Meteorol. Appl. 8, 209–219 (2001)

A skill score based on economic value for probability forecasts

D S Wilks, *Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, New York, USA*

An approach to evaluating probability forecasts for dichotomous events, based on their economic value over all possible cost/loss ratio decision problems, is proposed. The resulting Value Score (VS) curve shows non-dimensionalised relative economic value as a function of the cost/loss ratios for different decision-makers, over their full meaningful range. The VS curve is similar in terms of computational mechanics and graphical display to the Relative Operating Characteristic (ROC) curve, but the ROC curve is shown to be insensitive to either conditional or unconditional biases and thus to reflect potential rather than actual skill. The possibility of collapsing the VS curve into a single scalar score is addressed, and it is shown that the results can depend very strongly on the assumed distribution of cost/loss ratios in the community of forecast users.

1. Introduction

An important issue in forecast verification is the problem of evaluating probability forecasts for dichotomous events, for example occurrence of precipitation versus no precipitation. Almost universally, the accuracy of such forecasts is evaluated using the Brier Score (e.g., Murphy, 1997; Murphy & Daan, 1985; Stanski *et al.*, 1989; Wilks, 1995):

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2 \qquad (1)$$

where f_t is the probability forecast on occasion t, o_t is a binary variable indicating whether the event occurred ($o_t = 1$) or not ($o_t = 0$), and N is the number of forecasts being evaluated. Thus, the Brier Score is a mean-squared error measure for probability forecasts. Equation (1) is sometimes called the half-Brier score, because the score as originally introduced by Brier (1950) was larger by a factor of 2. Brier Scores (as well as other accuracy measures) are often expressed on a relative basis using skill scores of the form:

$$SS = \frac{BS - BS_{\text{clim}}}{BS_{\text{perf}} - BS_{\text{clim}}} = \frac{BS_{\text{clim}} - BS}{BS_{\text{clim}}} \qquad (2)$$

Here BS_{clim} indicates the Brier Score received by constant forecasts of the climatological probability, and BS_{perf} is the Brier Score for perfect forecasts. The second equality in equation (2) follows because $BS_{\text{perf}} = 0$. Equation (2) is often interpreted as the proportional improvement in accuracy between climatological and perfect forecasts.

Both BS and SS are scalar measures of the association between the forecasts and corresponding observations. As such, they necessarily (and in a somewhat arbitrary manner) collapse into a single number the full relationship between the forecasts and observations that is contained in the joint frequency distribution of the forecasts and observations, $p(f_i, o_j)$ (Murphy & Winkler, 1987). Here f_i, $i = 1, \ldots, I$, are the allowable forecasts; and o_j, $j = 0, 1$, are the possible observations for dichotomous outcomes. If forecast probabilities are rounded to tenths, then $I = 11$, with $f_1 = 0.0, f_2 = 0.1, \ldots,$ and $f_{11} = 1.0$. In that case the joint distribution $p(f_i, o_j)$ can be visualised as a table with two rows and 11 columns, each entry of which specifies one of the 22 joint probabilities in the distribution. Depending on the forecast format there may be more or fewer than $I = 11$ allowable forecasts, and for such cases the development in the following generalises directly.

Even for comparatively simple problems (small I and J) the information in $p(f_i, o_j)$ may be difficult to digest. In practice it has been found that two factorisations of this joint distribution may be helpful in understanding the performance characteristics of a set of forecasts. These factorisations are:

$$p(f_i, o_j) = p(o_j|f_i)p(f_i) \qquad (3a)$$

$$p(f_i, o_j) = p(f_i|o_j)p(o_j) \qquad (3b)$$

(a) *Calibration-refinement factorisation*. Equation (3a) is called the calibration-refinement factorisation, and consists of the I conditional distributions of observations conditional on each of the f_i (the 'calibration' distributions, $p(o_j|f_i)$), and one distribution quantifying the frequency of use of the f_i (the 'refinement' distribution $p(f_i)$).

(*b*) *Likelihood-base rate factorisation*. Equation (3b) is called the likelihood-base rate factorisation, and consists of conditional distributions of forecasts given subsequent observations (the 'likelihoods', $p(f_i | o_j)$), and the sample climatological or 'base rate' distribution $p(o_j)$.

For probability forecasts of dichotomous events, the calibration-refinement factorisation (equation (3a)) can be expressed graphically using a reliability diagram (e.g., Wilks, 1995), or an attributes diagram (Hsu & Murphy, 1986).These diagrams consist of plots of the conditional probabilities $p(o_1 | f_i)$ as a function of f_i (since the complementary probabilities $p(o_0 | f_i)$ are easily computed as $1 - p(o_1 | f_i)$), together with a histogram or other representation of the distribution $p(f_i)$.

From the perspective of a forecast user, the degree of forecast accuracy expressed in terms of mean squared error or some other scalar measure is of less interest than the value (economic or otherwise) that the user may derive from the forecasts. It is sometimes naively assumed that greater overall accuracy, as measured for example by *BS* or *SS*, will necessarily lead to greater value to users. Perhaps surprisingly, this is not necessarily the case: it is quite possible for forecasts of higher overall accuracy to impart less economic value for some decision problems (Ehrendorfer & Murphy, 1988, 1992; Murphy & Ehrendorfer, 1987; Murphy, 1997). Thus, summarising forecast performance in terms of *BS* or *SS* is deficient from the standpoint of forecast users, even if the users understand the structure of these statistics. At the opposite extreme, sophisticated forecast users with access to the joint distribution $p(f_i | o_j)$ (or one of its factorisations, for example in the form of a reliability diagram), and knowledge of their own loss functions can compute the expected economic value of the forecasts for their particular decision problems (e.g., Clemen, 1996; Wilks, 1997; Winkler & Murphy, 1985). However, typical forecast users would probably find a tabulation of $p(f_i | o_j)$ to be too complex and thus fairly uninformative.

This paper proposes an evaluation method for probability forecasts of dichotomous events which produces results in terms of economic value. This approach is directly relevant to users of forecasts, and will also be informative to forecast producers who are interested in the value of their forecasts to the users they serve. Because different forecast users can face quite different and sometimes quite complex decision problems, the method is based on a prototype decision problem known as the cost/loss ratio situation. While the cost/loss problem is an idealised analysis it nevertheless captures the essentials of optimal decision-making, and many real-world decision problems can be cast at least approximately in the cost/loss framework. The approach taken here is similar to that of Mylne (1999) and Thompson & Brier (1955), who characterised forecast performance in terms of economic value for particular cost/loss problems. Here the approach is extended to simultaneously consider economic value for all possible cost/loss problems, providing a more universal analysis that obviates the need for individual, user-by-user results.

