https://doi.org/10.25923/zamb-xc74

SPRING FORECASTING EXPERIMENT 2023

Conducted by the

EXPERIMENTAL FORECAST PROGRAM

of the

NOAA HAZARDOUS WEATHER TESTBED

https://hwt.nssl.noaa.gov/sfe/2023

Hybrid Experiment
1 May – 2 June 2023

Preliminary Findings and Results

Adam J. Clark[2,4], Israel L. Jirak[1], Tim Supinie[1], Kent Knopfmeier[2,3], Jake Vancil[1,3], David Jahn[1,3], David Harrison[1,3], Allie Brannan[1,3], Chris Karstens[1], Eric Loken[2,3], Nathan Dahl[1,3], Makenzie Krocak[3,4,5], David Imy[2], Andy Wade[1,3], Jeffrey Milne[1,3,4], Kimberly Hoogewind[2,3], Pamela Heinselman[2,4], Montgomery Flora[2,3], Joshua Martin[2,3], Brian Matilla[2,3], Joey Picca[1,3], Patrick Skinner[2,3], Patrick Burke[2]

(1) NOAA/NWS/NCEP Storm Prediction Center, Norman, Oklahoma
(2) NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma
(3) Cooperative In Institute for Severe and High-Impact Weather Research and Operations, University of Oklahoma, Norman, Oklahoma
(4) School of Meteorology, University of Oklahoma, Norman, Oklahoma
(5) Institute for Public Policy Research and Analysis, University of Oklahoma, Norman, Oklahoma

Table of Contents

List of Figures

List of Tables

Executive Summary

The Hazardous Weather Testbed (HWT) is a space in the National Weather Center Building in Norman, Oklahoma that facilitates forecasting experiments testing new concepts, tools, and algorithms developed at NOAA's National Severe Storms Laboratory (NSSL), Storm Prediction Center (SPC), and their partner institutions. Conducted annually during the peak severe weather season since 2000, the Spring Forecasting Experiment, or SFE, is the longest running HWT experiment. The SFEs are co-led by SPC and NSSL and aim to accelerate research to operations through testing new severe weather prediction tools and forecasting methods, studying how end-users apply severe weather guidance, and facilitating experiments for optimizing convection-allowing model ensemble design to inform NOAA's Unified Forecast System (UFS). The wealth of severe weather forecasting and research expertise at the National Weather Center, combined with state-of-the-art visualization tools, well-designed experiments, and valuable collaborations have made the annual SFEs one of the most productive and well-respected weather forecasting experiments in the world. SFE 2023 results have particular importance as NOAA's UFS initiative moves forward with the Rapid Refresh Forecast System (RRFS), NOAA's first formally designed convection-allowing model ensemble, which is scheduled for operational implementation in 2025.

Figure 1. Scenes from the 2023 NOAA Hazardous Weather Testbed Spring Forecasting Experiment.

1. Introduction

The 2023 Spring Forecasting Experiment (2023 SFE) was conducted from 1 May – 2 June by the Experimental Forecast Program (EFP) of the NOAA/Hazardous Weather Testbed (HWT), and was co-led by the NWS/Storm Prediction Center (SPC) and OAR/National Severe Storms Laboratory (NSSL). Additionally, important contributions of convection-allowing models (CAMs) were made by NOAA collaborators: Global Systems Laboratory (GSL), Environmental Modeling Center (EMC), and Geophysical Fluid Dynamics Laboratory (GFDL); and the National Center for Atmospheric Research (NCAR) and the National Aeronautics and Space Administration (NASA). Participants included over 127 forecasters, researchers, model developers, university faculty, and graduate students from around the world (see Table A1 in the Appendix). After three years of virtual experiments, SFE 2023 marked a return to in-person participation and was also the first hybrid SFE, with 50 of the 127 participants contributing remotely. As in previous years, the 2023 SFE aimed to test emerging concepts and technologies designed to improve the prediction of hazardous convective weather, consistent with the Forecasting a Continuum of Environmental Threats (FACETs; Rothfusz et al. 2018) and Warn-on Forecast (WoF; Stensrud et al. 2009) visions. Below are goals from the 2023 HWT SFE for product and service improvements and applied science activities.

Product and Service Improvements:

- Explore the ability to provide enhanced information on the conditional intensity of tornado, wind, and hail events by delineating areas expected to fall within four conditional intensity groups (CIG) defined as: no CIG, CIG 0, CIG 1, and CIG 2, for experimental outlooks covering Days 1, 2, & 3.
- Explore the ability to provide enhanced probabilistic information for Day 4 lead times by producing experimental outlooks for any type of severe hazard similar to current operational Day 3 outlooks.
- Test the utility of WoFS for updating coverage and conditional intensity full-period hazards forecasts valid 2100-1200 UTC.
- Explore how WoFS and other CAMs can be used in watch-to-warning scale forecasting applications with two separate activities focused on using this guidance for generating Mesoscale Discussions (MDs).

Applied Science Activities:

- Calibrated Guidance:
    - Evaluate the utility of several methods, including machine-learning approaches, for producing calibrated hazard guidance based on the HREF.

    - Compare and assess ML-based hazard probabilities using High-Resolution Rapid Refresh (HRRR) forecasts as input with and without convective mode information in the predictors.
    - Evaluate and compare two different methods for producing calibrated severe weather guidance at 3-7 day lead times using random forests with predictors from the Global Ensemble Forecast System (GEFS).
- Deterministic CAMs:
    - Scrutinize differences between the RRFS control member and the operational HRRR.
    - Conduct direct comparisons of storm attribute and environment fields in RRFS and HRRR for short lead times in which the data assimilation strongly impacts the forecasts, and longer lead times in which the data assimilation is less important.
    - Compare and assess the skill and utility of the primary deterministic CAMs provided by each SFE 2023 collaborator for Day 1 & 2 lead times.
    - Evaluate three configurations of MPAS runs initialized from HRRR or RRFS.
    - Examine whether decreasing horizontal grid-spacing from 3- to 1-km in Weather Research and Forecasting (WRF) model simulations provides benefits for tornado prediction and the strength of convective wind gusts.
- CAM Ensembles:
    - Compare various versions of the Rapid Refresh Forecast System (RRFS) ensemble to identify strengths and weaknesses of different configuration strategies. These comparisons were conducted within the framework of the Community Leveraged Unified Ensemble discussed below. Additional baseline comparisons were made using the operational High-Resolution Ensemble Forecast System version 3 (HREFv3).
    - Evaluate and compare the utility of global-with-nest CAM ensemble configurations using the Finite Volume Cubed Sphere (FV3) model and the Model for Prediction Across Scales (MPAS) for medium range severe weather prediction (i.e., Days 3-7).
- Analyses:
    - Compare and assess different versions of the 3D real-time mesoscale analysis (3D-RTMA) system that use different sources for the background first guess.
    - Test WoFS-based analyses of 80-m maximum winds, 2-5 km AGL updraft helicity, and column-maximum updraft speed as a potential verification source for severe weather.
- Funded Projects:

- Compare and assess different machine-learning approaches for estimating the likelihood of wind damage reports being associated with gusts ≥ 50 knots.
- Assess the utility of grid-based, ML-derived probabilities that use input from ProbSevere and WoFS to produce short-term calibrated severe hazard guidance at lead times up to 3 hours.

A suite of state-of-the-art experimental CAM guidance contributed by our large group of collaborators was critical to the 2023 SFE. For the eighth consecutive year, these contributions were formally coordinated into a single ensemble framework called the Community Leveraged Unified Ensemble (CLUE; Clark et al. 2018). The 2023 CLUE was constructed by having all groups coordinate as closely as possible on model specifications (e.g., version, grid-spacing, vertical levels, physics, etc.), domain, and post-processing so that the simulations contributed by each group could be used in controlled experiments. This design allowed us to conduct several experiments to aid in identifying optimal configuration strategies for CAM-based ensembles. The 2023 CLUE included 40 members using 3-km grid-spacing, as well as a single member using 1-km grid-spacing, which allowed for several unique experiments. The 2023 SFE activities also involved testing the WoFS for the seventh consecutive year. More information on all of the modeling systems run for the 2023 SFE is given below.

This document summarizes the activities, core interests, and preliminary findings of the 2023 SFE. More detailed information on the organizational structure and mission of the HWT, model and ensemble configurations, and information on various forecast tools and diagnostics can be found in the operations plan (https://hwt.nssl.noaa.gov/sfe/2023/docs/HWT_SFE2023_operations_plan_v2.pdf). The remainder of this document is organized as follows: Section 2 provides an overview of the models and ensembles examined during the 2023 SFE along with a description of the daily activities, Section 3 reviews the preliminary findings of the 2023 SFE, and Section 4 contains a summary of these findings and some directions for future work.

2. Description

2.1 Experimental Models and Ensembles

A total of 69 unique CAMs were run for the 2023 SFE, of which 41 were a part of the CLUE system. Other CAMs outside of the CLUE were contributed by NSSL (WoFS) and EMC (HREFv3). Forecasting activities during the 2023 SFE emphasized the use of CAM ensembles [i.e., HREF, Rapid Refresh Forecasting System (RRFS) prototypes, and WoFS] in generating experimental probabilistic forecasts of individual severe weather hazards. Additionally, the 2023 CLUE configuration enabled numerous scientific evaluations focusing on model sensitivities and various ensemble configuration strategies.

To put the volume of CAMs run for 2023 SFE into context, Figure 2 shows the number of CAMs run for SFEs since 2007, which was the first year CAM ensembles were

contributed to the SFE. In general, Figure 2 shows an increasing trend through 2019 and then stabilization around 75 CAMs. The consolidation of members into the CLUE has made this large volume of CAMs more manageable and has facilitated more controlled scientific comparisons.

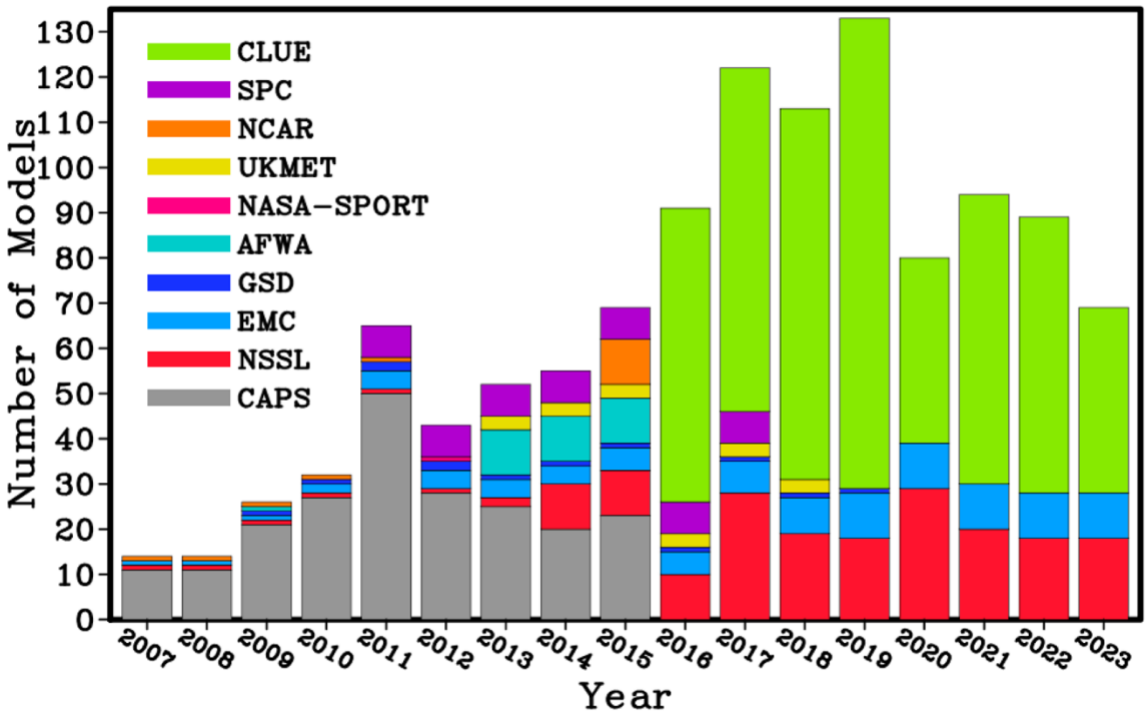

Figure 2. Number of CAMs run for SFEs since 2007. The different colored stacked bars indicate the contributing agencies.

2.1.1 The Community Leveraged Unified Ensemble (CLUE)

The 2023 CLUE is a carefully designed ensemble with subsets of members contributed by NOAA groups at NSSL, GFDL, GSL, and EMC, and the non-NOAA groups of NCAR and NASA. The 40 CLUE members with 3-km grid-spacing have a CONUS domain, while the single 1-km member has a 2/3 CONUS domain. Depending on the CLUE subset, forecast lengths range from 18 to 192 h. To ensure consistent post-processing, visualization, and verification, CLUE contributors output all model fields to the same grid using the Unified Post Processor (UPP; available at http://www.dtcenter.org/upp/users/downloads/index.php). All groups output a set of storm-based, hourly-maximum diagnostics including fields such as updraft helicity (UH) over various layers, updraft speed, and hail size, as well as standard CAM diagnostics like simulated reflectivity and precipitation. A full list of members, output fields, and further details on ensemble configurations are provided in the 2023 operations plan (https://hwt.nssl.noaa.gov/sfe/2023/docs/HWT_SFE2023_operations_plan_v2.pdf). Table 1 provides a summary of each CLUE subset.

Clue Subset	# of mems	IC/LBC perts	Mixed Physics	Data Assimilation	Dynamical Core	Agency	Init. Times (UTC)	Forecast Length (h)	Domain
RRFS	10	EnKF	no	Hybrid 3DEnVar	FV3	EMC/GSL	00-23	60/18	CONUS
RRFSphys	9	EnKF	yes	Hybrid 3DEnVar	FV3	EMC/GSL	00-23	60/18	CONUS
NSSL1	1	none	no	HRRR ICs	ARW	NSSL	00	36	2/3 CONUS
NSSL-MPAS	3	none	no	HRRR or RRFS ICs	MPAS	NSSL	00	48	CONUS
GFDL-FV3	1	none	no	GFS cold start	FV3	GFDL	00	126	CONUS
NASA-FV3	1	none	no	GEOS-DA	FV3	NASA	00	120	CONUS
NCAR-FV3	10	GEFS	no	GEFS cold start	FV3	NCAR	00	192	CONUS
NCAR-MPAS	5	GEFS	no	GEFS cold start	MPAS	NCAR	00	132	CONUS

Table 1. Summary of the 8 unique subsets that comprise the 2023 CLUE. For the RRFS and RRFSphys CLUE Subsets, 00, 06, 12, & 18 UTC initializations have 60 h forecast lengths & the entire 10-member ensemble is run; for all other RRFS & RRFSphys initialization times, only the control member is initialized with forecast lengths of 18-h.

The design of the 2023 CLUE allowed for several unique experiments that examined issues immediately relevant to the design of a NCEP/EMC operational CAM ensemble. The primary groups of experiments are listed as follows:

RRFS vs. HRRR/HREF

- **Description:** Deterministic and ensemble components of RRFS were compared to their operational counterparts HRRR and HREF, respectively, for Day 1 & 2 lead times. Additional comparisons were made during the first 12 h of the forecasts to evaluate the effectiveness of the data assimilation in each system.
- **Goal:** Evaluate RRFS skill and utility relative to HREF and HRRR to assess RRFS progress toward potential operational implementation in 2025.
- **CLUE subsets:** RRFS

RRFS Configuration Strategies

- **Description:** RRFS with 6- and 12-hour time lagging, as well as RRFS with mixed-physics were compared to RRFS with single physics.
- **Goal:** Identify a strategy within the UFS framework that performs as good as or better than HREFv3, so that it can serve as a replacement in NCEP's production suite.
- **CLUE Subsets:** RRFS and RRFSphys

Medium-Range CAM Ensembles

- **Description:** NCAR provided a 10-member, FV3-based, CAM ensemble with forecasts to 7 days, as well as a 5-member, MPAS-based CAM ensemble with forecasts to 5 days.
- **Goal:** Evaluate and compare the utility of CAM ensembles for medium-range severe weather forecasting.

- **CLUE Subsets:** NCAR-FV3 and NCAR-MPAS

Enhanced Resolution

- **Description:** NSSL ran two versions of WRF-ARW with 3- and 1-km grid-spacing.
- **Goal:** Examine grid-spacing sensitivity and assess whether enhanced resolution can provide improved severe weather guidance with particular attention given to depiction of storm structure and mode, as well as low-level rotation diagnostics.
- **CLUE Subsets:** NSSL3 and NSSL1

MPAS Configurations

- **Description:** NSSL ran three 00Z versions of convection-allowing MPAS runs over CONUS with varied ICs (HRRR or RRFS) and microphysics schemes (Thompson or NSSL).
- **Goal:** Assess sensitivities and performance differences in MPAS configurations with different initialization and microphysics.
- **CLUE Subsets:** NSSL-MPAS

3D-RTMA Background

- **Description:** Two hourly versions of 3D-RTMA that used a different background first-guess were compared.
- **Goal:** Assess the impact of the background first guess on the final analysis.
- **3DRTMA Versions:** HRRR and RRFS

2.1.2 The High-Resolution Ensemble Forecast System Version 3 (HREFv3)

HREFv3 is a 10-member CAM ensemble that was implemented in operations 11 May 2021 and forecasts can be viewed at: http://www.spc.noaa.gov/exper/href/. HREFv3 replaced HREFv2.1. The design of HREFv3 originated from the SSEO, which demonstrated skill for six years in the HWT and SPC prior to initial operational implementation in 2017. In HREFv3, the HRW NMMB simulations have been replaced with HRW FV3. The member configuration diversity in HREFv3 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 several years.

2.1.3 NSSL Cloud-Based Warn-on-Forecast System (cb-WoFS)

Cloud-based Warn-on-Forecast (cb-WoFS) is the next WoFS iteration, upgraded to use current technologies in containerization and cloud computing. The entire WoFS application was rebuilt on top of multiple Platform-as-a-Service and Infrastucture-as-a-Service technologies on the Azure platform and the WRF model itself rebuilt to run in containers optimized for HPC. With the new cb-WoFS interface, administrators can easily

configure the domain and dynamically create an HPC infrastructure for the run, and upon completion, tear it down, thereby reducing costs by only paying for used resources. Another benefit is that as Azure continues to add new, updated computer core types from chip manufacturers, these options are passed down to Azure customers, giving cb-WoFS operators the choice of running on the latest technologies. All parts of WoFS have been rebuilt for scalability: the containerized WRF can be executed on any node, the post-processing is built on high performance queues and containerized, so any number of post-processing jobs can run concurrently.

The cb-WoFS is a rapidly-updating 36-member, 3-km grid-spacing WRF-based ensemble data assimilation and forecast system. The cb-WoFS forecasts are initialized every 30 minutes and used to produce very short-range (0-6/0-3 h at top/bottom of the hour) probabilistic forecasts of individual thunderstorms and their associated hazardous weather phenomena such as supercell hail, high winds, flash flooding, and supercell thunderstorm rotation. The 900-km x 900-km daily cb-WoFS domain targeted the primary region where severe weather was anticipated. For SFE 2023, WoFS has the capability to run over two different domains. A second domain was only implemented when there were two separate regions where severe weather was expected (e.g., Midwest and East Coast), or when there was a very large single area for which two domains were needed to cover the entire risk area.

The starting point for each day's experiment was the High-Resolution Rapid Refresh Data Assimilation System (HRRRDAS) and the 1200 UTC HRRR forecast provided by NCO/GSL. A 1-h forecast from the 1400 UTC, 36-member, hourly-cycled HRRRDAS analysis provided the ICs for cb-WoFS. Boundary conditions were perturbed HRRR forecasts, where perturbations from the 0600 UTC GEFS were added to the 1200 UTC HRRR forecasts. The GEFS perturbations were scaled such that the ensemble spread at the lateral boundaries was similar to that provided previously by the experimental HRRR ensemble.

2.2 Daily Activities

SFE 2023 activities were focused on forecasting severe convective weather and evaluating the previous day's model forecasts. A summary of evaluation activities and forecast products can be found below while a detailed schedule of daily activities is contained in the appendix (Table A2). Note, when referencing the times in this document at which experiment activities occurred, we use Central Daylight Time (CDT), which is the time zone in which the HWT facility and SFE organizers are based. However, it is worth noting that many of our virtual participants were located in different time zones as far away as the United Kingdom and Australia, so their local time was quite different.

2.2.1 Forecast and Model Evaluations

SFE 2023 featured a period of formal evaluations from 9-11am CDT Tuesday Friday for the first four weeks and Wednesday-Friday for the fifth week for a total of 19 days of evaluation. The evaluations involved comparisons of different ensemble diagnostics, CLUE ensemble subsets, HREFv3, and WoFS. Additionally, the evaluations of yesterday's experimental forecasts products were conducted during this time, which involved comparing the experimental products to observed local storm reports (LSRs), NWS warnings, and Multi-Radar, Multi-Sensor (MRMS; Smith et al. 2016) radar reflectivity and maximum estimated size of hail (MESH). Participants were split into Groups 1, 2, and 3, and each conducted a separate set of model evaluations. These groups were hybrid meaning that they contained a mix of in-person and virtual participants. The evaluations were categorized as "CAM (E)nsembles", "(D)eterministic CAMs", "(A)nalyses", "Funded (P)rojects", "(C)alibrated Guidance", or "(O)utlooks". The letter in parentheses combined with a number was used to label the individual evaluations in each category (e.g., E1 refers to the first CAM Ensemble evaluation). Each evaluation group conducted a mix of evaluations from each category. Participants rotated through each evaluation group at least once. Participants worked on all the surveys individually, with short discussion periods after completion of each survey. SFE facilitators were available to answer any questions, troubleshoot issues, and discuss subjective impressions of the day.

2.2.2 Experimental Forecast Products

The experimental forecasts covered a limited-area domain typically encompassing the primary severe threat area with a domain based on existing SPC outlooks and/or where interesting convective forecast challenges were expected. An exception was the Day 3 & 4 outlooks, which covered the entire CONUS. There were two periods of experimental forecasting activities during SFE 2023. The first occurred from 11:00am – 12:30pm CDT and focused on generating probabilistic outlooks for individual hazards for Days 1-3, as well as more precise information on the intensity of specific hazards. The Day 4 outlooks only covered total severe (i.e., no individual hazards or conditional intensity forecasts). Participants were split into three groups: (1) In-Person R2O, (2) In-Person Innovation, and (3) Virtual. As the naming convention suggests, in-person participants were in R2O and Innovation groups, while all virtual participants were in the Virtual group. The In-Person R2O group issued products for Day 1, the Virtual group issued products for Day 2, and the In-Person Innovation group issued products for Days 3 & 4.

In all groups, the morning forecasts were done collectively. The individual hazard forecasts mimicked the SPC operational Day 1 & 2 Convective Outlooks by producing individual probabilistic coverage forecasts of large hail, damaging wind, and tornadoes within 25 miles (40 km) of a point. The Day 1 outlooks covered the period 1800 UTC to 1200 UTC the next day, while the Days 2, 3, & 4 outlooks covered 1200 – 1200 UTC

periods. Additionally, for experimental outlooks covering Days 1, 2, & 3, conditional intensity forecasts of tornado, wind, and hail were issued, in which areas are delineated where reports that are expected to follow intensity distributions defined by conditional intensity groups. These conditional intensity forecasts are similar to those issued during SFEs 2019-2022. The four possible conditional intensity groups (CIG) included: no CIG, CIG 0, CIG 1, and CIG 2. In plain language, CIG 0 refers to a typical severe weather day, where significant severe weather is unlikely, CIG 1 areas indicate where significant severe weather is possible, and CIG 2 areas indicate where high impact significant severe weather is expected. All groups had access to all available operational and experimental guidance products for issuing their outlooks.

The second period of experimental forecasting activities occurred during the 2-4pm CDT time period. From 2-2:15pm CDT, a weather briefing led by Dave Imy was conducted for all participants during which an update on current weather was given. In the In-Person R2O group, the 2:15-3:15pm CDT time period was devoted to an activity in which each participant created their own Mesoscale Discussion (MD) Product using WoFS and other available CAM guidance within the SFE Drawing Tool. Then, during the 3:15-4pm CDT time period, each In-Person R2O participant used WoFS and other available guidance to update the Day 1 individual hazard coverage and conditional intensity forecasts done earlier as a group for the period 2100 – 1200 UTC.

During the 2:15-4pm CDT time period in the In-Person Innovation Group and Virtual Group, another activity was devoted to issuing short-term, meso-beta to meso-gamma scale predictions of severe weather. In this activity, each participant issued a forecast consisting of two parts: (1) a geographic threat area (i.e., graphic) and (2) a text discussion. The geographic threat area was created using the WoFS web viewer drawing tool and took one of three formats: (1) A single contour highlighting a region of expected severe weather along the track of an individual storm, (2) two contours, one encompassing a broader region where severe weather is expected and the second, smaller contour outlining what is perceived as the corridor of greatest risk, or (3) A single contour that highlights a broader region where severe weather is expected. Each participant issued their first set of predictions during the 2:15-3pm CDT time period, and then from 3-3:15pm CDT each participant had an opportunity to present and discuss their product. Then, from 3:15-3:45pm CDT the outlooks and text discussions were updated with a focus on how more recent observations and more up-to-date WoFS guidance was influencing the perceived threat and confidence in the forecast. For example, did WoFS indicate increasing or decreasing likelihood of an event relative to previous guidance, or does the more recent guidance simple reinforce earlier guidance? Finally, from 3:45-4pm each participant participated in a short survey with some targeted questions on WoFS products used, changes in forecasts between 1st and 2nd hours, and overall confidence.

3. Preliminary Findings and Results

3.1 Model Evaluation – (C)alibrated Guidance

3.1.1 (C1) Day 2 12Z HREF Calibrated Tornado Guidance

A number of probabilistic calibrated tornado guidance products generated from the 12Z HREF and valid for the Day 2 period (i.e., f24-f48) were evaluated during the HWT SFE. It is worth noting that May 2023 was an abnormally quiet period for severe weather, including tornadoes. Thus, one must be careful to not overgeneralize these results; though there were a few active tornado days that were evaluated, including 11 May 2023 (Fig. 3). There were five independent tornado guidance products that were evaluated and all of them utilize the 12Z HREF as their primary numerical weather prediction input. A new product, the ensemble mean of these five probabilistic products, was also evaluated in this suite of guidance. Owing to time constraints during the HWT SFE, only active tornado days were evaluated and included in the subjective results.

In terms of the subjective ratings for the Day 2 tornado guidance products, they all had reasonably similar rating distributions, especially at the upper end (Fig. 4). The HREF/GEFS Calibrated and STP Cal MCS-TF tended to have more lower-rated tornado forecasts than the other guidance products bringing down their mean and median ratings. The HREF/GEFS Calibrated tended to have the highest peak probability magnitudes on many days and was often thought to be an overforecast while the STP Cal MCS-TF tended to have the lowest peak probability magnitudes on many days with less spatial coverage of probabilities overall. The Nadocast and Cal Ensemble Mean products tended to have slightly higher mean ratings during the SFE, but the differences were rather small.

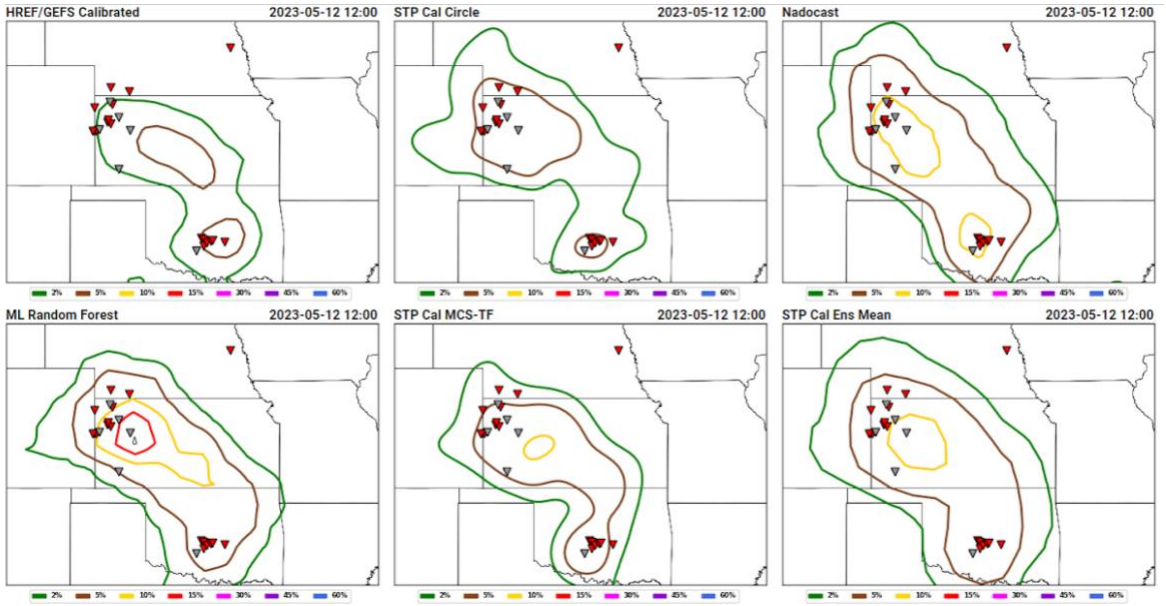

Figure 3. Example of the website comparison page for the Day 2 12 HREF calibrated tornado guidance during the 2023 HWT SFE. The different Day 2 guidance products valid for the convective day of 11 May 2023 are shown with tornado reports overlaid (gray symbols indicate "brief", "weak", and/or "landspout" tornadoes).

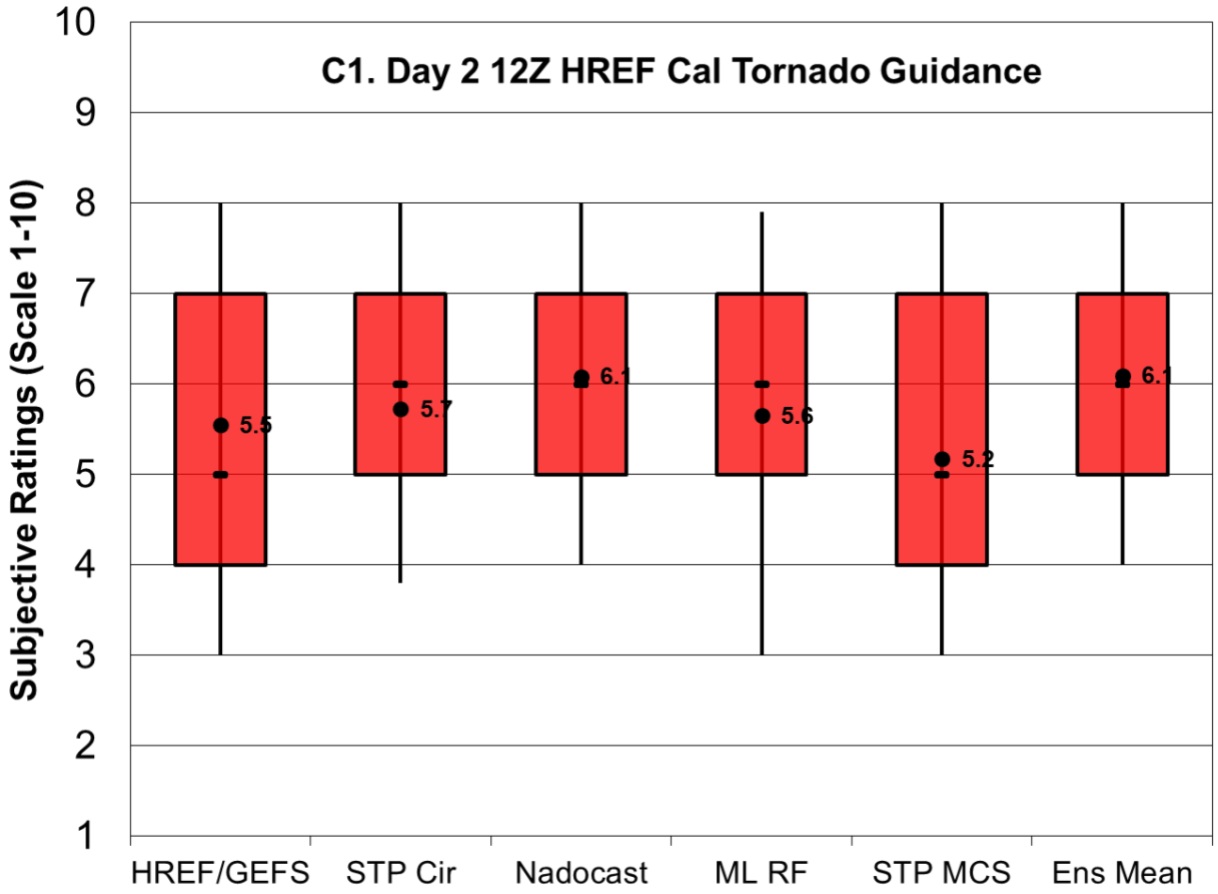

Figure 4. Distributions of subjective ratings (1-10; with 10 being best) by SFE participants of the Day 2 12Z HREF calibrated tornado guidance products: HREF/GEFS Calibrated, STP Cal Circle, Nadocast, ML Random Forest, STP Cal MCS-TF, and Cal Ensemble Mean.

3.1.2 (C2) Day 1 12Z HREF Calibrated Tornado Guidance

Similar to the evaluation of the Day 2 12Z HREF calibrated tornado guidance, the same tornado guidance products were also evaluated for the Day 1 period (i.e., f0-f24; Fig. 5). Not surprisingly, the subjective ratings were higher across the board for the shorter-range Day 1 products (c.f., Figs. 4 & 6). The STP Cal MCS-TF product was the lowest rated of the guidance suite on Day 1 with a median rating of 5 out of 10 (Fig. 6). Again, the lower probability magnitudes and more limited spatial coverage tended to hurt this product in the subjective ratings. Two products, Nadocast and the Cal Ensemble Mean, stood out as the best performing tornado guidance with the highest median ratings of 7 out of 10 during the 2023 HWT SFE.

