How to judge the quality and value of weather forecast products

John E Thornes, *School of Geography and Environmental Sciences, University of Birmingham, Birmingham, UK*
David B Stephenson, *Department of Meteorology, University of Reading, Reading, UK*

In order to decide whether or not a weather service supplier is giving good value for money we need to monitor the quality of the forecasts and the use that is made of the forecasts to estimate their value. A number of verification statistics are examined to measure the quality of forecasts – including Miss Rate, False Alarm Rate, the Peirce Skill Score and the Odds Ratio Skill Score – and a means of testing the significance of these values is presented. In order to assess the economic value of the forecasts a value index is suggested that takes into account the cost-loss ratio and forecast errors. It is suggested that a combination of these quality and value statistics could be used by weather forecast customers to choose the best forecast provider and to set limits for performance related contracts.

1. Introduction

Users of weather forecasts, particularly paying customers, are operating within an increasingly commercial environment and have to attempt to prove that they are getting value for money from all of their expenditure. The introduction of compulsory competitive tendering (CCT) may have reduced costs but it does not guarantee forecast quality and can lead to new players entering the market with no track record. One method to ensure customer satisfaction is to use performance related contracts. For example, a customer may set a target for the percentage of correct forecasts over the period of a contract: if the forecasts are better than the target the forecast provider receives a bonus and if they are worse than the target the forecast provider is paid less money, according to an agreed scale. If the target was 86% accuracy and the value of the contract was £20,000, for example, it could be agreed that for every percentage above 86% the forecast provider receives an extra £1,000 and for every percentage point below 86% they receive £2,000 less. These figures are arbitrary however. What is the true value of better or worse forecasts? This paper presents a more sophisticated set of potential verification targets which can be used to judge the quality of forecasts and chosen to suit the customer's commercial interests.

The judgement of weather forecast quality and/or value has received considerable attention in the literature in recent years, for example Mylne (1999), Stanski *et al.* (1989), Thornes (1995, 1996), Thornes & Proctor (1999), Richardson (2000), Stephenson (2000) and Wilks (1995). This paper demonstrates some of this knowledge in order to verify road weather forecasts and shows how that information can be used by highway engineers and other users, to keep a 'sharp eye' on their weather forecast suppliers.

2. Case study: road weather forecasts

More than £2 million pounds per winter is being spent on road weather forecasts in the United Kingdom out of a total budget of approximately £140 million for winter road maintenance. It is difficult to assess independently the quality and value of these road weather forecasts and most highway authorities rely on a simple set of statistics provided by the weather service providers. The current guidance specification for road weather forecasts issued by the Highways Agency, only calls for a Percent Correct of 86% for frost forecasts on nights when the minimum road surface temperature is 5° C or below. In this case, for simplicity, a frost is defined as when the road surface temperature falls to zero or below irrespective of surface moisture. Weather forecast providers are often required to produce a 2 × 2 contingency table at the end of the winter for each forecast site. The minimum road surface temperature for each night is noted at each forecast site and compared with the forecast minimum road surface temperature. The results are entered into the contingency table just for the nights when the actual road surface temperature fell to 5° C or less.

For example, during the winter of 1995/96 there were 77 such nights at a road weather site located at High Eggborough on the M62 motorway between Leeds and Hull. Both the Met. Office and Oceanroutes were providing forecasts for that site for different customers. The results for the Met. Office are given as an example in Table 1.

Percent Correct (*PC*) is simply the percentage of correct forecasts:

$$PC = \frac{a+d}{n} \times 100$$

For this case *PC* is 87%, which just exceeds the target of 86%.

There are two types of error in the forecast.

- A Type 1 error is defined as those nights when the road surface temperature was forecast to stay above zero when in fact it fell to zero or below. This is potentially dangerous for the road user as the maintenance engineer may decide not to salt the roads and if the road is wet, ice may form on the road surface and accidents take place. The number of nights with a Type 1 error are given in the contingency table as *c*.
- Type 2 errors occur when the road surface temperatures are forecast to go to zero or below but in reality they do not. The maintenance engineer may then salt the roads unnecessarily. This does not effect road safety but is a waste of salt and money. The number of nights with a Type 2 error are given in the contingency table as *b*.

