Aust. Met. Mag. 48 (1999) 159-167

An experiment to establish the limits of our predictive capability of weather elements for Melbourne

Harvey Stern
Bureau of Meteorology, Melbourne, Australia
(Manuscript received June 1998; revised November 1998)

The results of an experiment, which involves verifying a set of quantitative forecasts for Melbourne out to 14 days, are presented. The results are used to assess whether or not extending the period of the official forecasts beyond four days might be justified. The experimental forecasts are verified against 'climatology' and a randomly generated set of forecasts. The verification data derived using the methodology suggest that, at present, routinely providing or utilising day-to-day forecasts beyond day 4 would be inappropriate. The data also suggest, however, that it might be possible to provide some useful information about the likely weather up to about six days in advance for some elements, in some seasons and in some situations, for example, maximum temperature during summer. By contrast, in some circumstances it may not be possible to provide useful information even for day 1. Nevertheless, the data indicate that it might be possible to make useful statements about the expected average weather conditions over the 10-day period between days 5 and 14.

Introduction

Background

The Victorian Regional Office (VRO) of the Australian Bureau of Meteorology (BoM) currently provides one to four-day weather forecasts to the general public. These forecasts are issued each afternoon by the VRO Regional Forecasting Centre (RFC) in Melbourne. The forecasts are based upon an interpretation, in terms of local weather, of the output of various global and Australian region numerical weather prediction (NWP) models.

Corresponding author address: Harvey Stern, Victorian Regional Office, Bureau of Meteorology, GPO Box 1636M, Melbourne, Vic. 3001, Australia.
e-mail: H.Stern@bom.gov.au

This interpretation is carried out using a combination of the Generalised Analogue Statistics Model (GASM) (Stern, 1980a, 1980b; Dahni et al. 1984; Dahni 1988; Dahni and Stern 1995; Stern 1996), Model Output Statistics (MOS) (Woodcock 1984) and other guidance (including simple objective forecasting aids).

There is potential for extending the application of such forecast guidance schemes because interpreting local weather in terms of the synoptic flow is readily automated by statistical methodologies. Indeed, Brooks (1995) wrote that 'technology, which initially allowed humans to make routine weather forecasts, will soon close that avenue of human endeavour ... (and thereby permit) concentration on severe events'.

That this prediction by Brooks may soon become a reality is supported by Fig. 1, which presents forecast verification data for seventeen Victorian centres. Figure 1 suggests that, overall, whereas human forecasters are capable of significantly improving upon computer generated guidance for short-term predictions of temperature, that capability is much reduced for long-term predictions.

However, this reduction in performance may partly be a consequence of less attention being able to be directed towards the lower priority long-term predictions. Figure 2, which presents forecast verification data for Melbourne alone, indicates that where forecasters (particularly experienced forecasters, such as those in the VRO RFC) focus on a particular location, that capability is somewhat preserved. In the Victorian office, forecasters would be expected to focus on Melbourne, the State capital.

The United States National Centers for Environmental Prediction (NCEP) currently produce a 15-day global ensemble average prognosis (Toth and Kalnay 1993; Tracton and Kalnay 1993; Climate Diagnostics Center 1997). That the NCEP prognosis extends out to 15 days is consistent with the work of Lorenz (1963, 1969a, 1969b, 1993), which suggests a 15-day limit to day-to-day predictability of the atmosphere.

Fig. 1 **Chart illustrating the extent that official maximum temperature forecasts (squares) for seventeen Victorian centres, issued several days ahead, improved upon the guidance generated by GASM using the European Centre for Medium-range Weather Forecasts (ECMWF) global model output (diamonds). The January 1997 to June 1997 official VRO RFC data, upon which this chart is based, were provided by Setek (personal communication 1997).**

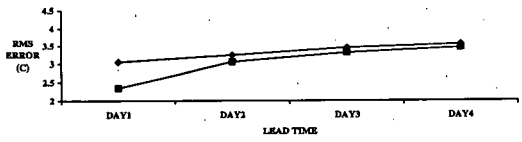
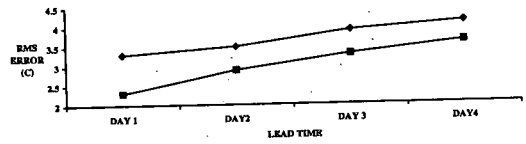