2. The 2 × 2 cost/loss ratio problem

The evaluation of forecasts proposed here is based on an idealised and extremely simple model of optimal decision-making, called the cost/loss ratio problem. The cost/loss problem was apparently introduced first by Thompson (1952), and has been studied extensively since that time (e.g., Katz & Murphy, 1997; Murphy, 1977; Murphy & Ehrendorfer, 1987; Thompson & Brier, 1955). While it is the simplest possible quantitative model of optimal decision-making, it nevertheless can approximate many real-world problems to a reasonable degree (Roebber & Bosart, 1996; Wilks, 1997).

The basic premise of the cost/loss problem is that of a decision-maker faced with the uncertain prospect of some kind of adverse weather. The decision-maker is able to protect against the effects of adverse weather by paying a cost C, whereas occurrence of adverse weather without benefit of this protection results in a loss L. The protection cost C is incurred whenever the decision is made to protect, whether or not adverse weather occurs, but if adverse weather does not occur and no protective action is taken the economic impact on the enterprise is zero. Figure 1(*a*) shows the full loss function for this problem.

The optimal decision whether or not to take protective action will be the one resulting in the least expected (i.e., probability-weighted average) expense. The information available to the decision-maker is that adverse weather will occur with probability p. If protection is chosen, the cost C will be incurred with probability 1. If not, the loss L will be suffered with probability p. Thus, protection against the adverse weather is optimal whenever:

$$C < pL$$

or

$$\frac{C}{L} < p$$

That is, protection is optimal when the cost/loss ratio is less than the probability of adverse weather. Note that this result holds whether or not the entry in the lower-right cell of Figure1(*a*) is zero because such loss tables can be transformed to that form (e.g., Roebber & Bosart, 1996). Note that if $C > L$ it is never worthwhile to protect, while if $C < 0$ (i.e., the decision-maker is paid to protect) protection will always be optimal. Thus meaningful cost/loss problems are confined to $0 < C/L < 1$.

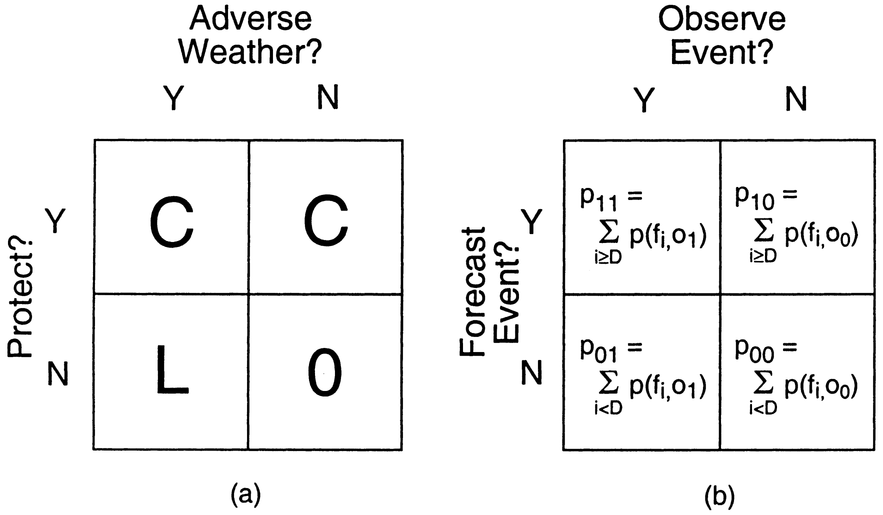

Figure 1. *(a) Loss function for the 2×2 cost/loss ratio situation. (b) Corresponding 2×2 verification table resulting from probability forecasts characterised by the joint distribution $p(f_i, o_j)$ being transformed to categorical forecasts according to a particular decision-maker's cost/loss ratio.*

If the decision-maker's only information is the climatological probability of adverse weather, π, then the optimal action will be either to always protect (if $C/L < \pi$) or never protect (if $C/L > \pi$). The expected expense given climatological information is thus:

$$EE_{\text{clim}} = \begin{cases} C \text{ if } C/L < \pi \\ \pi L \quad \text{otherwise} \end{cases}$$

On the other hand, if a decision-maker could have access to perfect forecasts, the decision to protect would be taken only on those occasions when adverse weather was about to occur, so that the expected expense would be:

$$EE_{\text{perf}} = \pi C$$

The expected expense when cost/loss decisions are based on imperfect probability forecasts for adverse weather depends on the performance characteristics of those forecasts. That is, the decision-maker pays the protection cost C whenever the forecast f is greater than C/L, incurs the loss L whenever $f < C/L$ but adverse weather nevertheless occurs, and suffers no loss when $f < C/L$ and the adverse weather does not occur. In effect, the decision-maker transforms each probability forecast to a categorical (i.e., yes/no) forecasts of adverse weather according to the magnitude of the forecast probability f in relation to a particular cost/loss ratio. The result is a categorical-forecast 2×2 contingency table that depends on the joint distribution of the probability forecasts and corresponding observations $p(f_i, o_j)$ as shown in Figure 1(*b*). Adverse weather is 'forecast' for all f_i, $i \geq D$, where D is the index of the smallest forecast probability that is larger than C/L. For example, if the forecast probabilities are rounded to tenths, with $I = 11$, a decision-maker with $C/L = 0.25$ would 'forecast' adverse weather (i.e., take protective action) when $i \geq 4$, since $f_4 = 0.3$. The proportion of occasions when this would occur in advance of adverse weather would be $p_{11} = \Sigma_{i \geq D} p(f_i, o_1)$, and the proportion of occasions when protective action would be taken needlessly would be $p_{10} = \Sigma_{i \geq D} p(f_i, o_0)$. The probability of adverse weather following a 'no' forecast, p_{01}, and the probability of no adverse weather following a 'no' forecast, p_{00}, are defined similarly as indicated in Figure 1(*b*). Thus the expected expense faced by a decision-maker consists of the entries in Figure 1(*a*) weighted by the probabilities in Figure 1(*b*), or:

$$EE_{\text{f}} = (p_{11} + p_{10})C + p_{01}L \tag{4a}$$

giving:

$$EE_{\text{f}} = C\sum_{j=0}^{1}\sum_{i \geq D} p(f_i, o_j) + L\sum_{i<D} p(f_i, o_1) \tag{4b}$$

This expected expense depends on the particular user's cost and loss, on the cost/loss ratio through its relationship to the decision threshold D, and on the performance characteristics of the forecasts on which the decisions are based, as encapsulated in the joint distribution of forecasts and observations $p(f_i, o_j)$.

