SPRING FORECASTING EXPERIMENT 2017

Conducted by the

EXPERIMENTAL FORECAST PROGRAM

of the

NOAA HAZARDOUS WEATHER TESTBED

http://hwt.nssl.noaa.gov/Spring_2017/

HWT Facility – National Weather Center
1 May - 2 June 2017

Preliminary Findings and Results

Adam Clark[2], Israel Jirak[1], Steven J. Weiss[1], Jack Kain[2], Kent Knopfmeier[2,3], Burkely Gallo[4], Robert Hepper[1,3], Chris Karstens[1], Andy Dean[1], Pam Heinselman[2], Jessica Choate[2,3], Patrick Skinner[2,3], Gerry Creager[2,3], Scott Dembek[2,3], and David Imy[2]

(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

1. Introduction

The 2017 Spring Forecasting Experiment (SFE2017) was conducted from 1 May – 2 June by the Experimental Forecast Program (EFP) of the NOAA/Hazardous Weather Testbed (HWT), and was co-led by the NWS/Storm Prediction Center (SPC) and OAR/National Severe Storms Laboratory (NSSL). Additionally, 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, Multi-scale data Assimilation and Predictability (MAP) Laboratory at the University of Oklahoma, Earth Systems Research Laboratory/Global Systems Division (ESRL/GSD), Geophysical Fluid Dynamics Laboratory (GFDL), United Kingdom Meteorological Office (Met Office), National Center for Atmospheric Research (NCAR), and NCEP's Environmental Modeling Center (EMC). Participants included more than 80 forecasters, researchers, model developers, university faculty and graduate students from around the world (see Table 1 in Appendix). As in previous years, SFE2017 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 severe weather outlooks than those issued operationally by SPC.
    - 4-h periods for individual severe hazards (tornado, hail, and wind)
    - 1-h periods for near-term total severe
- Explore the ability to generate experimental Day 2 severe weather 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.
- Explore methods to include more detailed timing information using isochrones to delineate the start-time of 4-h time windows with the highest total severe probabilities.
- Test the feasibility of generating short lead-time, 1-h time window convective outlooks using a prototype WoF system.

Applied Science Activities:

- Compare various CAM ensemble prediction systems to identify strengths and weaknesses of different configuration strategies. Most of these comparisons were conducted within the framework of the Community Leveraged Unified Ensemble (CLUE) discussed below. Additional comparisons were made using the Storm Scale Ensemble of Opportunity (SSEO) as a baseline.
- Examine CAM ensemble forecasts into Day 2 and assess their guidance for generating severe weather outlooks, including individual severe hazards.
- Compare and assess different approaches in CAMs for predicting hail size.
- Document characteristics of various microphysics schemes used with the WRF model.
- Compare strengths and weaknesses of convective forecasts from a High Resolution Ensemble Forecast system version 2 (HREFv2) to the SSEO; an almost identical configuration of HREFv2 became operational at EMC on 1 November 2017.
- Evaluate initial convective-scale versions of the Finite Volume Cubed Sphere model (FV3) using 3-km grid-spacing.
- Evaluate a prototype WoF system – the NSSL Experimental Warn-on-Forecast System for ensembles (NEWS-e) – for applications to short-term severe weather outlook generation.

As in previous experiments, a suite of state-of-the-art experimental CAM guidance contributed by our large group of collaborators was central to SFE2017. Additionally, for the second consecutive year, these contributions were formally coordinated into a single ensemble framework called the Community Leveraged Unified Ensemble (CLUE). The 2017 CLUE was constructed by having all groups agree on a set of model specifications (e.g., grid-spacing, vertical levels, domain size, etc.) so that the simulations contributed by each group could be used in controlled experiments. This design allowed us to conduct several experiments to aid in identifying optimal configuration strategies for CAM-based ensembles. The 2017 CLUE included 76 members using 3-km grid-spacing that allowed a set of five unique experiments. SFE2017 activities also involved testing of a Warn-on-Forecast prototype system called the NSSL Experimental WoF System for Ensembles (NEWS-e).

This document summarizes the activities, core interests, and preliminary findings of SFE2017. 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_2017/HWT_SFE2017_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 SFE2017 along with a description of the daily activities, Section 3 reviews the preliminary findings of SFE2017, and Section 4 contains a summary of these findings.

2. Description

a) Experimental Models and Ensembles

Building upon successful experiments of previous years, SFE2017 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. More than 100 unique CAMs were run for SFE2017, of which 76 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 volume of CAMs run for SFE2017 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 during the past two years has made the increase in members more manageable.

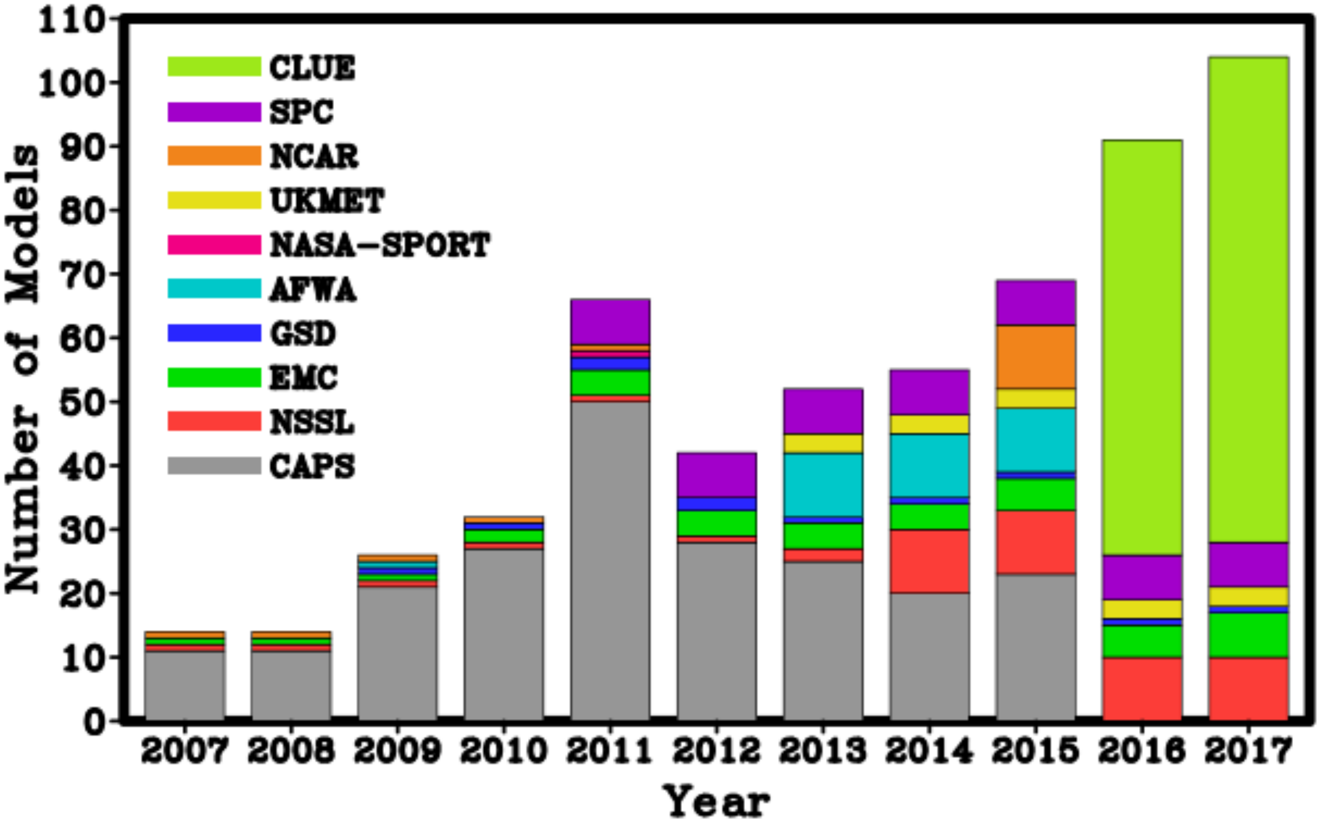


Figure 1 Number of CAMs run for SFEs since 2007. The different colored stacked bars indicate the contributing agencies.

More information on all the modeling systems run for SFE2017 is given below.

1) THE COMMUNITY LEVERAGED UNIFIED ENSEMBLE (CLUE)

The 2017 CLUE is a carefully designed ensemble with subsets of members contributed by NSSL, CAPS (OU), MAP (OU), GFDL, ESRL/GSD, and NCAR. In addition, the Developmental Testbed Center (DTC) provided support for post-processing, and configurations for NSSL runs that used stochastic physics. 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's 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 SFE2017 operations plan (Clark et al. 2017). Table 1 provides a summary of each CLUE subset. Note, because of last minute high performance computing issues, three of the ensemble subsets did not run in real-time. However, these were run after SFE2017 concluded to obtain a complete dataset for post-experiment analysis. The runs that were not produced in real-time are indicated in Table 1.

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-physcs, 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 stochastic physics (stoch-phys) ensembles. Thus, although the total number of members adds to 77, there were 75 unique members. The three ensembles marked with an asterisk were run in post-real-time because of last minute technical issues.

Clue Subset	# of mems	IC/LBC perturbations	Mixed Physics	Data Assimilation	Model Core	Agency
core	10	SREF	yes	ARPS-3DVAR	ARW	CAPS (OU)
single-phys*	10	SREF	no	ARPS-3DVAR	ARW	CAPS (OU)
caps-enkf	10	EnKF (CAPS)	yes	EnKF (CAPS)	ARW	CAPS (OU)
mp*	5	no	yes	ARPS-3DVAR	ARW	CAPS (OU)
stoch-phys*	10	SREF	no	ARPS-3DVAR	ARW	NSSL
HRRR36	1	no	no	RAP-GSI/DFI	ARW	ESRL/GSD
ncar-enkf	10	EnKF (DART)	no	EnKF (DART)	ARW	NCAR
gsi-enkf	10	EnKF (GSI)	no	EnKF (GSI)	NMMB	MAP (OU)
hrrre	9	EnKF	no	EnKF	ARW	ESRL/GSD
caps-fv3	1	no	no	cold start (GFS)	FV3	CAPS (OU)
gfdl-fv3	1	no	no	cold start (GFS)	FV3	GFDL

The design of CLUE allowed for 5 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 SFE2017.

Experiment Name	Description	CLUE subsets
Physics perturbation experiment	Three ensembles with perturbed ICs/LBCs were compared to test the effectiveness of different strategies for representing model error. One ensemble had single physics, one had mixed-physics, and one had single physics with stochastic perturbations.	core, single-phys, & stoch-phys
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
Data assimilation comparisons	3DVAR and several different 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, gsi-enkf, hrrre
Microphysics Sensitivities	The impact of different microphysical parameterizations on the resulting convective storm forecasts was examined.	mp
FV3	Two different version of FV3 were examined and compared to current well known CAMs (e.g., 3-km NAM, NSSL-WRF, etc.) to gauge performance at convective scales.	caps-fv3, gfdl-fv3

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 has provided a practical alternative to a formal/operational storm-scale ensemble, which did not become available until 1 November 2017 (the HREFv2). All members were initialized from the operational NAM or RAP models without additional data assimilation to produce the ICs.

3) HIGH RESOLUTION ENSEMBLE FORECAST SYSTEM VERSION 2 (HREFv2)

The HREFv2 is an 8-member, convection-allowing ensemble that was run in parallel during SFE2017. This version of HREFv2 was slightly different than the configuration implemented operationally by EMC on 1 November 2017, however, the performance of both are very similar. Even though half of the membership of HREFv2 consists of time-lagged runs, the design of HREFv2 closely follows that of the SSEO, which has demonstrated skill for the last five years in the HWT and SPC operations. All members, except for the NAM CONUS Nest, are initialized with a "cold-start". Forecasts to 36h are produced at 0000 and 1200 UTC. The HREFv2 is generally available before the SSEO, owing to moving the High Resolution Window (HiResW) runs earlier in the production suite more coincident with the NAM; however, 6-h old boundary conditions are used to allow for the earlier run time.

4) 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 fourth 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.

5) MET OFFICE CONVECTION-ALLOWING MODEL RUNS

Three nested, limited-area high-resolution versions of the Met Office Unified Model (UM) running once per day using 2.2 km grid-spacing were provided to SFE2017. 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 0000 UTC 17-km grid-spacing global configuration of the UM, the 2.2-km model was initialized without additional data assimilation and ran out to 120 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. There is no convection parameterization, and this model used the very latest UM model configuration internally designated Parallel Suite 39 (PS39). This has been extensively tested with parallel running and became the Met Office operational model configuration in June 2017.

The two experimental versions of the 2.2 km model used the “mid-latitude” and “tropical” variants of the new “Regional Atmosphere” (RA) configurations, designated RA1-M and RA1-T, respectively. The RA configurations are intended to provide standardised, portable, versions for use in other parts of the world on a longer (annual) development cycle than the internal UM. It is possible that RA1-M may not be that different from PS39 for the situations of interest in the HWT but it was run as a reference. RA1-M differs from PS39 in that aspects that are not easily portable to other parts of the world (e.g. the new urban scheme) are excluded. In addition, there are a number of other differences between PS39 and RA1-M such as turning on gravity wave drag.

The main differences between RA1-M and RA1-T are that the latter uses a prognostic scheme for cloud fraction (PC2), a larger Smagorinsky mixing length (0.5 of the grid-length compared to 0.2) and vertical resolution set with denser level spacing higher up in the atmosphere to take account of the deeper convection in the tropics. This configuration has been optimised for use in the maritime tropics but it is of interest to see how it compares to the mid-latitude configuration for convection over the continental US.

6) ESRL/GSD HIGH RESOLUTION RAPID REFRESH (HRRR) MODEL

The 3-km grid-spacing HRRR model developed by the ESRL/GSD continued to be examined in SFE2017. Both the NCEP operational HRRR (HRRRv2) and the ESRL developmental HRRR (HRRRv3) were evaluated. The developmental HRRR is scheduled to replace the operational HRRR in the spring of 2018. The HRRRv2 uses a 3-km grid with boundary conditions from the hourly updated, radar-DDFI-assimilated Rapid Refresh (RAPv3) model. The HRRR uses GSI hybrid data assimilation (instead of 3D-VAR), is initialized with the latest 3-D radar reflectivity and features a WRF-ARW core version 3.6.1 with Thompson microphysics. The operational HRRR is run every hour and produces hourly and sub-hourly forecasts out to 18 h.

The HRRRv3 runs every hour with output to 18-h (0100, 0200, 0400, 0500 UTC, …) or 36-h (0000, 0300, 0600 UTC…). The experimental HRRRv3 remains on a 3-km grid with hourly runs that are changed to the forecast lengths listed above. The HRRRv3 is initialized with an hour of 3-D radar reflectivity using a latent-heating specification technique including some refinements in this latent-heating from the parent RAPv4 model. The HRRRv3 uses grid-point statistical interpolation (GSI) hybrid GFS ensemble-variational data assimilation of conventional observations. Building upon the advancements in the operational HRRRv2 at NCEP, HRRRv3 includes assimilation of TAMDAR aircraft observations, refines assimilation of surface observations for improved lower-tropospheric temperature, dewpoint (humidity), winds and cloud base heights, and places more weight on the ensemble contribution to the data assimilation. HRRRv3 adds assimilation of lightning flash rates as a complement to radar reflectivity observations through a similar conversion to specified latent heating rates during a one-hour spin-up period in the model. HRRRv3 also contains numerous model changes including an update to WRF-ARW version 3.9 and the Thompson microphysics, transition to a hybrid sigma-pressure vertical coordinate for improved tropospheric temperature, dewpoint and wind forecasts, along with a higher resolution (15 second) land use dataset. Physics enhancements have also been made to the MYNN planetary boundary layer (PBL) scheme and RUC land surface model along with additional refinements to shallow cumulus/sub-grid-scale cloud parameterizations including enhanced interactions with the radiation and microphysics schemes for greater retention of cloud features.

7) HIGH RESOLUTION RAPID REFRESH ENSEMBLE (HRRRE)

In addition to the 0000 UTC initialized HRRRE runs that were a part of the 2016 CLUE, HRRRE forecasts were also provided at 1200, 1500, and 1800 UTC. These forecasts went out 18 h and were configured similarly to the 0000 UTC initializations. The experimental HRRRE consists of nine 3-km grid-spacing forecast members covering about 55% of the CONUS HRRR domain. The HRRRE is initialized at 0900 UTC each day from a combination of atmospheric RAPv4 mean and GFS data assimilation ensemble (GDAS) perturbations along with HRRRv3 land surface data. A total of 36 3-km HRRR members are initialized and then cycled hourly through 0000 UTC using an Ensemble Kalman filter to assimilate conventional and radar observations each hour followed by the application of the HRRR cloud analysis and soil adjustment to each member. At 0000 UTC, nine members produce 36-h forecasts. Stochastic soil moisture perturbations are introduced across all members at 0900 UTC and boundary layer parameter perturbations are applied at 0000 UTC along with lateral boundary perturbations and inflation during the cycled data assimilation to promote spread and represent both initial condition and model forecast uncertainties. The HRRRE uses WRF-ARW version 3.9 with the same physics configuration as the HRRRv3.

8) NSSL EXPERIMENTAL WARN-ON-FORECAST SYSTEM FOR ENSEMBLES (NEWS-E)

The NSSL Experimental Warn-on-Forecast System for ensembles (NEWS-e) is a 36-member WRF-based ensemble data assimilation system that was used to produce very short-range (0-4 h) probabilistic forecasts of supercell thunderstorm rotation, hail, high winds, and flash flooding. The starting point for the NEWS-e was the experimental HRRRE, provided by ESRL/GSD. The full ensemble is updated by hourly EnKF assimilation of conventional observations and Multi-Radar/Multi-Sensor (MRMS) radar reflectivity from 1000 UTC Day 1 to 0000 UTC Day 2. A 15-h ensemble forecast launched from the 1500 UTC HRRRE analysis is used to provide boundary conditions for the NEWS-e system for the period 1800 UTC Day 1 – 0300 UTC Day 2. Similarly, a 1-h ensemble forecast launched from the 1700 UTC HRRRE analysis is used to provide initial conditions for the NEWS-e system at 1800 UTC.

The daily NEWS-e domain location targeted the primary region where severe weather was anticipated, and covered a 1000-km wide region with very frequent 15-min updates. All ensemble members utilize the NSSL 2-moment microphysics parameterization and the RAP land-surface model, but the PBL and radiation physics options are varied amongst the ensemble members to address uncertainties in these model physics. Multi-Radar/Multi-Sensor (MRMS) radar reflectivity and Level II radial velocity data, cloud water path retrievals from the GOES-13 imager, and Oklahoma Mesonet observations (when available) were assimilated every 15 min using an EnKF approach, beginning at 1800 UTC each day. High-frequency ASOS were also assimilated at 15 and 45 minutes past each hour. A 4-h ensemble forecast was initialized from the 1900 UTC NEWS-e analysis for HWT product evaluation from 2000-2100 UTC. Then, beginning at 2000 UTC, a 180-min (90-min) ensemble forecast with 5-min output was launched at 00 (30) minutes past the hour, through 0300 UTC the next day. These forecasts were displayed in the web-based NEWS-e Forecast Viewer (http://www.nssl.noaa.gov/projects/wof/news-e/images.php) or the Probabilistic Hazard Information (PHI)-tool developed by NSSL.

b) Daily Activities

SFE2017 activities were focused on forecasting severe convective weather at two separate desks, one forecasting individual hazards (Severe Hazards Desk) and the other forecasting total severe (Innovation Desk), with different experimental forecast products being generated at different temporal resolutions. Forecast and model evaluations also were an integral part of daily activities. A summary of forecast products and evaluation activities can be found below while a detailed schedule of daily activities is contained in the appendix (Table A1).

1) EXPERIMENTAL FORECAST PRODUCTS

Similar to previous years, the experimental forecasts continued to explore the ability to add temporal specificity to longer-term SPC severe weather outlooks. The Severe Hazards Desk mirrored the SPC operational Day 1 severe weather 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. At the Innovation Desk, a separate Day 1 forecast was made for total severe (combined hail, wind, and tornado) probabilities valid over the same period. On each day, experimental forecasts were made for a re-locatable mesoscale region of interest where the greatest severe weather potential and/or specific forecasting challenges were identified.

Each desk then manually stratified their respective Day 1 forecasts into periods with higher temporal resolution. At the Severe Hazards Desk, individual 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 Innovation Desk drew hourly areas delineating the region(s) of anticipated severe weather occurrence (or geographic coverage of severe reports), followed by hourly isochrones of severe weather that delineated the start-time of the 4-h time window with the highest total severe probabilities. For example, an area encompassed by the 1800 and 1900 UTC isochrones would expect the start time of the 4-h time window with the highest severe weather probabilities to fall between 1800 and 1900 UTC. The isochrones were only drawn within areas where the experimental Day 1 total severe probabilities were 15% or greater. The isochrones activity was conducted for the second consecutive year to test methods for adding more detailed timing information to outlooks 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 short-term 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 comprehensive 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 3-h periods. The 1600-1200 UTC human forecasts for the Severe Hazards 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 Severe Hazards Desk, participants created their own short-time-window forecasts 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 at that desk using N-AWIPS. At the Innovation Desk, a separate lead forecaster drew the hourly areas of expected severe weather followed by hourly isochrones using the N-AWIPS machine. Meanwhile, participants split into five groups and issued the same type of timing outlooks, except using the web-based tool.

In the afternoon, experimental 24-h severe weather forecasts were also generated for Day 2 valid 1200-1200 UTC 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 experimental, longer-range CAMs such as FV3, Met Office, and the Model for Prediction Across Scales (MPAS).

Finally, the Severe Hazards 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 for the 2200-0200 UTC period. Unlike previous years, the Innovation Desk did not update their forecasts made earlier in the day. Instead, a forecasting activity using the WoF-prototype system, NEWS-e, was conducted. For this activity, the 1900 UTC initialized NEWS-e with 4-h forecasts was used to issue two 1-h time window forecasts of total severe valid 2100-2200 and 2200-2300 UTC. Then, these forecasts were updated using 2000 UTC initialized NEWS-e products. Innovation desk contributors also participated in a one-time scientific survey that explored interpretation of NEWS-e products.

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 and complementary component of SFE2017 was to look back and evaluate the forecasts and model guidance from the previous day. The former activity enables comparison of the perceived utility of various operational and experimental guidance systems as part of a simulated forecasting process, whereas the latter activity permits assessment of guidance performance from a post-event perspective. There were two periods of formal evaluations during SFE2017. 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 were compared to observed radar reflectivity, local storm reports (LSRs), 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 SFE2014-16, 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 calculate 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, grids were filtered spatially to ensure the presence of contiguous swaths in the MESH grids (Melick et al. 2014).

The second evaluation period occurred during the afternoon and focused on comparisons of different ensemble diagnostics and CLUE ensemble subsets. The Innovation and Severe Hazards Desks conducted two different sets of afternoon evaluations.

3. Preliminary Findings and Results

a) Evaluation of experimental forecast products – Innovation Desk

1) CONVECTIVE OUTLOOK EVALUATIONS

SFE2017 participants subjectively evaluated the previous day 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*." An example image used to conduct full period ratings is shown in Figure 2. This forecast was made the morning of 17 May and verified the next day. All six participants that rated this forecast assigned it 9/10.

