Initial Results of a Mesoscale Short-Range Ensemble Forecasting System over the Pacific Northwest

ERIC P. GRIMIT AND CLIFFORD F. MASS

Department of Atmospheric Sciences, University of Washington, Seattle, Washington

(Manuscript received 23 April 2001, in final form 25 September 2001)

ABSTRACT

Motivated by the promising results of global-scale ensemble forecasting, a number of groups have attempted mesoscale, short-range ensemble forecasting (SREF), focusing mainly over the eastern half of the United States. To evaluate the performance of mesoscale SREF over the Pacific Northwest and to test the value of using different initial analyses as a means of ensemble forecast generation, a five-member mesoscale SREF system was constructed in which the Pennsylvania State University–National Center for Atmospheric Research fifth-generation Mesoscale Model (MM5) was run with initializations and forecast boundary conditions from major operational centers. The ensemble system was evaluated over the Pacific Northwest from January to June 2000. The model verification presented in this study considers only near-surface weather variables, especially the observed 10-m wind direction. The ensemble mean forecast displays lower mean absolute wind direction errors than the component ensemble members when averaged over all cases. The frequency with which the ensemble mean forecast verifies best is no better than the frequency of any individual member forecast. The wind direction forecast errors for the 12-km ensemble mean forecasts are comparable to 4-km deterministic forecast errors. Ensemble mean forecasts are observed to retain much of the orographically forced mesoscale structure in the component forecasts while smoothing out phase differences for propagating features. The correlation between forecast spread and forecast error for wind direction is approximately 0.6 for most lead times. Spread–error correlations rise to roughly 0.8 when only cases with high or low spread are considered. Such high correlations suggest that the ensemble system possesses the ability to predict forecast skill for high- and low-spread cases. The tendency toward higher spread–error correlation when cases with medium spread are filtered out is also found for each component member of the ensemble.

1. Introduction

Short-range forecast accuracy has improved as increasing computational power has allowed the use of progressively finer resolution, more sophisticated model physics, and better data assimilation procedures. However, inherent limits in atmospheric predictability (Lorenz 1963, 1969) may restrict the value of further decreasing grid spacing in numerical weather prediction models. For example, real-time forecasts at the University of Washington over the past four years using the Pennsylvania State University–National Center for Atmospheric Research fifth-generation Mesoscale Model (PSU–NCAR MM5; Grell et al. 1994) suggest diminishing returns as grid spacing drops below 12 km, when evaluated using standard measures of forecast skill (Mass et al. 2002). Furthermore, numerical model forecasts can be very sensitive to slight changes in the larger-scale initial conditions (Brooks et al. 1992). Recognition of such predictability issues has led to increased interest in developing an alternative strategy for further improving numerical weather forecasts (Brooks and Doswell 1993), namely, ensemble forecasting.

Ensemble forecasting provides a practical way of addressing variability in the initial conditions (ICs), uncertainties in model physics, and the inherent uncertainty in atmospheric prediction. For example, using forecasts started from varying ICs, each being equally likely to match the actual atmospheric state, a collection of weather scenarios and their relative likelihood may be constructed. An approximation to the forecast probability density function (PDF) of model variables can be defined by the calibrated frequency distribution of the resulting ensemble forecasts. Thus, ensemble forecasting lends itself to prediction of forecast probability, an advantage over deterministic forecasting.

Ensemble forecasts can still provide additional value over deterministic forecasts even if the initial and forecast PDFs are poorly defined. On average, the ensemble mean forecast tends to be a better estimate of the verifying state than the individual forecasts that compose the ensemble (Thompson 1977; Toth and Kalnay 1993; Tracton and Kalnay 1993; Molteni et al. 1996; Hamill and Colucci 1997; Stensrud et al. 1999). This appears

Corresponding author address: E. P. Grimit, Department of Atmospheric Sciences, Box 351640, University of Washington, Seattle, WA 98195-1640.
E-mail: epgrimit@atmos.washington.edu

to be the case even when the forecast distribution is multimodal (Atger 1999). It is possible that information about forecast reliability may be derived from the spread of the ensemble forecasts (Buizza 1997; Whitaker and Loughe 1998), even though the spread may not match the variance of the true probability distribution.

The successful application of ensemble prediction systems (EPS) on the global scale at the National Centers for Environmental Prediction (NCEP; Toth and Kalnay 1993) and the European Centre for Medium-Range Weather Forecasts (ECMWF; Molteni et al. 1996) has motivated exploration of ensemble forecasting for shorter lead times on the mesoscale. Short-range ensemble forecasting (SREF) has been tested at NCEP with a combined Eta–Regional Spectral Model ensemble (Eta–RSM; Hamill and Colucci 1997, 1998; Stensrud et al. 1999; Du and Tracton 2001; Wandishin et al. 2001) and by a larger community during the Storm and Mesoscale Ensemble Experiment of 1998 (SAMEX; Droegemeier 1998; Hou et al. 2001). Preliminary results of these investigations have been mixed. The Eta–RSM ensemble run at 80-km horizontal grid spacing has been shown to match or even outperform the 29-km Meso Eta Model (Hamill and Colucci 1997; Stensrud et al. 1999; Wandishin et al. 2001). Ensemble mean forecasts in SAMEX outperformed their respective individual forecasts (Hou et al. 2001) over a limited number of synoptic situations.

Neither the Eta–RSM nor the SAMEX ensembles demonstrated an ability to skillfully predict forecast reliability. Correlations between the error of the ensemble mean and the spread of the component ensemble forecasts in both experiments were generally below 0.4 (Hamill and Colucci 1998; Hou et al. 2001), thus explaining less than 16% of the variance. However, even for a well-formulated ensemble the spread–error correlation is not required to be large, since the correlation depends primarily on the day-to-day variability of spread (Whitaker and Loughe 1998). Ensemble spread is likely to be a more useful predictor of skill when it is extreme (very large or very small) in comparison with its climatological mean value (Houtekamer 1993; Whitaker and Loughe 1998). The ability of short-term ensembles to predict forecast skill remains an open question.

Mesoscale ensemble experiments considering explosive cyclogenesis (Mullen and Baumhefner 1989, 1994; Du et al. 1997) and mesoscale convection (Stensrud and Fritsch 1994a,b; Stensrud et al. 1998, 2000) have focused primarily over the midwestern and eastern United States. Thus, many mesoscale ensemble studies have considered events in which deep convection plays an important role. In a convectively dominated situation, error growth due to model deficiencies may be comparable to error growth due to imperfect ICs, since simulated convection is very sensitive to its parameterization. Stensrud and Fritsch (1994a,b) suggest that varying model physics may be just as beneficial as varying the atmospheric ICs to create useful ensemble forecasts, especially for events with weak large-scale forcing.

The use of physics perturbations for the creation of short-term ensembles may be less valuable over orographic areas where mesoscale structures are determined predominantly by the interaction of synoptic-scale flow with resolved topography (Mass et al. 2002). Thus, mesoscale temporal and spatial variability is likely to be forced primarily by differences in the synoptic-scale flow. Over the Pacific Northwest, convection is typically weaker, shallower, and less frequent than over the Midwest. With a vast data-sparse region over the eastern Pacific, there is considerable uncertainty in the upstream initial conditions that can lead to large forecast errors over western North America (Silberberg and Bosart 1982; Grumm and Siebers 1989; Sanders 1992; Mullen 1994; Langland et al. 1999), especially when strong flow enters the United States from over the Pacific (Fritsch et al. 2000). Clearly, changing physics parameterizations, but using a basic state far from reality, is of minimal benefit for creating useful ensemble forecasts. Thus, over the Pacific Northwest it is desirable to sufficiently sample the initial state before moving on to address the concerns brought about by model deficiencies.

Selection of ICs for SREF has been based primarily on random perturbations (Mullen and Baumhefner 1989, 1994; Du et al. 1997; Stensrud et al. 2000; Hou et al. 2001) and breeding methods (see Toth and Kalnay 1993, 1997; Hamill and Colucci 1997, 1998; Stensrud et al. 1999; Du and Tracton 2001; Hou et al. 2001). Neither technique may be appropriate for SREF IC selection since it is not known how errors grow and organize on the mesoscale. Random perturbations are designed to be balanced, coherent structures whose size is scaled by the analysis uncertainty. The most rapidly growing modes during the analysis cycle, which are approximated by the breeding process, may not be optimal perturbations in many cases. Additionally, if the control IC is seriously in error, small perturbations about that analysis based on either random perturbations or the breeding method may not realistically capture the analysis uncertainty.

