https://doi.org/10.25923/r70a-aw71

# SPRING FORECASTING EXPERIMENT 2022

## Conducted by the

# EXPERIMENTAL FORECAST PROGRAM

## of the

# NOAA HAZARDOUS WEATHER TESTBED

**https://hwt.nssl.noaa.gov/sfe/2022**

**Virtual Experiment**
**2 May – 3 June 2022**

# Preliminary Findings and Results

Adam Clark[2], Israel Jirak[1], Burkely T. Gallo[1,3], Brett Roberts[1,2,3], Kent Knopfmeier[2,3], Jake Vancil[1,3], David Jahn[1,3], Makenzie Krocak[3,4,5], Chris Karstens[1], Eric Loken[2,3], Nathan Dahl[1,3], David Harrison[1,3], David Imy[2], Andy Wade[1,3], Jeffrey Milne[1,3,4], Kimberly Hoogewind[2,3], Montgomery Flora[2,3], Joshua Martin[2,3], Brian Matilla[2,3], Joey Picca[1,3], Corey Potvin[2], Patrick Skinner[2,3], Patrick Burke[2]

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

# Table of Contents

## List of Figures

## List of Tables

# Executive Summary

The Hazardous Weather Testbed (HWT) is a space in the National Weather Center Building in Norman, Oklahoma that facilitates forecasting experiments testing new concepts, tools, and algorithms developed at NOAA's National Severe Storms Laboratory (NSSL), Storm Prediction Center (SPC), and their partner institutions. Conducted annually during the peak severe weather season since 2000, the Spring Forecasting Experiment, or SFE, is the longest running HWT experiment. The SFEs are co-led by SPC and NSSL and aim to accelerate research to operations through testing new severe weather prediction tools and forecasting methods, studying how end-users apply severe weather guidance, and facilitating experiments for optimizing convection-allowing model ensemble design to inform NOAA's Unified Forecast System (UFS). The wealth of severe weather forecasting and research expertise at the National Weather Center, combined with state-of-the-art visualization tools, well-designed experiments, and valuable collaborations have made the annual SFEs one of the most productive and well-respected weather forecasting experiments in the world. SFE 2022 forecasting activities emphasized using experimental calibrated and machine learning (ML)-based products, with two data denial experiments withholding this guidance from control groups. Model evaluations used the 61-member Community Leveraged Unified Ensemble (CLUE) and included examinations of deterministic and ensemble prototypes for the Rapid Refresh Forecast System (RRFS).

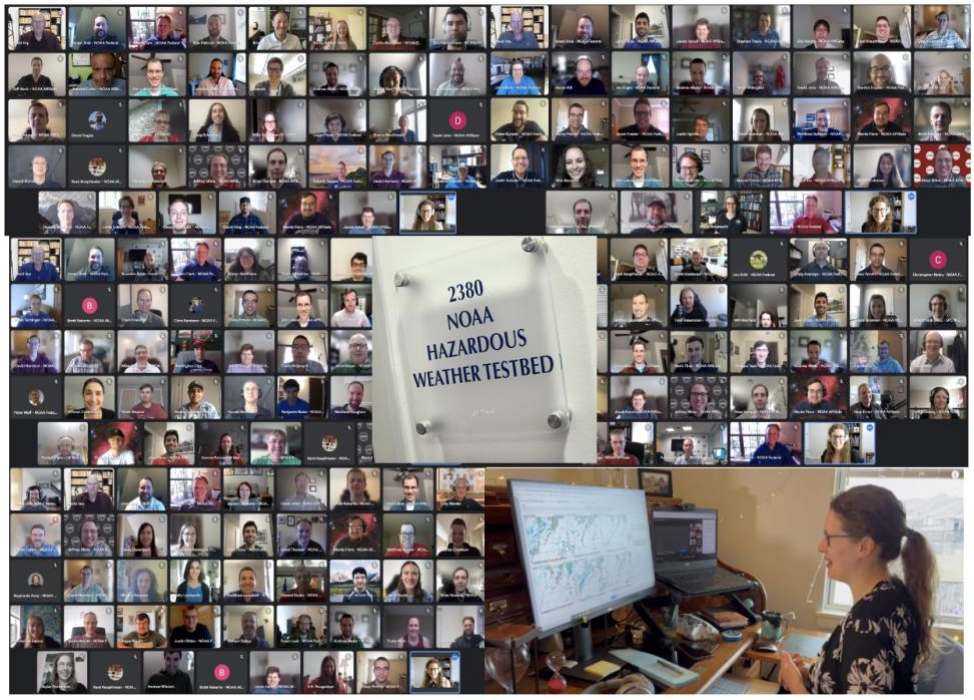


*Figure 1. Scenes and participant screenshots from each week of the 2022 NOAA Hazardous Weather Testbed Spring Forecasting Experiment.*

# 1. Introduction

The 2022 Spring Forecasting Experiment (2022 SFE) was conducted from 2 May – 3 June by the Experimental Forecast Program (EFP) of the NOAA/Hazardous Weather Testbed (HWT), and was co-led by the NWS/Storm Prediction Center (SPC) and OAR/National Severe Storms Laboratory (NSSL). Additionally, important contributions of convection-allowing models (CAMs) were made by NOAA collaborators: Global Systems Laboratory (GSL), Environmental Modeling Center (EMC), and Geophysical Fluid Dynamics Laboratory (GFDL); as well as University of Oklahoma collaborators: the Multi-scale data Assimilation and Predictability (MAP) group and the Center for Analysis and Prediction of Storms (CAPS). Participants included over 165 forecasters, researchers, model developers, university faculty, and graduate students from around the world (see Table A1 in the Appendix). Uncertainties related to the COVID-19 pandemic precluded an in-person experiment for the third consecutive year, but to maintain momentum in key areas of convection-allowing model development, the HWT EFP once again conducted the 2022 SFE virtually, building upon lessons learned from the two previous virtual experiments. As in previous years, the 2022 SFE aimed to test emerging concepts and technologies designed to improve the prediction of hazardous convective weather, consistent with the Forecasting a Continuum of Environmental Threats (FACETs; Rothfusz et al. 2018) and Warn-on Forecast (WoF; Stensrud et al. 2009) visions. Below are goals from the 2022 HWT SFE for product and service improvements and applied science activities.

Product and Service Improvements:

- Assess the utility of machine-learning (ML) guidance coupled with 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 the ML guidance, and surveying participants on the experience of using the ML guidance to issue these forecasts.
- 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, & 3.
- Test the utility of WoFS for updating coverage and conditional intensity full-period hazards forecasts valid 2100–1200 UTC.
- Explore the application and utility of calibrated guidance products for issuing Day 1 hazards forecasts valid 1800–1200 UTC by generating forecasts with and without calibrated guidance.
- 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).
- Conduct a focus group activity to gain insight on the conditional intensity products.

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 operational 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 three machine-learning techniques for producing probabilistic convective mode guidance using High-Resolution Rapid Refresh (HRRR) forecasts as input.
- Evaluate configurations of the limited area Finite Volume Cubed Sphere Model (FV3-LAM) with different data assimilation (DA) and physics suites.
- Examine whether increasing horizontal grid-spacing from 3- to 1-km in Weather Research and Forecasting (WRF) model simulations provides benefits for tornado prediction and the strength of convective wind gusts.
- Use an ensemble sensitivity-based ensemble subsetting approach to identify a small subset of members with the smallest errors out of 20 CLUE members and examine whether severe weather guidance derived from these subsets are improved relative to the full 20-member ensemble.
- Compare and assess different versions of the 3D real-time mesoscale analysis (3D-RTMA) system that use different sources for the background first guess.
- Test WoFS-based analyses of 10-m and 80-m wind speeds 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 2022 collaborator.
- Evaluate WoFS for applications to short-term severe weather product generation, and examine the impact of reducing the number of WoFS members, as well as different time-lagging approaches.
- Explore the “Threats in Motion” concept applied to county-based watch guidance derived from an ML model that uses HREF fields as predictors.

A suite of state-of-the-art experimental CAM guidance contributed by our large group of collaborators was critical to the 2022 SFE. For the seventh consecutive year, these contributions were formally coordinated into a single ensemble framework called the Community Leveraged Unified Ensemble (CLUE; Clark et al. 2018). The 2022 CLUE was constructed by having all groups coordinate as closely as possible on model specifications (e.g., version, grid-spacing, vertical levels, physics, etc.), domain, and post-processing so that the simulations contributed by each group could be used in controlled experiments. This design allowed us to conduct several experiments to aid in identifying optimal configuration strategies for CAM-based ensembles. The 2022 CLUE included 60 members using 3-km grid-spacing (one member with 1-km grid spacing), which allowed for several unique experiments. The 2022 SFE activities also involved testing the WoFS for the sixth consecutive year.

This document summarizes the activities, core interests, and preliminary findings of the 2022 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/2022/docs/HWT_SFE2022_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 2022 SFE along with a description of the daily activities, Section 3 reviews the preliminary findings of the 2022 SFE, and Section 4 contains a summary of these findings and some directions for future work.

## 2. Description

### 2.1 Experimental Models and Ensembles

A total of 89 unique CAMs were run for the 2022 SFE, of which 61 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 2022 SFE emphasized the use of CAM ensembles [i.e., HREF, Rapid Refresh Forecasting System (RRFS) prototypes, and WoFS] in generating experimental probabilistic forecasts of individual severe weather hazards. Additionally, the 2022 CLUE configuration enabled numerous scientific evaluations focusing on model sensitivities and various ensemble configuration strategies.

To put the volume of CAMs run for 2022 SFE into context, Figure 2 shows the number of CAMs run for SFEs since 2007, which was the first year CAM ensembles were contributed to the SFE. In general, Figure 2 shows an increasing trend through 2019 and then stabilization around 90 CAMs. The consolidation of members into the CLUE has made this large volume of CAMs more manageable and has facilitated more controlled scientific comparisons.

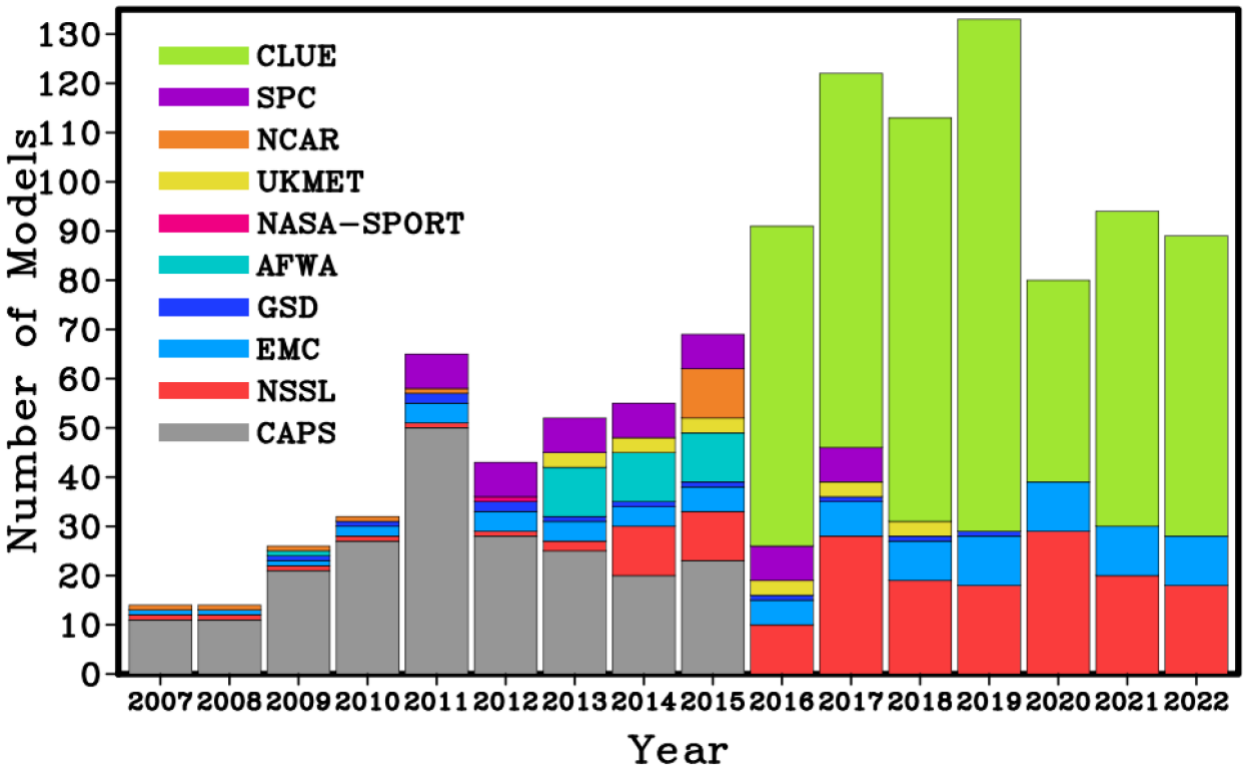


*Figure 2. Number of CAMs run for SFEs since 2007. The different colored stacked bars indicate the contributing agencies.*

More information on all of the modeling systems run for the 2022 SFE is given below.

### 2.1.1 The Community Leveraged Unified Ensemble (CLUE)

The 2022 CLUE is a carefully designed ensemble with subsets of members contributed by NOAA groups at NSSL, GFDL, GSL, and EMC, and the non-NOAA groups of OU-MAP and CAPS. The 60 CLUE members with 3-km grid-spacing have a CONUS domain, while the single 1-km member has a 2/3 CONUS domain. Depending on the CLUE subset, forecast lengths range from 30 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, output fields, and further details on ensemble configurations are provided in the 2022 SFE operations plan. Table 1 provides a summary of each CLUE subset.

| Clue Subset | # of mems | IC/LBC perts | Mixed Physics | Data Assimilation | Dynamical Core | Agency | Init. Times (UTC) | Forecast Length (h) | Domain |
|---|---|---|---|---|---|---|---|---|---|
| RRFSp1 | 1 | none | no | Hybrid 3DEnVar | FV3 | EMC/GSL | 00, 12 | 60 | CONUS |
| RRFSp2e | 10 | EnKF | no | Hybrid 3DEnVar | FV3 | EMC/GSL | 00 | 36 | CONUS |
| MAP-VTS-rad | 10 | GFS, GEFS | no | GSI-EnVar | FV3 | OU-MAP | 00 | 36 | CONUS |
| MAP-VTS-con | 10 | GFS, GEFS | no | GSI-EnVar | FV3 | OU-MAP | 00 | 36 | CONUS |
| MAP-VTS-bot | 10 | GFS, GEFS | no | GSI-EnVar | FV3 | OU-MAP | 00 | 36 | CONUS |
| NSSL-FV3-LAM | 1 | none | no | GFS cold start | FV3 | NSSL | 00 | 60 | CONUS |
| NSSL3 | 1 | none | no | GFS cold start | ARW | NSSL | 00 | 30 | CONUS |
| NSSL1 | 1 | none | no | GFS cold start | ARW | NSSL | 00 | 30 | 2/3 CONUS |
| GFDL-FV3 | 1 | none | no | GFS cold start | FV3 | GFDL | 00 | 126 | CONUS |
| RRFSp2eMP | 10 | EnKF | yes | Hybrid 3DEnVar | FV3 | CAPS | 00 | 84 | CONUS |
| RRFSphys | 6 | none | yes | Hybrid 3DEnVar | FV3 | CAPS | 00 | 36 | CONUS |

*Table 1. Summary of the 11 unique subsets that comprise the 2022 CLUE.*

The design of the 2022 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 below.

Valid Time Shifting Data Assimilation

- **Description:** The OU MAP group ran ensembles with Valid Time Shifting (VTS) applied to radar data only, as well as radar data and conventional observations. VTS 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 the value of the VTS approaches applied to different sets of observations.
- **CLUE subsets:** MAP-VTS-rad & MAP-VTS-bot

RRFS Configuration Strategies

- **Description:** Several different ensembles contributed by GSL, EMC, OU-MAP and CAPS 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.
- **CLUE Subsets:** RRFSp2e, RRFSp2eMP, MAP-VTS-rad, & MAP-VTS-bot

FV3-LAM Physics

- **Description:** CAPS ran several configurations of FV3-LAM that were identical except for their physics packages.
- **Goal:** Assess systematic differences and performance characteristics among the different physics suites.

- **CLUE Subsets:** RRFSphys

FV3-LAM Data Assimilation

- **Description:** EMC and GSL ran two deterministic RRFS prototypes. Prototype 1 used partially cycled (hourly) DA with GDAS (Global Data Assimilation System). Prototype 2 starts from GDAS, but then engages an hourly cycled storm scale ensemble EnKF-based system that informs hybrid deterministic analyses from which a deterministic forecast is launched.
- **Goal:** Determine the impact of the more sophisticated DA approach (similar to RAP/HRRR, but in UFS framework), with an emphasis on the first 12 h of the forecast.
- **CLUE Subsets:** RRFSp1 & RRFSp2

Enhanced Resolution

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

3DRTMA Background

- **Description:** Three hourly versions of 3D-RTMA were compared and each used a different background first-guess.
- **Goal:** Assess the impact of the background first guess on the final analysis.
- **3DRTMA Version:** 3DRTMAp1, 3DRTMAp2, & 3DRTMA HRRR

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

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

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

The cloud-based Warn-on-Forecast System (cb-WoFS) is the next WoFS iteration, upgraded to use current technologies in containerization and cloud computing on the Microsoft Azure platform. The cb-WoFS is a rapidly-updating 36-member, 3-km grid-spacing WRF-based ensemble data assimilation and forecast system. The cb-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. 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 High-Resolution Rapid Refresh Data Assimilation System (HRRRDAS) and the 1200 UTC HRRR forecast provided by NCO/GSL. A 1-h forecast from the 1400 UTC, 36-member, hourly-cycled HRRRDAS analysis provided the ICs for cb-WoFS. Boundary conditions were perturbed HRRR forecasts, where perturbations from the 0600 UTC GEFS were added to the 1200 UTC HRRR forecasts. The GEFS perturbations were scaled such that the ensemble spread at the lateral boundaries was similar to that provided previously by the experimental HRRR ensemble. All cb-WoFS forecasts were made available via the cb-WoFS Forecast Viewer at: https://cbwofs.nssl.noaa.gov/Forecast. Hereafter, cb-WoFS will simply be referred to as WoFS.

## 2.2 Daily Activities

SFE 2022 activities were focused on forecasting severe convective weather and evaluating the previous day's model forecasts. A summary of evaluation activities and forecast products can be found below while a detailed schedule of daily activities is contained in the appendix (Table A2). Note, when referencing the times in this document at which experiment activities occurred, we use Central Daylight Time (CDT), which is the time zone in which the HWT facility and SFE organizers are based. However, it is worth noting that many of our virtual participants were located in different time zones as far away as the United Kingdom and Australia, so their local time was quite different.

### 2.2.1 Forecast and Model Evaluations

SFE 2022 featured a period of formal evaluations from 9:15–11 am CDT Tuesday-Friday (except for Wednesday-Friday in the last week), for a total of 19 days of evaluation. The evaluations involved comparisons of different ensemble diagnostics, CLUE ensemble subsets, HREFv3, and WoFS. Additionally, the evaluations of yesterday's experimental forecasts products were conducted during this time, which involved comparing the experimental products to observed local storm reports (LSRs), NWS warnings, and Multi-Radar, Multi-Sensor (MRMS; Smith et al. 2016) radar reflectivity and maximum estimated size of hail (MESH). Participants were split into Groups A, B, C, and D, and each conducted a separate set of model evaluations. Participants stayed in their initial group for two days before switching to a different group for the second two days (one day during the last week), to balance building familiarity with product performance and exposure to multiple new CAMs and tools. 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, troubleshoot issues, and discuss the subjective impressions of the day. After completing the surveys individually, participants were encouraged to discuss their thoughts about the products evaluated as a group.

### 2.2.2 Experimental Forecast Products

The experimental forecasts covered a limited-area domain typically encompassing the primary severe threat area with a domain based on existing SPC outlooks and/or where interesting convective forecast challenges were expected. There were two periods of experimental forecasting activities during SFE 2022. The first occurred from 11:30 am–12:30 pm 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 2:15–4 pm CDT and focused on short-term forecasting applications with WoFS. Additionally, a focus group activity was conducted to gain insight on the conditional intensity products. Participants were split into two groups for the forecasting and focus group 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 calibrated guidance products including ML-based algorithms, while the other group did not use calibrated guidance. Both groups had access to various sets of numerical guidance such as the 1200 UTC initialized HREFv3, as well as numerous observational products (satellite, radar, mesoanalysis, surface observations, etc.). 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 severe reports that are expected to be significant, 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 (2:15–4 pm CDT), the R2O group conducted one forecasting activity from 2:15–3 pm 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 3–4 pm time period, the R2O group split into two sub-groups. In one group, each 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. In the other sub-group, a focus group activity was conducted to gain insight on the conditional intensity products.

In the Innovation Group, during the 2:15–4 pm CDT time period, participants generated severe hazard probabilities valid over 1-h time windows covering 2100–2200 and 2200–2300 UTC. Two initial forecasts were generated during the 2:15–3:15 pm period, and these forecasts were updated during the 3:15–4 pm time period. For both sets of initial and final forecasts, two expert forecasters used WoFS, WoFS-based ML algorithms, and any other available forecast data, while two other expert forecasters used WoFS and any other available forecast data, but *did not* use WoFS-based ML algorithms. Additionally, two other groups of non-expert forecasters issued forecasts with and without the WoFS-based ML algorithms similarly to the expert forecasters, which were combined into consensus forecasts.

## 3. Preliminary Findings and Results

### 3.1 Model Evaluation – Group A: Calibrated Guidance

SFE participants evaluated a series of severe weather hazard guidance forecasts including those for tornadoes, severe winds (≥ 50 kts), and hail (≥ 1 in.). A suite of forecasts was generated using variations of eight calibration methods (Table 2 and more details are included in the SFE 2022 operations plan) stemming from ML approaches or more traditional approaches based on severe weather hazard frequency given one or more storm or environmental parameters. Products represented guidance periods of 24-hours based on CAM data at 0000 UTC and valid at 1200 UTC for Day 1 (the current day), Day 2, or Day 3. Evaluation of guidance products was made relative to preliminary LSR observations that were available the day after an event, as well as WFO warning information and MRMS MESH.

| Method | Method Type | Forecast Type | Labels for plots |
|---|---|---|---|
| GEFS-based ML | ML | T/H/W | GEFS |
| HREF/GEFS calibrated | Traditional | T/H/W | HREF/GEFS |
| Significant Tornado Parameter (STP)-based calibrated | Traditional | T | STP Cal {Circle, MCS-TF, Inflow} |
| ML Random Forecast (RF) | ML | T/H/W | ML RF |
| Nadocast | ML | T | Nadocast |
| HREF/SREF and HREF/HREF | Traditional | T/H/W | HREF/{SREF Ops,SREF Para,HREF} |
| Flow-dependent training of RF models | ML | T/H/W | {Non-, Explicitly, Implicitly} Flow Dependent |
| HRRR-based ML Neural Network (NN), version 2 | ML | T/H/W | HRRR NN V2 |

*Table 2. Calibrated guidance methods specified by type (ML or traditional methods) and forecast type (tornado, T; hail, H; severe wind, W).*

### 3.1.1 Aggregate Evaluation of Calibrated Tornado Guidance

There was a total of 14 Day-1 calibrated tornado guidance products offered for daily evaluation during the SFE. Figure 3 provides a comparison of the aggregate subjective scores by evaluators across 19 cases. Of the 14 Day-1 products, Nadocast had the highest overall mean rating (6.08), although seven other methods shared its median rating of 6.0. The next highest performing methods were both ML models, GEFS_D1 and Explicitly_Flow_Dependent with mean scores of 5.81 and 5.79 respectively. The highest scoring of the traditional calibration methods for Day 1 forecasts was STP_Cal_Inflow with a mean score of 5.68 followed closely by STP_Cal_MCS-TF at 5.62. The similar scores for the Inflow and MCS-TF methods suggests that incorporating storm mode (specifically identifying MCSs) in the calculation of tornado probabilities did not make a large difference. The STP_Cal_Circle method scored lower than the other STP-calibrated methods suggesting a slight benefit in incorporating STP values strictly from the “inflow” region rather than the full (circular) near-storm environment for this set of cases.

