The Application of Signal Detection Theory to Weather Forecasting Behavior

LEWIS O. HARVEY JR., KENNETH R. HAMMOND, CYNTHIA M. LUSK, AND ERNEST F. MROSS

Department of Psychology, University of Colorado, Boulder, Colorado

(Manuscript received 26 December 1990, in final form 24 August 1991)

ABSTRACT

A variety of measures are used to judge the skill and accuracy with which forecasters predict the weather and to verify forecasts. Such measures can confound accuracy with decision strategy and sometimes give conflicting indications of performance. Signal detection theory (SDT) provides a theoretical framework for describing forecasting behavior and minimizing these problems. We illustrate the utility of signal detection theory in this context, show how it can be used to understand the effects of time pressure created by frequent weather activity on forecasting judgments, and illustrate how to achieve a specific social policy.

1. Introduction

In the meteorological literature, at least seven major concepts are discussed in the context of evaluating forecasting performance: accuracy, bias, calibration (or reliability), refinement, resolution, discrimination, and skill (Murphy 1985). Each of these concepts has, in addition, several measures associated with it. The resulting situation was summarized by Murphy and Winkler:

> Verification measures have been formulated with a variety of purposes in mind and for a multitude of different situations. . . . As a result, verification measures have tended to proliferate, with relatively little effort being made to develop general concepts and principles, to investigate the relationships between measures, or to examine their relative strengths and weaknesses (Murphy and Winkler 1987, p. 1330).

Not only has there been a proliferation of measures, there does not seem to be a broad consensus about which measures and concepts are appropriate to use. Fildes and Makridakis (1988), for example, noted that participants in the *M* competition did not agree on the meaning, if any, of the average mean-square error (MSE) as a measure of accuracy. Thompson (1990) discusses four of the major problems with using MSE as a measure of overall accuracy. Other examples of criticism of commonly used measures are found in the literature (Glahn 1985; Graedel and Kleiner 1985; Murphy 1985; Murphy and Daan 1985; Winkler and Murphy 1985).

Murphy and Winkler (1987) called for the development of a general framework within which to develop forecast evaluation:

> A need exists for a general framework for forecast verification. To be useful, such a framework should (inter alia) (i) unify and impose some structure on the overall body of verification methodology, (ii) provide insight into the relationships among verification measures, and (iii) create a sound scientific basis for developing and/or choosing particular verification measures in specific contexts. Moreover, such a framework should minimize the number of distinct situations that must be considered (Murphy and Winkler 1987, p. 1330).

The general framework proposed by Murphy and colleagues is based on the decomposition (or factorization) of joint probability distributions combined with performance measures based on the decomposition of mean-square errors (Clemen and Murphy 1986; Cronbach 1955; Hsu and Murphy 1986; Murphy 1988; Murphy et al. 1988; Murphy and Winkler 1987; Murphy and Winkler 1992).

The factorization of distributions is solidly grounded in probability theory, and while it provides an empirical basis for describing forecasting behavior, it does not provide a model of the forecasting process. A model of the process would give a framework within which to understand how the forecasts are generated and to predict forecasting behavior. We believe that signal detection theory (SDT) provides an appropriate model of forecasting behavior, achieves the three goals stated by Murphy and Winkler, is compatible with joint probability decomposition, and provides a simpler description of forecasting behavior.

The purpose of this paper, therefore, is threefold: to illustrate some problems that arise with certain measures of forecasting performance; to show how signal

Corresponding author address: Dr. Kenneth R. Hammond, Institute of Cognitive Science, Muenzinger Psychology Building, Campus Box 344, Boulder, CO 80309-0344.

detection theory offers relief from these problems and leads to an understanding of the effects of time pressure stress on forecasting; and to show how consumers of weather forecasts can make optimal use of those forecasts.

2. Forecast evaluation

Weather forecasting is one of many human activities where people make judgments and/or predictions about events. Therefore, we consider what sort of models of the judgment process have been developed and to what degree these models apply to weather forecasting. Virtually all models of human judgment have at least two components: an information-processing component and a decision-making, response-generating component (Baird and Noma 1978; Green and Swets 1974; Hammond and Adelman 1976; Hammond et al. 1980; Krantz 1969; Macmillan and Creelman 1991; Swets 1986a,b; Swets and Picket 1982; Thurstone 1927). Broadly speaking, the information-processing component builds internal representations in the mind based on information collected from the outside world and on knowledge already contained in memory. The decision-making component examines these internal representations and makes decisions about which of the possible responses to give, taking into account decision goals that might be appropriate to the situation.

In the model of forecasting to be considered here, the signal detection model, these two components operate independently of each other. To describe forecasting behavior in terms of this model two types of measures are used: One type of measure reflects the ability of the information-processing component to discriminate between the occurrence of the event and its nonoccurrence, and another type reflects the decision process that generates responses. Since the two components of the model are independent of each other, the two types of measures should also be independent of each other. Independent, in this context, means that the value of one type of measure does not predict the value of the other. If the model is appropriate for observed behavior, than the first type of measure computed from the data will be independent of a different decision used in generating responses (Egan 1975; Green and Swets 1974; Macmillan and Creelman 1991).

In the simplest forecasting situation the forecaster uses two forecast categories (e.g., "yes" and "no") to predict that an event will or will not occur [example A in Murphy and Winkler (1987)]. The four possible outcomes are illustrated in Table 1 as a 2 × 2 contingency table. There are two types of correct outcomes: a hit (cell A in Table 1) and a correct rejection (cell D). There are two types of errors: a false alarm or false positive (cell B) and a miss (cell C). To properly evaluate forecasting performance from such a contingency table, the frequencies of all four outcomes must be known (Swets 1988).

TABLE 1. A 2 × 2 contingency table used to evaluate weather forecasting performance. The definition of four meteorological performance indices (POD, CSI, CAR, FAR), one meteorological accuracy index (MSE), and two signal detection measures (HR and FAR) are presented.

	Event forecast	
Event	Will occur	Will not occur
Occurs	A (hit)	C (miss)
Does not occur	B (false alarm)	D (Correct rejection)

$$\text{POD} = \frac{A}{A + C} \quad \text{(Probability of detection)}$$

$$\text{CAR} = \frac{A}{A + B} \quad \text{(Correct-alarm ratio)}$$

$$\text{FAR} = \frac{B}{A + B} \quad \text{(False-alarm ratio)}$$

$$\text{CSI} = \frac{A}{A + B + C} \quad \text{(Critical success index)}$$

$$\text{MSE} = \frac{B + C}{A + B + C + D} \quad \text{(Mean-square error, Brier score)}$$

$$\text{SDT HR} = p(Y|s) = \frac{A}{A + C} \quad \text{(Hit rate)}$$

$$\text{SDT FAR} = p(Y|n) = \frac{B}{B + D} \quad \text{(False-alarm rate)}$$

Measures of forecasting performance may be computed from the cells of a 2 × 2 contingency table. Probability of detection (POD), critical success index (CSI), correct-alarm ratio (CAR), false-alarm ratio (FAR), and mean-square error are some of the more common terms (Brier 1950; Donaldson et al. 1975; Fildes and Makridakis 1988; Murphy and Winkler 1987; Thompson 1990). How these indices are computed is given in Table 1.

Forecasts are often made using more than two forecast categories: for example, rating the probability of the event occurring [example B in Murphy and Winkler (1987)]. The data from such forecasts form a 2 × nf joint frequency distribution table as illustrated in Table 2, where nf is the number of probability response categories used. The joint probabilities computed from these joint frequency distributions may be decomposed in various ways (Murphy and Winkler 1987; Murphy and Winkler 1992) that will be discussed later.

The measures indicated in Table 1 are computed from data in the form of Table 2 by combining the nf columns into two columns. A cutoff probability rating is established; the response frequencies below the cutoff for each row are added together to form one column, and those at and above the cutoff are added to form the other column. This procedure is equivalent to in-

TABLE 2. A joint distribution table used to represent weather forecasting performance. Seven forecast probabilities ranging from 0.00 to 1.00 are shown. The cells of the table, $f(f, x)$, represent frequency of each forecast f under the condition that the event occurs ($x = 1$) and does not occur ($x = 0$). Each of the cells may be converted into a joint probability by dividing each one by the total number of forecasts: $N = \sum_{x=0}^{1} \sum_{i=1}^{nf} f(f_i, x)$, where nf is the number of forecast categories (seven in this example).

	Event probability forecast						
Event	0.00	0.20	0.40	0.50	0.60	0.80	1.00
Occurs	$f(0.00, 1)$	$f(0.20, 1)$	$f(0.40, 1)$	$f(0.50, 1)$	$f(0.60, 1)$	$f(0.80, 1)$	$f(1.00, 1)$
Does not occur	$f(0.00, 0)$	$f(0.20, 0)$	$f(0.40, 0)$	$f(0.50, 0)$	$f(0.60, 0)$	$f(0.80, 0)$	$f(1.00, 0)$

terpreting all forecast values below the cutoff probability as a "no" forecast, and all forecasts at and above the cutoff as "yes" forecasts (Macmillan and Creelman 1991; Swets et al. 1961). Thus, a $2 \times nf$ joint frequency distribution table may be used to compute $nf - 1$ values of each of the measures given in Table 1. We now examine these measures for independence from decision or response factors using real weather forecasting data.

Mueller et al. (1987) studied the ability of research forecasters to forecast severe weather at approaches to Stapleton International Airport in Denver. Research forecasters forecast the probability of severe convection defined as radar reflectivity values greater than 30 dBZ_e. At 1-h intervals throughout the afternoon the forecasters assessed the probability that severe convection would occur during the following time periods from the time of the forecast: 0–15, 15–30, 30–45, and 45–60 min. Of the approximately 520 forecasts made, 470 did not result in severe convection, while 50 did. The actual numbers vary slightly for the four forecast intervals.

The measures commonly used by meteorologists (Table 1) were computed by constructing 2×2 contingency tables for each of a series of cutoff judgment probabilities. Forecaster judgments greater than or equal to each cutoff value were taken as "yes" forecasts of severe weather; judgments less than each cutoff were taken as "no" forecasts. The different forecast probabilities are associated with the response-generating component of the model and correspond to a series of internal decision criteria. Therefore, a measure of the information-processing component of the model should be independent of the forecast category used. The computed POD, CSI, CAR, and MSE are plotted in Fig. 1 as a function of cutoff probability for each of the four forecasting periods.

With increase in the cutoff probability, POD decreases, CAR increases, MSE decreases, and CSI first increases and then decreases. None of these indices is independent of the probability response used to define the 2×2 contingency table from which the indices are computed, and therefore, none is a good candidate for a measure of the information-processing component of the SDT forecasting judgment model (which requires independence). These various measures are also difficult to interpret unambiguously. Using POD or MSE, for example, one could argue that forecasting "ability" decreases at higher probability forecast categories but using CAR one could argue just the opposite, that "ability" improves. Also, in the case where only positive ("yes") forecasts are made, both CSI and CAR become equal to the event base rate.

The dependence of these indices on the probability response used to define them makes a meaningful comparison among forecasters and among forecasting conditions complicated. Choosing a fixed probability response (0.50, for example) for the computation of all indices does not solve the problem, because the same numerical probability forecast given by different forecasters, or even the same forecasters under different conditions, may be based on different values of an internal decision criterion. Under these conditions the actual posterior probabilities would not necessarily be equal to 0.5, and indeed they often are not, as is illustrated below in our discussion of calibration.

3. Signal detection theory

Signal detection theory provides a theoretical framework within which to understand the forecasting process. It also provides measures of the information-processing component of the model and measures of the decision component of the model. The potential value of SDT in meteorology has been described both by Mason (1982) and by Swets (1988), but SDT is not yet widespread in this field. The dual-Gaussian, variable-criterion signal detection model was originally introduced into psychology to describe the behavior of observers attempting to detect weak sensory signals (Green and Swets 1974; Swets 1961; Swets et al. 1961). Since its introduction, however, this model has been found to provide a good description of a much wider range of human judgment behavior. In his recent survey, for example, Swets (1986a) found that the Gaussian signal detection model describes recognition memory for words, recognition memory for odors, medical image evaluation, information retrieval, aptitude testing, polygraph lie detection, and weather forecasting. In addition, SDT is an appropriate model of expert judgment (Harvey 1992).

