SPRING FORECASTING EXPERIMENT 2018

Conducted by the

EXPERIMENTAL FORECAST PROGRAM

of the

NOAA HAZARDOUS WEATHER TESTBED

http://hwt.nssl.noaa.gov/sfe/2018

HWT Facility – National Weather Center
30 April - 1 June 2018

Preliminary Findings and Results

Adam Clark[2], Israel Jirak[1], Burkely Gallo[1,3], Brett Roberts[1,2,3], Kent Knopfmeier[2,3], Robert Hepper[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], Louis Wicker[2], Scott Dembek[2,3], and Jack Hales[2]

(1) NOAA/NWS/NCEP Storm Prediction Center, Norman, Oklahoma
(2) NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma
(3) Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma
(4) School of Meteorology, University of Oklahoma

1. Introduction

The 2018 Spring Forecasting Experiment (SFE2018) was conducted from 30 April – 1 June by the Experimental Forecast Program (EFP) of the NOAA/Hazardous Weather Testbed (HWT), and was co-led by the NWS/Storm Prediction Center (SPC) and OAR/National Severe Storms Laboratory (NSSL). Additionally, important contributions of convection-allowing models (CAMs) were made from collaborators including the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma, Multi-scale data Assimilation and Predictability (MAP) Laboratory at the University of Oklahoma, 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 more than 80 forecasters, researchers, model developers, university faculty and graduate students from around the world (see Table 1 in Appendix). As in previous years, SFE2018 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. 2014) and Warn-on Forecast (WoF; Stensrud et al. 2009) visions:

Operational Product and Service Improvements:

- Explore the ability to generate higher temporal resolution Day 1 severe weather outlooks than those issued operationally by SPC.
    - 4-h periods for individual severe hazards (tornado, hail, and wind)
    - 1-h periods for near-term total severe
- Explore methods to include more detailed timing information 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 short lead-time, 1-h time window convective outlooks using a prototype WoF system.
- Test the utility of a prototype WoF system as guidance for 4-h time window severe hazard outlooks.

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 (HREFv2) as a baseline.
- Compare and assess different approaches in CAMs for predicting hail size.
- Compare and assess the current version of HREFv2 with possible future configurations that include extended-length HRRRv3 forecasts, as well as elimination of time-lagged members.
- Evaluate 3-km grid-spacing, convective-scale global-nested versions of the Finite Volume Cubed Sphere model (FV3) that have different microphysics and boundary layer parameterizations.
- Evaluate a prototype WoF system – the NSSL Experimental Warn-on-Forecast System for ensembles (NEWS-e) – for applications to short-term severe weather outlook generation.
- Evaluate whether ensemble sensitivity-based subset probabilities provide improved guidance relative to the full ensemble from which the ensemble sensitivity was computed.

- Compare forecasts from a global and high-resolution configuration of the UK Met Office's Unified Model to diagnose errors in the high-resolution UM inherited from its parent global model.
- Evaluate the utility of an objective-based approach for efficiently visualizing and deriving probabilities from a CAM ensemble.

As in previous experiments, a suite of state-of-the-art experimental CAM guidance contributed by our large group of collaborators was central to SFE2018. Additionally, for the third consecutive year, these contributions were formally coordinated into a single ensemble framework called the Community Leveraged Unified Ensemble (CLUE). The 2018 CLUE was constructed by having all groups agree on a set of model specifications (e.g., grid-spacing, vertical levels, domain size, etc.) so that the simulations contributed by each group could be used in controlled experiments. This design allowed us to conduct several experiments to aid in identifying optimal configuration strategies for CAM-based ensembles. The 2018 CLUE included 82 members using 3-km grid-spacing that allowed a set of five unique experiments. SFE2018 activities also involved testing of a Warn-on-Forecast prototype system, the NEWS-e.

This document summarizes the activities, core interests, and preliminary findings of SFE2018. 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/2018/docs/HWT_SFE2018_operations_plan.pdf). The remainder of this document is organized as follows: Section 2 provides an overview of the models and ensembles examined during SFE2018 along with a description of the daily activities, Section 3 reviews the preliminary findings of SFE2018, and Section 4 contains a summary of these findings.

2. Description

a) Experimental Models and Ensembles

Building upon successful experiments of previous years, SFE2018 focused on the generation of experimental probabilistic forecasts of severe weather valid over shorter time periods than current operational SPC severe weather outlooks. This is an important step toward addressing a strategy within the National Weather Service (NWS) of providing nearly continuous probabilistic hazard forecasts on increasingly fine spatial and temporal scales (i.e., FACETs), in support of the NWS Weather-Ready Nation initiative. As in previous experiments, a suite of new and improved experimental CAM guidance including ensembles was central to the generation of these forecasts. For all of the models, hourly maximum fields (HMFs) of explicit storm attributes such as simulated reflectivity, updraft helicity, updraft speed, and 10-m wind speed, were examined as part of the experimental forecast and evaluation process. Ninety-five unique CAMs were run for SFE2018, of which 82 were a part of the CLUE system. Other deterministic and ensemble CAMs outside of the CLUE were contributed by NSSL, EMC, GSD and the UK Met Office. To put the volume of CAMs run for SFE2018 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.

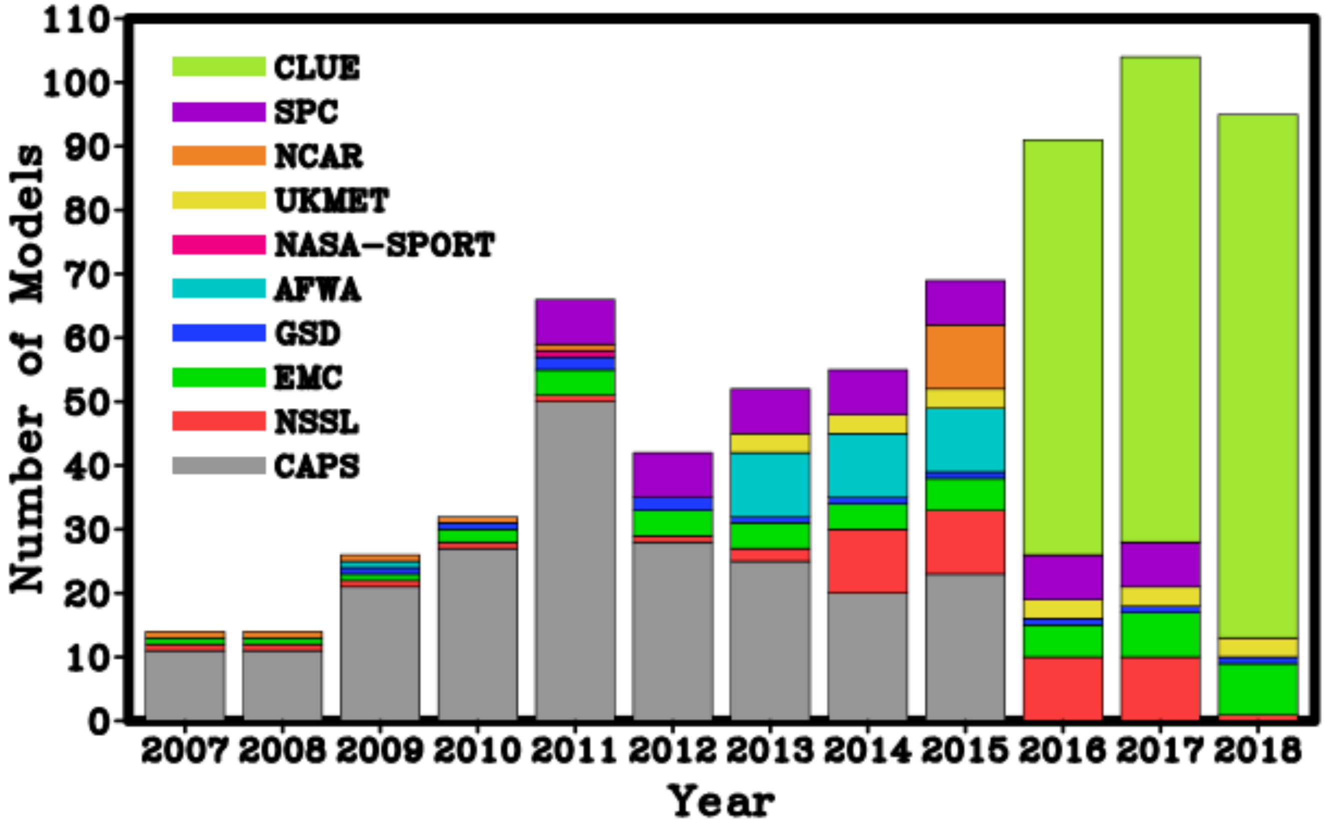


Figure 1 Number of CAMs run for SFEs since 2007. The different colored stacked bars indicate the contributing agencies.

More information on all of the modeling systems run for SFE2018 is given below.

1) THE COMMUNITY LEVERAGED UNIFIED ENSEMBLE (CLUE)

The 2018 CLUE is a carefully designed ensemble with subsets of members contributed by NOAA groups at NSSL, GFDL, and ESRL/GSD, and non-NOAA groups at CAPS (OU), MAP (OU), and NCAR. In addition, the Developmental Testbed Center (DTC) provided support for post-processing, and configurations for CAPS runs that used stochastic physics. To ensure consistent post-processing, visualization, and verification, all CLUE contributors used the same post-processing software to output the same set of model output fields on the same grid. An exception was some of the FV3 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. All CLUE members were initialized weekdays at 0000 UTC with 3-km grid-spacing covering a CONUS domain. A full description of all members and list of post-processed model fields are provided in the SFE2018 operations plan (Gallo et al. 2018). Table 1 provides a summary of each CLUE subset.

Table 1 Summary of CLUE subsets. IC/LBC perturbations labeled "SREF" indicate that IC perturbations were extracted from members of NCEP's Short-Range Ensemble Forecast system and added to 0000 UTC NAM analyses. In subsets with "yes" indicated for mixed-physics, the microphysics and turbulence parameterizations were varied. Note, a member in the mixed-phys ensemble was also used as a member in the single-phys ensemble. Thus, although the total number of members adds to 83, there were 82 unique members.

Clue Subset	**# of mems**	**IC/LBC perturbations**	**Mixed Physics**	**Data Assimilation**	**Model Core**	**Agency**
mixed-phys	11	SREF	yes	ARPS-3DVAR	ARW	CAPS (OU)
te14	1	none	no	ARPS-3DVAR	ARW	CAPS (OU)
single-phys	8	SREF	no	ARPS-3DVAR	ARW	CAPS (OU)
stoch-phys	8	SREF	no	ARPS-3DVAR	ARW	CAPS (OU)
caps-enkf	12	EnKF (CAPS)	yes	EnKF (CAPS)	ARW	CAPS (OU)
fv3-phys	11	none	yes	cold start (GFS)	FV3	CAPS (OU)
HRRR36	1	no	no	RAP-GSI/DFI	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)
hrrre	9	EnKF	no	EnKF	ARW	ESRL/GSD
nssl-fv3	1	no	no	cold start (GFS)	FV3	NSSL
gfdl-fv3	1	no	no	cold start (GFS)	FV3	GFDL

The design of CLUE allowed for 5 unique experiments that examined issues immediately relevant to the design of a NCEP/EMC operational CAM-based ensemble. These experiments are listed in Table 2.

Table 2 List of CLUE experiments for SFE2018.

Experiment Name	**Description**	**CLUE subsets**
Physics perturbation experiment	Three ensembles with perturbed ICs/LBCs were compared to test the effectiveness of different strategies for representing model error. One ensemble had single physics, one had mixed-physics, and one had single physics with stochastic perturbations.	mixed-phys, single-phys, & stoch-phys
Data assimilation comparisons	3DVAR and several different EnKF data assimilation approaches were compared. Note, this experiment was not as controlled as the others because there were other different aspects of the configurations in the subsets with different data assimilation.	map-hybrid, caps-enkf, ncar, & hrrre
Microphysics Sensitivities	The te14 member used a stochastically perturbed microphysics scheme, which was compared to a control member in the mixed-phys ensemble to evaluate sensitivities.	te14
FV3	Three versions of FV3 were examined and compared to current models, like HRRRv3, to gauge performance at convective scales.	fv3-phys01, nssl-fv3, gfdl-fv3, & HRRR36
FV3 physics	Convective-scale versions of FV3 were run with different microphysics and boundary layer parameterizations to examine sensitivities.	fv3-phys

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

The HREFv2 is an 8-member, convection-allowing ensemble that is run operationally at EMC. The version used for the HWT was slightly different than the configuration implemented operationally by EMC on 1 November 2017, however, the performance of both are very similar. HREFv2 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. All members, except for the NAM CONUS Nest, 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 (formerly called the SSEO). Thus, HREFv2 performance is considered the baseline against which potential future operational CAM ensemble configurations are compared.

3) MET OFFICE CONVECTION-ALLOWING MODEL RUNS

The operational configuration of the 2.2 km grid-spacing, nested high-resolution version of the UK Met Office's Unified Model [internally designated "Parallel Suite 41 (PS41)"] with forecasts to 120 h was initialized daily at 0000 and 1200 UTC and supplied to SFE2018. The UM forecasts had 70 vertical levels across a slightly sub-CONUS domain with initial and lateral boundary conditions from the 0000 and 1200 UTC initializations of the 10-km grid-spacing global configuration of the UM. This model configuration included a 3D turbulent mixing scheme using a locally scale-dependent blending of Smagorinsky and boundary layer mixing schemes with no convective parameterization. Stochastic perturbations were made to the low-level resolved-scale temperature field in conditionally unstable regimes (to encourage the transition from subgrid to resolved scale flows) and the microphysics was single moment. Partial cloudiness was diagnosed assuming a triangular moisture distribution with a width that is a universally specified function of height only.

In addition to the 2.2 km UM, data from the Met Office global model was provided to allow for comparison against the 2.2 km to gain more insight into the source of the errors in the convective scale model. The global data were provided from the OS40 version of the Met Office global UN, currently running at a horizontal grid-spacing of approximately 10 km in the mid-latitudes and using 70 vertical levels up to 80 km. It makes use of 4D-VAR hybrid data assimilation with the main scientific difference against the 2.2 km as follows: use of a mass flux convection scheme (based on Graham Rowntree), a Prognostic Cloud Scheme (PC2), a purely 1D non-local boundary layer scheme and schemes for parameterizing the effects of gravity wave drag and sub-grid orographic drag.

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

The 3-km grid-spacing HRRRv3 model developed by the ESRL/GSD, which became operational in July 2018, continued to be examined in SFE2018. The convection-allowing HRRR uses GSI hybrid data assimilation (instead of 3DVAR) with the latest 3-D radar reflectivity. The background ensemble for this assimilation is the 80-member GDAS (GFS) ensemble. The HRRRv3 runs every hour on a 3-km grid with output to 18 h, except at 0000, 0600, 1200, and 1800 UTC when it runs out to 36 hours. The HRRRv3 is initialized with an hour of 3-D radar reflectivity using a latent-heating specification technique including some refinements in this latent-heating from the parent RAPv4 model. The HRRRv3 uses grid-point statistical interpolation (GSI) hybrid GFS ensemble-variational data assimilation of conventional observations. Building upon the advancements in the operational HRRRv2 at NCEP, HRRRv3 includes assimilation of TAMDAR aircraft observations; refines assimilation of surface observations for improved lower-tropospheric temperature, dewpoint (humidity), winds, and cloud base heights; and places more weight on the ensemble contribution to the data assimilation. HRRRv3 also adds assimilation of lightning flash rates as a complement to radar reflectivity observations through a similar conversion to specified latent heating rates during a one-hour spin-up period in the model. Numerous model changes within the HRRRv3 include an update to WRF-ARW version 3.9, utilization of updated Thompson microphysics, transition to a hybrid sigma-pressure vertical coordinate for improved tropospheric temperature, and dewpoint and wind forecasts along with a higher resolution (15 second) land use dataset. Physics enhancements have also been made to the MYNN PBL scheme and RUC land surface model along with additional refinements to shallow cumulus/sub-grid-scale cloud parameterizations including enhanced interactions with the radiation and microphysics schemes for greater retention of cloud features.

5) HIGH RESOLUTION RAPID REFRESH ENSEMBLE (HRRRE)

In addition to the 0000 UTC initialized HRRRE runs that were a part of the 2018 CLUE, HRRRE forecasts were also provided at 1200 UTC. These forecasts were run across approximately the eastern 55% of the CONUS out to 24h across the same domain. This 1200 UTC ensemble was initialized from 3-km analyses in their data assimilation process rather than 15-km analyses, but are otherwise configured similarly to the 0000 UTC initialized HRRRE runs.

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

The NSSL Experimental Warn-on-Forecast System for ensembles (NEWS-e) is a 36-member WRF-based ensemble data assimilation system used to produce very short-range (0-6 h) probabilistic 18-member forecasts of supercell thunderstorm rotation, hail, high winds, and flash flooding. The starting point for each day's experiment was the experimental HRRRE provided by ESRL/GSD. A 6-h ensemble forecast launched from the 1200 UTC HRRRE analysis was used to provide initial conditions for the NEWS-e at 1800 UTC.

The daily NEWS-e domain location targeted the primary region where severe weather was anticipated and covered a 750-km wide region with very frequent 15-min DA cycles and forecasts every 30 minutes. All ensemble members utilized the NSSL 2-moment microphysics parameterization and the RAP land-surface model, but the PBL and radiation physics options were varied amongst the ensemble members to address uncertainties in model physics. MRMS radar reflectivity and Level II radial velocity data, cloud water path retrievals from the GOES-16 imager, and Oklahoma Mesonet observations (when available) were assimilated every 15 min using an EnKF approach, beginning at 1800 UTC each day. ASOS data was also assimilated at 15 minutes past each hour. A 6-h (5-h) ensemble forecast was initialized from the 1900 (2000) UTC NEWS-e analysis for HWT product evaluation from 2000 – 2100 UTC. Beginning at 2030 UTC, a 180-min ensemble forecast with 5-min output was launched every 30 minutes through 0300

UTC the next day. These forecasts were viewable using the web-based NEWS-e Forecast Viewer (https://www.nssl.noaa.gov/projects/wof/news-e/realtime/).

b) Daily Activities

SFE2018 activities were focused on forecasting severe convective weather at two separate desks, one forecasting individual hazards (Severe Hazards Desk) and the other forecasting total severe (Innovation Desk), with different experimental forecast products being generated at different temporal resolutions. Forecast and model evaluations also were an integral part of daily activities. A summary of forecast products and evaluation activities can be found below while a detailed schedule of daily activities is contained in the appendix (Table A1).

1) EXPERIMENTAL FORECAST PRODUCTS

Similar to previous years, the experimental forecasts explored the ability to add temporal specificity to longer-term SPC severe weather outlooks. All the forecast activities for SFE2018 focused on periods within the Day 1 time-frame. 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 severe hazards. At the Severe Hazards Desk, the first forecast mimicked the SPC operational Day 1 Convective Outlook, which consisted 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. The first forecast at the Innovation Desk also covered the 1600 to 1200 UTC period, but consisted of probabilities for total severe (combined tornado, hail, and wind). These experimental forecasts covered a limited-area domain where the primary severe weather threat for the day was expected to occur and/or where interesting forecast challenges were expected.

Each desk then manually stratified the Day 1 outlooks into periods with higher temporal resolution. The Severe Hazards Desk generated separate probability forecasts of large hail, damaging wind, and/or tornadoes for two 4 h periods: 1700-2100 and 2100-0100 UTC. As an alternative way of stratifying the Day 1 outlook, the Innovation Desk created a product aimed toward the emergency management community, designating areas and 4-h periods where severe convective hazards occurrence was expected throughout the day. These potential severe timing (PST) areas were designated within the areas of 15% probability in the Day 1 full-period outlook generated by the Innovation Desk. This approach built upon the isochrones approach during SFE2016 and SFE2017, with the timing and areal information both available on the final figure. Despite the different end products, 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.

During previous experiments, calibrated probabilistic severe guidance from the SREF/SSEO (Jirak et al. 2014) was used to temporally disaggregate a 1600–1200 UTC period human forecast. A scaling factor was formulated by matching the full-period calibrated severe SREF/SSEO guidance to the human forecast, and then this scaling factor (unique at every grid point) was applied to the calibrated severe guidance for each individual period. Finally, consistency checks were conducted to arrive at the final temporally disaggregated forecasts (Jirak et al. 2012). These automated forecasts from SFE2012 to SFE2017 fared favorably both in terms of objective metrics (e.g., CSI, FSS) and subjective impressions when compared to manually drawn forecasts. Similarly, for SFE2018, the 1600–1200 UTC human forecasts for the individual hazards were temporally disaggregated into the 4-h periods (1700–2100 UTC and 2100–0100 UTC) using HREF/SREF calibrated hazard guidance to provide a first guess for the two forecast periods.

The first set of short-time-window forecasts and timing forecasts were issued in the morning by both desks. At both the Severe Hazards Desk and the Innovation Desk, the lead forecaster generated the short-time-window forecasts on the N-AWIPS machines. However, the participants were split into five groups for each desk and used a web interface to generate their own short-time window probability forecasts using new Google Chromebooks. The redesigned web interface is similar to the Probabilistic Hazard Information (PHI) tool used in past experiments, but has been specifically designed to incorporate data from CLUE subsets and other experimental CAM ensemble guidance. Each Chromebook was associated with a specific ensemble or CLUE subset; participants were asked to base their forecast on that ensemble or CLUE subset. Additionally, the web interface has other important observational and model fields for participants to utilize in the forecast generation process. After issuing the high-temporal resolution individual forecasts based on model subsets and reporting the anonymous demographics of the group (e.g., the forecasting experience of the participants in each group), the desks regrouped and discussed the forecasts and behavior of the CLUE subsets. This approach more effectively engaged the participants directly with the CLUE subsets, since in prior years participants only interacted with CLUE subsets through facilitator-led discussions. After the teams issued and discussed the high-temporal resolution forecasts, there was a map discussion summarizing forecast challenges and highlighting interesting findings from the previous day open to all tenants of the National Weather Center. Each day of the week also featured a brief discussion of a special topic (Table A1).

After lunch, the Innovation Desk updated their PSTs, and each desk examined operational guidance as a group. Of the five ensemble subsets, three were updated at 1200 UTC to test the impact of updated CAM ensemble guidance on the timing forecasts. Participants that used CLUE subsets for which 1200 UTC guidance was not available updated their forecasts based on the most recent observational data and recent deterministic CAM guidance, such as the HRRR. Since the forecast process for these updates began in the afternoon, participants were instructed to only update their PSTs valid between 1900 and 1200 UTC. Participants at the Severe Hazards Desk followed a similar process, but generated a new forecast valid from 1900-2300 UTC.

Later in the afternoon, scientific evaluations were conducted (summarized in the next section). For the final activity of the day on Tuesday through Friday, forecast products using the WoF-prototype system, NEWS-e, were generated at both desks. For the Innovation Desk activity, the 1900 UTC initialized NEWS-e with 6-h forecast products available at the website https://www.nssl.noaa.gov/projects/wof/news-e/realtime/ were used to issue two 1-h time window forecasts of total severe valid 2100–2200 and 2200–2300 (i.e., 4–5PM and 5–6 PM CDT). Then, these forecasts were updated using 2000 UTC initialized NEWS-e products. Forecasts were drawn by facilitators (Clark and Gallo) and informed by small groups of participants interrogating NEWS-e data on their

Chromebooks, as well as by the forecast lead (Jack Hales). At the Severe Hazards Desk, participants used the NEWS-e data to update their probabilistic hail, wind, and tornado forecasts valid from 2100–0100 UTC.

To prepare participants for the NEWS-e activity, a training session was provided 3–4PM on Monday of each week. This training session included a description of NEWS-e and provided an overview of how to navigate the NEWS-e website and view forecast products. Following the presentation portion of the training, facilitators worked with smaller groups (of ~5 participants) and walked through a test case to become familiar with the NEWS-e activity. After practicing the issuance of a 1-hour outlook and update, participants were asked to view, answer, and ask for clarification on a short set of survey questions that were completed following each NEWS-e activity session. Twenty-nine questions were available in a Google survey form, and consisted of multiple-choice, ranking, and open-ended questions designed to capture participants' perceptions of the NEWS-e products specific to the forecast challenge presented in the activity. Finally, participants were made aware of additional NEWS-e-specific survey questions to be asked during the verification evaluation activity scheduled first thing on Tue-Fri mornings. These questions were appended to the Google survey form that was used for the verification evaluation activity that evaluated all experimental forecasts made on the previous day.

The training session was also used to obtain participants' consent (per IRB protocol) to take part in this activity and answer survey questions. Additionally, participants were asked to provide their subjective rating of forecasting experience on a scale of 1–3 (none/minimal, some, and extensive). Participants were given examples of what these different rating levels meant. The ratings were used to assign participants each day to either group 1 or group 2 of the activity to ensure a balance of forecasting experience for each of the outlooks that were issued, as well as to encourage discussion between participants of varying professional backgrounds (i.e., operational and research oriented).

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 SFE2018 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 SFE2018. 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: B. Gallo)

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. These outlooks were rated the next day by overlaying the forecast with Local Storm Reports (LSRs), watches, and warnings. A "practically perfect" forecast (Hitchens et al. 2013) was also generated from the LSRs and displayed alongside the experimental forecast 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. An example experimental outlook along with the practically perfect outlook is shown in Figure 2.

In general, participants thought that the Day-1 outlooks performed well (mean rating of 6.5/10; Fig. 3). Outlooks were given better ratings on high-end days, which are defined as days when the practically perfect forecast indicated a 45% or greater probability. 12 of the 24 experiment days were high-end according to this criterion, while 8 of these days included experimental outlooks with 45% or greater probabilities (Table 3). Forecast probability magnitudes generally matched the practically perfect forecasts within a categorical outlook category; only two days 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.

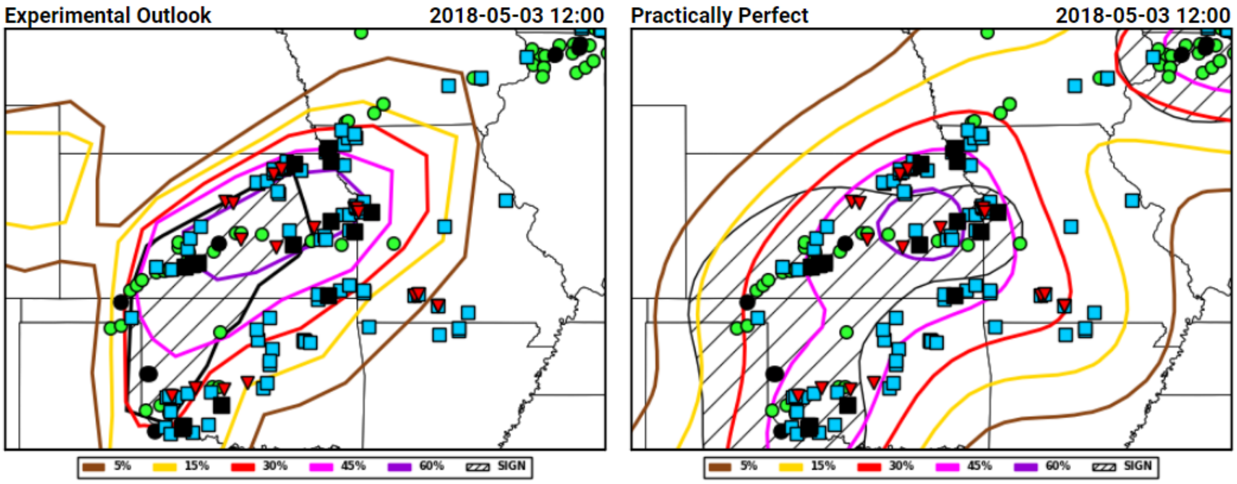


Figure 2 An experimental outlook (a) and practically perfect forecast (b) overlaid with wind (blue squares), significant wind (black squares), hail (green circles), significant hail (black circles), and tornado (red inverted triangles) reports. The brown, yellow, red, magenta, and purple contours indicate 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.

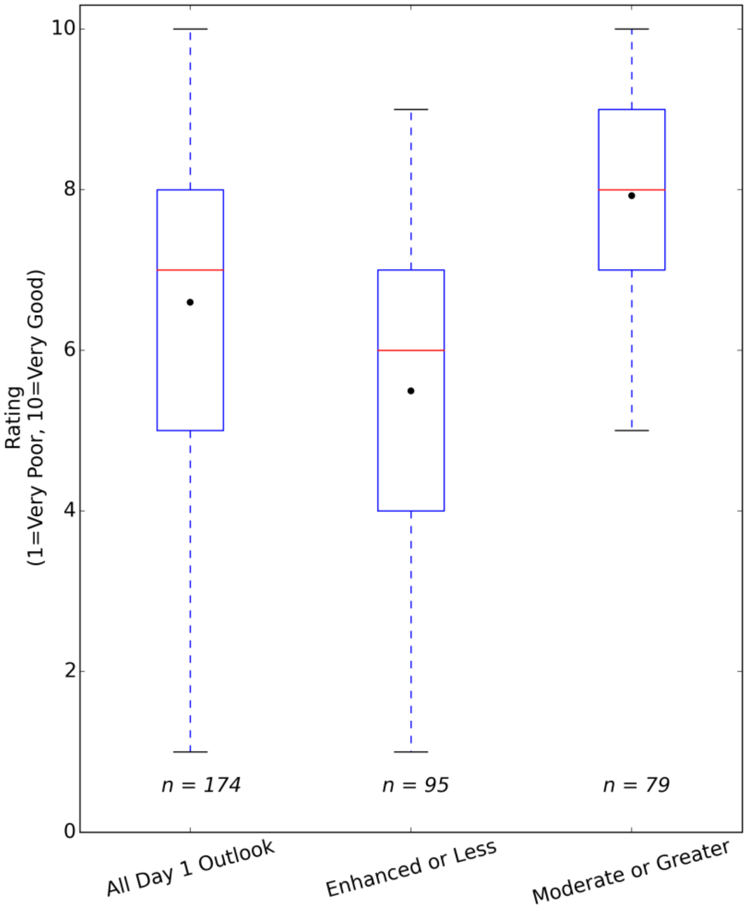


Figure 3 Subjective rating of the Day-1 outlook on all days, lower-end days, and high-end days.

Table 3 Probabilistic breakdown of lead and practically perfect forecasts during SFE 2018. Probabilities are assigned based on the maximum probability within the limited area domain of interest.

