopting this approach to forecast verification. Thus, the results presented here do not constitute a comprehensive verification of these forecasts. Rather, they represent the kinds of results—and modes of presentation of these results—that might be employed in a detailed diagnostic verification of forecasts of this type.

a. Data

The data used here to illustrate the diagnostic approach to forecast verification are objective and subjective temperature forecasts, and the corresponding observed temperatures for Minneapolis, Minnesota, for

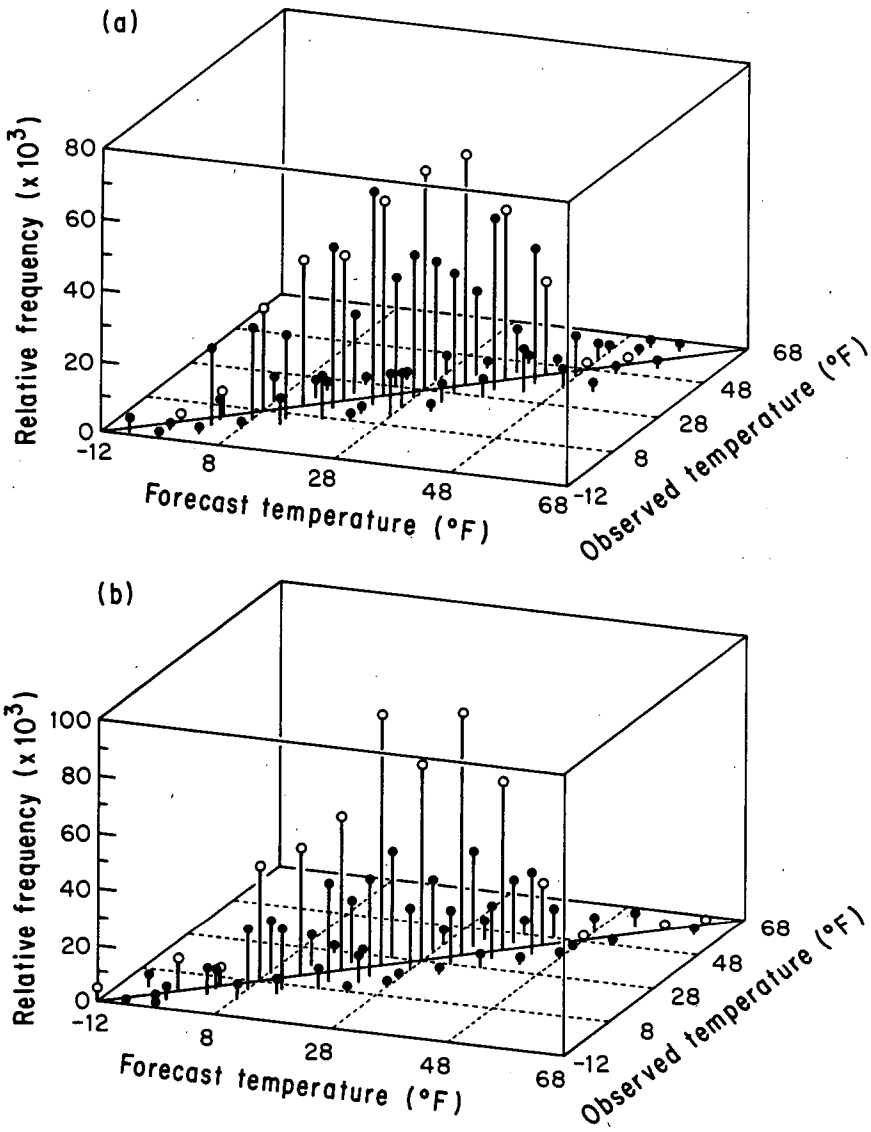

FIG. 1. Bivariate histograms of joint distributions of forecasts and observations [$p(f, x)$]. (a) 24-h objective forecasts in the winter season. (b) 24-h subjective forecasts in the winter season. (c) 24-h objective forecasts in the summer season. (d) 24-h subjective forecasts in the summer season. Values of $p(f, x)$ for which $f = x$ are indicated by open circles to facilitate identification. See text for additional details.

the period April 1980–March 1986. These data were obtained from the NWS Public Weather Verification Data Archive (see Carter and Polger 1986). The objective forecasts were produced by the model output statistics system (e.g., Glahn 1985), whereas the subjective forecasts were formulated by NWS forecasters. It should be noted that the objective forecasts generally were available to the forecasters prior to the formulation of their subjective forecasts. Both types of forecasts were prepared twice a day (cycle times of 0000 and 1200 UTC) for four lead times (approximately 24, 36, 48, and 60 h in advance) for maximum and minimum temperatures. Attention is focused in this paper primarily on the maximum temperature forecasts associated with the 0000 UTC cycle time for the 3-month winter (December, January, and February) and summer (June, July, and August) seasons. To conserve space, results are presented in many cases only for the winter season and for the 24-h lead time.

b. *Basic distributions and summary measures*

1) *Joint distribution, $p(f, x)$.* Examination of the joint distribution $p(f, x)$ provides overall insight into

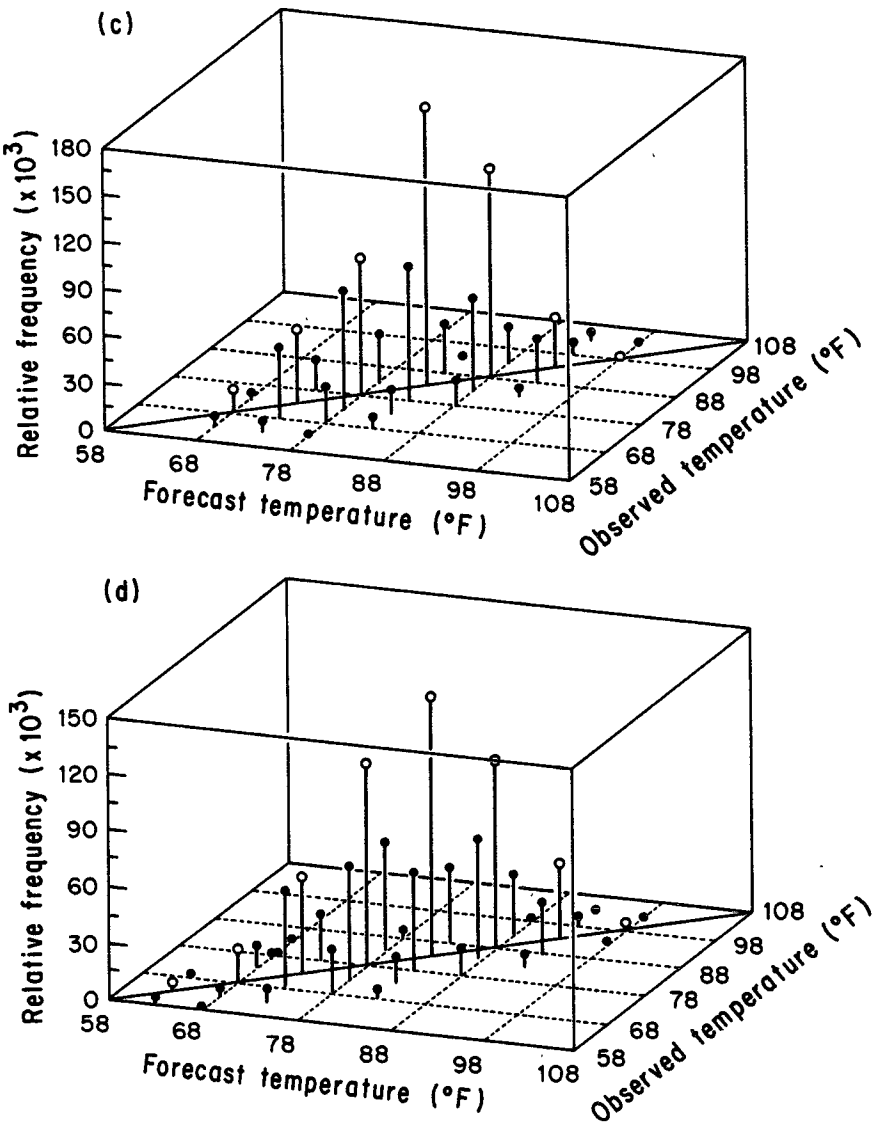

FIG. 1. (*Continued*)