A brief description of the signal detection model is

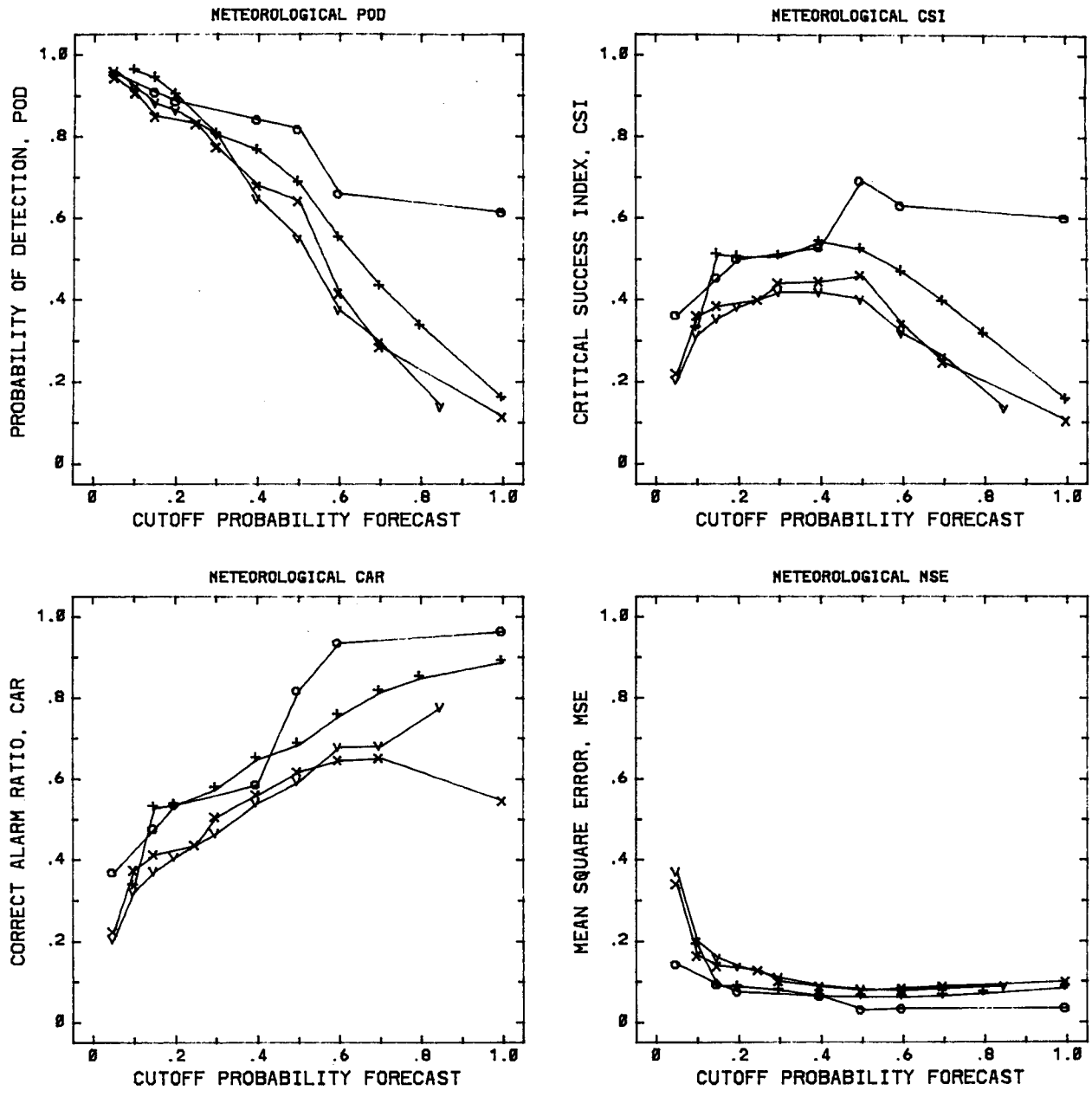

FIG. 1. Four meteorological measures of forecasting performance as a function of cutoff probability used to generate a 2 × 2 contingency table. "O," "+," "×," and "v" represent data for 15-, 30-, 45-, and 60-min forecasts, respectively.

given here. A fuller treatment is found in Green and Swets (1974), Egan (1975), Swets and Pickett (1982), Macmillan and Creelman (1991), and Harvey (1992). The theory asserts that the weather forecaster takes in information relevant to a weather event and forms a quantity that represents the strength of the evidence concerning the occurrence or nonoccurrence of the event. Because of uncertainty associated with making forecasts, the evidence strength fluctuates from occasion to occasion. This fluctuation may be described by two Gaussian probability distributions: One, the event (or signal) distribution, represents the probability distribution of evidence strength associated with the occurrence of the event; the other, the no-event (or noise) distribution, represents the probability distribution of evidence strength associated with the nonoccurrence of the event. Forecasters may not generally be consciously aware of these probability distributions and the process of combining information.

Each of these probability distributions is character-

ized by a mean μ and a standard deviation σ. Two examples of such a model are given in Fig. 2. The horizontal axis, the evidence strength axis or decision axis, is marked off in units of σ_n. By assumption, the no-event distribution has a mean of $\mu_n = 0.0$ and a standard deviation σ_n of 1.0. The mean and standard deviation of the event distribution, μ_s and σ_s, are thus two free parameters and constitute a complete description of the information-processing component of the model.

When a forecaster has combined information and has arrived at a single value representing the strength of evidence about the occurrence or nonoccurrence of an event, this value is compared with one or more internal decision criteria χ_c and decision rules are used to generate an opinion about the event. In the simplest case, example A (Murphy and Winkler 1987), a single decision criterion is used with a single decision rule: "If the evidence is greater than χ_c, say that the event will occur, otherwise say that it will not." The upper panel of Fig. 2 shows a signal detection model with a single decision criterion marked by the vertical line. This criterion partitions the decision axis into two response regions: Judgments that the event will not occur are to the left of χ_c and judgments the event will occur lie to the right of χ_c.

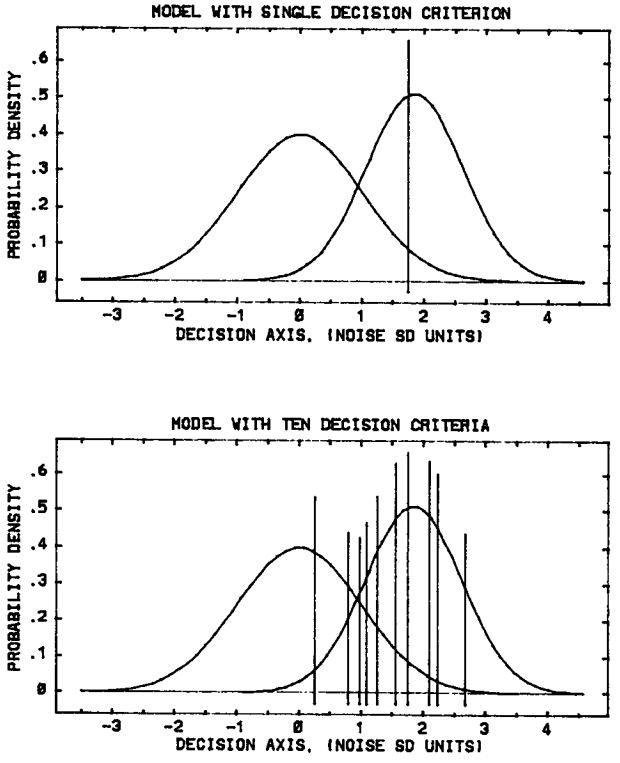

FIG. 2. Two Gaussian signal detection models. The upper model has a single decision criterion. The lower model has ten decision criteria.

Each decision criterion generates two independent conditional probabilities: $p(Y|s)$, the probability of predicting an event given that the event occurs (called the hit rate or HR); and $p(Y|n)$, the probability of predicting an event given that the event does not occur (FAR). When a forecaster uses nf probability forecast categories, $nf - 1$ HR and FAR values result. How HR and FAR are computed is illustrated in Table 1. Note that the false-alarm *ratio* computed by meteorologists (Donaldson et al. 1975) is not the same as the false-alarm *rate* computed for signal detection theory, although the abbreviations for these terms are the same—FAR. The signal detection hit rate (HR), however, is the same as the probability of detection (POD) (Donaldson et al. 1975).

The hit rate corresponds to the area under the event probability distribution lying to the right of χ_c; FAR corresponds to the area under the no-event probability distribution lying to the right of χ_c. The value of χ_c corresponding to a particular FAR may therefore be computed:

$$\chi_c = -z[p(Y|n)] \quad (1)$$

where z [ ] is the transformation of probability into a z score of the unit normal Gaussian distribution. The value(s) of χ_c constitute a complete description of the decision-making component of the model.

The exact value of the decision criterion is usually not under complete conscious control of the forecaster. Extensive experimental evidence indicates that three general factors influence the position of the decision criterion: instructions to and goals held by forecasters, beliefs about the base rate of the event being forecast, and beliefs about the costs and benefits associated with the various correct and incorrect outcomes of a forecast (Green and Swets 1974; Healy and Kubovy 1978; Swets et al. 1961).

Forecasts may be given in terms of probabilities that the event will occur, example B (Murphy and Winkler 1987). We will later discuss the issue of whether or not the forecast probabilities correspond to the actual posterior event probabilities (calibration); it is not relevant here because they are treated as confidence ratings, with higher forecast probabilities indicating a higher confidence that a weather event will occur. Although treating probability forecasts as confidence ratings may seem strange, this practice has a long history in the detection literature. The validity of this approach has been confirmed hundreds of times and forms one of the sources of strong empirical validation of the signal detection model (Egan 1975; Green and Swets 1974; Krantz 1969; Macmillan and Creelman 1991; Swets et al. 1961; Tanner and Swets 1954). In this case the model states that the forecaster holds a series of decision criteria that partition the decision axis into a series of different response categories. When a forecaster uses nf different response probabilities the model assumes that $nf - 1$

different values of χ_c are being held simultaneously. The lower panel of Fig. 2 illustrates a forecaster using 10 different decision criteria resulting in 11 different response categories.

The hit rate $p(Y|s)$ and the false-alarm rate $p(Y|n)$ covary as a function of the decision criterion value, and thus, neither is a suitable measure of the information-processing component of the model, as previously discussed. The relationship between these two conditional probabilities is called the receiver operating characteristic (ROC) and represents all the possible hit rate–false-alarm rate pairs that can be achieved using different decision criteria. The shape of the ROC depends uniquely on the mean and standard deviation of the event probability distribution, and thus is a characteristic of the information-processing component. The ROC generated by the model illustrated in Fig. 2 is shown in the left panel of Fig. 3.

The ten circles on the ROC represent the ten hit rate–false-alarm rate pairs generated by the ten decision criteria shown in the lower part of Fig. 2. Because the model assumes that the two probability density functions are Gaussian, when the hit rate and false-alarm rate probabilities are converted to z-score deviations under the unit normal Gaussian distribution, the ROC becomes a straight line, as is shown in the right panel of Fig. 3. The degree to which actual data lie along a straight line in z-score coordinates is a test of the model's validity (Swets 1986a,b) and may be evaluated by standard statistical methods, such as chi square.

Hit rate–false-alarm rate pairs were calculated from the forecasting data of Mueller et al. (1987). The best-fitting Gaussian signal detection models were also computed from the data using Dorfman's maximum-likelihood method (Dorfman and Alf 1969; Dorfman et al. 1973). A Fortran 77 version of this program may be obtained from the first author (see Acknowledgments). The ROCs formed by the data, and the ROCs predicted by the best-fitting Gaussian model, are plotted in Fig. 4 (on probability scales) and in Fig. 5 (on z-score scales). As the figures illustrate, these forecasting data are extremely well fit by the Gaussian signal detection model in agreement with previous reports (Mason 1982; Swets 1986a,b, 1988).

