SPRING FORECASTING EXPERIMENT 2016

Conducted by the

EXPERIMENTAL FORECAST PROGRAM

of the

NOAA HAZARDOUS WEATHER TESTBED

http://hwt.nssl.noaa.gov/Spring_2016/

HWT Facility – National Weather Center
2 May - 3 June 2016

Preliminary Findings and Results

Adam Clark[2], Israel Jirak[1], Kent Knopfmeier[2,3], Chris Melick[1,3], Gerry Creager[2,3], Burkely Gallo[2,4], Steven Weiss[1], Jack Kain[2], James Correia[1,3], Mackenzie Krocak[2,4], Harold Brooks[2], Louis Wicker[2], Kelton Halbert[2,4], David Imy, Bill Skamarock[5], and Brian Ancell[6]

(1) NOAA/NWS/NCEP Storm Prediction Center, Norman, Oklahoma
(2) NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma
(3) Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma
(4) School of Meteorology, University of Oklahoma
(5) National Center for Atmospheric Research, Boulder, Colorado
(6) Texas Tech University, Lubbock, Texas

1. Introduction

The 2016 Spring Forecasting Experiment (SFE2016) was conducted from 2 May – 3 June by the Experimental Forecast Program (EFP) of the NOAA/Hazardous Weather Testbed (HWT). SFE2016 was co-led by the Storm Prediction Center (SPC) and National Severe Storms Laboratory (NSSL). In addition, important contributions of convection-allowing models (CAMs) were made from collaborators including the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma, Earth Systems Research Laboratory/Global Systems Division (ESRL/GSD), University of North Dakota (UND), United Kingdom Meteorological Office (UK Met Office), National Center for Atmospheric Research (NCAR), and NCEP's Environmental Modeling Center (EMC). Participants included more than 80 forecasters, researchers, and model developers from around the world (see Table 1 in Appendix). As in previous years, SFE2016 aimed to test emerging concepts and technologies designed to improve the prediction of hazardous convective weather, with several primary goals consistent with the Forecasting a Continuum of Environmental Threats (FACETs; Rothfusz et al. 2014) and Warn-on Forecast (WoF; Stensrud et al. 2009) visions:

Operational Product and Service Improvements:

- Explore the ability to generate higher temporal resolution Day 1 convective outlooks than those issued operationally by SPC.
    - 4-h periods for individual severe hazards (tornado, hail, and wind), and all hazards combined
    - Isochrones were used to delineate the start-time of 4-h time windows with the highest total severe probabilities
- Generate experimental Day 2 convective outlooks containing probabilistic forecasts for individual hazards (tornado, hail, wind), to provide more specific threat information compared to current operational SPC Day 2 total severe storm outlooks.

Applied Science Activities:

- Compare various convection-allowing ensembles to identify strengths and weaknesses of different configuration strategies using the framework of the Community Leveraged Unified Ensemble (CLUE).
    - Compare the skill of ARW-based, NMMB-based, and multi-core ensembles.
    - Examine the impact of radar data assimilation using two similarly configured ensembles with and without radar data assimilation.
    - Assess the impact of ensemble size using three similarly configured multi-core ensembles (i.e., equal membership between WRF-ARW and NMMB) comprised of 6, 10, and 20 members.
    - Document characteristics of various microphysics schemes used with the WRF model.
    - Compare CLUE subsets to the Storm Scale Ensemble of Opportunity (SSEO) as a baseline.
- Utilize convection-allowing ensemble forecasts in generating convective outlooks for Day 2, including individual severe hazards.
- Compare and assess different approaches in CAMs for predicting hail size.
- Provide an assessment of the capability of the NCAR global Model for Prediction Across Scales (MPAS) with variable resolution (3 km grid-spacing over the CONUS) in generating realistic and operationally useful prediction of convective storms out to Day 5.
- Evaluate and compare forecasts from the NAM Rapid Refresh (NAMRR; an experimental, rapidly updating NCEP model), and the operational and development versions of the High Resolution Rapid Refresh (HRRR) model.
- Explore the use and application of ensemble sensitivity analyses in a CAM ensemble.

As in previous experiments, a suite of state-of-the-art experimental CAM guidance contributed by our large group of collaborators was central to SFE2016. However, this year a major effort was made to coordinate CAM-based ensemble configurations much more closely than in previous years. Specifically, instead of each group providing a separate, independently designed CAM-based ensemble, all groups agreed on a set of model specifications (e.g., grid-spacing, vertical levels, domain size, physics) so that the simulations contributed by each group could be used in carefully designed controlled experiments. This design allowed us to conduct several experiments geared toward identifying optimal configuration strategies for CAM-based ensembles, and was especially well timed to help inform the design of the first operational CAM-based ensemble for the US, which is planned for implementation by NOAA's NCEP/EMC in the upcoming years. This large number of CAM members has been termed the Community Leveraged Unified Ensemble, or CLUE, and included 65 members using 3-km grid-spacing that allowed for a set of eight unique experiments.

This document summarizes the activities, core interests, and preliminary findings of SFE2016. More detailed information on the organizational structure and mission of the HWT, model and ensemble configurations, and information on various forecast tools and diagnostics can be found in the operations plan (http://hwt.nssl.noaa.gov/Spring_2016/HWT_SFE2016_operations_plan_final.pdf). The remainder of this document is organized as follows: Section 2 provides an overview of the models and ensembles examined during SFE2016 along with a description of the daily activities, Section 3 reviews the preliminary findings of SFE2016, and Section 4 contains a summary of the preliminary findings.

2. Description

a) Experimental Models and Ensembles

Building upon successful experiments of previous years, SFE2016 focused on the generation of experimental probabilistic forecasts of severe weather valid over shorter time periods than current operational SPC severe weather outlooks. This is an important step toward addressing a strategy within the National Weather Service (NWS) of providing nearly continuous probabilistic hazard forecasts on increasingly fine spatial and temporal scales (i.e., FACETs), in support of the NWS Weather-Ready Nation initiative. As in previous experiments, a suite of new and improved experimental CAM guidance including ensembles was central to the generation of these forecasts. For all of the models, hourly maximum fields (HMFs) of explicit storm attributes such as simulated reflectivity, updraft helicity, updraft speed, and 10-m wind speed, were examined as part of the experimental forecast and evaluation process. About 90 unique CAMs were run for SFE2016, of which 65 were a part of the CLUE system. Other deterministic and ensemble CAMs outside of the CLUE were contributed by NSSL, GSD, SPC, and the UK Met Office. To put the number of CAMs run for SFE2016 into context, Figure 1 shows the number of CAMs run for SFEs since 2007. There is a clear increasing trend, but consolidation of members contributed by various agencies into the CLUE made the increase in members more manageable.

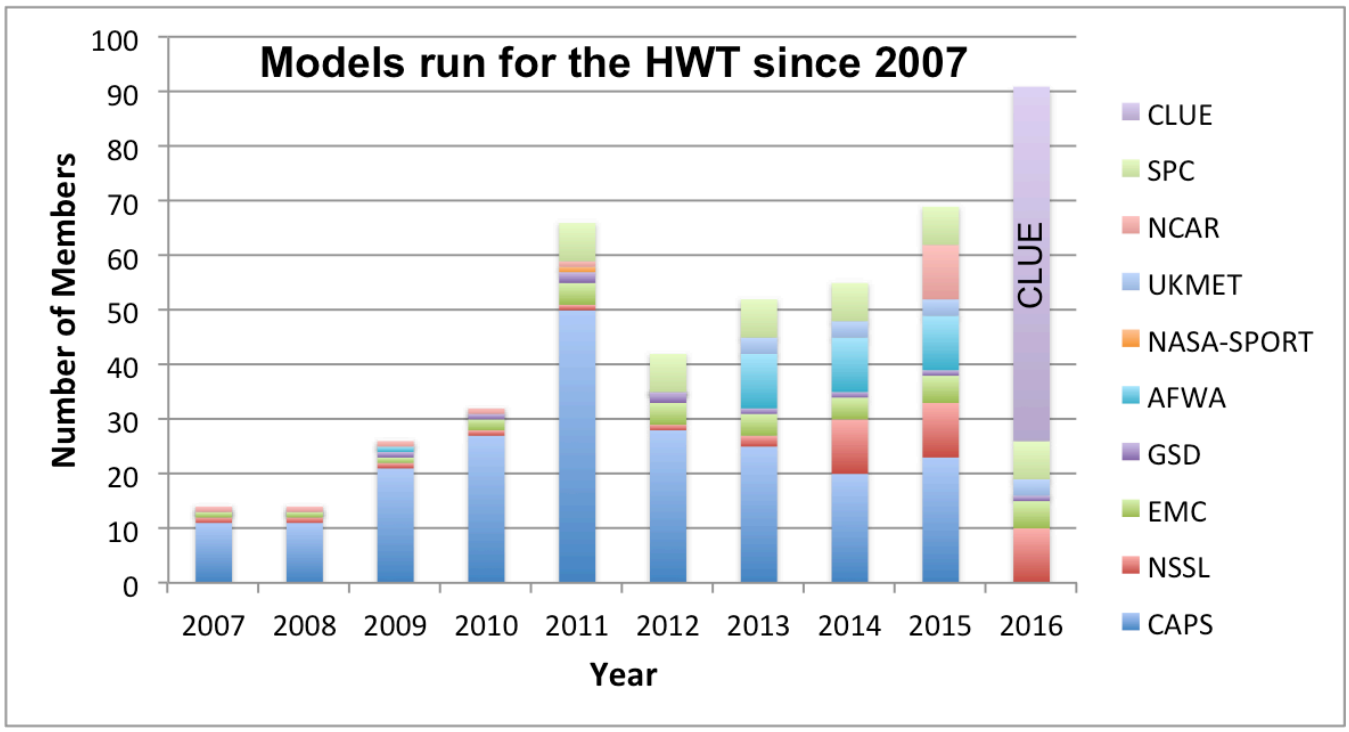

Figure 1 Number of CAMs run for SFEs since 2007. The different colored stacked bars indicate the contributing agencies. In SFE2016, the bar corresponding to the CLUE is marked by text.

More information on all the modeling systems run for SFE2016 is given below.

1) THE COMMUNITY LEVERAGED UNIFIED ENSEMBLE (CLUE)

The CLUE is a carefully designed ensemble with subsets of members contributed by NSSL, CAPS, UND, ESRL/GSD, and NCAR. In addition, EMC staff provided guidance on the NMMB configuration run by CAPS and NSSL, and the Developmental Testbed Center (DTC) provided support for post-processing. To ensure consistent post-processing, visualization, and verification, all CLUE contributors used the same post-processing software to output the same set of model output fields on the same grid. The post-processed model output fields are the same as the 2D fields output by the operational HRRR and were chosen because of their relevance to a broad range of forecasting needs, including aviation, severe weather, and precipitation. A small set of additional output fields requested by NCEP Centers, Weather Prediction Center (WPC), SPC, and Aviation Weather Center (AWC), were also included. All CLUE members were initialized weekdays at 0000 UTC with 3-km grid-spacing covering a CONUS domain. The ARW and NMMB members have matching horizontal and vertical grid specifications. A full description of all members and list of post-processed model fields are provided in the SFE2016 operations plan (Clark et al. 2016). Table 1 provides a summary of each CLUE subset.

During the first week of the experiment it was discovered that the caps-nmmb runs contained a bug that caused a very severe warm and dry bias in the near surface fields. It was found that the runs were being configured to ingest the International Geosphere-Biosphere Programme (IGBP) land surface parameter tables, but were reading in data from the United States Geological Table (USGS) tables. During the second week of SFE2016 this bug was fixed by having one of the participating EMC scientists, Jacob Carley, work with CAPS staff. The bug affected a total of 7 days, but since the conclusion of the experiment, all the affected runs have been rerun by NSSL. Figure 2 illustrates a forecast of 10-m dewpoint that contained the bug, and the same forecast after the bug fix was implemented.

Table 1 Summary of CLUE subsets. IC/LBC perturbations labeled "SREF" indicate that IC perturbations were extracted from members of NCEP's Short-Range Ensemble Forecast system and added to 0000 UTC NAM analyses. In subsets with "yes" indicated for mixed-physics, the microphysics and turbulence parameterizations were varied, except for subset mp, which only varied the microphysics. Note, the control member of the core ensemble was also used as the control member in the mp and s-phys-rad ensembles. Thus, although the total number of members adds to 67, there were 65 unique members. Further, one member planned for the core subset was not ready for real-time implementation, thus only 9 core members were actually run.

Clue Subset	# of mems	IC/LBC perturbations	Mixed Physics	Data Assimilation	Model Core	Agency
core	10 (9)	SREF	yes	ARPS-3DVAR	ARW	CAPS
s-phys-rad	10	SREF	no	ARPS-3DVAR	ARW	CAPS
caps-enkf	9	EnKF (CAPS)	yes	EnKF (CAPS)	ARW	CAPS
caps-nmmb-rad	1	none	no	ARPS-3DVAR	NMMB	CAPS
caps-nmmb	5	SREF	no	cold start	NMMB	CAPS
s-phys-norad	10	SREF	no	cold start	ARW	NSSL
nssl-nmmb	5	SREF	no	cold start	NMMB	NSSL
HRRR36	1	no	no	RAP-GSI/DFI	ARW	ESRL/GSD
ncar-enkf	10	EnKF (DART)	no	EnKF (DART)	ARW	NCAR
mp	5	no	yes	ARPS-3DVAR	ARW	UND

The design of CLUE allowed for 8 unique experiments that examined issues immediately relevant to the design of a NCEP/EMC operational CAM-based ensemble. These experiments are listed in Table 2.

*Table 2 List of CLUE experiments for SFE2016. Note, the GSD Radar vs. CAPS Radar Assimilation experiment (marked with *) was planned but not conducted because the core member that was going to be used was not ready for real-time implementation.*

Experiment Name	Description	CLUE subsets
ARW vs. NMMB	A direct comparison of subjective and objective skill of ARW and NMMB cores was conducted.	caps-nmmb, nssl-nmmb, & s-phys-norad
Multi-core vs. Single core	Three ensembles were compared to test the effectiveness of a single core vs. multi-core configuration. The first ensemble used 5 ARW and 5 NMMB members, the second 10 ARW members, and the third 10 NMMB members.	caps-nmmb, nssl-nmmb, & s-phys-norad
Single Physics vs. Multi-physics	An ensemble with perturbed ICs/LBCs was used to test whether there was a noticeable advantage when using multiple PBL and microphysics parameterizations vs. common physics in all members.	core, s-phys-rad
Ensemble radar vs. Ensemble No Radar	A single physics ensemble was used to test the influence of assimilating radar data. In particular, the longest forecast length at which the radar data had a noticeable influence was assessed.	s-phys-rad, s-phys-norad
3DVAR vs. EnKF	The 3DVAR and EnKF data assimilation approaches were compared. Note, this experiment was not as controlled as the others because there were other different aspects of the configurations in the subsets with different data assimilation.	core, caps-enkf, ncar-enkf
*GSD Radar vs. CAPS Radar Assimilation	Two methods for assimilating radar data were compared. One used ARPS-3DVAR and the other the DDFI system used in the HRRR.	core, HRRR36
Microphysics Sensitivities	The impact of different microphysical parameterizations on the resulting convective storm forecasts was examined.	mp
Ensemble Size Experiment	A comparison of ensembles with equal contributions of NMMB and ARW members using 6, 10, and 20 members was conducted to examine the impact of ensemble size.	caps-nmmb, s-phys-norad, nssl-nmmb

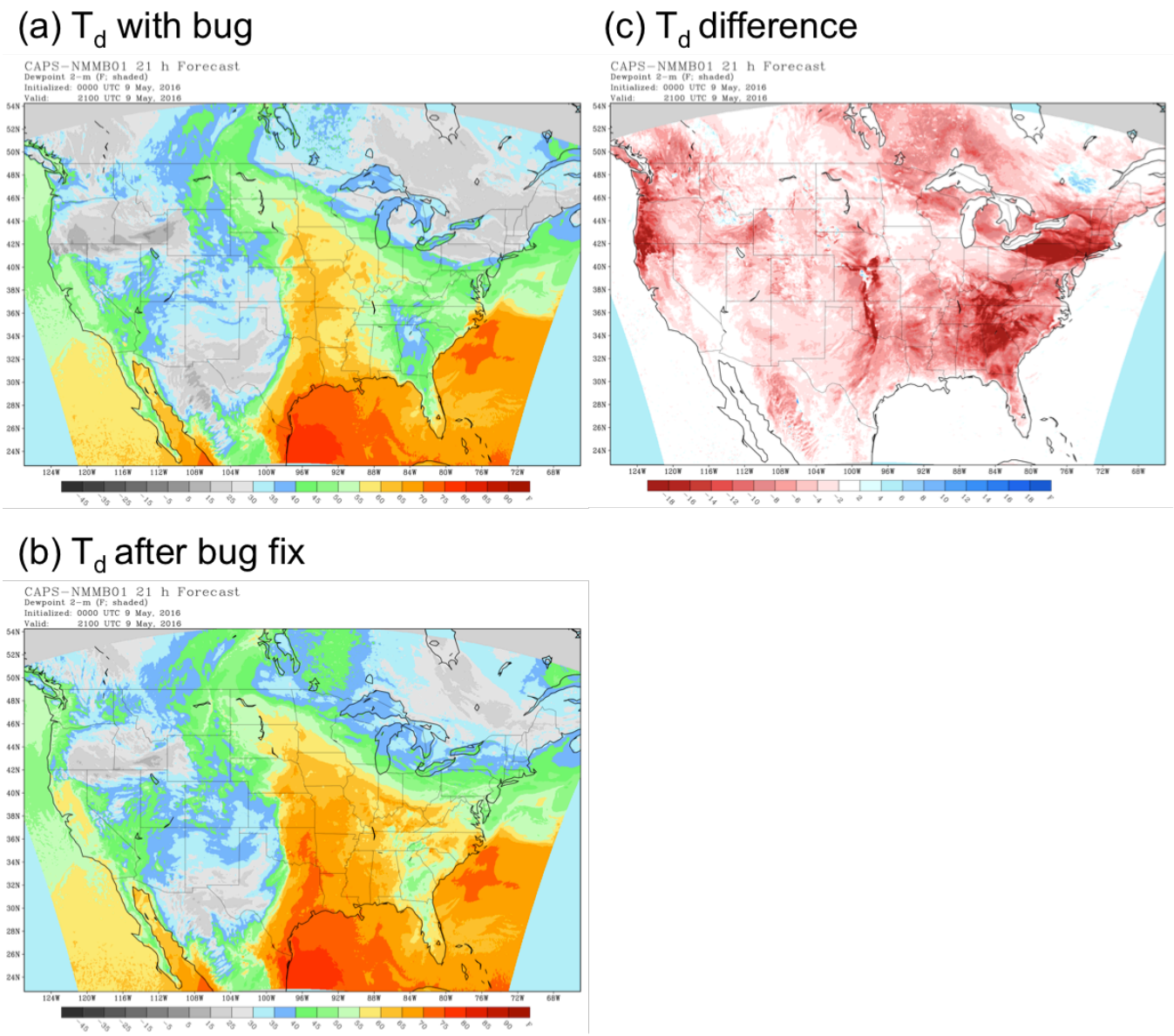

Figure 2 10-m AGL dewpoint temperatures at 21 h forecast lead time from a 0000 UTC 9 May, 2016 initialization of one of the caps-nmmb members (a) with the land surface bug, (b) without the land surface bug, and (c) the difference between them.

2) THE STORM SCALE ENSEMBLE OF OPPORTUNITY (SSEO)

The SPC Storm-Scale Ensemble of Opportunity (SSEO) is a 7-member, multi-model and multi-physics convection-allowing ensemble consisting of deterministic CAMs with ~4-km grid spacing available to SPC year-round. This "poor man's ensemble" has been utilized in SPC operations since 2011 with forecasts to 36 h from 0000 and 1200 UTC and provides a practical alternative to a formal/operational storm-scale ensemble, which has not been available operationally, owing to computational limitations in NOAA. All members were initialized as a "cold start" from the operational NAM or the RAP – i.e., no additional data assimilation was used to produce ICs.

3) THE NSSL-WRF AND NSSL-WRF ENSEMBLE

SPC forecasters have used output from an experimental 4-km grid-spacing WRF-ARW produced by NSSL (hereafter NSSL-WRF) since the fall of 2006. Currently, this WRF model is run twice daily at 0000 UTC and 1200 UTC throughout the year over a full-CONUS domain with forecasts to 36 hours.

For the third year, the NSSL-WRF ensemble was part of the experimental numerical guidance. This ensemble includes eight additional 4-km WRF-ARW runs that – along with the deterministic NSSL-WRF – comprised a nine-member NSSL-WRF-based ensemble. The additional eight members were initialized at 0000 UTC and use 3-h forecasts from the 2100 UTC NCEP Short Range Ensemble Forecast (SREF) system for initial conditions (ICs) and corresponding SREF member forecasts as lateral boundary conditions (LBCs). The physics parameterizations for each member are identical to the deterministic NSSL-WRF. Although the unvaried physics will have lower spread than a multi-physics ensemble, SPC forecasters and NSSL scientists are very familiar with the behavior of the NSSL-WRF physics, and this configuration will allow for the isolation of spread contributed only by varying the ICs/LBCs.

4) UKMET CONVECTION-ALLOWING MODEL RUNS

Three nested, limited-area high-resolution versions of the Met Office Unified Model (UM) running once per day were provided to SFE2016: two at 2.2 km grid spacing and one at 1.1 km. The operational 2.2-km version had 70 vertical levels across a slightly sub-CONUS domain. Taking its initial and lateral boundary conditions from the 00Z 17-km horizontal grid-spacing global configuration of the UM, the 2.2-km model was initialized without additional data assimilation and ran out to 48 hours. This model configuration included a 3D turbulent mixing scheme using a locally scale-dependent blending of Smagorinsky and boundary layer mixing schemes. Stochastic perturbations were made to the low-level resolved-scale temperature field in conditionally unstable regimes (to encourage the transition from subgrid to resolved scale flows) and the microphysics was single moment. Partial cloudiness was diagnosed assuming a triangular moisture distribution with a width that is a universally specified function of height only. A parallel version of the 2.2-km model was also run with a new scheme that addresses the moisture conservations issues in the model. The 1.1 km run was nested within the 2.2-km run with a slightly smaller domain centered over eastern Oklahoma. All UM simulations were run without convective parameterization.

5) NCAR MODEL FOR PREDICTION ACROSS SCALES (MPAS)

NCAR'S Model for Prediction Across Scales (MPAS; Skamarock et al. 2012) was examined for the second year during SFE2016. MPAS produced daily 0000 UTC initialized forecasts at 3-km grid-spacing over the CONUS with forecasts to 120 h (5 days). The MPAS horizontal mesh was based on Spherical Centriodal Voronoi Tesselations (SCVTs). These meshes allowed for both quasi-uniform discretization of the sphere and local refinement with smoothly varying mesh spacing between regions with differing resolutions. The smoothly varying mesh eliminates the major problems encountered with mesh transitions in forecast systems using traditional grid-nesting.

6) NORTH AMERICAN MESOSCALE RAPID REFRESH (NAMRR) SYSTEM

The NCEP experimental North American Mesoscale Rapid Refresh system (NAMRR) is an hourly-updated version of the North American Mesoscale (NAM) forecast system and its data assimilation system that assimilates radar data using the ESRL Diabatic Digital Filter Initialization (DDFI) technique. All NAMRR forecasts

were at least 18 h, and to maintain compatibility with the operational NAM, 60 h nested and 84 h large domain forecasts were issued at 0000, 0600, 1200, and 1800 UTC. The NAMRR was used during the SFE2016 forecast process and a formal evaluation activity was conducted comparing the NAMRR to the HRRR.

b) Daily Activities

SFE2016 activities were focused on forecasting severe convective weather at two separate desks, one forecasting individual hazards and the other forecasting total severe, with different experimental forecast products being generated at different temporal resolutions. Forecast and model evaluations also were an integral part of daily activities during SFE2016. A summary of forecast products and evaluation activities can be found below while a detailed schedule of daily activities is contained in the appendix.

1) EXPERIMENTAL FORECAST PRODUCTS

Similar to previous years, the experimental forecasts continued to explore the ability to add temporal specificity to longer-term convective outlooks. One desk mimicked the SPC operational Day 1 convective outlooks by producing separate probability forecasts of large hail, damaging wind, and tornadoes within 25 miles (40 km) of a point valid 1600 UTC to 1200 UTC the next day. On the other desk, a separate Day 1 forecast was made for total severe (combined hail, wind, and tornado) probabilities valid over the same period.

Each desk then manually stratified their respective Day 1 forecasts into periods with higher temporal resolution. Individual hazard probability forecasts of large hail, damaging wind, and tornadoes were generated for two four-hour periods: 1800-2200 UTC and 2200-0200 UTC. As an alternative way of stratifying the Day 1 forecast, the other desk generated five 4-h period outlooks of total severe at two-hour intervals covering the time periods: 1800-2200, 2000-0000, 2200-0200, 0000-0400, and 0200-0600 UTC. Additionally, the total severe desk drew isochrones of severe weather at two-hour intervals on top of the full period Day 1 total severe probabilities to delineate the start-time of the 4-h time window with the highest total severe probabilities. The 4-h time windows were chosen based on research finding that a strong majority (97%) of storm reports within 40-km of a point fall within a 4-h period. Another way to think of this is that, when SPC issues their first Day 1 convective outlook that is valid for a 24-h period, one can expect that almost all the severe weather potential will be confined to a much shorter time period (i.e. 4-h). Thus, if these time periods with the highest severe weather risk can be accurately delineated, it would provide very useful information to supplement the regular Day 1 convective outlooks. The isochrones were tested as an alternative (or supplement) to issuing more frequent outlooks valid for shorter time periods. The goals of testing these different approaches is to explore multiple ways of introducing probabilistic severe weather forecasts on time/space scales that are currently addressed with mostly categorical forecast products (i.e., SPC Mesoscale Discussions and Tornado/Severe Thunderstorm Watches), and to begin to explore ways of seamlessly merging probabilistic severe weather outlooks with probabilistic severe weather warnings as part of the NOAA WoF and FACETS initiatives.

