The Performance of Four Global Models over the Australian Region

L. M. LESLIE, G. D. HESS, AND E. E. HABJAN

Bureau of Meteorology Research Centre, Melbourne, Victoria, Australia

(Manuscript received 11 March 1993, in final form 6 January 1994)

ABSTRACT

National weather services now receive global model forecasts from a number of centers around the world. The existence of these forecasts raises the general question of how the operational forecaster can best use the information that the ensemble of predictions provides. The Australian Bureau of Meteorology receives four global model forecasts in real-time, but at present their performance is evaluated almost entirely in a subjective manner.

In this study, in addition to the standard objective measures (for example, bias and rms error), several alternative objective measures of model performance are calculated (such as the temporal forecast consistency of a given model and divergence between different models), in an attempt to provide the forecasters with more effective tools for model assessment. Both kinds of measures are applied to a two-year dataset (October 1989 to September 1991) of daily sea level pressure predictions from the four models.

There are two main outcomes of this study. First, the current subjective system of ranking the various models has been augmented with more objectively based performance measures. Second, these performance statistics provide guidance to the operational forecasters in a number of ways: geographical regions with large systematic errors can be identified for each model; case studies are presented that illustrate the utility of the regional maps of bias, consistency, and divergence computed in this study; and, finally, there are regions of uncertainty where no model is consistently superior, so forecasts over these regions should be treated with caution.

1. Introduction

Most of the major national weather centers receive medium-range global forecasts from a number of centers around the world in real-time by means of the Global Telecommunications Satellite. For instance, the Australian Bureau of Meteorology (ABM) receives real-time global forecasts from four centers: the European Centre for Medium-Range Weather Forecasts (ECMWF, Reading, England); the National Meteorological Center (Washington, D.C.); the UK Meteorological Office (Bracknell, England); and the local Australian global model (ABM, Melbourne, Australia). These models will be referred to as the EC, US, UK, and AG models, respectively. The availability of this ensemble of forecasts provides potentially important information to the operational forecaster, and the general question arises of how this information can be used best. There are several factors to be considered in using this ensemble of forecasts. For example, there is the question of timeliness. The forecasts from these models are received in Melbourne over the range of times shown in Table 1. The variation in the arrival times of these forecasts is an important factor in determining their usefulness. Also, as will be presented in detail herein, objective measures of the performance of the models from the centers can vary widely on a day-to-day basis, both individually and between each other. Finally, on a given day, forecasts from all the centers can vary widely in skill over various parts of the Australian region analysis/prognosis domain.

A brief summary of the configurations of the four models during the evaluation period is given in Table 2. The EC model remained static in terms of resolution during the period over which the dataset was compiled, whereas the UK and US models increased in resolution, as indicated in Table 1. (The EC model increased from T106 to T213 and 19 to 31 levels immediately following the completion of the dataset.)

Clearly, over any substantial period of time, changes in model configuration are likely, and are unavoidable in this kind of study. This factor stresses the need to continue the type of analysis presented here on a continuing basis. It should also be noted that the US model forecasts received in Melbourne extended only to 72 h (the early, so-called aviation forecast), whereas the other models extended to 120 h.

The factors mentioned above present operational forecast offices with difficulties in using the ensemble of model forecasts. An example in the Australian region is given in Fig. 1, which shows widely differing forecasts from global models whose analyses were almost identical. (This example will be discussed further.) As a consequence of this conflict between forecasts, the official forecast currently issued by the ABM is based

Corresponding author address: L. M. Leslie, Bureau of Meteorology Research Centre, P.O. Box 1289K, Melbourne, Victoria, Australia 3001.

TABLE 1. Daily arrival times in Melbourne of forecasts from base time 1200 UTC. Note that the EC model arrives in the early part of the new day, as indicated by the "+1." That is, it arrives 5 h after the local AG model.

Model	EC	UK	US	AG
Time (UTC)	0200 (+1)	2230	2230	2100

TABLE 2. Configurations of the four medium-range global models used in this study. Note that the UK model changed from $1.5^0 \times 1.875^0$, 15 levels in June 1991, and that the US model changed from T80, 18 levels in March 1991.

Model	EC	UK	US	AG
Type	Spectral	Grid point	Spectral	Spectral
Horizontal resolution	T 106	0.83×1.25	T 126	R 31
Levels	19	20	18	9
Forecast	5 days	5 days	3 days	5 days

largely on a subjective blend of the forecast ensemble. The relative weighting used essentially is a function of the confidence that individual duty forecasters have in the various models.

The present study is part of a larger program that has two steps: 1) to develop a methodology that minimizes the level of subjectivity by producing an extensive set of objective skill statistics for each model, and 2) eventually to develop a procedure for predicting the skill of the AG model forecasts over the Australian region by using the suite of forecasts available from the other global models. Here, the focus will be entirely on the first part of the research program, which, in turn, has two components: 1) the objective characterization of model performance through an enlarged range of statistical measures of the performance of the various models, and 2) their application in the operational forecasting process. They include the standard measures of skill, together with measures of other characteristics such as the temporal consistency of each of the models and the divergence between different models.

The latter group of measures has been used in studies aimed at predicting forecast skill for medium-range forecasts. This work has been pioneered largely by Kalnay and colleagues at the National Meteorological Center, Washington, D.C. (Hoffman and Kalnay 1983; Kalnay and Dalcher 1987; Toth and Kalnay 1993) and by Palmer and coworkers at the ECMWF (Palmer and Tibaldi 1989; Brankovic et al. 1990; Mureau et al. 1993). Studies carried out for the Australian region (Leslie et al. 1989; Leslie and Holland 1991) have thus far been preliminary and have focused only on short-

