Objective Verification of the SAMEX '98 Ensemble Forecasts

DINGCHEN HOU*

Center for Analysis and Prediction of Storms, Norman, Oklahoma

EUGENIA KALNAY

Department of Meteorology, University of Maryland at College Park, College Park, Maryland, and School of Meteorology, University of Oklahoma, Norman, Oklahoma

KELVIN K. DROEGEMEIER

Center for Analysis and Prediction of Storms, and School of Meteorology, University of Oklahoma, Norman, Oklahoma

(Manuscript received 6 June 1999, in final form 1 June 2000)

ABSTRACT

During May 1998, the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma coordinated a multi-institution numerical forecast project known as the Storm and Mesoscale Ensemble Experiment (SAMEX). SAMEX involved, for the first time, the real-time operation of four different ensembles of mesoscale models over the same region of the United States. The main purpose of this paper is the evaluation of the ensemble forecasts, performed at a relatively coarse resolution of 30 km. An additional SAMEX goal not discussed here is to compare the value of the ensemble forecasts against single forecasts made over smaller subregions of the Great Plains at both intermediate (10 km) and high (3 km) resolution.

The SAMEX '98 ensembles consisted of a single 36-h control forecast from the ARPS (at CAPS), the Penn State–NCAR fifth-generation Mesoscale Model (at NSSL), and the Eta Model and Regional Spectral Model (at NCEP), all with horizontal resolutions of approximately 30 km, and perturbed runs, resulting in a grand ensemble of 25 members. The forecasts of geopotential heights, temperatures, and moisture were verified against the Eta operational analyses, rather than observations. Unlike global ensembles, which tend to be useful in the medium range, the mesoscale SAMEX ensembles provided useful information in the short range. A major result is that the performance of the ensemble of *multiple* forecast systems is much better than that of each individual ensemble system, probably because it represents more realistically the current uncertainties in both models and initial conditions. A similar advantage from the use of multimodel, multianalysis systems has been observed with global ensembles. The SAMEX results also show that perturbations to model physics parameterizations, as well as the use of consistent perturbations in the boundary conditions, are important for regional ensemble forecasting. Efforts are now under way to compare the ensemble forecasts against those made using higher spatial resolution, and follow-on SAMEX experiments are anticipated in other geographical areas and weather regimes. Although the main results of this paper appear to be very robust, they were based on a small number of cases, and similar experiments carried out during other periods will help to test their significance.

1. Introduction

Since Leith (1974) showed that a number of numerical forecasts created from slightly different initial conditions can, if appropriately averaged, yield improved skill relative to a single control forecast, ensemble forecasting has grown into a major area of research. Global ensemble forecasting is now a cornerstone of several major operational prediction centers in the world [National Centers for Environmental Prediction (NCEP), the U.S. Navy, European Centre for Medium-Range Weather Forecasts (ECMWF), Japan Meteorological Agency (JMA), and U.K. Met. Office (UKMO)]. The ensemble approach is a computationally feasible method for estimating the probability density function of the atmospheric state as it evolves in time via the prediction of selected individual states, each physically plausible and distinct, and whose initial condition dispersion ideally should be representative of the initial analysis er-

* Current affiliation: School of Computational Sciences, George Mason University, Fairfax, Virginia.

Corresponding author address: Dr. Eugenia Kalnay, Department of Meteorology, University of Maryland at College Park, College Park, MD 20742.
E-mail: ekalnay@atmos.umd.edu

rors. It thus provides a quantitative basis for probabilistic forecasting. The averaging process serves as a nonlinear filter to selectively smooth more the unpredictable components of the flow, leaving behind those features for which the ensemble forecasts tend to agree (Toth and Kalnay 1997).

Leith (1974), Hollingsworth (1980), Mullen and Baumhefner (1989), and others have proposed using Monte Carlo methods for generating the initial conditions of the ensemble members. Recently, considerable attention has been given to methods for generating perturbations dynamically constrained by the ''flow of the day,'' including breeding (e.g., Toth and Kalnay 1993, 1997) and singular vectors (e.g., Buizza and Palmer 1995; Buizza 1997; see also Hamill et al. 2000). Such methods aim to introduce perturbations within the subspace of the growing errors. Breeding results in perturbations related to the leading Lyapunov vectors, which span the attractor of the dynamical system, and the leading singular vectors are the fastest growing perturbations given a choice of a norm (Ahlquist 2000, manuscript submitted to *J. Atmos. Sci.*). The control initial conditions represent the ''best'' estimate of the state of the atmosphere, and the added perturbations should be chosen so that they are representative of the expected analysis errors.

Following Houtekamer et al. (1996), Hamill et al. (2000) recently performed simulation experiments indicating that the most realistic set of ensemble initial conditions can be obtained from an ensemble of data assimilation cycles, where the observations are perturbed with random errors. In the perturbed observations (PO) method the ensembles contain dynamically constrained errors (contributed by the first guess) that project onto the subspace of leading Lyapunov vectors, and random errors (contributed by the random errors added to the observations). The PO results are therefore similar (but somewhat better) than breeding, and the perturbation growth is much smaller than that of total-energy-based singular vectors. More recently, perturbations to the models that account for uncertainties due to the use of imperfect models have also been introduced in ensemble forecasting by varying the model parameterizations of subgrid-scale physical processes (e.g., Stensrud et al. 1998; Houtekamer and Mitchell 1998; Andersson et al. 1998). Perturbations in the subgrid-scale ''physics'' may be more important for mesoscale short-range ensemble forecasting than in global models, since the higher resolution may allow a faster response from growing modes driven by convective instability.

Experiments with global models have shown that ensemble averaging applied to ''control'' forecasts from different centers, each of which is started from their own best initial condition (analysis), yields results significantly better than even the best of the individual forecasts (e.g., Kalnay and Ham 1989; Wobus and Kalnay 1995; Krishnamurti et al. 1999; Evans et al. 2000). For example, Kalnay and Ham (1989) pointed out that in the Northern Hemisphere extratropics, after only 12 h, ''consensus'' forecasts based on the average of multimodel–multianalysis systems (from ECMWF, UKMO, JMA, and NCEP) had a higher anomaly correlation than the best individual forecast (ECWMF) started from the same initial time. Similarly, M. Fritsch (1998, personal communication) found that the consensus of statistical forecasts based on several NCEP models outperformed the results from the best model. Mylne et al. (1999) and Evans et al. (2000) also show that ensembles including perturbations from both ECMWF and the UKMO, using both analyses and models from the two centers, outperform each individual system. Krishnamurti et al. (1999) showed that multisystems including a regression to correct systematic errors, which they denote ''superensembles,'' result in substantial skill improvements, including the forecast of hurricanes.

Although the multimodel, multianalysis approach does not fit squarely within the classic ensemble framework of perturbations to initial conditions, it is reasonable to assume that control forecasts produced by competing different state-of-the art data assimilation systems would represent fairly well the range of uncertainties present *both* in the *initial conditions* and in the *models.* Therefore it is perhaps not surprising that, when averaged, the ensemble of control forecasts yields greater skill, through a nonlinear filtering process, than even the best individual forecast.

Ensemble forecasting applied to the large-scale atmosphere became operational for the global forecasting systems at NCEP and ECMWF in December 1992 (Toth and Kalnay 1993; Molteni and Palmer 1993). The global ensemble forecasts are now widely used by forecasters to assess the reliability of the day-to-day forecasts (e.g., Toth et al. 1997). More recently, ensemble strategies have been used in limited-area models (e.g., at 60–80-km resolution). The potential advantages of this approach, denoted short-range ensemble forecasting (SREF), were first discussed during a workshop held at NCEP in 1994 (Brooks et al. 1995), where it was concluded that SREF should be especially useful for precipitation forecasting. As a result of the workshop, experimental ensemble forecasting systems were developed using the Eta Model and the Regional Spectral Model (RSM) at NCEP (Du et al. 1997; Hamill and Colucci 1997, 1998; Tracton et al. 1998; Du and Tracton 1999) and the Pennsylvania State University–National Center for Atmospheric Research (PSU–NCAR) fifth-generation Mesoscale Model (MM5) at the National Severe Storm Laboratory (NSSL) (Stensrud et al. 2000).

As numerical forecast systems and observational platforms (e.g., the Weather Surveillance Radar-1988 Doppler) continue to focus on smaller scales of the atmosphere, and as our understanding of physical processes continues to improve, greater emphasis will to be placed on the prediction of intense local weather using nonhydrostatic models at resolutions of 1 to 10 km. Indeed, the Weather Research and Forecasting model, being de-

veloped as a dual-purpose research and operational system by the national community (e.g., Dudhia et al. 1998), is targeted specifically at such resolutions. At this point in time, however, the specific data requirements, analysis and assimilation strategies, and spatial resolution and physics parameterizations needed to accurately predict the initiation, evolution, and decay of intense mesobeta and mesogamma weather systems are not well established. A fundamental question that remains to be answered is the relative value of an ensemble of mesoscale resolution forecasts (20–30 km), similar to the national models resolution, compared to a much smaller number of forecasts run at considerably higher resolution (1–3 km), and similar computational cost. The higher-resolution models would be run at much shorter-duration and smaller domains not only because of higher computational costs but also due to the lack of evidence for individual storm or mesoscale convective complex predictability at longer ranges. The answer to this and related questions has far-reaching implications for the manner in which the United States invests in future scientific research and technology acquisition, and efforts must be directed toward providing an answer so as to maximize available resources (e.g., Toth et al. 2000).

During the past decade, a number of groups in the United States and abroad have begun to experiment with mesoscale forecast models that seek to resolve explicitly, using high spatial resolution grids and observational data, the most important processes associated with intense convective and winter precipitation systems. Operationally, the Eta Model (e.g., Black 1994; Mesinger 1996) and the Rapid Update Cycle (Benjamin et al. 1996) have recently been implemented at mesoscale resolution (about 30 km) at NCEP for short-range predictions over North America. Several state-of-the-art nonhydrostatic models appropriate for storm-scale prediction have been developed as well, and some are also used for regional forecasting. For example, the Advanced Research and Prediction System (ARPS), developed by the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma, has been run on a daily basis for over two years with an emphasis on assimilating Doppler radar, satellite, and commercial aircraft observations to predict warm and cold season storms at spatial resolutions down to 3 km (Xue et al. 1995; Droegemeier 1997b; Carpenter et al. 1997, 1998, 1999). The MM5 (Dudhia 1993; Grell et al. 1994) also has been used for routine regional short-range predictions (e.g., Stensrud et al. 2000). The NCEP RSM is closely related to the NCEP global model and has been used for short-range forecast and climate applications (Juang et al. 1997).

In an attempt to build upon these many modeling activities, several groups joined forces in a regional numerical weather prediction experiment during the convective season of the spring of 1998. The Storm and Mesoscale Ensemble Experiment of May 1998 (Droegemeier 1997a) involved a real-time comparison of 25 ensemble forecasts, run at approximately 30-km resolution using four different models, against a much smaller number of forecasts run at both intermediate (10 km) and high (1–3 km) resolution over subsets of the ensemble domain.

Coordinated by CAPS, SAMEX involved the NSSL, the Air Force Weather Agency (AFWA), NCAR, and NCEP. Additionally, a number of other groups participated in the real-time forecast evaluation process, including several National Weather Service forecast offices, the Storm Prediction Center, Tinker Air Force Base, and the Aviation Weather Center.

SAMEX was novel in several ways. First, it provided for a direct comparison of several techniques for generating mesoscale ensemble initial conditions (bred perturbations, scaled-lagged average forecasting, Monte Carlo), and multiple model physics options. Second, the ensembles from each forecast system were themselves combined to create a multimodel/multianalysis ''grand ensemble.'' Third, it provided a framework for quantifying the quality and computational expense of low-resolution probabilistic and higher-resolution deterministic forecasts and developing techniques for verification and comparison (a process still under way). Fourth, it was conducted as a multiinstitution effort that included NCEP and leveraged several ongoing activities in experimental real-time NWP. And finally, it exposed operational forecasters, in real time, to technologies that are scheduled to become operational within the next several years.

