NOTES AND CORRESPONDENCE

Verification of Quantitative Precipitation Forecasts from Operational Numerical Weather Prediction Models over Australia

JOHN L. MCBRIDE AND ELIZABETH E. EBERT

Bureau of Meteorology Research Centre, Melbourne, Australia

17 February 1999 and 28 June 1999

ABSTRACT

Real-time gridded 24-h quantitative precipitation forecasts from seven operational NWP models are verified over the Australian continent. All forecasts have been mapped to a 1° latitude–longitude grid and have been verified against an operational daily rainfall analysis, mapped to the same grid. The verification focuses on two large subregions: the northern tropical monsoon regime and the southeastern subtropical regime. Statistics are presented of the bias score, probability of detection, and false alarm ratio for a range of rainfall threshold values. The basic measure of skill used in this study, however, is the Hanssen and Kuipers (HK) score and its two components: accuracy for events and accuracy for nonevents.

For both regimes the operational models tend to overestimate rainfall in summer and to underestimate it in winter. In the southeastern region the models have HK scores ranging from 0.5 to 0.7, and easily outperform a forecast of persistence. Thus for the current operational NWP models, the 24-h rain forecasts can be considered quite skillful in the subtropics. On the other hand, model skill is quite low in the northern regime with HK values of only 0.2–0.6. During the summer wet season the low skill is associated with an inability to simulate the behavior of tropical convective rain systems. During the winter dry season, it is associated with a low probability of detection for the occasional rainfall event. Thus it could be said that models have no real skill at rainfall forecasts in this monsoonal wet season regime.

Model skill falls dramatically for occurrence thresholds greater than 10 mm day^{-1}. This implies that the models are much better at predicting the occurrence of rain than they are at predicting the magnitude and location of the peak values.

1. Introduction

This study documents the level of skill at precipitation forecasts of the large-scale numerical weather prediction (NWP) models currently available at the National Meteorological Operations Centre Melbourne, Australia. Real-time gridded precipitation data are verified from seven models from Australia, the United Kingdom, Germany, Japan, the United States, and the European Centre for Medium-Range Weather Forecasts (ECMWF) over a one-year period. The verifications are for the first 24 h of the forecast period and are carried out over the Australian continent. The measures used are categorical statistics based on the rain contingency table (Table 1) applied at each analysis grid point over the verification period.

This study falls within the Techniques Development Initiative of the Australian Bureau of Meteorology, which has as one of its principal objectives the improvement of quantitative precipitation forecasts (QPF). The purpose of the current study is to present a concise and synthesized documentation of the current level of skill of operational numerical models at precipitation forecasts. Verification statistics are presented over a standardized 1° latitude–longitude grid over the continent of Australia as well as two large subregions thereof. It is hoped that the skill scores used over these standard grids can be used as a benchmark to measure improvements as NWP models develop further in coming years. It is also intended that the concise statistics presented here will provide a basic source of information to operational weather forecasters on the reliability of the precipitation forecasts from the various models used as input to their daily operations.

Daily precipitation patterns include large gradients over a range of scales, and forecasters and model developers are interested in quantitative verification on timescales from individual days up to seasons or years. Verification methods are not straightforward, however, and constitute an active area of research. Some discussions of current verification procedures in the United

Corresponding author address: Dr. John L. McBride, Bureau of Meteorology Research Centre, GPO Box 1289K, Melbourne 3001, Australia.
E-mail: J.Mcbride@bom.gov.au

© 2000 American Meteorological Society

TABLE 1. Rain contingency table applied at each verification grid point over the period of verification. A threshold value (e.g., 1 mm day^{-1}) is chose to separate rain from no-rain events. Here, Z is the number of correct predictions of rain amount below the specified threshold, F is the number of false alarms, M is the number of misses, and H is the number of correct rain forecasts or hits.

	Predicted	
Observed	No rain	Rain
No rain	Z	F
Rain	M	H

States are contained in the studies by Mesinger et al. (1990), Mesinger (1996), and Johnson and Olsen (1998). An internal report by the current authors (Ebert and McBride 1997) presents verification products over Australia including maps, time series, scatterplots, and tables of statistics over a number of time and space scales. In that report we also describe the various categorical statistics skill scores and discuss their relative advantages and disadvantages. On a more general level, excellent reviews of forecast verification methods have been carried out by Murphy and Winkler (1987), Stanski et al. (1989), and Wilks (1995).

The current paper chooses a small subset of the various measures of accuracy and skill that are available, and uses these measures to compare the performance at QPF of various NWP models in two different rainfall regimes for two different seasons (summer and winter). This exercise reveals a number of interesting aspects of the performance of the models. Thus a secondary purpose of the current paper is to indicate by example the type of information on model skill that can be quantified through the various categorical measures of skill.

Section 2 of the paper briefly describes the period of study, the data used for verification, and other aspects of the verification. In section 3, statistics are presented over the entire continent of Australia. One of the points made in that section is that the statistics are difficult to interpret as a number of very different rainfall regimes is involved. As a consequence of that, statistics are presented for two large subareas of the country. In section 4 the QPF verification is carried out over the Tropics, while in section 5 it is carried out over the subtropical southeast. The final section summarizes the performance of the seven models in tabular form for the two subregions and two seasons. The same information is compared against a persistence forecast through presentation of the performance of the individual models on phase-space diagrams where the dimensions are different measures of skill.

2. Verification procedures

Gridded precipitation forecasts were obtained for seven NWP models over the 12-month period September 1997–August 1998 inclusive. The models involved are all state of the science as they are the official operational forecast models of the respective meteorological agencies. The list of models, the owner meteorological agency, and the grid resolution of the precipitation data are given in Table 2.

For each model the gridded quantitative precipitation forecasts over the 0–24-h forecast period were collected for the one year of verification. These were verified against the gridded Australian operational daily rainfall analysis described by Weymouth et al. (1999). This uses a three-pass Barnes successive corrections technique to analyze rainfall observations from approximately 6000 synoptic, telegraphic, and observer sites onto a 0.25° latitude–longitude grid covering the land area of Australia, excluding some data-void areas in the central and western part of the country that have no rain gauges (Fig. 1). For details of data completeness and station locations, the reader is referred to Fig. 2 of Ebert and Weymouth (1999).

The rainfall analysis is based on 24-h accumulation observations made at 0900 local time. This corresponds to between 2300 and 0100 UTC in winter and between 2200 and 0100 UTC in summer (due to the spread of time zones across the country, and the fact that some states follow summer time and others do not). In this study we analyze the model output from the 0000 UTC run, which for each model has a base time of either 2300 or 0000 UTC. Thus in all cases the 24-h model forecasts and the 0900 local time accumulated precipitation observations are synchronous to within 2 h.

The precipitation output from the seven models and the operational rainfall analysis correspond to a variety of grid sizes. In order to produce a valid intercomparison

TABLE 2. NWP models for which 0–24-h quantitative precipitation forecasts were verified. Italics signify that the grid resolution of the model output received at the Bureau of Meteorology is coarser than the true resolution of the model.

Model	Operational center	Model type	Grid resolution (°lat–lon)
DWD	Deutscher Wetterdienst, Germany	Global spectral model	1.125°
ECMWF	European Centre for Medium-Range Weather Forecasts, United Kingdom	Global spectral model	0.50°
GASP	Bureau of Meteorology, Australia	Global spectral model	2.50°
JMA	Japanese Meteorological Agency, Japan	Global spectral model	***2.50°***
LAPS	Bureau of Meteorology, Australia	Nested regional gridpoint model	0.75°
AVN	National Centers for Environmental Prediction, United States	Global spectral model	***1.25°***
UKGC	Meteorological Office, United Kingdom	Global gridpoint model	***1.25°***

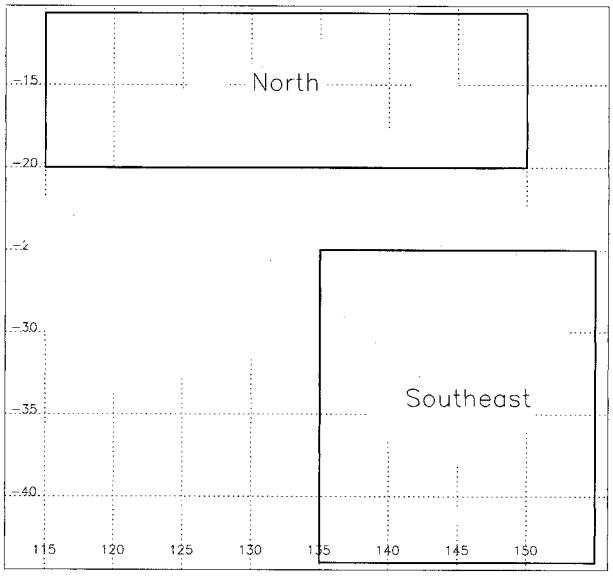

FIG. 1. Map of Australia showing the region of verification, which is the entire land area minus the blank data-void regions. Also shown are the northern and southeastern verification regions.

it was necessary to transform all fields to a standard grid size; an intermediate resolution of 1° latitude–longitude was chosen. The remapping was done by spatial averaging or by simple bilinear interpolation, depending on whether the grid size was being increased or reduced. This procedure resulted in 608 matched grid points in the analysis and in each of the prediction models. It is possible this remapping procedure may affect the comparative skills of models. The authors intend to investigate this aspect of verification in a later study.

Verifications were carried out over these grids for each individual month, as well as for the summer and winter seasons. Direct comparison can be made of mean values of *rain area, rain rate,* and *rain volume.* In addition simple point by point comparisons like *root-mean-square difference, mean absolute error,* and *coefficient of linear correlation* between forecast and analysis can be computed.

In addition to these simple measures a number of categorical statistics are applied. The term *categorical* refers to the yes/no nature of the forecast verification at each point. Some threshold (i.e., 0.1, 1, 2, 5, 10, 20, 50 mm day^{-1}) is considered to define the transition between a rain versus nonrain event. Then at each grid point, each verification time is scored as falling under one of the four categories of *correct nonrain forecasts, false alarms, misses,* or *hits* (*Z, F, M,* or *H* as shown in Table 1). Unless otherwise stated, in the following graphs and tables a rain event is defined to occur when the grid square average precipitation is equal to or greater than 1 mm day^{-1}.

