Probabilistic Precipitation Forecast Skill as a Function of Ensemble Size and Spatial Scale in a Convection-allowing Ensemble

Adam J. Clark[1], John S. Kain[1] , David J. Stensrud[1], Ming Xue[2,3], Fanyou Kong[2], Michael C. Coniglio[1], Kevin W. Thomas[2], Yunheng Wang[2], Keith Brewster[2], Jidong Gao[2], Xuguang Wang[2,3], Steven J. Weiss[4], and Jun Du[5]*

[1]NOAA/National Severe Storms Laboratory, Norman, OK 73072
[2]Center for Analysis and Prediction of Storms, and [3]School of Meteorology, University of Oklahoma, Norman, OK 73072
[4]NOAA/NWS/NCEP Storm Prediction Center, Norman, OK 73072
[5]NOAA/NWS/NCEP Environmental Modeling Center, Camp Springs, MD 20746

Submitted as an Expedited Article to
Monthly Weather Review
August 2010

* Corresponding author's address:
Adam J. Clark
National Weather Center, NSSL/FRDD
120 David L. Boren Blvd.,
Norman, OK, 73072
Email: adam.clark@noaa.gov
Phone: 405-325-6731

Abstract

Probabilistic quantitative precipitation forecasts (PQPFs) from the Storm-Scale Ensemble Forecast system run by the Center for Analysis and Prediction of Storms during spring 2009 are evaluated using area under the relative operating characteristic curve (ROC-area). ROC-area, which measure discriminating ability, is examined for ensemble sizes, *n*, from 1 to 17 members and for spatial scales ranging 4- to 200-km.

Expectedly, incremental gains in skill decrease with increasing *n*. Significance tests comparing ROC-areas for each *n* to those of the full 17 member ensemble revealed that more members are required to reach statistically indistinguishable PQPF skill relative to the full ensemble as forecast lead time increases and spatial scale decreases. These results appear to reflect the broadening of the forecast probability distribution function (PDF) of future atmospheric states associated with decreasing spatial scale and increasing forecast lead time. They also illustrate that efficient allocation of computing resources for convection-allowing ensembles requires careful consideration of spatial scale and forecast length desired.

1. Introduction

Motivated by promising results (e.g., Done et al. 2004; Kain et al. 2005) testing convection-allowing[1] configurations of the Weather Research and Forecasting (WRF) model (Skamarock et al. 2008) and a desire to quantify forecast uncertainty at convective scales, the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma began producing convection-allowing Storm Scale Ensemble Forecasts (SSEFs) for NOAA/Hazardous Weather Testbed Spring Experiments (Kain et al. 2003) in 2007. During 2007-2008, the SSEF systems were comprised of 10 WRF-ARW members (Kong et al. 2007, 2008 provides model configurations) and in 2009 ensemble size/diversification was increased by adding eight Nonhydrostatic Mesoscale Model (NMM; Janjic 2003) members and two Advanced Regional Prediction System (ARPS; Xue et al. 2001) members.

This study examines probabilistic quantitative precipitation forecasts (PQPFs) at different spatial scales and forecast lead times as a function of ensemble size, *n*, in the 2009 SSEF system. Specifically, we focus on one particular aspect of forecast skill – resolution or discriminating ability – and explore whether a point of "diminishing returns" is reached with increasing *n*. Clearly, such analyses are important in designing ensemble systems that efficiently utilize computing resources.

2. Data and methodology

a. Ensemble configuration

The 2009 SSEF members used 4-km grid-spacing, were initialized at 0000 UTC, and integrated 30 hours over a 3600 x 2700 km domain covering most of the contiguous US. A smaller sub-domain (~ 2000 x 2200 km; Fig. 1) centered over the central US is used in subsequent analyses to avoid lateral boundary condition (LBC) effects (e.g., Warner et al. 1997) and emphasize regions climatologically favored for strong, organized springtime convection. Results are aggregated over 25 cases between April 27 and June 5 (Fig. 1). Table 1 provides ensemble member configurations (see also Xue et al. 2009). Radial velocity and reflectivity data from WSR-88D radars and surface observations were assimilated into initial conditions (ICs) of 17 members using the ARPS 3DVAR (Xue et al. 2003; Gao et al. 2004) data and cloud analysis (Hu et al. 2006; Xue et al. 2008) system. Analyses from the 0000 UTC operational 12-km North American Mesoscale (NAM) model (Janjic 2003) were used as the analysis background. Three other members did not assimilate radar data so that impacts of the radar data assimilation could be isolated (Kain et al. 2010; only the 17 members assimilating radar data are

1 The term "convection-allowing" refers to simulations using the maximum grid-spacing (or below) at which convection can be treated explicitly and mid-latitude MCSs can be adequately resolved, which is generally thought to be ~ 4-km (Weisman et al. 1997).

used herein). IC/LBC perturbations were derived from evolved (through 3 hours) bred perturbations of 2100 UTC NCEP operational Short-Range Ensemble Forecast (SREF; Du et al. 2006) members and added to the ARPS 3DVAR analyses. Corresponding SREF forecasts were used for LBCs. To account for model physics uncertainty, different boundary layer, microphysics, radiation, and land surface schemes were used (see Table 1 for schemes and references).

b. Verification methods

NCEP's stage IV (Baldwin and Mitchell 1997) multi-sensor rainfall estimates are used to verify rainfall forecasts. The 4-km stage IV grids are remapped to the model grid using a neighbor-budget interpolation (e.g., Accadia et al. 2003). Statistical significance is determined using Hamill's (1999) resampling methodology.

PQPFs for 6-h accumulated rainfall computed using 1 to 17 members for different spatial scales are computed following Hamill and Colucci (1997, 1998), which basically involves finding the location of verification thresholds within the distribution of ensemble member forecasts. PQPFs for thresholds beyond the highest ensemble member forecast are obtained by assuming that the PDF in this region follows a Gumbel distribution (Wilks 1995). Different spatial scales are examined by averaging grid-points within circular regions with radii varying between 2-km (the raw model grid) and 200-km, similar to the "upscaling" methodology described by Ebert (2009). For the raw model grids, the 0.10-, 0.25-, and 0.50-in rainfall thresholds are verified. For the "upscaled" model grids, verification thresholds corresponding to the 0.10-, 0.25-, and 0.50-in. quantiles in the non-upscaled stage IV rainfall distribution (aggregated over all cases and grid-points) are used to allow equitable comparison among the different spatial scales. Thus, exceedance forecasts of constant rainfall quantiles, rather than amounts, are evaluated (e.g., Jenkner et al. 2008). Figure 2k illustrates how this procedure changes the verification thresholds with increasingly smoothed rainfall fields, and Figures 2a-j show how varying degrees of smoothing affect the appearance of the forecast probabilities and observed precipitation fields.