The use of Percent Correct is an over simplistic check on forecast quality that does not take into account the proportion of Type 1 and Type 2 errors. Also a forecast accuracy of greater than 86% may not be of greater value than the loss suffered if the forecast accuracy is less than 86%. The costs and losses associated with Type 1 and Type 2 errors are discussed below.

The highway engineer is concerned about more than just road surface temperature. Forecasts of road wetness and snow are also of considerable importance. It should be possible to also monitor the quality of such forecasts in any verification scheme and the assessment of snow forecasts is discussed below.

3. What makes a good weather forecast? Quality and/or value?

There needs to be a clear link between quality and value, especially when considering road weather forecasts. It has been traditionally accepted in the industry that a slight bias (explained below) in the forecast of road surface temperature should be present. This is due to consequences of the Type 1 and Type 2 errors discussed above. A Type 1 error in the forecast which leads to the roads not being salted could lead to the local authority being sued if a motorist skids on an icy road. This could cost the local authority millions of pounds in compensation if the driver is badly injured and the local authorities insurance 'excess' is high (Mead, 1998). A Type 2 error will only cost the local authority tens of thousands of pounds if the roads are salted unnecessarily. Hence there is a tendency to 'err on the side of caution' and 'over forecast' the number of frosts or the occurrence of snow. It is only a matter of time before the weather forecast provider is also sued as a result of an incorrect forecast (Millington, 1987).

The definitions of quality and value are discussed below with examples.

4. Quality of a forecast

Stanski *et al.* (1989) review six attributes of a weather forecast that make up the total quality: reliability, accuracy, skill, resolution, sharpness and uncertainty. They also make the important point that no single verification measure provides complete information about the quality of a product.

A number of measures of forecast quality are therefore required, but in order to avoid confusion their use must be obvious, they must be easy to calculate and their statistical significance should be testable. Of the six attributes mentioned above, the first three – reliability, accuracy and skill – are the easiest to measure and will be considered here. Resolution is important in the forecasting of precipitation – being able to distinguish between, for example, snow, sleet, freezing rain, hail, drizzle and rain. Sharpness is a measure of the spread of the forecasts away from climatology, e.g. a forecast method that can predict frosts in summer as well as winter shows high sharpness whereas a forecast method that can only predict frosts in winter has low sharpness. Uncertainty relates to the climate, for instance some areas of the United Kingdom have comparatively few road frosts (e.g. Cornwall) in comparison to others (e.g. Highlands of Scotland). This may effect the achievability of performance targets (Halsey, 1995) and if frost or snow are rare events then the 'base rate' effect (Mathews, 1997) comes into play as discussed below. Another important attribute of the forecast is the 'precision' with which the forecast can hit the right side of a threshold (e.g. 0° C). There are many other thresholds that are important to forecast users and sometimes the customer only needs to know whether or not a threshold will be crossed, e.g. high winds affecting overhead cables for railways (Thornes, 1997).

4.1. Reliability

The reliability of a forecast can be measured by calculating the bias. This will show if the forecasters are consistently over-forecasting the number of frosts or snow. The bias tells us whether or not more forecasts of frost are being issued than frosts are observed. It is normal to find a positive bias in frost forecasts in order to hedge the chance of a Type 1 error. The target limits to bias will be discussed later when it is related to value.

Table 1. *Contingency table for the analysis of Met. Office road surface temperature forecasts at High Eggborough for the winter of 1995/96*

Forecast	Observed		
	Frost	No frost	Total
Frost	$a = 29$	$b = 6$	$a + b = 35$
No frost	$c = 4$	$d = 38$	$c + d = 42$
Total	$a + c = 33$	$b + d = 44$	$n = a + b + c + d = 77$

a: Frost forecast and frost observed (29 nights): Forecast Correct
c: No frost forecast but frost observed (4 nights): Type 1 Error
b: Frost forecast but no frost observed (6 nights): Type 2 Error
d: No frost forecast and no frost observed (38 nights): Forecast Correct
Type 1 Error: Possibility of severe road accidents as roads may not be salted
Type 2 Error: Possibility of wasted salt as roads may be salted unnecessarily
W = a + c Total number of frosts i.e. a measure of winter severity

The bias (B) is calculated as follows using the notation of Table 1.

$$B = \frac{a+b}{a+c}$$

When $B = 1$ then the forecasts are said to be perfectly reliable. When the $B > 1$ then this indicates over-forecasting and when the $B < 1$ this indicates under-forecasting. The bias of the Met. Office forecast given in Table 1 is 1.06 (i.e. slight over-forecasting). A bias of 1 does not necessarily mean that the forecasts are accurate, however.

