Value of Human-Generated Perturbations in Short-Range Ensemble Forecasts of Severe Weather

Victor Homar and David J. Stensrud

NOAA/National Severe Storms Laboratory, Norman, Oklahoma

Jason J. Levit and David R. Bright

NOAA/NWS Storm Prediction Center, Norman, Oklahoma

(Manuscript received 4 March 2005, in final form 25 August 2005)

ABSTRACT

During the spring of 2003, the Storm Prediction Center, in partnership with the National Severe Storms Laboratory, conducted an experiment to explore the value of having operational severe weather forecasters involved in the generation of a short-range ensemble forecasting system. The idea was to create a customized ensemble to provide guidance on the severe weather threat over the following 48 h. The forecaster was asked to highlight structures of interest in the control run and, using an adjoint model, a set of perturbations was obtained and used to generate a 32-member fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5) ensemble. The performance of this experimental ensemble is objectively evaluated and compared with other available forecasts (both deterministic and ensemble) using real-time severe weather reports and precipitation in the central and eastern parts of the continental United States. The experimental ensemble outperforms the operational forecasts considered in the study for episodes with moderate-to-high probability of severe weather occurrence and those with moderate probability of heavy precipitation. On the other hand, the experimental ensemble forecasts of low-probability severe weather and low precipitation amounts have less skill than the operational models, arguably due to the lack of global dispersion in a system designed to target the spread over specific areas of concern for severe weather. Results from an additional test ensemble constructed by combining automatic and manually perturbed members show the best results for numerical forecasts of severe weather for all probability values. While the value of human contribution in the numerical forecast is demonstrated, further research is needed to determine how to better use the skill and experience of the forecaster in the construction of short-range ensembles.

1. Introduction

The numerical forecasting of mesoscale phenomena and severe convective weather poses one of the most challenging problems faced today in the atmospheric community. Model physics, resolution, and data assimilation techniques are continuously improving and examples of promising simulations of severe convective systems can be found (e.g., Fowle and Roebber 2003). However, models still do not provide consistently reliable guidance for operations about important aspects of severe weather such as initiation, mode, intensity, and evolution of convection (Weiss et al. 2004). Admittedly, short-range mesoscale numerical forecasts are hampered by the largely unknown observational sampling errors at the meso- and small scales, as well as by the deficiencies in the models from such sources as physical parameterization schemes (e.g., Davis et al. 2003; Baldwin et al. 2002; Gilmore et al. 2004; Zamora et al. 2003). Additionally, little is known about the limits of predictability at the spatial and temporal scales of intermittent weather systems responsible for producing severe weather (Stensrud and Wicker 2004). The perception that multiple sources of uncertainty may largely degrade the forecast decreases the confidence of the forecaster in the output produced by mesoscale numerical models, even when they provide highly realistic looking forecasts (Weiss et al. 2004). Inevitably, observational

Corresponding author address: V. Homar, NOAA/National Severe Storms Laboratory, 1313 Halley Circle, Norman, OK 73069.
E-mail: victor.homar@uib.es

dataset errors and model deficiencies, as well as the predictability concerns, introduce inherent uncertainties that always are present in the forecast.

Ensemble techniques are one method that can be used to explicitly account for uncertainties in the numerical forecasting system and their use may assist forecasters in assessing appropriate levels of confidence. However, identifying, quantifying, and representing these uncertainties in the forecast system is a complex task. Ideally, one should consider a multivariate probability density function (pdf) defined in the model state phase space with each component representing uncertainty at each grid point of the analysis dataset (Epstein 1969). This pdf should then be evolved in time with a Fokker–Plank equation (Penland 2003; Ehrendorfer 1994) that accounts for model uncertainties through stochastic nonlinear dynamics. Currently, this method is intractable and, in practice, a much more modest approach to the problem is used. Modern computational resources still require the use of the standard deterministic set of equations and, hence, the selection of a limited number of realizations of the analysis pdf and model configurations. Still, combining the solutions of a number of slightly different numerical simulations not only produces a forecast that is more skillful than each individual simulation when examined over many cases (Leith 1974), but also provides a quantitative indication of forecast uncertainty (Tracton and Kalnay 1993). How these realizations (i.e., ensemble members) are constructed is currently the subject of significant attention in the weather research community (Shapiro and Thorpe 2004).

One reason for that attention is that the relative effect on forecast errors from the observational dataset errors versus model deficiencies is yet unclear (Stensrud et al. 2000; Bright and Mullen 2002). Recently, to cope with model uncertainty for mesoscale predictions of sensible weather, various model and physics parameterization perturbations have been considered, showing significant improvements over single-model systems (Wandishin et al. 2001; Du et al. 2004). These systems focus primarily on those components that likely have the largest effect on the sensible weather forecast, such as the dynamic core (Wandishin et al. 2001) and physical process parameterizations (Stensrud et al. 2000; Bright and Mullen 2002). On the other hand, multiple methods to choose an *optimum* ensemble of realizations from the analysis pdf have been proposed. For forecasts in the medium range, two well-established strategies have been adopted by the major operational centers in the United States and Europe. The breeding (Toth and Kalnay 1993) and singular-vector (Buizza and Palmer 1995) techniques have provided notable improvements in the skill of the medium-range forecasts, even without considering model deficiencies (Kalnay 2003). However, medium-range ensembles that include model uncertainties are found to be more skillful than ensembles that do not include model uncertainty (Evans et al. 2000)

Unfortunately, sampling the analysis pdf for applications on the mesoscale becomes more complex due to the larger and less known analysis error, the large role that physical process parameterization schemes play in model forecasts of sensible weather, and the end user's more sensitive dependence upon reliable forecasts. Indeed, the European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction System has virtually no skill in predicting probabilistic warnings of severe weather events for the United Kingdom for lead times of 1–2 days (Legg and Mylne 2004), suggesting that we have much to learn about designing ensemble strategies for short-range forecasts of severe weather. Currently, the breeding method is used to generate initial conditions perturbations in the National Centers for Environmental Prediction (NCEP) operational Short-Range Ensemble Forecasting system (SREF; Hamill and Colucci 1997; Wandishin et al. 2001. Despite being based on the full nonlinear perturbation growth, the breeding modes tend to project upon the structures of synoptic systems (Toth and Kalnay 1997), and are not necessarily linked to the structures of most concern in the sensible weather forecast.

Xu et al. (2001, hereafter Xu01) present a method aimed at identifying realizations of the pdf that focus the ensemble on specific areas of concern during the first 48 h of the forecast. They propose a method to generate members for a short-range ensemble that benefits from a forecaster's guidance in identifying areas where threatening weather is likely in the forecast and the atmospheric features that can influence the development of the threatening weather. The method, described in detail in Xu01, involves running a tangent linear adjoint model. Adjoint models track the gradient of a forecast aspect with respect to the model state vector backward in time to determine its sensitivity to the initial conditions (ICs) state vector. To do so, a linear operator is constructed tangent to the phase space trajectory followed by the forward nonlinear deterministic forecast. The transposition of such a linear operator results in the adjoint model [see Errico (1997) for a comprehensive overview of adjoint models]. Adjoint models have been extensively used in adaptive observation campaigns, such as the Fronts and Atlantic Storm-Track Experiment (FASTEX) and the North Pacific Experiment (NORPEX) (e.g., Langland et al. 1999a,b), and they are also used operationally to compute ECMWF singular vectors for the medium-range Ensemble Prediction System (Gelaro et al. 1998). Xu01 use the adjoint model to select perturbations to the initial conditions that produce the largest influence on the forecast of manually defined atmospheric features; these features are identified to impact the development and evolution of a mesoscale convective system in the central plains of the United States.

This approach of Xu01 assumes that the experience and skill of human weather forecasters is a useful addition to the process of creating ensemble systems. It is well known that forecasters routinely improve upon numerical guidance, as is clearly seen in skill scores for precipitation (Funk 1991; Olson et al. 1995). In addition, forecasters at the Storm Prediction Center regularly identify mesoscale-sized regions of significant severe weather threat through the issuance of outlooks and severe weather watches with a high level of skill (Leftwich et al. 1998; McCarthy et al. 1998). There is no reason to assume that this human knowledge and experience, although subjective, cannot be made useful in the creation of ensemble members and thereby benefit the operational forecast process, particularly for rare and significant events.

With the aim of assessing the value of short-range numerical forecast ensembles to assist in the operational forecasting of severe weather, the Storm Prediction Center (SPC) and the National Severe Storms Laboratory conducted the 2003 Spring Program (SP03) experiment focused primarily on the generation and interpretation of mesoscale short-range ensembles. Encouraged by the promising conclusions of Xu01, the SP03 included a subexperiment to test their method for a larger number of cases using operational forecasters as the main drivers of the system. The underlying idea was to create a daily, customized ensemble to provide guidance on the severe weather threat over the following 48 h. Essentially, the ensemble dispersion was intended to be generated in specific areas, and focused upon specific fields of interest, as opposed to everywhere in the domain, or following fast-growing modes under global generic norms.

In this paper we present verification results of this SP03 test ensemble. Severe weather reports and observed precipitation amounts are used to assess the quality of the experimental ensemble forecasts. In addition, results from the SP03 ensemble are compared to operational forecasts from the NCEP SREF system and the Eta Model for the same period. Finally, we test the skill of mixed ensembles constructed with breeding multimodel members from the SREF and the forecaster-generated members.

Section 2 presents the experimental design and the verification datasets. The automated model diagnosis and verification of severe weather are detailed in section 3. Precipitation forecasts are evaluated in section 4. Results from the mixed ensemble configurations are presented in section 5 and a summary of conclusions and recommendations for implementation are presented in section 6.

