Meteorol. Appl. 12, 343–355 (2005) doi:10.1017/S135048270500191X

Verification and value of the Australian Bureau of Meteorology township seasonal rainfall forecasts in Australia, 1997–2005

A. L. Vizard[1], G. A. Anderson[1] & D. J. Buckley[2]
[1]*School of Veterinary Science, University of Melbourne, 250 Princes Highway, Werribee, 3030 Australia.*
[2]*School of Agricultural and Veterinary Sciences, Charles Sturt University, Locked Bag 588, Wagga Wagga, 2678 Australia*
Email: a.vizard@unimelb.edu.au

We verified the Bureau of Meteorology's seasonal rainfall forecasts for 262 townships throughout Australia, from its inception in June 1997 to May 2005. The results indicate that the forecasting system had low skill. Brier Skill Score and the receiver operating characteristic values were uniformly close to the no skill value. Forecast variances were consistently small. The overall observed variance was 0.0048, 2.1% of the variance of a perfect system. The estimate of the gradient of the outcome against forecast was 0.42 and was imprecise. Definitive statements about bias cannot be made. The value of the forecasts for decision-makers was estimated using value score curves, calculated for six forecast scenarios. All curves indicated that no economic benefit could have been reliably derived by users of the seasonal rainfall forecasts, with the exception of users with decisions triggered by a small shift in the forecast from climatology, in which case small economic gains may have occurred. Small value scores were associated with the observed forecast variance, not the observed bias. We examined the expected change in value scores associated with any future increase in forecast variance. This showed that a moderate increase from the observed variance would bring limited benefits. Substantial value to a broad range of users will only occur with a large increase in forecast variance. To deliver this, new lead indicators with markedly better predictive characteristics may need to be developed for the seasonal rainfall forecasting system.

1. Introduction

Many Australian industries, such as agriculture, face high levels of uncertainty and may suffer substantial losses caused by variations in seasonal rainfall. Seasonal rainfall forecasts have the potential to improve decision-making within affected industries and consequently reduce the losses associated with variation in seasonal rainfall.

Each month since June 1997, the Australian Bureau of Meteorology has issued three-month seasonal rainfall forecasts in two formats. One is a map of Australia depicting the probability of exceeding median rainfall or the probability that the seasonal rainfall will fall into each of three rainfall terciles. The other takes the form of a publication in which the probabilities of 'dry' and 'wet' seasons are published for 262 townships throughout Australia. For this purpose, 'dry' is defined as the lower tercile and 'wet' as the upper tercile. The forecasts are issued about 15 days prior to the commencement of the forecast period.

The statistical model used to derive these forecasts was based on a method of linear discriminant analysis (Wilks 1995) and is described fully in Drosdowsky & Chambers (2001). Predictor data used in the model have changed since its inception. Most importantly, in October 1998 indices of sea surface temperature replaced the Southern Oscillation Index as the principal predictor (Fawcett et al. 2005).

Information provided with the forecasts states that the forecasts should be useful in many farm planning operations and may assist decisions concerning which crops to plant, stocking rates over the coming season or management of drought. It is also claimed that the forecasts have a number of business applications, such as agribusiness, garden hose manufacture or futures market.

The utility of such forecasts to assist with rational decision-making is highly dependent upon the relationship between outcome and forecast as well as the distribution of the forecast probabilities. In this paper we

verify the Bureau's seasonal township rainfall forecasts against the observed rainfall from June 1997 to May 2005. Several standard verification methods, such as the Brier Skill Score (BSS) (Brier 1950; Toth et al. 2003), the receiver operating characteristic (ROC) value (Mason 1982) and the reliability curve (Wilks 1995) are presented. The value of the forecasts for decision-makers was estimated using value score curves (Wilks 2001). We also present the expected improvement in value scores associated with increased forecast variance and offer a new standard diagnostic verification parameter.

2. Materials and methods

2.1. Data

2.1.1. Forecasts

Each month, commencing in June 1997, forecasts of the probability of 'dry' and 'wet' seasons were provided in the Australian Bureau of Meteorology's Seasonal Climate Outlook publications, prepared by the Bureau of Meteorology's National Climate Centre, for 262 townships located throughout Australia. A 'dry' season was defined as less than the 33rd percentile (lower tercile) of seasonal rainfall for each particular township. A 'wet' season was defined as greater than the 67th percentile (upper tercile) of seasonal rainfall for each particular township. A separate analysis was conducted for 'dry' and 'wet' forecasts.

The permissible forecast values were (0.00, 0.01, 0.02, . . . 0.98, 0.99, 1.00). In order to have forecast periods that were mutually exclusive, we included only the autumn (March to May), winter (June to August), spring (September to November) and summer (December to February) forecasts in this analysis. The series considered in this paper started in winter 1997 and ended in autumn 2005. Consequently there were a total of 32 forecasting periods, each containing 262 township forecasts that were potentially available for analysis.

2.1.2. Rainfall outcomes

Rainfall records from the 262 towns for which the Bureau provided seasonal forecasts were used to verify the forecasts. The rainfall records for each township were obtained from the weather station that the Australian Bureau of Meteorology had designated to the township. There were 74 designated weather stations that had all or substantial parts of their rainfall records unavailable for the duration of the study. An alternative weather station within 20 kilometres was sought for each of these towns. Suitable alternative stations were found for 68 towns, 63 of which were less than 10 kilometres from the designated weather station leaving 256 valid towns for verification. There were some forecasting periods that did not have a rainfall record, resulting in 7598 valid observations. For each of the 256 valid weather stations the 33rd and 67th rainfall percentile for each season was obtained from the records of the Australian Bureau of Meteorology. These records were then used to calculate if the observed rainfall was in the lower or upper tercile.

Figure 1. *Map of Australia showing the eight regions used in the analysis.*

The 256 towns were allocated to one of eight regions, based on a previous empirical orthogonal function analysis of the rainfall data (Drosdowsky & Chambers 2001). The eight regions are shown in Figure 1. There were 75, 54, 14, 25, 41, 16, 23 and 8 towns in regions one to eight, respectively.

2.2. Statistical analysis

Descriptive statistics of forecasts were calculated for the overall results, for each of the eight regions and for each of the four seasons.

The Brier Skill Scores (Brier 1950; Toth et al. 2003) were calculated by using the sample climatology as the reference (Murphy & Winkler 1992). Empirical receiver operating characteristic values (Mason 1982) were derived using the Stata statistical program (StataCorp 2005).

The reliability curve was constructed after allocating the forecasts into probability categories with an interval range of 0.05. The nine categories were <0.205, 0.205–0.255, 0.256–0.305, 0.306–0.355, ..., 0.506–0.555, >0.555.

