Aust. Met. Mag. 43 (1994) 143–151

Assessments of categorical rainfall predictions

Ian Smith
CSIRO Division of Atmospheric Research, Aspendale, Australia
(Manuscript received September 1993; revised March 1994)

The problems involved in formulating and evaluating categorical rainfall predictions are discussed. Equitable skill scores, based on predicted/observed contingency tables, provide a relatively simple means of assessing performance and effectively distinguish between true and artificial skill. These have been used to assess the seasonal rainfall predictions issued by two agencies over recent years – the National Climate Centre (NCC), 1989–1992, and Austweather Pty Ltd, 1984–1992. Although the NCC predictions performed well in 1991, the predictions have not yet demonstrated any overall significant skill. The Austweather predictions, although confined to Western Australia, exhibit significant overall skill. This is particularly noteworthy since these predictions are issued with lead times up to eight months.

Introduction

Seasonal rainfall predictions for Australia have been available for a number of years now. In 1984 Austweather Pty Ltd, a private consulting firm, began issuing seasonal rainfall outlooks for the wheatbelt of Western Australia, while the National Climate Centre (NCC) of the Bureau of Meteorology has been issuing outlooks on a monthly basis since 1989. The outlooks in both cases are generally in the form of categorical predictions for rainfall districts and refer to three-month periods. The predictions are based on past relationships between rainfall and various climatic indices including the Southern Oscillation Index (SOI), sea-surface temperatures, etc. The strength of these relationships can often be measured in terms of linear regression correlation coefficients, but this measure of skill does not always satisfy many users, such as primary producers who may use the outlooks to assist them in their livestock or pasture management decisions. The basic question they often ask, and which is addressed by this study, is: How successful have the categorical rainfall predictions been for Australia over recent years? The literature is very sparse on this topic and only a report by NCC (1992) and a letter by Lamond (1993) attempt to quantify the skill associated with the published categorical predictions issued by the two agencies. Both studies adopted different and unexplained assessment methods and it would appear that there exists a need for at least one objective assessment method which can be applied to the range of outlooks issued by different agencies, for different regions at different times, and which yields an unambiguous measure of skill which is relatively simple to interpret. There is certainly a need to establish a basis of assessing rainfall predictions if only to monitor the effect of any improvements with time.

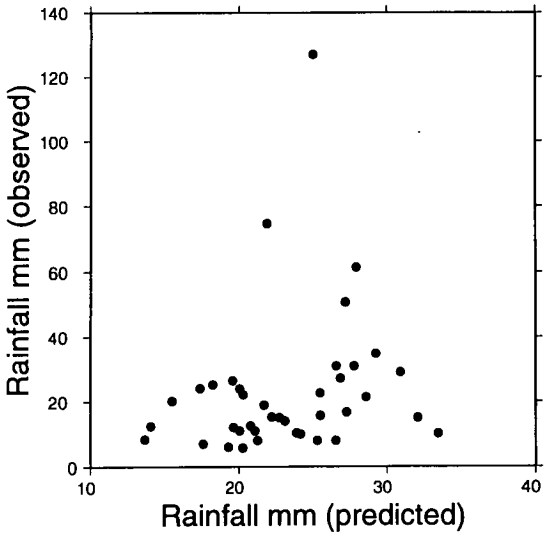
Ward and Folland (1991) discuss a number of methods and difficulties associated with assessing rainfall predictions, while Gandin and Murphy (1992) describe 'equitable skill scores' which effectively distinguish between true and artificial skill and are relatively simple to calculate. This paper describes the application of these scores to the Australian categorical predictions over recent time and accompanies them with estimates of the existence (or lack) of any statistical significance.

Equitable skill scores

In may cases, predictions of rainfall or temperature are based on regression fits between past data and climatic indices. An example of this type of approach is provided by Fig. 1, which shows predictions of winter rainfall amounts (R) for a

Corresponding author address: Dr I. Smith, CSIRO Division of Atmospheric Research, Private Bag No.1, Mordialloc, Vic 3195, Australia.

Fig. 1 An example of regression-based predictions of winter rainfall totals of a district in north Queensland versus observed totals.



district in north Queensland compared to observations. The predictions are generated as follows

$$R = a*SST + b$$

where the coefficients a and b are based on a linear regression fit between rainfall and an index of sea-surface temperatures (SST). This figure illustrates a well-known difficulty with regression-based predictions. It can be seen that the distributions of the predictions and observations are very different and this contributes to a relatively low correlation of 0.21. The predictions are biased away from extreme events and tend to be clustered about the mean (the predicted totals lie within the range 13 mm to 34 mm while the observations lie within the range 1 mm to 130 mm). One method of overcoming this problem is to 'inflate' the predictions (i.e. scale the predictions in order that they have the same standard deviation as the observed data (Ward and Folland 1991)) but this does not improve the poor fit to the data. If the correlation coefficient is used as a measure, these predictions would not be described as skilful.