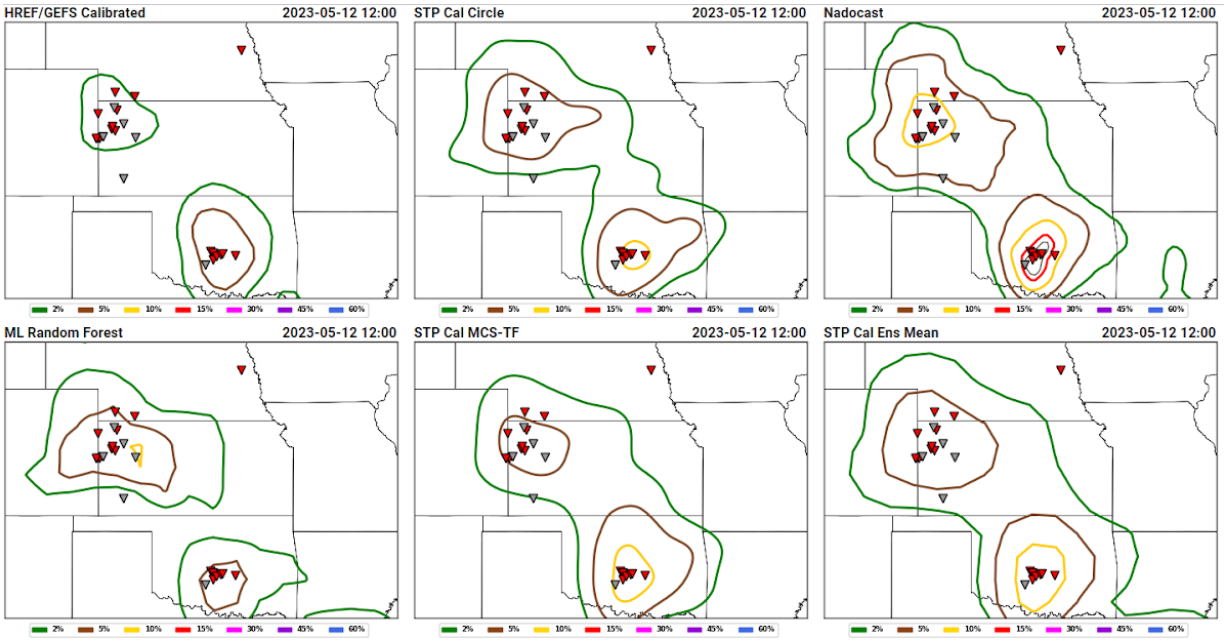

Figure 5. Example of the website comparison page for the Day 1 12 HREF calibrated tornado guidance during the 2023 HWT SFE. The different Day 1 guidance products valid for the convective day of 11 May 2023 are shown with tornado reports overlaid (gray symbols indicate "brief", "weak", and/or "landspout" tornadoes).

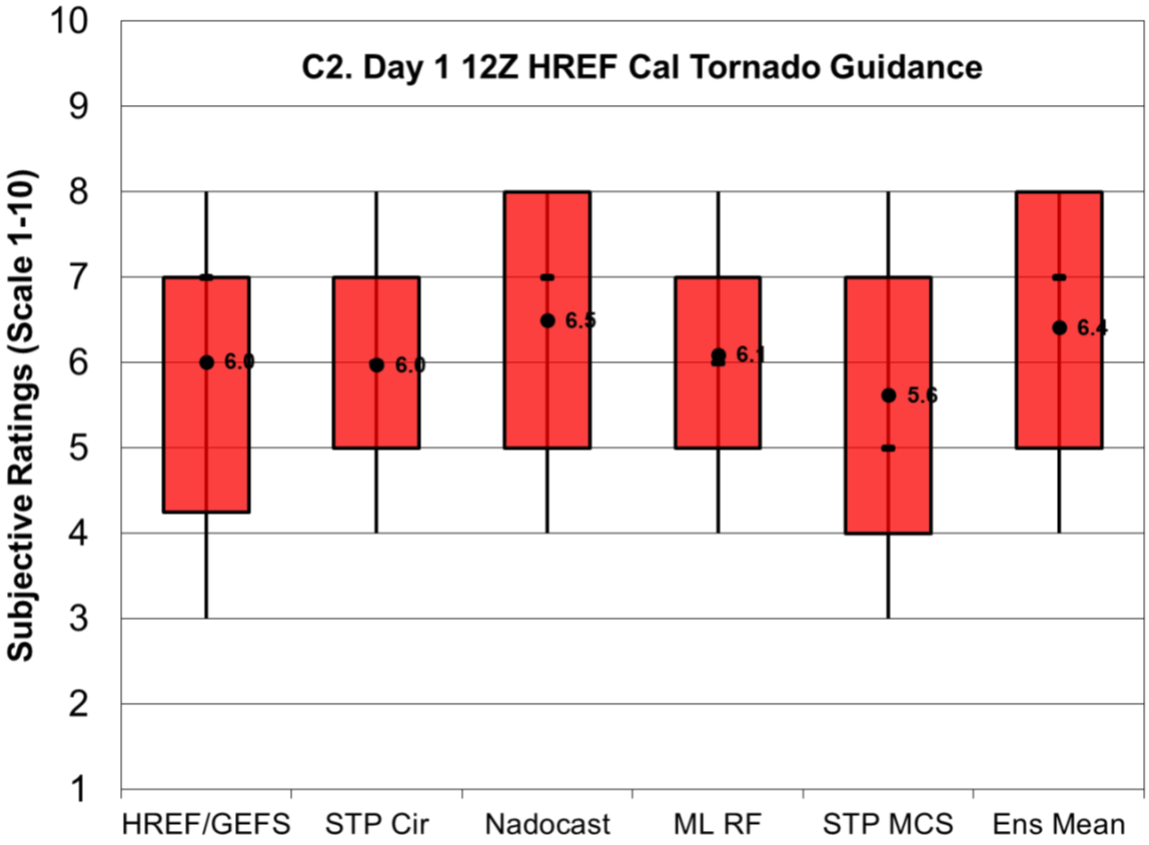

Figure 6. Distributions of subjective ratings (1-10; with 10 being best) by SFE participants of the Day 1 12Z HREF calibrated tornado guidance products: HREF/GEFS Calibrated, STP Cal Circle, Nadocast, ML Random Forest, STP Cal MCS-TF, and Cal Ensemble Mean.

3.1.3 (C3) 1630Z 4-h SPC Tornado Timing Guidance (hourly 20-12Z)

In an effort to add more specific temporal information to the Day 1 Outlook, SPC has developed Severe Timing Guidance products, which are hourly 4-h severe weather probabilities through the convective day. The Severe Timing Guidance products are consistent with and constrained by the human-issued SPC Convective Outlooks and uses HREF-based guidance to disaggregate the probabilities throughout the convective day. The current real-time SPC Timing Guidance probabilities leverage the operational HREF/SREF calibrated hazard probabilities, but with the planned retirement of the SREF in the coming years, other HREF-based guidance products were tested in the algorithm during the 2023 HWT SFE to determine the effect on the probabilistic output.

For the Tornado Timing Guidance, using Nadocast as the input to the algorithm produced the highest-rated timing guidance products (Fig. 7). The only version of the Tornado Timing Guidance that was a degradation over the baseline product (HREF/SREF) was the one using the HREF Calibrated Thunder (HREFCT). The HREFCT-based timing guidance tended to extend the probabilities too late into the overnight period when the tornado threat had diminished.

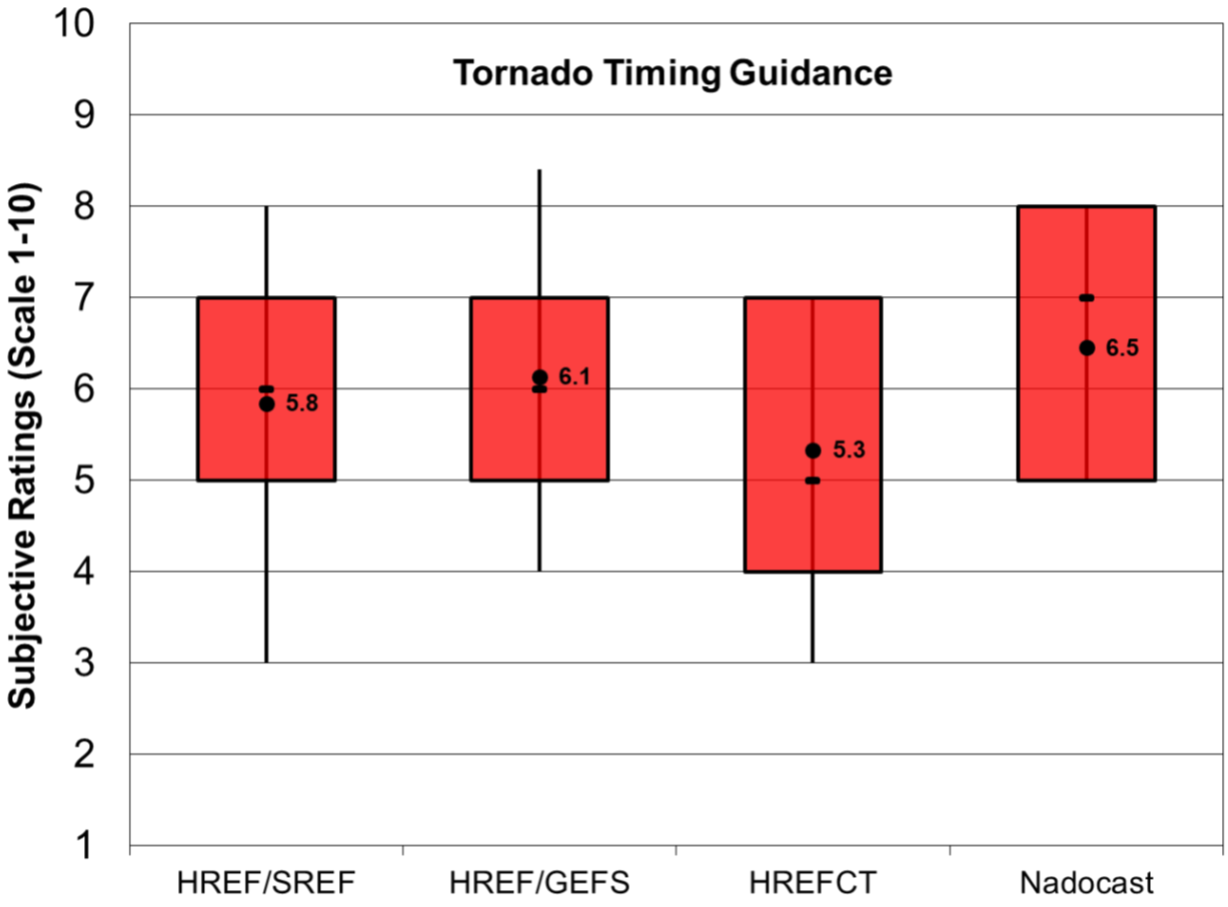

Figure 7. Distributions of subjective ratings (1-10; with 10 being best) by SFE participants of the 1630Z 4-h Tornado Timing Guidance Products based on the HREF/SREF, HREF/GEFS, HREFCT, and Nadocast.

3.1.4 (C4) Day 2 12Z HREF Calibrated Hail Guidance

Three different methods were examined that produced calibrated hail guidance for the Day 2 time period using 1200 UTC initialization HREF fields. For each method, participants were asked to evaluate the hail probabilities based on (1) magnitude, (2) areal coverage, and (3) placement, relative to the practically perfect hindcast. Then, participants assigned an overall rating on a scale of 1 (very poor) to 10 (very good). An example forecast is shown in Figure 8.

For HREF/GEFS Cal, magnitudes were most often rated about right or too low, coverages were most often rated too small or about right, and placement was usually somewhat displaced or nearly colocated (Fig. 9a). For Nadocast, magnitude, coverage, and placement were most frequently rated about right, about right, and nearly collocated, respectively (Fig. 9b). Finally, ML Random Forest had the highest frequency of about right responses for magnitude and coverage (Fig. 9c), but Nadocast placement was had the most nearly collocated responses. For the overall ratings, Nadocast had the highest average subjective rating, followed by ML Random Forest and HREF/GEFS Cal (Fig. 10).

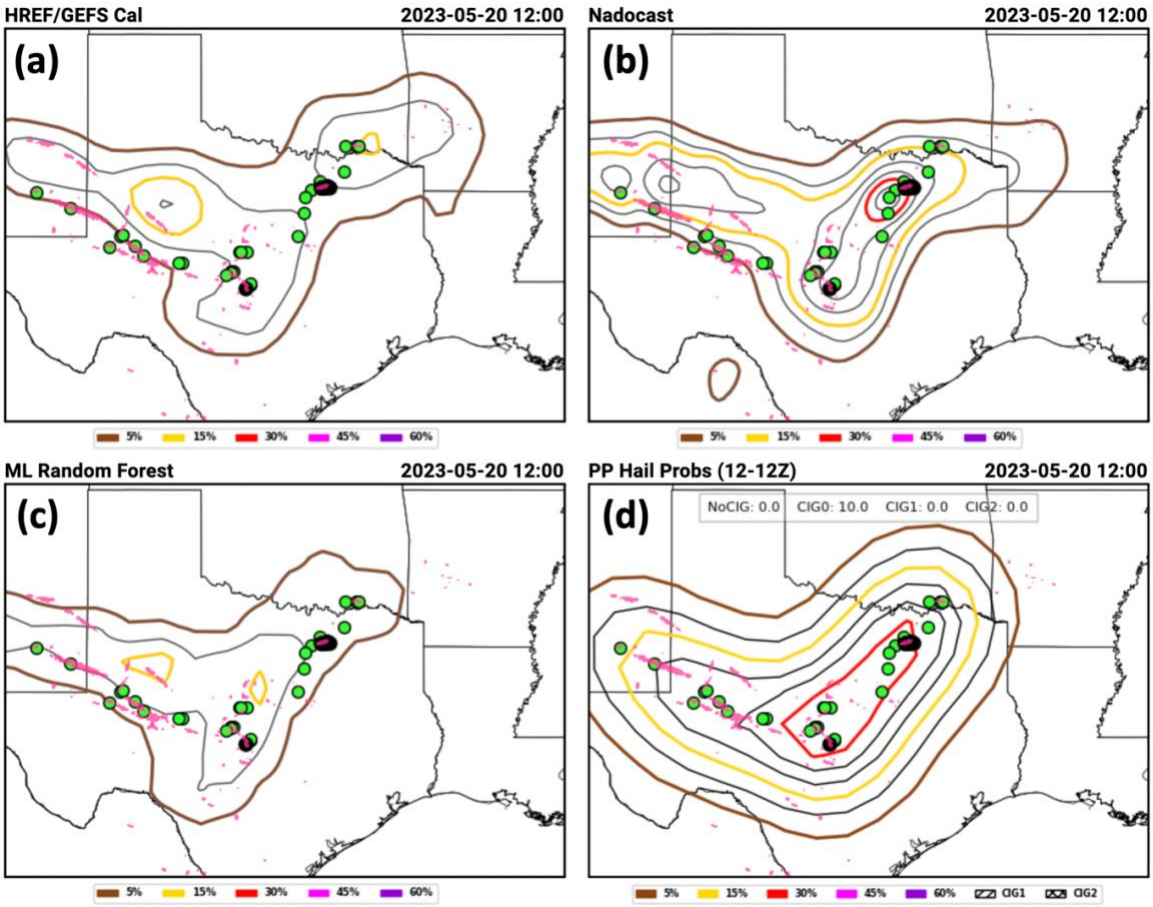

Figure 8. Day 2 Severe hail probabilities valid 1200 – 1200 UTC 19-20 May 2023 from (a) HREF/GEFS Cal, (b) Nadocast, (c) ML Random Forest, and (d) practically perfect hindcasts. In each panel, hail (green circles) and significant hail (black circles) LSRs, as well as areas of MESH ≥ 1.0-in. (pink shading) are overlaid.

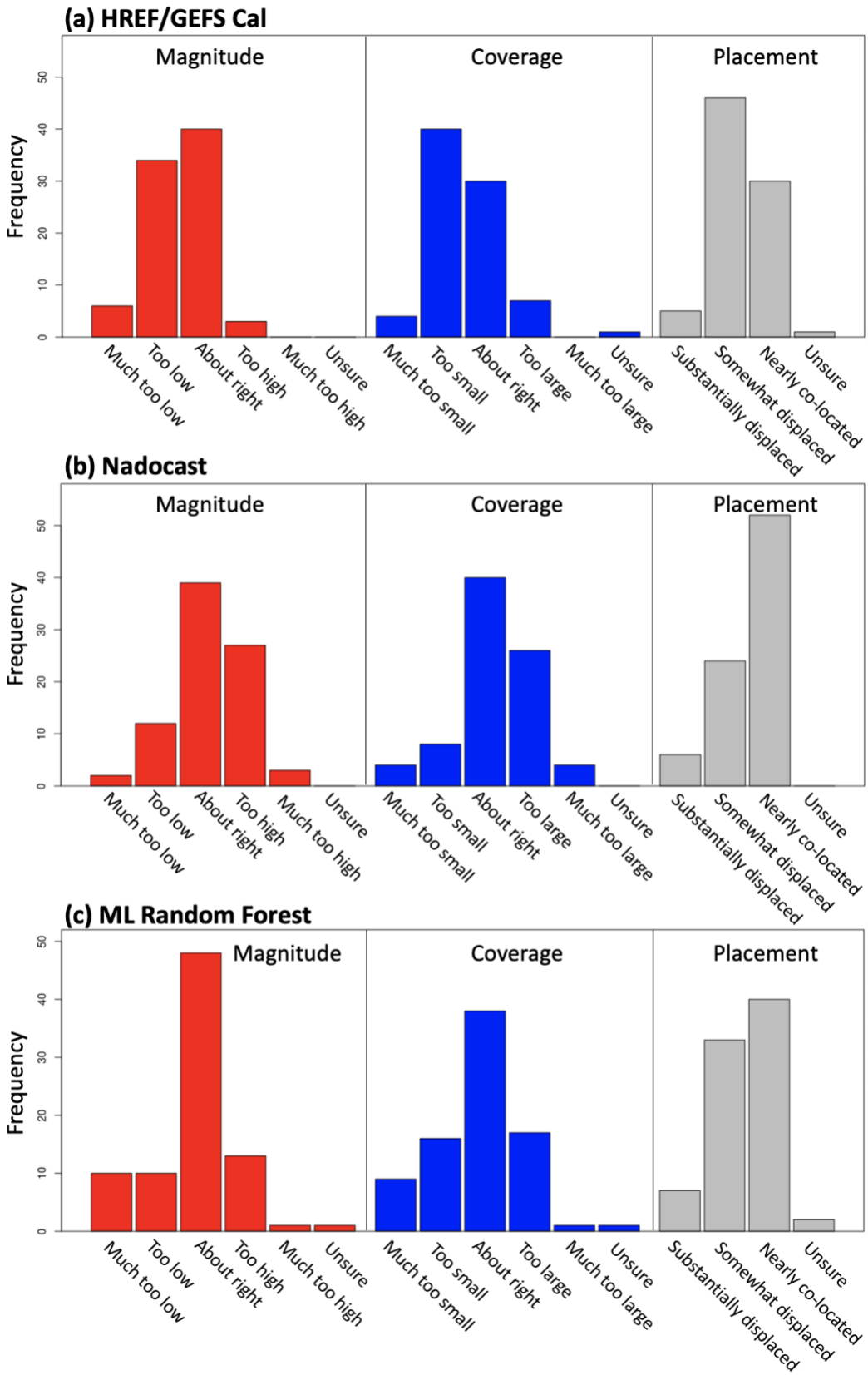

Figure 9. Response frequencies relating to magnitude, coverage, and placement of 1200 UTC HREF-based Day 2 hail probabilities derived from (a) HREF/GEFS Cal, (b) Nadocast, and (c) ML Random Forest.

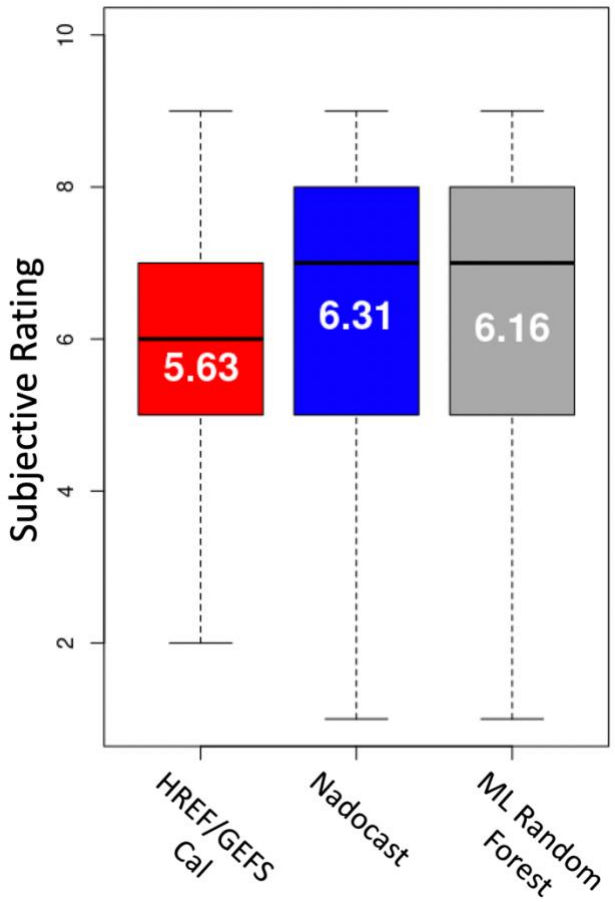

Figure 10. Box plots showing the distributions of subjective rankings by SFE 2023 participants for the overall forecast quality of 1200 UTC HREF-based Day 2 hail probabilities from HREF/GEFS Cal, Nadocast, and ML Random Forest.

3.1.5 (C5) Day 1 12Z HREF Calibrated Hail Guidance

Similar to C4, three different methods were examined that produced calibrated hail guidance using 1200 UTC initialization HREF fields, except this survey examined the Day 1 time period. For each method, participants were asked to evaluate the hail probabilities based on (1) magnitude, (2) areal coverage, and (3) placement, relative to the practically perfect hindcast. Then, participants assigned an overall rating on a scale of 1 (very poor) to 10 (very good).

For HREF/GEFS Cal, magnitudes were most often rated too low or about right, coverages were most often rated about right, and placement was usually somewhat displaced or nearly co-located (Fig. 11a). For Nadocast and ML Random Forest, magnitude, coverage, and placement were most frequently rated about right, about right, and nearly collocated, respectively (Fig. 11b & c). However, Nadocast had more frequent responses relative to ML Random Forest for about right, about right, and nearly collocated, for magnitude, coverage, and placement, respectively. For the overall ratings, Nadocast had the highest average subjective rating, followed by ML Random Forest and HREF/GEFS Cal (Fig. 12). The relative results for Day 1 closely followed those of Day 2.

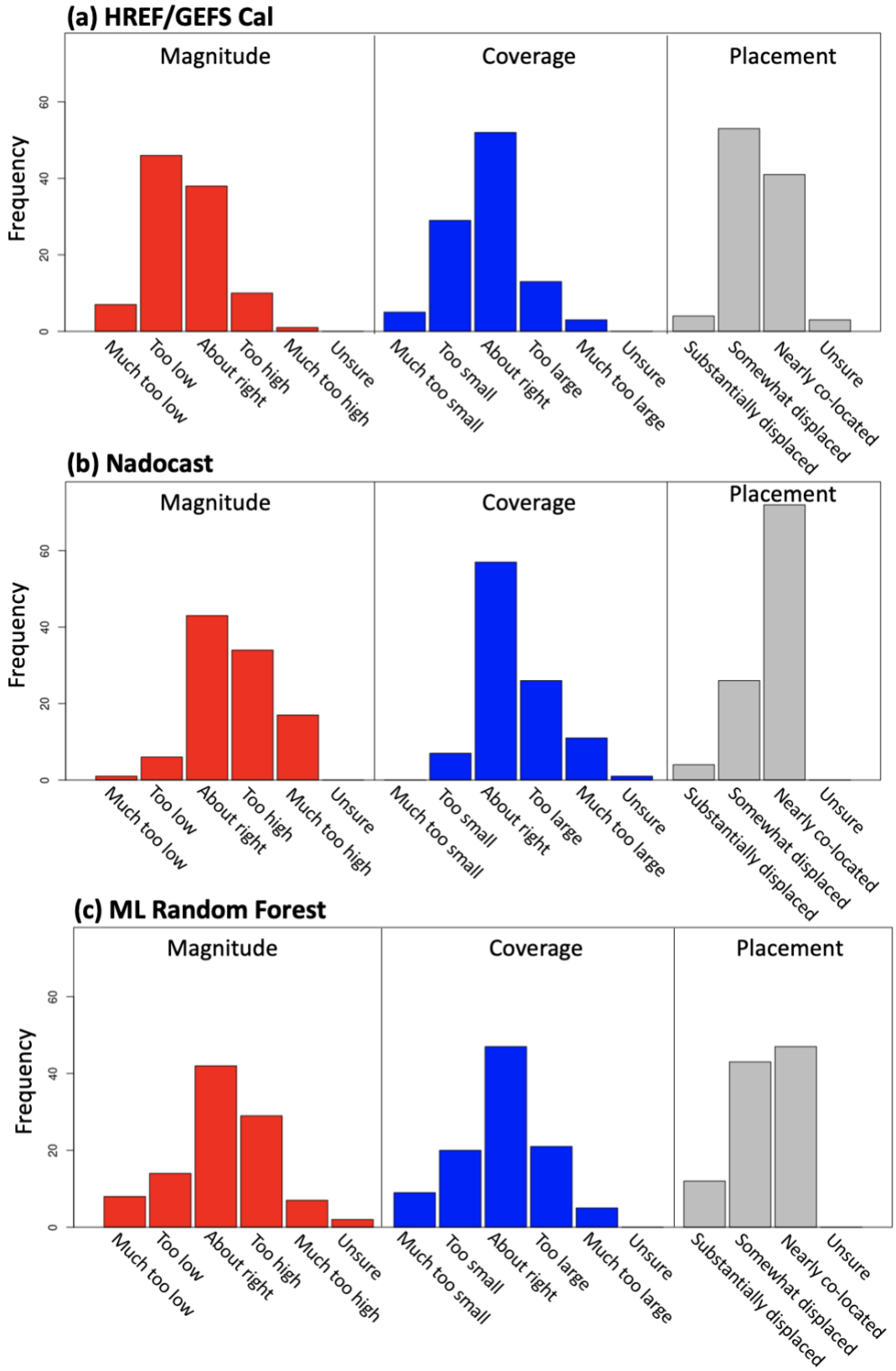

Figure 11. Response frequencies relating to magnitude, coverage, and placement of 1200 UTC HREF-based Day 1 hail probabilities derived from (a) HREF/GEFS Cal, (b) Nadocast, and (c) ML Random Forest.

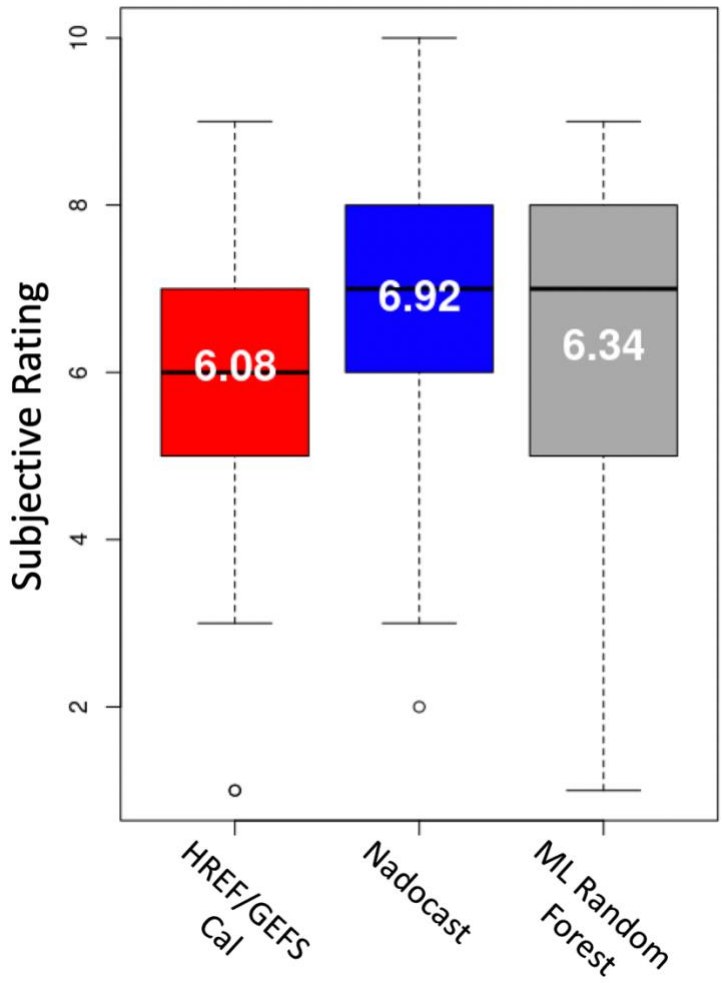

Figure 12. Box plots showing the distributions of subjective rankings by SFE 2023 participants for the overall forecast quality of 1200 UTC HREF-based Day 1 hail probabilities from HREF/GEFS Cal, Nadocast, and ML Random Forest.

3.1.6 (C6) Day 1 12Z HREF Calibrated Hail Guidance: MESH (Maximum Estimated Size of Hail)

In this survey, a version of HREF/GEFS Cal (referred to as HREF/GEFS MESH) was evaluated that was calibrated based on MESH instead of LSRs. Similarly, comparisons were made to practically perfect hindcasts computed from both MESH and LSRs, as well as only LSRs. An example forecast is shown in Figure 13. Magnitude, coverage, and placement were most frequently rated about right, about right, and nearly collocated, respectively (Fig. 14a). In addition, the mean subjective rating for HREF/GEFS MESH was 6.76 (Fig. 14b), which is an improvement relative to the 6.08 mean subjective rating of HREF/GEFS Cal. It appeared that most participants compared HREF/GEFS MESH to practically perfect hindcasts computed from only hail LSRs (e.g., Fig. 13b vs. Fig. 13c) when deciding on their ratings. Many of the survey comments noted that practically perfect hindcasts computed from both MESH and LSRs were way too high (e.g., Fig. 13d).

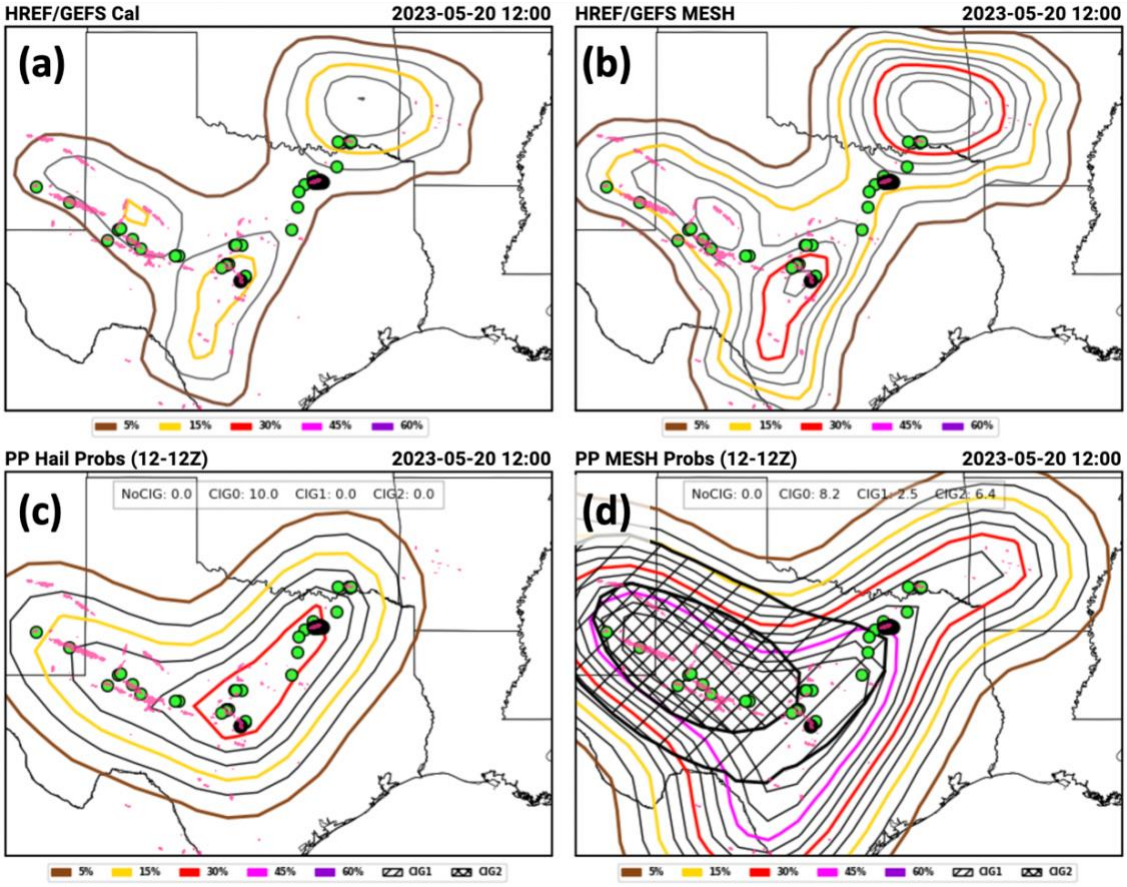

Figure 13. HREF/GEFS Cal severe hail probabilities valid 1200 – 1200 UTC 19-20 May 2023. (b) Same as (a), except for HREF/GEFS MESH. (c) Practically perfect hindcasts computed using hail LSRs, and (d) same as (c) except practically perfect hindcast computed using both LSRs and MESH. In each panel, hail (green circles) and significant hail (black circles) LSRs, as well as areas of MESH ≥ 1.0-in. (pink shading) are overlaid.

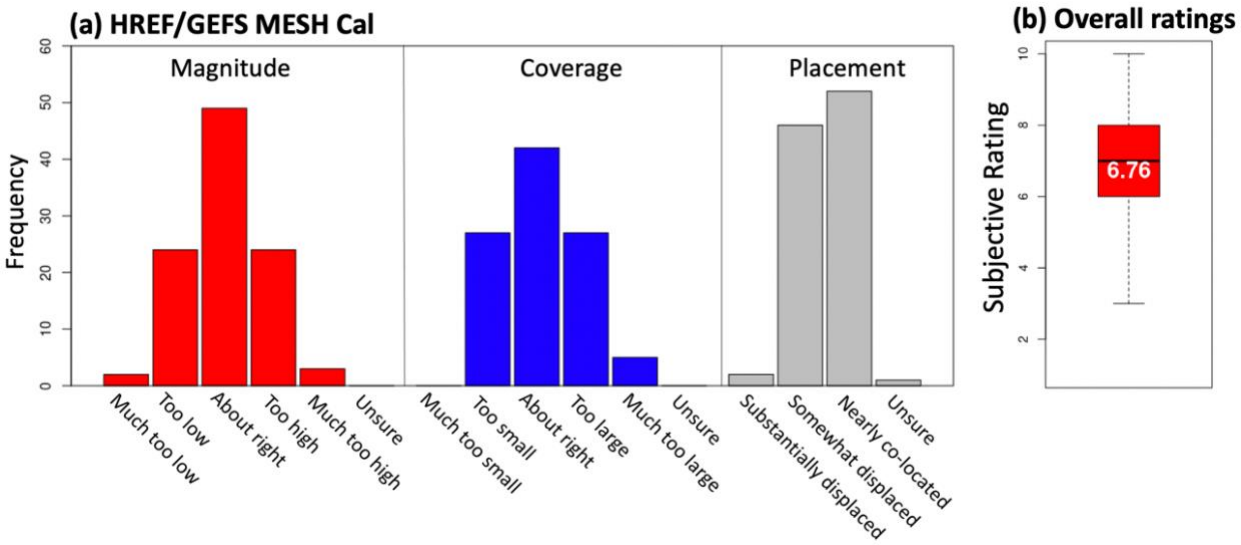

Figure 14. Response frequencies relating to magnitude, coverage, and placement of 1200 UTC HREF-based Day 1 hail probabilities derived from HREF/GEFS MESH. (b) Box plots showing the distributions of subjective rankings by SFE 2023 participants for the overall for the overall forecast quality of 1200 UTC HREF-based Day 1 hail probabilities from HREF/GEFS MESH.

