Aust. Met. Mag. 30 (1982) 291-303

551.509.51

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

perhaps, but not necessarily, a

ether or not to predict the event is
parison of this number with a
' which is predetermined. The
vith statistical hypothesis testing,
f a test variable (z, t, χ^2, etc.) is
or 99 per cent values of these
to decide whether to accept or
sis.

the paper falls into three parts.
n of a decision criterion is
he use of a variable decision
erate the ROC for a set of
asts is described. Thirdly, the
model from signal detection
d as a descriptive model for the
dices of forecasting performance
del are given. Signal detection
extensive literature, and the
paper refers only to those parts
relevant to weather forecast
ful entry to the field is Swets'
ailed presentations can be found
nd Swets (1974), Egan (1975) or
(1982).

on

nponents of a decision-making

n alternatives, for example
ecasts of occurrence or non-
f the event, or assertion that
nt alone or there is a signal as
;
events, for example rain or no
ture above or below zero, signal
sent;
available to the decision-maker
vents; and
terion, some specific value x_c of
riable X.
e value depends on the available
e given as the conditional
vent given current data, or in
od ratio:

ons event occurs},
ons event does not occur}

onotonic function of likelihood
he log likelihood ratio or log
thought of as analogous to a
n, or quite generally just as a
hat *provides information* about

en supposed to be made on the
alue x_c of the decision variable.
ion is required and for $x < x_c$
to the case of probabilistic
partition of the range of values
ritical values $\{x_i\}$, i = 1 to M-1,
er 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.**

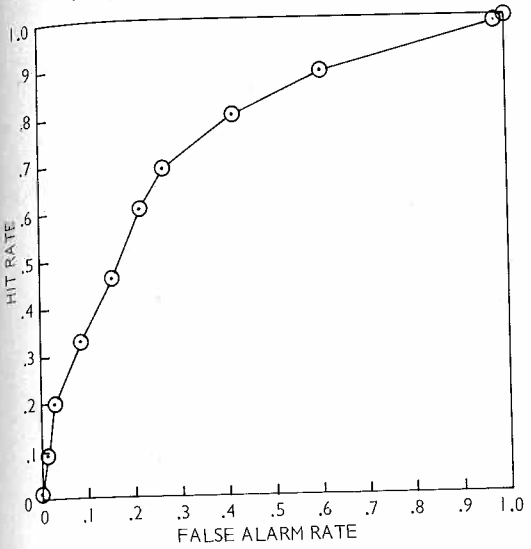
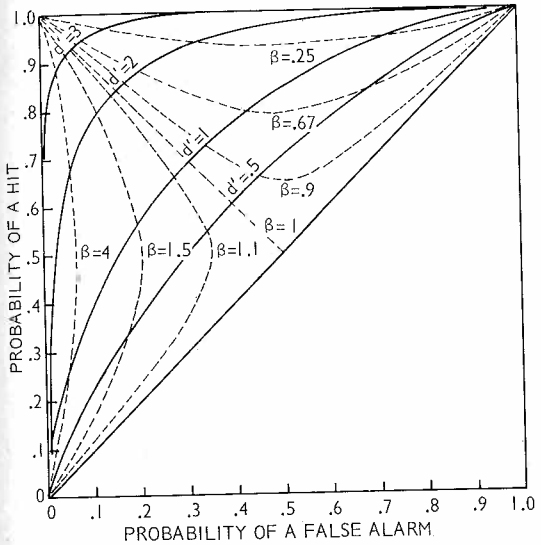


Fig. 3 **Relative operating characteristics generated by Gaussian distributions with equal variance for d' = 0.5, 1.0, 2.0 and 3.0. Broken lines are isopleths of the likelihood ratio $\beta = f_1(x)/f_0(x)$.**



for probabilities. If $\{p_i\}$, i = 1 to M, is the set 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.**

