

A model for assessment of weather forecasts

I. Mason, Regional Office, Bureau of Meteorology, ACT

(Manuscript received April 1982; revised August 1982)

A general paradigm for assessment of ability to discriminate between two alternatives is described in the context of weather forecast verification. The paradigm is based on the relative operating characteristic (ROC), a graph of the variation of hit rate with false alarm rate as decision criterion changes. A model for the ROC based on the mathematical theory of signal detection is shown to provide a good fit to verification data from weather forecasts for a wide variety of predictands. A basis is thus provided for the use of some indices of forecast quality derived from the model. These indices are relatively independent of calibration (i.e. the correspondence between estimated probability and relative frequency) and can be evaluated for forecasts expressed in yes/no form, as ratings of risk (e.g. low, moderate, high) or explicitly as probabilities, facilitating direct comparison of these different types of forecast.

Introduction

Meteorologists have given much attention to assessment of the quality of weather forecasts and a wide variety of procedures have been used to this end, corresponding with the various purposes for which assessment is required and the variety of formats in which forecasts are issued. Some recent work includes that of Murphy in Murphy and Williamson (eds) (1976) on probabilistic forecasts, Woodcock (1976, 1980) on yes/no forecasts and of Gulezian (1981) and Colls, Mason and Daw (1981) on routine weather forecasts, among many.

This variety of practices creates difficulties when it is desired to compare forecasts issued in different formats. For example, to compare probabilistic with yes/no forecasts it is necessary to reduce the probabilistic forecasts to yes/no form, usually by selection of a 'cut-off' probability which maximises one of the many yes/no scores (e.g. Bryan and Enger 1967; Mason 1979). This is unsatisfactory because all scores for yes/no forecasts confound accuracy with decision criterion, so that variations in a score may not be related to variations in the skill of the methods used to produce the forecasts (Mason 1982). Also, reducing probabilities to zeros and ones loses much information, and the resulting set of yes/no forecasts will be sub-optimal for most users.

Methods for the assessment of purely probabilistic forecasts are well developed. Murphy's (1973) three-component partition of the probability score appears to be the method of choice at present, providing three separate measures for variability in the data, resolution, and reliability respectively.

Resolution here refers to the ability shown by the forecaster (or forecasting method) to discriminate between situations that will be followed by the predictand and those that will not.

Reliability is the correspondence between forecast

probabilities and observed relative frequencies. High reliability is quite compatible with low resolution, for example in a forecast set consisting only of predictions of the climatological probability on every occasion, and high resolution can be achieved with low reliability. The term calibration will sometimes be used as a synonym for reliability in this note.

Yates (1982) has recently described another method of partitioning the probability score.

The Brier score, and in fact all currently available scoring rules for probabilistic forecasts, can only be evaluated when the forecasts are expressed as numerical probabilities. Forecasts given as risk-ratings (for example low, moderate or high risk for some event) cannot be assessed using these scores unless numerical probabilities are assigned to the ratings.

Comparisons with verbal forecasts that include 'chance of . . .' statements as well as yes/no predictions are further complicated by the lack of quantitative definition of the probability range corresponding to 'chance of'. There is clearly a need for a measure of forecast quality that can be evaluated for all these types of forecast.

A situation with some formal similarities to that of forecast assessment has been studied in the psychological theory of signal detection. The process of forecasting a discrete meteorological event is in some respects analogous to that of detection of a signal against a background of noise. In both cases the task is essentially to assign a conditional probability to some defined event on the basis of data which is insufficient to provide certainty. The outcome of a series of trials may be represented in both cases by formally identical verification arrays.

From the point of view of weather forecast

verification, signal detection theory (SDT) contains two features that may be useful. One is a very general paradigm for the assessment of the quality of predictions. This paradigm is exemplified by the relative (or receiver) operating characteristic (ROC), a graphical display of the relation between hit and false alarm rates as decision criterion varies. It has been applied successfully to the evaluation of performance in fields as diverse as clinical diagnosis (Swets and Pickett 1982), vigilance (Broadbent and Gregory 1963), information retrieval (Swets 1979), and the study of conditioned responses in pigeons (McCarthy and Davison 1980), among others.

The second interesting feature of SDT is a model which describes the relative operating characteristic in terms of the parameters of hypothetical probability distributions, and which forms the basis for several indices of performance.

The purpose of this paper is, firstly, to show that the ROC paradigm can be applied to the assessment of forecast quality and that it provides an informative way of presenting this kind of data. Secondly, it will be shown that the SDT model fits weather forecast data quite closely, and hence that the use of SDT-based indices to describe forecast quality is valid. These indices can be evaluated, subject to some weak constraints, for any set of forecasts for a dichotomous predictand, whether given as numerical probabilities, risk ratings, yes/no forecasts, or verbally as in routine public weather forecasts.

The structure of the paper is, firstly, an outline of the method for assessment of performance based on the ROC, and of the signal detection theory model for the ROC. Then, ROCs are presented for a variety of predictands together with model-based ROCs for each case. Some discussion and conclusions follow.

The weather forecast as a statistical decision: a model based on signal detection theory

In this section the weather forecaster is considered as a decision-maker whose task is to decide whether to forecast occurrence or non-occurrence of some meteorological event. For the purpose of this paper there are supposed to be only these two possibilities. Extension to predictands that may have more than two values is possible, by considering the final decision as the result of a sequence of yes/no decisions, so the simplicity of this situation does not make it too restrictive.

The data on which the forecaster bases his decision is the usual multivariate vector of values for weather-related variables that all forecasters are presented with during the day's work (although most would not think of it in precisely this way). It is hypothesised that the implications of this data for prediction of some particular event (rain, thunderstorm, tornado, etc.) can be summarised as

a single number, perhaps, but not necessarily, a probability.

The decision whether or not to predict the event is based on a comparison of this number with a 'decision criterion' which is predetermined. The analogy is drawn with statistical hypothesis testing, in which a value of a test variable (z , t , χ^2 , etc.) is compared with 95 or 99 per cent values of these variables, in order to decide whether to accept or reject the hypothesis.

This section of the paper falls into three parts. Firstly the notion of a decision criterion is elaborated then the use of a variable decision criterion to generate the ROC for a set of probabilistic forecasts is described. Thirdly, the 'normal-normal' model from signal detection theory is introduced as a descriptive model for the ROC, and some indices of forecasting performance based on this model are given. Signal detection theory has an extensive literature, and the presentation in this paper refers only to those parts that are directly relevant to weather forecast assessment. A useful entry to the field is Swets' review (1973). Detailed presentations can be found in texts by Green and Swets (1974), Egan (1975) or Swets and Pickett (1982).

The decision criterion

The minimum components of a decision-making situation are:

- (i) two decision alternatives, for example between forecasts of occurrence or non-occurrence of the event, or assertion that noise is present alone or there is a signal as well as noise;
- (ii) two possible events, for example rain or no rain, temperature above or below zero, signal present or absent;
- (iii) information available to the decision-maker about these events; and
- (iv) a decision criterion, some specific value x_c of a decision variable X .

X is a scalar whose value depends on the available data, and may be given as the conditional probability of the event given current data, or in terms of the likelihood ratio:

$$\frac{\Pr\{\text{current observations event occurs}\}}{\Pr\{\text{current observations event does not occur}\}}$$

or in terms of any monotonic function of likelihood ratio, for example the log likelihood ratio or log odds. X can also be thought of as analogous to a discriminant function, or quite generally just as a function of the data that provides information about the event of interest.

The decision is then supposed to be made on the basis of the critical value x_c of the decision variable. For $x \geq x_c$ one decision is required and for $x < x_c$ the other. Extension to the case of probabilistic forecasts is done by a partition of the range of values of X using a set of critical values $\{x_i\}$, $i = 1$ to $M-1$, where M is the number of discrete values permitted

Fig. 1



Fig. 3



perhaps, but not necessarily, a
 whether or not to predict the event is
 comparison of this number with a
 which is predetermined. The
 with statistical hypothesis testing,
 of a test variable (z , t , χ^2 , etc.) is
 or 99 per cent values of these
 to decide whether to accept or
 reject.

the paper falls into three parts.
 of a decision criterion is
 the use of a variable decision
 to generate the ROC for a set of
 forecasts is described. Thirdly, the
 model from signal detection
 is used as a descriptive model for the
 analysis of forecasting performance
 models are given. Signal detection
 theory and extensive literature, and the
 paper refers only to those parts
 relevant to weather forecast
 performance. A useful entry to the field is Swets'
 (1982).

Decision

components of a decision-making
 process are alternatives, for example
 forecasts of occurrence or non-
 occurrence of the event, or assertion that
 the event alone or there is a signal as
 opposed to no signal; events, for example rain or no
 rain; a threshold above or below zero, signal
 level; and a decision criterion available to the decision-maker
 based on the available data; and a decision criterion, some specific value x_c of
 the decision variable X . The value of x_c depends on the available
 information given as the conditional probability of the event given current data, or in
 terms of the odds ratio:

$$\frac{\text{Pr(event occurs)}}{\text{Pr(event does not occur)}}$$

is a monotonic function of likelihood
 ratio. The log likelihood ratio or log
 odds ratio is thought of as analogous to a
 signal, or quite generally just as a
 quantity that provides information about

the event supposed to be made on the
 basis of the value x_c of the decision variable.
 A decision is required and for $x < x_c$,
 the event is predicted to the case of probabilistic
 partitioning of the range of values
 into critical values $\{x_i\}$, $i = 1$ to $M-1$,
 where M is the number of discrete values permitted

Fig. 1 Relative operating characteristic generated by a set of 341 estimates of the probability of rain in an area around Canberra.