Fig. 2 **As Fig. 1, but for Melbourne alone. The January 1997 to June 1997 official VRO RFC data, upon which this chart is based, were provided by Setek (personal communication 1997).**



Purpose

This paper has a two-fold purpose:

(a) To present preliminary results of an experiment, which involves verifying a set of quantitative forecasts for Melbourne out to 14 days. These forecasts are based on a subjective interpretation of the NCEP predictions of mean sea-level pressure (MSLP), 500 hPa height and 1000-500 hPa thickness distributions, when available, for days 5 to 14, and on the official forecasts for days 1 to 4; and,

(b) To use the results to assess whether or not extending the period of the official forecasts beyond four days might be justified. After all, 'verification allows forecasters to know, quantitatively and objectively, how well they are doing, and in what ways they can improve their product' (Doswell 1995).

The work presented is an update of a paper by Stern (1998) which was mainly based on data from the winter season only.

Verification methodology

Verification measures

The experimental forecasts are verified against 'climatology' and a randomly generated set of forecasts. The 'climatology' of a particular weather element is regarded as the monthly mean value of that element – the calendar-day climatology displays day-to-day fluctuations that are too great for it to be considered a stable measure of climatology. The randomly generated set of forecasts are the official day 1 forecasts that were issued 15 days prior to the verifying day, given the 15-day theoretical limit on the relationship between weather patterns and, hence, atmospheric predictability (Lorenz 1963, 1969a, 1969b, 1993).

Several verification measures are used, with a view to establishing a possible limit to actual predictive capability.

These measures are:

(a) root mean square (rms) error of the minimum temperature forecasts, '*af*';

(b) rms error of the maximum temperature forecasts, '*bf*';

(c) rms error of the quantitative precipitation forecasts (QPF) in rainfall ranges, as defined by Stern (1980a), ranges 0, 1, 2, 3, 4, 5, etc. being respectively, less than 0.2 mm, 0.2-2.5 mm, 2.6-5 mm, 5.1-10 mm, 10.1-20 mm, 20.1-40 mm etc., '*cf*';

(d) percentage rain/no rain (R/NR) forecasts correct, '*df*' ('*df*' has a strong dependence on the climate probability, which varies through the year, and this variation slightly reduces the usefulness of the measure); and,

(e) Brier score (Brier 1950) about probability of precipitation (PoP) forecasts, as modified in accordance with how it is now 'used almost universally' (Wilks 1995), '*ef*'.

For each of the forecast days 1 to 14 inclusive, these measures are calculated. They are then compared with corresponding measures of the performance of climatology (*ac, bc, cc, dc, ec*) and combined into a series of skill scores for each of the Melbourne day 1 to 4 official forecasts and day 5 to 14 experimental forecasts.

Verification skill scores

Table 1 summarises the definitions of the verification measures employed to evaluate the forecasts. The 'minimum temperature skill score' (Fig. 3), the 'maximum temperature skill score' (Fig. 4), the 'QPF skill score' (Fig. 5) and the 'R/NR skill score' (Fig. 6), and the 'Brier skill score' (Fig. 7), require little further explanation. In regard to the 'R/NR skill score' and the 'Brier skill score', the square roots of (*df/dc*) and (*ec/ef*) are required, in order that they be directly comparable to (*ac/af*), (*bc/bf*) and (*cc/cf*) (which intrinsically do include a square root function). It is also worthwhile to note that the 'R/NR skill score' appears to be the inverse of the others because, unlike the others, its components increase, rather than decrease, as accuracy increases.

The 'temperature skill score' (Fig. 8) is simply the mean of the two temperature skill scores. The 'rainfall skill score' (Fig. 9) is simply the mean of the three precipitation skill scores. The 'joint skill score' (Fig.

Table 1. The definitions of the verification measures employed to evaluate the forecasts.

