9.12 THE VALUE IN PROBABILITIES FOR AERODROME FORECASTS

Ross Keith
Bureau of Meteorology, Townsville, Australia and James Cook University, Townsville, Australia

and

Ian Mason
Bureau of Meteorology, Canberra, Australia (retired)

1. INTRODUCTION

The value of a weather forecast to a user is about much more than accuracy. The most important aspect of the design of a forecast delivery system is to optimise the flow of the opinion of the forecaster directly to the user. This paper demonstrates that considerable potential economic value is lost by using the traditional method of providing weather information in terminal aerodrome forecasts (TAFs) in categorical form, i.e. as a binary, Yes/No product

2. EXPERIMENT DESIGN

One of the purposes of the experiment was to validate use of a signal detection model for forecasters' decisions when formulating TAFs. Other studies, most notably Mason (1982), have shown that probabilistic forecasts of elements like rain, storms and temperature closely fit the signal detection model. Forecasters were asked to nominate their confidence, to the nearest 10%, that the weather at five different lead times would be below the Special Lowest Alternate Minimum (SLAM) for the aerodrome. The SLAM is that level of visibility and cloud base used to determine fuel carriage. The lead times are 1, 3, 6, 12 and 18 hours. Forecasters input these percentages at the same time that they formulated the four routine TAFs each day. This was done so that the lead time-skill relationship was not skewed. Non-routine amendments are usually issued to amend the TAF at short lead times, and omitting these removes any possible bias. The trial data shown here is from Townsville on the tropical east coast of Australia, from Melbourne on the east of the south coast, and Sydney on the subtropical east coast. The meteorological causes of below minimum weather at the three locations are quite different.

3. THEORY

3.1 *Signal Detection Theory*

Signal Detection Theory (SDT) assumes that, prior to a decision, there are two overlapping normal probability distributions, one for the weight of evidence that the event will occur and another that the event will not occur. This is illustrated in fig. 1. A more complete treatment of SDT as applied to forecast verification is given in Mason (1982). The separation of the means, d', can be used as a measure of forecast skill providing the ratio of the standard deviations of the two distributions is close to one.

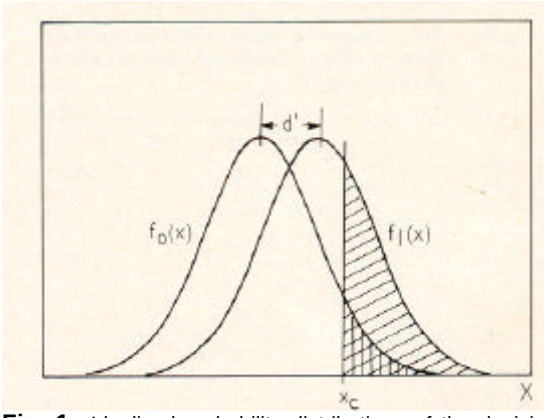

Fig. 1 Idealised probability distributions of the decision variable ÷, f_0 preceding non-occurrence of an event and f_1 preceding occurrence. The area marked by vertical hatching indicates the probability of a false alarm, and the diagonal hatching represents the probability of a hit.

The y axis is weight of evidence, and the x axis is values of ÷, the decision threshold. $f_1(÷)$ represents the evidence *for* the event, and $f_0(÷)$ the evidence *against* the event. $×_c$ is the critical decision threshold, above which the forecast is yes, below which it is no. Note that the formal definition of hit rate, *h*, is Pr(Forecast=Yes|Event =Yes) and of false alarm rate, *f*, is Pr(Forecast=Yes|Event=No).

For the experiment, the hit rate, *h*, and false alarm rate, *f*, were calculated for each of the 11 thresholds from 0% to 100%. For each threshold *h* and *f* are plotted against one another, the resulting graph is called a Relative Operating Characteristic and has a parabolic, concave shape with points at (0,0) and (1,1).

3.2 *Forecast Value*

---

* *Corresponding author address:* Ross Keith, Bureau of Meteorology, RAAF Base, Townsville, 4810, Australia. E-mail: r.keith@bom.gov.au

In any forced choice, binary outcome (Yes/No) forecast situation, the distribution of the outcomes can be summarised by four values: true positives (hits), true negatives (correct rejections), false negatives (misses) and false positives (false alarms). The expected value (EV) of a forecast can be calculated as the sum of the expected value (cost) of each of these four outcomes.

$$EV = h.p_C.V_{TP} + (1-p_C).f.V_{FP} + p_C.(1-h).V_{FN} + (1-p_C).(1-f).V_{TN}, \quad (1)$$

where p_C is the climatological rate of occurrence of the event, or Pr(Event = Yes), the Bayesian prior probability. V_{TP} is the value of a true positive, and similarly for V_{FN}, V_{FP}, V_{TN}.

