Short-Range Ensemble Forecasting of Quantitative Precipitation

JUN DU AND STEVEN L. MULLEN

Institute of Atmospheric Physics, The University of Arizona, Tucson, Arizona

FREDERICK SANDERS

Marblehead, Massachusetts

(Manuscript received 9 April 1996, in final form 2 January 1997)

ABSTRACT

The impact of initial condition uncertainty (ICU) on quantitative precipitation forecasts (QPFs) is examined for a case of explosive cyclogenesis that occurred over the contiguous United States and produced widespread, substantial rainfall. The Pennsylvania State University–National Center for Atmospheric Research (NCAR) Mesoscale Model Version 4 (MM4), a limited-area model, is run at 80-km horizontal resolution and 15 layers to produce a 25-member, 36-h forecast ensemble. Lateral boundary conditions for MM4 are provided by ensemble forecasts from a global spectral model, the NCAR Community Climate Model Version 1 (CCM1). The initial perturbations of the ensemble members possess a magnitude and spatial decomposition that closely match estimates of global analysis error, but they are not dynamically conditioned. Results for the 80-km ensemble forecast are compared to forecasts from the then operational Nested Grid Model (NGM), a single 40-km/15-layer MM4 forecast, a single 80-km/29-layer MM4 forecast, and a second 25-member MM4 ensemble based on a different cumulus parameterization and slightly different unperturbed initial conditions.

Large sensitivity to ICU marks ensemble QPF. Extrema in 6-h accumulations at individual grid points vary by as much as 3.00″. Ensemble averaging reduces the root-mean-square error (rmse) for QPF. Nearly 90% of the improvement is obtainable using ensemble sizes as small as 8–10. Ensemble averaging can adversely affect the bias and equitable threat scores, however, because of its smoothing nature. Probabilistic forecasts for five mutually exclusive, completely exhaustive categories are found to be skillful relative to a climatological forecast. Ensemble sizes of approximately 10 can account for 90% of improvement in categorical forecasts relative to that for the average of individual forecasts. The improvements due to short-range ensemble forecasting (SREF) techniques exceed any due to doubling the resolution, and the error growth due to ICU greatly exceeds that due to different resolutions.

If the authors' results are representative, they indicate that SREF can now provide useful QPF guidance and increase the accuracy of QPF when used with current analysis–forecast systems.

1. Introduction

It is well known that the atmosphere is a chaotic system (e.g., Lorenz 1963). As a consequence, small errors in the initial condition of any numerical weather prediction (NWP) model amplify as the forecast evolves, with the root-mean-square error (rmse) ultimately becoming saturated [i.e., forecast error variance is approximately twice the climatological variance (e.g., Anthes 1986)]. Because the atmospheric state can never be measured exactly, analyses will always contain errors whose size and structure can only be estimated. Hence, an infinite spectrum of plausible initial conditions exists, all of which are consistent with analysis uncertainty. A single model run gives only one possible solution to the future atmospheric state.

One approach to ensemble forecasting (EF) involves running multiple forecasts starting at the same time but from different, equally probable initial analyses. Ensemble forecasting is a finite approximation to the method of stochastic-dynamic prediction first proposed by Epstein (1969), which unfortunately is impractical except for very low order models. The advantage of EF over single-run deterministic forecasting was shown by Leith (1974) and Hoffman and Kalnay (1983), while Murphy (1988) and Palmer et al. (1990) gave early examples of its potential operational benefits. The practical advantages of EF depends upon the extent of a model's deficiencies, with the increase in skill obtainable from ensembles increasing with model skill (Murphy 1988; Palmer et al. 1990). In general, the benefits of EF will be maximized for any situation where the model's systematic error is relatively small compared to its initial condition sensitivity.

Several operational forecast centers now employ forms of ensemble prediction. Most research to date has

Corresponding author address: Dr. Steven L. Mullen, Department of Atmospheric Sciences, PAS Building 81, The University of Arizona, Tucson, AZ 85721.

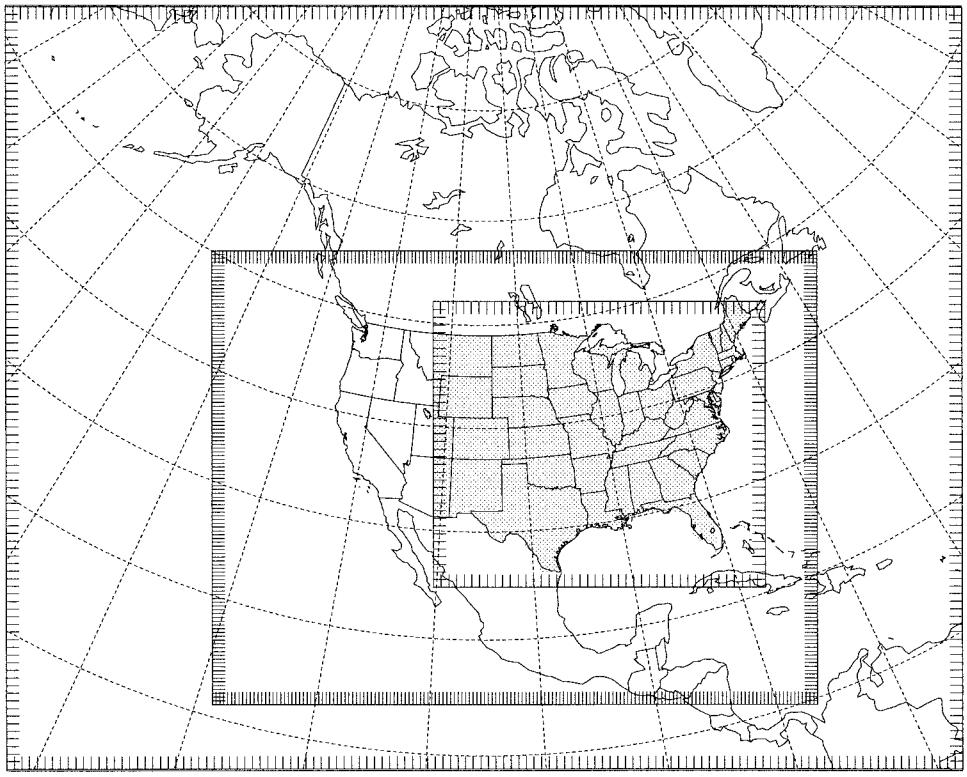

FIG. 1. The MM4 model forecasting domain (outermost panel, 126 × 101 grid points, 80-km mesh), the display domain (innermost panel, 44 × 38 grid points, 80-km mesh), and the domain for a higher-resolution run (middle panel with the finer grid marks, 161 × 121 grid points, 40-km mesh). The shading denotes the verification region for all QPF results.

focused on its application to extended range (6–10 days) forecasts (Tracton and Kalnay 1993; Toth and Kalnay 1993; Mureau et al. 1993, Molteni et al. 1996) and its impact on primary parameters (i.e., geopotential height, temperature, and horizontal wind). It has been recently proposed (Mullen and Baumhefner 1991, 1994; Brooks and Doswell 1993; Brooks et al. 1995) that ensemble methods could also benefit short-range (1–2 day) forecasts, perhaps even more than at extended ranges for certain weather elements such as precipitation.

Precipitation is an important surface weather element. Unfortunately, quantitative precipitation forecasts (QPFs) lose skill more rapidly with range than forecasts of any other surface element (Sanders 1986). Therefore, it is extremely desirable to learn how much short-range ensemble forecasting (SREF) can improve a QPF, given current model capability and computing power. This issue is addressed in this paper, which describes results for a 25-member ensemble of 36-h forecasts for a case of explosive cyclogenesis over land with widespread, significant precipitation.

We believe that rapid cyclogenesis is an appropriate phenomenon for a pilot study that explores the utility of SREF on QPF. The accuracy of operational NWP forecasts of explosive surface cyclogenesis (Sanders and Gyakum 1980), at least for primary variables such as sea level pressure, has increased dramatically over the past 15 years in terms of reduced systematic errors (Sanders 1992). This advance notwithstanding, the accuracy of model forecasts varies considerably from cyclone to cyclone and among successive forecasts for the same storm (Roebber 1990, 1993; Sanders 1992; Smith and Mullen 1993; Grumm 1993), and also among ensemble members of ''perfect model'' simulations for the same storm (Mullen and Baumhefner 1994). Moreover, it also appears that rapid cyclogenesis is more sensitive to initial condition uncertainty (ICU) than less volatile situations (Kallen and Huang 1988; Mullen and Baumhefner 1989, 1994; Kuo and Low-Nam 1990).

In this study, we use a limited-area model to forecast a case of explosive cyclogenesis over the contiguous United States where the surface rain gauge network allows for a reliable verification of QPF. To simulate the operational environment and allow for unbounded predictability error growth, lateral boundary conditions (LBCs) for the limited-area model are provided by *forecasts* from a global model. Because of their importance and difficulty, we focus on evaluating forecasts of precipitation accumulated during 6-h periods.

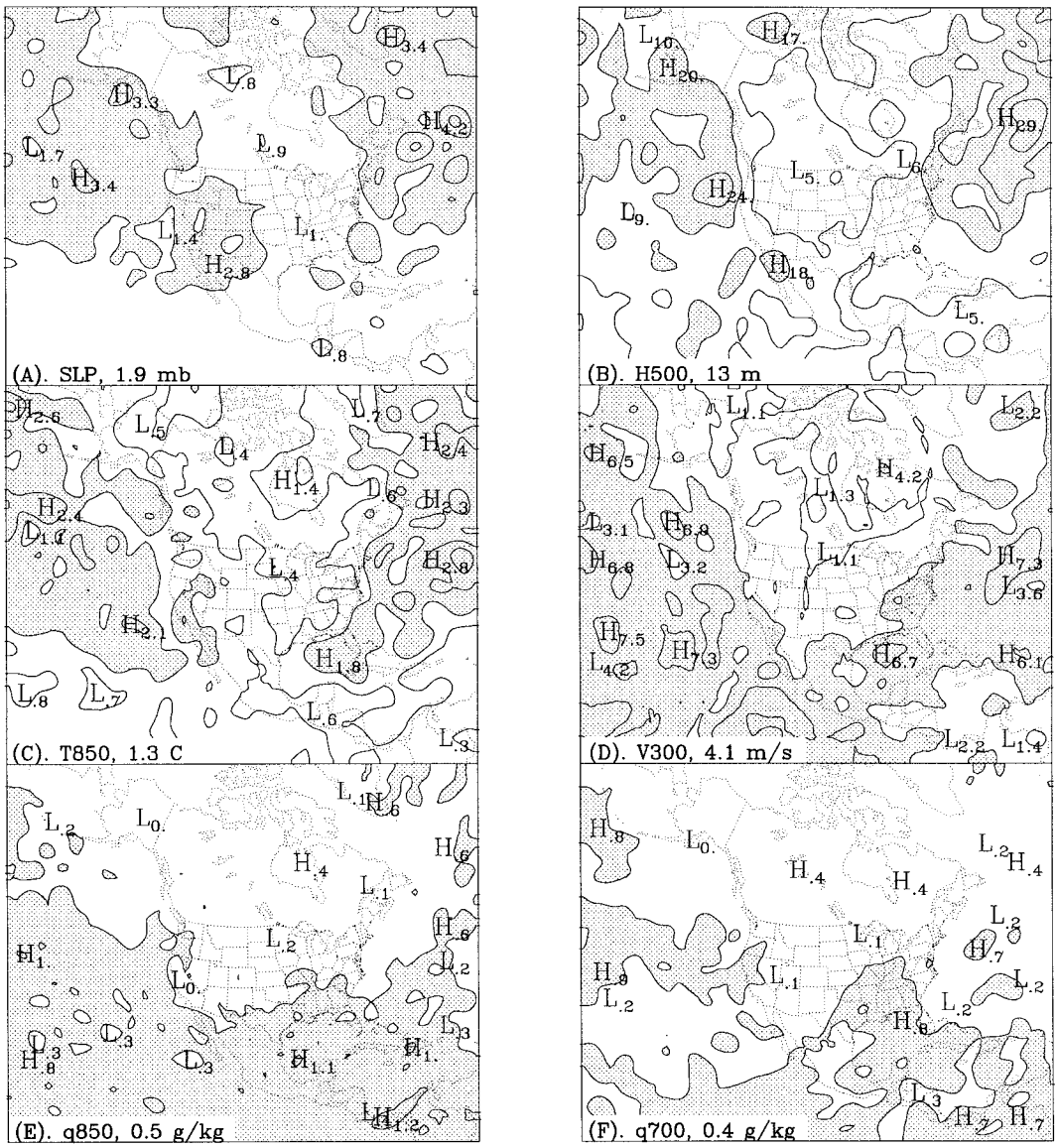

FIG. 2. Average rms values of the initial perturbation fields on the MM4 grid: (a) sea level pressure (SLP) at intervals of 1 mb (shading area ≥ 2 mb); (b) 500-mb height at intervals of 5 m (≥15 m); (c) 850-mb temperature at intervals of 0.5°C (≥1.5°C); (d) magnitude of 300-mb wind vector at intervals of 2.0 m s^{-1} (≥4 m s^{-1}); (e) 850-mb specific humidity at intervals of 0.5 g kg^{-1} (≥0.5 g kg^{-1}); and (f) 700-mb specific humidity at intervals of 0.5 g kg^{-1} (≥0.5 g kg^{-1}). The domain average of rms values is printed on the lower-left corner.

2. Methodology

a. Model descriptions

Two numerical forecast models are employed in this study: the Pennsylvania State University–National Center for Atmospheric Research (PSU–NCAR) Mesoscale Model version 4 (MM4) and the NCAR Community Climate Model version 1 (CCM1). The following is only a brief description of each model including the options that we selected. For details on the models, readers are referred to Anthes et al. (1987) and Zhang et al. (1988) for MM4 information and to Williamson et al. (1987) for CCM1 information.

Fifteen vertical sigma (σ) layers are used in the MM4 forecasts.[1] Within the planetary boundary layer, layers run about 10–30 mb thick, while in the upper and middle troposphere they are about 90 mb thick. A horizontal grid spacing of 80 km with 126 × 101 points (Arakawa and Lamb 1977) is used. The model is in-

[1] The 15 layers are σ = 0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.74, 0.81, 0.865, 0.91, 0.945, 0.97, 0.985, and 0.995, where $\sigma = (p - p_t)(p_s - p_t)^{-1}$, p is pressure at any level, p_s is surface pressure, and p_t is the constant pressure of the top of the model atmosphere. The model top is p_t = 100 mb for this case.

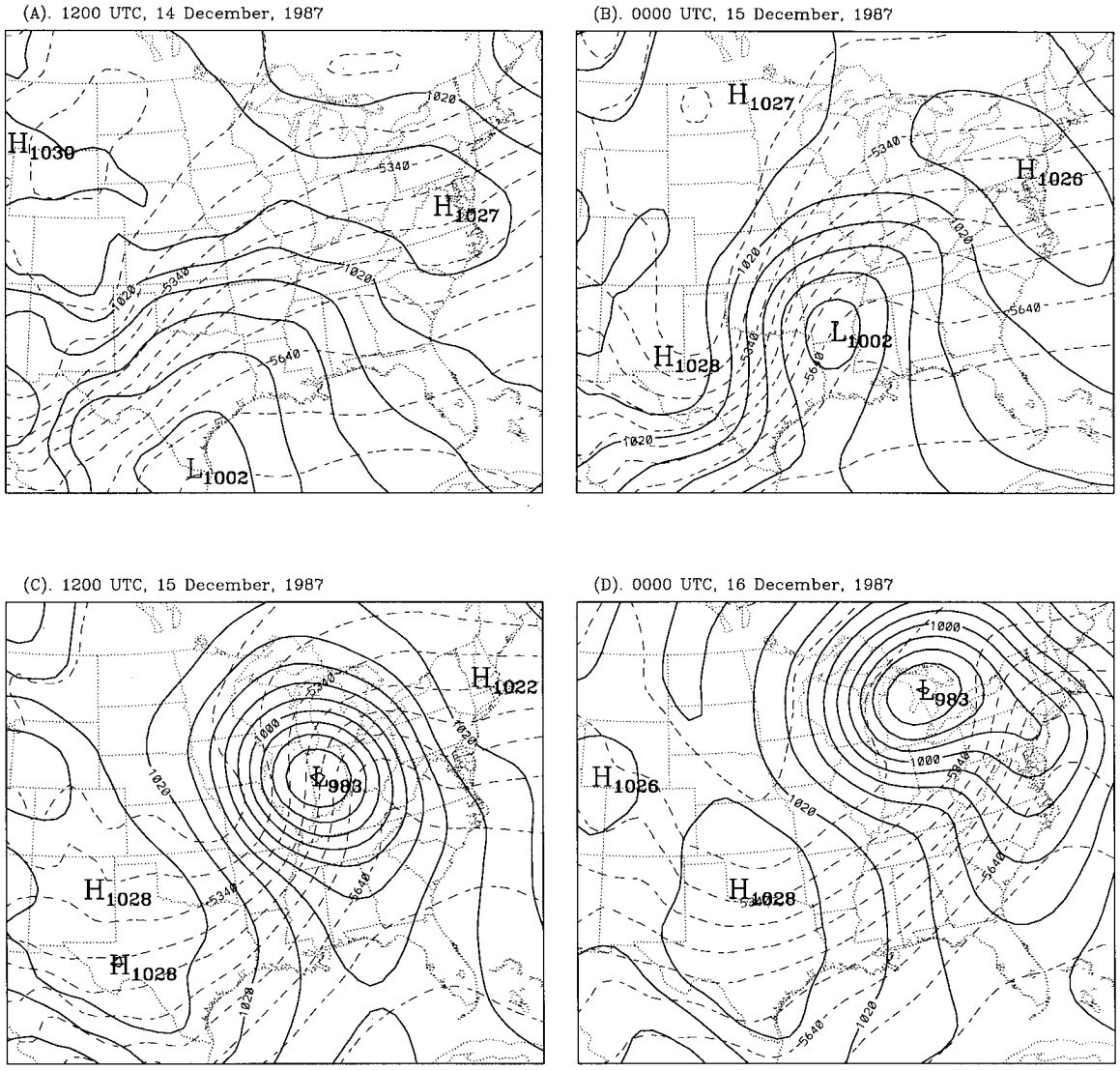

FIG. 3. Sea level isobars at intervals of 4 mb (solid) and isopleths of thickness of the layer 1000–500 mb at intervals of 60 m (dash) for (a) 1200 UTC 14 December 1987, (b) 0000 UTC 15 December 1987, (c) 1200 UTC 15 December 1987, and (d) 0000 UTC 16 December 1987.

itialized at 1200 UTC 14 December 1987 and is run for 36 h with a time step of 2 min. Model output is stored every 3 h.

The parameterizations of the surface and planetary boundary layers follow Blackadar (1979). Nonconvective precipitation is calculated from explicit prognostic equations for water vapor, cloud water, and rainwater (Hsie et al. 1984). Convective precipitation is parameterized with an Arakawa–Schubert (1974) scheme as modified by Grell et al. (1991) to include the effects of convective-scale downdrafts.

To simulate the operational environment, time-dependent LBCs for the MM4 (Anthes and Warner 1978) are provided by a parallel forecast from the NCAR CCM1, so-called one-way nesting. LBCs are updated every 3 h and are assumed to vary linearly with time between updates. To minimize the impact of error propagation across the lateral boundaries into the MM4 interior, the MM4 domain (126 $\times$ 101 points) covers the Northern Hemisphere from the central North Pacific to the western North Atlantic Ocean, while the much smaller verification domain (44 $\times$ 38 points) covers the eastern half of the contiguous United States (see Fig. 1). Given a typical error propagation speed of 20°–30° longitude per day across the inflow boundaries into the grid interior (Baumhefner and Perkey 1982), the impact of ''error sweeping'' on the verification domain (Fig. 1) should be minimal during a 36-h forecast.

