Evaluation of a Short-Range Multimodel Ensemble System

MATTHEW S. WANDISHIN AND STEVEN L. MULLEN

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

DAVID J. STENSRUD AND HAROLD E. BROOKS

NOAA/ERL/National Severe Storms Laboratory, Norman, Oklahoma

(Manuscript received 2 August 1999, in final form 7 September 2000)

ABSTRACT

Forecasts from the National Centers for Environmental Prediction's experimental short-range ensemble system are examined and compared with a single run from a higher-resolution model using similar computational resources. The ensemble consists of five members from the Regional Spectral Model and 10 members from the 80-km Eta Model, with both in-house analyses and bred perturbations used as initial conditions. This configuration allows for a comparison of the two models and the two perturbation strategies, as well as a preliminary investigation of the relative merits of mixed-model, mixed-perturbation ensemble systems. The ensemble is also used to estimate the short-range predictability limits of forecasts of precipitation and fields relevant to the forecast of precipitation.

Whereas error growth curves for the ensemble and its subgroups are in relative agreement with previous work for large-scale fields such as 500-mb heights, little or no error growth is found for fields of mesoscale interest, such as convective indices and precipitation. The difference in growth rates among the ensemble subgroups illustrates the role of both initial perturbation strategy and model formulation in creating ensemble dispersion. However, increase spread per se is not necessarily beneficial, as is indicated by the fact that the ensemble subgroup with the greatest spread is less skillful than the subgroup with the least spread.

Further examination into the skill of the ensemble system for forecasts of precipitation shows the advantage gained from a mixed-model strategy, such that even the inclusion of the less skillful Regional Spectral Model members improves ensemble performance. For some aspects of forecast performance, even ensemble configurations with as few as five members are shown to significantly outperform the 29-km Meso-Eta Model.

1. Introduction

Deficiencies in observational networks and data assimilation methods produce initial condition (IC) errors that ensure that a numerical model never starts from the same state as the atmosphere. Due to the chaotic nature of atmospheric flow, these IC errors grow unpredictably with time (Lorenz 1963, 1965, 1982). Eventually the errors in the model solution saturate and the forecast bears no more resemblance to the true state of the atmosphere than two analyses for different days picked at random (Lorenz 1963, 1965, 1982). Thus, IC error ultimately places a limit on the expected range of a useful numerical forecast, even under the assumption that the model is perfect. Model deficiencies from such sources as the physical parameterizations and numerical truncations, necessitated by an incomplete understanding of atmospheric processes and limited computational resources, reduce the limit of useful deterministic forecasts even further.

One way to deal with this inherent uncertainty is through ensemble forecasting. Typically, a single model is run repeatedly from a set of slightly different, equally viable ICs, and each resulting forecast solution is considered equally likely. One can then account for model uncertainty by running the model with different physical parameterizations or even taking forecasts from completely different numerical models. The ensemble mean of these different solutions yields a forecast more skillful than the individual ensemble members (Leith 1974), while the ensemble dispersion provides quantitative information on forecast uncertainty (e.g., Tracton and Kalnay 1993). This latter aspect of ensemble forecasting demands emphasis. As much as users and forecasters desire the simplicity of deterministic forecasts, such forecasts are clearly inferior in terms of quality or value to end users. In fact, neglecting uncertainty in the forecast process generally results in extreme reductions in both quality and value (Murphy 1993).

While medium-range (5–15 days) ensemble forecast

Corresponding author address: Matthew S. Wandishin, NSSL, 1313 Halley Circle, Norman, OK 73069.
E-mail: mwand@vicksburg.nssl.noaa.gov

systems have been run operationally since December 1992 (Tracton and Kalnay 1993; Molteni et al. 1996), the implementation of short-range (0–48 h) ensembles has lagged behind, along with research into the issues involved in this development. One of these issues is mesoscale predictability. If an attempt is to be made to forecast a particular phenomenon or weather event, it is important first to determine the limits imposed by the chaotic nature of the atmosphere on the potential skill of such forecasts and the length of time for which a skillful forecast is possible. Determination of predictability limits helps in assessing the current level of skill and the potential improvement that can be expected. Early estimates of predictability limits for the mesoscale were based on case studies with model resolutions much coarser than today's operational models (e.g., Anthes 1986; Errico and Baumhefner 1987; Vukicevic and Errico 1990). Computational advances now allow more sophisticated models to be run on larger domains at finer resolutions since those studies nearly a decade ago. These advances, combined with the success of the medium-range ensembles, sparked the creation in 1994 of an experimental program at the National Centers for Environmental Prediction (NCEP) designed to test the utility of short-range ensemble forecasting (SREF; Brooks et al. 1995).

One recommendation to emerge from the SREF workshop is the need to learn more about the predictability of different model fields (Brooks et al. 1995). The expected outcomes from finescale short-range modeling, namely accurate information on the details of sensible weather elements, are somewhat different than for medium-range modeling, where the focus tends more on planetary-scale flow regimes, synoptic wave patterns, and cyclone positions. Thus estimates of predictability limits for parameters other than geopotential height or temperature, which have been the focus of many previous studies of predictability (Lorenz 1982; Dalcher and Kalnay 1987; Errico and Baumhefner 1987; Tribbia and Baumhefner 1988; Molteni and Palmer 1993; Molteni et al. 1996; Buizza 1997), are desired. There is reason to expect that smaller-scale features, precipitation, and mechanisms that strongly modulate precipitation may have shorter predictability limits than the synoptic and planetary fields that are often examined in this context (e.g., Lorenz 1982; Baumhefner 1984; Anthes 1986; Stamus et al. 1992; Mullen et al. 1999).

Du et al. (1997) examined the impact of IC errors on quantitative precipitation forecasts for a case of strong cyclogenesis with two 25-member mesoscale ensembles. They found that 90% of the reduction in root-mean-square error afforded by ensemble averaging could be realized with as few as 8–10 members. However, ensemble averaging can have both positive and negative effects on traditional measures of forecast quality, such as the bias or equitable threat score. The improvement over climatology of probabilistic quantitative precipitation forecasts (PQPFs) for their ensemble was greater than the improvements gained from a single run of the model at twice the horizontal resolution.

Prior studies based on the data from the experimental NCEP SREF program (Hamill and Colucci 1997, 1998; Tracton et al. 1998; Stensrud et al. 1999) found that the spread of the ensemble members can become significant within a short period of time. However, in general the ensembles suffered from underdispersion; that is, the envelope of solutions generated by the ensemble is not, in all cases, sufficiently large to encompass reality. This underdispersion is hypothesized to be a major cause for the low degree of correlation between the ensemble spread and accuracy of the mean forecast (Hamill and Colucci 1998; Stensrud et al. 1999), that is, an inability of the ensemble to forecast the forecast skill of the ensemble mean. Despite these shortcomings, the ensembles possessed skill equal to or better than a single, higher-resolution run of the model for forecasts of precipitation, cyclone position, and mandatory level data. Furthermore, PQPFs derived from the ensemble were shown to be more skillful than the PQPFs for model output statistics from forecasts of the Nested Grid Model (Hamill and Colucci 1998).

In the present paper, the NCEP ensemble used by Hamill and Colucci (1997, 1998) and Stensrud et al. (1999) is employed primarily to address two areas: 1) to estimate the short-range predictability limits of forecasts of precipitation and fields relevant to the forecast of precipitation, and (2) to assess the quality of short-range ensemble forecasts of precipitation relative to forecasts from a single run of a higher-resolution model using similar computational resources. In addition, the data from the NCEP SREF project allows for the comparison of different ensemble methodologies.

This paper is organized as follows: section 2 describes the configurations of the various ensemble systems and verification procedures. Estimates of predictability limits for precipitation and convective indices are presented in section 3. The skill of various ensemble configurations in forecasting precipitation is examined in section 4. Section 5 presents a summary and discussion of results, conclusions and recommendations for future research.

2. Methodology

a. Description of ensemble forecast systems

The model forecasts of this study come from the experimental ensemble configuration assembled by the NCEP and run periodically between September 1994 and December 1997. The reader is referred to Hamill and Colucci (1997, 1998) and Stensrud et al. (1999) for details on the configuration of the ensemble. We only present a brief overview of the forecast system in this section.

The complete ensemble (FULL) consists of 10 members of the 80-km Eta Model (Janjić 1994) and five

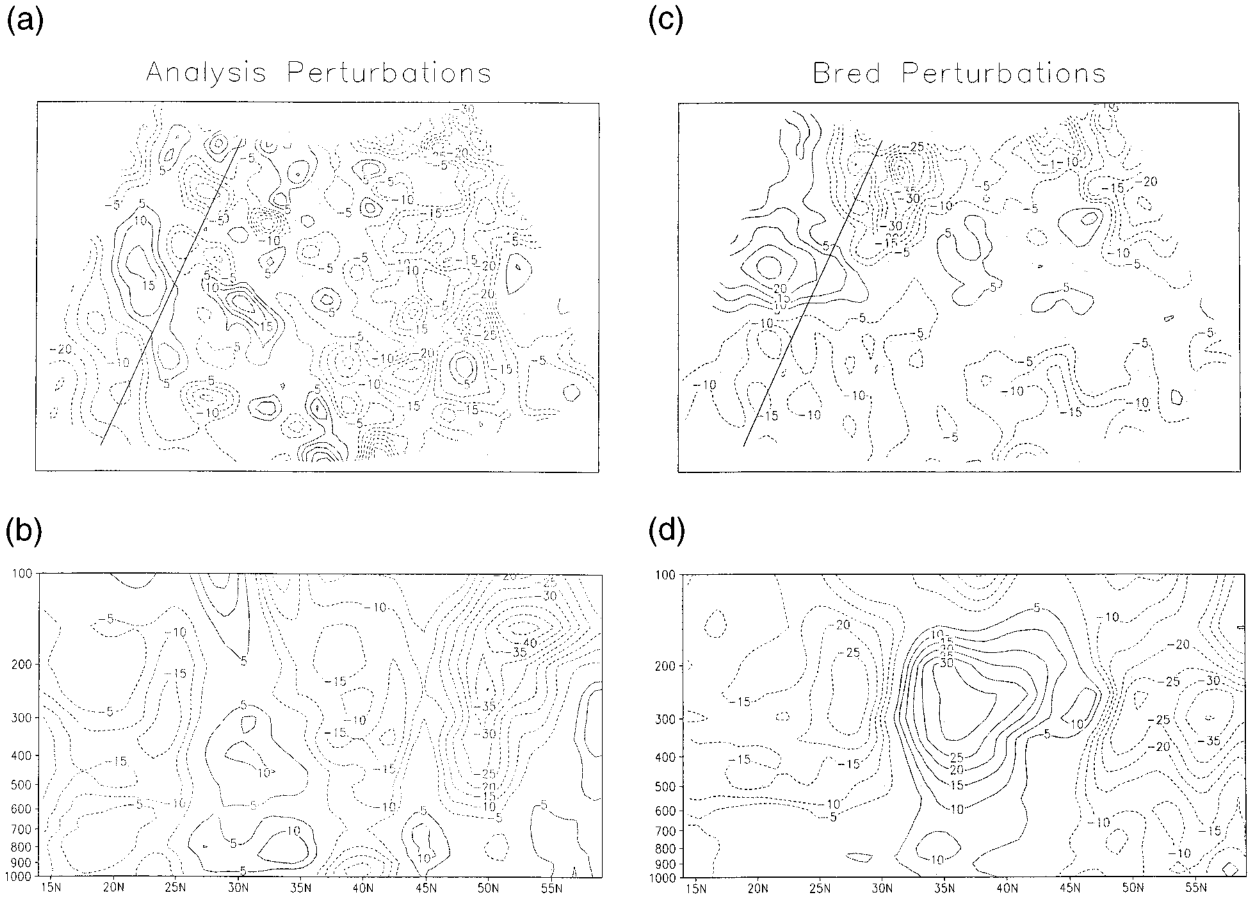

FIG. 1. Horizontal (500 mb) and vertical (along 125°W) view of height analysis differences (contour interval 5 dm, negative dashed, 0 contour omitted) at 1200 UTC 22 Jul 1997 for (a), (b) in-house analysis member and (c), (d) bred perturbation member.