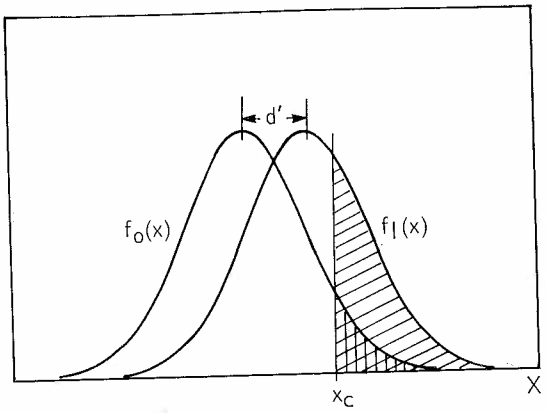
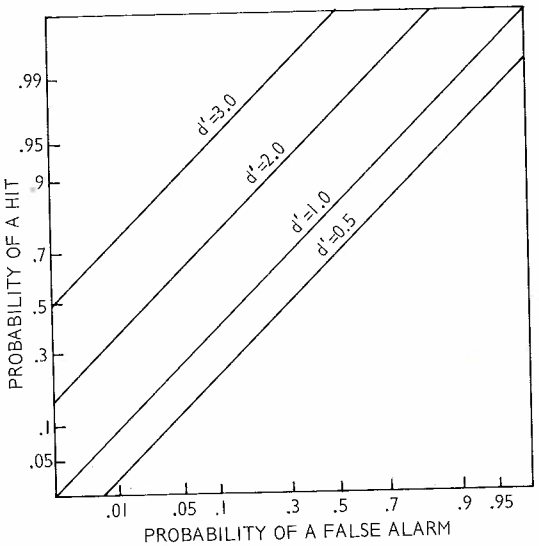


Fig. 4 **Relative operating characteristics for the Gaussian equal variance case with d' = 0.5, 1.0, 2.0 and 3.0, on double-probability axes.**



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:

hit rate, h = Pr{event predicted | event occurs} = d/(b+d) . . . 1

false alarm rate, f = Pr{event predicted | event does not occur} = c/(a+c). . . . 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 $f_1(x)$ is one standard deviation from that of $f_0(x)$ in all three cases.

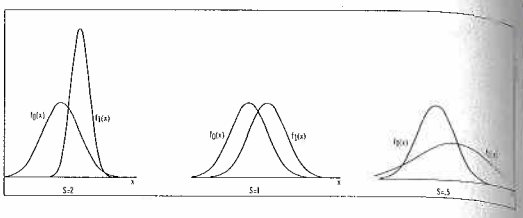


Fig. 5(b) Relative operating characteristics generated by the three pairs of distributions shown in Fig. 4, on linear axes.

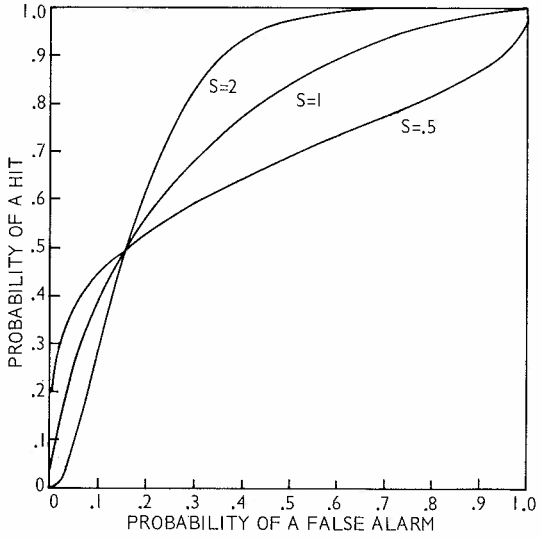


Fig. 5(c) Relative operating characteristics generated by the three pairs of distributions shown in Fig. 4, on double probability axes.