Despite its rapid organization and relatively short duration, SAMEX '98 resulted in an unprecedented dataset that provides the opportunity for a large number of model and ensemble intercomparisons. Plans now under way to complete higher-resolution runs for smaller areas embedded within the larger domain of SAMEX '98 should further enhance the utility of this dataset and allow an assessment of the relative value of high-resolution versus coarser ensemble forecasting in mesoscale modeling. Future coordinated SAMEX experiments in other regions and seasons are planned as part of the U.S. Weather Research Program, and should further contribute to our understanding of the characteristics of different models and their application.

The purpose of this paper is to present verifications of the ensemble forecasting systems utilized during SAMEX '98, and to explore the advantages and disadvantages of different perturbation methods as well as the use of different model and data assimilation systems. Operational Eta analyses were used for the verification of forecasts of geopotential heights, temperatures, wind, and dewpoint temperatures. Although in principle it is preferable to verify against observations, the software to do so is only now being planned for development at CAPS. The verification of precipitation forecasts, done against observations, has been performed in a separate study (Miller 2000) and is not included here. There is

TABLE 1. Summary of the SAMEX′98 mesoscale ensemble forecast system.

Name	Institution	Model	No. of members	Method of member generation
NCP1	NCEP	Eta	5	Bred perturbations from global system
NCP2	NCEP	RSM	5	Bred perturbations from global system
NSSL	NSSL	MM5	10	Perturbations in IC and change in physics
CAPS	CAPS	ARPS	5	Scaled lagged average forecasting
Full		Multimodel	25	Perturbations in IC, BC, and physics
Cntl		Multimodel	4	The four models' control runs

no intent to rank a given model's performance: all models used during SAMEX '98 are state of the art and demonstrated similar capability, each excelling in one or more measures. On the other hand, the identification of specific model problems should provide a basis for improvement.

The logistics of SAMEX '98 and the method by which the ensemble initial conditions were created are described in Section 2. An analysis of ensemble spread is presented in section 3, and in section 4 we verify the bias and standard deviation of the individual and ensemble mean forecasts. Rank histograms are presented in section 5, and section 6 includes extensive probabilistic verifications. A summary and discussion are given in section 7.

2. SAMEX '98 ensemble forecast system and dataset

As indicated in the introduction, SAMEX '98 was conducted during May 1998 over the continental United States and central-southern Great Plains. Performed without any additional funding, SAMEX '98 was a proof-of-concept effort to provide an initial quantitative assessment of the value of coarse- (30 km) resolution ensemble forecasts relative to a few intermediate- (10 km) and high- (1–3 km) resolution forecasts performed at shorter ranges and in smaller domains. It also exposed operational forecasters to mesoscale ensemble and explicit cloud-resolving numerical predictions.

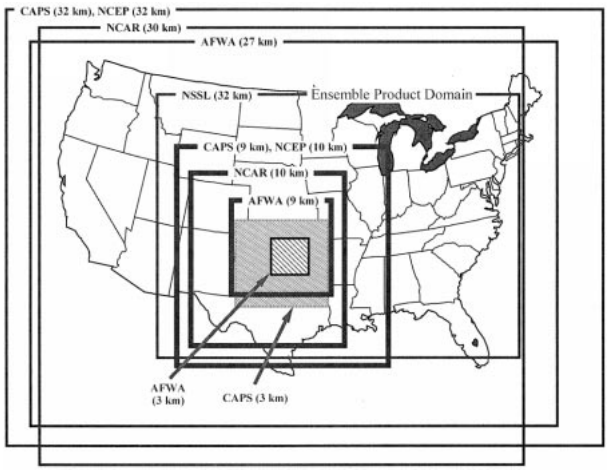

FIG. 1. The common ensemble domain used for forecast display and verification.

Three centers participated in this project by running four different numerical models: the CAPS ARPS, the NCEP Eta Model and RSM, and the MM5 used by NSSL. (Note that NCAR and the Air Force Weather Agency also participated by running the MM5 system, though not in an ensemble mode.) Each model was used to create a control run, along with a number of runs for which the initial conditions (and, in some cases, boundary conditions) were perturbed. The methods used to generate the perturbations in each system are given below and summarized in Table 1.

- NCP1: This five-member ensemble, created with the NCEP Eta Model, involved one control and four perturbed runs. The latter included initial condition (IC) and boundary condition (BC) perturbations generated by ''breeding'' fast growing modes (similar to Lyapunov vectors) in the global NCEP system (Toth and Kalnay 1993, 1997). Two different regional breeding perturbations were added (P1 and P2) and subtracted (N1 and N2) from the control (Tracton et al. 1998; Du and Tracton 1999). The boundary conditions were obtained from the NCEP global ensemble forecasting system.
- NCP2: This five-member ensemble, created with the NCEP RSM, involved one control and four perturbed runs. The same regionally bred IC and BC perturbations as in NCP1 were added and subtracted.
- NSSL: NSSL performed 10 forecasts with MM5 using both random perturbations in the IC (Mullen and Baumhefner 1989) and several combinations of changes in the model physical parameterizations [cumulus convection, moisture availability, and boundary layer, as described in Stensrud et al. (1998, 1999)]. The model was run on a North American domain at 96-km resolution, where the perturbations are introduced, and at 32 km in the domain of Fig. 1. The boundary conditions for the outer domain (from the Eta Model operational forecasts) were not perturbed and therefore were the same for all of the 10-member forecasts.
- CAPS: The perturbation members of the ARPS runs

TABLE 2. Dates in May 1998 for which the ensemble forecasts are available. The eight dates with complete ensembles are indicated in bold. The numbers on the top are the ensemble members and those in bold represent the control for each system (CT). For the CAPS system, +12, +24, −12, −24 are the SLAF ensembles based on 12- and 24-h perturbations added to or subtracted from the control. In the NCP1 and NCP2 systems, P1 and P2 are the first- and second-bred perturbations added to the control, and N1 and N2 the same perturbations subtracted from the control. In the NSSL system, P1–P9 represent different combinations of perturbed physical parameterization. See text for further discussion.

	CAPS					NCP1					NCP2					NSSL									
Day	**1** CT	2 +12	3 +24	4 −12	5 −24	**6** CT	7 P1	8 P2	9 N1	10 N2	**11** CT	12 P1	13 P2	14 N1	15 N2	**16** CT	17 P1	18 P2	19 P3	20 P4	21 P5	22 P6	23 P7	24 P8	25 P9
14	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	N	N	N
15	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
16	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	N	N	N	Y	Y	Y	Y	Y	N	N	N	N	N
17	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
18	Y	Y	N	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	N	N	N
19	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	N	N	N
20	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
21	Y	Y	N	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
22	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N	N	N
23	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
24	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
25	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
26	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
27	Y	Y	Y	Y	Y	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
28	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N
29	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
30	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
31	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

were generated by perturbing the IC and BC using the simple ''scaled lagged average forecasting'' (SLAF) method (Ebisuzaki and Kalnay 1991). Since SLAF was used for the first time in a mesoscale model, we discuss it here in more detail. SLAF is an extension of the lagged average forecasting (LAF) method (Hoffman and Kalnay 1983) in which the LAF initial perturbation, defined by the difference between a previous forecast and the verifying analysis, is scaled by its ''age'' (assuming approximately linear growth in time) and *added to* and *subtracted from* the initial analysis. Like LAF, breeding, and singular vectors, SLAF seeks to create dynamically growing perturbations. It includes consistent perturbed boundary conditions (constructed in the same fashion) and only requires the transfer of the control operational forecasts, a major advantage for centers not collocated with NCEP. By construction, the perturbations include ''errors of the day,'' presumably related to Lyapunov vectors. The use of positive and negative perturbations (which is done with breeding and singular vectors, but not with LAF) increases the chances of encompassing the truth, since in principle, positive and negative errors have an equal chance of occurrence. The Eta analysis was used as initial conditions for the control forecast, and the perturbations were generated from the 12- and 24-h-old Eta forecasts [the latter divided by two to account for the ''older'' (larger) initial errors]. These perturbations were added to and subtracted from the control, and the boundary conditions were obtained in a consistent way (scaling the older perturbations) from the Eta forecasts, resulting in one control and four perturbed forecasts.

- Full (or grand) ensemble: Included are all 25 forecasts (5 NCP1, 5 NCP2, 10 NSSL, and 5 CAPS) generated with different models and perturbation methods.
- Cntl ensemble: For some comparisons we also present the ensemble composed of the four control runs of the four forecast systems.

Note that at NCEP the Eta and RSM ensembles already comprise the multisystem NCEP regional ensemble (Tracton et al. 1998; Du and Tracton 1999). In this paper, however, we did not evaluate them together.

All forecasts in the ensembles were started at 0000 UTC and integrated for 36 h. The output from the forecasts, including the initial conditions, were interpolated to a common grid with horizontal spacing of 30 km for the purposes of display and verification. The common ensemble product domain is shown in Fig. 1 and includes the central and eastern parts of the United States. Numerous forecast products including mean and spread charts, ''spaghetti'' (single contour) diagrams, and conditional probability plots were available in real time on the World Wide Web. Also available were all of the individual forecasts composing the ensemble.

Despite the technical difficulties associated with an experiment of this magnitude, SAMEX '98 generated a major dataset for the period 14–31 May 1998, with only some members missing on particular days (Table 2). A complete ensemble dataset is available for 8 days, and for most of 18 days the majority of the forecasts were

completed. In all the comparisons that we made, we found that that *the average of the complete eight cases is very similar to the averages obtained with 18 days, so that the eight complete cases were used for the comparisons presented here.* As the verifying analysis we used the initial conditions from the Eta operational model interpolated to the SAMEX '98 common domain. This did not favor the NCEP forecasts, since NCP1 and NCP2 were initialized from the NCEP global system. The time evolution of ensemble spread is shown at 3-h intervals, while for other verification parameters only the 12-, 24-, and 36-h forecasts are presented.

3. Ensemble spread and its time evolution

In an ideal forecast ensemble, the verification should appear as a plausible member of the ensemble (Toth and Kalnay 1993). In order to satisfy this condition, a desirable feature of an ensemble system is that the perturbations introduced either in the IC or during the course of the forecast (as in the case of changes in the model physics) should grow at a rate comparable to the observed growth of forecast errors. It is easy to show that the average squared distance between two ensemble members is approximately twice the average squared distance between an ensemble member and the ensemble mean. Therefore, a convenient measure of the amplitude of the perturbations is the standard deviation or spread (SP) of the ensemble forecast members about the ensemble mean, defined as the rms difference from the ensemble mean of all of the ensemble members, averaged over the entire common domain; that is,

$$\mathrm{SP}(f) = \sqrt{\overline{\frac{1}{N}\sum_{n=1}^{N}(f_{i,j}^{n} - \tilde{f}_{i,j})^2}^{\,i,j}}\,, \qquad (1)$$

where f is a model-predicted variable, with subscripts i and j denoting the grid indices and the superscript n the ensemble member index, and N is the number of members in the ensemble. The tilde represents the ensemble mean and the overbar is the domain average:

$$\tilde{f}_{i,j} = \frac{1}{N}\sum_{n=1}^{N} f_{i,j}^{n}, \quad \text{and} \qquad (2)$$

$$\overline{f_{i,j}}^{\,i,j} = \frac{1}{I \cdot J}\sum_{i=1}^{I}\sum_{j=1}^{J} f_{i,j}. \qquad (3)$$

We found (not shown) that, for all forecast quantities, the spread results corresponding to the eight complete cases are extremely similar to those obtained for the entire 18-day dataset (Table 2). In light of this fact, and because some verification measures require complete datasets, only the results for the complete eight-case average will be discussed in the remainder of the paper. Because of the similarities, though, the results may be considered representative of the 18-day experiment.

