VERIFICATION OF HIGH-RESOLUTION PRECIPITATION FORECASTS FOR THE 1996 ATLANTA OLYMPIC GAMES

Lígia R. Bernardet*

NOAA Office of Oceanic and Atmospheric Research
Forecast Systems Laboratory
Boulder, Colorado

Abstract

The Local Analysis and Prediction System (LAPS), developed by the National Oceanic and Atmospheric Administration's Forecast Systems Laboratory (NOAA FSL), was an integral part of the Olympic Weather Support System (OWSS) designed by the NOAA National Weather Service (NWS) to supplement the forecasting operations in the Peachtree City, Georgia, NWS Forecast Office during the 1996 Atlanta Summer Olympic Games. This paper presents an objective hourly verification of some of the precipitation forecasts produced by the numerical modeling component of LAPS during the summer of 1996 for the southeastern United States.

The scores indicated underforecasting at all thresholds when the model was initialized at 0600 UTC. A later initialization improved the bias at lower thresholds, but caused overforecasting at higher thresholds. A comparison with the precipitation forecasts by the NWS 29-km Eta model showed that the high-resolution LAPS system was able to produce better precipitation forecasts, particularly when initialized with a high-resolution local analysis.

This paper also presents a discussion of the impact of different algorithms used to collocate observed and forecasted precipitation data. Higher bias scores (BSs) were obtained when the score was computed at the model grid points instead of at the station locations. For BSs computed at the station locations, higher scores were obtained when a larger number of grid points surrounding a station was used to compute forecasted precipitation at the station.

1. Introduction

Precipitation forecasts may have high economic value. Knowledge of upcoming precipitation events is important to economic activities such as transportation, irrigation, hydroelectric power, tourism and sports (Katz and Murphy 1997). To support the latter two activities, the Olympic Weather Support System (OWSS; Rothfusz et al. 1996) was designed by the National Weather Service (NWS) to operate during the 1996 Atlanta Summer Olympic Games. The OWSS was designed to produce high-quality local weather forecasts. The Local Analysis and Prediction System (LAPS), developed by the National Oceanic and Atmospheric Administration's Forecast Systems Laboratory (NOAA FSL), was an integral part of the OWSS. LAPS supplied the Peachtree City, Georgia, NWS Forecast Office with high-resolution, high frequency, surface and upper-air weather analyses (Albers 1995; Albers et al. 1996) and with local model forecasts (Snook et al. 1995). From here on, high-resolution refers specifically to high horizontal resolution.

This was one of the first attempts to use a high-resolution model in an operational environment. Other efforts (e.g., Colle et al. 1999) have indicated that fine model resolution does lead to improved precipitation forecasts. The forecasters in the Peachtree City, Georgia, NWS Forecast Office were pleased with the added benefit of LAPS in forecasting (Rothfusz and McLaughlin 1997). They pointed out that the model depicted well the development of the sea breeze and the onset of convection.

Quantitative verification of the model forecasts produced by LAPS was partially presented by Snook et al. (1998), who examined the model's performance in predicting surface temperature, dewpoint and winds. This paper examines the performance of the model in predicting precipitation. Scores are presented hourly during the 16-hour forecast period, so that model spin-up time and predictability can be addressed. The spatial distribution of scores is also shown, with the goal of identifying locations in the model domain where the forecasts are more or less reliable. A comparison of the observed and forecasted precipitation distributions at the end of the forecast period is also presented through the computation of quantiles of the distributions.

One of the features of the LAPS installation in the Peachtree City NWS Office that was most praised by the forecasters was the capability of starting a forecast whenever they decided it was necessary (Rothfusz and McLaughlin 1997). To understand the impact of different initialization times, forecasts initialized at 0600 UTC and 1500 UTC are examined and the precipitation forecasts in the afternoon and by the end of the 16-hour forecast period are compared. Furthermore, forecasts initialized with the LAPS analysis are compared with forecasts initialized with the NWS 29-km Eta model analysis (Black 1994), to assess the importance of the LAPS high-resolution analysis for model initialization. A comparison with the precipitation forecasts obtained from the 29-km Eta model itself is also presented. Due to difficulties in accessing the NOAA database, the number of days used in each

*also affiliated with the National Research Council analysis is limited and the days used in each analysis do not exactly coincide. This point is further addressed in Section 2. It should be stressed that these results reflect a snapshot of the models as they were in the summer of 1996. Since then, numerous significant changes have been implemented in each model, and the results herein do not necessarily reflect the current performance of each model.

Besides a description and interpretation of the precipitation verification, this paper discusses the methodology of forecast verification. Forecasted and observed precipitation must be collocated before the scores can be computed. Although the algorithms used to collocate the two datasets are seldom discussed in the literature of forecast verification, they can strongly impact the scores attained. In this paper, different algorithms to interpolate forecasted precipitation to the station locations are discussed, and results of verification done at the station locations are contrasted with verification done at the model grid points. It is shown that verification at the grid points produces consistently higher scores when the observational data network is sparser than the model mesh.

This paper is organized as follows. Section 2 presents the forecast model configuration and the dataset used for verification. Section 3 discusses the methodology for verification. Verification of the LAPS forecasts computed at model grid points and at the station locations is presented in Section 4. Section 5 presents the verification of the forecasts initialized at a later time and those initialized with the Eta model analysis. A comparison with the verification scores from the 29-km Eta model is presented in Section 6. A discussion of the results is presented in Section 7 and conclusions are in Section 8.

2. Model Setup and Dataset Used for Verification

The model used in this study is the non-hydrostatic Scalable Forecast Model (SFM), developed at Colorado State University (CSU) and at NOAA FSL, among other institutions. The setup of the model was identical to the one used during the 1996 Summer Olympic Games that took place in Georgia. During the Games, the model was run as part of LAPS (Snook et al. 1998) included in the OWSS designed by the NWS.

The model domain was configured with a polar-stereographic grid (Fig. 1) comprised of 85 points along each horizontal axis and 30 vertical levels. The horizontal grid spacing was 8 km. In the vertical a stretched grid was used with a grid spacing of 100 m near the ground, stretching gradually to 1000 m above 7000 m AGL, with the top of the domain at 17.3 km AGL.

For the control runs, the model was initialized from the real-time LAPS analysis (Snook et al. 1998), which used the same horizontal domain and grid spacing as the model. The LAPS vertical grid had 21 levels at 50-hPa increments. The LAPS forte is that it can incorporate a multitude of locally available data into a high-resolution local analysis. During the time period of the Olympic Games, there were numerous sources of data. LAPS used the 60-km Rapid Update Cycle (RUC; Benjamin et al. 1991) analyses as a first-guess for the upper-air analyses. Two WSR-88D Doppler radars were available within the LAPS domain, and three-dimensional, radar-derived winds were incorporated into the LAPS analyses. Upper-level winds from 120 to 3770 m were also derived from a boundary layer profiler. Hourly surface observations were obtained from about 60 standard surface stations, which reported in METAR format, along with 50 NWS-operated mesonet stations. Visible and infrared meteorological satellite data were used in the computation of surface temperatures and in the cloud analyses.

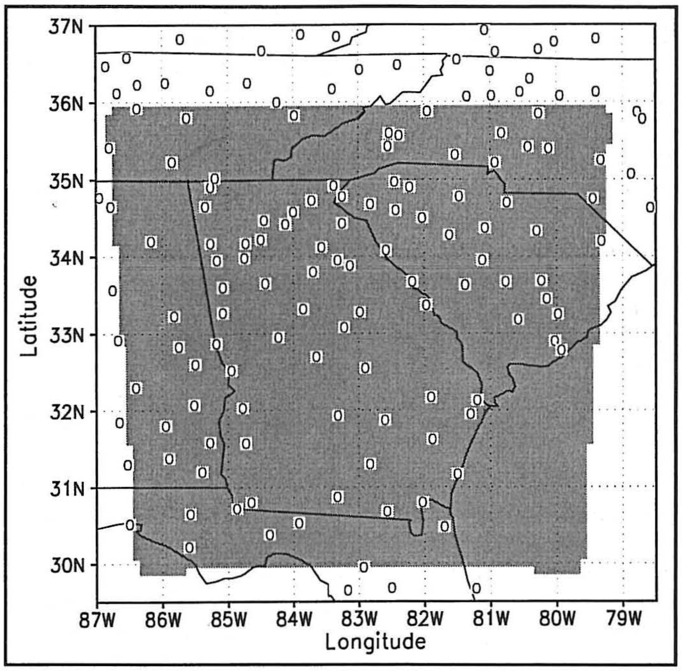

Fig. 1. LAPS domain (shaded area) and location of observation gauges (circles).

Since LAPS did not have a soil temperature and moisture analysis, nor a sea surface temperature (SST) analysis, the soil moisture was initialized at a constant 48% of saturation and the SST was set to its climatological value. The soil temperature was set to be identical to the temperature at the model's first level (48.3 m).

For lateral and top boundary conditions, the Eta model (Black 1994), run by the NWS/National Centers for Environmental Prediction at 29-km grid spacing, was used. Grids from the Eta were ingested into the SFM every three hours using nudging (Davies 1983) over five grid points for each lateral boundary, and over four grid points for the model top. The lower-boundary conditions were supplied by the model's surface parameterization (Louis 1979), which computes the fluxes of heat, moisture, and momentum. Prognoses of soil temperature and moisture content were made according to a parameterization by Tremback and Kessler (1985). The vegetation model (Avissar and Pielke 1989) was run using a variable vegetation initialization (Loveland et al. 1991), which characterizes the vegetation according to its leaf area index, roughness length, displacement height and root parameters.

For precipitation physics, the model employed the Walko et al. (1995) bulk microphysics scheme, which classifies the water substance in eight categories: vapor, cloud, rain, pristine ice, snow, aggregates, graupel and hail. Due to the subtropical summer characteristics of this study, rain was the only type of hydrometeor to reach the ground. No cumulus parameterization scheme was employed. The radiation scheme utilized was the Mahrer and Pielke (1977) formulation with a modification to account for the presence of clouds (Thompson 1993). It should be noted that this model setup was chosen to replicate NOAA FSL's forecasts. It is possible that other model configurations (use of a cumulus parameterization, variable soil moisture or observed SSTs) might have led to different and/or better results.

The model was initialized at 0600 UTC and integrated for 16 hours. Although the Olympic Games ran for 40 days, from 16 July through 26 August 1996, files for initialization and boundaries were not available for several days, limiting the sample used in this verification to 19 runs. Since our sample was limited to a small number of days, the results presented here should be understood as applicable only to this period (summer) in the southeastern U.S., and should not be extended to the performance of the SFM model in general.

Two experiments, discussed in Section 5, were set up in which the model configuration differed from the above. For the first experiment, the runs were initialized at 1500 UTC instead of 0600 UTC. The goal was to test whether a late morning initialization would lead to an improved forecast for the afternoon precipitation. The second set of runs was also initialized at 1500 UTC, but using the 29-km Eta data for initialization instead of the 8-km LAPS analysis, with the objective of testing for an improvement in the forecast due to the use of the high-resolution LAPS analysis. The first experiment will be referred to as $\text{LAPS}_{\text{init}}$ experiment, and the latter as Eta_{init} experiment.

For both experiments, the soil moisture and temperature initializations were altered to reflect afternoon conditions. The new values were typical numbers obtained at 1500 UTC from the model runs initialized at 0600 UTC. The altered soil moisture was set to 35%, and the altered soil temperature was set to a vertically variable profile, with a surface value 6°C warmer than the first model level decreasing to 2°C colder than the air at a depth of 50 cm. All other initialization procedures and parameterizations employed were identical to the ones used at 0600 UTC.

The results of the experiments were compared to the results of the 29-km Eta model itself, to identify the value added by using a high-resolution model for local precipitation forecasting. The days with initial and lateral boundary data available to run the experiments and with 29-km Eta outputs available were different than the days with data available to the control runs, since the experiments required LAPS analyses for different times (starting at 0600 or 1500 UTC) or from a different source. The choice of a set of days with data to initialize all three sets of runs would severely limit the size of the samples. Therefore, each experiment comprised a different set of days, all within the 40-day long Olympic Exercise. The size of the samples was 19 days for the control run, 21 days for the LAPS runs initialized at 1500 UTC and 22 days for the 29-km Eta runs. Since the number of days in each sample is relatively small and the days used for each experiment do not coincide, the conclusions derived from this study must be taken cautiously, and not be extended to other locations or time periods without further investigations. The results presented here should be interpreted qualitatively, since it is not possible to draw quantitative information due to the reduced size of the sample.