2. Experiment design

a. Generation of perturbations

The forecaster-generated ensemble consists of 32 forecasts produced using the nonhydrostatic fifth-generation Pennsylvania State University–National Center for Atmospheric Research (PSU–NCAR) Mesoscale Model (MM5, version 3; Dudhia 1993; Grell et al. 1994). The MM5 was selected because it has been shown to produce forecasts comparable to those of the Eta Model at similar horizontal grid spacing during the warm season (Colle et al. 2003) and because an adjoint version of the MM5 is available. This test ensemble of the SP03 experiment (MM5ADJ) ran weekdays from 28 April to 6 June (SP03 did not operate on weekends). A total of 27 cases are available for evaluating the ensemble guidance potential. To generate the set of different ICs for the ensemble, the method detailed in Xu01 was followed. Each day an experienced human severe weather forecaster was asked to identify 16 features of interest in the control run that were, in the forecaster's opinion, important to the potential development and/or evolution of severe weather on the following day (1200–1200 UTC day 2). The forecaster was able to select atmospheric structures at any time (in 6-h intervals) from the 48-h Eta control forecast for the following fields: horizontal and vertical wind components, temperature, specific humidity, geopotential height, sea level pressure, vertical relative vorticity and the lifted index. These fields were predetermined to allow an easier operational implementation for SP03, but in practice there is no other limitation on setting a *feature of interest* to initialize the adjoint model but to be differentiable with respect to the model state. Figure 1a shows examples of human-drawn features of interest for the forecast cycle of 5 May.

Table 1 shows the distribution of fields used by the forecasters during the entire experiment. The frequency of use is similar for all fields, with less preference for the vertical velocity and relative vorticity. This uniform distribution likely is a simple consequence of having a larger number of perturbations to create daily than the number of fields to choose from, inducing a tendency to use all available fields each day. Relative vorticity shows a lower use possibly because it is not

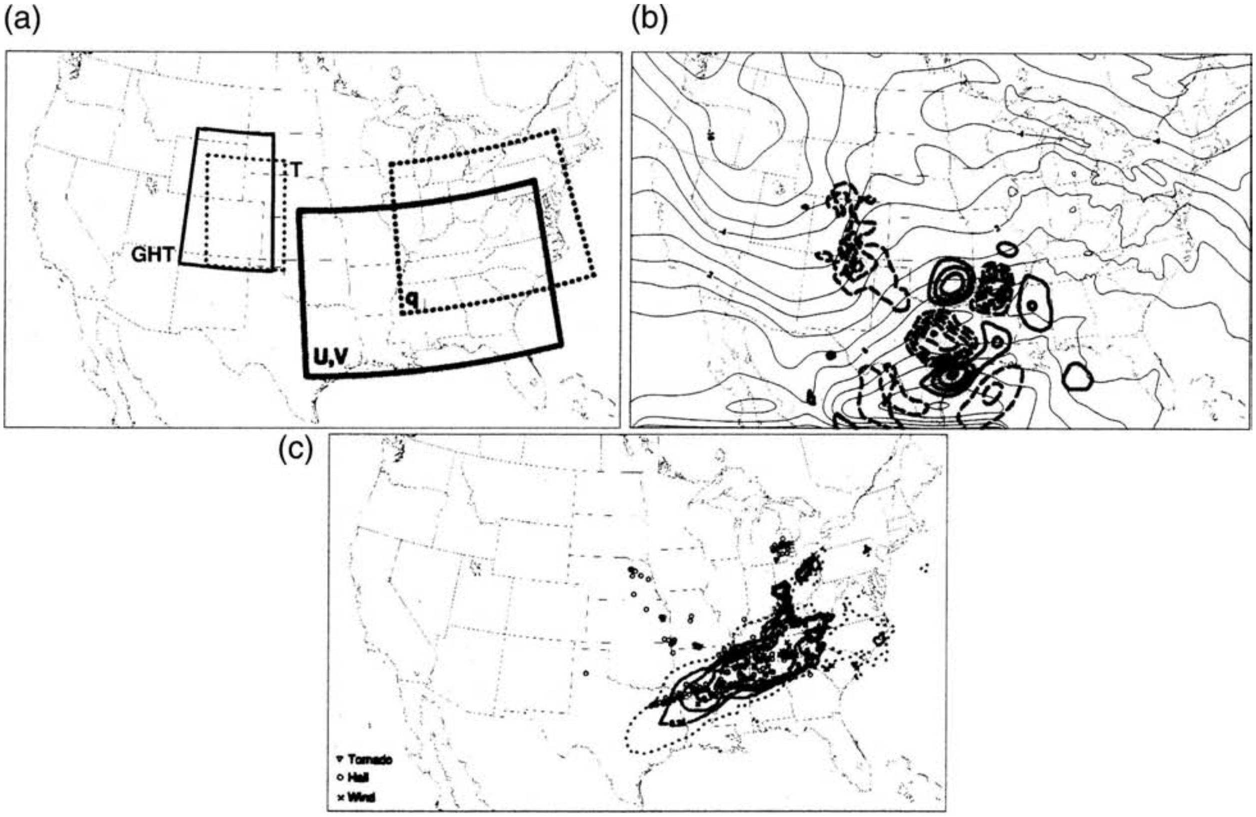

FIG. 1. Example steps from the ensemble test process for 5 May. (a) Areas and fields selected by the forecaster as important for the development of severe weather the next day. Geopotential height (GHT) and temperature T were selected at 500 hPa and t + 24 h lead time (1200 UTC 6 May); the wind components (U, V) at 500 hPa and the specific humidity q at the surface were specified at t + 36 h lead time (0000 UTC 7 May). (b) Example of initial condition perturbation of a single member: temperature at 700 hPa (°C; thin line) and perturbation (°C; thick line with 0.25°C intervals; the zero isoline is not shown and dashed lines show negative values) added to the original field as a result of the adjoint model run using the wind components feature selected by the forecaster [shown in (a)]. (c) Storm reports and probability of severe weather (black lines at 25% intervals) using the SCP parameter from the 32-member MM5 adjoint ensemble for the 24-h period beginning at 1200 UTC 5 May. Dotted lines depict the 5% probabilities.

traditionally used in mesoscale conceptual models for severe weather phenomena. The least frequent use of vertical velocity likely reflects the low confidence of the forecaster in its numerical forecast, selecting instead features related to the vertical velocity such as fronts, convergence zones, and jets. The distribution of vertical levels indicates that the surface, 850, and 500 hPa are preferred. The emphasis on low-level fields may be due to the forecasters focusing the ensemble on features related to convective initiation, which often is the significant concern for determining whether or not severe weather will occur (Johns and Doswell 1992; Stensrud and Fritsch 1993). Regarding the size of the structures identified by the forecasters, almost 80% of

TABLE 1. Relative frequency of fields, levels, areas, and forecast times selected by the forecasters and used to initialize the adjoint model integrations. A total of 432 perturbations (16 perturbations × 27 days) were defined during the experiment.

Field	U	V	T	Q	Vorticity	Height	Lifted index	SLP	W
Frequency (%)	12.7	11.3	13.2	11.3	7.2	13.0	14.4	11.8	5.1
Level	Surface	850 hPa	700 hPa	500 hPa	250 hPa				
Frequency (%)	36.1	20.6	9.9	28.5	4.9				
Area (10^3 km^2)	<200	200–400	400–600	600–800	800–1000	1000–1200	1200–1400	>1400	
Frequency (%)	12.3	32.4	23.8	13.9	4.9	3.7	3.0	6.0	
Forecast time (h)	+12	+18	+24	+30	+36	+42	+48		
Frequency (%)	0.5	0.9	11.3	26.4	47.0	8.3	5.6		

the areas are smaller than 800 10^3 km^2 (about the area of the state of Texas), which corresponds to the dimensions of large low-level mesoscale structures such as fronts, jets, and low pressure centers. However, larger areas were occasionally defined, with some of them being the size of half the conterminous United States (CONUS). These big areas likely rendered unreliable results from the adjoint model, and hence it is doubtful if these perturbations are useful, but they consist of only a few outliers during the whole experiment. The distribution in forecast hours shows that almost 75% of the time the forecaster defined the feature of interest at $t + 30$ h or $t + 36$ h, corresponding to the climatological peak in convective initiation for the day 2 forecast. This is in agreement with the forecast period under study in the SP03 experiment (day 2).

Each day, for each of the 16 selected features of interest, an adjoint model integration was correspondingly initialized and the sensitive areas of each forecaster-specified feature to the ICs were derived. The adjoint model used is the MM5 Adjoint Modeling System (Zou et al. 1997, 1998) developed by NCAR. The code is derived from a simplified version of the standard MM5. The adjoint runs have no parameterized convection but include explicit microphysics, radiation, and surface processes. Once the sensitivity fields were obtained from the adjoint, the horizontal wind components and temperature sensitivities were rescaled to a maximum amplitude of 8.0 m s^{-1} and 4.0 K, respectively. This rescaling, also used by Xu01, is intended to generate perturbations within the typical analysis error, and produced typical perturbation amplitudes of 2 m s^{-1} and 1.5 K at the 24–36-h forecast times. As shown later, these amplitudes are equivalent to the perturbation amplitudes found in the NCEP SREF ensemble and are consistent with observational uncertainty (Zapotocny et al. 2000). Finally, two MM5 simulations were run for each highlighted feature, each one using perturbations in both directions (positive and negative). Figure 1b shows an example of such perturbations for the temperature field at 700 hPa. Since the forecaster was requested to highlight 16 features each day, 32 perturbed simulations were produced to form the MM5ADJ ensemble.