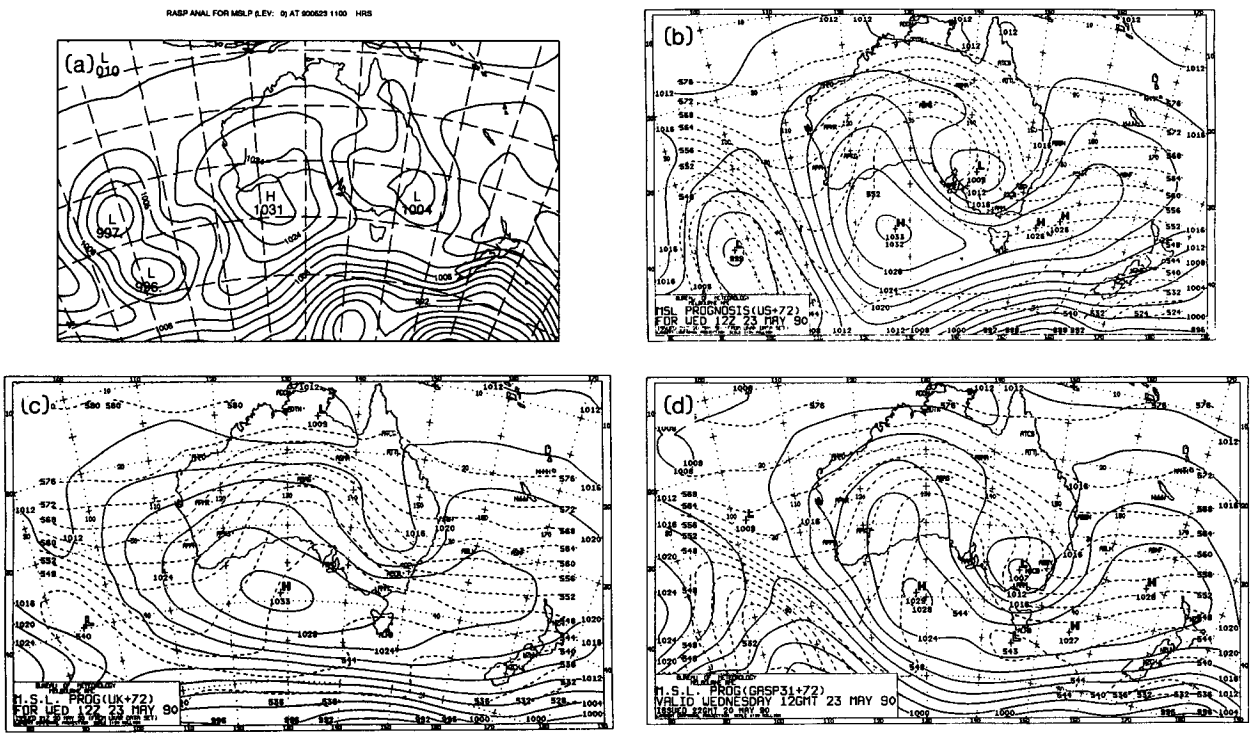


FIG. 1. An example of conflict in 72-h model predictions for 1200 UTC May 23 1990: (a) RASP verifying analysis, (b) US prediction, (c) UK prediction, and (d) AG prediction (also referred to as GASP).

range prediction. However, our attempt to apply these measures to the problem of the a priori estimation of model forecast errors will be considered in the next phase of this program.

To limit the paper to a reasonable length, the results presented here use one model only as the central member of the ensemble—namely, the EC model, as it has been the most skillful model in the Australian region during the decade or so prior to this study. This superiority is illustrated by the S_1 skill scores at 24 h over the period 1984–91, as shown in Fig. 2. The EC model (heavy line) stands out from the other models, although we note that the gap closes toward the end of the period, following upgrades of the US and UK model configurations.

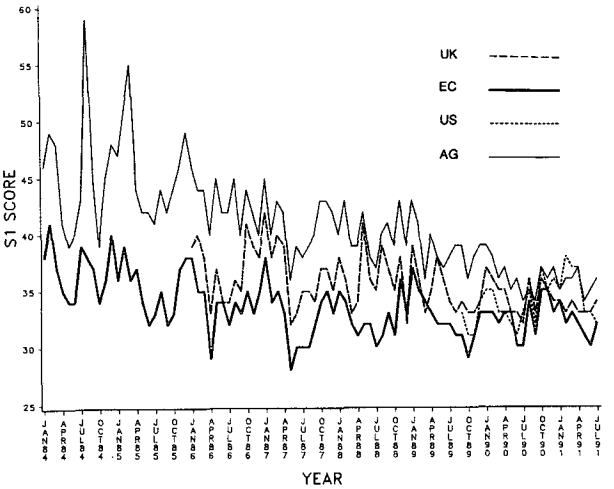


FIG. 2. The 24-h S_1 skill scores for the EC, UK, US, and AG global models, evaluated in the Australian region for 1984–91. Lower S_1 scores indicate greater skill.

2. Methodology

In this section the two groups of objective measures of model performance used in this study are described. The first group consists of the commonly used statistical quantities that characterize the performance of each model. The second group measures variations in individual model performance or in the behavior of one model relative to another model averaged over time and/or space. These measures are computed as values averaged over the whole domain, and also as gridpoint values to bring out the variations over the domain. The statistical measures were computed for all months of the two-year dataset of analyses and model forecasts (October 1989–September 1991). However, again for brevity, the focus in section 3 is on one month only—May, a transition season in Australia. We also note that, as was necessary for this study, there was close agreement between the domain-averaged initial global model analyses and the locally prepared Australian region analysis.

The performance statistics presented here concentrate on the sea level pressure (SLP) field, because in Australia it is by far the most widely used and disseminated of all the forecast charts, principally due to the poor upper-level network over Australia (particularly over the surrounding oceans), the associated dependence on satellite imagery interpretation, and the flatness of the land. This emphasis is changing, with improved data sources, analysis techniques, and model output.

a. Model performance statistics

The Australian limited-area analysis is obtained from the Australian region data assimilation system (RASP). The SLP field at 1200 UTC is used as the basis for the statistical calculations. The rms difference, correlation, bias, and S_1 skill score were chosen as the traditional measures of forecast model performance. The verification statistics such as correlation are not referred to as measures of skill, following the work of Murphy and Epstein (1989), who demonstrated that a correlation coefficient is not a measure of skill in the sense that it is not a measure of mean relative accuracy.

Some of the measures are defined differently depending upon whether they are domain-averaged plots or gridpoint plots. The domain-averaged measures are always positive:

$$\text{rms} = \{(1/N_tN_{xy}) \sum_t \sum_{xy} [P_M(D, T) - P_A(D, T)]^2 \cos^2\phi\}^{1/2}, \quad (1)$$

$$\text{cov} = (1/N_tN_{xy}) \sum_t \sum_{xy} P_M(D, T)P_A(D, T) \cos^2\phi - (1/N_t^2N_{xy}^2) \sum_t \sum_{xy} P_M(D, T) \times \cos\phi \sum_t \sum_{xy} P_A(D, T) \cos\phi, \quad (2)$$

$$\sigma_M = [(1/N_tN_{xy}) \sum_t \sum_{xy} P_M(D, T)P_M(D, T) \cos^2\phi - (1/N_t^2N_{xy}^2) \sum_t \sum_{xy} P_M(D, T) \times \cos\phi \sum_t \sum_{xy} P_M(D, T) \cos\phi]^{1/2}, \quad (3)$$

$$\text{cor} = \text{cov}/(\sigma_M\sigma_A), \quad (4)$$

$$\text{bias} = (1/N_tN_{xy}) \sum_t \sum_{xy} |[P_M(D, T) - P_A(D, T)] \cos\phi|, \quad (5)$$

$$S_1 = 100 \sum_{xy} |[P_M(D, T) - P_A(D, T)] \times \cos\phi| / \sum_{xy} |\max(d_0, d_f)|, \quad (6)$$

where the index D indicates the initial day, the index T indicates the forecast period in days, N_t is the total number of days, N_{xy} is the total number of grid points, and the subscripts M and A stand for "model" and "analysis," respectively. The sums are over time (for example, monthly averages) for the gridpoint calculations and over both time and grid points for the domain calculations. The latitude is denoted by ϕ. The term $\cos\phi$ allows for the convergence of the meridians. In the ABM, the S_1 skill score is still a widely used measure of skill, particularly for SLP forecasts.

The gridpoint definition of bias may be positive or negative:

$$\text{bias} = (1/N_t) \sum_t [P_M(D, T) - P_A(D, T)] \cos\phi. \quad (7)$$

b. Other measures of model performance

Model consistency and model divergence are defined below. Model consistency applies to a single model and is a measure of the difference between forecasts at a particular time, but initiated at different times. Model divergence is the rms difference between a particular model and the model chosen to be the central member of the ensemble.