members from the 80-km Regional Spectral Model (Juang and Kanamitsu 1994). The ICs for five of the Eta members (ANL) are interpolated from different in-house analyses, while the ICs for the remaining Eta members (BRD) are interpolated from the Medium Range Forecast (MRF) model control run (Parrish and Derber 1992) along with two positive and two negative bred perturbations (Toth and Kalnay 1993). The Regional Spectral Model (RSM) members are initialized from the same five bred ICs as the Eta members. An example of typical analysis and bred perturbation structure, which shows the smaller-scale structures apparent in the typical analysis differences as compared to the bred modes, is given in Fig. 1 (see also Errico and Baumhefner 1987). To allow for comparison of the attributes of the ensemble forecasts and higher-resolution deterministic forecasts, single runs of the 29-km Meso-Eta Model (MESO; Black 1994) are also examined. The different models share the same output domain, which contains the entire contiguous United States. There are 43 ensemble cases between September 1995 and December 1997 for which all of the relevant model fields are available. Of these 43 cases, 27 occur in the cool season (Nov–Apr) and 16 in the warm season (May–Oct).

To supplement estimates of predictability limits from the NCEP model, a 10-member ensemble is also produced from the Pennsylvania State University–National Center for Atmospheric Research fifth-generation Mesoscale Model (MM5; Dudhia 1993). The MM5 model is run to 36 h for 27 cases from October through December 1997. The nested MM5 configuration has an outermost grid domain of 80 × 45 grid points and horizontal grid spacing of 96 km, and an interior grid domain of 106 × 76 grid points and horizontal grid spacing of 32 km (Fig. 2). Both domains have the same 23 vertical sigma layers. The ensemble attempts to reflect the combined uncertainty of both initial analyses and model formulations. Initial analyses for the MM5 are provided by the 0000 UTC run of the Eta Model. Added to the initial analyses are randomly generated perturbations that contain amplitudes and scale-dependent, spatially correlated structures consistent with prior estimates of analysis uncertainty by Daley and Mayer (1986). More thorough descriptions of the scheme and examples of the perturbation structures can be found in

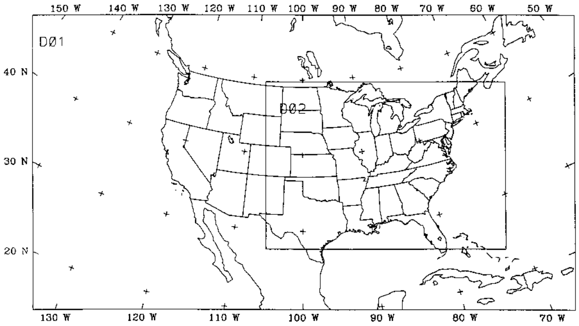

FIG. 2. Map of the MM5 outer domain (D01) at 96-km grid spacing and the model inner domain (DO2) at 32-km grid spacing.

TABLE 1. Model configurations used in the MM5 ensemble. Note that the Kain–Fritsch convective scheme in member 4 also includes delayed downdrafts (Spencer and Stensrud 1998).

Member no.	Moisture availability	PBL scheme	Convective scheme
1	Control	Blackadar	Kain–Fritsch
2	Control	Blackadar	Betts–Miller
3	Control	Burk–Thompson	Grell
4	Control	Burk–Thompson	Kain–Fritsch
5	Low	Burk–Thompson	Kain–Fritsch
6	Low	Blackadar	Grell
7	Low	Burk–Thompson	Betts–Miller
8	High	Blackadar	Kain–Fritsch
9	High	Burk–Thompson	Grell
10	High	Blackadar	Betts–Miller

Errico and Baumhefner (1987), Mullen and Baumhefner (1988, 1989), and Du et al. (1997).

Model uncertainty in the MM5 ensemble is sampled through the combination of different physical parameterization schemes. Specifically, the Kain–Fritsch (Kain and Fritsch 1990), Betts–Miller (Betts and Miller 1986), and Grell (Grell 1993) convective parameterization schemes are combined with the Blackadar (1976, 1979) and Burk–Thompson (Burk and Thompson 1989) planetary boundary layer schemes and three configurations of the MM5 moisture availability parameter (Anthes et al. 1987). The exact configuration for the 10 members is given Table 1.

b. Verification data and analyses

The data used for precipitation verification are 24-h accumulations assembled by the National Weather Service River Forecast Centers (RFCs). The precipitation data from roughly 5000 stations are placed onto the same 80-km grid used by the Eta Model through a ‘‘box averaging’’ technique in which observations are assigned to the nearest grid box; the gridded value is then the average of all observations within each box (Mesinger 1996, p. 2638). The same box-averaging technique is used to regrid the Meso-Eta to 80-km resolution before computing the verification statistics. The data for the verification of precipitation forecasts are available for 19 cases between October 1996 and December 1997, of which 10 cases occur in the warm season and 9 in the cool season.

Twenty-four-hour precipitation totals from the National Weather Service Summary of the Day archived by the National Climate Data Center are used to determine the climatological frequency of precipitation occurrence for various accumulations. Accumulations, accurate to 0.254 mm, are obtained for 353 stations in the contiguous United States. The period of record varies from station to station, but it exceeds 25 yr for all stations. For each station, the frequency of the occurrence of accumulations of at least 0.254, 2.54, 12.7, and 25.4 mm are computed along with the frequency of days for which there was no measurable precipitation. The box-averaging technique is not used to grid the climatological frequencies since a large number of the boxes do not contain a station, especially in the western United States. Therefore, fields of precipitation frequency are generated on the 80-km NCEP model grids for each of the four thresholds using a Cressman analysis scheme (Cressman 1959) with successive search radii of 5, 4, 3, 2, 1, and 0.5 grid points. For each verification time, the climatological frequency is defined as a 31-day mean centered on the verification day.

c. Accuracy measures and skill scores

Three scores are used in this paper to assess the accuracy and skill of precipitation forecasts: the equitable threat score (ETS), the ranked probability score (RPS), and the relative operating characteristic (ROC).

The ETS is commonly used by NCEP to evaluate gridpoint, precipitation forecasts. The ETS gives the skill in predicting the forecast area of precipitation at least equal to the threshold relative to the probability of achieving a correct forecast by chance (Schaefer 1990). An ETS $=0$ indicates a no-skill forecast, an ETS >0 denotes a forecast skillful relative to chance, while an ETS $=1$ indicates a perfect forecast. A shortcoming of the ETS, as discussed by Mason (1989) and Hamill (1999), is that the score is sensitive to model bias, with the wetter of two forecasts tending to have a higher ETS than if it was adjusted to match the bias of the drier forecast. Also, much of the information from ensembles is lost as the forecasts must be reduced to ensemble means for computation of the ETS.

A suitable measure of the accuracy for probabilistic categorical forecasts is the ranked probability score (Epstein 1969; Murphy 1971; Wilks 1995), which compares the cumulative distribution functions of the forecast and verification. The RPS is negatively oriented: an RPS $=0$ denotes a perfect forecast with the worst possible score being $J - 1$, where J is the number of mutually exclusive, collectively exhaustive categories. The categories used here follow the thresholds for determining the ETS: no measurable precipitation (pp $<$ 0.254 mm), 0.254 $\leq$

Table 2. Two-by-two contingency table for forecast and occurrence of binary events.

Event		Observed		
		Yes	No	Total
Forcast	Yes	*a*	*b*	*a* + *b*
	No	*c*	*d*	*c* + *d*
	Total	*a* + *c*	*b* + *d*	

pp < 2.54 mm, $2.54 \leq$ pp < 12.7 mm, $12.7 \leq$ pp $<$ 25.4 mm, and 25.4 mm $\leq$ pp. The ranked probability skill score (RPSS; Wilks 1995) measures the improvement of the RPS relative to some reference forecast. The reference forecasts used in this study are the climatological frequencies described in the previous subsection. Positive RPSS values denote a skillful forecast with respect to climatology, while an RPSS $=1$ represents a perfect forecast. The RPS and RPSS are computed at each model grid point and then averaged over the entire domain for which grid points with verification data exist.

A useful measure for assessing the ability of probabilistic forecasts to discriminate dichotomous events, and which has been proposed as suitable for comparing probabilistic and deterministic systems, is the relative operating characteristic (Swets 1973; Mason 1982). The ROC utilizes information from a 2 × 2 contingency table (Table 2) in which correct forecasts can be classified as ‘‘hits,’’ *a,* or ‘‘correct rejections,’’ *d,* and incorrect forecasts as ‘‘misses,’’ *b,* or ‘‘false alarms,’’ *c.* From the contingency table can be defined a hit rate (HR), which equals the proportion of events that are correctly forecast [$a/(a + c)$], and a false alarm rate (FAR), which equals the proportion of nonevents that are forecast as events [$b/(b + d)$]. For a probabilistic forecast, a set of contingency tables can be constructed using different probabilities as decision criteria, ranging from 0% (i.e., an event is always forecast) to 100% (i.e., an event is never forecast). Plotting HR against FAR over the range of decision criteria forms the ROC. A perfect forecast system yields an HR $=1$ and FAR $=0$, represented by the upper-left corner of the graph. A completely unskillful forecast system is unable to distinguish between event occurrence and nonoccurrence (HR = FAR), and thus the ROC lies along the diagonal from the point (0, 0) to (1, 1). Therefore an ROC in the upper-left half of the graph indicates a skillful forecast system.

Forecast quality is often summarized in terms of the area (A) under the ROC, with a skillful system having $A > 0.5$. Mason (1982) presents two methods for calculating this area. The first is through use of the trapezoid rule on the discrete data that come directly from experimental results. The area derived from the trapezoid method is necessarily dependent on the number of data points (i.e., the number of the ensemble members). The second method (often referred to as A_Z) involves fitting a line to the data transformed into normal deviate space. For further details on this method see Mason (1982), Harvey et al. (1992), and Richardson (2000). Transforming this fitted line back to linear probability scales yields a continuous curve, the area beneath which is no longer dependent on the number of decision criteria. The transformed curve is consistent with an assumption that the ensemble size is infinite and the interval between decision criteria dp is infinitesimal, dp $\rightarrow 0$ (Richardson 2000). Bamber (1975) notes, however, that the discrete ROC curve is meaningful completely apart from its being an approximation to the continuous curve in that it indicates the extent to which the model (or ensemble) is able to distinguish between events and nonevents. Furthermore, whereas A_Z, as derived from the continuous, fitted ROC curve, is a measure of the ensemble's discrimination capacity, A, as calculated by the discrete, trapezoid method, is a measure of the ensemble's discrimination performance (Bamber 1975).

