

The cost of uncertainty in weather prediction: modelling quality-value relationships for yes/no forecasts

Ian Mason

(Manuscript received April 2002; revised January 2004)

This paper investigates the relationship between the skill and economic value of forecasts using a modelling approach based on signal detection theory (SDT). The value of weather forecasts in a two-action, two-state decision situation can be modelled as a function of forecasting skill, decision threshold, climate and the economic parameters of the decision situation.

In general the value of forecasts increases monotonically with increasing skill, measured by the SDT parameter d' . In some circumstances, however, reversal of the expected relationship between forecasting skill and economic value is found, so that forecast value decreases as skill increases. This is associated with use of suboptimal decision thresholds. Thresholds on skill are found for some decision situations, such that forecasts of lower quality have no economic value even though they have some positive skill and are applied optimally.

These results reinforce the importance of using the optimal decision threshold in realising the full potential value of forecasts.

Introduction

Weather forecasts are produced to help people make decisions in weather sensitive situations, so an important aspect of forecast quality is the economic value of the forecasts to decision-makers in real applications. Assessing the economic value of weather forecasts in real-life operations can, however, be complex. The forecasts are usually only one of a number of factors influencing decisions, and it may be difficult to disentangle their effect from other factors. The relevant benefits from correct forecasts or losses from incorrect forecasts may not be known exactly, or may be 'non-monetary', for example the value of life or suffering. Information necessary for optimising forecast

value in the presence of uncertainty might be unavailable, so that inappropriate decisions are made even when the forecasts are quite accurate. In the case of public weather forecasts, which are used by many different decision-makers, the difficulty is multiplied. Recent studies on the economic evaluation of forecasts include Leigh (1995), Katz and Murphy (1997), Anaman et al. (1998), Zhu et al. (2002) and Richardson (2003).

It is plausible that more accurate forecasts will lead to better outcomes in a commercial sense, so verification has often focussed on accuracy, the association between forecasts and the relevant observations, rather than value. It has been found that some measures of accuracy may not be reliable indicators of value, so that forecasts which appear more accurate may not be more valuable to all users (Chen et al. 1987; Murphy and Ehrendorfer 1987). Also, some

Corresponding author address: Ian Mason, 32 Hensman St, Latham, ACT 2615, Australia.
Email: ibmason@webone.com.au

measures of quality may have thresholds such that the forecasts have no economic value below this threshold (Katz and Murphy 1987).

Murphy and Ehrendorfer (1987) explored the relationship between the accuracy and value of simple yes/no forecasts in the cost-loss decision model. They used a form of the Brier score as the measure of accuracy, and for the economic value of the forecasts used the difference between the expected expense associated with climatological (no-skill) forecasts and that of forecasts with some skill. They found that there was a range of possible economic value for forecasts for most values of the measure of quality, and that a decrease in economic value could be associated with improvement in forecast quality. Chen et al. (1987) investigated the relationship between the quality and value of forecasts in a three-action, three-event generalisation of the cost-loss model. Categorical forecasts calibrated by past performance were considered. The measure of quality was the ranked probability score, and value was assessed as the difference between the expected expense associated with climatological forecasts and expected expense for forecasts with non-zero skill. Again, they found that there was a range of economic value for forecasts for most values of the measure of quality, and that a decrease in economic value could be associated with improvement in forecast quality. Katz and Murphy (1987) used the cost-loss ratio model to examine the way in which the value of information changes as its quality increases. Their measure of forecast quality was a linear transformation of the conditional probability of the event being forecast, given that it was forecast. They considered yes/no forecasts assumed to be calibrated by past performance and constrained so that the unconditional probability of a forecast of occurrence was the same as the climatological probability of occurrence. The measure of economic value was, similarly to the above two studies, the difference between the expected expenses associated with climatology and with forecast sets with some skill. They showed that there may be a quality threshold below which forecasts are of no economic value, because the optimal action remains the same as when a no-skill climatological forecast is used. Recent quality/value studies, summarised by Richardson (2003), have extended those of Murphy and co-workers, but have been largely focused on demonstrating the improvement in economic value available from ensemble prediction systems.

The relationship between specific measures of forecast quality and the expected economic value of the forecasts is thus of some interest. In this paper, the relationship with forecast value of one such measure of

forecast quality is considered, the quantity denoted d' from signal detection theory, defined later in the text.

Signal detection theory (SDT) provides a general framework for evaluation of diagnostic systems. It is a statistical method which has been widely used in medical diagnosis and psychology and is increasingly used in other fields to evaluate the performance of systems which seek to detect or predict specified events on the basis of information which is insufficient to provide certainty (Swets 1988). Applications in weather forecast verification can be found in Mason (1982a,b; 1989), Levi (1985), Harvey et al. (1992) and Jolliffe and Stephenson (2003).

There are two main benefits from SDT-based methods in the present context. The first is that they provide measures of forecast quality that are independent of climatology and decision threshold. Secondly, SDT makes it possible to model the dependence of performance measures on decision threshold, through an empirical relationship between hit and false alarm rates as decision threshold varies. SDT-based methods can be related to the likelihood-base rate factorisation of the joint distribution of forecasts and observations in Murphy and Winkler's (1987) distributions-oriented framework for forecast verification.

The forecasts considered here are unequivocal predictions of occurrence or non-occurrence of a simple binary event, sometimes referred to as yes/no forecasts. Invention of measures of the skill of yes/no forecasts has a long history, dating back to at least the late nineteenth century (Finley 1884; Murphy 1996). While certain measures have become widely used, for example Probability of Detection, False Alarm Ratio, Critical Success Index, proportion (or per cent) correct, the Heidke score and Hanssen and Kuipers' score among many (Mason 2003), there has been little rationale for selection of one rather than another beyond precedent and plausibility, and no generally accepted body of theory to aid interpretation of specific values and differences in scores.

Most of the common scores have been criticised on various grounds. All have a dependence on decision threshold which can render comparisons meaningless (Mason 1982b, 1989). Many also depend on the sample climate, so that comparisons of forecasting systems in different climates may be invalid. Measures of skill based on SDT avoid most of the pitfalls associated with traditional measures.

To analyse economic value this paper uses a generalisation of the well-known cost-loss ratio model, a simple model of a decision situation in which the relevant economic quantities are specified. It is assumed that decision-makers behave rationally in the sense that they seek to take the action that has the greatest

expected value, and that they have enough information about the performance of the forecasts to do this.