PQPFs are evaluated using the area under the relative operating characteristic curve (ROC-area; Mason 1982). The ROC-area measures the ability to distinguish between events and non-events (or resolution) and is calculated by computing the area under a curve constructed by plotting the probability of detection (POD) against the probability of false detection (POFD) for specified ranges of PQPFs. The area under the curve is computed using the trapezoidal method (Wandishin et al. 2001). The ranges of PQPFs used for ROC curves in this study are $P < 0.05$, $0.05 \leq P < 0.15$, $0.15 \leq P < 0.25$ … $0.85 \leq P < 0.95$, and $P \geq 0.95$. The range of ROC-area is 0 to 1, with 1 a perfect forecast and areas greater than 0.5 having positive skill. A ROC-area of 0.7 is generally considered the lower limit of a useful forecast (Buizza et al. 1999). Because the method used to compute PQPFs allows for continuous (rather than discrete) values between 0 and 100%, the same set of probabilities used to define the ROC curves can be used for every ensemble size, and problems associated with comparing ROC-areas between different size ensembles are avoided.

To study the effect of increasing n on PQPF skill, ROC-areas were computed for 100 unique combinations of randomly selected ensemble members for $n = 2, 3, \ldots 15$. For $n = 1$, 16 and 17, ROC-areas were computed for all possible combinations of members because the number of unique member combinations for these n is smaller than 100.

3. Results

Results show that for each rainfall threshold and spatial scale examined, the ROC-areas generally increase with increasing n, but with lesser gains as n approaches the full 17 members (Fig. 3). To objectively define a "point of diminishing returns", significance tests were performed comparing ROC-areas for each n to that of the full 17 member ensemble. For each n, the combination of members

with the median ROC-area was used in significance tests (dark shading in Fig. 3 distinguishes significance). Clearly, for all three rainfall thresholds (or quantiles) examined, more members are required to reach statistically indistinguishable ROC-areas relative to the full ensemble as forecast lead time increases and spatial scale decreases. For example, at every spatial scale for the 0.25-in threshold at forecast hour 6 (Fig. 3f), only three members are needed to obtain ROC-areas statistically similar to those of the full ensemble. However, by forecast hour 30 at the smallest spatial scale, nine members are needed to obtain ROC-areas not significantly different than those of the full ensemble, and fewer members are needed with increasing spatial scale (Fig. 3j).

These results can be viewed as reflecting the gain in PQPF skill as the forecast probability distribution function (PDF) of future atmosphere states is better sampled by larger *n*. Because more members are required to effectively sample a wider forecast PDF, the *n* at which skill begins to flatten increases with a wider PDF. These apparent changes in the point of diminishing returns are consistent with two aspects of our analysis associated with a widening forecast PDF: 1) increasing forecast lead time (because model/analysis errors grow) and 2) decreasing spatial scale [because errors grow faster at smaller scales (e.g., Lorenz 1969)]. Alternatively, the results could be viewed as reflecting the typical *n* required to adequately encompass observed precipitation – as lead time increases and scale decreases, the resulting error growth means that individual ensemble member solutions become less likely to verify. Therefore, more members are needed to "capture" the observations. These results are consistent with Richardson (2001) who found that for lower predictability more members were required to reach maximum possible skill (i.e., skill obtained using ∞ members), and Du et al. (1997) who also found that the majority of PQPF skill could be obtained with ~ 10 members.

4. Discussion

Although Figure 3 appears to identify the point of "diminishing returns" for *n*, additional considerations should be made in future convection-allowing ensemble design. First, for cases with below average predictability, larger *n* is required to effectively sample the forecast PDF. Second, the rainfall forecasts are under-dispersive (i.e., observations often fall outside all ensemble members) as implied by the U-shaped rank-histograms (e.g., Hamill 2001)[2] in Figure 4a, and statistical consistency analyses finding that ensemble variance is less than the mean-square-error (MSE) of the ensemble mean (Fig. 4b)[3]. The under-dispersion means that the forecast PDF is too narrow and a statistically consistent ensemble (i.e., no under-dispersion) would require more members to effectively sample the forecast PDF. The under-dispersion, which is most pronounced at early forecast lead times (6 to 18-hrs), is likely related to under-sampling of model errors and inadequate IC/LBC perturbation methods. The IC/LBC perturbations are extracted from relatively coarse (30- to 45-km Δx) SREF members and do not account for smaller scale errors on the 4-km grids (Nutter et al. 2004). At early forecast lead times when error growth is dominated by these smaller scales, the SREF perturbations may not be able to generate enough spread to accurately depict forecast uncertainty.

To illustrate impacts of under-dispersion on the apparent *n* required to effectively sample an

2 Hamill (2001) suggests that U-shaped rank histograms can be somewhat ambiguous and shows that other factors like observational errors, conditional biases, and nonrandom sampling of mis-calibrated distributions can contribute to the appearance of under-dispersion. In our case, a statistical consistency analysis also implied under-dispersion as well as daily subjective examinations comparing SSEF system precipitation forecasts to observations during the 2009 NOAA/HWT Spring Experiment. Thus, most of the U-shape in the rank histograms likely reflects under-dispersion.

3 MSE and ensemble variance are computed using Eqs. (B6) and (B7), respectively, in Eckel and Mass (2005). In a statistically consistent (i.e., reliable) ensemble MSE and variance are approximately equal when averaged over many cases. To eliminate bias influences on MSE and variance, forecast precipitation distributions are replaced with those from stage IV estimates as described in Clark et al. (2009). The bias-elimination results in forecasts with the same spatial patterns as the raw fields with no bias.

idealized PDF, two Gaussian PDFs with an arbitrarily chosen mean (μ = 10-mm) are shown (Fig. 4c). The forecast PDF with standard deviation, σ = 1.87-mm (green), corresponds to the average ensemble variance at forecast hour 6, while the forecast PDF with σ = 3.0-mm (blue) corresponds to the average MSE of the ensemble mean at forecast hour 6 (shown in Fig. 4b). Each PDF is randomly sampled using *n* from 2 to 100. For each *n*, 1000 sets of synthetic "members" are drawn and probabilities for rainfall exceeding 12.7-mm (marked by vertical line in Fig. 4c) are computed for each set using the method to compute PQPFs described in Section 2b. Then, using the actual probabilities from the PDFs, the average probability error (i.e., sampling error) for each *n* is computed (Fig. 4d). In this idealization, if tolerable error is considered $\leq$ 0.05 (marked by horizontal line in Fig. 4d), then average errors associated with probabilities derived from the PDF with σ = 1.87-mm would fall below the tolerable error using a minimum of 20 members. However, if the ensemble was statistically consistent (i.e., σ = MSE = 3.0-mm), about 60 members would be required to fall below the tolerable error.