The version of CCM1 used here has 12 vertical σ

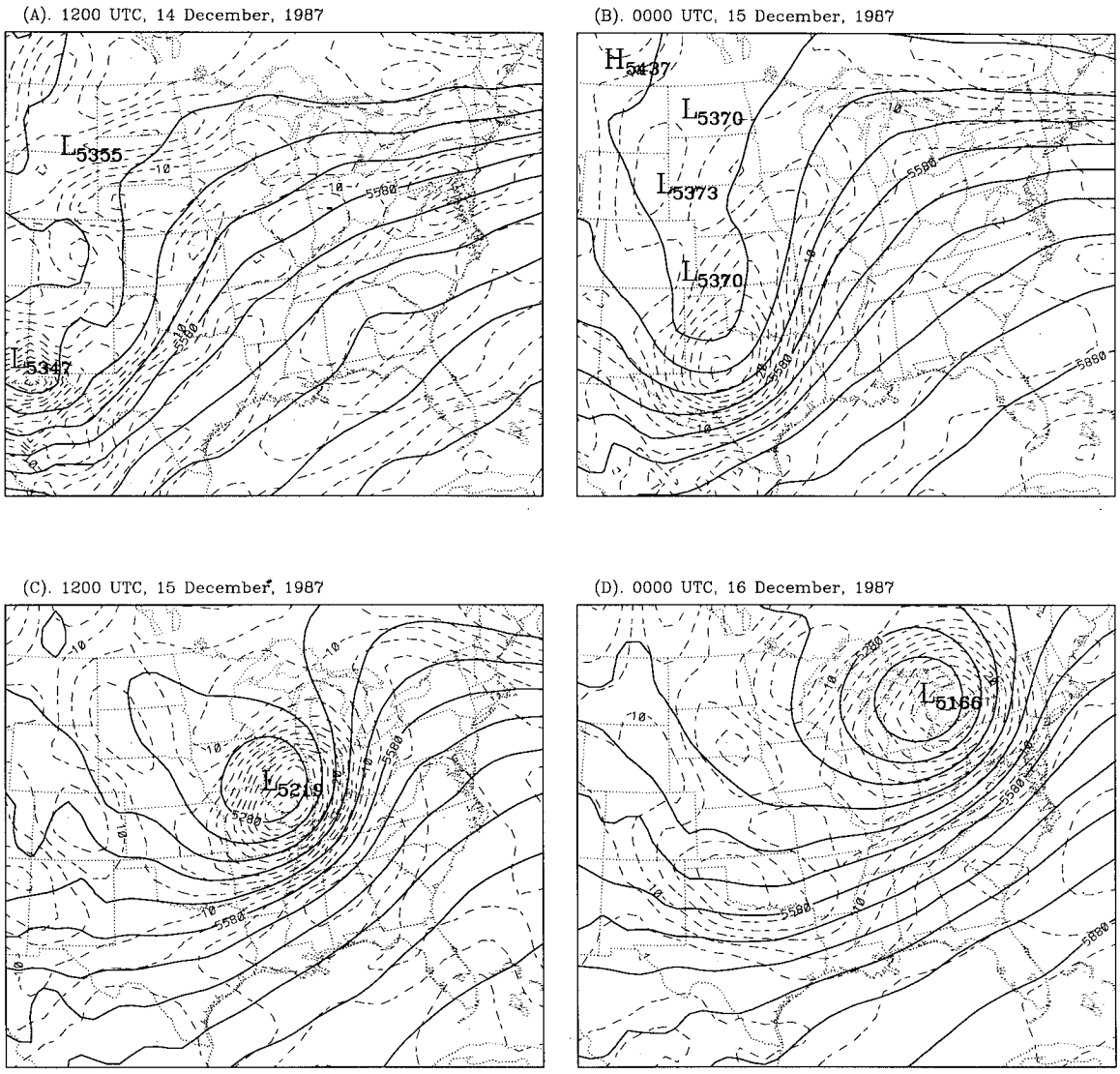

FIG. 4. The 500-mb geopotential height at intervals of 60 m (solid) and absolute vorticity at intervals of 2×10^{-5} s^{-1} (dashed) for (a) 1200 UTC 14 December 1987, (b) 0000 UTC 15 December 1987, (c) 1200 UTC 15 December 1987, and (d) 0000 UTC 16 December 1987.

layers.[2] The model is global and uses the spectral transform method to compute horizontal derivatives and perform linear operations. The spectral resolution is triangular 42 (T42), and the associated transform grid has 64 Gaussian latitudes between the poles and 128 grid points along each latitude. The equivalent gridpoint resolution is about 2.8° of latitude and longitude, and the smallest resolvable wavelength is about 800 km. The coarseness of the Gaussian grid in a spectral model, relative to a gridpoint model, is compensated in part by the elimination of aliasing in the computations. The CCM1 is run for 36 h with a time step of 15 min, and output is archived every 3 h.

CCM1 includes the following parameterized physical processes: convection; condensation; shortwave and longwave radiative transfers; surface fluxes of heat, moisture, and momentum; and interaction with subgrid-scale motions through diffusion. Clouds are formed in the model and can be convective or stratiform. They are radiatively active. In middle latitudes, clouds are allowed in all tropospheric layers except the lowest one. If the relative humidity exceeds 100%, clouds are formed and the excess water vapor is precipitated without evaporation of the condensate in the layers below.

The CCM1 includes a spectrally analyzed represen-

[2] The 12 layers are $\sigma = 0.017$, 0.0425, 0.085, 0.1375, 0.205, 0.300, 0.4275, 0.582, 0.7375, 0.8685, 0.9585, and 1.0. Note that $p_t = 0$ for the CCM1.

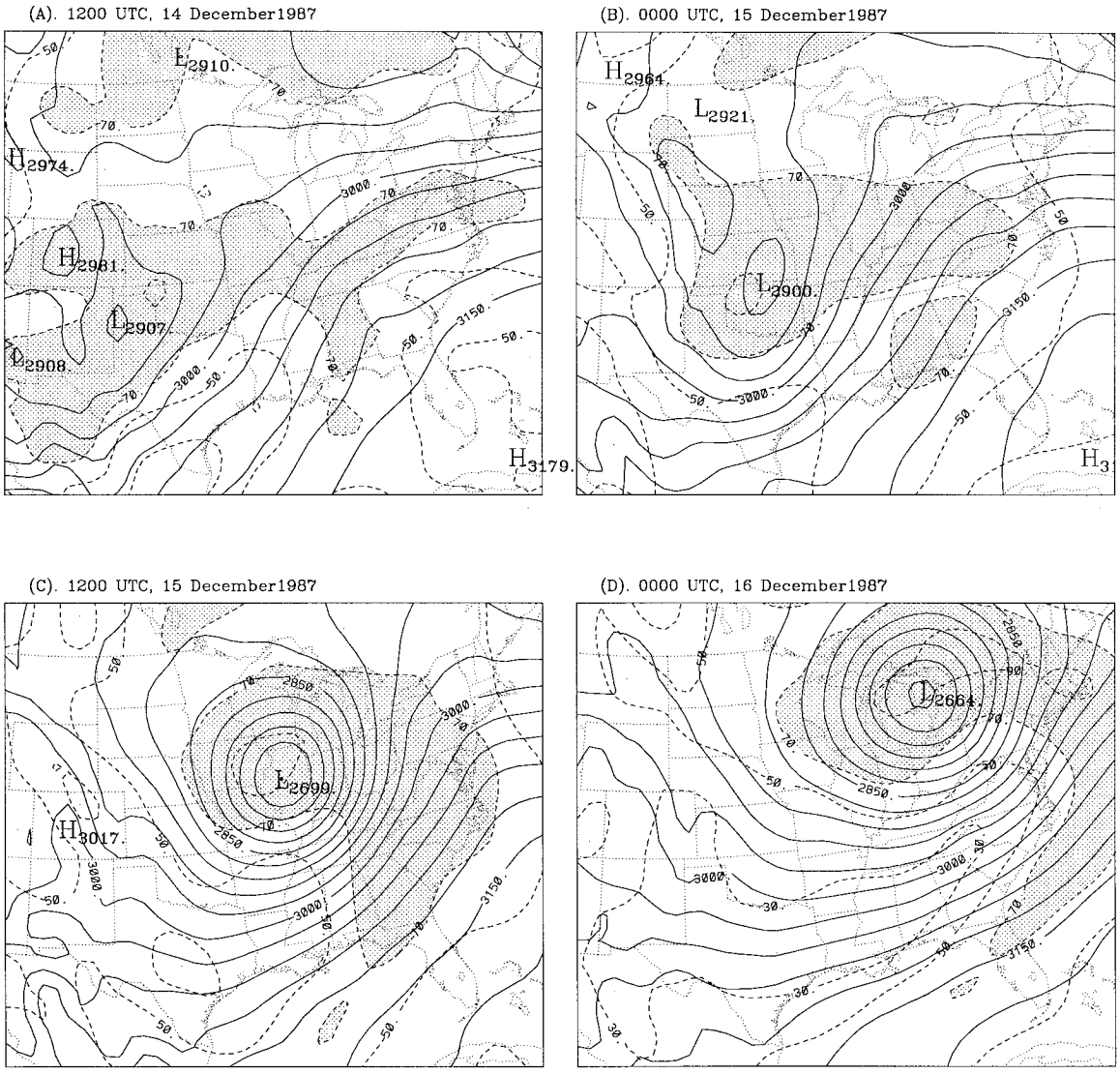

FIG. 5. The 700-mb geopotential height at intervals of 30 m (solid), and isopleths of mean relative humidity for surface to 500-mb layer at intervals of 10%, 30%, 50%, 70%, and 90% for (a) 1200 UTC 14 December 1987, (b) 0000 UTC 15 December 1987, (c) 1200 UTC 15 December 1987, and (d) 0000 UTC 16 December 1987. Shading denotes region of relative humidity greater than or equal to 70%.

tation of the earth's orography. Sea surface temperature, sea-ice distribution, and snow cover are externally prescribed and vary in accordance with the long-term seasonal averages. Incoming solar radiation varies daily according to a solar year of 365 days. Full radiation calculations are performed every 12 h, but CCM1 does not contain a diurnal variation in insolation.

b. Initial perturbation design

Rather than choosing perturbations that yield the most rapid error growth over a finite time interval for a prespecified metric such as singular vectors (e.g., Molteni et al. 1996), or another type of dynamically determined pattern such as a bred mode (Toth and Kalnay 1993), we prefer to choose perturbation fields that represent equally probable estimates of truth consistent with estimates of analysis uncertainty. For that reason, the method of perturbing initial fields is the same as that used by Mullen and Baumhefner (1989, 1994) with one important exception: the size of the perturbations reflects such important features of analysis uncertainty as land–sea differences and latitudinal variations.

Perturbation fields are first created for the CCM1 global model. Perturbations are independently applied to the initial fields of temperature, u and v wind components, and specific humidity using the method described by Errico and Baumhefner (1987). The amplitude and saturation point of the perturbation fields are adjusted to match corresponding spectra of Daley and

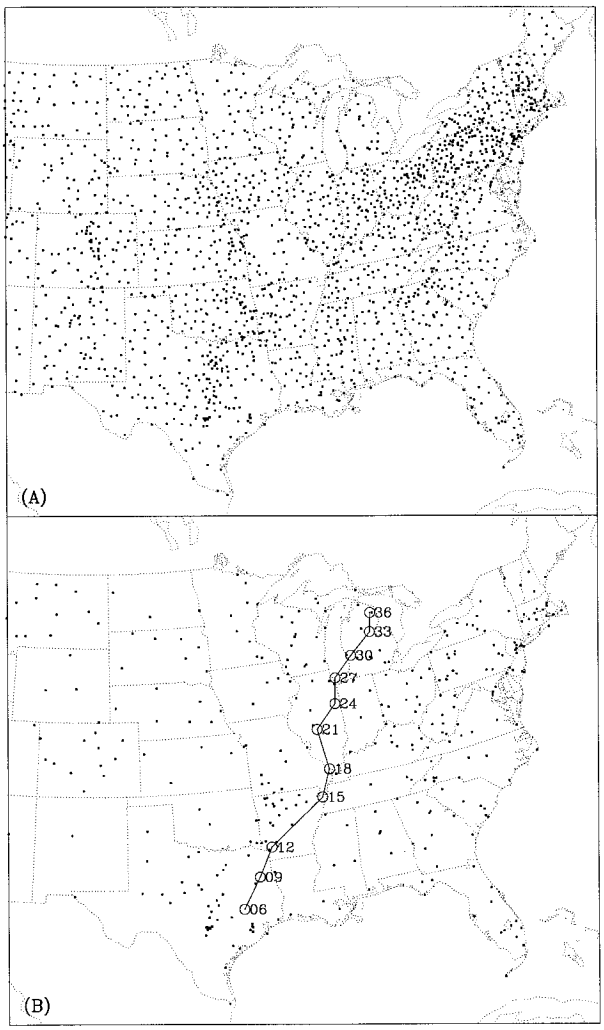

FIG. 6. Locations of the selected rain gauges and the low center: (a) 1839 usable rain gauges that record tenths of inches (0.1″); (b) 340 gauges that record hundredths of inches of rain (0.01″) and the position of the surface cyclone from 1800 UTC 14 December to 0000 UTC 16 December at intervals of 3 h.

Mayer (1986), who demonstrate that no reliable information is contained in global analyses at scales smaller than total wavenumber 30 (T30). Thus, for horizontal scales larger than T30, the perturbations are characterized by white noise (i.e., perturbation size independent of scale); the white noise portion of the spectrum is created by producing a field of uniformly distributed random values. For horizontal scales smaller than T30, the spectral components of the initial field are replaced by components of equal amplitude but random phase; this procedure produces small scales having identical variance spectra for the control and perturbed fields that are uncorrelated in space.

In an absolute sense, analysis uncertainty is larger over the oceans than over the continents, and it is larger over the midlatitudes than over the Tropics (Augustine et al. 1991). Analysis uncertainty for the wind field is largest at the tropopause level and above, while it is smallest for the temperature field in the midtroposphere (Augustine et al. 1991). To include such spatial variations in the CCM1 perturbations, a simple weighting mask is applied to scales larger than T30 that alters their amplitude locally in a manner that is roughly consistent with the results of Augustine et al. (1991). Thus, the final perturbed fields are not in a rigorous sense a white noise spectrum for scales larger than T30 since they contain a slight bias in the planetary scales that properly reflects estimates of analysis uncertainty. Perturbations for the MM4 forecasts are constructed by bilinearly interpolating the CCM1 ones from the T42 transform grid to the MM4's 80-km grid. This procedure produces perturbed MM4 fields consistent with the CCM1 perturbations but lacking amplitude in the smallest MM4 scales below the resolution of the T42 CCM1. In such a manner, a total of 24 different perturbation fields with equal globally averaged rms values are created for both models.

Because the perturbations are unbalanced, nonlinear normal model initialization (Errico 1983) is used for both models to remove most of the energy associated with inertial-gravitational modes. The initialization procedure reduces the rms size of the mass field by approximately a factor of 2, but the wind and moisture fields remain essentially unaltered. After initialization, the perturbations over the midlatitude oceans have a typical rms size of 25 m for 500-mb geopotential height field, 1.5°C for the tropospheric temperature field, 5 m s^{-1} for the upper-tropospheric wind field, and 1 g kg^{-1} for the lower-tropospheric specific humidity. Over land, the perturbations run about one-third to one-half as large. Figure 2 shows distributions of the rms values of selected perturbation fields on the MM4 grid for the 25 ensemble members. These initialized values are in close agreement with prior estimates of analysis uncertainty (Baumhefner 1984; Daley and Mayer 1986; Augustine et al. 1991).

With the National Meteorological Center [NMC, now known as the National Centers for Environmental Prediction (NCEP)] global analyses used as basic input fields, a forecast ensemble is thus created with 25 equally likely initial conditions (24 perturbed and one unperturbed) for the coupled CCM1–MM4 forecast system.

3. The storm of 14–16 December 1987

Mass and Schultz (1993) presented a detailed description of this cyclone and its simulation by the MM4 model (run with a 40-km mesh length), while Schneider (1990) earlier presented a detailed analysis of some aspects of the surface pattern. Powers and Reed (1993) diagnosed surface gravity wave activity associated with the storm using simulations with MM4 (run with a 20-km mesh length).

The storm formed on the Texas Gulf Coast at 1200

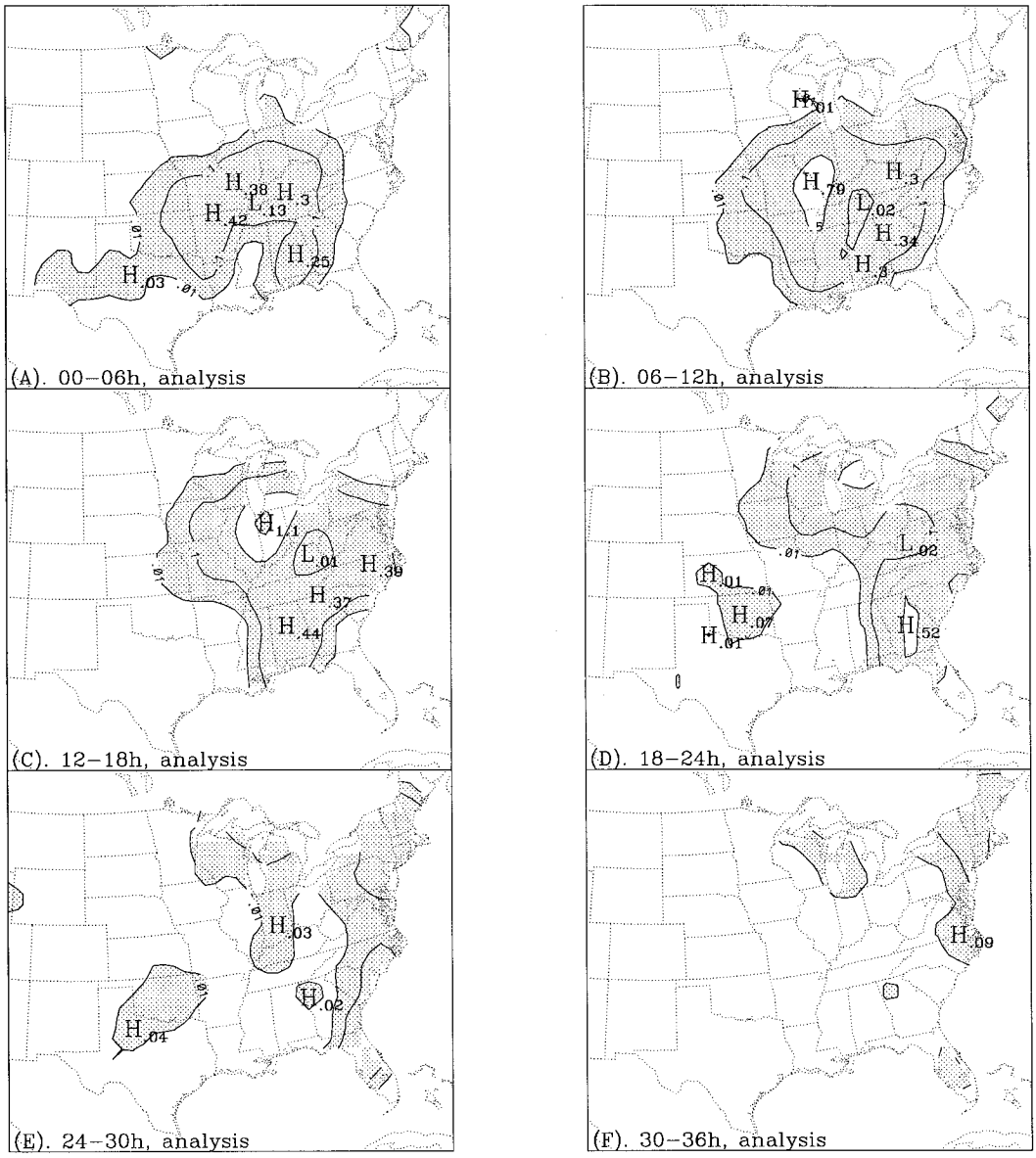

FIG. 7. Isohyets of objectively analyzed rainfall for the 6-h periods ending at (a) 1800 UTC 14 December 1987, (b) 0000 UTC 15 December 1987, (c) 0600 UTC 15 December 1987, (d) 1200 UTC 15 December 1987, (e) 1800 UTC 15 December 1987, and (F) 0000 UTC 16 December 1987. Isohyets are 0.01″, 0.1″, 0.5″, 1.0″, and 1.5″, etc. The shaded areas are 0.01″–0.5″, 1.0″–1.5″, and 2.0″–2.5″, etc. See text for analysis method.

UTC 14 December 1987 (Fig. 3a). It deepened rapidly while moving northeastward to the lower Great Lakes region by 0000 UTC on the 16th (Figs. 3b–d). The storm track at sea level is depicted in Fig. 6b. The low center qualified as an explosive deepener with a 19-mb drop between 1200 UTC of the 14th and 15th, yielding a rate of 1.2 bergerons (Sanders and Gyakum 1980). Virtually all of this deepening occurred during the second 12 h of this interval (e.g., Fig. 1 of Mullen and Du 1994). During this period of intensification, a powerful 500-mb vorticity maximum (Fig. 4) moved from a position about 1000 km west-southwest of the surface low into virtual coincidence with it.