This study is thus in the spirit of Murphy’s (1994) ‘prescriptive’ approach to assessment of the value of forecasts.

The specific measure of value used here is the difference between the expected value of perfect forecasts and that of the actual forecasts, that is, the reduction in (expected) value due to the uncertainty in the forecasts. This is different from the usual approach in value-of-forecast studies, which has been to consider the increase in the value of the forecasts over the value achieved by a naive, no skill, forecasting strategy, usually based on climatology (e.g. Winkler and Murphy 1985). The latter approach requires explicit consideration of two cases, for threshold decision probabilities, respectively, less than and greater than the climatological probability of the predictand. These two cases appear naturally in the analysis in this paper when the skill parameter tends to zero, without requiring separate consideration.

A salient implication of this paper is the central importance of decision criterion in assessment of the quality and value of forecasts, and in the application of forecasts to real-world decisions. The decision criterion in this context is the threshold level of certainty about the predictand at which the forecast changes from ‘no’ to ‘yes’. In general, optimising the value of forecasts to different decision-makers requires different decision thresholds. Yes/no forecasts, implicitly produced at a single threshold, are likely to be significantly sub-optimal for many decision-makers. If forecasts are provided in a form which communicates the level of uncertainty, for example as probabilities, then the potential value of the forecasting system is enhanced. Realising the potential value of probabilistic forecasts requires good communication between forecasters and users, especially with regard to the definition of the event to which the probabilities refer.

The next section defines some notation, and the theoretical expression for the cost of uncertainty is derived. The sections following define skill and decision threshold in the SDT framework, and derive an expression for the optimal decision threshold in terms of the decision-maker’s economic sensitivity to weather and forecasts. Results shown are (a) the dependence of expected cost on threshold probability for fixed values of the economic sensitivity parameter, at various levels of skill, (b) the dependence of cost on threshold probability when decisions are made at the optimal threshold, (c) the dependence of cost on skill for optimal thresholds and (d) the dependence of cost on skill when suboptimal thresholds are used. The paper ends with some discussion and conclusions.

Notation

The four possible combinations of forecast and event are set out in Table 1.

The probability of each of these outcomes can be expressed in terms of conditional and marginal probabilities using standard results on factorisation of joint probabilities. If a variable F represents the forecasts, taking one of the values $\{Y,N\}$ for ‘yes’ and ‘no’ respectively, E with the same set of values represents the events, and ‘ Pr ’ is read as ‘the probability of’ then

$$Pr(TP) = Pr(F=Y,E=Y) = Pr(E=Y)Pr(F=Y|E=Y) \dots 1$$

$$Pr(TN) = Pr(F=N,E=N) = Pr(E=N)Pr(F=N|E=N) \dots 2$$

$$Pr(FP) = Pr(F=Y,E=N) = Pr(E=N)Pr(F=Y|E=N) \dots 3$$

$$Pr(FN) = Pr(F=N,E=Y) = Pr(E=Y)Pr(F=N|E=Y) \dots 4$$

Murphy and Winkler (1987) refer to this as the likelihood/base rate factorisation of the joint distribution.

A hit rate, h , false alarm rate, f , and climatological probability, p_c , are defined as follows.

$$h = Pr(F=Y|E=Y) = 1 - Pr(F=N|E=Y) \dots 5$$

$$f = Pr(F=Y|E=N) = 1 - Pr(F=N|E=N) \dots 6$$

$$p_c = Pr(E=Y) = 1 - Pr(E=N) \dots 7$$

The familiar verification statistic Probability of Detection (POD) is a sample estimate of h , and Probability of False Detection is a sample estimate of f . The False Alarm Ratio (FAR), often encountered in forecast verification, is not an estimate of f , but of $Pr(E=N|F=Y)$, which is equal to $\{1 + [p_c/(1-p_c)](h/f)\}^{-1}$.

Using the above definitions Eqns 1 to 4 become

$$Pr(TP) = p_c h \dots 8$$

$$Pr(TN) = (1 - p_c)(1 - f) \dots 9$$

$$Pr(FP) = (1 - p_c) f \dots 10$$

$$Pr(FN) = p_c(1 - h) \dots 11$$

Table 1. The four possible combinations of forecast and event for yes/no forecasts.

		EVENT	
		NO	YES
FORECAST	NO	True Negative (TN)	False Negative (FN)
	YES	False Positive (FP)	True Positive (TP)

Skill: the signal detection model for the forecasting process

The model described in this section is based on the elementary theory of detection of signals in noise, which is related to the classical Neyman-Pearson approach to statistical hypothesis testing (Swets 1973). It provides a general framework for evaluation of diagnostic systems which has been widely applied, including in medical diagnosis, non-destructive testing of metals, information retrieval and vigilance of radar operators in a military setting, among many others. Central features for present purposes are a model for the process of discriminating between occurrence and non-occurrence of weather events, and measures of diagnostic skill and decision threshold based on the model. There is an extensive literature in fields other than meteorology, particularly psychology and medical diagnosis. Texts by Swets and Pickett (1982) and a collection of papers by Swets (1996) provide an introduction to other applications. The following outline uses the simplest form of the model and includes only aspects relevant to the present application.

It is assumed that a forecaster decides between prediction of occurrence or non-occurrence of a binary weather event on the basis of the accumulated weight of evidence for the event, which is represented as a continuous scalar quantity X in Fig. 1. The forecaster has a decision threshold x^* on the X scale such that occurrence is forecast if $X \geq x^*$ and non-occurrence if $X < x^*$ (the equality is arbitrary), Fig. 1.

X is further assumed to have a specific, known, probability density $f_N(x)$ before non-occurrences, when ‘noise alone’ is present, and a different, known, density $f_Y(x)$ before occurrences, i.e. when the ‘signal’ for the event is present in addition to noise, illustrated in Fig. 2.

Recalling that $h = Pr(F=Y|E=Y)$ and that occurrence is forecast when $X \geq x^*$, it follows that h is the probability of obtaining a value of X greater than or equal to x^* from $f_Y(x)$, which represents occasions on which the event occurs, so

$$h = \int_{x^*}^{\infty} f_Y(x) dx \quad \dots 12$$

equal to the area under $f_Y(x)$ to the right of x^* , illustrated by the oblique hatching in Fig. 2. Similarly,

$$f = \int_{x^*}^{\infty} f_N(x) dx \quad \dots 13$$

the area under $f_N(x)$ to the right of x^* , represented by the horizontal hatching in Fig. 2.