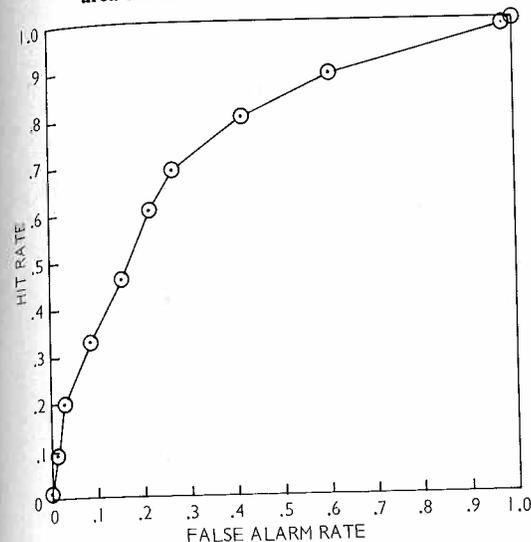
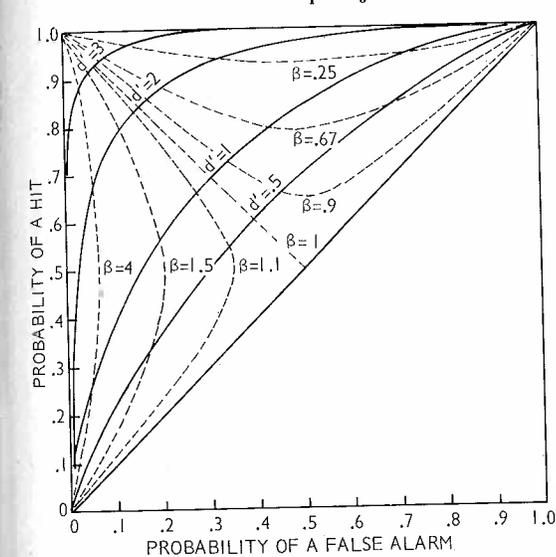


Fig. 3 Relative operating characteristics generated by Gaussian distributions with equal variance for $d' = 0.5, 1.0, 2.0$ and 3.0 . Broken lines are isopleths of the likelihood ratio $\beta = f_1(x)/f_0(x)$.



for probabilities. If $\{p_i\}$, $i = 1$ to M , is the set
 of permitted probabilities then the particular value p_k
 is used when x is in the half-open interval $[x_{k-1}, x_k)$.
 Choice of the cut-off values $\{x_k\}$ is clearly of some
 importance in practice, but is not directly relevant to
 this note; some discussion of this aspect can be
 found in Green and Swets (1974).

Returning for the present to the case of just two
 decision alternatives, performance in a series of N
 cases can be represented as a 2×2 array, the
 verification matrix at Table 1.

There are obviously two different ways to be right
 and also two ways to be wrong. Correct forecasts

Fig. 2 Probability distributions for the decision variable X preceding occurrence of the predictand, $f_1(x)$, and preceding non-occurrence, $f_0(x)$. x_c is the decision criterion. The diagonally hatched area is equal to the probability of a hit, and vertical hatching indicates the probability of a false alarm. The separation of the means, d' , is the fundamental signal detection index of discrimination.

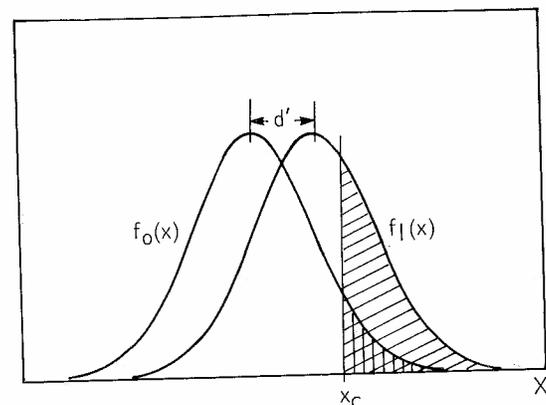
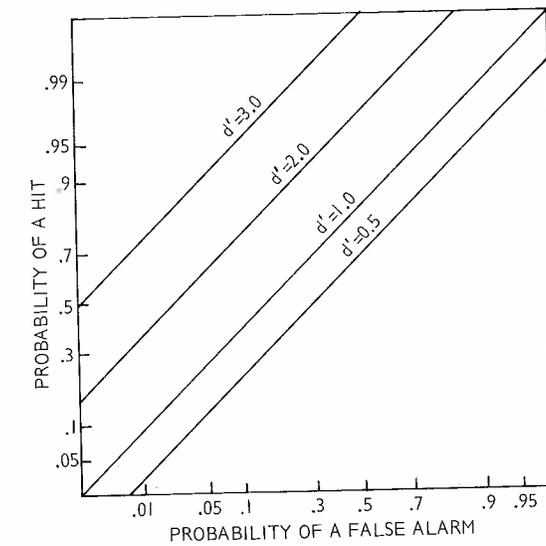


Fig. 4 Relative operating characteristics for the Gaussian equal variance case with $d' = 0.5, 1.0, 2.0$ and 3.0 , on double-probability axes.



may be 'hits', identified by d in Table 1, or 'correct
 negatives', a in Table 1. Wrong forecasts may be
 'false alarms', c , or 'misses', b .

It is convenient to describe the quality of the
 forecasts represented in Table 1 in terms of two
 parameters. These are hit rate and false alarm rate,
 defined as follows:

$$\text{hit rate, } h = \frac{d}{b+d} \dots 1$$

$$\text{false alarm rate, } f = \frac{c}{a+c} \dots 2$$

Hit rate defined in this way is equal to the quantity

Fig. 5(a) Probability densities for X for three different values of the variance ratio $s = \sigma_0/\sigma_1$. The mean of $f_0(x)$ is one standard deviation from that of $f_1(x)$ in all three cases.

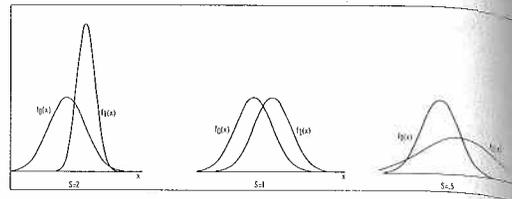


Fig. 5(b) Relative operating characteristics generated by the three pairs of distributions shown in Fig. 4, on linear axes.

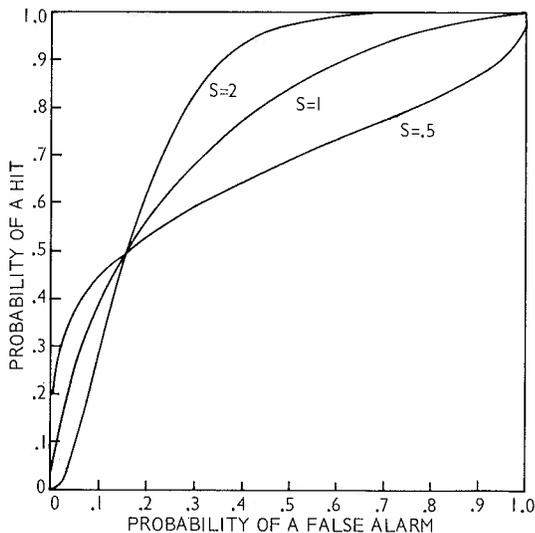


Fig. 5(c) Relative operating characteristics generated by the three pairs of distributions shown in Fig. 4, on double probability axes.

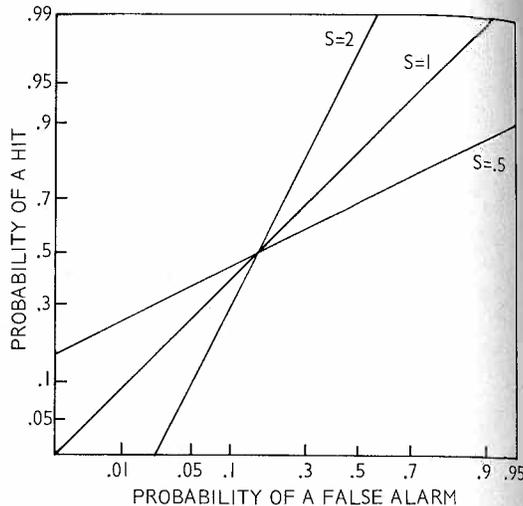


Fig. 6 Relative operating characteristics generated by a set of 341 estimates of the probability of rain in an area around Canberra. Scales linear in the standard normal deviate are superimposed.