Of the Day-1 guidance products, several were generated using variations of the same method. Of those guidance products that use HREF data for storm attribute fields (UH for tornado guidance), the current operational method that uses STP as a calibration agent provided by the SREF (HREF/SREF_Ops) was evaluated as superior with a mean score of 5.42 as compared to scores below 4.2 for the HREF/SREF_Para and HREF/HREF_Cal methods. From their comments, evaluators often preferred the HREF/SREF_Ops method because it did not over forecast as did the other two methods.

The flow dependent method that explicitly trained a ML model for unique regimes of the large-scale flow was evaluated with a mean score of 5.79, which is slightly higher than the method for which the model was trained without regard to the large-scale flow (mean of 5.42). The implicit flow method scored the lowest with a mean score of 5.29. From evaluator comments, the explicit method tended to overall reduce the area and magnitude of forecast tornado probabilities, which, for these HWT cases, was a positive effect, especially for the non-event cases.

A Day-1 product that was evaluated somewhat separately from the other products was generated with a ML neural network model that was trained based on deterministic HRRR forecasts. Version 2 of this ML model was trained with a few more predictors than version 1 and incorporates two node layers of much fewer nodes as compared to the one node layer design of version 1. Figure 4 shows that version 2 (‘HRRR NN V2 D1’) is evaluated with a mean score of 5.09, which places the evaluated performance of this product roughly in the middle of the suite of Day-1 products. Version 1 was not explicitly scored, but its performance was compared to version 2. The top violin plot of Fig. 4 has a mean score of 3.09 indicating that version 2 provided only slightly better forecasts than version 1.

Of the Day-2 guidance methods, three of the methods had similar ratings while the HREF/GEFS stood out with noticeably lower mean and median ratings. Despite using a coarse global ensemble as input, the GEFS tornado guidance received relatively high ratings from Day 3 to Day 1. In aggregate, the GEFS Day-3 product scored identically to

the GEFS Day-2 product. Figure 5 reveals a high consistency among GEFS tornado forecasts primarily between days 1 and 2 such that for the respective violin plot there is a cluster of responses around a '4' rating. The other two violin plots show for GEFS tornado products a decrease in forecast consistency with longer lead time. However, there is evidence of consistency even out to day 3 given that the mean score of the day 3 to day 1 forecasts is 3.53 (i.e., a score greater than '3' indicates a degree of forecast consistency).

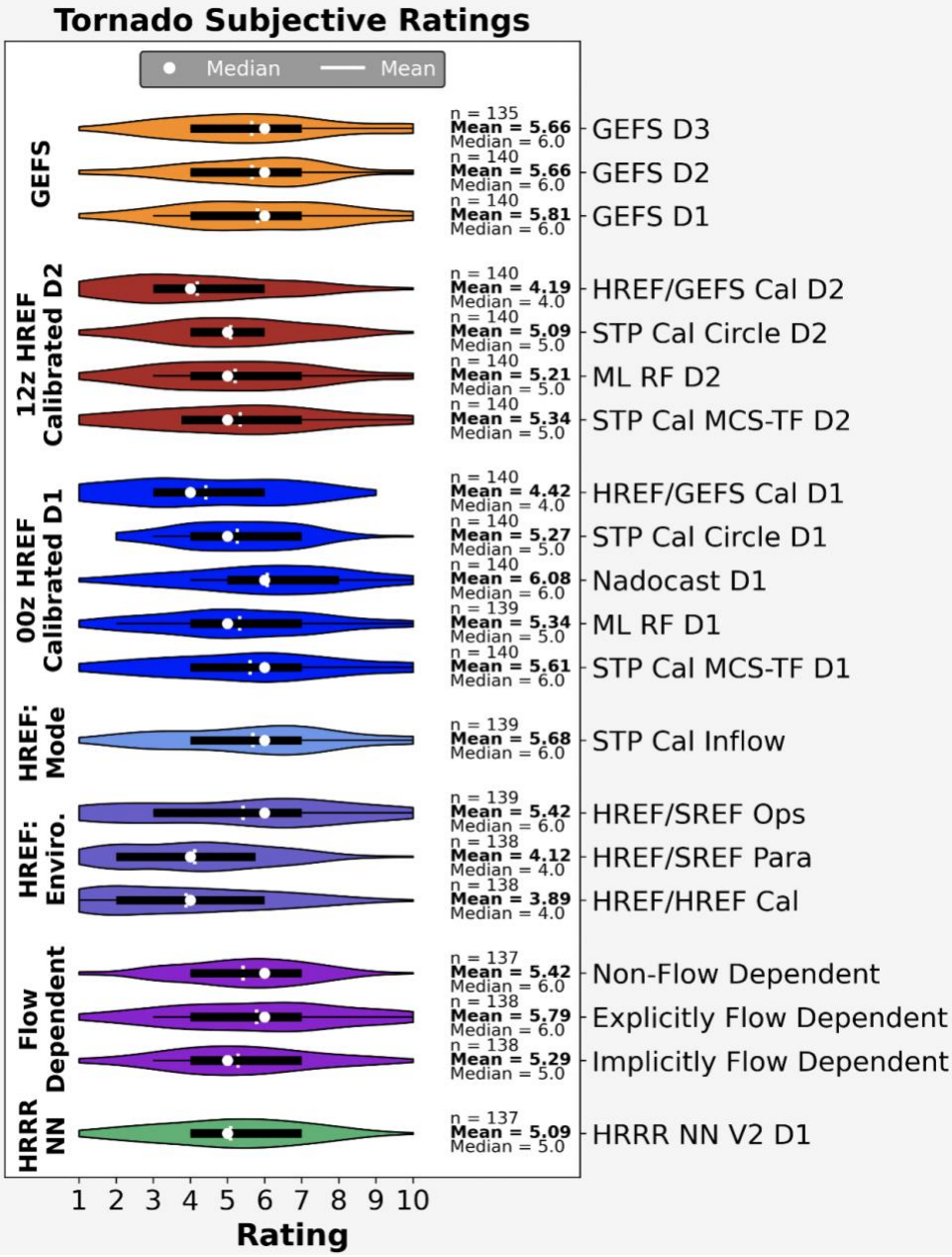


*Figure 3. Violin plots showing aggregate (across 19 SFE cases) evaluation ratings along with mean (dashed white line) and median (white dot) values for 20 tornado calibration methods as listed. A rating of '10' indicates a very good forecast.*

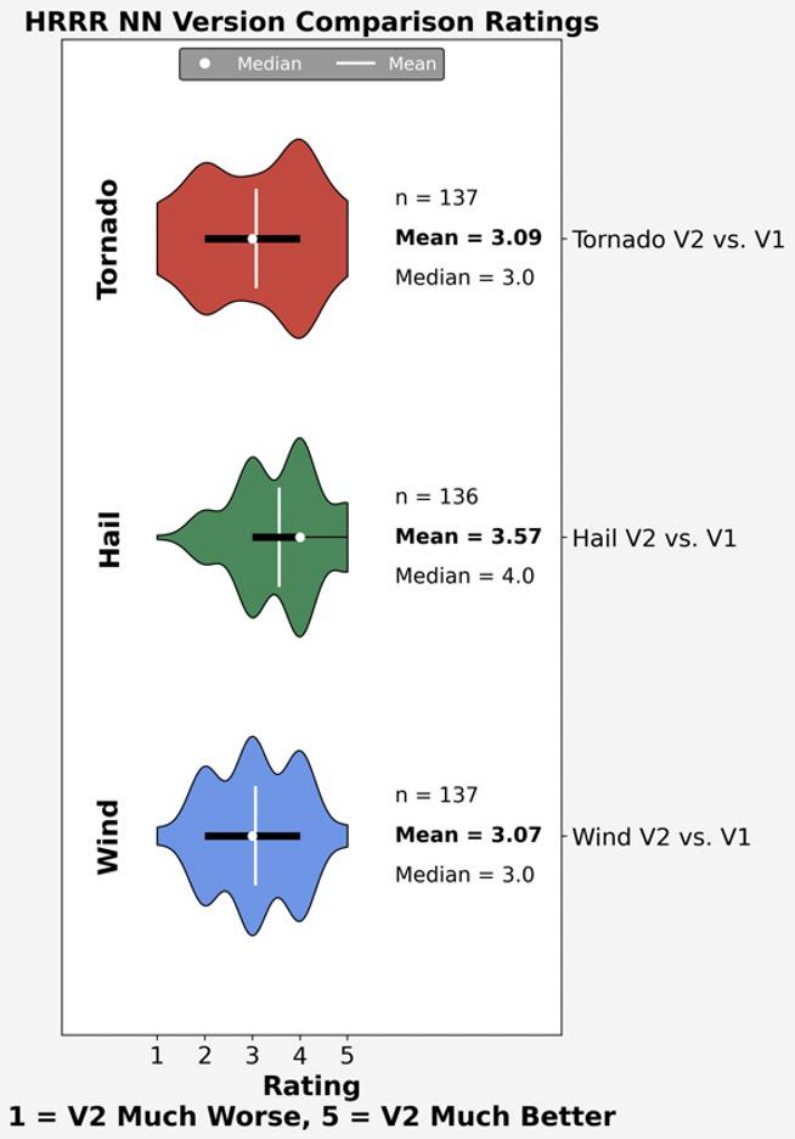


*Figure 4. Violin plots showing the distribution of rankings regarding the comparison between HRRR NN versions 1 and 2 (i.e., V1 and V2) for tornadoes (red), severe hail (green), and severe wind (blue). Higher ratings indicate version 2 is much better. The mean, median, and number of forecasts (n) are shown for each distribution.*

Of the Day-2 guidance methods, three of the methods had similar ratings while the HREF/GEFS stood out with noticeably lower mean and median ratings. Despite using a coarse global ensemble as input, the GEFS tornado guidance received relatively high ratings from Day 3 to Day 1. In aggregate, the GEFS Day-3 product scored identically to the GEFS Day-2 product. Figure 5 reveals a high consistency among GEFS tornado forecasts primarily between days 1 and 2 such that for the respective violin plot there is a cluster of responses around a '4' rating. The other two violin plots show for GEFS tornado products a decrease in forecast consistency with longer lead time. However, there is evidence of consistency even out to day 3 given that the mean score of the day 3 to day 1 forecasts is 3.53 (i.e., a score greater than '3' indicates a degree of forecast consistency).

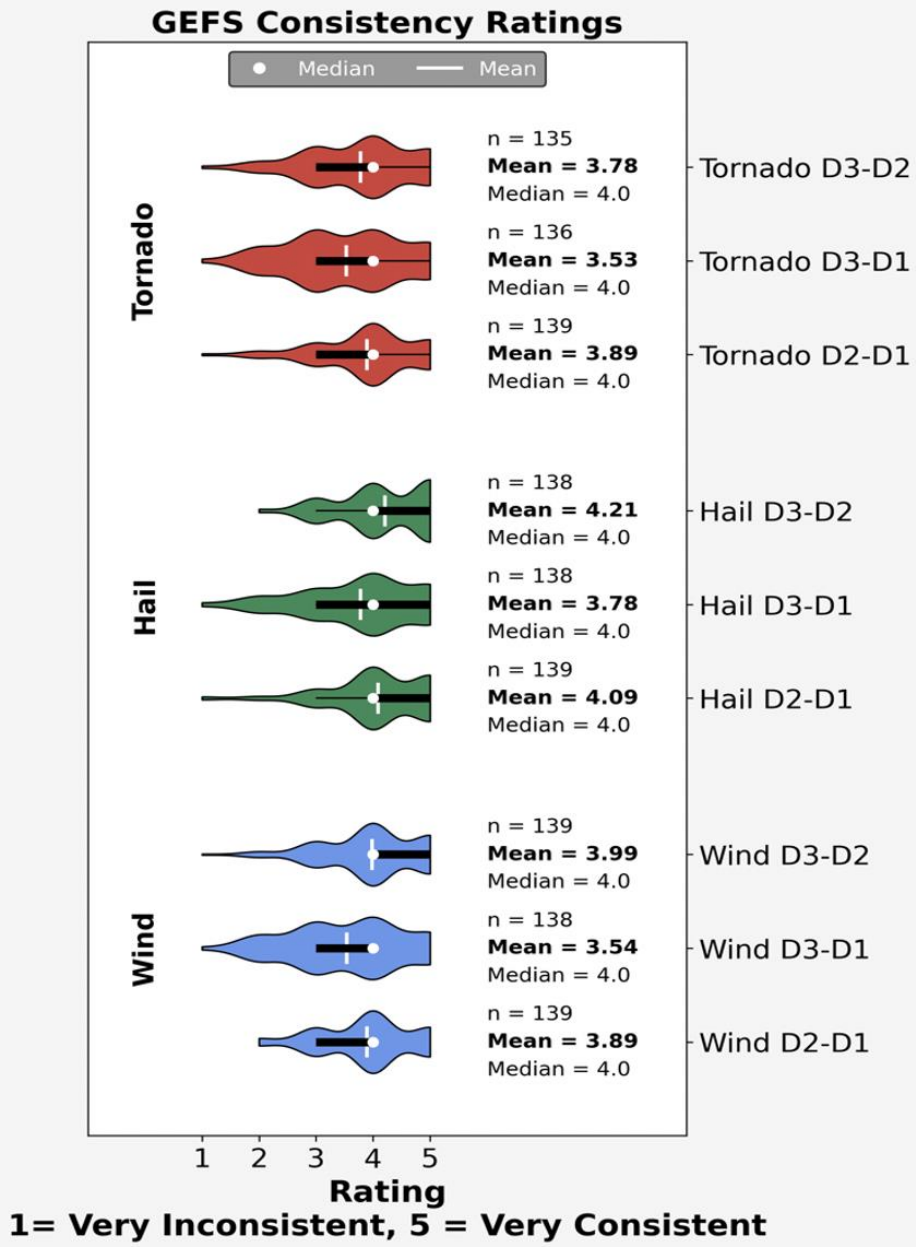


*Figure 5. Violin plots showing the distribution of rankings regarding the consistency of GEFS forecasts from day 3 to 2 (D3-D2), day 3 to 1 (D3-D1), and day 2 to 1 (D2-D1) for tornadoes (red), severe hail (green), and severe wind (blue). Higher ratings indicate more consistent forecasts, and the mean, median, and number of forecasts (n) are shown for each case.*

Analyzing results separately for either more active events (e.g., an SPC outlook of 5% tornado probability or higher) or less active events (e.g., an SPC outlook less than 5%), there is a notable performance difference among the calibrated methods. Figure 6 shows results for six of the Day-1 guidance methods. Nadocast and ML_RF score well for less active events while STP_MCS-TF and STP_inflow score nearly as well as Nadocast for more active days. ML_RF was evaluated lower for more active days likely because it tended to over forecast tornado probabilities.

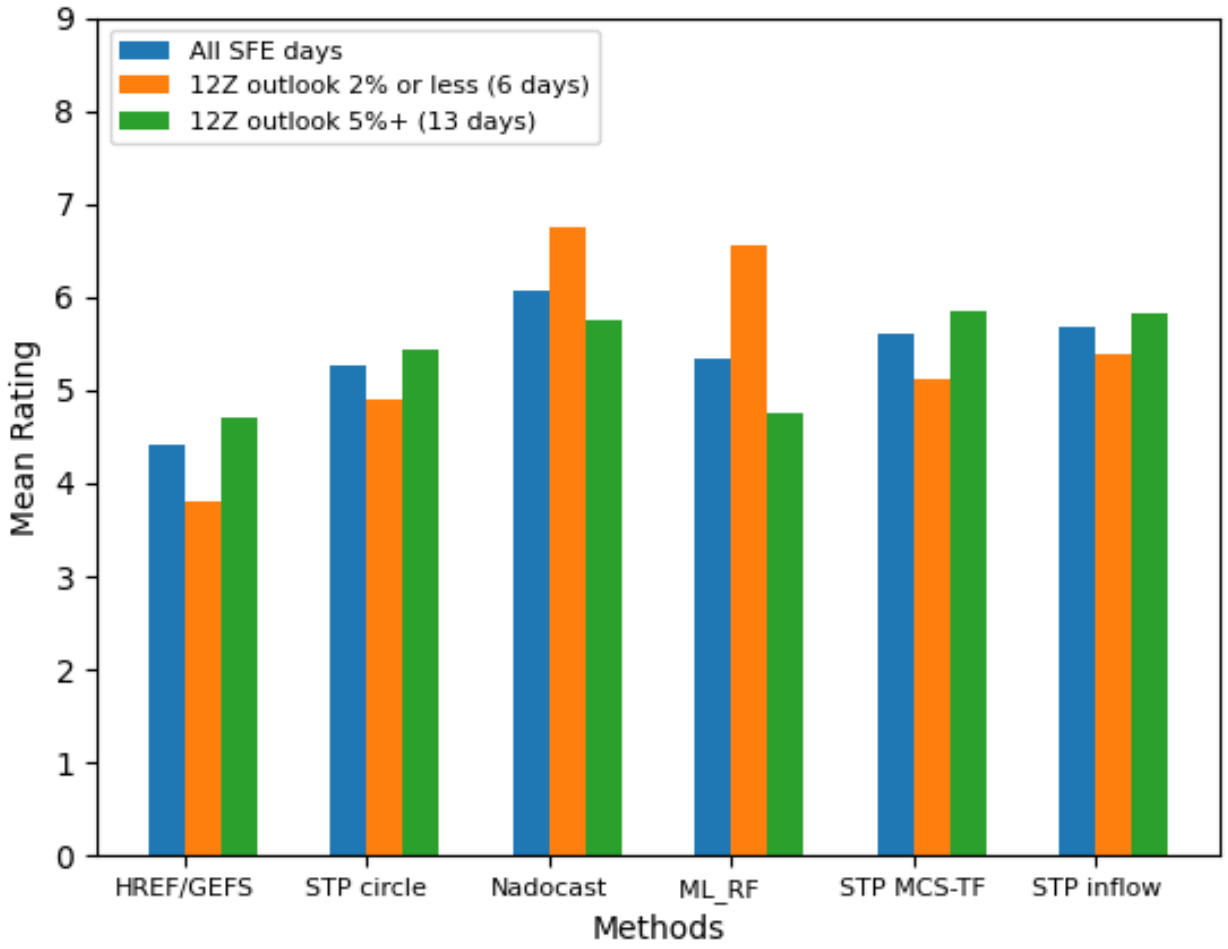


*Figure 6. Bar graph representing mean scores of six calibrated tornado guidance methods for all 19 SFE cases (blue), and days with SPC maximum outlook tornado probability 2% or less (orange) or 5% and greater (green).*

### 3.1.2 Analysis of Tornado Evaluation Philosophy

Evaluators were asked a series of questions as to what factors they considered most important in the process of scoring a tornado forecast. Overall, participants ranked high probability of detection (POD) within the highest probability contours (i.e., greater than 5%) as the most important factor in their ratings (mean ranking 2.24 out of 5; Fig. 7), followed by low false alarm ratio (FAR) for probabilities greater than 5% (mean ranking 2.75) and correct maximum tornado probabilities (mean ranking 2.83). Meanwhile, the two least important factors were low FAR (mean ranking 4.02) and high POD (mean ranking 3.17) for the lowest probabilities (5% and less). These results suggest participants emphasized the correctness of the highest tornado probabilities in their ratings. Participants were less concerned about large FAR in the lowest probabilities but, interestingly, were more split on the importance of POD in the low probability areas, indicated by a bimodal distribution of rankings (second violin from top in Fig. 7).

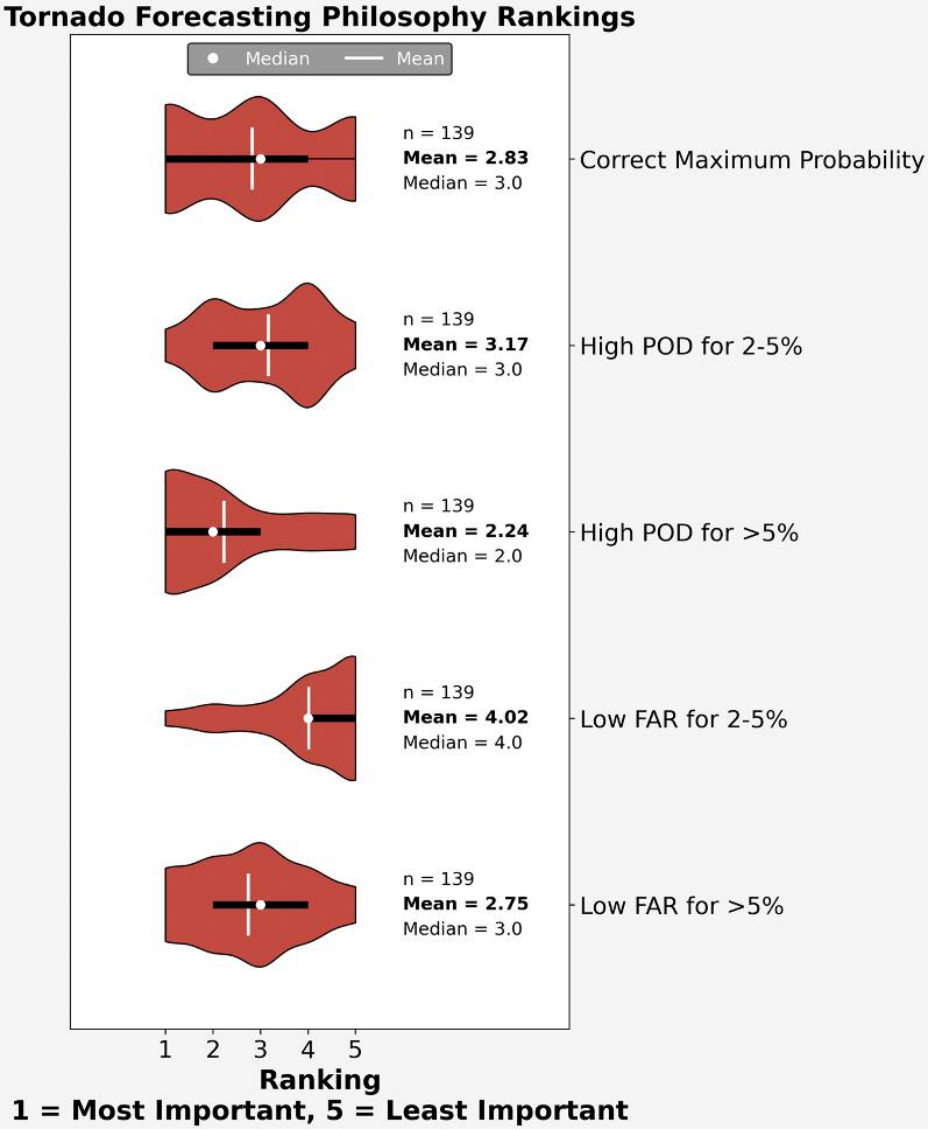


*Figure 7. Responses for a list of questions related to tornado forecast evaluation philosophy (see text).*

One factor that might explain this result is that different evaluators had different perspectives on how much a forecast should be rated on skill (i.e., based on official observations) versus usefulness (i.e., how a forecast represents the possibility of severe weather occurrence). An example of this dilemma is shown in Fig. 8, which depicts the HRRR_NN version 1 and 2 calibrated 24-h forecasts at initial time 1200 UTC on 5/18/22. No official tornado observations were recorded, but a series of NWS tornado warnings were issued for southeast Kentucky during the period. If evaluators scored strictly based on skill, they would favor version 1, which forecast no tornadoes; however, they would favor version 2 for usefulness as a forecast aid because it communicates some possibility of tornado occurrence that a forecaster should take into account when assessing the level of a severe weather threat. Consideration will be given in future SFEs to allow for scoring forecasts in consideration of both “skill” and “usefulness”.

