

# A verification of publicly issued seasonal forecasts issued by the Australian Bureau of Meteorology: 1998-2003

R.J.B. Fawcett, D.A. Jones and G.S. Beard

National Climate Centre, Bureau of Meteorology, Australia

(Manuscript received April 2003; revised June 2004)

**This article presents verification results for the current sea-surface temperature based seasonal forecasting system of the Australian Bureau of Meteorology's National Climate Centre. Publicly issued probability forecasts are verified using a range of scoring techniques, including the Linear Error in Probability Space skill-score, correct forecast rates and reliability curves. Of the three seasonal forecast variables (maximum temperature, minimum temperature and rainfall) routinely forecast, the maximum temperature outlooks have been the most successful, while the rainfall forecasts have almost everywhere performed better than climatology. Minimum temperature forecasts have shown mixed success, largely as a result of a sequence of poorly verifying forecasts during 2001 and early 2002, although the most recent forecasts have performed considerably better.**

## Introduction

The Australian Bureau of Meteorology's National Climate Centre (NCC) commenced its operational seasonal forecasting of Australian rainfall in 1989, with the first forecast (outlook) being issued for the two-month period June-July 1989. Initially, the outlooks were for a mixture of two and three-month periods. The operational service fairly quickly stabilised onto partly overlapping three-month periods, which have continued to the present day. Twelve seasonal outlooks are issued each year.

Across the years, a range of statistical forecasting techniques has been employed by the NCC, with the Southern Oscillation Index (SOI) being the principal

predictor for the first 10 years of the service (Bureau of Meteorology 1992; Casey 1995, 1998; Drosowsky and Chambers 1998). This changed in October 1998, with indices of sea-surface temperature (SST) replacing the Southern Oscillation Index (SOI) (Drosowsky and Chambers 1998, 2001). The shift to empirical forecasts based on SSTs reflected the growing availability of comprehensive and near-real-time ocean analyses (e.g. Reynolds and Smith (1994)) and the increasing understanding that the tropical oceans provide the major source of predictability on seasonal time-scales (see, for example, Davis (1976); Simmonds and Rocha (1991); Ward and Folland (1991); Lau and Nath (1994); Rowell (1998)).

A major aspect of the shift to SSTs in 1998 was the inclusion of a predictor that described variations in the eastern Indian Ocean. This move reflected the findings of studies such as Simmonds and Rocha

---

*Corresponding author address:* Dr R.J.B. Fawcett, National Climate Centre, Bureau of Meteorology, GPO Box 1289K, Melbourne, Vic. 3001, Australia.  
Email: r.fawcett@bom.gov.au

(1991) and Drosdowsky (1993) that Indian Ocean SSTs might affect seasonal climate over Australia, independently of the more familiar forcing associated with the El Niño-Southern Oscillation (ENSO) phenomenon. The first forecast issued using SST predictors was for the NDJ 1998/99 period\*. In February 2000, seasonal maximum and minimum temperatures were added to the set of forecast variables, with the first temperature forecast being for MAM 2000.

In choosing whether or not to upgrade its statistical forecasting schemes, NCC uses cross-validated hindcasting as the principal method for the assessment of forecast model skill (Michaelsen 1987). Such an approach is generally accepted in the climate forecasting community, as independent climate forecast realisations accumulate far too slowly to allow a timely and accurate assessment of forecast skill. Hindcast results for the current operational models are made available in a range of ways (Watkins 2002), and form the basis for communication of expected model skill to the general public. The interested reader is referred to the Bureau's web site ([www.bom.gov.au/silo/products/verif/](http://www.bom.gov.au/silo/products/verif/)) for further details. Nevertheless, it is important that the scientific community periodically assess the true forecast performance of climate outlooks to ensure the robustness of the underlying scientific assumptions, and strengths and weaknesses in these systems. This is particularly so, given that they represent the culmination of scientific efforts to understand climate variability and predictability.

This article represents the second published review of operational climate forecasts for Australia. Smith (1994) reviewed the performance of early (largely) categorical rainfall forecasts for Australia issued by the Bureau of Meteorology for 1989 to 1992, finding forecasts to be skillful during the El Niño of 1991, but to be unskilled during the other three years. Such a result is not surprising, given that these early forecasts were based on the SOI. A review of publicly issued seasonal outlooks was undertaken by Mr R. Seaman of the Bureau of Meteorology Research Centre in early 1994, but the results of this review were not published.

With nearly ten years having passed since these early studies, it seems pertinent and desirable to assess whether scientific advances in the knowledge of climate have flowed into improved forecast accuracies for Australia. In addition, with seasonal forecasts expected to be soon drawn from fully dynamical coupled modelling systems, it is important to provide a benchmarking of present-day statistical schemes, which should be exceeded or at least matched before a shift to dynamical forecast models is made.

## Seasonal forecasting systems

In this study we investigate the accuracy of publicly issued seasonal forecasts issued by the NCC for the period NDJ 1998/99 to JJA 2003. Through this period, the underlying statistical model has remained essentially unchanged, though the predictor data used to drive the forecasts has varied, as described below. This statistical model is based on the method of linear discriminant analysis (LDA) described by Wilks (1995). The model is comprehensively described in the literature by Drosdowsky and Chambers (1998, 2001) and Jones (1998). It replaced the idealised conditional climatology/phase model described by Casey (1995, 1998), which was largely based on the SOI.

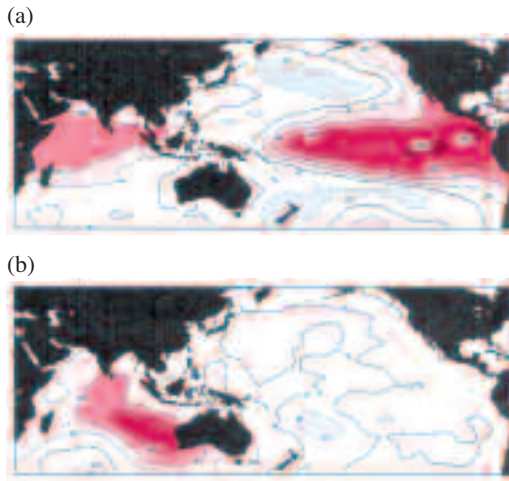
The predictand variables for which forecasts have been publicly issued are seasonal rainfall, maximum and minimum temperatures. A novel characteristic of the empirical model is that the predictand variables are spatially smoothed using truncated empirical orthogonal function (EOF) analysis, as described in Drosdowsky and Chambers (1998, 2001) and Jones (1998). It is these spatially smoothed versions of historical rainfall and temperature data which form the predictand dataset for the forecast model. The forecast model is a probabilistic analogue for principal component regression (Yu et al. 1997). The underlying LDA model is highly modular, and has been applied to the problem of seasonal prediction in Indonesia, Vietnam and Fiji (see, for example, Walsh et al. (2001)).

