SPRING FORECASTING EXPERIMENT 2019

Conducted by the

EXPERIMENTAL FORECAST PROGRAM

of the

NOAA HAZARDOUS WEATHER TESTBED

http://hwt.nssl.noaa.gov/sfe/2019

HWT Facility – National Weather Center
29 April - 31 May 2019

Preliminary Findings and Results

Adam Clark[2], Israel Jirak[1], Burkely Gallo[1,3], Brett Roberts[1,2,3], Kent Knopfmeier[2,3], Jake Vancil[1,3], Andy Dean[1], Pam Heinselman[2], Makenzie Krocak[2,3,4], Jessica Choate[2,3], Katie Wilson[2,3], Patrick Skinner[2,3], Yunheng Wang[2,3], Gerry Creager[2,3], Larissa Reames[2,3], Louis Wicker[2], Scott Dembek[2,3], and Steve Weiss[3]

(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

Table of Contents

Scenes from the 2019 NOAA Hazardous Weather Testbed Spring Forecasting Experiment (photo credit: James Murnan, NSSL)

1. Introduction

The 2019 Spring Forecasting Experiment (2019 SFE) was conducted from 29 April – 31 May 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, NOAA Earth Systems Research Laboratory/Global Systems Division (ESRL/GSD), NOAA Geophysical Fluid Dynamics Laboratory (GFDL), United Kingdom Meteorological Office (Met Office), National Center for Atmospheric Research (NCAR), and NOAA/NCEP's Environmental Modeling Center (EMC). Participants included about 100 forecasters, researchers, model developers, university faculty and graduate students from around the world (see Table 1 in the Appendix). As in previous years, the 2019 SFE aimed to test emerging concepts and technologies designed to improve the prediction of hazardous convective weather, consistent with the Forecasting a Continuum of Environmental Threats (FACETs; Rothfusz et al. 2018) 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 provide enhanced information on the conditional intensity of tornado, wind, and hail events by delineating areas expected to follow "normal", "hatched", or "double-hatched" distributions.
- Explore methods to include more detailed timing information within Day 1 & 2 severe weather outlooks by issuing potential severe timing (PST) areas, which are enclosed areas valid for 4-h periods that highlight the time window when the majority of severe weather reports are expected to occur.
- Test the feasibility of generating several types of short-term (0-6 h) convective outlooks valid over different time windows and lead times using a prototype WoF system.
- Test the utility of a prototype WoF system for updating full period hazard forecasts valid 2100-1200 UTC.

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 operational High-Resolution Ensemble Forecast System Version 2.1 (HREFv2.1) as a baseline.
- Compare and assess different approaches in CAM ensembles for predicting hail size.
- Compare and assess the current version of HREFv2.1 with possible future configurations that include forecasts from the EMC Stand-Alone-Regional Finite Volume Cubed Sphere Model (SARFV3) with and without HRRRv3, and without NMMB members.

- Evaluate 3-km grid-spacing SARFV3 configurations that have different microphysics, boundary layer, and land surface schemes.
- Evaluate the prototype WoF ensemble for applications to short-term severe weather outlook generation, and compare forecast utility to a time-lagged HRRR ensemble.
- Using the CLUE, evaluate whether ensemble sensitivity-based subset probabilities provide improved guidance relative to the full CLUE ensemble.
- Gauge forecast skill from a 2.2 km grid-spacing ensemble run by the UK Met Office relative to HREFv2.1 and other ensembles in the CLUE. Additionally, compare two unique physics configurations within the UK Met Office members.
- Evaluate the utility of an object-based approach for efficiently visualizing and deriving probabilities from a CAM ensemble.
- Evaluate the severe weather forecasting utility of a SARFV3 ensemble.

As in previous experiments, a suite of state-of-the-art experimental CAM guidance contributed by our large group of collaborators was central to the 2019 SFE. Additionally, for the fourth consecutive year, these contributions were formally coordinated into a single ensemble framework called the Community Leveraged Unified Ensemble (CLUE). The 2019 CLUE was constructed by having all groups coordinate as closely as possible on model specifications (e.g., grid-spacing, vertical levels, physics, etc.), domain, and post-processing 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 2019 CLUE included 104 members using 3-km grid-spacing (except for the 2.2 km grid-spacing UK Met Office members) that allowed a set of six unique experiments. The 2019 SFE activities also involved testing a prototype WoF system for the third consecutive year.

This document summarizes the activities, core interests, and preliminary findings of the 2019 SFE. 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 (https://hwt.nssl.noaa.gov/sfe/2019/docs/HWT_SFE2019_operations_plan.pdf). The remainder of this document is organized as follows: Section 2 provides an overview of the models and ensembles examined during the 2019 SFE along with a description of the daily activities, Section 3 reviews the preliminary findings of the 2019 SFE, and Section 4 contains a summary of these findings.

2. Description

a) Experimental Models and Ensembles

Building upon successful experiments of previous years, the 2019 SFE 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 was central to the generation of these forecasts. 133 unique CAMs were run for the 2019 SFE, of which 104 were a part of the CLUE system. Other deterministic and ensemble CAMs outside of the CLUE were contributed by NSSL (WoF system), EMC (HREFv2.1), and GSD (HRRRv4). To put the volume of CAMs run for 2019 SFE 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 three years has made the increase in members more manageable and has allowed for more controlled scientific comparisons.

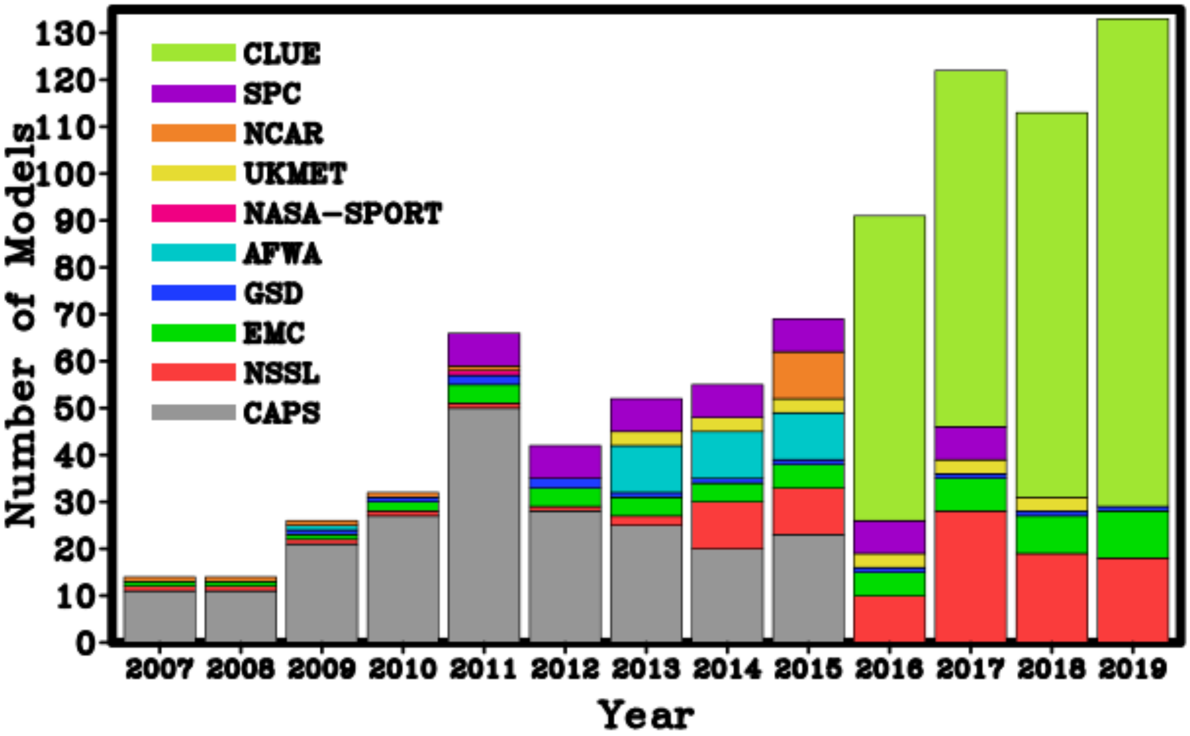


Figure 1 Number of CAMs run for SFEs since 2007. The different colored stacked bars indicate the contributing agencies.

More information on all of the modeling systems run for the 2019 SFE is given below.

1) THE COMMUNITY LEVERAGED UNIFIED ENSEMBLE (CLUE)

The 2019 CLUE is a carefully designed ensemble with subsets of members contributed by NOAA groups at NSSL, GFDL, ESRL/GSD, and EMC, and non-NOAA groups at CAPS (OU), MAP (OU), NCAR, and the UK Met Office. To ensure consistent post-processing, visualization, and verification, all CLUE contributors used the similar post-processing software to output the same set of model output fields on the same grid. Exceptions were the FV3 and UK Met Office runs, which required different software for post-processing, but the fields were output to the CLUE 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. The FV3 runs did not contain the full set of fields as all the other CLUE runs since development of FV3 diagnostics and post-processing remains in progress, and the UK Met Office runs also only included a subset of the output fields that were available from the other members. All CLUE members, except for the UK Met Office ones, were initialized weekdays at 0000 UTC with 3-km grid-spacing covering a CONUS domain. The UK Met Office members included 9 members with a 6-h time lag (i.e., 1800 UTC initialization) and all the UK Met Office members used an approximately 3/4 CONUS domain. A full description of all members and list of post-processed model fields are provided in the 2019 SFE operations plan (Clark et al. 2019). Table 1 provides a summary of each CLUE subset.

Table 1 Summary of 2019 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.

Clue Subset	**# of mems**	**IC/LBC perturbations**	**Mixed Physics**	**Data Assimilation**	**Model Core**	**Agency**
CAPS FV3	9	SREF	yes	cold start	FV3	CAPS (OU)
fv3-phys	7	none	yes	cold start	FV3	CAPS (OU)
wrf-exp	4	none	no	3DVAR	ARW	CAPS (OU)
CAPS EnKF	10	EnKF (CAPS)	yes	EnKF	ARW	CAPS (OU)
HRRRv3	1	none	no	GSI Ens-Var	ARW	ESRL/GSD
HRRRv4	1	none	no	GSI Ens-Var	ARW	ESRL/GSD
gsd-sarfv3	1	none	no	cold start	FV3	ESRL/GSD
HRRRE	9	EnKF	no	EnKF	ARW	ESRL/GSD
ncar	10	EAKF (DART)	no	EAKF (DART)	ARW	NCAR
map-hybrid	10	EnKF-Var hybrid (GSI)	no	EnKF-Var hybrid (GSI)	ARW	MAP (OU)
map-ICpert	10	EnKF-Var hybrid (GSI) w/ GEFS	no	EnKF-Var hybrid (GSI) w/ GEFS	ARW	MAP (OU)
HRRRE-nospp	9	EnKF	no	EnKF	ARW	NSSL
nssl-fv3	1	none	no	cold start	FV3	NSSL
nssl-sarfv3	1	none	no	cold start	FV3	NSSL
gfdl-fv3	1	none	no	cold start (GFS)	FV3	GFDL
ukmet-sphys	9	MOGREPS-G	no	cold start	UM	UK Met Office
ukmet-mphys	9	MOGREPS-G	yes	cold start	UM	UK Met Office
emc-fv3	1	none	no	cold start	FV3	EMC
emc-sarfv3	1	none	no	cold start	FV3	EMC

The design of the 2019 CLUE allowed for 6 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 the 2019 SFE. The CLUE subsets listed are from Table 1.

Experiment Name	**Description**	**CLUE subsets**
Physics perturbation experiment	The HRRRE uses stochastic parameter perturbations (SPP) applied to land-surface, PBL, and microphysics. A control ensemble run by NSSL was configured identically to the HRRRE, but had the SPP settings turned off. **Goal: Examine whether SPP results in practical forecast differences, and objectively quantify gains in ensemble spread and skill post-experiment.**	HRRRE & HRRRE-nospp
UM physics configurations	Two sets of members were contributed by the UK Met Office. One used the operational physics suite from MOGREPS-UK and the other used a development suite tuned for the tropics. **Goal: compare the physics suites and gauge skill against US systems such as the HREF.**	ukmet-sphys & ukmet-mphys
SARFV3 physics sensitivities	CAPS provided SARFV3 members that used the same ICs/LBCs, but different physics. Three microphysics, three PBL, and two land-surface models were tested using CPPP (Community Common Physics Package). **Goal: Gauge skill and quantify sensitivities in SARFV3.**	fv3-phys
Global-with-nest vs. SARFV4	EMC ran a global FV3 configuration with a high-resolution nest over the CONUS and a SARFV3 configurations with ICS/LBCs from the GFS. **Goals: Examine whether SARFV3 performance is similar to global-with-nest. Examine whether there is noticeable degradation at later hours from LBC errors in SARFV3.**	emc-fv3 & emc-sarfv3
HRRRv3 vs. HRRRv4 vs. gsd-sarfv3	HRRRv4 is the last implementation of HRRR that uses WRF-ARW. Subsequent versions will use SARFV3 once performance reaches HRRRv4. **Goal: Compare performance in HRRRv3, HRRRv4 and gsd-sarfv3.** ***NOTE: Configuration changes were made to HRRRv4 during the experiment and gsd-sarfv3 was only available for a limited period near the end of the experiment.***	HRRRv3, HRRRv4, & gsd-sarfv3
Data assimilation comparisons	Several ensembles used EnKF-based data assimilation. In the most controlled comparison, OU/MAP ran two ensembles that shared a control member with ICs from GSI-based hybrid EnVar where ensemble covariance and storm scale static covariances were combined. In one ensemble, ensemble member ICs were generated from the GSI EnKF analyses (multi-scale perturbations), in the other, ICs were generated by applying 0000 UTC GEFS perturbations to the re-centered GSI EnKF analyses (large-scale perturbations). **Goal: Examine which DA techniques are optimally suited for severe weather forecasting.**	map-hybrid, map-ICpert, CAPS EnKF, ncar, & hrrre

2) HIGH RESOLUTION ENSEMBLE FORECAST SYSTEM VERSION 2.1 (HREFv2.1)

The HREFv2.1 is a 10-member CAM ensemble currently running at EMC with forecasts that can be viewed at: http://www.spc.noaa.gov/exper/href/. HREFv2.1 members use different physics, model cores (ARW and NMMB), initial and lateral boundary conditions (NAM and RAP), and half of the members are 12-h time lagged. HREFv2 was implemented operationally on 1 November 2017 and was recently updated to include two HRRR members (one 6-h time lagged). The design of HREFv2.1 originated from the SSEO, which demonstrated skill during the previous six years in the HWT and SPC prior to HREFv2 operational implementation. All members, except for the NAM CONUS Nest and HRRR, are initialized with a "cold-start". Forecasts to 36 h are produced at 0000 and 1200 UTC. The diversity in HREFv2 has proven to be a very effective configuration strategy, and it has consistently outperformed all other CAM ensembles examined in the HWT during the last few years. Thus, HREFv2 performance is considered the baseline against which potential future operational CAM ensemble configurations are compared.

3) NSSL EXPERIMENTAL WARN-ON-FORECAST SYSTEM (WoFS; credit: Kent Knofpmeier)

The WoFS is a 36-member WRF-based ensemble data assimilation system used to produce very short-range (0-6 h) probabilistic 18-member forecasts to assist in the prediction of hazardous weather, such as tornadoes, hail, high winds, and flash flooding. The WoFS starts from the HRRRE and is updated by hourly GSI-EnKF data assimilation of conventional observations and Multi-Radar/MultiSensor (MRMS) radar reflectivity from 0300 UTC to 1800 UTC Day 1. A 36-h ensemble forecast launched form the 1200 UTC HRRRE analysis was used to provide boundary conditions for the WOFS for the period 1800 UTC Day – 0300 UTC Day 2. Similarly, a 1-h ensemble forecast launched form the 1700 UTC HRRRE analysis was used to provide ICs for the WOFS at 1800 UTC.

The daily WoFS domain targeted the primary region where severe weather was anticipated and covered a 900-km square region with frequent updates. All ensemble members used the NSSL 2-moment microphysics parameterization and the RAP land-surface model, but the PBL and radiation physics options were varied. MRMS radar reflectivity and NEXRAD Level II radial velocity data, GOES-16 cloud water path retrievals, and Oklahoma Mesonet observations (when available) were assimilated every 15-min also using the GSI-EnKF method, beginning at 1800 UTC each day. ASOS data was assimilated at 15 minutes past each hour. 6-h (3-h) ensemble forecasts were initialized from the WoFS analysis each hour (half-hour) from 1900 UTC Day 1 through 0300 UTC Day 2. These forecasts were viewable using the web-based WoFS Forecast Viewer (https://www.nssl.noaa.gov/projects/wof/WoFS/realtime/).

b) Daily Activities

SFE 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 A2).

1) EXPERIMENTAL FORECAST PRODUCTS

Similar to previous years, the experimental forecasts explored the ability to add temporal specificity to traditional SPC severe weather outlooks within the Day 1 & 2 time period. Additionally, the feasibility of providing more precise information on the intensity of specific hazards was explored. The participants were split into two desks, with those at the Innovation Desk forecasting the total severe threat (combining hail, wind, and tornado hazards) and those at the Severe Hazards Desk forecasting individual hazards. The experimental forecasts covered a limited-area domain where the primary severe weather threat was expected based on existing SPC outlooks and/or where interesting convective forecast challenges were expected.

At the Severe Hazards Desk, the first forecast was completed as a group and mimicked the SPC operational Day 1 Convective Outlooks comprised of individual probabilistic forecasts of large hail, damaging wind, and tornadoes within 25 miles (40 km) of a point valid 1600 to 1200 UTC the next day. Then, the Severe Hazards Desk issued conditional intensity forecasts of tornado, wind, and hail, in which areas were delineated where reports were expected to follow a "normal", "hatched", or "double-hatched" distribution, which corresponded to significant severe weather being unlikely, possible, and expected, respectively (Table A3 contains more information on category definitions). After these two sets of forecasts were completed as a group, the 1600-1200 UTC outlook for individual hazards was temporally disaggregated into 4-h periods using HREF/SREF calibrated hazards guidance to provide automated timing information on the severe weather threat, as has been provided in previous SFEs. Then, the Severe Hazards Desk participants split into five groups and used a web interface to generate their own set of coverage and intensity forecasts using Google Chromebooks. Each group was assigned a specific CLUE subset for this task to assess how the different subsets would influence the human-generated forecasts. The subsets included CAPS EnKF, HRRRE, CAPS FV3, map-hybrid, and HREFv2.1.

At the Innovation Desk, the first forecast also covered 1600-1200 UTC and was conducted as a group. Rather than individual hazards, the Innovation Desk issued probabilistic outlooks for total severe (combined tornado, hail, and wind). Similar to the 2018 SFE, the Innovation Desk then issued Potential Severe Timing areas (PSTs), which were delineated within areas of 15% total severe probabilities from the previous activity. The PSTs indicate the 4 h time period over which a highlighted region is expected to experience the highest severe weather risk within the full outlook time period, and are aimed toward communicating more detailed timing information to the emergency management community. The PSTs were issued independently by the lead forecaster at the Innovation Desk on the N-AWIPS machine and by five participant groups using a web interface to generate their own set of PSTs using Google Chromebooks. Similar to the Severe Hazards Desk activity, each group was assigned a specific CLUE subset to use for this task that included CAPS EnKF, HRRRE, CAPS FV3, map-hybrid, and HREFv2.1. After issuing the PSTs, the Innovation Desk regrouped and discussed the forecasts and behavior of the CLUE subsets. This approach was meant to engage the participants more directly with the CLUE subsets, since in prior years participants only interacted with CLUE subsets through facilitator-led discussions. Despite the different end products at each desk, the goals of the activities were the same as in prior years – namely to explore different ways of introducing probabilistic severe weather forecasts on time/space scales that are not currently addressed with categorical forecast products (e.g., SPC Mesoscale Discussions and Severe Thunderstorm/Tornado Watches), and to begin to explore ways of seamlessly bridging probabilistic severe weather outlooks and probabilistic severe weather warnings as part of the NOAA WoF and FACETS initiatives.

After both desks issued their morning outlooks, there was a map discussion open to all tenants of the National Weather Center summarizing forecast challenges and highlighting interesting findings from the previous day. Each day of the week also featured a brief discussion of a "special topic" (Table A2). After lunch, the Severe Hazards Desk issued Day 2 full-period (i.e., 1200-1200 UTC the next day) probabilistic forecasts of tornado, wind, and hail, as well as conditional intensity forecasts, over a regional area of interest, which was done as a group activity. Similarly, the Innovation Desk issued full-period probabilistic forecasts of total severe, as well as 4-h PSTs, which were both conducted as a group. Later in the afternoon, scientific evaluations were conducted (summarized in the next section).

For the final activity of the day on Tuesday through Friday, forecasting activities using the WoFS were conducted on both desks from 3-4pm. On Mondays, a training activity for the WoF activity occurred from 3-4pm for SFE participants at both desks. At the Severe Hazards Desk, on Tuesday-Friday, participants updated their full period (2100-1200 UTC) hazard probability and conditional intensity forecasts in the same small groups as the morning activity and using the Chromebooks. At the Innovation Desk, forecasts were drawn by facilitators (Adam Clark and Burkely Gallo) and informed by small groups of participants interrogating the WoFS data on their Chromebooks or personal laptops. Probabilistic total severe outlooks were issued that were valid for short (1-h) and long (4-h) time windows. Additionally, an outlook was issued that covered a "targeted" 1-h time window valid 8-9pm (0100-0200 UTC). The short (1-h) outlook was issued during the 3-3:30pm time period using the 1930 UTC initialization[1] of the WoFS and was valid 2100-2200 UTC (4-5pm). The long (4-h) outlook was issued during the 3:30-4pm time period using the 2000 UTC initialization of the WoFS and was valid 2200-0200 UTC. Finally, the targeted (1-h) outlook was also issued during the 3:30-4pm time period using the 2000 UTC WoFS initialization and was valid 0100-0200 UTC. After these outlooks at both desks were issued, SFE activities concluded for the majority of participants. However, for two forecasters that were selected for a WoFS-based evening activity, additional sets of outlooks were issued each hour from 4-8pm. The evening forecasters issued these outlooks individually on the Chromebooks. Additionally, evaluation activities were conducted at 6, 7, and 8pm. NSSL facilitators were on hand every evening to assist in the forecast generation and evaluation process. Table 3 summarizes the Innovation Desk WoFS activities.

Table 3 Schedule for experimental outlooks and evaluations based on the WoFS. The yellow, green, and blue shaded cells indicate the short, long, and targeted outlooks, respectively, while the unshaded cells indicate times at which subjective evaluations of earlier forecasts were conducted.

	Experiment Time	**Outlook Valid Time**	**WoF Ensemble Initialization**
All Participants (Tues – Fri)	3:00 – 3:30 PM	4:00 – 5:00 PM	2:30 PM (1930 UTC)
	3:30 – 4:00 PM	5:00 – 9:00 PM	3:00 PM (2000 UTC)
	3:30 – 4:00 PM	8:00 – 9:00 PM	3:00 PM (2000 UTC)
Evening Participants (Mon – Thurs)	4:00 – 4:30 PM	5:00 – 6:00 PM	3:30 PM (2030 UTC)
	4:30 – 5:00 PM	6:00 – 10:00 PM	4:00 PM (2100 UTC)
	4:30 – 5:00 PM	8:00 – 9:00 PM	4:00 PM (2100 UTC)
	5:00 – 5:30 PM	6:00 – 7:00 PM	4:30 PM (2130 UTC)
	5:30 – 6:00 PM	7:00 – 11:00 PM	5:00 PM (2200 UTC)
	5:30 – 6:00 PM	8:00 – 9:00 PM	5:00 PM (2200 UTC)
Evaluation	6:00 PM	4:00 – 5:00 PM	2:30 PM (1930 UTC)
	6:00 – 6:30 PM	7:00 – 8:00 PM	5:30 PM (2230 UTC)
	6:30 – 7:00 PM	8:00 PM – 12:00 AM	6:00 PM (2300 UTC)
	6:30 – 7:00 PM	8:00 – 9:00 PM	6:00 PM (2300 UTC)
Evaluation	7:00 PM	5:00 – 6:00 PM	3:30 PM (2030 UTC)
	7:00 – 7:30 PM	8:00 – 9:00 PM	6:30 PM (2330 UTC)
	7:30 – 8:00 PM	9:00 PM – 1:00 AM	7:00 PM (0000 UTC)
Evaluation	8:00 PM	6:00 – 7:00 PM	4:30 PM (2130 UTC)

[1] Oftentimes, 1900 UTC initializations were used because of slight delays in the 1930 UTC runs.

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 the 2019 SFE 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 the SFE. 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.

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. These evaluations are discussed in detail in Sections 3c and 3d.

3. Preliminary Findings and Results

a) Evaluation of experimental forecast products – Innovation Desk

1) CONVECTIVE OUTLOOK EVALUATIONS (credit: A. Clark)

The first forecasting activity of each day at the Innovation Desk was the generation of a Day 1 group probabilistic forecast of any severe hazard valid 1600 – 1200 UTC. A similar Day 2 outlook valid 1200 – 1200 UTC was issued in the afternoon. These outlooks were rated the next day by overlaying the forecast with Local Storm Reports (LSRs) and warnings. A "practically perfect" forecast (Hitchens et al. 2013) was also generated from the LSRs and displayed alongside the experimental forecasts for reference. Contours matching current SPC operational probability thresholds (5, 15, 30, 45, and 60%) could be issued, as well as 10% or greater probability of a significant severe weather event within 25 miles of a point. Example Day 1 and Day 2 experimental outlooks along with the corresponding practically perfect outlooks are shown in Figure 2.

In general, participants thought that the Day 1 outlooks performed well (mean rating of 7.56/10; Fig. 3a). The Day 2 outlooks also received fairly high ratings (mean rating of 6.13/10), but lower ratings were much more frequent than Day 1, and none of the Day 2 outlooks were rated 10/10. Outlooks were given slightly better ratings on high-end days, which are defined as days when the practically perfect forecast indicated a 45% or greater probability (Fig. 4). 15 of 23 experiment days were high-end according to this criterion, while 7 of these days included experimental outlooks with 45% or greater probabilities. Forecast probability magnitudes generally matched the practically perfect forecasts within a categorical outlook category; only one day had a category that differed from the verification by two categories (Table 4). Comments indicated that the participants focused most on the location and magnitude of the probabilities, penalizing large extents of false alarm, but rewarding outlooks that captured all or most of the reports.

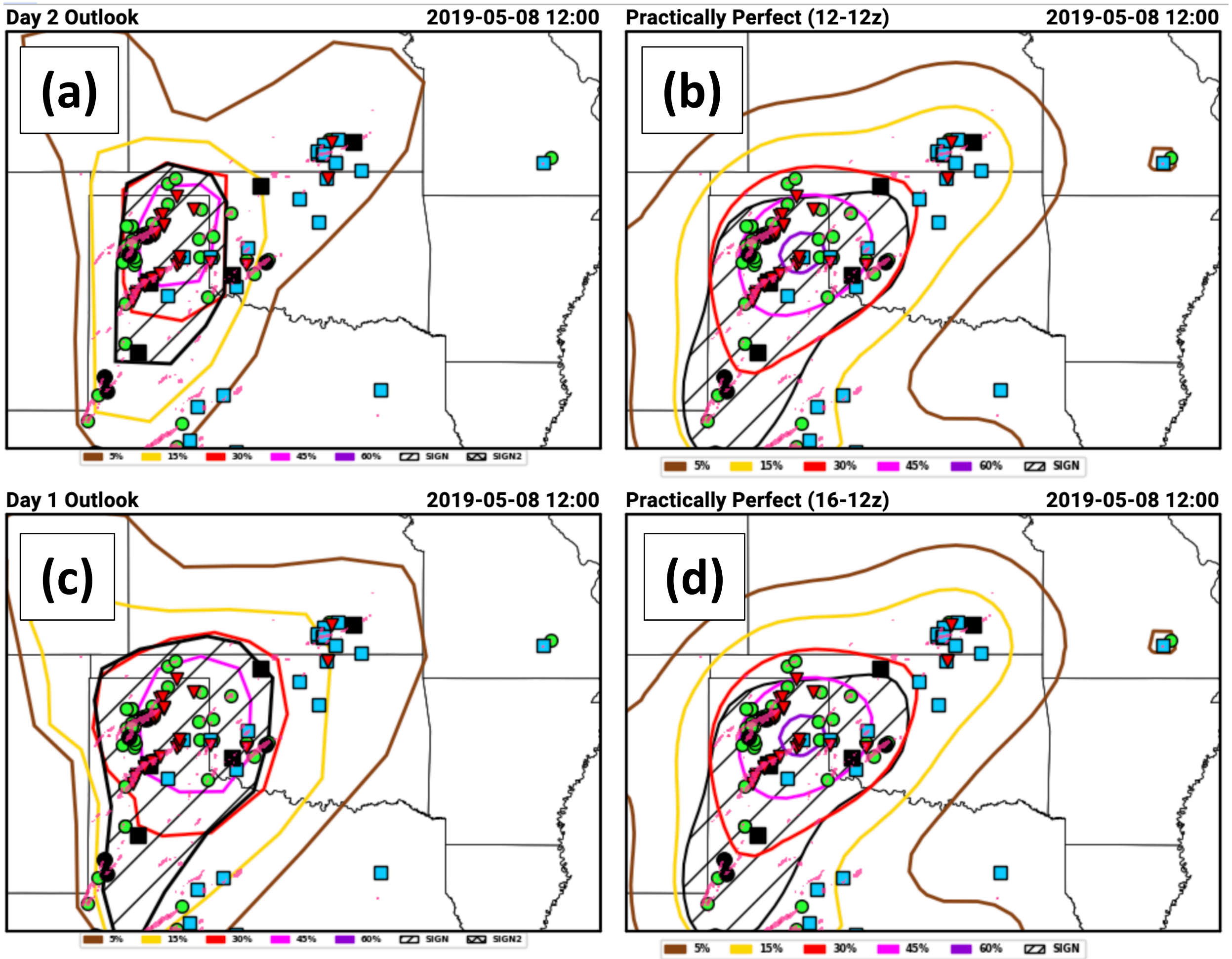


Figure 2 An experimental Day 2 outlook (a) and practically perfect forecast (b) overlaid with wind (blue squares), significant wind (black squares), hail (green circles), significant hail (black circles), and tornado (red inverted triangles) reports for 8 May 2019. The brown, yellow, red, magenta, and purple contours indicate 5%, 15%, 30%, 45%, and 60% probability of severe weather within 25 miles of a point. Hatched areas indicate a 10% or greater chance of a significant severe report within 25 miles. (c) - (d) same as (a) - (b), except for Day 1.

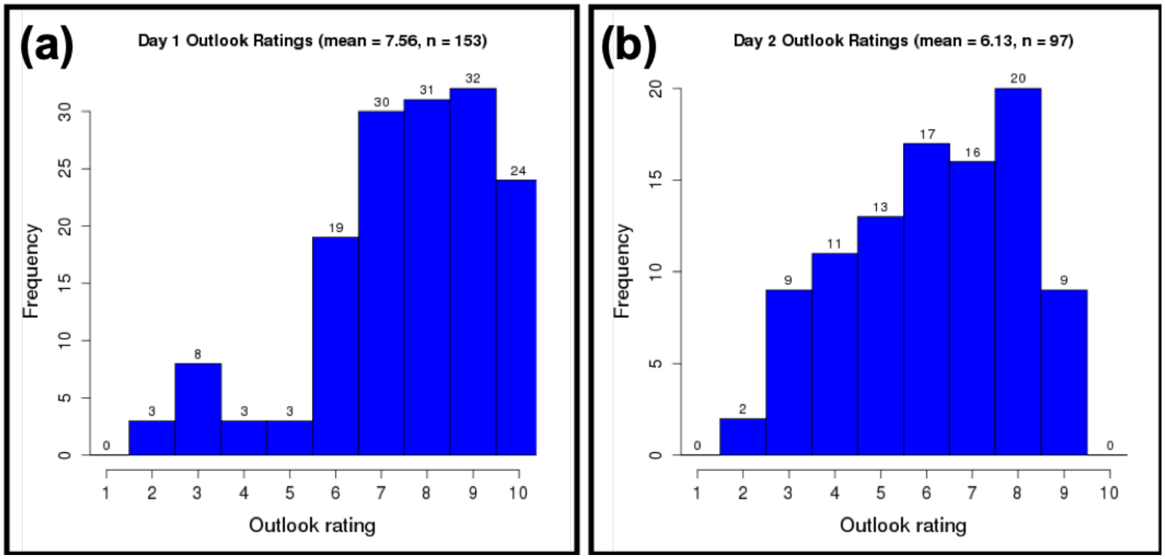


Figure 3 Histograms showing the distributions of full period outlook ratings for (a) Day 1, and (b) Day 2.

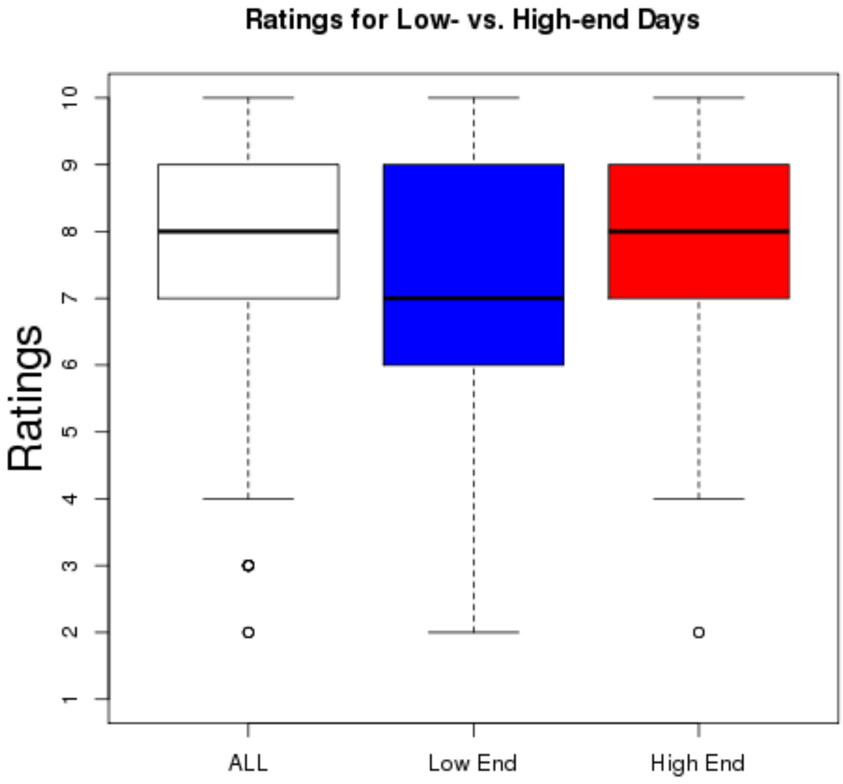


Figure 4 Boxplots for the distributions of subjective ratings on all experiment days, low-end days, and high-end days, where high-end days are defined as when the maximum practically perfect probabilities were equal to or greater than 45%.