A deterministic system can be compared with a probabilistic one either by visual inspection by plotting the HR and FAR of the former and noting whether it falls above or below the ROC for the probabilistic system (i.e., does the ensemble have a higher or lower HR for the given FAR of the deterministic model), or by comparing the area under the curve for the deterministic forecast that goes through its (HR, FAR) point and terminates at the default points (1, 1) and (0, 0). The area beneath the ROC curve for a deterministic system can be calculated only by a discrete method, since the single point is insufficient for fitting a line in the normal deviate space. One would expect, therefore, that the area for deterministic system would suffer in comparison to areas calculated from ensemble systems, but, as noted above, this reflects the additional ability to discriminate between rain events and nonevents associated with having multiple decision criteria.

d. Statistical significance testing

The Wilcoxon signed-rank test (Wilks 1995) is used to ascertain the differences in skill scores between the different ensemble groups that are significant at the 5% confidence level. The 15 ensemble members are grouped into seven pairings and compared as follows:

MESO-FULL to evaluate the skill of a mixed ensemble system with 15 members versus high-resolution deterministic forecasts;

MESO-OPNL (the operational 80-km Eta) to evaluate the impact of model resolution on the skill of deterministic forecasts;

BRD-ANL to evaluate the impact of different perturbation strategies on forecasts by ensemble systems with the same model configuration;

BRD-RSM to evaluate the impact of different model formulations on forecasts by ensemble systems with the same initial perturbations and lateral boundary conditions;

ETA-PHYS-MIX (three unique pairings), where the ETA is a 10-member ensemble composed of the BRD and ANL members, PHYS combines the BRD and RSM members, and MIX combines the ANL and RSM members; these 10-member groups are compared to evaluate the impact of a mixed-perturbation ensemble versus a mixed-model ensemble versus an ensemble incorporating both of these strategies, with the same number of ensemble members.

Because comparisons of the ETS from forecasts with different biases can be misleading (Mason 1989), we apply an adjustment for different biases following the recommendation of Hamill (1999) when estimating the significance of difference between ETS values. In order to apply this adjustment, the minimum precipitation threshold is raised to 0.76 mm for calculations of bias and ETS.

3. Predictability estimates

The method of determining predictability in this study follows Lorenz (1965) who looked at the growth rate of the difference between two or more solutions having similar but different initial conditions. Typically, after an initial period of exponential growth, the growth rate slows as the difference approaches a value representative of the magnitude of the difference between two randomly selected states (Lorenz 1982). However, these error growth curves can vary substantially between individual cases, and previous studies with limited-area models found somewhat different behavior, primarily in the form of more subdued growth (e.g., Errico and Baumhefner 1987; Vukicevic and Errico 1990). Output from the 43 ensemble cases is now examined.

a. 500-mb height

To establish a point of reference with previous predictability studies, the root-mean-square (rms) differences for the 500-mb heights from the full ensemble (heavy curve) are shown in Fig. 3. The values for this and subsequent error growth curves are found by computing the average rms difference between a given member taken as the control (or ‘‘truth’’) and each of the other members. For each case day, each member is taken, in turn, as the truth and the average of these values computed. Finally, the rms difference is averaged over all case days. Results indicate that the doubling time for the 500-mb height field is approximately 1.5 days. This doubling time is consistent with that found by Simmons et al. (1995) for forecasts of 500-mb height and somewhat shorter than the 2.5 days found by Lorenz (1982). This reduction in doubling time is consistent with the pessimistic view for extended-range forecasts presented by Lorenz (1982) in which error growth curves derived from models are shown to overestimate the predictability of the atmosphere. According to Lorenz (1982), as the models improve and resolutions become finer, doubling times will decrease as the modeled predictability limits approach those of the real atmosphere. Error growth curves for forecasts of sea level pressure (not shown), another field commonly used in medium-range predictability studies, behave very similarly to the 500-mb height fields with a doubling time of ~1.5 days.

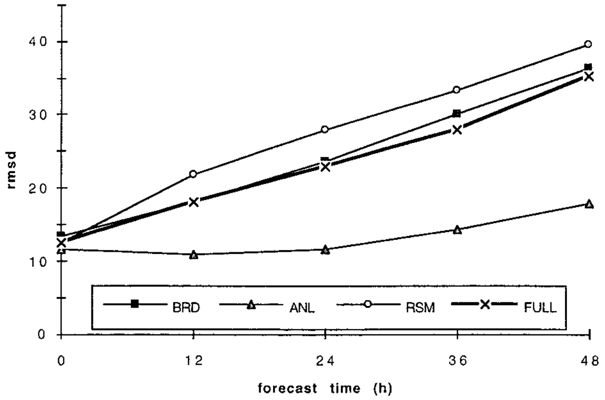

FIG. 3. Rms differences of 500-mb height forecasts. The thick curve is for the complete 15-member ensemble (FULL); thin curves are for the three 5-member subgroups: the Eta runs initialized by in-house analyses (ANL), the Eta runs initialized by interpolated bred modes (BRD), and the Regional Spectral Model runs (RSM).

Of particular interest is the error growth from the different IC perturbation techniques (Fig. 3). These curves are generated in the same manner as for the full ensemble except that only five-member subsets are used to compute the rms differences. The three subgroups are the Eta Model runs that are initialized from the in-house analyses (ANL), the Eta Model runs that are initialized from the four bred modes together with the control run (BRD), and the RSM runs. The error growth of the BRD and RSM subgroups are fairly similar to that of the full ensemble, but the ANL subgroup shows dramatically different behavior. Error growth does not begin until the second 24-h period. In fact, the rms differences decrease slightly in the first 24 h, and the slope during the rest of the period is still less than for the other curves.

Errico and Baumhefner (1987, Fig. 9) show results similar to the behavior of the ANL subgroup. One possible explanation that they cite for initial error decay, or weak growth, is dissipation by the fast-moving inertial–gravity waves. The bred mode technique filters the small-scale, fast modes in the global MRF ensemble (Toth and Kalnay 1993; Houtekamer and Derome 1995), but such unbalanced structures are more prevalent in the in-house analyses (Toth and Kalnay 1993), which use several different models for the assimilation cycle,

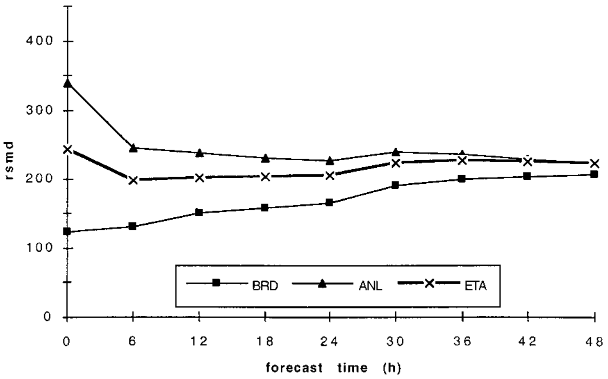

FIG. 4. As in Fig. 3 but for CAPE. The thick curve is for 10 Eta runs (ETA).

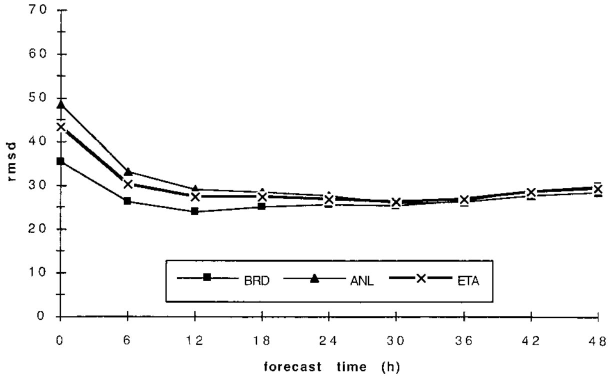

FIG. 5. As in Fig. 4 but for CIN.

with varying degrees of balance, and the initial fields are then transplanted into the Eta Model.

Another fundamental difference between the BRD and ANL configurations lies in the specification of their lateral boundary conditions. The control run of the MRF supplies the boundary conditions for each of the ANL members, whereas the boundary conditions for each of the BRD members are supplied by the corresponding run from the global ensemble forecast. The common lateral boundaries of the ANL members constrain error growth in that differences between members are swept out of the downstream boundaries while identical values are propagated inward from the upstream boundaries (Errico and Baumhefner 1987; Vukicevic and Errico 1990; Warner et al. 1997), leading to reduced dispersion as forecast time increases. For these two reasons, smaller spread between the ANL members than between the BRD members should be anticipated.

b. Convective indices

As discussed previously, predictive information for fields other than just pressure and height is desired from limited-area mesoscale models. For this reason, we computed the error growth curves for two convective indices, convective available potential energy (CAPE) and convective inhibition (CIN), which are shown in Figs. 4 and 5, respectively. CAPE and CIN (e.g., Bluestein 1993, 444–447) are vertically integrated fields that, as measures of instability and stability, respectively, provide guidance on the likelihood and severity of thunderstorms. (These fields were not available for the RSM ensemble members, and so the ETA ensemble is used in place of the FULL ensemble.) The CAPE forecast from the BRD subgroup (Fig. 4) shows weak error growth during the first 24 h (doubling time ~3 days), which slows even further during the final 18 h. The ETA ensemble, however, appears to be dominated by the sharp initial decay in the ANL members that is followed by a complete lack of error growth throughout the forecast period. This pattern of initial decay followed by a flat error growth curves holds for all groups for forecasts of CIN as well (Fig. 5).

A time series of the domain-average precipitation (averaged over all 43 cases) for forecasts from the ANL and BRD groups (Fig. 6) provides insight into the behavior of the CAPE and CIN error growth curves. The convergent nature of the error growth curves, evident especially during the initial 6–12 h, is mirrored in the domain-average precipitation forecasts. The 8% larger value for domain-averaged precipitation for the ANL group at 6 h reflects the ability of the Eta Model's convective scheme to remove the instability associated with the fast modes in the in-house analyses. This in turn results in a rapid reduction of the initial CAPE differences between the in-house members (Fig. 4). The flatness of the error growth curves through the rest of the period may reflect the quasi-equilibrium assumption of the Betts–Miller–Janjić convective scheme of the Eta Model (Betts and Miller 1986; Janjić 1994). The nonlinear saturation level for fields such as CAPE and CIN will be heavily dependent upon the model's convective scheme. With the Betts–Miller–Janjić scheme, this saturation level appears to be about 225 J kg^{-1}. For the ANL ensemble, the removal of the fast modes reduces the initially larger CAPE differences nearly to the sat-

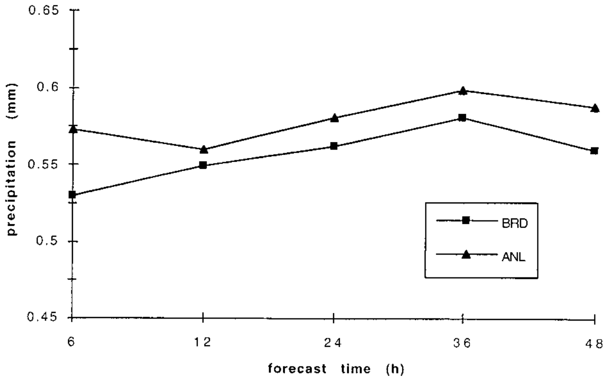

FIG. 6. Domain-averaged precipitation for the ANL and BRD ensemble groups.