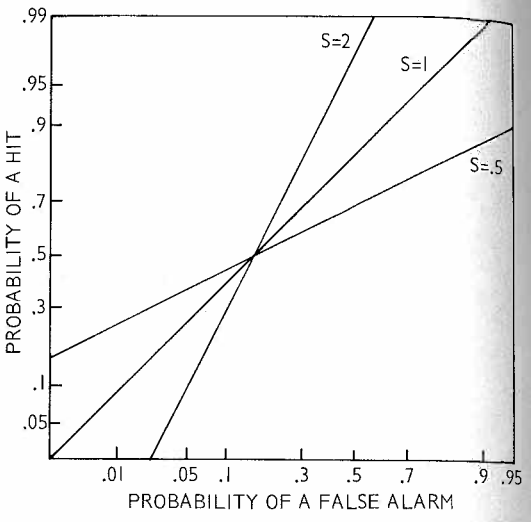


Fig. 6 Relative operating characteristics generated by a set of 341 estimates of the probability of rain in an area around Canberra. Scales linear in the standard normal deviate are superimposed.

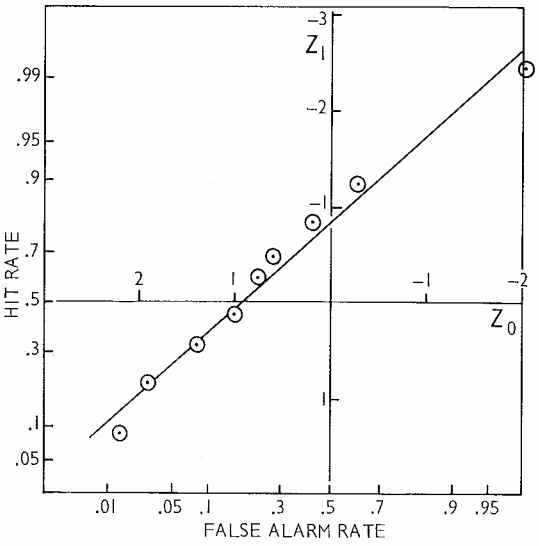
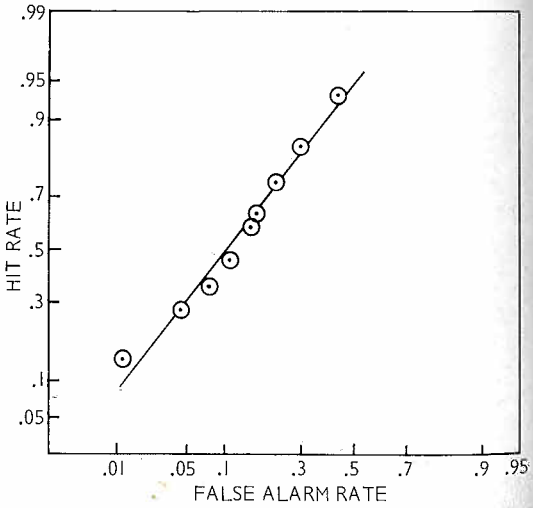
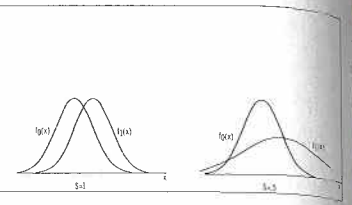
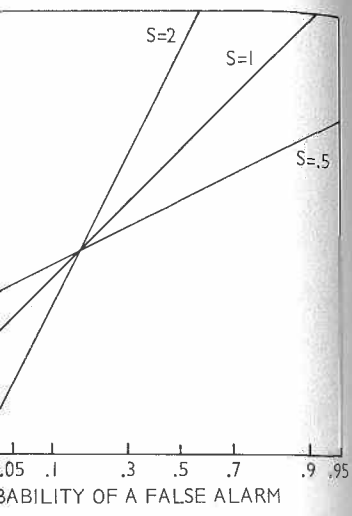


Fig. 7 ROC for estimates of probability of occurrence of measurable precipitation at Canberra Airport during the 12 hours 9.00 pm to 9.00 am local time. N = 335.