In addition to the complete suite of observational and model data available in SPC operations, first-guess guidance for individual severe weather hazards was available to assist in generating the higher temporal resolution outlooks. Calibrated guidance for the individual hazards, as derived from the SREF (environment information) and SSEO (explicit storm attributes; Jirak et al. 2014), was available in 4-h periods. The 1600-1200 UTC human forecasts for the SPC Desk were also temporally disaggregated (Jirak et al. 2012) into the 4-h periods (1800-2200 UTC and 2200-0200 UTC) using SSEO guidance to provide additional timing information for the four-hour periods.

At the individual hazards desk, participants created their own short-time-window forecasts on Google Chromebooks using a web-based tool to draw severe weather probability lines. The participant forecasts were compared to one another and to a "control" forecast issued by the lead forecaster using N-AWIPS. At the total severe desk, the full period and short-time-window forecasts were issued as a group. However, the isochrones were drawn by five small groups using the Chromebooks. The lead forecaster at the total severe desk drew isochrones independently using N-AWIPS.

Severe weather forecasts were also generated for Day 2 to explore the feasibility of issuing forecasts of individual severe storm hazards beyond Day 1, where current SPC operational forecasts for Day 2 (and beyond) only consider probabilities of total severe. In particular, operational and experimental CAM guidance were examined to assist in the individual hazard forecasts for Day 2. Forecasts for total severe were also generated for Day 2 and/or Day 3 if time and interest allowed. This provided an opportunity to explore convection-allowing guidance from MPAS into Day 3.

Finally, each desk examined observational trends and morning/afternoon model guidance to update (or add to) their respective short-time-window forecasts made earlier in the day. The individual hazard forecasts were updated for the 2200-0200 UTC period while the total severe forecasts were updated for the 2200-0200, 0000-0400, and 0200-0600 UTC periods. In addition, the total severe desk updated their total severe isochrones, contouring the 2200, 0000, and 0200 UTC times.

2) FORECAST AND MODEL EVALUATIONS

While much can be learned from examining model guidance and utilizing it to help create experimental forecasts in real time, an important component of SFE2016 was to look back and evaluate the forecasts and model guidance from the previous day. There were two periods of formal evaluations during SFE2016. The first was during the morning when experimental outlooks from the previous day generated by both forecast teams were examined. In these next-day evaluations, the team forecasts and first-guess guidance was compared to observed radar reflectivity, severe weather reports, NWS warnings, and Multi-Radar Multi-Sensor (MRMS) radar estimated hail sizes.

Objective verification metrics were also computed for some of the experimental outlooks and first-guess guidance. Similar to SFE2014 and SFE2015, experimental probabilistic forecasts of tornado, wind, and hail were evaluated using the Critical Success Index (CSI) and Fractions Skill Score (FSS) based on the local storm reports (LSRs) as the verification event. Supplemental observations for hail from the MRMS-based Maximum Estimated Size of Hail (MESH) were also used in near real-time to calculated skill scores and gauge the usefulness of alternative sources for verification. A quality control measure was applied to the hourly MESH grids, which ensured the existence of nearby CG lightning flashes. Further, only spatially filtered grids were considered to ensure the presence of contiguous swaths in the MESH (Melick et al. 2014).

The second evaluation period occurred during the afternoon, which was focused on comparisons of different ensemble diagnostics and CLUE ensemble subsets. The total severe and individual hazards desks conducted two different sets of evaluations.

3. Preliminary Findings and Results

a) Evaluation of experimental forecast products – Total Severe Desk

SFE2016 participants subjectively evaluated the full period probabilistic forecasts of total severe each morning on a scale of 1-10. Specifically, participants were asked to, "*Use a rating scale from Very Poor (1) to Very Good (10). Areas with greater severe storm occurrence higher forecast probabilities, and the forecast or*

occurrence of significant reports, should be given more weight in the rating process. Also, take into account radar-derived severe weather proxy products in assessing the quality of the forecasts." An example image used to conduct full period ratings is shown in Figure 3. This forecast was made the morning of 17 May and verified the next day. Nine participants rated this forecast 7/10 while one rated it 6/10.

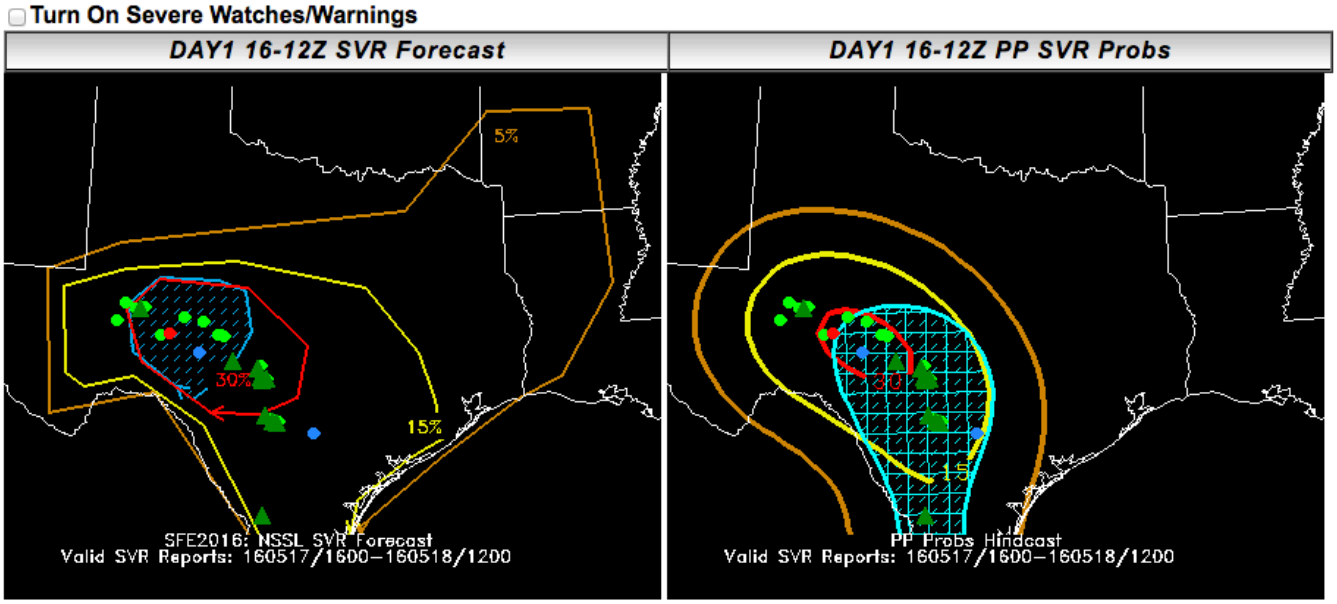

Figure 3 Left panel: Experimental Day 1 outlook for total severe weather valid 1600 – 1200 UTC 17-18 May 2016 with locations of storm reports overlaid. Right panel: Practically perfect hindcast probabilities with the locations of storm reports overlaid.

The distribution of subjective ratings for all of the full period outlooks (including Days 1, 2, and 3) is shown in Figure 4. It is important to note that each day of the experiment had a Day 1 full period outlook, while Day 2 and Day 3 outlooks were issued based on the anticipated severity and forecast uncertainty of the long-range weather. Day 2 forecasts were issued nearly every day, while Day 3 forecasts were only issued on 7 days. Most Day 1 outlooks had relatively high ratings, generally above a 5/10. The Day 2 outlooks tended to have more broadly distributed ratings, with nearly as many low ratings as high ratings. Since the issuance of Day 3 forecasts was reserved for days in which the weather was anticipated to be severe, the sample of Day 3 forecasts does not encompass the marginal severe weather days experienced in SFE2016. However, a majority of the forecasts were rated as either a 5 or a 6. Participants noted both displacement and magnitude issues in their comments. Participants also discussed whether to rate these the Day 3 forecasts as they would a Day 1 forecast. Because predictability generally increases with decreasing lead-time, a "good" Day 3 might not be considered a "good" forecast at Day 1 lead-time. Based on the subjective ratings and comments, most participants opted to rate the forecasts based on the correspondence with reports (i.e., not taking into consideration whether they were rating a Day 1 or Day 3 forecast).

In addition to the full-period outlooks, lead forecasters and participants generated a set of five, 4-h time window probabilities for total severe each morning. Three of these periods were updated in the afternoon using the most recent model guidance and observations. The distributions of subjective ratings for these forecasts are shown in Figure 5. The distributions of these rating are very similar and each time period has a median rating of 7. However, the afternoon updates did tend to have higher ratings than their morning counterparts, particularly for the 2200-0200 and 0000-0400 UTC periods. Similar to the full-period outlooks, participants comment on the extent and intensity of forecast probabilities, but mentioned more nuanced

forecast considerations such as convective initiation and the maintenance of cloud layers that inhibited convection.

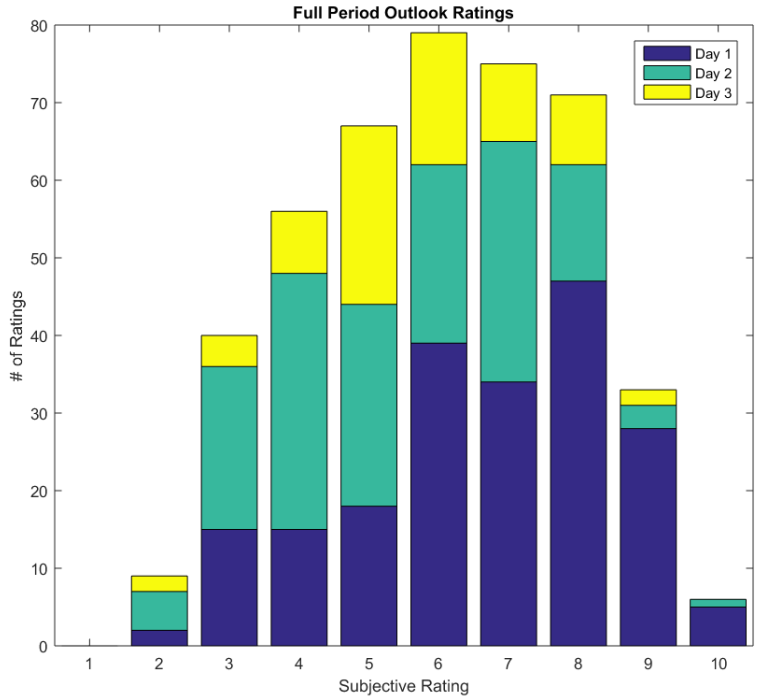

Figure 4 Distribution of subjective ratings (1-10) for the full-period hourly experimental forecasts issued by 1600 UTC. Day 1 ratings (blue) covered 1600 UTC – 1200 UTC the following day, while Day 2 (teal) and Day 3 (yellow) forecasts covered 1200 UTC – 1200 UTC.

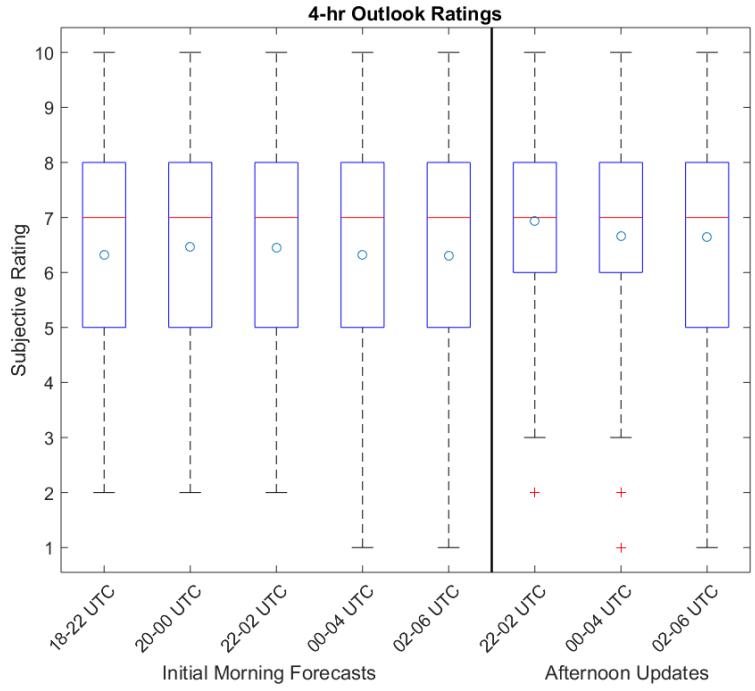

Figure 5 Box plots showing distributions of subjective ratings for the initial morning 4-h forecast periods and afternoon updates of those periods. The red line indicates the median rating for each period, the teal circle indicates the mean rating, and the red crosses are outliers.

Finally, the participants and lead forecaster drew isochrones of severe weather at two-hour intervals to delineate the start-time of the 4-h time window with the highest total severe probabilities. Recall, the lead forecaster drew their own isochrones, while participants broke down into 5 small groups that each generated their own set of isochrones. An example of isochrone forecasts and verifying storm reports is shown in Figure 6. For the next-day subjective evaluations, participants were asked to, "*Use a rating scale from Very Poor (1) to Very Good (10). Consider whether the majority of reports for each period fell within the correct areas indicated by the isochrones*". Note, these subjective ratings were only for the lead forecaster's isochrones forecasts. Participants were given the options to make comments about their own isochrones forecasts, but there was no formal rating assigned. For the initial isochrone forecasts that were issued in the late morning by the lead forecaster, the distribution of subjective ratings was relatively broad and non-skewed with a median rating of 7. For the final isochrones forecasts that were issued in the afternoon, the subjective ratings improved and the distribution was slightly right-skewed with a median rating of 8 (Fig. 7).

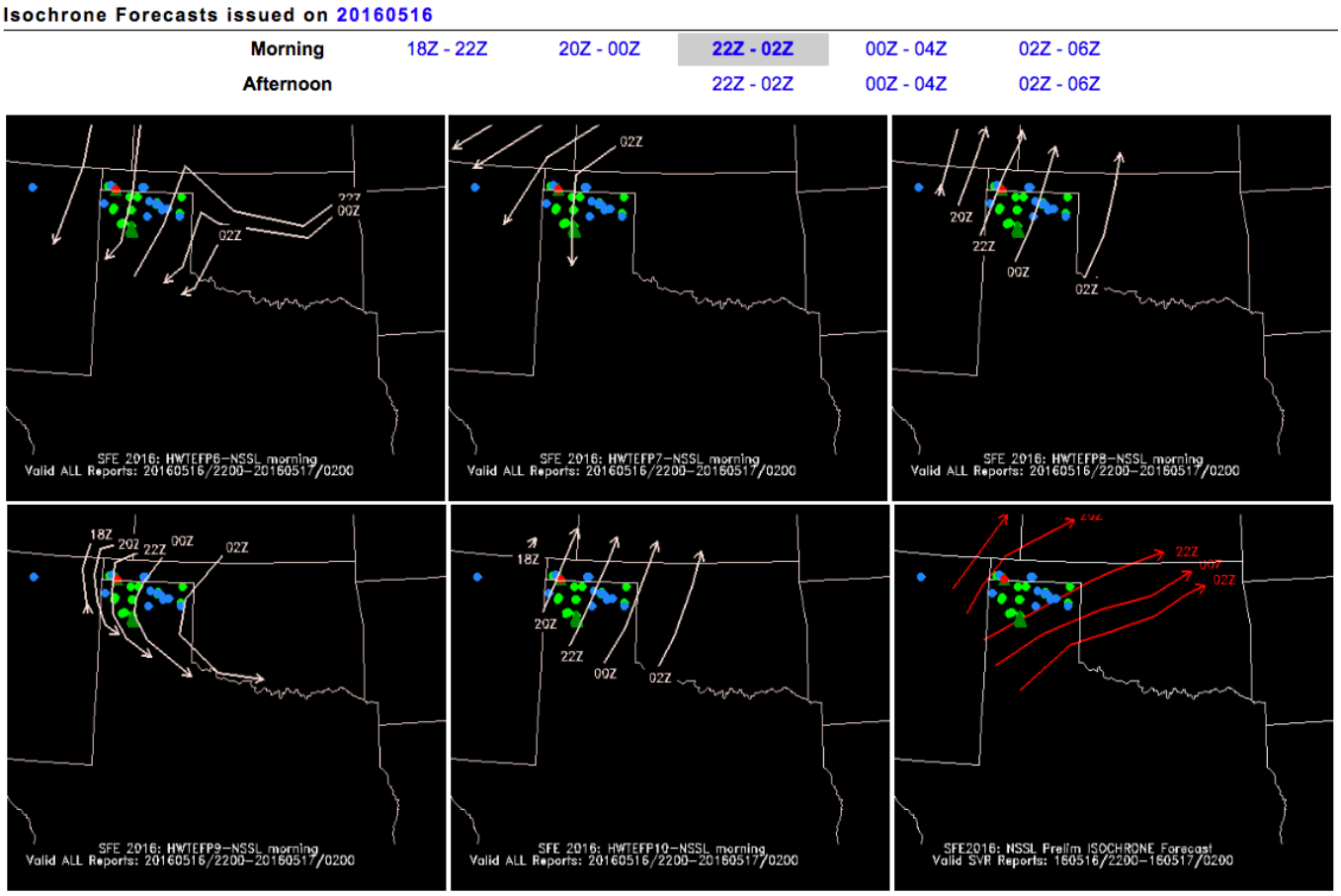

Figure 6 Isochrone forecasts for 5 May 2016. Storm reports that occurred during the 2200 – 0200 UTC period are overlaid. The lead forecaster of the total severe team generated the bottom right panel with the red isochrones, while all other panels display forecasts generated by participants.

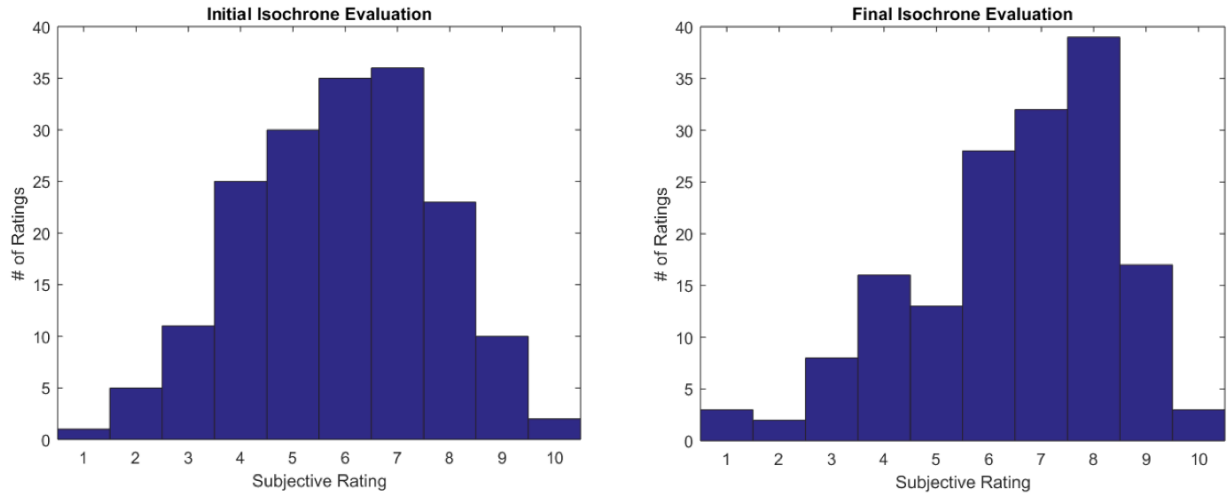

Figure 7 Distribution of subjective ratings for the initial (left) and final (right) isochrone forecasts.

Preliminary objective verification of isochrone forecasts has also been conducted by mapping forecast isochrones to an 80-km grid and comparing that grid to observed isochrones generated based on storm reports. To generate the grid of observed isochrones, the spatial distribution of reports that occurred within the 4-h time windows 1800-2200, 2000-0000, 2200-0200, 0000-0400, and 0200-0600 UTC mapped to an 80-km grid was spatially smoothed using a Gaussian kernel with a smoothing parameter of 120-km. Then each grid-point was assigned the time period with the highest smoothed probability, creating areas of timeframes that were contoured like isochrones. An example of forecast and observed isochrones is shown in Figure 8, and distributions of differences between forecast and observed severe weather timing derived from the isochrones is shown in Figure 9. Clearly, most forecasts indicated later time windows for the highest probability of severe weather than was observed. From examining the maps on individual days it appeared that the start period for the events was generally well forecast, but progression was often forecast to be too slow. As discussed during the recent "Probability of What workshop", this bias may be related to a time-centric forecaster mindset (i.e., where will severe weather be at a given time?), as opposed to location-centric mindset (i.e., when will the time period with the maximum probability of severe weather occur for a given point?).

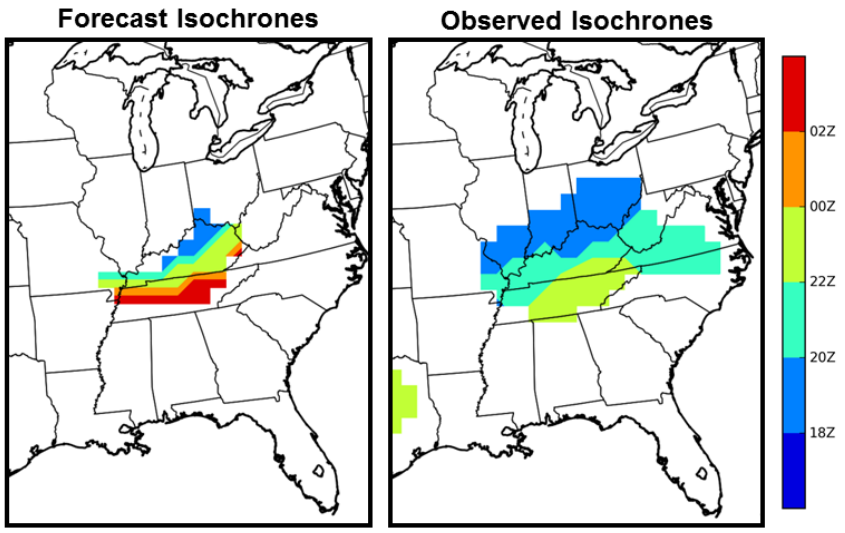

Figure 8 Example of forecast and observed isochrones valid 10 May 2016.

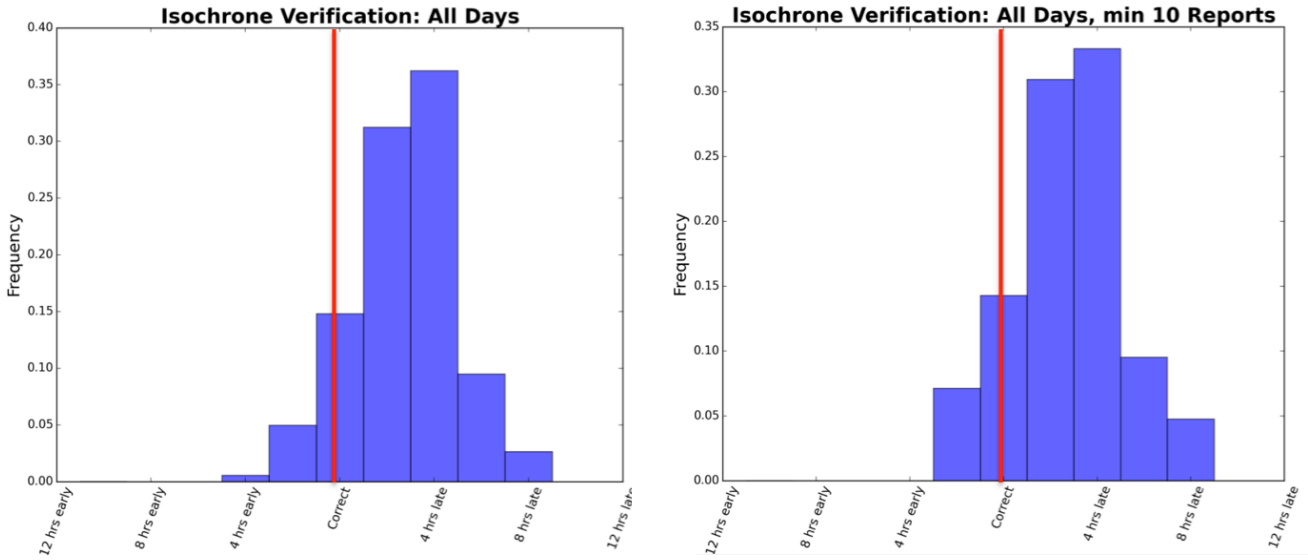

Figure 9 Relative frequencies of differences between forecast and observed severe weather timing as computed from grids derived from forecast and observed isochrones. Differences in timing were only computed for 80-km grid-points that fell within isochrones for both forecasts and observations. The left panel is for all of these points, while the right panel is only for days in which at least 10 reports occurred.

Automated isochrone forecasts were also generated using the NSSL-WRF ensemble. These were constructed by mapping maximum forecast UH within 4-h time windows from each ensemble member to an 80-km grid. Then, at each 80-km grid-point and time window, severe weather probabilities were derived by finding the ratio of ensemble members that forecast UH ≥ 40 m^2s^{-2} and applying a Gaussian kernel with σ = 90 km. Finally, the isochrones were derived by finding the 4-h time window at which these severe weather probabilities were highest. An example of these model-derived isochrones is shown in Figure 10. This automated product was not formally evaluated, but was generally very well received. In the SFE2016 report from the total severe desk lead forecaster Dave Imy, he states, "*In my opinion, the [model-derived guidance] provided excellent and broader guidance than our individual forecast. If isochrones were to be implemented as an operational forecast product in the future, the most efficient way to produce them might be by choosing an 'ensemble of choice' to generate them. Then a forecaster could make adjustments based on the diagnostic weather and other model guidance*."