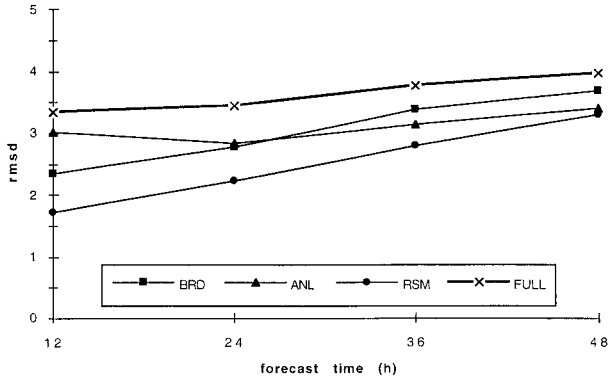

FIG. 7. As in Fig. 3 but for 12-h precipitation accumulations.

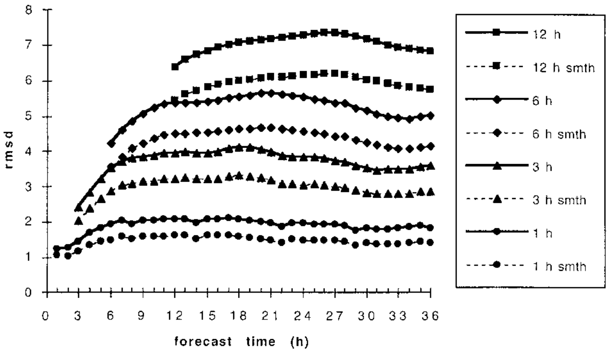

FIG. 8. As in Fig. 7 but for 1-, 3-, 6-, and 12-h accumulations (solid) from a 10-member MM5 ensemble. Dashed curves are for spatially smoothed fields.

uration level and so little further change occurs in the rms differences for this group. The initial rms differences for the BRD ensemble are substantially smaller (~125 J kg^{-1}) and so weak growth is seen through the forecast period as the curve asymptotes toward the saturation level. Therefore, whereas the perturbation strategy may affect how quickly error growth curves become nonlinear or reach saturation, the influence of the perturbation strategy becomes less important relative to a model's variability characteristics once nonlinear error growth begins. Activation-based schemes, such as the Kain–Fritsch (Kain and Fritsch 1990) cumulus parameterization, might behave quite differently. Curiously, the domain-averaged precipitation for the ANL group remains ~4% larger than for the BRD members from 18 to 48 h. This result suggests that the different perturbation strategies could affect forecasts of rainfall beyond the initial adjustment period for reasons that we do not yet fully understand.

c. Precipitation

Involving at the bare minimum delicate interactions between instability, moisture, and lift (Doswell 1987), precipitation is one of the more complex atmospheric processes, and thus it may be one of the more unpredictable. Unfortunately, accurate precipitation forecasts are also perhaps the most sought after by users, whose concerns range from flash flooding to agricultural demands to simply whether to grab an umbrella.

The error growth curves for 12-h precipitation accumulations show that the full ensemble exhibits a nearly complete lack of error growth through the 2-day forecast period (doubling time ~8 days; Fig. 7). The doubling times for the two bred mode groups are significantly shorter than for the ANL group, but still longer than would be expected for mesoscale phenomena, significantly longer than for 500-mb height even (cf. Fig. 3). It is not clear whether the flat curves denote saturation or mainly reflect a shortcoming in our methodology.

Since little insight can be gleaned from the flat rms curves for NCEP ensembles, it is of interest to observe similar curves for accumulation periods shorter than 12 h. Unfortunately forecast output at a temporal frequency higher than 6 h is not available for the NCEP short-range ensemble. So we turn to the MM5 ensemble described in section 2, for which hourly rain rates are available for a 36-h forecast period. One-, 3-, 6-, and 12-h totals are assembled from these data, so that the effect of the length of the accumulation period and temporal smoothing could be examined. The error growth curves for each of these accumulation periods are shown in Fig. 8 (solid). The increase in rms differences with the length of the accumulation period is due mainly to the fact that higher precipitation amounts occur for 12 h than for 1 h, and thus a wider range for error is possible. Of greater interest, the 1-h rain rate forecasts appear to become flat by ~6 h, a signal consistent with nonlinear saturation. The 3- and 6-h rain rate curves appear to saturate by 12 h and the 12-h rain rate by 24 h. Recall that the MM5 ensemble contains the combined impact of analysis and model uncertainty, which yields greater dispersion (Harrison et al 1999; Buizza et al. 1999) and shorter saturation times than the use of just initial perturbations. Regardless, the relationship between shorter accumulation period and faster saturation time is undoubtedly independent of ensemble configuration.

The fact that lengthening the accumulation period increases predictability is not surprising and has been observed in prior studies of tropical convection (e.g., Islam et al. 1993). For example, if a model forecast missed the timing of a particular storm by 4 h, then this rainfall would be missed entirely by some of the 1- and 3-h accumulation forecasts. Conversely, a 6-h accumulation forecast might capture part of the storm and the 12-h accumulation forecast would be considered completely accurate. In much the same way, spatial smoothing of the precipitation fields can also improve the accuracy of a forecast, in effect by increasing the envelope of what is considered an accurate forecast. To explore the issue of spatial averaging on precipitation, a nine-point

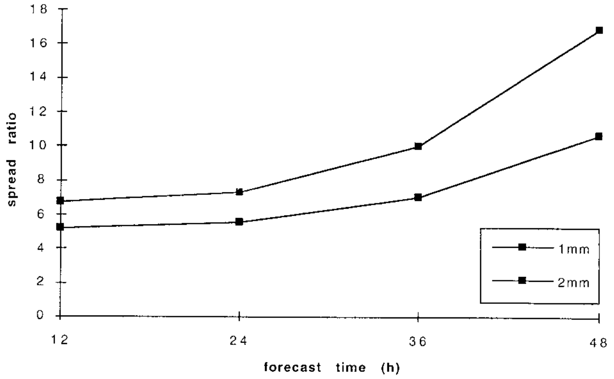

FIG. 9. Spread ratio for 1- and 2-mm thresholds from the FULL ensemble.

low-pass filter [Shapiro 1970, Eq. (31)] with a half-power cutoff of 4Δx is applied to the precipitation fields to simulate a coarser-resolution forecast (Fig. 8, dashed lines). Rather than lengthening the saturation time, the spatial smoothing lowers the rms differences nearly equally throughout the forecast period. In terms of the percent decrease in the errors (not shown), the shorter accumulation periods receive the greatest benefit from spatial smoothing. Overall, spatial averaging appears to have a smaller impact on error growth and saturation time than temporal averaging.

Comparison of quantitative precipitation amounts at each grid point, as judged from rms differences, is just one measure of forecast similarity. Different metrics can be defined. One such measure is a "spread ratio" [see the appendix and Stensrud and Wandishin (2000), hereafter SW], which in essence provides an estimate of the forecast dispersion of the precipitation shield for a given accumulation threshold. Examination of the evolution of the spread ratio for the 1- and 2-mm thresholds for the full NCEP ensemble (Fig. 9) suggests that error growth lasts well beyond 48 h by this measure. The curve for the 1-mm threshold, which is little more than a yes/no forecast of rain, has a doubling time of just over 2 days, or longer than even the rms estimate for 500-mb height. The 2-mm threshold has a doubling time of roughly 42 h. This reduction in doubling time for the higher threshold reflects the difficulty in predicting higher precipitation amounts. The longer doubling times for the rms estimates compared to the spread ratios is consistent with the notion that forecasting precipitation amount is more difficult than forecasting coverage because the predictability limits for coverage are significantly longer.

4. Precipitation forecasts

a. Equitable threat scores

The ETS values, averaged over the 19 cases for which verifying RFC analyses are available, are shown in Table

TABLE 3. Bias and ETS calculated for the 80-km operational Eta model (OPNL) and the 29-km Meso Eta model (MESO) at 24 h into the model forecasts of precipitation. The same is calculated at 24 and 48 h into the forecasts for the five Eta runs initialized from the in-house analyses (ANL), the five Eta runs initialized from the bred modes (BRD), and the five Regional Spectral Model runs (RSM), the 10-member mixed-perturbation (ETA), the 10-member mixed-physics (PHYS), the 10-member mixed-perturbation, mixed-physics (MIX), and the 15-member NCEP short-range ensemble (FULL). Matching bold, italics, or underline indicates pairs for which the differences are statistically significant. See section 2d for list of seven pairings for which significance was tested.

Pcat	OPNL	MESO	FULL	ETA	PHYS	MIX	ANL	BRD	RSM
Bias									
24 h									
0.76 mm	0.96	0.84	1.15	1.06	1.15	1.23	1.10	1.01	1.24
2.54 mm	1.16	1.08	1.25	1.22	1.20	1.29	1.28	1.15	1.29
12.7 mm	1.47	1.75	1.01	1.06	1.06	1.14	1.29	0.98	1.21
25.4 mm	1.15	1.08	0.66	0.83	0.83	0.67	0.97	0.70	0.76
48 h									
0.76 mm			1.22	1.07	1.27	1.29	1.08	1.04	1.39
2.54 mm			1.27	1.18	1.28	1.34	1.20	1.12	1.51
12.7 mm			1.05	1.08	1.00	1.07	1.09	1.04	1.01
25.4 mm			0.54	0.66	0.50	0.56	0.82	0.53	0.71
ETS									
24 h									
0.76 mm	0.47	0.44	0.47	0.48	*0.45*	*0.46*	0.48	**0.47**	**0.41**
2.54 mm	0.52	0.50	0.53	0.52	0.52	0.52	0.51	0.51	0.48
12.7 mm	0.35	0.37	0.39	0.42	0.38	0.40	0.39	0.36	0.39
25.4 mm	0.22	0.27	0.24	0.22	0.22	0.22	0.26	0.21	0.22
48 h									
0.76 mm			0.40	0.40	0.38	0.39	0.40	0.40	0.35
2.54 mm			0.41	0.40	0.40	0.40	0.40	0.38	0.34
12.7 mm			0.40	0.34	0.41	0.38	0.40	0.36	0.28
25.4 mm			0.27	0.31	0.20	0.26	0.29	0.25	0.18

TABLE 4. As in Table 3 but for calculations of RPS, RPSS, and the percentage of verifying grid points showing skill (RPSS>0) with respect to a climatological probabilistic forecast. Matching bold, italics, single or double underlines, and/or single of double overbars indicates pairs (of RPS and RPSS) for which differences are statistically significant. See section 2d for listing of seven pairs for which significance was tested.

h	OPNL	MESO	FULL	ETA	PHYS	MIX	ANL	BRD	RSM
RPS									
24	$\underline{\underline{\underline{0.49}}}$	$\overline{\overline{\underline{\underline{\underline{0.42}}}}}$	$\overline{\overline{0.29}}$	$\overline{0.31}$	$\underline{0.30}$	$\overline{\underline{0.30}}$	*0.32*	***0.33***	**0.39**
48			0.28	0.31	0.30	0.30	0.33	**0.32**	**0.39**
RPSS									
24	$\overline{\overline{0.14}}$	$\overline{\overline{\underline{\underline{\underline{0.22}}}}}$	$\overline{\overline{0.46}}$	0.42	$\underline{0.44}$	$\underline{0.45}$	*0.41*	***0.39***	**0.26**
48			0.42	0.37	0.39	0.39	0.33	**0.34**	**0.17**
Percent of grid points skillful relative to climatology									
24	58.2	63.4	62.5	65.7	53.9	53.0	63.0	64.7	36.6
48			54.0	56.9	46.1	45.9	55.7	56.6	31.8

3 for the FULL ensemble and several ensemble subgroups, as well as single model runs from the 80-km operational Eta and the 29-km Meso-Eta. The RFC verification data contain 24-h accumulations, so scores can be computed only for 1- and 2-day forecasts. The Meso-Eta forecast period is 36 h, so it is verified only at the 24-h time. The operational Eta is included merely for comparison with the Meso-Eta, so its 48-h forecast scores are not given. Ensemble group pairings with statistically significant differences are indicated in the table. The range of bias scores exhibited by the different ensemble groups for matching forecast times and threshold values (Table 3) indicate the need for adjusting the ETS when conducting significance tests, as discussed in section 2d.