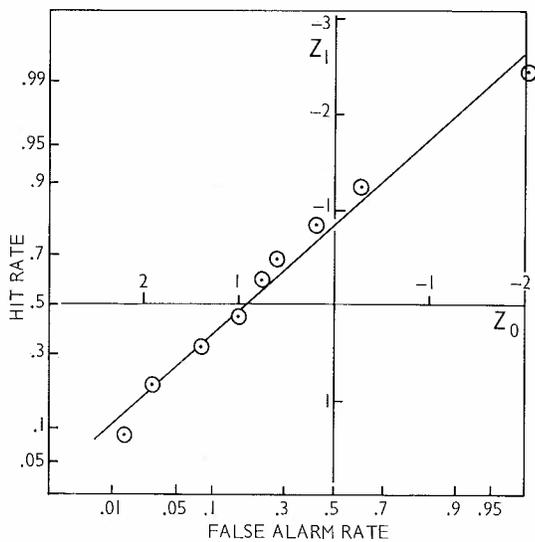
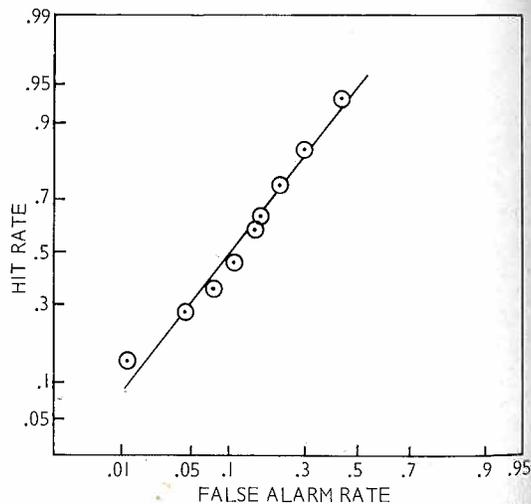


Fig. 7 ROC for estimates of probability of occurrence of measurable precipitation at Canberra Airport during the 12 hours 9.00 pm to 9.00 am local time. N = 335.



Mason: A n

Fig 8 RC 2.5 ho

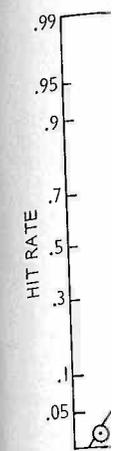
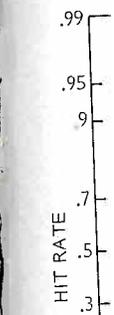
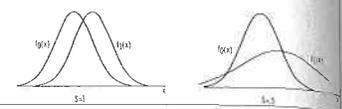


Fig. 10 1

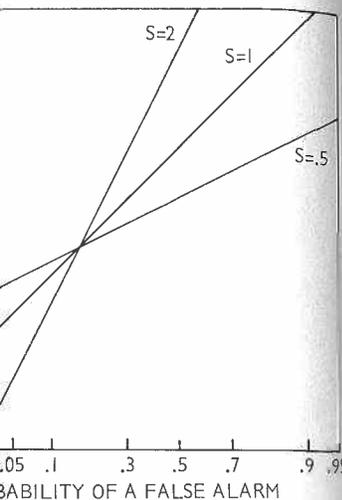


Table

DEC



operating characteristics generated by the pairs of distributions shown in Fig. 4, the probability axes.



estimates of probability of occurrence of precipitation at Canberra Airport 12 hours 9.00 pm to 9.00 am local time.

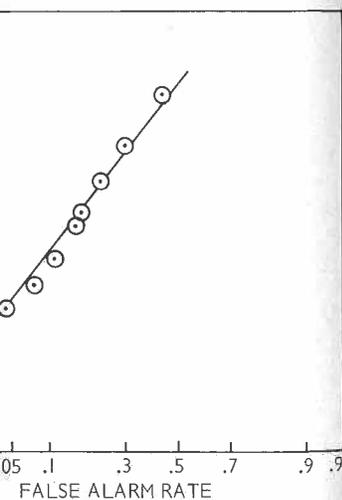


Fig 8 ROC for estimates of the probability of more than 2.5 mm of rain at Canberra Airport during the 12 hours 9.00 pm to 9.00 am local time. N = 335.

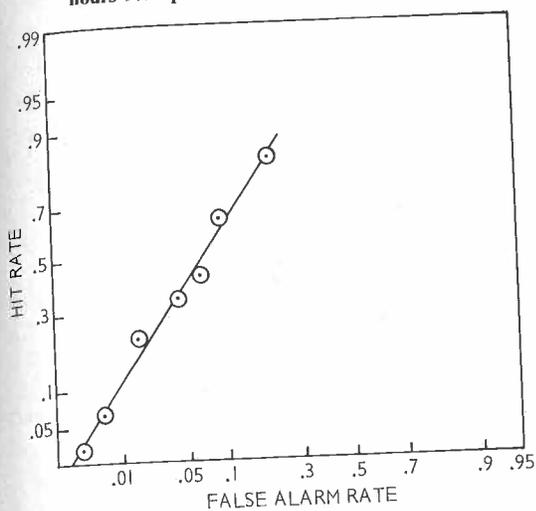


Fig. 10 ROC for individual forecaster D in Canberra, a sub-set of the data that produced Fig. 5. N = 83.

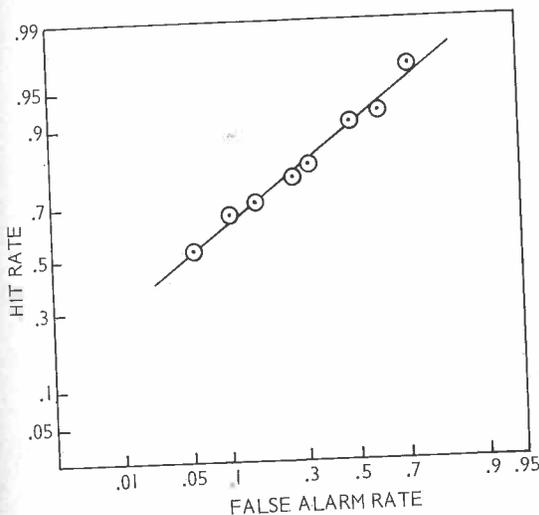


Fig. 9 ROC for individual forecaster C in Canberra, a sub-set of the data that produced Fig. 5. N = 96.

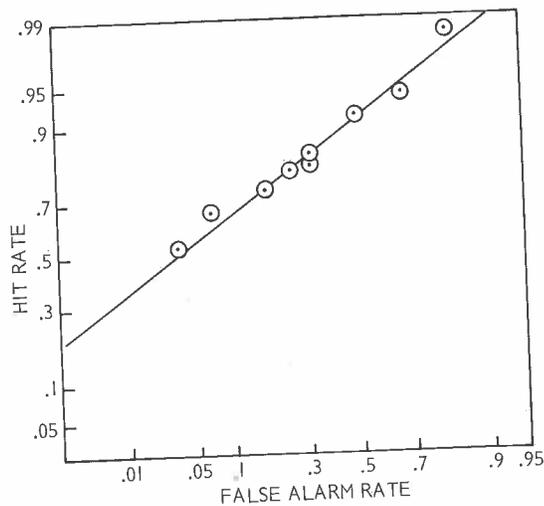


Fig. 11 ROC for estimates of the probability of rain near Great Falls, Mont., USA from data published by Murphy and Winkler (1974). N = 2646.

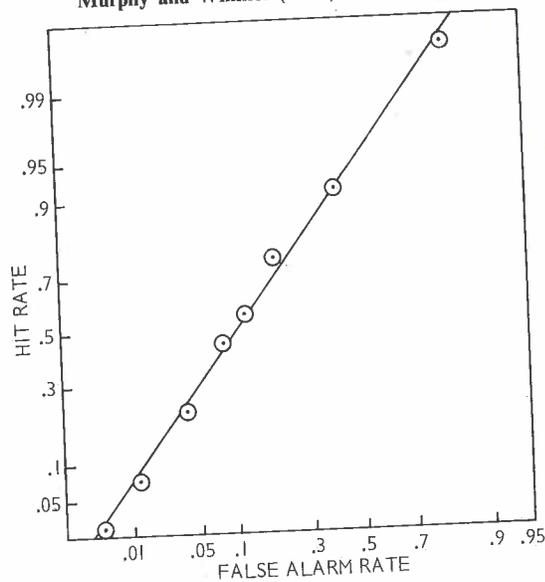


Table 1. Verification matrix for a series of N yes/no forecasts. a, b, c and d are the total frequencies of each possible combination of forecast and event.

		EVENT		
		does not occur	occurs	total
DECISION	forecast non-occurrence	a	b	a + b
	forecast occurrence	c	d	c + d
Total		a + c	b + d	N

Fig. 12 ROC for estimates of the probability of rain near Seattle, Wash., USA, from data published by Murphy and Winkler (1977). N = 948.

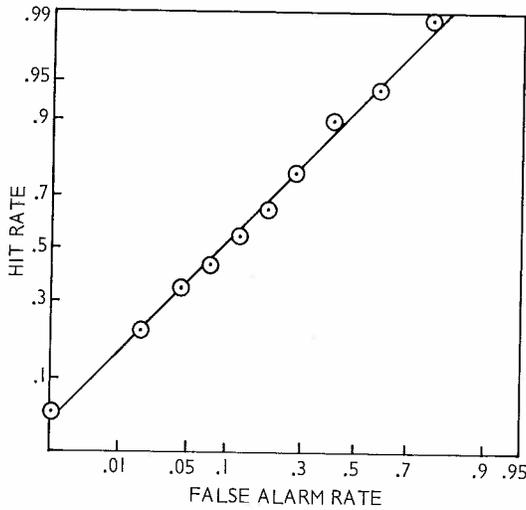


Fig. 13 ROC for US National Weather Services estimates of the probability of rain at Chicago, Illinois from data published by Murphy and Winkler (1977). N = 17 154.