During each week of the experiment, it seemed to take about two days for the new participants to figure out exactly what was being forecast using the isochrones. It was found that the isochrones were relatively easy to draw for progressively organized systems. However, in situations with weak flow in which mesoscale features instead of strong dynamical forcing dominate convective evolution, the isochrones became much more difficult to draw. For example, in situations with "back building" convection, a "perfect prog" would have had little or no spacing between the isochrones, which was not very intuitive. For testing isochrones during SFE2017, the total severe desk lead forecaster report brings up several questions/issues to address: the phrase "a majority or greatest severe coverage" has to be better defined, along with the number of reports needed to constitute an "accurate" isochrones forecast. Also, a "perfect prog" verification of isochrones should be generated to help participants calibrate their isochrone forecasts. Finally, it is recommended that 15% severe weather probability be the minimum probability for drawing isochrones, because lower probabilities usually imply only marginal and/or isolated reports.

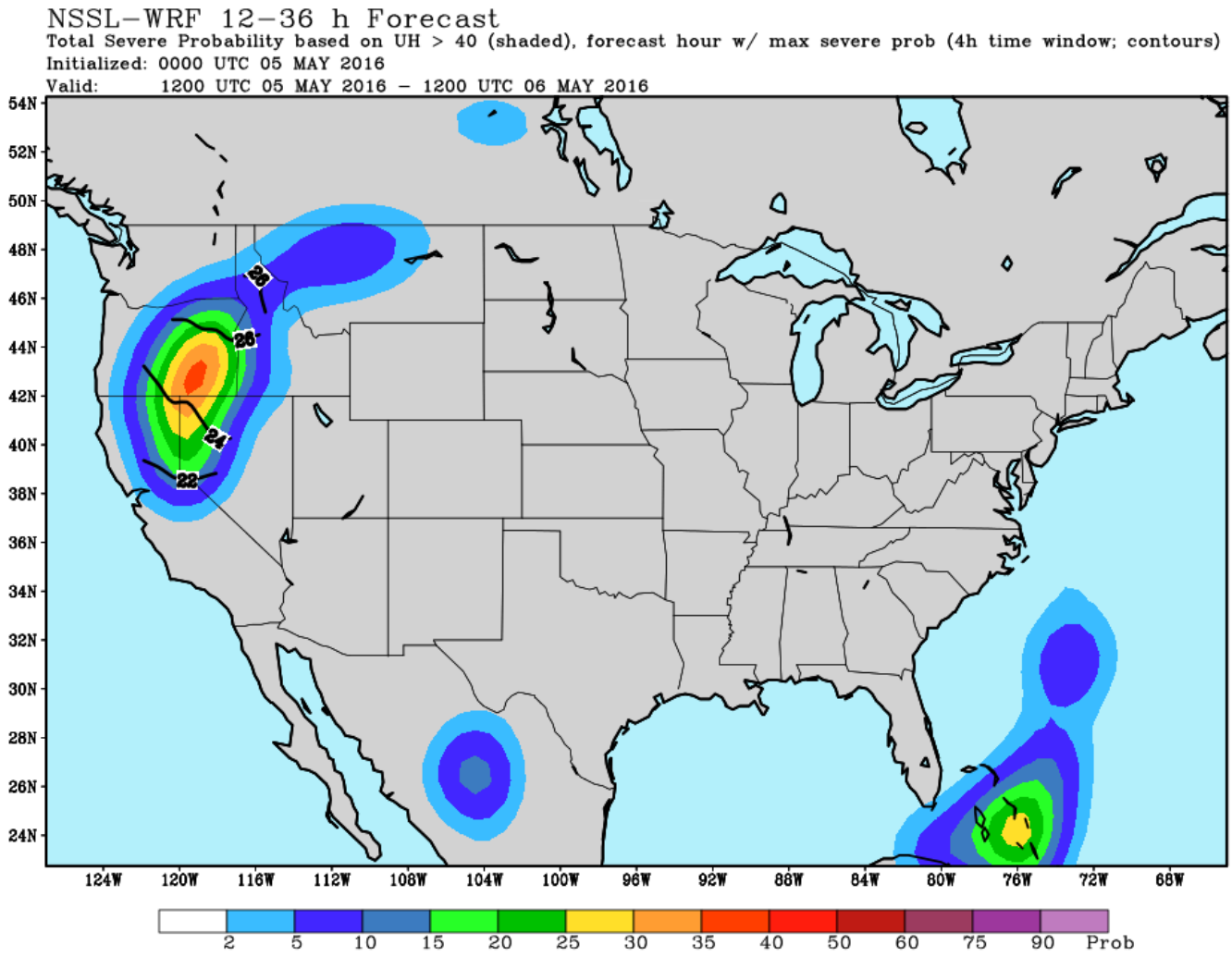

Figure 10 Example of automated isochrone forecast generated from the NSSL-WRF ensemble for the forecast initialized 0000 UTC 5 May 16 and valid over the 1200-1200 UTC time frame (forecast hours 12-36). Contours indicate the start times of the 4-h time windows with highest severe weather probabilities, while the shading indicates the total probability of severe weather over the 1200 – 1200 UTC period.

b) Evaluation of experimental forecast products – Severe Hazards Team

The severe hazards team conducted daily evaluations comparing temporally disaggregated 4-h first guess guidance for severe hazards (i.e., tornado, hail, and wind) to calibrated guidance generated from the SREF and SSEO (Jirak et al. 2014). The first-guess probabilities for the 4-h periods were generated using the temporal disaggregation technique (Jirak et al. 2012) by incorporating the full-period hazard outlook to constrain and scale the magnitude and spatial extent of the 4-h calibrated probabilities. This comparison provides an indication of how incorporating the human full-period outlook can improve upon the 4-h calibrated model guidance. An example forecast for tornadoes is shown in Figure 11.

During the 1800-2200 UTC period, the disaggregated first-guess guidance was rated similarly to slightly better than the calibrated guidance (Fig. 12; rating around 0 on a scale of -3 to +3). For the 2200-0200 UTC period, however, the disaggregated first-guess guidance was generally rated as an improvement over the calibrated model guidance.

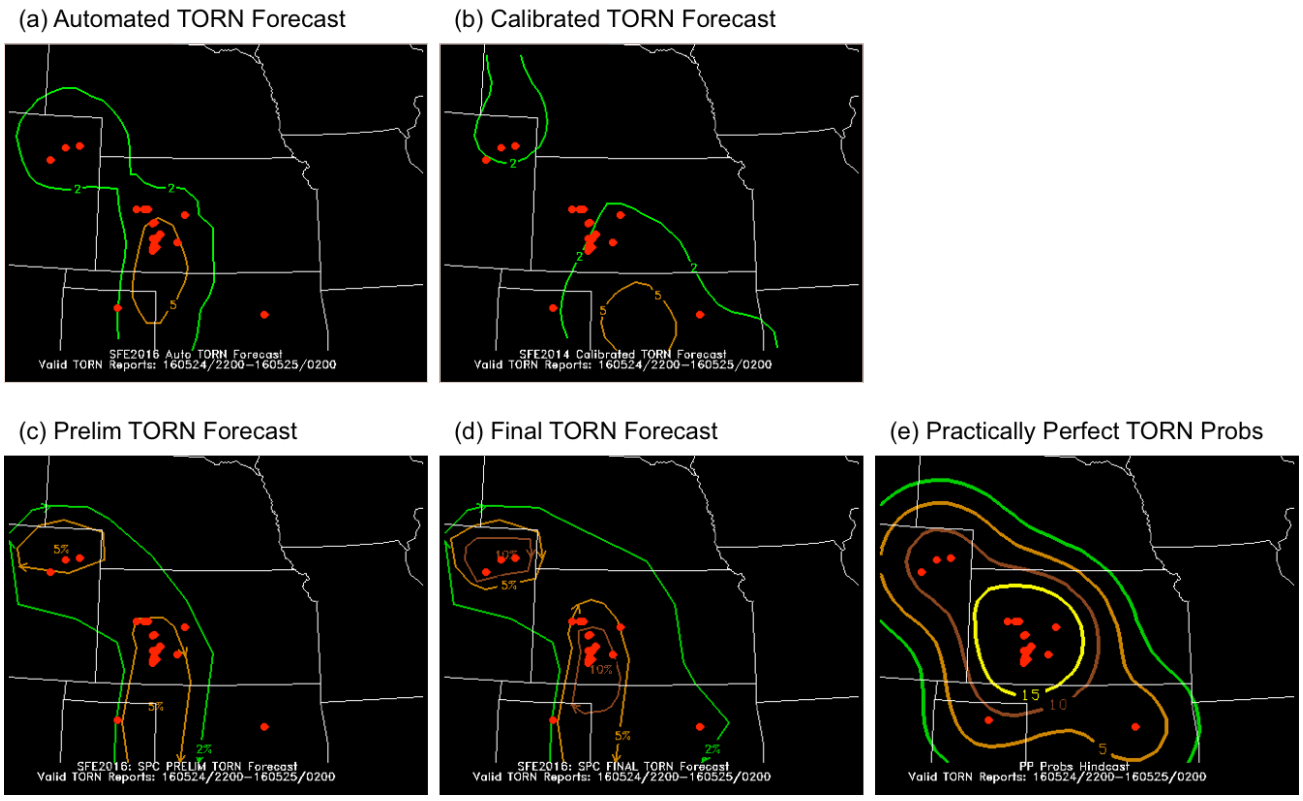

Figure 11 Example probabilistic tornado forecasts issued from the severe hazards desk on 24 May 2016 and valid for the 2200-0200 UTC time period with tornado reports overlaid. (a) Automated forecast using temporal disaggregation, (b) calibrated forecast using SREF and SSEO, (c) preliminary forecast issued by the severe hazards team in the morning, and (d) the final forecast issued in the afternoon. (e) Practically perfect probabilities derived from the distribution of observed tornadoes.

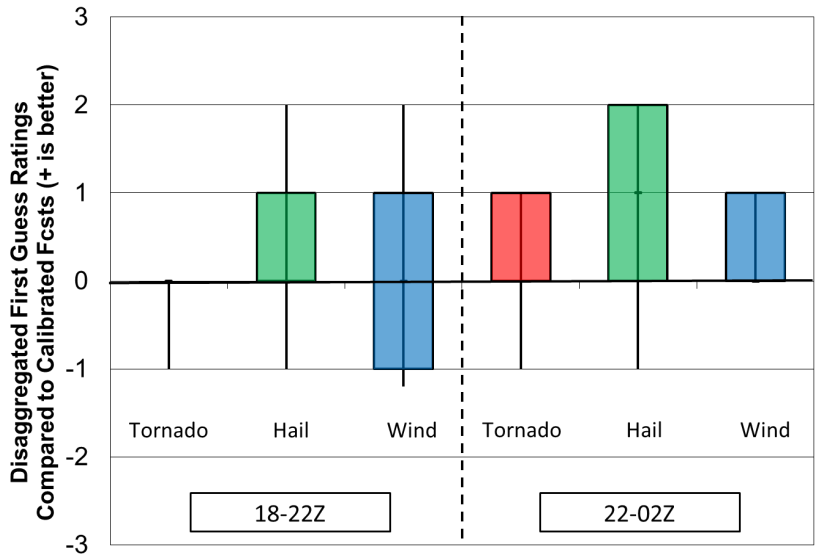

Figure 12 Box plots showing the distributions of subjective ratings (-3 to +3) of the temporally disaggregated first-guess guidance compared to the calibrated guidance for tornado (red), hail (green), and wind (blue) during the 1800-2200 UTC (left) and 2200-0200 UTC (right) periods.

The preliminary and final tornado, wind, and hail forecasts for the 2200-0200 UTC period were subjectively compared to determine the relative value of the afternoon forecast updates (Fig. 13). Overall, updating the forecasts in the afternoon generally resulted in similar or better forecast quality. Although the improvement was marginal (i.e. typically 0 to +1 rating) and often provided later confirmation of the existing threat, it was rare for the afternoon updates to result in degraded forecast quality.

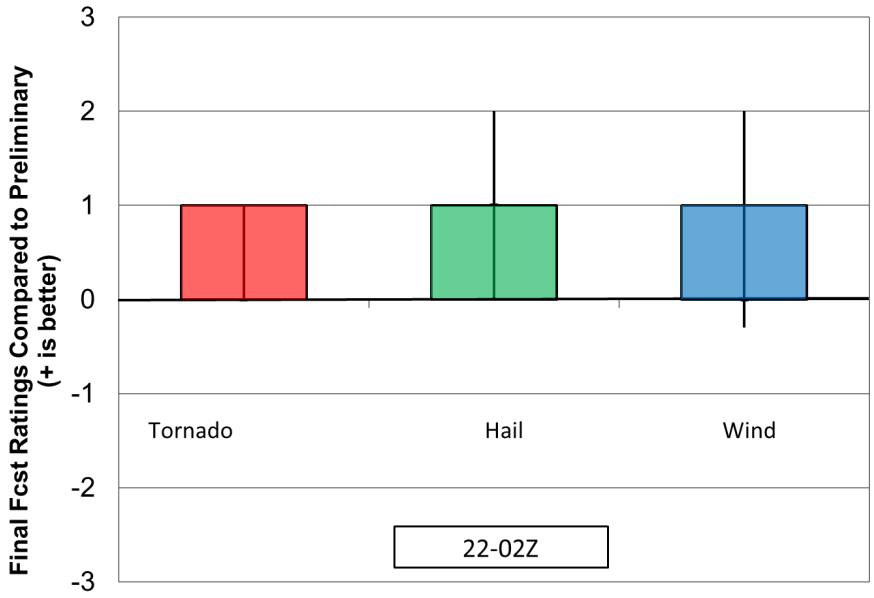

Figure 13. As in Fig. 12, except for the distribution of subjective ratings (-3 to +3) for the final forecast compared to the preliminary forecast for tornado (red), hail (green), and wind (blue) during the 2200-0200 UTC period.

c) Model Evaluations – Total Severe Desk

1) ARW VS. NMMB

Participants were asked to compare two members of the CLUE that used the same ICs/LBCs, but had different dynamic cores; specifically, NMMB and ARW. During the evaluations, participants assigned separate subjective ratings to reflectivity, hourly maximum fields such as updraft helicity, and thermodynamic fields such as temperature. Subjective ratings were taken from 9 May onward because of the previously mentioned bug in the NMMB members run by CAPS. An example 24 h forecast and corresponding observations of reflectivity for forecasts initialized 23 May 2016 is shown in Figure 14.

Ratings of simulated reflectivity had roughly the same distribution characteristics (Fig. 14), however, a closer look at these distributions reveals slightly different shapes, with a generally broader distribution for the NMMB core (Fig. 15b and c). However, both NMMB and ARW cores performed similarly for reflectivity, most often receiving ratings of 5/10. The hourly max fields tended to be rated more highly with the ARW member, though, the median rating was the same. The thermodynamic fields, however, were rated much more highly in the ARW. These variables included temperature, dew point, and surface-based CAPE. Participants' comments revealed that the members tended to trend in similar ways. For example, one participant said for 16 May 2016, *"Both the ARW and NMMB were too slow with the front and both models do not dry out the air mass behind the front fast enough. NMMB did better with the instability axis into MO than the ARW".*

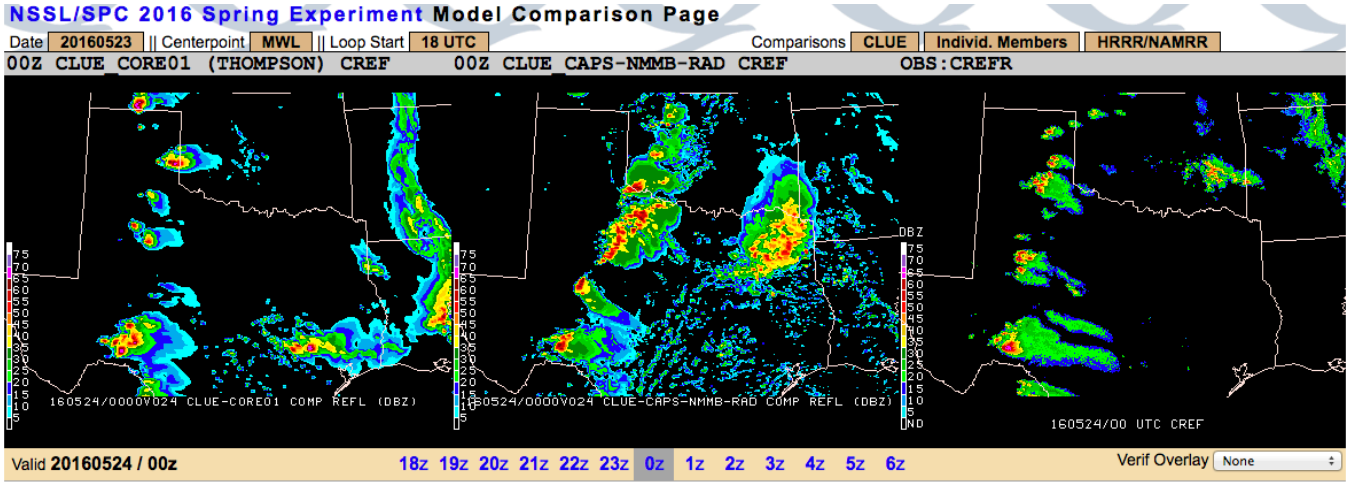

Figure 14 (left) Simulated reflectivity at 24 h forecast lead-time from the ARW member, (middle) same as left, except for NMMB, and (right) corresponding observations.

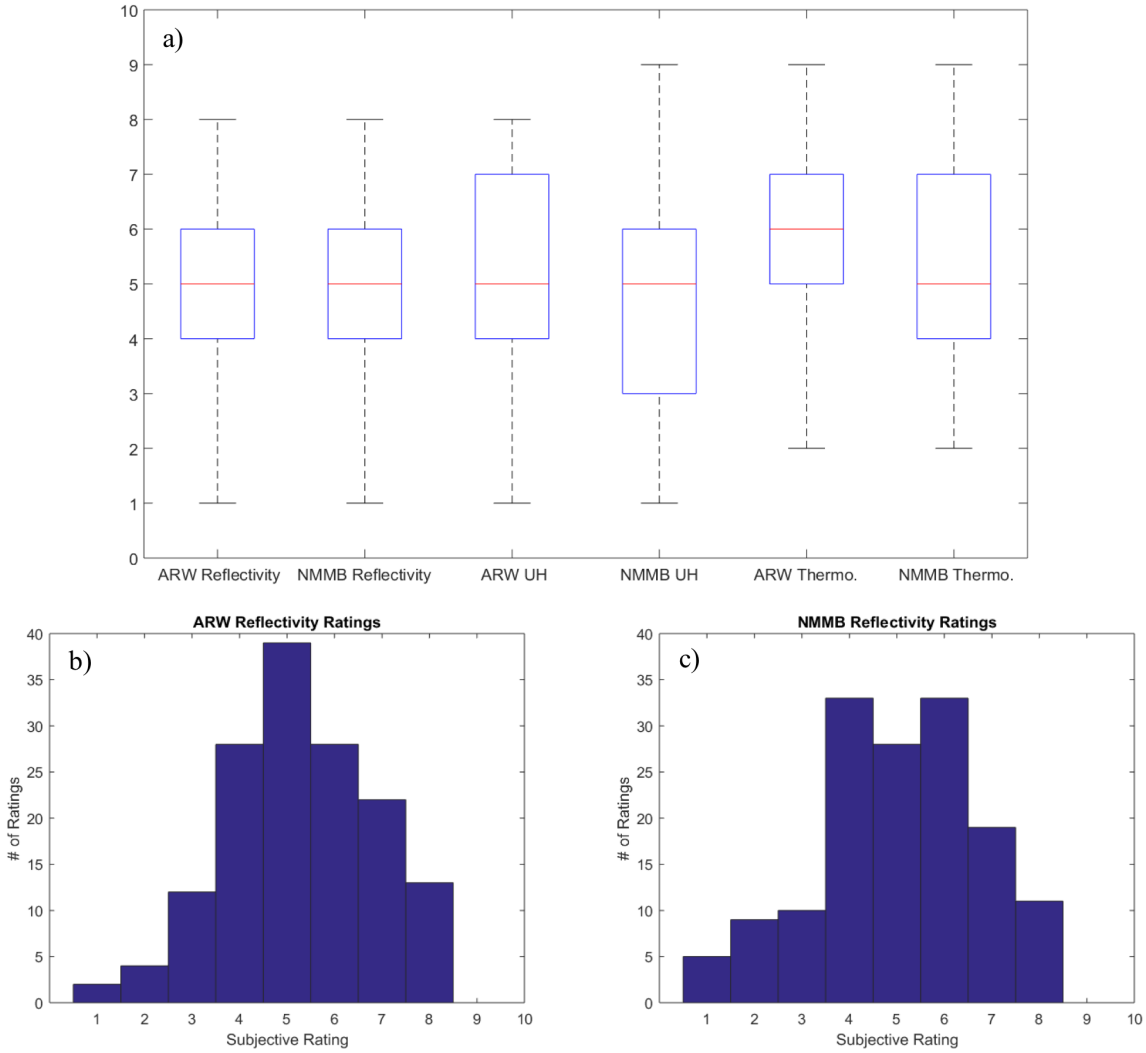

Figure 15 (a) Box plots for three categories of subjective ratings for ARW and NMMB simulations. The red line indicates the median. (b) Distribution of subjective ratings for the ARW on a scale of 1-10. (c) Same as (b) except for NMMB.

This type of comparison, where the participant noted that both were off, but one was off more so than others, was fairly typical and often accompanied numerical ratings that were identical, as participants did not consider the differences to be significant enough to warrant different ratings.

2) SINGLE CORE VS MULTI-CORE

Three ensembles were compared to test the effectiveness of a single core vs. multi-core configuration. The first ensemble consisted of 5 ARW and 5 NMMB members, the second was comprised of 10 ARW members, and the third consisted of 10 NMMB members. Comparisons were made using probabilities of reflectivity greater than 40 DbZ, maximum from any member UH, and probabilities of UH >25 and >100 m^2s^{-2}. An example comparison is shown in Figure 16.

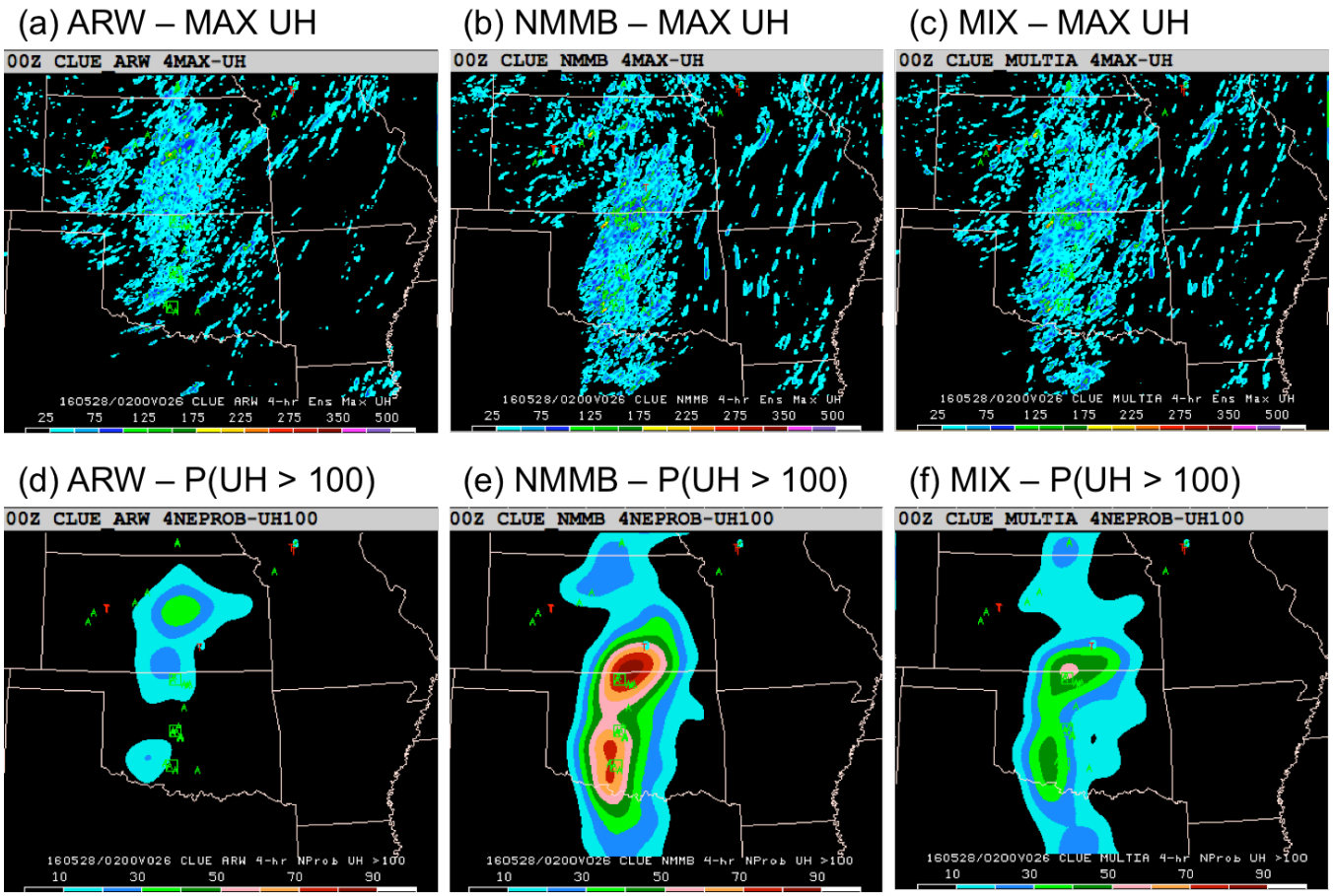

Figure 16 Example forecast imagery for the single core vs. multi-core ensemble comparisons for forecasts initialized 0000 UTC 27 May 2016. Ensemble maximum from any member UH for the 2200-0200 UTC period with storm reports overlaid for ensembles comprised of (a) ARW, (b) NMMB, and (c) a mix of NMMB and ARW members. Probabilities of UH > 100 with storm reports overlaid derived from ensembles comprised of (a) ARW, (b) NMMB, and (c) a mix of NMMB and ARW members.