To quantify the linear association between outcome and forecast, an ordinary least squares regression using the individual 7598 observations was conducted (Murphy & Winkler 1992). The robust standard error

Table 1. *Six scenarios used to calculate value score curves according to the forecast distribution and linear relationship between the outcome and the forecast of a dry outcome.*

Scenario	Forecast distribution	Linear relationship between outcome and forecast
1 (overall, as observed)	Overall observed ($\sigma_f^2 = 0.0048$)	Overall observed Pr(Dry) = 0.20 + 0.42 Forecast
2 (overall, modelled)	Beta model of overall observed ($\sigma_f^2 = 0.0048$)	Overall observed Pr(Dry) = 0.20 + 0.42 Forecast
3 (overall, recalibrated)	Beta model of recalibrated overall forecast ($\sigma_f^2 = 0.00085$)	Perfect reliability Pr(Dry) = Forecast
4 (overall, perfect reliability)	Beta model of overall observed ($\sigma_f^2 = 0.0048$)	Perfect reliability Pr(Dry) = Forecast
5 (best case regional or seasonal scenario)	Beta model of region or season with maximum variance ($\sigma_f^2 = 0.0094$)	Perfect reliability Pr(Dry) = Forecast
6 (conservative case regional or seasonal scenario)	Beta model of region or season with minimum variance ($\sigma_f^2 = 0.0023$)	Overall observed Pr(Dry) = 0.20 + 0.42 Forecast

(Williams 2000) of the regression coefficients was calculated using the cluster option of the regress command in Stata software (StataCorp 2005). The robust standard error takes into account the non-independence of observations within a forecast period (cluster).

An estimate of the correlation of outcomes within each of the 32 forecasting periods was provided by the intra-class correlation coefficient (ICC). The analysis of variance estimator was used to calculate this (Ridout et al. 1999). The reciprocal of the variance inflation factor (VIF) was used to multiply the observed sample size of 7598 to obtain an estimate of the number of independent observations (Donner & Klar 2000).

$$\mathrm{VIF} = 1 + (\mathrm{m} - 1)\,\mathrm{ICC} \tag{1}$$

in which m is the mean number of townships per forecast period (m = 237).

The mean forecast given that the outcome did not occur was subtracted from the mean forecast given that the outcome occurred. This parameter, the forecast difference, is offered as a diagnostic tool since it is simple to calculate and has a fundamental relationship with other important verification parameters as outlined below.

$$\mu_{f|x=1} - \mu_{f|x=0} = \beta\left(\sigma_f^2/\sigma_x^2\right) \tag{2}$$

in which
$\mu_{f|x=1}$ is the mean forecast when the outcome occurred
$\mu_{f|x=0}$ is the mean forecast when the outcome did not occur
β is the gradient of the ordinary least squares regression of outcome versus forecast
σ_f^2/σ_x^2 is the observed variance/variance of a perfect forecasting system. This is called the variance ratio
Other diagnostic verification parameters (Murphy & Winkler 1992) were also calculated.

2.3. Value score

Value scores are based on a simple cost/loss model (Wilks 2001). The value score (VS) of a forecast system 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 (Wilks 2001):

$$\text{Value Score} = \frac{\text{Expected Expenses Forecast System} - \text{Expected Expenses Climatology}}{\text{Expected Expenses Perfect Forecast} - \text{Expected Expenses Climatology}}$$

The value score was estimated using the following equations (Wilks 2001).

$$VS = \frac{\frac{C}{L}(p_{11} + p_{10} - 1) + p_{01}}{\frac{C}{L}(\pi - 1)} \quad if \quad \frac{C}{L} < \pi \tag{3a}$$

$$VS = \frac{\frac{C}{L}(p_{11} + p_{10}) + p_{01} - \pi}{\pi\left(\frac{C}{L} - 1\right)} \quad if \quad \frac{C}{L} > \pi \tag{3b}$$

in which
C = the cost of protecting against the effects of adverse weather
L = the losses that result from the occurrence of adverse weather without the benefit of protection
π = the probability of adverse weather (climatology)
p_{11} = the probability of adverse weather occurring following a 'yes' forecast
p_{10} = the probability of adverse weather not occurring following a 'yes' forecast
p_{01} = the probability of adverse weather occurring following a 'no' forecast

Value score curves were derived for six forecast scenarios. These scenarios varied in the assumption of the forecast distribution and also in the relationship between outcome and forecast, as described in Table 1.

Table 2. *Means and variances of outcomes (x) and forecasts (f). The gradient ± robust standard error and intercept of the ordinary least squares regression of outcome versus forecast.*

	Mean					Variance		Var Ratio		
	μ_x	μ_f	$\mu_{f\|x=0}$	$\mu_{f\|x=1}$	Diff	σ_f^2	σ_x^2	σ_f^2/σ_x^2	Gradient	Intercept
Overall										
'Dry'	0.34	0.34	0.336	0.345	0.0091	0.0049	0.2252	0.0216	0.42 ± 0.36	0.20
'Wet'	0.31	0.33	0.328	0.342	0.0148	0.0053	0.2122	0.0252	0.59 ± 0.26	0.11
Region ('Dry' forecasts)										
1	0.35	0.33	0.325	0.327	0.0016	0.0044	0.2282	0.0193	0.08 ± 0.57	0.33
2	0.34	0.35	0.344	0.350	0.0062	0.0046	0.2230	0.0205	0.30 ± 0.76	0.23
3	0.29	0.33	0.322	0.334	0.0115	0.0030	0.2062	0.0144	0.80 ± 0.55	0.03
4	0.24	0.34	0.341	0.343	0.0016	0.0028	0.1805	0.0156	0.10 ± 0.62	0.20
5	0.41	0.37	0.363	0.368	0.0045	0.0050	0.2425	0.0204	0.22 ± 0.67	0.33
6	0.36	0.33	0.323	0.347	0.0237	0.0094	0.2318	0.0407	0.58 ± 0.49	0.17
7	0.34	0.33	0.325	0.352	0.0274	0.0054	0.2238	0.0243	1.13 ± 0.56	−0.04
8	0.31	0.34	0.332	0.353	0.0205	0.0023	0.2156	0.0109	1.88 ± 0.78	−0.32
Season ('Dry' forecasts)										
Autumn	0.43	0.33	0.326	0.328	0.0018	0.0029	0.2452	0.0117	0.16 ± 0.58	0.38
Winter	0.38	0.36	0.357	0.368	0.0115	0.0058	0.2365	0.0243	0.47 ± 0.65	0.21
Spring	0.26	0.35	0.341	0.367	0.0260	0.0062	0.1919	0.0323	0.81 ± 0.45	−0.02
Summer	0.30	0.32	0.321	0.322	0.0005	0.0036	0.2094	0.0171	0.03 ± 0.52	0.29

In five of the scenarios it was assumed that the distribution of forecasts followed a beta distribution. The two parameters of the beta distribution, a and b, were estimated by the method of maximum likelihood, using the Stata software package (StataCorp 2005). In Scenario 1, value scores were estimated for cost/loss ratios between zero and one at intervals of 0.01; in all other scenarios value scores were estimated at intervals of 0.001.