Commencing with the forecast for NDJ 1998/99, the seasonal forecasts have been made using the two leading modes of Indian/Pacific Ocean SST variability as the predictors in the LDA (described in Drosdowsky and Chambers (2001)). These leading modes were obtained objectively using VARIMAX rotated principal component analysis (Richman 1986), applied to monthly SST data fields for the period 1949-1991. The first component (SST1) describes the canonical spatial structure associated with the peak phase of El Niño/La Niña (Tourre and White 1995), and correlates significantly with the SOI. The second mode (SST2) chiefly describes variability in the eastern Indian Ocean, and is related to the Indian Ocean dipole as described by Drosdowsky (1993), among others.

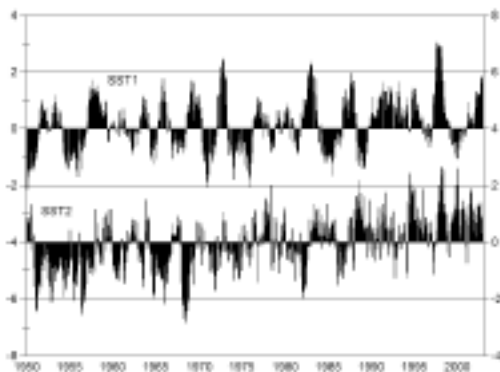
Figure 1(a) shows the EOF pattern for SST1, the tropical Pacific Ocean index, while Fig. 1(b) shows the EOF pattern for SST2, the tropical Indian Ocean index. Darker shades in the figures indicate oceanic areas with greater impact on the final index values, with red shades indicating positive impact and blue shades negative impact. Figure 2 shows time series of the two SST indices from 1950 to 2002. SST1 is typ-

\* Three-month seasons will be indicated throughout this article by JFM (January-February-March), FMA (February-March-April), ..., DJF (December-January-February).

**Fig. 1** EOF patterns for the SST1 (tropical Pacific, top) and SST2 (tropical Indian, bottom) forecast indices.



**Fig. 2** Values of the SST1 (top, left axis) and SST2 (bottom, right axis) forecast indices (1950–2002).



ically positive during El Niño years and typically negative during La Niña years, although the presence of a weak to moderate contribution to SST1 from the tropical Indian Ocean means that the correlation of SST1 with more traditional ENSO indices is not perfect. Even so, the very strong El Niño of 1997 and the protracted warm conditions of the early 1990s are clearly evident. The second index (SST2) shows a strong upward trend in recent decades. This reflects the long-term warming trend in the tropical Indian Ocean which itself is likely to be partly a consequence of the enhanced greenhouse effect (IOCI 2002).

From NDJ 1998/99 to FMA 2000, the predictions were based on a ‘best of four’ arrangement using the two SST predictors at lags of 1 and 3 months, with the possibility of issuing a climatology (‘none of four’) forecast in some seasons and parts of the country as the ‘most skillful’ forecast. With the introduction of seasonal temperature forecasting (MAM 2000), the ‘best of four’ arrangement for seasonal rainfall was abandoned, and all four SST predictors used instead (SST1 and SST2 at lags of 1 and 3 months; Drosowsky and Chambers (2001)). So, for example, forecasts for the summer season (DJF) are based on SST values in the preceding months of August and October, and are generally made available to the public around the middle of November. A major reason for this slight change in the rainfall forecasting procedure was to adopt a consistent approach across all forecast variables\* (the three forecast variables issued publicly – rainfall, maximum and minimum temperature – and other forecast variables investigated experimentally – mean temperature, diurnal temperature range and mean sea-level pressure). For three years now, the operational forecast model has remained unchanged and it is expected that this system will remain the Bureau’s primary operational climate forecast model for the short to medium term. In the longer term, it is expected that this empirical system will be replaced by a fully dynamical system based on the Predictive Ocean Atmosphere Model for Australia (POAMA) being jointly developed by the Bureau of Meteorology and CSIRO Marine Research (Alves et al. 2003).

The forecasts as publicly issued take two forms. The primary form is the probability of exceeding median seasonal rainfall/temperature (sometimes recast as the probability of not exceeding median values), with the secondary form being the probability that the seasonal outcome will fall into each of three terciles (obtained by splitting the climatological distribution for the season under consideration into three equally likely categories). While these two forms can be approached as separate statistical forecasting schemes (and prior to MAM 2000 this was the case), such a method does not necessarily give statistical consistency in all possible forecasts. In consequence, from MAM 2000 onwards, the terciles forecasts have been derived from the above/below median forecasts in such a manner that statistical consistency is preserved. The conceptual model underlying this forecast system consists of conditional shifts of the climatological probability distribution function without change of shape or variance, as described in Kumar and Hoerling (1999).

The evolution of the forecast systems (1998 to present) is summarised in Table 1.

\* A forecast system optimised for rainfall could not necessarily be expected to perform ideally for maximum or minimum temperature.

**Table 1. Evolution of NCC's seasonal forecasting systems (1998-2003).**

<i>Forecast periods</i>	<i>Predictors</i>	<i>Commentary</i>
NDJ 1998/1999 – FMA 2000	one-month SST indices (SST1 and SST2) at lags 1, 3 months	LDA system for seasonal rainfall only, using up to four SST-based predictors. Selection of which of the four predictors used varies seasonally and spatially. Model dataset is 1950-1993.
MAM 2000 – present	one-month SST indices (SST1 and SST2) at lags 1, 3 months	LDA system for seasonal rainfall, maximum and minimum temperatures using all four predictors. Three-category forecasts derived from two-category forecast model, rather than being generated directly as a separate forecasting problem. Model dataset expanded to 1950-1999. Historical time series of SST1, SST2 recalculated using GISST3 SST data (JJA 2000).

## Forecast and verification data

For the purposes of this article, the following sets of forecasts have been explored. All forecasts are issued on  $1^\circ \times 1^\circ$  grids across the country. The forecasts typically exhibit a high degree of smoothness, reflecting the fact that the rainfall (temperature) predictand fields are filtered using truncated EOF analysis as described previously. This means that the number of degrees of freedom represented in the forecasts is considerably less than the number of points on the  $1^\circ \times 1^\circ$  grid.