A number of categorical statistics skill measures are

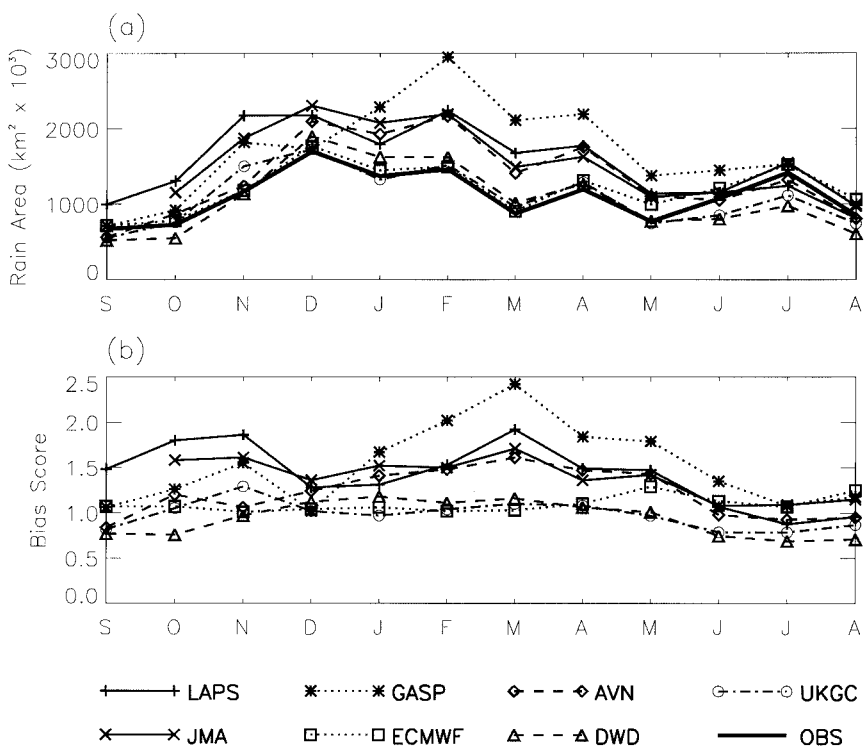

FIG. 2. Time series of monthly average values of (a) total rain area and (b) bias score over the Australian continent for the seven models discussed in the text. The solid heavy line shows the observed mean rain area from the operational daily rainfall analysis system.

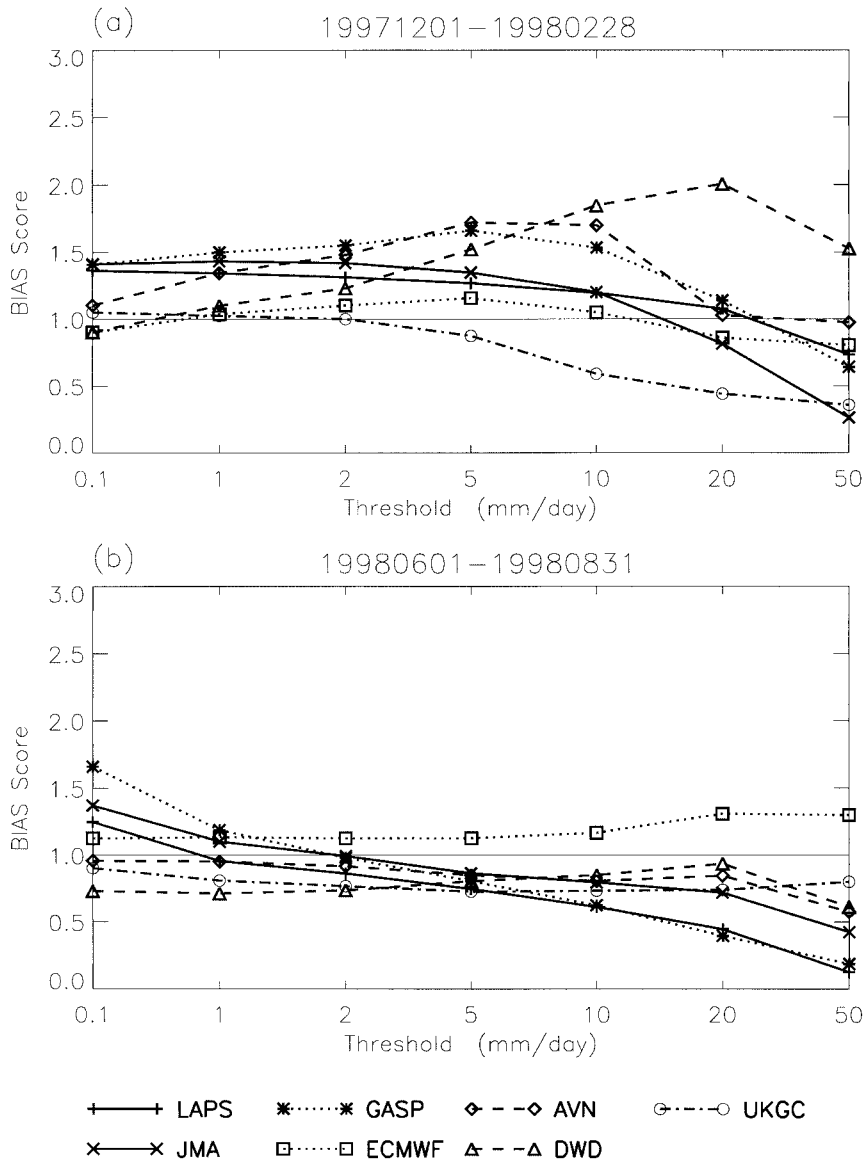

FIG. 3. Space–time average values of bias score over the Australian continent for the seven models as a function of rain threshold for (a) summer (Dec–Feb) and (b) winter (Jun–Aug).

used, computed from the elements of this rain/no-rain contingency table. They include bias score (bias):

$$\text{bias} = \frac{F + H}{M + H}. \tag{1}$$

The bias score is equal to the number of rain forecasts divided by the total number of observations of rain. Thus the bias score is a measure of the relative frequency of rain forecasts compared with observations. Averaged over a number of forecasts it is equal to the average rain area in the forecasts divided by the area in observations.

The probability of detection (POD) is equal to the number of hits divided by the total number of rain observations; thus it gives a simple measure of the proportion of rain events successfully forecast by the model:

$$\text{POD} = \frac{H}{M + H}. \tag{2}$$

The false alarm ratio (FAR) is equal to the number of false alarms divided by the total number of times rain was forecast; thus it gives a simple proportional measure of the model's tendency to forecast rain where none was observed:

$$\text{FAR} = \frac{F}{F + H}. \tag{3}$$

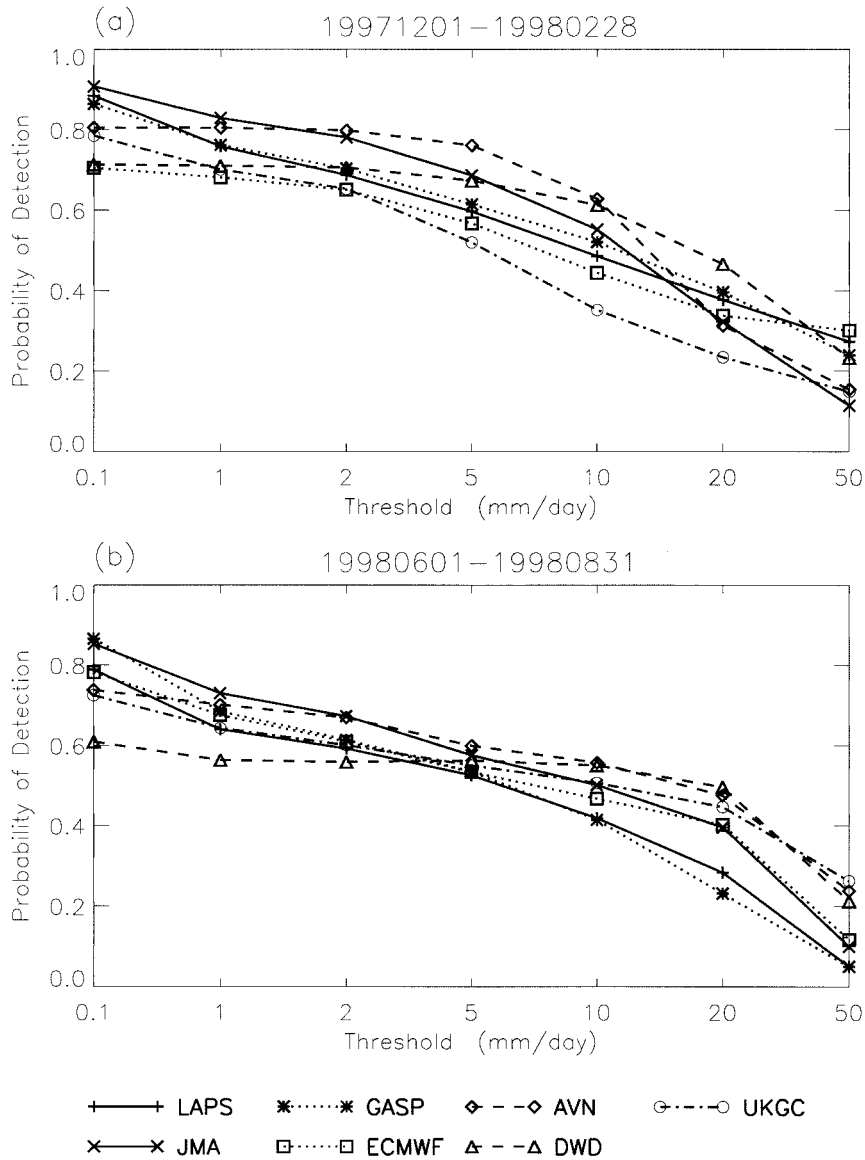

FIG. 4. As in Fig. 3 but for POD.

The Hanssen–Kuipers score (HK), also known as the *true skill statistic,* was developed by Hanssen and Kuipers (1965) and is used by the current authors as a result of the ideas advocated by Woodcock (1976) and discussed further by Ebert and McBride (1997):

$$\mathrm{HK} = \frac{(\mathrm{ZH} - \mathrm{FM})}{(Z + F)(M + H)}. \tag{4}$$

We can interpret the POD as being equal to a measure of *accuracy for rain events,* that is, the number of correct rain forecasts divided by the number of times rain was observed [$H/(M + H)$]. We can similarly define an *accuracy for nonrain events,* equal to the number of correct forecasts of nonrain divided by the number of nonrain events observed [$Z/(Z + F)$]. Through algebraic manipulation, it can be shown that the HK score is equal to the sum of these two accuracies minus one, where the last factor simply normalizes the score to the value range −1 to 1. Thus,

$$\begin{aligned}\mathrm{HK} &= \text{accuracy for events (POD)} \\ &\quad + \text{accuracy for nonevents} \\ &= \frac{H}{M + H} + \frac{Z}{Z + F} - 1.\end{aligned} \tag{5}$$

We prefer this scale because it is independent of the distribution of events and nonevents in the sample set. This facilitates the comparison of model skill across geographical and seasonal regimes. It also has an appeal in the fact that it can be expressed in terms of the gen-