A third consideration is the systematic wet bias implied by the right-skewness of the rank histograms (Fig. 4a), and explicitly shown by time series of average bias for each ensemble member in Figure 5. As discussed by Eckel and Mass (2005), systematic biases introduce "bogus" uncertainty because systematic errors are not uncertain. Thus, calibration – which was not attempted given the limited number of "training cases" available – typically reduces spread. Clark et al. (2009) also discusses the impact of wet biases on spread-error metrics.

Fourth, the change in ROC-area as a function of *n* does not necessarily reflect changes in potential value (e.g., Richardson 2000) to end-users. Richardson (2001) illustrated that for rare events, users with low cost-lost ratios (e.g., Murphy 1977) benefit significantly in terms of potential value by increasing *n* from 50 to 100, despite little change in Brier Skill Score (Wilks 1995).

In summary, average ROC-areas for PQPFs from the 2009 SSEF system were quite skillful, especially considering the relatively low predictability typical of May-June (Fritsch and Carbone 2004). For the full ensemble, ROC-areas range between 0.88 and 0.95 to forecast hour 30 for even the finest spatial scales. Additionally, relatively small *n* (3 to 9 members) had statistically indistinguishable average ROC-areas relative to the full 17 member ensemble, and the *n* at which skill began to level increased (decreased) with increasing forecast lead time (spatial scale). However, larger spread that better matched the MSE would have resulted in larger *n* to reach the point of "diminishing returns". Additionally, low predictability regimes and/or rare events would require more members to reach "diminishing returns". Nevertheless, clearly spatial scale and forecast lead time desired require careful consideration for future convection-allowing ensemble design. Future work should address improving statistical consistency of convection-allowing ensembles, and further evaluations are needed for weather regimes with varying degrees of predictability and/or rare events.

Acknowledgments

A National Research Council Post-doctoral Award supported AJC. SSEF forecasts were primarily supported by the NOAA CSTAR program, and were produced at the Pittsburgh Supercomputing Center, and the National Institute of Computational Science at the University of Tennessee. Supplementary support was provided by NSF-ITR project LEAD (ATM-0331594), NSF ATM-0802888, and other NSF grants to CAPS. Comments from two anonymous reviewers improved the manuscript.

References

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting*, **18**, 918–932.

Baldwin, M. E., and K. E. Mitchell, 1997: The NCEP hourly multisensor U. S. precipitation analysis for

operations and GCIP research. Preprints, *13th Conf. On Hydrology*, Long Beach, CA, Amer. Meteor. Soc., 54-55.

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.

Chen, F., and J. Dudhia, 2001: Coupling an advanced land-surface/ hydrology model with the Penn State/ NCAR MM5 modeling system. Part I: Model description and implementation. *Mon. Wea. Rev.*, **129**, 569–585.

Chen, S.-H., and W.-Y. Sun, 2002: A one-dimensional time dependent cloud model. *J. Meteor. Soc. Japan*, **80**, 99-118.

Chou M.-D., and M. J. Suarez, 1994: An efficient thermal infrared radiation parameterization for use in general circulation models. NASA Tech. Memo. 104606, 3, 85pp.

Clark, A. J., W. A. Gallus, M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Wea. Forecasting*, **24**, 1121–1140.

Done, J., C. A. Davis, and M. L. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the Weather Research and Forecast (WRF) Model. *Atmos. Sci. Lett*., **5**, 110-117.

Du, J., J. McQueen, G. DiMego, Z. Toth, D. Jovic, B. Zhou, and H. Chuang, 2006: New dimension of NCEP Short-Range Ensemble Forecasting (SREF) system: Inclusion of WRF members, Preprint,WMO Expert Team Meeting on Ensemble Prediction System, Exeter, UK, Feb. 6-10, 2006, 5 pp.

Ebert, E. E., 2009: Neighborhood verification: A strategy for rewarding close forecasts. *Wea. Forecasting*, **24**, 1498-1510.

Eckel, F. A., and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350.

Ek, M. B., K. E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, and J. D. Tarpley, 2003: Implementation of Noah Land Surface Model advances in the National Centers for Environmental Prediction operational mesoscale Eta Model, *J. Geophys. Res*., 108(D22), 8851, doi:10.1029/2002JD003296.

Fels, S. B. and M. D. Schwarzkopf, 1975: The Simplified Exchange Approximation: A New Method for Radiative Transfer Calculations, *J. Atmos. Sci.*, 32, 1475–1488.

Ferrier, B. S., Y. Jin, Y. Lin, T. Black, E. Rogers, and G. DiMego, 2002: Implementation of a new grid-scale cloud and rainfall scheme in the NCEP Eta Model. Preprints, *15th Conf. On Numerical Weather Prediction*, San Antonio, TX, Amer. Meteor. Soc., 280-283.

Fritsch, J. M., and R.E. Carbone, 2004: Improving quantitative precipitation forecasts in the warm season: A USWRP research and development strategy. *Bull. Amer. Meteor. Soc.*, 85, 955–965.

Gao, J., M. Xue, K. Brewster, and K. K. Droegemeier, 2004: A three-dimensional variational data analysis method with recursive filter for Doppler radars. *J. Atmos. Ocean. Tech*., **21**, 457-469.

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

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

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

——, 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560.

Hong, S.-Y, and J.-O. J. Lim, 2006: The WRF single-moment 6-class microphysics scheme (WSM6). *J. Korean Meteor. Soc.*, **42**, 129-151.

Hu, M., M. Xue, and K. Brewster, 2006: 3DVAR and cloud analysis with WSR-88D level-II data for the

prediction of Fort Worth tornadic thunderstorms. Part I: Cloud analysis and its impact. *Mon. Wea. Rev*., **134**, 675-698.

Janjic, Z. I., 2003: A nonhydrostatic model based on a new approach. *Meteorol. Atmos. Phys*., 82, 271-285.

Jenkner, J., C. Frei, and C. Schwierz, 2008: Quantile-based short-range QPF evaluation over Switzerland. *Meteorol. Z*., **17**, 827-848.