4. Forecasting accuracy

There are two common ways to express accuracy and we must take care not to confuse one with the other. One way is in probability terms. The probability of making a correct judgment is an example: The higher the probability the higher the accuracy. The other way is in terms of a computed difference between a given response and a desired response. The mean-square error is an example of this type of accuracy measure. In the signal detection literature, accuracy is a term used to describe the ability of the information-processing component of the SDT model to discriminate the event from the nonevent, and we use it in that sense here. Although this component of the model is completely

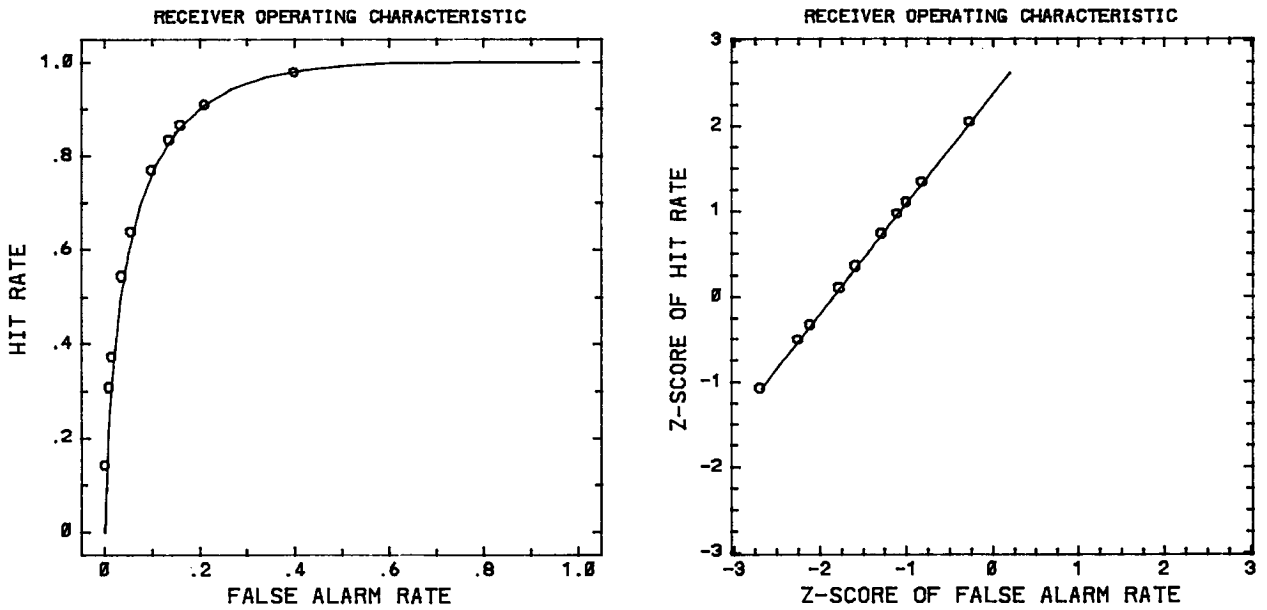

FIG. 3. The receiver operating characteristic (ROC) generated by the model shown in Fig. 2. The hit rate and false-alarm rate pairs correspond to the ten decision criteria shown in the lower panel of Fig. 2. In the left panel, hit rate and false-alarm rates are plotted as probabilities. In the right panel, these probabilities have been transformed into z scores.

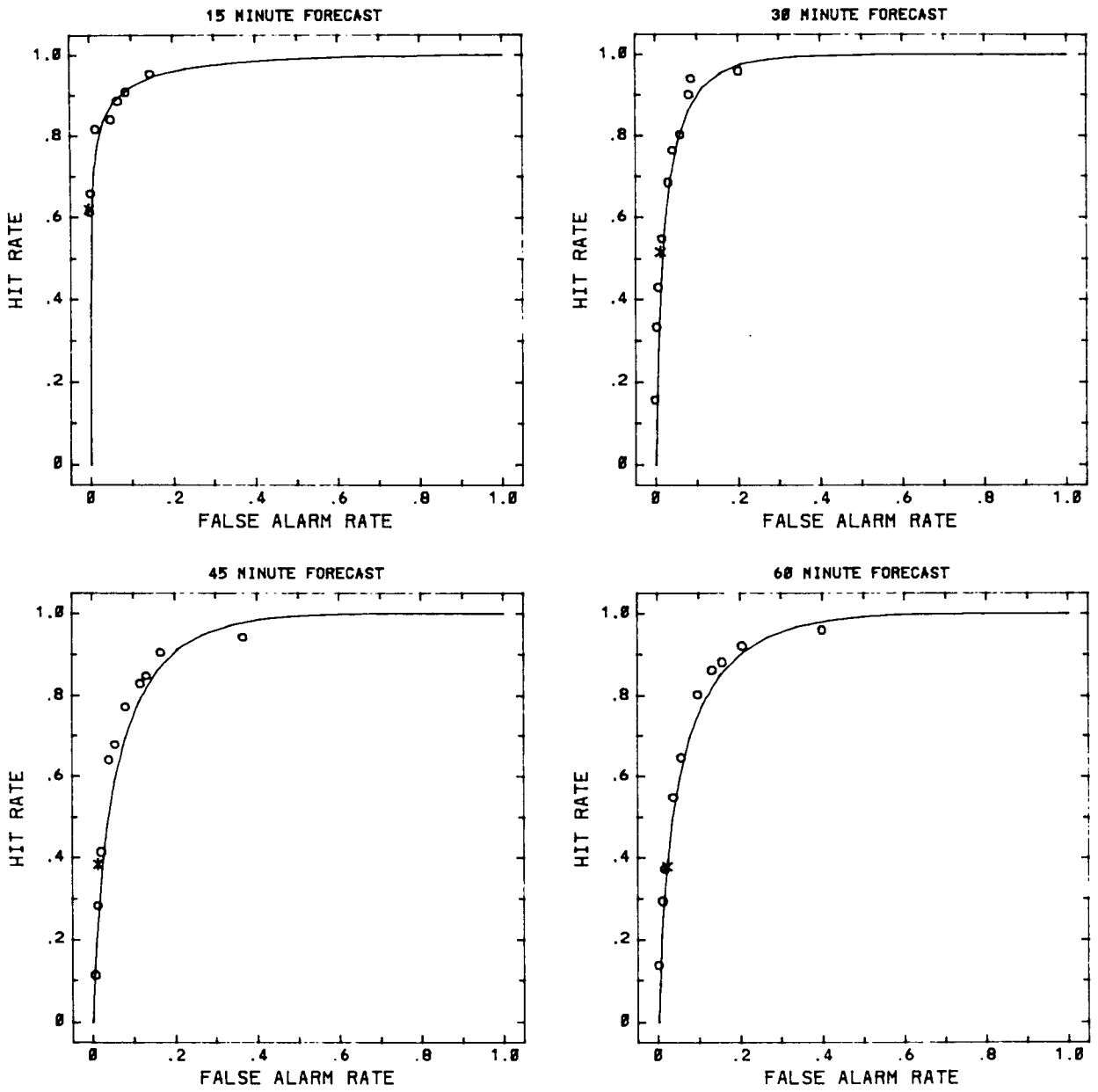

FIG. 4. Hit rate as a function of false-alarm rate for 15-, 30-, 45-, and 60-min forecast intervals. The solid lines are the predictions of the best-fitting Gaussian signal detection model. The asterisk represents the persistence forecast.

specified by μ_s and σ_s, it can be characterized by a single measure of accuracy. A widely used measure of this kind is the area under the ROC, A_z, which may be interpreted as a percent correct score. Chance accuracy corresponds to $A_z = .5$; perfect accuracy to $A_z = 1.0$. It is independent of the decision criterion used and of the event base rate. The variable A_z may be computed from another SDT accuracy measure, d_a, which may in turn be computed from the values of the best-fitting parameters μ_s and σ_s ($\mu_n = 0$ and $\sigma_n = 1$, by assumption):

$$d_a = (\mu_s - \mu_n)\left(\frac{\sigma_s^2 + \sigma_n^2}{2}\right)^{-1/2} \quad (2)$$

$$A_z = z^{-1}\left[\frac{d_a}{\sqrt{2}}\right] \quad (3)$$

where $z^{-1}[\ ]$ is the inverse z transform based on the unit normal Gaussian distribution. The variable A_z is independent of both the decision criterion and the event base rate, and thus provides a basis for directly comparing different forecasters and forecasts made

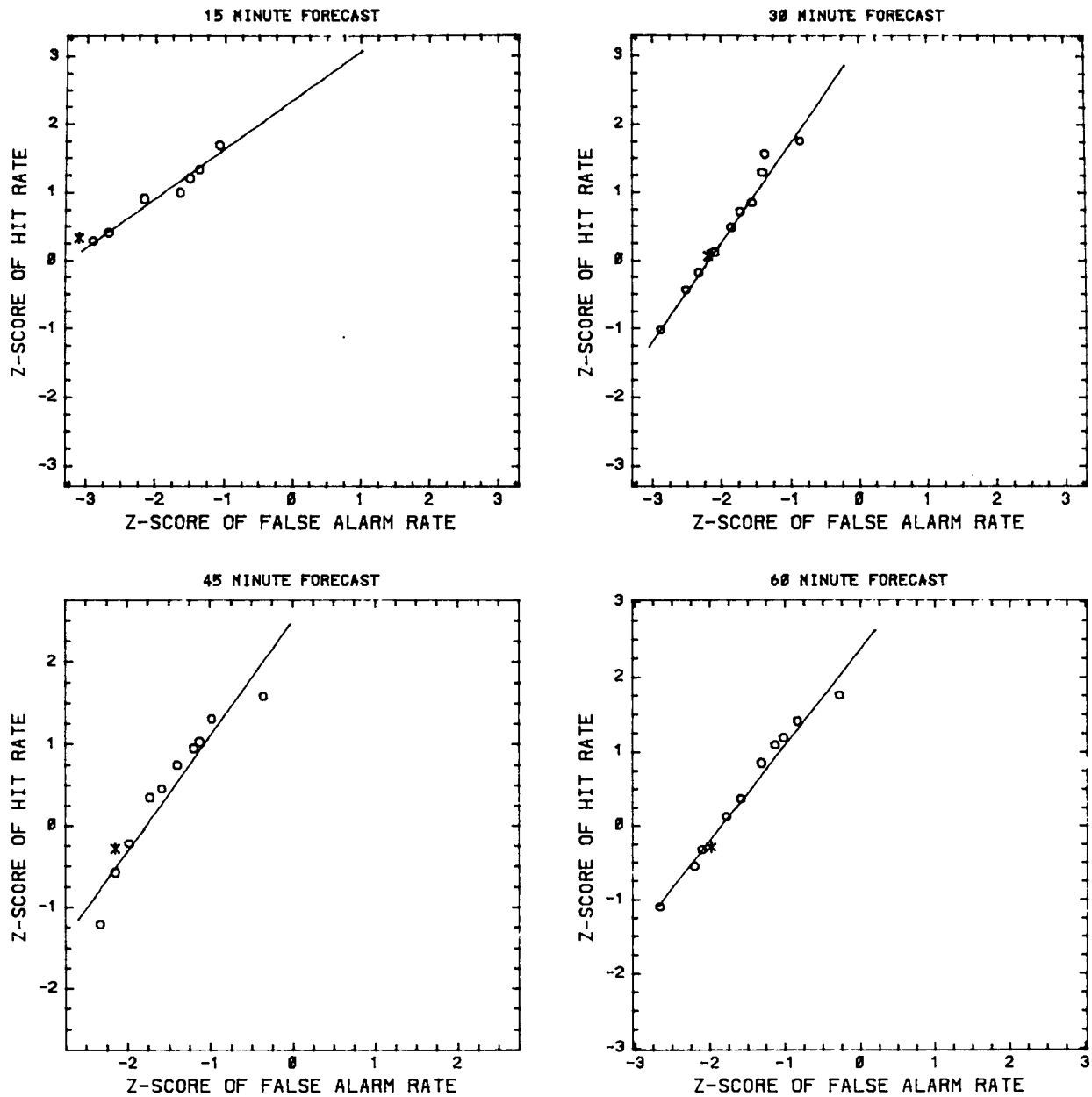

FIG. 5. The z-score hit rate as a function of z-score false-alarm rate for 15-, 30-, 45-, and 60-min forecast intervals. The solid lines are the predictions of the best-fitting Gaussian signal detection model. This prediction is a straight line in z-score coordinates. The asterisk represents the persistence forecast.

under different conditions. The forecasters in the Mueller et al. (1987) experiment achieved A_z of .97, .96, .93, and .93 for the 15-, 30-, 45-, and 60-min forecasts, respectively. These are high levels of accuracy that decrease only slightly as the forecast interval increases, in agreement with the conclusions of those authors.