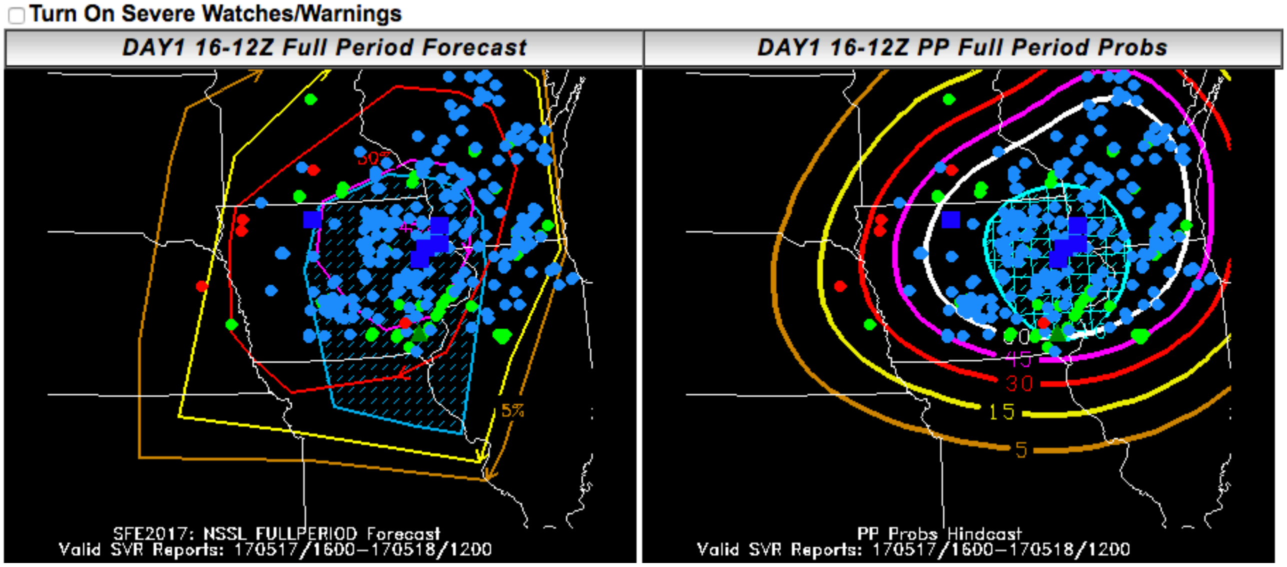

□ Turn On Severe Watches/Warnings
DAY1 16-12Z Full Period Forecast
DAY1 16-12Z PP Full Period Probs
30%
5%
SFE2017: NSSL FULLPERIOD Forecast
Valid SVR Reports: 170517/1600–170518/1200
60
45
30
15
5
PP Probs Hindcast
Valid SVR Reports: 170517/1600–170518/1200

Figure 2 Left panel: Experimental Day 1 outlook for total severe weather valid 1600 – 1200 UTC 17-18 May 2017 with locations of storm reports overlaid. Right panel: Practically perfect hindcast probabilities with the locations of storm reports overlaid.

The Day 1 full period forecasts were valid 1600 UTC – 1200 UTC, while the Day 2 and Day 3 forecasts covered the 1200 UTC – 1200 UTC time period. Day 2 forecasts were issued Monday through Thursday, but Day 3 forecasts were only issued as time permitted, typically when Day 3 appeared to have higher severe weather potential than Day 2 after considering the numerical guidance. Generally, forecasts performed well, with a median of 7.0/10.0 for each of the forecasts (Fig. 3). The performance of the daily forecasts also had similarly shaped distributions indicating that on some days the forecasts often had room for improvement. In the comments, participants cited both the location and magnitude of the probabilities as reasoning for their scores. Participants also noted that total severe forecasts were occasionally correct for the wrong reasons (e.g., anticipating a high wind threat and verifying the probabilistic contours with mostly hail reports instead). Debates also arose regarding how to rate the Day 2 and Day 3 forecasts compared to the Day 1 forecast. For example, some participants felt that Day 3 forecasts should be rated relative to the typical level of skill at Day 3, which could mean that a forecast with a rating of 9 for Day 3 may have only received a rating of 6 for Day 1, since the typical level of skill at Day 3 is lower. However, others felt the forecasts should be rated without consideration of the lead time, which would mean that a forecast with a rating of 9 for Day 3 would have also received a rating of 9 if it was valid for Day 1. As a result of these debates, the phrasing of this question will be clarified in future SFEs.

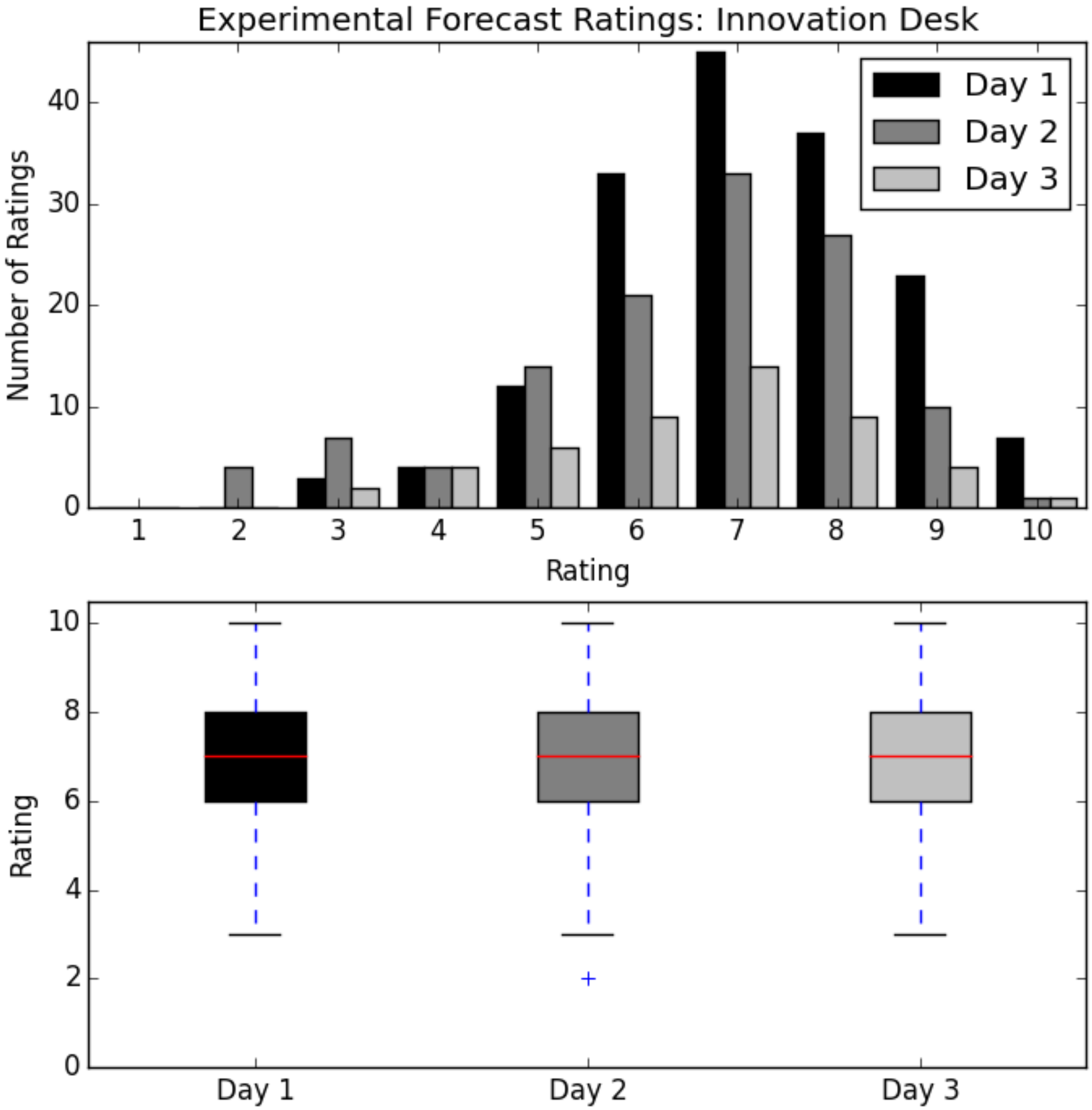


Figure 3 Distribution of participant ratings of experimental full-day total severe forecasts issued at three different lead times.

Participants were also asked what impact longer-range CAM guidance (beyond Day 1) had on the experimental forecasts, as SFE 2017 had a number of CAMs extending through the Day 2 forecast period, and at least five CAMs that extended to 120 h, meaning that they were available for the Day 2 and the Day 3 forecast periods. Participants overwhelmingly stated that the longer-range CAM guidance either improved (52%) or had no impact (30.6%) on their forecasts (Fig. 4), supporting continued efforts to extend the range of CAM guidance. Many of the benefits of CAMs seen in the near-term, such as initiation and mode guidance, help the longer-range forecasts. However, occasional difficulty with how the CAMs handled overnight convection warrants caution in circumstances where overnight convection is expected to affect the subsequent convective development during the next diurnal cycle, and this topic deserves closer study.

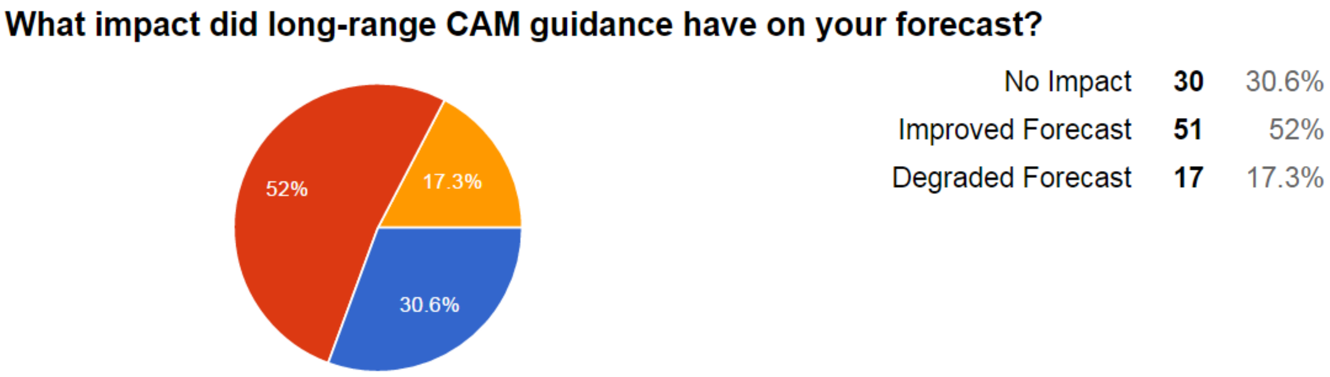


Figure 4 Participant responses to the impact of longer-range CAM guidance on the experimental full-period Day 2 and Day 3 forecasts.

2) ISOCHRONE EVALUATION

Finally, the Innovation Desk participants and lead forecaster drew hourly areas of expected severe reports, followed by isochrones of severe weather at hourly intervals to delineate the start-time of the 4-h time window with the highest total severe probabilities. Previous work has shown that a majority (97%) of reports within 40 km of a point will fall within a 4-h period of the 24-h convective outlook day (i.e., 1200-1200 UTC). Thus, researchers formulated an experimental forecasting activity based on this finding to SFE2017 to test human ability to predict the start time of this most active 4-h period. Similar to SFE2016, forecasters were instructed to draw isochrones on a map to indicate the start time of the 4-h period that would capture most of the severe reports for that day. However, the training and process for this experiment differed from that for SFE2016 in that it was much more in-depth and comprehensive. Participants were only asked to draw isochrones within the 15% total severe areas since events associated with lower probabilities are typically less coherent and focused, and thus more difficult to assign timing information. After the full day 1600-1200 UTC total severe outlook was drawn, participants drew hourly report areas within the 15% total severe forecast area (e.g., Fig. 5). These hourly report areas helped participants identify the severe threat areas and spatiotemporal evolution of potential severe reports. After the hourly areas were delineated, participants moved on to drawing isochrones on top of the areas (e.g., Fig. 6). This breakdown of the forecasting process helped participants understand the concept and forecast more accurate timing products than in SFE2016.

In the next-day subjective evaluations, participants rated the hourly report area forecasts from the desk lead on a scale of 1 (Very Poor) to 10 (Very Good). The distribution of these ratings (Fig. 7) indicate that forecasts were typically best at the start of the forecast period associated with shorter lead times, as well as toward the end of the period. The latter result can be explained because forecasts at the end of the period were often comprised of correct nulls, due to the cessation of severe convection. However, all hours except the 0300–0400 UTC period had median ratings higher than a 5.0/10.0, suggesting that hourly identification of geographic areas likely to experience severe convection was possible. This result is consistent with previous experiments, which found that hourly probabilistic forecasts were quite reliable (Gallo et al. 2017). Additionally, having the participants draw hourly report areas rather than hourly probabilistic forecasts allowed participants to issue their forecasts much more quickly than in previous SFEs that required forecasters to stratify graphical forecasts into hourly time periods. Large variability is also seen in the ratings, with the distributions for almost every hour containing both the highest (10) and the lowest (1) rating. The variability in these ratings highlights the challenge of issuing forecasts valid for hourly periods, when the correct timing of storm initiation and evolution is critical to forecast success. Participants noted this difficulty in their comments, which often mentioned small displacements between the report occurrence and the location of the forecast area, or the challenge of having multiple distinct areas of threat within the domain. In the end, however, participants also seemed to like the high temporal resolution forecasts, with many positive comments about the opportunity to issue experimental forecasts valid for hourly periods.

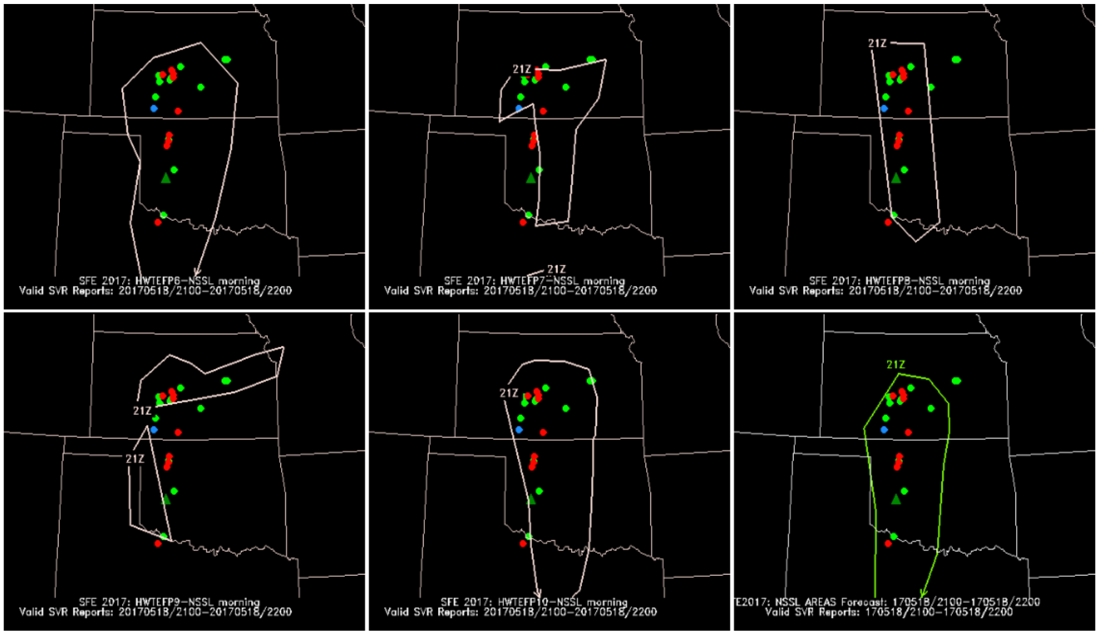

SFE 2017: HWTEFP6-NSSL morning
Valid SVR Reports: 20170518/2100-20170518/2200
21Z
SFE 2017: HWTEFP7-NSSL morning
Valid SVR Reports: 20170518/2100-20170518/2200
21Z
SFE 2017: HWTEFP8-NSSL morning
Valid SVR Reports: 20170518/2100-20170518/2200
21Z
21Z
SFE 2017: HWTEFP9-NSSL morning
Valid SVR Reports: 20170518/2100-20170518/2200
21Z
SFE 2017: HWTEFP10-NSSL morning
Valid SVR Reports: 20170518/2100-20170518/2200
21Z
E2017: NSSL AREAS Forecast: 170518/2100-170518/2200
Valid SVR Reports: 170518/2100-170518/2200

Figure 5 Areas of expected report occurrence valid 21-22 UTC 18 May 2017. The lead forecaster of the Innovation Desk generated the bottom right panel, while the other panels display participant-generated forecasts.

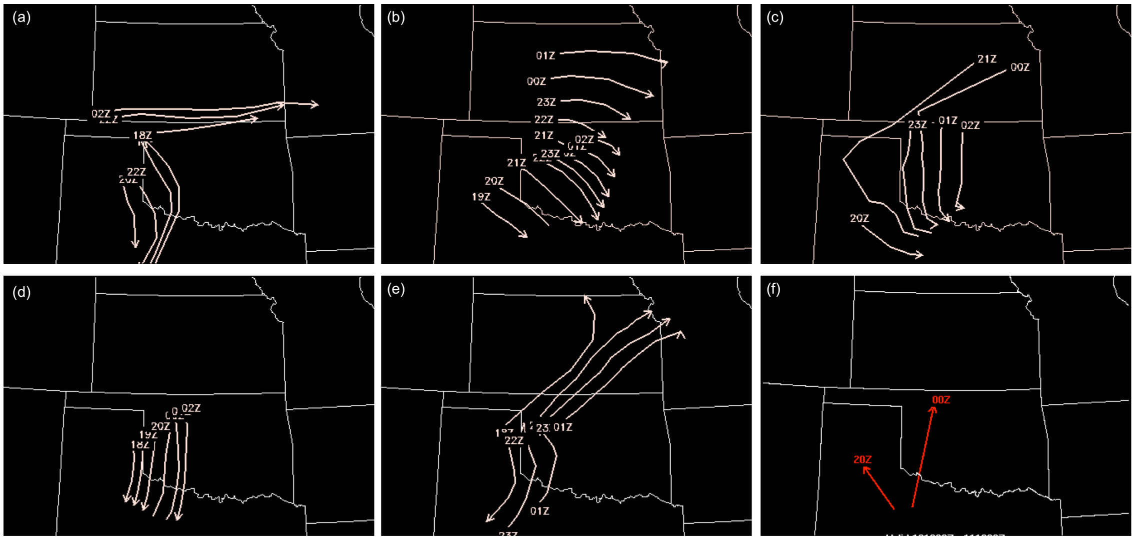


Figure 6 Isochrone forecasts for 10 May 2017. The lead forecaster of the Innovation Desk generated the bottom right panel with the red isochrones, while all other panels display forecasts generated by participants.

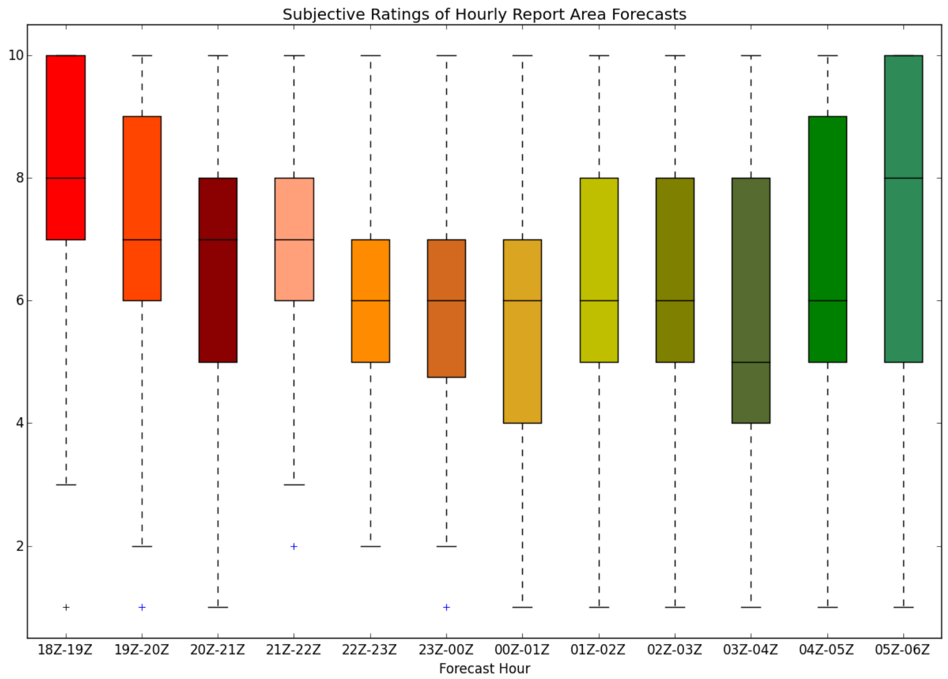


Figure 7 Subjective ratings of hourly report area forecasts issued by the Innovation Desk Lead Forecaster.

Next-day subjective evaluations of isochrones were verified against objective isochrones (see below) that were created based on the time of report occurrence, using a scale of 1 (Very Poor) to 10 (Very Good). Participants rated the Desk Lead's isochrones 6/10 or higher most of the time (Fig. 8), which follows from the skill shown in the hourly forecasts (Fig. 7). If the forecaster is able to delineate the hourly report areas, the four-hour period when the most reports will occur should be derived more directly, without first drawing the areal evolution of the report progression. Therefore, the skill level in the hourly area forecasts is reflected in the skill in the isochrones.

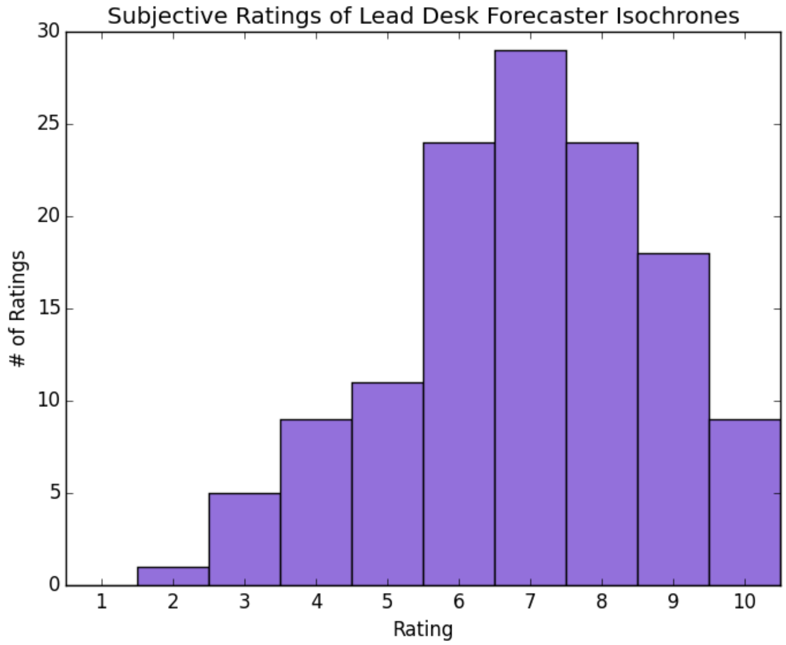


Figure 8 Subjective ratings for isochrones issued by the Innovation Desk Lead Forecaster.

Objective isochrone verification was performed by plotting the forecaster-produced isochrones on an 80 km grid and comparing that grid to objectively-generated isochrones based on observed severe wind, hail, and tornado reports. These reports were plotted on the grid corresponding to the time of occurrence (grids were created for 1800-2200, 2000-0000, 2200-0200, 0000-0400, and 0200-0600 UTC) and then smoothed spatially using a Gaussian kernel with a smoothing parameter (σ) of 120 km. Then each grid point was assigned the time period with the highest smoothed probability, creating areas of timeframes that were contoured analogously to isochrones.

Results from SFE2017 showed that participants had a much easier time understanding the concept behind the isochrones experiment and what was being tested (i.e., whether forecasters could identify the 4-h period with the most reports) compared to SFE2016. While the majority of forecasters saw this product being potentially helpful for emergency managers and members of the public, there was still concern over whether or not isochrones were the best way to display this type of timing information (see comments in Table 3). Compared to SFE2016, the accuracy of the isochrones created by the lead forecaster increased in SFE2017, especially over the last two weeks of the 2017 experiment (Fig. 9). Many more of the forecasted points fell within the one-hour-early to one-hour-late categories than in SFE2016. For future applications, there are plans to alter the visualization of the product to address concerns about its usability by forecasters and partners.

Table 3 Selected quotes from participants after the isochrone experiment.

