Forecasting the Skill of a Regional Numerical Weather Prediction Model

L. M. LESLIE, K. FRAEDRICH[+] AND T. J. GLOWACKI

Bureau of Meteorology Research Centre, Melbourne, Australia

(Manuscript received 8 June 1988, in final form 6 September 1988)

ABSTRACT

It is demonstrated that the skill of short-term regional numerical forecasts can be predicted on a day-to-day basis. This was achieved by using a statistical regression scheme with the model forecast errors (MFE) as the predictands and the initial analysis, together with the model forecast, at proximate points, as the predictors.

In a first attempt to assess the utility of the method, the technique was applied in a long-term quasi-operational trial to 24 h forecasts of mean sea level pressure in two seasonal periods (one summer and one winter period) on the Australian region forecast domain. Correlation coefficients were computed between the predicted and observed root-mean-square (rms) MFE and were found to be 0.54 and 0.51, respectively, averaged over the full region, for the 90-day summer and winter periods. Using standard Student's *t*-tests these correlations were shown to be highly significant. In addition, the regional forecasts were divided into four categories of rms MFE, and were verified against the observed rms MFE. Using a contingency table skill score (relative to chance), it was demonstrated that the category forecasts exhibited a very high level of skill.

The procedure also was applied to subdomains of the Australian region grid and it was found that the predictions of model forecast skill were improved further for these local forecasts.

1. Introduction

In addition to providing a numerical forecast, it is desirable that a numerical weather prediction (NWP) system also provides a prediction of the skill of that forecast on a routine basis. Various methods for predicting the skill of numerical forecasts have been developed and these have taken two main forms, *statistical–dynamical* and *ensemble forecast* predictions of skill.

The statistical–dynamical approach recognizes that there are uncertainties in the specification of the initial data and in the model itself, and attempts to recast the purely deterministic analysis-forecast problem in probabilistic terms (Epstein 1969). The stochastic–dynamical models therefore provide an explicit, a priori, estimate of the model forecast skill because they predict the forecast mean and variance. However, the statistical–dynamical methods have not been considered for routine use because the amount of computation they require is still beyond current computer technology, other than for the simplest models (Hoffman and Kalnay 1983).

An alternative approach to statistical–dynamical forecasting is the ensemble forecast method in which a weighted average is made of an ensemble of forecasts and predictions of the skill of the forecast is estimated a priori from the dispersion of the ensemble. The specification of the initial conditions from which the forecast ensemble is obtained has been made in a number of ways. Leith (1974) proposed a Monte Carlo technique in which the ensemble of initial conditions was obtained by a random perturbation to the objective analysis, and suggested that a set of about eight ensemble members should be sufficient. However, the requirement of providing multiple forecasts also is computationally unattractive in an operational context. Hoffman and Kalnay (1983), Kalnay and Dalcher (1987) and Dalcher et al. (1988) presented an alternative ensemble forecast method in which the members of the ensemble are forecasts based on previous initial analysis times. Because these forecasts are already available, this so-called lagged average forecasting (LAF) technique has almost no computational cost. Furthermore, Hoffman and Kalnay (1983) showed that they could predict forecast skill, from the ensemble dispersion, by obtaining good predictions of the timing of forecast breakdown. It should be noted that predictive methods other than ensemble forecast techniques also have been established. For example, Branstator (1985) and Palmer (1988) have demonstrated that predictability is closely related to sea-surface temperature anomalies and the Pacific/North American (PNA) index, respectively.

All of the attempts to predict forecast skill described above were concerned with medium and extended range forecasts. Hitherto, little effort has been made to

[+] Permanent affiliation: Institut fur Meteorologie, Freie Universitat Berlin, Berlin, West Germany.

Corresponding author address: Dr. Lance M. Leslie, BMRC, Bureau of Meteorology, GPO Box 1289K, Melbourne, Victoria, 3001, Australia.

develop schemes for predicting the skill of short-term (24–48 h) forecasts. Prediction of forecast skill in the medium and extended range are concerned with such problems as the limits of predictability and forecast breakdown. Apart from a relatively small number of rapidly changing mesoscale systems (that have large error doubling growth rates), predictions of short-term forecast skill are largely unconcerned with predictability limits. The emphasis is much more on the precise predictions of model errors.

The application of statistical techniques to short-term numerical forecasts generally has been limited to developing statistical regression relationships between the dynamical model output and the desired predictand, such as precipitation or temperature. The model output statistics (MOS) procedure (Glahn and Lowry 1972) is, perhaps, the best known example. However, some statistical procedures have been developed for short-term NWP models in which a priori estimates of model forecast error (MFE) are calculated. These are the *statistical correction* schemes, of which these are two main kinds, those that correct and reinitialize the numerical model (see, for example, Schemm and Faller 1986), and those that simply correct the output from a dynamical model at various time intervals (see Bennett and Leslie 1981).