Although the adjoint model is tangent linear, and hence the perturbations were defined strictly to change the forecaster-selected feature in a linear sense, the nonlinear evolution of the perturbation can be interpreted as a stochastic perturbation to the initial model state trajectory. However, this stochastic component of the perturbation will likely be confined about the area of concern in the forecast at the forecast time selected. In essence, by using both positive and negative perturbations the feature of interest likely is both enhanced and reduced equally in the linear sense. The nonlinear evolution of the positive and negative perturbations, however, may yield unexpected results since the specified feature of interest likely is not enhanced and reduced symmetrically in the two nonlinear forecasts. This nonlinear behavior is viewed as a positive attribute of the system, ensuring a rich diversity of solutions among the ensemble members over the forecaster-defined regions of concern as opposed to the trivial effects of the purely linear evolution of the linearly derived perturbations.

b. Model description

All simulations in the experiment are run with two domains interacting via a two-way nesting strategy. The coarser domain has 66 × 46 grid points, 90-km grid spacing, and covers the CONUS, southern Canada, the Gulf of Mexico, the eastern North Pacific, and the western North Atlantic. This is the domain used to run the adjoint model and define the IC perturbations. The inner domain has 157 × 97 grid points, 30-km grid spacing, and covers basically the CONUS. All simulations contain 24 sigma levels, with higher concentration at lower altitudes to better resolve boundary layer and near-ground processes. Subgrid moist convection is parameterized using the Kain and Fritsch (1990) scheme. A simple microphysics scheme that allows for ice concentration at temperatures below freezing is used (Dudhia 1989). Boundary layer processes are parameterized using the Eta planetary boundary layer (Janjić 1994) scheme together with the Dudhia (1996) five-layer simple soil model. Cloud and clear air radiative effects, as well as water vapor, carbon dioxide, and ozone concentrations, are considered in the radiation scheme. Both coarser and inner domains use the same parameterizations for all simulations. The Eta Model analysis at 1200 UTC is used to provide the ICs for the MM5ADJ. Time-dependent lateral boundary conditions are also provided by the 1200 UTC Eta Model results and are supplied to the simulation by means of a relaxation inflow–outflow five-point sponge frame. An upper radiative condition is used to minimize spurious noise reflection at the model top.

c. Verification and comparison datasets

The evaluation of the MM5ADJ is based on observations of severe weather and precipitation over the continental United States, east of the Rockies. All verification and forecasts are remapped to the MM5 30-km domain, in order to facilitate and ensure a fair comparison among them. Two observational datasets are used for verification:

1) SPC severe weather reports: The severe weather probabilistic forecasts are verified using the SPC severe weather reports database. This database contains a real-time list of tornado, large hail (larger than 20 mm), and convective wind (stronger than 50 kt; 1 kt = 0.5144 m s^{-1}) damage reports in the United States with information about the intensity of the event and its location in space and time. Figure 1c shows an example of the reports in the SPC database. A gridded field on the MM5 domain is created by setting the grid points with at least one report of severe weather in its grid box to the value of 1. This field does not contain information about the type, intensity, or density of reports within the grid box but it is consistent with the probabilistic forecast that it verifies. The model forecasts strictly refer to the occurrence of severe weather within a grid box rather than to the type, intensity, or density of events.
2) NCEP/CPC stage IV precipitation: To verify the precipitation forecasts the NCEP/Climate Prediction Center (CPC) 4-km stage IV data (Baldwin and Mitchell 1997) are used at 6-h intervals. This dataset is based on a multisensor hourly analysis and it is quality controlled manually. The precipitation remapping from the 4-km Hydrologic Rainfall Anaysis Project (HRAP) grid to the 30-km MM5 grid is performed while maintaining the original amount of precipitation as done by NCEP for grid interpolation and quantitative precipitation forecast (QPF) verification (Mesinger 1996).

In addition to the objective verification against the observational datasets, the relative value of the MM5ADJ is assessed by comparing it against the operational short-range forecasts available for the same period:

1) Subjective day 2 outlooks: After reviewing deterministic model guidance, the SP03 forecaster issued an experimental severe weather outlook for day 2, following the same guidelines used for the routine operational SPC outlooks (Kay and Brooks 2000). The SPC outlooks are issued to forecast the probability of severe weather within 25 mi of a point, which is equivalent to a square area of about 80 km on each side. Admittedly, remapping the subjective outlook probabilities to a 30-km grid produces a shift toward overforecasting since the original probability values that are implicitly calibrated through the forecaster experience and verification results have been shown to be quite reliable. However, it is uncertain how sensitive the forecaster is to this definition when rendering each individual outlook. In addition, the SPC outlooks are issued using five discrete probability categories: 0.00, 0.05, 0.15, 0.25, and 0.35. Again, in order to ensure fair comparison among the forecasts, the model forecasts are truncated into the SPC categories. For instance, all severe weather probability forecasts above 0.35 are considered 0.35 in the verification.
2) Operational Eta: The operational 1200 UTC daily run from the NCEP Eta Model is included to add a reference from a deterministic model into the comparison. Probabilistic forecasts from this model are trivially calculated by setting the field to 1 (0.35 for severe weather forecasts, as this is the highest probability allowed) when the condition to forecast is satisfied and 0 otherwise.
3) NCEP SREF system: The NCEP ensemble for short-range forecasting (e.g., Hamill and Colucci 1997; Wandishin et al. 2001) during SP03 consisted of 10 members: 5 Eta and 5 Regional Spectral Model (RSM) members. The SREF forecasts provide a unique opportunity to compare the experimental MM5ADJ ensemble, which uses human-perturbed ICs, against the dynamical method of breeding of growing modes used in the SREF. Unfortunately, owing to problems with the data archive, only 11 days are available for comparison during the period that the SP03 lasted. However, the available days correspond mainly to the first 2 weeks of May 2003, which was a historically active period of severe weather in the central plains and eastern United States (Schneider et al. 2004). All results obtained from such a small sample of 11 days are complemented with a statistical significance test.
4) Practically perfect prog: Although this field is not a forecast, it is used as a measure of the upper limit of a probabilistic forecast provided some realistic bounds in generating the forecast (e.g., smoothness, size, and spatial continuity of significant probability areas). Brooks et al. (2003) discuss the concept of the practically perfect progs (PPPs) and present the main characteristics. Essentially, the PPP field is constructed by using a nonparametric density estimation function with a two-dimensional Gaussian kernel for each grid point with a report in the observational dataset. This function spreads the probability of occurrence of the event around the grid point. The parameters that define the kernel are calculated by Brooks et al. (2003) from the statistical properties of the climatology of SPC operational outlooks. This hypothetical forecast is as accurate as

could be expected for a forecaster already aware of the reports, given the limitations of real-world forecasting.

3. Verification of severe weather forecasts

Unlike the SPC outlooks, current models do not explicitly forecast severe weather. The diagnosis of severe weather from analysis or models that do not explicitly resolve convection can be inferred, at least in part, through indices that characterize the environment and may allow some basic discrimination of the type or intensity of convective phenomena supported (Thompson et al. 2002). Most severe weather indices consist of a combination of convective instability and shear. The mechanisms of convection triggering are typically overlooked because they often act at scales not resolved by the data. However, models incorporate parameterized convection, which by definition includes a triggering (or activation) function. Although the trigger function is a complex part of the convective schemes (e.g., Kain and Fritsch 1992), it provides an additional component to be considered together with the environmental indices for use in the diagnostic evaluation of severe weather from the model output. Thus, in this study, *severe weather* is defined to occur within a grid box when both the supercell composite parameter (SCP; Thompson et al. 2002) >1 and the triggering of the model's convective scheme occur simultaneously. Together, these two quantities specify regions in which the model jointly predicts an environment that is favorable for supercell thunderstorms, and in which convection develops. The SCP is a nondimensional normalized parameter and is calculated as a combination of the most unstable convective available potential energy (muCAPE) in the column, 0–3-km storm relative helicity (SRH) using the Bunkers et al. (2000) storm motion algorithm, and the bulk Richardson number (BRN) shear (Bluestein 1993):

$$\mathrm{SCP} = \frac{\mathrm{muCAPE}}{1000\ \mathrm{J\ kg^{-1}}} \times \frac{0\text{–}3\ \mathrm{km\ SRH}}{100\ \mathrm{m^2\ s^{-2}}} \times \frac{\mathrm{BRN\ shear}}{40\ \mathrm{m^2\ s^{-2}}}.$$

Hence, the probability of occurrence of severe weather during a 24-h period at every grid point is simply defined as the number of ensemble members having an SCP larger than 1 and simultaneous convective precipitation at that grid point anytime during that 24-h period, divided by the total number of ensemble members. We use the threshold of SCP larger than 1 as it is the value suggested by Thompson et al. (2003) in order to discriminate supercell storm environments in both observed and Rapid Update Cycle-2 analysis–forecast model proximity soundings. For each case, two time periods for the probability of occurrence of severe weather are produced and evaluated from each of the four available forecasts used in the comparison: one for the 0–24-h forecast period and another for the 24–48-h forecast period. Figure 1c shows an example of this probability field from the MM5ADJ, showing for 5 May forecast probabilities of severe weather up to 80% from the lower Mississippi Valley to the Ohio and Tennessee Valleys.

Verification of the probabilistic forecasts for all cases is done by using the attributes diagram. This diagram shows the observed frequency of an event as a function of the forecast category and allows an interpretation of skill for each forecast category separately. The attributes diagram also allows one to interpret the reliability, resolution, and uncertainty of each forecast interval (Wilks 1995). Figure 2 shows the attributes diagram for all the forecasts compared in this study. The sample climatological frequency is 0.016 for the 27 cases and 0.024 for the 11 SREF cases. Not surprisingly for the prediction of unlikely events, all forecasts in the comparison show good skill at predicting no occurrence of severe events (0.00 probabilities), with the human outlooks showing the highest reliability in this category. For low (0.05) and moderate (0.15) probabilities, the MM5ADJ is the only forecast showing some skill, with especially good reliability at the low category. The fact that the MM5ADJ ensemble is underforecasting for both days 1 and 2 in the low category may indicate that a more adequate threshold for the SCP parameter may be needed for this system rather than SCP $>$ 1. For higher probabilities (when a majority of the ensemble members agree), the MM5ADJ shows no skill in predicting severe weather, although some resolution still exists between the 0.25 and 0.35 forecasts. This lack of skill at higher probabilities could be a consequence of overforecasting both convective activity in the model or SCP values, but also may be an indication of underdispersion in the ensemble, perhaps from using a single-model setup. The human outlooks, however, show skill at the high probability categories, revealing the skill of the forecasters when they show high confidence in the intensity of the situation of the day and decide to use high probabilities in the outlook. On the other hand, Eta forecasts are clearly hampered in this type of probabilistic verification, with the model overforecasting severe weather. The deterministic Eta probability of detection is only 11%–12% when SCP $>$ 1 and convective rain occurs in a grid point during the 24-h period.