the relationship between the forecasts and observations. This distribution can be depicted in several different ways: for example, graphically in the form of a bivariate histogram or numerically in the form of a contingency table. Fig. 1 contains bivariate histograms for the Minneapolis 24-h maximum temperature forecasts for the winter and summer seasons. In formulating the joint distributions shown in these diagrams, the raw data have been assigned to 5°F categories (i.e., 3°–7°F, 8°–12°F, 13°–17°F, etc.). The histograms are depicted with the forecast value (f) on the x-axis, the observed value (x) on the y-axis, and the joint relative frequency [$p(f, x)$] on the z-axis. The diagonal 45° line in the x–y plane represents one-to-one correspondence between the forecast and observed categories, namely, forecasts with errors ≤4°F. Thus, off-diagonal "elements" of $p(f, x)$ correspond to forecasts that are associated with even larger errors.

Bivariate histograms can provide qualitative (and even quasi-quantitative) information regarding various characteristics of the joint and marginal distributions, such as central tendency, variability, and symmetry, as well as information concerning differences in these characteristics among types of forecasts (e.g., objective/subjective), seasons, and lead times. For example, the joint distributions shown in Fig. 1 appear to be quite symmetric about the 45° line. That is, the largest joint relative frequencies are associated with categories for which $f = x$; otherwise, the observations are approximately equally likely to be greater or less than the forecasts. Comparison of the diagrams for winter and summer suggests that the primary difference between

the distributions for the two seasons relates to the number of temperature categories with nonzero relative frequencies. This number is much smaller in summer than in winter (i.e., the range of maximum temperatures at Minneapolis is evidently much greater in winter than in summer). Additional information could be obtained from the bivariate histograms by careful examination and comparison of specific "sectors" of the respective distributions.

The correlation coefficient is a traditional measure of the degree of (linear) association between two quantities, and it represents a quantitative summary measure of the joint distribution. It can be used to compare the quality of different types of forecasts, as well as to compare forecasts for different seasons or lead times. (However, it should be kept in mind that this measure ignores any conditional or unconditional biases in the forecasts—see section 2b and appendix D.) In the case of the Minneapolis maximum temperature forecasts, a line diagram depicting the correlations between the forecasts and observations for all lead times (Fig. 2) reveals that the degree of association between the forecasts and observations decreases with increasing lead time. Differences between the correlation coefficients

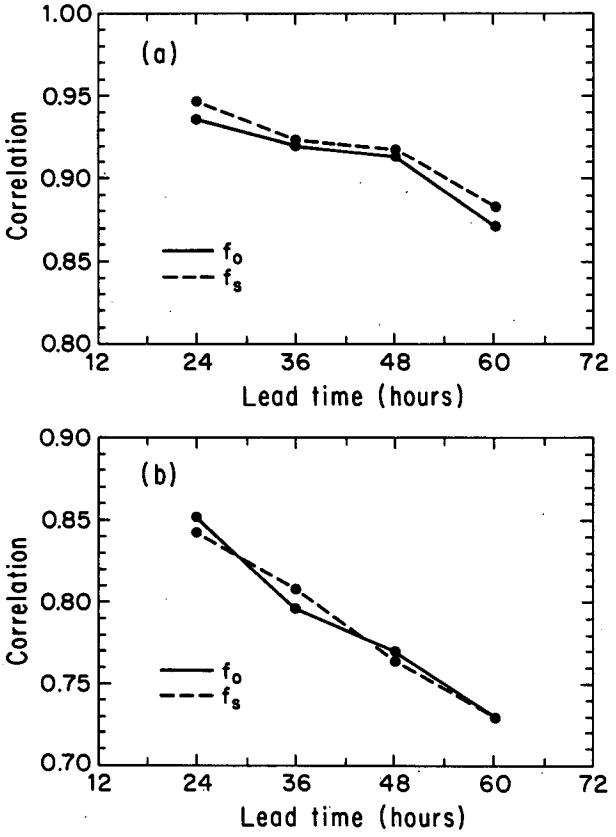

FIG. 2. Correlation coefficients (r_{fx}) between forecasts and observations as a function of lead time for (a) the winter season and (b) the summer season (f_o: objective forecasts, f_s: subjective forecasts).

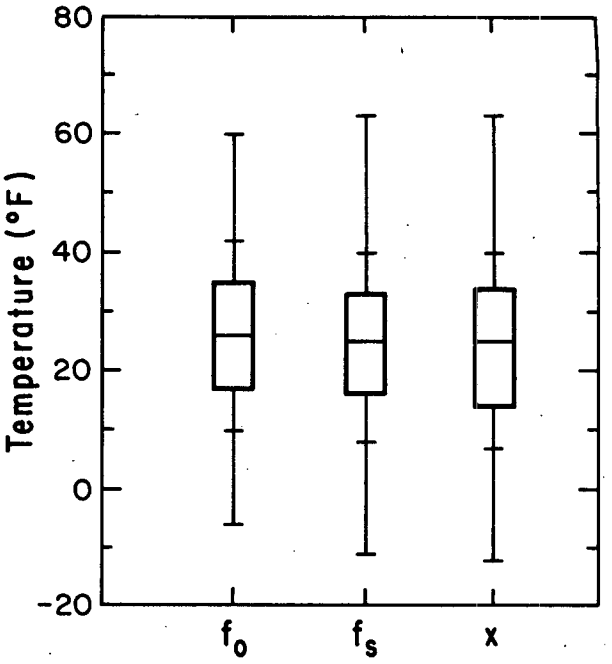

FIG. 3. Box plots of marginal distributions of forecasts and observations [$p(f)$ and $p(x)$, respectively] for the 24-h lead time in the winter season (f_o: objective forecasts, f_s: subjective forecasts).

for the two types of forecasts are generally small, with the correlations for the subjective forecasts in some cases exceeding (by a small margin) the correlations for the objective forecasts. Comparison of the correlations for the two seasons indicates that the forecasts and observations are more strongly associated in winter than in summer.

2) *Marginal distributions, $p(f)$ and $p(x)$.* Box plots provide a convenient means of summarizing the marginal distributions of the forecasts and observations (see Fig. 3). These plots describe distributional characteristics such as central tendency, variability, and symmetry. Moreover, box plots facilitate the comparison of these characteristics among distributions. Traditional practices in forecast verification seldom extend beyond the comparison of the means of the respective distributions.

The horizontal line inside each box designates the median (i.e., the 0.50th quantile), an alternative measure of central tendency. Upper and lower quartiles (i.e., the 0.75th and 0.25th quantiles, respectively) of the distribution are represented by the top and bottom of the box, and the difference between them—the length of the box—is the interquartile range (IQ), an alternative measure of variability. Maximum and minimum values, displayed at the ends of the "whiskers," as well as the 0.90th and 0.10th quantile values, are estimates of the extremes of the distribution. The difference between the 0.90th and 0.10th quantiles is also a measure of variability. In addition, information

concerning the degree of asymmetry of such a distribution is provided by the shape of the box plot.