A method which avoids these problems is to consider the success at predicting rainfall categories rather than raw rainfall totals. This can be done by ranking the predicted values and the observations independently so that, in the long term, the predicted categories have the same distribution as the observed categories – an important prerequisite for defining skill levels (Livezey 1987). Categories adopted by the Bureau of Meteorology are based on decile ranges and comprise: below average – deciles 1 to 3, average – deciles 4 to 7, and above average – deciles 8 to 10. Figure 2 compares the predictions and observations presented in Fig. 1 in terms of these categories. Table 1, a contingency table, quantifies the correct and incorrect categorical predictions. Note that the predictions and observations have the same distribution (i.e. approximately 30%, 40%, 30%). Despite the low correlation between the raw values, the categorical predictions now appear relatively skilful since 20 out of the total of 39 (or 51%) are correct whereas a series of random selections would be expected to yield, on average, only 13 (or 34%) correct.

Fig. 2 The same comparison as in Fig. 1 except that the data have been ranked into percentile ranges.

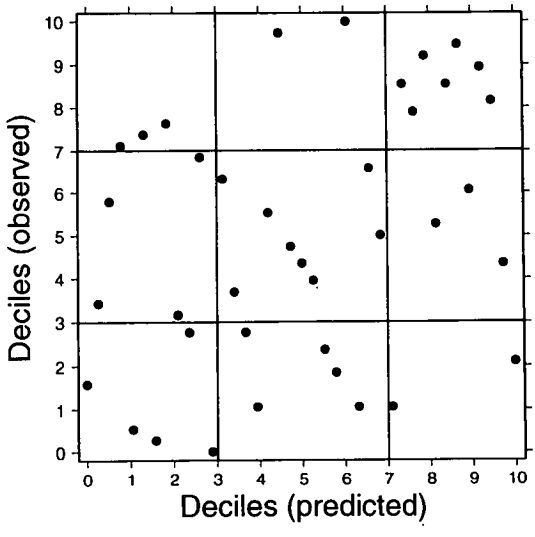


Table 1. Categorical rainfall predictions for North Queensland. Categories: below average, deciles 1–3 (BA), average, deciles 4–7 (AV) and above average, deciles 8–10 (AA).

	Predicted			*Total*
	BA	*AV*	*AA*	
BA	5	5	2	12
Observed AV	4	8	3	15
AA	3	2	7	12
Total	12	15	12	39

Percentage 'hits' = 51.3%
Skill score (using Table 2, see text) = 25.5%

The percentage of correct 'hits' provides a relatively simple measure of the effectiveness of the predictions but it does not discriminate between the different types of correct predictions, nor does it discriminate between the different types of incorrect predictions. The conservative strategy which relies on average-only conditions for every prediction will, in the long term, score 40%. This is greater than chance, but is obviously not skilful and arises because of the choice of categories. The consequence of this is 'artificial skill'. As already noted, raw linear regression-based predictions will tend to yield values clustered about the observed long-term mean value. If they are not ranked independently of the observations, they will appear to represent a cluster of average-only predictions and end up with a percentage 'hits' score of about 40% – better than chance but obviously without skill.

This problem can be avoided by formulating a skill score which takes into account the incorrect predictions. Gandin and Murphy (1992) present a method of formulating what are described as equitable skill scores which discriminate between true and artificial skill. Table 2 shows one of many possible scoring matrices associated with the three decile ranges used here. This particular matrix is constructed on the basis that if a correct prediction of average conditions is given a weighting of one, then a correct prediction of either below or above average conditions (assuming it to be twice as valuable) is given a weighting of two. Assuming they are symmetric, the penalties for the incorrect predictions are easily solved for. The net result is such that a random selection of categories will, in the long term, yield a score of zero. This is also true of any of the simple strategies which adopt either climatology, below average (the 'pessimistic' approach) or above average (the 'optimistic' approach) conditions each time. Note that there are an infinite number of possible scoring combinations with these desirable properties and the reader is referred to Gandin and Murphy (1992) for a more general discussion. Scaling (in this case by 62.5/N) of the final score is such that, in the long term, perfect predictions score 100%. The resultant score therefore provides a more satisfactory measure than does the percentage 'hits' score. In the example provided by Fig. 2, the skill score is 27.4% and reflects the fact that only five of the 39 predictions were seriously in error (i.e. predicted above average when observed was below average and vice-versa).