In order to proceed, some particular form needs to be specified for $f_N(x)$ and $f_Y(x)$. The usual assumption

Fig. 1 Forecasting a simple event under uncertainty. The decision to forecast occurrence or non-occurrence is made by reference to a decision threshold x^* on the ‘weight of evidence’ axis, X .

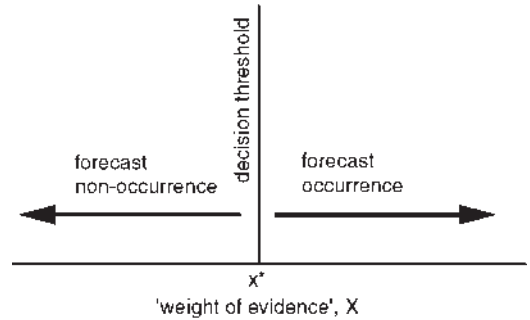
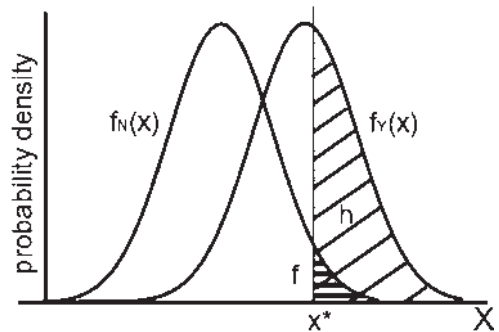


Fig. 2 Assumed distributions of ‘weight of evidence’ X before occurrences, $f_Y(x)$, and non-occurrences, $f_N(x)$. x^* is the decision threshold. The horizontal hatching represents false alarm rate corresponding to the threshold x^* and diagonal hatching, hit rate.



is that they are Gaussian and, in the case of yes/no forecasts, of equal variance. Gaussian distributions are consistent with empirical verification data (Mason 1982a) and have been found to be appropriate in many diverse applications (Swets and Pickett 1982). Other distributional forms are possible (Swets 1996) and may be worth investigating in a meteorological setting.

The assumption of equal variances is usually only approximately supported by data but is necessary in the present context because this paper is concerned

only with yes/no forecasts. The ratio of the variances of the two distributions can be estimated if forecasts are done at at least two decision thresholds. Further discussion of this parameter in the SDT model can be found in Swets (1996).

A brief discussion of the empirical basis for the model is in an appendix to this paper.

Without loss of generality, the X axis can be scaled so that the mean of $f_N(x)$ is zero and the common variance of $f_N(x)$ and $f_Y(x)$ is one.

Under these assumptions, the separation of the means of the two distributions in units of the common standard deviation is conventionally denoted d' , shown on Fig. 3, and this quantity is the measure of forecasting skill used in this paper.

It can be seen from Fig. 3 that if $f_N(x)$ and x^* are held constant while d' increases, i.e. $f_Y(x)$ moves away from $f_N(x)$, then h increases while f is unchanged. Alternatively, if x^* moves with $f_Y(x)$ so that h is held constant then f is reduced as d' increases.

Decision threshold

The location of the decision threshold x^* is clearly important in determining the realised performance of the system. If the distributions $f_N(x)$ and $f_Y(x)$ are fixed, then h and f can vary through their whole range from zero to one as a result of changes in x^* only. Values of x^* towards the left in Fig. 2 represent 'lenient' decision criteria. h approaches 1, so most of the occurrences will be correctly forecast, but f is also near 1 so there will be a higher level of false alarms. As x^* moves towards the right f decreases towards zero, so the number of false positives decreases, but the number of true positives tends to zero also; there will be more 'misses'.

Decision threshold is usually defined in the meteorological literature in terms of a critical conditional probability for the event given the evidence, denoted p^* , so

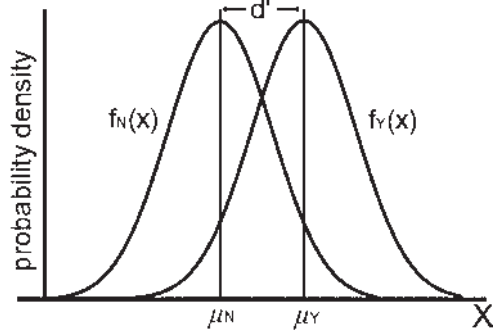
$$p^* = Pr(E=Y|X=x^*) \quad \dots 14$$

Thus, in the case of yes/no forecasts the event is forecast if its current probability, p , is greater than p^* , and forecast not to occur if p is less than p^* . A formal relation between p^* and x^* is provided by Bayes' rule which, in the case of a binary predictand, can be written in the odds form

$$w^* = w_c [f_Y(x^*)/f_N(x^*)] \quad \dots 15$$

where w^* is the odds ratio $p^*/(1-p^*)$ and w_c is the climatological odds ratio $p_c/(1-p_c)$.

Fig. 3 Assumed distributions of 'weight of evidence' X before occurrences, $f_Y(x)$, and non-occurrences, $f_N(x)$. μ_Y and μ_N are the means of $f_Y(x)$ and $f_N(x)$ respectively. $d' = \mu_Y - \mu_N$ is a measure of forecasting skill.



In the SDT context, decision threshold is usually indexed by the likelihood ratio on the right hand side of Eqn 15, denoted β , so that

$$\beta^* = f_Y(x^*)/f_N(x^*) \quad \dots 16$$

or

$$\beta^* = w^*/w_c \quad \dots 17$$

In the Gaussian equal variance model, x^* can be related to β^* and d' as follows.

Substituting the relevant functional forms for $f_Y(x^*)$ and $f_N(x^*)$, i.e.

$$f_Y(x^*) = (2\pi)^{-1/2} \exp(-(x^* - d')^2/2) \quad \dots 18$$

and

$$f_N(x^*) = (2\pi)^{-1/2} \exp(-x^{*2}/2) \quad \dots 19$$

so that

$$\beta^* = \exp(d' (x^* - d'/2)) \quad \dots 20$$

and

$$x^* = \ln\beta^*/d' + d'/2 \quad \dots 21$$

Zero skill corresponds to $d' = 0$, so that $h = f$. That this is a reasonable definition for zero skill can be seen from the following. Using the definitions of h and f (Eqns 5 and 6 above), the ratio $Pr(F=Y|E=Y)/Pr(F=Y|E=N) = 1.0$. In this case Bayes' formula in the odds form gives

$$Pr(E=Y|F=Y)/Pr(E=N|F=Y) = p_c/(1-p_c) \quad \dots 22$$

which implies $Pr(E=Y|F=Y) = p_c$, i.e. the conditional

probability of an occurrence of the weather event of interest given a forecast of occurrence is equal to the climatological probability of occurrence, so the forecast provided no information. Zero skill is equivalent to using the climatological probability for the event on every occasion.