Accuracy A_z was computed using the empirical hit rate $p(Y|s)$ and false-alarm rate $p(Y|n)$ by first calculating μ_s:

$$\mu_s = \sigma_s z[p(Y|s)] - z[p(Y|n)] \quad (4)$$

where z [ ] is the transformation of probability into a z score of the Gaussian unit normal probability distribution, σ_s is from the maximum-likelihood analysis of the forecast data (Harvey 1992) and then using Eqs. (2) and (3) to compute A_z. These values of A_z are plotted in Fig. 6 as a function of the cutoff probability used to define the contingency table. One sees in Fig. 6 that accuracy is independent of the cutoff probability,

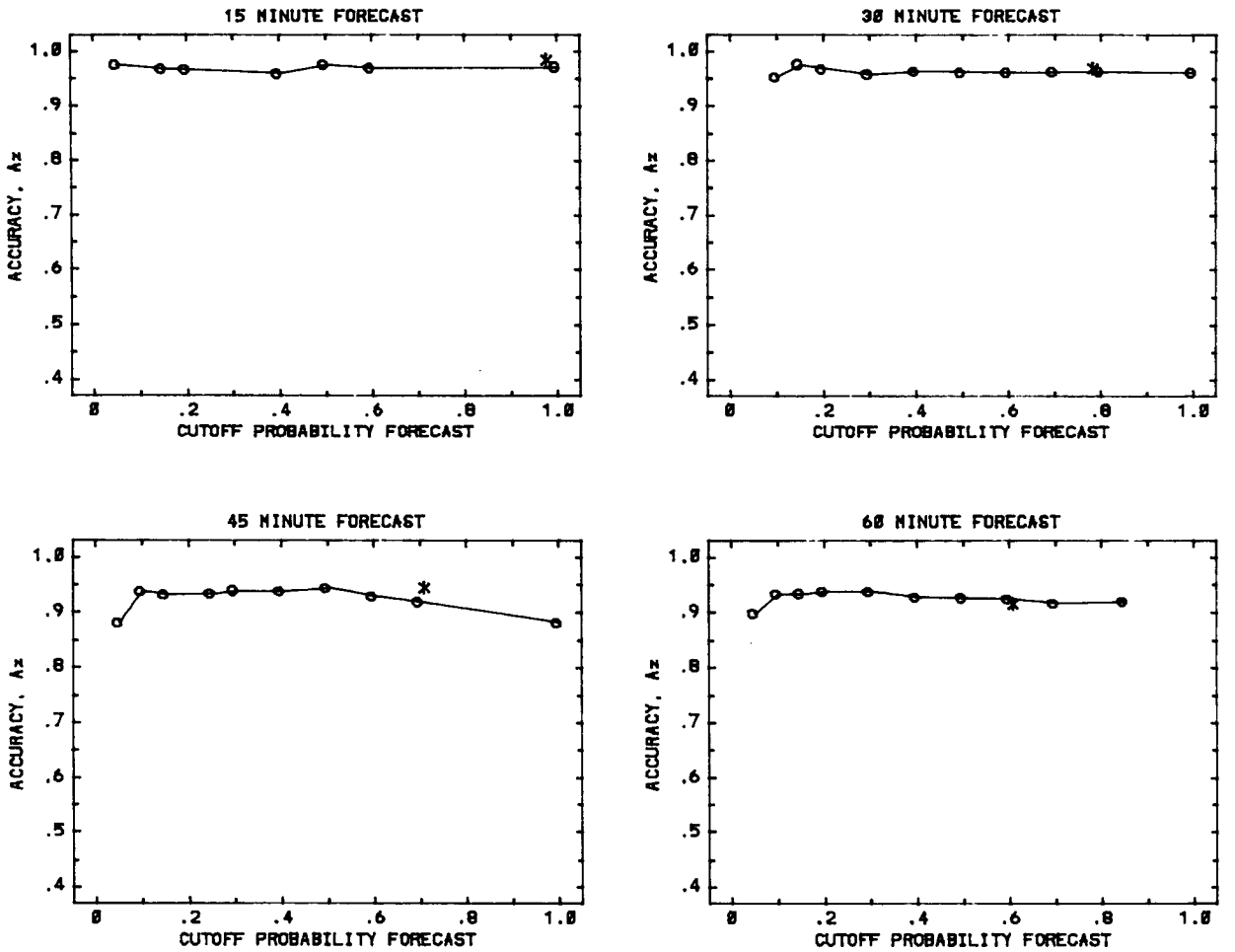

FIG. 6. Accuracy A_z as a function of cutoff probability forecast for the 15-, 30-, 45-, and 60-min forecast intervals. Each A_z was computed from the 2 × 2 table defined by the cutoff probability and the σ_s from the best-fitting signal detection model. The asterisk represents the persistence forecast.

as predicted by the signal detection model, and quite the opposite of the strong dependence found with the indices plotted in Fig. 1.

5. Forecasting performance

Signal detection accuracy A_z is *not* a measure of performance. Accuracy is independent of the decision criteria used to make forecasts and of the base rate of the event being forecast and should be viewed as a measure of *potential* performance. It does permit the comparison of forecasters with each other without contamination by these factors. The decision component of the model (measured by χ_c), on the other hand, is affected by belief about the base rate of the event, as well as the other factors mentioned here. Actual performance is affected both by accuracy (which is independent of base rate) and by decision criteria (which in turn are influenced by base rate). One measure of performance is the posterior hit probability $p(s|Y)$, the probability that the event will occur given that the forecaster says "yes" it will. It is a direct measure of the believability of the forecast, and is therefore a proper measure of performance (Harvey 1992).

The posterior hit probability is computed from the hit rate and the false-alarm rate by means of Bayes' theorem (Hayes 1973) in combination with the prior probabilities of the event happening and not happening:

$$p(s|Y) = \frac{p(Y|s)p(s)}{p(Y|s)p(s) + p(Y|n)p(n)} \qquad (5)$$

where $p(Y|s)$ is the hit rate, $p(Y|n)$ is the false-alarm rate, $p(s)$ is the prior probability of the event occurring (base rate), and $p(n)$ is the prior probability of the event not occurring. The posterior hit probability is formally equivalent to the correct-alarm ratio (CAR)

defined in Table 1. The posterior hit probabilities predicted by the best-fitting signal detection models for the four forecast intervals are plotted in Fig. 7 as a function of the cutoff probability. The posterior hit probability (designated here as performance) increases as the cutoff probability increases. Performance is actually better on the 60-min forecast than on the 45-min forecast for forecast probabilities higher than 0.40, even though accuracy in these two conditions is equal. This situation arises because different values of the decision criteria χ_c are being used to generate the same forecast probabilities in the two time periods. Even though performance increases at the higher forecast probabilities, this increase is achieved at the inevitable cost of lowering the hit rate (Harvey 1992).

6. Persistence forecasts

Mueller et al. (1987) concluded that the forecasters were better than the persistence forecast (the forecast that the future weather will be the same as the current weather). They based this conclusion on a comparison of the probability of detection (POD) and the false-alarm *ratio* (FAR) generated by the persistence forecast with the POD and FAR computed from the 2 × 2 contingency table of the forecasters based on a cutoff forecast probability of 0.50.

We have analyzed the persistence forecast data given in Table 1 of Mueller et al. (1987) and draw a different conclusion. The hit rate and false-alarm rate were computed for the 15-, 30-, 45-, and 60-min forecasts and are plotted as an asterisk on the ROCs of Figs. 4 and 5. The persistence forecast hit and false-alarm rates fall on the ROCs of the research forecasters, indicating that the accuracy of the persistence forecast is equal to that of the forecasters. The accuracy A_z of the persistence forecasts were .98, .96, .94, and .91 for the 15-, 30-, 45-, and 60-min forecasts. These values are not significantly different from the accuracies achieved by the forecasters themselves.

How should the persistence forecast be compared to the human forecast? The hit rate and false-alarm rates achieved by the persistence forecast can be thought of as being made by a perfectly calibrated forecaster; that is, nature is perfectly calibrated (Hsu and Murphy 1986; Murphy and Winkler 1987, 1992). If nature were perfectly calibrated, the probability forecast given by nature would correspond to the posterior probability of a hit. The computed posterior hit probabilities of the persistence forecasts are .9818, .7881, .7146, and .6132 for the forecast intervals of 15, 30, 45, and 60 min, respectively. These probabilities are all higher than the .5 assumed by Mueller et al. (1987) in making the comparison with the forecasters. When the persistence forecast is plotted (see asterisks in the figures) at the appropriate cutoff probability, both accuracy (Fig. 6) and posterior probability performance (Fig. 7) are the same as achieved by the forecasters.

Just because the persistence forecast and the researchers' forecasts are of equal accuracy does not mean that they are of equal utility. The persistence forecast is based on only two response categories, "yes" and "no." Forecasters, on the other hand, typically use ten or more different probability categories. As we shall illustrate at the end of this paper, the more categories that a forecaster can use reliably, the greater the possibility that the consumer of the forecast can make optimal decisions based on the forecast.

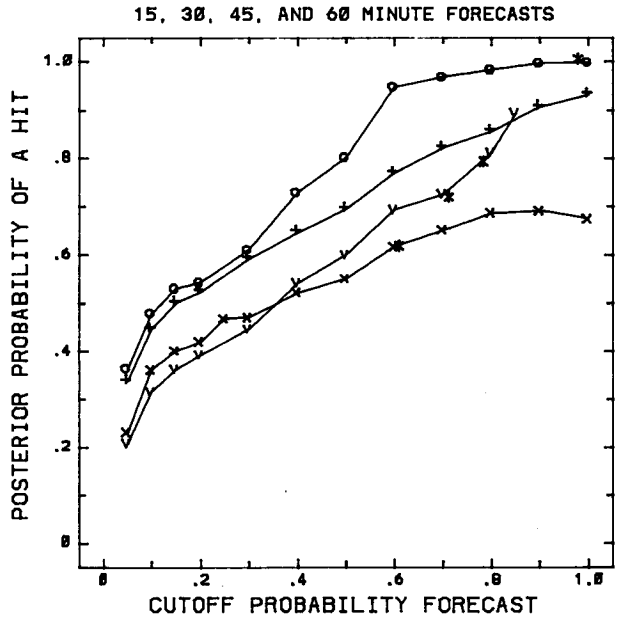

FIG. 7. Posterior hit probability of the weather event occurring as a function of cutoff probability forecast. Symbols "O," "+," "×," and "v" represents data for 15-, 30-, 45-, and 60-min forecasts, respectively. The asterisk represents the persistence forecasts.

7. Forecast calibration

Forecasters and end users are interested in the posterior probability of the event occurring given that a particular forecast probability was made. The relationship between this type of posterior probability and the probability forecast is called calibration or reliability (Murphy and Daan 1985; Murphy and Winkler 1987, 1992). The frequency of severe weather events is often low; therefore, the frequency of certain forecast probability categories is also low, making reliable estimation of the posterior probabilities directly from the observed data difficult (the reliability of probability estimates increases with the square root of the number of observations on which they are based). The signal detection model that best fits a set of data can be used to generate predicted posterior probabilities to evaluate calibration. The model is completely specified by μ_s, σ_s, and the

$N - 1$ values of the decision criteria χ_c derived from the observed data using the maximum-likelihood method referenced earlier. The predicted posterior response probabilities may be computed using the joint probabilities predicted by the model (Harvey 1992; Murphy and Winkler 1987; Murphy and Winkler 1992). The calibration curves based on predicted posterior probabilities for the Mueller et al. (1987) forecasters are shown in Fig. 8. Because the SDT model fits these data very well, the predicted probabilities are a good representation of the observed probabilities and the features reported below are not an artifact of the SDT model.

If the posterior probabilities matched the probability categories used by the forecasters, the data points in Fig. 7 would lie along the straight diagonal line. One sees in this figure that calibration is pretty good for all the forecast intervals except for 15 min. In the latter case, the forecasters overestimate low posterior probabilities and underestimate high posterior probabilities.