4.2. Accuracy

Percent Correct has already been discussed above and is a measure of forecast accuracy. It relates to the terms *a* and *d* in the contingency table. There are several other measures of accuracy that attempt to look at the incorrect forecasts *b* and *c* and the two independent measures that are recommended here are Miss Rate and False Alarm Rate, which are both calculated from the actual number of 'frosts' and 'no frosts' observed (i.e. the columns of the contingency table).

(a) Miss Rate

The Miss Rate (M) is an important statistic as it indicates how many of the observed frosts were not forecast (i.e. it relates directly to the number of Type 1 errors). We want this number to be as close to zero as possible. If *c* is zero (i.e. no Type 1 errors) then M will be zero.

$$M = \frac{c}{a+c}$$

The Miss Rate of the Met. Office forecast given in Table 1 is 0.12.

The Hit Rate (H) given by:

$$H = (1 - M)\frac{a}{a+c}$$

can be derived from the Miss Rate and both statistics are not therefore needed. The Hit Rate and the Miss Rate by themselves can be misleading; for instance if a frost was forecast every night then the H would be 1 and the M would be zero even though the forecast was of very poor quality.

(b) False Alarm Rate

The False Alarm Rate (F) is also an important statistic as it considers the number of Type 2 errors, i.e. the number of nights that a frost was forecast but did not occur. These nights are when roads may be salted unnecessarily. If *b* is zero then F is zero. The smaller the value of F the better. There is some confusion in the literature over the definition of the False Alarm Rate but for our purposes it is defined as:

$$F = \frac{b}{b+d}$$

The False Alarm Rate for the Met. Office data given in Table 1 is 0.14. The Miss Rate and False Alarm Rate correspond to the two columns of data in the contingency table. It is better to examine the columns of the contingency table rather than the rows because it is the observations of frost or no frost that determine the quality of the forecasts.

4.3. Skill

There are many different skill scores that attempt to assess how much better the forecasts are than those which could be generated by climatology, persistence or chance. A forecast based on climatology would take, for example, the likelihood of frost based on the minimum road surface temperatures that have been observed on that day over the last 30 years. For road weather forecasts it is very unlikely that climatology will be of any use as the mean minimum road surface temperature in winter is above zero across most of the UK (Thornes, 1995). Hence climatology would never predict a frost in most parts of the UK. Climatology can tell us that on average we can expect so many frosts in a winter but cannot tell us when. Persistence is a very simple forecast method; for example, persistence would tell us that if there was a frost last night then there will be a frost tonight. This would score quite well for long periods of frost but would always be incorrect when the weather changes from frosty to non-frosty nights and vice versa. Chance can be used to see if the distribution of scores in the contingency table is as expected from the frequencies of forecast and observed frosts (e.g. the Heidke Skill Score discussed in Stanski *et al.*, 1989) but the resulting statistic is difficult to interpret. It is proposed therefore to use two other measures of skill that are easy to calculate, are testable for significance and are appropriate for use in performance-related contracts.

(a) Peirce Skill Score

The Peirce Skill Score (*PSS*) was first published in 1884 and has since been rediscovered as Kuipers Performance Index and the True Skill Statistic (*TSS*) as discussed in Stephenson (2000). It is simply calculated from the Miss Rate (*M*) and the False Alarm Rate (*F*) as follows:

$$PSS = 1 - M - F$$

The closer the value of *PSS* to 1 the better. For the Met. Office data given in Table 1, *PSS* = 0.74. The significance of the Peirce Skill Score is discussed below in the Appendix. A weakness of *PSS* is that it treats *M* and *F* equally, irrespective of their likely differing consequences.