Comparison of the isobars and the thickness lines shows a region of prominent warm advection extending from the eastern Gulf Coast in an arc around the nascent low center and about 500 km distant from it to the northwest of it. It moved northeastward with the low and lost its westernmost extension. In fact, by the end of the period (Fig. 3d), the cyclone had moved to the cold edge of the major thickness gradient and was almost a barotropic vortex. At this time, a new low center

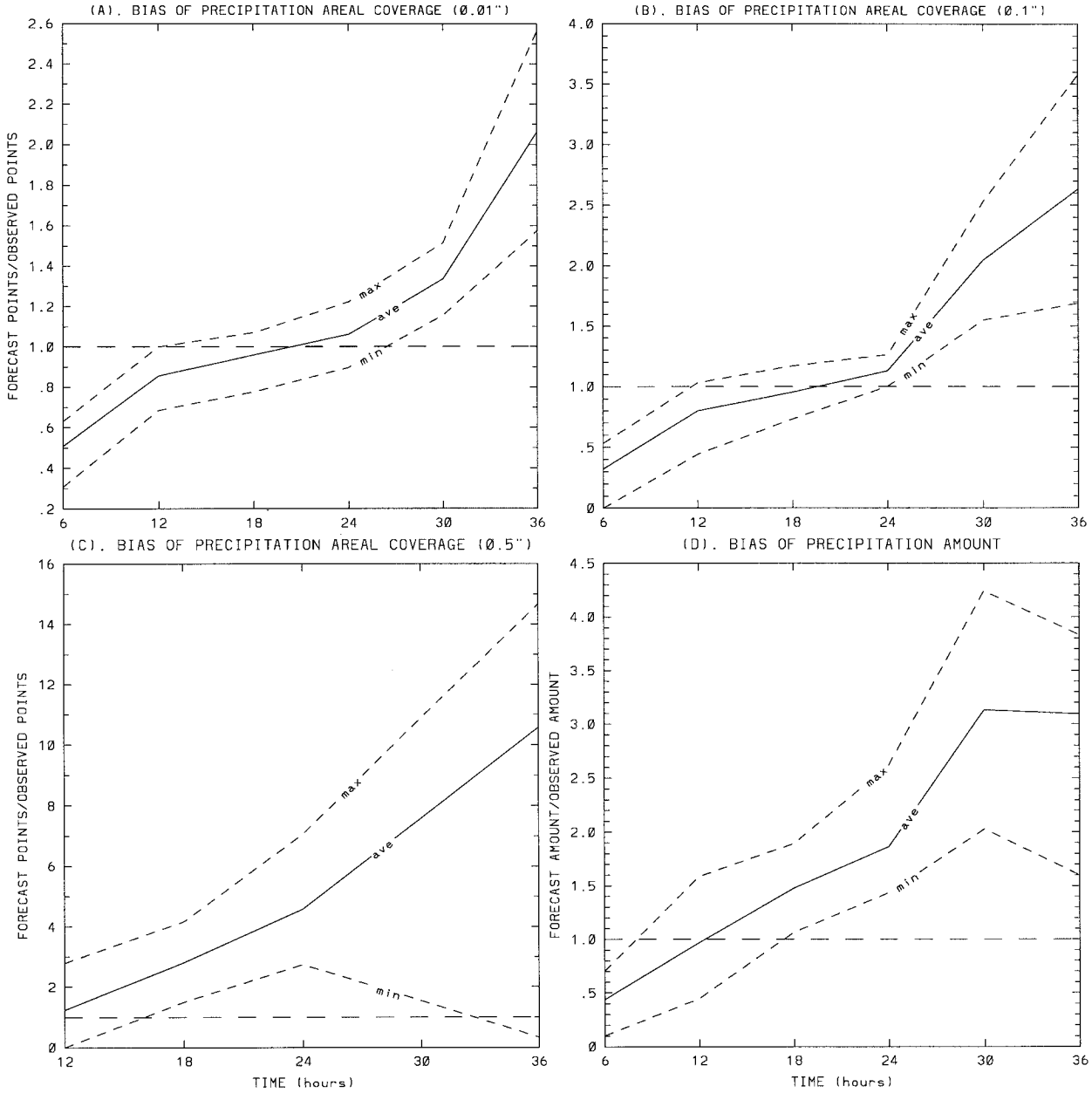

FIG. 8. Bias scores of 6-h precipitation forecast for 0.01″ (a), 0.1″ (b), and 0.5″ (c) thresholds. (d) The bias in precipitation amounts. Dash curves are for the driest (min) or wettest (max) individual cases; solid curves are for the ensemble average of 25 forecasts (ave).

was forming over New Jersey, in the region of major thickness gradient. It subsequently became the deeper system.

The storm was associated with a large precipitation shield at the start, more than 2000 km zonally and 1000 km meridionally at its greatest extent (Fig. 7a). This area had grown substantially during the 24 h preceding the onset of cyclogenesis (not shown). During the ensuing 24 h, the zonal extent shrank to less than 2000 km while the meridional extent remained more or less constant. (Fig. 3d). The area of observed precipitation corresponded reasonably closely to the area of greater than 70% mean relative humidity in the layer from the surface to 500 mb (Figs. 5a–d). This area of nominal deep saturation shrank during the 36-h period of cyclogenesis. It could be argued that the reduction of area of precipitation is an apparent one, due to the paucity of observations over the western Atlantic, but it is more likely attributable to the shortening of the distance between the surface ridge line and the 500-mb vorticity maximum, the likely eastern and western bounds of the region of quasigeostrophic forcing of ascent.

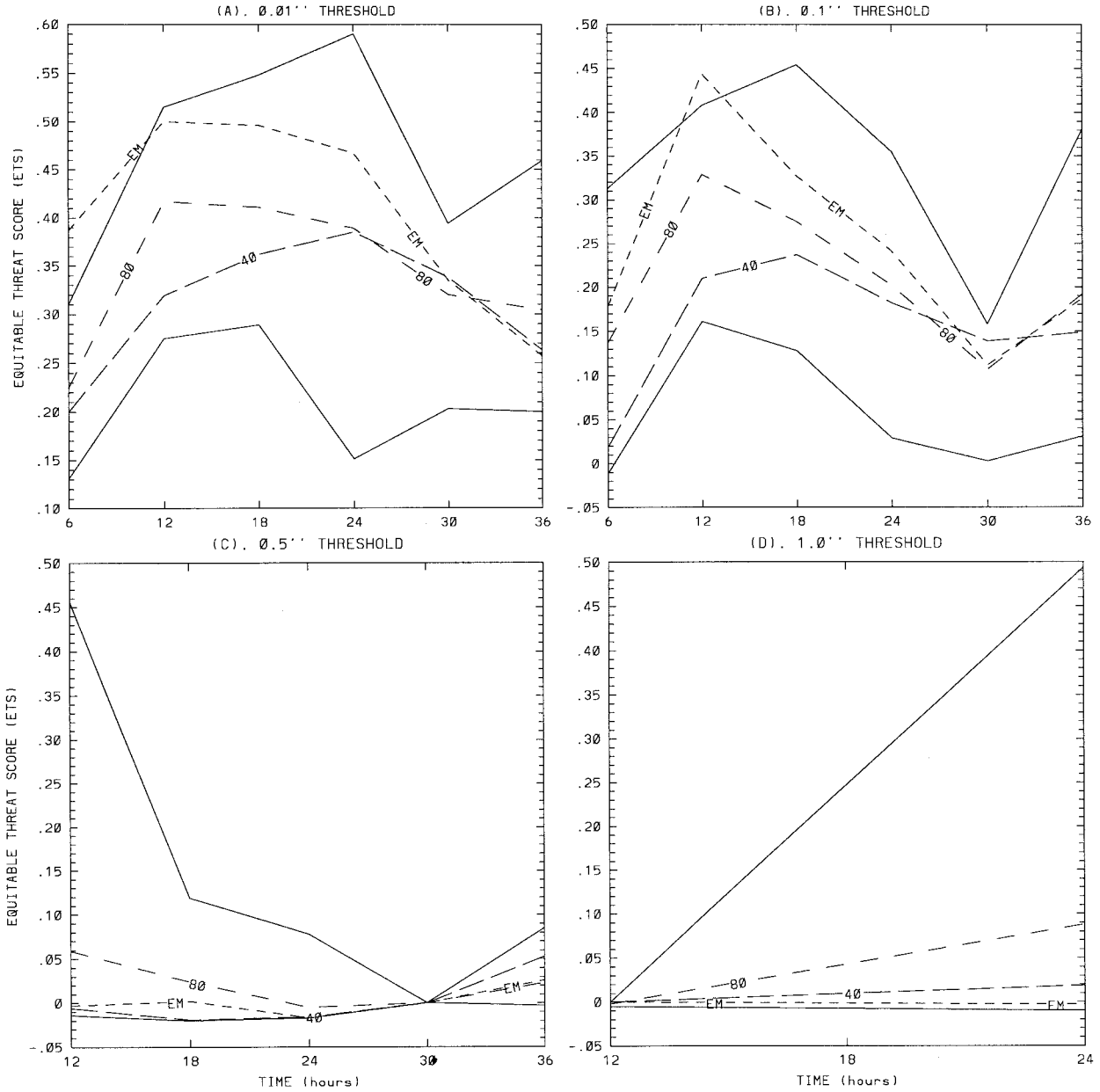

FIG. 9. ETSs for 0.01″, 0.1″, 0.5″, and 1.0″ thresholds, (a)–(d). Solid curves are for best and worst cases, curves ''EM'' the ensemble mean forecast, curves ''80'' the average of 25 forecasts, and curves ''40'' the higher-resolution (40 km) run. Note: the ETS for 0.01″ is based on the gridded data interpolated from the 340 rain gauges in Fig. 6b, while the ETSs for 0.1″, 0.5″, and 1.0″ are based on the gridded data interpolated from the 1839 rain gauges in Fig. 6a. All results are based on 6-h accumulations except (d), which is based on 12-h accumulations.

Note also in Fig. 5 the substantial growth of the region of very dry air in the wake of the cyclone and to the right of its track as it intensified. This appears to be due not only to advection but to descent predicted by quasigeostrophic theory as shown, for example, by Sanders (1971). The rapid eastward advance of this dry air with a history of descent was primarily responsible for the weakening of the band of precipitation extending southeastward from the cyclone center (Figs. 7d–f).

Vigorous convection was embedded in the precipitation area (not shown). Thunder was heard at 1200 UTC on the 14th from western Arkansas to northeast Texas. Twelve hours later, thunderstorms occurred from northern Louisiana to eastern Kentucky. Thereafter there was no large area of thunder, but sporadic thunderstorms or showers with moderate or heavy rain were observed in the southward-extending band of precipitation prior to its weakening after 1800 UTC of the 15th (Figs. 7d–f). Lightning damage was reported as far north as Wisconsin (NOAA/NCDC 1987).

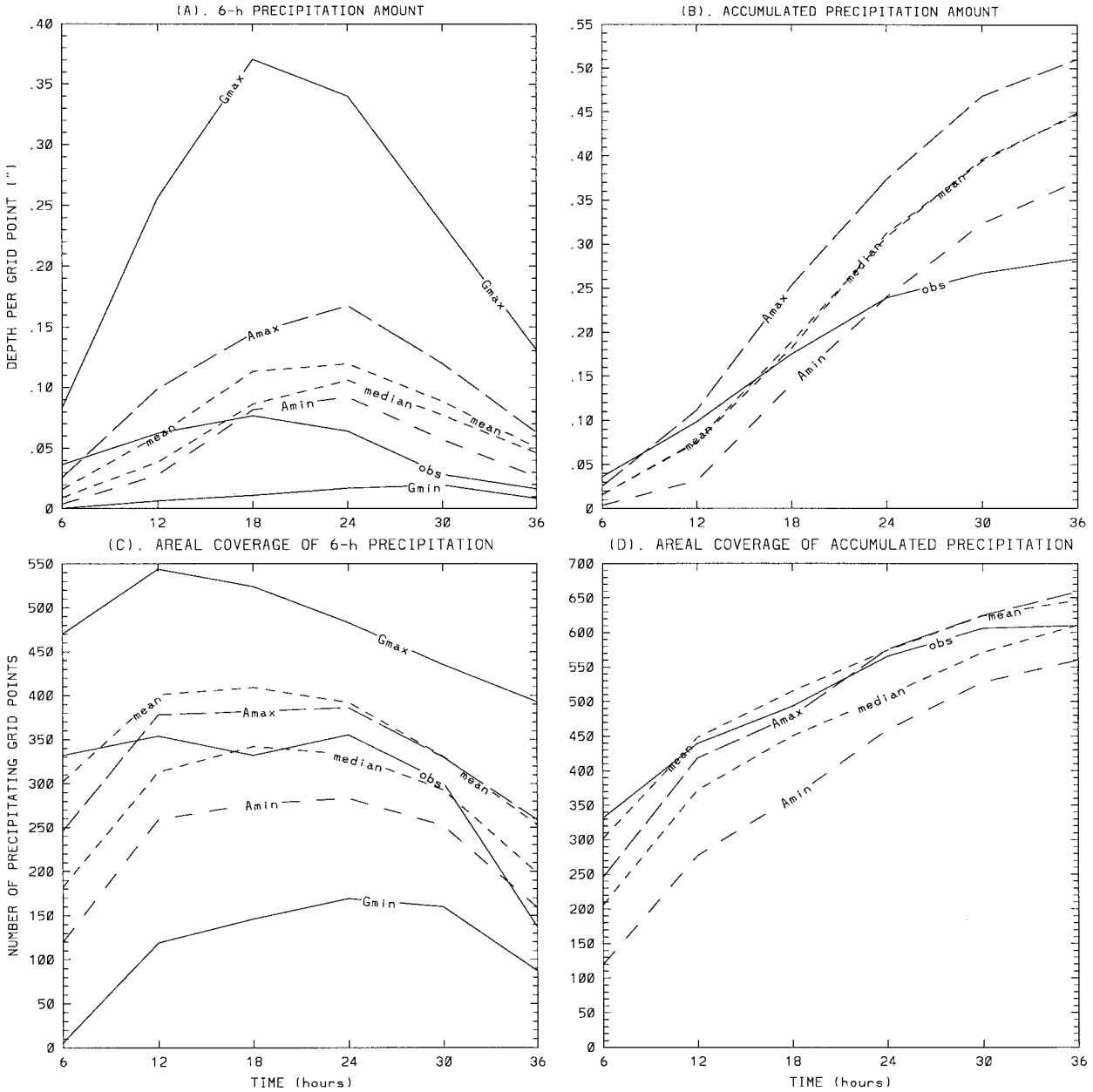

FIG. 10. (a) The 6-h accumulated precipitation per grid point; (b) storm accumulated precipitation per grid point; (c) number of grid points with measurable precipitation (≥0.01″) every 6-h period; and (d) number of grid points with measurable precipitation (≥0.01″) accumulated during the forecast period ending at the indicated time. Curves ''Amax'' and ''Amin'' are based on the individual cases that deposited the maximum and minimum precipitation over the verification area and correspond to the distributions of Fig. 11. Curves ''Gmax'' and ''Gmin'' are based on the extreme values of precipitation at the individual grid points and correspond to the distributions of Fig. 12. Curve ''mean'' denotes the 25-member ensemble mean forecast, ''median'' the median of the distribution, and ''obs'' the observations.

Radar summaries (not shown) provided evidence of convection embedded within much of the major rain area, where there was presumably large-scale saturation, minimizing the effect on area-average precipitation. In addition, a line of severe thunderstorms developed over easternmost Texas shortly before 2100 UTC of the 14th, prompting the issuance of a number of severe thunderstorm watches as the system moved across the Gulf Coast states, reaching central Florida and the adjacent western Atlantic by 0000 UTC of the 16th. This convective system had little impact on the rainfall pattern but produced a number of tornadoes (including an F3 one that killed 6 and injured 200 in West Memphis, Arkansas), besides widespread damaging winds and large hail (NOAA/NCDC 1987).

Verification of the forecasts of 6-h accumulations of

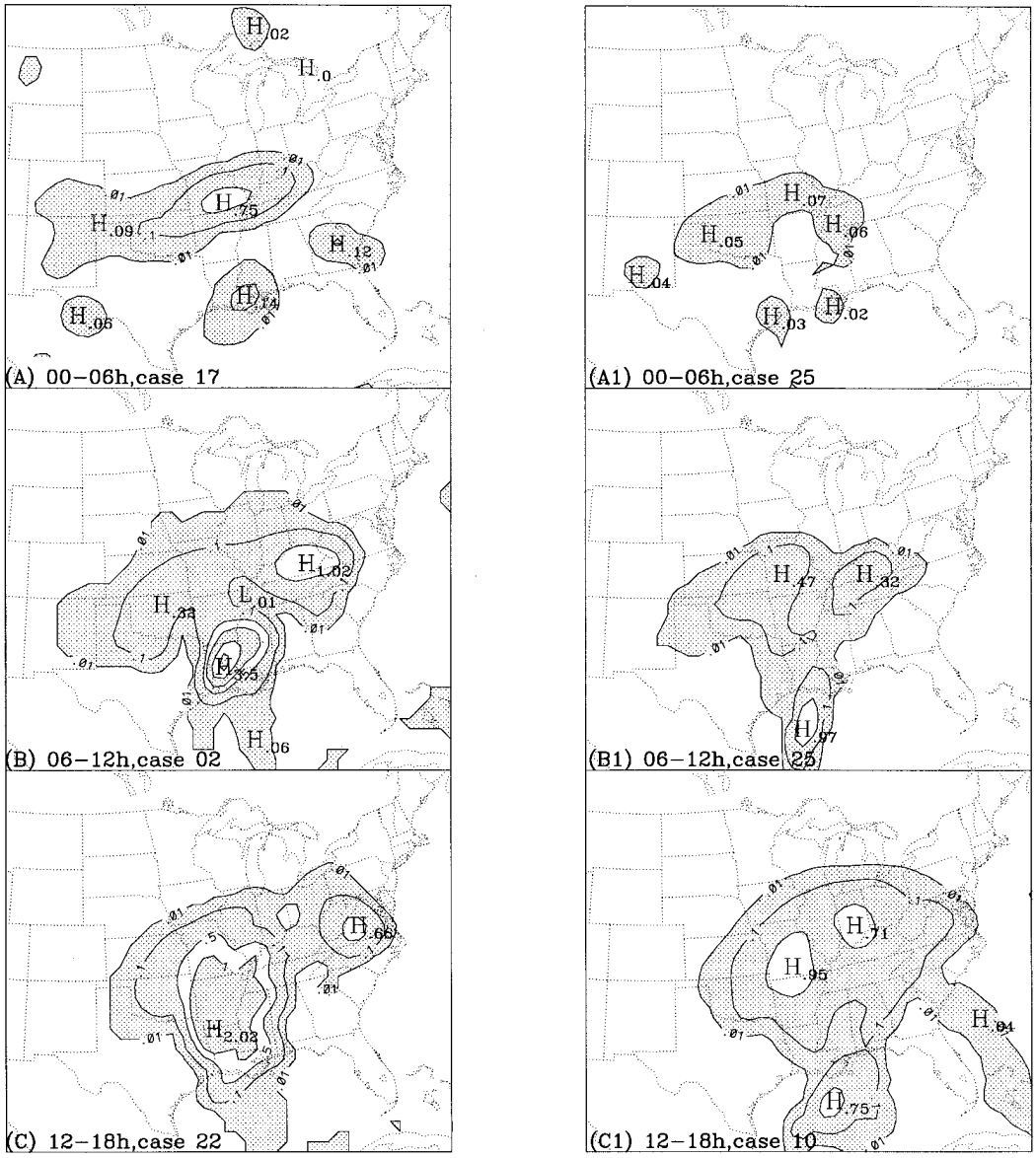

FIG. 11. Extreme forecasts of 6-h precipitation for a 6–36-h period, based on the individual cases that deposited the maximum and minimum precipitation over the verification area. These distributions correspond to curves ''Amax'' and ''Amin'' of Fig. 10. Isohyets are the same as Fig. 7.

precipitation was carried out not by comparison with the routine surface observations but rather by analysis of all the hourly rain gauge data from the United States east of 111°W (Fig. 6), as collected on a CD-ROM by EarthInfo, Inc. (1990) and provided by NOAA National Climate Data Center (NCDC). There were more than 2278 such gauges, but a close inspection of them showed that many were not appropriate for use. For 439 of them, either the location was not in operation at the time of this case, or the data were otherwise missing, delayed, or accumulated in such a manner as to negate their usefulness for our purposes. Of the 1839 remaining gauges (Fig. 6a), only 340 observed hundredths of inches of rain (0.01″)[3], and these were far from uniformly distributed over the states influenced by the storm (Fig. 6b). The rest of the gauges registered only tenths of inches (0.1″) and thus could not be used to verify the position of 0.01″ isohyet. Other thresholds of interest were for 0.1″ (the smallest amount for which the larger

[3] The conversion 1.00″ = 25.4 mm and the 0.01″ amount is commonly used as the threshold of measurable precipitation.

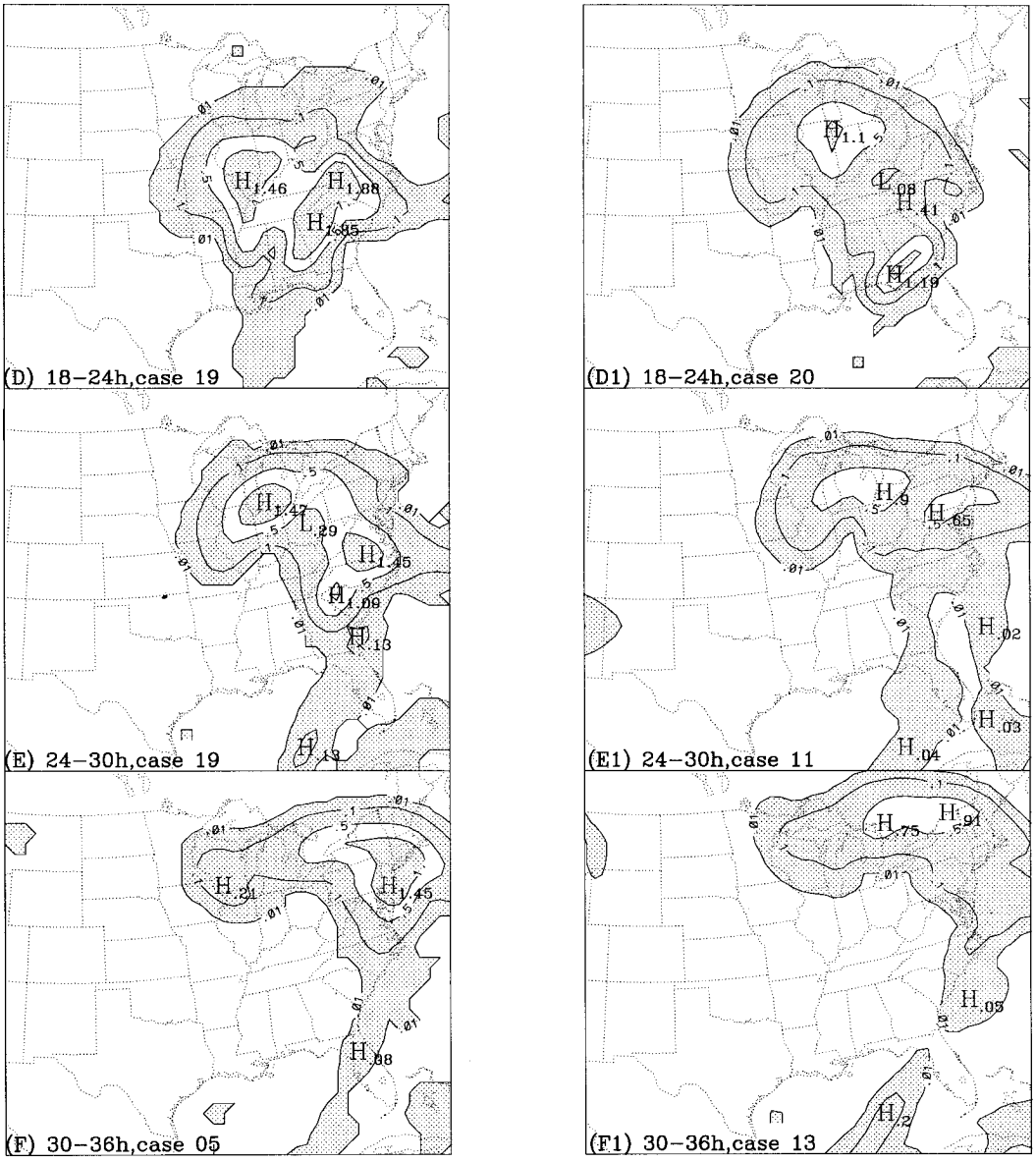

FIG. 11. (*Continued*)

number of gauges could be used), 0.5″, 1.0″, and 2.0″ (standard values for which the NCEP makes forecasts).

