# SPRING FORECASTING EXPERIMENT 2021

## Conducted by the

# EXPERIMENTAL FORECAST PROGRAM

## of the

# NOAA HAZARDOUS WEATHER TESTBED

https://hwt.nssl.noaa.gov/sfe/2021

**Virtual Experiment**
**3 May – 4 June 2021**

# Preliminary Findings and Results

Adam Clark[2,4], Israel Jirak[1], Burkely T. Gallo[1,3], Kent Knopfmeier[2,3], Brett Roberts[1,2,3], Makenzie Krocak[1,3,5], Jake Vancil[1,3], Kimberly Hoogewind[2,3], Nathan Dahl[1,3], Eric D. Loken[2,3,4], David Jahn[1,3], David Harrison[1,3], Dave Imy[2], Patrick Burke[2], Louis Wicker[2,4], Patrick Skinner[2,3], Pam Heinselman[2,4], Patrick Marsh[1], Katie Wilson[2,3], Andy Dean[1], Gerry Creager[2,3], Thomas Jones[2,3], Jidong Gao[1], Yunheng Wang[2,3], Montgomery Flora[2,3], Corey Potvin[2,4], Chris Kerr[2,3], Nusrat Yussouf[2,3], Joshua Martin[2,3], Jorge Guerra[2,3], and Brian Matilla[2,3], and Thomas J. Galarneau[2,3,4]

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

Table of Contents

Scenes and participant screenshots from each week of the 2021 NOAA Hazardous Weather Testbed Spring Forecasting Experiment

## 1. Introduction

The 2021 Spring Forecasting Experiment (2021 SFE) was conducted from 3 May – 4 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 collaborators including the NOAA Global Systems Laboratory (GSL), NOAA/NCEP Environmental Modeling Center (EMC), NOAA Geophysical Fluid Dynamics Laboratory (GFDL), National Center for Atmospheric Research (NCAR), and the Multi-scale data Assimilation and Predictability (MAP) group at the University of Oklahoma. Participants included over 130 forecasters, researchers, model developers, university faculty, and graduate students from around the world (see Table A1 in the Appendix). Because of the COVID-19 pandemic, restrictions on travel and gatherings precluded an in-person experiment for the second consecutive year. However, to maintain momentum in key areas of convection-allowing model development, the HWT EFP once again conducted the 2021 SFE virtually, and expanded upon the virtual activities of the prior year. As in previous years, the 2021 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:

Product and Service Improvements:

- Assess the utility of a prototype Warn-on-Forecast system (WoFS) by issuing 1-h time window outlooks for individual severe hazards (tornado, hail, and wind) with and without access to WoFS.
- Test the utility of WoFS for updating full period hazards forecasts valid 2100-1200 UTC.
- Explore the ability to provide enhanced information on the conditional intensity of tornado, wind, and hail events by delineating areas expected to follow “normal”, “hatched”, or “double-hatched” intensity distributions in Convective Outlooks covering Days 1 & 2.
- Explore how WoFS and other CAMs can be used in watch-to-warning scale forecasting applications with an activity focused on using this guidance for generating Mesoscale Discussions (MDs).
- Quantify the value of CAM guidance for generating Day 2 Convective Outlooks with individual hazard probabilities by issuing this product with and without CAMs.

Applied Science Activities:

- Compare various CAM ensemble prediction systems to identify strengths and weaknesses of different configuration strategies. Most of these comparisons were conducted within the framework of the Community Leveraged Unified Ensemble discussed below. Additional baseline comparisons were made using the High-Resolution Ensemble Forecast System version 3 (HREFv3).
- Compare and assess different machine-learning approaches for estimating the likelihood of wind damage reports being associated with gusts ≥ 50 knots.
- Compare and assess two machine-learning techniques for producing probabilistic convective mode guidance from deterministic 3-km grid-spacing CAMs.
- Evaluate configurations of the limited area Finite Volume Cubed Sphere Model (FV3-LAM) with different data assimilation (DA), physics suites, domain sizes, and stochastic physics.

- Compare and assess different versions of the 3D real-time mesoscale analysis (3D-RTMA) system that use different sources of background first guesses and DA frequencies, and test WoFS-based analyses of 10-m and 80-m as a potential verification source for severe winds.
- To assess the possible impact of retiring the Short-Range Ensemble Forecast system (SREF), evaluate ensemble forecasts of environmental parameters, as well as calibrated thunder and severe probabilities, for the Global Ensemble Forecast System (GEFS) and SREF at Days 2 & 3 lead times.
- Evaluate the utility of several methods, including machine-learning approaches, for producing calibrated hazard guidance.
- Compare and assess the skill and utility of the primary deterministic CAMs provided by each SFE 2021 collaborator.
- Evaluate WoFS for applications to short-term severe weather product generation, and explore the potential value provided by an enhanced resolution (1.5 km grid-spacing) deterministic configuration of WoFS that uses hybrid DA, a machine-learning approach to generate calibrated guidance from WoFS, and the application of an objective verification scorecard comparison between WoFS and a time-lagged HRRR ensemble.

A suite of state-of-the-art experimental CAM guidance contributed by our large group of collaborators was critical to the 2021 SFE. For the sixth consecutive year, these contributions were formally coordinated into a single ensemble framework called the Community Leveraged Unified Ensemble (CLUE; Clark et al. 2018). The 2021 CLUE was constructed by having all groups coordinate as closely as possible on model specifications (e.g., grid-spacing, vertical levels, physics, etc.), domain, and post-processing so that the simulations contributed by each group could be used in controlled experiments. This design allowed us to conduct several experiments to aid in identifying optimal configuration strategies for CAM-based ensembles. The 2021 CLUE included 64 members using 3-km grid-spacing that allowed for several unique experiments. The 2021 SFE activities also involved testing the WoFS for the fifth consecutive year.

This document summarizes the activities, core interests, and preliminary findings of the 2021 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/2020/docs/HWT_SFE2021_operations_plan.pdf). The remainder of this document is organized as follows: Section 2 provides an overview of the models and ensembles examined during the 2021 SFE along with a description of the daily activities, Section 3 reviews the preliminary findings of the 2021 SFE, and Section 4 contains a summary of these findings and some directions for future work.

## 2. Description

### *a) Experimental Models and Ensembles*

A total of 94 unique CAMs were run for the 2021 SFE, of which 64 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 2021 SFE emphasized the use of CAM ensembles (i.e., HREF, HRRRE, and WoFS) in generating experimental probabilistic forecasts of individual severe weather hazards. Additionally, the 2021 CLUE configuration enabled numerous scientific evaluations focusing on model sensitivities and various ensemble configuration strategies.

To put the volume of CAMs run for 2021 SFE into context, Figure 1 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 1 shows an increasing trend. The consolidation of members into the CLUE has made this increase more manageable and facilitated more controlled scientific comparisons.

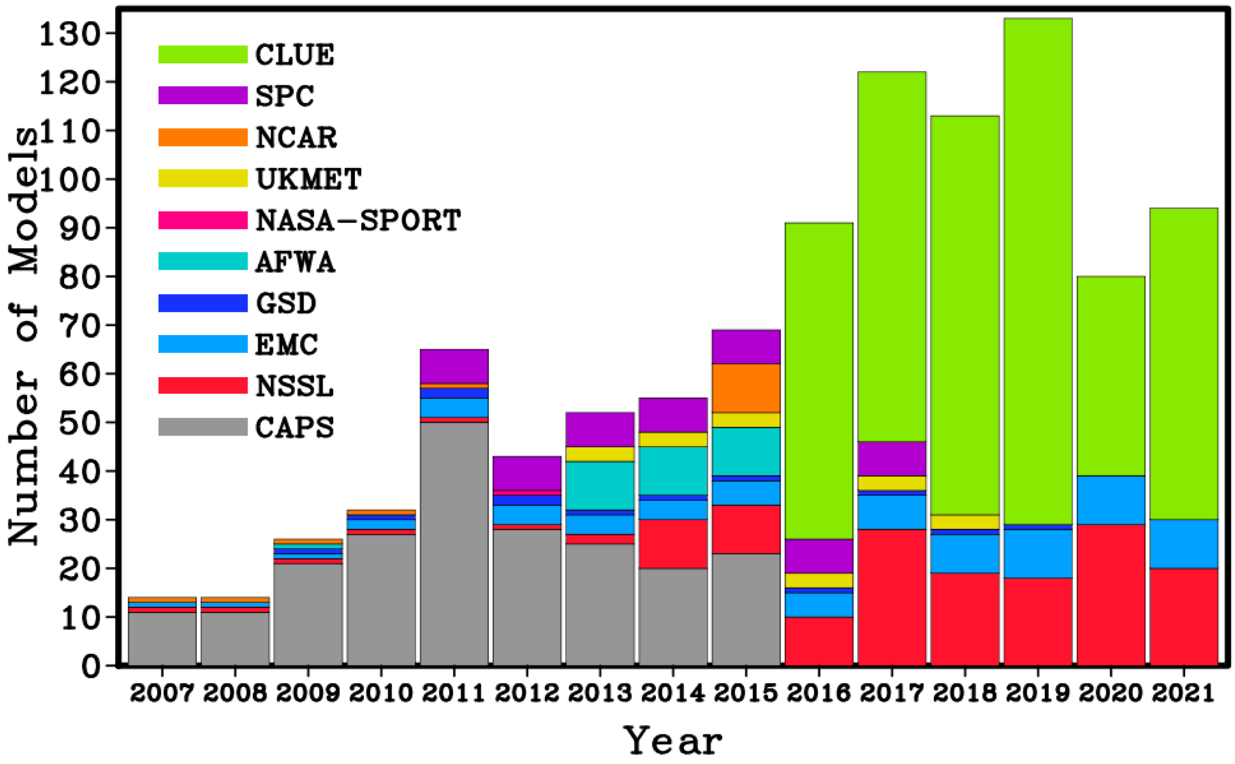


*Figure 1 Number of CAMs run for SFEs since 2007. The different colored stacked bars indicate the contributing agencies.*

More information on all of the modeling systems run for the 2021 SFE is given below.

## 1) THE COMMUNITY LEVERAGED UNIFIED ENSEMBLE (CLUE)

The 2021 CLUE is a carefully designed ensemble with subsets of members contributed by NOAA groups at NSSL, GFDL, GSL, and EMC, and the non-NOAA group of OU-MAP. CLUE members have 3-km grid-spacing and either a CONUS or North America domain. Depending on the CLUE subset, forecast lengths range from 18 to 126 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 and further details on ensemble configurations are provided in the 2021 SFE operations plan. Table 1 provides a summary of each CLUE subset.

*Table 1 Summary of the 14 unique subsets that comprise the 2021 CLUE.*

| Clue Subset | # of mems | IC/LBC perts | Mixed Physics | Data Assimilation | Model Core | Agency | Init. Times (UTC) | Forecast Length (h) | Domain |
|---|---|---|---|---|---|---|---|---|---|
| GSL RRFS | 9 | HRRRDAS/ GEFS | no | EnKF | FV3 | GSL | 00, 12 | 60, 48 | CONUS |
| HRRRE-S | 9 | HRRRDAS/ GEFS | no | EnKF | ARW | GSL | 12 | 24 | CONUS |
| HRRRE-M | 9 | HRRRDAS/ GEFS | yes | EnKF | ARW | GSL | 12 | 24 | CONUS |
| GSL FV3-LAM | 1 | none | no | Hybrid 3DEnVar (GDAS Ensemble) | FV3 | GSL | 00-23 (hourly) | 20x18h, 4x48h | CONUS |
| GSL FV3-LAM-NA | 1 | none | no | cold start from GFS | FV3 | GSL | 00, 12 | 60, 60 | N. America |
| EMC FV3-LAM | 1 | none | no | cold start from GFS | FV3 | EMC | 00, 12 | 60, 60 | CONUS |
| EMC FV3-LAMX | 1 | none | no | cold start from GFS | FV3 | EMC | 00, 12 | 60, 60 | N. America |
| EMC FV3-LAMDAX | 1 | none | no | Hybrid 3DEnVar (GDAS EnKF) | FV3 | EMC | 00, 12 | 60, 60 | CONUS |
| HRRRv4 | 1 | none | no | GSI-EnVar | ARW | EMC | 00-23 (hourly) | 20x18h, 4x48h | CONUS |
| RRFS Cloud | 9 | GFS, GEFS | yes | cold start from GFS, GEFS | FV3 | EMC/GSL | 00 | 60 | N. America |
| MAP RRFS | 10 | GFS, GEFS | no | GSI-EnVar | FV3 | OU-MAP | 21, 00 | 39, 36 | CONUS |
| MAP RRFS VTS | 10 | GFS, GEFS | no | GSI-EnVar | FV3 | OU-MAP | 21, 00 | 39, 36 | CONUS |
| NSSL FV3-LAM | 1 | none | no | cold start from GFS | FV3 | NSSL | 00 | 60 | CONUS |
| GFDL FV3 | 1 | none | no | cold start from GFS | FV3 | GFDL | 00 | 126 | CONUS |

The design of the 2021 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 in Table 2.

*Table 2 List of CLUE experiments for the 2021 SFE. The CLUE subsets listed are from Table 1.*

| **Experiment Name** | **Description** | **CLUE subsets** |
|---|---|---|
| RRFS Configuration Strategies | Several different ensembles contributed by GSL, EMC, & OU-MAP were evaluated against HREFv3. **Goal: Identify a strategy within the UFS framework (i.e., single-model, FV3-LAM) that performs as good as or better than HREFv3, so that it can serve as a replacement in NCEP's production suite.** | GSL RRFS, RRFS Cloud, MAP RRFS, & MAP RRFS VTS |
| FV3-LAM Configurations | GSL, NSSL, EMC, and GFDL ran various configurations of FV3. **Goal: Assess the impact of FV3-LAM configuration aspects including physics, data assimilation, initial conditions, domain, and stochastic physics.** | GSL FV3-LAM, GSL FV3-LAM-NA, EMC FV3-LAM, EMC FV3-LAMX, EMC FV3-LAMDAX, RRFS Cloud, NSSL FV3-LAM, & GFDL FV3 |
| Day 2 FV3-LAM performance | FV3-LAM configurations contributed by GSL and EMC were evaluated for the Day 2 forecast period (i.e., forecast hours 36-60). **Goal: Assess model performance at Day 2 lead times and evaluate whether expanded North American domains provide any performance improvements relative to smaller CONUS domains at these lead times.** | EMC FV3-LAM, EMC FV3-LAMX, GSL FV3-LAM, and GSL FV3-LAM-NA |
| Valid Time Shifting Data Assimilation | The OU MAP group ran ensembles with and without Valid Time Shifting (VTS), which is a cost-effective data assimilation approach that increases the membership (by a factor of three) for the background ensemble in convective scale, hybrid EnVar data assimilation. **Goal: Assess whether VTS provides value at 0-12 h lead times in both ensemble and deterministic forecasts.** | MAP RRFS & MAP RRFS VTS |
| CAM Ensemble Physics | Two configurations of the 12Z HRRRE were compared. One used a single physics scheme with SPP and SPPT stochastic perturbations (HRRRE-S), while the other replaces four of the members from HRRRE-S with a different physics configuration based on the NSSL-WRF (HRRRE-M). **Goal: Assess the role of a mixed-physics approach for increasing the spread and diversity of CAM ensemble forecasts.** | HRRRE-S & HRRRE-M |
| 3D-RTMA Background | Two hourly versions of 3D-RTMA were compared. One version from EMC used the operational HRRRv4 as the background while the other from GSL used the FV3-LAM as background. **Goal: Assess the impact of the background first guess on the final analysis.** | HRRRv4 & FV3-LAM |

### 2) 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 operational implementation. 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 few years.

### 3) NSSL EXPERIMENTAL WARN-ON-FORECAST SYSTEM

The NSSL Warn-on-Forecast System (WoFS) is a rapidly-updating 36-member, 3-km grid-spacing WRF-based ensemble data assimilation and forecast system. The WoFS is cycled every 15 minutes with forecasts initialized every 30 minutes and produces very short-range (0-6 h) probabilistic forecasts of individual thunderstorms and their associated hazards. In addition, a dual-resolution hybrid data assimilation and forecast system, WoFS-Hybrid was used to produce a single 1.5-km deterministic forecast. The 900-km x 900-km daily WoFS domain targeted the primary region where severe weather was anticipated.

The starting point for each day's experiment was the operational High-Resolution Rapid Refresh Data Assimilation System (HRRRDAS) provided by GSL from NCO and the HRRRE-S provided by GSL. A 1-h forecast from the 1400 UTC, 36-member, hourly-cycled HRRRDAS analysis provides the initial conditions for both the WoFS and WoFS-Hybrid. Boundary conditions were provided by 1200 UTC HRRRE-S forecasts, initialized from the 1200 UTC HRRRDAS analysis and valid for the period 1500 UTC Day 1 – 0300 UTC Day 2. All WoFS forecasts were made available via the WoFS Forecast Viewer at https://wof.nssl.noaa.gov/realtime.

## *b) Daily Activities*

SFE 2021 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.

### 1) FORECAST AND MODEL EVALUATIONS

SFE 2021 featured a period of formal evaluations from 9:15-11am CDT Tuesday-Friday, 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 radar reflectivity, local storm reports (LSRs), NWS warnings, and Multi-Radar, Multi-Sensor (MRMS; Smith et al. 2016) estimated hail sizes. Participants were split into Groups A, B, C, and D, and each conducted a separate set of model evaluations. The forecast product evaluations were similar across the groups, but the specific questions were dependent on which forecast products the participants issued, and some of the questions were randomized to reduce participant workload. Participants worked on all the surveys individually, but typically stayed in the virtual meeting where SFE facilitators were available to answer any questions or troubleshoot the model evaluation webpage.

### 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. There were two periods of experimental forecasting activities during SFE 2021. The first occurred from 11:30am – 12:30pm CDT and focused on providing individual hazard guidance, as well as more precise information on the intensity of specific hazards. The second forecasting period occurred from 2:15-4pm CDT and focused on short-term forecasting applications of WoFS. Participants were split into two groups for the forecasting activities: R2O & Innovation.

During the first forecasting period, the R2O group issued Day 1 Outlook hazard probabilities for the period 1800 – 1200 UTC. Within the R2O group, one set of participants used 1200 UTC initialized HREFv3 guidance, while another used 1200 UTC initialized HRRRE guidance. The Innovation Group issued Day 2 Outlook hazard probabilities (1200 – 1200 UTC). Within the Innovation Group, one set of participants used CAM guidance to generate their outlook, while another did not use CAMs. 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. Additionally, both groups generated conditional intensity forecasts, which delineate areas that are expected to follow a "normal", "hatched", or "double-hatched" intensity distribution. In plain language, "normal" refers to a typical severe weather day, where significant severe weather is unlikely, "hatched" areas indicate where significant severe weather is possible, and "double-hatched" areas indicate where high-impact significant severe weather is expected. These forecasts could also be thought of as indicating the proportion of observed reports that are expected to be severe, where going from "normal", to "hatched", to "double-hatched" would indicate an increasing proportion of significant-severe reports (see Fig. A3 of Appendix for more detailed information on each hazard).

During the second forecasting period (2:15-4pm CDT), the R2O group conducted one forecasting activity from 2:15-3pm in which each participant issued their own Mesoscale Discussion (MD) Product using WoFS and other available CAM guidance within the SFE Drawing Tool, followed by a group discussion of the MDs. Then, during the 3-4pm time period, each R2O group participant used WoFS and other available guidance to update the Day 1 individual hazard coverage and conditional intensity forecasts that were issued earlier in the day for the period 2100 – 1200 UTC. Four expert forecasters updated forecasts individually, while the remainder of the non-expert forecasters issued separate outlook updates that were combined into consensus forecasts. Current or retired operational NWS forecasters were considered expert forecasters in the context of this activity, as were facilitator-forecasters.

In the Innovation Group, during the 2:15-4pm CDT time period, participants generated severe hazard probabilities valid over 1-h time windows covering 2200-2300 UTC, 2300-0000 UTC, and 0000-0100 UTC. Two initial forecasts were generated during the 2:15-3:15pm period, which covered the 2200-2300 UTC and 2300-0000 UTC time windows. Then, during the 3:15-4pm period, the 2200-2300 UTC and 2300-0000 UTC periods were updated, and one more outlook covering 0000-0100 UTC was generated.

For both sets of initial and final forecasts, two expert forecasters used all available datasets including WoFS (Forecaster WOF 1 & 2), while two other expert forecasters used all available datasets except for WoFS (Forecaster NOWOF 1 & 2). Additionally, two other groups of non-expert forecasters issued forecasts with and without WoFS similarly to the expert forecasters, which were combined into consensus forecasts (ConWoFS and ConNoWoFS, respectively).

## 3. Preliminary Findings and Results

### *a) Model Evaluations - Group A: Calibrated Guidance*

#### *A1) Calibrated Guidance*

SFE participants evaluated a series of probabilistic severe weather hazard guidance forecasts, including those for tornadoes, severe winds (> 50 kts) and hail (> 1 in.). These guidance products covered the convective day (i.e., 1200-1200 UTC the following day; forecast hours 13-36) in 24-h periods on Day 1 and Day 2, as well as 4-h periods on Day 1. All of the guidance products were based on 0000 UTC model/ensemble runs, except for CAM-based guidance on Day 2, which was based on 1200 UTC runs to cover the full convective day (i.e., f024-f048). Evaluation of guidance products were made relative to preliminary local storm reports (LSRs) that were available the day after an event and a practically perfect hindcast (Hitchens et al. 2013) based on those LSRs, as well as WFO warning information and MRMS MESH data (for hail forecasts). Forecast skill was scored by participants using a scale of 1 to 10 such that a 10 indicated a superior forecast.

#### *i) Day 2 Calibrated Tornado Guidance*

A suite of guidance forecasts valid for a period coincident with the SPC Day 2 Outlook were evaluated and are products of various calibration approaches: HREF/SREF calibrated (HREF/SREF, Jirak et al. 2014), machine-learning (ML) global ensemble forecast system (GEFS) developed by Colorado State University (GEFS CSU; Hill et al. 2020), NCAR ML system based on deterministic 12Z HRRR data (HRRR NCAR; Sobash et al. 2020), and HREF calibrated based on an STP distribution within a 40-km circular radius (STP Cal Circle, Gallo et al. 2018). The products using HREF data were based on 12Z HREFv3.1 while 00Z GEFS data are generated using FV3 run on a global scale.

HREF/SREF and HRRR NCAR guidance scored the highest with mean skill scores of 5.98 and 6.06 respectively (Fig. 2a). Evaluators noted that these two guidance products did well in forecasting no tornadoes for non-event days. It should be noted that both guidance products forecast low probabilities of tornadoes (i.e., less than 2% in the SFE domain) for a majority of days (57% for HREF/SREF and 68% for HRRR NCAR) even though only 42% of days were non-tornadic.

When considering only days for which at least one tornado was observed in the SFE domain (58% of days, Fig. 2b), STP Cal Circle was evaluated with the highest mean skill score (5.62), all other guidance forecasts were given mean skill scores less than 4.6. Participants noted that STP Cal Circle did well in forecasting regions of tornado activity, but also with a relatively high FAR. HREF/SREF and HRRR NCAR did best in forecasting non-events (relatively wide violin plot for high-end skill in Fig. 2a) but did not score as well on tornadic days (wide violin plot for low-end skill in Fig. 2b) primarily because they forecast low (i.e., <2%) probabilities on 36% of tornadic days. GEFS CSU forecasts did not perform better than any other method when considering either all days or only tornadic days, but it was handicapped in terms of model resolution and initialization time compared to the HREF- and HRRR-based guidance.

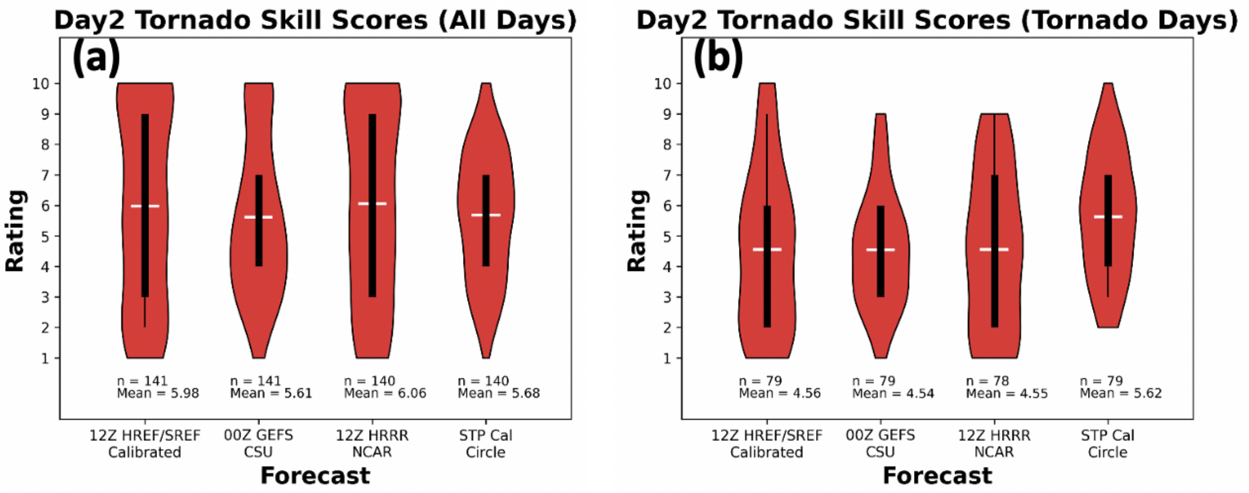


*Figure 2 Violin plots for the given calibrated Day 2 tornado forecast guidance methods showing mean values (short white line) and range of 25th to 75th quartiles (solid black line). Skill scores are aggregated over all 19 SFE evaluation days (left plot) and for 11 days, for which at least one tornado was observed in the SFE domain (right plot).*

*ii) Day 1 Calibrated Tornado Guidance*

Similar trends for Day 1 forecast guidance skill scores are given as for Day 2 in that HREF/SREF and HRRR NCAR are evaluated with a relatively large number of high scores (wide top-end violin plot Fig. 3a) when considering all SFE days, and relatively large number of low scores (wide bottom-end plot Fig. 3b) when considering only tornadic days. As with Day 2 forecasts, HREF/SREF and HRRR NCAR over-forecast non-events. For the set of tornadic days, GEFS CSU performs the best with a mean skill score of 5.39 as compared to the other two methods, which have mean score values less than 4.8.

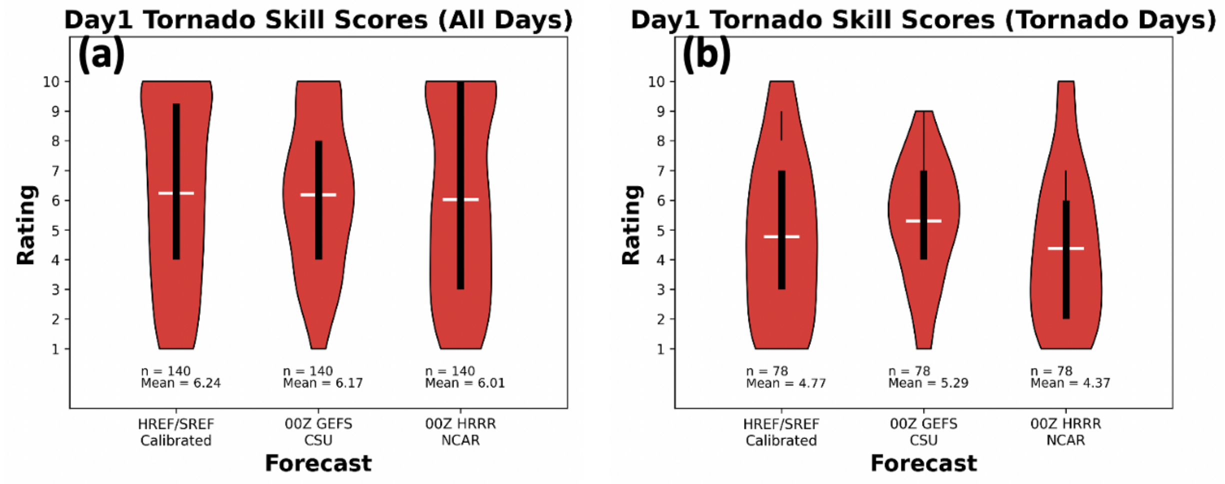


*Figure 3 Same as Figure 2 except for Day 1 tornado forecast guidance methods.*

### *iii) 00Z 24-h HREF Calibrated Tornado Guidance*

Five calibrated tornado forecast guidance methods based on 0000 UTC HREF data were evaluated for the Day 1 period (i.e., f012-f036). Along with the HREF/SREF Calibrated and STP Cal Circle methods, which were also used to generate Day 2 forecasts, two ML methods were evaluated, one based on a random forecast approach developed by E. Loken (RF Loken; Loken et al. 2020), and one on a neural network (NN) approach that was trained using the native HREF grid at 3-km spatial resolution developed by D. Jahn (NN Jahn; Jahn et al. 2020). In addition, an approach highly similar to STP Cal Circle was evaluated (STP Cal Inflow), which considered an STP distribution over an adjacent quadrant corresponding to storm inflow rather than a near-storm circular region. All products used HREFv3 data except for STP Cal Inflow, which used HREFv2.1 (and thus did not include FV3 data, using the HRW-NMMB member prior to 11 May 2021, and having 8 members starting on 11 May 2021 coincident with the operationalization of HREFv3).

Of the five guidance products, none was distinguished as significantly superior when evaluating over all SFE days (Fig. 4a). HREF/SREF and STP Cal Inflow registered respective mean skill scores of 6.28 and 6.11, but the mean scores of the other methods were not much lower with values between 5.82 and 5.94. Considering only tornado days (Fig. 4b), the skill of HREF/SREF was decreased more than the other methods (a consequence of underforecasting tornado events). STP Cal Circle performed the best with a mean skill score of 5.63 with the scores of STP Cal Inflow and RF Loken only slightly lower at 5.54 and 5.44 respectively. In comparing STP Cal methods, evaluators often noted that the Circle method produced larger areas of tornado probabilities than the Inflow method, which tended to increase its POD, but also its FAR. Among the two ML methods, the RF Loken received higher subjective ratings than the NN Jahn especially for tornado days, with respective mean skill scores of 5.44 and 5.04. Overall, there was not a noteworthy difference in the scores among ML methods (e.g., RF Loken and NN Jahn) and more traditional calibration methods (e.g., STP Cal Circle and Inflow).

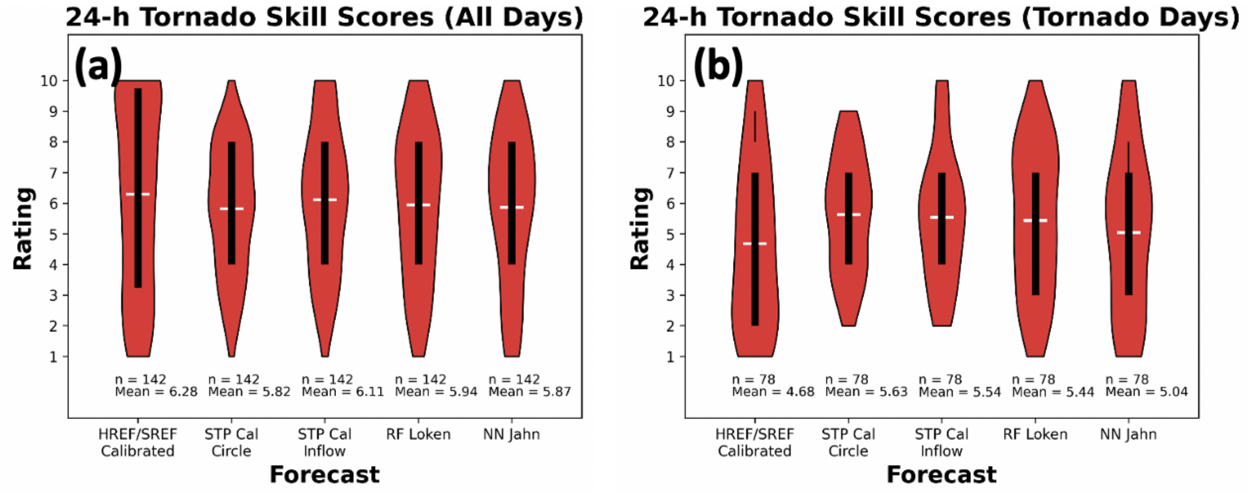


*Figure 4 Same as Figure 2 except for 24-hour tornado forecast guidance methods.*

*iv) 00Z 4-h HREF Calibrated Tornado Guidance*

As with previous tornado guidance products, the HREF/SREF on average scored well across all SFE days (with a mean of 6.34, Fig. 5a), but not so well for only tornado days (mean of 4.92, Fig. 5b). The 06Z and 13Z SPC timing guidance products (Jirak et al. 2020) followed this same trend, primarily because these products are dependent on the HREF/SREF calibrated data in order to time-disaggregate the 24-hour SPC forecast into 4-hour segments. It should be noted that higher ratings (score difference greater than 0.3, Fig. 5) were given to the SPC timing guidance products than the HREF/SREF 4-h guidance indicating the added value in combining SPC (human-produced) and ensemble model data. Further, the 13Z timing guidance was evaluated higher than the 06Z version, which points to the importance of using the latest forecast information for assessing the potential threat. When considering only tornado days, STP Cal Circle performed the best (5.66 mean skill) followed by the 13Z SPC timing guidance (5.53 mean skill).

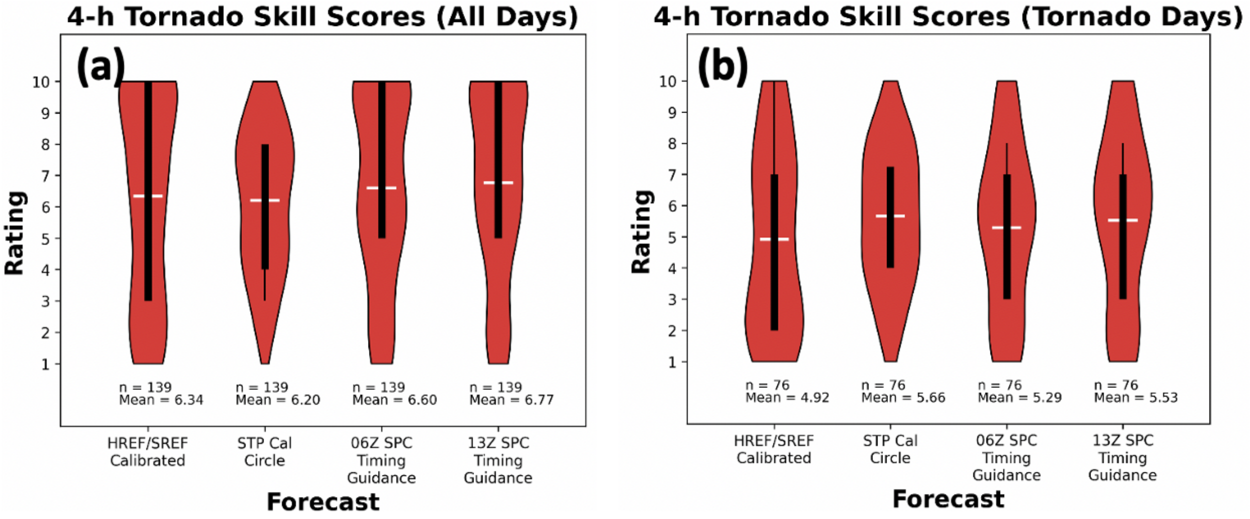


*Figure 5 Same as Figure 2 except for 4-hour tornado forecast guidance methods.*

*v) Day 2 Calibrated Hail Guidance*

Three probabilistic severe hail guidance products were evaluated at Day-2 lead times: HREF/SREF, GEFS CSU, and HRRR NCAR. GEFS CSU received the largest mean subjective rating (5.88) from participants, followed by HRRR NCAR (5.38) and HREF/SREF (4.56; red violins in Fig. 6). Participants noted that GEFS CSU frequently had the best probability of detection (POD), since it often had greater probability magnitudes and areal coverage compared to the other methods. Meanwhile, participants felt that HREF/SREF guidance generally produced probabilities too small in magnitude and areal coverage. Subjective ratings for the HRRR NCAR were more varied; the method received more 10s than GEFS CSU but also more 1s, 2s, and 3s (Fig. 6).

*vi) Day 1 Calibrated Hail Guidance*

As with the Day 2 hail guidance, the HREF/SREF/ GEFS CSU, and HRRR NCAR methods were evaluated in the Day 1 timeframe. Again, GEFS CSU received the greatest mean participant rating (6.32), followed by HRRR NCAR (5.51) and HREF/SREF (4.85; yellow violins in Fig. 6). As expected, all three products received greater mean ratings for day 1 compared to day 2 (Fig. 6), suggesting improvement with shorter lead times. At the same time, in their comments, participants noted that each product's day 1 and day 2 characteristics were similar. Namely, GEFS CSU (HREF/SREF) generally had the greatest (smallest) probability magnitudes and areal coverage, while the subjective performance of HRRR NCAR was more varied compared to GEFS CSU, owing to its more focused probabilities. Given that the HRRR NCAR method is based on a single deterministic CAM rather than a full ensemble, the more focused probabilities from this method are not surprising.

*vii) 00Z HREF 24-h Calibrated Hail Guidance*

Four 24-h lead-time probabilistic hail guidance products based on the 00z HREF were evaluated. These included: the HREF/SREF, a deep-learning method designed by Amanda Burke (ML DL), the RF Loken, and a random forest method developed by Amanda Burke (RF Burke; Burke et al. 2020). Of these four methods, participants clearly favored RF Loken, both in their subjective ratings (median rating two points higher than any other guidance; blue violins in Fig. 6) and comments. Compared to the other methods, RF Loken generally had greater probability magnitudes and areal coverage, which tended to increase POD but also, occasionally, false alarm. In contrast, ML DL and RF Burke produced more focused probability areas, which sometimes hurt POD.

*viii) 00Z HREF 4-h Calibrated Hail Guidance*

Five products were evaluated at 4-h lead times: HREF/SREF, ML DL, RF Burke, and 06z and 13z SPC timing guidance. In their subjective ratings and comments, participants expressed a clear preference for the 06z (mean rating of 6.43) and 13z (mean rating of 6.39) SPC timing guidance (purple violins in Fig. 6). They felt that these methods generally had better probability coverage and magnitudes compared to the other methods (despite HREF/SREF being a primary input to these products). After the SPC timing guidance, RF Burke received the next greatest mean rating (5.23), followed by ML DL (4.97), and HREF/SREF (4.60).

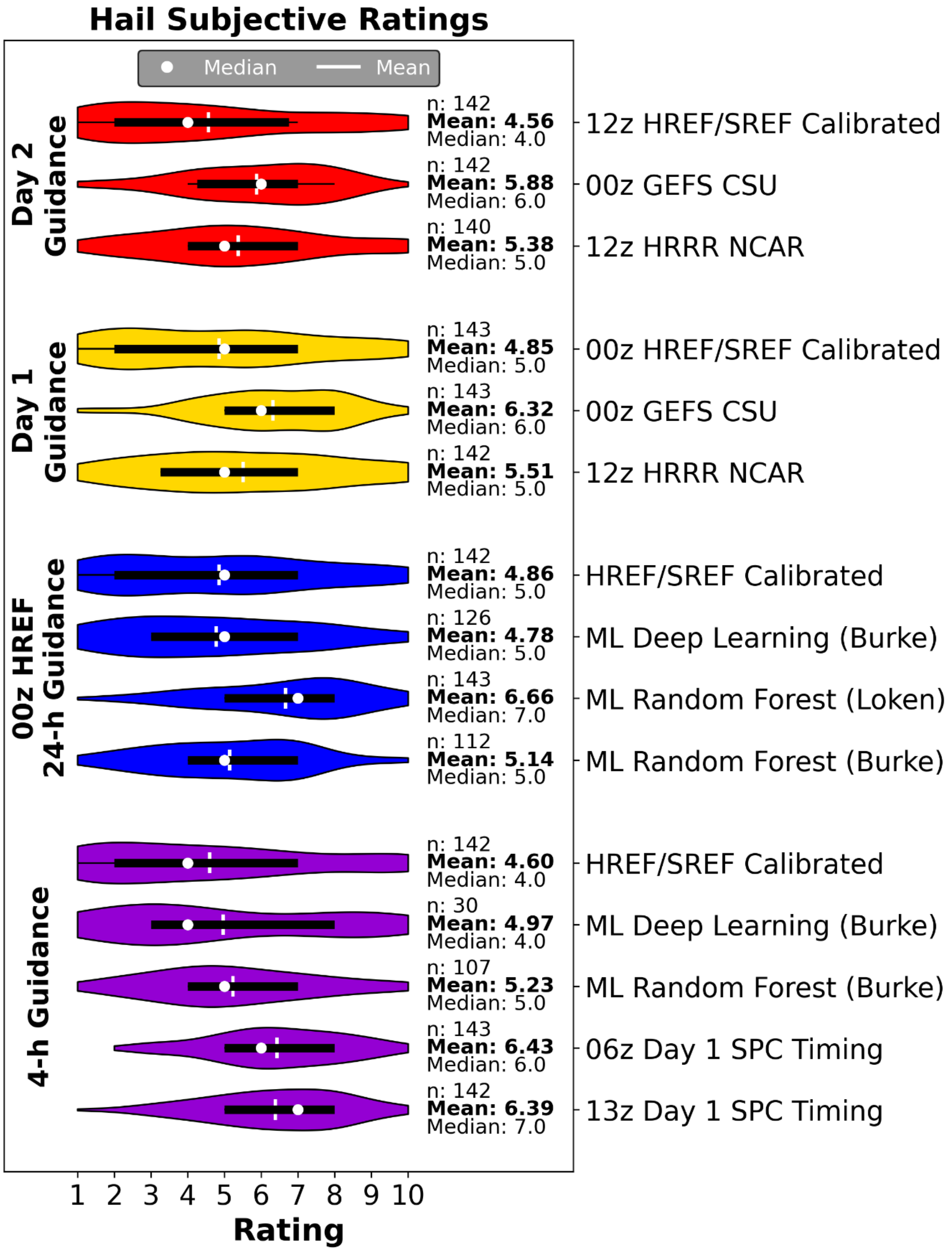


*Figure 6 Subjective participant ratings for calibrated severe hail guidance products, organized by evaluation type. Red, yellow, blue, and purple violins, respectively, correspond to day 2, day 1, 00z HREF, and 4-h products. White dots (bars) on each violin indicate median (mean) ratings. The number of responses (n), mean, and median ratings are also displayed to the right of each violin.*

### *ix) Day 2 Calibrated Wind Guidance*

As with hail, participants evaluated three methods for producing Day 2 probabilistic severe wind guidance: HREF/SREF, GEFS CSU, and HRRR NCAR. On average, participants ranked HRRR NCAR the highest (mean rating of 5.87), followed by GEFS CSU (mean rating of 4.98) and HREF/SREF (mean rating of 4.64; red violins in Fig. 7). Participants felt HRRR NCAR probabilities provided the best areal coverage but also noted that their magnitudes were sometimes too high. In contrast, participants thought HREF/SREF probability magnitudes and areal coverage were often too small. Participants generally found GEFS CSU useful but mentioned that the areal coverage of its probabilities did not always align with observed severe wind reports.