A combined multimodel multianalysis (MMMA) technique appears to be the most promising ensemble approach. Ensemble mean forecasts from SAMEX, which comprised forecasts from multiple modeling sources using both bred ICs and random ICs based on a number of basic states, displayed the lowest error scores (Hou et al. 2001). The Eta–RSM ensemble developed for the SREF pilot project follows the idea of the MMMA approach as well, and a 10-member version is now considered an operational part of NCEP's production suite (Du and Tracton 2001). The advantages of MMMA ensembles have also been demonstrated for the larger spatial and temporal scales (Evans et al. 2000; Ziehmann 2000; Richardson 2001). Although MMMA ensembles do not fit within the classic Monte Carlo

approach of generating perturbations to ICs, state-of-the-art analysis and forecast systems from various operational forecast centers represent the community's best attempts at simulating the atmosphere and may provide insight into the range of uncertainties present in both the ICs and the models. Multimodel approaches using a single analysis attempt to overcome modeling system deficiencies while using multiple analyses and a single modeling system may diagnose sensitivity to the ICs. The combination of the two approaches maximizes the benefits of each by compensating for deficiencies in the ICs and in the modeling systems.

The relative contributions of multimodel and multianalysis approaches to improving ensemble forecasts has not been quantified. Richardson (2001) found that up to 80% of the improvement over control forecasts produced using an MMMA ensemble was realized by just the multianalysis part of the ensemble. Thus, a majority of the benefits created by running an MMMA ensemble can be realized by using a single model and ICs from analyses produced by different centers. The improvement of multianalysis ensemble mean forecasts over control forecasts in a different study (Fritsch et al. 2000) is more modest because of the use of interrelated, in-house NCEP analyses rather than those collected from different operational forecast centers. Since we believe that the primary concern over the Pacific Northwest regards initial condition uncertainty, it seems logical to first construct an ensemble from multiple analyses before incorporating multiple modeling systems or varying model physics.

This paper presents an initial evaluation of a short-range ensemble forecasting system using the PSU–NCAR MM5 driven by five operational initializations over a limited-area domain encompassing the northwestern United States and the eastern Pacific Ocean. Using an outer domain with 36-km horizontal grid spacing and an inner nest of 12-km grid spacing over the states of Washington and Oregon, the ensemble forecast system has sufficient resolution to capture mesoscale weather phenomena forced by regional terrain features. The initializations and lateral boundary conditions for the MM5 ensemble are obtained from NCEP, Fleet Numerical Meteorology and Oceanography Center (FNMOC), and the Canadian Meteorological Centre (CMC). By drawing on analyses from different centers, the multianalysis approach to ensemble IC selection is evaluated. Consistency is maintained by keeping the domains and the model physics identical for all ensemble members, thus varying only the ICs and lateral boundary conditions (LBCs). This work attempts to answer several questions:

- Is a multianalysis ensemble approach using ICs and LBCs from different operational forecast systems viable?
- Does the ensemble mean possess greater skill than its component forecasts in terms of standard measures of forecast skill?

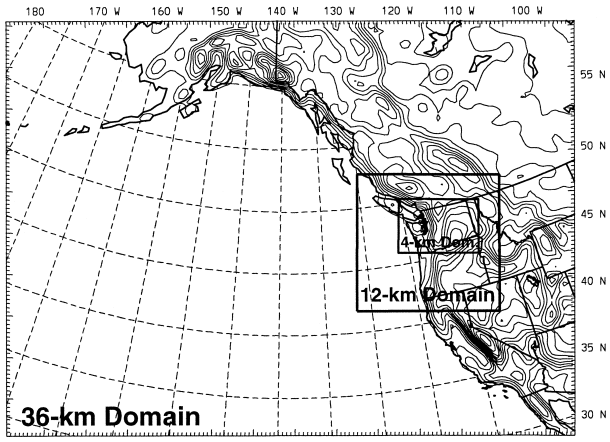

FIG. 1. Geographic coverage of the outer (36-km grid spacing) and inner (12-km grid spacing) domains for the MM5 short-range ensemble forecasts. The 4-km domain is included for reference to contemporaneously run, high-resolution, deterministic forecasts.

- How does the ensemble mean forecast skill compare with higher-resolution deterministic forecasts using the same model?
- Can the mesoscale ensemble predict forecast skill? Specifically, is there a significant correlation between ensemble spread and ensemble mean skill and/or the skill of the component forecasts?

Section 2 reviews the ensemble system in detail. A description of the verification tools and methodology used to answer the above questions is presented in section 3. Ensemble performance is evaluated in section 4, with conclusions in section 5. Ensemble results presented in this paper are compiled from real-time MM5 forecasts during the first phase of the project (January–June 2000). Further research into the forecast probability characteristics of this particular ensemble system will be presented in a future work using additional real-time forecasts from the cool season of 2000/01.

2. Description of the mesoscale ensemble system

a. Model configuration

To focus on the applicability of SREF to mesoscale prediction over the northwestern United States, the PSU–NCAR MM5 (version 2.12) numerical model is used. For the 0000 UTC cycle only, ensemble forecasts out to 48 h are run over an outer domain (137 × 101 grid points; 36-km horizontal grid spacing) that includes the west coast of the United States, western Canada, southern Alaska, and the eastern Pacific Ocean extending out to near the international date line. The one-way nested domain (100 × 103 grid points; 12-km horizontal grid spacing) includes the states of Washington and Oregon (Fig. 1).[1] The extent of the 36-km domain is

[1] The 36-km/12-km horizontal grid spacing configuration for the MM5 over the northwestern United States and eastern Pacific Ocean has been used since late 1997 at the University of Washington.

TABLE 1. Summary of initial condition and lateral boundary condition sources. The ''computational resolution'' is the grid spacing (in km) or truncation (Tnumber) and the number of vertical levels (Lnumber) on which the native model is computed. The format of the model output that was gathered at the University of Washington and used to initialize the MM5 is summarized under ''analysis grid spacing.''

Model (source)	Computational resolution	Analysis grid spacing	Data assimilation system	Data cutoff time
AVN (NCEP)	T170/L42	1.25° × 1.25°/L10	Spectral statistical interpolation/GDAS	2 h 45 min
GEM (CMC)	T199/L28	0.9° × 0.9°/L16	3DVAR	3 h 00 min
Eta (NCEP)	32 km/L45	90 km/L37	3DVAR/EDAS	1 h 30 min
NGM (NCEP)	80 km/L16	82 km/L19	RDAS*	2 h 00 min
NOGAPS (FNMOC)	T159/L24	1° × 1°/L19	Optimal interpolation	3 h 00 min

* Changed to EDAS on 15 March 2000.

sufficiently large that features at the western boundary do not generally make it to the 12-km nested domain within 48 h. Additionally, Steed et al. (2000) demonstrate that during an event where strong synoptic forcing and cross-boundary flow are evident, interior solutions exhibit very small differences due to changes in inner domain extent.

In the vertical, there are 33 unevenly spaced sigma levels with maximum resolution in the boundary layer.[2] The model top ($\sigma = 0$) is located at 100 hPa. The radiation scheme accounts for longwave and shortwave radiation interactions with both explicit cloud and clear air. The Kain–Fritsch convection parameterization scheme (Kain and Fritsch 1990), Medium-Range Forecast (MRF) planetary boundary layer scheme (Hong and Pan 1996), and the explicit moisture scheme of Hsie et al. (1984) with improvements for ice-phase microphysics (Dudhia 1989) are used. The sea surface temperatures are obtained from the FNMOC's Optimal Thermal Interpolation System (OTIS) data available at 1/4° resolution. Snow cover data are provided by the U.S. Air Force at roughly 48-km resolution. Consequently, initial conditions and time-dependent lateral boundary conditions are the only degrees of freedom between ensemble members.

b. Initial and boundary condition selection

Each day, five forecasts are initialized using different 0000 UTC synoptic-scale analyses. Sources for the five initializations are: the Aviation Model (AVN) Global Data Assimilation System (GDAS; Parrish and Derber 1992), the Nested Grid Model (NGM) Regional Data Assimilation System (RDAS; Petersen et al. 1991), the Eta Data Assimilation System (EDAS; Nelson 1999), FNMOC's U.S. Navy Operational Global Atmospheric Prediction System (NOGAPS) analysis, and CMC's Global Environmental Multiscale (GEM) analysis (Mitchell et al. 1996). Using these initializations, the MM5 forecasts are run out to 48 h with updated boundary conditions drawn from the corresponding, dynamically evolving, synoptic-scale forecasts. The lateral boundary conditions are supplied with an update frequency of once every 6 h, except for the LBCs from the Eta Model, which are updated every 3 h. Data cutoff times and the grid spacing of the five initializations vary. A comparison of the analysis grid spacing, data assimilation techniques, and data cutoff times for each model is found in Table 1. Although three of the initializations originate from NCEP, those analyses are created using different data assimilation systems for the first half of the experiment. The NGM RDAS was modified by NCEP on 15 March 2000 to share the Eta data assimilation system as well as boundary conditions from the AVN model. Thus, the only difference between the Eta and NGM initializations for the latter half of the experiment is resolution.