The analyzed positions of these isohyets for each 6-h period starting at 1200 UTC 14 December are shown in Figs. 7a–f. The positions were determined by interpolating objectively from the raw rain gauge data to the model grid points and then analyzing the resulting fields. The method is based on the Barnes objective analysis scheme (Barnes 1964, 1973), with the smoothing parameter selected a priori to give a half-power response at a wavelength of 320 km or a four-gridpoint wavelength for the 80-km model grid. This procedure is in fact a smoothing of the raw observations that removes small-scale irregularities from gauge to gauge. This small-scale variability is utterly beyond the ability of the model to predict, and it should be borne in mind that verification based on observations interpolated to the model grid will be, in general, better than verification in which the forecast values at the grid points are interpolated to the locations of the rain gauges. This reasoning has been proven true by comparing results based on the model output bilinearly interpolated to the station locations with the results discussed in this paper (results of the comparison are not shown). While interpolating the model values to the gauge locations did fare worse than interpolating the gauge values to the model grid points, the comparative benefits of SREF techniques relative to a deterministic forecast remain unaltered.

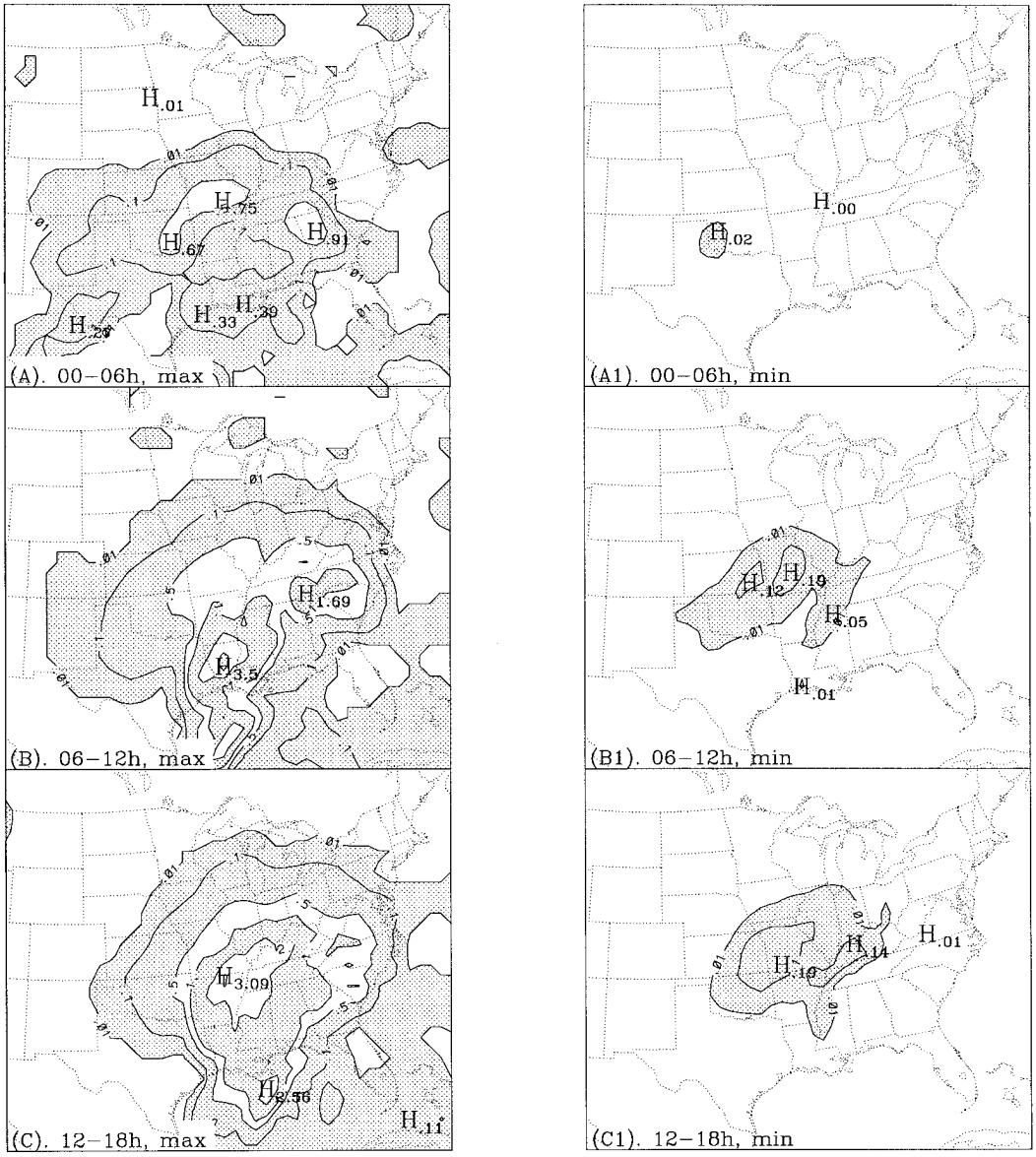

FIG. 12. Extreme values at each grid point of 6-h precipitation amount for the 6–36-h period. Left panel: maxima, where the region with amounts less than 0.01″ also defines 0% probability of measurable, etc.; right panel: minima, where the region with amounts greater than or equal to 0.01″ also defines 100% probability of measurable, etc. These distributions correspond to curves ''Gmax'' and ''Gmin'' of Fig. 10. Isohyets are the same as Fig. 7.

There is an uncertainty associated with many of the gauges reporting with resolution of 0.1″, which may be of the Fischer–Porter type. It is not clear whether some of the rain collected is discarded before being recorded. Such a practice, in the case of a major rainstorm, would lead to a modest underestimate of the rainfall. The isohyets in Fig. 7 represent our best estimate of the precipitation in the storm of 14–15 December based on the 1839 rain gauges. During the first 6 h (Fig. 7a), much of the area of precipitation shows amounts in excess of 0.01″. Although a 0.50″ isohyet is not analyzed in Fig. 7a, a region with numerous reports of amounts greater than 0.50″ is centered in western Arkansas and extends into Missouri (Fig. 20a), evidently reflecting the thunderstorm activity noted earlier. Other small areas of this amount lie to the east-northeast and to the southeast, while single gauges in Arkansas and Oklahoma report more than 1.0″ (Fig. 21a).

Six hours later (Fig. 7b), the major area of heavy precipitation has grown substantially, extending from northeast Louisiana to central Illinois and including a number of gauges registering more than 1.0″ (Fig. 21b). Again, sporadic reports of amounts greater than 0.5″ are found to the east and southeast (Fig. 20b).

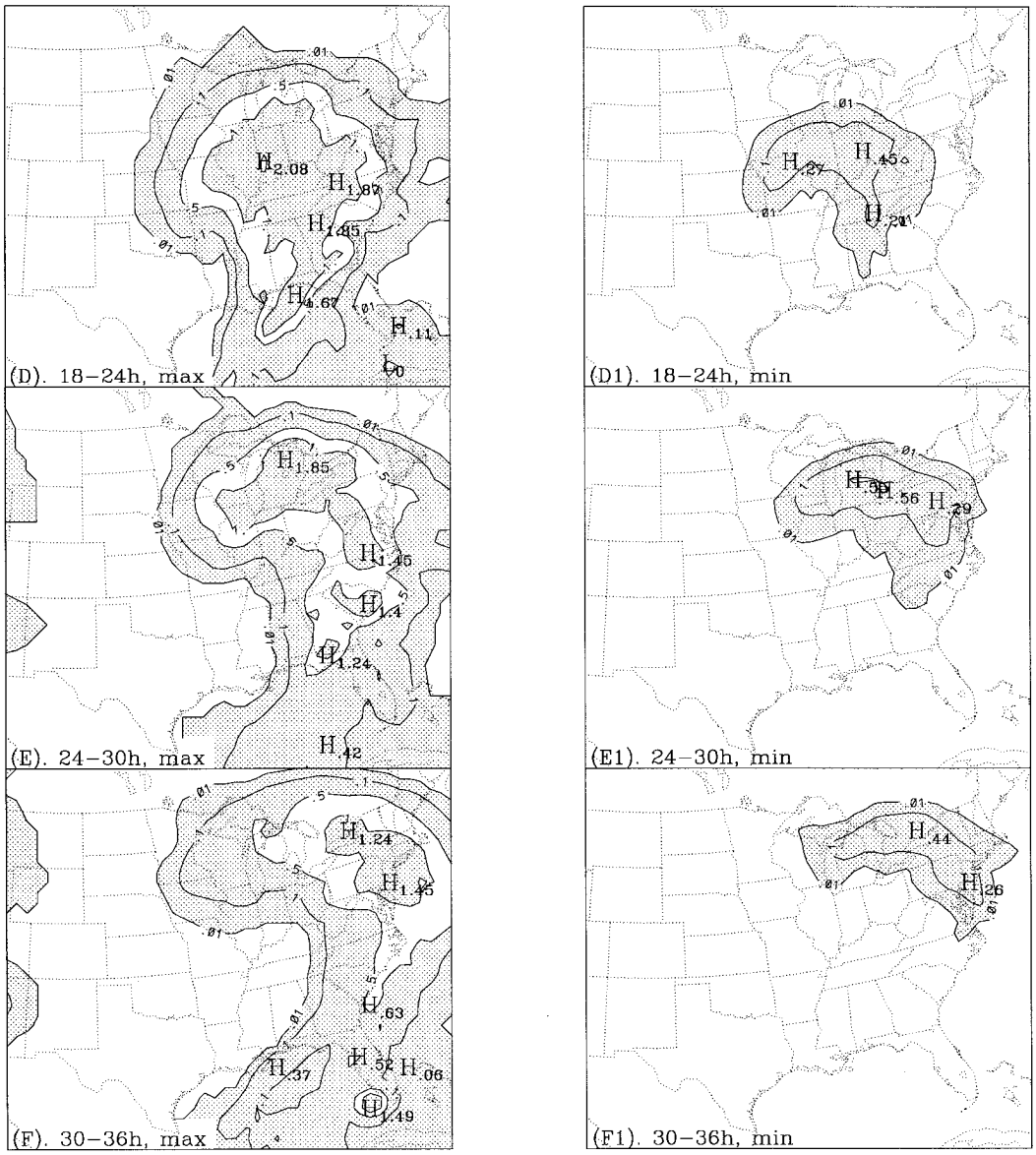

FIG. 12. (*Continued*)

During the period 0000–0600 UTC on the 15th (Fig. 7c), the major region of heavy rain, now in northern Indiana, grows no more in extent but gains appreciably in intensity with 33 contiguous gauges in this area recording more than 1.0″ (Fig. 21c). The other smaller areas gained slightly in intensity and coverage. The deepening low center at the middle of this period was along the Arkansas–Tennessee border, some 600 km south-southwest of the rainstorm (Fig. 6b).

Immediately after this time, as seen in Fig. 7d, the major rainstorm weakens markedly as it moves northeastward into lower Michigan, while the sporadic areas in the southeast grow and become well organized. Further lessening of rainfall in both regions occurred between 1200 and 1800 UTC of the 15th (Fig. 7e). Amounts of 0.5″ or more are present only in the northeast (Fig. 20e), evidently in response to the beginning of secondary cyclogenesis.

During the final 6 h, amounts of more than 0.5″ are analyzed (Fig. 7f) and recorded only in New England (Fig. 20f) as a secondary cyclone is about to appear by 0000 UTC of the 16th. In summary, the major aspects of the precipitation to be considered in evaluating the forecasts include:

1) The development, during the first 18 h, of a 1″ per 6 h rainstorm associated with the cyclone, moving northeastward in advance of the low center;
2) Its sudden weakening and disappearance during the next 12 h:

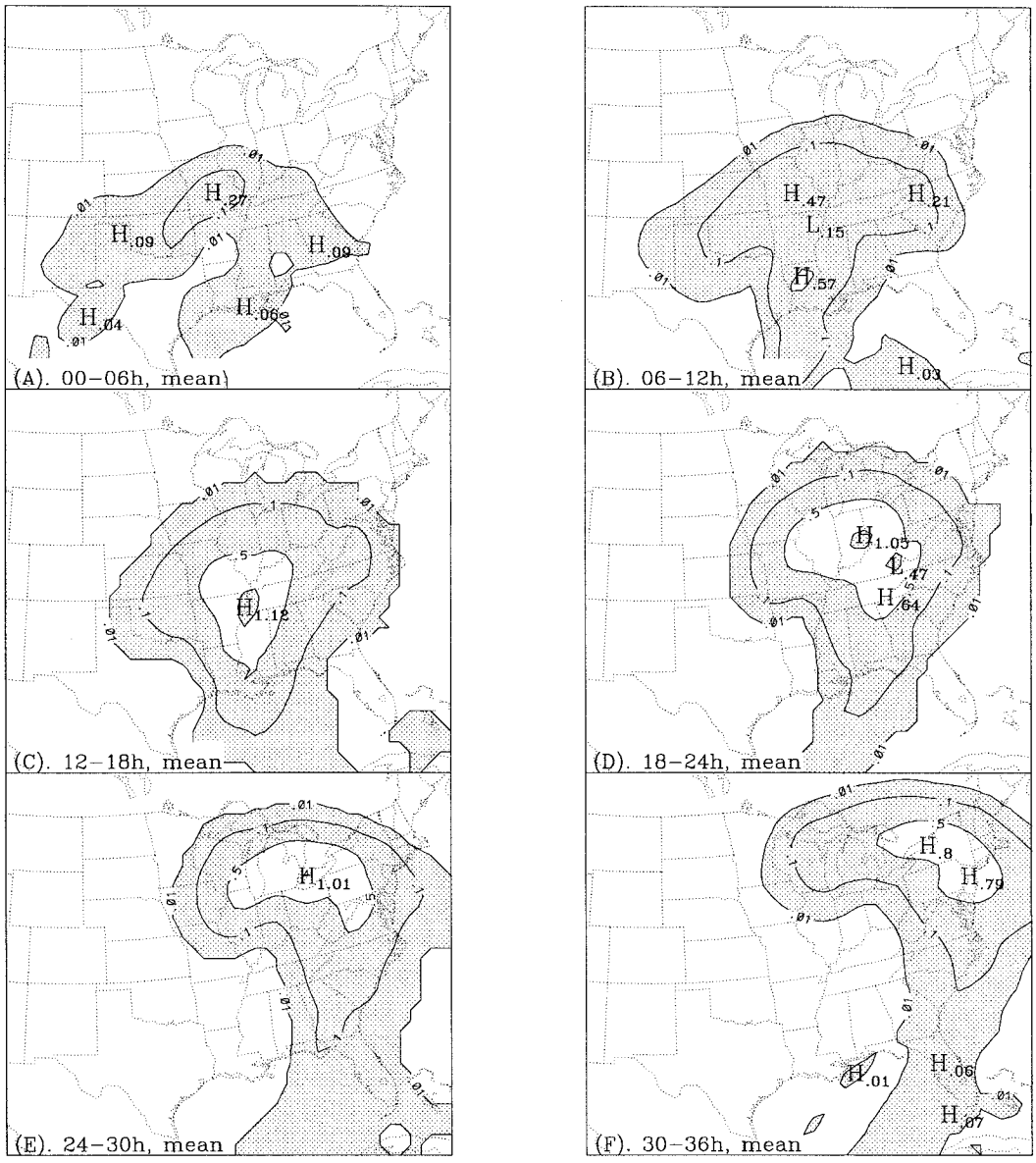

FIG. 13. The 25-member, ensemble mean forecast of precipitation for each of 6-h periods (a)–(f). Isohyets are the same as Fig. 7.

3) The development of a separate 0.5″ rainstorm in New England during the last 6 h of the 36-h period; and
4) The development, during the last 27 h of the 36-h period, of a 0.5″ rainstorm over the Gulf Coast states associated with a line of convection (with severe thunderstorms), moving eastward in advance of the surface cold front.

4. Verification and evaluation

a. Overview of mean model performance and forecast variability

Figures 8 and 9 show the bias score and equitable threat score (ETS), respectively, for three 6-h thresholds (0.01″, 0.10″, and 0.50″, panels a–c, respectively) along with one 12-h threshold (1.0″, Fig. 9d), averaged over the 25 individual forecasts. (The bias score and ETS are defined in the appendix.) Since reliable gridded analyses of precipitation, interpolated from rain gauge stations, are available only over the contiguous United States (Fig. 7), all QPF results are verified over the shaded region of Fig. 1.

The MM4 exhibits a mean error, or wet ''bias,'' at all thresholds by 24 h (Figs. 8a–c). This wet bias is largest and starts earliest for the highest accumulations. The bias score for the 0.50″ category increases from 2 to 10 between 18 and 36 h and is even greater for the 1.00″ per 12 h category (not shown). While the bias

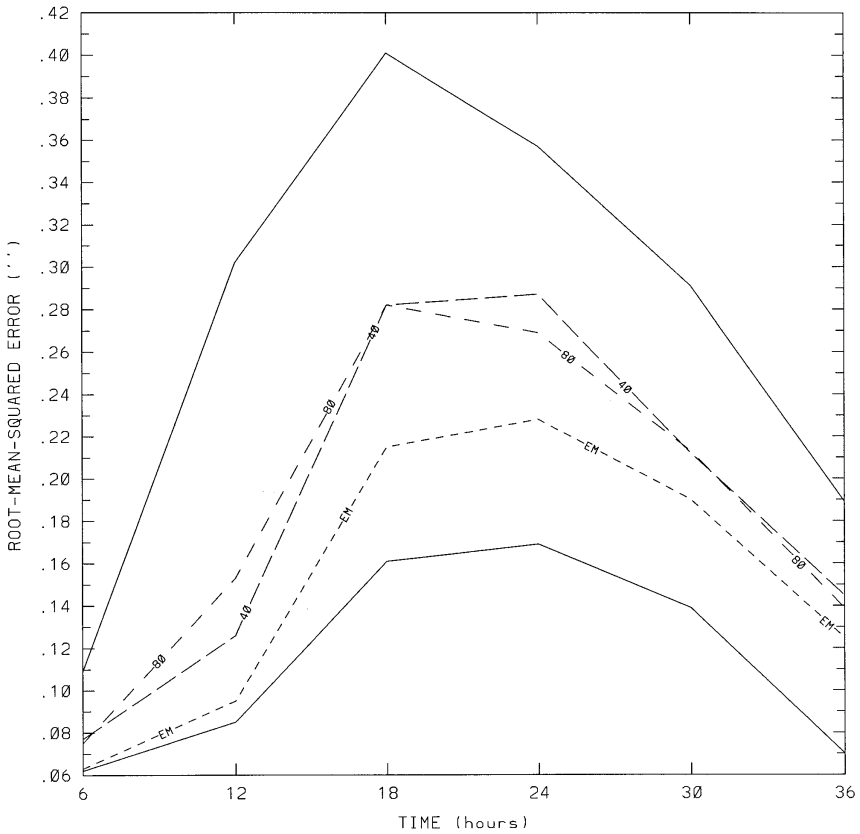

FIG. 14. As in Fig. 9 but for rmse.

score for the two smallest thresholds exceeds 2 by 36 h, it is close to one between 12 and 24 h and is less than one (dry bias) at 6 and 12 h. The range of the bias scores for the light thresholds (0.01″ and 0.10″) is relatively small; extrema are typically within a factor of 2 of the mean value. Extrema for the heavy amounts (0.50″) show much wider spreads, varying by an order of magnitude at 30 h and 36 h.

In terms of the storm total precipitation (Fig. 8d), spatially averaged over the verification region, the mean of the individual QPFs is three times too wet by 30 h, and all ensemble members are too wet at all by 18 h; early in the forecast (6 h), however, all members are too dry. The dry bias at the early stages is due in part to the ''spinup'' associated with our use of a static initialization, while the wet bias at the longer ranges is due in part to the forecast precipitation shield becoming increasingly displaced to the southwest of the observed shield, which begins to move outside the verification region by 24 h. The differences between extrema grow with time and reach a value of 2.2 by 30 and 36 h. It is important to note that the biases in spatially averaged, storm total precipitation, unlike those for specific thresholds, are not sensitive to our choice of Barnes coefficients used to generate the precipitation analyses.

In spite of these biases in areal coverage, the ETS (Fig. 9), which includes the bias score as a component (see appendix), reveals skillful forecasts for the 0.01″ and 0.10″ isohyets at all projections and indicates marginal skill at 12 h and 18 h for 0.5″ and at 24 h for 1.0″ (12-h forecast). The ETS for the individual forecasts exhibits considerable variability, however. The ETSs of all ensemble members for 0.01″ (Fig. 9a) are skillful at all projections; and except for 6 h, the ETSs for 0.10″ (Fig. 9b) are too. The range in the ETS between extreme ensemble members typically runs about 0.30 for both the 0.01″ and 0.10″ cutoffs, although it can be considerably larger at specific times (e.g., 24 h for 0.01″). The range in the ETSs for 0.50″ (Fig. 9c) or 1.00″ (Fig. 9d) at the time of highest skill (12 or 24 h, respectively) is also quite large (~0.50). Interestingly, 30 h is the only time at which no ensemble members are skillful at 0.50″.