(b) Odds Ratio Skill Score

The 'odds' or 'risk' of an event happening is the ratio of the probability that the event will happen to the probability of it not happening. In other words, the odds of an event that has a probability p of occurring is given by $p/(1–p)$ and ranges from zero to infinity. For example, if an event has a probability of 0.8 it has odds of 0.8/(1 – 0.8) = 4 (which equals '4 to 1 on' in bookmaker's jargon). Forecast skill can be assessed by comparing the odds of making a good forecast (a hit) with those of making a bad forecast (a false alarm). The Odds Ratio Skill Score has not been used for road weather forecasts verification before and therefore a detailed discussion of its significance is given in the Appendix and its derivation is discussed in Stephenson (2000). The Odds Ratio (*OR*) is defined as the ratio of the multiple of correct forecasts (i.e. $a \times d$) to the multiple of the incorrect forecasts (i.e. $b \times c$).

$$OR = \frac{ad}{bc}$$

The Odds Ratio for the Met. Office data given in Table 1 is 45.9.

The Odds Ratio Skill Score (*ORSS*) is given by:

$$ORSS = \frac{ad - bc}{ad + bc}$$

The Odds Ratio Skill Score for the Met. Office data given in Table 1 is 0.96. The *ORSS* varies between +1 and –1 where a score of 1 represents perfect skill and a score of zero represents no skill. Negative numbers imply that the forecasts were opposite to what was observed.

It is possible to assess the statistical significance of these verification statistics as shown in the Appendix. This also means that it is possible to show whether or not one service provider is significantly better than another.

Before targets can be considered for *PSS* and/or *ORSS* the value of the forecasts must considered: and these can then be incorporated into a performance related contract. Note that both the *PSS* and *ORSS* assume that Miss Rates and False Alarm Rates are of equal consequence. They are not of equal value however as we shall see.

5. The value of a forecast

Unlike skill, the value of a forecast depends on user requirements. Thompson & Brier (1955) proposed the simple cost/loss ratio for judging value. It can be applied to the following situations:

(a) where the effects of adverse weather on an operation and the cost of taking action to avoid weather damage are known in monetary terms;
(b) where the decision-maker's dissatisfaction with a loss is a linear function of the monetary value of the loss; and
(c) where the probability of occurrence of adverse weather is known precisely.

For winter road maintenance it should be possible to make reasonable estimates of (a) and (b) whereas assumption (c) is known after the event.

It is normal to denote the cost of taking action as *C*, in this case to salt the roads, and to denote the loss incurred as *L*, if the roads are not salted and accidents and delays occur, taking into account the savings made by not salting. On a given night if p is the expected probability of adverse weather (i.e. frost or snow) then:

if $p > C/L$ it will pay to salt the roads
if $p < C/L$ it will not pay to take action
if $p = C/L$ it doesn't matter either way

Obviously it is assumed that $0 < C/L < 1$ (i.e. that $C < L$). Relating this to the contingency table of Table 1 we find that:

Type 1 Errors = c with costs incurred = $c\ L$
Type 2 Errors = b with costs incurred = $b\ C$

Following Thornes (1999), if we take a benefit/cost ratio of 8:1 for winter maintenance of roads and assume that the cost C to a local authority of salting the roads for one night is £20,000, then the loss L incurred by not salting is likely to be £160,000, where $C/L = 0.125$. The frequencies of error given in Table 1 thus produce the following costs:

$$cL = 4 \times £160{,}000 = £640{,}000$$
$$bC = 6 \times £20{,}000 = £120{,}000$$

The total costs to the local authority due to errors in the forecast are thus estimated at £760,000. The cost for the nights when the roads were salted correctly is:

$$aC = 29 \times £20{,}000 = £580{,}000$$

The total cost for the winter therefore stands at £1.34 million.

It is usual to compare this to the costs that would have been incurred if no forecasts were issued and the roads were salted on all 77 marginal nights. This is given by 77 × £20,000 = £1.54 million. The forecasts therefore saved the local authority £1.54 million – £1.34 million = £0.2 million.

