

# A methodology for numerical prognosis evaluation using objective local weather retrieval

F. Woodcock\* and N. Nicholls†

\*Head Office, Bureau of Meteorology, Melbourne

†Australian Numerical Meteorology Research Centre, Melbourne

(Manuscript received July 1981; revised September 1981)

**A general method of evaluating numerical model prognoses in terms of local weather forecasts using the perfect prognosis approach is described. The main advantages of the method are that clean impact studies are feasible and only a short prognosis archive period is required. Information is given concerning which prognosis fields need most improvement to benefit local forecasts. Using a novel skill score to relate the objectively derived weather forecasts to the operational ones, the methodology is applied, as an example, to the operational Australian region primitive equation low-level prognoses' impact on Sydney, January daily maximum temperature forecasts. The prognosis fields are shown to require improvement particularly in the low-level temperature fields before they can provide operationally useful forecasts of Sydney, January maximum temperatures.**

## Introduction

Operational local weather forecasts usually contain a large subjective component and are therefore unsuitable for clean impact studies. Clearly an objective link between broadscale prognoses and local weather is required. The two procedures widely used in this link are the perfect prognosis (PP; e.g. Klein 1963) and model output statistics (MOS; Glahn and Lowry 1972). United States evidence suggests that MOS yields more accurate operational forecasts than PP (e.g. Klein and Hammons 1975). MOS, however, is not particularly useful in diagnostic or impact studies because the forecasts are optimised for a particular model. Any variation to the model, if it affects the forecasts at all, should deteriorate them unless the MOS equations are recomputed, which can only be done with a long series of the altered model broadscale numerical predictions.

PP forecasts in contrast are eminently suitable for impact studies. The PP procedure statistically relates broadscale analysis fields to near concurrent weather and in an operational mode the model prognosed predictors required in the PP equations are used to forecast the local weather. The strength of the PP method for impact studies is because the closer the model broadscale prognoses match the verifying analyses in the elements required to produce local weather forecasts, the better the local weather forecasts will be. Further, since the PP

equations are derived from analysis data, there is no requirement for a long period of prognosis records to evaluate the impact of model change.

The potential of PP forecasts in model diagnosis does not appear to have been exploited in the literature. For example, Klein and Marshall (1973) in a report on the testing of improved PP predictors for the Shuman and Hovermale (1968) United States six-layer primitive equation model noted that the local weather forecasts in a simulated operational test were worse than expected. They suggested that the poor performance could probably be attributed to defects in the model prognostic data, particularly the low-level temperature and dew-point depression prognosed predictors. If the diagnostic value of the PP forecasts had been fully exploited, this hypothesis could have been tested.

Apart from describing a method which enables model impact on local weather forecasts to be ascertained, this paper uses a novel skill score which enables PP forecasts to be compared easily with the corresponding operational manual forecasts.

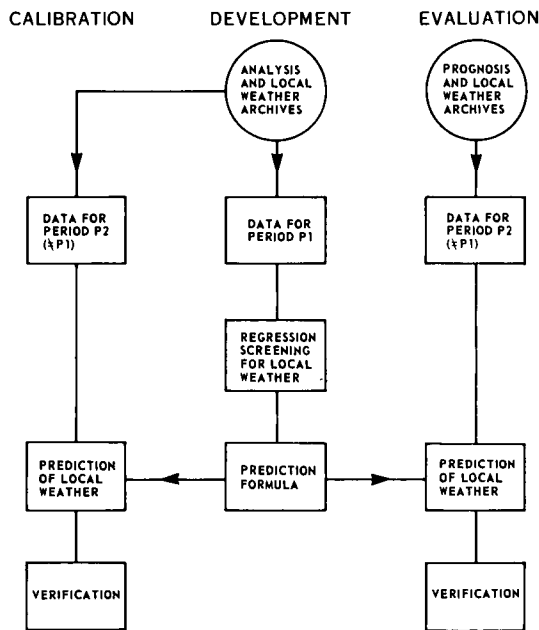
## Strategy

A flow diagram of the strategy is shown in Fig. 1. There are three modes in the diagram: development, calibration and evaluation. In the developmental mode a series of broadscale analyses and near