Table 4 Maximum probability contour issued for the Day 1 Outlooks and resulting from the practically perfect forecasts. Colors indicate the probabilities associated with each forecast. Days in the grey box with orange font are Fridays, and those outlooks were not subjectively evaluated by participants on the following day.

Date	Day 1 Max Probability (Forecast)	Day 1 Max Probability (PP)
29-Apr	15	5
30-Apr	45	45
1-May	45	45
2-May	30	15
3-May	15	n/a
6-May	30	45
7-May	45	60
8-May	30	45
9-May	30	45
10-May	15	5
13-May	15	5
14-May	5	15
15-May	15	30
16-May	15	45
17-May	45	60
20-May	60	45
21-May	30	45
22-May	30	45
23-May	30	45
24-May	30	30
28-May	45	45
29-May	45	60
30-May	30	30
31-May	30	60

2) POTENTIAL SEVERE TIMING (PST) AREA EVALUATIONS (credit: M. Krocak)

One of the most frequent questions that forecasters are asked is *when* a specific weather event is going to occur. Researchers and forecasters have been working in the SFE for a few years to identify how to best describe and communicate hazard timing information. This work continues that from previous SFEs by testing products that provide timing information on a sub-daily, regional scale.

During the 2019 SFE, participants were asked to create Potential Severe Timing areas (PSTs) that indicate the peak 4-h time period when they thought severe weather would occur within the 15% contour of the Day 1 full period forecast. Participants issued PSTs in small groups, ranging from one to three people, and were each assigned an experimental ensemble to incorporate into their forecast process. CLUE subsets used were the CAPS EnKF, CAPS FV3, HRRRE, map-hybrid, and the operational HREF.

Participants had access to a plethora of forecast fields from each subset available in the drawing tool and they could draw their forecasts directly on top of the fields. Some of the available environmental fields included CAPE, CIN, SRH, and storm attribute fields such as simulated reflectivity and updraft helicity (UH). The purpose of having participants use ensemble subsets was twofold. First, it allowed participants to explore the output of the single ensemble more deeply than if they were tasked with using multiple ensemble subsets. Second, it had the participants creating forecasts in an environment that more closely simulated operations, when likely only one new tool would be introduced at a time. A set of example forecasts is displayed in Figure 5.

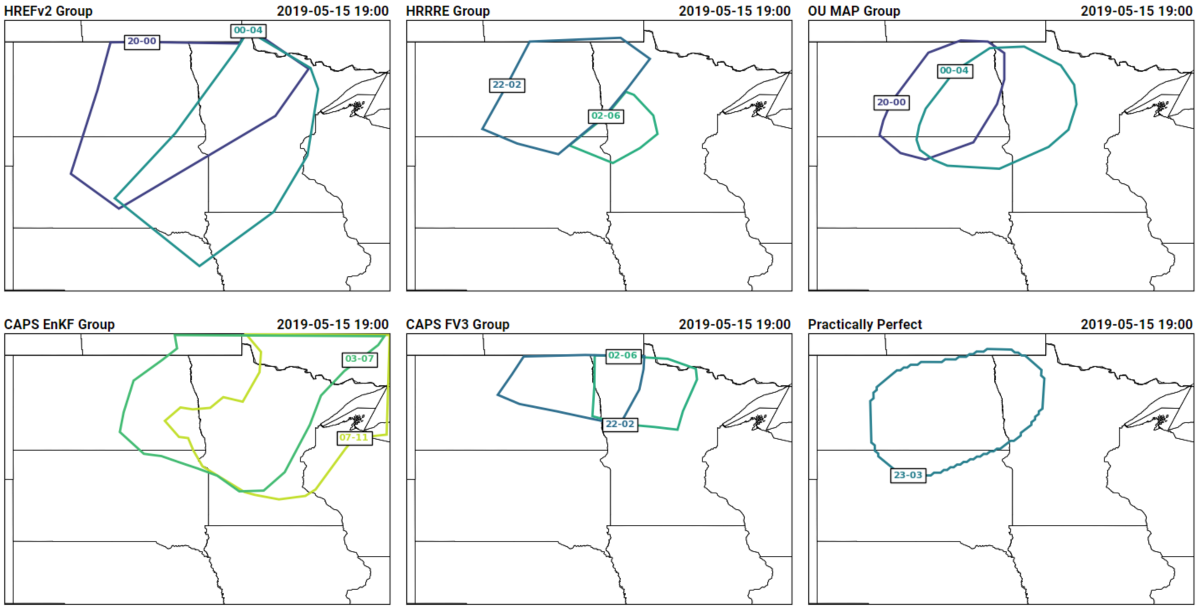


Figure 5 PST areas issued by each group and the PST verification on 15 May 2019.

During the experiment, forecasters were presented with background conceptual reasoning for the product and instructions on how to best create PSTs. Previous work has shown that a majority of convective outlook day events occur within just 4 h of the 24 h day (Krocak and Brooks 2017). Therefore, the idea behind the PST product is that forecasters can, in theory, identify a 4-h period in which the threat indicated by the convective outlook probabilities will be concentrated. After synthesizing all of the guidance information for the day, participants were broken into small groups and given laptops to draw their PSTs. They were all given the same instructions and the following "best practices" – (1) cover the 15% area, (2) don't draw an area for every hour, (3) minimize overlap between areas, and (4) keep it simple.

In addition to the forecaster-created products, a few automated products were also introduced to forecasters in the 2019 SFE. First, an experimental "automatic PST" was tested. This tool uses reports from the previous day to automatically draw PSTs based on where and when the reports occurred. An example of the automatic PST area is shown in the bottom right panel of Figure 5. In addition to the automatic PST, "first guess" PSTs were also created using a similar algorithm to the automatic PSTs, except the input was UH over a certain threshold based on the model climatology. Forecasters were asked to evaluate the first guess and the automatic PSTs as part of their daily subjective evaluation activities.

The overall performance of the lead forecaster PSTs was similar to what was seen in the 2018 SFE (Fig. 6). Note that the median and mean duration that each point on the grid was under a PST forecast is noted in Figure 6 as well. This was calculated by finding the earliest start time and the latest end time that was forecasted for each point. Then the mean and median duration was calculated over all points and all days for each kind of PST (automatic or forecaster-drawn).

Overall, the probability of detection was high for the forecaster PSTs, but the success ratio was relatively low, which indicates that the PSTs are drawn broadly to cover environments that are favorable for future storm development. In other words, the PSTs captured many of the reports on any given day, but suffered from a large false alarm area, which was unavoidable since precise locations for severe weather at Day 1 lead times are not predictable.

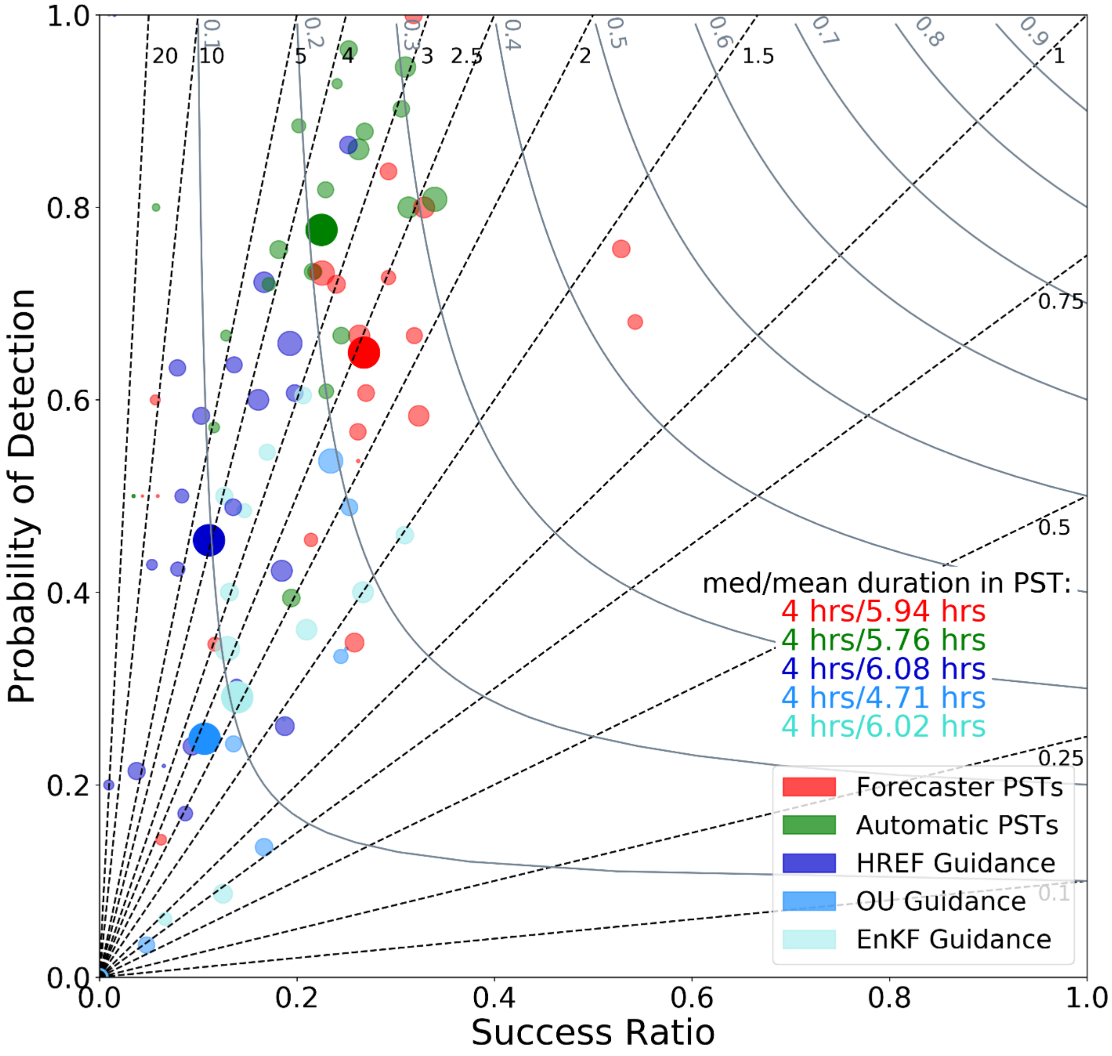


Figure 6 Performance diagram for the PST forecasts from 29 April to 31 May. The size of the dots is proportional to the number of local storm reports received on the forecast day.

Each morning during the experiment, participants were asked to evaluate the performance of the previous day's lead forecaster PSTs. They were also asked to evaluate the performance of the Day 2 PSTs as though they were Day 1 PSTs. Not surprisingly, the lead forecast for Day 1 generally scored higher than the same for Day 2, with the Day 1 median equal to 8/10 and the Day 2 median equal to 6/10 (Fig. 7). Additionally, there was more variation in the Day 2 forecasts (IQR=4) than the Day 1 forecasts (IQR=3). To provide context for the subjective forecast ratings, participants were also asked to rate the difficulty of the previous day's forecast (Fig. 8). Most ratings were within the difficult, neutral, or easy categories, but participants did identify most days as being "difficult" to draw PSTs for. It isn't surprising to see so many responses identifying the PST activity as difficult since participants are not asked to explicitly identify timeframes of severe weather in daily operations. Perhaps with repeated exposure and the development of strategies, forecasting timing would become less difficult.

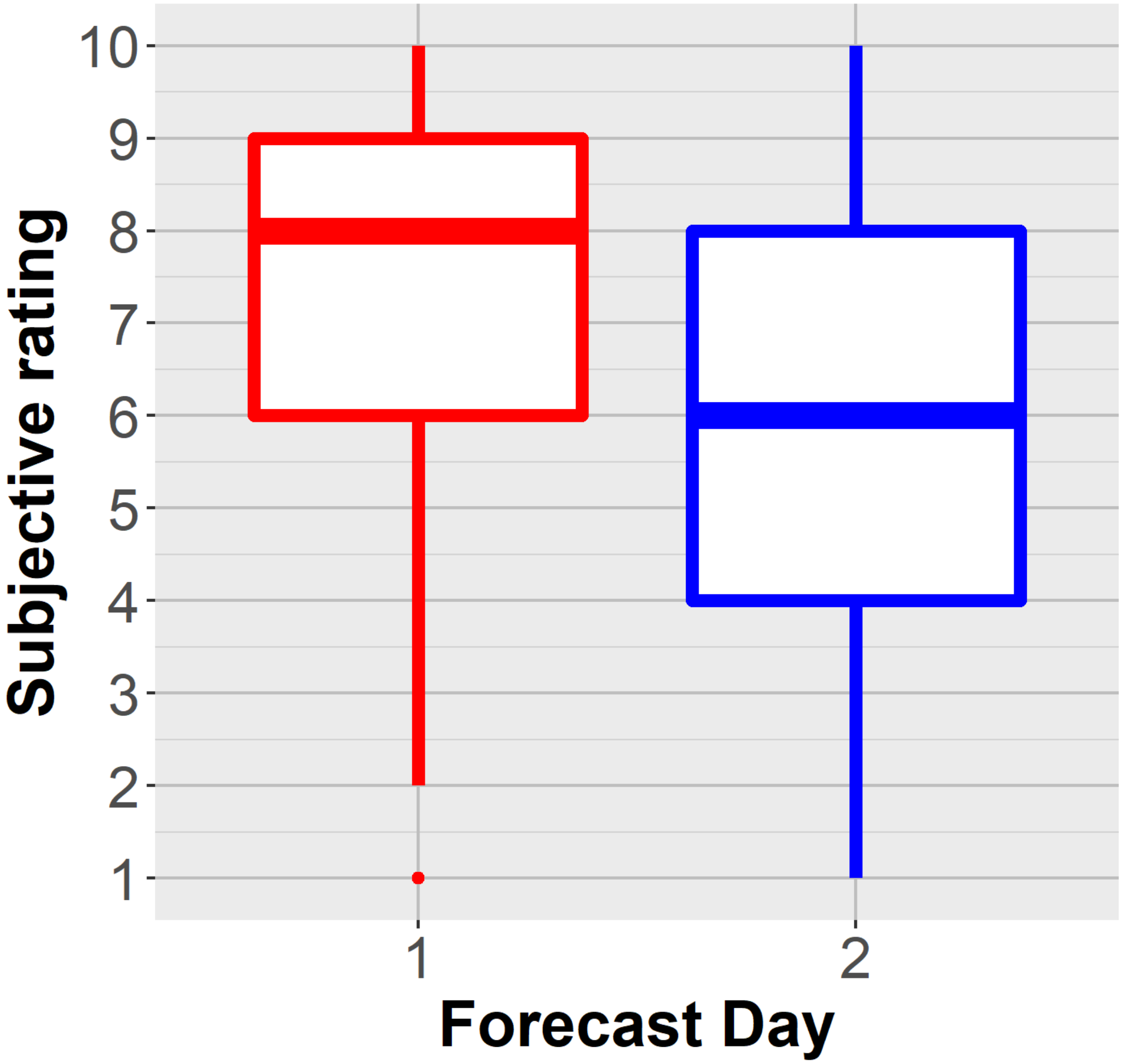


Figure 7 Boxplot of the subjective evaluation ratings for the lead forecaster PSTs on days 1 & 2.

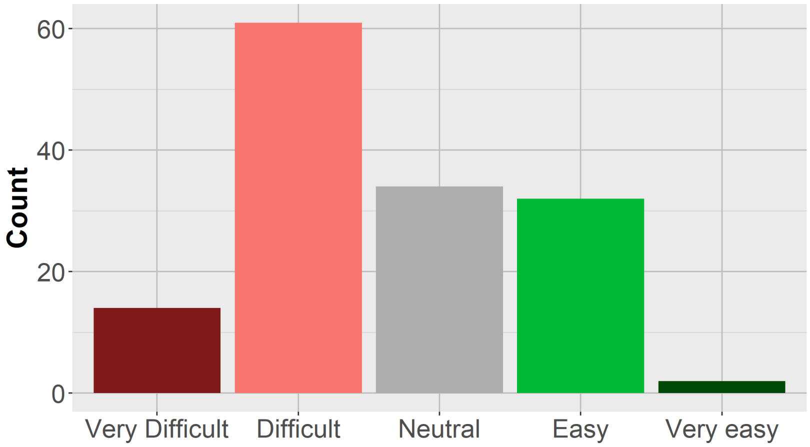


Figure 8 Subjective evaluations of the difficulty of drawing PST areas for the previous day.

Finally, participants were also asked to complete a survey at the end of the week to answer more in-depth questions about the PST forecast process and utility. While on a majority of the days participants thought that the process of drawing the PSTs was difficult, they also thought that the added value outweighed this added workload (Fig. 9). These results indicate that while there is much work to be done to streamline the forecast process, there is reason to pursue the addition of timing info in hazardous weather forecasts. When asked about whom the audience for the PST product should be, most forecasters thought that partners like emergency managers and broadcast meteorologists would benefit from having this timing information. Some participants also thought that the public should see this information as well (Fig. 10). Qualitative feedback indicates that while the forecast process and some visualization strategies should be improved upon, there are many stakeholders that would benefit from having information like what is provided in the PST product.

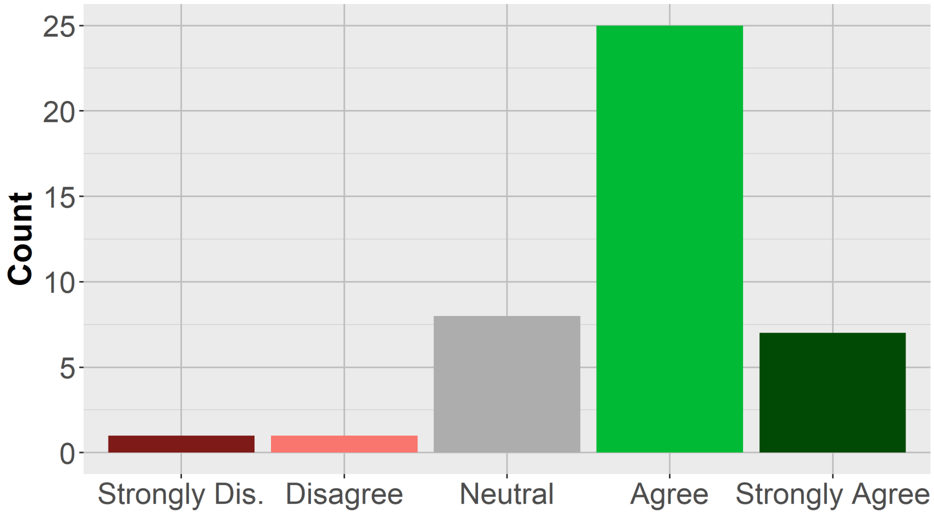


Figure 9 Participant ratings of their agreement with the following statement, "The added value of the PST product is greater than the added workload".

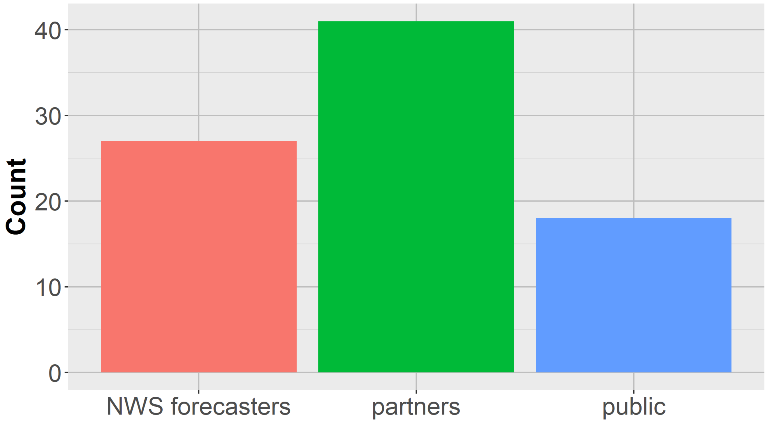


Figure 10 Participant evaluation of who the audience of the PST product should be.

3) WOFS TRAINING EXERCISE (credit: K. Wilson)

Overview

Prior to SFE 2019, the WoFS research group developed a training module to educate users on how to interpret WoFS guidance products. The content of this training was based on findings from the 2017 SFE survey that assessed meteorologists' interpretations of storm-scale ensemble forecast guidance. This training was designed to be shared with collaborating local NWS forecast offices and the Weather Prediction Center. Additionally, it was shared with the 2019 SFE participants to prepare them for the daily 3–4 pm WoFS activity.

To measure the impact of the training module on participants' understanding of WoFS guidance, a series of test questions was also administered to the 2019 SFE participants. Half (N=41) of the participants completed the training module followed by the test questions (Group 1, G1), and the other half (N=43) completed the test questions followed by the training module (Group 2, G2). Demographics were well balanced between G1 and G2; approximately 75% of participants in each group were male, approximately 80% of participants in each group had postgraduate degrees, and 40–50% of participants in both groups had prior experience using WoFS guidance.

Group Comparison

The training module consisted of content that described concepts such as neighborhood probabilities, probability exceedance thresholds, percentiles, 90th percentile versus maximum values, and paintball plots. The test questions focused on these same topics. The test activity took G1 participants on average less time to complete (mean=24.3 min, SD=13.5 min) than G2 participants (mean=28.1 min, SD=7.6 min). Over a total of 19 questions, participants could accumulate a total of 29 points. Single response answers were required for 12 questions, while seven questions required the selection of multiple responses. Each response scored one point. Summary score results of G1 and G2 suggest almost no difference between the two groups' test performance: G1 achieved a median score of 25 (SD=2.5) and G2 achieved a median score of 24 (SD=2.5). The nonparametric Mann Whitney U test indicates that this difference is not significant at the *p*<0.05 level.

To inspect differences between groups on a question-by-question basis, the distribution of participants obtaining a fully correct answer for each question was plotted (Fig. 11). In total, G1 obtained a higher percentage correct value than G2 for 12 questions. G2 obtained a higher percentage correct value than G1 in the remaining seven questions. Assessing the percentage difference between the two groups for each question, there are three occasions when the difference exceeded 10% (i.e., accounting for 5–8 participants). For each of these three occasions (relating to Q3, Q12, and Q13), G1 scores exceeded G2 scores (Fig. 11). Therefore, those that received training first performed notably better in questions related to the neighborhood probability concept and the 90th versus ensemble maximum concept.

Specifically, Q3 asked participants "In what scenario would you choose to utilize a larger neighborhood for your forecast?" The correct response was "When you are forecasting farther out in time, because the uncertainty is higher and ensemble spread is larger." While all G1 participants answered this question correctly, six G2 participants did not. Both Q12 and Q13 required multiple response answers. This question type was overall more challenging for both groups, as noted by the lower percentage correct values ranging between 12–90% (Fig. 11). Q12 asked participants why the 90th percentile and ensemble maximum representations of UH had different values, while Q13 required participants to retrieve, interpret, and compare wind values from the 90th percentile and ensemble maximum products. Despite these questions being more challenging for both groups, G1 participants—who had completed the training module first—had a higher success rate of selecting all correct responses, whereas G2 participants had a greater tendency to select just one correct response.

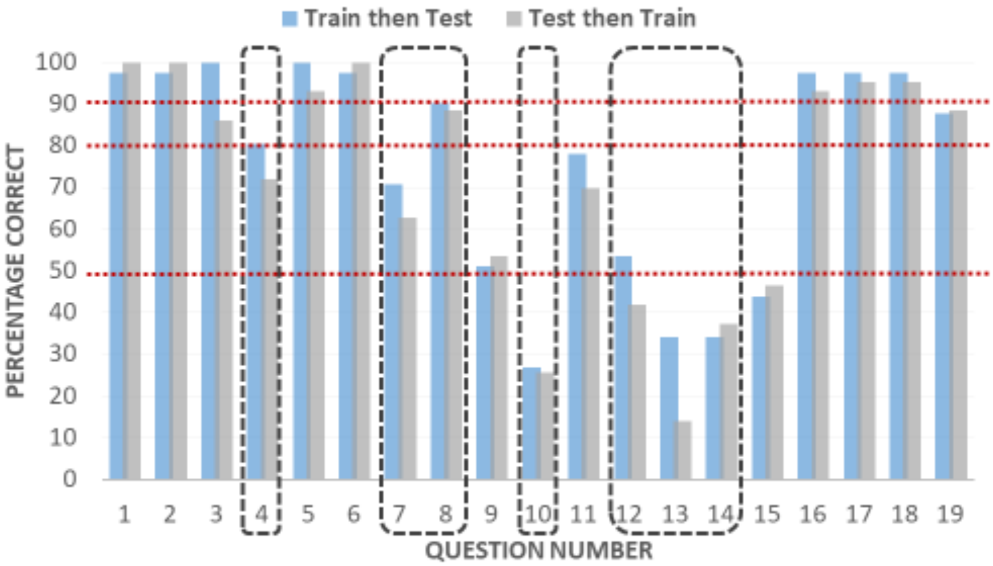


Figure 11 Percentage of participants obtaining a fully correct answer in G1 (Train then Test) and G2 (Test then Train). Red dotted lines indicate percentage correct levels at 50%, 80%, and 90%. Vertical black dotted lines indicate questions requiring multiple responses.

While questions with lower percentage correct values are typically associated with multiple response style questions, two questions (Q9 and Q15) requiring single response also had low percentage correct values. First, in Q9, participants were asked to view a UH probability of exceedance threshold product and report the expected value at a given location. In this instance, the low percentage correct value can be attributed to both the misleading nature of the question wording (such that participants may have been influenced to select an absolute value), and the example used a color scale that straddled two different answers (a mistake on the test designer's part). However, in Q15, the low percentage correct value can be attributed to misunderstanding. Participants were asked to view a paintball plot and select the correct probability range for UH reaching a given threshold. Whereas the question intended participants to count the number of different colors (each indicating a different ensemble member) in the paintball plot and relate it to the total number of ensemble members in the forecast, it seems that participants instead counted the number of paintball contours. This mistake was not possible to make in subsequent paintball plot questions, hence the more successful responses provided in Q15–Q17. This finding suggests that the training did not address aspects of paintball plots successfully: namely what contours represent, how they relate to individual ensemble members, and how together they can represent an overall probability.

Concept Comparison

The test performance of G1 and G2 can also be considered with respect to the different types of concepts being tested in each question (Fig. 12). As discussed above, participants in both groups performed well answering the paintball plot questions except for Q15. Additionally, both groups did well answering neighborhood probability concept questions, with an exception being Q10. In this question, participants were presented with a nine-panel plot that presented three probability of accumulated rainfall exceedance thresholds at three different neighborhood values (associated with the October 2018 Hurricane Michael event). In this multiple response question, the low percentage correct value can be attributed to most participants selecting only two of the three correct answers, that "The ensemble suggests a greater than 80% likelihood of rainfall accumulations >1 inch", and "The neighborhoods increase from left to right." Participants were unlikely to select the third correct response that "During extremely widespread events, the use of neighborhoods are more impactful in locations that experience larger impacts."

Performance in both groups was lower for the questions testing understanding of the 90th percentile and ensemble maximum values (Q12 and Q13, also discussed above). Despite this low performance, G1 performed comparably better than G2, suggesting that the training material did have a positive impact on some participants' understanding of this concept. Performance was similar but also low for both groups in Q14. This question tested the percentile concept, and low performance was tied to participants not selecting both correct answers and/or selecting an incorrect answer. When combined with probability information in Q19, participants were generally more successful at applying percentile information.

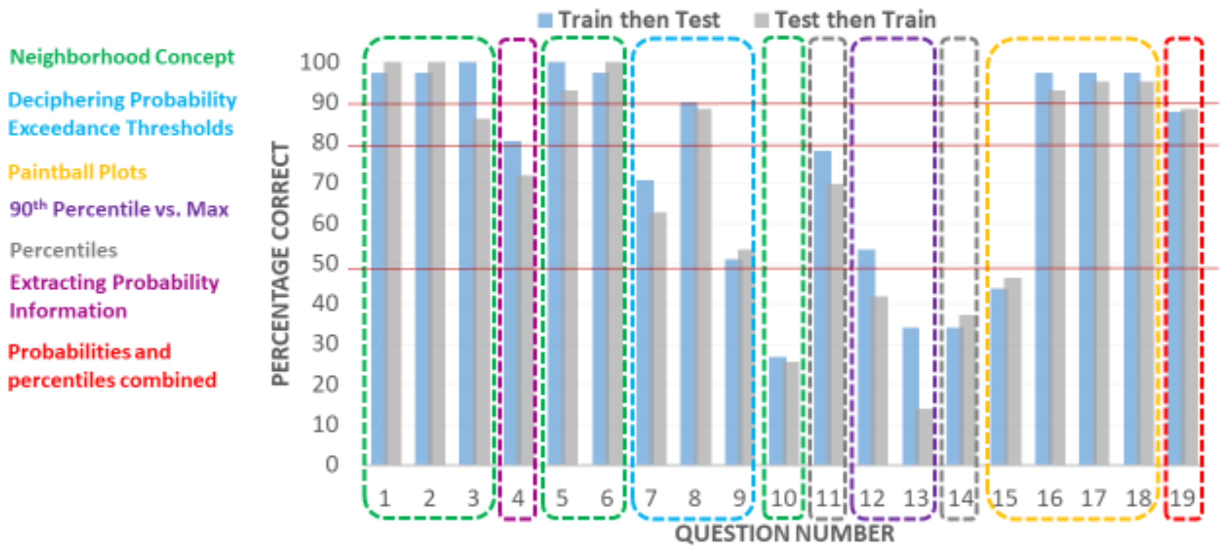


Figure 12 Same as in Fig. 11, but vertical colored dotted lines indicate the type of concept each question tested.

Prior WoFS Experience

To investigate whether prior experience with WoF had an impact on the training and test results, G1 and G2 were each broken down to consist of two subgroups: G1a (N=21) and G2a (N=19) encompassed participants that had prior WoF experience, and G1b (N=20) and G2b (N=24) encompassed participants with no prior WoF experience. The group assignment was based on participants' response to a survey question that enquired about prior exposure to WoF. Experience describes participants who participated in the past 2017 and/or 2018 SFE (and therefore took part in previous WoF activities) or participants who had previously viewed and used the WoF website. No experience describes participants who reported either only being aware of the WoF program or having no experience with WoF at all.

Participants' responses were compared within G1 and within G2. Overall, prior experience with WoF was found to impact test results, but this impact was notably larger within G2 i.e., for those who took the test *before* completing the training module. In both G1 and G2, 12 questions were answered correctly more often by subgroups that had experience with WoF than by subgroups without prior experience (Table 5). However, this difference was amplified in G2, such that there were twice as many occasions when the difference in proportion of correct participant responses exceeded 10% (and in some instances was greater than 30%). This result suggests that experience had a greater effect when training was not available prior to completing the test. Likewise, this result also suggests that the training module was effective at bridging the knowledge gap between participants with and without prior WoF experience.

Table 5 The number of questions in which a subgroup with prior WoF experience (G1a, G2a) or without prior WoF experience (G1b, G2b) had a higher proportion of participants answering a question correctly (by a difference of 10% “>”, or more than 10% “>>”).

	Train first	**Test first**
No exp> Exp	1	3
No exp >> Exp	3	1
Exp>No Exp	8	4
Exp >> No Exp	4	8
Equal	3	3

Although many participants did not answer Q15 correctly (the paintball plot question, as discussed above), the difference in proportion of correct responses was greatest between subgroups, such that subgroups with prior experience had a 20–30% higher percentage correct value than subgroups without experience (Fig. 13). Therefore, in this instance, while training did not necessarily improve understanding, prior experience had the most notable impact out of any of the questions for both groups. Those with prior experience also outscored those without prior experience in both groups in two other paintball plot questions (Q16 and Q18; Fig. 13).

In contrast, the non-experienced subgroups outperformed experienced subgroups in Q10, more notably so for G1a and G1b (by 25%) than G2a and G2b (by 8%) (Fig. 3). Combining concepts of probability of exceedance and neighborhood probabilities, these results suggest that neither training nor experience aided participants in answering this question correctly. The multiple response style of this question, with 3 of 4 answers requiring selection, was seemingly challenging to most participants.

For those who completed the training module first (G1), there are two more instances (Q13 and Q19) where the proportion of participants without experience (G1a) answered correctly more often by a difference of 15–20% than those with experience (G1b). These questions may highlight concept areas where training was particularly effective. In the several instances where the non-experienced subgroup (G2b) outperformed the experienced subgroup (G2a), the difference in percentage correct values were mostly <10% (Table 5), and therefore reflected a difference of only 1 or 2 participants.

Finally, differences were not found in G1’s subgroups and G2’s subgroups percentage correct values in a few other instances (Table 5). Both groups achieved 100% percentage correct values for three (not always the same) questions (Fig. 12). These instances can be explained by either participants’ prior education providing them with the knowledge to answer the questions correctly, or training and/or experience helping to close the knowledge gap between subgroups.