From the subjective ratings, the ARW and multi-core ensemble tended to receive higher ratings than the NMMB ensemble for hourly maximum fields. For reflectivity, all ensembles were rated similarly, and they did not receive as high ratings as the hourly maximum fields. The ARW had a larger IQR in the reflectivity than the other ensembles (Fig. 17). Regarding the hourly maximum fields, the largest difference between the subjective ratings of the ARW and NMMB ensembles on any given day was three points, with the ARW outperforming the

NMMB by three points eight times in the ratings and the NMMB besting the ARW by three points twice. The differences between the single core ensembles was larger than the difference between either of the single-core ensembles and the multi-core ensemble: the largest difference with the ARW ensemble occurred when one participant rated the ARW ensemble three points higher than the multi-core ensemble, and the largest difference with the NMMB ensemble occurred when one participant rated the multi-core ensemble four points better than the NMMB ensemble. Besides these two occurrences, the difference between either of the single-core ensembles and the multi-core ensemble was two points or fewer.

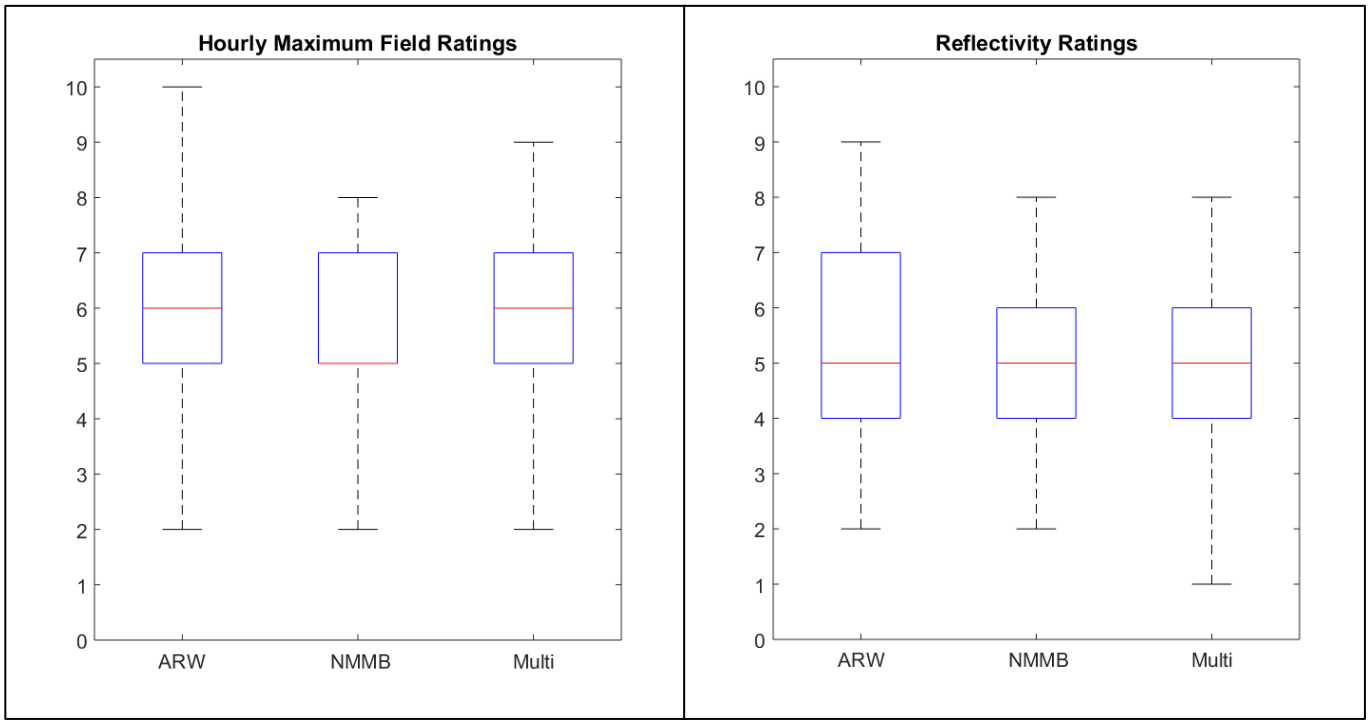

Figure 17 Box plots comparing the subjective ratings of (left) hourly maximum fields and (right) reflectivity between the single core ensembles and the multi-core ensemble. A red line indicates the median.

Objective comparisons of the single vs. multi-core ensemble strategies were also performed for severe weather and QPF forecasting. For severe weather, UH was used as a severe weather proxy to create Day 1 severe weather probabilities following the methods of Sobash et al. (2011, 2016), and then probabilistic skill metrics were used to verify against storm reports. One complication of comparing these surrogate severe probability forecasts (SSPFs) is that the distributions of UH in the ARW and NMMB cores is different. Thus, UH percentiles (as opposed to fixed thresholds) were computed after re-gridding UH to 80-km grids. The percentiles used were: 0.925, 0.95, 0.955, 0.96, 0.965, 0.97, 0.975, 0.98, 0.985, 0.99, 0.995, 0.9975, 0.9999, and 0.99999. Then, to create SSPFs, a range of smoothing parameters (i.e., the standard deviation of the Gaussian kernel) was used (9 total): 40, 60, 80, 100, 120, 140, 160, 180, and 200 km. A set of example SSPFs is shown in Figure 18.

One interesting result came from comparing the values of UH in the ARW and NMMB for the different UH percentiles (Fig. 19). After re-gridding the UH to the 80-km grid by finding the maximum UH from all the 3-km grid-points that fell within each 80-km box, the NMMB had higher UH at all percentiles. From a similar comparison, but using the raw UH from the 3-km grids, ARW had higher UH up to the 99.999 percentile, after which the NMMB had higher UH. This result implies that higher UH for the NMMB on the 80-km grid comes

from extremely small-scale, perhaps event grid-point scale, maxima. This difference may be explained by differences in how UH is computing within the NMMB and ARW cores, but this has yet to be confirmed.

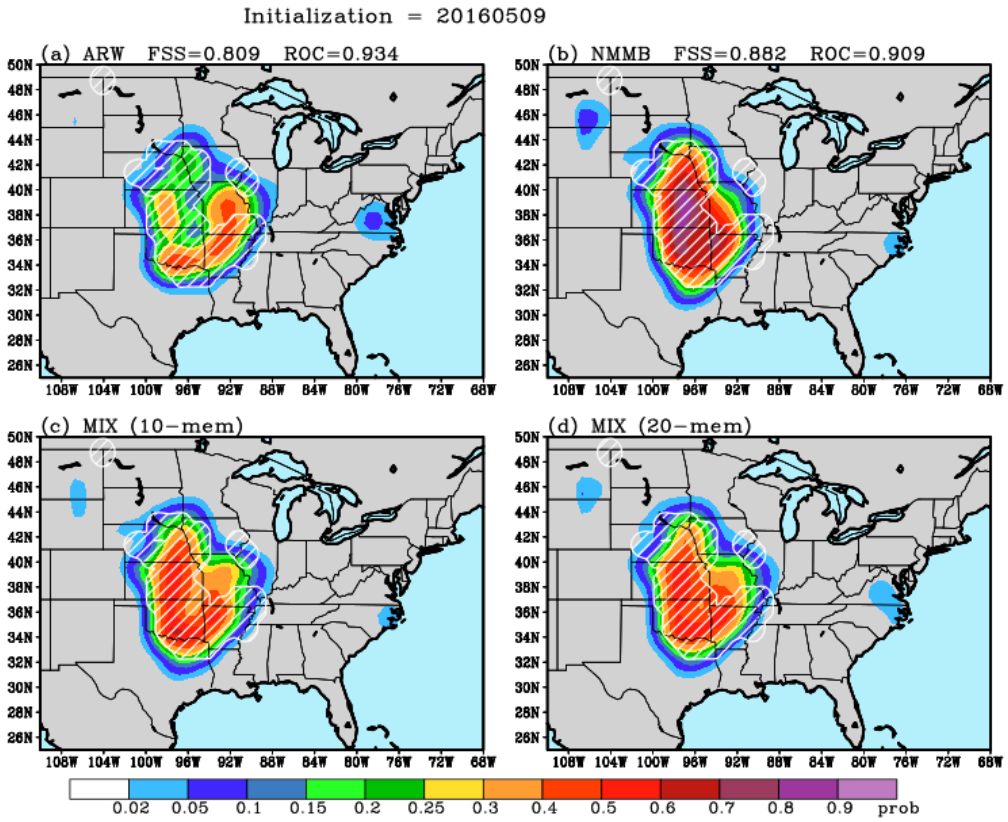

Figure 18 SSPFs (shaded) and the areas over which storm reports occurred (white hatching) for forecasts initialized 0000 UTC 9 May 2016 and valid 1200-1200 UTC (forecast hours 12-36) from ensembles using (a) 10 ARW members, (b) 10 NMMB members, (c) a mix of 10 ARW and NMMB members, and (d) a mix of 20 ARW and NMMB members. FSS and ROC areas are indicated at the top of each panel.

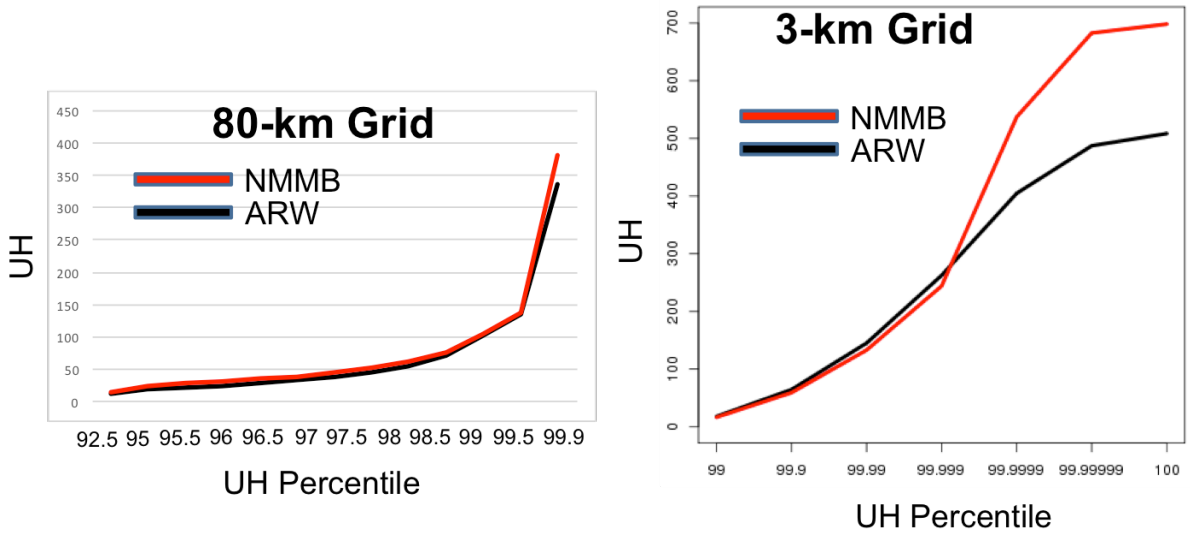

Figure 19 UH values as a function of percentile in the NMMB and ARW for the 80-km grid (left) and the 3-km grid (right).

To visualize ensemble forecast skill, ROC area and FSS are plotted as a function of UH percentile and smoothing parameter for each of the three ensembles in Figure 20. The highest ROC areas occur at smoothing parameters from 60 to 120 km and UH percentiles from 95% to 96%, which corresponds to UH values in the range 20-35 m^2s^{-2}. The ARW ensemble has slightly higher peak ROC areas than NMMB, but the mixed ensemble of 5 ARW and 5 NMMB members has the highest ROC areas. A similar plot, but for FSS, is shown in Figure 21. For FSS, scores are maximized at larger smoothing parameters and UH percentiles than the ROC areas. FSSs for the NMMB and ARW ensembles are very similar, but similar to the ROC areas, the mixed ensemble has the highest scores. For both FSS and ROC areas, an ensemble comprised of a mix of 10 ARW and 10 NMMB members had almost identical forecast skill as the 5 ARW and 5 NMMB members (not shown).

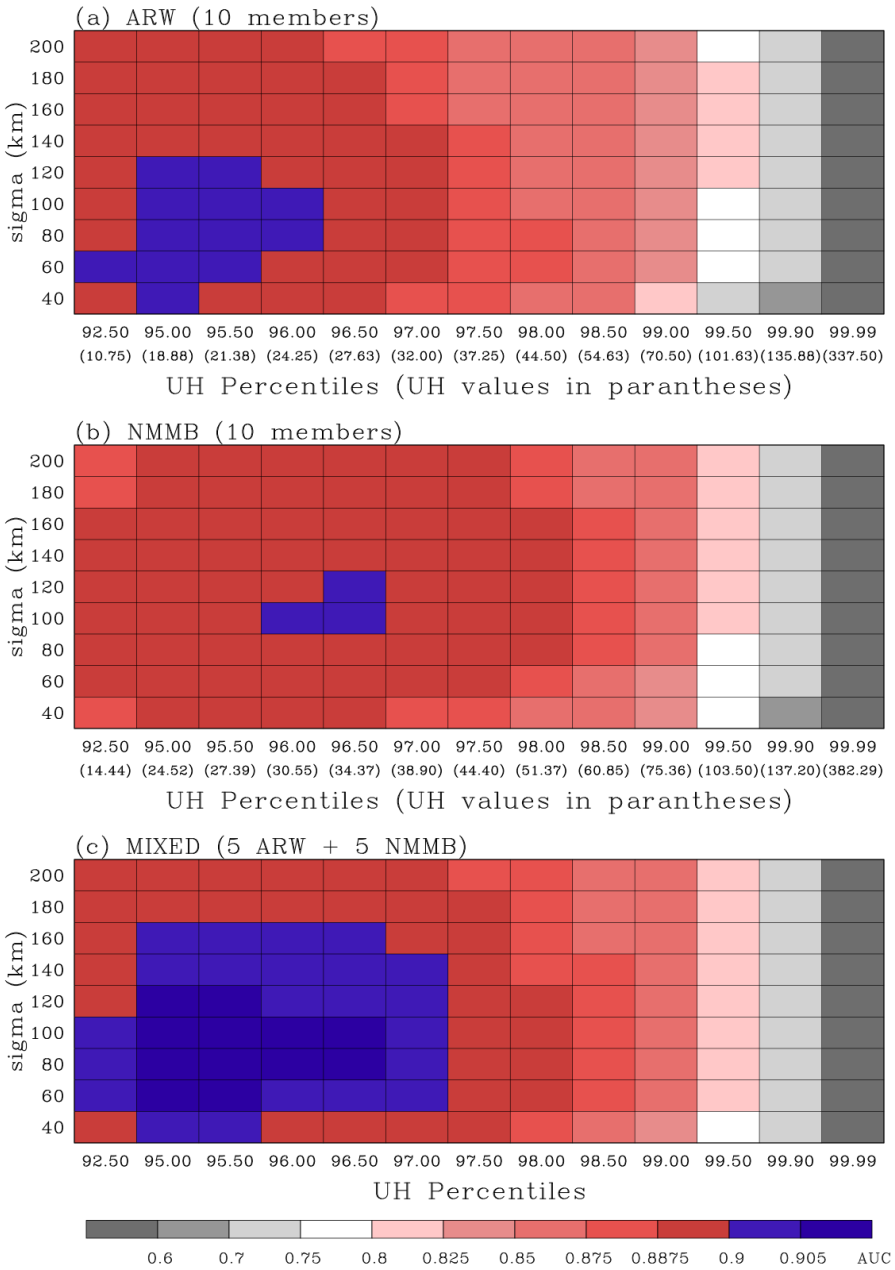

Figure 20 ROC areas as a function of smoothing parameter and UH percentile for ensemble comprised of (a) ARW members, (b) NMMB members, and (c) a mix of ARW and NMMB members.

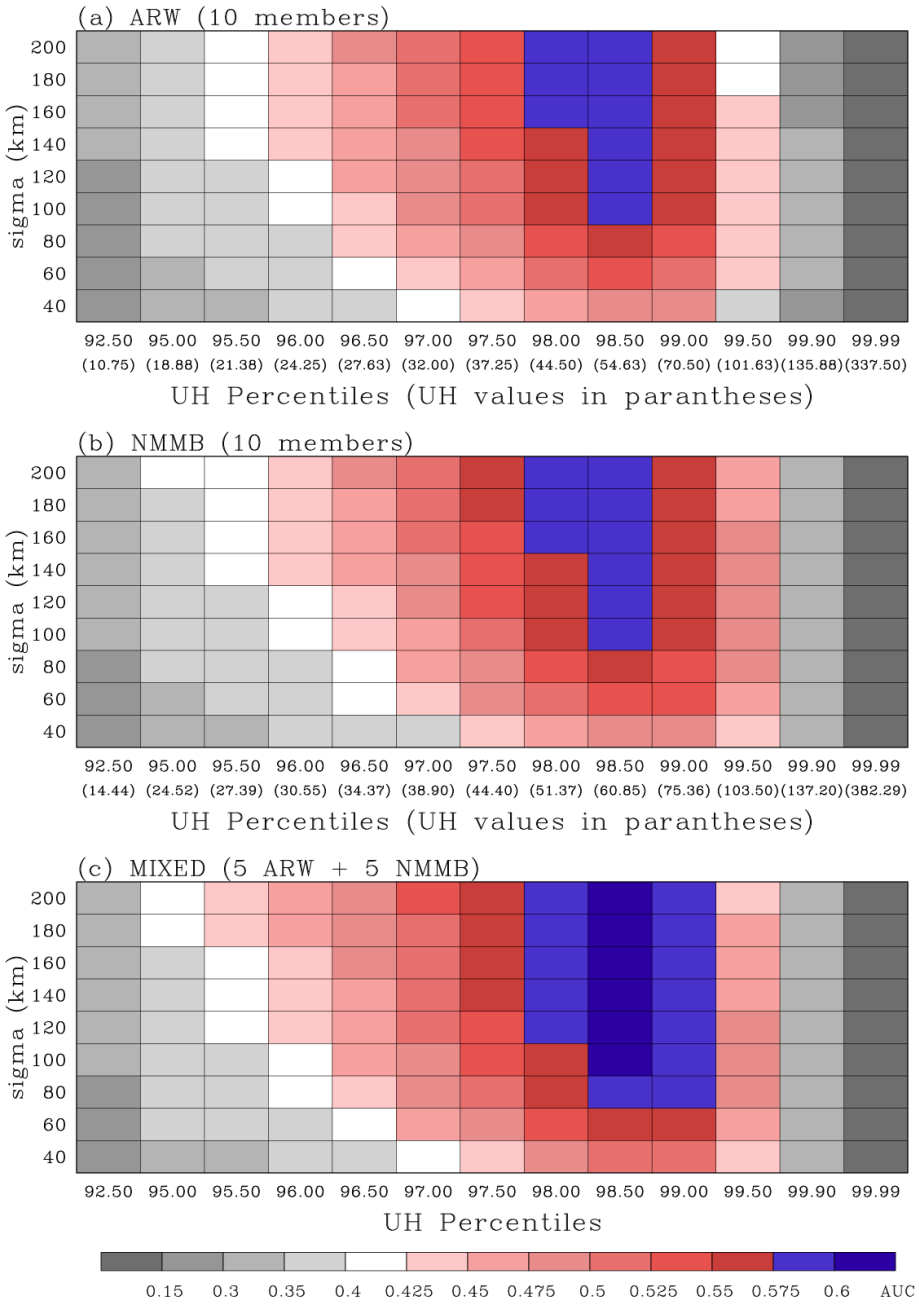

Figure 21 Same as Fig. 20, except for FSS.

To examine the reliability of the three ensembles, reliability diagrams for seven of the UH percentiles and five of the smoothing parameters are shown in Figure 22. In general, as the smoothing parameter and UH percentile increases, the observed relative frequencies increase. The best reliabilities occur with the UH percentile of 99% (UH = 70.5 m^2s^{-2} in the ARW ensemble) and smoothing parameters in the range 80 to 120 km. There is not a noticeable difference in reliability between any of the three ensembles. Note, the FSS scores maximize at smoothing parameters and UH percentiles very close to the best reliability. However, the ROC areas, which are generally not affected by reliability, maximize at smoothing parameters and UH percentiles at which there is strong over-forecasting.

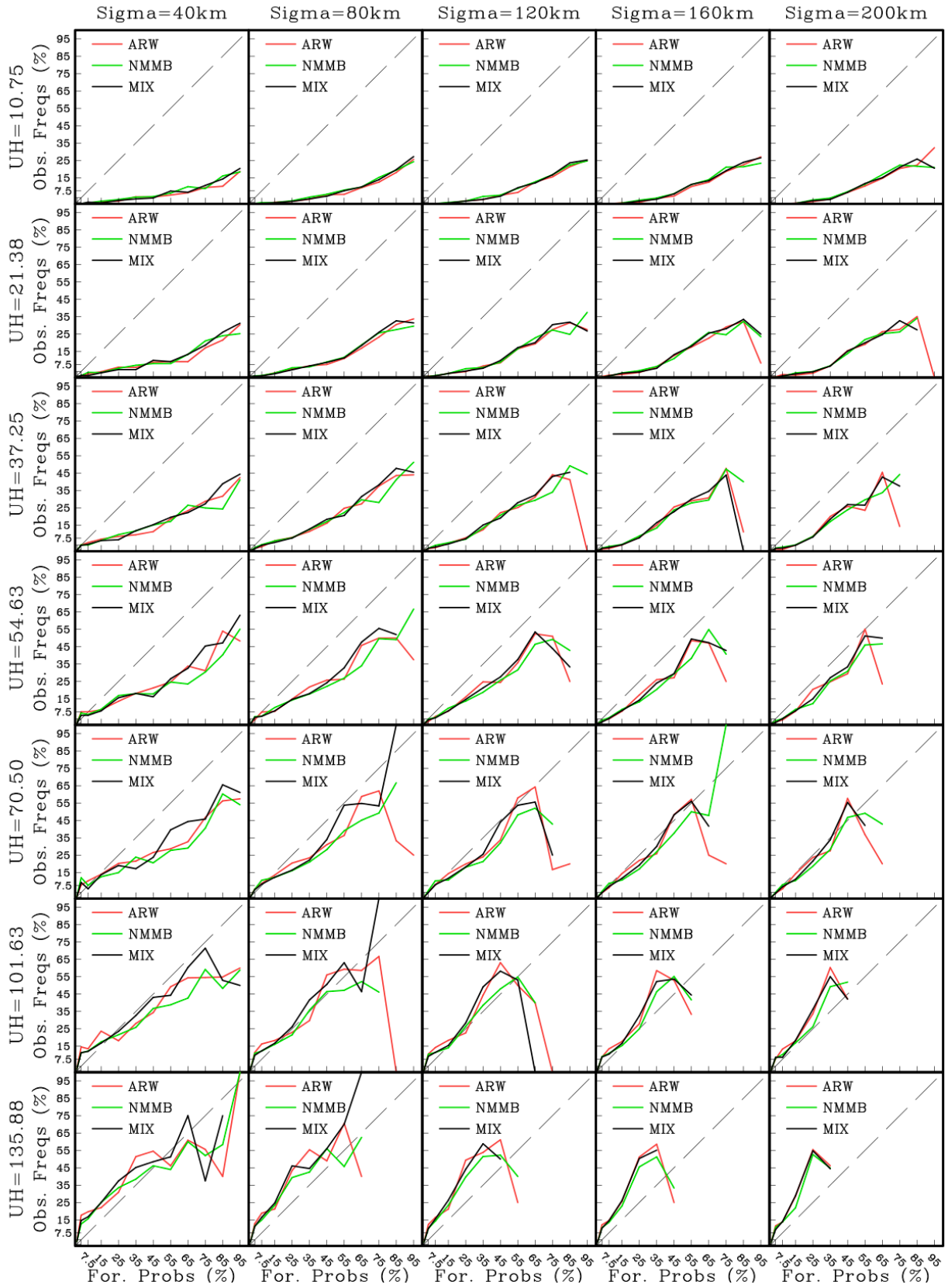

Figure 22 Reliability diagrams for various UH thresholds and sigmas (i.e., smoothing parameters) for SSFPs derived from ARW-only, NMMB-only, and a mix of NMMB and ARW members.

For objective QPF verification, a 2/3 CONUS domain (east of Rockies) is used with a land mask. NCEP's Stage IV precipitation estimates are used as observations. The Stage IV data and forecasts are both re-gridded to a common 4-km grid and bias, equitable threat score (ETS), and ROC areas are calculated. Additionally, time-longitude plots of diurnally averaged precipitation are constructed to examine the depiction of the diurnal cycle averaged over all 24 days of SFE2016. The time-longitude plots for ARW and NMMB members are shown in Figures 23 and 24, respectively. Clearly both sets of members capture the main features, (i.e., propagating axis of rainfall at night, and non-propagating area during afternoon), but there is obvious over-prediction in the NMMB, especially for the first 12 h of the forecast in the axis of propagating rainfall. To quantify how well each set of members depicts the diurnal cycle, spatial correlations in time-longitude space between each member and observations were computed and the distributions of correlations over forecast hours 3-36 and 18-36 are shown in Figure 25. For both of these periods the ARW members have a clear advantage over the NMMB members.