concurrent weather is screened to derive a regression formula relating the analysis information as predictors to the local weather predictand. In the calibration mode the formula is applied to the corresponding predictors from a separate period of *analyses* and corresponding local weather. The errors in the local forecasts during this calibration period are those which perfect prognoses would achieve with that formula. In the evaluation mode the same formula is applied to the corresponding *prognosed* predictors supplied by the model for the same period used in calibration. The resulting errors can be considered as comprising both a perfect and an imperfect prognosis component. Hence the difference between the calibration and evaluation errors can be attributed to imperfections in the prognoses.

One advantage of this strategy is its relative insensitivity to inhomogeneity in the developmental data.

**Fig. 1** Flow diagram of strategy for using perfect prognosis predictions of local weather in model evaluation.



**Application**

The Australian Region Primitive Equation model (ARPE; McGregor, Leslie and Gauntlett 1978) is known to produce poor temperature and moisture prognoses over the continent in the low levels, particularly in summer. The reasons for these are briefly:

- (i) no surface heating scheme combined with inadequate vertical resolution in the low levels (Leslie 1980);
- (ii) inadequate horizontal spatial resolution (Leslie, Mills and Gauntlett 1981; Gauntlett and Leslie 1981);

- (iii) inadequate horizontal and vertical resolution of topography along the east coast of Australia (Gauntlett 1981; Gauntlett and Leslie 1981);
- (iv) poor initial data due to the data-void ocean areas (Fitt, Whitby and Brown 1979; Leslie, Mills and Gauntlett 1981); and
- (v) prognostic temperature and moisture fields at 1000 mb are extrapolated from the 0.7 and 0.9 sigma surfaces.

The archived summer ARPE prognoses therefore provide an ideal data set to demonstrate the diagnostic value of PP forecasts.

Following the strategy outlined above, the particular application in this demonstration uses the real-time archived analyses and near concurrent observed Sydney December, January and February daily maximum temperatures from the period December 1971 to February 1978 in the development mode. The PP regression formula is applied to 62 January 1979 and 1980 analyses to estimate the corresponding Sydney daily maximum temperatures in the calibration mode and applied to the corresponding 62 prognoses in the evaluation mode. In relating the error characteristics from the calibration and evaluation modes to each other it has been found useful to develop a skill score based on the operational errors of forecasting of 62 Sydney daily January maximum temperatures.

**Skill score**

Operational daily January maximum temperature forecast errors are readily available for the period January 1970 to January 1980 inclusive (341 days). The forecasts were issued at 1600 EST the previous day. For the purpose of this study, error statistics for a 62-day sequence are required. In order to gain information on the likely distribution of them over a 62-day sequence, the error characteristics of all the 341 possible different sequences of 62 were obtained. Distribution parameters of these errors are shown in Table 1. There is a wide range evident in the distribution, for example, the smallest number of errors equal to or greater than 4°C was one forecast out of 62 (1.61 per cent in Table 1), while the greatest was eighteen.

The skill score was devised by ranking the mean absolute error, mean square error, percentage of errors equal or exceeding 4°C and the correlation coefficient between observed and forecast maxima separately from one (the worst) to 341.

Each set of 62-day forecasts has therefore four ranks associated with it. These ranks were then averaged and the 341 set of forecast sequences reranked (R) from worst to best. The skill (S) as a percentage is then defined as

$$S = 100R/341 \dots 1$$

**Table 1** Distribution statistics of characteristic errors for 341 sequent runs of 62 January days of Sydney maximum temperature forecasts issued by the Bureau of Meteorology.