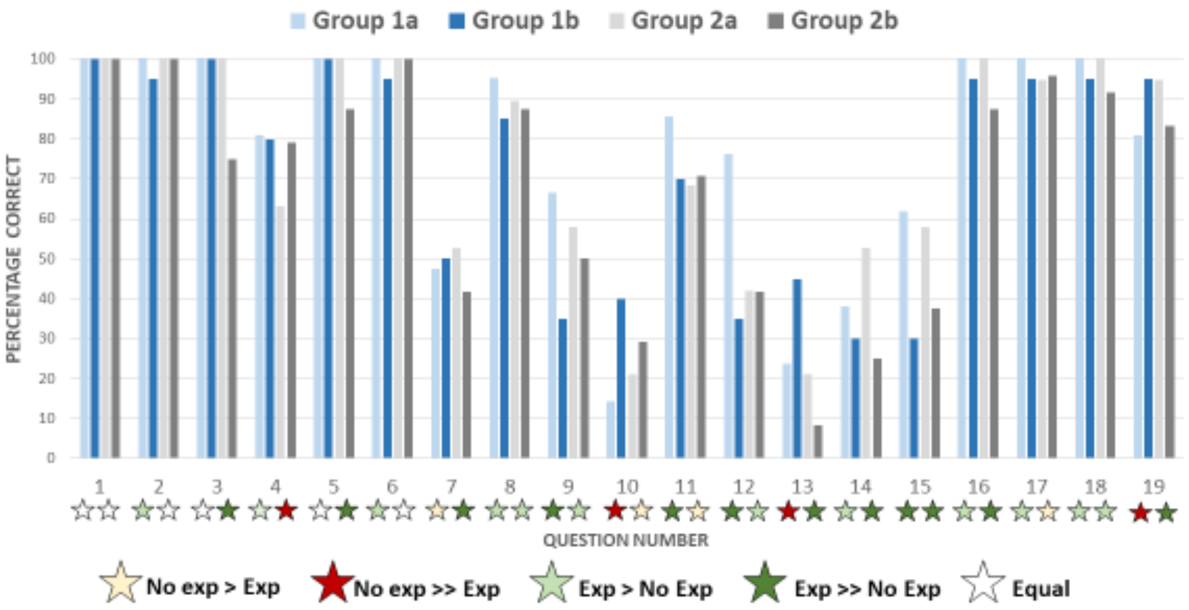


Figure 13 Similar to Fig. 11, but G1 and G2 are broken down into two subgroups based on prior experience with WoF. Each question is assigned two stars. The left and right stars reflect the difference between subgroups in G1 and subgroups in G2, respectively. The color of the star reflects the magnitude of the difference (same as in Table 5).

4) WOFS OUTLOOK ACTIVITY (credit: J. Choate)

Overview

During the 2019 SFE, a WoFS activity ran from 3 pm until 8 pm Monday through Thursday. On Mondays, the 3-4 pm time period was occupied by the training material referenced in the previous section. During the rest of the week, the 3-4 pm period was an outlook activity led by two facilitators where two groups of participants including two NWS forecasters produced 3 outlooks for each group using WoFS guidance. The outlooks were as follows: a 1-hour outlook valid from 2100-2200 UTC, a 1-hour targeted outlook valid from 0100-0200 UTC, and a 4-hour outlook valid from 2200-0200 UTC.

After the group activity, the two NWS forecasters continued issuing forecasts until 8 pm. The schedule they followed is listed in Table 3. They continued the same pattern as the group activity with one outlook always being approximately an hour ahead of the initialization time, the targeted outlook (0100-0200 UTC) staying the same during each period, and the 4-hour outlook always being valid for the last 4 hours of the WoFS forecast (or hours 2-6 of the WoFS forecast). Forecasters also filled out evaluations at the end of each set of outlooks. The following discussion will only focus on the outlooks produced during the 4-8 pm period.

All Outlooks

There were 19 days during the 2019 SFE during which NWS forecasters completed this activity. Ideally, each forecaster would have submitted 11 outlooks during this period each day. Given that there were two forecasters participating during each of these days, a complete set of outlooks would contain 418 submissions. In reality, 403 final outlooks were submitted, meaning 15 outlooks were never submitted. This can be attributed to human error with the outlook submission website, unclear direction or understanding with the submission process or timeline, or other factors.

Forecasters were able to draw 15%, 30%, 45%, and 60% contours anywhere within the WoFS domain for that day. Hatched areas for significant weather could also be drawn but future work is planned to analyze those outlooks. Multiple outlook contours (contours, hereafter) could be drawn over the domain. An example of multiple contours of the same value (three 30% contours) from Forecaster 1 on 30 April 2019 is shown in Figure 14.

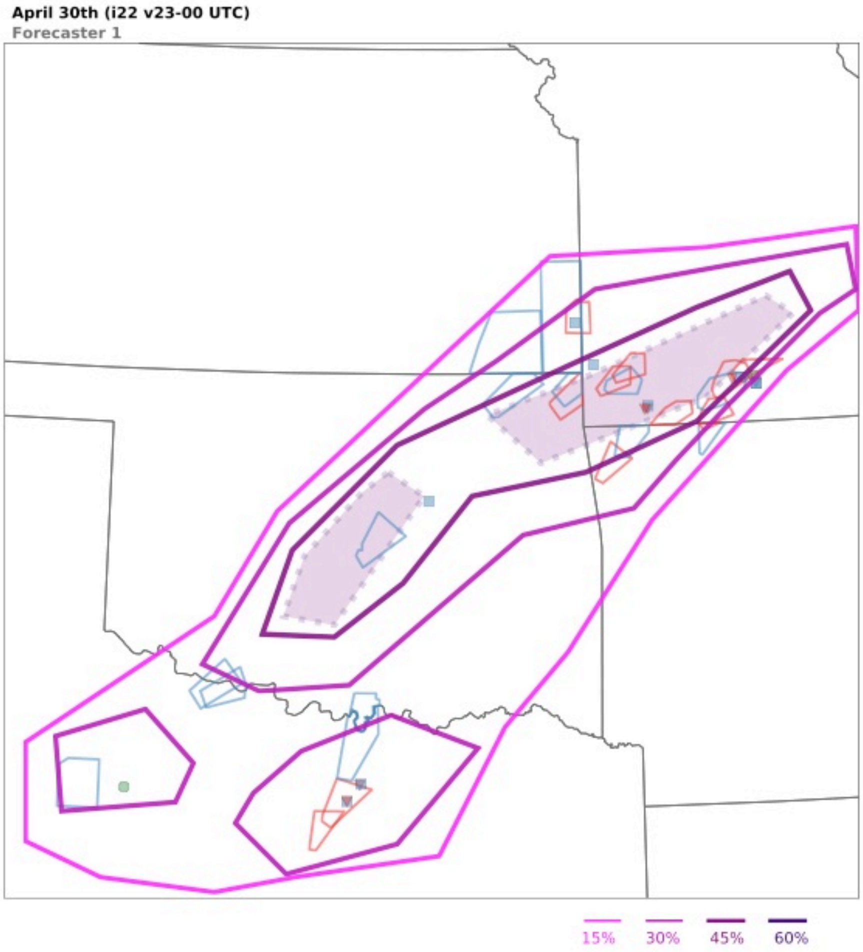


Figure 14 Forecaster 1 outlook submission during the 30 April 2019 activity. Outlook was made using the WoFS 2200 UTC initialization and is valid for the 2200-0000 UTC period. This outlook shows an example of a forecasters drawing three 30% contours.

Of the 403 final outlooks submitted, a total of 1261 contours were drawn. The number of 15% and 30% contours are similar (484 and 416, respectively) which implies that the majority of the time when a 15% contour was drawn a 30% contour was also drawn within it. The number of 45% contours drops significantly down to 250 and only 66 60% contours drawn within these final outlooks. A graph of the distribution is shown in Figure 15.

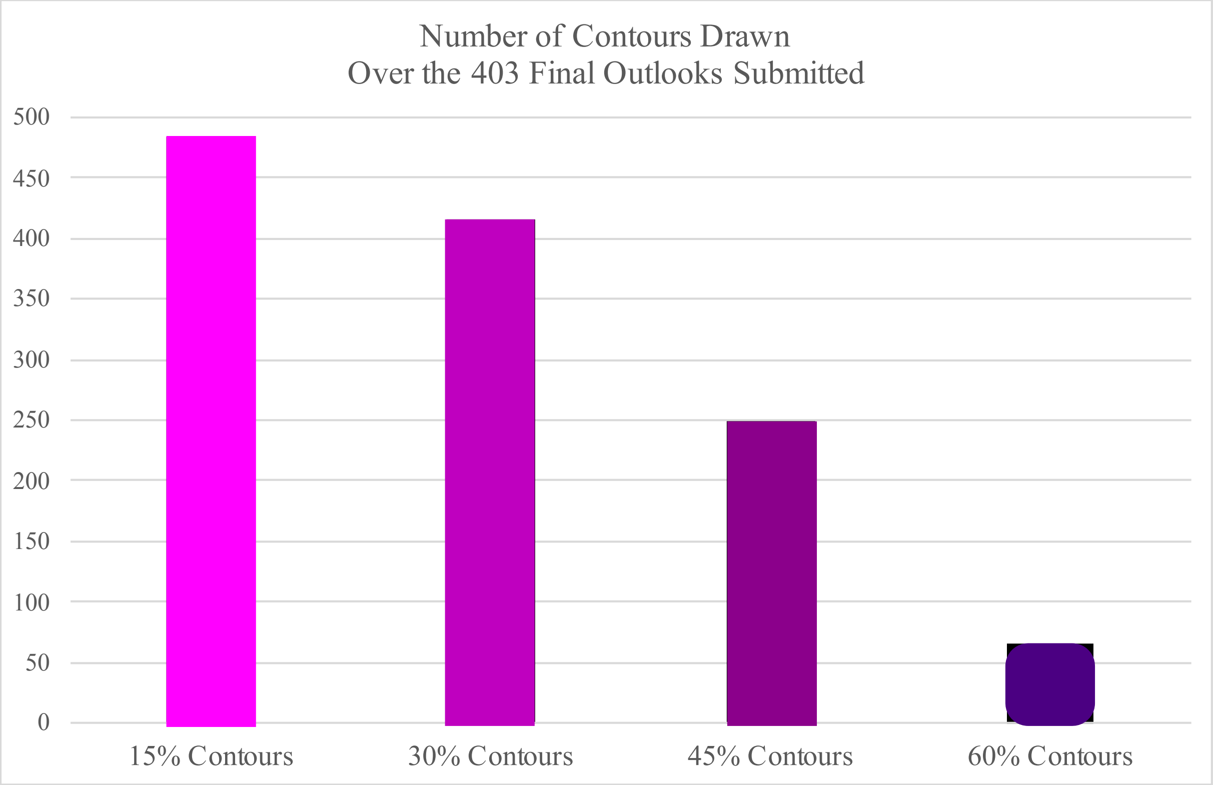


Figure 15 Comparison of the number of contours drawn for each value over the 403 final outlooks submitted.

The area encompassed by each outlook was also calculated. Figure 16 shows the average area enclosed by each contour value for the total outlook in squared kilometers. This means, if there were two 15% contours within an outlook, each of their areas were added together for a total 15% area enclosed and that value was used for the day. Each day's total was then used to calculate an average over the 19 days (if applicable). A day's contour area was only included in the average if there was a contour of that value drawn during that day's activity. For example, there were no 60% contours drawn on 30 April, instead of including a "0" in the average calculation, it wasn't included. The difference between areas is also noted between the bars.

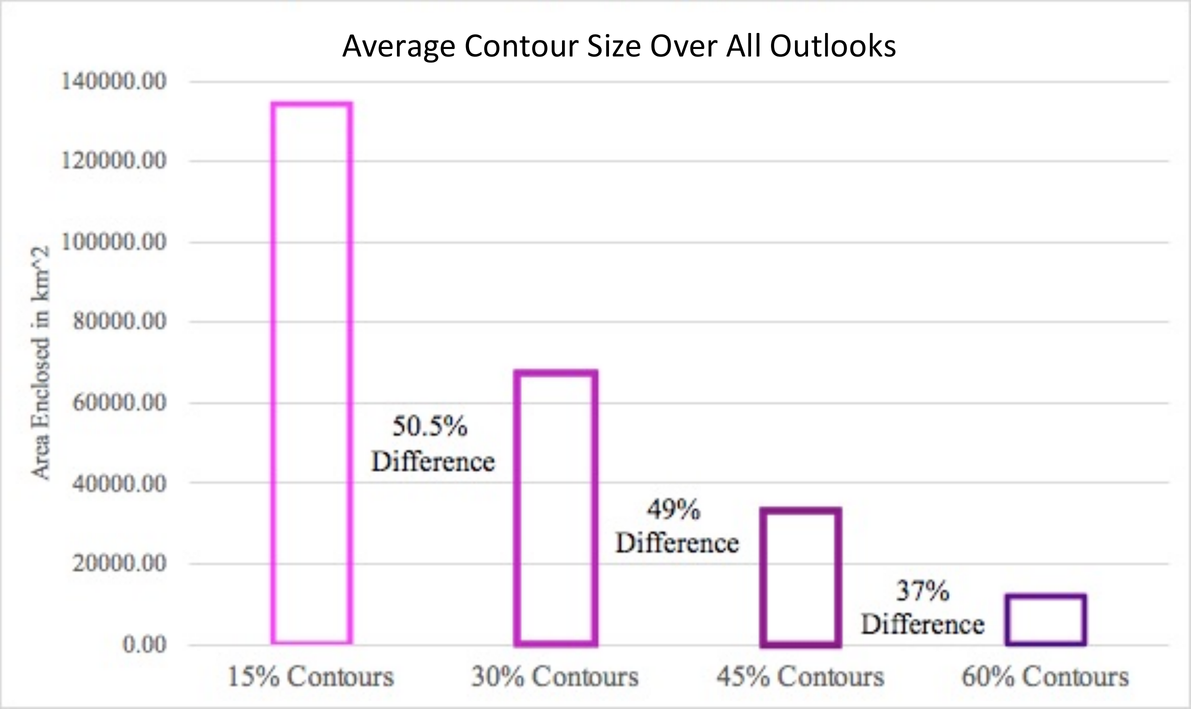


Figure 16 Average area enclosed by each contour value over all activity days.

Targeted Outlooks

During each hour of the WoFS activity, targeted outlooks were drawn (e.g., Fig. 14). These outlooks were always drawn for the 0100-0200 UTC period regardless of the forecast initialization time being used. Ideally, over the 19 days, there would be 152 targeted outlooks submitted (4 per forecaster per day). In actuality, there were 144 targeted outlooks submitted, meaning eight 0100-0200 UTC outlooks were never submitted for the same reasons mentioned above. Within the 144 outlooks submitted, 434 contours were drawn with a similar distribution to the overall count (Fig. 15).

The areas encompassed by the targeted outlooks were calculated the same way as the total outlook. The results of those areas can be seen in Figure 17. The areas for each contour are very similar between the overall and the targeted. The targeted 15% and 30% average areas are slightly smaller than the overall average areas and the 30% and 45% targeted average areas are slightly larger than the overall average areas. The difference between each contour value’s size is also similar between the target and overall outlooks (Fig. 16).

The evolution of outlooks for the targeted periods was of interest to researchers. Figure 18 looks closer at one forecaster’s outlook submissions for the 0100-0200 UTC period on 22 May 2019. This image shows the evolution of Forecaster 1’s 0100-0200 UTC outlook as he progressed through the new WoFS initializations.

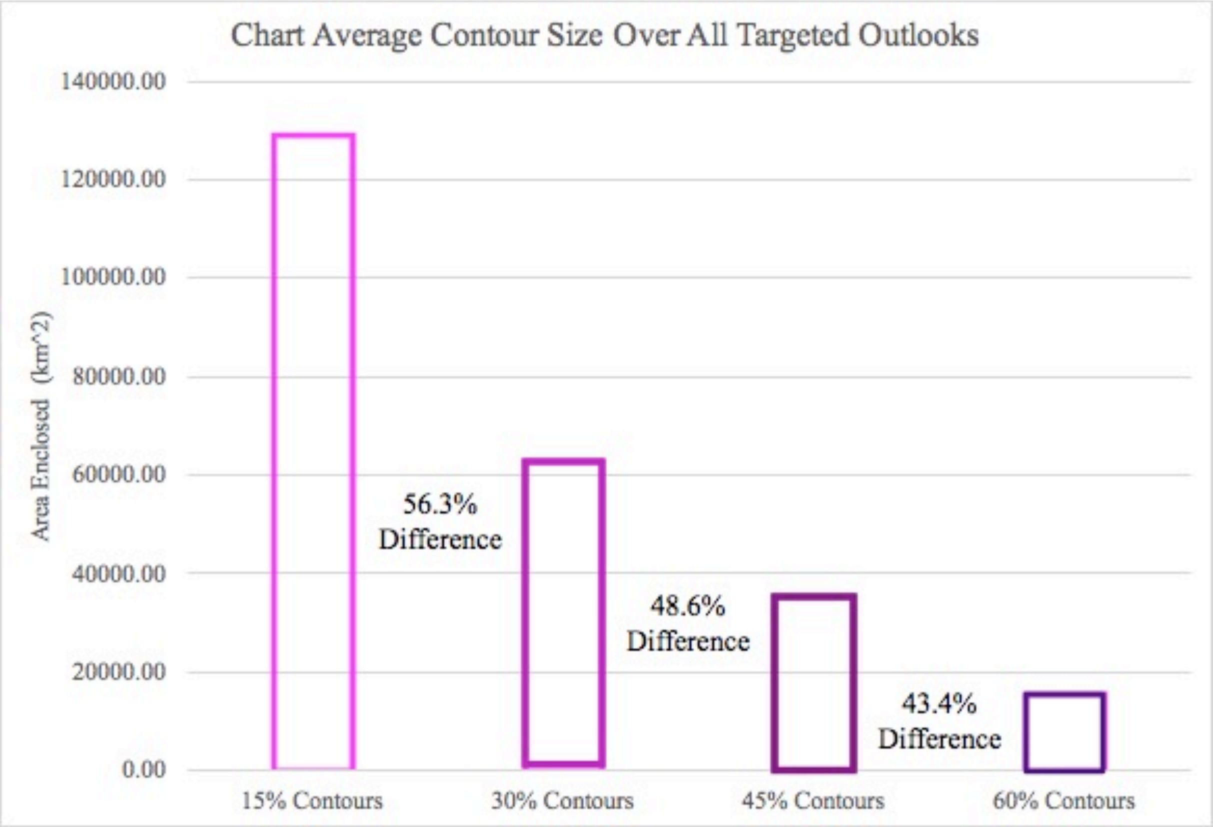


Figure 17 Average area enclosed by each contour value in the targeted outlooks over all activity days.

The area enclosed by each value of contour is in the upper left of each panel. The bold numbers indicate that that outlook has the largest area for that contour value. The outlooks initialized at 2100 UTC have the largest areas for all 3 values of contours drawn. The smallest contours are in bold italic. The outlook initialized with the 2200 UTC forecast has the smallest areas for the 15% and 30% contours but the smallest 45% contour is within the outlook initialized with the 0000 UTC forecast.

One would assume that the later WoFS initializations, which are closer to the valid time, would provide forecasts that would guide forecasters to produce more specific, or smaller, area outlooks. That was the case for many days during the 2019 SFE but not for this example when considering the 15% and 30% contours. What can be seen is that both these contours become narrower with more recent initializations, specifying a corridor of severe weather. Although the overall area increases with later initializations, the forecaster was able to identify the area of severe weather on the Oklahoma-Texas border that was not captured in the first two outlooks submitted (2100 and 2200 UTC initializations). The largest decrease in area is in the 0000 UTC 45% contour. Using the most recent forecast, the forecaster felt confident in locations where he thought the most severe weather would occur and broke up his 45% contour into three smaller 45% areas which included the area of multiple tornado reports near the Missouri, Kansas, Oklahoma boarder. Continuing analysis by WoFS researchers is ongoing into the evolution of the targeted outlooks throughout the SFE.

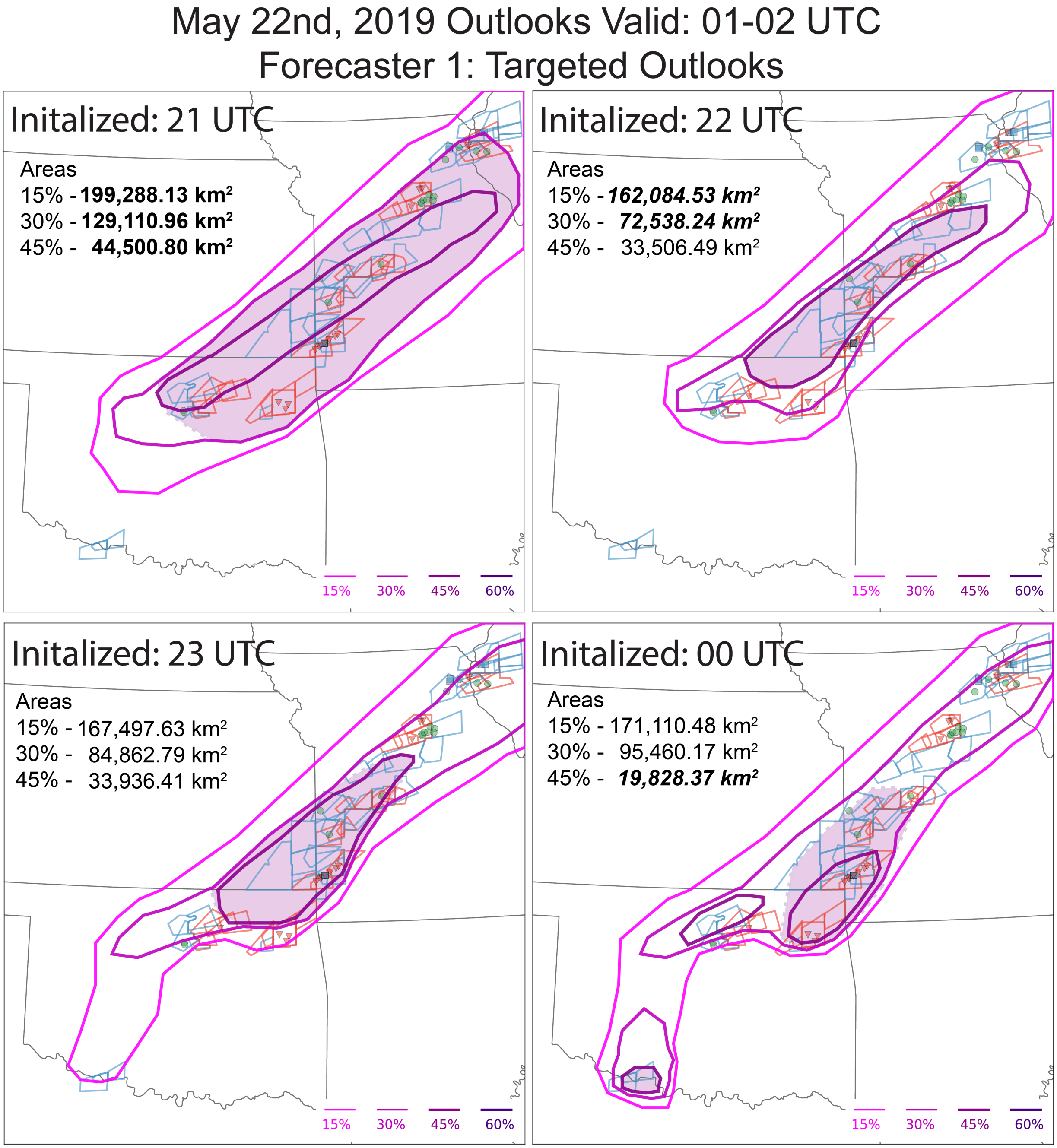


Figure 18 Evolution of targeted outlooks from Forecaster 1 on 22 May 2019. Areas in bold (italic) are the largest (smallest) area for that contour value. Shaded regions indicate probabilities of significant severe weather of 10% or greater (i.e., EF2+ tornadoes, 2-in+ hail, and/or wind gusts of 65 knots or greater). Blue and orange polygons show severe thunderstorm and tornadoes warnings, respectively, and red triangles, green circles, and blue boxes denoted locations of observed tornado, hail, and wind reports.

b) Evaluation of experimental forecast products – Severe Hazards Desk (credit: B. Gallo)

1) EXPERT FORECASTS

New to the 2019 SFE, forecasts were issued on the severe hazards desk for both coverage and conditional intensity. These forecasts were issued by the expert lead forecaster, with input from the group at the severe hazards desk. Conditional intensity reflects the expected intensity distribution of the storms should they form, rather than the current operational probabilistic definition, which is coverage-based. Through a conditional-intensity approach, nuance of different forecast scenarios can be better captured. For example, a wind event that is anticipated to generate many reports but few significant severe reports could have high coverage probabilities but a lower conditional intensity forecast. Conversely, an event where only one or two high-end storms are anticipated to occur could have low coverage probabilities but a higher conditional intensity forecast, suggesting that any storms that do form would cause significant severe weather. These coverage and conditional intensity forecasts were issued as separate layers for each convective hazard: tornadoes, hail, and wind. Three forecasts covering most of the convective day were issued: the Day 2 forecast, valid from 1200 UTC to 1200 UTC the following day; the Day 1 initial forecast, valid from 1600 UTC to 1200 UTC the following day; and the Day 1 updated forecast, valid from 2100 UTC to 1200 UTC the following day.

Each day during the 2019 SFE, participants rated the forecasts issued by the expert forecaster on a scale of 1 (Very Poor) to 10 (Very Good). Day 2 forecasts were only available for three of the five days per week, leading to a slightly smaller sample size for those than for both Day 1 forecasts. However, the trend in the ratings for both conditional intensity and coverage is the same: ratings were higher with shorter lead times (Fig. 19). Coverage and conditional intensity ratings had similar distributions, although for Day 2 the conditional intensity forecast ratings tended to be higher for a given hazard than the coverage forecast ratings. The Day 1 comparison of the coverage and conditional intensity forecasts for each hazard show more mixed results depending on the hazard and aspect of the distribution (mean, median, 75th percentile, etc.). The tornado ratings had higher median ratings than the other two hazards for the Day 1 updates, achieving a score of 8/10 for the coverage and a 7/10 for the conditional intensity. Median ratings were the same for all hazards for the initial Day 1 forecasts, with the hail having the lowest mean ratings for both coverage and conditional intensity.

Participants were also asked to comment daily as to whether or not the forecasts for each hazard improved with decreasing lead time. For the most part, participants indicated that the forecasts improved as lead time decreased, although there were cases noted while either the forecasts became worse (i.e., 21 May 2019) or where the Day 2 to initial Day 1 forecast decreased skill, but then the updated Day 1 forecast again increased skill (i.e., 30 April 2019). Participants also noted that occasionally a refining of the areas decreased the POD and the FAR simultaneously, and thus they had to evaluate the tradeoff.

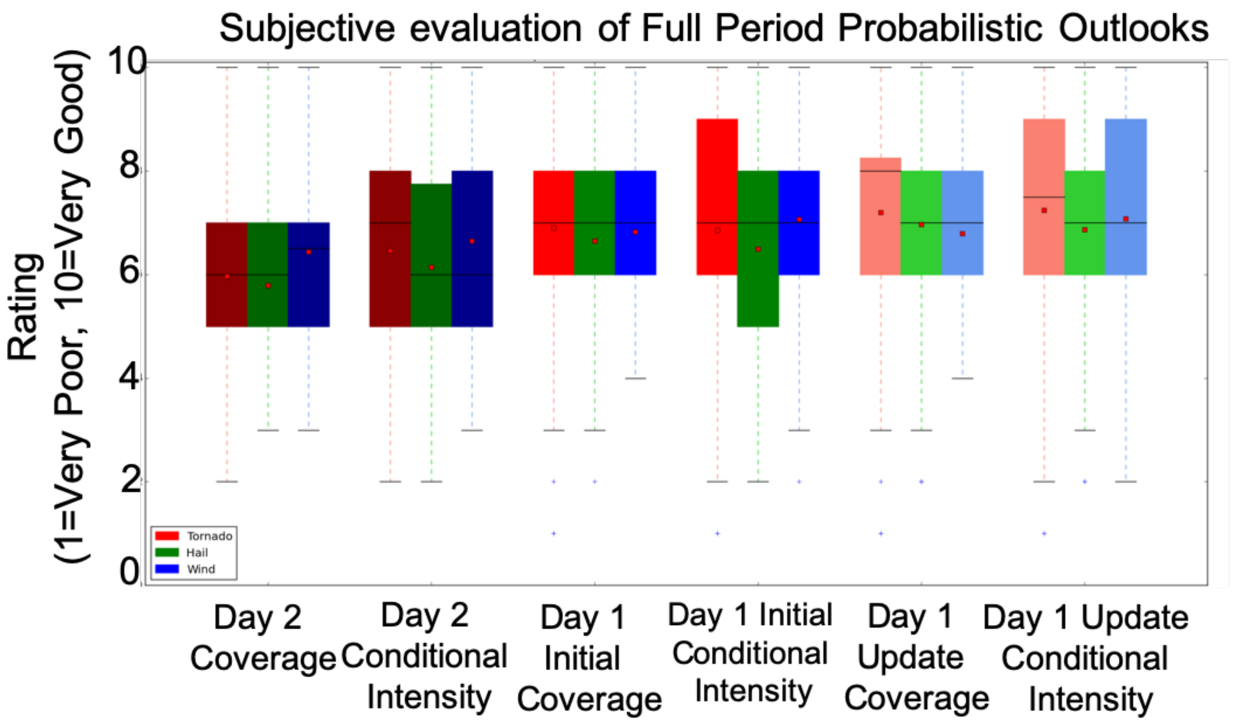


Figure 19 Participant subjective ratings of full period forecasts issued by the expert lead forecaster on the Severe Hazards desk. Darker colors indicate longer lead times, and lighter colors indicate shorter lead times. The different colors indicate the different hazard types being forecast.

2) GROUP FORECASTS

Once participants worked with the expert lead forecaster to create the full period forecast, they were given specific subsets of the CLUE to interrogate more deeply. The groups then generated their own full period coverage and conditional intensity forecasts using data from their specific CLUE subsets. In the afternoon, they updated their group forecasts using observations, deterministic CAM guidance such as information from the latest runs of the HRRRv3, and ensemble guidance provided from the WoFS. This activity allowed participants to examine their specific CLUE subset beyond the cursory glance at top-level fields necessitated by time constraints as a large group. It also allowed participants to try drawing conditional intensity forecasts, to begin developing best practices for drawing this type of forecast guidance. Participants were asked what subset they used to draw their forecasts, to rate each of their forecasts as they did for the expert forecasts, and then to select which group overall had the best forecasts for a given day. They were then asked about how WoFS influenced their forecast updates. Forecast updates were issued on Tuesday through Friday, as the WoFS training activity took place each Monday from 3–4 PM.

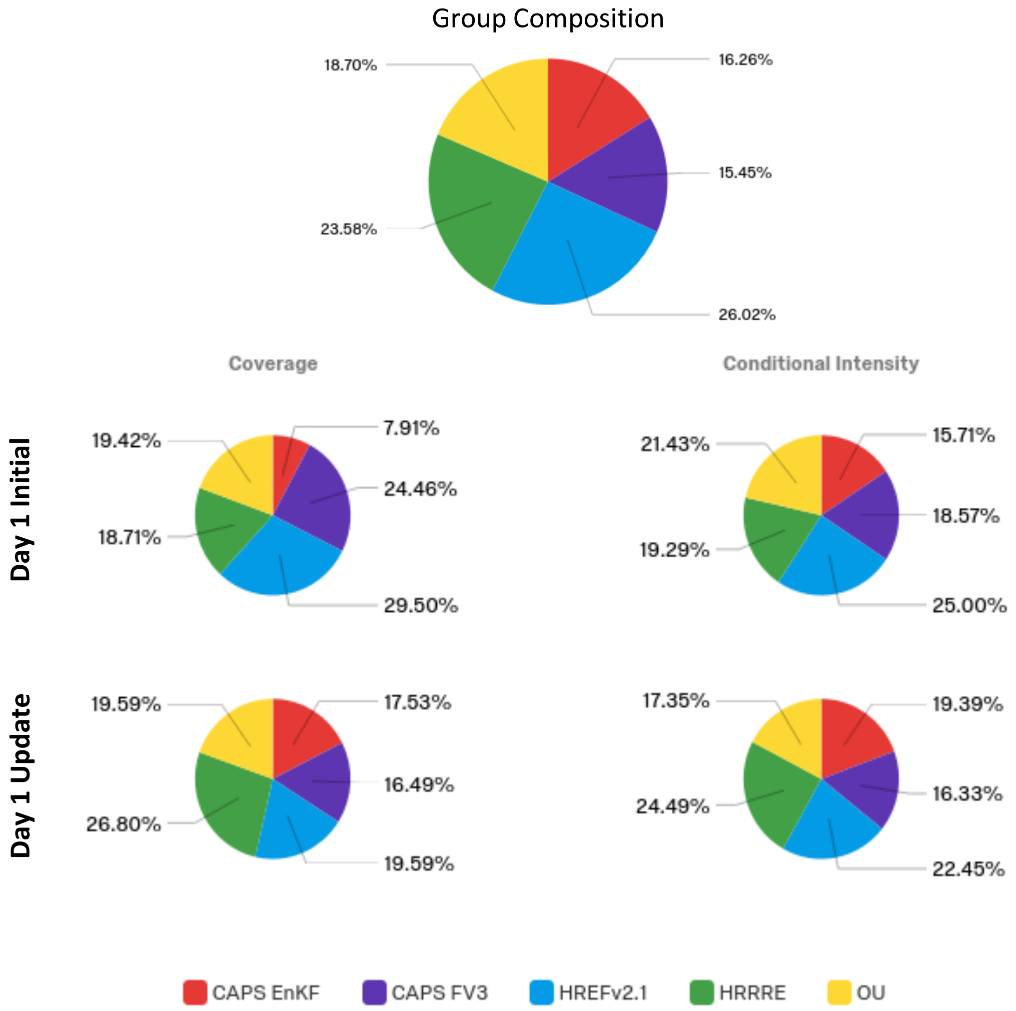


Figure 20 Pie charts showing (top) the division of participants amongst the available ensemble subsets and (bottom) which group had the best coverage (left) and intensity (right) forecasts for the initial and updated forecast products. Participants were asked to consider all hazards when generating these ratings, with higher weight given to the "hazard of the day", or most impactful hazard.

The division of participants into groups was roughly equivalent (Fig. 20), with participants most often using the HREFv2.1 or the HRRRE. Group composition was dependent on data availability, and so the group composition does vary slightly day-to-day during the experiment. For the initial forecasts, the HREFv2.1 group and the CAPS FV3 group were most often selected to have the best coverage forecast, and the HREFv2.1 group and the HRRRE group were most often selected to have the best conditional intensity forecast. The CAPS EnKF group's forecast was selected as the best least often for both the coverage and the conditional intensity forecasts. The group with the best forecast was divided more evenly for the afternoon updates, perhaps reflecting the added influence of observations and the WoFS guidance. Having all of the groups use the same information to update forecasts likely lessened differences between groups for the afternoon updates. For these updates, the HRRRE group tended to be selected as having the best forecast most often for coverage and conditional intensity alike.