It is a variation of the method of Bennett and Leslie (1981), as modified by Glowacki (1988), that will be used as the basis for this study. Bennett and Leslie predicted the MFE of the Australian region operational NWP model forecasts of mean sea level pressure by a statistical regression of the MFE against the first eight terms in the empirical orthogonal function (EOF) expansions of both the initial analysis and the model forecasts (detrended in time). Thus, the number of predictors was reduced to 16 for each grid point on the Australian region grid at which the MFE of mean sea level pressure was to be estimated a priori. The predicted MFE was then simply added to the model forecast as a statistical correction. Glowacki (1988) improved upon the Bennett and Leslie technique by replacing the EOF expansions with values of the initial analysis and forecast at a small number (about 10) of proximate grid points because, as he demonstrated, for forecasts out to 36 h the MFE is uncorrelated with predictors at distant locations. The Australian Bureau of Meteorology has been using this method operationally, twice-daily, since 2 December 1987 for correcting numerical forecasts of mean sea level pressure. The impact on the skill level of the forecasts has been substantial, with record low rms errors and S_1 skill scores being registered for each of the 5 months since the scheme was implemented. It is planned to use the method for fields other than mean sea level pressure, in the near future.

It is the aim of this study to show that the predictions of MFE from the statistical correction scheme of Glowacki (1988) can be used on a day-to-day basis in an operational system to provide accurate and reliable a priori estimates of the skill of the model forecast. This is achieved by correlating the predicted rms MFE with the actual (or observed) MFE and showing that these correlations are very significant. Moreover, the predictions of MFE also will be divided into four categories according to the size of the rms MFE and, using a contingency table skill score, these category predictions will be tested for skill relative to chance. Finally, the predicted rms MFE field will be divided into four regional subdomains and the predicted model forecast skill assessed for these "local" forecasts.

It is important to point out that the forecasts of skill presented below apply to the uncorrected model. Clearly, there can be no further information about the skill of the model after the correction has been applied. In order to make a forecast of the skill of the corrected model, an independent technique is required. Work is currently being undertaken on this problem.

2. Methodology

The technique used in this study is, first, to calculate the predicted MFE from the statistical regression scheme developed by Glowacki (1988). The scheme provides an estimate of the rms MFE and this estimate can then be used as a measure of the skill of the forecast.

a. Statistical prediction of the MFE

The statistical prediction of the MFE in the technique devised by Glowacki (1988) uses multiple linear regression methods, with the MFE as predictands, and the initial analysis together with the model forecast (both detrended) as predictors. Predictions of the MFE are made at the nodes of a 104 point (13 × 8) subgrid of the Australian region analysis-forecast domain, as shown in Figs. 1(a) and (b). This subgrid is then divided into overlapping subregions [see Fig. 1(c)] such that a prediction within a given subregion is regressed against all analysis and forecast predictors that lie within the subregion. The subdivision into overlapping subregions enables the number of predictors to be reduced dramatically, especially as those predictors which were located at large distances from the predictand were shown by Glowacki (1988) to be uncorrelated with the predictand.

At a gridpoint, j, of the 104-point Australian region subgrid, the MFE, e_j, is estimated (predicted) by Glowacki (1988) to be

$$e_j = b_j + \sum_{s=1}^{k} e_{j,s} R^t_{j,s}, \tag{1}$$

where

$$e_{j,s} = C_0 + \sum_{i=1}^{n} C_i P_i{}^a + \sum_{i=n+1}^{2n} C_i P_i{}^m. \tag{2}$$