To verify the precipitation forecasts, we used the Hourly Precipitation Data (HPD) managed by NOAA's National Climatic Data Center (NCDC), which include amounts obtained from recording rain gauges located at National Weather Service, Federal Aviation Administration, and cooperative observer stations. NCDC performs both automated and interactive quality control on HPD data. Preliminary screening of the data is based on gross error and neighboring stations' checks, and collocated cooperative summary of the day observations from standard 8-in. gauges. An NCDC quality control specialist makes the final determination on the validity of suspect data.

3. Methodology

The problem of forecast verification is complex and multidimensional. Ideally, one would want to analyze the joint distribution of forecasts and observations. Since the amount of data for such analysis is huge, a simplification of the problem is necessary, and scores that summarize the distributions are commonly used. However convenient it might sound, there is not a single number that can completely characterize a forecast. Different indices and scores evaluate different aspects of the forecast. In that light, we have chosen to use several measures, which we describe below.

a. Comparison of observed and forecasted data distributions

For each distribution of forecasted and observed precipitation, quantiles were computed as a form of exploratory data analysis. Since the distributions were typically characterized by a large number of zeroes (representing no-rain events), cases with rain amounts smaller than 0.5 mm were eliminated from the distribution. Although this procedure alters the original distribution, it makes the results more meaningful, since otherwise all quantile values would be close to 0.0 mm. The distributions' quantiles are shown in quantile-quantile (QQ) plots (Wilks 1995), in which only the 0.25-mm and greater quantiles were plotted, since the lower quantiles were too close to zero mm.

b. Quantitative methods for forecast verification

All scores employed involve categorical precipitation forecasts. The precipitation forecasts are assumed to be in discrete categories with a lower and an upper limit. Four categories were defined for this study: 0.0 mm $\leq$ p < 2.5 mm, 2.5 mm $\leq$ p < 12.5 mm, 12.5 mm $\leq$ p < 25.0 mm, and p $\geq$ 25.0 mm, where p is the precipitation amount accumulated since the beginning of the forecast. Verification scores will be presented for individual thresholds (2.5, 12.5 and 25.0 mm), which comprise precipitation events at or above the given amount. No verification will be discussed for thresholds smaller than 2.5 mm, since the majority of the rain gauges used only registers precipitation values of

Table 1. Sample contingency table. 'a' represents the number of correctly forecast events, 'b' represents the number of erroneously forecast non-events, 'c' represents the number of missed events (i.e., observed but not forecast), and 'd' represents the number of correctly forecast non-events.

		OBSERVATION		
FORECAST		yes	no	
	yes	a	b	a + b
	no	c	d	c + d
		a + c	b + d	a + b + c + d

2.5 mm and larger. We note that this choice of categories was arbitrary. It was not based on the characteristics of the observed or modeled precipitation distribution, but chosen to match precipitation amounts for which meteorologists usually make forecasts (0.1 in., 0.5 in., etc.).

Bias Scores (BSs), Threat Scores (TSs), and Equitable Skill Scores (ESSs) are the performance measures used in this study to reduce the comparison of the forecasted and observed distributions of precipitation to single numbers. To compute these scores, a four-category contingency table was first created.

1) Contingency tables

A contingency table relates the number of points with observed precipitation in each discrete category to the number of points with forecasted precipitation in each category. Each row corresponds to a category of forecast, and each column to a category of observations. A generic contingency table based on two precipitation categories is shown in Table 1. In the example shown, the event was successfully forecasted to occur **a** times, and erroneously forecasted to occur **b** times. The forecast missed the event **c** times, and **d** times the no-event was correctly forecasted. At the edges of the table, the marginal distributions of the forecasts and of the observations are displayed. These are simply the sums of the rows or columns of the table, and represent the total number of events that fall in each category. The sum of all marginal distributions of the forecasts is equal to the sum of all marginal distributions of the observations, and represents the total number of events studied.

2) Measure of bias

The BS is used to assess the bias of the forecasts. It is the ratio of the number of points at which an event has been forecasted to the number of points at which it has occurred. Unbiased forecasts have a BS value of 1. When overforecasting (underforecasting) occurs, BS is greater (less) than 1. The BS is computed as

$$BS = \frac{a+b}{a+c}. \tag{1}$$

3) Measure of accuracy

The TS is used as a measure of forecast accuracy. This score complements the BS since it considers the correspondence between each pair of forecasts and observations. Unlike the BS, it does not reward a correct number of forecasted events if their location is incorrect. The TS is the ratio between the number of points with correct forecasts (a) to the union of the number of points where the event was forecast and where the event occurred (a+b+c), and is computed as

$$TS = \frac{a}{a+b+c}. \tag{2}$$

It should be noted that the TS is not an ideal measure of local-scale forecast accuracy. The TS only considers as a correct forecast an event in which forecasted and observed rain are collocated. Therefore, it gives no reward to a good forecast (correct timing, intensity etc.) that is displaced from the observed location. The TS was used in this study because more sophisticated verification measures are yet to be developed.

4) Measure of skill

To measure skill, we adopted the ESS, as described by Schaefer (1990). This score assesses the accuracy of the model relative to a forecast of chance, and is computed as

$$ESS = \frac{a - chance}{a+b+c-chance}, \tag{3}$$

where

$$chance = \frac{(a+c)(a+b)}{a+b+c+d}. \tag{4}$$

c. Treatment of hourly precipitation data

Each of the measures described above was computed for each hour of each day, and also for the set of days being studied. For each hour, both observed and modeled precipitation amounts were accumulated from the beginning of the forecast.

The temporal distribution of the scores enables one to investigate aspects of model spin-up and model predictability. If the scores improve with time, one can assume that the model goes through an initial spin-up period, in which forecast quality is impacted, whereas if the scores decrease with time, one assumes that the model loses predictability with time.

d. Collocation of observations and model output

All statistical measures described above require that forecasted and observed precipitation be at the same location. Therefore, either the observed values must be represented on the model's grid or the forecasted precipitation must be interpolated to the station locations. The

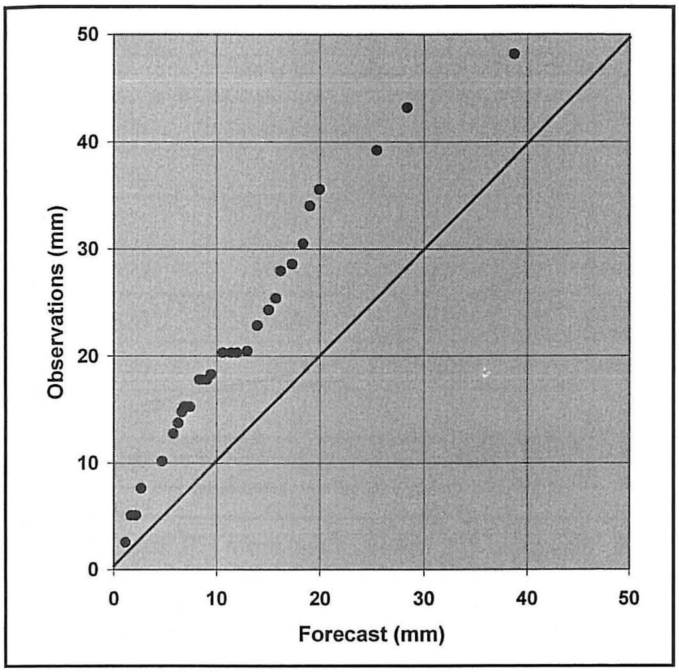

Fig. 2. Quantile-quantile plot for the control run ending at 2200 UTC.

difficulty in converting one type of information into the other is that observed amounts represent point values, while amounts at model grid points represent area averages. Although these analysis procedures have a large impact on the scores obtained from verification, they are seldom discussed in the literature.

Except where specified, the results presented in this paper are from verification done at the station locations, as recommended in the First Workshop on Model Verification, which took place in Boulder, Colorado, in June 1998. Model output was analyzed to the station locations using a bilinear interpolation involving the four points surrounding a station. Section 4c presents a comparison of verification performed at the grid and at the stations. For such a comparison, the station data were analyzed to the model grid points using a Barnes (1973) analysis. Ninety percent (90%) of the amplitude was retained for waves of 120-km wavelength, and a smaller retaining response was used for shorter waves. Section 4c also presents results using the 36 model grid points surrounding a station to compute forecasted precipitation at the station location. This method is referred to as 6x6, in contrast to the bilinear method used throughout the paper, referred to as 2x2.

4. Results for the Control Runs

a. Verification at 16 hours

The observed and forecasted distributions can be initially compared through the Q-Q plot shown in Fig. 2. The curve in the figure is always above the 1:1 line. This indicates that the forecasts allocated insufficient probability at the high rain values located at the right tail of the distribution, and allocated too much probability at the low rain values.

Table 2. Contingency table for the control run at 2200 UTC.

		OBSERVATIONS (mm)				
		0.0-2.5	2.5-12.5	12.5-25.0	>25.0	
FORECAST (mm)	0.0-2.5	1342	279	72	42	1735
	2.5-12.5	25	30	11	5	71
	12.5-25.0	11	7	3	3	24
	>25.0	8	1	0	0	9
		1386	317	86	50	1839

A contingency table at 16 hours into the run for all days and all 124 stations is shown in Table 2. The maximum possible sample size is 19 days x 124 stations = 2356 cases. In actuality there were 1839 cases, which accounts for stations that did not report. As expected, most of the observed precipitation was concentrated in the lower precipitation categories, with only a few events of higher precipitation amounts. At 2200 UTC (16 h into the forecast), 453 (24.6 %) of the cases had observed precipitation higher than 2.5 mm, versus 104 (5.6 %) cases with forecasted precipitation higher than that threshold. An inspection of observed and modeled precipitation in the higher categories shows that the model had many fewer cases with high precipitation than the observations. For example, the model had 71 cases in which precipitation occurred between 2.5 and 12.5 mm, while precipitation was observed in this category in 317 cases. Figure 3 summarizes the scores for each threshold by the end of the forecast period. Underprediction occurred at all thresholds (as suggested by the Q-Q plot) and was worse at the highest threshold. The BS was 0.23 for the 2.5 mm threshold and 0.18 for the 25.0-mm threshold. The best BS occurred for the 12.5 mm threshold, with 0.24. Precipitation placement was also worse at higher thresholds, with the ESS falling from 0.08 for the 2.5-mm threshold to 0.0 at the 25.0-mm threshold.

In the higher thresholds, there was a decoupling between the model and the observations. As an example, consider in Table 2 the category of precipitation above 25.0 mm. The model had nine cases in this category, while the observations had 50; therefore, the model severely underpredicted at that threshold. But more interesting, on the nine occasions that the model predicted in that category, the observations did not show precipitation in the same category in any. Nine times the observations showed less rain, of which eight observations were for precipitation amounts under 2.5 mm. The converse was also true. Observations showed 86 points in the 12.5-25.0 mm category. But on those occasions, the model produced precipitation higher than 12.5 mm only three times, and on 72 occasions (83.7 % of the 86 events), the model did not forecast above 2.5 mm. This indicates that the predictive ability of the model at the high thresholds was very limited. Although BSs for the set of all days may not be very low for the moderate and extreme rain events (since there is a large number of both predicted and

Table 3. Bias score (BS) for each threshold (mm) for each day at 2200 UTC.