In its current form, this product is rather complicated, and likely something that non-meteorologist users would not comprehend without training…I am afraid this product would not be beneficial without some adjustments (e.g. polygons instead of isochrones).	*I do not mean to sound too critical of this project. I think adding timing information is a great idea, and isochrones may still be the best way to do it, but in its current state, I find it somewhat confusing.*
This is a great idea as many customers to WFOs are really interested in both if severe weather will occur and when it will occur. In my office, showing timing spatially has been very challenging so I am glad to see there is an effort to spatially depict the timing for weather threats.	*These are great in more straight-forward scenarios but there are times when…it can be nearly impossible to convey timing... This product has tremendous value for the users though! I look forward to seeing how this will involve. Timing is one of the most frequently asked questions we get from users so there is definitely a need*
I could see the utility of using this for not only the forecast but also for emergency management briefings.	*I'd show animations of 4 hour threat polygons instead of isochrones.*

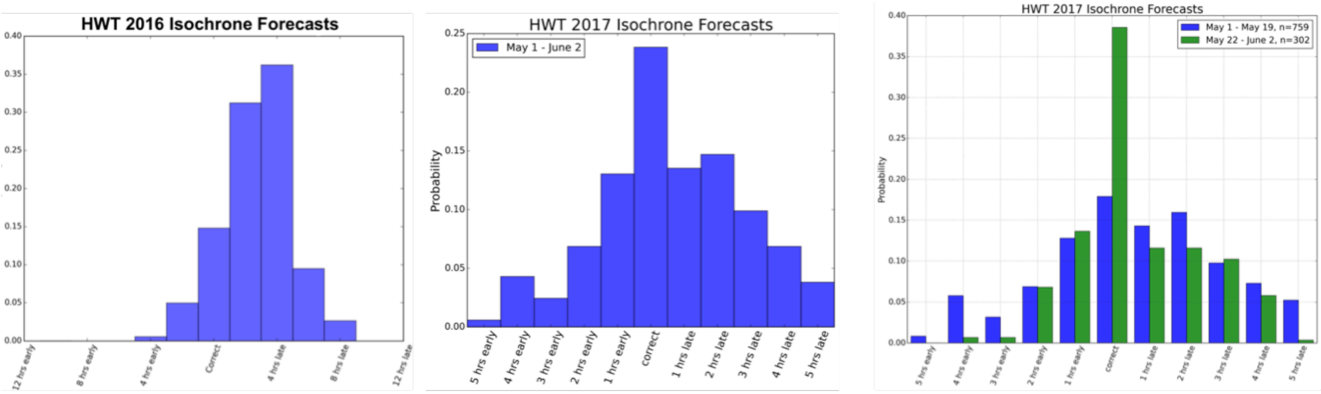


Figure 9 Verification of isochrone forecasted points by the Innovation Desk lead forecaster for the SFE2016 full period (left), SFE2017 full period (center), and SFE2017 broken down into forecasts during the first three weeks, and last two weeks of the five-week experiment (right).

Automated isochrone forecast guidance was also generated using the NSSL-WRF ensemble for the second consecutive year. 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. As in SFE2016, this automated product was not formally evaluated, but was generally well received as potential guidance.

3) NEWS-E EVALUATIONS

The NSSL Experimental Warn-on-Forecast (WoF) System for ensembles (NEWS-e; Wheatley et al. 2015, Jones et al. 2016) was a new addition to the SFE2017 activities. This prototype WoF system is a frequently updated, regional-scale, on-demand convection-allowing ensemble analysis and prediction system, nested within an experimental hourly CAM ensemble forecast system (currently HRRRe). This system produces 0–4-h predictions of individual convective storms and mesoscale environments that provide probabilistic forecast guidance, such as the probability of simulated reflectivity above a threshold at a grid point, and ensemble percentile values (e.g., 90th) of fields such as accumulated rainfall, 2–5-km updraft helicity, and 0–2-km vertical vorticity. Participants contributed to two activities assessing NEWS-e products during their time at the Innovations Desk. These two activities will be described as the “Survey Activity” and the “Outlook Activity”.

Survey Activity

Before participating in the outlook activity the participants completed an online survey on their interpretation of probabilistic forecast products. Operational and experimental CAM ensembles and probabilistic forecast guidance are becoming increasingly available to forecasters and as a result, the paradigm for interpreting forecast guidance is evolving from one that is deterministic to one that is probabilistic. The strengths and limitations of CAMs were being tested and evaluated within several experiments during SFE2017 with a focus on verification statistics and subjective evaluations. While these evaluations provide information useful for evidence-based decision-making, they do not provide insight into how probabilistic forecast guidance is interpreted or used by meteorologists. Therefore, the goal of the survey was to sample and document meteorologists' interpretations of probabilistic forecast guidance through various types of NEWS-e products. More information about the survey activity is found in Appendix B.

Outlook Activity

Once the survey was completed, participants were asked to join the lead forecaster at the Innovation Desk for the outlook activity. The primary goals of this part of the experiment were to explore how short-term ensemble forecast guidance from NEWS-e could be used by the lead forecaster to produce a series of 1-h severe weather outlooks and observe how the forecaster's understanding, use, and attitudes about NEWS-e guidance evolved through the experiment. Each morning, subjective verification of the previous afternoon outlooks was performed by comparing them to "practically perfect" hindcasts.

The outlook activity consisted of producing two 1-h outlooks of severe probabilities over the NEWS-e domain (decided jointly by the WoF researchers and SPC forecasters) between 2100-2200 and 2200-2300 UTC. These outlooks were produced using only the 1900 UTC NEWS-e 4-h forecast (valid 1900-2300 UTC) and then updated using the 2000 UTC NEWS-e 3-h forecast (valid 2000-2300 UTC) along with current observations including radar, satellite, and surface observations. An overview of the 2017 NEWS-e configuration is provided in Figure 10. Initial outlooks (produced from the 1900 UTC forecast) were submitted to an internal database by 2030 UTC and updated outlooks (produced from the 2000 UTC forecast) were submitted by 2100 UTC, resulting in four total outlooks. During the hour-long experiment, several forms of data were collected. These included the four outlooks, a recording of the lead forecaster's screen where the NEWS-e forecast viewer was being used, observer notes, and a "daily wrap-up" worksheet completed each day by the lead forecaster.

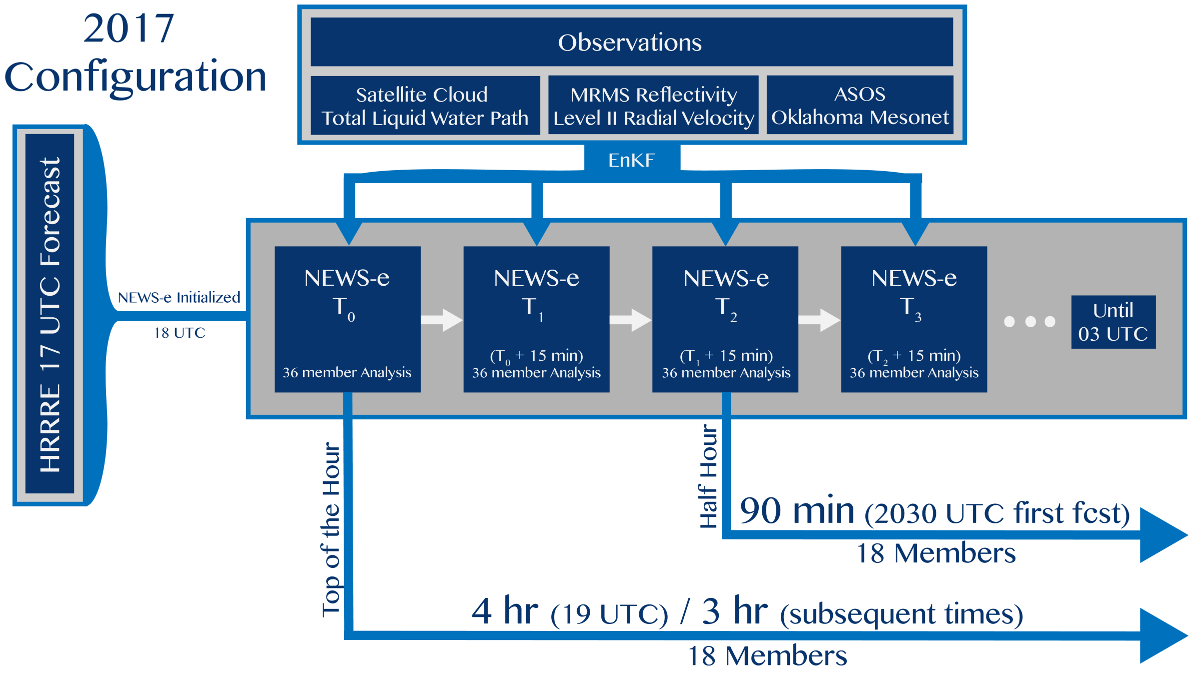


Figure 10 The 2017 NEWS-e configuration. 1900 UTC was the only 4-h forecast. 90 minute forecasts began at 2030 UTC.

As this was the first time NEWS-e has been used to issue forecasts in real-time, the outlook activity evolved over the course of the SFE to accommodate participant requests and suggestions. First, technical issues resulting in delayed NEWS-e forecasts prevented completion of all outlooks for 17 cases, which are marked "incomplete". Thus, the sample size for the outlook evaluations is somewhat limited. Additionally, practically perfect hindcasts, initially developed for full period SPC outlooks, were tuned to better reflect a higher level of precision that should be possible with very short lead-time and short time window outlooks. Specifically, instead of the standard smoothing level of σ = 120 km used to verify full period severe weather outlooks, the NEWS-e activity switched to σ = 40 km, which resulted in higher probabilities over smaller areas. Lastly, forecast guidance products were regularly updated throughout the 5 weeks to accommodate participant requests.

Examining the number of missed reports (i.e., reports within the forecast domain that were not associated with severe weather probabilities) revealed the same number of misses for the initial and updated outlooks valid for the 2100-2200 UTC time period. However, for the 2200-2300 UTC time period updated outlooks there were slightly fewer misses than the initial (Fig. 11). The lead forecaster at the Innovation Desk commented, "In probably about half of the cases, either only minor or no changes were made to the update. However, when changes were made, the update typically resulted in a better forecast using the latest NEWS-e run." The subjective ratings of the four sets of severe weather outlooks had distributions very similar to one another with a median rating of 7/10 (Fig. 12).

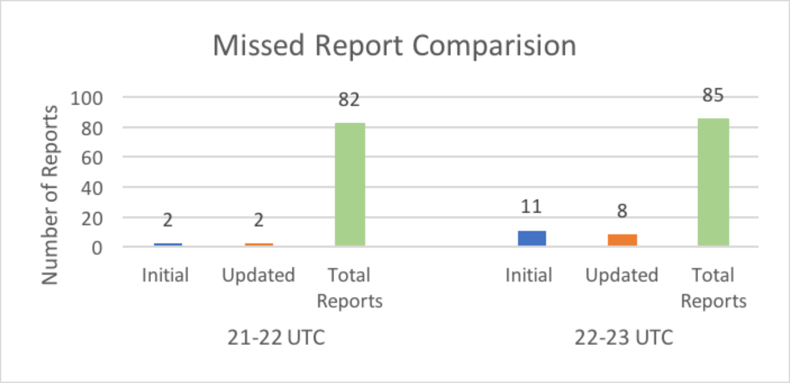


Figure 11 Preliminary missed reports. "Total reports" refers to all reports within the NEWS-e domain.

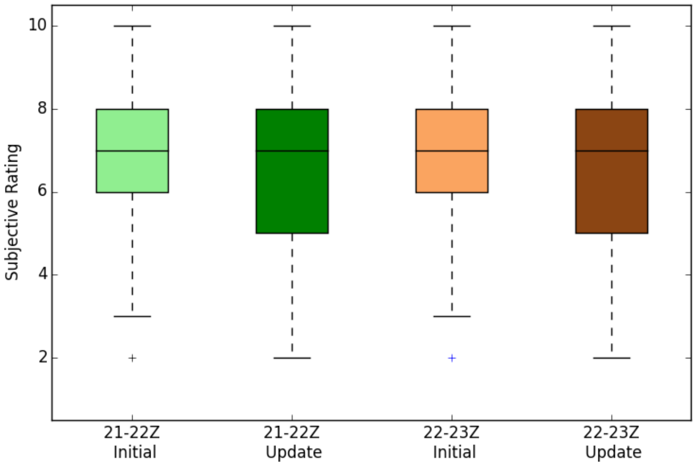


Figure 12 Boxplots of the ratings distributions for the hourly probabilistic forecasts issued using the NEWS-e guidance.

Variation in NEWS-e performance was noted across different cases as well as different aspects of individual forecasts. Object-based verification of NEWS-e forecasts using observations from the Multi-Radar Multi-Sensor system reveals more consistency in general thunderstorm forecasts, using composite reflectivity, than mesocyclone forecasts, using 2-5 km AGL rotation tracks (Fig. 13). This variation in NEWS-e forecasts of severe thunderstorms is evident comparing cases where NEWS-e performed particularly well, including a "tornado case" on 16 May and a "wind case" on 17 May, to cases where NEWS-e did not perform as well, including a "missed case" on 23 May and a "false alarm" case on 26 May. Possible explanations for degraded forecast performance include timing errors, missed convection initiation, or other factors such as forecaster influence from earlier CAM output, or forecast deadline time pressures. One set of outlooks [initial, updated, and practically perfect (PP)] from these four cases is shown in Figure 14. Despite the variations in forecast quality, the lead forecaster commented in his final report that, "On most days, NEWS-E was exceptional at identifying which storms had the greatest potential to become severe."

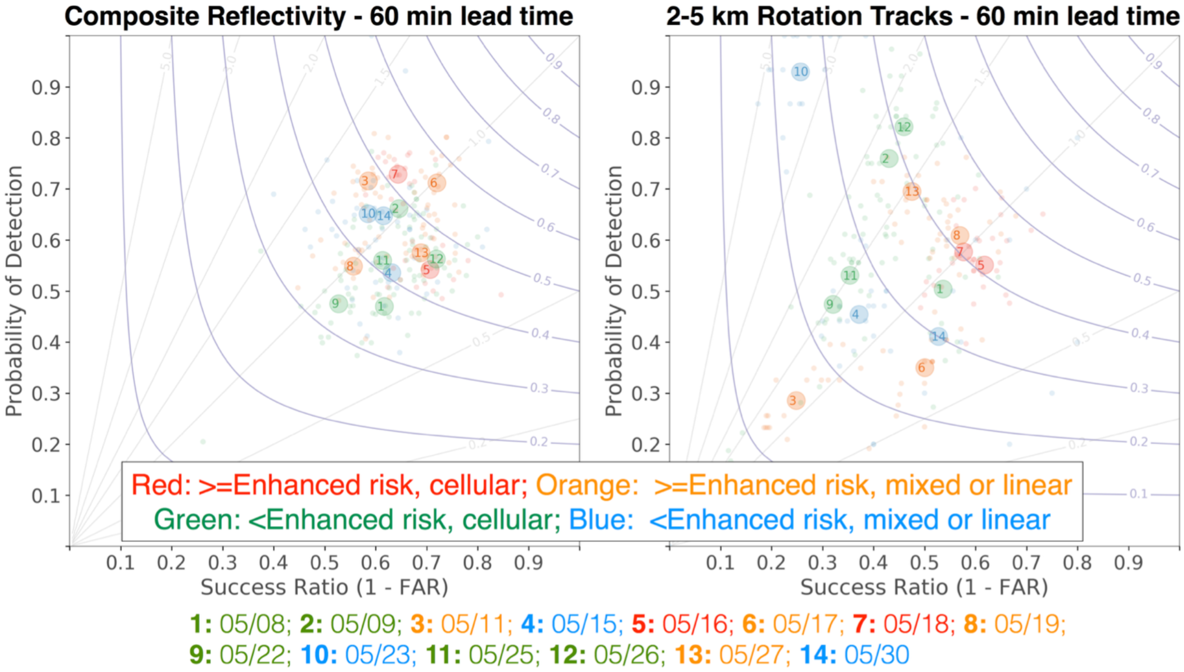


Figure 13 Performance diagrams of object-based NEWS-e verification for (left) composite reflectivity and (right) 30-minute rotation tracks for 14 cases during May 2017. NEWS-e forecast objects are verified against corresponding objects in MRMS observations. Cases are color coded according to maximum SPC categorical risk in the 1630 UTC update and subjectively identified storm mode. Ensemble mean values are plotted as large circles with individual members small circles. The number inside each ensemble mean value corresponds to the case dates listed below the plots.

Ongoing research includes quantitative analysis of local storm reports within 20 miles from contoured outlooks, direct comparisons to PP outlooks, and analysis of the forecaster's screen recordings to track changes in product usage as familiarity with products unique to NEWS-e increased. As a final thought, the lead forecaster stated, "*The goal is to provide more information on severe weather trends between the watch and warning. In my opinion, NEWS-E has successfully accomplished this mission and is already a tool that would be a great aid to WFO warning operations as well as providing storm-scale guidance for Storm Prediction Center Mesoscale Discussions.*"

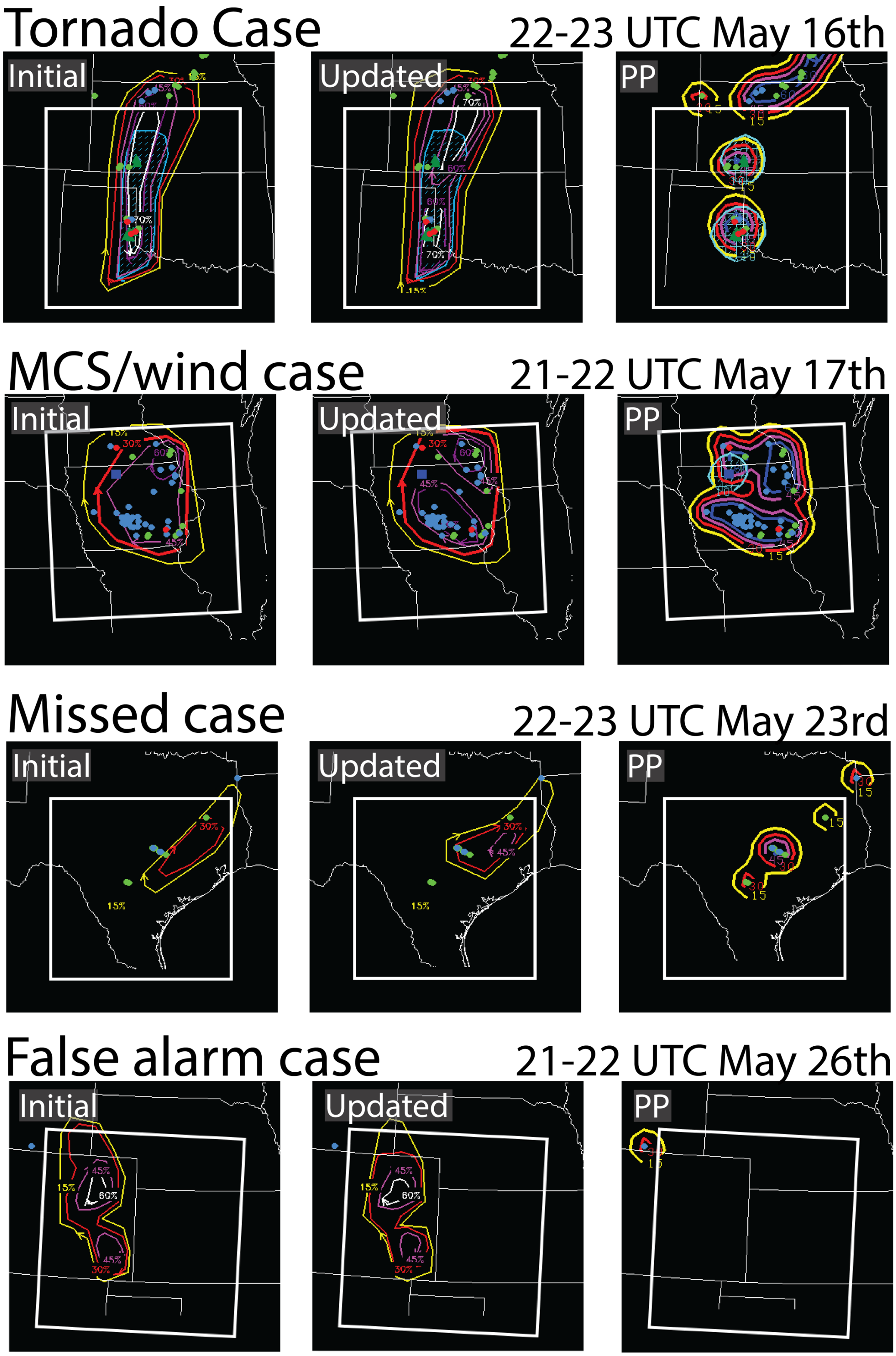


Figure 14 Example cases including a tornado case, MCS/wind case, missed case, and false alarm case, respectively. "PP" is the "practically perfect" hindcast used every morning for subjective verification.

b) Evaluation of experimental forecast products – Severe Hazards Desk

1) DAY 2 FORECASTS

SPC currently issues probabilistic forecasts of total severe (including all hazards) in the Day 2 Convective Outlook. With an increasing number of experimental and operational CAM forecasts extending into the Day 2 period, there is additional guidance regarding convective mode and storm intensity to assist in generating individual hazard forecasts. To explore creating this type of forecast product, experimental Day 2 probabilistic forecasts of individual severe weather hazards were generated during SFE2017, using traditional operational model guidance as well as experimental CAMs and CAM ensembles. These Day 2 individual hazard forecasts were then evaluated and compared with their respective Day 1 forecasts valid for the same convective day (Fig. 15). Overall, the Day 1 forecasts were rated higher than the Day 2 forecasts (except for wind) as expected, but there is a large overlap in the forecast ratings. These results suggest that useful forecasts can be made for individual severe storm hazard forecasts for Day 2 during the spring.

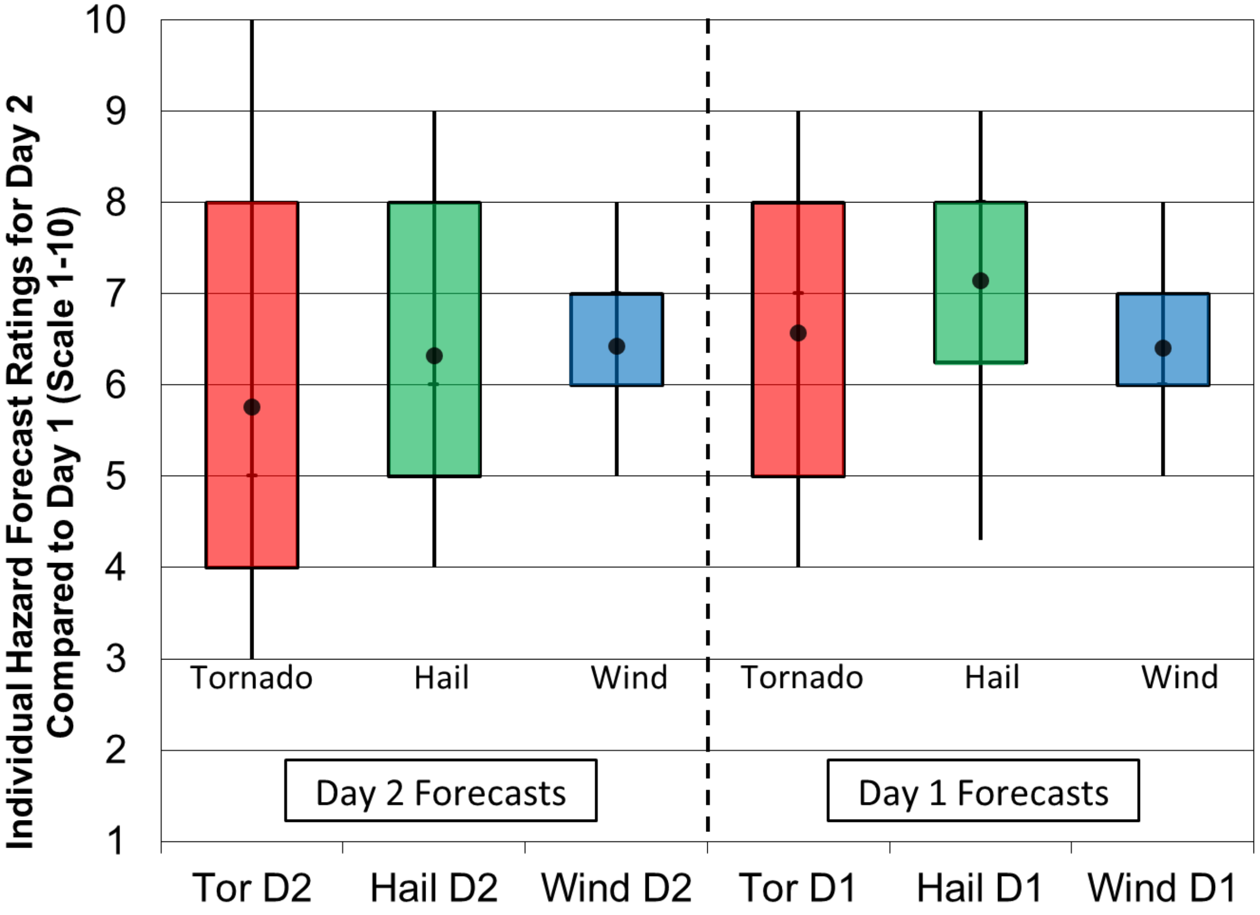


Figure 15 Distribution of subjective ratings (1-10) for experimental probabilistic tornado (red), hail (green), and wind (blue) outlooks for Day 2 (left) and Day 1 (right). The boxes span the interquartile range while the whiskers extend to the 10th and 90th percentiles. The horizontal dash (-) indicates the median rating, and the circle (●) indicates the mean rating.