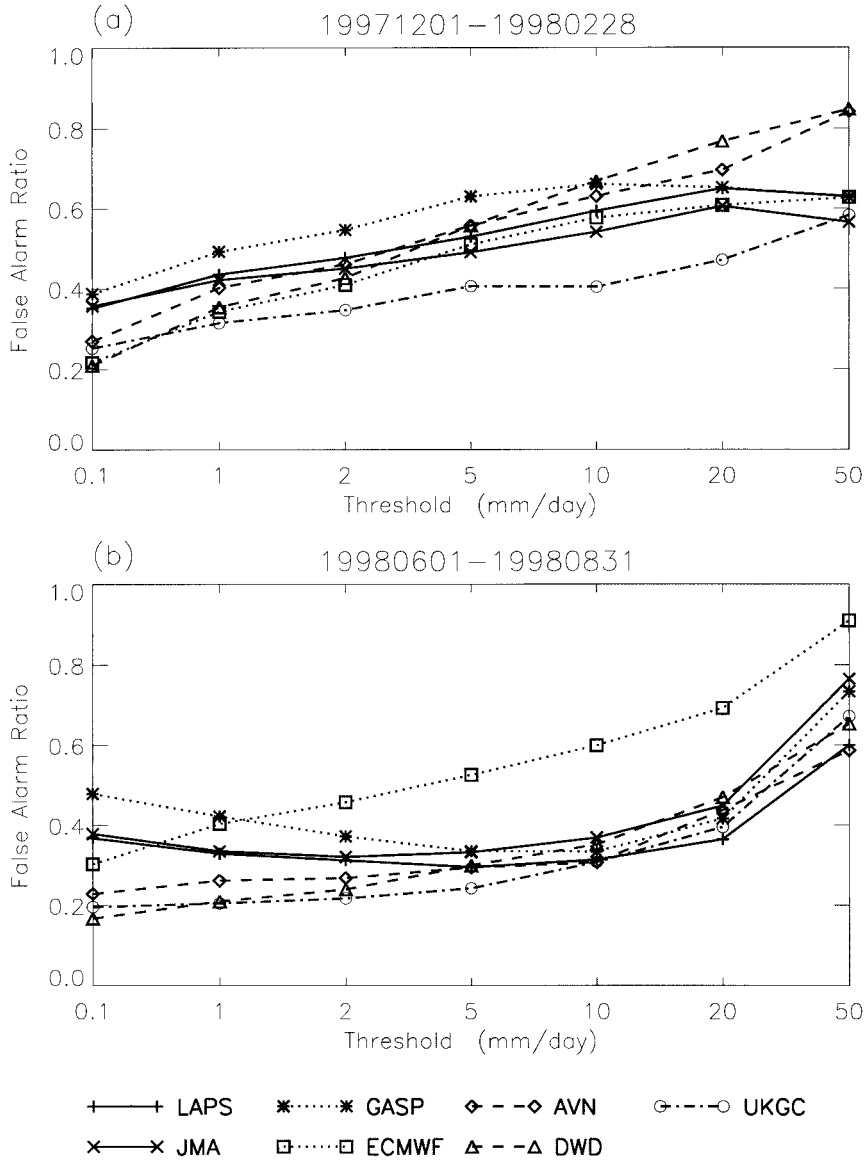

FIG. 5. As in Fig. 3 but for FAR.

eralized expression for skill as described by Wilks (1995):

$$\text{skill} = \frac{\text{accuracy}_{\text{forecast}} - \text{accuracy}_{\text{reference}}}{\text{accuracy}_{\text{perfect}} - \text{accuracy}_{\text{reference}}}. \quad (6)$$

When the HK score is expressed in this context, the accuracy is that for all forecasts, that is, both rain and nonrain events. In the numerator the reference forecast is that obtained by random chance with a model with the same distribution of yes/no (or event/nonevent) forecasts as that being verified. In the denominator the accuracy of a perfect forecast is set to 1.0, while the reference accuracy in the denominator is that obtained by chance by a hypothetical forecast model with the same distribution of yes/no events as occurs in the observational sample. It is this fact, that the denominator is totally independent of the characteristics of the particular forecast model, that gives the HK score an aesthetic or intellectual appeal for the intercomparison of different forecast models. For further details and discussion of HK in terms of Eq. (6), the reader is referred to Wilks (1995, chapter 7).

Referring back to Eq. (5), it is seen that the HK score gives equal emphasis to the ability of the forecast model to correctly predict events and nonevents. It has been criticized by Stanski et al. (1989) and Doswell et al. (1990) as being overly dependent on the probability of detection, and therefore less appropriate than other measures of skill in forecasting rare events (where Z, the correct forecasts of nonevents, dominates the contin-

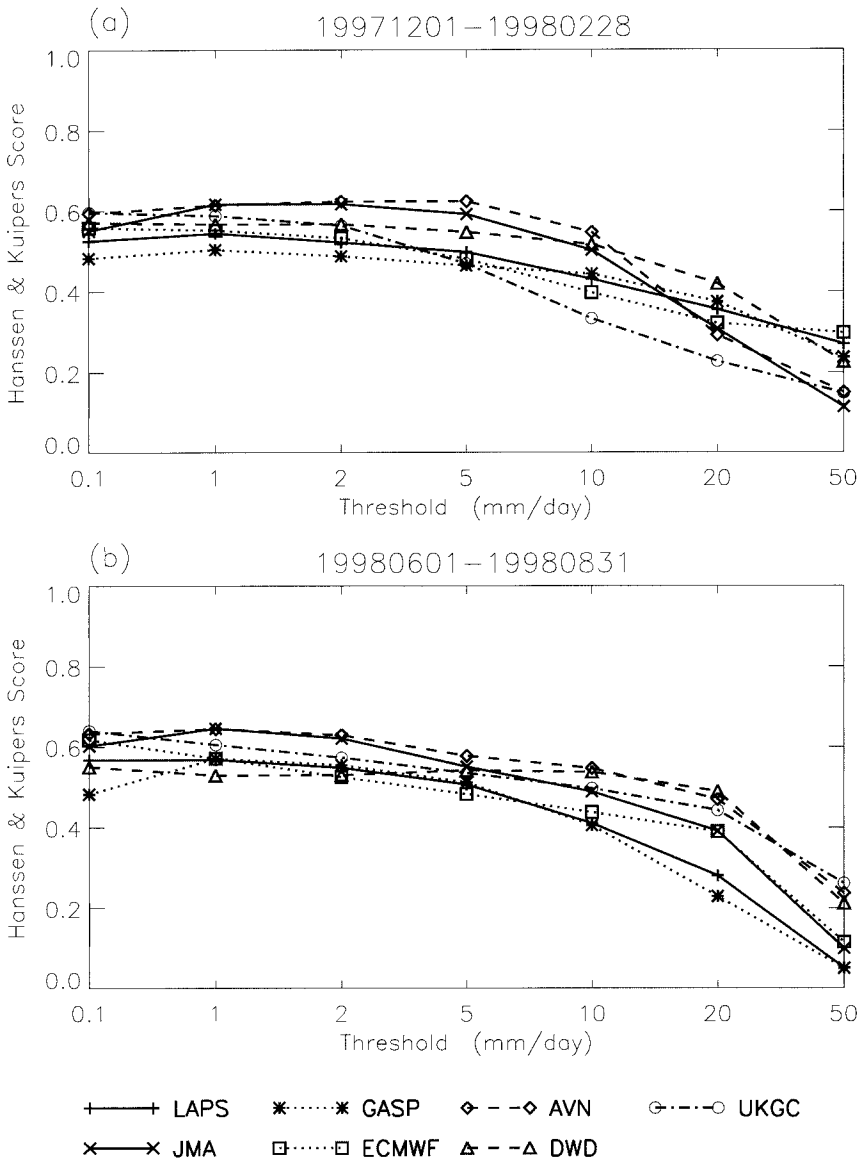

FIG. 6. As in Fig. 3 but for the HK score.

gency table). It is worth noting that the formulation of the HK score contains an inherent scale. As described above, the perfect forecast system ($M = 0$, $F = 0$) has an HK score of $+1$, while a constant forecast (i.e., either $Z = M = 0$, or $F = H = 0$) has an HK score of zero.

3. Verification over the Australian continent

Figure 2 shows a time series of monthly mean values of total rain area and bias score averaged over the continent for the year of verification. The upper panel (rain area) also shows the value from the analysis scheme. From inspection of the upper panel it is seen that for much of the year (October–May) four models [the Global Assimilation and Prediction System (GASP), the Japanese Meteorological Agency global spectral model (JMA), the Limited Area and Prediction System (LAPS), and the National Centers for Environmental Prediction's (NCEP) Aviation Model (AVN)] consistently overestimate the rain area. The ratio of the forecast to observed rain area is equal to the bias score shown in the second panel. This shows the overestimation of rain area to be of the order of 50% for three models (JMA, LAPS, and AVN) and approximately 100% for the GASP model. In contrast the remaining three models [U.K. Meteorological Office Global Model (UKGC), ECMWF, and the Deutscher Wetterdienst global spectral Model (DWD)] reproduce the total rain area to within 10% for most of the year, but two of them (UKGC and DWD) underestimate it by about 20% over

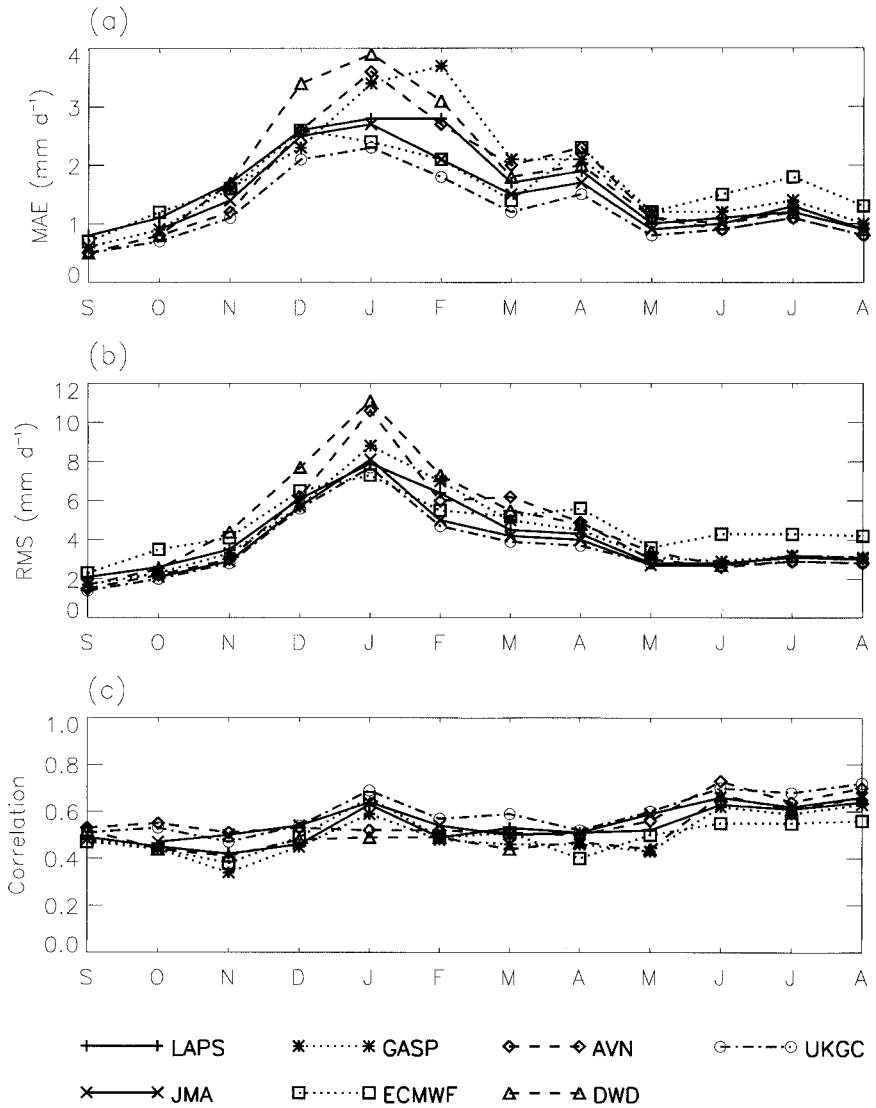

FIG. 7. Time series of monthly average values of (a) mae, (b) rmse, and (c) coefficient of correlation between model and analysis, averaged over the Australian continent for the seven models discussed in the text.

the winter months of September 1997 and June–August 1998.

