

Dependence of the Critical Success Index on sample climate and threshold probability

I. Mason

Regional Office, Bureau of Meteorology, Canberra, Australia

(Manuscript received August 1988; revised November 1988)

The Critical Success Index (CSI) is used as a measure of the accuracy of yes/no forecasts, particularly for rare events or in situations in which the not forecast, not observed frequency is unavailable. It is also used to select threshold probabilities to convert probabilistic forecasts to categorical forecasts. CSI depends on the sample relative frequency of the predictand and on the threshold probability, and this dependence itself varies with the accuracy of the forecasts. Use of CSI as a measure of forecast quality can be misleading unless forecast sets to be compared have the same relative frequency of the event predicted and CSI is calculated at the optimum threshold probability for that level of skill. It is recommended that reported values of CSI be accompanied by values for the sample relative frequency of the predictand and the threshold probability. The indices d' or A_2 from signal detection theory are recommended for evaluation of single sets of yes/no forecasts. Some remarks and suggestions are made on the problem of accounting for the not forecast, not observed occasions.

Introduction

The Critical Success Index was recommended by Donaldson et al. (1975) as a measure of accuracy for yes/no forecasts. It is used to compare sets of forecasts (e.g. McCoy 1986) and also as a criterion for selection of threshold probabilities to convert probabilistic forecasts to categorical forecasts (e.g. Arritt and Frank 1985). Its earliest appearance seems to have been in Gilbert's (1884) discussion of Finley's tornado forecasts, where it was called the ratio of verification. A form of CSI, the Threat Score, was recommended by Palmer and Allen (1949).

If the forecast/observed contingency table is represented by Table 1 then CSI is defined by

$$CSI = d/(b + c + d) \quad \dots 1$$

Other quantities usually presented with CSI are Probability of Detection (POD), False Alarm Ratio (FAR) and bias, defined by

$$POD = d/(b + d) \quad \dots 2$$

$$FAR = c/(c + d) \quad \dots 3$$

$$\begin{aligned} \text{bias} &= (c + d)/(b + d) && \dots 4 \\ \text{so that} &&& \\ CSI &= POD/(1 + FAR \cdot \text{bias}) && \dots 5 \\ \text{or} &&& \\ CSI &= [POD^{-1} + (1 - FAR)^{-1} - 1]^{-1} && \dots 6 \end{aligned}$$

Table 1. General forecast/observed verification table. a,b,c and d represent the number of times each combination of forecast and observation occurred.

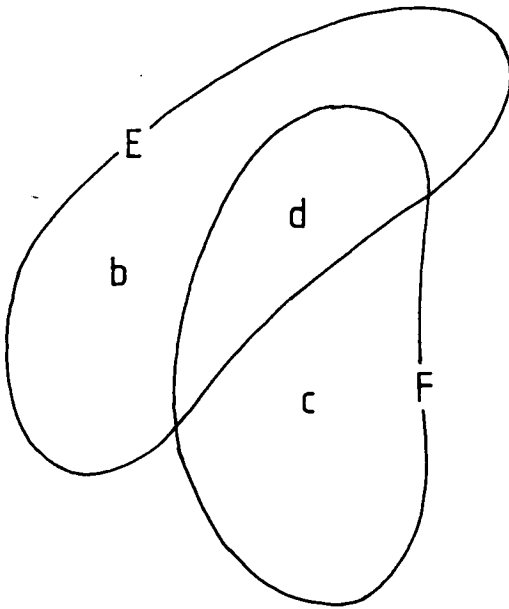
		Observed	
		No	Yes
Forecast	No	a	b
	Yes	c	d

POD and FAR may be regarded as sample estimates of interesting conditional probabilities. POD is the probability of a forecast of occurrence

given that the event occurs, and is related to Murphy and Winkler's (1987) likelihood-base rate factorisation of the joint distribution of forecasts and observations. FAR is the probability that the event does not occur given that it was forecast, and is related to the calibration-refinement factorisation. The quantity referred to as bias is the number of forecasts of occurrences per actual occurrence. CSI is intended to be a summary single-number index of forecast quality.

This score is also applied to areal forecasts of spatially distributed phenomena (e.g. rainfall, severe weather), when the frequencies in Table 1 are replaced by the corresponding areas as in Fig. 1. Weiss (1977) suggested a modification of CSI for this purpose by redefining FAR in terms of forecast densities of severe weather.