3.1.7 (C7) 1630Z 4-h SPC Hail Timing Guidance (hourly 20-12Z)

For the Hail Timing Guidance, the HREF/GEFS product was rated similarly to the Nadocast product as the best performing versions of Hail Timing Guidance with median ratings of 7 out of 10. Both versions appear to be slight improvements over the current HREF/SREF baseline (Fig. 15). Similar to the Tornado Timing Guidance results, only the HREFCT version of the Hail Timing Guidance was rated subjectively lower than the HREF/SREF version, owing to a slower-than-observed decrease in probabilities regarding the risk of large hail.

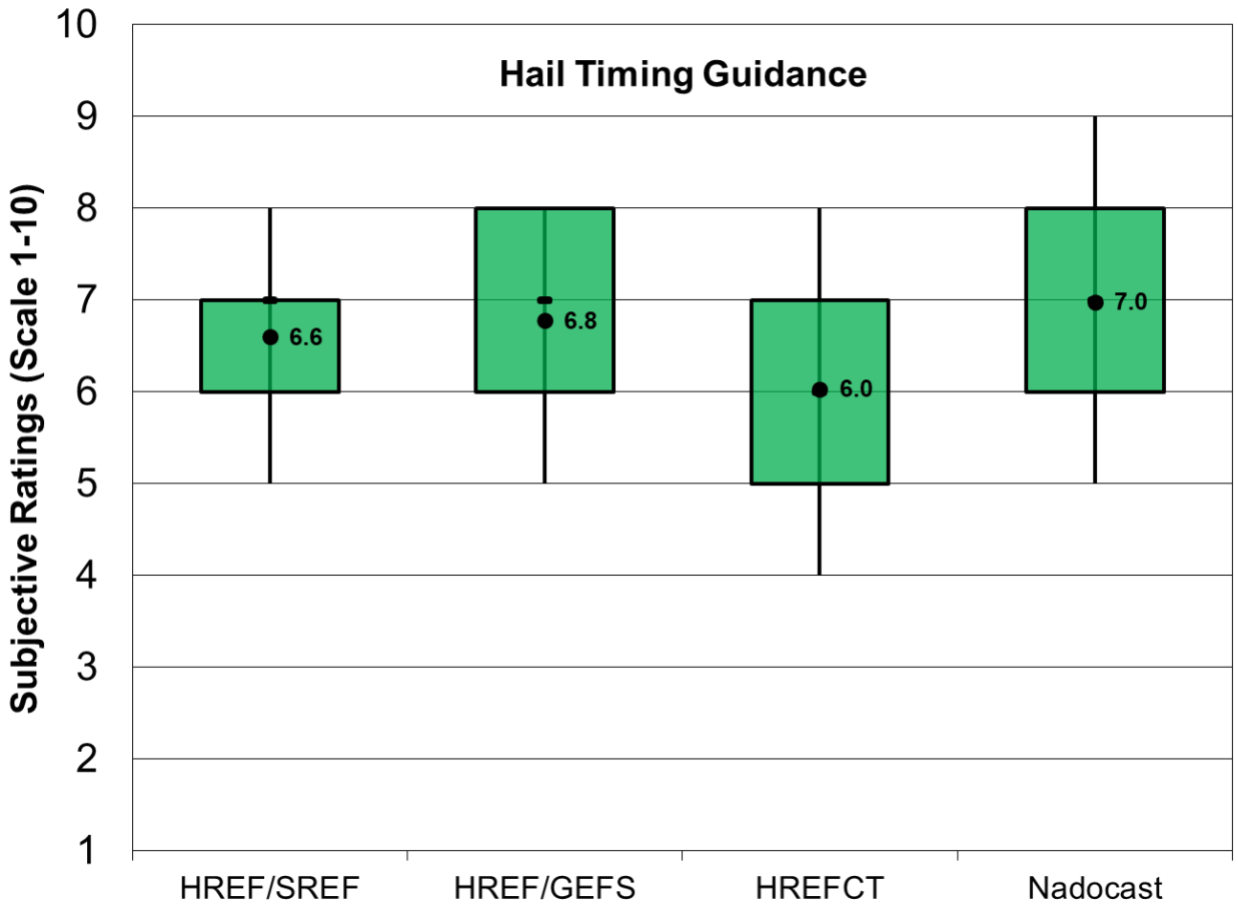

Figure 15. Distributions of subjective ratings (1-10; with 10 being best) by SFE participants of the 1630Z 4-h Hail Timing Guidance Products based on the HREF/SREF, HREF/GEFS, HREFCT, and Nadocast.

3.1.8 (C8) Day 2 12Z HREF Calibrated Wind Guidance

SFE participants evaluated four severe wind (> 50 kts) calibrated guidance products for a 24-hr period starting at 1200 UTC for the Day 2 forecast period. All products were derived from the 1200 UTC HREF. The HREF/GEFS calibration method is based on the historical frequency of severe wind reports as related to forecast storm

and environmental parameters. The ML Random Forecast (MLRF) method, uses a random forest machine-learning algorithm to generate severe wind probabilities from 22 derived storm attributes and environmental parameters. Nadocast is another machine learning method that incorporates more than 10,000 predictors derived from storm-scale and environmental parameters. Finally, the Nadocast 'adjusted' (NadoAdj) model is an alternate version of Nadocast that attempts to correct for biases in wind damage reports based on historical ratios of wind damage reports to nearby measured severe winds.

Participants subjectively evaluated the guidance products compared to observed local storm reports and NWS-issued warnings. On average, over the 19-day SFE evaluation period, MLRF and Nadocast were the top performers at Day 2 lead times with a mean score of 6.40 and 5.98, respectively (Fig. 16). This result demonstrates that the MLRF method, which uses a limited but judiciously selected list of ML predictors, performed as well as or even better than the much more complex Nadocast. The relatively low ratings attributed to NadoAdj suggests that the attempt to adjust for biases of measured wind reports was not effective within the experiment domain. Although NadoAdj was often successful in tempering over-forecast severe wind probabilities in the eastern US (where measured wind LSRs are potentially overestimated), it resulted in a potential overprediction of severe wind occurrence in the Great Plains.

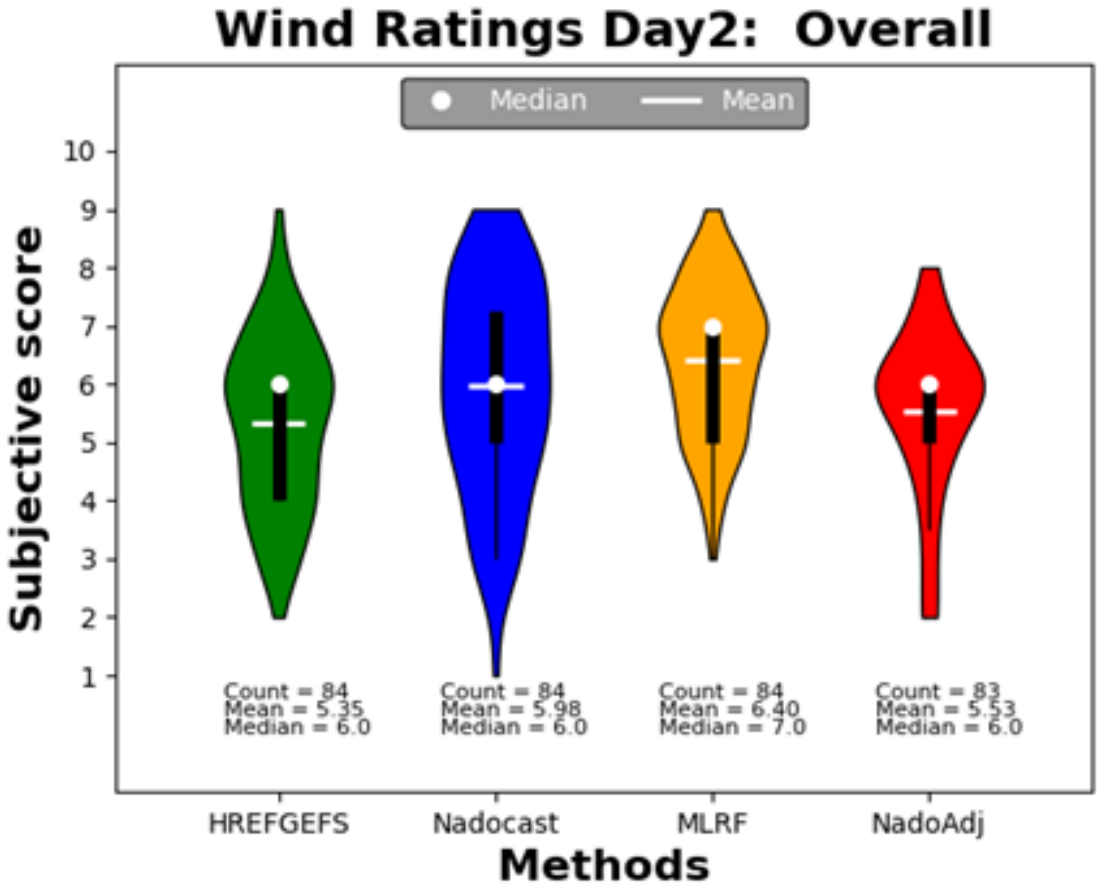

Figure 16. Distribution of subjective ratings for the C8 Day 2 Calibrated Wind Guidance evaluation. The white dots represent the mean ratings for each ensemble, and the white bars indicate the 95% confidence intervals for each mean.

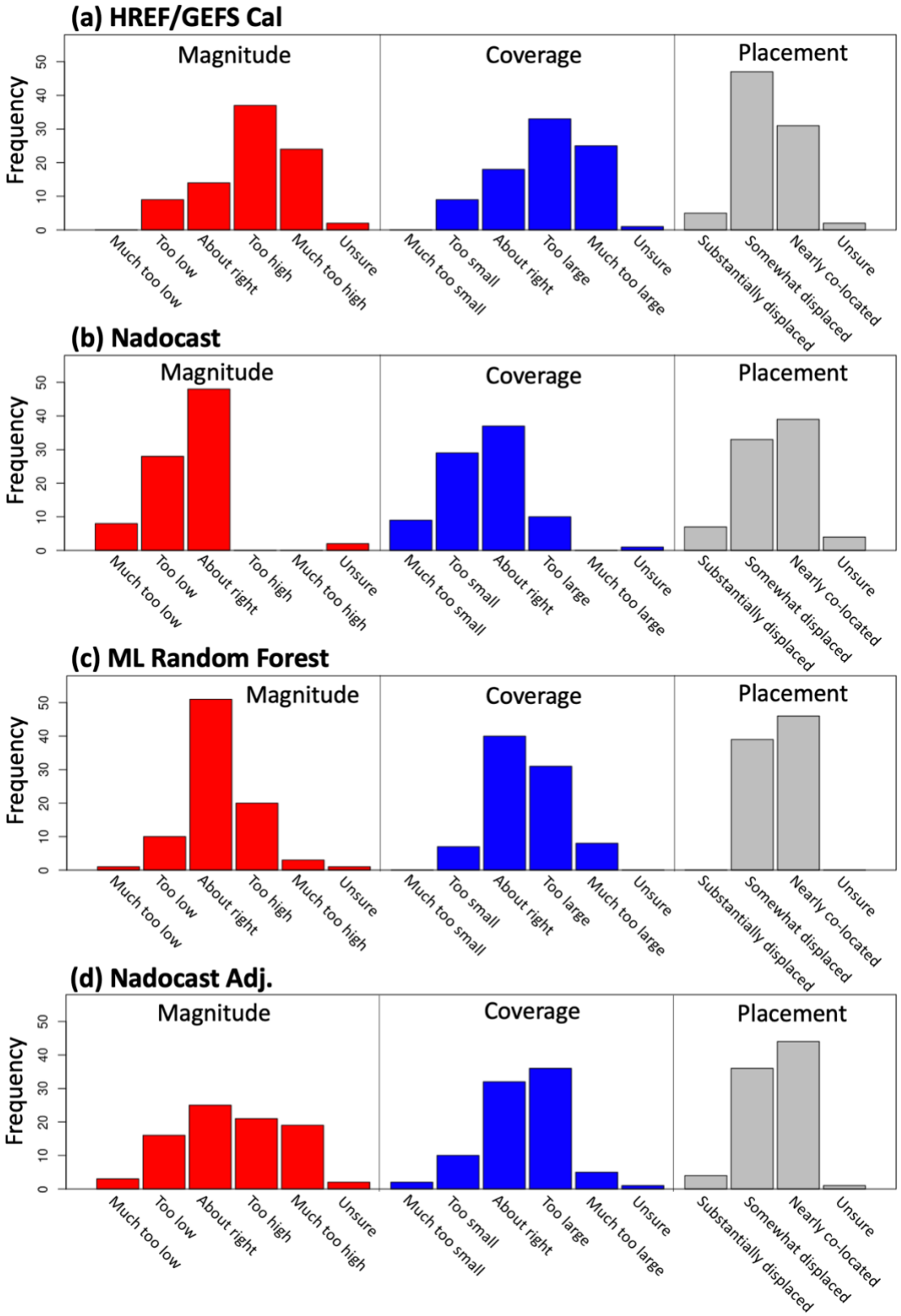

Figure 17. Response frequencies relating to magnitude, coverage, and placement of 1200 UTC HREF-based Day 2 wind probabilities derived from (a) HREF/GEFS Cal, (b) Nadocast, (c) ML Random Forest, and (d) NadoAdj.

To help identify dominant factors used by participants in assigning forecast skill scores, they were asked to evaluate each calibration method in specific consideration of placement, areal coverage, and magnitude of maximum forecast probability contours as compared to the practically perfect hindcast (PPH). In consideration of forecast magnitude, MLRF and Nadocast were both similarly evaluated as being "*about right*" (~50 frequency responses, Fig. 17). A majority of responses evaluated HREF/GEFS and Nadocast Adj. as being "*too high*" or "*much to high*". Similarly for coverage, MLRF and Nadocast were similarly evaluated as being "*about right*" (~40 frequency responses) while the majority of responses evaluated HREF/GEFS and Nadocast Adj. as being "*too large*" or "*much too large*", suggesting that coverage also had bearing on evaluation of overall skill. These results are consistent with the overall evaluated order of preferred methods (Fig. 16) for which MLRF and Nadocast were evaluated as the top performing methods, suggesting that both magnitude and coverage were dominant factors considered by SFE participants when rating overall method performance. Placement, however, for Nadocast, NadoAdj, and MLRF were evaluated with similar responses as being either "*nearly collocated*" or "*somewhat displaced*" (Fig. 17), but these methods had dissimilar overall skill scores (Fig. 16) suggesting that placement was not considered as much an evaluation factor compared to magnitude and coverage.

3.1.9 (C9) Day 1 12Z HREF Calibrated Wind Guidance

The same set of Day 2 calibrated wind guidance products was also evaluated for Day 1. The Day 1 overall ratings were quite similar to Day 2 with MLRF receiving the highest average subjective ratings, followed by Nadocast, NadoAdj, and HREF/GEFS Cal (Fig. 18).

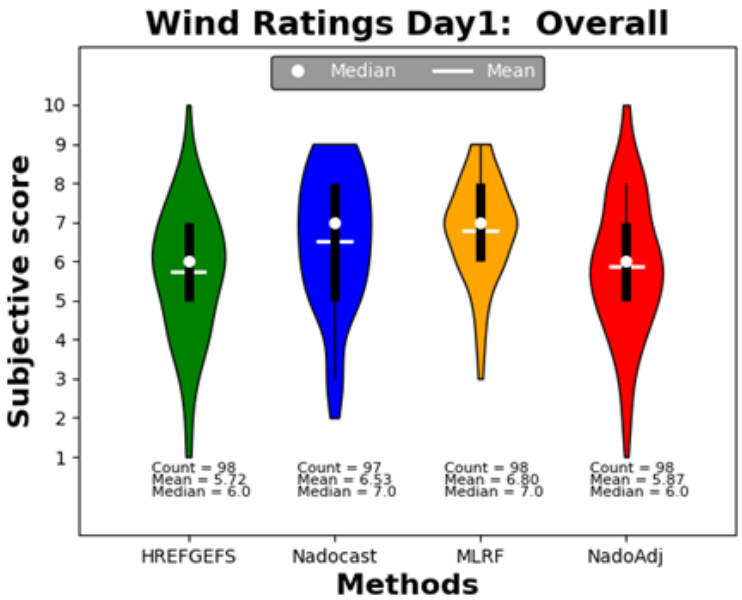

Figure 18. Distribution of subjective ratings for the C9 Day 1 Calibrated Wind Guidance evaluation. The white dots represent the mean ratings for each ensemble, and the white bars indicate the 95% confidence intervals for each mean.

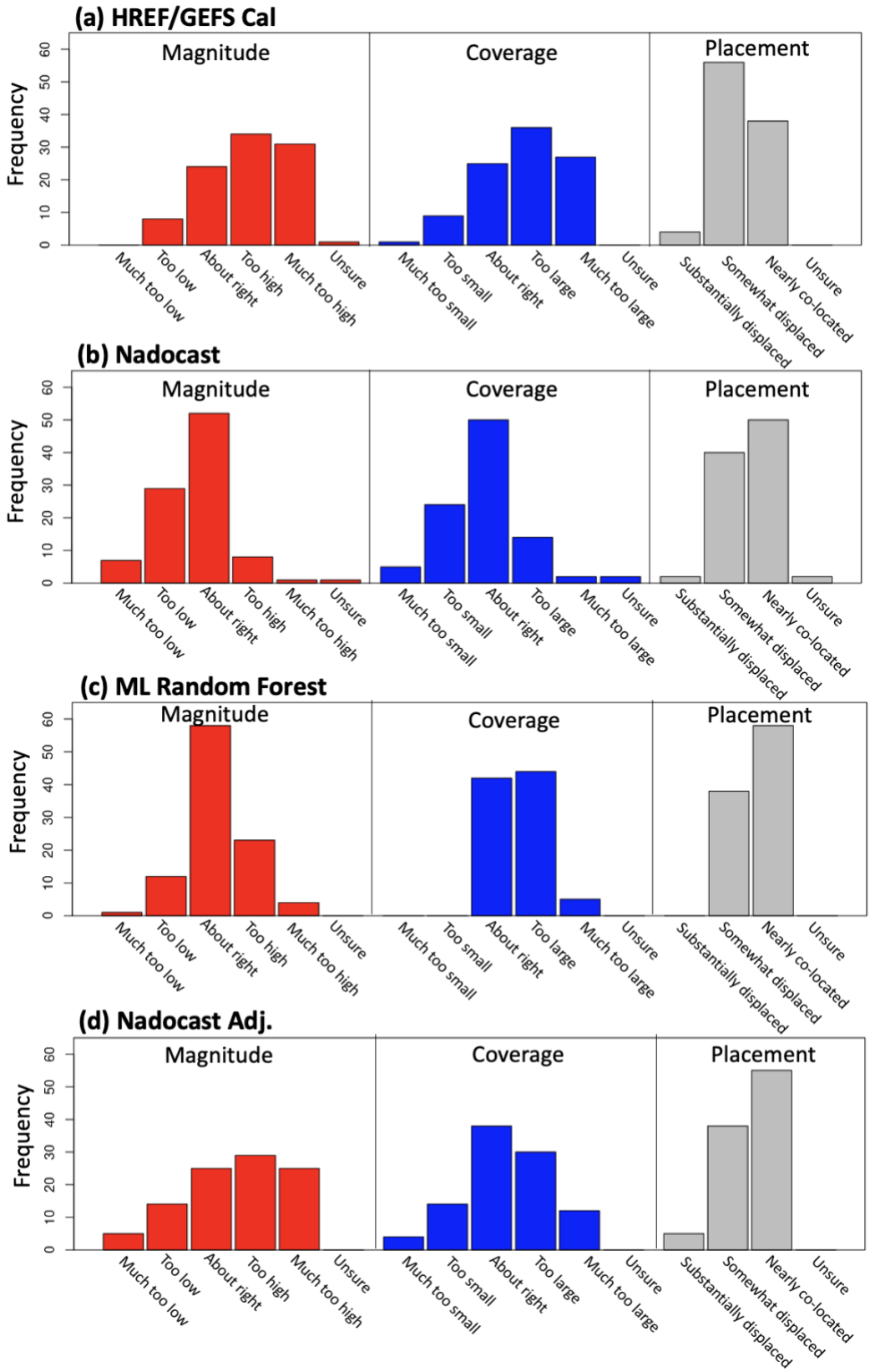

Figure 19. Response frequencies relating to magnitude, coverage, and placement of 1200 UTC HREF-based Day 1 wind probabilities derived from (a) HREF/GEFS Cal, (b) Nadocast, (c) ML Random Forest, and (d) NadoAdj.

Participant evaluation of forecast magnitude, coverage, and placement yielded similar results for Day 1 forecasts as for Day 2 forecasts as discussed above. Participants rated Nadocast and MLRF Day 1 forecast magnitude high with most frequent responses (greater than 50, Fig. 19) given as "*About right*", Nadocast and MLRF also had the highest overall skill score (mean value of greater than 6.5, Fig. 18), which suggests that forecast magnitude was an important factor in skill evaluation. HREF/GEFS Cal and NadoAdj usually had magnitudes that were rated less favorably, being either "*Too high*" or "*Much too high*", which is consistent with their overall lower skill scores (less than 5.9). For coverage, Nadocast forecasts were considered better than MLRF with most coverage forecasts evaluated as "*About right*", while MLRF coverage had the same frequency of responses being either "*About right*" or "*Too large*". These results are slightly different than the order of overall skill that favors MLRF, which are 6.8 and 6.5 for MLRF and Nadocast respectively. These results for Day 1 suggest that coverage is possibly a lesser factor than magnitude. Finally, for placement, Nadocast, MLRF, and NadoAdj had nearly the same frequency of responses for "*Nearly collocated*" and "*Somewhat collocated*", even though their overall skills differed, suggesting that placement was considered less influential in overall evaluation as were other factors.

3.1.10 (C10) 1630Z 4-h SPC Wind Timing Guidance (hourly 20-12Z)

As with the Tornado and Hail Timing Guidance (C3 & C7), the Wind Timing Guidance was highest rated for the version using Nadocast as the input, followed closely by the HREF/GEFS version. Again, both of these products were rated higher (both mean and median) than the HREF/SREF baseline version, but slightly lower mean ratings for wind as compared to hail (c.f. Figs. 15 & 20). The HREFCT version of the Wind Timing Guidance was once again the lowest-rated product among the suite, largely owing to a slower ramp up and ramp down in probabilities than the other versions. Generally speaking, this more gradual increase in severe wind probabilities preceding the peak of severe weather followed by a more gradual decrease in probabilities overnight did not match the timing of the severe wind threat as well as the other versions of the guidance.

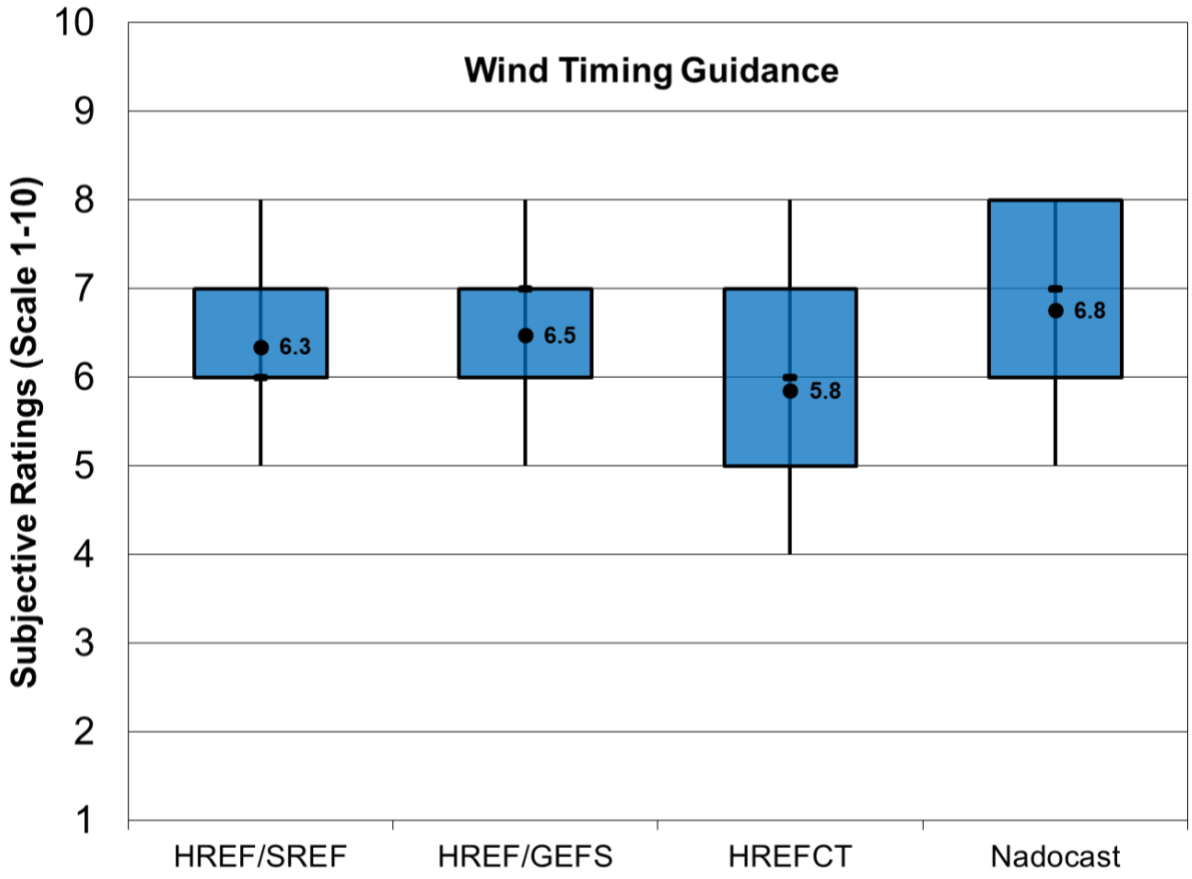

Figure 20. Distributions of subjective ratings (1-10; with 10 being best) by SFE participants of the 1630Z 4-h Wind Timing Guidance Products based on the HREF/SREF, HREF/GEFS, HREFCT, and Nadocast.

3.1.11 (C11) Medium Range 00Z GEFS Total Severe

Three algorithms for producing extended-range forecasts of total severe (tornado, wind, or hail) were assigned subjective ratings for Days 3-7. GEFS Reforecast ML is a random forest algorithm that uses environmental predictors from GEFS ensemble medians and is trained using 5-member GEFS reforecasts from Colorado State University. GEFS Reforecast ML has been tested in previous SFEs with very promising results; more info can be found in Hill et al. (2013). GEFS Operational ML from NSSL is similar to GEFS Reforecast ML, but it is trained using just over 2 years of the most recent GEFS Operational forecasts, which contain 31 members. Finally, GEFS Reforecast Cal from NSSL is a simple calibration method similar to what SPC has applied for many years to SREF, which uses environmental predictors from GEFS.

At each lead time, GEFS Operational ML was clearly the best performing algorithm, with statistically significant differences at all times. GEFS Reforecast ML was rated second, and GEFS Reforecast Cal was rated third (Fig. 21). An example case illustrating differences in GEFS Reforecast ML and GEFS Operational ML is shown in Figure 22. The GEFS Operational ML tends to generate higher probabilities at longer lead times that often correspond quite well with observed severe weather.

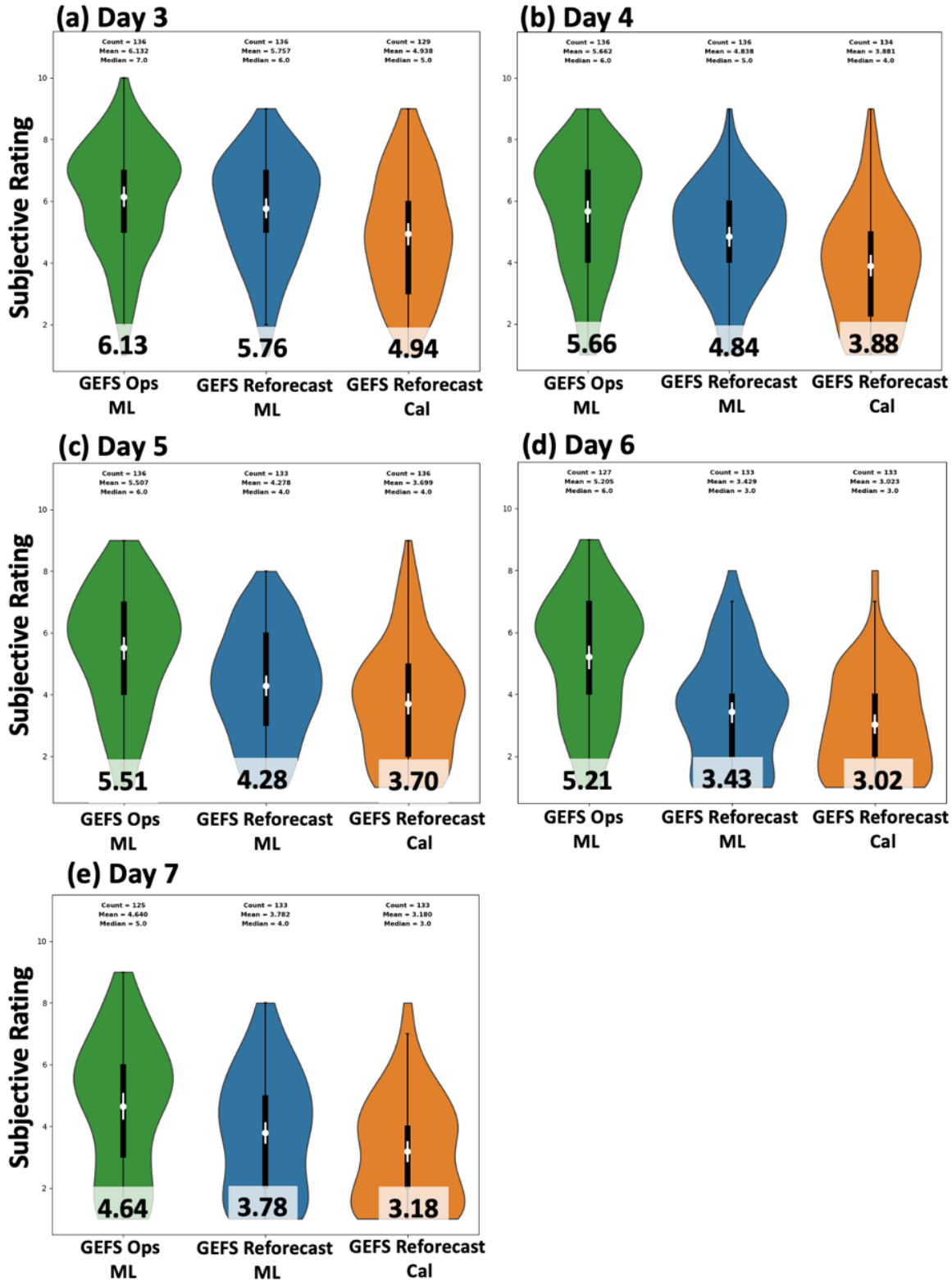

Figure 21. Distributions of subjective ratings for the C11 Medium-Range GEFS Total Severe evaluation for (a) Day 3, (b) Day 4, (c) Day 5, (d) Day 6, and (e) Day 7. The white dots represent the mean ratings for each ensemble, and the white bars indicate the 95% confidence intervals for each mean. The mean ratings are also shown at the bottom of each violin plot.

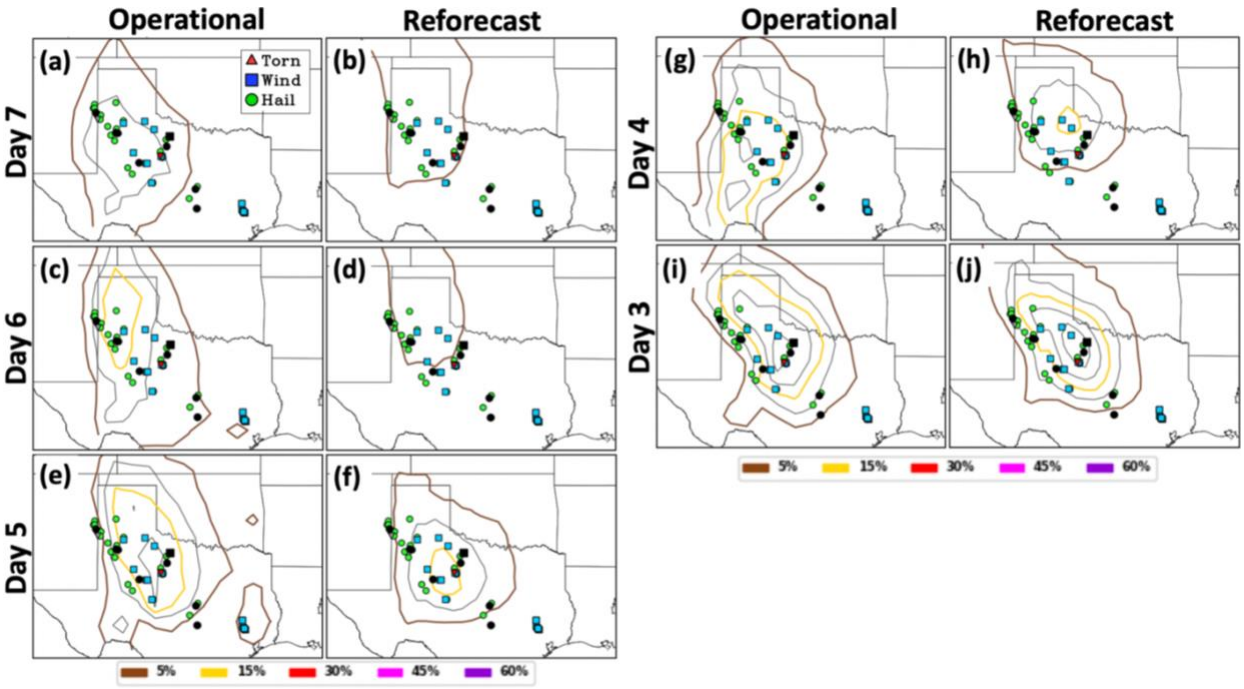

Figure 22. Severe weather probabilities at Day 7 lead time from (a) GEFS operational ML, & (b) GEFS reforecast ML. (c)-(d), (e)-(f), (g)-(h), and (i)-(j), same as (a)-(b), except for lead times of 6, 5, 4, & 3 days, respectively. Locations of observed storm reports are overlaid.

3.1.12 (C12) 00Z HRRR NCAR NN Tor/Hail/Wind Guidance

Probabilistic convective hazard guidance for tornado, hail, and severe wind forecasting is generated using a neural network (NN) algorithm. The initial version (v1) of this algorithm was trained with 42 diagnostics based on forecasted fields of the operational HRRR. The updated version (v2) includes 6 additional predictors that are related to convective mode. During the SFE, participants were asked to evaluate and compare the tornado, hail, and severe wind probabilistic guidance produced by both versions of the NN algorithm.