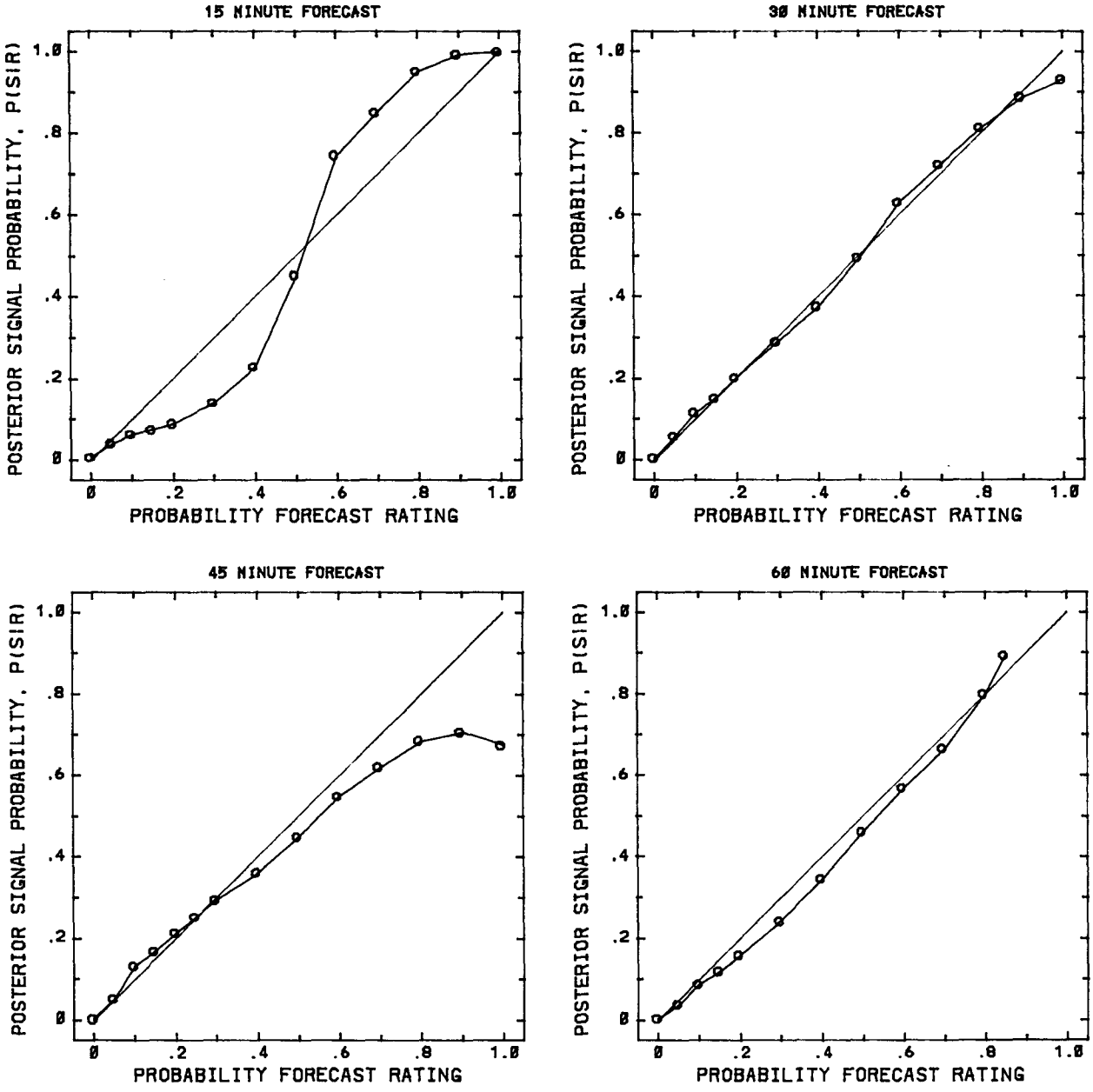

FIG. 8. Calibration curves for the 15-, 30-, 45-, and 60-min forecasts. The posterior probability of the event given a particular prediction response category is plotted as a function of the probability forecast rating used by the forecasters.

Methods for quantifying the degree of calibration discussed in the literature may be effectively applied to the data plotted in Fig. 8 (Murphy and Winkler 1987, 1992).

A point that we wish to emphasize is that calibration is independent of accuracy A_z. Knowing accuracy carries no information about calibration. A well-calibrated forecaster could have high or low accuracy. A poorly calibrated forecaster could at the same time be highly accurate. In principle, for a constant level of accuracy, the decision criterion values χ_c could be manipulated to achieve any degree of calibration desired. This independence allows the separate evaluation of these two aspects of forecasting behavior: the accuracy of the information-processing component and the calibration of the response-generating component.

8. Additional comments

Murphy has developed measures of forecast evaluation based on decompositions of probability forecast data of the kind illustrated in Table 2 (Murphy and Winkler 1987, 1992). The joint frequencies are first converted into joint probabilities of the form $p(f, x)$. There are two ways to factor these joint probabilities into conditional probabilities and marginal probabilities: likelihood–base-rate factorization:

$$p(f, x) = p(f|x)p(x) \tag{6}$$

and calibration–refinement factorization:

$$p(f, x) = p(x|f)p(f) \tag{7}$$

where, as before, f is the forecast probability and $x = 1$ if the event occurs and $x = 0$ if the event does not occur, $p(x)$ is the base rate of the event, and $p(f)$ is the rate of giving a specific forecast probability. These decompositions represent different aspects of forecasting performance and are in no way incompatible with the signal detection model. Indeed, once the SDT model is fit to a particular set of data it can then generate all the joint and conditional probabilities contained in Eqs. (6) and (7).

The likelihood–base-rate factorization of Eq. (6) contains the information used to define the accuracy of the SDT information processing component. This measure, A_z, is the area under the ROC; the ROC is the relationship between hit rate and false-alarm rate as decision criterion is varied; the hit rates are composed of the conditional probabilities of Eq. (6) when the forecast event occurs $[p(f|x = 1)]$ and the false-alarm rates are composed of the conditional probabilities of Eq. (6) when the forecast event does not occur $[p(f|x = 0)]$. The accuracy A_z has the virtue that it is not contaminated by changes in decision criteria nor is it affected by the event base rate, and therefore, it is a relatively pure measure of the ability to discriminate the occurring event from the nonoccurring event.

In the psychological literature as well as the meteorological literature, many measures of accuracy have been proposed in addition to d_a and A_z discussed here. The MSE first proposed by Brier in this context (Brier 1950) and developed more fully by others (Murphy 1973, 1986, 1988; Murphy and Winkler 1987; Thompson 1990; Yates and Curley 1986) has been widely used as the basis for accuracy measures. These measures are based on linear combinations of probabilities (i.e., sums and differences). Such linear combinations of probabilities are inadequate as measures of the accuracy of the information-processing component of forecasting behavior for two reasons: They are not independent of decision processes and response bias (see Fig. 1); and they imply theoretical models of the judgment and forecasting process that are incompatible with observed data (Krantz 1969; Macmillan and Creelman 1991; Mason 1982; Swets 1961, 1986a,b; Swets et al. 1961). Linear decomposition of MSE scores into various components also suffers from the same two problems.

The calibration–refinement factorization [Eq. (7)] is also closely related to the signal detection model. The marginal forecast rate $p(f)$ of Eq. (7) is directly determined by the position of the decision criteria χ_c of the signal detection model. The conditional probabilities $p(x|f)$ are the posterior event probabilities previously discussed under calibration and represent important information about whether or not to take a forecast probability at face value. The measure of forecasting performance that we proposed, the posterior hit probability, is based on the conditional probabilities of Eq. (7). Actually, as Murphy and Winkler have pointed out, the posterior probabilities of Eq. (7) may be computed from the conditional probabilities of Eq. (6) and the base rates of Eq. (6) using Bayes' theorem [Eq. (5)]. We are in complete agreement with Murphy and Winkler, who wrote:

> We believe that these factorizations constitute complementary rather than alternative ways to approach the verification problem. After all . . . the two factorizations are concerned with different attributes of the forecasts and/or observations. Thus a complete verification study would necessarily involve the evaluation of factors associated with both factorizations (Murphy and Winkler 1987, p. 1335).

The signal detection model allows a complete description of forecasting behavior that contains the important elements of both types of decompositions. Like the factorization described above, the model also decomposes observed data into two components. Unlike the factorization method, these two components are independent of each other. The model provides the basis for imposing structure on the overall body of forecast evaluation measures in the form of a widely accepted model of the human judgment process. It

provides insight in the relationships among evaluation measures by dividing forecasting into two components: an information-processing component and a response-generating or decision component. The validity of this model is strongly supported by a wide range of empirical data, and thus creates a sound scientific basis for developing specific measures to be used in specific contexts.

We have analyzed the forecasting data reported by Murphy and Winkler (1987). The data were 2820 probability-of-precipitation (PoP) forecasts formulated by National Weather Service (NWS) forecasters at Chicago, Illinois, during the period July 1972–June 1976. The forecasts were for the 12-h period from 12 to 24 h following the forecast. The dual-Gaussian signal detection model fits these data well and cannot be rejected on statistical grounds. The upper two graphs in Fig. 9 are the ROC computed from the data. The solid line in each graph is the ROC predicted by the best-fitting model. The accuracy A_z of the PoP forecasts is .85, lower than the over-.90 accuracy achieved by the research forecasters discussed above (Mueller et al. 1987).

The lower left panel of Fig. 9 represents forecasting performance (posterior hit probability) as a function of cutoff probability corresponding to different decision criteria. The behavior of the data (open circles) is in strong agreement with the prediction of the signal detection model (solid function). The posterior event probability is close to the base-rate probability (the horizontal line) for low-decision criteria (corresponding to low-probability forecasts) and steadily increases as the decision criterion increases. The panel on the lower right represents the calibration of the forecaster. The open circles are the posterior probabilities conditional on forecast response categories [$p(x|f)$ in Eq. (7)] and the solid functions, are the predictions of the best-fitting model. Both the data and the predictions fall close to the 45° line, indicating very good calibration. We conclude that the signal detection model allows one to capture the important aspects of PoP forecasting behavior.

9. Effects of stress on forecasting behavior

As meteorologists become more closely involved in aviation weather forecasting, stress increasingly becomes a factor in defining their work load. Because the weather information supplied to air traffic controllers often determines air traffic patterns, the timeliness of weather advisories can become critical. Significant weather activity also increases demand on forecasters for rapid assimilation of vast quantities of information and for making rapid judgments. The consequences of error have sharply increased because of the increase in air traffic and the introduction of large jet aircraft. In short, as aviation weather forecasters become an even more integral part of air safety, stress becomes an increasingly significant element of their work. Unfortunately, knowledge regarding the effects of stress on cognitive activity lacks a strong scientific base (Hammond 1990).

We analyzed the research forecast data (Mueller et al. 1987) to learn whether stress affects accuracy, decision criteria, or both. The definition of stress was based on discussions with the forecasters themselves and was defined in terms of the amount of the weather activity predicted by the forecaster. Low-activity days were those for which forecasters predicted little activity and as a consequence did not feel much stress; high-activity days were those for which forecasters predicted much activity. On these days the forecasters reported feeling stressed by the time pressure while making forecasts. Days were therefore divided into low- and high-activity days by a median split of amount of weather activity. Each sample included 194 forecasts spanning 13 days.

The Gaussian signal detection model was fit to the data with the maximum-likelihood method used previously. The SDT model provides excellent fits to these data and the resulting ROCs are shown in Figs. 10 and 11. Accuracy of forecasting A_z is higher under the high-stress condition than under the low-stress condition for all four of the forecast intervals. The ROCs of the high-stress condition have steeper slopes than those of the low stress. This slope difference means that under high-stress conditions σ_s is smaller than under low stress. We note that this difference is found for all four forecast intervals and therefore is a reliable finding.

Forecast performance, as measured by the posterior hit probability, for different cutoff response probabilities is higher for the high-stress condition, although this performance difference is reduced at the very highest forecast probabilities as is shown in Fig. 12. In fact, for the 15-min forecast, performance is equally good under the two conditions for forecast probabilities above 0.5.

Calibration also changes in the high- and low-activity conditions. With the exception of the 15-min interval forecast, generally forecasters overestimated the posterior probability on low-activity days and underestimated it on high-activity days, as is shown in Fig. 13. Thus, the good calibration curves seen in Fig. 8 are really based on a combination of over- and underestimation under the two stress conditions. When the data are grouped together, these effects cancel each other out. The calibration depends on the value of the decision criteria χ_c, which as outlined in section 3, are themselves influenced by three general factors, including belief about the base rates. Generally the decision criteria are lowered when the event to be predicted becomes more likely (Healy and Kubovy 1978).