Category	# of Days (Lead)	# of Days (PP)
5%	1	3
15%	7	3
30%	8	6
45%	7	8
60%	1	4

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 (Lead)	Day 1 Max Probability (PP)
30-Apr	30	15
1-May	45	45
2-May	60	60
3-May	30	15
4-May	45	60
7-May	15	30
8-May	30	5
9-May	15	45
10-May	30	45
11-May	15	5
14-May	30	60
15-May	45	60
16-May	5	5
17-May	30	45
18-May	45	45
21-May	15	30
22-May	15	45
23-May	30	45
24-May	15	30
25-May	15	15
29-May	45	60
30-May	45	45
31-May	45	60
1-Jun	45	45

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

One of the challenges facing the FACETs paradigm is the gap in hazard information between the long-lead-time convective outlook probabilities and the short-lead-time warning-scale probabilities. Often, the only information between a convective outlook and a warning is a watch or mesoscale discussion. To address this gap in hazard information, this work continues that from previous SFEs testing timing products that provide information on a sub-daily, regional scale. Specifically, participants were asked to create PSTs indicating 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 Mixed Physics, CAPS Stochastic Physics, HRRRE, and NCAR. Also, one group used the operational HREFv2. Many more fields were available from each subset for participants to draw over than in previous years, including environmental fields such as CAPE and storm attribute fields such as simulated reflectivity and updraft helicity (UH). Participants issued a preliminary set of forecasts after the Day 1 outlook and prior to the morning forecast discussion, and then updated the forecasts in the early afternoon. Three subsets (the HRRRE, NCAR, and HREFv2 ensembles) had 1200 UTC forecast cycles available in the afternoon; the other participants were asked to solely use the original guidance, plus observed environmental trends and operational deterministic guidance, such as later runs of the HRRR. 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 4.

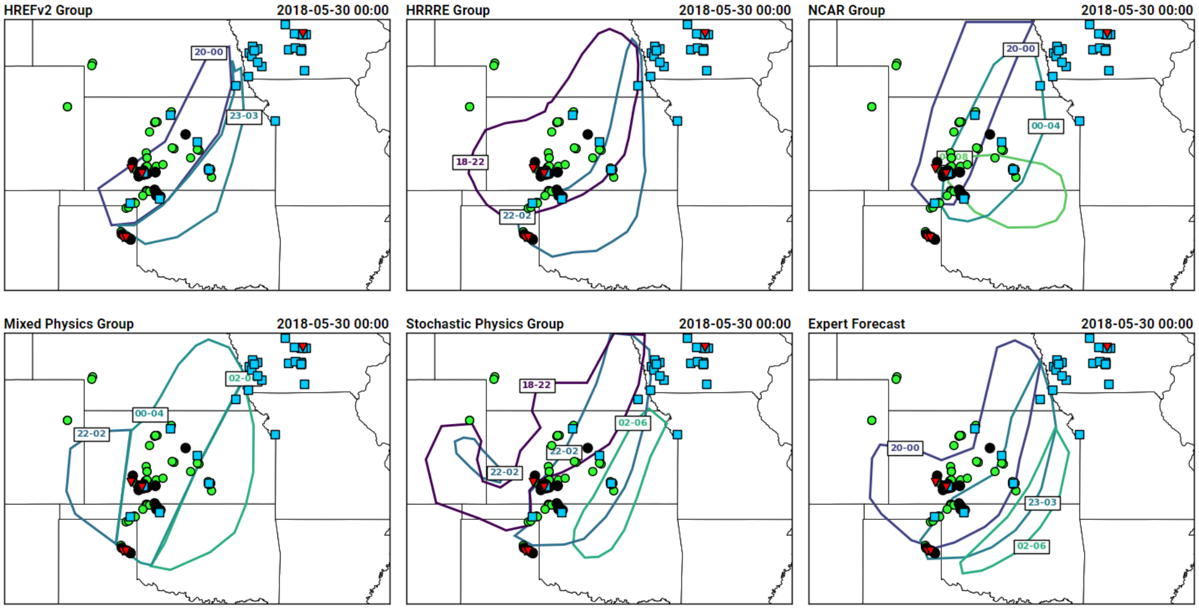


Figure 4 PST areas issued by each group and the lead forecaster on 29 May 2018. Reports shown occurred from 20-00 UTC, with green circles, black circles, blue squares, black squares, and red inverted triangles indicating hail, significant hail, wind, significant wind, and tornado reports.

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.

Throughout the experiment, it became evident that there were two prominent theories on how best to draw the PSTs. The first includes 2 areas that overlap temporally but not spatially (Fig. 5a). This indicates some uncertainty in where the severe reports will be occurring between 0000 and 0100 UTC, as either box would be valid. The second theory (Fig. 5b) has both temporal and spatial overlap. The justification was that there is some uncertainty about when the severe threat will be occurring in the overlapping area, hence the longer time period (2000 to 0200 UTC).

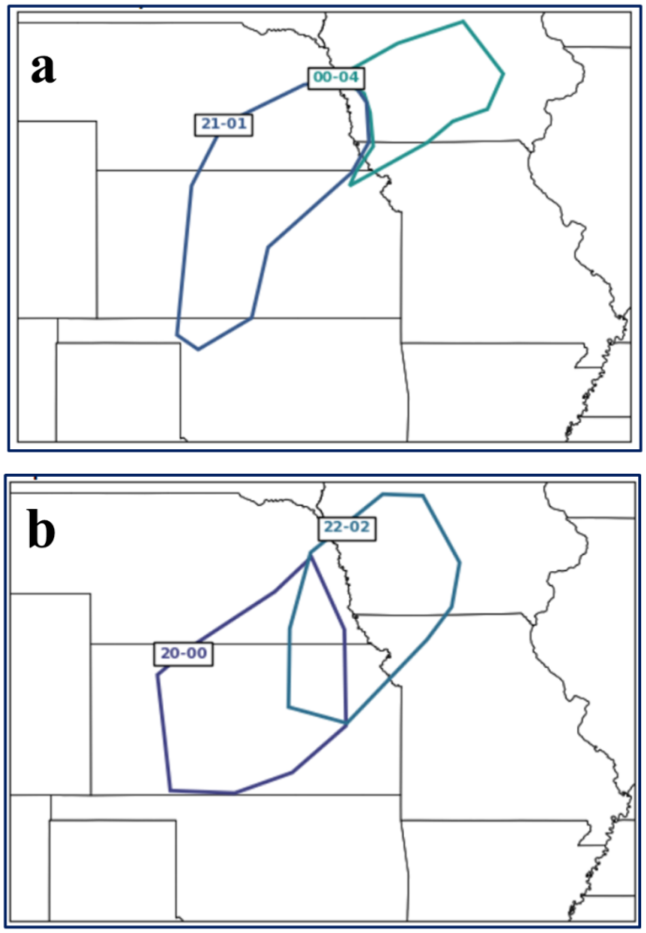


Figure 5 PST forecast philosophies (a) temporal overlap with no spatial overlap, and (b) temporal and spatial overlap.

Feedback from forecasters in SFE2018 included many comments about this year's visualization compared to the isochrone visualization (which was tested in the 2016 and 2017 SFE). Most people liked the PST visualization better, as they thought areas were easier to interpret than contours. Representative quotes about the strengths and weaknesses of the product are shown in Table 5.

Table 5 Selected quotes from forecasters about the Potential Severe Timing (PST) Product

Strengths	**Weaknesses**
I am seeing the applications of this tool being extremely useful, especially for EMs.	*We need to eliminate overlap when possible*
I could see this being really useful operationally to provide greater information on the timing windows.	*It's challenging to identify just 4 hours, and to differentiate between convective initiation and severe report initiation.*
I think we have the ability to do this, and guidance is only going to get better.	*We need ensemble guidance products that show timing better.*

For quantitative evaluation of the PST forecasts, two-by-two contingency tables were constructed using local storm reports that occurred between 1600 and 1200 UTC the following day. The reports and areas were gridded onto an 80 km grid and then grid points were tallied up for each box in the two-by-two table. This process was done for all 24 case days during SFE 2018, and then the probability of detection and success ratio was calculated for each day and the full experiment (Fig. 6).

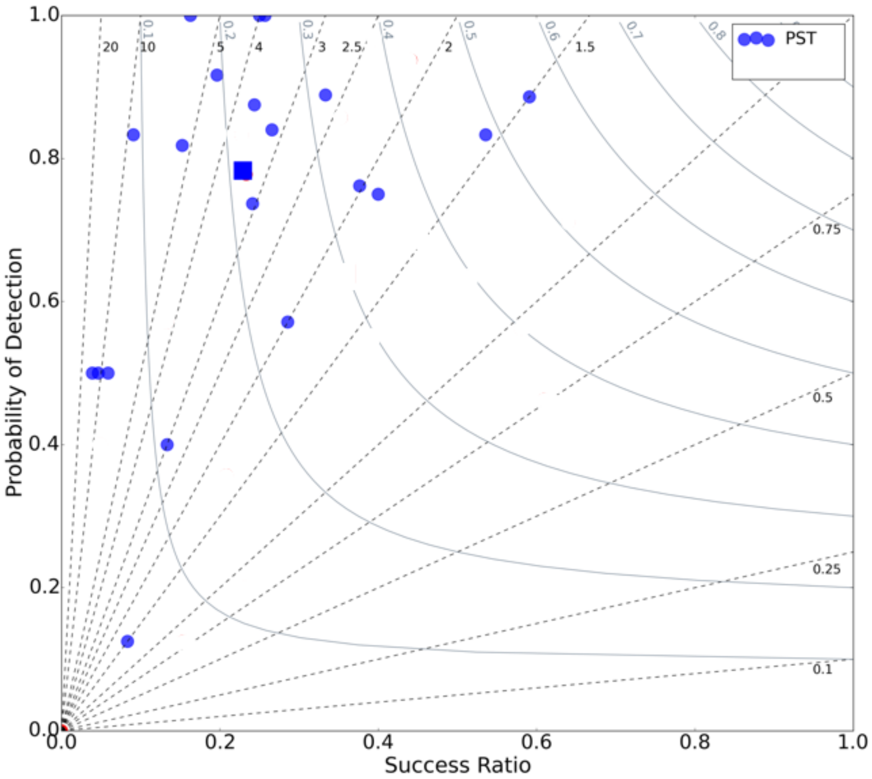


Figure 6 Performance diagram of the PST forecasts (blue) for SFE forecast days between April 30 and June 1.

Overall, the PSTs had relatively high PODs, but success ratios were generally below 0.50. Since PSTs were drawn in the morning, often 4-6 hours before the first severe report of the day, they were relatively large. Thus, they captured many of the reports on any given day, but suffered from a large false alarm area.

For the subjective evaluations, each day participants evaluated forecasts issued the previous day. They were asked to rate the PSTs issued by the lead forecaster, as well as the forecasts issued by their own group. Finally, participants indicated which ensemble subsets they used, and which group they thought performed the best overall. Typically, the lead forecaster scored quite highly on both the preliminary and final PSTs, with a median score of 7/10 for both time frames (Fig. 7). However, the mean ratings show a large increase from the preliminary to the final PST areas, and the 25th percentile rating increases from 5/10 in the preliminary to 6/10 in the final. The time of the first and last PST issued remained constant between the preliminary and final forecasts over the entire sample, with a median first PST time of 2000 – 0000 UTC and a median last PST time of 2300 – 0300 UTC. Nine of the twenty-four days had a shift in the time of the first PST, with a tendency to shift the outlooks later, and eleven of the twenty-four days had a shift in the time of the last PST, which also tended to be shifted later. Most of the timing adjustments between the preliminary and final outlooks were only a one-hour change, with one two-hour change each in the time of first PST and time of last PST, and one three-hour change in the time of last PST. Despite the small sample size, we would expect more shifts in the timing of the last PST area due to it being closer to the end of the forecast period. Between the preliminary and final outlooks, contours could also be added or removed if the forecaster's certainty increased or decreased.

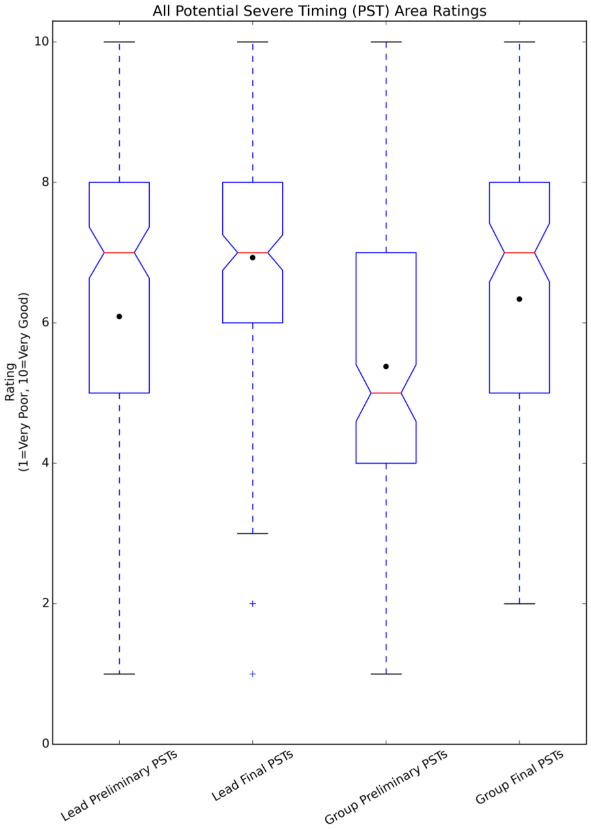


Figure 7 Subjective evaluation distributions for the PST areas from the lead forecaster and from the groups evaluating their own PST areas.

Overall, the participant PSTs were not rated as highly as those of the lead forecaster, especially the preliminary areas (Fig. 7). The updated, final areas had a similar ratings distribution to the preliminary PSTs issued by the lead forecaster. It is possible that, after seeing the lead forecaster's preliminary PSTs, participants may have modified their own PSTs towards those of the lead forecaster. Each group typically showed improvement from the preliminary to the final outlooks, with the median rating improving one point in all groups (Fig. 8). Groups using the HREFv2 and the HRRRE generally gave their forecasts higher ratings than groups using the other three ensembles (recall that participants were asked to only rate their own group's forecast). However, when participants were asked to choose which group performed the best (looking at both preliminary and final PSTs), the CAPS Stochastic Physics ensemble group was second, after the HREFv2 (Fig. 9). The group using the NCAR ensemble had the largest variation in their preliminary PST ratings. Further analysis is needed to determine whether there was an improvement in the forecasts beyond the perceived improvement shown in the subjective results, particularly since the improvement occurred for groups using all ensemble subsets, including those that did not have an updated 1200 UTC cycle of guidance. Further analysis is also needed to determine whether the "best forecasts" chosen subjectively align with the best forecasts indicated by objective metrics. Finally, demographic data collected on the participants will be analyzed to determine whether group composition played a role in which group's forecasts performed the best (i.e., was the group using the HREFv2 always composed of forecasters, who typically have more experience forecasting than model developers?). Participants also brought up potential issues in generating this product that will influence the design of next year's experiment. These comments include the potential difficulty of collaborating with a partner to create a forecast that both people agree on, and needing increased clarity on how closely participants are to follow their assigned experimental model subset when generating their forecasts. Other future plans for the PSTs include looking at different visualization options (polygons vs. shading), developing automated "first guess" PSTs based on convection allowing model output, and developing verification methods that fairly evaluate the quality of PST forecasts.

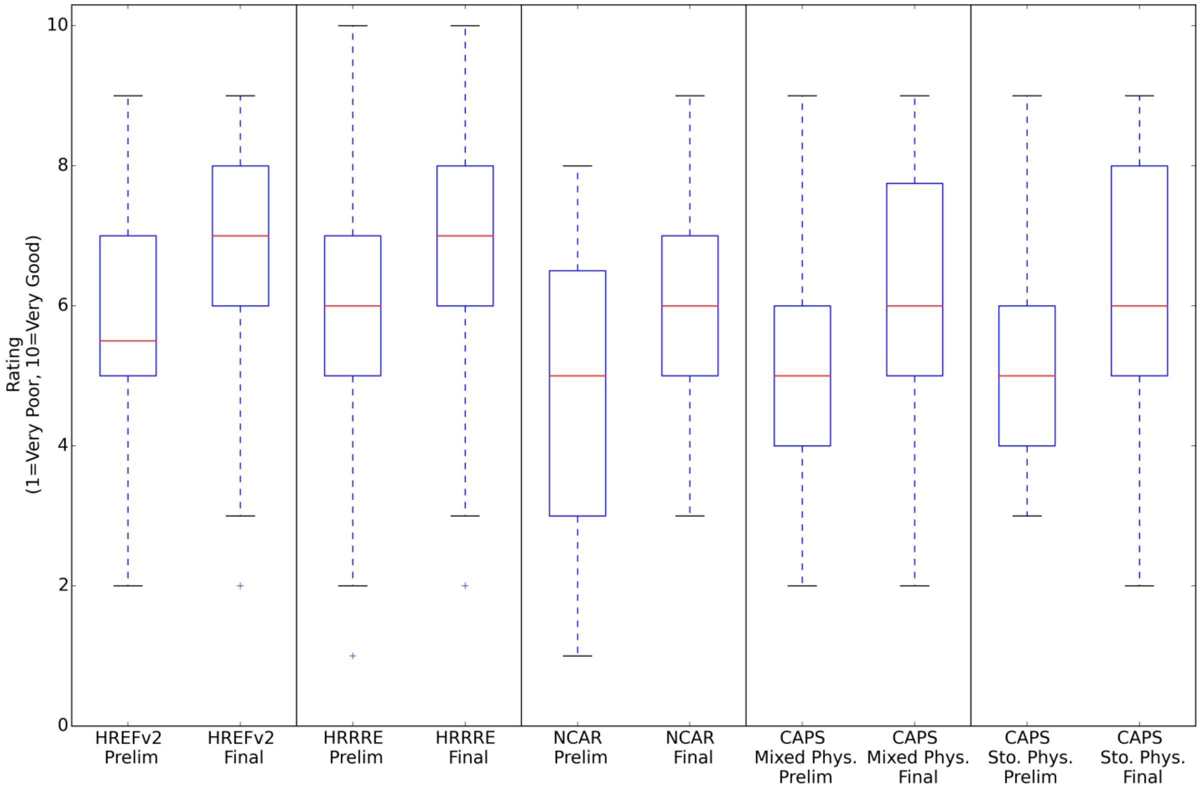


Figure 8 Subjective evaluations of each group's initial and final forecasts.

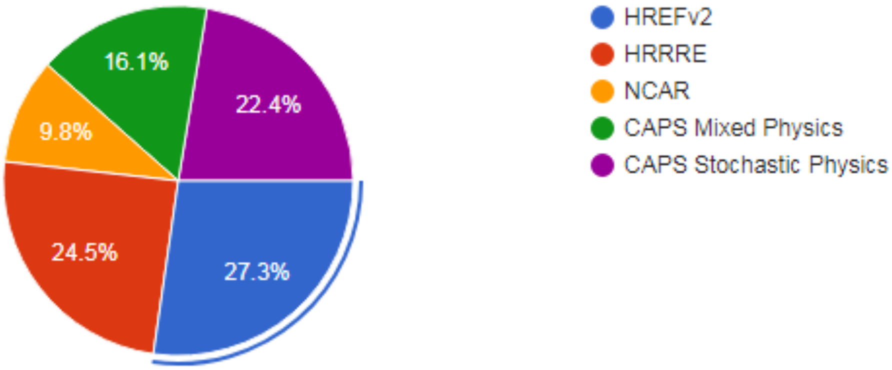


Figure 9 Responses to the question "Which group's forecasts performed best overall?"

3) NEWS-E EVALUATIONS (credit: J. Choate)

The NEWS-e (Wheatley et al. 2015, Jones et al. 2016) was tested in the SFE for the second year during the 2018 season. This prototype WoF system is a frequently updated, regional-scale, on-demand convection-allowing ensemble analysis and prediction system, nested within an experimental hourly CAM ensemble forecast system (currently the HRRRE). The 2018 configuration changed slightly from 2017 to produce longer forecasts for 0–6 h predictions of individual convective storms and mesoscale environments that provide probabilistic forecast guidance. This guidance includes products such as the probability of simulated reflectivity above a threshold at a grid point, and ensemble percentile values (e.g., 90th) of fields such as accumulated rainfall, 2–5-km UH, and 0–2-km vertical vorticity. Participants were given training on the system before they utilized NEWS-e for the outlook activity. They were also asked to respond to a series of questionnaires at different points throughout the activity. These sections will be referred to as Training, Outlook Activity, and Questionnaires.

Training

On Monday of each week, all SFE participants completed training on the NEWS-e system during the last hour of the day (3-4pm). This training session was chosen because researchers found that during SFE2017, participants spent a majority of their first day at the Innovation Desk trying to orientate themselves to the NEWS-e Viewer and were not able to fully contribute to the activity. Furthermore, since participants joined the Innovation Desk on different days during the week, the entire group was rarely on the same page. The Monday afternoon training allotted time for all participants to interact with several members of the NEWS-e development team and ask any questions that arose before they were expected to complete a forecast.

The Monday training session consisted of an explanation of the Warn-on-Forecast program, the NEWS-e configuration and verification, and how to navigate the new NEWS-e Viewer. Participants were able to follow along with a demonstration on how to navigate the webpage and where certain products could be found. Participants were also introduced to the outlook activity and the three questionnaires they would be asked to complete each day while working in smaller groups on a test case. This allowed participants to navigate the website and become comfortable with the location and meaning of different products. This test case also gave participants the opportunity to ask any clarifying questions on the survey questions they would answer each day.

Outlook Activity

Starting on Tuesday, participants worked within their Innovation or Severe Hazards Desk groups. At the Severe Hazards Desk, the NEWS-e was used to update their probabilistic hail, wind, and tornado forecasts valid from 2100-0100 UTC. At the Innovation Desk, new 1-h severe weather outlooks were issued. The primary goals of these outlooks were to 1) explore how short-term ensemble forecast guidance from NEWS-e could be used by groups of forecasters to produce a series of 1-h severe weather outlooks and 2) observe how the forecasters' understanding, use, and attitudes about NEWS-e guidance evolved through the experiment. Each morning, subjective verification of the previous afternoon outlooks was performed by comparing them to "practically perfect" hindcasts.

The outlook activity consisted of producing two 1-h outlooks of total severe probabilities over the NEWS-e domain (decided jointly by the WoF researchers and SPC forecasters) between 2100–2200 and 2200–2300 UTC. These outlooks were produced using only the 1900 UTC NEWS-e 6-h forecast (valid 1900–0100 UTC) and then updated using the 2000 UTC NEWS-e 5-h forecast (valid 2000–0100 UTC) along with current observations including radar, satellite, and surface observations. An overview of the 2018 NEWS-e configuration is provided in Figure 10. Participants were separated into either Group 1 (led by A. Clark) or Group 2 (led by B. Gallo) based on self-ranked forecast experience to try to balance forecast knowledge between groups. A third group consisted only of the lead forecaster and a researcher (J. Hales and J. Choate, respectively). Initial outlooks produced using the 1900 UTC NEWS-e guidance were submitted to an internal database by 2030 UTC and updated outlooks produced from the 2000 UTC NEWS-e guidance were submitted by 2100 UTC, resulting in four total outlooks per team, or twelve outlooks total. An outlook breakdown can be seen in Figure 11.

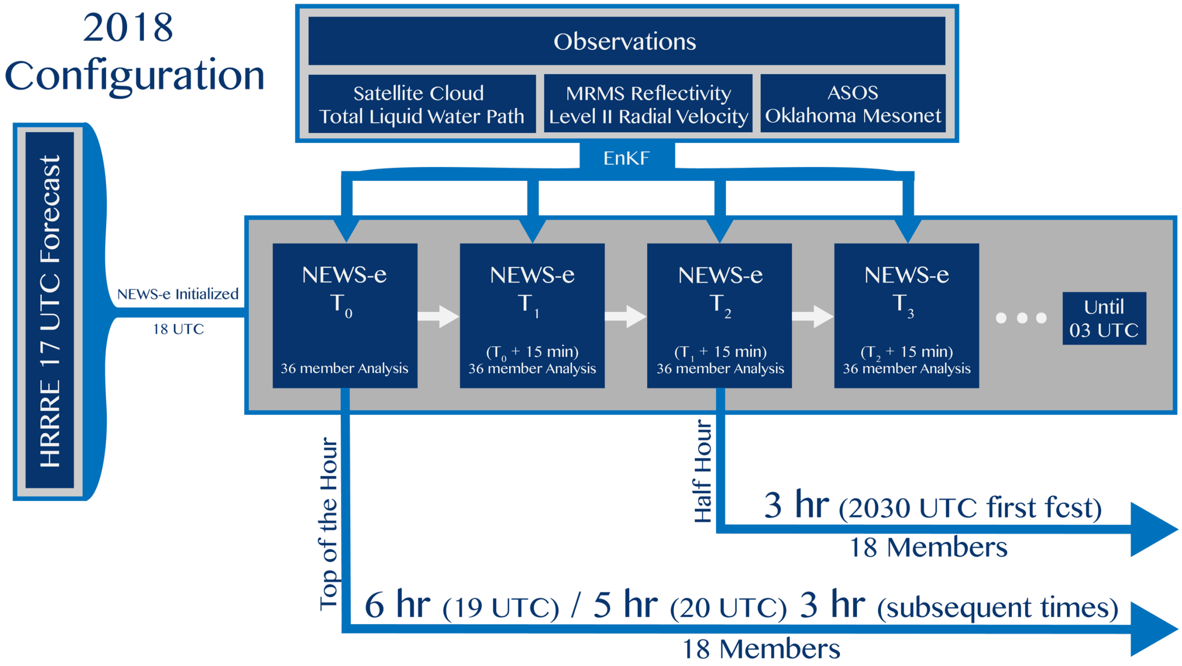
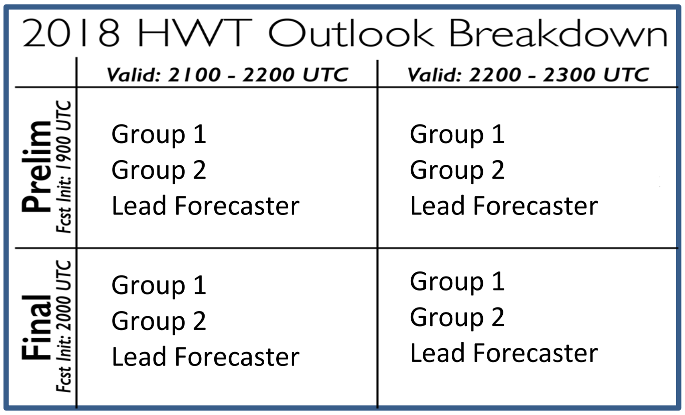


Figure 10 The 2018 NEWS-e configuration. 6 and 5 h forecasts were initialized at 1900 and 2000 UTC, respectively. Starting with 2030 UTC and at every half hour interval until 0300 UTC, 3 h forecasts were initialized.



2018 HWT Outlook Breakdown

	Valid: 2100 - 2200 UTC	Valid: 2200 - 2300 UTC
Prelim Fcst Init: 1900 UTC	Group 1 Group 2 Lead Forecaster	Group 1 Group 2 Lead Forecaster
Final Fcst Init: 2000 UTC	Group 1 Group 2 Lead Forecaster	Group 1 Group 2 Lead Forecaster

Figure 11 Summary of Innovation Desk NEWS-e outlooks.

For this experiment, there were three identical NEWS-e Viewer webpages made and labeled for each group of forecasters. Each team was asked to only use the NEWS-e Viewer that was assigned to their team. This instruction was given so researchers could track what type of products the participants were using during the forecast activity. Observations forecasters used were not tracked. The tracked information was used to inform researchers on what types of products were preferred for creating the outlooks, how often certain types of products were used, and how some products could possibly influence a group's outlook. For a quick overview of product usage during the activity, products were ranked as the "top" products used to create outlooks each day by how often they were requested to the server (Fig. 12). Of note is that paintball products which provide both deterministic and probabilistic information as well as ensemble spread were used most often. Also, environmental products were used more during the prelim rather than final outlook process and member viewer usage dropped by half during the final outlook drawing. These products were also able to show us how they influence the drawing of outlooks. For example, Figure 13 shows the most used product for each group overlaid by that group's outlook for May 4th. Group 1 drew their highest contour around the probability of UH while Group 2 focused more on the leaded edge of the percentile graphic of simulated composite reflectivity.

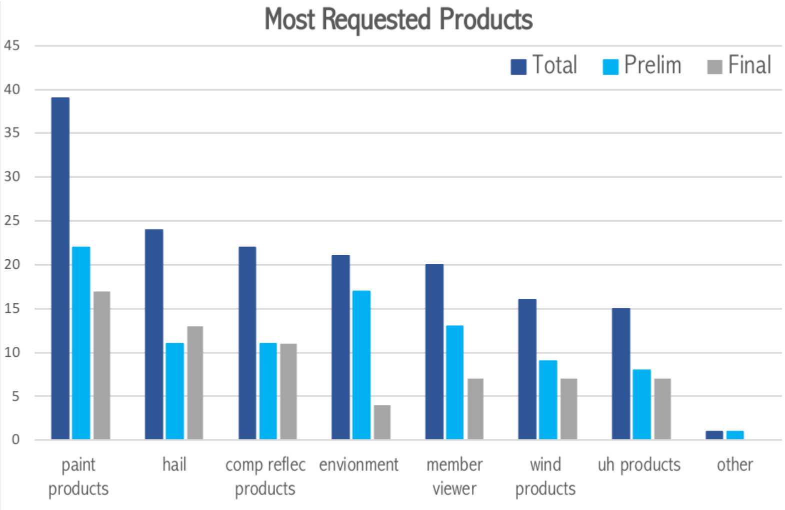


Figure 12 Most requested products from the web-based NEWS-e forecast viewer.