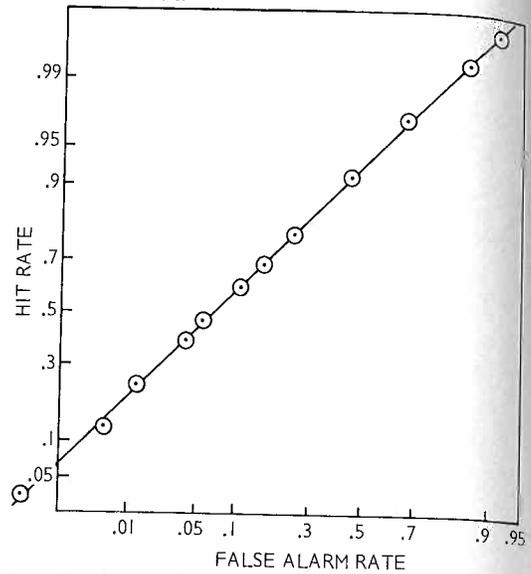


Fig. 14 ROC for individual forecaster A at Chicago. A sub-set of the data in Fig. 12. N = 2916.

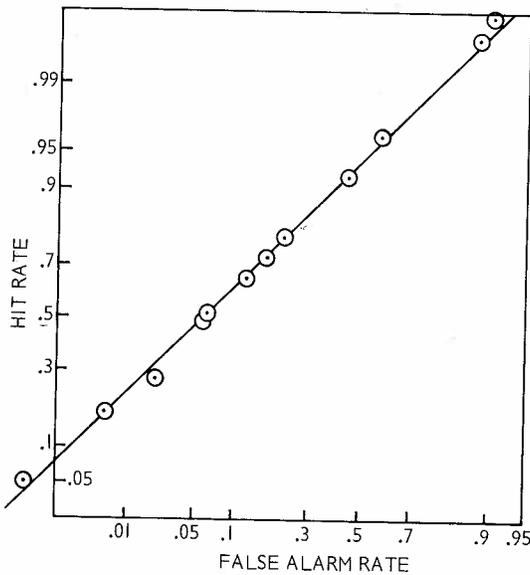
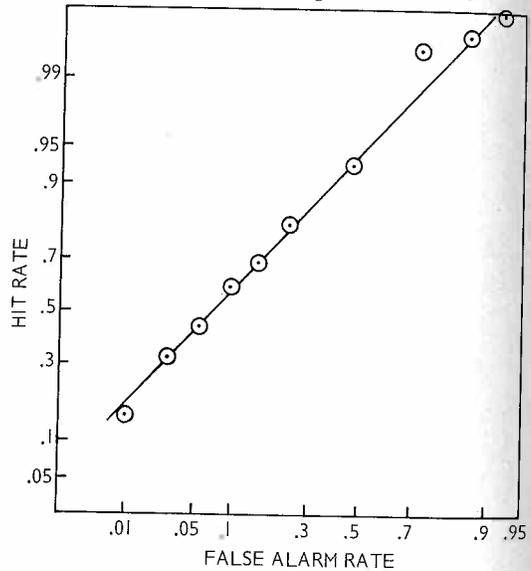


Fig. 15 ROC for individual forecaster B at Chicago. A sub-set of the data in Fig. 12. N = 2820.



'prefiguration on Yes forecasts' suggested by Olson (1965), and Olson's 'prefiguration on No forecasts' is equal to 1 - false alarm rate.

The relative operating characteristic

Given a set of probabilistic forecasts, a sequence of verification matrices of the form of Table 1, and hence a sequence of hit rate, false alarm rate pairs, can be generated by stepping a decision probability p_c through the range of values used in the forecasts. The graph of hit rate against false alarm rate as decision criterion varies is called the relative (or receiver) operating characteristic (ROC). Reasons

for this terminology can be found in Swets' 1973 review.

Table 2 shows an example of this procedure for a set of probability forecasts of rain over Canberra city or suburbs. Hit rates and false alarm rates are shown in rows (f) and (g), and Fig. 1 is the ROC for this set of forecasts.

The ROC shows how hit rate may be 'traded-off' against false alarm rate by varying the decision probability. Perfect performance is represented on ROC axes by the upper left-hand corner, $f = \theta$, $h = 1$, and perfectly wrong forecasts by the opposite point, $f = 1$, $h = \theta$. A ROC lying along the major

ason: A model fo

Fig. 16 ROC for temperature Murphy and

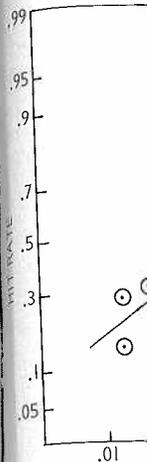


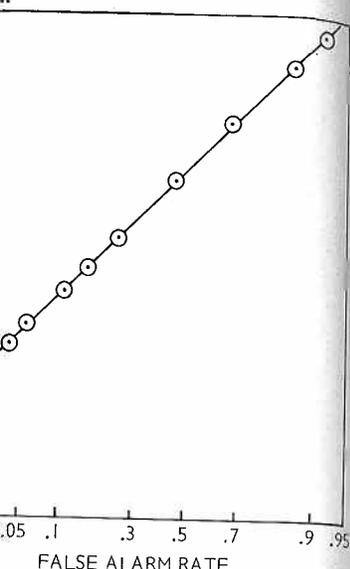
Fig. 18 ROC for minimum 28°F ne publishe



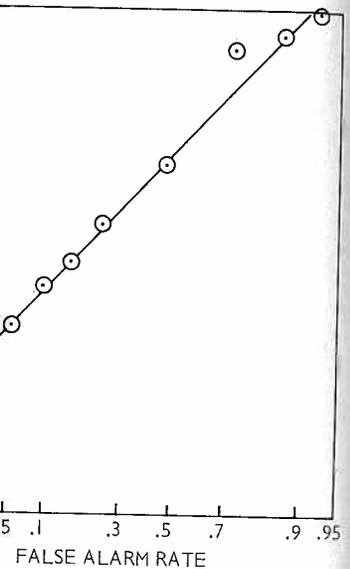
diagonal fro forecasting, likely to pre than it is to forecast sets this sense wi and better fo left corner.

The econ cost/loss m and Brier 1 hit and fals appropriate Table 2 s correspond: rate is 0.8

US National Weather Services estimates probability of rain at Chicago, Illinois from data published by Murphy and Winkler (1977). N = 122.



individual forecaster B at Chicago. A the data in Fig. 12. N = 2820.



can be found in Swets' 1973

An example of this procedure for a forecasts of rain over Canberra hit rates and false alarm rates are and (g), and Fig. 1 is the ROC for is. how hit rate may be 'traded-off' n rate by varying the decision t performance is represented on upper left-hand corner, $f = \theta$, $h =$ wrong forecasts by the opposite θ . A ROC lying along the major

Fig. 16 ROC for 5F 'fixed width credible interval' temperature probabilities from data published by Murphy and Winkler (1974). N = 122.

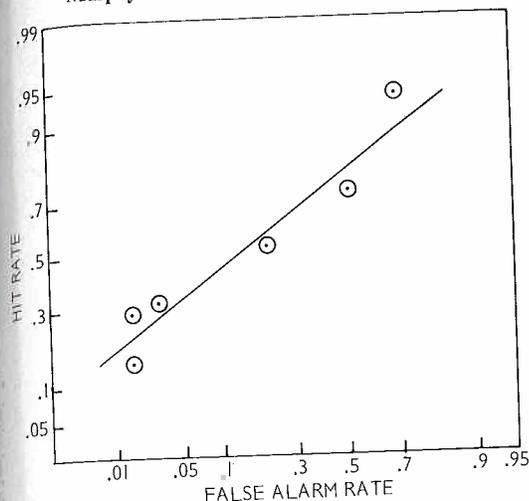
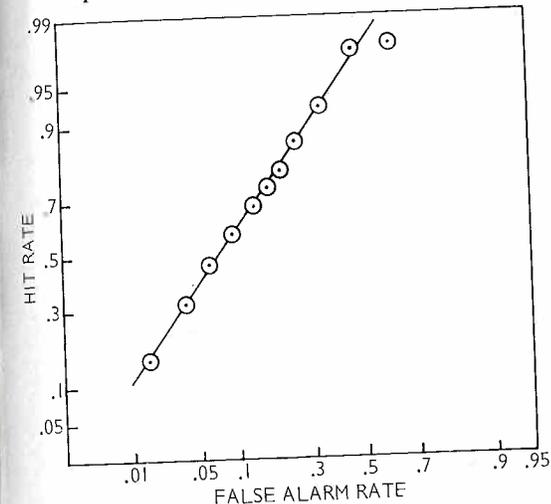


Fig. 18 ROC for estimates of the probability that minimum temperature will be less than or equal to 28°F near Albuquerque, New Mexico, from data published by Murphy (1977). N = 1443.



diagonal from 0, 0 to 1, 1, represents random forecasting, in which a forecast of 'yes' is no more likely to precede an occurrence of the predictand than it is to precede a non-occurrence. Hence all forecast sets with some positive skill over chance in this sense will have ROCs in the upper left triangle, and better forecasts have ROCs nearer to the upper left corner.