Regarding the results from the 11 SREF cases, similar scores to those from the 27 days sample are obtained for the SPC outlooks, MM5ADJ, and Eta Model, with some minor but notable differences. The human-

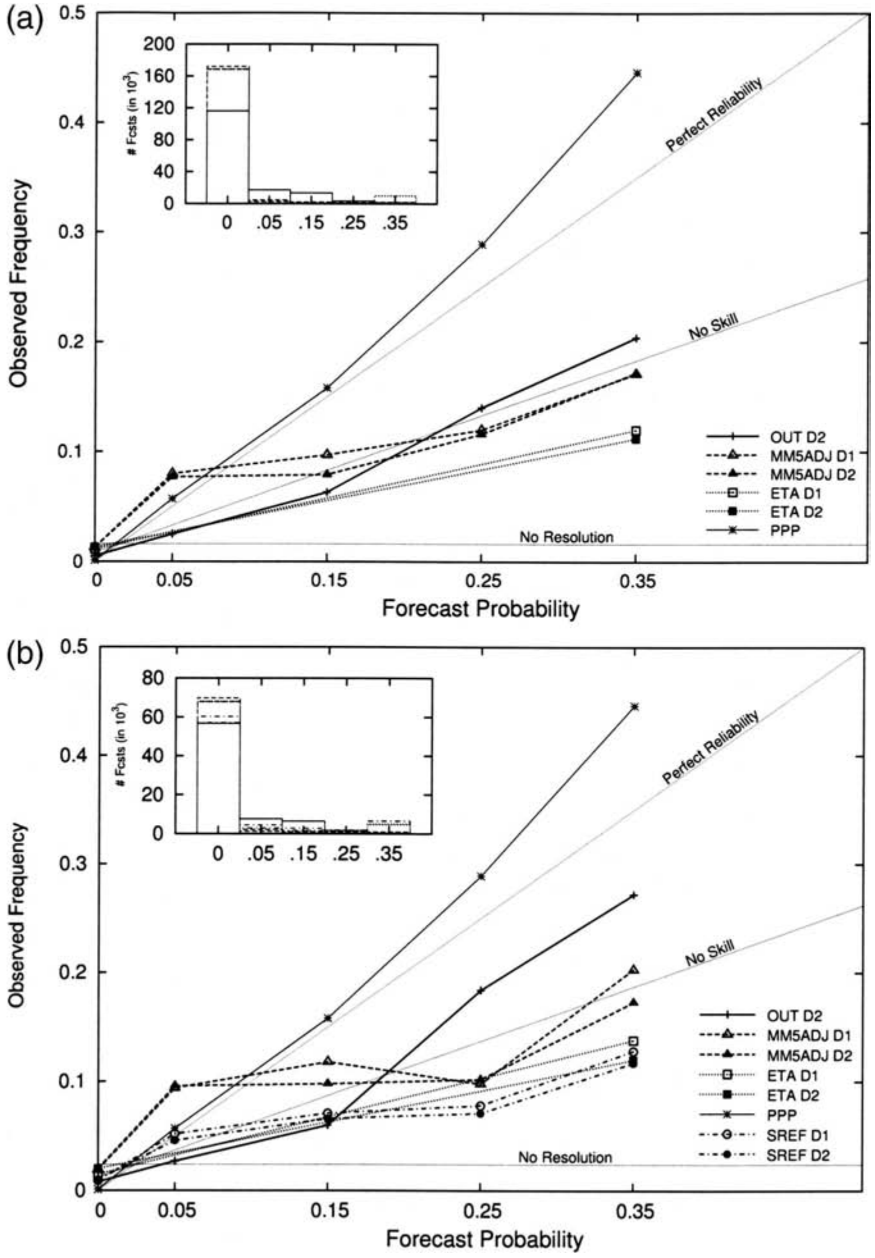

FIG. 2. Attributes diagrams for the probability of severe weather as obtained from SP03 preliminary day 2 outlooks, t + 24 and t + 48 h MM5ADJ, Eta, and SREF. (a) All 27 days of the MM5ADJ experiment are used and (b) results from the subset of the 11 days the SREF output is available.

rendered outlooks again show a remarkable increase in skill and reliability for the high probability categories during this unusually active severe weather period. Also, the MM5ADJ is the only model showing some skill at the 0.35 probability category. Focusing on the SREF results, it shows almost perfect reliability (even better than the PPP for day 1) for the low category but has no skill for higher-probability categories. The significance of the differences between the MM5ADJ and SREF results is assessed using a bootstrap nonparametric test with 10 000 resamples (Wilks 1995). Differences between MM5ADJ and SREF visible in Fig. 2b are all significant to the 99% confidence level, except for the 25% category for D1. For this particular case, observed frequency is significantly larger than SREF at a 95% confidence level. This result clearly shows the advantage of the MM5ADJ over the SREF in forecasting probabilities of severe weather at and above 0.15, usu-

TABLE 2. Mean of the std dev computed at sounding sites. Global values are averaged over the CONUS east of the Rockies, and the targeted include only the areas delineated by the forecaster when defining the perturbations. Values in parentheses indicate the percent change from the global to the targeted standard verification. Here, Q_{ll} refers to the average of the std dev of Q at 1000, 850, and 700 hPa.

	MM5				SREF			
	Global		Targeted		Global		Targeted	
Variable	24 h	36 h	24 h	36 h	24 h	36 h	24 h	36 h
T_{850}	0.63	0.83	1.24 (+96%)	1.40 (+68%)	0.87	1.01	1.00 (+15%)	1.25 (+24%)
T_{700}	0.45	0.62	0.75 (+66%)	0.96 (+56%)	0.80	0.97	0.77 (−3%)	1.01 (+5%)
T_{500}	0.39	0.52	0.52 (+33%)	0.71 (+37%)	0.64	0.74	0.58 (−10%)	0.78 (+6%)
T_{250}	0.41	0.51	0.48 (+17%)	0.61 (+19%)	0.76	0.87	0.78 (+2%)	0.899 (+3%)
Q_{ll}	0.55	0.78	1.06 (+95%)	1.29 (+64%)	0.79	0.88	0.86 (+9%)	0.96 (+8%)
Q_{850}	0.71	0.99	1.36 (+90%)	1.63 (+64%)	0.86	0.93	1.03 (+20%)	1.07 (+15%)
Q_{700}	0.50	0.72	0.91 (+83%)	1.06 (+49%)	0.72	0.84	0.73 (+2%)	0.87 (+3%)
U_{850}, V_{850}	1.54	1.97	2.90 (+88%)	3.31 (+68%)	1.95	2.09	2.27 (+16%)	2.46 (+18%)
U_{700}, V_{700}	1.51	1.96	2.55 (+69%)	2.91 (+48%)	1.93	2.17	2.06 (+7%)	2.36 (+9%)

ally associated with the more intense and damaging episodes. This is most likely a consequence of the customized design of the MM5ADJ to focus on the areas of severe weather threat, whereas the SREF system is designed to cover a wide range of mesoscale forecast aspects and shows its strength at the low-probability range.

Note that in both panels of Fig. 2, differences between results from day 1 (D1) and day 2 (D2) forecasts are small for all models. Finally, as expected, besides the SREF is forecasting the low category with great success for this small sample size, all forecasts considered in this comparison are far from the hypothetical limit set by the PPP field.

Targeted spread

To better understand the differences between the MM5ADJ and SREF systems in forecasting higher-probability (≥25%) episodes of severe weather, we analyze the ability of the MM5ADJ to generate spread specifically over the areas of concern defined by the forecaster. The spread is evaluated using the standard deviation around the mean of various model prognostic fields. Two versions of the spread for each model are computed (Table 2): the *global* spread is the mean of the spread calculated at each sounding site[1] within the CONUS, east of the Rocky Mountains; the *targeted* spread is computed considering only sounding sites within the areas of concern and times designated by the forecaster in constructing the ensemble. Table 2 shows the spread at 24 and 36 h (within SPC day 2) for the fields of temperature, specific humidity, and wind. While the global spread generally is larger in the SREF than in the MM5ADJ, the values of targeted spread between the two ensemble systems are very similar in size. This suggests that the amplitudes of the original perturbations applied to the ICs in the MM5ADJ are reasonable and do not yield spread that is too large. Both the MM5ADJ and SREF systems also show an increase in spread from 24 to 36 h, revealing that the dispersion is not yet saturated in either of the ensembles and continues to grow after 24 h. However, the relative increase of spread from the global to the targeted spread is much larger in the MM5ADJ than in the SREF, especially in low levels where increases in spread ranging from 65% to 95% are obtained. In addition, the spread in the targeted regions is always larger than the global spread in the MM5ADJ, whereas this is not true for the SREF. Therefore, the breeding vectors technique (as well as the model diversity) in the SREF system produces larger dispersion in a global sense, whereas the customized MM5ADJ successfully targets ensemble dispersion both spatially and temporally over the region selected by the forecaster.

4. Verification of precipitation forecasts

a. *Bias and equitable threat score*

The 6-hourly accumulated precipitation forecasts from the MM5ADJ, Eta, and SREF are verified using the NCEP/CPC stage IV dataset. Although probabilistic fields can be derived from an ensemble of precipitation forecasts, we first compute verification scores of deterministic-like fields from the ensembles, such as the mean and the probability mean matched (PMM) precipitation. The PMM (Ebert 2001) is calculated as the

[1] The calculation of the standard deviation does not involve the observed sounding data, but it is found that averaging the model fields at just the sounding locations reduces problems that can arise from the spatial correlation of the data.