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



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

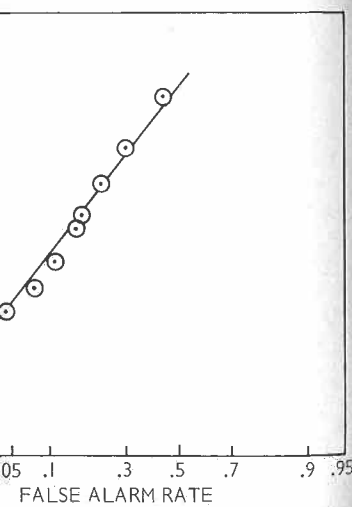


Fig 8 ROC for estimates of the probability of more than 2.5 mm of rain at Canberra Airport during the 12 hours 9.00 pm to 9.00 am local time. N = 335.

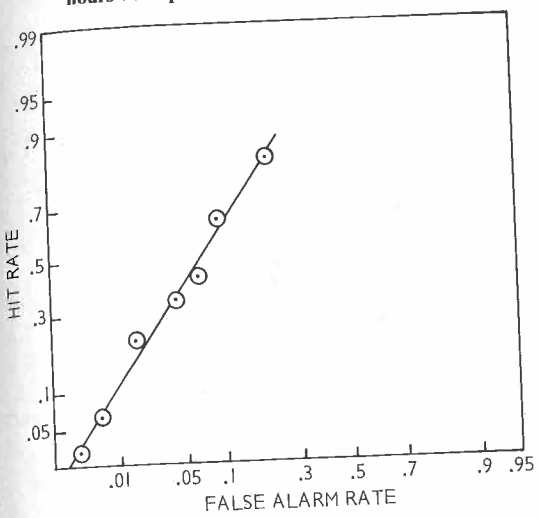


Fig. 9 ROC for individual forecaster C in Canberra, a sub-set of the data that produced Fig. 5. N = 96.

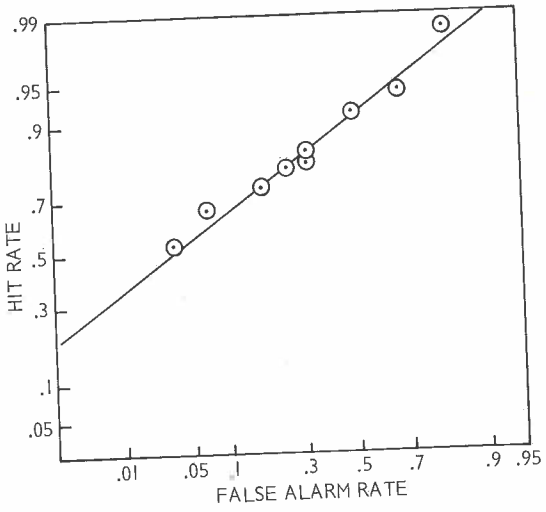


Fig. 10 ROC for individual forecaster D in Canberra, a sub-set of the data that produced Fig. 5. N = 83.

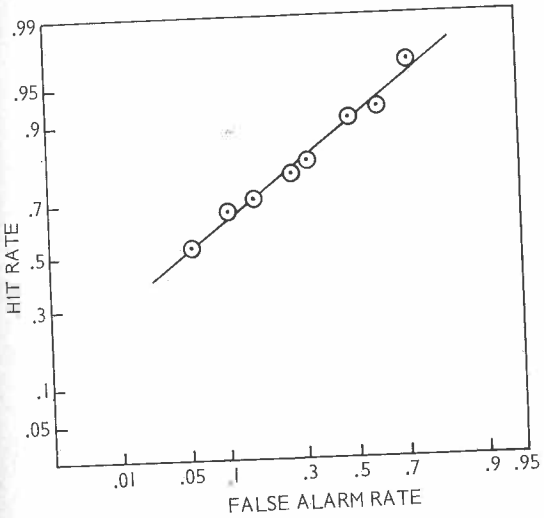


Fig. 11 ROC for estimates of the probability of rain near Great Falls, Mont., USA from data published by Murphy and Winkler (1974). N = 2646.