Figure 2 shows the averaged spread for the geopotential heights at 250 (HGT250), 500 (HGT500), and 850 hPa (HGT850), as well as mean sea level pressure (MSLP), for the eight complete cases. These ''pressure'' or ''height'' variables show similar features, of which the most distinctive is a large peak in the full system at 3 h, which does not appear in the temperature, thickness, or moisture variables (cf. Fig. 3). This suggests that the large spread at 3 h is associated with an early

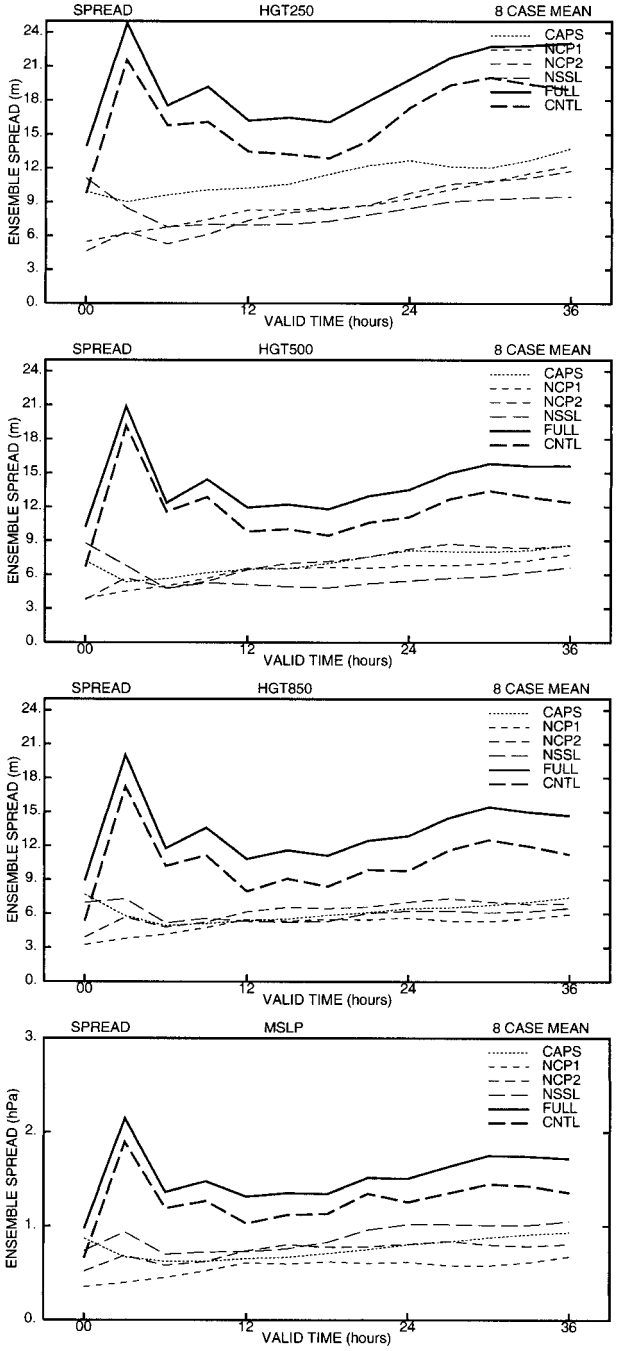

FIG. 2. Time series of the ensemble spread for MSLP, 850- (HGT850), 500- (HGT500), and 250-hPa heights (HGT250), averaged over the eight cases with complete datasets.

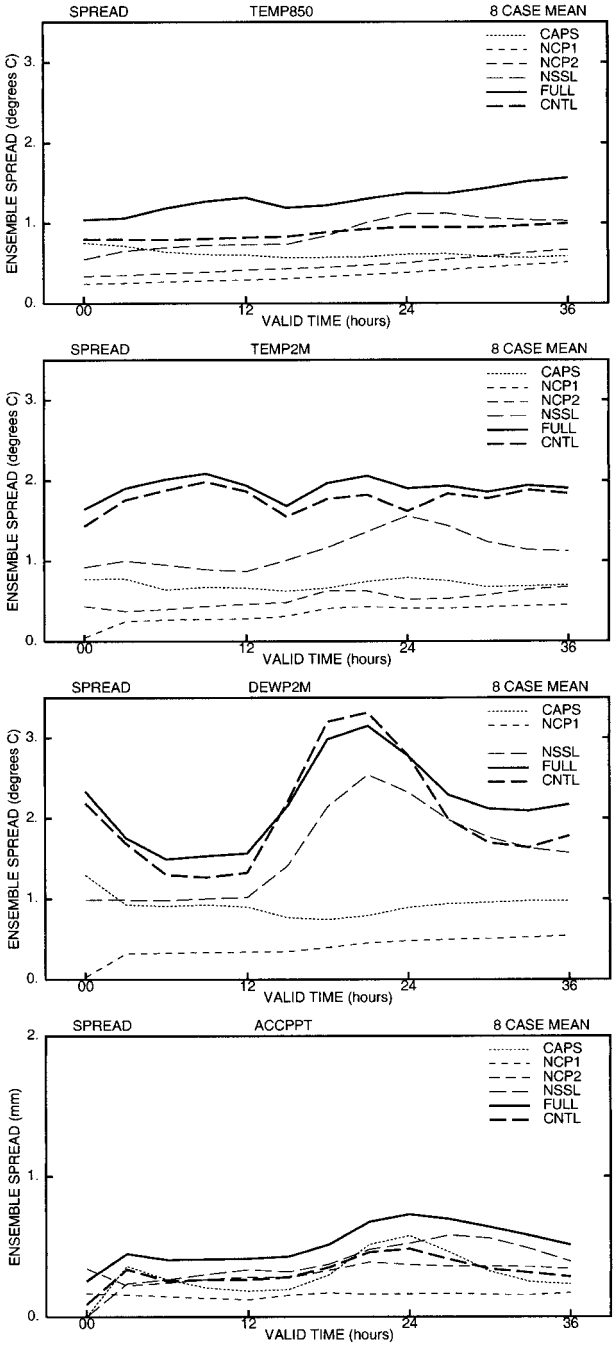

FIG. 3. Same as in Fig. 2 but for temperature at 850 hPa (TEMP850), the 2-m surface air and dewpoint temperature (TEMP2M and DEWP2M), and 3-h accumulated precipitation (ACCPPT). As with other products, the accumulated 3-h precipitation for each model was interpolated to the common grid. Dewpoint data for NCP2 were not available.

imbalance affecting the MSLP and, through hydrostatic balance, the geopotential heights at upper levels. This imbalance must be generating fast external inertia–gravity waves, since most of the geostrophic adjustment takes place within 6 h and the spread subsides. Since the imbalance does not appear in individual ensembles spread, it must be related to an initial imbalance *shared by all the members of at least one ensemble system.* In order to determine the system originating the imbalance, we computed the spread of three of the four systems, and found that the spread computed *without* the NSSL forecasts (not shown) does not have a peak at 3 h, indicating that the initial imbalance is associated with the NSSL ensemble.

Figure 2 also shows that the ensemble spread of the mass variables grows at different rates at different levels. NSSL and CAPS have some initial decay, suggesting that the initial perturbations are not as well balanced as in the NCP1 and NCP2 systems. At lower levels (MSLP and HGT850) NSSL has the largest spread and some growth (defined as absolute, not relative growth, i.e., the slope of the spread), but at the upper levels (HGT500 and HGT250), after the first 6 h, NSSL shows less growth than the other models. NCP1, NCP2, and CAPS have similar slopes in the spread at the upper levels, indicating dynamical growth of the initial perturbations. The lower slope of the NSSL ensemble at the upper levels may be due to the fact that the boundary conditions in the outer domain are the same for all of its members and because of the use of initial random perturbations. The lower levels may be more influenced by the spread in the precipitation due to the perturbations in the physics.

The most notable characteristic in Fig. 2, however, is that the full ensemble spread is much larger than those of the individual ensembles throughout the integration. Furthermore, it tends to show a larger slope (absolute growth) after the initial oscillations during the first 12 h. This suggests that the use of different analyses and different models leads to an ensemble with larger spread growth than attained by individual systems, which may help to make the ensemble more effective. A quantitative comparison of the forecast spread with the forecast error (presented in the next section, cf. Fig. 6) supports this conjecture. The spread of the ensemble of four control runs (cntl) is a close second, also supporting the conjecture.

The spread of several variables associated with temperature and moisture, that is, temperatures at 850 hPa (TMP850), 2-m temperature (TEMP2M), dewpoint (DEWP2M), and 3-h accumulated precipitation (ACCPPT), is shown in Fig. 3. The spread of these variables evolves quite differently from the height fields spread of Fig. 2. The NSSL ensemble, which includes different combinations of physical parameterizations for the boundary layer and for the cumulus convection, has much larger spread than the other models. In the surface temperature and dewpoint spread (but not in the precipitation), the NSSL ensemble has a very strong diurnal signal, with the maximum spread in the dewpoint at 2100 UTC (about 1500 local time), and in the temper-

ature at 2400 UTC (about 1800 local time). The CAPS ensemble, on the other hand, has a large initial spread, but no further growth except in precipitation, where it also has a large diurnal cycle. The ARPS model (CAPS) used a Kuo parameterization in this experiment, which when triggered changes only minimally the soundings at low levels.

As with most other variables, the spread of the 3-h ACCPPT for the full ensemble is significantly greater than for the individual ensembles. The Eta ensemble (NCP1) precipitation spread is surprisingly small and devoid of growth. This is an undesirable characteristic for this ensemble, suggesting that the Eta precipitation parameterization may be less sensitive to dynamical perturbations than the other models. For the thermal variables, the spread of the full ensemble is also much larger than those of individual ensembles (Fig. 3) and shows a steady increase with time for 850-hPa temperature and 500–1000-hPa thickness (not shown), but not in the 2-m temperature.

We can define the growth of the spread either in relative terms (percentage growth) or in absolute terms (slope of the spread). The results in this section indicate that the full ensemble has overall a larger absolute growth, whereas, for many of the variables, the two NCEP ensembles have a larger relative growth. It will be shown in the next section that the full system is the only ensemble for which the spread is similar in magnitude to the ensemble mean error, a desirable feature for an ensemble. As a result, the full system may be superior to the individual ensembles in that it provides overall a better chance to encompass the truth. It is important to recall that the models' domain-averaged bias was not included in the calculation of the spread. The larger spread and growth of the multisystem is due to both the use of different models and of different initial conditions, which may reflect better the current uncertainties in initial conditions and model formulation than a ''synthetic'' ensemble created using a single system, no matter how carefully it is designed. This efficiency of a multimodel ensemble, already noted for global models (e.g., Kalnay and Ham 1989), provides a benchmark that improvements in the perturbation generation schemes for single models may be able to achieve in the future.

