15B.5 Investigation of the Impact of Convection-Allowing Ensemble Size for Severe Weather Forecasting

Israel L. Jirak[1*], Adam J. Clark[2], Christopher J. Melick[1,3], and Steven J. Weiss[1]
[1]NOAA/NWS/NCEP/Storm Prediction Center, Norman, OK
[2]NOAA/OAR/National Severe Storms Laboratory, Norman, OK
[3]CIMMS, University of Oklahoma, Norman, OK

1. INTRODUCTION

One primary goal of annual Spring Forecasting Experiments (SFEs), which are co-organized by NOAA's National Severe Storms Laboratory and Storm Prediction Center and conducted in NOAA's Hazardous Weather Testbed (HWT), is documenting performance characteristics of experimental, convection-allowing modeling systems (CAMs). Since 2007, the number of CAMs (including CAM ensembles) examined in the SFEs has increased dramatically – six different CAM ensembles were examined in 2015 – and major advances have been made in creating, importing, processing, verifying, and providing analysis and visualization tools for these large and complex datasets. However, progress toward identifying optimal CAM ensemble configurations has been inhibited because the different CAM systems have been independently designed by our diverse collaborators, making it difficult to attribute differences in performance characteristics.

Given this background and recent recommendations to NOAA by the international UCACN Model Advisory Committee to unify model development through a collaborative, evidence-driven approach, a much more coordinated effort was established for SFE2016 with regard to convection-allowing ensemble design. This was achieved by working with collaborators on a common set of model specifications (e.g., model version, grid-spacing, domain size, physics, etc.) so that the simulations contributed by each collaborator could be combined to form one large, carefully designed ensemble known as the Community Leveraged Unified Ensemble (CLUE). The CLUE was comprised of 65 members contributed by five research institutions, and represents an unprecedented effort to help guide NOAA's operational modeling efforts. Eight unique experiments were designed within the CLUE framework to examine issues directly relevant to the design of NOAA's future operational CAM-based ensembles.

This paper will focus on one of the experiments from the CLUE that explored the impact of ensemble size on convection-allowing forecasts. The basic configuration of the ensemble size experiment can be found in the following section. Results from the comparison of ensembles of different sizes during SFE2016 are presented in the third section, followed by conclusions and discussion.

* *Corresponding author address*: Israel L. Jirak, NOAA/NWS/NCEP/Storm Prediction Center, 120 David L. Boren Blvd., Norman, OK 73072; e-mail: Israel.Jirak@noaa.gov

2. ENSEMBLE CONFIGURATION

One of the CLUE experiments examined during the 2016 HWT SFE was the impact of ensemble size, which involved comparing mixed-core ensembles with equal contributions of NMMB and ARW members of 2, 4, 6, 10, and 20 total members. Each model core used constant physics in this design: ARW members used the Thompson microphysics scheme and MYJ planetary boundary layer (PBL) scheme while NMMB members used the Ferrier-Aligo microphysics scheme with the MYJ PBL scheme. The model runs were initialized at 0000 UTC, and none of the members included radar data assimilation. The initial conditions (ICs) and lateral boundary conditions (LBCs) for the first two members (i.e., one ARW and one NMMB member) were from the NAM model while IC perturbations for additional members were extracted from various Short-Range Ensemble Forecast (SREF) system members and applied to the NAM analysis with LBCs provided by the corresponding SREF member forecasts (Table 1). Ensembles of larger size were simply constructed by adding members to the ensembles of smaller size (e.g., the 4-member ensemble adds 2 perturbed members to the 2-member ensemble and so on).

Table 1. CLUE ensemble size configuration. NAMa and NAMf refer to 12-km NAM analysis and forecast, respectively. The model names appended with "pert" refer to perturbations and forecasts extracted from the 16-km SREF member.

#	Model	IC	BC
1	ARW	NAMa	NAMf
2	NMMB	NAMa	NAMf
3	ARW	NAMa + arw-p1_pert	arw-p1
4	NMMB	NAMa + nmmb-n1_pert	nmmb-n1
5	ARW	NAMa + nmmb-p2_pert	nmmb-n2
6	NMMB	NAMa + arw-p2_pert	arw-p2
7	ARW	NAMa + arw-n2_pert	arw-n2
8	NMMB	NAMa + arw-n1_pert	arw-n1
9	ARW	NAMa + nmmb-n1_pert	nmmb-n1
10	NMMB	NAMa + nmmb-p2_pert	nmmb-p2
11	ARW	NAMa + arw-n1_pert	arw-n1
12	NMMB	NAMa + arw-p3_pert	arw-p3
13	ARW	NAMa + arw-p2_pert	arw-p2
14	NMMB	NAMa + nmmb-p1_pert	nmmb-p1
15	ARW	NAMa + arw-p3_pert	arw-p3
16	NMMB	NAMa + nmmb-n2_pert	nmmb-n2
17	ARW	NAMa + nmmb-p1_pert	nmmb-p1
18	NMMB	NAMa + arw-p1_pert	arw-p1
19	ARW	NAMa + nmmb-n2_pert	nmmb-n2
20	NMMB	NAMa + arw-n2_pert	arw-n2

