

THE WGNE ASSESSMENT OF SHORT-TERM QUANTITATIVE PRECIPITATION FORECASTS

BY ELIZABETH E. EBERT, ULRICH DAMRATH, WERNER WERGEN, AND MICHAEL E. BALDWIN

Even as operational numerical weather prediction models become more accurate, improvements in rain forecasts are hard to come by.

Of all the weather elements for which forecasts are provided to the public, rainfall is perhaps of the greatest interest. While most people simply want to know whether they will need an umbrella that day, there is growing demand from industry, agriculture, government, and many other sectors for more detailed rainfall predictions. Unfortunately, rainfall is certainly among the most difficult weather elements to predict correctly. Rainfall has greater spatial and temporal variability than most other meteorological quantities of interest. Many processes can lead to rain, including large-scale ascent of moist air, convection caused by heating of moist air near the surface, con-

vergence of moist air in a baroclinic zone, and orographic lifting. These processes must all be represented in numerical weather prediction (NWP) models, whose output forms the basis for most rainfall forecasts. It is of great interest to assess how well we can meet the need for timely and accurate rainfall forecasts using operational NWP models.

In the mid-1990s, the Working Group on Numerical Experimentation (WGNE), established under the World Meteorological Organisation's World Climate Research Programme (WCRP) and Commission for Atmospheric Sciences (CAS), turned its attention to quantitative precipitation forecasts (QPFs). Since accurate prediction of rainfall depends critically on the accurate prediction of atmospheric motion and moisture content, it is reasonable to expect that a good forecast of rainfall over a large domain indicates a good forecast overall. Indeed, many operational centers use QPF skill as a critical measure of model health. Accumulated precipitation can be verified (albeit imperfectly, given its highly variable nature) using rain gauge networks. Knowledge of a model's QPF behavior not only helps model developers but also users of the QPFs to understand the reliability of the model output.

At the 10th annual WGNE meeting it was recommended that QPFs from several operational NWP models be evaluated in different areas of the globe

AFFILIATIONS: EBERT—Bureau of Meteorology Research Centre, Melbourne, Australia; DAMRATH AND WERGEN—Deutscher Wetterdienst, Offenbach am Main, Germany; BALDWIN—National Severe Storms Laboratory, Norman, Oklahoma
A supplement to this article is available online (DOI: 10.1175/BAMS-84-4-Ebert)

CORRESPONDING AUTHOR: Dr. Elizabeth Ebert, Bureau of Meteorology Research Centre, GPO Box 1289K, Melbourne 3001, Australia

E-mail: e.ebert@bom.gov.au

DOI: 10.1175/BAMS-84-4-481

In final form 11 October 2002
©2003 American Meteorological Society

(WCRP 1995). As a result, starting in 1995 verification of QPFs from a number of global and regional operational NWP models was undertaken at the National Centers for Environmental Prediction (NCEP) in the United States, and the Deutscher Wetterdienst (DWD) in Germany. The Bureau of Meteorology Research Centre (BMRC) in Australia began verifying a number of NWP QPFs in 1997. The Met Office (UKMO) joined the effort in 2000, MeteoFrance in 2001, and the Japan Meteorological Agency (JMA) in 2002. These countries extend over tropical, subtropical, and mid latitudes, and include a variety of regimes ranging from very wet to very dry, with rainfall scales ranging from thunderstorm to monsoon scale.

This paper reports on the findings of the WGNE QPF assessment over the United States, Germany, and Australia from 1997 to 2000. We concentrate on the evaluation of deterministic forecasts of 24-h accumulated rain made from operational NWP models since that is the quantity most easily verified using existing rain gauge data. Although ensemble forecasting is playing an increasing role in QPF, particularly in probabilistic forecasts, it has not been evaluated in the WGNE study to date. The emphasis is on global NWP models, although a few regional models were also included. High-resolution mesoscale models were not considered in this study, even though they have the potential to provide more

TABLE 1a. Eight global NWP models verified in this study. The plain type refers to the NCEP verification over the United States, the bold type to the DWD verification in Germany, and the italics to the BMRC verification over Australia. An asterisk signifies that the resolution received is coarser than the original model resolution.

Center/global model	Horizontal resolution received (°lat, °lon)	Initial time (UTC)	Forecast periods (h)	Verified since
Australian Bureau of Meteorology (ABOM)	<i>0.75° × 0.75°^a</i>	2300	24, 48, 72	<i>Jul 1997</i>
Canadian Meteorological Centre (CMC)	0.9° × 0.7° 1.0° × 1.0°*	0000, 1200 0000	24, 36, 48, 60, 72, 84, 30, 54, 78	Jan 1996 Jan 1997
Deutscher Wetterdienst (DWD)	0.75° × 0.75° ^b 0.75° × 0.75°^b <i>0.75° × 0.75°^b</i>	0000, 1200 0000 <i>0000</i>	24, 36, 48 30, 54, 78 <i>24, 48</i>	Apr 1995 Jan 1997 <i>Sep 1997</i>
European Centre for Medium Range Forecasts (ECMWF)	0.5° × 0.5° 0.5° × 0.5° <i>0.5° × 0.5°^c</i>	1200 1200 <i>0000</i>	24, 48, 72 42, 66, 90 <i>24, 48</i>	Feb 1996 Jan 1997 <i>Sep 1997</i>
Japan Meteorological Agency (JMA)	1.25° × 0.25° <i>2.5° × 2.5°*</i>	1200 <i>0000</i>	42, 66, 90 <i>24, 48, 72</i>	Aug 2000 <i>Oct 1997</i>
Meteo France (FRA)	0.25° × 0.25°^d	0000	30, 54, 78	Jul 1998
Met Office (UKMO)	0.83° × 0.56° <i>1.25° × 1.25°*</i>	0000 <i>0000</i>	30, 54, 78 <i>24, 48</i>	Jul 1998 <i>Sep 1997</i>
National Centers for Environmental Prediction (NCEP) Aviation (AVN) model	0.7° × 0.7° ^e 0.94° × 0.94° <i>1.25° × 1.25°*</i>	0000, 0600 1200, 1800 0000 <i>0000</i>	24, 30, 36, 42 48, 54, 60, 66, 72 30, 54, 78 <i>24, 48</i>	Dec 1992 Jan 1997 <i>Sep 1997</i>

^aGrid resolution of 2.5° received until Nov 1998.

^bGrid resolution of 1.5° received until Dec 1999.

^c36-h forecast from 1200 UTC run replaced 0000 UTC based 24-h and 48-h forecasts starting Nov 1999.

^dVariable grid is 0.25° resolution over Europe, coarser resolution elsewhere.