<i>Error characteristic</i>	<i>Smallest value</i>	<i>Mean and 95 per cent confidence interval</i>		<i>Standard deviation</i>	<i>Median</i>	<i>Largest value</i>
Bias (°C)	-0.78	-0.65	-0.05	0.48	0.3	-0.08
Mean absolute error (°C)	1.13	1.23	1.80	2.26	0.3	1.84
Mean squared error (°C <sup>2</sup> )	2.58	2.86	6.50	10.16	2.15	6.63
Per cent of absolute errors $\geq 4^\circ\text{C}$	1.61	3.23	11.73	20.97	4.93	12.10
Correlation between observed and forecast maxima	0.592	0.597	0.698	0.815	0.058	0.689
			0.815	0.058	0.689	0.835

A forecast set as good as the best operational set would score 100 per cent and one as good as the median would score 50 per cent etc. The operational forecasts over the January 1979, January 1980 sequence scored 57 per cent.

## Data

For the development of the regression formula, the analysis information was obtained from the Australian region real-time 2300 GMT analysis grid point archives for the months December, January and February from December 1971 to February 1978 inclusive. From this data set of 632 days, 11 were missing. Sydney daily maximum temperature data for the same period were obtained from separate Bureau archives.

For calibration of the regression formula using independent data, the analyses and maximum temperatures were obtained for January 1979 and January 1980.

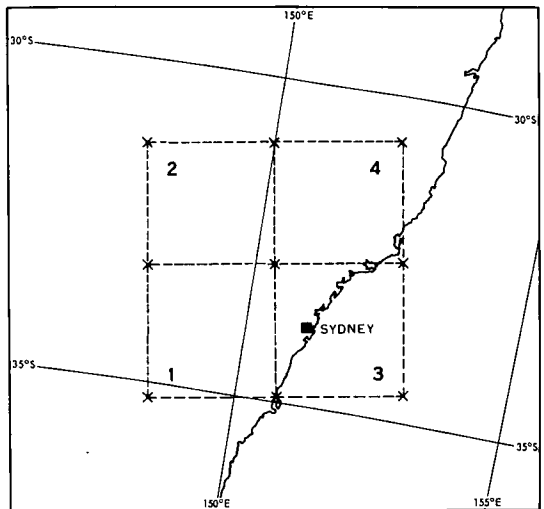
In the evaluation mode it was necessary to rerun (using operational parameters) the ARPE model for January 1979 and January 1980 in order to obtain the required prognosis data.

Since the regression equation was to predict Sydney, January daily maximum temperatures from information valid approximately six hours earlier, only grid point data close to Sydney was offered to the stepwise screening procedure. Consequently the mean sea-level pressure, 1000 mb and 850 mb temperature and dew-point and 850 mb geopotential height information was extracted for each of the nine grid points indicated by crosses in Fig. 2.

To further reduce the number of potential predictors offered to the stepwise screening, the nine grid point values were reduced to four by averaging the grid point values at the four corners of each of the four smaller squares (labelled 1 to 4 in Fig. 2). There is therefore a strong spatial coherence between these potential predictors.

No correction for serial correlations in the data was attempted, but the auto-correlation (one-day lag) for two Januaries of Sydney maximum temperatures was 0.06 and so the reduction in the effective number of degrees of freedom in the data

due to temporal correlations (e.g. Davis 1976) is unlikely to be serious.

**Fig. 2** Grid used in the extraction of data.

## Development of regression formula

To derive the regression formula the following potential predictors, from analyses for each of the 621 days in the developmental sample, were offered to the stepwise screening procedure: (1) eastward component of the mean sea-level geostrophic wind; (2) northward component of the mean sea-level geostrophic wind; (3) mean sea-level pressure (MSL); (4) approximate mean sea-level geostrophic wind convergence (the first four predictors were computed from the combined values of the four squares in Fig. 2); (5, 6, 7, 8) 1000 mb temperatures for each of the four squares in numerical order respectively; (9-24) corresponding 1000 mb dew-points, 850 mb geopotential heights, temperatures and dew-points respectively. These, together with the same-day maximum temperature, were subjected to Efroymsen's (1960) forward stepwise screening procedure using Miller's (1962) F test as a stopping criterion.