- (a) Above median seasonal maximum temperature forecasts. These cover the period MAM 2000 to JJA 2003 (40 forecasts), verified against climatological medians obtained from the 50-year dataset (1950-1999) used to develop the forecast model (49 years – 1950/51 to 1998/99 – in the case of the seasons NDJ and DJF). The analyses used for both the climatology and the verifications are the  $1^\circ \times 1^\circ$  Barnes successive correction analyses routinely generated by NCC using high-quality stations. They are described in Jones (1999).
- (b) Above median seasonal minimum temperature forecasts. As for maximum temperature.
- (c) Above median seasonal rainfall forecasts. These cover the period JJA 2000 to JJA 2003 (37 forecasts), likewise verified against climatological medians obtained from the 50-year dataset (1950-1999) used to develop the forecast model. The analyses used for both the climatology and the verifications are  $1^\circ \times 1^\circ$  sub-sampled versions of the Bureau's operational  $0.25^\circ \times 0.25^\circ$  rainfall analyses described by Jones and Weymouth (1997). (For technical reasons associated with the computer software used to generate the earlier forecasts, above median seasonal rainfall forecasts for periods prior to JJA 2000 have not been included in the present study.)

- (d) Three-category (terciles) seasonal rainfall forecasts. These cover the period NDJ 1998/99 to JJA 2003 (56 forecasts). These forecasts are verified against climatological fields calculated from a 98-year (1900-1997) dataset consisting of NCC's  $0.25^\circ \times 0.25^\circ$  Barnes analyses, with the forecasts being interpolated onto  $0.25^\circ \times 0.25^\circ$  versions for verification. (The use of a different resolution for this last set of forecasts simply reflects an older set of verification software.)

For the first three of these sets, essentially single forecast models are being verified (although there is a real sense in which each season represents a separate model, as the LDA model parameters are seasonally varying). The last set of forecasts includes forecasts from essentially two slightly different forecast models as described in the previous section, the earlier forecasts in the sequence being derived from 44 (43) year training periods, the later ones being derived from 50 (49) year training periods.

## Verification methodology

As described by Wilks (2001) and Potgieter et al. (2003), no single skill measure is sufficient for describing the multi-faceted nature of probability forecast quality. Failure to recognise this multi-dimensionality can lead to misleading conclusions about the accuracy (or skill) of probabilistic forecasts, and for users to inappropriately treat probabilistic forecasts as statements of categorical fact. For the purposes of this study, three principal forecast scoring techniques have been adopted for the verification of the seasonal forecasts. These techniques are well established in the verification literature, though (surprisingly) they have not featured significantly in past published analyses undertaken for Australia. (Brier skill-scores (see, for example, Wilks (1995)) have also been calculated for

the two-category forecasts. The results though have not been included in this article primarily for reasons of space, but also because the spatial distributions of the Brier skill-scores for the three forecast variables are qualitatively similar to the corresponding spatial distributions of the LEPS skill-scores.)

The first technique is the Linear Error in Probability Space (LEPS) scoring technique, primarily through the calculation of the probability weighted LEPS skill-score (Potts et al. 1996). This was the principal scoring technique used in assessing the expected skill of the models through scoring of cross-validated hindcasts (Drosowsky and Chambers 1998, 2001; Jones 1998). Details of how the LEPS scoring technique is applied in the present circumstances are given in Appendix 1.

The second technique is derived from the contingency table of outcomes (Table 2). For above and below median forecasting, the model is treated as producing a definite (categorical) forecast of an above median result if its calculated probability exceeds 0.5, and likewise a definite forecast of a below median result if its calculated probability for above median is less than 0.5.

Out of a set of  $n$  forecasts, on  $a$  occasions the forecast model predicts an enhanced probability of a below median outcome which is subsequently observed, whereas on  $b$  occasions the model predicts an above median outcome which is not subsequently observed. The other two counts  $c$  and  $d$  are similarly defined. For the purposes of this article, the ratio  $(a + d)/n$  will be called the Correct Forecast Rate (CFR), although within the context of the Bureau's SILO project and in response to user feedback, this ratio has been called the 'percent consistent'. The base rate (zero skill rate) for CFR over a long series of forecasts is 0.5.

The last technique considered here is that of 'reliability' (see Hartmann et al. (2002)). This involves conflating forecasts and the associated verifying observations across many different grid-points into a group and stratifying that group according to forecast probability. The intent is to assess for a given forecast probability of an above median outcome (say) the rate at which that outcome actually occurs. While this technique can

of course be applied in principle to a single grid-point, the number of forecasts available in the present study is very much smaller than would be necessary to make such an attempt meaningful. In an effort to minimise sampling variability and to give a whole-of-model estimate, the conflation has been performed across all Australian grid-points and all issued forecasts.

In performing this study our explicit goal is to describe and understand the performance of operationally issued climate forecasts over the past few years. We have considered and rejected the idea of estimating statistical significance for the presented results. The spatial dependence of both forecasts and observations not only strongly reduces the degrees of freedom but also makes those degrees of freedom difficult to estimate. The degrees of freedom are further reduced by auto-correlation in both forecasts and observations arising from the issuing of forecasts for overlapping periods. Further, the small sample of forecasts represents a likewise small set of instantiations of the climate system which therefore do not adequately characterise the full climatological range of possibilities. These factors combine to make the estimation of statistical significance very difficult and potentially misleading.

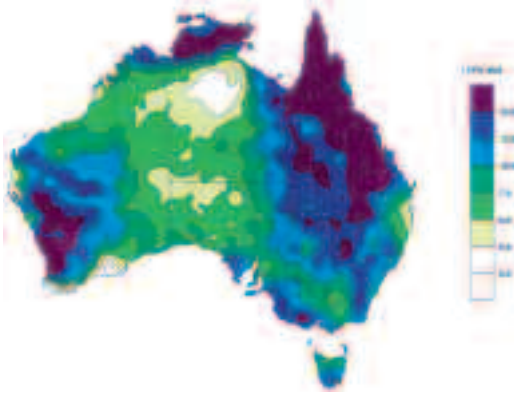
## Results

Figure 3 shows the spatial distribution of LEPS skill-scores for the 40 above/below median seasonal maximum temperature forecasts (MAM 2000 to JJA 2003), while Fig. 4 shows the corresponding distribution for seasonal minimum temperature. The temporal evolution of forecast skill is highlighted in Figs 5 and 6, which show the field-averaged (Australian) scaled LEPS scores (the solid line) for the 40 above/below median maximum and minimum temperature forecasts. The field-wide averages were computed from 687 grid-points at  $1^\circ$  intervals. To give an indication of the magnitude of the probability departures predicted, overlaid on these two figures (the dashed line) is the mean absolute forecast departure from climatology (0.5), across Australia.