An examination of the decision criteria under the two stress conditions reveals that they shift in the ex-

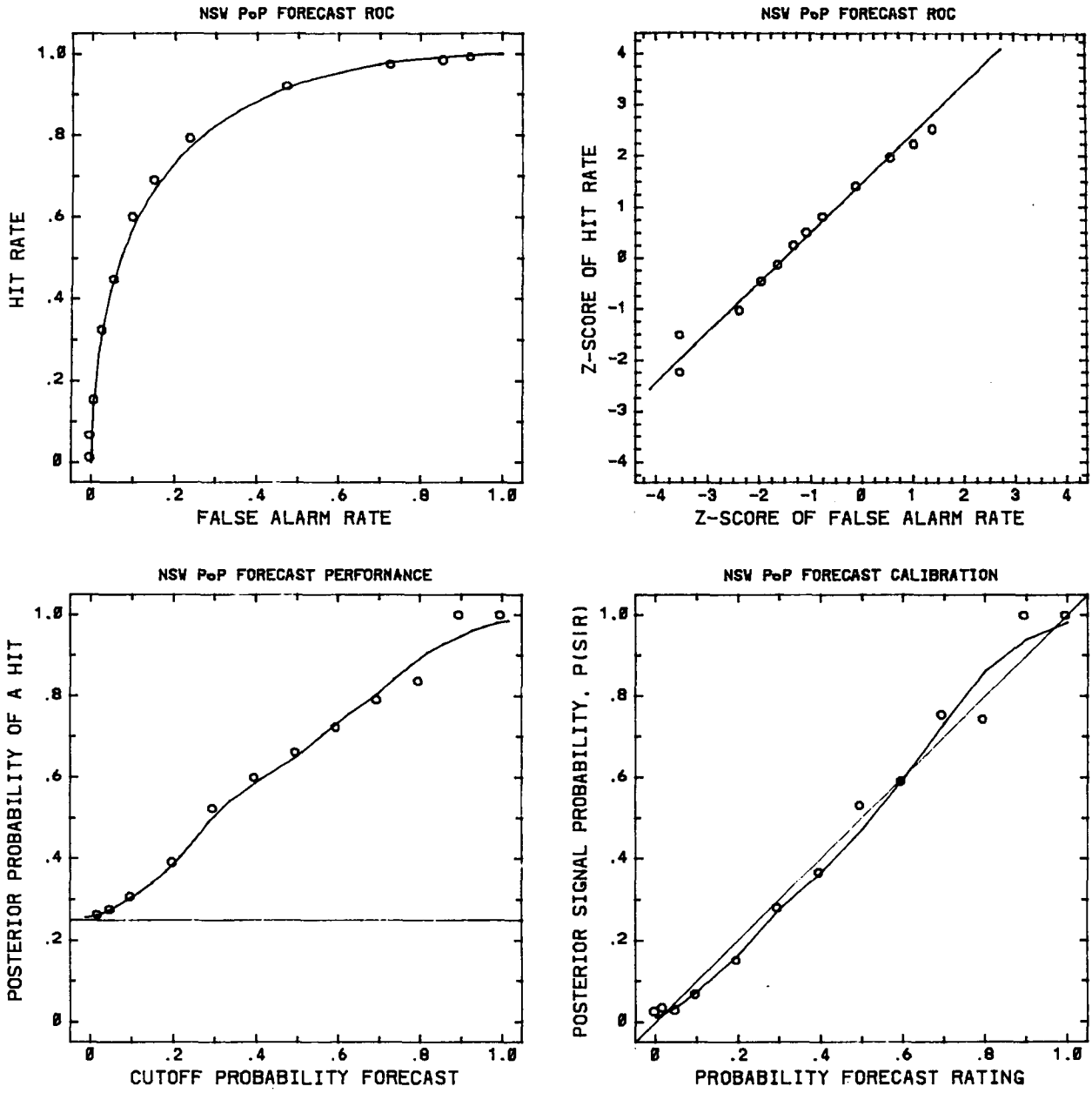

FIG. 9. Description of the PoP forecast data taken from Murphy and Winkler (1987). The upper left panel is the ROC in probability units. The area under the ROC is .85. The upper right panel is the ROC in *z*-score probability units. The solid line in both panels is the prediction of the best-fitting dual-Gaussian signal detection model. The lower left panel represents forecasting performance and the lower right panel represents calibration. See text for details.

pected direction. This shift is illustrated in Fig. 14 where the different decision criteria χ_c are plotted as points at the accuracy level A_z of the high- and low-stress conditions. As was shown in Figs. 10 and 11, the low-stress condition forecasts are less accurate than those made under the high-stress conditions. In Fig. 14, decision criteria that represent boundaries between identical response categories (e.g., between .4 and .5, or between .7 and .8) are each connected by a straight line. One sees in Fig. 14 that the decision criteria shift to the left to lower values when the forecasters change from low- to high-stress conditions. Lowering the decision criterion increases the false-alarm rate, but since this shift is accompanied by an increase in accuracy, the hit rate must be increasing by a much larger amount than the false-alarm rate.

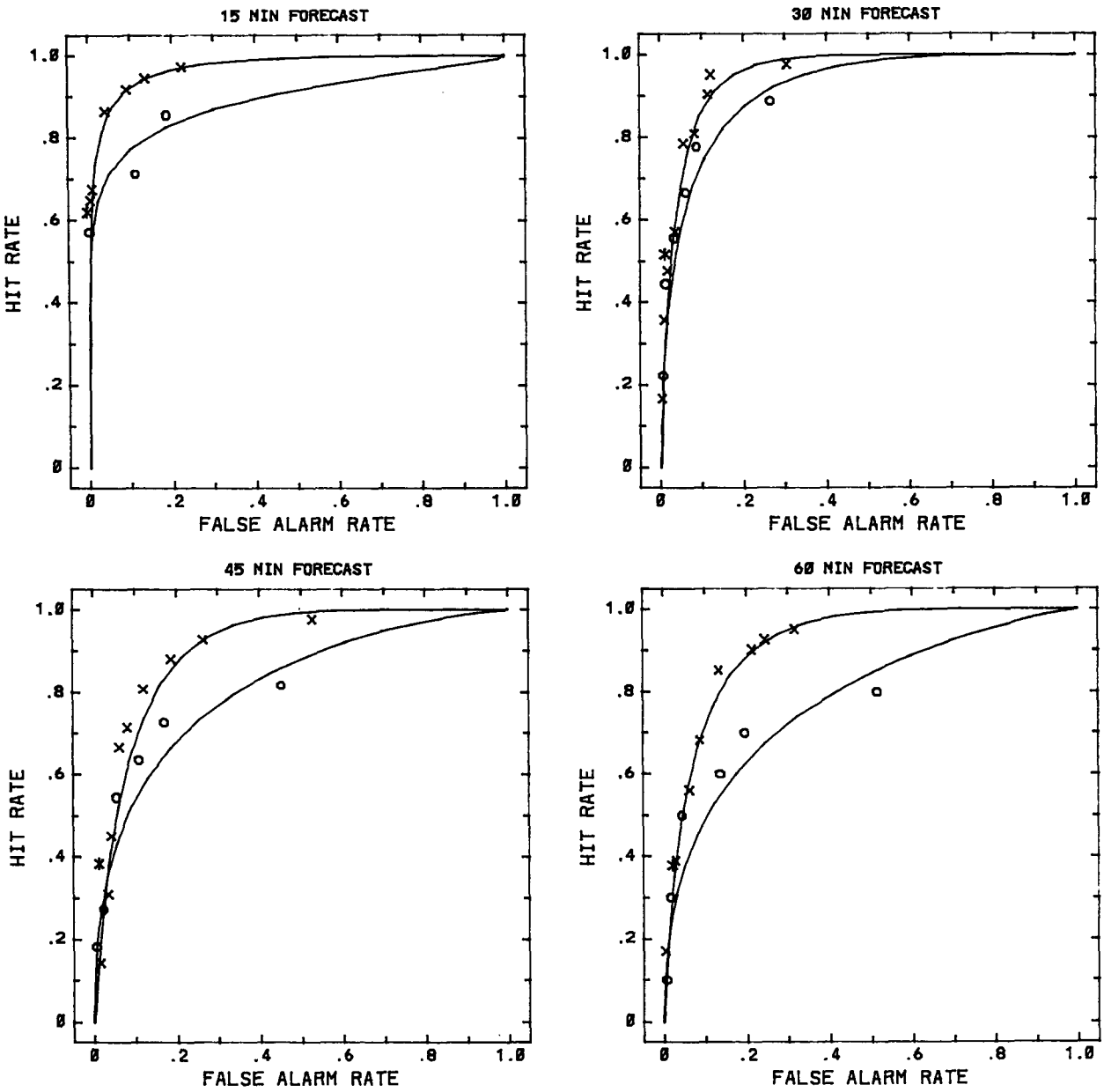

FIG. 10. Hit rate as a function of false-alarm rate for 15-, 30-, 45-, and 60-min forecast intervals. High-activity days are marked by "×"; low-activity days by "O." The solid lines are the predictions of the best-fitting Gaussian signal detection model. The asterisk represents the persistence forecast.

10. Using weather forecasts

The consumer of a weather forecast is usually a person who must decide whether or not to take action based on the forecast. For example: Given that the forecaster says there is an .80 probability of rain, should you wear a raincoat? Given that the forecaster says there is a .60 probability of severe weather, should a flight controller change the landing pattern of aircraft at an airport?

The consumer may have goals that take into account social policy. There are many different possible goals: to achieve maximum percent correct; to achieve the highest hit rate for a specified false-alarm rate; to achieve a specified ratio of hits to misses; to minimize cost or maximize benefits. We wish to emphasize the point that the goals of the consumer may be quite different from the goals of the forecaster. The analysis presented next is part of the field of decision making, whose principles are well developed (Forgionne 1986).

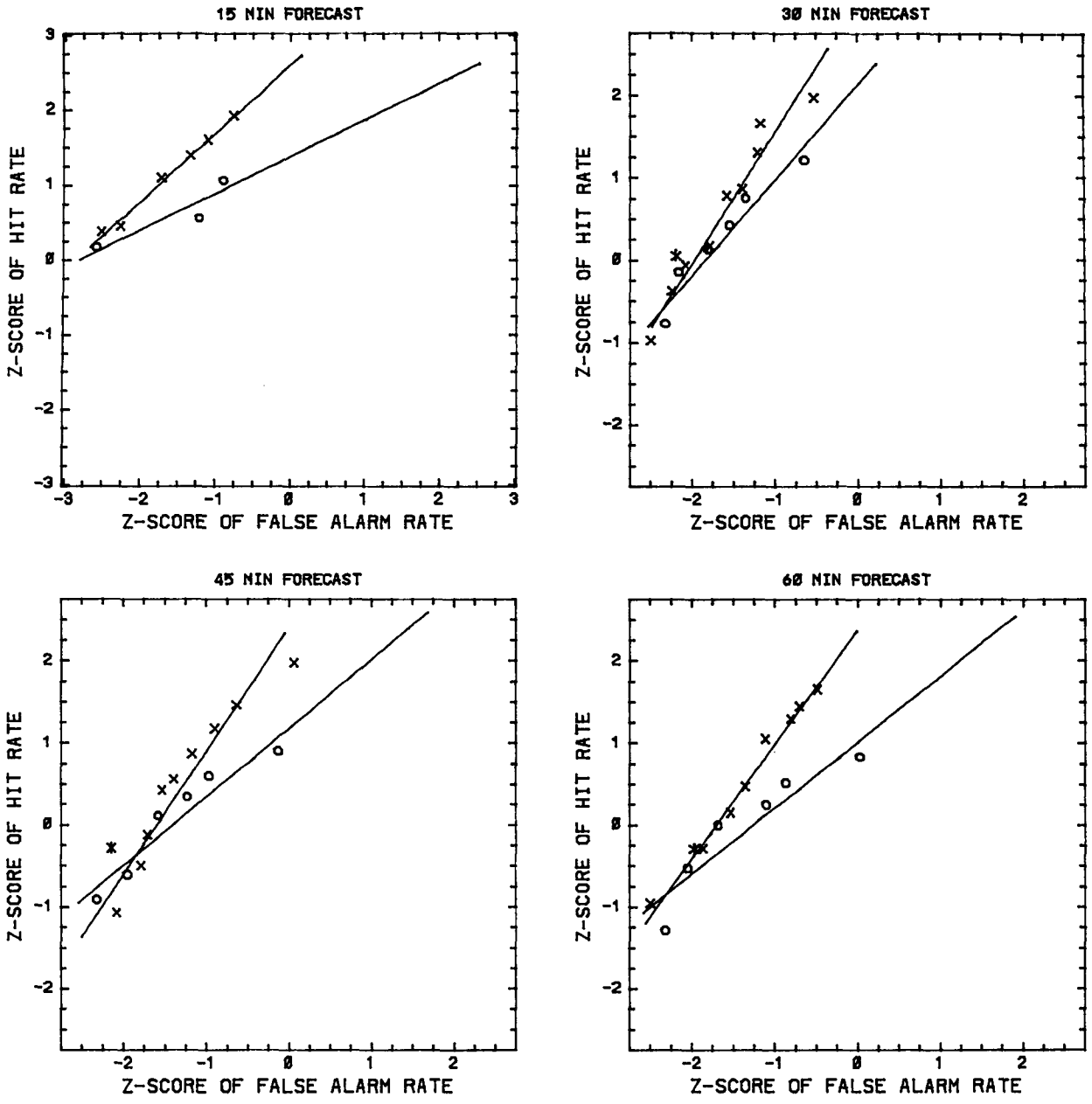

FIG. 11. The z-score hit rate as a function of z-score false-alarm rate for 15-, 30-, 45-, and 60-min forecast intervals. High-activity days are marked by "×"; low-activity days by "○." The solid lines are the predictions of the best-fitting Gaussian signal detection model. This prediction is a straight line in z-score coordinates. The asterisk represents the persistence forecast.