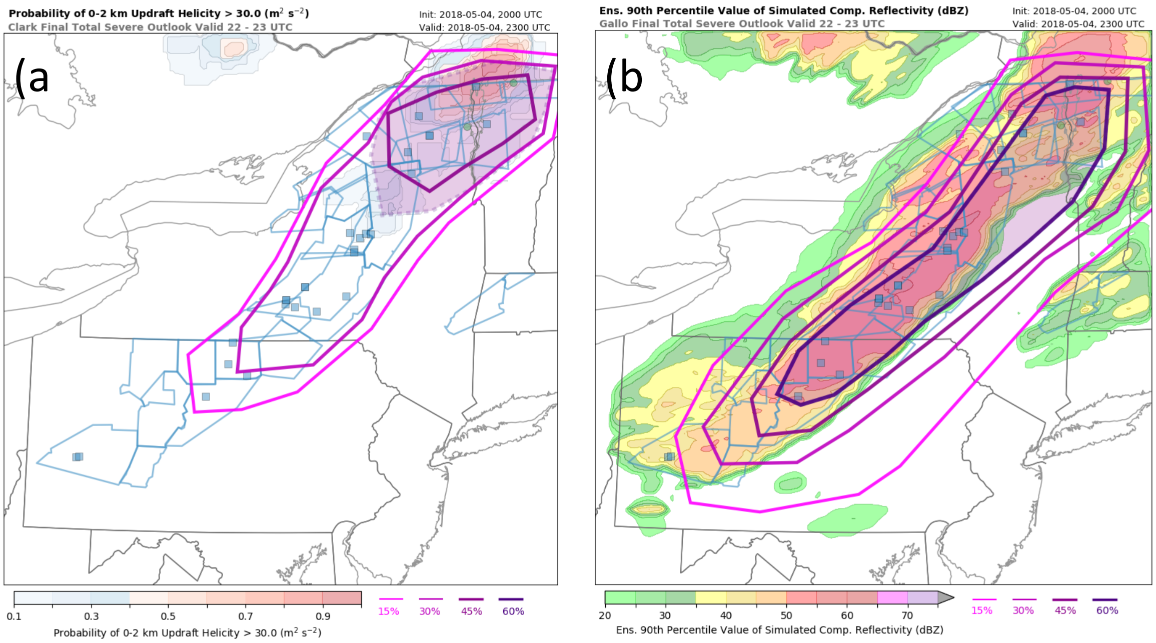

Probability of 0-2 km Updraft Helicity > 30.0 (m^2 s^{-2})
Clark Final Total Severe Outlook Valid 22 - 23 UTC
Init: 2018-05-04, 2000 UTC
Valid: 2018-05-04, 2300 UTC
(a
0.1
0.3
0.5
0.7
0.9
Probability of 0-2 km Updraft Helicity > 30.0 (m^2 s^{-2})
15%
30%
45%
60%
Ens. 90th Percentile Value of Simulated Comp. Reflectivity (dBZ)
Gallo Final Total Severe Outlook Valid 22 - 23 UTC
Init: 2018-05-04, 2000 UTC
Valid: 2018-05-04, 2300 UTC
(b
20
30
40
50
60
70
Ens. 90th Percentile Value of Simulated Comp. Reflectivity (dBZ)
15%
30%
45%
60%

Figure 13 (a) Group 1 (Clark) outlook for total severe valid 2200 – 2300 UTC 4 May 2018 indicated by contours with the transparent purple shading (outlined by dotted contour lines) indicating 10% or greater probability of significant severe. The shading shows NEWS-e forecasts for probability of 0-2 km AGL updraft helicity > 30 m^{2}s^{-2}. (b) Same as (a), except for the Group 2 (Gallo) outlook and the shading shows NEWS-e forecasts of the 90th percentile value of simulated composite reflectivity.

The next morning the outlooks were verified against “practically perfect” probabilities and verification questions were answered by the participants. An example of what would be seen each morning is shown in Figure 14. Forecast days were only included in these verification diagrams if outlooks from both groups were successfully submitted. When comparing all of Group 1’s outlooks to all of Group 2’s outlooks using reliability diagrams, Group 1 seemed to perform better (Fig. 15a). This comparison is based off the Practically Perfect guidance and it is important to remember that skill is a function of the scale at which you verify. A forecast would be considered perfectly reliable if it followed the 45-degree dashed line. If a point is above the 45-degree line, it represents an under-forecasted event, if it is below the 45-degree line, then it was an over-forecasted event. There are further questions to explore when considering the verification of these outlooks. An interesting note is that when comparing all of the preliminary outlooks to the final outlooks, the preliminary outlooks seems to perform slightly better. One would assume that the updated, or final, outlooks would be better predictors of the areas of total severe hazards, but this result shows the opposite (Fig. 15b). This may be due to participants feeling like a change needed to be made for the sake of the activity. More research needs to be put into understanding their decision process while drawing the outlook contours.

Questionnaires

Participants were asked to answer three sets of questions. Two surveys (the Prelim and Final) were taken during the outlook activity and were linked to each group’s webpage. The third survey was taken the next morning as part of the daily verification activity. The Prelim survey was completed after each group submitted their preliminary outlooks (i.e., the outlooks using the 1900 UTC NEWS-e forecast). The Final survey was completed after the final outlooks were submitted (i.e., the outlooks using the 2000 UTC NEWS-e forecast). These two sets of questions asked about items such as participants' confidence in their group's forecast, the quality of the NEWS-e forecasts, and to what extent they thought their group's forecast differed from NEWS-e output. The Final survey, taken after the final outlooks, also asked questions about that day's activity overall (e.g., how much participants changed their forecasts from preliminary to final, how difficult the forecast was, and how satisfied participants were with the forecast overall). Examples of Question 3 from the Prelim and Final questionnaires and the participants' responses can be seen in Figures 16 and 17. An example of one of the questions asked later in the Final survey is shown in Figure 18. The verification surveys the next morning were meant to test the participants' perceptions after they had seen how their outlooks performed. This survey asked similar questions as the Prelim and Final surveys to see if their thoughts on the outlook performance were influenced by the verification contours. An example is shown in Figure 19. The responses to these questionnaires will help researchers understand some of the thoughts forecasters had while completing their outlooks, how these different perceptions could have affected group performance, and how these perceptions changed after participants saw verification results.

Ongoing research continues to evaluate group outlooks based on different sets of practically perfect probabilities. Researchers are also considering different effects on group performance, such as: group-leader dynamic, imbalance of group experience, and size and scale of each group's outlook contours.

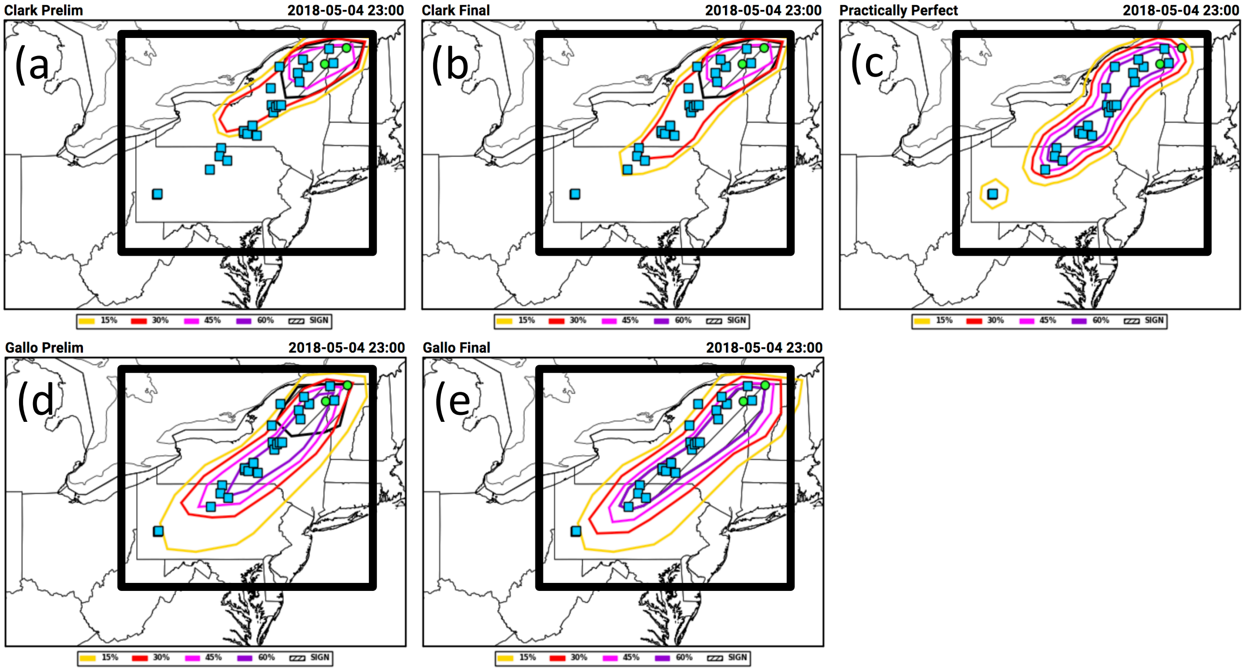


Figure 14 Group 1 (a) Prelim and (b) Final outlooks valid 2200 – 2300 UTC 4 May 2018 (contours), and (c) corresponding practically perfect observations (contours). (d) and (e), same as (a) and (b), except the Group 2 outlooks. Locations of storm reports marked in each panel and the black boxes indicate the NEWS-e domain.

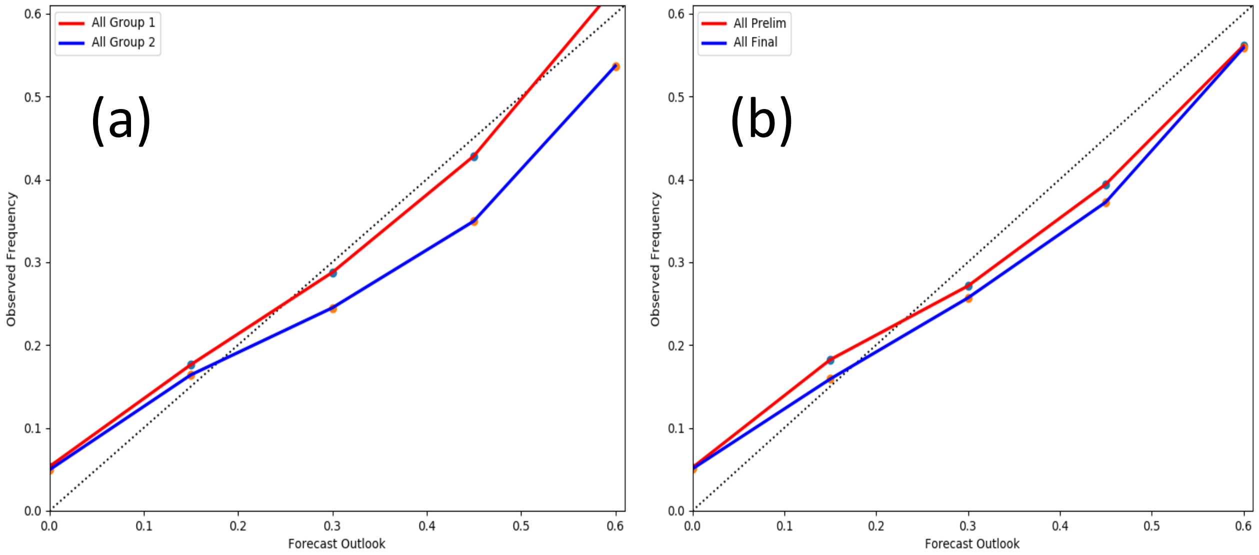


Figure 15 (a) Reliability diagrams for all Group 1 (red) and Group 2 (blue) outlooks. (b) Same as (a), except for all Prelim (red) and Final (blue) outlooks.

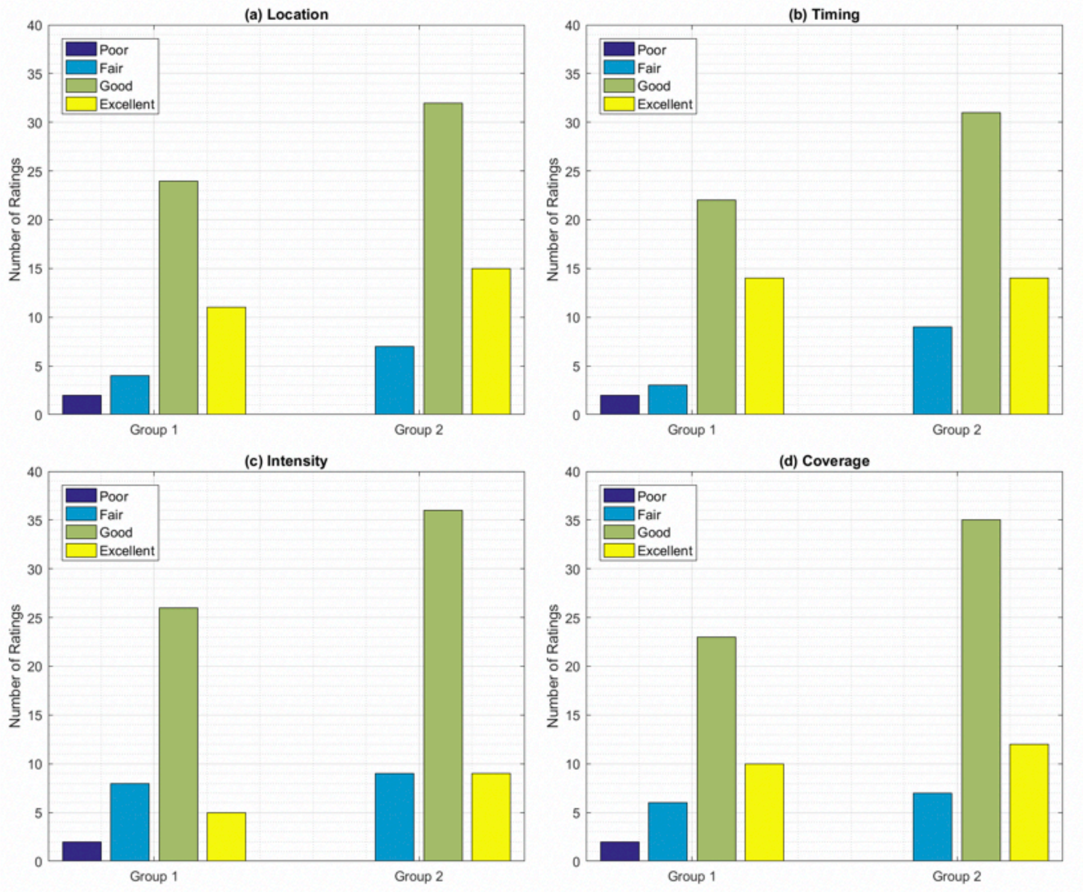


Figure 16 Survey responses to the question, "Compared to your expectations, how would you rate the quality of the 1900 UTC NEWS-e forecast in terms of the following?" (a) location, (b) timing, (c) intensity, and (d) coverage.

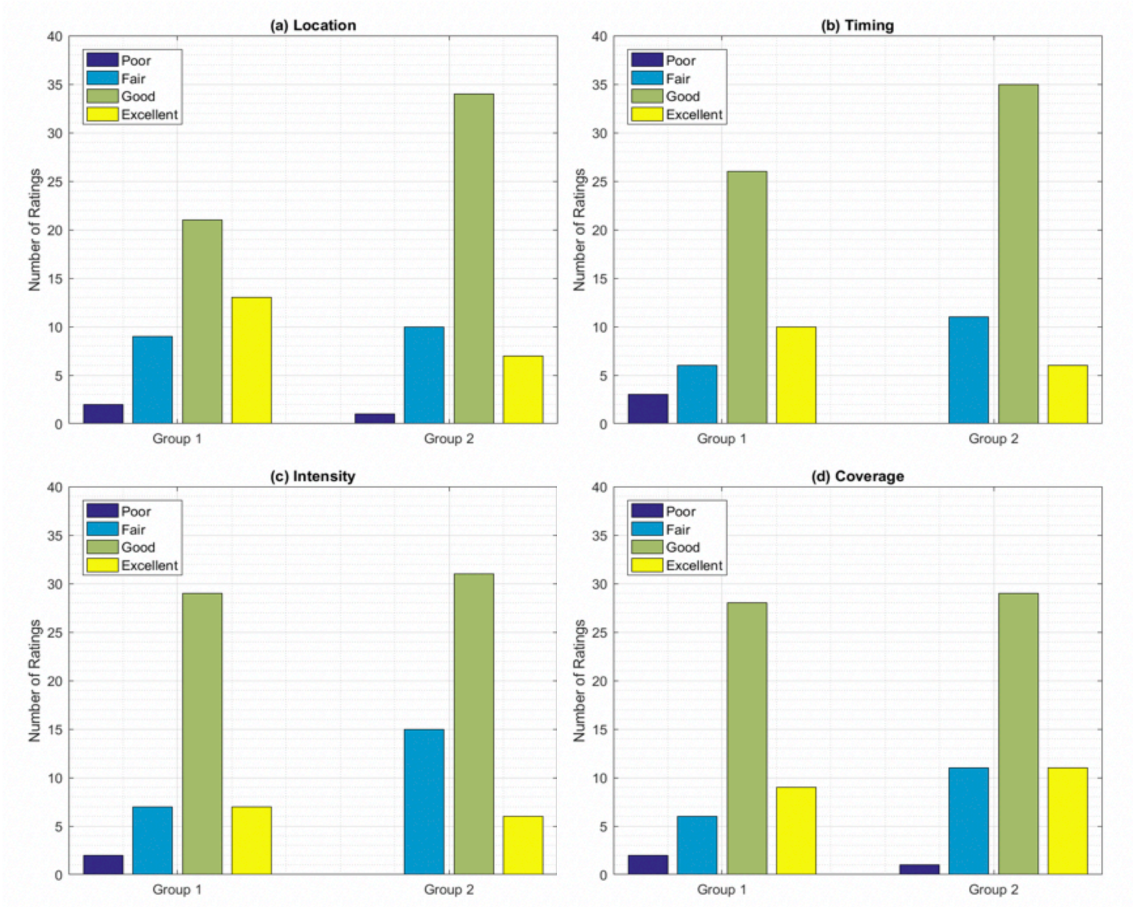


Figure 17 Same as Fig. 16, except for the 2000 UTC NEWS-e forecast.

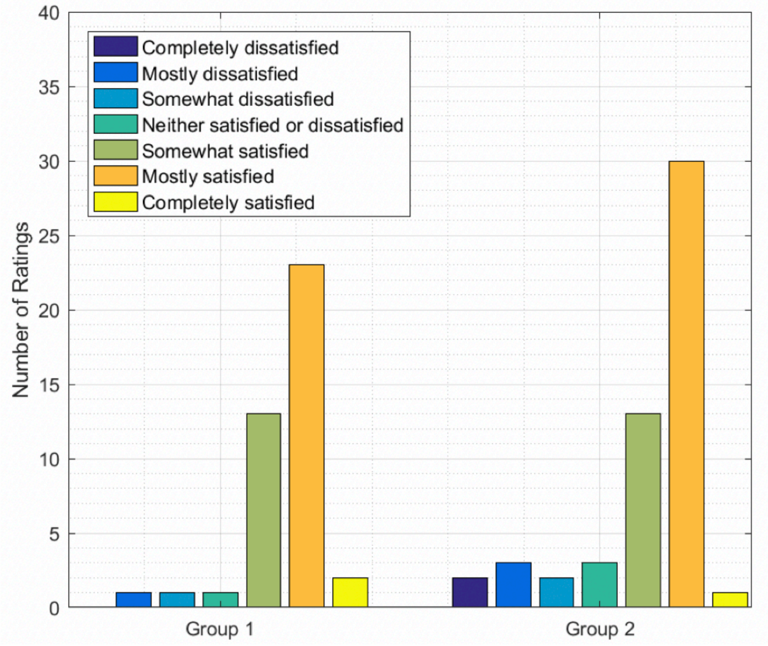


Figure 18 Summary of responses to the question, "How satisfied are you with the overall forecast performance?"

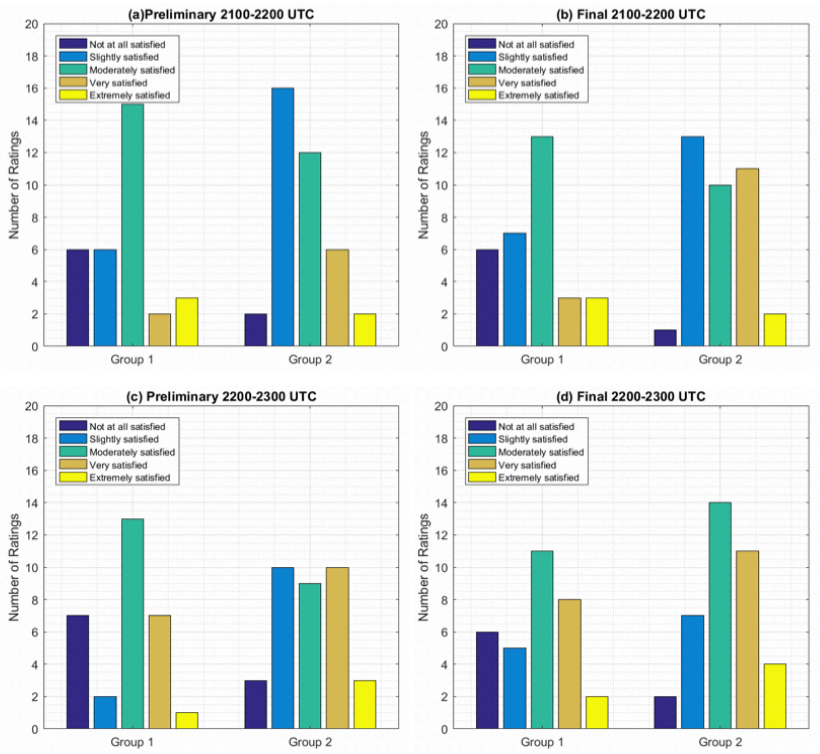


Figure 19 Summary of responses to the question, "How satisfied are you with the overall performance of your team's outlook?"

b) Evaluation of experimental forecast products – Severe Hazards Desk (credit: I. Jirak)

1) DAY 1 FULL-PERIOD GUIDANCE AND OUTLOOKS

Experimental probabilistic tornado, hail, and wind outlooks were generated for the Day 1 period (i.e., 1600-1200 UTC) during the morning activities of SFE2018. These experimental outlooks and HREF-based hazard guidance were subjectively evaluated (using a rating scale from 1-10) on the following day using local storm reports, NWS warnings, and radar-derived products as verification sources. The HREF-based guidance included the HREF/SREF calibration approach (Jirak et al. 2014) for tornadoes, hail, and wind, the STP calibration approach for tornadoes (Gallo et al. 2018a), and a machine-learning approach for hail (Gagne et al. 2017). An example of the web-based interface used to make these evaluations and comparisons is shown in Fig. 20.

For the full-period probability forecasts of tornadoes and severe hail, the HREF/SREF calibrated guidance compared favorably in terms of subjective ratings to the experimental outlooks (Fig. 21), which is noteworthy considering that the HREF/SREF guidance was available to the forecasters while generating the outlooks. The STP calibrated guidance for tornadoes and the machine-learning (ML) guidance for hail generally received lower subjective ratings for the full-period forecasts compared to the HREF/SREF guidance and the experimental outlooks (Fig. 21). For the probabilistic severe wind outlooks, the forecaster-generated outlooks were often rated higher than the HREF/SREF guidance, indicating the calibrated guidance for severe wind is not as mature as the guidance for tornadoes and hail.

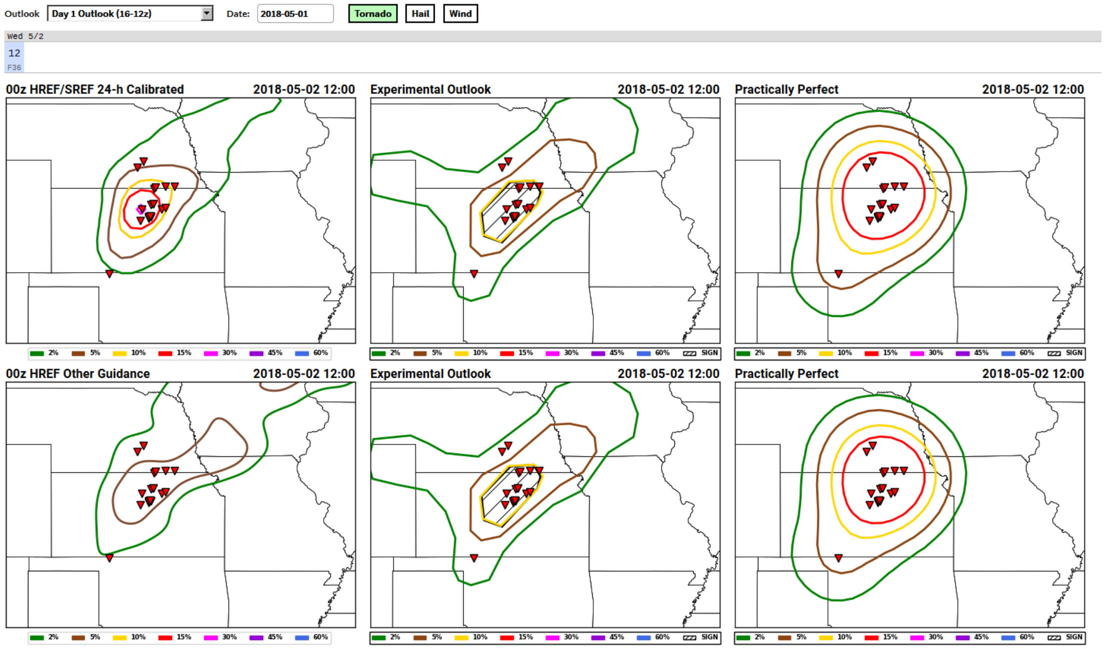

Outlook
Day 1 Outlook (16-12z)
Date: 2018-05-01
Tornado
Hail
Wind
Wed 5/2
12
F36
00z HREF/SREF 24-h Calibrated 2018-05-02 12:00
Experimental Outlook 2018-05-02 12:00
Practically Perfect 2018-05-02 12:00
00z HREF Other Guidance 2018-05-02 12:00
Experimental Outlook 2018-05-02 12:00
Practically Perfect 2018-05-02 12:00
2%
5%
10%
15%
30%
45%
60%
SIGN

Figure 20 Six-panel comparison plot used to conduct the evaluation of the calibrated guidance and experimental outlooks issued during the 2018 HWT SFE. The full-period guidance and outlooks for tornadoes on 1 May 2018 are shown for the HREF/SREF calibrated guidance (top-left panel), experimental outlook (top-middle panel), practically perfect hindcast (Hitchens et al. 2013; top-right panel), HREF STP calibrated guidance (bottom-left panel), experimental outlook (repeated; bottom-middle panel), and practically perfect hindcast (repeated; bottom-right panel). The observed tornado reports (upside-down red triangles) are overlaid as a reference for subjective verification.

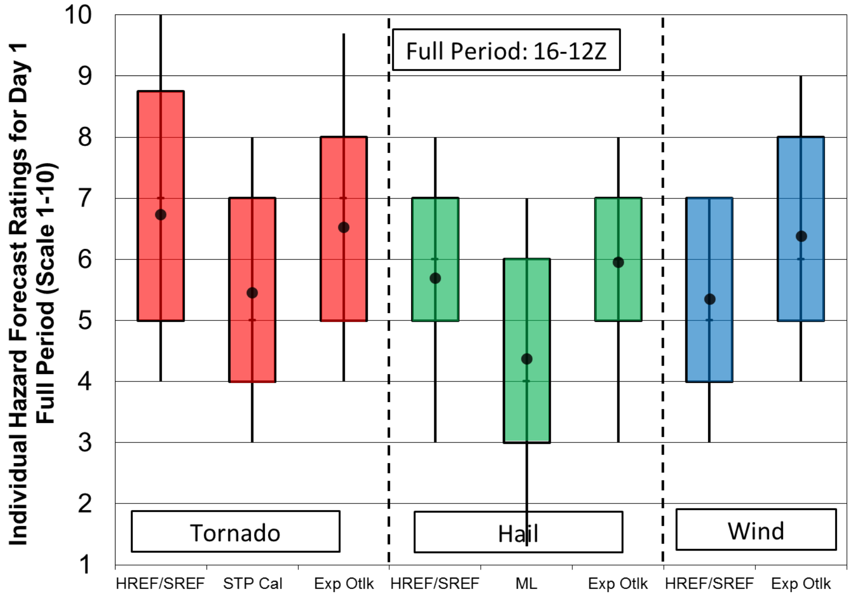


Figure 21 Distribution of subjective ratings (1-10) for calibrated hazard guidance and experimental probabilistic outlooks for tornado (red; left), hail (green; middle), and wind (blue; right) for Day 1 (i.e., 1600-1200 UTC). The boxes span the interquartile range while the whiskers extend to the 10th and 90th percentiles. The horizontal dash (-) indicates the median rating, and the circle (●) indicates the mean rating.

2) DAY 1 4-H FORECASTS

Experimental, overlapping 4-h probabilistic outlooks were also generated for the Day 1 period during SFE2018. First-guess 4-h probabilities for tornadoes, hail, and wind were generated using the temporal disaggregation technique (Jirak et al. 2012) by incorporating the full-period, forecaster-generated hazard outlook to constrain and scale the magnitude and spatial extent of the 4-h HREF/SREF calibrated probabilities. These first-guess probabilities were available during the forecast process and then compared in the next-day evaluation to the forecaster-issued probabilities, providing an indication of how much a forecaster can improve upon the 4-h first-guess guidance. For the 1700-2100 UTC outlook (Fig. 22), which was issued in the morning, forecasters were generally able to improve upon the disaggregated first-guess guidance for hail and wind (i.e., fewer lower rated forecasts) while the tornado forecasts were rated about the same as the first-guess guidance during an overall climatologically below-average period for tornadoes (i.e., the high ratings are largely a result of low/no probabilities with no tornado occurrence). In general, the overlapping distribution of ratings suggests that the guidance is a reasonable first guess that can be used by the forecasters.

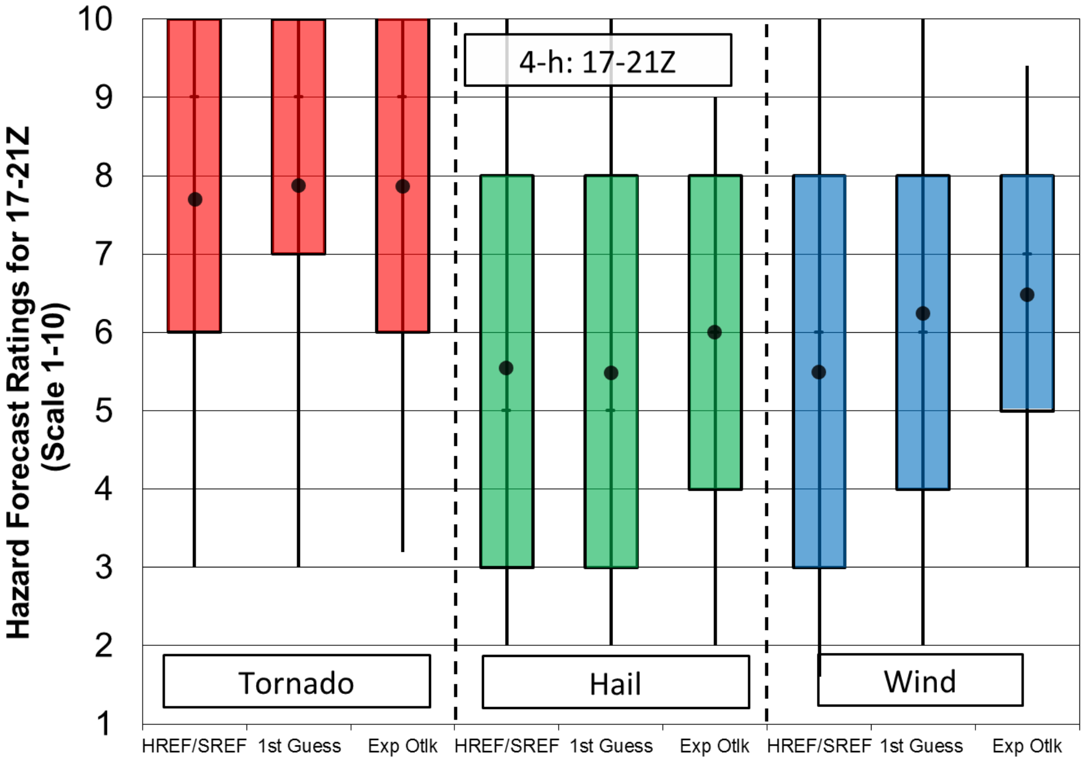


Figure 22 Same as Fig. 21, except for 4-h outlooks valid 1700-2100 UTC for the HREF/SREF calibrated guidance (left), first-guess guidance (middle), and the forecaster-issued outlook (right) for each hazard.