For the Minneapolis temperature data, the box plots for the 24-h forecasts in the winter season (Fig. 3) indicate that the distributions are quite symmetric. In particular, the upper and lower whiskers are approximately equal in length and the two halves of the boxes are approximately the same size. Small differences between the respective medians suggest that the forecasts may be slightly biased. In addition, the observed temperatures exhibit somewhat greater variability than that exhibited by the objective and subjective temperature forecasts, as indicated by the lengths of the boxes (i.e., the IQs) and the differences between the 0.90th and 0.10th quantile values.

Marginal distributions can also be described in terms of traditional summary measures, such as the mean and variance (or standard deviation). Since the forecast and observed temperature distributions appear to be fairly symmetric, the use of such measures is not unreasonable. In the cases of other weather variables (e.g., precipitation, wind speed) for which the distributions are unlikely to be symmetric, more "robust" summary measures such as the median and IQ would be more appropriate (such measures are robust if they are not unduly influenced by a few outlying values).

These traditional statistics can also be depicted in graphical form, as in Figs. 4 and 5, to facilitate comparisons among lead times and between the objective and subjective forecasts. Relatively large differences between the mean values of the forecasts and observations are evident in some cases. For example, the means of the objective forecasts for Minneapolis in winter (Fig. 4) are about 2°F higher than the means of the observed temperatures at each lead time, which suggests that the objective forecasts are appreciably biased overall. On the other hand, the means of the corresponding subjective forecasts are quite similar to

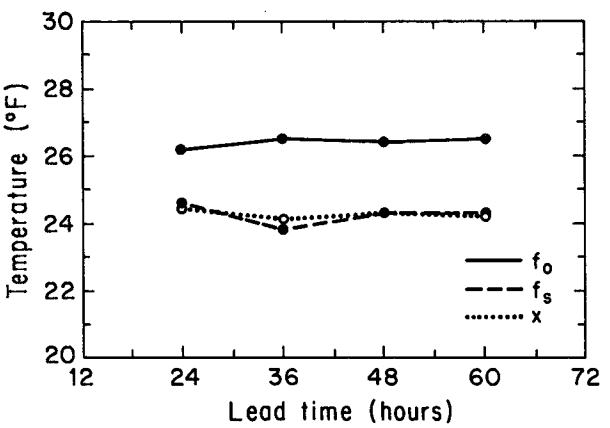

FIG. 4. Means of the forecasts and observations ($\langle f \rangle$ and $\langle x \rangle$, respectively) as a function of lead time for the winter season (f_o: objective forecasts, f_s: subjective forecasts).

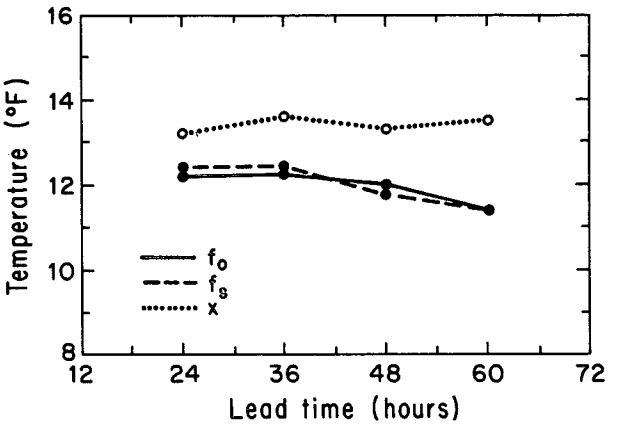

FIG. 5. Standard deviations of the forecasts and observations (s_f and s_x, respectively) as a function of lead time for the winter season (f_o: objective forecasts, f_s: subjective forecasts).

the mean observed temperatures, indicating that these forecasts are relatively unbiased. Comparison of the standard deviations of forecast and observed temperatures (Fig. 5) suggests that in winter the standard deviations of the two types of forecasts are quite similar. However, the observed temperatures exhibit considerably more variability than the forecast temperatures.

3) *Conditional distributions,* $[p(x|f)]$. The conditional distributions provide various kinds of information regarding the relationship between the forecasts and observations, including information concerning several dimensions of forecast quality. This information is generally not considered in traditional forecast verification studies. Quantiles—in particular, medians—of the conditional distributions provide information about conditional bias (or calibration). These quantiles also describe the way in which the variability in the observations changes as a function of the forecast. Specifically, the variability of the conditional distributions can be measured quantitatively using conditional values of the interquartile range. Moreover, it should be noted that these conditional interquartile range values are inversely related to (conditional) forecast accuracy. Finally, the symmetry of the distributions around the conditional medians can be measured in the manner defined in section 2b. Diagrams displaying these characteristics of the conditional distributions of the observed temperatures given the 24-h temperature forecasts for the winter season at Minneapolis are presented in Fig. 6. Results for other lead times and for the summer season are qualitatively similar to those for the 24-h lead time in the winter season.

The conditional quantile plots in Figs. 6a and 6b display various quantiles of the conditional distributions of observed maximum temperature given forecast maximum temperature. These diagrams include "running" values of the 0.10th, 0.25th, 0.50th (median), 0.75th, and 0.90th conditional quantiles. The respective