### *x) Day 1 Calibrated Wind Guidance*

At Day 1 lead times, participants ranked HRRR NCAR (mean rating of 5.93) and GEFS CSU (mean rating of 5.72) noticeably higher than HREF/SREF (mean rating of 5.21; yellow violins in Fig. 7). Again, participants felt HREF/SREF tended to under-forecast in terms of probability magnitude and areal coverage, while they noted that GEFS CSU and HRRR NCAR had different strengths and weaknesses. HRRR NCAR probabilities generally had good areal coverage but relatively high magnitudes, while GEFS probabilities sometimes covered too broad of an area, resulting in too much false alarm. Interestingly, participants noted that HRRR NCAR and GEFS CSU often complemented each other well by correctly highlighting different severe wind threat areas.

### *xi) 00Z HREF 24-h Calibrated Wind Guidance*

Two probabilistic guidance methods were evaluated for predicting 24-h severe wind based on 00z HREF data: HREF/SREF and RF Loken. On average, participants ranked RF Loken (mean rating of 6.23) higher than HREF/SREF (mean rating of 5.25; blue violins in Fig. 3), owing to its better POD. However, participants noted that both methods had flaws. While RF Loken had better POD, its probability magnitudes tended to be too high, and it produced more areas of false alarm than HREF/SREF. In contrast, participants felt that the areal coverage and magnitudes of HREF/SREF probabilities were frequently too low.

### *xii) 00Z HREF 4-h Calibrated Wind Guidance*

Three 4-h probabilistic severe wind forecasts were evaluated: HREF/SREF, 06z SPC, and 13z SPC. As with severe hail, participants expressed a strong preference for the SPC timing guidance products (mean ratings of 6.19 and 6.01 for the 06z and 13z products, respectively) compared to HREF/SREF (mean rating of 4.39; purple violins in Fig. 7). Participants wrote that the SPC products generally had better areal coverage of probabilities, resulting in greater POD, as well as better timing and probability magnitudes compared to HREF/SREF. Participants felt that the 06z and 13z SPC guidance were often similar or identical but noted that sometimes, surprisingly, the 06z guidance outperformed the 13z product.

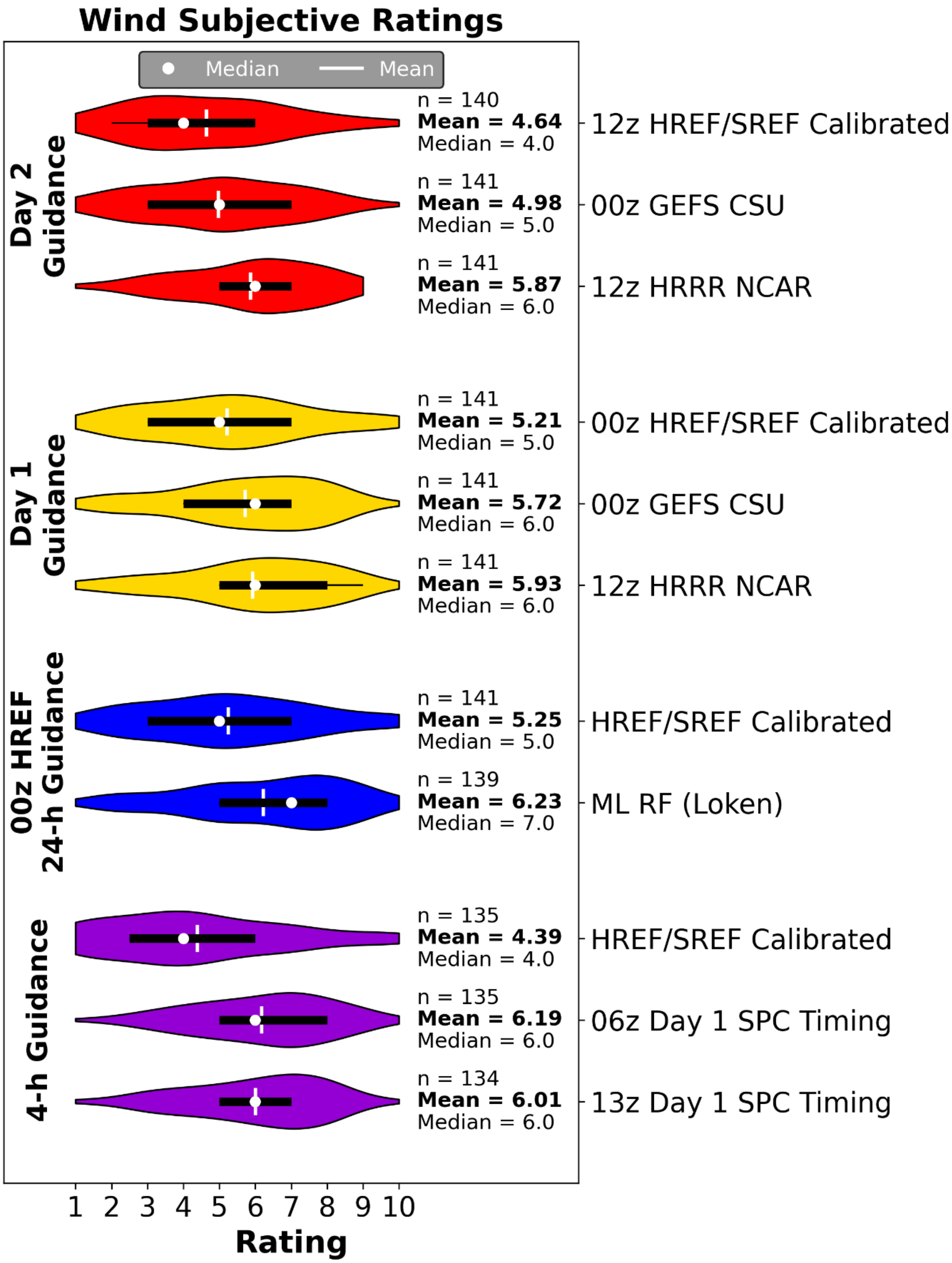


*Figure 7 As in Figure 6 but for severe wind guidance products.*

### *b) Model Evaluations – Group B: Deterministic CAMs*

#### *B1) Deterministic Flagships*

As in previous years, the Deterministic Flagships comparison evaluated experimental guidance from contributing agencies, with each agency contributing one model to this comparison. These models can be thought of as potential next iterations of the High-Resolution Rapid Refresh model (HRRRv4), an operational deterministic CAM as of this writing. Given the differing configurations, physics parameterizations, and data assimilation strategies, the focus on this comparison is less direct than other comparisons and instead serves to evaluate the state-of-the-art guidance. Simulated composite reflectivity at three times (f18, f24, and f30; or 1800 UTC, 0000 UTC, and 0600 UTC respectively) was examined and environmental fields were examined at one time (f18; 1800 UTC). Participants randomly were told to evaluate temperature, dewpoint, and surface-based CAPE to lessen participant workload; however, each field was assigned to approximately the same number of participants. These results are shown collectively, but specific comments about fields will be discussed. The observations used for the reflectivity are drawn from the MRMS composite reflectivity, and the environmental observations are drawn from the GSL 3D-RTMA Analysis (see section D3 for a discussion of the performance of these analysis fields). Finally, forecast soundings from two of the experimental models (the NSSL FV3-LAM and the EMC FV3-LAM) were evaluated for their depictions of inversions. Participants were instructed to find a time and location where inversions were present; as such, these results are less controlled than the other results in this comparison due to differing times and locations evaluated.

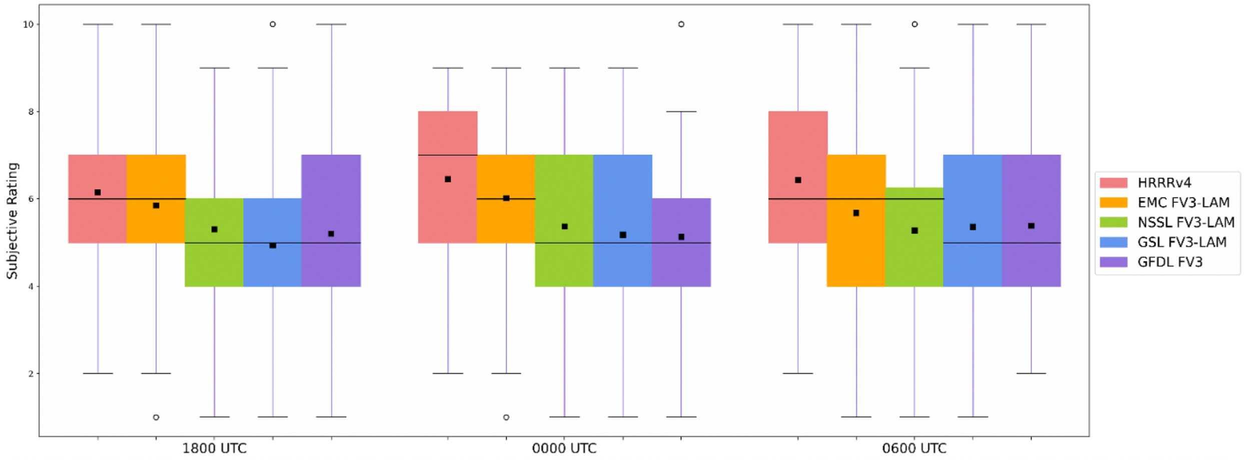


*Figure 8 Deterministic flagship simulated reflectivity results at 1800 UTC, 0000 UTC, and 0600 UTC. Mean subjective evaluation ratings are shown in the black square on each bar.*

At all hours, the HRRRv4 performed the best of any guidance (Fig. 8), with differences between the HRRRv4 and the other experimental guidance increasing later in the forecast period, at and past the diurnal afternoon convective peak. The EMC FV3-LAM performs best of any of the experimental FV3 guidance throughout, with average ratings of approximately 6/10. All model guidance performs very similarly throughout the evaluation times. After a few weeks of examining the model reflectivity in this comparison, it became clear that the NSSL FV3-LAM, which is configured the same as EMC FV3-LAM, except for the microphysics, had a bug in the reflectivity depiction, leading to high reflectivity values in areas of stratiform precipitation. This bug dominated the feedback on the NSSL FV3-LAM reflectivity performance, so the bug was corrected for the last two weeks of the experiment, starting with the 24 May 2021 run.

Figure 9 shows the ratings distributions prior to 24 May 2021 (Fig. 9a) and after 24 May 2021 (Fig. 9b). The NSSL-FV3's rating improved relative to the pre-fix distributions by about one point at 1800 UTC and 0000 UTC, and at all times the ratings distributions were higher once the fix was implemented. This is also reflected in the choice of which model performed best over the entire run. While the HRRRv4 and the EMC FV3-LAM were most frequently chosen before and after the NSSL FV3-LAM bug fix was implemented, the NSSL FV3-LAM went from being chosen as the best-performing model for a given day the least, to being selected 3rd most of any of the models after the fix.

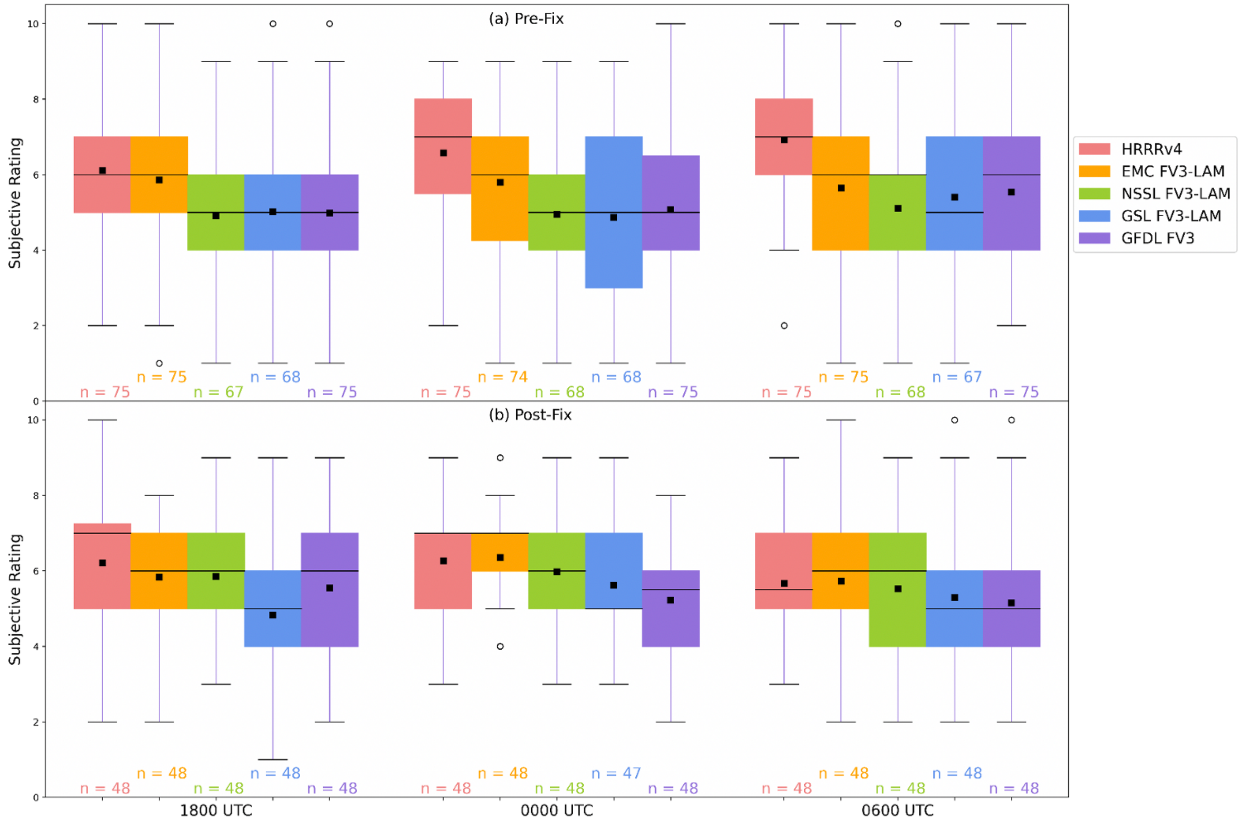


*Figure 9 As in Figure 8, but (a) before and (b) after 24 May 2021. Sample sizes are shown below each plot.*

Participants were also asked several open-ended questions about the guidance, including: (1) if there were any additional significant times to note, (2) what differences there were between the HRRRv4 and the FV3-based CAMs, and (3) if they had anything else to note. A few common themes emerged in the responses to these questions. First, the participants frequently noted convective characteristics of the FV3-based CAMs, such as their tendency to have more coverage of storms and very intense updrafts. A few participants suggested that the circular nature of the intense updrafts in FV3-based CAMs indicated explosive updrafts occurring in those models. Unlike prior years, a tendency was noted by a few participants that the FV3-based CAMs were less likely to produce MCSs, instead retaining discrete storms for too long. Multiple participants singled out the GFDL-FV3 as having very smooth reflectivity and a lack of intense convective cores, but they also noted that it had a good depiction of the stratiform region compared to some of the other CAMs. Across the board, three themes were noted for all CAMs in this comparison. First, many participants stated that they saw few differences between the FV3-based CAMs and the HRRRv4, remarking upon the great progress that has been made in the development of FV3-based CAMs over the past few years. Second, participants noted that all involved models (including the HRRRv4) struggled with convective initiation, often being too late with initiation on a given day. Finally, participants frequently wrote that models performing well at one time didn't necessarily guarantee that they would perform well at other times, and that the best-performing model at any given time would switch. While systematic variation in which model performed best did not show up across the experiment, it is important to recall that for individual case studies variation in relative model performance across the day is to be expected.

Participants were also asked to examine environmental fields important to convection at 1800 UTC, in hopes that that time would be prior to the majority of convection on any given day and provide for a cleaner comparison of background model environment than other times would. However, as is often the case, remnant convection from the previous day frequently affected the domain of interest and remnant boundaries or ongoing convection factored into several cases. Overall, subjective evaluations of the environmental fields (Fig. 10) followed similar patterns to the reflectivity, with the HRRRv4 performing best. The GSL FV3-LAM performed next best, followed by the EMC FV3-LAM and the NSSL FV3-LAM. These patterns held before and after the reflectivity bug fix in the NSSL FV3-LAM (not shown), which is expected given that the fix soley affected the reflectivity depiction.

Participant comments about the environmental fields had similar themes, noting biases in temperature, dewpoint, and SBCAPE. Cold biases in the temperatures were noted by many participants, especially outside of cold pools. Within cold pools, participants noted an occasional cold bias that led to a fast propagation of the cold pools. Cases with a high and low dewpoint bias (e.g., too moist or too dry) were both noted, but there were many more cases with too low of dewpoints relative to the observations (see results from the D3 evaluation regarding the moist bias in the GSL 3D-RTMA). Dryline placement was noted as being too far east in one case. The low dewpoint bias contributed to a low bias in the CAPE, particularly for the FV3-based models. While the GSL FV3-LAM and the HRRRv4 were frequently singled out as having better SBCAPE compared to the other models, no model in the B1. Deterministic Flagships comparison did particularly well with handling the magnitude of SBCAPE during the experiment.

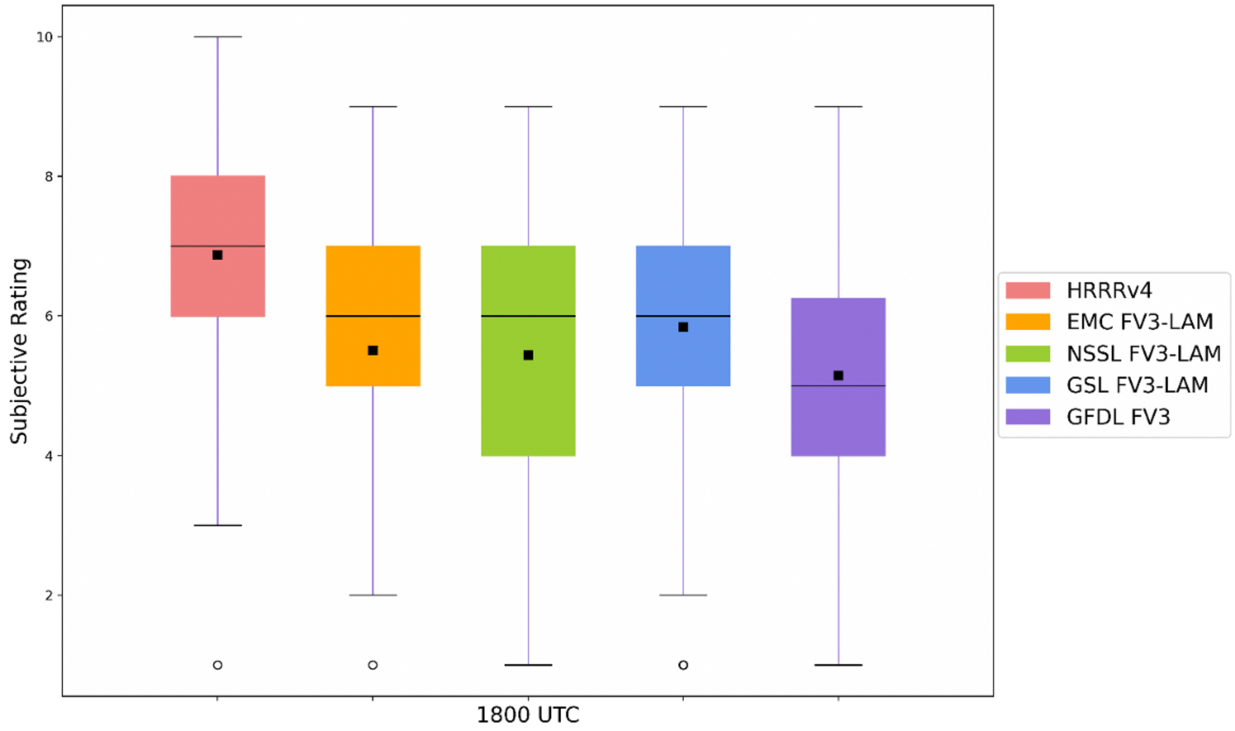
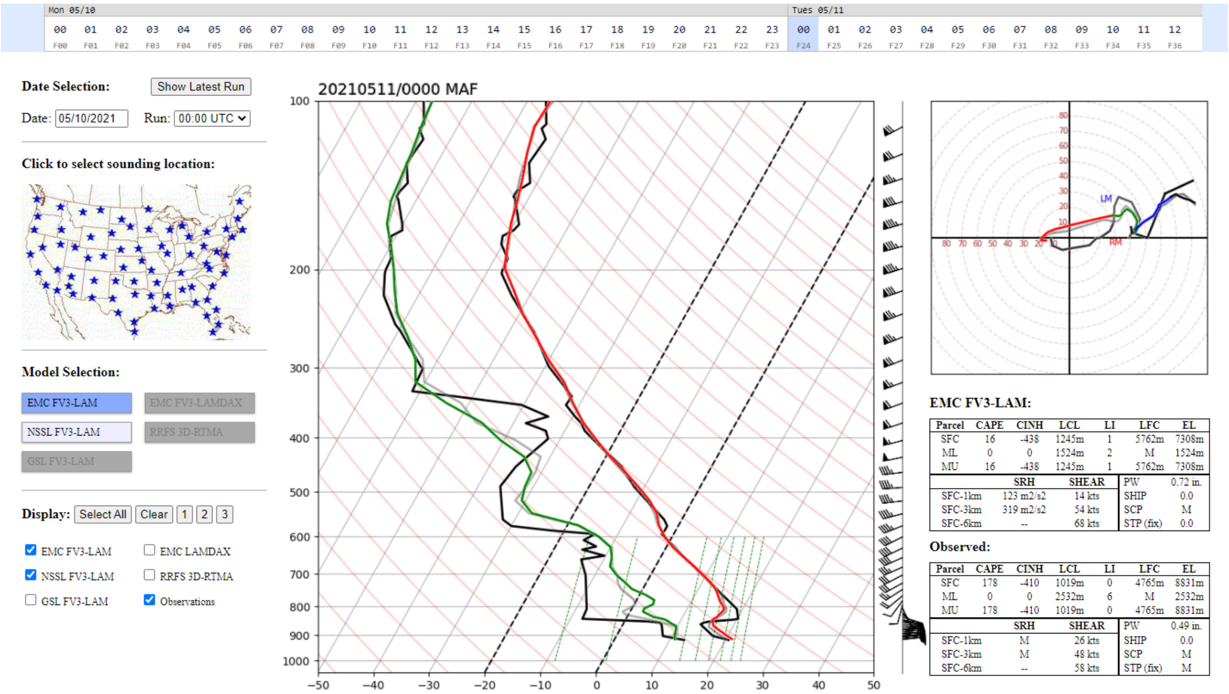


*Figure 10 As in Figure 8, but for the environmental fields at 1800 UTC.*

The final piece of the B1. evaluation focused on the sounding depiction, particularly on the depiction of inversions (Fig. 11). The EMC FV3-LAM and NSSL FV3-LAM soundings were available throughout the experiment for evaluation. Participants generally found the EMC FV3-LAM to better depict the inversion strength compared to the NSSL FV3-LAM, but the differences in the means were small and both models typically had weaker inversions than observations (Fig. 12). The inversion height more closely matched observations than the inversion strength in both models, with the mean inversion height nearly perfectly matching observations. The EMC FV3-LAM and the NSSL FV3-LAM were evenly chosen as the guidance with the best match to observed soundings at the chosen participant location and time; the EMC FV3-LAM was chosen 44 times (48.35% of the sample) and the NSSL FV3-LAM was chosen 45 times (49.45% of the sample).



**Date Selection:**

Date: 05/10/2021 Run: 00:00 UTC

**Click to select sounding location:**

**Model Selection:**

**Display:**

☑ EMC FV3-LAM ☐ EMC LAMDAX

☑ NSSL FV3-LAM ☐ RRFS 3D-RTMA

☐ GSL FV3-LAM ☑ Observations

**EMC FV3-LAM:**

| Parcel | CAPE | CINH | LCL | LI | LFC | EL |
|---|---|---|---|---|---|---|
| SFC | 16 | -438 | 1245m | 1 | 5762m | 7308m |
| ML | 0 | 0 | 1524m | 2 | M | 1524m |
| MU | 16 | -438 | 1245m | 1 | 5762m | 7308m |

| | SRH | SHEAR | | |
|---|---|---|---|---|
| SFC-1km | 123 m2/s2 | 14 kts | PW | 0.72 in. |
| SFC-3km | 319 m2/s2 | 54 kts | SHIP | 0.0 |
| SFC-6km | -- | 68 kts | SCP | M |
| | | | STP (fix) | 0.0 |

**Observed:**

| Parcel | CAPE | CINH | LCL | LI | LFC | EL |
|---|---|---|---|---|---|---|
| SFC | 178 | -410 | 1019m | 0 | 4765m | 8831m |
| ML | 0 | 0 | 2532m | 6 | M | 2532m |
| MU | 178 | -410 | 1019m | 0 | 4765m | 8831m |

| | SRH | SHEAR | | |
|---|---|---|---|---|
| SFC-1km | M | 26 kts | PW | 0.49 in. |
| SFC-3km | M | 48 kts | SHIP | 0.0 |
| SFC-6km | -- | 58 kts | SCP | M |
| | | | STP (fix) | M |

*Figure 11 The webpage participants used for sounding evaluation, showing the EMC FV3-LAM sounding in color, the NSSL FV3-LAM sounding in light grey, and the observed sounding in black.*

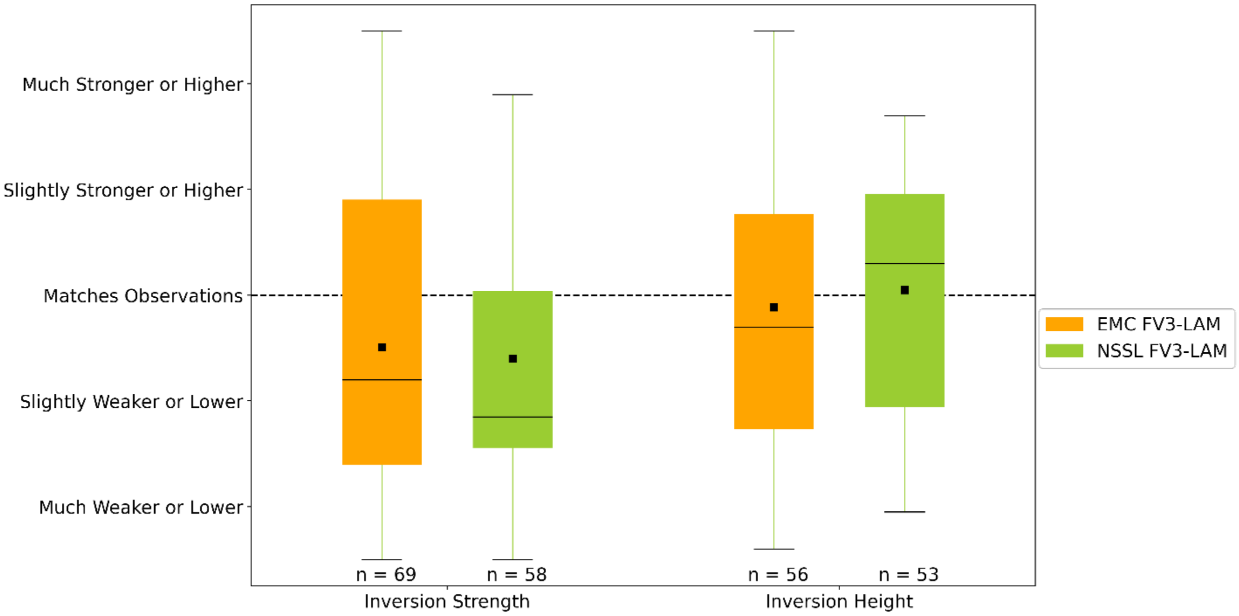


*Figure 12 Inversion strength and height characterization for the EMC FV3-LAM and the NSSL FV3-LAM. The dashed line indicates a perfect match with observations.*

### *B2) CLUE: FV3-LAM Domain Comparison*

The next comparison in the B group looked at the impact of differing domains of the limited area model (LAM) versions of the FV3. Two pairs of models were compared, one from EMC (the EMC FV3-LAM and the EMC FV3-LAMX) and one from GSL (the GSL FV3-LAM and the GSL FV3-LAM-NA). The sole difference in the EMC pair was the domain size, whereas the GSL pair differed in other aspects as well, such as the data assimilation strategy. The EMC FV3-LAM and the GSL FV3-LAM both used a CONUS domain; the EMC FV3-LAMX and the GSL FV3-LAM-NA both used a North America domain. Reflectivity and environment comparisons were done similarly to the B1. Deterministic Flagships comparison, but the in-depth examination was done at Day 2 (D2; e.g., 1800 UTC, 0000 UTC, and 0600 UTC corresponded to f42, f48, and f56, respectively). At Day 1, participants were merely asked about how similar the two pairs of models were on a scale of 1 (Least similar) to 5 (Most similar), to help highlight cases where domain differences may have been particularly impactful to the forecasts at short lead times.

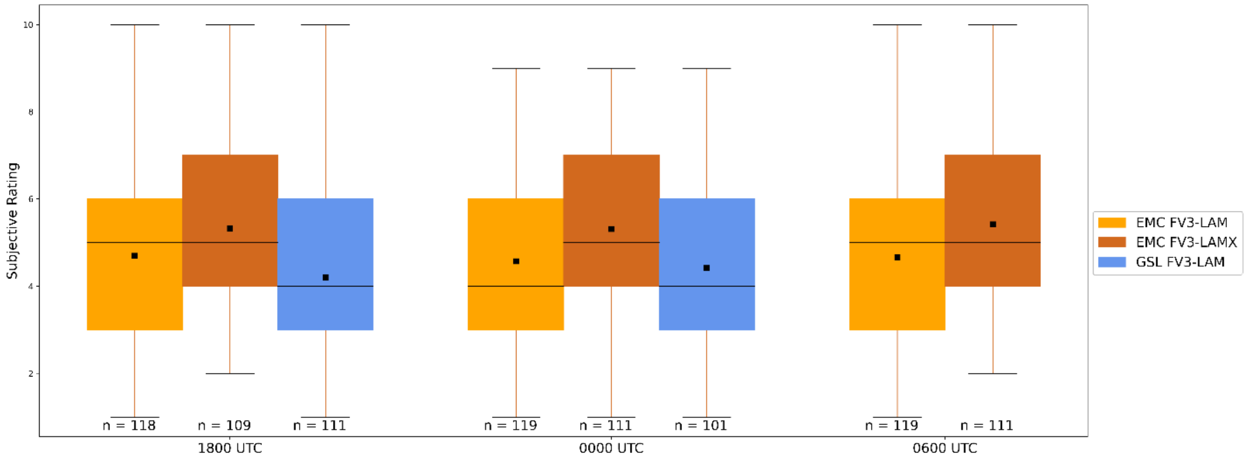


*Figure 13 Participant subjective evaluation of models in the FV3-LAM comparisons. Orange models are run by EMC and blue models are run by GSL. Darker colors are run over the North America domain; lighter colors are run over the CONUS domain. The mean of each distribution is marked by a black square.*

When looking at the reflectivity results (Fig. 13), participants generally preferred the North America domain to the CONUS domain at all times, with the mean difference slightly increasing as the day went on. The GSL FV3-LAM-NA had too small of a sample size to include in the analysis, and the length of the GSL FV3-LAM run prevented evaluation at 0600 UTC. Participants generally preferred the EMC FV3-LAMX to the other models in this comparison when asked to rate the overall convective evolution through the forecast period, with it being selected as the best 56% of the time (not shown). Participant comments focused less on comparisons between the specific pairs of models and more on model groupings (e.g., "The EMC Models were…") indicating that more differences occurred between GSL and EMC configurations, rather than between the two GSL *or* EMC configurations. The most common participant comment seemed to be that the models performed poorly across the board, with one participant even writing that none of the available guidance for this comparison would have helped their forecasts for severe convection. Perhaps the poor performance relative to other comparisons is in part due to the models being evaluated during the D2 timeframe; when comparing the reflectivity results for D2 and Day

1 (D1), there is an increase for the EMC FV3-LAM and the GSL FV3-LAM of 1-2 points with the later forecast initializations (Fig. 14).

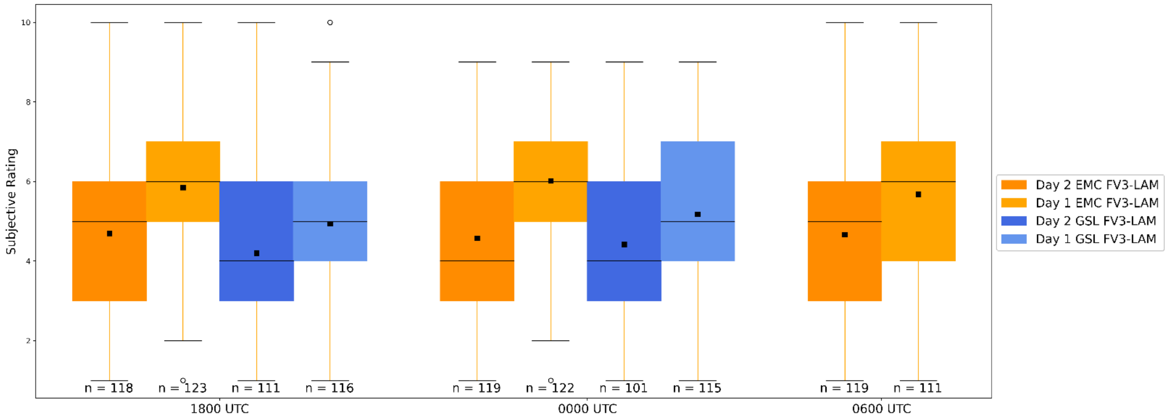


*Figure 14 Participant subjective evaluation of the EMC FV3-LAM and the GSL FV3-LAM using initializations from 0000 UTC on the day of the forecast evaluation period (Day 1; lighter colors) and from 0000 UTC the day prior to the forecast evaluation period (Day 2; darker colors).*

Evaluation of the D2 environmental fields (Fig. 15) show differences in the median ratings between the EMC FV3-LAM and the EMC FV3-LAMX, but the mean values are quite similar. At 0000 UTC, differences are larger than at 1800 UTC, likely in part due to the influence of convection. For the environmental fields at 1800 UTC, again the EMC FV3-LAMX is preferred to the EMC FV3-LAM, and the EMC FV3-LAMX score distribution is approximately equivalent to the GSL FV3-LAM score distribution. Participant comments for this evaluation focused more on the differences between the EMC FV3-LAM and the EMC FV3-LAMX relative to the reflectivity comparison. Participants noted a persistent low instability bias across all of the models, describing that while the location of the SBCAPE was fine, the magnitude was severely underdone (again, likely influenced by use of GSL 3D-RTMA as the observations). Participants noted variation in cases between how much CAPE the EMC FV3-LAM and the EMC FV3-LAMX depicted, though which model performed better was inconsistent from case to case. Additional interesting notes by participants were that the LAM showed a strong dependence on orographic features, and that the LAMX handled frontal boundary placement and location better than the LAM. These comments, together with the participant ratings for both reflectivity and environmental fields, indicate that the North American-domain runs offer improved forecasts over the CONUS-domain runs for the Limited-Area Model FV3.

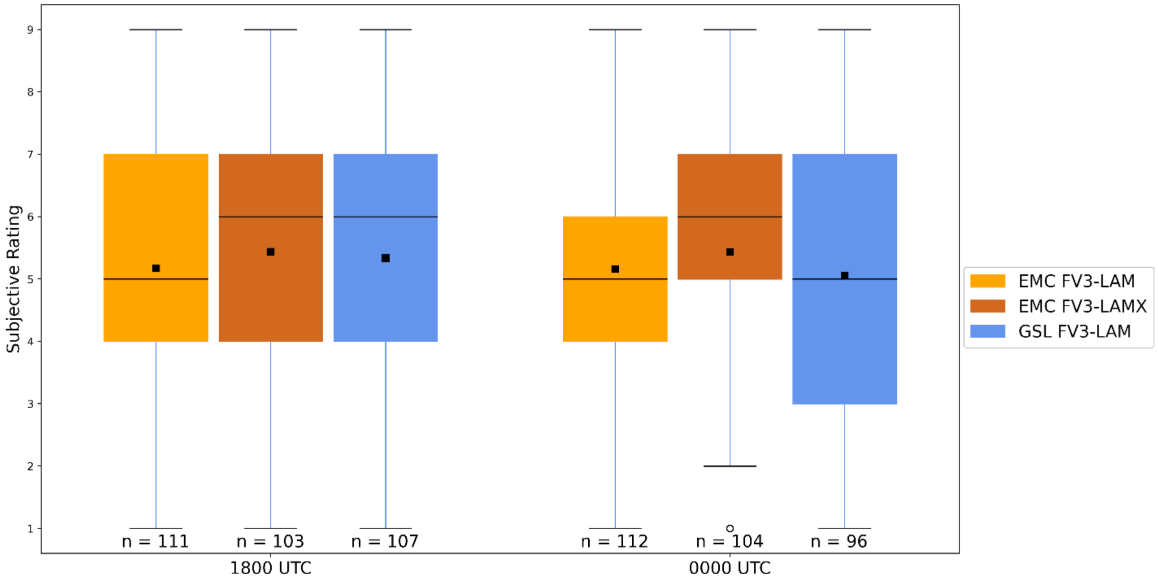


*Figure 15 As in Figure 13, but for temperature, dewpoint, and SBCAPE fields.*

Improvements to the environmental fields with later initializations (e.g., D1 vs. D2) were smaller than those associated with simulated reflectivity (Fig. 16), with most of the improvement seeming to come from an increase in the lower percentile ratings. The similarity between the EMC FV3-LAM and the EMC FV3-LAMX were very high during the D1 timeframe (Fig. 17), which is an expected result given that the differing boundary conditions may not have had a chance to propagate into the domain of interest for greatest severe weather. The GSL models, while less frequently available for comparison, were much more different from one another. This result is also somewhat unsurprising considering the differences in the GSL pair of models were a function of more than just the model domain.

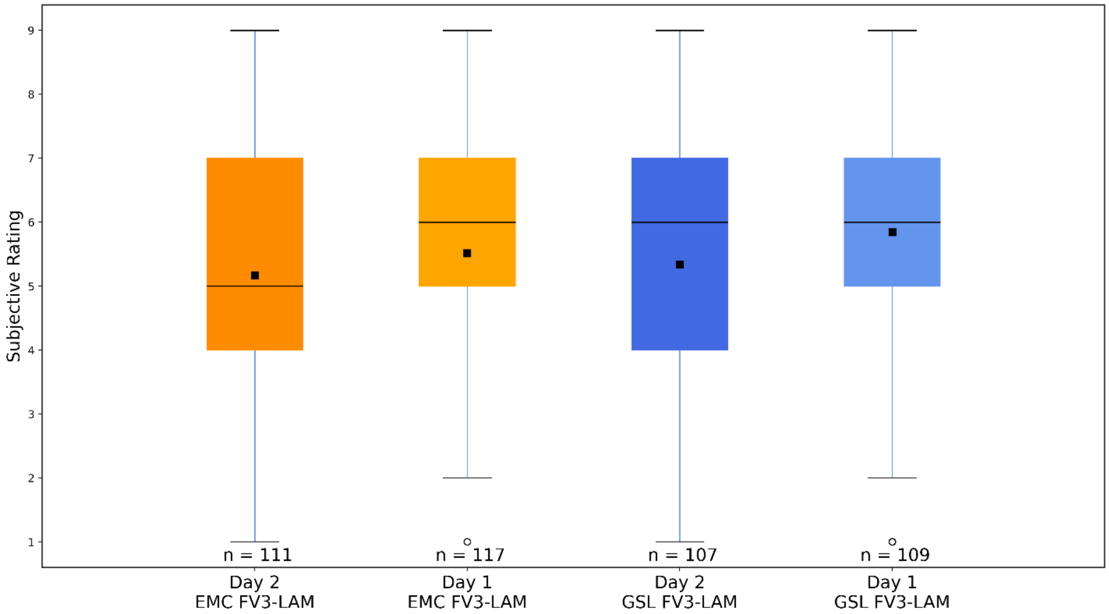


*Figure 16 As in Figure 14, but for the temperature, dewpoint, and SBCAPE fields.*

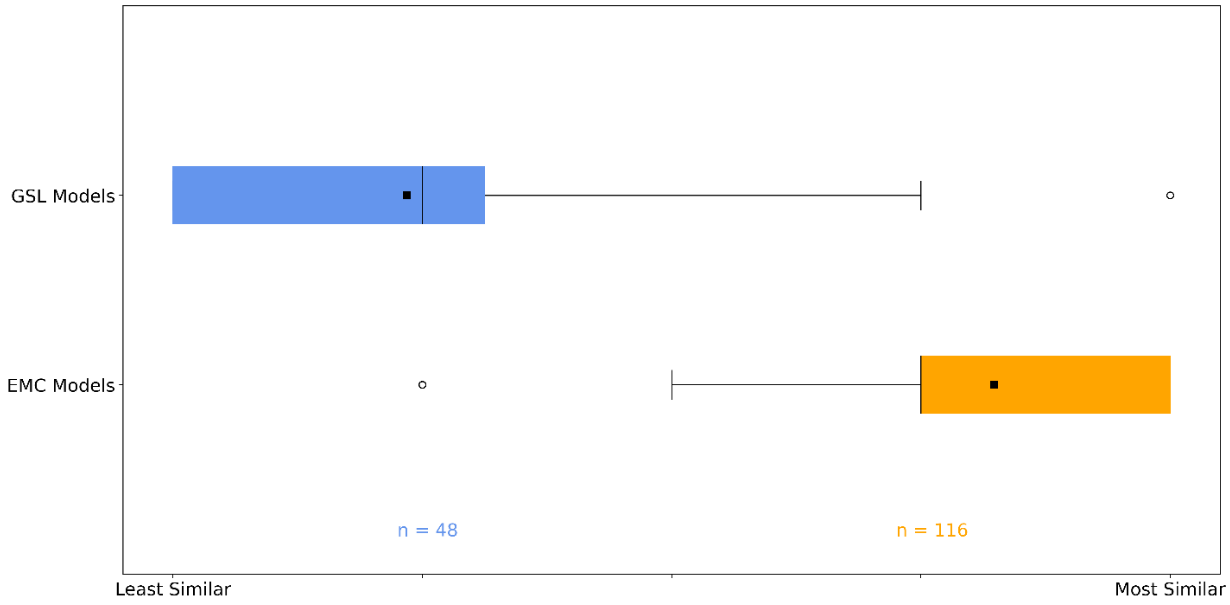


*Figure 17 Participant responses to the question, "On a scale of 1-5, with 1 being the least similar and 5 being the most similar, how similar are the reflectivity and UH forecasts from the following pairs of models?"*

### *B3) CLUE: FV3-LAM Data Assimilation Comparisons*

This comparison looked at various forms of Data Assimilation (DA) and their impact on the forecasts relative to the cold-start EMC FV3-LAM. The evaluation times remained the same as prior comparisons, with the earliest forecast hour evaluated being forecast hour 18. Since frequently the benefits of DA are most evident in the first few hours of a forecast, when participants were asked if there were any additional times of note, "(e.g., the first few hours)" was added to the question to prompt participants to consider those hours. Participants were also prompted with their prior ratings for the reflectivity of the EMC FV3-LAM and the GSL FV3-LAM to provide context for the new models being evaluated and to reduce the participant workload.