This aspect of model behavior is further explored in Fig. 3, which shows the bias score as a function of the threshold value chosen to separate the rain/no-rain events. The upper panel is for the three summer months of December to February, while the lower panel is for the winter months of June to August. The ''spaghetti diagram'' nature of the upper panel reveals a large diversity in model behavior. A noteworthy feature of the lower panel is that the two Australian models (LAPS and GASP) have approximately linear curves with a bias score well above one at low thresholds and well below one at higher thresholds. This means that in this winter regime these two models strongly overestimate the areal coverage of rainfall and strongly underestimate the peak values.

Summarizing the results of this diagram and the time series of bias score in Fig. 2, it seems that each model has consistent behavior within a particular (summer vs winter) regime, but there is a very large diversity (or lack of consistency) between models in this rather fundamental aspect of the moisture cycle. In general, however, throughout the range of thresholds most models significantly overestimate the area of rain throughout the summer months, while in winter all models except one underestimate the frequency (or area) of rain at thresholds above 2 mm day^{-1}.

The equivalent diagrams for probability of detection and false alarm ratio as a function of rain threshold

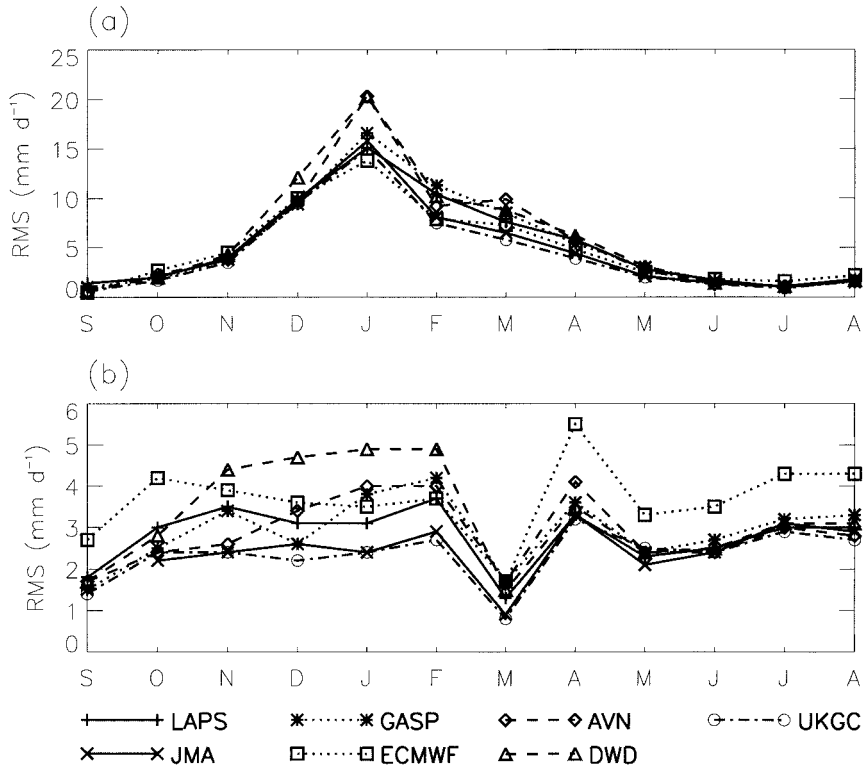

FIG. 8. Time series of monthly average values of rmse for the seven models discussed in the text, averaged over the (a) northern region and (b) southeastern region.

are shown in Figs. 4 and 5. For both figures it is seen that skill is a strong function of threshold, with the probability of detection decreasing from about 80% for rain/no rain (low thresholds) to about 30% or 40% for rain amounts above 20 mm day^{-1}. Consistent with this the false alarm ratio increases with threshold, from about 30% at low threshold to 60%–70% at high thresholds.

These figures also reveal some noteworthy characteristics of individual models. For example, in the summer regime the GASP model shows a much higher FAR than the other models through the range of thresholds, while the ECMWF model has the same characteristic during winter. In the latter case, this high false alarm ratio is apparently a reflection of the fact that the model overestimates the area of rain, as seen on the corresponding bias score diagram (Fig. 3b). In either case, it is clear that various characteristics of the behavior of individual models are brought out very clearly from intercomparisons of these types of simple scores.

Figure 6 shows the two seasonal curves for the HK score as a function of rain threshold. Interestingly, for most models the skill remains relatively constant as a function of threshold for low and moderate threshold values. Above about 10 mm day^{-1} the skill drops rapidly. These curves also show that two of the models (AVN and JMA) clearly outperform the others at rainfall prediction for both summer and winter for threshold values ranging from 0.1 to 5 mm day^{-1}.

Figure 7 shows time series of several point by point measures of skill: mean absolute error (mae), root-mean-square error (rmse), and coefficient of linear correlation. There are a number of comments to be made on interpretation of these curves. The first is that comparison with the HK diagrams in Fig. 6 shows that different scores measure quite different aspects of model behavior. For example the two models (AVN and JMA) that outscored the others on the HK diagram do not stand out as being particularly skillful on the mae, rmse, and correlation diagrams. In fact during the summer months of December to January, one of these (AVN) has among the worst (highest) values of mae and rmse. This suggests these two models have relatively greater success in forecasting the presence of rain, but that the amount of rainfall is incorrect.

The second comment is that all three scores in Fig. 7 reveal a large seasonal variation with highest skill in winter and lowest skill in summer. For the mae and rmse scores, this seasonal behavior is largely a result of poor model skill during the wet season of the summer/wet–winter/dry monsoonal regime over northern Australia. To illustrate, Fig. 8 shows rmse time series for two subregions of the continent. The upper curve is for tropical northern Australia while the lower curve

TABLE 3. Total number of points at which rain occurred (i.e., the sum of $M + H$ from Table 1) for each rainfall threshold, as plotted in Figs. 10, 11, 13, 14, 15, and 18.

Threshold (mm)	0.1	1	2	5	10	20	50
Southeast, summer	7521	3940	2873	1573	778	234	14
Southeast, winter	9321	5377	4025	2304	1201	379	38
North, summer	8984	6818	5753	3924	2409	1106	218
North, winter	1486	637	449	224	112	49	15
No. of days (summer)	72						
No. of days (winter)	92						
No. of grid points (southeast)	222						
No. of grid points (north)	147						

is for the high-rainfall subtropical southeastern portion of the country (as delineated in Fig. 1). In terms of number of grid points, these subregions occupy 24% and 37%, respectively, of the total domain. It is seen that in the southeast (Fig. 8b) rms errors are of the order of 2–3 mm day^{-1} throughout the year. Thus the high values of 6–10 mm day^{-1} in summer over the whole continent (Fig. 7b) are brought about by the large values during the summer wet season in northern Australia (Fig. 8a).

Due to this very different behavior of the models in the two different (tropical/subtropical) regimes, we decided the most sensible verification exercise would be to score the two different regions separately. This is done in the following two sections. To aid interpretation, Table 3 gives an indication of the sample size for each threshold value. For the southeastern region, for example, the total sample size equals the number of grid points (222) multiplied by the number of days (72 in summer) equaling a sample size of 15 984 (observation, forecast) pairs. The numbers in the table give the number of grid points in the sample at which rain occurred above the given threshold. It can be seen that for the first three rows the sample is quite large up to a threshold value of 10 mm of rain; thus, meaningful interpretation can be given to the graphs and tables to follow at least up to this value. For the northern region in winter, however, the sample of rain events is quite small, such that for this regime care should be taken in interpretation of the statistics above the lowest threshold.

4. Verification over northern Australia

Northern Australia is characterized by a tropical monsoonal regime. It has a very large seasonal cycle with almost all precipitation (as measured by volume) occurring in the six months from November to April (Fig. 9). The meteorology of the region has been reviewed by Manton and McBride (1992). The predominant rain-producing mechanism is through cumulonimbus convection and its associated altostratus anvils, and is often associated with weather systems such as tropical cyclones, monsoon depressions, and tropical squall lines. The ability of the models to accurately simulate precipitation in this regime will depend on such features as their initialization of tropical moisture, their initialization of divergence, and particularly on their representation of cumulonimbus convection.

Figure 9 shows time series for the seven models of mean rain area, rain rate, and their product, rain volume, through the verification period. As was the case over the continent, four models (GASP, JMA, LAPS, and AVN) consistently overestimate the areal extent or frequency of precipitation throughout the entire wet season. There is a wide variation in mean rain rate between models, though the three that perform well at rain area either strongly underestimate (UKGC and ECMWF) or strongly overestimate (DWD) rain rate.

The threshold-dependent curves of POD, FAR, and HK for the summer months where frequent convective rain predominates are shown in Fig. 10. For this regime the POD values are very high with two models (AVN and JMA) detecting more than 90% of rain occurrences at a threshold level of 1 mm day^{-1}. As before, POD decreases and FAR increases strongly as a function of model threshold. The HK scores, however, are relatively constant through the range of threshold values (Fig. 10c). Despite the high values for POD, the HK scores are very low, revealing a low level of skill of the models in this tropical convective regime. The high frequency of rain events in this regime means that Z is not significantly greater than F in Eq. (5), so the false alarms have a large effect on this score. For the general threshold of 1 mm day^{-1}, the models that perform best are the DWD and UKGC, primarily because they have the smaller false alarm ratios compared to the other models. The model that performs worst in this regime is GASP, with a high false alarm ratio again being the main factor separating it from the other models.