Skill score	*Definition*
minimum temperature	100((ac/af)-1)
maximum temperature	100((bc/bf)-1)
QPF	100((cc/cf)-1)
R/NR	$100((df/dc)^{1/2}-1)$
Brier	$100((ec/ef)^{1/2}-1)$
Temperature	100((ac/af)+(bc/bf)-2)/2
Rainfall	$100((cc/cf)+(df/dc)^{1/2}+(ec/ef)^{1/2}-3)/3$
joint	$((100((ac/af)+(bc/bf)-2)/2)+(100((cc/cf)+(df/dc)^{1/2}+(ec/ef)^{1/2}-3)/3))/2$

Fig. 3 Minimum temperature skill score compared to climatology. The predictions verified are the official forecasts, for days 1 to 4, and the forecasts based on a subjective interpretation of the NCEP output, for days 5 to 14. Positive values show skill better than climatology. A perfect score is 'infinite' (random = -21).

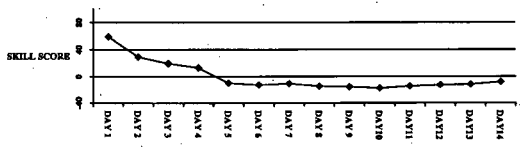


Fig. 4 As Fig. 3, but for maximum temperature skill score (random = -21).

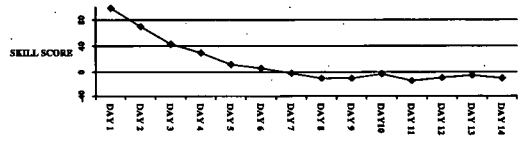


Fig. 5 As Fig. 3, but for QPF skill score (random = -11).

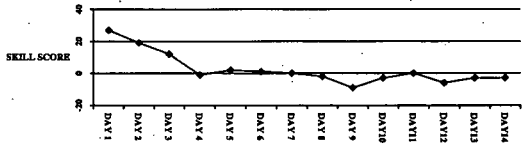


Fig. 6 As Fig. 3, but for R/NR skill score (random = -4).

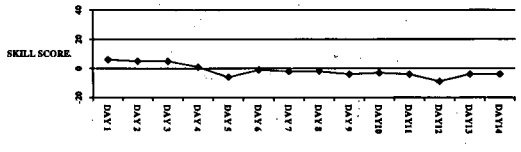


Fig. 7 As Fig. 3, but for Brier skill score (random = -4).

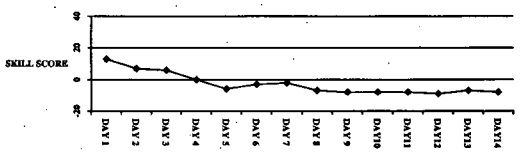


Fig. 8 **As Fig. 3, but for temperature skill score (random = -21).**

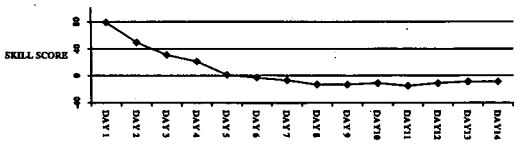


Fig. 9 **As Fig. 3, but for rainfall skill score (random = -6).**

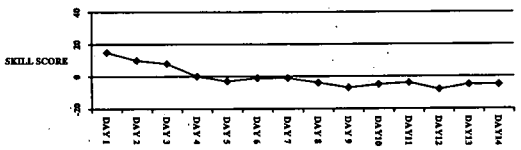
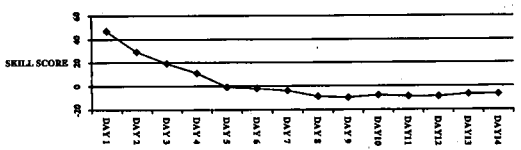


Fig. 10 **As Fig. 3, but for joint skill skill score (random = -13).**



10) is simply the mean of the temperature and rainfall skill scores (the 'joint skill score' should always be considered in the context of its components because, taken on its own, it represents a fairly radical reduction in forecast verification dimensionality – see Murphy (1991)).