For a perfect forecast $h = 1$ $f = 0$, the expected cost (EC) with respect to a perfect forecast is:

$$EC = (1-p_C).f.(\text{False Alarm Cost}) + p_C.(1-h).(\text{Miss Cost}), \quad (2)$$

where False Alarm Cost = $V_{TN} - V_{FP}$, the cost of incorrectly forecasting an event, and Miss Cost = $V_{TP} - V_{FN}$, the cost of not forecasting an event. Note that, by definition, Miss Cost *does not* include any costs which are already incurred by a hit, i.e. a correct forecast of an event.

Because the form of the normal distribution of the signal detection model is assumed, and *h* and *f* can be measured, d' and thus $\div_C$ can be deduced. From the results of the experiment, graphs of EC vs decision threshold were constructed for particular flights. The False Alarm Cost and Miss Cost were provided by QANTAS for these flights. It soon became apparent that there existed a value of the decision threshold which minimises EC. It can be shown that for $d(EC)/d\div = 0$, the optimal value of the decision threshold is p(opt), where

$$p(opt) = CR / (1 + CR) \quad (3)$$

CR is the Cost Ratio and equals False Alarm Cost divided by Miss Cost. Each flight is specified by a CR value. p(opt) is actually the same as the cost-loss ratio from economic utility analysis, derived here in a signal detection framework.

4. RESULTS AND DISCUSSION

Fig. 2 shows plots of *h* vs *f* for individual forecasters at Townsville, Melbourne (Vic RFC) and Sydney (SAMU). There are significant differences in attitude to risk between A and B, C and D, and E and F. A, C and E exhibit a less cautious (higher decision threshold) than B, D and F respectively. So in each group there exists a significant range of attitude to risk i.e. $\div_C$ varies among individuals.

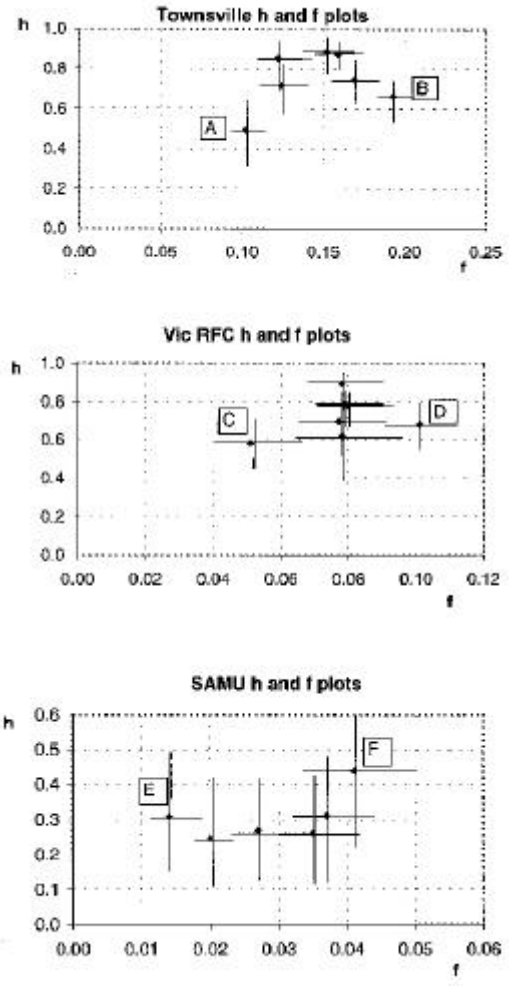

Fig. 2. h,f plots for individual forecasters at Townsville, Vic RFC (Melbourne) and SAMU (Sydney) with 95% confidence limits.

Fig. 3 shows plots, for each lead time at Townsville, of the normal deviates of *h* and *f*. The linear relationship between them is proof of the initial assumption that they are generated by normal probability distributions. This has been shown to be the case for other weather elements. Forecasts for the other two locations exhibit similar behaviour.

Fig. 4 shows plots, for Townsville and Melbourne, of *h* vs lead time for the forecasts and for persistence. The forecast *h* is calculated from a maximum likelihood ROC, and is at the same value of *f* as scored by persistence at each lead time. As can be seen the forecasts at Townsville fail to match persistence out to about 4 hours, and at Melbourne out to about 2 hours. At these short lead times, airlines would be better off using present weather for flight planning.