In order to make the best possible decisions in light of a specific goal, the consumer needs to know about both the forecaster who has generated the forecast and the events themselves. Decision making and its consequences have been an important part of both the signal detection literature (Egan 1975; Green and Swets 1974; Macmillan and Creelman 1991; Swets and Picket 1982) and the meteorological literature (Murphy 1977; Murphy and Katz 1985; Thompson 1952; Thompson and Brier 1955; Winkler and Murphy 1985). To use the signal detection model, the consumer needs to know μ_s, σ_s, and the decision criteria χ_c that describe the forecaster. The consumer also needs to know the prior probability of the event itself, $p(s)$. Armed with this information, the consumer can use the forecast to make an action decision that will implement whatever social policy or cost-benefit outcome is desired. The model offers the advantage of predicting probabilities

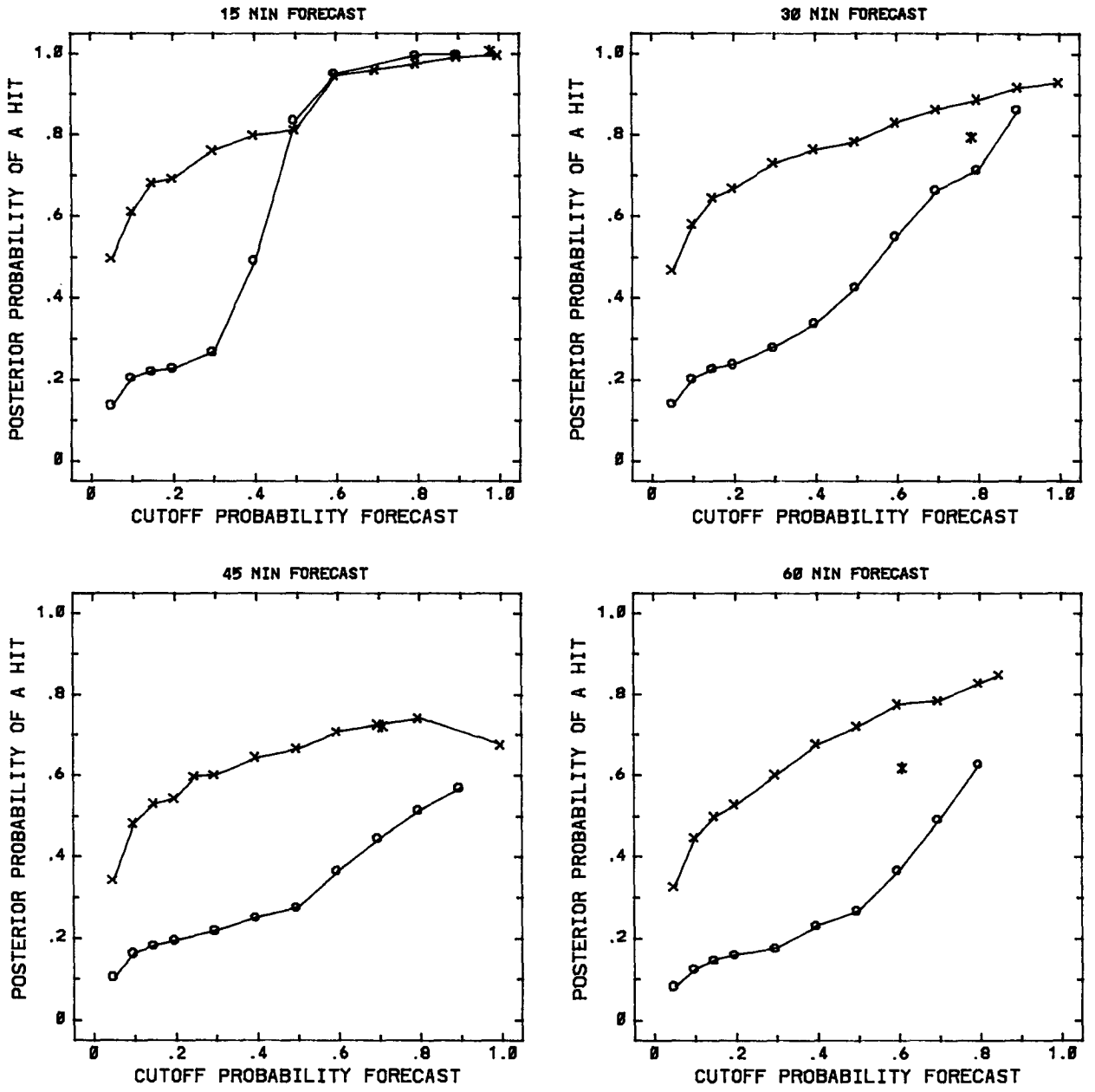

FIG. 12. Posterior hit probability of the weather event occurring as a function of cutoff probability forecast for high-activity ("×") and low-activity ("○") days for 15-, 30-, 45-, and 60-min forecast intervals. The solid lines are the predictions of the best-fitting Gaussian signal detection model. The asterisk represents the persistence forecast.

associated with low-frequency events, which are often difficult to estimate reliably from the observed data. In principle, however, the analyses below could be undertaken using observed data (Winkler and Murphy 1985).

An example illustrates how the consumer can maximize the expected value of a decision taken after receiving a weather forecast. The maximum-likelihood estimates of μ_s, σ_s, and the various χ_c decision criteria forming the boundaries between the response categories are given in Table 3 for the 30-min forecasts just discussed. With $A_z = .9629$, the accuracy of the forecasts is quite high.

For the consumer, who must take some action before severe weather occurs, there are four possible outcomes of a forecast: Action was taken, severe weather occurs (hit); action was taken, severe weather does not occur (false alarm); action was not taken, severe weather oc-

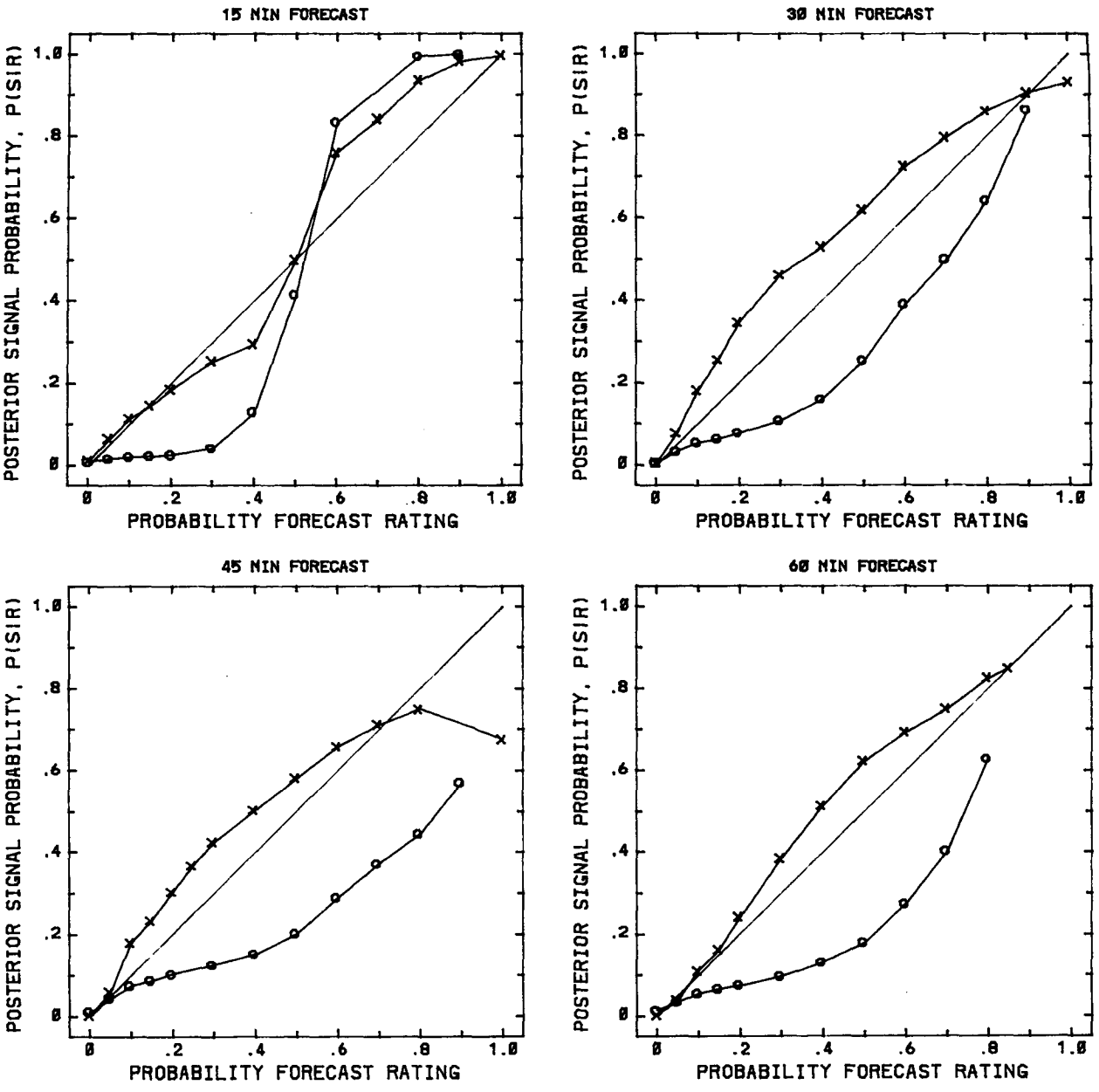

FIG. 13. Calibration curves for the 15-, 30-, 45-, and 50-min forecasts for high- ("×") and low-activity ("○") days. The posterior probability of the event given a particular response category is plotted as a function of the probability forecast rating used by the forecasters. The solid lines are the predictions of the best-fitting Gaussian signal detection model.

curs (miss); action was not taken, severe weather does not occur (correct rejection). If a cost or a benefit can be associated with each of the four outcomes, the consumer can use the information in Table 3 to decide when to take action in a way that will maximize the expected value of the decision. To compute the expected value, multiply the cost or benefit of each outcome by the probability of that outcome occurring, and then add the four values together (benefits are positive, costs are negative):

$$E(V) = p(Y|s)p(s)V_{s,Y} + p(Y|n)p(n)V_{n,Y} + p(N|s)p(s)V_{s,N} + p(N|n)p(n)V_{n,N}.$$

The four conditional probabilities needed in this equation may be computed for each decision criterion [see Harvey (1992)]. These, combined with the prior probabilities, $p(s)$ and $p(n)$, and the value of each outcome are all that are needed to compute the expected value.

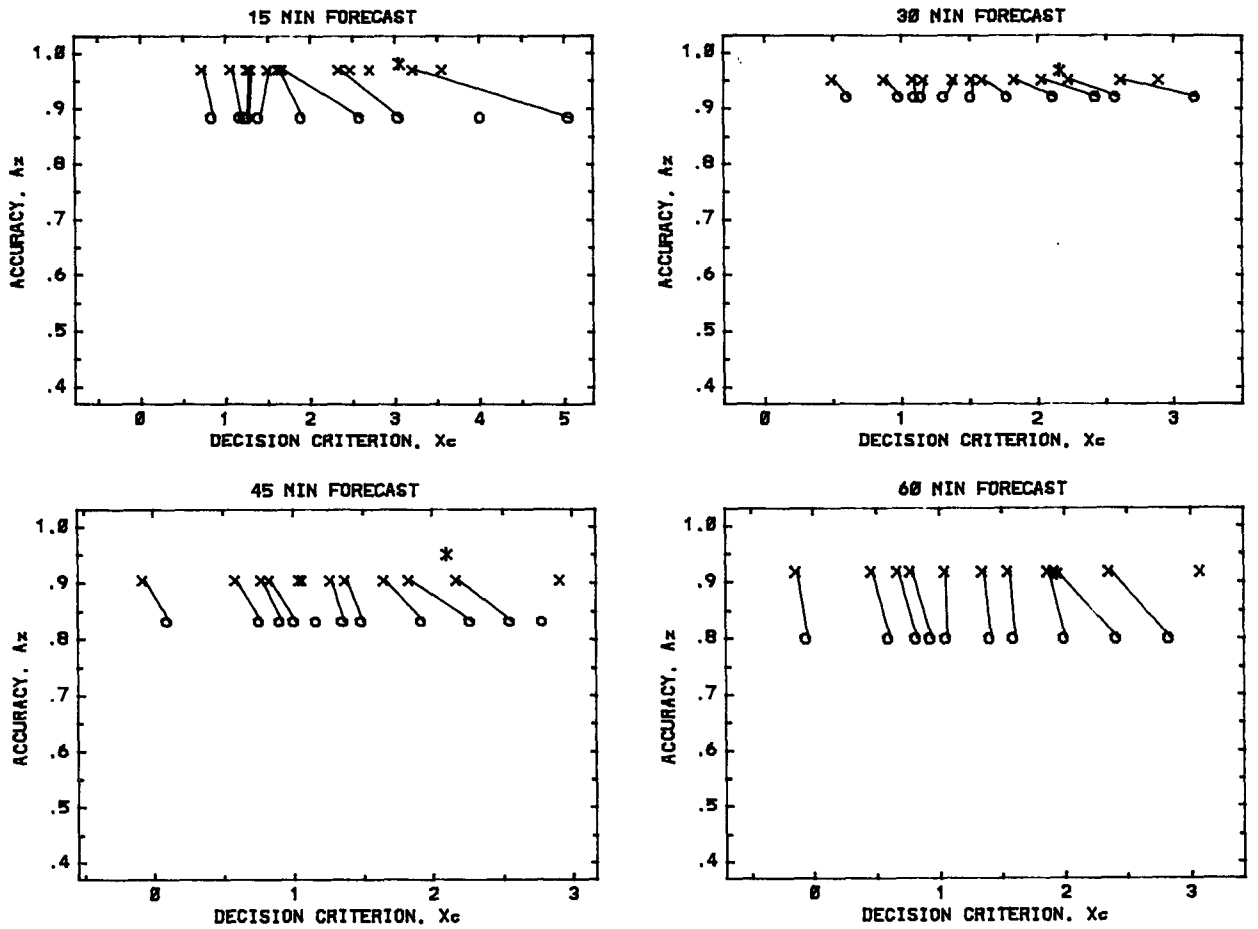

FIG. 14. Accuracy A_z as a function of decision criterion χ_c for high- ("×") and low-activity ("○") days for 15-, 30-, 45-, and 60-min forecasts. Criterion points that separate the same response categories are connected by solid lines. The asterisk represents the persistence forecast.