	2.5	12.5	25.0
Jul-18	0.00	-	-
Jul-19	0.33	0.00	0.00
Jul-21	0.03	0.09	0.00
Jul-23	0.71	0.67	2.00
Jul-24	0.21	0.00	0.00
Jul-25	0.13	0.00	0.00
Jul-26	0.23	0.67	-
Jul-28	0.22	0.00	0.00
Jul-30	0.30	0.33	-
Jul-31	0.23	0.31	0.00
Aug-01	0.32	0.46	0.67
Aug-04	0.54	1.33	0.00
Aug-08	0.00	0.00	0.00
Aug-09	0.00	0.00	0.00
Aug-11	0.44	0.55	0.13
Aug-16	0.00	-	-
Aug-17	0.00	-	-
Aug-18	0.00	0.00	0.00
Aug-24	0.00	0.00	0.00

observed points in the high categories), a high daily variability in bias scores is expected, since the observed and modeled extreme events do not coincide. Additionally, TSs, ESSs and other measures of rain location are expected to be low.

Table 3 shows the BS for each of the 19 days studied here. Each column corresponds to a threshold. Some table cells show a dash, which indicates that the BS could not be computed because no precipitation was observed at or above that threshold. This occurred more often for higher thresholds. Large differences existed from day to day, but in general the bias was less than one, indicating that the model tended to underforecast precipitation.

Table 4 shows the TSs, which also indicated a large spread from day to day. At the 2.5-mm threshold, the largest TS was 0.24 and there were nine days with TSs of zero. Larger TSs were attained for the lowest thresholds. For extreme rain events, with thresholds above 12.5 mm, the TSs never exceeded 0.13.

ESSs for different days are listed in Table 5. Several days had negative scores, indicating that the placement of precipitation by the model was worse than by chance. Again, a large spread in ESSs was observed from day to day, and scores were worse at higher thresholds.

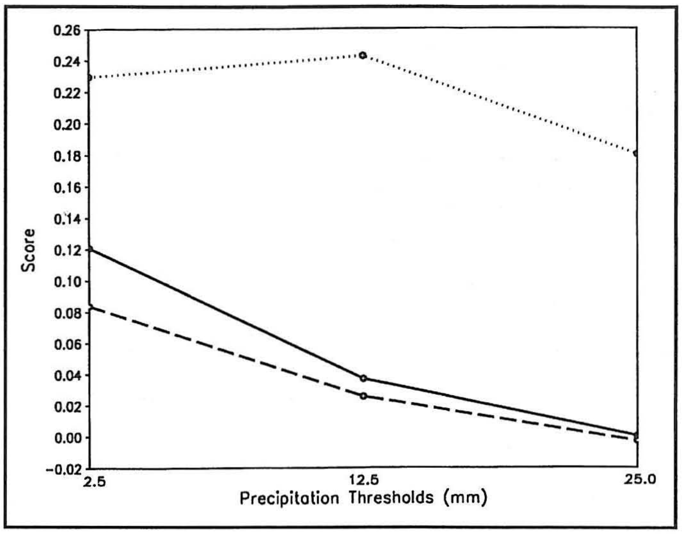

Fig. 3. Bias score (BS, dotted), threat score (TS, solid), and equitable threat score (ESS, dashed) for the 2.5-, 12.5-, and 25.0-mm thresholds for the control run ending at 2200 UTC.

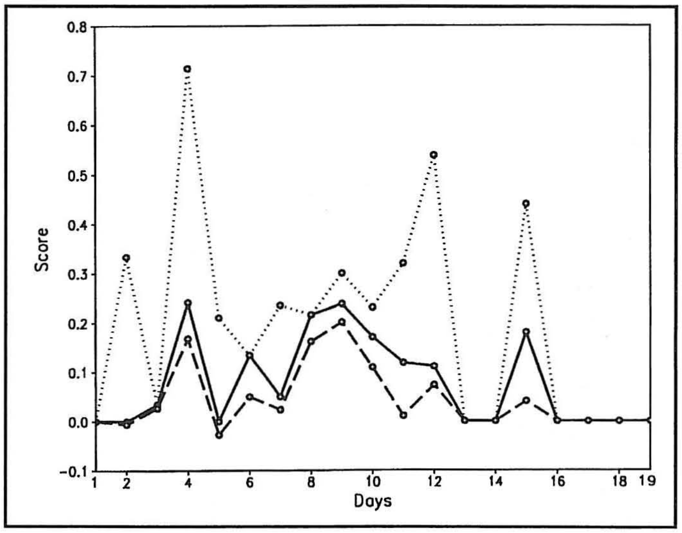

Fig. 4. Daily time series of bias score (BS, dotted), threat score (TS, solid), and equitable skill score (ESS, dashed) for the 2.5-mm threshold for the control run ending at 2200 UTC.

Figure 4 summarizes the daily variation for the BS, TS and ESS for all stations at 16 h for the 2.5-mm threshold. The extreme variability in scores from day to day was noteworthy, especially in the BS, depicting the decoupling between the model and the observations described previously.

b. Hourly distribution of scores

The evolution in time of the number of cases forecasted and observed for the 2.5-mm threshold shown in Fig. 5 is instrumental in the interpretation of the physical nature of the model's underprediction of precipitation. In the first hours of a forecast, the predicted precipitation was close to zero, since the model had not yet developed convective clouds and precipitation. Observed values were also low since the model started at 0600 UTC (0100 EST) normally before the daytime precipitation developed. While the observed numbers increased almost quadratically as the

Table 4. As in Table 3, except for threat score (TS).

	2.5	12.5	25.0
Jul-18	0.00	-	-
Jul-19	0.00	0.00	0.00
Jul-21	0.03	0.00	0.00
Jul-23	0.24	0.11	0.00
Jul-24	0.00	0.00	0.00
Jul-25	0.13	0.00	0.00
Jul-26	0.05	0.00	0.00
Jul-28	0.22	0.00	0.00
Jul-30	0.24	0.00	-
Jul-31	0.17	0.13	0.00
Aug-01	0.12	0.00	0.00
Aug-04	0.11	0.00	0.00
Aug-08	0.00	0.00	0.00
Aug-09	0.00	0.00	0.00
Aug-11	0.18	0.11	0.00
Aug-16	0.00	-	-
Aug-17	0.00	-	-
Aug-18	0.00	0.00	0.00
Aug-24	0.00	0.00	0.00

Table 5. As in Table 3, except for equitable skill score (ESS).

	2.5	12.5	25.0
Jul-18	0.00	-	-
Jul-19	-0.01	0.00	0.00
Jul-21	0.03	-0.01	0.00
Jul-23	0.17	0.09	-0.01
Jul-24	-0.03	0.00	0.00
Jul-25	0.05	0.00	0.00
Jul-26	0.02	-0.01	0.00
Jul-28	0.16	0.00	0.00
Jul-30	0.20	-0.01	-
Jul-31	0.11	0.11	0.00
Aug-01	0.01	-0.03	-0.02
Aug-04	0.07	-0.01	0.00
Aug-08	0.00	0.00	0.00
Aug-09	0.00	0.00	0.00
Aug-11	0.04	0.05	-0.01
Aug-16	0.00	-	-
Aug-17	0.00	-	-
Aug-18	0.00	0.00	0.00
Aug-24	0.00	0.00	0.00

day progressed, the forecasted numbers increased almost linearly. The result is a growing difference between the two curves indicating that the model failed to develop the diurnal cycle of observed precipitation, with its increase in areal coverage during the afternoon hours.

Figure 6 shows the verification scores by forecast hour for the 2.5-mm threshold. The BS was always less than one, since the model was consistently underpredicting for this threshold. The increasing trend in BS in the first hours of the model forecast can be used as a measure of spin-up time (Colle et al. 1999). From Fig. 6, we can infer that the model took approximately nine hours to 'spin-up,' that is, to develop clouds and precipitation. The BS reached a maximum of 0.34 at 1600 UTC. The TS and ESS gradually increased in time, indicating that the model increased its capability of forecasting precipitation location throughout the forecast period. The maximum TS and ESS were 0.12 and 0.08, respectively, indicating the poor accuracy of the model, even for the lowest precipitation thresholds.

c. Verification at the model grid points versus at the stations

The scores presented in the previous and following sections were computed at the observing stations, after the forecasted precipitation was interpolated to the station locations using the 2x2 method described in Section 3. In the literature, one finds studies in which scores were computed at the stations (e.g., Gaudet and Cotton 1998; Colle et al. 1999), and others in which the scores were computed at the model grid points (e.g., Black 1994; Zhao et al. 1997), and still others which do not mention where the scores were computed. Moreover, seldom does one find a discussion about the choice of methodology (Gaudet and Cotton 1998), or about the differences between results obtained with either methodology (Briggs and Zaretzki 1998; Bernardet 2000). Gaudet and Cotton (1998) justified the calculation of forecast verification scores at station locations as a measure to avoid smoothing the observed values with an interpolation to model grid points. The difference between their method and the one used in this paper is that, for a forecasted value, they used the nearest model grid point, and not an average of the surrounding points. This difference in method can certainly impact the verification scores. This impact was evaluated by Briggs and Zaretzki (1998), who discussed the influence on the scores of gridding errors introduced by different algorithms used to interpolate/extrapolate the observed data to the model grid points. They concluded that verification at the station locations instead of at the model grid points is the

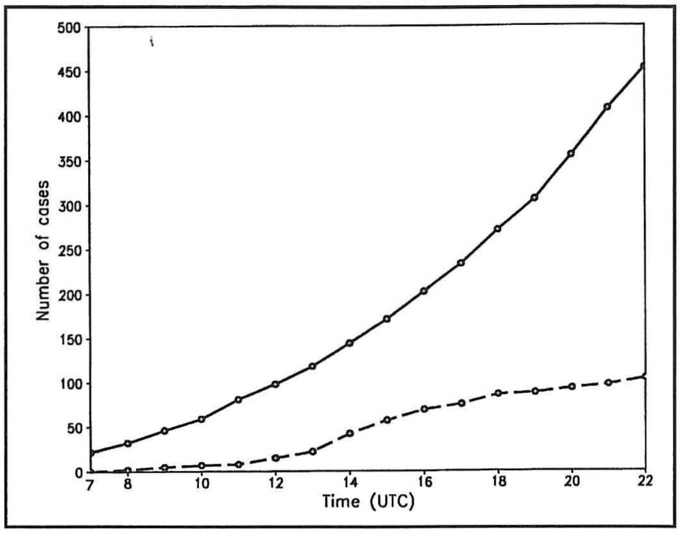

Fig. 5. Hourly time series of the number of cases with observed (solid) and forecast (dashed) precipitation ≥ 2.5 mm, for the control run.

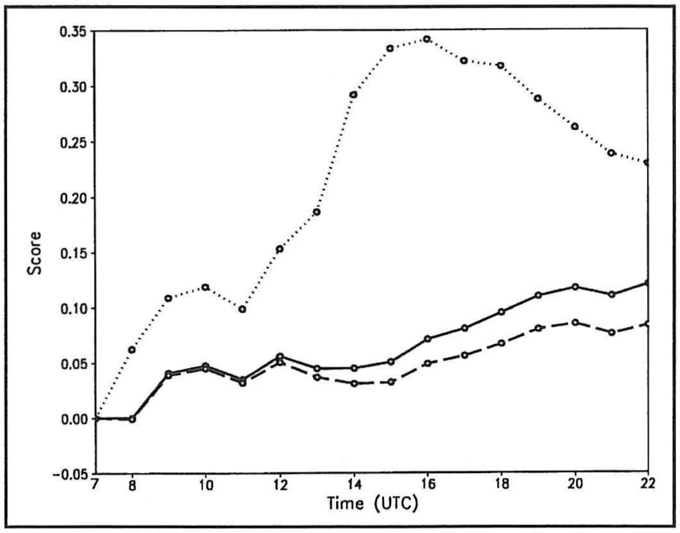

Fig. 6. Hourly time series of bias score (BS, dotted), threat score (TS, solid), and equitable skill score (ESS, dashed) for the 2.5-mm threshold for the control run.

method that introduces the least amount of errors. Bernardet (2000) showed that the BSs can be significantly different when verification is performed at the grid or at the stations. She also discussed the impact of different algorithms for interpolation of model forecasts to station locations and showed that the algorithms that use a larger number of model grid points to compute the forecast at a station location yielded higher BSs, since they increase the probability of a station being influenced by a non-zero forecast.