Fig. 1 Areas used to calculate CSI (threat score) for forecasts of spatially distributed events. *F* is the area over which the event was forecast to occur, and *E* the area over which it did occur. *b*, *c* and *d* are subsets of these areas.



CSI is particularly appealing when forecasts of very rare events are to be evaluated as it does not require knowledge of the frequency of the not forecast, not observed combination. Non-occurrence of rare events is usually very easy to forecast and is not usually forecast explicitly. If every non-forecast followed by a non-occurrence is counted as a correct forecast, the frequencies in this cell of the contingency table can be several orders of magnitude larger than the others. This loads the sample of forecasts with many trivially easy cases and gives an unrealistic impression of skill. In the case of areal forecasts the not forecast, not observed area is generally indeterminate.

CSI avoids this problem by omitting the not forecast, not observed frequency from consideration altogether. This is unsatisfying because correct forecasting of non-occurrence does sometimes require at least as much skill as correct forecasting of occurrences. The location of a line on a map delineating a forecast area is important for what it excludes almost as much as for what it includes, at least in its immediate vicinity. These considerations tend to make one a little uneasy about measures of forecast quality that do not involve 'a' in Table 1 in some way. Nevertheless, the attraction of a measure of accuracy that does not require the explicit definition of those occasions on which the event was not either forecast or observed has often been overwhelming, and it is likely that this index will continue to be used. This note is about some features of CSI as a measure of accuracy that need to be appreciated if misleading results are to be avoided.

Two specific problems are addressed in this article. One is the strong dependence of CSI on the sample relative frequency of the event being forecast, noted by Gilbert (1884). This may seem surprising, since the sample relative frequency cannot be calculated without the element 'a' of Table 1. It arises operationally through the association of 'a' with 'c', the number of false alarms. Each non-occurrence of the event has a certain probability of being incorrectly forecast as an occurrence. This probability varies from day to day but there are evidently some occasions on which it is high, since false alarms do occur. The number of false alarms actually realised, 'c' in Table 1, clearly depends in part on the number of opportunities for this kind of error presented to the forecaster, which is the total number of non-events, $a + c$.

Another problem with CSI is a dependence on threshold probability, the cut-off probability above which the event is forecast to occur and below which forecast not to occur, or just not forecast. This dependence is well known and is the basis for the use of CSI to select an optimal threshold, i.e. that threshold probability at which CSI is maximised. This is usually done by plotting CSI as a function of threshold probability in a developmental data set. It is shown in this article that this optimising probability depends on the accuracy of the forecasts, so that a threshold selected in a training set of forecasts may not be optimal in a different set of forecasts with possibly different accuracy. Comparisons of values of CSI from sets of forecasts with different accuracies will not be a reliable indication of relative quality if the same threshold has been used in both sets.

The nature of the variation of CSI with sample climate and threshold probability is investigated in more detail in the remainder of this article and some suggestions are made about assessment of

the number of not forecast, not observed occasions that might be used in a valid measure of performance.

A model for the forecasting process

Suppose the forecasts summarised in Table 1 were based on an underlying continuous quantity, for example by forecasting thunderstorms when the Total-Totals Index is over 45. Denote the underlying quantity by X , and the X -criterion for forecasting an event as x^* . X might represent a discriminant function or a forecaster's subjective judgement from which a forecast probability is derived. The value of X does not have to be available explicitly but can be inferred within a monotonic transformation from the forecasts and verifying observations alone.

Let the probability density of X when the event does not occur be $f_0(x)$, and the probability density when the event occurs $f_1(x)$. The elements a, b, c and d in Table 1 can be related to areas under $f_0(x)$ and $f_1(x)$. For example, if N is the total number of forecasts in the sample then d/N is the sample estimate of the probability that the event was both forecast and observed. Using indicator variables F for a forecast of occurrence ($F=1$) or non-occurrence ($F=0$) and E for an observation of occurrence ($E=1$) or non-occurrence ($E=0$),