The expected improvement in value scores associated with increased forecast variance was explored assuming perfectly reliable forecasts. It was also assumed that the distribution of forecasts followed a beta distribution. The integral of the value score curve and the mean value score for specific ranges of the curve were estimated across the entire possible range of variance ratios.

3. Results

All descriptive and inferential results from the separate analysis of 'dry' and 'wet' forecasts were very similar. Consequently, we present complete results of the analysis of the 'dry' forecasts, but only the overall results from the analysis of the 'wet' forecasts.

3.1. Descriptive statistics

Table 2 shows complete descriptive statistics of the 'dry' forecasts, including means, variances and estimates of the linear regression coefficients of 'dry' outcome versus forecast as well as the overall results from the analysis of the 'wet' forecasts. The proportion of 'dry' observations was 0.342 (sample climatology, n = 7598). The mean forecast was 0.340 (variance 0.0048, n = 8384). The mean forecast in the 2601 'dry' outcomes was 0.345 and was 0.336 in the 4997 'non-dry' outcomes.

The overall variance of the 'dry' forecasts was 0.0048. The observed variance of forecasts as a proportion of the variance of forecasts from a perfect forecasting system was 2.1%. The maximum variance of forecasts for any one town was 0.016 at Lancelin (Region 6), with 3% of towns having a variance of the forecasts of greater than 0.01. The minimum variance for any one town was 0.00048 at Giles (Region 2), with 17% of towns having a variance of the forecasts of less than 0.0025.

Figure 2 illustrates the observed distribution of the 'dry' forecasts. The zero, 5th, 10th, 25th, 50th, 75th, 90th, 95th and 100th percentiles of the forecasts for the probability of a 'dry' season were 0.13, 0.22, 0.25, 0.30, 0.33, 0.38, 0.43, 0.46 and 0.66, respectively.

The estimates of the two parameters of the beta distribution that was fitted to the overall 'dry' forecasts were a = 15.61 and b = 30.33.

The reliability curve is shown in Figure 3.

The intraclass correlation coefficient of 'dry' outcomes within each of the 32 forecasting periods was 0.18, indicating that observations within a forecasting period were not independent and 0.005 within each of the 256 townships, indicating that observations between forecasting periods were independent.

3.2. Diagnostic verification parameters

Table 3 provides the mean square error (MSE), a decomposition of the MSE, the BSS and the area under

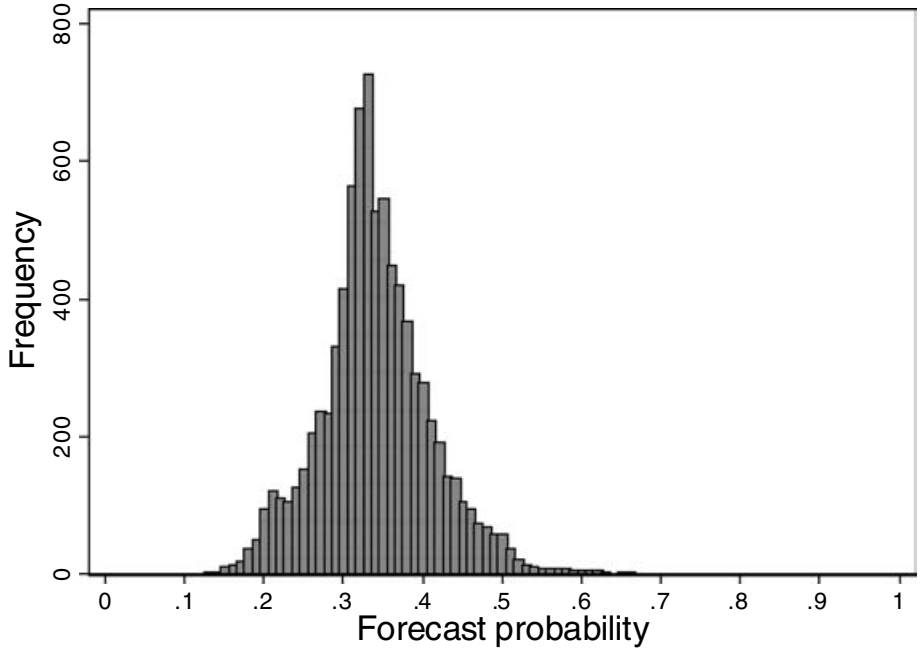

Figure 2. *Distribution of the 'dry' forecast probabilities from June 1997 to May 2005 (n = 8384). The interval range of the categories is 0.01.*

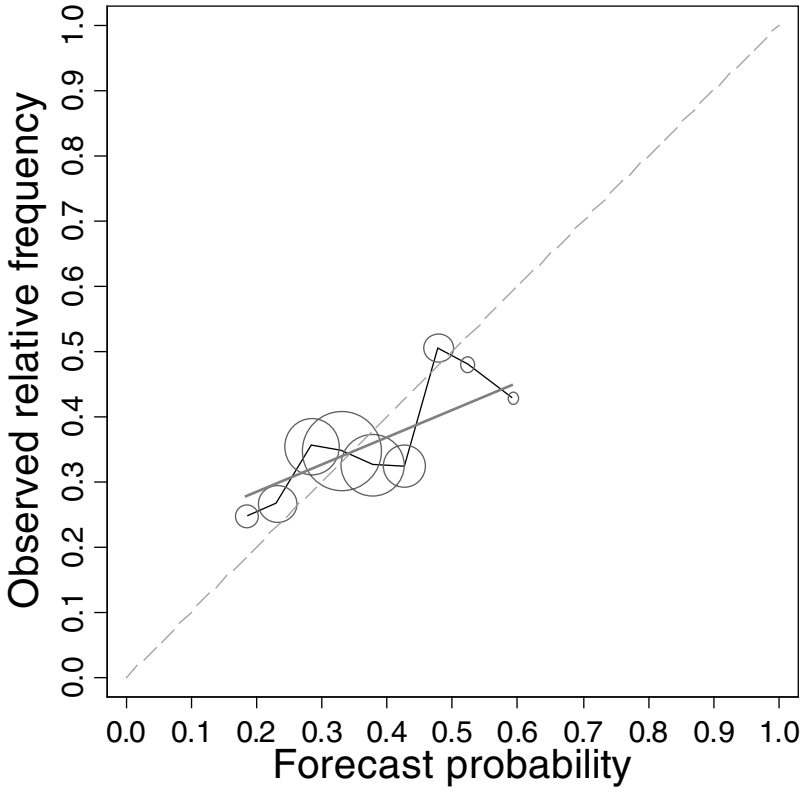

Figure 3. *Reliability curve for 'dry' forecasts, June 1997 to May 2005 (n = 7598). The size of each circle is proportional to the number of observations. The dashed line is the line of perfect reliability and the solid line is the ordinary least squares regression line based on the 7598 individual observations.*

the receiver operating characteristic curve (ROC values) for 'dry' forecasts. The overall results of the analysis for 'wet' forecasts are also provided in Table 3. All measures show a low level of skill.