**Table 2** Details of the stepwise screening output. The numbers in brackets in the predictor column indicate the square in Fig. 2 from which the predictor was chosen.

Predictor	Per cent of maximum temperature variance explained	Regression coefficient (units omitted)	Simple correlation coefficient
Mean sea level eastward component of geostrophic wind	9.3	0.45	0.71
1000 mb temperature (2)	0.4	-0.24	0.59
1000 mb temperature (3)	60.8	0.96	0.78
850 mb geopotential height (2)	1.1	0.01	-0.03
850 mb geopotential height (4)	0.2	-0.01	0.02
850 mb temperature (4)	0.5	0.20	0.65
850 mb dew-point (3)	0.3	-0.27	0.07
850 mb dew-point (4)	0.3	0.13	0.01

Table 2, which shows some details from the screening, indicates that the eastward component of the MSL geostrophic wind and the 1000 mb temperature in square three (Fig. 2) are the most important contributors to the prediction of the Sydney maximum temperature. Their selection and function are as expected, the 1000 mb temperature chosen is the nearest one to Sydney; the hotter it is at 2300 GMT, the higher the expected maximum, hence the positive regression coefficient is expected.

The positive regression coefficient for the eastward component of the mean sea-level geostrophic wind can also be justified. An eastward component of the near-surface wind, corresponding to offshore wind, would delay or prevent the onset of a cooling sea-breeze, enhance downslope subsidence warming and advect drier and hotter continental air over Sydney (thus reducing the chance of cloud). Conversely, onshore winds corresponding to a negative eastward component of the geostrophic wind, would be cool and moist due to their over-water trajectory.

### Calibration of regression formula

To establish the calibration error characteristics of the regression formula it was applied to the January 1979 and January 1980 analyses and maximum temperatures, this period being outside the developmental data period. Table 3 shows these error characteristics in the column headed 'Forecasts from independent analyses' (referred to as column 2).

Some perspective can be placed on these errors by comparing them with those where the regression formula is applied to the data from which it has been derived (Table 3, column 1) and the Bureau's subjectively produced forecasts (column 4). The regression forecasts show a small deterioration when independent rather than dependent data are used. The bias, mean absolute error, mean square error and percentage of absolute errors at least exceeding four degrees have all increased slightly, but the correlation between the observed and forecast temperature has remained similar. These independent six-hour statistical forecasts are better than the 24-hour subjective forecasts according to

**Table 3** Error characteristics of the different Sydney daily maximum temperature forecasts. The dependent analyses column refers to the December, January and February months from December 1971 to February 1978 inclusive, the other three columns refer to January 1979 and January 1980 only.

Mode	Forecasts from dependent analyses	Forecasts from independent analyses	Forecasts from independent prognoses	Subjective forecasts
	Development	Calibration	Evaluation	
Number of forecasts	621	62	62	62
Bias (°C)	0.0	0.1	-0.7	-0.2
Mean absolute error (°C)	1.4	1.6	4.9	1.8
Mean square error (°C <sup>2</sup> )	3.4	4.4	24.4	6.4
Per cent of absolute errors > 4°C	3.4	6.5	41.9	8.1
Correlation between forecast and observed maxima	0.9	0.9	0.5	0.8
Skill (per cent)	92	74	0	57
Approximate lead time (hrs)	6	6	30	24

the error characteristics in columns two and four. A *t*-test for difference between these and the Bureau's forecasts being greater than zero was significant at the 95 per cent level for both absolute and mean square error. The Bureau's skill for these forecasts from Eqn 1 was 56.5 per cent while that for the statistical forecast was 73.6 per cent. These statistics give an indication of the accuracy that 30-hour forecasts of Sydney temperature maxima could achieve if the low-level prognosis data were perfect.