2) 4-H FORECASTS

Experimental 4-h probabilistic outlooks were also generated for the Day 1 period during SFE2017. First-guess 4-h probabilities of severe hazards (i.e., tornado, hail, and wind) 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 SSEO/SREF calibrated probabilities (Jirak et al. 2014). These first-guess probabilities were available during the forecast process and then compared in the next-day evaluation to the forecaster-issued probabilities, providing an indication of how much a forecaster can improve upon the 4-h first-guess guidance. An example 4-h time window forecast for tornadoes in shown in Figure 16. During the 1800-2200 UTC period, forecasters were generally able to improve upon the disaggregated first-guess guidance for tornadoes and hail while the wind forecasts were rated about the same as the first-guess guidance (Fig. 17). In general, the overlapping distribution of ratings suggests that the guidance can provide useful first guess information that can be improved upon by forecasters.

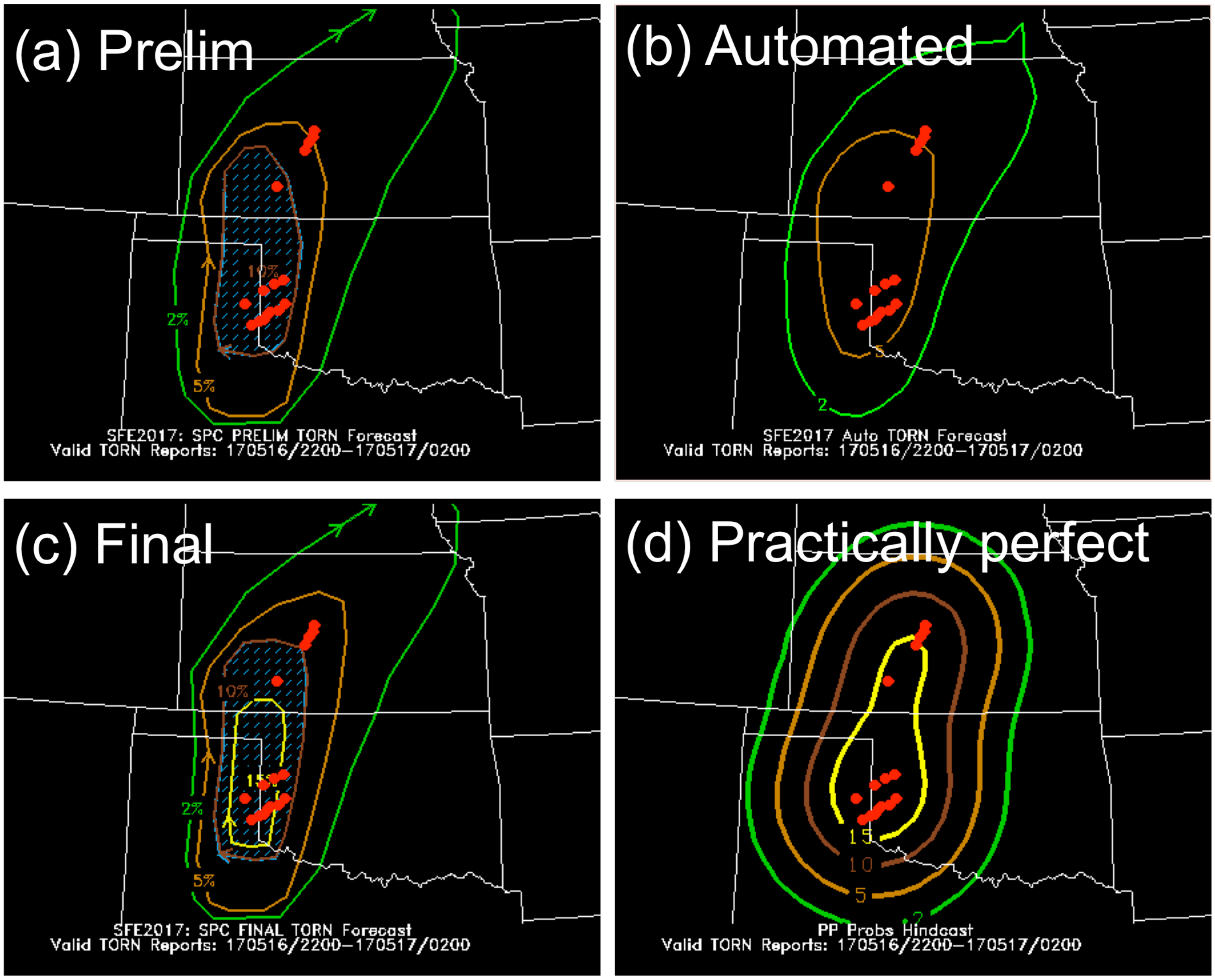

(a) Prelim
10%
2%
5%
SFE2017: SPC PRELIM TORN Forecast
Valid TORN Reports: 170516/2200–170517/0200
(b) Automated
5
2
SFE2017 Auto TORN Forecast
Valid TORN Reports: 170516/2200–170517/0200
(c) Final
10%
15%
2%
5%
SFE2017: SPC FINAL TORN Forecast
Valid TORN Reports: 170516/2200–170517/0200
(d) Practically perfect
15
10
5
2
PP Probs Hindcast
Valid TORN Reports: 170516/2200–170517/0200

Figure 16 Example probabilistic tornado forecasts issued from the severe hazards desk on 16 May 2017 and valid for the 2200-0200 UTC time period with tornado reports overlaid. (a) Preliminary forecast issued by the severe hazards team in the morning, (b) automated forecast using temporal disaggregation, (c) final forecast issued in the afternoon, and (d) practically perfect probabilities derived from the distribution of observed tornadoes.

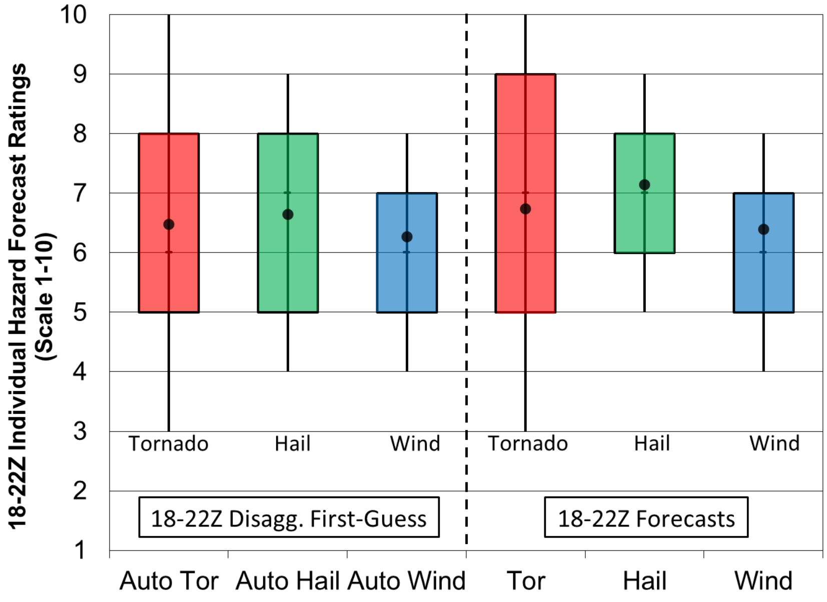


Figure 17 Same as Fig. 15, except for 4-h outlooks valid 1800-2200 UTC for the temporally disaggregated first-guess guidance (left) and the forecaster-issued outlook (right).

Similarly, experimental 4-h outlooks were generated for the 2200-0200 UTC period. In addition to the first-guess guidance and morning (i.e., preliminary) forecasts for 2200-0200 UTC, there was an afternoon update to this forecast period. The first-guess guidance, preliminary forecasts, and final forecasts of tornado, wind, and for this period were subjectively rated and compared (Fig. 18). In general, the forecaster was able to improve upon the first-guess guidance in the preliminary forecasts for this period. While updating the forecasts in the afternoon generally resulted in similar or slightly better forecast ratings, the improvement was fairly small in terms of subjective ratings.

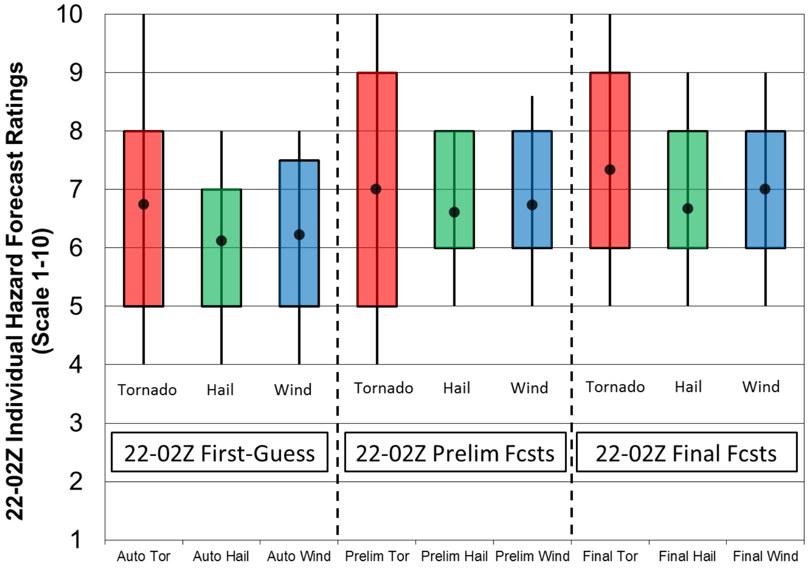


Figure 18 Same as Fig. 17, except for 4-h outlooks valid 2200-0200 UTC for the temporally disaggregated first-guess guidance (left), preliminary (morning) forecaster-issued outlooks (middle), and final (afternoon) forecaster-issued outlooks (right).

c) Model Evaluations – Innovation Desk

1) TORNADO PROBABILITIES

First-guess tornado forecast probabilities derived from CAM ensembles were evaluated during SFE2017. The first evaluation considered four different methods of first-guess forecast generation, which used only information from the 0000 UTC initialization of the NSSL-WRF ensemble. Two of the methods (2–5 km UH and 2–5 km UH requiring STP ≥ 1) were generated following the methods of Gallo et al. (2016). These probabilities were created based on the number of members solely exceeding 2–5 km UH ≥ 75m^2s^{-2}, or exceeding 2–5 km UH ≥ 75m^2s^{-2} coincident where STP ≥ 1 during the prior hour. The third method adapted the Gallo et al. (2016) approach by using low-level (0–3 km) UH and requiring it to exceed 33 m^2s^{-2}. The fourth and final method incorporated the climatological frequency of a tornado from a right-moving supercell given a specific value of STP (Thompson et al. 2017). STP values from each ensemble member were combined with the climatological frequency to provide a tornado probability from each member; the final product shown to participants was the average probability from the entire ensemble. Further details of this methodology can be found in Gallo et al. (2018).

The second evaluation examined the method that combines ensemble output with the STP climatology (Method 4) in both the NSSL-WRF ensemble and the HREFv2 ensemble, and subjectively compared those guidance forecasts to the operational Day 1 tornado forecast issued by SPC at 0600 UTC. Additionally, the current first-guess SSEO-SREF based guidance used by the SPC was also examined. This method incorporates environmental information (STP) from the SREF with explicit storm attribute information (UH) from the SSEO (see Jirak et al. 2014 for further details).

Participants were also asked to rate the guidance forecasts from the four different methods for two different qualities: the forecast area (or geographic coverage) and the forecast magnitude. This distinction stems from SFE2015, wherein participants commented that the forecasts highlighted the area of tornado occurrence properly but over-forecast the probabilities. Forecast area ratings between the four methods tested in the NSSL-WRF ensemble varied between methods, with the median scores ranging from 3.0/10.0 for the 0–3 km UH method to 6.0/10.0 for the STP-calibrated method (Fig. 19). **The methods incorporating the STP performed better than the methods that did not use any environmental information for the forecast area and the magnitude**. The STP-calibrated method also garnered the highest ratings for the magnitude of the probabilities, with a median score of 7.0/10.0. The UH-only methods both had very low median scores of 3.0/10.0, and participants commented on their tendency to produce unrealistically high probabilities.

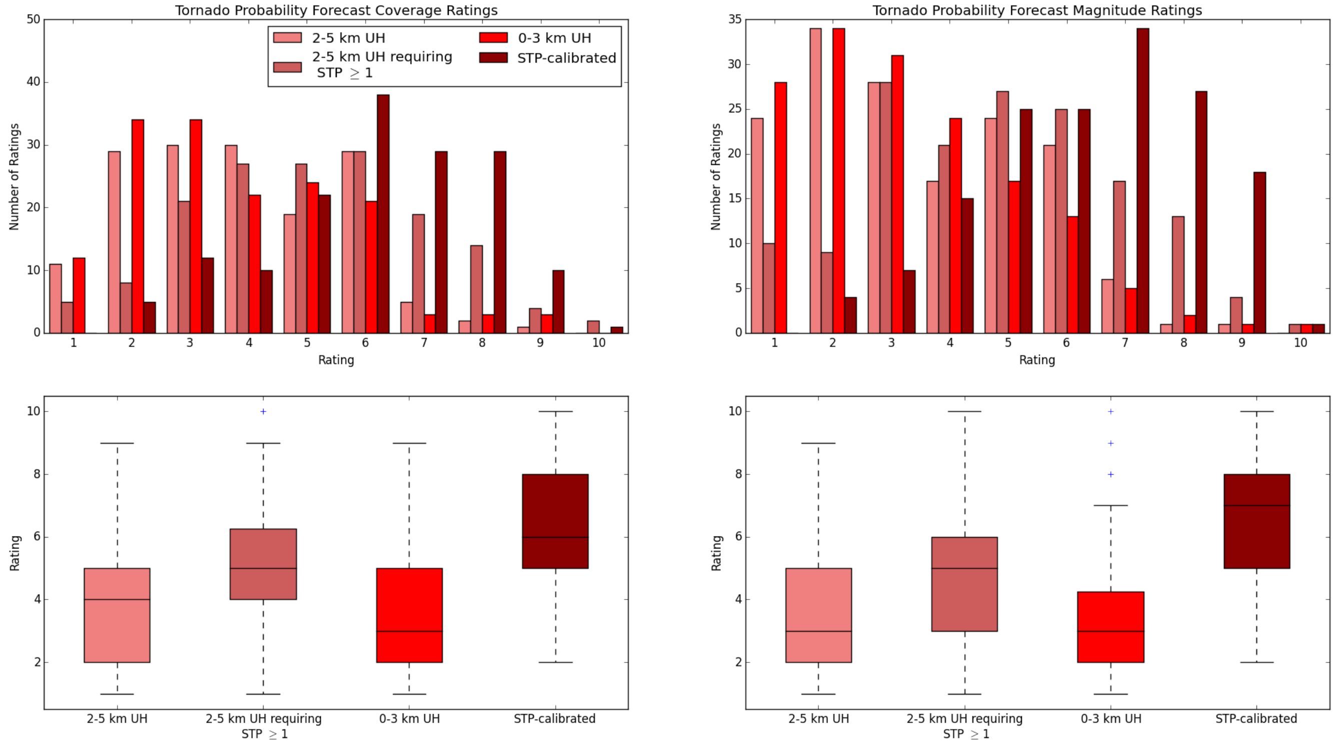


Figure 19 Subjective ratings of the forecast area (coverage) (left) and magnitude (right) of tornado probabilities generated using four different forecast methods.

After providing a numerical rating, participants were asked to choose the best and worst forecast (Fig. 20), and then explain their choices. Many participants cited false alarm as the reason for their worst forecasts, citing too much false alarm coverage as well as probability values that were too high. However, the STP-calibrated forecasts, which typically had lower magnitudes than the other forecast methods, were occasionally mentioned in the comments as having reduced the forecast area to an extent that reported tornadoes were excluded. Multiple participants also noted that including STP improved the forecasts in magnitude and coverage compared to solely using UH.

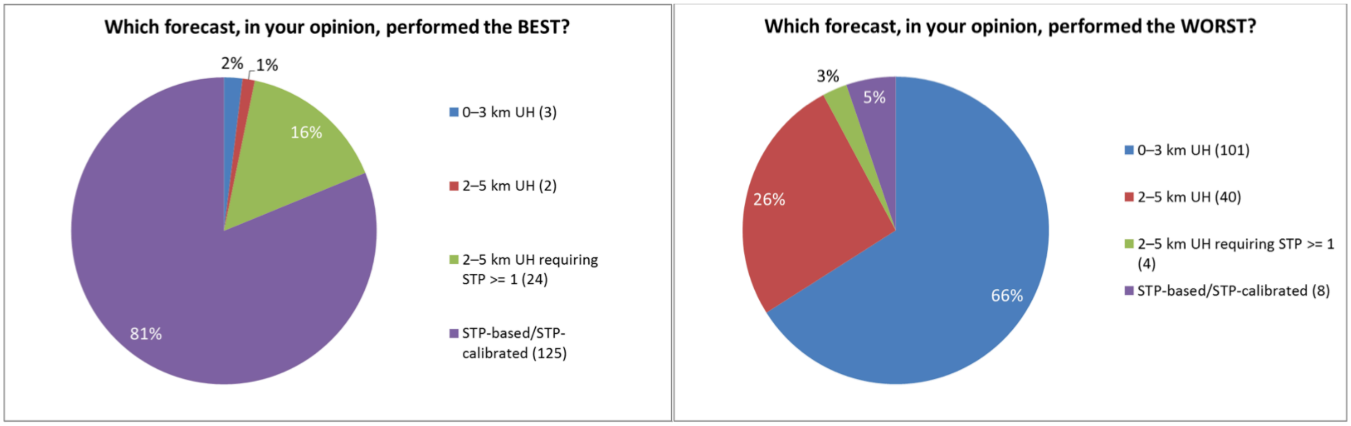


Figure 20 Summary of participant responses for the best (top) and worst (bottom) forecasts.

When comparing different CAM ensembles (NSSL-WRF, HREFv2, SSEO), the forecast ratings for forecast area and magnitude were extremely similar between ensembles, with a forecast area rating median of 6.0/10.0 for all ensembles (Fig. 21). The SSEO had higher ratings variability, but was also available during fewer days of the experiment compared to the guidance generated using either the NSSL-WRF ensemble or the HREFv2. When selecting the best and worst forecasts, participants also had the option of choosing the initial Day 1 SPC operational probabilistic tornado outlooks, which were issued at 0600 UTC. Forecasts were displayed alongside the practically perfect probabilities, with the tornado reports overlaid (e.g., Fig. 22). The SPC forecasts were most often rated the best of the four sets of forecasts, with 31% of these forecasts rated best, while the other three automated methods were each rated best about 23% of the time (Fig. 23). Participants were concerned with over-forecasting in some of the ensemble first-guess methods, but also noted a tendency of the SSEO probabilities to under-forecast some events. The NSSL-WRF ensemble was most often selected as the worst forecast, and the SPC outlook was the least often selected as the worst forecast. That the SPC is most often the best forecast is unsurprising, given the expertise of the forecasters and availability of observations and NWP guidance when making forecasts. However, the experimental first-guess forecasts were rated the best forecast on over 50% of the days, which is an encouraging result. This suggests that the highest rated STP-calibrated method in particular can provide forecasters with useful first-guess tornado probability guidance.

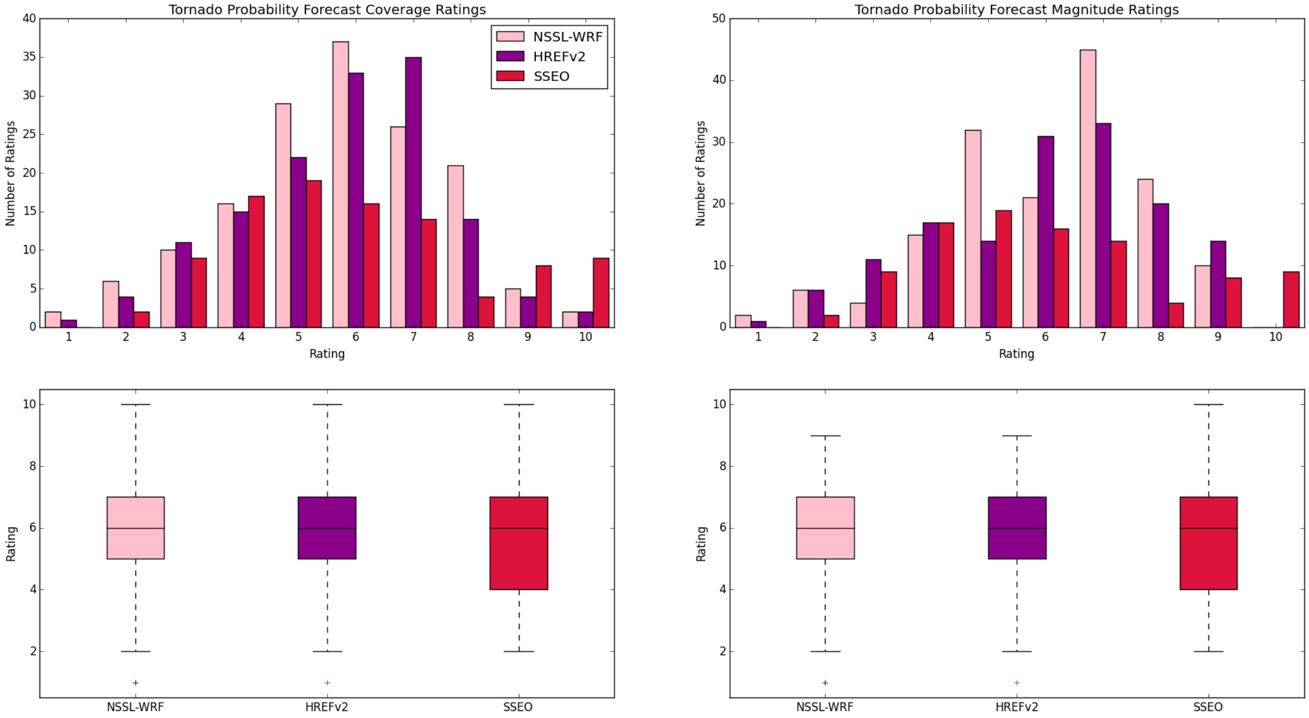
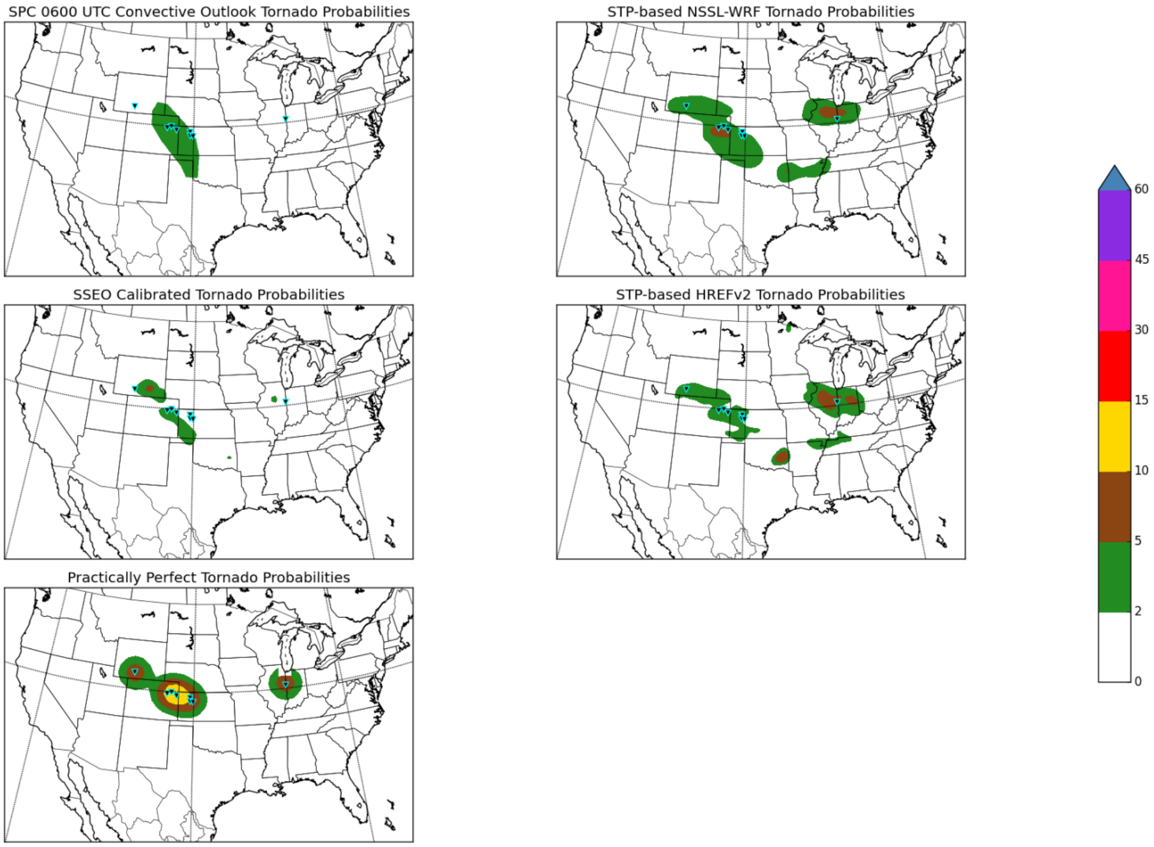
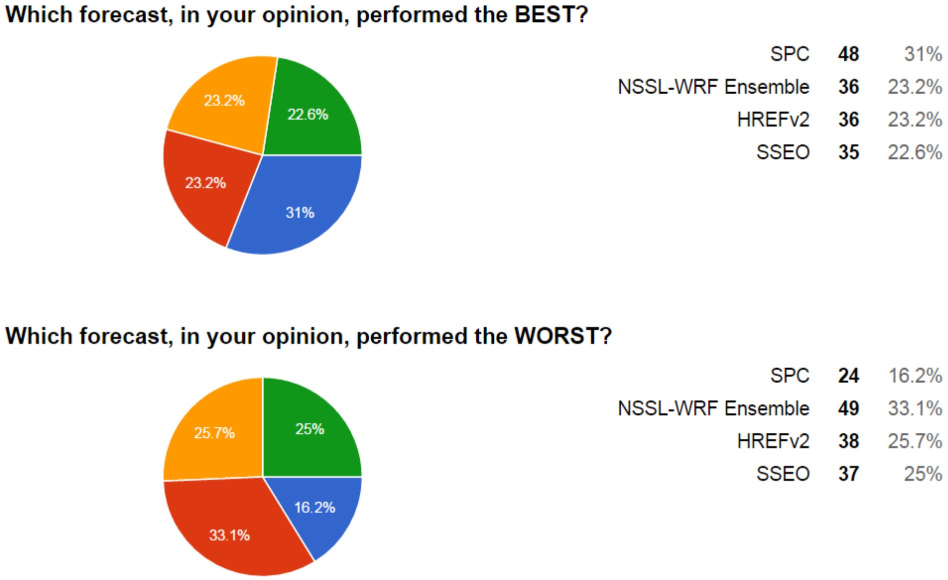