The reflectivity findings showed that the cold-start EMC FV3-LAM performed better than the models using DA at the specific forecast hours evaluated herein, though the EMC FV3-LAM-DAX had a very similar distribution to the EMC FV3-LAM at 1800 UTC (Fig. 18). By 0600 UTC, many of the models incorporating DA (all but the EMC FV3-LAM) were performing similarly. The MAP RRFS VTS control member performed about the same as the MAP RRFS control member at 1800 UTC, but did slightly better than the MAP RRFS control member at 0000 UTC and 0600 UTC. However, these forecast hours may not have represented the overall reflectivity forecasts for this set of models as well as prior comparisons; when asked to choose a best-performing model, the EMC FV3-LAM and the EMC FV3-LAMDAX were selected nearly the same number of times, and the MAP RRFS VTS Control member was also selected fairly frequently (Fig. 19). This result suggests that objective verification over all hours of these forecast models should be undertaken to determine the full impact of DA on FV3-LAM forecasts.

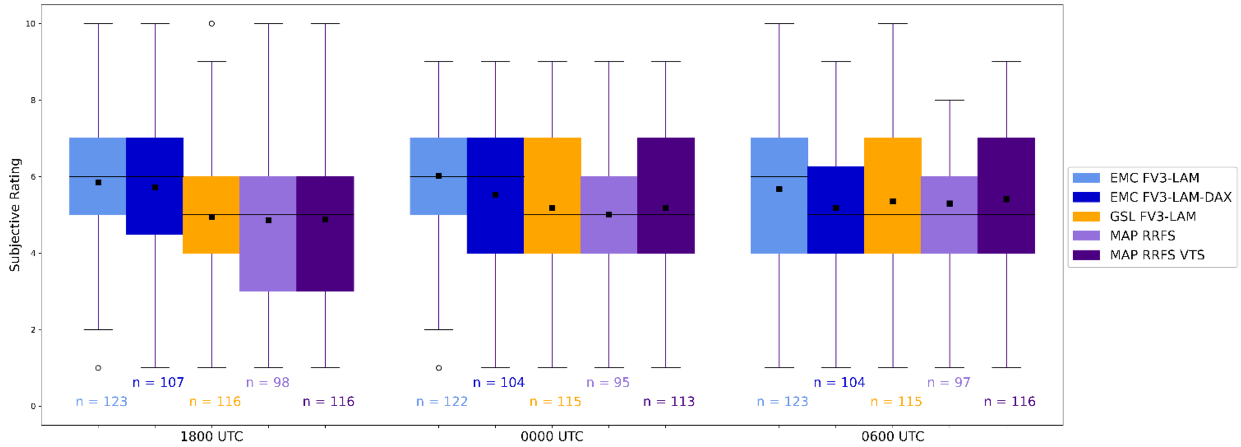


*Figure 18 Participant subjective evaluation of simulated reflectivity for models in the Data Assimilation comparison at three selected times. Sample sizes for each distribution are shown below the plots, and the black square shows the mean rating of each distribution.*

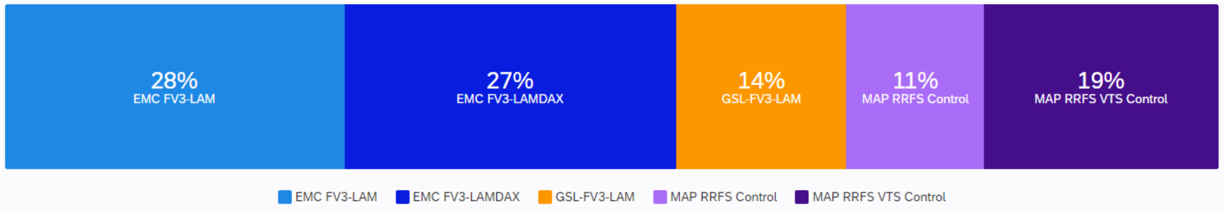


*Figure 19 Participant responses to the question "Which deterministic CAM(s) best captured the convective evolution (i.e., timing, CI, convective mode) through the entire forecast run?" Participants could choose more than one response on a given day.*

Participant comments indicated that larger differences frequently occurred in earlier hours, as was expected. One or two participants singled out the first forecast hour as being the most different, while others indicated the first 6 to 9 hours had large differences. However, other participants mentioned later forecast hours, such as the f12-f24, or f21 as having interesting differences, frequently concerning CI. Not all of these highlighted hours showed a positive impact of DA; in one case, a participant highlighted that there was actually *too much* convection in the models with DA at early hours.

Comments on the reflectivity forecasts overall seemed to be that the MAP runs had too much and too intense of convection; an early bug in the runs was corrected mid-experiment, and after this fix the frequency of those comments decreased. The MAP runs were also highlighted as better in the morning, but then overforecasting in the afternoon once convection initiated. Another participant indicated that convective initiation was too early in the EMC runs, but that the MAP runs did better with CI timing, and that the control RRFS MAP run did the best in that case. A broad comment was that many of the models missed or struggled with overnight convection, even when they employed DA, suggesting that work remains to develop successful DA strategies in the FV3-LAM cores to capture and develop overnight convection.

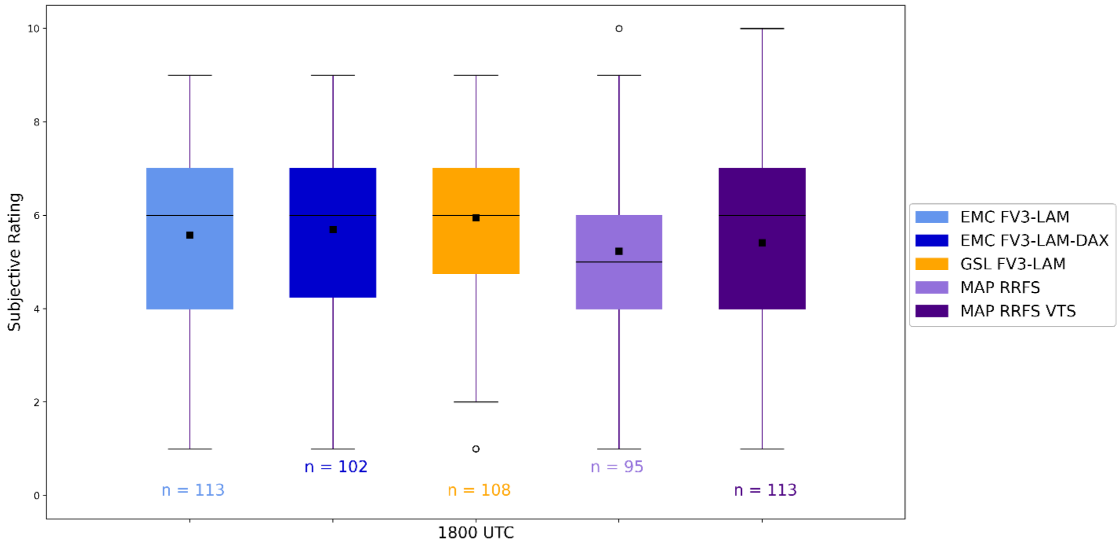


*Figure 20 Participant subjective evaluation of temperature, dewpoint, and SBCAPE for models in the Data Assimilation comparison at 1800 UTC. Sample sizes for each distribution are shown below the plots, and the black square shows the mean rating of each distribution.*

For the environmental fields, the models were rated much more similarly than in the reflectivity comparison. The GSL FV3-LAM had a slightly higher mean rating than the other models, and the mean of the EMC FV3-LAM-DAX was slightly higher than the mean of the EMC FV3-LAM. The largest outlier is the MAP RRFS Control member, which had lower scores for the environmental variables relative to the rest of the models. Participant comments again indicated an underprediction of SBCAPE magnitudes across all of the models. The EMC LAM-DAX was pointed out multiple times as having too strong of cold pools, while the MAP RRFS VTS member was the only member specifically mentioned by participants as having a good evolution of the cold pool (this occurred for the 18 May 2021 case). Other participants mentioned the diurnal cycle of thermodynamic variables, as in prior comparisons. The diurnal cycle seemed to be too amplified in the EMC runs, not amplified enough in the MAP runs, and about right for the GSL runs according to one participant.

### *B4) CLUE: FV3 Physics Suites Comparisons*

This comparison looked at different microphysics and planetary boundary layer (PBL) suites implemented in FV3-LAM configurations, to examine (1) which set of suites performed the best in forecasts of severe convection, and (2) how much spread could be achieved from an ensemble perspective by simply varying the microphysics and PBL schemes. As in the other "B" evaluations, participants looked at the simulated reflectivity and one of either temperature, dew point, or SBCAPE and rated each model from 1–10, where 1 indicates very poor performance and 10 indicates very good performance.

Reflectivity comparisons showed that the MYNN/Thompson schemes generally performed best, especially at later times in the forecast (Fig. 21). The MYNN/Thompson member was also selected as a best-performing model across the entire forecast run about 50% of the time (not shown). The Hybrid-EDMF/NSSL member was ranked lower than the other members, but this may in part be due to the bug that affected the NSSL FV3-LAM in B1. A fix was not implemented in the member used during this comparison, but should be re-evaluated in future experiments. Participant comments support the attribution of the poor ratings to the bug, with several participants throughout the experiment noting the very high reflectivity values in the Hybrid-EDMF/NSSL member. Other participant comments focused on the TKE-EDMF/GFDL member, with a few participants stating that the storms were not realistic looking due to the large storm sizes and low maximum reflectivity in those cells. However, one participant noted that on 13 May 2021 the TKE-EDMF/GFDL member was able to correctly initiate convection earlier than the other members, leading to a higher rating in that case.

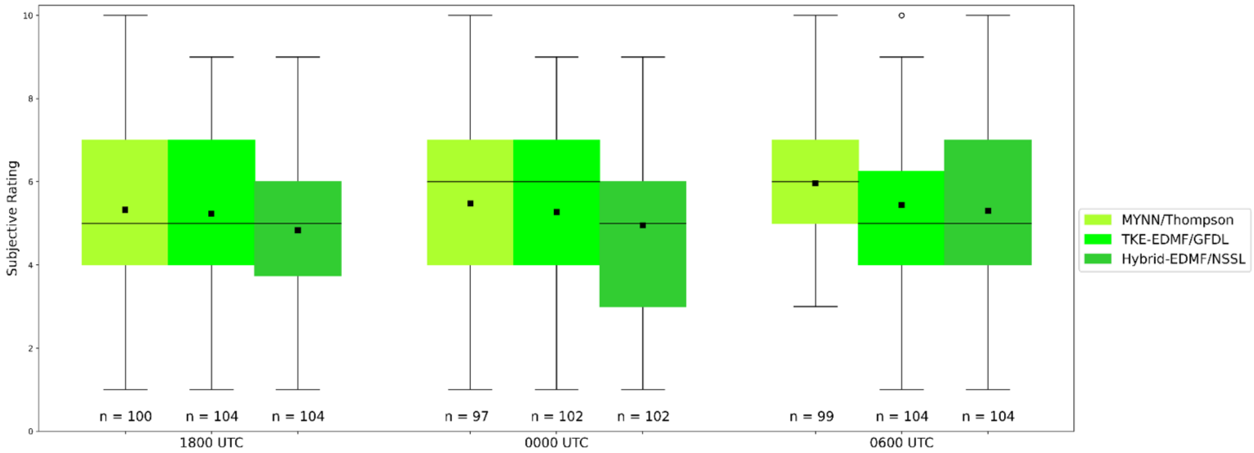


*Figure 21 Participant subjective evaluation of simulated reflectivity for models in the Physics Suite comparison at three selected times. Sample sizes for each distribution are shown below the plots, and the black square shows the mean rating of each distribution.*

Subjective ratings of the environment at 1800 UTC showed a similar distribution to the simulated reflectivity comparisons, with the MYNN/Thompson run performing better than the other two runs (Fig. 22). When asked about additional times of interest, many participants mentioned differences at 2100 UTC, suggesting that this may be an additional time worth examining in future experiments. Specific notes at this time period were that the MYNN/Thompson member correctly had higher CAPE than the other members, and that the Hybrid-EDMF/NSSL member was too cool relative to the other members (though this feedback varied; in later questions, the Hybrid-EDMF/NSSL member was noted as being too warm!). One participant noted that while the Hybrid-EDMF/NSSL member had a better representation of the location for a convective boundary, the magnitude of that boundary was still underdone relative to the observations for all models. Additional comments about the environment for this comparison noted that there was a dry bias across all models (again, impacted by moist bias in GSL 3D-RTMA), as well as a warm bias noted for the TKE-EDMF/GFDL and the Hybrid-EDMF/NSSL in the late afternoon. As for reflectivity, one participant similarly noted a smoothness to the TKE-EDMF/GFDL environment fields. Finally, many of the comments again summarized all three models (e.g., “All models had trouble forecasting the Td…”), suggesting that the environments did not show many differences between the models relative to the model difference with the observations. One participant even noted that they were surprised by the similarity in the environmental fields, given that the simulated reflectivity and UH were very different among the various parameterizations for the case that they examined.

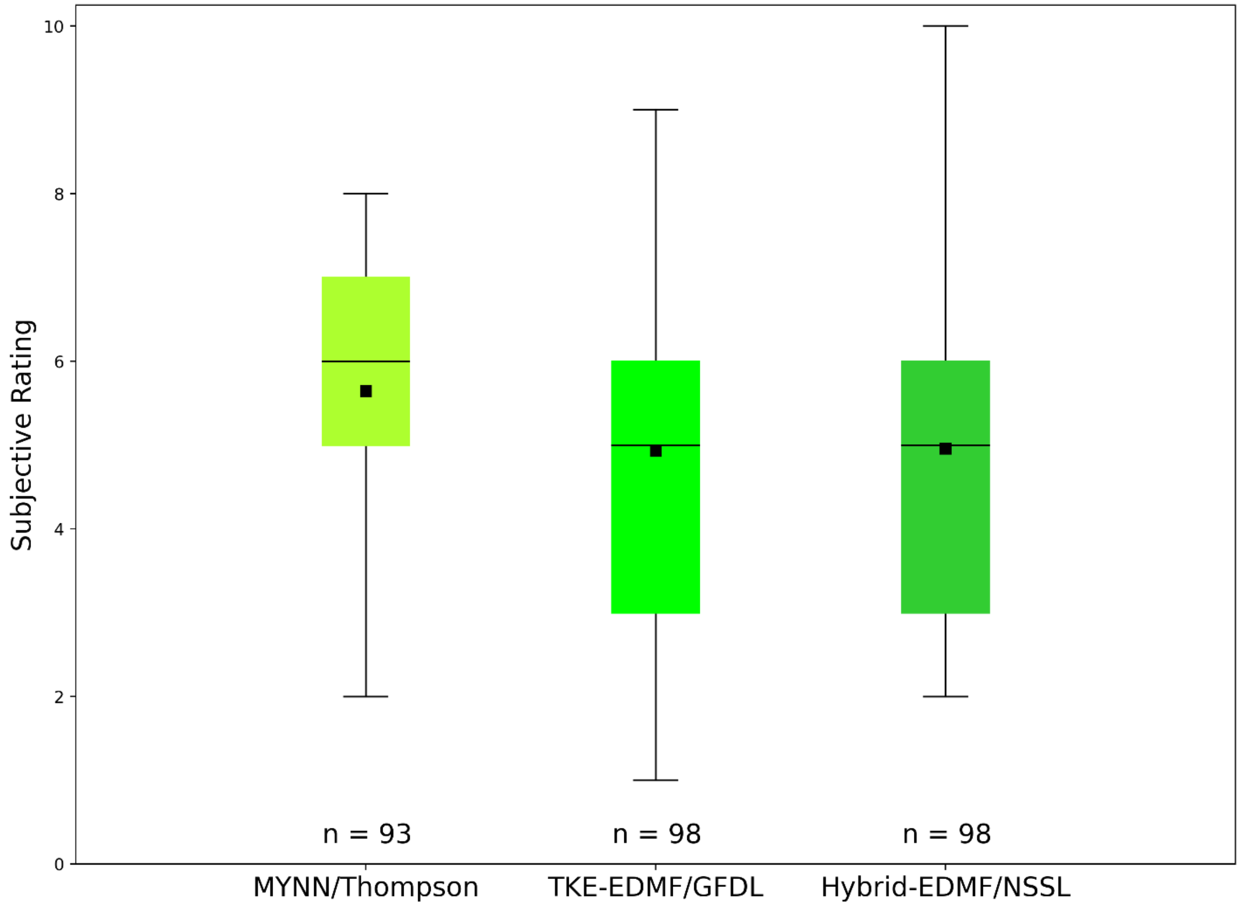


*Figure 22 Participant subjective evaluation of temperature, dewpoint, and SBCAPE for models in the FV3 Physics Suite comparison at 1800 UTC. Sample sizes for each distribution are shown below the plots, and the black square shows the mean rating of each distribution.*

### *B5) CLUE: FV3 Stochastic Physics Comparisons*

This comparison looked at different implementations of stochastic physics suites within the RRFS Cloud ensemble, with similar goals to the B4. comparison, namely to determine (1) which stochastic physics approach performed the best in forecasts of severe convection, and (2) how much spread could be achieved from an ensemble perspective by implementing stochastic physics. This comparison is not quite as clean as the B4. comparison, as the models herein also had varying initial conditions (ICs) and lateral boundary conditions (LBCs) that could have contributed to their differences. However, the ICs and LBCs for the two models with stochastic perturbations should be statistically indistinguishable over a sufficiently large sample. As in the other “B” evaluations, participants looked at the simulated reflectivity and one of either temperature, dew point, or SBCAPE and rated each ensemble from 1–10, where 1 indicates very poor performance and 10 indicates very good performance. The baseline model in this experiment did not contain any stochastic physics, and was compared to two models that each had different types of stochastic physics (SPPT and SPPT/SHUM/SKEB perturbations, respectively). The baseline model was the MYNN/Thompson member evaluated in the previous comparison, and participants were reminded of their score for that member while evaluating the two new members with stochastic physics perturbations.

The reflectivity results showed that the member without stochastic physics generally outperformed the members with stochastic physics, especially at 1800 UTC and 0000 UTC (Fig. 23). The SPPT member performed the worst early in the run, but improved steadily as the forecast hour increased. Conversely, the mean rating of the SPPT/SHUM/SKEB member increased only slightly as time went on, leading it to be the worst-performing model in this comparison by 0600 UTC.

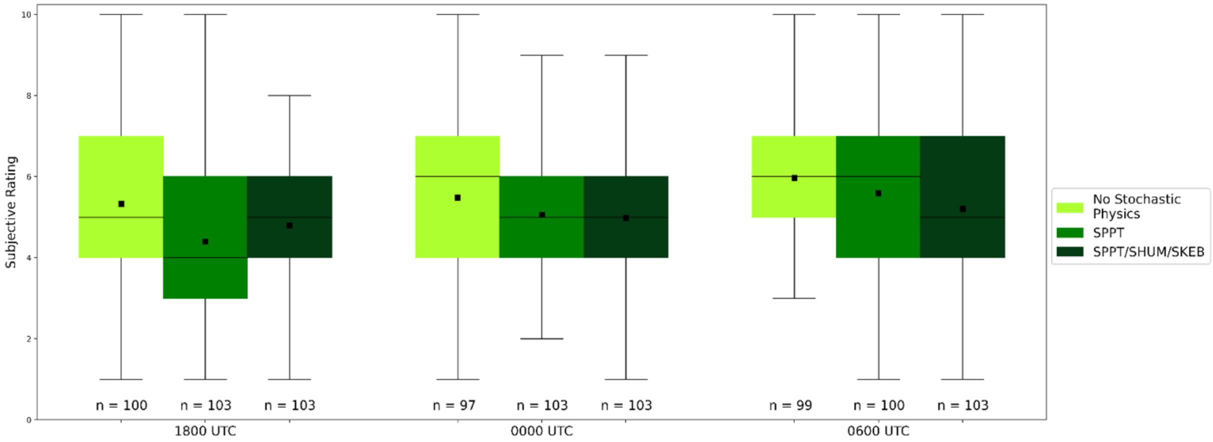


*Figure 23 Participant subjective evaluation of simulated reflectivity for models in the Physics Suite comparison at three selected times. Sample sizes for each distribution are shown below the plots, and the black square shows the mean rating of each distribution.*

Participants noted differences in the ability of the models to capture convective initiation for some cases, with the SPPT performing better than the SPPT/SHUM/SKEB on one day (1 June 2021) and the SPPT/SHUM/SKEP doing better on other days (13 May 2021). Similar to what is seen in the overall distributions, participants noted that the SPPT improved slightly as convection continued. However, there were plenty of comments indicating that the best-performing model shifted over the course of the forecast, similar to what was seen in the Deterministic Flagship comparison. This is supported by the participants' selection of which model(s) performed best during the full run, which showed the SPPT/SHUM/SKEP as a best-performing model most frequently, closely followed by the No Stochastic baseline model (Fig. 24). Multiple participants highlighted the SPPT/SHUM/SKEB as being able to handle the dominant convective mode relative to the other members. One participant suggested that if the three forecasts in this comparison were taken as an aggregate, they would have highlighted the important areas, supporting that stochastic physics could be a useful way to introduce spread in future FV3-based ensemble configurations.

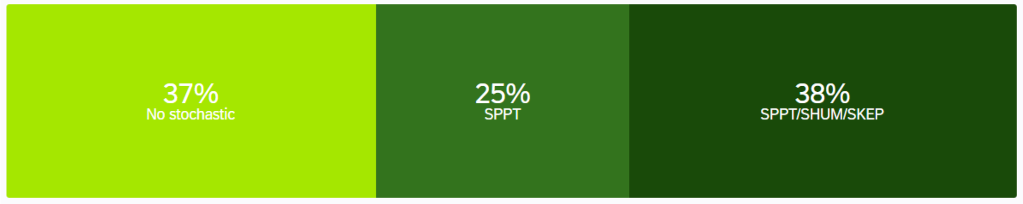


*Figure 24 Participant responses to the question "Which deterministic CAM(s) best captured the convective evolution (i.e., timing, CI, convective mode) through the entire forecast run?" Participants could choose more than one response on a given day.*

The environmental fields showed similar distributions to the reflectivity subjective evaluation scores in that the No Stochastic member performed better than the members with stochastic physics parameterizations (Fig. 25). However, there were very few differences between the SPPT and the SPPT/SHUM/SKEP member relative to the reflectivity and UH comparisons described earlier. Interestingly, despite the No stochastic member having a higher mean than the other two members, the highest score it received was an 8/10, compared to a 9/10 achieved by the SPPT/SHUM/SKEB and a 10/10 achieved by the SPPT. In the comments, once again participants were noticing systemic differences across all models, such as the insufficient magnitude of CAPE (albeit with decent coverage/location of CAPE) and a cool and dry bias. The SPPT/SHUM/SKEP member was highlighted by participants most frequently: for having too low of a dewpoint in one case, for a correctly higher dewpoint in another case, and for capturing a CAPE increase that the other members did not depict.

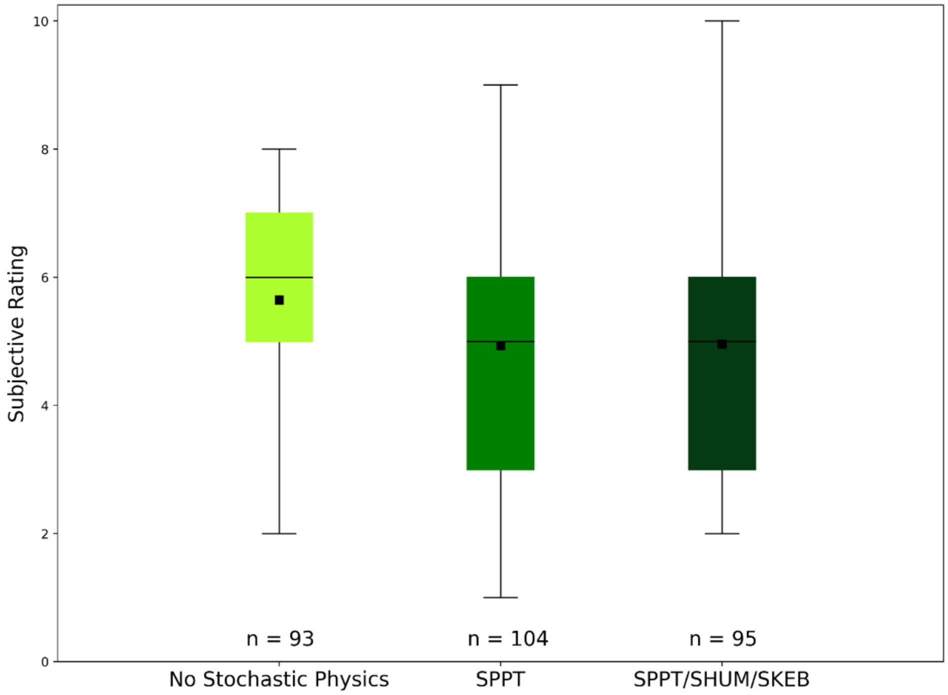


*Figure 25 Participant subjective evaluation of temperature, dewpoint, and SBCAPE for models in the FV3 Physics Suite comparison at 1800 UTC. Sample sizes for each distribution are shown below the plots, and the black square shows the mean rating of each distribution.*

*c) Model Evaluations – Group C: CAM Ensembles*

*C1) CLUE: 00Z Ensembles*

This evaluation compared four 0000-UTC initialized, FV3-LAM CAM ensembles to HREFv3: (1) GSL RRFS, (2) RRFS Cloud, (3) MAP RRFS, and (4) MAP RRFS VTS. Each of these ensembles has a unique configuration strategy, so the primary goals were to find which strategy provided the most skillful forecasts and how each performed relative to HREFv3. Note, the GSL RRFS leveraged the operational HRRRDAS for ICs. These evaluations were focused over a mesoscale area of interest with the greatest potential for severe weather over the CONUS during the convective day (i.e., 1200-1200 UTC; forecast hours 13-36). The forecast field most commonly examined during this severe weather evaluation was the 24-h summary of 2-5 km AGL hourly maximum UH. The ensemble maximum UH and neighborhood UH probabilities (>99.85th percentile) were displayed along with preliminary local storm reports (e.g., Fig. 26), and the participants rated the forecasts on a scale of 1-10 based on the quality of guidance provided to a severe weather forecaster.

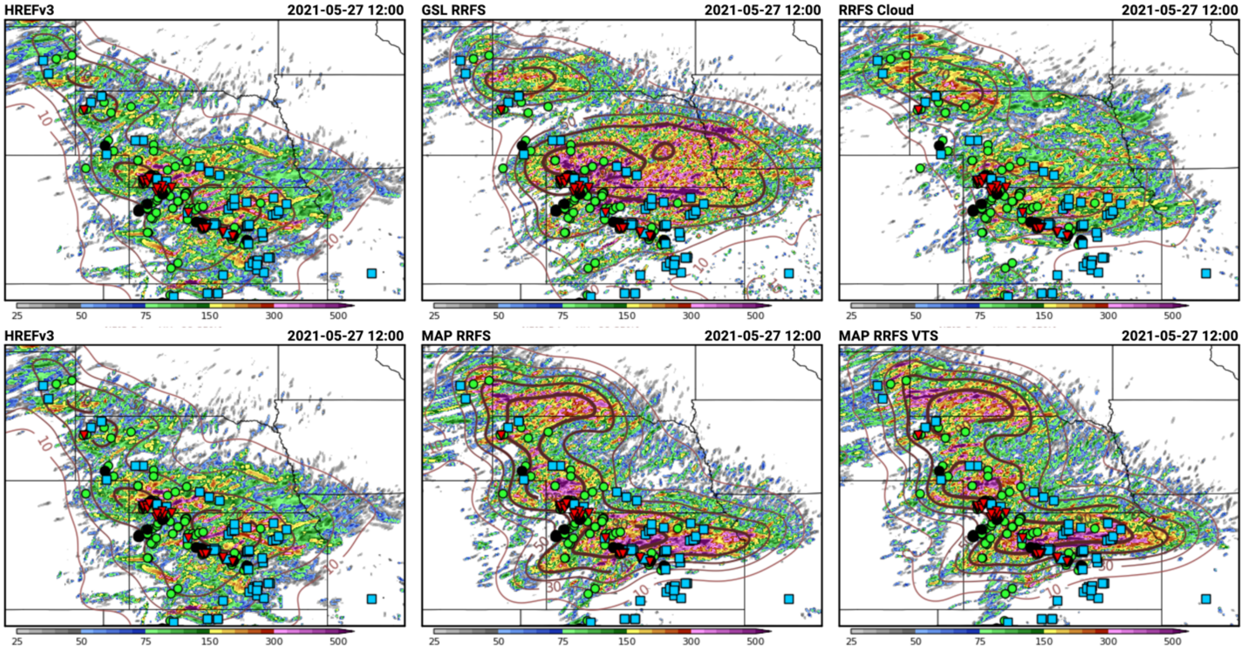


*Figure 26 Example of multi-panel comparison webpage for the 0000 UTC CAM ensemble C1 evaluation during the 2021 SFE. The 24-h ensemble maximum UH (shaded) and neighborhood probability of UH>99.85th percentile (contoured) is displayed for HREFv3 (upper left), GSL RRFS (upper middle), RRFS Cloud (upper right), HREFv3 (lower right), MAP RRFS (lower middle, and MAP RRFS VTS (lower right) for 26 May 2021. Preliminary severe storm reports are also overlaid (wind – blue squares, hail – green circles, and tornado – red upside-down triangles).*

The distribution of subjective ratings (Fig. 27) indicates that HREFv3 performed best (mean = 6.64), followed by GSL RRFS (mean = 6.32). The other systems - RRFS Cloud, MAP RRFS, and MAP RRFS VTS - trailed behind with mean ratings clustered around 5.7. Many participant comments noted that HREFv3 and GSL RRFS predicted locations of severe weather reports very accurately. There were also several comments noting that HREFv3 and RRFS Cloud had the largest spread or broadest coverage of probabilities. Finally, although comments sometimes noted that the MAP runs did quite well with storm placement and coverage, there were many instances where the MAP runs were too aggressive and/or forecast the placement of storms incorrectly.

As has been demonstrated in previous SFEs, the HREF continues to stand as a formidable baseline for experimental CAM ensembles. Although the GSL RRFS didn't perform quite as well as HREF, its performance was very encouraging, especially considering that this is the first year it has been tested in the HWT.

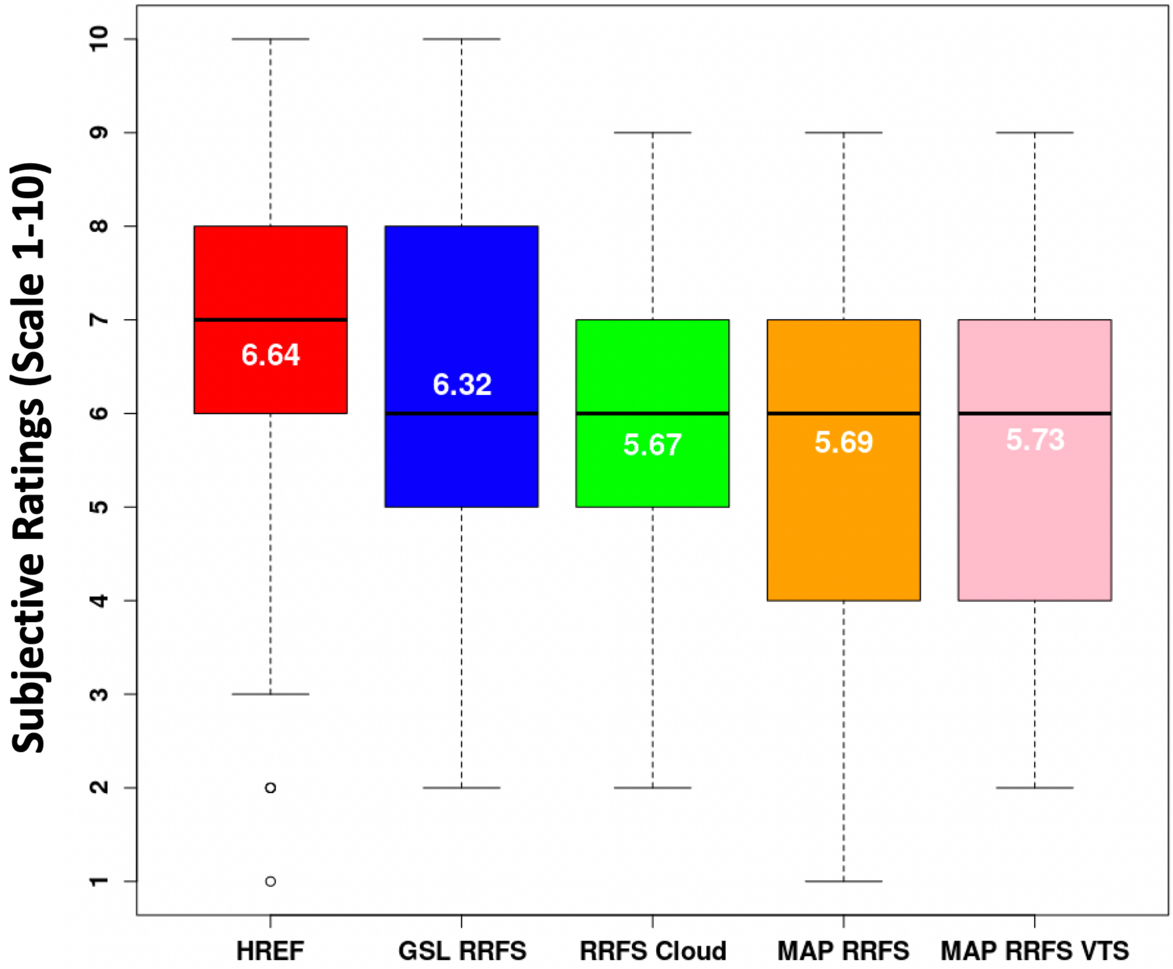


*Figure 27 Distributions of subjective ratings (1-10) by SFE participants of hourly maximum fields for severe weather forecasting over a mesoscale area of interest for the forecast hours 13-36 for the C1: CLUE 00Z Ensembles evaluation (HREF – red; GSL RRFS – blue; RRFS Cloud – green; MAP RRFS – orange; MAP RRFS VTS – pink). The numbers in white text indicate mean ratings, while the horizontal black lines indicate the median.*

### *C2) CLUE: 12Z Ensembles*

In this evaluation, three 1200-UTC initialized CAM ensembles were compared to HREFv3: (1) GSL RRFS, (2) HRRRE-S, and (3) HRRRE-M. The primary goals of this comparison were to examine how stochastic physics and multi-physics approaches compare in the HRRRE, and what effect that choice has on their performance relative to the HREFv3 and GSL RRFS. Similar to the 00Z evaluations, these were focused over a mesoscale area of interest and participants most frequently relied on 24-h maximum UH to make their subjective assessments. An example set of forecasts initialized 1200 UTC 17 May 2021 is shown in Figure 28.

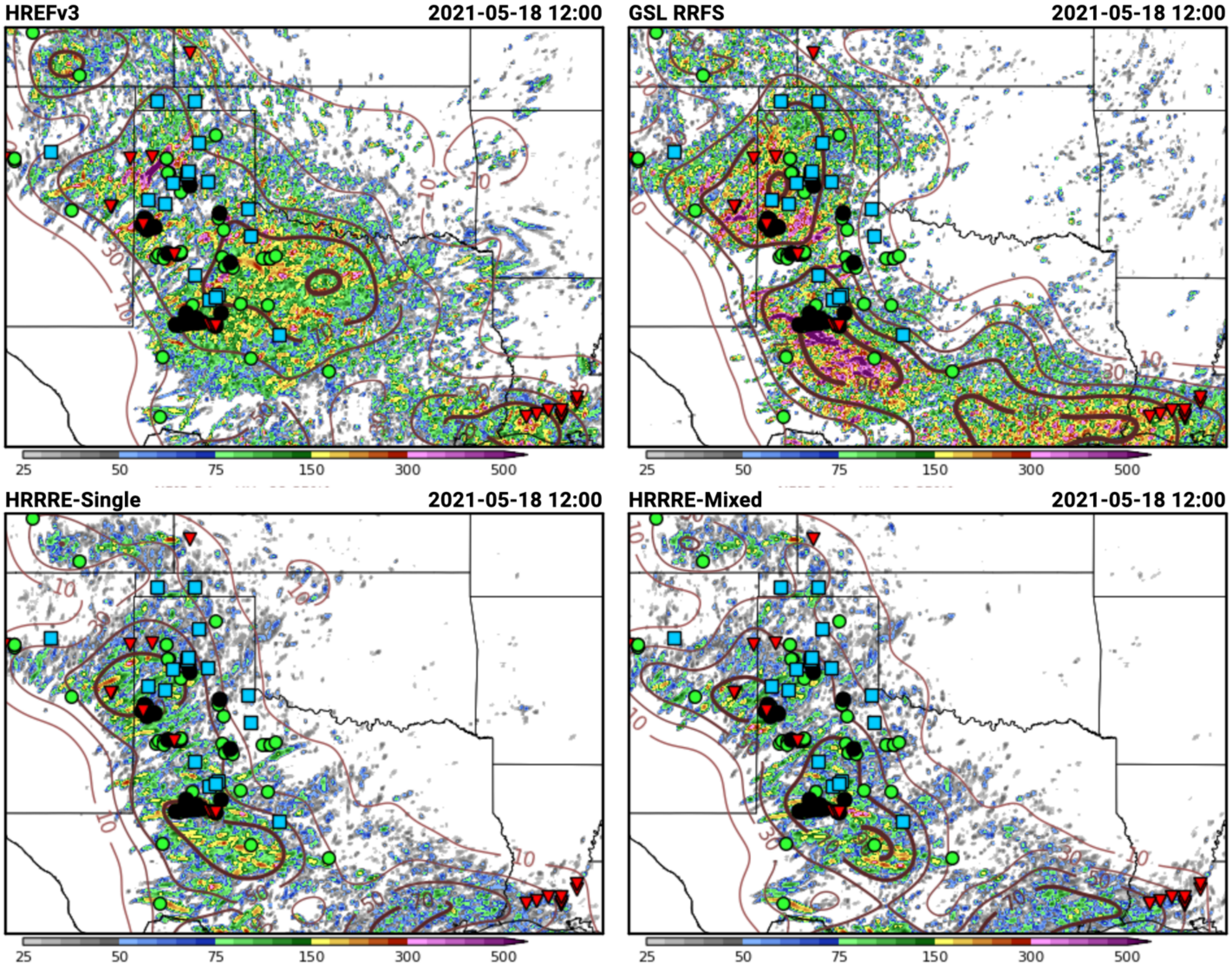


*Figure 28 Example of multi-panel comparison webpage for the 1200 UTC CAM ensemble C2 evaluation during the 2021 SFE. The 24-h ensemble maximum UH (shaded) and neighborhood probability of UH>99.85th percentile (contoured) is displayed for HREFv3 (upper left), GSL RRFS (upper right), HRRRE-S (lower left), and HRRRE-M (lower right) for 17 May 2021. Preliminary severe storm reports are also overlaid (wind – blue squares, hail – green circles, and tornado – red upside-down triangles).*

The ratings distributions in Figure 29 indicate that HREFv3 performed best, but GSL RRFS and HRRR-M were not far behind. The similar performance in GSL RRFS and HRRRE-M may be attributable to the fact that both are initialized from operational HRRRDAS analyses. The improved performance in HRRRE-M relative to HRRRE-S shows the skill that can be gained through a multi-physics ensemble configuration approach, even if it is simply two different tuned and tested physics suites. The comments often noted that the 1200 UTC initializations were markedly better than the 0000 UTC ones. Although the HREFv3 was often commended for accurately predicting the placement of storms, there were several comments that alluded to HREFv3 probabilities being too expansive, or having too much spread or false alarm.