3.3. Value scores

Table 4 provides a summary of the value scores for each of the six scenarios. Value score curves are shown in

Table 3. *Decomposition of the Mean Square Error (MSE), Brier Skill Score (BSS) and receiver operating characteristic (ROC) curve.*

	No. Obs	MSE (Accuracy)	= x (1−x) (Uncertainty)	+ $E_f(\mu_{x\|f} - f)^2$ (Reliability)	− $E_f(\mu_{x\|f} - \mu_x)^2$ (Resolution)	BSS	ROC
Overall							
'Dry'	7598	0.2259	0.2251	0.0054	0.0046	−0.003	0.53
'Wet'	7598	0.2120	0.2122	0.0046	0.0048	0.001	0.55
Region ('Dry' forecasts)							
1	2272	0.2325	0.2281	0.0156	0.0113	−0.019	0.51
2	1581	0.2248	0.2229	0.0152	0.0133	−0.009	0.50
3	411	0.2052	0.2057	0.0145	0.0149	0.002	0.55
4	670	0.1937	0.1802	0.0269	0.0134	−0.075	0.50
5	1187	0.2472	0.2422	0.0296	0.0247	−0.020	0.48
6	512	0.2307	0.2313	0.0267	0.0273	0.003	0.57
7	712	0.2166	0.2235	0.0229	0.0297	0.031	0.59
8	253	0.2090	0.2148	0.0263	0.0320	0.027	0.61
Season ('Dry' forecasts)							
Autumn	1878	0.2576	0.2451	0.0235	0.0109	−0.051	0.49
Winter	1917	0.2372	0.2364	0.0121	0.0113	−0.003	0.54
Spring	1902	0.1958	0.1918	0.0197	0.0156	−0.021	0.58
Summer	1901	0.2132	0.2093	0.0148	0.0109	−0.019	0.52

Table 4. *Integral of value score (VS) curve, maximum value score, minimum value score, mean value score for the following three ranges of the value score curve: 1) cost/loss ratio 0.0–0.14 and 0.67 and above; 2) cost/loss ratio 0.15–0.25 and 0.46–0.66; 3) cost/loss ratio 0.26–0.45 for the six scenarios that describe the forecast distribution and the linear relationship between the outcome and the forecast. See Table 1 for a description of the scenarios.*

Scenario	1	2	3	4	5	6
Integral	−0.0028	−0.0025	0.0020	0.0115	0.0238	−0.0011
Maximum VS	0.0428	0.0514	0.0514	0.1224	0.1732	0.0356
Minimum VS	−0.0254	−0.0267	0.0000	0.0000	0.0000	−0.0175
Mean VS range 1	−0.0000	−0.0000	0.0000	0.0000	0.0001	0.0000
Mean VS range 2	−0.0078	−0.0063	0.0000	0.0026	0.0135	−0.0010
Mean VS range 3	−0.0016	−0.0027	0.0103	0.0559	0.1015	−0.0040

Figures 4, 5 and 6. The integral of the value score over the entire cost/loss range is close to zero for all scenarios.

Figure 7 illustrates the expected improvement in the integral of the value score curve and the mean value score for various ranges of the value score curve with increased variance ratio. This relationship assumes the forecasts are perfectly reliable and that the forecast distributions follow beta distributions.

4. Discussion

4.1. Skill and distribution of forecasts

The results of this verification indicate that during the period of the study the seasonal rainfall forecasting system in Australia was not skilful. In contrast to the suggestion of Fawcett et al. (2005), there was no meaningful variation of skill either temporally or spatially. Brier Skill Scores and the ROC values were uniformly close to the values corresponding to no skill. The resolution of forecasts was particularly low (Table 3).

We suggest that the small variance of forecast probabilities was the principal cause for the low skill. The overall observed variance of 'dry' forecasts of 0.0048 was only 2.1% of the variance of a perfect forecasting system. Forecast variances were consistently small in every township, region and season. Since these results are independent of any estimates of outcome, are based on 32 mutually exclusive seasons and involve over 8000 forecasts, the finding of uniformly very small forecast variance is both unambiguous and fundamental.

Significantly, very few forecasts substantially deviated from climatology. For example, a forecast of 0.66 probability of a 'dry' outcome represents a doubling of the risk and a forecast of 0.17 represents a halving of the risk of a 'dry' outcome. Of the 8384 forecasts issued, there were no forecasts greater than 0.66 and only 30 forecasts (0.36%) less than 0.17. The 30 forecasts less

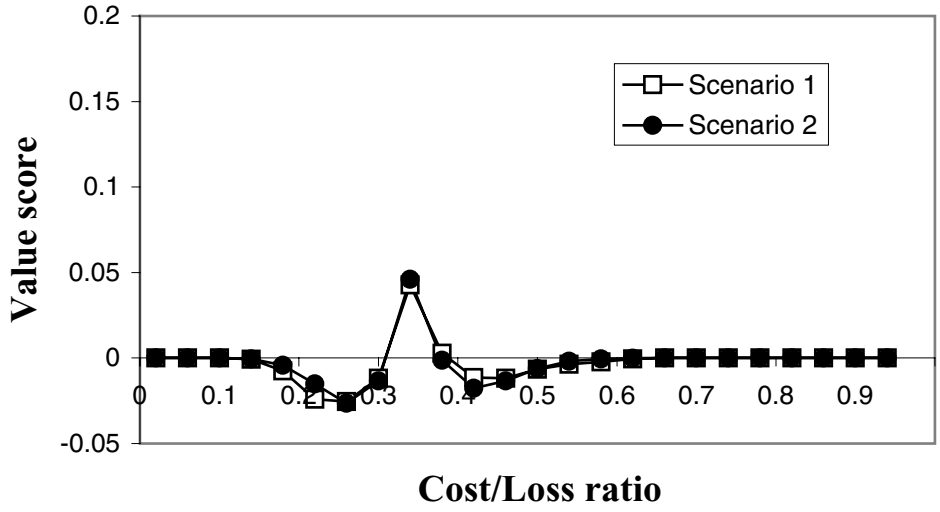

Figure 4. *Value score curves. Scenario 1 (open square) is the observed distribution of the forecasts and the observed overall linear relationship between outcome and forecast. Scenario 2 (filled circle) assumes a beta distribution with the observed overall variance of the forecasts and the observed overall linear relationship between outcome and forecast.*