Figure 21 Subjective ratings of the forecast area (coverage) (left) and magnitude (right) of tornado probabilities generated using three different ensemble systems. The STP-calibrated method was implemented in the NSSL-WRF and HREFv2, while the SSEO used current first-guess guidance.



SPC 0600 UTC Convective Outlook Tornado Probabilities
STP-based NSSL-WRF Tornado Probabilities
SSEO Calibrated Tornado Probabilities
STP-based HREFv2 Tornado Probabilities
Practically Perfect Tornado Probabilities
60
45
30
15
10
5
2
0

Figure 22 Example verification for the tornado probabilities generated using different ensemble systems.



Which forecast, in your opinion, performed the BEST?

SPC	**48**	31%
NSSL-WRF Ensemble	**36**	23.2%
HREFv2	**36**	23.2%
SSEO	**35**	22.6%

Which forecast, in your opinion, performed the WORST?

SPC	**24**	16.2%
NSSL-WRF Ensemble	**49**	33.1%
HREFv2	**38**	25.7%
SSEO	**37**	25%

Figure 23 Summary of participant responses for the best (top) and worst (bottom) forecasts.

d) Model Evaluations – Severe Hazards Desk

1) FV3 EVALUATION

During SFE2017, the FV3 was run at convection-allowing scales for the first time in real-time as part of the 2017 CLUE. Two different 0000 UTC experimental versions of FV3 at 3-km grid spacing were examined and compared to operational CAMs to gauge its performance at convective scales. These versions included the FV3-GFDL with GFS physics and GFDL microphysics and FV3-CAPS with GFS physics and Thompson microphysics. GFDL and CAPS put forth substantial effort to implement severe weather diagnostic variables into the FV3 code and to generate grib2 output. An example of the convection-allowing FV3 forecasts is provided in Fig. 24. Overall, subjective ratings from SFE participants indicated that the FV3 reflectivity forecasts compared favorably to operational CAMs (Fig. 25).

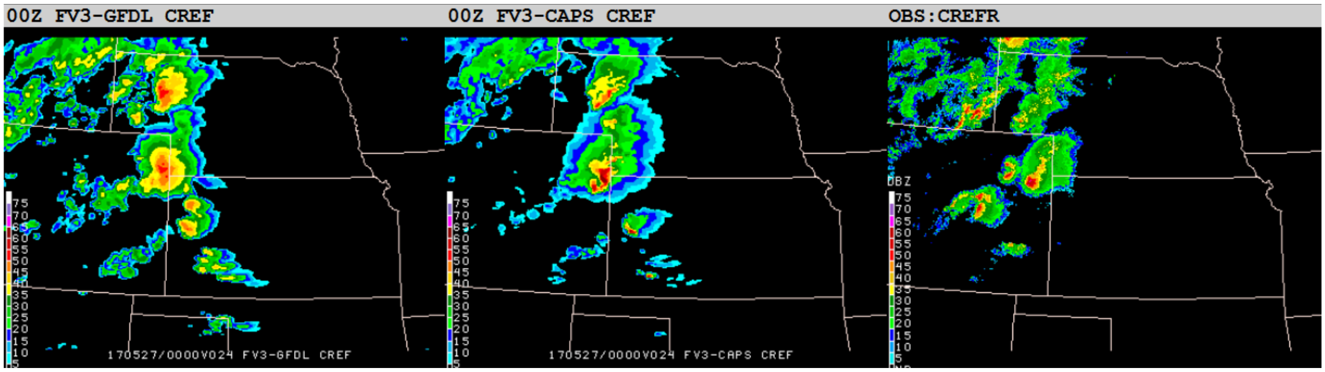


Figure 24 Example of subjective comparison plots used for rating CAM performance at convective scales. The left panel shows 24-h forecast of composite reflectivity of the FV3-GFDL, the middle panel shows the 24-h forecast of composite reflectivity of the FV3-CAPS, and the right panel shows the observed composite reflectivity at 0000 UTC on 27 May 2017.

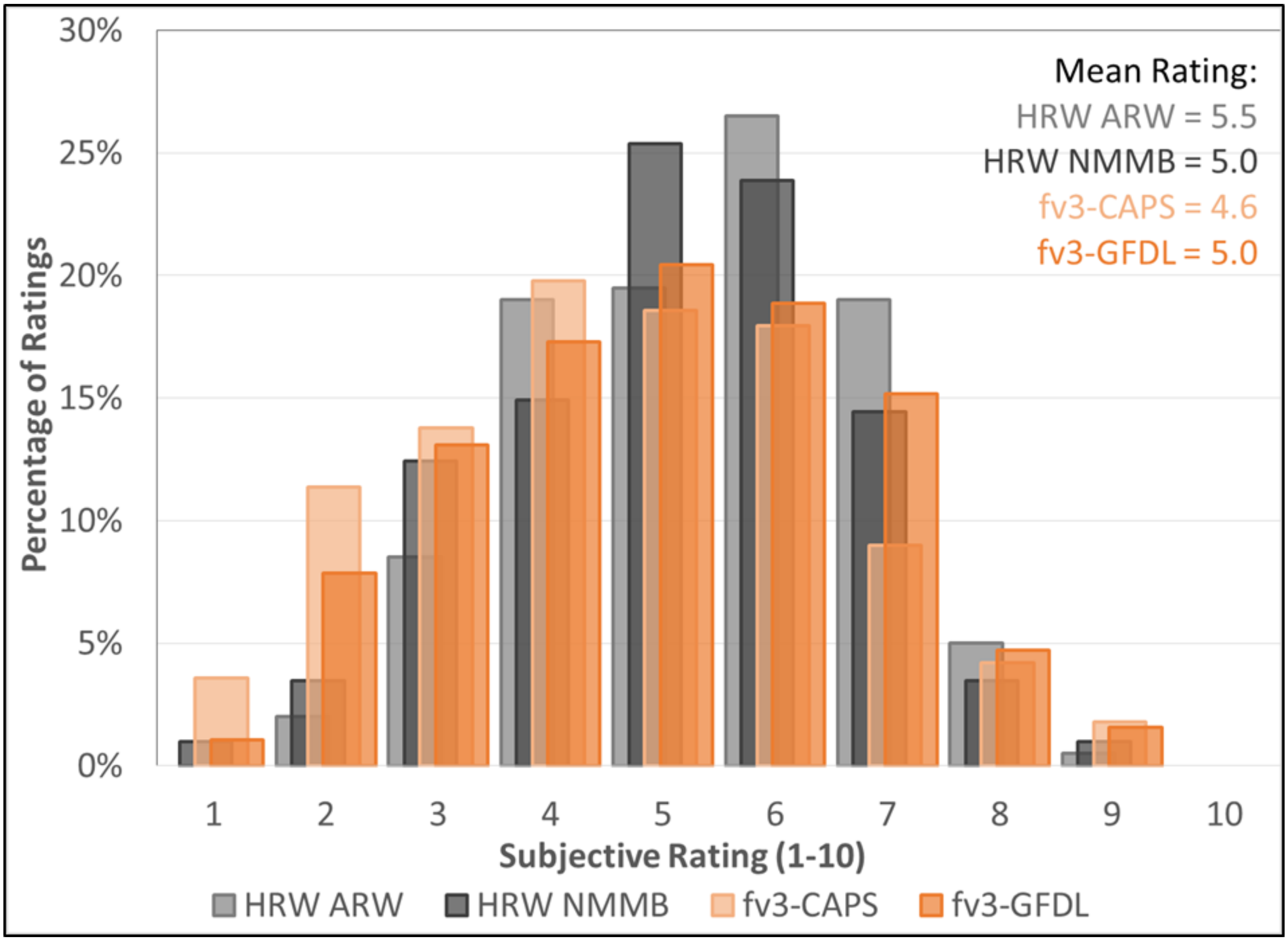


Figure 25 Normalized histogram of subjective ratings (on a scale of 1-10 with 10 being the highest rating) of the 0000 UTC 1-km AGL reflectivity forecasts (f018-f030) during SFE2017 for the FV3-GFDL (orange), FV3-CAPS (light orange), operational HiResW NMMB (gray), and operational HiResW WRF-ARW (light gray).

Additionally, an objective evaluation of FV3-GFDL was conducted using the surrogate severe methodology (Sobash et al. 2011, 2016), and comparisons were made to the HiResW 3-km NSSL-WRF configuration produced by EMC. The NSSL-WRF was chosen for comparison because it is highly regarded by severe weather forecasters, and its performance characteristics are well known. For application of the surrogate severe approach, the maximum UH at each grid-point was computed over the 24 h period 1200-1200 UTC for both models. Then, the maximum UH values were remapped to the 81 km NCEP 211 grid by assigning each 81 km grid box the maximum value of UH out of all 3-km grid-points within the 81 km boxes. Next, severe weather probabilities were computed by assigning grid-boxes a value of 1.0 (0.0) if a specified UH percentile was (not) exceeded, and then applying a two-dimensional Gaussian filter. The UH percentiles were computed separately for each model using the distribution of UH values from the 81 km grids over all 24 cases. ***The percentiles, rather than thresholds, were used to ensure equitable comparisons between the two models, which have different UH climatologies.***

The percentiles from 0.90 to 0.999 in increments of 0.01 (100 unique percentiles) were examined, and for each percentile, a range of standard deviations (sigma) from 40 to 300-km in increments of 5 km were tested (i.e., 53 unique sigma values). Thus, for each case and model, there were 100 x 53 = 5300 sets of severe probabilities. To verify the probabilities, preliminary observed storm reports from SPC were mapped to the same 81 km grid as the severe probabilities. Any grid-box with one of more reports over the 1200 – 1200 UTC time period was assigned 1.0 while boxes with zero reports were assigned 0.0. The metrics area under the relative operating characteristic curve (AUC; Mason 1982) and Fractions Skill Score (FSS; Roberts and Lean 2008) were used for objective verification.

Each skill metric is presented as a function of sigma and UH percentile in Figure 26. The NSSL-WRF had noticeably higher AUC and FSS indicating more skillful severe weather probabilities than the FV3. Further work is planned to explore these probabilities in more depth.

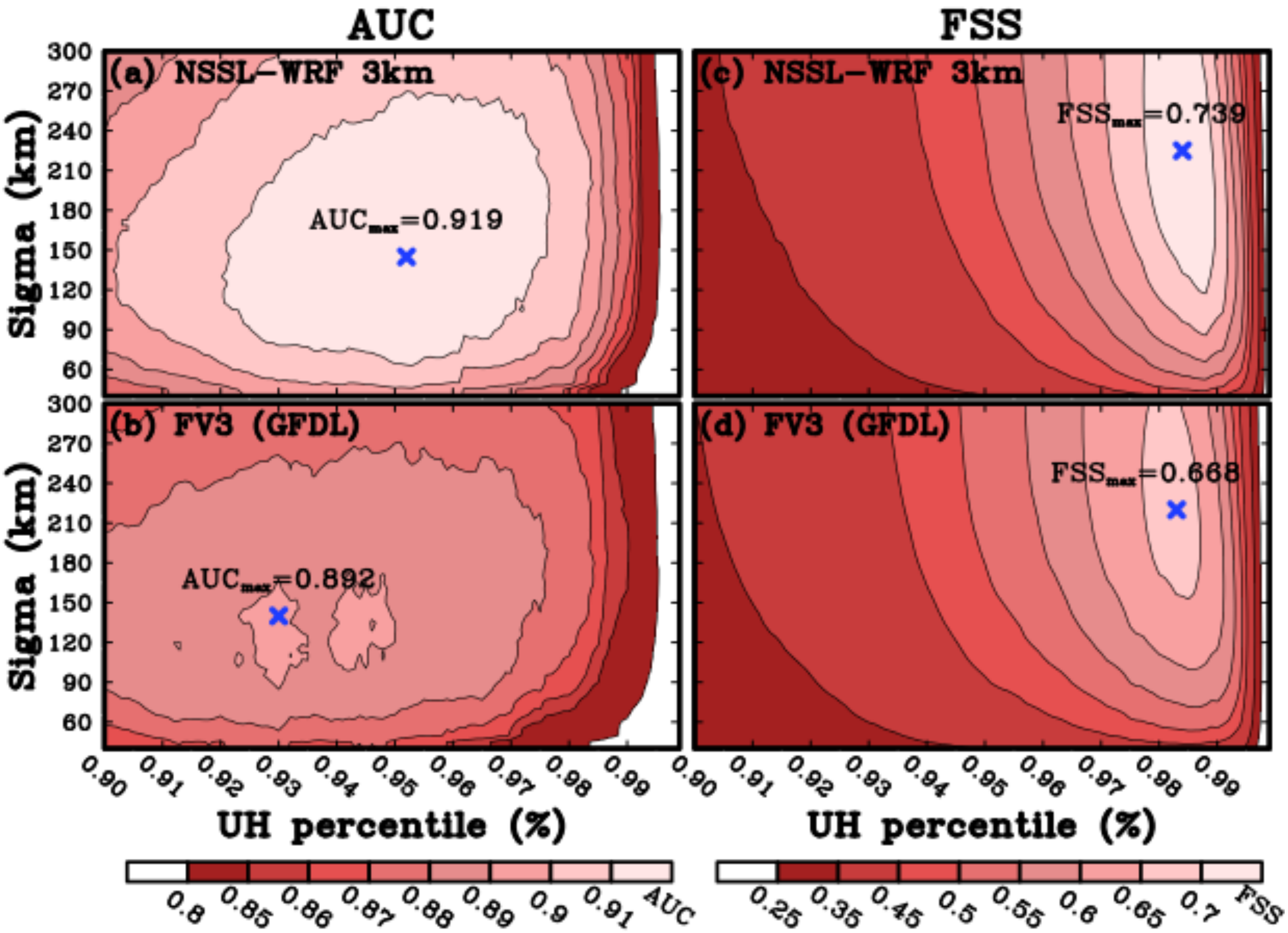


Figure 26 AUC as a function of sigma and UH percentile for the (a) NSSL-WRF and (b) FV3-GFDL. (c) and (d) same as (a) and (b), except for the FSS. In each panel, a blue "x" marks the best score, which is indicated in text.

2) HIRES WINDOW RUN EVALUATIONS

Parallel versions of the High Resolution Window (HiResW) runs were available from EMC during the SFE2017. These were very similar to the operational versions with the primary difference of being run at higher resolution (i.e., grid spacing of 3.2 km). For the HiResW NMMB, many of the forecasts were like the example in Fig. 27, where there are small-scale reflectivity differences between the operational and parallel versions, but the forecasts were qualitatively similar. The more detailed reflectivity structures were often considered to provide improved forecast guidance (Fig. 27). The subjective ratings show a distribution of scores shifted toward higher values with improved mean scores for the parallel HiresW NMMB.

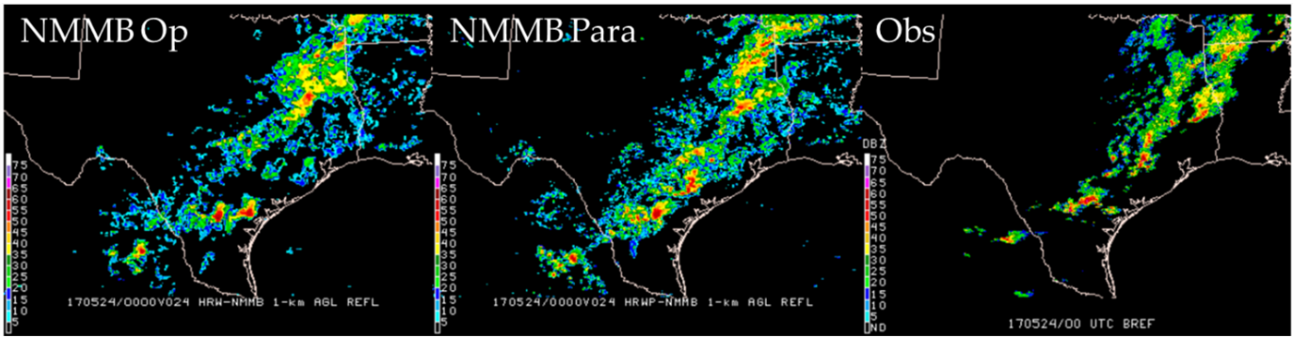

NMMB Op
NMMB Para
Obs
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
DBZ
ND
170524/0000V024 HRW-NMMB 1-km AGL REFL
170524/0000V024 HRWP-NMMB 1-km AGL REFL
170524/00 UTC BREF

Figure 27 Same as Fig. 24, except for 24-h forecast for the operational HiResW NMMB (left), parallel HiResW NMMB (middle), and observed reflectivity valid at 0000 UTC on 24 May 2017.

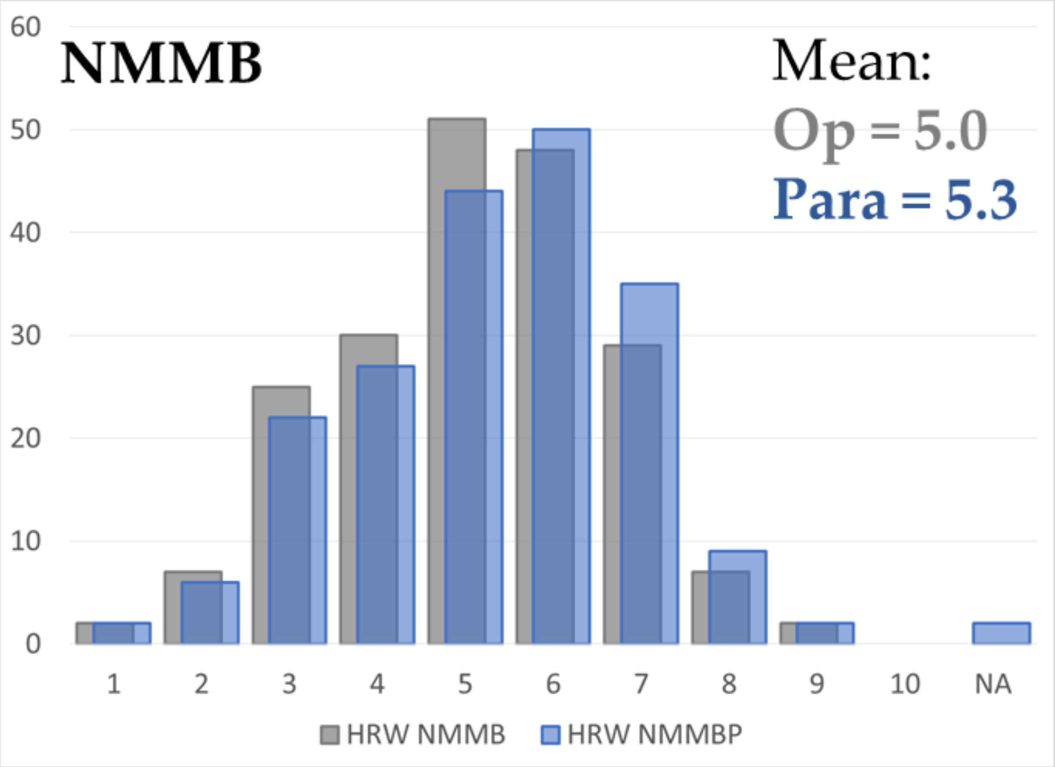


Figure 28 Histogram of subjective ratings (on a scale of 1-10 with 10 being the highest rating) of the 0000 UTC 1-km AGL reflectivity forecasts (f018-f030) during SFE2017 for the parallel HiResW NMMB (blue) and operational HiResW NMMB (gray).

The differences in performance between the operational and parallel HiResW ARW runs were generally smaller than those found between the operational and parallel HiResW NMMB. The two ARW forecasts were often qualitatively very similar. For example, Figure 29 shows a 24-h forecast where the parallel ARW mirrored the location, mode, and intensity of storms in the operational version, although smaller scale reflectivity details are seen in the higher resolution parallel run. Given the overall small differences between these runs, the subjective ratings were, not surprisingly, very similar between the operational and parallel versions (Fig. 30), with a very slight edge to the parallel HiResW ARW.