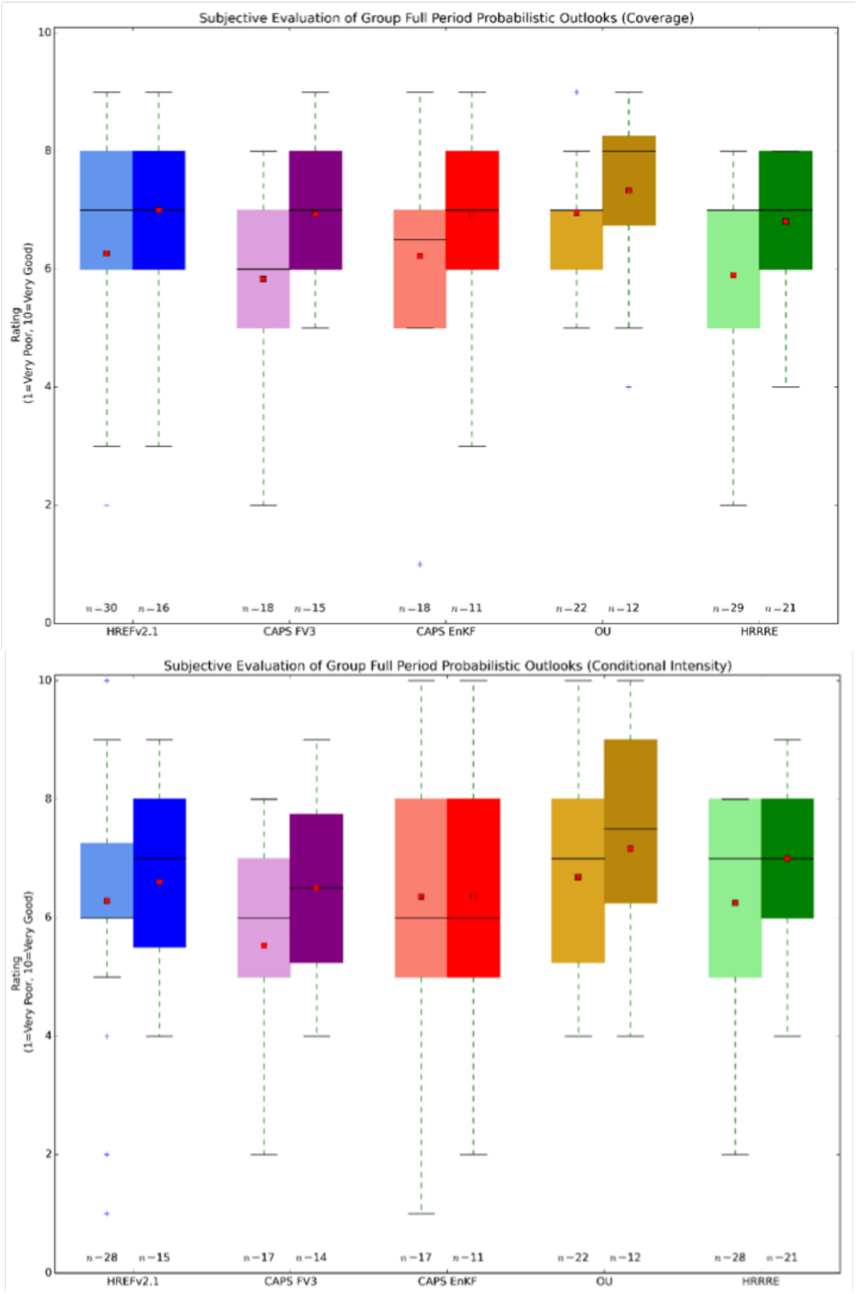


Figure 21 Participant subjective ratings of their group's (top) coverage and (bottom) conditional intensity forecasts. Lighter colors indicate the initial morning forecasts, and darker colors indicate the afternoon updated forecasts. All hazards were considered in assigning these ratings.

When looking at the distribution of ratings assigned to each group's forecasts (Fig. 21), the clearest trend is the increase in ratings for the afternoon update coverage forecasts compared to the initial morning coverage forecasts, with every group except for the HREFv2.1 and HRRRE groups having a 1-point increase in the median and a large shift in the distribution (Fig 21, top). All groups' mean ratings also increase. Overall, the HREFv2.1, HRRRE, and OU groups have the highest median rating for the initial coverage forecasts, and the OU group has the highest median rating for the updated coverage forecasts.

For conditional intensity forecasts issued by the groups (Fig 21, bottom), again the ratings typically increase going from the morning to the afternoon forecasts. However, unlike the coverage forecasts, this did not happen uniformly across the groups. The CAPS EnKF and HRRRE groups have the same median rating, and the CAPS EnKF group maintains the same mean rating as well. For the conditional intensity, the OU group had the highest ratings. Taken with the expert forecast ratings, these results seem to suggest that while forecasting the coverage and location of severe weather becomes easier as the event approaches, determining the significance of the severe weather doesn't necessarily become easier with decreasing lead time.

When asked specifically about how WoFS guidance affected their forecasts, approximately half of the time participants indicated that they didn't look at the WoFS guidance or relied far more on current observations, which at times may have been due to the WoFS guidance not being available. However, when WoFS was available and considered by forecasters, they most often said that it improved their forecasts. Some participants mentioned that they used it in concordance with observations to refine the area of their forecast coverage. Even more than the location and coverage, participants mentioned that WoFS was useful in identifying the severity of storms and improving their confidence in the occurrence of severe weather. As one participant stated, "the ensemble probabilities were a big factor in where I changed the extent of my conditional intensity boundaries and where I increased the conditional intensity values". Storm-scale attributes such as probabilities and percentiles of UH, as well as reflectivity, were mentioned most often as products utilized from the WoFS. One challenge mentioned by the participants was the limited area domain of the WoFS, which is smaller than the daily subdomain they were tasked with issuing their forecasts for. Another issue mentioned by participants was the relatively short length of the guidance, considering that the forecasts available while they were issuing their updates (valid until 1200 UTC) only ran to 0100 UTC or 0200 UTC depending on the initialization time. Eventually, the goal of the FACETs paradigm is to bridge the gap between convective outlooks, watch scale guidance, and warning scale guidance with probabilities, and these comments reflect some of the challenges to creating guidance and forecast products across different spatial and temporal scales.

To understand the drawing of the conditional intensity forecasts, participants were asked about how easy they were to draw and to describe any specific challenges encountered when drawing them. Participants typically were able to generate the conditional intensity forecasts without too much trouble, with a majority of participants stating that the forecasts were "neither difficult nor easy", "easy", or "very easy" (Fig. 22). Responses to the question "What was the biggest challenge associated with creating the conditional intensity forecasts yesterday?" followed three major themes focused on challenges. Participants often mentioned that knowing where exactly to draw the forecasts was challenging, similar to what we have previously found for the coverage probabilities. Convective mode and evolution were also a challenge, specifically because of the focus on hazard type and the potential to produce significant severe weather for each type. Finally, participants occasionally mentioned difficulty with the CAM guidance, wanting to look at more ensembles beyond their specific subset, but also pointing out that there are few reliable diagnostics that can distinguish significantly severe weather from severe weather overall. Developing some of these diagnostics would help participants issue the conditional intensity forecasts more easily. Participants also made several comments demonstrating that they grasped the concept and found it intuitive. Several comments were made to the effect of "there were no major challenges". There were also mentions of how the forecasts could be useful from a threat communication standpoint, motivating further development and testing of this product.

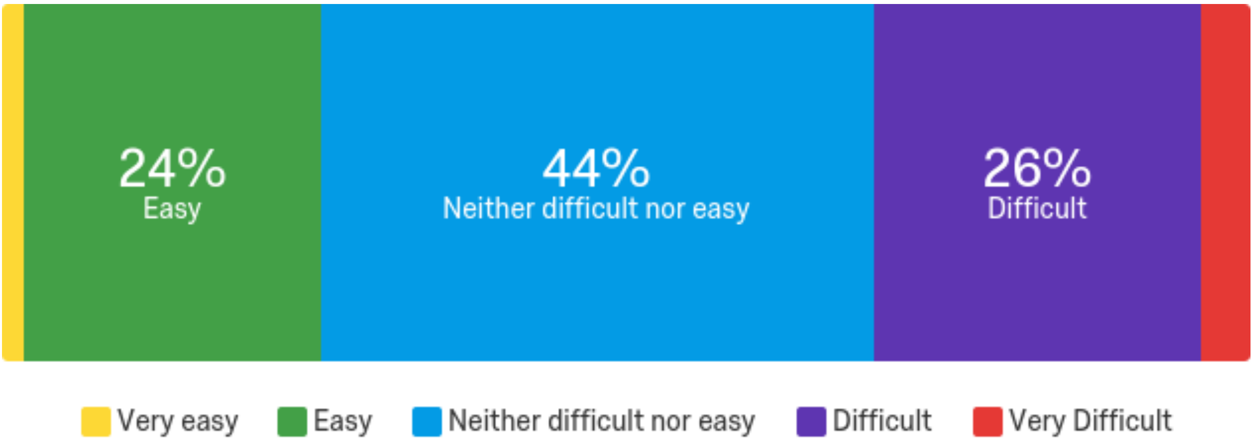


Figure 22 Participant responses to the question "How difficult was it to create the conditional intensity forecasts yesterday?" 2% of participants responded "Very easy", and 4% of participants responded "Very Difficult".

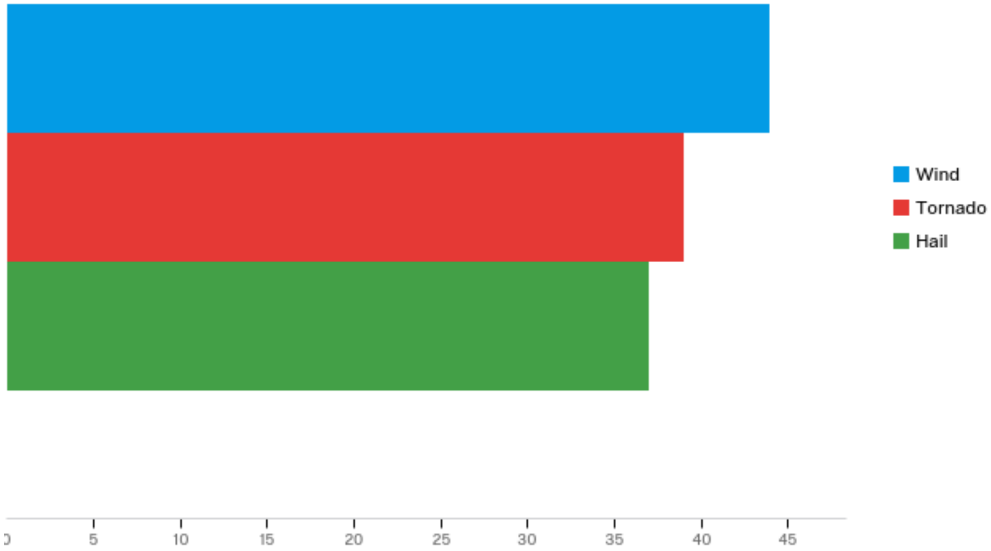


Figure 23 Participant responses to the question "Which hazard was most difficult to draw for, and why?".

When asked which hazard was most difficult to draw for, participants were split relatively evenly between the three hazards, with all three receiving over 35 responses across the 2019 SFE (Fig. 23). Wind was most often selected as the most difficult hazard to draw for. Five responses selected two hazards as being most difficult. Of these five, two selected both wind and hail, two selected tornado and hail, and one selected tornado and wind. When explaining why the forecasts were difficult, the tornado and wind response indicated challenges pinpointing the location with the tornado hazard, and difficulty uncertainty regarding the upscale growth of the system with the wind hazard. One hail and wind respondent indicated that storm mode was the main challenge, and the other respondent indicated that the entire situation for the day (which produced hail and wind reports) was challenging. The tornado and hail respondents indicated challenges with the size of the area under hail threat, and referenced the climatology of tornadoes for the region.

For participants that selected the tornado threat as being most difficult, explanations as to why included multiple factors. Some participants indicated marginal environments for tornadoes, others described difficulty in convective mode, and others mentioned issues with the CAMs. Specific comments acknowledged that tornadoes require the most "interpretation" from the CAM diagnostics. Participants who selected hail as the most difficult threat also reference the convective mode, but also discuss the location of convection more frequently than those that selected tornado as the most difficult hazard. Participants that indicated wind as the most difficult hazard were more likely to state the difficulty with observations, and not knowing whether or not the winds were significantly severe without direct measurements. One participant's comment represents the theme of these responses particularly well: "It's difficult to differentiate between it just being windy because there are storms, and the wind itself being the/a significant hazard".

3) CALIBRATED AND TEMPORALLY DISAGGREGATED GUIDANCE

After viewing the forecasts made on the previous day, participants at the severe hazards desk viewed calibrated guidance based around the HREFv2.1 for each hazard. This guidance was evaluated for both the entire forecast period and for 4-h chunks covering the entire day. Additionally, temporally disaggregated guidance was evaluated. This guidance took the individual hazard forecasts generated by the expert forecaster with the help of the group and used timing information from the HREF/SREF calibrated guidance to break down the probabilities into 4-h periods. Essentially, the probability locations and magnitudes were reliant on the forecaster-generated outlook, with the timing primarily driven by the HREFv2.1. Temporally disaggregated probabilities were computed using both the initial morning forecast and the afternoon update.

The full-period calibrated guidance garnered good ratings from participants, with a median score of 6/10 for all of the guidance (Fig. 24). When looking at mean ratings for each hazard, the STP-calibrated tornado guidance (Gallo et al. 2018) scored slightly higher than the HREF/SREF tornado guidance (Jirak et al. 2014), and the Machine Learning hail guidance scored slightly higher than the HREF/SREF hail guidance. Scores were relatively similar across types of guidance. The 4-h calibrated guidance and temporally disaggregated outlooks also most often achieved a rating of 6/10 (Fig. 25), although there was more variability. The 4-h HREF/SREF guidance for wind and hail scored lower than the full period

HREF/SREF guidance, perhaps indicating that this guidance is more useful/skillful on a longer timeframe. However, the temporally disaggregated guidance shows improvement relative to the HREF/SREF guidance (i.e., one of the inputs) in the mean for all hazards, and most especially for wind. A median score improvement of 2 points occurred from the calibrated guidance to the temporally disaggregated guidance for the wind hazard. Participant feedback on the temporally disaggregated guidance was extremely positive, mentioning that the guidance would be very useful in the field and help Decision Support Services (DSS). They also expressed interest at seeing these products over more seasons and mentioned that the disaggregation adds valuable information to standard SPC outlooks. When the temporally disaggregated guidance did not perform well, participants often linked it to a poor input forecast. For example, if the input forecast does not have an area where storms occur, the temporal disaggregation algorithm will not put probabilities there. Conversely, if the input forecast is too high, it can lead to overdone temporally disaggregated probabilities as well.

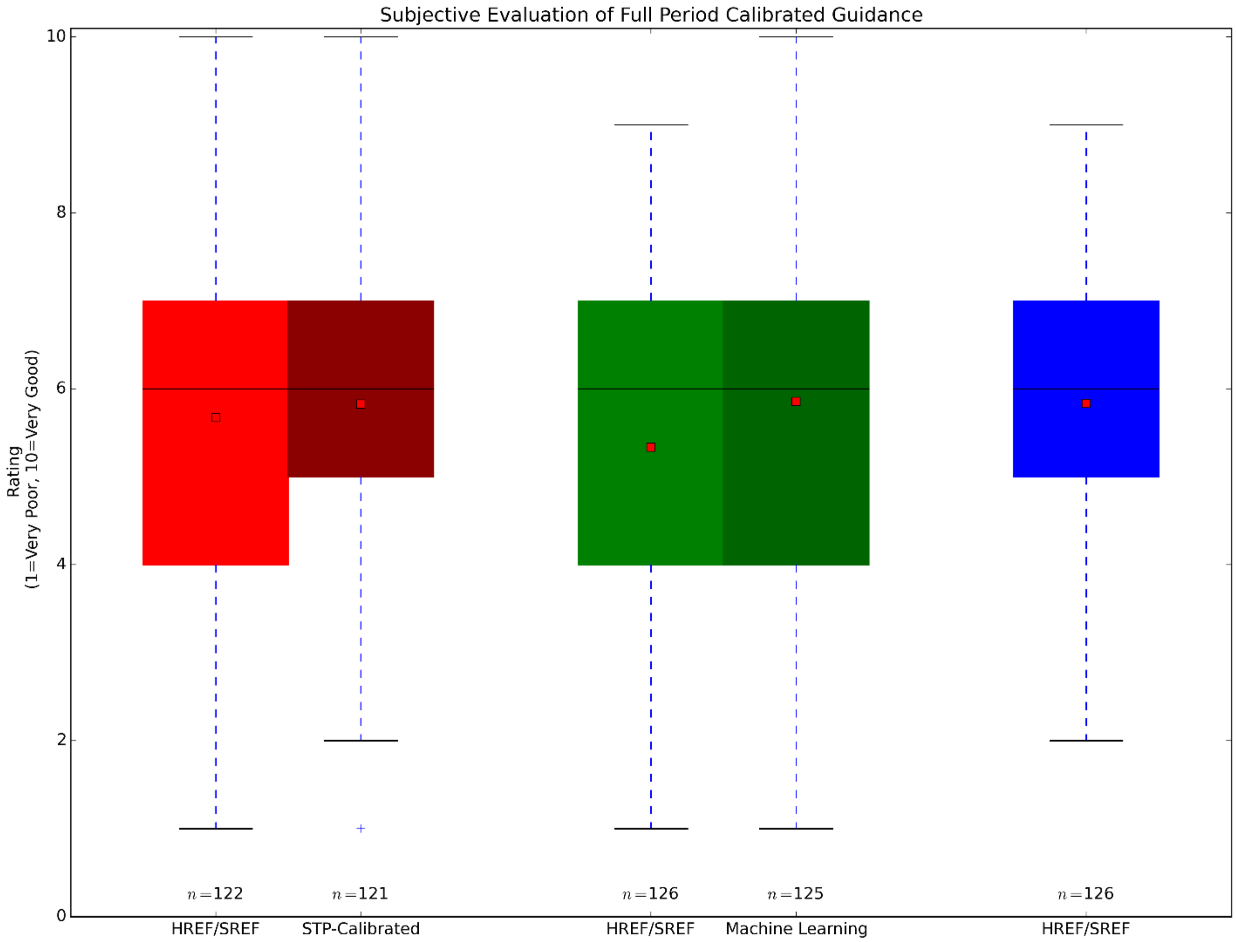


Figure 24 Participant subjective evaluation of full period calibrated guidance available for each hazard.

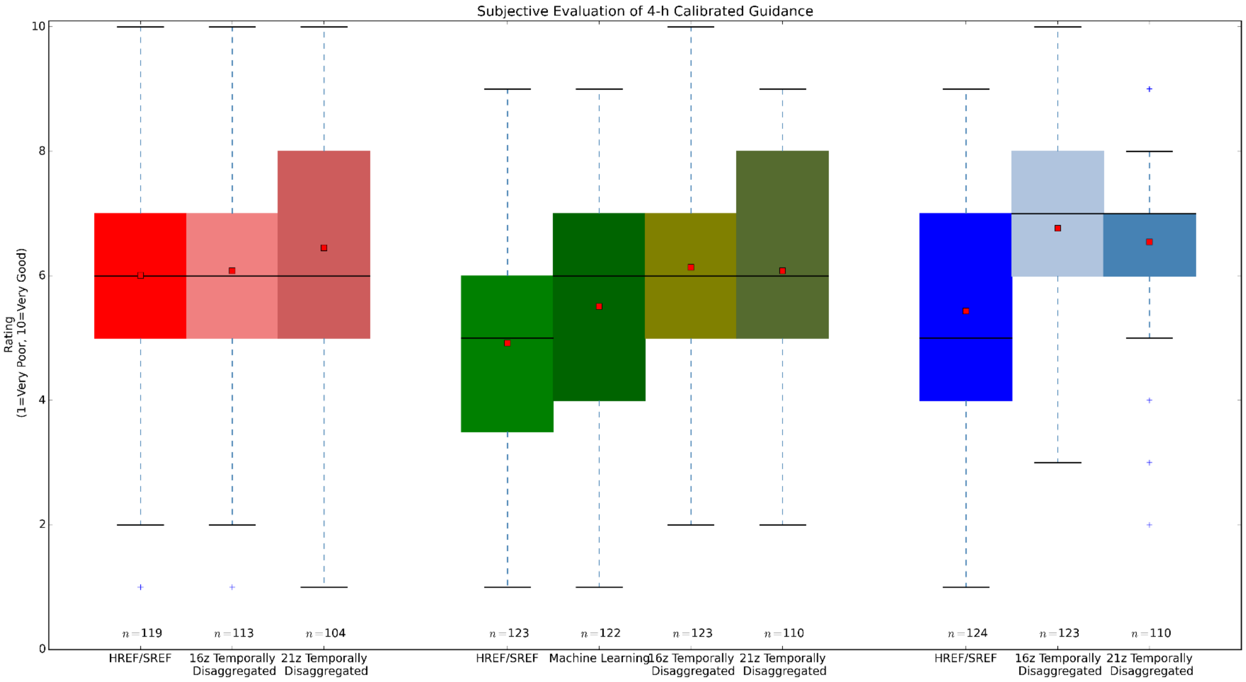


Figure 25 Participant subjective evaluation of 4-h calibrated guidance and temporally disaggregated outlooks for each hazard.

c) Model Evaluations – Innovation Desk (credit: B. Gallo)

1) FV3 COMPARISON

The first deterministic model comparison looked at five different versions of the FV3 provided by NSSL, EMC, and GFDL. Three versions were nested versions, with a high-resolution domain nested in the parent global FV3 domain. Two other versions were stand-alone regional (SAR) versions, with lateral boundary conditions rather than running the model across the full globe. Participants were asked to rate two groups of fields on a scale of 1 (Very Poor) to 10 (Very Good): the reflectivity and UH fields were grouped together, as were the thermodynamic fields of 2-m temperature, 2-m dewpoint, and surface-based convective available potential energy (SBCAPE). Participants were also asked to provide comments on the model performance via an open-ended text box.

Participants generally provided ratings between 4/10 and 7/10 for most of the models (Fig. 26). There are few differences between the ratings for the reflectivity/UH and the thermodynamic variables, with the exception of the GFDL nest and NSSL nest performing better for the thermodynamic variables as compared to the reflectivity/UH fields for the same model. Median scores for UH and reflectivity were tied at 6/10 for the EMC nest, EMC SAR, and NSSL SAR, but mean values show that the NSSL SAR generally had higher ratings than the two EMC models. The mean ratings for the EMC models were quite similar, with the EMC nest having a wider distribution in reflectivity and UH ratings than the EMC SAR.

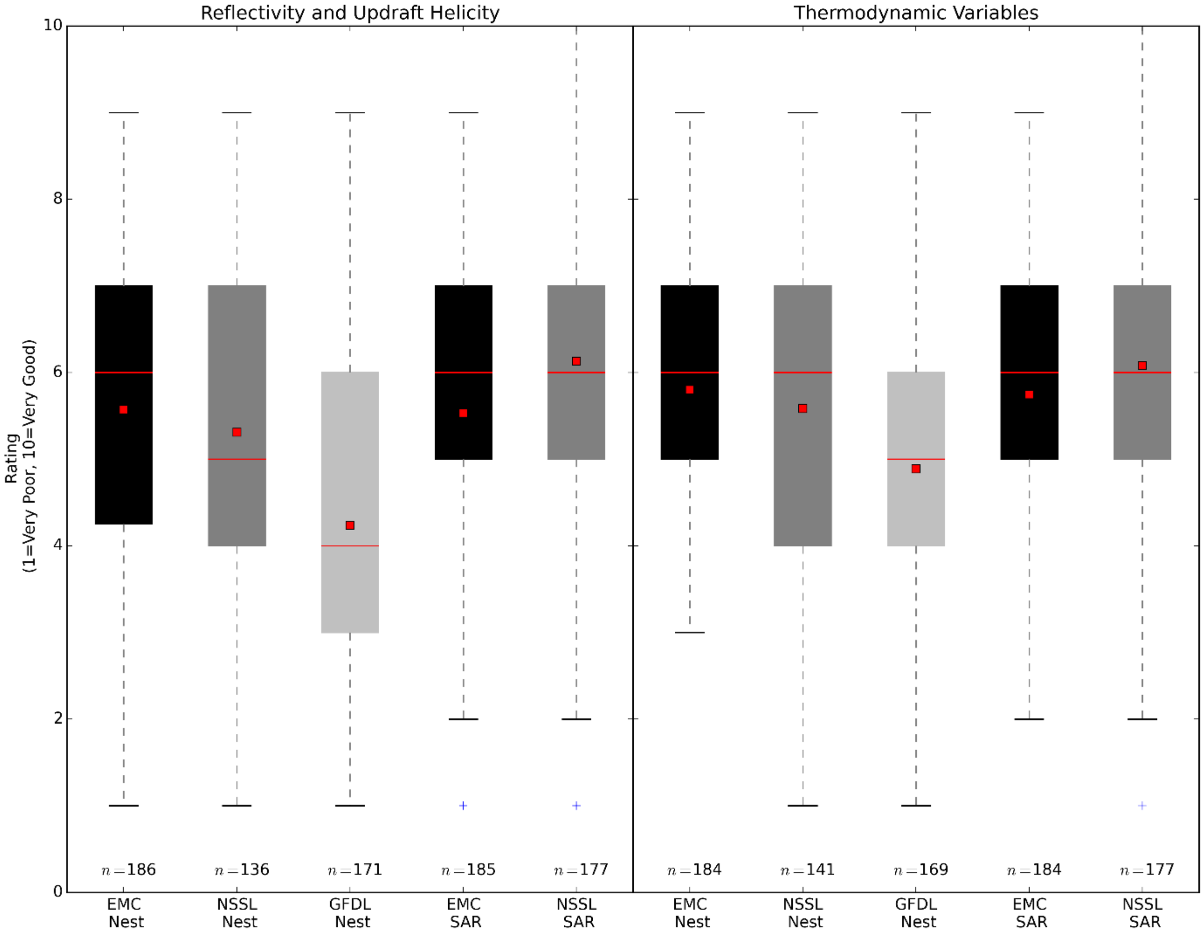


Figure 26 Participant subjective ratings of five nested and SAR FV3 configurations provided by three organizations. Red squares indicate the mean rating, and response sample size is located directly above the label. Reflectivity and UH ratings are on the left, and thermodynamic variable ratings are on the right.

Participant comments when asked about these configurations followed three main themes. First, in accordance with the objective ratings, the SAR configurations were often rated better than the nest configurations. The EMC nest was mentioned as “best” within the comments more often than any of the other nest versions, and was mentioned as best almost as much as the two SAR configurations. The GFDL nest was most often mentioned by participants as doing the worst. These best and worst rankings seemed to line up with how participants evaluated the convective mode and evolution of the storms. The NSSL SAR and two EMC versions were often described by participants as having a good mode and evolution of convection, while the GFDL nest and NSSL nest were often described as being too linear compared to the other models. Second, the NSSL-SAR was the only model that had mentions of good intensity in the reflectivity and UH, although comments that it was too intense were more frequent than the comments that the intensity was good. However, the reflectivity (and occasionally UH) intensity in all of the other models was often mentioned as being too intense. Finally, this set of models was deemed to be too cold at the surface across the board. The 2-m cold bias was seen both in the overall temperatures for the models, and occasionally within cold pools produced by simulated convection.

2) MIXED CORE COMPARISON

The second set of models examined within the deterministic model comparison included models with multiple dynamical cores. A main focus of the comparison was the HRRRv3 and the HRRRv4, the operational and developmental versions of the HRRR. These two models were compared with the HRRR-FV3, a version of FV3 configured to be as close to the physics of the HRRR as possible, the CAPS-FV3, and a member of the UK Met Office-provided ensemble, which uses the Unified Model (UM) core. The HRRR-FV3 was only available during the second half of the experiment, and thus has approximately half of the sample size of the other configurations. Also, development work was still being conducted for HRRRv4 during the SFE, so it's configuration did not remain constant.

In general, the HRRRv3 and HRRRv4 performed better than the HRRR-FV3, CAPS FV3 SAR, and the UM, with a median rating of 6/10 compared to 5/10 for the reflectivity and the UH fields (Fig. 27, left). The HRRRv4 achieved the highest median rating of the five models for the thermodynamic variables (Fig. 27, right), scoring 7/10. Otherwise, the median ratings are the same for each model between the reflectivity/UH comparisons and the thermodynamic comparisons. These distributions are similar to those found in the prior set of FV3 models. Mean ratings for the reflectivity and UH are slightly higher in the HRRRv3 than the HRRRv4, while the opposite is true for the thermodynamic variables. The HRRR-FV3 had the worst mean rating for the reflectivity and UH fields, while the CAPS FV3 SAR had the worst mean rating for the thermodynamic variables.

When participants were asked about the largest differences they saw between the operational HRRR (HRRRv3) and the developmental HRRR (HRRRv4), participants noted three main themes. First, many participants saw only small differences between the models, which occurred more often than either of the models being singled out as best. This concurs with the numerical ratings, which showed very similar distributions between the HRRRv3 and the HRRRv4. Second, when differences were observed participants noted a tendency for reflectivity to be too intense in the HRRRv4. Finally, participants also noted multiple times that the evolution and mode were better in the HRRRv4 than in the HRRRv3. Thermodynamic fields were also noted as improved in the HRRRv4 than in the HRRRv3, and there was a tendency for the HRRRv4 to be drier than the HRRRv3. When looking at participant comments of all five deterministic models in the second comparison, participants often noted a cool and moist bias in the CAPS-FV3, and little systemic difference in the models despite the different dynamical cores. Again, this fits relatively well with the numerical ratings, given the similarity in the distributions between the UK Met Office UM, CAPS-FV3, and the HRRR-FV3. Often, the participants noted that different models were better at different points in the 24-h cycle observed, leading to difficulty in rating. Difference plots for the thermodynamic fields and some objective measure of verification were suggested as possibilities for future experiments, to help participants with the rating process for the deterministic CAMs.

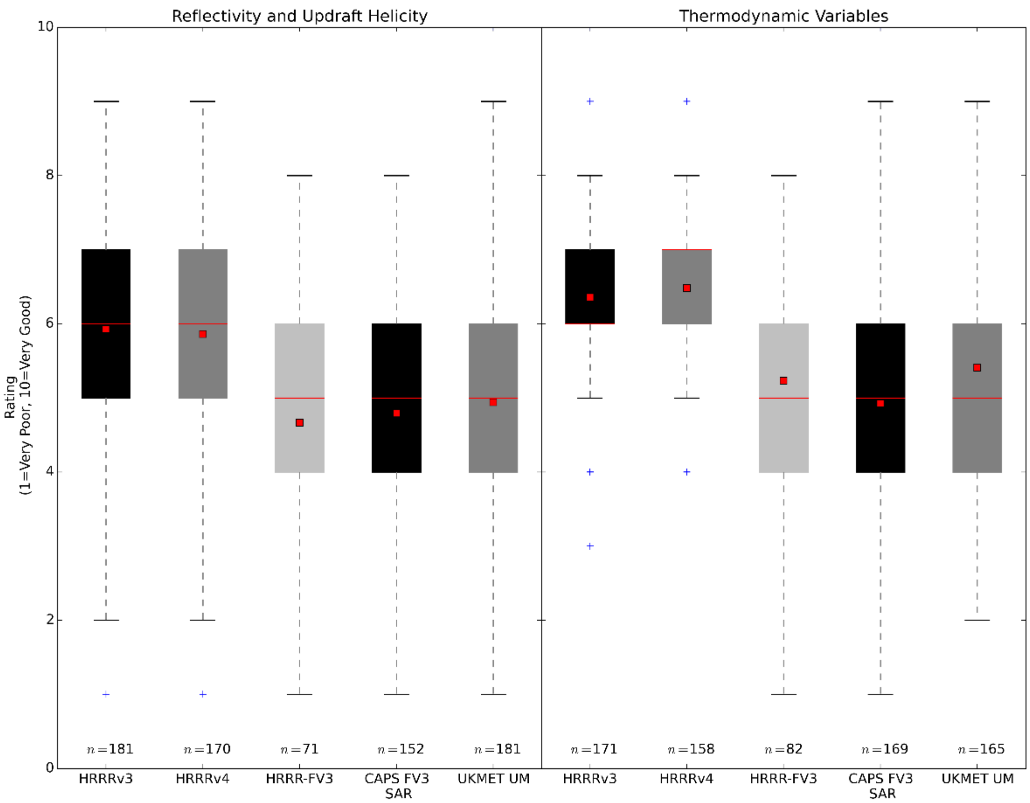


Figure 27 Participant subjective ratings of five deterministic model configurations provided by three organizations. Red squares indicate the mean rating, and response sample size is located directly above the label. Reflectivity and UH ratings are on the left, and thermodynamic variable ratings are on the right.

3) FV3 PHYSICS EVALUATION

For the second year, CAPS provided an ensemble of CAMs using the FV3 dynamical core, with differences between the microphysics schemes and PBL schemes. This ensemble allowed participants to evaluate the impacts of different physics parameterizations. Participants looked at thermodynamic variables from members with three varying PBL schemes and reflectivity and UH fields from members with three different microphysics schemes, as the largest differences between PBL schemes and microphysics schemes were anticipated to be in the thermodynamic fields and reflectivity/UH fields, respectively.

The member using the Morrison microphysics differed most from the observations compared to the Thompson and NSSL member, scoring the lowest in the participant subjective evaluations with a median rating of 4/10 (Fig. 28, left). Median ratings for the Thompson and NSSL members were the same, but the mean rating for the NSSL member was slightly higher. The Morrison scheme also had a larger range of responses than the other members, garnering ratings from 1/10 to 9/10 over the entire course of the SFE. Participants noted the relatively low intensity of the simulated reflectivity in the Morrison scheme, particularly in the convective cores. However, despite the weak intensity of the convective cores in the Morrison forecasts, participants noted that the timing and location of the storms was often in the correct place. Additionally, high reflectivity values across broad swaths of stratiform precipitation in the Morrison member was also mentioned by multiple participants. Therefore, several participants recommended adjusting the dBZ diagnostic within the Morrison scheme. Also, the NSSL and Thompson simulated reflectivity were noted to be too high for a number of cases. Many other cases were noted where all three models showed significant limitations, resulting in the median ratings of 4/10 or 5/10 for all three runs.