Kain, J. S., P. R. Janish, S. J. Weiss, M. E. Baldwin, R. S. Schneider, and H. E. Brooks, 2003: Collaboration between forecasters and research scientists at the NSSL and SPC: The Spring Program. *Bull. Amer. Meteor. Soc.*, **84**, 1797–1806.

——, S. J. Weiss, M. E. Baldwin, G. W. Carbin, D. A. Bright, J. J. Levit, and J. A. Hart, 2005: Evaluating high-resolution configurations of the WRF model that are used to forecast severe convective weather: The 2005 SPC/NSSL Spring Program. Preprints, *21st Conf. Weather Analysis and Forecasting and 17th Conf. Numerical Weather Prediction*, Washington, DC, Amer. Meteor. Soc., 2A.5. [Available online at http://ams.confex.com/ams/pdfpapers/94843.pdf.]

——, S. J. Weiss, J. J. Levit, M. E. Baldwin, and D. R. Bright, 2006: Examination of convection-allowing configurations of the WRF model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. *Wea. Forecasting*, **21**, 167–181.

Kong, F., M. Xue, K. K. Droegemeier, D. Bright, M. C. Coniglio, K. W. Thomas, Y. Wang, D. Weber, J. S. Kain, S. J. Weiss, and J. Du, 2007: Preliminary analysis on the real-time storm-scale ensemble forecasts produced as a part of the NOAA Hazardous Weather Testbed 2007 Spring Experiment. *Preprints, 22nd Conf. Weather Analysis and Forecasting/18th Conf. Numerical Weather Prediction*, Park City, UT, Amer. Meteor. Soc., 3B.2.

——, M. Xue, K. W. Thomas, Y. Wang, K. Brewster, J. Gao, K. K. Droegemeier, J. S. Kain, S. J. Weiss, D. Bright, M. Coniglio, and J. Du, 2009: A real-time storm-scale ensemble forecast system: 2009 Spring Experiment. *Preprints, 23rd Conf. Weather Analysis and Forecasting/19th Conf. Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., CD-ROM, 16A.3.

Lacis, A. A., and J. E. Hansen, 1974: A parameterization for the absorption of solar radiation in the earth's atmosphere. *J. Atmos. Sci.*, 31, 118–133.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289-307.

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

Mellor, G. L., and T. Yamada, 1982: Development of a turbulence closure model for geophysical fluid problems. *Rev. Geophys.,* **20,** 851–875.

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmosphere: RRTM, a validated correlated-k model for the long-wave. *J. Geophys. Res*., 102(D14), 16663-16682.

Mylne, K. R., 1999: The use of forecast value calculations for optimal decision making using probability forecasts. Preprints, *17th Conf. on Weather Analysis and Forecasting*, Denver, CO, Amer. Meteor. Soc., 235-239.

Noh, Y., W. G. Cheon, S.-Y. Hong, and S. Raasch, 2003: Improvement of the K-profile model for the planetary boundary layer based on large eddy simulation data. *Boundary Layer Meteor*., **107**, 401-427.

Nutter, P., D. Stensrud, M. Xue, 2004: Effects of Coarsely Resolved and Temporally Interpolated Lateral Boundary Conditions on the Dispersion of Limited-Area Ensemble Forecasts. *Mon. Wea. Rev.*, **132**, 2358-2377.

Richardson, D. S., 2000: Applications of cost-loss models. Proc. *Seventh ECMWF Workshop on Meteorological Operational Systems*, Reading, United Kingdom, ECMWF, 209-213.

——, 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the

effect of ensemble size. *Quart. J. Roy. Meteor. Soc.,* **127**, 2473-2489.

Roberts, N. M. and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev*., **136**, 78-97.

Schwarzkopf, M. D., and S. B. Fels, 1991: The simplified exchange method revisited — An accurate, rapid method for computation of infrared cooling rates and fluxes. *J. Geophys. Res.*, 96 (D5), 9075–9096.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, M. G. Duda, X.-Y., Huang, W. Wang, and J. G. Powers, 2008: A description of the Advanced Research WRF Version 2, NCAR Tech Note, NCAR/TN-475+STR, 113 pp. [Available at: http://www.mmm.ucar.edu/wrf/users/docs/arw_v3.pdf.]

Smirnova, T. G., J. M. Brown, and S. G. Benjamin, 1997: Performance of different soil model configurations in simulating ground surface temperature and surface fluxes. *Mon. Wea. Rev.*, **125**, 1870–1884.

——, ——, ——, and D. Kim, 2000: Parameterization of coldseason processes in the MAPS land-surface scheme. *J. Geophys. Res.*, **105** (D3), 4077–4086.

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.

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

Weisman, M. L., W. C. Skamarock, J. B. Klemp, 1997: The resolution dependence of explicitly modeled convective systems. *Mon. Wea. Rev*., **125**, 527–548.

——, C. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with 0–36-h explicit convective forecasts with the WRF-ARW model. *Wea. Forecasting*, **23**, 407–437.

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

Xue, M., K. K. Droegemeier, V. Wong, A. Shapiro, K. Brewster, F. Carr, S. Weber, Y. Liu, and D.-H. Wang, 2001: The Advanced Regional Prediction System (ARPS) – A multiscale nonhydrostatic atmospheric simulation and prediction tool. Part II: Model physics and applications. *Meteor. Atmos. Physics*. **76**, 143-165.

——, D. Wang, J. Gao, K. Brewster, and K. K. Droegemeier, 2003: The advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation. *Meteor. Atmos. Phys*., **82**, 139-170.

——, F. Kong, D. Weber, K. W. Thomas, Y. Wang, K. Brewster, K. K. Droegemeier, J. S. Kain, S. J. Weiss, D. Bright, M. Wandishin, M. Coniglio, and J. Du, 2007: CAPS realtime storm-scale ensemble and high-resolution forecasts as part of the NOAA Hazardous Weather Testbed 2007 Spring Experiment. *Preprints, 22nd Conference On Weather Analysis and Forecasting/18th Conference on Numerical Weather Prediction*, Park City, UT, Amer. Meteor. Soc., 3B.1.

——, F. Kong, K. W. Thomas, J. Gao, Y. Wang, K. Brewster, K. K. Droegemeier, X. Wang, J. S. Kain, S. J. Weiss, D. R. Bright, M. C. Coniglio, and J. Du, 2009: CAPS realtime multi-model convection-allowing ensemble and 1-km convection-resolving forecasts for the NOAA Hazardous Weather Testbed 2009 Spring Experiment. Preprints, *23rd Conference on Weather Analysis and Forecasting/19th Conference on Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., 16A.2.