The five-member ensemble applied in this study may be enough to realize a large portion of the improvements obtainable through ensemble averaging. For example, Leith (1974) and Buizza and Palmer (1998) found that 8–10 members provide substantial benefits. The precise number of ensemble members required for useful results may largely depend on the meteorological problem of interest as well as the quality of the constituents. For short-term mesoscale forecasts, Du et al. (1997) noted that five members improved the Ranked Probability Score (RPS; Epstein 1969; Murphy 1971) of probabilistic quantitative precipitation forecasts (PQPF) by about 63% (while 10 members yielded about 90%) of the potential improvement attainable by an infinite number of ensemble members.

By using the multianalysis approach to IC selection, it is theorized that the probability of capturing the initial state of the atmosphere over the eastern Pacific is enhanced. But even if the correct initial state is not captured, the differences between the ICs may reveal the degree of uncertainty in the analyses. As noted above, the multianalysis approach may realize a significant portion of the improvement associated with multimodel, multianalysis ensembles with the multianalysis approach being easier to implement.

There are some inherent limitations accompanying the multianalysis approach. First, the ICs are not equally likely. Some analyses may be consistently more accurate

[2] The 33 full-sigma levels used are: $\sigma = 1.0, 0.99, 0.98, 0.97, 0.96, 0.94, 0.92, 0.90, 0.88, 0.86, 0.83, 0.80, 0.77, 0.74, 0.71, 0.68, 0.64, 0.60, 0.56, 0.52, 0.48, 0.44, 0.40, 0.36, 0.32, 0.28, 0.24, 0.20, 0.16, 0.12, 0.08, 0.04, 0.0$.

TABLE 2. Number of cases by domain and period. A case is defined by having all five completed ensemble forecasts and their verification data. Complete cases were established on 102 days out of 178 possible. Ten fewer cases are available for the 36-km ensemble forecasts, since their model output was lost.

	36-km domain	12-km domain	Both	No. of days in period
Winter (Jan–Mar 2000)	46	53	43	87
Spring (Apr–Jun 2000)	49	49	49	91
Phase I (Jan–Jun 2000)	95	102	92	178

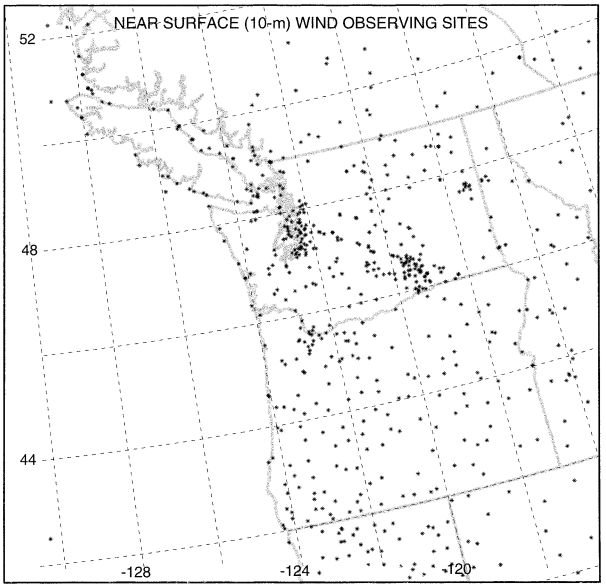

FIG. 2. Distribution of locations over the 12-km domain with near-surface wind-observing capabilities. Total number of sites is 673.

than others because of the different data assimilation procedures and model dynamics. Second, only a finite number of independent analyses are available. Third, the approach relies on products created by the various meteorological centers, which may evolve with time. For example, because of the changes to the NGM data assimilation procedure midway through the experiment, the Eta and NGM initializations became highly correlated. Such similarity among member ICs is undesirable when the goal is to independently and extensively sample the initial probability distribution. Last, since using multiple analyses does not constrain the differences between the ICs to be orthogonal, complete independence among the IC perturbations is not guaranteed. In fact, some similarities exist among the disparate data assimilation systems. For example, there are parallels between the Spectral Statistical Interpolation (SSI) technique used in the GDAS and the use of Three-Dimensional Variational Analysis (3DVAR) in the GEM analysis. In addition, the RDAS and EDAS use the GDAS analysis as a first-guess field.

3. Verification method

During the 178-day period from 5 January to 30 June 2000, complete ensemble forecasts (all five members) are available on 102 days. The full sample size of 102 cases is included for all statistics involving the 12-km domain ensemble forecasts. When intercomparison between the 36-km and 12-km domains is necessary, 10 cases are discarded since verification data for the 36-km domain on those days was lost (Table 2). In this experiment, the existing observation-based verification system at the University of Washington, used for validation of MM5 model forecasts over the past four years (Colle et al. 2000; Mass et al. 2002), also validates the ensemble forecasts.[3] Model forecast data at the four grid points surrounding each observation location are bilinearly interpolated to the observation site. Verification is limited to near-surface weather parameters [sea level pressure, surface (2 m) temperature, surface (2 m) relative humidity, and surface (10 m) wind], following recommendations from the most recent ensemble workshop (Hamill et al. 2000) and because of an insufficient number of verifying locations aloft.

Scarcity of data presents a difficult problem for mesoscale verification. Sea level pressure observations are generally located at low-elevation airports and fail to sample mesoscale pressure features that are located within or near major mesoscale terrain features. Relative humidity observations are also sparse. Near-surface temperature observation density compares favorably with wind observation density; however, temperature is relatively uniform over the Northwest (discounting elevation differences) and often changes little in response to strong synoptic-scale features. For example, wintertime frontal passages are often associated with little temperature change, because of the moderating ocean influence.

In this paper, 10-m wind direction is chosen for presentation because of 1) its more extensive areal coverage and greater number of reporting sites in comparison with the other surface variables over the domain of interest, 2) the fact that wind direction is greatly influenced by regional orography and synoptic-scale changes, and 3) the MM5's systematic biases in the other near-surface variables tending to dominate the error originating from the IC uncertainty. The geographic distribution of the 673 observation sites within the 12-km domain with surface wind-observing capabilities is presented in Fig. 2. The number of stations reporting each hour varies, but averages between 517 and 574. Wind direction forecast–observation pairs are excluded from the statistics when the observed wind speeds are less than 5 kt (2.57 m s^{-1}), since wind direction is poorly observed at such

[3] See http://www.atmos.washington.edu/mm5rt/verify.html for an introduction to the verification system at the University of Washington.

low wind speeds. The total number of observations used in the statistics ranges from 12 833 (for forecast hour 9, or F09) to 25 963 (F48) for the 92 cases available for both domains and from 14 166 (F09) to 27 237 (F24) for all 102 cases.

Standard measures of forecast skill are computed for each component member forecast and ensemble mean forecast, including mean error (bias), mean absolute error (mae), and root-mean-square error (rmse). For brevity, only the mean absolute error scores are presented here. The mae is calculated by taking an average of the absolute errors at each observational point within the domain. To allow for fair comparison between the 36-km forecasts and the 12-km forecasts, performance scores only include observations from the domain common to both forecasts—the 12-km domain in this case.

The relationship between ensemble spread and forecast error is calculated by the linear correlation between the standard deviation of the ensemble forecasts and the absolute error of the ensemble mean or component forecasts. The absolute errors and standard deviations are averaged over the 12-km domain for each case, yielding two time series. The spatially averaged wind direction error is calculated at every third forecast hour by

$$\mathrm{mae}_\theta(i) = \frac{1}{N_{\mathrm{obs}}(i)} \sum_{n=1}^{N_{\mathrm{obs}}(i)} [|\theta_n^{\mathrm{obs}}(i) - \overline{\theta}_n(i)|], \qquad (1)$$

where $\theta_n^{\mathrm{obs}}(i)$ denotes the observed wind direction at observation location n for case i, $\overline{\theta}_n(i)$ is the ensemble mean forecast at location n, and $N_{\mathrm{obs}}(i)$ is the total number of observations in the domain for case i. Directional differences are constrained to have a maximum of 180° [e.g., if $\theta_n^{\mathrm{obs}} = 5°$ (NNE) and $\overline{\theta}_n = 355°$ (NNW), then the difference is $-10°$, where negative differences denote a counterclockwise forecast bias]. Forecast spread is quantified by finding the spatially averaged standard deviation of the wind direction forecasts at every third forecast hour:

$$\overline{\sigma}_\theta(i) = \frac{1}{N_{\mathrm{obs}}(i)} \sum_{n=1}^{N_{\mathrm{obs}}(i)} \left\{ \frac{1}{M-1} \sum_{m=1}^{M} [\theta_n^m(i) - \overline{\theta}_n(i)]^2 \right\}^{1/2}, \qquad (2)$$

where $\theta_n^m(i)$ is the wind direction forecast by ensemble member m at location n for case i, and M is the total number of ensemble members. Correlation coefficients are then found that represent the linear relationship between the two time series. A high degree of correlation implies that the magnitude of forecast spread is indicative of the magnitude of forecast error, and thus an ability to predict forecast skill exists.

