12B.4 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[3], Michael C. Coniglio[1], Kevin W. Thomas[3], Yunheng Wang[3], Keith Brewster[3], Jidong Gao[3], Steven J. Weiss[4], and Jun Du[5]

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

1. INTRODUCTION

During the period 2003-2005 the NOAA/Hazardous Weather Testbed Spring Experiments (e.g., Kain et al. 2003) tested deterministic convection-allowing configurations of the Weather Research and Forecasting (WRF) model that used the Advanced Research WRF (ARW) dynamic core (Skamarock et al. 2005) in an operational forecasting environment. Promising results (e.g., Done et al. 2004; Kain et al. 2005; Weisman et al. 2008) from these years and a desire to quantify forecast uncertainty at convection-allowing scales motivated the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma to begin producing Storm Scale Ensemble Forecast (SSEF) systems (i.e. convection-allowing ensembles) for subsequent Spring Experiments beginning in 2007. During 2007-2008, 10 WRF-ARW members composed the SSEF systems (see Kong et al. 2007, 2008 for model configurations) and in 2009 the ensemble size and diversification was increased by including eight Nonhydrostatic Mesoscale Model (NMM; Janjic 2003) members and two Advanced Regional Prediction System (ARPS; Xue et al. 2000) members.

Given the relatively large size of the 2009 SSEF system and considerable computational expense (e.g., Xue et al. 2009), the purpose of this study is to examine the skill of probabilistic quantitative precipitation forecasts (PQPFs) at different spatial scales as a function of the ensemble size in the 2009 SSEF system. Specifically, we explore whether the gain in PQPF skill from each additional ensemble member becomes small, approaching a point of "diminishing returns". Clearly, such an analysis is important to consider in designing an ensemble system that efficiently utilizes available computing resources.

2. DATA AND METHODOLOGY

2.1 Ensemble Configuration

The 2009 SSEF system members used 4-km grid-spacing, were initialized at 0000 UTC, and integrated 30 hours over an approximately 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

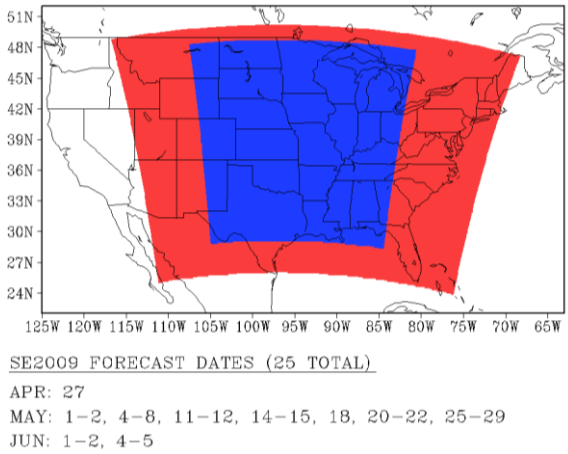

Figure 1 The outer domain (red shading) used for the SSEF system ensemble members and the inner sub-domain (blue shading) used for the analyses conducted in this study. The dates for the 25 cases analyzed are listed below the domain.

lateral boundary condition (LBC) effects (e.g., Warner et al. 1997) and to concentrate on regions climatologically favored for strong, organized springtime convection. Results are aggregated over 25 cases between April 27 and June 5 (Fig. 1). Ensemble member configurations are provided in Table 1 (see also Xue et al. 2009). Radial velocity and reflectivity data from WSR-88D radars and the 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 (12-km grid-spacing) from the 0000 UTC operational 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). However, only the 17 members assimilating radar data are used herein. To account for analysis uncertainty, 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 member forecasts were used for LBCs. To account for model physics