Figure 7 contrasts the BSs for the 2.5-, 12.5- and 25.0-mm thresholds obtained at the model grid with those obtained at the station locations. For all thresholds, the BS computed at the grid (BS_{grid}) was larger than the BS computed at the station locations (BS_{st}) at almost all times. The largest difference occurred for the highest threshold. At 1100 UTC it reached 1.04 for the 25.0-mm threshold. For the 12.5-mm threshold, the largest difference was 0.42 at 1500 UTC. By the end of the forecast period, the difference had diminished to 0.09 for the 12.5-mm threshold and 0.20 for the 25.0-mm threshold.

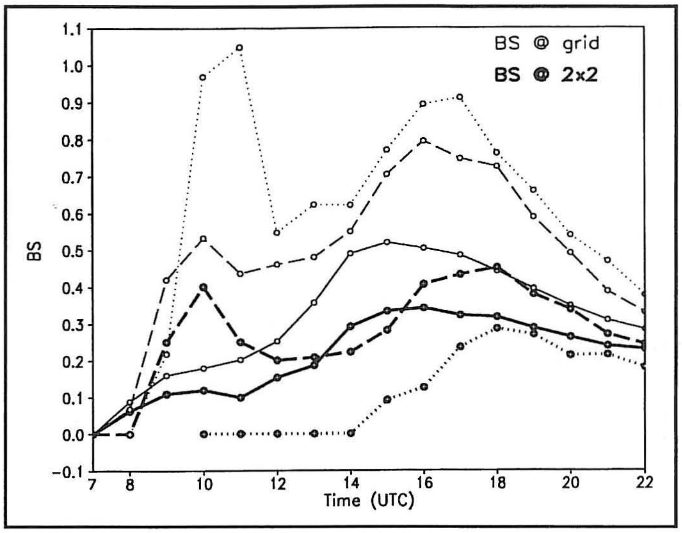

Fig. 7. Hourly time series of bias score (BS) for the 2.5- (solid), 12.5- (dashed), and 25.0-mm (dotted) thresholds computed at the model grid points (thin lines) and at the station locations (thick lines), for the control run.

The cause of the difference between BS_{st} and BS_{grid} can be understood using an idealized forecast and observing system, composed of a domain of 256 (16 x 16) grid points with nine rain gauges. Imagine a situation in which the model forecasted precipitation at 121 (11 x 11) grid points. When this forecast is interpolated to the stations using the 2x2 method, four stations are forecasted to have precipitation (Fig. 8a). Suppose, furthermore, that the verification dataset for that day had rain at four stations, and the objective analysis spread the observed rain over 121 grid points (Fig. 8b). In this case BS_{grid}=121/121 and BS_{2x2}=4/4, so $BS_{grid} = BS_{2x2} = 1$. A similar situation may be envisioned in which $BS_{grid} = BS_{2x2}$, even if BS $\neq 1$.

A distinct situation is now presented, for which $BS_{grid} \neq BS_{2x2}$. Assume that the model forecasted precipitation at 42 (7 x 6) grid points, or one station (Fig. 9a). Assume, furthermore, that rain was observed at four stations, 121 (11 x 11) grid points (Fig. 9b). The scores for this case are BS_{grid} = 42/121 = 0.35 and BS_{2x2} = 1/4 = 0.25, therefore $BS_{grid} > BS_{2x2}$.

These simple examples show that the relative magnitudes of the BS_{grid} and the BS_{2x2} can be determined by the spread of the model forecasted precipitation over the stations. Since the model grid spacing is eight km, a forecasted event using the 2x2 method can only extend for eight km. As a consequence, it is common that model forecasted events cover few observing stations, driving the BS_{2x2} down. To support this idea, the BS_{6x6} was computed. Through this method, forecasted precipitation events can extend over 3x8=24 km and potentially cover a larger number of stations. In the example showed in Fig. 9a, four stations receive precipitation if the 6x6 method is used, therefore BS_{6x6} = 4/4 = 1.0, $BS_{grid} = BS_{2x2}$, and $BS_{6x6} > BS_{2x2}$.

BSs computed using the 6x6 method for the actual forecasts are shown in Fig. 10 for the 2.5-, 12.5- and 25.0-

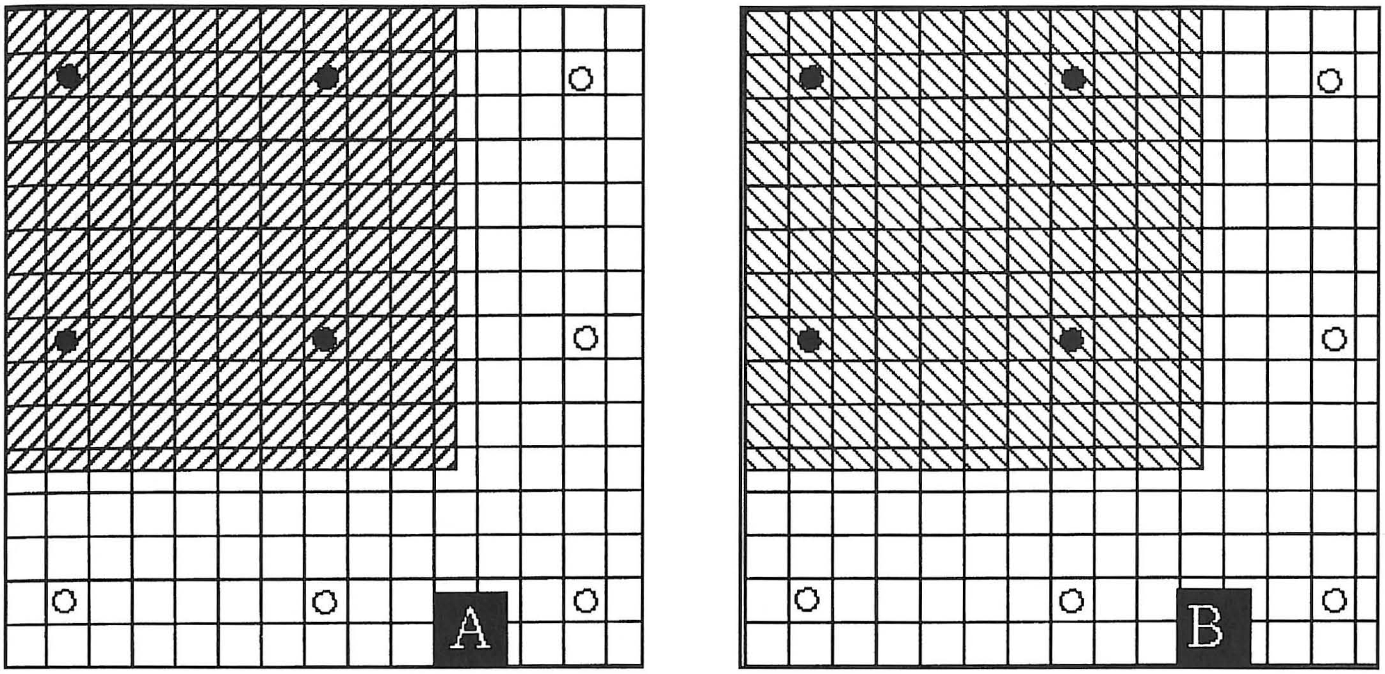

Fig. 8. Schematic model grid and gauge locations. a) The shaded region represents the area covered by forecasted precipitation. The filled (open) dots are gauges with (without) interpolated forecasted precipitation. b) The filled (open) dots represent gauges with (without) observed precipitation. The shaded region represents the area covered by observed precipitation analyzed to the model grid.

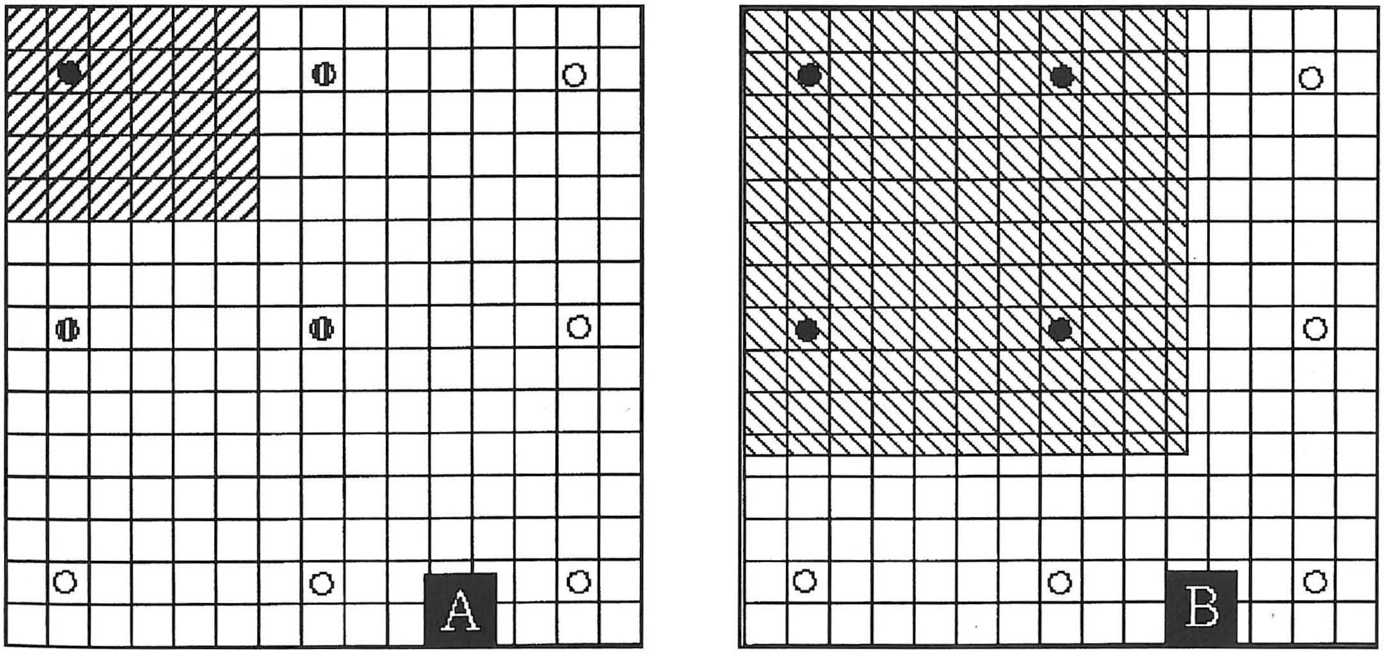

Fig. 9. As in Fig. 8, but for a different event. The half-filled dots in (a) represent gauges that receive interpolated precipitation when the 6x6 method is used.

mm thresholds. In general the scores were higher using the 6x6 method, since a larger number of stations with forecasted precipitation was generated.

The TS and the ESS are also dependent on the choice of location for verification, and on the algorithms used. Consider again the idealized model and observation system shown in Figs. 9a and b. One could consider "shifting" the location of the model forecasted precipitation in such a way as to keep it within the range of one station, but altering the number of model grid points with correct forecasts. This exercise would alter the TS and ESS computed at the grid (TS_{grid} and ESS_{grid}, respectively) without changing their counterparts computed at the station locations, and would yield situations in which the scores computed at the grid are different than the ones computed at the stations.

A comparison of ESSs computed at the grid and at the stations for the actual forecasts is presented in Fig. 11.

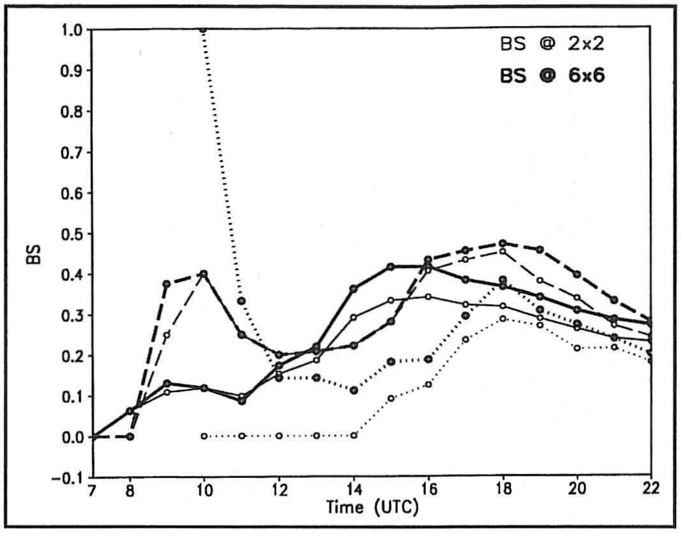

Fig. 10. Hourly time series of bias score (BS) for the 2.5- (solid), 12.5- (dashed), and 25.0-mm (dotted) thresholds computed at the stations when using the 2x2 (thin lines) and the 6x6 (thick lines) methods, for the control run.