## Evaluation of broadscale prognoses

The broadscale prognosis data relevant to the Sydney maximum temperature were evaluated by applying the regression equation to the prognoses corresponding to the analysis data used in the calibration, namely January 1979 and January 1980. If the prognosis had been perfect then the analysis data would be reproduced and the Sydney maximum temperature forecasts would be identical to the calibration set. Departure of the prognoses from the corresponding analyses should result in a deterioration in the forecast maxima. The extent of the deterioration induced by replacing the analyses with prognoses in this case can be seen by comparing column 3 in Table 3 with column 2. All the error characteristics indicate seriously degraded forecasts. These forecasts using prognosis data are so inaccurate as to be operationally useless as the comparison with the manual forecasts (column 4) shows.

This disappointing result warrants further investigation. It is possible to identify the causes of the poor forecasts by applying the regression formula to the errors in the prognosis data and recording the contribution of each term to the final difference in forecasts. When the absolute error in maximum temperature for each term is partitioned in this way Table 4 is obtained, from which it follows that the 1000 mb temperature in square three (Fig. 2) is the main source of error from this set of predictors. The contribution to error arises because of the poor prognosis of the term and the

magnitude of the coefficient associated with it (see Table 2).

In Fig. 3, two cases in 1979 are given in which there was a large discrepancy between the maximum temperature forecast from prognosis data compared to that forecast from analysis data. On 17 January the observed maximum was 29°C; the forecast from analysis data using the regression equation was 29°C and from the prognosis data 37°C. The 1000 mb temperature over the area was forecast to be far hotter than occurred. On 28 January the observed maximum temperature was 24°C, the analysis-based forecast maximum was 23°C and the prognosis-based forecast 17°C. The prognosis 1000 mb temperature field exhibits a 5°C centre over the Tasman Sea and temperatures near Sydney of 8°C whereas the corresponding analysis field has temperatures near Sydney closer to 10°C warmer.

In an attempt to improve the quality of the Sydney maximum temperature forecasts obtainable from prognosis data in the PP mode, the preceding experiment was repeated using subsets of the data. The second experiment, with all three modes, had all the 1000 mb potential predictors withheld from the regression screening, while the third experiment offered only MSL potential predictors to the screening.

The consequence of these two further experiments was that the accuracy of the forecasts in the calibration mode decreased as fewer potential predictors were offered, but in the evaluation mode the forecasts improved slightly. For example, Table 5 shows the error characteristics when only MSL pressure information is used in the regression equation. Comparison of Tables 3 and 5 shows the marked degradation of the accuracy of forecasts in the development and calibration modes and the improvement in the evaluation mode.

## Discussion of results

The application of PP forecasts in the diagnostic study of the ARPE model in relation to its impact on Sydney, January daily maximum temperatures

**Table 4** Frequency distribution of the absolute temperature error contributed by errors in the prognosis terms in the regression equation.

Predictor	0	1	2	3	4	5	6	7	8	9	10
Mean sea level eastward component of geostrophic wind	18	25	17	1	1	0	0	0	0	0	0
1000 mb temperature (2)	21	35	5	1	0	0	0	0	0	0	0
1000 mb temperature (3)	6	9	12	5	11	9	2	5	2	0	1
850 mb geopotential height (2)	33	28	1	0	0	0	0	0	0	0	0
850 mb geopotential height (4)	41	21	0	0	0	0	0	0	0	0	0
850 mb temperature (4)	40	22	0	0	0	0	0	0	0	0	0
850 mb dew-point (3)	23	32	4	1	0	2	0	0	0	0	0
850 mb dew-point (4)	38	22	1	1	0	0	0	0	0	0	0

Fig. 3 Two 1000 mb temperature analysis and prognosis sets corresponding to poor evaluation mode forecasts.