^eGrid resolution of 0.94° received until Jan 2000.

detailed and accurate rain forecasts in limited domains.

QPFs from eight global and three regional NWP models, listed in Tables 1a and 1b, were examined. In addition to the operational models run in the United States, Germany, and Australia, models from the UKMO, European Centre for Medium-Range Forecasts (ECMWF), MeteoFrance, Canadian Meteorological Centre (CMC), and JMA were also verified. The QPFs were received either over the Global Telecommunications System (in which case the spatial resolution was lower than the resolution at which the model was run), or by individual FTP arrangement with the originating operational center. Note that not all models were verified over all three domains, nor were the spatial resolutions and initial times consistent across domains. The difference in local rain gauge observing times dictates that a different set of forecast periods be verified in each domain. Thus, even within a given domain not all models start on an even footing. We therefore do not attempt to judge which model performed “best” or “worst.”

The forecasts were examined at daily, monthly, and seasonal timescales, over both national and regional domains. Aspects of the forecasts that have been verified include the spatial and temporal frequency of rain, the rain-rate distribution, and the correspondence of the forecast rain pattern with the observed pattern. This comprehensive and continuing assessment of model QPFs has revealed much about the skill and behavior

of the models in different regimes, and provides the participating agencies with useful information to guide them in improving their operational models.

VERIFICATION METHODS. *Verification measures.* Unlike most other meteorological fields, rainfall has both an on/off component and a quantitative component. Thus we need to assess the model’s ability to predict rain occurrence as well as rain amount. Stanski et al. (1989) and Wilks (1995) thoroughly describe and discuss the many statistics applicable to QPF verification. This study will focus primarily on the bias score, which measures the ratio of the frequency of forecast rain to the frequency of observed rain, and the equitable threat score (ETS), which measures the fraction of observed and/or forecast events that were correctly predicted, adjusted for correct predictions due to random chance. (For a discussion of the most commonly used statistics, see <http://dx.doi.org/10.1175/BAMS-84-4-Ebert.>)

To put the skill of the model QPFs into context, model verification statistics should also be compared to verification statistics for an unskilled forecast such as persistence or climatology. The persistence forecast is simply the observed rainfall from the previous forecast period, and is the most commonly used unskilled forecast when evaluating QPFs. “Climatology” can be either some long-term mean rainfall, or the most probable value. The model forecasts can be considered useful only if they outperform these two simple forecasts.

TABLE 1b. As in Table 1 for three regional NWP models.

Center/regional model	Horizontal resolution received	Initial time (UTC)	Forecast period (h)	Verified since
Australian Bureau of Meteorology (ABOM) Limited Area Prediction System (LAPS) model	0.375° × 0.375° ^a	2300	24, 48	Jul 1997
Canadian Meteorological Centre (CMC)	24 km × 24 km ^b	0000, 1200	24, 36, 48	Mar 1995
National Centers for Environmental Prediction (NCEP) Eta model	22 km × 22 km ^c	0000, 1200	24, 36, 48, 60	Jun 1992

^aGrid resolution of 0.75 received until Feb 1999.

^bGrid resolution of 35 km received until Sep 1998.

^cGrid resolution of 80 km until Oct 1995, then 48 km until Feb 1998, then 32 km until Sep 2000.

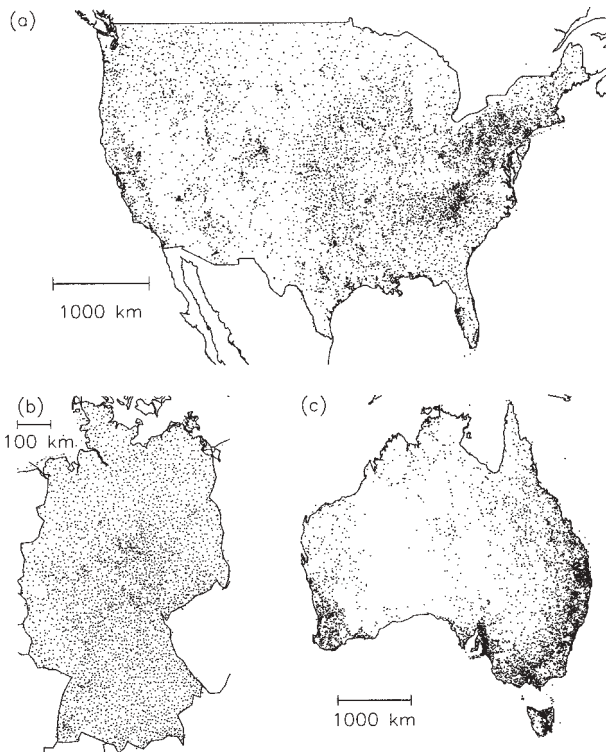


FIG. 1. Distribution of rain gauges in (a) the United States (7664 points), (b) Germany (3954 points), and (c) Australia (5177 points).

Verification data. The forecast rainfall must be verified against the best available estimates of the true rainfall. Quantitative precipitation measurements are made from rain gauges, radar, and satellite. Of these, rain gauges have the greatest accuracy and are generally considered to provide the best available estimates of rain distribution, at least in developed countries where the density of gauge sites is reasonably high. The gauge density is quite high throughout Germany, roughly one gauge per $(9 \text{ km})^2$, while in the United States it is one per $(32 \text{ km})^2$ (see Fig. 1), and the Australian gauge density is one per $(35 \text{ km})^2$. In both the western United States and western Australia, large (sparsely populated) regions have few or no rain gauges. Gauge-void regions are not verified.

Routine rain gauge measurements for meteorological purposes may be recorded as frequently as every hour or as infrequently as every 24 h. All three countries have large networks of volunteers who measure daily rainfall. In order to make best use of the data we chose a 24-h accumulation period as the basis for the QPF verification.

The particular 24-h period verified for the model QPFs varies from country to country. The local reporting time in Germany is 7:30 a.m., corresponding

to 0530 and 0630 UTC in winter and summer, respectively. Therefore, model forecasts valid at 0600 UTC are verified over Germany. Rainfall observations in the United States are made at 1200 UTC, and are therefore used to verify model forecasts valid at 1200 UTC. In Australia rainfall observations are made at 9 a.m. local time, which corresponds to between 2200 and 0100 UTC, depending on location and daylight savings. Over Australia model forecasts valid at 0000 UTC are verified.