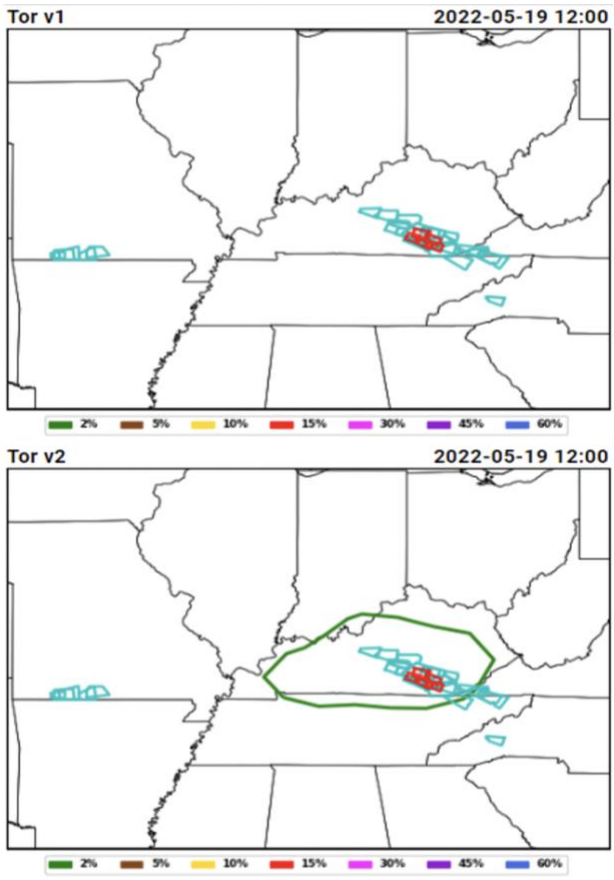


*Figure 8. Plots of 24-hr tornado probability for 5/18/22 using HRRR_NN calibrated method version 1 (top) and version 2 (bottom). No tornadoes observed. Red polygons indicate NWS tornado warnings.*

### 3.1.3 Aggregate Evaluation of Calibrated Hail Guidance

Fifteen calibrated hail guidance products were tested in the 2022 SFE. These products were created using different methods and underlying dynamical models, and they forecast for lead times ranging from 1 to 3 days (Table 3).

As with tornadoes, the GEFS products had the longest lead times, with 1-, 2-, and 3-day forecasts evaluated. Participants observed a general trend toward increasing skill as lead time decreased, both in their written comments and numerical ratings. The day 3, day 2, and day 1 GEFS hail forecasts (GEFS D3, GEFS D2, and GEFS D1) received mean ratings of 5.97, 6.25, and 6.42, respectively (orange violins in Fig. 9). One characteristic of all three forecasts was their tendency to cover broad areas with non-zero probabilities. Indeed, participants found the locations highlighted by the three forecasts to be very consistent (with median subjective consistency ratings of 4 out of 5 for days 3 to 2, 3 to 1, and 2 to 1; green violins in Fig. 5) but noted also as lead time decreased that the forecast probability magnitudes tended to be appropriately adjusted, either higher or lower as consistent with the increase or waning of the forecasted favorability of severe weather over time. The day 1 and 2 GEFS forecasts performed particularly well relative

to the other guidance products, especially considering the coarse global ensemble input (Fig. 9). With that said, some participants felt that the GEFS products covered too large of areas, particularly at shorter lead times.

Two other hail forecasting methods were evaluated out to 2-day lead times: the HREF/GEFS calibrated product and the ML RF. Both of these considered storm attribute information from the 1200 UTC initialization of the HREF. In their written comments, participants noted that the probabilities from the day 2 ML RF (ML RF D2) tended to be higher and cover a larger area than those from the day 2 HREF/GEFS product (HREF/GEFS Cal D2), giving the ML RF D2 greater probability of detection (POD) but also greater false alarm. While participants found both methods useful, in general there was a slight preference for ML RF D2, with participants preferring its POD and areas of maximum probabilities. This finding was reflected in the subjective ratings; both methods received a median rating of 6, but the ML RF D2 had a greater mean rating (6.06) than HREF/GEFS Cal D2 (5.42; Fig. 9).

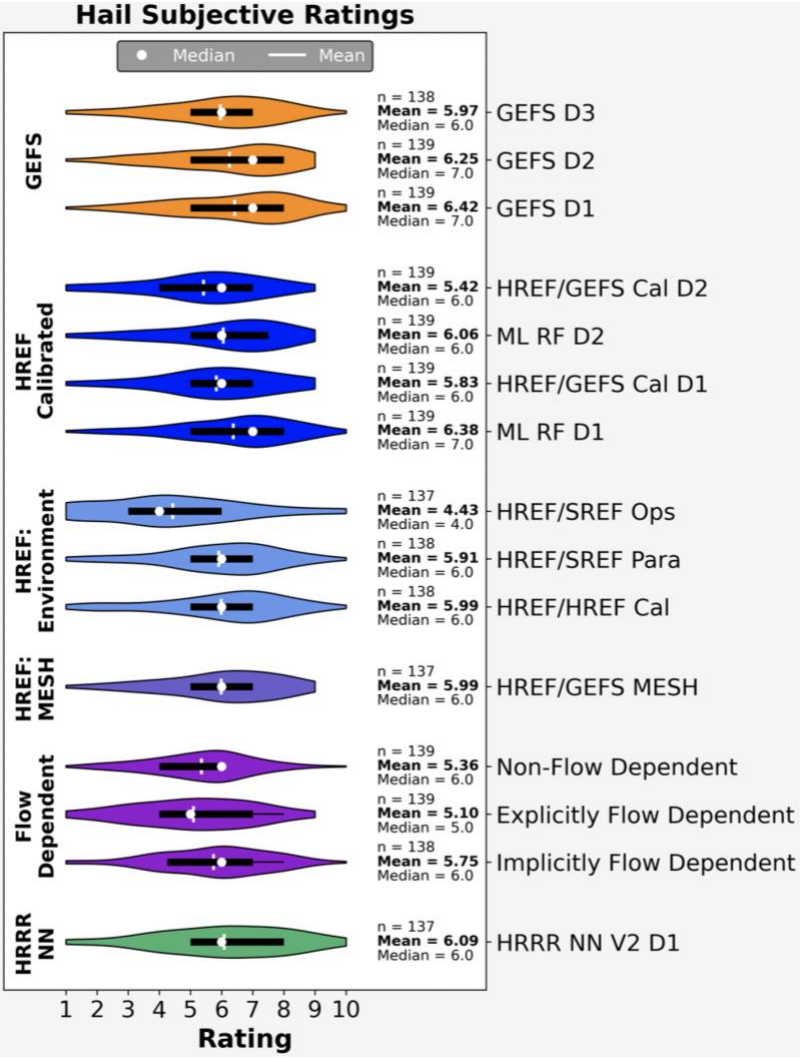


*Figure 9. As in Figure 3 but for the 15 calibrated hail guidance methods evaluated.*

As with the GEFS forecasts, participants found the HREF/GEFS D2 and ML RF D2 products to be consistent with their day-1-lead-time analogues (i.e., HREF/GEFS D1 and ML RF D1, respectively), with the day 1 forecasts having similar characteristics but slightly greater skill. The mean ratings for HREF/GEFS Cal D1 (5.83; third dark blue violin in Fig. 9) and ML RF D1 (6.38; fourth dark blue violin in Fig. 9) were higher than their corresponding day 2 ratings.

Other day-1 lead-time hail products based on HREF storm attributes were evaluated in different survey subsections. The current non-machine-learning operational standard, HREF/SREF Ops, received the lowest mean (4.43) and median (4) ratings of all hail methods evaluated (Fig. 9), with participants noting its tendency to under-forecast and have a low POD. However, recalibrating the product's exceedance thresholds produced a substantially better forecast (HREF/SREF Para, mean rating 5.91). Interestingly, using the GEFS (HREF/GEFS Cal D1; mean rating 5.83), SREF (HREF/SREF Para; mean rating 5.91), and HREF (HREF/HREF Cal; mean rating 5.99) ensembles for the method's environmental information did not produce substantially different ratings. Participants felt those three forecasts (i.e., HREF/GEFS Cal D1, HREF/SREF Para, and HREF/HREF Cal) tended to be quite similar with only minor differences, indicating that the HREF storm-attribute inputs tend to dominate the guidance products.

Calibrating the HREF/GEFS product based on MRMS MESH instead of observed hail reports resulted in forecasts with larger probability magnitudes and areal coverage, which gave better POD but worse false alarm. In general participants found value from HREF/GEFS MESH (mean rating 5.99; indigo violin in Fig. 9) and especially liked the greater areal coverage of HREF/GEFS MESH's probabilities, although they generally felt the probability magnitudes were too high. Incidentally, participants felt the same way about the practically perfect MESH product (not formally rated). Overall, calibrating the HREF/GEFS product based on MESH was found to be useful and a promising idea for future development, especially if over-forecasting bias could be reduced.

To test if the characteristics of the large-scale flow could be used to improve forecast guidance, three flow-dependent methods were evaluated. In both their ratings and comments, participants expressed a slight preference for the implicitly flow dependent guidance (median 6, mean 5.75), as it generally highlighted the best areas and probability magnitudes of the three methods. However, participants wrote that the non-flow dependent forecasts (median 6, mean 5.36) tended to be qualitatively very similar to the implicitly flow dependent forecasts on many days. Meanwhile, the explicitly flow dependent forecasts generally were generally inferior (median 5, mean 5.10). Participants felt all three products tended to under-forecast probability magnitudes while producing non-zero probability areas that were relatively large, such that the forecasts often appeared overly smoothed, despite no spatial smoothing being applied.

Finally, a new neural network-based product using the HRRR was evaluated (HRRR NN). Compared to an earlier version (i.e., version 1), participants found the new version (i.e., HRRR NN or version 2) tended to produce higher probabilities, giving it better POD than version 1. When asked to rate how version 2 compared to version 1, participants gave a median rating of 4.0 out of 5.0 (green violin in Fig. 4), showing a

preference for version 2. Indeed, version 2 received a mean rating of 6.09 (green violin in Fig. 9), one of only five hail methods to receive a mean rating above 6.0. Compared to other top-performing hail methods, HRRR NN received more ratings of 9s and 10s, but also more 1s and 2s (Fig. 9). It is possible that the greater variance in its ratings is at least partially due to its reliance on a single model (the HRRR) as opposed to an ensemble (e.g., the GEFS or HREF) for its forecast inputs.

Overall, participants generally found all of the hail products useful, if imperfect, and noted that differences between methods could be valuable from a forecasting perspective. Finally, participants felt that most products could benefit from additional calibration to improve their usefulness.

### 3.1.4 Aggregate Evaluation of Calibrated Wind Guidance

Fourteen calibrated guidance products were evaluated for severe wind. These included all methods evaluated for severe hail except the HREF/GEFS MESH.

As with severe hail, the GEFS wind forecasts were found to be relatively consistent from day 3 to day 1 lead times, with median consistency ratings of 4 out of 5 for days 3 to 2, 3 to 1, and 2 to 1 (blue violins in Fig. 5). Participants noted the three GEFS forecasts generally highlighted the same, relatively broad, areas with their nonzero probabilities, but trended toward higher and more accurate probability magnitudes at shorter lead times. As a result, the day 1 GEFS wind forecasts were rated the highest, on average (mean rating 5.78; Fig. 10), followed by the day 2 (mean rating 5.58) and day 3 (mean rating 5.17) forecasts.

The HREF/GEFS Cal and ML RF methods also provided good consistency between their day 2 and day 1 wind forecasts, with both methods' forecasts having similar characteristics and covering similar areas on days 2 and 1. However, as expected, both methods tended to be sharper on day 1. These sharper probabilities led to a slight increase in mean rating from day 2 to day 1 for HREF/GEFS Cal (mean rating 5.65 for day 2, 5.82 for day 1; Fig. 10) but not for ML RF (mean rating of 6.12 for both days), although ML RF D1 had a higher median rating (7) than ML RF D2 (6). Participants explained that, while they liked the areas outlined by the ML RF method, ML RF tended to over-forecast probability magnitudes, so the increased sharpness from day 2 to day 1 did not always lead to subjectively better forecasts. Indeed, ML RF D1 received more ratings of 8s, 9s, and 10s than ML RF D2, but also more 1s and 2s (Fig. 10). Despite its over-forecasting bias, participants expressed a slight preference for ML RF compared to HREF/GEFS Cal at both lead times due to its better areal coverage, and the ML RF D2 and ML RF D1 products received the highest mean ratings of all wind products evaluated.

Conversely, the current operational standard product, HREF/SREF Ops, received the lowest mean rating (4.35; Fig. 10) of the calibrated wind products evaluated because of its strong under-forecasting bias. However, applying a new calibration and updated algorithm to the product (HREF/SREF Para) resulted in a substantially improved forecast (mean rating 5.86) that was among the top-performing methods for severe wind prediction. Participants noted only minor differences in the forecast when the method's

environmental information came from GEFS, SREF, or HREF, as the mean ratings for HREF/GEFS Cal D1 (5.82), HREF/SREF Para (5.86), and HREF/HREF Cal (5.74) were all similar.

When evaluating the flow-dependent severe wind products, participants generally found only minor differences between the non-, explicitly-, and implicitly-flow dependent forecasts (mean ratings of 5.22, 5.17, 5.16, respectively; Fig. 10). Like the flow-dependent hail products, all flow-dependent wind products tended to have broad areas of relatively low probabilities, a characteristic that led to mixed reviews. In some cases—particularly for more widespread wind events—participants liked the broad-coverage, low-magnitude probabilities, finding them easier to interpret than other wind guidance products. However, in cases with more localized threat areas, participants found the probabilities too diffuse to give meaningful guidance. Despite the products' similar tendencies, at least one self-identified forecaster enjoyed seeing the output from the multiple methods. The forecaster felt the differences between products helped show the importance of the large-scale flow characteristics (e.g., magnitude, orientation) to the day's severe weather threat.

The new version of HRRR NN (version 2) was among the top-performing methods for severe wind, receiving a mean rating of 6.07 (green violin, Fig. 10). Compared to version 1, participants observed that version 2 tended to have higher probabilities that encompassed slightly larger areas. Although many participants preferred the larger highlighted areas, many felt that version 2's probability magnitudes were too high. Consequently, when asked to rate version 2 compared to version 1, participants gave a mean rating of 3.07 (median of 3) out of 5 with a tri-modal distribution (blue violin in Fig. 4), indicating no clear overall preference for one version or another.

Overall, participants felt that most of the wind guidance products had utility but could benefit from additional calibration, with most products having too high of magnitudes. Nevertheless, the highlighted threat areas were commonly cited as being very useful, and the consistency of the products out to multiple day lead times was seen as encouraging.

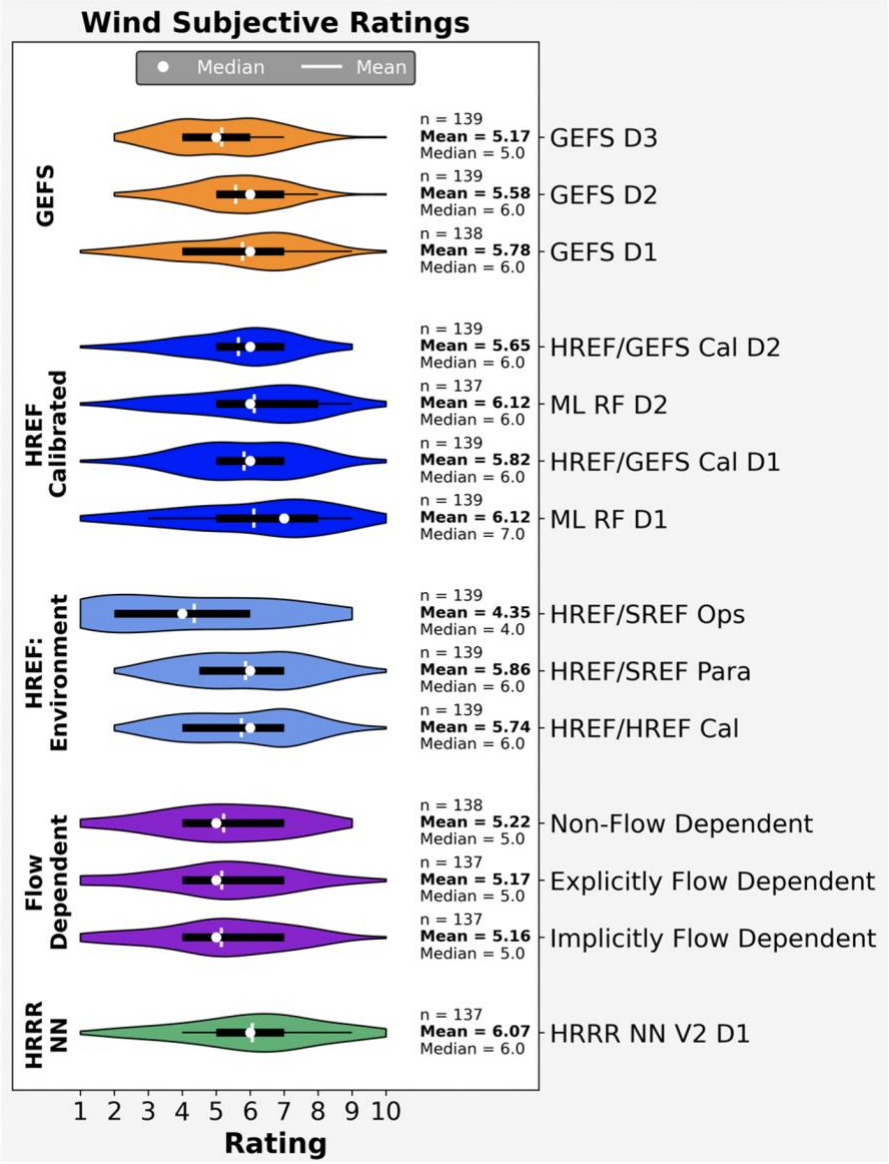


*Figure 10. As in Figure 3 but for the 14 calibrated wind guidance methods evaluated.*

## 3.2 Model Evaluations – Group B: Deterministic CAMs

### 3.2.1 Deterministic Flagships

The first evaluation in the B group focused on the “Deterministic Flagships” comparison, which consists of cutting-edge model guidance contributed by various agencies. The guidance provided herein entails models with different dynamical cores, data assimilation strategies, physics parameterizations, and whether they are nested or global configurations. As such, the main goal of this comparison is to see which configuration is performing best rather than to do a specific comparison between models

with slightly different configuration strategies. Models contributed here frequently have been iterated on by their agencies, and are relatively advanced in their development.

For the first time, these models were evaluated blindly; participants were not able to see which model was producing which images. In addition to being blinded, the location of each particular model was randomized each day, so participants could not count on a model being in the same panel day-to-day. Models were unblinded during the discussion of the results, after the surveys were submitted. Participants compared the reflectivity and 2–5 km updraft helicity (UH) from each configuration, along with one of the following environmental variables: 2-m temperature, 2-m dewpoint, and surface-based convective available potential energy (SBCAPE). The evaluation strategy also differed from prior years in that rankings were used rather than 1-10 ratings, forcing participants to distinguish between model performance. Models were ranked from best (i.e., ranking of 1) to worst (i.e., ranking of 5). One limitation to the ranking methodology, which is more robust than unlabeled 1–10 scales, is that all models must be present for the rankings to be compared. As such, we had 12 cases out of 19 that met this criterion, with 92 responses from participants across these cases. Participants were asked to consider forecast hours 13–36 in their evaluations, corresponding with 1200–1200 UTC.

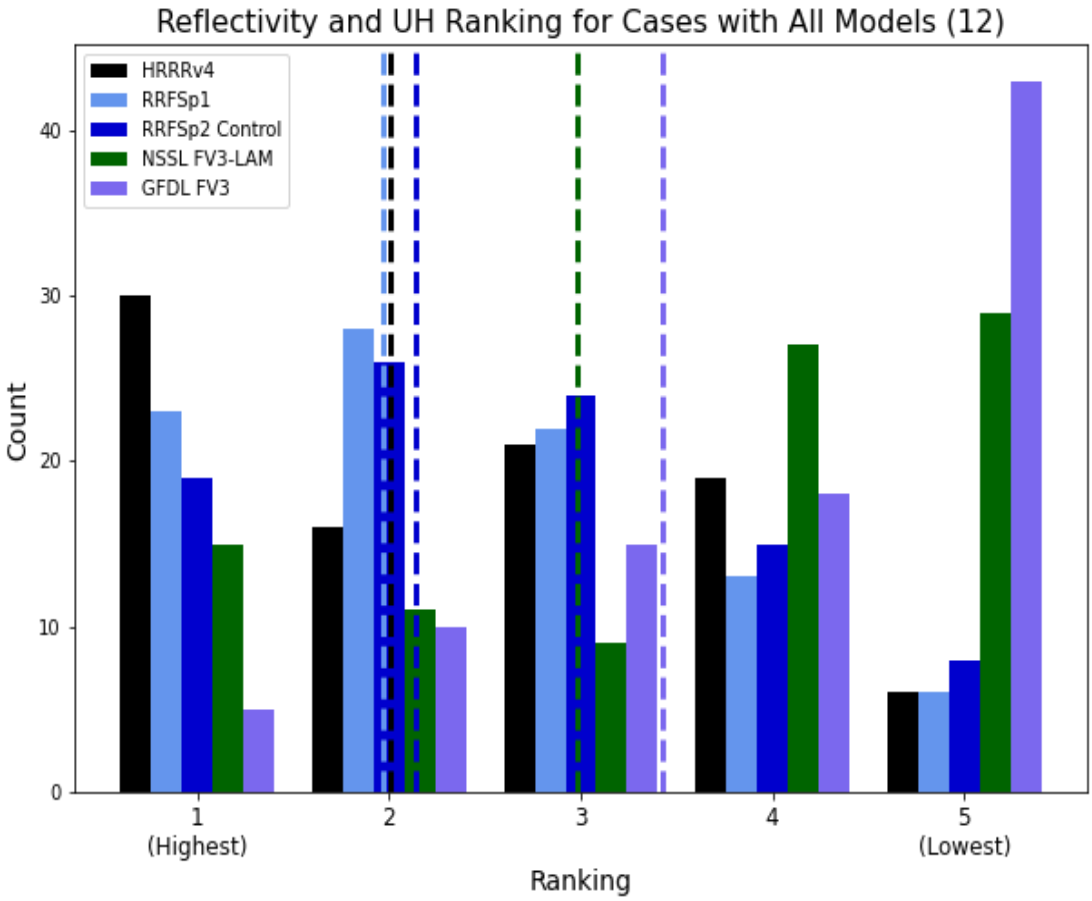


*Figure 11. Reflectivity and UH rankings for models in the Deterministic Flagship comparison. Dashed lines indicate the mean ranking (lower numbers are better).*

Rankings for the reflectivity and UH show two groupings of model performance (Fig. 11). The HRRRv4, RRFSp1, and RRFSp2 Control were ranked relatively similarly with regards to the mean ranking, followed by the NSSL FV3-LAM, and then the GFDL FV3. The HRRRv4 was most frequently ranked first, followed by the RRFSp1 and the RRFSp2 Control. The RRFS models were most frequently rated second or third, leading the RRFSp1 to a slightly higher overall mean ranking than the HRRRv4, though these differences are likely not significant. The NSSL FV3-LAM was most frequently rated fourth or fifth, and the GFDL FV3 was most frequently rated last. When asked what characteristics of the simulated reflectivity and UH forecasts were most important to the participants when ranking the models, participants broadly cited forecasting challenges such as the convective initiation, progression of storms, location of storms, intensity of storms. In a word cloud of participant responses, timing, location, and storm mode showed up frequently. Convective coverage also came up in some participant responses.