4. Ensemble forecast errors

In this section we perform forecast verifications and investigate the improvement in the forecast due to the use of ensemble average compared with individual members, and the impact of using full ensembles compared with individual ensembles. Because our verification involves 2D atmospheric fields, we start with a widely used measure of horizontal average error, namely, mean-square error (mse). If f is a single forecast variable and v the corresponding verifying analysis, mse is defined (Wilks 1995) as

$$\text{mse} = \overline{(f_{i,j} - v_{i,j})^2}^{ij}, \tag{4}$$

where the overbar indicates, as before, domain average. The mean-squared error can be separated into the square of the domain bias and the error variance var [square of the standard deviation of the error (sde)]. The variance, in turn, can be separated into systematic and random components, following Takacs (1985) and Murphy (1988):

$$\text{mse} = \text{bias}^2 + \text{sde}^2 = \text{mnbias}^2 + \text{sdbias}^2 + \text{disp}^2, \tag{5}$$

where

$$\text{mnbias} = \overline{f_{i,j}}^{ij} - \overline{v_{i,j}}^{ij} = \frac{1}{I \cdot J} \sum_{i=1}^{I} \sum_{j=1}^{J} (f_{i,j} - v_{i,j}),$$

$$\text{sde} = \sqrt{\frac{1}{I \cdot J} \sum_{i=1}^{I} \sum_{j=1}^{J} (f'_{i,j} - v'_{i,j})^2},$$

$$f(f, v) = \frac{\sqrt{\dfrac{1}{I \cdot J} \displaystyle\sum_{i=1}^{I} \sum_{j=1}^{J} f'_{ij} v'_{ij}}}{\text{sd}(f)\text{sd}(v)},$$

$$\text{sdbias} = [\text{sd}(f) - \text{sd}(v)], \quad \text{and}$$

$$\text{disp}^2 = 2[1 - r(f, v)]\text{sd}(f)\text{sd}(v). \tag{6}$$

Here, $f'_{i,j} = f_{i,j} - \overline{f_{i,j}}^{ij}$, $v'_{i,j} = v_{i,j} - \overline{v_{i,j}}^{ij}$, sd is the standard deviation about the domain average of the forecast or the verification, and $r(f, v)$ is the pattern correlation (corr) between the forecast and verification fields. The bias of the mean (mnbias) and the bias of the standard deviation (sdbias) in Eq. (6) measure the systematic components of the domain-averaged forecast error and, for Gaussian variables, can be corrected a posteriori by subtracting the bias and by inflating or deflating the standard deviation of the forecast. The last term in (5) is the dispersion error (disp), proportional to 1-corr. Because it cannot be ''calibrated out'' it can be considered the most important measure of the forecast skill. As pointed out by Takacs (1985) the dispersion errors are due to phase errors, rather than amplitude errors.

As indicated before, the Eta Model initial conditions (0-h forecast) were used as verifying analysis. The mnbias, sdbias, disp, and corr were calculated for each case and averaged for the eight complete cases. As with the spread, these results were representative of the average obtained when using all 18 almost-complete cases. The time series of the mnbias, sdbias, disp, and corr of the individual ensemble averages as well as the full and cntl ensemble averages for HGT500 and TMP850 are shown in Fig. 4 and Fig. 5, respectively. The plot also includes the 0-h forecast, that is, the initial conditions, although at that time they are not representative of forecast errors but of the choice of initial conditions and verification. The CAPS and NSSL have much smaller

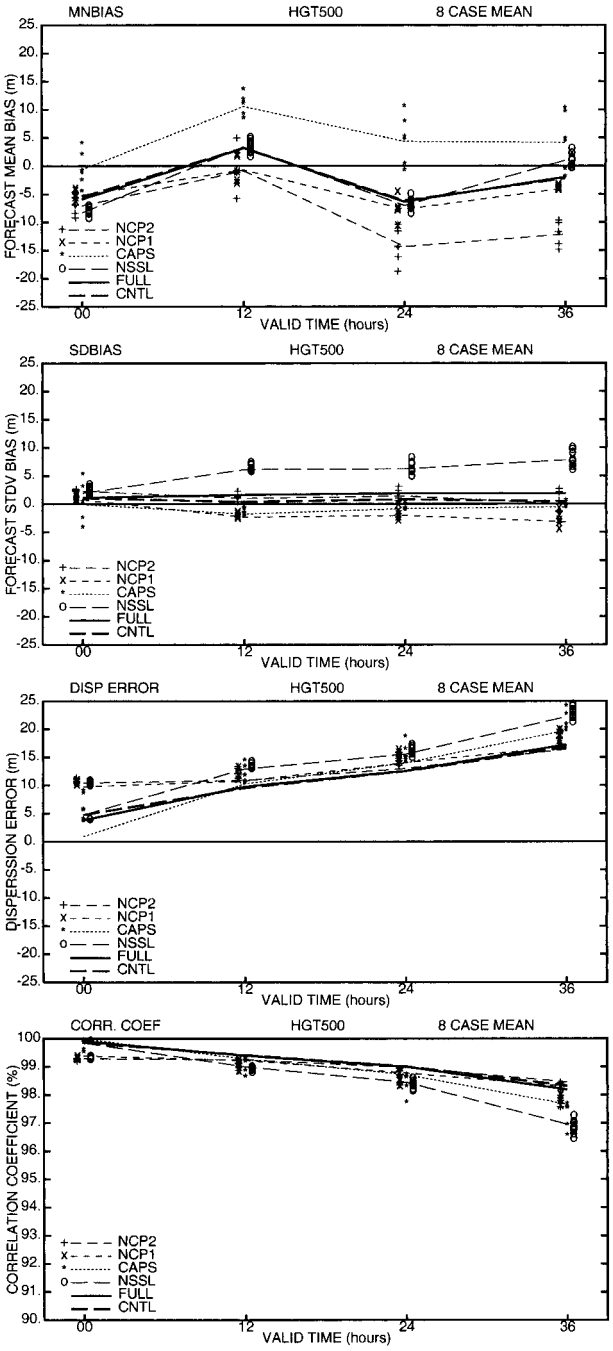

FIG. 4. Time series of average error statistics for 500 hPa for individual ensembles and the full and cntl ensembles: (a) Mean bias error, mnbias; (b) bias of the standard deviation, sdbias; (c) dispersion (disp) error; and (d) forecast–analysis correlations. The average score of the individual ensemble members are also indicated with different symbols.

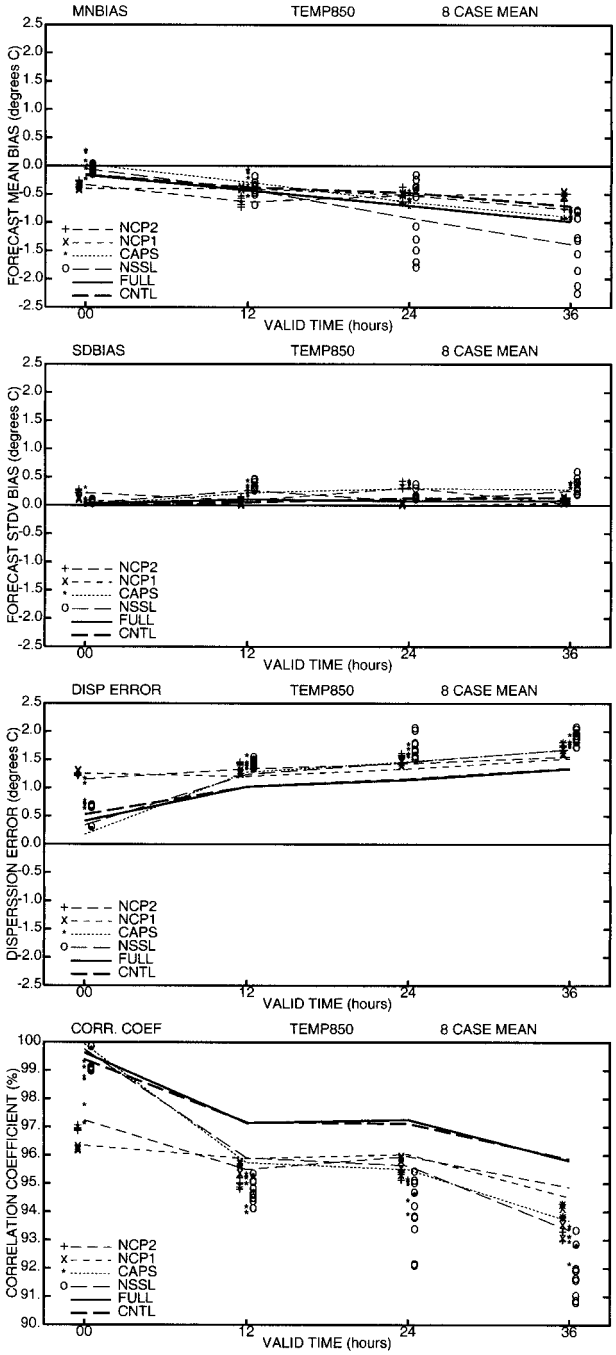

FIG. 5. Same as in Fig. 4 except for TMP850.

initial SDE compared with the NCEP ensembles. This is because the former systems used the operational Eta Model output as the basis of initial conditions, the same as the verifying analysis, whereas the control NCEP forecasts were obtained from the global system. The control CAPS system has larger initial apparent ''errors'' than its corresponding ensemble (see Fig. 7) because (unlike the other CAPS ensemble members) it underwent an additional regional analysis.

Clear differences between the mass variables and the thermal variables are noticed in the mnbias. Figure 4a shows that MNBIAS of HGT500 has a strong diurnal

cycle in all the ensembles. Because the minima are at 0 and 24 h (about 1800 local time), and maxima at 12 and 36 h (about 6000 local time), respectively, the diurnal cycle indicates not enough elevation of the 500-hPa surface during the daytime. In contrast, the mnbias in TEMP850 (Fig. 5a) displayed a clear tendency of cooling with time. The two NCEP ensemble averages show less of a tendency to cool in the mnbias (0.5 K or less up to 36 h). The CAPS ensemble has a stronger cooling tendency (about 0.8 K over 36 h) and NSSL cools the most (about 1.4 K in 36 h). The cntl ensemble shows some cancellation of errors in the bias, being smaller than all the individual bias except for NCP1.

The sdbias in Figs. 4b and 5b is a relatively small error term compared to the overall mnbias or to the dispersion component of the standard deviation of the error. It indicates that all the models are fairly good in representing the atmospheric variability during this period, although the NSSL slightly overestimates the variability in height, and NCP1 tends to underestimate it.

As indicated above, the most useful measure of the actual forecast skill is the dispersion component of the error (Figs. 4c and 5c). First we note that among the individual systems, both NCEP models have smaller standard deviation of height errors, and the Eta Model has the best temperature errors, compared to the non-operational CAPS and NSSL forecasts. The eight-case average for each of the individual ensemble members is also indicated in Figs. 4 and 5 with symbols. It can be observed that (with the exception of the heights in the NSSL ensemble), the dispersion error for each ensemble average is lower than the individual members dispersion error (including the control forecast). This is an important result indicating that nonlinear ensemble filtering of errors occurs very early in these short-range ensemble forecasts (see also discussion about Fig. 7 later in this section).

The most remarkable result in Figs. 4c and 5c, however, is that the full ensemble average has disp errors substantially smaller than all of the individual ensemble averages, and the cntl ensemble with only four members is a very close second. This is true for HGT500 (except for the 36-h forecast for which the NCEP forecasts are similar to the full) and for all other variables (not shown). It is especially clear with the temperature forecast errors at 850 hPa. Figure 5c shows that the disp error of the full ensemble is less than the NCP1 (Eta Model), the smallest among the individual ensembles, by about 0.3°C. The time correlation plots (Figs. 4d and 5d) also show that the multiple-model ensembles are far superior to any individual ensemble, having an advantage of 12–24 h in forecast skill using this measure, at least for the TEMP850. The cntl ensemble does also quite well, at a level essentially comparable to the full ensemble. Note that the differences in the apparent initial error in Figs. 4 and 5 are due to the choice of verification analysis (the Eta analysis) but this does not seem to affect the verifications at later times, when the errors are larger than the uncertainty in the analysis.