The economic value of the forecasts in the cost/loss meteorological decision model (Thompson and Brier 1955) is in large measure determined by hit and false alarm rates (Mason 1980). Hence an appropriate choice of p_c is economically important. Table 2 shows that if the decision criterion corresponds to a probability of say, 0.3 then the hit rate is 0.8 and the false alarm rate 0.42. It may be

Fig. 17 ROC for 9F 'fixed width credible interval' temperature probabilities, from data published by Murphy and Winkler (1974). N = 122.

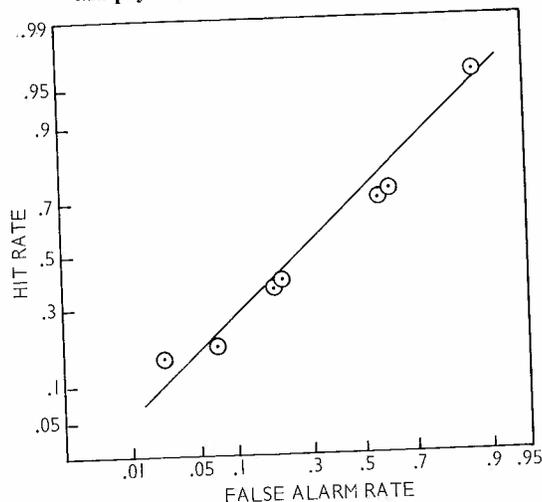
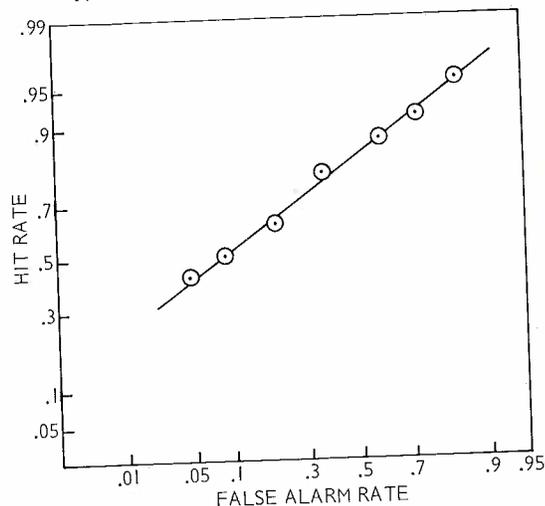


Fig. 19 ROC for estimates of the probability that one or more tornadoes will occur in the severe weather areas delineated by the US National Severe Storms Forecast Centre, Kansas City, Missouri. Data published by Murphy and Winkler 1977(b). N = 92.



that a false alarm rate of 0.42 is too large, for example if the cost of precautions against rain (in this case) is high relative to the protectable loss if rain occurs unforecast. Table 2 and Fig. 5 show that the false alarm rate can be reduced, but only at the expense of a reduction in the hit rate. If the false alarm rate has to be less than say, 5 per cent, then the forecasts shown above would achieve this level with a decision probability of about 0.77, and the corresponding hit rate would be 0.25. The ROC can be used in this way to determine any two of the three variables p_c , hit rate, and false alarm rate given one.

Note that the numerical values of probability, row (a) in Table 2, are not used in the calculation of hit

Fig. 20 ROC for estimates of the probability that there will be ten or more tornadoes anywhere in the USA on a given day. Data published by Murphy and Winkler 1977(b). N = 92.

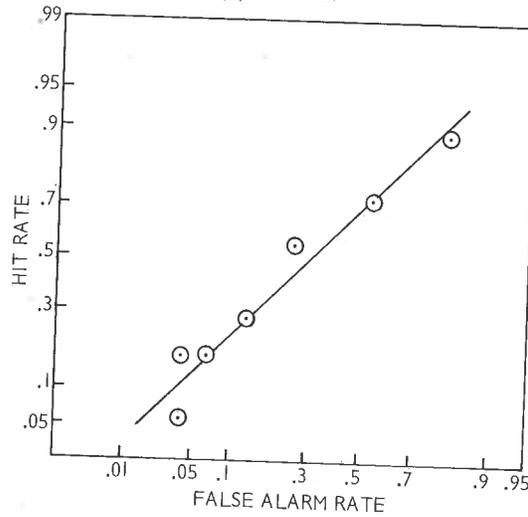


Fig. 21 ROC for lightning risk forecasts for the Australian Capital Territory. N = 813.

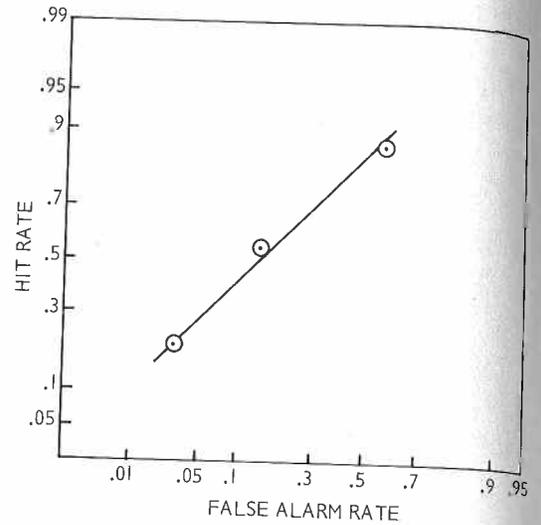


Table 2. Calculation of hit rate and false alarm rate for a set of probabilistic forecasts for rain in Canberra.

(a)	0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
(b)	1	14	13	16	11	21	18	18	16	11	1
(c)	4	75	37	30	10	12	16	11	3	3	0
(d)	140	139	125	112	96	85	64	46	28	12	1
(e)	201	197	122	85	55	45	33	17	6	3	0
(f)	1.0	.99	.89	.80	.69	.61	.46	.33	.17	.06	.03
(g)	1.0	.98	.61	.42	.27	.22	.16	.09	.03	.015	0

- (a) Estimated probability, p.
- (b) Number of occurrences following estimated probability given.
- (c) Number of non-occurrences following estimated probability given.
- (d) Accumulated occurrences following probability $\geq p$.
- (e) Accumulated non-occurrences following probability $\geq p$.
- (f) Hit rate.
- (g) False alarm rate.

and false alarm rates; their sole use is to locate the decision criterion. Hence analyses involving only hit and false alarm rates will be relatively unaffected by the calibration of the probabilities, except insofar as their rank order is important.

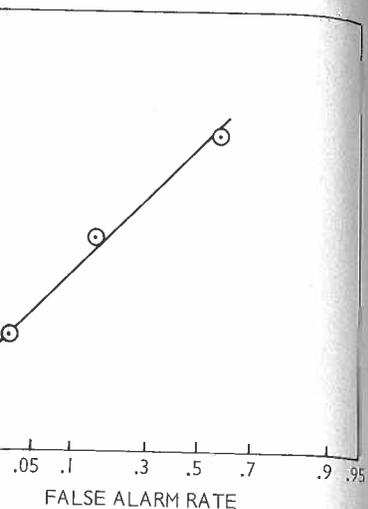
An index of performance that suggests itself from the fact that ROCs nearer to the 0,1 corner represent superior performance, is the area beneath the ROC. It can be calculated by joining the points with straight lines and summing areas of the resulting trapezoids. This index is referred to as P(A) by Green and Swets (1974). The range of P(A) is from 0.5 for random forecasting to 1.0 for perfect forecasts. It is non-parametric in that it does not depend on the assumptions about underlying probability distributions to be introduced in the next part of this section. However it does depend to some extent on the number of points on the ROC, that is on the number of discrete values allowed to the forecasters for their probability estimates, and gives

no indication at all of the shape of the ROC. The parametric indices suggested below are for these reasons more satisfactory.

The normal-normal model for the ROC

It is assumed that the observation variable X has a specific and known probability distribution on occasions preceding occurrence of the event, denoted $f_1(x)$, and a different distribution $f_0(x)$ preceding non-occurrence. Figure 2 represents these distributions. Their parameters, specifically the difference between their mean values, and their variances, characterise the quality of the information about the event. If the distributions are identical then the observational data X provides no information. If they are very far apart then an observation of a value of X determines occurrence or non-occurrence of the event with high probability. (In signal detection theory $f_0(x)$ is the distribution of X when noise alone is present and $f_1(x)$ when the signal is present in addition to noise.)

lightning risk forecasts for the Australian Territory. N = 813.



istic forecasts for rain in Canberra.