If the forecasts had been perfect then the roads would have been salted only on the nights when there was a frost (i.e. $a + c = 33$ nights). This would have cost $33 \times C = 33 \times £20{,}000 = £660{,}000$. Perfect forecasts would have saved the local authority £1.54 million – £0.66 million = £0.88 million.

To put these figures into perspective, note that if the roads were never salted the total loss would be:

$$(a + c)\, L = 33 \times £160{,}000 = £5.28 \text{ million}$$

To summarise, therefore, the expense E of the various options is as follows:

- with perfect weather forecasts, the cost to a local authority would be $E(P)$ = £0.66 million
- with the quoted accuracy of Table 1, the cost would be E(A) = £1.34 million
- if the roads were salted every marginal night, the cost would be $E(S)$ = £1.54 million
- if the roads were never salted, the cost would be $E(N)$ = £5.28 million

These figures are only illustrative but they show the value of accurate forecasts and that Type 1 errors are to be avoided. There is still much to be gained by increasing the accuracy of the forecasts. One way to reduce Type 1 errors is to issue more forecasts of frost, but that will increase the chance of a Type 2 error. This is acceptable up to a limit because the cost of a Type 2 error is so much less than that of a Type 1 error. In order to compare the quality and value of forecast providers we need an index that takes into account the number of Type 1 and Type 2 errors as well as the size of the Cost/Loss ratio. The relative value V of a forecast system, as defined by Richardson (2000), compares the mean expense ME of using a forecast with the mean expense caused by the climate such that:

$$V = \frac{ME(climate) - ME(forecast)}{ME(climate) - ME(perfect)}$$

V will have a value of 1 if the forecast system is perfect and will have a value of zero if the forecast is no better than climatology. For the purpose of this article, to present a simple approach that can be understood by users, the Value Index (V) is defined as:

$$V = \frac{E(without\,forecast) - E(forecast)}{E(without\,forecast) - E(perfect)}$$

Where $E(without\ forecast)$ can relate to climate, persistence or chance, or whatever is used to compare with the forecast. For example one could compare the expense of salting all marginal nights or salting all nights or not salting at all, whichever is the cheapest method that does not use a forecast. In the example used above it is cheaper to salt all marginal nights ($E(S)$ = £1.54 m) than not to salt at all ($E(N)$ = £5.28 m). Therefore we can state that:

$$V = \frac{E(S) - E(A)}{E(S) - E(P)}$$

Using the figures from above:

$$V = \frac{1.54 - 1.34}{1.54 - 0.66} = 0.23$$

It can be shown that the V can be simply calculated as follows

$$V = \frac{(c + d) - c / p}{n - W}$$

where $p = C/L$.
n = total number of nights RST ≤ 5° C
W = winter severity (a + c)

Let us summarise the proposed quality and value statistics in Table 2.

The Value Index normally varies between zero and 1. If the V is negative it means that the forecasts are so poor that it would be more cost effective to salt the roads every marginal night.

Care should be taken to define the size of d, in other words to ensure that only marginal decisions are included in the contingency table. For example, in Tables 1 and 2 above, only 77 out of 151 winter nights (1 November to 31 March) were considered when the minimum road surface temperature was observed to fall to 5° C or below. Otherwise d would be very large and make the calculations less

Table 2. *Forecast quality and value statistics for High Eggborough for 77 nights during the winter of 1995/96*

Statistic	Value
Percent Correct (*PC*)	87%
Bias (*B*)	1.06
Miss Rate (*MR*)	0.12
False Alarm Rate (*FAR*)	0.14
Peirce Skill Score (*PSS*)	0.74
Odds Ratio Skill Score (*ORSS*)	0.96
Total Cost ($aC + bC + cL$)	£1.34 m
Potential cost with no forecast (nC)	£1.54 m
Potential cost of perfect forecast ($(a + c)C$)	£0.66 m
Value Index (*V*)	0.23

meaningful. The term (n–W) will be relatively constant and allow spatial comparisons of the Value Index across regions.

The size of *V* will depend upon $p = C/L$. Table 3 shows the effect of varying p between 0.1 and 1.0 for the Met. Office values given in Table 2.