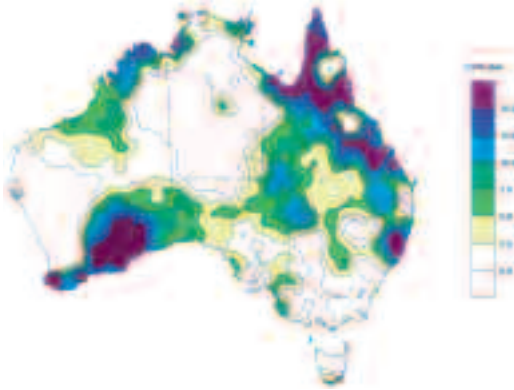
**Table 2.** Contingency table for scoring two-category seasonal forecasts.

		Forecasts		Totals
		Below median	Above median	
Observations	below median	$a$	$b$	$e = a + b$
	above median	$c$	$d$	$f = c + d$
	totals	$p = a + c$	$q = b + d$	$n = e + f$

**Fig. 3** LEPS skill-scores for above/below median seasonal maximum temperature (40 forecasts).



**Fig. 4** LEPS skill-scores for above/below median seasonal minimum temperature (40 forecasts).



Several things are apparent from these figures. The maximum temperature forecasts have shown positive skill across almost all of the country, with the best skill across the southwest and most of the east of the country. In contrast, the minimum temperature forecasts have been conspicuously less successful over the period for which operational forecasts have been issued. Skill has been particularly low across subtropical Western Australia and the Northern Territory.

The time series of Australia-wide averages for the individual forecasts indicate that it is the most recent forecasts which have scored highest. This can be attributable in large part to the El Niño event that began in May 2002, and is consistent with the experimental results of Jones (1998) for the 1997/98 El Niño. Such a result is very important, in so far as El

**Fig. 5** Australia-averaged LEPS scores (solid line) and mean absolute forecast departure from climatology (dotted line) for 40 seasonal maximum temperature forecasts. The spatial averages comprise 687 grid-points.



**Fig. 6** Australia-averaged LEPS scores (solid line) and mean absolute forecast departure from climatology (dotted line) for 40 seasonal minimum temperature forecasts. The spatial averages comprise 687 grid-points.



Niño events have traditionally been responsible for many of Australia's most damaging drought episodes, and indeed the 2002/03 El Niño drought was quite possibly Australia's most severe on record (Jones 2002).

The minimum temperature forecasts so far have been much less successful than the results from the cross-validated hindcasts would lead one to expect (Jones 1998; see also Appendix 2). The low skill largely reflects a sequence of forecasts in 2001 and early 2002 which verified poorly (see Fig. 6). Forecasts during this period tended to strongly favour above median minimum temperatures (largely as a result of very warm SSTs in the eastern Indian Ocean). Rather large probability departures were widespread, as evidenced by the mean absolute forecast departure

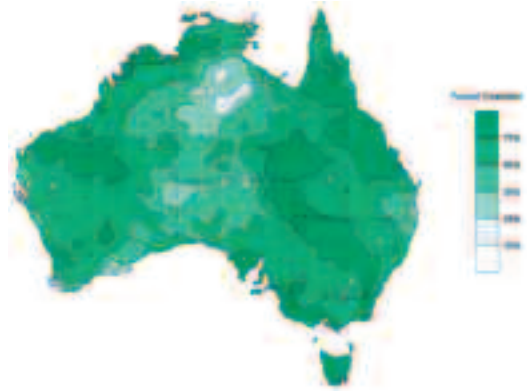
curve. The physical reasons for this sequence of failures are unclear, but likely reflect a fraction of weather noise. In an aggregate sense, the lower than expected skill of the minimum temperatures outlooks may be simply a question of sampling variability, which would be a justifiable conclusion if future performance yields lasting improvement (such as has been observed over the last ten months of the forecast sequence). On the other hand, continued poor performance over a long series of forecasts could suggest that (a) the forecast model is exhibiting artificial skill\*, or (b) the assumptions of the forecast model are being invalidated by climate change or low frequency variability.

In both these figures (and subsequently in Fig. 10), there is an observable tendency for strong (weak) field-averaged LEPS scores, whether positive or negative, to be associated with strong (weak) field-averaged absolute forecast departures from climatology. This follows directly from the LEPS scoring method (see Appendix 1), as the LEPS scores for an individual forecast at each grid-point are proportional (in magnitude) to the absolute forecast departures from climatology. In physical terms, greater forecast departures from climatology are possible in times and places of greater forecast skill (see for example, Kumar and Hoerling (1999)).

Figures 7 and 8 show the correct forecast rates as percentages for the seasonal maximum and minimum temperature forecasts respectively. These results are broadly consistent in terms of the amplitude of the signal with the LEPS skill-score results of Figs 3 and 4. Over much of New South Wales and Victoria, the maximum temperature model correctly preferred the subsequently observed outcome category more than 75 per cent of the time, and indeed has approached 100 per cent in small regions. Under this metric, the minimum temperature forecasts again demonstrate mixed skill, though we note that the area of positive skill is somewhat larger than that evident in the LEPS skill maps. This largely reflects the fact that the forecasts during 2001 and early 2002 were both rather emphatic and ‘categorically’ wrong, and were consequently punished more severely by the LEPS skill-score than by the correct forecast rate.

Figure 9 shows the LEPS skill-scores for the above/below median seasonal rainfall forecasts (JJA 2000 to JJA 2003). As expected from the cross-validated hindcast skill estimates (Drosowsky and

**Fig. 7** Correct forecast rate for above/below median seasonal maximum temperature forecasts (40 forecasts).



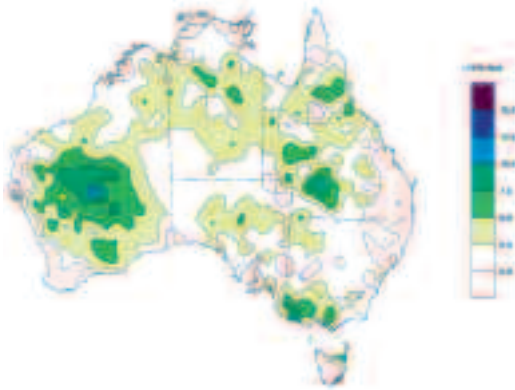
**Fig. 8** Correct forecast rate for above/below median seasonal minimum temperature forecasts (40 forecasts).