Figure 23 presents the subjective evaluation of v2 in the prediction of tornado, hail, and winds as well as the perceived improvement of v2 over v1 for all three convective hazards. Improvement is indicated by a mean value greater than zero. The hail guidance received a subjective improvement score of 0.01, thus there was no perceived change in performance over v1. Conversely, tornado and wind guidance showed a slight increase in skill for v2 with improvement scores of 0.15 and 0.11 respectively. These results are consistent with participant comments which generally considered v2 and v1 forecasts as similar. If some differences were noted, v2 was favored as the slightly better performer. This was mentioned as true in particular for tornado forecasts.

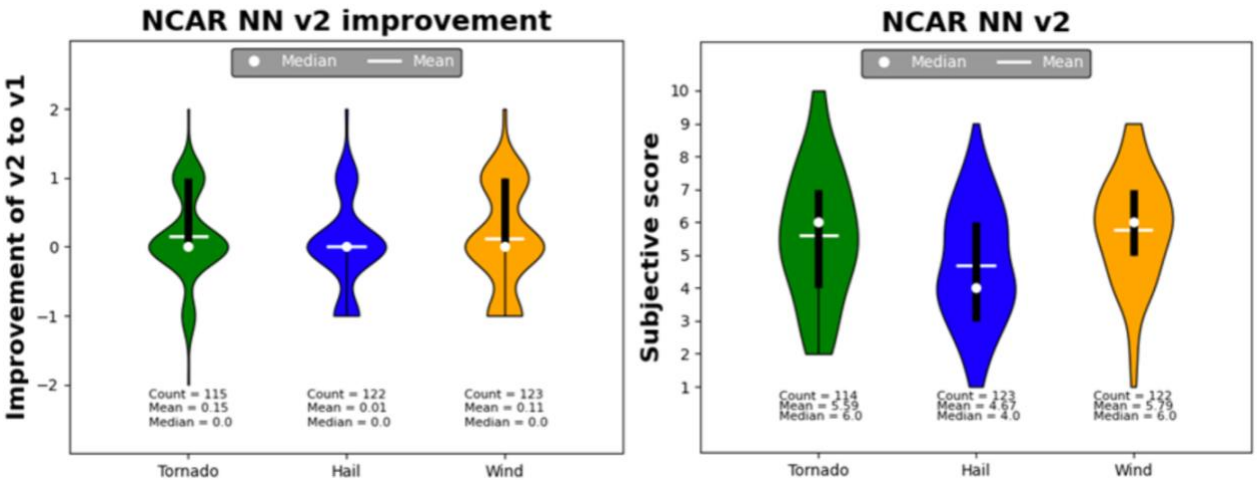

Figure 23. Distribution of ratings assigned to the NCAR NN algorithm v2 forecast guidance for tornadoes, hail, and severe wind (left plot). Perceived improvement of the NN algorithm v2 as compared to v1 (right plot; values -2 to 2 indicate respectively: much worse, worse, the same, better, much better).

3.2 Model Evaluation – (D)eterministic CAMs

3.2.1 (D1) CLUE: 00Z Day 1 Deterministic Flagships

This evaluation focused on comparing deterministic convection-allowing models which have been iterated on by their respective agencies and are relatively advanced in their development. Models included in this year's experiment consist of the GFDL FV3, NSSL MPAS RT, RRFS, and NASA GEOS FV3, each representing unique combinations of dynamical cores, data assimilation strategies, and physics parameterizations. The operational HRRRv4 was also included as a point of comparison for the other models. Only the 0000 UTC model initializations were assessed in this evaluation, and participants were asked to only look at forecast hours 12 - 36 when completing their surveys. This limited the evaluation to the Day 1 (1200 – 1200 UTC) time period. All models were evaluated blindly such that participants were not able to see which model produced which forecast. Additionally, the order of each model was randomized daily so that participants could not anticipate a model being in the same panel day-to-day. Models were unblinded following discussion of the results and after all surveys were submitted.

Participants compared the reflectivity and UH fields from each configuration, along with one environmental variable randomly selected from 2-m temperature, 2-m dewpoint, or surface-based convective available potential energy (SBCAPE). All participants then assessed and compared the 6-h quantitative precipitation forecast (QPF) produced by each model. Each model and field were independently rated on a scale of 1 (Very Poor) to 10 (Very Good), and participants had the option to provide additional insights via an open response box following each survey question.

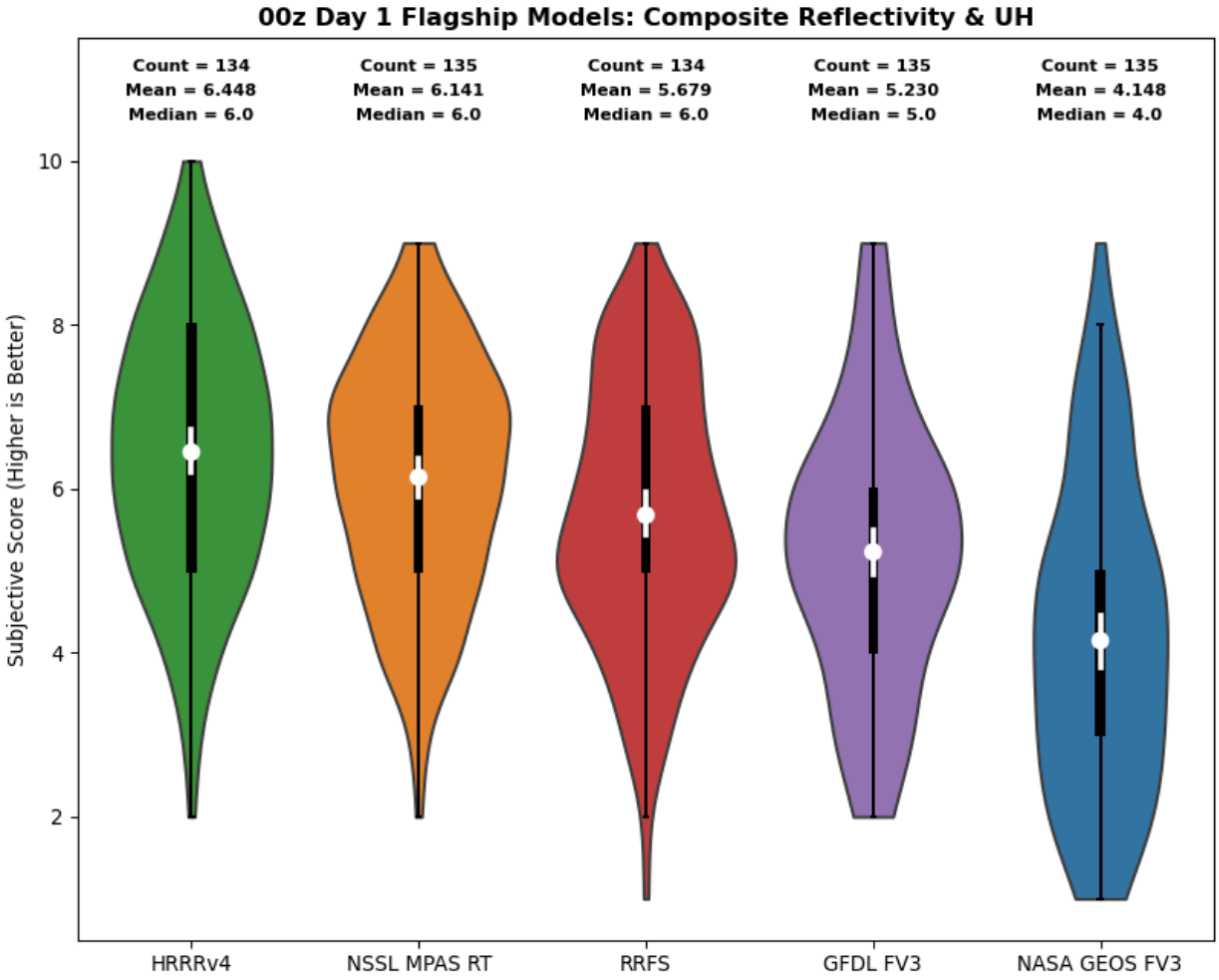

Figure 24. Distribution of subjective scores received by each deterministic flagship model at Day 1 lead times during the 5-week experiment. The white dots represent the mean scores for each ensemble, and the white bars indicate the 95% confidence intervals for each mean.

The HRRRv4 received the highest ratings on average when evaluating the structure, evolution, location, and timing of simulated storm reflectivity and UH at Day 1 lead times (Fig. 24), with a mean rating of 6.448 followed by the NSSL MPAS RT at 6.141. The RRFS received a mean rating of 5.679, the GFDL FV3 was given a 5.230, and the NASA GEOS FV3 saw the lowest mean rating of 4.148. The HRRRv4 mean rating was found to be significantly higher (at the 95% confidence level) than that of the RRFS, GFDL FV3, and NASA GEOS FV3, but was not significantly different from the NSSL MPAS RT. Conversely, the mean rating of the NASA GEOS FV3 was significantly lower than all other model configurations at the 95% confidence level. The HRRRv4 was the only model to receive a score of 10 at some point during the experiment, while the other models had a maximum rating of 9. The HRRRv4, NSSL MPAS RT and GFDL FV3 had minimum scores of 2, while the RRFS and NASA GEOS FV3 each received ratings of 1. When asked what characteristics of the simulated reflectivity and UH forecasts were most important to the participants when ranking the models, participants highlighted the timing and location of convective initiation, storm mode and evolution, and realistic storm

structure as their main points of focus. Simulated storm coverage and intensity were also discussed as influential factors.

Objective neighborhood statistics computed over the SFE 2023 domains mirrored the average subjective ratings very closely, as illustrated by the performance diagram in Figure 25. In this plot, probability of detection (POD) is plotted against the success ratio (SR), and the Critical Success Index (CSI) increases towards the upper right part of the plot. All five flagship models have similar SRs, but PODs are much higher in the HRRR and MPAS runs, which are followed by RRFS and then GFDL FV3 and NASA GEOS FV3. This results in CSIs that follow the same relative ranking.

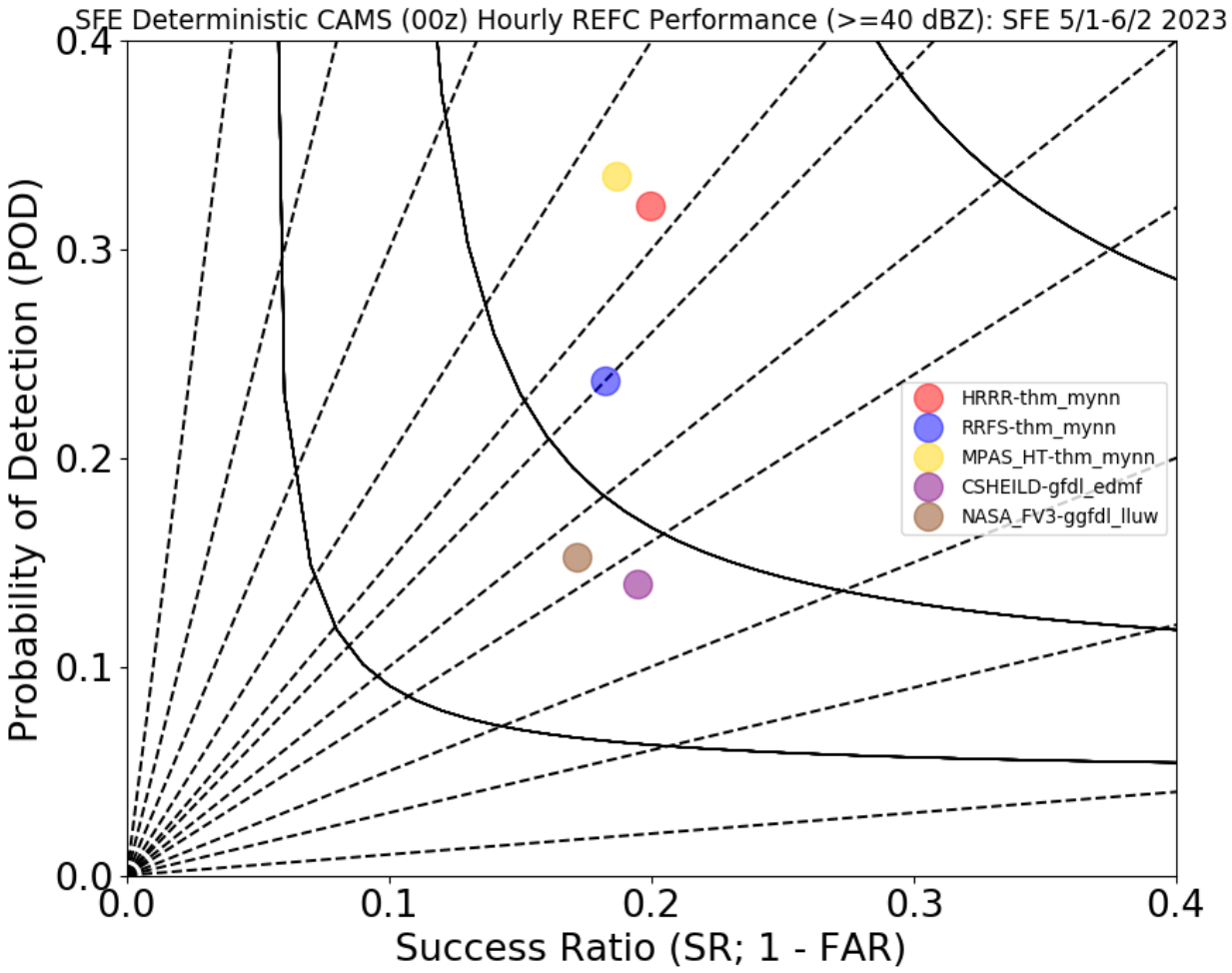

Figure 25. Performance diagram for hourly forecasts of simulated composite reflectivity ≥ 40 dBZ within 40-km neighborhoods computed over SFE 2023 domains during the Day 1 forecast period (i.e., f12-36).

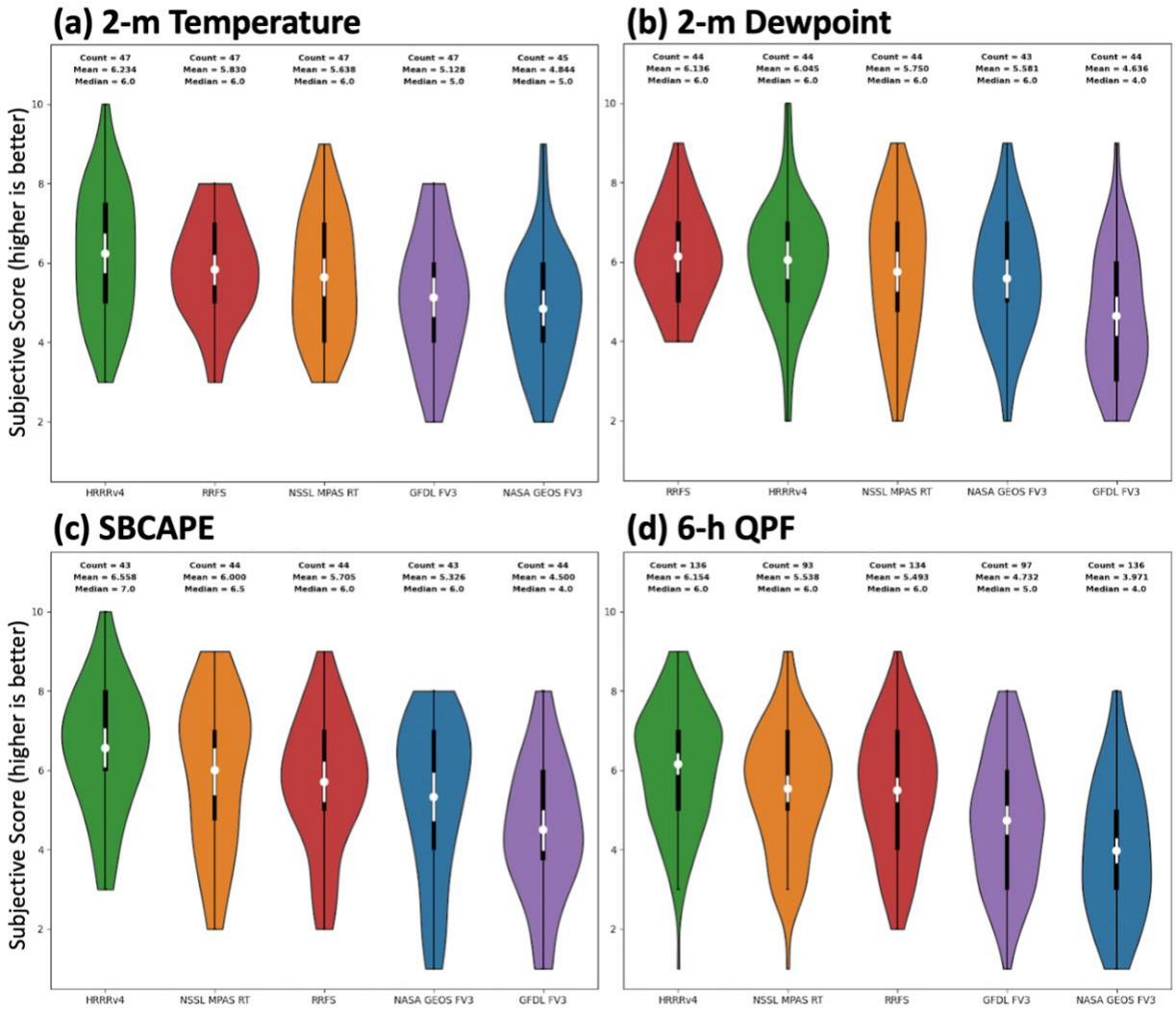

Figure 26. Response distributions for (a) 2-m temperature, (b) 2-m dewpoint, (c) SBCAPE, and (d) 6-h QPF at Day 1 lead times. The white dots represent the mean scores for each ensemble, and the white bars indicate the 95% confidence intervals for each mean.

Participants rated the models much closer on average when assessing the three environmental fields and QPF (Fig. 26). The HRRRv4 again received the highest mean score for 2-m temperature, SBCAPE, and 6-h QPF, but the RRFS saw a higher mean rating for 2-m dewpoint. In general, the HRRRv4, RRFS, and NSSL MPAS RT received very similar ratings in all four fields, and any differences were not significant at the 95% confidence level. Conversely, the GFDL FV3 and NASA GEOS FV3 models consistently received the lowest mean ratings in each comparison, and these scores were found to be significant when compared to the highest rated models. The HRRRv4 was the only model to receive a rating of 10 at some point during the 5-week experiment in the 2-m temperature, 2-m dewpoint, and SBCAPE fields. Participants cited the magnitude and location of cold pools and mesoscale boundaries as the most influential factors contributing to their ratings for each environmental field, but systematic biases in the models were also noted as points of concern. For example, respondents frequently observed a strong dry bias in the GFDL FV3's 2-m dewpoint which adversely affected its

rating for that field. Participants primarily considered the coverage and magnitude of estimated rainfall when assessing each model's 6-h QPF, and many respondents commented that they did not place much emphasis on location error. Some concerns were raised about the difficulty of providing a single rating for multiple 6-h time frames, and participants suggested that 24-h QPF may be easier to evaluate in future SFEs.

3.2.2 (D2) CLUE: 00Z Day 2 Deterministic Flagships

This evaluation was similar to the previous, except participants were asked to evaluate the models at Day 2 lead times (forecast hours 37-60). As before, only the 0000 UTC model initializations were assessed in this evaluation, and all models were blinded until after the surveys were submitted. Participants were again asked to evaluate the reflectivity and UH forecasts from each model, the same randomly selected environmental field from the Day 1 evaluation, and the 6-h QPF on a scale of 1 (Very Poor) to 10 (Very Good). Because the HRRRv4 is only available through forecast hour 48, it was necessary to replace it with another operational model for the Day 2 evaluation. As such, the NAM CONUS Nest was chosen to serve as an operational point of comparison for the other models. Participants were not informed of this change from the Day 1 evaluation until after all surveys had been submitted.

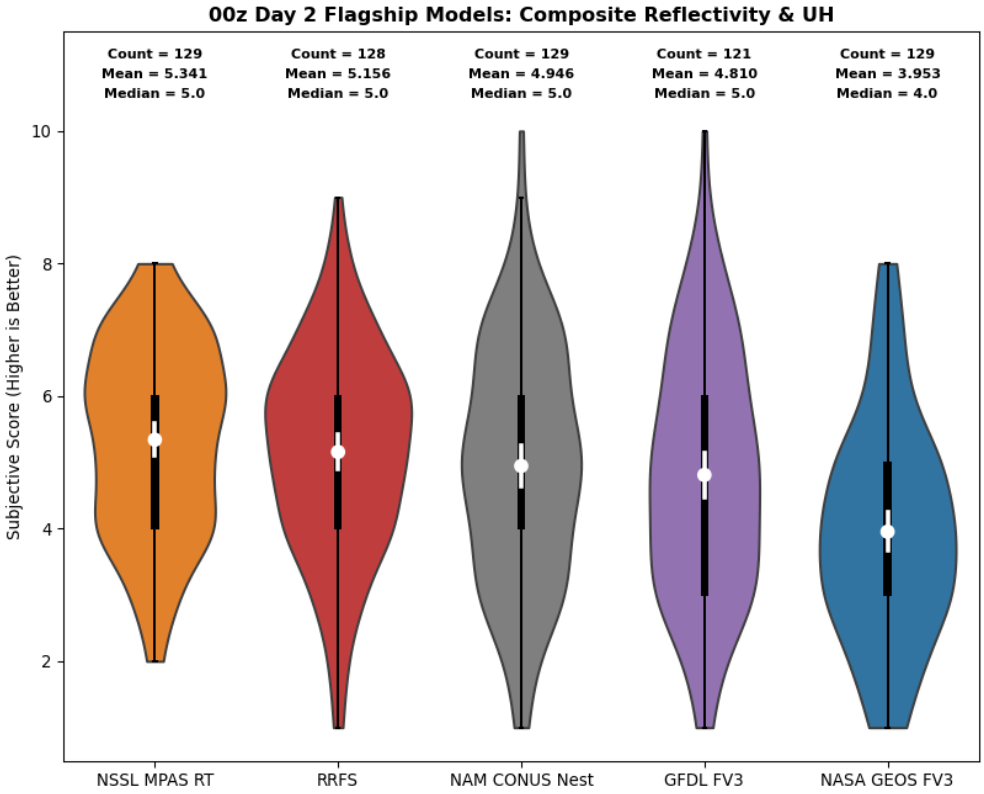

Figure 27. Distribution of subjective scores received by each deterministic flagship model at Day 2 lead times during the 5-week experiment. The white dots represent the mean scores for each ensemble, and the white bars indicate the 95% confidence intervals for each mean.

In the absence of the HRRRv4, the NSSL MPAS RT received the highest mean rating when assessing the structure, evolution, location, and timing of simulated storm reflectivity and UH (Fig. 27). The NSSL MPAS RT had a mean score of 5.341, followed by the RRFS (5.156), NAM CONUS Nest (4.946), GFDL FV3 (4.810), and NASA GEOS FV3 (3.953). The NASA GEOS FV3 was again found to have a significantly lower (at the 95% confidence level) mean rating than all other model configurations, but the NSSL MPAS RT, RRFS, NAM CONUS Nest, and GFDL FV3 received statistically similar ratings on average. The NAM CONUS Nest and GFDL FV3 were the only models to receive a score of 10 during the experiment, while the RRFS had a maximum rating of 9. The NSSL MPAS RT and NASA GEOS FV3 both had the lowest maximum rating of 8 during the Day 2 evaluations. Conversely, the NSSL MPAS RT was the only model to receive a minimum rating of 2, while all other configurations saw a rating of 1 at some point during the experiment.

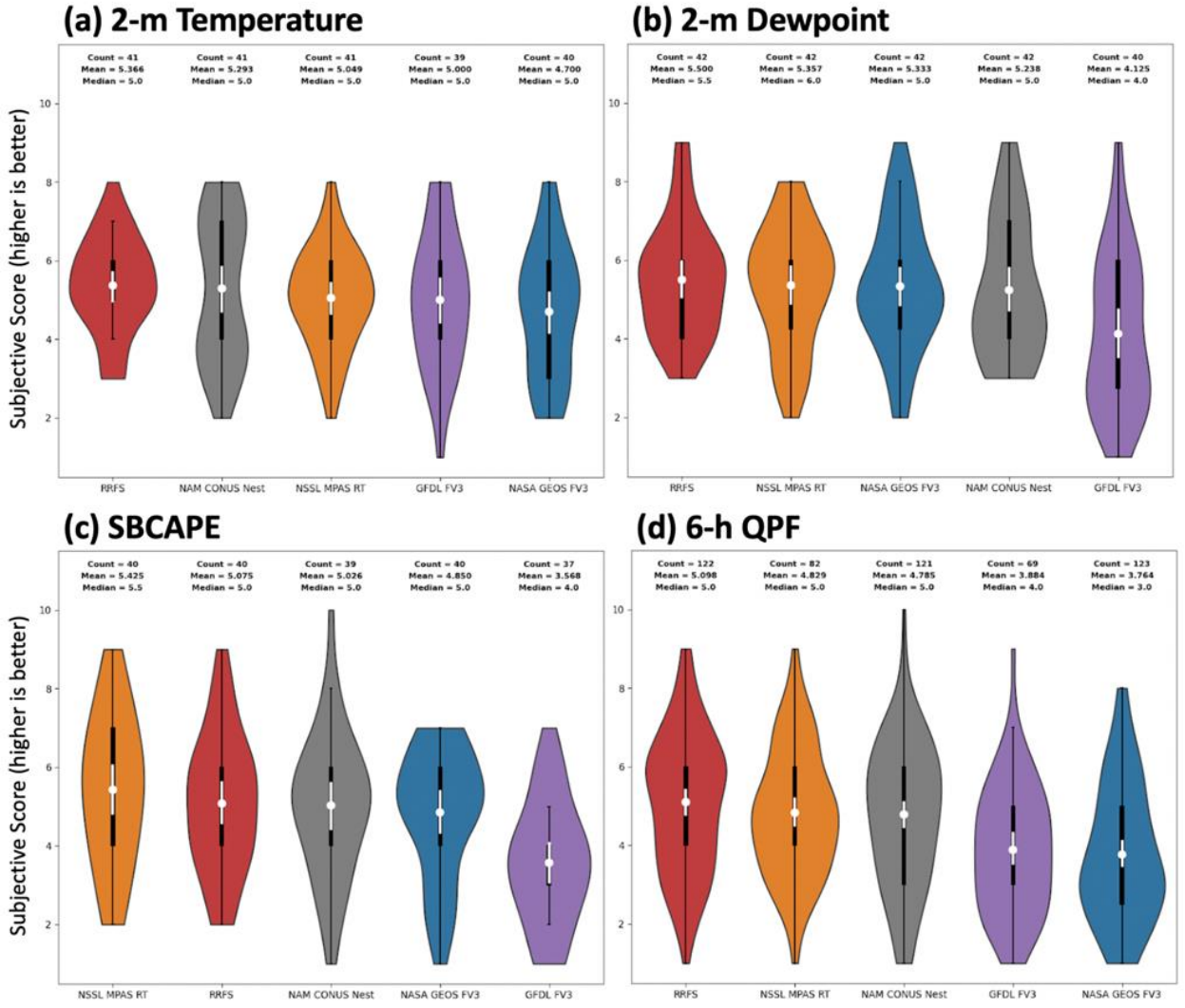

Figure 28. Response distributions for (a) 2-m temperature, (b) 2-m dewpoint, (c) SBCAPE, and (d) 6-h QPF at Day 2 lead times. The white dots represent the mean scores for each ensemble, and the white bars indicate the 95% confidence intervals for each mean.

Participants gave the RRFS the highest mean rating for 2-m temperature, 2-m dewpoint, and 6-h QPF at Day 2 lead times, while the NSSL MPAS RT was rated the best for SBCAPE (Fig. 28). In general, all five models received very similar ratings in each of the environment evaluations, though there were a few notable exceptions. The GFDL FV3 was found to have a significantly lower mean score than the best-rated models in the 2-m dewpoint and SBCAPE evaluations. Similarly, the GFDL FV3 and NASA GEOS FV3 were found to have statistically lower mean ratings than the other three configurations when assessing the 6-h QPF. These results are consistent with those shown in the Day 1 evaluations.

3.2.3 (D3) CLUE: RRFS vs. HRRR

One of the critical evaluations during the 2023 HWT SFE was comparing the deterministic RRFS control member to the operational HRRR. This was done for both the 00Z and 12Z runs to assess the readiness of the RRFS to replace the HRRR for operational convective forecasting applications on Day 1. Participants were asked to examine storm-attribute fields (e.g., Fig. 29), including composite reflectivity and UH, updraft speed, 10-m wind speed, and 6-h QPF, and provide a single rating for the convective day (i.e., f12-f36 for the 00Z runs, and f01-f24 for the 12Z runs). For this evaluation, a five-point Likert scale was used to rate the RRFS as much worse, slightly worse, about the same, slightly better, or much better than the HRRR for each cycle and each field. For example, the 00Z RRFS forecast was generally rated slightly worse than the 00Z HRRR forecast for the derecho-producing MCS in Kansas on 9 May 2023 while the 12Z RRFS forecast was rated slightly better than the 12Z HRRR forecast (Fig. 29).

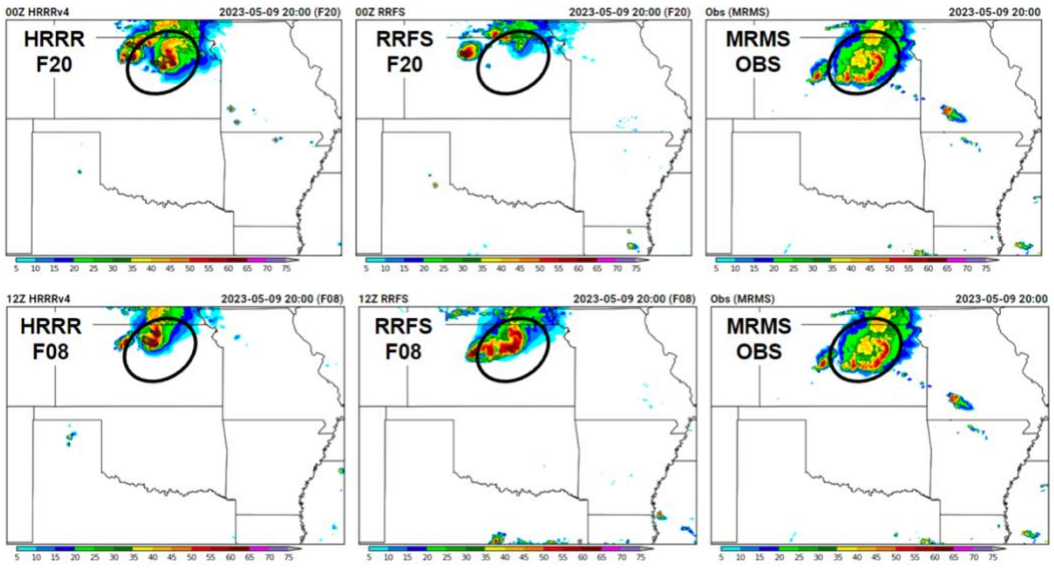

Figure 29. Example of the 2023 HWT SFE model comparison page for the RRFS vs. HRRR valid at 20Z on 9 May 2023. The composite reflectivity forecasts are shown for the 00Z HRRR (upper-left panel), the 00Z RRFS control (upper-middle panel), the 12Z HRRR (lower-left panel), and the 12Z RRFS control (lower-middle panel). The observed MRMS composite reflectivity is shown in both the upper-right and lower-right panels.

For the 00Z storm-attribute fields, the HRRR was rated slightly better for simulated reflectivity/UH, updraft speed, and QPF than the RRFS by the SFE participants (Fig. 30). Meanwhile, the RRFS was very slightly favored for severe convective 10-m winds to occasionally have a better signal for strong winds in the vicinity of local storm reports of damaging winds. The most common comments from SFE participants included that the RRFS developed storms that were too intense and too numerous/widespread compared to observations, which distracted from primary threat regions and/or disrupted the downstream environment for convection.

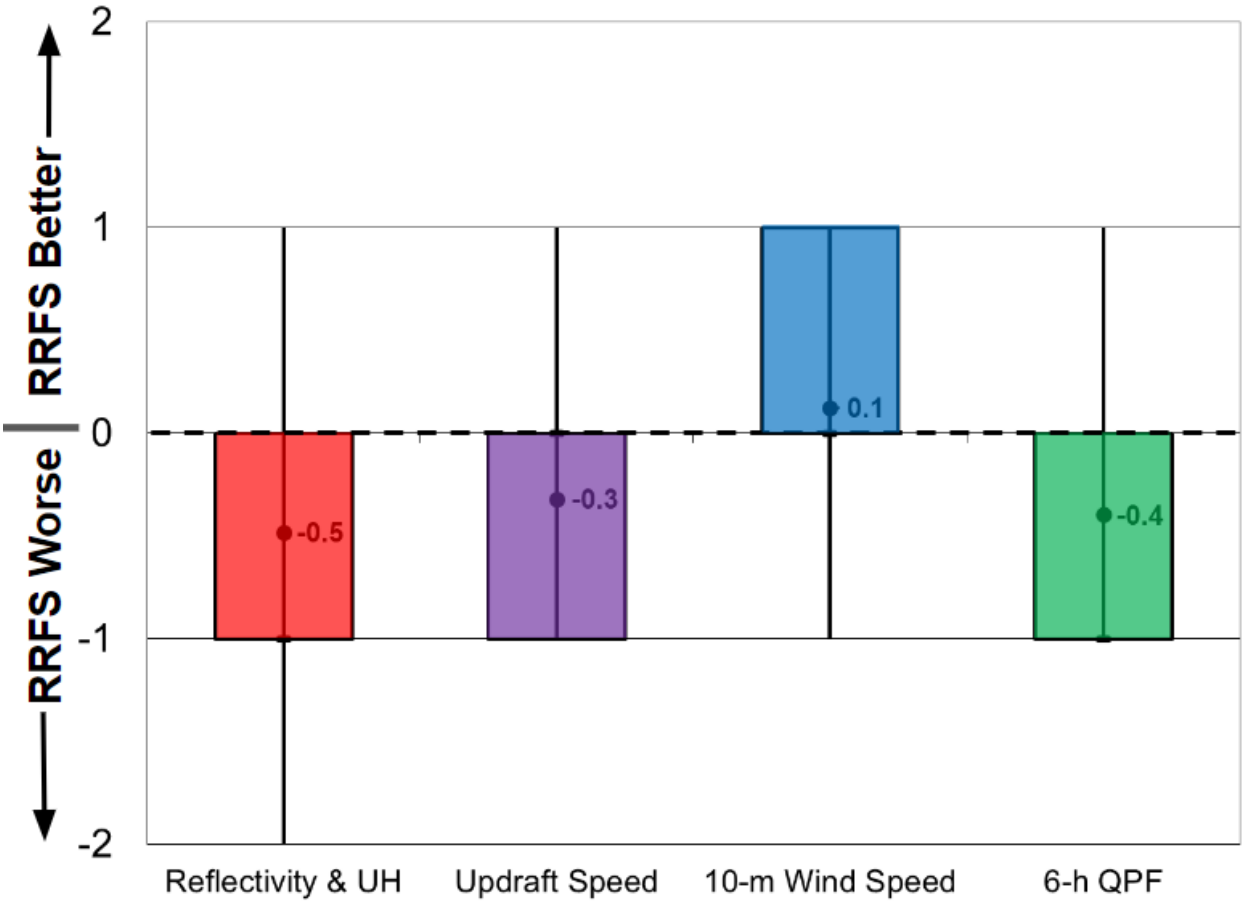

Figure 30. Distributions of subjective ratings (-2 to +2) by SFE participants of the 00Z RRFS compared to the HRRR for composite reflectivity and UH (red), updraft speed (purple), 10-m wind speed (blue), and 6-h QPF (green). The ratings represent the RRFS compared to the HRRR -2: Much Worse; -1: Slightly Worse; 0 – About the Same; +1 Slightly Better; +2: Much Better.