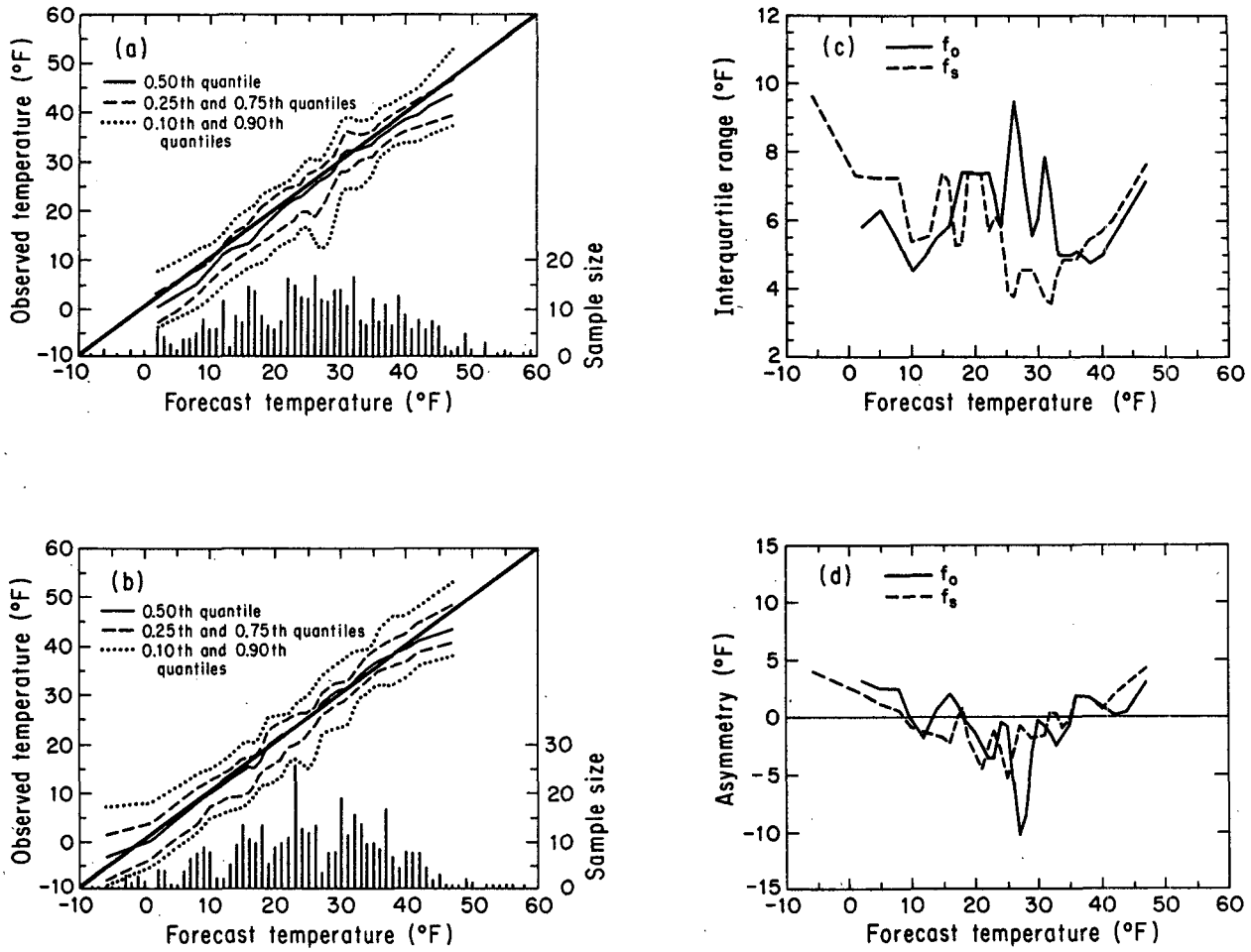

FIG. 6. Quantiles of the conditional distributions of observations given forecasts [$p(x|f)$] and histogram of the marginal distribution of forecasts [$p(f)$] for the winter season: (a) 24-h objective forecasts, (b) 24-h subjective forecasts, (c) interquartile ranges of $p(x|f)$ for 24-h lead time (f_o: objective forecasts, f_s: subjective forecasts), and (d) asymmetry values for $p(x|f)$ for 24-h lead time (f_o: objective forecasts, f_s: subjective forecasts). See text for additional details.

sets of quantile values have been smoothed using a three-point smoothing algorithm called hanning (Tukey 1977). The 45° line in the conditional quantile plots (the line of one-to-one correspondence between forecasts and observations) is included for purposes of comparison with the median values. For these temperature forecasts, it is reasonable to assume that deviations of the conditional medians from the 45° line indicate that the forecasts are conditionally biased. Histograms showing the frequency with which each forecast was used [i.e., $p(f)$] are presented along the x-axis. These frequencies are important for evaluating the credibility of the conditional quantiles. That is, the conditional quantile estimates are generally less credible for small subsamples (i.e., in the tails or extremes of the distribution of the conditioning variable) than they are for large subsamples (i.e., near the center of the conditioning distribution).

The diagrams in Fig. 6 provide a variety of information about the conditional distributions, $p(x|f)$, and they permit comparisons between the objective and subjective forecasts. For example, the objective forecasts (Fig. 6a) exhibit a tendency toward overforecasting, as indicated by the location of the conditional median line somewhat below the 45° line. Specifically, in the case of an objective forecast of 10°F in winter, the median observed temperature is 8°F. On the other hand, the subjective forecasts appear to be relatively unbiased over the entire range of forecast values (Fig. 6b).

The marginal frequency distributions of the forecasts presented in Figs. 6a and 6b provide information about the central tendency, variability, etc., of the distributions of forecasts. In addition, they indicate those ranges of forecast temperatures for which variations in the conditional quantiles may simply be an artifact of small subsample sizes. For the Minneapolis data, the marginal distributions in some cases (e.g., see Fig. 6b) ap-

pear to exhibit more than one mode (or peak). Otherwise, these distributions generally are quite symmetric, as was suggested by the box plots in Fig. 3. As expected, the frequencies associated with high and low forecast values are generally smaller than those associated with forecasts in the middle of the range of values.

The conditional IQ diagram (Fig. 6c) indicates how the variability in the observations changes as a function of the numerical value of the forecast. For example, in winter for Minneapolis, the conditional IQ values for the subjective forecasts decrease from about 10°F for very low forecast temperatures to less than 4°F for forecasts around 30°F, and then they increase again for forecasts of higher temperatures. Thus, subjective forecasts of maximum temperature in the vicinity of 30°F are associated with less variability than are forecasts of higher or lower maximum temperatures. In addition, differences in these values between the objective and subjective forecasts can be evaluated using this diagram. For these forecasts such differences appear to be relatively small.

Patterns associated with the conditional asymmetry statistic can indicate how the shapes of the conditional distributions change as a function of the forecast. For example, the asymmetry values in winter (Fig. 6d) are negative for forecasts between about 20° and 30°F, with the values being more strongly negative for the objective forecasts than for the subjective forecasts. This result suggests that observations associated with forecasts in this range tend to have somewhat negatively skewed distributions, a fact that is supported by examination of the conditional quantiles (Figs. 6a and 6b); i.e., most observations associated with forecasts in this range represent relatively high temperatures, although occasionally much lower temperatures are observed.

4) *Conditional distributions, $p(f|x)$.* The likelihood-base rate factorization leads to another set of conditional distributions. As in the case of the distributions $p(x|f)$, the conditional distributions based on the likelihood-base rate factorization can be described using various quantiles, the interquartile range, and the measure of asymmetry defined previously. In addition to these statistics, it is of interest in the case of $p(f|x)$ to consider the relative amount of discrimination provided by the forecasts, as described in section 2b. Information concerning all of these characteristics of the conditional distributions $p(f|x)$ for the 24-h forecasts of maximum temperature for Minneapolis in the winter season is presented in Figs. 7–9. Equality of forecasts and observations in the conditional quantile diagrams is once again represented by the 45° line. Moreover, these diagrams contain a line (the heavy dashed line) defined by the linear regression of f on x. In the case of $p(f|x)$, this line is more appropriate than the 45° line as a standard of reference for comparison with the conditional medians, for reasons that were discussed in section 2b (see also appendix A).

In the conditional quantile diagrams (Figs. 7a and 7b) the orientation of the conditional median line relative to the regression line indicates the degree to which the regression model represents the relationship between the observations and the conditional medians of the forecasts. In this case, the running conditional medians appear to be quite well approximated by the regression lines. It should not be overlooked, however, that this correspondence necessarily implies that the conditional mean temperature forecast has a tendency to be larger (smaller) than the observed temperature for relatively low (high) observed temperatures. Thus, a conditional bias exists in the sense that $E(f|x) \neq x$ for all x.

Differences in the variability and asymmetry of the forecasts given particular observations can be compared using diagrams such as those presented in Figs. 7c and 7d. Little systematic variation in conditional IQ is apparent in these data for Minneapolis. The asymmetry diagrams suggest that the conditional distributions of forecast temperatures given observed temperatures are relatively symmetric for most values of x.

To investigate the ability of the forecasts to discriminate among the observations, we have examined the conditional distributions $p(f|x)$ for different values of the observed temperature. Examples of these distributions for the 24-h objective and subjective forecasts in the winter season are presented in Figs. 8a and 8b, respectively. In preparing these figures we have chosen three representative values of x: the lower quartile, the median, and the upper quartile of the marginal distribution $p(x)$ (for the Minneapolis data, these observed temperatures are 14°, 25°, and 34°F, respectively). Moreover, to reduce the sampling variability of the results, all forecasts and observations associated with 5°F categories of observed temperature centered on the quantiles have been considered in defining the conditional distributions. In addition, we have smoothed these distributions using hanning.