The three forecasts with varying PBL schemes showed very little difference in their subjective ratings, with almost identical distributions (Fig. 28, right). These results agree with results from SFE 2018, which found that the members of the CAPS FV3 ensemble that solely varied PBL schemes had very similar forecasts. When asked about the largest differences between these three models, participants said that the models were quite similar – that the difference between any of the models and the observations were larger than the differences between the models themselves. A persistent moist bias was seen throughout the three runs evaluated (occasionally as large as 5–10°F), although there were a few comments that singled out the Shin-Hong as having the highest moisture bias.

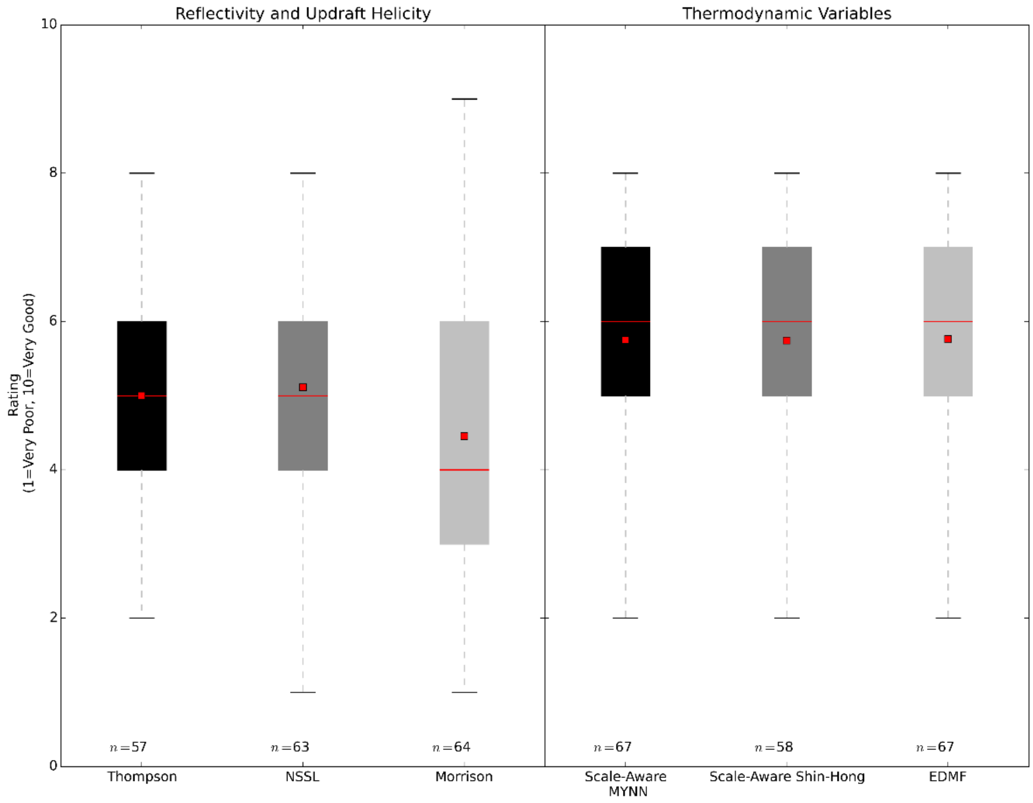


Figure 28 Participant subjective ratings of different FV3-based configurations with different microphysics parameterization schemes (left), and PBL schemes (right). Red squares indicate the mean rating, and response sample size is located directly above the label.

4) WOFS EVALUATION

As a transition into the WoFS activity, the final afternoon evaluation at the Innovation Desk focused on the WoFS system from the previous day and compared it to two versions of a time-lagged HRRR ensemble (HRRR-TL). The first version of the HRRR-TL used the previous four runs of the HRRR, and the second version used six previous runs of the HRRR. Participants evaluated hourly and 4-hourly neighborhood probabilities of UH and underlying UH values compared to observed storm reports within the 900 km x 900 km WoFS domain. Comparisons were made for two initializations: the 1900 and 2100 UTC. The 1900 UTC initialization was the first available WoFS forecast during the 2019 SFE, while the 2100 UTC forecast was launched as the peak of the daily convective cycle approached.

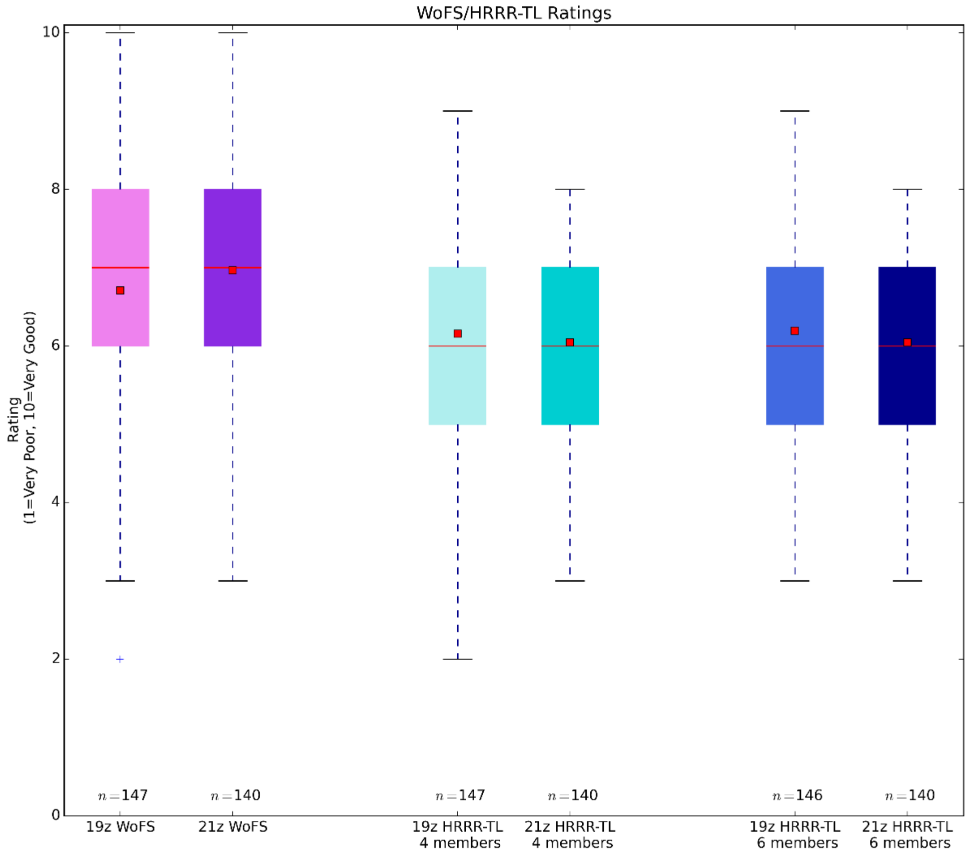


Figure 29 Participant subjective ratings of the Warn-on-Forecast System and two versions of a HRRR-Time-Lagged ensemble. The 1900 UTC cycle ratings are in the lighter colors, and the darker colors indicate the 2100 UTC cycle ratings. Red squares on each plot indicate the mean subjective rating, and the sample size is indicated above the x-axis label for each distribution.

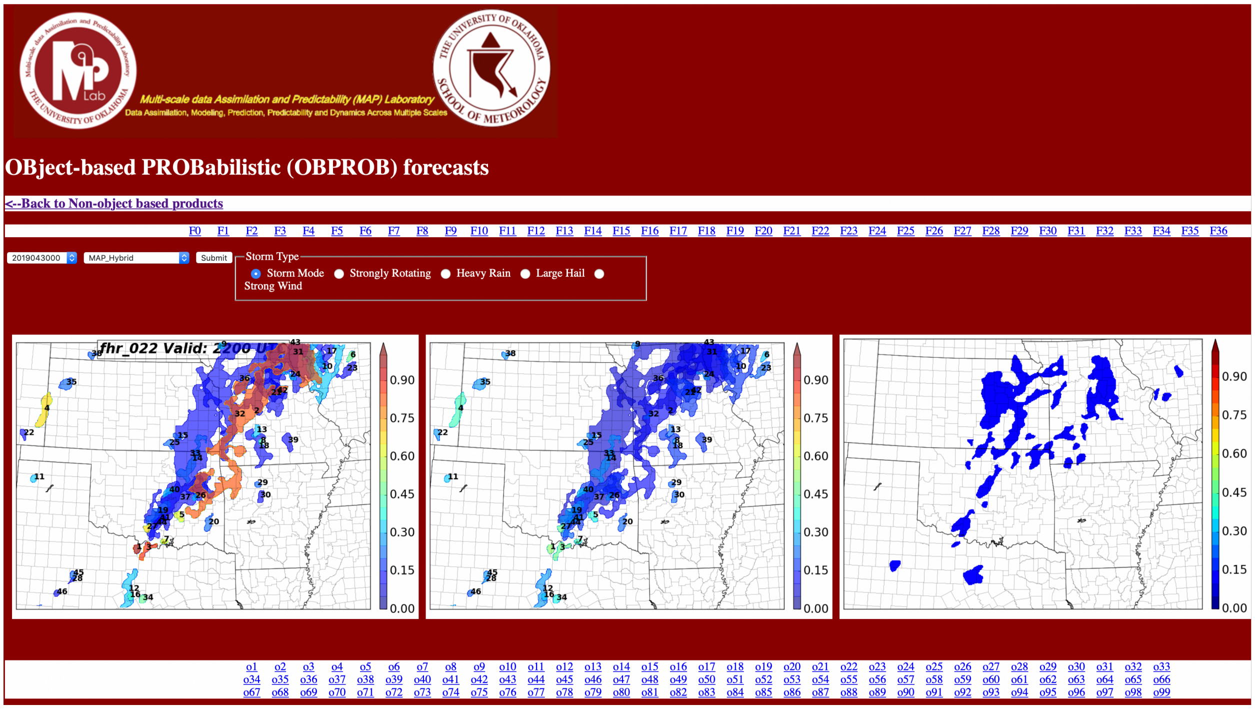
Through both the 1900 and 2100 UTC forecasts, the WoFS outperformed the HRRR-TL ensembles of both sizes (Fig. 29). The median ratings were on average one point higher for the WoFS runs than for the HRRR-TL counterparts, and median ratings did not shift for any ensemble between the 1900 and the

2100 UTC runs. However, mean ratings reflected different patterns for the WoFS compared to the HRRR-TL ensembles. The WoFS mean rating increased from 1900 to 2100 UTC, perhaps reflecting the effect of the rapid and more advanced data assimilation methods, which we would expect to improve the forecasts once storms initiate. The HRRR-TL ensembles conversely saw their mean ratings decrease from 1900 to 2100 UTC, though the absolute difference was smaller than the increase realized by the WoFS. When asked what the largest differences were between the ensembles at 1900 UTC, participants mentioned a better representation of the outliers in the WoFS than the HRRR-TL. Given that WoFS has 18 members compared to the 4- and 6-member HRRR-TLs, this result would be expected for a well-performing ensemble. Also mentioned were false alarms in WoFS compared to the HRRR-TL ensembles – in some cases, WoFS lessened instances of false alarms, but in other cases the WoFS had more widespread probabilities compared to the HRRR-TL and increased areas of false alarm. Participants also often mentioned when different systems were capturing different areas of convection better or worse, or when one system captured early storms better and another captured later storms better – with the active 2019 spring season, there were often multiple rounds of convection even within the relatively limited WoFS domain. Very little difference was seen in both the numerical ratings and the text comments between the 4-member and 6-member HRRR-TL. Looking at the 2100 UTC ensembles, participants noted either minimal differences with the 1900 UTC runs or a slight improvement in the WoFS compared to the 1900 UTC initialization.

5) OBJECT-BASED PROBABILISTIC (OBPROB) FORECAST EVALUATION (credit: Aaron Johnson)

Another daily evaluation focused on the OBPROB products in the map-hybrid ensemble web interface (e.g., Fig. 30) created using the forecasts for the previous day. Using the forecast from the previous day allowed participants to evaluate not only how well the OBPROB products corresponded to their subjective interpretations of the convective scale ensemble forecast, but also how well the OBPROB products corresponded to the verifying observation objects. Comparisons were also made between the raw ensemble OBPROB products, where probability for each object is determined by the percentage of ensemble members with a matching object, and a (simple preliminary) calibrated OBPROB product, where the probability for each object is determined by a logistic regression model based on various object attributes trained using the 2018 map-hybrid ensemble OBPROB forecasts. Individual plotted objects could also be further interrogated using the arbitrary number plotted at each object's centroid by clicking that number at the bottom of the page.

An example of the focus of the daily evaluations is illustrated by Figures 30-32. Figure 30 indicated that at the 22-hour lead time (2200 UTC) the high probability of a large linear object from eastern Oklahoma into Missouri and far western Illinois (left panel) indicated that the ensemble was predicting upscale growth of the cellular and multi-cellular storms into an organized MCS already at this time. In the observations (right panel), this upscale growth did not occur until a couple of hours later. The calibrated object probabilities (center panel) decreased the probability in this large object, which was more consistent with its absence in the observations at this time. There was also at least one high probability cellular object near the Oklahoma-Texas border at this time (left panel of Fig. 30). Further examination of this object (Fig. 31) allowed participants to better visualize the uncertainty of the forecast of such an object in terms of both spatial location and convective structure in the ensemble of matched objects.


Multi-scale data Assimilation and Predictability (MAP) Laboratory
Data Assimilation, Modeling, Prediction, Predictability and Dynamics Across Multiple Scales
MAP Lab
The University of Oklahoma School of Meteorology
<--Back to Non-object based products
F0 F1 F2 F3 F4 F5 F6 F7 F8 F9 F10 F11 F12 F13 F14 F15 F16 F17 F18 F19 F20 F21 F22 F23 F24 F25 F26 F27 F28 F29 F30 F31 F32 F33 F34 F35 F36
Submit
Storm Type
● Storm Mode ○ Strongly Rotating ○ Heavy Rain ○ Large Hail ○ Strong Wind
fhr_022 Valid: 2200 UTC
0.90
0.75
0.60
0.45
0.30
0.15
0.00
o1 o2 o3 o4 o5 o6 o7 o8 o9 o10 o11 o12 o13 o14 o15 o16 o17 o18 o19 o20 o21 o22 o23 o24 o25 o26 o27 o28 o29 o30 o31 o32 o33
o34 o35 o36 o37 o38 o39 o40 o41 o42 o43 o44 o45 o46 o47 o48 o49 o50 o51 o52 o53 o54 o55 o56 o57 o58 o59 o60 o61 o62 o63 o64 o65 o66
o67 o68 o69 o70 o71 o72 o73 o74 o75 o76 o77 o78 o79 o80 o81 o82 o83 o84 o85 o86 o87 o88 o89 o90 o91 o92 o93 o94 o95 o96 o97 o98 o99

OBject-based PROBabilistic (OBPROB) forecasts

2019043000 | MAP_Hybrid

Figure 30 Web interface to the map-hybrid ensemble OBPROB products for the 22-hour forecast valid at 2200 UTC 30 April 2019. The three panels from left to right show the raw ensemble OBPROB product, calibrated OBPROB product, and verifying observation objects, respectively.

Three hours later, at 0100 UTC (Figure 32), the OBPROB forecast was indicating high confidence that the most likely extent of a sub-synoptic scale squall line would be from the upper Midwest down to eastern Oklahoma (left panel). In the observations there was such a squall line, but it extended all the way down into northwest Texas (right panel). In contrast, the ensemble forecast indicated that the highest probability objects in the vicinity of the Red River would have a discrete cellular mode.

When asked the question “Did the OBPROB guidance provide unique information about the forecast challenge of the day and/or more concisely confirm your interpretation of non-object- based products?”, there was a lot of variability among different participants and among different forecast cases. Overall, based on subjective categorization of the answers to this open-ended question, 42.2 % (out of 147 total responses) of respondents said that OBPROB provided unique information and/or more concisely confirmed interpretation from manual or non-OBPROB evaluation of the ensemble, while 27.2 % of respondents said that it did not and 29.3 % of respondents were unsure or did not directly answer the question.

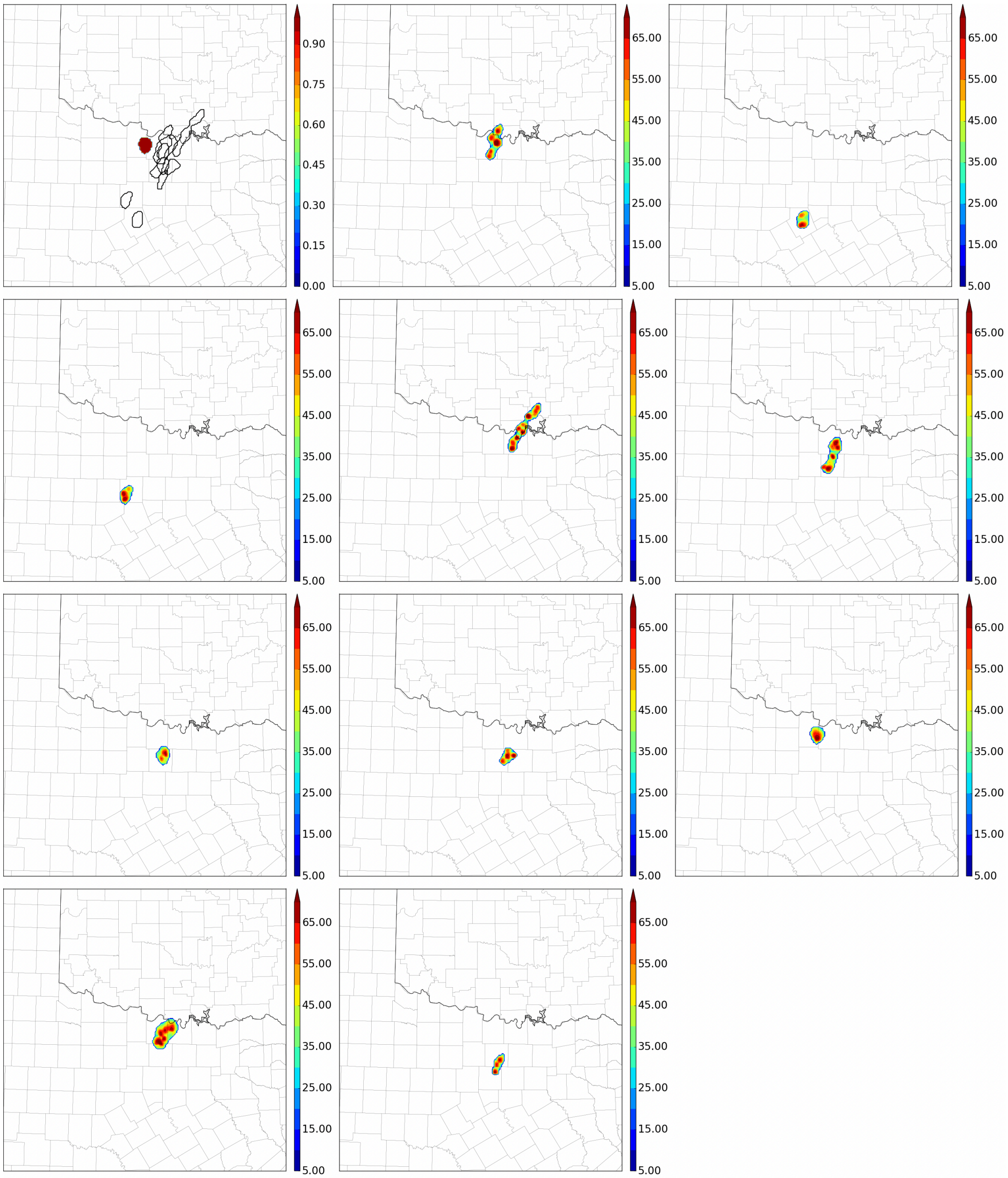
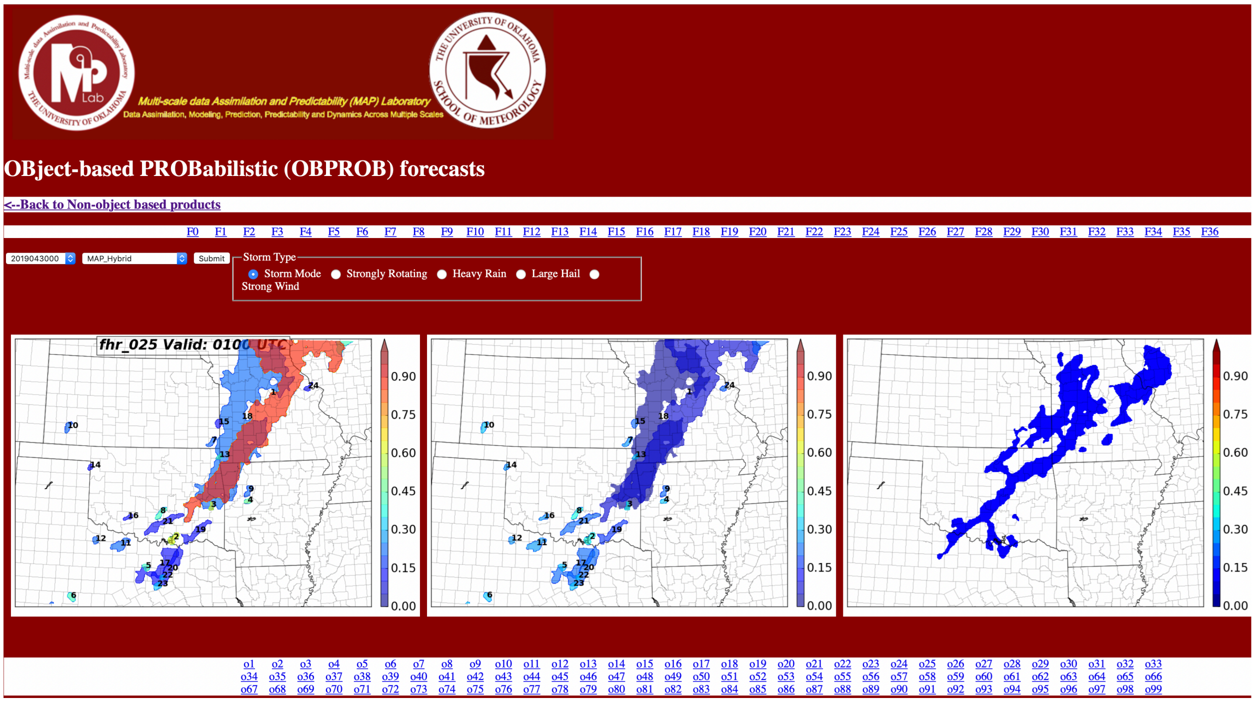


Figure 31 Pop-up image that would appear when clicking the "o1" link at the bottom of Figure 30 to see further details of the high-probability object in north-central Texas.

Most respondents thought that the preliminary (i.e., not thoroughly investigated or optimized) calibration applied to the OBPROB products did not add to the usefulness of OBPROB, with 65% of respondents indicating that the calibrated products were equally (32.5%) or less (32.5%) useful than the uncalibrated products. However, many respondents (19.7%) did find the calibrated products more useful (15.5% were unsure or did not answer).



OBject-based PROBabilistic (OBPROB) forecasts

Multi-scale data Assimilation and Predictability (MAP) Laboratory
Data Assimilation, Modeling, Prediction, Predictability and Dynamics Across Multiple Scales
<--Back to Non-object based products
F0 F1 F2 F3 F4 F5 F6 F7 F8 F9 F10 F11 F12 F13 F14 F15 F16 F17 F18 F19 F20 F21 F22 F23 F24 F25 F26 F27 F28 F29 F30 F31 F32 F33 F34 F35 F36
2019043000
MAP_Hybrid
Submit
Storm Type
● Storm Mode ○ Strongly Rotating ○ Heavy Rain ○ Large Hail ○ Strong Wind
fhr_025 Valid: 0100 UTC
0.90
0.75
0.60
0.45
0.30
0.15
0.00
o1 o2 o3 o4 o5 o6 o7 o8 o9 o10 o11 o12 o13 o14 o15 o16 o17 o18 o19 o20 o21 o22 o23 o24 o25 o26 o27 o28 o29 o30 o31 o32 o33
o34 o35 o36 o37 o38 o39 o40 o41 o42 o43 o44 o45 o46 o47 o48 o49 o50 o51 o52 o53 o54 o55 o56 o57 o58 o59 o60 o61 o62 o63 o64 o65 o66
o67 o68 o69 o70 o71 o72 o73 o74 o75 o76 o77 o78 o79 o80 o81 o82 o83 o84 o85 o86 o87 o88 o89 o90 o91 o92 o93 o94 o95 o96 o97 o98 o99

Figure 32 Same as in Figure 30, except at the 25-hour forecast valid 0100 UTC 1 May 2019.

Comparing the OBPROB forecasts to observations, among the choices of “very inaccurate”, “slightly inaccurate”, “slightly accurate”, “very accurate” or “not sure”, 61.6% considered them slightly accurate, 21.9% responded slightly inaccurate, 6.2 % responded very inaccurate, 5.5% responded very accurate, and 4.8% responded not sure.

While this feedback was encouraging, it is clear that there were still a lot of participants who were a little unsure of what to make of the OBPROB products. In unstructured discussions there were three comments that either came up repeatedly or were particularly salient as things that could be addressed to further improve the utility of OBPROB products to operational severe weather forecasters. For example, some forecasters believed that in order to really add unique and useful information compared to what can already be obtained from realistically sized (~10 member) CAM ensembles there should be some incorporation of underlying physical mechanisms into the object-based probabilistic display of storm modes and severe hazards. Second, it was often noted that the use of various thresholds (e.g., storm rotation, hail size, etc) to filter out “irrelevant” objects was an important advantage of OBPROB and should be further emphasized and optimized. Third, many forecasters found the interpretation of spatially overlapping objects to be a bit confusing. These valuable responses are being considered while developing future iterations of the OBPROB technique, and are expected to improve the utility to forecasters in future forecasting experiments.

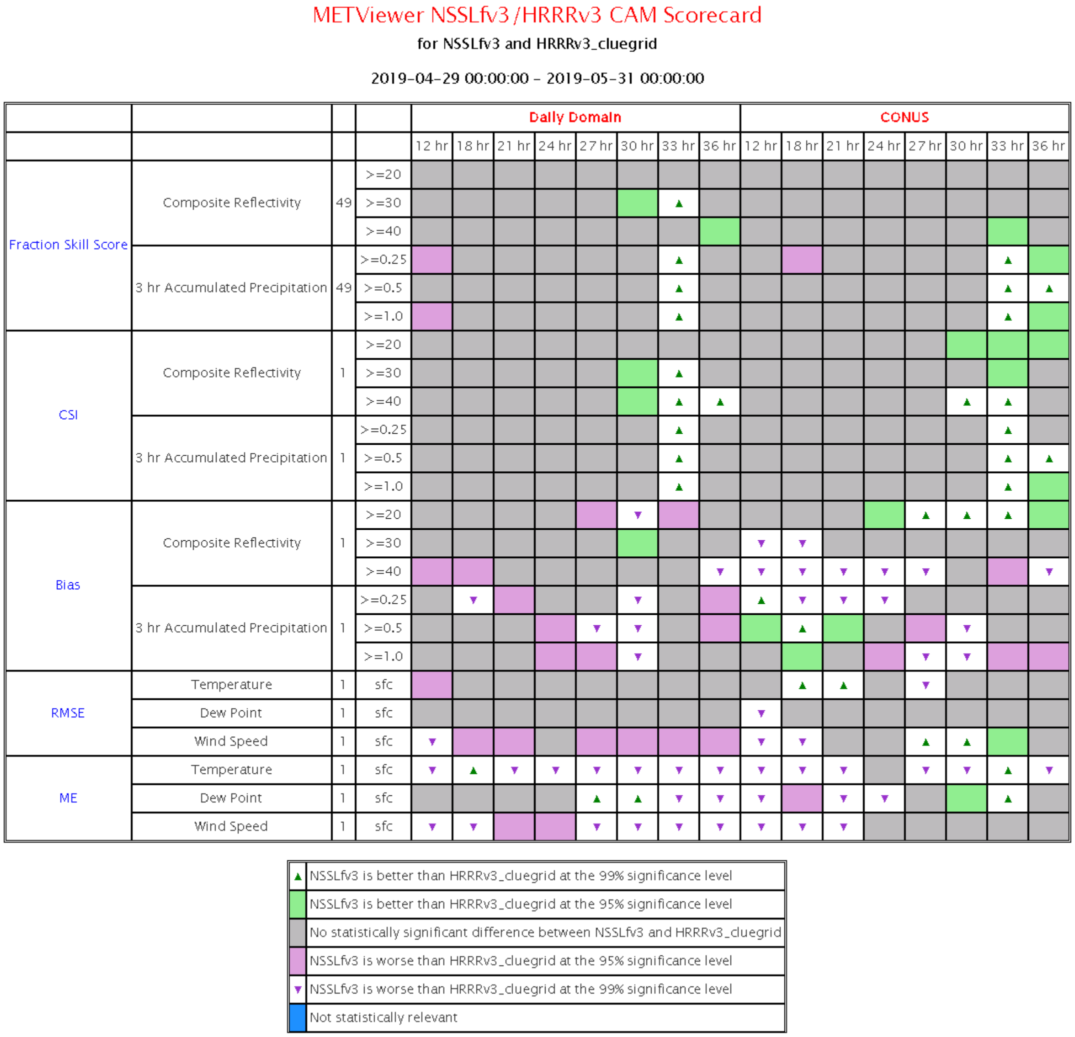
6) CAM SCORECARD

During the 2019 SFE, real-time scorecards were available at https://hwt.nssl.noaa.gov/sfe_viewer/2019/verification/scorecards.php for comparing objective verification of different models. Results from the scorecard were shown each Friday during the 11:30 AM (CDT) forecast briefing alongside a discussion of the background and aims of the scorecard.

Two ensembles (the HREFv2.1 and the HRRRE) and three deterministic CAMs (the HRRRv3, HRRRv4, and the NSSL-FV3 SAR) were compared using the scorecard after re-gridding to a common grid: The HRRRv3/HRRRv4 comparison updated daily, while the remaining comparisons updated twice weekly. Two scorecards were generated for each comparison for different fields, depending on whether the product was a full day (24-h) product or an hourly product. For the 24-h products, accumulated precipitation and surrogate severe fields based on a percentile of UH were evaluated. Hourly products included the simulated composite reflectivity, 3 h accumulated precipitation, temperature, wind speed, and dewpoint.

Due to configuration changes within some of the models being compared during the verification period, the NSSL-FV3 SAR/HRRRv3 scorecards are the only ones that objective conclusions should be drawn from at the time of this writing. Pending reruns of the relevant models, updated scorecards may be generated at a later date.

The NSSL-FV3 SAR/HRRRv3 scorecard comparing hourly fields shows mixed results (Fig. 33), with the composite reflectivity and 3-h accumulated precipitation in the NSSL-FV3 SAR outperforming the HRRRv3 at later forecast hours in Fractions Skill Score (FSS) and Critical Success Index (CSI) for both the contiguous United States (CONUS) and the daily domain of interest. However, the difference in bias between the two models at these hours is either statistically insignificant or slightly favors the HRRRv3. The high reflectivity threshold (40 dBZ) and 3-h accumulated precipitation threshold (≥ 1”) generally shows higher scores for the HRRRv3. Environmental variables generally favored the HRRRv3 over the NSSL-FV3 SAR, with the mean error (ME) having statistically significant differences for most of the hours across all three variables. The root mean squared error (RMSE) showed the same trend, but the wind speed was the only environmental variable that maintained a majority of the hours having a statistically significant difference. For all fields, larger statistical significance was often seen across the CONUS than the individual daily domains.