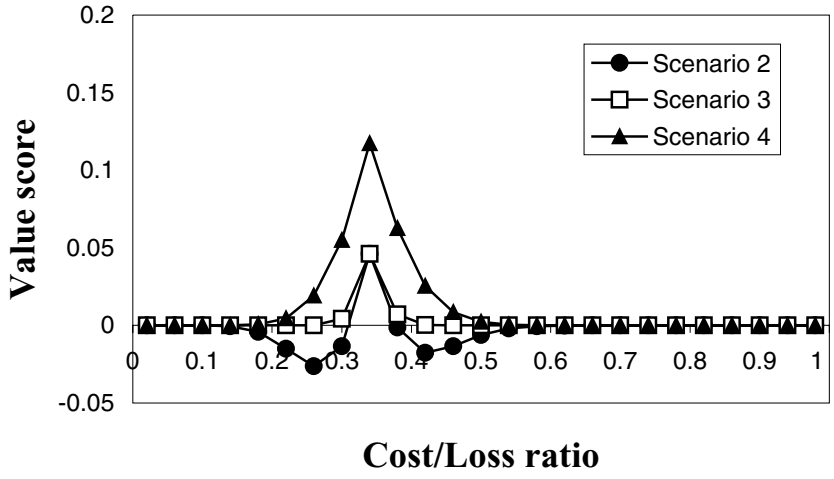

Figure 5. *Value score curves. Scenario 2 (filled circle) assumes a beta distribution with the observed overall variance of the forecasts and the observed overall linear relationship between outcome and forecast. Scenario 3 (open square) assumes a beta distribution of the recalibrated forecasts and assumes perfect reliability. Scenario 4 (filled triangle) assumes a beta distribution of the forecasts with the observed overall variance and assumes perfect reliability.*

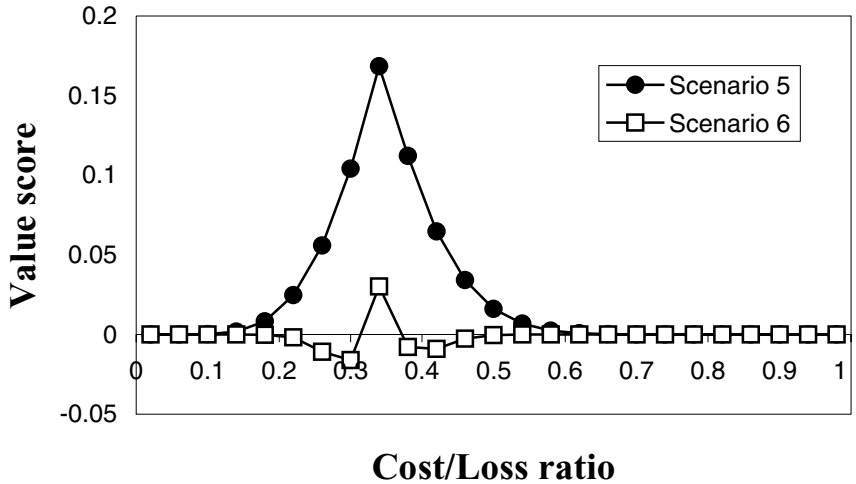

Figure 6. *Value score curves. Scenario 5 (filled circle) assumes a beta distribution of the forecasts with the maximum observed variance for any region or season and assumes perfect reliability. Scenario 6 (open square) assumes a beta distribution of the forecasts with the minimum observed variance for any region or season and assumes the observed overall linear relationship between outcome and forecast.*

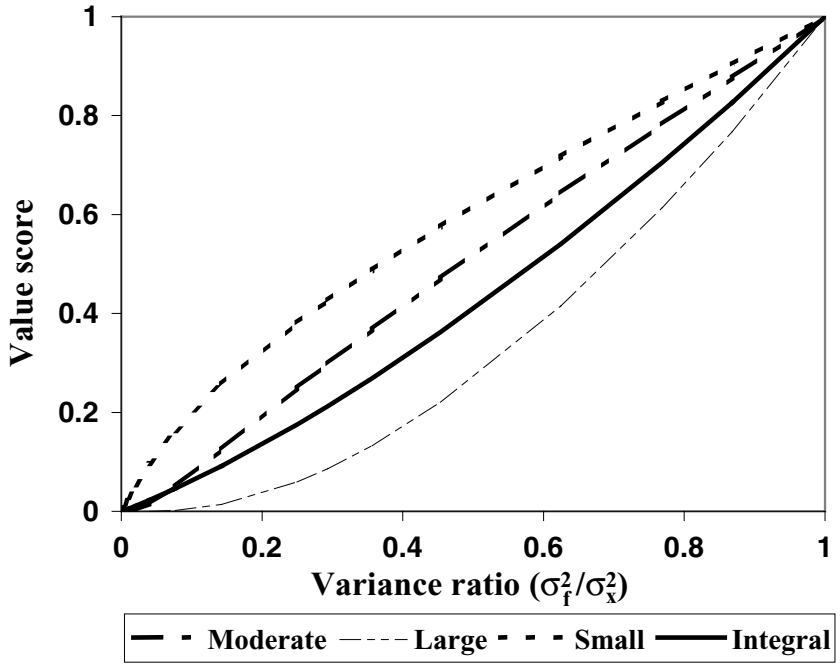

Figure 7. *Value score as a function of the variance ratio (σ_f^2/σ_x^2) for perfectly reliable forecasts in which the forecast distribution follows a beta distribution. Integral (solid line) is the integral of the value score curve. Moderate (thick dash dot) is the mean value score over the cost/loss ratio range 0.15–0.25 and 0.46–0.66. Large (fine dash dot) is the mean value score over the cost/loss ratio range 0.0–0.14 and 0.67 and above. Small (dotted line) is the mean value score over the cost/loss ratio range 0.26–0.45.*

than 0.17 were distributed across a range of townships, regions and seasons.

Although this is the first published verification of township seasonal rainfall forecasts in Australia, Fawcett et al. (2005) verified gridded tercile seasonal rainfall forecasts for Australia in a 4.5-year period commencing November 1998. Linear Error in Probability Space (LEPS) scores were close to zero for most of Australia, consistent with the results presented here.

The observed difference between the mean forecast given a dry outcome and the mean forecast given a not dry outcome (forecast difference) was consistently close to zero (Table 2). This also indicates low skill.

The forecast difference (Equation 2) is offered as a standard diagnostic verification parameter. The forecast difference has a domain between -1 and 1. A perfect forecast always has a forecast difference of 1; a valueless forecast has a forecast difference of 0 and a totally incorrect forecasting system always has a forecast difference of -1. In the case of an unbiased forecast, the forecast difference is equal to the Brier Skill Score and is also equal to the ratio of the variance of the observed forecast to the variance of a perfect forecast (variance ratio). Since there is a relationship between the integral of the value score curve and the variance ratio (Figure 7), the same relationship holds true for the integral of the value score curve and the forecast difference for perfectly calibrated forecasts.