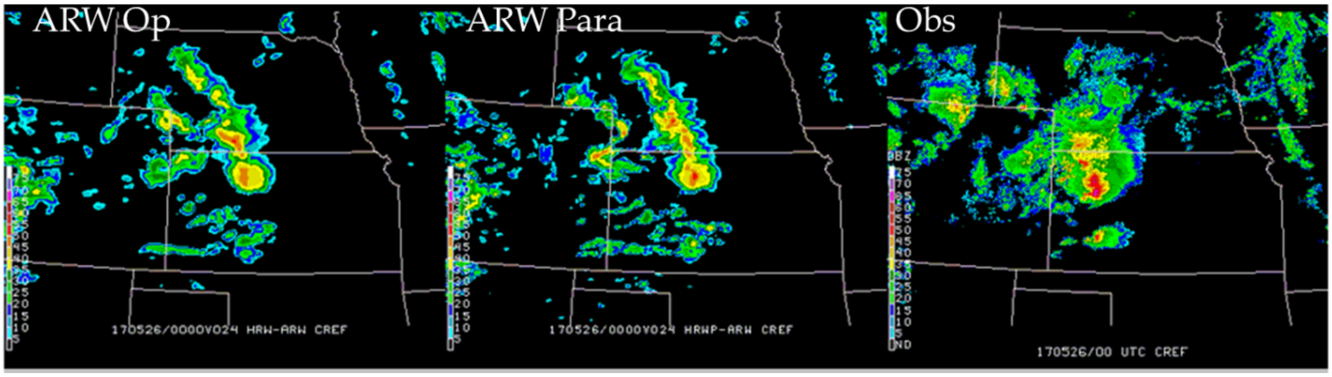


Figure 29 Same as Fig. 24, except for 24-h forecast for the operational HiResW ARW (left), parallel HiResW ARW (middle), and observed reflectivity valid at 0000 UTC on 26 May 2017.

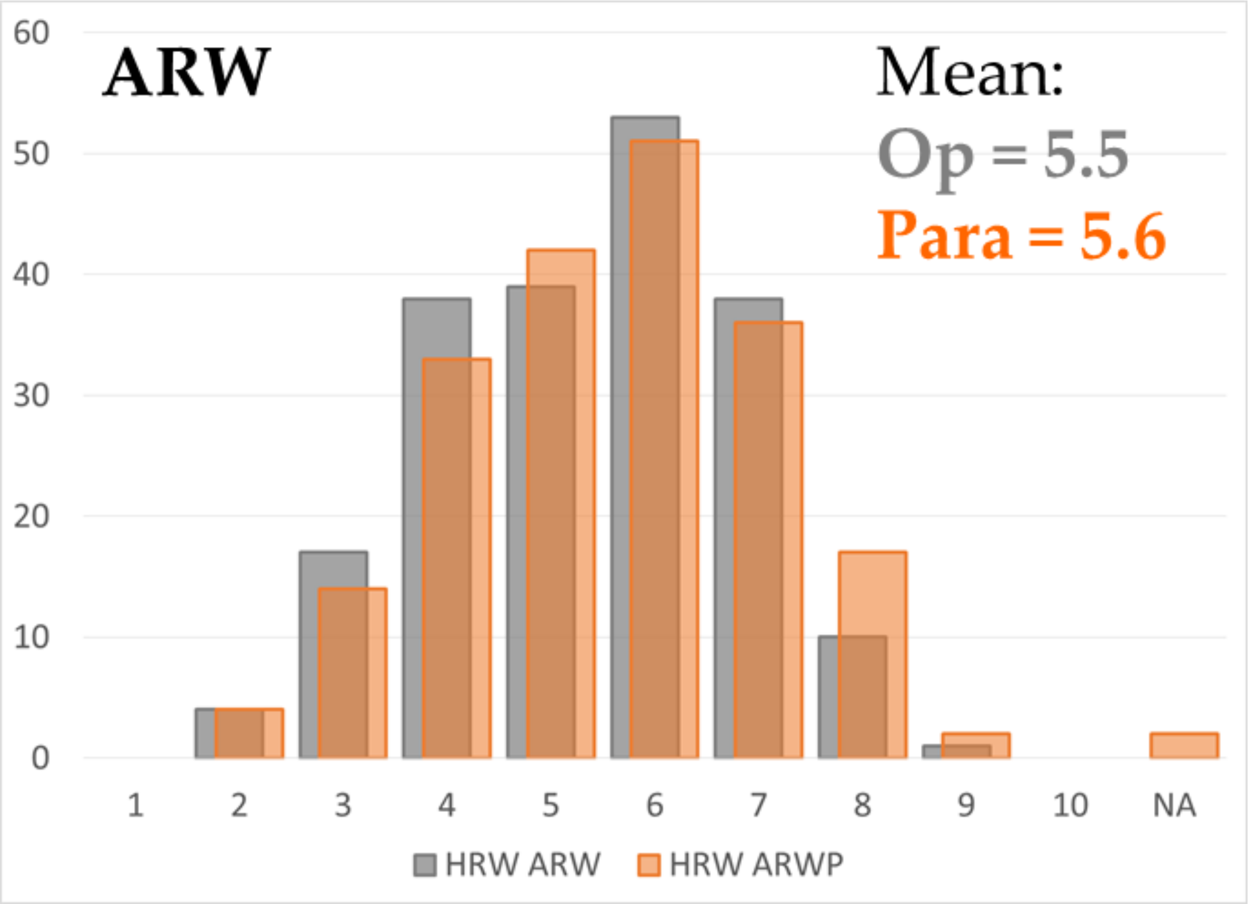


Figure 30 Same as Fig. 28, except for the parallel HiResW ARW (orange) and operational HiResW ARW (gray).

A third HiResW run that is configured like the experimental NSSL-WRF was also examined during SFE2017. The HiResW NSSL-WRF version was often as good as or slightly better than the NSSL-WRF. Figure 31 shows an example where the HiResW NSSL run provided an improved forecast of tornadic supercells in southwestern Oklahoma. Overall, the HiResW NSSL run was the highest rated 0000 UTC deterministic CAM examined during SFE2017 with slightly higher ratings than the experimental NSSL-WRF (Fig. 32).

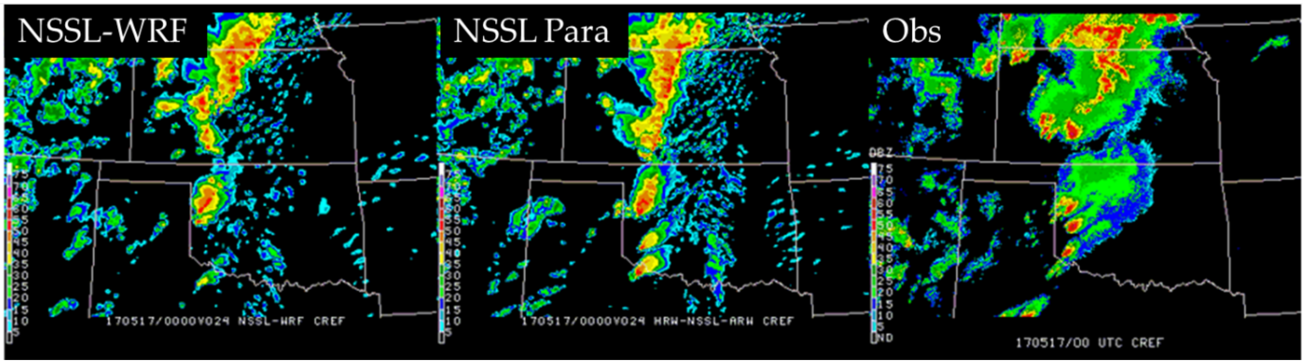

NSSL-WRF
NSSL Para
Obs
170517/0000V024 NSSL-WRF CREF
170517/0000V024 HRW-NSSL-ARW CREF
170517/00 UTC CREF

Figure 31 Same as Fig. 24, except for 24-h forecast for the experimental NSSL-WRF (left), parallel HiResW NSSL (middle), and observed reflectivity valid at 0000 UTC on 17 May 2017.

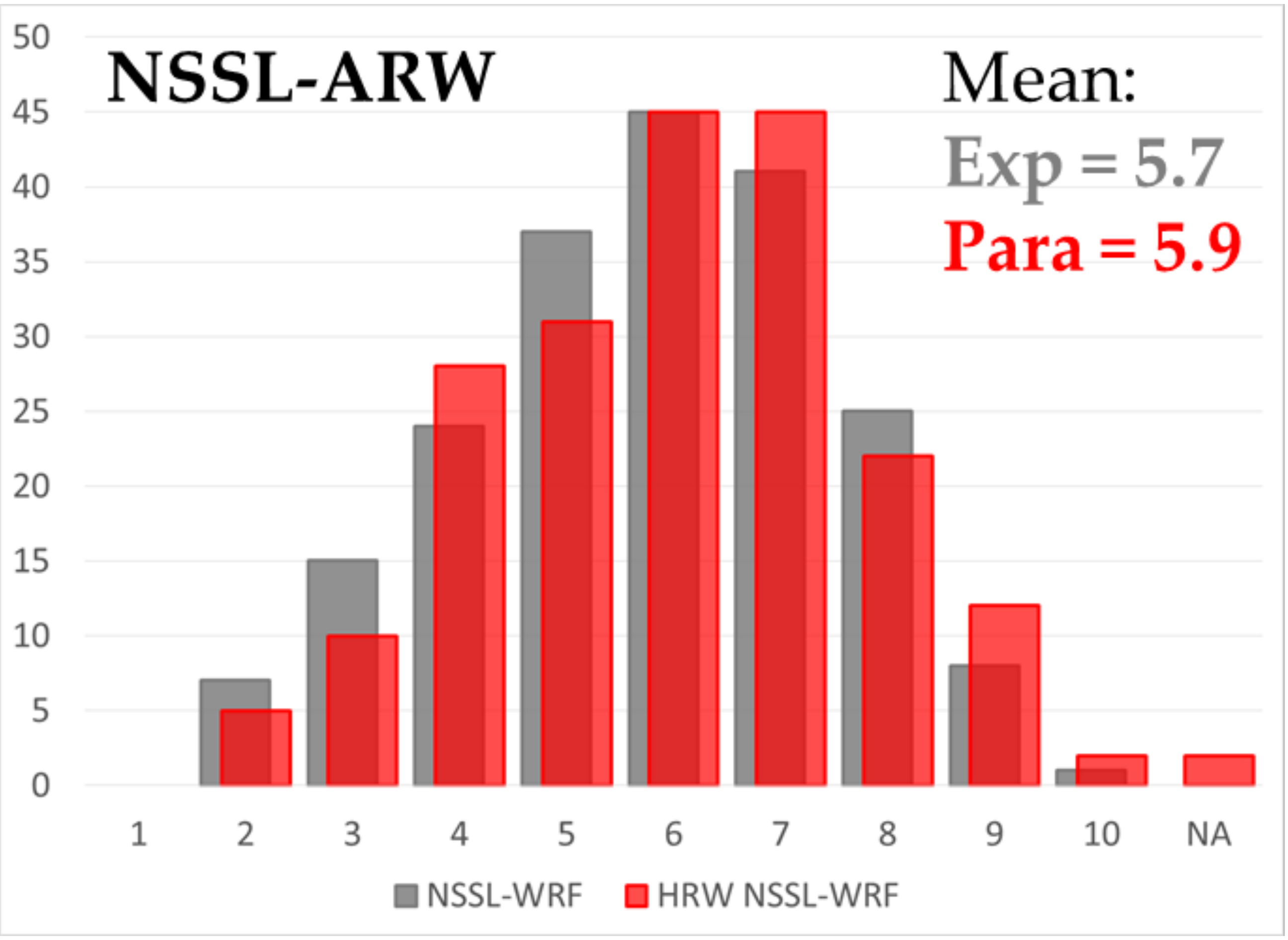


Figure 32 Same as Fig. 28, except for the parallel HiResW NSSL (red) and experimental NSSL-WRF (gray).

The parallel HiResW NMMB, ARW, and NSSL-ARW runs were implemented operationally on 1 November 2017 as part of the HREFv2 package.

3) HRRR EVALUATIONS

An experimental version of the HRRR (i.e., HRRRv3) produced by GSD was examined for comparison with the operational HRRR (i.e., HRRRv2). This evaluation primarily focused on 1500 UTC forecasts valid 1800-0600

UTC, but the 1200 UTC or 1800 UTC cycles were examined if the 1500 UTC cycle was not available. These runs often revealed differences in the forecasts between HRRRv2 and HRRRv3 (e.g., Fig. 33), but it was not always clear if one version was better than the other. Overall, the subjective ratings were similar for these runs (e.g., mean rating was the same), but the HRRRv3 had a distribution shifted toward higher ratings (e.g., mode rating of 7) than the operational HRRR (Fig. 34).

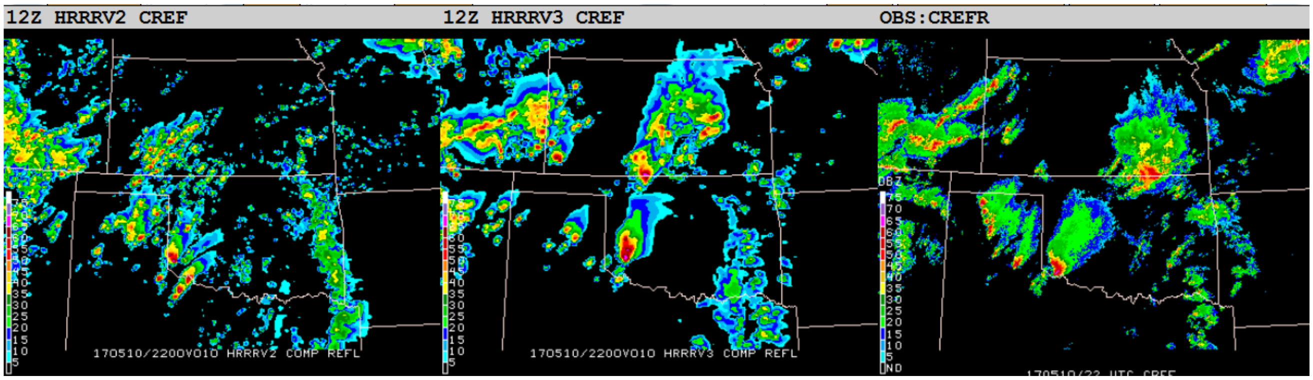


Figure 33 Same as Fig. 24, except for 10-h forecast for the operational HRRRv2 (left), experimental HRRRv3 (middle), and observed reflectivity valid at 2200 UTC on 16 May 2017.

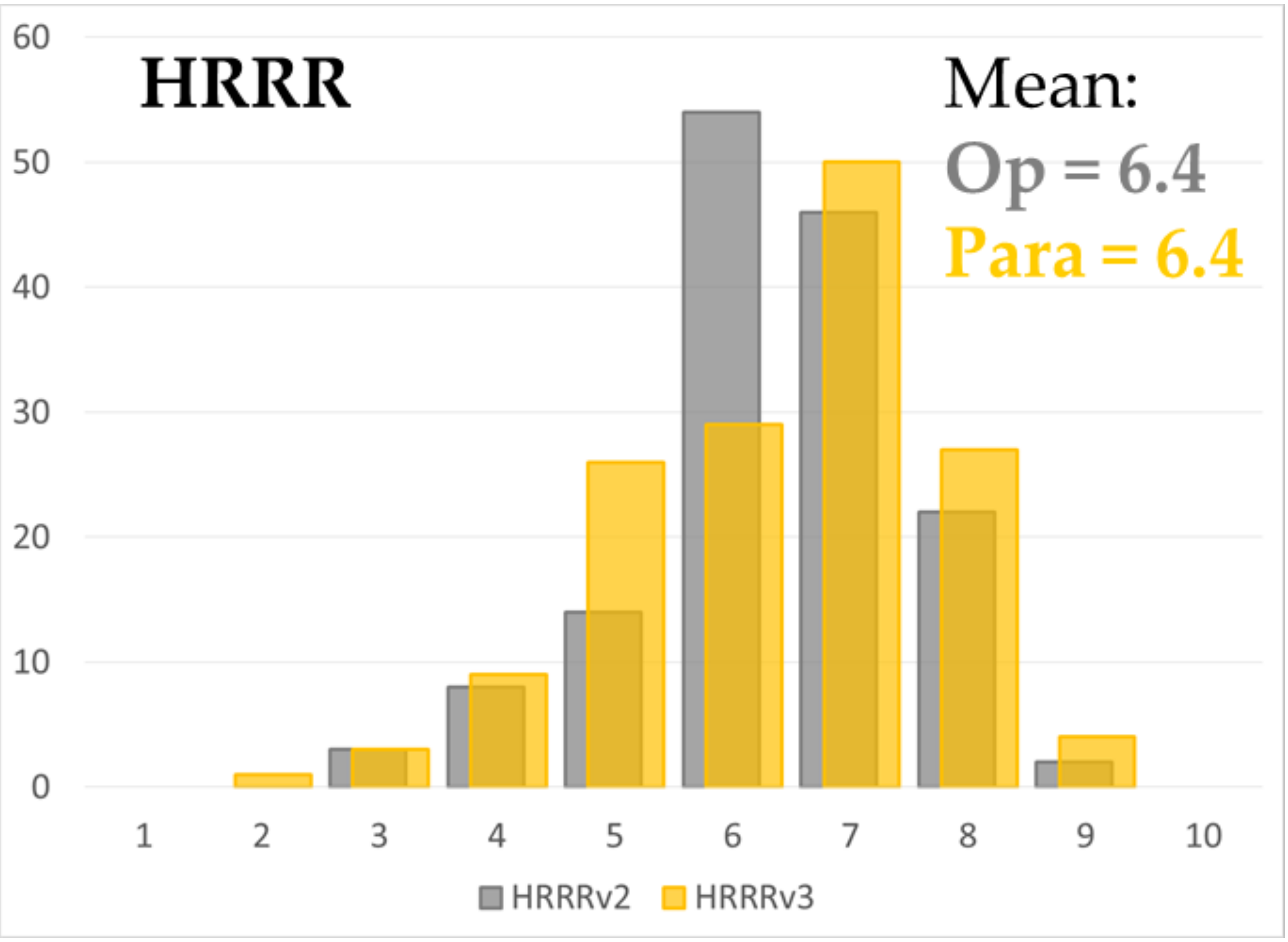


Figure 34 Same as Fig. 28, except for 1200, 1500, or 1800 UTC forecasts (depending on availability) of the experimental HRRRv3 (yellow) and operational HRRRv2 (gray).

4) UM EVALUATIONS

The Met Office provided three separate CAM runs over the CONUS, and two of these were evaluated on a daily basis (when available): the operational UM configuration and the experimental mid-latitude configuration. This evaluation primarily focused on the 18-30 hour reflectivity forecasts from the 0000 UTC runs (i.e. valid 1800-0600 UTC; Fig. 35). The UM runs often looked similar as anticipated, and the subjective ratings revealed this similarity (Fig. 36). The UM CAMs compared favorably to the best-performing CAM (i.e., the 3-km HiResW NSSL), but revealed a bi-modal rating distribution with more lower-rated forecasts than the HiResW-NSSL.

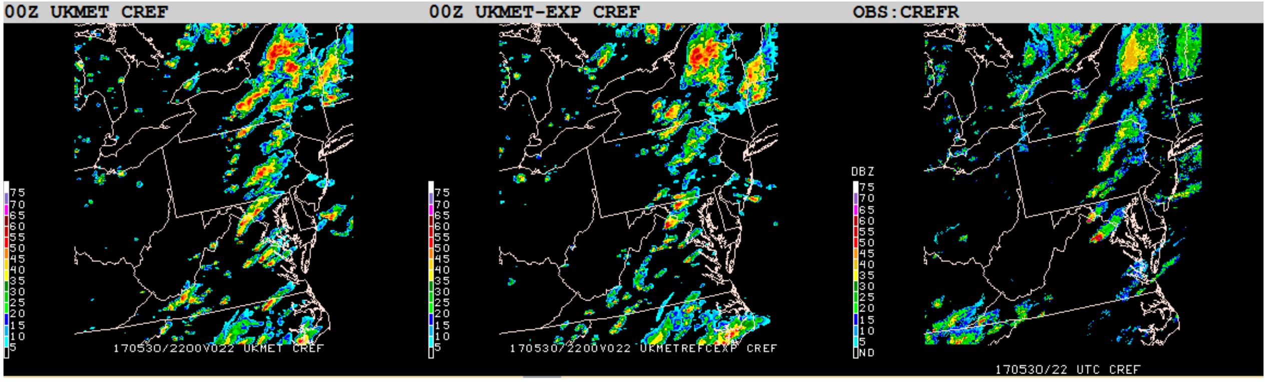


Figure 35 Same as Fig. 24, except for 22-h forecast for the operational UM (left), experimental UM (middle), and observed reflectivity valid at 2200 UTC on 30 May 2017.

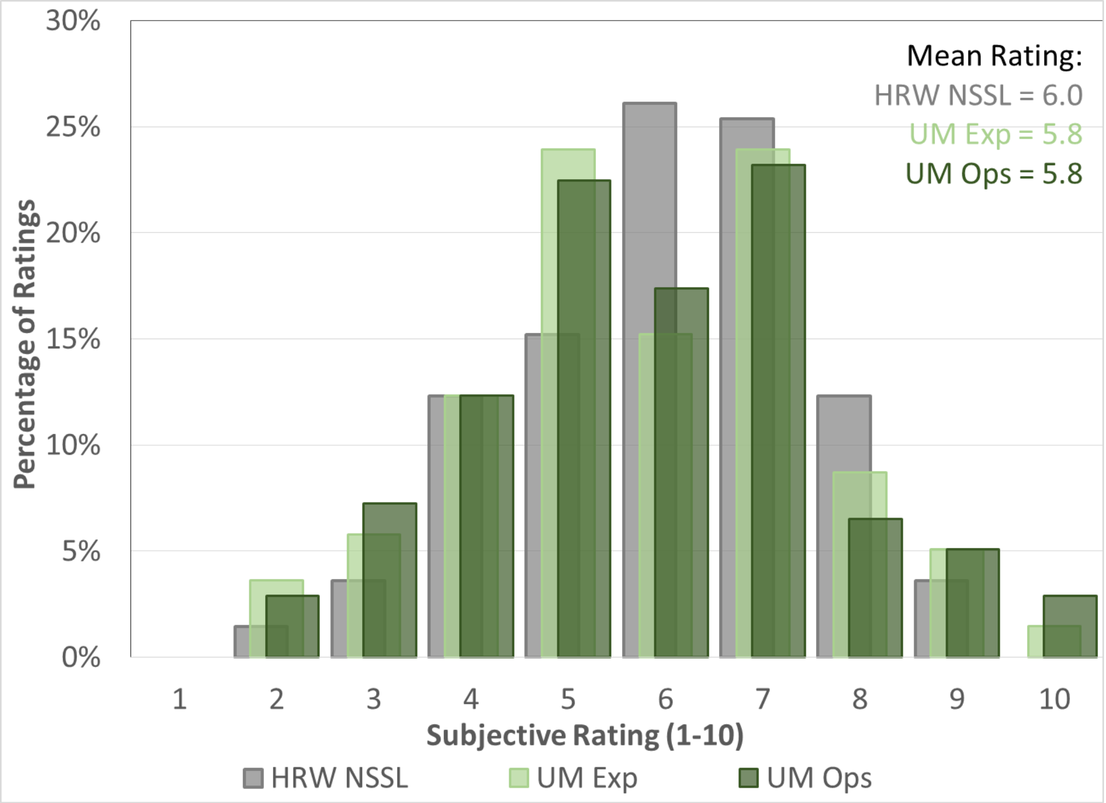


Figure 36 Same as Fig. 28, except for the HiResW NSSL (gray), UM experimental midlatitude configuration (UM exp; light green), and UM operational configuration (UM Ops; dark green).

4) CLUE: COMPARISON TO SSEO/HREFv2 AS A BASELINE

The HREFv2, comprised of the aforementioned parallel HiResW runs along with the NAM CONUS Nest, is an operational version of the experimental SSEO. Given the utility and success of the SSEO in forecasting hazardous weather since 2011, it has been used as a baseline to assess the performance of other experimental CAM ensembles. With the operational implementation of HREFv2 on 1 November 2017, it can be used now as the performance baseline for experimental CAM ensemble configurations being considered for operational implementation.

An important aspect of SFE2017 was to compare the HREFv2 to the SSEO. While there are some differences between these ensembles (e.g., membership, horizontal resolution, etc.), the HREFv2 was designed to be a multi-model, multi-physics, multi-IC operational version of the SSEO, so the expectation was that the HREFv2 would perform similarly to the SSEO. Figure 37 provides an example of the similarity between the HREFv2 and SSEO forecasts. Note that many of the HREFv2 members are run at higher resolution (i.e., 3-3.2 km grid spacing) than the SSEO members (3-4 km grid spacing), which impacts the magnitudes of UH forecasts. Thus, neighborhood probabilities of UH ≥75 m^2s^{-2} were examined for the HREFv2 along with other 3-km CAM ensembles while neighborhood probabilities of UH ≥25 m^2s^{-2} were examined for the SSEO.