Perfect skill is the limiting case as $d' \rightarrow +\infty$, so $h = 1.0$ and $f = 0.0$.

There are SDT-based measures of skill which are more satisfactory than d' when sufficient data are available to estimate the ratio of the variances of f_N and f_Y . A detailed discussion of other SDT-based measures of performance is in Swets (1996).

The value of forecasts

The general case is considered of a decision-maker who has to decide between just two mutually exclusive courses of action, A0 and A1. If weather event E occurs then the decision-maker prefers to have taken action A1, and if E does not occur then A0 is preferred. The value of each of the four possible combinations of weather event and action is shown in Table 2.

The elements U_{ij} in Table 2 can be regarded as monetary benefits (positive) and costs (negative), or more generally as utilities, which take into account the decision-maker's attitude to risk (von Neumann and Morgenstern 1947; Winkler and Murphy 1985; Winterfeldt and Edwards 1986).

If the choice between A0 and A1 is made on the basis of forecasts as represented in Table 1 then the overall expected utility of the forecasts, EU , is given by

$$EU = Pr(TN)U_{00} + Pr(FN)U_{01} + Pr(FP)U_{10} + Pr(TP)U_{11} \quad \dots 23$$

Substituting from Eqns 8 to 11 into Eqn 23 gives

$$EU = (1-p_c)(1-f)U_{00} + p_c(1-h)U_{01} + (1-p_c)fU_{10} + p_chU_{11} \quad \dots 24$$

Perfect forecasts have $h = 1.0$ and $f = 0.0$, so the expected utility of perfect forecasts is

$$EU_{\text{perf}} = (1-p_c)U_{00} + p_cU_{11} \quad \dots 25$$

Subtracting Eqn 24 from 25 provides an expression for the reduction in expected utility due to the uncertainty or imperfection of the forecasts, denoted CoU for cost of uncertainty.

$$CoU = (1-p_c)f(U_{00}-U_{10}) + p_c(1-h)(U_{11}-U_{01}) \quad \dots 26$$

$U_{00}-U_{10}$ is the cost penalty for false alarms; the reduction in value which results from forecasting

Table 2. Utilities of the four possible combinations of action and weather event.

		EVENT	
		NO	YES
ACTION	A0	U_{00}	U_{01}
	A1	U_{10}	U_{11}

occurrence when the event does not occur. $U_{11}-U_{01}$ is the cost penalty for misses; the cost of forecasting non-occurrence when the event does occur.

Some simplification can be obtained by defining a 'penalty ratio' R as the ratio of the false alarm penalty to the miss penalty, so that

$$R = (U_{00}-U_{10})/(U_{11}-U_{01}) \quad \dots 27$$

and also defining a 'relative' CoU , C , such that

$$C = CoU/(U_{11}-U_{01}) = (1-p_c)fR + p_c(1-h) \quad \dots 28$$

C is thus the (relative) reduction in value of the actual forecasts compared with perfect forecasts. If the U_{ij} are dollars then the units of C are dollars per forecast per dollar of miss penalty.

The optimal decision threshold

The optimal threshold probability, p^*_{opt} , is the value of p^* which minimises C . It can be derived as follows. Differentiating C (Eqn 28) with respect to p^* and equating to zero gives

$$(1-p_c)df/dp^*R - p_c dh/dp^* = 0 \quad \dots 29$$

so that

$$dh/df = R(1-p_c)/p_c \quad \dots 30$$

It can be shown (Green and Swets 1966) that dh/df is equal to the likelihood ratio β at the corresponding value of X , defined in eqn 17 above, leading to

$$R \frac{1-p_c}{p_c} = \beta^* = \frac{p^*}{1-p^*} \frac{1-p_c}{p_c} \quad \dots 31$$

and hence

$$p^*_{\text{opt}} = R/(1+R) \quad \dots 32$$

In the cost-loss model for meteorological decision making (Thompson and Brier 1955; Murphy 1977), $U_{00}=0$, $U_{11} = U_{10} = -C$, $U_{01} = -L$ and p^*_{opt} is then equal to C/L .

Equations 12 and 13 provide the form of the covariation of h and f with x^* , and x^* is related to p^* by Eqn 15. It is then possible to graph the relative cost of uncertainty, C , against p^* for fixed values of p_c and R , using Eqn 28.

In outline the procedure is as follows. The values of p_c , d' and R are fixed at the outset. For a given value of p^* , β^* is found by Eqn 17 and the corresponding value of x^* by Eqn 21. The hit and false alarm rates h and f are then calculated by Eqns 12 and 13 using the standard normal distribution function and hence a value for C is found using Eqn 28.

Results

Dependence of C on threshold probability

Figures 4 and 5 show the variation of C with p^* , using specific values for two cases. These are

- (a) $p^*_{opt} > p_c$, specifically $p_c = 0.2$ and $R = 0.3$, giving $p^*_{opt} = 0.23$ (Fig. 4) and
- (b) $p^*_{opt} \gg p_c$, specifically $p_c = 0.2$ and $R = 2.0$, giving $p^*_{opt} = 0.67$ (Fig. 5).

The range of values of d' graphed is from $d' = 0.0$ (no skill) to $d' = 3.0$ indicating a high level of skill. $d' = 3.0$ corresponds to percentage correct above 90 per cent, the exact value depending on p^* and p_c .

In Fig. 4 the optimal threshold probability is 0.23 and in Fig. 5, 0.67. The model correctly produces minima for C at these values for p^* .