The large ranges in the biases and ETSs suggest that large variability exists in the accumulated precipitation. The 6-h totals, spatially averaged over the verification region (Fig. 10a, curves labeled Amax and Amin), vary by as much as 0.10″ per grid point at the time when the mean and median values are largest (e.g., 18 h and 24 h). If normalized by the verifying amount, the range corresponds to approximately 50%–200% of the observations. Storm total accumulations (Fig. 10b, curves Amax and Amin) also exhibit appreciable spreads, with

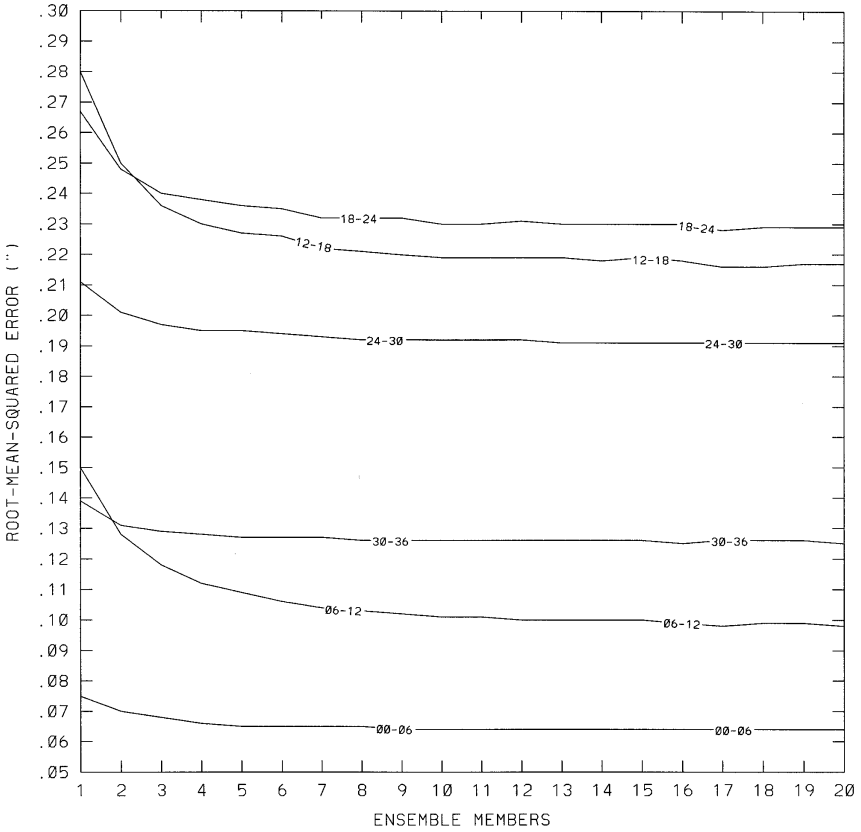

FIG. 15. Variation of rmse with the ensemble size for each 6-h period as determined from Monte Carlo permutation sampling. The results for ensemble sizes between 3 and 22 members are the average of 2000 trials selected from the 25 ensemble members without replacement; results for ensemble sizes of one and two are the average of all unique combinations.

36-h extrema of 0.32″ and 0.51″. Even in the presence of such variability, the extreme events do not envelop the verifying accumulation after 24 h because of the large bias: all ensemble members are too wet.

The variability in the area covered by 0.01″ or greater amounts (Figs. 10c and 10d) behaves somewhat differently. Extrema of 6-h accumulations (Fig. 10c, curves Amax and Amin) span observations from 12 to 30 h, inclusive, with the coverage being too big (small) for all members at 36 h (6 h). The behavior of the storm total area coverage (Fig. 10d, curves Amax and Amin) opposes that of the QPF itself (Fig. 10b): prior to 24 h the coverage for all members is too small, while afterward the MM4 ensemble spans the observations.

The spatial distributions of 6-h accumulations for the extreme ensemble members, based on total precipitation within the verification domain, are given in Fig. 11. Differences can be extreme indeed. At 12 h, for example, case 02 (Fig. 11b) displays a region of heavy precipitation over central Louisiana with a maximum of 3.50″ and a secondary maximum of about 1.00″ over eastern Kentucky. Case 25 (Fig. 11b1), on the other hand, reveals amounts less than 0.10″ over most of Louisiana and a maximum of only 0.32″ near the Kentucky–Tennessee border. At 30 h, case 19 (Fig. 11e) has an area with amounts greater than 1.00″ centered over lower Lake Michigan and another area over the North Carolina–Virginia border, whereas case 11 (Fig. 11e1) contained no regions with greater than 1.00″. The other times in Fig. 11 exhibit similar differences.

The spatial distribution of 6-h extrema, defined as the outliers at each grid point (Fig. 12), exhibits even larger variations than the area-averaged acculumations. The 18-h maxima panel (Fig. 12c), for example, shows amounts greater than 3.00″ situated over the Arkansas–Missouri border, while the minima panel (Fig. 12c1) shows values of about 0.10″. The area-averaged QPF for the gridpoint outliers (Fig. 10a, curves Gmax and Gmin) indicates a factor of 30 difference between maxima and minima panels at 18 h, while the area covered by measurable precipitation (Fig. 10c, curves Gmax and Gmin) differs by more than three times. In spite of such large ranges in extrema, the MM4 ensemble predicts maximum values of only about 0.50″ over northwestern Indiana, the one observed region with amounts greater than 1.00 at 18 h (Fig. 7c).

To summarize, the MM4 forecasts on average show skill in terms of the ETS, especially for the smallest

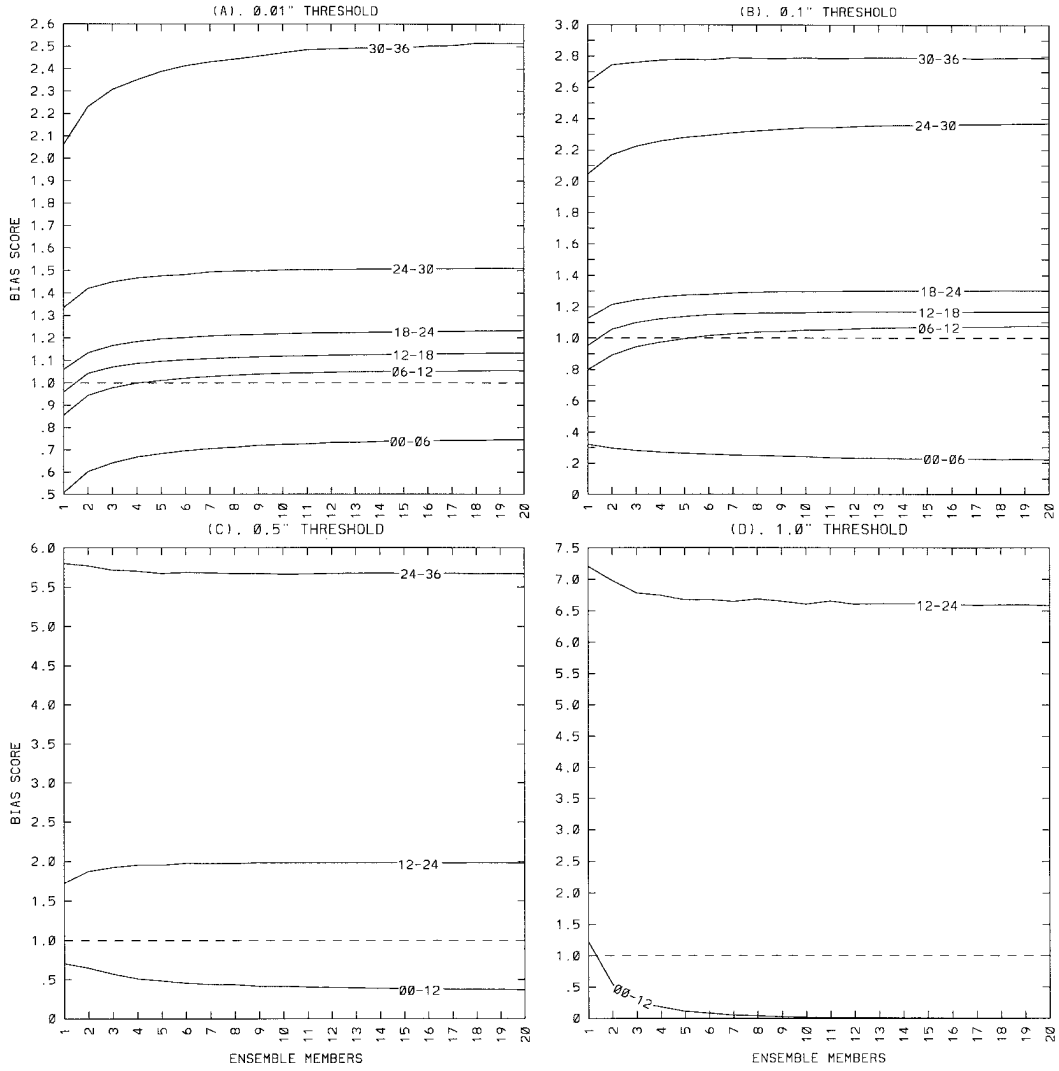

FIG. 16. Variation of the bias score in areal coverage with the ensemble size for (a) 0.01″ per 6 h, (b) 0.1″ per 6 h, (c) 0.5″ per 12 h, and (d) 1.0″ per 12 h thresholds. Results obtained as in Fig. 15.

thresholds (0.01″ and 0.10″). There is a dry bias during the first 6 h, due in part from our use of a static initialization; after 18 h there is a wet bias, owing in part to the forecast precipitation shield lagging behind the observed one, which moved more outside the verification region. Most importantly, large variability characterizes the QPF, both in terms of area-averaged accumulations and especially spatial distributions. Evidently, the sensitivity of short-range QPF to ICU, even over the data-rich contiguous United States, can be quite large.

b. Impact of ensemble averaging

Leith (1974) demonstrates how ensemble averaging can reduce the forecast error variance for an initial sample of normal, random initial analyses. His theoretical results indicate that the error variance for an ensemble mean forecast varies as $E_m = 0.5(1 + m^{-1})E_1$, where m is the number of ensemble members and E_m and E_1 are the error variances for an ensemble mean and the single deterministic forecasts, respectively. He explicitly notes that using an ensemble size as small as eight would appreciably increase the accuracy of the operational forecasts.

In view of Leith's theoretical result, it is of interest to examine how ensemble averaging impacts accuracy for precipitation and determine whether such small ensemble sizes can also lead to major improvements. Figure 13 shows the spatial distribution of QPF, averaged over all 25 ensemble members, while Fig. 14 shows the evolution of the 6-h rmse for the 25-member ensemble mean along with the average rmse for the individual forecasts and the extreme rmse scores. As Leith's results indicate, averaging over all 25 members indeed

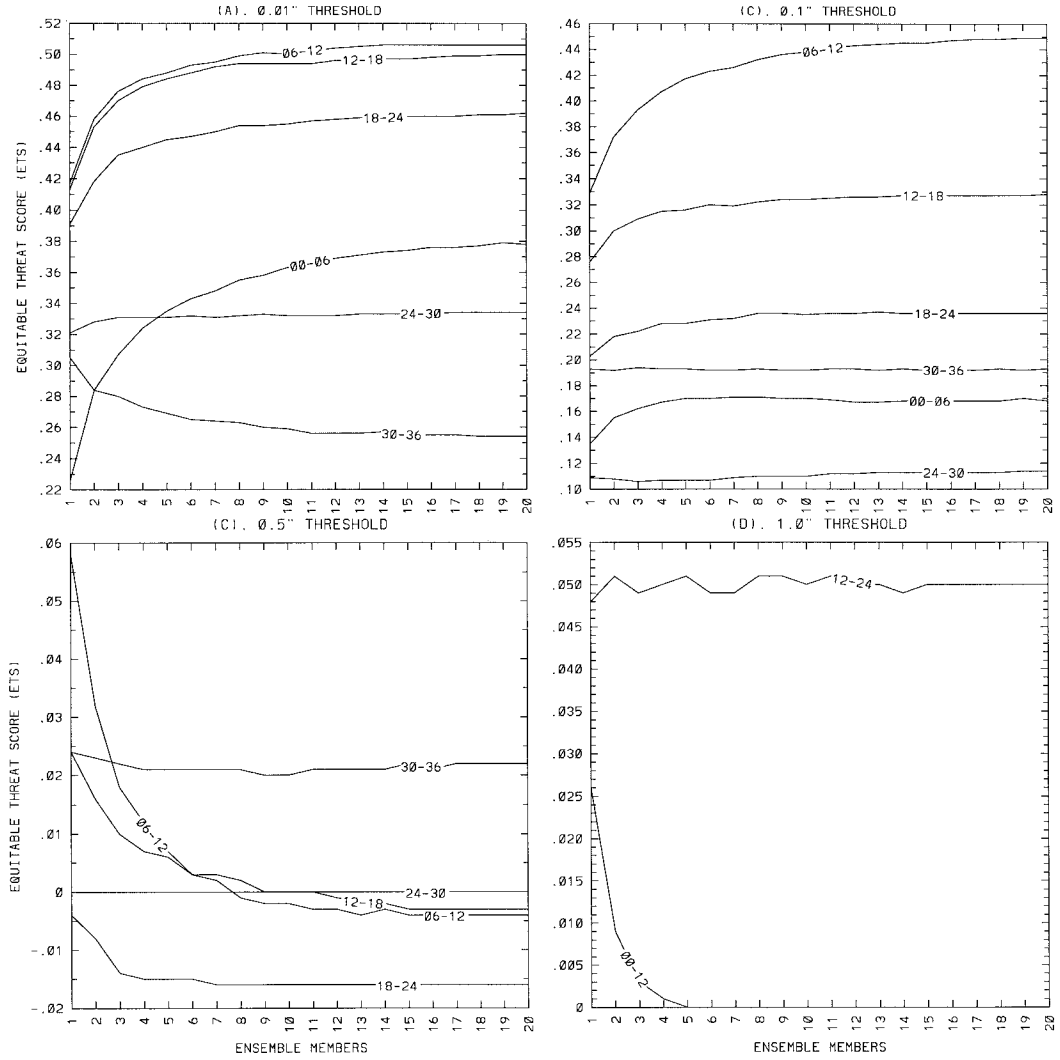

FIG. 17. Variation of the ETS with the ensemble size for (a) 0.01″ per 6 h, (b) 0.1″ per 6 h, (c) 0.5″ per 6 h, and (d) 1.0″ per 12 h thresholds. Results obtained as in Fig. 15.

lowers the rmse relative to the average rmse of the individual forecasts at all projections (Fig. 14), but the improvement is less than the $0.5(1 + m^{-1})$ coefficient suggests, *at least for this single forecast, which includes model error*. The rmse for a 25-member ensemble should be about 72% of the single value and does approach that level at 12 and 18 h, the times at which the bias for the area-averaged QPF is relatively small (Fig. 8d); typically, however, it runs approximately 85% or higher.

An issue of practical importance is the minimum number of ensemble members required to reap most of the benefit of ensemble averaging. Figure 15, which shows the variation in the rmse with number of ensemble members as determined by Monte Carlo permutation techniques, indicates that using as few as 8–10 members, as Leith suggests, yields 90% of the improvement obtainable from 25 members for this case.

While ensemble averaging *always* improves the rmse, Figs. 16–17 reveal that the behavior of the bias score and ETS is more equivocal. Figure 16 indicates that at the lightest thresholds (0.01″ and 0.10″) averaging over all 25 members produces a wetting effect compared to a single forecast, while at our heaviest threshold (1.00″ per 12 h) it produces a drying effect. Because averaging is a smoothing operation, ensemble averaging will, in general, enlarge (shrink) a precipitation area and create a wetter (drier) bias score for light (heavy) thresholds. Thus, whether ensemble averaging improves or degrades the bias score will depend upon the bias characteristics of the model forecasts. For our case, where the ensemble area coverage is too large for all thresholds after 18–24 h, ensemble averaging improves (degrades) the bias score for heavy (light) amounts but degrades (improves) bias scores for light (heavy) amounts. Figure 16 also indicates that this statistically generated com-

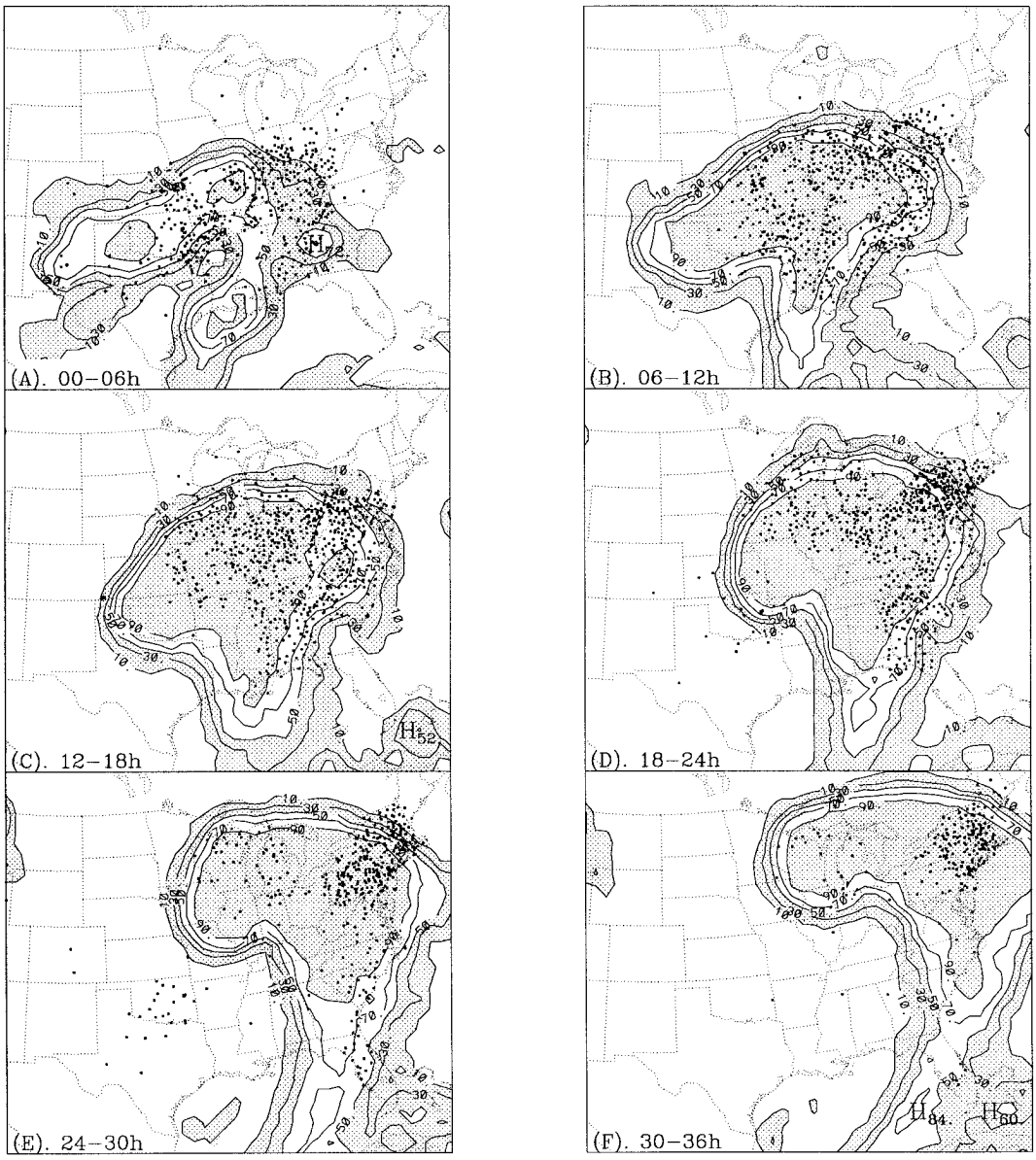

FIG. 18. Probability of 6-h accumulated precipitation exceeding 0.01″ for each of 6-h periods (a)–(f). Contour lines are 10%, 30%, 50%, 70%, and 90%. Shading areas are 10%–50%, 90%–100%. The small dots represent the rain gauges where the given category (e.g., ≥0.01″ for this forecast) is verified.

ponent of the bias score reaches its full impact with ensemble sizes of 5–6 members.

Whether ensemble averaging improves or deteriorates the ETS relative to the score for an individual forecast also depends upon the threshold and forecast range. Figures 9a,b show that the ETSs for the 25-member ensemble average at 0.01″ and 0.10″ thresholds are better than the average of the individual ETSs through 24 h, even exceeding the best individual ETS early in the forecast, but by 30 h they become no better and are even worse at 36 h for 0.01″. In fact, the degradation at 36 h for 0.01″ begins with ensemble sizes of only two (Fig. 17a). The ETSs of the 25-member ensemble mean for the 0.50″ and 1.00″ (per 12 h) isohyets never exceed the average of individual ETSs (Figs. 9c,d or Figs. 17c,d). Like the case for the rmse (Fig. 15), when ensemble averaging improves the ETS for the 0.01″ and 0.10″ thresholds, most of the benefit is achieved with only 5–10 members (Figs. 17a,b).

c. Probabilistic QPFs

As Brooks et al. (1995) discuss, the major advantage of SREF methods over deterministic forecasting is that explicit information about the probability density function (pdf), and thus forecast uncertainty, can be obtained

for any model parameter. For the purposes of verification and illustration of SREF as applied to QPF, we selected a priori five mutually exclusive, collectively exhaustive (MECE) categories of 6 h accumulations that correspond to some of the categories used for QPF verification by NCEP/National Weather Service (NWS)[4] and are consistent with the 0.10″ resolution of our dense rain gauge network. These five MECE categories are no measurable precipitation (pp $<$ 0.01″), 0.01″ $\leq$ pp $<$ 0.10″, 0.10 $\leq$ pp $<$ 0.50″, 0.50″ $\leq$ pp $<$ 1.00″, and pp $\geq$ 1.00″. A probability forecast at each grid point is constructed for each of the five categories based on the 25 ensemble members. The MECE categories are then used to produce a cumulative distribution function (cdf) for four thresholds, pp $\geq$ 0.01″, pp $\geq$ 0.10″, pp $\geq$ 0.50″, and pp $\geq$ 1.00″. Maps of the forecast cdf for the four categories are shown in Figs. 18–21, respectively, for all six 6-h periods.