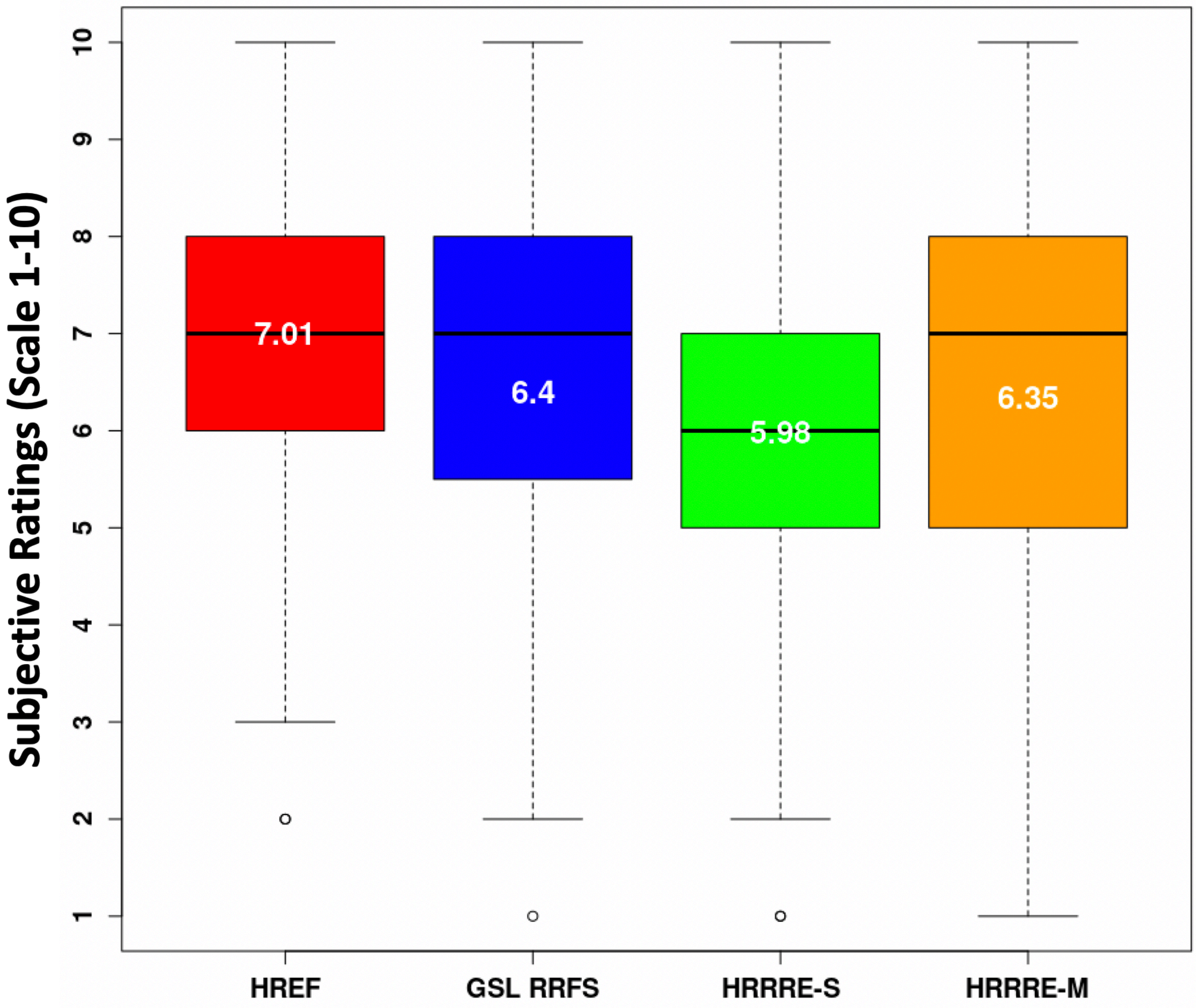


*Figure 29 Distributions of subjective ratings (1-10) by SFE participants of hourly maximum fields for severe weather forecasting over a mesoscale area of interest for the forecast hours 1-24 for the C2: CLUE 12Z Ensembles evaluation (HREF – red; GSL RRFS – blue; HRRRE-S – green; HRRRE-M – orange). The numbers in white text indicate mean ratings, while the horizontal black lines indicate the median.*

*C3) Hourly Updating CAM Ensembles*

In this evaluation, various strategies for producing CAM ensemble guidance after 12Z, but before 00Z, were examined. Specifically, five different strategies were compared to 12Z HREF: (1) 12Z HRRR-TL, (2) 15Z HRRR-TL, (3) 18Z HRRR-TL, (4) 18Z HREF/HRRR-TL Blend, and (5) 18Z Error-Weighted Blend. The HRRR-TL comprises the four most recent hourly runs of the HRRR, weighted equally. In the time-based blend, the HREF's weight is the ratio of the HRRR-TL forecast's lead time to that of the HREF forecast, so that each new run of the HRRR-TL receives linearly increasing weight as the HREF ages. For example, the 18Z blend valid at 00Z is 50% 12Z HREF, 50% 18Z HRRR-TL. The error-based blend combines the 10 HREF members and 4 HRRR-TL members into a single ensemble. Each member receives weight based on the sum of its normalized domain-wide RMSEs in 2-m temperature, 2-m dewpoint, and 10-m wind component fields using RTMA at the blend initialization time as truth. The member with the largest errors on the SFE domain receives no weight and the member with the smallest errors receives maximum weight. This was motivated by a set of test cases in which these short-term errors were weakly negatively correlated with convective forecast skill at later times. The primary goal of this evaluation was to assess whether there are optimal ways to produce updated and improved CAM ensemble guidance within the Day 1 forecast period in between HREF updates (i.e., available at 15Z to 03Z daily). An example case from 4 May 2021, which was used in one of the evaluations, is shown in Figure 30.

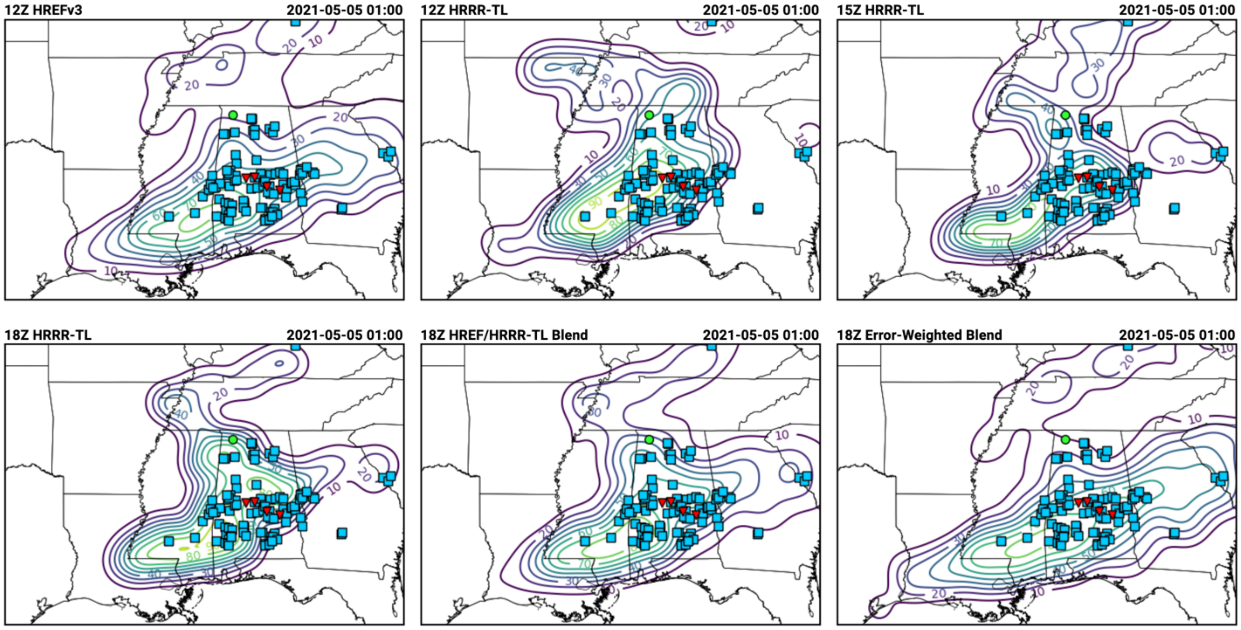


*Figure 30 Neighborhood maximum ensemble probability of 2-5 km AGL UH > 75 $m^2s^{-2}$ (contours) covering the 4-h period ending 0100 UTC 5 May 2021 for forecasts initialized 0000 UTC 4 May 2021, with preliminary severe storm reports overlaid (wind – blue squares, hail – green circles, and tornado – red upside-down triangles): 12Z HREFv3 (upper left), 12Z HRRR-TL (upper middle), 15Z HRRR-TL (upper right), 18Z HRRR-TL (lower left), 18Z HREF/HRRR-TL Blend (lower middle), and 18Z Error-Weighted Blend (lower right).*

The subjective ratings distributions (Fig. 31) indicate that the 18Z Blend and 18Z Blend (error weighted) clearly performed better than the probabilities derived from HRRR-TL. Although the 18Z Blend (error weighted) had a slightly higher mean rating than the 18Z Blend, their ratings distributions were almost identical. Several comments noted that the two 18Z blends seemed to improve upon the HRRR-TL configurations because of how they smoothed out over-confident areas.

As expected, each of the more recent HRRR-TL initializations had slightly higher ratings than the previous initializations, although the comments noted that there were exceptions when the older runs performed better. These results indicate that the blending strategies could be a useful strategy for generating skillful probabilistic guidance in between the times that new HREFv3 forecasts become available.

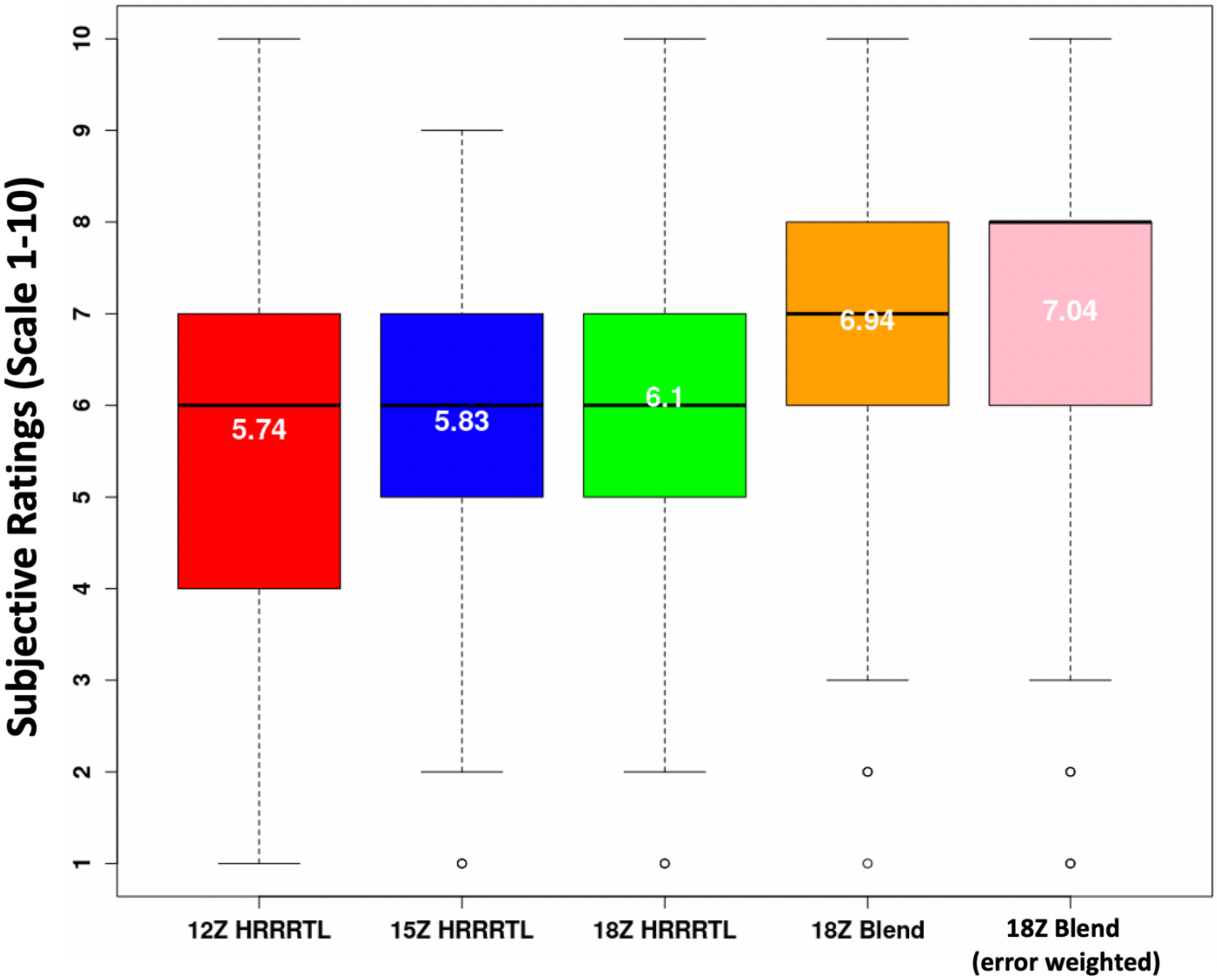


*Figure 31 Distributions of subjective ratings (1-10) by SFE participants of hourly maximum fields for severe weather forecasting over a mesoscale area of interest valid 2200 – 0300 UTC for the C3: Hourly Updating CAM Ensemble evaluation (12Z HRRR-TL – red; 15Z HRRR-TL – blue; 18Z HRRR-TL – green; 18Z Blend – orange; 18Z error-weighted Blend - pink). The numbers in white text indicate mean ratings, while the horizontal black lines indicate the median.*

### *C4) CLUE: VTS DA*

Ensembles initialized at 2100 and 0000 UTC with and without the valid-time-shifting (VTS) data assimilation strategy were compared for the first 12 hours of the forecasts in this evaluation. VTS is a cost-effective way to increase the membership (by a factor of three) for the background ensemble in convective scale, hybrid EnVar data assimilation. The increased membership is achieved by populating the background ensemble with analyses valid at slightly different times. Thus, the primary goal of this evaluation was to assess whether the VTS strategy results in improved performance during the first 12 hours of the forecast. Similar to the other Group C evaluations, these were focused over a mesoscale area of interest. Participants most frequently relied on 4-h maximum UH to make their subjective assessments. An example set of forecasts initialized 0000 UTC 3 May 2021 is shown in Figure 32.

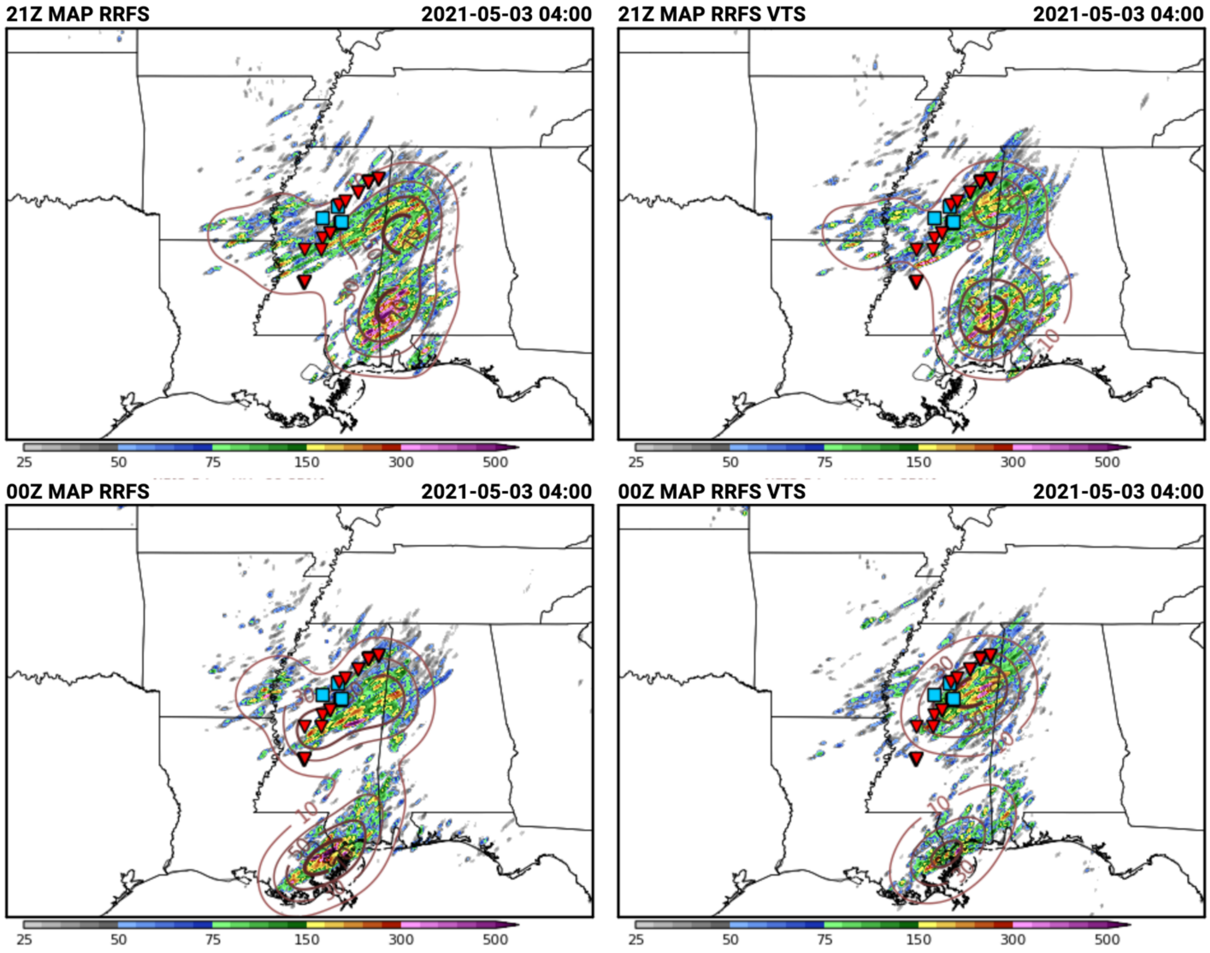


*Figure 32 Example of multi-panel comparison webpage for the CLUE: VTS DA C4 CAM Ensemble evaluation during the 2021 SFE. The 4-h ensemble maximum UH (shaded) and neighborhood probability of UH>99.85th percentile (contoured) is displayed for 21Z MAP RRFS (upper left), 21Z MAP RRFS VTS (upper right), 00Z MAP RRFS (lower left), and 00Z MAP RRFS VTS (lower right) for 17 May 2021. Preliminary severe storm reports are also overlaid (wind – blue squares, hail – green circles, and tornado – red upside-down triangles).*

The subjective ratings distributions (Fig. 33) indicate the difference in performance between the pairs of runs with and without VTS were very small, although the differences were in the direction of very slight improvements for VTS. Additionally, the 0000 UTC initialized runs were slightly better than the 2100 initialized runs, which is expected given slightly shorter forecast lead times in the 0000 UTC initializations. The comments generally reflected the similar performance in the runs with and without VTS, and several participants noted that the differences between the 2100 and 0000 UTC initialization times were bigger than the differences in runs with and without VTS. The small impact of VTS on the forecasts may not be too surprising since the VTS is only directly applied to the control member initial conditions. The other members are only indirectly impacted by VTS in that their perturbations are recentered around the control member. Since the VTS does not degrade the forecasts, and may result in very slight improvements, the fact that it could save computational resources makes it a technique with potential applications in future CAM ensembles. However, an important caveat here is that when evaluated at longer lead times alongside HREFv3 and other experimental ensembles (Fig. 27), the MAP runs were not rated as high as other CAM esnembles; thus, additional work is needed.

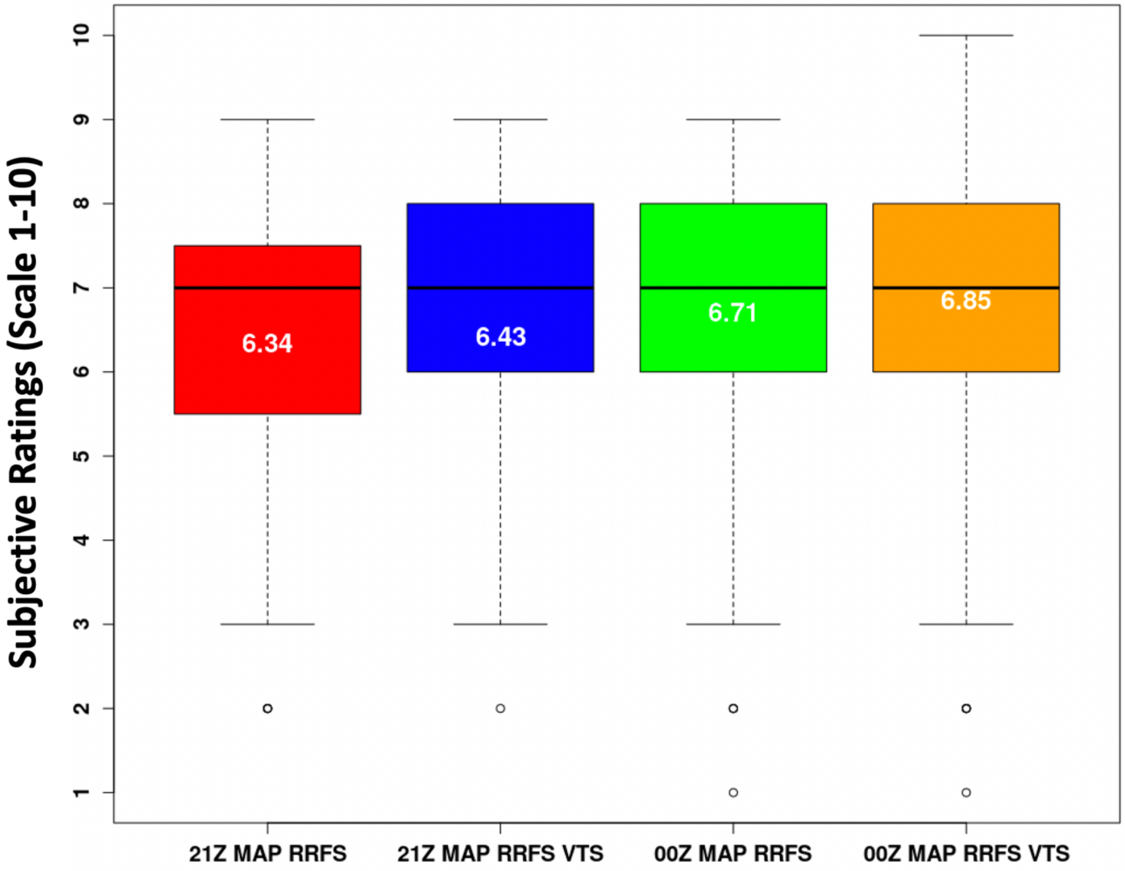


*Figure 33 Distributions of subjective ratings (1-10) by SFE participants of hourly maximum fields for severe weather forecasting over a mesoscale area of interest for the forecast hours 0-12 for the C4: CLUE VTS DA Ensembles evaluation (21Z MAP RRFS – red; 21Z MAP RRFS VTS – blue; 00Z MAP RRFS – green; 00Z MAP RRFS VTS – orange). The numbers in white text indicate mean ratings, while the horizontal black lines indicate the median.*

### *C5) WoFS Evaluations*

#### *i) WoFS vs. HRRR-TL*

WoFS initializations at 2100 and 2300 UTC were compared to HRRR-TL ensembles based at the same times. The goal of this evaluation was to assess how the WoFS performs relative to systems that are currently available in operations. Like the other evaluations, these focused over a mesoscale area of interest. Within the web-viewer, participants were instructed to enable an overlay of the WoFS domain bounds and only consider the forecast performance enclosed within that region. Participants relied on 4-h and 1-h fields of maximum UH, as well as paintball plots of simulated composite reflectivity > 40 dBZ to make their ratings assessments. An example set of forecasts from 14 May 2021 is shown in Figure 34.

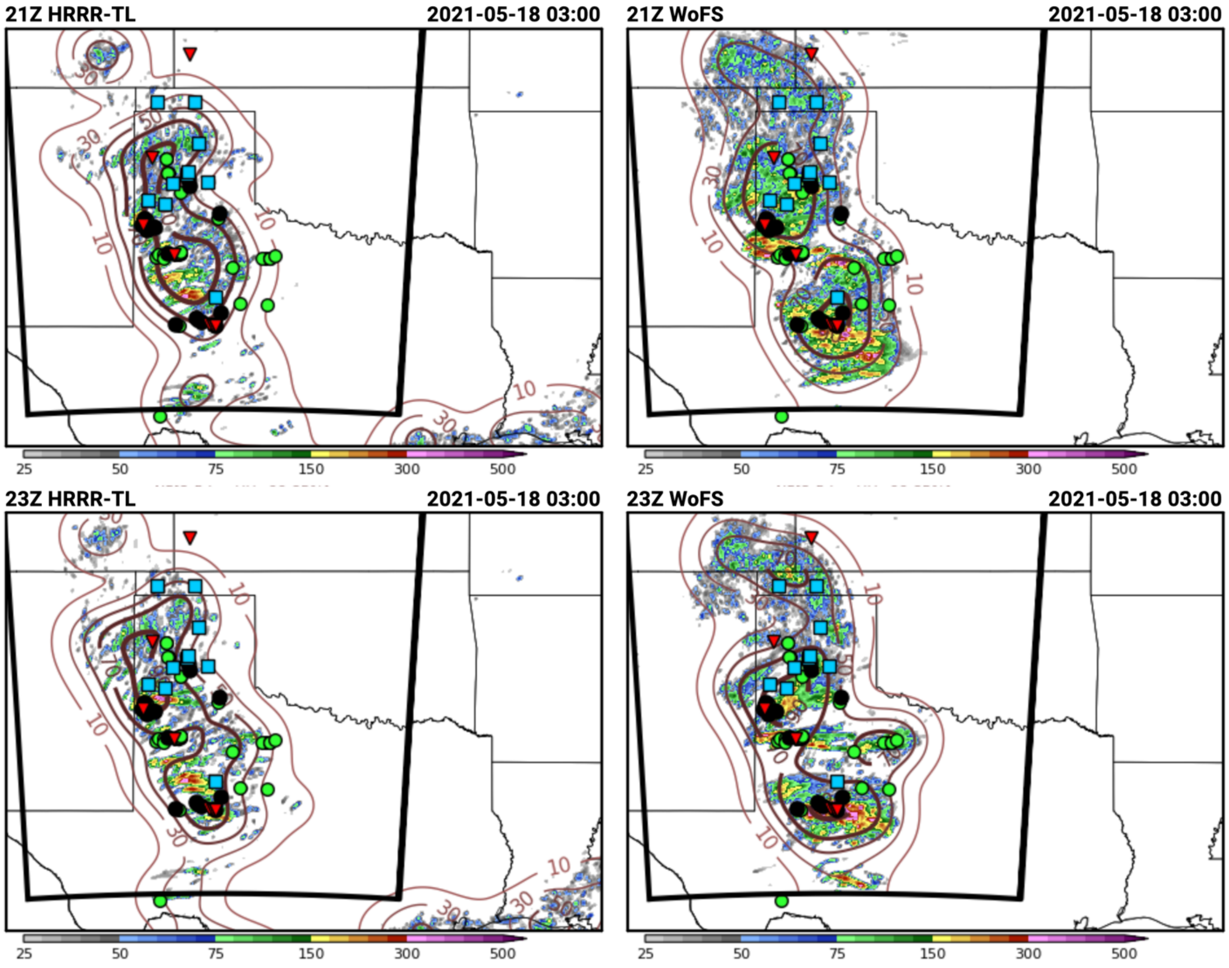


*Figure 34 Example of multi-panel comparison webpage for the C5a WoFS vs. HRRR-TL evaluation during the 2021 SFE. The 4-h ensemble maximum UH (shaded) and neighborhood probability of UH>99.85th percentile (contoured) is displayed for 21Z HRRR-TL (upper left), 21Z WoFS (upper right), 23Z HRRR-TL (lower left), and 23Z WoFS (lower right) for 17 May 2021. Preliminary severe storm reports are also overlaid (wind – blue squares, hail – green circles, and tornado – red upside-down triangles).*

The subjective ratings distributions (Fig. 35) indicate slightly better performance for WoFS, especially for the 23Z initializations. The more dramatic differences for the 23Z initializations might be attributable to there being more ongoing convection at that time, for which the rapid DA of WoFS is better equipped to realistically depict and evolve. From the comments, it was clear that the main advantage of WoFS was its ability to highlight very specific threat areas. For example, one participant noted, "*WoFS seemed to be aggressive on intensity of the storms, but pinpointed locations quite well, especially the 23Z WoFS run*". Another comment stated, "*The WoFS was able to identify more bullseyes, which were more accurate than the broad areas highlighted by the HRRR-TL*", and finally, "*The corridor of higher probabilities by WoFS tended to be better placed than the HRRR-TL (at least relative to where reports and MESH tracks were located)*". These results confirm the potential value provided by WoFS if it were available on a regular basis to forecasters.

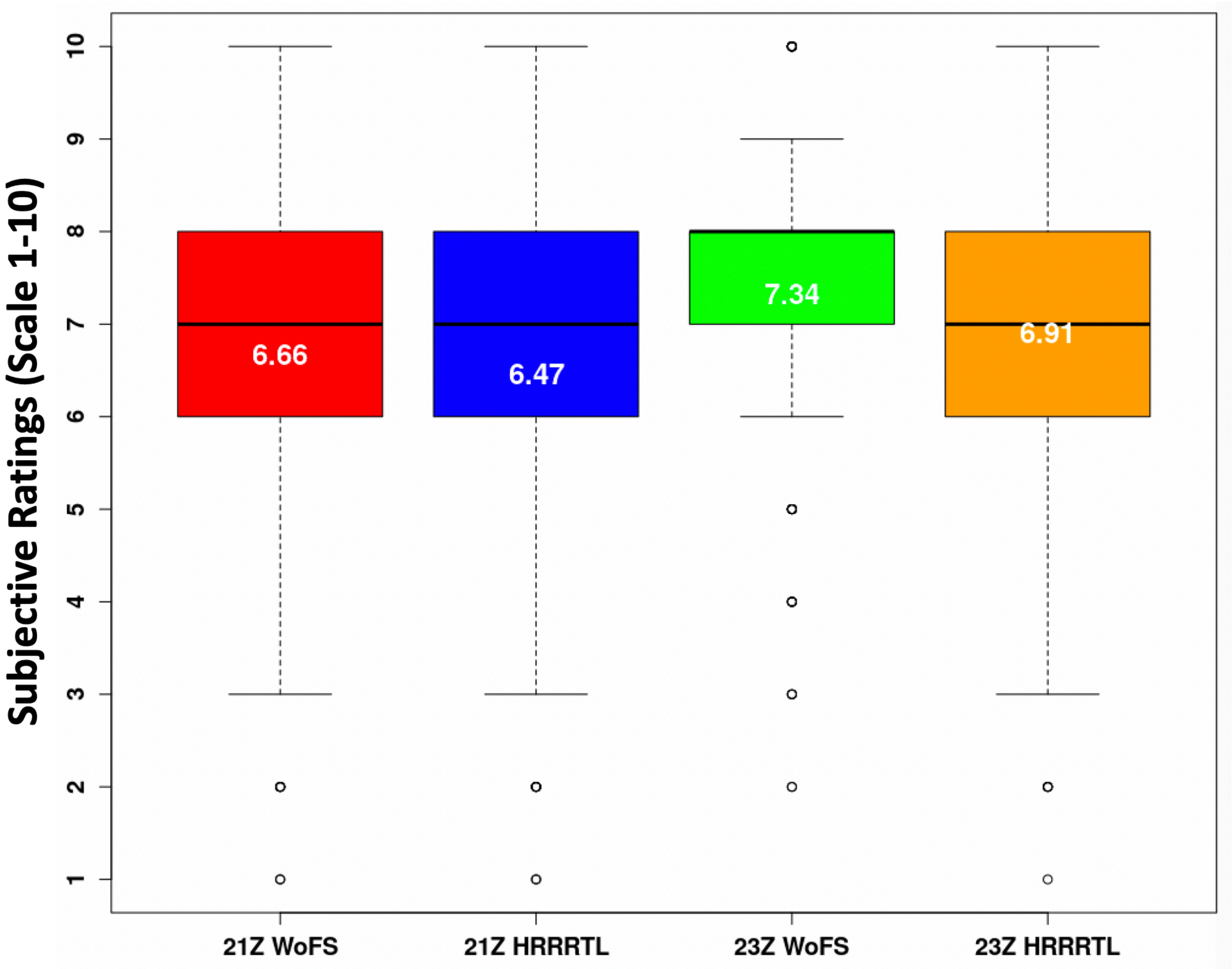
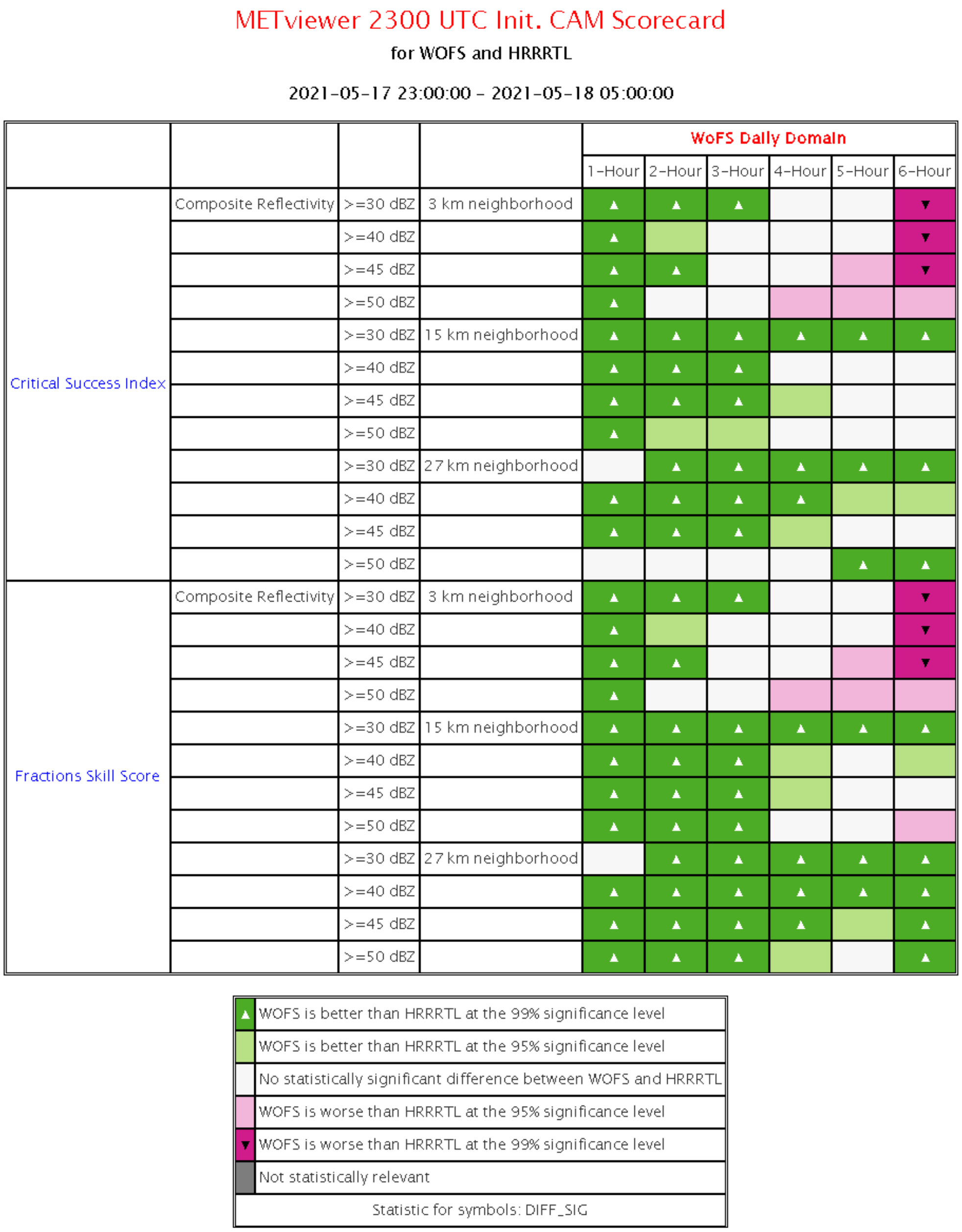


*Figure 35 Distributions of subjective ratings (1-10) by SFE participants of hourly maximum fields for severe weather forecasting over a mesoscale area of interest for the forecast hours 0-6 for the C5a WoFS vs. HRRR-TL ensemble evaluation (21Z WoFS – red; 21Z HRRR-TL – blue; 23Z WoFS – green; 23Z HRRR-TL – orange). The numbers in white text indicate mean ratings, while the horizontal black lines indicate the median.*

Another component of the WoFS vs. HRRR-TL comparisons involved application of a CAM scorecard, which displays objective information on the relative performance of two modeling systems. MET was used to construct these scorecards, and critical success index (CSI) and fractions skill score (FSS) were computed for several thresholds of composite reflectivity using neighborhood sizes of 3-km 15-, and 27-km. These scores were computed every hour at 0-6 h lead times for both WoFS and HRRR-TL. The symbols and colors in the scorecard indicate the direction and significance level of the differences in each metric. An example scorecard for WoFS and HRRR-TL 2300 UTC initializations on 17 May 2021 is shown in Figure 36.



METviewer 2300 UTC Init. CAM Scorecard

for WOFS and HRRRTL

2021-05-17 23:00:00 - 2021-05-18 05:00:00

| | | | | WoFS Daily Domain | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | 1-Hour | 2-Hour | 3-Hour | 4-Hour | 5-Hour | 6-Hour |
| Critical Success Index | Composite Reflectivity | >=30 dBZ | 3 km neighborhood | ▲ | ▲ | ▲ | □ | □ | ▼ |
| | | >=40 dBZ | | ▲ | 🟩 | □ | □ | □ | ▼ |
| | | >=45 dBZ | | ▲ | ▲ | □ | □ | 🟥 | ▼ |
| | | >=50 dBZ | | ▲ | □ | □ | 🟥 | 🟥 | 🟥 |
| | | >=30 dBZ | 15 km neighborhood | ▲ | ▲ | ▲ | ▲ | ▲ | ▲ |
| | | >=40 dBZ | | ▲ | ▲ | ▲ | □ | □ | □ |
| | | >=45 dBZ | | ▲ | ▲ | ▲ | 🟩 | □ | □ |
| | | >=50 dBZ | | ▲ | 🟩 | 🟩 | □ | □ | □ |
| | | >=30 dBZ | 27 km neighborhood | □ | ▲ | ▲ | ▲ | ▲ | ▲ |
| | | >=40 dBZ | | ▲ | ▲ | ▲ | ▲ | 🟩 | 🟩 |
| | | >=45 dBZ | | ▲ | ▲ | ▲ | 🟩 | □ | □ |
| | | >=50 dBZ | | □ | □ | □ | □ | ▲ | ▲ |
| Fractions Skill Score | Composite Reflectivity | >=30 dBZ | 3 km neighborhood | ▲ | ▲ | ▲ | □ | □ | ▼ |
| | | >=40 dBZ | | ▲ | 🟩 | □ | □ | □ | ▼ |
| | | >=45 dBZ | | ▲ | ▲ | □ | □ | 🟥 | ▼ |
| | | >=50 dBZ | | ▲ | □ | □ | 🟥 | 🟥 | 🟥 |
| | | >=30 dBZ | 15 km neighborhood | ▲ | ▲ | ▲ | ▲ | ▲ | ▲ |
| | | >=40 dBZ | | ▲ | ▲ | ▲ | 🟩 | □ | 🟩 |
| | | >=45 dBZ | | ▲ | ▲ | ▲ | 🟩 | □ | □ |
| | | >=50 dBZ | | ▲ | ▲ | ▲ | □ | □ | 🟥 |
| | | >=30 dBZ | 27 km neighborhood | □ | ▲ | ▲ | ▲ | ▲ | ▲ |
| | | >=40 dBZ | | ▲ | ▲ | ▲ | ▲ | ▲ | ▲ |
| | | >=45 dBZ | | ▲ | ▲ | ▲ | ▲ | 🟩 | ▲ |
| | | >=50 dBZ | | ▲ | ▲ | ▲ | 🟩 | □ | ▲ |

| | |
|---|---|
| ▲ | WOFS is better than HRRRTL at the 99% significance level |
| 🟩 | WOFS is better than HRRRTL at the 95% significance level |
| □ | No statistically significant difference between WOFS and HRRRTL |
| 🟥 | WOFS is worse than HRRRTL at the 95% significance level |
| ▼ | WOFS is worse than HRRRTL at the 99% significance level |
| ⬛ | Not statistically relevant |
| | Statistic for symbols: DIFF_SIG |

*Figure 36 Example CAM scorecard for WoFS and HRRR-TL 2300 UTC initialization on 17 May 2021. A legend is provided at the bottom.*

For the scorecard evaluation, participants were asked to comment on whether the scorecard reflects your subjective impression of model performance. Specifically, the participants were asked to consider their ratings from the preceding evaluation and then examine the scorecards for the corresponding set of 2100 and 2300 UTC WoFS and HRRR-TL initializations. Since this was an open-ended question (i.e., comment box), each response was manually classified as "yes", "no", "mixed", or "not sure". The response frequencies (Fig. 37) indicate that participants generally believed that the scorecard results reflected their subjective impressions, although there were a notable number of instances where it did not. Also, there were many responses classified as "mixed". In these cases, participants indicated that for some of the lead times, thresholds, or metrics the scorecard matched their subjective impressions, but for others it did not match. Participant comments often noted that the differences seemed to be very dependent on the neighborhood size.