Finally, the significance of the above score needs to be considered since the relatively high skill score could have been achieved simply by chance. In this example, the sample comprises 39 effectively independent predictions/observations. The question as to whether the skill score is statistically significant can be addressed by considering the distribution of skill scores associated with sets of observations and predictions, each comprising 39 randomly generated numbers. Ten thousand such series of pseudo-observations and predictions were generated and skill scores calculated in each case. The results revealed that less than 2% of the total achieved skill scores exceeding 25% – indicating that the skill score of 27.4% is significant at close to the 99% level. This method is used to estimate significance levels in all the cases which follow.

Table 2. Equitable scoring matrix.

		Predicted	
	BA	*AV*	*AA*
BA	2.00	−0.67	−1.11
Observed AV	−0.67	1.00	−0.67
AA	−1.11	−0.67	2.00

Scaling factor = 62.5/N (%)

NCC seasonal outlooks 1989–1992

Interannual fluctuations in seasonal rainfall over large parts of Australia have been linked to changes in the Southern Oscillation Index (SOI) and sea-surface temperatures (Nicholls 1989) and these form the basis of seasonal rainfall outlooks issued by the NCC for regions of Australia. Figure 3 shows the behaviour of the monthly SOI from July 1989 to January 1993. Between late 1989 and the end of 1990 the SOI varied between positive and negative values characteristic of non-ENSO years. In 1991 it entered a negative phase in conjunction with warm sea-surface temperature anomalies in the eastern equatorial Pacific. In August-September of 1992 there was some evidence of a breakdown in the warm event, but the SOI reverted to negative values in October which persisted well into 1993.

In June 1989 the National Climate Centre of the Bureau of Meteorology began issuing a series of seasonal rainfall outlooks which included predicted rainfall deciles for selected Australian districts (National Climate Centre 1989). The districts selected were those where a statistically significant linear regression relationship was found to exist between rainfall and pre-season SOI. The number of districts in each seasonal period are shown in Fig. 4(a) and differ according to the time of year but peaked during September-November when the SOI-rainfall relationship is known to be strongest. The numbers also differ from year to year because different cut-offs were adopted by the NCC when distinguishing between significant and non-significant regression rela-

tionships. Between 1989 and 1991 no predictions were issued for the period January to May, since the relationships are generally weakest at this time of year. The periods for which categorical predictions were issued comprise the following groups of months:

1989 - JJ (i.e. June-July), JAS, ASO, SON and OND.
1990 - JJ, JAS, SON, NDJ (i.e. November, December 1990 and January 1991) and DJF.
1991 - JJ, JAS, ASO, SON, OND, NDJ and DJF

Fig. 3 Monthly SOI (July 1989 to January 1993).

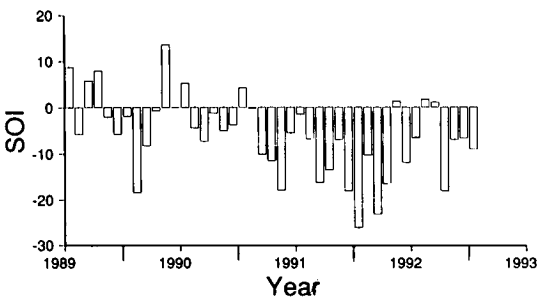
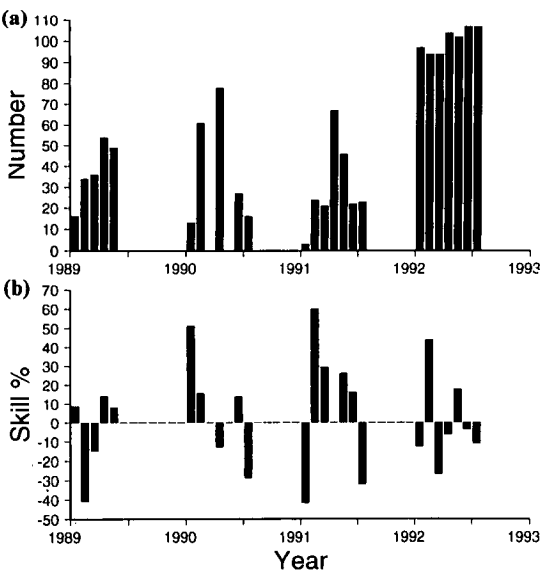


Fig. 4 NCC categorical seasonal predictions: (a) number per season; and (b) associated skill scores for the period June–July 1989 to December 1992–February 1993.