Verification of forecasts inevitably involves a compromise between obtaining a sufficient number of observations for verification and maintaining the relevance of the verification to current users of the forecasts. Although 7598 observations were used in this analysis, these were not independent. The main correlation was between towns within a forecasting period. The intraclass correlation coefficient was 0.18. The effective sample size was estimated, using the variance inflation factor (Donner & Klar 2000), as being equivalent to about 174 independent observations. This was considered sufficient for most verification purposes.

During the eight-year period of this study the Bureau substantially altered its forecasting methodology. In March 2000, the use of predictors was changed and the model data set was expanded (Fawcett et al. 2005). This analysis is indifferent to the methodology used to generate forecasts, but we note that the variance of the 'dry' forecasts prior to March 2000 was 0.0065 and was 0.0038 afterwards. This suggests that there was no improvement in this important metric following this alteration to forecasting methodology.

A number of the 262 towns for which the Bureau provided seasonal forecasts did not have corresponding rainfall records from their designated weather station. Most of the unavailable records were from designated weather stations that had been closed and replaced with another weather station in the same township. Consequently, rainfall records for alternative weather

stations, usually within the same town and always within 20 kilometres were generally available (for example, Bunbury Racecourse substituted for Bunbury Post Office). Since all substitute weather stations were well within the forecast resolution (Fawcett et al. 2005), the use of these stations leads to increased precision of estimates without causing bias.

Of the 7598 valid observations, 49 had a recorded rainfall of zero mm and a lower tercile seasonal rainfall of zero mm. In the results presented here, these observations were categorised as 'not dry'. Very similar results were obtained when an analysis was conducted in which these observations were classified as 'dry'.

4.2. Reliability

For both 'dry' and 'wet' forecasts the point estimates of the gradient of the linear regression between overall outcome and forecast probability were close to the no skill gradient of 0.5 (Hsu & Murphy 1986). This suggests that the forecasts were overconfident. That is, for any given forecast probability except climatology, the outcome was closer to the climatological value. Overconfident forecasting can impart negative economic value to users of the forecasts (Wilks 2001).

The estimates of the gradients were imprecise, as indicated by the large standard errors associated with each estimate (Table 2). Consequently, definitive statements about bias cannot be made. It is inherently difficult to obtain precise estimates of the gradient whenever forecasts have small variance, because the standard error of the gradient is inversely proportional to the spread of the forecasts (Snedecor & Cochran 1980). The lack of independence of the 7598 observations also contributed to the large standard error of the gradient.

During the three-year period (12 mutually exclusive forecasting periods) commencing in June 2000 the gridded above-median seasonal rainfall forecasts of Australia were shown to be reliable (Fawcett et al. 2005). Accurate estimates of the reliability of a forecasting system will not be obtained from 12 mutually exclusive forecasting periods. During the same period we estimated that the gradient of the linear regression between overall outcome and forecast probability was 0.59 for the lower tercile township forecasts. Although this is greater than the overall gradient of 0.42, it is still well short of perfect reliability.

This analysis was conducted assuming a linear relationship between outcome and forecast probability. Since the relationship between observed and forecast in a perfect forecasting system is linear with a gradient of one and an intercept of zero and the relationship in a forecasting system with no resolution is also linear with a gradient of zero and an intercept of the climatological base, a linear model has good utility for hypothesis testing. A logistic regression may also be considered appropriate for the analysis because of the binary nature of the outcome and, unlike linear regression, logistic regression constrains the outcome to be between zero and one (Hosmer & Lemeshow 2000). Exploratory analyses using logistic regression showed that there was generally good agreement between the predicted values from the linear model and the logistic model. Furthermore, in the linear model all of the predicted values of the outcome for any given forecast were within the domain of zero to one. The minimum predicted value from any of the 14 linear regressions was 0.08 and the maximum was 0.73.

4.3. Value of forecasts

The traditional measures of forecast skill, such as Linear Error in Probability Space and ROC values, do not provide direct information about the value of a forecasting system to users. In contrast, value score curves model the scaled economic value of a forecasting system for all economically rational decisions that users may face (Wilks 2001).

When the observed overall forecast distribution and the observed overall reliability were used to calculate value scores (Scenario 1), estimates ranged between −0.0254 and 0.0428, with the great majority of value scores approximating zero (Figure 4). The integral of the value score curve across the entire rational range of cost/loss ratios was also close to zero (–0.00285). The small negative value scores resulted from the lack of perfect reliability and indicate that some users would have been better off ignoring the forecast and using climatology.

The value score curve derived from Scenario 2, in which the overall forecast distribution was fitted to a beta distribution and the observed linear relationship between outcome and forecast was used to determine value scores, almost exactly matched the value score curve from Scenario 1 (Figure 4). This indicates that the beta distribution was an adequate descriptor of the forecast distribution. Beta distributions have often been used to model forecast distributions, both in long- and short-term forecasting (Wilks 2001).

Biased forecasts can be recalibrated to remove the bias. Scenario 3 provides the value scores from a recalibrated forecast. Since the original forecasts were overconfident, the recalibrated forecast distribution has a smaller variance than the observed distribution. The value scores in this scenario are never negative, but are highly constrained by the very small variance (0.00085) of the recalibrated forecasts (Figure 5).

The small value scores were not caused by the observed bias. In Scenario 4, in which it was assumed that the

forecasting system had perfect reliability but retained the observed overall variance, value scores remained low (Figure 5). Consequently, although estimates of reliability are imprecise, we conclude that even in the absence of bias, the forecasting system could have delivered little value to users.

An upper limit of the value scores associated with the forecasting system was estimated by assuming that the system had perfect reliability and by using the maximum observed regional or seasonal variance (Scenario 5). Even in this best-case scenario, although value scores corresponding to decisions with a cost/loss ratio of about 0.33 were moderate, all other value scores remained small or zero (Figure 6). A more conservative estimate of the value scores associated with the forecasting system was obtained by using the minimum observed regional or seasonal variance and using the observed overall relationship between outcome and forecast (Scenario 6). In this scenario, value scores were consistently close to zero (Figure 6).

Similar results were obtained by Gagnon & Verret (2002) in their study of the seasonal rainfall forecasts issued by the Canadian Meteorological Centre. They estimated the maximum value score associated with the Canadian seasonal rainfall forecasts as 0.09, comparable to the estimates of maximal value scores derived in this study. Their conclusion was that there was little or no value associated with seasonal precipitation anomaly forecasts in Canada.