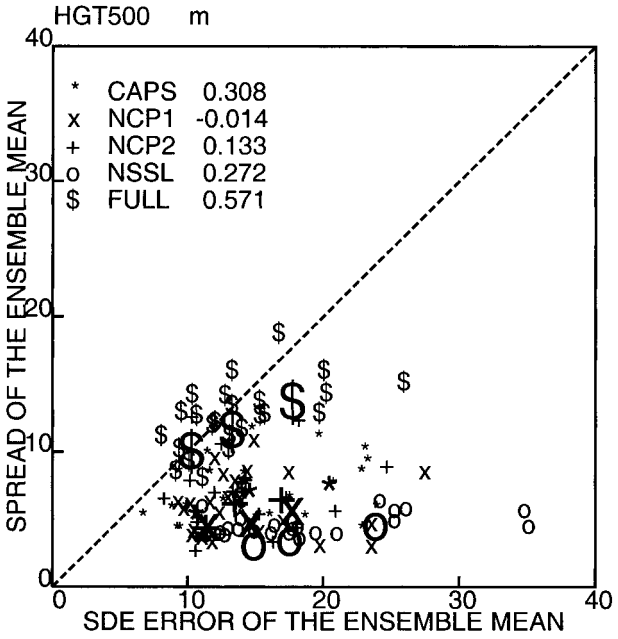

FIG. 6. Scatter diagram of the ensemble spread (sp), and the error standard deviation of the ensemble average (emsd) for the eight complete forecasts at 12-, 24-, and 36-h forecasts. The average for each forecast lead is indicated with larger symbols. The correlation between sp and emsd for each ensemble is shown in the inset.

We now compare the ensemble forecast error with the forecast spread of the different ensemble systems as discussed in the previous section. In a perfect ensemble the verifying field is an indistinguishable member of the ensemble. For a perfect ensemble we can assume that the pattern correlation between any two members of the ensemble, or between one member of the ensemble and the verification, is approximately the same. We can also assume that the variance of the field is the same for both ensemble members and the verification:

$$\overline{f_{i,j}^n f_{i,j}^m}^{i,j} = Vr \qquad \overline{f_{i,j}^n v_{i,j}}^{i,j} = Vr$$
$$\overline{f_{i,j}^n f_{i,j}^n}^{i,j} = \overline{v_{i,j} v_{i,j}}^{i,j} = V. \tag{7}$$

Under these assumptions the square of the spread of the ensemble members, and the ensemble mean error variance for a perfect ensemble, should be given by

$$\mathrm{sp}^2(f) = \frac{1}{N}\sum_{n=1}^{N} \overline{(f_{i,j}^n - \tilde{f}_{i,j})^2}^{i,j} = \frac{N-1}{N} V(1-r) \quad \text{and}$$

$$\mathrm{emsd}^2 = \overline{(\tilde{f}_{i,j} - v_{i,j})^2}^{i,j} = \frac{N+1}{N} V(1-r). \tag{8}$$

Therefore, *for a perfect ensemble, the spread of the ensemble should be similar to the forecast error of the ensemble mean*:

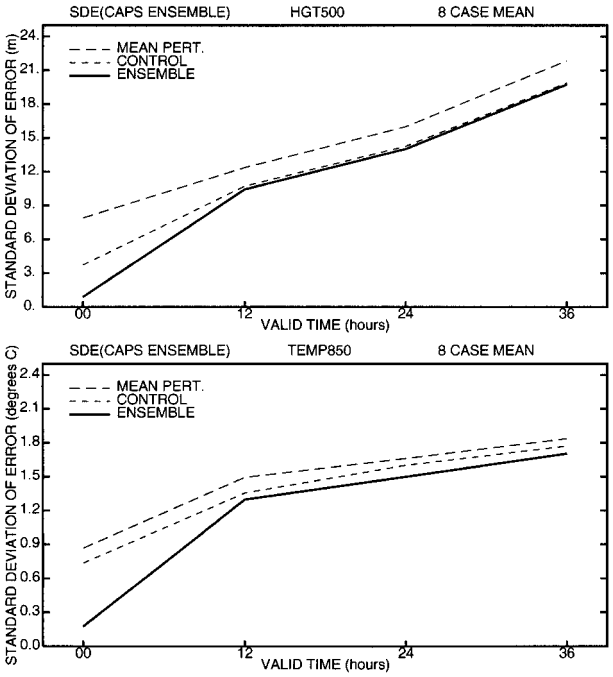

FIG. 7. Time evolution of sd of errors (sde) of (top) HGT500 and (bottom) TMP850, for CAPS ensemble, averaged over the eight cases of complete datasets.

$$\frac{\text{sp}}{\text{emsd}} \approx \sqrt{\frac{N-1}{N+1}}, \tag{9}$$

where N is the number of ensemble members.

Comparing Fig. 2c with Fig. 4c, it can be seen that the full ensemble spread at 500 hPa after the first 12 h, with the transient errors, is similar in magnitude to the dispersion errors that dominate the forecast errors. The spread of the individual ensembles, on the other hand, is much smaller than their corresponding errors. The same is true comparing the 850-hPa temperature spread and dispersion error (Figs. 3a and 5c). This relationship is shown very clearly in Fig. 6, which compares the spread and the standard deviation of the errors for the individual and full ensemble forecasts of HGT500, with the averages of the eight forecasts for 12, 24, and 36 h indicated with large symbols. The individual ensembles have spreads that are much smaller than the standard deviation of the error (even accounting for their smaller number of members). As noted in the previous section, the individual ensemble systems grow slower than the error in an absolute sense. The full ensemble lies closer to the ideal diagonal that would be expected for a perfect ensemble from (9). Moreover, the correlation between the spread sp and emsd shown in the inset in Fig. 6 shows again a very clear advantage for the full ensemble.

We now focus on the error characteristics of individual ensemble members, the control forecast, and the ensemble mean, for the CAPS system. The ensemble average errors for the five-member CAPS ensemble are smaller than the control error, despite the fact that the control forecast has smaller errors than the four CAPS individual perturbed forecasts. Figure 7 shows the sde errors associated with the control run, the error averaged over the four perturbation members, and the error of the ensemble average. Since the control is the best estimate of the initial conditions, it is not surprising that the four perturbed runs have individually larger rms errors than the control run. Nevertheless, the ensemble average including these perturbations provides a better forecast than the control. This is especially clear with the temperature (Fig. 7, lower panel), indicating that even at this short range, there is a beneficial effect from nonlinear ensemble filtering of the errors. We also note again that in the CAPS ensemble, the control was subjected to a mesoscale analysis using the ARPS Data Assimilation System (Brewster 1996), which increases its initial difference with the operational Eta initial conditions, used for verification.

We also looked at histograms of the number of forecasts for which an individual ensemble member is the best, defined as the forecast that has the smallest bias or sde (not shown). The most interesting result was that the control runs had the smallest sde; nevertheless, they were *not* the most frequent best ensemble member. This may be because in breeding and SLAF there is a relationship between perturbed forecasts over succeeding days. Therefore, if for example the second negative perturbation yields the best forecast on a given day, it is likely that the same will remain true for several successive days (Z. Toth and E. Kalnay 1992, personal communication). Obviously this would not hold true for a series of forecasts much longer than the ones we have available in SAMEX (Z. Toth 2000, personal communication).

5. Rank histograms

Rank histograms, also known as Talagrand diagrams, provide a necessary but not sufficient test for evaluating whether the forecast and verification are sampled from the same probability distribution (O. Talagrand 1996, personal communication; Anderson 1996; Zhu et al. 1996; Hamill and Colucci 1997), and therefore it is a useful measure of the realism of an ensemble (but not of its skill; see next section). The rank histograms are generated by ordering at each grid point the forecast values from each of the ensemble members from smallest to largest. For our full ensemble, with 25 members, this creates 26 intervals or bins. The value of the verifying analysis at this point then falls into one of the 26 categories or bins. If the analysis is less than the smallest value of the member forecasts, it falls in category 1, and if greater than the largest of the forecast values it falls into category 26. The same is done for all the grid points and all of the eight complete cases,

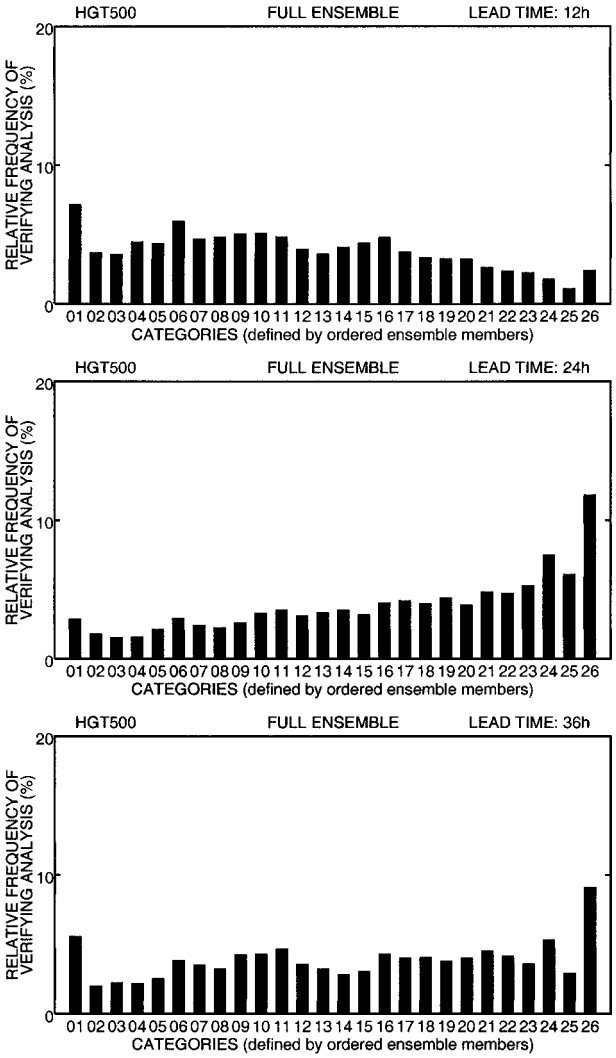

FIG. 8. Rank histogram or Talagrand diagram showing the relative frequencies at which the verifying analysis falls in each of the 26 categories, defined by the 25 ordered ensemble members at each grid points, for the full ensemble HGT500 forecast: (a) 12-, (b) 24-, and (c) 36-h forecasts.

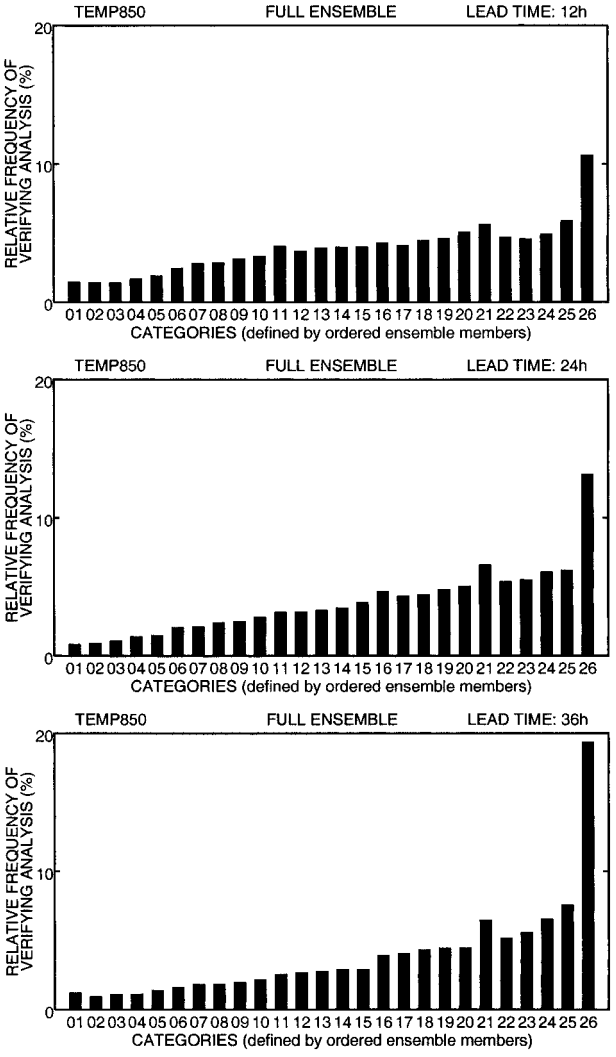

FIG. 9. Same as in Fig 7 except for TEMP850.

and the average frequencies at which the verifying analysis falls into each of the 26 categories are determined.