Strategies for using rain gauge data. QPFs can be verified either directly against the observations themselves, or against a gridded analysis of the observations. Because rainfall varies spatially at small (subgrid) scales, the spatial smoothing by analyses can produce rain rates that differ markedly from the original observations. The maxima are flattened, and the area with low rain rates is artificially increased. However, rain gauge reports can suffer severe errors of representativeness (e.g., if a thunderstorm occurred over an isolated rain gauge).

One could argue that if the model output is to be used directly (i.e., interpolated) to make predictions at point locations, then the observations themselves are the most appropriate “truth” for verification. However, model values are inherently grid-scale quantities, and it is clear that model rainfall fields are much smoother than observed rainfall fields.

A recent study by Cherubini et al. (2002) showed that QPF verification scores computed by comparing model grid box values to gridded rainfall data were more favorable than those computed by comparing interpolated model output to the original point observations. Their results were not sensitive to the method used to assign the observations to the grid but larger grid boxes led to better verification scores.

DWD, NCEP, and BMRC all use different verification strategies. DWD verifies directly against the rain gauge data, with the model fields interpolated bilinearly to the station locations. No height corrections or other adjustments are applied. Offline experiments have shown that verification using the nearest grid point gives very similar overall results.

NCEP treats the forecast and observations as areal averages of the precipitation over an entire grid box. The verification grid over the United States corresponded to each model’s own grid until March 1999, after which it was changed to a standard Lambert conformal grid with a resolution of 80 km for the global models and 40 km for the regional models. At the same time the reporting practice for zero precipitation improved, resulting in a more accurate rain

gauge analysis. In the current system any data more than three standard deviations from the overall mean value are eliminated. The analysis is then simply the average of all rain gauge observations within each grid box. Grid boxes with no gauge data are not verified. The model QPFs are remapped to the verification grid by area-weighting the contributions from each model grid box, thus preserving the original rain area and amount.

The verification dataset over Australia is also a gridded rain gauge analysis. After range and buddy checking to eliminate erroneous data, the daily observations are analyzed to a 0.25° grid using a three-pass inverse distance-weighting (Barnes) scheme (Weymouth et al. 1999). The first pass uses a long length scale to obtain the large-scale pattern, then the second and third passes use a shorter length scale to capture the fine detail. The actual length scales used in the analysis vary according to data density; that is, they are shorter where the average interstation distance is less. The analyses are averaged to 1° resolution for the purpose of the WGNE verification, and the model QPFs are remapped to a 1° grid using bilinear interpolation or averaging, whichever is appropriate.

In central Australia, and to a lesser extent in the western United States, there are regions with no rain gauge coverage. QPF verification is not possible in these regions, and therefore the verification results will be unavoidably biased. Even where there are gauges, the “truth” data contain some amount of error due to sampling and/or smoothing, as discussed above. However, the differences between the forecast and observed rain fields are much greater than the errors in the verification data themselves, and so the latter can be used as a reasonable representation of the true rainfall.

Because each operational center handles the verification data differently it will not be possible to compare the quantitative verification statistics directly with each other to determine, for example, in which country the NWP QPFs have greatest skill. However, station and grid box verifications both reveal similar overall trends in QPF ability, as well as regional and seasonal trends. It is therefore possible to *qualitatively* compare model QPF behavior across the various domains.

RESULTS FOR CURRENT OPERATIONAL NWP MODELS. *QPFs for the United States.* Figure 2 shows the seasonal variation of the bias score and equitable threat score for 24-h forecasts during March 1999, when verification on a standard grid began, through February 2001. Four global models (solid

lines) and two regional models (dashed lines) were verified. Scores for the persistence forecast, verified from 2000 onward, are also shown. Most models were biased high during all seasons, indicating that rain was predicted more frequently than it was observed. The equitable threat score shows a clear seasonal dependence. ETS values of 0.4–0.5 were achieved in winter, when rain is associated mainly with synoptic-scale systems. The summertime values were lower, typically about 0.3, reflecting the greater difficulty in predicting convective rainfall. All of the models clearly outperformed persistence, which had quite low skill. Although QPF verification over the United States began in 1995, because the verification grid was changed in 1999 it is not possible to say with certainty whether model QPF skill has improved significantly in the last 5 yr.

For 24-h forecasts, all NWP models overestimated the frequency of light rain during January–December 2000 (Fig. 3). However, for rain exceeding 25 mm most models underestimated the frequency. Interestingly, two models (DWD and ECMWF) overestimated the frequency of rain exceeding 75 mm in both the 24- and 48-h forecasts, while the NCEP (eta) and CMC (global) models seriously underestimated heavy rain frequency. The model bias tended to be higher for 48-h forecasts than 24-h forecasts.

The equitable threat score (lower diagrams) peaked for rain threshold of 2.5 mm, with values in the range 0.37–0.41 for 24-h QPFs and 0.31–0.36 for 48-h QPFs. The skill then decreased for higher rain thresholds, with most models having an ETS of less than 0.20 for rain exceeding 40 mm for the 24-h forecasts and 20 mm for the 48-h forecasts. The low skill

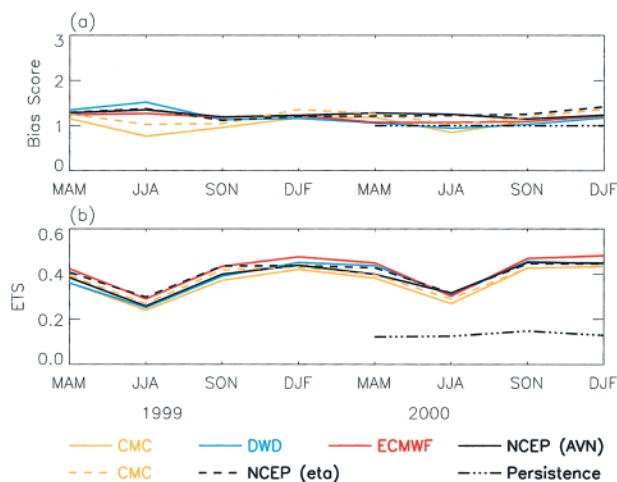


FIG. 2. Time series of seasonal (a) bias score, and (b) equitable threat score for 24-h QPFs over the United States, using a rain threshold of 0.1" (2.5 mm) d⁻¹.