List of Tables

Table 1 Model configuration for SSEF system members. All WRF members used version 3.0.1.1 NAMa and NAMf indicate NAM forecasts and analyses, respectively; ARPSa refers to ARPS 3DVAR and cloud analysis; em_pert, nmm_pert, etaKF_pert, and etaBMJ_pert are perturbations from different SREF members; and em-n1, em-p1, nmm-n1, nmm-p1, etaKF-n1, etaKF-p1, etaBMJ-n1, and etaBMJ-p1 are different SREF members that are used for LBCs. Note, the SREF members obtain their LBCs from NCEP's Global Ensemble Forecast System. Boundary layer schemes include Mellor-Yamada-Janjic (MYJ; Mellor and Yamada 1982; Janjic 2002), YonSei University (YSU; Noh et al. 2003), and a 1.5-order closure scheme developed for ARPS (Xue et al. 2001). Microphysics schemes include Thompson et al. (2004), WRF single-moment 6-class (WSM-6; Hong and Lim 2006), Ferrier et al. (2002), and Purdue Lin (Chen and Sun 2002). Radiation schemes include the RRTM short-wave (Mlawer et al. 1997), Goddard long-wave (Chou and Suarez 1994), and GFDL short-wave (Lacis and Hansen 1974) and long-wave (Fels and Schwarzkopf 1975; Schwarzkopf and Fels 1991). Land-surface models include the NOAH (Chen and Dudhia 2001), RUC (Smirnova et al. 1997, 2000), and Force-restore (Xue et al. 2001).

List of Figures

Figure 1 The outer domain (light shading) used for the SSEF system ensemble members and the inner sub-domain (dark shading) used for the analyses conducted in this study. The dates for the 25 cases analyzed are listed below the domain.

Figure 2 (a) Stage IV precipitation estimates interpolated to the raw model grid for the 6-hr period ending 0000 UTC 16 May 2009 with the black contours marking the 0.75-in rainfall threshold; and (b) corresponding SSEF system 24-hr forecast probabilities of rainfall greater than 0.75-in (shaded) with areas of stage IV rainfall greater than 0.75-in hatched. (c) - (d), (e) - (f), (g) - (h), and (i) – (j), same as (a) – (b), except stage IV estimates and the forecasts used to generate probabilities are smoothed over grid-points within radii of 10-, 60-, 100-, and 200-km, respectively, of each grid-point. (k) The quantiles for a range of rainfall thresholds in the unsmoothed stage IV rainfall distribution are marked (e.g., p = 0.980 for 0.50-in.), and each lines shows how the values corresponding to each of these quantiles changes for grids smoothed over radii from 10- to 200-km.

Figure 3 ROC-areas with increasing ensemble size at different spatial scales for 6-hr accumulated precipitation at the 0.10-in rainfall threshold for forecast hours (a) 6, (b) 12, (c) 18, (d) 24, and (e) 30. (f) – (j), and (k) – (o) same as (a) – (e), except for the 0.25- and 0.50-in rainfall thresholds, respectively. The range of values encompassed by each color corresponds to the range of ROC-areas for each ensemble size within the "whiskers" of a standard box-plot (i.e., the most extreme values within 1.5 times the inter-quartile range). The dark shaded areas denote ensemble sizes for which the ROC-areas are significantly less ($\alpha = 0.05$) than that of the full 17 member ensemble. The legend in panel (a) shows the spatial scales that correspond to each color of shading.

Figure 4 (a) Rank histograms from the SSEF system for 6-hr accumulated precipitation ending at forecast hours 6-30. (b) Average mean-square-error (MSE) of ensemble mean 6-hr precipitation (blue), and corresponding ensemble variance (green) from the SSEF system at forecast hours 6-30. (c) Idealized, normally distributed forecast probability distribution functions (PDFs) with μ = 10-mm and σ = 1.87-mm [green; corresponding to the variance at forecast hour 6 in (b)] and σ =

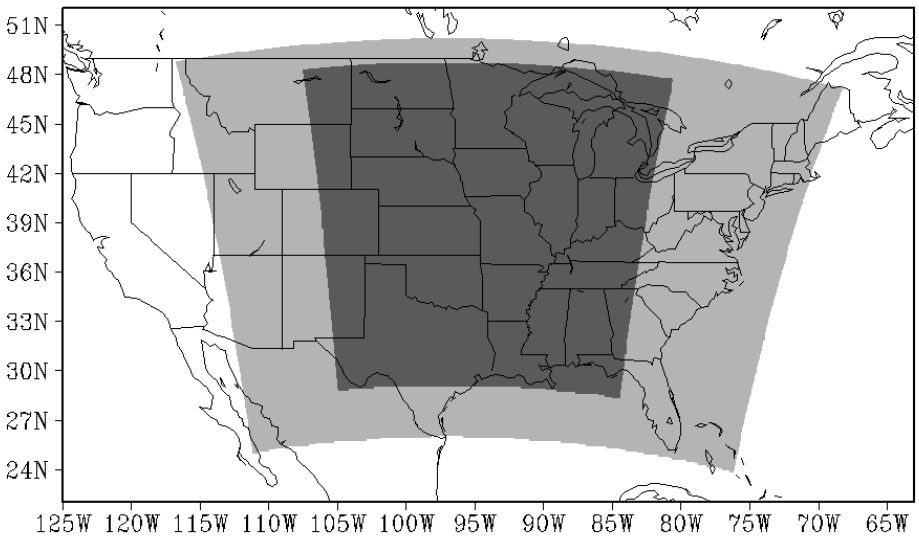
3.0-mm [blue; corresponding to the MSE of the ensemble mean at forecast hour 6 in (b)]. The black vertical line marks the 12.7-mm (0.5-in) rainfall threshold. (d) Average error in forecast probabilities for rainfall greater than 12.7-mm derived from randomly sampling the PDFs in (c) using an increasing number of samples.

Figure 5 Average forecast precipitation bias for 6-hourly accumulation intervals from each SSEF system member [legend in panel (a)] for the thresholds (a) 0.10- , (b) 0.25-, (c) 0.50-, (d) 1.00-, and (e) 2.00-in. The dashed horizontal line in each panel marks a bias of 1.0.