Domain averages are defined as

$$\text{con} = \{(1/N_t N_{xy}) \sum_t \sum_{xy} [P_M(D-1, T+1) - P_M(D, T)]^2 \cos\phi\}^{1/2}; \quad (8)$$

$$\text{div} = \{(1/N_t N_{xy}) \sum_t \sum_{xy} [P_M(D, T) - P_{EC}(D, T)]^2 \cos\phi\}^{1/2}. \quad (9)$$

The gridpoint definition of consistency is positive or negative:

$$\text{con} = (1/N_t) \sum_t [P_M(D-1, T+1) - P_M(D, T)] \cos\phi, \quad (10)$$

whereas the definition of gridpoint divergence is defined in a manner that retains the metric character:

$$\text{div} = \{(1/N_t) \sum_t [P_M(D, T) - P_{EC}(D, T)]^2 \cos\phi\}^{1/2}. \quad (11)$$

3. Results

a. Conflicting predictions

As mentioned in the introduction, the forecaster often is confronted with a selection of widely differing numerical predictions from various weather centers. This divergence between forecasts frequently is very obvious even as early as day 3, despite the models having been run from almost identical initial fields. The example presented in Fig. 1 and mentioned in the introduction compares 72-h forecasts from the US, UK, and AG models [the AG model also is referred to as GASP (Global Assimilation and Prognosis)] valid at 1200 UTC 23 May 1990. The particular event is an important one for southeastern Australia, as it involves a cutoff low pressure system (see Fig. 1a). Such systems are a common cause of flood rains and strong winds. There is a large disparity over southeastern Australia, particularly over southern New South Wales (NSW). The US model (Fig. 1b) predicts a cutoff low with a central pressure of 1009 hPa to the northwest of Melbourne. The UK prediction (Fig. 1c) has a 1028-hPa contour passing just south of Melbourne and an inland trough of intensity approximately 1014 hPa located in central NSW (with no closed contour). Finally, the AG 72-h prediction (Fig. 1d) has an intense cutoff low with a central pressure of 1007 hPa just to the north of Melbourne. The information provided to weather forecasters by these three predictions has very different implications in terms of local weather conditions for the areas influenced by the cutoff low. Also, there are major differences in other parts of the charts that are immediately obvious from inspection. The verifying analysis reveals that the cutoff low is actually straddling the southern portion of the east coast. Clearly, in this case all four models (including the EC, which is not shown) are incorrect. This is frequently the case when most or all of the models do not agree with each other.

b. Model statistics

1) Biases, rms Differences, and correlations: Domain-averaged

In Fig. 3 the biases, rms differences, and correlations between regional analyses and the four forecast models are shown. There are several features of note. The model rankings for May 1990 and 1991 combined are the same for each of the statistics. The EC model ranks first, followed by the UK and the US almost equal, and then AG. At $t = 0$ the local model (AG) is favored slightly, probably because it uses the same type of analysis scheme as the regional system and because it includes PAOBS (bogus observations) over the no-data area that are given the same weight as conventional observations by the Australian Bureau of Meteorology Research Centre (BMRC) analysis scheme. The remaining models did not include Australian PAOBS in the period of this study. A second feature is that the growth rate in the biases and rms differences is close to zero over the first 24 h for all models except AG. The error pattern then grows quasi-linearly out to about day 3 before exhibiting slight signs of leveling off.

2) Biases: Spatial patterns

The average EC gridpoint SLP bias for May 1991, defined by (5), is shown in Fig. 4. Figure 4a shows the

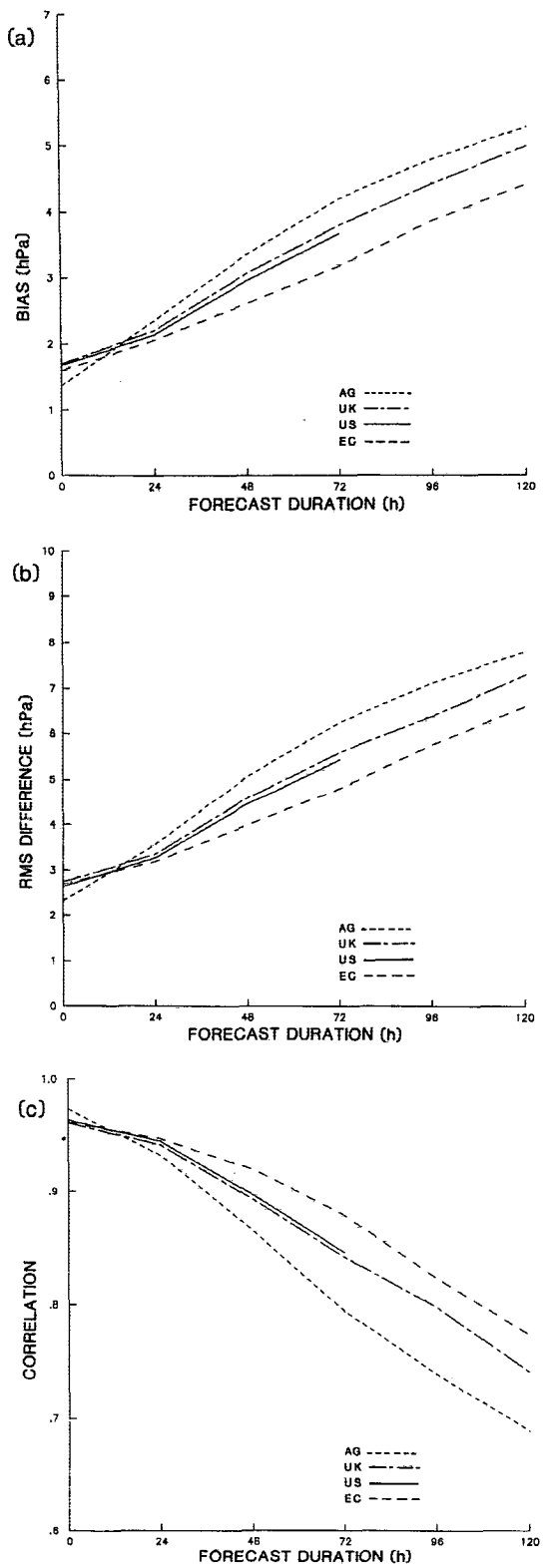


FIG. 3. The domain-averaged values for May (1990 and 1991) of three standard forecast performance statistics for each of the four models: (a) bias, (b) rms difference between forecasts and Australian region operational analysis, and (c) correlation coefficient.

average bias in the analyses, and Figs. 4b–d show the average bias for the forecasts over the next 5 days, at 2-day intervals. The overall pattern of the bias changes very little from the analysis through to day 5. However, the magnitude of the bias to the southwest of Western Australia, which is a region of critical importance, increases steadily from a maximum of 3.8 hPa in the analysis to a very large value of 8.7 hPa at day 5. Examination of Figs. 4b–d indicates that the bias is so large that it merits attention. The synoptic interpretation of this pattern is that the model is overforecasting the intensity of the subtropical ridge, a semipermanent feature of the Indian Ocean/Southern Ocean in this area. Similar patterns (not shown) were found for the 1990 data, although the SLP bias at day 5 to the southwest of Western Australia is considerably smaller and does not exceed 3.5 hPa.