The maximum (worst) values for C correspond to zero skill ($d' = 0.0$). At zero skill C has different constant values for thresholds above and below p_c , with the lower value on the side on which p^*_{opt} lies. For $p^* < p_c$, Eqn 17 with $h = f = 1.0$ gives $C = (1 - p_c)R$. For $p^* > p_c$, $h = f = 0.0$ and $C = p_c$. For the special case of $p^* = p_c$, $h = f = 0.5$ and C is the mean of its values for $p^* > p_c$ and $p^* < p_c$. These values of C are those that would be obtained using climatology as a forecast, i.e. do A1 if $p_c > p^*_{opt}$ and do A0 if $p_c < p^*_{opt}$.

The range of values of p^* over which forecasting skill has some value decreases as skill decreases. In Fig. 4, at the modest level of skill of $d' = 0.5$, an operation with $R = 0.3$ would find the forecasts commercially useful only if they were produced using p^* between about 0.05 and 0.55 (with minimum cost from forecasts produced at $p^* = 0.23$ as expected). At $d' = 0.2$, a level of skill which would be difficult to detect in practice, the range of p^* over which even slightly useful forecasts could be produced for this operation is from approximately 0.15 to 0.3. The range of p^* over which useful forecasts can be produced increases as skill increases, so that for $d' = 1.0$ the useful range is from near 0.0 to about 0.85, and $d' \geq 1.5$ effectively covers the whole range of p^* from 0 to 1.0.

Fig. 4 Variation of relative cost with threshold probability for $p^*_{opt} \geq p_c$, for d' from 0.0 to 3.0. $R = 0.3$, $p_c = 0.2$, $p^*_{opt} = 0.23$.

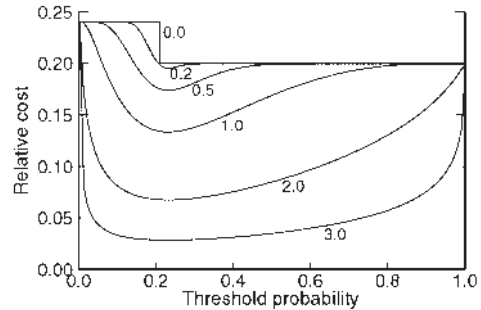
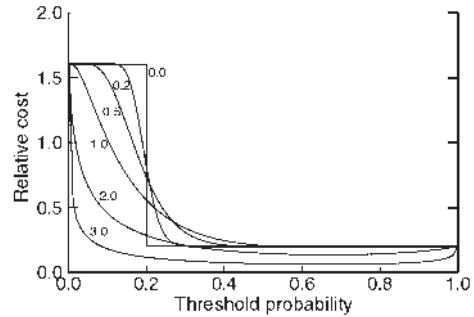


Fig. 5 Variation of relative cost with threshold probability for $p^*_{opt} \gg p_c$, for d' from 0.0 to 3.0. $R = 2.0$, $p_c = 0.2$, $p^*_{opt} = 0.67$.



For forecasts of moderate to low quality there are thus thresholds on p^* above and below p_c , between which the forecasts can have some value, and outside which they have none. The range of values of p^* over which yes-no forecasts can have some value for some users contracts steadily towards p_c from above and below as quality decreases.

Figure 5 illustrates the consequences of using very suboptimal decision thresholds. In this case $R = 2.0$, so that $p^*_{opt} = 0.667$ and p_c is again set to 0.2. If a business with this value for R used forecasts produced with p^* between 0.2 and about 0.4, at levels of skill up to at least $d' = 1.5$, the business would be worse off than using a naive strategy based on climatology. A value of 1.5 for is typical of those encountered in real forecasts.

For $p^* < p_c$ these forecasts are less costly than climatology, only because the cost of the no-skill climatological strategy jumps to high values. All forecasts produced at these p^* are however more costly than

those at p^*_{opt} and, at the lower levels of skill and thresholds further from p^*_{opt} , much more costly. Forecasts with the quite high skill of $d' = 2.5$ would have less value to a decision-maker with $R = 2.0$ using $p^* = 0.2$ than simply taking action A0 on every occasion regardless of the forecasts. This shows that in a yes-no forecasting situation the reversal of the quality/value relationship noted by Ehrendorfer and Murphy (1988) is a consequence of the use of suboptimal decision thresholds.

The curves of C against p^* are rather flat near p^*_{opt} , suggesting that the penalty for using mildly suboptimal decision thresholds may not be severe, and Figs 4 and 5 indicate that this 'flatness' increases as skill increases. The insensitivity of expected utility to suboptimal decision rules is well known (the flat maximum effect; see for example Winterfeldt and Edwards (1986)). This happens because mathematical expectation is equivalent to taking a long-term average, which tends to reduce variation. On single occasions, the actual benefit or cost will be one of the U_{ij} in Table 2, not the expected value beforehand, which lies between the U_{ij} since it is a weighted mean with weights between zero and unity. It follows that some suboptimality in p^* may not be important in terms of the average cost over a large enough number of forecasts (so long as the operation survives the occasional losses).

Variation of C with threshold probability: optimal decisions

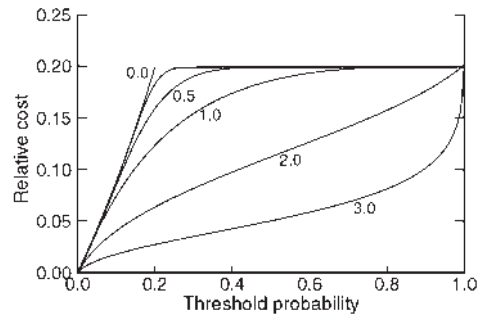
Decisions are made optimally when the threshold probability p^* is equal to $R/(1+R)$. Figure 6 shows the variation of C with p^*_{opt} in this case, for $p_c = 0.2$.

For zero skill and $p^* < p_c$, $h = f = 1.0$ and Eqn 17 gives $C = (1 - p_c)p^*/(1-p^*)$ so the curve of C against p^* from $p^* = 0$ to $p^* = p_c$ is slightly concave upwards. For zero skill and $p^* > p_c$, $h = f = 0.0$ and $C = p_c$.

Figure 6 shows that even when decisions are made optimally at low levels of skill there is again a limited range of thresholds over which some value can be achieved. A forecasting system operating at $d' = 0.5$ could only provide useful forecasts to businesses having p^*_{opt} between about 0.1 and 0.45, corresponding to R between about 0.1 and 0.8. Skill needs to reach $d' = 1.5$ before useful forecasts can be provided over the whole range of p^* .