Experimental 4-h outlooks were also generated in the afternoon for the 1900-2300 UTC period (i.e., 0-4 h period). For this period, the experimental outlooks for hail and wind were an improvement over the calibrated and first-guess guidance (Fig. 23) while the experimental tornado outlooks were rated slightly lower overall than the first-guess guidance (again during a relatively quiet period of tornado activity).

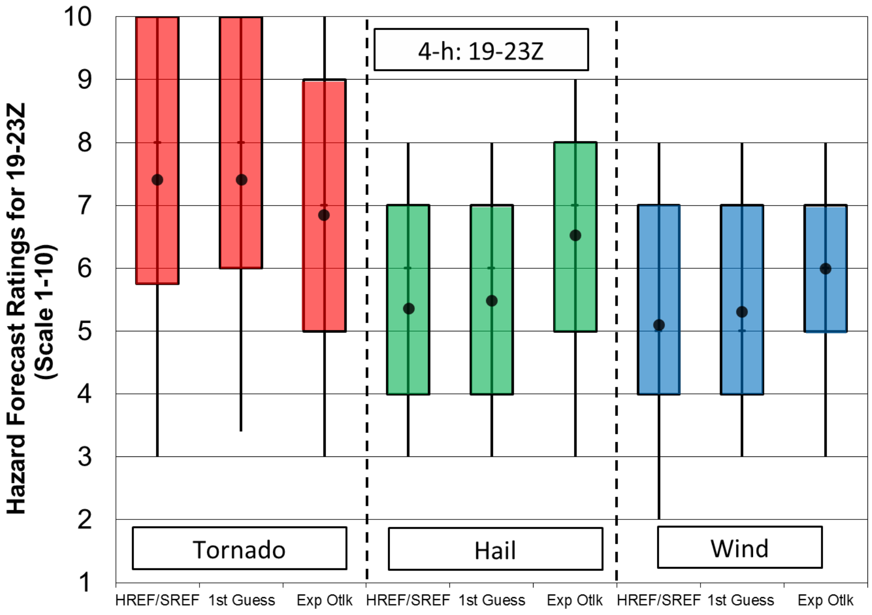


Figure 23 Same as Fig. 22, except for the period valid 1900-2300 UTC.

Finally, experimental 4-h outlooks were generated for the 2100-0100 UTC period. In addition to the HREF/SREF calibrated and first-guess guidance, a morning (i.e., preliminary) forecast was made for 2100-0100 UTC, along with an afternoon update to this forecast period that included the incorporation of the latest CAM guidance (e.g., HRRR, NEWS-e). The HREF/SREF calibrated guidance, first-guess guidance, preliminary forecasts, and final forecasts of tornado, hail, and wind for this period were subjectively rated and compared (Fig. 24). In general, the forecaster was able to improve upon the first-guess guidance in the preliminary forecasts for this period. Updating the forecasts in the afternoon with the latest observations and CAM guidance also generally resulted in further improvements in the experimental outlooks, as evidenced by the higher subjective ratings.

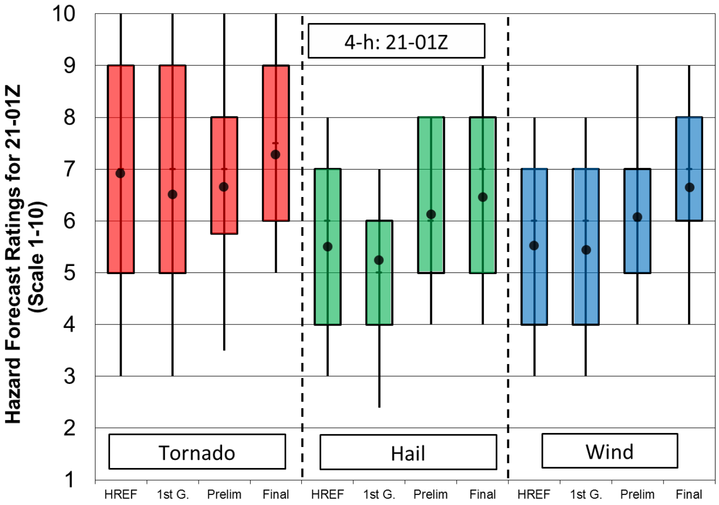


Figure 24 Same as Fig. 3, except for the period 2100-0100 UTC and including a preliminary and final forecaster-issued outlook.

A new aspect to issuing experimental outlooks this year was dividing the participants into small groups and assigning them to use a specific CAM ensemble in generating the 4-h outlooks. While there are a number of uncontrolled aspects in doing this as part of an experiment, it did allow participants to investigate the individual CAM ensembles in more detail. Overall, the groups that used the HREF and HRRRE in generating the 4-h experimental outlooks were more likely to have the highest rated forecast than any of the other groups (Fig. 25).

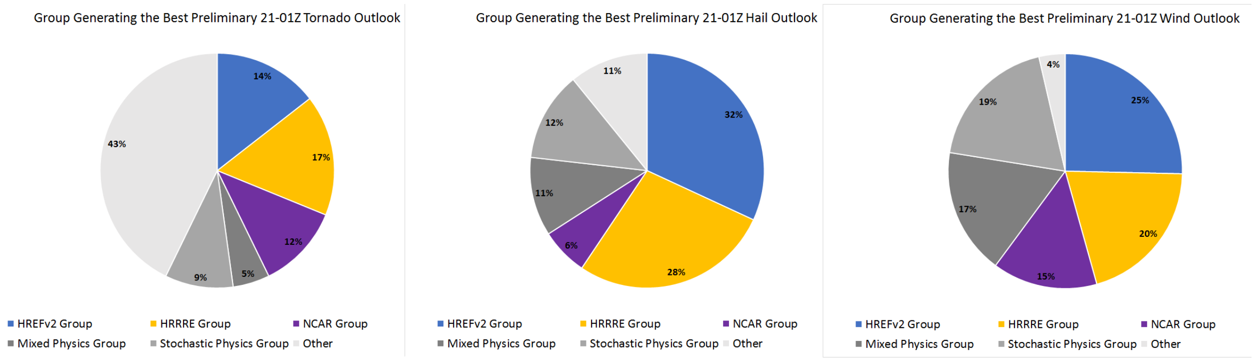


Figure 25 Pie charts for the CAM ensemble groups (HREFv2, HRRRE, NCAR, Mixed Physics, Stochastic Physics and other – no "best" group indicated) with the highest rated preliminary tornado (left), hail (middle), and wind (right) outlooks for 2100-0100 UTC.

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

1) FV3 PHYSICS COMPARISON

CAPS ran multiple convection-allowing configurations of the FV3 model, experimenting with varied microphysics and planetary boundary layer schemes. To compare these subjectively, participants evaluated 2-D storm attribute fields such as simulated reflectivity and UH, 2-D storm environment fields such as CAPE, temperature, and dewpoint, and point soundings to examine vertical structure within the model.

Overall, not much difference was seen between the different boundary layer schemes, in both storm attribute fields such as UH and reliability (Fig. 26a, top) and environmental fields such as thermodynamic variables (Fig. 26a, bottom). Though the EDMF scheme does have a median that is one point higher than the other schemes, it also has a smaller sample size, so it is unknown if this difference is meaningful. When participants were asked which thermodynamic variable showed the largest difference between members with different PBL schemes, most often CAPE was selected, either alone or in conjunction with another variable such as temperature or dewpoint (Fig. 26b).

The sounding evaluation showed similar results to the boundary layer evolution, with a median rating of 6/10 (Fig. 26c). Participants were also asked what features they noticed differed most between the ensemble members, and their answers most often referenced low-level moisture, followed by the temperature and the inversion structure (Fig. 27). In comments, participants often noted that there wasn't much diversity in the soundings and that they were all often too moist.

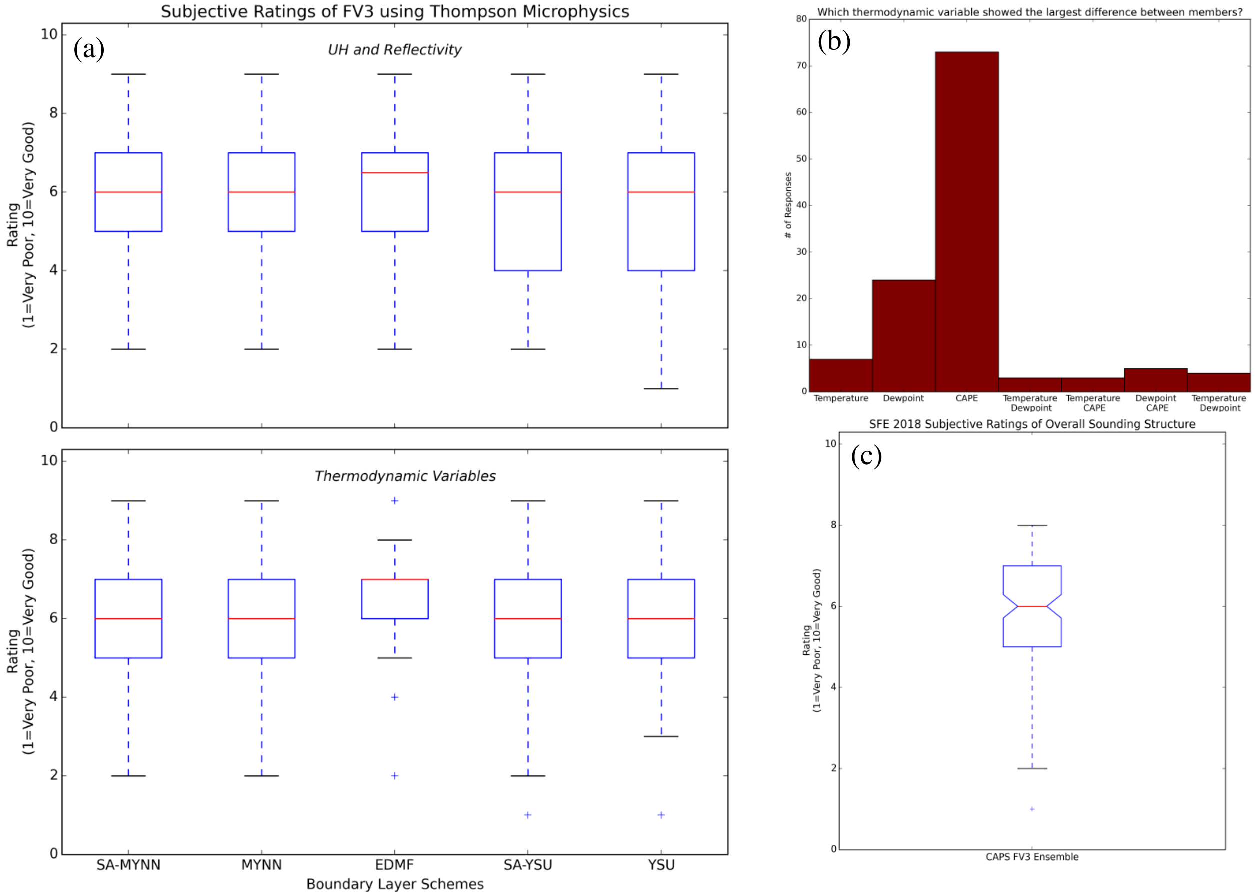


Figure 26 Subjective evaluation results from the FV3 comparison, looking at (a) different PBL schemes, (b) what thermodynamic variable showed the largest differences within the ensemble, and (c) the sounding structure compared to observed soundings.

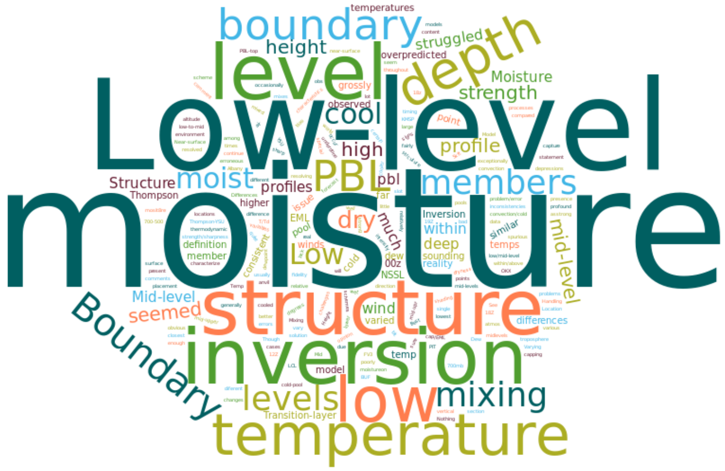


Figure 27 A word cloud highlighting participants' responses to the question, "Which aspects of the soundings differed most within the ensemble? (E.g., inversion structure, low-level moisture, etc.)". The larger a word is in the cloud, the more often it was mentioned.

In comparing the two microphysics schemes tested within the CAPS FV3 ensemble (Thompson and NSSL), participants were shown two members using each of the schemes. One of the members used the MYNN PBL scheme, and one showed the YSU PBL scheme. This part of the evaluation focused on the reflectivity and UH location, reflectivity magnitude, and storm mode, to try to determine which aspects of the storms were being well-captured by each scheme. When looking at the location of storms, most often the Thompson members were indicated as better than the NSSL members (Fig. 10). However, with respect to reflectivity magnitude, the NSSL members were more often better than the Thompson members, or the members were rated as roughly equivalent. Thus, opposite results were found for reflectivity/UH location and reflectivity magnitude. Regarding storm mode, the most common response was that the two schemes were roughly equivalent. When one scheme was favored regarding storm mode, it was the Thompson members.

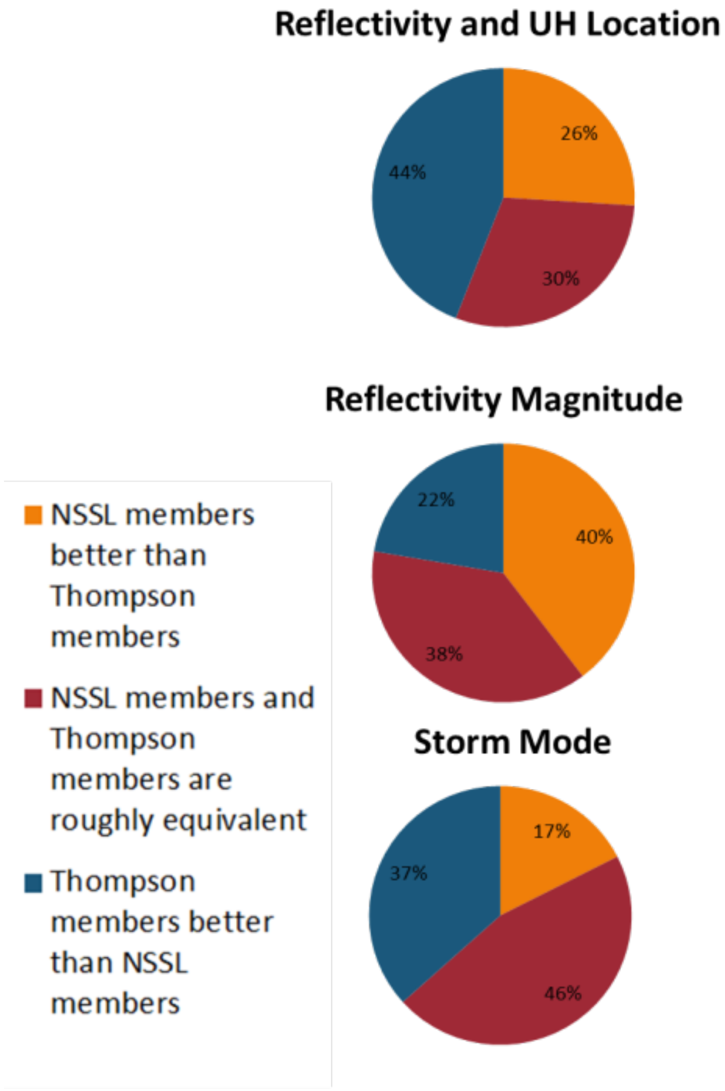


Figure 28 Subjective evaluation responses comparing the two microphysics schemes implemented in the FV3 ensemble.

2) UK MET OFFICE EVALUATION

The UK Met Office provided two versions of their Unified Model (UM) to the 2018 SFE, a global version with grid spacing of approximately 10 km at the mid-latitudes, and a high-resolution version with 2.2 km grid spacing. These two versions were compared to gain more insight into the source of errors in the regional convective-scale model. Environmental fields such as temperature and dewpoint were examined, as were the rain rates for each model. In addition, soundings were also compared between these two configurations.

Similarities between the two models can be seen in the overlapping distributions of the soundings (Fig. 29d), which aligns with comments from the participants. However, the regional soundings had slightly higher ratings overall than the global soundings. The most pronounced differences occurred in the dewpoint (Fig. 29b) and rain rate (Fig. 29c) fields. In both of these fields, the regional model garnered higher ratings than the global model, with the global having many more rankings in the 3/10–5/10 range than the regional model. The temperature ratings (Fig. 29a) had smaller differences between the models than the other fields, but the regional again scored slightly higher than the global model. Comments from participants often cited the dew points as the field that showed the largest difference, and also often discussed the advantage of the higher-resolution model explicitly depicting convection and its influence.

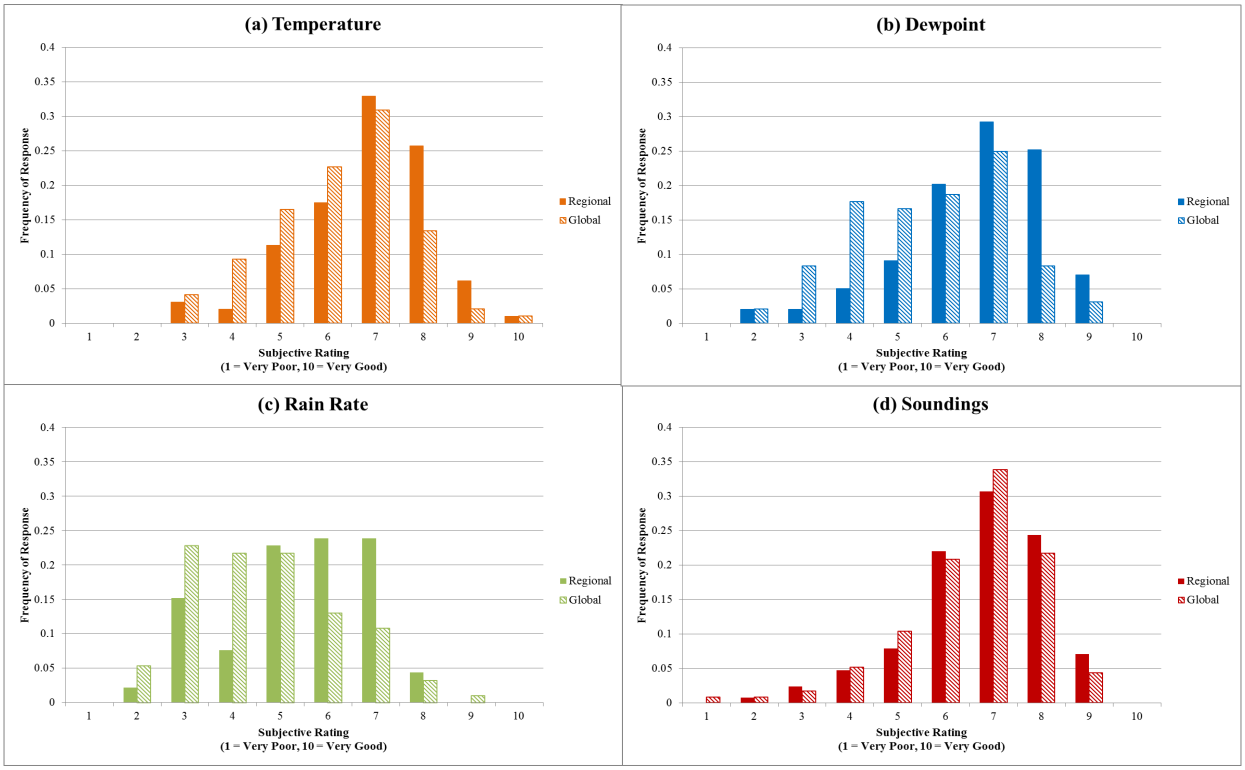


Figure 29 Subjective evaluation responses for the regional (solid bars) and global (striped bars) UM (a) temperature, (b) dewpoint, (c) and rain rate fields, and (d) soundings. Distributions are plotted as the frequency of responses to account for different sample sizes between the regional sounding sample (n=127) and the global sounding sample (n=115).

As mentioned previously, participants noted several occasions where the global and regional looked extremely similar as far as the sounding structure goes, and even noted multiple instances where the global sounding structure was better than the sounding structure from the high-resolution model. This was reflected in participant responses to questions asking specifically about the degree of difference between the soundings (Fig. 30a) and rain rate fields (Fig. 30b,c). More major differences were seen between the magnitude and placement of important features in the rain rate than between the sounding characteristics. This result is somewhat unsurprising given the difference in grid spacing. As in previous years, participants commented on the inversion structure depicted in both versions of the UM, noting that the UM often was able to capture the sharpness and magnitude of the inversion structures.

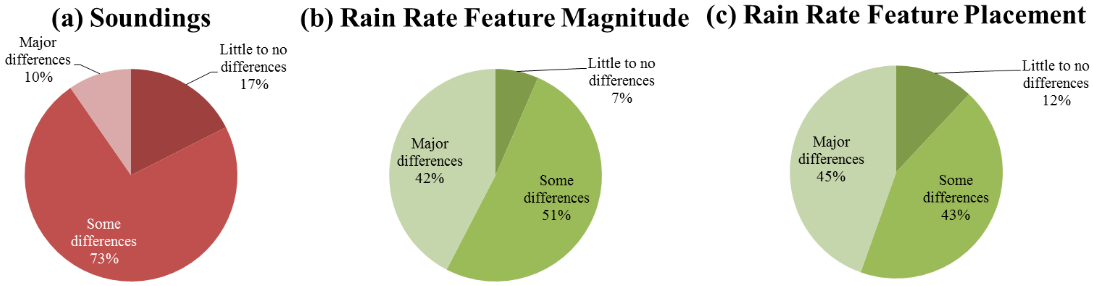


Figure 30 Responses indicating the degree of difference between the regional and global UM configurations regarding (a) soundings, (b) rain rate important feature magnitude, and (c) rain rate important feature placement.

3) STOCHASTIC, SINGLE, and MIXED-PHYSICS EVALUATION

Three WRF-ARW-based ensembles were compared that all had the same perturbed initial conditions and lateral boundary conditions, but different sets of physics. One ensemble had mixed physics, one had single physics, and one had stochastic physics. The physics configurations for the single and stochastic physics ensembles were the same, and were designed based on the HRRR/RAP physics suite. Stochastic perturbations were applied to multiple parameters within the MYNN PBL scheme and Thompson microphysics in the stochastic physics ensemble. Participants subjectively evaluated three severe hazard fields within the experiment: 2–5 km UH, maximum updraft speed, and 10 m wind speed. Objective evaluation of the reflectivity and surrogate severe fields based on UH took place after the experiment. A bug in the code was also discovered post-experiment that caused the stochastic perturbations to be smaller in magnitude than intended, so caution should be exercised in the general applicability of the results in this section.

Reflectivity at each hour was examined using the fractions skill score (FSS; Roberts and Lean 2008), to highlight ensemble performance during the peak afternoon convective period. These forecasts were verified using MRMS reflectivity data. The mixed-physics ensemble performed better at the peak Day 1 convective period (forecast hours 18-26), but the skill drops off more sharply in the overnight periods (Fig. 31a). The single and stochastic ensembles perform quite similarly, with the single physics ensemble reflectivity performing slightly better than the stochastic physics ensemble. The reliability diagram (Fig. 31b) reveals the advantage of increased spread in a mixed-physics ensemble that results in less of an overconfident/underdispersive ensemble when compared to the single- and stochastic-physics ensembles.

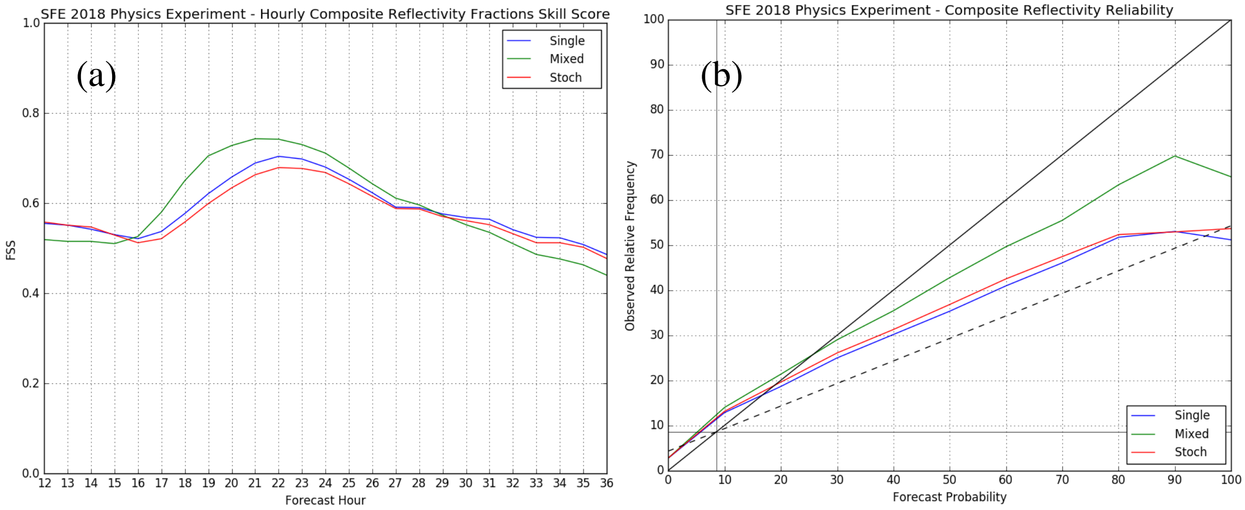


Figure 31 Composite reflectivity results for (a) hourly FSS values and (b) reliability of probabilistic forecasts of reflectivity greater than 40 dBZ within a 20 km radius.

Different UH thresholds were used to test the surrogate severe fields, ranging from 25 m^2s^{-2} to 150 m^2s^{-2}. The highest FSS occurred at thresholds of 50 m^2s^{-2} and 75 m^2s^{-2}, and the mixed physics achieved a higher FSS than the single or the stochastic physics ensembles at any given threshold (Fig. 32). The threshold differences were relatively consistent between the forecast thresholds except for the 25 m^2s^{-2} threshold, which had smaller differences between the mixed and the stochastic physics ensembles than the other thresholds.

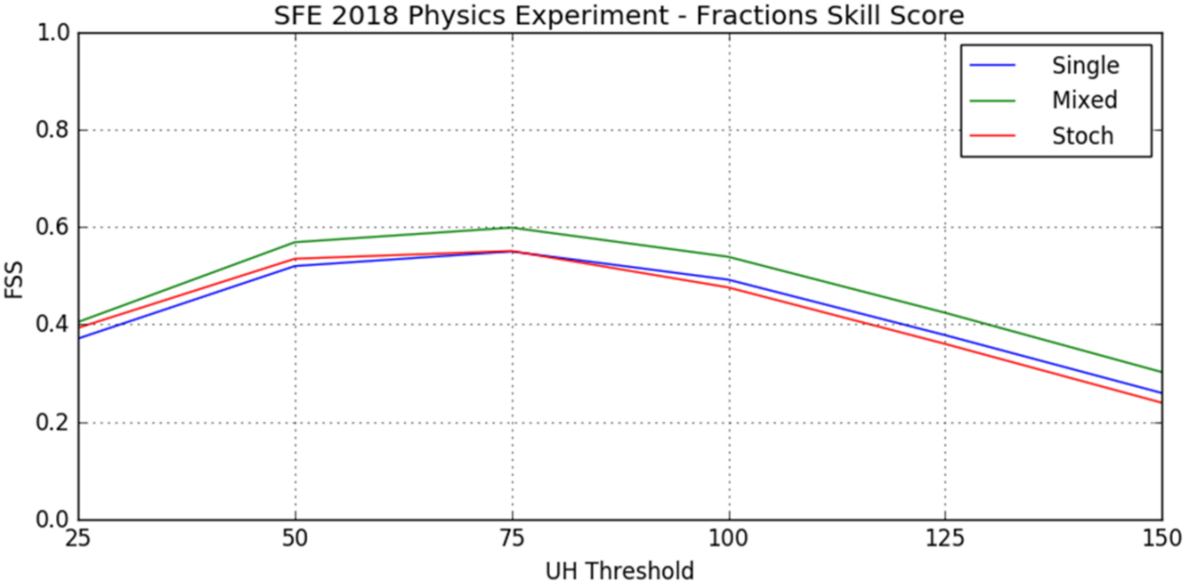


Figure 32 Ensemble FSS for surrogate severe fields generated using different UH thresholds.

Overall, objective differences were relatively small between all three ensembles. This trend was also seen in the subjective evaluations, which showed quite similar rating distributions (Fig. 33), particularly for UH and updraft speed. Mean subjective ratings for UH were actually *highest* in the stochastic physics ensemble, contrary to the objective results, but the differences in the means between all of the ensembles were small enough that this difference is likely insignificant. Participants also remarked on the similarity of the ensembles in the comment section of the evaluation, with many comments first describing the behavior of all of the ensembles and later commenting on the specifics that may have led them to give slightly different ratings to each ensemble. The largest differences between the subjective evaluations occurred with 10 m wind forecasts. For these forecasts, the mixed physics performed better than the single and stochastic physics, with a higher mean and 75th percentile value.