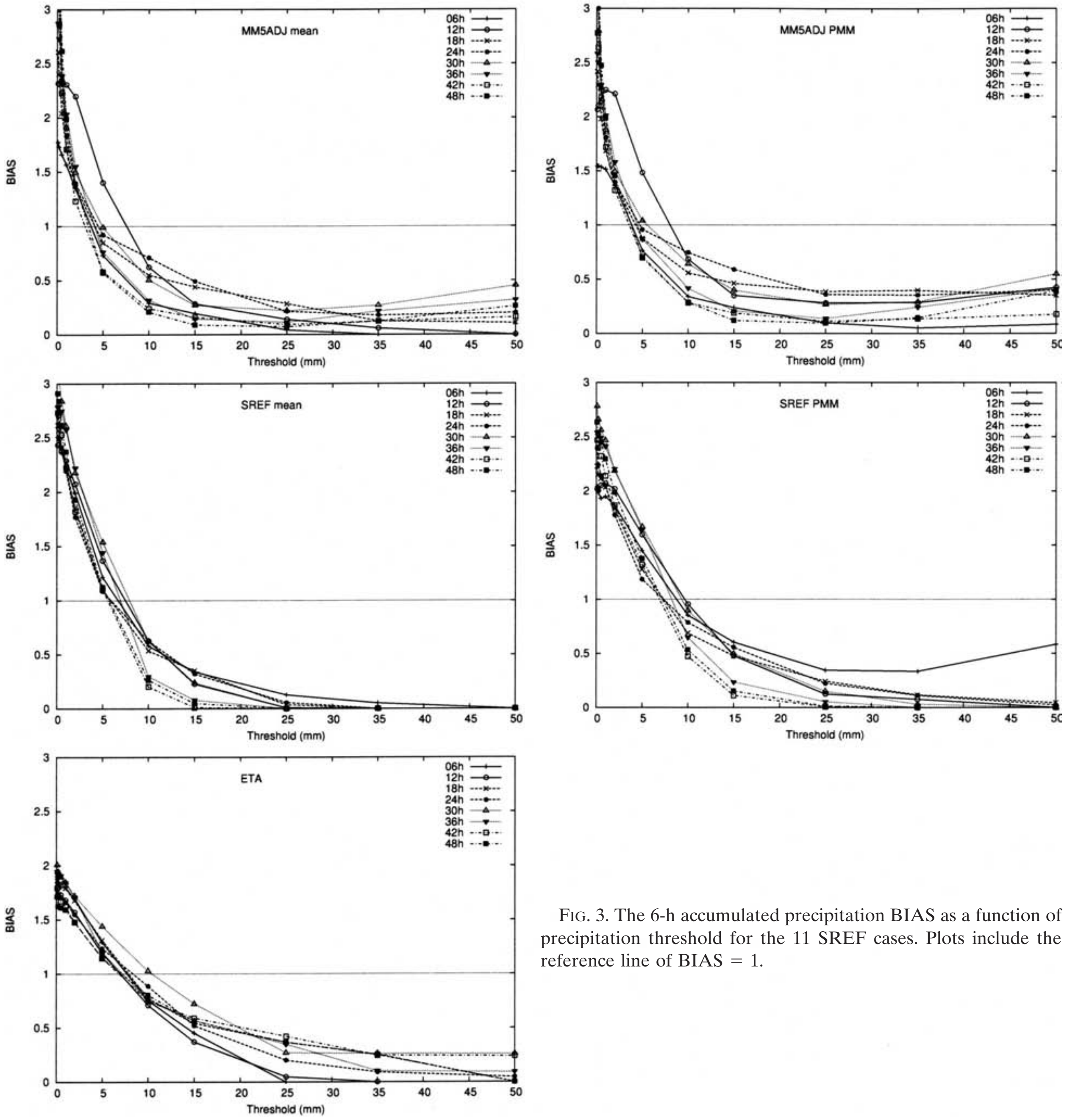

FIG. 3. The 6-h accumulated precipitation BIAS as a function of precipitation threshold for the 11 SREF cases. Plots include the reference line of BIAS = 1.

ensemble mean, rescaled locally using the global (all ensemble members) distribution of precipitation amounts. This field is intended to be representative of, and to exhibit the properties (in terms of smoothing out uncertain features from the individual members) of the ensemble mean, but produces a distribution of precipitation amounts similar to the precipitation field of an individual deterministic member.

To provide global skill scores for the single-field precipitation forecasts, we compute the bias (BIAS) and equitable threat score (ETSs). For the sake of brevity, only results for the 11 SREF cases are presented. The BIAS for 11 precipitation thresholds is calculated for each 6-h interval as

$$\mathrm{BIAS} = \frac{F}{O},$$

where F is the number of forecast points above the threshold and O the number of observed points above the threshold.

All systems overpredict (BIAS > 1) amounts less than 5 mm, though the Eta Model shows the lowest bias (Fig. 3). As expected, for larger amounts, all models underpredict (BIAS < 1) precipitation, generally decaying to smaller bias with increasing threshold. Noticeably, the bias of both the mean and PMM SREF precipitation fields decay rapidly toward zero whereas the Eta, and especially the MM5ADJ, fields do not decay as fast for larger precipitation thresholds (>25 mm 6 h^{-1}). It is apparent by comparing the mean and PMM results, from both MM5ADJ and SREF systems, that the PMM in general provides a slightly better BIAS than the mean, especially at larger thresholds where the mean is reduced more strongly through averaging.

To ensure a fair comparison of the precipitation forecasts among the models, the bias is corrected using a rescaling technique on the local precipitation amounts. For each case, threshold, and 6-h period, the bias (*F/O*) of all remaining cases in the dataset (10 days) is computed. Then, an unbiased threshold for the model precipitation is searched for iteratively until a value is found where $F_{\text{unbiased Thrs}} = O_{\text{biased Thrs}}$ and a correction factor (equal to bias-corrected threshold/original threshold) is obtained for each precipitation threshold. The bias correction is then applied by multiplying the original precipitation amount by this correction factor. A linear interpolation is used to calculate the correction factor for values between thresholds. As expected, the bias correction basically decreases (increases) the precipitation amounts below (above) 4–5 mm. Not surprisingly, mean fields require a larger bias correction than PMM at higher thresholds due to the smoothing effect of averaging.

The day 1 results from the MM5ADJ system indicate that even for ensemble mean PMM-adjusted precipitation forecasts, the MM5ADJ appears to occasionally improve upon the SREF. This is promising, since the experiment was not designed to focus upon precipitation forecasts. However, in retrospect this result perhaps is not too surprising, since heavy precipitation amounts are usually linked to active mesoscale convective systems that are often associated with severe weather. Also note that most of the MM5ADJ improvements over the SREF occur in the first 30 h and the forecasters most often selected forecast times of 36 h or less in creating the ensemble perturbations. Considering the smaller amounts of spread in the MM5ADJ seen outside of the forecaster-targeted areas in comparison with the SREF, this positive result is encouraging.

The remaining bias in the precipitation fields after the correction is shown in Fig. 4. Besides the very good bias values for low precipitation amounts from the corrected Eta forecasts, the PMM fields show generally better results from the correction at low precipitation values than the ensemble mean. However, for high thresholds, the bias correction is less effective, with the MM5ADJ mean and the SREF PMM resulting in slightly better biases than the other models in that range. Unfortunately, the number of observation–forecast pairs available to estimate the bias correction factor is limited and the significance of these differences is weak, producing undesired results at high amount thresholds.

The ETS provides a good global skill score intended to minimize the impact of biases in the evaluation of precipitation forecasts. The score is computed as

$$\text{ETS} = \frac{C - E}{F + O - C - E} \text{ with E} = \frac{F \times O}{T},$$

where *C* is the number of points with both forecast and observations above a threshold and *T* is the total number of grid points in the forecast. Values of ETS range from 0 to 1, with larger values implying a more accurate forecast. Since the lower thresholds of precipitation are the most populated and the PMM field allows for a better bias correction than the ensemble mean at these thresholds, only the PMM fields are shown in the comparison of ETS values (Fig. 5). The SREF PMM and the Eta produce higher ETSs than the MM5ADJ for low (<2 mm 6 h^{-1}) precipitation amounts. As expected, the ETS decreases for all models as the threshold increases; however, the loss of skill is more rapid for the SREF PMM and Eta, such that the MM5ADJ PMM produces the best forecasts for amounts above 25 mm 6 h^{-1} for day 1. Table 3 summarizes the results of the bootstrap significance test using 10 000 resamples for the differences between MM5ADJ and SREF ETS scores. Deterministic precipitation forecasts from the MM5ADJ system are clearly degraded on day 2 for all precipitation thresholds. Day 1 results from the MM5ADJ system are consistent with the conclusions in the previous section, since heavy precipitation amounts are usually linked to active mesoscale systems that often produce severe weather reports.