It is interesting to note that when decisions are made at the optimal p^* , there is no reversal of the quality/value relationship. Forecasting systems with higher values of d' always have the same or higher economic value than systems with lower d' . This substantiates the comment in the previous section, that the reversal of quality/value relationships is a consequence of the use of suboptimal decision thresholds.

Fig. 6 Variation of relative cost with threshold probability for optimal decisions; for d' from 0.0 to 3.0, $p_c = 0.2$.



Variation of C with d' : optimal decisions

Figures 7 and 8 show C as a function of the SDT measure of skill d' for various values of p_c and R . Both figures assume that decisions are made optimally, i.e. values of h and f used to calculate C were those obtained at a threshold probability equal to $R/(1+R)$. In both figures $p_c = 0.2$. Figure 7 graphs C for values of R corresponding to $p^*_{\text{opt}} < p_c$, and in Fig. 8, $p^*_{\text{opt}} > p_c$.

C is a monotonically decreasing function of d' for all penalty ratios and climatological probabilities. There is thus no reversal of the quality/value relationship when (a) the equal variance SDT model applies, (b) d' is the measure of quality and, (c) decisions are made at the optimal threshold probability.

There is also no sharp quality threshold, i.e. a single specific d' such that C is reduced for higher values but not reduced for lower values. However, for values of R such that $p^*_{\text{opt}} \neq p_c$ there is a range of values of d' near zero over which the forecasts only reduce C very slightly if at all. As d' increases C is gradually reduced, i.e. the forecasts become more valuable, at a rate which depends on R , p_c and d' itself. For example, in Fig. 8 ($p_c = 0.2$), for an operation with $R = 0.1$ corresponding to $p^*_{\text{opt}} = 0.11$, the curve of C against d' is flat at the no-skill value of about 0.8 until d' reaches 0.5, above which the forecasts would have sufficient skill to be of some value. Thus, there is a diffuse quality threshold when d' is the measure of quality, and the location of this threshold depends on the nature of the operation (R), the climate (p_c), and the level of skill (d').

Fig. 7 Variation of relative cost with d' for optimal decisions and $p^*_{opt} \leq p_c$; for R from 0.05 to 0.25. $p_c = 0.2$.

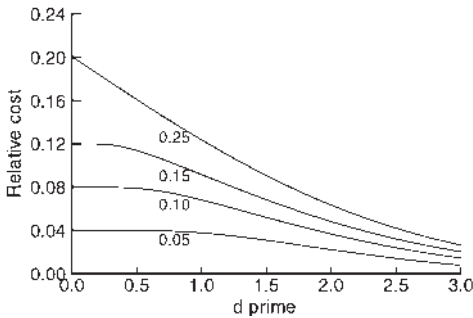
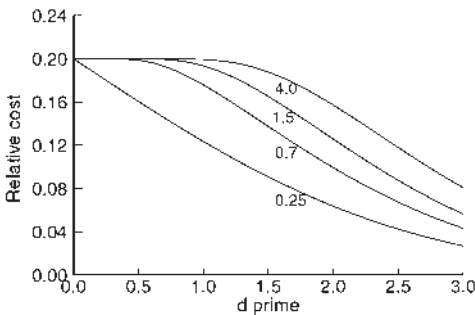


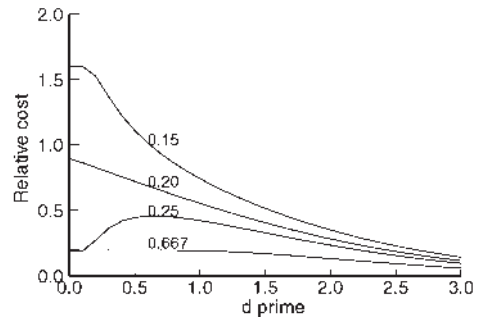
Fig. 8 Variation of relative cost with d' for optimal decisions and $p^*_{opt} \geq p_c$; for R from 0.25 to 4.0. $p_c = 0.2$.



Variation of C with d' : sub-optimal decisions

Figure 9 illustrates the reversal of the quality-value relationship that can occur for some parameter values with substantially suboptimal decision making. The curves represent the variation of C with d' for an operation with $R = 2.0$ and with $p_c = 0.2$, as in Fig. 5. The lowest curve corresponds to optimal decisions, with $p^*_{opt} = R/(1+R) = 0.67$. The other curve corresponds to $p^*=0.25$, well below the optimal value and above the climatological probability. The cost of these forecasts to a decision-maker with $R = 2.0$ actually rises with increasing skill up to d' around 0.7, then decreases. This is a surprising result, and a numerical example follows to illustrate the effect. Note that this example does not represent real forecasts, although it is not unrealistic. The forecast sets have been devised

Fig. 9 Variation of relative cost with d' for suboptimal decisions; for p^* from 0.15 to 0.667. $p_c = 0.2, R = 2.0, p^*_{opt} = 0.667$.



to illustrate the way cost can deteriorate although accuracy may improve, as a result of sub-optimal decision-making.

Table 3 shows a set of yes/no forecasts (set A) artificially constructed to have a low level of skill corresponding to $d' = 0.193$, with $p^* = 0.25, p_c = 0.2$ and $N = 1000$. The corresponding values of h and f are 0.090 and 0.062, respectively. Table 4 shows another set of forecasts (set B) constructed with the same parameters except that $d' = 1.0$, a moderate level of skill, giving $h = 0.585$ and $f = 0.215$. Table 5 shows a number of measures of performance for both these sets of forecasts. Note that C in Table 5 is the relative cost to an operation with $R = 2.0$.

Forecast set B has better scores than set A on POD, FAR, Critical Success Index, Heidke score, Hansen & Kuipers' score and d' , and bias closer to 1.0. It has a slightly lower percentage correct, and higher false alarm rate, f . From most points of view B is a better set of forecasts than A. Nevertheless, a user with $R = 2.0$ would find set A, with lower skill, preferable, as his average relative cost would be 28.2 cents per forecast, whereas the cost of set B to this user would be 42.7 cents per forecast. The reason is that $R = 2.0$ implies that the false alarm penalty is twice as large as the miss penalty, and B has a substantially higher rate of false alarms. The fact that set B is superior to set A on most measures of forecast quality shows that the reversal of the quality/value relationship illustrated in Fig. 9 is not just an artefact of the SDT model. As noted in the previous section, it is associated with use of a suboptimal threshold.