In 1992 the seasonal outlooks were issued for the first time in the form of maps delineating regions where the probabilities of below average and above average rainfall were predicted to be greater than the 30% expected from climatology. Initially only two maps were issued – each indicating just those regions where the predicted probabilities exceeded 40% for each category. These are now issued throughout the year and provide a prediction for most districts throughout Australia. The exceptions are the monsoon regions of northern Australia which are seasonally dry during winter. The predicted probabilities are still mainly based on the distribution of past rainfall amounts as a function of SOI, but this method tends to avoid the conservative nature of linear regression methods discussed previously. Furthermore, as Fig. 4(a) indicates, a far greater number of districts are now included in the outlooks. However, there are many districts where there is no appreciable rainfall-SOI relationship and in these cases the probabilities tend to reflect climatology (i.e. 30%, 40%, 30%). The periods for which predictions were issued in 1992 and which are used in this assessment comprised the following groups of months: JJA, JAS, ASO, SON, OND, NDJ and DJF.

The predictions for 1989 through 1991 can easily be assessed since they were issued as categorical predictions. However, the assessment of predictions during 1992 is slightly more difficult. Although the individual categorical probabilities are available from the NCC, only the maps as described were published. If a map indicated that the probability of below (or above) average rainfall for a particular district exceeded the climatological expectation of 30% – by at least 10% – then the approximation is made that the prediction was a below (or above) average categorical prediction (this approximation leads to skill scores which do not differ greatly from those based on the actual probabilities, see Appendix 1). A slight difficulty occurs when dealing with predictions for average conditions, since no maps were shown where the predicted probability for this category exceeded the climatological expectation of 40%. This means that districts where the rainfall-SOI relationship indicates a strong likelihood of average conditions are not easily distinguished from districts where no relationship exists and climatology was assumed. This situation was eventually remedied with the outlook issued in December 1992, in which predictions of average-only rainfall were distinguished from predictions that were no more than statements of climatology.

Two assessment methods are adopted here. The first method deals with all categorical predictions (1989–1992) whereas the second method deals with just the extreme (below and above average) categories. The second method avoids both the

conservative bias in the regression-based predictions for 1989–1991 and the lack of distinction between average-only predictions and statements of climatology in 1992.

Assessment method 1 – all predictions

The first set of predictions were for the June–July 1989 period and involved a total of 16 districts. Table 3 is the contingency table for this set and shows that the percentage 'hits' score is only 31.3% – less than the 34% expected in the long term by a random selection strategy. On the other hand, the weighted skill score (using Table 2) is +8.6% – indicating some skill and reflecting the fact that only one of the 19 predictions was seriously in error. Assuming that the 26 predictions and observations are independent, it is estimated that this score is only significant at the 73% level. It can be easily seen that neither the predictions nor observations are completely independent and so the effective sample size is less than 16. Consequently little statistical significance can be associated with this score.

Determining the effective sample size is made difficult when assessing sets comprising predictions and observations that are correlated in space and time. Hereafter, where positive skill scores are quoted they are accompanied (in brackets) by two estimates of significance levels based on trials with random number series. The first level is estimated by assuming the data are independent while the second is estimated by assuming that dependence within the data reduces the effective sample size by half. The first level will almost always represent an overestimate while the second value, although likely to be more realistic in the majority of cases, may still represent an overestimate. This may be the case when the data refers to a large number of small districts with similar rainfall patterns or the predictions, which sometimes refer to overlapping three-month periods, are strongly correlated in time. This limitation needs to be kept in mind but in most cases these two estimates are sufficient to judge the significance of the skill scores.

Table 3. NCC predictions for June–July 1989.

	Predicted			Total
	BA	*AV*	*AA*	
BA	0	0	1	1
Observed AV	0	0	9	9
AA	0	1	5	6
Total	0	1	15	16

Percentage 'hits' = 31.3%
Weighted skill score = 8.6%

The performance of the predictions for each period for the entire four years is shown in Fig. 4(b). The skill scores vary considerably and, except for two seasons, were consistently positive throughout 1991. Tables 4 to 7 show the contingency tables that apply to all predictions in each of these years. A feature of the annual skill scores is the fact that the percentage 'hits' score consistently exceeds 34%, but only during 1991 is the skill score positive and apparently significant – +10.8% (97%, 92%). Note that in 1989 and 1990 not one prediction was for below average rainfall, which contrasts with 1991, when not one prediction was for above average rainfall. During 1992 the probabilities yielded predictions for both below and above average rainfall but the implied predictions of average conditions can be seen to have dominated. As discussed above, this occurs because nearly all districts are included in the assessment, including those where climatology applies by default.

Table 4. NCC predictions for 1989.

	Predicted			Total
	BA	*AV*	*AA*	
BA	0	36	32	68
Observed AV	0	51	42	93
AA	0	11	17	28
Total	0	98	91	189

Percentage 'hits' = 36.0%
Skill score = −3.4%

Table 5. NCC predictions for 1990.