Figure 11 shows the same set of threshold-dependent curves for the winter months of June to August. As seen previously (Fig. 9) this is a time of very low rainfall in this region. The most noticeable aspect of Fig. 11 is its very different character from the equivalent graphs for the whole continent (Figs. 4b, 5b, and 6b) and from the equivalent graphs for summer (Fig. 10). In particular, the characteristic decrease of POD and increase of FAR with increasing threshold is only weakly present. The graph of the true skill or HK score is very similar in appearance to that for POD reflecting the very large

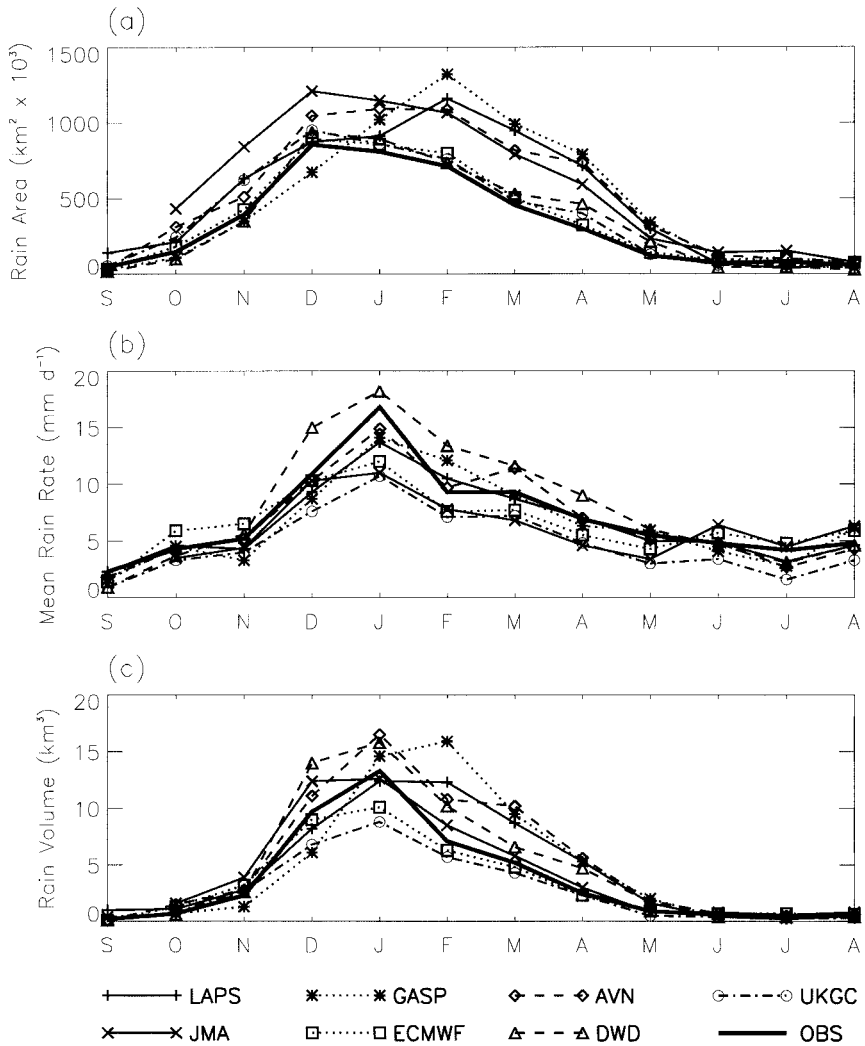

FIG. 9. Monthly mean time series for the seven models averaged over the northern region of (a) rain area, (b) rain rate, and (c) rain volume. Also shows are the observed (analysis) values.

value of Z compared to F in Eq. (5). This is consistent with the observation by Wilks (1995) that ''the contributions made to the HK score by a correct 'no' or 'yes' forecast increases as the event is more or less likely, respectively.'' Thus when this score is applied to human forecasters, they ''are not discouraged from forecasting rare events on the basis of their low climatological probability alone'' (Wilks 1995). The cluster of HK curves in the lower panel of Fig. 11 indicates an overall level of skill that is greater than for the high-rainfall regime in summer (Fig. 10c). There is a also a large difference between the various models for this regime, with two models (AVN and JMA) clearly outscoring the others. Despite the relative rareness of a rainfall event in this regime, the better skill of AVN and JMA is probably significant as they were consistent in outscoring the other models for each of the three months making up the average in Fig. 11.

5. Verification over southeastern Australia

Southeastern Australia is the most densely populated part of the country and has the greatest density of rainfall observations, especially around the coastal fringe. As seen in the time series of rain volume (Fig. 12), rain amounts are much smaller than for the northern region wet season. However, considerable rain falls all year round, with winter having approximately 50% more than summer. The meteorology of the region is subtropical, and rain-bearing weather systems include cold fronts, tropical–extratropical cloud bands, cutoff lows, east coast lows, summertime convection in the humid trade easterlies, and wintertime convection in the cold air behind a front. Some important studies of major rain events in the region have been carried out by Tapp and Barrell (1984), Zhao and Mills (1991), Mills and Russell (1992), McInnes and Hess (1992), Mills and Wu (1995), and Wright (1989, 1997).

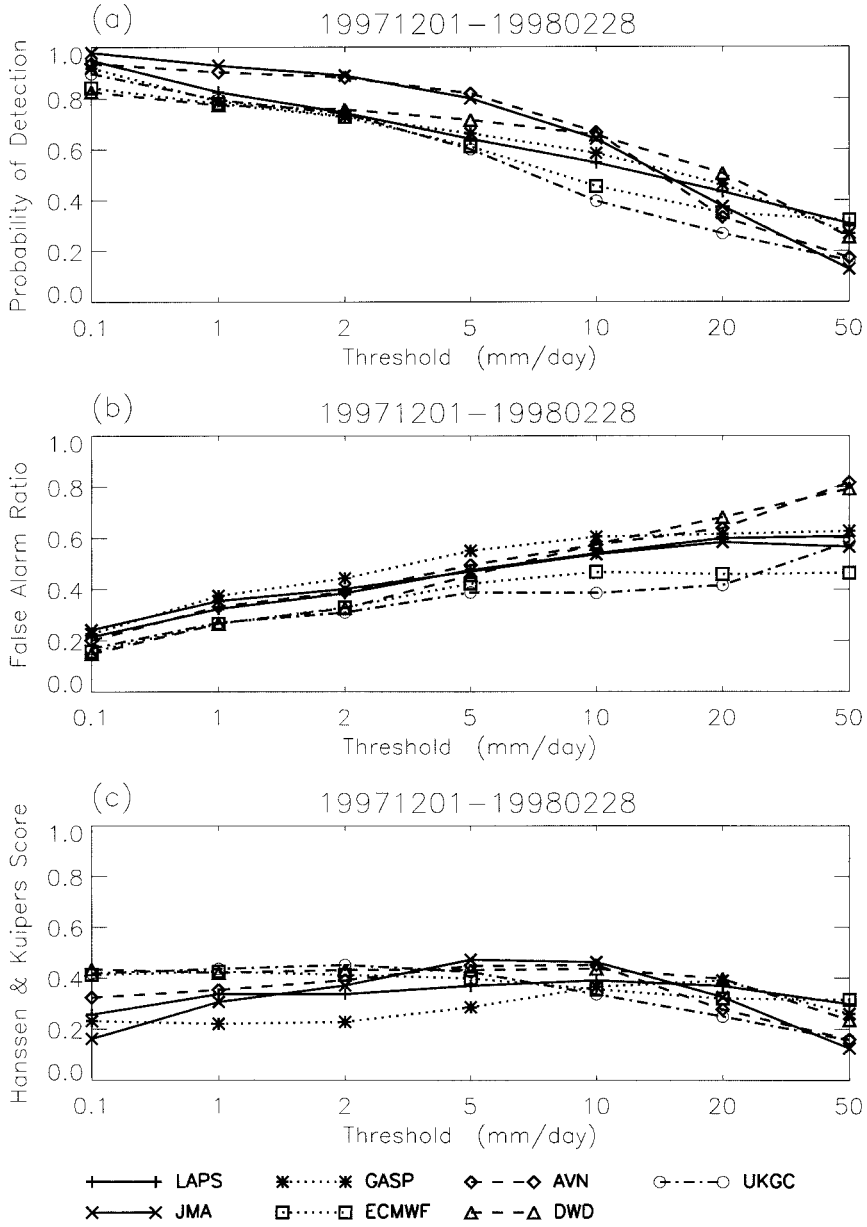

FIG. 10. Space–time average values over the northern region for the summer wet season (Dec–Feb) of (a) probability of detection, (b) false alarm ratio, and (c) HK score for the seven models.

As seen in Fig. 12, in general the models overestimate rainfall in summer and underestimate it in winter; this is reflected in the values of bias score for the two seasons (Fig. 13), which have summer values greater than one (Fig. 13a) and winter values less than one (Fig. 13b). An interesting aspect of Fig. 13a is the very large divergence between models, which increases with increasing threshold. This implies that there are major deficiencies in the operational models in the simulation of major or heavy rain events in summer.

Figure 14 shows the threshold-dependent POD, FAR, and HK curves for summer. At thresholds above 2 mm day^{-1}, the POD (Fig. 14a) and HK (Fig. 14c) curves parallel one another, implying that the POD or accuracy for events is the main determiner of skill. The overall skill (as measured by HK) is much higher than for the tropical north in the previous section, implying a greater ability of the models to simulate rainfall associated with subtropical weather systems. For most of the range of thresholds, the AVN model outperforms most others, and this is essentially due to its better POD scores. It is interesting that this is one of the two models that