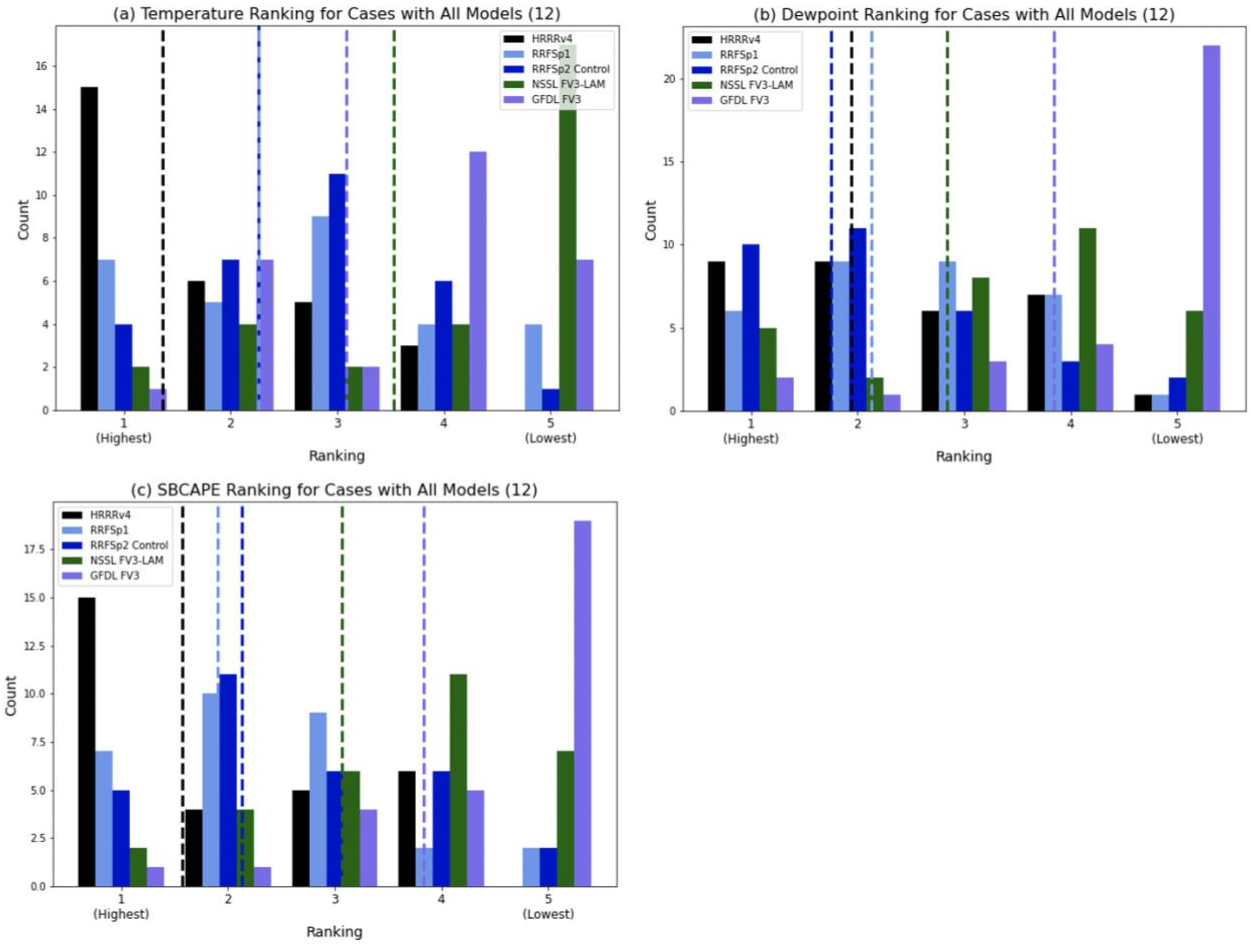


*Figure 12. Rankings of environment for the Deterministic Flagship models. Rankings were completed for (a) 2-m Temperature, (b), 2-m Dewpoint, and (c) SBCAPE. Dashed lines indicate the mean ranking for the model in question (lower numbers are better), and the dashed blue lines in (a) indicate that the RRFSp1 and the RRFSp2 Control had the same mean ranking. Note that the y-axes on these comparisons are scaled to each individual subplot.*

Rankings for the environmental fields followed similar patterns to the reflectivity and UH rankings, although the HRRRv4 easily received the highest mean ranking in temperature and SBCAPE (Fig. 12a,c). The HRRRv4 was most frequently rated the highest of all of the models considered in those fields, while the RRFSp2 Control was most frequently ranked first for dewpoint (Fig. 12b). Overall, the pattern of the HRRRv4, RRFSp1, and RRFSp2 Control ranking the best continued for all environmental fields considered, followed by the GFDL FV3 and the NSSL FV3-LAM. The GFDL FV3 placed fourth in terms of highest ranking for temperature, but was most frequently rated last for dewpoint and SBCAPE. For temperature, the NSSL FV3-LAM was most frequently ranked last. When evaluating the 2-m temperature, participants looked more closely at boundaries, gradients, and mesoscale areas of bias in making their rankings. Cold pools were also considered. Similar considerations applied for the 2-m dewpoint and the SBCAPE, although the shape and orientation of boundaries were specifically cited with regards to any drylines that may have been in the SFE domain of interest. Horizontal distribution of large areas of SBCAPE (e.g., warm sectors) also played a role for some participants assigned the SBCAPE field.

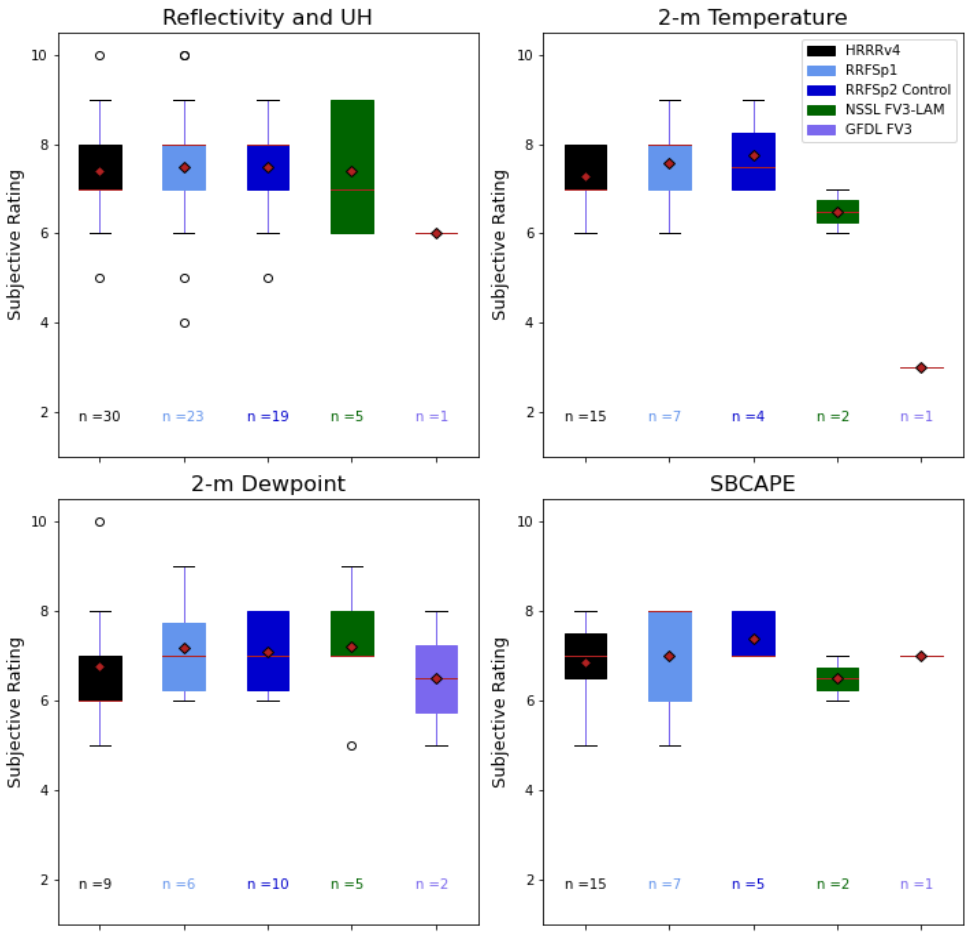


*Figure 13. Subjective rating scores for the highest-ranking model in each comparison. Medians are shown by the brown line; brown diamonds indicate the mean. The sample size is listed at the bottom of each subplot.*

Finally, participants were asked to rate their highest-ranked model on a scale of 1–10 for both the UH and reflectivity fields and their environmental fields of choice, to indicate how the models were performing on a given day (Fig. 13). Note that their best-performing model for reflectivity/UH and the environmental field could differ. Results show that the best performing model was frequently ranked similarly, with median ratings around 7 or 8 out of 10 in most cases. Since the environmental fields have quite small sample sizes, strong conclusions cannot be drawn from them. However, looking at the distributions of the HRRRv4, RRFSp1, and RRFSp2 Control members in the reflectivity and UH ratings show very similar distributions, indicating very similar performance on days where this set of models are performing their best.

### 3.2.2 RRFS vs. HRRR

Our next comparison in the B group looked at an in-depth comparison between the HRRRv4 (currently operational), and the RRFSp2 Control (a candidate for future HRRR replacement under Unified Forecasting System efforts). This comparison was not blinded, so participants knew which model guidance that they were considering. During this survey, participants were asked to select which model performed best for a variety of fields, or if they performed about the same. Participants were asked to select two storm attribute fields out of a possible five to evaluate, were next randomly assigned one of (a) 2-m temperature, 2-m dewpoint, or SBCAPE, and finally were randomly assigned two out of a possible five additional environmental fields to comment upon differences for. The final five environmental fields did not have corresponding observations, so participants were asked solely to remark upon the differences between the fields.

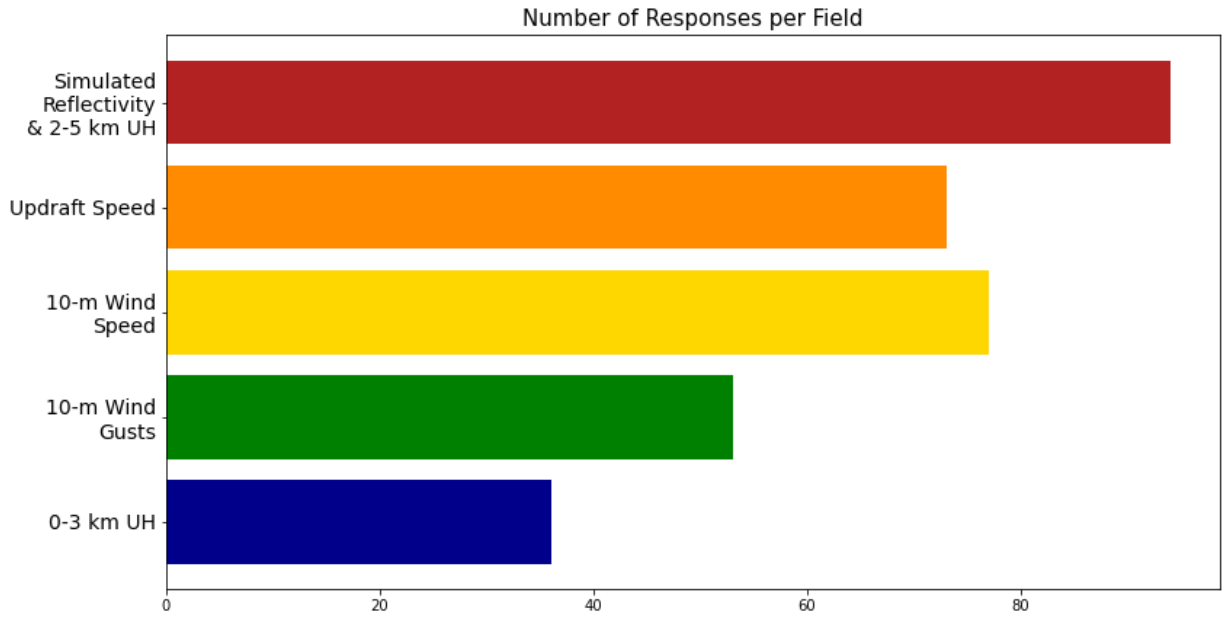


*Figure 14. Number of times each storm attribute field was selected for evaluation. Note: 0-3 km UH was unavailable for the first few weeks of SFE 2022.*

Participants most frequently selected the reflectivity/2–5 km UH and 10-m wind speed to evaluate, although the updraft speed was a close third (Fig. 14). The 10-m wind gusts, although only evaluated fourth most often, were frequently a topic of discussion after completion of the survey. Storm attribute field performance varied regarding which model was selected as the best performer (Fig. 15). For simulated reflectivity and updraft speed, the HRRRv4 was selected as the better-performing model more frequently than the RRFSp2 Control. However, the 10-m wind speeds and the 0–3 km UH were frequently better in the RRFSp2 Control relative to the HRRRv4. The 10-m wind gust performance was similar across all categories. When commenting on the 2–5 km UH and simulated reflectivity, participants frequently noted different performance at different time periods, as exemplified by comments such as: "*HRRR did better first half of the period by far, but RRFS did better with the bigger event, derecho later in period*". Comments such as these highlight the necessity of objective verification across the entire convective day, which is time-prohibitive to do subjectively in the context of the SFE. Participants frequently commented that the 10-m wind speed was too low, particularly in the HRRRv4. The 10-m wind gust product, which is not constrained by having to meet a reflectivity criterion, was noted by the participants to show swaths of strong wind gusts in the RRFSp2 Control that appeared to be synoptically driven rather than associated with convection. However, during one discussion session, a WFO forecaster mentioned that it wouldn't necessarily be bad for the model to show high synoptic gusts, as they currently faced a forecast challenge in getting good guidance for gusty winds that were not associated with convection.

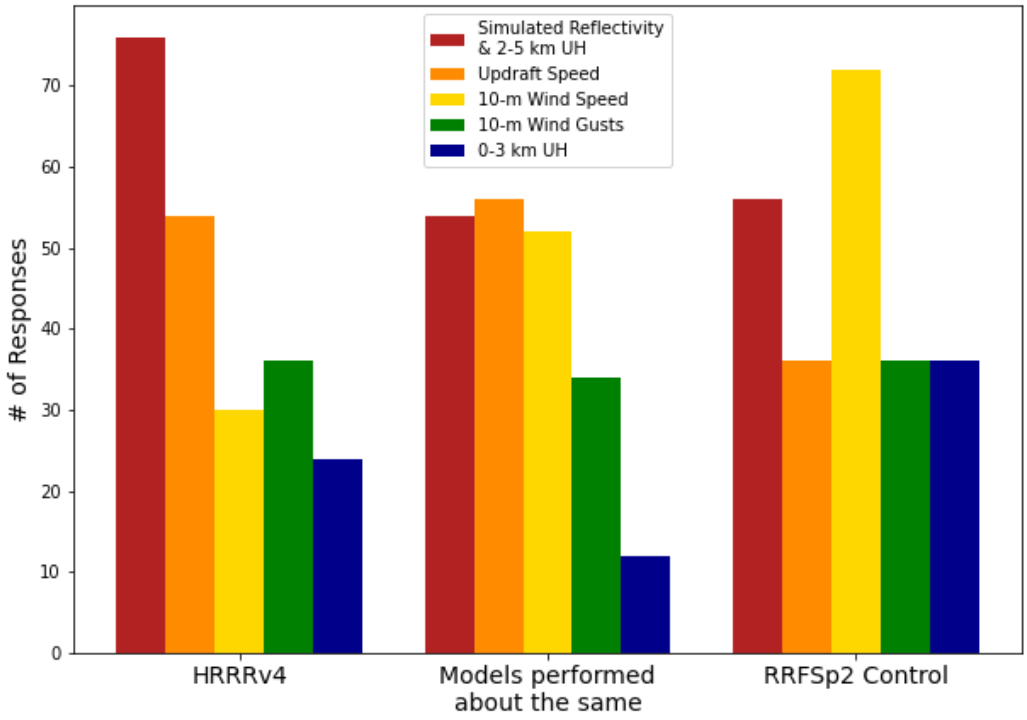


*Figure 15. Answers to the question, "Which model performed best for this field?", in which participants were asked to select at least two of the five fields presented to evaluate.*

Participants were also asked, “How important does this field seem to be to the evolution of severe weather on <date>?” This question checked the assumptions that we utilized regarding the importance of different storm attribute variables on forecasting severe convective storms. Responses here were normalized based on the number of responses received for each variable (Fig. 16). Simulated reflectivity and 2–5 km UH were most frequently rated as either “Very important” or “Extremely important”, and were by far the most frequently selected as “Extremely important”. Updraft speed and 10-m wind speed were most frequently rated moderately important. The 10-m wind gusts were most frequently rated “Very important”, and 0–3 km UH was most frequently rated “Slightly important”. These findings may be linked to the type of severe convective weather that took place during SFE 2022. Severe wind was one of the most frequent hazards, while tornadoes were relatively infrequent.

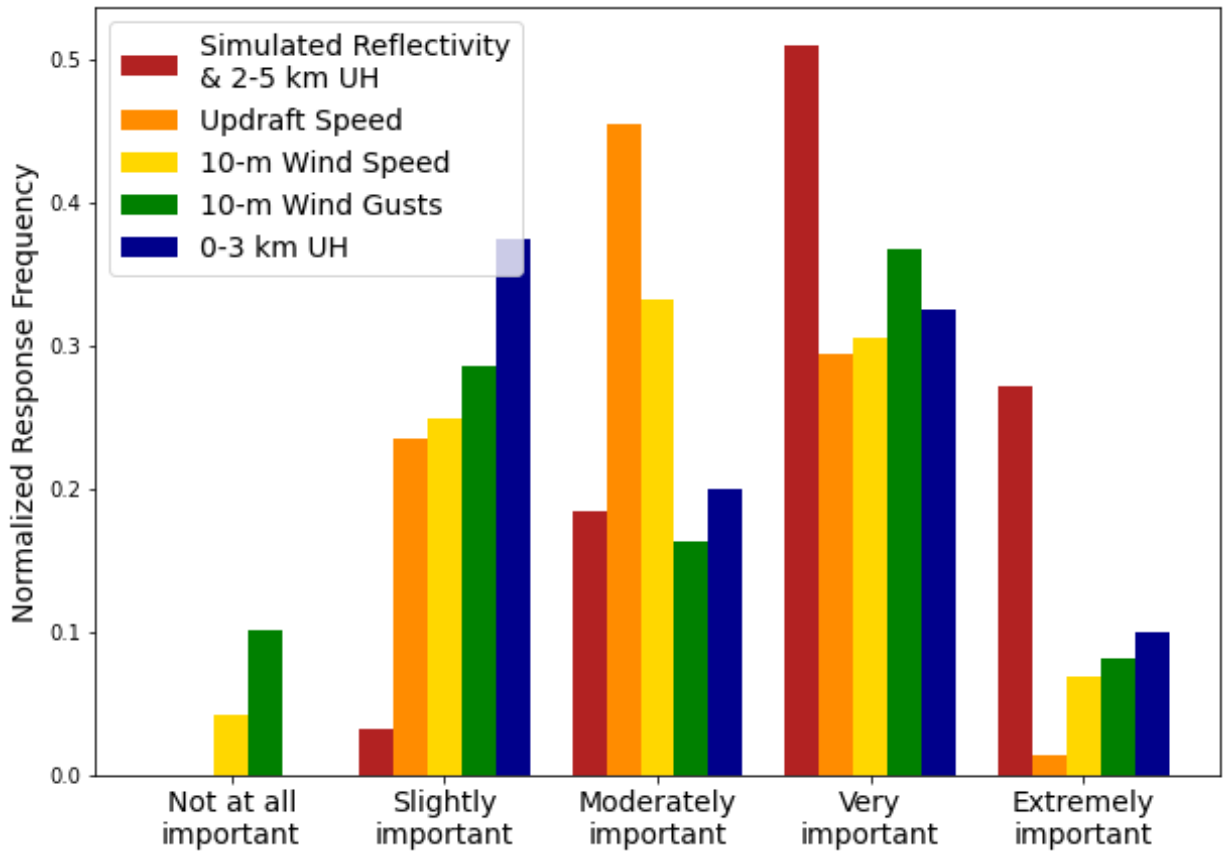


*Figure 16. Participant indications of how important particular storm-attribute fields were to the forecast of severe convection on the day they were evaluating. Responses are normalized by the total amount of responses for each given variable as shown in Fig. 14.*

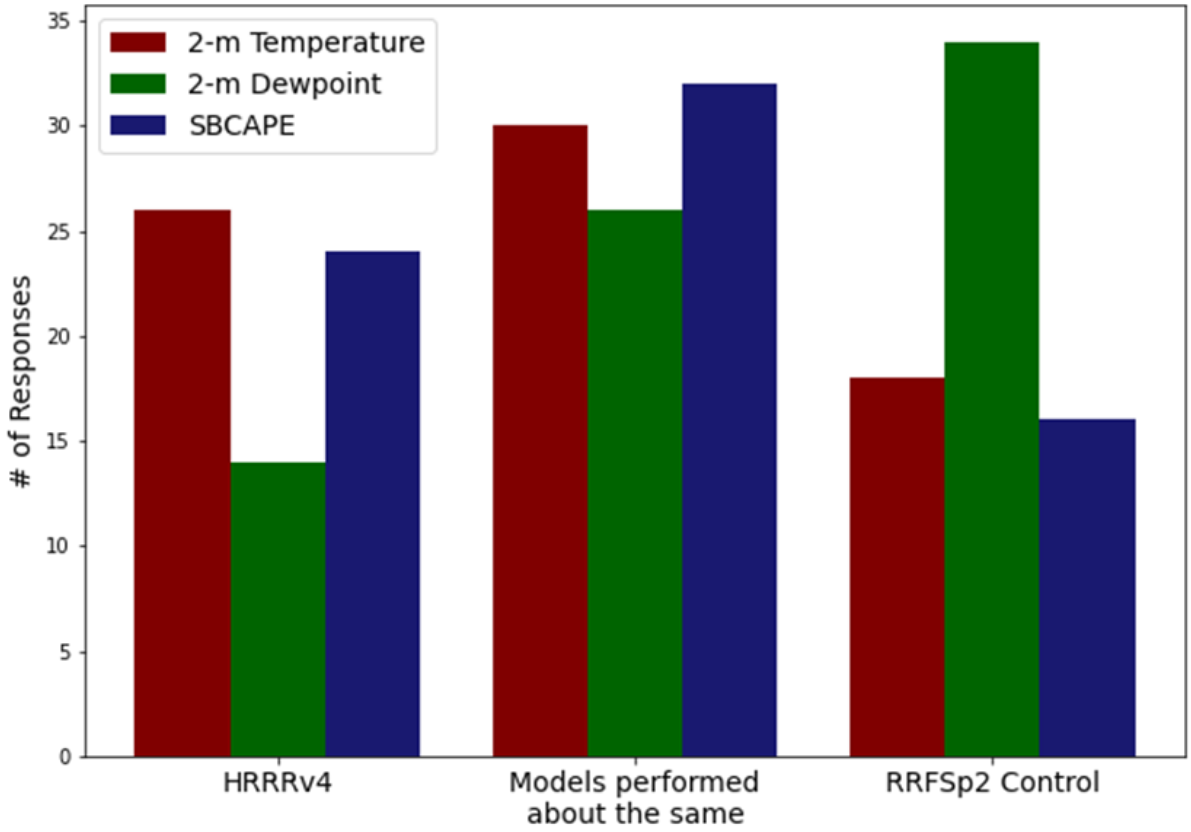


*Figure 17. As in Fig. 15, but with environmental fields that were randomly assigned.*

Participants next evaluated a randomly selected environmental field. Fields were evenly assigned between 2-m temperature, 2-m dewpoint, and SBCAPE. For temperature and CAPE, the most frequent response from participants was that the HRRRv4 and the RRFSp2 Control performed about the same (Fig. 17). For dewpoint, however, the RRFSp2 Control being better was the most frequent response. Overall, the HRRRv4 appears to perform better with regards to the 2-m temperature and the SBCAPE, but the RRFSp2 Control handles the 2-m dewpoints better than the HRRRv4. Participant comments surrounding the temperature spoke to the placement and intensity of boundaries and cold pools, and some participants focused in on the reasoning why specific biases may be preferred: "*The RRFSp2 appeared slightly closer to true values but given it was cool biased compared to HRRRv4 warm bias, I preferred the warmer solution given the impacts of the day may be made more significant with a warmer boundary layer*." Participants frequently commented on a dry bias in the HRRRv4's 2-m dewpoints, and the comments surrounding the CAPE showed no clear trends. Environmental fields were generally rated as less important to the day's forecast of severe convection (Fig. 18) relative to the storm attribute fields, but overall, the SBCAPE was indicated as the most important, followed by the dewpoint, and then the temperature.