$$d/N = P\{F=1 \text{ and } E=1\} \quad \dots 7$$

or, factorising as a product of conditional and marginal probabilities,

$$d/N = P\{E=1\} \cdot P\{F=1 | E=1\} \quad \dots 8$$

Since the threshold for a forecast of occurrence is x^* ,

$$P\{F=1 | E=1\} = P\{X > x^* | E=1\} \quad \dots 9$$

or

$$P\{F=1 | E=1\} = \int_{x^*}^{\infty} f_1(x) dx \quad \dots 10$$

The integral on the right-hand side of Eqn 10 is just the area under $f_1(x)$ to the right of x^* . Denoting this area as A_{11} , where the first subscript refers to the forecast and the second to the event, and writing $P\{E=1\}$ as r , the sample estimate of the climatological probability of the event, gives

$$d/N = r \cdot A_{11} \quad \dots 11$$

Similarly,

$$a/N = (1-r) \cdot A_{00} \quad \dots 12$$

$$b/N = r \cdot A_{01} \quad \dots 13$$

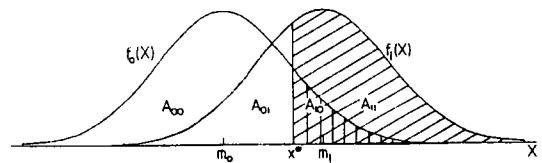
$$c/N = (1-r) \cdot A_{10} \quad \dots 14$$

Figure 2 illustrates these areas.

If the climatological probability is converted to its odds ratio w , so that $w = r/(1-r)$, CSI can be expressed

$$CSI = A_{11}/(1 + A_{10}/w) \quad \dots 15$$

Fig. 2 Probability densities for a generalised forecasting index X . $f_0(x)$ is the density of X prior to non-occurrence of the predictand, and $f_1(x)$ prior to occurrence. A_{10} (vertical hatching) represents the probability of a forecast of occurrence prior to a non-occurrence. A_{11} (diagonal hatching) represents the probability of a forecast of occurrence prior to an occurrence.



The areas A_{11} depend only on the location of the threshold x^* , if f_0 and f_1 are fixed. Equation 15 shows that CSI depends on x^* , through the effect of x^* on A_{11} and A_{10} , and also shows that CSI depends on r .

The threshold on the forecast index x^* is readily converted to a threshold probability p^* using Bayes' formula, which in this case can be expressed

$$p^*/(1 - p^*) = w \cdot f_1(x)/f_0(x) \quad \dots 16$$

As long as the nature and parameters of the distributions of X are fixed, it is reasonable to regard the skill of the forecasting system as constant, since the performance of forecasts based on X is then determined for all possible thresholds and climatic probabilities. The further apart $f_0(x)$ and $f_1(x)$ are the better the accuracy with which the event can be forecast. Any sensible measure of the difference between the distributions $f_0(x)$ and $f_1(x)$ could serve as a measure of skill. One such measure which can be calculated from Table 1 under certain assumptions is described below.

Suppose that $f_0(x)$ is a Gaussian density with mean m_0 and variance s_0^2 and $f_1(x)$ also Gaussian with mean m_1 and variance s_1^2 . The difference between these distributions, which determines the achievable performance of the system for all r and p^* , can be described by the separation of m_0 and m_1 in units of the standard deviation of $f_0(x)$, denoted Dm , and the ratio of the standard deviation of $f_0(x)$ to $f_1(x)$, denoted s .

$$Dm = (m_1 - m_0)/s_0 \quad \dots 17$$

$$s = s_0/s_1 \quad \dots 18$$

When $s_0 = s_1$, $s = 1.0$ and Dm is conventionally denoted d' . d' is calculated as the difference between the standard normal deviates of the conditional hit rate $d/(b + d)$ and the conditional false alarm rate $c/(a + c)$ (Mason 1982). d' is readily converted into the performance measure A_z that has been recommended as a good single-number index of forecasting accuracy (Swets 1988), discussed further below.

Variation of CSI with sample climate

It is possible to calculate A_{11} and A_{10} for given probability densities f_1 and f_0 and any fixed threshold, and hence graph CSI as a function of r using Eqn 15 for various values of Dm and s . Figure 3 shows the variation of CSI with r at a constant threshold probability of 0.1, for values of d' from 0.5 to 2.5 ($s = 1.0$). Figure 4 shows the

Fig. 3 Variation of CSI with climatological probability of event to be forecast. Threshold probability 0.10. a, $d'=0.5$; b, $d'=1.0$; c, $d'=1.5$; d, $d'=2.0$; e, $d'=2.5$.