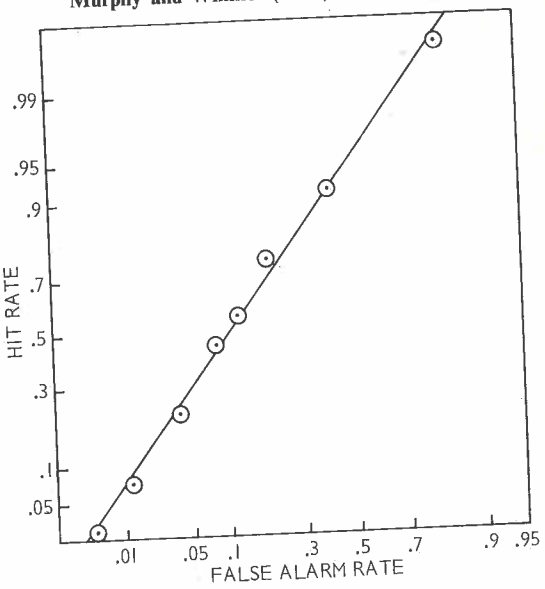


Table 1. Verification matrix for a series of N yes/no forecasts. a, b, c and d are the total frequencies of each possible combination of forecast and event.

		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.

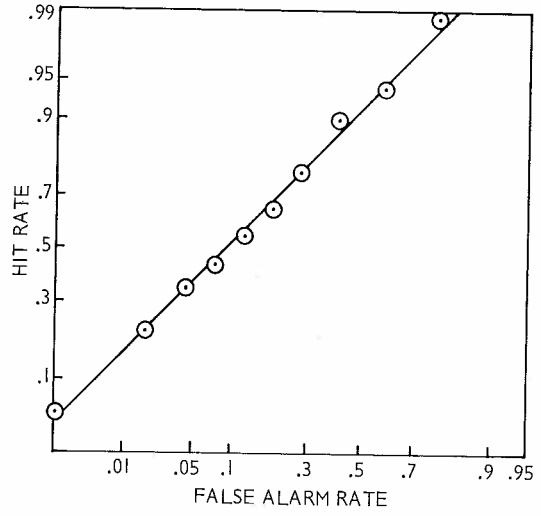


Fig. 13 ROC for US National Weather Services estimates of the probability of rain at Chicago, Illinois from data published by Murphy and Winkler (1977). N = 17 154.

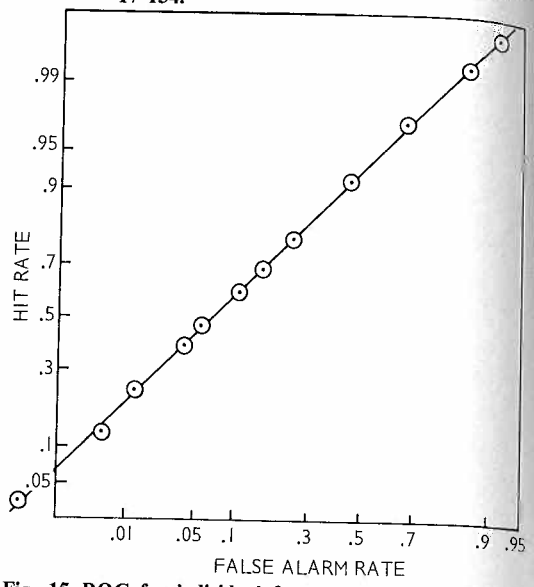


Fig. 14 ROC for individual forecaster A at Chicago. A sub-set of the data in Fig. 12. N = 2916.