Regarding the ratings of the 00Z environment fields, SFE participants gave a slight edge to the HRRR for SBCAPE, 2-m temperature, and 2-m dewpoint over the RRFS control (Fig. 31). The SBCAPE forecasts were the most common environmental field to be favored for the HRRR with a median rating of the RRFS being slightly worse. Overall, the RRFS tends to have a low bias (frequency and magnitude) in forecasts of CAPE.

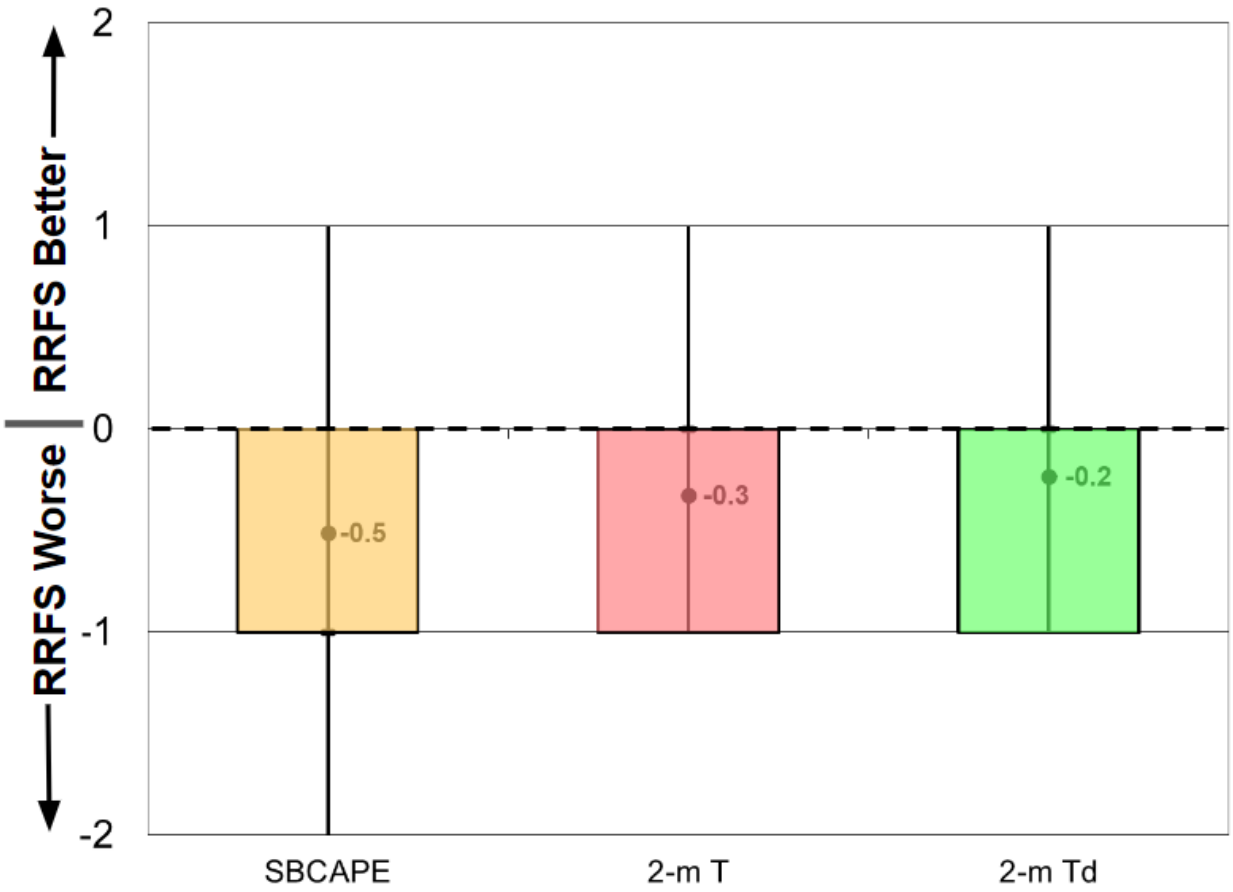

Figure 31. Same as Fig. 30, except for environmental fields of SBCAPE (yellow), 2-m temperature (pink), and 2-m dewpoint (light green).

Surprisingly, evaluations of the 12Z runs revealed different results. The subjective ratings of storm-attribute fields (Fig. 32) indicate that the performance of the 12Z RRFS control member was much closer to that of the 12Z HRRR compared to the respective 00Z runs. This subjective difference in performance based on model initialization is supported by objective metrics as well. A performance diagram for convective storms (i.e., ≥40 dBZ composite reflectivity) reveals that the 12Z RRFS control and 12Z HRRR have very similar performance characteristics while the 00Z HRRR has a clear advantage in POD and CSI over the 00Z RRFS (Fig. 33). While it is difficult to identify the primary cause for this performance dependence on initialization time, it is speculated that the data assimilation issues noted in the next section (D4) have a stronger impact on the 00Z runs when the coverage of deep convection is greater at initialization as compared to the 12Z runs.

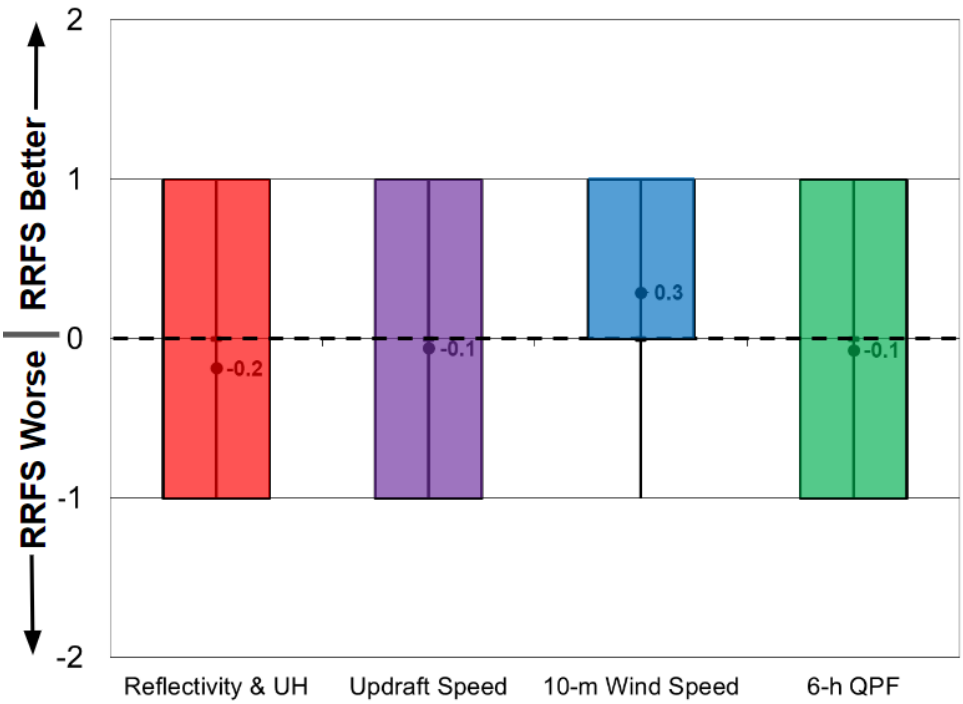

Figure 32. Same as Fig. 30, except for comparing the 12Z runs of the RRFS to the HRRR.

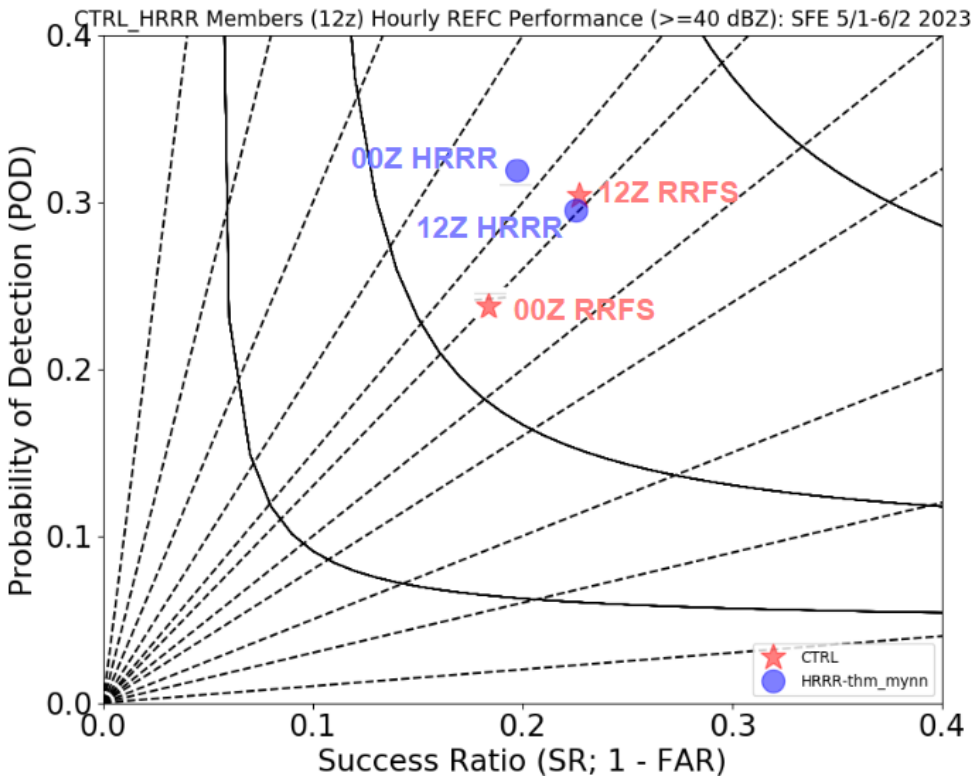

Figure 33. Performance diagram for hourly composite reflectivity ≥40 dBZ covering the 24-h convective day (i.e., 12-12Z) over the five-week period of the HWT SFE. The 00Z and 12Z HRRR (blue circle) and RRFS (red star) performance characteristics are labeled on the diagram. The statistics are only calculated over the primary mesoscale domain used each day for evaluation activities.

3.2.4 (D4) CLUE: RRFS vs. HRRR DA

HRRRv4 and RRFS were examined during the first 12 hours of their forecasts for 2100 and 0000 UTC initialization to assess forecast skill at times during which the data assimilation has a large impact. The times 2200, 0100, and 0600 UTC were considered. Participants first compared forecast UH and simulated composite reflectivity to observed reflectivity and were asked, "*Please rate on a scale of 1 (very poor) to 10 (very good) how well each model depicts storms that were ongoing at [22z, 01z, or 06z]. Consider aspects like storm retention, strength, and location in your answer*." An example graphic from the model comparison webpage is shown in Figure 34.

For all lead times at both initializations examined, the mean subjective ratings for HRRRv4 were higher than RRFS (Fig. 35). For the 2100 UTC initialization, differences were statistically significant at 2200 and 0600 UTC, but not 0100 UTC. For the 0000 UTC initialization, differences were statistically significant at 0100 UTC, but not 0600 UTC. The most common theme from the survey comments was that the simulated reflectivity in RRFS was too high, and that RRFS often had spurious storms. This is clearly reflected in the bottom-middle panel of Figure 34 where the line of storms in western Texas is clearly too intense, and to the east and south of this line there are spurious storms. Some representative comments included, "*RRFS initializations were hot*", "*Both models were a bit more cellular than the obs, especially the RRFS. The RRFS also had way more storms than the HRRR and obs had*.", "*Overall, the two models produced similar depictions of convective evolution, although the RRFS had a higher reflectivity bias, including several areas of spurious storms.*", "*RRFS echoes are too intense, especially compared to the HRRRv4 and the observations*", "*Sig. difference between RRFS and HRRRv4 - RRFS misleads with spurious convection in the south (TN et al.); difference between the two models quite pronounced. Interestingly, difference between the two DA cycles less pronounced*", "*RRFS has stronger and more storms at each time than HRRR, and consistently has erroneous convection in TN and stronger convection than realized. HRRR does much better at the intensity and location of storms, particularly one hour after initialization times*", "*RRFS starts hot at initialization and seems to carry some of this 'excess' convection over into the start of the forecast. HRRR 21z seemed to do well with overall evolution...00z HRRR not quite as good*".

Next, participants were asked to evaluate one of three randomly selected environment fields, and assign a subjective rating on a scale of 1-10 reflecting the overall skill during the first 12 hours of the forecast. Again, for each field and at both initialization times, the mean subjective ratings in HRRRv4 were higher than RRFS. The largest differences were with surface-based CAPE. The differences for 2-m dewpoint and surface-based CAPE were statistically significant, but the differences for 2-m temperature were not (Fig. 36).

For 2-m temperature, comments frequently noted that HRRRv4 was too warm while RRFS was too cool. It was also commented on several times that convectively generated cold pools in RRFS were too strong. For example, one participant wrote, "*HRRR appeared to be warmer than RRFS, which tends to better agree with obs later in the period. RRFS cold pools appear to be too cold*". For 2-m dewpoint, comments most

frequently mentioned that RRFS had a moist bias. Finally, for surface-based CAPE the common themes were that RRFS was too low and that the HRRRv4 magnitudes and spatial patterns were better than RRFS.

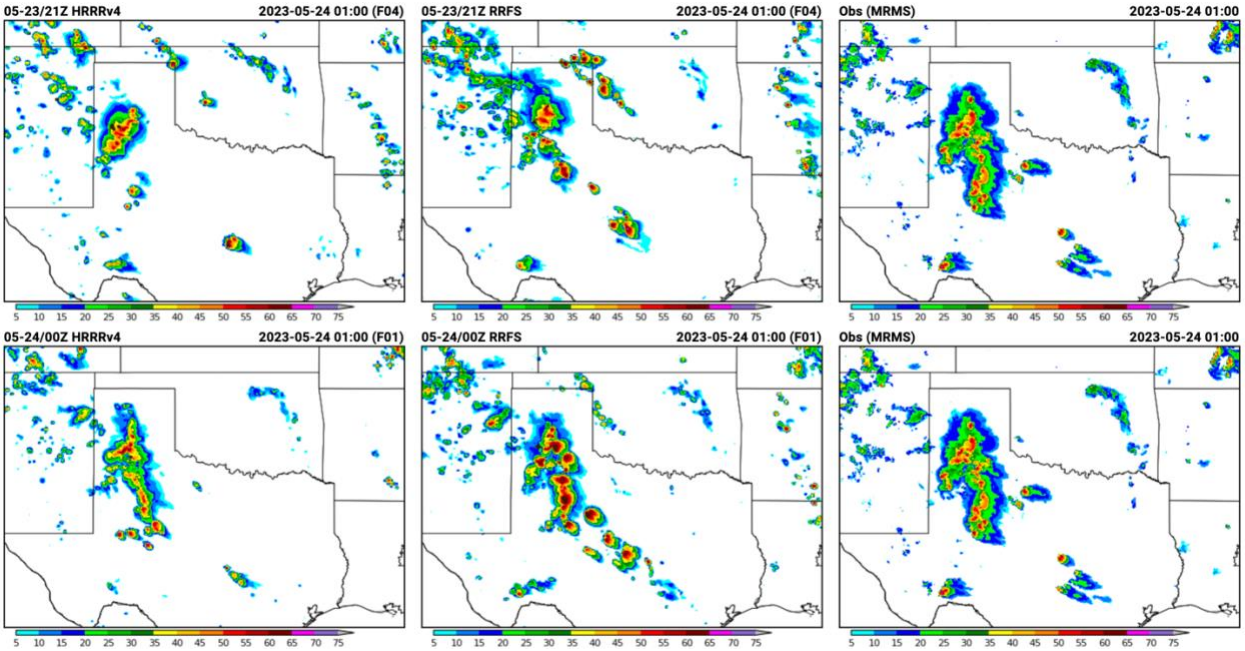

Figure 34. Example of multi-panel comparison webpage for the D4 RRFS vs. HRRR DA evaluation. The top row displays simulated composite reflectivity from 2100 UTC initializations of HRRRv4 (left) and RRFS (middle) valid at 0100 UTC compared to MRMS observations (right). The bottom row displays the same as the top, except for 0000 UTC initializations.

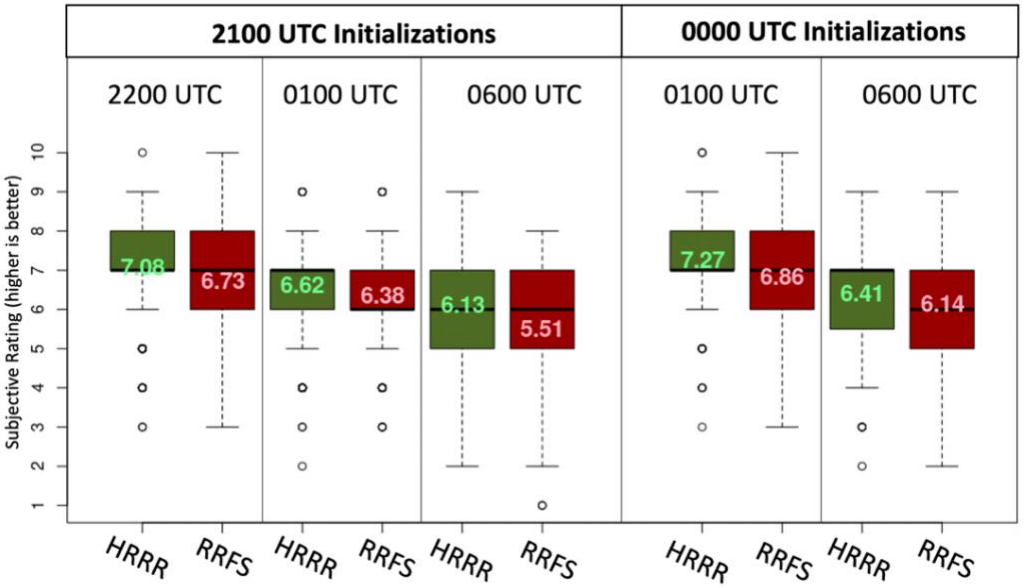

Figure 35. Boxplots of subjective rating distributions for reflectivity and UH forecasts from 2100 and 0000 UTC initializations of the HRRR (green) and RRFS (red) valid at 2200, 0100, and 0600 UTC. The mean value for each distribution is overlaid on the corresponding boxplots.

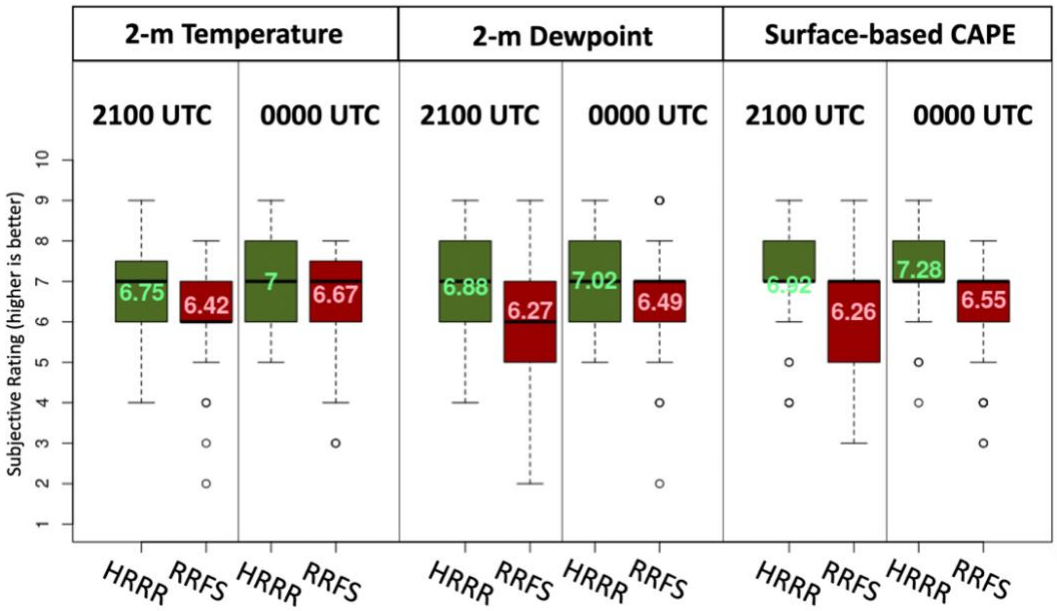

Figure 36. Boxplots of subjective rating distributions for 2-m temperature, 2-m dewpoint, and surface-based CAPE forecasts from 2100 and 0000 UTC initializations of the HRRR (green) and RRFS (red). The mean value for each distribution is overlaid on the corresponding boxplots.

3.2.5 (D5) CLUE: 00Z MPAS

Three configurations of MPAS run by NSSL were assigned subjective ratings. These configurations included: (1) NSSL MPAS HT, (2) NSSL MPAS HN, and (3) NSSL MPAS RT. "HT" refers to HRRR initialization with Thompson microphysics, "HN" is HRRR initialization with NSSL microphysics, and "RT" is RRFS initialization with Thompson microphysics. Participants were asked to focus on the entire 36 h forecast period and focus on how the models depicted the timing, location, and mode of thunderstorms within the domain and how those forecasts compared to observations. An example of the model comparison interface is shown in Figure 37.

Overall, the three configurations performed quite similarly and none of the differences between pairs of MPAS runs were statistically significant, although, MPAS HN had the highest mean subjective rating (Fig. 38). Some of the survey comments reflecting MPAS HN performance included, "*The HN performed better with placement, mode, and intensity later in the period*", "*The HN was almost always most faithful to the actual storm evolution. The HT often slightly overdid convection, while the RT appeared to underdo nocturnal convection*", "*HN preferred for organisation and structure. RT and HT both offered hot reflectivity, RT especially so, while both also offered excess structure and organization (although they had the right idea) - both holding on to storms too long into the post-00Z period*", and "*RT actually handled the previous overnight better in Arkansas. But for the main event in Kansas during the day these were all fairly impressive, but especially HT and HN which had a clear bowing structure. HN better depicted the narrow stratiform region with the primary bow, but both HT and HN looked very good*".

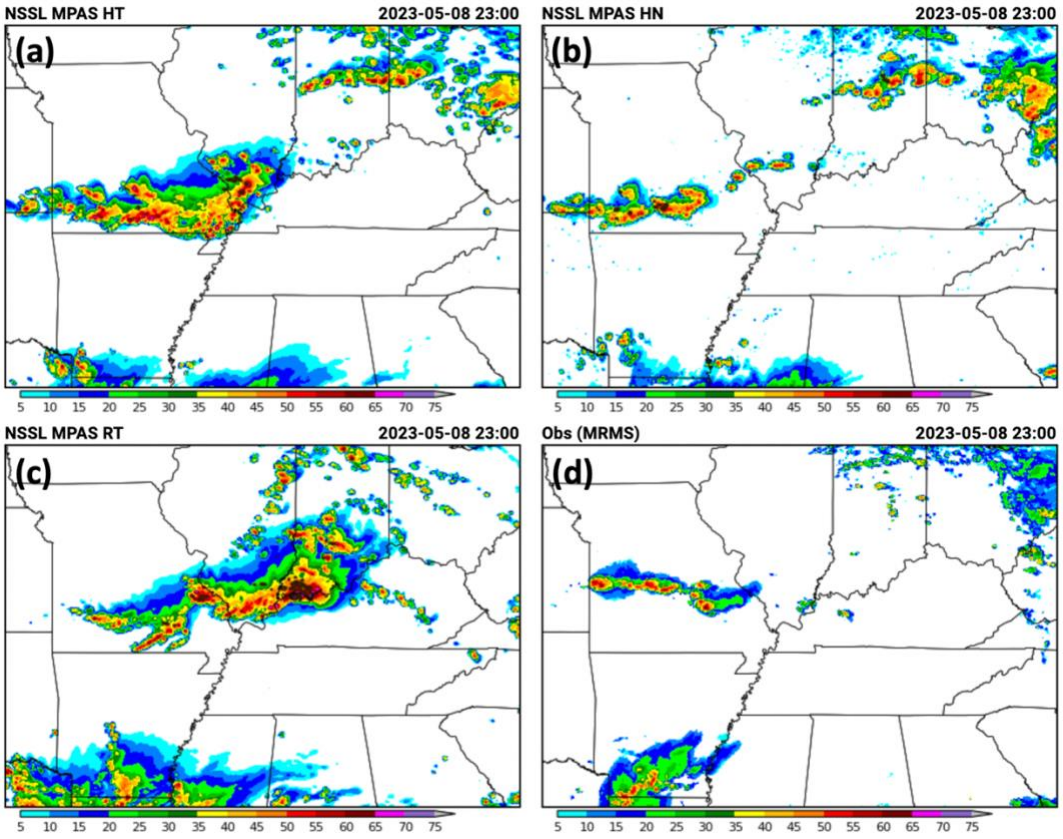

Figure 37. Example of multi-panel comparison webpage for the D5 00Z MPAS evaluation. The panels show simulated composite reflectivity from 0000 UTC MPAS initializations valid at 2300 UTC 8 May 2023 from (a) NSSL MPAS HT, (b) NSSL MPAS HN, (c) NSSL MPAS RT, and (d) the corresponding MRMS observations.

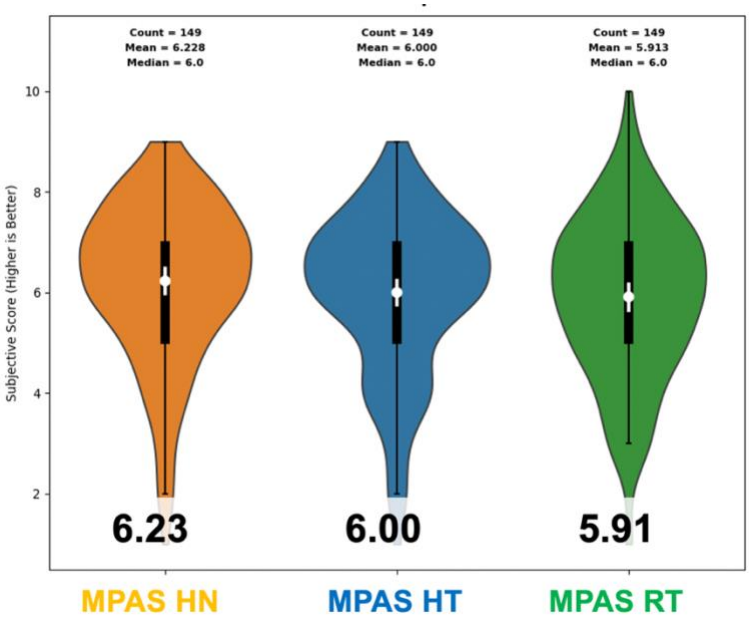

Figure 38. Response distributions shown with violin plots for the D5 evaluation of MPAS configurations. The white dots represent the mean scores for each ensemble, and the white bars indicate the 95% confidence intervals for each mean. The number at the bottom of each violin plot indicates the mean subjective rating.

3.2.6 (D6) CLUE: NSSL1 vs. HRRR

This evaluation focused on comparing the NSSL1 (1-km grid-spacing WRF model configuration) and HRRRv4. Particular attention was given to unique storm attribute fields such as 0-1 km AGL UH and 0-2 km AGL maximum wind. It is hypothesized that for these fields, the enhanced resolution of NSSL1 could provide improved guidance for hazards like tornadoes, whose parent mesocyclones and associated low-level rotation are better resolved using 1-km grid-spacing, and wind, which is better resolved at higher resolutions. Specifically, there were three survey questions: (1) "*Please rate on a scale of 1 (Very Poor) to 10 (Very Good) how well each model captured the convective evolution compared to observations. Consider factors such as the number of storms depicted, the structure and evolution of those storms, and the timing of convective initiation*", (2) "*Please rate on a scale of 1 (very poor) to 10 (very good) how well the 0-2 km UH field delineates the tornado threat in each model*", and (3) "*Please rate on a scale of 1 (very poor) to 10 (very good) how well the hourly maximum 10-m wind speed delineates the wind threat in each model*".

For the convective evolution and 0-2 km AGL UH, differences in mean subjective ratings were quite similar, and although HRRRv4 was slightly higher than NSSL1, the differences were not significant (Fig. 39). For 0-2 km AGL UH, there were not many tornado events during SFE 2023, so the similar ratings likely reflect many null cases. The results for maximum 10-m wind were different, though. NSSL1 mean subjective ratings were notably higher than HRRR and this difference was significant (Fig. 39). There were several cases in which the NSSL1 had a much better signal in the 10-m maximum wind gusts associated with severe-wind-producing mesoscale convective systems or small clusters of storms producing severe wind. One such example is shown in Figure 40. In this case, NSSL1 had an improved evolution and structure of a bowing MCS that moved through eastern Kansas and western Missouri, producing many severe wind and some significant wind gusts. Additionally, the NSSL1 maximum 10-m winds better delineated the areas where severe wind gusts were observed. A few of the survey comments reflecting the improved wind guidance from NSSL1 are highlighted as follows: "*They both had trouble with the convection behind and on the edge of the MCS and didn't really develop it for a while. The 1-km though did handle the speed of the MCS much better. As far as hazards go, the 1-km NSSL-WRF did a great job*", "*A lot of times models have a nice accurate depiction, but to be in the exact right place with the exact right forward propagation is very rare. The NSSL1 forecast looked pretty amazing in that regard. This shows up in the reflectivity and in the 10-m winds. HRRR was playing catch up a bit, and structure wasn't as good. I thought the UH swath in NSSL1 was near perfection. Strongest track goes right over the 1 tornado report, and then the signal really dissipates in SE KS, whereas HRRR has some spotty strong tracks later in time*", and "*Biggest difference is in depiction in wind. NSSL-WRF did far better in depicting wind reports associated with the MCS, while the HRRRv4 was displaced well to the northwest relative to reports*".

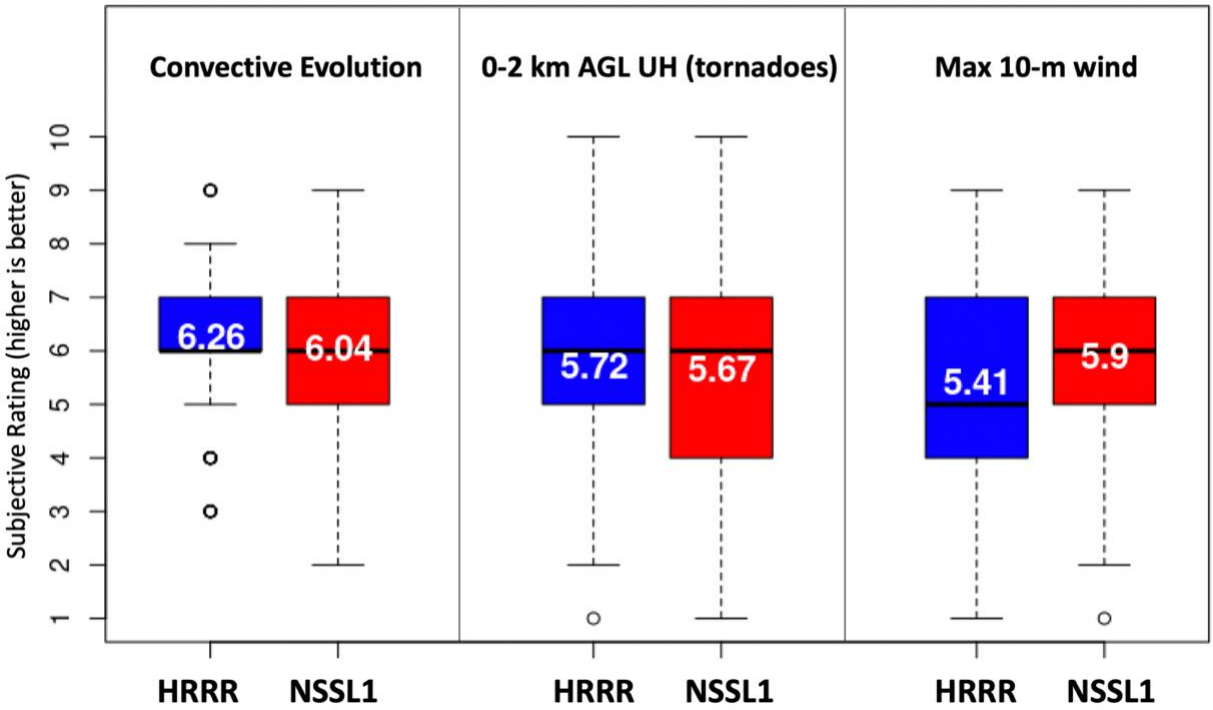

Figure 39. Boxplots of subjective rating distributions for overall convective evolution, 0-2 km AGL UH, and maximum 10-m wind speeds from HRRRv4 and NSSL1.

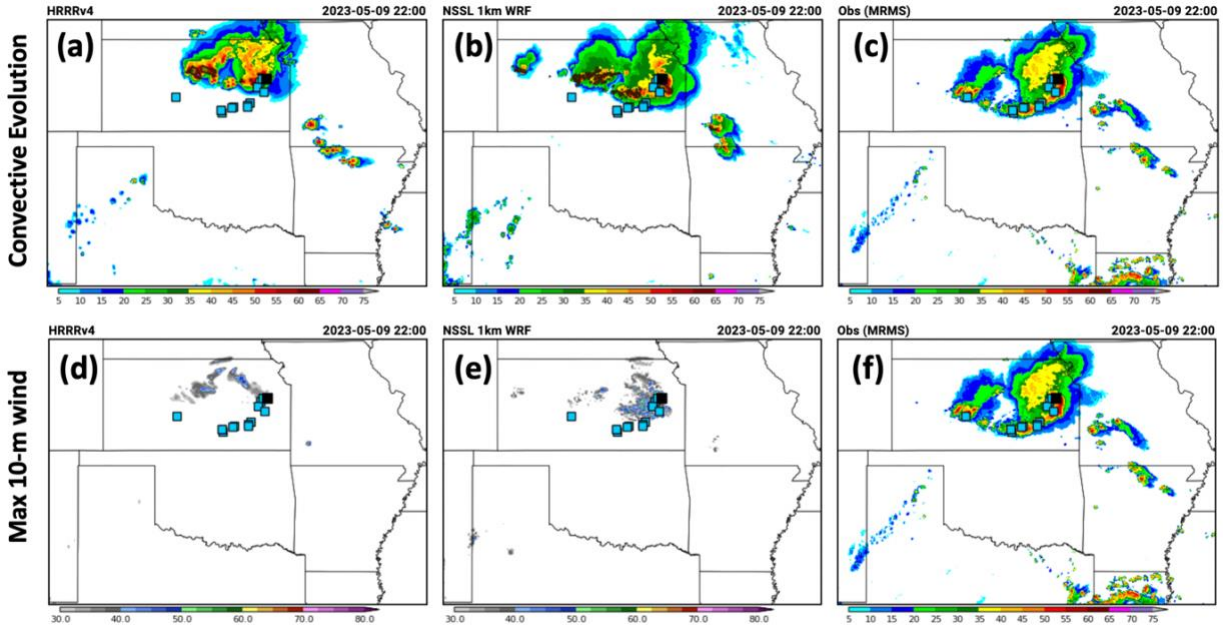

Figure 40. Simulated composite reflectivity with LSRs of severe wind gusts overlaid (blue squares) for 0000 UTC initializations valid 2200 UTC 9 May 2023 from (a) HRRRv4, (b) NSSL1, and (c) MRMS observations. (d)-(f) same as (a) and (c), except (d) and (e) show 4-h maximum 10-m winds.