3. The Value Score

3.1. Definition

Conventionally the economic value of forecasts is computed as the difference of expected expenses between some baseline (often climatological) information and the expected expenses given the forecasts under consideration, $EE_{\text{clim}} - EE_{\text{f}}$ (e.g., Wilks, 1997). In raw form this difference would be unsuitable as a general measure of forecast performance because of its dependence on particular values of C and L. However, normalising this difference by the corresponding value of perfect information yields a score in the form of equation (2) for the economic value that depends only on the cost/loss ratio:

$$VS = \frac{EE_{\mathrm{f}} - EE_{\mathrm{clim}}}{EE_{\mathrm{perf}} - EE_{\mathrm{clim}}} \tag{5a}$$

giving:

$$VS = \begin{cases} \dfrac{\frac{C}{L}\left(p_{11} + p_{10} - 1\right) + p_{01}}{\frac{C}{L}\left(\pi - 1\right)} & \text{if } \dfrac{C}{L} < \pi \quad \text{(5b)} \\[2ex] \dfrac{\frac{C}{L}\left(p_{11} + p_{10}\right) + p_{01} - \pi}{\pi\left(\frac{C}{L} - 1\right)} & \text{if } \dfrac{C}{L} > \pi \quad \text{(5c)} \end{cases}$$

This Value Score (VS) can be interpreted as the expected economic value of the forecasts of interest as a fraction of the value of perfect forecasts relative to climatological forecasts, or as a percentage improvement in value between climatological and perfect information; as a function of the cost/loss ratio, for $0 < C/L < 1$. Its maximum value is $VS = 1$ for perfect forecasts, while $VS = 0$ when the forecasts impart no more value that does optimal use of the climatological probability π. The VS may be negative in cases where the cost/loss decision-maker would be better served by ignoring the forecasts, and acting optimally according to π. Note that in the restricted but desirable case of well-calibrated ('reliable') forecasts (i.e., $p(o_1|f_i) = f_i, \forall\ i$) the minimum VS is zero: perfectly calibrated forecasts are always at least as useful as the climatological information (Murphy, 1977). This occurs because economic value can be non-zero only if some decisions are changed as a result of knowledge of a forecast rather than the climatological probability (e.g., Winkler & Murphy, 1985), and these changes are made only if the expected expense (equation (4)) according to that probability is more favorable. Therefore, equation (5) can be negative only if forecast probabilities do not correspond to event relative frequencies. Thus the VS is sensitive to unconditional and conditional biases, as well as to non-systematic deviations from perfect calibration, in a way that reflects the losses suffered by decision-makers who take the forecasts at face value.

In practice the verification data will consist of a finite number I of allowable forecast probabilities f_i, so that VS can be evaluated at I+1 values of C/L. The joint distribution $p(f_i, o_j)$ can be partitioned non-trivially as shown in Figure 1(b) in I–1 distinct ways, by taking the decision index D between all adjacent pairs of forecasts, while interpreting the corresponding cost/loss ratio for decision-makers whose optimal actions would follow the resulting contingency table $(C/L)_i = (f_i + f_{i+1})/2$. In addition, if $C/L = 0$ then the optimal decision-maker always protects (because protection is free) so that $p_{01} = p_{00} = 0$. Since this is also the optimal action under climatological information when $C/L < \pi$, the resulting $EE_{\mathrm{f}} = EE_{\mathrm{clim}}$ so that $VS = 0$. Similarly if $C/L = 1$ the decision-maker should never protect, which is the same optimal action as under climatological information since $C/L > \pi$, leading to $p_{11} = p_{10} = 0$ and again $VS = 0$.

Different forecast users with different decision problems will realise different levels of economic value from optimal use of the same forecasts. Thus the VS is best expressed graphically as a function of C/L. Note that ideas similar to the VS have been proposed both by Mylne (1999) and by Thompson & Brier (1955), but for individual C/L levels rather than over the full meaningful range of C/L.

3.2. Illustration

It is useful to illustrate the VS using hypothetical sets of forecasts. Twelve such forecasts sets will be considered, which have been constructed as all possible combinations of the four calibration functions $p(o_1|f_i)$ in Figure 2(a) with the three refinement distributions $p(f_i)$ in Figure 2(b). The two panels in Figure 2 comprise reliability diagrams for these forecasts, and constitute full specifications of the respective joint distributions $p(f_i, o_j)$ through equation (3a). The climatological probability for these forecasts is $\pi = 0.3$, except $\pi \approx 0.2$ for the over-forecast case, and $\pi \approx 0.4$ for the under-forecast case.

Figure 2(a) indicates that none of the forecasts are well-calibrated, because none of the four calibration functions coincide with the dashed 1:1 line, i.e. $p(o_1|f_i) \neq f_i$, except in two cases when $f_i = \pi$. The 'under-confident' forecasts have been constructed using $p(o_1|f_i) = 2\,f_i - \pi$ and thus show conditional bias: the larger probabilities are under-forecast and the smaller probabilities are over-forecast. The 'over-confident' forecasts have the opposite conditional biases, since $p(o_1|f_i) = (f_i + \pi)/2$. This over-confident calibration distribution is an interesting case because it is equidistant between the dashed perfect calibration line and the 'no resolution' line $p(o_1|f_i) = \pi$ (which is not plotted), and thus these forecasts have zero skill according to the Brier Skill Score in equation (2) (Hsu & Murphy, 1986; Wilks, 1995). The remaining two calibration functions, for which $p(o_1|f_i) = f_i \pm 0.1$, exhibit unconditional biases, or consistent under- and over-forecasting. In all these cases $p(o_1|f_i)$ has been constrained to be on the unit interval.

The refinement distributions $p(f_i)$ are represented here using beta distributions:

$$p(f) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}(f)^{a-1}(1-f)^{b-1}, 0 \le f \le 1 \tag{6}$$

The two distribution parameters a and b are positive real numbers, and $\Gamma(\ )$ denotes the gamma function.

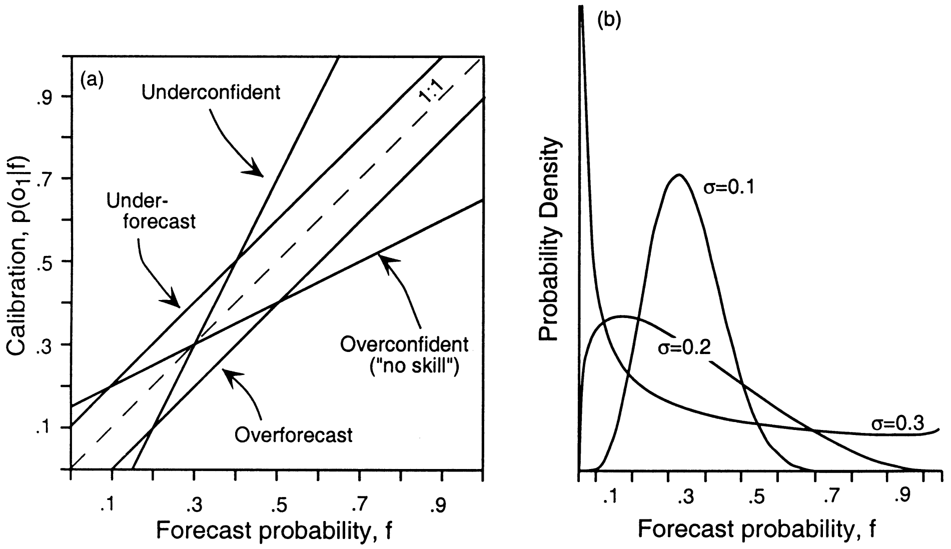

Figure 2. *(a) Calibration functions and (b) refinement distributions for hypothetical forecasts used to illustrate the VS in Figure 3. The 'perfect reliability' line is dashed in (a). The values of σ in (b) refer to the standard deviation of the beta distribution (see equation (6)). In all cases the average forecast is 0.30.*