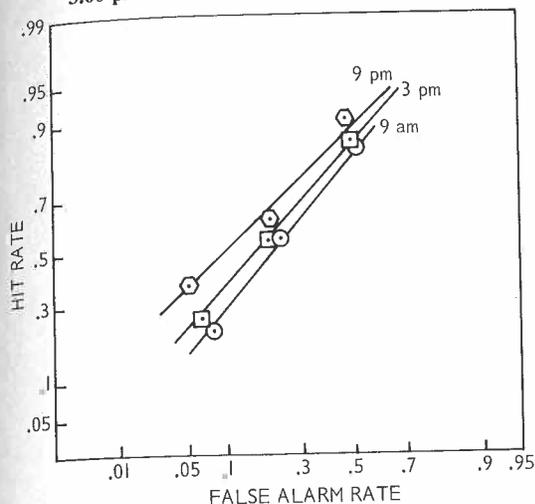
.7	.8	.9	1.0
18	16	11	1
11	3	3	0
46	28	12	1
17	6	3	0
.33	.20	.09	.01
.09	.03	.015	0

all of the shape of the ROC. The es suggested below are for these isatisfactory.

mal model for the ROC

t the observation variable X has a own probability distribution on ding occurrence of the event, and a different distribution $f_0(x)$ occurrence. Figure 2 represents these their parameters, specifically the en their mean values, and their racterise the quality of the t the event. If the distributions are observational data X provides no hey are very far apart then an value of X determines occurrence nce of the event with high gnal detection theory $f_0(x)$ is the when noise alone is present and ial is present in addition to noise.)

Fig. 22 ROCs for fog risk forecasts for Canberra Airport. \odot Issued 9.00 am (local time), N = 332; \square Issued 3.00 pm. N = 356; \ominus Issued 9.00 pm, N = 336.



It is usually convenient to assume that f_0 and f_1 are both Gaussian, and sometimes the further assumption is made that they are of equal variance. Normality of the distributions is 'a highly robust, empirical result, which is now substantiated in dozens of diverse applications' (Swets and Pickett 1982).

Exact equality of variances is in general not the case. They do not, however, usually differ greatly; Swets and Pickett (1982) state that almost all empirical ROCs imply ratios of the variance of f_0 to that of f_1 between 0.5 and 1.5, and this also seems to apply to weather forecast ROCs.

In Fig. 2, the decision criterion is represented by the point x_c , so that observation of a value for X greater than x_c requires a forecast of occurrence of the predictand, and less than x_c , non-occurrence. The model hence provides the following expressions for the probability of hits and false alarms.

$$\Pr\{\text{hit}\} = \Pr\{\text{event forecast} \mid \text{event occurs}\}$$

$$= \int_0^{\infty} f_1(x_c) dx \quad \dots 3$$

$$\Pr\{\text{false alarm}\} = \Pr\{\text{event forecast} \mid \text{event does not occur}\}$$

$$= \int_0^{\infty} f_0(x_c) dx \quad \dots 4$$

represented by the diagonally and vertically hatched areas respectively in Fig. 2. Note that a false alarm is similar to a type I error in statistical hypothesis testing, and the probability of a hit to the power of a statistical test, or, one minus the probability of a type II error.

If the decision criterion x_c is allowed to vary through its range then Eqns 3 and 4 show that hit and false alarm probabilities vary together (in the sense that both increase as x_c decreases and decrease as x_c increases) and hence trace out a ROC. The precise form of the ROC is determined by the nature and parameters of f_1 and f_0 . Since a Gaussian distribution is completely specified by its mean and variance, only one parameter is required to describe the model in the equal variance case. This is the separation of the mean values, denoted d' when given in units of the common standard deviation. In Fig. 3 the solid curves show ROCs generated by Gaussian distributions with equal variance for several different values of d' . The broken lines are isopleths of decision criterion, here given as the likelihood ratio

$$\beta = \frac{f_1(x_c)}{f_0(x_c)} \quad \dots 5$$

Isopleths of β show how the probabilities of a hit and false alarm co-vary if the decision criterion is kept constant while d' changes. It proves to be more convenient to plot ROCs on double-probability axes; since the distributions are Gaussian these ROCs are straight lines, and Fig. 4 shows the ROCs of Fig. 2 plotted on such axes.

β is related to p_c through Bayes' formula, in the 'odds' form

$$\frac{p_c}{1-p_c} = \frac{p_0}{1-p_0} \beta \quad \dots 6$$

where p_0 is prior probability.

If the variance of f_1 and f_0 are not equal then two parameters are required to specify the model. These are the separation of the means, conventionally denoted by Δm in the unequal variance case, and the ratio of the standard deviations, denoted by $s = \sigma_0/\sigma_1$ where σ_0 is the standard deviation of f_0 and σ_1 that of f_1 .

The effect of unequal variances on the ROC is shown in Fig. 5 on both linear and double probability axes (based on Fig. 3-3 in Green and Swets 1974). Note that changes in s affect the slope of the ROC in the latter case, but changes in Δm do not.

Plotting on double probability axes is facilitated by transforming hit and false alarm rates to the corresponding value of the standard normal deviate. Double probability axes are equivalent to axes with a scale linear in the normal deviate. Figure 6 shows the data of Fig. 1 plotted in this way. Formally, h is transformed to a value z_1 using the relation

$$h = \int_{z_1}^{\infty} f_1(z_c) dz_c \quad \dots 7$$

where $z_c = (x_c - \bar{x}_1)/\sigma_1$, \bar{x}_1 being the mean of $f_1(x)$. A similar expression relates f_0 , the false alarm rate, and z_0 .

This procedure has an advantage over the use of axes linear in probability, in that the intercept of the line of best fit on the Z_0 axis is an estimate (with sign reversed) of the model parameter Δm , and the slope, dz_1/dz_0 is an estimate of s .

Indices of performance based on the SDT model

Several indices of performance can be evaluated from the fitted normal-normal ROC. Swets and Pickett (1982) recommend A_z , the proportion of the ROC unit square that lies beneath the fitted ROC (on linear probability scales). Similarly to $P(A)$, it ranges from 0.5 for a ROC along the major diagonal, indicating performance at a chance level, to 1.0 for perfect performance. A_z is superior to $P(A)$ because it is much less affected by the number or scatter of the data points that define the ROC.

There are also some indices similar to d' which are essentially measures, subject to various scale changes, of the distance between the means of the f_1 and f_0 distributions. Swets and Pickett prefer $z(A)$, the normal deviate value that corresponds to the area measure A_z when A_z is taken as a probability. Multiplying $z(A)$ by $\sqrt{2}$ gives the value of an index called d_a which has some advantages from a statistical point of view (Simpson and Fitter 1973).

Single-number indices inevitably lose some of the information in the ROC, since a straight line requires two numbers for its definition. Particular values of A_z or d_a can result from linear ROCs of different slopes and intercepts. The two parameter measure $D(\Delta m, s)$ fixes the entire ROC, so that it can be reconstructed and the variation in h and f assessed for various criteria. If $\Delta m = 1.5$ and $s = 0.9$ then this measure is written $D(\Delta m, s) = (1.5, 0.9)$. The range of Δm is effectively from 0 to 4 or 5; s is almost always between 0.5 and 1.5.

Values for these indices can be estimated directly from the graph, or using a computer program developed by Dorfman and Alf (1969), listed in Swets and Pickett (1982).

Validation of the normal-normal model for probabilistic forecasts

This section of the paper presents data to show that the SDT model using normal distributions can be used to describe subjectively formulated probabilistic forecasts. For this it is sufficient to demonstrate that ROCs calculated from such forecasts are linear on Gaussian double-probability axes. (The model has implications for the calibration of probabilistic forecasts but this aspect will not be pursued in detail. Only the linearity or otherwise of empirical ROCs is of concern.)

A number of sets of subjective probability forecasts are presented as ROCs on double probability axes, and the straight line of best fit to

the points is shown. The straight line is the ROC which would be predicted using the normal-normal model described above with the appropriate fitted values for Δm and s . These values can be estimated from the empirical ROC, but are not of direct concern in this paper. The line was fitted using conventional least-squares methods on values of Z corresponding to hit and false alarm rates.

Correlation coefficients and significant levels were calculated but are not given. Tested against the null hypothesis of no linear correlation all the correlation coefficients are significant well beyond the 0.01 level.

Perusal of the Figures indicates that this conclusion is trivial. A straight line obviously fits the data well in all cases. The only kind of hypothesis worth testing linearity against is that there is a curved line that would fit better, equivalent to the hypothesis that the underlying distributions have some systematic deviation from normality, or that other distributions would be more appropriate. This possibility may be worth further investigation. Green and Swets (1974) suggested that exponential distributions have some theoretical advantages, and the psychological literature contains discussions of other distributions, summarised largely in Egan's book (1975). The Figures to follow make the case for normal distributions in the meteorological context and other possibilities will not be pursued.

Predictands are rainfall as point and area probabilities, temperature as fixed width credible interval probabilities, and frost, tornadoes, fog and lightning. The latter two were forecast in four risk categories (negligible, slight, moderate and high) and thus provide only three points on the ROC. Fitting three points with a two-parameter model is admittedly a procedure of questionable validity; nevertheless, it is evident (Figs 20 and 22) that, at the least, they do not conflict with the model, and in view of the good fit shown for predictands for which more categories are available, provide some further support.

Most of the forecast sets shown here are aggregates, produced by a number of different individuals. It has been suggested by Craig (1977) that the smoothly curved shape of the ROC on linear axes is an artifact of this aggregation, and that a 'threshold' decision model might be more appropriate for individuals considered separately, leading to a ROC composed of linear segments (on linear axes). For this reason four sets of forecasts from individual forecasters are included; they also support the normal-normal model (Figs 9,10,14,15).