The three conditional distributions (corresponding to the three 5°F categories of x) depicted in Fig. 8 are indicative of a substantial degree of discrimination, as reflected by the relatively modest amount of overlap among the distributions. In particular, the values of $p(f|x)$, for fixed f, are quite different for different values of x for both types of forecasts. Careful examination of these (and other similar) displays suggests that the subjective forecasts exhibit a somewhat greater degree of discrimination than the objective forecasts.

This conjecture regarding the relative amount of discrimination provided by the objective and subjective forecasts can be investigated using the quantitative measures of discrimination defined in appendix B. Values of the discrimination measure DIS(f) [see (B2)] are depicted in Fig. 9 as a function of f. These values have also been smoothed using hanning. Com-

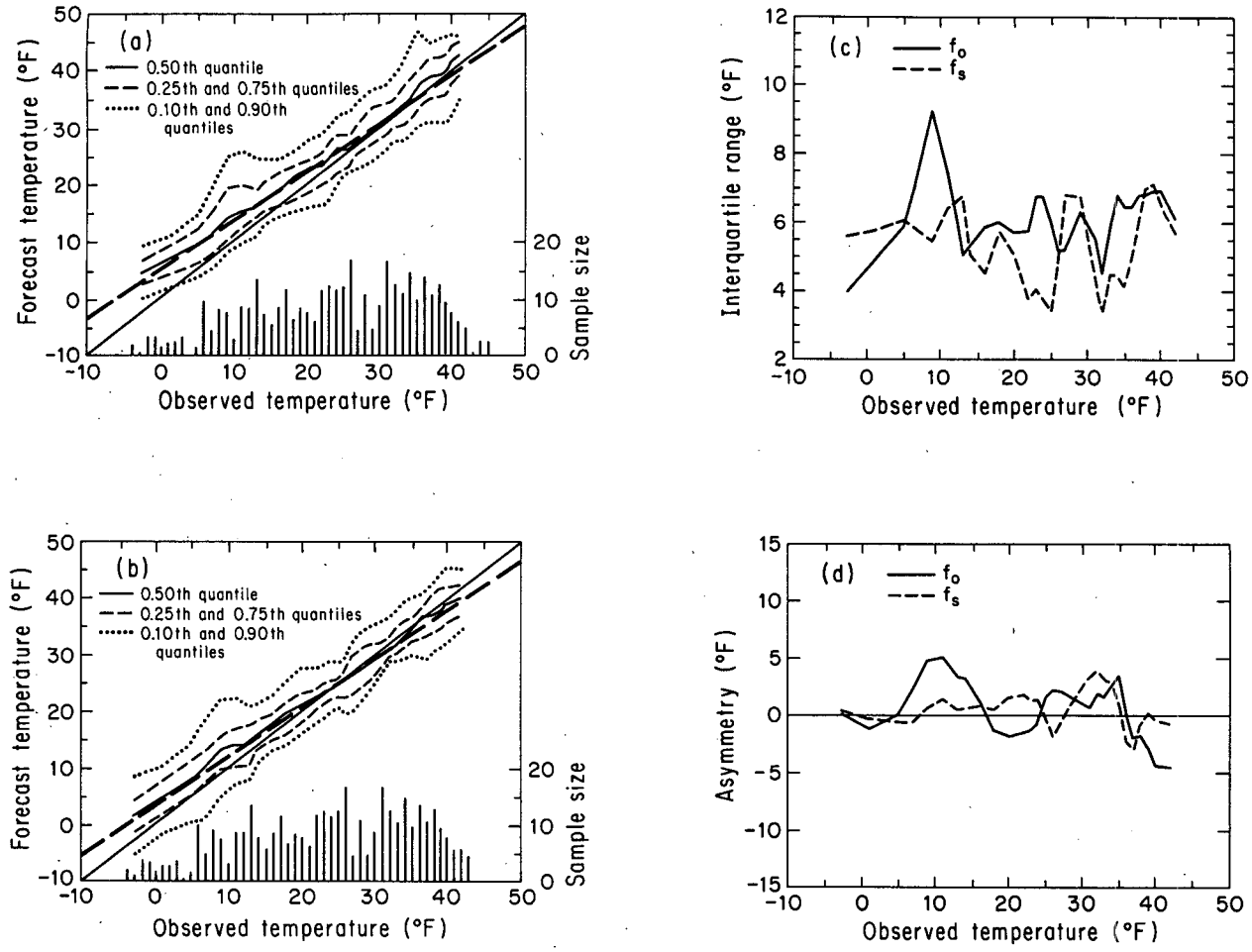

FIG. 7. Quantiles of the conditional distributions of forecasts given observations $[p(f|x)]$ and histogram of the marginal distribution of observations $[p(x)]$ for the winter season: (a) 24-h objective forecasts and (b) 24-h subjective forecasts. Dashed straight line in panels (a) and (b) represents linear regression of f on x. (c) Interquartile ranges of $p(f|x)$ for 24-h lead time (f_o: objective forecasts, f_s: subjective forecasts). (d) Asymmetry values for $p(f|x)$ for 24-h lead time (f_o: objective forecasts, f_s: subjective forecasts). See text for additional details.

parison of the curves for the objective and subjective forecasts reveals that the latter do indeed exhibit greater discrimination than the former for a majority of the forecast values. Moreover, the overall quantitative measure of discrimination [DIS in (B3)] for the 24-h lead time yields values of 1.446 for the objective forecasts and 1.479 for the subjective forecasts. Comparison of these values with the values of DIS for the 48-h lead time—1.404 for the objective forecasts and 1.401 for the subjective forecasts—reveals that (as expected) discrimination decreases as lead time increases.

c. *Performance measures*

As indicated in section 2b, traditional performance measures represent an important class of diagnostic verification methods. Moreover, these measures essentially are functions of summary statistics that describe the joint, conditional, and marginal distributions (see appendix C). In addition, the MSE and the SS can be decomposed into terms that are based on the means and variances of the forecasts and observations and on the correlation between the forecasts and observations (see appendix D).

Performance measures can be displayed in simple line diagrams, as in Fig. 10 for the Minneapolis winter maximum temperature data. These diagrams facilitate comparisons among lead times and between the objective and subjective forecasts. For example, the subjective forecasts are evidently somewhat more accurate than the objective forecasts, as indicated by the differences in RMSE values for the two types of forecasts. Moreover, accuracy decreases with increasing lead time. The ME values depict the bias in the forecasts, either in terms of overforecasting (positive bias) or in terms of underforecasting (negative bias). In the case