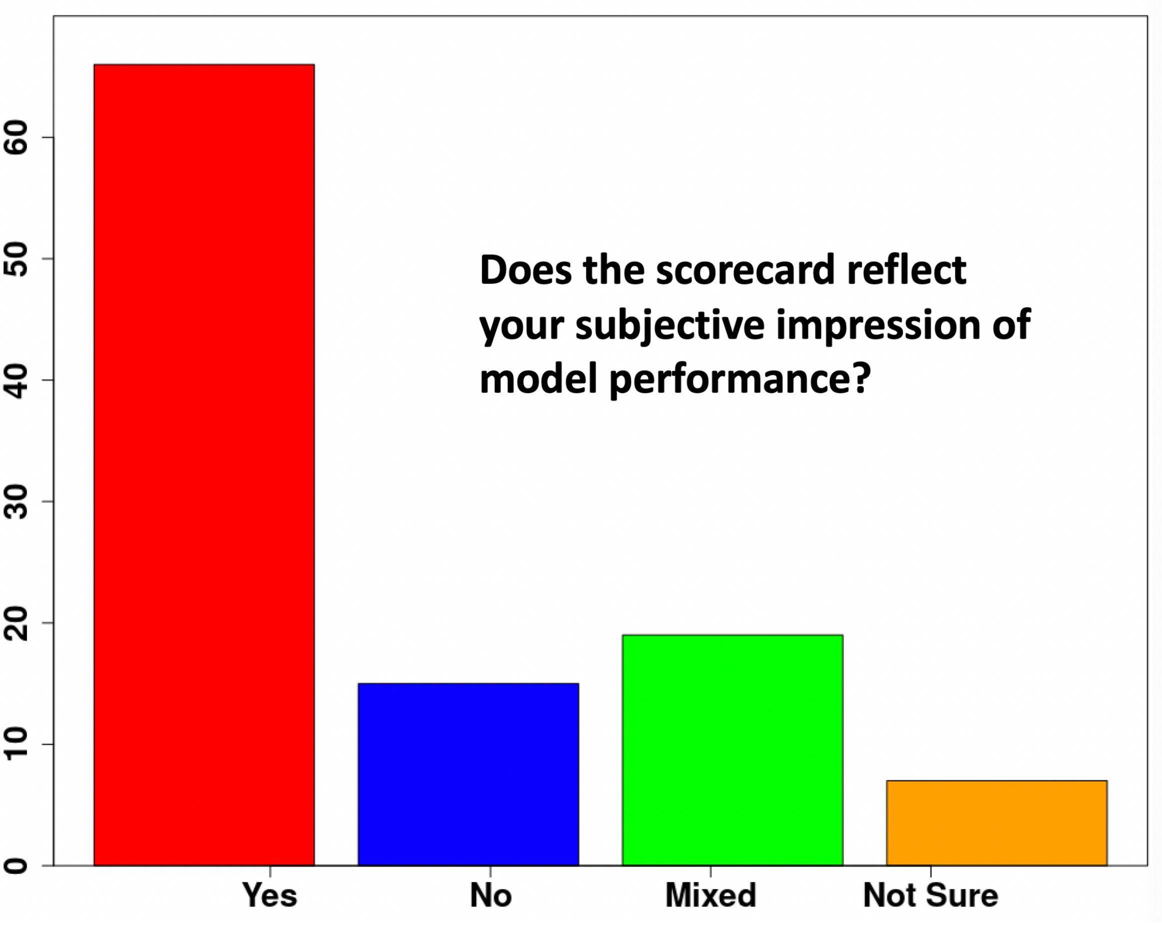


*Figure 37 Response frequencies for the CAM scorecard evaluation. Specifically, participants were asked, "Does the scorecard reflect your subjective impression of model performance?".*

Finally, participants were asked, "What forecast fields do you think are the most important to have on a CAM ensemble scorecard?". This was also an open-ended question, so participants could choose as many fields as they wanted. Thus, the comments were coded as one of the categories indicated in Figure 38. Note, participants often mentioned different versions of the same field; in these instances, they were assigned to the same classifier. For example, several different types of reflectivity were mentioned (e.g., low-level, composite, maximum, etc.), so these were all assigned to the "Reflectivity" classification. By far, the fields most frequently mentioned were reflectivity and updraft helicity. Other storm-attribute fields such as precipitation, max winds, hail, and updraft speed were also mentioned quite frequently, but environmental fields were rarely mentioned (Fig. 38).

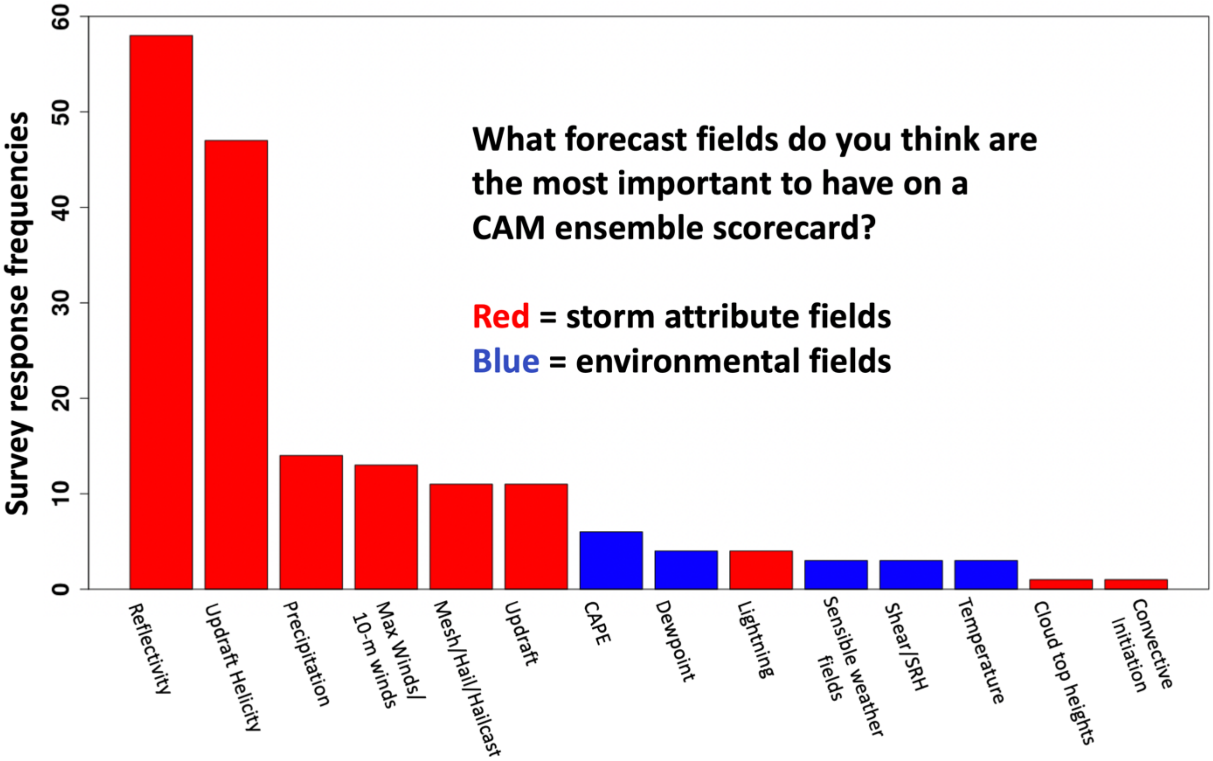


*Figure 38 Response frequencies for the CAM scorecard evaluation. Specifically, participants were asked, "What forecast fields do you think are the most important to have on a CAM ensemble scorecard?".*

*ii) Deterministic WoFS*

These comparisons examined the WoFS deterministic 1.5 km grid-spacing hybrid data assimilation runs initialized at 2100 and 2300 UTC, which were compared to a random member from the 3-km baseline WoFS configuration with the same physics configuration. The WoFS forecast viewer was used for additional comparisons between the WoFS and WoFS-Hybrid systems. The goal of this evaluation was to assess whether the WoFS-Hybrid, on average, performs better than an individual member of WoFS, and if the WoFS-Hybrid provides additional value relative to WoFS. Simulated composite reflectivity and UH was compared over a mesoscale region of interest to the MRMS composite reflectivity and LSRs. An example is shown Figure 39 for forecasts initialized 17 May 2021.

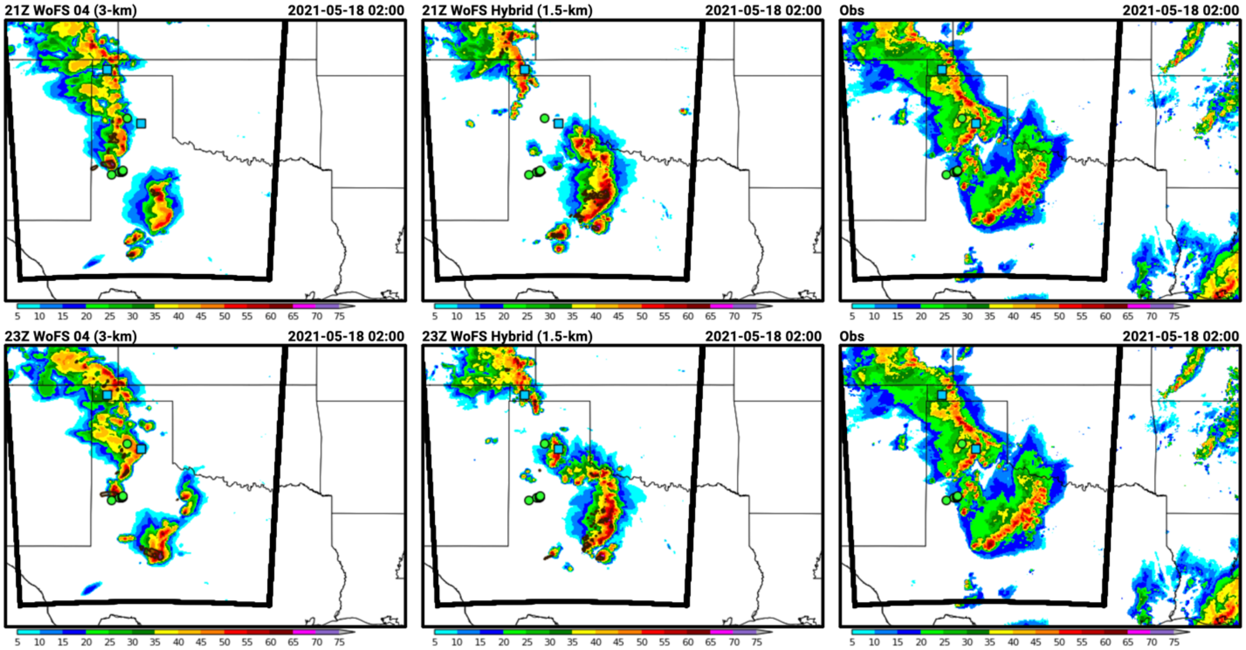


*Figure 39 Example of multi-panel comparison webpage for the Deterministic WoFS C5b evaluation during the 2021 SFE. Simulated composite reflectivity (shaded) and UH (black contours) is displayed for member 04 of the 2100 UTC WoFS (upper left), the 2100 UTC WoFS-Hybrid (upper middle), member 04 of the 2300 UTC WoFS (lower left), and the 2300 UTC WoFS-Hybrid (lower middle). The two panels on the right display MRMS composite reflectivity. The black box in each panel shows the WoFS domain. Preliminary severe storm reports are also overlaid (wind – blue squares, hail – green circles, and tornado – red upside-down triangles).*

The subjective ratings distributions (Fig. 40) indicate that for 2100 UTC initializations, the WoFS-Hybrid performed slightly better than the random member of WoFS. However, for the 2300 UTC initializations the WoFS and WoFS-hybrid forecasts received similar ratings, on average. It's possible that the WoFS-Hybrid is able to better depict storms in their early development stages compared to WoFS, and that this is reflected in the better ratings for the 2100 UTC compared to 2300 UTC initializations, which would be more likely to have more mature storms.

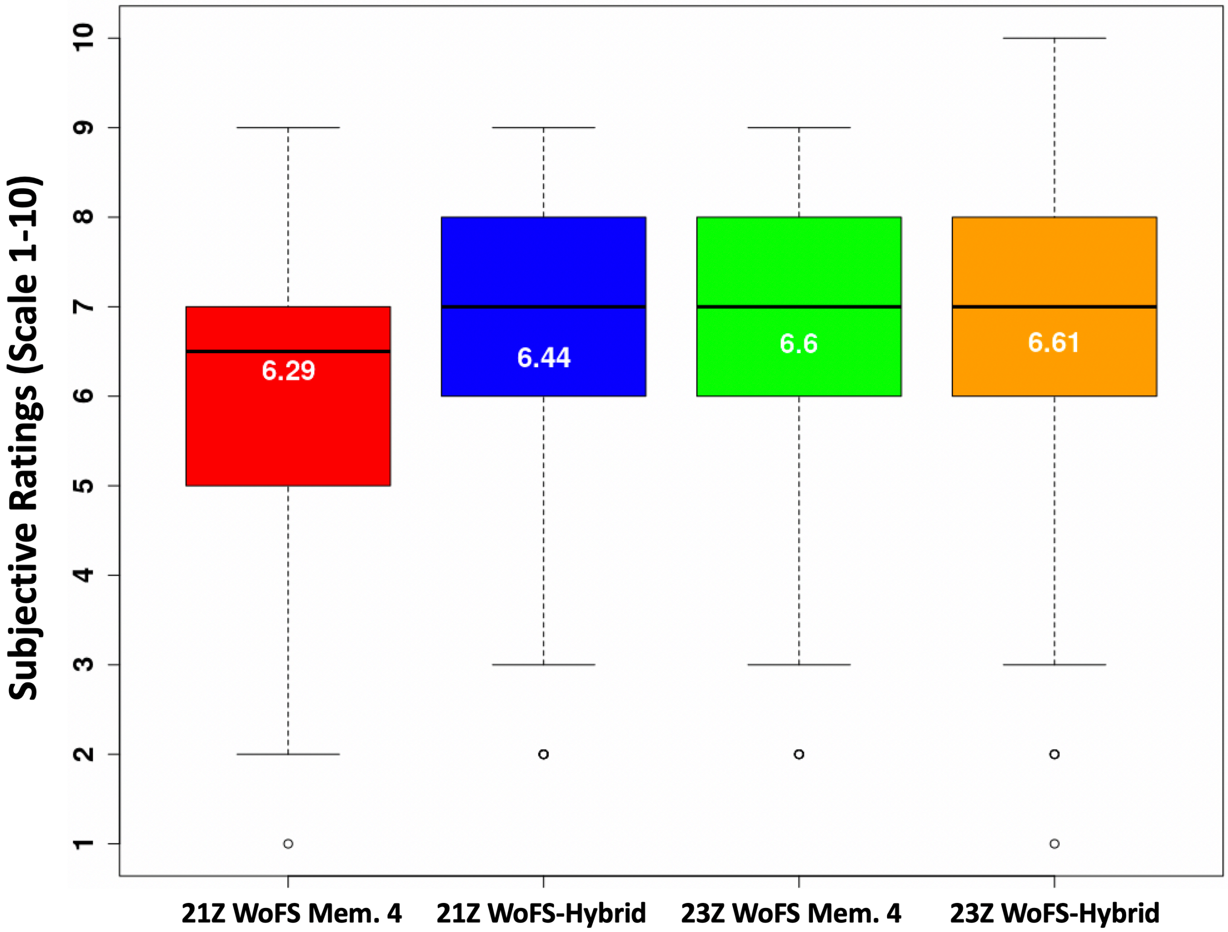


*Figure 40 Distributions of subjective ratings (1-10) by SFE participants of composite reflectivity and UH over a mesoscale area of interest for the forecast hours 0-6 for the C5b Deterministic WoFS evaluation (21Z WoFS Mem. 4 – red; 21Z WoFS-Hybrid – blue; 23Z WoFS Mem. 4 – green; 23Z WoFS-hybrid – orange*). *The numbers in white text indicate mean ratings, while the horizontal black lines indicate the median.*

In the next component of the Deterministic WoFS evaluation, participants were given the following set of instructions: "*Navigate to the WoFS real-time web viewer and select the date for yesterday's WoFS forecasts. Then, choose a run that captures the major severe activity of the day. Next, bring up the member viewer by going to the Composite reflectivity under the Rad/Sat menu. At the bottom right of the page under the 'Verification' heading, there are several options for overlaying MRMS-based observations on top of the member viewer forecasts. Use these verification overlays and peruse the members giving particular attention to the WoF Hybrid member and its performance relative to the 'regular' members. Remember, at any time the keyboard shortcut 'o' allows you to flip the MRMS reflectivity on and off, which is very useful for comparing reflectivity 'snapshots' at various times.*" Then, based on their observations, the participants were asked to evaluate whether the WoFS-Hybrid provides value relative to the WoFS in terms of storm depiction and storm location, and whether its forecasts fall within the envelope of the WoFS members. An example of the imagery from the WoFS web-viewer for this comparison is shown in Figure 41. In this particular case, the WoFS-Hybrid member was slightly slow in moving a line of storms to the east, while the 3-km WoFS members were slightly faster and thus had better placement of the observed storms.

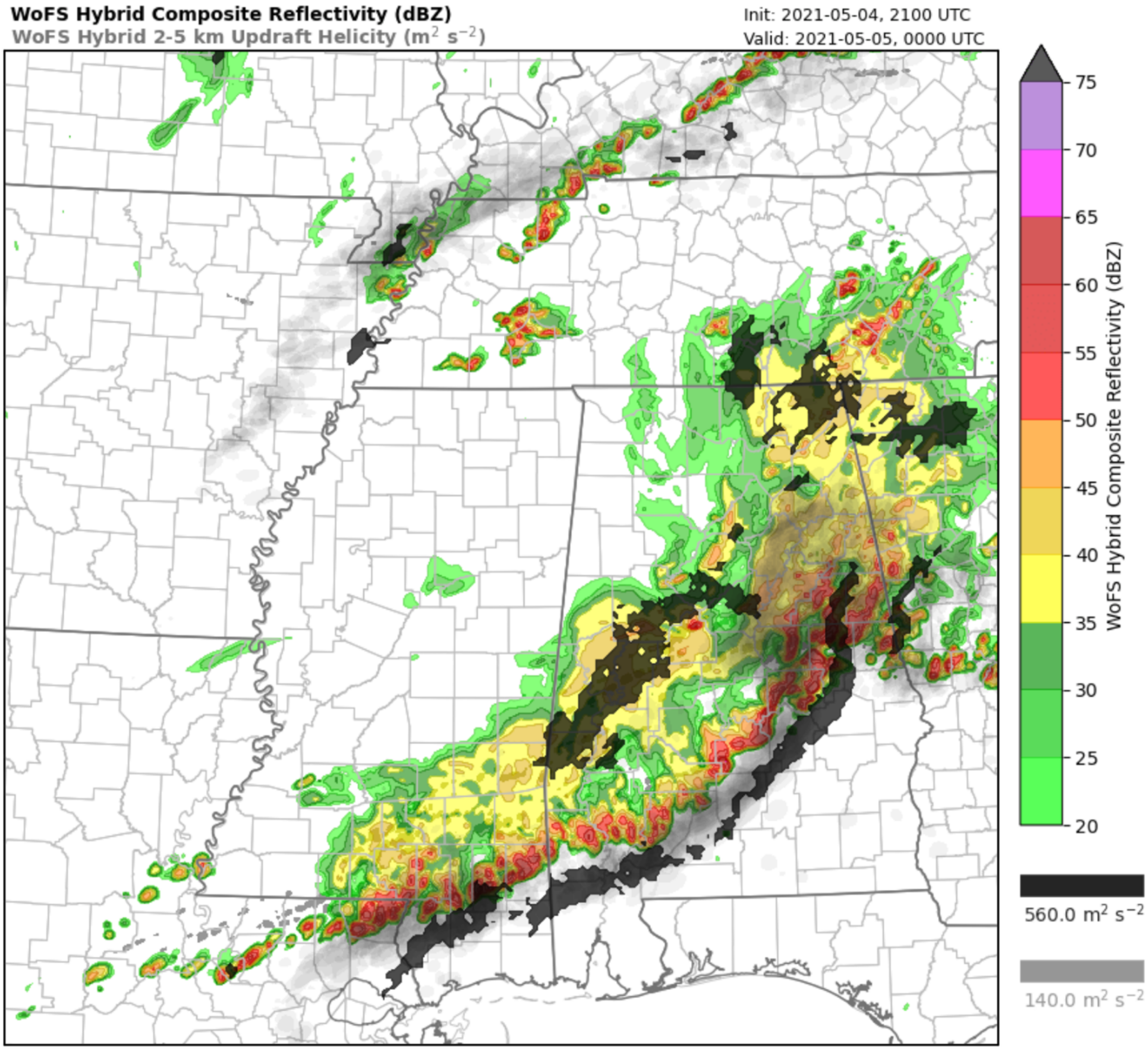

WoFS Hybrid Composite Reflectivity (dBZ)
WoFS Hybrid 2-5 km Updraft Helicity (m² s⁻²)
Init: 2021-05-04, 2100 UTC
Valid: 2021-05-05, 0000 UTC
75
70
65
60
55
50
45
40
35
30
25
20
WoFS Hybrid Composite Reflectivity (dBZ)
560.0 m² s⁻²
140.0 m² s⁻²

*Figure 41 Forecast of simulated composite reflectivity at 3-h lead time initialized from 2100 UTC of the WoFS-Hybrid (shaded). The slightly transparent black shaded regions indicate MRMS composite reflectivity > 40 dBZ. The light gray shaded regions indicate where members of the 3-km WoFS forecast composite reflectivity > 40 dBZ.*

The response frequencies indicated that, most of the time, participants believed that the WoFS-Hybrid member provided value in terms of storm structure. However, it was also quite common for participants to indicate that WoFS-Hybrid did not provide value or they were not sure (Fig. 42a). For storm location, the most frequently chosen response was that WoFS-hybrid did not provide value, but there were also many cases in which it did (Fig. 42b). Finally, when asked whether WoFS-Hybrid forecasts fell within the envelope of the 3-km WoFS members, participants most frequently selected "Within", but this was closely followed by "Outside" (Fig. 42c). From the comments, there were some instances in which it was noted that the WoFS-Hybrid forecasts fell outside the envelope, resulting in an improvement, but other situations in which falling outside the envelope resulted in degradation. In general, the results could be interpreted as positive for the WoFS-Hybrid runs, but further work is needed to gauge their utility. Also, there is a clear problem with spurious convection in the WoFS-Hybrid analyses, so this should be a high priority area for improving the forecasts.

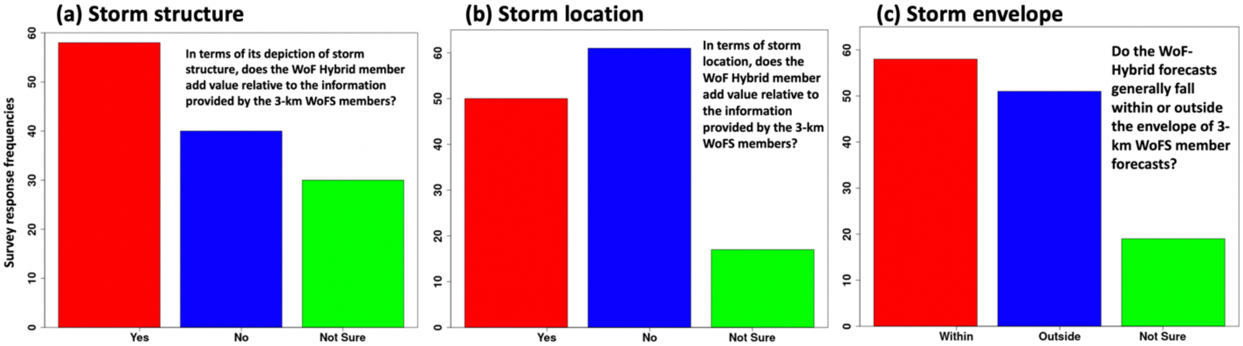


*Figure 42 Response frequencies for (a) "In terms of its depiction of storm structure, does the WoFS Hybrid member add value relative to the information provided by the 3-km WoFS members.", (b) "In terms of storm location, does the WoFS-Hybrid member add value relative to the information provided by the 3-km WoFS members, and (c) "Do the WoFS-Hybrid forecasts generally fall within or outside the envelope of 3-km WoFS member forecasts?".*

*iii) Machine-Learning Calibrated WoFS probabilities*

Hazard probabilities were derived using predictors from the WoFS output. The activity examined the utility of these probabilities and participants were given the opportunity to comment on the visualization strategy within the WoFS web-viewer. The primary goal was to assess whether these probabilities provide value on top of the already available WoFS guidance products. Before participants conducted this evaluation, a short primer was given by the developers of the product. For the evaluation, participants were instructed, "*In the WoFS web viewer (linked to above), select the date for yesterday's WoFS forecasts. Then, choose the 23:00 run from the drop-down menu at the top of the page. Next, use the 'ML products' tab at the top to examine the tornado, wind, and hail probabilities using logistic regression with a 3-km neighborhood. Overlay the local storm reports using the toggle on the right-hand-side.*" Example imagery from the web-viewer is shown in Figure 43.

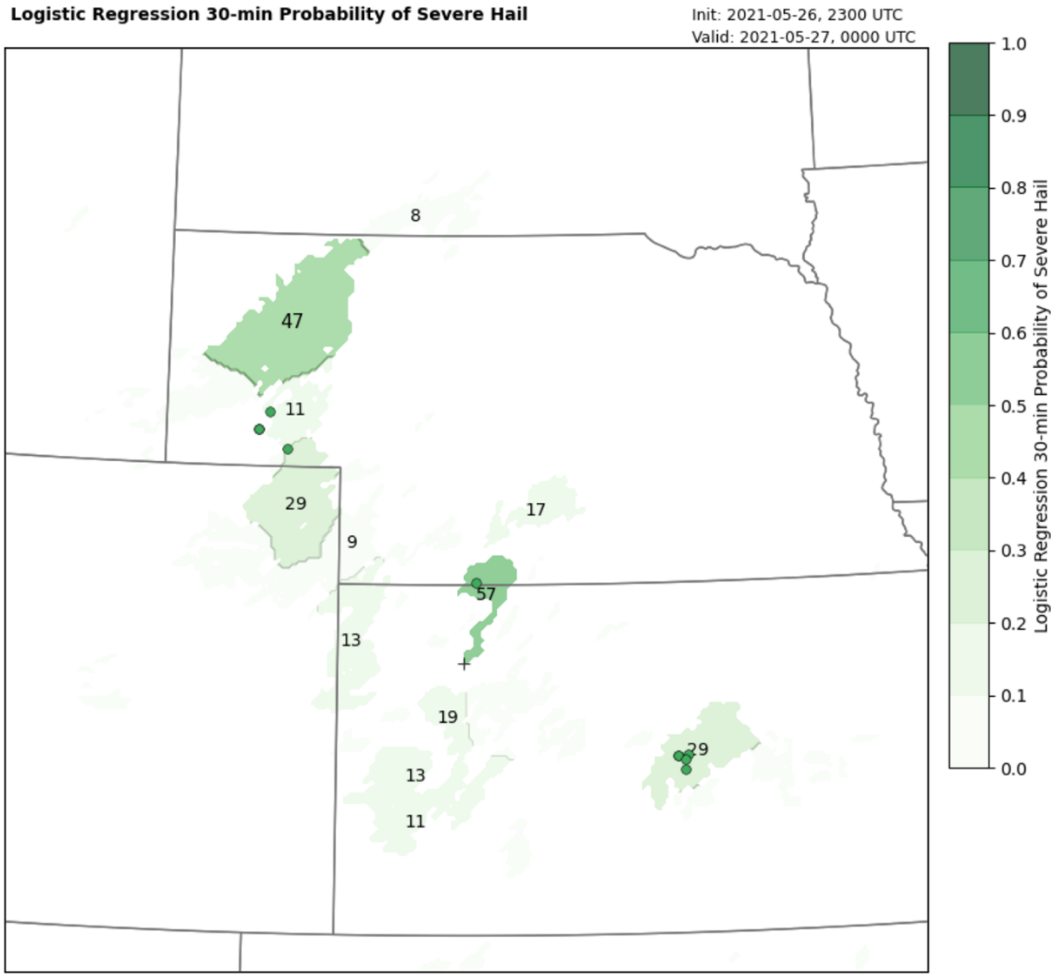

Logistic Regression 30-min Probability of Severe Hail
Init: 2021-05-26, 2300 UTC
Valid: 2021-05-27, 0000 UTC
8
47
11
29
9
17
57
13
19
13
11
29
Logistic Regression 30-min Probability of Severe Hail
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

*Figure 43 Object-based, ML-calibrated hail probabilities (shaded objects) for 1-h WoFS forecasts initialized 2300 UTC 26 May 2021. The observed hail reports are indicated by the green circles.*

Based on their impressions of these products, participants were asked: "*In what way is this guidance useful or not useful, and why?*". This was an open-ended question, and, although a variety of responses were received, they were generally positive and touched on some recurring themes. First, it was often noted that, if WoFS wrongly predicted the locations of storms, the ML-based probabilities were of little use. Second, it was noted that the guidance was very skillful at discriminating which hazard would be dominant. For example, in cases where wind was the dominant threat, the probabilities were highest for wind. Finally, there were several comments noting that the product could be useful for Impact-Based Decision Support Services (IDSS), but several comments were skeptical of its use for warning operations because the polygons were so large.

The next question asked, "*How well does the object identification algorithm highlight distinct storms/regions? (i.e., are there too many separate objects, or too few large/merged objects?)*". Participants were presented with a slider bar where they could indicate their impression within the spectrum of "Too much separation" (left), "Objects look about right" (center), and "Too much merging" (right). The results of this survey indicate that the responses skewed toward "Too much merging", but they weren't far off from "Objects look about right" (Fig. 44).

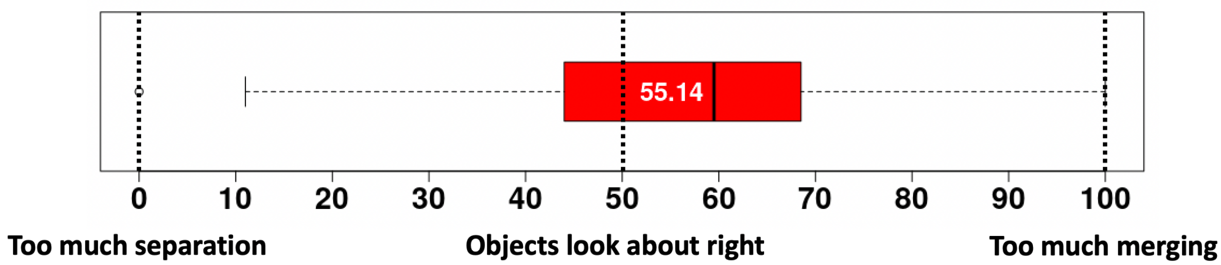


*Figure 44 Distribution of where participants placed the slider bar for the question, "How well does the object identification algorithm highlight distinct storms/regions? (i.e., are there too many separate objects, or too few large/merged objects?)".*

Next, participants were asked, "*What suggestions do you have to improve the graphics? (e.g., is filling the object with a single color confusing?)*". Since this was an open-ended question, there were a variety of answers. For the most part, people did not believe that major changes were needed for the graphics. Several participants mentioned that the lighter colors were hard to see, and there was a mix of participants that thought a single color should be used within each object and those that thought some kind of gradient in color should be used within the objects. Other comments suggested that some type of smoothing over time could be implemented so that the probabilities don't jump around so much from one time to the next, and another comment suggested finding a way to composite the hazard types on one plot.

The next question asked participants to consider an alternative version of the object presentation, which is shown in Figure 45. Specifically, participants were asked, "Do you prefer the raw unaltered objects, or the smooth elliptical objects?". 68% (80) preferred the raw, 16% (19) the elliptical, and 15% (18) liked both.

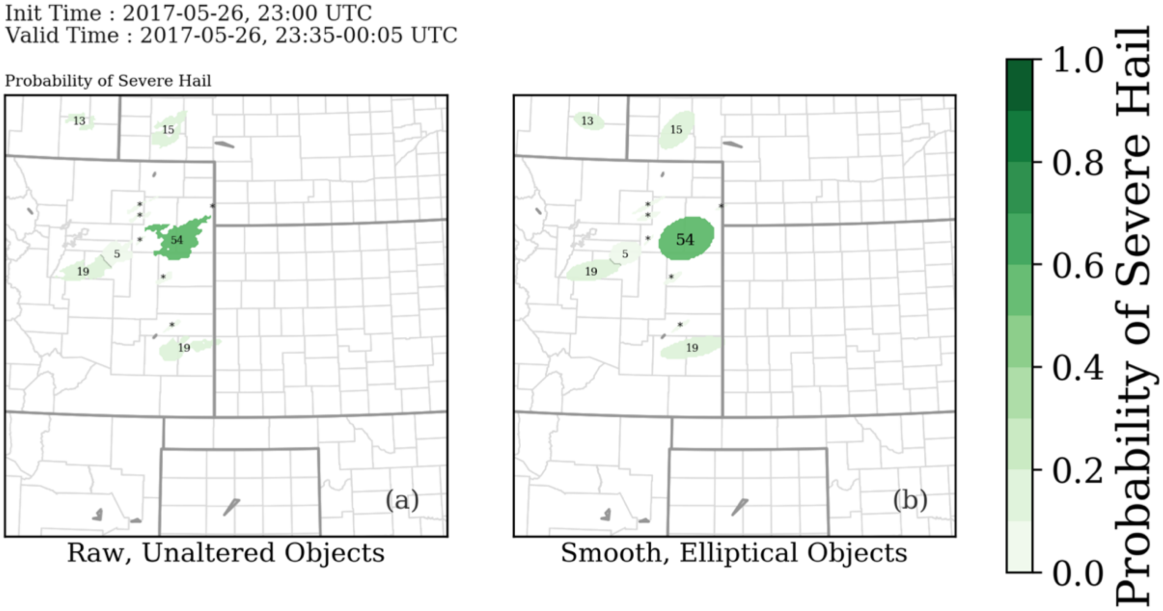
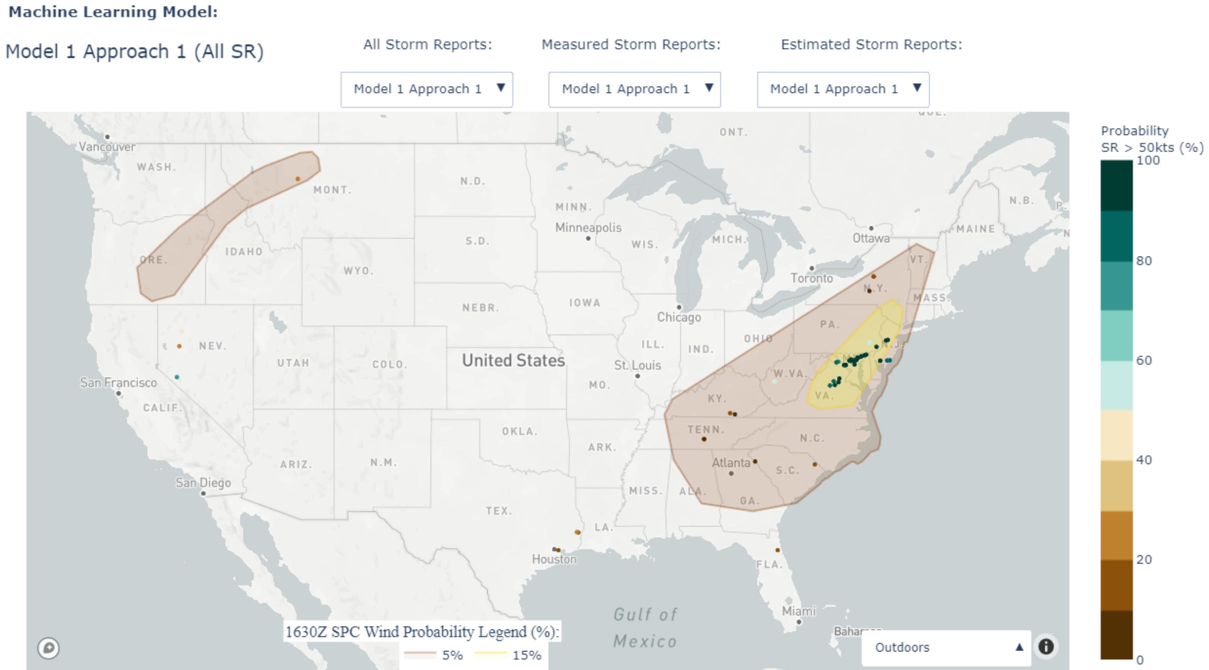

Init Time : 2017-05-26, 23:00 UTC
Valid Time : 2017-05-26, 23:35-00:05 UTC
Probability of Severe Hail
13
15
54
5
19
19
(a)
Raw, Unaltered Objects
13
15
54
5
19
19
(b)
Smooth, Elliptical Objects
1.0
0.8
0.6
0.4
0.2
0.0
Probability of Severe Hail

*Figure 45 Two different versions of ML-derived object probabilities*

*d) Model Evaluations – Group D: Medley*

*D1) ISU ML Severe Wind Probabilities*

Machine-learning algorithms were used to derive probabilities that thunderstorm wind damage reports were associated with severe intensity winds (i.e., 50 knots or more). Three training approaches were utilized: one including radar data, one without radar data, and one using regional training with radar data. For each of the three approaches, output from two different algorithms were presented. One was an average ensemble (using a stack generalized linear model (GLM)], while the other was the best single model determined from objective measures in testing [i.e., gradient boosted machine (GBM)]. Severe wind probabilities derived from each of these six machine learning models were available for yesterday's preliminary wind reports for evaluation on an interactive webpage developed by Iowa State University (ISU; Fig. 46).



**Machine Learning Model:**

Model 1 Approach 1 (All SR)

All Storm Reports: Model 1 Approach 1 ▼

Measured Storm Reports: Model 1 Approach 1 ▼

Estimated Storm Reports: Model 1 Approach 1 ▼

*Figure 46 Example of interactive webpage for the D1. ISU Machine-Learning Severe Wind Probability evaluation during the 2021 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.*

Participants were asked to evaluate (on a scale of 1 to 10) how well the machine-learning algorithms provided useful and accurate probabilistic information regarding the likelihood that wind damage reports were associated with winds >= 50 knots. Given the subjective nature of the evaluation, participants were asked to consider an assessment of the environment and storm mode, agreement with severe wind probabilities from the SPC Day 1 Outlook, and the ML probabilities assigned to *measured* wind reports. This evaluation was done without the participants knowing which model or approach was being displayed during the evaluation. The distribution of subjective ratings by participants during the five-week evaluation of the ISU ML severe wind probabilities reveals a relatively narrow rating range (i.e., 4-8) regardless of model or approach (Fig. 47). The primary findings from the subjective ratings include 1) the regionally trained ML models generally received lower ratings than full CONUS-trained models, 2) the impact of including radar data was relatively small in the subjective ratings with a slight improvement when radar was used to train the GBM model, and 3) the single GBM model received slightly higher subjective ratings than the stack GLM ensemble approach. The SFE participants commonly noted that the single model (i.e, GBM) often produced dichotomous output (i.e., very high or very low probabilities) while the ensemble approach (i.e., stack GLM) often produced moderate probabilities (i.e., 40-60%) for most reports.

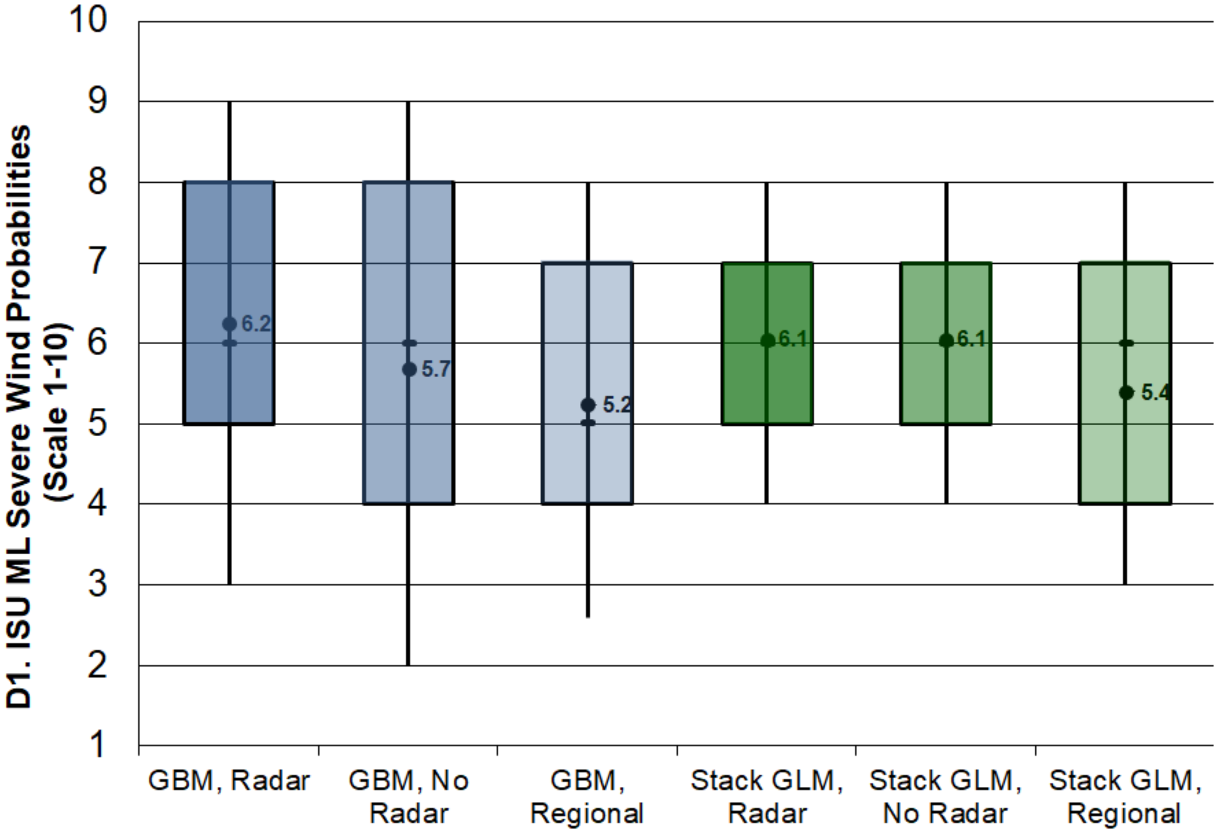


*Figure 47 Distributions of subjective ratings (1-10) by SFE participants of the ISU ML severe wind probabilities for preliminary wind reports for two models (GBM - blue; stack GLM - green) and three different approaches (radar - darkest shade; no radar - medium shade; regionally trained with radar - lightest shade).*

*D2) NCAR ML Convective Mode Probabilities*

Machine-learning algorithms were trained to provide probabilistic guidance of simulated storm mode using convection-allowing model (CAM) output. Specifically, two trained ML models were evaluated: 1) a supervised ML system that trains a convolutional neural network (CNN) to predict the mode of CAM storms using a hand labeled dataset of ~2000 CAM storms, and 2) a partially supervised CNN system, that is trained with UH and clustered using a Gaussian mixture model (GMM). Both systems output probabilistic predictions of supercells, quasi-linear convective systems (QLCSs), and disorganized modes for storm objects from two 3-km, 36-hr, deterministic, 00 UTC-initialized WRF-ARW CAM forecasts: one generated locally at NCAR and the other being the operational HRRRv4. The two ML models applied to these two CAMs were evaluated based on the subjective impressions of the participants on estimating convective mode probabilities using an interactive website developed by NCAR (Fig. 48).