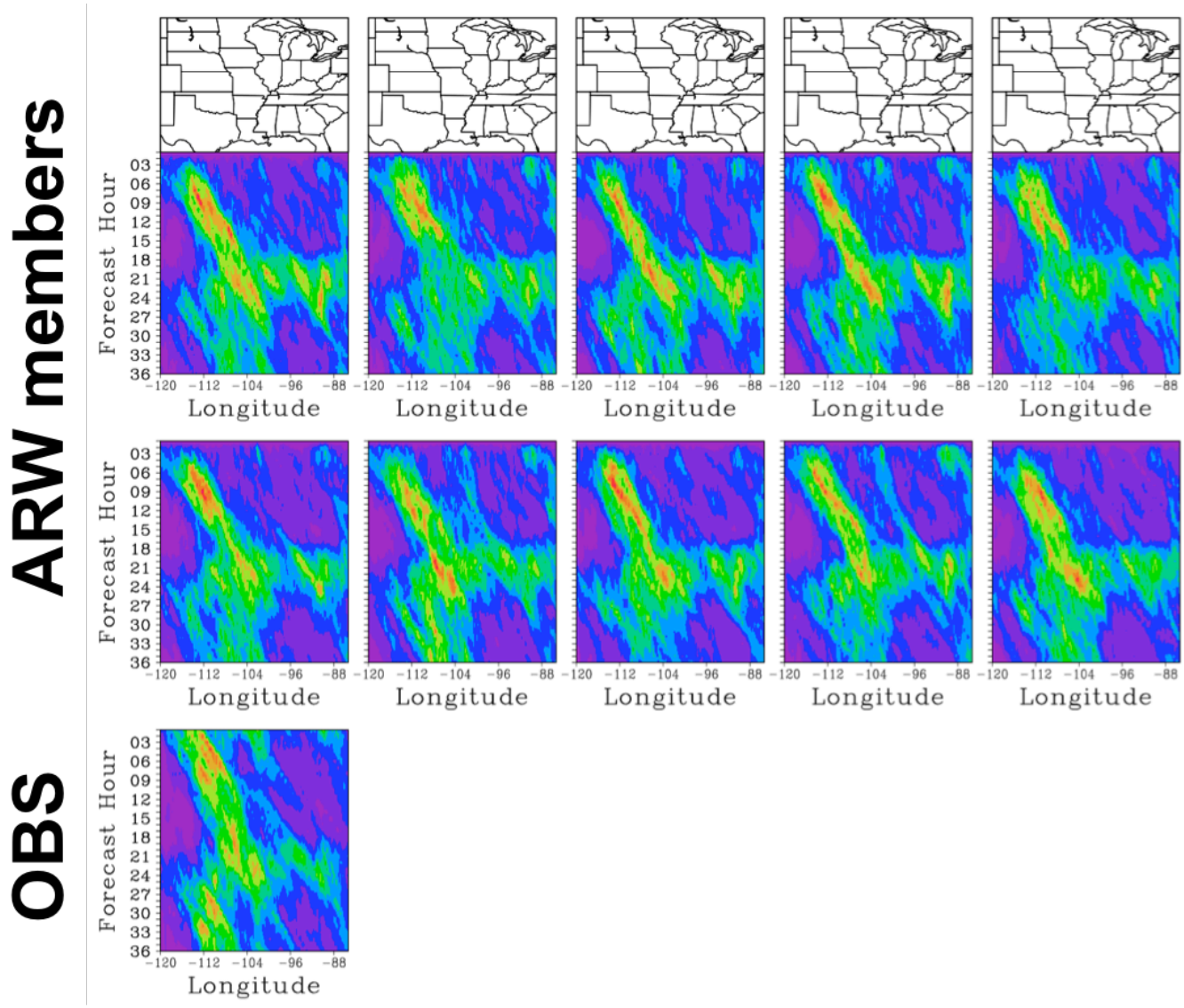

Figure 23 Time-longitude diagram of precipitation averaged over each forecast hour for all 24 days of SFE2016 for all ARW members (top two rows) and observations (bottom left).

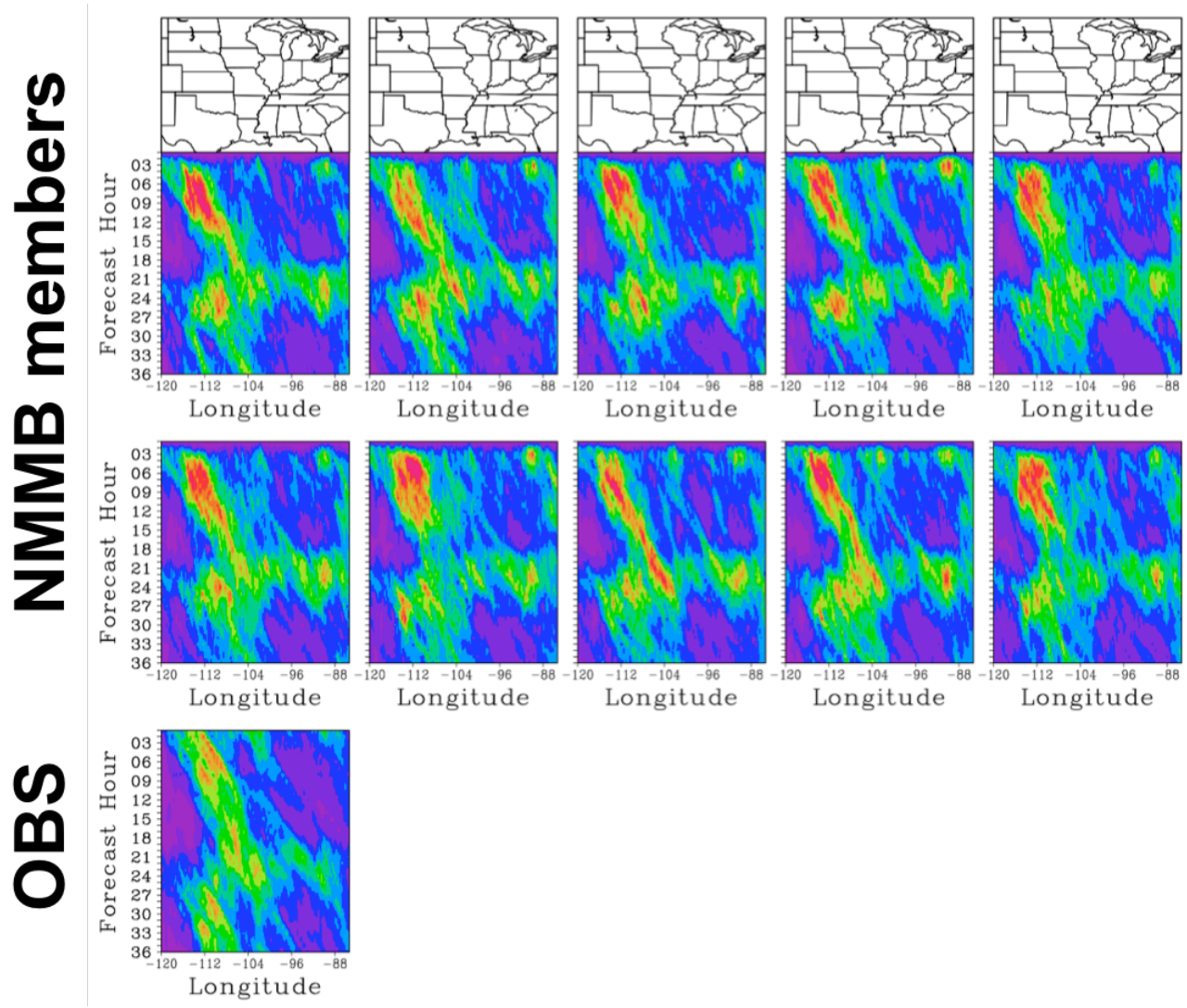

Figure 24 Same as Fig. 23, except for NMMB members.

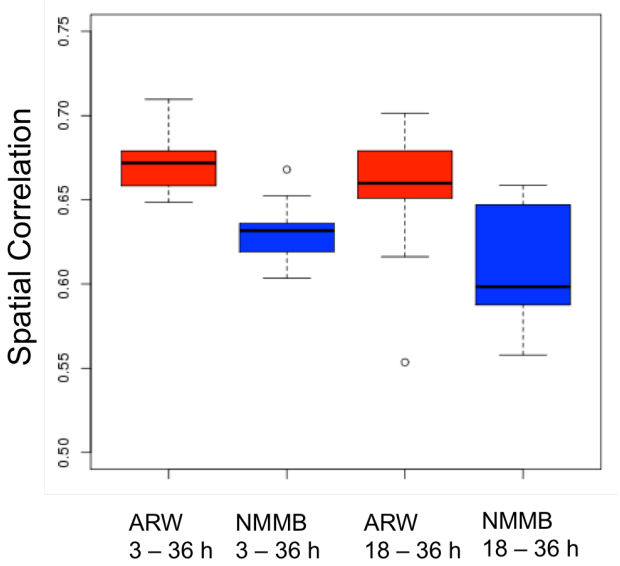

Figure 25 Distribution of spatial correlation coefficients of precipitation in time-longitude space for ARW and NMMB members during the forecast periods 3-36 h and 18-36 h.

Bias and ETS for 24 h accumulated precipitation (valid 1200-1200 UTC or forecast hours 12-36) was calculated for seven thresholds ranging from 0.10 to 4.00-in (Fig. 26). The bias for the ARW members is very close to 1.0 (perfect bias) at all rainfall thresholds. The NMMB members have slight under-prediction at low thresholds and over-prediction at high thresholds. For 24 h ETS (Fig. 27), the ARW members have a clear advantage at thresholds up to 1.0-in. At thresholds above 1.0-in, all scores are very low and similar to one another.

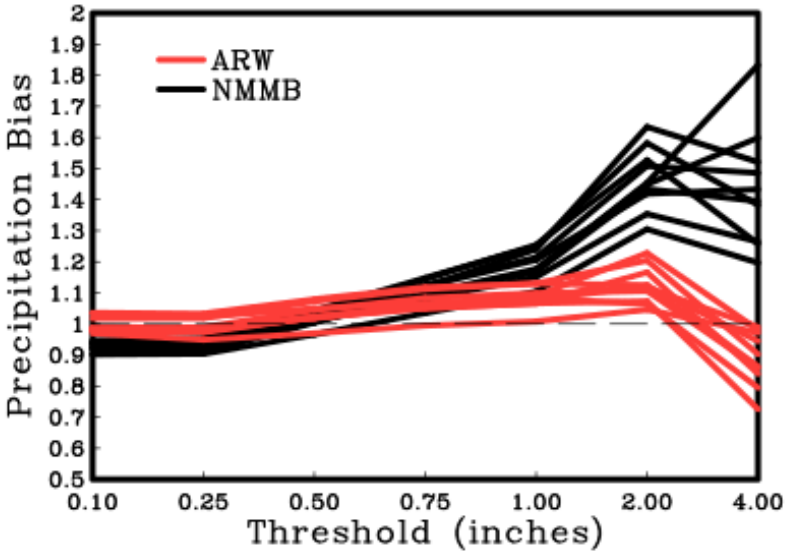

Figure 26 Bias as a function of threshold for 24 h accumulated precipitation from the ARW (red) and NMMB (black) members.

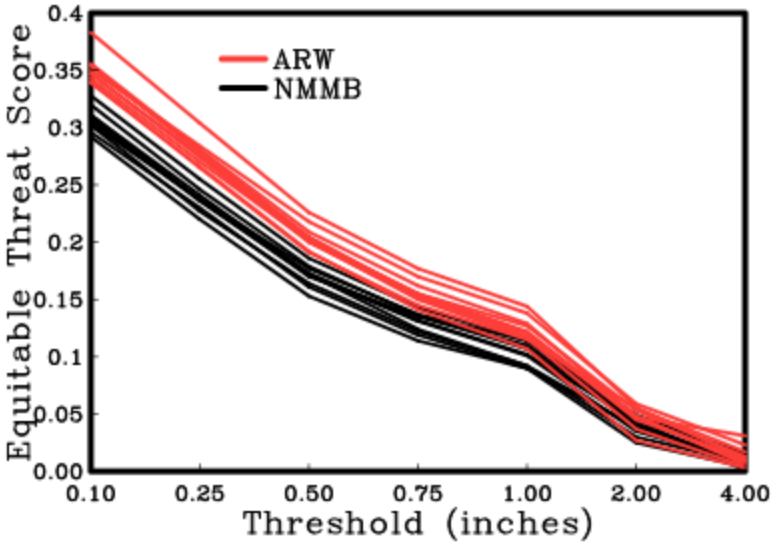

Figure 27 Same as Fig. 25, except for equitable threat score (ETS).

To verify probabilistic precipitation forecasts (PQPF), probabilities were computed on the 4-km grid by finding where the rainfall threshold fell within the distribution of ensemble members. For points at which the rainfall threshold was greater than the amounts from all ensemble members, a Gumbel distribution (for extreme values) was used to estimate the probabilities (e.g., Clark et al. 2009). In addition, the probabilities were

smoothed using a Gaussian kernel with smoothing parameters of 10, 25, and 50 km. The ROC areas for PQPF as a function of threshold for the three sets of ensembles are shown in Figure 28. From this plot, it is clear that increasing the smoothing helps the most of the highest thresholds. Also, the ARW ensemble performs best, even slightly better than the mixed ensemble.

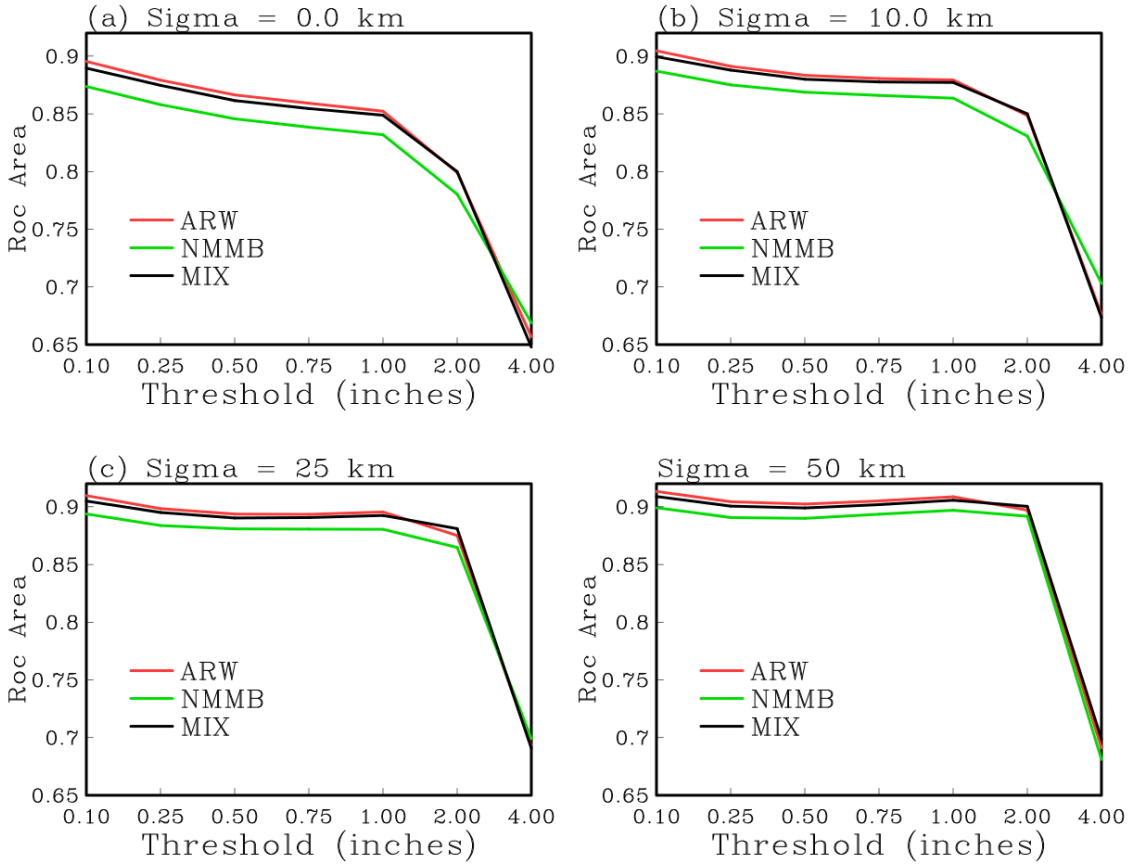

Figure 28 ROC area as a function of precipitation threshold for ensembles comprised of ARW, NMMB, and a mix of ARW and NMMB members using smoothing parameters (sigma) of (a) 0 km, (b) 10 km, (c) 25 km, and (d) 50 km.

3) HAIL SIZE FORECASTS

Three methods of hail size forecasting were tested in SFE2016: (1) HAILCAST, an algorithm that predicts maximum hail size using a hila growth model coupled to WRF, (2) the Thompson method, which is based directly on microphysical parameterization information such as the graupel size distributions, and (3) the Gagne method, which is based on a machine-learning algorithm. More details on each method can be found in the operations plan (Clark et al. 2016). Probabilities were derived from a single ensemble using each method. Both the 1-in and 2-in hail size thresholds were examined. Verification was performed with hail reports and MESH.

Participants rated HAILCAST the highest in forecasting both the probability of hail greater than 1- and 2-in. (Fig. 29). The Gagne machine-learning method also performed well when forecasting hail greater than 1-in. The Thompson method was generally rated the lowest of the three methods. All three methods obtained a broad variety of ratings throughout the experiment, particularly for the probability of hail greater than 2-in as can be seen in the wide IQRs (Fig. 29). HAILCAST was noted to have generally higher coverage of 1-in hail

probabilities compared to the other two methods, and the Thompson method was noted by participants to have less spatial coverage than the other methods. This worked to the Thompson method's favor on days with less hail, while hindering it on days with more hail. The Gagne method overall received favorable comments, with a seemingly even split in participant comments as to whether it was too high or too low in its probabilities at both size thresholds. Participant comments note the Thompson method's tendency to produce a lot of hail greater than 2-in, particularly early in the experiment.

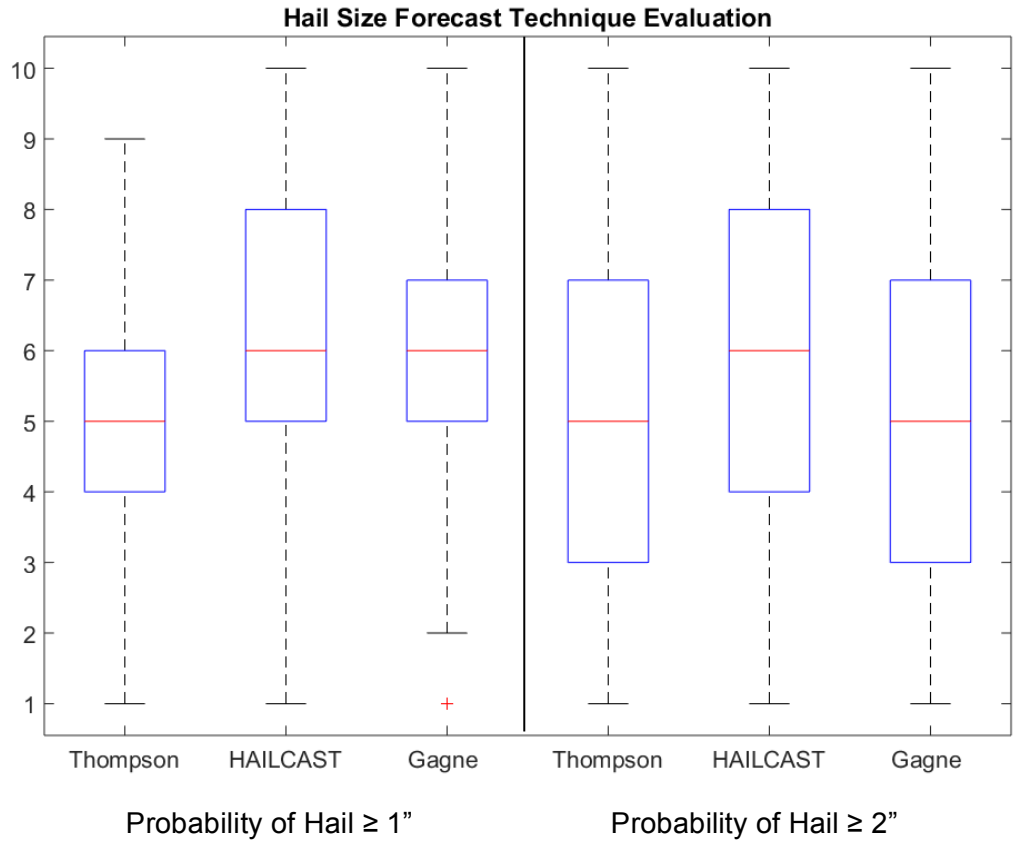

Figure 29 Box plots of subjective ratings for the three hail size forecasting methods examined in SFE2016. The median of each distribution is indicated by a red line.

4) MICROPHYSICS COMPARISONS

The microphysics evaluation was an open-ended question prompting the participants to comment on differences in reflectivity, updraft helicity, temperature, dew point, and SBCAPE for five members of CLUE that were identically configured except for their microphysics parameterizations. Parameterization schemes tested were: Morrison, Thompson, P3, WSM6, and M-Y. An example comparison is shown in Figure 30. In general, participants thought that all microphysics members were quite similar, and details of which scheme performed best varied throughout the day. Cold pool strengths were also similar between the schemes. P3 tended to develop higher updraft speeds than other members, which occasionally caused difficulties in completing the P3 because of numerical instability issues.

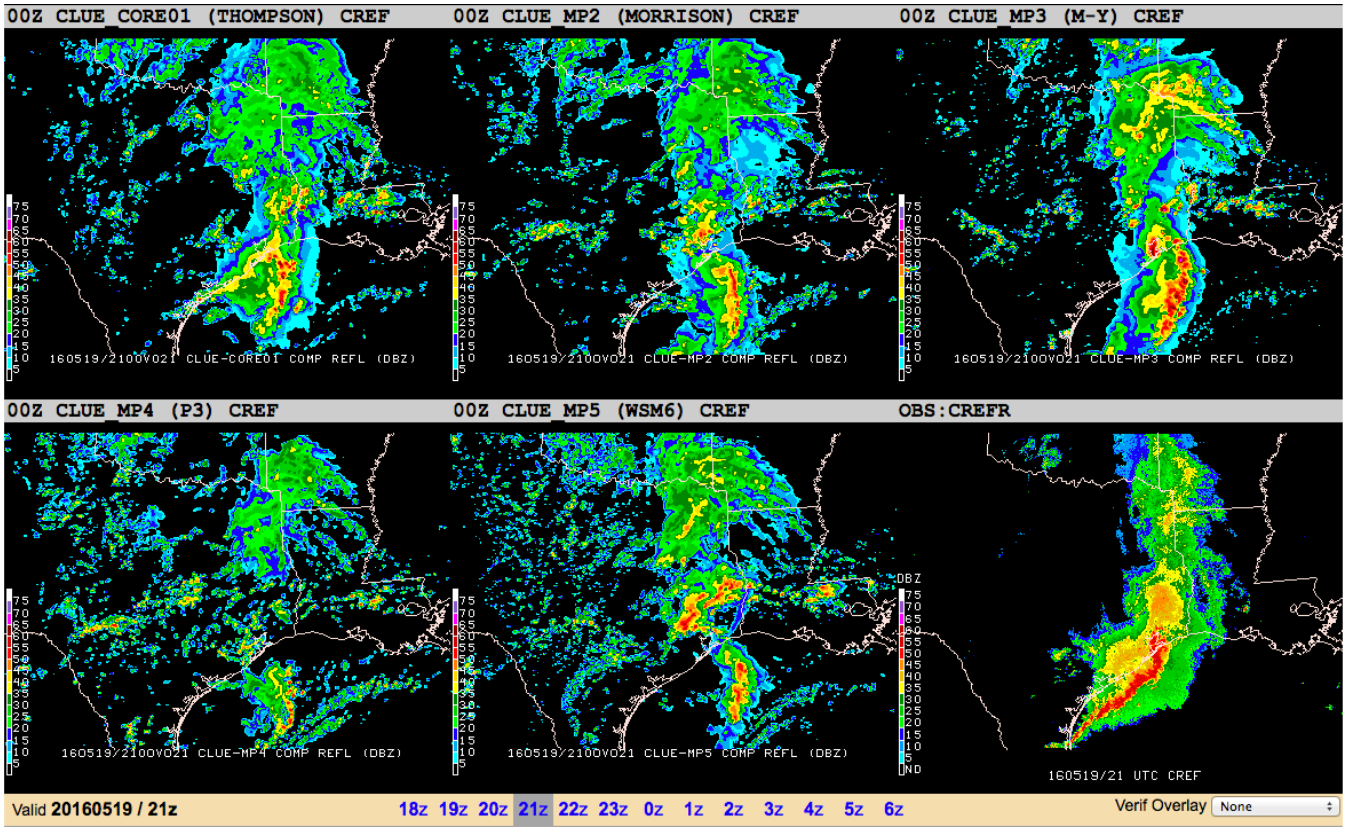

Figure 30 Forecasts of simulated composite reflectivity valid 2100 UTC 19 May 2016 from CLUE members with different microphysics parameterizations (scheme indicated by label on top of each panel). The corresponding observed composite reflectivity is in the bottom left panel.