Chambers (2001); see also Appendix 2), the rainfall forecasts have shown less skill than the maximum temperature forecasts. Even so, the observed LEPS skill is above the climatological or no-skill value (zero) for most of the country. Figure 10 shows the Australia-wide average scaled LEPS scores for the 37 forecasts in the sample, along with the mean absolute forecast departures from climatology. Consistent with the tendency for seasonal rainfall to show considerable spatial variability and lower predictability (e.g. Rowell (1998)), the rainfall forecasts show smaller probability swings than do the temperature forecasts. It is noteworthy that the forecasts showed positive skill during both the extremely wet summer of 2000/01 (Fawcett 2002) and the severe drought of 2002/03 (Jones 2002), suggesting that users of this service would have had some advanced warning of these two extreme events.

\* Essentially this implies that the statistical relationships between predictors and predictand within the training period are basically chance relationships which do not survive into operational prediction because they are not dependent on underlying meteorological relationships.

**Fig. 9** LEPS skill-scores for above/below median seasonal rainfall forecasts (37 forecasts).



**Fig. 10** Australia-averaged LEPS scores (solid line) and mean absolute forecast departure from climatology (dotted line) for 37 seasonal rainfall forecasts. The spatial averages comprise 687 grid-points. Note: this graph is plotted on a different scale from Figs 3 and 4.

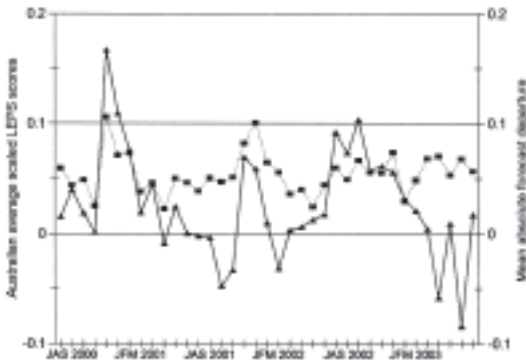


Figure 11 shows the correct forecast rate for the seasonal rainfall forecasts. The results are broadly consistent with the LEPS skill-scores mapped in Fig. 9. The greatest success has been over Victoria, Western Australia and the Northern Territory. This is not entirely consistent with the cross-validated hindcast data which suggest that over the long term, the greatest success should be in northeast Australia. This discrepancy likely reflects the fact that for much of the verification period the Pacific Ocean showed near-neutral ENSO conditions, whereas it is known that much of the predictability in northeastern parts arises from the influence of El Niño and La Niña episodes on rainfall.

**Fig. 11** Correct forecast rate for above/below median seasonal rainfall forecasts (37 forecasts).

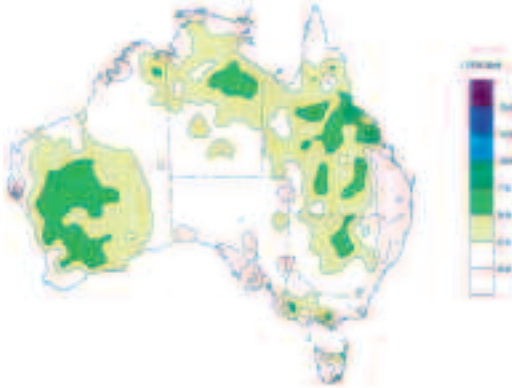


Figure 12 shows the LEPS skill-scores for the 56 terciles seasonal rainfall forecasts (NDJ 1998/99 to JJA 2003). The general patterns are similar to the equivalent map for above/below median rainfall (Fig. 9), though the forecast skill has tended to be a little higher for the most recent period. The tercile verifications agree somewhat better with the hindcast skill estimates (Drosowsky and Chambers 1998, 2001). Correct forecast rates for the terciles (not shown) are above the climatological rates across most of the country, and particularly so over Western Australia and the Northern Territory. The rates across the eastern States and the southern half of South Australia are less impressive, however.

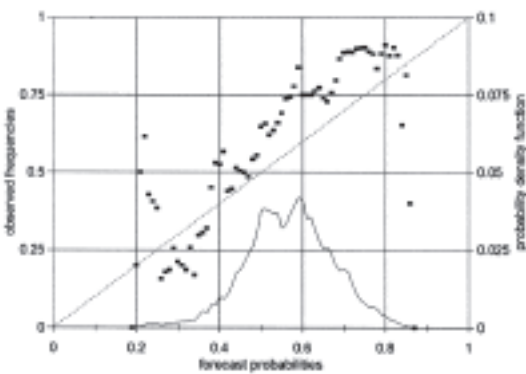
Figure 13 shows the reliability data (plotted as rectangles) for the 40 seasonal above median maximum temperature forecasts aggregated across all 687  $1^\circ \times 1^\circ$  Australian grid-points (in effect, 27,480 individual forecasts). The horizontal axis shows the forecast probabilities for above median outcomes, aggregated into bins of integer percentage probability. For each probability bin, the left-hand side vertical axis shows the observed outcome rate conditional on the forecast probability. The solid line (right-hand side vertical axis) shows the distribution (probability density function) of forecast probabilities across all grid-points and all forecasts.

Over a long sequence of forecasts, the reliability data for a 'perfectly reliable' forecast model should asymptote to  $y = x$ , shown as a dashed line in the figure. In contrast, the zero skill outcome for these reliability calculations is the horizontal line  $y = 0.5$ . For the extreme probabilities (small and large), no reliability data are shown because no Australian grid-points have shown those forecast probabilities within the sequence of forecasts under assessment. This cor-

**Fig. 12** LEPS skill-scores for the terciles seasonal rainfall forecasts (56 forecasts).



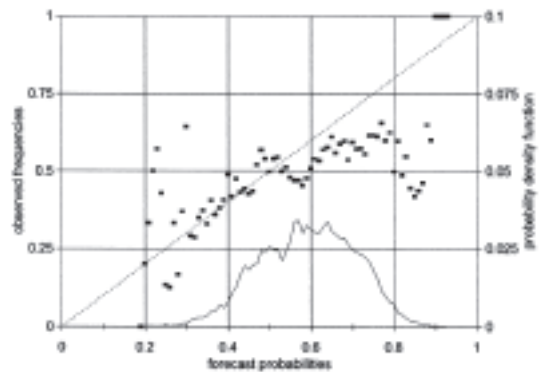
**Fig. 13** Reliability data (rectangles: climatological rate dotted line) and density function of forecast probabilities (solid line) for 40 above median maximum temperature forecasts.