At the 0.01″ threshold (Fig. 18), a widespread region of probabilities $P > 90\%$ stretches across the Mississippi River watershed by 12 h. The region of $P > 90\%$ translates to the northeast, maintaining a rather constant areal extend through 36 h. There is substantial overlap between the observed pattern of measurable precipitation and high probability regions (cf. Figs. 7 and 18), but the forecast pattern clearly lags behind the observed shield. With $P > 90\%$ over southwestern Ontario at 30 and 36 h, it is also clear that the observed cessation of precipitation over the region was missed by virtually all ensemble members. Every forecast member missed an observed area of light precipitation over Oklahoma and northern Texas during the 24–30-h period. The forecast characteristics of the 0.10″ isohyet (Fig. 19) are similar to those for 0.01″.

Amounts in excess of 0.50″ appear in the ensembles within the first 6-h period and last for the remainder of the forecast (Fig. 20). Although amounts above 0.50″ are not analyzed at 6 h, the forecast region of greatest probability does coincide with the analyzed region of maximum precipitation (Fig. 7a). By 12 h, however, maximum probabilities begin to lag behind the observed area. At 24 h an axis of enhanced probabilities extends from eastern Tennesee along the Alabama–Georgia border; this axis appears to be associated with an observed line of convective rainfall that produced the 0.50″ isohyet and was situated through central Georgia (Fig. 20d). The secondary cyclogenesis over southern New England at 36 h also yielded amounts above 0.50″ near Cape Cod, but the MM4 ensemble placed its highest probabilities of amounts greater than 0.50″ well to the southwest of the verifying position.

QPFs in excess of 1.00″ do not appear until 12 h, with the area of maximum likelihood (10%–30%) being centered over northern Louisiana (Fig. 21b). Nonzero probabilities exist throughout the remainder of the forecast, with an area of $P > 50\%$ propagating from western Tennessee at 18 h into southeastern Indiana by 24 h (Figs. 21c,d). This area is shifted southward of the observed coverage over northwestern Indiana at 18 h, the only 6-h period during which amounts of at least 1.00″ are analyzed (Fig. 7c). Unlike the observed storm where the heaviest precipitation occurs several hundred kilometers ahead (northeast) of the surface low during explosive deepening (cf. Figs. 6b and 7c,d), the MM4 forecast ensembles tend to place the heaviest precipitation much closer to the cyclone (cf. Figs. 6b and 21c,d).

A verification measure of probabilistic, categorical forecasts that is both strictly proper (i.e., does not award hedged forecasts; see Wilks 1995, 267–268) and sensitive to distance is the ranked probability score (RPS; Epstein 1969; Murphy 1971; Wilks 1995; the RPS is defined in the appendix). The best possible RPS that corresponds to a perfect categorical forecast is zero, and the worst possible score is $J - 1$, where J is the number of MECE categories. [Note that for $J = 2$, the RPS reduces to the well-known half–Brier score (Brier 1950).] To judge the accuracy of our QPF predictions, we first calculate the RPS based on the forecast pdf's at each grid point ($\mathrm{RPS}_{\mathrm{fcst}}$). The skill of MM4 QPF's relative to a climatological control forecast is then evaluated by computing a ranked probability skill score (RPSS), defined as

$$\mathrm{RPSS} = 1 - \frac{\mathrm{RPS}_{\mathrm{fcst}}}{\mathrm{RPS}_{\mathrm{clim}}}, \tag{1}$$

where $\mathrm{RPS}_{\mathrm{clim}}$ is the RPS based on long-term frequencies for December.[5] Any value of RPSS > 0 denotes a skillful QPF relative to climatology, with RPSS $= 1$ representing a perfect forecast. We believe that a climatological control is probably more difficult to beat than persistence for relatively short-lived events like rainstorms.

The spatial distribution of the RPSS for every 6-h period (Fig. 22) shows that the ensemble QPFs are skillful over most, but not all, of the verification region. For example, on average approximately 77% of the region exhibits skill relative to climatology (Table 1). More specifically, about 70% of the precipitating area and 80% of the nonprecipitating area show skill. In view of the skillful area greatly exceeding the unskillful area, we conclude that the MM4 ensembles provide useful QPF guidance for this cyclogenetic event.

A fundamental issue of SREF that is crucial to its effective operational implementation is the determination of the minimum number of ensemble members re-

[4] NCEP/NWS verifies 0.01″, 0.10″, 0.25″, 0.50″, 0.75″, 1.00″, 1.50″, 2.00″, etc.

[5] The frequencies for the five QPF categories are taken from NOAA (1987).

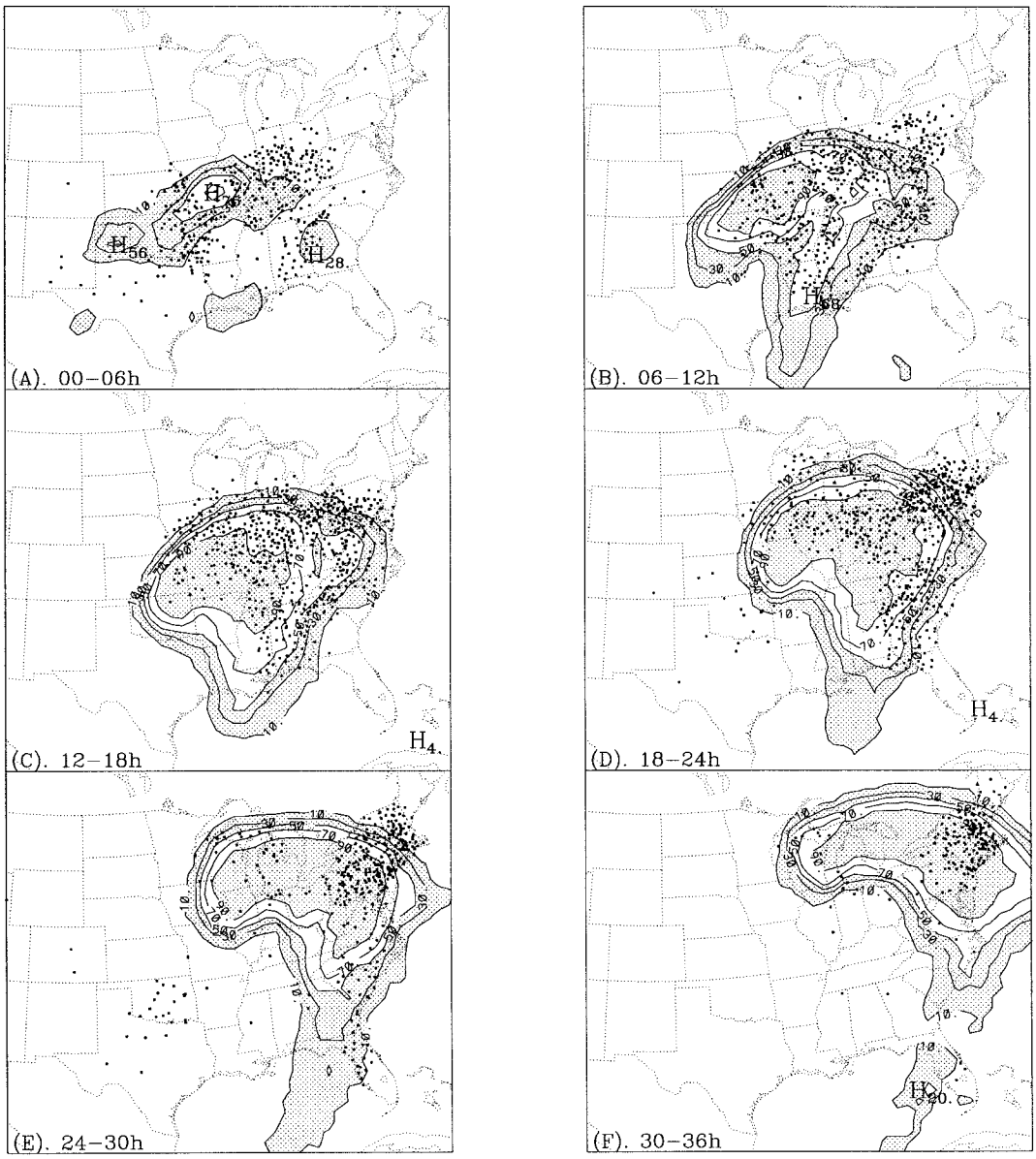

FIG. 19. Same as in Fig. 18 except for amounts exceeding 0.1″.

quired to estimate robustly the forecast evolution of the pdf. To examine this question in the context of our single MM4 ensemble, we again employed Monte Carlo permutation techniques to test the sensitivity of the RPS to variations in the ensemble size. The results are shown in Fig. 23 where the mean RPS, $\overline{\text{RPS}}$, is obtained by averaging over all grid points within the verification domain. The figure reveals that RPS behaves as a damped exponential with respect to ensemble size—that is, $\text{RPS}(m) \approx \alpha \exp(-\beta m) + \gamma$, where α, β, and γ are constants, and m is the number of ensemble members. The e-folding value of m averages five members for the all forecast projections. Hence, even with ensemble sizes as small as 5 (10) members, about 63% (90%) of the improvement relative to $m_{\max} = 25$ members is realized. The results for this cyclogenesis, if representative of cases with major cyclogenesis, suggest that ensemble size as small as 5–10 could greatly benefit wintertime QPFs if approximately five MECE categories are employed. The result is consistent with the preliminary conclusions of Hamill and Colucci (1996, 1997), who also find improvements in QPF accuracy using a 10-member ensemble based on the 80-km version of the NCEP Eta Model relative to a single forecast by the 29-km meso–Eta Model. The applicability of our finding to a finer stratification of the QPF categories is clearly dubious, however, since as the number of MECE categories increases, the number of ensemble members

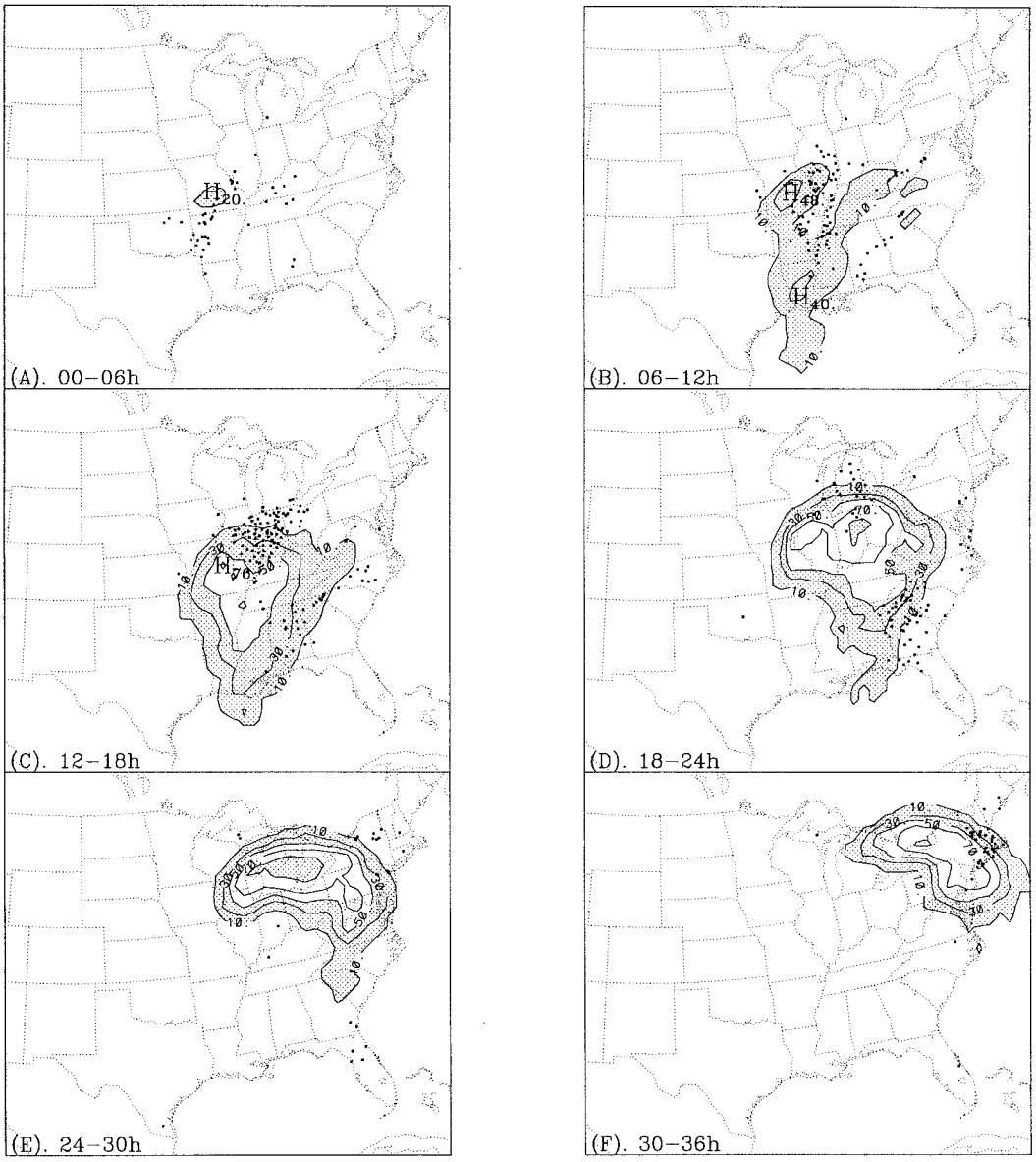

FIG. 20. Same as in Fig. 18 except for amounts exceeding 0.5″.

needed to achieve a given fraction of the potential RPSS also increases.

Comparison of Fig. 23 and Table 1 suggests an apparent discrepancy. Figure 23, which is based on $\overline{\text{RPS}}$, indicates that the QPFs for the two 6-h periods ending 24 and 30 h are not skillful relative to climatology. Conversely, Table 1 indicates that all periods are skillful in terms of the majority of grid points being skillful. The difference arises because Fig. 23 retains the full distance sensitivity of the RPS, whereas Table 1 does not. The synoptic interpretation is clear: at those 23% of grid points where the RPS is not skillful, the MM4 ensemble is far less accurate than at the skillful grid points.

d. Comparison with other model forecasts and ensembles

1) THE OPERATIONAL NGM FORECAST

It is also useful to compare the MM4 ensemble forecast with the state-of-the-art operational model in use at the time of the storm, in this case the Nested Grid Model (NGM), a component of the Regional Analysis and Forecasting System (Hoke et al. 1989). Since the NGM precipitation forecast is in the form of 12-h accumulations, it was necessary to combine successive 6-h observed and MM4-predicted amounts in order to compare the performance of the two models. Results for the

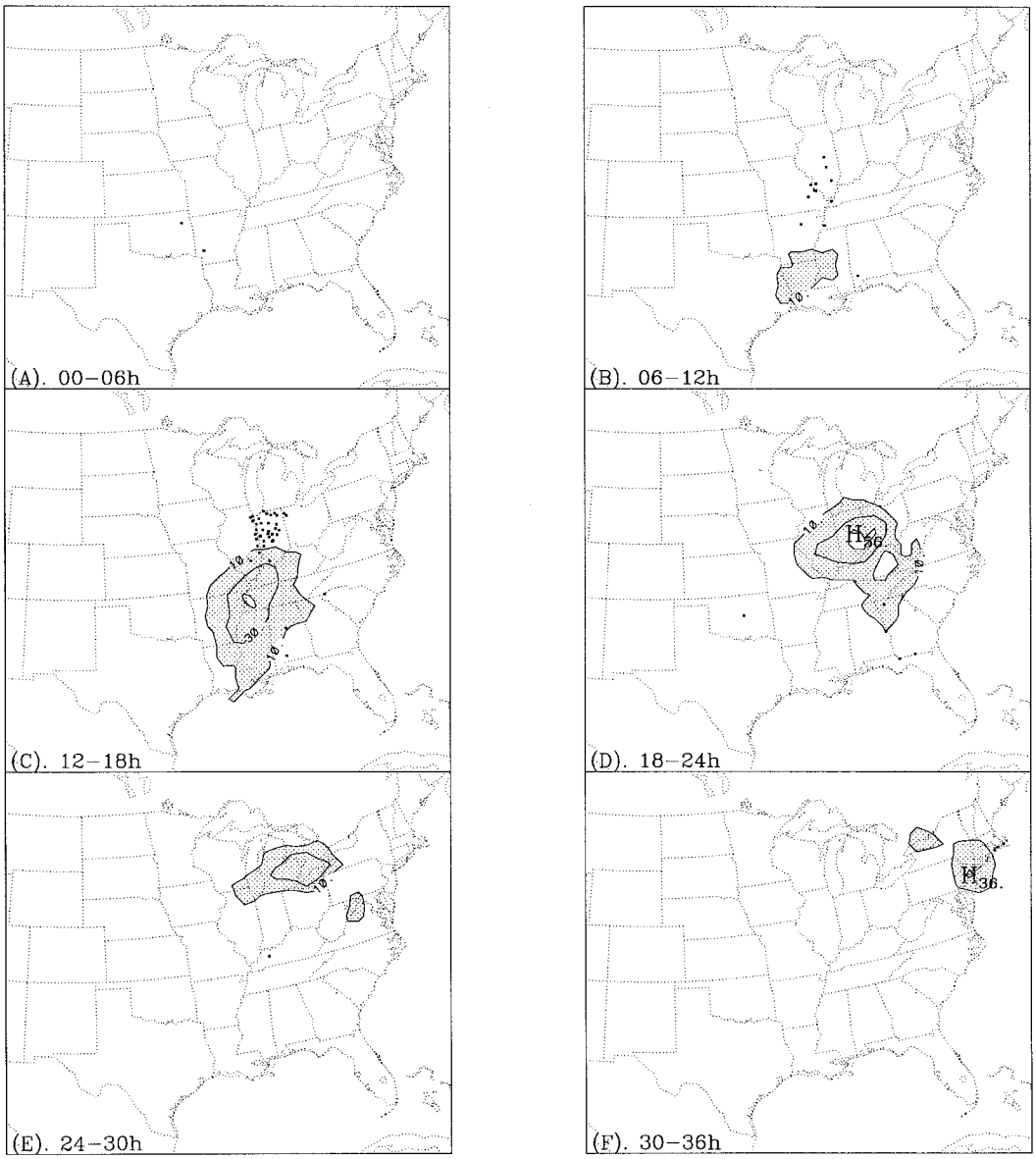

FIG. 21. Same as in Fig. 18 except for amounts exceeding 1.0″.

three 12-h periods in the 36-h interval are shown in Fig. 24.

During the first 12 h, both the NGM (Fig. 24a2) and the MM4 ensemble mean (Fig. 24a1) predict a maximum north of the cyclone center. The MM4 position, over southern Missouri, is much better. Neither model predicts enough precipitation, the NGM showing a somewhat greater discrepancy. Both models produce maxima in the Gulf Coast states, where only a small 0.5″ maximum is analyzed over Alabama (Fig. 24a3). The amount in the NGM is about twice the observed, while the MM4 amount is approximately correct. The NGM position, however, is better. The MM4, on the other hand, is obviously more accurate in precipitation areal coverage, especially in the placement of the northeastern edge of the precipitation, while the NGM predicts too small a precipitation region.