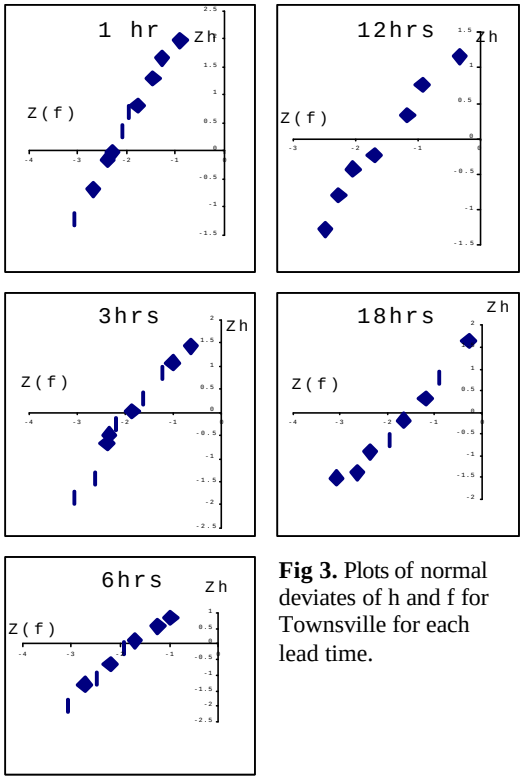

Fig 3. Plots of normal deviates of h and f for Townsville for each lead time.

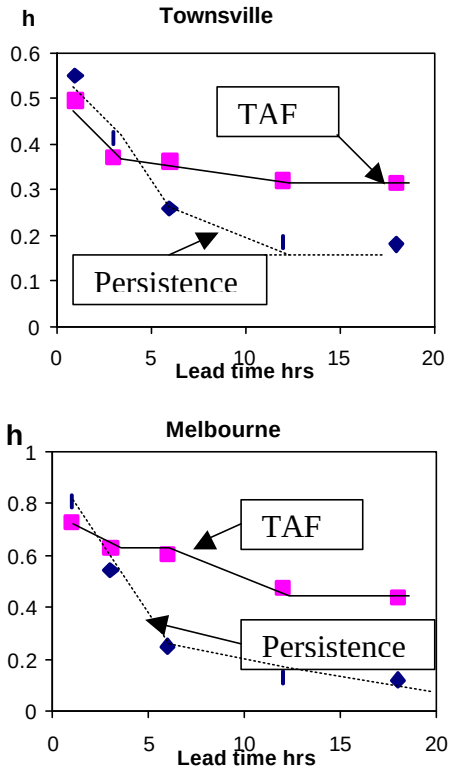

Fig.4. h vs lead time for Townsville and Melbourne, for the TAF and persistence.

Fig. 5 shows reliability diagrams for Melbourne. A degree of over-forecasting is apparent, though there is obviously significant skill shown in forecasting the probabilities. Similar results occurred for Townsville, but with greater overforecasting.

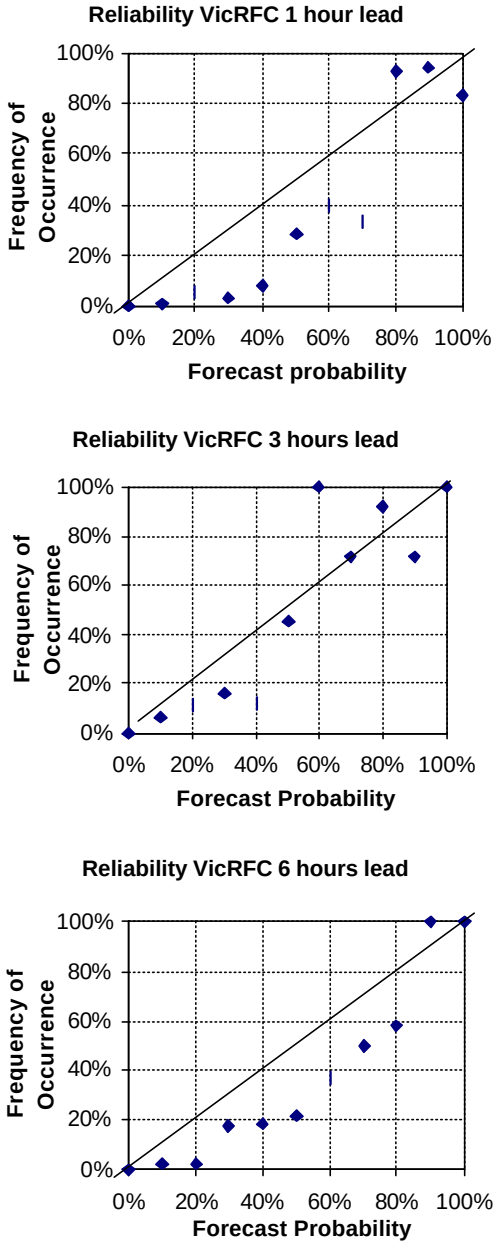

Fig. 5. Reliability diagrams for Melbourne for 1, 3 and 6 hour lead times. The diagonal is perfect calibration.