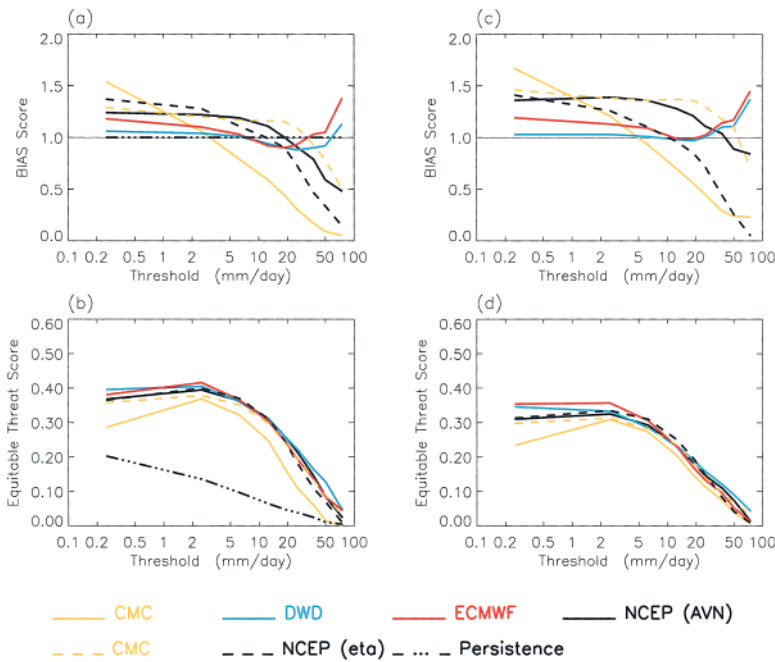


FIG. 3. (a), (c) Bias score and (b), (d) equitable threat score as a function of rain threshold for (a), (b) 24-h and (c), (d) 48-h forecasts over the United States for Jan–Dec 2000. The thresholds used were 0.25, 2.5, 6.25, 12.5, 18.75, 25, 37.5, 50, and 75 mm d⁻¹.

for the higher rain rates results largely from the difficulty in correctly predicting the location of the heaviest rain. The superiority of the model QPFs over the persistence forecast was greatest for moderate rain rates.

Except for one model, there were only small differences in skill among the models. It appears that models that overforecast rain frequency score more hits but also more false alarms (errors). Conversely, models that underforecast rain score fewer hits and fewer false alarms, with the net result that the equitable threat score has little dependence on the model’s forecast bias.

QPFs for Germany. The bias scores for January–December 2000 (Fig. 4a) show that for amounts up to 2 mm d⁻¹, all models have a tendency to overpredict rain frequency over Germany. For higher observed amounts, the UKMO, ECMWF, CMC, and DWD model underpredicted the rain frequency. The threat score (Fig. 4b) shows a rapid decline in quality for forecasts of heavy rain.

(Fig. 5a). As was the case in the United States, the equitable threat score shows a marked annual cycle in

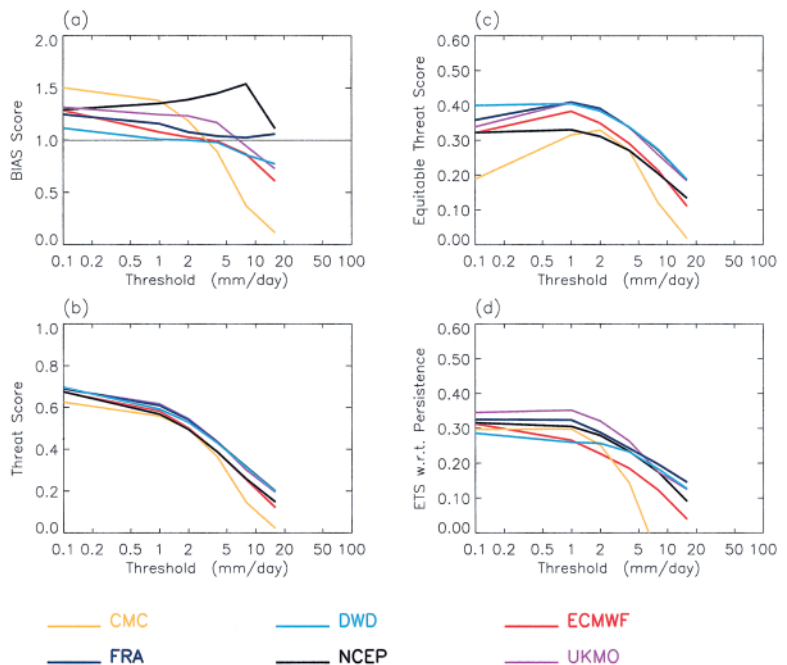


FIG. 4. (a) Bias score, (b) threat score, (c) equitable threat score, and (d) equitable threat score with a persistence reference, as a function of rain threshold for 24-h accumulated precipitation valid 42 h into the forecast (ECMWF) or 30 h (other models) for Jan–Dec 2000 over Germany. The thresholds used were of 0.1, 1, 2, 4, 8, and 16 mm d⁻¹.

The equitable threat scores computed with respect to a random forecast and a persistence forecast are shown in Figs. 4c, and 4d, respectively. Model performance was most similar for the 2 mm d⁻¹ threshold, with the spread between the different models being considerably larger for light and heavy rain. The apparent difficulty in predicting heavy rainfall is partly related to the comparison of model resolution QPFs with point rainfall observations (see Ghelli and Lalaurette 2000). The best forecast quality is achieved for rain exceeding 1 mm d⁻¹. Figure 4d shows that the persistence forecast is more difficult to beat than a random forecast. There is still considerable scope for improvement, as none of the global models approach the optimum value of 1.0 even for short-range forecasts.

Some centers managed to reduce systematic rainfall overprediction by their models over the last 3 yr

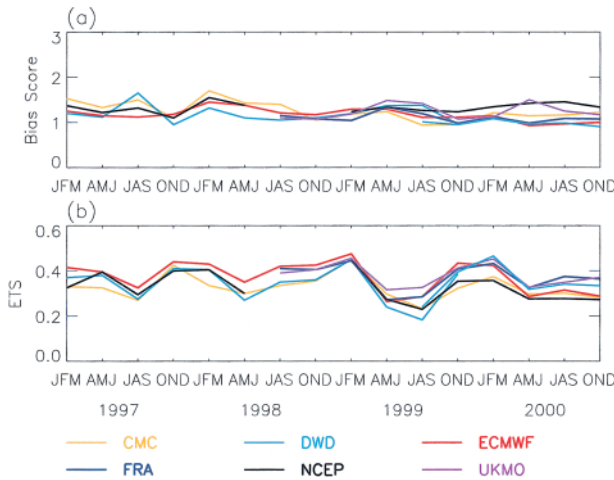


FIG. 5. (a) Time evolution of the bias score and (b) the equitable threat score over Germany between Jan 1997 and Jun 2000 for a rain threshold of 2 mm d^{-1} . The forecast valid time is 42 h (ECMWF) and 30 h (other models). Because of technical problems, some values for the NCEP model are missing.

forecast quality with predictions being better in winter than in summer (Fig. 5b). Apparently, the summer of 1999 was more difficult to predict than other summers; this period also coincided with a spell of lower than normal medium-range forecast quality over Europe.