For the setting of performance targets it is necessary to agree on a realistic *C/L* ratio first and then set reasonable targets based on the likely number of Type 1 and Type 2 errors.

6. The Base Rate Effect and snow

Some weather events occur much more frequently than others and this can affect the quality of the forecasts. This is called the Base Rate Effect. Mathews (1997), for example, has shown that that when 'rain' is forecast it is less likely to be accurate than when the forecast is 'no rain'. To illustrate this effect, we can use the snow forecasts produced by two forecast providers for the same High Eggborough site during the winter of 1995/96 (see Table 4).

The number of days with snow falling in the UK is many less than the number of days when the road surface temperature falls below zero. Thus, during the winter of 1995/96 at High Eggborough, there were 77 nights when the road surface temperature fell to 5° C or below; frosts occurred on 33 of these nights and snow was recorded on 16 days. In most winters in this part of the UK there are less than 10 days with snow. The *C/L* ratio and the *V* are therefore not normally considered for snow in the UK but are important forecast value indicators in climates where snow is more prevalent.

In Table 4, although the Percent Correct of Provider A looks better at 81% than that of Provider B at 79% and the Bias of Provider A is 1.0 compared to the Bias of 1.88 for Provider B, the rest of the statistics tell a very different story. Provider A has a Miss Rate of 0.44 compared to Provider B's Miss Rate of only 0.06. Thus one has to be very careful in interpreting the snow forecasts. These contrasting results for Provider A are a consequence of the small Base Rate of only 16 days when snow fell out of 77 days. The results for Provider B show the benefits of a large positive bias, i.e. 'over-forecasting' snow, which reduces the chance of a Type 1 error. The Peirce Skill Score, the Odds Ratio Skill Score and the Value Index are much higher for Provider B and clearly show that Provider B provided better snow forecasts than Provider A. There is still much room for improvement on the part of both Providers.

Table 3. *Values of V for a variety of C/L values*

C/L	V
0.1	0.05
0.125	0.23
0.2	0.50
0.4	0.73
0.6	0.80
0.8	0.84
1.0	0.86

Table 4. *Snow forecasts for High Eggborough for 77 nights during the winter of 1995/96*

Provider A		
	Observed	
Forecast	Snow	No snow
Snow	9	7
No snow	7	54
Provider B		
	Observed	
Forecast	Snow	No snow
Snow	15	15
No snow	1	46

Statistic	Provider A	Provider B
Percent Correct	81%	79%
Bias	1.00	1.88
Miss Rate	0.44	0.06
False Alarm Rate	0.12	0.25
Peirce Skill Score	0.45	0.69
Odds Ratio Skill Score	0.82	0.96
Value (*C/L*=0.125)	0.08	0.64

7. Conclusion

With the use of a simple contingency table a number of very useful statistics can be calculated by the customer. These results can be written into performance-related contracts, or at the very least be demanded from the weather forecast service providers at the end of the season. Also, these statistics can be used to choose the best forecast provider, and if a new provider enters the market, then a performance-related contract would safeguard against possible poor performance.

The Value Index should make the setting of value targets more understandable but it should be noted that it is very dependent upon the cost-loss ratio. It is important therefore that both the customer and the forecast provider agree on this value before entering into a contract. The Peirce Skill Score and the Odds Ratio Skill Score should only be used for setting performance targets if C/L is close to 1 (in other words, if the Miss Rates and False Alarm Rates are of similar economic consequence to the user).

Although the examples used in this paper have been drawn from road weather forecasts the general results can be easily tailored for other forecast services.

Appendix. Is forecast skill simply due to chance?

Forecasting is a game of chance in which one uses inside information to try to reduce the odds of an erroneous forecast. Skill scores calculated from the contingency tables are only long-run estimates of the true skill of the forecasts and often contain sampling uncertainties. For this reason, impressively good scores can sometimes be obtained purely by chance (flukes), especially if one compiles the score using only a small number of forecasts. For example, winning at roulette a few times does not imply that one has any real skill that will enable one to win in the future! Statistics can be used to help reject flukey skill scores that could have happened by chance sampling fluctuations. This section will discuss briefly how statistical error estimates (confidence limits) can be used to judge the Miss and False Alarm Rates and the Peirce and Odds Ratio Skill Scores.