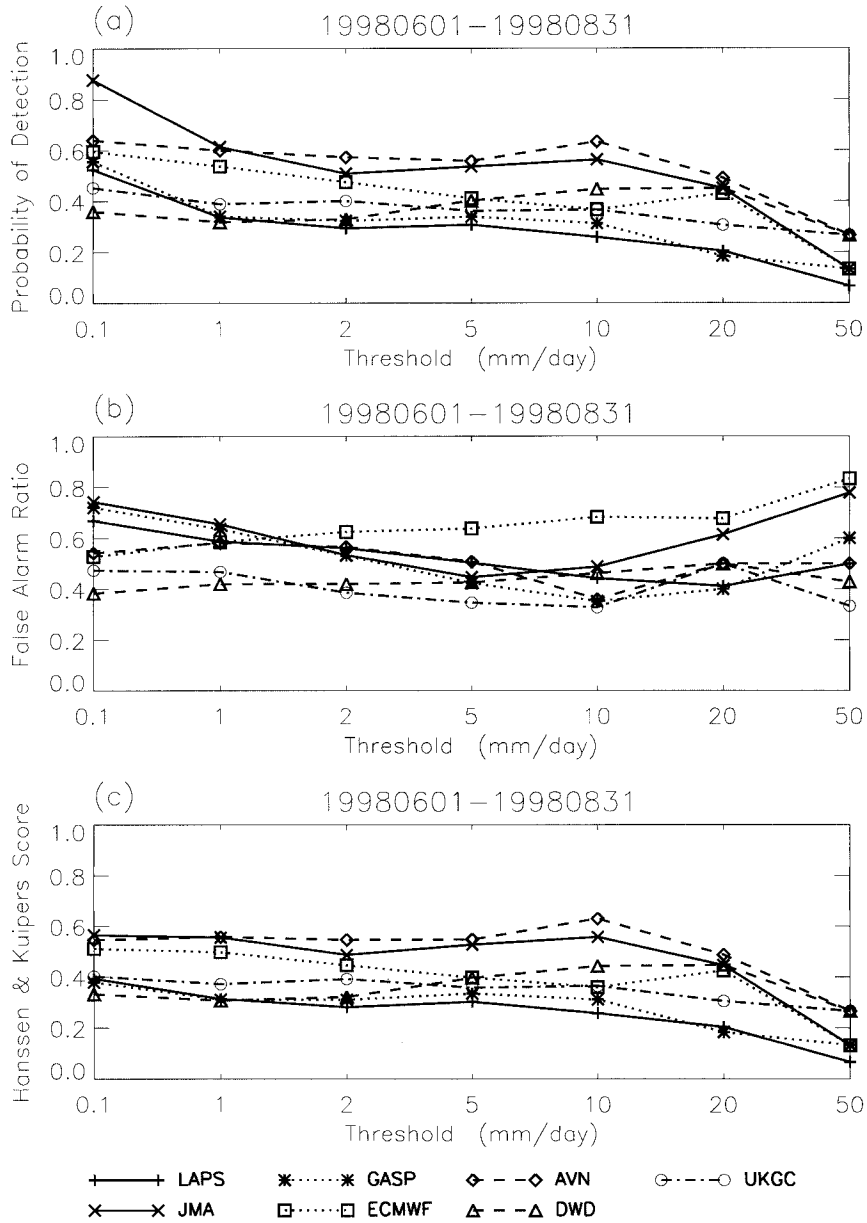

FIG. 11. As in Fig. 10 but for the winter dry season (Jun–Aug) in the northern region.

demonstrated superior skill in the northern winter (previous section).

Figure 15 shows the same set of verification statistics for the winter regime. The skill level is quite similar to that for the summer (Fig. 14), though the HK scores for the better models are slightly higher in winter. This improved skill is associated with lower false alarm ratios in winter. This mimics the behavior in the northern summer wet season and, presumably, reflects model deficiencies in simulating summertime convective systems. Once again the AVN model is superior to the others for most of the range of thresholds.

As an aside, Fig. 15 gives an example of how these categorical statistics can clearly reveal characteristics of particular models. This is seen in the FAR graph (Fig. 15b), where the ECMWF model has a consistently higher value than the other models. This behavior penalizes that model in the HK score (Fig. 15c). It is associated with the fact that in this winter regime, the ECMWF model overestimates the area (or frequency) of rain as reflected by the bias score (Fig. 13b).

6. Summary and discussion

Verification statistics have been presented for 24-h precipitation forecasts over the Australian continent, for

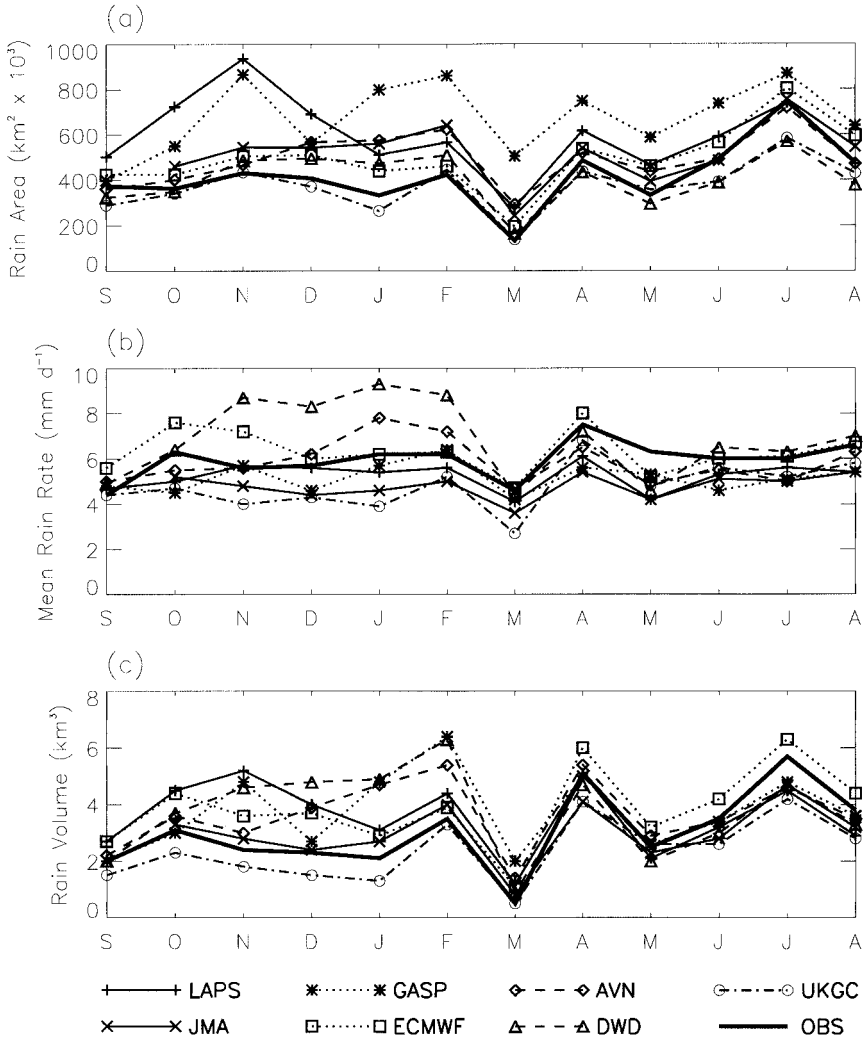

FIG. 12. As in Fig. 9 but for the southeastern region.

seven operational NWP models. All model rainfall forecasts have been mapped to a 1° latitude–longitude grid and have been verified against an operational rainfall analysis, mapped to the same grid. The verification has focused on two large subregions: the northern tropical monsoon regime, and the southeastern subtropical regime.

For both regimes the operational models tend to overestimate the rainfall in summer and to underestimate it in winter. Model skill is much greater for the subtropical regime and is actually quite low in the northern regime. During the summer wet season the low skill is associated with an inability to simulate the behavior of tropical convective rain systems. During the winter dry season, it is associated with a low probability of detection for the occasional rainfall event.

For both regimes, model skill (as measured by the HK score) falls dramatically for thresholds greater than 10 mm day^{-1}. This implies that the models are much better at predicting the occurrence of rain than they are at predicting the magnitude and location of the peak values.

For comparison with other studies, and for a baseline for future models, Tables 4 and 5 summarize the current level of skill for the seven models. Table 4 is for a threshold value of 1 mm day^{-1}, while Table 5 is for a threshold of 10 mm day^{-1}. To put these scores into perspective, the HK score for a persistence forecast at each grid point is given in the last column of both tables.

For a subtropical (southeastern) regime, current models have POD values between 0.6 and 0.8 in both winter and summer. The FAR values have a wide range between models and FAR seems to be the main factor differentiating models in the subtropics. The HK scores are in the range of 0.5–0.7 and are slightly higher for winter than for summer. The HK scores very easily outscore persistence in the subtropics for both the 1 mm day^{-1} (Table 4) and the 10 mm day^{-1} threshold (Table 5). Thus

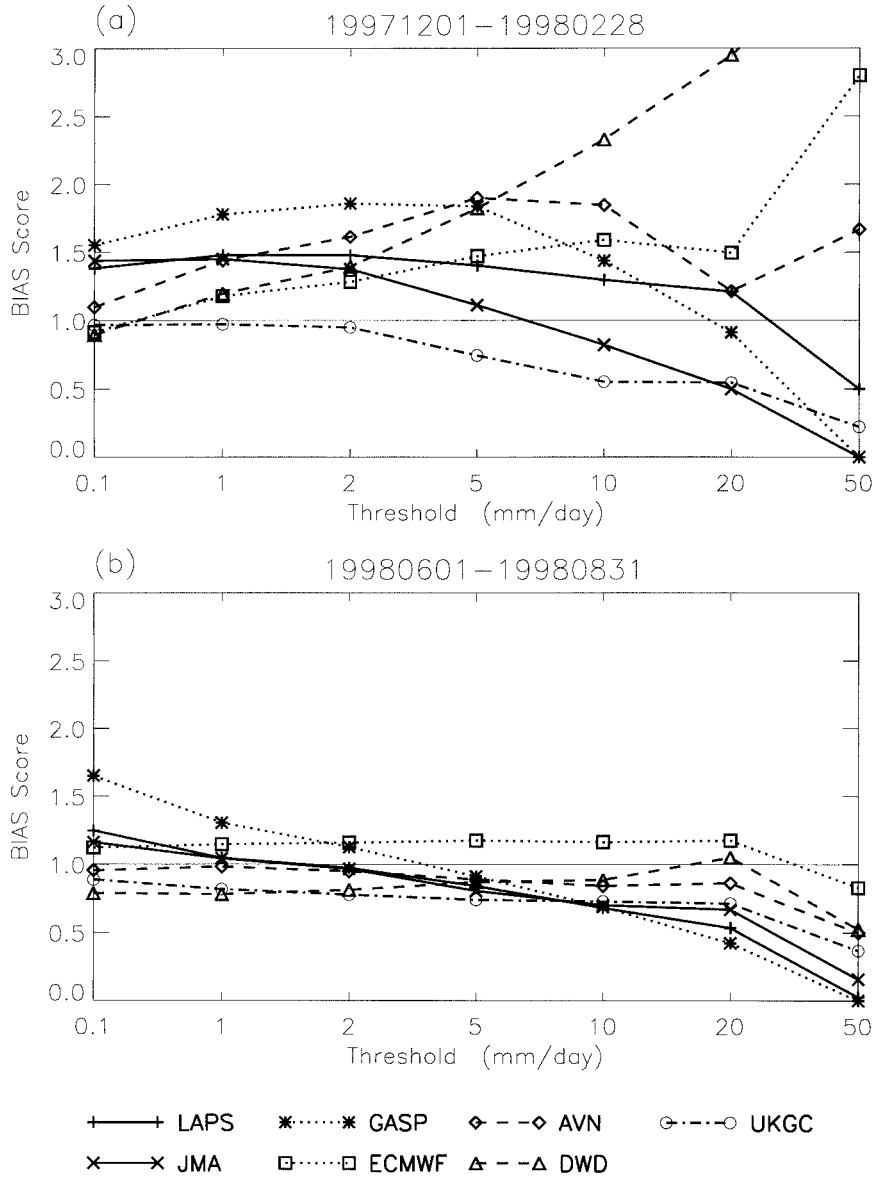

FIG. 13. Space–time average values of bias score over the southeastern region as a function of rain threshold for (a) summer (Dec–Feb) and (b) winter (Jun–Aug).

for the current operational NWP models, the 24-h rain forecasts can be considered quite skilful.