3.3 Evaluation – CAM (E)nsembles

3.3.1 (E1) CLUE: 00Z RRFS vs. HREF

Similar to the deterministic evaluation of the RRFS control member to the operational HRRR (section D3), the RRFS single-physics ensemble was compared directly to the HREF, which is the operational CAM ensemble in the NWS. This evaluation was done for the 00Z run to assess the readiness of the RRFS ensemble to replace the HREF for operational convective forecasting applications on Day 1. Participants were asked to examine probabilistic storm-attribute fields, including updraft helicity, updraft speed, 10-m wind speed, and composite reflectivity, and provide a single rating for the convective day (i.e., f12-f36). A five-point Likert scale was also used in this evaluation to rate the RRFS ensemble as much worse, slightly worse, about the same, slightly better, or much better than the HREF for each field. For example, the 00Z RRFS forecast was generally rated slightly worse than the 00Z HREF forecast for the derecho-producing MCS across Kansas on 9 May 2023, owing to the HREF having better orientation and centering of probabilities on the preliminary local storm reports (Fig. 41).

For the storm-attribute fields, the 00Z HREF has slightly higher ratings for updraft helicity, where the median rating was slightly worse for the RRFS compared to the HREF (Fig. 42). In the subjective comments, SFE participants noted that the RRFS ensemble is more likely to generate high probability (i.e., ≥50%) false alarm forecasts (with no severe weather reported) when compared to the HREF. The rating distributions are more neutral for updraft speed, 10-m wind speed, and composite reflectivity. These "*about the same*" ratings are really more indicative of positives and negatives of the RRFS and HREF forecasts balancing out than the forecasts looking similar.

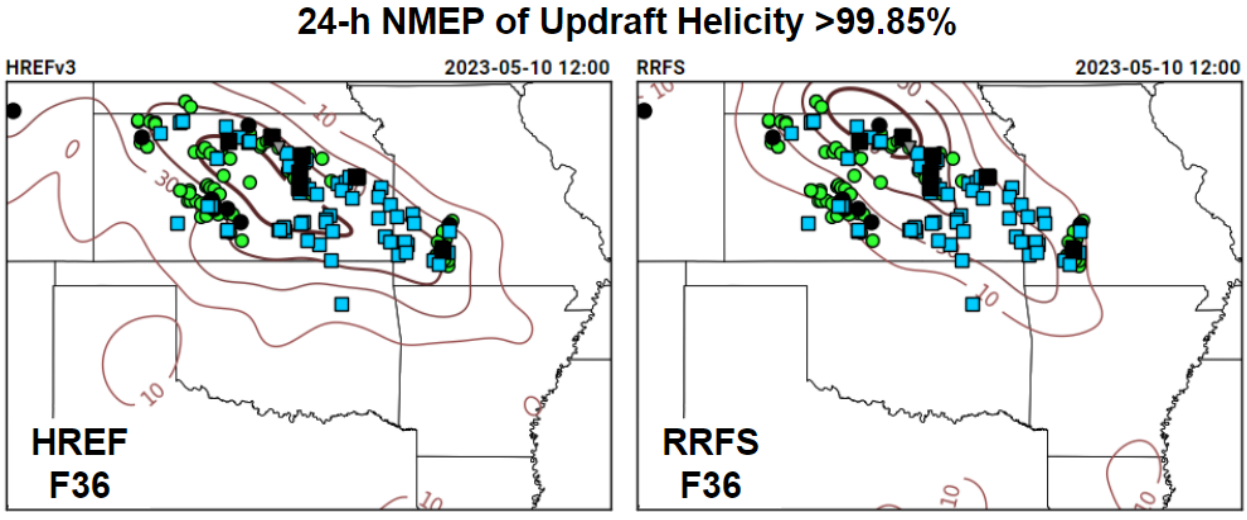

Figure 41. Example of the 2023 HWT SFE model comparison page for the RRFS vs. HREF valid for the convective day of 9 May 2023. The 24-h neighborhood maximum ensemble probability (NMEP) forecasts of UH are shown for the 00Z HREF (left panel) and the 00Z RRFS ensemble (right panel). The observed preliminary local storm reports (wind – blue boxes; sig wind – black boxes; hail – green circles; sig hail – black circles) are overlaid in both panels.

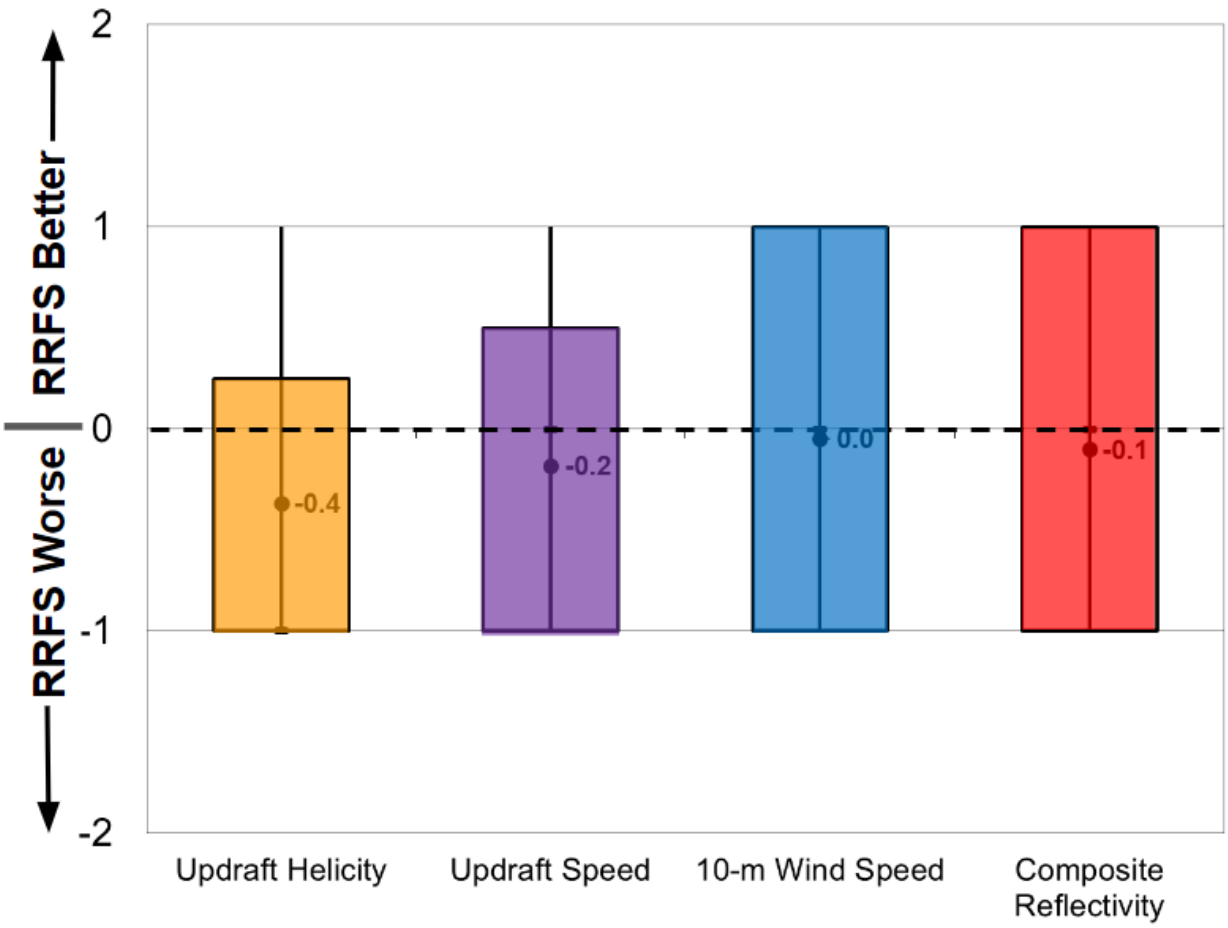

Figure 42. Distributions of subjective ratings (-2 to +2) by SFE participants of the 00Z RRFS ensemble compared to the HREF for updraft helicity (yellow), updraft speed (purple), 10-m wind speed (blue), and composite reflectivity (red). The ratings represent the RRFS ensemble compared to the HREF -2: Much Worse; -1: Slightly Worse; 0 – About the Same; +1: Slightly Better; +2: Much Better.

For the ensemble mean environmental fields, the RRFS was typically rated about the same to slightly worse than the HREF (Fig. 43). While the deterministic RRFS CAPE forecast was rated slightly worse more often than the other environmental fields, the RRFS ensemble mean CAPE forecast actually received higher average ratings than the 2-m temperature and dewpoint forecasts. For 2-m dewpoint, the RRFS was typically more moist than the HREF in the warm sector, which was perceived by the SFE participants to be a slightly worse forecast. Another frequent comment was that the RRFS mean environmental fields displayed more detailed structure and features than the HREF mean environmental fields, which is not surprising given the greater diversity in the HREF in terms of model core and physics.

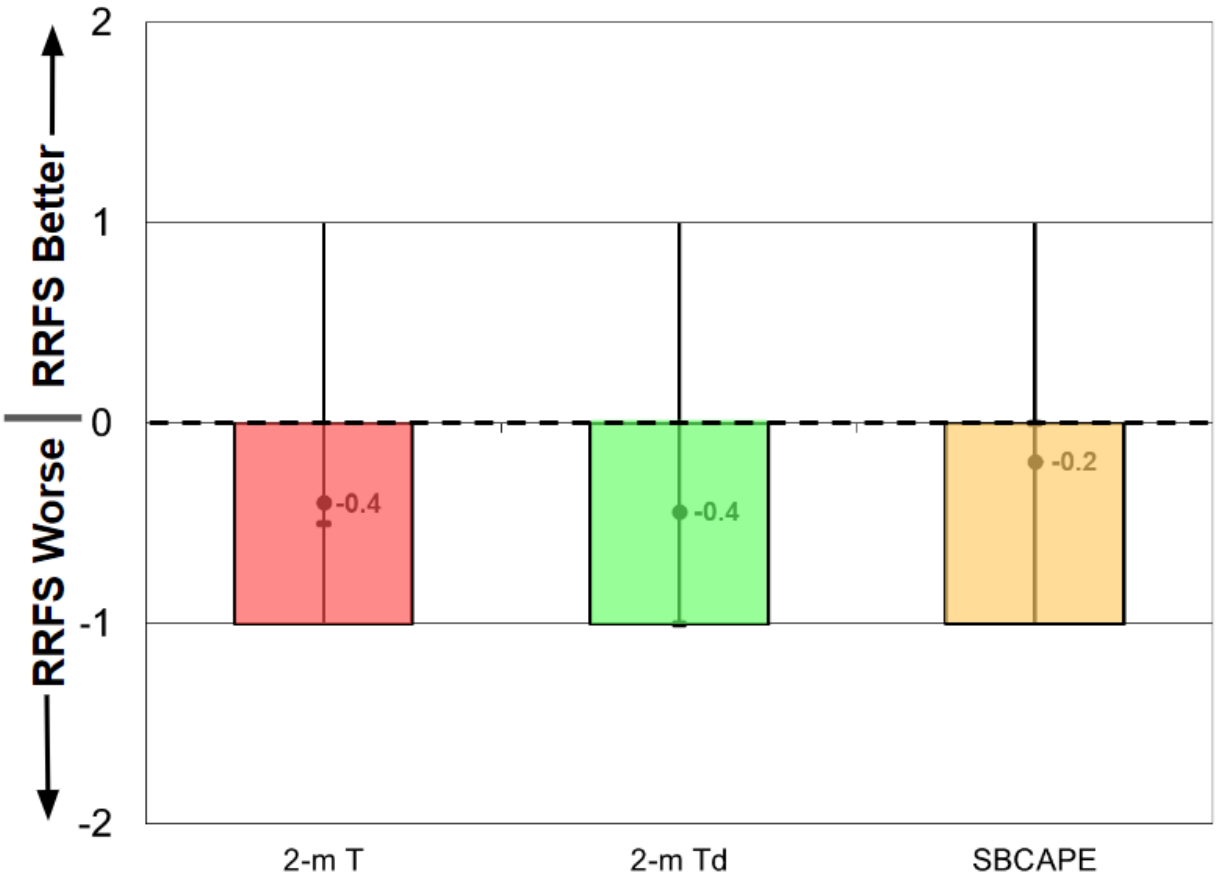

Figure 43. Same as Fig. 42, except for environmental mean fields of 2-m temperature (pink), 2-m dewpoint (light green), and SBCAPE (yellow).

3.3.2 (E2) CLUE: 12Z Day 1 RRFS Physics & Time-Lagging vs. HREF

This evaluation assessed the skill of multiple RRFS ensemble configurations compared to the operational HREFv3 at Day 1 lead times. At the time of this experiment, it was unknown how many RRFS ensemble members would be computationally feasible at 6-hour initialization times. Therefore, multiple time-lagging strategies were tested to determine if they could meet or exceed the skill of a 10-member RRFS ensemble initialized at a single time (t). Additionally, this evaluation tested the impact of a mixed-physics approach within the ensemble to determine if the increased member diversity could improve forecast quality compared to a single-physics (with stochastic perturbations) approach. Participants were asked to assess and compare the skill of the following six ensembles initialized at 1200 UTC:

1. RRFS (single physics; 10 members at t)
2. RRFS-TL6 (single physics; 6 members at t and 6 members at t-6h)
3. RRFS-TL12 (single physics; 6 members at t and 6 members at t-12h)

4. RRFSphys (mixed physics; 10 members at t)

5. RRFSphys-TL6 (mixed physics; 6 members at t and 6 members at t-6h)

6. HREF (operational HREFv3)

Respondents were provided 4-h and 24-h updraft helicity, updraft speed, and 10-m wind speed neighborhood probabilities, as well as 1-h composite reflectivity and 6-h QPF fields with which to base their assessment of the ensembles. MRMS MESH, local storm reports, NWS warnings, and NLDN lightning flashes were provided as ground truth observations. For the first part of this evaluation, participants were asked to "*Subjectively rate on a scale of 1 (Very Poor) to 10 (Very Good) the 24-h ensemble storm-attribute products during the Day 1 12-12Z period with regard to the quality of guidance for severe weather forecasting. Focus primarily on Updraft Helicity, but Updraft Speed & 10-m Winds can be used to supplement the ranking, especially on days without supercells*." Respondents were further instructed to rate each ensemble independently, such that the assessment of one ensemble should not directly consider the performance of another ensemble. This method enabled participants to rate different ensemble forecasts equally when warranted and was found to be more robust against missing data than a traditional ranking system. Upon completing these assessments, participants were given an opportunity to share their thoughts about any differences in the ensembles via an optional open response question. Participants then shared and elaborated on these insights during a group discussion period immediately following the survey.

The HREF received the highest rating on average during the 5-week experiment with a mean subjective score of 6.774 (Fig. 44). The RRFS ranked second at 6.730, followed by the RRFSphys (6.600), RRFS-TL6 (6.341), RRFSphys-TL6 (6.171), and RRFS-TL12 (6.032). It is notable that there was only a 0.742 difference between the highest and lowest mean ratings, suggesting that all six ensembles performed similarly on average. This is further supported by the response distributions shown in Figure 44 which demonstrate similar characteristics across all configurations. Indeed, only the difference between the HREF and RRFSphys-TL12 mean scores was found to be statistically significant at the 95% confidence level. All ensembles except the HREF and RRFSphys-TL6 received a minimum rating of 2 at some point during the experiment, and all ensembles except the RRFSphys-TL6 received a maximum rating of 10. Note that some ensembles experienced outages during the experiment, and so the number of responses is not uniform across all ensembles.

Objective statistics (ROC area and reliability) were also computed over the SFE domains for probabilities of simulated reflectivity ≥ 40 dBZ in the HREF, RRFS, and RRFSphys. The statistics follow the subjective ratings quite closely, with HREF have the highest ROC area (0.821), followed by RRFS (0.813) and RRFSphys (0.804; Figs. 45a & 46a). Additionally, HREF had improved reliability over RRFS, while RRFS and RRFSphys were quite similar in terms of reliability (Figs. 45b and 46b).

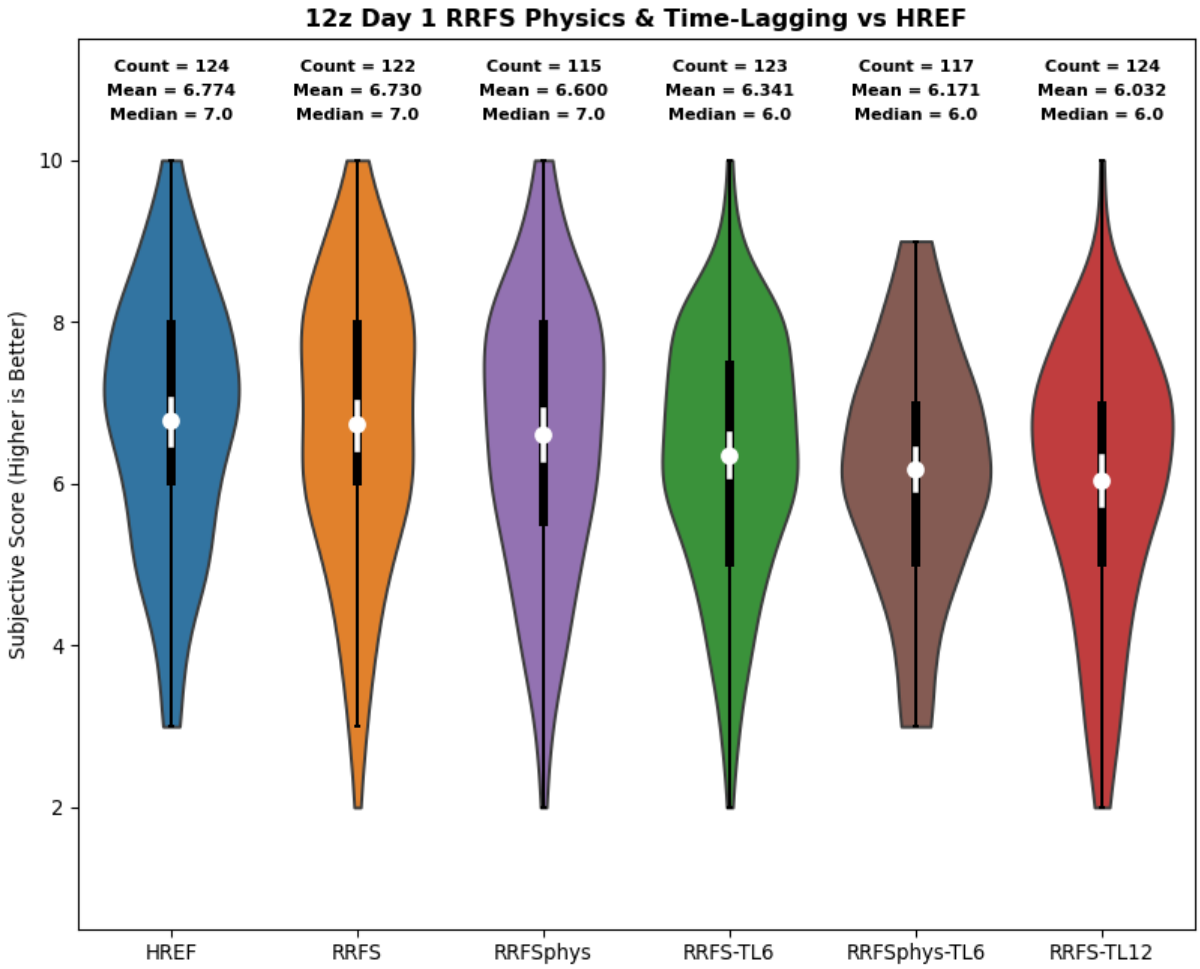

Figure 44. Distribution of subjective scores received by each ensemble at Day 1 lead times during the 5-week experiment. The white dots represent the mean scores for each ensemble, and the white bars indicate the 95% confidence intervals for each mean.

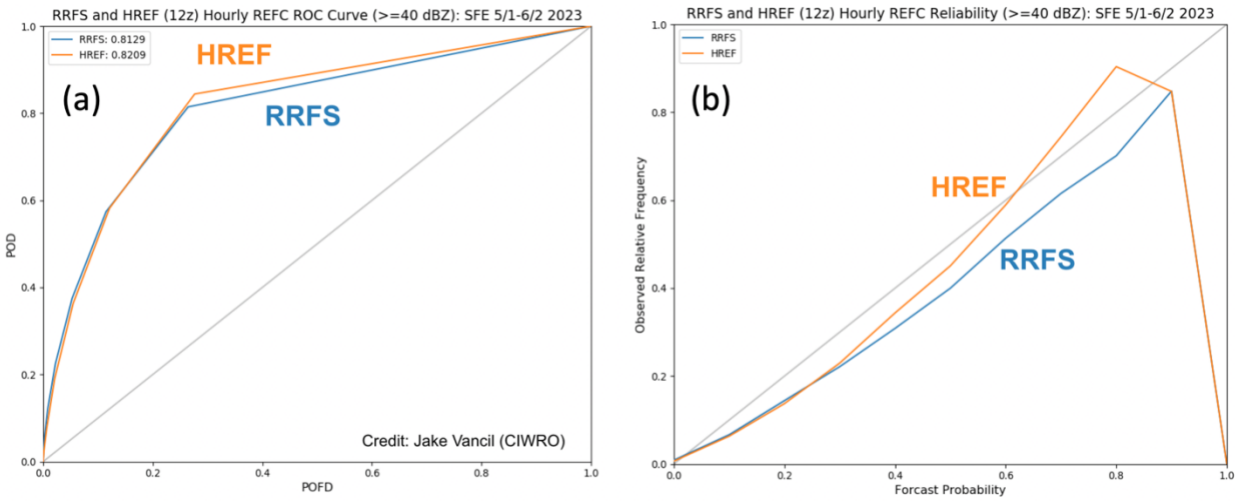

Figure 45. (a) ROC curve for reflectivity probabilities ≥ 40 dBZ over SFE 2023 domains for HREF (orange) and RRFS (blue). (b) Reliability diagrams for reflectivity probabilities ≥ 40 dBZ.

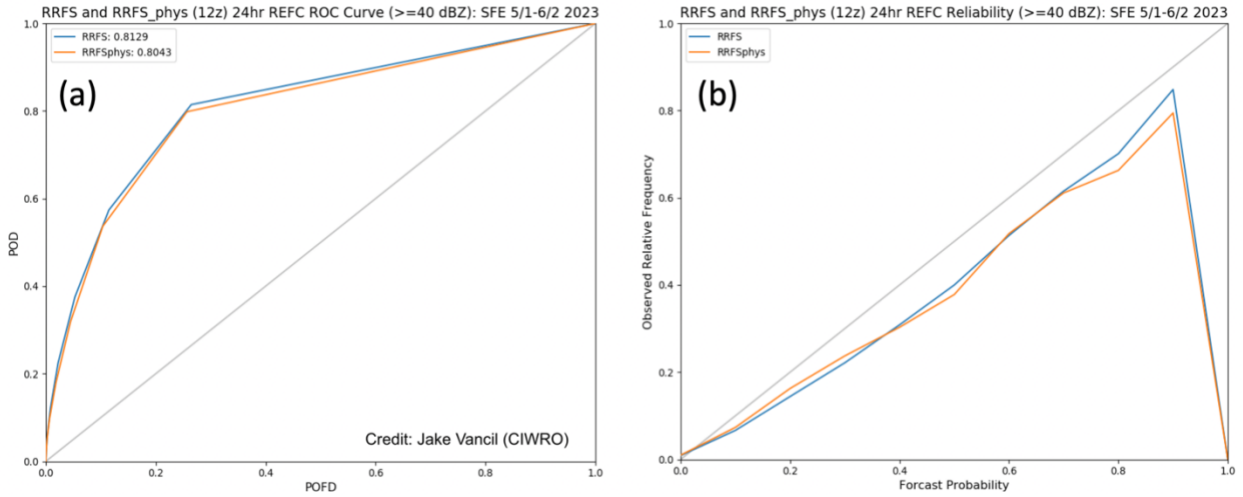

Figure 46. Same as Figure 27, except for RRFS and RRFSphys.

In the open response question and post-survey discussion, participants again commented on how similar the 12Z ensembles were on average at Day 1 lead times. When differences in the forecasts were observed, respondents noted that the RRFS ensembles tended to produce neighborhood probabilities that covered a smaller and more focused spatial extent than the HREF. These "bullseye" forecasts were often praised by respondents for a perceived reduction in false alarm area, but the smaller probability fields also occasionally missed or underforecast severe reports near the periphery of the events (Fig. 47). Participant opinions of these differences varied greatly each day, and post-survey discussion frequently revealed that ratings were closely tied to whether the respondent placed greater value on detection or false alarm when making their assessments. Overall, the RRFS ensembles were found to provide similar or slightly reduced forecast quality to the HREF at Day 1 lead times.

Respondents indicated some surprise at how similar the time-lagged ensembles were on average when compared to the ensembles initialized at a single time. One participant commented, "[T]*he RRFS time-lagged solutions didn't significantly change for each lag. Not sure if that is/was an issue with dispersion (underdispersed)?*" while another stated, "*The non-time-lagged ensembles tended to do better, but the differences are pretty small.*" In general, the time-lagged solutions were perceived to be slightly worse than the non-lagged ensembles, and the 12-h lag was viewed less favorably than the 6-h lag. That said, there were a few cases during the experiment when participants felt the time-lagging strategies helped increase the dispersiveness of the RRFS and improved the forecast: "*The RRFS was quite underdispersive [...]. However, time-lagging improved this a bit, especially on day 1. In particular, the RRFS-TL6 was somewhat close to the impressive HREF output for day 1.*"

Finally, participants were undecided on the benefits of a mixed-physics ensemble. In some cases, the mixed-physics ensembles produced more dispersive forecasts which increased the areal extent of the neighborhood probabilities without adversely affecting the probability magnitudes. One participant summarized this well, stating, "The advantage of time-lagged ensembles is not clear, but the multi-physics ensemble expands the region

of high probability and improves the probability forecasts." However, these differences were typically small and did not have much apparent impact on forecast quality as noted by another participant: "*Generally speaking, RRFS and RRFSphys seemed to perform similarly*." Based on these results, it appears that the time-lagging and mixed-physics strategies are at least comparable in quality to a 10-member RRFS ensemble initialized at a single time. As such, these strategies may be viable alternatives if computational limitations preclude a larger operational RRFS ensemble.

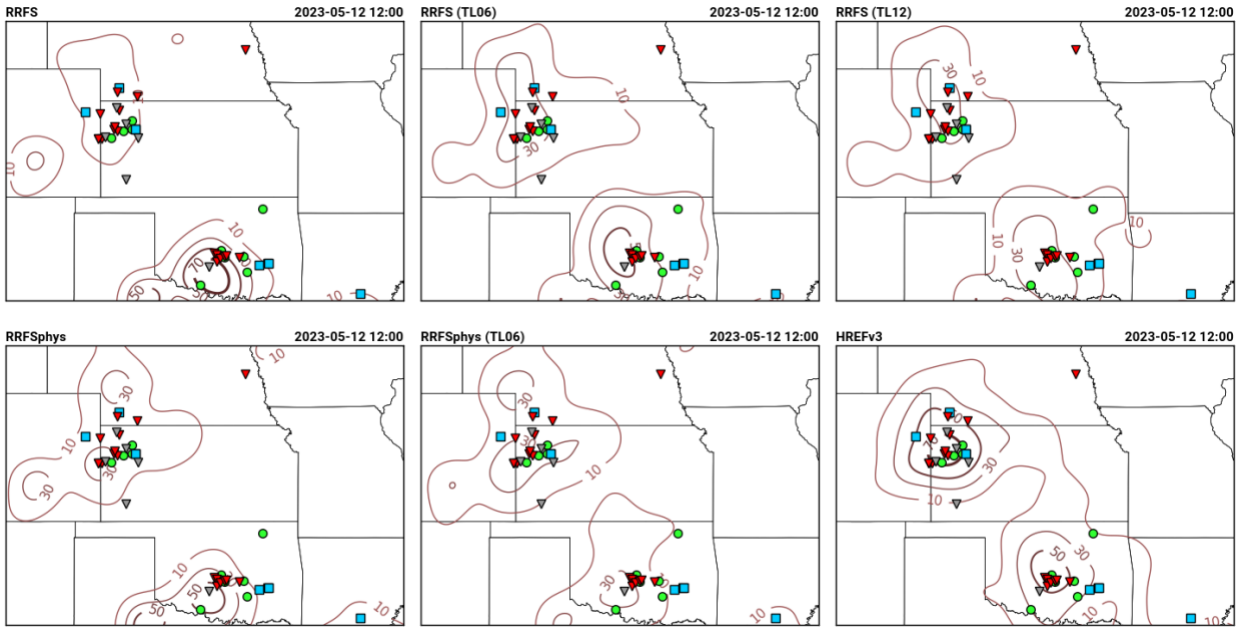

Figure 47. 24-h neighborhood probabilities of updraft helicity exceeding the 99.85th percentile for the period 1200 – 1200 UTC 11-12 May 2023. Red triangles represent tornado reports, blue squares are wind reports, and green circles are hail reports.

3.3.3 (E3) CLUE: 12Z Day 2 RRFS Physics & Time-Lagging vs. HREF

This evaluation was identical to the previous one, except participants evaluated the HREF and RRFS ensembles at Day 2 lead times. Specifically, respondents were asked to "*Subjectively rate on a scale of 1 (Very Poor) to 10 (Very Good) the 24-h ensemble storm-attribute products during the Day 2 12-12Z period with regard to the quality of guidance for severe weather forecasting. Focus primarily on Updraft Helicity, but Updraft Speed & 10-m Winds can be used to supplement the ranking, especially on days without supercells*." Participants were given access to the same fields and observations as before, but the forecast products were derived from the previous day's 12Z ensemble runs. As such, this evaluation focused on ensemble forecast quality at forecast hours 24 - 48.

As before, the HREF once again received the highest ratings overall, with a mean subjective score of 6.500 (Fig. 48). This is only a decrease of 0.274 from the Day 1 scores, suggesting impressive consistency in forecast quality at longer lead times. In contrast, all

five RRFS ensembles saw notably degraded performance at Day 2 lead times. The RRFS received the second highest mean rating of 5.653, followed by the RRFSphys (5.466), RRFSphys-TL6 (5.388), RRFS-TL6 (5.349), and RRFS-TL12 (5.315). The RRFS mean rating fell a considerable 1.074 points compared to its Day 1 score, and this was found to be statistically significant when compared to the HREF at the 95% confidence level. Indeed, the HREF Day 2 mean rating was significantly higher than that of all RRFS ensembles at the same lead time. Otherwise, the RRFS ensembles all performed similar to each other on average, and their mean scores were not significantly different at the 95% confidence level. All six ensembles received a low rating of 1 at some point during the 5-week experiment, and the HREF was the only ensemble to receive a rating of 10. The RRFSphys and RRFS-TL6 ensembles both received a maximum rating of 9, while the other RRFS ensembles peaked at a rating of 8.

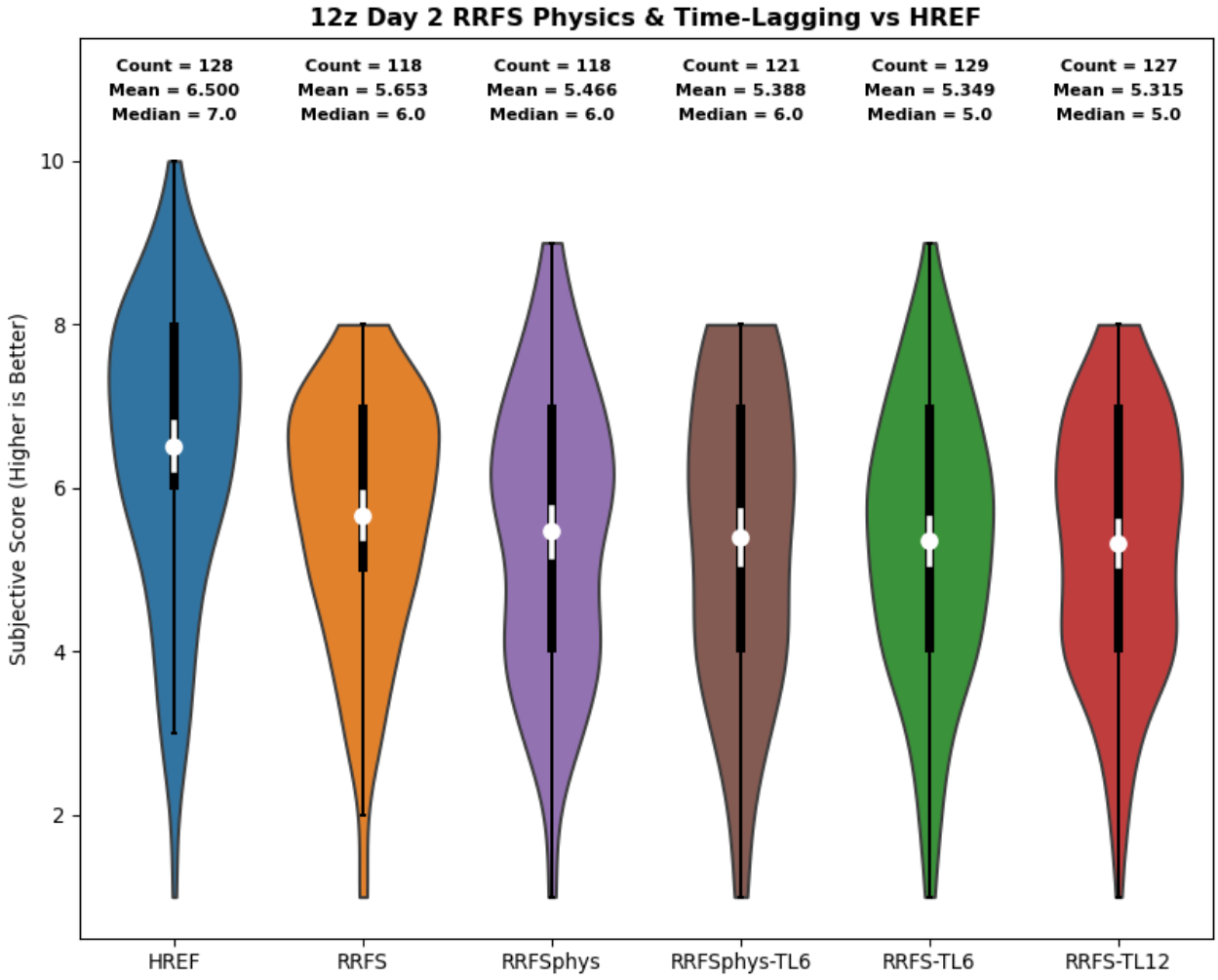

Figure 48. Same as Fig. 44, but for Day 2 lead times.