Admittedly, the differences in skill of precipitation forecasts among the systems may be a consequence of the different microphysics and convective schemes used in the models. The Kain–Fritsch convective scheme (Kain and Fritsch 1990) has been shown to produce useful forecasts of maximum rainfall amounts, although the location of the heaviest rainfall may be displaced by several hundred kilometers (Gallus 1999). In contrast, the Betts–Miller–Janjić convective adjustment (Janjić 1994) used in five of the SREF members has been shown to produce broad areas of rainfall with better

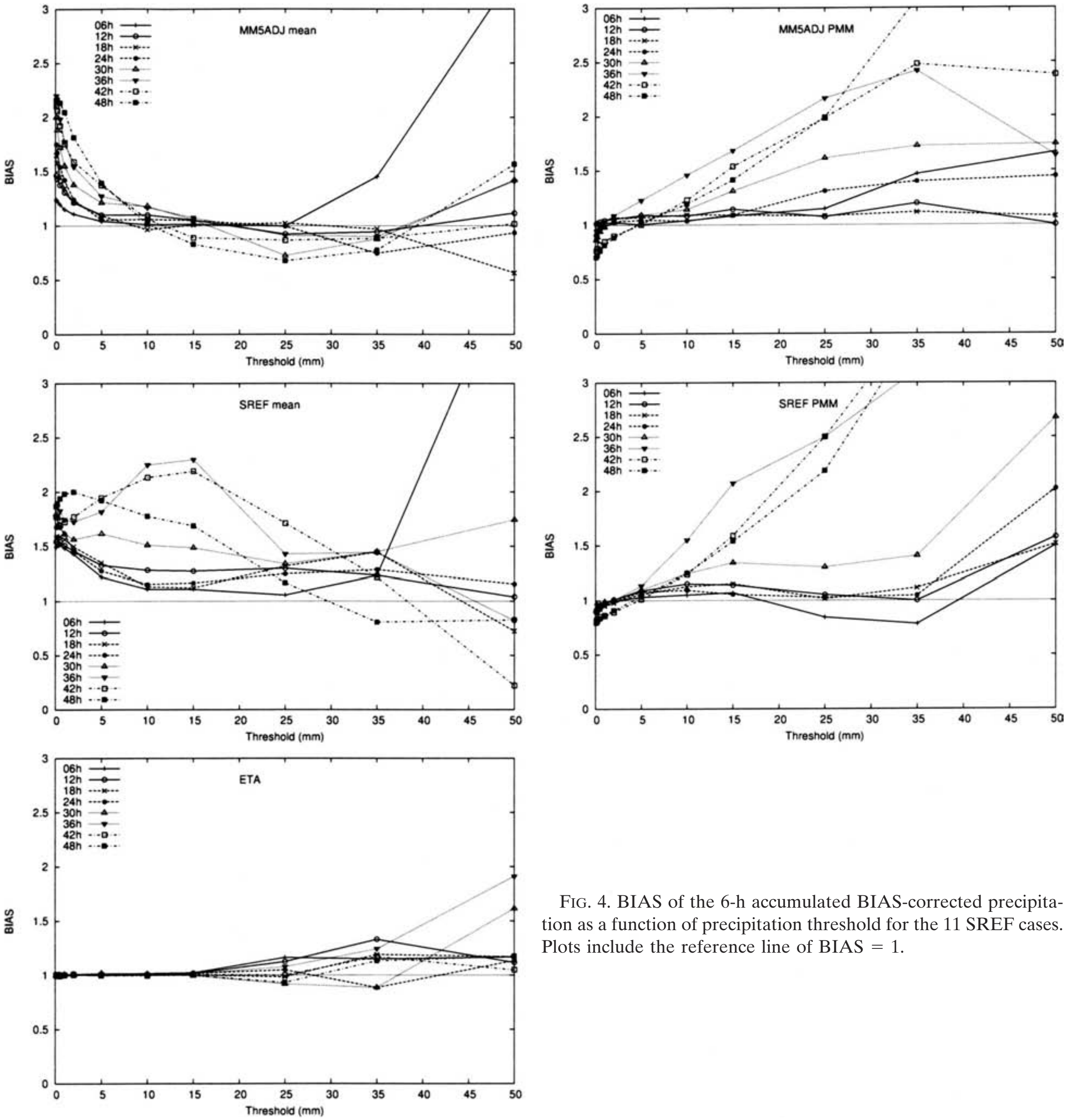

FIG. 4. BIAS of the 6-h accumulated BIAS-corrected precipitation as a function of precipitation threshold for the 11 SREF cases. Plots include the reference line of BIAS = 1.

skill for lower precipitation amounts, although it may miss the heavier rainfall amounts (Bright and Mullen 2002; Wang and Seaman 1997). Nevertheless, our results suggest that the targeting of ensemble dispersion over areas with predicted active convection tends to improve the forecasts of high precipitation amounts more than the forecasts of low precipitation amounts, as reflected in the ETS score results. Since high precipitation amounts present a more significant flood threat to the public, the MM5ADJ appears to be providing useful deterministic information on the precipitation threats for day 1.

b. Probabilistic forecasts

Besides the single-field (deterministic like) precipitation forecasts, ensemble results can account for the uncertainty in the forecast. A common way to express the uncertainty is to forecast the probability of an event as determined by the frequency of occurrence among the ensemble members. Thus, the skill of the forecasting

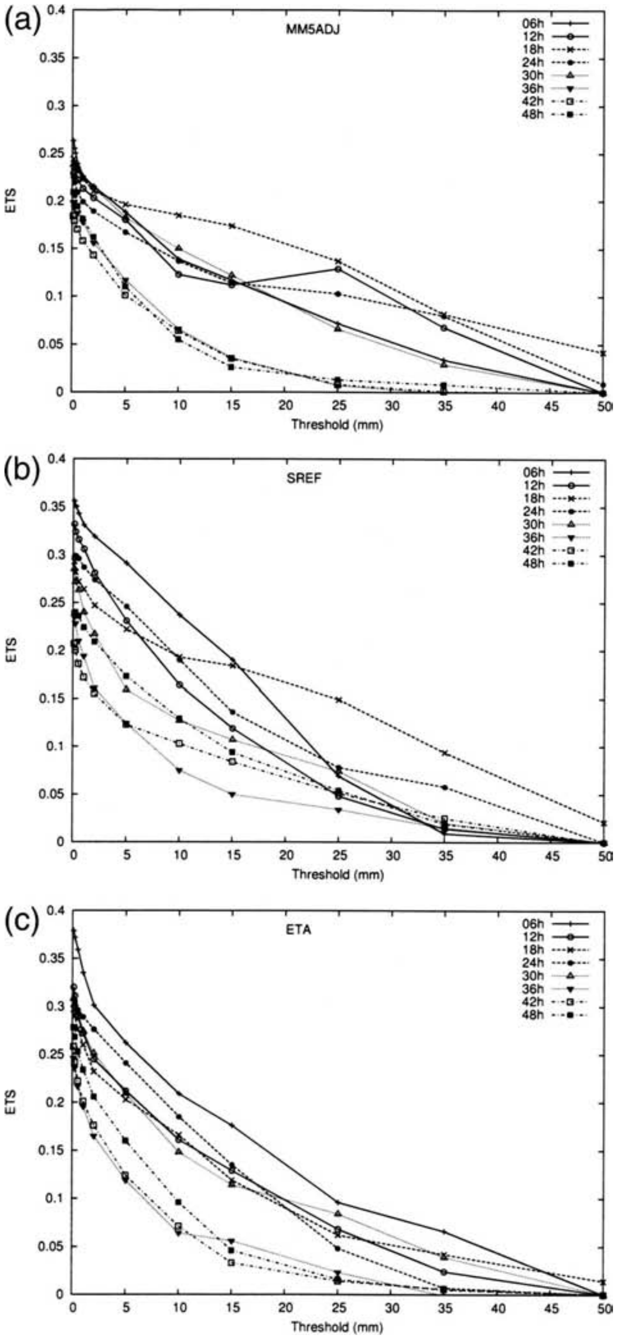

FIG. 5. ETS of the BIAS-corrected PMM 6-h accumulated precipitation as a function of precipitation threshold for the 11 SREF cases.

TABLE 3. Significance test results for ETS values shown in Figs. 5a and 5b. The S or M indicates that SREF or MM5ADJ produces a significantly larger ETS score than the other at a level of 95% confidence. Dashes indicate not significantly different ETS values for the two systems.

	Precipitation threshold (mm)								
Forecast	0.1	1	2	5	10	15	25	35	50
06 h	S	S	S	S	S	S	—	M	—
12 h	S	S	S	S	S	—	M	M	—
18 h	S	S	S	S	—	—	—	—	M
24 h	S	S	S	S	S	S	M	M	M
30 h	S	S	—	M	M	M	—	—	—
36 h	S	S	—	S	S	S	S	S	—
42 h	S	S	S	S	S	S	S	S	—
48 h	S	S	S	S	S	S	S	S	—

systems in predicting the probabilities of the bias-adjusted 6-h accumulated precipitation above 1, 20, and 35 mm is evaluated. The skill of the probabilistic forecasts is compared using an attributes diagram for each threshold (Fig. 6) and the statistical significance of the results is tested by means of a bootstrap nonparametric method to a 95% confidence level for the differences. The probability of 6-h precipitation above 1 mm has good reliability in both ensembles for forecasts up to the 30% category, with the MM5ADJ being significantly more reliable than SREF for the 10% and 30% categories. For higher probabilities, the SREF forecasts are better, with skill and resolution for all categories. Results further indicate that as the accumulated precipitation amounts increase, the skill of all the systems decreases. In fact, for all forecast categories, none of the systems produce skillful forecasts for precipitation amounts greater than 20 mm 6 h^{-1} (Figs. 6b and 6c). The MM5ADJ show small but significant differences with respect to the SREF forecasts for the 10%, 30%, and 50% categories of 20 mm 6 h^{-1}. For higher probabilities, the SREF shows a remarkable resolution, providing valuable forecasts (i.e., observed frequency corresponding to a forecast category) of 40% when 100% is predicted by the ensemble.

Forecasts of more than 35 mm are less reliable than those of 20 mm, but the MM5ADJ shows some resolution for forecasts up to 50% and provides significantly more reliable forecasts than SREF for the 30%, 50%, and 70% category. However, the skill of the high probability forecasts from the MM5ADJ is low. This reduced reliability is a reflection of high agreement among ensembles on an erroneous forecast. One reason for this result may be that the model incorrectly forecasts heavy precipitation amounts in areas far from the area of concern defined by the forecaster, so that the dispersion in the model is low and high probabilities are forecasted.