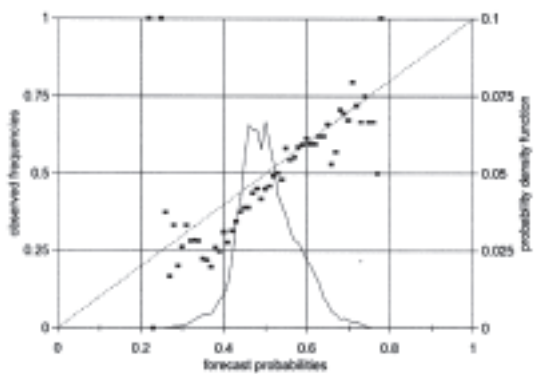
responds to zero values of the associated probability density function. In interpreting the reliability data (e.g. comparing them with base rates), it is important to take into consideration the distribution of forecast probabilities. For ‘wide’ probability bins, meaningful comparisons with base rates cannot be done without weighting the results according to the distribution of forecast probabilities. Integer probability bins as used here are sufficiently narrow to mean that this is not an issue, but suffer the disadvantage of significant sampling variability, particularly for the more extreme and poorly sampled probabilities.

Figures 14 and 15 show the corresponding results for seasonal above median minimum temperature (40 forecast periods) and seasonal above median rainfall (37 forecast periods), respectively. In all three cases, the corresponding results for seasonal below median

**Fig. 14** Reliability data (rectangles: climatological rate dotted line) and density function of forecast probabilities (solid line) for 40 above median minimum temperature forecasts.



**Fig. 15** Reliability data (rectangles: climatological rate dotted line) and density function of forecast probabilities (solid line) for 37 above median rainfall forecasts.



variables are obtainable simply by rotating the reliability data as graphed by  $180^\circ$  about the centre of graph (the point (0.5,0.5)).

The reliability data for seasonal maximum temperature (Fig. 13) and rainfall (Fig. 15) are both remarkably consistent with the ‘perfect reliability’ curves, and clearly much closer to them than to the ‘zero skill’ curves, given the relatively small number of forecasts involved (even when the aggregation of grid-points across Australia is taken into account\*). In both cases, it appears that the forecasts are slightly

\* That the number of forecasts is ‘small’ as far as the reliability data is concerned is evidenced by the considerable variation in the graphs from probability bin to probability bin, even though the underlying forecast grids are themselves quite smooth, as are the verifying temperature grids. The verifying rainfall grids are of course more detailed.

'under-confident' in terms of their departures from climatology. That is, they tend to underpredict probability swings. For example, when a forecast probability of above median of 60 per cent is issued, on average the above median outcome will be observed slightly more than 60 per cent of the time, with corresponding results for raised probabilities for below median outcomes. The fact that the forecasts are relatively reliable (or a little 'under confident') means that users with suitable cost/loss ratios can be expected to obtain long-term benefit from judicious use of the seasonal forecast information (Wilks 2001). However, the relatively tight distribution of forecast probabilities around the base rate of 0.5 for rainfall, as indicated by the rather narrow density function in Fig. 15, means that benefit is likely to accrue rather slowly when viewed across Australia as a whole. Of course, the skill of forecast varies both spatially and temporally, and blanket statements about forecast utility which ignore this fact are likely to be at best misleading and potentially costly to forecast users.

The fact that the forecasts appear a little 'under-confident' is consistent with the manner in which the forecast probabilities for the underlying rainfall and temperature principal components, generated internally within the LDA forecast model, are distributed amongst the grid-points according to the fraction of variance explained (always less than 1). The 'remedy', an artificial after-the-fact inflation of the forecast probability variance, has been considered and rejected as having no sound justification.

The reliability data for seasonal minimum temperature (Fig. 14) are obviously less consistent with the base rate curve. This outcome agrees with the overall lesser success of the minimum temperature forecasts, as shown previously. There is a suggestion in the figure that probability shifts towards below median minimum temperatures are more 'reliable' than the probability shifts towards above median minimum temperatures. Such a result is consistent with the run of forecasts issued during 2001 and early 2002, which favoured warm conditions but verified poorly.

As evidenced by the reliability diagrams and the associated distributions of forecast probabilities, forecasts of increased likelihood of above median temperatures have been in the great majority, largely reflecting the fact that the eastern Indian Ocean has been dominated by large positive anomalies for the past three years. An interesting result is that over the study period, night-time temperatures across Australia have averaged close to the long-term average, whereas SSTs in the eastern Indian Ocean (as measured by SST2) have been continuously at near-record values. The exceptionally warm Indian Ocean temperatures during this period largely reflect a continuation of an

approximately century-long warming trend (Hoerling and Kumar 2003) which is likely to be at least partly attributable to anthropogenic climate change (IPCC 2001). Since 1950, Australian minimum temperatures have similarly trended upwards, with the run of relatively normal values during the past few years being unusual by recent historical standards.

## Discussion and conclusions

This article has presented verification results for the current operational seasonal forecasting systems of the Bureau of Meteorology's National Climate Centre. In doing so, this represents the first rigorous assessment of seasonal forecast skill for Australia in nearly a decade. The current forecast variables comprise seasonal maximum and minimum temperature and rainfall. These forecasts are routinely verified, using calculation of LEPS skill-scores, correct forecast rates and reliability curves. Verification results are typically available within a few days of the end of the forecast period, and distributed within the Bureau of Meteorology via a website.

The seasonal maximum temperature forecasts have shown positive skill since their inception in autumn 2000. While the operational rainfall forecasts across this period, and over the longer term since SST-based forecasts began in 1998, have shown less success, the results have exceeded climatological expectation over almost all of Australia. The minimum temperature forecasts have so far been less successful, largely as a result of poor forecasts in 2001 and early 2002, although there has been considerable improvement in outcomes towards the end of the forecast sequence. For each variable, the operational forecast service performed with considerable skill during the severe 2002/03 drought, highlighting the potential value of climate forecasts when the anomalous climate forecast signal is strong.

While the statistical significance of the results described in this paper must necessarily be tempered by the relatively short period available for forecast verification, it is interesting to compare and contrast our results with those described in Smith (1994). In contrast with this early study, we have found the seasonal rainfall forecasts to be skillful over most of Australia, and encouragingly these appear to be well calibrated (or reliable). Importantly, forecast reliability ensures that users with suitable cost/loss ratios can be expected to obtain long-term benefit from judicious use of the seasonal forecast information (Wilks 2001).

The good calibration of the forecasts suggests that forecast skill is being principally limited by the underlying predictability captured by the two SST

predictors. Clearly, the future challenge is to capture a greater fraction of the climate variability in the predictor data. Given the shortness of potential training periods for empirical models (typically about 50 years) and recent large-scale climate change (e.g. IPCC 2001, IOCI 2002), the authors believe significant further advance will likely require a shift to coupled model-based prediction techniques.