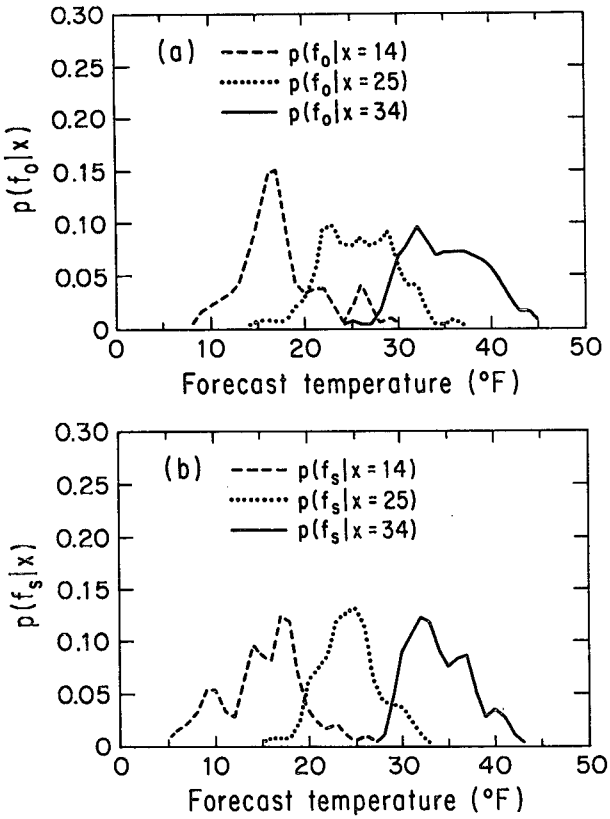

FIG. 8. Conditional distributions of forecasts given observations $[p(f|x)]$ for lower quartile (x = 14°F), median (x = 25°F), and upper quartile (x = 34°F) of $p(x)$ for the 24-h lead time in the winter season: (a) objective forecasts, and (b) subjective forecasts. See text for additional details.

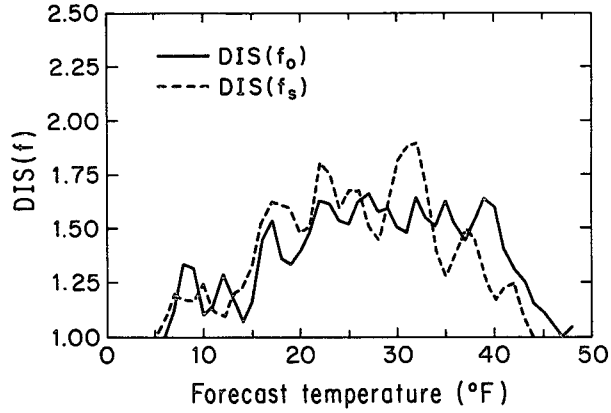

FIG. 9. Discrimination score DIS(f) as a function of f for the objective forecasts (f_o) and subjective forecasts (f_s) for the 24-h lead time in the winter season.

of the forecasts for Minneapolis, the ME values in winter suggest that the objective forecasts are appreciably biased overall, with an ME value of about 2°F for all lead times.

The decompositions of the MSE and the SS can identify the relative contributions of various terms to the overall accuracy and skill of the forecasts. These terms can also be compared among lead times and between the objective and subjective forecasts. The values of the terms in these decompositions for the Minneapolis forecasts are shown in Tables 2 (MSE) and 3 (SS). In this case, it is apparent that the terms increase or decrease as a function of lead time in a manner that generally would be expected. For example, the factor $2s_f s_x r_{fx}$ in Table 2—an indicator of the covariability or degree of linear association between the forecasts and observations—decreases as lead time increases. Bias also makes a larger contribution to the MSE for the objective forecasts than to the MSE for the subjective forecasts, as indicated by the respective magnitudes of the $(\langle f \rangle - \langle x \rangle)^2$ term in Table 2. It is of interest to note that the second term in the decomposition of the SS is essentially a measure of the appropriateness of the regression model (described in section 2b) as a description of the conditional distribution of observations given forecasts. In the case of these forecasts, the regression model apparently provides a good fit to the data because this term is quite small for both the objective and subjective forecasts for all lead times (Table 3).

4. Summary, discussion, and conclusion

In this paper we have described a diagnostic approach to forecast verification. This approach is based on a general framework for forecast verification and is designed to provide detailed insight into the basic characteristics of the forecasts, the observations, and their relationship. A body of diagnostic verification methodology has been presented and the utility of these methods has been illustrated by an application involving a sample of NWS temperature forecasts.

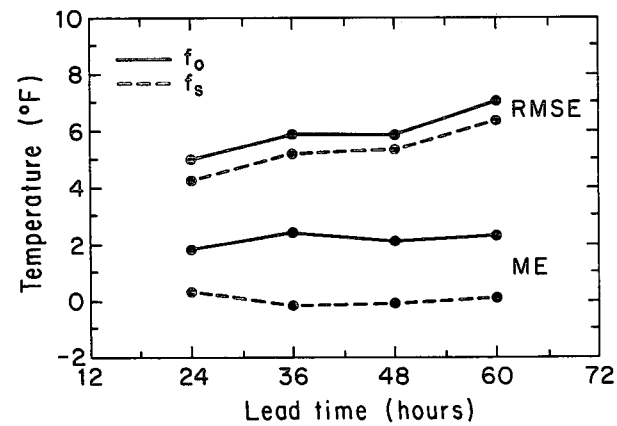

FIG. 10. Performance measures (ME and RMSE) as a function of lead time for the winter season (f_o: objective forecasts, f_s: subjective forecasts).

TABLE 2. MSE and terms in its decomposition for the winter season (f_o: objective forecasts, f_s: subjective forecasts). See text for additional details.

Lead time (h)	Type of forecast	Sample size (n)	MSE*	$(\langle f\rangle - \langle x\rangle)^2$	s_f^2	s_x^2	$2s_f s_x r_{fx}$
24	f_o	417	24.9	3.2	148.6	174.9	302.0
	f_s		18.0	0.0	154.0		310.8
36	f_o	405	34.4	5.8	149.3	184.1	304.8
	f_s		26.9	0.1	154.4		311.6
48	f_o	416	33.9	4.4	143.8	177.8	292.0
	f_s		28.4	0.0	137.9		287.4
60	f_o	397	49.6	5.3	129.8	182.7	268.2
	f_s		40.5	0.0	129.6		271.8

* $\mathrm{MSE} = (\langle f\rangle - \langle x\rangle)^2 + s_f^2 + s_x^2 - 2s_f s_x r_{fx}$.

The joint distribution of forecasts and observations—and its factorizations into conditional and marginal distributions—represent the basic elements of the general framework, and these distributions constitute the first (and fundamental) class of diagnostic verification methods. Summary measures of these distributions represent the second class of verification methods, and they include both "standard" measures of central tendency and variability (e.g., the mean and standard deviation) as well as robust measures of these and other characteristics (e.g., the median and interquartile range) based on quantiles of the distributions. Performance measures—such as the ME, MSE (or RMSE), and SS—constitute the third class of diagnostic verification methods. This latter class also includes decompositions of measures such as the MSE and SS that provide quantitative information concerning specific characteristics of the forecasts and/or observations. Finally, in the interpretation of the summary and performance measures, we found it useful to appeal to simple linear regression models in which the observations are regressed on the forecasts and vice versa.

With regard to the insights provided by the diagnostic verification methods, the joint and marginal distributions contribute both overall insights into the individual and joint "behavior" of the forecasts and observations, as well as detailed insights into basic characteristics of the forecasts, observations, and their relationship. Moreover, the summary measures yield *quantitative* information concerning these fundamental characteristics. For example, the correlations presented in Fig. 2 illustrate the decrease in the level of association between forecasts and observations with increases in lead time. Such insights and information generally cannot be obtained from traditional verification methods.

The conditional distributions play a special role in diagnostic verification because they describe the relationship between the forecasts and observations. In particular, these distributions provide detailed insights into both conditional behavior and performance, through characteristics such as conditional bias and conditional variability. For example, the conditional IQ values presented in Fig. 6 suggest that the distributions of observations associated with the subjective forecasts are least variable for maximum temperature

TABLE 3. SS and terms in its decomposition for the winter season (f_o: objective forecasts, f_s: subjective forecasts). See text for additional details.