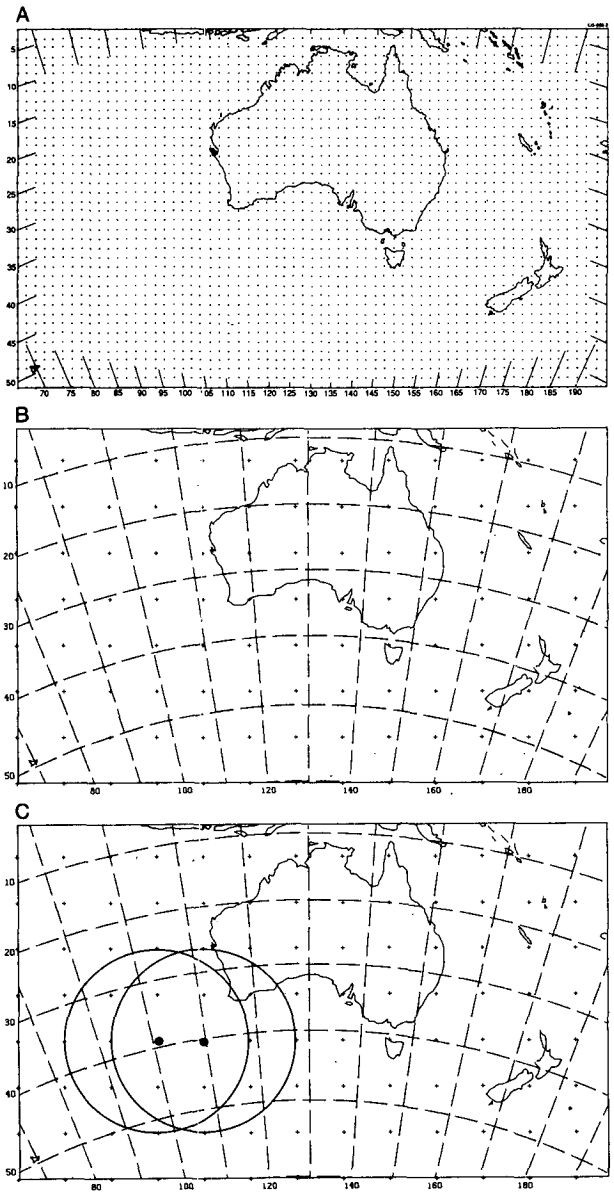

FIG. 1. (a) The 65 × 40 point Australian region NWP grid of resolution 150 km. (b) The 104 point (13 × 8) subgrid, on which the predictions of MFE are made. (c) The overlapping subregions used for the multiple linear regressions of the predictors contained within a circle, against the predictand at the center of the circle.

In (1) and (2), b_j is the bias, $e_{j,s}$ is the estimated MFE at the point j using predictors from subregion s, $R_{j,s}$ is the multiple correlation coefficient, k is the number of subregions that include point j, t is a weighting power factor, n is the number of gridpoints within subregions s, P_i^a and P_i^m are the analysis and model predictors, and C_i are the regression coefficients (for each j and s). The values of C_i are calculated for each month and obtained from 5 years (1980–1984) of archived operational numerical analyses and forecasts.