	Predicted			Total
	BA	*AV*	*AA*	
BA	0	75	0	75
Observed AV	0	65	5	70
AA	0	42	8	50
Total	0	182	13	195

Percentage 'hits' = 37.4%
Skill score = −0.7%

Table 6. NCC predictions for 1991.

	Predicted			Total
	BA	*AV*	*AA*	
BA	36	61	0	97
Observed AV	39	50	0	89
AA	14	6	0	20
Total	89	117	0	206

Percentage 'hits' = 41.7%
Skill score = +10.8%

Table 7. NCC predictions for 1992.

	Predicted			
	BA	*AV*	*AA*	*Total*
BA	5	61	26	92
Observed AV	30	173	64	267
AA	56	213	77	346
Total	91	447	167	705

Percentage 'hits' = 36.2%
Skill score = −0.1%

Figure 5 shows the annual skill scores compared to the skill scores associated with persistence (using the previous season's observed category) and the conservative strategy of adopting climatology. It can be seen that the NCC predictions performed best during 1991 but were outperformed by the simpler predictive strategies in the remaining years. Persistence can be seen to have performed well during 1989 (score = +15.9% (>99%, 98%)) and 1992 (score = +21.8% (>99%, >99%)) but poorly during 1990 and 1991. Climatology performed moderately well in 1989 (score = +9.3% (97%, 93%)) but was of little use otherwise.

The overall skill score based on a total of 1295 predictions for the four years is +1.3% (77%, 69%) compared to +8.8% (>99%, >99%) for persistence and −0.1% for climatology.

Assessment method 2 – extreme predictions only

If we ignore average-only predictions then equitable skill scores can be calculated using the scoring matrix show in Table 8. This is simply derived from the original matrix (Table 2) by setting the weights for average-only predictions to zero and modifying the scaling factor to ensure a perfect set of predictions would score 100%. Again, a random selection will, in the long term, yield a score of zero and this is also true if the strategy adopted is either 'pessimistic' or 'optimistic'.

Figure 6 compares the performance of the NCC predictions to persistence. The number of predictions differ in each case depending on the number of extreme predictions in the original set. Although the extreme predictions performed poorly in 1989, they did much better in 1990 (score = +48.7%, for only eight cases and therefore not regarded as highly significant) and 1991 (score = +17.0% (>99%, 95%)) but only moderately well in 1992 (score = +1.9% (insignificant)). Persistence showed little skill in 1989 (score = +2.7% (insignificant)), performed poorly in 1990 and 1991, but very well in 1992 (score = +32.2% (>99%, >99%)).

Fig. 5 NCC categorical predictions: (a) number per year; and (b) annual skill scores 1989–1992. Dark bars correspond to the predictions; shaded bars correspond to persistence-based predictions; and enclosed bars correspond to average-only predictions.

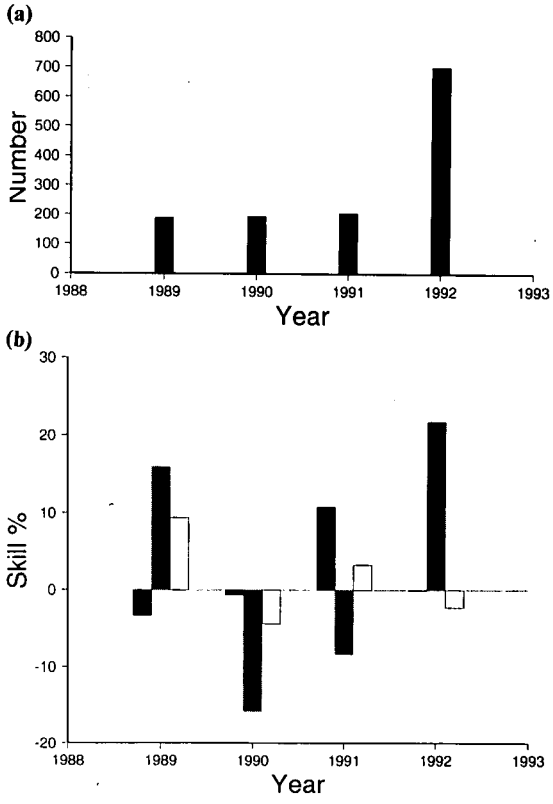


Table 8. Equitable scoring matrix for extreme predictions.

	Predicted		
	BA	*AV*	*AA*
BA	2.00	0.00	−1.11
Observed AV	−0.67	0.00	−0.67
AA	−1.11	0.00	2.00

Scaling factor = 50/N (%)

The overall skill score based on all 451 extreme predictions is +2.6% but the significance levels are not high (76%, 73%). Persistence yielded a total of 762 predictions and overall skill score of +10.9% (>99%, >99%).

Fig. 6 NCC extreme categorical predictions: (a) number per year; and (b) annual skill scores 1989–1992. Dark bars correspond to official predictions; shaded bars to persistence-based predictions.