Figure 2(*b*) shows three beta distributions with mean (μ) of 0.3 and standard deviation (σ) of 0.1, 0.2, and 0.3 where:

$$\mu = \frac{a}{a+b}$$

$$\sigma = \frac{1}{(a+b)}\left(\frac{ab}{a+b+1}\right)^{1/2}$$

For σ = 0.1 the forecasts resemble long-lead seasonal forecasts (Wilks, 2000), deviate only rarely from their mean value, and thus have the least potential to be informative. For σ = 0.3 the forecasts resemble short-range weather forecasts (Murphy & Wilks, 1998), are comparatively sharp, and forecast large deviations from the mean value fairly frequently. The discrete probabilities $p(f_i)$ have been obtained by integrating these functions over the eleven intervals shown on the abscissa, with the nominal forecast probabilities f_i taken to be the interval midpoints 0.025, 0.10, . . . , 0.90, and 0.975.

The solid lines in Figure 3 show *VS* curves for the refinement distributions in Figure 2(*b*); for (a) the under-confident, (b) the over-confident, (c) under-forecast, and (d) the over-forecast calibration functions in Figure 2(*a*). Clearly the under-confident forecasts in (a) are the best of these, as they are able to forecast a comparatively large number of events with either probability zero or probability one, which happens more frequently for the larger σ. Figure 3(*b*) shows that the over-confident, or 'no skill' forecasts can impart positive economic value for decision-makers whose action threshold is near the climatological probability π = 0.3. These are decision-makers, who have been called 'ideal users' (Gandin *et al.*, 1992), whose best choice is least obvious given only the climatological information and so benefit most from forecasts. For other decision-makers these forecasts impart negative value, since users with small cost/loss ratios are persuaded not to protect on occasions when the probability of adverse weather $p(o_1|f_i)$ is actually larger than *C*/*L* for the smaller f_i, and users with large cost/loss ratios are induced to incur protection costs when $p(o_1|f_i) < C/L$ for the larger f_i. These negative economic values are magnified for larger σ because these extreme f_i occur more frequently in those cases. Similarly, forecasts with simple unconditional biases (Figures 3(*c*) and 3(*d*)) show generally positive *VS*, except for those decision-makers who are badly misled because of their cost/loss ratios. Decision-makers with very small *C*/*L* will protect too rarely given the under-forecasting of the small probabilities (Figure 3(*c*)), and decision-makers with large *C*/*L* will protect too often when the largest probabilities are over-forecast (Figure 3(*d*)).

The dashed curves in Figure 3 show the corresponding results after recalibration of the forecasts: all the forecast characteristics displayed in Figure 2 are retained, except that the forecast probabilities f_i are relabelled to be consistent with the calibration functions. That is, the recalibrated forecast probabilities are defined as $f_i \equiv p(o_1|f_i)$ so that the calibration functions for the f_i are coincident with the dashed 1:1 line (with f_i constrained to lie within the unit interval), and the abscissa of Figure 2(*b*) is appropriately rescaled while the integrated probabilities in each of the 11 bins remain unchanged. In each case the *VS* responds to removal of bias with a higher score, with the minimum *VS* now being zero. Note that in Figure 3(*a*) the dashed curves are below the solid curves in the vicinity of *C*/*L* = π solely because of the discretisation of the forecasts: after recalibration of the under-confident forecasts only the revised probabilities f_i = 0.05, 0.20, 0.40, 0.60, 0.80, and 0.95 are 'issued', so the *VS* curve is only coarsely defined.