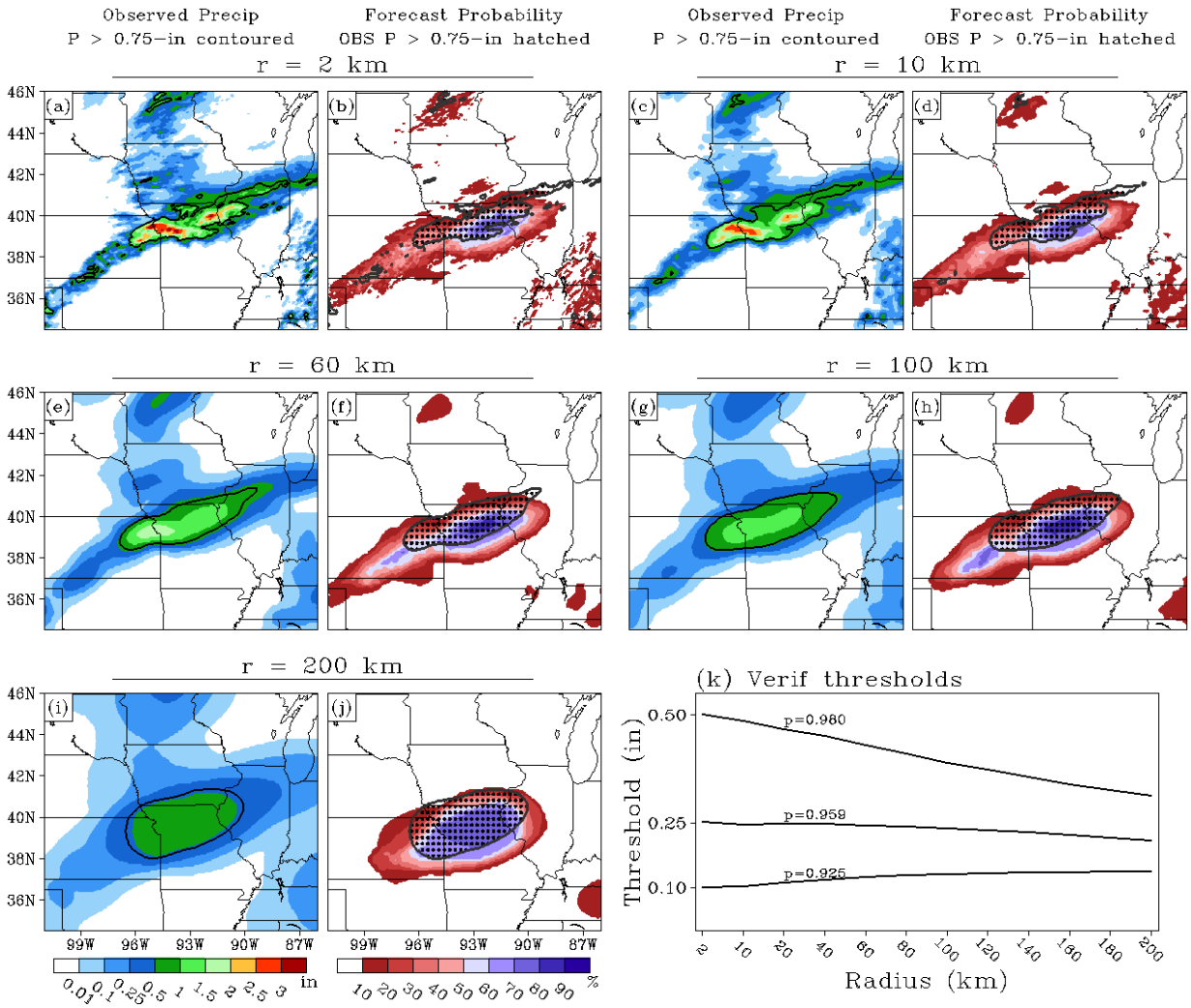

Figure 2 (a) Stage IV precipitation estimates interpolated to the raw model grid for the 6-hr period ending 0000 UTC 16 May 2009 with the black contours marking the 0.75-in rainfall threshold; and (b) corresponding SSEF system 24-hr forecast probabilities of 6-hr rainfall greater than 0.75-in (shaded) 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.

uncertainty, different boundary layer, microphysics, radiation, and land surface schemes were used (see Table 1 for schemes and references).