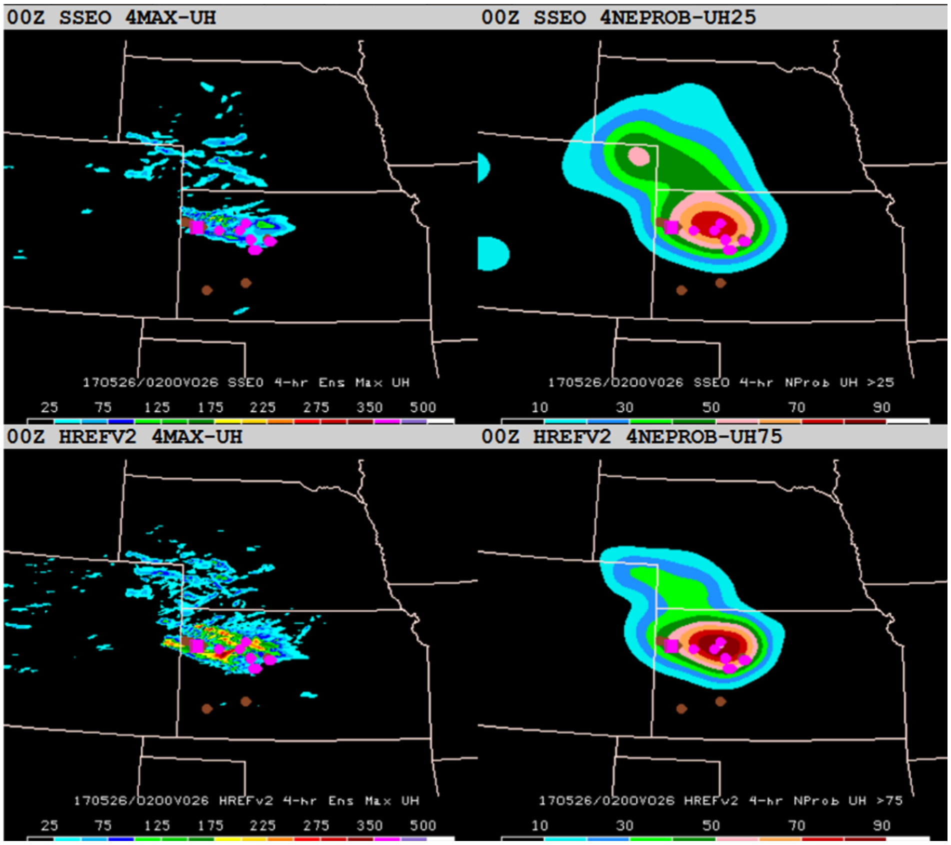


Figure 37 Example of subjective comparison plots used for rating CAM ensemble performance. The left column shows the 26-h forecast of the 4-h ensemble max updraft helicity (UH) from the SSEO (top) and HREFv2 (bottom), and the right column shows the 26-h forecast of neighborhood probability of UH ≥25 m^2s^{-2} for the SSEO (top) and UH ≥75 m^2s^{-2} for the HREFv2 (bottom). The preliminary local storm reports (hail – brown, wind – magenta, tornado – white) are shown for the 4-h valid period of the forecasts: 2200 UTC 25 May – 0200 UTC 26 May 2017.

The subjective component of the evaluation examined ensemble forecasts (i.e., 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. The distribution of forecast ratings for the SSEO and HREFv2 were indeed very similar during SFE2017 (Fig. 38). Both had a mean rating of 6.6 with a median and mode of 7.

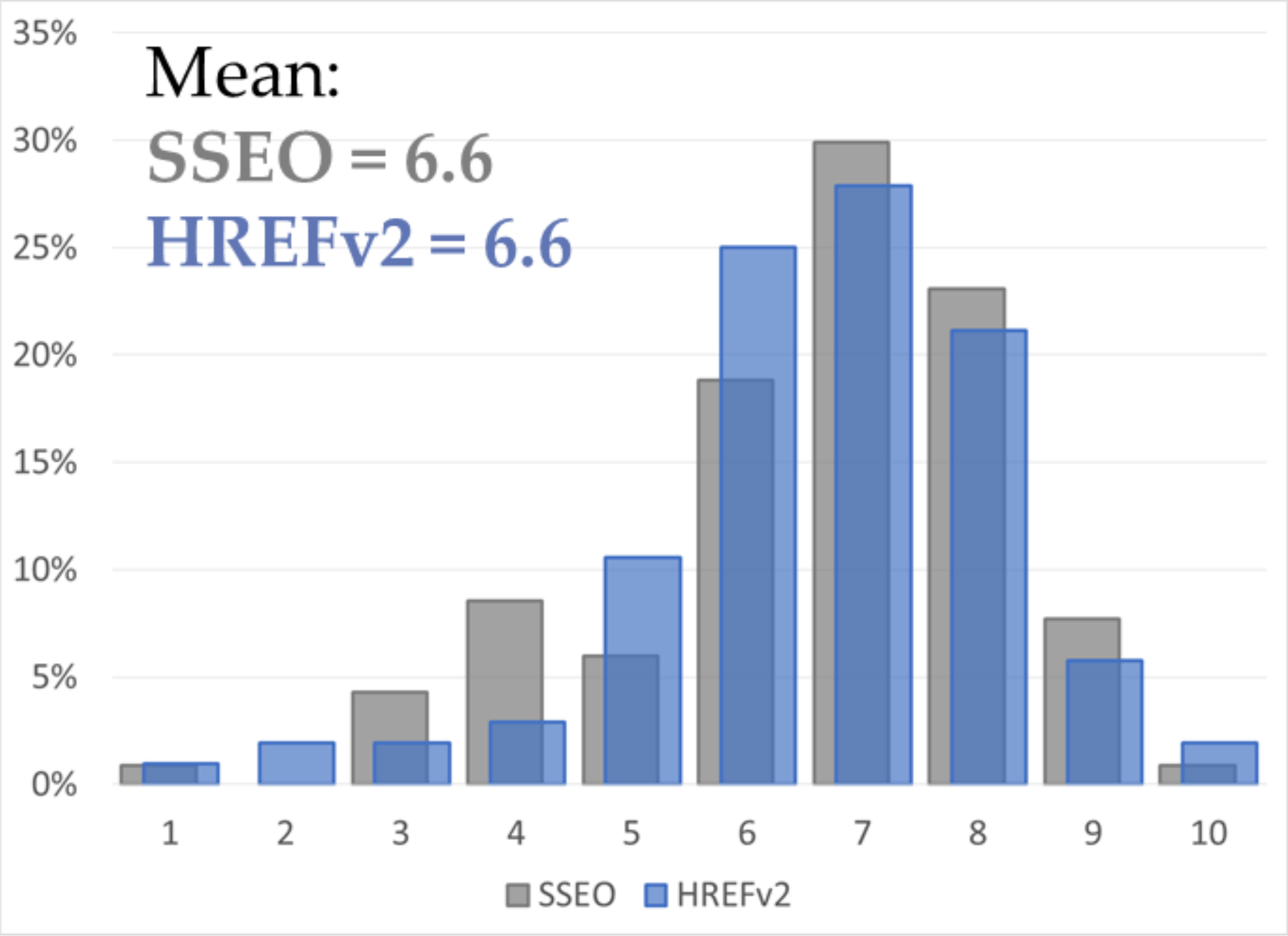


Figure 38 Same as Fig. 28 , except for CAM ensemble ratings of HMF forecasts from the parallel HREFv2 (blue) and experimental SSEO (gray).

The SSEO and HREFv2 were also compared to other ensemble subsets from the 2017 CLUE, including ensembles with advanced ensemble-based data assimilation: NCAR EnKF, HRRRE, CAPS EnKF, and GSI EnKF. The SSEO and HREFv2 tended to have the highest subjectively rated forecasts compared to the other CLUE ensembles (Fig. 39). The NCAR ensemble was overall the next-highest-rated ensemble followed by the HRRRE, CAPS EnKF, and GSI EnKF. These subjective results suggest that the HREFv2 will serve as a meaningful baseline against which experimental and next-generation CAM ensembles should be compared for consideration of operational implementation.

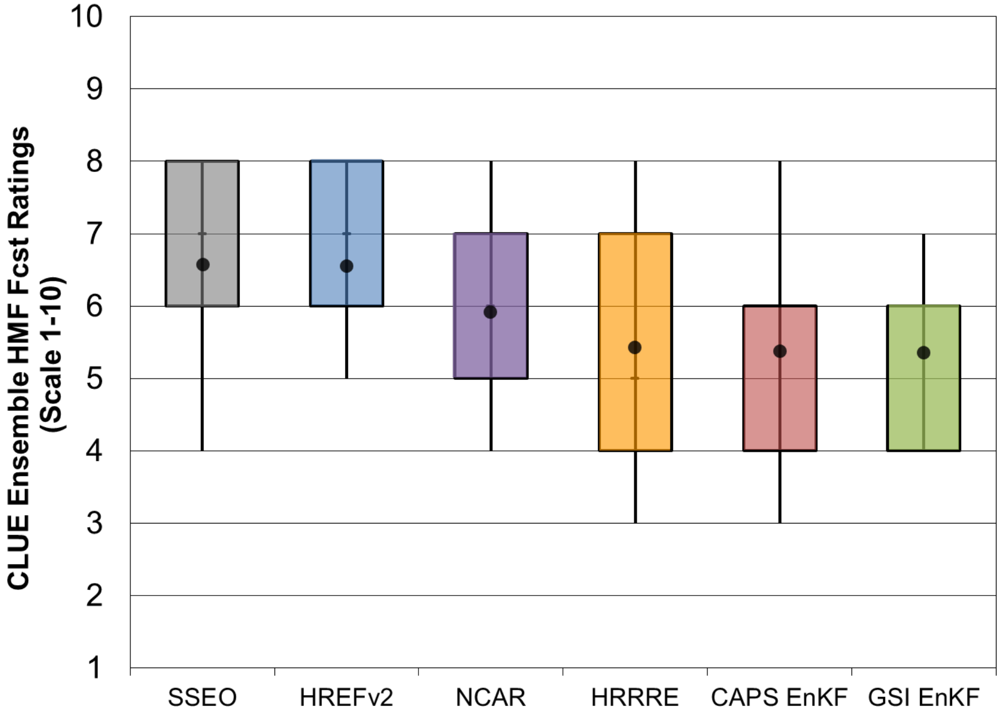


Figure 39 Distributions of subjective ratings (1-10) by SFE participants of HMFs over a mesoscale area of interest for the forecast hours 13-30 for various CLUE ensembles compared to the SSEO and HREFv2.

The objective verification results of reflectivity forecasts (i.e., neighborhood probabilities ≥40 dBZ) from the CAM ensembles generally agree with subjective ratings of HMF forecasts. The fractions skill scores (FSS) of the SSEO, HREFv2, and NCAR ensembles are generally higher than those of the other CAM ensembles, especially during the period of peak convective activity from 2100 UTC to 0400 UTC (Fig. 40).

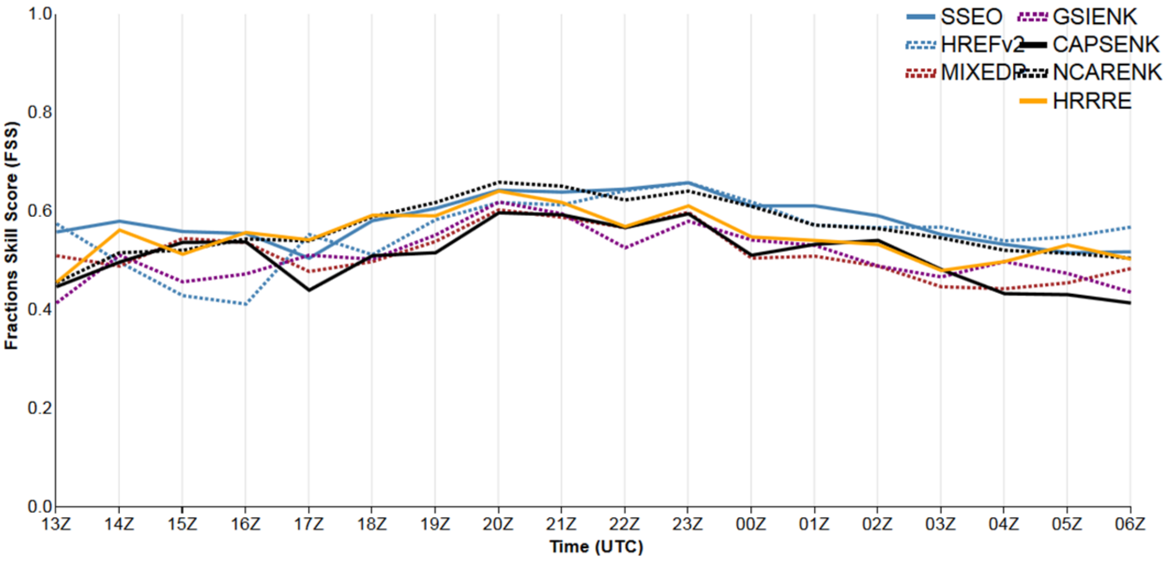


Figure 40 Accumulated fractions skill score of 0000 UTC reflectivity forecasts ≥40 dBZ by forecast hour (fh13-fh30) during SFE2017 over the daily mesoscale area of interest for the various CLUE ensembles and the SSEO and HREFv2.

4) SPC HAZARD GUIDANCE

Various forms of calibrated hazard guidance from SPC were available for examination and evaluation during SFE2017. The SSEO/SREF approach calibrates probabilistic environment forecasts from the SREF and storm-attribute forecasts from the SSEO based on the observed historical frequency of LSRs within 25 miles of a forecast location (Jirak et al. 2014). This guidance has been examined in the HWT SFEs for the past few years and is utilized in the temporally disaggregated first-guess guidance. The Statistical Severe Convective Risk Assessment Model (SSCRAM; Hart and Cohen 2016) was also evaluated using RAP forecasts as environmental input and HRRR reflectivity forecasts to remove the condition of thunderstorm occurrence. A third type of guidance referred to as the RAP STP approach was also available for tornado forecasting. This guidance assigns the observed climatological frequency of tornadoes produced from supercells (Thompson et al. 2017) based on RAP forecasts of STP as the conditional 4-h tornado probability. This conditional probability is then multiplied by the 4-h, 40-km neighborhood probability of UH≥100 m^2s^{-2} from a time-lagged HRRR (i.e., proxy supercell probability), resulting in the unconditional 4-h probability of a tornado within 25 miles of a point.

The 4-h tornado guidance received a wide distribution of ratings during SFE2017 (Fig. 43). The RAP STP approach received the highest overall ratings, followed by the SSEO/SREF and the SSCRAM approaches. For the 4-h severe hail (Fig. 44) and severe wind (Fig. 45) probabilities, the SSEO/SREF guidance was also subjectively rated higher overall than the SSCRAM guidance. Using the SSEO/SREF calibration approach, the wind guidance received the highest mean rating, while hail was the highest-rated hazard guidance using the SSCRAM approach.

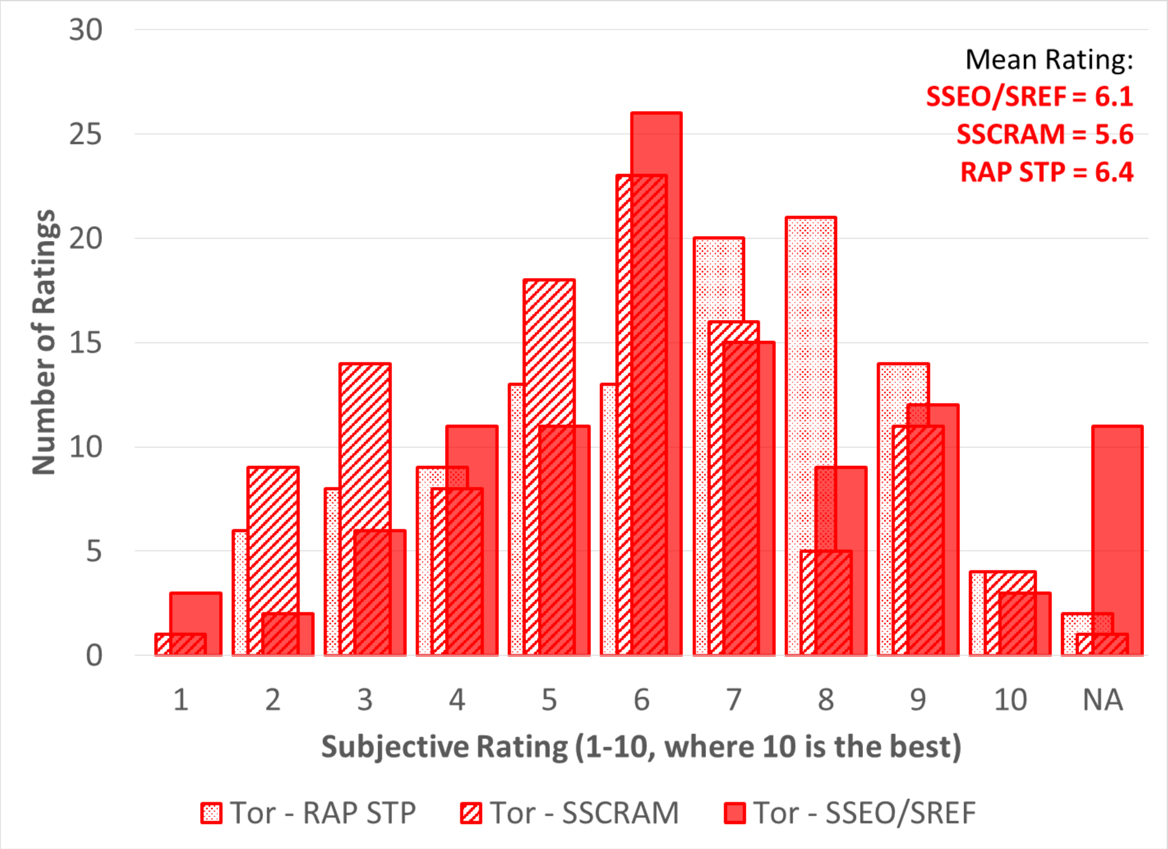


Figure 41 Same as Fig. 28, except for 1500 UTC 4-h tornado guidance valid from 2000 UTC to 0600 UTC for calibrated SSEO/SREF (solid fill), RAP STP/HRRR UH (pattern fill), SSCRAM (diagonal line).

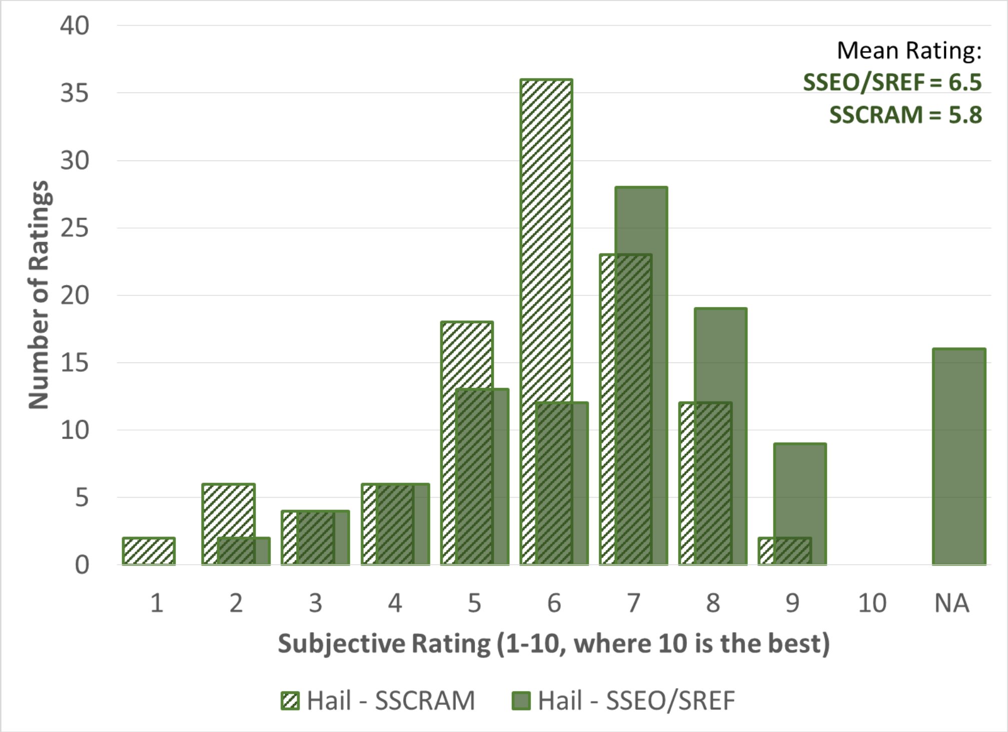


Figure 42 Same as Fig. 28, except for 1500 UTC 4-h severe hail guidance valid from 2000 UTC to 0600 UTC for calibrated SSEO/SREF and SSCRAM (diagonal line).

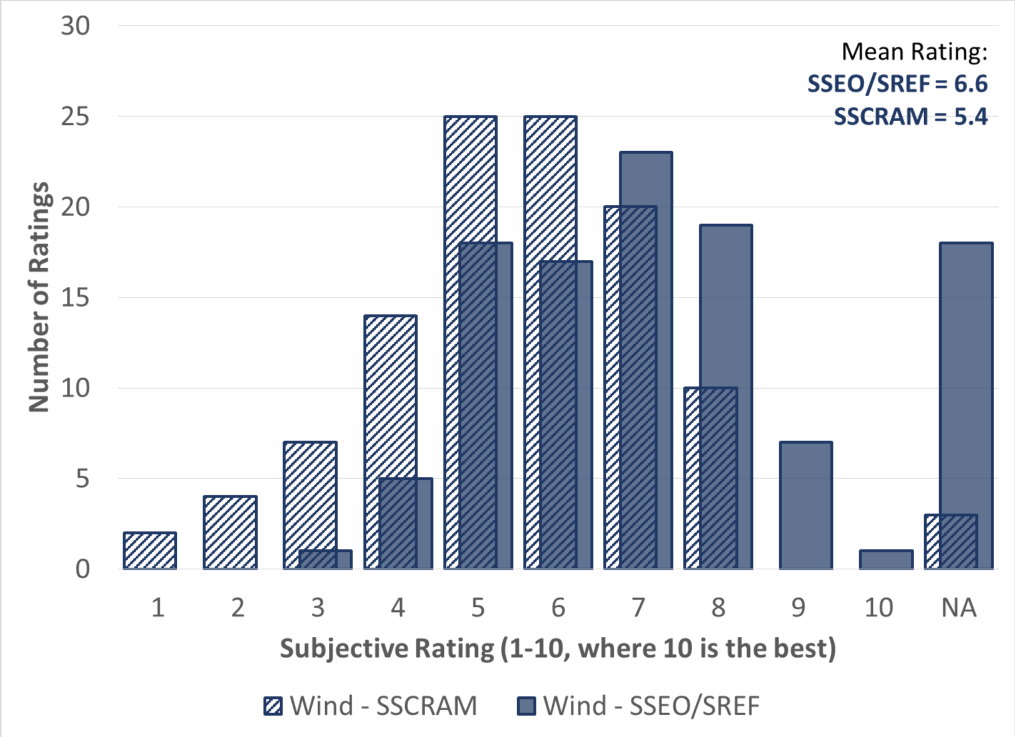


Figure 43 Same as Fig. 28, except for 1500 UTC 4-h severe wind guidance valid from 2000 UTC to 0600 UTC for calibrated SSEO/SREF (solid fill) and SSCRAM (diagonal line)

4. Summary

The 2017 Spring Forecasting Experiment (SFE2017) was conducted at the NOAA Hazardous Weather Testbed from 1 May – 2 June by the SPC and NSSL with participation from forecasters, researchers, model developers, university faculty and graduate students from around the world. The primary theme of SFE2017 was to utilize convection-allowing model and ensemble guidance in creating experimental high-temporal resolution probabilistic forecasts of severe weather hazards, including extension into the Day 2 and occasionally Day 3 periods. Furthermore, this was the second year that a major effort was made to closely coordinate CAM-based ensemble configurations into the Community Leveraged Unified Ensemble (CLUE). The CLUE allowed several carefully designed controlled experiments to be conducted that were geared towards identifying optimal configuration strategies for CAM-based ensembles.