TABLE 3. The mean μ_s, standard deviation σ_s, and the decision criterion values χ_c of the best-fitting signal detection model for weather forecasters predicting the occurrence of severe weather 15–30 min from the time of the forecast. The prior probability of the severe weather was 0.098. Forecasts were given as probability values. Also included is the accuracy index A_z.

Judgment boundaries	Decision criteria χ_c	Expected value scenario 1	Expected value scenario 2
0.00|0.05	0.8198	114.2	92.4
0.05|0.10	1.1564	141.2	115.8
0.10|0.15	1.3041	149.3	121.3
0.15|0.20	1.3623	151.8	122.6
0.20|0.30	1.5457	157.5	123.6
0.30|0.40	1.6986	159.8	121.1
0.40|0.50	1.8413	160.2	116.4
0.50|0.60	2.0962	157.7	103.4
0.60|0.70	2.3171	153.3	89.5
0.70|0.80	2.5037	149.0	77.4
0.80|0.90	2.9106	140.3	55.0
0.90|1.00	3.2432	135.5	43.3
$\mu_s = 2.20$			
$\sigma_s = 0.72$			
$A_z = 0.26$			

Assume the following: The value of taking action before severe weather strikes ($V_{s,Y}$) is 100, the value of not taking action when severe weather does not occur ($V_{n,N}$) is 200, the value of taking action when severe weather does not occur ($V_{n,Y}$) is −200, the value of not taking action when severe weather strikes ($V_{s,N}$) is −500. The expected value associated with each χ_c for a prior probability of severe weather of .098 is shown in the second column of Table 3. The maximum expected value of 160.2 would be attained when the consumer took the action whenever the forecasters predicted severe weather with a probability of .5 or higher. Note that the expected value is determined by the value of χ_c and not by the probability forecast given by the forecaster, which may or may not be accurate in terms of the true posterior probability. Although it would be nice if the forecaster were perfectly calibrated and the forecasts could be taken at face value, this condition is not necessary.

Another consumer, having different costs and benefits from the one in the preceding example, can use

the same forecast to maximize his/her expected value. Consider a flight controller who must decide whether or not to change the runway of incoming flights, based on the forecast of severe weather. Assume the following costs and benefits: The value of changing the runway before severe weather strikes (thus avoiding a plane crash, but disrupting airline schedules) ($V_{s,Y}$) is -100, the value of not changing the runway when severe weather does not occur ($V_{n,N}$) is 200, the value of taking action when severe weather does not occur (thus disrupting airline schedules) ($V_{n,Y}$) is -200, the value of not taking action when severe weather strikes (thus causing a plane crash) ($V_{s,N}$) is -1500. The resulting expected values are listed in the third column of Table 3. The maximum expected value is 123.6 and would be achieved by changing the landing runway when severe weather is forecast with a predicted probability of .3 or higher.

11. Conclusions

The above analysis demonstrates that the Gaussian signal detection model provides an excellent basis for understanding forecasting behavior. The model accurately describes the relationship between the hit rates and the false-alarm rates (ROC) achieved under different decision criteria. It provides a criterion-free measure of accuracy A_z, the area under the ROC, that allows the potential performance of different forecasts and different forecasters to be meaningfully compared. It provides a clear way of defining performance in terms of the posterior hit probabilities achieved using different decision criteria. And it provides a way of evaluating how well the forecast probabilities issued by the forecasters correspond to the posterior probabilities corresponding to each response. It is compatible with the probability decomposition methods developed by Murphy (Murphy and Daan 1985; Murphy and Winkler 1987, 1992). The signal detection model leads to an understanding of the effect of base rate–related stress on forecasting behavior. It both increases forecasting accuracy and causes a lowering of decision criteria. Using the predictions of the signal detection model, the consumer of weather forecasts can make choices to implement policy.

Acknowledgments. We thank Allan Murphy for his challenging and critical comments that have greatly helped the authors make the revisions to the manuscript. Requests for reprints should be sent to Lewis O. Harvey, Jr., or Kenneth R. Hammond, Department of Psychology, Campus Box 345, University of Colorado, Boulder, Colorado 80309. Copies of the maximum likelihood computer program, RSCORE, written in Fortran 77, may best be obtained via e-mail request to the first author at lharvey@clipr.colorado.edu.

REFERENCES

Baird, J. C., and E. Noma, 1978: *Fundamentals of Scaling and Psychophysics.* Wiley, 287 pp.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78,** 1–3.

Clemen, R. T., and A. H. Murphy, 1986: Objective and subjective precipitation probability forecasts: Statistical analysis of some interrelationships. *Wea. Forecasting,* **1,** 56–65.

Cronbach, L. J., 1955: Processes affecting scores on "understanding of others" and "assumed similarity." *Psychol. Bull.,* **52,** 177–193.

Donaldson, R. J. Jr., R. M. Dyer, and M. J. Kraus, 1975: An objective evaluator of techniques for predicting severe weather events. Preprints, *Ninth Conf. on Severe Local Storms.* Boston, Amer. Meteor. Soc., 321–326.

Dorfman, D. D., and E. Alf, Jr., 1969: Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—rating method data. *J. Math. Psychol.,* **6,** 487–496.

——, L. L. Beavers, and C. Saslow, 1973: Estimation of signal detection theory parameters from rating-method data: A comparison of the method of scoring and direct search. *Bull. Psychon. Soc.,* **1,** 207–208.

Egan, J. P., 1975: *Signal Detection Theory and ROC Analysis.* Academic Press, 277 pp.

Fildes, R., and S. Makridakis, 1988: Forecasting and loss functions. *Int. J. Forecasting,* **4,** 545–550.

Forgionne, G. A., 1986: *Quantitative Decision Making.* Wadsworth, 901 pp.

Glahn, H. R., 1985: Statistical weather forecasting. *Probability, Statistics, and Decision Making in the Atmospheric Sciences,* A. H. Murphy and R. W. Katz, Eds., Westview Press, 289–335.

Graedel, T. E., and B. Kleiner, 1985: Exploratory analysis of atmospheric data. *Probability, Statistics, and Decision Making in the Atmospheric Sciences,* A. H. Murphy and R. W. Katz, Eds., Westview Press, 1–43.

Green, D. M., and J. A. Swets, 1974: *Signal Detection Theory and Psychophysics* (A reprint, with corrections of the original 1966 ed.). Robert E. Krieger Publishing Co., 479 pp.

Hammond, K. R., 1990: The effects of stress on judgment and decision making: An overview and arguments for a new approach. Tech. Rep. No. 320, 77 pp. [Center for Research on Judgment and Policy, Institute of Cognitive Science, Campus Box 345, University of Colorado, Boulder, Colorado 80309-0345.]

——, and L. Adelman, 1976: Science, values, and human judgment. *Science,* **194,** 389–396.

——, G. H. McClelland, and J. Mumpower, 1980: *Human Judgment and Decision Making: Theories, Methods, and Procedures.* Praeger, 258 pp.

Harvey, L. O., Jr., 1992: The critical operating characteristic and the evaluation of expert judgment. *Organizational Behavior and Human Decision Processes,* in press.

Hayes, W. L., 1973: *Statistics for the Social Sciences* (2d ed.). Holt, Rinehart and Winston, 954 pp.

Healy, A. F., and M. Kubovy, 1978: The effects of payoffs and prior probabilities on indices of performance and cutoff location in recognition memory. *Memory and Cognition,* **6,** 544–553.

Hsu, W., and A. H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting,* **2,** 285–293.

Krantz, D. H., 1969: Threshold theories of signal detection. *Psychological Review,* **76,** 308–324.

Macmillan, N. A., and C. D. Creelman, 1991: *Detection Theory: A User's Guide.* Cambridge University Press, 407 pp.

Mason, I. B., 1982: A model for assessment of weather forecasts. *Aust. Meteorol. Mag.,* **30,** 291–303.

Mueller, C. K., J. W. Wilson, and B. Heckman, 1987: Evaluation of the TDWR aviation nowcasting experiment. *Third Int. Conf.*

on the Aviation Weather System. Boston, Amer. Meteor. Soc., 212–216.

Murphy, A. H., 1973: A new partition of the probability score. *J. Appl. Meteor.*, **12,** 595–600.

——, 1977: The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Monthly Weather Review,* **105,** 803–816.

——, 1985: Probabilistic weather forecasting. *Probability, Statistics, and Decision Making in the Atmospheric Sciences,* A. H. Murphy and R. W. Katz, Eds., Westview Press, 337–377.

——, 1986: A new decomposition of the Brier score: Formulation and interpretation. *Mon. Wea. Rev.*, **114,** 2671–2673.

——, 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, 1985: Forecast evaluation. *Probability, Statistics, and Decision Making in the Atmospheric Sciences,* A. H. Murphy and R. W. Katz, Eds., Westview Press, 379–437.

——, and R. W. Katz, Eds., 1985: *Probability, Statistics, and Decision Making in the Atmospheric Sciences.* Westview Press, 547 pp.

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

——, and ——, 1992: Diagnostic verification of probability forecasts. *Int. J. Forecasting,* **6,** in press.

——, Y.-S. Chen, and R. T. Clemen, 1988: Statistical analysis of interrelationships between objective and subjective temperature forecasts. *Mon. Wea. Rev.*, **116,** 2121–2131.

Swets, J. A., 1961: Is there a sensory threshold? *Science,* **134,** 168–177.

——, 1986a: Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychol. Bull.*, **99,** 181–198.

——, 1986b: Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychol. Bull.*, **99,** 100–117.

——, 1988: Measuring the accuracy of diagnostic systems. *Science,* **240,** 1285–1293.

——, and R. M. Picket, 1982: *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory.* Academic Press, 253 pp.

——, W. P. Tanner, Jr., and T. G. Birdsall, 1961: Decision processes in perception. *Psychol. Rev.*, **68,** 301–340.

Tanner, W. P., Jr., and J. A. Swets, 1954: A decision-making theory of visual detection. *Psychol. Rev.*, **61,** 401–409.

Thompson, J. C., 1952: On the operational deficiencies in categorical weather forecasts. *Bull. Amer. Meteor. Soc.*, **33,** 223–226.

——, and G. W. Brier, 1955: The economic utility of weather forecasts. *Mon. Wea. Rev.*, **83,** 249–254.

Thompson, P. A., 1990: An MSE statistic for comparing forecast accuracy across series. *Int. J. Forecasting,* **6,** 219–227.

Thurstone, L. L., 1927: A law of comparative judgment. *Psychol. Rev.*, **34,** 273–286.

Winkler, R. L., and A. H. Murphy, 1985: Decision analysis. *Probability, Statistics, and Decision Making in the Atmospheric Sciences,* A. H. Murphy and R. W. Katz, Eds., Westview Press, 493–524.

Yates, J. F., and S. P. Curley, 1986: Conditional distribution analyses of probabilistic forecasts. *J. Forecasting,* **4,** 61–73.