In addition, skill scores for sets of randomly generated forecasts are quoted in the figure captions.

Interpreting worded forecasts

In order to verify the official forecasts objectively, it is necessary to interpret worded components of the one to four-day forecasts in terms of QPF, R/NR and PoP.

Stern (1980a) developed a standard system of terminology linked directly with the coded observations, as recorded in the official observation book (BoM 1977) and this is presented, in full, in Stern's (1980a) Appendix A. The format for the system of terminology is a description based on weather and cloud, followed by a description based on wind, followed by minimum and maximum temperatures and rainfall amount, followed by a precis and clarification of preceding components. Stern's (1980a) Appendix C includes an objective scheme to 'translate' official forecasts into that terminology and a consequential objective means to evaluate the official forecast.

Stern's (1998) Table 1 interpretation of worded components of the one to four-day forecasts was based on the aforementioned scheme. However, Stern (1998) reports 'that some of the words might better have been assigned ... (different interpretations) ... than they were'. The present paper's Table 2 represents an adjustment of that earlier interpretation, the adjustment being based on data gathered during the period of the present experiment. As a result, some of the words relating to precipitation probability are assigned lower probabilities in the present paper than they were by Stern (1998).

Figure 11 depicts the frequency with which precipitation occurred in association with various PoPs. It provides support to Table 2's interpretation of the worded components in terms of PoP as well as to the forecasters' skill at composing these components.

Figure 12 depicts the mean precipitation that occurred in association with various QPFs. It provides support to Table 2's interpretation of the worded components in terms of QPF as well as to the forecasters' skill at composing these components.

Fig. 11 **Percentage of occasions day 1, 2, 3 and 4 PoPs were associated with precipitation. Skill is demonstrated by the fact that the percentage increases as the PoP estimates increase.**

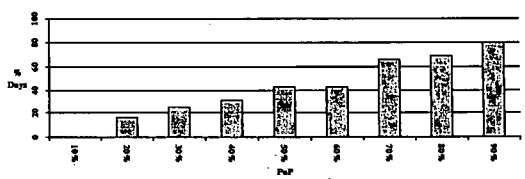


Fig. 12 **Observed mean precipitation associated with day 1, 2, 3 and 4 QPFs. Skill is demonstrated by the fact that the observed mean precipitation increases as the QPFs increase.**

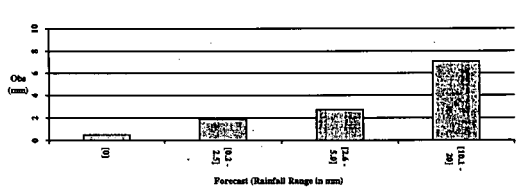


Table 2. Interpretation of words in terms of PoP (%) and QPF (Ranges). R/NR is defined by PoP being above or below 50%. It is set at a 50/50 chance if PoP is 50%. The second figure in a column applies if precipitation is referenced only in one of the pre-noon / post-noon periods (for example, 'showers clearing', 'morning drizzle').

Words	*PoP (%)*	*QPF (Ranges: 0, 1, 2, 3, 4 are respectively <0.2mm, 0.2-2.5mm, 2.6-5mm, 5.1-10mm, 10.1-20mm)*
Sunny	10	0
Mainly sunny; Fine	20	0
Becoming sunny; Partly cloudy; Becoming cloudy; Mainly cloudy; Cloudy	30	0
Chance of precipitation; Chance of thunder; Unsettled; Cool change (with no precipitation reference); Local showers; Local thunder; Fog or drizzle; Drizzle patches; Little drizzle; Becoming fine	40	0
Few showers; Little rain; Mainly fine; Shower or two	50, 40	1, 0
Drizzle	80, 60	1
Showers	80, 70	2, 1
Snow; Sleet	90, 80	2, 1
Thunderstorms	80, 70	3, 2
Rain	90, 80	3, 2
Showers, heavy at times	90	3
Rain at times	90	3
Rain, heavy at times; Thunderstorms, heavy at times	90	4

Discussion

Forecast performance

Figures 3 to 10 show that skill at predicting both temperature and rainfall decline, albeit unsteadily, as one moves from day 1 to day 14. The unsteady character of some of the declines is probably a consequence of the experiment being based on only one year's data. The experiment began in May 1997 – the first forecast verified was that based on 20 May 1997 data; the most recent forecast verified was that based on 19 May 1998 data. The declines might be expected to become smoother as the numbers of data increase.