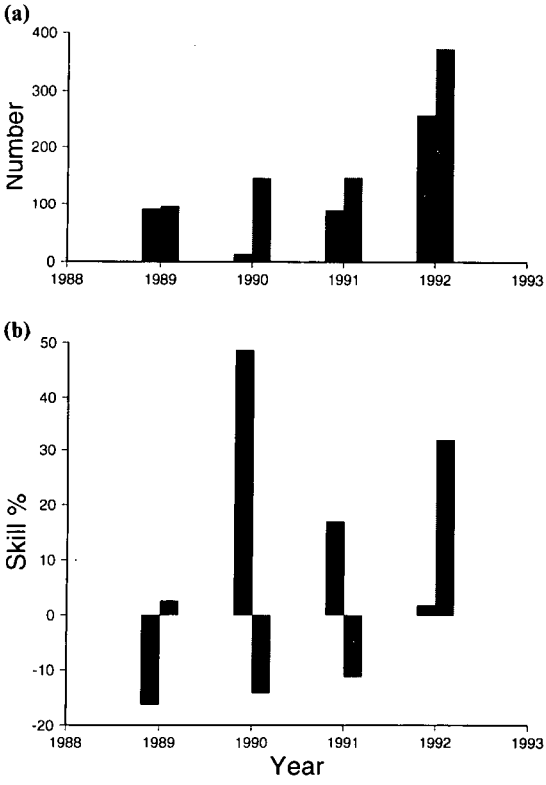
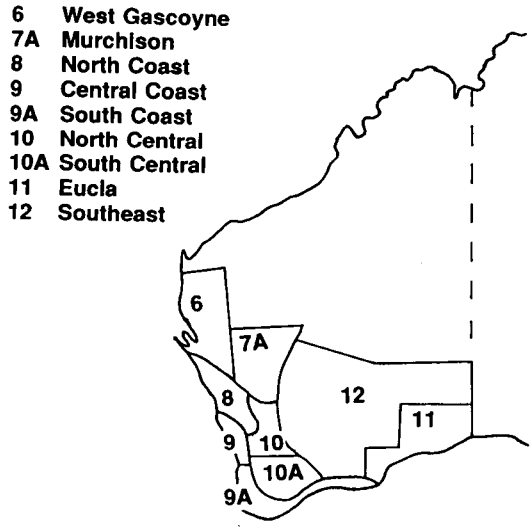


Fig. 7 Districts where Austweather predictions were issued 1984–1992.



Austweather seasonal outlooks 1984–1992

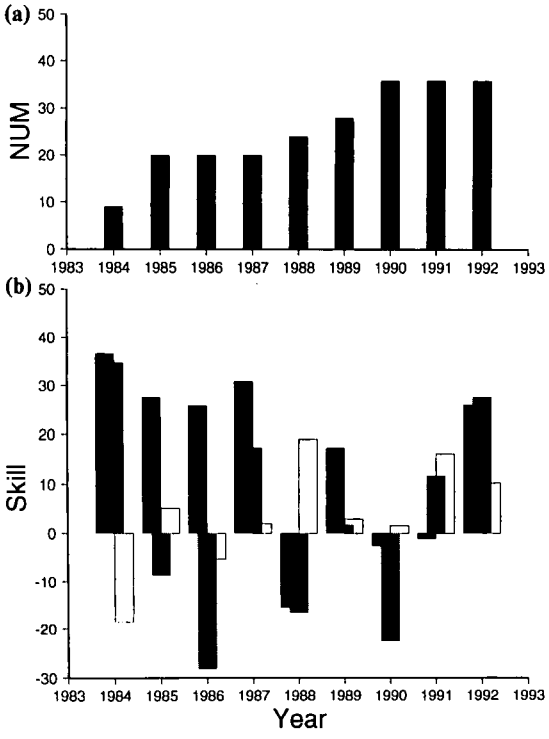
The Austweather outlooks are issued regularly near the beginning of each year and refer to the four groups of months: JFM, AMJ, JAS and OND. Updates are also issued throughout the year but these are not assessed in this study. The original outlooks covered only three districts in Western Australia (North Central, South Central and North Coastal) but by 1990, outlooks were being issued for a further six districts (West Gascoyne, Murchison, Central Coast, South Coast, Eucla and South East), see Fig. 7. In later years the outlooks included districts further east but these are not assessed here.

The original predictions were mainly descriptive and referred to broad categories such as 'near to or above normal' but by 1986 referred to five precise decile ranges (1 to 2, 3, 4 to 6, 7, 8 to 10) in order to cater for client perceptions. In the case of the South Coast district, predictions were issued for separate sub-regions and these have been combined into single predictions without any great loss of information.