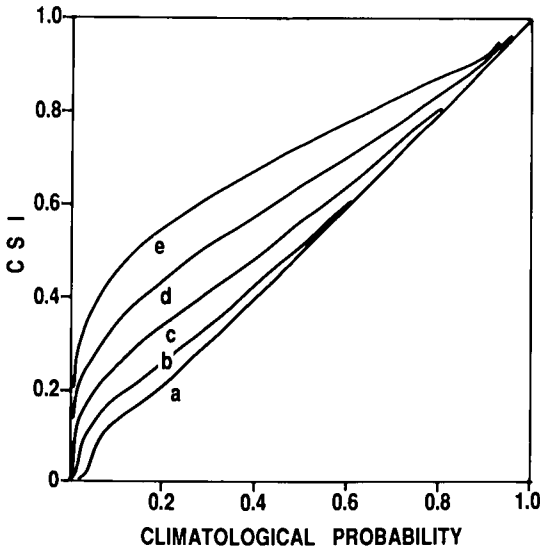
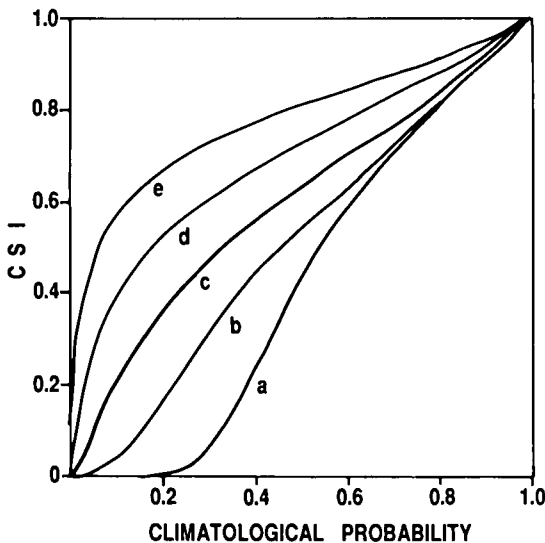


Fig. 4 Variation of CSI with climatological probability of event to be forecast. Threshold probability 0.50. a, $d'=0.5$; b, $d'=1.0$; c, $d'=1.5$; d, $d'=2.0$; e, $d'=2.5$.



variation of CSI with r when the threshold probability is 0.5, otherwise for the same parameters as Fig. 3. The salient point is that CSI can vary through its whole range as a result only of changes in the sample relative frequency of the predictand. This variation is particularly strong when r is small, often the case when CSI is used.

When forecast sets are compared using CSI it is clearly essential that they have the same value of r . If values of CSI are quoted the value of r should also be specified.

Variation of CSI with threshold probability

Using the same assumptions as in the previous section, Figs 5 and 6 show the variation of CSI with threshold probability for the same range of values of d' . Figure 5 has r equal to 0.05 and Fig. 6 has r equal to 0.20, typical of the values encountered in practice.

The locus of optimal CSI is indicated by the broken line. The relation between the optimal value of CSI, CSI^* , and the corresponding threshold probability p^* can be calculated using the method detailed by Mason (1979) and is

$$p^* = CSI^*/(1 + CSI^*) \quad \dots 19$$

The main point of Figs 5 and 6 is that the maximum value of CSI occurs at different threshold probabilities for different levels of skill (d'). Hence a threshold probability that has been

Fig. 5 Variation of CSI with threshold probability. Climatological probability of event 0.05. Curves labelled with d' . Broken line indicates optimal CSI.