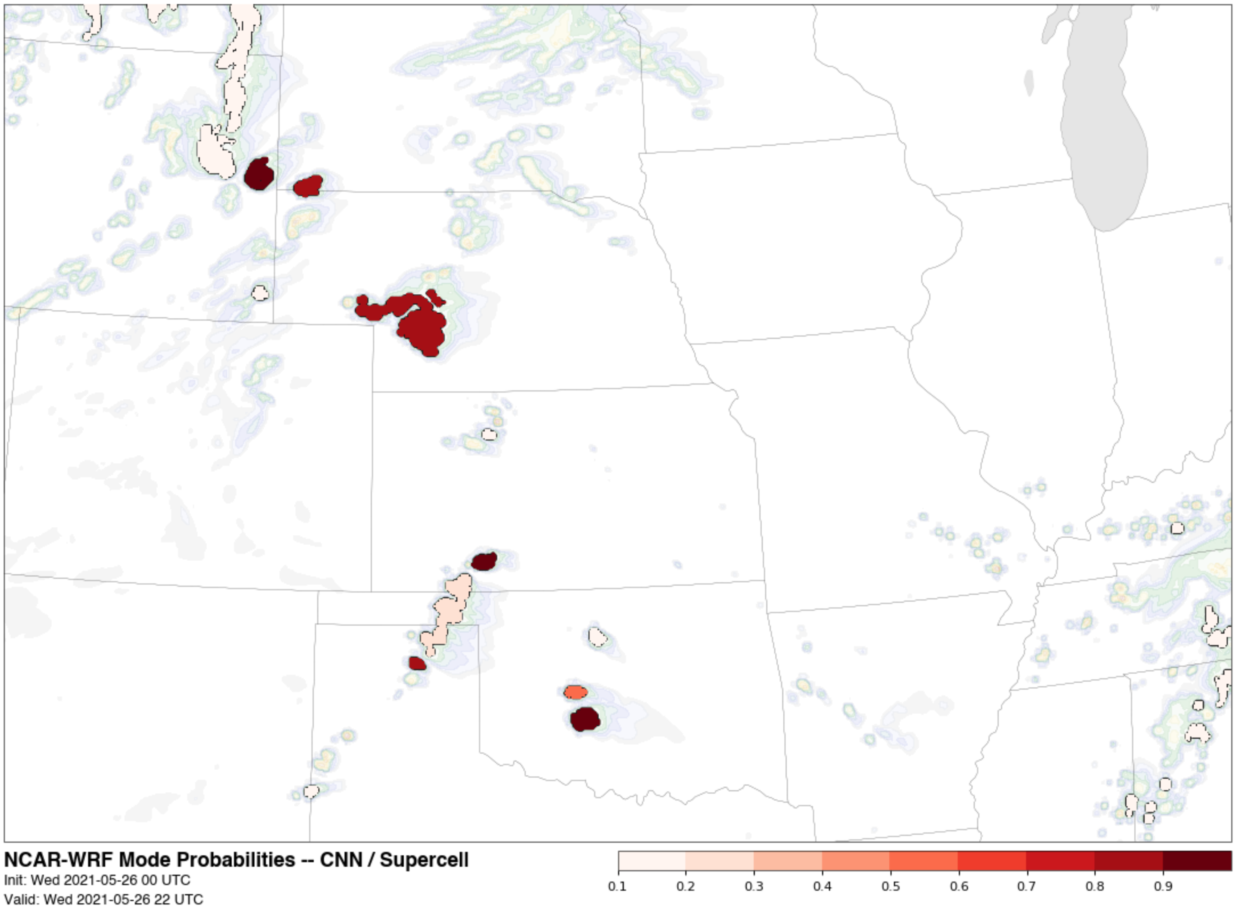

NCAR-WRF Mode Probabilities -- CNN / Supercell
Init: Wed 2021-05-26 00 UTC
Valid: Wed 2021-05-26 22 UTC
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

*Figure 48 Example of interactive webpage for the D2. NCAR Machine-Learning Convective Mode Probability evaluation during the 2021 SFE. Storm objects from the CAMs are shaded with the probability of being a supercell, QLCS, or disorganized convective mode with composite reflectivity lightly shaded in the background.*

Participants were asked to evaluate (on a scale of 1 to 5) how well the machine-learning algorithms provided useful and accurate probabilistic estimates of convective mode (i.e., supercell, QLCS, or disorganized) over a regional domain. The distribution of subjective ratings by participants during the five-week evaluation of the NCAR ML severe wind probabilities (Fig. 49) reveals a couple of primary findings: 1) the supervised CNN convective mode output generally received higher subjective ratings than the partially supervised GMM convective mode output and 2) the ratings between the two CAMs were very similar with perhaps a very slight edge to the NCAR WRF, which was the model used for the hand-created labels in the CNN output. Overall, SFE participants found promise in this type of output, especially if applied to a CAM ensemble. The most common concern from SFE participants was hour-to-hour inconsistency for the convective mode probabilities, especially the GMM output. For example, it was commonly noted that probabilities for the same storm object would flip between high and low probabilities with time in a manner not warranted by the environment or storm structure.

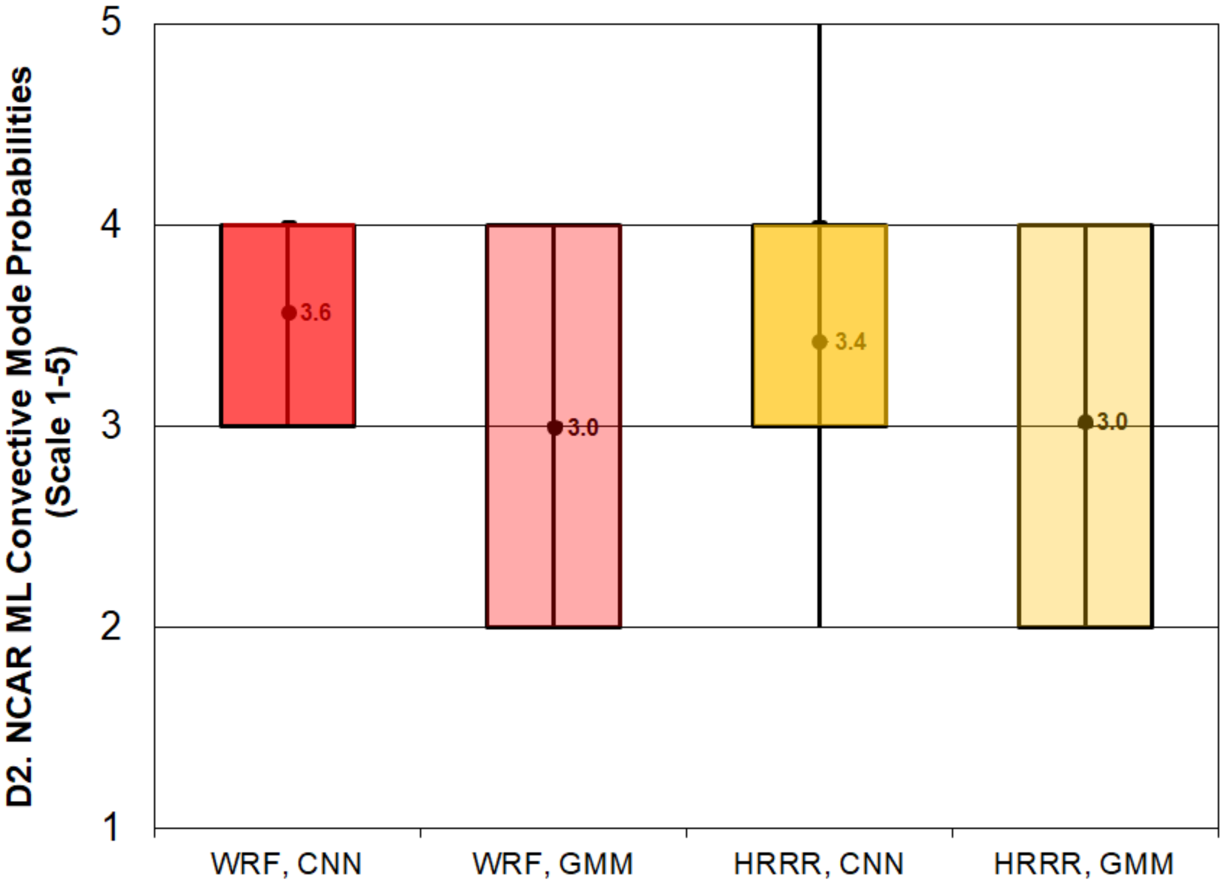


*Figure 49 Distributions of subjective ratings (1-5) by SFE participants of the NCAR ML convective mode probabilities for storm objects from two models (NCAR WRF - red; HRRR - yellow) and two ML algorithms (supervised CNN - darkest shade; partially supervised GMM - lightest shade).*

### *D3) Mesoscale and Storm-Scale Analyses*

#### *i) Mesoscale Analysis Background*

Two hourly versions of 3D-RTMA with different backgrounds were subjectively evaluated by participants during the 2021 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 GSL version used the FV3-based RRFS as the first-guess background information in the hybrid DA system, and the EMC version used the operational HRRR for first-guess background. Both versions of 3D-RTMA used the GDAS for background error covariance information. The hourly analyses for 2-m temperature, dewpoint, SB/ML/MUCAPE, and effective-layer STP and SRH were examined during the 18-03 UTC period on the following day (Fig. 50).

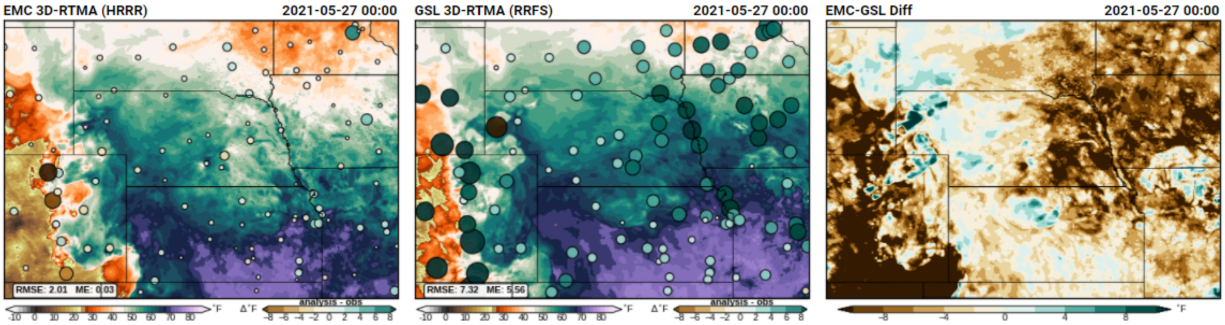


*Figure 50 Example of the website comparison page for the 3D-RTMA during the 2021 HWT SFE. The EMC version of 3D-RTMA is shown in the left panel, the GSL version of 3D-RTMA in the middle panel, and the difference (EMC-GSL) in the right panel. The 2-m dewpoint temperature analysis valid at 0000 UTC on 27 May 2021 is shaded in the left two panels. The difference (analysis-obs) at METAR sites is shown by the size and shading of the dots. The corresponding 2-m dewpoint temperature analysis difference is shaded on the right panel.*

The goal was to assess the impact of the first-guess background on analyses for short-term severe weather forecasting applications. Overall, the FV3-based GSL version of the 3D-RTMA was rated subjectively "much worse" to "slightly worse" than the HRRR-based EMC version (Fig. 51). Specifically, participants commonly noted issues in the FV3-based version in the composite reflectivity field for too much/intense convection and too moist in the 2-m dewpoint field. GSL was unable to implement all of the components of the cloud analysis routine into the FV3-based version prior to the SFE that already existed in the HRRR-based version, so that had a notable impact on the difference between the versions. Upscaled 40-km versions of the full resolution 3D-RTMA systems were also compared to the SPC mesoanalysis (Fig. 52). The RAP-based SPC mesoanalysis is the operational standard in the NWS for situational awareness and short-term forecasting applications. The upscaled 40-km FV3-based GSL version was most commonly rated "slightly worse" to "about the same" as the SPC mesoanalysis. Of course, the previously mentioned issues with the 3-km GSL version apply to the upscaled version, so fixing those issues would likely place the upscaled FV3-based 3D-RTMA on par with the SPC mesoanalysis.

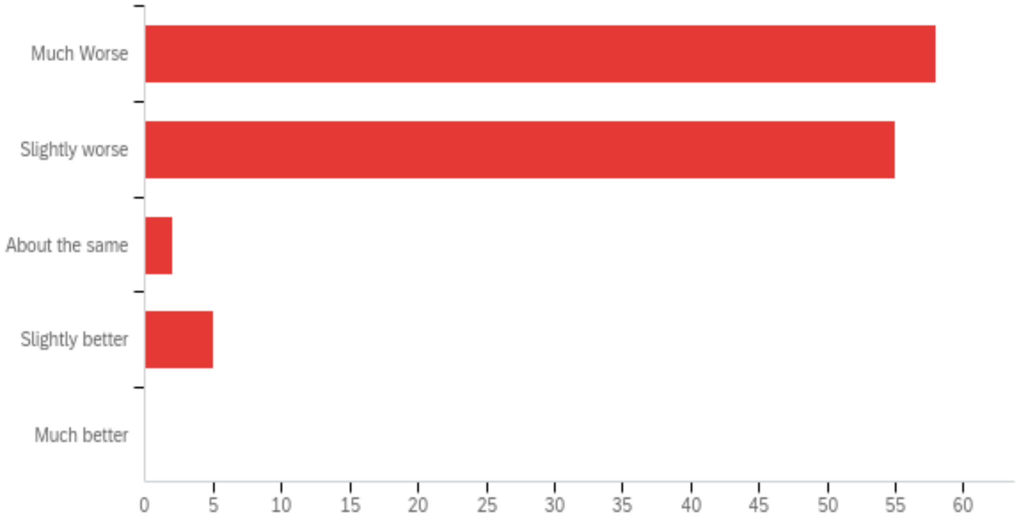


*Figure 51 Subjective daily rating counts by 2021 HWT SFE participants regarding whether the FV3-based GSL version of 3D-RTMA was "much better", "slightly better", "about the same", "slightly worse", or "much worse" than the HRRR-based EMC version of 3D-RTMA.*

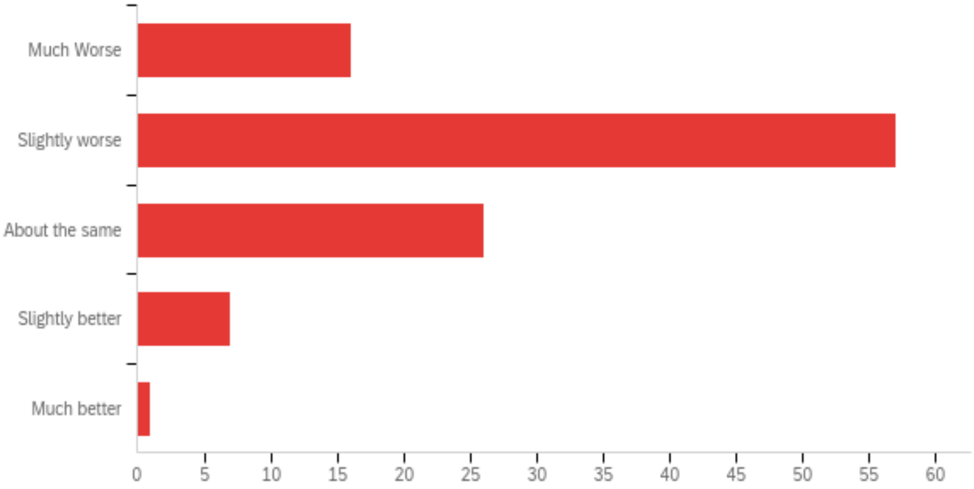


*Figure 52 Subjective daily rating counts by 2021 HWT SFE participants regarding whether the upscaled 40-km FV3-based GSL version of 3D-RTMA was "much better", "slightly better", "about the same", "slightly worse", or "much worse" than the SPC Mesoanalysis.*

*ii) Mesoscale Analysis Data Assimilation Frequency*

A sub-hourly version of 3D-RTMA was also run by EMC for comparison with the EMC hourly version during the 2021 SFE. The sub-hourly EMC version also used the operational HRRR for its first-guess background, but used the HRRRDAS instead of the GDAS for background error covariance information. The analyses for 2-m temperature, dewpoint, SB/ML/MUCAPE, and effective-layer STP and SRH were examined at 15-minute intervals during the 18-03 UTC period on the following day (Fig. 53).

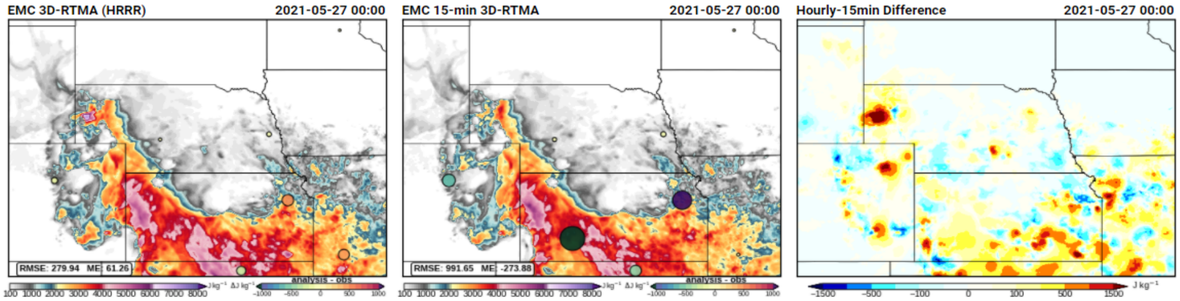


*Figure 53 Example of the website comparison page for the 3D-RTMA during the 2021 HWT SFE. The hourly EMC version of 3D-RTMA is shown in the left panel, the sub-hourly EMC version of 3D-RTMA in the middle panel, and the difference (hourly-subhourly) in the right panel. The surface-based CAPE analysis valid at 0000 UTC on 27 May 2021 is shaded in the left two panels. The difference (analysis-obs) at radiosonde sites is shown by the size and shading of the dots. The corresponding SBCAPE analysis difference is shaded on the right panel.*

Overall, the 15-minute version of the 3D-RTMA was rated subjectively "about the same" to "slightly better" than the hourly version (Fig. 54). Specifically, participants commonly noted that the versions looked similar at the top of the hour and found the higher frequency of analysis updates useful. Despite a discontinuity in the temporal evolution of the analysis fields going from the 15-minute updates at 15, 30, and 45 minutes past to the top-of-the-hour analysis (i.e., drift and reset), over 70% of participants thought the 15-minute version provides useful information over the hourly version. Another noteworthy comment was that the updated quality-control package in the sub-hourly 3D-RTMA seemed to reduce the frequency of erroneous local maxima that arise in the presence of bad surface observations.

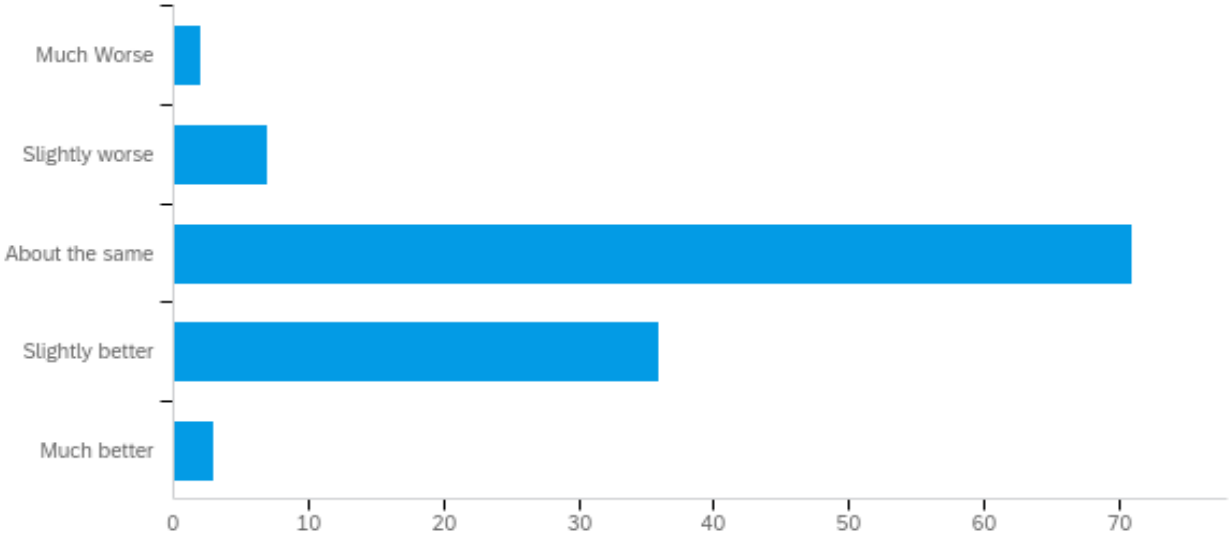


*Figure 54 Subjective daily rating counts by 2021 HWT SFE participants regarding whether the 15-minute EMC version of 3D-RTMA was "much better", "slightly better", "about the same", "slightly worse", or "much worse" than the hourly EMC version of 3D-RTMA.*

*iii) 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 winds. Specifically, the 15-minute forecasts of 10-m and 80-m winds from WoFS (cycled every 15 minutes) were used as a proxy for the analysis (i.e., ground truth) of severe wind. The WoFS ensemble maximum 10-m and 80-m wind analyses were accumulated from 1800 UTC through 0300 UTC for comparison with preliminary local storm reports, especially measured gusts (Fig. 55).

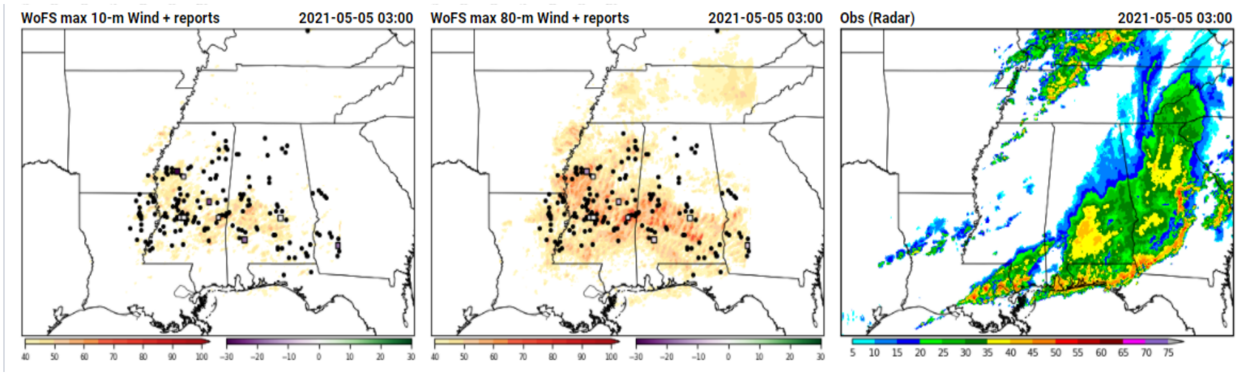


*Figure 55 Example of the website comparison page for the WoFS analyses during the 2021 HWT SFE. The 1800-0300 UTC accumulated ensemble maximum 10-m wind is shown in the left panel, the ensemble maximum 80-m wind in the middle panel, and the observed composite reflectivity in the right panel. The wind damage reports are the black circles on the left two plots while the measured gusts are the open squares shaded by the difference (analysis-obs) of the gust measured at that location.*

The goal of the evaluation was to assess the current capability of WoFS to produce output to diagnose severe and damaging winds. Overall, the WoFS ensemble maximum winds were positively viewed in terms of lining up with preliminary severe wind reports and a subjective assessment of severe wind based on environment and radar characteristics, as more than three-fourths of the participants gave neutral or positive ratings (Fig. 56). Overall, the 80-m winds received higher subjective ratings than the 10-m winds and often better matched the magnitudes of any measured gusts (i.e., the 10-m winds were always a larger underestimate of the measured gusts). Although this evaluation was exploratory, the participants found this to be an interesting and promising approach and use of a rapidly cycling convection-allowing ensemble system.

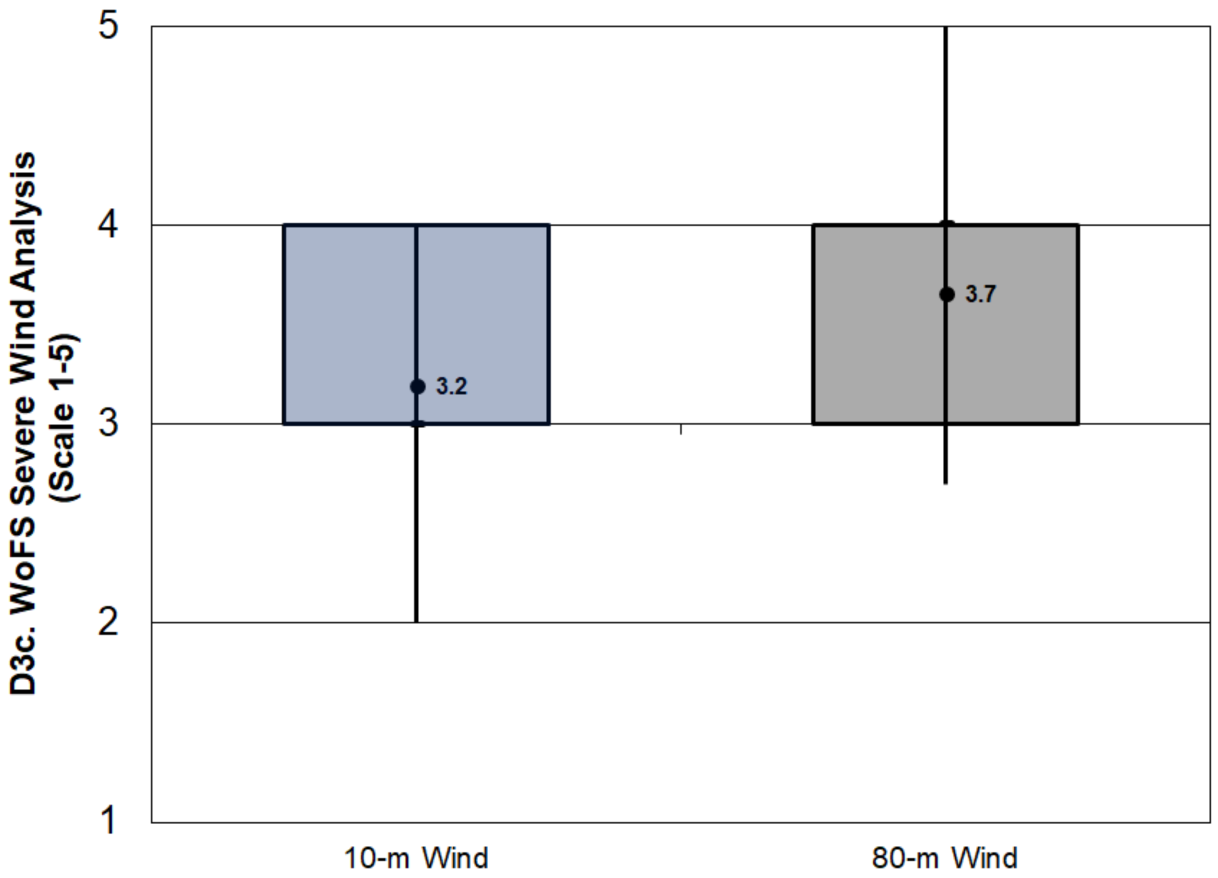


*Figure 56 Distributions of subjective ratings (1-5) by SFE participants of the WoFS storm-scale severe wind analysis for ensemble maximum 10-m winds (blue) and 80-m winds (gray), where the ratings represent how well the WoFS maximum wind analyses align with the preliminary severe wind reports and overall assessment of severe winds: 1 - Very Poorly; 2 - Poorly; 3 - Neutral, neither poorly nor well; 4 - Well; 5 - Very Well.*

### *D4) GEFS vs. SREF – Severe Weather Forecasting*

With a plan to develop a Unified Forecast System (UFS) in NOAA, legacy operational systems, like the Short-Range Ensemble Forecast (SREF) system are slated for retirement in the next few years. To assess the readiness of the Global Ensemble Forecast System (GEFS) to replace the SREF for severe weather forecasting applications, an evaluation was performed during the 2021 HWT SFE. Several relevant fields for severe weather forecasting were examined, including 2-m dewpoint, MLCAPE, CAPE and shear combined probabilities, and the significant tornado parameter (STP), along with calibrated thunder and severe probabilities. These fields were examined at 3-h intervals during the convective Day 3 and Day 2 periods using a multi-panel webpage with the SPC mesoanalysis as the verification standard (Fig. 57).

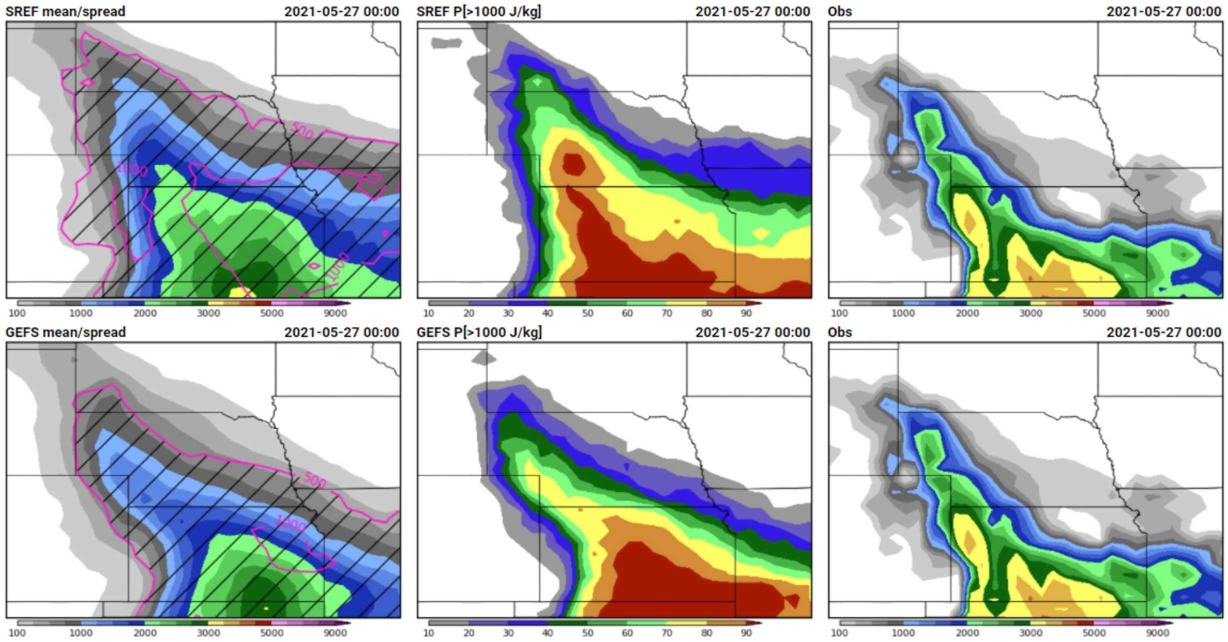


*Figure 57 Example of the website comparison page for the GEFS comparison to the SREF during the 2021 HWT SFE. The SREF forecasts are shown in the top row with the GEFS forecasts in the bottom row. The Day 3 forecasts of MLCAPE mean/spread (left column) and probability of exceeding 1000 J/kg (middle column) are shown for comparison with the SPC Mesoanalysis (right column) as the "observation" valid at 0000 UTC on 27 May 2021.*

For the Day 3 environment forecasts, the GEFS severe weather fields were subjectively rated similar to SREF overall (Fig. 58). There are some days/locations where the GEFS does better than the SREF and vice versa, with the median and mean ratings centered on "about the same". The minor exception was for STP, which was favored in the GEFS forecasts over the SREF forecasts, which tended to overestimate the STP on some days. The Day 3 calibrated guidance offers a different perspective on the GEFS performance relative to the SREF (Fig. 59). The GEFS calibrated thunder and severe guidance more frequently was rated better than the SREF when compared to the raw environment output from the ensembles. Given that the methodologies for generating the calibrated guidance are very similar between the GEFS and SREF, it is hypothesized that the 20-year reforecast dataset with the GEFS offers the ability to improve upon SREF calibrated guidance, which is trained on only one year of data.

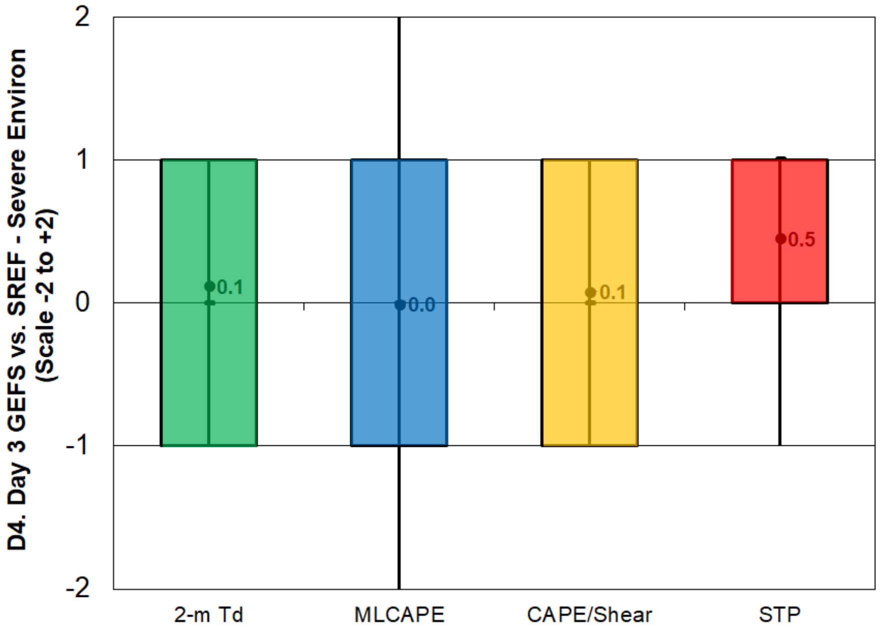


*Figure 58 Distributions of Day 3 subjective ratings (-2 to +2) by SFE participants of the GEFS environment forecasts compared to the SREF forecasts for ensemble fields of 2-m dewpoint (green), MLCAPE (blue), CAPE & shear (orange), STP (red). The ratings represent how the GEFS compared to the SREF for these environment fields from -2: Much Worse; -1: Slightly Worse; 0 - About the Same; +1: Slightly Better; +2: Much Better.*

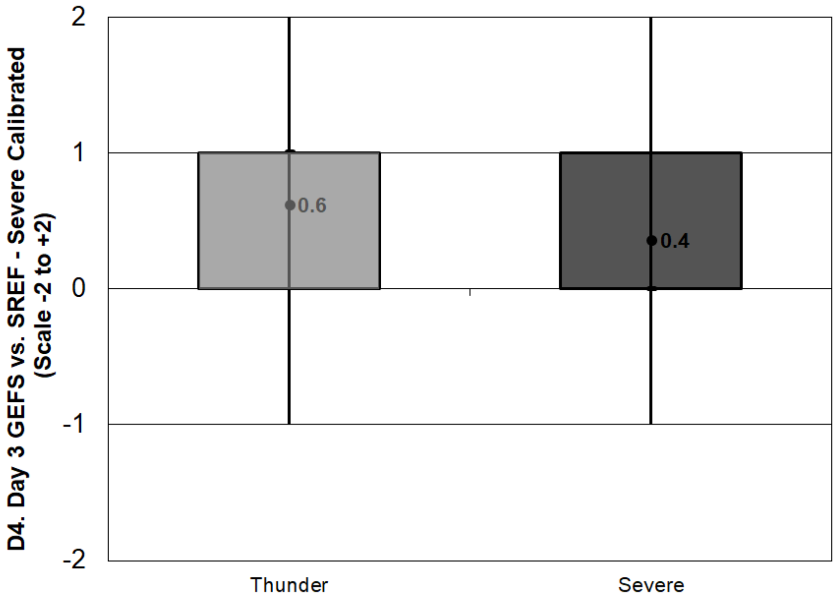


*Figure 59 Distributions of Day 3 subjective ratings (-2 to +2) by SFE participants of the GEFS calibrated forecasts compared to the SREF calibrated forecasts for thunder (light gray) and severe (dark gray). The ratings represent how the GEFS calibrated guidance compared to the SREF calibrated guidance from -2: Much Worse; -1: Slightly Worse; 0 - About the Same; +1: Slightly Better; +2: Much Better.*

For the Day 2 environment forecasts, the results are similar to those seen on Day 3. Overall, the GEFS and SREF forecasts for severe weather fields were comparable (Fig. 60), with the GEFS forecasts slightly favored for 2-m dewpoint and STP and the SREF forecasts slightly favored for MLCAPE. The Day 2 calibrated guidance evaluation also reveals higher subjective ratings for the GEFS calibrated thunder and severe guidance compared to the SREF (Fig. 61).

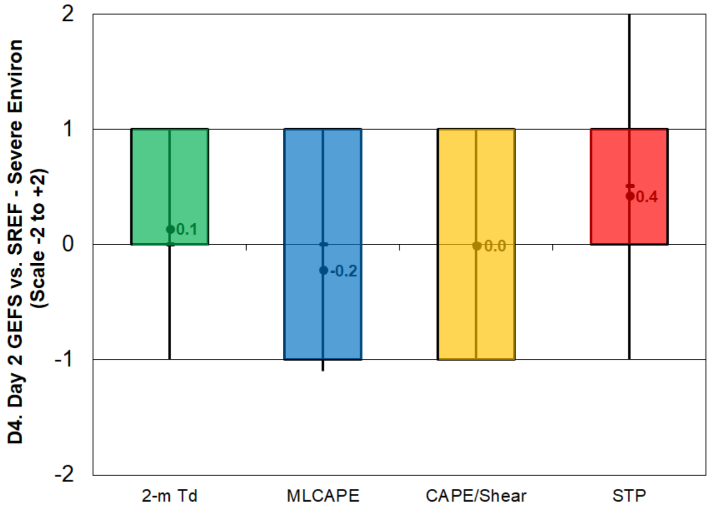


*Figure 60 Distributions of Day 2 subjective ratings (-2 to +2) by SFE participants of the GEFS environment forecasts compared to the SREF forecasts for ensemble fields of 2-m dewpoint (green), MLCAPE (blue), CAPE & shear (orange), STP (red). The ratings represent how the GEFS compared to the SREF for these environment fields from -2: Much Worse; -1: Slightly Worse; 0 - About the Same; +1: Slightly Better; +2: Much Better.*

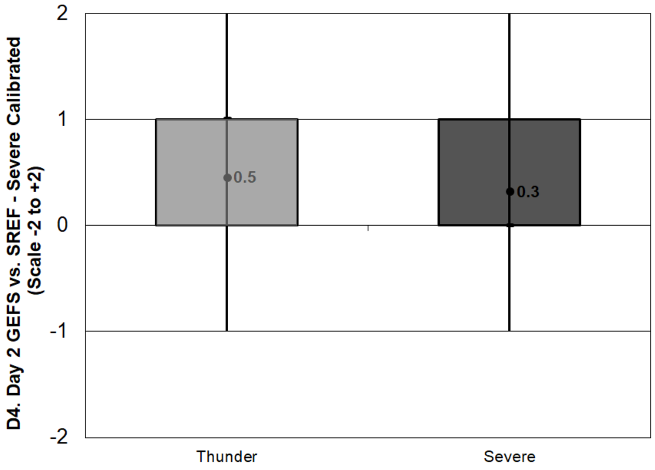


*Figure 61 Distributions of Day 2 subjective ratings (-2 to +2) by SFE participants of the GEFS calibrated forecasts compared to the SREF calibrated forecasts for thunder (light gray) and severe (dark gray). The ratings represent how the GEFS calibrated guidance compared to the SREF calibrated guidance from -2: Much Worse; -1: Slightly Worse; 0 - About the Same; +1: Slightly Better; +2: Much Better.*

Related to the calibrated thunder evaluation, an evaluation of the Extended Range Convective Forecast Product (ECFP) was also conducted for the Aviation Weather Center (AWC). The ECFP Planning Tool is a graphical representation of the forecast probability of thunderstorms, which is produced by AWC. The product identifies graphically where in the US thunderstorms are likely over the next 87 hours to support the long-range planning for the National Airspace System. The ECFP planning tool has been developed in response to FAA and Industry needs in planning for weather hazards, specifically convection, one to three days in advance. Currently, the ECFP depiction is an automated graphical forecast created from the SREF calibrated thunderstorm guidance with forecast times valid in 3-hour intervals. Probability of thunderstorm contours are depicted at 30, 50, and 70% probabilities. With the retirement of the SREF approaching, the AWC wanted to evaluate the GEFS calibrated thunder product as a possible replacement for the SREF in generating the ECFP.

The results for the GEFS-based ECFP were overwhelmingly positive during the SFE (Fig. 62). The GEFS-based ECFP received, on average, subjective ratings that were more than one point (on a ten-point scale) higher than the SREF-based ECFP. While the participants noted that the higher probabilities from the GEFS often led to higher POD & FAR of convective activity, this may also be a function of the fixed probability thresholds utilized by the ECFP algorithm (i.e., some product calibration may be required). Nevertheless, the GEFS calibrated thunder product looks promising as a potential replacement for the SREF calibrated thunder in producing the AWC ECFP.