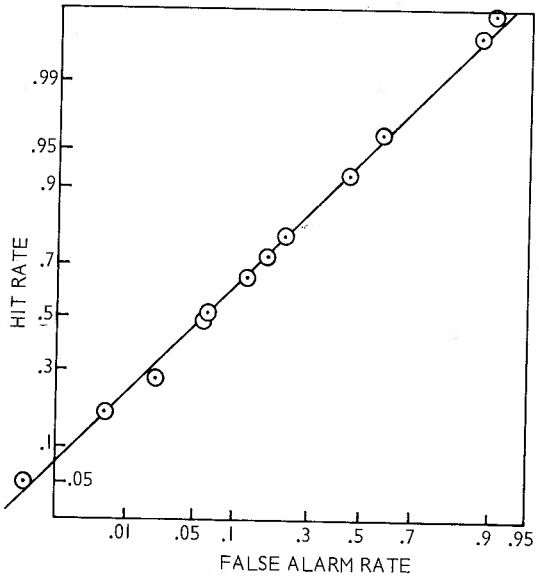
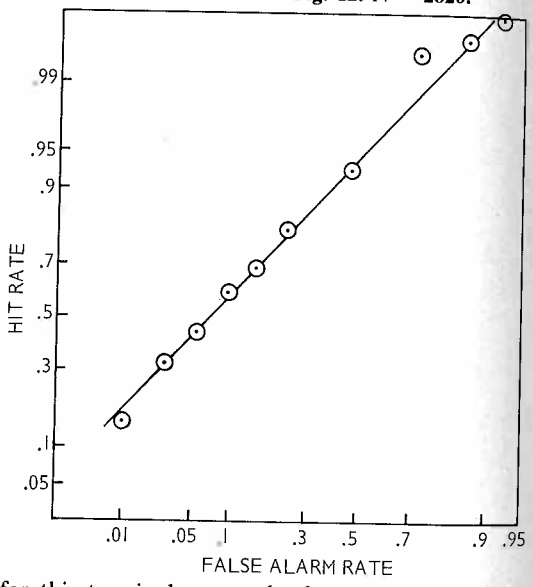


Fig. 15 ROC for individual forecaster B at Chicago. A sub-set of the data in Fig. 12. N = 2820.



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

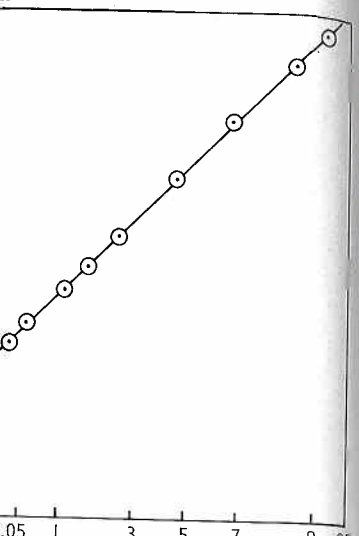
The relative operating characteristic

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.

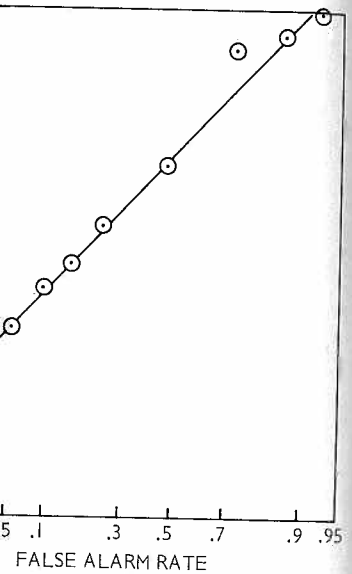
Table 2 shows an example of this procedure for a set of probability forecasts of rain over Canberra city or suburbs. Hit rates and false alarm rates are shown in rows (f) and (g), and Fig. 1 is the ROC for this set of forecasts.

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 = θ, h = 1, and perfectly wrong forecasts by the opposite point, f = 1, h = θ. A ROC lying along the major

US National Weather Services estimates
obability of rain at Chicago, Illinois from
ished by Murphy and Winkler (1977). N
1.



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



gy can be found in Swets' 1973

n example of this procedure for a
forecasts of rain over Canberra
it rates and false alarm rates are
nd (g), and Fig. 1 is the ROC for
ts.

how hit rate may be 'traded-off'
n rate by varying the decision
t performance is represented on
pper left-hand corner, f = θ, h =
vrong 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.