In most cases the method of assessment has been to simply map both the descriptive and the precise outlooks to the three conventional decile ranges for comparison with observations. Where some of the early outlooks are not precise (e.g. 'near to or above normal') they are treated as categorical probabilities (i.e. 50% in each category mentioned, see the appendix). Note that this method does not do justice to the precision of the more recent outlooks, but is adopted mainly for ease of application and consistency over all nine years. Other methods are available which are better able to take account of the precision of these types of predictions. For example, a 'difference from correct' scheme (D. Stephens, private communication) was used in the assessments reported by Lamond (1993).

The results have been aggregated by constructing contingency tables based on predictions for all districts for all four periods during each year. These annual skill scores (and total number of predictions) are shown in Fig. 8 compared to the equivalent scores for persistence and climatology. The predictions performed reasonably well in six out of the nine years: 1984 (score = 36.6% but for only nine predictions and so the significance

Fig. 8 Austweather categorical predictions for the nine-year period 1984–1992; legend as for Fig. 5.



levels (92%, 85%) are only moderate), 1985 (score = 27.5% (94%, 87%)), 1986 (score = 25.8% (92%, 84%)), 1987 (score = 30.8% (96%, 90%)), 1989 (score = 17.2% (88%, 79%)) and 1992 (score = 26.0% (97%, 91%)). Overall, the total number of predictions was 229 and the associated skill score was as 12.2% (99%, 94%). Note that because nine years worth of predictions have been assessed, the data are characterised by a greater degree of independence compared to the NCC set which encompass only four years. As a result, the effective sample size is more likely to be greater than 115 (N/2) in which case the actual significance level is estimated to be greater than 94%. By way of comparison, both persistence (score = 0.8%, insignificant) and climatology (score = 6.0% (88%, 80%)) exhibited far less skill.

There does not appear to be a problem with the distribution of the Austweather predictions (i.e. there is no bias towards an excessive number of average-only predictions) and this is confirmed by the overall skill score for extreme-only predictions (11.7%) which is effectively unchanged.

Summary and discussion

Equitable skill scores provide a relatively simple method for assessing predictions of categories and probabilities. They effectively discriminate between true and artificial skill such that simple strategies including climatology, optimism (always above average), pessimism (always below average) and random selections will, in the long term, score zero. One particular scoring matrix has been used to assess the seasonal rainfall predictions issued by two agencies over recent years: NCC and Austweather.

Although the NCC predictions performed well in 1991, the predictions have not yet demonstrated any overall significant skill. One reason for this may be the use of regression-based predictions for the first few years and the interpretation of probability maps for 1992 which tend to yield an abundance of average-only predictions. However, when these are ignored the extreme-only predictions show little evidence of improvement. Another problem is that four years may not be sufficient to reveal any skill associated with the SOI-rainfall relationships observed over the past century. During 1991, at the start of an El Niño episode, the outlooks performed very well but the following 1992–93 period was unusual. During this time the SOI partly recovered and then reverted to a negative phase (see Fig. 3). This behaviour is almost unprecedented and corresponds to the persistence of an unusually long warm event. As a result, the historical record provided relatively poor guidance on this occasion whereas persistence can be seen to have preformed well. This may only be a short-term factor but, in the longer term, there remains a possibility that SOI-rainfall relationships can break down – as they apparently did between 1920 and 1940 (Allan 1993). An assessment in the future, after several El Niño and La Niña episodes, may therefore reveal greater skill than found here. It is also expected that climatic indices other than just the SOI (e.g. sea-surface temperature indices (Lough 1992; Smith 1993; Drosdowsky 1993) or even sea-level changes (Allan et al. 1990)) may form the basis of future outlooks.

It has often been observed that forecasts of 'near-normal' conditions tend to contribute negligibly to overall skill levels. Van den Dool and Toth (1991) have attempted to explain this phenomenon with a number of examples but utilise a skill score which effectively represents a percentage 'hits' score and does not take into account (as they acknowledge) the off-diagonal elements of the contingency matrix as equitable skill scores do. Therefore any apparent lack of skill associated with the average category revealed by this study is not related to the definition of the skill score. It is more likely related to the fact that linear regression relationships are often fitted to data distributed almost randomly about the means, yet show some signs of linearity at the extremes. Predictions in the form of probabilities, as issued by NCC in 1992, should overcome this problem.

The Austweather predictions show definite evi-

dence of skill over nine years since the overall skill score (+12.2%) is estimated to be significant at better than the 94% level. This is noteworthy since the predictions assessed here were issued at the beginning of each year and the lead time for the OND outlooks was almost eight months. This relatively high level of skill was achieved despite relatively poor performances in three of the nine years.