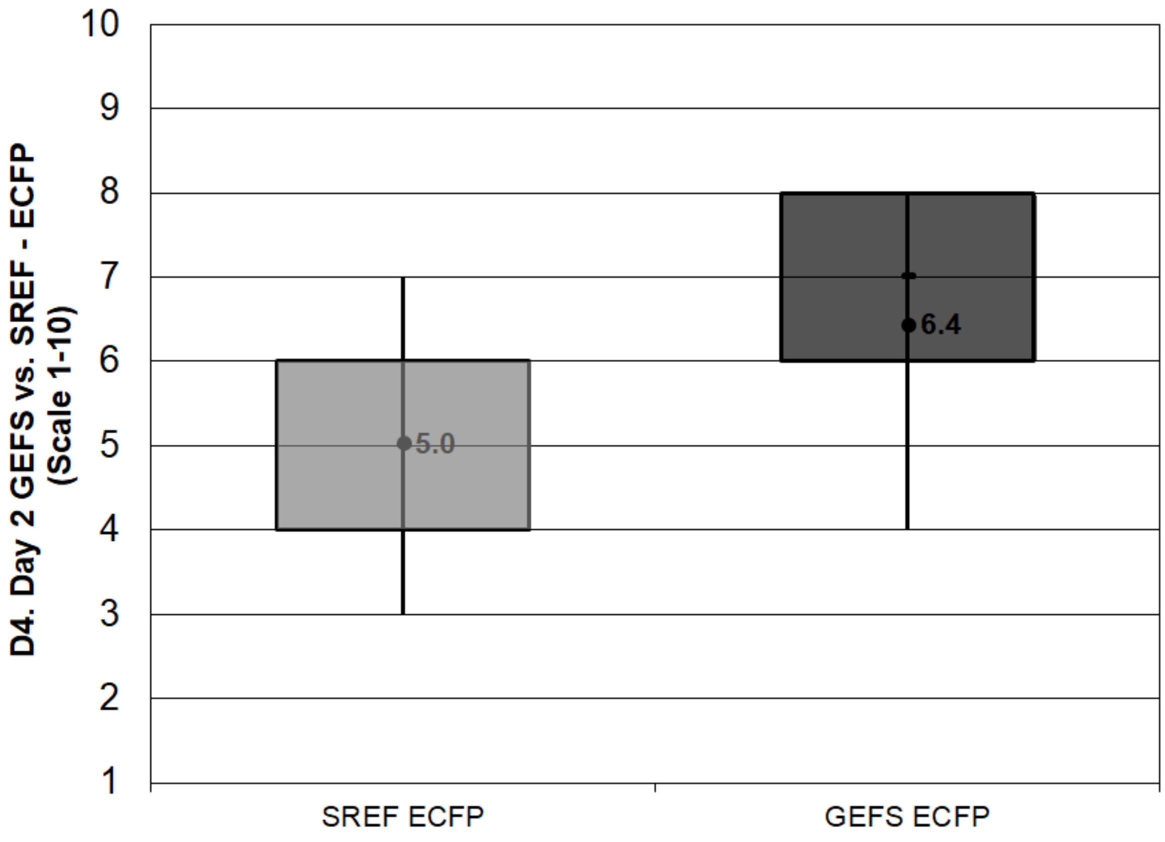


*Figure 62 Distributions of* ***Day 3*** *subjective ratings (1-10) by SFE participants of the GEFS-based ECFP (light gray) and SREF-based ECFP (dark gray).*

*e) Evaluation of Experimental Forecast Products – Innovation Group*

During the first period of experimental forecasting, which occurred from 11:30am – 12:30pm CDT, the Innovation Group issued coverage and conditional intensity forecasts for individual hazards (tornado, wind, and hail) covering the Day 2 period (1200 – 1200 UTC). Within the Innovation Group, one set of participants used CAM guidance to generate their outlook (All Data), while another did not use CAMs (No CAMs). Both sets of forecasts were issued as a group, with each group being led by an SFE facilitator. The primary goal of this activity was to quantify the value of CAM guidance for issuing Day 2 products. An example set of forecasts is displayed in Figure 63.

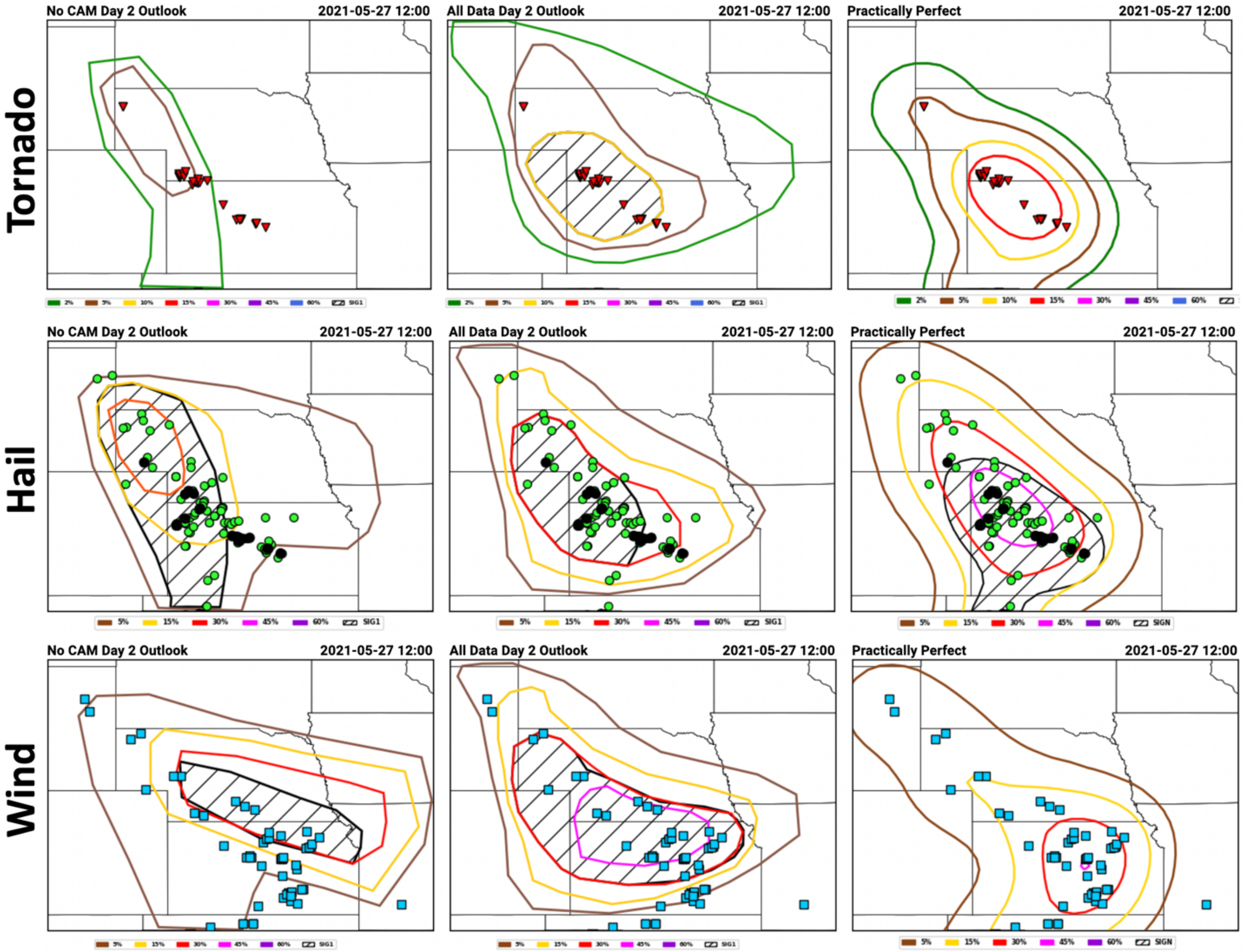


*Figure 63 Experimental Day 2 outlooks for tornado (top), hail (middle), and wind (bottom) issued by the "No Cam" group (left column) and the "All Data" group (middle column). Practically perfect hindcasts are displayed in the right column. In each panel, the LSRs are overlaid.*

Overall, the differences in distributions of ratings between the All Data and No CAM forecasts for each hazard were small, with one exception: the All Data group's wind coverage forecasts were rated much higher than the No CAM forecasts (Fig. 64). In fact, using a Welch's t-test, the wind coverage forecasts were the only pair of All Data and No CAM forecasts with statistically significant differences.

This was very likely due to several instances in which CAMs were indicating that convection would evolve into a linear MCS and potential severe wind threat. This type of evolution can be difficult, if not impossible, to infer from non-CAM operational models at Day 2 lead times.

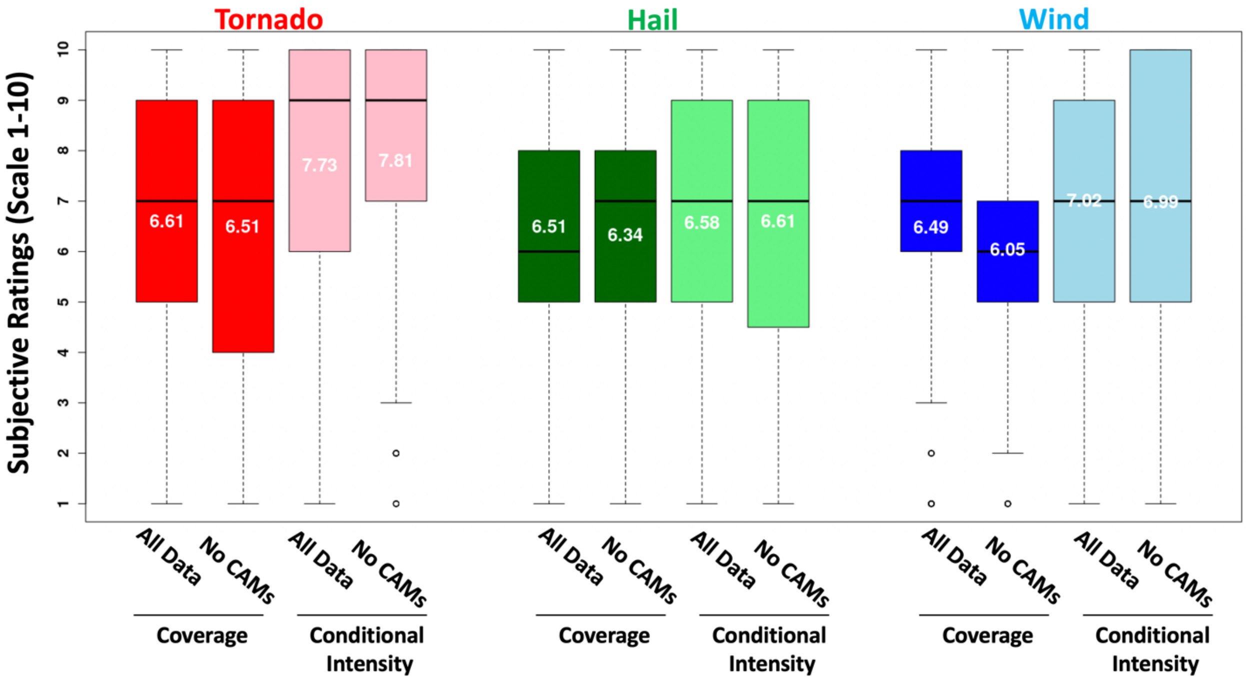


*Figure 64 Distributions of subjective ratings (1-10) by SFE participants of the Day 2 All Data and No CAM group forecasts for coverage and conditional intensity of tornado, wind, and hail. The numbers in white text indicate mean ratings, while the horizontal black lines indicate the median.*

For the Innovation Group forecasting activity conducted from 2:15-4pm CDT, participants generated severe hazard probabilities valid over 1-h time windows covering 2200-2300, 2300-0000, and 0000-0100 UTC. Two initial forecasts were generated during the 2:15-3:15pm period, which covered the 22-23Z and 23-00Z time windows. Then, during the 3:15-4pm period, the 22-23Z and 23-00Z periods were updated, and one more outlook covering 00-01Z was generated. For both sets of initial and final forecasts, two expert forecasters used all available datasets including WoFS (Forecaster WOF 1 & 2), while two other expert forecasters used all available datasets except for WoFS (Forecaster NOWOF 1 & 2). The NOWOF forecasters used the SFE viewer (https://hwt.nssl.noaa.gov/sfe_viewer/2021/forecast_tool) to generate forecasts, while the WoFS forecasters used an internal version of the WoFS viewer (https://wof.nssl.noaa.gov/realtime/). Forecasters using the SFE viewer had access to the WoFS domain bounds, so that the forecast domain was the same between the two groups. Additionally, two other groups of non-expert forecasters issued forecasts with and without WoFS similarly to the expert forecasters. These non-expert forecasts were combined into consensus forecasts (WoFS Consensus and No WoFS Consensus, respectively). The consensus forecasts were created by gridding the outlooks and converting them to continuous spatial probabilities using a method developed at SPC (Karstens et al. 2019). Non-expert numerical coverage probability forecasts were averaged to create the consensus forecasts. If non-expert forecasters drew a significant severe contour, indicating a greater than 10% probability for significant severe weather, a significant contour was drawn for the consensus forecasts at a point if at least half of participants drew a significant severe contour at that gridpoint.  On the first day that participants were engaged in the activity, facilitators made a brief presentation of training material. It was emphasized that the 1-h time window outlooks should not be treated as the longer time window outlooks.  Given the short lead times, the outlooks should in theory be more accurate and precise, meaning that highlighted areas should have higher probabilities and cover smaller areas relative to SPC's Day 1 Convective Outlooks.  An example set of forecasts from 3 June 2021 is shown in Figure 65.

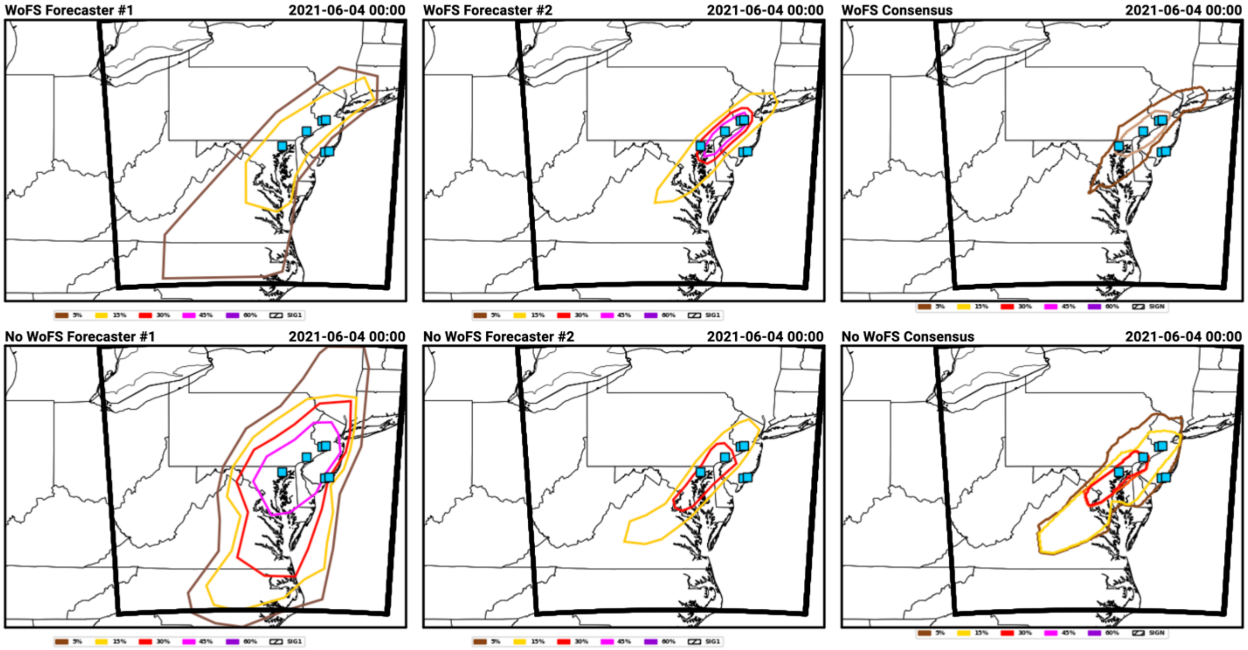


*Figure 65 Innovation Group outlooks generated as part of the afternoon forecasting activity highlighting the probability of severe wind gusts covering the 1-h period 23-00Z on 3 June 2021:  WoFS Forecaster #1 (upper left), WoFS Forecaster #2 (upper middle), WoFS Consensus (upper right), No WoFS Forecaster #1 (lower left), No WoFS Forecaster #2 (lower middle), and No WoFS Consensus (lower right).  Observed wind reports are indicated by the blue boxes.*

For the evaluation of these forecasts, participants categorized each outlook as "Excellent", "Above Average", "Average", "Below Average", and "Poor".  Comparisons were made to the observed storm reports, MESH, NWS warnings, and practically perfect hindcasts, which were tuned with a smaller standard deviation to give higher amplitude and smaller areas.  There was a total of 90 outlooks that were evaluated each day (3 hazards x 5 times x 6 forecasts = 90).  These outlooks were split so that two of the evaluation groups examined the initial forecasts, and the two other evaluation groups examined the final ones. The primary goal of this exercise was to quantify the value of WoFS to the experimental outlooks by comparing the WoFS and No WoFS outlooks, as well as test the concept of the Consensus outlook.  To present the evaluation statistics quantitatively, the categories listed above were converted to a 1-5 rating scale where 5 corresponds to excellent, 4 to above average, and so on.  Then, the average ratings were computed for each hazard and forecaster. The WoFS Forecasters #1 and #2 were averaged together along with No WoFS Forecasters #1 and #2. Furthermore, initial forecasts, which were issued during the 2:15-3:15pm time period, were averaged together, and the same was done for the final forecasts. Plots showing the results for each of the five forecast sets separately (i.e., 22-23 & 23-00Z initial, and 22-23, 00-23, 00-01 final) are available in the appendix.

For the initial forecasts (Fig. 66), the WoFS outlooks were rated higher than No WoFS, with the most dramatic differences for wind. The consensus WoFS and No WoFS forecasts were rated very similarly. The final forecasts (Fig. 67) exhibited similar behavior with the WoFS rated higher than No WoFS and consensus forecasts rated similarly. To examine whether differences were statistically significant, a Welch's t-test was applied to all six pairs of expert WoFS and No WoFS forecasts. Using a significance level of $\alpha = 0.05$, differences within each pair were significant, except for the initial tornado forecasts. None of the differences between the Consensus WoFS and No WoFS forecasts were significant. These results largely confirm the potential value provided by WoFS for issuing short-range severe weather products.

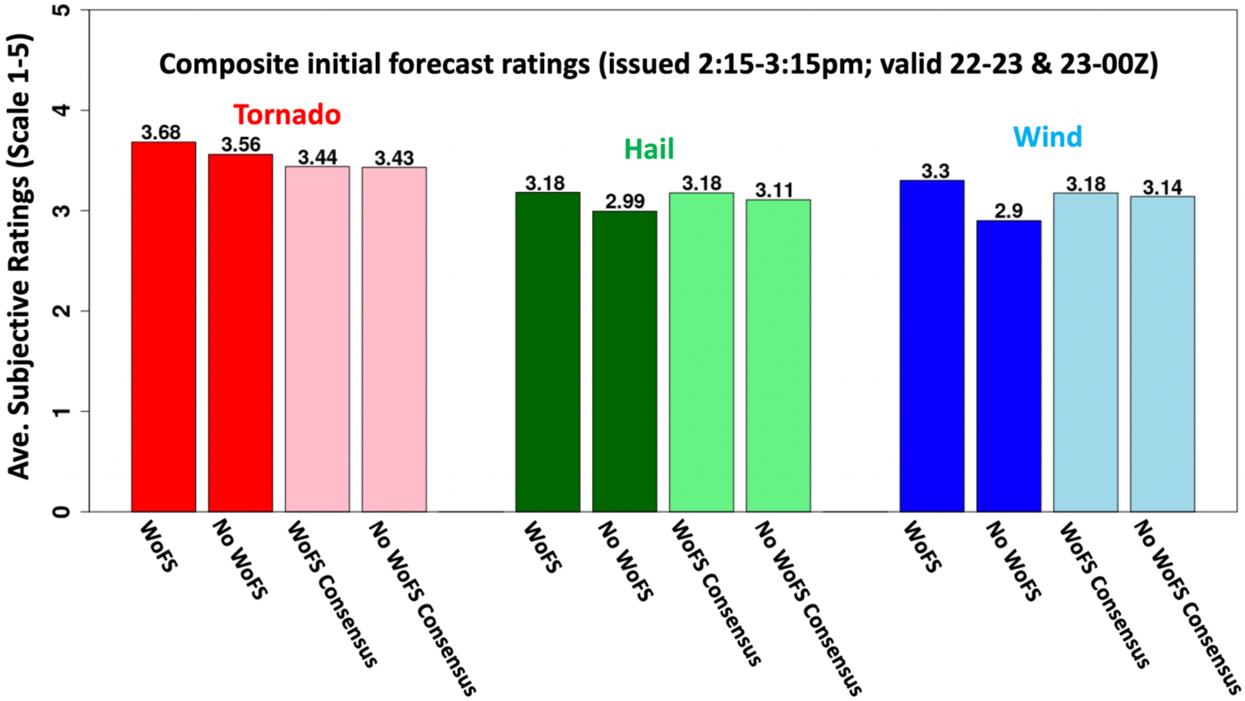


*Figure 66 Average subjective ratings of initial forecasts valid 22-23 & 23-00Z for the Innovation Group afternoon forecasting activity during SFE 2021.*

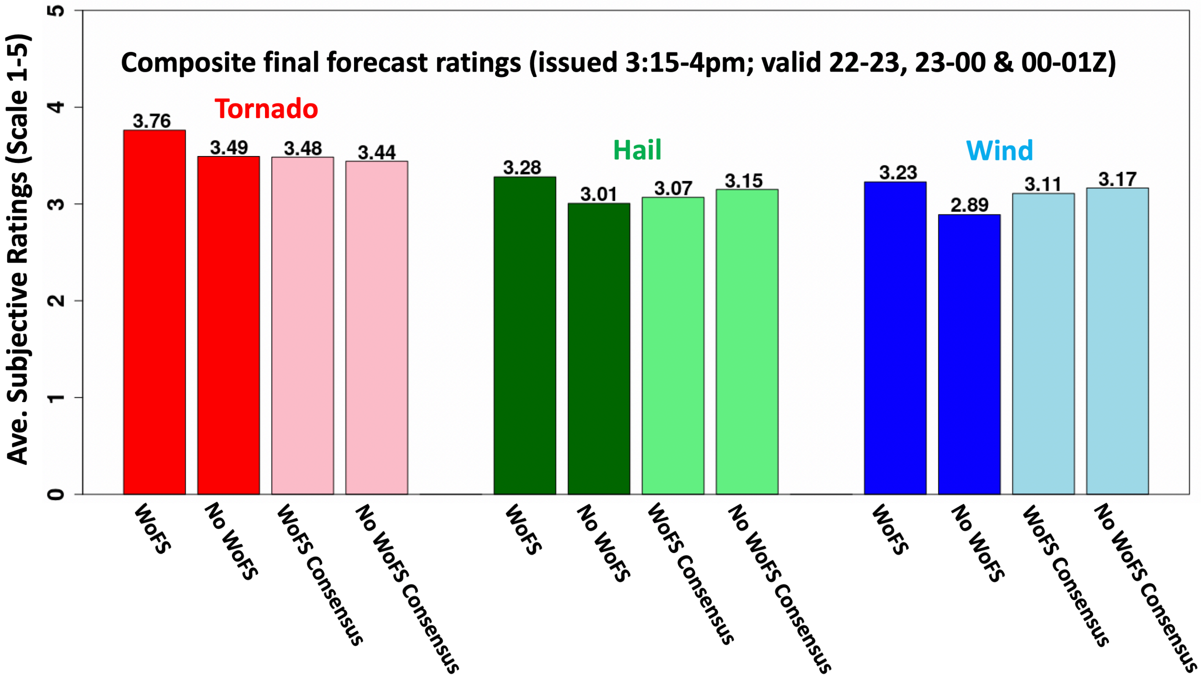


*Figure 67 Same as Figure 66, except for final forecasts valid 22-23, 23-00, & 00-01 UTC.*

After rating the forecasts, participants were asked, "How useful was WoFS to you yesterday in issuing your forecasts?", and then they were given choices that ranged from "Extremely useful" to "Not at all useful". The response frequencies to this question are shown in Figure 68. By far, participants most often selected "Very useful", and 81% of the responses indicated "Moderately useful" or better.

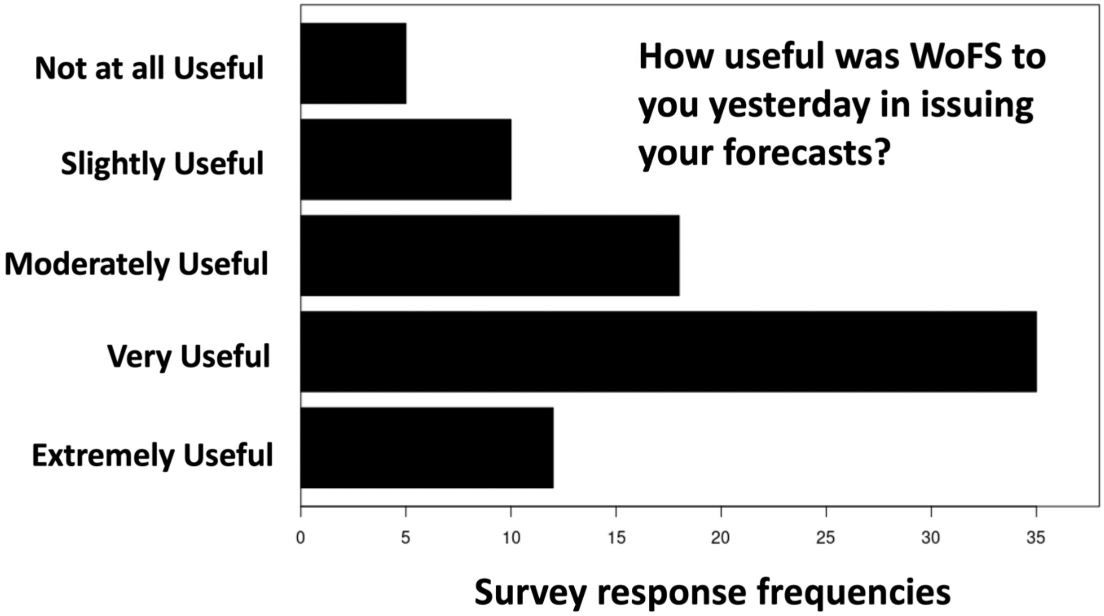


*Figure 68 Survey response frequencies to the question, "How useful was WoFS to you yesterday in issuing your forecasts?".*

Then, participants were asked, “Please indicate the hazard(s) for which you found WoFS to be most useful”, and could choose tornado, wind, and/or hail. Wind was chosen most frequently, which was closely followed by hail. Tornado was chosen about half as frequently as hail and wind (Fig. 69). These results are consistent with the average subjective ratings (Figs. 66 & 67) showing that the biggest differences between WoFS and No WoFS forecasts occurred with the wind forecasts.

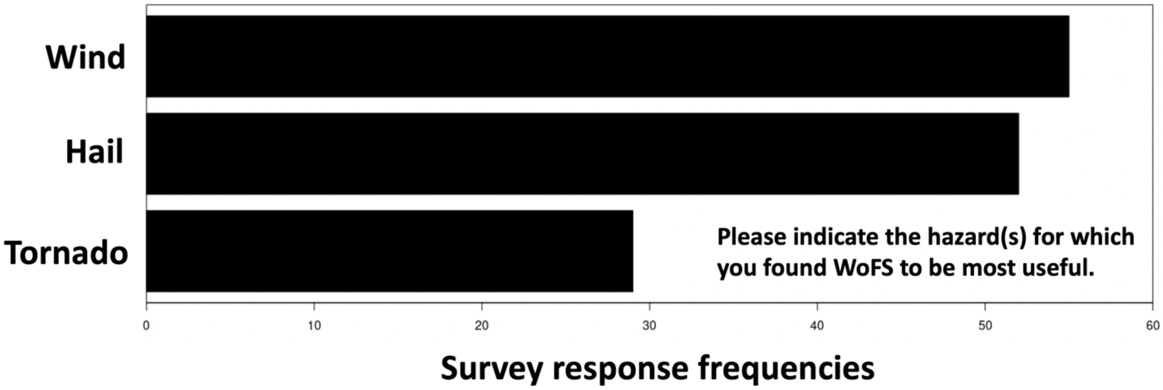


*Figure 69 Survey response frequencies to the question, “Please indicate the hazard(s) for which you found WoFS to be most useful.”*

Finally, participants were presented with a list of 12 WoFS products and asked to, “Please check the WoFS products that you found to be most useful yesterday”. The top five products (in order) were: 2-5 km UH probabilities, Wind speed probabilities, Reflectivity paintballs, Hail probabilities, and Hail percentiles. The only environmental products on the list, MLCAPE and STP, were selected the least frequently (Fig. 70).

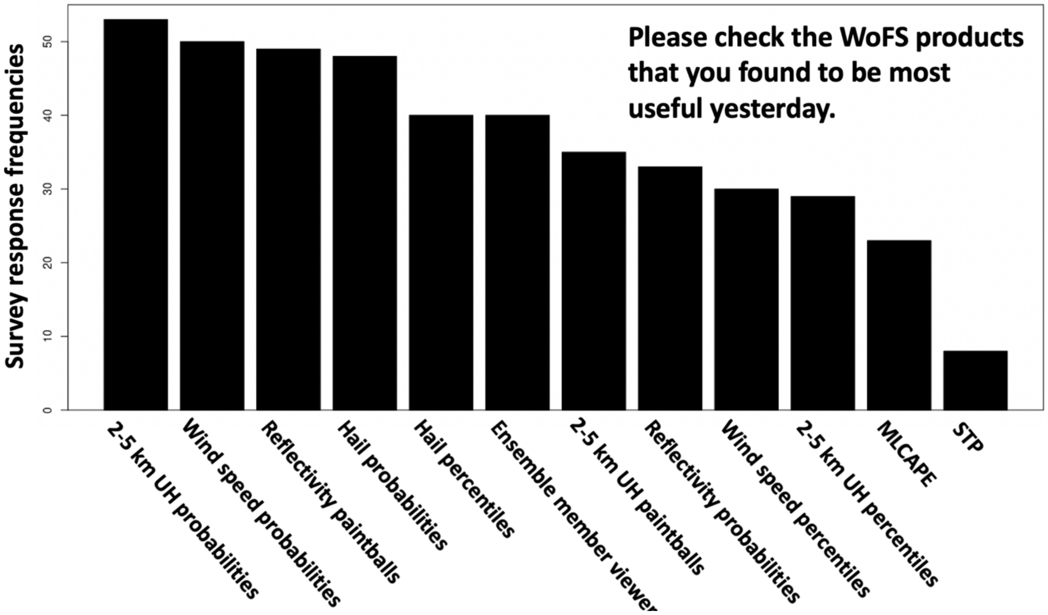


*Figure 70 Survey response frequencies to the question, “Please check the WoFS products that you found to be most useful yesterday”.*

*f) Evaluation of Experimental Forecast Products – R2O group*

During the morning forecasting activities, the R2O Group issued coverage and conditional intensity forecasts for individual hazards (tornado, wind, and hail) covering the Day 1 period (1800 – 1200 UTC). Within the R2O Group, the participants were divided into two subgroups to reduce group sizes and encourage engagement from all participants. One subgroup of participants used the 12Z HREF as the primary source of CAM guidance to generate their outlook, while another subgroup used the 12Z HRRRE as the primary source of CAM guidance. Both sets of forecasts were issued as a group (i.e., one outlook from each subgroup), with each subgroup being led by an SFE facilitator. Later during the afternoon forecasting activities from 3:00-4:00 p.m. CDT, these group outlooks were updated individually by the participants (i.e., separate outlooks for each R2O participant) using the latest observations and guidance, including WoFS. The primary goal of this activity was to engage the participants in examining and using CAM ensemble data in a relevant forecast process before knowing the outcome, which aids in the model evaluations on the following day. In addition, the outlook updates allow for the quantification of the value of updating outlooks in the afternoon using the latest observations and guidance. An example set of morning and afternoon update outlooks is displayed in Figure 71.

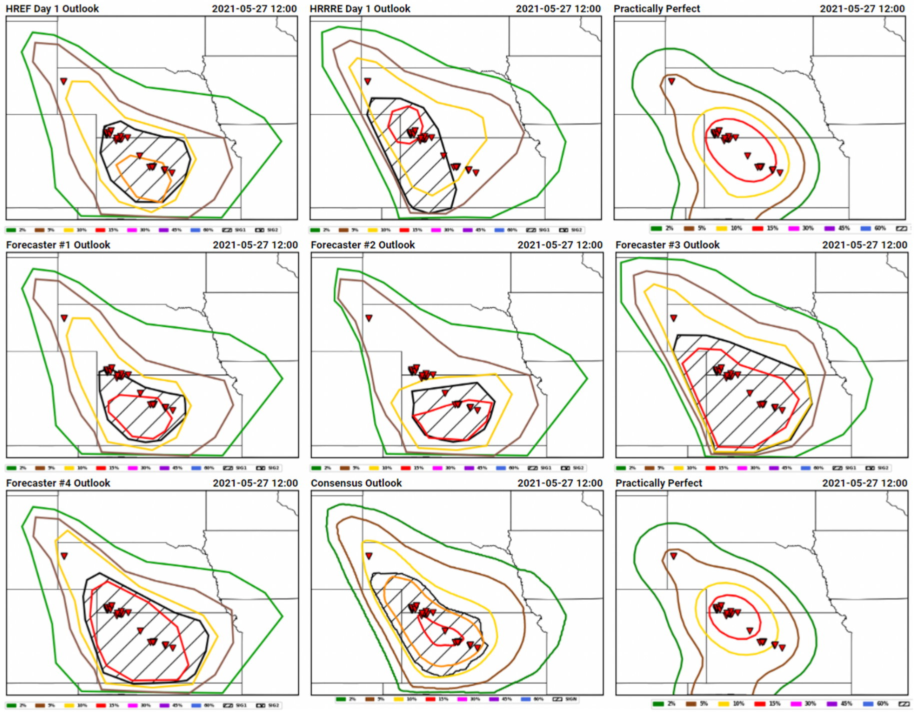


*Figure 71 Experimental Day 1 tornado morning outlooks for 26 May 2021 by the "HREF" subgroup (top left) and the "HRRRE" subgroup (top middle). The practically perfect hindcast is shown for the 1800-1200 UTC verification window (top right). The afternoon outlook updates for the same day are shown for individual NWS forecasters (middle row and bottom left), as well as the consensus of the rest of the R2O group (bottom middle). The practically perfect hindcast is shown for the 2100-1200 UTC verification window (bottom right). In each panel, the preliminary tornado LSRs are overlaid.*

Each initial experimental Day 1 outlook was rated the following day by all participants (regardless of group) while the afternoon updates were only rated by the R2O participants who generated the outlook updates. The primary purpose of the subjective ratings was to determine if the final, updated outlook that included use of WoFS data was an improvement over the initial outlook (i.e., without WoFS data). The afternoon outlook updates received higher mean subjective ratings across all hazards (tornado, hail, wind) and for coverage & conditional intensity layers (Fig. 72). The improvement in the mean ratings largely appears to be associated with improving the poorer morning outlooks (i.e., an increase in the lower quartile ratings). Overall, the participants found the WoFS to be useful in the outlook-update process by increasing confidence in various aspects of the forecast: CI, location, coverage, and intensity.

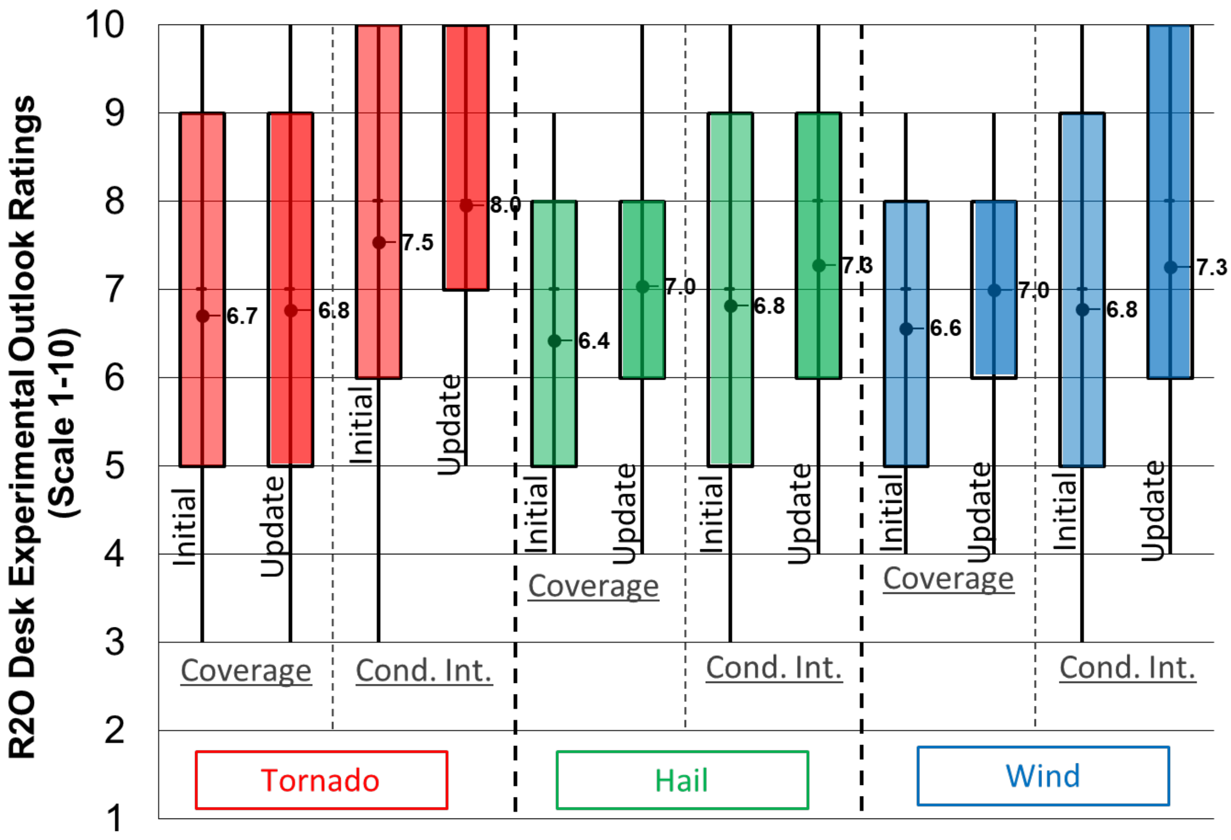
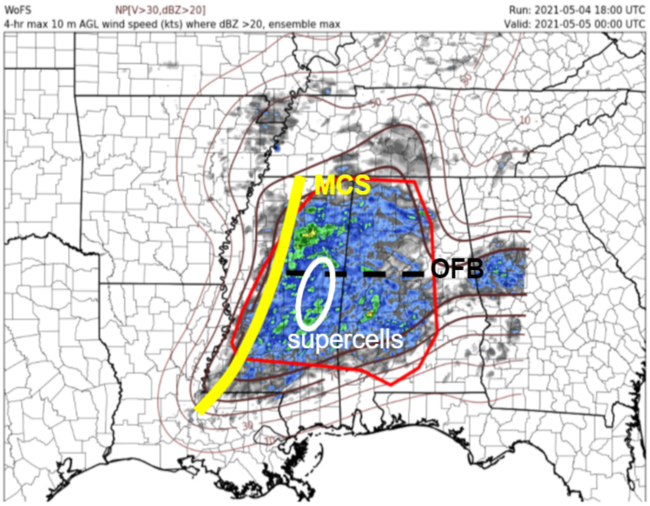


*Figure 72 Distribution of subjective ratings of experimental initial/final coverage and conditional intensity outlooks issued by SFE participants for tornado (red), hail (green), and wind (blue). The mean rating is shown by the circle with numerical value.*

As part of the afternoon forecasting activities on the R2O Desk, experimental mesoscale discussions (MDs) were generated during the 2021 HWT SFE. These MDs were generated daily in Google Slides (Fig. 71) 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: 1) focus on a meso-beta corridor with the greatest potential for severe weather over the next few hours and 2) explore the utility of WoFS to inform these MD products within the watch-to-warning time frame. In a feedback survey following the SFE, the forecasting activities were often cited by participants as their favorite SFE activity. Participants noted that these activities provided an opportunity to use and experience the models and products first-hand, which often led to an appreciation of the challenges faced by SPC forecasters in generating short-fused forecast products. Of all of the different forecast activities, the experimental MD generation was commonly mentioned by participants as his/her favorite activity of the SFE.

Participant MCD


WoFS
NP[V>30,dBZ>20]
4-hr max 10 m AGL wind speed (kts) where dBZ >20, ensemble max
Run: 2021-05-04 18:00 UTC
Valid: 2021-05-05 00:00 UTC
MCS
OFB
supercells

**MESOSCALE CONVECTIVE DISCUSSION**

AREAS AFFECTED…MS and AL

CONCERNING...WWs 144, 146, 147

VALID...20-22Z

SUMMARY…Increasing severe wind potential

DISCUSSION…Severe thunderstorms along a line from the central MS-TN border through the SW corner of MS have grown upscale and are absorbing scattered cellular convection ahead of the line. These storms have already produced severe wind across N LA prior to mesoscale organization. Progressive MCS will encounter 2000-3000+ J/kg MLCAPE across much of MS and W AL, along with ample shear in both the 0-3 and 0-6-km layers for continued MCS maturation. Most WoFS members depict a dominant bow echo emerging within this line in varying locations from C MS into W AL. WoFS produces its highest 10-m wind speeds near and just north of the bow apex, in the part of the convective line oriented most normal to the shear vector, with relatively less severe wind to the south along the more shear-parallel segments (despite dense coverage of very short UH tracks within these segments). The most likely mesoscale focus for this potentially significant severe mode appears to be the outflow boundary extending roughly W-E across central MS and AL and/or the supercells just E of KJAN and S of the OFB merging with the line. This MCD also accounts for the possibility of underforecast rightward propagation into the highly supportive (derecho composite 6-8) airmass south of the OFB.