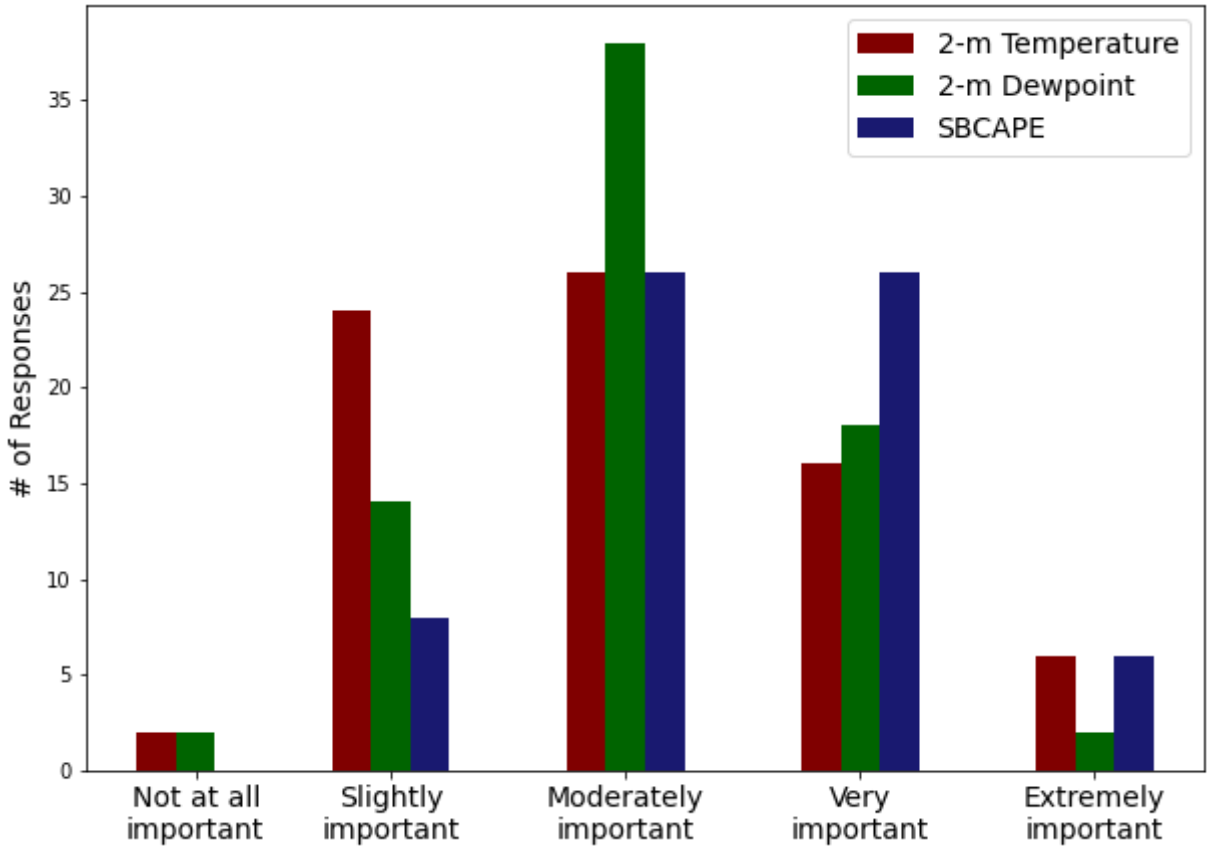


*Figure 18. As in Fig. 16, but for environmental attributes. Responses are not normalized due to the evenly distributed random assignment of environmental variables to participants.*

Finally, participants were asked to evaluate fields that were new to formal subjective evaluation, and were asked to comment on differences between three of the following fields that were randomly assigned: 500 mb height/wind, 700 mb height/wind, 850 mb height/wind, MUCAPE, and MLCAPE. Comments on the 500 mb fields were frequently that the models were similar, but for some cases participants were able to highlight details of the evolution of the upper-air fields. One such example reads, "*HRRRv4 had a weaker trough with the main core of winds mainly centered over AMA. RRFSp2 has a stronger trough with a wind core extending south of Lubbock. This also might explain the resulting differences in convective products*." Case-based analysis of these upper-air CAM fields can help developers identify systematic differences that may be linked to sensible weather. At 700 mb, participants frequently noted that the RRFSp2 Control had stronger winds than the HRRRv4. This comment was less frequent at 850 mb relative to 700 mb, but participants also noticed more small perturbations in the 850 mb height lines in the RRFSp2 Control. MUCAPE magnitudes were a mixed bag, although the spatial extent was not as widespread in the RRFS according to some participants. MLCAPE, however, almost always was noted to be higher in the HRRRv4 relative to the RRFSp2 Control. This impression was conveyed by participants commenting on not only higher maximum values, but also broader areal coverage of large CAPE.

### 3.2.3 Data Assimilation Strategies

The third comparison in the B group evaluations looked at the impact of data assimilation strategies in five deterministic models, focusing on the first twelve hours of the forecast. The domain for this comparison was shifted relative to the other comparisons, reverting to the previous day's domain to ensure that convection was captured within the area being evaluated. Participants answered questions about storm structure, retention, and location using the simulated reflectivity and 2–5 km UH fields at forecast hours 1 and 6, and rated one of three environmental fields (2-m temperature, 2-m dewpoint, and SBCAPE) on a scale of 1 (Very Poor) to 10 (Very Good). Participants were also asked to provide comments twice; first on the storm structure, location, and retention in the first 12 h of these runs, and then on the environmental field that they were assigned. Responses shown herein encompass the 15 cases where all model data was available; RRFSp2 control runs were unavailable on four days. Thus, 107 participant responses were collected for all models.

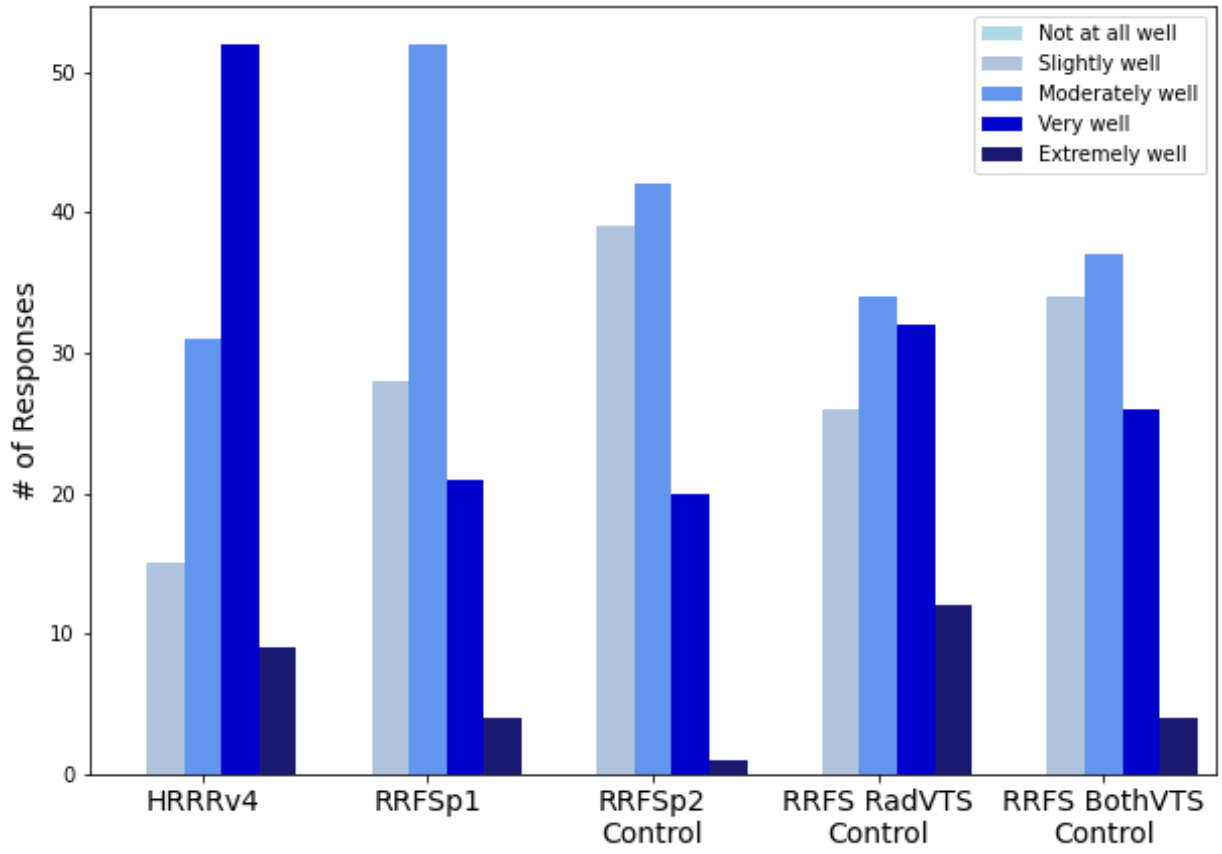


*Figure 19. Participant responses to the question, "At forecast hour 1, how well do the following models depict storms that were ongoing at the model initialization time? Consider aspects like storm retention, strength, and location in your answer."*

In the first forecast hour, participants most frequently selected the options of “Very well” or “Moderately well” when asked, “At forecast hour 1, how well do the following models depict storms that were ongoing at the model initialization time? Consider aspects like storm retention, strength, and location in your answer.” (Fig. 19). The HRRRv4 had the best performance of the models selected, with a majority of its responses being “Very well”. The most frequent response for the other models was “Moderately well”, though all models had participants respond “Very well” or “Extremely well” in some cases. In fact, the model with the largest number of “Extremely well” responses was the RRFS RadVTS Control. Overall, the RadVTS control seemed to perform better in the first forecast hour relative to the RRFS BothVTS Control, with more “Very well” responses too. No models elicited the response “Not at all well”, suggesting that in most cases the models are able to assimilate storms successfully. The RRFSp1 appeared to slightly outperform the RRFSp2 Control in the first 12 hours, similar to findings from B1 which covered forecast hours 13 through 36.

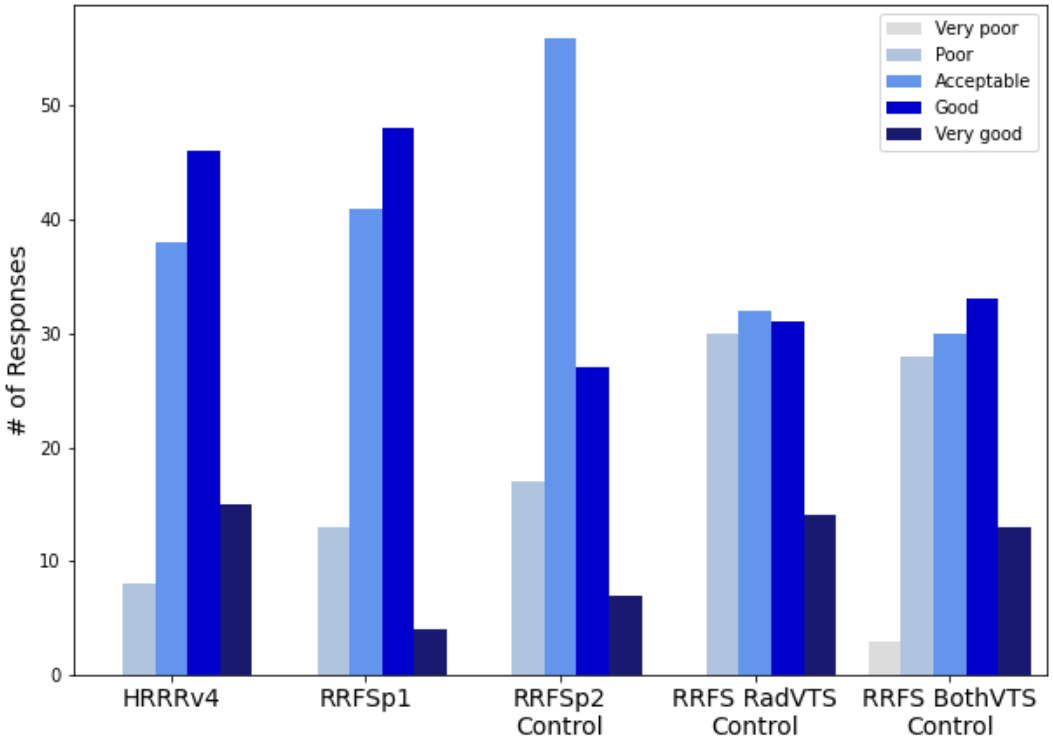


*Figure 20. Participant responses to the prompt, “Please evaluate the structure and location of storms at forecast hour 6 in the following models.”*

At forecast hour 6, the HRRRv4 and the RRFSp1 perform quite similarly, although the HRRRv4 has more “Very good” ratings than the RRFSp1 (Fig. 20). “Good” was the most frequent response for the HRRRv4, the RRFSp1, and the RRFS BothVTS Control runs, and was a very close second to “Acceptable” for the RRFS RadVTS Control run. The RRFS RadVTS Control and RRFS BothVTS Control runs performed more similarly at forecast hour 6 relative to their performance at forecast hour 1 (Fig. 10). Overall, the

HRRRv4 and the RRFSp1 perform best relative to the other three models at forecast hour 6, with fewer "Poor" responses than the other configurations. Participants systematically noted a lot of convective coverage early in the VTS simulations, which sometimes benefited the runs, and sometimes hurt the participant impression of the runs. This is exemplified by the participant comment: "*RRFS VTS runs seem hot compared to others. Too much coverage of storms in south Texas at 1 h. Interestingly enough, however, these were the best forecasts at 6 hours. All of the models did well with the MCS in TX/OK, but the VTS runs had arguably the best structure and also best captured the storms on the TX side of the Rio Grande.*" Similarly, a few days earlier a participant had commented "*Interesting switch around between HRRR and RRFS Rad/Both VTS Ctl, the latter representing the structures well at T+1 but degenerating into a mess by T+6, not really recognisable as the obs bar the general envelopes.*" Frequently, the RRFS RadVTS Control and the RRFS BothVTS Control were referred to collectively in the comments (e.g., "The VTS runs…"), suggesting that this comparison may also benefit from blinding in future SFEs.

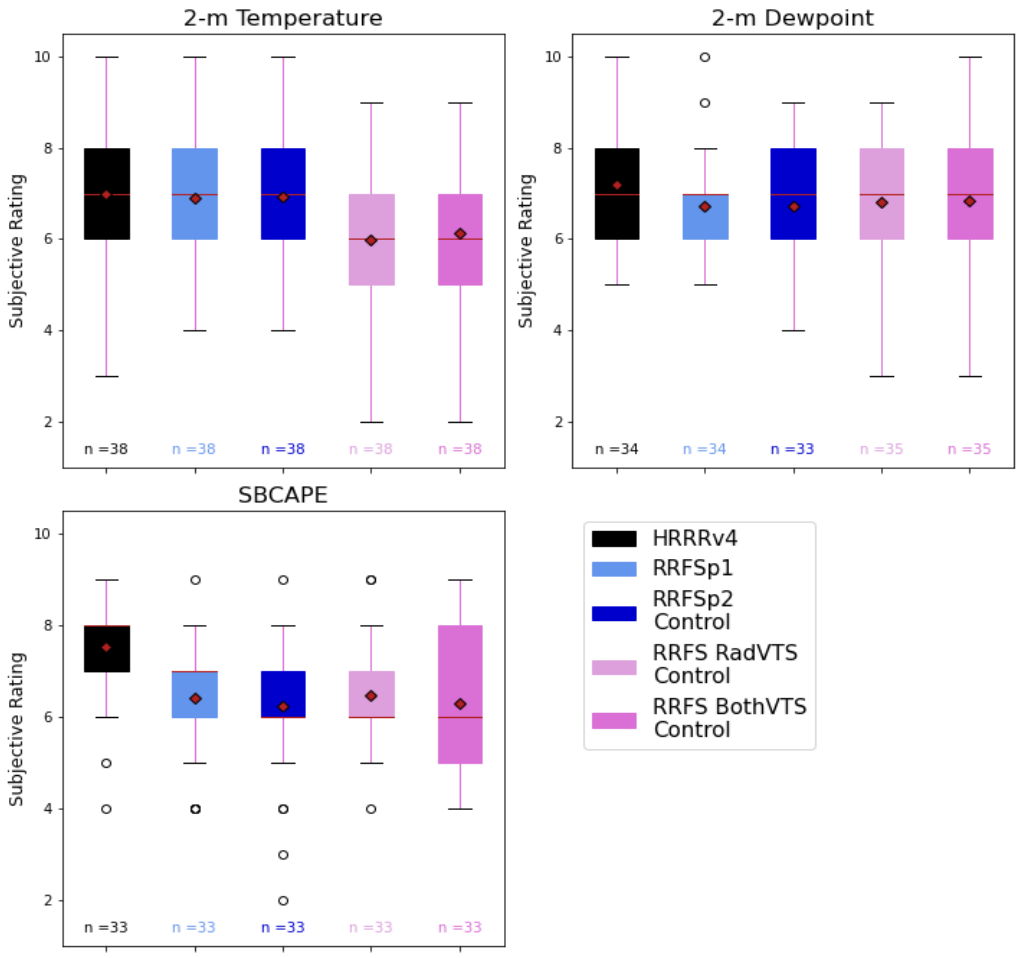


*Figure 21. Participant ratings of assigned environmental fields. Medians are indicated by the brown lines, and mean values are indicated by the brown diamonds. The number of samples in each box is indicated by the text at the bottom of each subplot. Note that different n-values in the 20m dewpoint are due to two participants not rating all models in their response.*

Finally, participants provided a 1–10 rating encompassing all 12 hours of the simulations for one of three environmental fields (Fig. 21), which were randomly assigned such that an approximately equal number of people evaluated each field. The HRRRv4, RRFSp1, and RRFSp2 Control performed slightly better than the RRFS RadVTS Control and the RRFS BothVTS Control for 2-m temperature, across the entire distribution of ratings. There were essentially two groups of models; the RRFS RadVTS Control and RRFS BothVTS Control performed approximately the same, as did the HRRRv4, RRFSp1, and the RRFSp2 Control. All models performed similarly with respect to dewpoint, although the HRRRv4 had the highest mean rating. For SBCAPE, the HRRRv4 performed best, having a higher mean and median than the other configurations. All of the other configurations performed similarly to one another, although the RRFS BothVTS Control member had high variability in its ratings as seen by the span of its box plot in Figure 21. Participants consistently noted a warm bias in the HRRRv4, RRFSp1, and RRFSp2 Control, and a cold bias in the RRFS RadVTS Control and RRFS BothVTS Control 2-m temperature forecasts. Comments on the 2-m dewpoints were mixed, with participants noting a wet bias on some days and a dry bias on others. One participant noted that the guidance was almost all too moist over the Great Lakes, which should be investigated by the developers to ensure this is not a systemic issue. Regarding SBCAPE, the RRFS VTS models were occasionally remarked to improve later in the forecast. Participants also noted an occasional underforecast of CAPE by the RRFS (FV3-based) models relative to the WRF-based HRRRv4, as in the following comment: "*RRFSs continue to have lower CAPE than obs and HRRR. However, the RadVTS and BothVTS hold onto higher CAPE values over KS longer than the other RRFSs.*"

### 3.2.4 FV3 Physics Suites

This comparison explored the impact of different physics parameterizations on the depiction of simulated reflectivity, 2–5 km UH, and environmental attributes including 2-m temperature, 2-m dewpoint, and SBCAPE. Five different models were evaluated, which utilized two different microphysics schemes, two planetary boundary layer (PBL) schemes, and three land-surface models (LSMs). These physics suites were implemented in different combinations such that effects from the individual schemes could be isolated (e.g., one model used Thompson microphysics, MYNN PBL, and NOAH LSM, another model used NSSL microphysics but MYNN PBL and NOAH LSM, and yet another model used Thompson microphysics and MYNN PBL, but NOAH-MP LSM). These runs were mainly available for the initial three weeks of SFE 2022, and a configuration change at the start of the fourth week led to all comparisons after that change to be excluded from the analysis herein. Participants ranked the five component models from best to worst, and assigned the best-performing model a rating from 1 (Very Poor) to 10 (Very Good).

Results for the 2–5 km UH and simulated reflectivity show the RRFSphys_02 and RRFSphys_04 configurations as consistently being ranked the highest, with both the highest mean ranking and the most frequent number one rankings amongst all of the members (Fig. 22). These two configurations were the two that utilized the NSSL microphysics, as opposed to the Thompson microphysics. Given that the simulated

reflectivity is heavily dependent on the microphysics schemes utilized, it is not surprising that there is some separation by microphysics. The RRFSphys_05 member consistently ranked the lowest. This member used the Thompson microphysics, but solely differed from the RRFSphys_03 members in that it used the TKE-EDMF PBL scheme. RRFSphys_04 also utilized the TKE-EDMF PBL scheme, confirming that the microphysics played a larger role in determining rankings for the simulated reflectivity and UH than the PBL schemes did. As with the B1 evaluations, participants noted the storm coverage, storm timing, and storm mode as important factors in completing their rankings and ratings; however, reflectivity gradients and location of significant features were also mentioned. Some participants noted spotty showers in warm advection areas, which they linked to the Thompson microphysics scheme. Other participants commented that most of the configurations were “too hot” (i.e., too intense) or otherwise overdid the amount of convection. A few participants also commented that differences were minimal between the runs on some days.