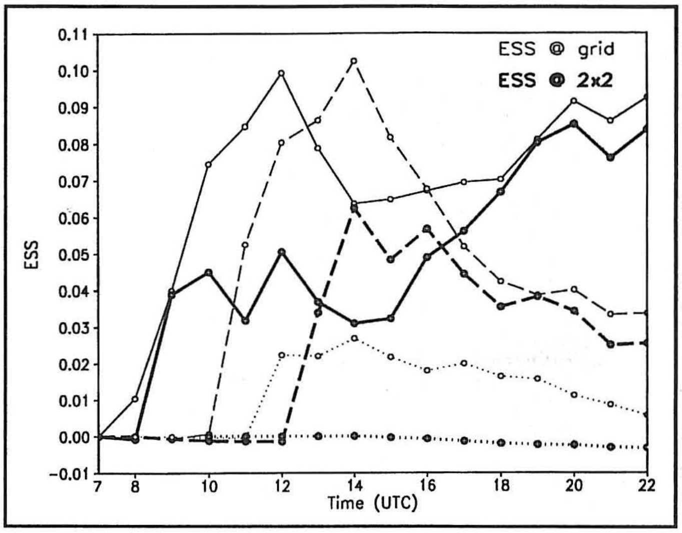

Fig. 11. Hourly time series of equitable skill score (ESS) for the 2.5- (solid), 12.5- (dashed), and 25.0-mm (dotted) thresholds computed at the model grid points (thin lines) and at the station locations (thick lines), for the control run.

For the 2.5-, 12.5- and 25.0-mm thresholds, $ESS_{grid} > ESS_{2x2}$ at almost all times. This indicates that for a given number of stations with correct forecasted precipitation, the area with forecasted precipitation in the model largely superposed the area with observed analyzed precipitation.

d. Spatial distribution of scores

The spatial distribution of scores at the 2.5-mm threshold is shown in Fig. 12 for two different times: 1600 UTC, 10 h into the run, and 2200 UTC, 16 h into the run. These times correspond to the maximum BS and to the end of the forecast period as seen in Fig. 6.

The spatial distribution of ESSs showed a correlation with topography. The highest values of the ESS at 1600 UTC were found along the Appalachian Mountains along the Tennessee-North Carolina border, on the South Carolina-North Carolina border and in the Savannah River valley along the border between Georgia and South Carolina. Throughout the rest of the domain, ESSs were close to zero. At 2200 UTC, the higher ESSs were still located on the Appalachian Mountains and in the Savannah River valley, but there was also a maximum that extended northeastward from the Savannah River valley, parallel to the Atlantic Coast, approximately 120 km inland, covering central South Carolina. A local maximum was also present in central Alabama, in the westernmost part of the domain.

The BSs also had a high spatial variability that could be correlated both with topography and continentality. At 1600 UTC, values varied between zero and 2.0. The maximum BSs were located in a band oriented parallel to the Atlantic Coast. This region coincides with the gentle slope of the terrain, leading from the coast to the Appalachians along northern Georgia and South Carolina, and with the location where the highest ESSs were found at 2200 UTC. Relatively high BSs were also found on the northern boundary of the model domain, in the Appalachian Mountains. Near zero BSs were found near the coast, indicating that the model never produced precipitation at these stations by this time.

At 2200 UTC, BS values continued to vary between zero and 2.0. BSs near 1.0 were found along the Appalachians, and a band of relatively high BSs (up to 2.0) extended through central Georgia and South Carolina, oriented parallel to the Atlantic Coast, as it did at 1600 UTC.

5. Forecasts Made *A Posteriori*

In this section we discuss two sets of retrospective runs made with model configurations different than the control forecasts. Details of the configurations were discussed in Section 2. Experiment $LAPS_{init}$ was initialized at 1500 UTC with the LAPS analysis and experiment Eta_{init} was initialized at 1500 UTC with the Eta analysis. All verification scores were computed at the station locations, and the 2x2 method was used to analyze the forecasted data to the stations.

a. Runs initialized at 1500 UTC with the LAPS analysis

The Q-Q plot in Fig. 13 shows a behavior quite different than the one displayed by the control run. At the left tail of the distribution, the points fell close to the 1:1 line, indicating that the model had the correct number of points forecasted at the low thresholds. However, for higher quantiles, the curve fell below the 1:1 line, indicating overforecasting of precipitation by the model.

Figure 14 shows the hourly evolution of the number of points with forecasted and observed precipitation at the 2.5-mm threshold for the $LAPS_{init}$ experiment. Note that the number of observed points was different than the one for the control experiment (Fig. 5). This happened because the hours of accumulation for each experiment

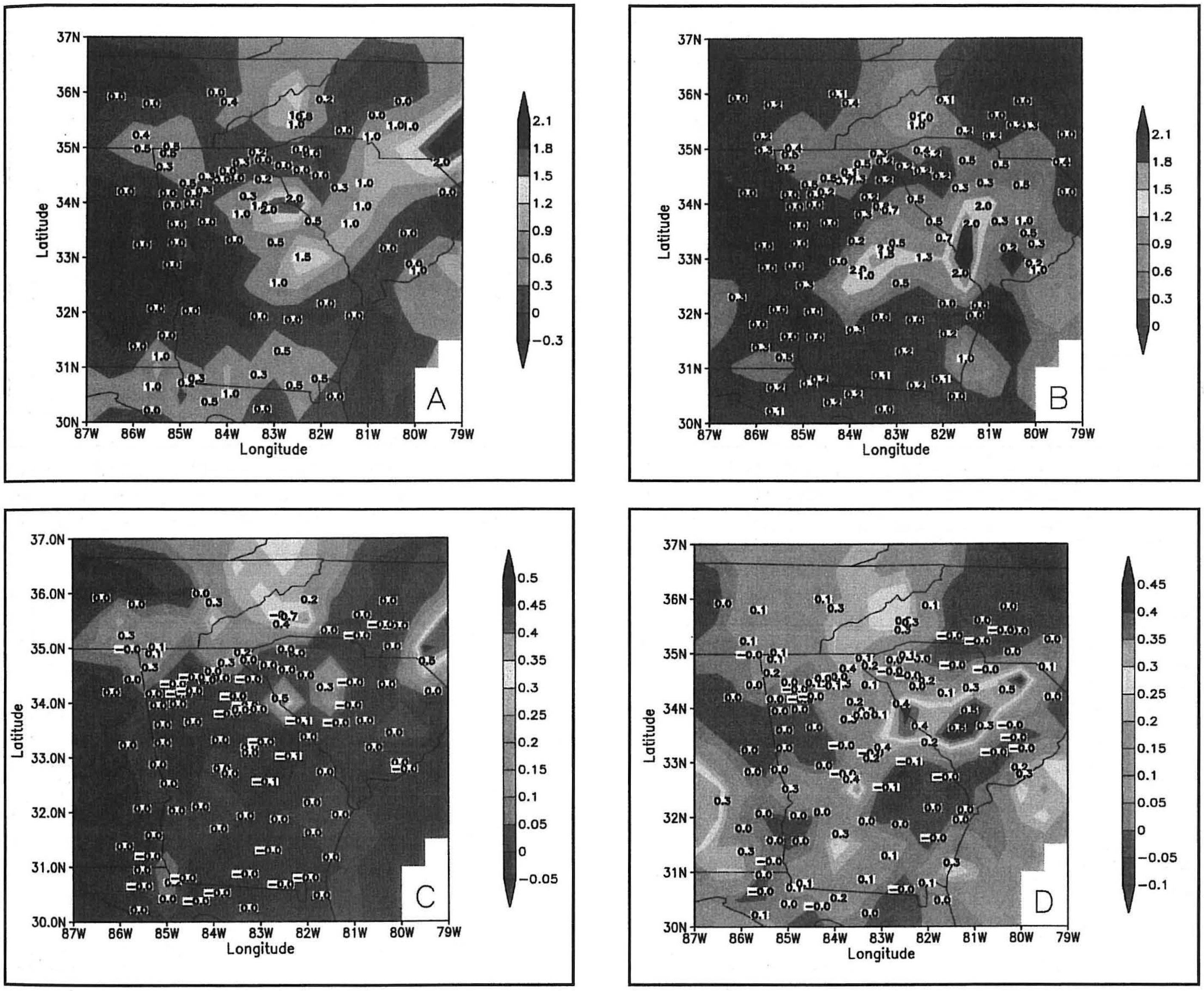

Fig. 12. Spatial distribution of bias score (BS) and equitable skill score (ESS) for the 2.5-mm threshold at the 10-h and 16-h forecast times in the control run: (a) 1600 UTC BS; (b) 2200 UTC BS; (c) 1600 UTC ESS; and (d) 2200 UTC ESS.

were different, since the control experiment was initialized at 0600 UTC and the LAPS$_{\text{init}}$ experiment, at 1500 UTC. At 16 hours into the forecast, 559 points registered observed precipitation, 106 more than for the control experiment. The difference is due to the precipitation regime, which is characterized by afternoon clouds and precipitation development. The number of points with forecasted precipitation was always inferior to the observed number, yielding a BS lower than one.

A comparison of Figs. 6 and 15 indicates that the LAPS$_{\text{init}}$ BS was always larger than the BS for the control forecasts, suggesting that the model was capable of producing a more realistic number of precipitation points when it was initialized at 1500 UTC. Although there was consistent underforecasting, the BS increased monotonically with time, denoting that the model was keeping up with the precipitation development. The BS for the 2.5-mm threshold ended the forecast period at 0.66. The ESS for this threshold at the end of the forecast period (Fig. 15) was slightly higher (0.10) than the one for the control forecasts. The TS was larger than the one for the control forecasts, which reflects the influence of a higher bias in that score (Schaefer 1990). The scores indicated that a better positioning of the late afternoon precipitation was achieved with a 1500 UTC initialization rather than with a 0600 UTC initialization. At 2200 UTC, LAPS$_{\text{init}}$ showed a BS of 0.26 and a ESS of 0.04, while the control forecasts had a BS of 0.23 and an ESS of 0.08. This very small difference is partially caused by the fact that the ESS increased in time and 2200 UTC is just seven hours into the LAPS$_{\text{init}}$ forecasts, still within the spin-up period.

Figure 16 shows that the BS at 16 h into the run increased with threshold. The BS was larger than 1 for the 12.5- and 25.0-mm thresholds, reflecting excessive areas of forecasted rain at the higher thresholds. The Q-Q plot also pointed to overforecasting at the right end of

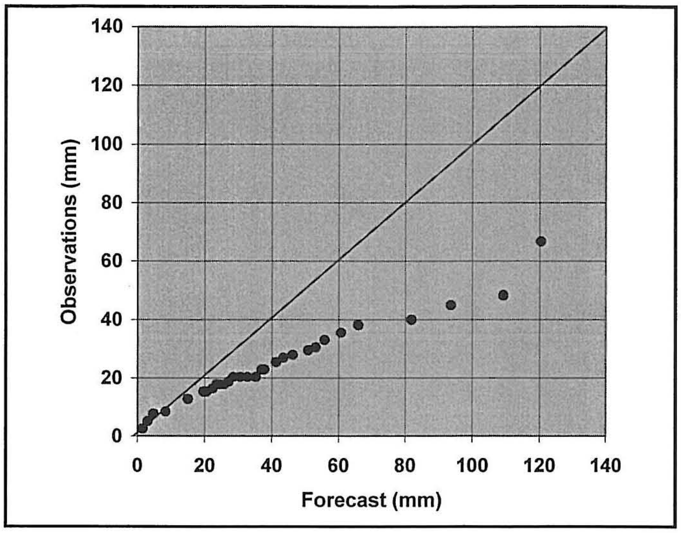

Fig. 13. Quantile-quantile plot as in Fig. 2, except for the LAPS$_{\text{init}}$ experiment with forecast period ending at 0700 UTC.