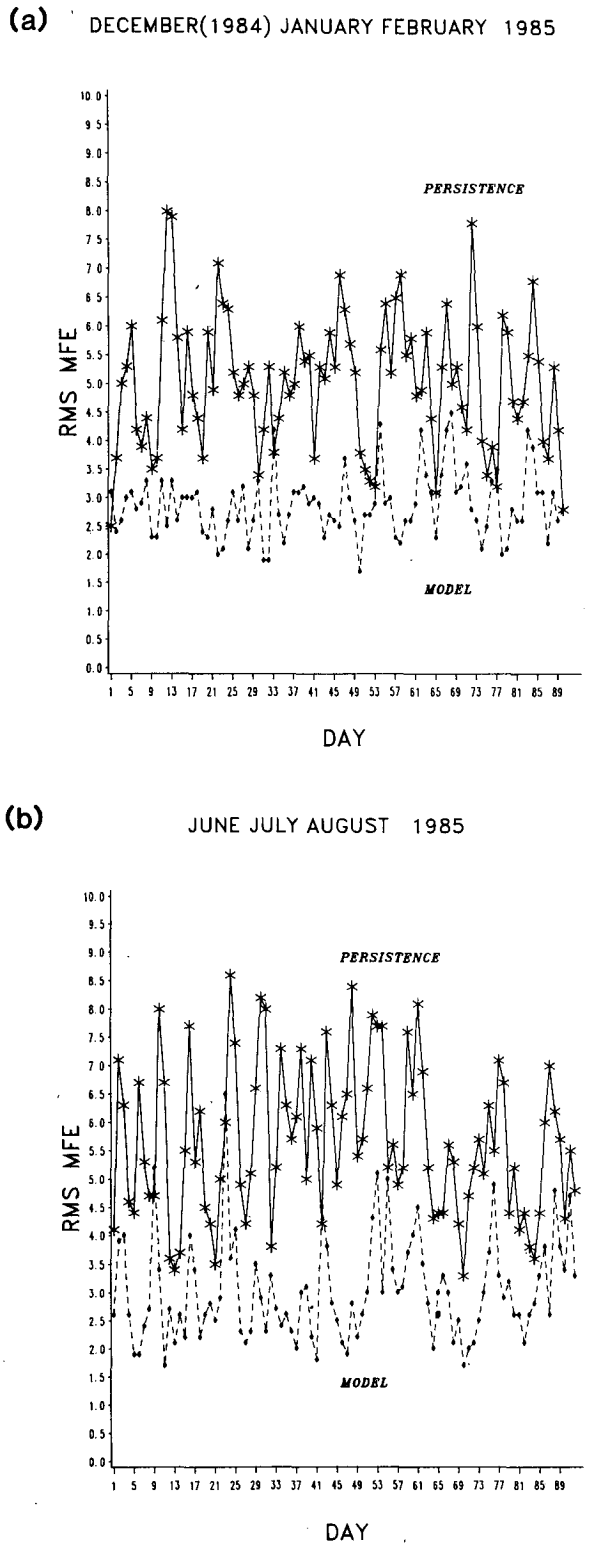

FIG. 2. (a) Actual rms 24-h model forecast errors (dashed line) for the Australian Bureau of Meteorology operational regional model for the summer period (December–February) 1984/85. For reference purposes the rms errors of persistence forecasts also are shown (heavy line). (b) As in (a) but for winter (June–August) 1985.

b. Prediction of forecast skill

The measure of skill of a model forecast used in this study will be the root-mean-square (rms) MFE. Along with the S_1 skill score this is a standard measure of skill employed by the Australian Bureau of Meteorology for the verification of its numerical forecast products. Figures 2(a) and (b) show the daily values of the actual rms MFE from the operational 24 h Australian region NWP system for the summer of 1984/85 and the winter of 1985, respectively. For reference purposes, rms errors from persistence forecasts also are shown. It is worth noting that for short-term NWP model forecasts, persistence clearly is no longer a suitable reference relative to which skill can be defined, as its accuracy is far worse than that of the NWP model.

1) CORRELATIONS BETWEEN PREDICTED AND OBSERVED MFE

Having obtained a prediction of the MFE from Eq. (1), the skill of these predictions must be assessed. This is achieved by correlating the predicted and observed rms MFE.

2) SIGNIFICANCE OF CORRELATIONS

The correlations obtained between the predicted and observed rms MFE must be tested for significance before the predicted rms MFE can be used as an a priori measure of forecast skill. If the correlations are highly significant, then the predicted rms MFE can be used directly to predict the forecast quality. If the correlations are not significant, either for the entire forecast domain or for subsets of it, the predicted rms MFE will have no practical value over those areas.

Tests of significance of the correlations between predicted and actual rms MFE were carried out at each grid point of the 104-point subgrid using the null hypothesis that the correlation coefficient is zero. The Student's t-test was employed, using the t-statistic

$$t = \frac{r\sqrt{(N-2)}}{\sqrt{(1-r^2)}}, \tag{3}$$

where r is the correlation coefficient and N the length of the time series (in this case, the number of days in the summer or winter sample).

3) CATEGORY PREDICTIONS

In addition to assessing the skill of the predictions of MFE it was decided to divide the total range of the MFE into a number of subintervals. The predictions of rms MFE can then be represented in a contingency table and the accuracy of the forecasts assessed from a skill score based on frequencies relative to chance. Four categories of rms MFE were chosen, namely, <2.5, 2.5 to 2.9, 3.0 to 3.4, and ≥3.4 mb. For the Australian region forecast domain, these intervals correspond approximately to forecasts rated very good, good, mediocre, and poor to very poor. The skill score used is that of Panofsky and Brier (1958) and is defined by

$$S = \frac{C-E}{T-E}, \tag{4}$$

where C is the number of correct forecasts by chance, E the expected number of correct forecasts, and T the total number.

It is noted that the category predictions also provide a measure of the bias in the predicted rms MFE. As the correlation coefficients are insensitive to bias, the

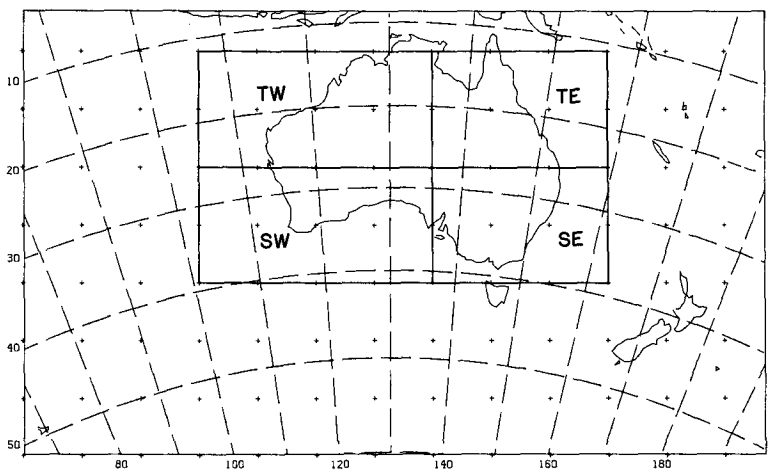

FIG. 3. The local subregions for which local predictions of forecast still will be made. The subregions are denoted TW (tropics west), TE (tropics east), SW (southwest), and SE (southeast).