In an ideal ensemble all the perturbations in initial conditions, boundary conditions, and changes in model physics should represent equally likely scenarios, and if the verifying analysis is a plausible ensemble member, the rank histogram should be flat. A biased ensemble is indicated by a sloping rank histogram; insufficient spread among the ensemble members is indicated by a U-shaped histogram; an inverted U shape indicates excessive spread.

Figures 8 and 9 show the rank histograms for HGT500 and TMP850, respectively, that is, the histograms of the frequencies as a function of the category index. For the 500-hPa heights, the distribution is indeed quite flat, although the two extreme categories are somewhat higher than their adjacent categories. The 36-h forecast is particularly good in this respect, even though the accuracy of the forecast decreases with time. There is a shift of frequencies of the verifying analysis from the lower categories to higher categories between 12 and 24 h, consistent with the positive mnbias of the full ensemble average at 12 h, and the negative mnbias at 24 h, as shown in Fig. 4.

The rank histogram of TEMP850 shows a systematic bias toward the higher categories, increasing with forecast lead time. This could be expected from the systematic cold bias of the temperature forecasts from the full ensemble shown in Fig. 5. If the mnbias had been corrected (Hamill and Colucci 1997), this plot would have been also fairly flat (not shown).

The rank histograms of HGT500 and TEMP850 from individual ensemble systems (not shown) are much less

flat than the full ensemble distributions, and they show larger proportions of the verifications falling outside the ensemble (in the extreme categories). The sum of the frequencies of the two extreme categories represents the relative frequency of cases in which the ensemble failed to encompass the truth and is therefore referred to as missing rate. However, with different numbers of members in each ensemble, the missing rate cannot be directly compared because an ensemble with a larger number of members would have a better chance to encompass the truth. Following Zhu et al. (1996) the missing rate is adjusted for ensemble size by subtracting the expected missing rate if the verification is evenly distributed among all of the categories. The expected missing rate for a perfect ensemble of 25 members, is 2/(25 + 1) or about 7.7%. For the individual ensembles, this correction is 33% (NCP1, NCP2, CAPS) or 18% (NSSL). The cntl ensemble has only four members, so its adjustment is 40%. The ''adjusted missing rates'' for all six ensembles are shown in Fig. 10. It is clear that for all of the variables and all forecast leads, the full ensemble has an adjusted missing rate smaller than any of the individual ensembles and much smaller than their average. In fact, the adjusted missing rate is less than 10% in most cases. This indicates that the members of the full ensemble are more representative of all the possible scenarios, compared with the individual ensembles. The adjusted missing rate associated with the cntl ensemble is even lower, and mostly negative for the sea level pressure. Negative adjusted missing rates indicate that the members of the ensemble have an excessive spread. Except for this, the cntl ensemble has very good scores for adjusted missing rates, suggesting that the improvement of the forecast by using the full ensemble average is due, to a large extent, to the use of multiple models. Further improvements can be expected with better methods of generation of perturbation members. Comparing the individual ensembles, it is apparent that the two operational models do fairly well in the MSLP and 500-hPa heights, but as we saw before, they do not have enough spread in the temperature. This is also true for the CAPS ensemble and is also suggested by the fact that their temperature spread is considerably smaller than the dispersion error. The NSSL system, the only one with perturbations in the physics, does the best for the temperatures, but the worst for 500 hPa heights. The CAPS system does the worst in MSLP.

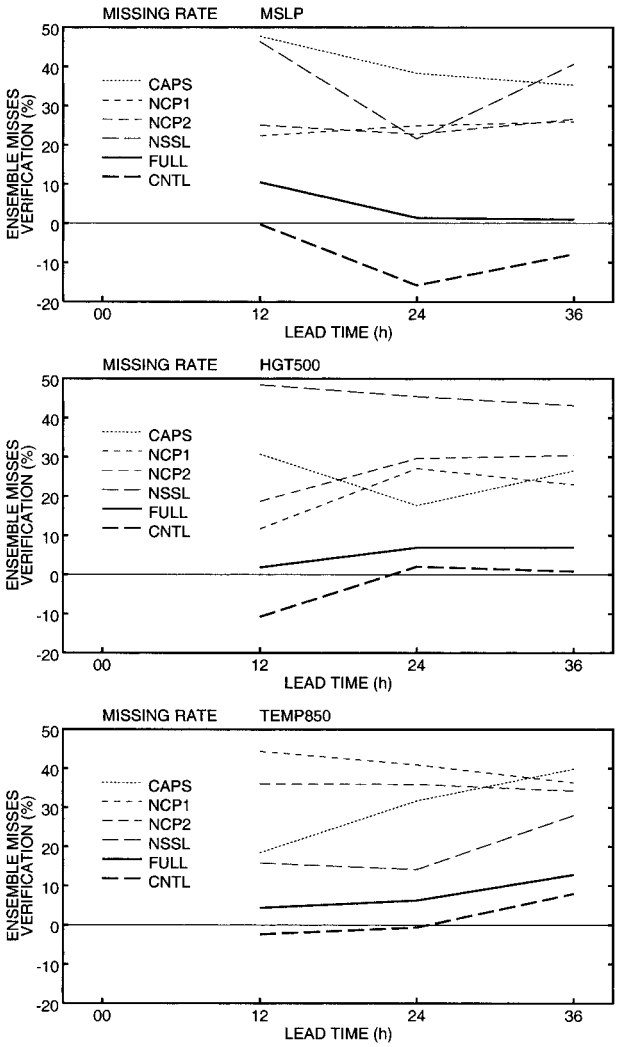

FIG. 10. Frequencies at which the ensemble does not encompass the verifying analysis, adjusted according to the size of the ensembles (see the text for detail).

6. Probability verification measures

One of the most important applications of the ensemble forecasts is their use for generation of probabilistic forecasts. The rank histograms (Talagrand diagrams) of the previous section show the extent to which the ensembles encompass the truth in the probability distributions, as well as indicating their bias. However, they cannot measure the usefulness of the forecast: *an ensemble created from a random climatological distribution of states of the atmosphere would result in a perfectly flat Talagrand diagram for a sufficient number of forecasts, but would have no useful information beyond that contained in climatology.* In this section, we evaluate the performance of a simple probabilistic forecast of HGT500 and TEMP850 based on the full ensemble. For brevity the individual ensembles and cntl ensemble are not included in this evaluation.

The probabilistic forecasts were created by determining the percentage of the ensemble members that fall into any of 10 climatologically equally likely bins. These bins were determined by Zhu et al. (1996) using the climatological database from the NCEP–NCAR reanalysis (Kalnay et al. 1996), and were kindly made available to us for this study. The original climatological bins were computed on the 2.5° lat × 2.5° long global grid of the reanalysis, and a distance-weighed interpo-

TABLE 3. Comparison of the Brier score and its components [reliability (rel.) and resolution (res.)] calculated from the eight complete cases. The third component of the Brier score (uncertainty, not shown) is equal to 0.090 and is the same for all forecasts.

		HGT500			TEMP850		
		BS	Rel.	Res.	BS.	Rel.	Res.
CAPS	t = 12 h	0.061	0.009	−0.038	0.059	0.005	−0.036
	t = 24 h	0.053	0.005	−0.042	0.065	0.009	−0.034
	t = 36 h	0.075	0.011	−0.026	0.094	0.018	−0.014
NCP1	t = 12 h	0.046	0.003	−0.047	0.078	0.015	−0.027
	t = 24 h	0.063	0.007	−0.035	0.068	0.010	−0.033
	t = 36 h	0.068	0.008	−0.030	0.086	0.015	−0.019
NCP2	t = 12 h	0.053	0.004	−0.042	0.076	0.012	−0.026
	t = 24 h	0.071	0.009	−0.028	0.071	0.011	−0.030
	t = 36 h	0.080	0.011	−0.021	0.089	0.015	−0.016
NSSL	t = 12 h	0.071	0.014	−0.033	0.059	0.004	−0.035
	t = 24 h	0.065	0.011	−0.035	0.051	0.002	−0.041
	t = 36 h	0.074	0.012	−0.028	0.076	0.005	−0.019
Cntl	t = 12 h	0.043	0.001	−0.048	0.055	0.003	−0.038
	t = 24 h	0.049	0.002	−0.043	0.051	0.003	−0.042
	t = 36 h	0.058	0.003	−0.035	0.073	0.006	−0.023
Full	t = 12 h	0.045	0.001	−0.045	0.052	0.001	−0.039
	t = 24 h	0.048	0.000	−0.042	0.048	0.001	−0.042
	t = 36 h	0.056	0.001	−0.035	0.069	0.001	−0.024

lation scheme was used to generate the same bins for the SAMEX '98 grid.

A forecast probability y_i for each of the 10 bins is given by the relative number of forecast members falling within that bin, and since there are 25 independent forecasts it can only take one of the following 26 values: 0/25, 1/25, . . . , 24/25, 25/25, or 0%, 4%, . . . , 96%, 100%. On the other hand, the observed probability of the event, o, can only be $o_1 = 1$ (when the observation does falls within the bin) or $o_2 = 0$ otherwise. Similar to the mean-square error in Eq. (4), the Brier score (BS) is essentially the mean-squared error of the probability forecasts:

$$\mathrm{BS} = \frac{1}{n}\sum_{k=1}^{n}(y_k - o_k)^2, \tag{10}$$

where the summation takes place over all n forecast–observation pairs available for all grid points and all eight cases. Murphy (1973) showed that the Brier score can be separated into three components:

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{26} n_i(y_i - \overline{o}_i)^2 - \frac{1}{n}\sum_{i=1}^{26} n_i(y_i - \overline{o})^2 + \overline{o}(1 - \overline{o}), \tag{11}$$

where n_i is the number of times a forecast y_i is used in the collection of forecast–verification pairs, $\overline{o} = 0.1$ is the observed average frequency of the verification falling into each bin, and $\overline{o}_i = p(o_1|y_i) = (1/n_i)\,\Sigma_{k\in n_i} o_k$ is the conditional frequency of the verification falling into a bin when a forecast y_i has been issued.

The three terms on the right-hand side of (11) are known as reliability, resolution, and uncertainty, respectively. The uncertainty $\overline{o}(1 - \overline{o}) = 0.09$ is independent of the forecast method. Because the reliability can be increased by calibration (e.g., Zhu et al. 1996; Toth et al. 1998), the resolution is the most important term. Table 3 shows the reliability, resolution, and total Brier score for each of the ensemble systems for the eight complete days. It shows that the full and the cntl ensembles had very good reliability for both the heights and the temperatures. Since the full ensemble has 25 members, and the cntl ensemble only 4 members, this suggests that the cntl ensemble itself would be considerably better than 4 randomly chosen subensemble members of the full ensemble (Du et al. 2000). NSSL had relatively poor reliability for the heights, but it was the best individual system for the temperatures. In general, the individual systems had comparable resolution, which decayed with forecast length. CAPS was the best individual system for the heights at 24 h.