(a) Sampling error in Hit and Miss Rates

Estimates of false alarm rates and miss rates contain sampling errors. Table A1 gives estimates of these errors obtained using the 'score confidence interval' discussed in Agresti & Coull (1998).

For example, the Miss Rate for the Met. Office forecasts is 0.12 (= 4/33) calculated with 33 events, and so from Table 2 has a standard error of about 0.057. Therefore the Miss Rate is slightly more than 2 standard errors above zero and so is significantly different from a Miss Rate of zero (perfect forecast) at 95% confidence.

(b) Standard error in the Peirce Skill Score

Assuming independence of the false alarm and miss rates the standard error in the Peirce Skill Score is simply the square root of the sum of the squared standard errors in the Miss and False Alarm Rates. For example, the Met. Office forecasts have similar Miss and False Alarm Rates of 0.12 and 0.14 respectively with typical standard errors of about 0.057 and 0.049 and therefore the standard error in the estimated Peirce Skill Score of 0.74 is given by $(0.057^2+0.049^2)^{\frac{1}{2}} = 0.075$. The Peirce Skill Score for the Met. Office forecasts is therefore much more than two standard errors above zero and at 95% confidence the forecasts have skill.

(c) Significance of the Odds Ratio Skill Score

The Odds Ratio Skill Score can be tested by considering that the logarithm of the Odds Ratio is approximately Gaussian distributed. Using the data calculated in section 4.3(b), the log(e) of the Odds Ratio (i.e. ln 45.9 = 3.83) and the Asymptotic Standard Error (ASE) on the log of the Odds Ratio is given by $ASE = 1/\sqrt{m}$ where m is the harmonic mean of a, b, c, d given by:

$$\frac{1}{m}=\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}$$

Therefore using the values in Table 1 gives:

$$\frac{1}{m}=\frac{1}{29}+\frac{1}{6}+\frac{1}{4}+\frac{1}{38}$$

Therefore $m = 2.08$ and $ASE = 1/\sqrt{m} = 1/\sqrt{2.08} = 0.69$.

The log(e) of the odds ratio 3.83 is greater than 1.96 ASE (i.e. 1.36) and therefore we can state with 95% confidence that the observed and forecast values are not independent.

Table A2 gives the minimum values of Odds Ratio Skill

Table A1. *Standard error in estimated miss or false alarm rate calculated using the 95% score confidence interval as discussed in Agresti & Coull (1998).*

Events	Estimated Miss or False Alarm Rate										
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
5	0.111	0.134	0.150	0.160	0.166	0.168	0.166	0.160	0.150	0.134	0.111
10	0.071	0.099	0.116	0.126	0.132	0.134	0.132	0.126	0.116	0.099	0.071
20	0.041	0.070	0.086	0.095	0.101	0.102	0.101	0.095	0.086	0.070	0.041
30	0.029	0.057	0.071	0.080	0.084	0.086	0.084	0.080	0.071	0.057	0.029
40	0.022	0.049	0.062	0.070	0.074	0.076	0.074	0.070	0.062	0.049	0.022
50	0.018	0.043	0.056	0.063	0.067	0.068	0.067	0.063	0.056	0.043	0.018
100	0.009	0.030	0.040	0.045	0.048	0.049	0.048	0.045	0.040	0.030	0.009
500	0.002	0.013	0.018	0.020	0.022	0.022	0.022	0.020	0.018	0.013	0.002
1000	0.001	0.010	0.013	0.014	0.015	0.016	0.015	0.014	0.013	0.010	0.001