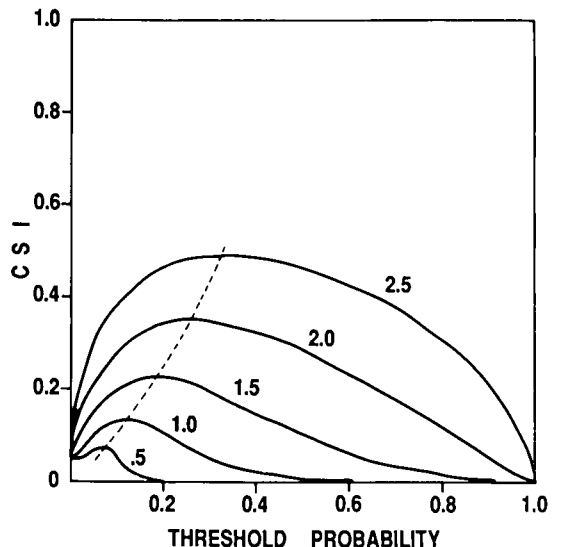
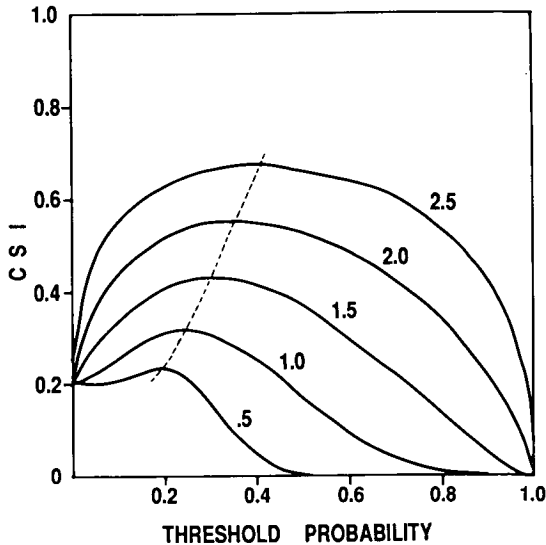


Fig. 6 Variation of CSI with threshold probability. Climatological probability of event 0.20. Curves labelled with d' . Broken line indicates optimal CSI.



found empirically in a developmental set of forecasts may not be optimal in an independent set. CSI will change in the appropriate direction as d' changes, and ranks on CSI will be the same as ranks on d' only if the threshold is optimal and sample climate is the same in both forecast sets. If these conditions are not met, or it is not known whether they are met, then CSI is not a reliable indicator of forecast quality.

Figures 5 and 6 also show that a CSI as high as r can be obtained with no skill at all simply by using a threshold of zero, i.e. by forecasting occurrence all the time.

Values of CSI should always be accompanied by the threshold probability used to generate the forecasts. If this is not available then Figs 5 and 6 illustrate the fact that it is difficult to give much meaning to CSI.

Discussion

CSI is attractive as a summary measure of forecast quality because it can be calculated without using the frequency of the not forecast, not observed category. When the event being forecast is rare this category can be several orders of magnitude larger than the other categories. Since many of these correct forecasts of non-occurrence are trivially easy, measures of skill that take them into account are unrealistically inflated. However, CSI is severely compromised by a dependence on sample climate as shown in Figs 3 and 4 and a dependence on threshold probability as shown in Figs 5 and 6. Use of CSI to compare forecast sets that do not have the same

values for these quantities is likely to be misleading.

The other common use for CSI, to select optimal threshold probabilities to convert probabilistic to yes/no forecasts, is also problematic. The optimal threshold depends on both sample climate and skill. Values of CSI calculated from sets of yes/no forecasts generated at the same threshold but in forecast sets with different r and different d' do not reliably indicate a difference in skill.

Although the detail of the curves in Figs 3 to 6 is determined by the assumption that f_0 and f_1 are specifically Gaussian densities, the curves are similar for practically any more or less bell-shaped distributions. Any forecasting system whose output is stationary in a time-series sense implies fixed distributions of some kind and the areas A_{11} and A_{10} will vary together so as to produce curves with similar features to those shown. CSI will have an optimal value related to p^* by Eqn 19, hence dependent on the value of CSI, and will increase monotonically with r .

All the scores commonly used in meteorology have similar problems, showing a more or less marked dependence on p^* and usually on r as well. The only well-known score that does not depend on r is Hansen and Kuipers (1965; originally due to Pierce 1884). This score does have a strong dependence on p^* . It is optimised at a threshold probability equal to r (Mason 1979). Murphy (1986) recently suggested using a form of the Brier probability score to evaluate categorical forecasts. The dependence of the Brier score on sample climate has been known for some time (Glahn and Jorgensen 1970) and the two-state form also has this property, although its dependence on p^* is relatively slight.