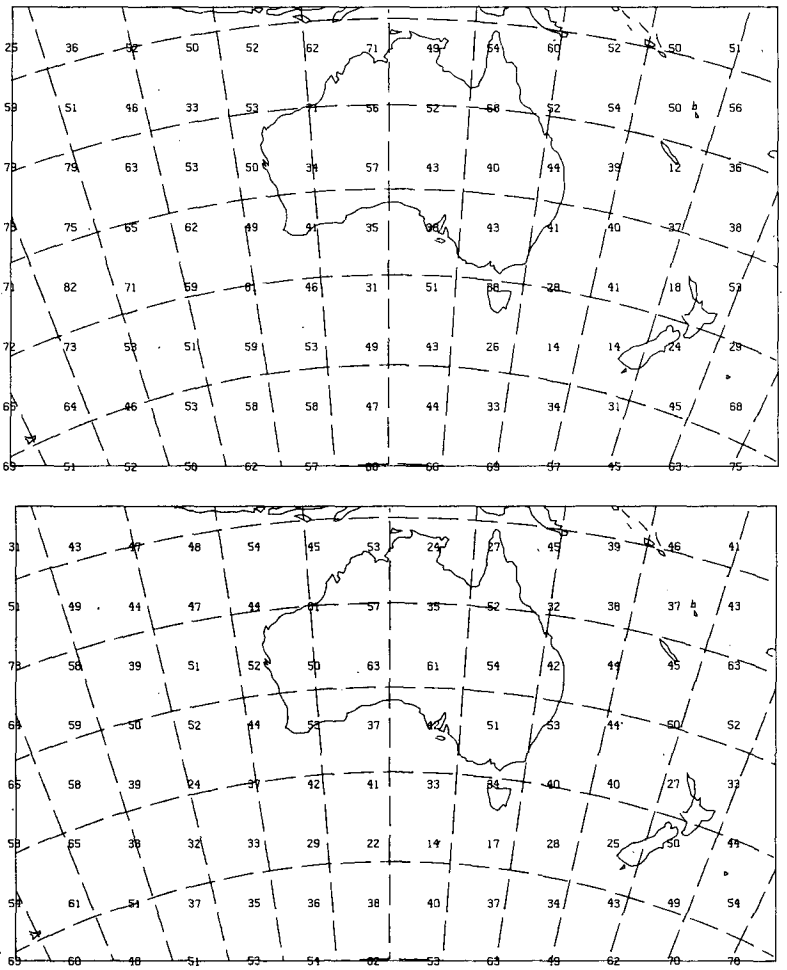

FIG. 4. (a) Correlation coefficients between predicted and actual rms 24-h model forecast errors for summer 1984/85. (b) As in (a), but for winter 1985.

category predictions add to the assessment of the skill of the predicted rms MFE.

4) TOTAL REGION VERSUS LOCAL MODEL FORECAST SKILL

Currently, the skill of the operational Australian region NWP model is assessed over (almost) the entire forecast domain. However, it would be useful to have predictions of forecast skill not only for the total region but also for local subsets of the total region. This would provide forecasters with valuable guidance concerning the quality of the forecasts in specific localities.

Therefore it was decided to evaluate the predictions of forecast skill not only for the full 104 point subgrid but also for the four subregions shown in Fig. 3. These subregions were labeled tropics west (TW), tropics east (TE), southwest (SW), and southeast (SE). Predictions of forecast skill were assessed for all four areas, in section 3c.

3. Results

The method described in section 2 was applied to two seasonal three month periods, December 1984–February 1985 (summer) and June 1985–August 1985 (winter). The periods were chosen to be 90 days in extent.

a. Correlations between predicted and observed rms MFE

The predicted rms MFE were calculated at each grid point of the 104-point subgrid for each day of the two seasonal periods. Correlation coefficients were then calculated at each grid point and are shown in Figs. 4a and 4b. The average correlation coefficients over the entire grid were found to be 0.54 and 0.51 for summer and winter, respectively.