4. Results

To evaluate whether the ensemble mean forecasts exhibit greater mesoscale skill than the individual forecasts that compose the ensemble, mean absolute error in 10-m wind direction is used as the metric. The case-averaged mae for the forecasts with 36-km and 12-km grid spacing for both the ensemble mean and the component members are shown every 3 h in Fig. 3. It is apparent that the ensemble means exhibit lower mae values than the component forecasts for nearly all lead times at both grid spacings. The separation between the 12-km ensemble mean errors and the 12-km member forecast errors is noticeably greater than the separation between the 36-km mean and member forecast errors, suggesting that ensemble averaging helps more at higher resolution. Differing skill of the individual member forecasts is also evident. The MM5 forecast forced by the AVN model (hereinafter AVN–MM5) appears superior at lead times greater than about 30 h, but it is in the middle of the pack at the earlier lead times. The NOGAPS–MM5 member tends to exhibit the highest error scores beyond about F15, but not from F00–F12. Since all runs began at 0000 UTC, the MM5's systematic diurnal biases are apparent, with errors tending to be higher during the early morning and lower in the late afternoon.

To investigate how often the ensemble mean forecasts verify better or worse than the component forecasts, the forecasts are tallied on a case-by-case basis. Fig. 4a shows the percentage of time each 12-km component forecast or the 12-km ensemble mean forecast verified with the lowest mae in wind direction. While the percentages fluctuate by forecast hour, the ensemble mean forecast verifies as the best forecast with about the same frequency as any individual member forecast. The lone exception is at the initialization time, when the Eta initialization seems to draw closest to the observed 10-m wind directions. The significantly higher resolution computational grid employed by the Eta Model may partially account for the lower error scores seen here. It is interesting to note that, at all times except initialization, the various members have nearly equal frequencies of being the best forecast.

In contrast, examining the frequency of forecasts having the highest (least skillful) mae (Fig. 4b) shows that the ensemble mean forecasts never verify worst, with the component forecasts having nearly equal frequencies. That the ensemble mean forecast never verifies worst is an expected result that can be shown to be universally true for any ensemble system. The NOGAPS–MM5 tends to be the worst member after F12, but not prior to that lead time. As seen in the best-member frequencies, all member forecasts appear to have a nearly equal opportunity to verify with the highest mae, except at the initialization time.

The relative increase in forecast skill due to ensemble averaging can be compared with the increase in skill provided by increasing horizontal resolution. In Fig. 5, the colored solid lines represent the 10-m wind direction mae values from Eta–MM5 forecasts for 36-, 12-, and 4-km grid spacings. The 36- and 12-km statistics presented in Fig. 5 are calculated in the same way as those presented in Fig. 3, except that the errors are averaged only over stations within the smaller 4-km domain en-

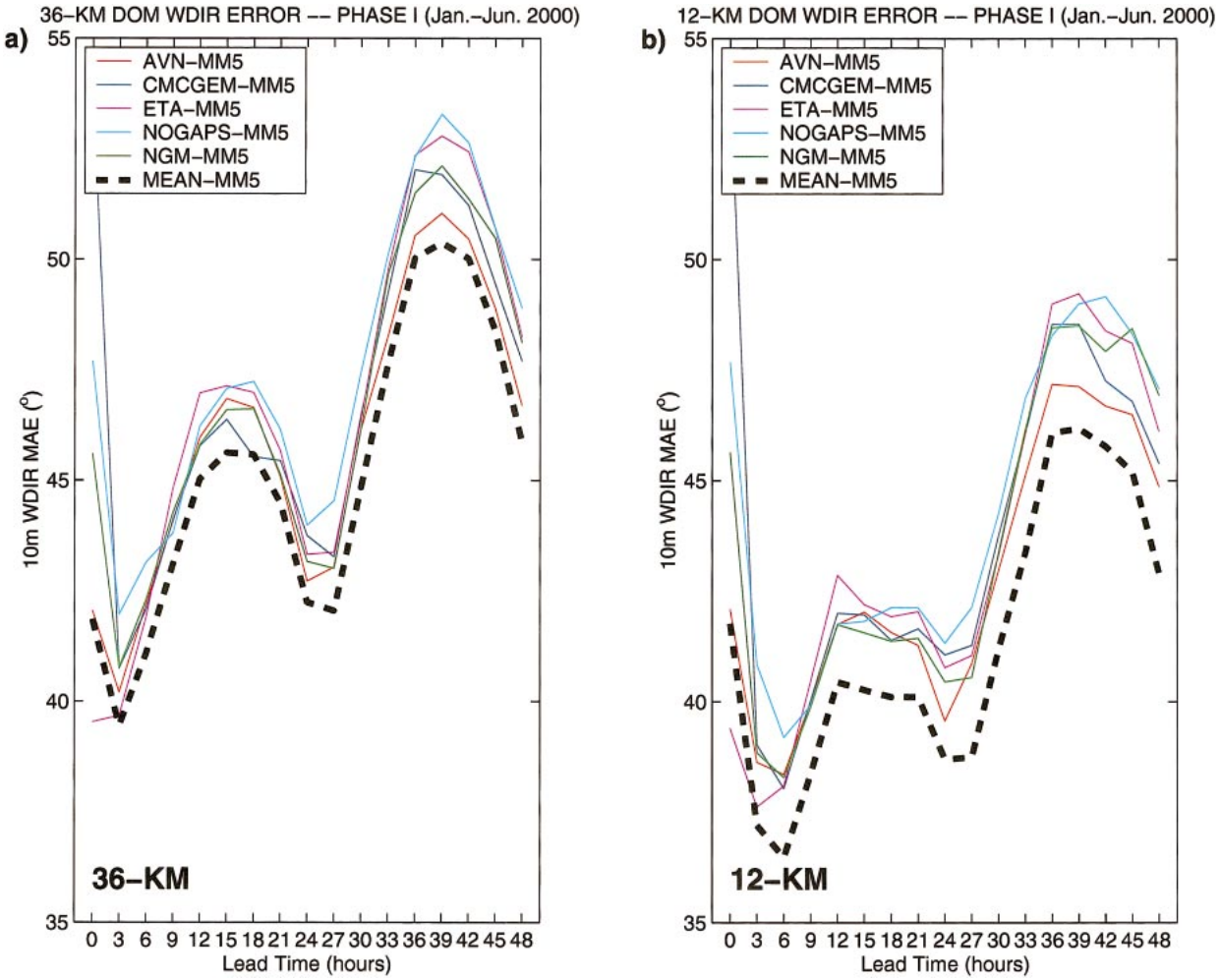

FIG. 3. Mean absolute errors for 10-m wind direction over the 12-km domain for the (a) 36-km ensemble member forecasts and 36-km ensemble mean forecasts, and (b) 12-km ensemble member forecasts and 12-km ensemble mean forecasts.

compassing Washington State and only for F06–F36. As noted by Mass et al. (2002), the increase in skill going from 36 to 12 km is far larger than the change from 12 to 4 km. The ensemble mean forecasts (Fig. 5, black lines) for both 36- and 12-km grid spacing improve upon the Eta–MM5 forecasts at the same grid spacing. On average, 12-km ensemble mean forecasts are as good as the 4-km Eta–MM5 forecasts. At lead times beyond about 21 h, the 12-km ensemble means exhibit lower error scores than the 4-km forecasts.