The least unsatisfactory measure of forecasting performance for a single set of yes/no forecasts appears at this stage to be either d' or the related index A_2 from signal detection theory. Detailed discussion of the basis for the choice of A_2 is beyond the scope of this note, and may be found in papers by Swets (1986a,b; 1988). It is equal to the area under the relative operating characteristic (ROC) fitted to a Gaussian probability model, and is calculated for yes/no forecasts as the area under a normal curve to the left of $d'/2^{1/2}$ taken as a standard normal deviate. A_2 has no variation with r and when $s = 1.0$ has no variation with p^* . It is not possible to calculate s from a single set of yes/no forecasts, so $s = 1.0$ has to be assumed.

Underlying the characteristics of CSI described above is the fact that a single set of yes/no forecasts is not sufficient to completely describe the performance of a forecasting system, because it can only show the performance of the system at a single threshold. Use of methods based on signal detection theory shows that at least two par-

ameters are required to describe the discrimination capacity of any diagnostic system (Swets 1986a). These parameters are the slope and intercept of the ROC on bi-normal axes. A single set of yes/no forecasts can only provide one of these, for example the intercept, if the slope is assumed. The ROC is the only available means of assessing accuracy not troubled by variations in sample climate or differing thresholds. Construction of the ROC does require knowledge of the not forecast, not observed frequency, but it is difficult to see how any valid measure of forecast quality can be obtained without it. Correct forecasts of non-occurrence are sometimes neither trivially easy nor unimportant. The problem is to distinguish these significant 'true negatives' from those which are so easy that no skill is involved.

Murphy and Ehrendorfer (1987), examining the relationship between forecast accuracy as indicated by the Brier score and value in the cost-loss ratio situation, have recently pointed out that no single-number index can adequately summarise all aspects of forecast quality.

Estimating the no-no frequency

The following remarks are a tentative approach to the problem of estimating a realistic value for the number of not forecast, not observed occasions.

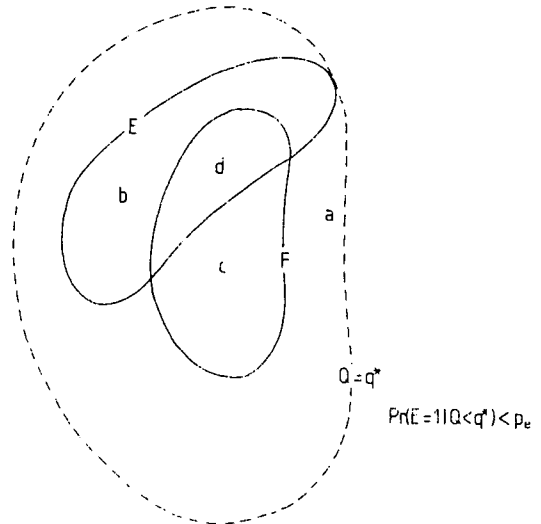
A start might be made by defining the set of cases over which forecast quality is to be evaluated. An objective selection rule is required which eliminates from the sample as many as possible of the very simple no-no forecast/event pairs, while keeping as many as possible of the events and preferably all of the erroneous forecast/event pairs, which almost by definition were not easy. Given some predictor Q a threshold q^* is sought such that

$$P\{E=1 \text{ or } F=1\} < p_0 \quad \dots 20$$

where p_0 is a sufficiently small probability. q^* is chosen so that p_0 is as large as possible but no forecasts of occurrence or actual occurrences have $Q < q^*$. The not forecast, not observed frequency is then the number of occasions on which $Q > q^*$ and the event was neither forecast nor observed. For example, it might be observed that there has never been a thunderstorm when some index Z is less than say 30, although there have been some for $Z=31$. The set of forecasts to be evaluated would then contain only those days on which $Z \geq 30$, unless some forecasts of occurrence had been issued when Z had a lower value. This lower value would then be the threshold z^* .

In the two-dimensional case construction of q^* might proceed, given isopleths of Q , by starting with a remote value and shrinking the region enclosed by the isopleth until it contacts either the forecast or the observed area. Figure 7 illustrates this suggestion.