3. RESULTS

Forecasts from the 6, 10, and 20-member CLUE ensembles were available for evaluation during SFE2016, providing an opportunity for comparisons among the convection-allowing ensembles of different sizes. It was apparent during the first few days of SFE2016, both subjectively and objectively, that the forecasts from the larger-member ensembles were not much different than those from the smaller-member ensembles. In fact, the distribution of subjective ratings of the usefulness of the guidance from the 6, 10, and 20 member ensembles is essentially identical during the five-week SFE2016 (Fig. 1). This initial finding led to the exploration of the statistical performance of ensembles with 2 and 4 members following SFE2016. There were two primary components to the evaluation of the convection-allowing ensembles: 1) objective verification of ensemble neighborhood probabilities of reflectivity ≥40 dBZ and 2) objective verification of 4-hour ensemble neighborhood probabilities of hourly maximum updraft helicity (UH; Kain et al. 2008), relative to preliminary storm reports.

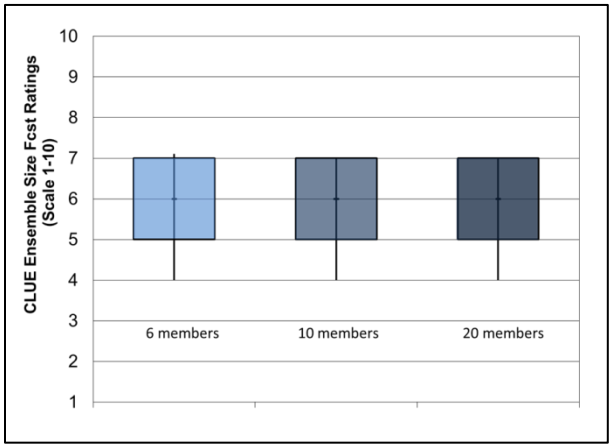

Figure 1. Subjective ratings (on a scale of 1-10 with 10 being the highest rating) of ensemble neighborhood probabilistic reflectivity forecasts ≥40 dBZ from the CLUE ensemble size experiment during SFE2016.

3.1 Objective Verification of Reflectivity Forecasts

The fractions skill score (FSS; Roberts and Lean 2008; Schwartz et al. 2010) was calculated for the ensemble neighborhood probability of 1-km AGL simulated reflectivity ≥40 dBZ using observed radar reflectivity for verification. When looking at the FSS for reflectivity by forecast hour during SFE2016 (Fig. 1), increasing the ensemble size had only small positive impact on FSS. The biggest improvement in skill occurred when increasing the number of members from two to four with relatively small improvements for further addition of members. Overall, there was only ~13% increase in FSS by increasing the membership tenfold. At least for this particular configuration (Section 2), the improvement in forecast skill is not worth the additional computing resources required to run more than six CAM members.

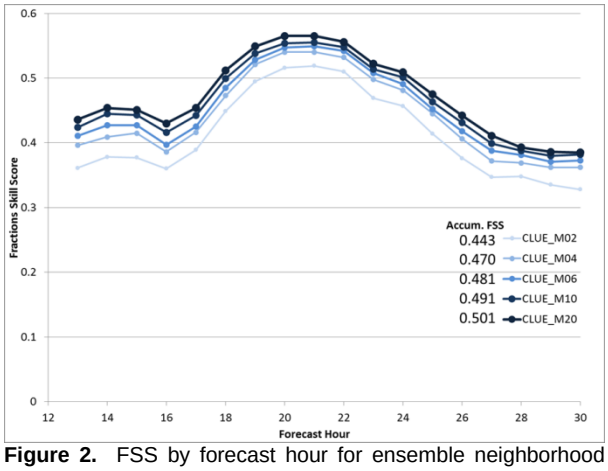

Figure 2. FSS by forecast hour for ensemble neighborhood probabilistic reflectivity forecasts ≥40 dBZ from the CLUE ensemble size experiment during SFE2016.