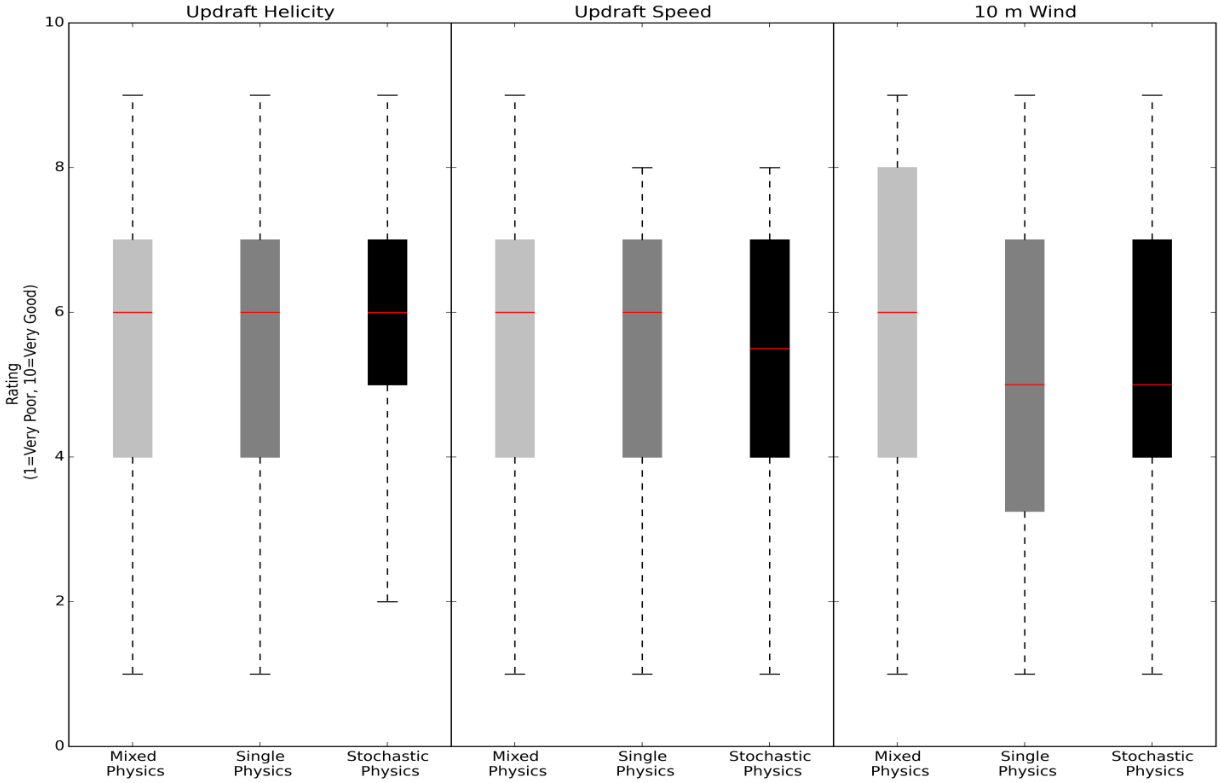


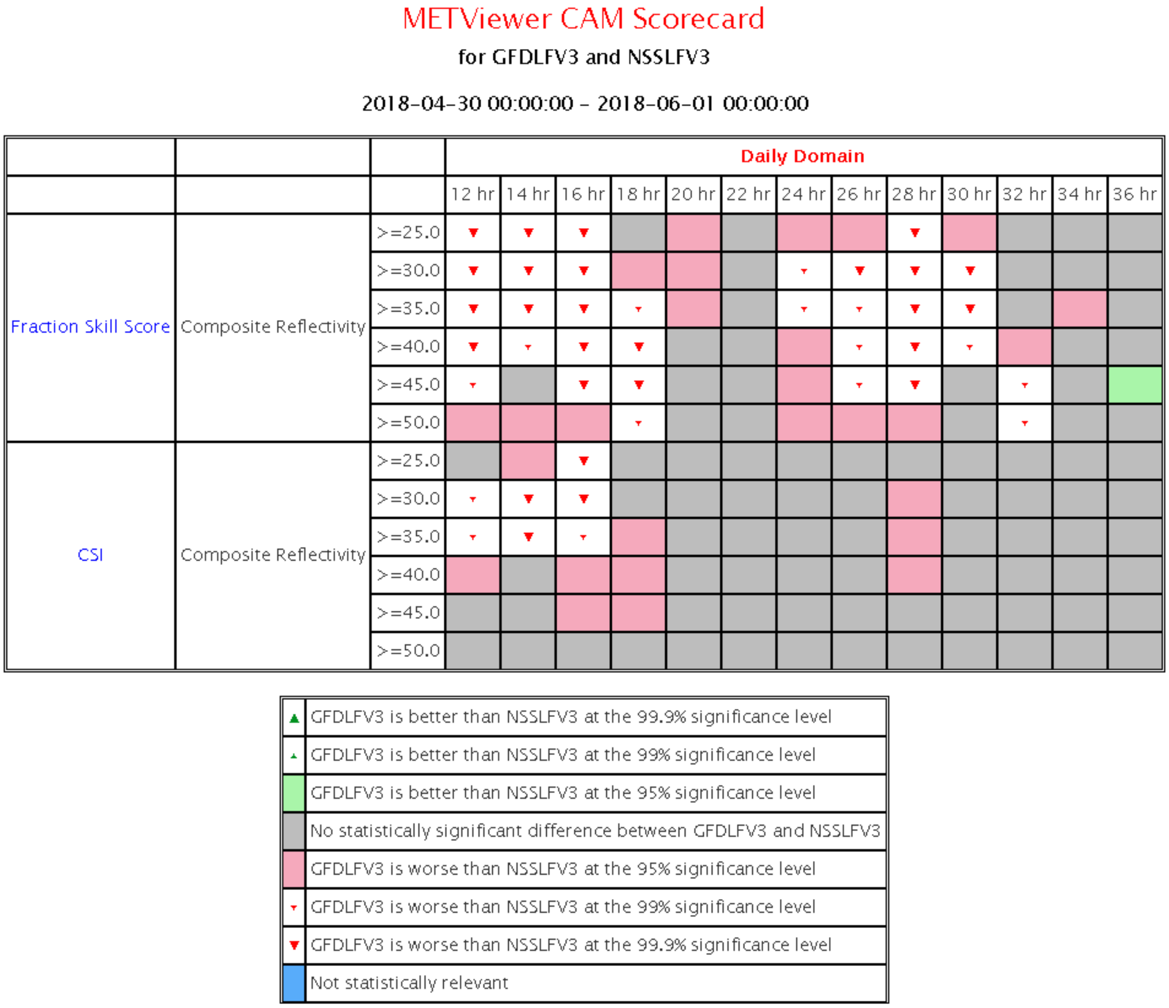
Figure 33 Subjective evaluation results for the ensemble physics experiment, looking at hourly maximum fields. Median values are highlighted by the red lines.

4) STOCHASTIC MICROPHYSICS COMPARISON

One member of the 2018 CLUE implemented a stochastically perturbed microphysics scheme. Reflectivity, UH, and simulated satellite were examined in this comparison, which was only done on a small subset of the SFE 2018 cases (*n* = 10 days) due to time constraints. Participants saw very small differences between the stochastically perturbed and the non-perturbed member, in the reflectivity and UH fields, as well as in the temperature and dewpoint fields. The simulated satellite fields showed that both members were producing simulated clouds that were about the correct size. Results from this year will be used to inform the design of the stochastic perturbations in the microphysics scheme, and a more thorough investigation of the stochastically perturbed microphysics scheme is planned for SFE 2019.

5) CAM SCORECARD

In collaboration with the Developmental Testbed Center (DTC), a convection-allowing model (CAM) scorecard verification technique was implemented in SFE 2018. The deterministic models initially tested were the NSSL-FV3, GFDL-FV3, and HRRRv3, and the initial ensemble systems were the HREFv2 and the HRRRE. Fields included in this first iteration of the CAM scorecard were simulated reflectivity and surrogate severe fields. The surrogate severe fields used a sigma value of 120 km in the Gaussian smoother and multiple UH thresholds. Initial verification metrics included the CSI and the FSS. Results for individual metrics were available online in graph and table formats, and at the end of the experiment, summary scorecards (Fig. 34) encompassing the entire SFE were presented to participants at the morning forecast discussion. Ongoing work is underway to add environmental fields such as temperature, dewpoint, and wind into the scorecard. Additional work on evaluating how best to determine model thresholds of variables such as UH is needed, given differing model climatologies. From the threshold approach used in SFE 2018, future experiments will move to a percentile approach. Finally, efforts to generate the scorecard real-time are planned for SFE 2019.



METViewer CAM Scorecard

for GFDLFV3 and NSSLFV3

2018-04-30 00:00:00 - 2018-06-01 00:00:00

			Daily Domain												
			12 hr	14 hr	16 hr	18 hr	20 hr	22 hr	24 hr	26 hr	28 hr	30 hr	32 hr	34 hr	36 hr
Fraction Skill Score	Composite Reflectivity	>=25.0	▼	▼	▼	▒	🟥	▒	🟥	🟥	▼	🟥	▒	▒	▒
		>=30.0	▼	▼	▼	🟥	🟥	▒	▾	▼	▼	▼	▒	▒	▒
		>=35.0	▼	▼	▼	▾	🟥	▒	▾	▾	▼	▼	▒	🟥	▒
		>=40.0	▼	▾	▼	▼	▒	▒	🟥	▾	▼	▾	🟥	▒	▒
		>=45.0	▾	▒	▼	▼	▒	▒	🟥	▾	▼	▒	▾	▒	🟩
		>=50.0	🟥	🟥	🟥	▾	▒	▒	🟥	🟥	🟥	▒	▾	▒	▒
CSI	Composite Reflectivity	>=25.0	▒	🟥	▼	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
		>=30.0	▾	▼	▼	▒	▒	▒	▒	▒	🟥	▒	▒	▒	▒
		>=35.0	▾	▼	▾	🟥	▒	▒	▒	▒	🟥	▒	▒	▒	▒
		>=40.0	🟥	▒	🟥	🟥	▒	▒	▒	▒	🟥	▒	▒	▒	▒
		>=45.0	▒	▒	🟥	🟥	▒	▒	▒	▒	▒	▒	▒	▒	▒
		>=50.0	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒

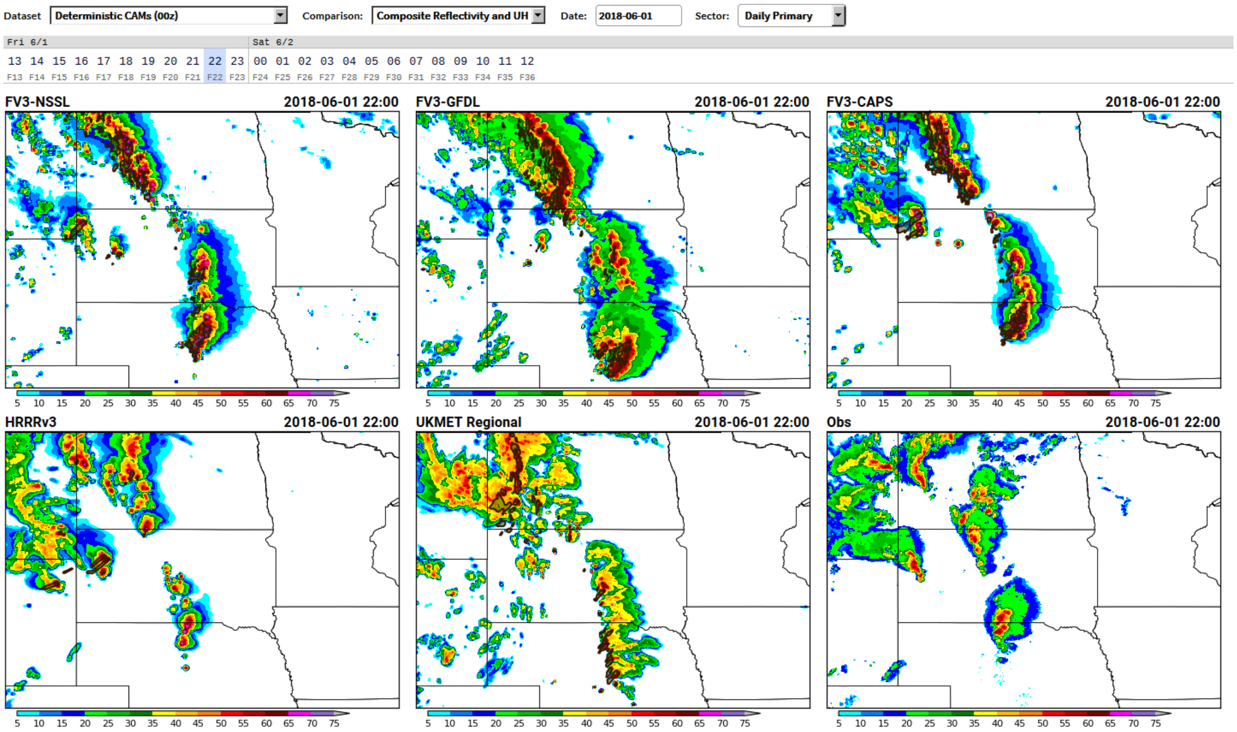
▲	GFDLFV3 is better than NSSLFV3 at the 99.9% significance level
▴	GFDLFV3 is better than NSSLFV3 at the 99% significance level
🟩	GFDLFV3 is better than NSSLFV3 at the 95% significance level
▒	No statistically significant difference between GFDLFV3 and NSSLFV3
🟥	GFDLFV3 is worse than NSSLFV3 at the 95% significance level
▾	GFDLFV3 is worse than NSSLFV3 at the 99% significance level
▼	GFDLFV3 is worse than NSSLFV3 at the 99.9% significance level
🟦	Not statistically relevant

Figure 34 The CAM scorecard comparing the GFDL FV3 and NSSL-FV3 simulated composite reflectivity fields.

d) Model Evaluations – Severe Hazards Desk

1) EVALUATION of DETERMINISTIC CAMS (credit: B. Gallo and I. Jirak)

For the 2018 HWT SFE, there were several global versions of FV3 run with convection-allowing nests over the CONUS. Three different 0000 UTC experimental versions of FV3 with ~3-km grid spacing over the CONUS were examined and compared to the now-operational HRRRv3 and the UK Met Office convection-allowing version of the UM to gauge performance at convective scales. An example of the reflectivity forecasts from these deterministic CAMs is provided in Figure 35.



Dataset: Deterministic CAMs (00z) Comparison: Composite Reflectivity and UH Date: 2018-06-01 Sector: Daily Primary

Figure 35 Example of subjective comparison plots used for rating CAM performance at convective scales valid at 2200 UTC on 1 June 2018. The 21-h forecasts of composite reflectivity are shown for the a) FV3 NSSL (upper-left panel), b) FV3-GFDL (upper-middle panel), c) FV3-CAPS (upper-right panel), d) HRRRv3 (lower-left panel), and e) UK Met Office UM. The bottom-right panel shows the observed composite reflectivity at 2100 UTC on 1 June 2018.

Top-of-the-hour composite reflectivity fields from the deterministic CAMs were subjectively evaluated by SFE participants for correspondence with timing, intensity, coverage, and mode of observed composite reflectivity from 13-12Z daily and assigned a rating on a scale of 1-10, with 10 being best. The HRRRv3 was the highest-rated deterministic CAM during the five-week SFE (Fig. 36) while the GFDL version of the FV3 was the lowest rated CAM during the SFE, largely owing to an overforecast of coverage and intensity of convective storms (see Figure 35 as an example). The other three deterministic CAMs (i.e., FV3-NSSL, FV3-CAPS, and UK Met Office UM) fell in the middle of the pack with similar mean and median ratings (i.e., ~5/10).

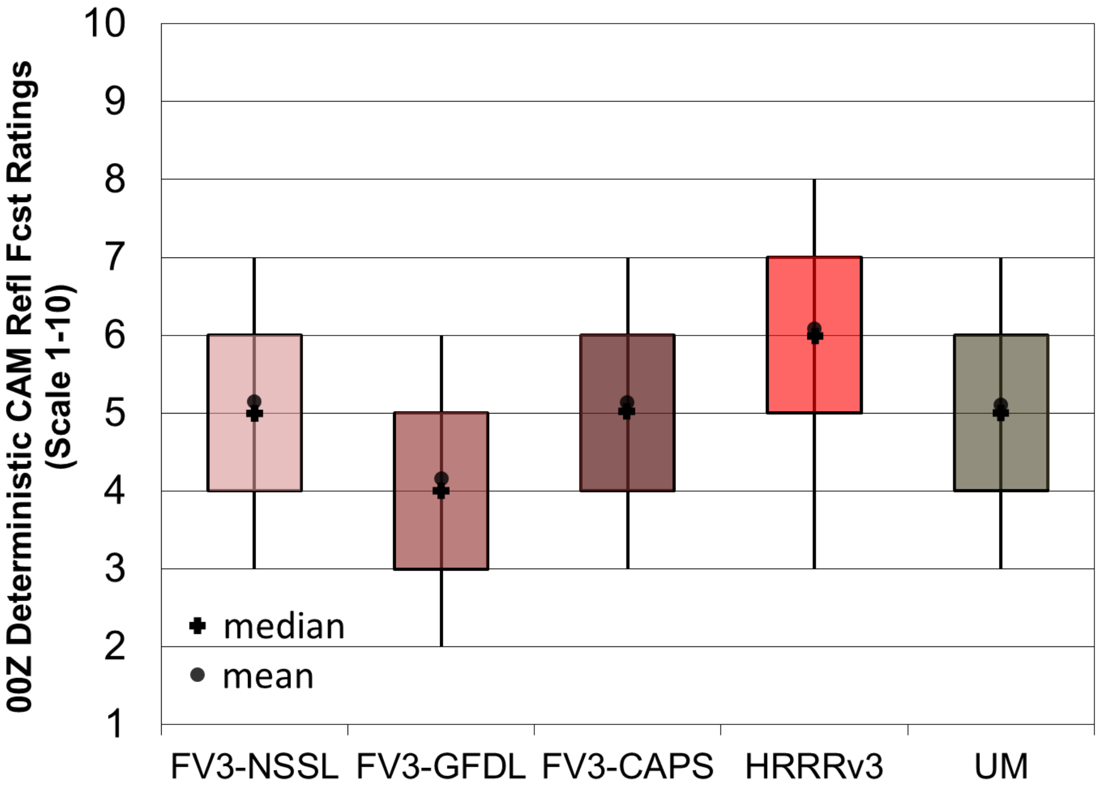


Figure 36 Box-and-whiskers plot of subjective ratings (1-10) for deterministic CAM reflectivity forecasts from 0000 UTC during the five-week 2018 HWT SFE. The boxes represent the interquartile range, and the whiskers represent the 10th and 90th percentiles. The crosses represent the median ratings, and the circles represent the mean ratings.

Model climatologies of UH differed drastically over the course of the experiment. At both the native grid resolution (Fig. 37, dashed lines) and a coarsened 80-km grid resolution (Fig. 37, solid lines), FV3 UH values were consistently larger than those of the HRRRv3 at a given percentile, often by nearly an order of magnitude. This trend was noted first in SFE 2017, and continues for SFE 2018. These results again highlight the need to use UH percentiles rather than fixed thresholds to ensure fair comparisons between dynamical cores and prevent a skewing of verification scores due to systemic climatological differences between models.

To measure severe weather forecasting skill, surrogate severe fields were calculated at 100 different percentiles and 53 different σ values for the deterministic FV3 versions and the HRRRv3, and then verified against LSRs. For each of these percentile-σ sigma combinations, contingency tables were constructed at 20 different percentage thresholds, starting at 2% and then in 5% increments from 5% to 95%. Each of these points was then plotted on a performance diagram (Roebber 2009; Fig. 38) to get an overall view of the model performance. Across most thresholds, the HRRRv3 outperforms all FV3 configurations in terms of CSI. Of the FV3 versions, the NSSL FV3 and the CAPS FV3 are often close, with the NSSL-FV3 slightly outperforming the CAPS FV3 in terms of CSI. The GFDL-FV3 performs worse than the other FV3 configurations shown.

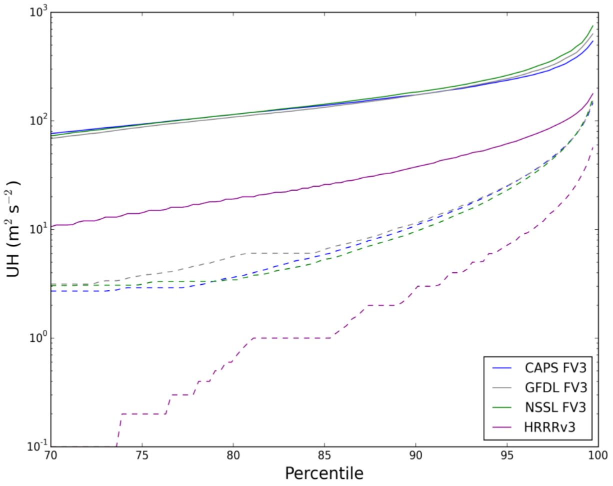


Figure 37 Model UH climatologies for the FV3 versions and HRRRv3 during SFE 2018. Solid lines are the climatologies at the 80-km grid used to construct surrogate severe fields, and dashed lines are the climatologies at the native grid resolutions (~3 km).

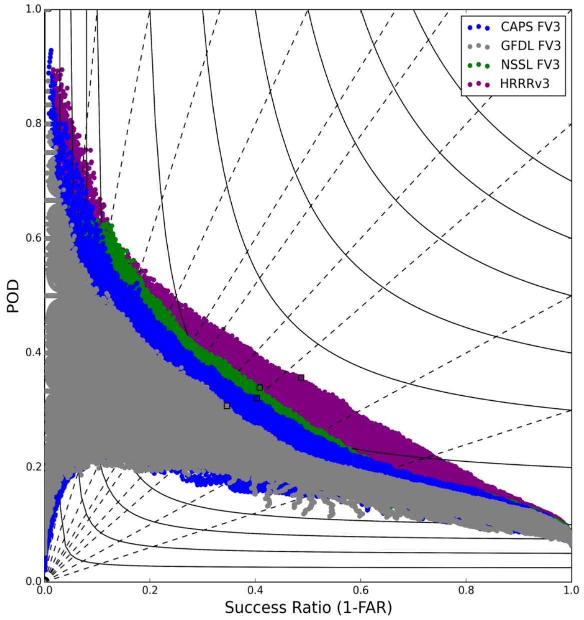


Figure 38 Performance diagram summary of three deterministic FV3 models and HRRRv3. Solid lines are lines of constant CSI, and dashed lines are lines of constant bias.

These objective results match participants' impressions as conveyed by the subjective evaluations (Fig. 36). For example, participants often noted that the GFDL-FV3 was creating too many areas of intense reflectivity and UH compared to observations. While participants often noted that the models were behaving similarly, the GFDL-FV3 was often singled out as having the wrong mode (i.e., developed convection into a linear system far too early) or being too convectively active overall. It should also be noted that the UH was displayed once a certain threshold of UH was exceeded – some participants indicated that the visualization thresholds should likely be adjusted to account for climatology as was done in the objective verification. However, the relatively consistent climatologies between the FV3 members seen in Figure 37 suggest that the subjective results showing a lower rating for the GFDL-FV3 compared to the NSSL-FV3 and the CAPS-FV3 are not due to climatology alone.

2) CLUE: 0000 UTC CAM ENSEMBLES (credit: I. Jirak)

Several experimental CAM ensembles initialized at 0000 UTC were compared to the operational HREF, which serves as the performance baseline for experimental CAM ensembles being considered for operational implementation. The subjective component of the evaluation examined ensemble forecasts (i.e., ensemble maximum and neighborhood probabilities) of hourly maximum fields (HMFs) of UH, updraft speed, and 10-m wind speed relative to LSRs of hail, wind, and tornadoes. The HREF was compared to ensemble subsets from the 2018 CLUE with advanced ensemble-based data assimilation: HRRRE, NCAR EnKF, CAPS EnKF, and the OU MAP Hybrid ensemble system.

An example of how these subjective comparisons were conducted during the 2018 HWT SFE is shown in Fig. 39 for 1 May 2018. In the 24-hour ensemble probability plots of updraft helicity, the HREF (Fig. 39, top/bottom left panels) best captured the highest density of severe weather reports within the area of the highest UH probabilities, resulting in higher subjective ratings from participants.   In contrast, the HRRRE (Fig. 39, top middle panel) UH probabilities are offset to the northeast of the where the severe weather occurred, garnering lower subjective ratings from participants for this forecast. This type of subjective comparison and evaluation of 0000 UTC CAM ensembles occurred daily during the five-week SFE.

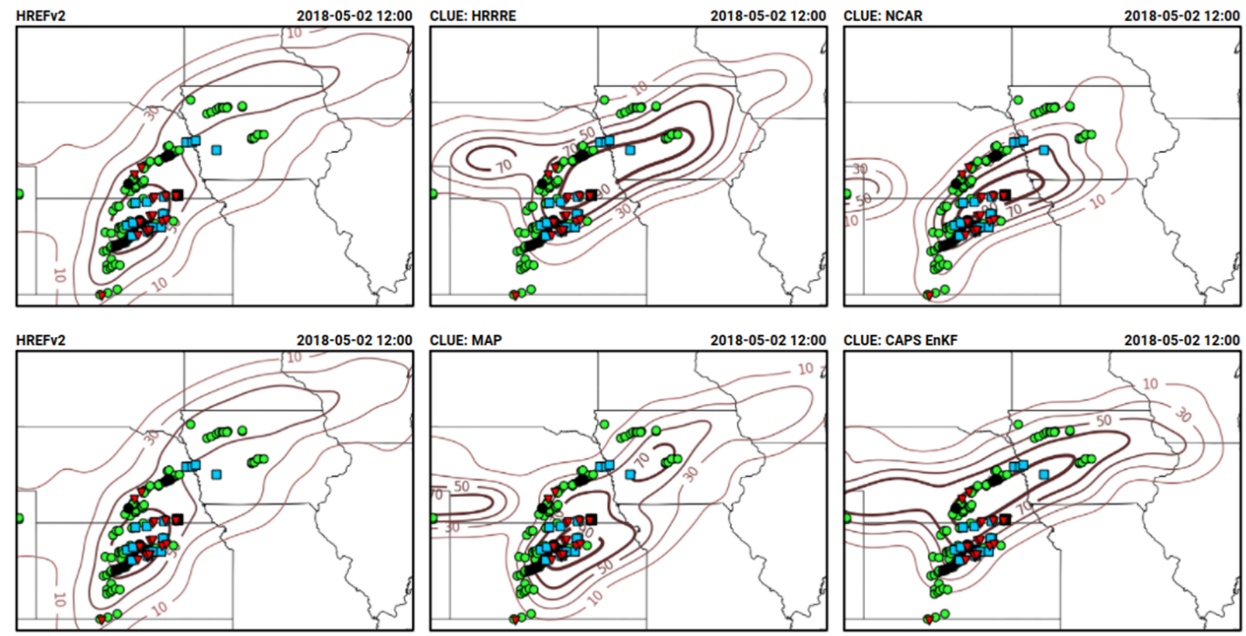


Figure 39 Six-panel comparison plot used to conduct the evaluation of the 0000 UTC CLUE CAM ensembles during the 2018 HWT SFE. The 24-h neighborhood UH probability forecasts exceeding 75 m2/s2 valid for 1 May 2018 are shown for the operational HREFv2 configuration (top-left panel), HRRRE (top-middle panel), NCAR ensemble (top-right panel), HREFv2 (repeated; bottom-left panel), OU MAP ensemble (bottom-middle panel), and CAPS EnKF (bottom-right panel). The observed tornado reports (upside-down red triangles), severe hail reports (green circles), and severe wind reports (blue squares) are overlaid as a reference for subjective verification.

A summary of the distribution of subjective ratings for the 0000-UTC initialized CAM ensembles is shown in Figure 40 for the 2018 HWT SFE. The HREF tended to have the highest subjectively rated forecasts (i.e., highest mean, median, and mode ratings) for severe weather guidance compared to the CLUE ensembles (Fig. 40). The OU MAP and CAPS EnKF ensembles were the next-highest-rated ensembles followed by the HRRRE and NCAR ensemble.

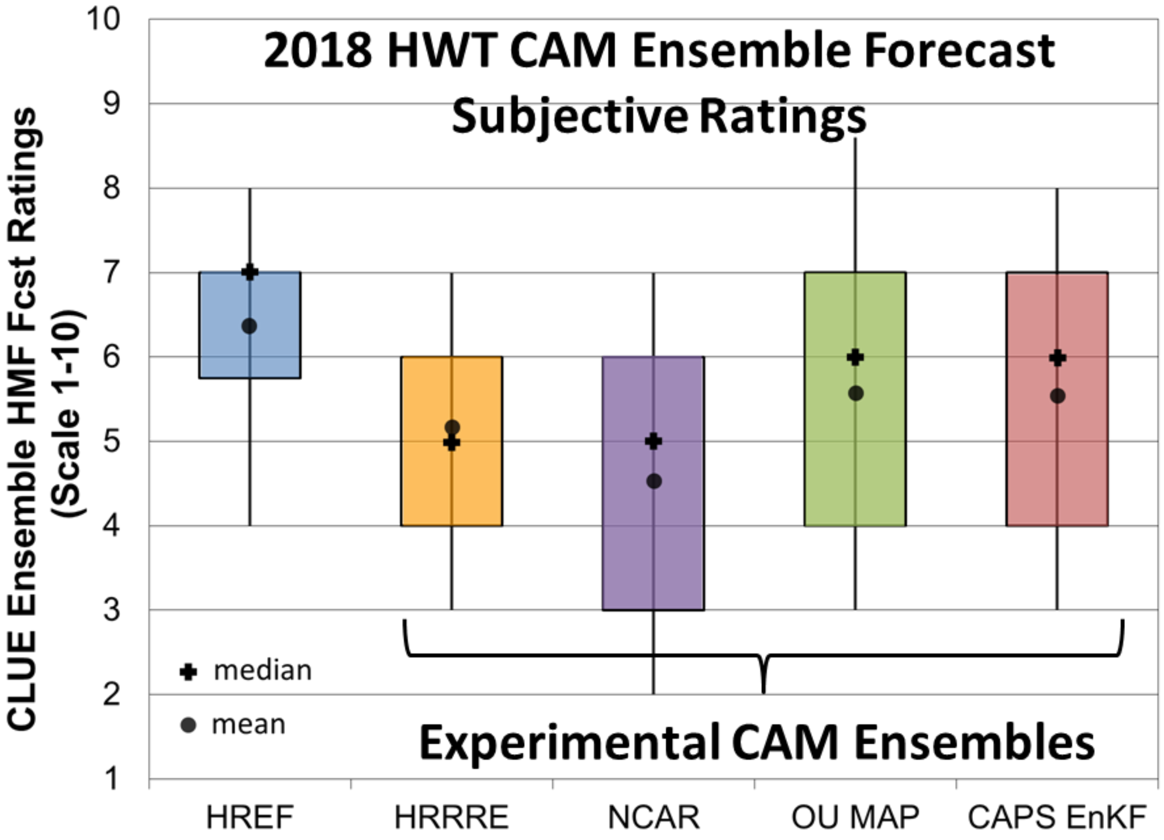


Figure 40 Distributions of subjective ratings (1-10) by SFE participants of HMFs over a mesoscale area of interest for the forecast hours 13-36 for various 0000 UTC CLUE CAM ensembles compared to the HREF.