Table 1 Model configuration for SSEF system members. All WRF members used version 3.0.1.1 NAMa and NAMf indicate NAM forecasts and analyses, respectively; ARPSa refers to ARPS 3DVAR and cloud analysis; em_pert, nmm_pert, etaKF_pert, and etaBMJ_pert are perturbations from different SREF members; and em-n1, em-p1, nmm-n1, nmm-p1, etaKF-n1, etaKF-p1, etaBMJ-n1, and etaBMJ-p1 are different SREF members that are used for LBCs. Note, the SREF members obtain their LBCs from NCEP's Global Ensemble Forecast System. Boundary layer schemes include Mellor-Yamada-Janjic (MYJ; Mellor and Yamada 1982; Janjic 2002), YonSei University (YSU; Noh et al. 2003), and a 1.5-order closure scheme developed for ARPS (Xue et al. 2001). Microphysics schemes include Thompson et al. (2004), WRF single-moment 6-class (WSM-6; Hong and Lim 2006), Ferrier et al. (2002), and Purdue Lin (Chen and Sun 2002). Radiation schemes include the RRTM short-wave (Mlawer et al. 1997), Goddard long-wave (Chou and Suarez 1994), and GFDL short-wave (Lacis and Hansen 1974) and long-wave (Fels and Schwarzkopf 1975; Schwarzkopf and Fels 1991). Land-surface models include the NOAH (Chen and Dudhia 2001), RUC (Smirnova et al. 1997, 2000), and Force-restore (Xue et al. 2001).

Ensemble Member	**ICs**	**LBCs**	**Radar data**	**Micro-physics scheme**	**Short-wave Radiation scheme**	**Long-wave Radiation scheme**	**Land-surface model**	**Boundary Layer scheme**
arw_cn	00Z ARPSa	00Z NAMf	yes	Thompson	Goddard	RRTM	Noah	MYJ
arw_c0	00Z NAMa	00Z NAMf	no	Thompson	Goddard	RRTM	Noah	MYJ
arw_n1	arw_cn – em_pert	21Z SREF em-n1	yes	Ferrier	Goddard	RRTM	Noah	YSU
arw_p1	arw_cn + em_pert	21Z SREF em-p1	yes	WSM6	Dudhia	RRTM	Noah	MYJ
arw_n2	arw_cn – nmm_pert	21Z SREF nmm-n1	yes	Thompson	Dudhia	RRTM	RUC	MYJ
arw_p2	arw_cn + nmm_pert	21Z SREF nmm-p1	yes	WSM6	Dudhia	RRTM	Noah	YSU
arw_n3	arw_cn – etaKF_pert	21Z SREF etaKF-n1	yes	Thompson	Dudhia	RRTM	Noah	YSU
arw_p3	arw_cn + etaKF_pert	21Z SREF etaKF-p1	yes	Ferrier	Dudhia	RRTM	Noah	MYJ
arw_n4	arw_cn – etaBMJ_pert	21Z SREF etaBMJ-n1	yes	WSM6	Goddard	RRTM	Noah	MYJ
arw_p4	arw_cn + etaBMJ_pert	21Z SREF etaBMJ-p1	yes	Thompson	Goddard	RRTM	RUC	YSU
nmm_cn	00Z ARPSa	00Z NAMf	yes	Ferrier	GFDL	GFDL	Noah	MYJ
nmm_c0	00Z NAMa	00Z NAMf	no	Ferrier	GFDL	GFDL	Noah	MYJ
nmm_p1	nmm_cn + em_pert	21Z SREF em-p1	yes	WSM6	GFDL	GFD:	RUC	MYJ
nmm_n2	nmm_cn – nmm_pert	21Z SREF nmm-n1	yes	Ferrier	Dudhia	RRTM	Noah	YSU
nmm_p2	nmm_cn + nmm_pert	21Z SREF nmm-p1	yes	Thompson	GFDL	GFDL	RUC	YSU

nmm_n3	nmm_cn – etaKF_pert	21Z SREF etaKF-n1	yes	WSM6	Dudhia	RRTM	Noah	YSU
nmm_n4	nmm_cn – etaBMJ_pert	21Z SREF etaBMJ-n1	yes	WSM6	Dudhia	RRTM	RUC	MYJ
nmm_p4	nmm_cn + etaBMJ_pert	21Z SREF etaBMJ-p1	yes	Ferrier	Dudhia	RRTM	RUC	YSU
arps_cn	00Z ARPSa	00Z NAMf	yes	Lin	Goddard	Goddard	Force-restore	TKE
arps_c0	00Z NAMa	00Z NAMf	no	Lin	Goddard	Goddard	Force-restore	TKE



SE2009 FORECAST DATES (25 TOTAL)

APR: 27

MAY: 1–2, 4–8, 11–12, 14–15, 18, 20–22, 25–29

JUN: 1–2, 4–5

Figure 1 The outer domain (light shading) used for the SSEF system ensemble members and the inner sub-domain (dark shading) used for the analyses conducted in this study. The dates for the 25 cases analyzed are listed below the domain.