In the next 12 h, both models produce a large and substantial rainstorm; the MM4's (Fig. 24b1) center is in southern Indiana and the NGM's (Fig. 24b2) two centers are in western Illinois and eastern Kentucky. In the analysis (Fig. 24b3) there is a single center in northern Indiana, the amount, 1.26″, intermediate between the 1.50″ for the MM4 and the 1.12″ for the NGM. The convective system discussed earlier produced a separate and distinct maximum in the Southeastern states, while the models show only a southward extention of the main rain area too far west, the MM4 especially so. The MM4,

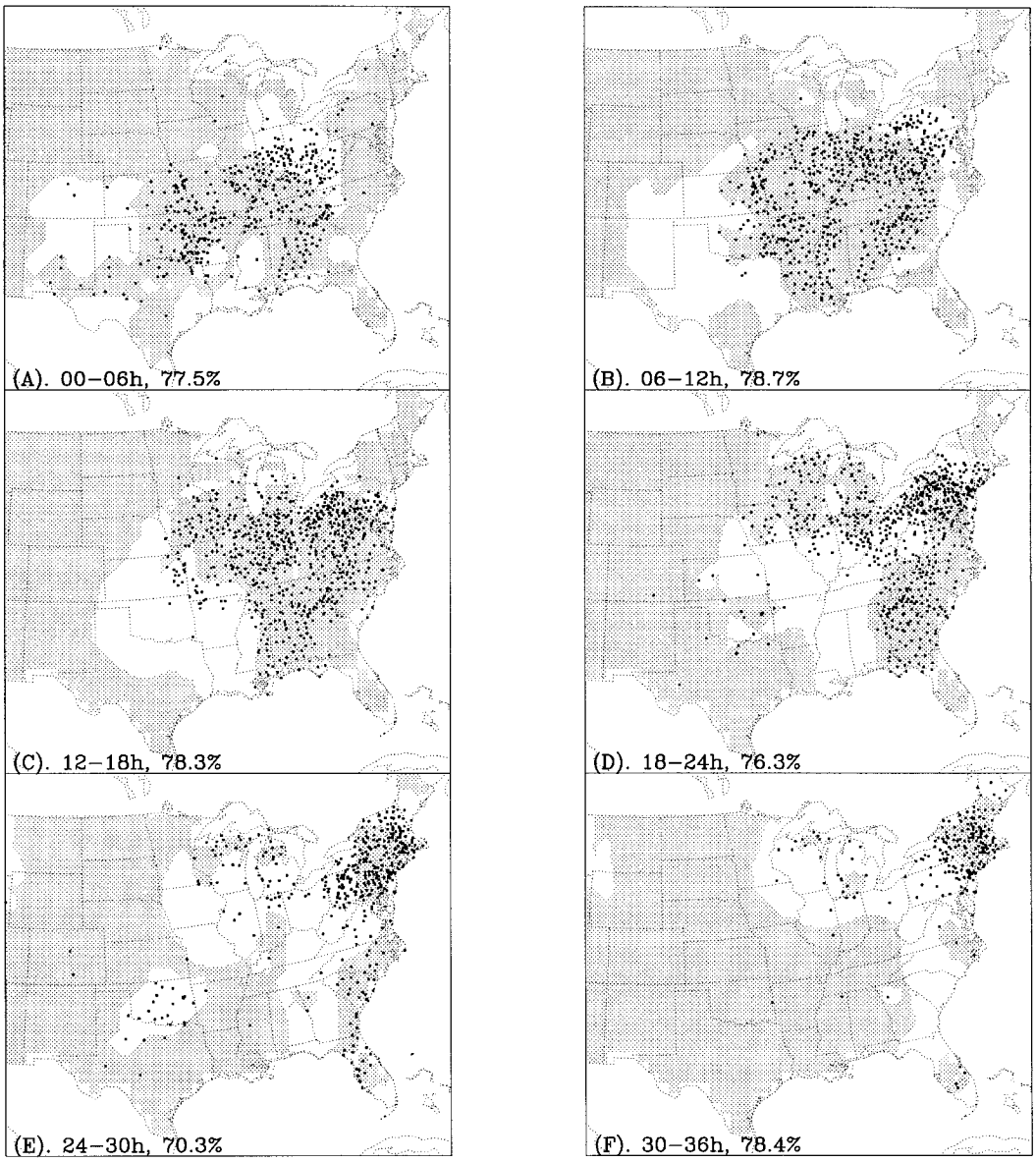

FIG. 22. Region where the ranked probability skill score (RPSS) is skillful during every 6-h period (a)–(f). Shading indicates region where the RPSS is skillful with respect to the climatological control forecast. Small dots represent the rain gauges where measurable rainfall was observed (≥0.01″) during the 6-h period. Percentage value at the bottom of each plot is the fraction of verifying area with RPSS > 0.0 for the indicated 6-h period.

TABLE 1. The fraction of a certain verifying area (grid points), showing skill (RPSS > 0) with respect to the corresponding climatological control forecast in probabilistic forecasting. Three types of area are considered based on the observations: all verifying, precipitating, and nonprecipitating areas. Columns 2–7 are for every 6-h period and the last column gives the 36-h average. All values are in percentage (%).

	0–6 h	6–12 h	12–18 h	18–24 h	24–30 h	30–36 h	Mean
All verifying	77.5	78.7	78.3	76.3	70.3	78.4	76.6
Precipitating	72.9	82.8	78.9	68.2	50.0	61.3	69.0
Nonprecipitating	79.9	76.4	78.0	81.1	79.6	81.3	79.4

on the other hand, is somewhat more accurate in the placement of the eastern edge of the precipitation in the mid-Atlantic states. Overall, the models are more accurate during this 12-h period than during the others.

Figures 24c1, 24c2, and 24c3 show that both models perform poorly during the last 12-h period. They are equally poor in failing to weaken the major precipitation area over the Great Lakes states as discussed earlier. Neither shows a maximum over New England, although irregularities in the shapes of the 0.5″ isohyets show a tendency in that direction. However, quanti-

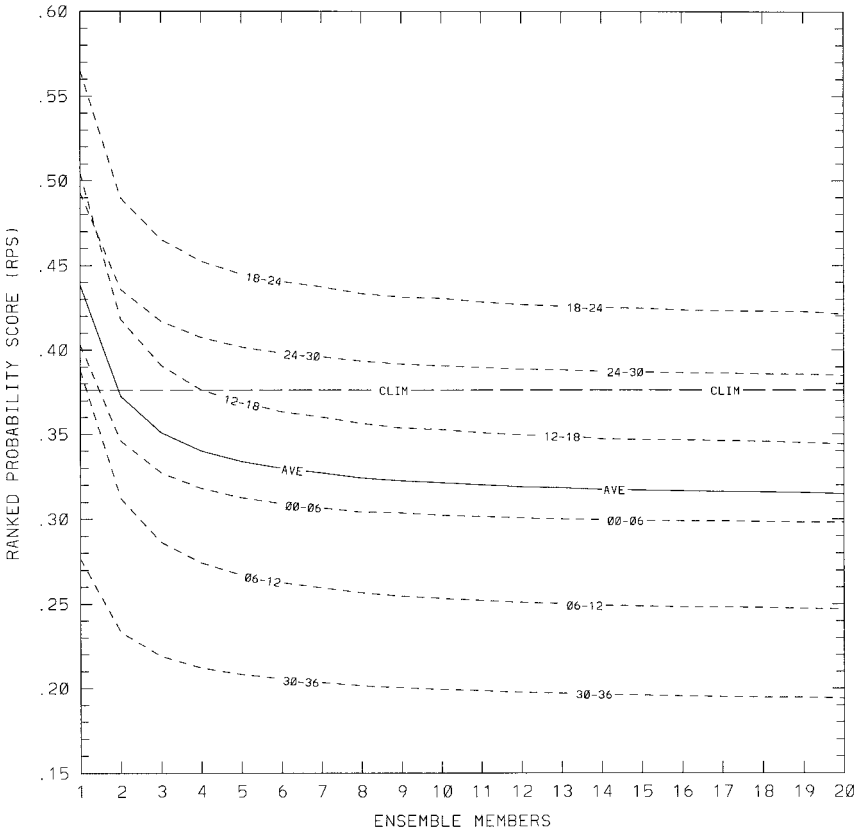

FIG. 23. Variation of the ranked probability score (RPS) with the ensemble size for each 6-h period. Horizontal line marked ''CLIM'' is RPS value for the climatological control forecast. Results obtained using Monte Carlo permutation techniques as in Fig. 15.

tatively speaking, the MM4 is slightly better over that region. Neither model places the tongue of precipitation associated with the convective system far enough south on the Florida penninsula and neither terminates the precipitation far enough east in the mid-Atlantic states.

2) TWO HIGHER-RESOLUTION MM4 FORECASTS

To compare QPF for the 80-km ensemble with higher-resolution forecasts, an MM4 forecast for a (1) 15-layer, 40-km mesh and (2) a 29-layer, 80-km mesh were run for the domain shown in Fig. 1. The σ layers for the 29-layer run were chosen to minimize the spurious generation of internal gravity waves due to inconsistent vertical and horizontal resolutions (e.g., Lindzen and Fox-Rabinowitz 1989; Persson and Warner 1991).[6] The 40-km (29 layer) forecast was run for 36 h (24 h). The initial fields for the higher resolution forecasts were obtained by linearly interpolating the unperturbed 80-km analyses to the finer grids. All other model options for the higher-resolution forecasts were set to the same values as the 80-km forecast ensemble.

The rmse and ETS for the 40-km run, which are also given in Figs. 14 and 9, are close to, but typically less accurate than, the mean scores of individual 80-km forecasts. The spatial distribution of the 6-h accumulations for the 40-km forecast (Fig. 25), with the exception of about 0.5% of the grid points, lies within extrema of the 80-km ensemble (Fig. 12). Table 2 indicates that the differences between the 80-km and 40-km control forecasts are typically much smaller than those due to the error growth of ICU among the ensemble members, especially for projections 24 h and shorter. Table 3 and Fig. 25 indicate in even smaller impact from doubling the number of vertical layers. (Accuracy measures for the 29-layer run are not shown.) It appears that doubling the horizontal or vertical resolution for a single forecast did not improve the QPF for this case of major cyclogenesis, much less match any improvements in accuracy and skill in the 80-km/l5-layer runs due to SREF methods (Figs. 9 and 14). It is possible that different cases or synoptic events might behave differently, or that further increases in

[6] The 29 layers are σ = 0.025, 0.075, 0.125, 0.175, 0.21, 0.23, 0.25, 0.27, 0.29, 0.3125, 0.3375, 0.3625, 0.3875, 0.4167, 0.45, 0.4833, 0.525, 0.575, 0.625, 0.675, 0.725, 0.775, 0.825, 0.87, 0.91, 0.945, 0.97, 0.985, and 0.995.

model resolution could yield greater improvements than our results indicate.

3) ANOTHER 25-MEMBER MM4 ENSEMBLE

The results of Stensrud and Fritsch (1994a,b) demonstrate that QPF is strongly affected by both ICU and changes in physical parameterizations. They argue that SREF should consider the impacts of both sensitivities. Hence, it is of interest to compare the above MM4 forecast ensemble with another MM4 forecast ensemble that is based on a different cumulus parameterization and set of perturbed initial analyses. For this reason, we also ran a 25-member ensemble for the 14–16 December 1987 cyclogenesis, substituting the Kuo–Anthes cumulus scheme (KA) (Kuo 1974; Anthes 1977) for the Grell/explicit schemes (Grell et al. 1991; Hsie et al. 1984). We also by-passed the objective analysis program (Manning and Haagensen 1992) in the MM4 preprocessing package for this second ensemble, a decision that yields somewhat smoother (i.e., poorer fit to the observations) analyses over the data-rich continents. The experimental design crudely mimics the type of differences that might arise in the operational environment between two forecast ensembles based on different analysis–forecast systems having identical resolutions. Only the highlights of the comparison are discussed here.[7]

A comparison of the average rmse score for the two ensembles indicates that the Grell ensemble produces, on average, more accurate QPFs for this case (Fig. 26a). The reduction in rmse due to the averaging of the 25 KA members (Fig. 26b) exceeds that due to the use of the Grell/explicit schemes and the enhanced (and presumably more accurate) initial state at all times. Comparison of the 0.01″ ETS gives a similar result (except for 6 and 36 h, Figs. 26c,d). These findings suggest that the potential benefit from SREF, at least for this case of a strongly forced cyclonic situation, is comparable to or can exceed that obtainable from changes or improvements in an analysis-forecast system.

5. Summary and concluding remarks

In this paper, the impact of ICU and SREF on QPF was examined for a case of explosive cyclogenesis that occurred over land. A version of the PSU–NCAR MM4, a limited-area model with 80-km horizontal resolution and 15 layers, was used to produce a 25-member 36-h forecast ensemble. Lateral boundary conditions for the MM4 model were provided by ensemble forecasts from a global spectral model, the NCAR CCM1, a situation that ensures unbounded predictability error growth. The initial perturbations of the ensemble members possessed a magnitude and spatial decomposition, which closely matched estimates of global analysis error, but they were not dynamically conditioned (e.g., Mureau et al. 1993). Results for an 80-km ensemble forecast were compared to forecasts from the then operational NGM, a single 40-km MM4 run, a single 29-layer MM4 run, and a second 25-member MM4 ensemble based on a different cumulus parameterization and slightly different initial conditions.

The primary findings of this study are as follows.

1) Acute sensitivity to initial condition uncertainty marks QPF. Extrema in 6-h accumulations at grid points varied by as much as 3.00″ by 12 h. Such variations occur even without the use of dynamically conditioned perturbations.
2) Ensemble averaging reduces the rmse for precipitation. The degree of improvement was less than that indicated by Leith (1974), owing in part to the presence of model error. Consistent with Leith (1974), nearly 90% of the improvement was obtainable using ensemble sizes as small as 8–10.
3) Bias scores for the forecast ensemble relative to a single forecast are wetter for light thresholds but drier for heavy thresholds. This relationship is a direct manifestation of the smoothing nature of the averaging process. Equitable threat scores, which contain the bias score as a component, can also be adversely affected by ensemble averaging. This result is generalizable to any ensemble-averaged QPF.
4) The RPS decreases exponentially with increasing ensemble size. The *e*-folding value averages five members for all forecast projections for the five MECE accumulation categories employed here.
5) The majority of grid points (about 77% on average) exhibit skill (RPSS) with respect to the climatology at all times, even in spite of the tendency for MM4 QPFs to be too wet and lie to the southwest of observations. The numerical value of the RPSS, averaged over the verification region, however, is not skillful for the 18–24-h and 24–30-h periods. This indicates that the RPS at approximately 23% of grid points with no skill is substantially larger then the RPS at the skillful locations. These large errors are believed not to be characteristic of this model. Nevertheless, the MM4 ensembles could provide useful QPF guidance for this case.
6) The ensemble QPF with the 80-km grid model is more accurate than a single forecast with the 40-km run in terms of better rmse and ETS statistics. Except at 0.5% of all grid points, the 40-km QPF values are contained within the bounds defined by the 80-km ensemble. The predictability error growth and forecast dispersion due to ICU exceeded that due to either doubling the horizontal or vertical resolution.
7) SREF techniques can provide increased accuracy in QPF even without further improvements in analysis–forecast systems.

[7] A detailed comparison of the two is the subject of research still in progress.

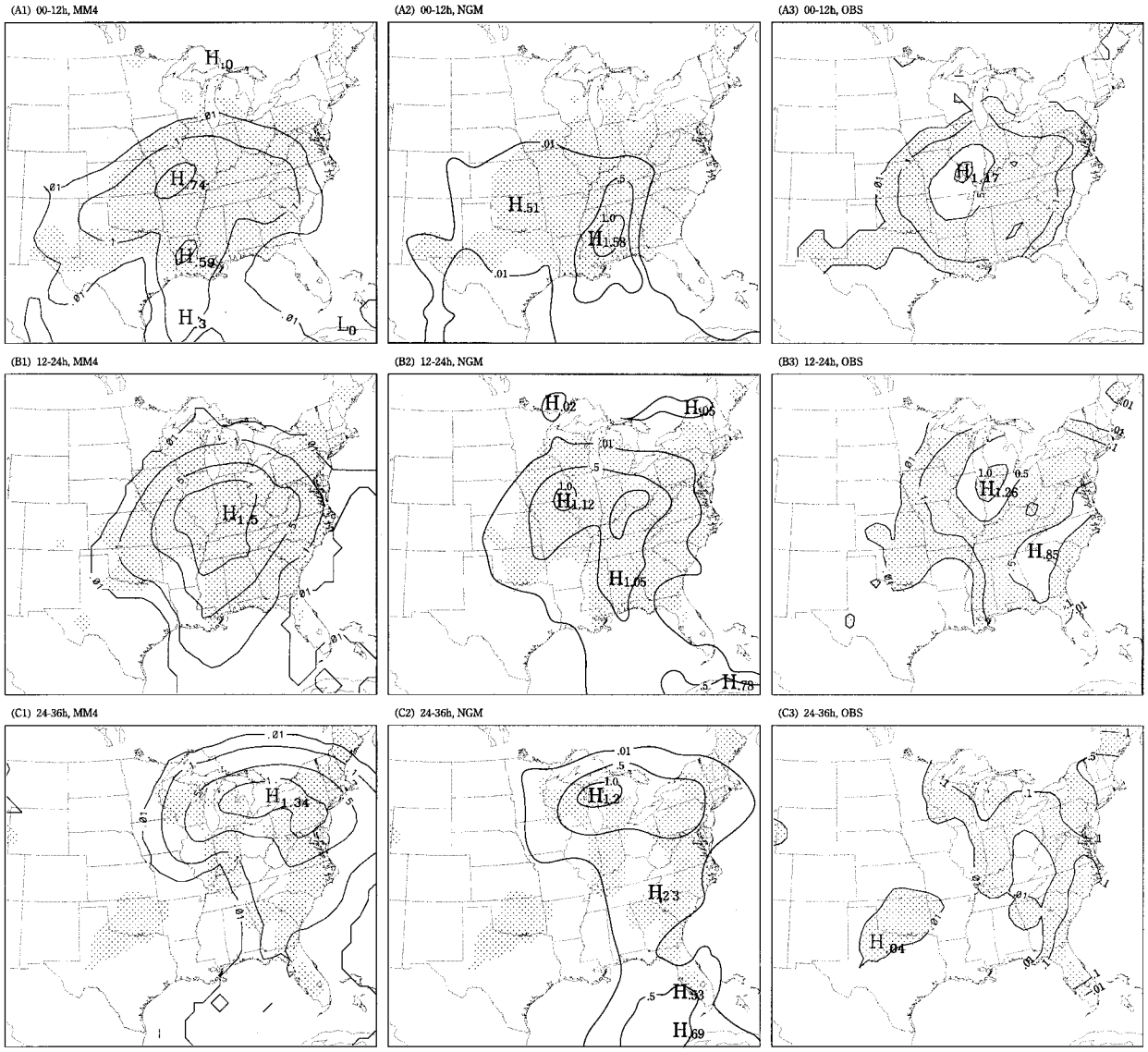

FIG. 24. Isohyets of observed precipitation (inches) accumulated for the 12-h periods (a3) 0–12 h, (b3) 12–24 h, and (c3) 24–36 h. Predicted amounts for the same periods appear in (a1), (b1), and (c1) for the MM4 ensemble average, (a2), (b2), and (c2) for the NGM forecast. The shaded areas in the leftmost and middle panels indicates the corresponding observed 0.01″ thresholds, which are in the rightmost panel. Selected isohyets are the same as Fig. 7.

Although the results for this single case represent perhaps the most encouraging evidence to date of the potential benefit of SREF and its application to QPF, our study suggests many future research paths. Of course, ensemble QPF needs to be examined within the context of many more cases. We sampled only one type of system, wintertime explosive cyclogenesis. We recognize that other explosive cyclones could behave differently from the one we examined, and, obviously, other weather events could too. Particularly important to a thorough examination of SREF and QPF is the sampling of extreme precipitation events, especially convective events under weak synoptic forcing that characterize the warm season.

Although our results indicate that ensemble sizes as small as 5 (10) could provide about 65% (90%) of the benefit obtainable through SREF in terms of reduced RPS for categorical, probabilistic QPFs, we again emphasize that our findings are applicable only to this case and particular forecast model. As noted earlier, the use of finer stratifications than five QPF categories would require more members to reach the same level of improvement. Because the operational NCEP currently uses a finer stratification of precipitation categories than we used here, the problem of the minimum number of ensemble members clearly warrants additional research for a number of weather situations. In view of the finite limitations of computational

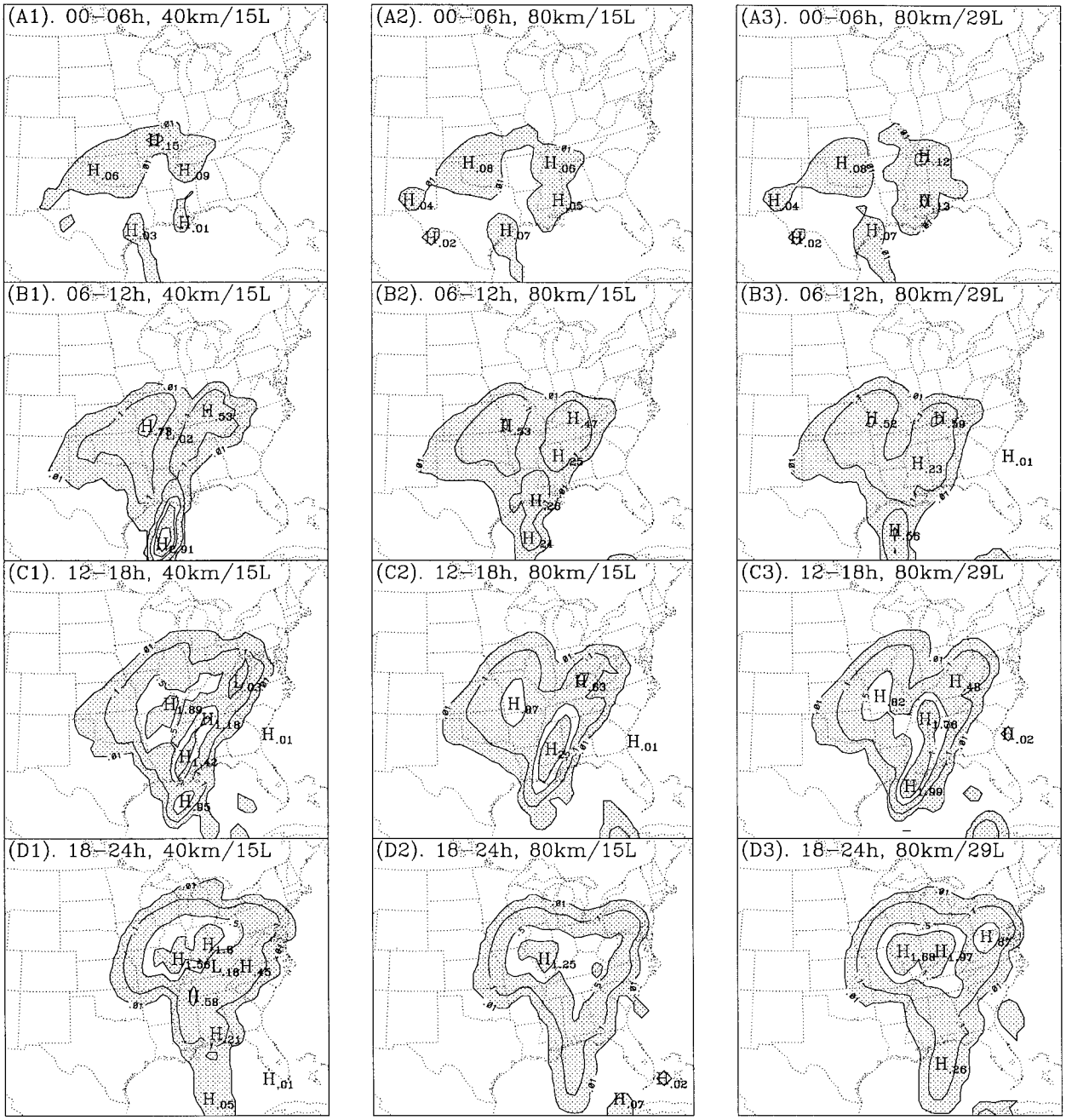

FIG. 25. QPFs for each 6-h period by the MM4 with 40-km mesh length. To faciliate comparision with the 80-km ensemble results, the results have been interpolated from the 40-km grid to the 80-km grid with a nine-point filter of Holloway (1958, his Fig. 10). Isohyets are the same as in Fig. 7.

resources in the operational environment, the inverse problem is perhaps even more important: given a fixed number of ensemble members (e.g., 10 members for the once weekly 80-km Eta ensembles currently run by NCEP on an experimental basis), what is the maximum number of QPF categories that can be robustly estimated?