While the ensemble mean forecasts display lower error scores over a number of cases and never verify as the worst forecast, one may ask whether the ensemble mean is simply a smoothed forecast devoid of realistic mesoscale features. Is mesoscale forecast information maintained by the ensemble mean forecasts? To illustrate an example of mesoscale information retention by the ensemble mean, the forecast 3-h accumulated precipitation over the 12-km domain at forecast hour 30 from the ensemble run initialized at 0000 UTC 13 March 2000 is shown in Fig. 6. Sea level pressure isopleths are also shown to indicate the location and intensity of the forecast surface trough. It is obvious that the location, timing, and intensity of the trough and associated precipitation features vary substantially among the component forecasts. The ensemble mean forecast smoothes out the differences in precipitation due to the phase differences in the trough location but retains the features common to all the component members, such as the orographic precipitation on the windward (southwest) side of the Olympic Mountains and the precipitation minimum to the northeast of the Olympics. Differences among the component members at the finer scales are also smoothed. It appears that the ensemble mean forecasts retain useful mesoscale information that can be beneficial to forecasters.

The mesoscale ensemble system's ability to *predict* forecast skill is summarized in Figs. 7–9. All calculations are based on the full set of ensemble forecasts at 12-km grid spacing (102 total cases). As shown in Fig.

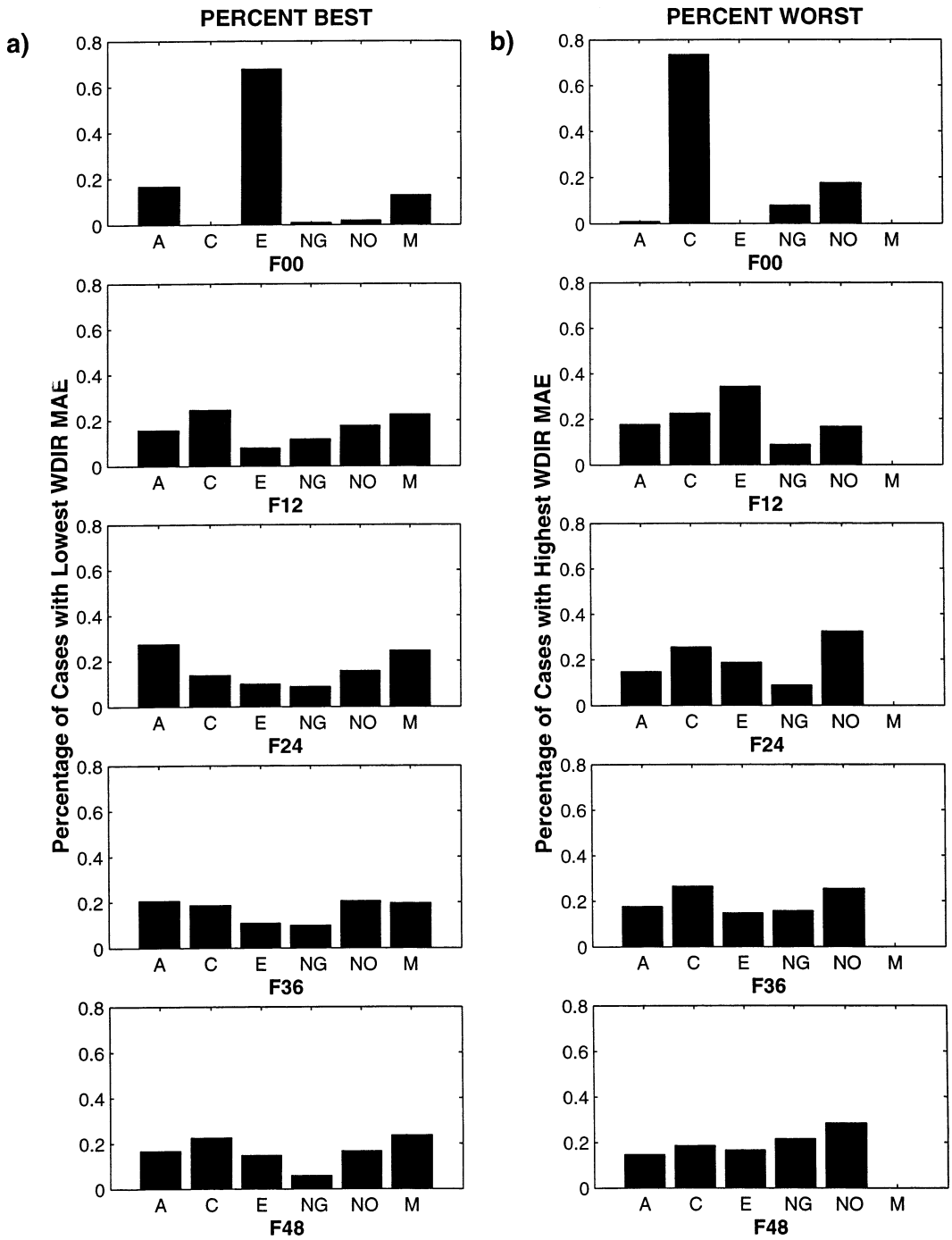

FIG. 4. (a) The percentage of cases (out of 102) that each 12-km member forecast and 12-km mean forecast had the lowest mean absolute wind direction error. (b) Same as (a) except that the percentages represent the number of cases that each ensemble member or the ensemble mean had the highest mean absolute wind direction error. (A = AVN, C = CMC–GEM, E = Eta, NG = NGM, NO = NOGAPS, M = Mean).

7a, correlations between ensemble spread and the mae of the ensemble mean for all wind direction forecasts (solid line) are approximately 0.6 at most lead times, explaining about 36% of the variance in the data. This number is encouraging since most previous studies have shown the highest spread–error correlations to be around 0.4. Houtekamer (1993) and Whitaker and Loughe (1998) both suggested that spread–error correlation tends to be a better predictor of skill when the spread is extreme. To test this hypothesis, the ensemble dataset was divided according to the spread in the ensemble forecasts. When one-third of the cases with medium

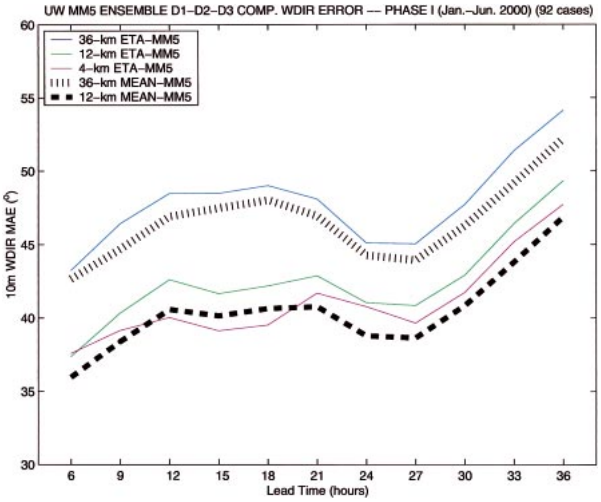

FIG. 5. Mean absolute wind direction forecast errors for varying horizontal grid spacings and for the ensemble means. The mae is calculated only over the 4-km domain (Washington State) illustrated in Fig. 1.

spread are taken out of the dataset, leaving only the extreme two-thirds of the cases, the spread–error correlations improve noticeably, reaching slightly higher than 0.7 (Fig. 7a, dash–dotted line). The spread–error correlations for the nonextreme cases (the middle one-third) are lower than for the full dataset, generally below 0.4 (Fig. 7a; dashed line).

Plotting 95% confidence limits for both the extreme and nonextreme cases shows that nearly all of the spread–error correlations for the extreme cases are significantly higher than those for the nonextreme cases (Fig. 7b). The same confidence limits also show that about one-half of the spread–error correlations from the middle one-third are not significantly different from zero. For the nonextreme cases, the forecast spread gives little useful guidance regarding error magnitude.

To further explore the correlation between forecast spread and error for wind direction, cases are separated by spread into three categories and their mae values are averaged (Fig. 7c). As expected, cases with large (small) spread also have large (small) forecast errors. Specifically, the extreme high-spread group has wind direction errors that average 7°–15° larger than the extreme low-spread group.