## Appendix 1

Specific details are given below of how the LEPS scoring technique has been applied in this study to probabilistic forecasts for two and three climatologically equally likely categories, along with how to calculate the LEPS skill-scores.

### Above/below median (two-category) forecasts

Suppose the probability of an above median outcome at a given point is given as  $p_2$ , with the corresponding probability of a below median outcome being  $p_1 = 1 - p_2$ . If the outcome is below median then the forecast is given the LEPS score  $s = 1/6(p_1 - p_2)$ , whereas if the outcome is above median then the forecast is given the LEPS score  $s = 1/6(p_2 - p_1)$ . These scores range from  $-1/6$  to  $+1/6$ , but can be multiplied by 6 to obtain a scaled LEPS score which ranges from  $-1$  to  $+1$  for an individual forecast. Field-averaged LEPS scores scaled in just this way are shown in Figs 5, 6 and 10. These field averages can therefore also range from  $-1$  to  $+1$ . For a sequence  $\{s_1, \dots, s_n\}$  of these LEPS scores, the LEPS skill-score is calculated as

$$\text{LEPS SKILL} = \frac{6}{n} \sum_{i=1}^n s_i .$$

As constructed, the LEPS skill-score can range from  $-1$  to  $+1$ .

### Terciles (three-category) forecasts

Suppose the probability of a tercile 1 outcome at a given point is given as  $p_1$ , with the corresponding probabilities for terciles 2 and 3 being  $p_2$  and  $p_3$  respectively. Obviously,  $p_1 + p_2 + p_3 = 1$ . If the outcome is in tercile 1 then the forecast is given the LEPS score  $s = 8/27 p_1 - 1/27 p_2 - 7/27 p_3$ , if the outcome is in tercile 2 then the forecast is given the LEPS score  $s = -1/27 p_1 + 2/27 p_2 - 1/27 p_3$ , and if the outcome is in tercile 3 then the forecast is given the LEPS score  $s = -7/27 p_1 - 1/27 p_2 + 8/27 p_3$ . These scores range from  $-7/27$  to  $+8/27$ , but can be multiplied by  $27/8$  to give a scaled LEPS score which ranges from  $-7/8$  to  $+1$ . For a sequence  $\{s_1, \dots, s_n\}$  of these LEPS scores, calculate two additional sequences,  $\{u_1, \dots, u_n\}$  which is the sequence of maximum possible LEPS scores given

the observed outcomes, and  $\{l_1, \dots, l_n\}$  which is the sequence of minimum possible LEPS scores given the observed outcomes. Obviously this implies that  $l_i \leq s_i \leq u_i$  for  $i = 1, \dots, n$ . If the outcome for the  $i$ th forecast is a tercile 1 or tercile 3 outcome, then  $u_i = 8/27$  and  $l_i = -7/27$ , whereas if the outcome for the  $i$ th forecast is a tercile 2 outcome, then  $u_i = 2/27$  and  $l_i = -1/27$ . The LEPS skill-score for the sequence of forecasts is calculated as

$$\text{LEPS SKILL} = \frac{\sum_{i=1}^n s_i}{\sum_{i=1}^n u_i}$$

if the mean LEPS score is non-negative, and as

$$\text{LEPS SKILL} = \frac{\sum_{i=1}^n s_i}{-\sum_{i=1}^n l_i}$$

otherwise. As in the two-category case, the LEPS skill-score can range from  $-1$  to  $+1$ . While the calculation of the LEPS skill-score for the three-category case appears to be much more complicated than in the two-category case, the same principles have in fact been applied in both cases. The fact that the scoring coefficients all have magnitude  $1/6$  in the two-category case causes the apparent simplification.

The LEPS skill-scores\* mapped in Figs 3, 4, 9 and 12 are expressed as percentages, rather than as fractions of 1, and can therefore theoretically range from  $-100\%$  to  $+100\%$ . It is a property of the LEPS skill-scoring technique that skill-scores for terciles forecasts may be compared more or less directly with those for above/below median forecasts even though the categories involved are qualitatively different. Accordingly, the skill-scores in Fig. 12 are presented on the same scale as those for Figs 3, 4 and 9. The LEPS score base rate for individual forecasts is zero in both the two-category and three-category cases, this being the score awarded to a climatological forecast ( $p_1 = p_2 = 1/2$  in the two-category case and  $p_1 = p_2 = p_3 = 1/3$  in the three-category case). The base rate for the LEPS skill-score is also zero in both cases.

\* For purposes of comparison, a simple linear regression model with an underlying correlation coefficient of  $+0.4$  between predictor and predictand, when converted into a model for predicting above and below probabilities (two categories), would possess a LEPS skill-score slightly in excess of  $+10\%$  over an infinitely long series of forecasts. For an underlying correlation coefficient of  $+0.5$ , the corresponding LEPS skill-score would be slightly in excess of  $+16\%$ . The same simple linear regression model converted instead into a model for predicting tercile probabilities (three categories) would generate LEPS skill-scores slightly higher than those of the two-category model for the same underlying correlation coefficients. This reflects the additional forecast information conveyed by the three-category model over the corresponding two-category model.

## Appendix 2

For purposes of comparison, the distributions of correct forecast rates for above median seasonal forecasts obtained by cross-validated hindcasting across the 50-year dataset used to develop the latest version of the forecast model are shown in Figs 16 (seasonal maximum temperature), 17 (seasonal minimum temperature) and 18 (seasonal rainfall). These maps have been supplied to the authors by Dr Andrew Watkins (personal communication). The hindcast dataset consists of the 598 seasons from JFM 1950 to OND 1999.

Comparing Fig. 16 with Fig. 7 (seasonal maximum temperature), the recent independent forecasts have performed better over eastern Australia and most of Western Australia than would be expected from the cross-validated hindcast results, although there are some similarities to the overall patterns. Figure 17 (minimum temperature) is very similar to Fig. 16, but bears little similarity to Fig. 8 (see previous discussion).

Figure 18 (seasonal rainfall) shows the most skill across Queensland and the Northern Territory as far as the cross-validated hindcasts are concerned, but the recent skill (Fig. 9) shows better outcomes across much of Western Australia. Across the rest of the country there are weak similarities in the patterns of Figs 9 and 18, but Fig. 9 is much patchier reflecting typical rainfall anomaly length scales and the much shorter period of forecasts.