For the tropical regime (upper two rows of each table), model skill is much lower. For the summer wet season, POD values are high (0.8–0.9), but in this tropical convective regime, false alarms play an important role so that HK scores are only around 0.2–0.4 and are comparable to persistence. Thus it could be said the models have no real skill at rainfall forecasts in this monsoonal wet season regime.

It is noteworthy that over the four regimes, one model stands out as being the most skillful. This is the AVN model. For the eight upper-range values of HK listed in the HK column in Tables 4 and 5, five were scored by the AVN. Of the remaining three ''best scores,'' two were obtained by the JMA model. As follow-up work toward the improvement of model rainfall forecasts, it would be worth investigating the differences between the moisture cycles of the AVN and JMA models and the other five operational models.

A cautionary note is made to the extent that these conclusions are all in the context of using the HK score as the measure of skill. The referees of this paper advocated the use of the equitable threat score (or ETS), which was proposed by Schaefer (1990) and is in common usage, particularly in the United States (e.g., Mesinger et al. 1990; Mesinger 1996). The ETS has some philosophical differences to the HK score in that it focuses more on event

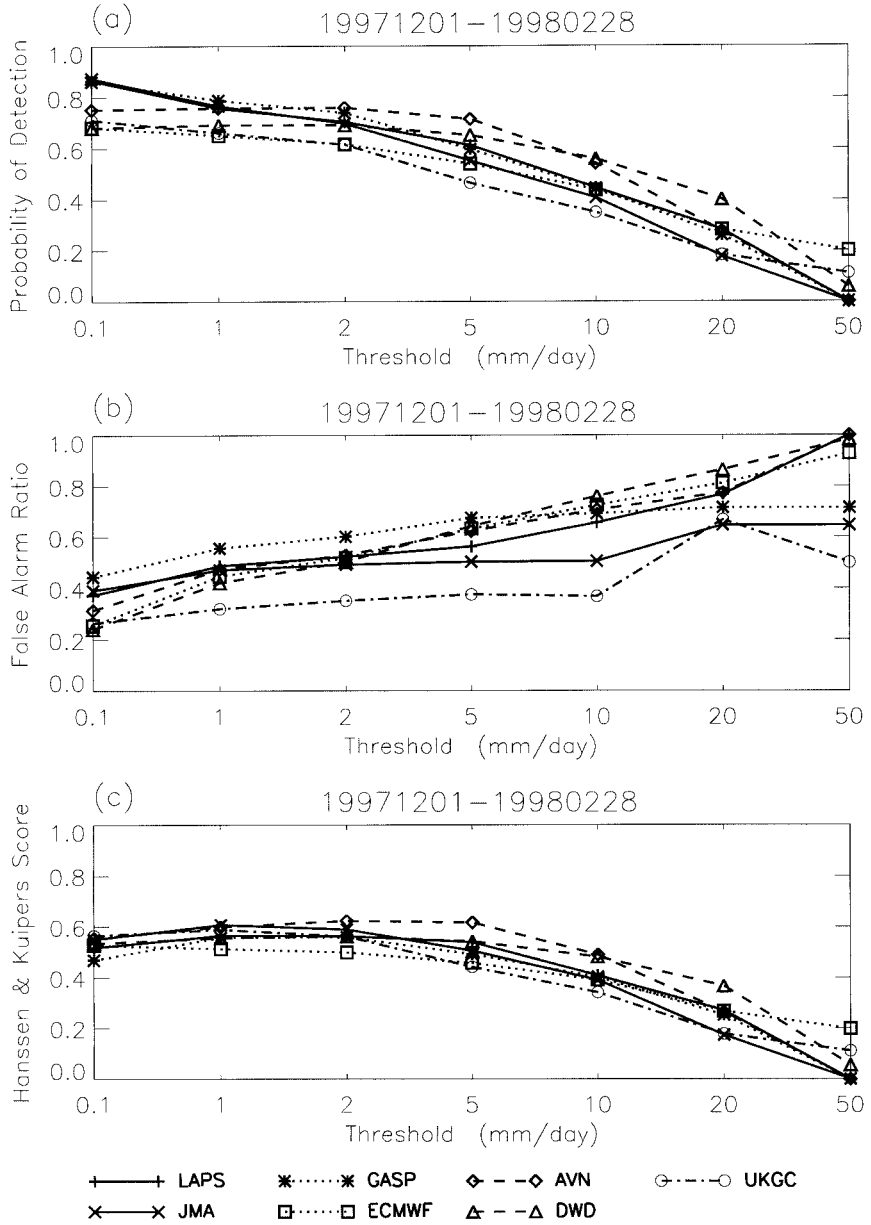

FIG. 14. Space–time average values over the southeastern region for the summer season (Dec–Feb) of (a) probability of detection, (b) false alarm ratio, and (c) HK score for the seven models.

forecasts and has less emphasis on successful nonevent forecasts. One disadvantage perceived by the current authors is that the reference accuracy for a random forecast in the ETS is dependent on the properties of the model being verified. The relative advantages of various measures of skill is a complex issue, however, and is beyond the scope of the current study.

Given the common usage of the ETS score, ETS scores for the four regimes were calculated for completeness. These are presented in Fig. 16, which are the equivalent figures to Figs. 10c, 11c, 14c, and 15c, but with ETS replacing the HK score in each case. The numerical values of ETS differ from those of HK, but otherwise the general character of the comparison between regimes and between models follows that already discussed. The main difference is for the southeastern regime in summer (Fig. 16c) where, compared with the equivalent figure for HK, the UKGC has increased greatly in skill and the AVN has decreased in skill. Inspection of Fig. 14b shows that these differences can be attributed to the greater sensitivity of the ETS to the number of false alarms, *F*.

Returning to the HK score as the measure, an alter-

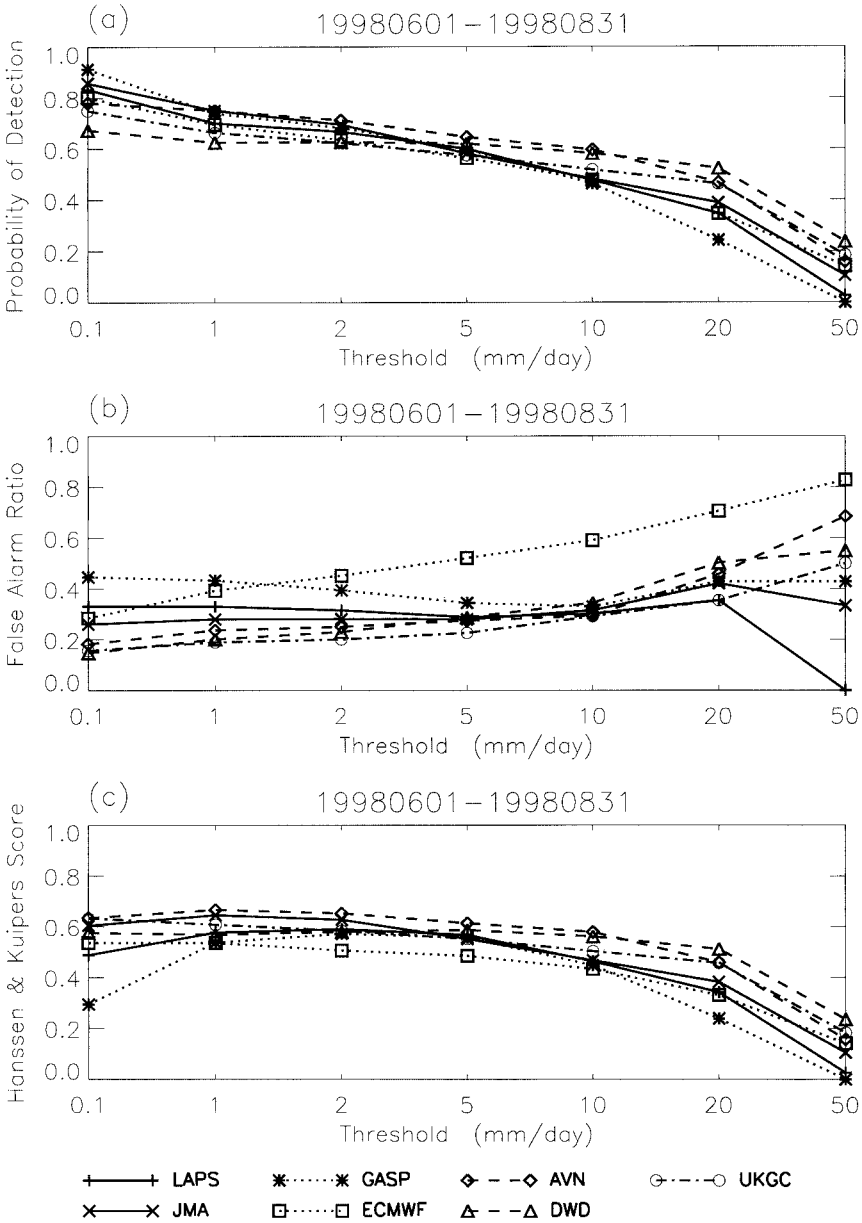

FIG. 15. As in Fig. 14 but for the winter season (Jun–Aug) in the southeastern region.

TABLE 4. The range of values of seven operational models over the summer and winter seasons for two subdomains, for a rain threshold value of 1 mm day^{-1}. Also shown in the last column are values of the HK score for a persistence forecast.

	POD	FAR	HK	HK for persistence
North, summer (wet)	0.78–0.93	0.27–0.38	0.21–0.44	0.34
North, winter (dry)	0.32–0.61	0.42–0.65	0.31–0.56	0.34
Southeast, summer	0.63–0.79	0.32–0.57	0.51–0.62	0.31
Southeast, winter	0.62–0.75	0.19–0.44	0.54–0.67	0.33

TABLE 5. The range of values of seven operational models over the summer and winter seasons for two subdomains, for a rain threshold value of 10 mm day^{-1}. Also shown in the last column are values of the HK score for a persistence forecast.