5) Objective Verification of MPAS Forecasts

The NCAR Model for Prediction Across Scales (MPAS) was run for the second consecutive year during SFE2016. MPAS produced 5-day forecasts (120 hours) initialized daily at 0000 UTC using a mesh that varies in horizontal resolution from 15-km globally to 3-km over the CONUS. Probabilistic forecasts of severe weather were generated using the Surrogate Severe Probabilistic Forecast (SSPF) method detailed by Sobash et al. (2011, 2016). The forecasts were verified against observed storm reports to compute CSI and ROC curves for forecasts initiated May 1 through May 31 using 1200-1200 UTC windows on Days 1-4 for verification. An example probabilistic forecast is shown in Figure 31.

The SSPF methodology requires choosing a threshold of UH to compute the gridded probabilities. This value was determined following the methodology in Sobash et al. (2016) by choosing the threshold that produced a bias closest to 1 with respect to the gridded storm reports. For MPAS, this value was determined to be 175 m^2s^{-2} based on the 2016 dataset. Using this threshold, the performance diagrams and ROC curves were computed over the whole month for the Day 1 through Day 4 forecasts and are shown in Figures 32 and 33, respectively.

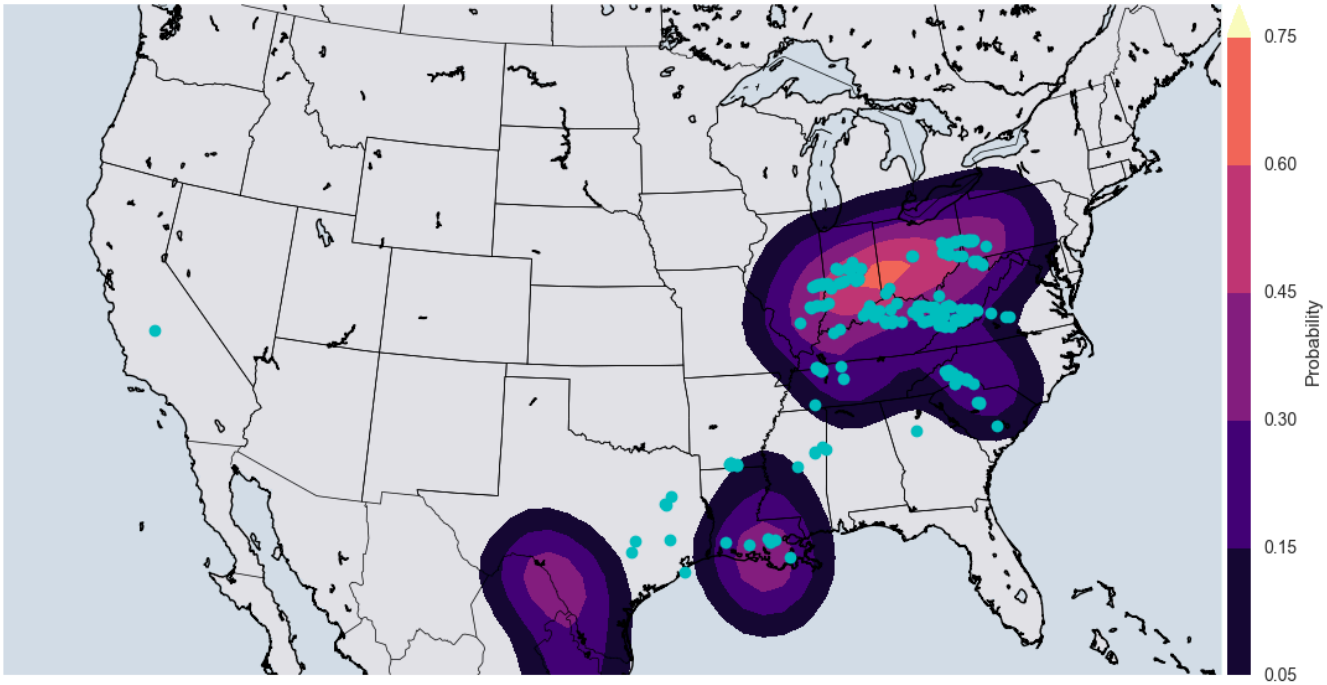

Figure 31 Day 1 Severe weather probabilities derived from a 1 May 2016 0000 UTC initialization of MPAS valid from 1200-1200 UTC 1-2 May (forecast hours 12-36) with storm reports indicated by blue dots.

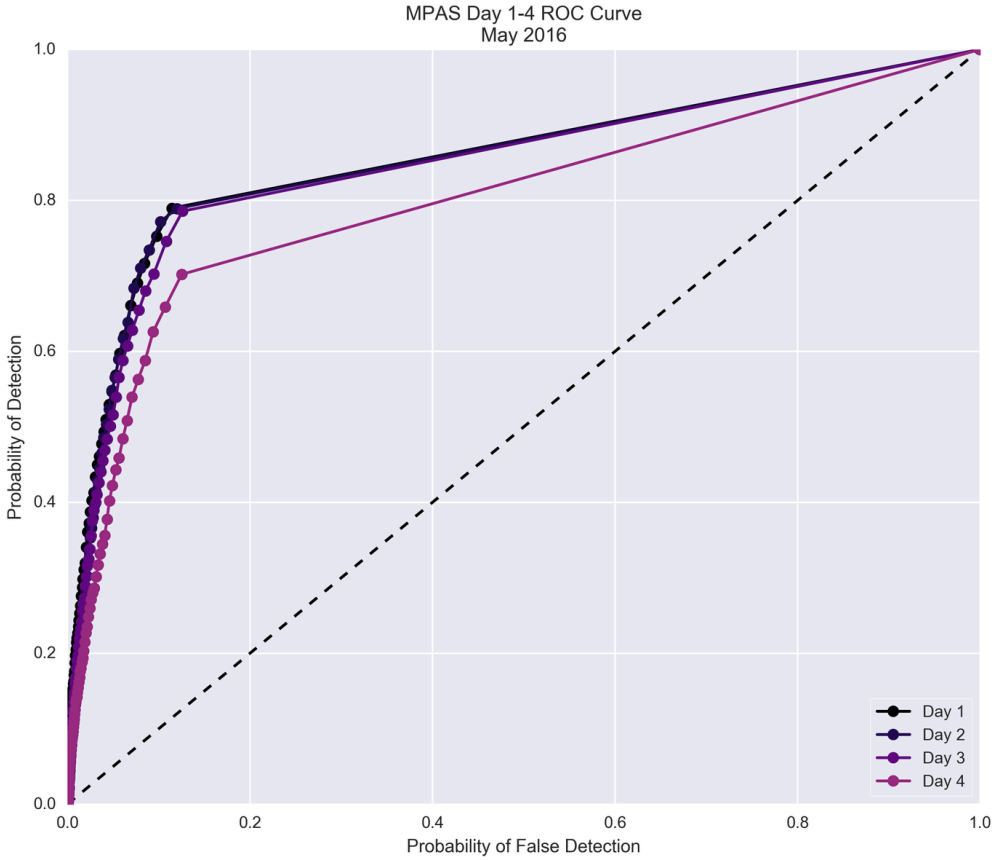

Figure 32 ROC curves for MPAS-derived probabilistic severe weather forecasts at Days 1-4 lead times for the 1-31 May 2016 time period.

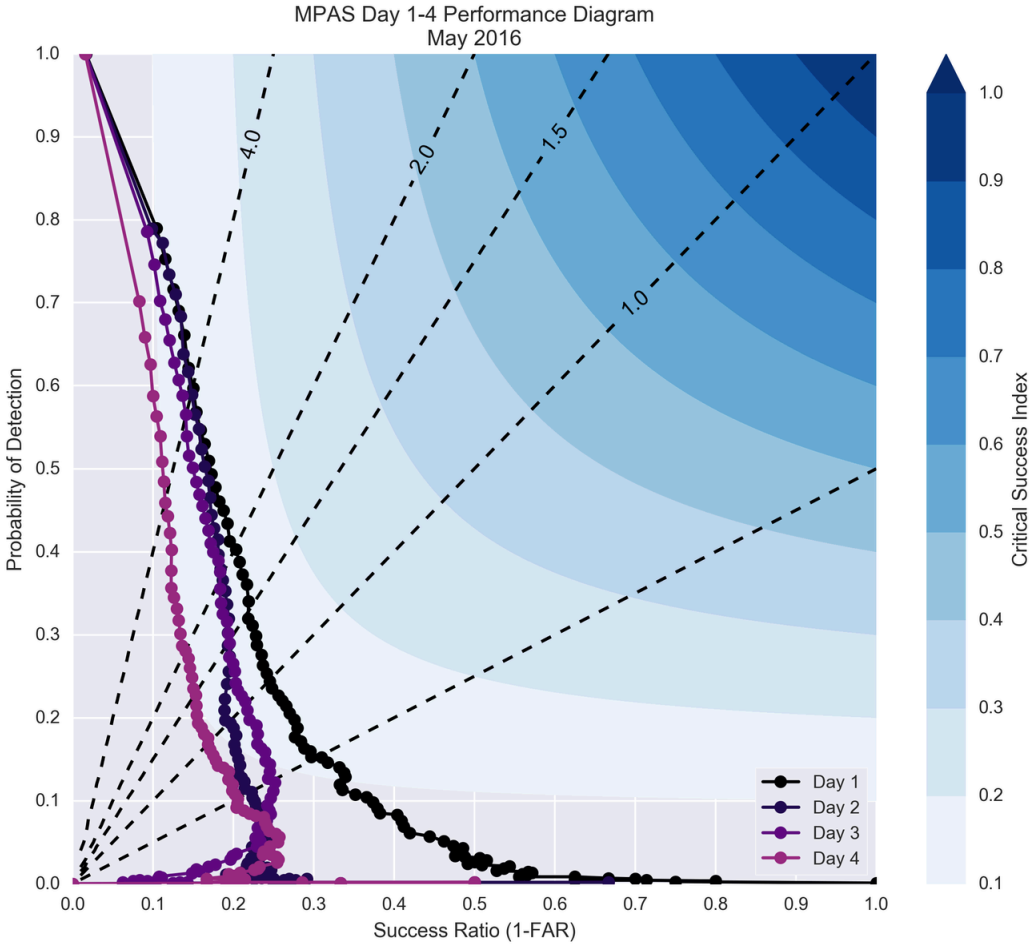

Figure 33 Same as Fig. 31, except for a performance diagram.

The Day 1 forecasts exhibit the greatest CSI and highest resolution ROC curves and overall have the best performance at all probability thresholds (Fig. 33). The Day 2 and Day 3 forecasts have non-zero CSI values, slightly less than the Day 1 forecasts, but are comparable with each other. The Day 2 forecasts exhibited slightly better skill at the lower probability thresholds but Day 3 exhibited higher skill at the mid-range probability thresholds. The least skilled forecasts came from Day 4, although at the higher probability thresholds, fell in-line with the Day 2 and Day 3 CSI values.

Currently, results are pending on the companion 3-km WRF 120 hour forecasts that will be used to compare forecast skill beyond Day 1 between MPAS and existing modeling frameworks in order to determine if additional skill is being added to the forecast. While the results are still pending, some preliminary work has been done comparing the Day 1 MPAS forecasts to the Day 1 forecasts from the 10 member 3-km NCAR Ensemble using the same verification methods discussed previously in order to determine if the model has value at Day 1. It should be noted that the NCAR Ensemble forecasts are treated as an Ensemble-SSPF, or E-SSPF, as detailed by Sobash et al. (2016). The ROC curves and performance diagram for both the MPAS and NCAR Ensemble forecasts are shown in Figures 34 and 35, respectively.

The NCAR Ensemble SSPFs have higher values of CSI in the lower to mid-range probabilities, but MPAS trends closely and even passes the NCAR Ensemble at the higher probability thresholds. This is likely due to the fact that MPAS has a UH distribution that produces a longer tail into the higher UH values. Additionally, the ROC curve for the NCAR Ensemble SSPFs displays slightly higher resolution than the SSPFs for MPAS during the 2016 season.

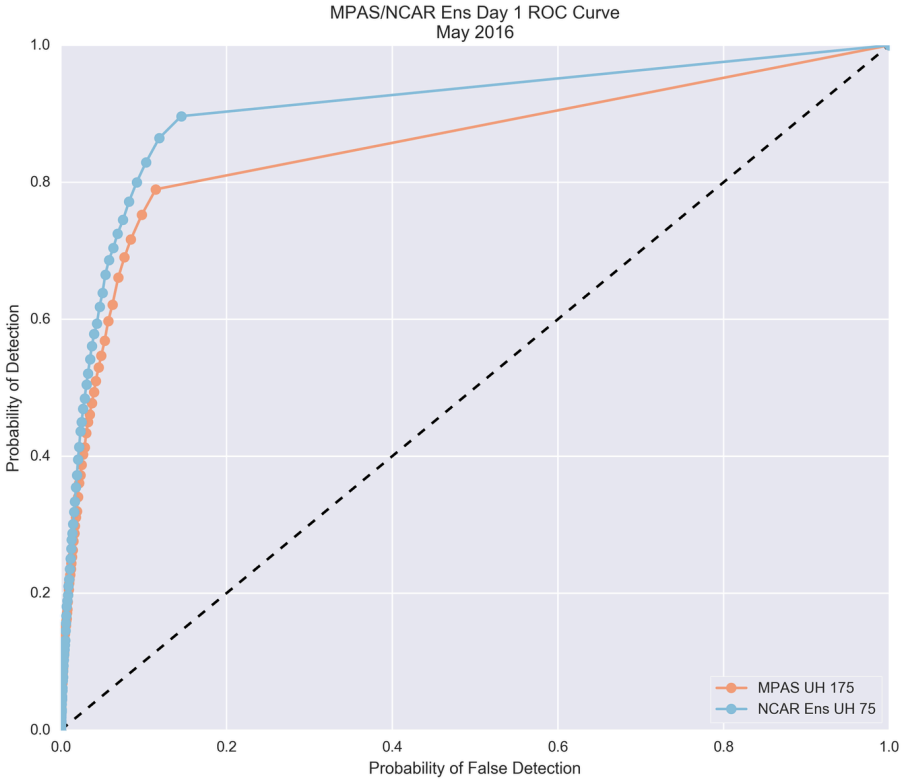

Figure 34 ROC curve for MPAS-derived (blue curve) and NCAR-ensemble-derived (orange curve) probabilistic severe weather forecasts at Day 1 for the 1-31 May 2016 time period.

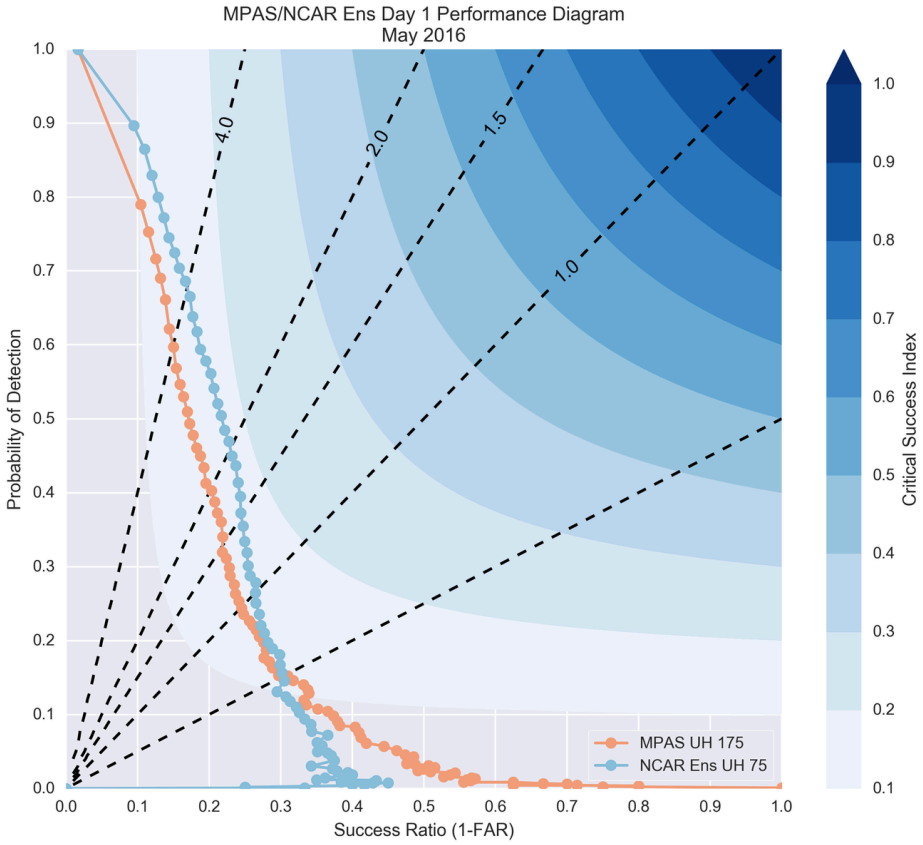

Figure 35 Same as Fig. 22, except for a performance diagram.

d) Model Evaluations – Severe Hazards Desk

1) HRRR VS NAMRR

During SFE2016, the 1500 UTC run of the 3-km hourly NCEP experimental NAMRR Nest was examined for the first time in the HWT and compared to the HRRRv2, which became the operational version of the HRRR at NCEP on 23 August 2016. This evaluation activity focused on a regional area of interest and evaluated how well these deterministic runs depicted storms in the initial conditions and the subsequent evolution of storms during the 15-h forecast. SFE participants subjectively rated the reflectivity forecasts of the 1500 UTC cycle of the HRRRv2 and NAMRR. Overall, the HRRRv2 forecasts received higher subjective ratings than the NAMRR forecasts during SFE2016 (Fig. 36).

Objective neighborhood verification (using a 40-km radius of influence) was also performed on composite reflectivity forecasts from the 1500 UTC cycle of the HRRRv2 and NAMRR (Fig. 37). The HRRRv2 shows a much higher POD and CSI along with lower FAR than the NAMRR at the 30, 40, and 50 dBZ thresholds for all forecast hours of the 1500 UTC cycle. Overall, these objective verification results agree well with the subjective evaluation by SFE participants in showing the better-quality reflectivity forecasts of the HRRRv2 compared to the NAMRR.

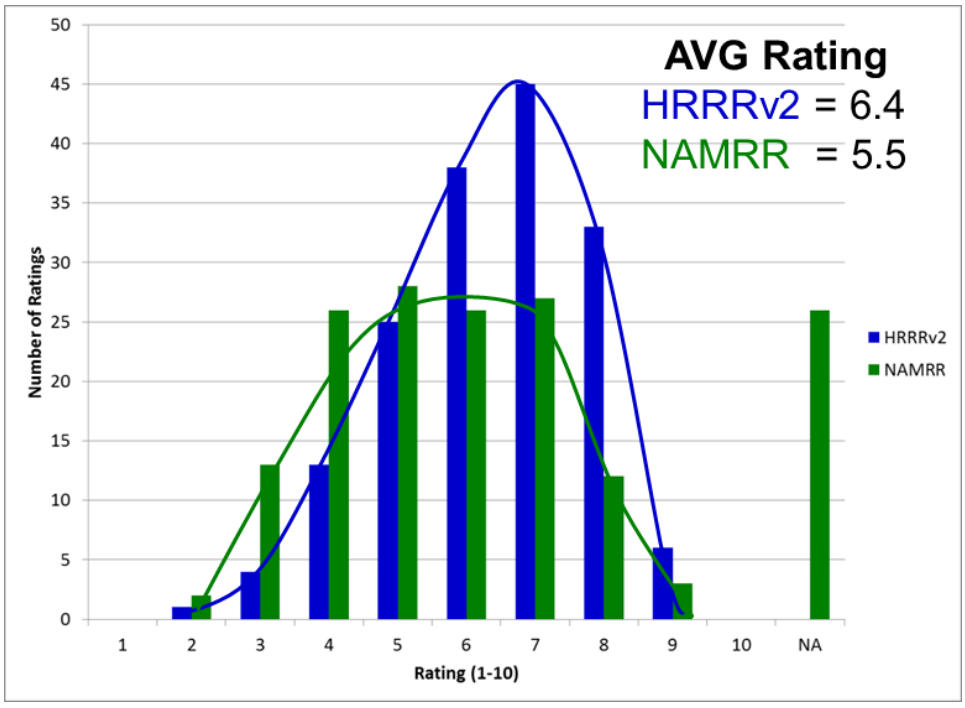

Figure 36 Histogram of subjective ratings (1-10) from SFE participants for reflectivity forecasts over a mesoscale area of interest from the 1500 UTC of the HRRRv2 (blue) and NAMRR Nest (green) during SFE2016.

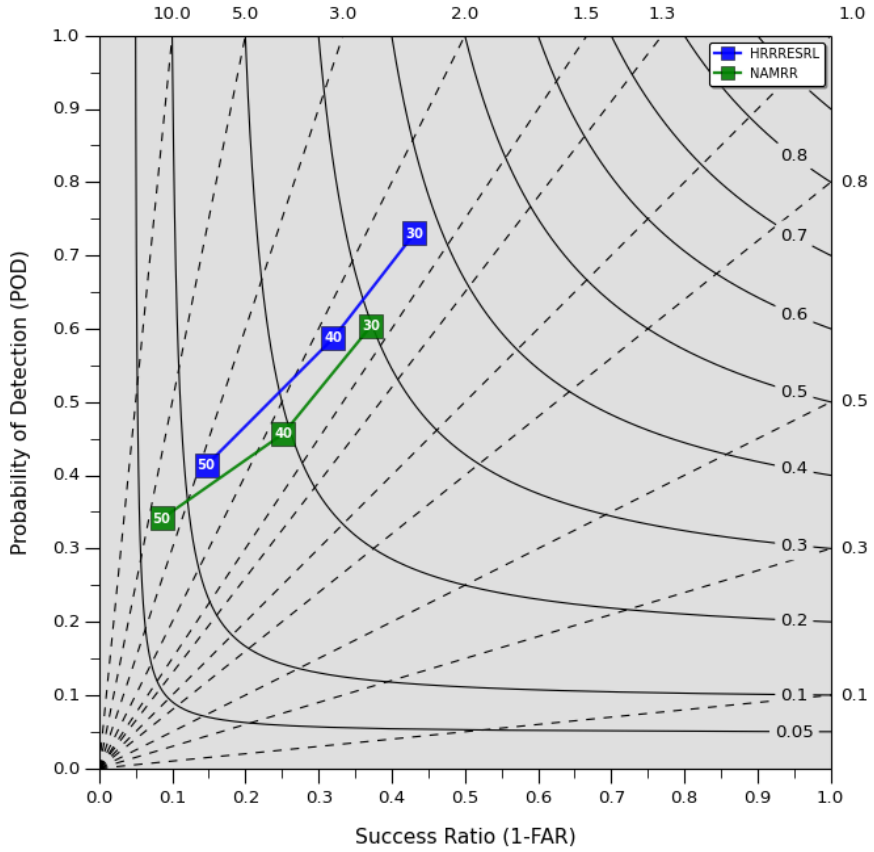

Figure 37 Performance diagram of 40-km neighborhood verification of composite reflectivity forecasts over the CONUS from the HRRRv2 (blue) and NAMRR (green) for the 1500 UTC cycle from 2 May 2016 through 3 June 2016.

2) RADAR DATA ASSIMILATION COMPARISONS

Two 10-member WRF-ARW ensembles with single physics from the CLUE were compared where the only difference is the assimilation of radar data. The members of one ensemble had radar data assimilated in the 0000 UTC NAM analysis using the CAPS 3DVar method, followed by applying SREF perturbations for IC/LBC diversity. The other ensemble did *not* include radar data assimilation, but still applied the same SREF perturbations to the 0000 UTC NAM analysis. Overall, the ensemble with radar data assimilation was subjectively rated slightly better than the ensemble without radar data assimilation for the first 24 hours of the forecast (Fig. 38).

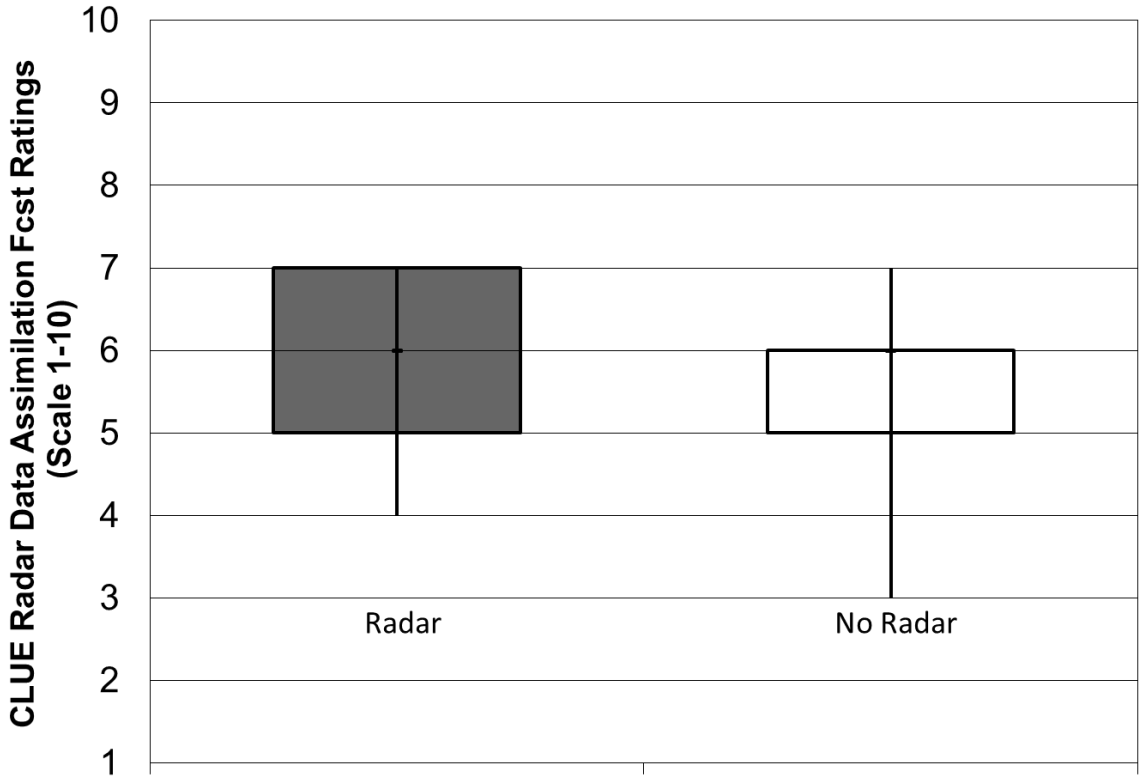

Figure 38 Distributions of subjective ratings (1-10) by SFE participants of probabilistic reflectivity forecasts ≥40 dBZ over a mesoscale area of interest for the first 24 hours of the single-physics CLUE ensembles with and without radar data assimilation.