Figure 5 also shows that the introduction of the new global model (GME) at DWD in mid-1999 resulted in a marked improvement, whereas changes made to the CMC model did not appear to have the desired effect for precipitation over Germany. It is difficult to detect a systematic longer-term trend toward improved model QPF skill as there were large year-to-year variations in the predictability for precipitation over central Europe.

In Fig. 6, QPF bias shows little sensitivity to forecast length. For light precipitation, all forecasts (42, 66, and 90 h) show virtually the same overprediction. As also seen in Fig. 5 there is a trend toward smaller biases over the 4-yr period. There is stronger dependence of the equitable threat score on the forecast period, with shorter-range forecasts having greater skill. The quality of the 90-h forecast is closer to the 66-h quality than the 66 h is

to the 42 h, possibly indicating the beginning of saturation at the medium range. An apparent decline in the forecast quality of the ECMWF model QPFs (Figs. 6e,f) may be due to interannual variability in precipitation predictability, as mentioned earlier.

QPFs for Australia. QPFs for Australia were verified separately for two domains, a tropical domain extending equatorward of 20°S , and a midlatitude southeastern domain extending poleward and eastward of (25°S , 135°E), following McBride and Ebert (2000). Australian tropical rainfall is governed by a monsoon cycle with heavy rainfall and frequent deep convection during summer, and almost no rainfall during winter. In the midlatitude region the cool season rainfall is primarily caused by synoptic disturbances and onshore advection of moist air, while summer has greater proportion of convective rainfall.

For northern (tropical) Australia (Fig. 7) the models tend to predict the rain frequency quite well during the rainy season [December–January–February (DJF), hereafter all 3-month combinations will be set as such] underestimate it during the dry season (JJA), and overestimate it in the transition seasons. Because so little rain falls in winter these verification scores are highly sensitive to even small errors in predicted rainfall, and much of the wintertime error is related to the

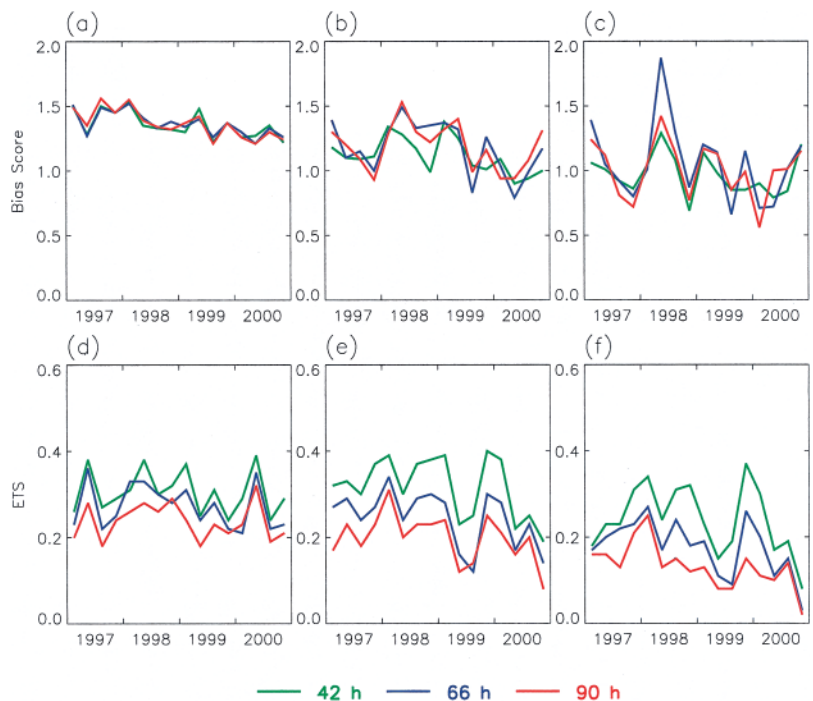


FIG. 6. (a)–(c) Bias score and (d)–(f) equitable threat score over Germany for three different thresholds (left to right, respectively, 0.1, 4, and 8 mm d^{-1}) and for three forecast periods (42: green; 66: blue; 90 h: red) for the ECMWF model.

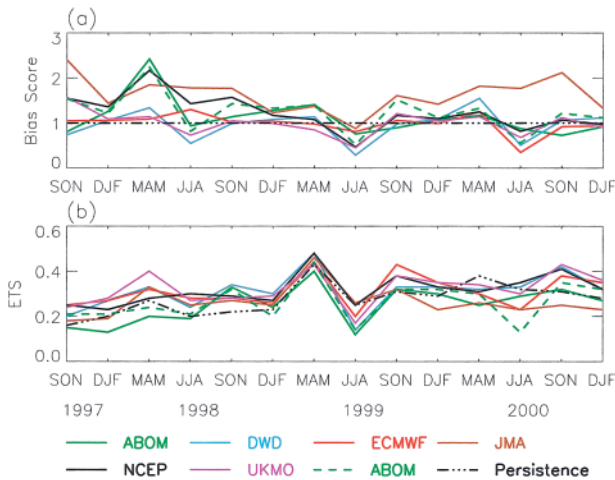


FIG. 7. Time series of seasonal (a) bias score and (b) equitable threat score for northern (tropical) Australia for 24-h QPFs, for a rain threshold of 1 mm d⁻¹.

failure of the models to capture orographic rainfall in far northeastern Australia. The particularly high bias shown by the JMA model in the later part of the time series may be related to a model change implemented in December 1999. No significant change in behavior was noted for the DWD model, which also underwent an upgrade in December 1999.

The 24-h ETS averaged about 0.3 for northern Australia and did not vary systematically by season. To put these results into perspective, the 24-h persistence forecast scored equally well, indicating that the NWP models provided little additional useful QPF information in tropical Australia. This was also noted by McBride and Ebert (2000). Although model ETS values appear to grow slightly over time, this is most likely due to the tropical weather in later years being easier to forecast since the ETS for the persistence forecast also had positive trend.

The differences in scores among models can be considered statistically significant if the confidence intervals associated with the individual scores do not overlap. The 95% confidence intervals for the seasonal values of bias and ETS were estimated using a bootstrap (resampling) method (Mason and Mimmack 1992). The mean confidence interval for bias was ± 0.2 , while for ETS

the mean confidence interval was ± 0.06 . This suggests that the differences among most models during a given season, or in the annual performance of a single model over time, were not statistically significant.