METViewer NSSLfv3 /HRRRv3 CAM Scorecard

for NSSLfv3 and HRRRv3_cluegrid

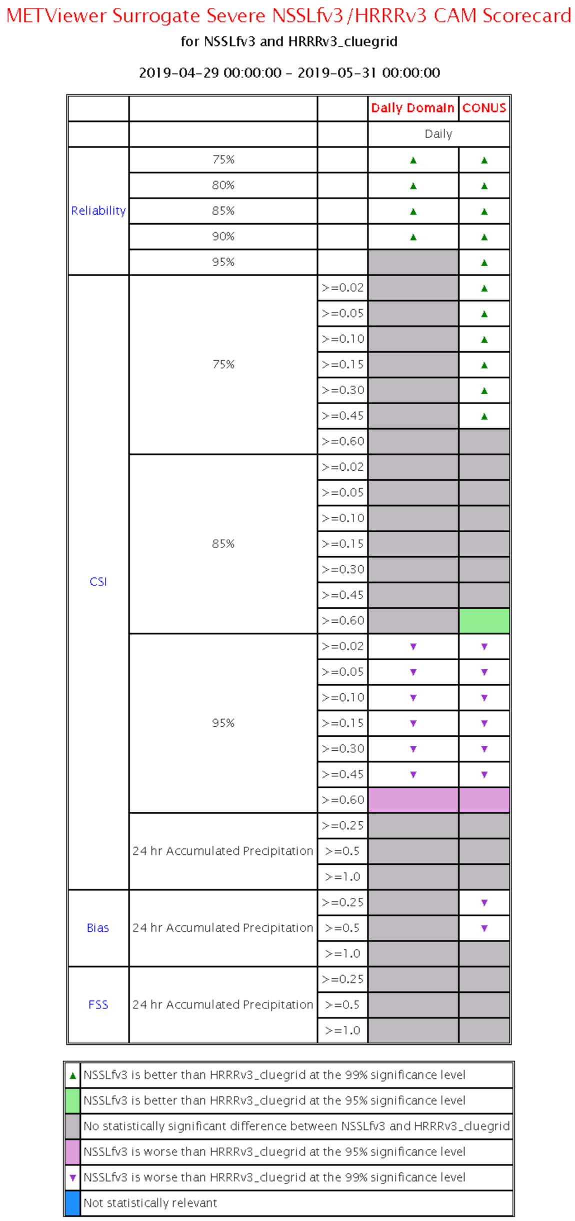
2019-04-29 00:00:00 - 2019-05-31 00:00:00

				Daily Domain								CONUS							
				12 hr	18 hr	21 hr	24 hr	27 hr	30 hr	33 hr	36 hr	12 hr	18 hr	21 hr	24 hr	27 hr	30 hr	33 hr	36 hr
Fraction Skill Score	Composite Reflectivity	49	>=20	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
			>=30	▨	▨	▨	▨	▨	🟩	▲	▨	▨	▨	▨	▨	▨	▨	▨	▨
			>=40	▨	▨	▨	▨	▨	▨	▨	🟩	▨	▨	▨	▨	▨	▨	🟩	▨
	3 hr Accumulated Precipitation	49	>=0.25	🟪	▨	▨	▨	▨	▨	▲	▨	▨	🟪	▨	▨	▨	▨	▲	🟩
			>=0.5	▨	▨	▨	▨	▨	▨	▲	▨	▨	▨	▨	▨	▨	▨	▲	▲
			>=1.0	🟪	▨	▨	▨	▨	▨	▲	▨	▨	▨	▨	▨	▨	▨	▲	🟩
CSI	Composite Reflectivity	1	>=20	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	🟩	🟩	🟩
			>=30	▨	▨	▨	▨	▨	🟩	▲	▨	▨	▨	▨	▨	▨	▨	🟩	▨
			>=40	▨	▨	▨	▨	▨	🟩	▲	▲	▨	▨	▨	▨	▨	▲	▲	▨
	3 hr Accumulated Precipitation	1	>=0.25	▨	▨	▨	▨	▨	▨	▲	▨	▨	▨	▨	▨	▨	▨	▲	▨
			>=0.5	▨	▨	▨	▨	▨	▨	▲	▨	▨	▨	▨	▨	▨	▨	▲	▲
			>=1.0	▨	▨	▨	▨	▨	▨	▲	▨	▨	▨	▨	▨	▨	▨	▲	🟩
Bias	Composite Reflectivity	1	>=20	▨	▨	▨	▨	🟪	▼	🟪	▨	▨	▨	▨	🟩	▲	▲	▲	🟩
			>=30	▨	▨	▨	▨	▨	🟩	▨	▨	▼	▼	▨	▨	▨	▨	▨	▨
			>=40	🟪	🟪	▨	▨	▨	▨	▨	▼	▼	▼	▼	▼	▼	▨	🟪	▼
	3 hr Accumulated Precipitation	1	>=0.25	▨	▼	🟪	▨	▨	▼	▨	🟪	▲	▼	▼	▼	▨	▨	▨	▨
			>=0.5	▨	▨	▨	🟪	▼	▼	▨	🟪	🟩	▲	🟩	▨	🟪	▼	▨	▨
			>=1.0	▨	▨	▨	🟪	🟪	▼	▨	▨	▨	🟩	▨	🟪	▼	▼	🟪	🟪
RMSE	Temperature	1	sfc	🟪	▨	▨	▨	▨	▨	▨	▨	▨	▲	▲	▨	▼	▨	▨	▨
	Dew Point	1	sfc	▨	▨	▨	▨	▨	▨	▨	▨	▼	▨	▨	▨	▨	▨	▨	▨
	Wind Speed	1	sfc	▼	🟪	🟪	▨	🟪	🟪	🟪	🟪	▼	▼	▨	▨	▲	▲	🟩	▨
ME	Temperature	1	sfc	▼	▲	▼	▼	▼	▼	▼	▼	▼	▼	▼	▨	▼	▼	▲	▼
	Dew Point	1	sfc	▨	▨	▨	▨	▲	▲	▼	▼	▼	🟪	▼	▼	▨	🟩	▲	▨
	Wind Speed	1	sfc	▼	▼	🟪	🟪	▼	▼	▼	▼	▼	▼	▼	▨	▨	▨	▨	▨

▲	NSSLfv3 is better than HRRRv3_cluegrid at the 99% significance level
🟩	NSSLfv3 is better than HRRRv3_cluegrid at the 95% significance level
▨	No statistically significant difference between NSSLfv3 and HRRRv3_cluegrid
🟪	NSSLfv3 is worse than HRRRv3_cluegrid at the 95% significance level
▼	NSSLfv3 is worse than HRRRv3_cluegrid at the 99% significance level
🟦	Not statistically relevant

Figure 33 The hourly scorecard comparing the NSSL-FV3 SAR and the HRRRv3. Purple (green) colors indicate that the HRRRv3 is performing better (worse) than the NSSL-FV3 SAR. Solid shading of boxes indicates a difference that is significant at the 95% level, while the arrows indicate statistical significance at the 99% level. Grey shaded boxes indicate no statistically significant difference between the NSSL-FV3 SAR and the HRRRv3. The first four columns indicate (from left to right) the metric, field, gridpoint-based neighborhood, and threshold being compared.

The NSSL-FV3 SAR/HRRRv3 full-period scorecard also shows mixed results (Fig. 34). The NSSL-FV3 SAR scores higher than the HRRRv3 for reliability across both the daily domain and the CONUS for four of the five percentile-based UH thresholds used to generate surrogate severe fields. However, the 95th percentile field (indicated on the scorecard as 95%) use thresholds most similar to what is looked at subjectively in the SFE (~50-75 m^2s^{-2}), and so should be weighed more heavily than the lower percentiles. For this higher percentile, the HRRRv3 performs better than the NSSL-FV3 SAR field at all forecast percentages. Accumulated precipitation scores only showed statistically significant differences for the bias across the CONUS, where the HRRRv3 scored better than the NSSL-FV3 SAR.



METViewer Surrogate Severe NSSLfv3/HRRRv3 CAM Scorecard

for NSSLfv3 and HRRRv3_cluegrid

2019-04-29 00:00:00 - 2019-05-31 00:00:00

			Daily Domain	CONUS
			Daily	
Reliability	75%		▲	▲
	80%		▲	▲
	85%		▲	▲
	90%		▲	▲
	95%		▨	▲
CSI	75%	>=0.02	▨	▲
		>=0.05	▨	▲
		>=0.10	▨	▲
		>=0.15	▨	▲
		>=0.30	▨	▲
		>=0.45	▨	▲
		>=0.60	▨	▨
	85%	>=0.02	▨	▨
		>=0.05	▨	▨
		>=0.10	▨	▨
		>=0.15	▨	▨
		>=0.30	▨	▨
		>=0.45	▨	▨
		>=0.60	▨	🟩
	95%	>=0.02	▼	▼
		>=0.05	▼	▼
		>=0.10	▼	▼
		>=0.15	▼	▼
		>=0.30	▼	▼
		>=0.45	▼	▼
		>=0.60	🟪	🟪
	24 hr Accumulated Precipitation	>=0.25	▨	▨
		>=0.5	▨	▨
		>=1.0	▨	▨
Bias	24 hr Accumulated Precipitation	>=0.25	▨	▼
		>=0.5	▨	▼
		>=1.0	▨	▨
FSS	24 hr Accumulated Precipitation	>=0.25	▨	▨
		>=0.5	▨	▨
		>=1.0	▨	▨

▲	NSSLfv3 is better than HRRRv3_cluegrid at the 99% significance level
🟩	NSSLfv3 is better than HRRRv3_cluegrid at the 95% significance level
▨	No statistically significant difference between NSSLfv3 and HRRRv3_cluegrid
🟪	NSSLfv3 is worse than HRRRv3_cluegrid at the 95% significance level
▼	NSSLfv3 is worse than HRRRv3_cluegrid at the 99% significance level
🟦	Not statistically relevant

Figure 34 The full period scorecard comparing the NSSL-FV3 SAR and the HRRRv3. Purple (green) colors indicate that the HRRRv3 is performing better (worse) than the NSSL-FV3 SAR. Solid shading of boxes indicates a difference that is significant at the 95% level, while the arrows indicate statistical significance at the 99% level. Grey shaded boxes indicate no statistically significant difference between the NSSL-FV3 SAR and the HRRRv3. The first three columns indicate (from left to right) the metric, field, and threshold being compared. For the surrogate severe fields, the percentage indicates the percentile of UH from climatology used to generate the surrogate severe fields, and the threshold is the forecast percentage.

d) Model Evaluations – Severe Hazards Desk (credit: I. Jirak)

1) MESOSCALE ANALYSES

For the first time in the SFEs, formal comparisons of different mesoscale analysis systems were conducted. The analyses examined were the SPC surface objective analysis (sfcOA) based on the operational RAP, the three-dimensional Real-Time Mesoscale Analysis (3D-RTMA) upscaled to the sfcOA 40-km grid, the 3D-RTMA on its native 3-km grid, the 15-km High Resolution Rapid Refresh Ensemble (HRRRE) mean, and the 3-km Warn-on Forecast System (WoFS) ensemble mean. Web-based comparison displays (e.g., Fig. 35) of 2-m temperature, 2-m dewpoint, and CAPE fields were used for the evaluation, with color-coded dots used to depict local differences between observations and the analysis fields.

A direct comparison between the SPC sfcOA and the upscaled 40-km 3D-RTMA was conducted as part of the next-day afternoon evaluations during the SFE. The evaluation was focused over a mesoscale domain with the greatest potential for severe weather across the CONUS from 18-03Z. Three fields were primarily examined during this evaluation: 2-m temperature, 2-m dewpoint, and MLCAPE. The MLCAPE field was calculated identically using SPC mesoanalysis post-processing code for mixed-layer parcel characteristics within the lowest 100 mb above ground level. During the five-week SFE, the quality of these 3D-RTMA fields were typically rated by participants as “about the same” as or “slightly better” than the SPC sfcOA fields (Fig. 32). This was a very encouraging and positive result, as the SPC sfcOA (https://www.spc.noaa.gov/exper/mesoanalysis/) is widely used operationally on the web for real-time situational awareness by NWS forecasters.

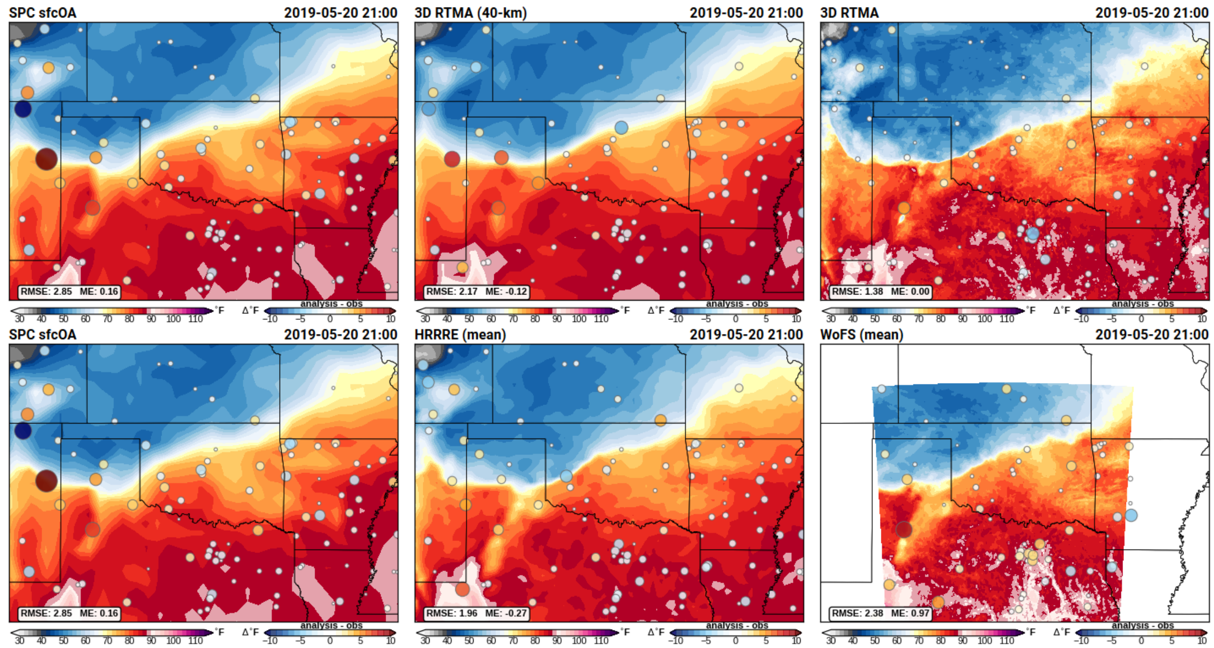


Figure 35 Example of multi-panel comparison webpage for the mesoscale analysis evaluation during the 2019 SFE. The 2-m T analysis with ASOS observation differences (circles; sized and shaded by magnitude, where cool shades indicate analysis is cooler than obs) for SPC sfcOA (upper left), upscaled 40-km 3D-RTMA (upper middle), native 3-km 3D-RTMA (upper right), SPC sfcOA (repeated; lower left), 15-km HRRRE mean (lower middle), and WoFS mean (lower right) at 2100 UTC on 20 May 2019.

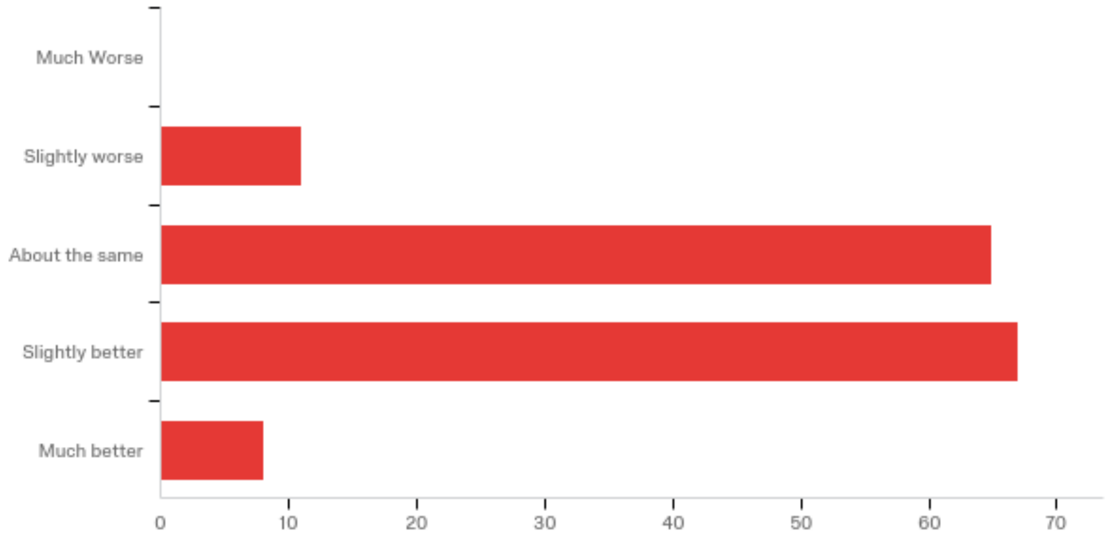


Figure 36 Distribution of subjective survey responses (counts) from participants during the five-week 2019 SFE to the following question: "Provide your overall subjective impression of the quality of the 3D-RTMA (40-km) analyses compared to the SPC sfcOA:"

Another direct comparison was made during the next-day evaluations between the upscaled 40-km 3D-RTMA and the native resolution 3-km 3D-RTMA. There are many reasons for performing this type of comparison focused on analysis resolution that involve aspects of legacy, familiarity, accuracy, and dissemination/display considerations. When focused on comparing instability fields, the participants slightly favored the native 3-km 3D-RTMA over the coarsened 40-km 3D-RTMA, though the individual perspective (e.g., WFO or NCEP) did seem to influence the preference (Fig. 37).

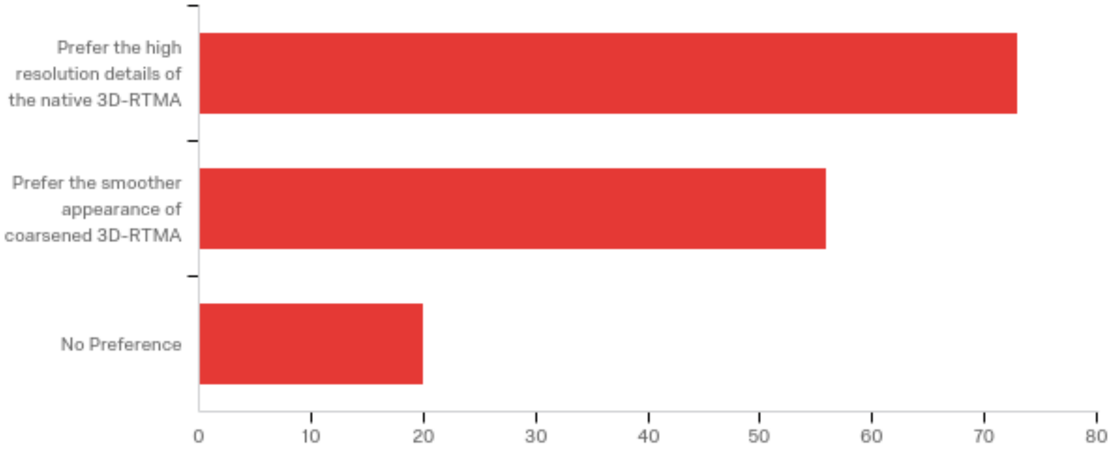


Figure 37 Distribution of subjective survey responses (counts) from participants during the five-week 2019 SFE to the following question: "Provide your overall subjective impression of the native resolution 3D-RTMA (3-km) compared to the coarsened 3D-RTMA (40-km), especially with regard to instability fields:"

Finally, for all of the available analyses, including ensemble-based analyses, a comparison of the 2-m T/Td and CAPE fields was conducted. Participants were instructed to indicate which analysis system provided the overall highest-quality analysis for that particular day (i.e., 18-03Z) and mesoscale domain. The native resolution 3-km 3D-RTMA was the overwhelming favorite and was rated the highest-quality analysis more than the other analysis systems combined (Fig. 38). Again, this is a very promising result for the 3D-RTMA, especially considering this was the first time the output was examined and scrutinized systematically in real-time.

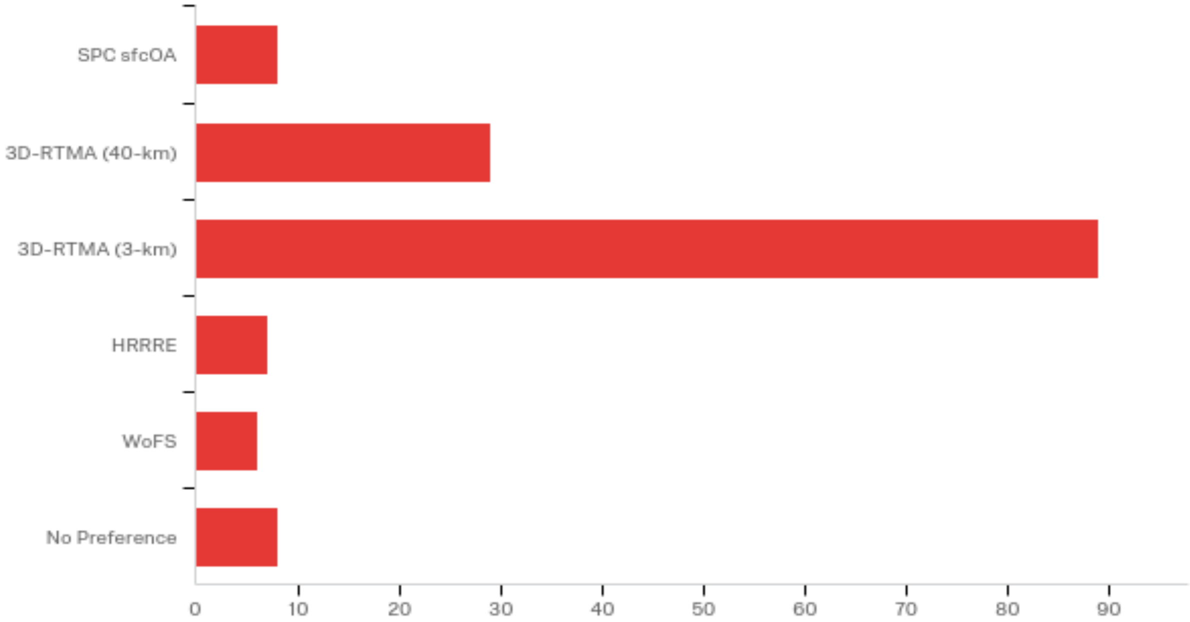


Figure 38 Distribution of subjective survey responses (counts) from participants during the five-week 2019 SFE to the following question: "For all of the available analyses, including ensemble-based analyses, provide your overall subjective impression of the highest-quality analysis:"

2) STORM-SCALE ANALYSES

Following the evaluation of mesoscale analysis systems, a brief, exploratory look at storm-scale analyses was performed for the 3D-RTMA on its native 3-km grid and the 3-km Warn-on Forecast System (WoFS) ensemble probability-matched mean. Web-based comparison displays (e.g., Fig. 39) of composite reflectivity, UH and updraft speed were examined during the evaluation.

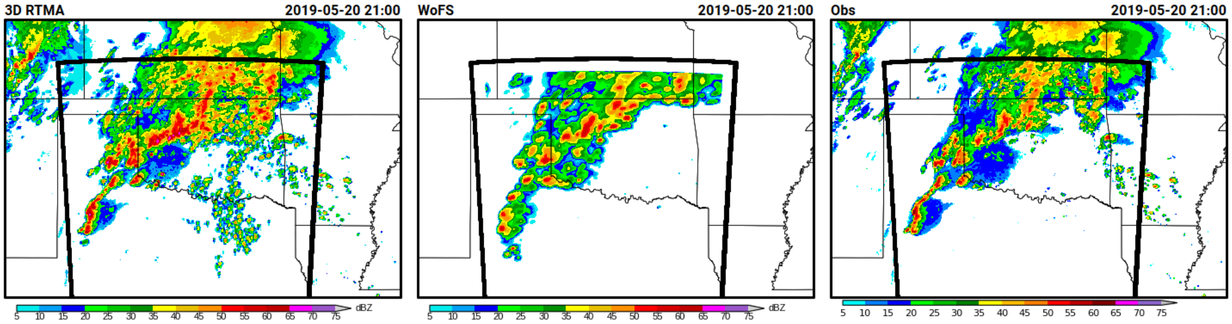


Figure 39 Example of multi-panel comparison webpage for the storm-scale analysis evaluation during the 2019 SFE. The composite reflectivity analysis for the native 3-km 3D-RTMA (left), WoFS probability-matched mean (middle), and radar observations (right) at 2100 UTC on 20 May 2019.

The primary motivation for exploring and evaluating storm-scale analysis fields was to get developers and forecasters thinking about using analysis systems to synthesize information for forecasters from a situational awareness perspective regarding trends in storm intensity and severity (e.g., updraft strength and rotation). While there is much room for improvement in this area regarding the analysis systems examined in the 2019 SFE, most participants found the storm-scale analysis fields to at least be somewhat useful as a situational awareness tool regarding storm intensity and severity (Fig. 40).

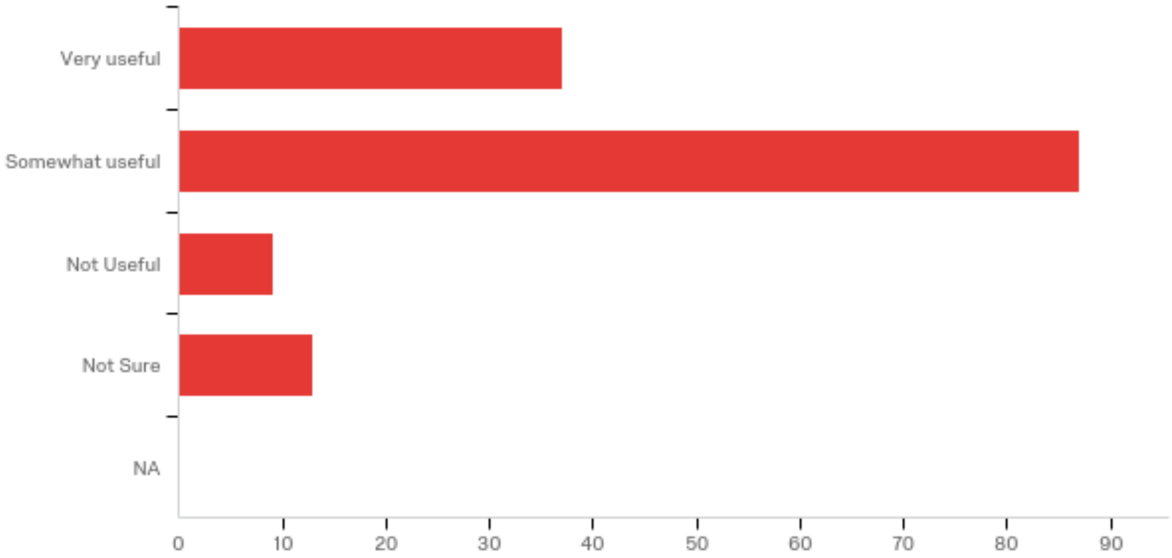


Figure 40 Distribution of subjective survey responses (counts) from participants during the five-week 2019 SFE to the following question: "Provide your overall subjective impression of the usefulness of storm-scale analysis fields for situational awareness of storm intensity and severity:"

3) CLUE: 0000 UTC CAM ENSEMBLES – WRF-ARW

During the 2019 SFE, several real-time 0000 UTC WRF-ARW-based CAM ensemble forecasts were examined and compared to the 0000 UTC HREFv2.1 forecasts. This evaluation was focused over a mesoscale area of interest with the greatest potential for severe weather over the CONUS during the convective day (i.e., 1200-1200 UTC). The output field most commonly examined during this severe weather evaluation was the 2-5 km AGL hourly maximum UH. The ensemble maximum UH and neighborhood UH probabilities (>75 m^2/s^2 & > 150 m^2/s^2) were displayed along with preliminary local storm reports (Fig. 41), and participants rated the forecasts (on a scale of 1-10) based on the quality of guidance provided to a severe weather forecaster.

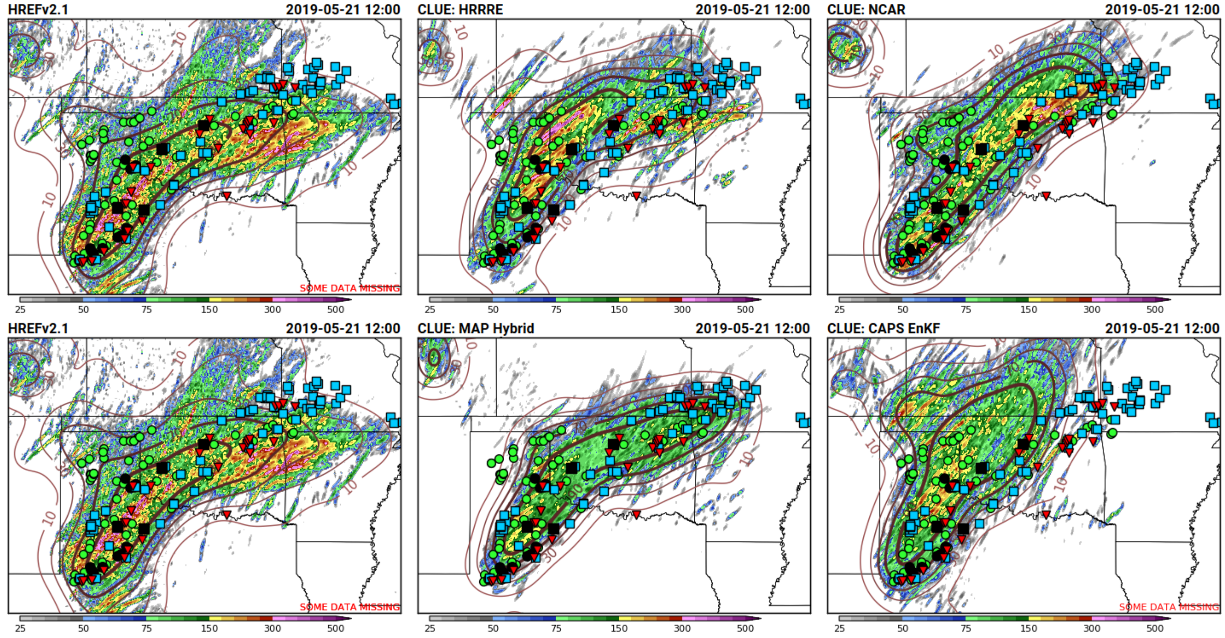


Figure 41 Example of multi-panel comparison webpage for the 0000 UTC CAM ensemble evaluation during the 2019 SFE. The ensemble maximum UH (shaded) and neighborhood probability of UH>75 m^2/s^2 (contoured) is displayed for HREFv2.1 (upper left), HRRRE (upper middle), NCAR (upper right), HREFv2.1 (repeated; lower left), map-hybrid (lower middle), and CAPS EnKF (lower right) for 20 May 2019. Preliminary severe storm reports are also overlaid (wind - blue squares, hail - green circles, and tornado - red upside-down triangles).

For the 0000 UTC-initialized forecasts, the HRRRE, NCAR, and map-hybrid ensembles all had a similar distribution of subjective ratings during the five-week SFE (Fig. 42). The NCAR ensemble had a slightly higher mean rating than the HRRRE, but the HRRRE had higher mean and median ratings than the map-hybrid ensemble. Overall, however, the single-core WRF-ARW ensembles had notably lower subjective ratings during the SFE than the HREF (v2.1, which adds two HRRR members to the operational HREFv2; Fig. 42). These results continue to highlight the challenge for a single-core, single-physics CAM ensemble to match the skill/utility of a multi-model, multi-physics CAM ensemble (i.e., HREF) for severe weather forecasting.

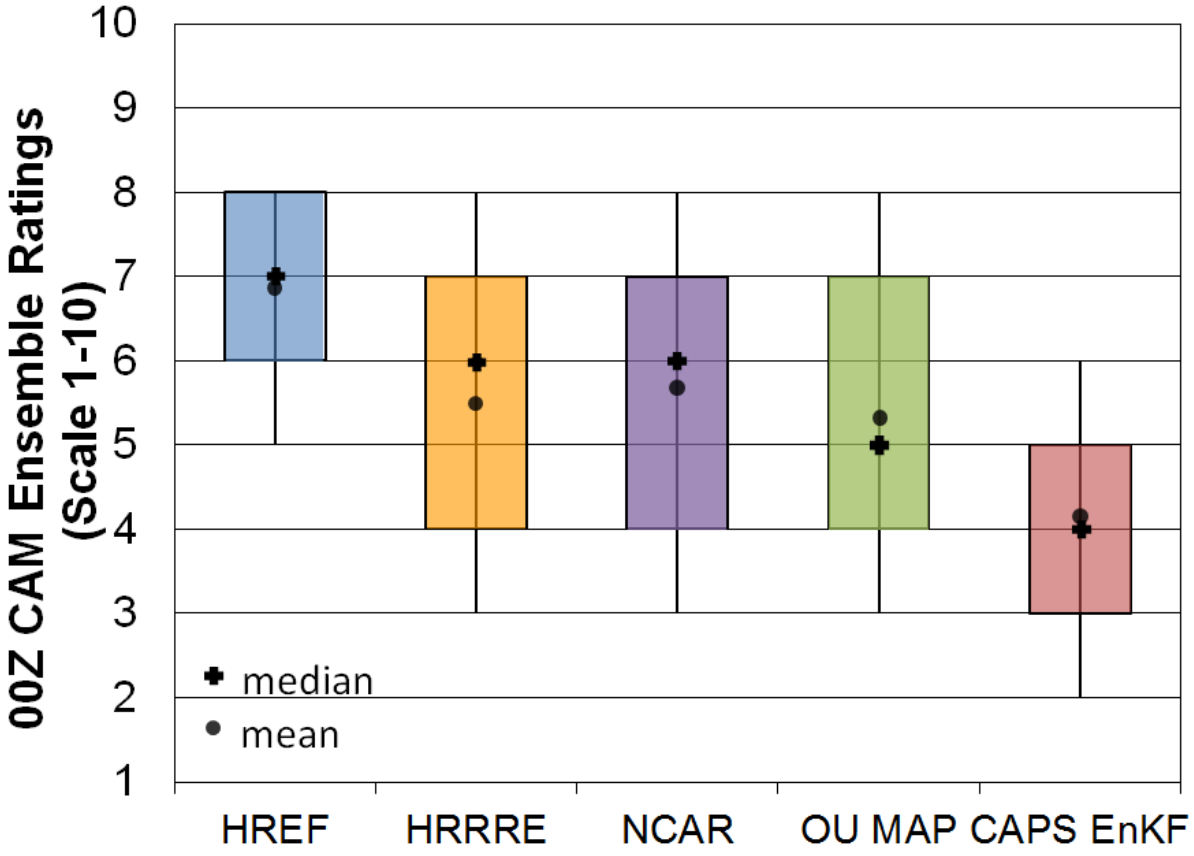


Figure 42 Distributions of subjective ratings (1-10) by SFE participants of hourly maximum fields for severe weather forecasting over a mesoscale area of interest for the forecast hours 13-36 for various 0000 UTC CLUE CAM ensembles (HRRRE - orange; NCAR - purple; map-hybrid - green; CAPS EnKF - red) compared to the HREF (blue).

4) CLUE: 0000 UTC CAM ENSEMBLES – FV3 and UM

Additional 0000-UTC initialized CAM ensembles with different dynamic cores were run over the CONUS for the first time to compare with the operational HREF. CAPS ran an FV3-based CAM ensemble and the UK Met Office ran a UM-based CAM ensemble. Similar to the other 0000 UTC CAM ensemble evaluation, this evaluation was focused over a mesoscale area of interest with the greatest potential for severe weather over the CONUS during the convective day (i.e., 1200-1200 UTC). The output field most commonly examined during this severe weather evaluation was the 2-5 km AGL hourly maximum UH. However, owing to the differences in model UH climatology for the different dynamic cores, a percentile-based exceedance probability (>99.85th percentile) was used instead of a fixed-threshold exceedance probability (e.g., >75 m^2/s^2) for evaluating the quality of guidance for severe weather forecasting.

Overall, the UM-based CAM ensemble performed well over the CONUS for severe weather applications (Fig. 43). It was a comparable performer to the HRRRE and NCAR ensemble (c.f. Fig. 42), but did not perform quite as well overall as the HREF. The FV3-based CAM ensemble had lower mean/median ratings (~5 out of 10) than most of the other 0000-UTC initialized CAM ensembles, suggesting that more development work is needed to improve the performance of an FV3-based CAM ensemble for severe weather applications.