Value score curves illustrate that the value derived from using a forecasting system is dependent upon the decision threshold selected by the forecast user. Decisions that are only triggered when a large deviation of the forecast from climatology occurs are associated with lower value scores than decisions that are triggered by a small deviation of the forecast from climatology. In Table 4 the mean value score for three decision threshold ranges are provided:

(1) decisions that are triggered by a small shift in the forecast from climatology (forecast probability between 0.26 and 0.45);
(2) decisions triggered by a moderate shift of the forecast from climatology (forecasts between 0.15 and 0.25 and forecasts between 0.46 and 0.66) and
(3) decisions that are only triggered by a large shift in the forecast from climatology (less than 0.15 or greater than 0.66).

The mean value scores derived using Scenarios 1 to 6 indicate that no value could have been derived if the forecasts were being used for decisions requiring a large shift in forecast probability (Table 4). This is because it is extremely unlikely that such decision thresholds will be triggered by the observed seasonal forecasting system and consequently the user will always fail to take protection. Similarly, the mean value score for decisions triggered by a moderate shift in forecast probability also approximated zero under all six scenarios. Once again, this is because it is highly unlikely that such decision thresholds will be triggered by the observed forecasting system. The uniformity of value scores from all scenarios indicates that these are robust results.

There are numerous methods other than the cost/loss model that users of seasonal forecasts may employ to choose a decision threshold and many divergent user views of utility, but these findings can be broadened to all such cases. Put simply, if a decision threshold of greater than about 0.51, or less than about 0.19, is chosen by any method whatsoever, then the expected utility of the observed forecasting system will approximate zero for all such users, regardless of their view on utility. This is because it is highly unlikely (probability less than 0.01) that such decision thresholds will be triggered by the observed forecasting system and consequently the user will invariably fail to take protection.

Small positive mean value scores were observed for decisions triggered by a small shift in the forecast probability from climatology. The small value scores indicate that the expected value accruing from the observed forecast is highly eroded when compared to the expected value that would accrue from perfect information. This is because at decision thresholds close to climatology, users will fail to take protection in a high proportion of 'dry' events and additionally on those occasions when protection is taken, a large proportion of the outcomes will not be 'dry' and the user will have unnecessarily incurred the cost of protection. A practical example that illustrates the lack of value for decisions that have a threshold close to climatology is provided below (section 4.3.1).

The value score curve is based on a simple cost/loss ratio model. Despite its simplicity, the cost/loss model does approximate many real-world problems to a reasonable degree (Roebber & Bosart 1996; Wilks 1997). Our experience with the types of decisions faced by Australian farmers when considering the risk of dry seasonal conditions is that the cost/loss model is appropriate. Changing supplementary feed purchasing policies, altering fertiliser applications or considering sale of animals are all well described by the cost/loss model. In Australian agriculture it is usual to describe these decisions using decision trees and expected monetary values (Vizard 1994), but mathematically these descriptions are identical to the cost/loss model.

It has been assumed in most meteorological applications of the cost/loss model that the costs and losses must be measured in monetary units and consequently the cost/loss model does not consider an individual's risk tolerance (Rollins & Shaykewich 2003). However, units other than money can be used within the cost/loss model to estimate the utility of a forecasting system for any end

user. For example, costs and losses can be measured in units of utility that reflect the subjective value that a decision-maker may place on risk and other intangible factors (Friedman & Savage 1948). Since the value score curve is showing the relationship between relative value and the cost/loss ratio, regardless of the units that costs and losses are measured in, the value score curve remains unchanged. This indicates that the value score curve is a robust model capable of incorporating an individual's risk tolerance.

Another limitation of the cost/loss model in its simplest form is that it is constrained to dichotomous decisions. However, sequential decisions and graduated responses can be deconstructed to a series of dichotomous events that can be analysed using the cost/loss model.

The cost of protecting against adverse weather, like any expenditure, is an investment upon which the decision-maker expects a return. If the decision criterion is to maximise expected monetary value, the return on the investment is the increase in revenue or reduction in loss when adverse weather occurs but protection has been sought. The net benefit of the protection will be the increased revenue (L) minus the costs (C). The rate of return on investment in protecting against adverse weather as a percentage (RR%) is given by:

$$\mathrm{RR}\% = \frac{\mathrm{L} - \mathrm{C}}{\mathrm{C}} \times \frac{100}{1} \tag{4}$$

This RR% is a ratio that measures the efficiency with which additional expenditure on protecting against adverse weather avoids expected losses.

The cost/loss ratio can be converted to the RR% and vice versa.

$$RR(\%) = \left(\frac{1}{\left(\frac{C}{L}\right)} - 1 \right) \times 100 \tag{5}$$

The value of the forecast system to a particular user will depend on the distribution of the cost/loss ratios of their decisions. The true distribution of these decisions is not well known, although there are indications that it is unlikely to be uniform (Richardson 2003). Expressed as rate of return on investment rather than the cost/loss ratio enables direct comparisons with the rates of return offered by decisions in the economy.

Decisions with rates of return of about 10% are common but often rejected because of the poor return compared to the opportunity cost of the investment. Decisions with rates of return greater than 10% are intrinsically more interesting but are also less common. Our experience in Australian agricultural industries is that when a dry season is defined as less than the 33rd percentile of seasonal rainfall, most decisions of interest offer rates of return between 10% and 100%. This is equivalent to cost/loss ratios between 0.91 and 0.5. The mean value score across the range of cost/loss ratios between 0.5 and 0.91 may therefore provide a more meaningful measure of the value of the forecasting system to decision makers in Australian agriculture than the integral of the value score curve. In all six scenarios, value scores across this range approximated zero. We therefore conclude that the seasonal rainfall forecasts could not have been used to reliably increase economic value for any decision in this important range.

The cost/loss ratio is usually described in its simplest case, which assumes that if action is taken, the cost of protection is fixed irrespective of the outcome. In many decisions this may not be the case, as often part of the cost of protection, C, may be recoverable if adverse weather fails to occur. In the more general case of costs differing depending on outcome of weather, we note that the value score methodology still applies, with the decision threshold being calculated by:

$$\mathrm{Threshold} = \mathrm{C}_2/(\mathrm{L} + \mathrm{C}_2 - \mathrm{C}_1) \tag{6}$$

in which C_1 is the cost of protecting against the effects of adverse weather when adverse weather occurs, C_2 is the cost of protecting against the effects of adverse weather when adverse weather does not occur and L is the loss that results from the occurrence of adverse weather without the benefit of protection. This reduces to C/L when C_1 equals C_2.

Value score curves examine the value of forecasts as a business production input. Forecasts may also have consumption value to some users. There will always be some users that perceive 'value' from a forecast regardless of its attributes. For example, some users may derive value from finding the forecast interesting, or finding it reassuring that researchers are working on the forecasts. These and other consumption sources of value are not assessed in value score curves.

In summary, we conclude that no economic benefit could have been reliably derived by users of the Bureau's seasonal rainfall forecasts for Australia, between 1997 and 2005, with the exception of users with decisions triggered by a small shift in the forecast from climatology, in which case small economic gains may have occurred.