Lead time (h)	Type of forecast	Sample size (n)	SS*	r_{fx}^2	$[r_{fx} - (s_f/s_x)]^2$	$[(\langle f\rangle - \langle x\rangle)/s_x]^2$
24	f_o	417	0.858	0.876	0.000	0.018
	f_s		0.897	0.897	0.000	0.000
36	f_o	405	0.813	0.846	0.001	0.031
	f_s		0.854	0.854	0.000	0.000
48	f_o	416	0.809	0.834	0.000	0.025
	f_s		0.840	0.843	0.001	0.000
60	f_o	397	0.728	0.759	0.001	0.029
	f_s		0.778	0.780	0.002	0.000

* $\mathrm{SS} = r_{fx}^2 - [r_{fx} - (s_f/s_x)]^2 - [(\langle f\rangle - \langle x\rangle)/s_x]^2$.

forecasts around 30°F. Once again, summary measures of the distributions yield quantitative information regarding these (conditional) characteristics.

Performance measures are used to obtain quantitative information concerning overall dimensions of forecast quality such as bias, accuracy, and skill. In addition, these measures can (in some cases) be decomposed into measures of other characteristics of performance, such as conditional bias (or calibration). The terms in these decompositions are functions of summary measures of the joint and marginal distributions of forecasts and observations.

Diagnostic verification, as described and illustrated in this paper, focuses on the fundamental aspects of forecast quality. In particular, it can identify the *situations*—defined in terms of individual or joint values of the forecasts and observations—in which forecasting performance may be especially weak or strong. Identification of such situations is obviously an essential first step in the process of improving forecast quality. Moreover, this knowledge can provide modelers or forecasters (who presumably may be familiar with the meteorological conditions that lead to such situations) with clues as to ways in which forecasts would be improved. For example, in the case of the objective maximum temperature forecasts for Minneapolis in the winter season, examination of the conditional distributions $p(x|f)$ (see Fig. 7a) revealed that the forecasts were biased for high and low forecast temperatures but relatively unbiased for intermediate forecast temperatures. To modelers familiar with the numerical-statistical models on which such forecasts are based, these results might provide clues as to ways in which these models might be improved. Of course, supplementary diagnostic studies of a *meteorological* nature would also be desirable to provide further insight into the atmospheric conditions under which such biases occur and/or the reasons for their occurrence.

In the context of subjective weather forecasting, forecasters generally exhibit individual strengths and weaknesses. Diagnostic verification offers the possibility of obtaining more detailed information concerning individual distributional and performance characteristics than is usually available from traditional verification programs. Such information, when provided to forecasters as feedback, can have a beneficial impact on forecast quality (e.g., see Murphy and Daan 1984).

In this paper we have focused on diagnostic verification as a means of providing detailed information to modelers and forecasters, in order to improve forecasting performance. However, the information produced by the diagnostic approach to forecast verification should also be of considerable importance to actual and potential users of forecasts. Such individuals need quantitative information concerning the basic dimensions of forecast quality—information that generally is not provided by traditional verification methods—in order to make optimal use of the forecasts.

Notwithstanding the apparent benefits of the verification methods described in this paper, we believe that it will be possible to develop even more useful forms of diagnostic verification in the future. For example, considerably greater use could be made of methods of exploratory data analysis (e.g., Graedel and Kleiner 1985), and such methods might be especially valuable in applications involving probability forecasts and/or variables with asymmetric distributions (e.g., precipitation probability forecasts and wind speed forecasts). Moreover, more imaginative use could be made of graphical displays (e.g., response surface plots, other bivariate displays), including colors, to facilitate insights into basic distributional and performance characteristics.

In conclusion, traditional verification generally has consisted of characterizing forecast quality in terms of a few overall performance measures (e.g., measures of accuracy and/or skill). Since *many* basic characteristics of forecasts and observations—both individually and jointly—exist, traditional practices are necessarily incomplete and potentially misleading. In the context of comparative verification, for example, the fact that one forecasting system is more accurate (as measured by the mean square error) than another forecasting system is no guarantee that the quality—or value—of the former equals or exceeds that of the latter (see Murphy and Ehrendorfer 1987; Ehrendorfer and Murphy 1988). This fact, when considered in conjunction with the potential benefits of diagnostic verification to both producers and users of forecasts, underscores the desirability of adopting a diagnostic approach to forecast verification in the future.

Acknowledgments. The comments of three reviewers were helpful in preparing a revised version of this paper. The work described herein was supported in part by the National Science Foundation (Division of Atmospheric Sciences) under Grant ATM-8714108. The National Center for Atmospheric Research is sponsored by the National Science Foundation.

APPENDIX A

Regression Models

Simple linear regression models are used in this paper to facilitate the interpretation of the summary measures of the conditional distributions, $p(x|f)$ and $p(f|x)$. In the case of $p(x|f)$, in which the observations are regressed on the forecasts, the linear regression equation describing the relationship between the expected value of the observations given a particular forecast, $E(x|f)$, and the forecast, f, can be written as follows:

$$E(x|f) = a + bf, \tag{A1}$$

where $a = \langle x \rangle - b\langle f \rangle$ and $b = (s_x/s_f)r_{fx}$ are the ordinary least squares estimates of the (unknown) regression coefficients. Here $\langle f \rangle$ is the sample mean of the forecasts, $\langle x \rangle$ the sample mean of the obser-

vations, s_f the sample standard deviation of the forecasts, s_x the sample standard deviation of the observations, and r_{fx} is the sample (product moment) correlation coefficient between the forecasts and observations.

In a verification context, it is obviously desirable for the forecasts of interest to be conditionally and unconditionally unbiased. In terms of the notation employed here, these "requirements" are represented by $E(x|f) = f$ for all f and $\langle f \rangle = \langle x \rangle$, respectively. Thus, the concept of conditionally unbiased forecasts is identical to that of perfectly calibrated—or completely reliable—forecasts (see section 2a). Moreover, it should be noted that forecasts that are conditionally unbiased for all forecast values are necessarily also unconditionally unbiased. That is, $E_f[E(x|f)] = E(x) = \langle x \rangle = E(f) = \langle f \rangle$.

With reference to the regression model in (A1), it can be seen that the requirements of conditionally and unconditionally unbiased forecasts will be satisfied only if $a = 0$ and $b = 1$. Ideally, then, the regression line in this case will have its intercept at zero and will possess unit slope [i.e., it will correspond to the 45° line in a diagram in which $E(x|f)$ is plotted against f]. The latter condition (i.e., $b = 1$) implies that $s_f = s_x r_{fx}$ and, since r_{fx} is always less than or equal to 1, the standard deviation of the forecasts must be less than or equal to the standard deviation of the observations.

In the case of $p(f|x)$, in which the forecasts are regressed on the observations, the linear regression equation describing the relationship between the expected value of the forecasts given a particular observation, $E(f|x)$, and the observation, x, can be expressed as follows:

$$E(f|x) = c + dx, \tag{A2}$$

where $c = \langle f \rangle - d\langle x \rangle$ and $d = (s_f/s_x)r_{fx}$ are the ordinary least squares estimates of these regression coefficients. Furthermore, since $r_{fx} = (s_f/s_x)b$, it follows that $c = \langle f \rangle - (s_f/s_x)^2 b\langle x \rangle$ and $d = (s_f/s_x)^2 b$. Thus, when the forecasts are conditionally and unconditionally unbiased (i.e., when $a = 0$ and $b = 1$) and yet are still imperfect (i.e., $r_{fx} < 1$), the intercept and slope of the regression line in this case will be greater than zero and less than one, respectively (i.e., $c > 0$ and $d < 1$). Moreover, this regression line rather than the 45° line represents the ideal relationship between $E(f|x)$ and x in the situation described by (A2).