2.2 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) that conserves the total volume of liquid over the domain. Statistical significance is determined using Hamill's (1999) resampling methodology.

PQPFs for 6-h accumulated rainfall computed using 1 to 17 ensemble members for different spatial scales are examined. PQPFs are computed by finding the location of the verification thresholds within the distribution of ensemble member forecasts (e.g., Hamill and Colucci 1997, 1998). This method for computing probabilities is preferred because it results in a continuous range of probabilities rather than a set of discrete values that are obtained by simply tallying the fraction of members that forecast an "event". 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

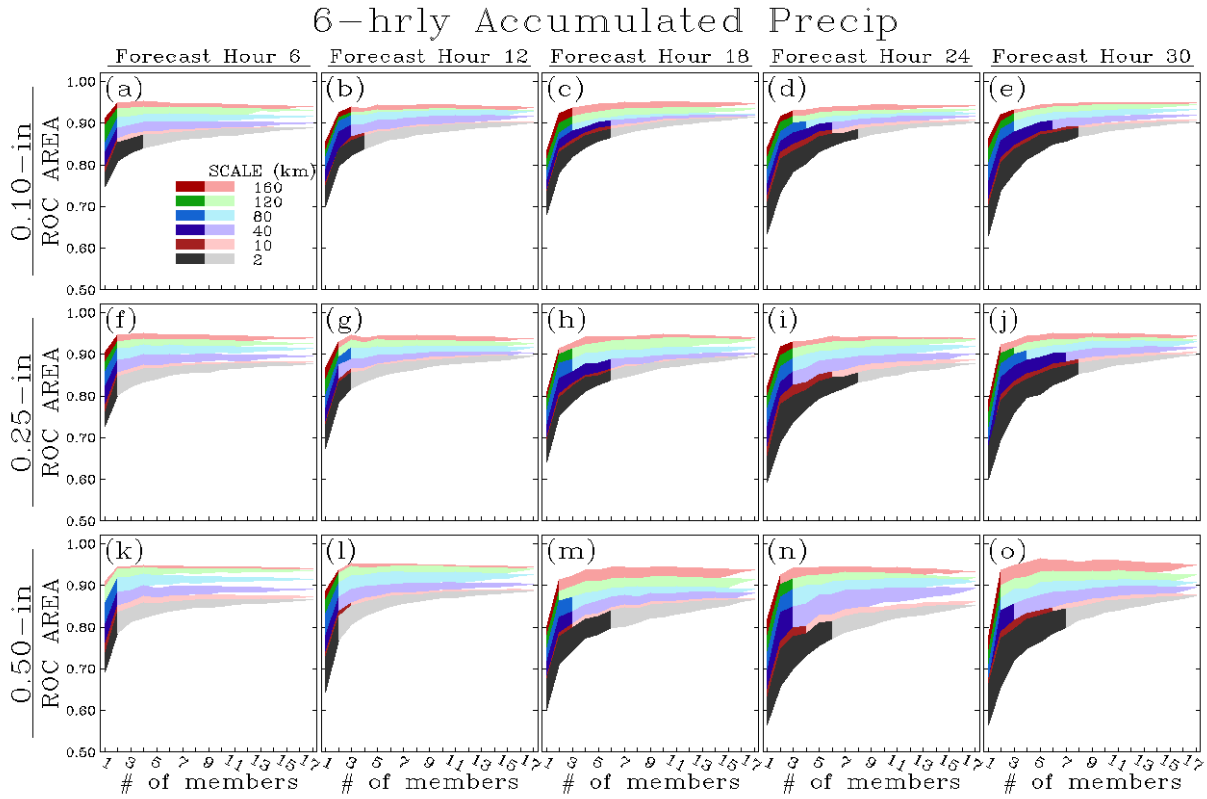

Figure 3 ROC areas with increasing ensemble size at different spatial scales for 6-hr accumulated precipitation at the 0.10-in rainfall threshold for forecast hours (a) 6, (b) 12, (c) 18, (d) 24, and (e) 30. (f) – (j), and (k) – (o) same as (a) – (e), except for the 0.25- and 0.50-in rainfall thresholds, 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 (α = 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.