Statistically significant differences in ETS exist for only three pairings, and these are all for the lowest precipitation threshold at 24 h. However, some further inferences may be drawn. If the current signal holds for larger sample sizes, the RSM forecasts appear they might be inferior to those from the Eta groups, particularly at 48 h. Meanwhile, the perturbation method did not affect the accuracy of the ensemble forecasts in our sample, as is suggested by the similarity in ETS between the BRD and ANL. Among the 10-member ensembles, it is important to note that inclusion of the inferior RSM members does not degrade skill relative to the ETA ensemble. The improvements gained from including different model physics balance the lower skill of that model. Finally, while the increase in skill gained from combining all 15 members relative to the 10-member ensembles is not significant, the skill of the FULL ensemble is significantly greater than the 29-km MESO forecasts.

b. *Ranked probability skill scores*

As mentioned previously, a primary advantage of ensemble forecasting is its inherent probabilistic nature. Table 4 shows the RPS and the RPSS, with respect to a climatological forecast, for the 19 case days with verification data. Ensemble group pairings with statistically significant differences are indicated in the table.

Surprisingly, the domain-averaged RPSs do not appreciably change from the 24- to 48-h forecast. This counterintuitive result is related to the difficulty of forecasting heavy precipitation amounts and to our limited sample. The main area of most frequently observed heavy precipitation (12.7 mm or greater) is centered over the Mississippi and Ohio River valleys at 24 h (Fig. 10), but by 48 h (Fig. 11) this area decreases as the heaviest precipitation begins to move off the East Coast. The RPSS for the ensemble groups do show an appreciable loss of skill with time, however, which indicates that long-term climatology performs better for day 2 than for day 1. In other words, the weather events of day 2 appear to be easier to forecast than the day 1 events for our limited sample of 19 case days. This points to the need for larger, more representative samples to judge model performance.

Table 4 also reveals some significant differences between the RPS and RPSS values for the various pairings of ensemble groups. All of the five-member subgroups are much more skillful than the deterministic OPNL forecast, a clear indication of the value of ensembles with few members for PQPF in general. Of the three ensemble subgroups, the RSM ensemble is considerably less skillful than the two five-member Eta Model subgroups. Nonetheless, its inclusion in the 10-member PHYS and MIX ensembles leads to an apparent increase in skill relative to the ETA group, which points to the utility of a mixed-model ensemble configuration for PQPF. On the other hand, the BRD and ANL groups exhibit comparable skill, as do PHYS and MIX; thus, it appears that the different perturbation methods did not have a major impact on 24-h PQPFs.

The spatial distribution of the RPSS for four ensemble groups, shown in Fig. 12 for 24-h forecasts, indicates that skill can vary substantially across the domain. Five distinct areas of high skill (RPSS > 0.5) are evident in

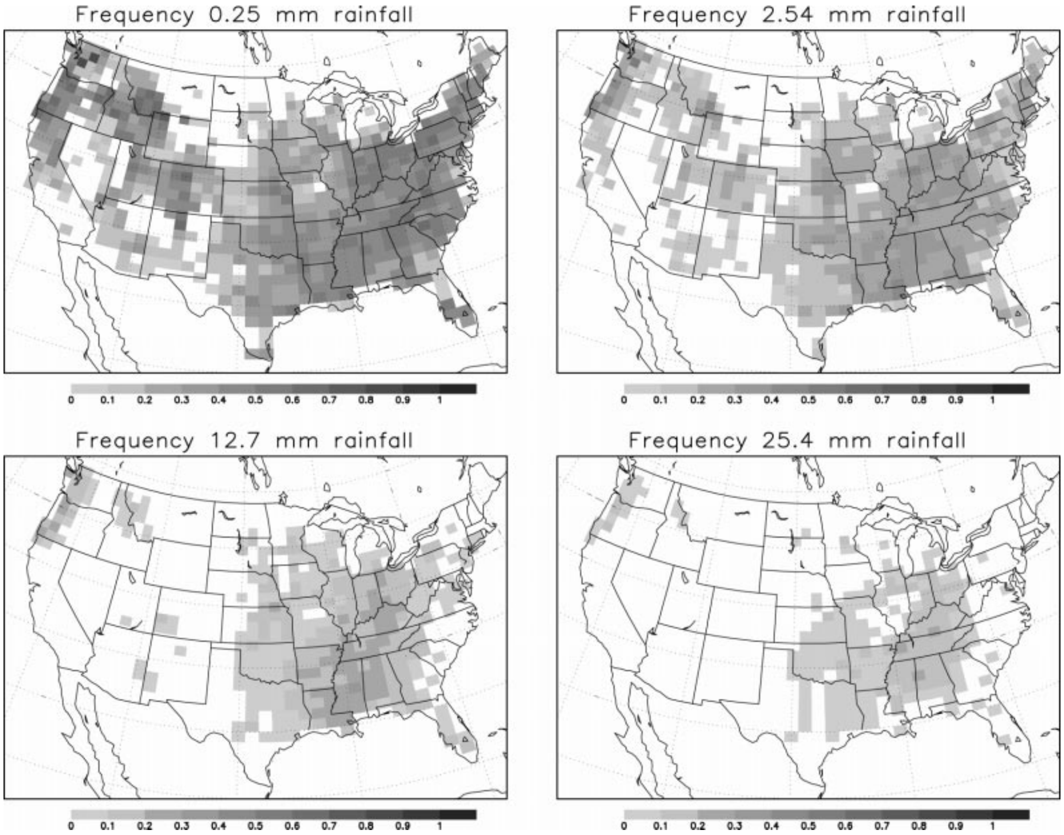
Frequency 0.25 mm rainfall
Frequency 2.54 mm rainfall
Frequency 12.7 mm rainfall
Frequency 25.4 mm rainfall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

FIG. 10. Map of the frequency of occurrence of 24-h accumulations of precipitation greater than 0.25, 2.54, 12.7, and 25.4 mm for the 19 cases for which the RPSS was calculated. White areas indicate regions with no verifying data or where precipitation amounts did not exceed the threshold.

our sample: southern New England, central Appalachia, the southern Mississippi River valley, northwest Arizona, and central California. The RPSS distributions for three Eta Model groups are very similar, whereas the RSM group differs substantially from them, which attests to the importance of model configuration over perturbation design for 24-h PQPFs. The Eta groups are skillful over ~65% of the grid points, whereas the RSM ensemble is skillful over less than half (~35%) of the area.

The spatial distributions of the RPSS at 48 h (not shown), though shifted somewhat, are very similar to the 24-h distributions. The percentage of verifying grid points that are skillful drops to ~30% for the RSM and to ~55% for the three other groups. Consequently, while all ensemble groups retain skill in a spatially averaged sense, they are less useful in that the unskillful area becomes as large as or larger than the skillful area.

An important practical issue concerns the relative merit of ensemble forecasts at coarser resolutions versus deterministic forecasts with higher-resolution models of comparable computational cost. Prior studies (Du et al. 1997; Hamill and Colucci 1998; Stensrud et al. 1999) found that skill for a short-range ensemble was comparable to or greater than skill for a single run of the higher-resolution model. Here we compare the performance of the 80-km ensembles and 29-km Meso-Eta Model for 24-h PQPFs.

The FULL ensemble, the 10-member groups, and even the two 5-member Eta groups, are roughly twice as skillful as the MESO (Table 4). Even the RSM ensemble is significantly more skillful than the MESO. Note that the extent to which the ensemble is an improvement over the MESO, in terms of RPSS, exceeds the improvement gained from the increased resolution of the Meso-Eta over the operational Eta. However, it is important to note that the superiority of the MESO over the OPNL implies that an ensemble based upon a higher-resolution model would likely perform significantly better than the ensemble formulation in this study. The percentage of skillful grid points for the MESO is ~65%, or nearly the same value as for the Eta ensembles. Comparison of the spatial distribution of the RPSS for the MESO (Fig. 13) and for the five-member Eta ensembles (Fig. 12) shows that the regions of high skill and no skill, respectively, tend to coincide. Even in the West, where the distribution of precipitation depends greatly on orographic effects and a finer grid spacing provides a better representation of this complex terrain,

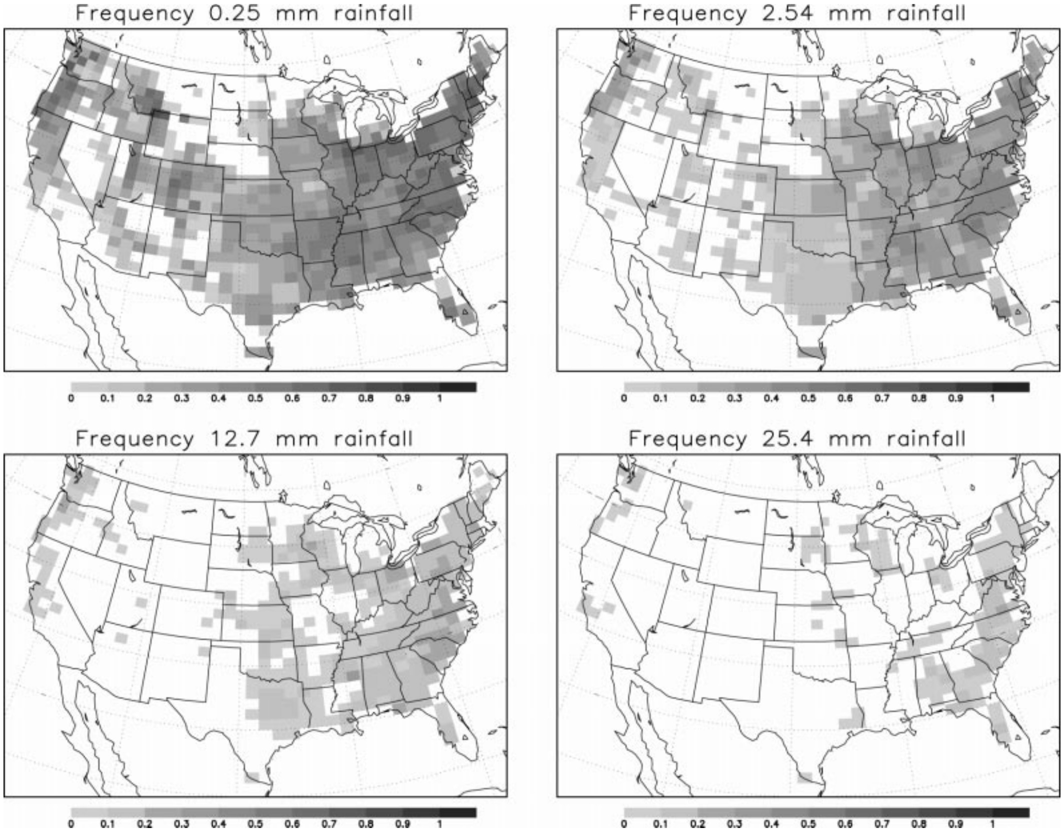
Frequency 0.25 mm rainfall
Frequency 2.54 mm rainfall
Frequency 12.7 mm rainfall
Frequency 25.4 mm rainfall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

FIG. 11. As in Fig. 10 but for 48 h.

the Meso-Eta evinces no apparent improvement over the coarser Eta ensembles.

c. Relative operating characteristic

Comparison of the areas under the ROC curve, from the trapezoid rule, for the different ensemble groups along with the OPNL and MESO (Table 5) yields conclusions largely consistent with those drawn from the other scoring metrics examined. One notable exception is that the OPNL slightly outperforms the MESO for all but the highest threshold, both of which are roughly on par with, or slightly below, the RSM. The ANL and BRD groups have somewhat larger areas than the RSM, particularly at the highest threshold. The 10-member groups show improvement over the 5-member groups, while the FULL ensemble is largely indistinguishable from the MIX group. Note that all models show areas above 0.5 for all thresholds and, thus, possess the ability to discriminate precipitation events.