The correlation coefficients were assessed for statistical significance using the Student's *t*-test, with the *t*-

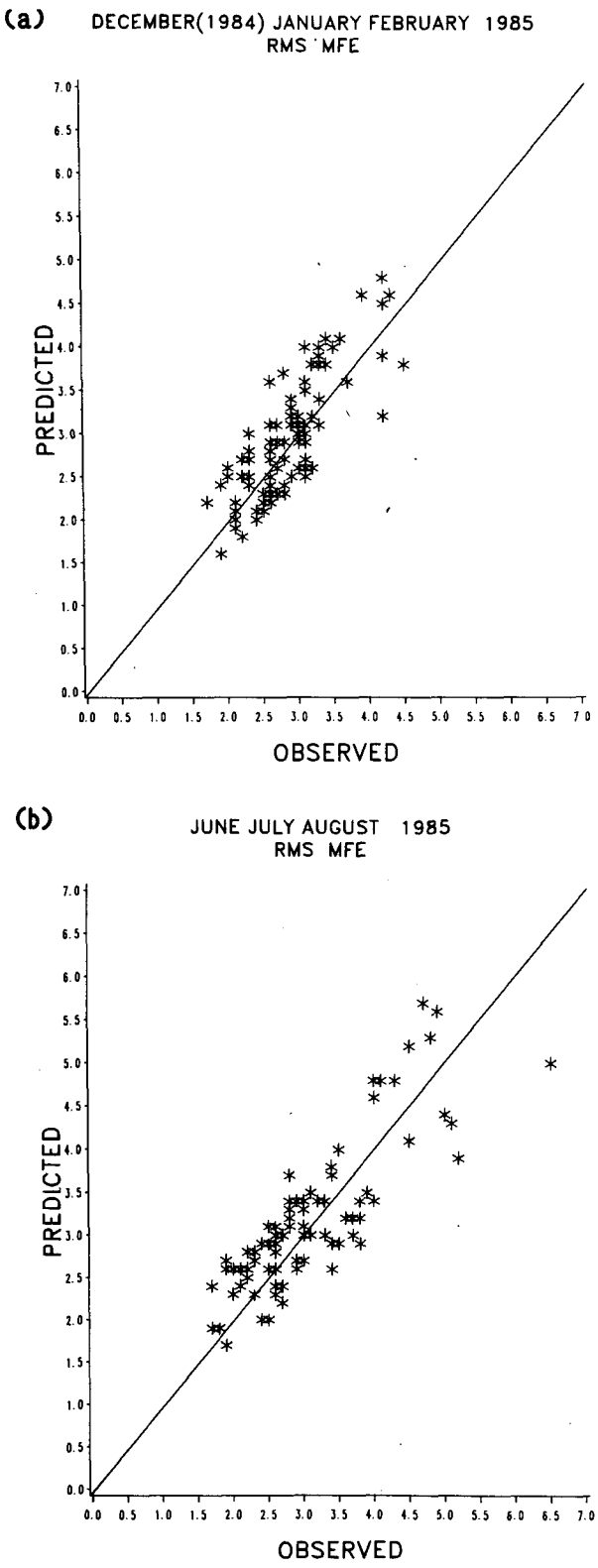

FIG. 5. (a) Scatter diagram for predicted and actual rms 24-h model forecast errors, for summer 1984/85. (b) As in (a), but for winter 1985.

statistic defined in Eq. (3). The null hypothesis that the correlation coefficient between predicted and observed rms MFE is zero was rejected at the 99% confidence level, provided $r > 0.20$, and at the 99.9% level if $r > 0.28$. An inspection of the correlation coefficients in Figs. 3a and 3b shows that at most of the 104 grid points, for both summer and winter, the correlations are highly significant. In particular, the correlation coefficients of 0.54 and 0.51, respectively, averaged over the full 104-point subgrid are very highly significant for both summer and winter and confirm that the predicted rms MFE is closely related to the observed rms MFE and therefore provides a direct a priori measure of forecast skill.

A scatter diagram was prepared, to illustrate further the high level of correlation between the predicted and observed skill of the forecast model. In Figs. 5a and 5b the predicted and observed rms MFE are plotted as the y-coordinate and x-coordinate, respectively, for both the summer and winter periods, for the entire forecast domain of 104 points. The high correlation is seen clearly in the small amount of scatter around the straight line $y = x$. It also appears that there is very little bias in the predictions as they are quite evenly spread on either side of the $y = x$ line.

b. Category forecasts

The assessment of the predicted rms MFE in terms of the correlations between predicted and observed rms MFE forms only part of the verification. Correlation coefficients are insensitive to bias and therefore it is important to have a quantitative measure of the skill of the predicted rms MFE that includes bias. This is achieved by dividing the predicted rms MFE into four categories according to the magnitude of the rms error. The categories chosen were <2.5, 2.5–2.9, 3.0–3.4, and ≥3.5 mb. These intervals correspond approximately to the assessment categories of very good, good, moderate, and poor to very poor.