3) CONSISTENCY AND DIVERGENCE: DOMAIN-AVERAGED

In Figs. 5a,b the model forecast consistency, as defined above in (7), is the difference between forecasts valid at the same time but initiated at different times. Figure 5a is a plot of the domain-averaged consistency as a function of forecast duration. The measure of forecast consistency of each of the models grows in time (that is, the model consistency deteriorates). Features of note are the slow deterioration of the consistency of the UK model, which interchanges position with the EC model after 72 h, and the relatively rapid deterioration of the EC model, to the last position after 96 h. Both of these features are observed in other months (not shown here).

The spatial distribution of model consistency for all forecast durations (not shown) is primarily a function of latitude and grows steadily with time. Also, there is little difference between the various months in the spatial patterns, although the maximum amplitudes vary by around 50%.

The divergence between models is presented in Fig. 5b, with the EC model as the central member of the ensemble. The model divergence as defined in (8) is the growth in time of the difference between the EC model and the other models. The divergence is seen to increase with the length of the forecast for all models. Apart from slight differences between the US and UK models at the initial time, their divergences are nearly equal, while the AG model is somewhat larger at all times. The results for months other than May are very similar to Fig. 5b.

The spatial distribution of divergence for May, and indeed for all months, (not shown) again reveals an expected pattern of a tendency to zonality and values increasing with latitude until by day 5 the highest values exceed 12 hPa in the Southern Ocean. Another notable feature of the spatial patterns of divergence is that the EC and AG models separate steadily with forecast duration over almost all of the domain.

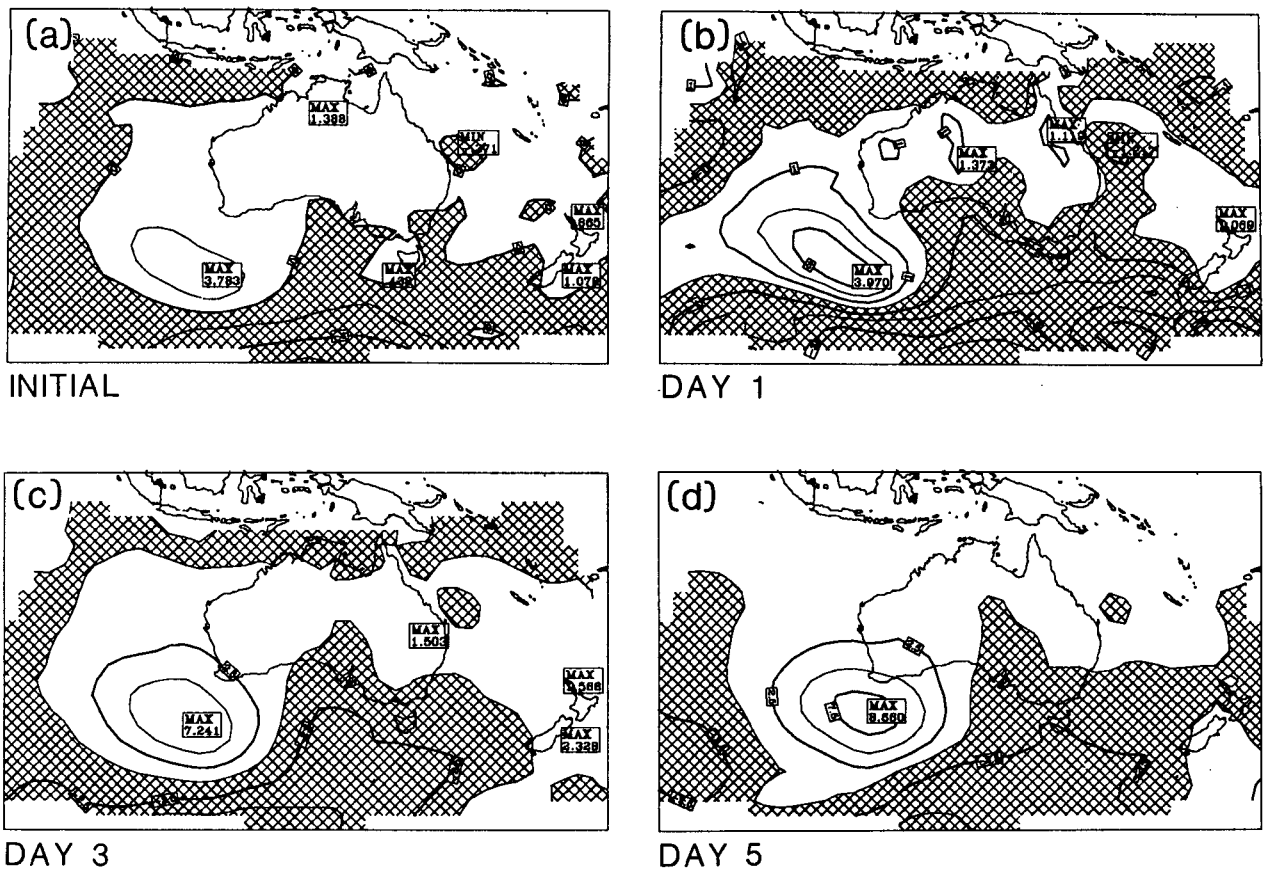


FIG. 4. Geographical distribution of the EC model bias for May (1990 and 1991): (a) initial, (b) day 1 forecast, (c) day 3 forecast, and (d) day 5 forecast.

4. Applications to forecasting

In this section the value (calculated in section 3) of the statistics to the operational forecasters will be demonstrated. It has been found that the fields of most interest are the geographical distribution (monthly averages) of systematic errors, such as those revealed by the maps of bias and the fields of consistency and divergence, which can be used to provide information about nonsystematic errors. First, a brief summary will be presented of the more important weaknesses in the model forecasts, as claimed by the operational forecasters. Then it will be shown how the model performance statistics assist the forecaster in general, and by the presentation of particular examples.

a. Forecaster-identified model deficiencies

During the period of the study, the operational forecasters at the ABM were asked to complete a survey in which they indicated their subjective assessment of deficiencies in each of the models. The survey was conducted by a member of the operations branch of the ABM, whose major responsibility is to monitor existing models and to assess new models. Presented herein is a summary of the main problems that were identified in at least one of the four models (T. C. L. Skinner, Operations Branch, 1993, personal communication).

(i) The forecasters claimed that generally, frontal systems, moved too slowly in the models when compared with observations.

(ii) Anticyclones also were regarded as moving too slowly, particularly the larger ones. Moreover, the subtropical ridge, which occupies a latitudinal band over the oceanic areas just to the south of Australia between latitudes 30° and 45°S is an important controlling feature of the weather over southern Australia. The subtropical ridge is regarded as being significantly weakened in intensity by the models to the southwest of the continent. In particular, the EC model is seen as systematically underestimating the strength of the subtropical ridge.