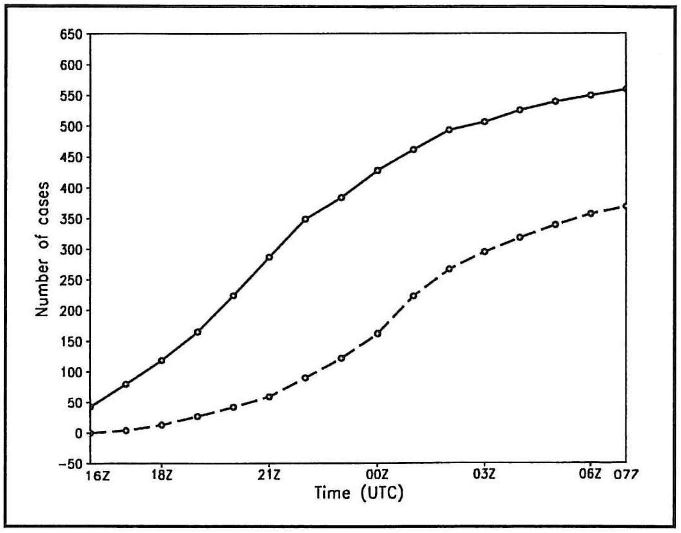

Fig. 14. Hourly time series of the number of cases with observed (solid) and forecast (dashed) precipitation ≥ 2.5 mm, as in Fig. 5, except for the LAPS$_{\text{init}}$ experiment with forecast period ending at 0700 UTC.

the distribution. The TS and the ESS decreased for higher thresholds, indicating that the model had the best performance placing precipitation at the lower thresholds.

b. Runs initialized at 1500 UTC with the 29-km Eta model analysis

The Q-Q plot for this experiment is shown in Fig. 17. The behavior in this case was different than both cases discussed previously. For the lower thresholds, the curve fell above the 1:1 line, indicating that the model was allocating too much probability to low values. For values in the center of the distribution, the model allocated the correct amount of probability. However, at the right tail of the distribution, the curve fell below the 1:1 line. This would suggest, as discussed for the LAPS$_{\text{init}}$ experiment, that the model had too many points with high precipitation amounts. However, this curve must be interpreted in

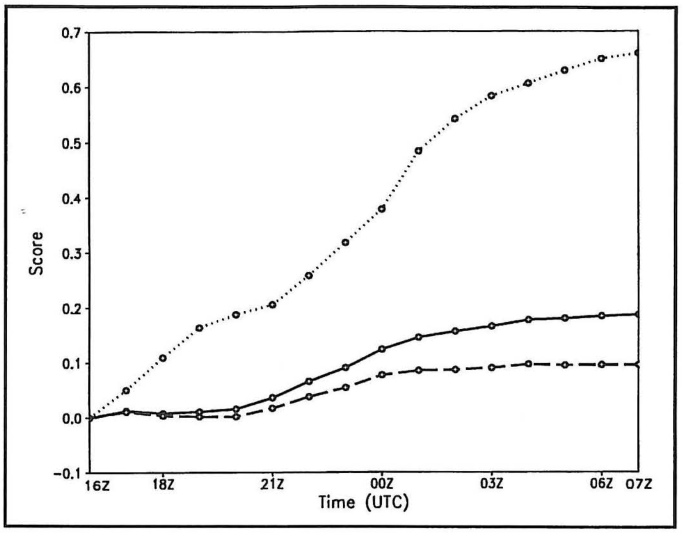

Fig. 15. Hourly time series of bias score (BS, dotted), threat score (TS, solid), and equitable skill score (ESS, dashed) for the 2.5-mm threshold as in Fig. 6, except for the LAPS$_{\text{init}}$ experiment.

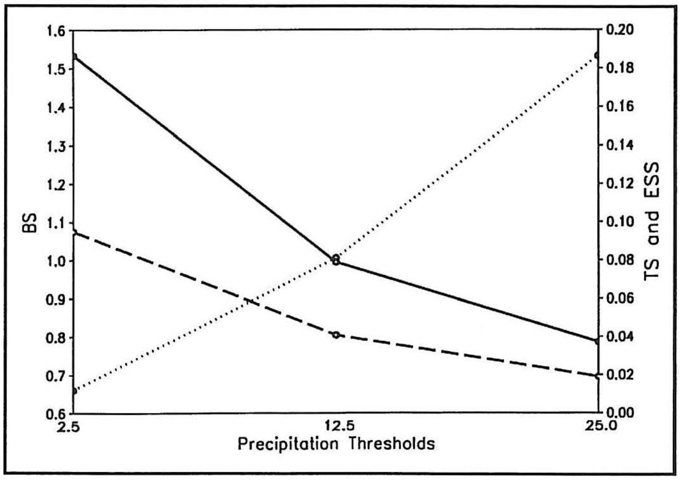

Fig. 16. Bias score (BS, dotted), threat score (TS, solid), and equitable threat score (ESS, dashed) for the 2.5-, 12.5-, and 25.0-mm thresholds as in Fig. 3, except for the LAPS$_{\text{init}}$ experiment.

the light of the methodology used to select the data to compose it. As discussed in Section 3, all cases with precipitation amounts less than 0.5 mm were excluded from the data series. Although in the previous cases discussed, this exclusion left the observed and forecasted series with a similar number of cases, for the Eta$_{\text{init}}$ experiment the modeled series was left with approximately half the number of the observed series. Consequently, caution is needed to interpret the right tail of the distribution in the Q-Q plot. The verification scores described below will aid this interpretation.

The hourly evolution of the number of points with forecasted and observed precipitation for the Eta$_{\text{init}}$ experiment is shown in Fig. 18. Note that the number of observed points was similar to the one for the LAPS$_{\text{init}}$ experiment (Fig. 14). This was a coincidence, since although both experiments comprised 21 days, the days chosen for each experiment were different, because the

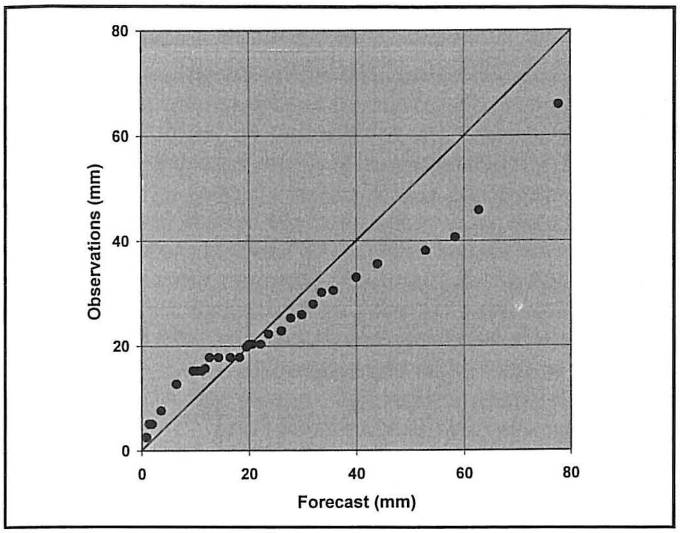

Fig. 17. As in Fig 13, except for the Eta$_{\text{init}}$ experiment.

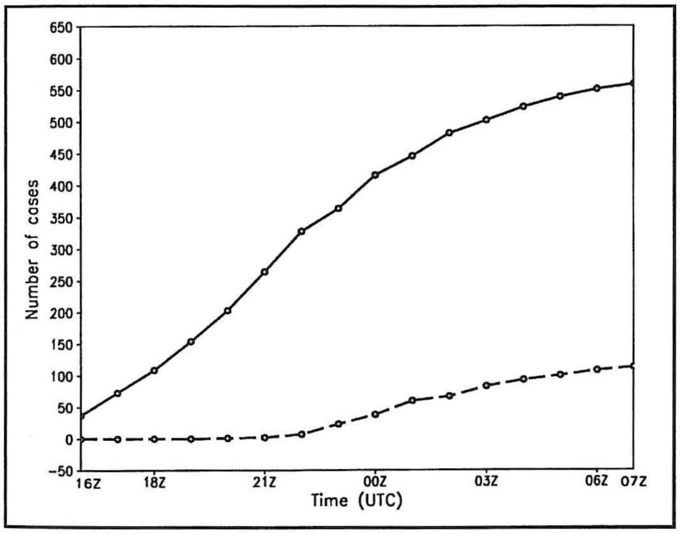

Fig. 18. As in Fig. 14, except for the Eta$_{\text{init}}$ experiment.

days with initialization data available were different. The number of points with forecasted data in the Eta$_{\text{init}}$ experiment, on the other hand, was significantly different than the one from the LAPS$_{\text{init}}$ experiment. The Eta-initialized model had no points with forecasted precipitation at or above the 2.5-mm threshold up to six hours into the run, and thereafter started producing precipitation quite slowly. The result was a BS that never got above 0.20.

The time series of the BS, TS and ESS at the 2.5-mm threshold is shown in Fig. 19. The BS increased monotonically with time, but the underforecasting was more pronounced than for the LAPS$_{\text{init}}$ experiment (Fig. 15) and for the control forecasts (Fig. 6). The TS also increased with time, to end the forecast period at 0.07. The ESS peaked at 0300 UTC with 0.04 and ended the forecast period at 0.03. The values for TS and ESS were considerably lower than their counterparts for the LAPS$_{\text{init}}$ experiment (Fig. 15), indicating that for this precipitation threshold the LAPS$_{\text{init}}$ experiment produced a superior forecast.

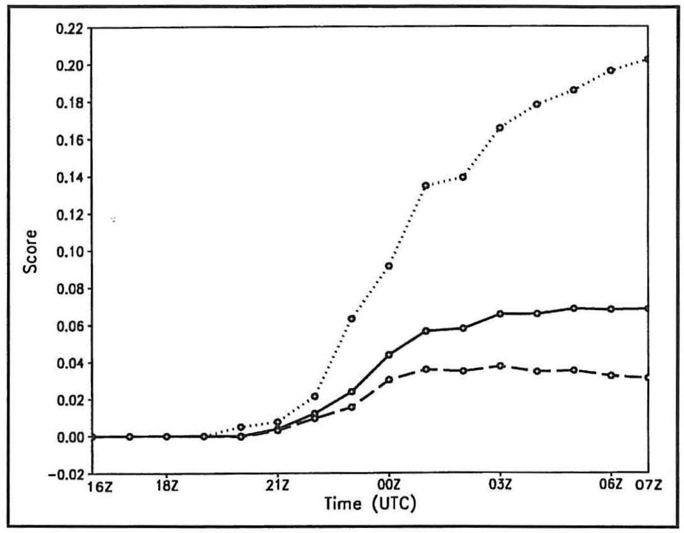

Fig. 19. As in Fig. 15, except for the Eta$_{\text{init}}$ experiment.

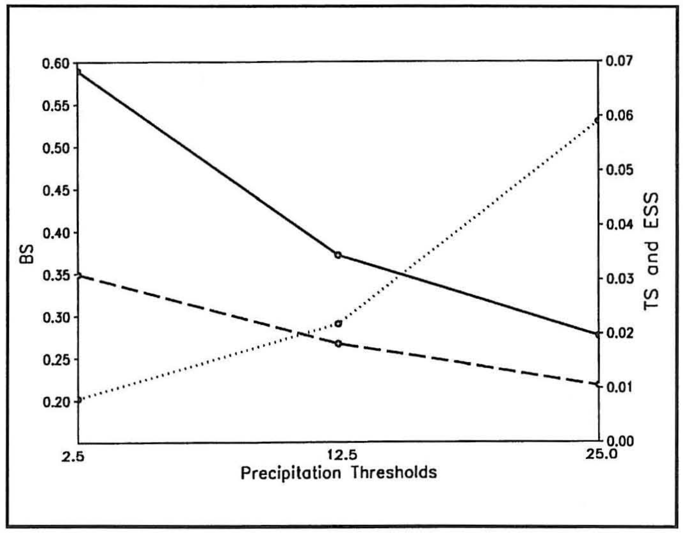

Fig. 20. As in Fig. 16, except for the Eta$_{\text{init}}$ experiment.