The ensembles were also compared to identify the length of time that radar data assimilation had a noticeable positive impact on convection-allowing ensemble forecasts. During this part of the evaluation, the ensembles were compared side-by-side with the observations to determine the length of the positive impact of radar data assimilation. The results show a wide variation in the length of impact of radar data assimilation, but most SFE participants felt the impact often fell between four hours and thirteen hours into the forecast (Fig. 39). The objective results in terms of fractions skill score (FSS) generally agree that the cumulative positive impact is generally lost after forecast hour fifteen for the ensemble with radar data assimilation (Fig. 40).

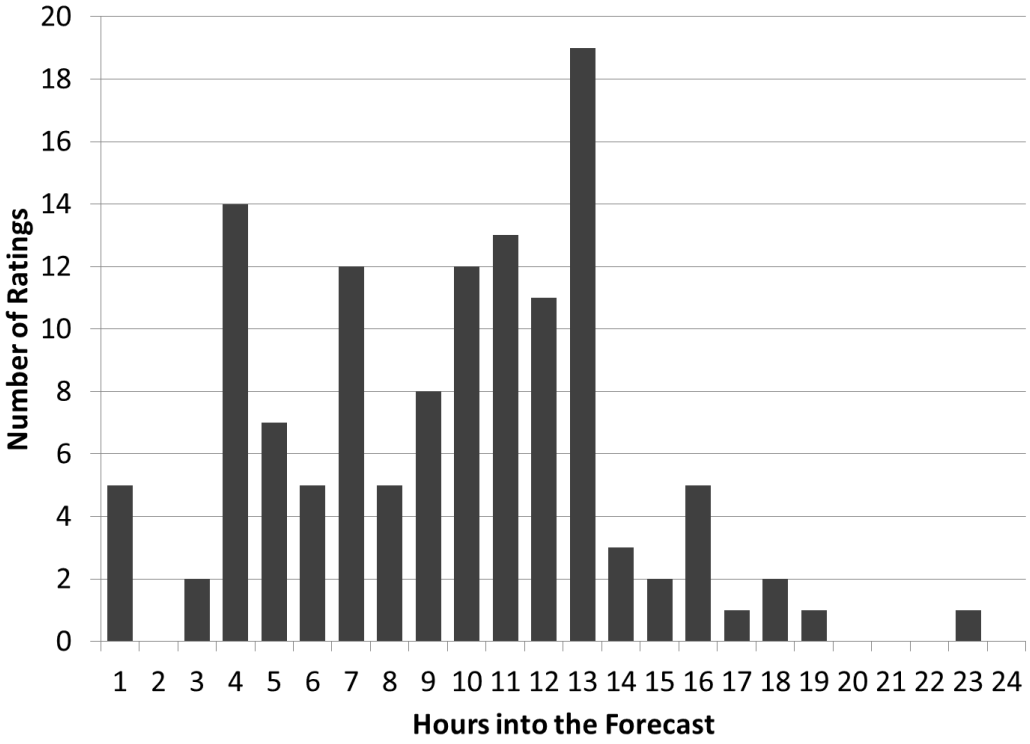

Figure 39 Histogram of the subjective assessment by SFE participants on how long into the forecasts the assimilation of radar data has a positive impact on the ensemble forecast.

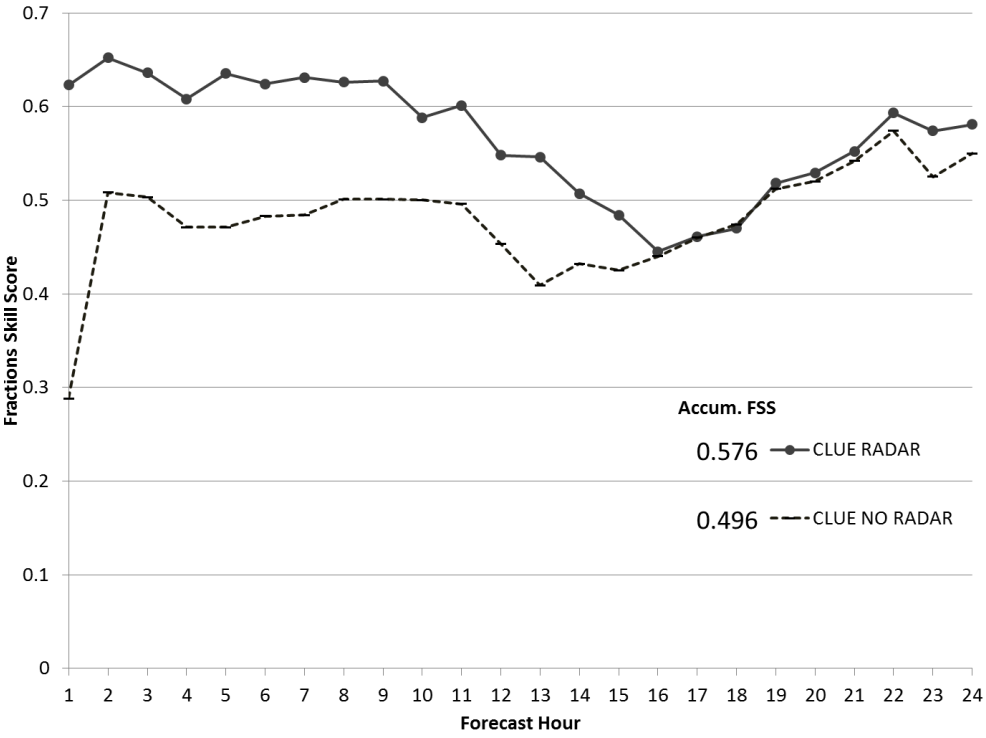

Figure 40 Accumulated fractions skill score of reflectivity forecasts ≥40 dBZ by forecast hour (fh01-fh24) during SFE2016 over a mesoscale area of interest for the CLUE ensembles with and without radar data assimilation.

3) ENSEMBLE SIZE COMPARISONS

Three multi-core ensembles (i.e., equal membership between WRF-ARW and NMMB) were compared where the only difference was the number of ensemble members. A six-member ensemble subset was compared to a ten-member ensemble subset and the full twenty-member multi-core ensemble with single physics (per core) and no radar data assimilation. This experiment was performed to determine the impact of increasing membership on the skill of forecasts from a convection-allowing ensemble. In terms of the reflectivity forecasts, SFE participants rated the ensembles of varying sizes nearly identical during SFE2016 (Fig. 41).

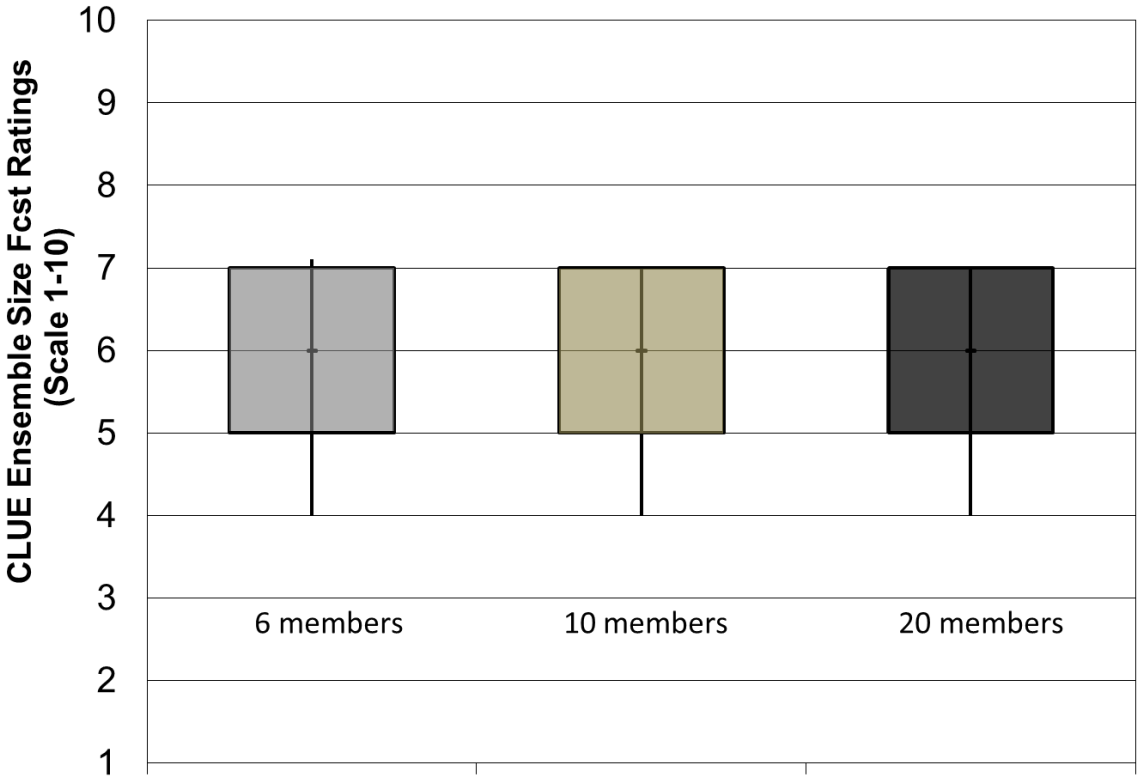

Figure 41 Distributions of subjective ratings (1-10) by SFE participants of probabilistic reflectivity forecasts ≥40 dBZ over a mesoscale area of interest for the forecast hours 13-30 for mixed-core CLUE ensembles of different sizes (i.e., 6, 10, and 20 members).

Objective verification of reflectivity forecasts for the CLUE ensembles of different sizes generally agrees with the subjective evaluation that there is not much difference among the forecasts. In terms of FSS (Fig. 42) and ROC area (Fig. 43), the 20-member ensemble only has a slight statistical advantage over the 10-member ensemble, which has slightly better statistical metrics than the 6-member ensemble. Give the similar subjective ratings among the ensembles and small statistical improvement by adding additional members; a thorough cost-benefit analysis is needed to determine optimal ensemble size for an operational convection-allowing ensemble.

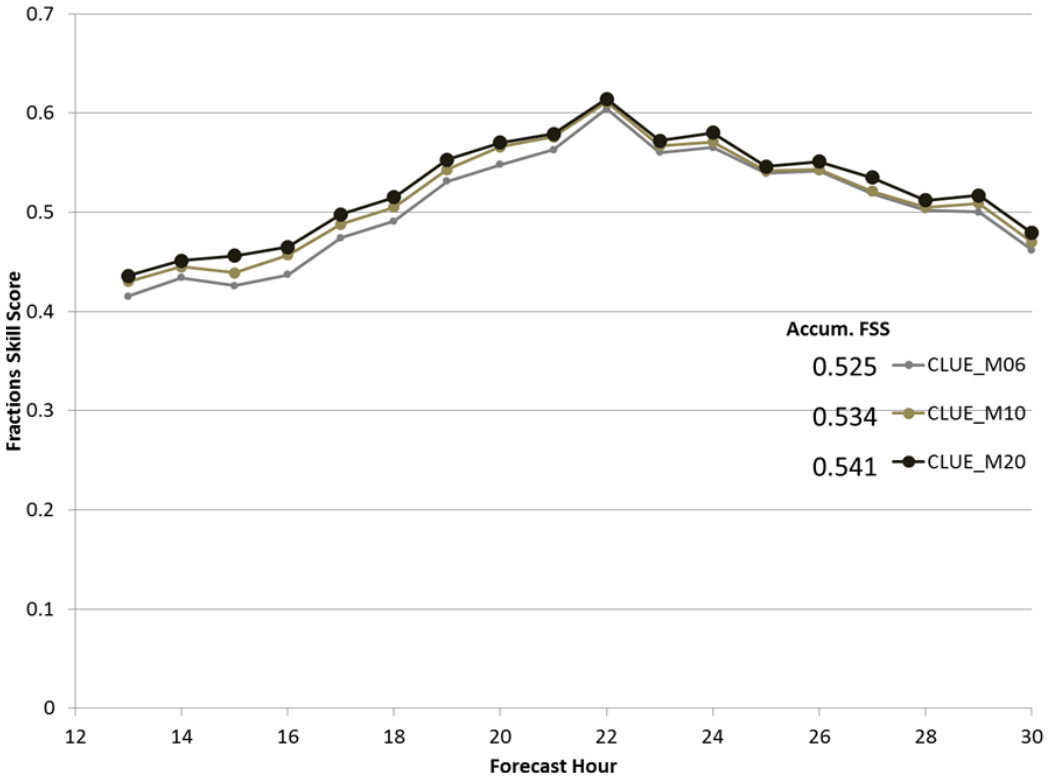

Figure 42 Accumulated fractions skill score of reflectivity forecasts ≥40 dBZ by forecast hour (fh13-fh30) during SFE2016 over a mesoscale area of interest for the CLUE ensembles of different sizes (i.e., CLUE_M06 = 6 members, CLUE_M10 = 10 members, and CLUE_M20 = 20 members).

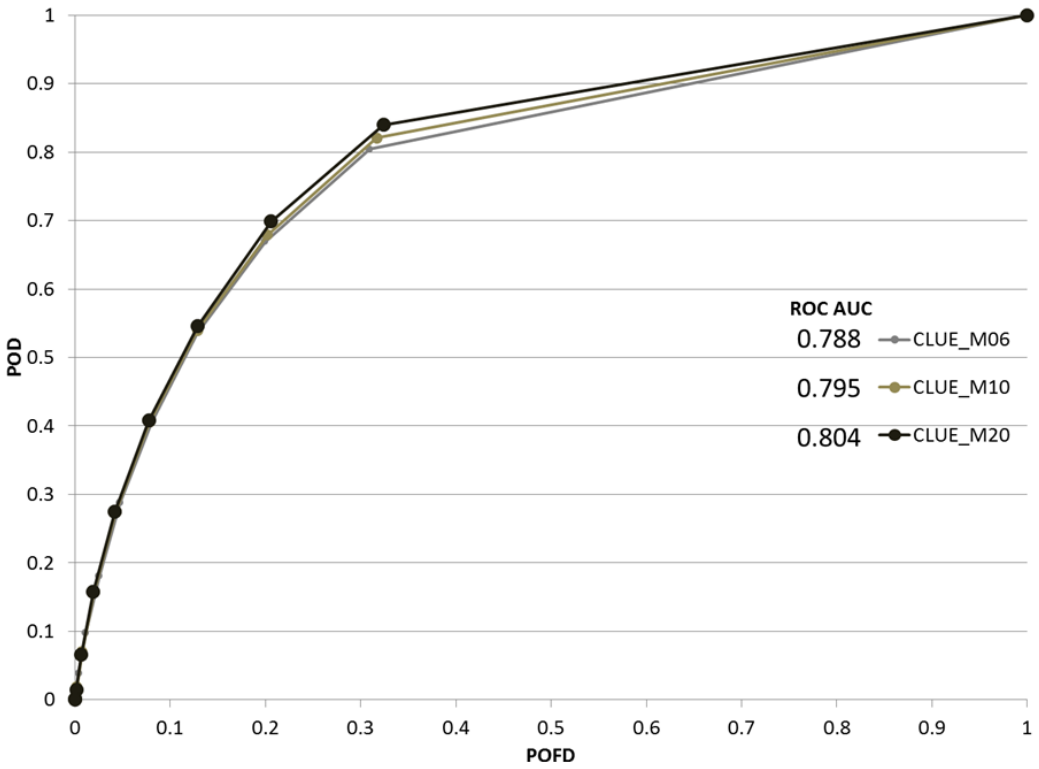

Figure 43 Same as Fig. 42, except for the relative operating characteristic (ROC) curve for probabilistic reflectivity forecasts ≥40 dBZ.

4) COMPARISON TO SSEO AS A BASELINE

Various ensemble subsets from the CLUE, including a multi-core ensemble and two EnKF-based WRF-ARW ensembles, were compared to the SSEO during SFE2016. Given the utility and success of the SSEO in forecasting hazardous weather since 2011, it was used as a baseline to assess the performance of other CAM ensembles with the goal of informing the design of the initial configuration of an operational convection-allowing ensemble. The subjective component of this evaluation examined ensemble forecasts (ensemble maximum and neighborhood probabilities) of hourly maximum fields (HMFs) of UH, updraft speed, and 10-m wind speed relative to LSRs of hail, wind, and tornadoes, as well as ensemble neighborhood probabilities of reflectivity ≥40 dBZ. For the subjective rankings of HMFs (Fig. 44), the SSEO tended to have fewer lower-rated forecasts than the other ensembles while the CAPS EnKF tended to have fewer higher-rated forecasts compared to the other ensembles. For subjective ratings of probabilistic reflectivity forecasts (Fig. 45), the 10-member mixed-core CLUE ensemble and NCAR EnKF had more higher-rated forecasts than the SSEO while the CAPS EnKF remained as the lowest rated ensemble.

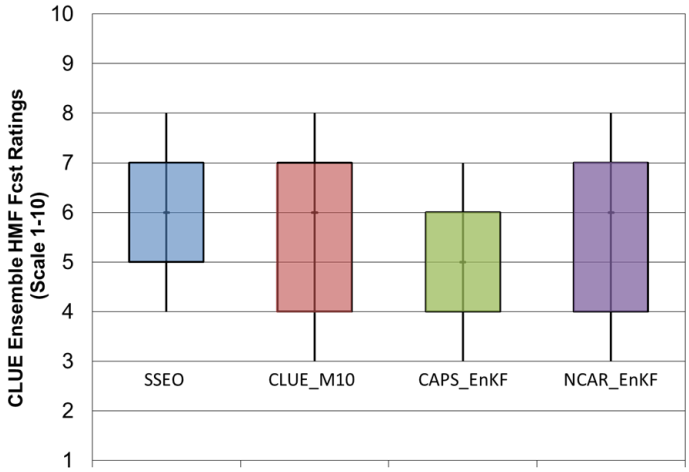

Figure 44 Distributions of subjective ratings (1-10) by SFE participants of HMFs over a mesoscale area of interest for the forecast hours 13-30 for various CLUE ensembles compared to the SSEO.

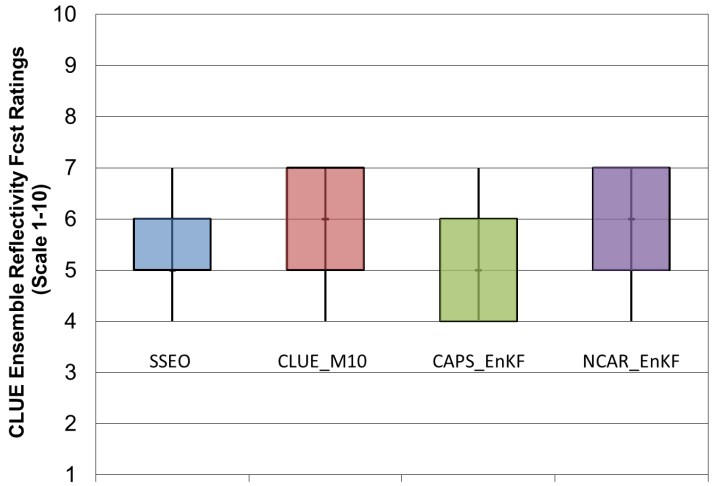

Figure 45 Distributions of subjective ratings (1-10) by SFE participants of probabilistic reflectivity forecasts ≥40 dBZ over a mesoscale area of interest for the forecast hours 13-30 for various CLUE ensembles compared to the SSEO.

The objective verification results of reflectivity forecasts from the SSEO do not necessarily agree with subjective ratings of SFE participants. The FSS (Fig. 46) and ROC curves (Fig. 47) indicate that the SSEO produced more skillful probabilistic reflectivity forecasts than the other CLUE ensembles during SFE2016. The objective results do agree, however, with the subjective ratings of the CAPS EnKF as producing the least skillful reflectivity forecasts of the CLUE ensembles examined. Also of note, the multi-core ensemble (i.e., CLUE_M10) produced slightly better statistical results than the single-core ARW EnKF systems (i.e., NCAR_EnKF and CAPS EnKF).

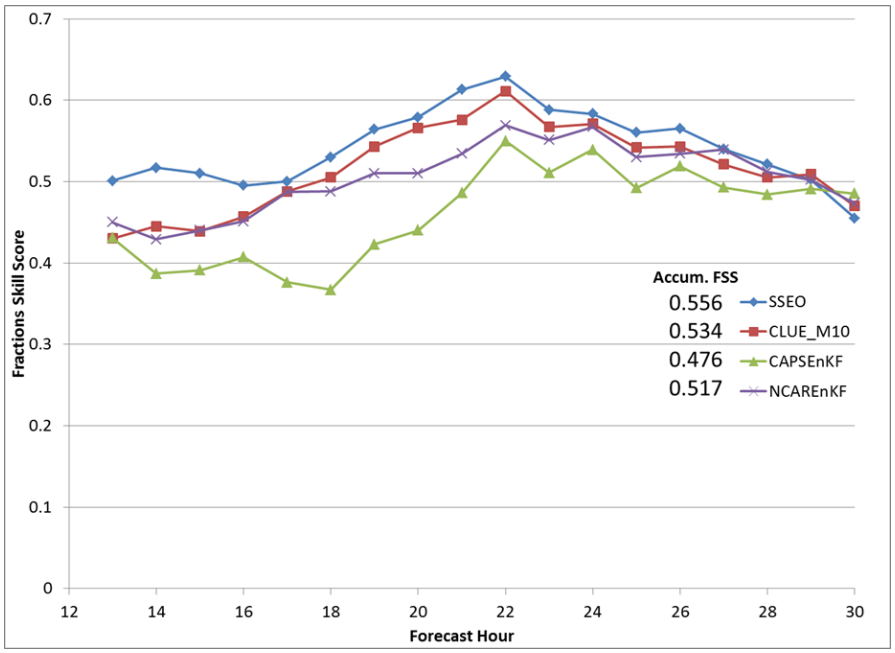

Figure 46 Accumulated fractions skill score of reflectivity forecasts ≥40 dBZ by forecast hour (fh01-fh24) during SFE2016 over a mesoscale area of interest for various CLUE ensembles compared to the SSEO.

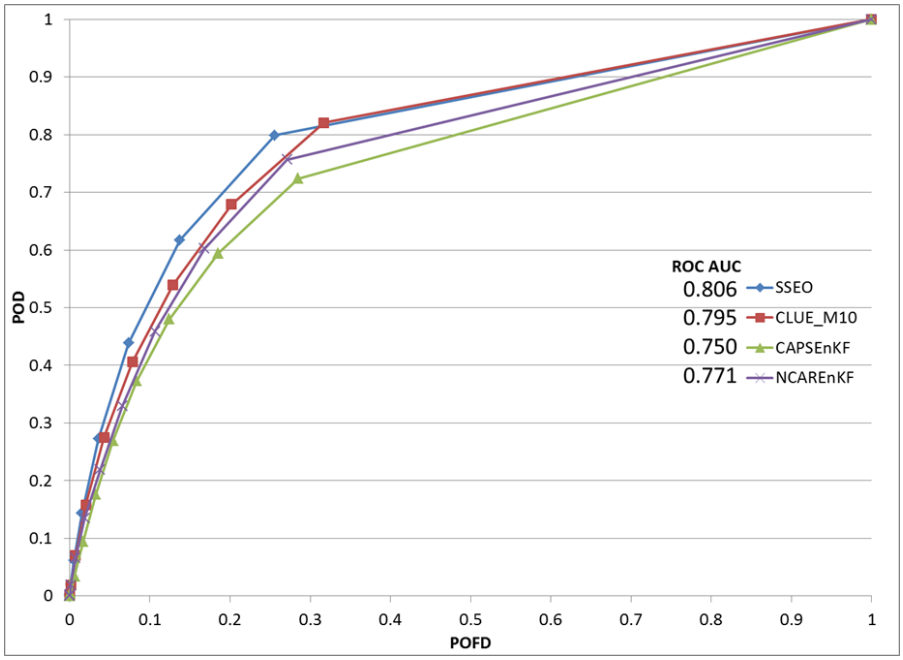

Figure 47 Same as Fig. 45, except for the relative operating characteristic (ROC) curve for probabilistic reflectivity forecasts ≥40 dBZ.

Perhaps the biggest difference in ensemble performance can be seen in the reliability diagram (Fig. 48). While still an over-forecast of reflectivity probabilities, the SSEO produces probabilistic reflectivity forecasts closer to perfect reliability than the other ensembles. The under-dispersiveness of the CLUE ensembles is evident in the strong over-forecasts at nearly aLL probability thresholds with the forecasts typically falling below the "no skill" line. The better reliability of the SSEO reflectivity forecasts highlights the benefit of having greater diversity in a convection-allowing ensemble through a multi-model, multi-initial condition, and multi-physics approach.