(iii) Analyzed tropical depressions tend to be weakened very rapidly. In particular, small tropical depressions frequently disappear from the model forecast within the first 24 h. These systems frequently bring flood rains and strong winds, and can develop into

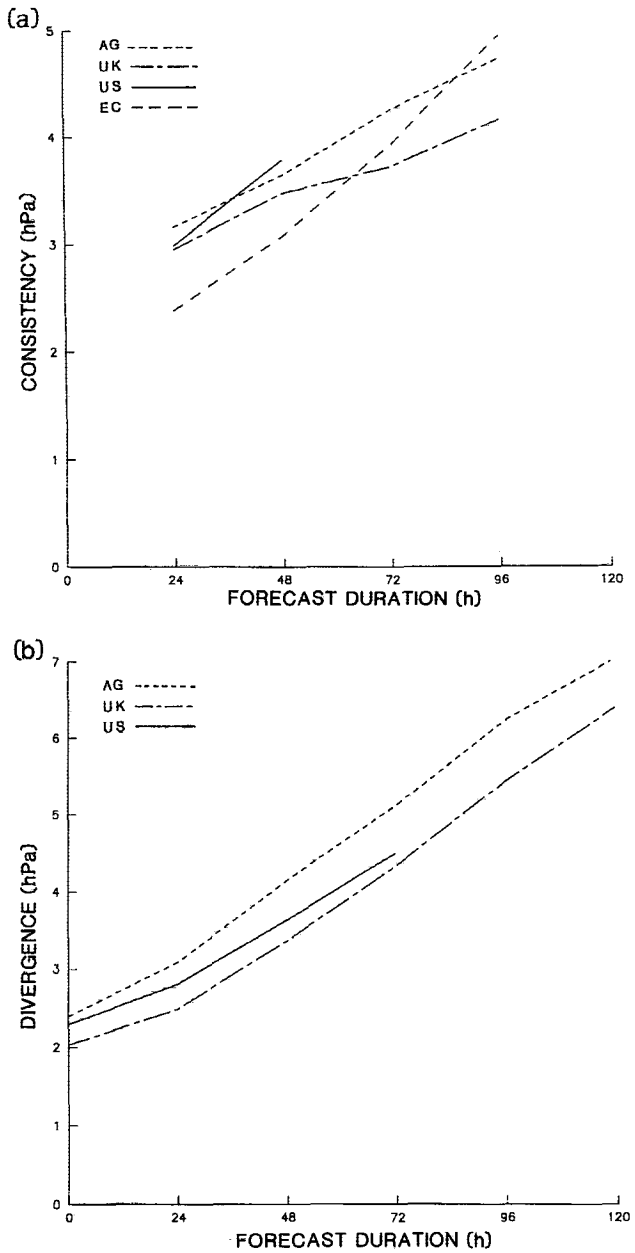


FIG. 5. As in Fig. 3 except for the model performance statistics: (a) consistency and (b) divergence.

tropical cyclones, so they are systems of primary concern.

(iv) Tropical cyclones are regarded as being poorly forecast by all models. This conclusion is supported by mean forecast track errors routinely calculated at the end of each tropical cyclone season. Initially, tropical cyclones are too weak. They weaken further during the forecast period, and their tracks are far more erratic than the observations indicate.

(v) The continental heat lows that form over the warmer months (late August to early April) are consistently underestimated in intensity and location by all models. This error appears to be very systematic and related to inadequate resolution and representation of physical processes in the model, particularly radiative processes.

(vi) The consistency in time of forecasts from each of the models with itself is regarded as being inadequate, particularly beyond 48 h.

(vii) The forecasters regarded the number of forecast conflicts as being far too great, once again particularly after 48 h. Such conflicts, as already pointed out, produce considerable concern and frequently indicate that none of the models has skill on that day.

b. Application of the performance statistics in operations

At the end of this study, the ensemble of forecasts from the four global models, and the performance statistics described in section 2, both in domain-averaged form and as regional maps, were available to the operational forecaster. One of the most important outcomes of the present study is that the deficiencies in the model forecasts identified above by the operational forecasters were seen in the performance statistics. This provides the forecasters with objective support for their

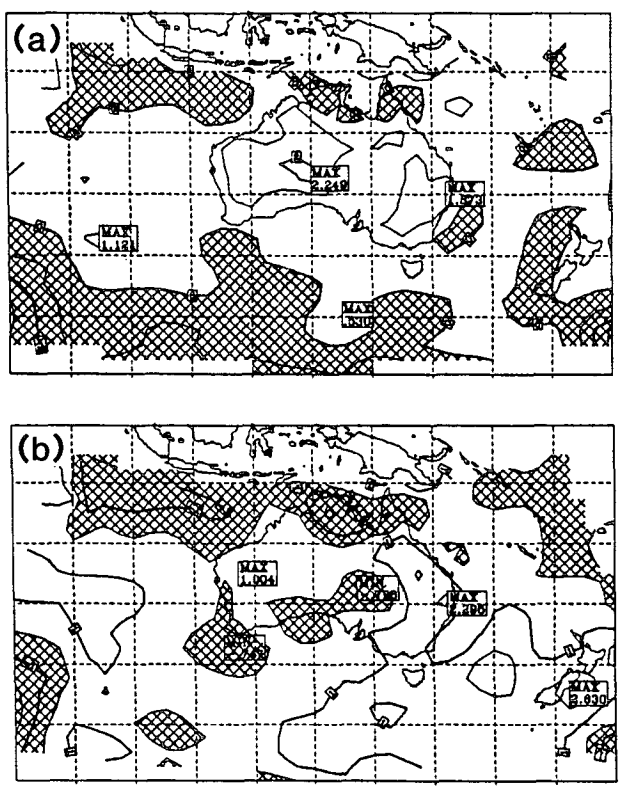
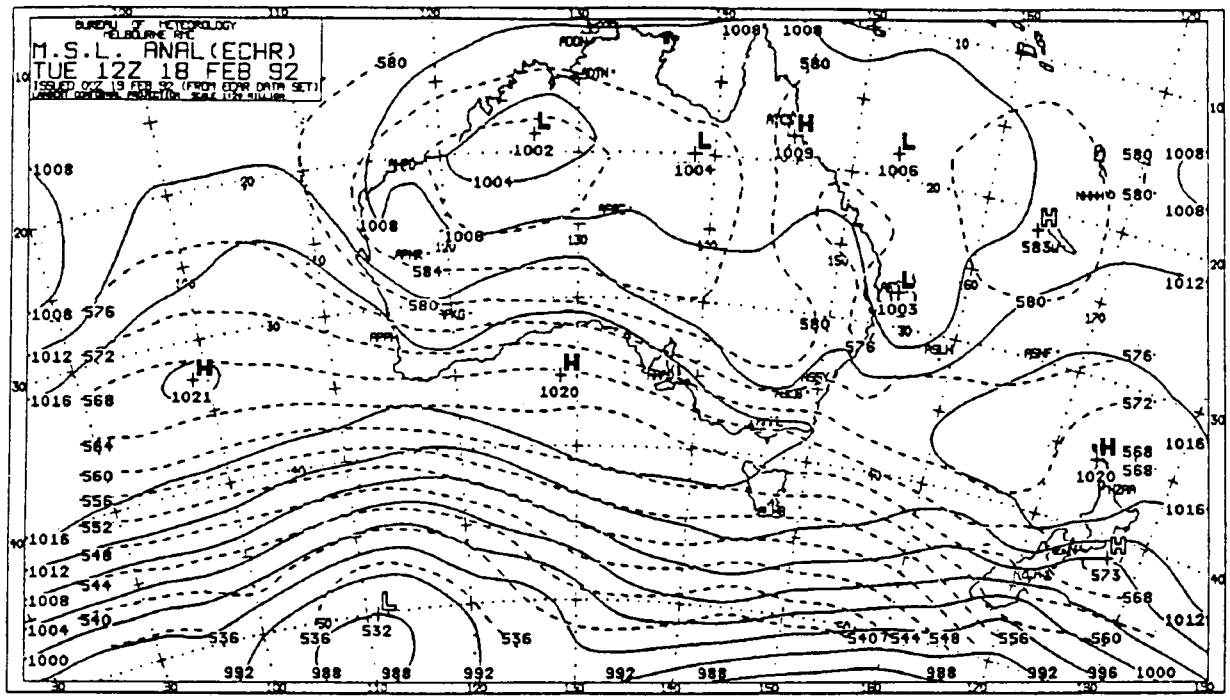