We state once again that the potential advantages of SREF increase with increasing accuracy of the model/analysis system. To reap the maximum benefits from SREF, especially as applied to QPF, requires both further improvements in the NWP models and data analysis systems and more research into ways in which to use the output from forecast ensembles most effectively.

Acknowledgments. This paper is based on the first author's Ph.D. dissertation research. Preliminary results were presented in the 1994 NMC SREF Workshop (Camp Springs, Maryland, July 1994; see Brooks et al. 1995) and

TABLE 2. The percentage of 80-km forecast pairs among the 300 unique pairings with smaller rmse's (first row) and larger spatial correlation coefficients (second row) than the difference between the 80-km control and 40-km forecast. A value less than 5.0% is akin to statistical significance at the 5.0% level.

	0–6 h	6–12 h	12–18 h	18–24 h	24–30 h	30–36 h
rmse	0.0%	35.3%	0.3%	0.0%	8.3%	48.3%
Correl. coef.	0.0%	1.3%	0.3%	0.0%	15.0%	46.0%

TABLE 3. As in Table 2 except for the 29-layer forecast. Values for forecast projections beyond 24 h are not available (N/A).

	0–6 h	6–12 h	12–18 h	18–24 h	24–30 h	30–36 h
rmse	0.0%	0.0%	2.3%	5.0%	N/A	N/A
Correl. coef.	0.0%	0.0%	2.3%	0.3%	N/A	N/A

the Second Workshop on Adjoint Applications in Dynamic Meteorology (Visegrad, Hungary, May 1994; see Prager et al. 1995). The project was supported by the National Science Foundation (NSF) under Grants ATM-9118898, ATM-9318751, and ATM-9419411. All model forecasts were performed at the Scientific Computing Division (SCD) of NCAR. NCAR is supported by NSF. We thank the following people who offered us help with various aspects of the research: David P. Baumhefner, Ronald M. Errico, Joseph J. Tribbia, Tomislava Vukicevic, Sue Chen, Tom Mayer, Gary Bates of NCAR, and Brian Auvine of the University of Arizona.

We also thank the anonymous reviewers whose suggestions led to substantial improvements in the manuscript.

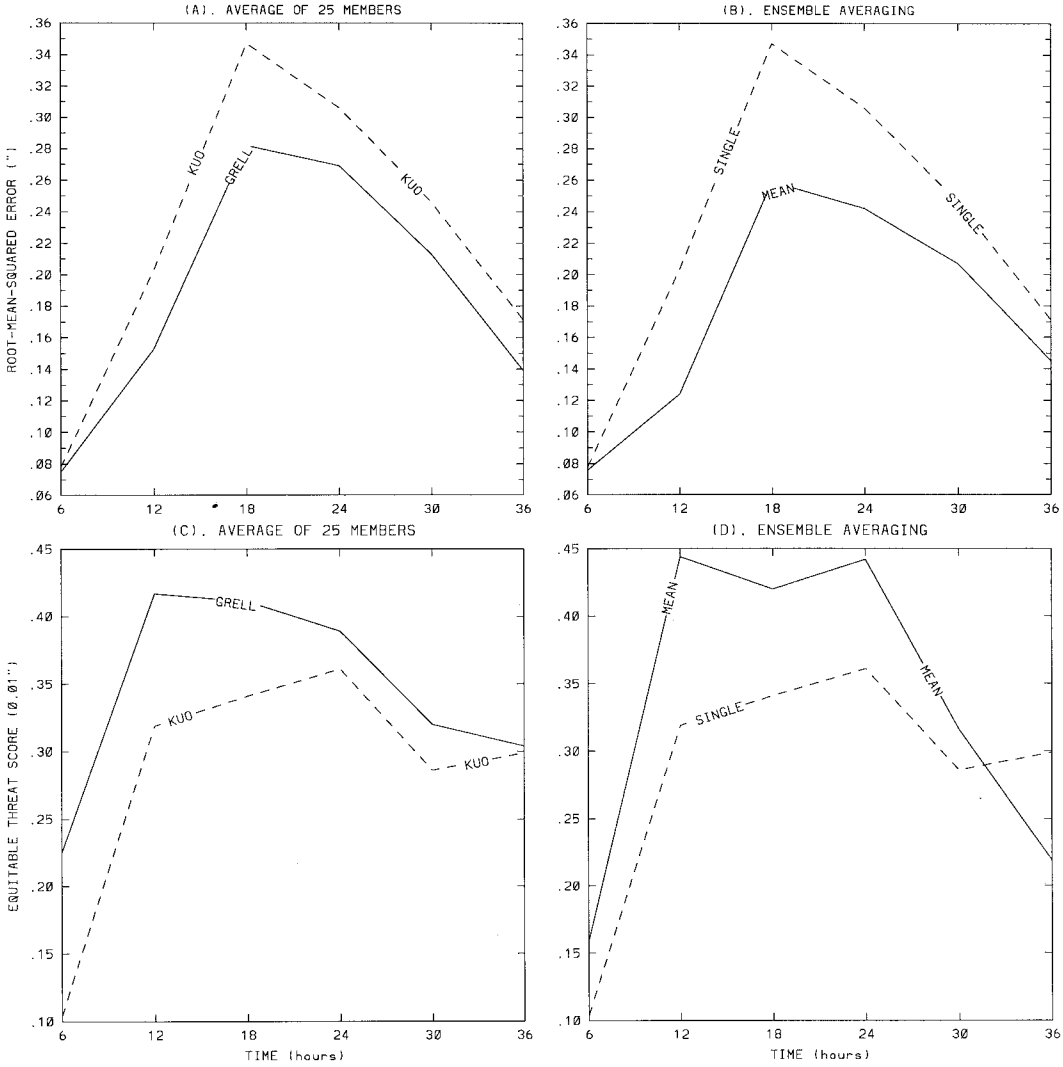

FIG. 26. Comparison of the impact of ensemble averaging with a change in the analysis–forecast system. (a) Average rmse for the 25 individual forecasts with ''Grell'' and ''Kuo'' schemes; (b) average rmse for the 25 ''single'' forecasts and the rmse for the 25-member ensemble ''mean'' forecast with the ''Kuo'' scheme. (c) and (d) As in (a) and (b), respectively, except for the ETS at 0.01″ threshold.

APPENDIX

Verification Scores

a. Bias score

The bias score is the ratio of the forecast area to observed area of precipitation amounts over any given threshold (Anthes 1983). It is defined as

$$\text{BIAS} = \frac{\text{FA}}{\text{OA}}, \tag{A1}$$

where FA is the forecast area and OA is the observed area. Note that this bias score is defined for areal coverage; an analogous bias score could be defined for precipitation amount.

b. Equitable threat score (ETS)

The traditional threat score (TS, e.g., Anthes 1983) measures the accuracy in predicting area of precipitation amounts over any given threshold. The TS is defined

$$\text{TS} = \frac{\text{CFA}}{\text{FA} + \text{OA} - \text{CFA}}, \tag{A2}$$

where CFA is the correctly forecast area bounded by a given precipitation amount, FA is the forecast area, and OA is the observed area. Here, TS value ranges from 0.0 (zero accuracy) to 1.0 (perfect forecast).

The equitable threat score (ETS, Schaefer 1990) negates the reward achieved in TS through increasing the hit rate from just increasing the forecast areal coverage FA. The ETS measures the skill in predicting the area of precipitation amounts over any given threshold with respect to a random (no skill) control forecast and is defined

$$\text{ETS} = \frac{\text{CFA} - \text{CHA}}{\text{FA} + \text{OA} - \text{CFA} - \text{CHA}}, \tag{A3}$$

where the CHA term denotes the expected area of hits in a random forecast of FA area given OA area or

$$\begin{aligned}\text{CHA} &= (\text{probability of correctly hitting a unit area by chance}) \times (\text{observed area}) \\ &= \frac{\text{FA}}{\text{VA}} \times \text{OA}\end{aligned} \tag{A4}$$

with VA the verification area. The accuracy of a forecast is directly proportional to the ETS value. A perfect forecast has an ETS = 1, while an ETS $>$ 0.0 denotes a skillful forecast relative to a random forecast. When ETS $\leq$ 0.0, a forecast has no skill.

c. Ranked probability score (RPS)

The ranked probability score (RPS) is a scoring rule for evaluating categorical, probabilistic forecasts at a grid point or a station. It was first proposed by Epstein (1969) and simplified by Murphy (1971). For J MECE categories, the RPS can be written:

$$\text{RPS}(\mathbf{r}, \mathbf{d}) = \sum_{i=1}^{J} \left(\sum_{k=1}^{i} r_k - \sum_{k=1}^{i} d_k \right)^2, \tag{A5}$$

where the vector $\mathbf{r} = (r_1, \ldots, r_k)$ ($r_k \geq 0$ and $\Sigma_{k=1}^{J} r_k = 1$) represents the forecast probability distribution and the vector $\mathbf{d} = (d_1, \ldots, d_k)$ (d_k equals 1 if class k occurs and zero otherwise) represents the observation; $\Sigma_{k=1}^{i} r_k$ and $\Sigma_{k=1}^{i} K_k$ describe the forecast and the observed cumulative probabilities, respectively. Thus, the RPS represents the sum of the squares of the differences between the forecast and the observed cumulative probabilities.

The RPS is inversely proportional to the accuracy of a probabilistic, categorical forecast. An RPS of zero denotes a perfect forecast, that is, the forecast probability is 1 in the correct category. The maximum value of the RPS is $J - 1$.

REFERENCES

Anthes, R. A., 1977: A cumulus parameterization scheme utilizing a one-dimensional cloud model. *Mon. Wea. Rev.*, **105,** 270–286.

——, 1983: Regional models of the atmosphere in middle latitudes. *Mon. Wea. Rev.*, **111,** 1306–1335.

——, 1986: The general question of predictability. *Mesoscale Meteorology and Forecasting*, P. S. Ray, Ed., Amer. Meteor. Soc., 636–656.

——, and T. T. Warner, 1978: Development of hydrostatic models suitable for air pollution and other mesometeorological studies. *Mon. Wea. Rev.*, **106,** 1045–1078.

——, E.-Y. Hsie, and Y.-H. Kuo, 1987: Description of the Penn State/NCAR Mesoscale Model Version 4 (MM4). NCAR Tech. Note NCAR/TN 282+STR, 66 pp. [Available from NCAR, P.O. Box 3000, Boulder, CO 80303.]

Arakawa, A., and W. H. Schubert, 1974: Interaction of a cumulus cloud ensemble with the large-scale environment. Part I. *J. Atmos. Sci.*, **31,** 674–701.

——, and V. R. Lamb, 1977: Computational design of the basic dynamical process of the UCLA general circulation model. *Methods Comput. Phys.*, **17,** 173–265.

Augustine S. J., S. L. Mullen, and D. P. Baumhefner, 1991: Examination of actual analysis differences for use in Monte Carlo forecasting. Preprints, *16th Annual Climate Diagnostics Workshop*, Los Angeles, CA, NOAA/NWS/NMC/CAC, 375–378.

Barnes, S. L., 1964: A technique for maximizing details in numerical weather map analyses. *J. Appl. Meteor.*, **3,** 396–409.

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

Baumhefner, D. P., 1984: Analysis and forecast intercomparisons using the FGGE SOP1 data base. *Proceedings of the First National Workshop on the Global Weather Experiment*, Vol. 2, Part 1, National Academy Press, 228–246.

——, and D. J. Perkey, 1982: Evaluation of lateral boundary errors in a limited-domain model. *Tellus*, **34,** 409–428.

Blackadar, A. K., 1979: High resolution models of the planetary boundary layer. *Adv. Environ. Sci. Eng.*, **1**(1), 50–85.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78,** 1–3.

Brooks, H. E., and C. A. Doswell III, 1993: New technology and numerical weather prediction wasted opportunity? *Weather*, **48,** 173–177.

——, M. S. Tracton, D. J. Stensrud, G. J. DiMego, and Z. Toth, 1995: Short-range ensemble forecasting (SREF): Report from a workshop. *Bull. Amer. Meteor. Soc.*, **76,** 1617–1624.

Daley, R., and T. Mayer, 1986: Estimates of global analysis error from the global weather experiment observational network. *Mon. Wea. Rev.*, **114,** 1642–1653.

EarthInfo, Inc., 1990: Climatedata—Hourly precipitation over the United States. EarthInfo, Inc.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.,* **8,** 985–987.

Errico, R. M., 1983: A guide to transform software for non-linear normal-mode initialization of the NCAR Community Forecast Model. NCAR Tech. Note NCAR/TN-217+IA. 86 pp. [Available from NCAR, P.O. Box 3000, Boulder, CO 80303.]

——, and D. P. Baumhefner, 1987: Predictability experiments using a high-resolution and limited-area model. *Mon. Wea. Rev.,* **115,** 488–504.

Grell, G., Y.-H. Kuo, and R. Pasch, 1991: Semiprognostic tests of cumulus parameterization schemes in the middle latitudes. *Mon. Wea. Rev.,* **119,** 5–31.

Grumm, R. H., 1993: Characteristics of surface cyclone forecasts in the aviation run of the global spectral model. *Wea. Forecasting,* **8,** 87–112.

Hamill, T. M., and S. J. Colucci, 1996: Random and systematic error in NMC's ETA short-range ETA ensembles. Preprints, *13th Conf. on Probabililty and Statistics in the Atmospheric Sciences,* San Francisco, CA, Amer. Meteor. Soc., 51–56.

——, and ——, 1997: Verification of Eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.,* **125,** 1312–1327.

Hoffman, R. N., and E. Kalnay, 1983: Lagged-average forecasting, an alternative to Monte Carlo forecasting. *Tellus,* **35,** 100–118.

Hoke, J. E., N. A. Phillips, G. J. DiMego, J. J. Tucillo, and J. G. Sela, 1989: The regional analysis and forecast system of the National Meterological Center. *Wea. Forecasting,* **4,** 323–334.

Holloway, J. L., Jr., 1958: Smoothing and filtering of time series and space fields. *Advances in Geophysics,* Vol. 4, Academic Press, 351–389.

Hsie, E.-Y., R. A. Anthes, and D. Keyser, 1984: Numerical simulation of frontogenesis in a moist atmosphere. *J. Atmos. Sci.,* **41,** 2581–2594.

Kallen, E., and X.-Y. Huang, 1988: The influence of isolated observations on short-range numerical weather forecasts. *Tellus,* **40A,** 324–336.

Kuo, H. L., 1974: Further studies of the parameterization of the effect of cumulus convection on large-scale flow. *J. Atmos. Sci.,* **31,** 1232–1240.

Kuo, Y.-H., and S. Low-Nam, 1990: Prediction of nine explosive cyclones over the western Atlantic with a regional model. *Mon. Wea. Rev.,* **118,** 3–25.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.,* **102,** 409–418.

Lindzen, R. S., and M. Fox-Rabinowitz, 1989: Consistent vertical and horizontal resolution. *Mon. Wea. Rev.,* **117,** 2575–2583.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.,* **20,** 130–141.

Manning, K. W., and P. L. Haagenson, 1992: Data ingest and objectives analysis for the PSU/NCAR modeling system: Programs DATAGRID and RAWINS. NCAR. Tech. Note NCAR/TN 376+STR+IA, 209 pp. [Available from NCAR, P.O. Box 3000, Boulder, CO 80303.]

Mass, C. F., and D. M. Schultz, 1993: The structure and evolution of a simulated midlatitude cyclone over land. *Mon. Wea. Rev.,* **121,** 889–917.

Molteni, F., T. N. Palmer, R. Buizza, and T. Pertroliagis, 1996: The ECMWF ensemble prediction system: Methodology and verification. *Quart. J. Roy. Meteor. Soc.,* **122,** 73–121.

Mullen, S. L., and D. P. Baumhefner, 1989: The impact of initial condition uncertainty on numerical simulations of large scale explosive cyclogenesis. *Mon. Wea. Rev.,* **117,** 2289–2329.

——, and ——, 1991: Monte Carlo simulations of explosive cyclogenesis using a low-resolution, global spectral model. Preprints, *Ninth Conf. on Numerical Weather Prediction,* Denver, CO, Amer. Meteor. Soc., 750–751.

——, and ——, 1994: Monte Carlo simulations of explosive cyclogenesis. *Mon. Wea. Rev.,* **122,** 1548–1567.

——, and J. Du, 1994: Monte Carlo forecasts of explosive cyclogenesis with a limited-area, mesoscale model. Preprints, *10th Conf. on Numerical Weather Prediction,* Portland, OR, Amer. Meteor. Soc., 638–640.

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

Murphy, A. H., 1971: A note on the ranked probability score. *J. Appl. Meteor.,* **10,** 155–156.

Murphy, J. M., 1988: The impact of ensemble forecasts on predictability. *Quart. J. Roy. Meteor. Soc.,* **114,** 463–494.

NOAA, 1987: Monthly relative frequencies of precipitation for the United States for 6-, 12-, and 24-h periods. NOAA Tech. Rep. No. 39, NWS/NOAA/US Department of Commerce, 262 pp.

NOAA/NCDC, 1987: *Storm Data.* Vol. 29, No. 12, National Oceanic and Atmospheric Administration/National Climate Data Center, 47 pp.

Palmer, T. N., R. Mureau, and F. Molteni, 1990: The Monte Carlo forecast. *Weather,* **45,** 198–207.

Persson, P. O. G., and T. T. Warner, 1991: Model generation of spurious gravity waves due to inconsistency of the vertical and horizontal resolution. *Mon. Wea. Rev.,* **119,** 917–935.

Powers, J. G., and R. J. Reed, 1993: Numerical simulation of the large-amplitude mesoscale gravity-wave event of 15 December 1987 in central United States. *Mon. Wea. Rev.,* **121,** 2285–2308.

Prager, T., T. Vukicevic, J.-N. Thepaut, J.-F. Louis, P. Gauthier, R. Errico, J. Derber, and P. Courtier, 1995: Second workshop on adjoint applications in dynamic meteorology, Visegrad, Hungary, 2–6 May 1994. *Bull. Amer. Meteor. Soc.,* **76,** 375–379.

Roebber, P. J., 1990: Variability in successive operational model forecasts of maritime cyclogenesis. *Wea. Forecasting,* **5,** 586–595.

——, 1993: A case study of self-development as an antecedent conditioning process in explosive cyclogenesis. *Mon. Wea. Rev.,* **121,** 976–1006.

Sanders, F., 1971: Analytic solutions of the non-linear omega and vorticity equations for a structurally simple model of disturbances in the baroclinic westerlies. *Mon. Wea. Rev.,* **99,** 393–408.

——, 1986: Trends in skill of Boston forecasts made at MIT 1966–84. *Bull. Amer. Meteor. Soc.,* **67,** 170–176.

——, 1992: Skill of operational models in cyclone prediction out to five-days during ERICA. *Wea. Forecasting,* **7,** 3–25.

——, and J. R. Gyakum, 1980: Synoptic-dynamic climatology of the ''bomb.'' *Mon. Wea. Rev.,* **108,** 1589–1606.

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

Schneider, R. S., 1990: Large-amplitude mesoscale wave disturbances with the intense Midwestern extratropical cyclone of 15 December 1987. *Wea. Forecasting,* **5,** 533–558.

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

Stensrud, D. J., and J. M. Fritsch, 1994a: Mesoscale convective systems in weakly forced large-scale environment. Part II: Generation of a mesoscale initial condition. *Mon. Wea. Rev.,* **122,** 2068–2083.

——, and ——, 1994b: Mesoscale convective systems in weakly forced large-scale environment. Part III: Numerical simulations and implications for operational forecasting. *Mon. Wea. Rev.,* **122,** 2084–2104.

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

Tracton, M. S., and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting,* **8,** 378–398.

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

Williamson, D. L., J. T. Kiehl, V. Ramanathan, R. E. Dickinson, and J. J. Hack, 1987: Description of the NCAR Community Climate Model (CCM1). NCAR Tech. Note NCAR/TN-285+STR, 112 pp. [Available from NCAR, P.O. Box 3000, Boulder, CO 80303.]

Zhang, D.-L., E.-Y. Hsie, and M. W. Moncrieff, 1988: A comparison of explicit and implicit prediction of convective and stratiforrn precipitating weather systems with a meso-β scale numerical model. *Quart. J. Roy. Meteor. Soc.,* **114,** 31–60.