In the tropical region during the wet season December 1999–February 2000, most models slightly overestimated the total rain area exceeding 1 mm d⁻¹, while there was large variability in the predicted frequency of moderate and heavy rainfall (Fig. 8). The models tended to achieve their highest ETS for light rainfall, decreasing with increasing rain rate. In summer the models generally outperformed persistence, and all but two of the methods, JMA and NCEP, performed better in summer than in winter for tropical rainfall (not shown). The 48-h forecasts had only slightly less skill than the 24-h forecasts.

In contrast to the Tropics, the model bias scores in southwestern Australia showed little seasonal variation, with some models (e.g., ABOM and JMA) consistently overestimating rain frequency and others (e.g., UKMO and DWD) consistently underestimating rain frequency (Fig. 9). The equitable threat scores were slightly higher than average during winter and lower in summer, similar to the findings over Germany and the United States. ETS values were significantly greater in this midlatitude region than in the Tropics, while the skill of “persistence” was

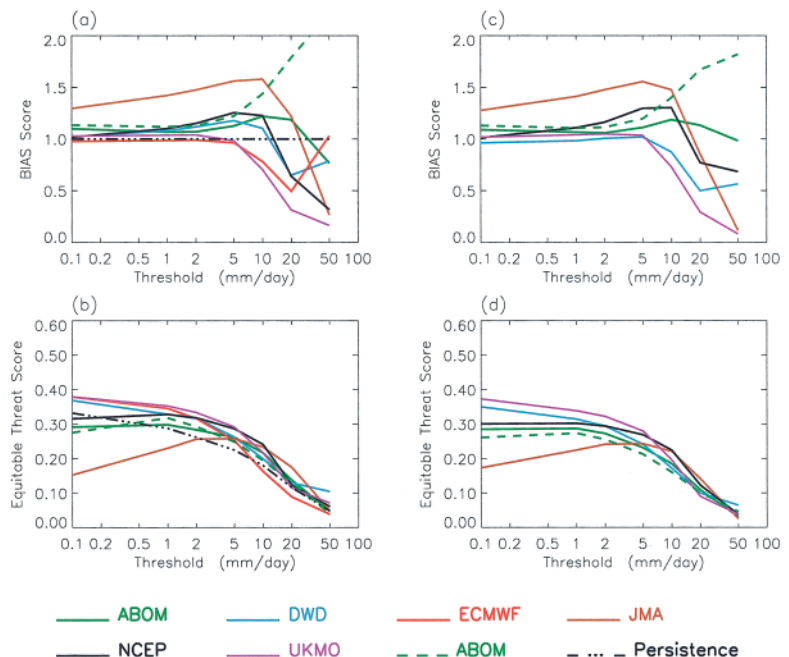


FIG. 8. (a), (c) Bias score and (b), (d) equitable threat score as a function of rain threshold for (a), (b) 24- and (c), (d) 48-h forecasts in the tropical region of Australia during Dec 1999–Feb 2000. The thresholds used were 1, 2, 5, 10, 20, and 50 mm d⁻¹.

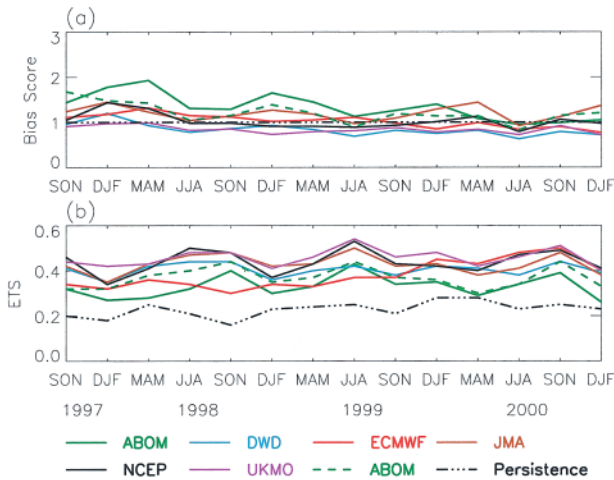


FIG. 9. Time series of seasonal (a) bias score and (b) equitable threat score for southeastern (midlatitude) Australia for 24-h QPFs, for a rain threshold of 1 mm d⁻¹.

slightly lower. The relative advantage of model QPFs over a persistence forecast is clearly much greater in the midlatitudes than in the Tropics.

The confidence associated with the midlatitude scores was greater than in the Tropics, with a mean 95% confidence interval of ± 0.1 for the bias score, and ± 0.03 for the equitable threat score. This means that some of the differences in performance among models in midlatitude Australia were statistically significant.

As seen in other domains there was large intermodel variability in the predicted frequency of heavy rainfall in midlatitude Australia, and the ETS peaked for light to moderate rain rates (Fig. 10). At high rain rates the DWD and ECMWF models displayed the same high bias as was seen over the United States. This can be attributed to overprediction of heavy rain in a few low pressure systems during autumn and spring. For moderate rainfall in particular the model QPFs showed significantly greater skill than persistence, and greater skill at 24 than at 48 h. The mean model ETS for rain greater

than 10 mm d⁻¹ was 0.30 for 24-h QPFs and 0.24 for 48-h QPFs.

The position errors of the forecast rain systems were determined using the entity-based, pattern matching verification method of Ebert and McBride (2000), where a contour of 5 mm d⁻¹ was used to delineate “significant” rain systems within contiguous areas (see Fig. 11). The most common location errors for over 600 Australian rain systems were in the range of 0.5°–1.2° latitude–longitude, represented by the peak at 1°. Fewer than 10% of forecasts predicted exactly the correct location. The mean location error was 2.5° for 24-h forecasts and 2.9° for 48-h forecasts. The 95% confidence interval for the mean location errors was $\pm 0.15^\circ$ for both forecasts periods.

The most accurate forecasts of rain location occurred for winter rain systems in Australian midlatitudes where the mean 24-h location error was 1.9° and 17% of forecasts predicted the correct location. In this case the location errors accounted for about one-third of the total error, rain volume errors accounted for about 10% of the total error, with the remainder being due to errors in finescale structure (see Fig. 12). In the Tropics less than 5% of the total error could be attributed to rain volume errors, with the remainder split evenly between location and pattern error.