The areas under the fitted ROC curves (Fig. 14) display very similar behavior. (Error bars are constructed such that a separation between the error bars of two groups indicates that the areas for the two groups are significantly different to the 95% confidence level.) The RSM group is significantly worse than all other groups for most categories and worse than the FULL ensemble for all categories save the 25.4-mm threshold at 48 h. The performance capability of the other five-member groups to discriminate precipitation events varies greatly. For example, the potential accuracy of the BRD group is good for the 0.254-mm threshold at 24 h, relative to the FULL ensemble, but not at 48 h, and is poor for the 12.7-mm threshold but strong for the 25.4-mm threshold at both forecast times. The 10-member groups and the FULL ensemble are indistinguishable for most categories.

The ROC provides more information, however, than can be summarized by a single score such as the area. Plots of the ROC at 24 h for the five-member groups, the FULL ensemble, and the deterministic runs are shown in Fig. 15. Recall that in place of calculating areas, a deterministic system can be compared to an ensemble by noting whether its (HR, FAR) point falls above or below the ensemble ROC curve. Figure 16 shows the HR (with 95% confidence bounds) of the ensemble groups at the FARs of the deterministic runs. The HRs of the MESO and OPNL runs are indicated by the dotted lines in the figures. Again, the RSM group is the only ensemble to possess a discriminating capability inferior to the deterministic forecasts. Aside from the significantly larger HRs of the FULL, MIX, and PHYS groups at 12.7 mm, the other ensemble groups exhibit a capability on a par with the deterministic runs.

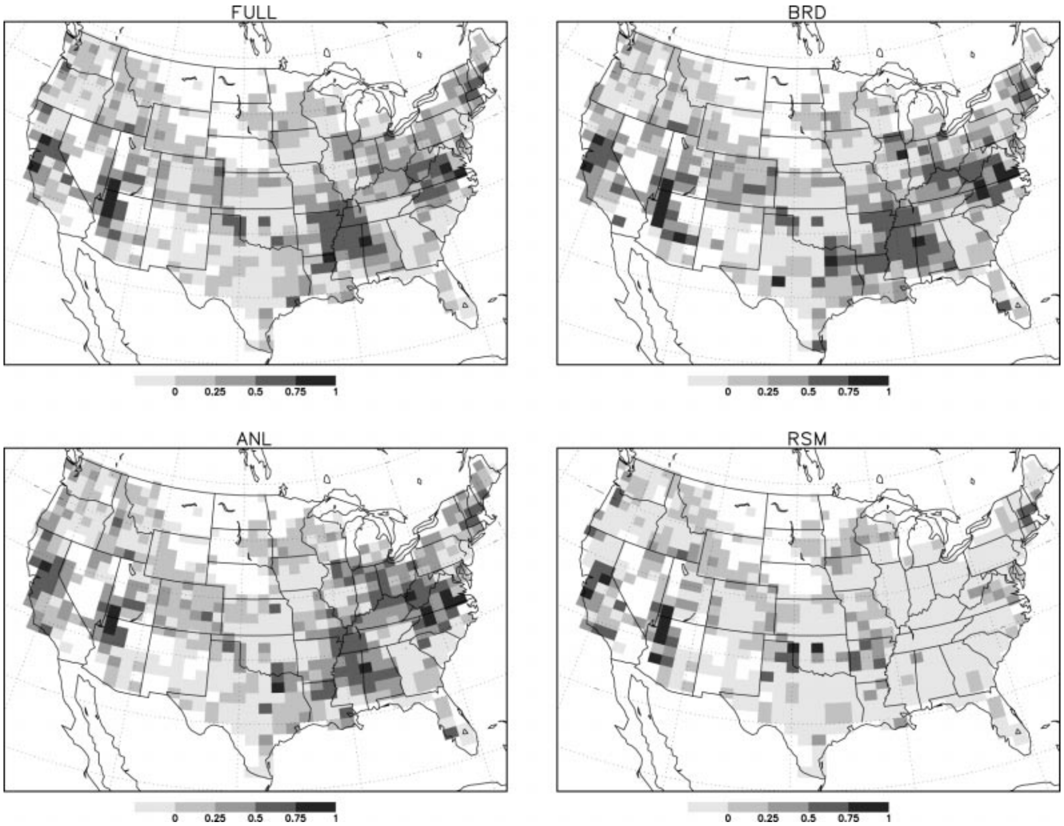

FIG. 12. Map of the RPSS for 24-h precipitation forecasts from FULL and the three five-member ensemble subgroups (BRD, ANL, RSM). White areas indicate regions with no verifying data.

This method of comparing forecasts should be seen as roughly equivalent to using the fitted ROC curves for the ensembles (this technique cannot be applied to the signal data point of the deterministic models). Thus, it should not be surprising that the OPNL scored on a level with the ensemble groups since the ensembles are largely constructed from the same model and so one might expect that they would possess similar discriminating capability. Perhaps more surprising is that the MESO does not appear to improve this capability. In examining the discriminating performance, the inability to choose between difference decision points from the deterministic runs becomes increasingly important at the higher thresholds, where the HR of the ensembles is raised substantially with only a slight increase in the FAR (Fig. 15); that is, by lowering the decision criterion for a ‘‘yes’’ forecast, more events can be correctly forecast with very little penalty. Indeed, the greater the size of the ensemble, the more decision criteria can be used, thereby providing the potential for greater utility to the end user (Richardson 2000).

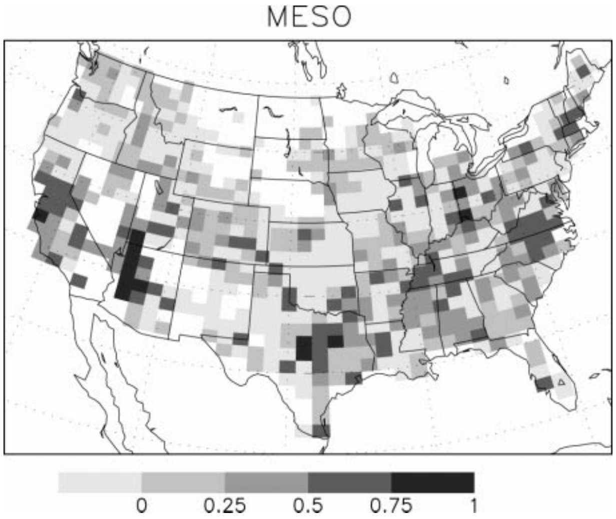

FIG. 13. As in Fig. 12 but for forecasts from the Meso-Eta (MESO).

5. Summary and discussion

In this paper, forecasts from the experimental NCEP short-range ensemble forecast system, supplemented by forecasts from the MM5 modeling system, were used to estimate deterministic predictability limits for precipitation and convective indices and to assess the accuracy and skill of ensemble forecasts of precipitation. Results were computed for a number of case days, 43

TABLE 5. Area beneath the discrete ROC curve (A), calculated for the same forecast systems as in Table 3.

Pcat	OPNL	MESO	FULL	ETA	PHYS	MIX	ANL	BRD	RSM
24 h									
0.254 mm	0.79	0.77	0.85	0.85	0.85	0.85	0.84	0.83	0.79
2.54 mm	0.85	0.83	0.92	0.91	0.91	0.91	0.90	0.89	0.87
12.7 mm	0.79	0.77	0.91	0.89	0.90	0.91	0.88	0.84	0.82
25.4 mm	0.71	0.72	0.80	0.77	0.77	0.79	0.76	0.74	0.70
48 h									
0.254 mm			0.85	0.84	0.85	0.85	0.83	0.82	0.80
2.54 mm			0.88	0.87	0.87	0.88	0.85	0.84	0.83
12.7 mm			0.90	0.87	0.87	0.89	0.85	0.83	0.81
25.4 mm			0.78	0.74	0.76	0.75	0.71	0.73	0.63

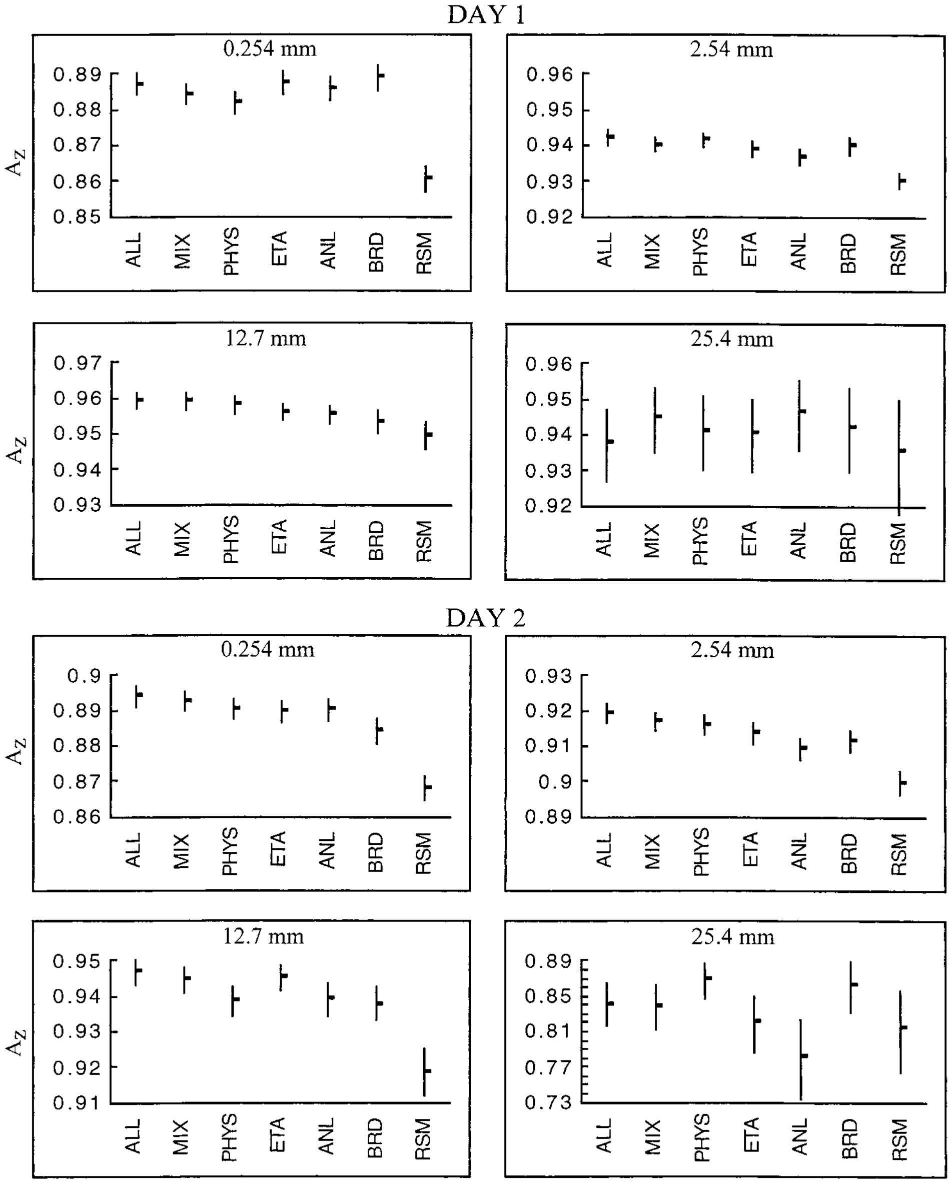

FIG. 14. Areas (A_z) under fitted ROC curves for all of the ensemble groups for the four precipitation thresholds: 0.254, 2.54, 12.7, and 25.4 mm, for day 1 and day 2 forecasts. Error bars are computed such that ensemble groups with nonoverlapping bars are significantly different with 95% confidence.