FIG. 6. (a) Geographical distribution of EC model bias for February at day 0 (analysis); (b) as in (a) except for day 1.

(a)



(b)

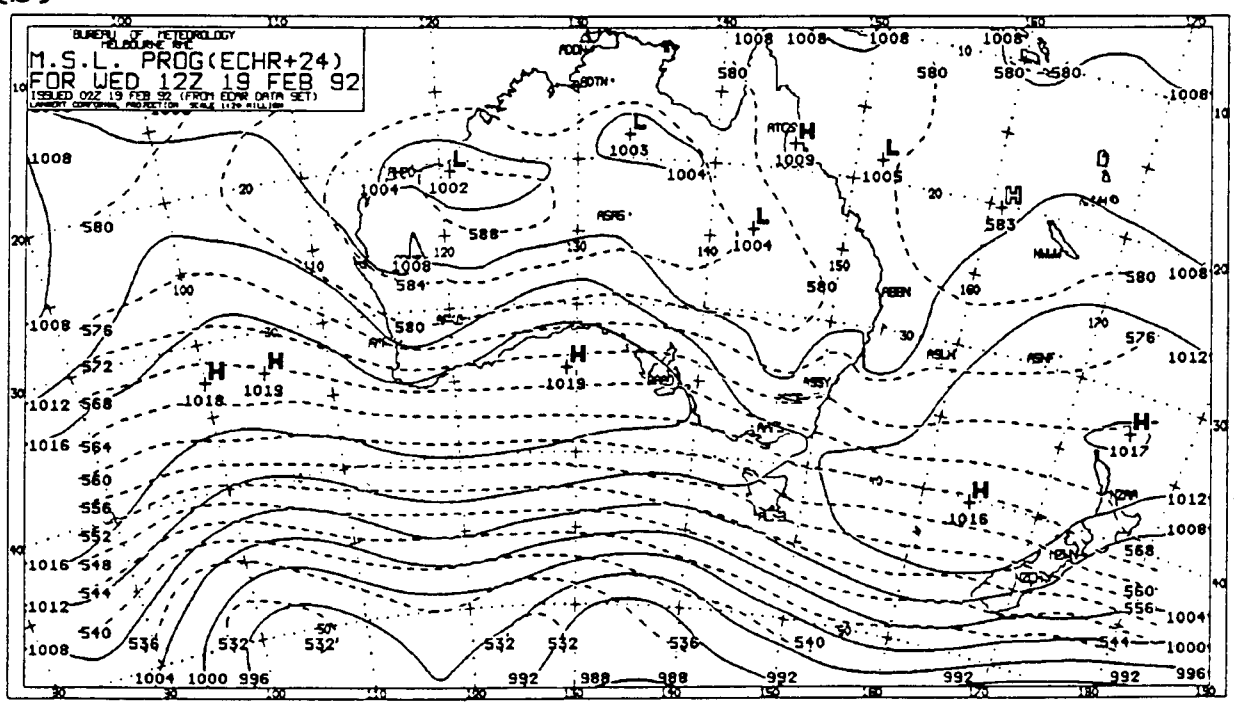
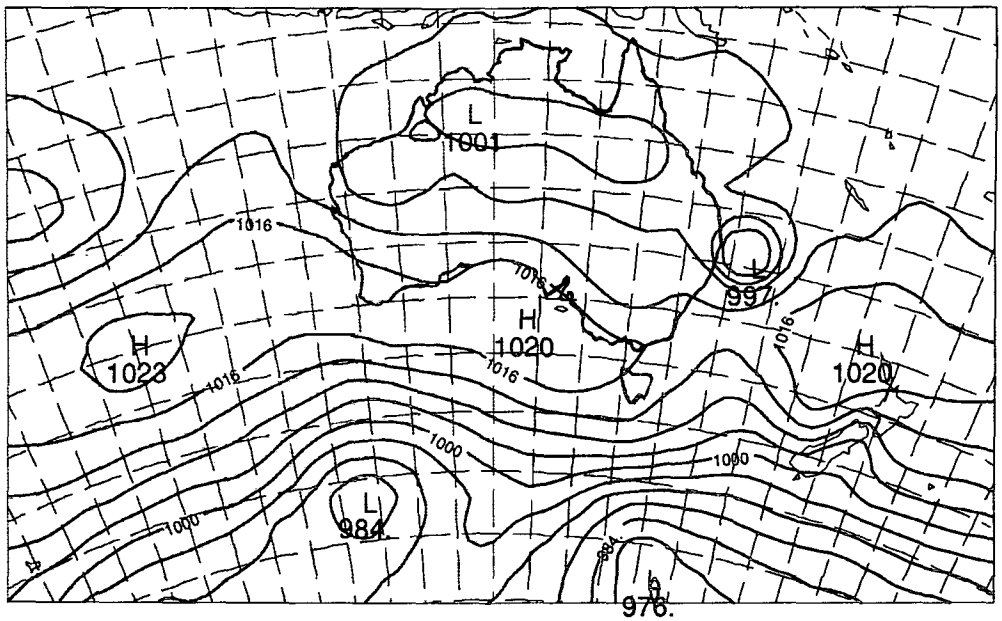


FIG. 7. (a) EC model analysis (day 0), 1200 UTC 18 February 1992; (b) EC model day 1 forecast; (c) RASP initial analysis (Day 0); and (d) RASP verifying analysis (day 1). Note that the tropical low pressure system off the central east coast has been markedly weakened.