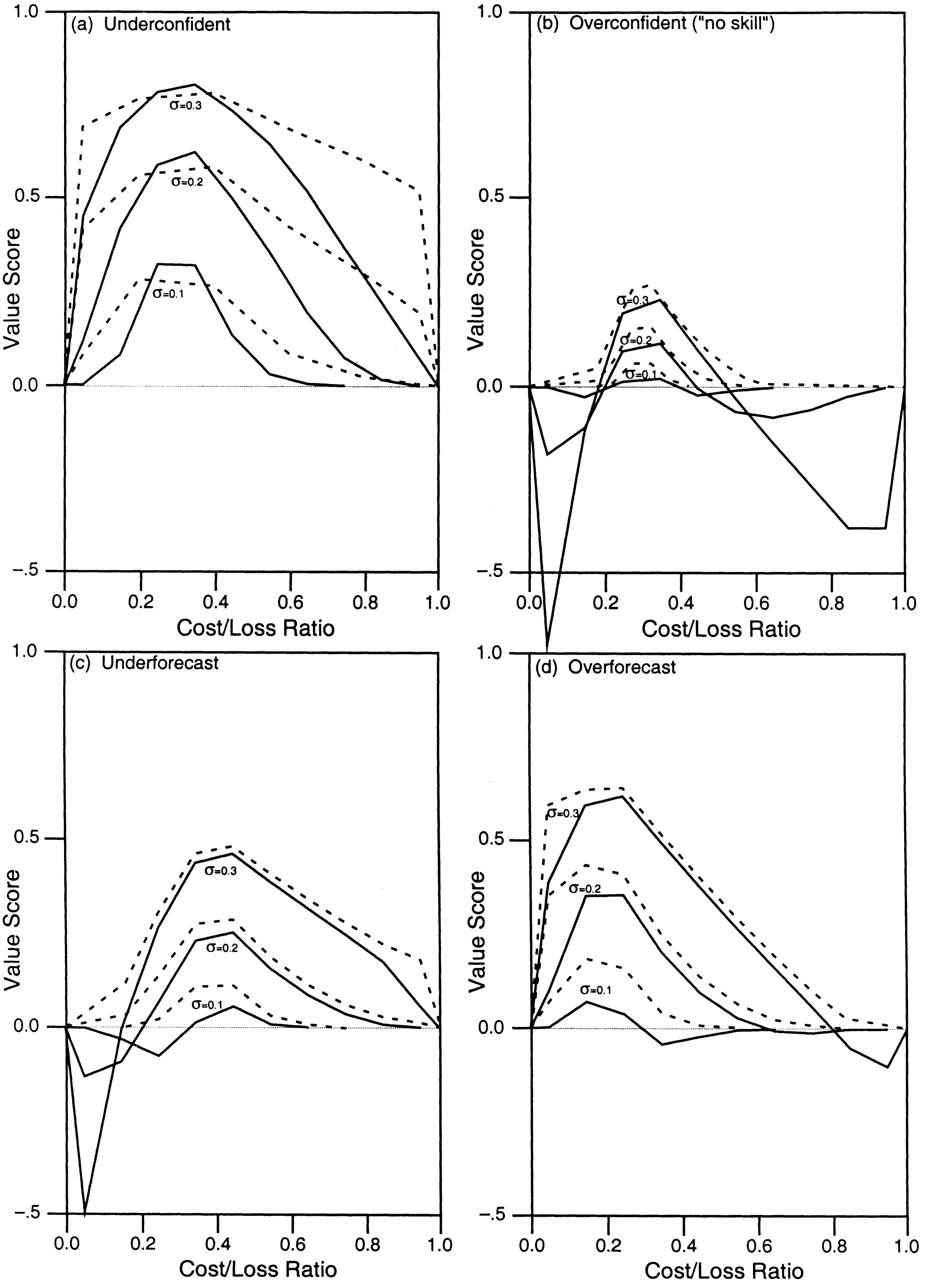

Figure 3. *VS curves for the (a) under-confident, (b), over-confident, (c) under-forecast, and (d) over-forecast calibration functions in Figure 2(a); for each of the three refinement distributions shown in Figure 2(b). Solid lines show VS for the forecasts as portrayed in Figure 2. Dashed lines show the improvements resulting from recalibrating (i.e., correcting the biases in) the forecasts.*

4. Connections to Relative Operating Characteristic (ROC)

The mechanics of calculating *VS* curves as described above overlap in part with those for the Relative Operating Characteristic (ROC) curve for probability forecasts. The ROC curve has a long history in psychology (e.g., Swets, 1973) as an element of signal detection theory. It has been described in the meteorological literature by Harvey *et al.* (1992), Mason (1982) and Stanski *et al.* (1989), and is currently used for verifying probability forecasts (e.g., Buizza *et al.*, 1999; Mason & Graham, 1999).

Both ROC and VS curves are constructed by transforming the joint distribution of forecasts and observations into a series of 2×2 contingency tables as shown in Figure 1(*b*), by imposing *I*-1 decision thresholds between adjacent pairs of the *I* forecasts, f_i. Computation of the points on the *VS* curve then proceeds through equations (4) and

(5) for each of the contingency tables. For the ROC curve each of the contingency tables is summarised by computing the signal-detection hit rate and the signal-detection false alarm rate, defined as:

$$SD\ \ HR = \frac{p_{11}}{p_{11} + p_{01}} = \frac{p_{11}}{\pi}$$

and

$$SD\ \ FAR = \frac{p_{10}}{p_{10} + p_{00}} = \frac{p_{10}}{1 - \pi}$$

Note that these statistics are different from the conventional meanings of hit rate and false alarm rate (e.g., Wilks, 1995), and so are distinguished here with the prefix '*SD*' for 'signal detection'. The *SD HR* and *SD FAR* are conditional probabilities of a 'yes' forecast given either occurrence or non-occurrence of the event, and so relate to the likelihood-base rate factorisation (equation (3b)). As such they reflect discrimination exhibited by the forecasts (Murphy, 1997; Murphy & Winkler, 1987), without regard to the particular numerical values of the f_i.

The ROC curve consists of a plot of *SD HR* as a function of *SD FAR*. As is the case for the *VS* curve, there are *I*–1 nontrivial points; plus the two points corresponding to never forecasting the event (*SD HR* = *SD FAR* = 0), and always forecasting the event (*SD HR* = *SD FAR* = 1). For example, Figure 4 shows ROC curves for the forecasts whose *VS* curves are plotted in Figures 3(*a*) and 3(*b*). The ROC curves for better forecasts are more strongly bowed and closer to the upper left-hand corner, approaching the limit of perfect forecasts for which *SD FAR* = 0 and *SD HR* = 1. The ROC curves for poorer forecasts are closer to the dashed *SD HR* = *SD FAR* line, which corresponds to forecasts with no discrimination.

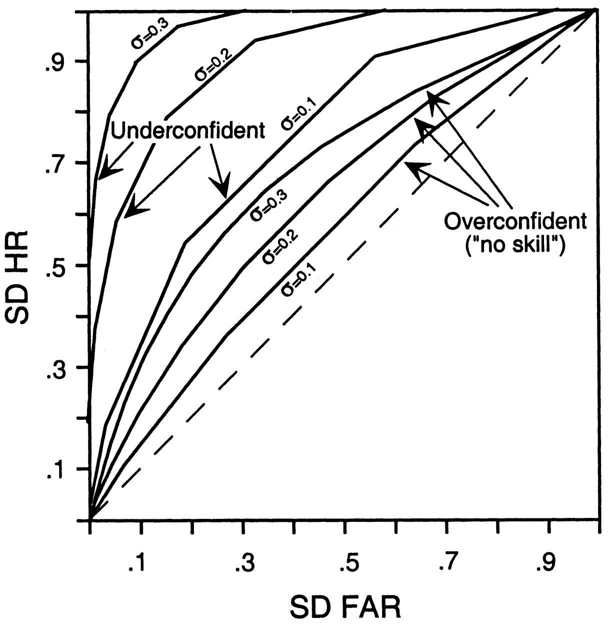

Figure 4. *ROC curves for the under-confident and over-confident forecasts, corresponding to the VS curves in Figures 3(a) and 3(b). The dashed curves for recalibrated forecasts are coincident with the solid curves because the ROC does not respond to forecast biases.*

Note that the counterparts of the dashed curves in Figures 3(*a*) and 3(*b*), corresponding to forecasts whose conditional biases (evident in Figure 2(*a*)) have been corrected, are not visible in Figure 4. The ROC curves for the pairs of original and bias-corrected forecasts are identical. Because the actual numerical value of the forecast probabilities f_i are not considered in the computations for the ROC curve, systematic errors in the f_i (e.g., conditional or unconditional biases as shown in Figure 2(*a*)) do not affect either *SD HR* or *SD FAR*. The result is that the ROC curve is blind to such biases, and forecast improvements that alleviate them will not be credited as improvements by verification systems based on the ROC curve. Thus, in common with other bias-insensitive forecast performance measures such as correlations (Murphy & Epstein, 1989), the ROC curve is best interpreted as reflecting *potential* rather than actual skill. ROC scores are reportedly under consideration by the World Meteorological Organization as the standard for reporting skill of probabilistic weather and climate forecasts (Mason & Graham, 1999). Their evident insensitivity to forecast biases indicates that this would be a poor choice.

5. An instructive example

This section compares verification results for probability-of-precipitation forecasts at Las Vegas (36° N, 115 ° W), a desert location, for the period April 1980 through March 1987. The forecasts consist of objective statistical forecasts from Model Output Statistics (MOS) (Glahn & Lowry, 1972), and the corresponding subjective probability forecasts (SUB) made by US National Weather Service forecasters (who generally had access to the MOS forecasts before formulating their own). In both cases the event being forecast is the occurrence of at least 0.254 mm of precipitation over a 12-hour period. The results are tabulated separately for the 'cool' (October through March) and 'warm' (April through September) seasons.