As with the LAPS$_{\text{init}}$ experiment, the BS for the Eta$_{\text{init}}$ experiment increased with precipitation threshold (Fig. 20). BSs for all thresholds were smaller than one. Referring back to the Q-Q plot, we note that the low values of probability allocated at the right end of the distribution were related to the large number of forecasted cases with rain amounts less than 0.5 mm, which were excluded from the distribution, and not to overforecasting at high thresholds. The TS and ESS decreased with threshold. The TS and ESS for all thresholds at 16 h were lower than their LAPS$_{\text{init}}$ counterparts.

6. Comparison with Forecasts from the 29-km Eta Model

In this section a verification of the forecasts from the 29-km Eta model initialized at 1500 UTC is performed, and a comparison with the results from the last section is presented. Details of the Eta model can be obtained from Black (1994). Verification results are available only every

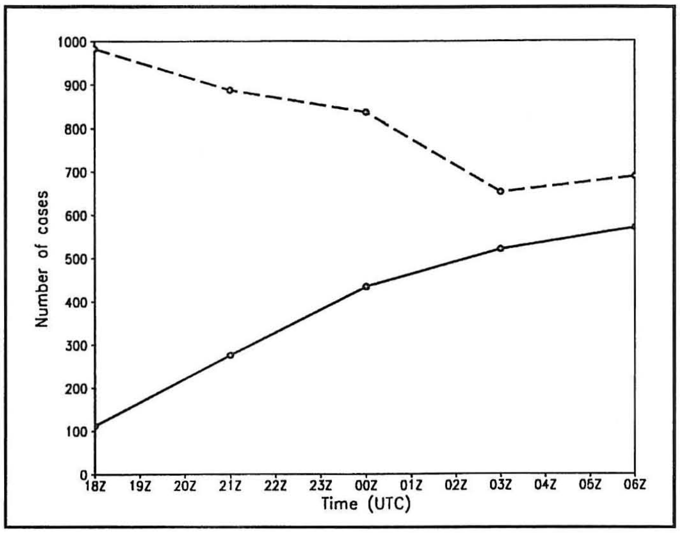

Fig. 21. As in Fig. 14, except for the 29-km Eta model.

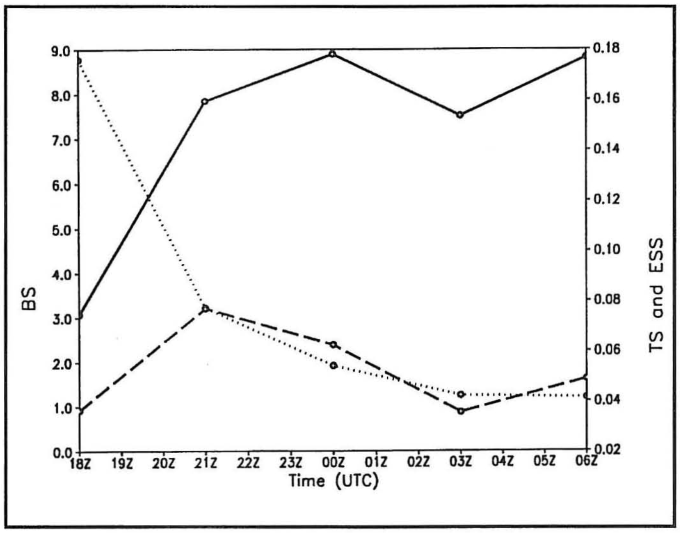

Fig. 22. As in Fig. 15, except for the 29-km Eta model.

three hours, since that was the frequency with which model output was stored.

All verification scores were computed at the station locations, and the 2x2 method was used to analyze the model data to the stations. Twenty-two days during the Olympic Exercise were available for the computation of the verification scores.

Figure 21 shows the evolution in time of the number of cases with observed and modeled precipitation at or above the 2.5-mm threshold. The curve of observed cases is similar to the one of experiment LAPS$_{\rm init}$; the number of cases increased throughout the forecast period, to reach 570 by 0600 UTC. The forecasted curve, however, behaved quite differently than the results described for the previous experiments. It decreased until 0300 UTC, and increased thereafter. The number of cases with forecasted precipitation was much higher than the observed number, especially in the early hours. Further investigation of the data (not shown) indicated that the large majority of forecasts early in the period was associated with convective precipitation. This suggests that the convective parameterization of the Eta model may have been activated spuriously in the first hours of forecast, possibly by gravity waves or other imbalanced circulations, a typical problem of the first few hours after model initialization.

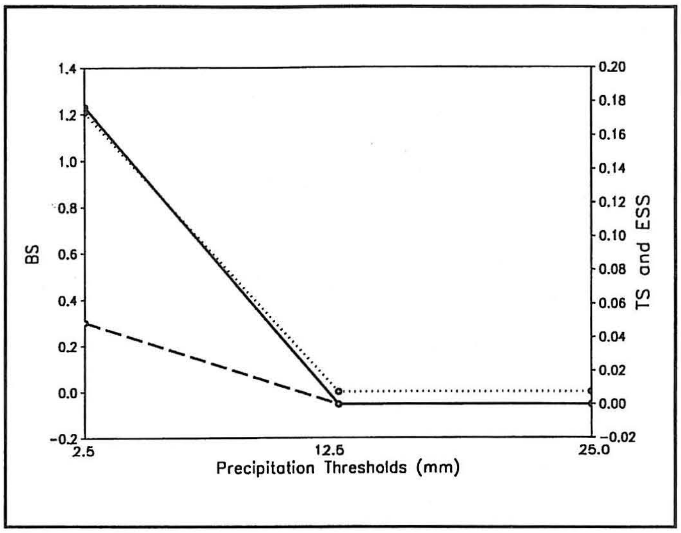

Fig. 23. As in Fig. 16, except for the 29-km Eta model.

Figure 22 shows the hourly evolution of the BS, TS and ESS for the 2.5-mm threshold. As expected, the BS was large especially in the first hours of the forecast, it reached 8.78 at 1800 UTC, and overforecasting during the whole period. The TS increased in the first hours of forecast, had a minimum at 0300 UTC and increased again by 0600 UTC. The ESS peaked at 2100 UTC, and then decreased, to peak again at 0600 UTC. The ESS was comparable to the ones for LAPS$_{\rm init}$ and Eta$_{\rm init}$. A comparison indicated that the ESS for the Eta model was higher than its counterparts in the beginning of the forecast period (until 2100 UTC) and at 0600 UTC, the ESS was 0.05, which fell between the value of 0.1 for LAPS$_{\rm init}$ and 0.03 for Eta$_{\rm init}$ The values of TS were three times higher than the ones for Eta$_{\rm init}$, which reflected the influence of the bias in the TS (Schaefer 1990). The TSs were higher than the ones for LAPS$_{\rm init}$ up to 0000 UTC, after which the LAPS$_{\rm init}$ TSs were higher.

The variation in 0600 UTC scores with threshold is shown in Fig. 23. The overforecasting noted previously for the 2.5-mm threshold was limited to that threshold. There was actually underforecasting at the higher thresholds, with BSs close to zero. Due to the lack of forecasted rain, the TS and ESS were near zero at high thresholds.

7. Discussion

The NWS demonstrated its state-of-the-art techniques for weather forecasting during the 1996 Atlanta Olympic Games. One of the goals of this paper was to verify the precipitation forecasts produced by LAPS during that period. The observed and forecasted precipitation distributions were initially compared using Q-Q plots, and subsequently three scores were used for verification: the BS, to verify the degree of areal overforecasting or underforecasting of precipitation, the TS, to check whether the precipitation was forecasted in the correct location, and the ESS, to compare the model forecasts with those obtained by chance. It must be pointed out that the use of the TS and the ESS for verification of mesoscale forecasts is problematic. These scores are low when the forecasted and observed precipitation areas do not overlap. In a regime of convective precipitation, as is prevalent in the southern U.S. in the summer months, it is possible that the forecast area of precipitation might be a few kilometers offset from the observed precipitation area. Such forecasts will lead to low TSs and ESSs, although they may contain valuable information to operational forecasters about the actual development of precipitation and its characteristics (severity, timing, duration, translation speed, etc.). However, since scores more appropriate for the verification of mesoscale precipitation have yet to be developed, these traditional scores were resorted to in this study.

One of the main findings of verification of the forecasts initialized by the LAPS analysis at 0600 UTC was the consistent underprediction of precipitation. The model took about nine hours to spin up (start producing clouds and precipitation), and at the 2.5-mm threshold, the BS peaked at 0.34 at 1600 UTC and decreased after that. At the end of the period, the highest BS occurred for the 12.5-mm threshold, and the lowest for the 25.0-mm threshold. Forecasts initialized at 1500 UTC with the LAPS analysis showed a different behavior. They developed clouds and precipitation quicker and displayed a BS for the 2.5-mm threshold that continuously increased with time, to peak at the end of the forecast period at 0.66. The amount of precipitation at the 12.5- and 25.0-mm thresholds, however, was too large. The BS at higher thresholds was excessively large, reaching 1.53 for the 25.0-mm threshold. When the Eta model was used for initialization at 1500 UTC, the spin-up of precipitation was much slower. The model did not start producing precipitation until 2100 UTC. This could indicate that a local high-resolution analysis for model initialization better describes the mesoscale boundary layer convergence zones that lead to cloud and precipitation development and thus significantly impacts models that use a 'cold start' (i.e., initialized without clouds and precipitation). The fact that $\text{LAPS}_{\text{init}}$ had a faster spin-up is possibly attributable to the fact that the majority of the supplemental data LAPS used was surface data. By 1500 UTC this data better depicted a mixed boundary layer and thus connectivity to the upper atmosphere.

Verification of the forecasts obtained from the 29-km Eta model itself showed that overforecasting occurred at all times for the 2.5-mm threshold, being worse at the early hours of forecasting, when the BS was as high as 8.78. By 0600 UTC the BS had decreased to 1.21. For higher thresholds, the BS was always close to zero, indicating that the 29-km Eta model did not produce realistic areas of precipitation at that threshold. In summary, at the 2.5-mm threshold, there was underforecasting by the LAPS model, whichever way it was initialized, and overforecasting by the Eta model. Both these biases were worse at the first hours of forecast and improved with time. At higher thresholds, the Eta model severely underestimated precipitation, while $\text{LAPS}_{\text{init}}$ overestimated, and Eta_{init} slightly underestimated. Since the Eta was run with a 29-km horizontal grid spacing, it was somewhat expected that it underestimated the larger amounts of precipitation, because the modeled amounts represent an average over a grid cell area. Models with larger grid spacings, therefore, must produce smaller amounts of precipitation.

The forecasted location of precipitation by these models, expressed by the ESS, was quite poor. The ESS at the end of the forecast period for the control forecasts was 0.08, similar to the ESS for the forecasts initialized at 1500 UTC with the LAPS analysis (0.10). The ESS was lower for the runs initialized with the Eta model analysis (0.03) and for the runs with the 29-km Eta model (0.05).

A large variability of scores was observed from day to day, especially at the higher thresholds, for which the forecasted and observed precipitation was decoupled. The 16-h ESS for the 2.5-mm threshold for the control forecasts varied from -0.01 to 0.20. Large variability was also present in the spatial distribution of scores. The control runs produced the BSs closest to one and had better performance in precipitation placement in the Appalachians and on the gentle slopes of South Carolina and central Georgia, which connect the Atlantic Coast with the mountains to the northwest. The low BSs attained near the shore on the South Carolina-Georgia border suggest that the model is too slow in developing rain as the convective systems forced by the sea breeze move inland.

The verification scores presented are quite low. For the control case, the BS shows that the area covered by forecasted precipitation was not even half of that covered by observed precipitation, and the area covered by correct forecasts was less than 10% of that covered by forecasted plus observed precipitation. This indicates that a lot of improvement is still necessary in the forecasts of summertime precipitation in the Southeast.