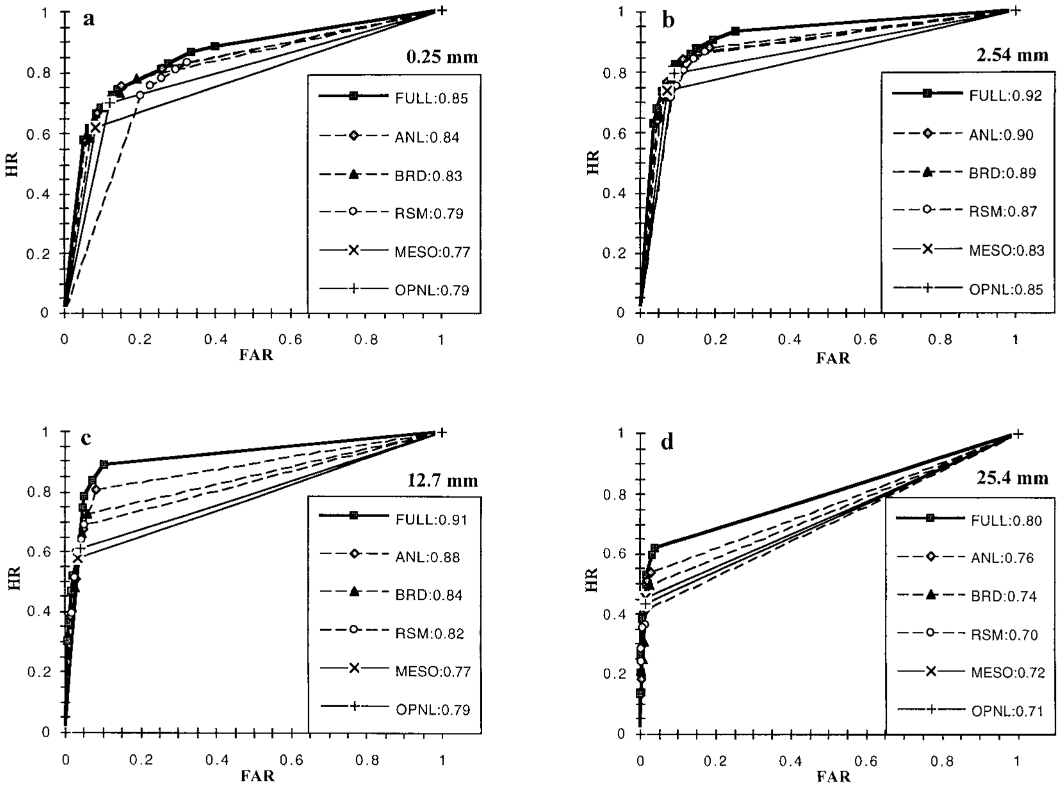

FIG. 15. Discrete ROC curves and area (*A*) under the ROC curves for the FULL, ANL, BRD, and RSM ensemble groups and the single-run Meso MESO and operational Eta Model (OPNL) calculated for the thresholds (a) 0.25, (b) 2.54, (c) 12.7, and (d) 25.4 mm.

for the predictability estimates and 19 for the forecast verification. The statistical significance of differences in skill between different model configurations was assessed.

Predictability error growth of the 500-mb height field shows general agreement with the prior results from global models, and is consistent with the pessimistic view presented by Lorenz (1982) that the predictability of the true atmosphere will be overstated by the models. The NCEP Eta ensemble has a more difficult time forecasting more complex fields, such as CAPE and CIN. An initial 6–12 h of error decay, most likely associated with adjustments by the models' convective schemes to initialized instabilities, is followed by a nearly flat error growth curve for the duration of the forecast period. It is likely that this behavior is due, at least in part, to the quasi-equilibrium assumption in the models' convective schemes.

The NCEP ensembles exhibit only weak error growth for 12-h rainfall totals. The weak growth in the NCEP ensembles is related in part to the long accumulation period. Error growth curves for a mixed-physics ensemble with the MM5 model reveal that the predictability limit for precipitation decreases significantly as the accumulation period becomes shorter. The NCEP ensembles exhibit more dispersion when the similarity measure was altered to emphasize coverage. The doubling time for the spread ratio is greater than 2 days. Our results indicate that deterministic forecasts with current analysis–forecast systems are able to provide only limited guidance for precipitation amount, but ensemble forecast systems can provide useful guidance for the areal coverage of precipitation and PQPF.

The NCEP ensembles at 80 km are significantly more skillful than the 29-km Meso-Eta forecast. The improvement of the full ensemble over the operational Eta for PQPF, as judged by the RPSS, is nearly twice the improvement gained from the increased resolution of the Meso-Eta. Even the least skillful five-member ensemble, the RSM, is significantly more skillful than the Meso-Eta while requiring about half of the computation time (Hamill and Colucci 1997). Comparison of the ensemble subgroups indicates that model configuration is more important than perturbation strategies for PQPF and points to the potential value of using mixed-model or mixed-physics ensembles for precipitation forecasts. Once again, it should be noted that the significant increase in skill of the MESO compared with the operational Eta implies that an ensemble system based on a higher-resolution model would yield still greater improvements in skill over the ensemble used in this study, an ensemble that already provides a nearly 50% improvement over climatology.

The above conclusions about ensemble skill were

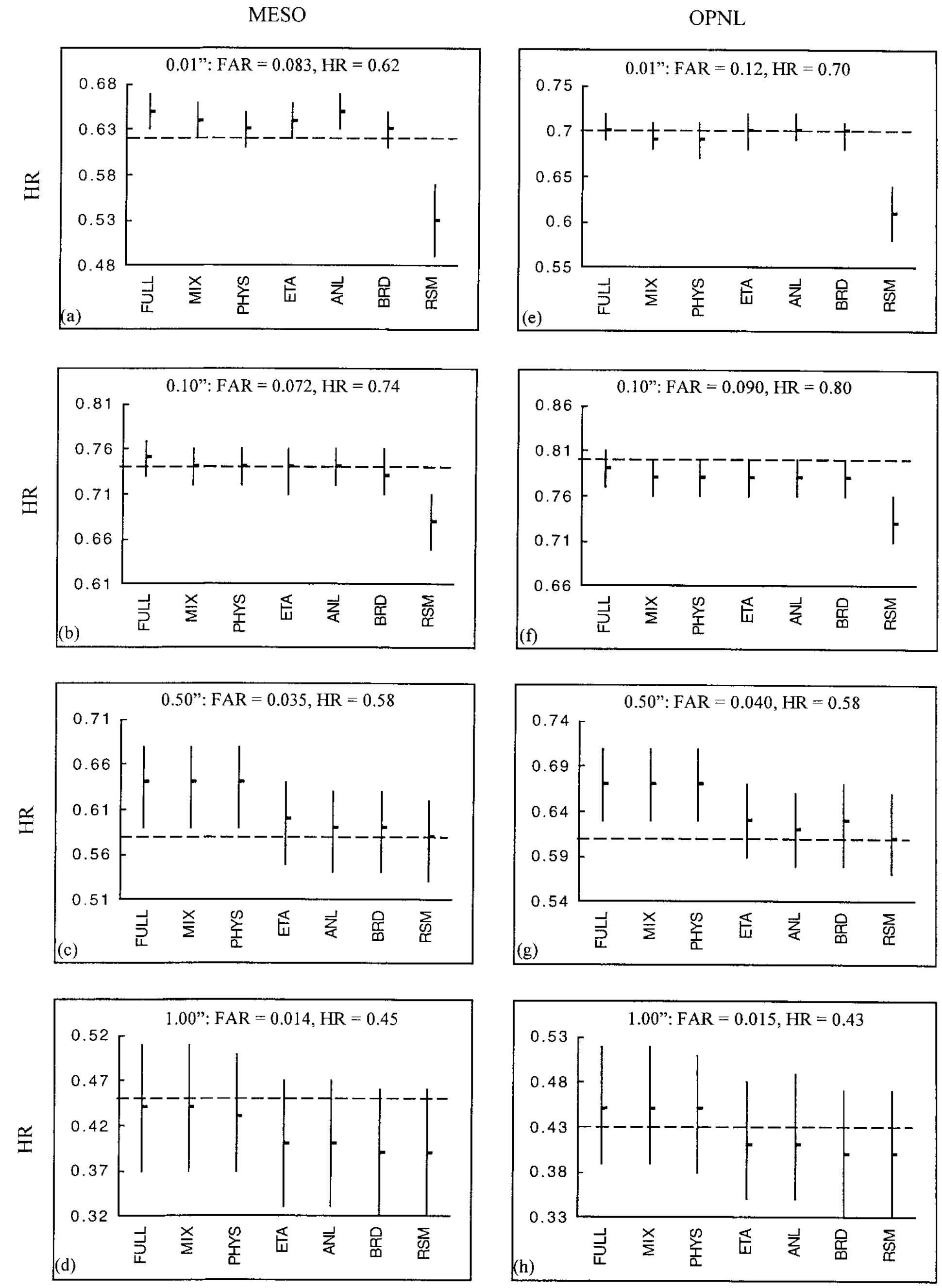

FIG. 16. Hit rates (HR), with 95% confidence intervals, for all of the ensemble groups computed from the fitted ROC curves at the same FAR as the deterministic runs: (a)–(d) MESO and (e)–(h) OPNL, for the same thresholds as in Fig. 15. FARs are given in each panel. HRs of the deterministic runs are given in each panel and denoted by the horizontal dotted line.

based solely on the RPSS, owing to the lack of useful information obtained from the ETS. As mentioned previously, this may be due, in part, to the small sample size that was available. However, it also highlights the fact that the ETS is not well suited as a verification measure for ensemble forecasting, as it reduces a probabilistic ensemble forecast to a deterministic one, thus discarding much useful information. In the same manner, using the ensemble mean as a tool for evaluating forecasts is insufficient.