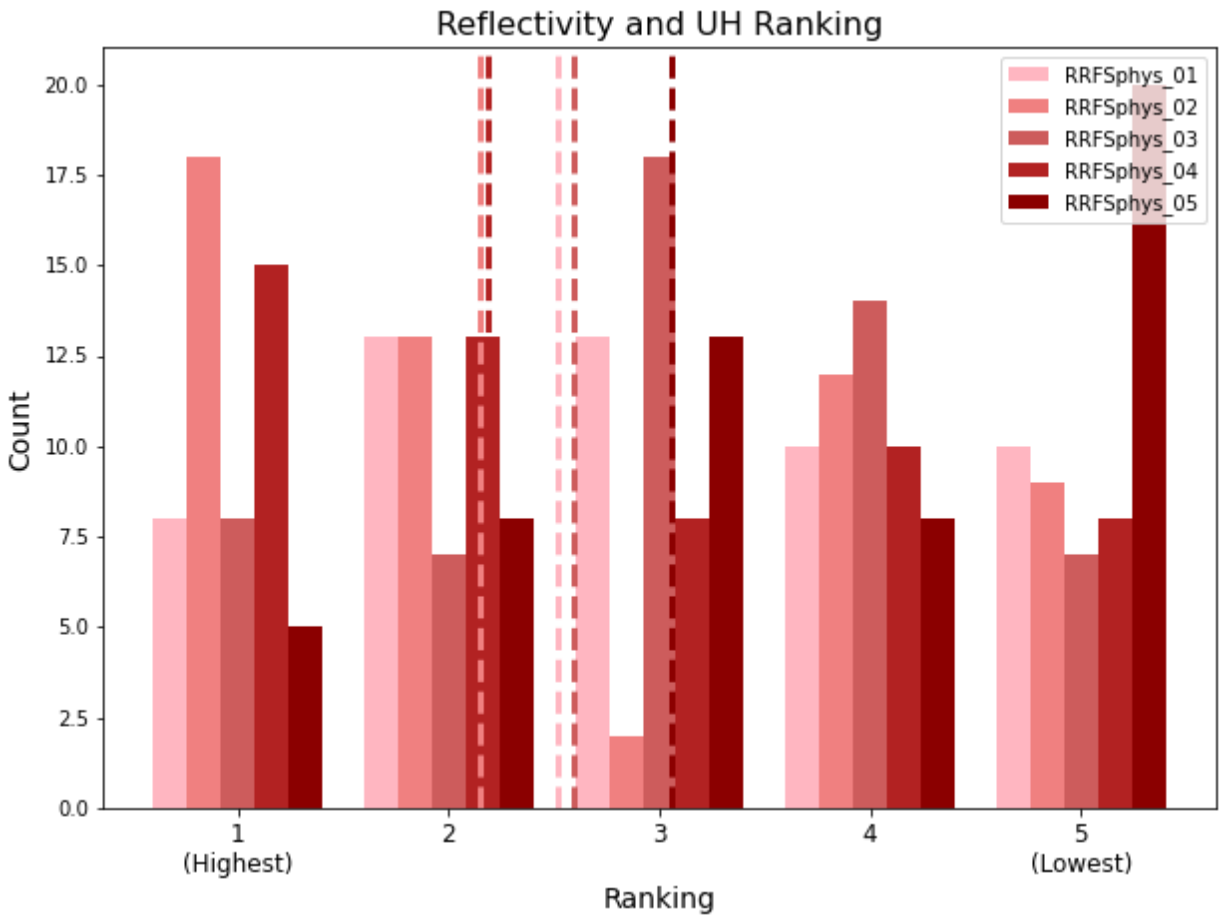


*Figure 22. Reflectivity and UH rankings for models in the FV3 Physics suites comparison. Dashed lines indicate the mean ranking (lower is better).*

Environmental fields tell different stories depending on the field (Fig. 23). For the temperature, the RRFSphys_04 again is the top-ranked model, and its mean ranking is

quite far ahead of the remainder of the configurations (Fig. 23a). In the comments, participants noted that RRFSphys_04 was frequently cooler than the other configurations, which evidently benefitted the forecasts. Cold pools were also noted as being too cold on occasion. However, for dewpoint, the RRFSphys_04 ranks third, after the RRFSphys_02 and the RRFSphys_01 configurations (Fig. 23b). Participants noted that RRFSphys_04 was frequently much too moist, while RRFSphys_03 and RRFSphys_05 were noted as having dry biases. These two configurations both use the MYNN PBL scheme and the NOAH LSM, and differ only in their microphysics configurations. For SBCAPE, the best performing configuration in terms of mean ranking and number of times it was ranked first was far and away the RRFSphys_01 (Fig. 23c), though participant comments frequently noted that all SBCAPE was underdone. These results combined with the prior reflectivity and UH comparisons seem to suggest that the NSSL microphysics, MYNN PBL scheme, and NOAH LSM perform best overall (RRFSphys_02). However, elements of the Thompson microphysics scheme (the sole difference between RRFSphys_01 and RRFSphys_02) may lead to it performing better in terms of SBCAPE.

Ratings for the top-ranked model in each of these fields (Fig. 24) showed more stratification than the same plots for the B1 comparison (Fig. 4), perhaps due to a smaller sample size. For Reflectivity and UH, the RRFSphys_02 and the RRFSphys_05 had the highest median ratings, and the RRFSphys_02 had the highest mean. The RRFSphys_04, while chosen second most frequently as the best model, had a wider distribution of ratings when selected as the best model. The 2-m temperature, 2-m dewpoint, and SBCAPE all had relatively small sample sizes (below n=10 in all but one case), precluding robust comparisons and indicating that objective verification of these fields would likely be a useful supplement to the subjective data collected herein.

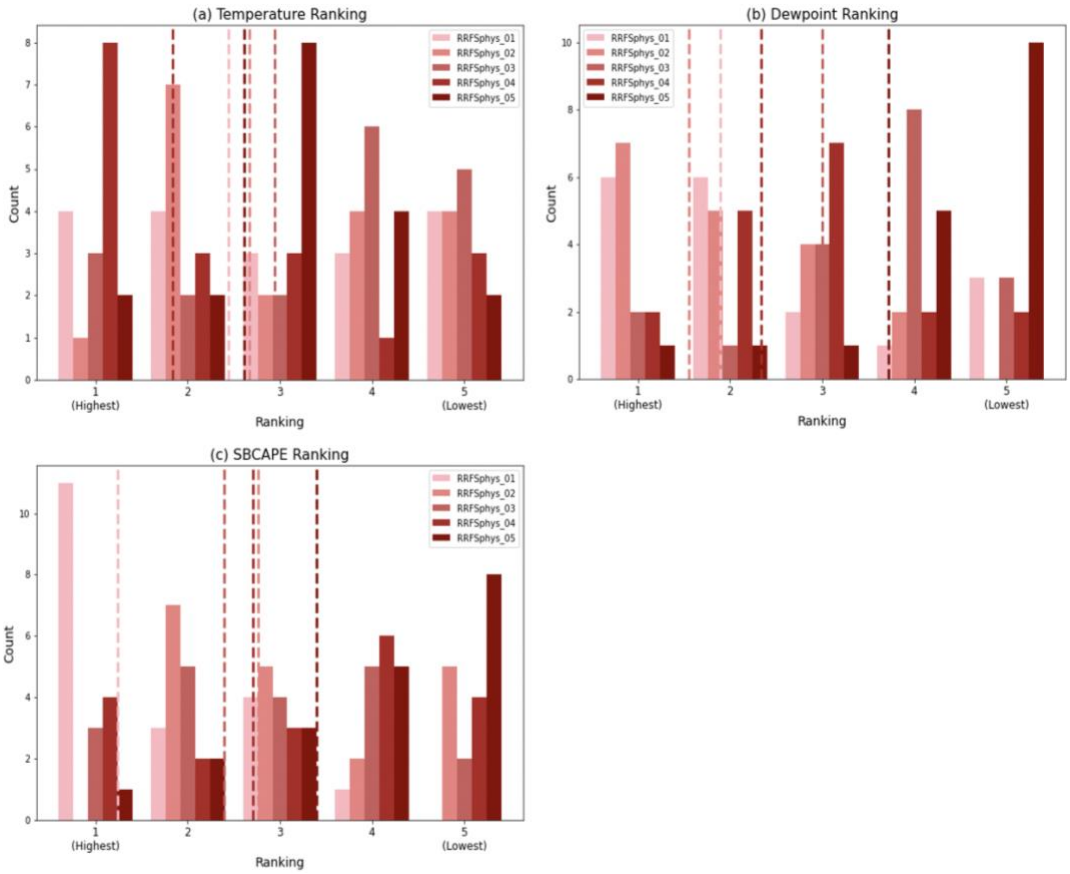


*Figure 23. Rankings of environment for the FV3 Physics Suites comparisons. Rankings were completed for (a) 2-m Temperature, (b), 2-m Dewpoint, and (c) SBCAPE. Dashed lines indicate the mean ranking for the model in question (higher is better).*

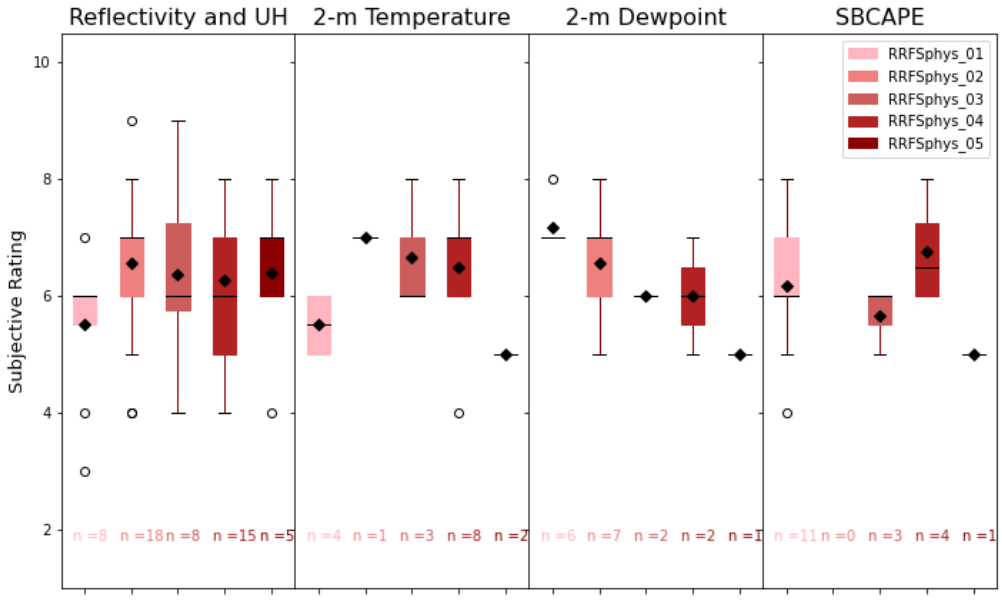


*Figure 24. Participant ratings for the top-ranked model in each comparison. Medians are shown by the black line; black diamonds indicate the mean. The sample size is listed below each box-and-whisker plot.*

### 3.2.5 1-km vs. 3-km NSSL-WRF

The final evaluation unique to the B group looked at the impact of increased horizontal grid spacing on severe convective forecasts, focusing specifically on UH at two different levels, low-level wind speed depiction, and various storm characteristics. Participants compared a 1-km version nested within a 3-km version of the same NSSL-WRF configuration at forecast hours 12–30 directly for simulated reflectivity characteristics, and answered questions about both configurations with regard to UH and hourly maximum 10-m wind speed. Due to a typo in the question about the maximum 10-m wind speed, data for this question is only available after 12 May 2022.

For all characteristics except storm structure, the most frequent response to "*Which model best depicts the following aspects of severe convective storms?*" was that the 1-km and 3-km NSSL-WRF runs performed about the same (Fig. 25). However, the 1-km NSSL-WRF was selected as performing the best far more frequently than the 3-km NSSL-WRF for the number of storms, storm structure, storm evolution, and timing of convective initiation. The 1-km NSSL-WRF appeared to be most skilled at capturing the number of storms and the storm structure relative to the 3-km NSSL-WRF.

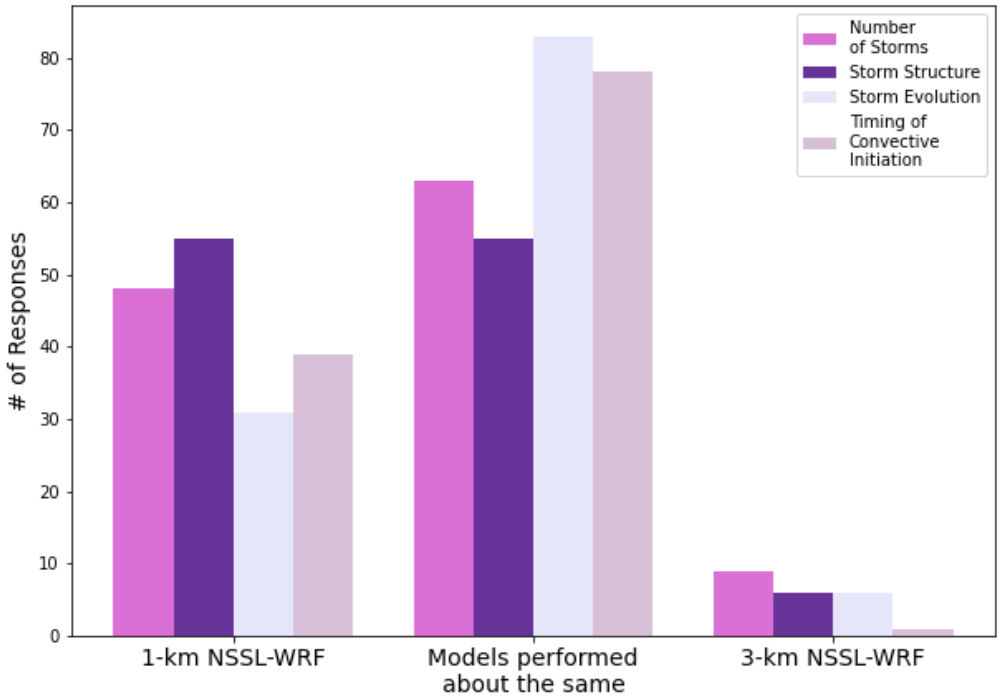


*Figure 25. Participant responses to the question "Which model best depicts the following aspects of severe convective storms?"*

When asked specifically how well severe weather proxies performed for each of the models, participants were asked about 2–5 km UH's delineation of a total severe threat in each model, 0–2 km UH's delineation of a tornado threat in each model, and the

hourly maximum 10-m wind speed's delineation of a wind threat in each model. Overall, participants indicated that the 1-km NSSL-WRF performed better than the 3-km NSSL-WRF for all severe weather proxies indicated (Fig. 26). The 2–5 km UH achieved higher scores as a proxy for total severe than the 0–2 km UH as a proxy for tornadoes or the 10-m wind speed as a proxy for severe convective winds. Participant evaluation indicated that these proxies had relatively little value, with the distribution for each of the models and fields centered around the middle (1-km NSSL-WRF 2–5 km UH), second-lowest (3-km NSSL-WRF 2–5 km UH and both models' 0–2 km UH), or the lowest (both models' 10-m Wind) option available to them in the survey. Causes for these discrepancies and overall poor performance should be investigated. One participant offered a potential direction for investigation in their comments, noting that "*The 1-km NSSL run storm characteristics are very similar to those seen in the 3-km NAM nest. The NAM nest diffusion on hydrometeors and Smagorinsky diffusion were reduced in the latest version*." Participants also noted the relative discontinuities of the 1-km horizontal grid spacing relative to the 3-km horizontal grid spacing, and that this visualization can influence impressions of the model performances. Unfortunately, many participant comments indicated that these two models did not provide useful guidance during the cases that they were evaluating. Utilizing 1-km models will require further calibration of forecasters and post-processed guidance due to differences with regards to the storm features. As one participant stated, "*1 km version looked better due to more resolution, but not sure if it was a better forecast. Would definitely need to recalibrate for the much higher UH values from the 1 km model!*"

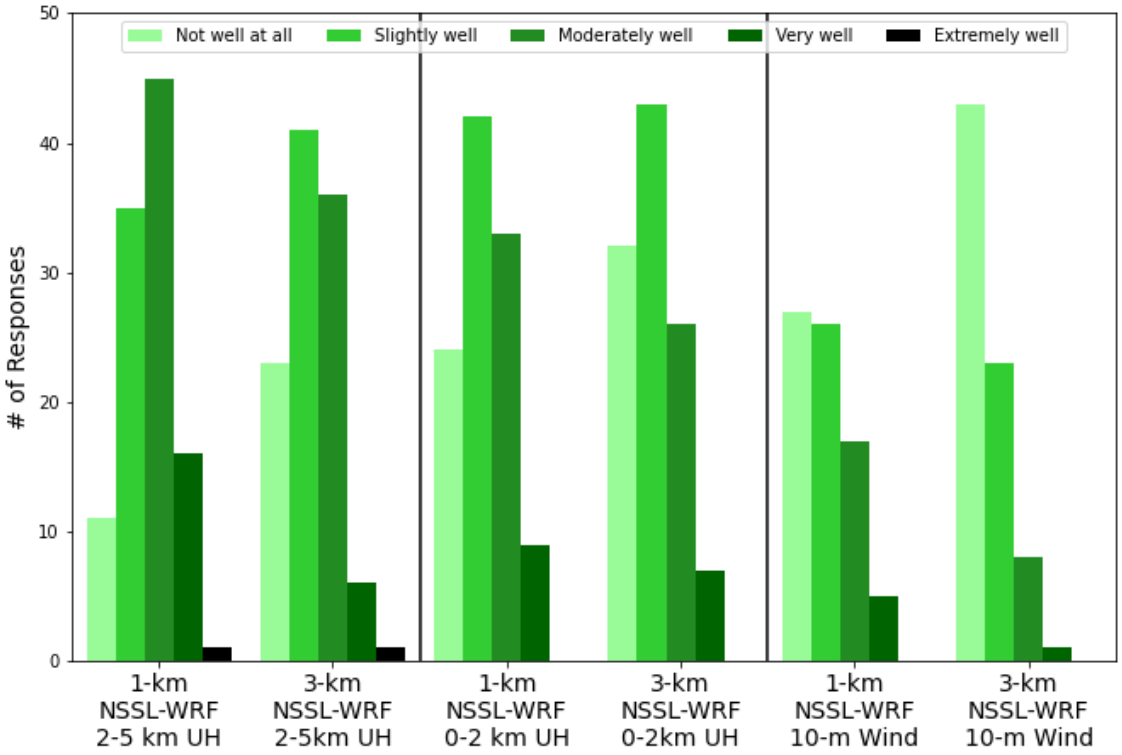


*Figure 26. Participant responses to the question of how well storm proxies delineated the overall severe threat (2–5 km UH), tornado threat (0–2 km UH), and wind threat (10-m Wind).*

## 3.3 Model Evaluations – Group C: CAM Ensembles

### 3.3.1 CLUE: 00Z Ensembles

This evaluation compared four 0000-UTC initialized, FV3-LAM CAM ensembles to HREFv3: (1) RRFSp2e, (2) MAP-VTS-rad, (3) MAP-VTS-bot, and (4) RRFSp2eMP. 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, RRFSp2e is a prototype for what will eventually be implemented operationally as RRFSv1, replacing HREF and subsuming several other regional systems to simplify NCEP's production suite. 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 (>99.85$^{th}$ percentile) were displayed along with preliminary local storm reports (e.g., Fig. 27). A significant change relative to previous years was that these were blind evaluations. The panels were labeled as "Model A", "Model B", etc., and the configuration of the panels was randomized daily. This helped to negate any implicit or explicit biases of participants and facilitators. The models were revealed by the facilitators to the participants after the evaluations were submitted. Another change relative to previous years was that the ensembles were evaluated using rankings (best = 1; worst = 5) instead of the 1-10 ratings, where 1 was worst and 10 was best. The rankings allow for a more consistent range of scores, unlike the ratings where there can be significant variability depending on how individual participants perceive and score "good" or "bad" forecasts.

The boxplots showing the distributions of subjective rankings are displayed in Figure 28. Note, because rankings were used, the results had to be divided according to model availability. So, rankings for cases where all ensembles were available are shown in Fig. 28a, where all ensembles except RRFS MixPhys were available are shown in Fig. 28b, and where all ensembles except RRFSp2e and RRFS MixPhys were available are in Fig. 28c. This accounts for all 19 days over which evaluations were conducted. The main takeaway from this comparison was that HREFv3 and RRFSp2e consistently were ranked highest and performed quite similarly. Out of the RRFS MixPhys, RRFS RadVTS, and RRFS BothVTS ensembles, none stood out as being ranked better or worse, but they were consistently ranked lower than HREF and RRFSp2e.