The data depicted in Figs 3 to 10 show that:

(a) overall forecast performance, as measured by the 'joint skill score' (Fig. 10) declines rapidly from day 1, the skill displayed falling to a level that is no better than climatology at day 5;

(b) in regard to the individual components of the 'joint skill score', the skill displayed by the 'QPF skill score' (Fig. 5) and the 'Brier skill score' (Fig. 7) both fall to a level that is no better than climatology at day 4;

(c) the skill displayed by the 'minimum temperature skill score' (Fig. 3) and the 'R/NR skill score' (Fig. 6) both fall to a level that is no better than climatology at day 5:

(d) by contrast, when compared to climatology, it is not until day 7 that skill displayed by the 'maximum temperature skill score' (Fig. 4) falls to a level that is no better than climatology (the apparent skill here could be due to a small number of correctly forecast extreme values);

(e) overall, forecast performance declines more slowly from day 5 to about day 9, at which point it is not only substantially inferior to climatology, but is also close to that performance to be expected from a randomly generated forecast; and,

(f) overall forecast performance remains close to that performance to be expected from a randomly generated forecast from day 9 to day 14, although there is a slight increase in skill after day 9 (although not statistically significant, this trend appears to be 'real' and is explained in the next subsection).

Explanation

The superior performance of the maximum temperature forecasts (when compared with the forecasts of other elements) is reflected in it not being until day 7 during the warmer half of the year (October to March) that this set of forecasts display no skill (this is explained later). By contrast, during the cooler half of the year (April to September), the skill displayed falls to a level that is inferior to climatology at day 5. This is even though

Melbourne maximum temperature variability is less during the cooler half of the year (Stern 1996).

Interestingly, it is also not until day 7 during the warmer half of the year that the minimum temperature forecasts display no skill. However, the skill displayed by the minimum temperature forecasts on days 5 and 6 during the warmer half is slight. For this reason, the 'minimum temperature skill score' calculated over the complete dataset reaches a skill level that is no better than climatology at day 5.

That the overall forecast performance (Fig. 10) declines to a level that is no better than climatology at day 5 may be attributed to there being limited useful skill displayed by the associated NCEP ensemble prognoses. Although there are no verification data over the Australian region about the accuracy of the NCEP ensemble prognoses, there are such data collected (by the BoM's National Meteorological Centre (NMC)) about the accuracy of a number of other deterministic prognoses that are routinely available to Australian forecasters. These data suggest that the ECMWF MSLP prognosis (base time 10 pm Australian Eastern Standard Time the night before) is the most accurate of those prognoses that are available at the time of issue of the official late afternoon Melbourne forecast for the next four days. This prognosis recorded an average anomaly correlation (AC) over the Australian region during 1997 of 0.585 at 144 hours (the 144-hour output would need to be used for a forecast valid on the fifth day).

The work of Hollingsworth et al. (1980), Murphy and Epstein (1989) and Wilks (1995) all suggests that an anomaly correlation of 0.6 may be employed as a lower cut-off for useful forecast skill. This suggests that even the ECMWF MSLP forecast would be considered to display insufficient skill to justify its use for a weather prediction beyond four days. This is notwithstanding Wilks (1995) having reported higher (than 0.6) ACs for 500 hPa prognoses given that 'Melbourne's weather is highly correlated with the surface circulation' (Stern, 1980a).

The decline in overall forecast performance to a level that is inferior to climatology is attributed to the forecasts based on climatology not including errors as large as the forecasts based on the full range of possibilities, as would a set of randomly generated forecasts.

The slowly declining overall forecast performance between day 5, when it reaches a level no better than climatology, to day 9, when it reaches a level close to the level of a randomly generated forecast, deserves explanation. This is attributed to there being some residual skill present in forecasts for some elements and in some synoptic situations during the day 5 to 9 component of the forecast period. That there is this skill is supported by the data presented in Fig. 13.