What happens to the spread–error correlations when even more cases are filtered out and only the cases with the most extreme spread are considered? By halving the number of cases in the extreme categories (from two-thirds to one-third of all cases), the spread–error correlations jump to near 0.8 (Fig. 8a, dashed line). Spread–error correlations for the remainder of events also rise, becoming significantly higher than zero, since some of

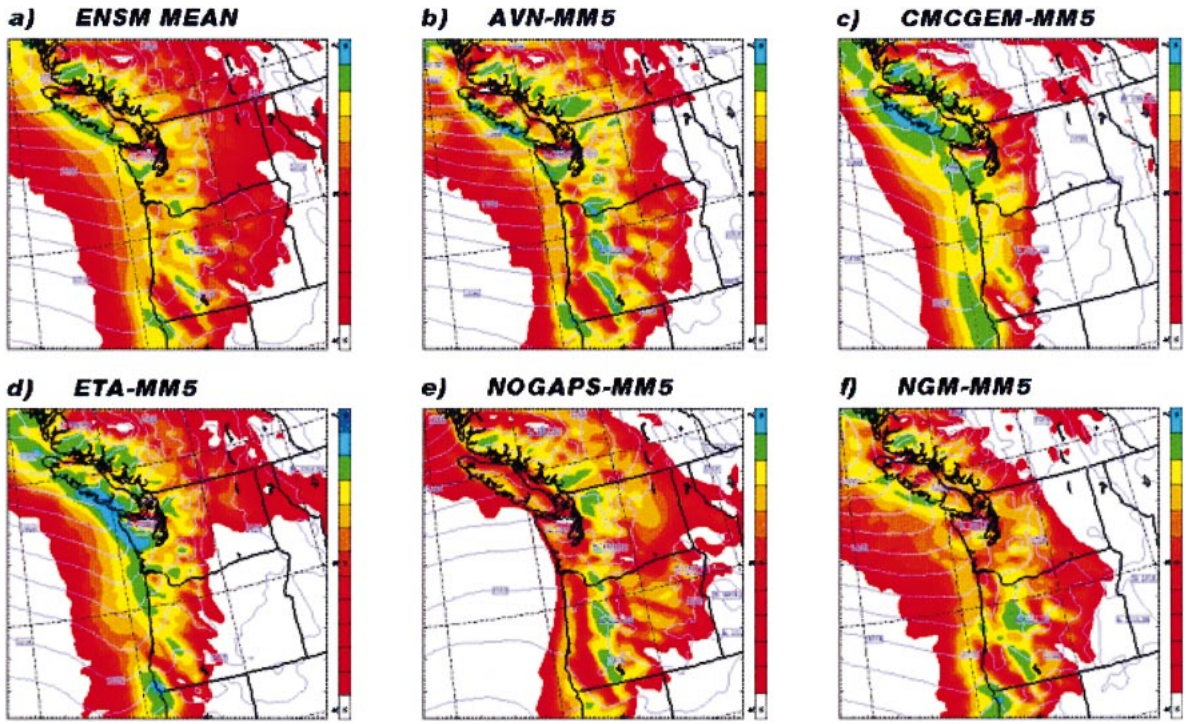

FIG. 6. The 3-h accumulated precipitation (color-filled) and sea level pressure (contours) for forecast hour 30 of the ensemble run initialized at 0000 UTC 13 Mar 2000 for: (a) the ensemble mean, (b) the AVN–MM5, (c) the CMCGEM–MM5, (d) the Eta-MM5, (e) the NOGAPS–MM5, and (f) the NGM–MM5. The scale for precipitation amount ranges from trace amounts (0.1 mm) in red to heavy amounts (~30 mm) in blue.

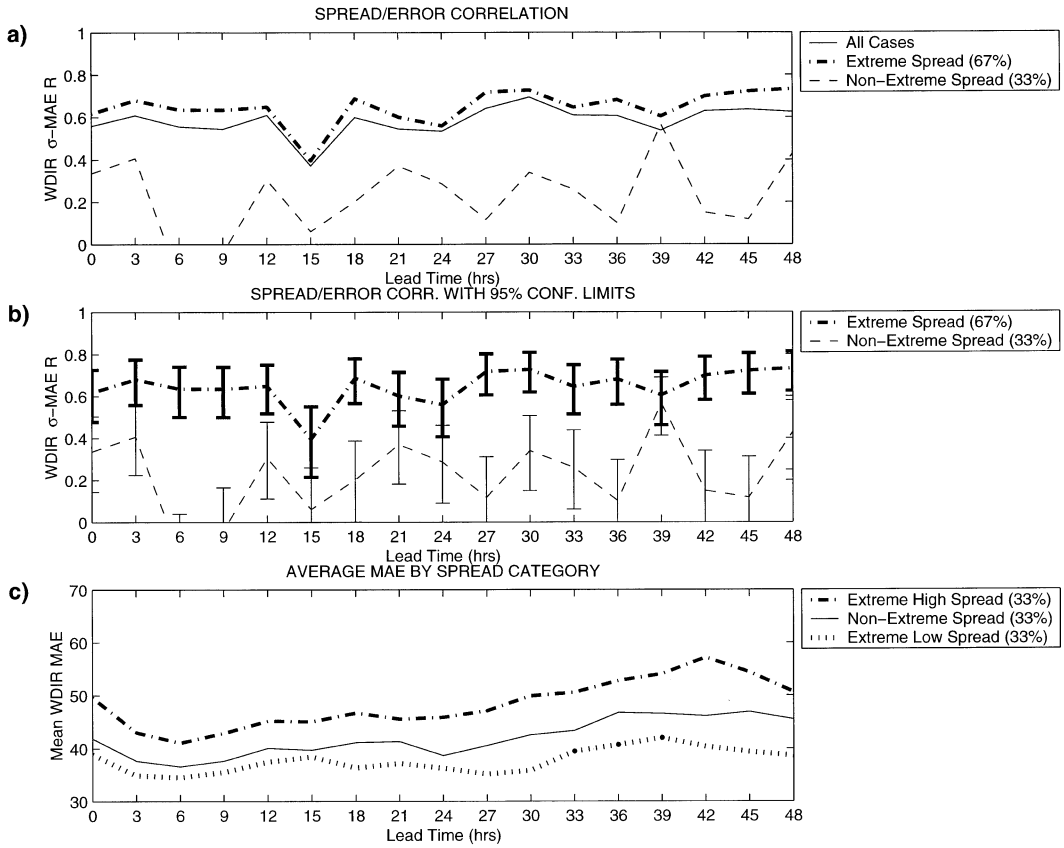

FIG. 7. (a) Spread–error correlations for 12-km wind direction forecasts every 3 h into the forecast. All 102 cases are included in the ''All Cases'' solid line. The dash–dotted line denotes the spread–error correlations for cases with extreme spread (highest and lowest one-third). The dashed line denotes the spread–error correlations for the nonextreme cases (middle one-third). (b) Spread–error correlation coefficients plus their 95% confidence limits for the extreme and nonextreme cases. (c) The case-averaged ensemble mean mae for the one-third of cases with extreme high spread, the one-third of cases with extreme low spread, and the middle one-third (nonextreme cases).

the previous extreme spread events are now included (Fig. 8b, dash–dotted line). Even though the spread–error correlations for the nonextreme cases are higher than before, they are still significantly lower than the correlations for the extreme cases. In fact, the 95% confidence limits show that there is a significant difference between extreme and nonextreme cases at nearly all forecast lead times (Fig. 8b).

When the wind direction forecast errors are separated by spread according to the more highly filtered configuration, the differences between spread categories become even more pronounced (Fig. 8c). The wind direction errors for the extreme high-spread group are 10°–25° larger than for the extreme low-spread group. The average forecast errors for the highest one-sixth are larger than those for the highest one-third (Fig. 8c vs Fig. 7c). As suggested by the high spread–error correlations for extreme spread cases, the magnitude of the spread is likely to be indicative of the magnitude of forecast error.

An important question is whether the errors of the individual ensemble members, not just the ensemble mean, are correlated with the forecast spread. Spread–error correlations for the component members of the ensemble are similar in magnitude to the spread–error correlations for the ensemble mean (not shown). Filtering out the intermediate-spread cases increases the spread–error correlations for all the component forecasts. As seen in Fig. 9, the average mae values separated by spread magnitude for each ensemble member follow the same pattern found for the ensemble mean. Thus, the ensemble mean is more (less) skillful for extremely low (high) spread situations because each component member is also more (less) skillful. *This implies that low-spread events are essentially more predictable and that high-spread events are essentially less predictable.*

A qualitative evaluation of the synoptic patterns associated with extreme-spread events reveals that for high-spread events there is no particular associated flow pattern. The most common flow pattern associated with low-spread events is ridging at 500 hPa either over the offshore waters or directly over the Pacific Northwest. A number of low-spread events are characterized by a