Figure 5 shows *VS* curves summarising the forecasts. These are interesting cases to examine because for both the cool (Figure 5(*a*)) and warm (Figure 5(*b*)) seasons there are ranges of cost/loss ratios for which each of the two forecast sources are more valuable. In the cool season, decision-makers with *C*/*L* < 0.2 would prefer the SUB forecasts, while those with *C*/*L* > 0.4 would prefer the MOS forecasts. Similarly in the warm season forecast users with smaller cost/loss ratios would prefer the MOS forecasts while users with larger cost/loss ratios would prefer the SUB forecasts. The warm-season case is especially interesting since the MOS system never forecast probabilities greater than 0.5; and while the human forecasters were able to successfully forecast some larger

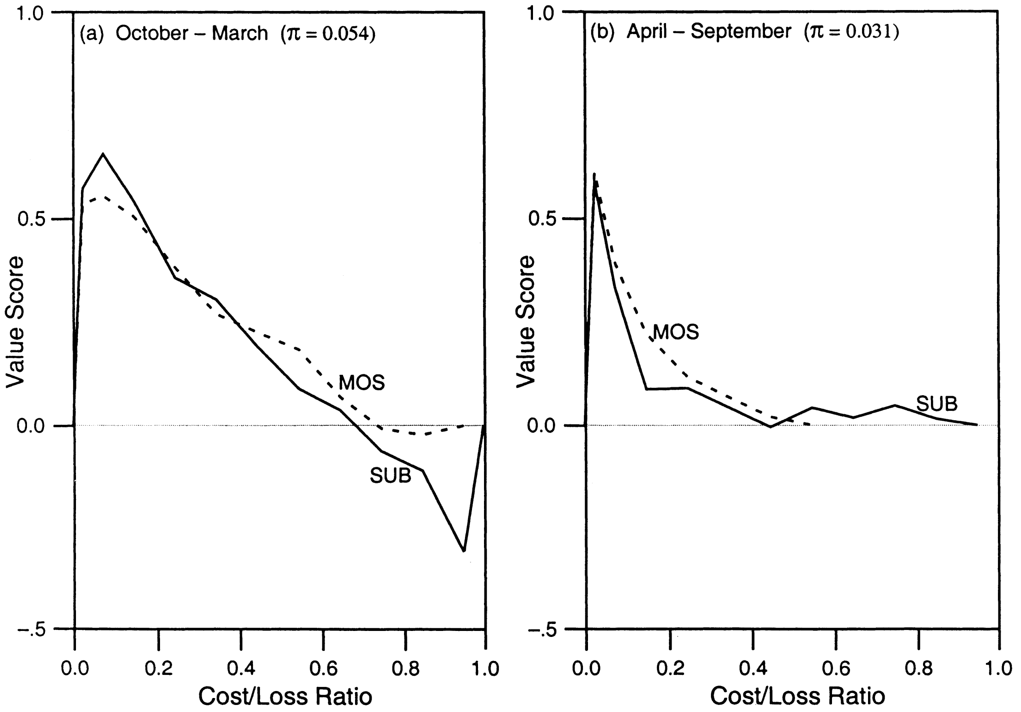

Figure 5. *VS curves for objective (MOS) and subjective (SUB) probability of precipitation forecasts at Las Vegas, Nevada, for the period April 1980–March 1987.*

probabilities, this was apparently done at the expense of forecast quality for the smaller probabilities.

It is common practice to attempt to summarise the performance characteristics of a set of forecasts using a single number, for example *SS* (equation (2)), or the area under the ROC curve, A_{ROC}. Table 1 compares these two statistics for the Las Vegas data. For both the cool and warm seasons, the MOS forecasts are judged to be better according to the *SS* statistic, while the SUB forecasts are superior in both cases according to the A_{ROC}. This conflicting result should not be especially surprising in view of the fact that compressing a high-dimensional verification problem into a single scalar value will necessarily obscure some important information (Murphy, 1991; Murphy & Winkler, 1987).

In each case here there are some forecast users for whom the 'inferior' (according to one or the other of *SS* or A_{ROC}) forecasts are actually superior. These are examples of quality/value reversals (Ehrendorfer & Murphy, 1988; Murphy & Ehrendorfer, 1987) that can easily occur when the evaluation of forecast quality is collapsed into a scalar measure. It is sometimes possible to demonstrate clear superiority of one set of forecasts over another in terms of economic value for all users (Krzysztofowicz & Long, 1991; Murphy & Ye, 1990), a condition which is known as sufficiency. However, often when comparing closely competitive forecasts (as in the present case) neither will be clearly superior for all users. Since cost/loss decision-makers are a subset of all forecast users, *VS* curves for forecasts that are unambiguously superior (i.e., sufficient) will be everywhere larger than the *VS* curves for their inferiors (e.g., the pairs of dashed and solid lines in Figure 3), although the converse might not be true.

Table 1. *Comparisons of skill score (SS, equation (2)) and the area under the ROC curve (A_{ROC}) for objective (MOS) and subjective (SUB) probability of precipitation forecasts at Las Vegas, Nevada, April 1980–March 1987*

Forecast of probability of precipitation	*SS*	A_{ROC}
October–March		
MOS	0.316	0.928
SUB	0.313	0.931
April–September		
MOS	0.142	0.905
SUB	0.108	0.920

Particular forecast users will want to examine *VS* curves at cost/loss ratios relevant to their own decision problems. On the other hand, it is possible that a forecaster or forecasting agency might want to collapse a *VS* curve into a single number that integrates relative economic value of the forecasts over a collection of users. While the spectrum of cost/loss ratios in a particular user community will be difficult to evaluate, one approach to exploring the effects of different distributions of cost/loss ratios within the user population is to model their relative frequencies using beta distributions (Murphy, 1969; Roebber & Bosart, 1996), i.e., equation (6) with C/L rather than f as its argument.

Figure 6 shows differences (SUB – MOS) in weighted-average *VS*, where the weights characterising the relative importance of different cost/loss ratios are provided by beta distributions with $0 < a < 8$ and $0 < b < 8$. Figures 6(*a*) and 6(*b*) contain results for the Las Vegas cool- and warm-season forecasts, respectively. The zero line is heavy, the contour interval is 0.02, and dashed contours indicate negative values (i.e., MOS forecasts deemed more valuable in aggregate). The raw integrated *VS* results underlying these figures range from near zero to – ¼ for large *a* and small *b*, to approximately ½ for small *a* and large *b*. Figure 6(*c*) shows thumb-nail sketches of selected beta distributions for perspective. The case of $a = b = 1$ yields the uniform distribution, for which the *VS* at all cost/loss ratios are weighted equally. This seemingly attractive choice is equivalent to use of the *BS* (equation (1)) to evaluate the forecasts, since the expected value of EE_f according to the uniform distribution is a linear function of *BS* (Murphy 1966), and therefore so also is the expected value of *VS*. As could have been anticipated, the MOS forecasts have a greater integrated *VS* when the larger cost/loss ratios are emphasised in the cool season, and when the smaller cost/loss ratios are emphasised during the warmer months. Figure 6 shows very clearly that the result of collapsing the *VS* curve into a single score depends very strongly on the weighting function that is assumed to represent the spectrum of forecast users.