The relative operating characteristic (ROC) curves generated for probabilistic forecasts of 1-km AGL simulated reflectivity ≥40 dBZ provided another statistical perspective on the performance of the ensembles (Fig. 3). Similar to the results for FSS, the area under the ROC curves were comparable for the ensembles of different sizes. While increasing ensemble membership led to greater ROC areas, the improvement was relatively small, especially considering the additional computational expense.

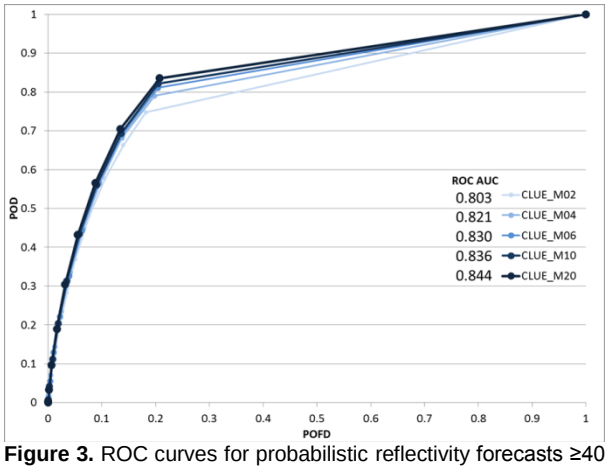

Figure 3. ROC curves for probabilistic reflectivity forecasts ≥40 dBZ from the CLUE ensemble size experiment during SFE2016.

Since ROC diagrams and areas are not sensitive to forecast biases (Wilks 2006), reliability diagrams were also examined for probabilistic forecasts from the ensembles (Fig. 4). The underdispersive nature of the CAM ensembles (of any size) resulted in a strong overforecast bias (e.g., 50% forecast probability only verified ~18% of the time). Even though the larger-member ensembles showed better reliability for forecasts ≥50% than the smaller-member ensembles, all

of them fell below the theoretical "no-skill" line. Again, these results were a function of the configuration of the ensembles, which had limited diversity and forecast spread.

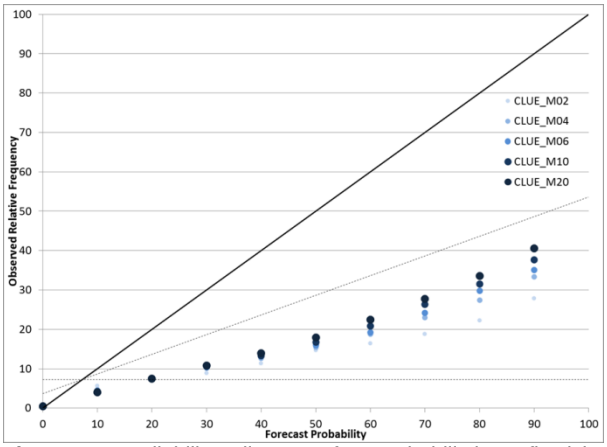

Figure 4. Reliability diagram for probabilistic reflectivity forecasts ≥40 dBZ from the CLUE ensemble size experiment during SFE2016.

With regard to the similarity of the forecasts from the ensembles of different sizes, one hypothesis for the resemblance was that the size of the neighborhood utilized (i.e., 40-km) was too large, resulting in similar probabilities regardless of ensemble size. To explore that hypothesis, ensemble neighborhood probabilities using a 20-km ROI were verified and compared to the 40-km neighborhood results (Fig. 5). In using a smaller neighborhood size, the addition of members resulted in only slightly larger percentage improvement over using a larger neighborhood. In addition, the overall FSS were much lower using a 20-km neighborhood compared to using a 40-km neighborhood. This finding suggests that the similar verification results among different ensemble sizes was likely not an artifact of the neighborhood size utilized, but more likely a function of limited spread in the ensemble forecasts.

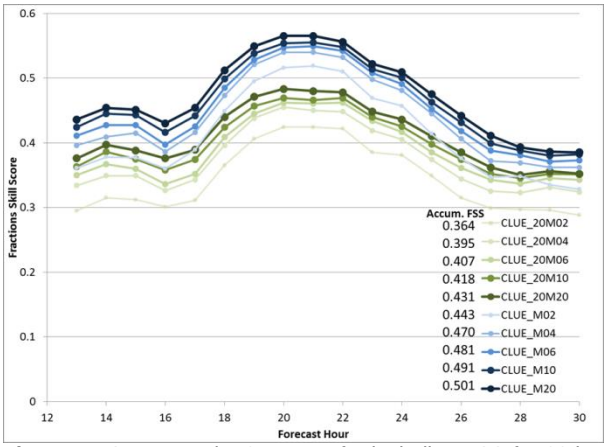

Figure 5. Same as Fig. 2, except for including FSS for 20-km ROI neighborhood probabilities (green lines) for the different ensemble sizes.