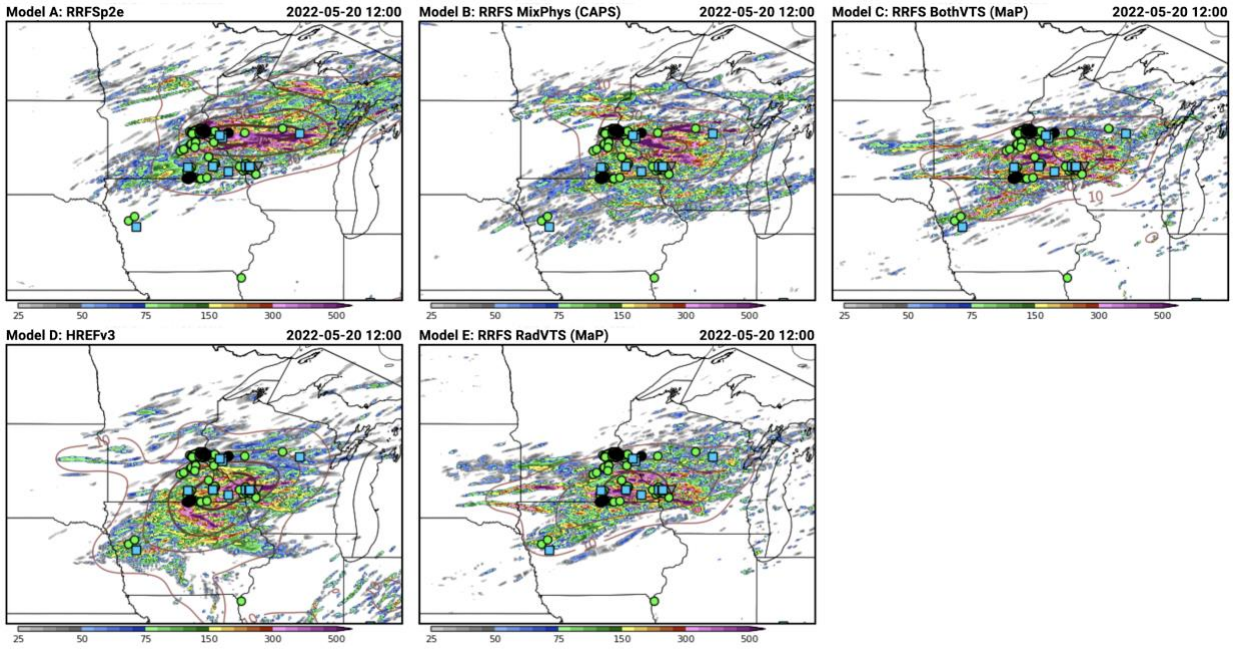


*Figure 27. Example of multi-panel comparison webpage for the 0000 UTC CAM ensemble C1 evaluation during the 2022 SFE. The 24-h ensemble maximum UH (shaded) and neighborhood probability of UH>99.85th percentile (contoured) is displayed for RRFSp2e (upper left), RRFS MixPhys (upper middle), RRFS BothVTS (upper right), HREFv3 (lower left), and RRFS RadVTS (lower middle) for 19 May 2022. Preliminary severe storm reports are also overlaid (wind – blue squares, hail – green circles, and tornado – red upside-down triangles. Significant reports are filled in black). Note, only the "Model A", "Model B", etc., labels were displayed during evaluations.*

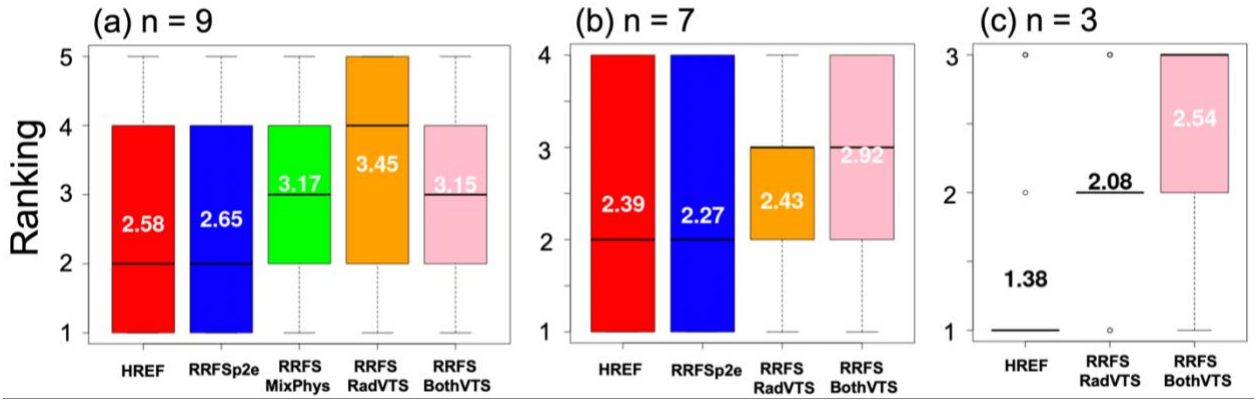


*Figure 28. Box plots showing the distributions of rankings by SFE participants of hourly maximum fields for severe weather forecasting over a mesoscale area of interest for the forecast hours 13-36 for the C1: CLUE 00Z Ensembles evaluation (HREF – red; RRFSp2e – blue; RRFS MixPhys – green; RRFS RadVTS – orange; RRFS BothVTS – pink). The numbers overlaid on each bar indicate the value of the average ranking and the horizontal line indicates the median. (a) Rankings distributions for the days that all five ensembles were available, (b) Same as (a) except for days when only RRFS MixPhys was missing, (c) same as (a) and (b) except for days when RRFS MixPhys and RRFSp2e were missing.*

Participants were also asked to subjectively rate on a scale of 1–10 (1 = worst; 10 = best), the one ensemble that they ranked best. The distributions of these ratings are displayed in Figure 29. By far, the HREF and RRFSp2e were the top ranked ensemble

most frequently, with HREF being best 43% of the time and RRFSp2e 31% of the time. The average ratings for the top ranked ensembles were tightly clustered between 7 and 8, and RRFSp2e had the highest average at 7.86.

Participant comments most frequently pointed out timing and probability magnitude differences in the ensembles. When timing errors were noted in any ensemble, convection initiation or translation speed was most often too slow. In a couple cases, it was noted that HREFv3 had the most ensemble spread and that the RadVTS and BothVTS runs had the least spread. It was also noted several times that performance depended on whether UH or wind diagnostics were compared to the distribution of LSRs.

Although the HREF has been a formidable baseline for several years, the performance of the RRFSp2e in this year's SFE was very encouraging with generally very similar subjective scores compared to HREF. These results give some confidence that with further research and development, RRFSv1 can be transitioned into operations – a very important step for NOAA's UFS initiative.

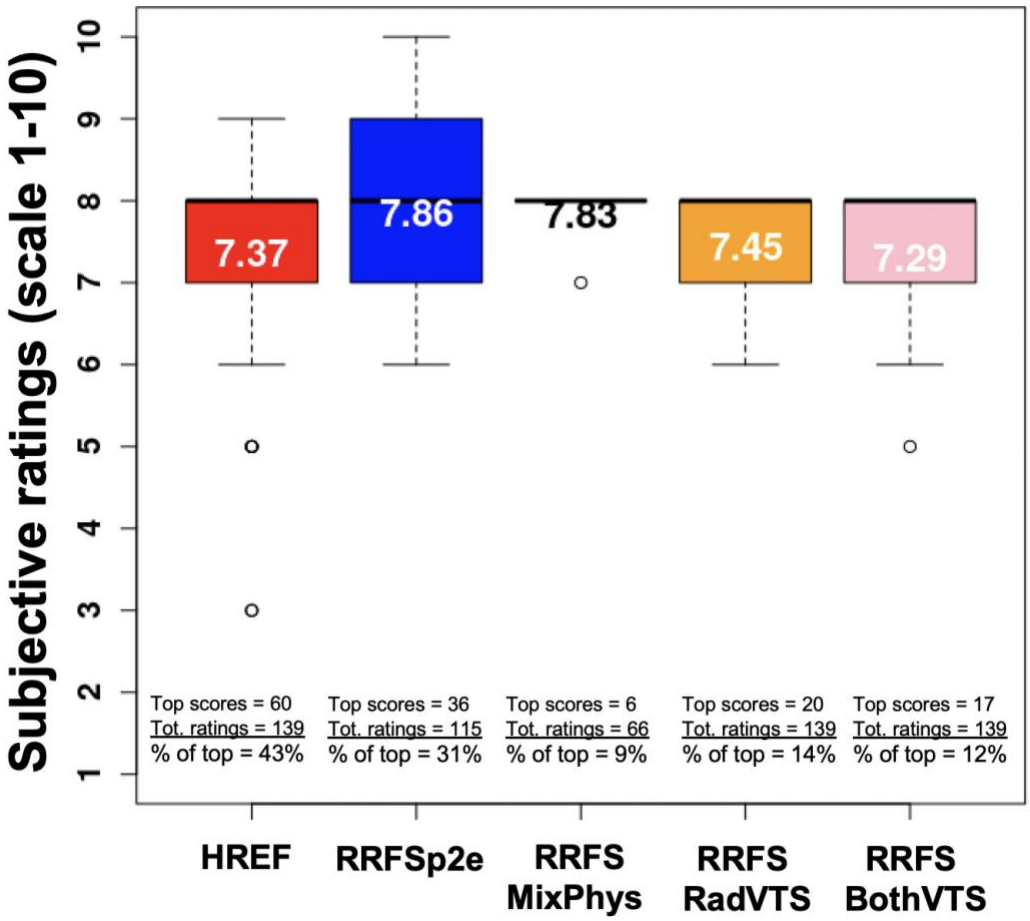


*Figure 29. Distributions of subjective ratings for each ensemble when it was ranked as the best performing. Numbers overlaid on the boxplots indicate mean ratings, while the horizontal black line indicates the median. At the bottom of the panel, "Top scores" refers to the number of top ranked forecasts for the corresponding ensemble, "Tot. ratings" is the total number of times that particular ensemble received any ranking, and the "% of top" is the percentage of time that ensemble was ranked the top for the cases that ensemble was available.*

### 3.3.2 RRFSp2e vs. HREF

This evaluation featured an in-depth examination of several storm attribute and environmental fields from 00Z initializations of RRFSp2e and HREFv3. These direct comparisons served to unearth ways in which the current operational CAM ensemble (HREFv3) differs from a candidate to replace it (the RRFSp2e). Specifically, participants were asked to compare mean environmental fields (2-m temperature, 2-m dewpoint, and surface-based CAPE) in the HREF and RRFSp2e ensembles, as well as UH aggregated over 4-h time windows, for the periods 1700-2000, 2100-000, and 0100-0400 UTC. For each time period and field, participants indicated if RRFSp2e was *much worse, worse, about the same, better,* or *much better* than HREFv3. The mean environmental fields were compared to analyses from the 3D-RTMA, while UH was compared to LSRs. An example of the displays used in this evaluation from 2000 UTC 20 May 2022 is shown in Figure 30.

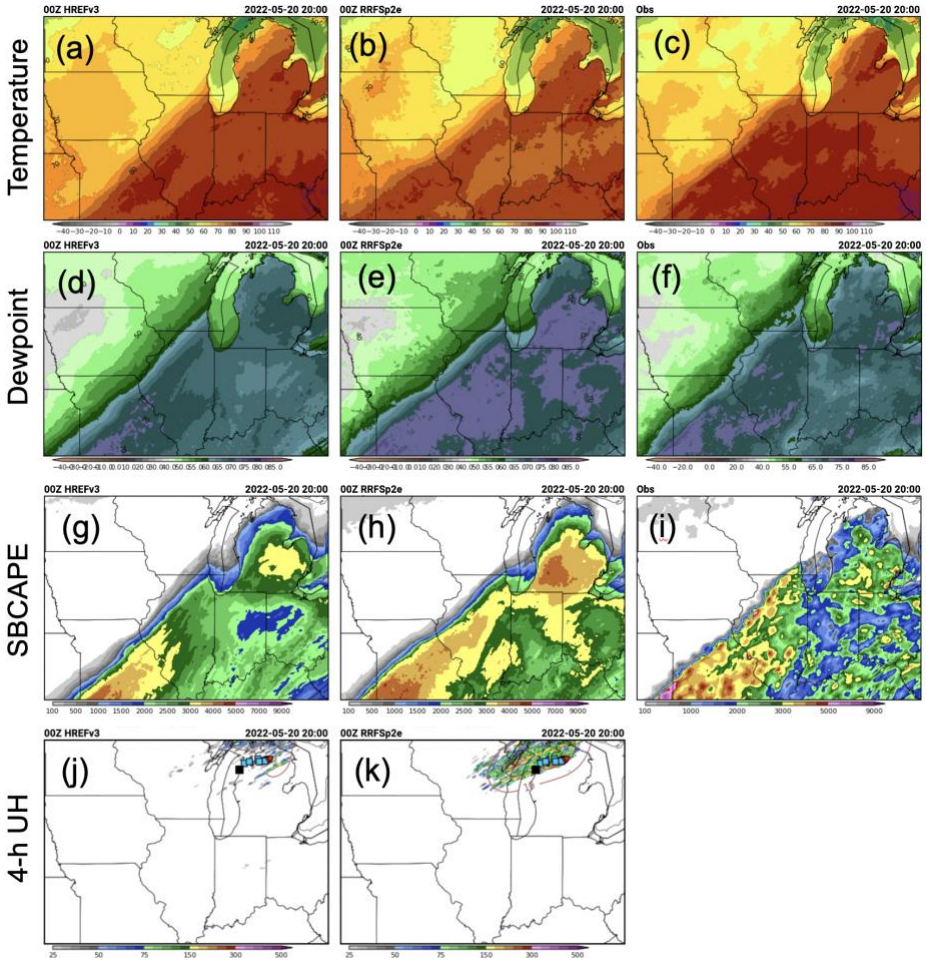


*Figure 30. The ensemble mean 2-m temperature fields valid 2000 UTC 20 May 2022 from (a) HREFv3, (b) RRFSp2e, and (c) 3D-RTMA. (d) – (f) and (g) – (i) same as (a) – (c), except for 2-m dewpoint and surface-based CAPE, respectively. 4-h maximum UH and neighborhood maximum probabilities of UH ≥ 99.85$^{th}$ percentile with LSRs overlaid for (j) HREFv3 and (k) RRFSp2e.*

The aggregate results from these comparisons are shown in Figure 31. The participant selections were converted to numerical values so that *much worse* = -2, *worse* = -1, *about the same* = 0, *better* = 1, and *much better* = 2. The results were highly dependent on the field examined and the time window. The largest differences occurred at the earlier times (i.e., 1700–2000 UTC), and the amplitude of these differences decreased with lead time. For temperature, RRFSp2e was generally worse than HREFv3, while for dewpoint and surface-based CAPE, RRFSp2e was generally better. The 4-h time window UH forecasts were generally rated about the same. From the participant comments, it was consistently noted that RRFSp2e had a distinct cool bias that was most prevalent at the earlier times in the forecasts. For dewpoint, RRFSp2e was closer to observations, and HREFv3 had a dry bias. For surface-based CAPE, RRFSp2e had magnitudes that more closely matched observations than HREF. Finally, for UH it was often noted that differences were mostly small.

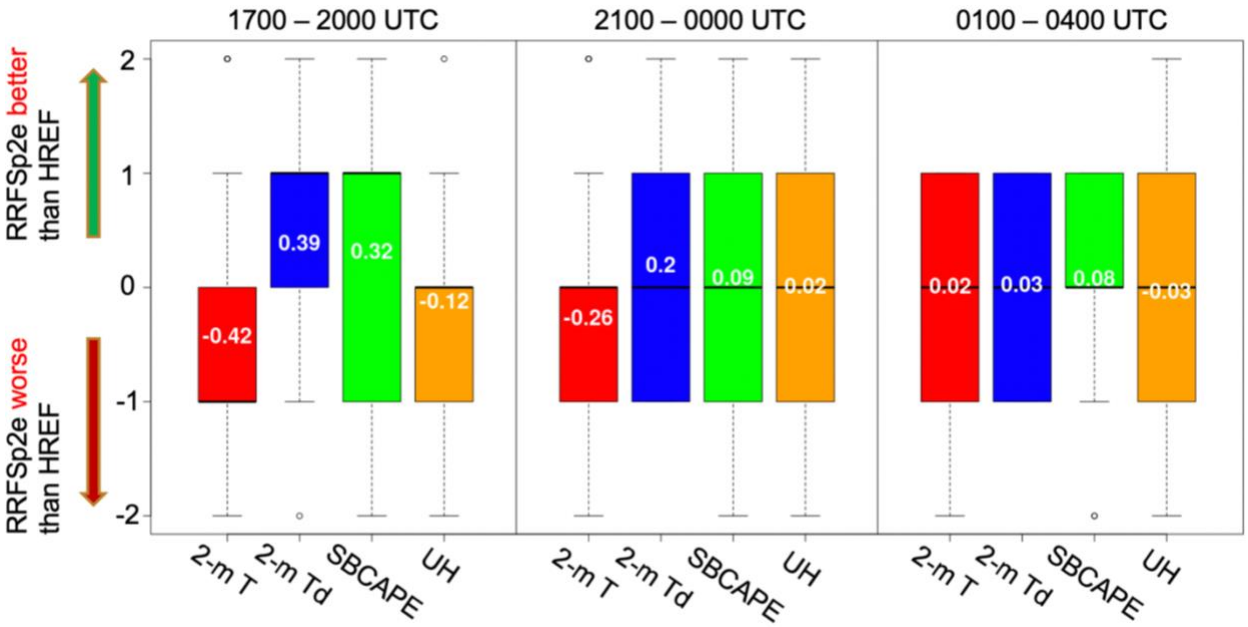


*Figure 31. Aggregate results for the comparisons between HREFv3 and RRFSp2e. The participant selections were converted to numerical values so that much worse = -2, worse = -1, about the same = 0, better = 1, and much better = 2, and then boxplots of the distributions were plotted.*

### 3.3.3 CLUE: Data Assimilation

This evaluation focused on the first 12 h of 0000 UTC initialized forecasts from three CAM ensembles that employed different data assimilation strategies and compared their forecasts to HREFv3. Participants were asked, "*Focusing on the first 12 hours of the forecast, evaluate four 00Z-initialized ensembles with different data assimilation approaches. Three ensembles use a hybrid EnVar approach, but two also use a technique called valid-time-shifting (VTS) to increase the membership of the background ensemble by a factor of 3. One of the ensembles only applies VTS to radar data (RRFS*

*RadVTS), while another applies VTS to both radar data and conventional observations (RRFS BothVTS). The RRFSp2e does not use VTS (note, there are slight differences between RRFSp2e and the VTS ensembles, so the RRFSp2e comparisons are not strictly controlled). For each ensemble, evaluate 1-h composite reflectivity paintballs and probabilities, 4-h Updraft Helicity, and an environmental field (2-m Temperature, 2-m Dewpoint, or Surface-based CAPE will be randomly selected) for the 0000–0400, 0500–0800, and 0900–1200 UTC time periods*." An example of the display of composite reflectivity probabilities used for these evaluations is shown in Figure 32.

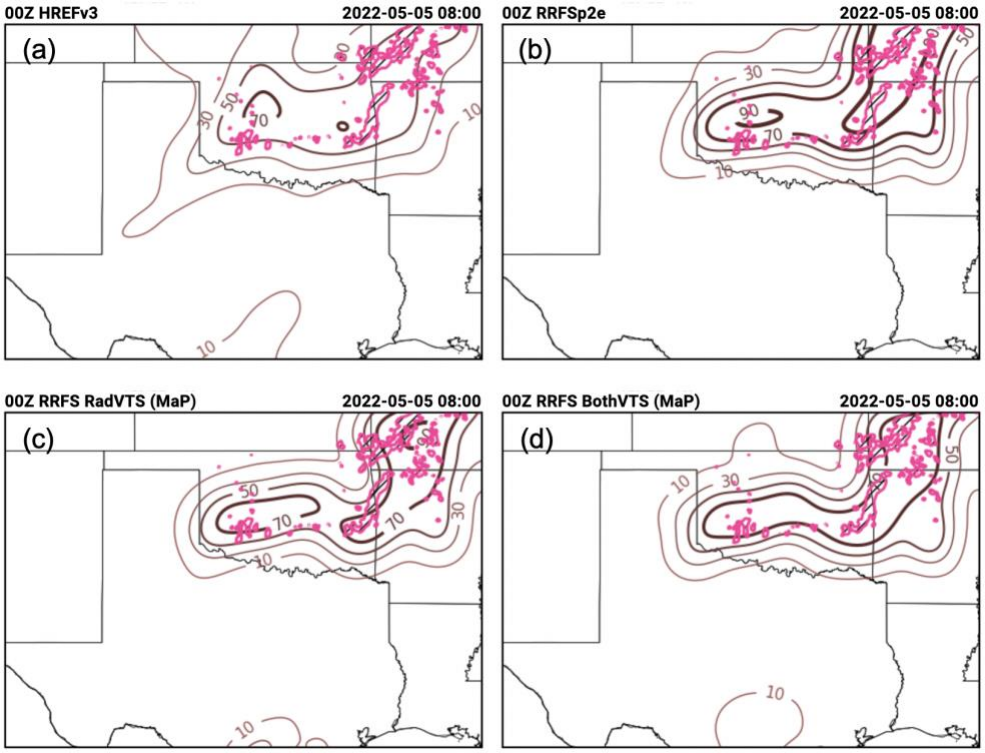


*Figure 32. Neighborhood maximum probabilities of composite reflectivity ≥ 40 dBZ valid 0800 UTC 5 May 2022 for 0000 UTC initializations of (a) HREFv3, (b) RRFSp2e, (c) RRFS RadVTS, and (d) RRFS BothVTS. In each panel observed composite reflectivity ≥ 40 dBZ is indicated by the pink contours with hatching inside.*

For UH, RRFS RadVTS had a slight advantage over RRFS BothVTS at forecast hours 0-4, but this advantage decreased at later lead times (Fig. 33). The UH RRFSp2e scores were generally similar to both of the VTS configurations across all three lead times. For composite reflectivity, once again RRFS RadVTS had a slight advantage at forecast hours 0–4, but at later forecast hours RRFS BothVTS and RRFSp2e ratings were similar. For 2-m temperature, results were very similar for all three models in each time period, and for 2-m dewpoint the BothVTS runs had an advantage for the 0–4 and 5–8 h time periods, but for hours 9–12 results were similar between all three models. Finally, the surface-based CAPE results jumped around quite a bit. The VTS runs were notably better than RRFSp2e for the 0–4 h period, but the next time period this result switched with the RRFSp2e being notably better than each of the VTS runs.

Several themes emerged from the comments. The first was that at many times the differences were relatively minor between the three systems. At other times, superior performance was noted in RRFSp2e, but there were just as many cases where the VTS runs were noted as performing better than RRFSp2e. Thus, the comments did not reveal consistent differences in performance. Several comments noted a cool temperature bias for all three models. When comparing the two VTS runs, oftentimes it was noted that they performed very similarly. There were also times when either Rad or Both was noted as performing best, so again the comments did not reveal systematic differences in forecast quality.

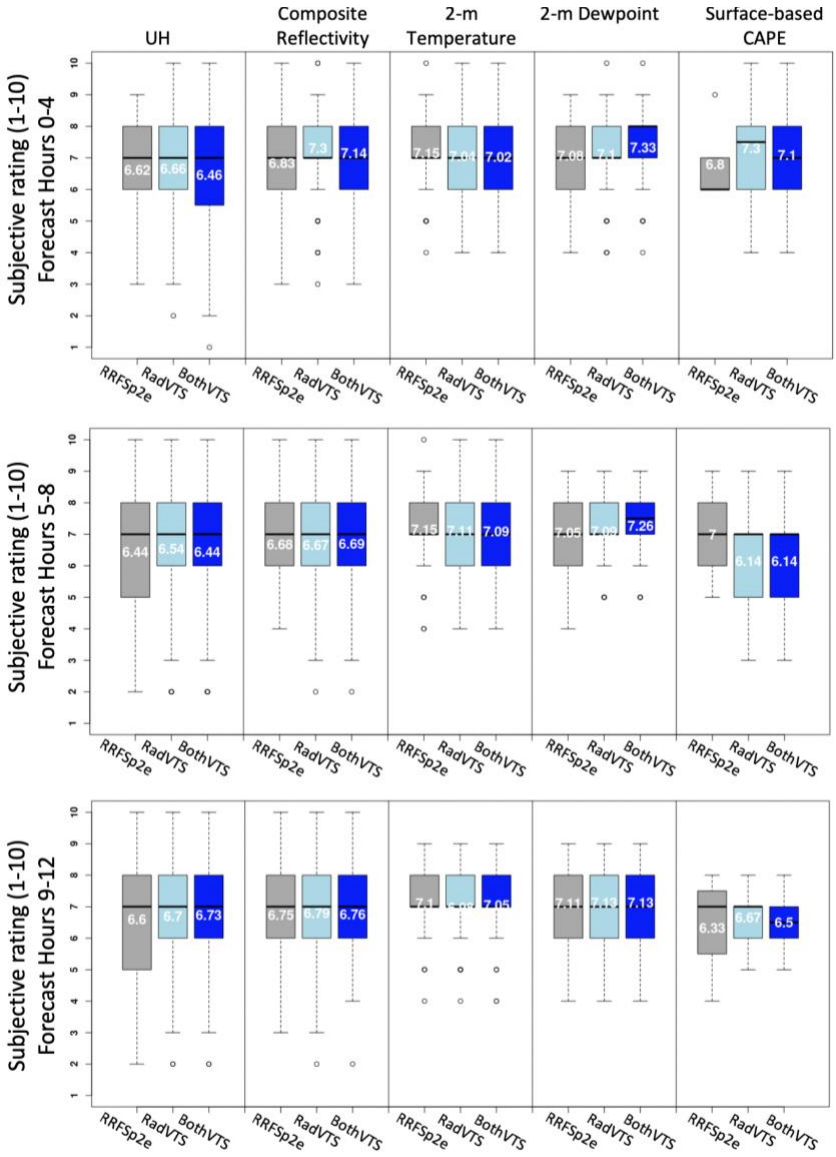


*Figure 33. Distributions of subjective ratings (1–10) by SFE participants for the C3 CLUE: Data Assimilation evaluation. The top, middle, and bottom set of boxplots are for forecast hours 0–4, 5–8, and 9–12 h, respectively. In each row, distributions are shown for RRFSp2e (gray), RRFS RadVTS (light blue), and RRFS BothVTS (dark blue) from left to right for the variables UH, composite reflectivity, 2-m temperature, 2-m dewpoint, and surface-based CAPE. The numbers in white text indicate mean ratings. The horizontal black lines indicate the median.*

### 3.3.4 TTU Ensemble Subsetting

This evaluation involved a collaboration with Texas Tech University to look at how ensemble sensitivity analysis could potentially be used to optimize ensemble-derived forecast probabilities. Specifically, ensemble sensitivity was used to identify a subset of 6 members with the smallest errors from a 20-member ensemble composed of RRFSp2e and RRFSp2eMP. Neighborhood probabilities of UH $\geq$ 100 $m^2s^{-2}$ and composite reflectivity $\geq$ 40 dBZ within 40-km of a point for a 6-h time window were derived from the 6-member subset, as well as the full 20-member ensemble. In the survey, participants were asked to determine how the subset probabilities compared to those from the full ensemble (i.e., were the subset probabilities *much worse*, *worse*, *about the same*, *better*, or *much better*). Each day, before the survey was administered, participants worked with the facilitators to choose the area and 6-h time interval over which to apply the ensemble subsetting, which would be evaluated the next day. The strategy here was to choose an area with uncertainty since the subsetting is designed to increase certainty by increasing or decreasing probabilities in the "right" direction (i.e., decrease where storms do not occur, and increase where they do occur). An example of these comparisons is shown in Figure 34 and the summary of results is shown in Figure 35.