5. COST ANALYSIS.

From costs supplied by QANTAS for a flight from Singapore to Melbourne, the False Alarm Cost (FAC) was $1,390 and the Miss Cost (MC) was $10,535. This produces a CR value of 0.132, and so p(opt) = 0.117. From *h* and *f* values measured in the trial, the average critical decision threshold for the Melbourne forecasters is about 0.02, or in other words they forecast Yes once they think below minimum weather is >= 2% likely. This extreme degree of conservatism is caused by forecasters' perception of the consequences of a missed event. Fig. 6 is a graph of EC vs $\div_C$ for that flight, using d' measured in the trial and the climatological rate of below SLAM weather of 0.02.

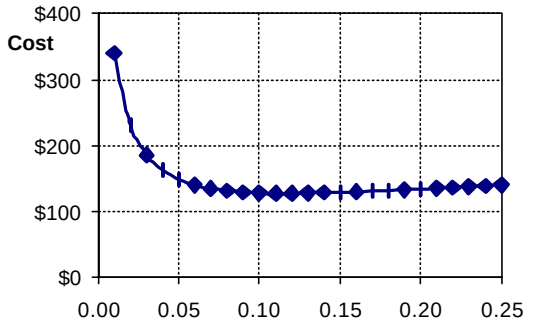

Fig. 6 Cost vs Decision Threshold
FAC $1,390 MC $10,535 d'=2.1, Pc=0.02

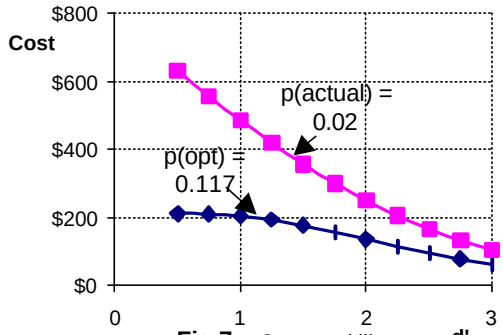

Fig 7. Cost vs skill
FAC = $1,390 MC=$10,535 Pc=0.02

At a decision threshold of 0.02, the cost of the errors in the forecast is on average $231 per flight. If the forecast was reliably made at the optimum decision threshold of 0.117, the cost would be $128. So a perfectly reliable forecast, made at the appropriate decision threshold, would save about 45% of the total cost of the errors. Using the reliability diagrams in Fig.5, if a forecaster was asked to use 0.117 as his or her decision threshold, the effective decision threshold would be about 0.07. If this is used, the cost is $135. So, for this flight, even the moderately reliable forecasts as currently produced would provide most of the savings (41%) gained by the perfectly reliable forecasts (45%).

Fig. 7 shows, again for the Singapore to Melbourne flight, how EC varies with d', the index of skill. The two plots are for the optimum decision threshold of 0.117 and for the measured decision threshold of 0.02. The significant difference in EC for the two different decision probabilities is obvious, especially at low skill. Note also that if the decision threshold is optimal, a decrease in skill does not matter all that much. So one could suggest that the issue of risk management is at least as important as forecaster skill.

Of course the value of p(opt) is different for each flight. For example another flight into Melbourne, a short flight originating in Sydney, has a p(opt) of 0.008. Remembering that the flight from Singapore into Melbourne has p(opt) of 0.117, if the two flights arrived in Melbourne at the same time, at least one would have planned on a TAF formulated with a highly sub-optimal decision threshold.

6. CONCLUSION

Considerable economic benefit is potentially available to airlines if TAFs were expressed as estimated probability of below minimum weather.

Such a system would unlock the value of forecasters' ability to provide reasonably reliable estimates of the probability of occurrence of these events. The amount of benefit would depend on three factors:

1. The ability of airlines to specify False Alarm Costs and Miss Costs,
2. The degree to which regulators would allow airlines to incorporate this approach into flight planning,
3. The ability of forecasters to provide reliable estimates of the probability of events.

7. REFERENCE

Mason, I. B. 1982. A model for assessment of weather forecasts. *Aust. Met. Mag.* **30.** 291-303.