It should be stressed that the approach adopted in this study is one of many that could be used. We have not considered the *value* associated with skilful predictions (see, for example, Hammer et al. (1991), who describe how to assess the economic value of seasonal predictions). Nevertheless, the results presented here are believed to provide a reliable indication of performance which could serve as a reference against which future improvements may be judged.

Acknowledgments

The author would like to thank Mal Lamond and David Stephens for providing the Austweather outlooks (these are available from the author) and comments on the assessments; Grant Beard, Willem Bouma, Peter Whetton and an anonymous referee for suggestions on improving the manuscript.

Appendix

Assessing categorical probabilities

An example of the unpublished categorical probabilities calculated by the NCC during 1992 is: 20% below average, 20% average and 60% above average. This can be treated as an above average prediction since the 60% exceeds the expected probability (of 30%) by at least 10%. However, this is an approximation which disregards the information inherent in the probabilities for the other two categories. A more accurate assessment can be made by applying the scoring matrix to either of the following contingency tables:

If observed below average

$$\begin{bmatrix} 0.2 & 0.2 & 0.6 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Score = −0.4 (−25%)

If observed average

$$\begin{bmatrix} 0 & 0 & 0 \\ 0.2 & 0.2 & 0.6 \\ 0 & 0 & 0 \end{bmatrix}$$

Score = −0.34 (−21%)

If observed above average

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0.2 & 0.2 & 0.6 \end{bmatrix}$$

Score = +0.78 (+49%)

If the probabilities are approximated as an above-average prediction, the corresponding scores are −60%, −42% and +125%, and it can be seen that the categorical probabilities are penalised less for failures but, at the same time, rewarded less for successes. With a large enough sample, these tend to cancel out so that there is no significant difference between the methods used. When the individual NCC predictions for 1992 are assessed using both methods, relatively large (i.e. up to 5%) differences are found between the seasonal scores but the difference in the annual skill scores is less than 1%. If available, the probabilities should be assessed in the manner described above. Otherwise, the approximate method can be used since the results are similar given a sufficiently large enough sample.

An example of an early prediction issued by Austweather is: 'near to or above normal'. The scores associated with these types of predictions were calculated by assigning equal probabilities of 50% to each of the categories mentioned and adopting the precise method as described above.

References

Allan, R.J. 1993. Fluctations in ENSO and teleconnection structure. *Fourth Internatioanl Conference on Southern Hemisphere Meteorology and Oceanography*, American Meteorological Society, Boston, 185–6.

Allan, R.J., Beck, K. and Mitchell, W.M. 1990. Sea level and rainfall correlations in Australia: Tropical links. *Jnl climate, 3*, 838–46.

Drosdowsky, W. 1993. Potential predictability of winter rainfall over southern and eastern Australia using Indian Ocean sea-surface temperature anomalies. *Aust. Met. Mag., 42*, 1–6.

Gandin, L.S. and Murphy, A.H. 1992. Equitable skill scores for categorical forecasts. *Mon. Weath. Rev., 120*, 361–70.

Hammer, G.L., McKeon, G.M., Clewett, J.F. and Woodruff, D.R. 1991. Usefulness of seasonal climate forecasts in crop and pasture management. *Bull. Aust. Met. Ocean. Soc., 4(6)*, 104–109.

Lamond, M.H. 1993. Letter to the editor. *Bull. Aust. Met. Ocean. Soc., 6(3)*, p. 43.

Livezey, R.E. 1987. The evaluation of skill in climate predictions. In: *Toward Understanding Climate Change* (ed U. Radok), Westview Press, 200 pp.

Lough, J.M. 1992. Variations of sea-surface temperatures off north-eastern Australia and associations with rainfall in Queensland: 1956–1987. *Int. J. Climatol., 12*, 765–82.

National Climate Centre 1989 onwards. *Seasonal Climate Outlook* (by season), No. 1 onwards. Bur. Met., Australia.

National Climate Centre 1992. *A review of the National Climate centre seasonal outlook service during the El Niño episode of 1991/92.* Bur. Met., Australia, 7 pp.

Nicholls, N. 1989. Sea surface temperatures and Australian winter rainfall. *Jnl climate, 2*, 965–73.

Smith, I.N. 1994. Indian Ocean sea-surface temperature patterns and Australian winter rainfall. *Int J. Climatol., 14*, 287–305.

Van den Dool, H.M. and Toth, Z. 1991. Why do forecasts for "near normal" often fail? *Weath. forcasting, 4*, 76–85.

Ward, N.M. and Folland, C.K. 1991. Prediction of seasonal rainfall in the north Nordeste of Brazil using eigenvectors of sea-surface temperature. *Int. J. Climatol., 11*, 711–43.