Fig. 13 A comparison of the R/NR skill scores for high-confidence forecasts (diamonds) and low-confidence forecasts (squares). High-confidence forecasts are those with PoPs of 10-20% or 80-90%. Low-confidence forecasts are those with PoPs of 30-70%.

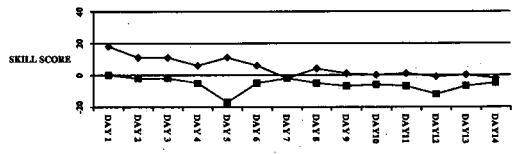


Figure 13 shows that, for example, when there is confidence about whether or not precipitation is likely, either:

(a) on the part of the official forecaster (up to day 4); or

(b) due to a strong signal from the NCEP MSLP, 500 hPa, and 1000-500 hPa ensemble averages (from day 5);

then – this confidence is justified.

Figure 13 also shows that, when there is little confidence about whether or not precipitation is likely, this reduced confidence is justified by estimates that are no better than climatology even at day 1 (it might be expected that forecasts with no confidence should possess the same skill as climatology).

The slight improvement in overall forecast performance from about day 9 to day 14 is attributed to the NCEP ensemble average prognosis tending towards climatology at the end of the forecast period, probably as a result of there being less energy through cascading.

That most of the precipitation-related skill scores decline to a level inferior to climatology, sooner than do the temperature-related skill scores, also deserves explanation. The relatively more rapid decline of the precipitation-related skill scores might be due to the combined effect of:

(a) temperature, particularly maximum temperature during summer, being highly correlated with features of the broadscale flow;

(b) useful information about the broadscale flow able to be readily derived from the prognoses for a relatively long period ahead;

(c) difficulty in timing onset and duration of precipitation at longer lead times;

(d) the occurrence of precipitation being not as highly correlated (as is temperature) with features of the broadscale flow; and (as a consequence)

(e) useful information about the complex physical processes associated with both broadscale and convective precipitation able to be derived from the prognoses for only a relatively short period ahead.

That the less rapid decline in the temperature skill scores are not observed in the winter data might be due to winter temperature, particularly winter maximum temperature, not being as highly correlated with features of the broadscale flow as is temperature in the other seasons. This is on account of the greater daytime solar insolation in summer leading to mixing of the low-level air to a greater depth. Another factor might be the reduced inter-diurnal temperature variability in the winter, resulting in a reduced opportunity to demonstrate skill.

The somewhat negative picture painted by the overall performance might be attributed to the experimental forecasts being based on a subjective interpretation of only one set of numerical prognoses. During parts of the period that the experimental NCEP-based forecasts have been prepared, forecasts of temperature have also been prepared by five other sources. These sources are the ECMWF GASM guidance, the GASP GASM guidance, the experimental VRO RFC forecasts, independent predictions by a 'group' hereafter referred to as Forecaster (1), and independent predictions by a 'group' hereafter referred to as Forecaster (2). Table 3 depicts the percentage improvement that these sources achieved over corresponding experimental NCEP-based predictions. Verification of these other temperature forecasts can be expected to shed some light upon whether or not the NCEP-based forecasts under-estimate the level of potential skill.

Table 3 suggests that all five sources were inferior to the NCEP-based performance at predicting maximum temperature. However, application of two-tail tests indicates that in no case is the inferiority significant at the five per cent level.

However, Table 3 suggests that only one of the sources was inferior to the NCEP-based performance at predicting minimum temperature and that the superiority of the ECMWF and GASP guidance over the NCEP-based forecasts is significant at the five per cent level.

In summary, Table 3 indicates that the performance of the NCEP-based maximum temperature forecasts is probably a reasonable measure of what can be achieved in predicting that element. However, Table 3 tells us that the performance of the NCEP-based minimum temperature forecasts may be an under-estimate of what can be achieved in predicting that element.