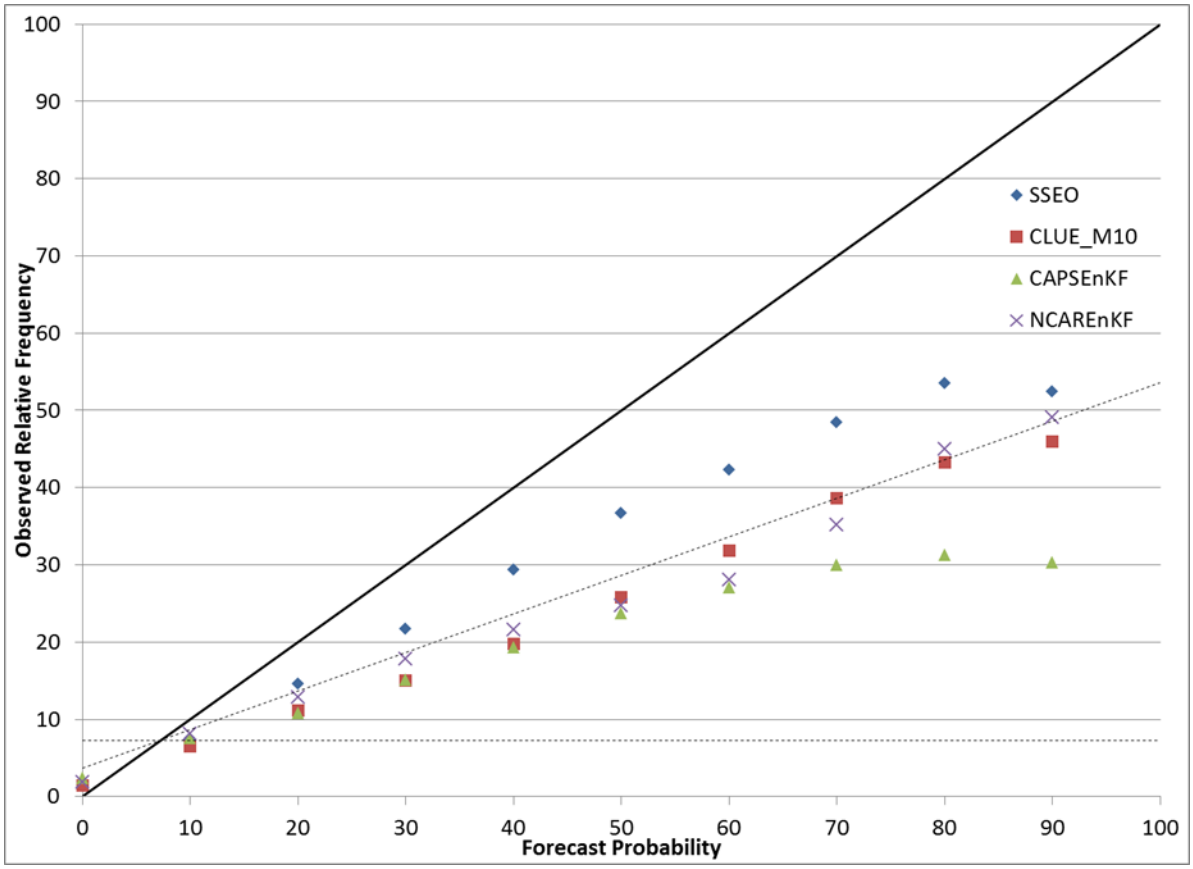

Figure 48 Same as Fig. 47, except for the reliability diagram for probabilistic reflectivity forecasts ≥40 dBZ.

e) Evaluation of Texas Tech University Sensitivity Products

Ensemble sensitivity was calculated within the TTU real time ensemble system for each 0000 and 1200 UTC 48-hr forecast initialization during SFE2-16. A response function location was chosen during the experiment each day to capture the expected areas of severe convection within Day 1 and Day 2 forecasts. Sensitivity of a number of severe convective parameters within the response location was calculated hourly in real time within the 12-48 hr forecast window, plotted, and evaluated. These response functions included "magnitude" functions defined as the 1- and 6-hr maximum 2-5 km updraft helicity, surface wind speed, and simulated lowest model level reflectivity. Response functions also included "coverage" functions defined as the 1- and 6-hr number of grid points exceeding 40 dBZ simulated lowest model level reflectivity, 50 m^2/s^2 2-5 km updraft helicity, and 40 mph surface wind speed. The sensitivity of these response functions were calculated with respect to 300-hPa wind speed, 500-hPa geopotential height, 700-hPa temperature and dew point, 850-hPa

temperature, surface temperature and dew point, and sea level pressure. The purpose of this initial evaluation was to learn the consistency of coherent sensitivity signals across many cases of severe convection, understand the nature of these signals, and simulate how the sensitivity fields could be used operationally to improve forecasts.

Sensitivity patterns nearly always showed coherent signals aloft with respect to 300-hPa wind speed and 500-hPa geopotential height, and tended to reveal significant sensitivity to the positions and magnitudes of local minima and maxima (for 300-hPa wind speed) and ridges, troughs, and gradients (for 500-hPa geopotential height) in the field. Figure 49 shows examples of these typical sensitivity fields for two independent cases. A clear positional sensitivity signal exists at forecast hour 18 in the case initialized on May 7, indicating that shifts in the jet core toward the northwest (reduces 48-hr maximum 6-hourly 2-5 km updraft helicity because higher wind speed values would exist in areas of negative sensitivity) or southeast (increases 48-hr maximum 6-hourly 2-5 km updraft helicity because higher wind speed values would exist in areas of positive sensitivity) are associated with differences in the response function in the green box. The sensitivity pattern also shows a number of broad features in the case shown in Figure 49 initialized on May 3 with respect to the 500-hPa geopotential height field. The features shown in Figure 49, and those frequently observed, generally propagated upstream backward in time in a coherent way, highlighting the dynamical relevance of the flow structures (e.g. a 500-hPa trough) to which they were attached and suggesting they weren't present as a result of statistical noise. Sensitivity fields were, however, much noisier closer to the surface, and rarely provided useful signals when calculated with respect to temperature or dew point at 700- and 850-hPa. Sensitivity to sea level pressure was slightly more useful, but not nearly as much as the 500- and 300-hPa sensitivity fields.

Sensitivity patterns also generally spanned the entire domain as shown in Figure 49. This highlights the ability of the ensemble to pick out relationships between early forecast variables and response functions that exhibit no direct dynamical link. In other words, features typically revealed through the sensitivity fields, such as those described above, commonly existed downstream of the response function location at a time prior to the valid time of the response itself. This is because within the ensemble, the evolution of different features can be related to the same processes without directly affecting each other.

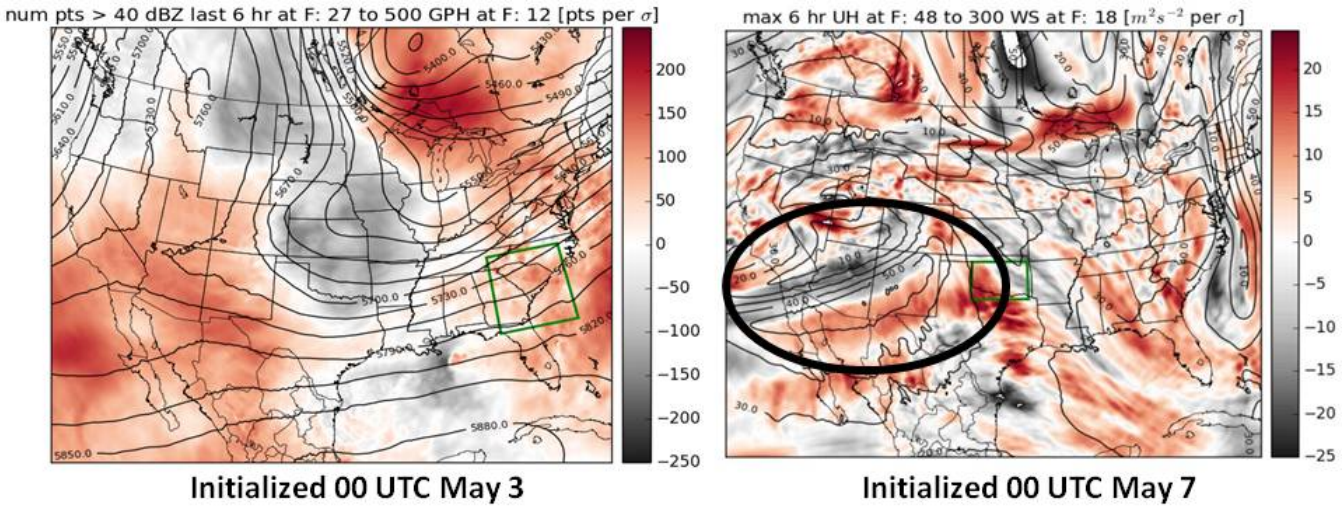

Figure 49 Ensemble sensitivity of 27-hr forecast number of grid points exceeding 40 dBZ simulated reflectivity over the last 6 hours with respect to 500-hPa geopotential height at forecast hour 12 (forecast initialized 00 UTC May 3), and sensitivity of 48-hr forecast maximum 6-hourly 2-5 km updraft helicity with respect to 300-hPa wind speed at forecast hour 18 (forecast initialized 00 UTC May 7). The green box in both plots shows the response function location.

The sensitivity fields associated with the different response functions over the spring forecasting experiment were generally different to a degree such that they suggested variations in early-forecast flow features affected forecasts of different hazards in different ways. Figure 50 shows an extreme example of this for a 22-hr forecast for which the response location was located over the majority of Oklahoma (not shown). Sensitivity with respect to 3-hr sea level pressure was observed to have very different signals at and near the precursor midlatitude cyclone in the Northern Plains for coverage response functions associated with updraft helicity, surface winds, and simulated reflectivity. In particular, the position of the cyclone was relevant to the coverage of high winds, while either positive (for updraft helicity) or negative (for simulated reflectivity) sensitivity existed over much of the area occupied by the cyclone, suggesting the intensity of the cyclone itself was more relevant to these response functions. Interestingly, the opposite sign of the sensitivity pattern for reflectivity and updraft helicity indicates a change in the intensity of the cyclone would have opposite effects on the coverage of these parameters. More commonly the differences in sensitivity patterns were not this extreme, but almost always showed differences in position and magnitude that indicate how the atmospheric state affects different convective hazards in different ways. There were numerous events when the sensitivity field of certain response variables was near zero across the domain, but showed clear, coherent features for other response functions.

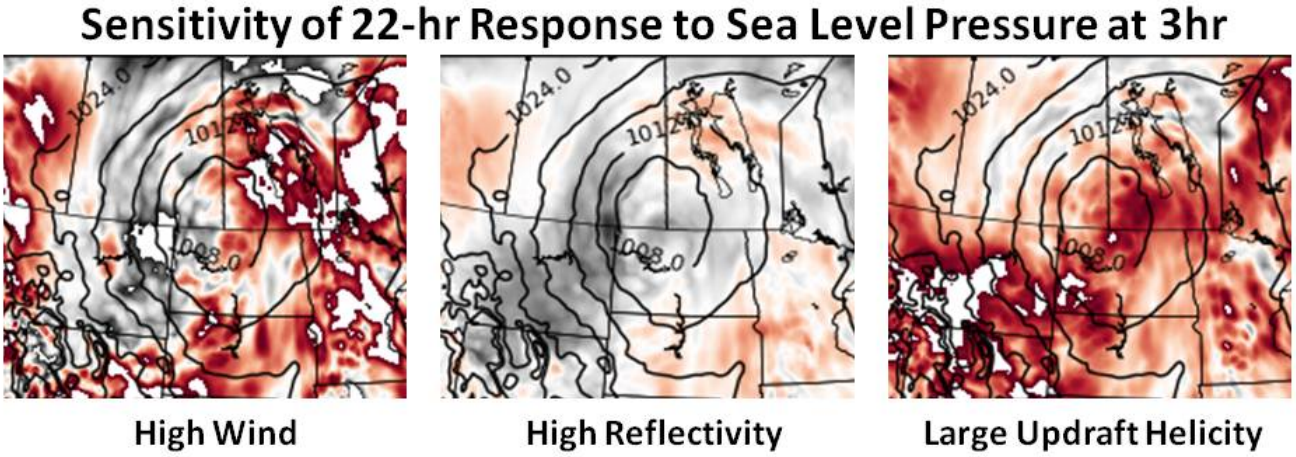

Figure 50 Ensemble sensitivity of the coverage of high surface wind speed (greater than 40 mph), high simulated reflectivity (greater than 40 dBZ), and large updraft helicity (greater than 50 m2/s2) at 22-h forecast time with respect to sea level pressure at forecast hour 3. Red values indicate positive sensitivity, grey values indicate negative sensitivity. The response function location at 22-hr forecast time is not show, but exists over the majority of Oklahoma. Note: areas of white within areas of large sensitivity values indicate large magnitudes that exceeded the maximum values within the chosen range of plotted colors.

The initial conclusions from the evaluation of the sensitivity products are that wind and geopotential height sensitivities aloft are probably useful in diagnosing the important features relevant to forecasts of severe convective hazards, while sensitivity with respect to moisture and temperature closer to the surface are probably less useful. Further, sensitivities of different convective hazards are able to discriminate different features that affect the hazards in different ways. Participant feedback from the experiment was in agreement as to this usefulness of the sensitivity field, but forecasters overwhelmingly thought the subjective interpretation of these fields in real time would not be feasible. As a result, these results have collectively driven the next phase in this work, which is to use sensitivity to objectively modify the ensemble of forecasts (e.g. through choosing ensemble subsets) to provide a new forecast distribution with greater skill than the original ensemble.

4. Summary

The 2016 Spring Forecasting Experiment (SFE2016) was conducted at the NOAA Hazardous Weather Testbed from 2 May – 3 June by the SPC and NSSL with participation from forecasters, researchers, and developers from around the world. The primary theme of SFE2016 was to utilize convection-allowing model and ensemble guidance in creating high-temporal resolution probabilistic forecasts of severe weather hazards, including extension into the Day 2 period. Furthermore, this year a major effort was made to coordinate CAM-based ensemble configurations much more closely than in previous years, which was done in the context of the Community Leveraged Unified Ensemble (CLUE). The CLUE allowed us to conduct several experiments geared toward identifying optimal configuration strategies for CAM-based ensembles, and was especially well timed to help inform the design of the first operational CAM-based ensemble for the US, which is planned for implementation by NOAA's NCEP/EMC in the upcoming years.

Several preliminary findings/accomplishments from SFE2016 are listed below:

- Four-hour outlooks for individual severe hazards were generated using first-guess guidance from a temporally disaggregated full-period outlook created with calibrated guidance from the SREF and SSEO.
- Severe weather isochrones were explored to add enhanced timing information of the severe weather threat. The isochrones were drawn to delineate the start time of the 4-h time window with the highest severe weather probability.
- The CLUE allowed for an unprecedented number of controlled experiments on convection-allowing ensemble design. This is closely aligned with UMAC recommendations for evidence-based model development decision-making and utilization of a unified collaborative strategy that better leverages capabilities of the larger community.
    - The multi-core ensemble strategy provided better probabilistic forecasts for severe weather while a single-core ARW ensemble verified better for QPF and PQPF.
    - For reflectivity and updraft helicity forecasts, the statistical improvement in forecast skill by increasing the size of convection-allowing ensemble membership was relatively small.
    - The ensemble comprised of members with radar DA verified better than the ensemble without radar DA through ~15 hours into the forecast cycle.
    - The SSEO verified objectively as well as or better than any CLUE subset for probabilistic reflectivity forecasts. An operational version of the SSEO is being developed by EMC in FY17 to serve as a baseline for future CAM ensemble improvements.
- Examined and documented the characteristics of three algorithms for predicting hail size. Post-experiment analysis has found that simply using UH as a hail predictor may still have the most skill, especially if hourly minimum UH (i.e., anti-cyclonic mesocyclones) are considered.
- The HRRRv2 generated better reflectivity forecasts than the NAMRR both subjectively in terms of participant ratings and objectively through neighborhood verification.
- The variable-resolution MPAS runs at convection-allowing scale over the CONUS were explored subjectively and objectively in generating realistic simulated storm structures out to Day 5.

- The sensitivities of different microphysics schemes were documented, including differences in cold pool strengths and updraft speeds..
- Ensemble sensitivity analysis revealed that the wind and geopotential height sensitivities aloft are the most useful for diagnosing features relevant to forecasts of convective hazards (i.e., reflectivity, updraft helicity, and 10-m wind speed).

The CLUE was a successful venture during SFE2016 in pulling together five research institutions in an unprecedented effort to help guide NOAA's operational modeling efforts at the convective scale. Eight unique, controlled experiments were designed within the CLUE framework to examine issues directly relevant to the design of NOAA's future operational CAM-based ensembles. SFE2016 was also successful in testing new forecast products and other modeling systems to address relevant issues related to the prediction of hazardous convective weather. The findings and questions exposed during SFE2016 directly promote continued progress to improve forecasting of severe weather in support of the NWS Weather-Ready Nation initiative.

Acknowledgements

SFE2016 would not have been possible without dedicated participants and the support and assistance of numerous individuals at SPC and NSSL. In addition, collaborations with CAPS, NCAR, the United Kingdom Met Office, ESRL/GSD, the University of North Dakota, and EMC were vital to the success of SFE2016. In particular, Ming Xue (CAPS), Fanyou Kong (CAPS), Kevin Thomas (CAPS), Keith Brewster (CAPS), Youngsun Jung (CAPS), Glen Romine (NCAR), Craig Schwartz (NCAR), Ryan Sobash (NCAR), Kate Fossell (NCAR), Steve Willington (Met Office), Curtis Alexander (ESRL/GSD), Aaron Kennedy (UND), Joshua Markel (UND), Xiquan Dong (UND), Geoff Dimego (EMC), Jacob Carley (EMC), Brad Ferrier (EMC), and Eric Aligo (EMC) were essential in generating and providing access to model forecasts examined on a daily basis.

References

Adams-Selin, R. 2013: In-line 1D WRF hail diagnostic. AFWA Internal Tech. Memo, SEMSD.21495.

Brimelow, J.C., 1999: Modeling maximum hail size in Alberta thunderstorms. *Wea. Forecasting*, **17**, 1048-1062.

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

Jewell, R., and J. Brimelow, 2009: Evaluation of Alberta hail growth model using severe hail proximity soundings from the United States. *Wea. Forecasting*, **24**, 1592-1609.

Jirak, I. L., C. J. Melick, A. R. Dean, S. J. Weiss, and J. Correia, Jr., 2012: Investigation of an automated temporal disaggregation technique for convective outlooks during the 2012 Hazardous Weather Testbed Spring Forecasting Experiment. Preprints, *26th Conf. on Severe Local Storms*, Nashville, TN, Amer. Meteor. Soc., 10.2.

Jirak, I. L. C. J. Melick, and S. J. Weiss, 2014: Combining probabilistic ensemble information from the environment with simulated storm attributes to generate calibrated probabilities of severe weather hazards. Preprints, *27th Conf. on Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., 2.5.

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.

Rothfusz, L. P., P. T. Schlatter, E. Jacks, and T. M. Smith, 2014: A future warning concept: Forecasting A Continuum of Environmental Threats (FACETs). *2nd Symposium on Building a Weather-Ready Nation: Enhancing Our Nation's Readiness, Responsiveness, and Resilience to High Impact Weather Events,* Atlanta, GA, Amer. Meteor. Soc., 2.1.

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.Skamarock, W. C., Klemp, J. B., Duda, M., Fowler, L. D., Park, S.-H., and T. Ringler, 2012: A multiscale nonhydrostatic atmospheric model using centroidal Voronoi tesselations and c-grid staggering. *Mon. Wea. Rev*, ***140***, 3090–3105.

Stensrud, D. J., and Co-authors, 2009: Convective-scale warn-on-forecast system. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1499.

APPENDIX

Daily activities schedule in local (CDT) time

Severe Hazards Desk	*Total Severe Desk*
0800 – 0845: **Evaluation of Experimental Forecasts & Guidance**	
Subjective rating relative to radar evolution/characteristics, warnings, and preliminary reports and objective verification using preliminary reports and MESH	
• Day 1 & 2 full-period probabilistic forecasts of tornado, wind, and hail • Day 1 4-h period forecasts and guidance for tornado, wind, and hail	• Days 1, 2, & 3 full-period probabilistic forecast of total severe • Day 1 4-h period probability forecasts and isochrones
0845 – 1115: **Day 1 Convective Outlook Generation**	
Hand analysis of 12Z upper-air maps and surface charts	
• Day 1 full-period probabilistic forecasts of tornado, wind, and hail valid 16-12Z over mesoscale area of interest • Day 1 4-h probabilistic forecasts of tornado, wind, and hail valid 18-22 and 22-02Z*	• Day 1 full-period probabilistic forecast of total severe valid 16-12Z over mesoscale area • Day 1 4-h probabilistic total severe forecasts valid 18-22, 20-00, 22-02, 00-04, and 02-06Z. • Day 1 isochrones for 4-h periods (every 2 h) with highest probability of total severe*
1115 – 1130: **Break**	
Prepare for map discussion	
1130 – 1200: **Map Discussion**	
Brief discussion of today's forecast challenges and products Highlight findings from previous days	
1200 – 1300: **Lunch**	
Brief EWP participants at 1245 if needed	
1300 – 1345: **Day 2 Convective Outlook Generation**	
• Day 2 full-period probabilistic forecasts of tornado, wind, and hail valid 12-12Z over mesoscale area of interest	• Day 2 or Day 3 full-period probabilistic forecasts of total severe valid 12-12Z over mesoscale area of interest
1345 – 1515: **Scientific Evaluations**	
• CLUE (4): Radar data assimilation • CLUE (8): Ensemble size comparisons • CLUE: SSEO as baseline • NAMRR Nest, HRRRv2, HRRRv3 • Ensemble sensitivity (TTU)	• CLUE (1,2): Model core (det. & ens.) • CLUE (6): Radar data assimilation approaches • CLUE (7): Microphysics sensitivity • Explicit hail size forecast comparison
1515 – 1600: **Short-term Outlook Update**	
• Update 4-h probabilistic forecasts of tornado, wind, and hail valid 22-02Z*	• Update 4-h probabilistic forecasts of total severe for 22-02, 00-04, and 02-06Z. • Update total severe isochrones (22, 00, & 02Z)*

* Denotes forecasts also made by participants using the PHI tool on Chromebooks.

Table A1 List of weekly participants (with affiliation) during SFE2016.

Week 1	Week 2	Week 3	Week 4	Week 5
May 2-6	**May 9-13**	**May 16-20**	**May 23-27**	**May 31-June 3**
Brian Ancell (TTU)	Brian Ancell (TTU)	Brock Burghardt (TTU)	Brock Burghardt (TTU)	Brian Ancell (TTU)
Brock Burghardt (TTU)	Brock Burghardt (TTU)	Jack Kain (NSSL)	Andy Taylor (WFO FGZ)	Aaron Kennedy (UND)
Mike Evans (WFO BGM)	Bill Gallus (Iowa State)	Tom Workhoff (FirstEnergy)	Pete Wolf (WFO JAX)	Brooke Hagenhoff (UND)
Madalina Surcel (McGill Univ.)	Brian Squitieri (Iowa State)	Nathan Hitchens (Ball State)	Nathan Wendt (SPC)	Joshua Markel (UND)
Greg Gallina (WPC)	Sean Stelten (Iowa State)	Lance Bosart (SUNYA)	Mark Rodwell (ECMWF)	Jingyu Wang (UND)
Ben Albright (WPC/HMT)	Jim Nelson (WPC)	Kyle Pallozzi (SUNYA)	Vince Agard (MIT)	John Stoppkotte (WFO LBF)
Mike McClure (WFO DVN)	Marc Chenard (WPC)	Bruno Ribeiro (SUNYA/Brazil)	Aaron Johnson (WFO DDC)	Binbin Zhou (EMC)
Pat Spoden (WFO PAH)	Stan Czyzyk (WFO VEF)	Bill Martin (WFO GSP)	Ivan Tsonevsky (ECMWF)	Jeff Beck (GSD)
Tom Holtquist (WFO MPX)	Robert Hepper (SPC)	Glen Romine (NCAR)	Matt Pyle (EMC)	Isidora Jankov (GSD)
Stephen Bieda (WFO AMA)	Jacob Carley (EMC)	Corey Guastini (EMC)	John Brown (GSD)	Jeff Milne (SPC)
Ryan Sobash (NCAR)	Trevor Alcott (GSD)	Curtis Alexander (GSD)	Ed Szoke (GSD)	Greg Thompson (NCAR) W-F
Geoff Manikin (EMC)	David Dowell (GSD) Th,F	David Dowell (GSD) M,T	Steve Willington (UK Met)	Hugh Morrison (NCAR) W-F
Terra Ladwig (GSD)	Becky Adams-Selin (USAF)	Anke Finnenkoetter (UK Met)	Mark Bevan (UK Met)	Jason Milbrandt (EC) W-F
Eric James (GSD)	Humphrey Lean (UK met)	Humphrey Lean (UK Met) M	Glenn White (EMC)	Steve Willington (UK Met)
Makenzie Krocak (OU)	Anke Finnenkoetter (UK Met) F	Tracey Dorian (EMC)	David John Gagne (OU)	Mark Conder (WFO LUB)
	Bill Skamarock (NCAR) W-F	Pamela Eck (SUNYA)	Nick Nauslar (SPC)	Ron Miller (WFO OTX)
	Corey Potvin (NSSL)	Paula Davidson (NWS) T,W	Bruce Entwistle (AWC)	Pam Heinselman (NSSL)