The category forecasts were carried out for both the summer and winter periods and the results are shown in the histograms of Figs. 6a and 6b respectively. The histograms show that the numbers and distribution of predicted rms MFE appears to match with the observed rms MFE to a high degree. The degree of correspondence can be assessed more accurately by casting the predictions into contingency tables, as shown in Tables 1 and 2, and calculating the skill score defined by Eq. (5). The expected number of correct predictions, E, based on chance occurrence is given by

$$E = \sum_{i=1}^{4} P_i O_i / T, \qquad (5)$$

where P_i and O_i are the subtotal number of predicted and observed values for each category.

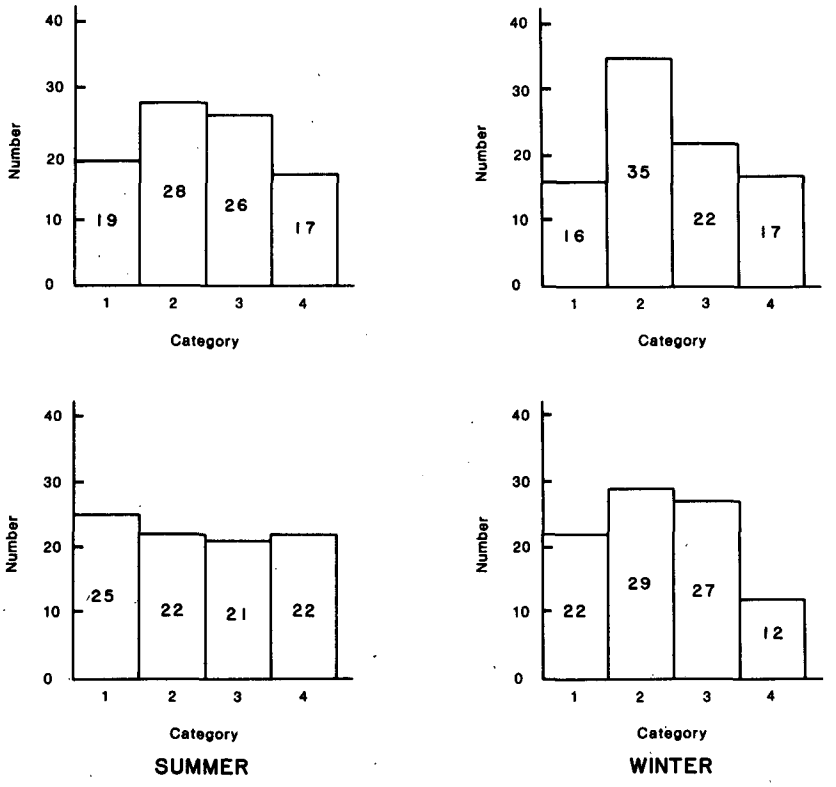

FIG. 6. (a) Histogram showing numbers of predicted and observed rms 24-h model forecast errors in the four categories <2.5, 2.5–2.9, 3.0–3.4, and ≥3.5 mb for summer 1984/85. (b) As in (a), but for winter 1985.

For the summer of 1984/85 the skill score as calculated by substituting from Table 1 into Eqs. (6) and (5) was found to be 0.27. The corresponding value for winter, obtained from Table 2 was 0.25. Both these skill score values are high. A chi-squared test of the null hypothesis that the predicted values were obtained by chance was rejected with 99.9% confidence, and demonstrates that the predictions of MFE into categories are very skillful.

c. *Predictions of local forecast skill*

In addition to predicting model forecast skill for the entire forecast domain, it is very useful to have available forecasts of model skill on a local basis. The skill of these local forecasts can depend on a number of factors. They may be significant systematic errors that result from a consistent failure to forecast, for example, the exact location of an orographically induced flow pattern. Alternatively, the error may be related to a particular weather system occurring on a given day. Regardless of the source of the error, it would be of considerable benefit to weather forecasters if they had a measure of the quality of the forecast for a given locality.

As described in section 2d, four subsets of the Australian region forecast domain were chosen in an initial experiment to examine variations in the regional prediction of model forecast skill. These subsets were denoted by TW (tropics west), TE (tropics east), SW (southwest), and SE (southeast). For each of these four local regions, correlations were made between the daily predicted and observed values of the rms MFE. The correlation coefficients are shown in Table 3 and each was shown to be very highly significant, using the *t*-test. It is seen from Table 3 that in each of the four local regions, the correlation coefficients are larger than

TABLE 1. Contingency table for category predictions of rms model forecast error (MFE), for summer 1984/85.