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. In other words, exceedance forecasts of constant rainfall quantiles, rather than amounts, are evaluated (e.g., Jenkner et al. 2008). For example, in the distribution of stage IV rainfall estimates on the non-upscaled grid for all of the 6-hr accumulation periods considered in this study, the verification threshold 0.50-in represents the p=0.98 quantile of the distribution. So, for comparison to one of the upscaled grids, the value corresponding to the p=0.98 quantile of the upscaled grid is used. Figure 2k illustrates how using constant quantiles 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 ability to distinguish between events and non-events and is closely related to the economic value of a forecast system (e.g., Mylne 1999; Richardson 2001). The ROC area 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 $P < 0.05$, $0.05 \le P < 0.15$, $0.15 \le P < 0.25$ … $0.85 \le P < 0.95$, and $P \ge 0.95$. 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).

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, … 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.

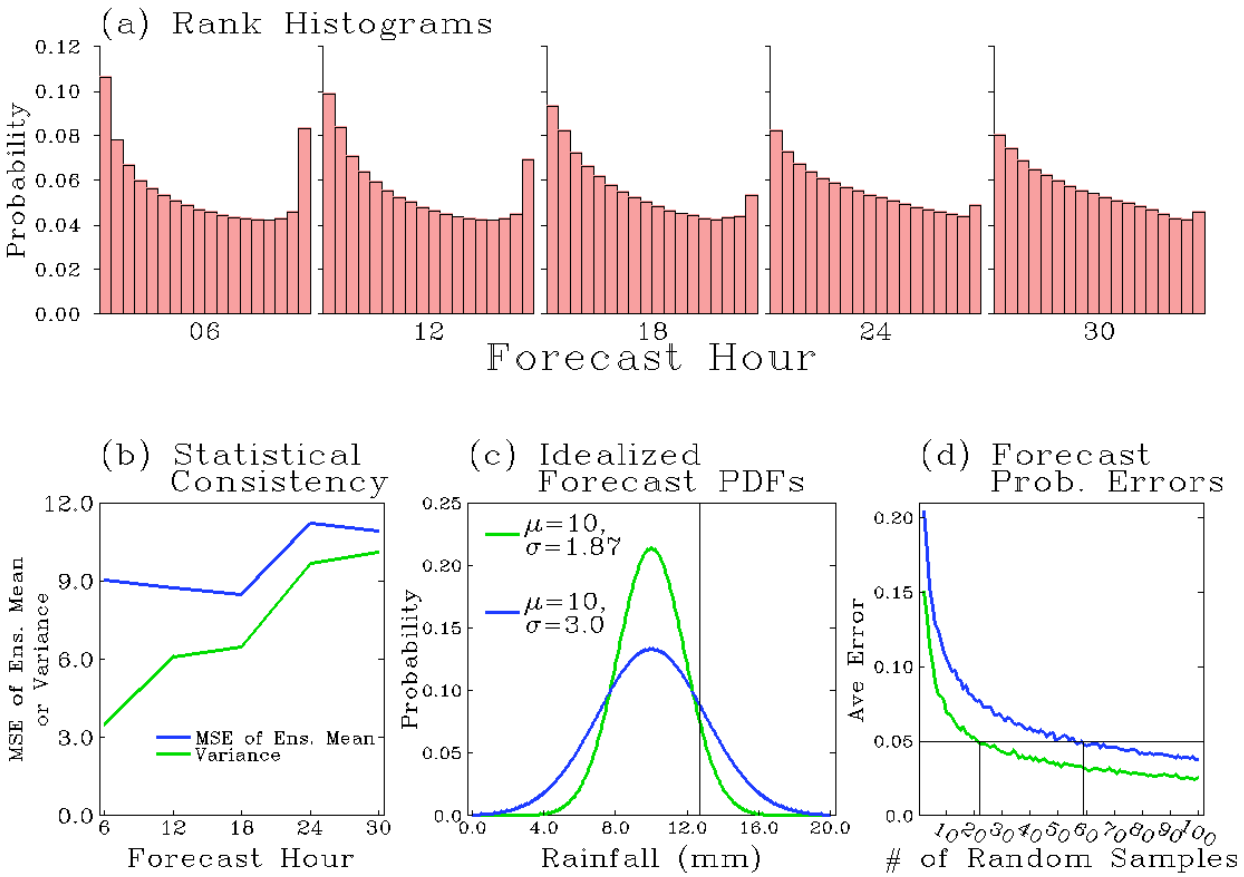

Figure 4 Rank histograms from the SSEF system for 6-hr accumulated precipitation ending at forecast hours 6-30. (b) Average mean-square-error (MSE) of ensemble mean 6-hr precipitation (blue), and corresponding ensemble variance (green) from the SSEF system at forecast hours 6-30. (c) Idealized, normally 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.

3. RESULTS

Figure 3 illustrates how the ROC areas change with increasing *n*. Not surprisingly, for each rainfall threshold and spatial scale examined, the ROC areas generally increase with increasing *n*, but with lesser gains as *n* approaches the size of the full ensemble. 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 the significance test (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 that are not significantly different than 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 number of members required to adequately encompass the 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 accounted for in future convection-allowing ensemble design. First, for cases with lower than 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 the range of ensemble member solutions) which is implied by the U-shape in rank-histograms (e.g., Hamill 2001) at each forecast hour (Fig. 4a), as well as a statistical consistency analysis showing that the ensemble variance is less than the mean-square-error (MSE) of the ensemble mean (Fig. 4b). The under-dispersion means that the forecast PDF is too narrow and a reliable 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. For example, the IC/LBC perturbations are extracted from relatively coarse (30- to 45-km Δx) SREF members