(c)



(d)

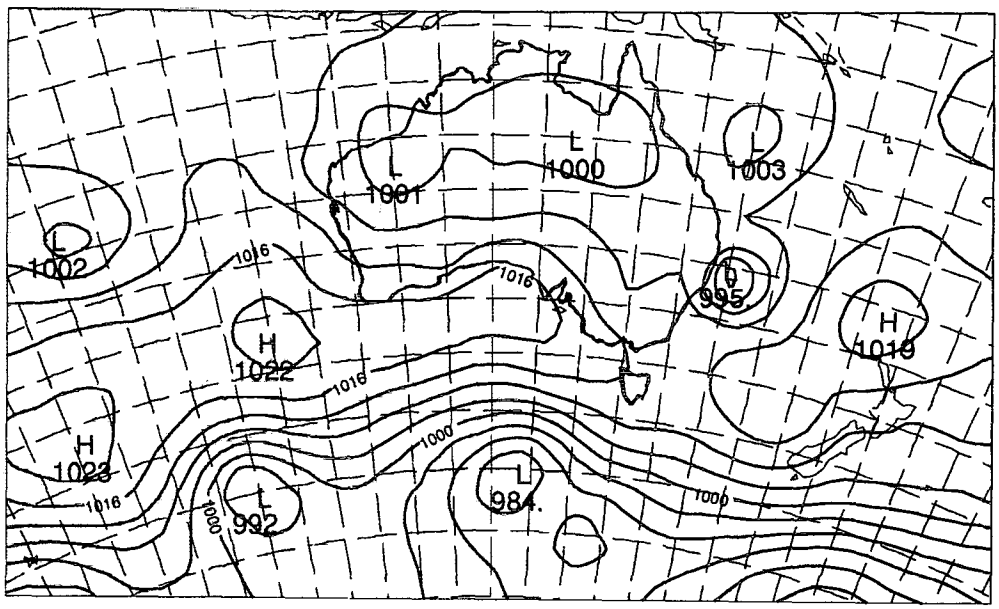


FIG. 7. (*Continued*)

predictions in general. The following small selection of three examples then demonstrates in detail the operational utility of the performance statistics.

1) BIAS

Figures 6a,b show the spatial pattern of the bias in the EC model, averaged over the months of February 1990 and 1991. There are a number of areas of interest, but here the focus will be on the central east coast of Australia. Figure 6a shows the bias in the analysis, and Fig. 6b the bias in the 24-h forecast. There is a maximum in the bias, indicating that the SLP is overestimated in the analysis and that the overestimation increases over the 24-h period. This feature is contrary to the perception of the operational forecasters, de-

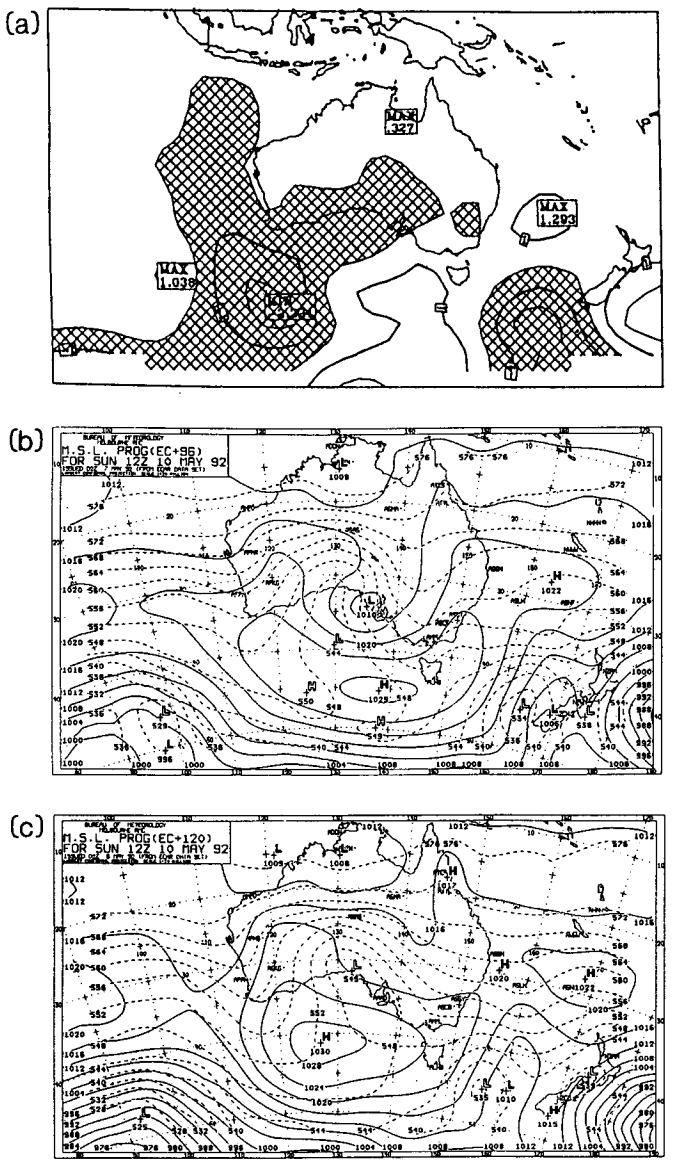


FIG. 8. Illustration of the use of the geographical distribution of model consistency: (a) the regional pattern of consistency between a 4- and 5-day forecast valid at the same time; (b) 4-day EC forecast valid at 1200 UTC 10 May 1992; and (c) as in (b) except for a 5-day EC forecast.

scribed above: the tropical depressions over the central east coast are underestimated in intensity in the analysis and are then unrealistically weakened further in the first 24-h.

A specific example is given in Fig. 7. In Fig. 7a, a small tropical low is present in the EC analysis with a central pressure of 1003 hPa. In the 24-h forecast the small tropical depression has virtually disappeared. The corresponding ABM numerical analyses shown in Figs. 7c,d reveal that the small low had an initial central pressure of 997 hPa and deepened to 995 hPa 24 h later.

2) CONSISTENCY

The spatial distribution of consistency at day 5 is shown in Fig. 8a for the EC model averaged over May 1990 and 1991. The field displayed in this case is simply the algebraic part of (7) and does not include the squared or square-root symbols. The feature of interest here is the area of minimum consistency to the southwest of Australia. Figures 8b,c support this distribution, with the day 4 forecast having considerably lower pressure values than the day 5 forecast. In this case, the forecaster would be guided better by the day 5 forecast,

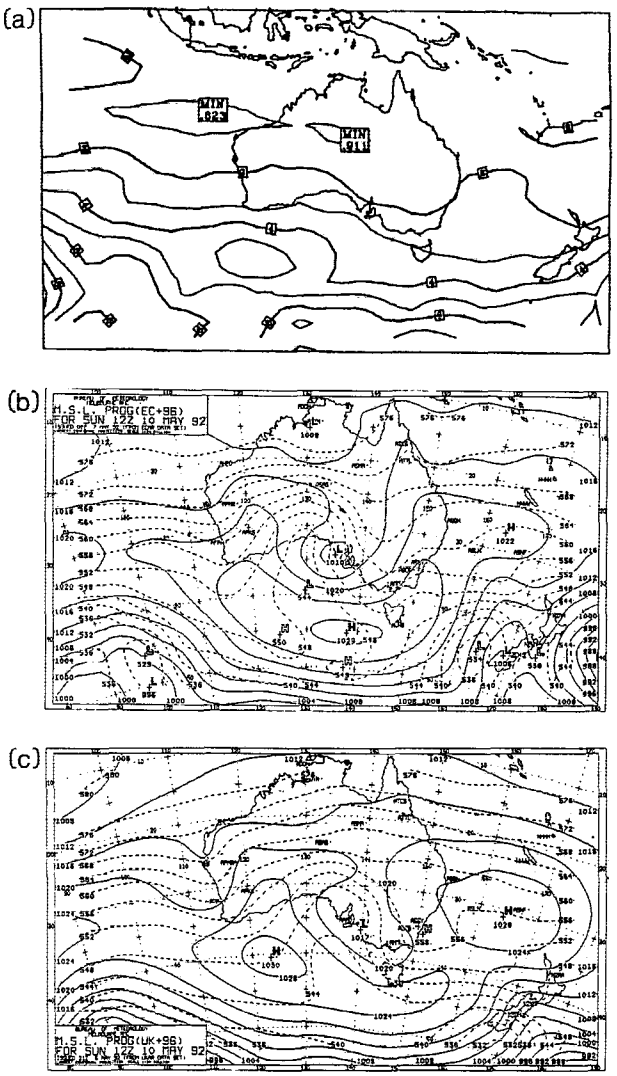


FIG. 9. As in Fig. 8 except for the model performance statistic, divergence between the UK and EC models at 4 days for May: (a) field of divergence; (b) 4-day EC model forecast; and (c) as in (b) except for the UK model.