Participant comments were somewhat critical of the RRFS ensembles at the Day 2 lead times, and particularly noted the degraded performance compared to the Day 1 forecast. For example, a couple representative comments from the experiment read,

"*RRFS had basically zero signal for the day 2 event, while the HREF showed good coverage of modest probabilities*," and, "*The RRFS configurations were comparable to HREF on Day 1, but I thought the HREF was the clear winner on Day 2.*" On average, the RRFS ensembles tended to underforecast storm attribute neighborhood probabilities at Day 2 lead times and often exhibited large spatial errors compared to the HREF. Some participants indicated that the timelagged ensembles potentially improved the forecast quality by increasing the spatial coverage of the probabilities: "*For the Day 2 - the timelagged members seemed to perform better (particularly the TL6) with greater spatial errors in the RRFS/RRFSphys.*" However, differences between the time-lagged and non-lagged ensembles varied greatly from day to day, and participant opinions did not reach a consensus on their skillfulness at Day 2 lead times.

The mixed-physics ensembles were again found to be similar to or worse than the single-physics ensembles in this evaluation. Participants noted the mixed-physics ensembles often produced lower neighborhood probabilities than the other ensembles, and the probability fields tended to cover much larger areas. This effectively increased the false alarm of the ensemble while also reducing probability magnitudes where severe weather was observed. However, the greater spatial coverage of the neighborhood probabilities sometimes captured severe events that were missed by the single-physics ensembles. Respondents varied in how favorably they viewed these differences, and it is unclear from these results whether the current mixed-physics scheme is viable at Day 2 lead times. Overall, this survey found that more work is needed for the RRFS ensemble to reach an equivalent forecast quality to the HREF at these longer lead times. Time-lagging strategies showed some potential for improved forecasts at the Day 2 time-frame, but results were mixed and further investigation is needed.

3.3.4 (E4) CLUE: Medium-Range Lead Time/Core/Members

In this survey, subjective ratings of forecast skill were assigned to CAM ensemble guidance from a 5-member subset of the NCAR FV3 and 5-member MPAS ensemble at lead times of 3-5 days. Additionally, subjective ratings were assigned to the 10-member NCAR FV3 for lead times of 3-7 days. Specifically, SFE participants were asked, "*Subjectively rate on a scale of 1 (Very Poor) to 10 (Very Good) the 24-h ensemble storm-attribute products during the 12-12Z period with regard to the quality of guidance for severe weather forecasting. Focus primarily on Updraft Helicity, but Updraft Speed & 10-m Winds can be used to supplement the ranking, especially on days without supercells. In addition, the 4-h storm-attribute products and 1-h composite reflectivity paintballs and probabilities can be utilized to adjust or fine-tune the overall ratings*." An example of the forecasts is shown in Figure 49.

In comparisons between the NCAR MPAS and 5-member NCAR FV3 ensemble subset (Fig. 50), the NCAR MPAS ensemble performed best at every lead time it was available, and the differences in mean subjective ratings were statistically significant (Student's t-test with α = 0.05). One representative comment was, "*The NCAR-FV3 looked like it split into two areas of focus shortly after the day 5 forecast which didn't really represent the swath of storm report we actually saw. MPAS pretty much kept to*

one bullseye for days 3-5 and did generally better than the FV3." However, participants sometimes noted over-forecasting in NCAR MPAS.

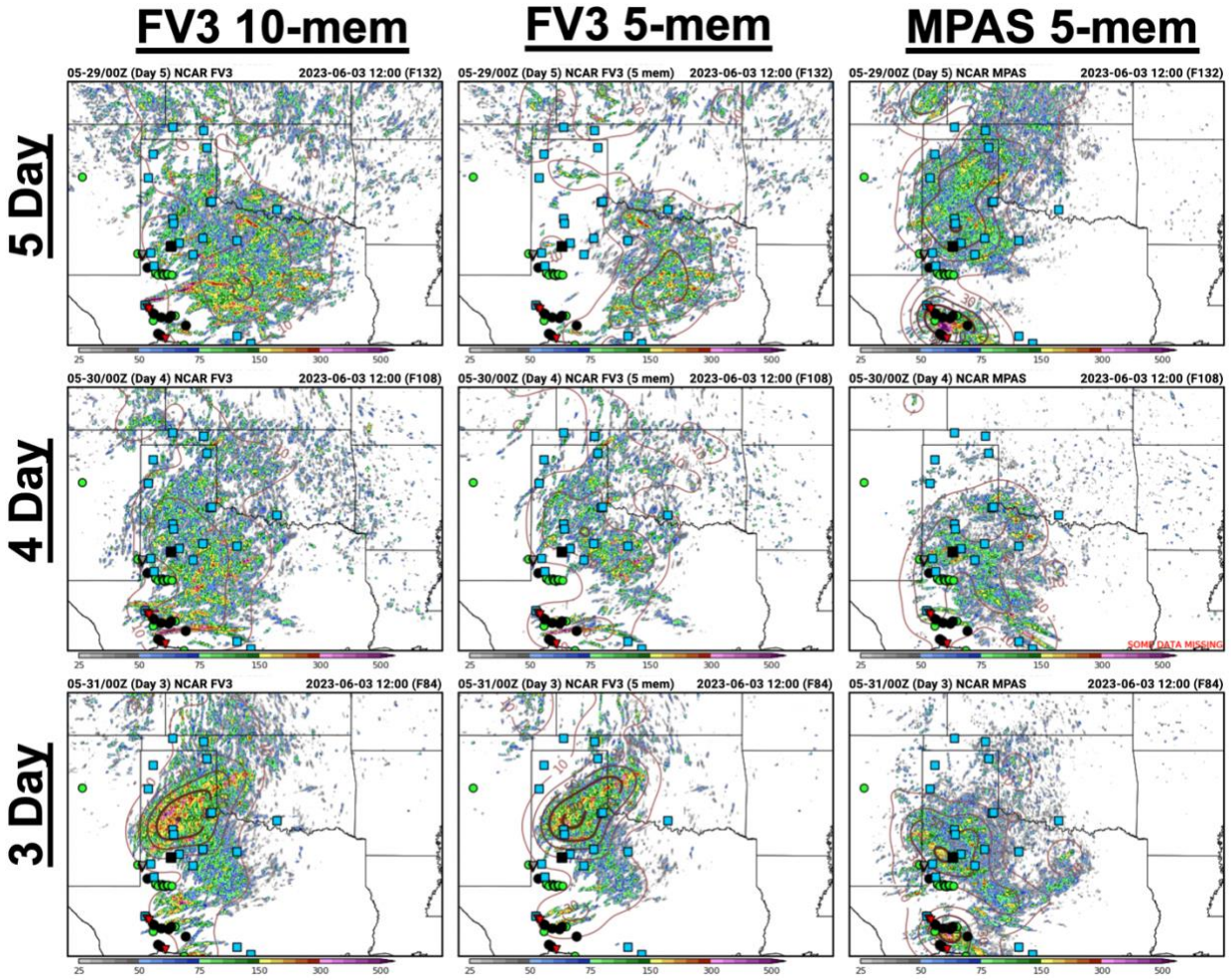

Figure 49. Example of multi-panel comparison webpage for the E4 Medium-Range Lead Time/Core/Members evaluation. In each panel, 24 h maximum UH (shaded) and neighborhood probability of UH ≥ 99.85th percentile (contours) is displayed. LSRs are also overlaid (wind – blue squares, hail – green circles, and tornado – upside-down triangles; significant reports are filled in black).

More generally, there were several days in which value was highlighted all the way out to Day 7. For example, for one case a participant noted, "*The general region was highlighted out to Day 7, which was impressive. There was an eastward displacement of the probs in all models, though.*" Also, it was common for there to be inconsistent forecast quality with decreasing lead time. In other words, sometimes the later lead times actually performed better than the shorter lead times. For example, in one case a participant commented, "*NCAR FV3 was very impressive at long ranges around 6-7 days, but there tended to be poorer performance and some jumpiness by Day 5 forward*."

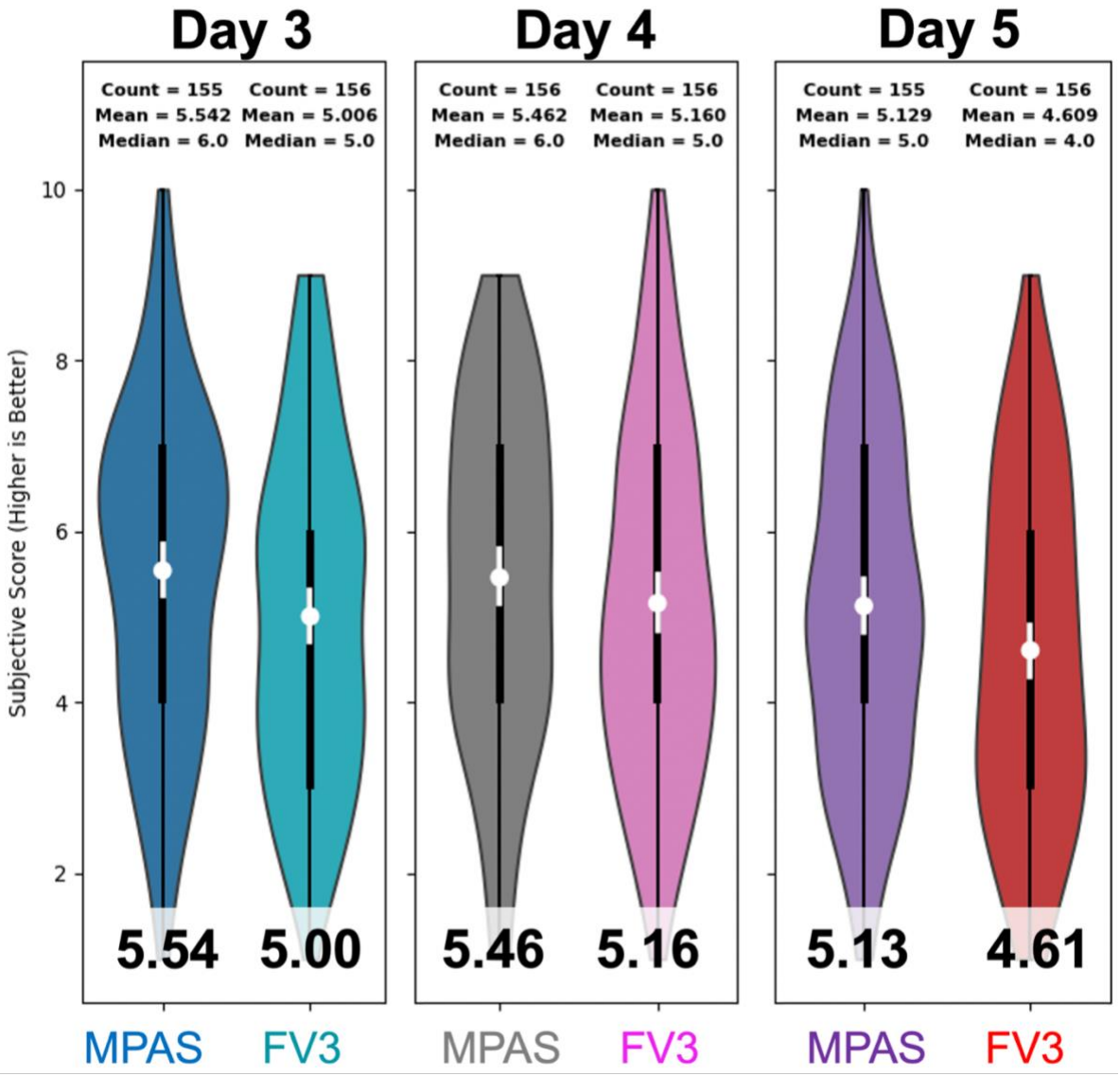

Figure 50. Distributions of subjective ratings for the NCAR MPAS and 5-member NCAR FV3 ensemble subset at Day 3 (left), Day 4 (middle), and Day 5 (right) lead times. Mean subjective ratings are indicated at the bottom of each violin plot. The white dots represent the mean scores for each ensemble, and the white bars indicate the 95% confidence intervals for each mean.

For the 10-member NCAR FV3, there was a gradual degradation in mean subjective ratings with increasing lead time from Days 3 to 6. By Day 6, the mean subjective ratings appeared to level out with Day 6 and 7 both achieving a mean subjective rating of 4.50 (Fig. 51). Additionally, there were only small differences in mean subjective ratings between the 5- and 10-member NCAR-FV3 ensembles. For example, at Day 5 the 10-member mean rating was 4.64 (Fig. 50), while the 5-member was 4.61 (Fig. 50).

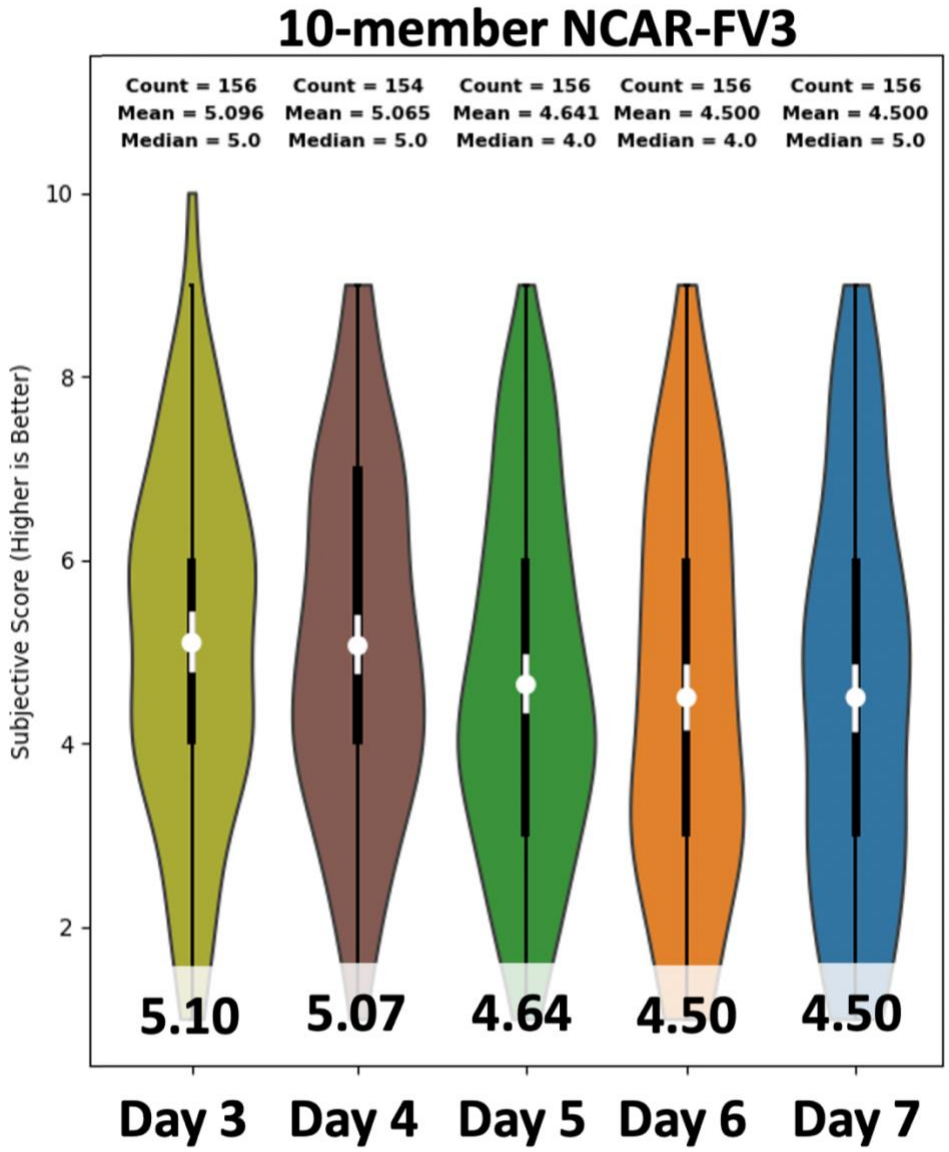

Figure 51. Distributions of subjective ratings for the 10-member NCAR FV3 ensemble for lead times of Day 3 to Day 7. Mean subjective ratings are indicated at the bottom of each violin plot. The white dots represent the mean scores for each ensemble, and the white bars indicate the 95% confidence intervals for each mean.

3.4 Evaluation – (A)nalyses

3.4.1 (A1) Mesoscale Analysis Background

Two hourly versions of 3D-RTMA with different backgrounds were subjectively evaluated by participants during the 2023 HWT SFE. The evaluation was performed to assess the quality and utility of these analysis systems for situational awareness and short-term forecasting of convective-weather scenarios. The 3D-RTMA RRFS used the FV3-based RRFS as the first-guess background while the 3D-RTMA HRRR used the operational HRRR for first-guess background and serves as the baseline for this evaluation. The hourly analyses for composite reflectivity, 2-m temperature (e.g., Fig. 52),

dewpoint, SB/ML/MUCAPE, and the significant tornado parameter (STP) were examined during the 18-03 UTC period on the following day. The SFE participants were tasked with looking through all of these fields during this period and arrive at a single rating of the quality of the 3D-RTMA RRFS compared to the 3D-RTMA HRRR.

In general, the two versions of 3D-RTMA were typically similar to one another with the 3D-RTMA RRFS having slightly larger errors over the domains in 2-m temperature and dewpoint. As seen in prior years, the biggest differences in the 2-m temperature field were most commonly associated with effects from convection. In general, the HRRR-based version handled the effects of convection on 2-m temperature better than the RRFS-based version through more accurate representation of the size, shape, and magnitude of cold pools and thunderstorm outflows (e.g., eastern Kansas in Fig. 52). In terms of the overall subjective ratings from SFE participants, the majority of responses indicated the 3D-RTMA RRFS was about the same to slightly worse than the HRRR-based version (Fig. 53). The participants noted some common issues in the RRFS-based version: overall 2-m moist bias across the domain, too moist in dry air (e.g., behind dryline), low CAPE bias overall, horizontal convective-roll-like structures in the CAPE field were rather prominent and distracting, thunderstorm outflows were often too early, cold, and/or expansive, and spurious convection in the background 1-h forecast could often disrupt derived environmental fields (e.g., STP) in a negative fashion.

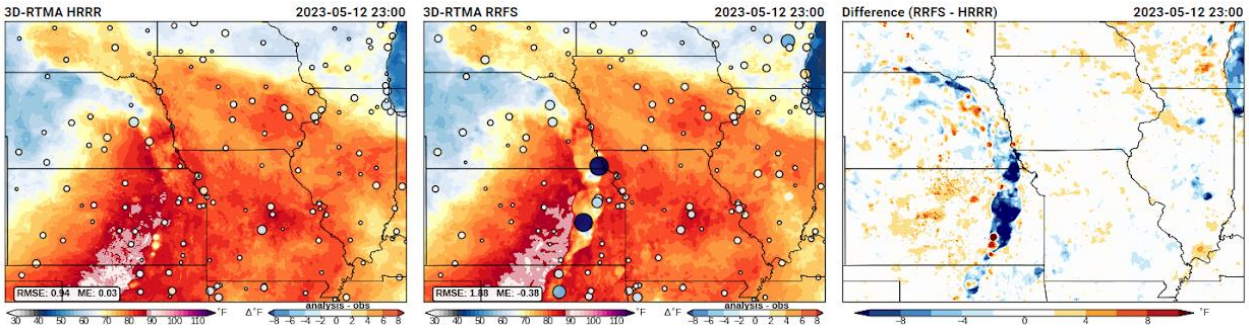

Figure 52. Example of the website comparison page for the 3D-RTMA during the 2023 HWT SFE. The 3D-RTMA HRRR baseline is shown in the left panel, the 3D-RTMA RRFS is in the middle panel, and the difference plot (3D-RTMA RRFS - 3D-RTMA HRRR) is shown in the right panel. The 2-m temperature analysis valid at 2300 UTC on 12 May 2023 is shaded in the left and middle panels. The difference (analysis-obs) at METAR sites is shown by the size and shading of the dots in the left and middle panels.

In addition to examining the 3D-RTMA systems at their native resolution (i.e., 3-km grid spacing), participants also examined upscaled versions (i.e., 40-km grid) for comparison to the widely used SPC RAP-based mesoanalysis. This was a practical exercise to determine the readiness of the 3D-RTMA systems to replace the functionality and capacity currently served by the SPC mesoanalysis. SFE participants were asked to rank (from best to worst; i.e., 1 to 3) the overall quality of the analysis for situational awareness and short-term forecasting, despite the limitation of not having observational truths for all of the fields. The 3D-RTMA HRRR had the lowest (i.e., best) mean ranking of 1.7, while the SPC mesoanalysis came in second with a mean ranking of 2.0, and the 3D-RTMA RRFS had the highest mean ranking of 2.3.

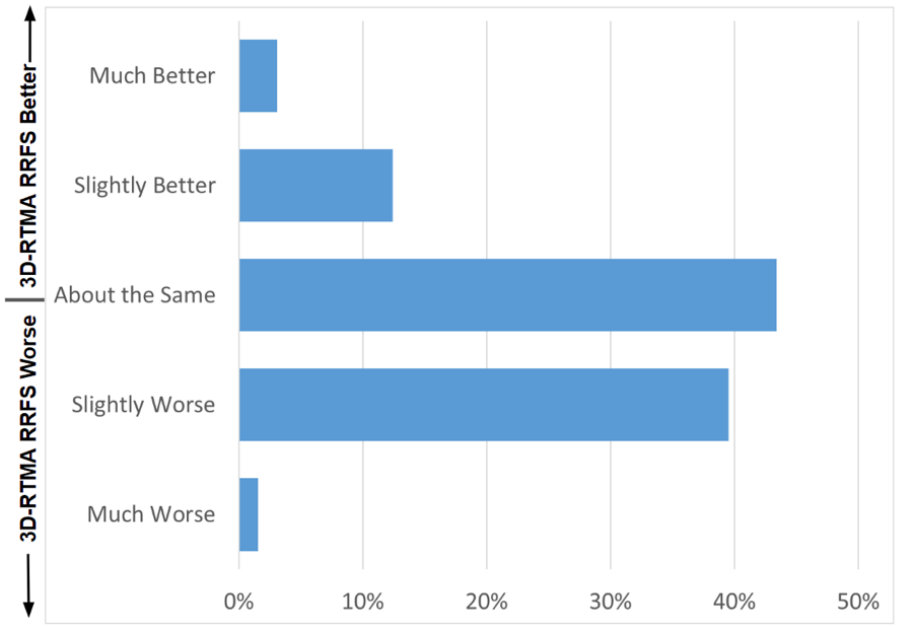

Figure 53. Percentage of subjective ratings by SFE participants for each rating category (Much Worse, Slightly Worse, About the Same, Slightly Better, and Much Better) of the 3D-RTMA RRFS compared to the 3D-RTMA HRRR.

3.4.2 (A2) Storm-scale Analyses

The Warn on Forecast System (WoFS) was used to explore whether a high resolution, rapidly updating ensemble DA system can serve as a verification source for severe weather. Specifically, the 15-minute maximum forecasts of 80-m winds, 2-5 km AGL UH, and column-maximum updraft speed from WoFS (cycled every 15 minutes) were used as a proxy for the analysis (i.e., ground truth) of severe weather. The WoFS ensemble analysis fields were accumulated from 1800 UTC through 0300 UTC for comparison with MRMS-derived products [composite reflectivity, midlevel rotation tracks, and maximum estimated size of hail (MESH)] and preliminary local storm reports, (Fig. 54).

The goal of the evaluation was to assess the current capability of WoFS to produce output for diagnosing severe weather. Overall, the WoFS ensemble analysis fields were positively viewed in terms of lining up with radar-derived proxies of severe weather, preliminary local storm reports, and a subjective assessment of severe weather based on the environment. Overall, the WoFS analyses of 2-5 km AGL UH and 80-m winds received higher subjective ratings than the column-maximum updraft speed (Fig. 55) in terms of alignment with severe-weather occurrence. The 80-m wind analyses have been examined in previous years, so the slightly higher mean ratings for the 2-5 km AGL UH analyses are surprising and encouraging. Based on participant comments, there is room for improvement in terms of optimizing the analysis products, which simply use the ensemble maximum for UH and updraft speed. Overall, the participants found this to be an interesting and promising approach for using a rapidly cycling convection-allowing ensemble system.

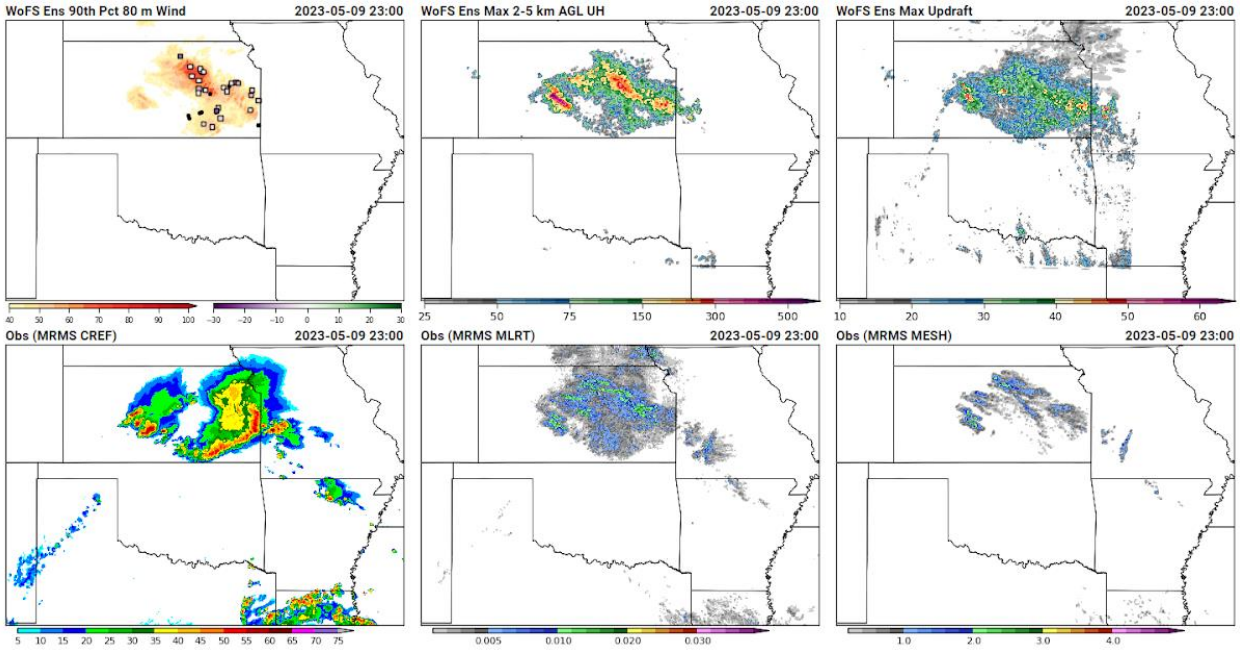

Figure 54. Example of the website comparison page for the WoFS analyses during the 2023 HWT SFE. The 9 May 1800-2300 UTC accumulated ensemble 90th percentile 80-m wind is shown in the upper-left panel, the ensemble maximum 2-5 km AGL UH in the upper-middle panel, and the ensemble maximum column-maximum updraft speed in the upper-right panel. The observed MRMS composite reflectivity is in the bottom-left panel, observed MRMS midlevel rotation tracks are in the bottom-middle panel, and the MRMS MESH is in the bottom-right panel. In the upper-left panel, the wind damage reports are the black circles while the measured gusts are the open squares shaded by the difference (analysis-obs) of the gust measured at that location.

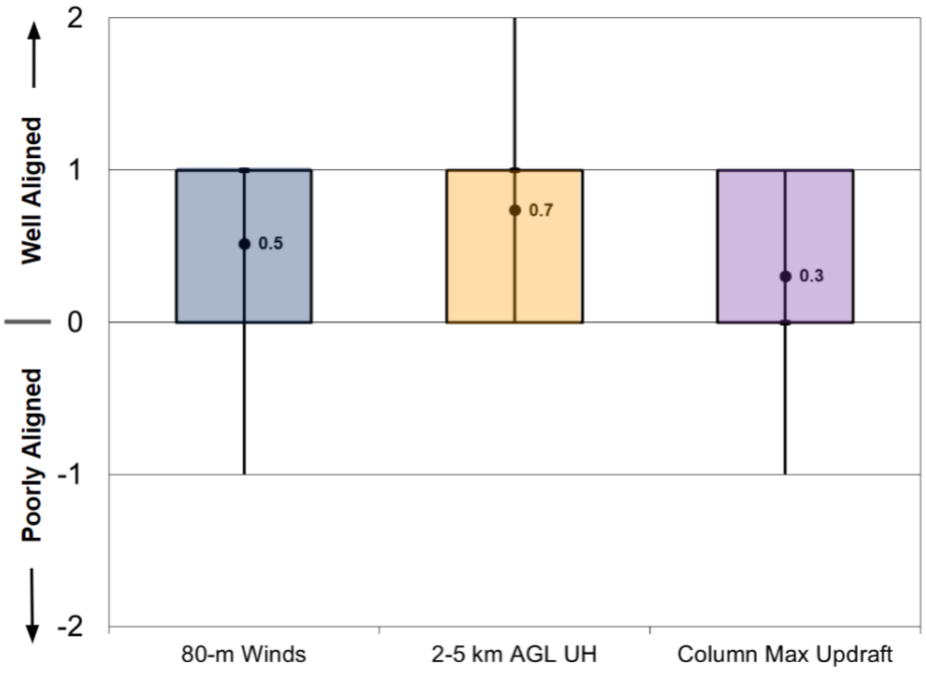

Figure 55. Distributions of subjective ratings (-2 to +2) by SFE participants of the WoFS storm-scale analysis for ensemble 90th percentile 80-m winds (blue), 2-5 km AGL UH (light orange), and column-maximum updraft speed (light purple), where the ratings represent how well the WoFS analyses align with the MRMS observed fields and preliminary severe wind reports: -2 – Very Poorly; -1 – Poorly; 0 – Unsure/Neutral, neither poorly nor well; 1 – Well; 2 – Very Well.

3.5 Evaluation – Funded (P)rojects

3.5.1 (P1) ISU ML Severe Wind Probabilities

This evaluation assessed two machine-learning models which estimate the probability that a wind damage report was associated with severe-intensity winds (> 50 kts). The first version of the guidance utilizes a stack generalized linear model (GLM), which is an ensemble of multiple models. The second version employs a gradient boosted machine (GBM) which was determined to be the best single model via objective measurements on an independent test set. The probabilities produced by either machine learning model were displayed alongside observed wind reports on an interactive website designed for this evaluation (Fig. 56). Participants were asked to evaluate on a scale of 1 (very poor) to 10 (very well) how well either machine learning model provided useful and accurate probabilistic information regarding the likelihood that wind damage reports were associated with winds ≥ 50 knots.

The results show a relatively small difference in evaluated performance between the two models (Fig. 57). The GLM received a mean subjective score of 6.93, which was slightly higher than the GBM's mean score of 6.67. Additionally, the violin plots indicate a larger number of scores at or above 8 were assigned to the GLM model compared to the GBM. These results suggest that participants perceived the GLM method as slightly better than the GBM when identifying severe wind reports during this experiment.

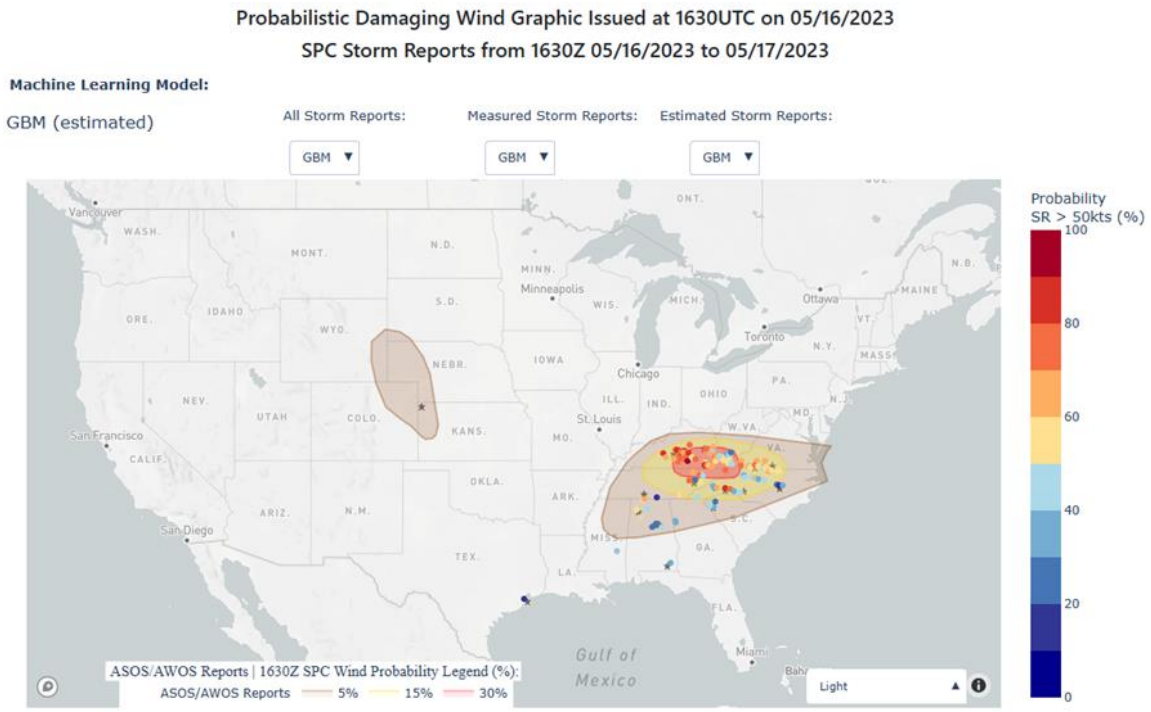

Figure 56. Example of the interactive webpage developed for the ISU Machine-Learning Severe Wind Probability evaluation during the 2023 SFE. The preliminary wind reports are shaded with the probability that the report was associated with a wind gust of ≥50 knots from the various ML algorithms. The user has the option to zoom/roam, hover over a report to see associated probabilities and report text, and choose to view all reports, just measured reports, or just damage reports.