The TSs and ESSs attained in this study are also somewhat lower than the ones from other studies discussed in the literature. One reason is that TSs and ESSs tend to be higher for forecasts computed using larger grid spacings because the misplacement of precipitation is not as evident. Precipitation misplacements are only accounted for if they are larger than the model grid spacing. Therefore high-resolution configurations, such as the one used in this paper, are very sensitive to displacements of just a few kilometers. Gaudet and Cotton (1998) presented the verification of 17 24-h precipitation forecasts done by the Regional Atmospheric Modeling System (RAMS) at 16-km grid spacing for April 1995 over Colorado. On average, a BS of 0.93 and a TS of 0.48 were obtained for the 2.5-mm threshold. Colle et al. (1999) also examined high-resolution precipitation forecasts over a particular region, the Pacific Northwest. They compared the performances of the 24-h forecasts by the 12-km Mesoscale Model version 5 (MM5), the 36-km MM5 and the 10-km Eta model during the winter of 1996-1997. They presented their results every six hours, and noted that the model took about 12 h to spin up. During that time, the BSs increased, and then settled around 1.2 for the 2.54-mm threshold, around 1.0 for the 7.62-mm threshold and around 0.7 for the 12.7-mm threshold. The

ESSs peaked 12 h into the forecast and then decreased. For the 2.5-mm threshold, the highest ESS was about 0.34. Higher thresholds yielded lower ESSs. Black (1994) presented scores for 24-h forecasts by the 80-km Eta model and by the 40-km meso-Eta model for November 1993, and showed that the higher resolution meso-Eta model performed better at all thresholds. The ESS peaked at approximately 0.4 for the 6.35-mm threshold. The BS was slightly above 1.0 for thresholds up to 12.70 mm, and quickly decreased for higher thresholds. It is possible that if more sophisticated model physics had been employed (e.g., use of a cumulus parameterization, observed SSTs, and variable soil moisture initialization) the scores obtained might be higher.

The low ESSs obtained for the Olympic forecasts are partially related to the difficulty of the forecasting problem. The forecasts are for summer, when convection is less organized, lowering the predictability. Junker et al. (1989) showed that the TSs for the 190.5-km Limited-area Fine-mesh Model (LFM) and for the 90-km Nested Grid Model (NGM) run by the NWS National Centers for Environmental Prediction (NCEP; then National Meteorological Center, NMC) displayed a pronounced reduction during the summer months. Schwartz and Benjamin (1998) presented the BS and the ESS for 13 24-h forecasts done by the 60-km Rapid Update Cycle-2 (RUC-2) model and by the 48-km Eta model during summer 1997. The ESSs peaked at 0.16 for both the 6.35-mm threshold in the 48-km Eta model and the 2.5-mm threshold in the RUC-2 model.

The low BSs obtained in this study may be partially attributed to two other factors. First, unless where specified, the scores presented in this paper were computed at the station locations using four model grid points surrounding a station to compute the forecasted precipitation at that station. It was shown that when a larger number of model grid points is used or when the verification is performed at the model grid points instead, larger BSs are obtained. Du et al. (1997) mentioned (but did not show or discuss) that better verification results were obtained when the scores for an extreme precipitation case study were computed at the model grid. Second, the rain gauges used for the verification dataset only register amounts at 2.5 mm increments. It is possible, therefore, that in the beginning of the forecast period, there was already up to 2.4 mm accumulated in the gauge. In that case, 0.1 mm of rain may fall, and the model may correctly forecast this amount, but the gauge will register 2.5 mm, thus producing an underforecast. Colle et al. (1999) provide a thoughtful discussion of this problem.

8. Conclusions

The 16-h forecasts by LAPS initialized at 0600 UTC during the 1996 Olympic Games underpredicted precipitation at all threshold amounts studied. In general, less than half of the area covered by precipitation was forecasted. Location of precipitation was correctly forecasted in less than 10% of the area of forecasted plus observed precipitation. The underforecasting was related to a long spin-up time, and to the inability of the model to keep up with the increased area of observed precipitation in the afternoon hours. Scores that measure placement of precipitation improved in time, indicating that the model did not lose predictability during the forecast period. Initialization of the model at 1500 UTC reduced the problem of underforecasting at the light thresholds, although it created overforecasting at the high thresholds.

Forecasts initialized with the 29-km Eta model analysis at 1500 UTC took much longer to develop precipitation, stressing the importance of a high-resolution initialization for models with a cold start. The forecasts initialized with the LAPS analysis had better performance, as measured by BS and ESS, than the ones initialized with the Eta model analysis.

A comparison of the high-resolution forecasts, initialized with the 8-km LAPS data with the forecasts by the 29-km Eta model, both initialized at 1500 UTC, showed that the placement of precipitation by the LAPS model was slightly better at the end of the forecast period. At the higher thresholds, the LAPS model was able to produce precipitation, which the 29-km Eta was not. At the 2.5-mm threshold both models had problems at the first hours of forecasts. The LAPS model took several hours to start producing precipitation, while the 29-km Eta model, which used a cumulus precipitation, had excessive amounts of rain. Later in the forecast, the BS of the LAPS model increased and that of the Eta model decreased, so both ended the forecast period closer to one, but the placement of precipitation by the LAPS model was better.

A comparison of verification scores using different algorithms to interpolate the forecasted precipitation at the station locations was presented. It was shown that the BS increased when a larger number of grid points surrounding a station was used to compute the forecasted precipitation at that location. It was also shown that the BS computed at the station locations was lower than the BS computed at the model grid points.

Acknowledgments

Mr. Mark Govett and Dr. Edward Tollerud provided the observed precipitation data used for verification. The LAPS analyses for forecast initialization were products of the group effort of the Local Analysis and Prediction Branch of FSL, headed by Dr. John McGinley. The 29-km Eta model analyses for model initialization were obtained from Dr. Adrian Marroquin. Dr. Edward Tollerud, Dr. John McGinley, Mr. James P. Edwards and Ms. Nita Fullerton reviewed earlier versions of this manuscript and provided helpful comments. Reviewers made numerous suggestions that led to great improvement of the manuscript.

Author

Lígia R. Bernardet has been involved in many different areas of the Atmospheric Sciences. During her College years, in Sao Paulo, Brazil, she worked for a Television Station in broadcast meteorology. After graduation, she moved on to work forecast shifts at a State Forecast Center in the countryside of Brazil. The Center focused on meteorological support for farming. She obtained her Masters Degree from the University of Sao Paulo in 1992 accomplishing a numerical study of the regional circulation in that area. She obtained her Ph.D. in Atmospheric Sciences from Colorado State University in 1997, focusing on a multi-scale numerical study of strong winds generated by thunderstorms (derechoes). She then stayed for two years at NOAA Forecast Systems Laboratory (FSL), holding a National Research Council Associateship, and studying the performance of a local model in precipitation forecasting. During 2000 and 2001 she worked for the Brazilian National Weather Service (INMET) in Brasília, Brazil, implementing their first operational numerical forecast system. She currently works for NOAA FSL (325 Broadway R/FS1, Boulder, CO 80305; ligia.bernardet@noaa.gov), testing and evaluating the new Weather Research and Forecasting (WRF) model.

References

Albers, S. C., 1995: The LAPS wind analysis. *Wea. Forecasting*, 10, 342-352.

________, J. A. McGinley, D. L. Birkenheuer and J. R. Smart, 1996: The Local Analysis and Prediction System (LAPS): Analysis of clouds, precipitation, and temperature. *Wea. Forecasting*, 11, 273-287.

Avissar, R., and R. A. Pielke, 1989: A parameterization of heterogeneous land surface for the atmospheric numerical models and its impact on regional meteorology. *Mon. Wea. Rev.*, 114, 330-343.

Barnes, S. L., 1973: Mesoscale objective map analysis using weighted time series observations. NOAA Tech. Memo. ERL NSSL-62, 60 pp. [NTIS COM-73-10781.]

Benjamin, S. G., K. A. Brewster, R. Brummer, B. F. Jewett, T. W. Schlatter, T. L. Smith and P. A. Stamus, 1991: An isentropic three-hourly data assimilation system using ACARS aircraft data. *Mon. Wea. Rev.*, 119, 888-906.

Bernardet, L. R., 2000: QPF verification at the model grid versus at the stations. Preprints, *15th Conf. on Hydrology,* Long Beach, CA, Amer. Meteor. Soc., 365-368.

Black, T. L., 1994: The new NMC mesoscale Eta Model: description and forecast experiments. *Wea. Forecasting*, 9, 265-278.

Briggs, W. M., and R. Zaretzki, 1998: The effect of randomly spaced observations on field forecast errors. Preprints, *14th Conf. on Probability and Statistics in Atmospheric Sciences*, Phoenix, AZ, Amer. Meteor. Soc., 5-8.

Colle, B. A., K. J. Westrick and C. F. Mass, 1999: Evaluation of MM5 and Eta-10 precipitation forecasts over the Pacific Northwest during the cool season. *Wea. Forecasting*, 14, 137-154.

Davies, H. C., 1983: Limitations of some common lateral boundary schemes used in regional NWP models. *Mon. Wea. Rev.*, 111, 1002-1012.

Du, J., S. L. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.*, 125, 2427-2459.

Gaudet, B., and W. R. Cotton, 1998: Statistical characteristics of a real-time precipitation forecasting model. *Wea. Forecasting*, 13, 966-982.

Katz, R. W., and A. H. Murphy, 1997: *Economic Value of Weather and Climate Forecasts*. Cambridge University Press, 222 pp.

Junker, N. W., J. E. Hoke and R. H. Grumm, 1989: Performance of NMC's regional models. *Wea. Forecasting*, 4, 368-390.

Louis, J. F., 1979: A parametric model of vertical eddy fluxes in the atmosphere. *Bound.-Layer Meteor.*, 17, 187-202.

Loveland, T. R., J. W. Merchant, D. O. Ohlen and J. F. Brown, 1991: Development of a land-cover characteristics database for the conterminous U.S. *Photo. Eng. Rem. Sens.*, 57, 1453-1463.

Mahrer, Y., and R. A. Pielke, 1977: A numerical study of the airflow over irregular terrain. *Beitr. Phys. Atmos.*, 50, 98-113.

Rothfusz, L. P., and M. R. McLaughlin, 1997: Weather support for the XXVI Olympiad. NOAA Tech. Memo. NWS SR-184, National Weather Service, Southern Region, Fort Worth, TX, 70 pp.

________, J. T. Johnson, L. C. Safford, M. R. McLaughlin and S. K. Rinard, 1996: The Olympic Weather Support System. Preprints, *2nd Intl. Conf. on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography and Hydrology*, Atlanta, GA, Amer. Meteor. Soc., 1-6.

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

Schwartz, B. E., and S. G. Benjamin, 1998: Verification of RUC-2 and Eta model precipitation forecasts. Preprints, *16th Conf. on Weather Analysis and Forecasting*, Phoenix, AZ, Amer. Meteor. Soc., J103-J105.

Snook, J. S., J. M. Cram and J. M. Schmidt, 1995: LAPS/RAMS: A nonhydrostatic mesoscale numerical modeling system configured for operational use. *Tellus*, 47A, 864-875.

________, P. A. Stamus, J. Edwards, Z. Christidis and J. A. McGinley, 1998: Local-domain mesoscale analysis and forecast model support for the 1996 Centennial Olympic Games. *Wea. Forecasting,* 13, 138-150.

Thompson, G., 1993: Prototype real-time mesoscale prediction during 1991-92 winter season and statistical verification of model data. Atmospheric Science Paper No.

521. [Available from Department of Atmospheric Science, Colorado State University, Fort Collins, CO 80523].

Tremback, C. J., and R. Kessler, 1985: A surface temperature and moisture parameterization for use in mesoscale numerical models. Preprints, *7th Conference on Numerical Weather Prediction*, Montreal, QC, Amer. Meteor. Soc., 355-358.

Walko, R. L., W. R. Cotton, J. L. Harrington, and M. P. Meyers, 1995: New RAMS cloud microphysics parameterization. Part I: The single-moment scheme. *Atmos. Res.*, 38, 29-62.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 464 pp.

Zhao, Q., T. L. Black, and M. E. Baldwin, 1997: Implementation of the cloud prediction scheme in the Eta model at NCEP. *Wea. Forecasting*, 12, 697-712.