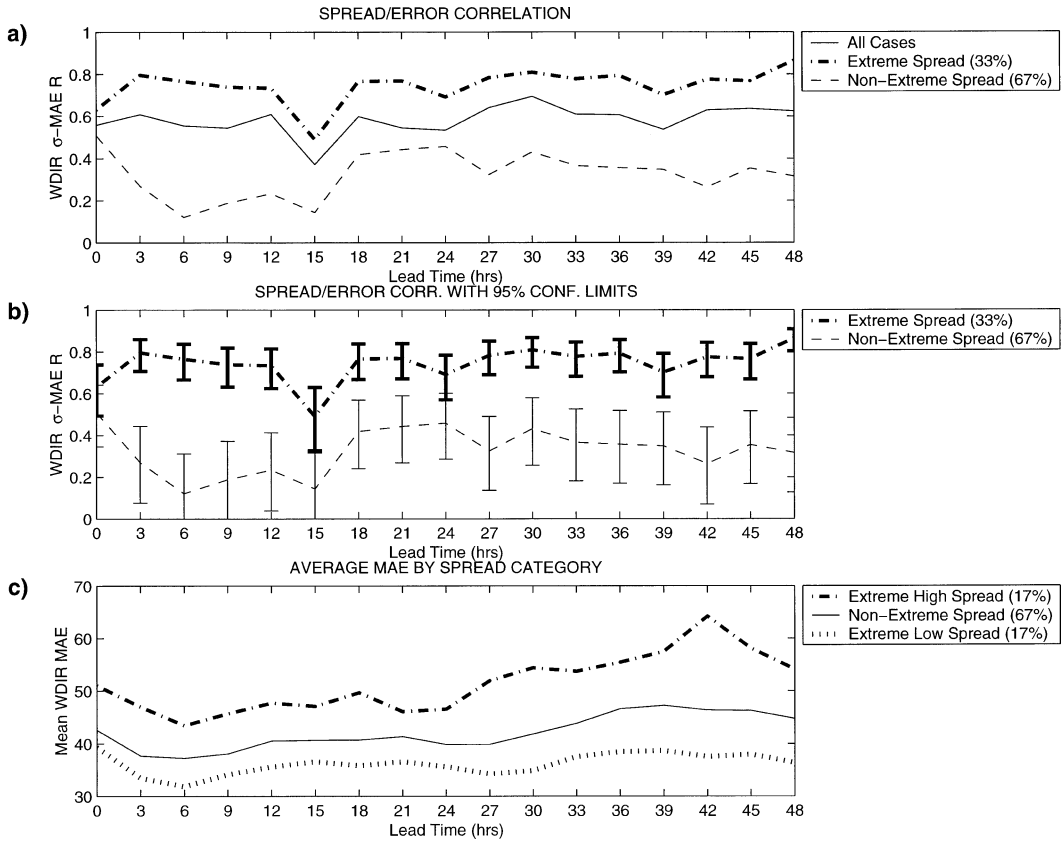

FIG. 8. Same as in Fig. 7 except that the middle cases now make up two-thirds of the data. (c) The case-averaged ensemble mean mae is for the one-sixth of cases with extreme high spread, the one-sixth of cases with extreme low spread, and the middle two-thirds (nonextreme cases).

small-amplitude, short-wave ridge building over the region as a short-wave trough propagates downstream. This pattern is generally associated with a post–cold frontal environment and northwesterly flow at the surface. However, low-spread events also occur in conjunction with strong southwesterly flow and embedded upper level troughs in which all ensemble members are in close agreement.

5. Summary and conclusions

An initial, six-month evaluation of a five-member MM5 ensemble prediction system over the Pacific Northwest suggests that mesoscale short-range ensemble forecasting in which one model is driven by multiple initializations and forecast lateral boundary conditions from several major operational weather centers is a useful forecast tool for the region. The ensemble mean forecasts verify with lower mean absolute wind direction errors than the component members of the ensemble. The 12-km grid spacing ensemble mean forecasts perform as well as 4-km deterministic forecasts, suggesting diminishing returns as grid spacing is decreased. Although MM5 ensemble mean forecasts generally have lower average wind direction error scores than the component forecasts, for any individual event the ensemble mean forecast displays no tendency to verify as the best forecast. While ensemble mean forecasts verify as the best forecast with about the same frequency as each component forecast, the ensemble mean forecasts never verify as the worst forecast. Ensemble mean forecasts in this experiment retain much of the orographically forced mesoscale structure in the component forecasts while smoothing out finer-resolution details and differences in structures associated with propagating features such as fronts. Even though the ensemble mean is only a minimal usage of ensemble predictions, forecasters can benefit from the useful information retained by the mean.

The real advantage of ensemble predictions over deterministic forecasts hinges upon their potential for providing probability forecasts and information about the predictability of forecast parameters. The University of Washington MM5 ensemble results suggest that a short-range ensemble system with only five members *can* predict forecast skill, as evidenced by high spread–error correlations in 10-m wind direction. Cases with very high or very low spread exhibit positive spread–error correlations as large as 0.8, in contrast to the lower correlations found in previous studies (Hamill and Col-

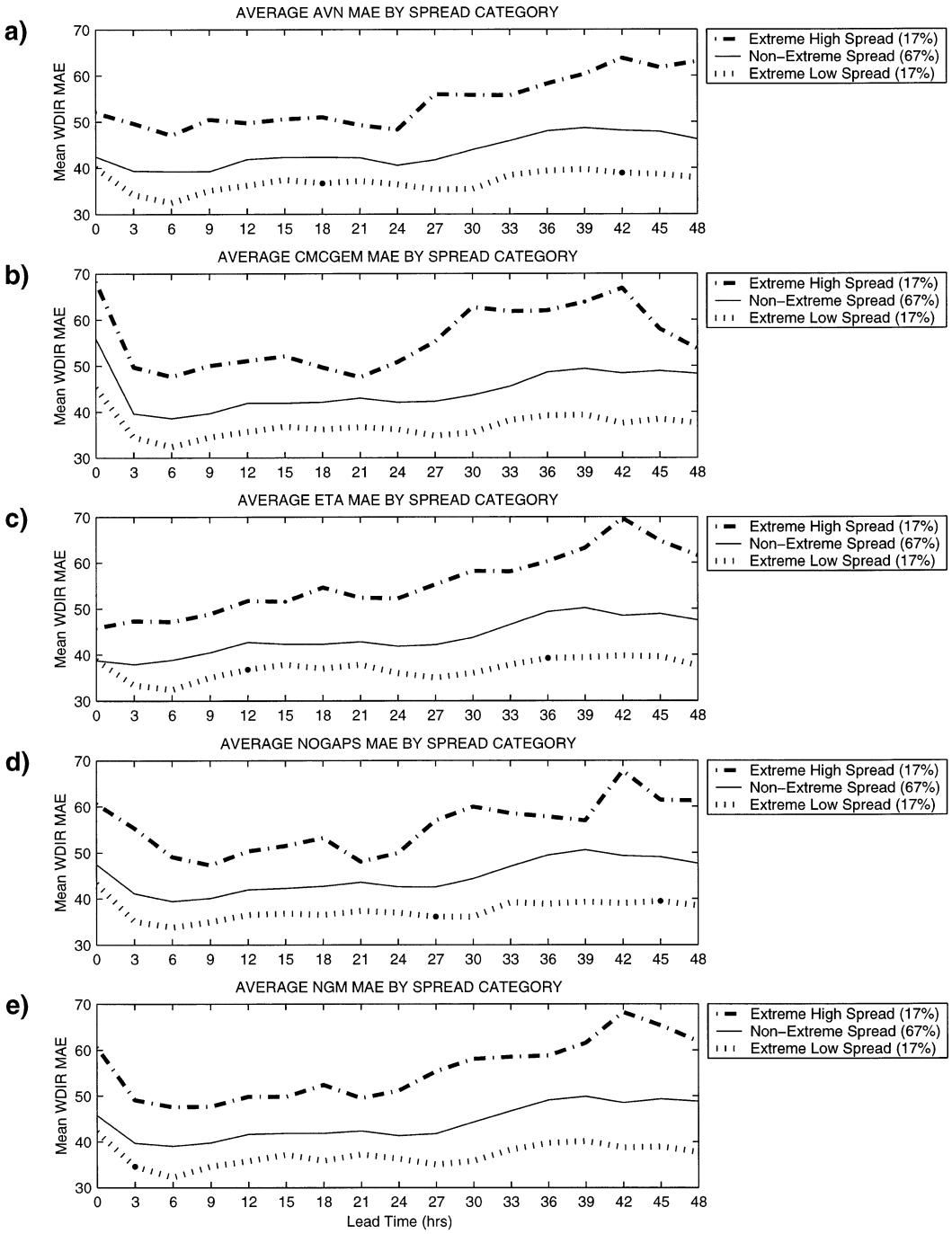

FIG. 9. Case-averaged mae for the (a) AVN–MM5 (b) CMCGEM–MM5 (c) Eta–MM5 (d) NGM–MM5 and (e) NOGAPS–MM5 forecasts. Errors are divided by spread categories as in Fig. 8.

ucci 1997; Stensrud et al. 1999; Hou et al. 2001). Thus, events characterized by small spread are likely to verify better than cases with large spread. For the cases characterized by an intermediate amount of spread, the spread–skill relationship is weak, with spread–error correlations generally less than 0.4 and often not significantly different from zero. In those situations, the ensemble does not give any useful information about the reliability of the forecast. Because the ensemble spread is also well correlated with the skill of each component member of the ensemble, low- (high) spread events appear to be essentially more (less) predictable.