Figures 6, 7 and 8 are ROCs for precipitation probabilities done in the Canberra RFC. In Fig. 6 the predictand was occurrence of measurable precipitation at any one of 25 official daily rain gauges in the vicinity of Canberra City during the 24-hour period 9.00 am to 9.00 am local time. The estimates were done at about 5.00 pm on the

previous day. The probability that an area. In Figs occurrence of precip report during th 9.00 pm.

In Fig. 7 the am 8 more than between hit rate a Figures 9 and recasters in Canl ta plotted in Fig 6 and 83 respo reasonable hypoth Figures 11 ar probabilistic forec the Great Falls experiment cond (1974). The exper ffect of guidan precipitation pro own in this p examining the g normal-normal

Figures 13, 14 Weather Service issued to the publ was published by These are large s of Fig. 12, and t and hence to the excellent. Figure individual foreca Fig. 12.

Figures 16 an credible interval set was publishe The forecaster probability that maximum or mi of 5F (Fig.16) o of their subjecti sizes are not lar more scatter ab of rainfall prob

Figure 18 shc the probability or equal to 28F Albuquerque, Murphy (1977)

Figures 19 probabilities re in the USA. T the National Kansas City, Murphy and generated by e more tornado areas delineat question, and occurrence of

is shown. The straight line is the ROC that would be predicted using the normal-normal model described above with the appropriate fitted parameters Δm and s . These values can be estimated from the empirical ROC, but are not of direct use in this paper. The line was fitted using the ordinary least-squares methods on values of Δm and s to hit and false alarm rates. The coefficients and significant levels are not given. Tested against the hypothesis of no linear correlation all the coefficients are significant well beyond the 5% level.

Figure 7 indicates that this is a trivial case. A straight line obviously fits the data in all cases. The only kind of hypothesis being tested against linearity is that there is a better fit that would fit better, equivalent to the hypothesis that the underlying distributions have systematic deviation from normality, or that other distributions would be more appropriate. This may be worth further investigation. Swets (1974) suggested that exponential distributions have some theoretical advantages, and the meteorological literature contains discussions of other distributions, summarised largely in Egan (1974). The Figures to follow make the case for the normal-normal model. Other possibilities will not be pursued. The data are rainfall as point and area, temperature as fixed width credible intervals, and frost, tornadoes, fog and the letter two were forecast in four risk categories (negligible, slight, moderate and high). Only three points on the ROC are plotted with a two-parameter model is a procedure of questionable validity. It is evident (Figs 20 and 22) that, although they do not conflict with the model, and in good fit shown for predictands for which data are available, provide some further

the forecast sets shown here are produced by a number of different methods. It has been suggested by Craig (1977) that the smoothly curved shape of the ROC is an artifact of this aggregation, and a 'threshold' decision model might be more appropriate for individuals considered separately. The ROC composed of linear segments (on the left) is a good fit. For this reason four sets of forecasts from individual forecasters are included; they also fit the normal-normal model (Figs

7 and 8 are ROCs for precipitation done in the Canberra RFC. In Fig. 6 the predictand was occurrence of measurable rainfall at any one of 25 official daily rain gauges in the vicinity of Canberra City during the period 9.00 am to 9.00 am local time. The forecasts were done at about 5.00 pm on the

previous day. The task is essentially to estimate the probability that an event will occur somewhere over the area. In Figs 7 and 8, the predictand was occurrence of precipitation in the gauge at Canberra Airport during the 12-hour period 9.00 am to 9.00 pm.

In Fig. 7 the amount was a trace or more, and in Fig. 8 more than 2.5 mm. A strong linear relationship between hit rate and false alarm rate is evident.

Figures 9 and 10 are ROCs for individual forecasters in Canberra, and are both subsets of the data plotted in Fig. 5. Sample sizes are rather small (25 and 83 respectively) but linearity is still a reasonable hypothesis.

Figures 11 and 12 show ROCs for some probabilistic forecasts of precipitation produced at Great Falls and Seattle WSOs during an experiment conducted by Murphy and Winkler (1974). The experiment was intended to assess the effect of guidance (PEATMOS) forecasts upon precipitation probability forecasts. The forecasts shown in this paper are those produced before examining the guidance. Both sets clearly fit the normal-normal model well.

Figures 13, 14 and 15 are ROCs for US National Weather Service precipitation probability forecasts issued to the public in Chicago, Illinois. The data set is published by Murphy and Winkler (1977(a)). These are large sets of forecasts, 17 514 in the case of Fig. 12, and the closeness of fit to a straight line is hence to the normal-normal model is obviously excellent. Figures 14 and 15 are ROCs for two individual forecasters, and are subsets of the data in Fig. 12.

Figures 16 and 17 are ROCs for 'fixed width credible interval' temperature forecasts. The data were published by Murphy and Winkler (1974). The forecasters were asked to estimate the probability that the observed temperature (either maximum or minimum) would lie within an interval of 5F (Fig. 16) or 9F (Fig. 17) centred on the median of their subjective probability distribution. Sample sizes are not large (122 forecasts each) and there is more scatter about the straight line than in the case of rainfall probabilities.

Figure 18 shows the ROC for a set of estimates of the probability of a minimum temperature less than or equal to 28F, formulated by NWS forecasters at Albuquerque, New Mexico, and published by Murphy (1977).

Figures 19 and 20 are ROCs for two sets of probabilities related to the occurrence of tornadoes in the USA. They were produced by forecasters at the National Severe Storms Forecast Centre in Kansas City, Missouri and were published by Murphy and Winkler (1977(b)). Figure 18 was generated by estimates of the probability that one or more tornadoes would occur in the severe weather areas delineated in the outlook on the day in question, and Fig. 20 by probabilities for the occurrence of ten or more tornadoes anywhere in the

USA on that day. Sample sizes were relatively small (92 in each case) and there is some scatter evident in Fig. 20. These were not particularly good forecasts, either from the point of view of reliability or resolution. Murphy and Winkler state that 'the forecasters were not experienced in making such forecasts, and they did not receive any feedback concerning their performance during the period of the experiment, ...'. They still fit the normal-normal model well.

Figure 21 shows the ROC for some lightning risk forecasts done for the Australian Capital Territory during bushfire seasons from 1973/74 to 1980/81. The forecasts are not issued as numerical probabilities but in four categories of risk; nil, slight, moderate and high. This procedure gives only three points on the ROC. It can be seen, however, that the points lie close enough to the line of best fit to make a linear relationship very plausible.

Finally, Fig. 22 shows ROCs for three sets of fog risk forecasts for Canberra Airport. The points identified by circles correspond to forecasts issued at 9.00 am local time for the following morning, squares to forecasts issued at 3.00 pm and hexagons, 9.00 pm. A linear relationship between different degrees of risk is evident at all three times. It is interesting also to note that Δm increases as the lead time decreases, showing that the forecasters were more successful in discriminating between occurrence and non-occurrence of fog as more relevant data became available, as one would hope.

Discussion

Figures 6 to 22 provide substantial support for the proposition that the signal detection theory model with normal distributions is an excellent descriptive model for ROCs derived from weather forecasts. A linear relationship between hit and false alarm rates as decision criterion varies (on double-probability axes), is, at the very least, an acceptable hypothesis for all the forecast sets presented here.

It follows that the signal detection theory model with Gaussian probability distributions is a valid descriptive model for weather forecasts, at least of the type considered here (subjective probability forecasts). Indices derived from the SDT model can therefore be used to describe weather forecasts, and hence to make comparisons between different forecast sets.

Perhaps the greatest advantage of SDT indices is that comparisons can be made between yes/no and probabilistic forecasts without the need for some more or less arbitrary method of reducing the probabilities to a categorical form. Bryan and Enger (1967) identified strategies for converting sets of probabilistic forecasts to yes/no forecasts so as to maximise certain skill scores in the long run and Miller and Best (1979), Bermowitz and Best (1979) and Mason (1980) have looked at this problem. Reducing probabilities to zeros and ones is

somewhat unsatisfactory as a means of comparing yes/no and probabilistic forecasts, as much information is lost when probabilistic forecasts are treated in this way and the resulting set of yes/no forecasts is optimal for only a small sub-set of users. This comparison between probabilistic and yes/no forecasts can be made by plotting the empirical ROC for the probabilistic forecasts and also the point representing the hit and false alarm rates achieved by the yes/no forecasts; the relationship is immediately clear. At a lower level of detail d' can be calculated for the yes/no forecasts, and either d' or $D(\Delta m, s)$ for the probabilities, giving a good indication of relative accuracy. ROCs can also be plotted for forecasts given as risk ratings (e.g. low, moderate, high) as in Figs. 21 and 22, and for public weather forecasts that include 'chance of ...' statements. The ROC provides a framework for the display of any type of forecast for a dichotomous predictand. Its use in this form is not dependent on the SDT model, and forecast sets generated by definitely non-normal distributions can be compared in this way (Mason 1980).