Fig. 7 A suggestion for the area to be credited as a correct forecast of non-occurrence. Q is some meteorological quantity related to the predictand and q^* is an isopleth of Q such that $Q < q^*$ is the smallest area that contains all of both the forecast area F and the observed area E . The unconditional probability of an occurrence outside q^* is less than some small value p_0 .



Some such effort to define an evaluation set seems to be necessary. Restricting evaluation to only those occasions on which the predictand was either forecast or observed eliminates many skilful forecasts of non-occurrence for which the system should be given credit, and introduces an uncontrolled source of variation into CSI.

Conclusion

CSI has a strong dependence on sample climate and also on the threshold probability for forecasting an occurrence. The threshold probability at which CSI is optimised depends on the accuracy of the forecasts. A value of CSI has little meaning unless these quantities are known.

All scores for yes/no forecasts have similar problems. The indices d' or A_z from signal detection theory are least affected by sample climate or threshold probability and may be worth investigation by meteorologists concerned with evaluation of categorical forecasts.

Satisfactory assessment of the quality of yes/no forecasts requires knowledge of the frequency of the not forecast, not observed category. This requires some restriction of the set of occasions to be evaluated so as to eliminate trivially easy forecasts of non-occurrence. This note has suggested a general approach to this problem but it clearly needs a good deal more attention.

Acknowledgments

This work was supported by the Australian Bureau of Meteorology, the Center for International Resource and Environmental Studies at the University of Colorado, and the US NOAA Environmental Research Laboratories, Boulder, Colorado. Discussions with Duane Haugen, Herb Winston, Mary McCoy, Denice Walker and John Flueck, all at NOAA Environmental Research Laboratories, Boulder, were valuable.

References

- Arritt, R.W. and Frank, W.M. 1985. Experiments in probability of precipitation amount forecasting using model output statistics. *Mon. Weath. Rev.*, 113, 1837-51.
- Donaldson, R.J., Dyer, R.M. and Kraus, R.M. 1975. An objective evaluator of techniques for predicting severe weather events. Preprint volume, *Ninth AMS Conference on Severe Local Storms*. Norman, Okla. Published by American Meteorological Society, Boston, Mass., 321-6.
- Gilbert, G.K. 1884. Finley's tornado predictions. *American Meteorological Journal*, 8, 166-72.
- Glahn, H.R. and Jorgensen, D.L. 1970. Climatological aspects of the Brier P-score. *Mon. Weath. Rev.*, 98, 136-41.
- Hansen, A.W. and Kuipers, W.J.A. 1965. On the relationship between the frequency of rain and various meteorological parameters. Koninklijk Nederlands Meteorologisch Instituut, *Meded. Verhand.*, 81, 2-15.
- McCoy, M.C. 1986. Evaluation of PROFS 1985 convective weather forecasts. Preprint volume, *Eleventh Conference on Weather Forecasting and Analysis*, June 17-20 1986, Kansas City, Mo. Published by American Meteorological Society, Boston, Mass., 340-4.
- Mason, I.B. 1979. On reducing probability forecasts to yes/no forecasts. *Mon. Weath. Rev.*, 107, 207-11.
- Mason, I.B. 1982. A model for assessment for weather forecasts. *Aust. Met. Mag.*, 30, 291-303.
- Murphy, A.H. 1986. Comparative evaluation of categorical and probabilistic forecasts: two alternatives to the traditional approach. *Mon. Weath. Rev.*, 114, 245-9.
- 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, 1130-8.
- Palmer, W.C. and Allen, R.A. 1949. Note on accuracy of forecasts concerning the rain problem. Unpublished manuscript, U.S. Weather Bureau, Washington DC, 4pp.
- Pierce, C.S. 1884. The numerical measure of the success of predictions. *Science*, 4, 453-4.
- Swets, J.A. 1986a. Indices of discrimination or diagnostic accuracy: their ROCs and implied models. *Psychological Bulletin*, 99, 100-27.
- Swets, J.A. 1986b. Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. *Psychological Bulletin*, 99, 181-98.
- Swets, J.A. 1988. Measuring the accuracy of diagnostic systems. *Science*, 340, 1285-93.
- Weiss, S.J. 1977. Objective verification of the severe weather outlook at the National Severe Storms Forecast Center. Preprints, *Tenth Conference on Severe Local Storms*, Omaha. Published by American Meteorological Society, Boston, Mass., 395-402.