The NWP models were more successful at forecasting the location of storms with very heavy rain-

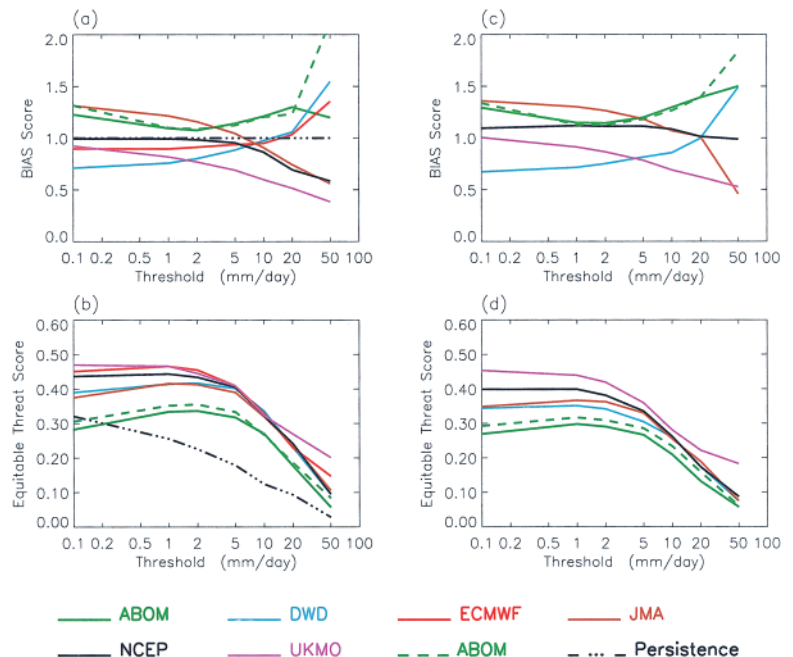


FIG. 10. (a), (c) Bias score and (b), (d) equitable threat score as a function of rain threshold for (a), (b) 24-h and (c), (d) 48-h forecasts in the Australian midlatitude region during Dec–Jan 2000. The thresholds used were 1, 2, 5, 10, 20, and 50 mm d⁻¹.

¹ The pattern matching technique can only estimate displacements to the nearest grid point, so it is not possible to resolve displacement errors between 0° and 1°. Errors between 0° and 0.5° will be assigned to the 0° bin, and errors between 0.5° and 1.2° will be assigned to the 1° bin.

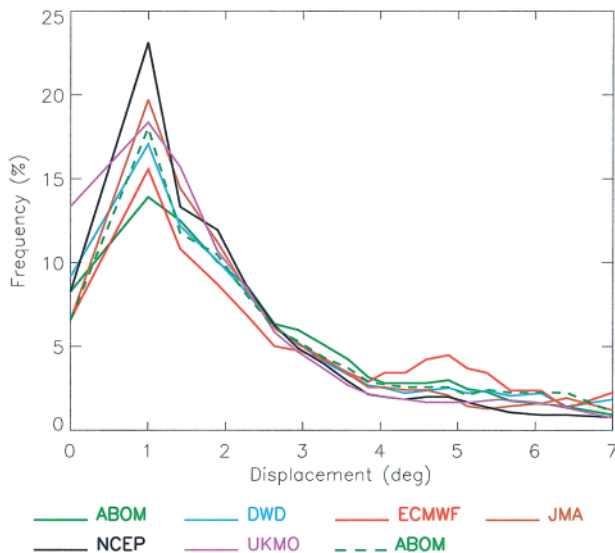


FIG. 11. Frequency distribution of location errors in latitude for 24-h QPFs over Australia during Sep 1997–Dec 2000.

fall. For cases in which the maximum observed rainfall exceeded 100 mm d^{-1} the mean displacement was 1.2° for 24-h forecasts and 2.1° for 48-h forecasts. This improvement probably reflects the greater degree of synoptic forcing that would be expected to accompany heavy rain systems. This result does not contradict the earlier findings showing the lowest equitable threat scores for heavy rainfall (Figs. 8 and 10). The entity-based verification indicates that from a “rain system” point of view, the models do a good job of predicting the location of the rainfall. However, from a “gridpoint” point of view, the low ETS values for rain exceeding 50 mm d^{-1} reflect the difficulty in predicting the precise location of the heaviest rain.

DISCUSSION. This study complements a number of earlier studies of QPF skill for operational models (Roads and Maisel 1991; Junker et al. 1992; Olson et al. 1995; Gartner et al. 1998; McBride and Ebert 2000; Damrath et al. 2000). The small number of verification statistics reported here give only a partial picture of the ability of current NWP models to make accurate and useful rainfall predictions. The WGNE QPF verification study produces a much greater array of scientific verification results, and interested readers may contact the authors to obtain further information if desired.

The results of the WGNE QPF assessment show that operational NWP models are still a long way from producing perfect QPFs. However, except for the tropical regions, the models easily outperform a forecast of “persistence” and thus clearly provide

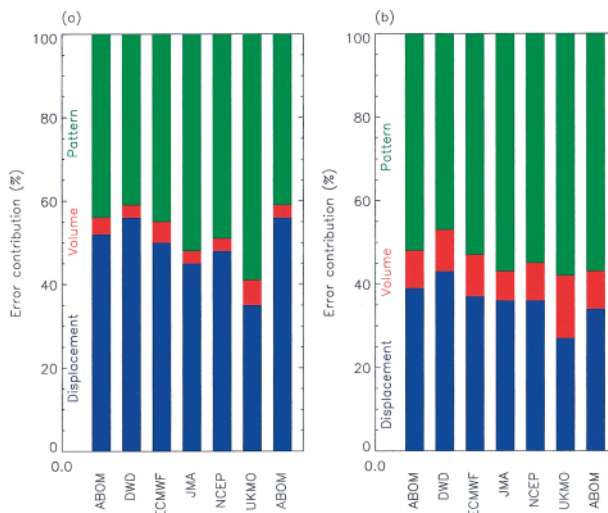


FIG. 12. Error decomposition (%) for rain systems occurring in (a) the Australian Tropics during summer and (b) southeastern Australia during winter for the different models.

useful routine guidance. Other studies have shown that operational NWP models, particularly the high-resolution regional models, are certainly capable of producing high-quality forecasts for specific cases of heavy rain (e.g., Mesinger 1996; Damrath et al. 2000).

Unfortunately, we cannot conclude that, generally speaking, upgrades to the operational NWP models have led to significantly improved QPFs during the last four to five years. Not all new model versions have enhanced QPF performance, although further regional verification should be done to check this conclusion. The blessing and the curse of quantitative precipitation forecasting using NWP models is that the forecast rain depends so sensitively on the model’s predicted atmospheric and surface conditions. While a good rain forecast strongly suggests a good forecast of all other atmospheric variables, a bad rain forecast can result from errors that have little or nothing to do with the rainfall parameterization itself. If a model is tuned to optimize a field other than rainfall (e.g., the mass field is often optimized using the S1 skill score), then model errors may actually cause the rain forecasts to deteriorate. The process of improving model numerics and physics is a complicated juggling act. Unless the accurate prediction of rainfall is made a top priority then improvements in NWP model QPF will continue to be realized slowly.