Comparison of the three panels in Fig. 6 reveals that, although the loss of skill as the precipitation amount increases is observed in both systems, the MM5ADJ produces significantly better probabilistic forecasts of heavy precipitation amounts for cases with moderate probabilities (indicating that there is some dispersion,

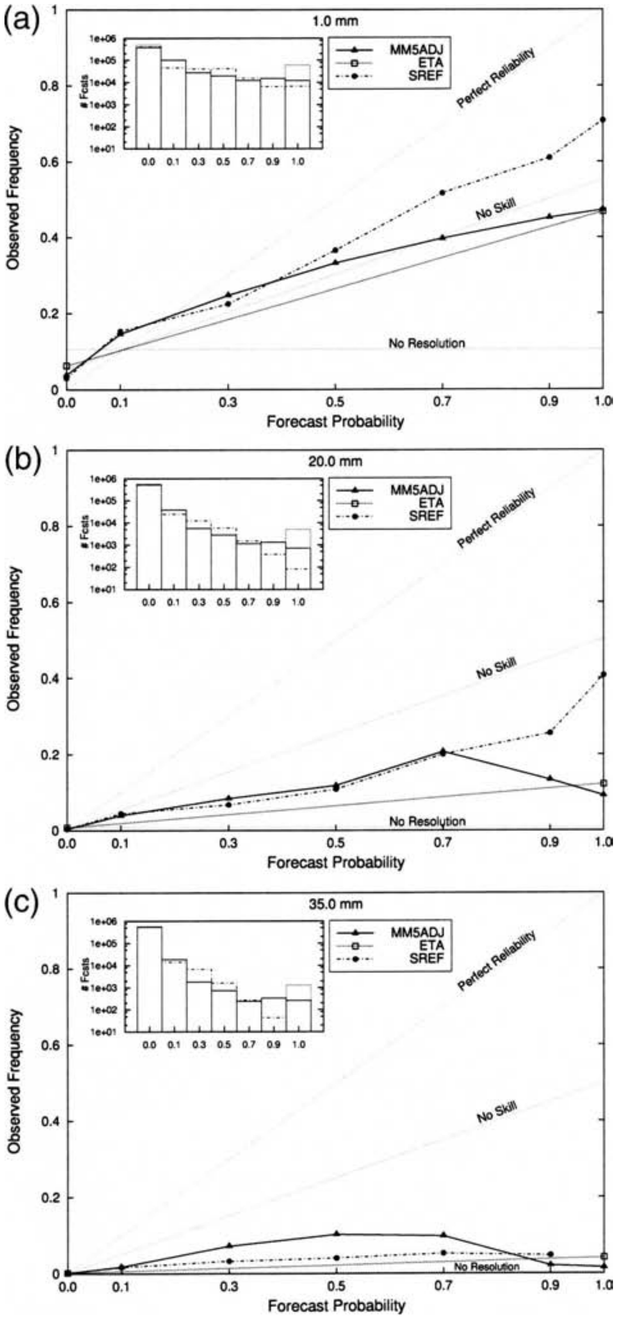

FIG. 6. Attributes diagrams for 6-h accumulated precipitation thresholds at (a) 1, (b) 20, and (c) 35 mm. Note the logarithmic scale in the inset histograms.

or uncertainty, in the ensemble). This is not unexpected from an ensemble constructed with a single model and designed to generate dispersion over specific areas with the potential for severe convective activity (and likely heavy precipitation amounts). Areas not targeted by the forecaster when defining the perturbations will likely have little or no dispersion.

5. Mixed ensemble test

Severe weather and precipitation verification scores for the MM5ADJ experimental ensemble reveal the value of creating the ensemble that targets areas of specific concern during the forecast time interval. The experimental ensemble forecasts severe weather with some skill and produces some better ETS and probabilistic forecasts for heavy precipitation amounts than the operational SREF system. However, we hypothesize that the lack of dispersion in areas not selected by the forecaster results in inferior forecasts of low-intensity severe episodes and low precipitation amounts. On the other hand, the SREF system shows better skill in forecasting low-intensity events but produces unreliable high probability forecasts of severe weather and worse probabilistic forecasts of heavy precipitation amounts than MM5ADJ.

To test the effect of adding members to the MM5ADJ system that provide spread across the entire domain, we evaluate the forecast skill of an ensemble generated by combining the 32 MM5ADJ and 10 SREF members to produce a 42-member ensemble (42ENS). This ensemble not only will benefit from a large number of members but also from being multimodel and including two initial conditions perturbation techniques. This ensemble is still primarily focused on targeting severe weather but may also benefit from the globally better scores of the 10 SREF members.

Severe weather forecasts are produced for the 42ENS following the same method presented in section 3 (Fig. 7). The bootstrap nonparametric test is also used to assess the significance of the differences between the 42ENS and MM5ADJ results. The attributes diagram curve for the 42ENS forecasts shows the superior skill of this configuration as compared to the MM5ADJ for almost all probabilities. For the 15% and 30% categories only, the 42ENS does not produce results significantly better than MM5ADJ for both days 1 and 2. It is noteworthy that the 42ENS produces the best model forecasts for the 25% category and on day 1 the 42ENS forecast is the only one of all considered forecasts in this comparison with some skill at all forecast categories.

These results suggest that the combination of systems produces a positive synergism in the forecast of severe weather. One mechanism to correct the overprediction initially obtained at high probabilities for the MM5ADJ is by adding members that do not forecast as much convective activity so that the resulting probabilities are generally lower and the overprediction is to some extent alleviated.

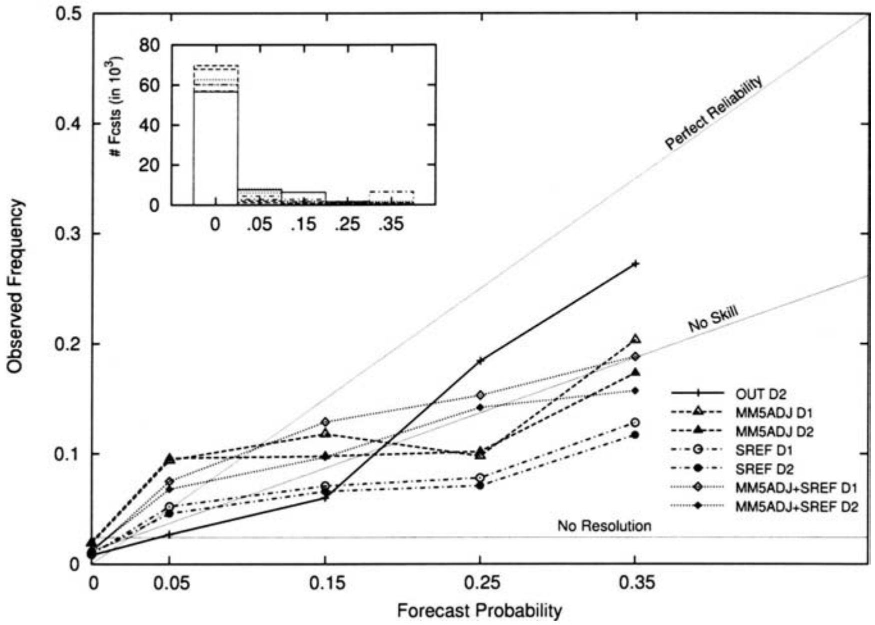

FIG. 7. As in Fig. 2b but for the 42ENS system.

6. Discussion

The SPC/NSSL SP03 included an experimental ensemble aimed at testing for an extended period of time the ensemble generation method of Xu01 who proposed using human forecasters to identify atmospheric features they believed to be important to the development and evolution of severe weather during the 24–48-h forecast period. Using an adjoint model, perturbations to the forecast model initial conditions that would influence these forecaster-selected atmospheric features are identified and used to create an ensemble of model forecasts. The performance of this experimental ensemble is evaluated by using severe weather reports and 6-h accumulated precipitation observations.

The experiment was designed to run in real time, with the initial hope that forecasters would have time to examine and verify the ensemble forecasts and gain experience in selecting the perturbation fields, vertical levels, and areas. Unfortunately, computer limitations did not allow for this learning experience to happen as the forecasts were available too late in the day. Thus, the forecasters were only given basic guidance on how to generate the perturbations. Many other aspects of the experiment also are imperfect and should be improved upon in future experiments. Yet the initial results are promising and warrant careful consideration.

Verification results show value in the experimental ensemble forecasts compared to the operational SREF system, despite the multiple improvements still possible to the experimental system. A single model is used in the experiment, with the human-selected perturbations the only source of dispersion in the ensemble system. Although basic training was provided at the beginning of each experimental week of SP03 covering the selection of fields, levels, sizes, and time of the targeted structures, no definitive rules were made available to the forecasters on the construction of perturbations, because this had never before been conducted as a real-time experiment. Additionally, the forecasters had no previously experience with this type of ensemble creation and no quantitative feedback was provided to them during the experiment. Further research might indicate whether certain sizes, fields, and levels are more appropriate to define the perturbations for specific types of predicted weather.

Despite the lack of previous knowledge and experience using this technique, the experimental ensemble is shown to improve the numerical forecasts of severe weather, arguably because it successfully generates dispersion over the areas of concern selected by the forecaster. The system also produces better probabilistic forecasts of heavy precipitation than the SREF. However, the experimental ensemble forecasts of low probability severe weather and low amounts of precipitation have less skill than those of the SREF and the operational Eta. A clear conclusion from these results is that this ensemble, customized to exclusively focus on high-intensity and damaging weather, lacks global dispersion and produces unreliable forecasts for nonhazardous weather events. Results from an ensemble constructed by combining globally perturbed members (from SREF) and humanly perturbed members (from MM5ADJ) show promising skill for the forecast of severe weather. While the experimental setup was not perfect, the results indicate that the value of human beings in the creation of ensembles designed to target specific weather threats is potentially large. Further investigation of the potential value of humans being part of the ensemble process is strongly recommended, even if the end result is to learn how forecasters can provide real-time input into an automated ensemble generation system. We still have a lot to learn about how to create ensembles for short-range forecasts of high-impact weather, and we need to make better use of the skill and experience of human forecasters in this learning process.

Acknowledgments. We thank the effort of all SPC forecasters who thoughtfully collaborated on this project. We are also grateful to Dr. Harold Brooks for the assistance provided on the practically perfect prog fields and to Dr. Michael Baldwin for his helpful comments on bias correction and for providing the code to process precipitation fields. The adjoint calculations were performed on the Linux supercomputer located at the Oklahoma Supercomputer Center for Education and Research (OSCER) at the University of Oklahoma. This work was performed while the lead author held a National Research Council Research Associateship Award at the NOAA/National Severe Storms Laboratory.