Some implications

Traders in agricultural commodities regularly utilise longer term day-to-day predictions in their work. For example, the 22 June 1992 *Wall Street Journal* reported that 'the possible development of a high pressure ridge', depicted in the 10th day of the US National Weather Service's (NWS) model, sparked 'renewed fears of a drought in the central Midwest (and) drove grain futures prices higher at the Chicago Board of Trade'. If the data presented for Melbourne are regarded as representative of other locations, this practice cannot be justified. Indeed, the NWS model prediction, referred to earlier, proved to be incorrect and the *Wall Street Journal* of 3 July 1992 reported that 'heavy rain ... helped alleviate short-term drought fears and drove grain futures prices lower at the Chicago Board of Trade'.

Table 3. Ratio of temperature variance not explained by NCEP-based predictions, over that not explained by ECMWF GASM guidance, GASP GASM guidance, VRO RFC forecasts, independent predictions by Forecaster (1) and independent predictions by Forecaster (2). Values below 1.0 suggest that the source is inferior to that of the NCEP-based predictions. Values above 1.0 suggest that the source is superior to that of the NCEP-based predictions. Significant values are indicated with an asterisk.

Source	*Days*	*Element*	*Number*	*Ratio*
ECMWF	Day 5	Max	61	0.90
GASP	Days 5, 6 & 7	Max	132	0.91
VRO	Days 5, 6, & 7	Max	45	0.79
Forecaster (1)	Days 5 & 6	Max	270	0.86
Forecaster (2)	Days 5, 6, & 7	Max	199	0.89
ECMWF	Day 5	Min	61	1.98*
GASP	Days 5 & 6	Min	131	1.74*
VRO	Days 5, 6 & 7	Min	45	1.77
Forecaster (1)	Days 5 & 6	Min	270	0.84
Forecaster (2)	Days 5, 6 & 7	Min	N/A	N/A

It may be asserted, however, that even when the NCEP-based forecasts for the individual 10 days between days 5 and 14 display no skill, they nevertheless may indicate the overall weather conditions during that 10-day period. In order to assess the validity of this assertion, each of the forecast sets was evaluated in terms of whether or not a correct indication was given of whether or not the number of rain days during the period was more than the climatological normal. Sixty-one per cent of the forecast sets gave such a correct indication.

Employing a one-tail test on this proportion, and regarding each of the ten-day forecast sets as independent from each other, suggests the statement that 'the proportion is greater than 50%' is significant at the 0.2 per cent level. However, the forecast sets are not completely independent from each other – they partially overlap. Taking into account this overlapping, the level of significance reduces to 13 per cent. Given that overlapping implies only partial dependence, on account of the NCEP prognoses having different initial conditions, the true level of significance is between 0.2 per cent and 13 per cent – an encouraging outcome, but one that is not conclusive.

Conclusion

Findings

A rigorous forecast verification methodology has been described, which may be used to guide decision-makers about the validity (or otherwise) of providing (or utilising) longer-term day-to-day weather predictions.

The verification data derived using the methodology suggest that, at present:

(a) routinely providing or utilising day-to-day forecasts beyond day 4 would be inappropriate; but,

(b) it might be possible to provide some useful information about the likely weather up to about six days in advance for some elements, in some seasons and in some situations (notwithstanding that the level of skill displayed at these longer lead times is limited). For example, electricity companies, wishing to plan for the heavy loadings that are caused by excess use of air conditioners during heatwaves in summer, would benefit from temperature forecasts out to six days during that season. By contrast, in some circumstances it may not be possible to provide useful information even for day 1; and,

(c) it might also be possible to make useful statements about the expected average weather conditions over the ten-day period between days 5 and 14.

Future work

The conclusions presented here are largely based on only one year's data for one place. It is therefore planned to continue the experiment, to extend it to other places because local features, such as distance from the ocean, may influence the level of potential forecast skill associated with long lead-time predictions.

Finally, it is also planned to explore the use of MOS as a tool to provide longer term day-to-day predictions. A MOS-based forecast would tend towards climatology as forecast skill decreases to zero. However, it also would provide a 'best-guess' of the extent to which a forecaster could depart from climatology, where potential forecast skill does exist.