The optimal approach for mesoscale ensemble forecasting over the Pacific Northwest may be different than

for other parts of the country. In the Pacific Northwest, convection is typically weaker, shallower, and less frequent, and mesoscale variability is primarily forced by the regional orography. Model deficiencies are hypothesized to be less important than over the eastern United States since mesoscale models appear to do well with orographic circulations (Mass et al. 2002). However, with a vast data-sparse region over the eastern Pacific, large uncertainty in the upstream synoptic-scale initial conditions often exists. For the Pacific Northwest, it is desirable to sufficiently sample the initial state before moving on to address the concerns brought about by model deficiencies. Therefore, the multianalysis approach is considered a logical first step for ensemble forecast generation in this region.

As noted by Fritsch et al. (2000), ensemble mean forecasts are adversely affected by the particularly poor performance of one of the component members. If one member forecast consistently performs worse than others, the quality of the ensemble mean forecast is degraded. In this experiment, since all of the constituent forecasts are not equally likely solutions, the intrinsically different levels of skill among members (apparent in Fig. 3) should be considered. It is possible that an improved forecast might be created using a weighted mean based on the relative skill of the individual members (Van den Dool and Rukhovets 1994). In addition, the relative skill of each member forecast could be broken down by flow regime. However, judging ensemble performance solely based on the skill of its mean forecast is insufficient. Wandishin et al. (2001) show that the addition of less skillful members actually improves the ranked probability skill scores and relative operating characteristic curves (Stanski et al. 1989). Another approach might be to initialize a high-resolution forecast with the same ICs as the ensemble member that either performs the best overall or is performing better than the rest over a recent period or on a given day. In this way, high-resolution forecasts and ensemble forecasts can complement each other to maximize the benefits of each.

Creating a temporal ensemble by augmenting the current component forecasts with forecasts using older initial and boundary conditions from the same synoptic-scale modeling systems may have additional benefits. A temporal ensemble, formulated similarly to lagged-average forecasting (Hoffman and Kalnay 1983), might be constructed in such a way as to derive a measure of model consistency. Model consistency is often used as an informal tool for operational weather forecasting, and its relative merits have not been quantified.

In future work we will examine the value of the MM5 ensemble system for creating forecast PDFs for the Pacific Northwest. Owing to the unequal likelihood of forecasts produced by a multianalysis ensemble such as this one, obtaining reliable estimates of the PDF may not be feasible. Even if it is not possible to use the ensemble system to produce calibrated probability forecasts, valuable information about the reliability of the forecasts may be gleaned from the ensemble spread. If ensemble spread is higher (lower) than its climatological average, then a forecaster may place less (more) emphasis on the information given by the high-resolution forecast. To develop a prediction system for forecast skill, more robust spread–error statistics are needed. A record longer than the six-month period studied here is needed to establish an annual climatology for both spread and error. If reference levels of spread can be found, those thresholds can be used to generate an operational forecast skill prediction system for each forecast parameter. Additionally, spread–error relationships must be further investigated for near-surface variables other than just wind direction.

Acknowledgments. This research was supported by the Marine Meteorology Program at the Office of Naval Research (N00014-98-1-0193) and the National Weather Service C-STAR Program (NOAA Cooperative Agreement NA67RJ0155). We especially thank Brad Colman (NOAA/NWS Seattle) for his encouragement to get this project started as well as his careful review of this manuscript. F. Anthony Eckel (U.S. Air Force and University of Washington) reviewed a prior version of this work and provided helpful input. Additionally, Jun Du (NCEP) and two anonymous reviewers contributed to the final form of this paper with insightful suggestions and comments.

REFERENCES

Atger, F., 1999: Tubing: An alternative to clustering for the classification of ensemble forecasts. *Wea. Forecasting,* **14,** 741–757.

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

——, ——, and R. A. Maddox, 1992: On the use of mesoscale and cloud-scale models in operational forecasting. *Wea. Forecasting,* **7,** 120–132.

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. N. Palmer, 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.,* **126,** 2503–2518.

Colle, B. A., C. F. Mass, and K. J. Westrick, 2000: MM5 precipitation verification over the Pacific Northwest during the 1997–99 cool seasons. *Wea. Forecasting,* **15,** 730–744.

Droegemeier, K. K., 1998: SAMEX: The storm and mesoscale ensemble experiment. *CAPS News Funnel,* Spring, Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, OK, 1–2.

Du, J., and M. S. Tracton, 2001: Implementation of a real-time short range ensemble forecasting system at NCEP: An update. Preprints, *Ninth Conf. on Mesoscale Processes,* Ft. Lauderdale, FL, Amer. Meteor. Soc., 355–360.

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

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.

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

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

Fritsch, J. M., J. Hilliker, and J. Ross, 2000: Model consensus. *Wea. Forecasting,* **15,** 571–582.

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, 121 pp. [Available from MMM Division, NCAR, P.O. Box 3000, Boulder, CO 80307.]

Grumm, R. H., and A. L. Siebers, 1989: Systematic surface cyclone errors in NMC's nested grid model November 1988–January 1989. *Wea. Forecasting,* **4,** 246–252.

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

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

——, C. Snyder, S. L. Mullen, Z. Toth, and D. P. Baumhefner, 2000: Ensemble forecasting in the short to medium range: Report from a workshop. *Bull. Amer. Meteor. Soc.,* **81,** 2653–2664.

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

Hong, S.-Y., and H.-L. Pan, 1996: Nonlocal boundary layer vertical diffusion in a medium-range forecast model. *Mon. Wea. Rev.,* **124,** 2322–2339.

Hou, D., E. Kalnay, and K. K. Droegemeier, 2001: Objective verification of the SAMEX '98 ensemble forecasts. *Mon. Wea. Rev.,* **129,** 73–91.

Houtekamer, P. L., 1993: Global and local skill forecasts. *Mon. Wea. Rev.,* **121,** 1834–1846.

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

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.

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

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

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.,* **20,** 131–140.

——, 1969: The predictability of a flow which possesses many scales of motion. *Tellus,* **21,** 289–307.

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce better forecasts? The results of two years of real-time numerical weather prediction in the Pacific Northwest. *Bull. Amer. Meteor. Soc.,* 407–430.

Mitchell, H. L., C. Chouinard, and C. Charette, 1996: Impact of a revised analysis algorithm on an operational data assimilation system. *Mon. Wea. Rev.,* **124,** 1243–1255.

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

Mullen, S. L., 1994: An estimate of systematic error and uncertainty in surface cyclone analysis over the North Pacific Ocean: Some forecasting implications. *Wea. Forecasting,* **9,** 221–227.

——, 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.

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

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

Nelson, J. A., 1999: The Eta data assimilation system. WR-Tech. Attachment 99-14, 6 pp. [Available from National Weather Service Western Region, P.O. Box 11188, Salt Lake City, UT 84147.]

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

Petersen, R. A., G. J. DiMego, J. E. Hoke, K. E. Mitchell, and J. P. Gerrity, 1991: Changes to NMC's regional analysis and forecast system. *Wea. Forecasting,* **6,** 133–141.

Richardson, D. S., 2001: Ensembles using multiple models and analyses. *Quart. J. Roy. Meteor. Soc.,* **127,** 1847–1864.

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

Silberberg, S. R., and L. F. Bosart, 1982: An analysis of systematic cyclone errors in the NMC LFM-II Model during the 1978–79 cool season. *Mon. Wea. Rev.,* **110,** 254–271.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. Environment Canada Research Rep. 89-5, 114 pp. [Available from Atmospheric Environment Service, Forecast Research Division, 4905 Dufferin St., Downsview, ON M3H 5T4, Canada.]

Steed, R., N. Bond, and C. Mass, 2000: The effects of domain size and lateral boundary condition update frequency on high resolution NWP. Preprints, *The Tenth PSU/NCAR Mesoscale Model Users' Workshop,* Boulder, CO, NCAR MMM Division, 66–69.

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

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

——, 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.

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

——, 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, P. D., 1977: How to improve accuracy by combining independent forecasts. *Mon. Wea. Rev.,* **105,** 228–229.

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,** 379–398.

Van den Dool, H. M., and L. Rukhovets, 1994: On the weights for an ensemble-averaged 6–10 day forecast. *Wea. Forecasting,* **9,** 457–465.

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,** 729–747.

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

Ziehmann, C., 2000: Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models. *Tellus,* **52A,** 280–299.