REFERENCES

Baldwin, M. E., and K. E. Mitchell, 1997: The NCEP hourly multisensor U.S. precipitation analysis for operations and GCIP research. Preprints, *13th Conf. on Hydrology,* Long Beach, CA, Amer. Meteor. Soc., 54–55.

——, J. S. Kain, and M. P. Kay, 2002: Properties of the convection scheme in NCEP's Eta Model that affect forecast sounding interpretation. *Wea. Forecasting,* **17,** 1063–1079.

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

Bright, D. R., and S. L. Mullen, 2002: Short-range ensemble forecasts of precipitation during the southwest monsoon. *Wea. Forecasting,* **17,** 1080–1100.

Brooks, H., C. A. Doswell III, and M. P. Kay, 2003: Climatological estimates of local daily tornado probability for the United States. *Wea. Forecasting,* **18,** 626–640.

Buizza, R., and T. N. Palmer, 1995: The singular vector structure of the atmospheric general circulation. *J. Atmos. Sci.,* **52,** 1434–1456.

Bunkers, M. J., B. A. Klimowski, J. W. Zeitler, R. L. Thompson, and M. L. Weisman, 2000: Predicting supercell motion using a new hodograph technique. *Wea. Forecasting,* **15,** 61–79.

Colle, B. A., J. B. Olson, and J. S. Tongue, 2003: Multiseason verification of the MM5. Part I: Comparisons with the Eta Model over the central and eastern United States and impact of MM5 resolution. *Wea. Forecasting,* **18,** 431–457.

Davis, C. A., K. W. Manning, R. E. Carbone, S. B. Trier, and J. D. Tuttle, 2003: Coherence of warm-season continental rainfall in numerical weather prediction models. *Mon. Wea. Rev.,* **131,** 2667–2679.

Du, J., and Coauthors, 2004: The NOAA/NWS/NCEP short-range ensemble forecast (SREF) system: Evaluation of an initial condition vs multi-model physics ensemble approach. Preprints, *16th Conf. on Numerical Weather Prediction,* Seattle, WA, Amer. Meteor. Soc., CD-ROM, 21.3.

Dudhia, J., 1989: Numerical study of convection observed during the Winter Monsoon Experiment using a mesoscale two-dimensional model. *J. Atmos. Sci.,* **46,** 3077–3107.

——, 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.

——, 1996: A multi-layer soil temperature model for MM5. Preprints, *Sixth PSU/NCAR Mesoscale Model User's Workshop,* Boulder, CO, 49–50.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.,* **129,** 2461–2480.

Ehrendorfer, M., 1994: The Liouville equation and its potential usefulness for the prediction of forecast skills. Part I: Theory. *Mon. Wea. Rev.,* **122,** 703–713.

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

Errico, R. M., 1997: What is an adjoint model? *Bull. Amer. Meteor. Soc.,* **78,** 2577–2591.

Evans, R. E., M. S. J. Harrison, R. J. Graham, and K. R. Mylne, 2000: Joint medium-range ensembles from the Met. Office and ECMWF systems. *Mon. Wea. Rev.,* **128,** 3104–3127.

Fowle, M. A., and P. J. Roebber, 2003: Short-range (0–48 h) numerical prediction of convective occurrence, mode, and location. *Wea. Forecasting,* **18,** 782–794.

Funk, T. W., 1991: Forecasting techniques utilized by the forecast branch of the National Meteorological Center during a major convective rainfall event. *Wea. Forecasting,* **6,** 548–564.

Gallus, W. A., Jr., 1999: Eta simulations of three extreme precipitation events: Sensitivity to resolution and convective parameterization. *Wea. Forecasting,* **14,** 405–426.

Gelaro, R., R. Buizza, T. N. Palmer, and E. Klinker, 1998: Sensitivity analysis of forecast errors and the construction of optimal perturbations using singular vectors. *J. Atmos. Sci.,* **55,** 1012–1037.

Gilmore, M. S., J. M. Straka, and E. N. Rasmussen, 2004: Precipitation uncertainty due to variations in precipitation particle parameters within a simple microphysics scheme. *Mon. Wea. Rev.,* **132,** 2610–2627.

Grell, G. A., J. Dudhia, and D. R. Stauffer, 1994: A description of the fifth-generation Penn State/NCAR Mesoscale Model (MM5). NCAR Tech. Note NCAR/TN-398+STR, 117 pp.

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

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.

Johns, R. H., and C. Doswell III, 1992: Severe local storms forecasting. *Wea. Forecasting,* **7,** 588–612.

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.

——, and ——, 1992: The role of convective "trigger function" in numerical forecasts of mesoscale convective systems. *Meteor. Atmos. Phys.,* **49,** 93–106.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability.* Cambridge University Press, 342 pp.

Kay, M. P., and H. E. Brooks, 2000: Verification of probabilistic severe storm forecasts at the SPC. Preprints, *20th Conf. on Severe Local Storms,* Orlando, FL, Amer. Meteor. Soc., 285–288.

Langland, R. H., R. Gelaro, G. D. Rohaly, and M. Shapiro, 1999a: Targeted observations in FASTEX: Adjoint-based targeting procedures and data impact experiments in IOP17 and IOPl8. *Quart. J. Roy. Meteor. Soc.,* **125,** 3241–3270.

——, and Coauthors, 1999b: The North Pacific Experiment (NORPEX-98): Targeted observations for improved North American weather forecasts. *Bull. Amer. Meteor. Soc.,* **80,** 1363–1384.

Leftwich, P. W., Jr., J. T. Schaefer, S. J. Weiss, and M. Kay, 1998: Severe convective storm probabilities for local areas in watches issued by the Storm Prediction Center. Preprints, *19th Conf. on Severe Local Storms,* Minneapolis, MN, Amer. Meteor. Soc., 548–551.

Legg, T. P., and K. R. Mylne, 2004: Early warnings of severe weather from ensemble forecast information. *Wea. Forecasting,* **19,** 891–906.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.,* **102,** 401–418.

McCarthy, D. J., J. T. Schaefer, and M. Kay, 1998: Watch verification at the Storm Prediction Center 1970–1997. Preprints, *19th Conf. on Severe Local Storms,* Minneapolis, MN, Amer. Meteor. Soc., 548–551.

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

Olson, D. A., N. W. Junker, and B. Korty, 1995: Evaluation of 33 years of quantitative precipitation forecasting at the NMC. *Wea. Forecasting,* **10,** 498–511.

Penland, C., 2003: A stochastic approach to nonlinear dynamics. *Bull. Amer. Meteor. Soc.,* **84,** doi:10.1175/BAMS-84-7-Penland.

Schneider, R. M., H. E. Brooks, and J. T. Schaefer, 2004: Tornado outbreak day sequences: Historic events and climatology (1875–2003). Preprints, *22d Conf. on Severe Local Storms,* Hyannis, MA, Amer. Meteor. Soc., CD-ROM, P12.1.

Shapiro, M. A., and A. J. Thorpe, 2004: THORPEX International Science Plan., Version III. WMO Tech. Rep. WMO/TD-No. 1246, WWRP/THORPEX No. 2, World Meteorological Organization, Geneva, Switzerland, 51 pp.

Stensrud, D. J., and J. M. Fritsch, 1993: Mesoscale convective systems in weakly forced large-scale environments. Part I: Observations. *Mon. Wea. Rev.,* **121,** 3326–3344.

——, and L. Wicker, 2004: On the predictability of mesoscale convective systems. *Int. Conf. on Storms,* Brisbane, Australia, Australian Meteorological and Oceanographic Society, 62–67.

——, J.-W. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.,* **128,** 2077–2107.

Thompson, R. L., R. Edwards, and J. A. Hart, 2002: Evaluation and interpretation of the supercell composite and significant tornado parameters at the Storm Prediction Center. Preprints, *21st Conf. on Severe Local Storms,* San Antonio, TX, Amer. Meteor. Soc., J11–J14.

——, ——, ——, K. L. Elmore, and P. Markowski, 2003: Close proximity soundings within supercell environments obtained from the Rapid Update Cycle. *Wea. Forecasting,* **18,** 1243–1261.

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

——, and ——, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.,* **125,** 3297–3319.

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

Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev.,* **129,** 129–141.

Wang, W., and N. L. Seaman, 1997: A comparison study of convective parameterization schemes in a mesoscale model. *Mon. Wea. Rev.,* **125,** 252–278.

Weiss, S. J., J. S. Kain, J. J. Levit, M. E. Baldwin, and D. R. Bright, 2004: Examination of several different versions of the WRF model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. Preprints, *20th Conf. on Severe Local Storms,* Hyannis, MA, Amer. Meteor. Soc., CD-ROM, 17.1.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction.* International Geophysics Series, Vol. 59, Academic Press, 467 pp.

Xu, M., D. J. Stensrud, J.-W. Bao, and T. T. Warner, 2001: Applications of the adjoint technique to short-range ensemble forecasting of mesoscale convective systems. *Mon. Wea. Rev.,* **129,** 1395–1418.

Zamora, R. J., and Coauthors, 2003: Comparing MM5 radiative fluxes with observations gathered during the 1995 and 1999 Nashville southern oxidant studies. *J. Geophys. Res.,* **108,** 4050, doi:10.1029/2002JD002122.

Zapotocny, T. H., and Coauthors, 2000: A case study of the sensitivity of the Eta Data Assimilation Scheme. *Wea. Forecasting,* **15,** 603–621.

Zou, X., F. Vandenberghe, M. Pondeca, and Y.-H. Kuo, 1997: Introduction to adjoint techniques and the MM5 adjoint modeling system. NCAR Tech. Note NCAR/TN-435+IA, 120 pp.

——, W. Huang, and Q. Xiao, 1998: A user's guide to the MM5 adjoint modeling system. NCAR Tech. Note NCAR/TN-437+IA, 45 pp.