Observed rms MFE (mb)	Predicted rms MFE (mb)				
	<2.5	2.5–2.9	3.0–3.4	≥3.5	Total
<2.5	13	9	3	0	25
2.5–2.9	5	11	5	1	22
3.0–3.4	1	6	10	4	21
≥3.5	0	2	8	12	22
Total	19	28	26	17	90

for the entire Australian region domain. In fact, for the TE and SE areas the correlation coefficients are >0.55, suggesting that the predictions of model forecast skill for these regions would be very useful.

No attempt was made in this study to perform category forecasts over the local regions. Of course, the categories chosen should be dependent on factors such as the latitude of the local region and the season. For example, a local region in the tropics during winter would have much lower rms model forecast errors than a local region in the midlatitudes. In the near future, extensive trials will be carried out in a real-time assessment of the method described in this article for predicting model forecast skill over the entire Australian region and selected local regions. The regions selected for assessing the performance on a local basis have yet to be chosen and will be made only after consultation with the major regional forecasting centers around Australia.

4. Summary and conclusions

It has been shown that it is possible to predict the forecast skill of a regional NWP model, on a day-to-day basis, by the use of a statistical correction scheme developed by Glowacki (1988). This scheme uses multiple linear regression to predict the model forecast error (MFE) from neighboring values of the model initial analysis and the forecast itself.

One of the measures of forecast skill adopted for verification of the Australian region 24-h numerical forecasts is the rms error. By demonstrating that the predicted rms MFE of two seasonal sets of 90 consecutive 24-h forecasts are highly correlated (with correlation coefficients of 0.54 and 0.51, respectively) with the observed MFE, it is shown that the predicted rms MFE can be used directly as a prediction of model forecast skill.

The utility of the approach was underlined further by an assessment of the ability of the scheme to predicted categories of model forecast skill. Four categories of rms MFE were defined and contingency table skill scores were calculated, using a skill score of Panofsky and Brier (1958). These contingency table skill scores, which are measured relative to chance, were found to be very highly significant.

TABLE 2. As in Table 1, but for winter 1985.

Observed rms MFE (mb)	Predicted rms MFE (mb)				
	<2.5	2.5–2.9	3.0–3.4	≥3.5	Total
<2.5	10	5	1	0	16
2.5–2.9	9	17	8	1	35
3.0–3.4	3	6	10	3	22
≥3.5	0	1	8	8	17
Total	22	29	27	12	90

TABLE 3. Correlation coefficients between predicted and observed rms model forecast error (MFE) for the TW, TE, SW and SE regions shown in Fig. 2.

Region	Correlation coefficients	
	Summer	Winter
TW	0.549	0.542
TE	0.553	0.571
SW	0.550	0.521
SE	0.568	0.553

Finally, when the predictions of model forecast skill are made for local subsets of the Australian region, the correlations between predicted and observed rms MFE are increased. This indicates that for local forecasts, the prediction of model forecast skill should prove to be even more useful than for the entire Australian region forecast domain.

The scheme described in this paper will be tested in an operational trial sometime in the near future, over the entire Australian region, and for selected subregions.

Acknowledgments. The authors are grateful to Dr. N. Nicholls of BMRC for advice on certain statistical aspects of this paper and to Dr. M. J. Manton, also of BMRC for helpful comments on the manuscript. We also wish to thank Lucie Crivera for typing the manuscript.

REFERENCES

Bennett, A. F., and L. M. Leslie, 1981: Statistical correction of the Australian region primitive equation model. *Mon. Wea. Rev.*, **109**, 453–462.

Branstator, G., 1985: Analysis of general circulation model sea-surface temperature anomalies using a linear model. Parts I and II. *J. Atmos. Sci.*, **42**, 2225–2254.

Dalcher, A., E. Kalnay and R. N. Hoffman, 1988: Medium range lagged average forecasts. *Mon. Wea. Rev.*, **116**, 402–416.

Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus*, **21**, 739–759.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211.

Glowacki, T., 1988: Statistical corrections to dynamical model predictions. *Mon. Wea. Rev.*, **116**, 2614–2627.

Hoffman, R. N., and E. Kalnay, 1983: Lagged average forecasting. *Tellus*, **35A**, 100–118.

Kalnay, E., and A. Dalcher, 1987: Forecasting forecast skill. *Mon. Wea. Rev.*, **115**, 349–356.

Leith, C., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418.

Palmer, T. N., 1988: Medium and extended range predictability and stability of the Pacific North American mode. *Quart. J. Roy. Meteor. Soc.*, **114**, 691–714.

Panofsky, H. A., and G. W. Brier, 1958: *Some Applications of Statistics to Meteorology.* Pennsylvania State University, 224 pp.

Schemm, J-K. E., and A. J. Faller, 1986: Statistical corrections to numerical predictions. *Mon. Wea. Rev.*, **114**, 2402–2417.