Several preliminary findings/accomplishments from SFE2017 are listed below:

- Generated high temporal resolution outlooks for individual severe hazards (tornado, hail,wind) using first-guess guidance from a temporally disaggregated full-period outlook created with calibrated probabilistic guidance from a convection-allowing ensemble.
- Explored adding enhanced timing information by drawing severe weather isochrones, which delineated the start time of the 4-h time windows with the highest severe weather probability.
- Examined various convection-allowing ensemble systems within the CLUE using the SSEO as a baseline.
  - As designed, the HREFv2 performed similarly to the SSEO.
  - While all of the ensembles provided similar, useful guidance for Day 1 severe weather forecasting, the SSEO and HREFv2 received higher subjective ratings and verified slightly better in terms of objective metrics than the other systems.
  - These results suggest that the now operational HREFv2 will serve as a meaningful baseline against which experimental and next-generation CAM ensembles should be compared for consideration of operational implementation.
- Tested a prototype Warn-on-Forecast short-term prediction system for the first time in real-time during an afternoon forecasting activity with very promising results.
- Utilized several convection-allowing models and ensembles for creating Day 2 and Day 3 severe weather outlooks for individual severe hazards, noting utility beyond the Day 1 period.
- Examined CAM ensemble-based first-guess tornado probability guidance and found that inclusion of environment information improved storm attribute-only methods, and output derived from an environment (STP)-tornado climatology statistical approach performed best.
- Found that HRRRv3 had higher subjectively rated reflectivity forecasts than HRRRv2, supporting the eventual operational implementation of HRRRv3.

- Examined real-time, storm-scale FV3 simulations for the first time during SFE2017.
    - Subjective ratings revealed that FV3 reflectivity forecasts were often comparable to operational CAMs.
    - An objective comparison to the 3-km grid-spacing NSSL-WRF using the surrogate UH severe method found that FV3 forecasts exhibited lower skill compared to the NSSL-WRF, which is historically one of the better-performing CAMs.
    - These results support continued research to refine and improve FV3 for storm-scale applications before it is implemented operationally as part of an emerging unified NOAA model production suite.
- Evaluated 4-h calibrated hazard guidance from SPC using a variety of approaches.

Overall, SFE2017 was successful in testing new forecast products and modeling systems to address relevant issues related to the prediction of hazardous convective weather. The findings and questions generated during SFE2017 directly promote continued progress to improve forecasting of severe weather in support of the NWS Weather-Ready Nation initiative.

Acknowledgements

SFE2017 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 Oklahoma, and EMC were vital to the success of SFE2017. In particular, Ming Xue (CAPS), Fanyou Kong (CAPS), Kevin Thomas (CAPS), Keith Brewster (CAPS), Youngsun Jung (CAPS), Tim Supinie (CAPS), Glen Romine (NCAR), Craig Schwartz (NCAR), Ryan Sobash (NCAR), Kate Fossell (NCAR), Steve Willington (Met Office), Curtis Alexander (ESRL/GSD), Geoff Dimego (EMC), Jacob Carley (EMC), Brad Ferrier (EMC), Eric Aligo (EMC), S. J. Lin (GFDL), Lucas Harris (GFDL), and Xuguang Wang (OU/MAP) were essential in generating and providing access to model forecasts examined on a daily basis.

References

Adams-Selin R. and C. L. Ziegler, 2016: Forecasting hail using a one-dimensional hail growth model within WRF. *Mon. Wea. Rev.*, **144**, 4919-4939.

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.

Brooks, H. E., M. Kay, and J. A. Hart, 1998: Objective limits on forecasting skill of rare events. 19th Conf. on Severe Local Storms, Minneapolis, MN, Amer. Meteor. Soc., 552–555.

Clark, A. J., I. Jirak, S. J. Weiss, J. Kain, J. Correia, A. Dean, K. Knopfmeier, C. Karstens, B. Gallo, P. Heinselman, R. Hepper, G. Creager, and S. Dembek, 2017: Spring Forecasting Experiment 2017 Program Overview and Operations Plan. Available online at: http://hwt.nssl.noaa.gov/Spring_2017/HWT_SFE2017_operations_plan_FINAL.pdf.

Gallo, B. T., A. J. Clark, and S. R. Dembek, 2016: Forecasting tornadoes using convection-permitting ensemble. *Wea. Forecasting*, **31**, 273-295.

Gallo, B.T., and Coauthors, 2017: Breaking new ground in severe weather prediction: The 2015 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Wea. Forecasting*, **32**, 1541–1568, https://doi.org/10.1175/WAF-D-16-0178.1

Gallo, B. T., A. J. Clark, B. T. Smith, R. L. Thompson, I. Jirak, and S. R. Dembek, 2018: Blended probabilistic tornado forecasts: Combining climatological frequencies with NSSL-WRF ensemble forecasts. *Wea. Forecasting*, in review.

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.

Jones, T. A., K. Knopfmeier, D. Wheatley, G. Creager, P. Minnis, and R. Palikondo, 2016: Storm-scale data assimilation and ensemble forecasting with the NSSL Experimental Warn- on-Forecast System. Part 2: Combined radar and satellite data experiments. *Wea. Forecasting*, 31, 297–327.

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.

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

Melick, C. J., I. L. Jirak, J. Correia Jr., A.R. Dean, and S.J. Weiss, 2014: Exploration of the NSSL Maximum Expected Size of Hail (MESH) Product for Verifying Experimental Hail Forecasts in the 2014 Spring Forecasting Experiment. Preprints, 27th Conf. Severe Local Storms, Madison, WI.

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.

Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme phenomena in convection-allowing model forecasts. *Wea. Forecasting*, **26**, 714-728.

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

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

Wheatley, D. M., K. H. Knopfmeier, T. A. Jones, and G. J. Creager, 2015: Storm-scale data assimilation and ensemble forecasting with the NSSL Experimental Warn-on-Forecast System. Part 1: Radar data experiments. *Wea. Forecasting*, 30, 1795–1817.

APPENDIX A

Table A1 Daily activities schedule in local (CDT) time

Severe Hazards Desk	*Innovation 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 hourly total severe areas and isochrones • NEWS-e based initial and final total severe forecasts
0845 – 1115: **Day 1 Convective Outlook Generation**	
Hand analysis of 12Z upper-air maps and surface charts and select domain by 10 a.m.	
• 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 hourly coverage areas and total severe isochrones* for full-period total severe ≥15%
1115 – 1130: **Break**	
Prepare for map discussion	
1130 – 1200: **Map Discussion**	
Brief discussion of today's forecast challenges and products Topic of the day: CLUE, 3D Vis, Met Office, FV3, SPC Short-term Guidance	
1200 – 1300: **Lunch**	
Brief EWP (PHI prototype) 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** (as small groups on Chromebooks)	
• CLUE: SSEO (HREFv2) as baseline • HRRRv2, HRRRv3 • Deterministic CAMs (FV3, UM, HRW) • SPC Short-Term Guidance	• CLUE: Physics Experiment • Hail guidance • Tornado guidance • Microphysics (optional)
1500 – 1600: **Short-term Outlook Updates and NEWS-e**	
• Update 4-h probabilistic forecasts of tornado, wind, and hail valid 22-02Z* • Utilize SPC Short-Term Guidance	• NEWS-e evaluation • Utilize NEWS-e to generate preliminary and final hourly probabilistic forecasts of total severe valid 21-22Z and 22-23Z.

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

Table A2 List of weekly participants (with affiliation) during SFE2017. Facilitators/leaders for SFE2017 included: Adam Clark (NSSL), Kent Knopfmeier (CIMMS/NSSL), Israel Jirak (SPC), Dave Imy (retired SPC), Andy Dean (SPC), Jessica Choate (CIMMS/NSSL), Steve Willington (UKMO), Burkely Gallo (OU/NSSL), and MacKenzie Krojac (OU/NSSL).

Week 1	Week 2	Week 3	Week 4	Week 5
May 1-5	**May 8-12**	**May 15-19**	**May 22-26**	**May 30-June 2**
Nathan Wendt (SPC)	Eli Dennis (PSU)	Neil Taylor (MSC)	Austin Harris (WDTD)	Clark Evans (UWM)
Becky Adams-Selin (AER)	Bruce Entwistle (AWC)	Trevor Mitchell (UMan)	Lance Bosart (SUNYA)	Clark Evans student (UWM)
David Gagne (NCAR)	Bill Gallus (ISU; M-Th)	Kelly Lombardo (UConn)	Andrew Winters (SUNYA)	Steve Willington (UK Met)
Victor Gensini (COD)	Brian Squiteri (ISU; M-Th)	Kwinten Van Weverberg (UK Met)	Tomer Burg (SUNYA)	Katie Howard (UK Met)
Matt Pyle (EMC)	John Allen (CMICH)	Steve Willington (UK Met)	Harald Richter (BoM)	Paul Kocin (EMC)
Ben Blake (EMC)	Jeff Craven (MDL)	Katie Howard (UK Met)	Michael Colbert (PSU)	Jamie Wolff (DTC)
Brian Kolts (FirstEnergy)	Joshua Kastman (WPC)	Michael Bush (UK Met)	Steve Willington (UK Met)	Isidora Jankov (DTC/GSD)
Ryan Sobash (NCAR)	Lucas Harris (GFDL)	Jacob Carley (EMC)	Katie Howard (UK Met)	Lee Carlaw (WFO FWD)
Glen Romine (NCAR)	S-J Lin (GFDL; M-W)	Mallory Row (EMC)	Ben Albright (WPC; M-Th)	Evan Kuchera (USAF)
Terra Ladwig (GSD)	Matt Morin (GFDL; W-F)	Greg Thompson (NCAR; W-F)	Sarah Perfater (WPC; M-Th)	Bill Bua (UCAR/EMC)
Eric James (GSD)	Geoff Manikin (EMC)	Jason Milbrandt (MSC; W-F)	Corey Guastini (EMC)	Mike Evans (WFO BGM)
Todd Kluber (WFO MQT)	Tracey Dorian (EMC)	John Brown (GSD)	Michelle Harrold (DTC)	Eric Loken (OU)
Rich Otto (WPC)	Curtis Alexander (GSD)	David Dowell (GSD)	Jeff Beck (GSD)	Jeff Milne (OU/CIMMS/SPC)
	Ed Szoke (GSD)	Evan Bentley (WFO PQR)	Ryan Ellis (WFO RAH)	
	Phil Schumacher (WFO FSD)	Randy Bowers (WFO OUN)	Hendrik Tolman (NWS)	
	Corey Potvin (CIMMS/NSSL)	Monique Sellers (OCS; M-W)		

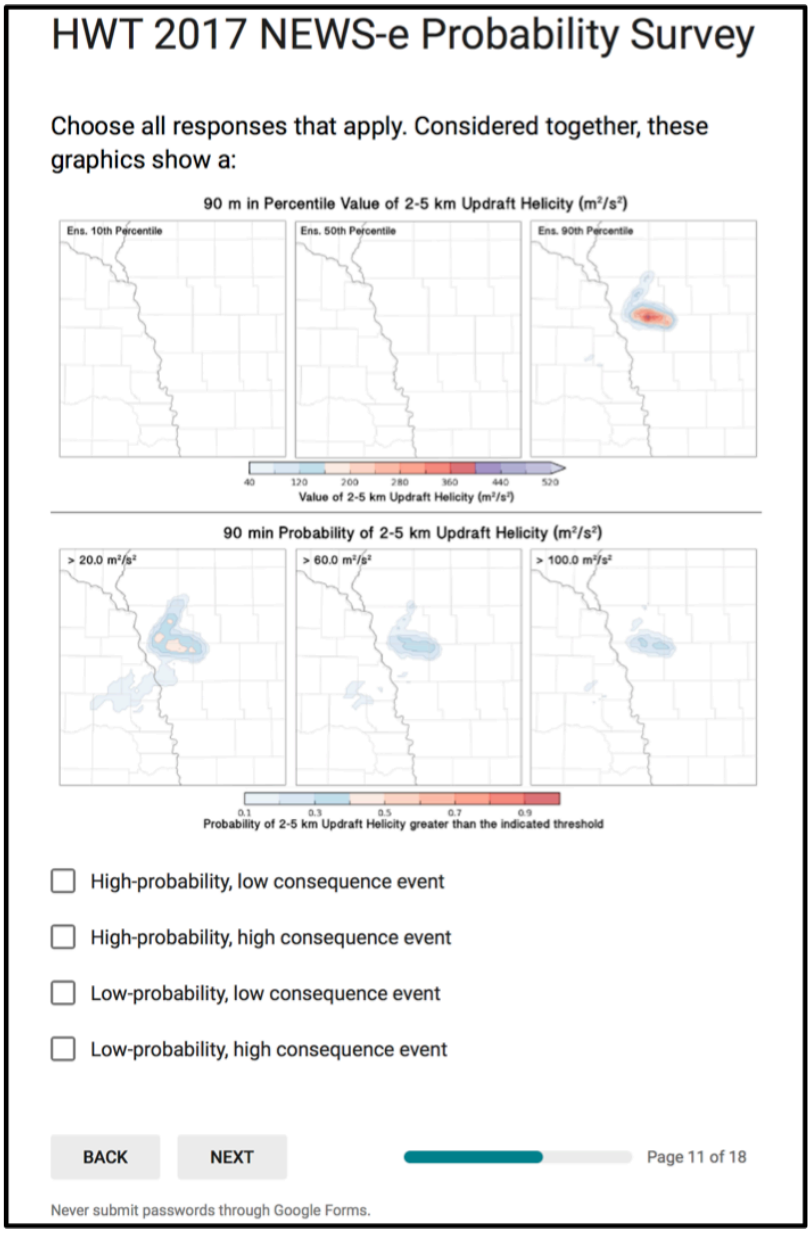
APPENDIX B. NEWS-e Survey Activity

This survey was administered to 62 participants as a part of the first evaluation of the experimental WoF system. A sample of participant demographic information is provided in Table B1.

Table B1. Sample results from the demographic questions asked of participating meteorologists for the NEWS-e survey. States in each region are: East North Central (Ohio, Indiana, Illinois, Michigan, and Wisconsin), Mid-Atantic (New York, New Jersey, and Pennsylvania), Mountain (Montana, Idaho, Wyoming, Colorado, New Mexico, Arizona, Utah, and Nevada), New England (Main, New Hampshire, Vermont, Massachusetts, Rhode Island, and Connecticut), Pacific (Washington, Oregon, California, Alaska, and Hawaii), South Atlantic (Delaware, Maryland, District of Columbia, Virginia, West Virginia, North Carolina, South Carolina, Georgia, and Florida), West North Central (Minnesota, Iowa, Missouri, North Dakota, South Dakota, Nebraska, and Kansas), West South Central (Arkansas, Louisiana, Oklahoma, and Texas).

What is your job title?	**# of responses**	**In which US region do you work?**	**# of responses**	**Describe your forecast experience**	**# of responses**
Assistant Professor	2	East North Central	6	Professional	15
Professor	4	Mid-Atlantic	7	Some/Hobby	16
Associate Professor	4	Mountain	11	Little to none	27
Meteorologist	8	New England	2		
Research Scientist	6	Pacific	1		
Support Scientist	4	South Atlantic	11		
Scientist	3	West North Central	6		
Physical Scientist	1	West South Central	9		
Forecaster	3	Not in US	7		
Graduate Student	8				
Postdoc	2				
Research Meteorologist	8				
Project Scientist	2				
Science and Operations Officer	2				
Hydrometeorologist	1				
Scientific Manager	1				
Chief Meteorologist	1				

Meteorologists were presented 12 open-ended and multiple-choice questions that queried their interpretation of NEWS-e ensemble products that provided measures of likelihood or severity at varying time and space scales. Sample survey questions include: "In an ensemble-based probabilistic forecast, what do you think the 70th percentile value of accumulated rainfall represents?" "Given the information presented, what is the probability of exceeding 0.5" of rainfall within box A?" Questions on the joint interpretation of probabilistic and percentile products were also included. An example of question format and visualization is seen in Figure B1. The analysis of responses to these and the other questions is in progress.


90 m in Percentile Value of 2-5 km Updraft Helicity (m²/s²)
Ens. 10th Percentile
Ens. 50th Percentile
Ens. 90th Percentile
40
120
200
280
360
440
520
Value of 2-5 km Updraft Helicity (m²/s²)
90 min Probability of 2-5 km Updraft Helicity (m²/s²)
> 20.0 m²/s²
> 60.0 m²/s²
> 100.0 m²/s²
0.1
0.3
0.5
0.7
0.9
Probability of 2-5 km Updraft Helicity greater than the indicated threshold
BACK
NEXT

HWT 2017 NEWS-e Probability Survey

Choose all responses that apply. Considered together, these graphics show a:

□ High-probability, low consequence event

□ High-probability, high consequence event

□ Low-probability, low consequence event

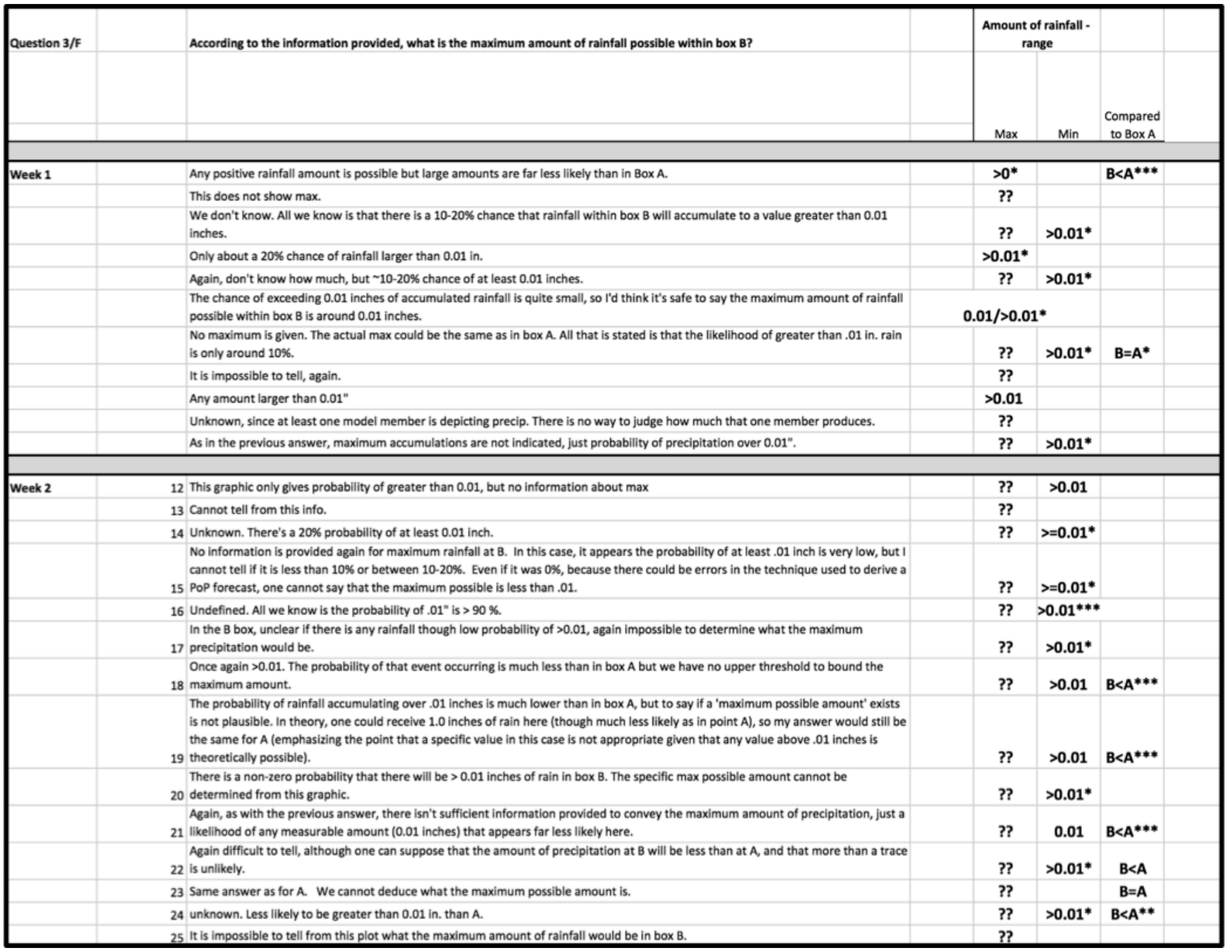
□ Low-probability, high consequence event

Page 11 of 18

Never submit passwords through Google Forms.

Figure B1. A screen capture of a question from the NEWS-e survey.

Initial analyses have focused on coding the various responses for each of the open-ended questions. Once a formal coding methodology was established, each open-ended question was assigned to two researchers to code all participants' responses for that question for all 5-weeks of the testbed. The coders would convene after independent coding to come to a consensus on how each participant's response to that assigned question should be accurately coded. An example of a coding scheme is provided in Figure B2. Current research is focused on providing a measure of inter-coder reliability. Results from the survey activity will be presented as an oral presentation during the 2017 Annual AMS Conference in Austin, Texas.



Question 3/F		According to the information provided, what is the maximum amount of rainfall possible within box B?	Amount of rainfall - range		
			Max	Min	Compared to Box A
Week 1		Any positive rainfall amount is possible but large amounts are far less likely than in Box A.	>0*		B<A***
		This does not show max.	??		
		We don't know. All we know is that there is a 10-20% chance that rainfall within box B will accumulate to a value greater than 0.01 inches.	??	>0.01*	
		Only about a 20% chance of rainfall larger than 0.01 in.	>0.01*		
		Again, don't know how much, but ~10-20% chance of at least 0.01 inches.	??	>0.01*	
		The chance of exceeding 0.01 inches of accumulated rainfall is quite small, so I'd think it's safe to say the maximum amount of rainfall possible within box B is around 0.01 inches.	0.01/>0.01*		
		No maximum is given. The actual max could be the same as in box A. All that is stated is that the likelihood of greater than .01 in. rain is only around 10%.	??	>0.01*	B=A*
		It is impossible to tell, again.	??		
		Any amount larger than 0.01"	>0.01		
		Unknown, since at least one model member is depicting precip. There is no way to judge how much that one member produces.	??		
		As in the previous answer, maximum accumulations are not indicated, just probability of precipitation over 0.01".	??	>0.01*	
Week 2	12	This graphic only gives probability of greater than 0.01, but no information about max	??	>0.01	
	13	Cannot tell from this info.	??		
	14	Unknown. There's a 20% probability of at least 0.01 inch.	??	>=0.01*	
	15	No information is provided again for maximum rainfall at B. In this case, it appears the probability of at least .01 inch is very low, but I cannot tell if it is less than 10% or between 10-20%. Even if it was 0%, because there could be errors in the technique used to derive a PoP forecast, one cannot say that the maximum possible is less than .01.	??	>=0.01*	
	16	Undefined. All we know is the probability of .01" is > 90 %.	??	>0.01***	
	17	In the B box, unclear if there is any rainfall though low probability of >0.01, again impossible to determine what the maximum precipitation would be.	??	>0.01*	
	18	Once again >0.01. The probability of that event occurring is much less than in box A but we have no upper threshold to bound the maximum amount.	??	>0.01	B<A***
	19	The probability of rainfall accumulating over .01 inches is much lower than in box A, but to say if a 'maximum possible amount' exists is not plausible. In theory, one could receive 1.0 inches of rain here (though much less likely as in point A), so my answer would still be the same for A (emphasizing the point that a specific value in this case is not appropriate given that any value above .01 inches is theoretically possible).	??	>0.01	B<A***
	20	There is a non-zero probability that there will be > 0.01 inches of rain in box B. The specific max possible amount cannot be determined from this graphic.	??	>0.01*	
	21	Again, as with the previous answer, there isn't sufficient information provided to convey the maximum amount of precipitation, just a likelihood of any measurable amount (0.01 inches) that appears far less likely here.	??	0.01	B<A***
	22	Again difficult to tell, although one can suppose that the amount of precipitation at B will be less than at A, and that more than a trace is unlikely.	??	>0.01*	B<A
	23	Same answer as for A. We cannot deduce what the maximum possible amount is.	??		B=A
	24	unknown. Less likely to be greater than 0.01 in. than A.	??	>0.01*	B<A**
	25	It is impossible to tell from this plot what the maximum amount of rainfall would be in box B.	??		

*Figure B2. Sample coding scheme used to sort response to the survey questions. Not shown: description of tags specifying different characteristics of a response. For example, "***" refers to a participant with a high level of certainty.*