Even with a perfect model, errors in the initial conditions would still lead to imperfect forecasts. To get a rough feeling for the magnitude of QPF errors resulting from initial conditions alone, a simple ex-

periment was conducted using 24- and 48-h forecasts of rainfall over Australia during January and July 1998 from the ECMWF Ensemble Prediction System (EPS) at 1.25° spatial resolution. The 51 members of the ensemble differed only in the specification of their initial conditions. The imposed dynamical perturbations were of comparable magnitude to the uncertainties in the analysis of the mass field. By verifying the control (unperturbed) forecast against a randomly chosen ensemble member, one can remove the effect of model bias. In this case the model is assumed to be perfect and the QPF “errors” result solely from differences in initial conditions. Such a verification gives the *potential skill* of the forecast (Buizza 1997).²

The Australia-wide value of the *potential* equitable threat score for the ECMWF model for those two months was 0.89 at 24 h and 0.78 at 48 h. These values are much higher than the values of 0.3–0.5 that the models currently achieve. This suggests that the development of a perfect NWP model would close most of the gap between the current model QPF skill and perfect skill.

This is one of the approaches advocated at a recent U.S. Weather Research Program (USWRP) Workshop on Warm Season Precipitation (Fritsch and Carbone 2003). Although model resolution has increased dramatically over recent years, fairly crude parameterizations of convection that were designed for coarser models continue to be used in NWP. To amend this situation it may be necessary to run operational models on grid scales that can explicitly resolve clouds so that cloud physical and dynamical processes can be more accurately represented. This will require an improved understanding of the life cycle of precipitation events, including the initiation and evolution of convection, and cloud microphysical development, on which to base new moist physics schemes. Field experiments using state-of-the-art observing systems will be needed to address these issues. Continuous four-dimensional assimilation of an increasing variety of nonsynoptic data will improve the specification of model initial conditions. The USWRP will promote research in these areas over the next decade.

² A few caveats are warranted with regard to the estimation of potential skill using this method. A perfect model would have improved initial conditions through provision of a better first-guess field, increasing the value of the potential skill. On the other hand, the perturbations in the EPS initial conditions did not account for uncertainties in diabatic processes, which would tend to increase the initial spread and lower the apparent potential skill.

In the meantime, one of the most promising and practical ways to improve quantitative precipitation forecasting using existing NWP models is the use of ensembles to generate multiple rain scenarios and probabilistic forecasts. Most global ensemble predictions systems were designed for medium-range forecasts, and do not tend to show optimal performance in the 1–2-day range. In the United States the short-range ensemble forecast (SREF) system combines ensembles from the Eta and regional spectral models and has been shown to produce skillful probabilistic and deterministic (ensemble mean) forecasts of precipitation (Stensrud et al. 1999; Wandishin et al. 2001). A “poor man’s ensemble” of QPFs from a number of operational centers has been shown to perform well in the 1–2-day range (Ebert 2001). While improvements in our understanding of rainfall process, numerical models, and data assimilation are important steps toward improving quantitative precipitation forecasting, ensemble prediction may offer the most effective means of making best use of the imperfect QPFs available to us at present.

ACKNOWLEDGMENTS. The authors would like to acknowledge many helpful comments from the three anonymous reviewers and the editor. Roberto Buizza kindly provided ensemble QPFs from the ECMWF ensemble prediction system.

REFERENCES

- Buizza, R., 1997: Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.*, **125**, 99–119.
- Cherubini, T., A. Ghelli, and F. Lalaurette, 2002: Verification of precipitation forecasts over the Alpine region using a high-density observing network. *Wea. Forecasting*, **17**, 238–249.
- Damrath, U., G. Doms, D. Fruehwald, E. Heise, B. Richter, and J. Steppeler, 2000: Operational quantitative precipitation forecasting at the German Weather Service. *J. Hydrol.*, **239**, 260–285.
- Ebert, E. E., 2001: Ability of a poor man’s ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480.
- , and J. L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrol.*, **239**, 179–202.
- Fritsch, J. M., and R. E. Carbone, 2003: Research and development to improve quantitative precipitation forecasts in the warm season. Synopsis of the March 2002 USWRP workshop and statement of priority

- recommendations, 130 pp. [Available from Dr. Richard Carbone, NCAR, P.O. Box 3000, Boulder, CO 80307-3000.
- Gartner, W. E., M. E. Baldwin, and N. W. Junker, 1998: Regional analysis of quantitative precipitation forecasts from NCEP's "Early" Eta and meso-Eta models. Preprints, *16th Conf. on Weather and Forecasting*, Phoenix, AZ, Amer. Meteor. Soc., 187–189.
- Ghelli, A., and F. Lalaurette, 2000: Verifying precipitation forecasts using upscaled observations. ECMWF Newsletter No. 87, 9–17.
- Junker, N. W., J. E. Hoke, B. E. Sullivan, K. F. Brill, and F. J. Hughes, 1992: Seasonal and geographic variations in quantitative precipitation prediction by NMC's Nested-Grid Model and medium-range forecast model. *Wea. Forecasting*, **7**, 410–429.
- Mason, S. J., and G. M. Mimmack, 1992: The use of bootstrap confidence intervals for the correlation coefficient in climatology. *Theor. Appl. Climatol.*, **45**, 229–233.
- McBride, J. L., and E. E. Ebert, 2000: Verification of quantitative precipitation forecasts from operational numerical weather prediction models over Australia. *Wea. Forecasting*, **15**, 103–121.
- 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.
- Olson, D. A., N. W. Junker, and B. Korty, 1995: Evaluation of 33 years of quantitative precipitation forecasting at the NMC. *Wea. Forecasting*, **10**, 498–511.
- Roads, J. O., and T. N. Maisel, 1991: Evaluation of the National Meteorological Center's medium range forecast model precipitation forecasts. *Wea. Forecasting*, **6**, 123–132.
- 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, WMO, 114 pp.
- Stensrud, D. J., H. E. Brooks, J. Du, M. S. Tracton, and E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev.*, **127**, 433–446.
- Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of short-range multimodel ensemble system. *Mon. Wea. Rev.*, **129**, 729–747.
- WCRP, 1995: Report of the tenth session of the CAS/JSC Working Group on Numerical Experimentation. WMO/TD No. 678, WMO, 43 pp.
- 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.