Table A2. *Minimum value of the Odds Ratio Skill Score required in order to have real skill at various confidence levels. Values have been estimated using the asymptotic Gaussian distribution of the logarithm of the odds (Agresti, 1996) where m is the harmonic mean of the numbers of events in the contingency table (i.e. 1/m=1/a+1/b+1/c+1/d).*

m	Confidence that there is skill					
	50%	70%	90%	95%	99%	99.9%
1	0.000	0.256	0.565	0.676	0.822	0.913
2	0.000	0.183	0.424	0.524	0.676	0.798
3	0.000	0.150	0.354	0.442	0.586	0.712
4	0.000	0.130	0.310	0.390	0.524	0.648
5	0.000	0.117	0.279	0.352	0.478	0.599
10	0.000	0.083	0.200	0.254	0.352	0.453
20	0.000	0.059	0.142	0.182	0.254	0.332
30	0.000	0.048	0.116	0.149	0.209	0.275
40	0.000	0.041	0.101	0.129	0.182	0.240
50	0.000	0.037	0.090	0.116	0.163	0.215
100	0.000	0.026	0.064	0.082	0.116	0.153
500	0.000	0.012	0.029	0.037	0.052	0.069
1000	0.000	0.008	0.020	0.026	0.037	0.049

Score (Q) needed in order to signify real forecast skill at different levels of confidence.

The value m is the harmonic mean of the number of events, and is *always* smaller than the smallest number of events in the contingency table. Take care not to forget the reciprocal on the left-hand side! For example, the Met. Office forecasts have $a = 29$, $b = 6$, $c = 4$ and $d = 38$ and so the harmonic mean is 2.09. Their *ORSS* is 0.96 which therefore exceeds the 99.9% confidence limit in Table A2 for $m = 2$. In other words, at 99.9% confidence the skill of the Met. Office forecasts is not due to chance sampling of the events in the contingency table. More discussion about the sampling errors of skill scores can be found in Stephenson (2000).

Acknowledgements

The authors would like to thank the editor and the reviewers for their valuable comments on the first draft of this paper.

References

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. John Wiley & Sons, Inc., 290 pp.

Agresti, A. & Coull, B. A. (1998). Approximation is better than 'Exact' for interval estimation of binomial proportions, *The Am. Stat.*, **52:** 1–7.

Halsey, N. G. J. (1995). Setting verification targets for minimum road temperature forecasts, *Meteorol. Appl.*, **2:** 193–197.

Matthews, R. (1997). How right can you be?, *New Scientist*, **2072:** 28–31.

Mead, J. (1998). There's a killer on the loose. In *Proc. of the Cold Comfort 98 Conference*, September 1998, Northampton. (Organised by the *Surveyor Magazine*.)

Millington, (1987). Weather forecasting and 'The Limitless Seas'. *The Law Quarterly Review*, **103:** 234–245.

Mylne, K. R. (1999). The use of forecast value calculations for optimising decision making using probability forecasts. In *Proc. of 17th Conference on Weather Analysis and Forecasting,* 13–17 September 1999, Denver, Colorado, 235–239.

Richardson, D. (2000). Skill and economic value of the ECMWF ensemble prediction system. *Q. J. R. Meterol. Soc.*, **126:** 649–668

Stanski, H. R., Wilson, L. J. & Burrows, W. R. (1989). Survey of common verification methods in meteorology. *WMO/TD-No. 358, World Meteorological Organisation*, Geneva, Switzerland, 114 pp.

Stephenson, D. B. (2000). Use of the 'odds ratio' for diagnosing forecast skill. *Wea. and Forecasting*, **15:** 221–232.

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

Thornes, J. E. (1995). A comparative real-time trial between the Met. Office and Oceanroutes to predict road surface temperatures. *Meteorol. Appl.*, **2:** 113–119.

Thornes, J. E. (1996). The quality and accuracy of a sample of public and commercial weather forecasts in the UK. *Meteorol. Appl.*, **3:** 63–74.

Thornes, J. E. (1997). Transport. In *Applied Climatology*, A. Perry & R. Thomson (eds), Routledge, ch. 12.

Thornes, J. E. (1999). UK road salting – an international benefit/cost review. *Journal of the Institute of Highways and Transportation*, July/August: 22–26.

Thornes, J. E. & Proctor, E. A. J. (1999) Persisting with persistence: the verification of Radio 4 weather forecasts. *Weather*, **54:** 311–321.

Wilks, D. S. (1995). *Statistical Methods in the Atmospheric Sciences*. Academic Press, 465 pp.