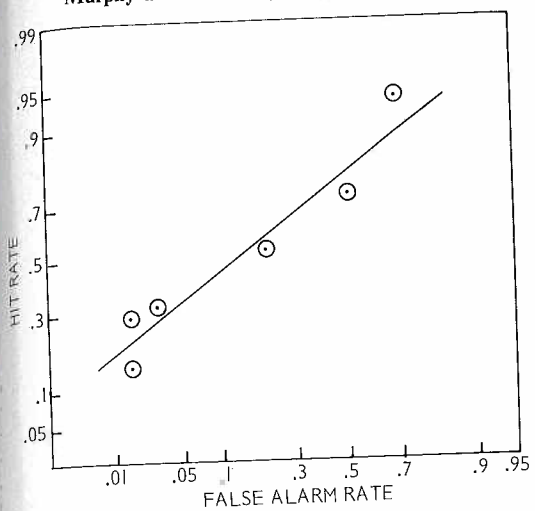


Fig. 17 ROC for 9F 'fixed width credible interval' temperature probabilities, from data published by Murphy and Winkler (1974). N = 122.

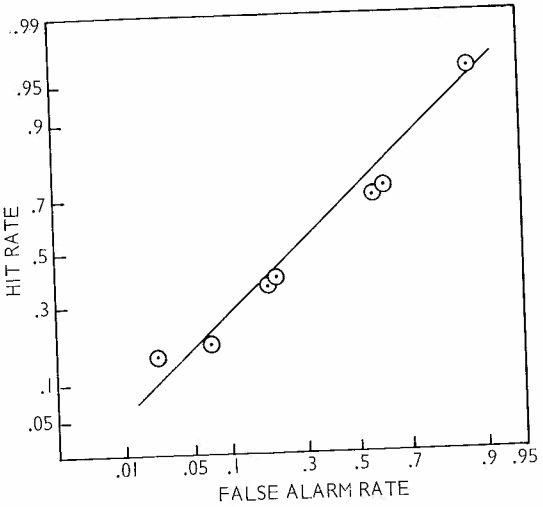


Fig. 18 ROC for estimates of the probability that minimum temperature will be less than or equal to 28°F near Albuquerque, New Mexico, from data published by Murphy (1977). N = 1443.

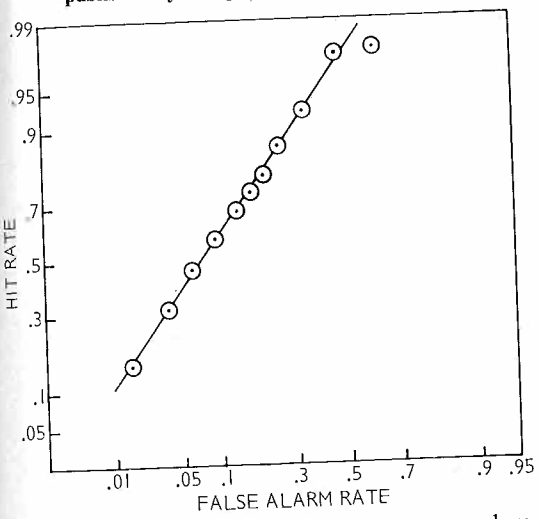
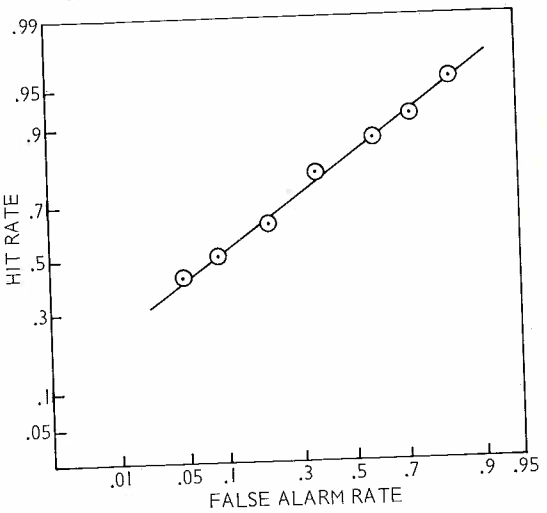


Fig. 19 ROC for estimates of the probability that one or more tornadoes will occur in the severe weather areas delineated by the US National Severe Storms Forecast Centre, Kansas City, Missouri. Data published by Murphy and Winkler 1977(b). N = 92.



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 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