3) CLUE: 1200 UTC CAM ENSEMBLES (credit: I. Jirak)

Three 1200-UTC initialized CAM ensembles were compared to time-lagged (TL) ensembles generated from HRRRv3 during the 2018 NOAA HWT SFE: the operational HREF, HRRRE, and NCAR ensemble. 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 based on 4-hour hourly maximum field (HMF) forecasts (e.g., UH) for severe weather guidance.

An example of how these subjective comparisons for the 1200 UTC ensembles were conducted during the 2018 HWT SFE is shown in Figure 41 for the 1 May 2018 severe-weather event. In the 4-hour ensemble probability plots of updraft helicity, the HREF (Fig. 41, top/bottom left panels) nicely captures the highest density of severe weather reports within the area of the highest UH probabilities while the HRRRE (Fig. 41, top middle panel) UH probabilities are elongated too far to the northeast of the where the majority of severe weather reports occurred. The NCAR ensemble (Fig. 41, top right panel) is underdispersive leading to an overconfident, yet accurate, forecast, and the HRRR-TL ensembles (Fig. 41, bottom middle/right panels) are also underdispersive/overconfident, but still nicely encompass the reports within the UH probability envelope.

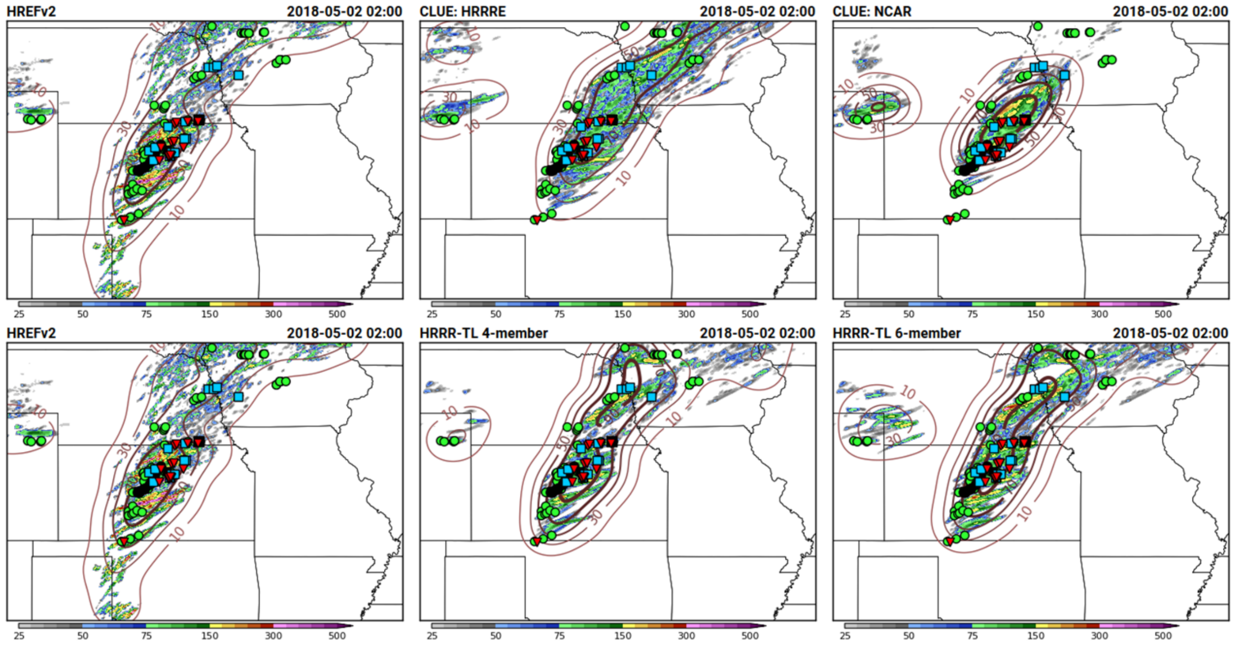


Figure 41 Six-panel comparison plot used to conduct the evaluation of the 1200 UTC CLUE CAM ensembles during the 2018 HWT SFE. The 4-h neighborhood UH probability forecasts exceeding 75 m^2/s^2 valid at 0200 UTC on 2 May 2018 are shown for the operational HREFv2 configuration (top-left panel), HRRRE (top-middle panel) , NCAR ensemble (top-right panel), HREFv2 (repeated; bottom-left panel), HRRR-TL4 ensemble (bottom-middle panel), and HRRR-TL6 ensemble (bottom-right panel). The observed tornado reports (upside-down red triangles), severe hail reports (green circles), and severe wind reports (blue squares) are overlaid as a reference for subjective verification.

Ensemble maximum and neighborhood probabilities of HMF fields (typically UH and 10-m wind speed) were subjectively evaluated by SFE2018 participants for correspondence with severe weather reports from 16-03Z and assigned a rating on a scale of 1-10, with 10 being best. HREF routinely had the highest subjectively rated forecasts (Fig. 42), likely owing to a more diverse ensemble forecast represented by broader probabilistic fields. The formal ensembles, HRRRE and NCAR, generally had lower subjective ratings than the HREF for the 1200-UTC initialized forecasts. The HRRR-TL ensembles fared well in subjective ratings, commonly outperforming the HRRRE, a formal initial-condition ensemble with ensemble DA, using the same model configuration. This result highlights the current usefulness of HRRR-TL ensembles, which are an underutilized resource in NWS severe weather operations, given that the data (i.e., HRRR) already exist operationally and are updated on an hourly basis.

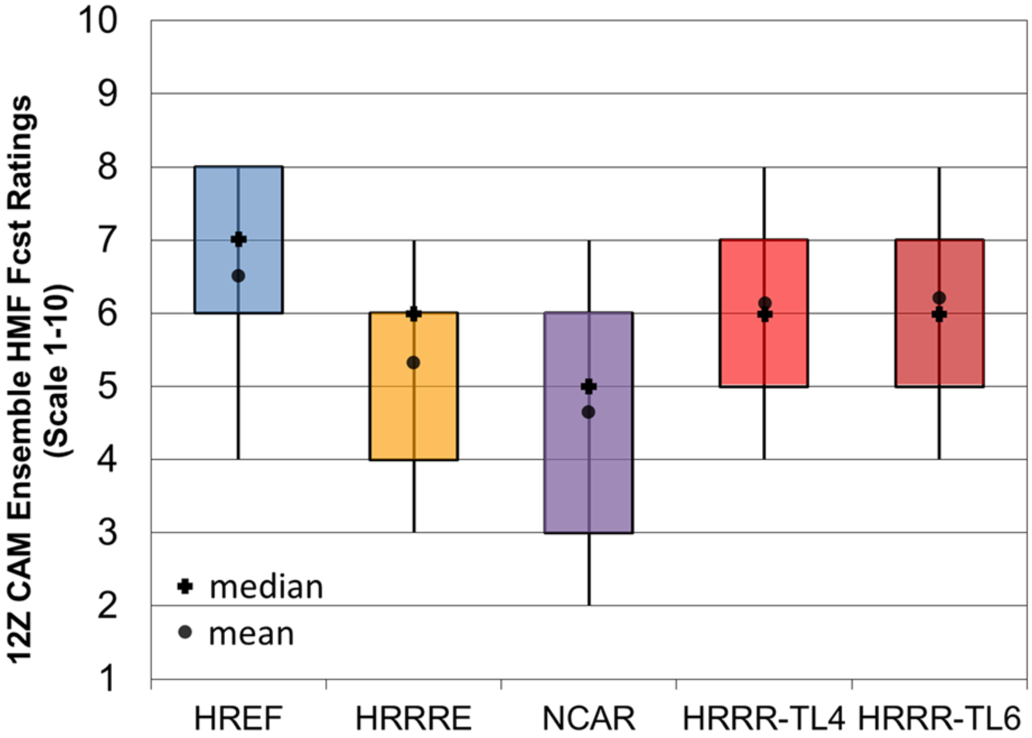
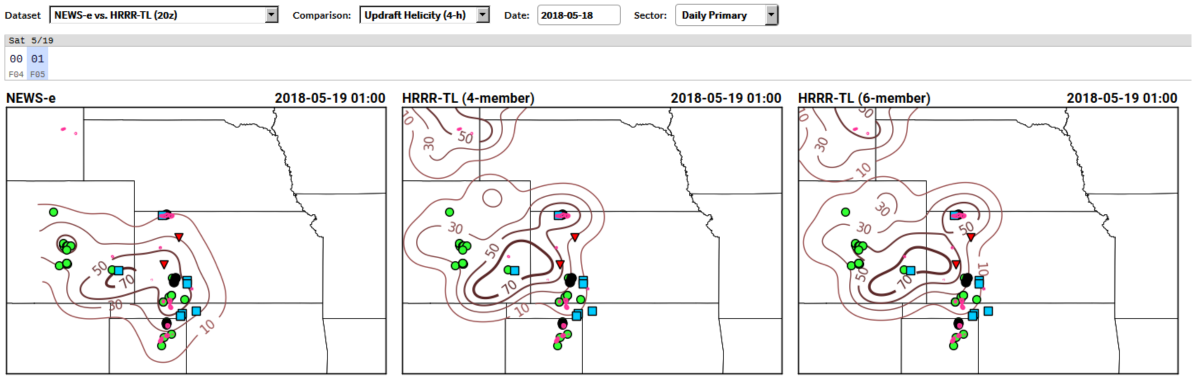


Figure 42 Distributions of subjective ratings (1-10) by SFE participants of HMFs over a mesoscale area of interest for the forecast hours 4-15 for various 1200 UTC CLUE CAM ensembles compared to the HREF.

4) NEWS-E FORECASTS (credit: I. Jirak)

After the first two weeks of the 2018 HWT SFE, it became apparent that forecasts from the NSSL Experimental Warn-on Forecast System for ensembles (NEWS-e) should be subjectively evaluated and compared to operationally available model output. Therefore, during the last three weeks of the 2018 HWT SFE, the 1900 and 2000 UTC-initialized NEWS-e forecasts were compared to HRRR-TL ensembles. The 1900 UTC and 2000 UTC HRRR-TL ensembles were constructed similarly to the 1200 UTC HRRR-TL4 and HRRR-TL6 discussed previously. The HRRR-TL4 consisted of the last four (4) 1-h time-lagged members (i.e., 19, 18, 17, and 16 UTC runs for the 1900 UTC version and 20, 19, 18, and 17 UTC runs for the 2000 UTC version) while the HRRR-TL6 added the 6- and 12-h time-lagged members to HRRR-TL4 (i.e., 19, 18, 17, 16, 13, and 07 UTC runs for the 1900 UTC version and 20, 19, 18, 17, 14, and 08 UTC runs for the 2000 UTC version). These CAM ensembles were evaluated subjectively based on 4-hour hourly maximum field (HMF) forecasts (e.g., updraft helicity – UH) for severe weather guidance.

An example of how these subjective comparisons for the 1200 UTC ensembles were conducted during the 2018 HWT SFE is shown in Fig. 43 for severe weather on 18 May 2018. In the 4-hour ensemble probability plots of updraft helicity, the limited-area domain NEWS-e (Fig. 43, left panel) better captures the severe weather reports within the UH probability envelope while the HRRR-TL ensembles (Fig. 43, top middle and right panels) do not have UH probabilities extending into the Texas panhandle, where severe weather was observed.



Dataset: NEWS-e vs. HRRR-TL (20z)
Comparison: Updraft Helicity (4-h)
Date: 2018-05-18
Sector: Daily Primary

Figure 43 Three-panel comparison plot used to conduct the evaluation of the 1900 and 2000 UTC NEWS-e during the 2018 HWT SFE. The 4-h neighborhood UH probability forecasts exceeding 75 m^2/s^2 valid for 0100 UTC on 19 May 2018 are shown for the 2000 UTC NEWS-e (left panel), HRRR-TL4 (middle panel), and HRRR-TL6 (right panel). The observed tornado reports (upside-down red triangles), severe wind reports (blue squares), severe hail reports (green circles), and observed radar-derived maximum estimated size of hail (MESH; pink swaths) are overlaid as a reference for subjective verification.

Ensemble maximum and neighborhood probabilities of HMF fields (typically UH and 10-m wind speed) from the NEWS-e and HRRR-TL ensembles were subjectively evaluated by SFE2018 participants during the final three weeks for correspondence with severe weather reports from 2000-0100 UTC and assigned a rating on a scale of 1-10, with 10 being best. The NEWS-e forecasts generally had slightly higher subjectively rated forecasts than the HRRR-TL ensembles (Fig. 44), but the updated 2000 UTC NEWS-e run did not typically result in a higher-rated forecast compared to the 1900 UTC run. For the HRRR-TL ensembles, the 6-member version did result in slightly higher-rated forecasts than the 4-member version. For many cases and events, the NEWS-e produced better ensemble forecasts for severe weather events than the HRRR-TL ensembles, highlighting its potential utility for operations. It is worth noting that the small domain of the NEWS-e (i.e., smaller than the daily mesoscale sectors examined during the 2018 HWT SFE) limits its utility for regional or larger-scale severe weather events.

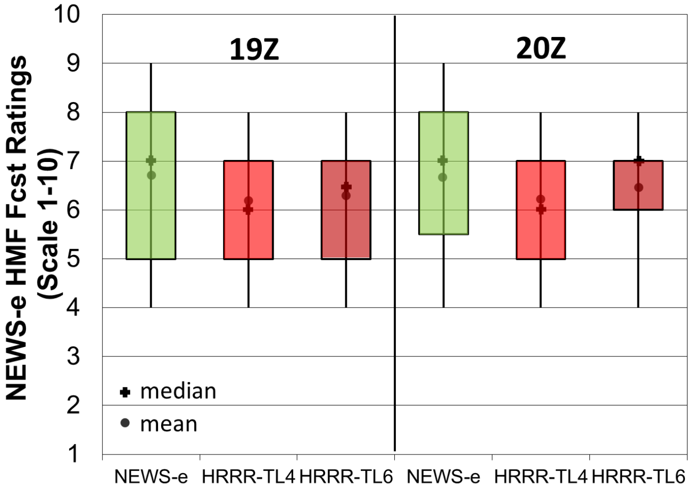


Figure 44 Distributions of subjective ratings (1-10) by SFE participants of HMF forecasts over a mesoscale area of interest for the forecast hours 1-6 for the 1900 UTC runs (on the left) and forecast hours 1-5 for the 2000 UTC runs (on the right) for the NEWS-e, HRRR-TL4, and HRRR-TL6.

5) HREF CONFIGURATIONS (credit: B. Gallo, I. Jirak, and B. Roberts)

The High-Resolution Ensemble Forecast system version 2 (HREFv2) was implemented in the NWS on 1 November 2017 as an operational version of the SPC Storm-Scale Ensemble of Opportunity (SSEO). Thus, the operational HREFv2 serves as a meaningful baseline against which experimental and future CAM ensembles should be compared for consideration of operational implementation. With the July 2018 operational implementation of the extended HRRR runs at 0000, 0600, 1200, and 1800 UTC to 36 hours, there is an opportunity to include the HRRR model as an additional HREF member. Multiple experimental configurations of the HREF were tested and evaluated during the 2018 HWT SFE to help inform how to best configure the next operational version of HREF (i.e., v2.1). The candidate HREFv2.1 configurations included versions that add the extended HRRR runs to the HREFv2, as well as versions that remove some or all of the time-lagged members.

The HREFv2 consists of eight members with half of the members being time-lagged runs. The models are run at ~3-km grid spacing, using a multi-model (WRF-ARW & NMMB), multi-initial condition (NAM & RAP), and multi-physics approach to diversify forecast solutions (Table 6). To provide an evidence-based approach for making configuration decisions at the Environmental Modeling Center (EMC), several potential HREF configurations were examined and evaluated during the 2018 HWT SFE, including the addition of the HRRR (Table 7). The evaluation focused on HREF configurations that would maintain forecast diversity (i.e., multi-core, multi-IC). These HREF configurations (Table 8) included the current HREFv2 configuration for comparison with five other candidate HREF configurations that added the HRRRv3, as well as four versions that removed selected time-lagged members.

*Table 6 HREFv2 member configuration showing ICs/LBCs, planetary boundary layer (PBL) schemes, and microphysics schemes. *SPC uses the 12-h time-lagged NAM Nest while NCO uses the 6-h time-lagged NAM Nest in HREFv2 products.*

Member	ICs/LBCs	PBL	Micro
HRW NSSL	NAM/NAM -6h	MYJ	WSM6
HRW NSSL -12h	NAM/NAM -6h	MYJ	WSM6
HRW ARW	RAP/GFS -6h	YSU	WSM6
HRW ARW -12h	RAP/GFS -6h	YSU	WSM6
HRW NMMB	RAP/GFS -6h	MYJ	F-A
HRW NMMB -12h	RAP/GFS -6h	MYJ	F-A
NAM Nest	NAM/NAM	MYJ	F-A
NAM Nest -12h*	NAM/NAM	MYJ	F-A

Table 7 Same as Table 6, except for the HRRRv3 configuration, as potential addition(s) to the HREFv2.1.

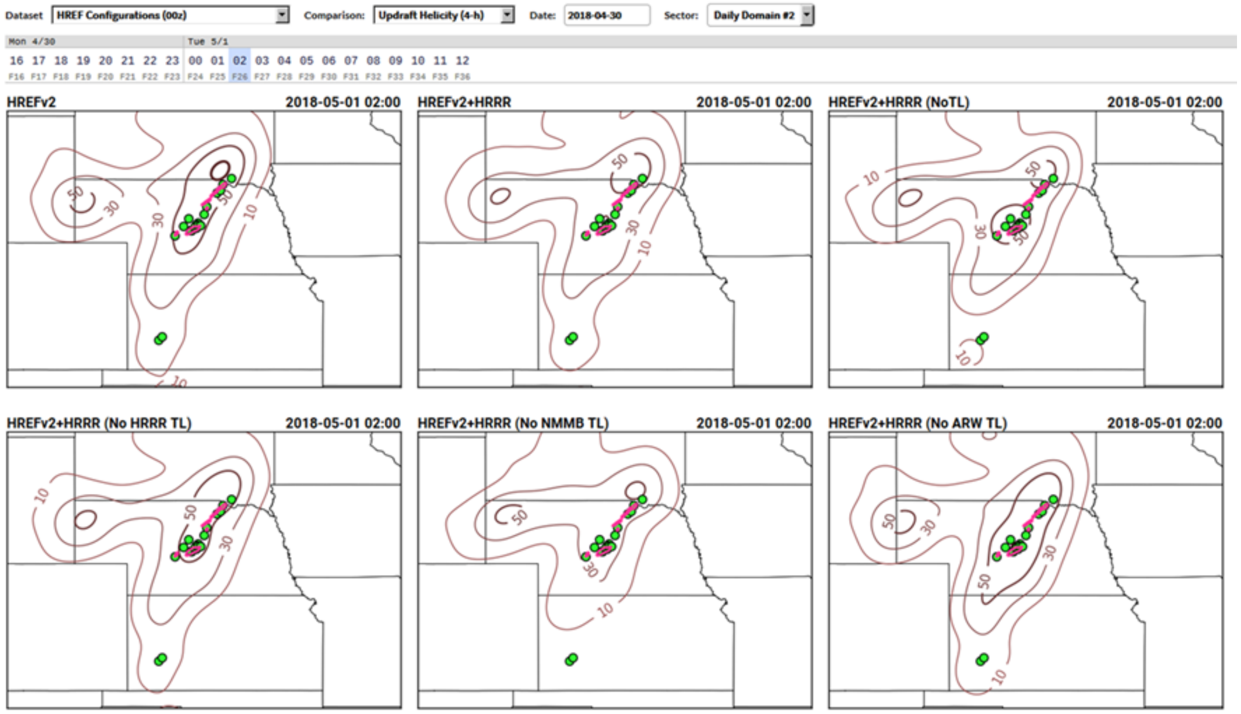
Member	ICs/LBCs	PBL	Micro
HRRRv3	RAP/RAP -3h	MYNN	Thompson
HRRRv3 -6h	RAP/RAP -3h	MYNN	Thompson

Table 8 Different HREF configurations explored during the 2018 HWT SFE. Left column includes the name of the configuration, including time-lagged (TL) members, a description of the configuration, and the total number of ensemble members.

HREF Config	Description	#
HREFv2	Current Config (Table 1)	8
HREFv2+HRRR	Add HRRR & HRRR TL	10
HREFv2+HRRR (No TL)	Remove **all** TL members	5
HREFv2+HRRR (No HRRR TL)	Remove HRRR TL member	9
HREFv2+HRRR (No NMMB TL)	Remove NMMB TL members	8
HREFv2+HRRR (No ARW TL)	Remove ARW TL members	7

Forecasts from the different HREF configurations were available for next-day evaluation during the 2018 HWT SFE, providing an opportunity for subjective comparisons among the configurations with regard to providing severe weather guidance. The HWT SFE participants examined the forecasts from the different HREF configurations using a multi-panel plot with observational overlays. Then, the participants provided a subjective rating of the forecasts based on their assessment of the utility of this guidance for a severe weather forecaster. Overall, the forecasts from the different HREF configurations appeared qualitatively similar on most days. While there were some differences among the forecasts, a careful examination was typically required.

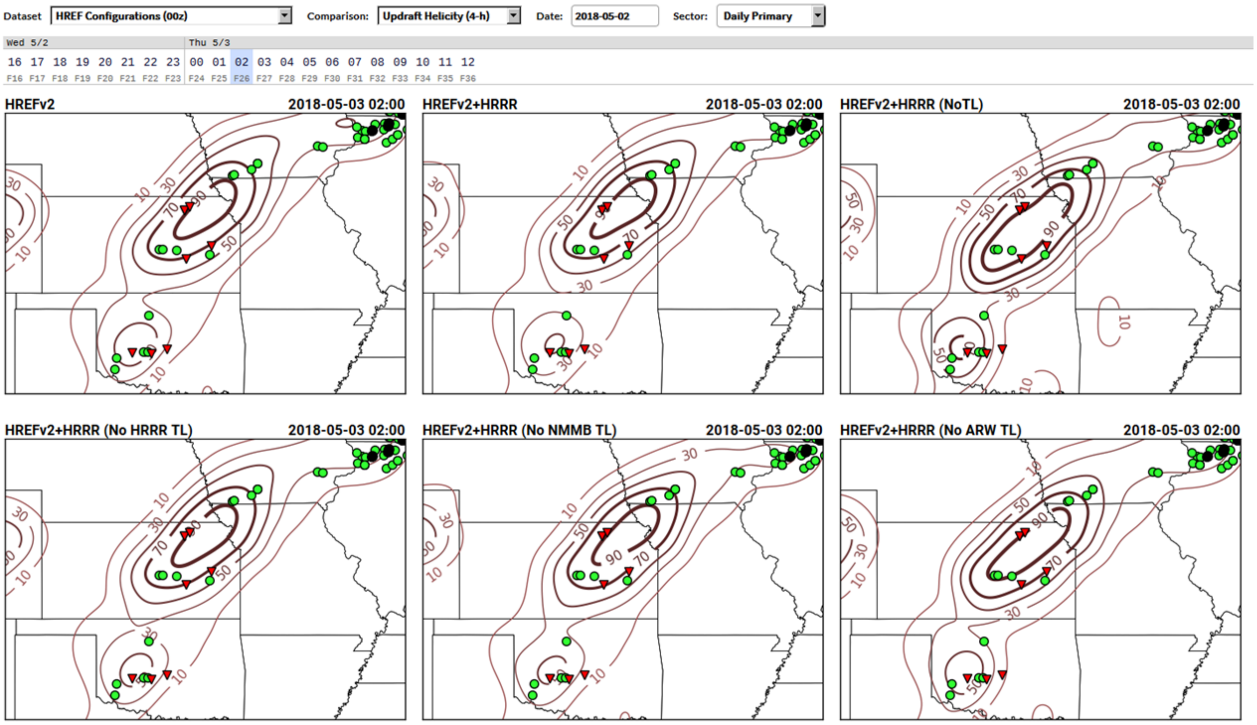
However, on the first day of the 2018 HWT SFE, there were notable differences in the forecasts from the HREF configurations (Fig. 45). In the 26-hour forecast valid at 0200 UTC on 1 May 2018, the forecast from the HREFv2+HRRR (No ARW TL) (Fig. 45, bottom right panel) was rated higher by most participants than the HREFv2+HRRR (No NMMB TL) (Fig. 45, bottom middle panel) forecast. The HREFv2+HRRR (No ARW TL) better captures the axis of severe hail across central Nebraska within higher UH probabilities and also has an extension of low UH probabilities into southwest Kansas, where isolated severe hail was reported.



Dataset: HREF Configurations (00z)
Comparison: Updraft Helicity (4-h)
Date: 2018-04-30
Sector: Daily Domain #2

Figure 45 Six-panel comparison plot used to conduct the evaluation of the 0000 UTC HREF configurations during the 2018 HWT SFE. The 4-h neighborhood UH probability forecasts exceeding 75 m^2/s^2 valid for 0200 UTC on 1 May 2018 are shown for the current HREFv2 configuration (top-left panel) , HREFv2+HRRR (top-middle panel) , HREFv2+HRRR (No TL) (top-right panel), HREFv2+HRRR (No HRRR TL) (bottom-left panel), HREFv2+HRRR (No NMMB TL) (bottom-middle panel), and HREFv2+HRRR (No ARW TL) (bottom-right panel). The observed severe hail reports (green circles) and observed radar-derived maximum estimated size of hail (MESH; pink swaths) are overlaid as a reference for subjective verification.

Other representative examples from the 2018 HWT SFE are shown in Figures 46 and 47. On 2 May 2018, the forecasts from the HREFv2+HRRR (No TL) (Fig. 46; top-right panel) were subjectively rated higher (primarily for higher UH probabilities in southwest Oklahoma) by most participants than forecasts from the other HREF configurations. More typical, however, were the forecasts for 23 May 2018, where all HREF configurations generated very similar forecasts of severe wind potential (Fig. 47).

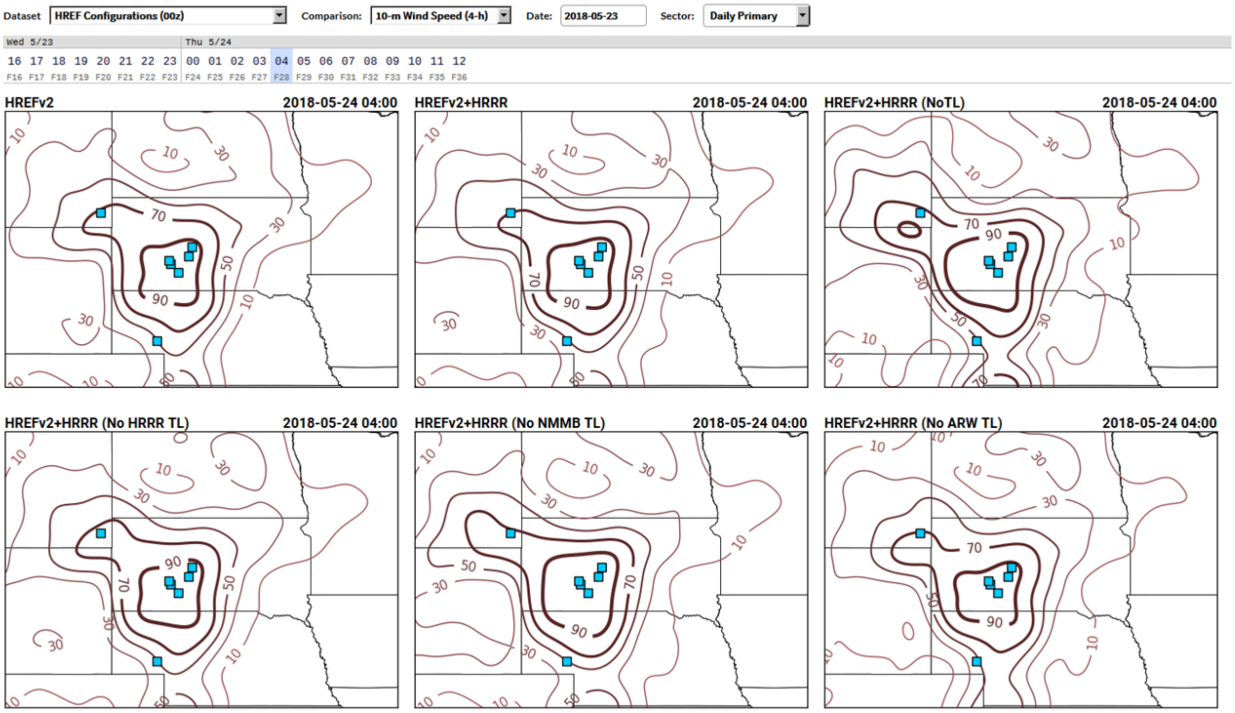


Dataset: HREF Configurations (00z) Comparison: Updraft Helicity (4-h) Date: 2018-05-02 Sector: Daily Primary

Wed 5/2
Thu 5/3
16 17 18 19 20 21 22 23 00 01 02 03 04 05 06 07 08 09 10 11 12
F16 F17 F18 F19 F20 F21 F22 F23 F24 F25 F26 F27 F28 F29 F30 F31 F32 F33 F34 F35 F36

HREFv2 2018-05-03 02:00
HREFv2+HRRR 2018-05-03 02:00
HREFv2+HRRR (NoTL) 2018-05-03 02:00
HREFv2+HRRR (No HRRR TL) 2018-05-03 02:00
HREFv2+HRRR (No NMMB TL) 2018-05-03 02:00
HREFv2+HRRR (No ARW TL) 2018-05-03 02:00

Figure 46 Same as Fig. 45, except for forecasts valid 0200 UTC on 3 May 2018. The upside-down red triangles represent tornado reports, and the black circles represent significant hail (i.e., ≥ 2" diameter) reports.



Dataset: HREF Configurations (00z) Comparison: 10-m Wind Speed (4-h) Date: 2018-05-23 Sector: Daily Primary

Wed 5/23
Thu 5/24
16 17 18 19 20 21 22 23 00 01 02 03 04 05 06 07 08 09 10 11 12
F16 F17 F18 F19 F20 F21 F22 F23 F24 F25 F26 F27 F28 F29 F30 F31 F32 F33 F34 F35 F36

HREFv2 2018-05-24 04:00
HREFv2+HRRR 2018-05-24 04:00
HREFv2+HRRR (NoTL) 2018-05-24 04:00
HREFv2+HRRR (No HRRR TL) 2018-05-24 04:00
HREFv2+HRRR (No NMMB TL) 2018-05-24 04:00
HREFv2+HRRR (No ARW TL) 2018-05-24 04:00

Figure 47 Same as Fig. 45, except for 4-h neighborhood probabilities of 10-m wind speeds exceeding 30 kts valid at 0400 UTC on 24 May 2018. The blue squares represent severe/damaging wind reports.