6. Conclusions

This paper has proposed a method to evaluate probability forecasts for binary predictands, based on the simple 2×2 cost/loss ratio problem. The result is a curve representing the variation of (scaled) economic value as a function of the cost/loss ratio; which respects the fact that different decision-makers will derive different benefits from a given set of forecasts, and that attempting to express forecast quality with a single number will fail to capture this variation. The *VS* curve is an extension of the approaches of Mylne (1999) and Thompson & Brier (1955), who considered forecast evaluation in terms of economic value for particular, individual cost/loss ratios. The approach is a rational way to approach verification for weather variables that directly affect human enterprises, and would probably be less attractive for forecasts of variables such as geopotential heights. Note that the 2×2 cost/loss ratio situation is essentially a toy decision problem which neglects such complications as multiple decision options, multiple relevant meteorological events, related sequential decisions, and different attitudes toward risk (e.g., Wilks, 1997), although a large number of decision problems

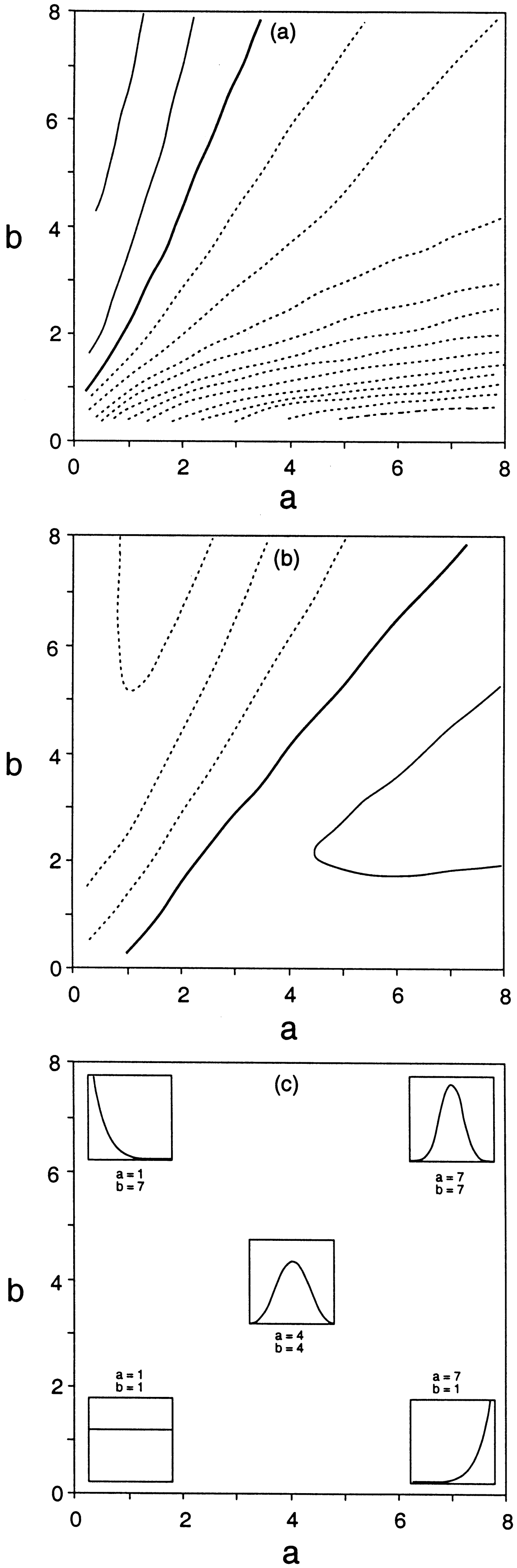

Figure 6. *(Left) Differences (SUB – MOS) in weighted-average VS, as a function of the beta distribution parameters a and b used to define relative importances of cost/loss ratios, for (a) cool season and (b) warm season Las Vegas forecasts of probability of precipitation. The heavy contours indicate zero difference, the contour interval is 0.2, and negative contours are dashed. (c) Illustration of selected beta densities to aid interpretation.*

can be viewed at least approximately in terms of the simple cost/loss ratio situation. While it might be possible to devise similar scores based on more complex and realistic decision problems, these would either pertain to quite specific user subsets, or increase the dimensionality of the results (e.g., Figures 3 and 5) in ways that would make concise display difficult.

The *VS* is non-negative for calibrated forecasts (i.e., $p(o_1|f_i)= f_i$). To the extent that there is miscalibration (e.g., conditional or unconditional biases) the score reflects penalties endured by decision-makers who take the forecasts at face value. It is sometimes assumed in this context that the forecast user can and will recalibrate forecasts based on past performance of those forecasts (Krzysztofowicz, 1992; Krzysztofowicz & Long, 1990), a process which is equivalent to the bias corrections yielding the dashed curves in Figure 3. However, it is arguably unreasonable to expect real-world decision-makers to perform this task when it is the forecaster or forecasting agency who should have the information required to correct forecast miscalibration. Such biases can be corrected in operational settings (Atger, 1999; Eckel & Walters, 1998), and it seems reasonable and desirable for a verification system to reward such efforts.

Certain of the mechanics of the computation of the *VS* parallel those for the Relative Operating Characteristic (ROC), and both are displayed as functions rather than reported as scalar scores. However, it has been shown that ROC curves do not penalise forecast biases, and thus are better regarded as expressing potential rather than actual skill. Similarly the area under the ROC curve will not be affected by systematic forecast biases. Any number of scalar scores can also be constructed through statistical expectations of the *VS* curve with respect to probability densities over user cost/loss ratios, but the result can depend very strongly on the nature of the assumed user community.

Finally, the *VS* could also be used to compare probabilistic and non-probabilistic forecasts for binary events. The *VS* curve is constructed by transforming probabilistic forecasts into a categorical yes/no protection decision, according to the magnitude of the forecast probability in relation to a decision-maker's cost/loss ratio. Similarly, categorical forecasts effectively transform probability judgements of the forecaster into yes/no forecasts (that may reflect a presumed decision threshold relevant to forecast users), effectively shifting the responsibility for decision-making from the users to the forecaster (Thompson & Brier,1955). Within the framework presented here, categorical forecasts would yield a single verification table (Figure 1(*b*)), regardless of a particular decision-maker's *C*/*L*. Different decision-makers would realise different economic benefits (or losses) from such forecasts, with magnitudes depending on the differences between their cost/loss ratios and the forecaster's single yes/no threshold. These scaled economic values could also be portrayed graphically as a function of *C*/*L*, and would be equal to (for *C*/*L* equal to the forecaster's yes/no threshold) or less than the value achieved through optimal forecast use (Krzysztofowicz, 1983; Murphy, 1977) as indicated by the *VS* curve.