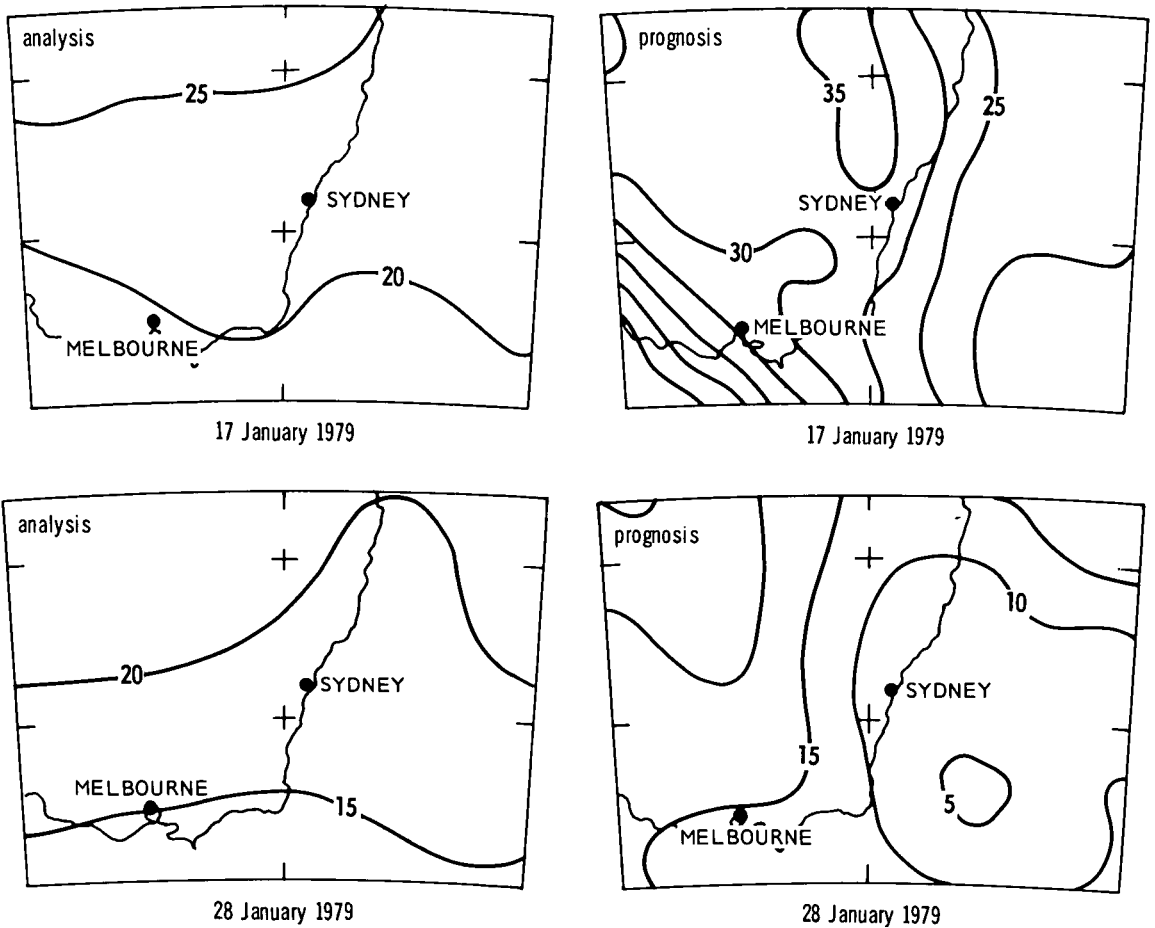


Table 5 Error characteristics of the Sydney daily maxima objective forecasts using MSL predictors only.

	<i>Forecasts from dependent analyses</i>	<i>Forecasts from independent analyses</i>	<i>Forecasts from independent prognoses</i>
Number of forecasts	621	62	62
Bias (°C)	0.0	-0.4	0.3
Mean absolute error (°C)	1.8	1.7	2.3
Mean square error (°C <sup>2</sup> )	6.1	6.2	8.5
Per cent of absolute errors $\geq 4^\circ\text{C}$	13.8	11.3	16.1
Correlation between forecast and observed maxima	0.7	0.8	0.7
Skill (per cent)	48	59	13
Approximate lead time (hrs)	6	6	30

has unambiguously shown that the verifying analysis fields contain very useful predictive information that the prognostic fields are failing to capture. The most influential predictors are the 1000 mb or 850 mb temperatures and these are poorly forecast by the model (at 850 mb) or derived from the model fields (1000 mb).