Overall, the different HREF configurations were rated similarly in terms of providing severe weather guidance during the five-week 2018 HWT SFE with mean subjective ratings ranging between 6.1 to 6.4 (Fig. 48). In fact, all of the HREF configurations had a median rating of 7 (out of 10), except for the HREFv2+HRRR (No NMMB TL) configuration, which had a lower median rating of 6. Subjectively, there was day-to-day variability in the performance of the various HREF configurations with forecasts on most days appearing similar enough to not provide a practical difference to a forecaster (i.e., differences not large enough to change an outlook). This result was not necessarily expected for the 10-member HREFv2+HRRR configuration compared to the 5-member HREFv2+HRRR (No TL) configuration, but it does highlight the resiliency of an ensemble to membership changes.

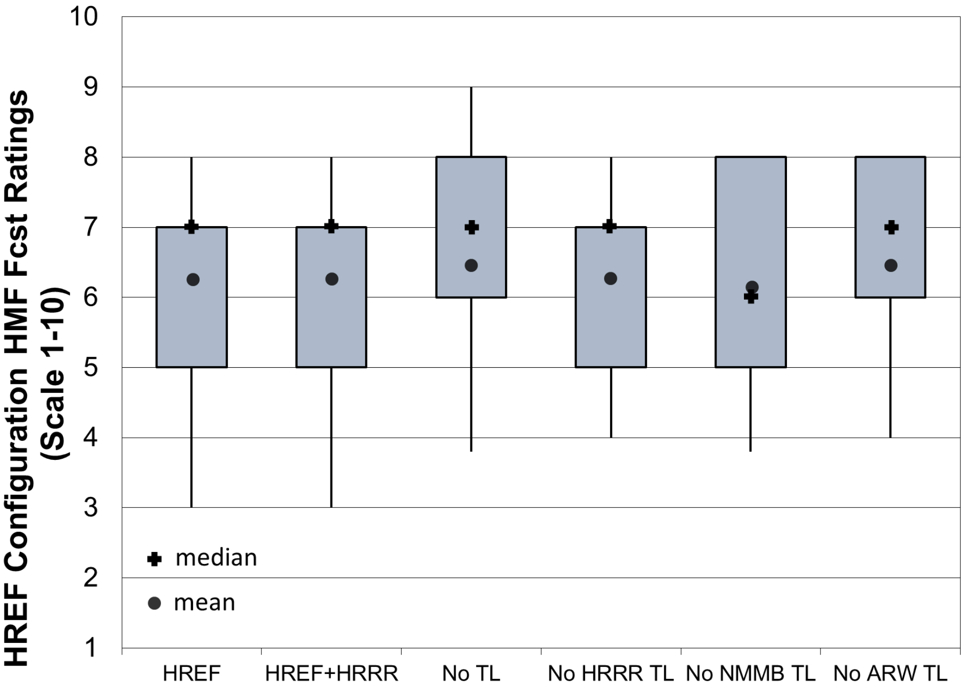


Figure 48 Box-and-whiskers plot of subjective ratings (1-10) for ensemble neighborhood probabilistic forecasts of hourly maximum fields from the HREF configuration experiment during the five-week 2018 HWT SFE. The boxes represent the interquartile range, and the whiskers represent the 10th and 90th percentiles. The crosses represent the median ratings, and the circles represent the mean ratings.

To investigate another perspective of the subjective ratings, the number of times that each HREF configuration was given the single-highest rating for a particular forecast was recorded. This indicated when an HWT SFE participant felt that one HREF configuration stood out as the top performer for a particular forecast. For the majority of forecasts, no HREF configuration stood out as the top performer (Fig. 49). The HREF configuration without any time-lagged members [HREFv2+HRRR (No TL)] was rated as the top performer more often than the other configurations, but it was only for a small percentage (~15%) of the forecasts. On occasions when older runs performed poorly, removing them from the HREF improved the probabilistic forecast. However, for most cases, the time-lagged members did not degrade and actually improved the probabilistic ensemble forecast.

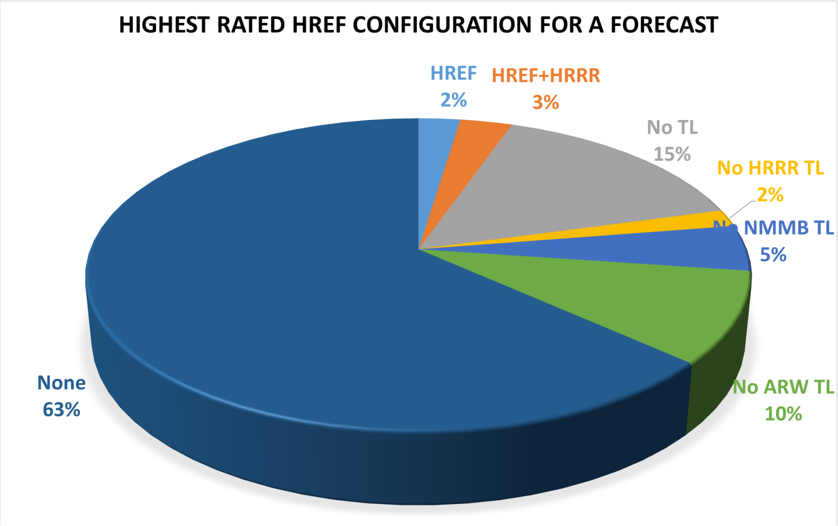


Figure 49 Pie chart showing the percentage of forecasts for which an HREF configuration received the highest subjective rating during the 2018 HWT SFE.

Similarly, the number of times that each HREF configuration was given the single-*lowest* rating for a particular forecast was documented. It was even more common for none of the HREF configurations to stand out as the worst performer, as more than three-fourths of the ratings did not highlight a single poorest-performing configuration (Fig. 50). The HREFv2+HRRR (No NMMB TL) configuration was rated as the worst performer more often than any other configuration (i.e., ~9% of the ratings), which is somewhat surprising given the perception that NMMB members do not perform as well as the ARW members for convective weather forecasting. The additional spread provided by time-lagged NMMB members occasionally contributed to improving the probabilistic severe weather guidance in HREF forecasts (e.g., convective initiation southward along a dryline).

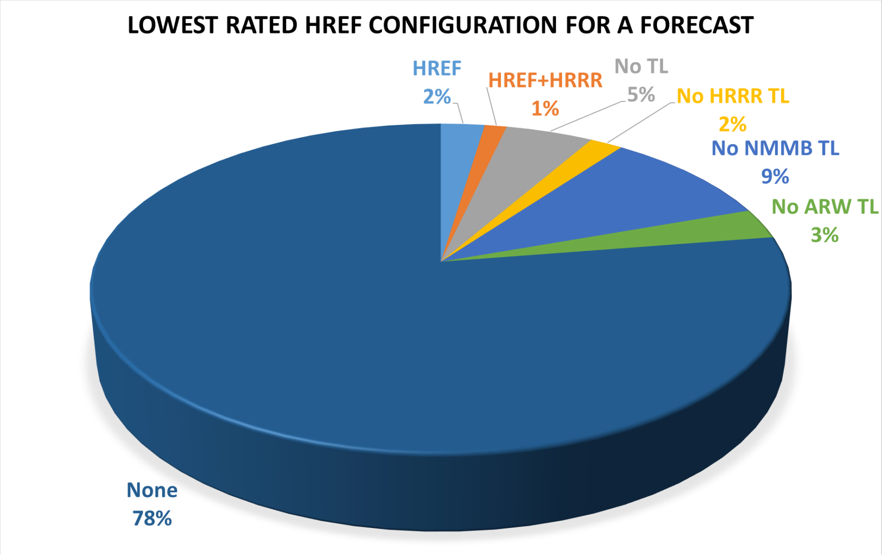


Figure 50 Same as Fig. 51, except for the percentage of forecasts in which a particular HREF configuration received the lowest subjective rating.

The main takeaway from the subjective HREF configuration comparison is that the various ***HREF configurations looked very similar overall*** on most days for severe weather guidance (i.e., the practical difference to a forecaster was small). On some days during the SFE, the time-lagged members did not perform as well as more recent convection-allowing model runs, so removing them improved the probabilistic ensemble forecast. Unexpectedly, the time-lagged NMMB members appeared to add more value (through increased ensemble diversity/spread) than the time-lagged ARW members during the 2018 HWT SFE, as the HREFv2+HRRR (No NMMB TL) configuration was rated the lowest overall.

For objective verification, surrogate severe fields were generated using the 0000 UTC cycle of six HREF configurations, following the procedure outlined by Sobash et al. (2011). Each member's UH fields were regridded to an 80-km grid, with the maximum value over the 1200 UTC – 1159 UTC the following day assigned to each gridbox. If a field exceeded a percentile threshold of UH, it was assigned a value of one. Finally, a Gaussian smoother was applied to the resulting binary field to generate a probabilistic forecast. Member fields were then averaged for each ensemble configuration to generate an ensemble surrogate severe field (Sobash et al. 2016). For each configuration, 100 different UH percentile thresholds and 53 different σ thresholds were tested.

To determine what values the UH percentiles corresponded to in each member of the HREF, climatologies were first generated for each of the models over the SFE (Fig. 51). The models with the NMMB dynamical core (the NAM nest and HRW NMMB members) had overall higher UH than the members with the ARW cores. As expected, the 0000 UTC members and 1200 UTC time-lagged members had very similar UH climatologies.

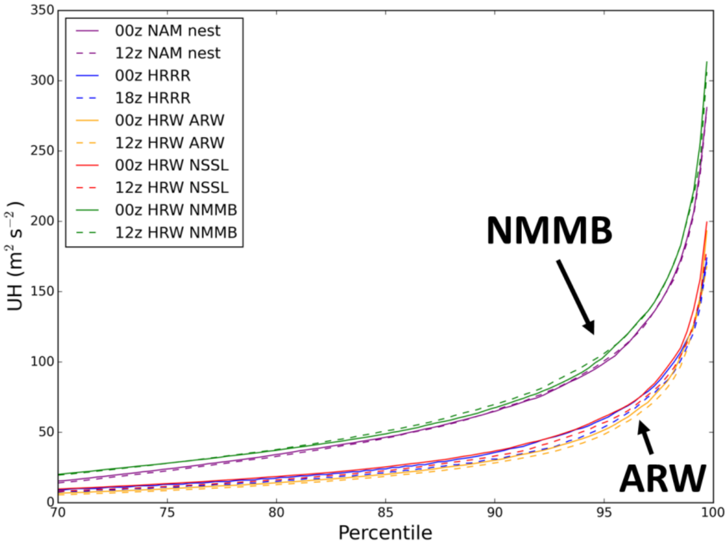


Figure 51 Model climatologies of the different members of the 0000 UTC experimental HREF configurations. Dashed lines indicate time-lagged members, and the different models are indicated by the different colors.

Once the surrogate severe fields were calculated, they were verified against practically perfect fields generated using LSRs, as in the practically perfect fields that were used to subjectively evaluated the full period Innovation Desk outlooks. Metrics were calculated over two sets of dates based on the SPC 0600 UTC Day 1 Convective Outlook: dates with a categorical slight risk or less (*n* = 15), and dates with a categorical enhanced risk or greater (*n* = 14). Objective verification metrics considered include the ROC area (Mason 1982), FSS, and the reliability term from the Brier Score.

FSS results from the two different categories showed different behavior depending on the severity of the day's weather (Fig. 52). On the lower-end days (categorical slight risk or less), the 10-member ensemble composed of all of the current members of the HREFv2, the HRRR, and a time-lagged version of the HRRR performed best of any configuration. The configuration that contained the HRRR but no time-lagged members (5 total members) performed the worst on these days. Conversely, on the high-end days categorized as an enhanced risk or greater, the FSS was *highest* for the no time-lagged member ensemble and *lowest* for the 8-member original HREFv2. Second-lowest, however, was the full 10-member ensemble. These results suggest that the added diversity provided by the time-lagged members helps the forecast in more weakly forced environments, but that the time-lagged members may be degrading forecast skill for high-end events. These results also show that adding the HRRR benefits the HREF across all types of events. Finally, better performance is seen in all configurations for high-end days compared to the lower-end days.

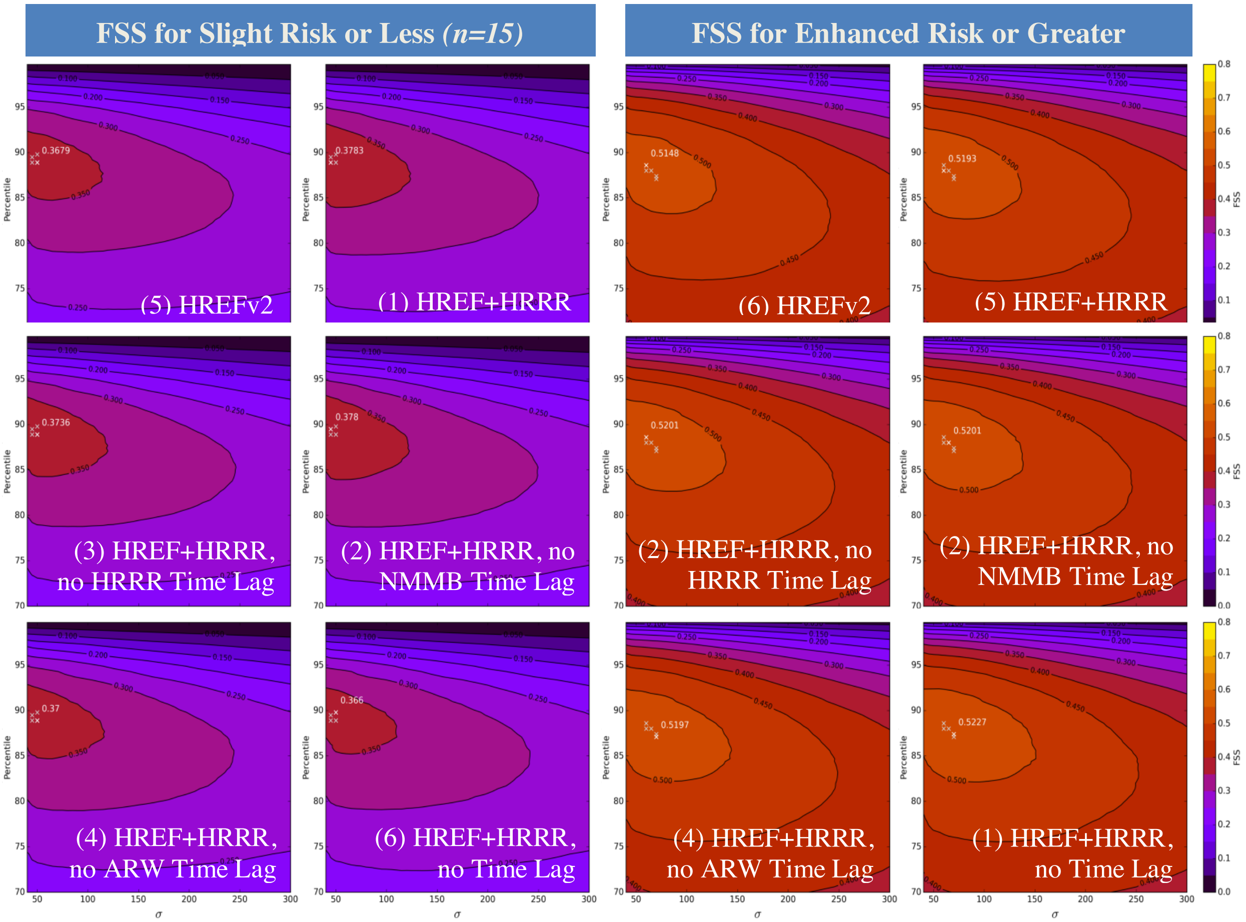


Figure 52 FSS contour plots for each potential configuration of the HREF. The best percentile and σ combination for each ensemble is highlighted by a white x, and the numerical ranking of the scores is by the configuration name in the lower right-hand corner of the plot. The left six figures show the FSS for slight risk or less days, and the right six figures show the FSS for enhanced risk or greater days.

In order to objectively compare HREF candidate configurations' performance in their depiction of the overall convective evolution, neighborhood maximum ensemble probability (NMEP) forecasts for simulated composite reflectivity (hereafter "REFC") exceeding 40 dBZ were generated for each configuration for each day's 0000 UTC HREF run during the 2018 SFE period. For each HREF run, these forecasts were generated hourly between forecast hours 12-30 (hours 31-36 were excluded because one member, the HRRR -6h, was not available after 30 hours). The neighborhood was an 80x80 km box centered on each grid point, and a Gaussian smoother (σ = 40 km) was applied to the grid point probabilities. These probabilistic forecasts were verified against the NSSL Multi-Radar/Multi-Sensor (MRMS) national composite reflectivity mosaic ("MergedReflectivityQCComposite") valid at the closest available analysis time (within ±5 minutes of the HREF forecast valid time). Specifically, for MRMS reflectivity, binary fields (0 = no, 1 = yes) were generated with respect to 40-dBZ threshold exceedance within the same 80x80 km neighborhood used for the HREF forecasts, then smoothed in the same manner. These pseudo-NMEP grids are the verification dataset used for the REFC.

Frequency biases for REFC were computed for each HREF member as an intermediate step toward producing bias-corrected probabilistic forecasts. The biases were calculated over the same set of forecasts and forecast times given above, and for each REFC threshold between 35-50 dBZ at an interval of 1 dBZ, plus 25 and 30 dBZ. Figure 53 presents a summary of these biases for each member; note that values for the time-lagged and non-lagged variants of each CAM configuration are combined here. For a given CAM configuration, the lagged and non-lagged biases were qualitatively very similar for all thresholds (not shown), indicating that model configuration differences (e.g., model core and microphysics) were dominant in controlling frequency bias.

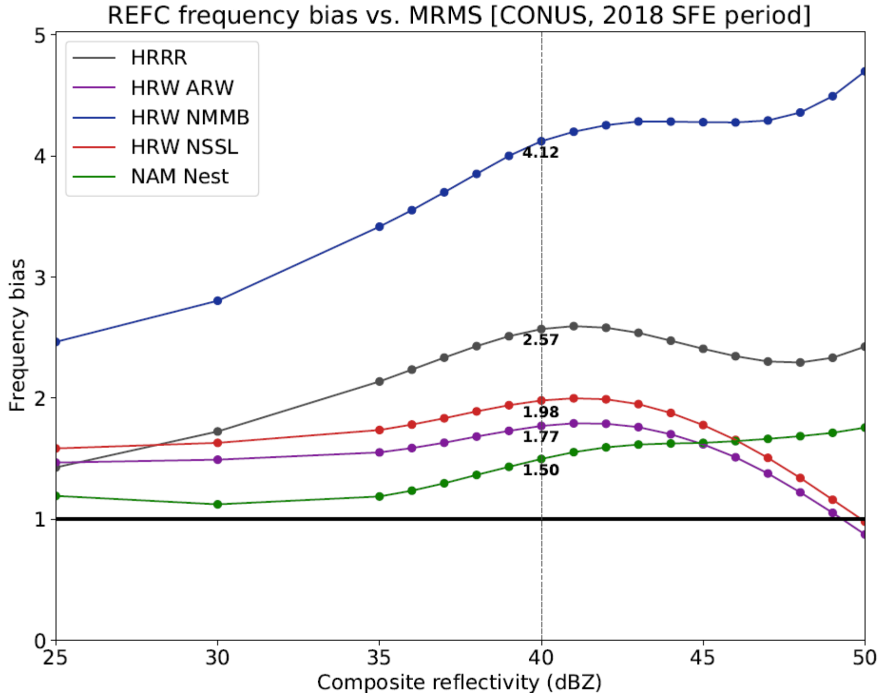


Figure 53 Frequency bias vs. REFC threshold for each HREF candidate member over the SFE 2018 dataset. Time-lagged members are included with their non-lagged counterparts in this chart (e.g., "HRRR" stats include both the lagged and non-lagged HRRR members). Verification domain is the CONUS.

At the 40-dBZ threshold, all members were high biased, with biases ranging from 1.50 to 4.12. The HRW NMMB member exhibits a far more severe bias than the other members. At REFC values approaching 50 dBZ associated with intense convection, the HRW ARW and HRW NSSL members exhibit biases nearer unity, while the remaining members remain quite high-biased. Overall, the CAMs participating in the candidate HREF configurations were too aggressive in their spatial coverage of REFC exceeding all thresholds between 30-50 dBZ, but the HRW NMMB is noteworthy for its anomalously severe bias.

Two sets of NMEP forecasts were produced for each configuration. In the "uncorrected" NMEPs, the 40-dBZ exceedance threshold was used in each member (such that their biases were not accounted for). In the "corrected" NMEPs, an exceedance threshold was chosen separately for each member, such that the frequency bias was ~1.0 with respect to the 40-dBZ threshold in the MRMS verification dataset. The REFC thresholds used for the "corrected" NMEPs ranged from 42.7 dBZ (NAM Nest) to 47.8 dBZ (HRW NMMB).

Figure 54 presents Brier Skill Score (BSS) for the corrected (solid) and uncorrected (hatched) NMEPs. Although the corrected NMEP scores are markedly higher than the uncorrected scores, the candidate configurations' rankings are identical between the two sets of scores. The No NMMB-TL configuration performs best, while the No ARW-TL configuration is worst; the current baseline HREFv2 (which excludes both HRRR members) is second-worst among the six configurations. Fig. 55 presents the Fractions Skill Score (FSS) for the same NMEPs. Although the quantitative difference between the corrected and uncorrected scores is generally much smaller for FSS, the same ranking seen in Fig. 54 is replicated here, except for rankings #5 and #6 switching in the corrected forecasts. Based on the 40-dBZ REFC NMEP verification, two inferences can be made about the candidate HREF configurations: (1) the HRRR is adding useful information to the baseline HREFv2 configuration; and (2) the NMMB-TL members are hurting BSS and FSS scores in configurations where they participate, suggesting whatever spread they add to the forecast is not useful enough to offset their relatively poor performance overall for REFC.

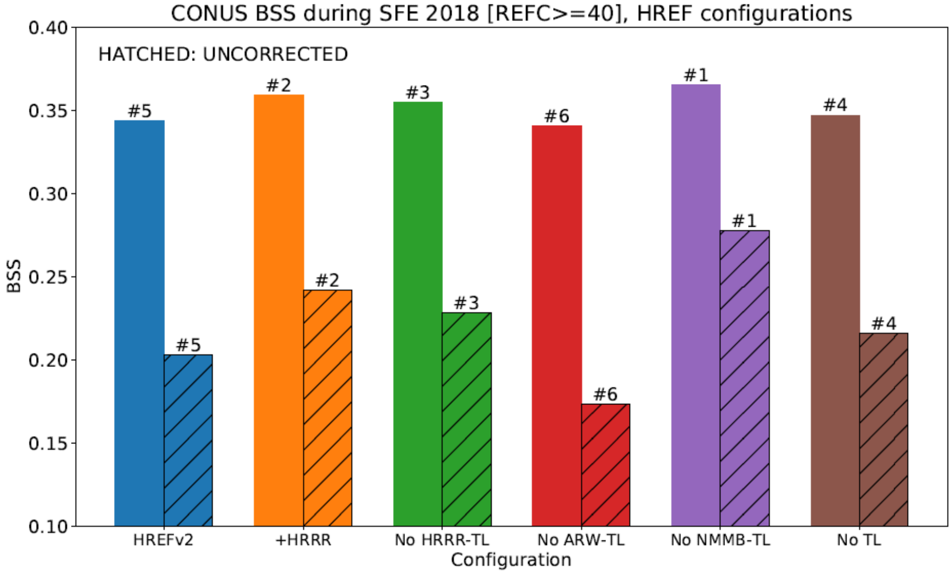


Figure 54 Brier Skill Score (BSS) for 40-dBZ REFC for HREF candidate configurations over the SFE 2018 dataset; corrected (solid) and uncorrected) hatched are given. Verification domain is the CONUS.

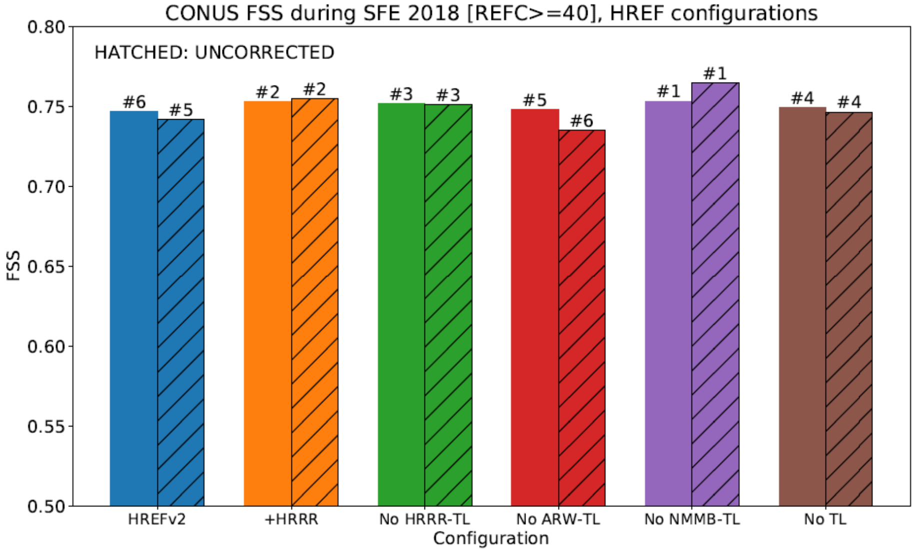


Figure 55 Fractions Skill Score (FSS) for 40-dBZ REFC NMEPs for HREF candidate configurations over the SFE 2018 dataset; corrected (solid) and uncorrected) hatched are given. Verification domain is the CONUS.

Fig. 56 presents an attributes diagram for the corrected configuration forecasts; in this case, scores are computed over the daily SFE domains, rather than the full CONUS. All configurations demonstrate remarkably good reliability after bias correction. The bias-corrected GSD HRRRE NMEPs are also included for context, showing considerably poorer reliability (with notably worse resolution for probabilities in the 0.2-0.8 range). This corroborates a frequent observation in our subjective evaluations from the SFE: that the HRRRE is consistently more under-dispersive in its depiction of convective evolution than the HREF (regardless of HREF configuration).

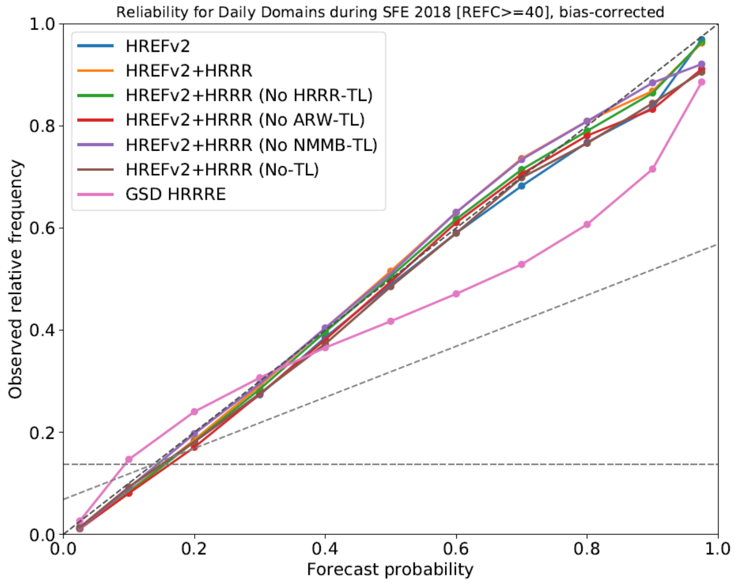


Figure 56 Attributes diagram for 40-dBZ REFC NMEPs for HREF configurations, plus GSD HRRRE, over the 2018 SFE dataset. Bias-corrected NMEPs are used. Verification domain for each day's forecast is the daily SFE domain.

6) HAIL GUIDANCE (credit: I. Jirak)

Several existing and new hail diagnostic fields were examined in the mixed-physics ensemble CLUE subset during the 2018 HWT SFE. The hail diagnostic fields examined as part of the mixed-physics ensemble included HAILCAST (Becky Adams-Selin; Adams-Selin et al. 2018), a microphysics-based approach (Greg Thompson), hail size estimated based on updraft speed (Nathan Wendt), and a machine-learning approach (David Gagne, Nathan Snook; Gagne et al. 2017), along with the standard CAM storm-attribute fields (i.e., UH and updraft speed). These hail proxy forecasts were evaluated and compared daily during the 2018 HWT SFE using the web-based interface shown in Fig. 57.