The effect is more striking for rare events. The curves in Fig. 10 show C as a function of d' for forecasts of an event with $p_c = .01$, for an operation with $R = 2.0$. The lower (better) curve uses the optimal threshold of $p^* = 0.667$.

Table 3. Forecast set A, constructed to have $d' = 0.2$, $p^* = 0.25$, $p_c = 0.2$ and $N = 1000$.

		EVENT	
		NO	YES
FORECAST	NO	750	182
	YES	50	18

Table 4. Forecast set B, constructed to have $d' = 1.0$, $p^* = 0.25$, $p_c = 0.2$ and $N = 1000$.

		EVENT	
		NO	YES
FORECAST	NO	628	83
	YES	172	117

Table 5. Some performance measures for forecast sets A and B. Asterisk denotes better score. C is the relative cost for $R = 2.0$.

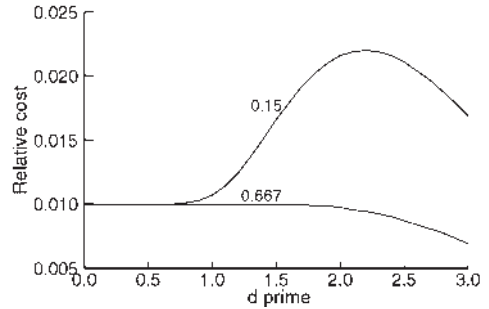
Measure	Forecast set A	Forecast set B
POD or h	.090	.585*
FAR	.735	.595*
f	.063*	.215
Bias	.340	1.445*
CSI	.072	.315*
Heidke score	.037	.317*
Hansen and Kuipers score	.028	.370*
d'	0.193	1.0*
Percent correct	76.8*	74.5
C	.282*	.427

For decisions made at the optimal threshold, C is constant at the no-skill level as skill increases from zero until d' reaches about 1.8, where the cost starts to decline slowly with increasing skill. The higher (worse) curve on Fig. 10 results when a threshold of $p^* = 0.15$ is used, well below the optimal threshold. The curves diverge for skill higher than $d' = 0.8$, and the cost to this operation of using forecasts produced at $p^* = 0.15$ increases with increasing skill up to the quite high level of $d' = 2.2$, where C is more than twice its value at $d' = 0.8$. This emphasises the central importance of decision threshold in forecasting rare events, and in using and verifying the forecasts.

Discussion

There appear to be two kinds of limits on the value of forecasts to decision-makers. One of these is a limit

Fig. 10 Variation of relative cost with d' for suboptimal decisions and a rare event, for $p^* = 0.15$ and $p^* = 0.667$. $p_c = 0.01$, $R = 2.0$, $p^*_{opt} = 0.667$.



on the range of threshold probabilities within which a forecasting system can possibly have some economic value to some users. At low levels of skill, this range may be much less than the whole probability interval $[0,1]$ and shrinks to p_c as skill tends to zero. Operations with R corresponding to p^*_{opt} outside this range will get no value from the forecasts, and would be better off using a strategy based on minimising expected cost using climatological probabilities, regardless of the forecasts.

The second kind of limit to the value of the forecasts is imposed by the level of skill itself, measured by d' in this study. There is a diffuse threshold on d' such that forecasting systems operating at lower levels of skill have no practical economic value, even though the actual value of d' may imply positive skill and the forecasts are used optimally. The location of this threshold on quality depends on the operation (R), the climate (p_c) and the level of skill (d').

The relationship between skill and expected cost is monotonic when the equal variance SDT model applies, skill is measured by d' , and decisions are made at the optimal decision threshold. Under these conditions, a system with higher d' is sufficient for a system with lower d' , in the sense of Ehrendorfer and Murphy (1988). One forecasting system is sufficient for another, in this sense, if its forecasts have equal or higher value for all users when applied optimally. Reversal of the quality-value relationship is possible when decisions are made at a threshold that is substantially different from the optimal threshold, so that increases in skill can result in worse decisions.

Even when there is no reversal of the quality/value relationship, decisions made at sub-optimal thresholds can be much more costly than those made optimally. Yes/no forecasts, produced at a single thresh-

old, must be sub-optimal to some degree for most users, and seriously sub-optimal for some users. Forecasts issued as probabilities provide users with information about the uncertainty of the event of interest, and enable different users to apply appropriate thresholds for decision making.

Acknowledgment

Ross Keith and two referees provided very useful comments on drafts of this paper.