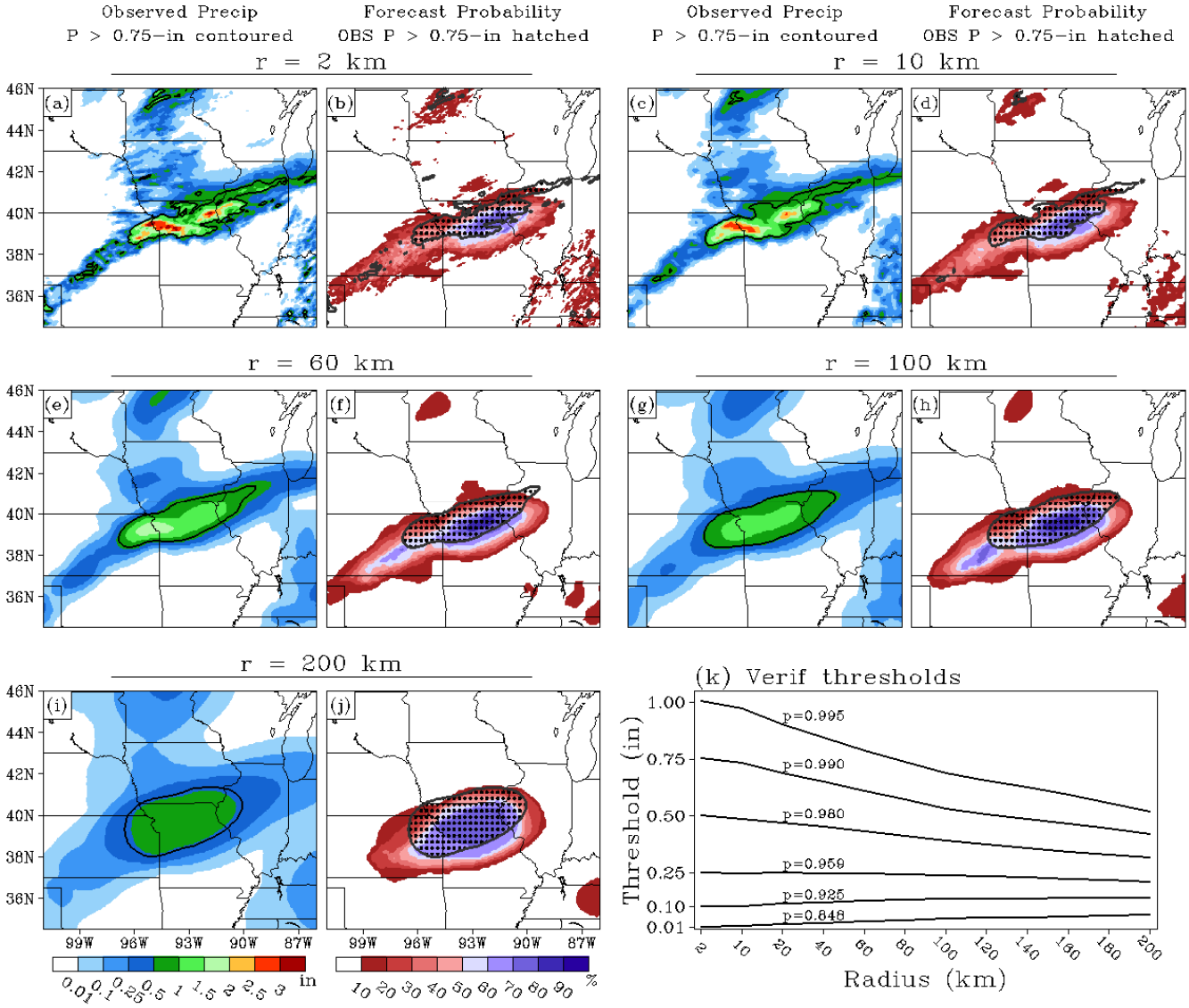


Figure 2 (a) Stage IV precipitation estimates interpolated to the raw model grid for the 6-hr period ending 0000 UTC 16 May 2009 with the black contours marking the 0.75-in rainfall threshold; and (b) corresponding SSEF system 24-hr forecast probabilities of 6-hr rainfall greater than 0.75-in (shaded) with areas of stage IV rainfall greater than 0.75-in hatched. (c) - (d), (e) - (f), (g) - (h), and (i) – (j), same as (a) – (b), except stage IV estimates and the forecasts used to generate probabilities are smoothed over grid-points within radii of 10-, 60-, 100-, and 200-km, respectively, of each grid-point. (k) The quantiles for a range of rainfall thresholds in the unsmoothed stage IV rainfall distribution are marked (e.g., p = 0.980 for 0.50-in.), and each line shows how the values corresponding to each of these quantiles changes for grids smoothed over radii from 10- to 200-km.

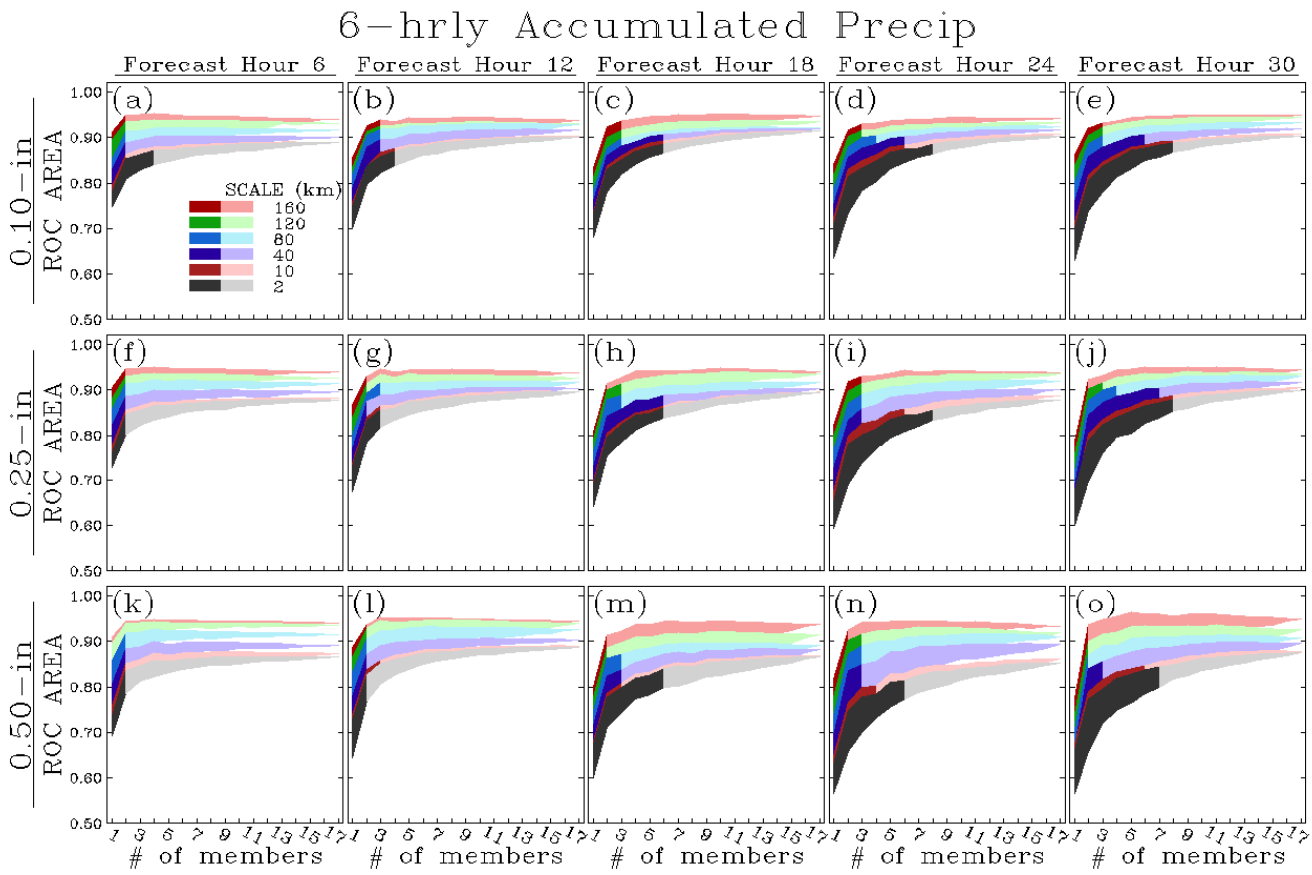


Figure 3 ROC-areas with increasing ensemble size at different spatial scales for 6-hr accumulated precipitation at the 0.10-in rainfall threshold for forecast hours (a) 6, (b) 12, (c) 18, (d) 24, and (e) 30. (f) – (j), and (k) – (o) same as (a) – (e), except for the 0.25- and 0.50-in rainfall thresholds, respectively. The range of values encompassed by each color corresponds to the range of ROC-areas for each ensemble size within the "whiskers" of a standard box-plot (i.e., the most extreme values within 1.5 times the inter-quartile range). The dark shaded areas denote ensemble sizes for which the ROC-areas are significantly less ($\alpha = 0.05$) than that of the full 17 member ensemble. The legend in panel (a) shows the spatial scales that correspond to each color of shading.