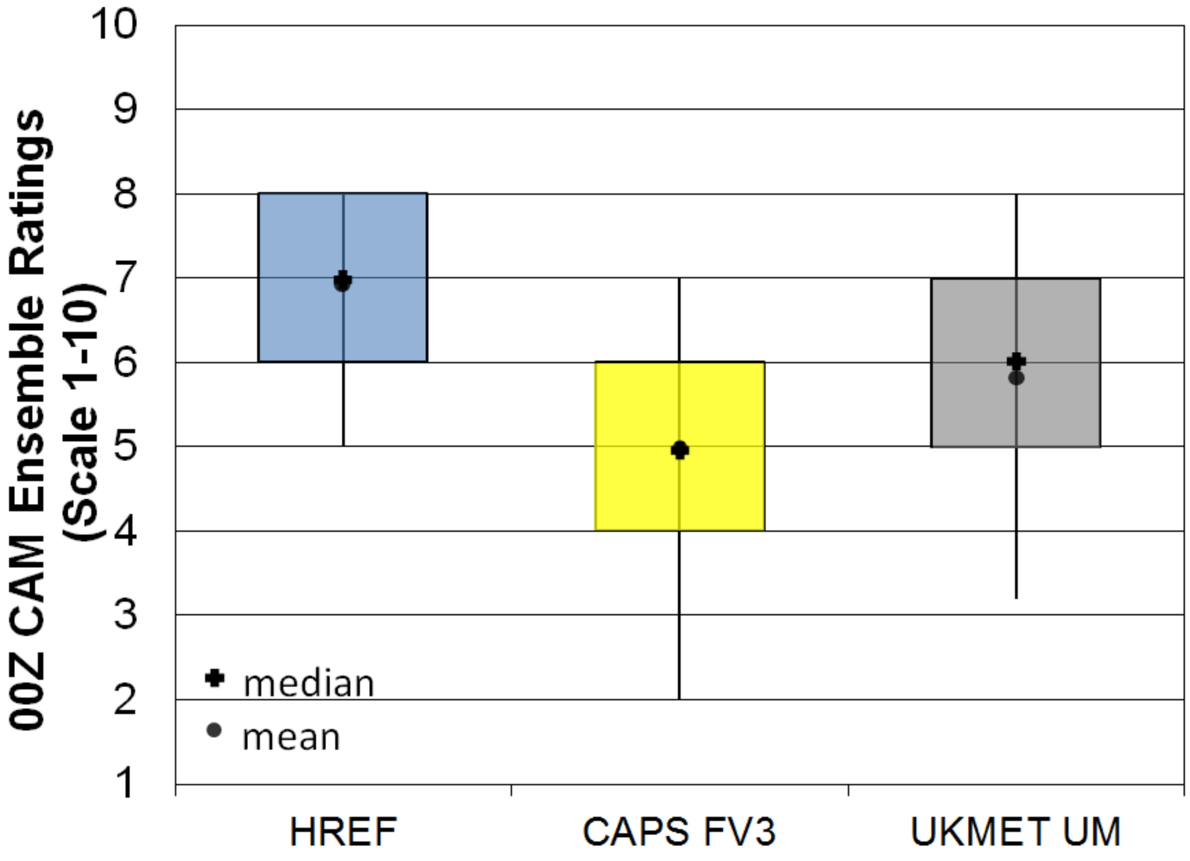


Figure 43 Distributions of subjective ratings (1-10) by SFE participants of hourly maximum fields for severe weather forecasting over a mesoscale area of interest for the forecast hours 13-36 for 0000 UTC CAPS FV3 (yellow) and UKMET UM (gray) compared to the HREF (blue).

5) 0000 UTC HREF CONFIGURATIONS

To move toward a unified forecasting system in the NWS, rapid development has been taking place to generate a SAR-FV3 CAM run over the CONUS. The most logical initial operational implementation of the SAR-FV3 would be as a member of the HREF, so a couple of HREF configurations that included the EMC SAR-FV3 were explored. Specifically, a nine-member 0000 UTC HREF configuration that added the 0000 UTC EMC SAR-FV3 to the HREFv2.1 (i.e., 10-member HREF with two HRRR members), but removed the two HRW-NMMB members (i.e., the worst-performing members of the HREF). A third configuration with the SAR-FV3 was also explored that did not include the two HRRR members (i.e., seven total members).

Overall, replacing the HRW-NMMB members with a SAR-FV3 member did not have a large impact on the subjective performance of the HREF for severe weather forecasting. The distribution of subjective ratings from SFE participants were nearly identical for HREFv2.1 and the configuration with the SAR-FV3 (Fig. 44). A larger impact was noted when the HRRR members were not included in the HREF, as this configuration resulted in lower subjective ratings from the participants (Fig. 44). These results support the EMC plans of moving forward with an HREF configuration that includes HRRR members and replaces the HRW-NMMB with the SAR-FV3.

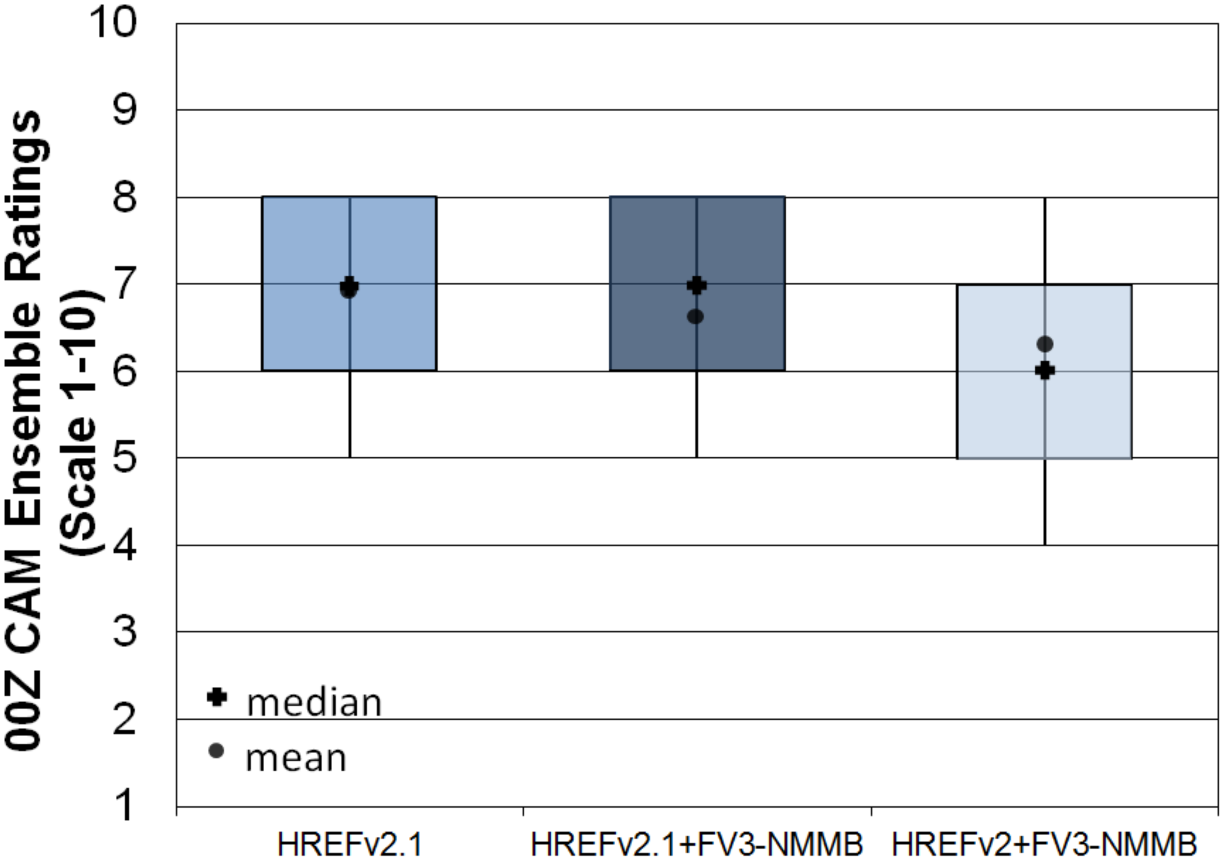


Figure 44 Distributions of subjective ratings (1-10) by SFE participants of hourly maximum fields for severe weather forecasting over a mesoscale area of interest for the forecast hours 13-36 for the 9-member HREFv2.1 configuration with the SAR-FV3 replacing the HRW NMMB (dark blue) and the 7-member HREFv2 (no HRRR members) with the SAR-FV3 replacing the HRW NMMB (light blue) compared to the 0000 UTC HREF (blue).

6) STOCHASTIC PHYSICS

One approach to increase the ensemble spread and improve probabilistic forecasts from a single-core, single-physics ensemble is to use a stochastic-physics approach. A controlled experiment to test this hypothesis was performed during the 2019 SFE using two versions of the HRRRE: one with stochastic parameter perturbations applied and one without stochastic physics (i.e., pure single-physics ensemble). For severe weather applications, the HRRRE with stochastic physics did not provide improved probabilistic forecasts according to subjective ratings from the participants (Fig. 45). In fact, the version of HRRRE without stochastic physics tended to have fewer lower-rated forecasts than the HRRRE with stochastic physics, as evidenced by higher mean, 25th, and 10th percentile ratings. As has been identified in previous SFEs, stochastic physics (in the current state-of-science) do not typically provide a practical improvement to forecasts over a single-physics approach for severe convective weather. Subjectively, the impact of stochastic physics appeared to lower the storm-attribute probabilities (without noticeably changing the spatial ensemble envelope) by removing/weakening storms.

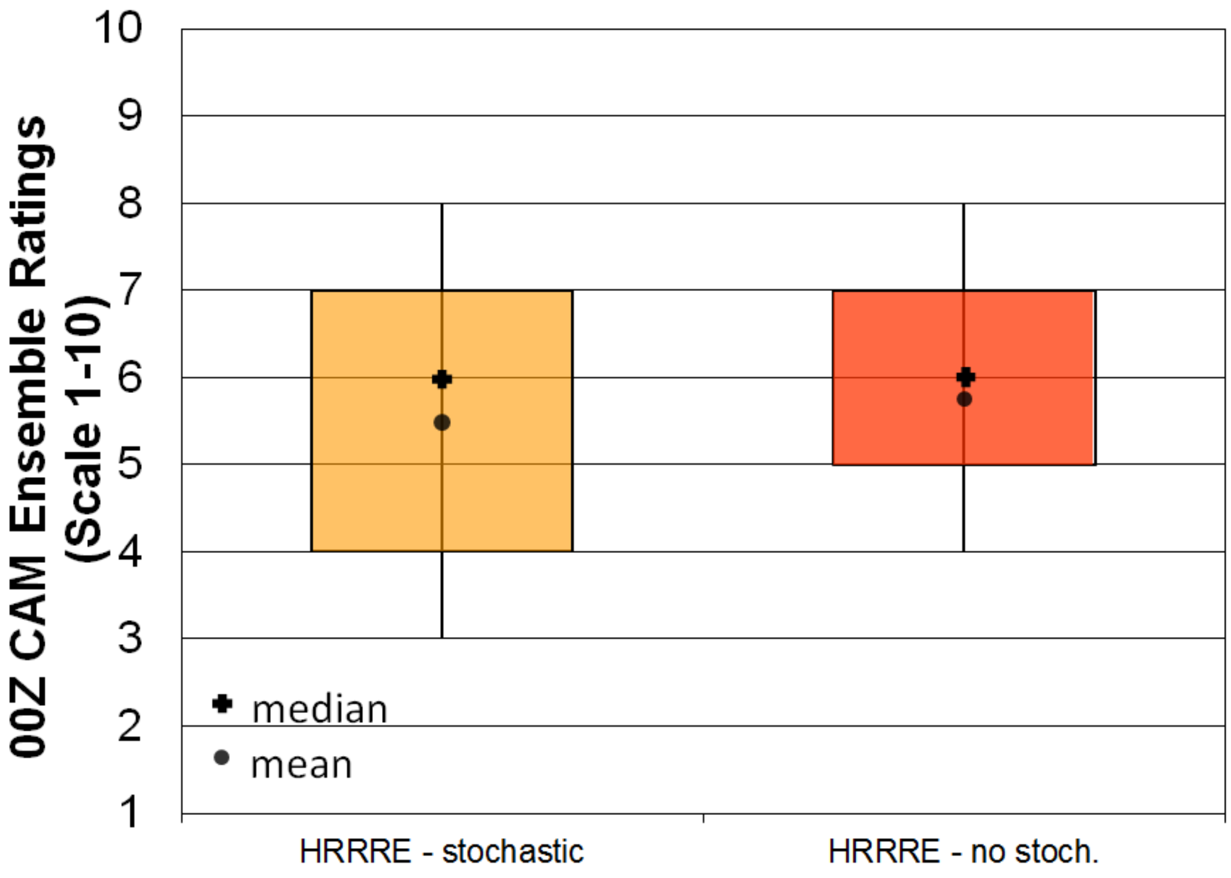


Figure 45 Distributions of subjective ratings (1-10) by SFE participants of hourly maximum fields for severe weather forecasting over a mesoscale area of interest for the forecast hours 13-36 for the HRRRE with stochastic physics (left; light orange) and the HRRRE without stochastic physics (right; dark orange).

7) UM PHYSICS

The primary CAM ensemble from the UK Met Office was the single-physics ensemble (discussed in subsection 4 earlier), which is configured with a physics suite tuned for the mid-latitudes. To investigate the impact of multiple physics schemes on ensemble performance, a different physics suite (tuned for the tropics) was utilized in some members of the mixed-physics ensemble. For severe weather applications, the UM mixed-physics ensemble was rated subjectively lower than the UM single-physics ensemble by SFE participants (Fig. 46). Generally, the UM mixed-physics ensemble resulted in lower probabilities and more weakly rotating storms, which often led to a worse forecast for severe weather. This is not necessarily an expected result, and may be indicative of a physics suite (i.e., tropical suite) not well suited for severe weather applications over the CONUS.

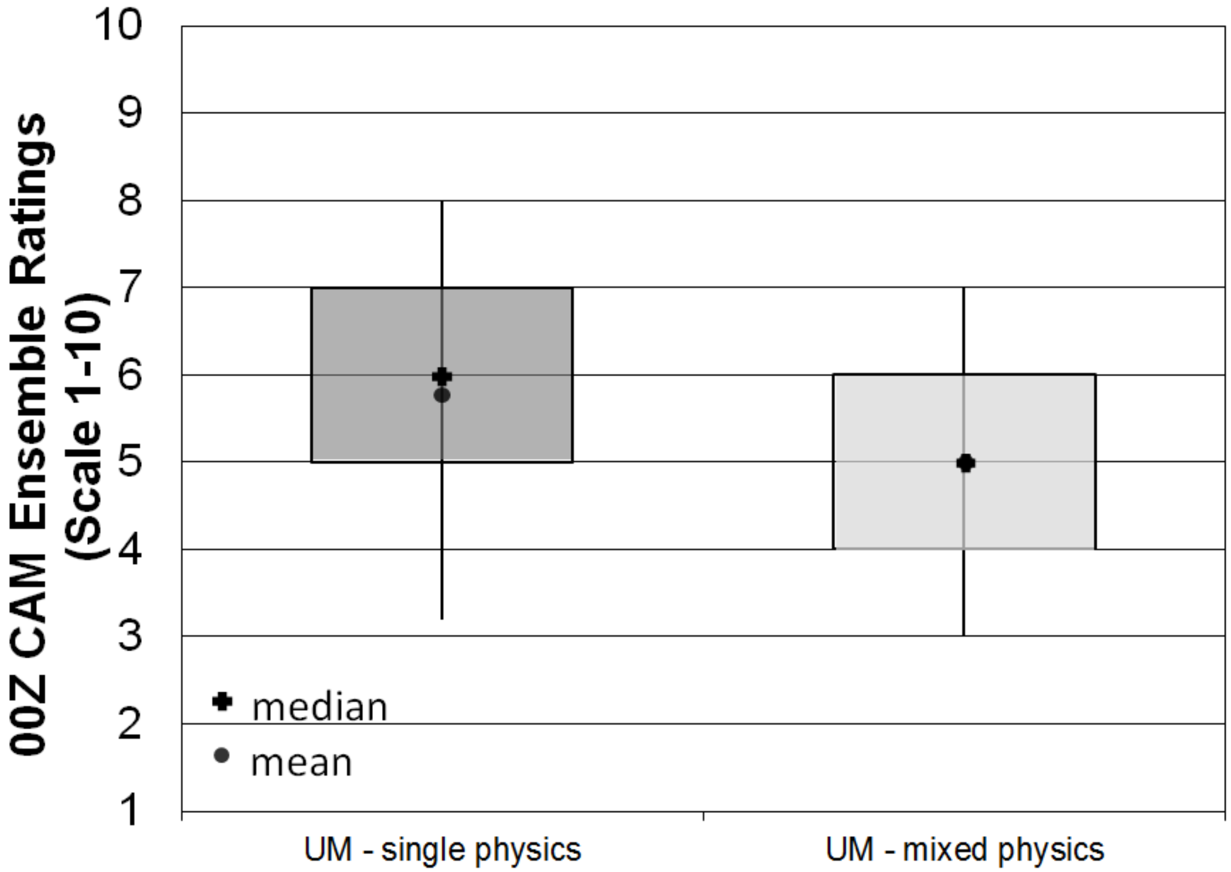


Figure 46 Distributions of subjective ratings (1-10) by SFE participants of hourly maximum fields for severe weather forecasting over a mesoscale area of interest for the forecast hours 13-36 for the UM single-physics CAM ensemble (left; dark gray) and the UM mixed-physics CAM ensemble (right; light gray).

8) IC PERTURBATIONS

The impact of initial condition (IC) perturbations was explored in a controlled experiment from the OU MAP group. Their primary ensemble (map-hybrid; discussed in subsection 3 above) uses multi-scale perturbations in the ICs of the ensemble forecast members while the experimental ensemble was centered on the hybrid mean analysis, but utilized GEFS IC perturbations for the forecast members (map-Icpert). The hypothesis was that the multi-scale perturbations in the hybrid ensemble would result in better forecasts than the ensemble with IC perturbations from the GEFS (i.e., only large-scale perturbations). In general, the forecasts from these ensembles were very similar during the SFE and provided very similar guidance for severe weather forecasting. The subjective rating distributions from the SFE participants were nearly identical during the SFE (Fig. 47).

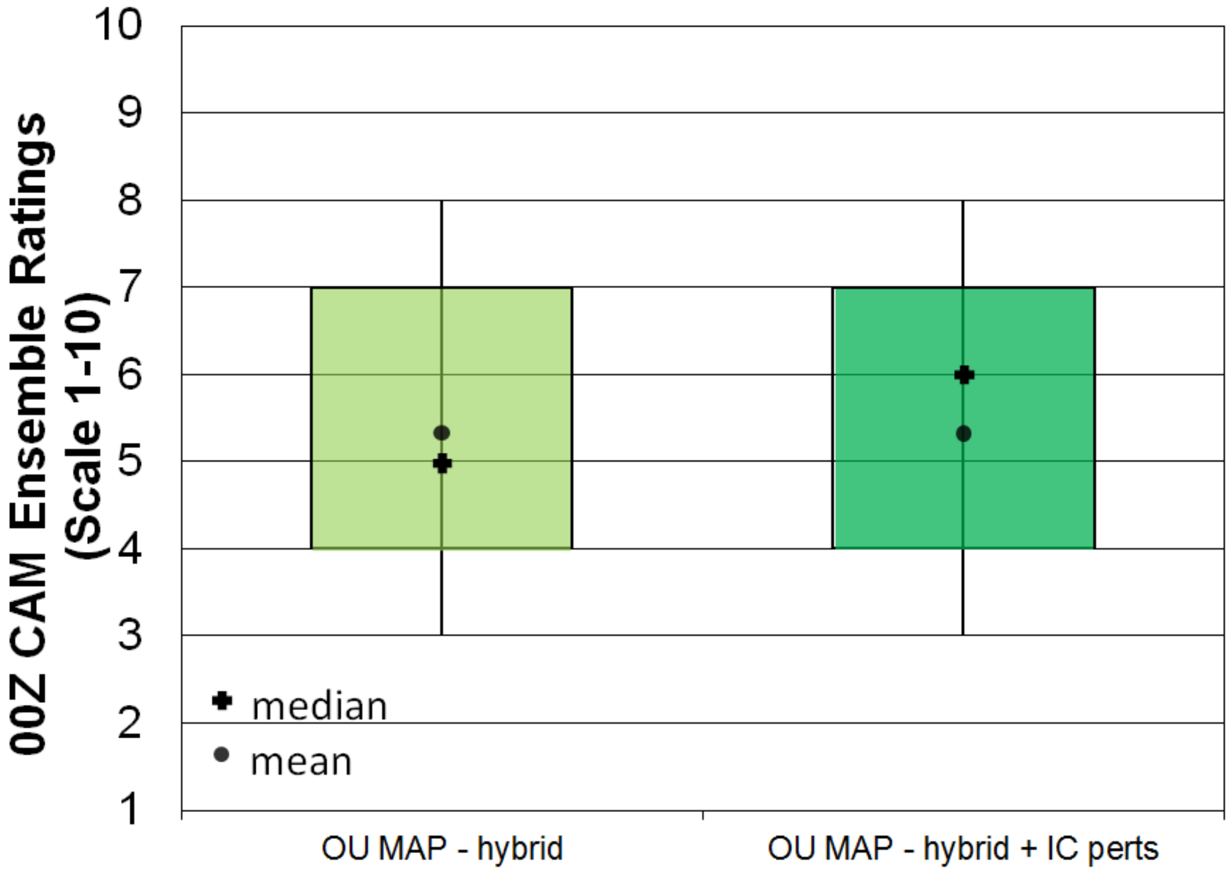


Figure 47 Distributions of subjective ratings (1-10) by SFE participants of hourly maximum fields for severe weather forecasting over a mesoscale area of interest for the forecast hours 13-36 for the map-hybrid ensemble (left; light green) and the map-ICpert ensemble with GEFS IC perturbations (right; green).

9) 1200 UTC CAM ENSEMBLES

For the 1200 UTC-initialized forecasts, the HRRRE was also compared to time-lagged (TL) ensembles generated from the operational High-Resolution Rapid Refresh (HRRRv3) during the SFE. Two HRRR-TL ensembles based at 1200 UTC were constructed: 1) HRRR-TL4: consisting of four (4) 1-h time-lagged members (i.e., 12, 11, 10, and 09 UTC runs) and 2) HRRR-TL6, which adds the 6- and 12-h time-lagged members to HRRR-TL4 (i.e., 12, 11, 10, 09, 06, and 00 UTC runs). These 1200 UTC CAM ensembles were evaluated subjectively on 4-hour hourly maximum field (HMF) forecasts (e.g., UH) from 16-03Z for severe weather guidance. The HRRR-TL ensembles fared well in subjective ratings, commonly outperforming the formal ensembles: HRRRE and NCAR ensemble (Fig. 48). Overall, the HREF once again had the highest subjectively rated forecasts of the 12-UTC initialized ensembles.

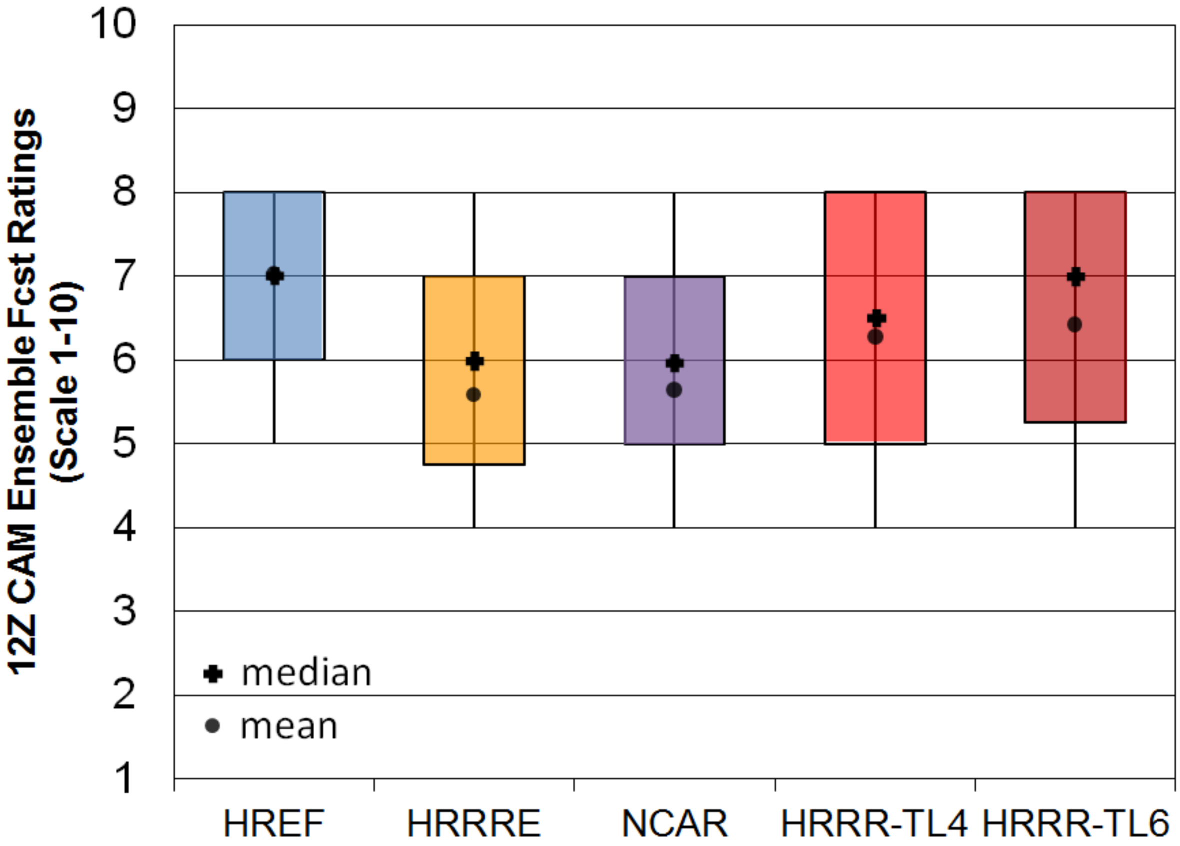


Figure 48 Distributions of subjective ratings (1-10) by SFE participants of hourly maximum fields over a mesoscale area of interest for the forecast hours 1-24 for 1200 UTC CLUE CAM ensembles compared to the HREF and HRRR TL ensembles.

10) HAIL GUIDANCE

Several different approaches for estimating/deriving severe hail probabilities were examined from the 0000 UTC HRRRE during the 2019 SFE. The hail diagnostic fields examined included a microphysics-based approach (Greg Thompson), a machine-learning approach (Nathan Snook and Amanda Burke; Gagne et al. 2017), along with the standard CAM storm-attribute fields [i.e., Max UH (cyclonic; +), Total UH (cyclonic and anticyclonic; +/-), and updraft speed]. These hail proxy forecasts were evaluated and compared daily during the 2019 SFE using the web-based interface shown in Figure 49.

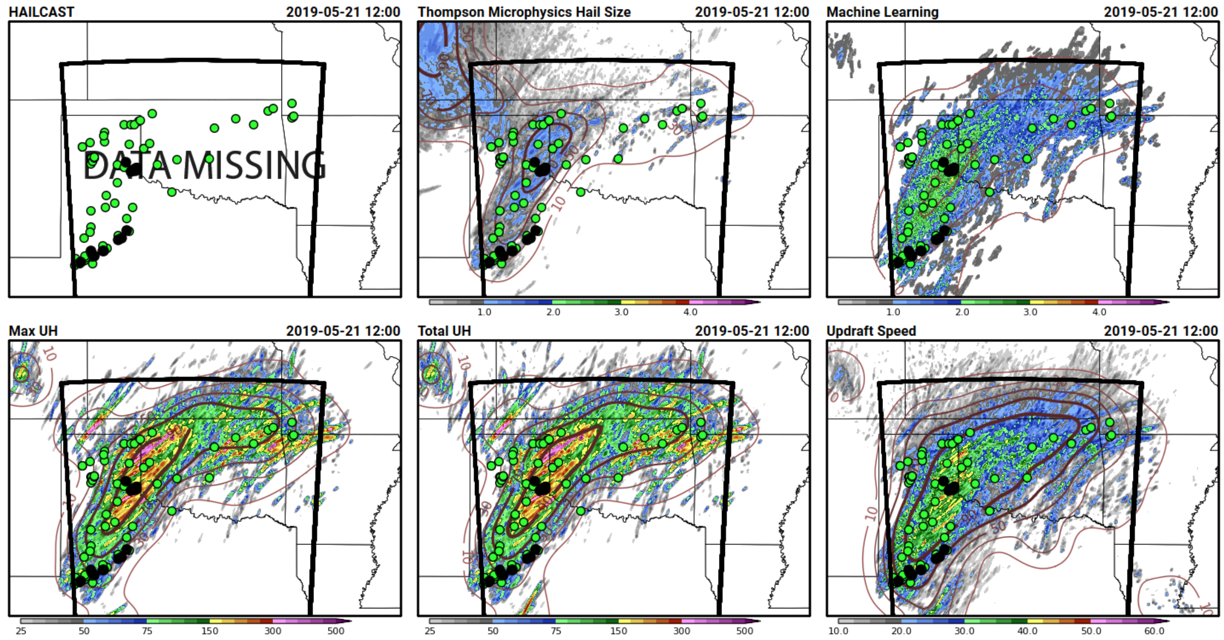


Figure 49 Six-panel comparison plot used to conduct the evaluation of the hail output variables from the HRRRE during the 2019 SFE. The 24-h neighborhood hail probability forecasts exceeding 1 inch valid for 1 June 2018 are shown for HAILCAST (top-left panel - NA), microphysics-based approach (top-middle panel), machine-learning technique (top-right panel), max UH (≥90 m2s-2; bottom-left panel), max/min UH (≥90 m2s-2; bottom-left panel), and updraft speed (≥20 ms-1, bottom-right panel). The observed severe hail reports (≥1 inch; green circles) and significant severe hail reports (≥2 inches; black circles) are overlaid as a reference for subjective verification.

During each afternoon of the 2019 SFE, participants would subjectively rate (on a scale of 1 to 10) the quality of the different hail-proxy forecasts from the HRRRE valid for the previous day. The observed hail reports and radar-derived MESH values were used as the verification sources to help assess the quality of the forecasts. None of the proxies stood out as the best method for extracting severe hail probabilities from the HRRRE (Fig. 50). The total UH (cyclonic plus anti-cyclonic) had a slight advantage in subjective ratings over max UH (cyclonic-only), updraft speed, and machine-learning method (which used these storm-attribute fields as inputs) while the microphysics-based approach had the lowest subjective ratings.

These subjective results provide supporting evidence of the difficulty in adding value over and above standard storm-attribute information from CAM ensembles.

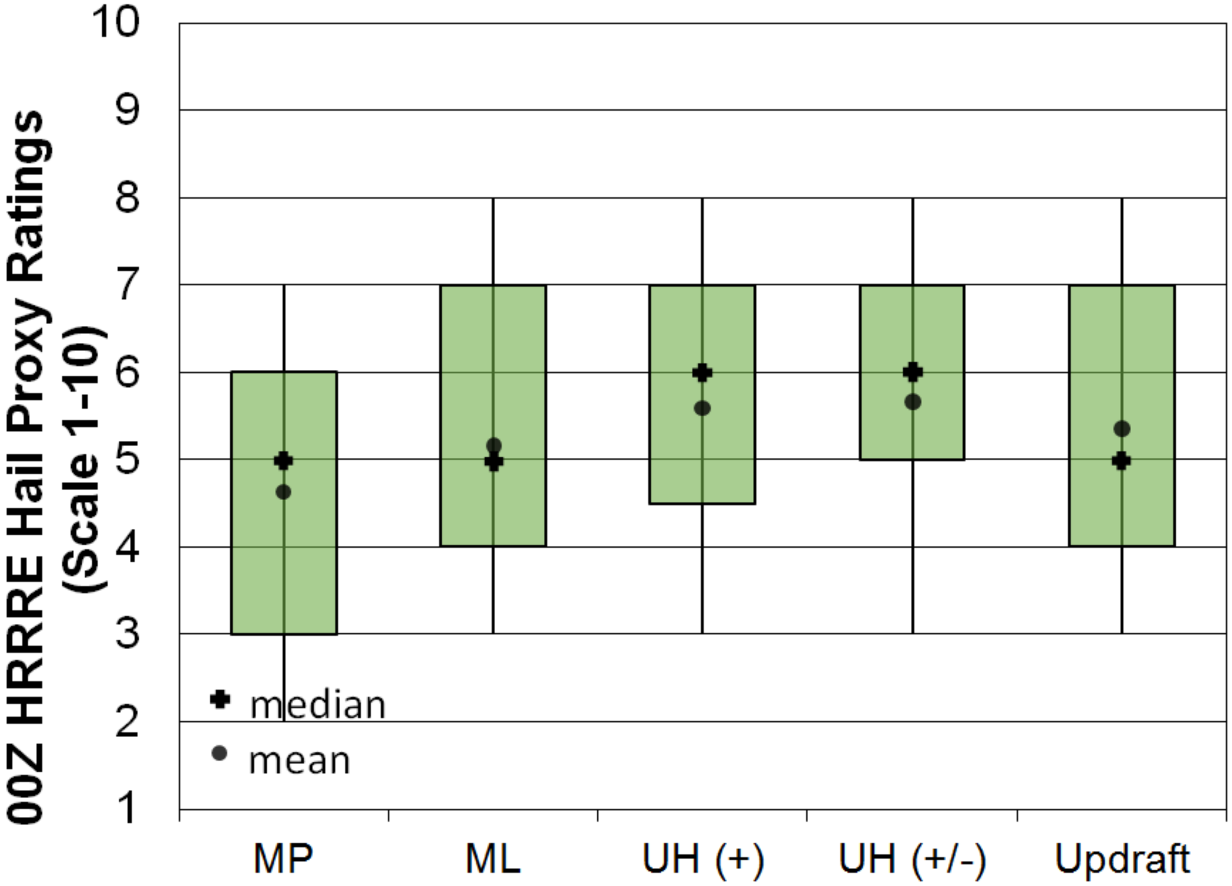


Figure 50 Distributions of subjective ratings (1-10) by SFE participants of hail proxy forecasts over a mesoscale area of interest for the forecast hours 12-36 for 0000 UTC HRRRE for: a) microphysics approach (MP), b) machine-learning technique (ML), c) cyclonic UH (UH+), d) cyclonic and anticyclonic UH (UH+/-), and e) updraft speed.