Acknowledgments

I thank reviewers Terry Skinner, Mark Williams, Clem Davis, Ian Mason, an anonymous reviewer, and *Australian Meteorological Magazine* Associate Editor, Ian Simmonds, for their helpful comments on the paper. I also thank my colleagues at the BoM for their valuable suggestions regarding this work, in particular, Ken Dickinson, Geoff Feren, Noel Fitt, Mary Voice and Richard Whitaker. I also thank Wilma Skinner of NMC for providing me with the anomaly correlation verification data.

References

Bureau of Meteorology 1977. *Recording and encoding weather observations*. Bur. Met., Australia.

Brier, G.W. 1950. Verification of forecasts expressed in terms of probabilities. *Mon. Weath. Rev., 78*, 1-3.

Brooks, H. 1995. Human forecasters and technology From http://www.nssl.uoknor.edu/~brooksl (available January 1998).

Climate Diagnostics Center 1997. CDC Map Room Weather Products. From http://www.cdc.noaa.gov/~map/maproom/ENS/ens_desc.html (available January 1998).

Dahni, R.R., de la Lande, J. and Stern, H. 1984. Testing of an operational statistical forecast guidance system. *Aust. Met. Mag., 32*, 105-6.

Dahni, R. R. 1988. The development of an operational analogue statistics model to produce weather forecast guidance. Ph. D. Thesis, Department of Meteorology, University of Melbourne.

Dahni, R. R. and Stern, H. 1995. The development of a generalised UNIX version of the Victorian Regional Office's operational analogue statistics model. *BMRC Res. Rep. 47*, Bur. Met., Australia.

Doswell, C.A. 3rd 1995. Meteorology and users. From http://www.nssl.uoknor.edu/~doswell/ (available January 1998).

Hollingsworth, A., Arpe, K., Tiedtke, M., Capaldo, M. and Savijärvi, H. 1980. The performance of a medium range forecast model in winter. *Mon. Weath. Rev., 108*, 1737-73.

Lorenz, E.N. 1963. Deterministic, non-periodic flow. *J. Atmos. Sci., 20*, 130-41.

Lorenz, E.N. 1969a. Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci., 26*, 636-46.

Lorenz, E.N. 1969b. The predictability of a flow which possesses many scales of motion. *Tellus, 21*, 289-307.

Lorenz, E.N. 1993. *The essence of chaos.* University of Washington Press.

Murphy, A.H. and Epstein E S. 1989. Skill scores and correlation coefficients in model verification. *Mon. Weath. Rev., 117*, 573-81.

Murphy, A.H. 1991. Forecast verification: its complexity and dimensionality. *Mon. Weath. Rev., 119*, 1590-1601.

Stern, H. 1980a. The development of an automated system of forecasting guidance using analogue retrieval techniques. M. Sc. Thesis, Department of Meteorology, University of Melbourne (published in 1985 as *Meteorological Study 35*, Bur. Met., Australia).

Stern, H. 1980b. A system for automated forecasting guidance. *Aust. Met. Mag., 28*, 141-54.

Stern, H. 1996. Statistically based weather forecast guidance. Ph. D. Thesis, School of Earth Sciences, University of Melbourne (published in 1999 as a *Meteorological Study 43*, Bur. Met., Australia).

Stern, H. 1998. An experiment to establish the limits of our predictive capability. *14th Conf. on Probability and Statistics / 16th Conf. on Weather Forecasting and Analysis*, Amer. Meteor. Soc., Phoenix, Arizona, 11-16 January, 1998.

Toth, Z. and Kalnay, E. 1993. Ensemble forecasting at NMC: The generation of perturbations. *Bull. Am. met. Soc., 74*, 2317-30

Tracton, M.S. and Kalnay, E. 1993. Ensemble forecasting at NMC: Operation implementation. *Weath. forecasting, 8*, 379-98.

Wilks, D. S. 1995. *Statistical methods in the atmospheric sciences. An introduction.* Academic Press, California, 468 pp.

Woodcock, F. 1984. Australian experimental model output statistics forecasts. *Mon. Weath. Rev., 112*, 2112-21.