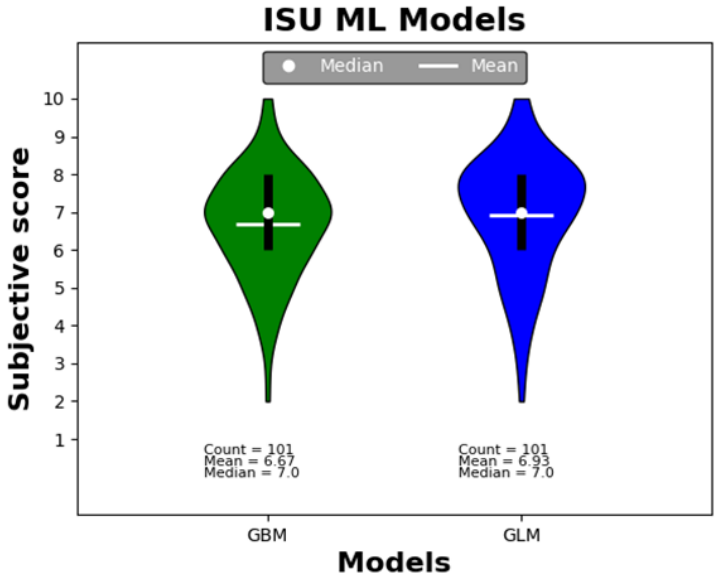

Figure 57. Violin plots representing the distribution of subjective ratings assigned to the GBM (green) and GLM (blue) models. A rating of 10 denotes excellent performance in identifying severe wind reports.

3.5.2 (P2) WoFS-PHI Spatial Hazard Probabilities

Machine-learning-based spatial hazard probabilities using predictors from WoFS and ProbSevere Version 2 (i.e., WoFS-PHI probabilities) were evaluated during a next-day evaluation activity. The primary goals of the activity were to: 1) determine the spatial radii (from 7.5 to 39 km) that participants most preferred as a function of lead time and WoFS initialization time, 2) assess the usefulness of the WoFS-PHI's 5-minute updates, and 3) solicit general feedback on the usefulness and design of WoFS-PHI.

Preliminary findings suggest that, overall, participants favored the 15- and 30-km radii at lead times of 1, 2, and 3 hours, as these radii received the most favorable rankings (smallest numbers; Fig. 58). As lead time increased, participants showed a slight preference toward larger radii. For example, at 1 hour lead times, the 15 km radius had a slightly smaller (i.e., better) mean ranking than the 30 km for both WoFS initialization times; meanwhile, at 3h lead times, the 30km received a slightly better ranking than the 15km. While participants generally ranked both the 7.5 and 39km poorly, they tended to prefer the 7.5km to the 39km radius at 1h lead times and the 39km at later lead times (Fig. 58). These results are unsurprising, since later lead times are associated with greater uncertainty of storm placement and intensity. More surprising was that participants' rankings did not seem to vary much between the early and late WoFS initialization times (Fig. 58). One possible explanation for these results was that both the early and late WoFS initialization times tended to be at or after storm initiation, resulting in similar underlying degrees of uncertainty in storm placement and intensity.

Participants were also asked to provide feedback on how the WoFS-PHI 5-minute updates impacted the forecast. Approximately 60% of participants felt the updates made the forecasts somewhat or much better, and another 30% of participants thought the updated and older forecasts were about the same. These results suggest that the rapid

5-minute WoFS-PHI updates would be at least somewhat useful and would very rarely degrade the forecast.

When asked to provide additional written feedback, many participants stated they liked the spatial precision of WoFS-PHI and mentioned they found the product most useful when storms were just beginning to initiate. However, many participants also expressed a desire to see higher-magnitude probabilities, especially for the smaller radii.

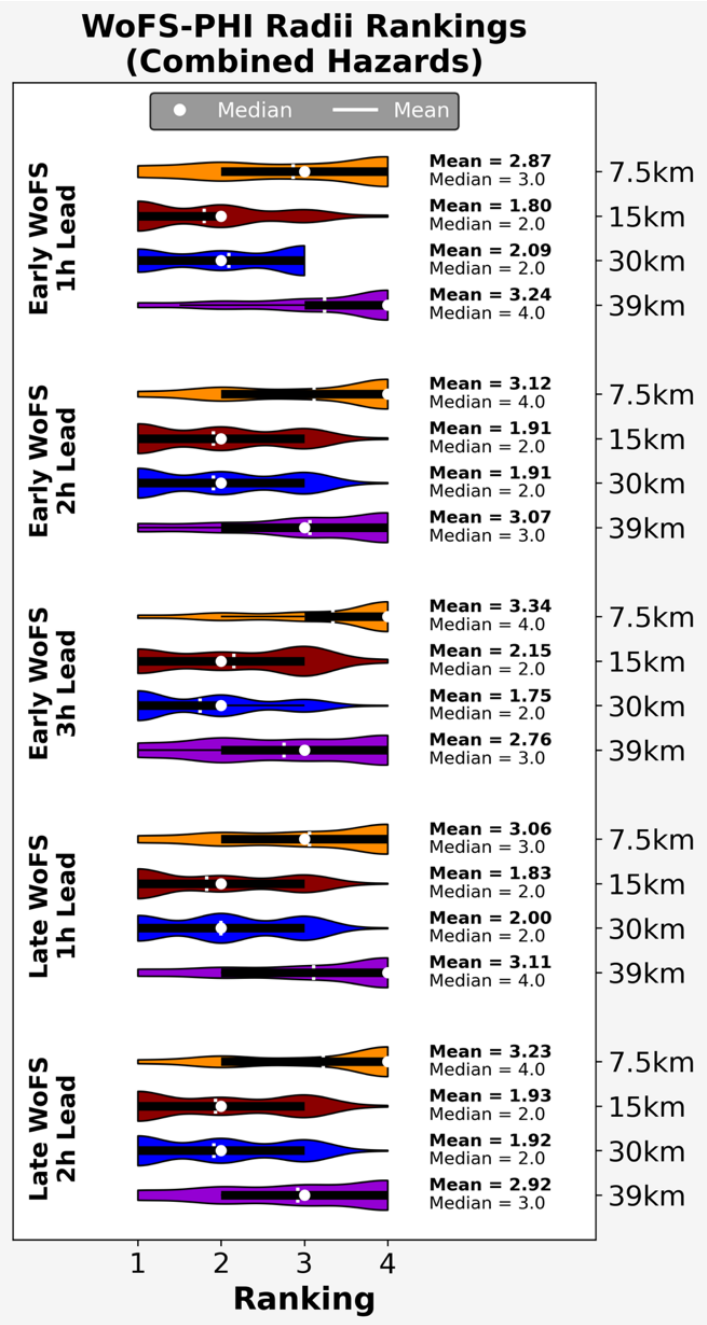

Figure 58. Violin plots of rankings of 7.5 km (gold), 15 km (red), 30 km (blue), and 39 km (purple) radii from WoFS-PHI spatial hazard probabilities for different sets of early and late WoFS initialization times and forecast lead times. Rankings from all hazards are aggregated. Lower rankings (i.e., smaller numbers are more favorable.

3.6 (O)utlook Evaluations and Mesoscale Discussions (MDs)

3.6.1 (O1) Day 1/2/3/4 Outlooks

In this evaluation, the experimental Day 1-3 outlooks for tornado, wind, and hail, and Day 4 outlook for total severe produced by SFE teams were subjectively rated and compared. Generally, average ratings decreased with increasing lead time, as expected (Fig. 59). The Day 4 probabilities were rated only slightly lower, on average, than the Day 3 individual hazard outlooks, suggesting that generating Day 4 outlooks similarly to Day 3 may be operationally feasible.

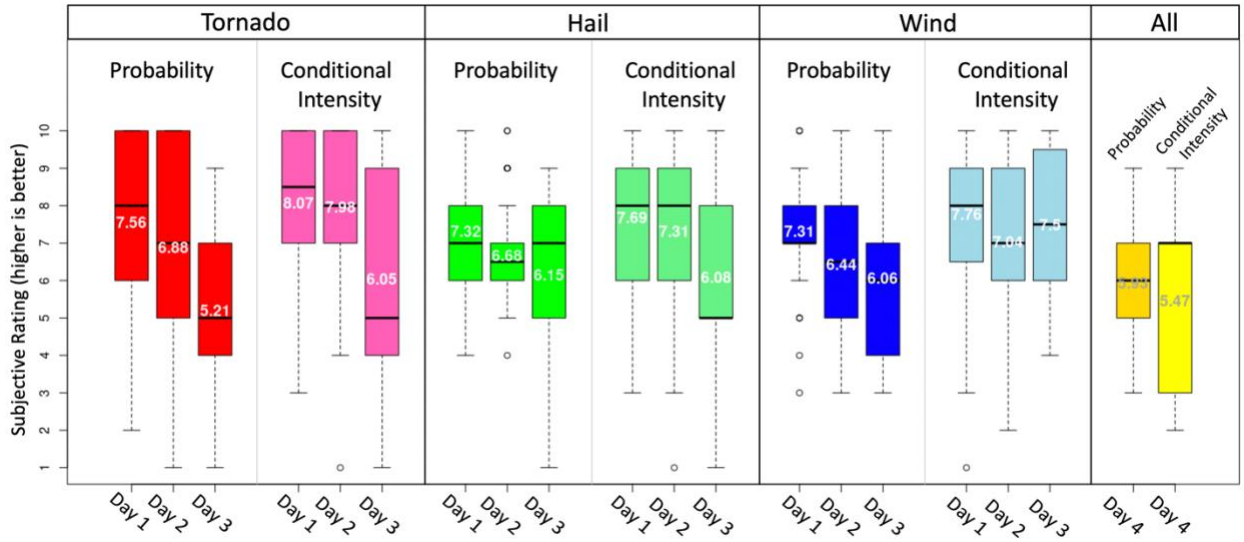

Figure 59. Boxplots depicting the distributions of subjective ratings assigned to the Days 1-3 Probability and Conditional Intensity outlooks for tornado (red), hail (green), and wind (blue); as well as Day 4 all hazards probability and conditional intensity (yellow) outlooks.

3.6.2 (O2) Day 1 Outlook Update (w/ WoFS)

In this evaluation, SFE participants were asked to, "*Subjectively compare the Day 1 Outlook Update (Forecaster 1) to the Day 1 Outlook issued by the group in the morning from much worse to much better*". The probability and conditional intensity outlooks for tornado, hail, and wind were compared. For the probability outlooks, the hail and wind outlook updates were most frequently rated "*about the same*" or "*slightly better*", while the tornado outlooks were dominated by "*about the same*" responses. For the conditional intensity outlooks, both tornado and wind outlooks were dominated by "*about the same*" responses, while for hail there was a slight tendency toward "*slightly better*" (Fig. 60).

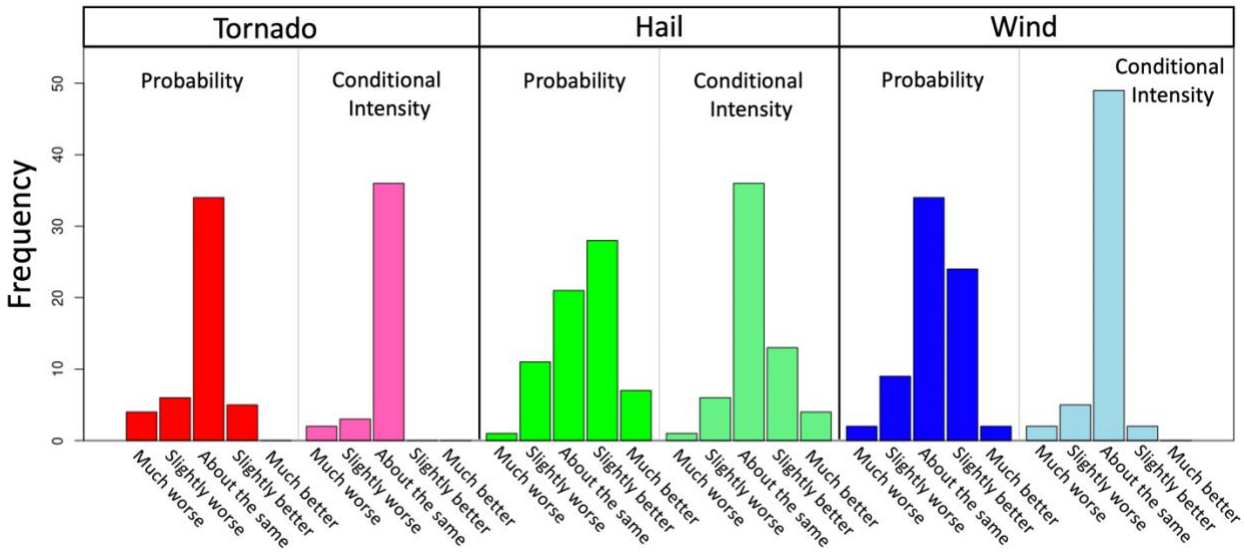

Figure 60. Response frequencies for the O2 evaluation comparing Day 1 Outlook updates to earlier group issued Day 1 outlooks.

3.6.3 (O3) SPC Impacts System: Day 1 Outlook Tornado Counts and Impacts

The SPC Impacts System was run on the Day 1 tornado outlooks with conditional intensity information to estimate the number of tornadoes by EF scale and the potential societal impacts. This tool includes information on population impacted by tornadoes of different intensities, number of schools and mobile homes potentially affected by any tornadoes, and EF2+ tornadoes, and estimates of casualties. Participants were asked to, "*Discuss whether the quantitative impact estimates are consistent and aligned with your expectation of potential tornado weather impacts for the day*", and then there was an optional comment box in which participants were asked to "*Comment on the visualization aspects of displaying the quantitative-impact information and offer any suggestions for improvement*".

Given the below average tornado numbers that occurred during SFE 2023, this evaluation was dominated by null events. There was one localized event that occurred 11 May 2023 in which two EF1 tornadoes affected central Oklahoma. For this event, one participant commented, "... *it seems to show expectation of 6 tornadoes and 1 of then significant. It seems to show 82 people would be affected? That actually seems like not a bad estimate of what occurred in Oklahoma*". For some of the null events, comments such as, "*Low probability of tornado impacts are captured well*", were common, as well as comments like, "*Not quite sure how to evaluate this product*".

3.6.4 (MD-R2O) R2O Group MD Activities

As part of the afternoon forecasting activities on the R2O Desk, experimental mesoscale discussions (MDs) were generated during the 2023 HWT SFE. These MDs

were generated daily in Google Slides (example provided in Fig. 61) by all R2O Group participants from 2:15-3:00 p.m. CDT covering a limited-area domain with the greatest severe potential across the CONUS. There were two items of emphasis on these experimental MDs in utilizing WoFS during the watch-to-warning time frame: 1) focus on a meso-beta corridor with the greatest potential for severe weather over the next few hours and 2) estimate the expected peak intensity of tornado, hail, and convective winds within that corridor. SPC forecasters developed a matrix of overlapping peak intensity bins for participants to select from in making these intensity forecasts for the MDs. Subjective evaluation of these intensity forecasts on the following day generally revealed skill in selecting the appropriate intensity bin. More rigorous evaluation will be done to assess the feasibility of implementing this approach in operational MDs at SPC.

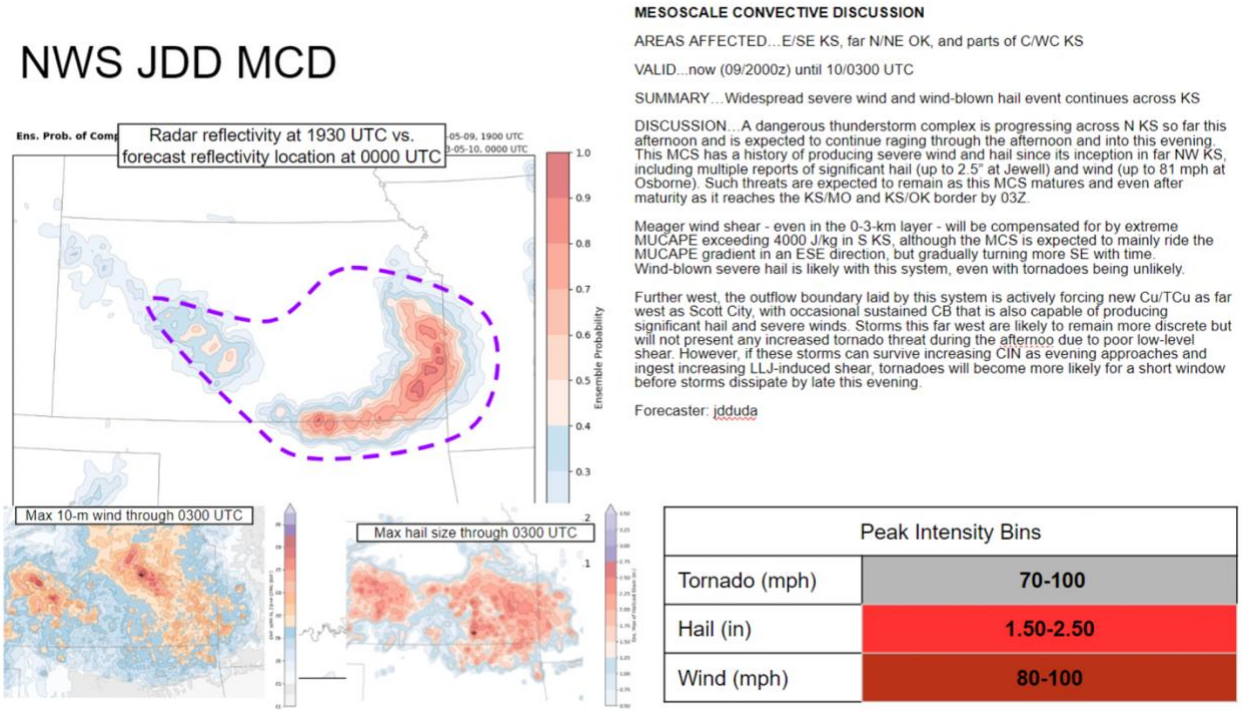

Figure 61. Example of an experimental MD created on 9 May 2023 using WoFS output. The table in the bottom right indicates the forecast of the peak intensity expected for tornadoes, hail, and convective wind within the MD area during the valid time.

3.6.5 (MD-Innovation) Innovation and Virtual Group MD Activities

Similar to the R2O desk, the Innovation and Virtual groups engaged in an afternoon mesoscale discussion activity. In this activity, each participant issued a forecast consisting of a geographic threat area, and a text discussion. The threat area was created using the WoFS web viewer drawing tool and took one of three formats: (1) A single contour highlighting a region of expected severe weather along the track of an individual storm, (2) two contours, one encompassing a broader region where severe weather is expected and the second, smaller contour outlining what is perceived as the

corridor of greatest risk, or (3) A single contour highlighting a broader region where severe weather was expected. Each participant issued their first set of predictions during the 2:15-3pm CDT time period, took turns discussing their product from 3-3:15pm CDT, and issued a second set of predictions from 3:15-3:45pm CDT. Finally, from 3:45-4pm each participant participated in a short survey with targeted questions on WoFS products used, changes in forecasts between 1st and 2nd hours, and overall confidence. The activity revealed many different ways in which WoFS guidance could be used in the watch to warning time frame. More rigorous evaluation will be done with the survey responses to assess most frequent products used, forecast changes, and overall confidence. An example outlook is provided in Figure 62.

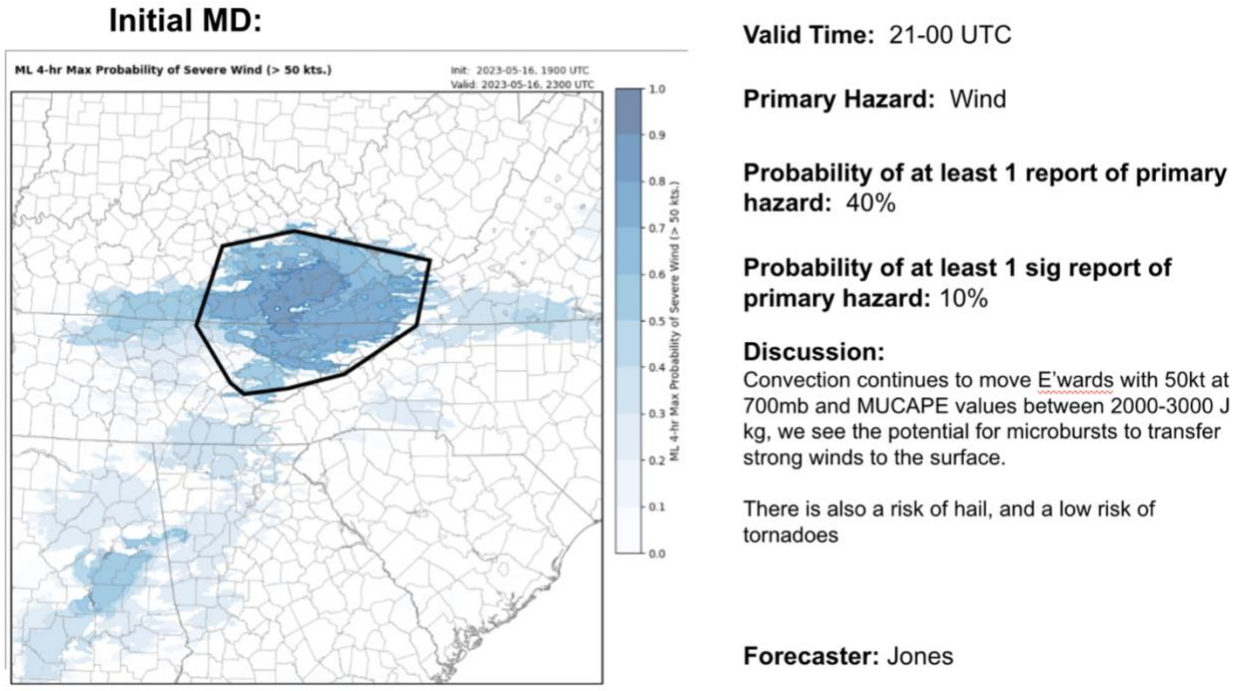

Figure 62. Example of an experimental MD created on 15 May 2023 using WoFS output.

4. Summary

The 2023 NOAA HWT Spring Forecasting Experiment (2023 SFE) was conducted virtually from 1 May – 2 June by the SPC and NSSL with participation from forecasters, researchers, model developers, university faculty, and graduate students from around the world. The primary goals of the 2023 SFE were to (1) evaluate convection-allowing model and ensemble guidance for identifying optimal configurations of convection-allowing versions of FV3 and CAM ensembles, including several carefully designed and controlled experiments as part of the Community Leveraged Unified Ensemble (CLUE), (2) study how forecasters and meteorologists utilize CAMs and CAM ensembles, such as WoFS, and evaluate various experimental severe weather outlooks generated using WoFS and other CAM ensembles for lead times from one hour to 4 days, and (3) evaluate different CAM ensemble post-processed guidance with an emphasis on those using machine-learning algorithms.

Several preliminary findings/accomplishments from the 2023 SFE are listed below:

- WoFS was used for updating full-period hazard forecasts valid 2100-1200 UTC and corresponding conditional intensity guidance, as well as generating experimental mesoscale discussions using WoFS and other CAM guidance.
    - Subjective evaluations indicated that updated hazard probabilities were generally improved for hail and wind, while improvement in the conditional intensity guidance was only noted for hail.
    - One of the most popular activities was generating experimental mesoscale discussions using WoFS and other CAM guidance. This activity provided an opportunity to synthesize a variety of information from the experimental models to generate forecast products first-hand, which often led to an appreciation of the challenges faced by SPC forecasters in generating short-fused forecast products.
- SFE 2023 marked the first time that Day 4 total severe outlooks were issued and the second year for individual hazard outlooks for Day 3. Early indications are that these products could be operationally feasible, and future SFEs will continue to test experimental products at these extended range lead times.
- Examined and assessed various methods to produce first-guess calibrated probabilistic hazard guidance based on forecast output from HREFv3, GEFS, and HRRRv4.
    - For active tornado days at both Day 1 and 2 lead times, the ML algorithm known as “Nadocast” performed best overall along with an ensemble of guidance products, although for the Day 2 lead times differences in mean subjective ratings were smaller relative to Day 1.

- For hail, Nadocast was the top performer at both Day 1 and 2 lead times, and the differences in mean subjective ratings were notable larger at the Day 2 lead times relative to Day 1.
- For wind, the HREF-based ML random forest algorithm performed best for Day 1 and 2 lead times.

- Examined various **deterministic** CAM systems within the CLUE using HRRRv4 as a baseline.
  - In blinded 00Z Day 1 evaluations, HRRRv4 was the clear top performer for simulated reflectivity and UH, 2-m temperature, SBCAPE, and 6-h QPF, while RRFS performed best for 2-m dewpoint.
  - At Day 1, NSSL MPAS RT performed notably better than RRFS for simulated reflectivity and UH, and performed similarly to RRFS for environment and QPF fields.
  - GFDL FV3 and NASA GEOS FV3 were the worst performing flagship models for every blinded evaluation at both Day 1 and 2 lead times. The only exception was that NASA GEOS FV3 performed slightly better than the NAM Nest for 2-m dewpoint at Day 2.
  - In direct comparisons between 0000 UTC initialized RRFS and HRRR, RRFS was on average rated worse than HRRR for reflectivity and UH, updraft speed, 6-h QPF, SBCAPE, 2-m temperature, and 2-m dewpoint. RRFS was rated slightly better than HRRR for 10-m wind speed. For 1200 UTC initializations, HRRR and RRFS performance was more comparable relative to the 0000 UTC initializations, and once again RRFS had the advantage for 10-m wind speed.
  - In direct comparisons between RRFS and HRRR from 2100 and 0000 UTC initializations focused on 0-12 h lead times, the HRRR had superior performance at every time examined at each initialization for reflectivity and UH, 2-m temperature, 2-m dewpoint, and SBCAPE.
  - In comparisons of three MPAS configurations run by NSSL, the one initialized from HRRR that used NSSL microphysics performed best (MPAS HN).
  - In comparisons between a 1-km grid-spacing WRF-ARW configuration (NSSL1) and the HRRR, the NSSL1 did not have an apparent advantage in forecasting convective evolution and tornadoes using 0-2 km AGL UH as a proxy, but the NSSL1 did have a significant advantage in forecasting severe wind using maximum 10-m wind speed as a proxy.
- Examined various ensemble CAM systems within the CLUE using HREFv3 as a baseline.
  - In direct comparisons between 0000 UTC initializations of RRFS and HREF, RRFS was rated slightly worse than HREF for UH and updraft speed, while 10-m wind speed and composite reflectivity were rated similarly. RRFS was

rated slightly worse than HREF for 2-m temperature, 2-m dewpoint, and SBCAPE.

- In comparisons of various Day 1 forecasts from 1200 UTC initialized RRFS configurations that included mixed-physics and time-lagging, mean subjective ratings were tightly clustered, but HREF still had the highest mean ratings. Mixed-physics, time-lagging, and combinations of mixed-physics and time-lagging did not result in any improvement relative to the single-physics RRFS.
- In comparisons of Day 2 forecasts from 1200 UTC initialized RRFS configurations, HREF significantly outperformed the all of the RRFS configurations. Similar to Day 1, mixedphysics, time-lagging, and combinations of the two strategies did not result in any improvement relative to the single physics RRFS for Day 2.
- In comparisons of NCAR FV3 and NCAR MPAS ensembles at Day 3-5 lead times, the MPAS ensemble received significantly higher mean subjective ratings. At times, forecast value was noted in these CAM ensembles all the way to Day 7, which was the longest lead time examined.

- Various other projects and products were assessed and evaluated related to severe weather prediction, including machine-learning approaches for severe wind and convective mode probabilities, mesoscale and storm-scale analyses, and global ensemble forecasts for severe weather applications.
  - Two machine-learning-based algorithms were used to diagnose the likelihood that severe wind reports were actually associated with winds ≥ 50 knots. The algorithm that used a stack generalized linear model (GLM) received slightly higher mean subjective ratings than a gradient-boosted machine (GBM) algorithm.
  - Three algorithms for producing extended-range total severe forecasts based on GEFS for Days 3-7 were examined. The GEFS Operational ML algorithm was the clear top performer relative to GEFS Reforecast ML and GEFS Reforecast Cal.
  - A neural-network algorithm was trained using predictors from the operational HRRR to produce tornado, wind, and hail probabilities. Two versions were tested: one that used convective mode information and one that did not. The version with convective mode information received slightly higher ratings for tornado and wind, while there was little difference for hail.
  - Two versions of 3D-RTMA with HRRR and RRFS backgrounds were evaluated. The RRFS background was most frequently rated "*about the same*" or "*slightly worse*" relative to the HRRR version. Additionally, the two versions of RTMA and SPC mesoanalysis were ranked from best to worst. 3D-RTMA HRRR received the best rankings, followed by SPC mesoanalysis and 3D-RTMA RRFS.

- 15-minute maximum forecasts of 80-m winds, 2-5 km AGL UH, and column-maximum updraft speed from WoFS were used as a proxy for the analysis of severe weather. Overall, the WoFS ensemble analysis fields were positively viewed in terms of lining up with radar-derived proxies of severe weather, preliminary local storm reports, and a subjective assessment of severe weather based on the environment.
- A random forest ML algorithm called WoFS-PHI was used to combine information from ProbSevere Version 2 and WoFS to produce spatial hazard probabilities at 0-3 h lead times. In comparisons of 7.5-, 15-, 30-, and 39-km radii used to generate the probabilities, participants generally favored the 15- and 30-km radii. Participants generally liked the spatial precision of WoFS-PHI and found the product most useful when storms were just beginning to initiate. However, participants wanted to see higher-magnitude probabilities, especially for smaller radii.

Overall, the 2023 SFE was successful in testing new forecast products and modeling systems to address relevant issues related to the prediction of hazardous convective weather. The findings and questions generated during the 2023 SFE directly promote continued progress to improve forecasting of severe weather in support of the NWS Weather-Ready Nation initiative. In subsequent years, we plan to continue exploring the potential forecasting applications of Warn-on-Forecast, continue examining strategies for CAM ensemble design, accelerate work with our partners to optimize the UFS for CAM forecasting applications, and explore new ways to leverage AI/ML-based strategies for calibrating and post-processing CAM output to aid forecasters. Additionally, we expect that this work will take on particular importance and assist with evidence-based decision making as NOAA moves forward with its plans for a Unified Forecasting System. SFE 2023 marked a return to in-person participation and was the first hybrid experiment (i.e., both in-person and virtual participation). We plan to continue with a hybrid format in subsequent experiments, as having in-person participation is much more conducive to science-based discussions and establishing new collaborations, while virtual participation enables people to participate that are unable to attend in-person, which expands the SFE scope.

Acknowledgements

The 2023 SFE would not have been possible without dedicated participants and the support and assistance of numerous individuals at SPC and NSSL. In addition, collaborations with NCAR, GSL, GFDL, NASA, and EMC were vital to the success of the 2023 SFE. In particular, Ryan Sobash (NCAR), Craig Schwartz (NCAR), David John Gagne (NCAR), Dave Ahijevych (NCAR), Charlie Becker (NCAR), Gabrielle Gantos (NCAR), Curtis Alexander (GSL), David Dowell (GSL), Christina Holt (GSL), Chris Harrop (GSL), Steve Weygandt (GSL), Terra Ladwig (GSL), Amanda Back (GSL), Guoqing Ge (GSL), Craig Hartsough (GSL), Ming Hu (GSL), Chunhua Zhou (GSL), Trevor Alcott (GSL), Jeff Beck (GSL), Jaymes Kenyon (GSL), Bob Lipschutz (GSL), Haidao Lin (GSL), Jacob Carley (EMC), Jili Dong (SAIC/EMC), Matt Pyle (EMC), Ben Blake (Lynker/EMC), Eric Aligo (SAIC/EMC), Xiaoyan Zhang (SAIC/EMC), Ting Lei (Lynker/EMC), Shun Liu (EMC), Manuel Pondeca (Lynker/EMC), Edward Colon (Lynker/EMC), Matthew Morris (SAIC/EMC), Gang Zhao (SAIC/EMC), Annette Gibbs (Lynker/EMC), Sho Yokota (JMA visitor at EMC), Donnie Lippi (Lynker/EMC), Andrew Benjamin (EMC), Kai-Yuan Cheng (GFDL), Lucas Harris (GFDL), Matthew Morin (GFDL), Linjiong Zhou (GFDL), William Putman (NASA), and Scott Rabenhorst (NASA) were essential in generating and providing access to model forecasts or products examined on a daily basis.

References

Clark, A. J. and Coauthors, 2018: The Community Leveraged Unified Ensemble (CLUE) in the 2016 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Bull. Amer. Meteor. Soc*., **99**, 1433-1448.

Hill, A. J., R. S. Schumacher, and I. L. Jirak, 2023: A new paradigm for medium-range severe weather forecasts: Probabilistic random-forest-based predictions. *Wea. Forecasting*, **38**, 251-272.

Karstens, C. D., R. Clark III, I. L. Jirak, P. T. Marsh, R. Schneider, and S. J. Weiss, 2019: Enhancements to Storm Prediction Center convective outlooks. Ninth Conf. on Transition of Research to Operations, Phoenix, AZ, Amer. Meteor. Soc., J7.3, https://ams.confex.com/ams/2019Annual/webprogram/Paper355037.html.

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

Smith, T. M., V. Lakshmanan, G. J. Stumpf, K. L. Ortega, K. Hondl, K. Cooper, K. M. Calhoun, D. M. Kingfield, K. L. Manross, R. Toomey, and J. Brogden, 2016: Multi-Radar Multi-Sensor (MRMS) Severe Weather and Aviation Products: Initial Operating Capabilities. *Bull. Amer. Meteor. Soc*., **97**, 1617-1630.

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

APPENDIX

Time (CDT)	
8:00 AM – 8:45 AM	*(Optional) Map Analysis, Data Loading, and Networking* *In-Person (Optional)*
8:45 AM – 9:00 AM	**Overview of Yesterday's Severe Weather** *Hybrid All* (David Imy)
9:00 AM – 10:30 AM	**Model & Outlook Evaluation** (Orientation, Surveys, and Discussion) *Hybrid Groups* (Group 1; Group 2; Group 3)
10:30 AM – 10:45 AM	***Break***
10:45 AM – 11:00 AM	***Evaluation Highlights*** *Hybrid All* (Group 1; Group 2; Group 3)
11:00 AM – 11:15 AM	**Weather Briefing** *Hybrid All* (David Imy)
11:15 AM – 12:30 PM	**Group Forecasting Activity** (Coverage and Conditional Intensity Outlooks) *In-Person R2O* (Day 1); *In-Person Innovation* (Days 3 & 4); *Virtual* (Day 2)
12:30 PM – 2:00 PM	***Lunch/Break*** ***Science Discussion (Wednesdays @ 1:15)***
2:00 PM – 2:15 PM	**Update on Today's Weather** *Hybrid All* (David Imy)
2:15 PM – 3:15 PM	**Individual Forecasting Activity** (Mesoscale Discussions and Discussion) *In-Person R2O* (Meso-beta MD); *In-Person Innovation* (WoFS MD); *Virtual* (WoFS MD)
3:15 PM – 4:00 PM	**Individual Forecasting Activity Continued** (MD & Day 1 Updates) *In-Person R2O* (Day 1 Update); *In-Person Innovation* (WoFS MD); *Virtual* (WoFS MD)

Table 2. Schedule for Tuesday – Friday. On Monday, the schedule is similar except the period 9-11am is devoted to training and introductory material.