APPENDIX B

Discrimination Measure

The measure of discrimination (DIS) employed in this paper is based on the likelihood ratio $\mathrm{LR}(f; x_i, x_j)$, where $\mathrm{LR}(f; x_i, x_j) = p(f|x_i)/p(f|x_j)$. Specifically, the discrimination between the observations x_i provided by the forecast f is denoted by $\mathrm{DIS}(f; x_i, x_j)$, where

$$\mathrm{DIS}(f; x_i, x_j) = \max[\mathrm{LR}(f; x_i, x_j), 1/\mathrm{LR}(f; x_i, x_j)]. \tag{B1}$$

Note that, as defined in (B1), $\mathrm{DIS}(f; x_i, x_j) \geq 1$, with larger values indicating greater discrimination.

To obtain an average value of discrimination for a particular forecast f, it is necessary to average $\mathrm{DIS}(f; x_i, x_j)$ in (B1) over all combinations of observations x_i and x_j. Let $\mathrm{DIS}(f)$ denote this average value. Then,

$$\mathrm{DIS}(f) = [1/\sum_{x \epsilon T} p(x)]^2 \times \sum_i \sum_j p(x_i)p(x_j)\mathrm{DIS}(f_j, x_i, x_j), \tag{B2}$$

where $p(x_i)$ and $p(x_j)$ represent the marginal probabilities (relative frequencies) of the respective observations, $p(x)$ denotes the generic marginal distribution of x, and the set T consists of all values of x for which $p(f, x) > 0$.

Finally, to obtain an overall measure of discrimination, it is necessary to average $\mathrm{DIS}(f)$ in (B2) over all forecasts. If we denote this overall average by DIS, then

$$\mathrm{DIS} = \sum_f p(f)\mathrm{DIS}(f), \tag{B3}$$

where $p(f)$ is the marginal probability distribution of the forecasts. Since $\mathrm{DIS}(f; x_i, x_j) \geq 1$, values of DIS in (B3) close to one indicate relatively little overall discrimination. Larger departures of DIS from this reference value are indicative of greater overall discrimination.

APPENDIX C

Performance Measures

The mean error (ME) for a sample of forecasts f and observations x is defined as follows:

$$\mathrm{ME} = \langle (f - x) \rangle = \langle f \rangle - \langle x \rangle. \tag{C1}$$

ME in (C1) is a measure of the unconditional (or systematic; or overall) bias in the forecasts, and ME = 0 for unconditionally unbiased forecasts.

In an analogous manner, the mean square error (MSE) for a sample of data is defined as follows:

$$\mathrm{MSE} = \langle (f - x)^2 \rangle. \tag{C2}$$

MSE in (C2) is a measure of the accuracy of the forecasts. The root-mean-square error of the forecasts, RMSE, is the square root of MSE in (C2); MSE = RMSE = 0 for completely accurate forecasts.

The skill score (SS) employed in this paper is based upon the MSE. Specifically, it is assumed that the standard of reference (for accuracy) is the MSE for forecasts based solely on sample climatology (i.e., on $\langle x \rangle$). Thus,

$$SS = 1 - (MSE/MSE_{\langle x \rangle}), \quad (C3)$$

where $MSE_{\langle x \rangle}$ denotes the MSE for the reference forecasts. Since $MSE_{\langle x \rangle} = s_x^2$ [see (C2)], it follows that

$$SS = 1 - [\langle (f - x)^2 \rangle / s_x^2]. \quad (C4)$$

SS in (C4) [or (C3)] is a measure of skill (or relative accuracy). In particular, SS = 0 when the forecasts of interest and the climatological reference forecasts are equally accurate, and it is positive when the accuracy of the former exceeds that of the latter.

APPENDIX D

Decompositions of Mean Square Error and Skill Score

The MSE for a sample of forecasts and observations, as defined in (C2), can be decomposed as follows:

$$MSE = (\langle f \rangle - \langle x \rangle)^2 + s_f^2 + s_x^2 - 2 s_f s_x r_{fx} \quad (D1)$$

(Murphy 1988). In this decomposition, the MSE is expressed in terms of summary measures of the marginal and joint distributions of f and x. The first term on the right-hand side (RHS) of (D1) represents a measure of overall bias, whereas the remaining terms—taken together—constitute the variance of the forecast errors (i.e., s_{f-x}^2). Separately, these latter three terms characterize the variability and covariability (or degree of linear association) of the forecasts and observations.

The SS based on the MSE is defined by (C3) [or (C4)]. Substitution of (D1) into (C4) yields

$$SS = r_{fx}^2 - [r_{fx} - (s_f/s_x)]^2 - [(\langle f \rangle - \langle x \rangle)/s_x]^2 \quad (D2)$$

(Murphy 1988). This decomposition expresses SS in terms of summary measures of $p(f, x)$, $p(f)$, and $p(x)$. It is evident, from the previously discussed regression models (see appendix A), that the first term on the right-hand side (RHS) of (D2) is a measure of the degree of (linear) association between the forecasts and observations and that the second and third terms on the RHS of (D2) are measures of conditional and unconditional bias, respectively. This decomposition reveals a fundamental deficiency in the correlation coefficient as a performance measure; namely, it ignores the conditional and unconditional biases in the forecasts (see Murphy 1988).

REFERENCES

Brier, G. W., and R. A. Allen. 1951. Verification of weather forecasts. *Compendium of Meteorology,* ed. T. F. Malone. Boston, MA: American Meteorological Society, 841–848.

Carter, G. M., and P. D. Polger. 1986. A 20-year summary of National Weather Service verification results for temperature and precipitation. Silver Spring, MD, NOAA, National Weather Service, NOAA Technical Memorandum NWS FCST 31. (NTIS PB88 235353/AS).

Ehrendorfer, M., and A. H. Murphy. 1988. Comparative evaluation of weather forecasting systems: sufficiency, quality, and accuracy. *Mon. Wea. Rev.* **116:** 1757–1770.

Glahn, H. R. 1985. Statistical weather forecasting. *Probability, Statistics, and Decision Making in the Atmospheric Sciences,* eds. A. H. Murphy and R. W. Katz. Boulder, CO: Westview Press.

Graedel, T. E., and B. Kleiner. 1985. Exploratory analysis of atmospheric data. *Probability, Statistics, and Decision Making in the Atmospheric Sciences* eds. A. H. Murphy and R. W. Katz. Boulder, CO: Westview Press.

Murphy, A. H. 1988. Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.* **116:** 2417–2424.

——, and H. Daan. 1984. Impacts of feedback and experience on the quality of subjective probability forecasts: comparison of results from the first and second years of the Zierikzee experiment. *Mon. Wea. Rev.* **112:** 413–423.

——, and ——. 1985. Forecast evaluation. *Probability, Statistics, and Decision Making in the Atmospheric Sciences* eds. A. H. Murphy and R. W. Katz. Boulder, CO: Westview Press.

——, and M. Ehrendorfer. 1987. On the relationship between the accuracy and value of forecasts in the cost-loss ratio situation. *Wea. Forecasting* **2:** 243–251.

——, and R. L. Winkler. 1987. A general framework for forecast verification. *Mon. Wea. Rev.* **115:** 1330–1338.

Tukey, J. W. 1977. *Exploratory Data Analysis.* Reading, MA: Addison-Wesley.