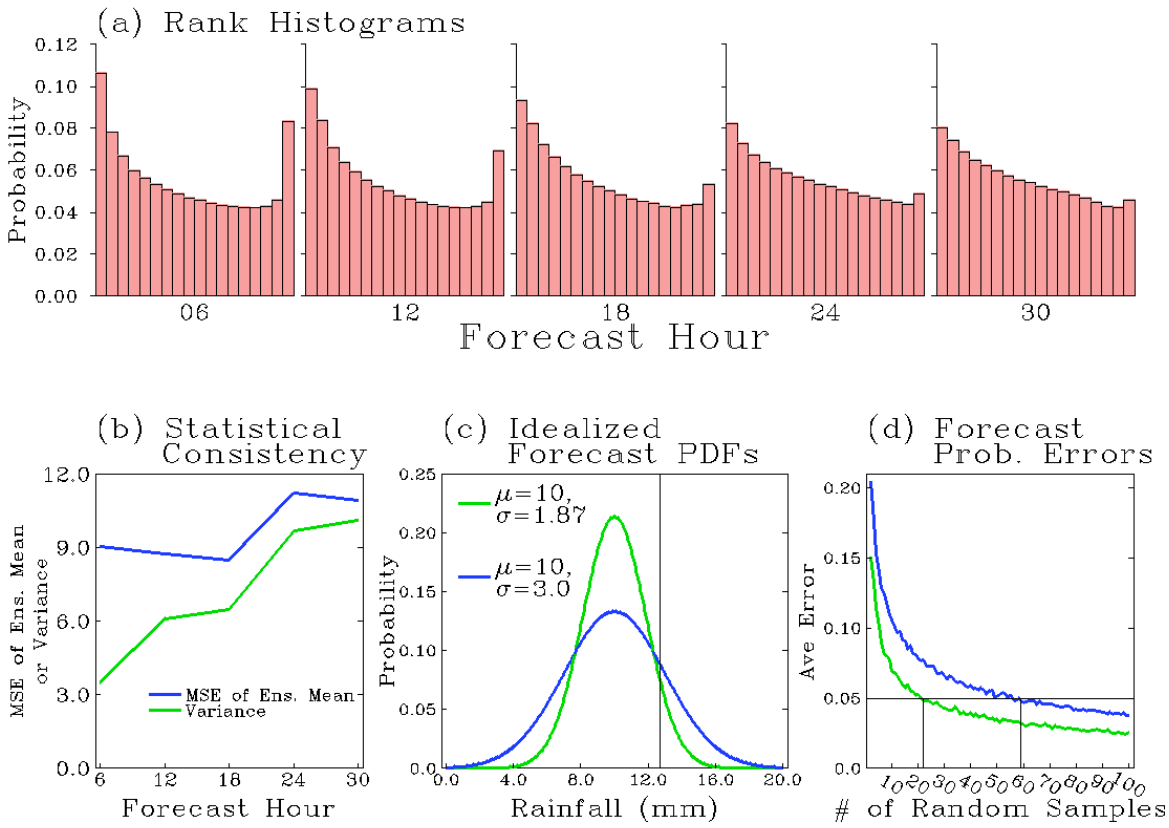


Figure 4 (a) Rank histograms from the SSEF system for 6-hr accumulated precipitation ending at forecast hours 6-30. (b) Average mean-square-error (MSE) of ensemble mean 6-hr precipitation (blue), and corresponding ensemble variance (green) from the SSEF system at forecast hours 6-30. (c) Idealized, normally distributed forecast probability distribution functions (PDFs) with μ = 10-mm and σ = 1.87-mm [green; corresponding to the variance at forecast hour 6 in (b)] and σ = 3.0-mm [blue; corresponding to the MSE of the ensemble mean at forecast hour 6 in (b)]. The black vertical line marks the 12.7-mm (0.5-in) rainfall threshold. (d) Average error in forecast probabilities for rainfall greater than 12.7-mm derived from randomly sampling the PDFs in (c) using an increasing number of samples.

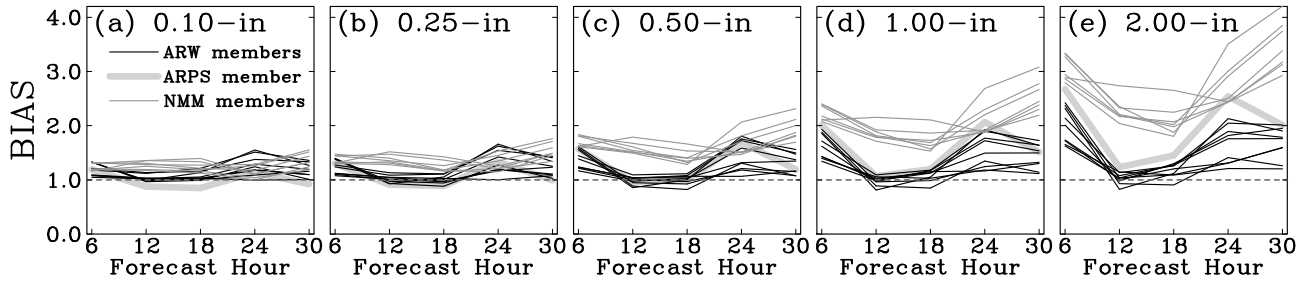


Figure 5 Average forecast precipitation bias for 6-hourly accumulation intervals from each SSEF system member [legend in panel (a)] for the thresholds (a) 0.10- , (b) 0.25-, (c) 0.50-, (d) 1.00-, and (e) 2.00-in. The dashed horizontal line in each panel marks a bias of 1.0.