References

- Anaman, K.A., Lelleyett, S.C., Drake, L., Leigh, R.J., Henderson-Sellers, A., Noar, P.F., Sullivan, P.J. and Thampapillai, D.J. 1998. Benefits of meteorological services: evidence from recent research in Australia. *Met. Apps.*, 5, 103-15.
- Chen, Y.-S., Ehrendorfer, M. and Murphy, A.H. 1987. On the relationship between the quality and value of forecasts in the generalized cost-loss ratio situation. *Mon. Weath. Rev.*, 115, 1534-41.
- Ehrendorfer, M. and Murphy, A.H. 1988. Comparative evaluation of weather forecasting systems: sufficiency, quality and accuracy. *Mon. Weath. Rev.*, 116, 1757-70.
- Finley, J.P. 1884. Tornado predictions. *American Meteorological Journal*, 1, 85-8.
- Green, D.M. and Swets, J.A. 1966. *Signal detection theory and psychophysics*. Wiley, New York. Reprinted by Kreiger, Huntington, NY, 1974. 479pp.
- Harvey, L.O. Jr, Hammond, K.R., Lusk, C.M. and Moss, E.F. 1992. The application of signal detection theory to weather forecasting behaviour. *Mon. Weath. Rev.*, 120, 863-83.
- Jolliffe, I.T. and Stephenson, D.B. (eds) 2003. *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley and Sons, England, 240 pp.
- Katz, R.W. and Murphy, A.H. 1987. Quality/value relationship for imperfect information in the umbrella problem. *The American Statistician*, 41, 187-9.
- Katz, R.W. and Murphy, A.H. (eds). 1997. *Economic Value of Weather and Climate Information*. Cambridge University Press, Cambridge, UK, 220 pp.
- Leigh, R.J. 1995. Economic benefits of terminal aerodrome forecasts at Sydney airport. Australia. *Met. Apps.*, 2, 239-47.
- Levi, K. 1985. A signal detection framework for the evaluation of probabilistic forecasts. *Organizational Behaviour and Human Decision Processes*, 36, 143-66.
- Mason, I. 1982a. A model for assessment of weather forecasts. *Aust. Met. Mag.*, 30, 291-303.
- Mason, I. 1982b. On scores for yes/no forecasts. *Preprints of papers delivered at the Ninth AMS Conference on Weather Forecasting and Analysis*, Seattle, Washington, 169-74.
- Mason, I. 1989. Dependence of the critical success index on sample climate and threshold probability. *Aust. Met. Mag.*, 37, 75-81.
- Mason, I. 2003. Binary events. In Ian T. Jolliffe and David B. Stephenson (eds). *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley and Sons, England, 240 pp.
- 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.H. 1992. Climatology, persistence and their linear combination as standards of reference in skill scores. *Mon. Weath. Rev.*, 120, 693-8.
- Murphy, A.H. 1994. Assessing the economic value of weather forecasts: an overview of methods, results and issues. *Met. Apps.*, 1, 69-73.
- Murphy, A.H. 1996. The Finley affair: a signal event in the history of forecast verification. *Weath. forecasting*, 11, 3-20.
- Murphy, A.H. and Ehrendorfer, M. 1987. On the relationship between the accuracy and value of forecasts in the cost-loss ratio situation. *Weath. forecasting*, 2, 243-51.
- Murphy, A.H. and Winkler, R.L. 1987. A general framework for forecast verification. *Mon. Weath. Rev.*, 115, 1330-8.
- Richardson, D.S. 2003. Economic value and skill. In Ian T. Jolliffe and David B. Stephenson eds. *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley and Sons, England, 240 pp.
- Swets, J.A. 1973. The relative operating characteristic in psychology. *Science*, 182, 990-1000.
- Swets, J.A. 1988. Measuring the accuracy of diagnostic systems. *Science*, 240, 1285-93.
- Swets, J.A. 1996. *Signal detection theory and ROC analysis in psychology and diagnostics: collected papers*. Lawrence Erlbaum and Associates Inc, 308pp.
- Swets, J.A. and Pickett, R.M. 1982. *Evaluation of diagnostic systems*. Academic Press, New York, 253 pp.
- Thompson, J.C. and Brier, G.W. 1955. The economic utility of weather forecasts. *Mon. Weath. Rev.*, 83, 249-54.
- von Neumann, J. and Morgenstern, O. 1947. *Theory of games and economic behaviour*. Princeton University Press, Princeton, NJ.
- Winkler, R.L. and Murphy, A.H. 1985. *Decision Analysis. In Probability, Statistics and Decision Making in the Atmospheric Sciences*, Allan H. Murphy and R.W. Katz, eds. Westview Press, Boulder and London, 545pp.
- Winterfeldt, D. von, and Edwards, W. 1986. *Decision Analysis and Behavioral Research*. Cambridge University Press, Cambridge, U.K., 604pp.
- Zhu, Y., Toth, Z., Wobus, R., Richardson, D. and Mylne, K. 2002. The economic value of ensemble-based weather forecasts. *Bull. Am. met. Soc.*, 83, 73-83.

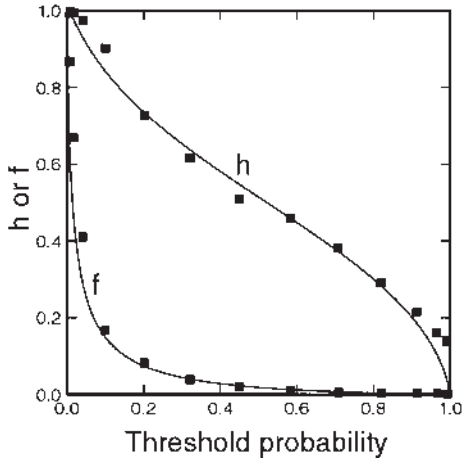
Appendix

Validation of the model

The validity of the modelling results shown in this paper rests on the observation that, when a number of sets of yes/no forecasts are produced by the same forecasting system at different p^* , the variation of h and f with p^* follows closely the prediction of the signal detection model with Gaussian distributions.

As a single illustrative example, Fig. 11 shows the modelled and empirical variation of h and f with p^* for a particular set of forecasts. The data points are values of h and f produced by setting threshold probabilities at successively increasing values in a set of daily probabilistic forecasts of rain at Canberra Airport. There were 3286 forecasts in the sample and the whole-sample relative frequency of rain was 0.109. The forecast probabilities were recalibrated by fitting a straight line to the reliability diagram with axes transformed to log odds, providing estimates of the 'true' probabilities corresponding to the forecast probabilities. The smooth curve shows h (upper curve) and f (lower curve) modelled by moving a

Fig. 11 Modelled and empirical variation of hit and false alarm rates with threshold probability. Data points generated from probabilistic forecasts for rain at Canberra Airport. Curve from fitted SDT model with separation of means 2.162 and ratio of variances 0.886. $p_c = 0.109$.



decision threshold through Gaussian distributions with means separated by 2.16 in units of the standard deviation of the f_N distribution, and a ratio of the variance of $f_N(x)$ to $f_Y(x)$ of 0.886. These values were estimated by fitting a straight line to the relative operating characteristic for the forecasts on 'bi-normal' axes using standard methods (Mason 1982a). The fit is evidently good. This illustrates the correspondence between SDT modelled and observed relationships between h, f and p^* .

Fig. 12 Modelled and empirical variation of relative cost with threshold probability for an operation using the forecasts in Fig. 11, with $R=1.0$. Data points from probabilistic forecasts, curve fitted using SDT model with parameter values as in Fig. 11.

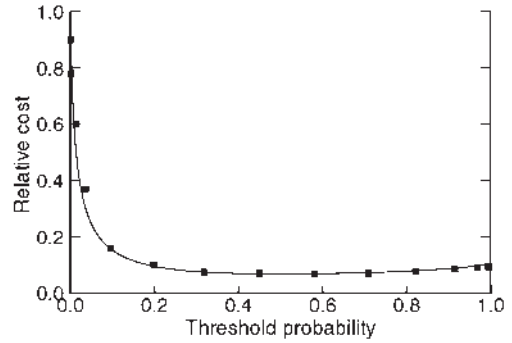


Figure 12 shows implied values of C , the relative cost of uncertainty, for $R=1$. The solid line is the prediction of the model with parameter values as specified in the previous paragraph, and the data points are values of C calculated from the same set of forecasts, plotted against the threshold probabilities.