An idea of the improvement to the operational forecasts which could be gained from independent forecasts using perfect prognoses can be seen from Table 6. Here the skill scores of the calibration mode in several experiments are shown separately and also with the average of them and the operational forecasts (consensus forecasts). An inference from these results is that perfect prognoses in conjunction with statistics are capable of substantially improving the operational forecasts (skill = 57 per cent) and some effort aimed at improving the low-level prognosis data is likely to be beneficial.

### Concluding remarks

It has been shown that statistical screening of broadscale numerical fields and local weather in a PP mode can provide a powerful method in impact studies. Its main advantages are that it is objective and thereby permits clean impact studies; it does not require a long history of prognoses to be applied; it relates the prognoses to the end result of the forecasting procedure, the local weather forecast, and the verifying data are available and directly measured.

The results of the specific application addressed in the example given show which elements of the broadscale prognoses require most attention for a particular local forecast element and yield sample estimates of the best likely improvement that perfect prognoses can give.

### Acknowledgments

The authors thank B. Shanahan, J. Brown, G. Mills and J. McGregor for their useful comments, W. Kininmonth and D. Gauntlett for their valuable reviews of an earlier draft of this paper, M. Navin for drafting and M. Smedley for typing the script.

### References

- Davis, R. E. 1976. Predictability of sea surface temperature and sea level pressure anomalies over the north Pacific ocean. *J. Phys. Oceanogr.*, **6**, 249-66.
- Efroymsen, M. A. 1960. *Multiple regression analysis. Mathematical Methods for Digital Computers*, 1st edn, Ralston, A. and Wilf, H. S. eds, John Wiley & Sons Inc., New York, 293 pp.
- Fitt, R. N., Whitby, F. and Brown, J. 1979. The impact of FGGE data on prognosis in the Australian region. *Proceedings Australian-New Zealand Global Atmospheric Research Programme Symposium, Melbourne, Australia*, 59-62.
- Gauntlett, D. J. 1981. The numerical simulation of intense frontal discontinuities over south eastern Australia. *Preprint International Association of Meteorological and Atmospheric Physics Nowcasting Symposium, Hamburg, 17-28 August*.
- Gauntlett, D. J. and Leslie, L. M. 1981. Numerical experiments in meso-scale analysis and prediction. Paper to be submitted to *Monthly Weather Review*.
- Glahn, H. R. and Lowry, D. A. 1972. The use of model output statistics (MOS) in objective weather forecasting. *Jnl appl. Met.*, **11**, 1203-11.
- Klein, W. H. 1963. Specification of precipitation from the 700 mb circulation. *Mon. Weath. Rev.*, **91**, 527-36.
- Klein, W. H. and Hammons, G. A. 1975. Maximum/minimum temperature forecasts based on model output statistics. *Mon. Weath. Rev.*, **103**, 796-806.
- Klein, W. H. and Marshall, F. 1973. Screening improved predictors for automated max/min temperature forecasting. *Preprints Third Conference on Probability and Statistics in the Atmospheric Sciences, Boston, American Meteorological Society*, 36-43.
- Leslie, L. M. 1980. Numerical modelling of the summer heat low over Australia. *Jnl appl. Met.*, **19**, 381-7.
- Leslie, L. M., Mills, G. A. and Gauntlett, D. J. 1981. The impact of FGGE data coverage and improved numerical techniques in numerical weather prediction in the Australian region. *Q. Jl R. met. Soc.*, **107**, 629-42.
- McGregor, J. L., Leslie, L. M. and Gauntlett, D. J. 1978. The ANMRC limited area model: consolidated formulation and operational results. *Mon. Weath. Rev.*, **106**, 427-38.
- Miller, R. G. 1962. Statistical prediction by discriminant analysis. *Met. Monogr.*, No. 25, American Meteorological Society, 45-7.
- Shuman, F. G. and Hovermale, J. B. 1968. An operational six layer primitive equation model. *Jnl appl. Met.*, **7**, 525-47.