Forecaster: Wade

*Figure 73 Example of an experimental MD created on 4 May using WoFS output.*

## 4. Summary

The 2021 NOAA HWT Spring Forecasting Experiment (2021 SFE) was conducted virtually from 3 May – 4 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 2021 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 products generated using WoFS and other CAM ensembles for lead times from one hour to 2 days, and (3) evaluate different CAM ensemble post-processing strategies with an emphasis on those using machine-learning algorithms.

Several preliminary findings/accomplishments from the 2021 SFE are listed below:

- Experimental short-term individual hazard outlooks were generated with and without WoFS. Additionally, WoFS was used for updating full-period hazard forecasts valid 2100-1200 UTC and corresponding conditional intensity guidance.
  - In the Innovation Group, subjective ratings indicated that WoFS provided a statistically significant advantage relative to No-WoFS outlooks for expert forecasters for each hazard and at both issuance times, with the exception of the initial tornado outlooks. The biggest advantage in the WoFS relative to the No-WoFS outlooks was in the initial wind forecasts.
  - In the R2O group, 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.
- Examined and assessed various methods to produce first-guess calibrated probabilistic hazard guidance based on forecast output from HREFv3, GEFS, and HRRRv4.
  - For tornadoes, when considering only days with one or more tornado in the SFE domain, “STP Cal Circle” (combination of UH with forecast STP and associated tornado climatologies) performed best for Day 1 & 2 lead times. STP Cal Circle also performed best over 4-h time windows within Day 1, even besting the 0600 and 1300 UTC SPC Severe Timing Guidance products, which has performed best in previous experiments.
  - For Day 2 hail guidance, the GEFS CSU product performed noticeably better than HREF/SREF calibrated and HRRR NCAR (ML algorithm based on 1200 UTC HRRR initialization). Out of the seven sets of Day 1 products, the Loken et al. (2020) hail guidance was notable for its exceptional performance, while the GEFS CSU product also performed quite well. For the 4-h time windows, the SPC Timing Guidance performed best.

  - For severe wind, the HRRR NCAR method had the highest mean ratings for Day 2, besting the GEFS CSU and HREF/SREF Calibrated. For Day 1, out of five sets of products, the Loken et al. (2020) wind guidance performed best. For the 4-h time windows within Day 1, the SPC Timing Guidance performed much better than the HREF/SREF Calibrated.
- Examined various **deterministic** CAM systems within the CLUE using HRRRv4 as a baseline.
  - HRRRv4 continues to perform better than experimental FV3 configurations; however, large improvement in the FV3 configurations relative to prior years is noted.
  - GFDL-FV3 reflectivity structures do not depict realistic looking convective cores, but do handle the extent of stratiform precipitation associated with convective systems well.
  - Strong updrafts and intense convective cores are a feature of FV3-LAM models worth investigating, as they may lead to greater perceived severity of events relative to the observed event.
  - Mean subjective ratings for the North American EMC-FV3 LAM domain were higher than the CONUS domain, so implementation of the North American domain should benefit forecasts especially for Day 2 and beyond.
  - Evaluated times for the Data Assimilation comparisons were likely insufficient for drawing strong conclusions; objective verification over all hours of these forecast models should be undertaken to determine the full impact of DA on FV3-LAM forecasts.
  - The MYNN/Thompson physics suite for FV3-LAM is currently performing the best for severe convective forecasting; however, there was a lot of spread in model solutions generated by using different physics parameterizations.
  - Further development and testing is recommended for stochastics parameterization strategies; while differences were seen between members, whether those differences contributed positively or negatively to the forecast was highly case-dependent.
  - Both physics parameterization differences and stochastic physics perturbations seem to be viable strategies for increasing ensemble spread to provide realistic depictions of severe convective storms.
  - Stability continues to be a struggle for the FV3-based forecasts, with SBCAPE fields uniformly being reported as too low relative to observations.
- Examined various **ensemble** CAM systems within the CLUE using HREFv3 as a baseline.
  - HREF continues to stand out as the best performing CAM ensemble. However, GSL RRFS was notable for its performance – coming in a fairly close second in terms of mean subjective ratings.
  - Improved performance in HRRRE-M relative to HRRRE-S showed that skill can be gained through combining stochastic physics perturbations with a multi-physics ensemble configuration approach, even if it is simply two different physics suites.
  - Blending strategies for HRRR and HREF could be a useful strategy to generate skillful probabilistic guidance in between the times that new HREFv3 forecasts become available.
  - Comparing ensembles at 0-12 h lead times with and without a valid-time-shifting (VTS) data assimilation approach revealed very small differences, but in the direction of improvement for the VTS runs. The improved computational efficiency makes VTS a

technique with potential applications in future CAM ensembles, but future work is needed to improve overall performance.

- Examined utility of WoFS for short-term severe weather forecasting application in the watch-to-warning timeframe.
    - Comparison of 2100 and 2300 UTC WoFS initializations to time-lagged HRRR forecasts indicated superior performance for WoFS, especially for the 2300 UTC initializations when the rapid DA of WoFS likely helps the most. WoFS was frequently praised for its ability to highlight very specific threat areas.
    - A survey on an experimental CAM scorecard indicated that the scorecard usually reflected participants' subjective impressions of relative forecast quality. Reflectivity and UH were most often listed as fields that participants believed should be included in the scorecard.
    - Subjective ratings of a 1.5-km WoFS-Hybrid compared to a random WoFS member with matching physics revealed that the biggest advantage for the WoFS-Hybrid was for the earlier 2100 UTC initializations, which is likely related to the ability of WoFS-Hybrid to spin up storms faster. However, WoFS-Hybrid has a problem with spurious convection early in the forecast, so further work is needed to address that issue.
    - The majority of the time participants believed that WoFS-Hybrid provided value relative to WoFS for storm structure, but did not provide value the majority of the time for storm location. The WoFS-Hybrid forecasts usually fell within the envelope of WoFS members.
    - ML-calibrated, WoFS-derived object-based hazard probabilities were very skillful at discriminating which hazard would be dominant. Forecasters commented that the product could be useful for Impact-Based Decision Support Services, but were skeptical of its use for warning operations because the polygons were so large.
- 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.
    - Machine-learning-based algorithms were used to diagnose the likelihood that severe wind reports were actually associated with winds ≥ 50 knots. The primary results were: (1) the regionally trained ML models generally received lower ratings than full CONUS-trained models, (2) the impact of including radar data was relatively small in the subjective ratings with a slight improvement when radar was used to train the GBM model, and (3) the single GBM model received slightly higher subjective ratings than the stack GLM ensemble approach.
    - ML algorithms were trained to provide probabilistic guidance on storm mode using CAM output. It was found that a supervised convolutional neural network (CNN) generally received higher ratings than a partially supervised Gaussian mixture model (GMM). The most common concern from participants was hour-to-hour inconsistency in the probabilities, especially those from GMM.
    - Evaluation of two hourly versions of 3D-RTMA with different backgrounds revealed that the FV3-based GSL version of the 3D-RTMA was rated subjectively "much worse" to

“slightly worse” than the HRRR-based EMC version, with the caveat that the GSL version is still under early stages of development.

- A sub-hourly version of 3D-RTMA, which used HRRRDAS for background error covariance, was run by EMC for comparison with the EMC hourly version, which uses GDAS for background error covariance. Although the 15-minute version of the 3D-RTMA was rated subjectively “about the same” to “slightly better” than the hourly version, over 70% of participants believed that the more frequent updates in the 15-minute version provided value relative to the hourly system.
- 15-minute forecasts of 10-m and 80-m winds from WoFS were used as a proxy for the analysis of severe wind. Overall, the WoFS ensemble maximum winds were positively viewed in terms of lining up with preliminary severe wind reports, as more than three-fourths of the participants gave neutral or positive ratings. The 80-m winds received higher subjective ratings and were found to better match the magnitudes of any measured gusts than the 10-m winds.
- To assess the readiness of the Global Ensemble Forecast System (GEFS) to replace the SREF, an evaluation was performed during the 2021 HWT SFE. For severe weather applications at Day 2 & 3 lead times, GEFS generally performed as well as the SREF, except for MLCAPE, and better than the SREF for calibrated thunder and severe products. For an evaluation of the Extended Range Convective Forecast Product (ECFP), which was conducted for the Aviation Weather Center, it was found that GEFS provided notable improvement over the SREF-based products.

Overall, the 2021 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 2021 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 FV3-SAR for CAM forecasting applications, and explore new ways to leverage AI-based strategies for calibrating and post-processing CAM output to aid forecasters. Additionally, we expect that this work will take on particular importance and aid with evidence-based decision making as NOAA moves forward with its plans for a Unified Forecasting System. In the second year of a virtual experiment, we emphasize that – although we have been successful at accomplishing our mission – science-based discussions and establishing new collaborations are more difficult in the virtual environment. Moving forward, we believe that the lessons learned from virtual experiments could benefit in a future hybrid approach involving both in-person and virtual participation.

**Acknowledgements**

The 2021 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, ESRL/GSL, GFDL, OU MAP, and EMC were vital to the success of the 2021 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), Jacob Carley (EMC), Rajendra Panda (EMC), Jim Abeles (EMC), Jili Dong (EMC), Matt Pyle (EMC), Ben Blake (EMC), Eric Rogers (EMC), Eric Aligo (EMC), Xiaoyan Zhang (EMC), Ting Lei (EMC), Shun Liu (EMC), Logan Dawson (EMC), Perry Shafran (EMC), Manuel Pondeca (EMC), Edward Colon (EMC), Matthew Morris (EMC), Gang Zhao (EMC), Annette Gibbs (EMC), Lucas Harris (GFDL), Matthew Morin (GFDL), Kai-Yuan Cheng (GFDL), Linjiong Zhou (GFDL), Xuguang Wang (OU MAP), Yongming Wang (OU MAP), and Nick Gasperoni (OU MAP), were essential in generating and providing access to model forecasts or products examined on a daily basis.

## References

Burke, A., N. Snook, D. J. Gagne, S. McCorkle, and A. McGovern, 2020: Calibration of Machine Learning-Based Probabilistic Hail Predictions for Operational Forecasting. *Wea. Forecasting*, **35**, 149-168.

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.

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

Hill, A. J., G. R. Herman, and R. S. Schumacher, 2020: Forecasting severe weather with random forests. *Mon. Wea. Rev.*, **148**, 2135–2161, https://doi.org/10.1175/MWR-D-19-0344.1.

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

Jahn, D. E., B. t. Gallo, C. Broyles, B. T. Smith, I. Jirak, and J. Milne, 2020: Refining CAM-based Tornado Probability Forecasts Using Storm-inflow and Storm-attribute Information. Preprints, *30th Conf. On Weather Analysis and Forecasting/26th Conf. on Num. Wea. Prediction*, Boston, MA, Amer. Meteor. Soc., 2A.4.

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

Jirak, I. L., M. S. Elliott, C. D. Karstens, R. S. Schneider, P. T. Marsh, and W. F. Bunting, 2020: Generating Probabilistic Severe Timing Information from SPC Outlooks using the HREF. Preprints, *30th Conf. On Weather Analysis and Forecasting/26th Conf. on Num. Wea. Prediction*, Boston, MA, Amer. Meteor. Soc., 3.1.

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.

Loken, E. D., A. J. Clark, and C. D. Karstens, 2020: Generating Probabilistic Next-Day Severe Weather Forecasts from Convection-Allowing Ensembles Using Random Forests. Wea. Forecasting (In Press).

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.

Sobash, R. A., G. S. Romine, and C. S. Schwartz, 2020: A comparison of Neural-Network and Surrogate-Severe Probabilistic Convective Hazard Guidance Derived from a Convection-Allowing Model. *Wea. Forecasting*, **35**, 1981-2000.

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

Wang, Y., J. Gao, P. S. Skinner, K. Knopfmeier, T. Jones, G. Creager, P. L. Heinselman, and L. J. Wicker, 2019: Test of a Weather-Adaptive Dual-Resolution Hybrid Warn-on-Forecast Analysis and Forecast System for Severe Weather Events. *Wea. Forecasting*, **34**, 1807-1827.

## *APPENDIX*

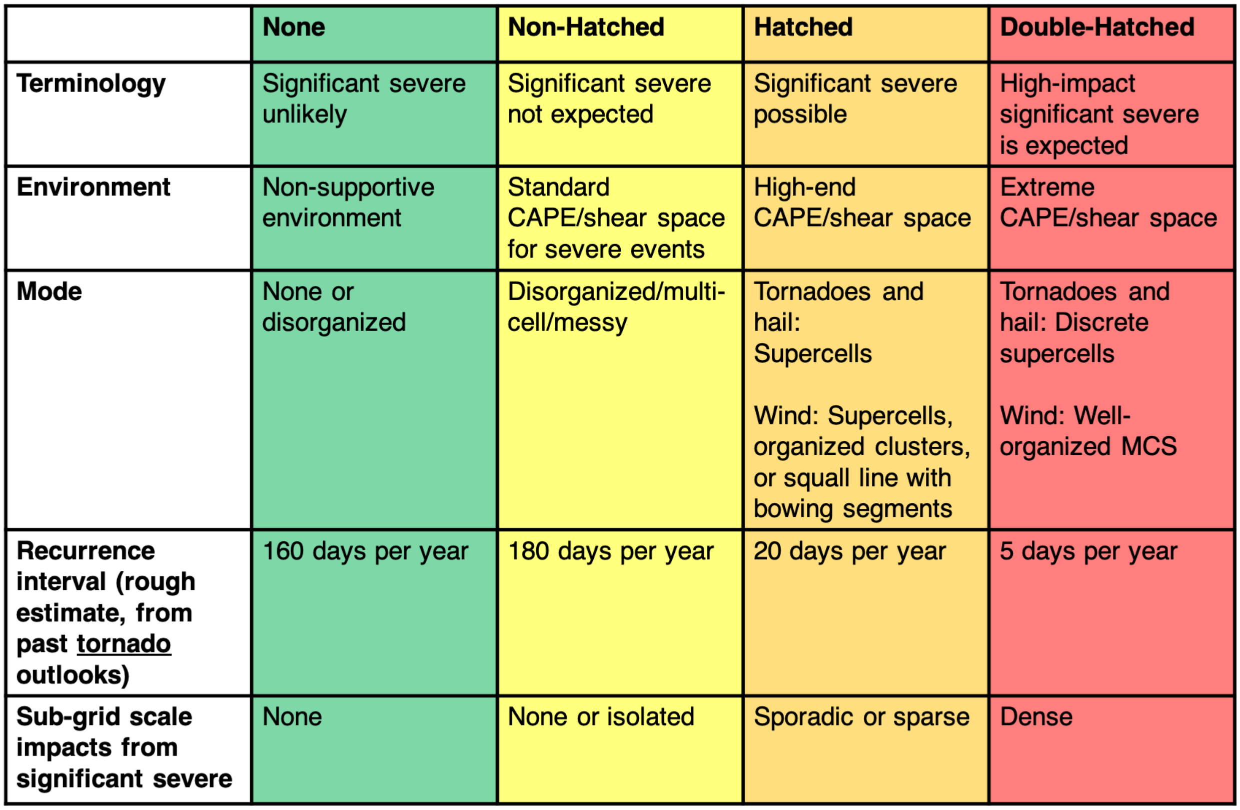
*Table A1 Weekly participants during the 2021 SFE. SFE facilitators included Adam Clark (NSSL), Israel Jirak (SPC), Dave Imy (retired SPC), Burkely Gallo (CIMMS/SPC), Kenzie Krocak (CIMMS/SPC/CRCM), Brett Roberts (CIMMS/SPC/NSSL), Kent Knopfmeier (CIMMS/NSSL), Andy Dean (SPC), Eric Loken (CIMMS/NSSL), David Harrison (CIMMS/SPC), David Jahn (CIMMS/SPC), Jacob Vancil (CIMMS/SPC), Jeff Milne (CIMMS/SPC), and Nathan Dahl (CIMMS/SPC).*

| **Week 1** | **Week 2** | **Week 3** | **Week 4** | **Week 5** |
|---|---|---|---|---|
| **3-7 May** | **10-14 May** | **17-21 May** | **24-28 May** | **1-4 June** |
| Rick Garuckas (WFO MRX) | Nick Hampshire (WFO EWX) | Anna Lindeman (WFO BOI) | Heather Kenyon (WFO BUF) | Andrew Zimmerman (WFO AKQ) |
| John Wetenkamp (WFO ARX) | Francis Kredensor (WFO TFX) | Kevin Huyck (WFO DLH) | Emily McGraw (WFO CHS) | Michael Sporer (WFO RNK) |
| Pat Spoden (WFO PAH) | Steve Zubrick (WFO LWX) | Chad Entremont (WFO JAN) | Nate McGinnis (WFO ILN) | Tara Dudzik (WFO IND) |
| Kristen Cassady (WFO ILN) | Matthew Brady (WFO EWX) | Jack Settlemaier (NWS SRH) | Linda Gilbert (WFO MQT) | Nick Vertz (WFO BYZ) |
| Lee Robertson (WFO PHI) | Keith Sherburn (WFO UNR) | Nicholas Fenner (WFO JAN) | James Wood (WFO MKX) | Aidan Kuroski (WFO MKX) |
| Dirk Peterson (WFO OAX) | Eswar Iyer (WFO AKQ) | Jaclyn Anderson (Ritzman) (WFO MKX) | Eric Bunker<br>(M-Th; WFO TAE) | Keith White (WFO EWX) |
| Stephen Harrison (WFO SJT) | Brian Carcione (WFO HUN) | Lizzie Tirone (ISU) | Maria Molina (NCAR) | Austin Coleman (TTU) |
| Evan Kuchera (USAF) | Ed Shimon (WFO ILX) | Clark Evans (UWM) | Lance Bosart (SUNY-Albany) | Chris Melick (USAF) |
| Greg Stumpf (NSSL) | Bill Gallus (ISU) | Dillon Blount (UWM) | Steve Weiss (ret. SPC) | Craig Schwartz (NCAR) |
| Chris Karstens (SPC) | Becky A.-Selin (AER) | Felicia Guarriello (WPO) | Harald Richter (BOM) | Mike Coniglio (NSSL) |
| Jamie Wolff (DTC) | Russ Schumacher (CSU) | Casey Davenport (UNCC) | Reid Strickler (USAF) | Derek Stratman (CIMMS) |
| Aaron Hill (CSU) | John Peters (Naval P.-Grad) | Roger Riggin (UNCC) | Gary Lackmann (NCSU) | Jidong Gao (NSSL) |
| Leigh Orf (Wisc) | J. Peters student #1 | Kyle Struckmann (NWS NAM) | Trevor Campell (NCSU) | Lewis Kanofsky (AWC) |
| Gabrielle Gantos (NCAR) | J. Peters student #2 | Nick Goldacker (NCSU; M. Parker student) | Jacob Radford (NCSU) | Kai-Chih Tseng (GFDL/Princeton) |
| Rob Hepper (AWC) | Dave Ahijevych (NCAR) | Andrew Winters (CU) | Jeff Beck (CIRA/GSL/DTC) | Tim Marchok (GFDL) |
| Chris Nowotarski (TAMU) | Ty Higginbotham (AWC) | Rebecca Baiman (CU) | Nat Johnson (GFDL) | Kelly Lombardo (PSU) |
| Matt Brown (TAMU) | Tomas Pucik (ESSL) | Alexandra A.-Frey (UW) | Kelton Halbert (Wisc) | Geoff Manikin (EMC) |
| Brice Coffer (NC State) | Francesco Battaglioli (ESSL) | Rohan Jain (UW) | Jacob Carley (EMC) | Matthew Pyle (EMC) |
| Chris MacIntosh (EMC) | Binbin Zhou (EMC) | Charlie Becker (NCAR) | Gang Zhao (EMC) | Kendall Junker (CAPS/OU) |
| Shun Liu (EMC) | Shannon Shields (EMC) | Logan Dawson (EMC) | Ben Blake (EMC) | Jana Houser (U. Ohio) |
| Xiaoyan Zhang (EMC) | Matthew Morris (EMC) | Annette Gibbs (EMC) | Nigel Roberts (UK Met) | Darby Johnson (U. Ohio) |
| Nick Silkstone (UK Met) | Stephen Gallagher (UK Met) | Travis Elless (EMC) | Matt Lehnert (UK Met) | Curtis Alexander (GSL) |
| Aurore Porson (UK Met) | Adrian Semple (UK Met) | Sebastian Cole (UK Met) | Steve Willington (UK Met) | Dan Dawson (Purdue) |
| David Dowell (GSL) | Chris Bulmer (UK Met) | Steve Willington (UK Met) | Eric James (GSL) | Allie Mazurek (CSU) |
| Sarah Trojniak (WPC) | Aaron Johnson (OU/MAP) | Nate Snook (CAPS) | John Brown (GSL) | Ben Henry (Princeton undergrad) |
|  | Jeff Duda (GSL) | Xuguang Wang (OU/MAP) | Mike Baldwin (Purdue) |  |
|  | John Allen (CMU) | Terra Ladwig (GSL) | Geeta Nain (Purdue) |  |
|  |  | Ed Szoke (GSL) |  |  |
|  |  | Jordan Dale (WPO) |  |  |

*Table A2 Schedule for Tuesday – Friday. On Mondays, the schedule is similar except the period 9-11:15am is devoted to training and introductory material.*

| Time (CDT) | **R2O Group** | | **Innovation Group** | |
|---|---|---|---|---|
| 9:00 AM – 9:15 AM | **Overview of Yesterday's Severe Weather**<br>David Imy | | | |
| 9:15 AM – 11:00 AM | **Evaluation Orientation, Individual Working Time, and Discussion** | | | |
| | **Group A: Calibrated Guidance** | **Group B: Deterministic CAM** | **Group C: CAM Ensembles** | **Group D: Medley** |
| 11:00 AM - 11:15 AM | ***Break*** | | | |
| 11:15 AM – 11:30 AM | **Weather Briefing**<br>David Imy | | | |
| 11:30 AM – 12:30 PM | **Issue *Day 1* Hazards Coverage and Conditional Intensity Forecasts (2 groups)** | | **Issue *Day 2* Hazards Coverage and Conditional Intensity Forecasts (2 groups)** | |
| | *12z HREF* | *12z GSL RRFS* | *No CAMs* | *All data (incl. CAMs)* |
| 12:30 PM – 2:00 PM | ***Lunch/Break*** | | | |
| 2:00 PM – 2:15 PM | **Update on Today's Weather**<br>David Imy | | | |
| 2:15 PM – 3:00 PM | **Issue MD Product** | | **Issue 1-h outlooks (22-23, 23-00Z)** | |
| | *WoFS & obs* | | *WoFS* | *No WoFS* |
| 3:00 PM – 4:00 PM | **Update Day 1 Outlook** | | **Issue 1-h outlooks (22-23, 23-00, 00-01Z)** | |
| | *WoFS & other guidance* | | *WoFS* | *No WoFS* |

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



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

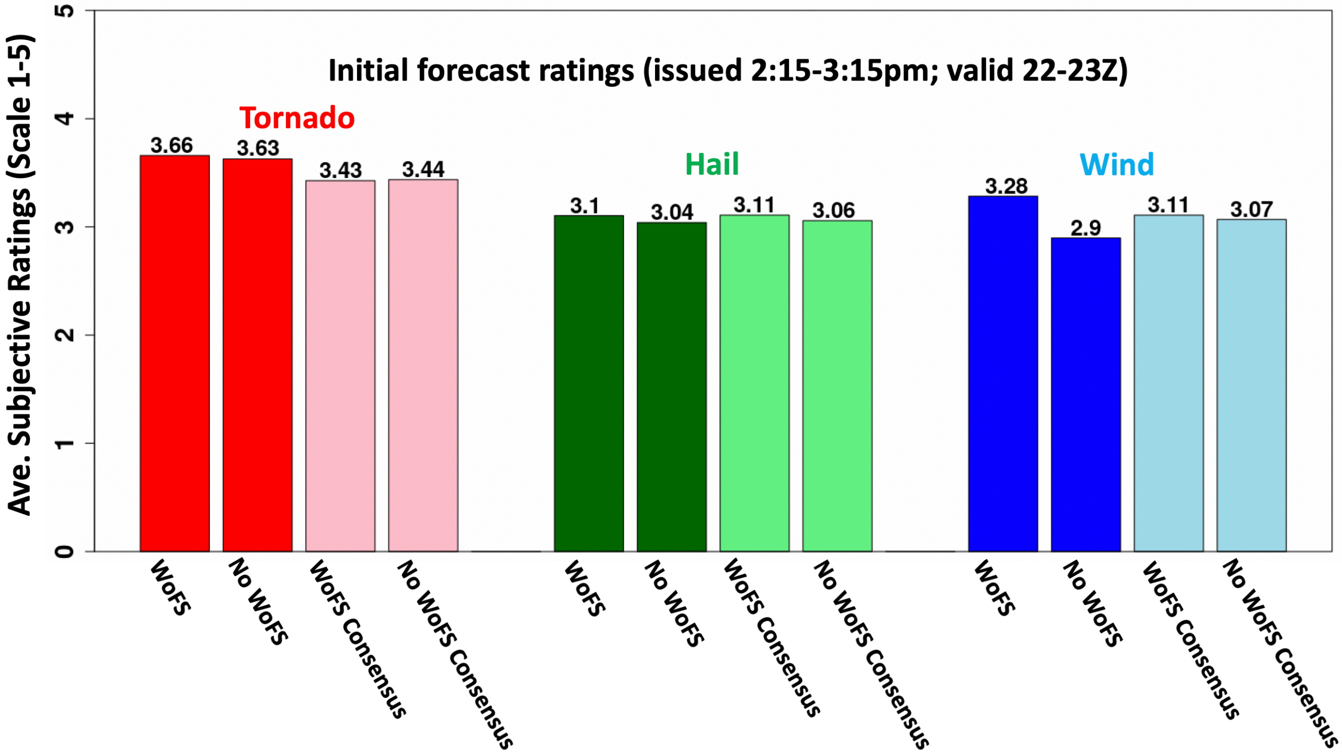


*Figure A1 Average subjective ratings of initial forecasts valid 22-23Z for the Innovation Group afternoon forecasting activity during SFE 2021.*

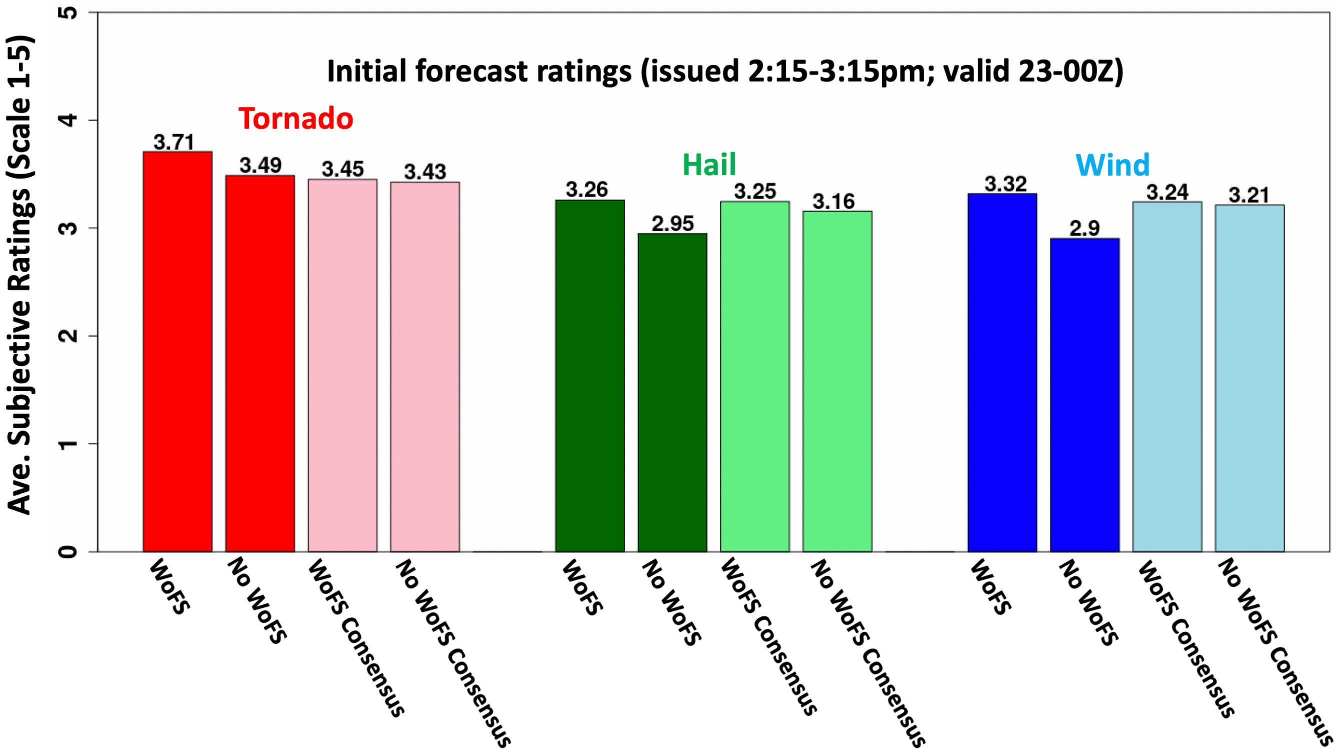


*Figure A2 Same as A1, except for forecasts valid 23-00Z.*

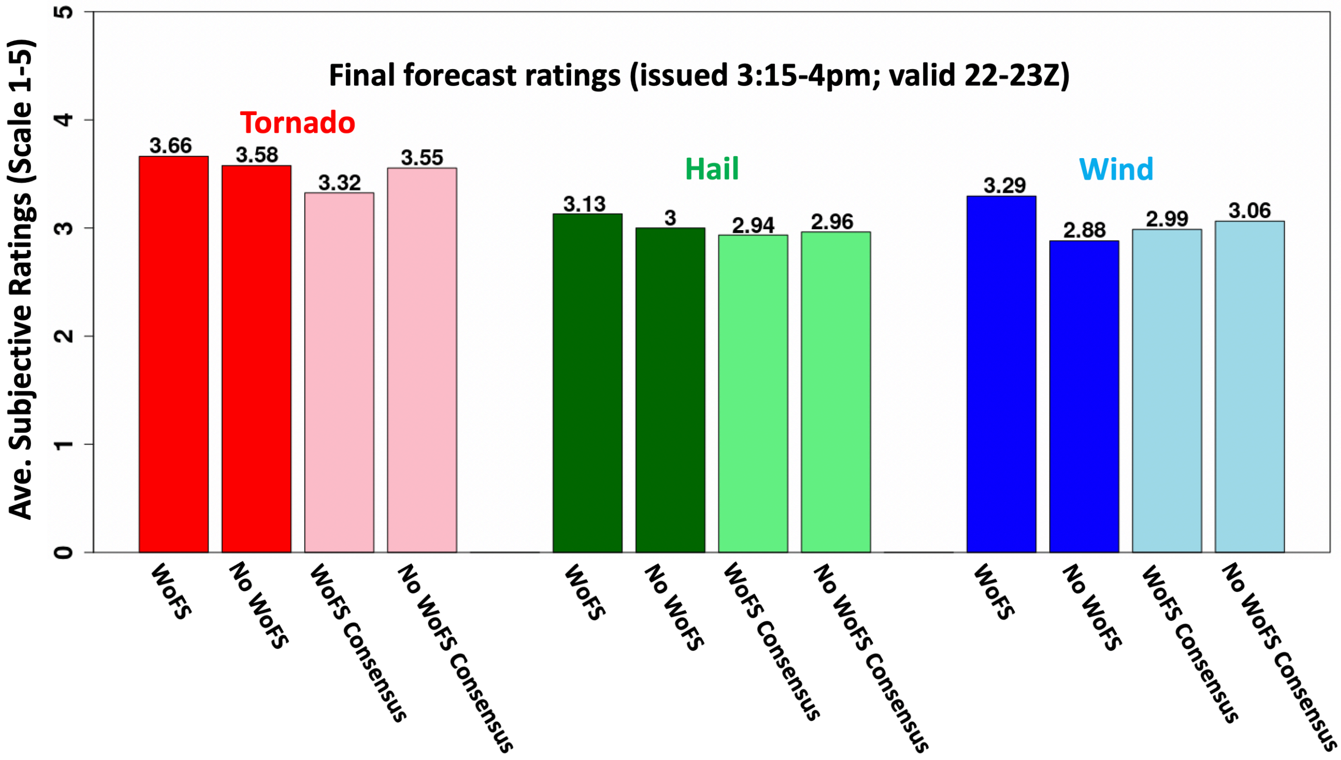


*Figure A3 Same as A1, except for final forecasts valid 22-23Z.*

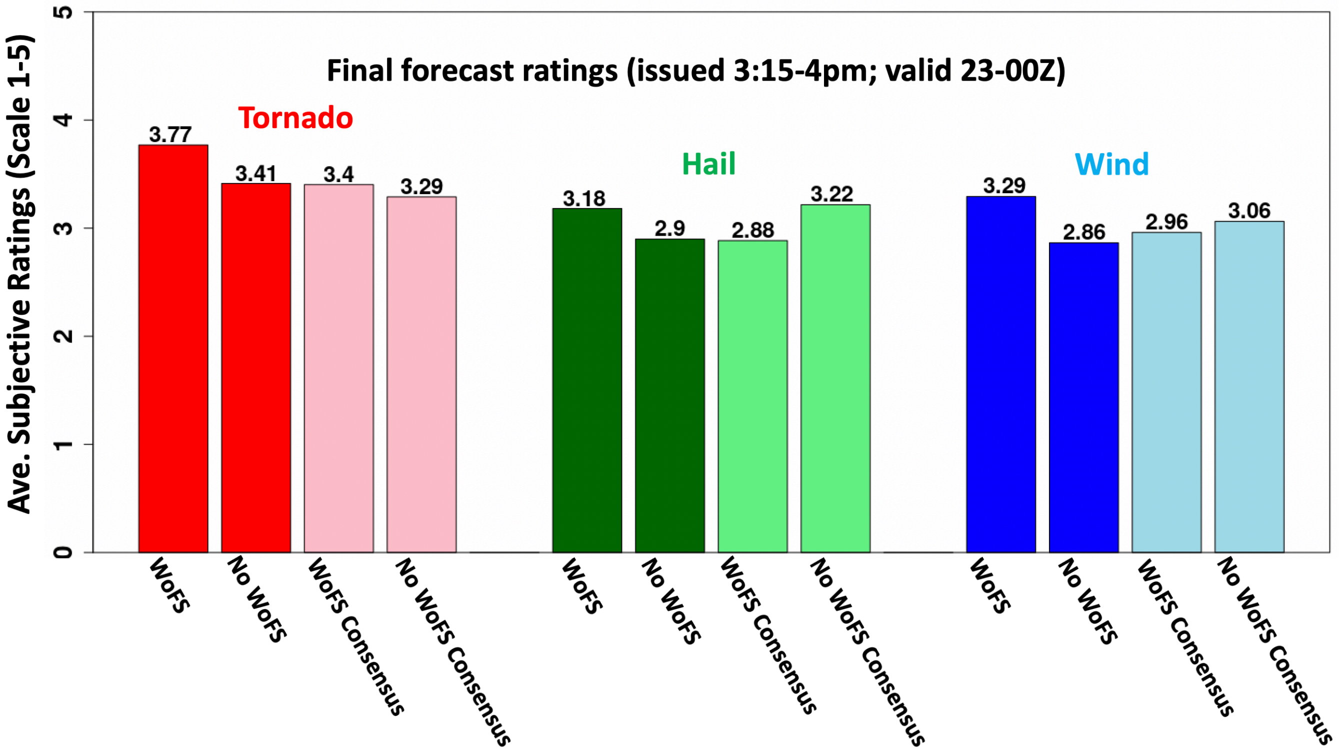


*Figure A4 Same as A1, except for final forecasts valid 23-00Z.*

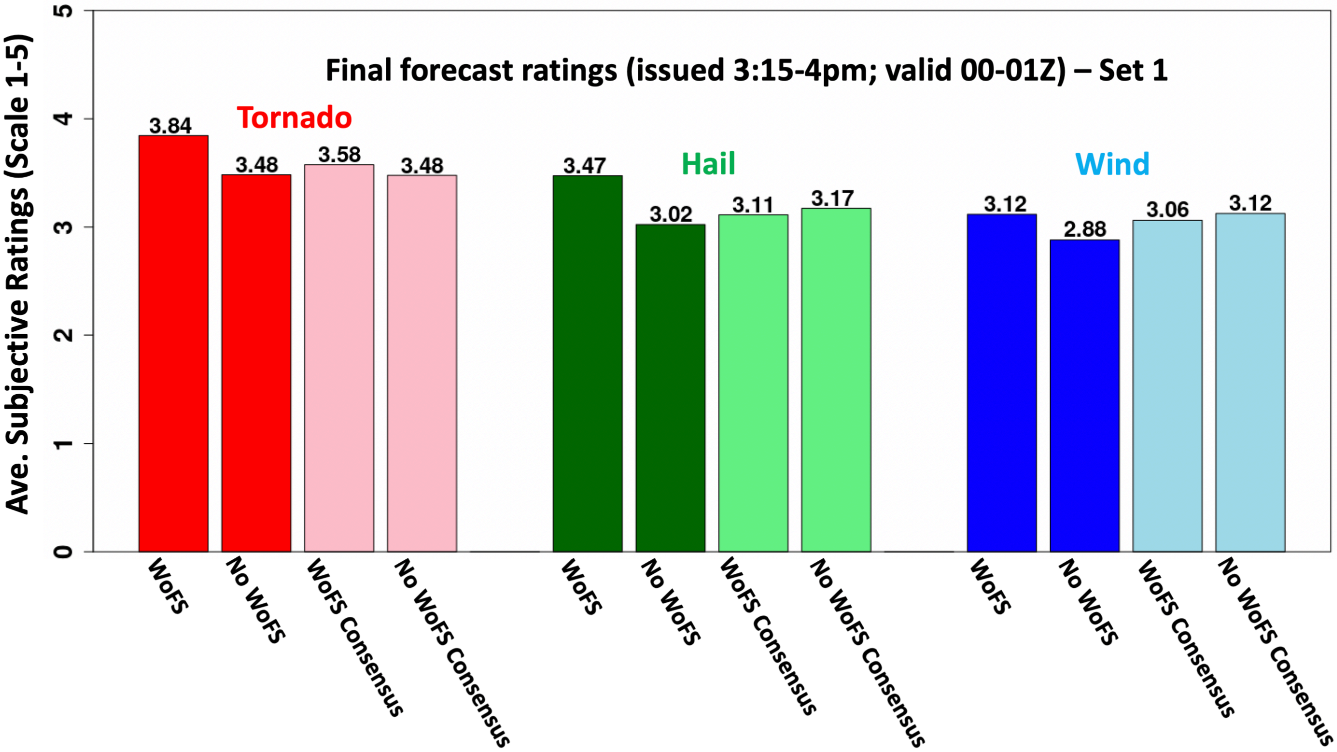


*Figure A5 Same as A1, except for final forecasts valid 00-01Z (Set 1).*

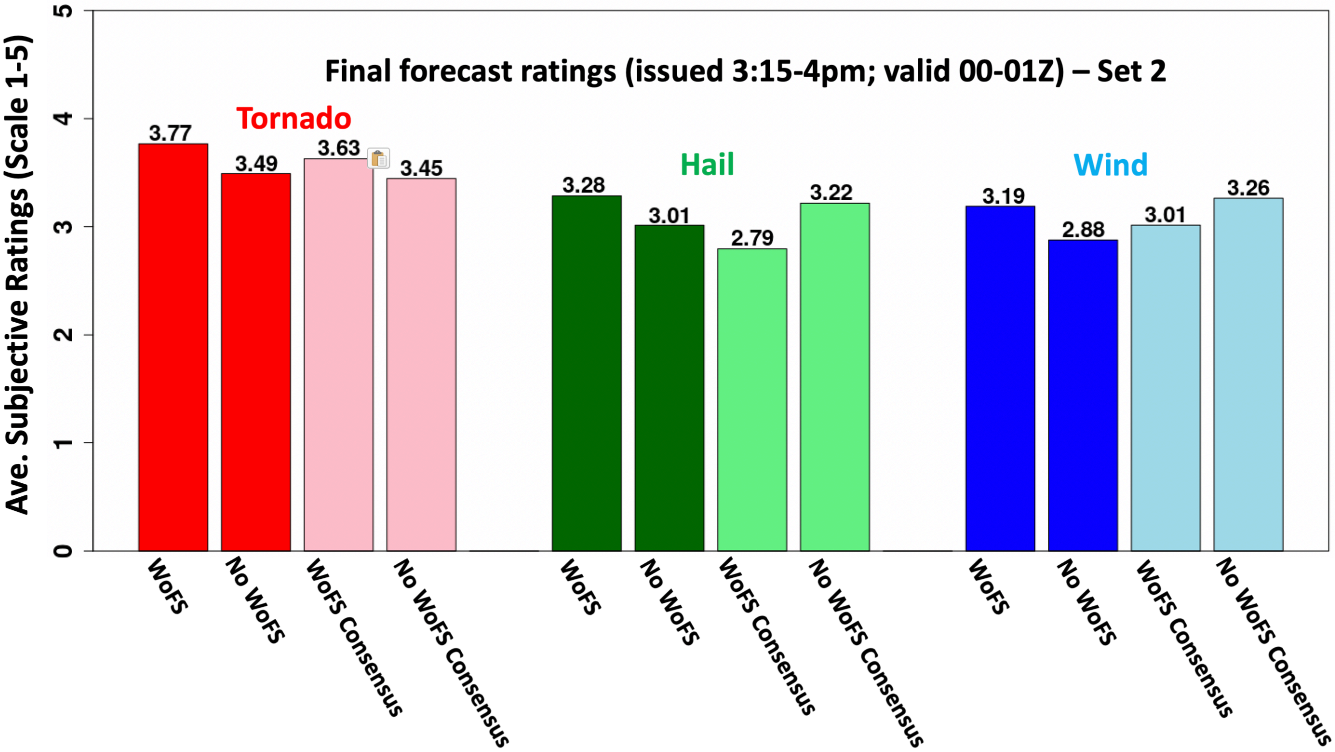


*Figure A6 Same as A1, except for final forecasts valid 00-01Z (Set 2).*