The ROC is a metric that measures the ability to discriminate dichotomous events and can be used to evaluate deterministic and ensemble forecasts. Comparisons of the areas under the ROC curves for the different ensemble groups along with the MESO and OPNL models largely agree with the RPSS results in terms of the discriminating performance, although the difference between the ensembles and the deterministic forecasts is much less pronounced. The deterministic runs are largely equivalent to the ensemble groups in terms of discriminating capability, but this does not reflect the information that forecasters derive from the models when forming their forecasts. The fact that the MESO and the Eta-based ensemble possess similar discriminating capability suggests that, at this point, additional CPU resources would be better used on increasing the number of ensemble members (decision criteria), and thus the discriminating performance, rather than on increasing model resolution. By any measure, the RSM yields an inferior ensemble group.

The NCEP short-range ensemble demonstrates that ensemble prediction systems with limited-area models can be useful for operational precipitation forecasts. Even the inclusion of members from a less skillful model can add skill in a mixed ensemble system. However, many basic questions concerning the design of an ensemble forecasting system remain unanswered. Previous studies indicated a need to increase the spread of short-range ensemble systems (Hamill and Colucci 1997, 1998; Stensrud et al. 1999). Our results indicate that increasing the spread of the 500-mb height field (Fig. 3) or sea level pressure (SLP) field through dynamically conditioned initial perturbations does not guarantee a better precipitation forecast. The ensemble based on the bred modes from the global ensemble system produces greater spread for 500-mb height and SLP than the ensemble based on different in-house analyses, but the two ensembles possess comparable skill in terms of PQPF. Therefore, examinations of the spread of a mesoscale model should focus on mesoscale phenomena. Also, attempts at increasing the spread to more reliably encompass reality must be done judiciously to avoid deteriorating the predictive skill of the ensemble.

The in-house analyses in this study are just one manifestation, which was readily available, of the Monte Carlo method. It is of interest to examine the utility of short-range ensemble forecasts based on other Monte Carlo paradigms such as an ensemble Kalman filter, as advocated by Anderson (1996, 1998) and Houtekamer and Mitchell (1998), relative to ensembles based on optimal perturbation schemes. Moreover, it is well known that the model configuration can affect estimates of predictability limits (Tribbia and Baumhefner 1988) and that model deficiencies can have a deleterious impact on accuracy of ensemble forecasts (e.g., Murphy 1988). Therefore, future work is also needed to quantify the roles of model formulation and initial condition uncertainty for PQPF and predictability estimates.

Acknowledgments. The authors would like to thank Mike Baldwin for his assistance with the RFC precipitation analyses and clarification of the box-averaging method. M. Wandishin and S. Mullen were supported by NSF Grant ATM-9612487. S. Mullen was further supported by Grants ONR N00014-98-0551 and ONR N00014-99-1-0181.

APPENDIX

Spread Ratio

In illustrating concepts of probability, the Venn diagram is often introduced. This geometric diagram illustrates the relationships among events in a sample space and further can be used to indicate the probability of the union of intersecting events (see Wilks 1995). The spread ratio is an attempt to reproduce numerically in a single number the information contained in the simple Venn diagram.

The spread ratio (SR) is defined as the area of the union (U) of all specified field values divided by the area of the intersection (I) of these same specified field values, such that

$$\mathrm{SR} = \frac{U}{I},$$

where the U and I areas are defined using either a threshold value, or a range of values, for a given parameter. Since the intersection of all forecasts is in the denominator, the spread ratio can be unbounded.

The SR is simply the inverse of the correspondence ratio (CR) introduced by SW. The SR is used here in place of the CR because the characteristics of the SR curve (namely, the increasing values as forecast time increases) comport with the behavior typically seen in predictability curves.

REFERENCES

Anderson, J. L., 1996: Impacts of dynamically constrained initial conditions on ensemble forecasts. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., 56–57.

——, 1998: Application of a fully non-linear filter and Monte Carlo techniques to atmospheric data assimilation. Preprints, *14th Conf. on Probability and Statistics in the Atmospheric Sciences,* Phoenix, AZ, Amer. Meteor. Soc., 143–145.

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

——, E.-Y. Hsie, and Y.-F. Li, 1987: Description of the Penn State/NCAR mesoscale model version 4 (MM4). NCAR Tech. Note NCAR/TN-282 + STR, 66 pp.

Bamber, D., 1975: The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Math. Psychol.,* **12,** 387–415.

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 I, National Academy Press, 228–246.

Betts, A. K., and M. J. Miller, 1986: A new convective adjustment scheme. Part II: Single column tests using GATE wave, BOMEX, ATEX and arctic air-mass data sets. *Quart. J. Roy. Meteor. Soc.,* **112,** 693–709.

Black, T., 1994: The new NMC mesoscale Eta model: Description and forecast examples. *Wea. Forecasting,* **9,** 265–278.

Blackadar, A. K., 1976: Modeling the nocturnal boundary layer. Preprints, *Third Symp. on Atmospheric Turbulence, Diffusion and Air Quality,* Raleigh, NC, Amer. Meteor. Soc., 46–49.

——, 1979: High resolution models of the planetary boundary layer. *Advances in Environmental Science and Engineering,* J. Pfafflin and E. Ziegler, Eds., Vol. 1, No. 1, Gordon and Breach, 50–85.

Bluestein, H. B., 1993: *Synoptic–Dynamic Meteorology in Midlatitudes.* Vol. 2. *Observations and Theory of Weather Systems.* Oxford University Press, 594 pp.

Brooks, H. E., M. S. Tracton, D. J. Stensrud, G. DiMego, and Z. Toth, 1995: Short-range ensemble forecasting: Report from a workshop, 25–27 July 1994. *Bull. Amer. Meteor. Soc.,* **76,** 1617–1624.

Buizza, R., 1997: Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. *Mon. Wea. Rev.,* **125,** 99–119.

——, M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.,* **125,** 2887–2908.

Burk, S. D., and W. T. Thompson, 1989: A vertically nested regional numerical weather prediction model with second-order closure physics. *Mon. Wea. Rev.,* **117,** 2305–2324.

Cressman, G. P., 1959: An operational objective analysis system. *Mon. Wea. Rev.,* **87,** 367–374.

Dalcher, A., and E. Kalnay, 1987: Error growth and predictability in operational ECMWF forecasts. *Tellus,* **39A,** 474–491.

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

Doswell, C. A., III, 1987: The distinction between large-scale and mesoscale contribution to severe convection: A case study example. *Wea. Forecasting,* **2,** 3–16.

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

Dudhia, J., 1993: A nonhydrostatic version of the Penn State–NCAR Mesoscale Model: Validation tests and simulation of an Atlantic cyclone and cold front. *Mon. Wea. Rev.,* **121,** 1493–1513.

Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus,* **21,** 739–759.

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

Grell, G., 1993: Prognostic evaluation of assumptions used by cumulus parameterizations. *Mon. Wea. Rev.,* **121,** 764–717.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting,* **14,** 155–167.

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

——, and ——, 1998: Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.,* **126,** 711–724.

Harrison, M. S. J., T. N. Palmer, D. S. Richardson, and R. Buizza, 1999: Analysis and model dependencies in medium-range ensembles: Two transplant case studies. *Quart. J. Roy. Meteor. Soc.,* **125,** 2487–2516.

Harvey, L. O., K. R. Hammond, C. M. Lusk, and E. F. Mross, 1992: The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev.,* **120,** 863–883.

Houtekamer, P. L., and J. Derome, 1995: Methods for ensemble prediction. *Mon. Wea. Rev.,* **123,** 2181–2196.

——, and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.,* **126,** 796–811.

Islam, S., R. L. Bras, and K. A. Emanuel, 1993: Predictability of mesoscale rainfall in the tropics. *J. Appl. Meteor.,* **32,** 297–310.

Janjić, Z. I., 1994: The step-mountain Eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.,* **122,** 927–945.

Juang, H.-M., and M. Kanamitsu, 1994: The NMC nested regional spectral model. *Mon. Wea. Rev.,* **122,** 3–26.

Kain, J. S., and J. M. Fritsch, 1990: A one-dimensional entraining/detraining plume model and its application in convective parameterization. *J. Atmos. Sci.,* **47,** 2784–2802.

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

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

——, 1965: A study of the predictability of a 28-variable atmospheric model. *Tellus,* **17,** 321–333.

——, 1982: Atmospheric predictability experiments with a large numerical model. *Tellus,* **34,** 505–513.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.,* **30,** 291–303.

——, 1989: Dependence of the critical success index on sample climate and threshold probability. *Aust. Meteor. Mag.,* **37,** 75–81.

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

Molteni, F., and T. N. Palmer, 1993: Predictability and finite-time instability of the northern winter circulation. *Quart. J. Roy. Meteor. Soc.,* **119,** 269–298.

——, R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.,* **122,** 73–119.

Mullen, S. L., and D. P. Baumhefner, 1988: The sensitivity of numerical simulations of explosive oceanic cyclogenesis to changes in physical parameterizations. *Mon. Wea. Rev.,* **116,** 2289–2339.

——, and ——, 1989: The impact of initial condition uncertainty on numerical simulations of large scale explosive cyclogenesis. *Mon. Wea. Rev.,* **117,** 1548–1567.

——, J. Du, and F. Sanders, 1999: The dependence of ensemble dispersion on analysis–forecast systems: Implication to short-range ensemble forecasting of precipitation. *Mon. Wea. Rev.,* **127,** 1674–1686.

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.

——, 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting,* **8,** 281–293.

Parrish, D. F., and J. C. Derber, 1992: The National Meteorological Center's spectral statistical-interpolation analysis system. *Mon. Wea. Rev.,* **120,** 1747–1763.

Richardson, D., 2000: Skill and economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.,* **126,** 649–667.

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

Shapiro, R., 1970: Smoothing, filtering and boundary effects. *Rev. Geophys. Space Phys.,* **8,** 359–387.

Simmons, A. J., R. Mureau, and T. Petroliagis, 1995: Error growth and estimates of predictability from the ECMWF forecasting system. *Quart. J. Roy. Meteor. Soc.,* **121,** 1739–1771.

Spencer, P. L., and D. J. Stensrud, 1998: Simulating flash flood events: Importance of the subgrid representation of convection. *Mon. Wea. Rev.,* **126,** 2884–2912.

Stamus, P. A., F. H. Carr, and D. P. Baumhefner, 1992: Application of a scale-separation verification technique to regional forecast models. *Mon. Wea. Rev.,* **120,** 149–163.

Stensrud, D. J., and M. S. Wandishin, 2000: Correspondence and spread ratios in forecast verification. *Wea. Forecasting,* **15,** 593–602.

—— H. E. Brooks, J. Du, M. S. Tracton, and E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev.,* **127,** 433–446.

Swets, J. A., 1973: The relative operating characteristic in psychology. *Science,* **182,** 990–1000.

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.

——, J. Du, Z. Toth, and H. Juang, 1998: Short-range ensemble forecast (SREF) at NCEP/EMC. Preprints, *12th Conf. on Numerical Weather Prediction,* Phoenix, AZ, Amer. Meteor. Soc., 269–272.

Tribbia, J. J., and D. P. Baumhefner, 1988: The reliability of improvements in deterministic short-range forecasts in the presence of initial state and modelling deficiencies. *Mon. Wea. Rev.,* **116,** 2276–2288.

Vukicevic, T., and R. M. Errico, 1990: The influence of artificial and physical factors upon predictability estimates using a complex limited-area model. *Mon. Wea. Rev.,* **118,** 1460–1482.

Warner, T. T., R. A. Peterson, and R. E. Treadon, 1997: A tutorial on lateral boundary conditions as a basic and potentially serious limitation to regional numerical weather prediction. *Bull. Amer. Meteor. Soc.,* **78,** 2599–2617.

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