4.3.1. An example

In Australia, sheep and cattle grazing pastures often require increased supplementary feed during a 'dry' season. Forecasts of a 'dry' season offer farmers the potential to purchase supplementary feed in advance of the 'dry' season and in doing so avoid the price premium that results from the increased demand for grain. For example, during winter, a farmer faced with a potentially 'dry' spring may be able to purchase extra grain immediately for \$175/tonne (C_1). If a 'dry' spring

occurs the farmer expects the price of grain to increase to $220/tonne (L). If the spring is not 'dry' the farmer does not need the extra grain but believes he can trade out of it for a total loss of $30/tonne ($C_2$). Using the decision criteria of maximising expected monetary value (EMV) and substituting the above values into Equation (6), gives a decision threshold of 0.40. Consequently, the farmer will only buy extra grain if a seasonal rainfall forecast of 0.40 or greater is given.

Consider 100 seasons in which the farmer uses this strategy. Thirty-three of the seasons are expected to be dry, with the remainder not dry. If the farmer had access to a perfect rainfall forecasting system, action would be taken in all 33 dry seasons and no action taken in all 67 'not dry' seasons. The EMV of the perfect forecasting system compared to climatology is thus (220–175)*0.33, or $14.85/tonne of supplementary feed/year.

The observed forecasting system, however, is not perfect. Assuming the observed overall forecast variance of 0.0048 and perfect reliability (Scenario 4), the probability of the farmer being spurred into action is 0.19. The farmer will therefore take no action in 81 of the 100 seasons. But 25 of these seasons are expected to be dry (31%) and the farmer will have taken no action to minimise losses. Further, of the 19 occasions when the farmer is convinced by the forecasting system to purchase extra grain, 8 are expected to be dry (42%), in which a $45/tonne benefit is gained, and the other 11 seasons are expected to be not dry, in which a $30/tonne loss is suffered. The estimated benefit of the observed forecasting system to the farmer is thus $0.30/tonne of supplementary feed/year, or 0.02 of the benefit of a perfect forecasting system, which, by definition, equals the value score.

4.4. Expected improvement in value scores associated with increased forecast variance

We examined the expected change in value scores associated with any increase in forecast variance (see Figure 7). In Figure 7 the x-axis is the variance ratio. In the unbiased situation, this also equals the forecast difference and the Brier Skill Score. The relationship between the integral of the value score curve and the variance ratio is exponential. This result can be generalised. Assuming forecasts follow beta distributions, the relationship between the integral of the value score curve and the variance ratio described in Figure 7 holds for all unbiased forecasting systems regardless of the climatological frequency of the adverse event.

Figure 7 illustrates that the first derivative is small at or about the observed variance ratio of 2.1% for the integral, the mean value score for decisions triggered by a large shift in the forecast from climatology and for the mean value score for decisions triggered by a moderate shift in the forecast from climatology. In contrast, the first derivative of the mean value score for decisions triggered by a small shift in the forecast from climatology is large. Consequently, any small or moderate increase from the observed variance is expected to bring limited benefits to users and will be generally restricted to users with decisions that are triggered by a small shift in the forecast from climatology. Uniform and substantial value to users will only occur with a large increase in forecast variance. For example, substantial value to a broad range of users would occur if an unbiased forecasting system captured about 35% of the variance of a perfect forecasting system. Assuming a beta distribution, such a forecasting system would have an integral of the value score curve of 0.27 and mean value scores of 0.49, 0.35 and 0.13 for decisions triggered by small, moderate and large shifts in the forecast from climatology, respectively. To provide this level of value to users requires a seventeen-fold increase in variance of the forecasts from the observed variance.

Any future minor improvement to the seasonal forecasting model that was used during the study period is therefore unlikely to generate significant and widespread benefits to users. To deliver uniform and widespread value to users of the forecasts, new lead indicators with markedly better predictive characteristics may need to be developed.

Acknowledgements

We dedicate this paper to Dr Fred Morley, who initiated this journey. We thank the Hermon Slade Foundation for funding this research.

References

Brier, G. W. (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**: 1–3.

Donner, A. & Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold.

Drosdowsky, W. & Chambers, L. E. (2001) Near-global sea surface temperature anomalies as predictors of Australian seasonal rainfall. *J. Climate* **14**: 1677–1687.

Fawcett, R. J. B., Jones, D. A. & Beard, G. S. (2005) A verification of publicly issued seasonal forecasts issued by the Australian Bureau of Meteorology: 1998–2003. *Aust. Meteorol. Mag*. **54**: 1–13.

Friedman, M. & Savage, L. J. (1948) The utility analysis of choices involving risk. *J. Political Econ*. **56**: 279–304.

Gagnon, N. & Verret, R. (2002) Relative economical value of CMC seasonal forecasts. In *Proc. of the 16th AMS Conference on Probability and Statistics in Atmospheric Sciences*, Orlando, Florida, 26–30.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression*. 2nd edn, New York: Wiley & Sons, Inc.

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

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

Murphy, A. H. & Winkler, R. L. (1992) Diagnostic verification of probability forecasts. *Int. J. Forecasting* **7**: 435–455.

Richardson, D. S. (2003) Economic value and skill. In I. T. Jollife & D. B. Stephenson (eds.), *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Chichester: John Wiley & Sons, 165–187.

Ridout, M. S., Demetrio, C. G. B. & Firth, D. (1999) Estimating intraclass correlation for binary data. *Biometrics* **55**: 137–148.

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.

Rollins, K. S. & Shaykewich, J. (2003) Using willingness-to-pay to assess the economic value of weather forecasts for multiple commercial sectors. *Meteorol. Appl*. **10**: 31–38.

Snedecor, G. W. & Cochran, W. G. (1980) *Statistical Methods*. 7th edn. Ames: Iowa State University Press.

StataCorp. (2005) Stata Statistical Software: Release 9. Stata Corporation, College Station, Texas.

Toth, Z., Talagrand. O., Candille. G. & Zhu, Y. (2003) Probability and ensemble forecasts. In I. T. Jollife & D. B. Stephenson (eds.), *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Chichester: John Wiley & Sons, 137–163.

Vizard, A. L. (1994) Finances and decisions. In F. H. W. Morley (ed.), *Merinos, Money & Management*, Post Graduate Committee in Veterinary Science, University of Sydney, 47–71.

Wilks, D. S. (1995) *Statistical Methods in the Atmospheric Sciences: An Introduction*. San Diego: Academic Press.

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

Wilks, D. S. (2001) A skill score based on economic value for probability forecasts. *Meteorol. Appl*. **8**: 209–219.

Williams, R. L. (2000) A note on robust variance estimation for cluster-correlated data. *Biometrics* **56**: 645–646.