	POD	FAR	HK	HK for persistence
North, summer (wet)	0.40–0.67	0.39–0.61	0.34–0.46	0.30
North, winter (dry)	0.26–0.63	0.33–0.68	0.26–0.63	0.31
Southeast, summer	0.35–0.56	0.37–0.76	0.34–0.48	0.18
Southeast, winter	0.46–0.60	0.29–0.59	0.43–0.58	0.15

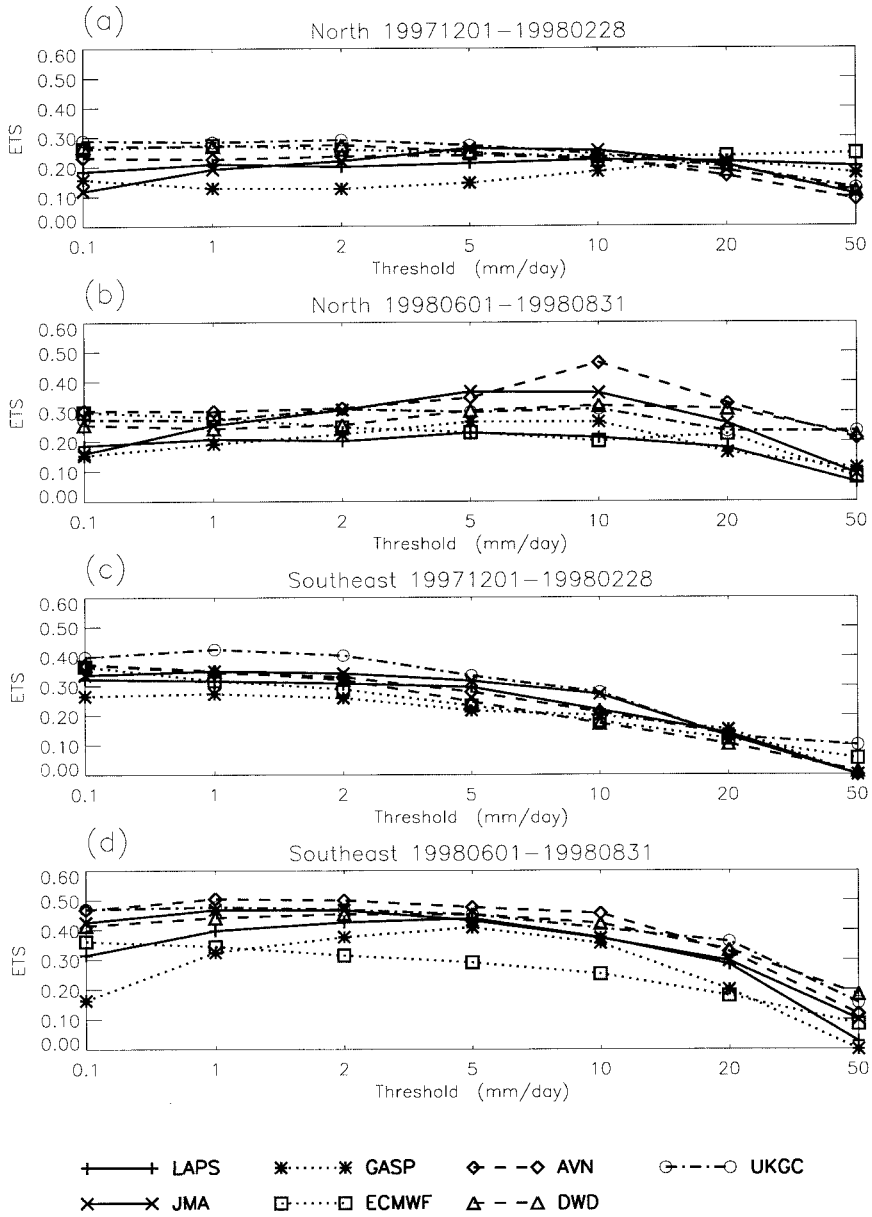

FIG. 16. ETS as a function of rain threshold for the seven forecast models: (a) for the north region in summer, (b) for the northern region in winter, (c) for the southeastern region in summer, (d) for the southeastern region in winter.

native framework for summarizing the results of this study is shown in Fig. 17. This plots the skill of every model for all four regimes in (*accuracy for events, accuracy for nonevents*) phase space. In this case the plot is for a rain threshold of 1 mm day^{-1}. The four regimes are denoted by different symbols, as given in the key. The overall model skill can be measured by HK value, which is marked by the labeled isopleths. The values for persistence are indicated by filled symbols.

This figure shows that for the two subtropical regimes HK scores are between 0.5 and 0.7, with both accuracy for events and accuracy for nonevents playing a role in determining comparative model skill. For the subtropics (southeast), the model superiority over persistence is totally associated with the models' skill at accuracy for events (or POD). The low values of HK score for the tropical wet season can be seen to be due to a poor value of accuracy for nonevents. For the tropical dry season, the situation is reversed with all models having a high accuracy for nonevents, so that POD is the determiner of skill.

The equivalent diagram is shown in Fig. 18 for a threshold of 10 mm day^{-1}. For this threshold an event is a much rarer occurrence so that Z (the number of accurate no-rain

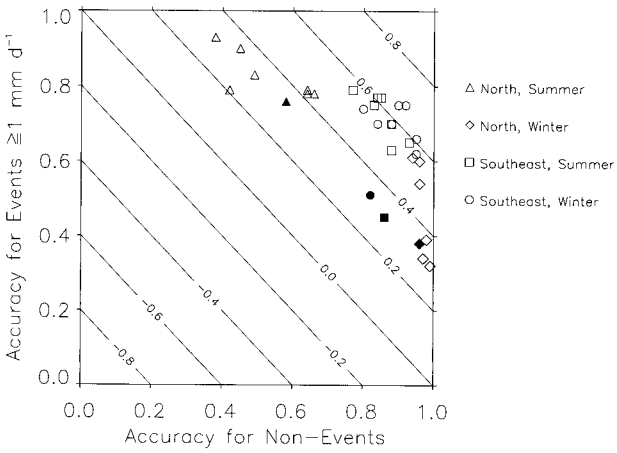

FIG. 17. Skill of the models for four rainfall regimes for a rain/no-rain threshold of 1 mm day^{-1}. The skill is plotted in (*accuracy for events, accuracy for nonevents*) phase space. The contours denote isolines of the HK score, with maximum skill (HK = 1) in the top-right corner. The value for persistence in each regime is marked by the equivalent solid symbol.

forecasts) far exceeds F (the number of false alarms) in Eq. (5). Thus the accuracy for nonevents is always greater than 90% and the accuracy for events (or POD) is the determiner of skill. The distribution of scores along the right-hand y axis of the box in Fig. 18 reveals a wide distribution between models at POD at this higher rain threshold. The exception to this discussion is the northern wet/summer regime, depicted by triangles. For this convective regime at high thresholds, the accuracy for nonevents is often near 80%. Thus, even at this high threshold the number of false alarms F is of the order of 25% of the number of correct no-rain forecasts Z.

As a final comment, this paper has shown the usefulness of categorical statistics, calculated as a function of rain threshold, in forecast verification. Tables 4 and 5, and Figs. 17 and 18, give a concise summary of many aspects of the skill of current NWP models at 24-h forecasts of precipitation. Using these values as baselines for comparison, questions of immediate interest are the degree to which model skill decreases as forecasts are extended to two days and beyond, and the comparative skills of mesocale models when mapped onto this standard 1° grid. These aspects will be the subject of immediate follow-on studies.

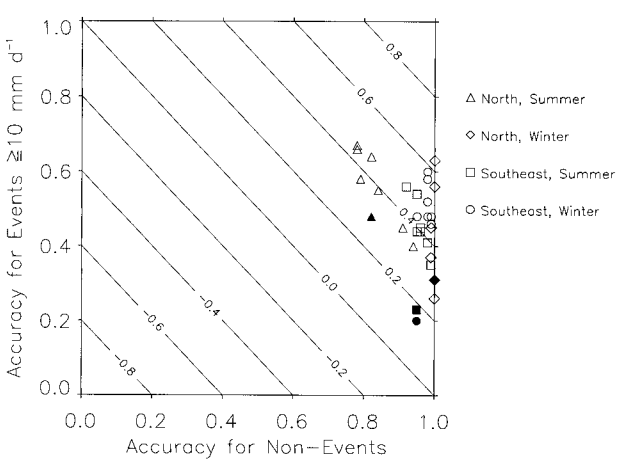

FIG. 18. As in Fig. 17 but for a rain threshold of 10 mm day^{-1}.

REFERENCES

Doswell, C. A., III, R. Davies-Jones, and D. L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting,* **5,** 576–585.

Ebert, E. E., and J. L. McBride, 1997: Methods for verifying quantitative precipitation forecasts: Application to the BMRC LAPS model 24-hour precipitation forecasts. BMRC Techniques Development Rep. 2, 87 pp. [Available from Bureau of Meteorology Research Centre, GPO Box 1289K, Melbourne 3001, Australia.]

——, and G. T. Weymouth, 1999: Incorporating satellite observations of ''no rain'' in an Australian daily rainfall analysis. *J. Appl. Meteor.,* **38,** 30–38.

Hanssen, A. W., and W. J. A. Kuipers, 1965: On the relationship between the frequency of rain and various meteorological parameters. *Meded. Verh.,* **81,** 2–15.

Johnson, L. E., and B. G. Olsen, 1998: Assessment of quantitative precipitation forecasts. *Wea. Forecasting,* **13,** 75–83.

Manton, M. J., and J. L. McBride, 1992: Recent research on the Australian monsoon. *J. Meteor. Soc. Japan,* **70,** 275–285.

McInnes, K. L., and G. D. Hess, 1992: Modifications to the Australian region limited area model and their impact on an east coast low event. *Aust. Meteor. Mag.,* **40,** 21–31.

Mesinger, F., 1996: Improvements in quantitative precipitation forecasts with the Eta regional model at the National Centers for Environmental Prediction: The 48-km upgrade. *Bull. Amer. Meteor. Soc.,* **77,** 2637–2649.

——, T. L. Black, D. W. Plummer, and J. H. Ward, 1990: Eta model precipitation forecasts for a period including Tropical Storm Allison. *Wea. Forecasting,* **5,** 483–493.

Mills, G. A., and I. Russell, 1992: The April 1990 floods over eastern Australia: Synoptic description and assessment of regional NWP guidance. *Wea. Forecasting,* **7,** 636–668.

——, and B.-J. Wu, 1995: The 'Cudlee Creek' flash flood—An example of synoptic-scale forcing of a mesoscale event. *Aust. Meteor. Mag.,* **44,** 201–218.

Murphy, A. H., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.,* **115,** 1330–1338.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting,* **5,** 570–575.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. World Weather Watch Tech. Rep. 8, WMO/TD No. 358, World Meteorological Organization, 114 pp.

Tapp, R. G., and S. L. Barrell, 1984: The north-west Australian cloud band: Climatology, characteristics and factors associated with development. *J. Climatol.,* **4,** 411–424.

Weymouth, G., G. A. Mills, D. Jones, E. E. Ebert, and M. J. Manton, 1999: A continental-scale daily rainfall analysis system. *Aust. Meteor. Mag.,* **48,** 169–179.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences. An Introduction.* Academic Press, 467 pp.

Woodcock, F., 1976: The evaluation of yes/no forecasts for scientific and administrative purposes. *Mon. Wea. Rev.,* **104,** 1209–1214.

Wright, W. J., 1989: A synoptic climatological classification of winter precipitation in Victoria. *Aust. Meteor. Mag.,* **37,** 217–219.

——, 1997: Tropical–extratropical rainbands and Australian rainfall: I. Climatology. *Int. J. Climatol.,* **17,** 807–829.

Zhao, S., and G. A. Mills, 1991: A study of a monsoon depression bringing record rainfall over Australia. Part II: Synoptic–diagnostic description. *Mon. Wea. Rev.,* **119,** 2074–2094.