Another advantage of the SDT approach is that accuracy can be assessed independently of calibration. The indices d' , Δm , d_s , A_z , (and some other related indices derived from the ROC; see Swets 1979) are measures of the inherent ability of the forecasting system to *discriminate* between situations that will be followed by occurrence of a predictand, and those that will be followed by non-occurrence. The SDT indices are relatively independent of the calibration of the probabilities, that is the relation between estimated probability and relative frequency. In fact the ROC can be determined without information about the numerical values of estimated probability. Hence we have a measure of accuracy for weather forecasts which is not confounded by individual variations in calibration of subjective probabilities.

Finally, signal detection theory is not a substantive model for the psychological processes taking place in the formulation of a forecast. It is asserted simply that it provides a very good descriptive model for weather forecast verification, with the advantages noted above. It appears that forecasters *behave as if* they are making statistical decisions on the basis of data in which the weight of evidence for the predictand varies from day to day, and this variation in weight of evidence is well represented by the normal-normal model of signal detection theory.

Conclusions

The relative operating characteristic, a graph of hit rate against false alarm rate as decision criterion varies, can be used to display and compare forecast sets for dichotomous predictands presented in any form, whether as numerical probabilities, risk ratings, yes/no, or verbally as in public weather

forecasts. It represents a very general paradigm for assessment of forecast quality.

ROCs for subjective probability forecasts are linear when plotted on double probability axes, supporting the normal-normal signal detection theory model. Hence indices based on the SDT model may be used to describe subjectively formulated weather forecasts, and have the advantage of being relatively independent of the locations of the forecasters' decision criterions. In the case of probability forecasts this provides indices of accuracy that are effectively independent of calibration.

Acknowledgment

The author wishes to thank Dr A. H. Murphy for encouragement with the work presented in this paper and the staff of the ACT Regional Office of the Australian Bureau of Meteorology for assistance in a wide variety of ways. Stephanie Monro typed numerous drafts of the manuscript.

References

- Bermowitz, R. J. and Best, D. L. 1979. An objective method for maximising threat score. *Sixth Conference on Probability and Statistics in Atmospheric Sciences, Banff Alta., Canada, 9-12 October 1979, American Meteorological Society*, 103-7.
- Broadbent, D. E. and Gregory, M. 1963. Vigilance considered as a statistical decision. *Brit. J. Psychol.*, 54, 309-23.
- Bryan, J. G. and Enger, I. 1967. Use of probability forecasts to maximise various skill scores. *Jnl appl. Met.* 6, 762-9.
- Colls, K. E., Mason, I. B. and Daw, F. A. 1981. A forecast verification procedure for public weather forecasts. *Aust. Met. Mag.*, 29, 9-23.
- Craig, A. 1977. Broadbent and Gregory revisited: vigilance and statistical decision. *Human Factors*, 19(1), 25-36.
- Dorfman, D. D. and Alf Jr., E. 1969. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals — rating method data. *J. Math. Psych.*, 6, 487-96.
- Egan, J. P. 1975. *Signal detection theory and ROC analysis*. Academic Press, New York.
- Green, D. M. and Swets, J. A. 1974. *Signal detection theory and psychophysics*. Wiley, N.Y. 1966, reprinted by Kreiger, Huntington, N.Y.
- Gulezian, D. P. 1981. A new verification score for public forecasts. *Mon. Weath. Rev.* 109 313-23.
- McCarthy, D. and Davison, M. 1980. Independence of sensitivity to relative reinforcement rate and discriminability in signal detection. *Jnl. of the Experimental Analysis of Behaviour*, 34, 273-84.
- Mason, I. 1979. On reducing probability forecasts to yes/no forecasts. *Mon. Weath. Rev.*, 107, 207-11.
- Mason, I. 1980. Decision-theoretic evaluation of probabilistic predictions. In *The collection of papers presented at the WMO Symposium on Probabilistic and Statistical Methods in Weather Forecasting, Nice, 8-12 September 1980*, 219-28.
- Mason, I. 1982. On scores for yes/no forecasts. *Preprints of papers delivered at the ninth conference on weather forecasting and analysis, 28 June - 1 July 1982, Seattle, Washington, USA. American Meteorological Society*.

represents a very general paradigm for forecast quality.

Subjective probability forecasts are plotted on double probability axes, the normal-normal signal detection model. Hence indices based on the SDT can be used to describe subjectively weather forecasts, and have the advantage of being relatively independent of the forecasters' decision criterions. In probability forecasts this provides indices that are effectively independent of

Judgment

I wish to thank Dr A. H. Murphy for his interest with the work presented in this paper. I am also grateful to the staff of the ACT Regional Office of the Australian Bureau of Meteorology for their assistance in a wide variety of ways. Stephanie and I have written numerous drafts of the manuscript.

S

Mason, I. B. and Best, D. L. 1979. An objective method for maximising threat score. *Sixth Conference on Probability and Statistics in Atmospheric Sciences, Banff, Alta., Canada, 9-12 October 1979, American Meteorological Society*, 103-7.

Swets, J. A. and Gregory, M. 1963. Vigilance and statistical decision. *Brit. J. Psychol.*, 54, 1-10.

Mason, I. B. and Enger, I. 1967. Use of probability forecasts to maximise various skill scores. *Jnl appl. Met.*, 6, 1-5.

Mason, I. B. and Daw, F. A. 1981. A forecast verification procedure for public weather forecasts. *Meteorol. Mag.*, 29, 9-23.

Broadbent and Gregory revisited: vigilance and statistical decision. *Human Factors*, 19(1), 25-36.

Mason, I. B. and Alf Jr., E. 1969. Maximum likelihood estimates of parameters of signal detection theory and of confidence intervals — rating method. *Psychol. Monographs*, 6, 487-96.

Mason, I. B. 1975. *Signal detection theory and ROC curves*. Academic Press, New York.

Mason, I. B. and Swets, J. A. 1974. *Signal detection theory and psychophysics*. Wiley, N.Y. 1966, reprinted 1974, Huntington, N.Y.

Mason, I. B. 1981. A new verification score for public weather forecasts. *Mon. Weath. Rev.* 109 313-23.

Mason, I. B. and Davison, M. 1980. Independence of the relative reinforcement rate and decision bias in signal detection. *Jnl. of the Experimental Psychology: Applied*, 34, 273-84.

Mason, I. B. 1979. On reducing probability forecasts to a single score. *Mon. Weath. Rev.*, 107, 207-11.

Mason, I. B. 1980. Decision-theoretic evaluation of weather forecasts. In *The collection of papers presented at the WMO Symposium on Probabilistic and Statistical Methods in Weather Forecasting*, Nice, 8-12 October 1980, 219-28.

Mason, I. B. 1982. On scores for yes/no forecasts. *Preprints presented at the ninth conference on weather forecast analysis*, 28 June - 1 July 1982, Seattle, Washington, USA. American Meteorological Society

Miller, R. G. and Best, D. L. 1979. A model for converting probability forecasts to categorical forecasts. *Sixth Conference on Probability and Statistics in Atmospheric Sciences, Banff, Alta., Canada, 9-12 October 1979, American Meteorological Society*, 98-102.

Murphy, A. H. 1973. A new vector partition of the probability score. *Jnl appl. Met.*, 12, 595-600.

Murphy, A. H. 1977. The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Mon. Weath. Rev.*, 105, 803-16.

Murphy, A. and Williamson, D. 1976. *Weather Forecasting and Weather Forecasts. Models, Systems and Users — Vol II*. Available from NCAR Publications Office, P.O. Box 3000, Boulder, CO 80307, USA.

Murphy, A. H. and Winkler, R. L. 1974. Subjective probability forecasting experiments in meteorology: some preliminary results. *Bull. Am. met. Soc.*, 55, 1206-16.

Murphy, A. H. and Winkler, R. L. 1977(a). Reliability of subjective probability forecasts of precipitation and temperature. *Jnl Roy. Statist. Soc. (c)*, 26, 41-7.

Murphy, A. H. and Winkler, R. L. 1977(b). Probabilistic Tornado Forecasts: some experimental results. *Preprints, Tenth Conference on Severe Local Storms, 18-21 October 1977, Omaha, N.E., American Meteorological Society*, 403-9.

Olson, R. H. 1965. On the use of Bayes' theorem in estimating false alarm rates. *Mon. Weath. Rev.*, 93, 557-8.

Simpson, A. J. and Fitter, M. J. 1973. What is the best index of detectability?. *Psychological Bulletin*, 80, 481-8.

Swets, J. A. 1969. Effectiveness of information retrieval methods. *American Documentation*, 20, 72-89.

Swets, J. A. 1973. The relative operating characteristic in psychology. *Science*, 182, 990-1000.

Swets, J. A. 1979. ROC analysis applied to the evaluation of medical imaging techniques. *Investigative Radiology*, 14, 109-21.

Thompson, J. C. and Brier, G. W. 1955. The economic utility of weather forecasts. *Mon. Weath. Rev.*, 83, 249-53.

Woodcock, F. 1976. The Evaluation of Yes/No forecasts for scientific and administrative purposes. *Mon. Weath. Rev.*, 104, 1209-14.

Woodcock, F. 1981. Hanssen and Kuipers' discriminant related to the utility of yes/no forecasts. *Mon. Weath. Rev.*, 109, 172-3.

Yates, J. F. 1982. External correspondence; decompositions of the mean predictability score. *Organizational Behaviour and Human performance*, 30, 132-56.