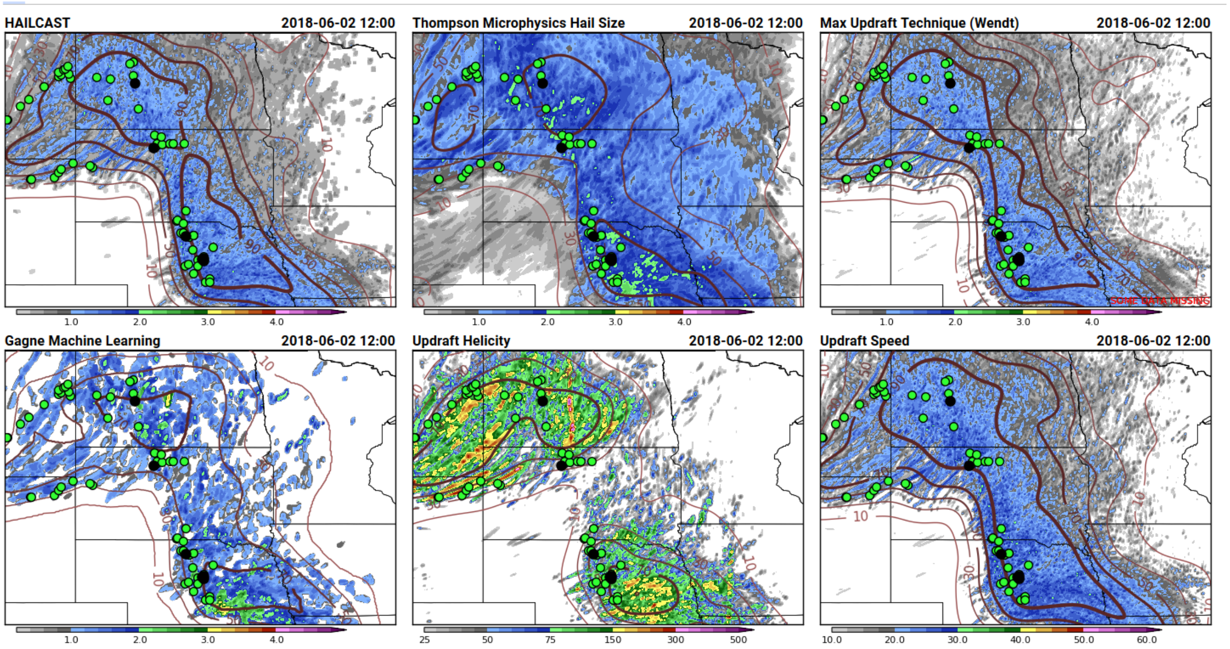


Figure 57 Six-panel comparison plot used to conduct the evaluation of the hail output variables from the CLUE mixed-physics ensemble during the 2018 HWT SFE. The 24-h neighborhood hail probability forecasts exceeding 1 inch valid for 1 June 2018 are shown for HAILCAST (top-left panel), microphysics-based approach (top-middle panel), maximum updraft approach (top-right panel), machine-learning technique (bottom-left panel), updraft helicity (≥90 m^2s^{-2}; bottom-middle 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 2018 HWT SFE, participants would subjectively rate (on a scale of 1 to 10) the quality of the different hail-proxy forecasts from the mixed-physics CAM ensemble valid for the previous day. The observed hail reports and MESH values were used as the verification sources to help assess the quality of the forecasts. The distribution of the subjective ratings for these hail proxies are shown in Fig. 58. The top-rated hail-proxy forecasts were from the machine-learning method, UH, and HAILCAST. The MAXHAILW and updraft speed subjective ratings followed closely behind with the microphysics-based approach having the lowest subjective ratings. The microphysics-based approach was primarily tuned for the Thompson microphysics scheme, so the results from this mixed-physics ensemble (i.e., using other microphysics schemes) were negatively impacted from not being properly tuned to each microphysics scheme. Another notable issue that impacted the updraft-speed based estimates/approaches was the use of the HRRR configuration setting in WRF to cap the latent heating rate (for operational stability purposes), which limits the updraft strength.

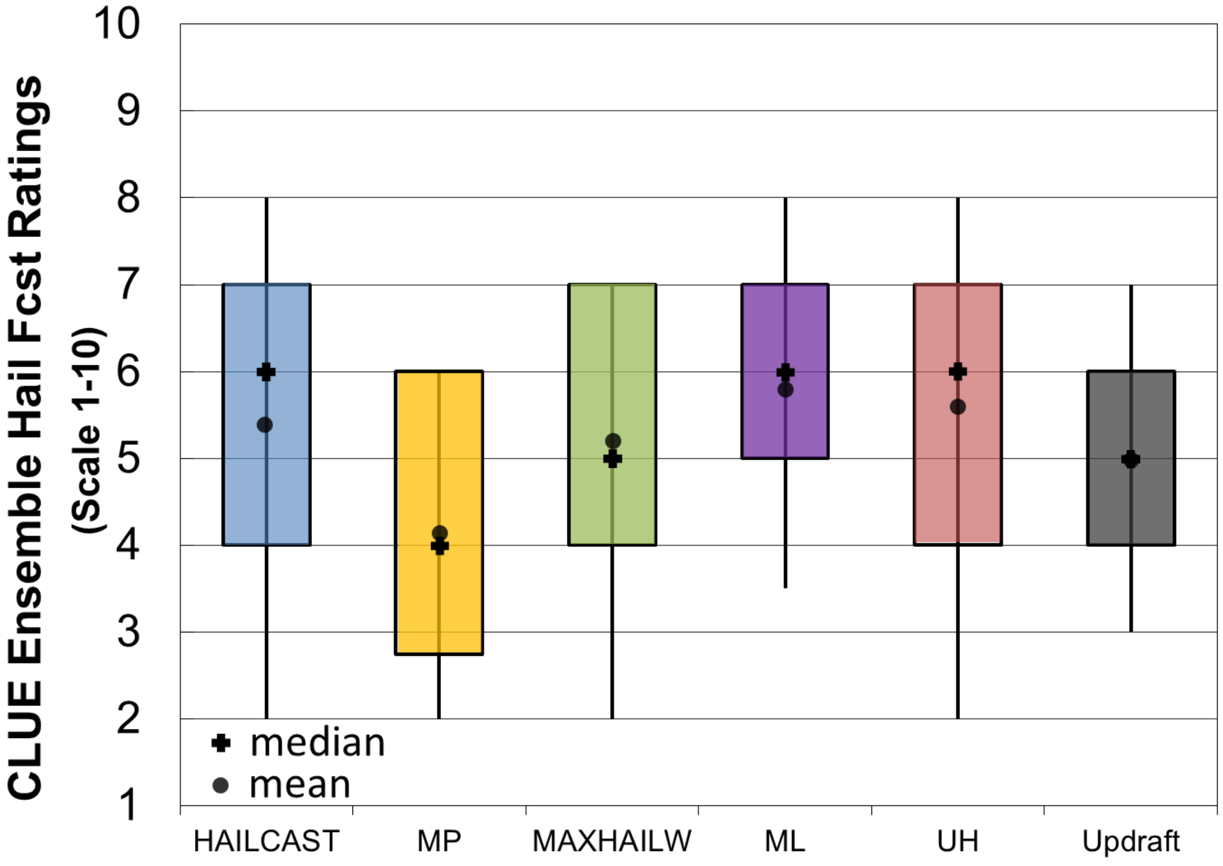


Figure 58 Subjective ratings (scale of 1-10) by participants during the 2018 HWT SFE of the hail proxy forecasts: a) HAILCAST – blue, b) microphysics (MP) approach – yellow, c) updraft speed approach (MAXHAILW) – green, d) machine-learning (ML) approach, e) updraft helicity (UH) – red, and f) updraft speed – gray.

7) ENSEMBLE SENSITIVITY-BASED SUBSETTING (credit: Brian Ancell)

A daily evaluation of probabilities from the operational 42-member Texas Tech ensemble against those based on 10-member ensemble subsets chosen objectively through the sensitivity-based subsetting technique (Ancell 2016) was performed. Each day, a response function location was chosen collectively with HWT participants through a web-based graphical user interface that identified areas of Day 1 severe convection within that day's 0000 UTC Texas Tech operational ensemble run. The Day 1 response function area was selected at a forecast hour between 1800 UTC (the 18hr forecast) and 1200 UTC (the 36hr forecast) in areas that exhibited high uncertainty over the prior 6hr period. This response selection process was performed by viewing the full 42-member probabilities of exceeding 25 and 100 m^2/s^2 2-5km UH valid over the 18-36hr forecast period. Once the response function time and location were chosen, the sensitivity of three independent response functions were automatically calculated: 1) maximum 2-5km UH, 2) number of grid points exceeding 25 m^2/s^2 2-5km UH, and 3) number of grid points exceeding 40 dBZ lowest-model-level simulated reflectivity. The sensitivities of the three response functions (chosen on the 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 CONUS domain all with respect to the 7-hr forecast state (valid 0700 UTC).

The 10 ensemble members from the 0000 UTC run 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. The analysis used to determine the errors was the 1hr forecast ensemble mean (valid at 0700 UTC) from the 0600 UTC Texas Tech ensemble initial conditions determined through the DART EAKF data assimilation procedure. The 1hr forecast at 0700 UTC was used in lieu of the analysis valid at 0600 UTC due to significant imbalance present after the assimilation procedure. Probability fields (specifically maximum 6-hourly 20-mile neighborhood exceedance probabilities of 25 m^2/s^2 2-5km UH and 40 dBZ simulated near-surface reflectivity) of Day 1 convection were generated for the 10-member ensemble subset and compared against probabilities from the full ensemble the following day after the severe event occurred. Differences between the full and subset probabilities were also calculated and evaluated, and SPC storm reports and practically perfect probability fields were generated to serve as the probabilistic "truth" against which both the full and subset ensemble probabilities were judged.

Figures 59 and 60 show two examples of the subsetting product during the 5-week experiment that participants evaluated. Figure 59 depicts a successful case for convection in southwest Oklahoma on May 2. Probabilities of simulated reflectivity exceeding 40 dBZ from the 10-member subset were increased over that of the full ensemble generally by 20-30%, reaching over 50% across a wide area and suggesting a higher probability of storms there. Using the simulated reflectivity coverage response function in this case demonstrates how the forecast of the presence of storms can be improved through the subsetting technique and its adjusted probability fields. Numerous storm reports reflected the presence of high wind, hail, and tornadoes in the same area, and was unanimously viewed as a success on this day by HWT participants. In contrast, Figure 60 shows a failure case for convection in southeast Colorado on May 10. In this case the UH coverage response function was used in an attempt to better predict the extent of rotating thunderstorms. Subset probabilities were increased beyond the full ensemble by around 20%, yet no storm reports were made in this area.

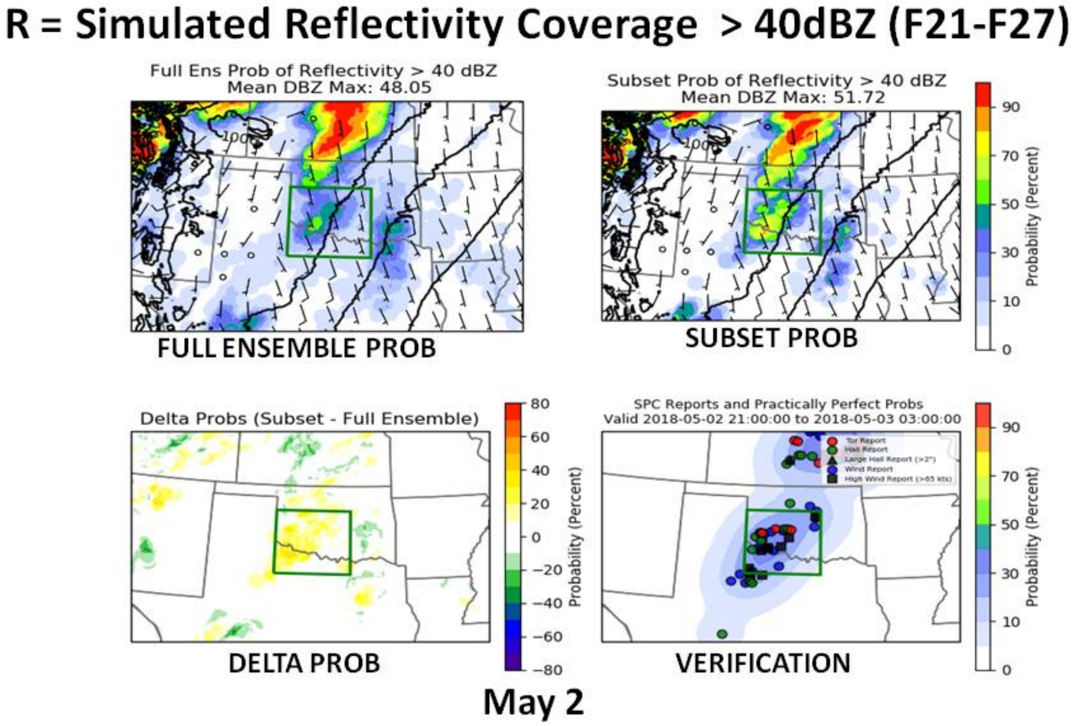

R = Simulated Reflectivity Coverage > 40dBZ (F21-F27)
Full Ens Prob of Reflectivity > 40 dBZ
Mean DBZ Max: 48.05
Subset Prob of Reflectivity > 40 dBZ
Mean DBZ Max: 51.72
FULL ENSEMBLE PROB
SUBSET PROB
Probability (Percent)
0
10
30
50
70
90
Delta Probs (Subset - Full Ensemble)
80
60
40
20
0
−20
−40
−60
−80
SPC Reports and Practically Perfect Probs
Valid 2018-05-02 21:00:00 to 2018-05-03 03:00:00
Tor Report
Hail Report
Large Hail Report (>2")
Wind Report
High Wind Report (>65 kts)
DELTA PROB
VERIFICATION
May 2

Figure 59 Full 42-member and subset 10-member ensemble probability of exceeding 40 dBZ within 20 miles (top row), and the difference in full and subset ensemble probabilities as well as storm reports with SPC practically perfect total severe probabilities (bottom row) for May 2 (success case).

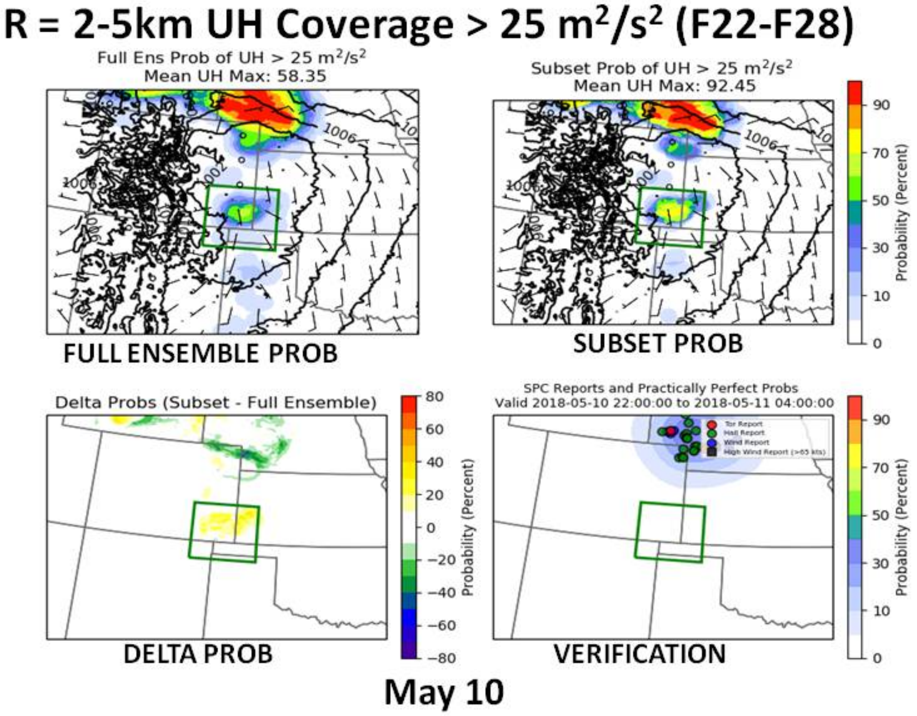
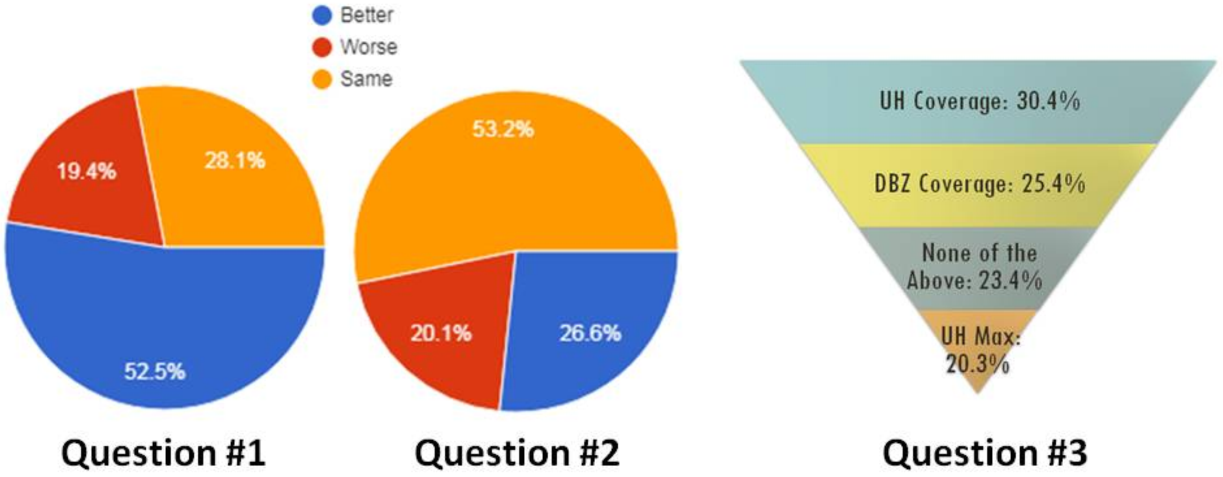

R = 2-5km UH Coverage > 25 m²/s² (F22-F28)
Full Ens Prob of UH > 25 m²/s²
Mean UH Max: 58.35
Subset Prob of UH > 25 m²/s²
Mean UH Max: 92.45
1006
1002
FULL ENSEMBLE PROB
SUBSET PROB
Probability (Percent)
0
10
30
50
70
90
Delta Probs (Subset - Full Ensemble)
80
60
40
20
0
−20
−40
−60
−80
SPC Reports and Practically Perfect Probs
Valid 2018-05-10 22:00:00 to 2018-05-11 04:00:00
Tor Report
Hail Report
Wind Report
High Wind Report (>65 kts)
DELTA PROB
VERIFICATION
May 10

Figure 60 Full 42-member and subset 10-member ensemble probability of exceeding 25 m^2/s^2 2-5km updraft helicity within 20 miles (top row), and the difference in full and subset ensemble probabilities as well as storm reports with SPC practically perfect total severe probabilities (bottom row) for May 10 (failure case).

Efforts at Texas Tech after the HWT have been focused on understanding the optimal parameters for the subsetting technique and the technique's success rate and degree of improvement. This is because over the entire 5-week period of the HWT, participant responses were favorable toward further development of the method (results shown in Figure 61). Three questions were asked of participants daily: 1) What is the skill of the ensemble subset relative to the full ensemble inside the response function box? 2) What is the skill of the ensemble subset relative to the full ensemble outside the response function box? 3) What response function produce the most skillful subset probabilities? The purpose of question #1 was to understand how participants viewed subset probability improvements inside the chosen response box, which is the area directly targeted by the subsetting technique. Question #2 was raised to understand whether participants thought areas outside the chosen response area were improved by ensemble subset probabilities, which could be achieved if those areas were correlated with the severe weather inside the response area (even though the technique does not directly target those areas outside the response area for improvement). Question #3 was aimed to reveal the most useful response function toward improving probabilistic skill. Just over half of all responses indicated probabilistic skill was improved within the response box, while only about 19% felt it was degraded. Similarly, about 20% of responses reflected a degradation outside the box, although most of the responses (just over half) indicated no change in skill there. UH coverage was viewed as the most beneficial response function to produce the adjusted subset probabilities, while maximum UH was perceived as the worst.



Question #1

Question #2

Question #3

Figure 61 Ensemble sensitivity-based subsetting survey question results for the entire 5-week period of the 2018 HWT.

4. Summary

The 2018 Spring Forecasting Experiment (SFE2018) was conducted at the NOAA Hazardous Weather Testbed from 30 April – 1 June by the SPC and NSSL with participation from forecasters, researchers, model developers, university faculty and graduate students from around the world. The primary theme of SFE2018 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 third 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 second year that a prototype Warn-on-Forecast system has been tested for issuing short-lead-time outlooks.

Several preliminary findings/accomplishments from SFE2018 are listed below:

- Generated high temporal resolution outlooks for individual severe hazards (tornado, hail, wind) using first-guess guidance from a temporally disaggregated full-period outlook created with calibrated probabilistic guidance from a convection-allowing ensemble.
- Explored methods to include more detailed timing information by issuing potential severe timing (PST) areas, which are enclosed areas valid for 4-h periods that highlight the expected timing of severe weather occurrence.
- Examined various convection-allowing ensemble systems within the CLUE using HREFv2 as a baseline.
    - While all of the ensembles provided similar, useful guidance for Day 1 severe weather forecasting, the HREFv2 received higher subjective ratings than the other systems, suggesting that HREFv2 is a skillful baseline for CAM ensemble forecasts.
- Tested a prototype Warn-on-Forecast short-term prediction system in real-time for the second year at the Innovation Desk and the first year at the Severe Hazards Desk during an afternoon forecasting activity with very promising results.
- Examined real-time, storm-scale FV3 simulations for the second year during SFE2018.
    - Subjective ratings revealed that FV3 reflectivity forecasts were often comparable to operational CAMs.
    - Subjectively, none of the PBL schemes in FV3 stood out as performing best. In comparisons of sounding structures between PBL schemes, participants most often noted large differences in low-level moisture.
    - In subjective comparisons of FV3 members with Thompson and NSSL microphysics, NSSL (Thompson) members most often had the best depiction of reflectivity and UH location (reflectivity magnitude), and the two schemes most frequently had similar depictions of storm mode.
    - In subjective comparisons between deterministic FV3 configurations, NSSL and CAPS runs performed best and GFDL runs performed worst, while HRRRv3 runs performed better than all the FV3 configurations.
    - These results support continued research to refine and improve FV3 for storm-scale applications before it is implemented operationally as part of an emerging unified NOAA model production suite.

- Evaluated an ensemble sensitivity-based subsetting technique. Subjectively, most frequently the ensemble subset had greater or equal forecast skill relatively to the full ensemble. Updraft helicity coverage was viewed as the most beneficial response function to produce the ensemble subset probabilities.
- Examined six candidate HREFv2.1 configurations that added extended HRRR runs and/or removed some of all of the time-lagged members. The various HREF configurations looked very similar overall on most days for severe weather guidance. In objective analyses, adding the HRRR was beneficial across a range of severe weather events.
- Several different hail size forecasting methods were examined. In subjective evaluations, the top-rated hail-proxy forecasts were from the machine learning method, UH, and HAILCAST.

Overall, SFE2018 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 SFE2018 directly promote continued progress to improve forecasting of severe weather in support of the NWS Weather-Ready Nation initiative.

Acknowledgements

SFE2018 would not have been possible without dedicated participants and the support and assistance of numerous individuals at SPC and NSSL. In addition, collaborations with CAPS, NCAR, the United Kingdom Met Office, ESRL/GSD, the University of Oklahoma, and EMC were vital to the success of SFE2018. In particular, Ming Xue (CAPS), Fanyou Kong (CAPS), Kevin Thomas (CAPS), Keith Brewster (CAPS), Youngsun Jung (CAPS), Tim Supinie (CAPS), Chunxi Zhang (CAPS), Glen Romine (NCAR), Craig Schwartz (NCAR), Ryan Sobash (NCAR), Kate Fossell (NCAR), Steve Willington (Met Office), Curtis Alexander (ESRL/GSD), Geoff Dimego (EMC), Jacob Carley (EMC), Brad Ferrier (EMC), Eric Aligo (EMC), S. J. Lin (GFDL), Lucas Harris (GFDL), 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 examined on a daily basis.

References

Adams-Selin, R. A., A. J. Clark, C. J. Melick, S. R. Dembek, I. L. Jirak, and C. L. Ziegler, 2018: Evolution of WRF-HAILCAST during the 2014-2016 NOAA/Hazardous Weather Testbed Spring Forecasting Experiments. *Wea. Forecasting*, Early Online Release.

Ancell, B.C., 2016: Improving High-Impact Forecasts through Sensitivity-Based Ensemble Subsets: Demonstration and Initial Tests. *Weather and Forecasting*, Vol. 31, No. 3, pages 1019-1036.

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.

Gallo, B. T., A. J. Clark, I. Jirak, S. J. Weiss, A. Dean, K. Knopfmeier, B. Roberts, L. Wicker, M. Krocak, N. Wendt, P. Skinner, J. Choate, P. Heinselman, K. Wilson, R. Heper, J. Correia, G. Creager, T. Jones, J. Gao, Y. Wang, S. Dembek, 2018b: Spring Forecasting Experiment 2018 Program Overview and Operations Plan. Available online at: https://hwt.nssl.noaa.gov/sfe/2018/docs/HWT_SFE2018_operations_plan.pdf.

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

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

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

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

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]

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

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

Roberts, N. M. and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97.

Roebber, P.J., 2009: Visualizing Multiple Measures of Forecast Quality. *Wea. Forecasting*, **24**, 601–608.

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

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

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

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

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

APPENDIX A

Table A1 Daily activities schedule in local (CDT) time

<table>
<tr><th><i>Severe Hazards Desk</i></th><th><i>Innovation Desk</i></th></tr>
<tr><td colspan="2">0800 – 0845: <b>Evaluation of Experimental Forecasts & Guidance</b>
Subjective rating relative to radar evolution/characteristics, warnings, preliminary reports, and MRMS MESH and rotation tracks</td></tr>
<tr><td>• Day 1 full-period probabilistic forecasts of tornado, wind, and hail
• Day 1 4-h period forecasts and guidance for tornado, wind, and hail</td><td>• Days 1 full-period probabilistic forecast of total severe
• Day 1 4-h areas for severe weather timing
• Day 1 1-h total severe outlooks</td></tr>
<tr><td colspan="2">0845 – 0915: <b>Map Analysis</b>
Hand analysis of 12Z upper-air and surface maps, discussion, and domain selection (from two areas)</td></tr>
<tr><td colspan="2">0915 – 1130: <b>Convective Outlook Generation</b></td></tr>
<tr><td>• Day 1 full-period probabilistic forecasts of tornado, wind, and hail valid 16-12Z over mesoscale area of interest
• Day 1 4-h probabilistic forecasts of tornado, wind, and hail valid 17-21Z and 21-01Z using CLUE subsets*</td><td>• Day 1 full-period probabilistic forecast of total severe valid 16-12Z over mesoscale area of interest
• Day 1 4-h timing areas (16-12Z) for full-period total severe ≥15% using CLUE subsets*</td></tr>
<tr><td colspan="2">1130 – 1200: <b>Map Discussion</b>
Brief discussion of today’s forecast challenges and products
Topic of the day: 3D Vis, Met Office, FV3, NEWS-e, CAM scorecard</td></tr>
<tr><td colspan="2">1200 – 1300: <b>Lunch</b></td></tr>
<tr><td colspan="2">1300 – 1345: <b>Convective Outlook Generation</b></td></tr>
<tr><td>• Day 1 4-h probabilistic forecasts of tornado, wind, and hail valid 19-23Z using 12Z CAM ensembles*</td><td>• Update Day 1 4-h timing areas (19-12Z) for full-period total severe ≥15% using 12Z CAM ensembles*</td></tr>
<tr><td colspan="2">1345 – 1500: <b>Scientific Evaluations</b></td></tr>
<tr><td>• HREF Configurations
• CLUE: HRRRE
• Hail Guidance
• Deterministic CAMs (FV3, UM, HRRR)
• TTU Sensitivity-Based Ensemble Subsetting</td><td>• CLUE: Physics Experiment
• CLUE: FV3 Physics
• Met Office UM Evaluation
• CLUE: Microphysics
• Ensemble Object-Based Visualization</td></tr>
<tr><td colspan="2">1500 – 1600: <b>Short-term Outlook Update</b></td></tr>
<tr><td>• Update 4-h probabilistic forecasts of tornado, wind, and hail valid 21-01Z using SPC Short-Term Hazard Guidance and NEWS-e*</td><td>• Utilize NEWS-e to generate preliminary and final hourly probabilistic forecasts of total severe valid 21-22, 22-23, and 23-00Z*</td></tr>
<tr><td colspan="2">* Denotes forecasts also made by participants using the web drawing tool on Chromebooks.</td></tr>
</table>

Table A2 Weekly participants during SFE2018. Facilitators/leaders for SFE2018 included: Adam Clark (NSSL), Kent Knopfmeier (CIMMS/NSSL), Israel Jirak (SPC), Jack Hales (retired SPC), Andy Dean (SPC), Jessica Choate (CIMMS/NSSL), Steve Willington (UKMO), Burkely Gallo (OU/NSSL), and MacKenzie Krojac (OU/NSSL).

Week 1	**Week 2**	**Week 3**	**Week 4**	**Week 5**
April 30-May 4	**May 7-11**	**May 14-18**	**May 21-25**	**May 29-June 1**
Eric Loken (OU)	Nate Snook (OU/CAPS)	Steve Willington (Met Office)	Steve Willington (Met Office)	Steve Willington (Met Office)
Christina Kalb (DTC)	Brad Grant (WDTD)	Sarah Bull (Met Office)	Sarah Bull (Met Office)	Sarah Bull (Met Office)
Brian Ancell (TTU)	Shannon Rees (GFDL)	Michael Lewis (Met Office)	Michael Lewis (Met Office)	Michael Lewis (Met Office)
Aaron Hill (TTU)	Andy Hazelton (GFDL)	Jason Otkin (CIMSS)	Harald Richter (BoM)	Justin Gibbs (WDTD)
Victor Gensini (NIU)	Bill Gallus (ISU)	Greg Thompson (NCAR)	Lance Bosart (SUNYA)	Tara Jensen (DTC)
Jamie Wolff (DTC)	Bill Gallus student (ISU)	Amanda Burke (OU)	Massey Bartolini (SUNYA)	Clark Evans (UWM)
Terra Ladwig (GSD)	Ryan Sobash (NCAR)	Brian Ancell (TTU)	Marshall Pfahler (SUNYA)	David Nevius (UWM)
Dave Turner (GSD)	Brian Ancell (TTU)	Austin Coleman (TTU)	Craig Schwartz (NCAR; M-W)	Austin Coleman (TTU)
Tracy Dorian (EMC)	Aaron Hill (TTU)	Brian Kolts (FirstEnergy)	Glen Romine ? (NCAR)	Pete Wolf (NWS JAX)
Scott Rentschler (557WW)	Eric James (GSD)	Becky Adams-Selin (AER)	Austin Coleman (TTU)	Jeff Beck (DTC/GSD)
Dan Leins (NWS TWC)	Trevor Alcott (GSD)	John Brown (GSD)	Ed Szoke (GSD)	Michelle Harrold (DTC)
Austin Harris (WDTD)	Alicia Bentley (EMC)	Jeff Duda (GSD)	Curtis Alexander (GSD)	Isidora Jankov (GSD) ?
Brittany Peterson (NWS FGF)	Geoff Manikin (EMC)	Eric Aligo (EMC)	Logan Dawson (EMC)	Ed Strobach (EMC)
Michael Strickler (NWS RAH)	Robert Hart (NWS CRP)	Glen Romine (NCAR)	Ben Blake (EMC)	Hugh Morrison (NCAR, T-W)
Dave Imy (retired SPC)	Darren Van Cleave (NWS SLC)	Jaret Rogers (NWS PSR)	Andy Hatzos (NWS ILN)	Matthew Jackson (NWS TFX)
Colby Neuman (NWS PQR)	John Allen (CMU)	Matthew Friedlein (NWS LOT)	Keith Sherburn (NWS UNR)	Jeff Milne (OU/CIMMS/SPC)
Caleb Grunzke (CIMMS/SPC)	John Gagan (NWS MKX)	Jason Davis (NWS BMX)	Brian Squitieri (SPC)	Ryan Solomon (AWC)
	Mike Evans (WFO ALY)	Brendon Ruben-Oster (WPC)		
	Nathan Wendt (SPC)			