Note added in proof: Recently Richardson (2000) has described a procedure similar to that presented here, but which assumes in effect that the forecasts will be calibrated.

Acknowledgements

I thank Ken Mylne for stimulating my interest in this topic. This paper is dedicated to the memory of Allan Murphy.

References

Atger, F. (1999). The skill of ensemble prediction systems. *Mon. Wea. Rev.*, **127:** 1941–1953.

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

Buizza, R., Hollingsworth, A., Lalaurette, F. & Ghelli, A. (1999). Probabilistic predictions of precipitation using the ECMWF ensemble prediction system. *Weather and Forecasting*, **14:** 168–189.

Clemen, R. T. (1996). *Making Hard Decisions: an Introduction to Decision Analysis.* Second edition, Duxbury Press, Belmont, California.

Eckel, F. A. & Walters, M. K. (1998). Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Weather and Forecasting*, **13:** 1132–1147.

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

Ehrendorfer, M. & Murphy, A. H. (1992). Evaluation of prototypical climate forecasts: the sufficiency relation. *J. Climate*, **5:** 876–887.

Gandin, L. S., Murphy, A. H. & Zhukovsky, E. E. (1992). Economically optimal decisions and the value of meteorological information. In *Preprints, 5th International Meeting on Statistical Climatology*, 22–26 June 1992, Toronto, Canada. J64–J71.

Glahn, H. R. & Lowry, D. A. (1972). The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteorol.*, **11:** 1203–1211.

Harvey, L. O. Jr., Hammond, K. R., Lusk, C. M. & Mross, E. F. (1992). The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev.*, **120:** 863–883.

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

Katz, R. W. & Murphy, A. H. (1997). Forecast value: prototype decision-making models. In *Economic Value of Weather and Climate Forecasts*, R. W. Katz and A. H. Murphy, eds., Cambridge University Press, 183–217.

Krzysztofowicz, R. (1983). Why should a forecaster and a decision maker use Bayes theorem. *Water Resources Res.*, **19:** 327–336.

Krzysztofowicz, R. (1992). Bayesian correlation score: a utilitarian measure of forecast skill. *Mon. Wea. Rev.*, **120:** 208–219.

Krzysztofowicz, R. & Long, D. (1990). Fusion of detection probabilities and comparison of multisensor systems. *IEEE Transactions on Systems, Man, and Cybernetics*, **20:** 665–677.

Krzysztofowicz, R. & Long, D. (1991). Forecast sufficiency characteristic: construction and application. *Int. J. Forecasting*, **7:** 39–45.

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

Mason, S. J. & Graham, N. E. (1999). Conditional probabilities, relative operating characteristics, and relative operating levels. *Weather and Forecasting*, **14:** 713–725.

Murphy, A. H. (1966). A note on the utility of probabilistic predictions and the probability score in the cost-loss ratio situation. *J. Appl. Meteorol.*, **5:** 534–537.

Murphy, A. H. (1969). Measures of the utility of probabilistic predictions in cost-loss ratio decision situations in which knowledge of the cost-loss ratios is incomplete. *J. Appl. Meteorol.*, **8:** 863–873.

Murphy, A. H. (1977). The value of climatological, categorical and probabilistic forecasts in the cost-loss situation. *Mon. Wea. Rev.*, **105:** 803–816.

Murphy, A. H. (1991). Forecast verification: its complexity and dimensionality. *Mon. Wea. Rev.*, **119:** 1590–1601.

Murphy, A. H. (1997). Forecast Verification. In *Economic Value of Weather and Climate Forecasts*, R. W. Katz and A. H. Murphy, eds., Cambridge University Press, 19–74.

Murphy, A. H. & Daan, H. (1985). Forecast evaluation. In *Probability, Statistics and Decision Making in the Atmospheric Sciences*, A. H. Murphy and R.W. Katz, eds., Westview Press, Boulder, Colorado, 379–437.

Murphy, A. H. & Ehrendorfer, M. (1987). On the relationship between the accuracy and value of forecasts in the cost-loss ratio situation. *Weather and Forecasting*, **2:** 243–251.

Murphy, A. H. & Epstein, E. S. (1989). Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.*, **117:** 572–581.

Murphy, A. H. & Wilks, D. S. (1998). A case study of the use of statistical models in forecast verification: precipitation probability forecasts. *Weather and Forecasting*, **13:** 795–810

Murphy, A. H. & Winkler, R. L. (1987). A general framework for forecast verification. *Mon. Wea. Rev.*, **115:** 1330–1338.

Murphy, A. H. & Ye, Q. (1990). Comparison of objective and subjective precipitation probability forecasts: the sufficiency relation. *Mon. Wea. Rev.*, **118:** 1783–1792.

Mylne, K. R. (1999). The use of forecast value calculations for optimal decision-making using probability forecasts. In *Preprints, 17th Conference on Weather Analysis and Forecasting*, American Meteorological Society, Boston, Massachusetts, 235–239.

Richardson, D. S. (2000). Skill and economic value of the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.*, **126:** 649–667.

Roebber, P. J. & Bosart, L. F. (1996). The complex relationship between forecast skill and forecast value: a real-world analysis. *Weather and Forecasting*, **11:** 544–559.

Stanski, H. R., Wilson, L. J. & Burrows, W. R. (1989). Survey of Common Verification Methods in Meteorology. *WMO World Weather Watch Tech. Report No. 8, WMOI TD No. 358*, 114 pp.

Swets, J. A. (1973). The relative operating characteristic in psychology. *Science*, **182:** 990–1000.

Thompson, J. C. (1952). On the operational deficiencies in categorical weather forecasts. *Bull. Am. Meteorol. Soc.*, **33:** 223–226.

Thompson, J. C. & Brier, G. W. (1955). The economic utility of weather forecasts. *Mon. Wea. Rev.*, **83:** 249–254.

Wilks, D. S. (1995). *Statistical Methods in the Atmospheric Sciences.* International Geophysics Series, Vol. 59, Academic Press, 464 pp.

Wilks, D. S. (1997). Forecast value: prescriptive decision studies. In *Economic Value of Weather and Climate Forecasts*, R. W. Katz and A. H. Murphy, eds., Cambridge University Press, 109–145.

Wilks, D. S. (2000). Diagnostic verification of the Climate Prediction Center long-lead outlooks. *J. Climate*, **13:** 2389–2403.

Winkler, R. L. & Murphy, A. H. (1985). Decision Analysis. In *Probability, Statistics and Decision Making in the Atmospheric Sciences*, A. H. Murphy and R. W. Katz, eds., Westview Press, Boulder, Colorado, 493–524.