The terms in (9) can also be interpreted geometrically with a reliability diagram (Wilks 1995) in which the observed frequency (subsample relative frequency) $\overline{o}_i$ is plotted as a function of y_i, the forecast probability. Reliability diagrams for HGT500 and TEMP850 are shown in Figs. 11a and 11b. The insert in each diagram shows the size of the corresponding subsample (histogram), n_i, as a function of y_i. Because the n_i values are calculated from all the 10 bins and the eight complete cases, the histogram shows that most of the forecast probabilities are zero or close to zero. However, there is also a noticeable maximum in the histogram for forecast probability 1, indicating that the forecasts show significant resolution. Since the reliability in (11) is the sum of the squared difference between y_i and $\overline{o}_i$, the diagonal line in Fig. 11 represents perfect reliability. For both variables and all lead times, the curves showing $\overline{o}_i$ as a function of y_i are quite close to the diagonal, indicating that the forecasts are quite reliable. The TEMP850 24-h forecasts, show overforecasting of medium probability and underforecasting of higher prob-

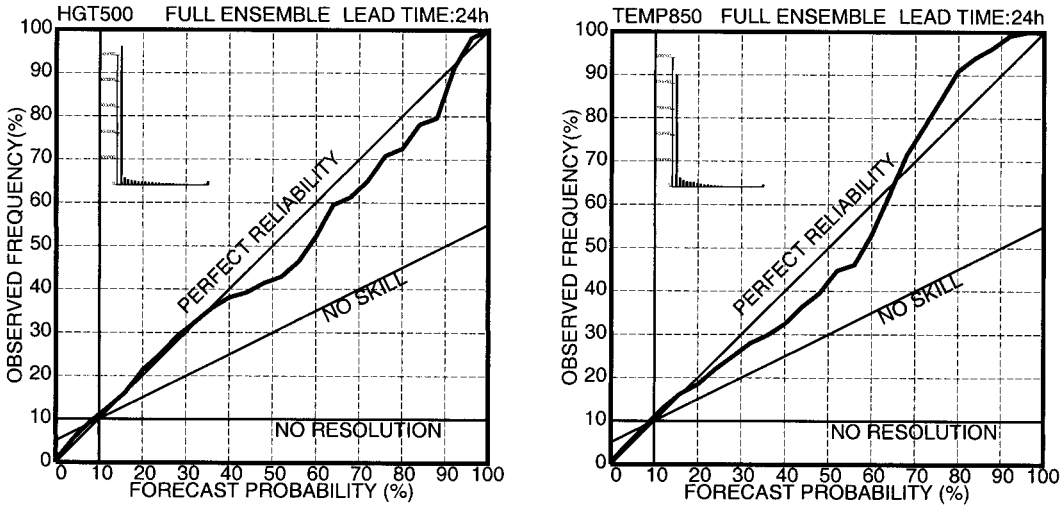

FIG. 11. The 24-h forecast reliability–attributes diagrams for (a) HGT500 and (b) T850.

ability (and similarly for the 36-h forecast, not shown). However, most of the forecasts are very good for low probability (30% or less) and highest probability (92% or over). Considering the fact that the results have not been calibrated, and that the sample is concentrated in these two ranges, the full ensemble is very successful. Note also that we have plotted the ''no skill'' line for which the resolution is equal to the reliability (Wilks 1995, p. 265). Since all the points in Fig. 11 are above the no skill line, all the subsamples of forecasts contribute positively to the overall skill.

The relative operating characteristic (ROC) is another verification measure based on signal detection theory. The ROC diagram shows the false alarm rate (F) as the abscissa and the hit rate (H) as the ordinate. ROC is especially useful in ensemble verification because it offers another way of comparing the performance of the control forecasts with that of the ensemble. For each of the four control forecasts, we check whether a forecast falls into a bin, and a hit rate and a false alarm rate are determined. For the ensemble forecast, the forecast probability of each bin is determined as before and a critical probability y^* can be used to interpret the probabilistic forecast to ''occur'' (the probability is greater than y^*) or ''not occur'' (the probability is less than y^*). Clearly, a different hit rate and false alarm rate can be found for each value of y^*. To plot the ensemble ROC, 25 values of y^* were used, ranging from 4% to 100% with even intervals of 4%.[1]

In the classic definition used by Swets (1973) and Stanski et al. (1989), the hit rate is defined as the probability that an event was forecasted given that it was

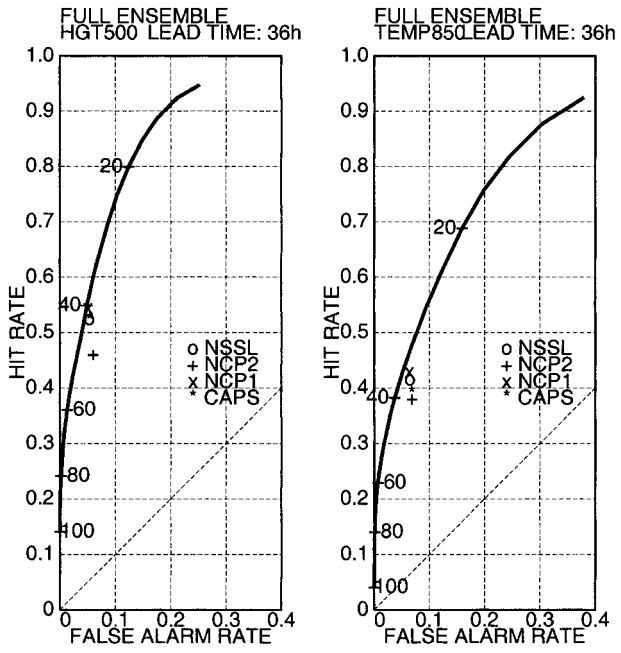

FIG. 12. Relative operating characteristic plots for the control forecasts (symbols) and the full Ensemble forecast (curves) of (a) HGT500 and (b) TEMP850. ROC using Swets' (1973) definition. The thin diagonal line is the critical line indicating positive skill with respect to chance under the classical definition of Swets (1973) and Stanski et al. (1989).

[1] Surprisingly, different definitions of hit rate and false alarm rate are used in the meteorological literature, so that the interpretation of the ROC diagram also varies with the definition. Mason (1982), Zhu et al. (1996), and Hamill et al. (2000) followed the classical definition used by Swets (1973) and Stanski et al. (1989). Wilks (1995) gives different definitions of hit rate and false alarm rate, which satisfies the principle of equivalence of events (a successful ''yes'' forecast is given as much value as a successful ''no'' forecast). The difference between these two sets of definitions and their interpretation are discussed in Hou et al. (2001, manuscript submitted to *Wea. Forecasting*), who demonstrate the following. 1) Under the principle of equivalence of events, the above-mentioned critical line $H = F$ should be replaced by the line $H = 9F - 4$. 2) Using $H = 0.5$ as the critical line, the ROC following Wilks' definition is equivalent to the Swets ROC in (1).

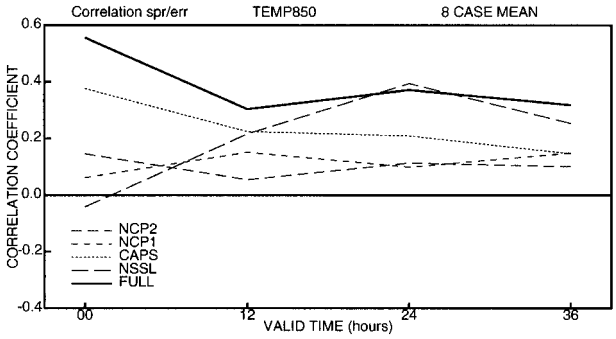

FIG. 13. Pattern correlation between ensemble spread and forecast error averaged for all complete eight cases. See text for details.

observed [$p(f = \text{yes} | o = \text{yes})$], and the false alarm rate is defined as the probability that an event was forecast given that it did not happen [$p(f = \text{yes} | o = \text{no})$]. In this definition all forecast sets with some positive skill over chance will have ROCs in the upper-left triangle (the diagonal is referred to as the critical line hereafter) and a forecast is better if it corresponds to a point closer to the top-left corner (Mason 1982; Zhu et al. 1996; Hamill et al. 2000).

We present the ROC diagrams for 36-h forecasts in Fig. 12a for the 500-hPa heights and Fig. 12b for the 850-hPa temperatures. Each of the control forecasts is shown as a single point and the ensemble forecast is represented by a series of points associated with different y^*. It can be seen that as y^* increases, the false alarm rate and the hit rate both approach 0. As expected from previous results, the four control forecasts are close to each other, indicating that these forecasting systems have comparable performance in both ROC measures. However, the NCP1 is, by a small amount, the best (slightly closer to the upper-left corner), and NCEP2 is, also by a small amount, the worst in this measure.

The traditional ROC definition using $H = F$ as the critical line would indicate that the best forecast is the ensemble with $y^* \sim 20\%$ and that the ensemble is better than the controls if $y^* < 40\%$. This interpretation would seem to be biased against the ensemble forecast with higher y^*. This is because, in the classical ROC definition, false alarms of rare events are considered to be less important than hits.

Note that the height forecast shows a higher hit rate for all of the control forecasts and all y^* values of ensemble forecast, when compared with the temperature forecast. The difference is greater with low y^* values. This supports the conclusion that the ensemble forecast for height is more successful than that of temperature.

Finally, we present in Fig. 13 the pointwise correlation between spread and SDE (standard deviation of the error) as a function of the forecast length. This is computed by first creating for each grid point (not underground) a pair of values corresponding to the absolute error and to the ensemble spread. For each forecast length the spread and errors of all eight complete cases are lumped together, and the two sets of data are then correlated. This measure can be used to assess the ability of the ensemble to predict the forecast skill a priori (Kalnay and Dalcher 1987; Palmer and Tibaldi 1988), but it is not necessarily an accurate measure since a large spread indicates potential for large errors, but actual errors may be small. For the full ensemble, the correlation is higher, about 0.4. Although this explains only 16% of the variance, it compares well with other systems, which tend to reach a similar maximum value only after several days into the forecast (Barker 1991; Wobus and Kalnay 1995; Whitaker and Laughe 1998). It is very noticeable that among the subsystems the NSSL subsystem starts with very low correlation (consistent with the use of random initial perturbations) but reaches the highest correlation at 24 h (presumably due to the use of multiple physical parameterizations).[2]

7. Summary and discussion

We summarize here the main results of this paper, with the caveat that since the SAMEX '98 experiments took place within a period of less than a month, additional experiments for other periods, verifying against observations, should be carried out before our results can be considered definitive.

- Some problems with individual systems have been clearly identified, and could be corrected in the future. For example, the NSSL system had a strong early imbalance; the Eta ensemble at NCEP showed a disturbingly small spread in the precipitation; the CAPS system had little growth in the spread of the temperature and a strong bias probably associated with problems in the soil model initialization that have since been identified and corrected.
- The height spread for the NSSL ensemble was smaller than for other three systems, possibly due to the lack of perturbed BC in its outer domain. However, the presence of perturbations in the physics was clearly beneficial, resulting in a larger spread of temperature and moisture in the NSSL system.
- The full ensemble including multiple systems had by far the largest spread after correcting for the models bias. The ensemble of four cntl forecasts also had larger spread than any individual ensemble system. A perfect ensemble would have an average ensemble spread similar to the error of the ensemble mean. The spread of the full ensemble system was much closer to fulfilling this condition than the individual ensembles. Considering its smaller size, the multisystem cntl ensemble was very competitive with the full ensemble.

[2] The correlation of the spread and the errors could be spuriously augmented by a latitudinal dependence. The fact that the NSSL system, with realistic but random initial perturbations, has zero initial correlation between spread and errors indicates that this effect is small.

- The full ensemble resulted in the best average forecast. This is perhaps not surprising, because a forecast ensemble system should reflect our present uncertainty in the initial conditions and in the model deficiencies. A multiple-system apparently does this more realistically than any present method used to create a synthetic ensemble.
- The correlation between spread and error is about 0.4 for the full system, comparable to the maximum value other larger-scale systems attain much later in the forecast. NSSL starts with a very low correlation, presumably because of the use of random initial perturbations but reaches the highest correlation of any system after 24 h because of the use of multiple physical parameterizations.
- Although for each system the control is, on the average, the forecast with smallest errors, it was not the most frequent best forecast, presumably because in the breeding and SLAF methods there is memory between successive runs.
- Rank histograms show rather ''flat'' distributions for the full ensemble for heights. The full ensemble has the lowest number of misses (not encompassing the verification). The cntl is also good in this measure, although it tends to overestimate the spread. The rank histogram for temperature reflects the cooling tendency of all models.
- Probabilistic forecasts were generated by using 10 bins with equal climatological probability derived from the NCEP–NCAR reanalysis (Zhu et al. 1996) and they were evaluated using the Brier score. As with other measures, the full and cntl were the best systems overall.
- The hit rate versus false alarm (relative operating characteristic) allows a comparison between deterministic and probabilistic forecasts. It shows that all the control forecasts have comparable skill, with the Eta control having a slightly better ROC. However, the full ensemble beats the control forecasts if the threshold probability is chosen to be 40% or less.