and do not account for smaller scale errors that would be present on the 4-km grids (Nutter et al. 2004). At the 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 the impact that under-dispersion has on the apparent *n* required to effectively sample a PDF in an idealized framework, two normally distributed 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 of these PDFs is randomly sampled using *n* from 2 to 100. For each *n*, 1000 sets of synthetic "members" are drawn and probabilities for rainfall greater than 12.7-mm (marked by vertical line in Fig. 4c) are computed for each set using the same 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 idealized case, if the tolerable error is considered ≤ 0.05 (marked by horizontal line in Fig. 4d), then the errors associated with probabilities derived from the PDF with σ = 1.87-mm would on average fall below the tolerable error using a minimum of 20 members. However, if the ensemble was reliable (i.e., σ = MSE = 3.0-mm), about 60 members would be required to fall below the tolerable error.

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

Not to be overlooked are the actual values of the 2009 SSEF system ROC areas (Fig. 3). For the full ensemble, ROC areas range between 0.88 and 0.95 out to forecast hour 30 for even the finest spatial scales. These values are quite skillful, especially considering the relatively low predictability typically associated with May-June (Fritsch and Carbone 2004) and the relatively short accumulation periods.

In summary, results of this study are encouraging because PQPFs derived from the 2009 SSEF system were found, on average, to be quite skillful. Additionally, relatively small *n* (~ 3 to 9 members) was found, on average, to have statistically indistinguishable ROC areas relative to the full 17 member ensemble, and the *n* at which skill began to level off increased (decreased) with increasing forecast lead time (spatial scale). However, if the SSEF system had larger spread that better matched the MSE, more members would be required to reach the point of "diminishing returns". In addition, relatively low predictability regimes and/or rare events would require more members to reach a point of "diminishing returns". Nevertheless, clearly the spatial scale and forecast lead time needed for end-users should be carefully considered in future convection-allowing ensemble designs. Future work is needed to improve reliability of convection-allowing ensembles, and further evaluations are needed for weather regimes with varying degrees of predictability and/or rare events.

5. 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. NCAR scientists Drs. Jimy Dudhia, Morris Weisman, Greg Thompson and Wei Wang provided guidance on WRF-model configurations.

6. 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.