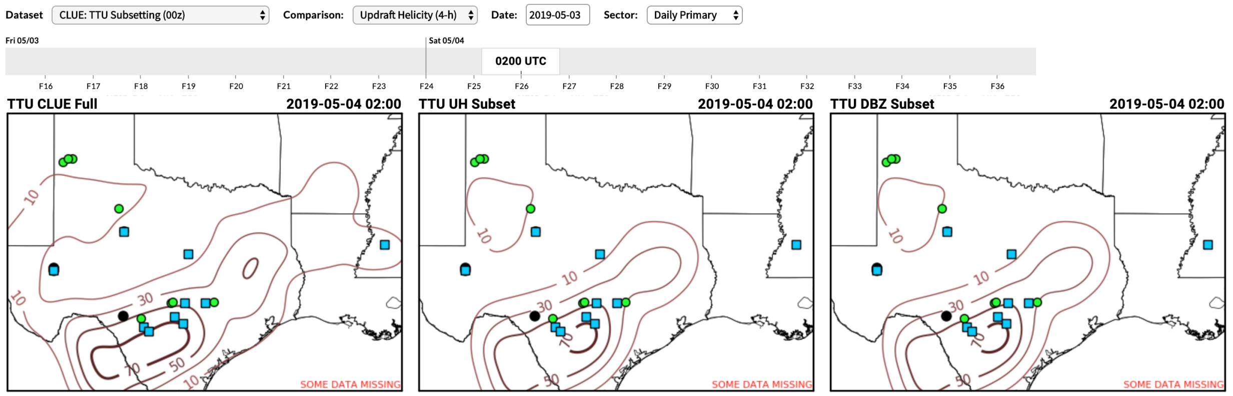
11) TTU ENSEMBLE SENSITIVITY-BASED SUBSETTING PRODUCT (credit: Brian Ancell)

Building off work from the 2018 SFE, a daily evaluation of probabilities was performed comparing a large ensemble and a subset of that ensemble chosen by selecting the members with the smallest errors in sensitive regions (following the ensemble sensitivity-based subsetting technique of Ancell 2016). This strategy employs ensemble sensitivity (Ancell and Hakim 2017) to estimate the regions and variables at early forecast times that have an influence on severe convection later in a forecast, and then selects a smaller number of ensemble members that are most accurate in those sensitive areas at early forecast time using analyses or observations. Theoretically this should improve forecast probability distributions of severe convection well before subsequent ensemble initializations are available. This procedure follows the evaluation at the 2018 HWT at which only the TTU ensemble members composed the full ensemble and the subset, but instead is executed using members from the wider HWT CLUE ensemble. The full ensemble (termed “TTU Full CLUE”) consisted of several ensemble systems within the CLUE, specifically the HRRRE (9 members), two OU MAP ensembles (map-hybrid and map-ICpert; 20 members), the CAPS FV3 ensemble (9 members), the NCAR ensemble (10 members), and the CAPS EnKF (10 members) for a total of 58 members. Individual forecast members (not individual ensemble systems) were chosen from the TTU Full CLUE ensemble to compose the subset for comparison. The TTU Full CLUE contains members that represent different model cores, physics parameterizations, and initial conditions. The reasons for conducting this evaluation using members from the CLUE system are 1) to provide a more consistent framework for evaluation alongside the other HWT comparisons using CLUE, 2) to test the operational capability using several different ensemble systems, and 3) to perform the subsetting procedure in an ensemble with relatively large spread (the 2018 HWT subsetting procedure showed limited results due to relatively small spread in the TTU system alone).

Each day a response function location was chosen collectively with HWT participants through a web-based graphical user interface that identified areas of uncertainty in a 6-hr window in the forecast of Day 1 severe convection within that day's 0000 UTC forecast. These areas of uncertainty were chosen based on inspection of ensemble probabilities from the different ensemble systems within the CLUE as well as the real-time TTU 42-member ensemble system. The Day 1 response function area was selected at a forecast hour between 1800 UTC (the 18-h forecast) and 1200 UTC the following day (the 36-h forecast). Once the response function time and location were chosen, the sensitivity of two independent response functions were automatically calculated: 1) number of grid points exceeding 50 m^2/s^2 2-5km UH, and 2) number of grid points exceeding 40 dBZ lowest-model-level simulated reflectivity. The sensitivities of the two response functions (chosen on the TTU 4-km nested domain over the Midwest and South Plains) were calculated with respect to 300- and 500-hPa temperature, winds, and geopotential height, and 700-hPa temperature on the 12-km TTU CONUS domain all with respect to the 7-hr forecast state (valid 0700 UTC).

Once the ensemble sensitivity fields were calculated, each 7-hr forecast from every TTU Full CLUE ensemble member (originally produced at 3-km grid spacing but interpolated to the 12km TTU domain) was compared with analyses. These analyses were generated from the 1-hr forecast ensemble mean from the 0600 UTC TTU DART EAKF data assimilation cycle on the 12-km grid. The 1-hr forecast at 0700 UTC was used in lieu of the analysis valid at 0600 UTC due to significant imbalance present after the assimilation procedure. The 20 ensemble members from the 0000 UTC TTU Full CLUE forecasts that possessed the smallest sensitivity-weighted errors (chosen using the sum resulting from the projection of the ensemble differences with the analysis onto the ensemble sensitivity field over the greatest 50% of sensitivity magnitudes) were chosen as the ensemble subset.

Probability fields (e.g. neighborhood probability of 2-5km UH exceeding 75 m^2/s^2) were calculated for both the TTU Full CLUE and the sensitivity-based subset for the subsets generated using sensitivity of both the UH and simulated reflectivity response functions (courtesy Brett Roberts, NSSL). These probabilities were compared and evaluated for the 6-hr period for which the response function was valid (as chosen by HWT participants). Figure 51 shows an example of the comparison of probabilities from the TTU Full CLUE and the sensitivity-based subsets from the ensemble forecast initialized at 0000 UTC May 3, 2019. For this and all cases participants were asked to provide their opinions and answer survey questions regarding the relative skill of the full ensemble and the ensemble subset. Most days only the UH-based subsets were used – the simulated reflectivity response function was only evaluated in depth when large UH values were not produced in the TTU Full CLUE ensemble (which a small minority of days). As with other HWT comparisons, the skill of the TTU Full CLUE and the subsets were judged against the location and density of storm reports, MRMS MESH maximum hail size estimates, and NWS warnings. For the case shown in Figure 51, false alarm area was reduced within the subset probabilities, which was a typical characteristic of the cases in which the subset was deemed superior. There were several days where the subset was deemed inferior as well, which seemed to occur when probability maxima were shifted away from storm reports relative to the full ensemble (suggested through comments made by survey participants).



Dataset: CLUE: TTU Subsetting (00z)
Comparison: Updraft Helicity (4-h)
Date: 2019-05-03
Sector: Daily Primary

TTU CLUE Full 2019-05-04 02:00

TTU UH Subset 2019-05-04 02:00

TTU DBZ Subset 2019-05-04 02:00

Figure 51 Example of HWT comparison product showing both the TTU Full CLUE ensemble probabilities (left) against those of the ensemble sensitivity-based subset based on a UH coverage response function (middle) and a simulated reflectivity coverage response function (right). These subsets were generated for a response function valid from 2000 UTC May 3 to 0200 UTC May 4 for the forecast initialized at 0000 UTC May 3 (specific response function area not shown).

Figure 52 shows the survey questions and results over the entire 5-week experiment (97 survey responses). About half (51.5%) of responses indicated that the skill of the full ensemble and subset were the same, while 29.9% felt the subset improved probabilities. 18.6% of responses indicated the full ensemble was better. In terms of the viability as an operational product, 74.2% of participant responses suggested the subsetting procedure could be used as an operational tool, with 25.8% indicating that it could not. These results indicate that while improvements of the subset outweighed degradations, it was perceived that both successes and failures occurred at relatively close frequency. Figure 53, which shows fractions skill scores of the subset and full ensemble against SPC practically perfect probability fields, also suggests day-to-day variability in the success of the subsetting procedure. Interestingly the majority of responses reflect similar skill of the subset and full ensemble, which we speculate results from the full ensemble being relatively skillful in the first place (making it difficult for the subset to provide any improvement). As mentioned above, the largest benefit was the reduction of false alarm area, while the largest disadvantage appeared to be shift in location of probability maxima relative to observed storm reports.

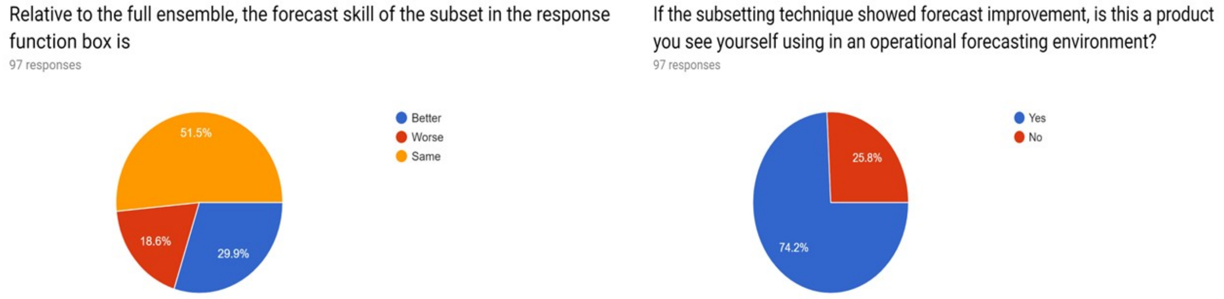


Figure 52 Cumulative survey results and questions valid over the entire 5-week experiment (97 responses).

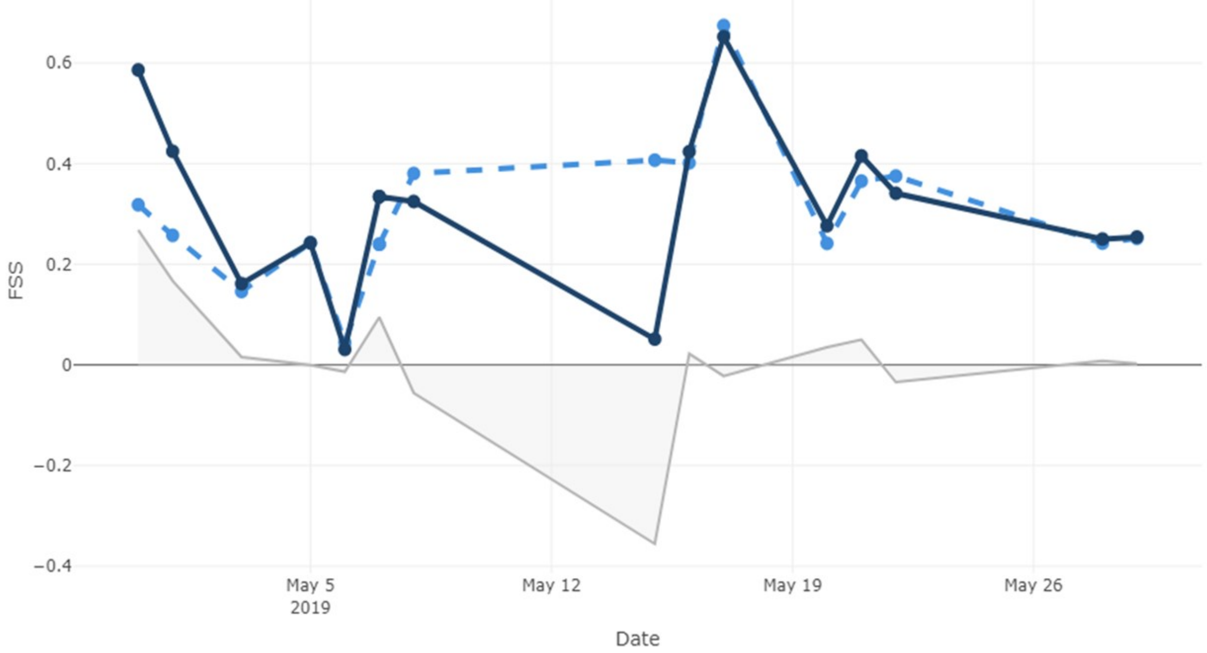


Figure 52 Fractions skill scores (ensemble probability field verified against SPC practically perfect probabilities) for the ensemble subset probability (solid line) based on the UH response function, as well as the TTU Full CLUE ensemble probabilities (dashed line) over the 5-week experiment. The grey line shows the difference between the subset and full ensemble probabilities.

Finally, Figure 53 shows how often specific ensemble members were chosen through the subsetting procedure. Clearly the HRRRE members were chosen most frequently, followed by NCAR members, the CAPS EnKF members, and the OU MAP members. Interestingly the CAPS FV3 members were never chosen in the subset, indicating that the pre-convective environment aloft in sensitive regions always had the largest errors. Whether this indicates a systematic issue with interpolation in the FV3 configuration or if the skill of the FV3 members aloft is indeed poor will be investigated in the future.

Future investigation will focus on understanding the nature of the subset success and failure cases toward improving the success rate of the procedure in general. Further, several assumptions were made in the HWT experimental setup which will need to be further vetted: 1) the sensitivity fields from the TTU ensemble apply generally to the flow, 2) improving forecasts with sensitivity-based subsetting using a UH coverage response function also improves the location and magnitude of UH tracks, and 3) model error with regard to the sensitivity variables used does not play a significantly larger role than initial condition error.

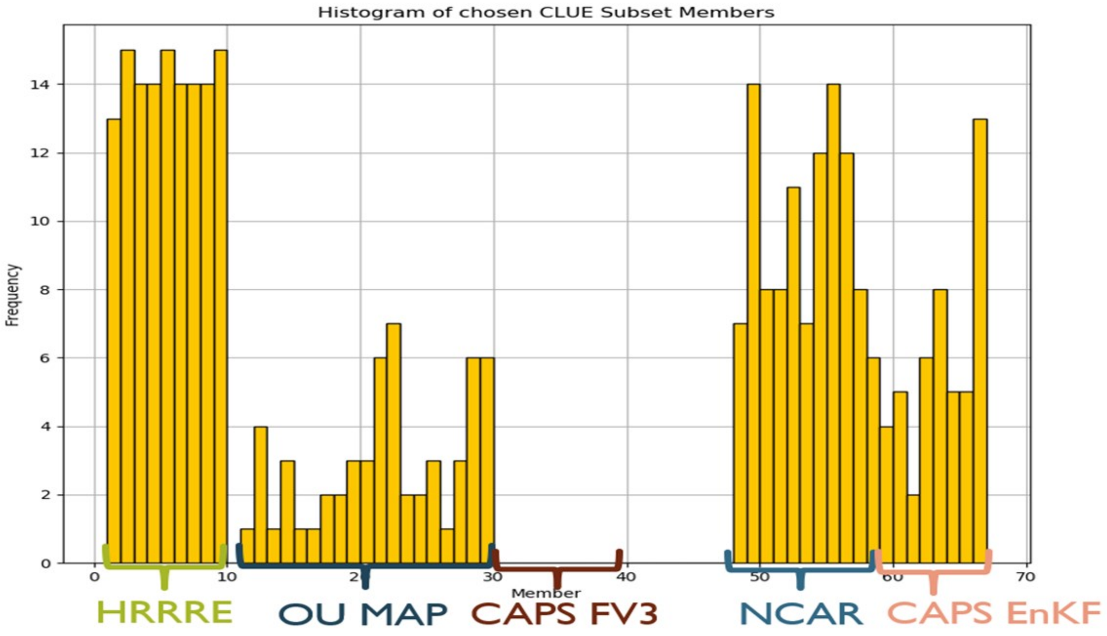


Figure 53 Histogram showing the frequency of the various TTU Full CLUE ensemble members chosen for the ensemble subset over the entire 5-week experiment using the UH response function.

4. Summary

The 2019 Spring Forecasting Experiment (2019 SFE) was conducted at the NOAA Hazardous Weather Testbed from 29 April – 31 May 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 the 2019 SFE was to utilize convection-allowing model and ensemble guidance in creating experimental high-temporal resolution probabilistic forecasts of severe weather hazards. Furthermore, this was the fourth 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. Additionally, this is the third year that a prototype Warn-on-Forecast system has been tested for issuing short-lead-time outlooks.

Several preliminary findings/accomplishments from the 2019 SFE are listed below:

- Explored different methods of generating high temporal resolution outlooks using experimental CAM ensemble guidance.
- Explored the ability to provide enhanced information on the conditional intensity of tornado, wind, and hail events by delineating areas expected to follow “normal”, “hatched”, or “double-hatched” distributions.
    - The majority of participants (70%) stated that generating the conditional intensity forecasts was “neither difficulty nor easy”, “easy”, or “very easy”.
    - Wind was the hazard often cited as being the most difficult to generate conditional intensity forecasts, which was in part due to the observed severe wind database being unreliable.
- Explored methods to include more detailed timing information by issuing potential severe timing areas (PSTs), which are enclosed areas valid for 4-h period that highlight the expected timing of severe weather occurrence.
    - When asked how difficult it was to draw PSTs, “difficult” was the most frequently chosen response (other choices were: “very difficult”, “neutral”, “easy”, and “very easy”).
- Examined various CAM ensemble systems within the CLUE using HREFv2.1 as a baseline.
    - As in other recent SFEs, while all of the ensembles provided useful guidance for Day 1 severe weather forecasting, the HREFv2.1 received higher subjective ratings than the other systems, which continues to highlight the challenge for a single-core, single-physics system to match the utility of a multi-model, multi-physics system.
    - The UM-based CAM ensemble had similar ratings to the HRRRE and NCAR ensemble, but didn’t perform as well as HREF.
    - The FV3-based CAM ensemble had lower mean/median ratings than most of the other 0000 UTC initialized ensembles suggesting more development work is needed to improve FV3-based CAM ensembles for severe weather applications.

    - Stochastic physics did not provide a practical improvement over a single-physics approach for severe weather forecasting.
    - OU MAP ensemble members using multi-scale perturbations (map-hybrid) performed very similar to a similarly configured members that used large scale perturbations derived from GEFS members (map-ICpert).
- In the HREF, replacing the HRW-NMMB members with a SAR-FV3 member did not have a large impact on the subjective performance of the HREF for severe weather forecasting, which supports EMC plans of moving forward with an HREF configuration that includes HRRR members and replaces the HRW-NMMB with the SAR-FV3.
- Tested a prototype WoF system in real-time for the third year at the Innovation Desk and the second year at the Severe Hazards Desk during an afternoon forecasting activity. Additionally, each week two NWS forecasters continued this activity until 8pm. Finally, the impact of a WoF system training exercise was examined.
- Examined real-time storm-scale FV3 simulations for the third year during the 2019 SFE.
    - It was found that a configuration of the Stand-Alone-Regional FV3 performed very similarly to a global-with-nest configuration (both run by EMC).
    - Subjective ratings revealed that FV3 reflectivity forecasts were often comparable to operational CAMs, and the best performing deterministic FV3 forecast was the NSSL SAR, which received average subjective ratings that were very similar to HRRRv3.
    - All of the deterministic SAR and global-with-nest versions of FV3 had a cool surface temperature bias.
    - Thompson and NSSL microphysics received very similar subjective ratings, while Morrison had lower ratings with more variability. Because of a low bias in convective cores with Morrison, it was suggested that the dBZ diagnostic needed to be adjusted in Morrison.
    - Scale-aware MYNN, Scale-aware Shin-Hong, and EDMF PBL schemes all performed very similarly. All PBL schemes had a persistent moist bias.
- Evaluated an ensemble sensitivity-based subsetting technique applied to the CLUE.
    - About half of the subjective evaluation responses indicated that the skill of the full ensemble and subset were the same, while 30% said the subset was an improvement. About 75% of responses suggested that the subsetting could be used as an operational product. While improvements outweighed degradations, successes and failures occurred at relatively close frequency.
- Hail diagnostics were examined which included a microphysics-based approach, a machine-learning approach, and the standard CAM storm-attribute fields: Max UH, Total UH (i.e., includes cyclonic and anti-cyclonic), and updraft speed.
    - None of the proxies stood out as best, but the total UH had a slight advantage over other methods and the microphysics-based approach stood out as performing worse than the other methods.
- An object-based approach (OBPROB) for visualizing and deriving probabilities from a CAM ensemble was evaluated.

- When asked whether OBPROB guidance provided unique information, 42% said that it did, 27% said it did not, and 29% were unsure.

Overall, the 2019 SFE 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 the 2019 SFE directly promote continued progress to improve forecasting of severe weather in support of the NWS Weather-Ready Nation initiative.

Acknowledgements

The 2019 SFE 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, GFDL, and EMC were vital to the success of the 2019 SFE. 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), Mark Weeks (Met Office), Curtis Alexander (ESRL/GSD), Jacob Carley (EMC), Eric Aligo (EMC), Lucas Harris (GFDL), Xuguang Wang (OU/MAP), Aaron Johnson (OU/MAP), Yongming Wang (OU/MAP), Tara Jensen (DTC), Tina Kalb (DTC), and Hank Fisher (DTC) were essential in generating and providing access to model forecasts or products examined on a daily basis.

References

Ancell, B.C., and G.J. Hakim, 2007: Comparing Adjoint- and Ensemble-Sensitivity Analysis with Applications to Observation Targeting. *Mon. Wea. Rev.*, **135**, 4117-4134.

Ancell, B.C., 2016: Improving High-Impact Forecasts through Sensitivity-Based Ensemble Subsets: Demonstration and Initial Tests. *Wea. Forecasting*, **31**, 1019-1036.

Clark, A. J., and Coauthors 2019: Spring Forecasting Experiment 2019 Program Overview and Operations Plan. Available online at: https://hwt.nssl.noaa.gov/sfe/2019/docs/HWT_SFE2019_operations_plan.pdf.

Gagne, D.J., A. McGovern, S.E. Haupt, R.A. Sobash, J.K. Williams, and M. Xue, 2017: Storm-Based Probabilistic Hail Forecasting with Machine Learning Applied to Convection-Allowing Ensembles. *Wea. Forecasting*, **32**, 1819–1840.

Gallo, B. T., A. J. Clark, B. T. Smith, R. L. Thompson, I. Jirak, and S. R. Dembek, 2018a: 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.

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.

Krocak, M. and H. Brooks, 2017: Towards Consistency in Forecasting Severe Weather Events across a Wide Range of Temporal and Spatial Scales in the FACETs Paradigm. 97th Annual AMS Meeting, Seattle, WA, Amer. Meteor. Soc. [Available online at https://ams.confex.com/ams/97Annual/webprogram/Paper308117.html]

Rothfusz, R. Schneider, D. Novak, K. Klockow-McClain, A. E. Gerard, C. Karstens, G. J. Stumpf, and T. M. Smith, 2018: FACETs: A Proposed Next-Generation Paradigm for High-Impact Weather Forecasting. Bull. Amer. Metor. Soc., 99, 2025-2043.

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

APPENDIX

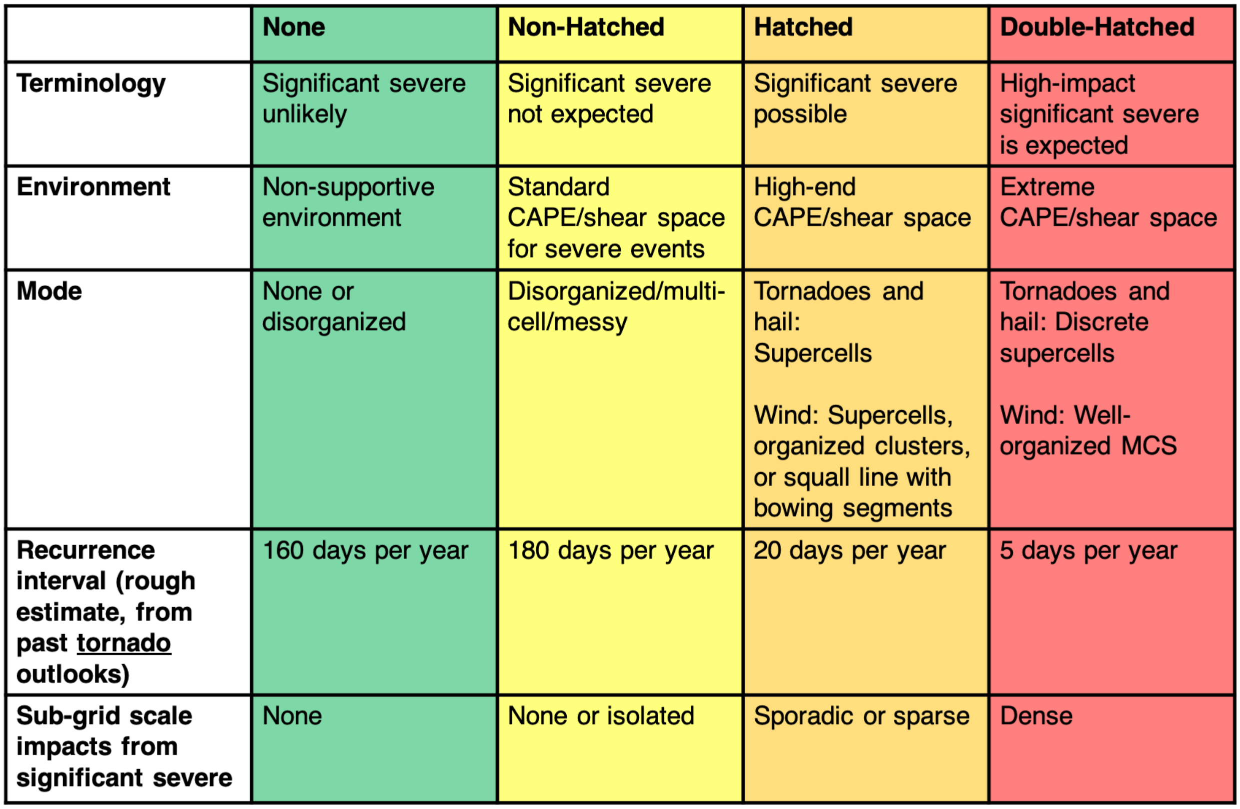
Table A1 Weekly participants during the 2019 SFE. Facilitators/leader for the 2019 SFE included Adam Clark (NSSL), Israel Jirak (SPC), Steve Weiss (retired SPC), Burkely Gallo (CIMMS/SPC/NSSL), Kenzie Krocak (CIMMS/NSSL/OU), Brett Roberts (CIMMS/SPC/NSSL), Kimberly Hoogewind (CIMMS/SPC/NSSL), Kent Knopfmeier (CIMMS/NSSL), and Andy Dean (SPC). WoF event activity facilitators included Pam Heinselman (NSSL), Kimberly Hoogewind (CIMMS/SPC/NSSL), Patrick Skinner (CIMMS/NSSL), Katie Wilson (CIMMS/NSSL), Jessie Choate (CIMMS/NSSL) and Corey Potvin (CIMMS/NSSL).

Week 1	**Week 2**	**Week 3**	**Week 4**	**Week 5**
April 29-May 3	**May 6-10**	**May 13-17**	**May 20-24**	**May 28-31**
SFE Participants (8am – 4pm)				
Brian Ancell (TTU)	Brian Ancell (TTU)	Austin Coleman (TTU)	Michael Brennan (NHC)	Ben Blake (EMC)
Austin Coleman (TTU)	Austin Coleman (TTU)	David Gagne (NCAR)	Clark Evans (UWM)	Jamie Wolff (DTC)
Willy Sedlacek (USAF)	Geoff Manikin (EMC)	Lance Bosart (SUNYA)	Jason Otkin (UW/CIMSS)	Curtis Alexander (GSD)
Shawn Corvec (ECCC)	Lara Pagano (WPC)	Tyler Leicht (SUNYA)	Greg Thompson (NCAR)	David Walters (UKmet)
Tracey Dorian (EMC)	Lindsay Blank (DTC)	Alex Mitchell (SUNYA)	Bill Gallus (ISU)	Steve Willington (UKmet)
Eric Aligo (EMC)	Glen Romine (NCAR)	Logan Dawson (EMC)	Zach Hiris (ISU)	Gordon Brooks (USAF)
Terra Ladwig (GSD)	Trevor Alcott (GSD)	Alicia Bentley (EMC)	Jacob Carley (EMC)	Amanda Burke (OU)
Christina Kalb (DTC)	Shin-Ping Kuan (CWB)	Ryan Sobash (NCAR)	Craig Schwartz (NCAR)	David Imy (SPC Ret.)
Eric Loken (OU)	Ping-Hsiang Wang (CWB)	John Brown (GSD)	Kai-Yuan Cheng (GFDL)	Jeff Milne (SPC)
David Jahn (SPC)	Jeff Duda (GSD)	Shin-Ping Kuan (CWB)	Ed Szoke (GSD)	Arianna Jordan (Howard U.)
David Harrison (SPC)	Victor Gensini (NIU)	Ping-Hsiang Wang (CWB)	Jon Petch (UKmet)	Andy Bollenbacher (WFO HNX)
Patrick Gilchrist (WFO GGW)	John Allen (CMU)	Aurore Porson (UKmet)	Paul Davies (UKmet)	Brandt Maxwell (WFO SGX)
Joseph Clark (WFO DTX)	Rachel North (UKmet)	Andy Hartley (UKmet)	Steve Willington (UKmet)	Michael Hill (WFO LIX)
Jimmy Correia (NWS AFS)	James Varndell (UKmet)	David Hayter (UKmet)	Neil Armstrong (UKmet)	Dan Hofmann (WFO LWX)
Seongmook Kim (CAPS - KMA)	David Hayter (UKmet)	Katie Deroche (AWC)	Arianna Jordan (Howard U.)	David Dowell (GSD)
	Becky Adams-Selin (AER)	David Stark (WFO OKX)	Sarah Trojniak (WPC)	Chris Stammers (Winnipeg MSC)
	Tom Hultquist (WFO MPX)	Brett Albright (WFO OAX)	Harald Richter (BoM)	
	Andy Wilkins (OU)	Daniel Zumpfe (WFO MSO)	Anders Jensen (NCAR)	
	Tom Galarneau (CIMMS/OU)		Rob Hepper (AWC)	
	John Henderson (AER)		*Rich Fulton (OWAQ)*	
	Chad Entremont (WFO JAN)		*Brian Oswiak (Toronto MSC)*	
Warn on Forecast Participants (12 – 8pm)				
David Cox (WFO JAN)	Brittany Newman (WFO GLD)	Andrew Moore (WFO FGF)	Larry Hopper (WFO PSR)	John Boris (WFO APX)
Joseph Cebulko (WFO ALY)	Suzanna Lindeman (WFO BOI)	Michael Hollan (WFO BIS)	Christina Leach (WFO JKL)	Aaron Mangels (WFO GID)

Table A2 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, preliminary reports, and MRMS MESH and rotation tracks	
• Days 1 & 2 full-period probabilistic forecasts of tornado, wind, and hail • Day 1 4-h period temporal disaggregation and guidance for tornado, wind, and hail	• Days 1 & 2 full-period probabilistic forecast of total severe • Days 1 & 2 4-h potential severe timing areas • Day 1 1-h and 4-h total severe outlooks
0845 – 0915: **Map Analysis** Hand analysis of 12Z upper-air & surface maps, discussion, and domain selection (from two areas)	
0915 – 1130: **Convective Outlook Generation**	
• Day 1 full-period probabilistic forecasts of tornado, wind, and hail valid 16-12Z over mesoscale area of interest* • Day 1 full-period (16-12Z) conditional intensity forecasts of tornado, wind, and hail using CLUE subsets*	• Day 1 full-period probabilistic forecast of total severe valid 16-12Z over mesoscale area of interest • Day 1 4-h potential severe timing (PST) areas (16-12Z) for full-period total severe ≥15% using CLUE subsets*
1130 – 1200: **Map Discussion** Brief discussion of today's forecast challenges and products Topic of the day: Ens. Subsetting (M), FV3 (T), WoF system (W), Met Office (R), CAM scorecard (F)	
1200 – 1300: **Lunch**	
1300 – 1345: **Convective Outlook Generation**	
• Day 2 full-period probabilistic forecasts of tornado, wind, and hail valid 12-12Z over mesoscale area of interest	• Day 2 full-period prob. forecast of total severe valid 12-12Z over mesoscale area of interest & 4-h PST areas (≥15% prob.)
1345 – 1500: **Scientific Evaluations**	
• Mesoscale Analyses • CLUE: CAM Ensembles • HREF Configurations w/FV3 • CLUE: Physics & IC Perturbations • Hail Guidance • Sensitivity-Based Ensemble Subsetting	• Ensemble Object-Based Probabilities • Deterministic CAMs (FV3 Nest, SAR) • Deterministic CAMs (HRRR, UM) • CLUE: FV3 Physics • WoF System Evaluation • WoF-based Outlook Evaluation
1500 – 1600 (2000 for WoF participants): **Short-term Outlook Update**	
• Update full-period prob/intensity forecasts of tornado, wind, and hail valid 21-12Z using observations and WoF system*	• Utilize obs. and WoF system to generate short (1-h), long (4-h), and targeted (1-h) probabilistic forecasts of total severe*
* Denotes forecasts also made by participants using the web drawing tool on Chromebooks.	

Table A3 Description of “non-hatched” (normal), “hatched”, and “double-hatch” conditional intensity forecasts for wind, hail, and tornadoes.



	None	**Non-Hatched**	**Hatched**	**Double-Hatched**
Terminology	Significant severe unlikely	Significant severe not expected	Significant severe possible	High-impact significant severe is expected
Environment	Non-supportive environment	Standard CAPE/shear space for severe events	High-end CAPE/shear space	Extreme CAPE/shear space
Mode	None or disorganized	Disorganized/multi-cell/messy	Tornadoes and hail: Supercells Wind: Supercells, organized clusters, or squall line with bowing segments	Tornadoes and hail: Discrete supercells Wind: Well-organized MCS
Recurrence interval (rough estimate, from past tornado outlooks)	160 days per year	180 days per year	20 days per year	5 days per year
Sub-grid scale impacts from significant severe	None	None or isolated	Sporadic or sparse	Dense