at least for this part of the domain. Over other parts of the chart, there is little guidance for the forecaster except for the midlatitude depression in Fig. 9b located on the central south coast. Here, in the absence of information to the contrary, the 4-day forecast would be preferred to the quite different 5-day forecast on the basis that 5-day forecasts are known to be almost always inferior to 4-day forecasts.

3) DIVERGENCE

Figure 9 is a specific example of divergent forecasts obtained from two of the models. In fact, all four models were different on this day in terms of their prediction of the location of the anticyclone in the subtropical ridge. Figure 9a is the spatial distribution of the monthly mean average of divergence at day 4 for the months May 1990 and 1991. Note the maximum to the southwest of Australia. This is yet another of the regions identified by the forecasters as being not well forecast by the models. Figures 8b,c are the 4-day forecasts from the EC and UK models, respectively, and they reveal that the anticyclone in the subtropical ridge is centered at approximately 45°S, 138°E in the EC model, while for the UK model it is centered at 40°S, 123°E. This is a very large difference and confirms the concerns of the operational forecasters in this region. In this case the forecasters would lean toward the UK model, as they are aware of the systematic bias already presented in Fig. 4.

There is another feature in Fig. 9 that merits attention. Although there is no real evidence of a maximum in the divergence field, the 4-day forecasts of the low pressure system over the central south coast are also very different, both in structure and intensity. In this case, the forecasters would look at the US and AG forecasts. If they were also different, there would be little reason for the forecasters to choose one model in preference to another and they would be forced to resort, without confidence, to some other approach. Typically, in this situation, the forecasters revert to some kind of "weighted average" of one or more of the models.

5. Discussion and conclusions

In this study the performance of the four global operational models received by the Australian Bureau of Meteorology has been assessed using a range of objective measures. The motivation was to quantify the current subjective means of producing the official forecast charts, which is forecaster dependent and ranges from a "favorite" model approach (most often the EC model) through to a blend of all of the charts. The blend consists of subjective weightings of how well the individual models are regarded as predicting particular events, or their perceived performance over specific geographical areas of the forecast domain. In this sense it is not like the simple objective ensemble averaging of the global and regional forecast models described by Smith and Mullen (1993) for the National Meteorological Center, Washington, D.C. Two types of objective measures were used to characterize the statistical properties of the global model forecasts. The first group were the standard statistical quantities that highlight in particular the amplitude and phase differences from the corresponding regional numerical analyses. These were the SLP bias and the associated rms differences, and the correlation coefficient. The second group was composed of intra- and intermodel comparisons. The intramodel measure was temporal consistency, while the intermodel measure involved calculating the divergence (often referred to elsewhere as the "separa-

tion") of the various model forecasts relative to a base model that we chose to be the EC model.

The standard statistical techniques showed a definite objective ranking of the performance of the models. For the period of the study the EC model performed best, followed by the UK and US models, which were almost equal, and then by the AG model. However, it is emphasized that the purpose of this study was to set up an objective, practical method of evaluating the skill of the models that has practical value to forecasters. Such a method must be applied regularly because changes in the model configurations could well alter the rankings and associated performance statistics in the Australian region.

Although the model comparison statistics were computed for all months of the two-year sample, results are presented only for one month, May. This month was chosen essentially for reasons of brevity and because it was possibly the best month in terms of domain-averaged agreement between the analyses from the different models and the local regional analysis.

Other findings of the study were the identification of a region of systematic bias in the subtropical ridge to the southwest of Western Australia. Southern Ocean anticyclones were systematically overestimated in intensity by all models for this period, and this overestimation increased with time to a large maximum (8.7 hPa) at 5 days. However, in other regions, notably the populous east coast, the errors were nonsystematic. In this region there is no means, at present, of estimating which of the models should be preferred. Results for the warmer months, which are not shown here, revealed that the models failed to capture the strength of the Western Australian heat trough. For example, the EC model overestimates the pressure in the heat trough by approximately 3 hPa at 48 h. Tropical cyclones also were weakened markedly, and mean forecast track errors were unacceptably large.

Now that the present study has developed objective measures of model performance, the next step, which has just commenced, is a search for predictors that correlate highly with these model forecast error fields. This program has a high priority because in the Australian region forecast conflicts occur frequently, and it is of great importance to develop a reliable measure of forecasting model forecast skill.

Acknowledgments. The authors express their appreciation to the reviewers of this manuscript for their valuable comments, to Terry Skinner of the Australian Bureau of Meteorology for assisting in the survey that identified operational forecasting problems, to Mike Manton of BMRC for careful reading of the manuscript, and to David Pike also of BMRC for drafting assistance. This work was partially supported by ONR Grant N00014-89-J-1737.

REFERENCES

Brankovic, C., T. N. Palmer, F. Molteni, S. Tibaldi, and U. Cubasch, 1990: Extended-range predictions with ECMWF models: Time-lagged ensemble forecasting. *Quart. J. Roy. Meteor. Soc.*, **116,** 867–912.

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.

Leslie, L. M., and G. J. Holland, 1991: Predicting regional forecast skill using single and ensemble forecast techniques. *Mon. Wea. Rev.*, **119,** 425–435.

——, K. Fraedrich, and T. J. Glowacki, 1989: Forecasting the skill of a regional numerical weather prediction model. *Mon. Wea. Rev.*, **117,** 550–557.

Mureau, R., F. Molteni, and T. N. Palmer, 1993: Ensemble prediction using dynamically conditioned perturbations. *Quart. J. Roy. Meteor. Soc.*, **119,** 299–323.

Murphy, A. H., and E. S. Epstein, 1989: Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.*, **117,** 572–581.

Palmer, T. N., and S. Tibaldi, 1989: On the prediction of forecast skill. *Mon. Wea. Rev.*, **116,** 2453–2480.

Smith, B. B., and S. L. Mullen, 1993: An evaluation of sea level cyclone forecasts produced by NMC's Nested-Grid Model and Global Spectral Model. *Mon. Wea. Rev.*, **8,** 37–56.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74,** 2317–2330.

Wobus, R. L., and E. Kalnay, 1991: Prediction of forecast skill for the NMC global model. Preprints, *Ninth Conf. on Numerical Weather Prediction,* Denver, CO, Amer. Meteor. Soc., 481–484.