In summary, the first major conclusion of this study is that an ensemble of multiple models, multiple analyses has a much better performance than individual system ensembles, probably because it represents most realistically the current uncertainties in both the models and in the initial conditions. This result supports previous similar results obtained for the global systems [e.g., Kalnay and Ham (1989) found that after 12 h the average of a small ensemble of four operational global models was a better forecast than the best global operational system]. The second major conclusion is that perturbations in the physics, and BC consistent with perturbations in IC, are both important for regional ensemble forecasting. Since our sample of forecasts covered less than a month, it is important that more SAMEX experiments over other regions and seasons be carried out in the future in order to confirm whether our results are robust.

Acknowledgments. We extend our deepest appreciation to all who participated in SAMEX '98. In particular, we acknowledge those who ran their models including Drs. David Stensrud (NSSL), Steve Tracton and Jun Du (NCEP), Randy Lefevre (AFWA), and Jimmy Dudhia (NCAR). The entire CAPS team was involved in SAMEX in one way or another, and we give a special thanks to Ming Xue, Fred Carr, Donghai Wang, Henry Neeman, Steve Weygandt, Vince Wong, Dan Weber, Richard Carpenter, Gene Bassett, Jian Zhang, Yvette Richardson, Jason Levit, and Yuhe Liu. In addition, we are very grateful to Y. Zhu (NCEP) for making available climatologically equally probable bins computed from the NCEP–NCAR reanalysis, and to Drs. Harold Brooks and David Stensrud for several helpful suggestions.

The CAPS and NCEP forecasts were made at the Pittsburgh Supercomputing Center (PSC), and exceptional assistance was provided by Ralph Roskies, Bruce Loftis, and others at the PSC. Drs. Bob Borchers and Richard Hirsh of the National Science Foundation worked tirelessly to ensure our access to PSC, and Jason Martin of the Oklahoma State Regents for Higher Education is gratefully acknowledged for providing high speed access to the Internet. This research was supported by the National Science Foundation under Grant ATM91-20009 to the Center for Analysis and Prediction of Storms at the University of Oklahoma.

Last, we are very grateful to three reviewers, Zoltan Toth, Tom Hamill, and an anonymous reviewer, whose very detailed and careful suggestions and corrections resulted in a substantially improved manuscript.

REFERENCES

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate,* **9,** 1518–1530.

Andersson, E., and Coauthors, 1998: The ECMWF implementation of three-dimensional variational assimilation (3D-Var). III: Experimental results. *Quart. J. Roy. Meteor. Soc.,* **124,** 1831–1860.

Barker, T. W., 1991: The relationship between spread and forecast error in extended-range forecasts. *J. Climate,* **4,** 733–742.

Benjamin, S. G., J. M. Brown, K. J. Brundage, D. Devenyi, B. Schwartz, T. G. Smirnova, T. L. Smith, and F.-J. Wang, 1996: The 40-km 40-level version of MAPS/RUC. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., 161–163.

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

Brewster, K., 1996: Application of a Bratseth analysis scheme including Doppler radar data. Preprints, *15th Conf. on Weather Analysis and Forecasting,* Norfolk, VA, Amer. Meteor. Soc., 92–95.

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.
——, and T. Palmer, 1995: The singular vector structure of the atmospheric general circulation. *J. Atmos. Sci.,* **52,** 1434–1456.
Carpenter, R. L., Jr., K. K. Droegemeier, G. M. Bassett, W. L. Qualley, and R. Strasser, 1997: Project Hub-CAPS: Storm-scale NWP for commercial aviation. Preprints, *Seventh Conf. on Aviation, Range, and Aerospace Meteorology,* Long Beach, CA, Amer. Meteor. Soc., 474–479.
——, and Coauthors, 1998: Storm-scale NWP for commercial aviation: Results from real-time operational tests in 1996–1997. Preprints, *12th Conf. on Numerical Weather Prediction,* Phoenix, AZ, Amer. Meteor. Soc., 213–216.
——, K. K. Droegemeier, G. M. Bassett, S. S. Weygandt, D. E. Jahn, S. Stevenson, W. L. Qualley, and R. Strasser, 1999: Storm-Scale numerical weather prediction for commercial and military aviation. Part 1: Results from operational tests in 1998. Preprints, *Eighth Conf. on Aviation, Range, and Aerospace Meteorology,* Dallas, TX, Amer. Meteor. Soc., 209–211.
Droegemeier, K. K., 1997a: Outline of plans for SAMEX '98: The 1998 Storm and Mesoscale Ensemble Experiment. Center for Analysis and Prediction of Storms Internal Rep., 4 pp. [Available from CAPS, 100 East Boyd Street, Norman, OK 73019.]
——, 1997b: The numerical prediction of thunderstorms: Challenges, potential benefits, and results from real-time operational tests. *WMO Bull.,* **46,** 324–336.
Du, J., and M. S. Tracton, 1999: Impact of lateral boundary conditions on regional-model ensemble prediction. *Research Activities in Atmospheric and Oceanic Modelling,* H. Ritchie, Ed., WMO/TD-No. 942, 6.7–6.8.
——, S. L. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.,* **125,** 2427–2459.
——, ——, and ——, 2000: Removal of distortion error from an ensemble forecast. *Mon. Wea. Rev.,* **128,** 3347–3351.
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.
——, J. Klemp, W. Skamarock, D. Dempsey, Z. Janjic, S. Benjamin, and J. Brown, 1998: A collaborative effort towards a future community mesoscale model (WRF). Preprints, *12th Conf. on Numerical Weather Prediction,* Phoenix, AZ, Amer. Meteor. Soc., 242–243.
Ebisuzaki, W., and E. Kalnay, 1991: Ensemble experiments with a new lagged average forecasting scheme. WMO Research Activities in Atmospheric and Oceanic Modeling Rep. 15, 308 pp. [Available from WMO, C.P. 2300, CH1211 Geneva, Switzerland.]
Evans, R. E., M. S. J. Harrison, and R. Graham, 2000: Joint medium-range ensembles from The Met. Office and ECMWF systems. *Mon. Wea. Rev.,* **128,** 3104–3127.
Grell, G., J. Dudhia, and D. R. Stauffer, 1994: A description of the fifth-generation Penn State/NCAR Mesoscale Model (MM5). NCAR/TN-398 + STR, 121 pp. [Available from MMM Division, NCAR, P.O. Box 3000, Boulder, CO 80307.]
Hamill, T. M., and S. J. Colucci, 1997: Verification of Eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.,* **125,** 1322–1327.
——, and ——, 1998: Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.,* **126,** 711–724.
——, C. Snyder, and R. E. Morss, 2000: A comparison of probabilistic forecasts from bred, singular-vector, and perturbed observation ensembles. *Mon. Wea. Rev.,* **128,** 1835–1851.
Hoffman, R. N., and E. Kalnay, 1983: Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus,* **35A,** 100–118.
Hollingsworth, A., 1980: An experiment in Monte Carlo forecasting. *Proc. Workshop on Stochastic-Dynamic Forecasting,* Reading, United Kingdom, ECMWF, 65–85.
Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.,* **126,** 796–811.
——, L. Lefaiver, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.,* **124,** 1225–1242.
Juang, H.-M. H., S.-Y. Hong, and M. Kanamitsu, 1997: The NCEP Regional Spectral Model: An update. *Bull. Amer. Meteor. Soc.,* **78,** 2125–2143.
Kalnay, E., and A. Dalcher, 1987: Forecasting forecast skill. *Mon. Wea. Rev.,* **115,** 349–356.
——, and M. Ham, 1989: Forecasting forecast skill in the Southern Hemisphere. *Extended Abstracts, Third Int. Conf. on Southern Hemisphere Meteorology and Oceanography,* Buenos Aires, Argentina, Amer. Meteor. Soc., 24–27.
——, and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.,* **77,** 437–471.
Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Willford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecast from multimodel superensemble. *Science,* **285,** 1548–1550;
Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.,* **102,** 409–418.
Mason, J., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.,* **30,** 291–303.
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; Corrigendum: **78,** 506.
Miller, M., 2000: Verification of precipitation and forecast usefulness for SAMEX mesoscale ensemble forecasts. M.S. thesis, School of Meteorology, University of Oklahoma, 95 pp. [Available from School of Meteorology, University of Oklahoma, 100 E. Boyd, Norman, OK 73019.]
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.
Mullen, S. L., and D. P. Baumhefner, 1989: The impact of initial condition uncertainty on numerical simulations of large-scale explosive cyclogenesis. *Mon. Wea. Rev.,* **117,** 2800–2821.
Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.,* **12,** 595–600.
——, 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.,* **116,** 2417–2424.
Mylne, K., R. T. Clark, and R. E. Evans, 1999: Quasi-operational multi-model multianalysis ensembles on medium-range timescales. Preprints, *13th Conf. on Numerical Weather Prediction,* Denver, CO, Amer. Meteor. Soc., 204–209.
Palmer, T. N., and S. Tibaldi, 1988: On the prediction of forecast skill. *Mon. Wea. Rev.,* **116,** 2453–2480.
Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. World Weather Watch Rep. 8, WMO TD/No.-358, 114 pp.
Stensrud, D. J., J.-W. Bao, and T. T. Warner, 1998: Ensemble forecasting of mesoscale convective systems. Preprints, *12th Conf. on Numerical Weather Prediction,* Phoenix, AZ, Amer. Meteor. Soc., 265–268.
——, ——, and ——, 2000: Using initial condition and model physics perturbations in short-range ensembles. *Mon. Wea. Rev.,* **128,** 2077–2107.
Swets, J. A., 1973: The relative operating characteristics in psychology. *Science,* **182,** 990–999.
Takacs, L., 1985: A two-step scheme for the advection equation with minimized dissipation and dispersion errors. *Mon. Wea. Rev.,* **113,** 1050–1065.
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: The breeding method. *Mon. Wea. Rev.,* **125,** 3297–3318
——, ——, S. Tracton, R. Wobus, and J. Irwin, 1997: A synoptic evaluation of the NCEP ensemble. *Wea. Forecasting,* **12,** 140–153.
——, Y. Zhu, T. Marchok, S. Tracton, and E. Kalnay, 1998: Verification of the NCEP global ensemble forecasts. Preprints, *12th*

Conf. on Numerical Weather Prediction, Phoenix, AZ, Amer. Meteor. Soc., 286–289.

——, ——, and R. Wobus, 2000: On the economic value of ensemble weather forecasts. Preprints, *15th Conf. on Probability and Statistics in the Atmospheric Sciences*, Asheville, NC, Amer. Meteor. Soc., 88–91.

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

Whitaker, J. S., and A. F. Loughe, 1998: The relationship between ensemble spread and ensemble mean skill. *Mon. Wea. Rev.*, **126,** 3292–3302.

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

Wobus, R., and E. Kalnay, 1995: Three years of operational prediction of forecast skill. *Mon. Wea. Rev.*, **123,** 2132–2148.

Xue, M., K. K. Droegemeier, V. Wong, A. Shapiro, and K. Brewster, 1995: ARPS version 4.0 user's guide. Center for Analysis and Prediction of Storms, 380 pp. [Available from Center for Analysis and Prediction of Storms, 100 East Boyd Street, Norman, OK 73019.]

Zhu, Y., G. Iyengar, Z. Toth, M. S. Tracton, and T. Marchok, 1996: Objective evaluation of the NCEP global ensemble forecasting system. Preprints, *15th Conf. on Weather Analysis and Forecasting*, Norfolk, VA, Amer. Meteor. Soc., J79–J82.