3.2 Objective Verification of UH Forecasts

The FSS was also calculated for severe weather forecasting using a methodology very similar to the surrogate-severe approach outlined by Sobash et al. (2016). Essentially, the 4-h ensemble neighborhood probabilities of UH ≥50 m^2s^{-2} were verified with the practically perfect hindcast (Hitchens et al. 2013) generated from preliminary local storm reports during the same 4-h period. FSS for two 4-h periods were specifically examined here corresponding with the typical peak in severe weather activity: 1800-2200 UTC and 2200-0200 UTC (Fig. 6). While increasing the ensemble size had a slightly larger effect on FSS for severe weather forecasting compared to the reflectivity forecasts, the impact of ensemble size was still relatively small.

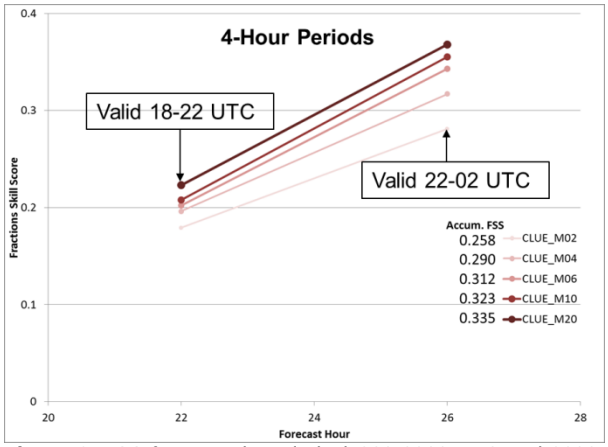

Figure 6. FSS for two 4-h periods (1800-2200 UTC and 2200-0200 UTC) for ensemble neighborhood probabilistic updraft helicity forecasts ≥50 m^2s^{-2} from the CLUE ensemble size experiment during SFE2016.

4. CONCLUSIONS

An unprecedented effort was made in the HWT during SFE2016 to coordinate CAM ensemble configurations much more closely than in previous SFEs, which was done in the context of the CLUE. The CLUE allowed for an experiment to explore the impact of ensemble size on CAM forecasts, which involved comparing mixed-core ensembles with equal contributions of NMMB and ARW members. Subjectively, SFE2016 participants rated the ensembles with 6, 10, and 20 members similarly, indicating little practical difference among the ensemble forecasts. Statistically, the improvement by increasing the ensemble membership was relatively small.

Overall, the ensembles were very underdispersive for reflectivity forecasts, and adding more members had limited benefit for this ensemble configuration (e.g., using single physics per core). In addition, applying SREF perturbations did not appear to provide adequate diversity for running more than six CAM members for this type of ensemble configuration. More work is needed in understanding and applying scale-appropriate

perturbations to provide sufficient spread for convection-allowing ensembles on Day 1 (i.e., forecast hours 0-36).

Acknowledgements. The CLUE would not have been possible during SFE2016 without dedicated participants and the support and assistance of numerous individuals at SPC and NSSL. In particular, collaboration with OU CAPS was vital to the success of the ensemble size experiment of CLUE. Ming Xue (OU CAPS), Fanyou Kong (OU CAPS), Kevin Thomas (OU CAPS), and Keith Brewster (OU CAPS) were essential in generating and providing access to ensemble forecasts examined on a daily basis.

REFERENCES

Hitchens, N. M., H. E. Brooks, and M. P. Kay, 2013: Objective limits on forecasting skill of rare events. *Wea. Forecasting*, **28**, 525–534

Kain, J. S., S. J. Weiss, D. R. Bright, M. E. Baldwin, J. J. Levit, G. W. Carbin, C. S. Schwartz, M. L. Weisman, K. K. Droegemeier, D. B. Weber, K. W. Thomas, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931-952.

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.

Schwartz, C. S., J. S. Kain, S. J. Weiss, M. Xue, D. R. Bright, F. Kong, K. W. Thomas, J. J. Levit, M. C. Coniglio, and M. S. Wandishin, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263–280.

Sobash, R. A., C. S. Schwartz, G. S. Romine, K. Fossell, and M. Weisman, 2016: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Wea. Forecasting*, **31**, 255–271

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences: An Introduction*. 2nd ed. Academic Press, 467 pp.