## References

- Alves, O., Wang, G., Zhong, A., Smith, N., Warren, G., Marshall, A., Tseitkin, F. and Schiller, A. 2003. POAMA: Bureau of Meteorology Operational Coupled Model Seasonal Forecast System. In *Proceedings of the ECMWF Workshop on the role of the upper ocean in medium and extended range forecasting*, ECMWF, Reading, UK, 13-15 November 2002.
- Bureau of Meteorology 1992. A review of the National Climate Centre Seasonal Outlook Service during the El Niño episode of 1991/92. *National Climate Centre Report*, Bur. Met., Australia, 7 pp.
- Casey, T. 1995. Optimal linear combination of seasonal forecasts. *Aust. Met. Mag.*, 44, 219-24.
- Casey, T. 1998. Assessment of a seasonal forecast model. *Aust. Met. Mag.*, 47, 103-11.
- Davis, R.E. 1976. Predictability of sea surface temperature and sea level pressure over the north Pacific ocean. *J. Phys. Oceanogr.*, 6, 249-66.
- Drosowsky, 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.
- Drosowsky, W. and Chambers, L. 1998. Near global sea surface temperature anomalies as predictors of Australian seasonal rainfall. *BMRC Research Report No. 65*, Bur. Met., Australia (available on the Bureau's website at [www.bom.gov.au/bmrc/cffor/cfstaff/wld/RESREP65/rr65.htm](http://www.bom.gov.au/bmrc/cffor/cfstaff/wld/RESREP65/rr65.htm)).
- Drosowsky, W. and Chambers, L.E. 2001. Near global sea surface temperature anomalies as predictors of Australian seasonal rainfall. *Jnl climate*, 14, 1677-87.

**Fig. 16** Correct forecast rate for above/below median seasonal maximum temperature (598 cross-validated hindcasts).



**Fig. 17** Correct forecast rate for above/below median seasonal minimum temperature (598 cross-validated hindcasts).



**Fig. 18** Correct forecast rate for above/below median seasonal rainfall (598 cross-validated hindcasts).



- Fawcett, R.J.B. 2002. Seasonal climate summary southern hemisphere (summer 2000/01): a third successive positive phase of the Southern Oscillation continues. *Aust. Met. Mag.*, 51, 49-57.
- Hartmann, H.C., Pagano, T.C., Sorooshian, S. and Bales, R. 2002. Confidence builders: evaluating seasonal climate forecasts from user perspectives. *Bull. Am. Met. Soc.*, 84, 683-98.
- Hoerling, M. and Kumar, A. 2003. The perfect ocean for drought. *Science*, 299, 691-4.
- IOCI 2002. *Climate variability and change in south west Western Australia*. Indian Ocean Climate Initiative Panel, Perth, September 2002, 34 pp.
- IPCC 2001. *Climate Change 2001: The scientific basis*. World Meteorological Organization/United Nations Environmental Programme. Cambridge Univ. Press, Cambridge, UK, 881 pp.
- Jones, D.A. 1998. The prediction of Australian land surface temperatures using near global sea surface temperature patterns. *BMRC Research Report No. 70*, Bur. Met., Australia (available on the Bureau's website at [www.bom.gov.au/climate/ahead/rr70/](http://www.bom.gov.au/climate/ahead/rr70/)).
- Jones, D.A. 1999. Characteristics of Australian land surface temperature variability. *Theor. Appl. Climatol.*, 63, 11-31.
- Jones, D.A. 2002. The 2002 El Niño and its impacts on Australia. *Bull. Aust. Met. Oceanog. Soc.*, 15, 91-5
- Jones, D.A. and Weymouth, G. 1997. An Australian monthly rainfall data set. *Technical Report No. 70*, Bur. Met., Australia.
- Kumar, A. and Hoerling, M.P. 1999. Analysis of a conceptual model of seasonal climate variability and implications for seasonal prediction. *Bull. Am. Met. Soc.*, 81, 255-64.
- Lau, N.-C. and Nath, M.J. 1994. A modeling study of the relative roles of tropical and extratropical SST anomalies in the variability of the global atmosphere-ocean system. *Jnl climate*, 7, 1184-207.
- Michaelsen, J. 1987. Cross-validation in statistical climate forecast models. *Jnl Clim. Appl. Met.*, 26, 1589-600.
- Potgieter, A.B., Everingham, Y.L. and Hammer, G.L. 2003. On measuring quality of a probabilistic commodity forecast for a system that incorporates seasonal climate forecasts. *Int. J. Climatol.*, 23, 1195-210.
- Potts, J.M., Folland, C.K., Joliffe, I.T. and Sexton, D. 1996. Revised "LEPS" scores for assessing climate model simulations and long-range forecasts. *Jnl climate*, 9, 34-53.
- Richman, M.B. 1986. Rotation of principal components. *J. Climatol.*, 6, 293-335.
- Reynolds, R.W. and Smith, T.M. 1994. Improved global sea surface temperature analysis. *Jnl climate*, 7, 929-48.
- Rowell, D.P. 1998. Assessing potential seasonal predictability with an ensemble of multidecadal GCM simulations. *Jnl climate*, 11, 109-20.
- Simmonds, I. and Rocha, A. 1991. The association of Australian winter climate with ocean temperatures to the west. *Jnl climate*, 4, 1147-61.
- Smith, I. 1994. Assessments of categorical rainfall predictions. *Aust. Met. Mag.*, 43, 143-51.
- Tourre, Y.M. and White, W.B. 1995. ENSO signals in global upper-ocean temperature. *J. Phys. Oceanogr.*, 25, 1317-32.
- Walsh, K.J.E., Bettio, L., Power, S., Fawcett, R. and Pahalad, J. 2001. Extended seasonal prediction of precipitation in Fiji. *Aust. Met. Mag.*, 50, 195-203.
- 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.
- Watkins, A. B. 2002. "Well how good is it?" – Seasonal outlook verification for the non-specialist. *AMOS 9th National Conference Abstract Volume*. Australian Meteorological and Oceanographic Society, p.77.
- Wilks, D.S. 1995. *Statistical methods in the atmospheric sciences*. Int. Geophysical Series, Academic Press, San Diego, USA, 467 pp.
- Wilks, D.S. 2001. A skill score based on economic value. *Meteorol. Appl.*, 8, 209-19.
- Yu, Z.-P., Chu, P.-S. and Schroeder, T. 1997. Predictive skills of seasonal to annual rainfall variations in the US Affiliated Pacific Islands: Canonical correlation analysis and multivariate principal component regression approaches. *Jnl climate*, 10, 2586-99.