One issue with this evaluation was that there were only 9 out of 19 days in which all the members of the full, 20-member ensemble were available. In these situations, it was usually only the RRFSp2e ensemble that was available, so the 6-member subset was drawn from only 10 members. Ideally, more members are desired to choose from to maximize the chances of finding members with small errors. Because of this issue, results were compiled from all the cases (Figs. 35a and c; i.e., including cases where data from the full ensemble was missing), and cases for which data from all members of the full ensemble were available (Figs. 35b and d). For UH probabilities (Figs. 35a and b), for both all days and for the subset of days in which all data was available, the subsetting most frequently resulted in either *about the same*, or *better* forecasts than the full ensemble. After converting to numerical ratings, it was found that the mean scores (0.41 for all days and 0.5 for only days with all data) were significantly greater than 0, where a one sample Student's t-test was used for hypothesis testing. This indicates that, on average, the subsetting resulted in improved UH probabilities. On the other hand, for the composite reflectivity probabilities, the mean subjective ratings were not significantly different than 0, indicating that subsetting did not result in improved reflectivity probabilities.

Examples of comments from participants on days when the 6-member subset performed better than the full ensemble included, "The subset moves the probabilities closer to the storm reports AND increases the probability magnitudes. Fascinated that only 6 members can make that difference...", "The subset had more realistic probabilities, but were slightly displaced to the west compared to where the storm reports occurred", and "Generally it is hard to distinguish large differences however I thought the 6-member subset narrowed down the solution a bit more and reduced the FAR."

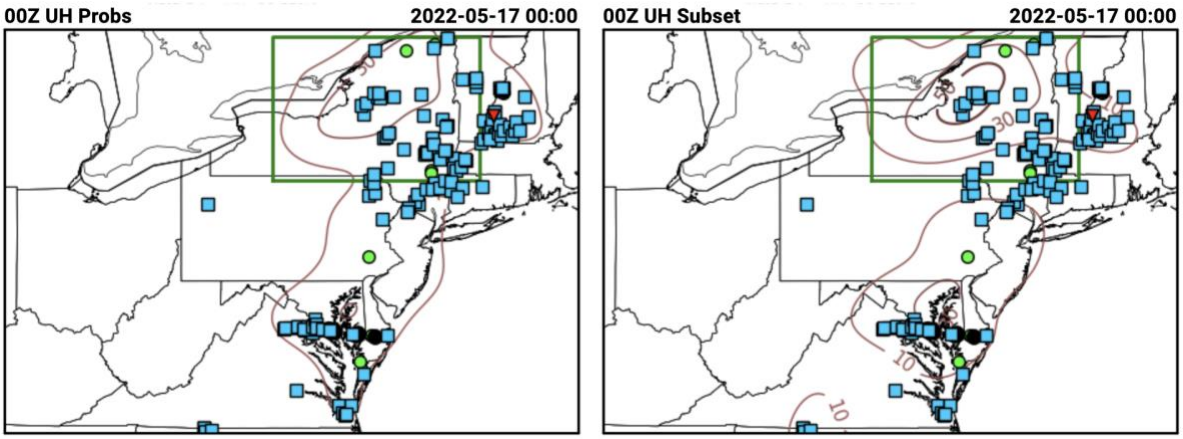


*Figure 34. Neighborhood Maximum Ensemble Probability of UH ≥ 100 $m^2s^{-2}$ derived from (a) the full 20-member ensemble, and (b) the 6-member subset. Locations of LSRs are overlaid in each panel. The forecasts are valid over a 6-h time window ending 0000 UTC 17 May 2022.*

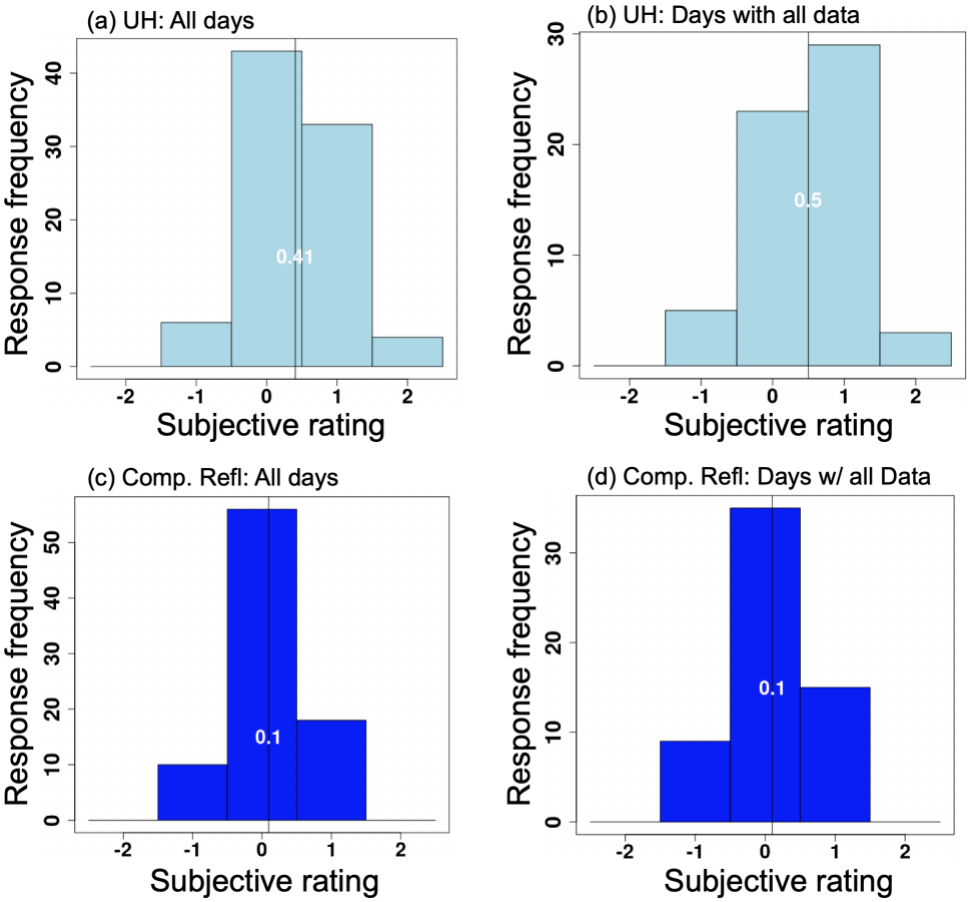


*Figure 35. Histograms showing the response frequencies to whether the subset probabilities were much worse (-2), worse (-1), about the same (0), better (1), or much better (2) than probabilities derived from the full ensemble for (a) UH probabilities calculated on all days, (b) UH probabilities calculated on days with all ensembles available, (c) composite reflectivity probabilities calculated on all days, and (d) composite reflectivity probabilities calculated on days with all ensembles available. In each panel, the number in white text and corresponding vertical line marks the mean subjective rating.*

### 3.3.5 WoFS: Number of Members

In this evaluation, UH- and reflectivity-based probabilities and paintball plots from 2100 and 2300 UTC WoFS initialization were compared for the full 18-member WoFS, as well as 9- and 13-member subsets. The purpose of this evaluation activity was to see whether it might be possible to run WoFS with fewer members and get the same forecast quality for hourly probabilistic forecasts. If it is possible to run WoFS with fewer members while maintaining the same level of forecast quality, it is possible that gains in skill could be achieved through reducing membership but using more advanced physics, data assimilation, and/or enhancing the resolution, while using the same number of computational resources. For both the 2100 and 2300 UTC WoFS initializations participants were asked, "*On a scale of 1-10 (1 = very poor; 10 = very well) rate the quality of storm attribute (1- and 4-hourly) and reflectivity forecasts from 9, 13, and 18 members... Feel free to compare whichever storm attribute fields you find most useful/relevant*...". An example display is shown in Figure 36 and results are summarized in Figure 37.

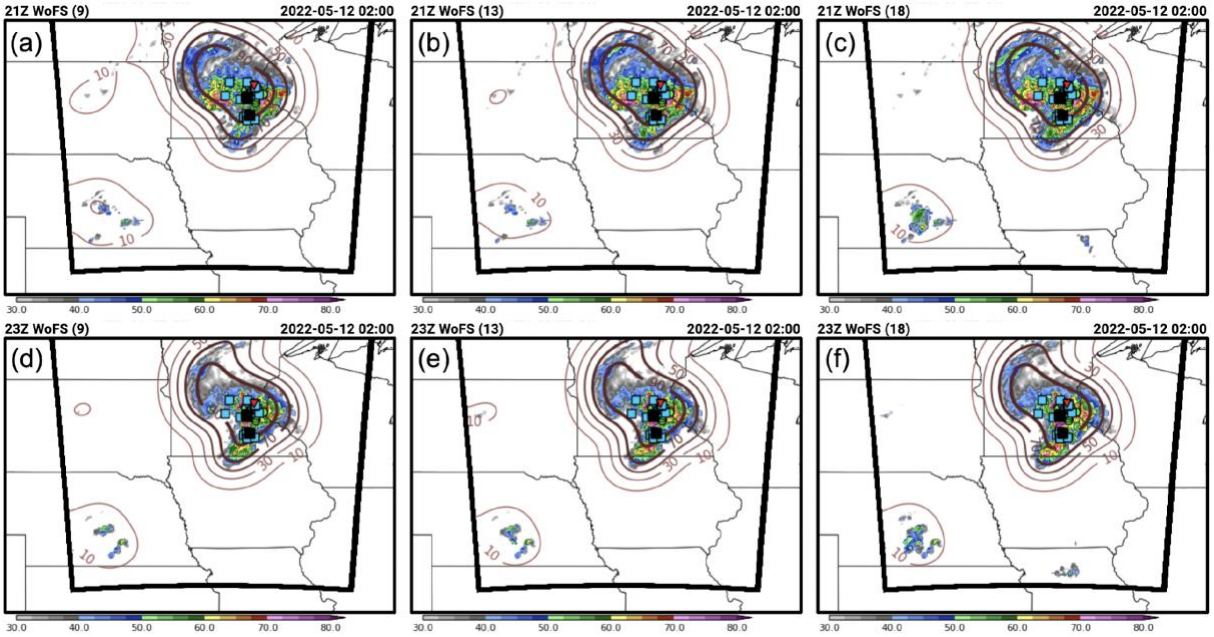


*Figure 36. Neighborhood maximum ensemble probabilities of hourly maximum 10-m wind speed ≥ 30 knots (contours) and the maximum from any member values of hourly maximum 10-m wind speed (shaded) from 2100 UTC WoFS initializations on 11 May 2022 with (a) 9, (b) 13, and (c) 18 members. (d) – (f) same as (a) – (c) except for 2300 UTC WoFS initializations. LSRs are overlaid in each panel.*

The results revealed that the 9, 13, and 18 member WoFS forecasts performed very similarly. Although there was a very slight increase in the mean subjective ratings as the number of members increased, none of these differences were statistically significant (Welch 2-sample t-test with $\alpha$ = 0.05). The overriding theme from the comments was how similar the forecasts were to one another, which likely reflects an under-dispersive ensemble.

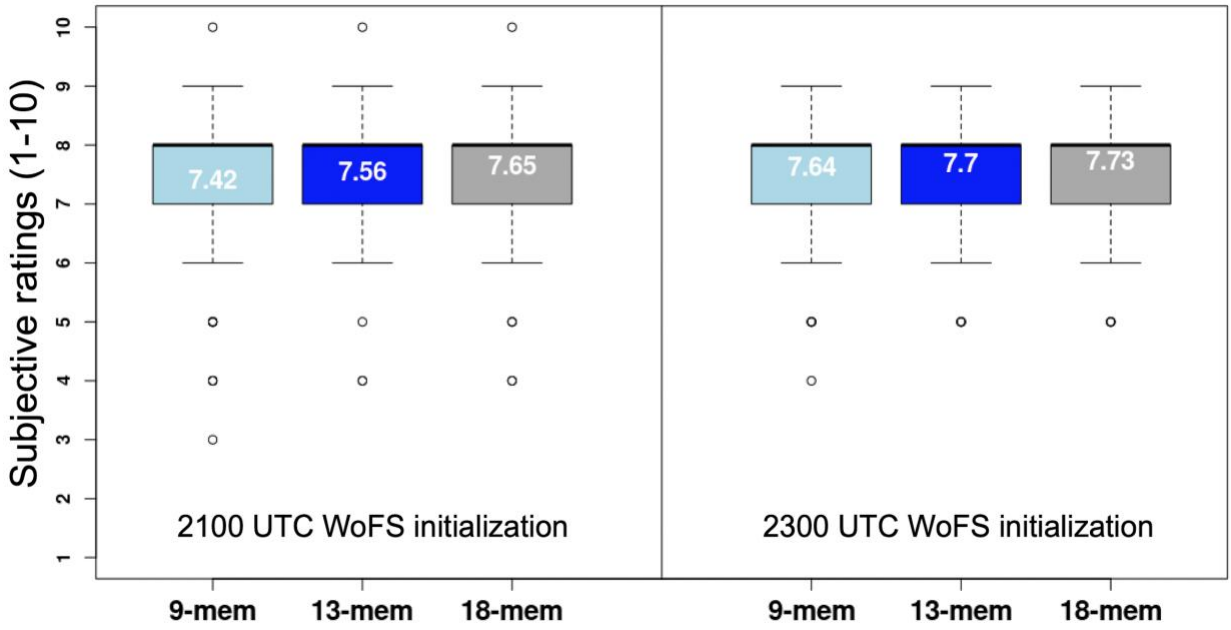


*Figure 37. Distributions of subjective ratings (1–10) by SFE participants for 2100 and 2300 UTC WoFS initializations where probabilities were derived from 9, 13, and 18 members.*

### 3.3.6 WoFS: Time Lagging

In this evaluation, 18-member WoFS guidance was compared where the 18 members came from a single initialization time versus different (i.e., time-lagged) initialization times. Specifically, one set of 18 members came from 6 members drawn from 1900, 2000, and 2100 UTC; another set of 18 members came from 9 members drawn from 2000 and 2100 UTC; and the final set of 18 members all came from 2100 UTC. The same comparisons were repeated, but for WoFS ensembles based at 2300 UTC. The primary science question was whether time-lagging could be a beneficial strategy for WoFS forecasts. For both the 2100 and 2300 UTC WoFS initializations participants were asked, "*To gauge the potential impact of time-lagging with the Warn-on-Forecast System (WoFS), probabilities and ensemble maxima for several storm attribute fields (updraft helicity, updraft speed, & 10-m wind gust) as well as reflectivity are computed from 3 different sets of WoFS members based at 2100 and 2300 UTC. The first set uses 6 members from t, t-1, and t-2, where t is initialization time. The second set uses 9 members from t and t-1, and the third uses 18 members from t. These configurations are referred to as WoFS (6/6/6), WoFS (9/9), and WoFS (18), respectively*." An example display is shown in Figure 38 and results are summarized in Figure 39.

The results revealed that WoFS (18) – i.e., the full WoFS ensemble initialized at 2100 and 2300 UTC – performed slightly better than the two time-lagged WoFS configurations. For 2100 UTC based forecasts, the differences in mean subjective ratings for both time-lagging configurations and WoFS (18) were statistically significant, but for 2300 UTC based forecasts, the differences were not quite significant (hypothesis tests

used Welch 2-sample t-test with $\alpha$ = 0.05). General themes from the participant comments were that differences were often not noticeable or very subtle. At other times, participants noted that the non-time-lagged ensemble was able to better "hone in" on events.

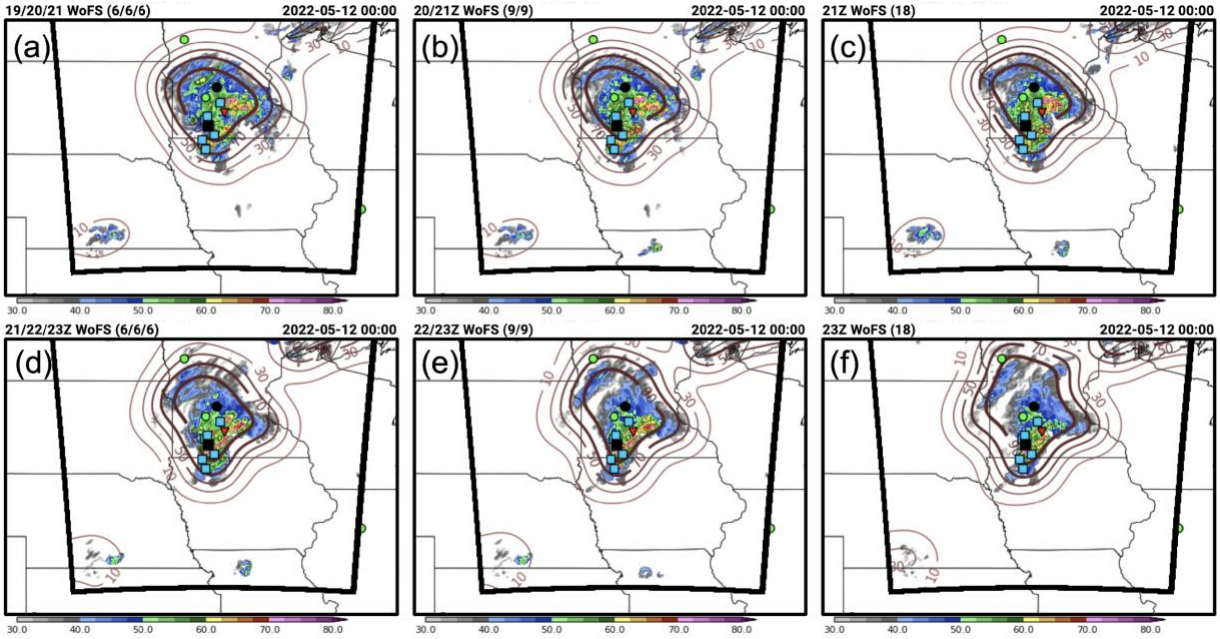


*Figure 38. Neighborhood maximum ensemble probabilities of hourly maximum 10-m wind speed ≥ 30 knots (contours) and the maximum from any member values of hourly maximum 10-m wind speed (shaded) from WoFS initializations based at 2100 UTC 11 May 2022 for (a) WoFS (6/6/6), (b) WoFS (9/9) and (c) WoFS (18). (d) – (f) same as (a) – (c) except for WoFS initializations based at 2300 UTC. LSRs are overlaid in each panel.*

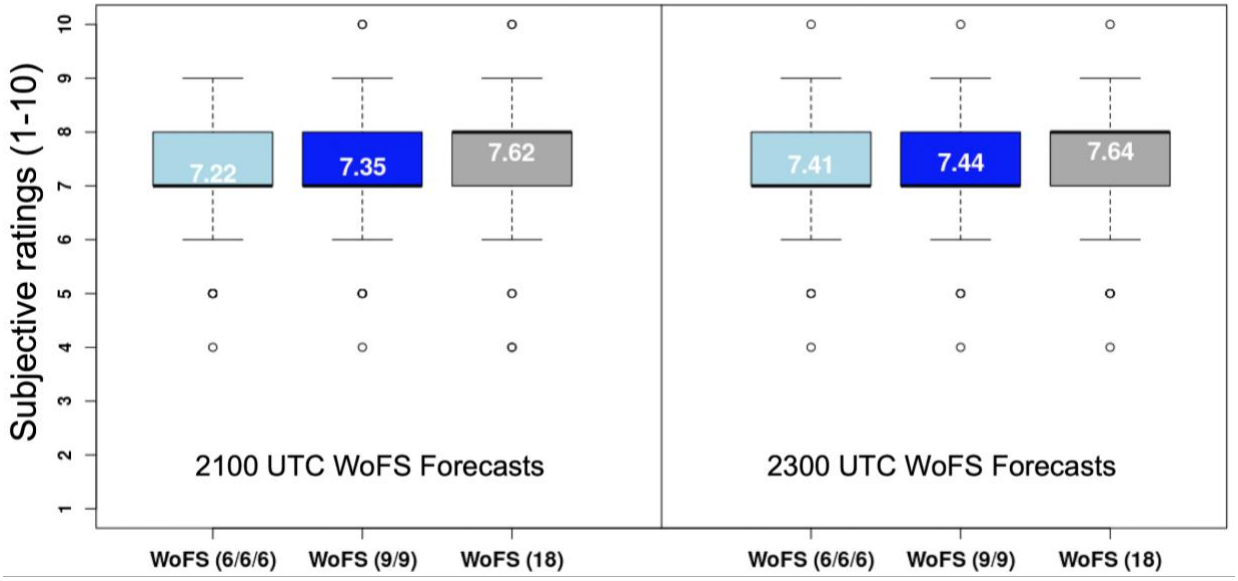


*Figure 39. Distributions of subjective ratings (1-10) by SFE participants for time-lagged WoFS forecasts based at 2100 and 2300 UTC where probabilities were derived from WoFS (6/6/6), WoFS (9/9), and WoFS (18) (see text for these definitions).*

## 3.4 Model Evaluations – Group D: Medley

### 3.4.1 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). Two training approaches were utilized: one using wind damage reports that had a measured wind value and one with an additional dataset of sub-severe thunderstorm wind measurements. For both of these approaches, output from two different algorithms were presented. One was an ensemble model [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 four machine learning models were available for yesterday's preliminary wind reports for evaluation on an interactive webpage developed by Iowa State University (ISU; Fig. 40).

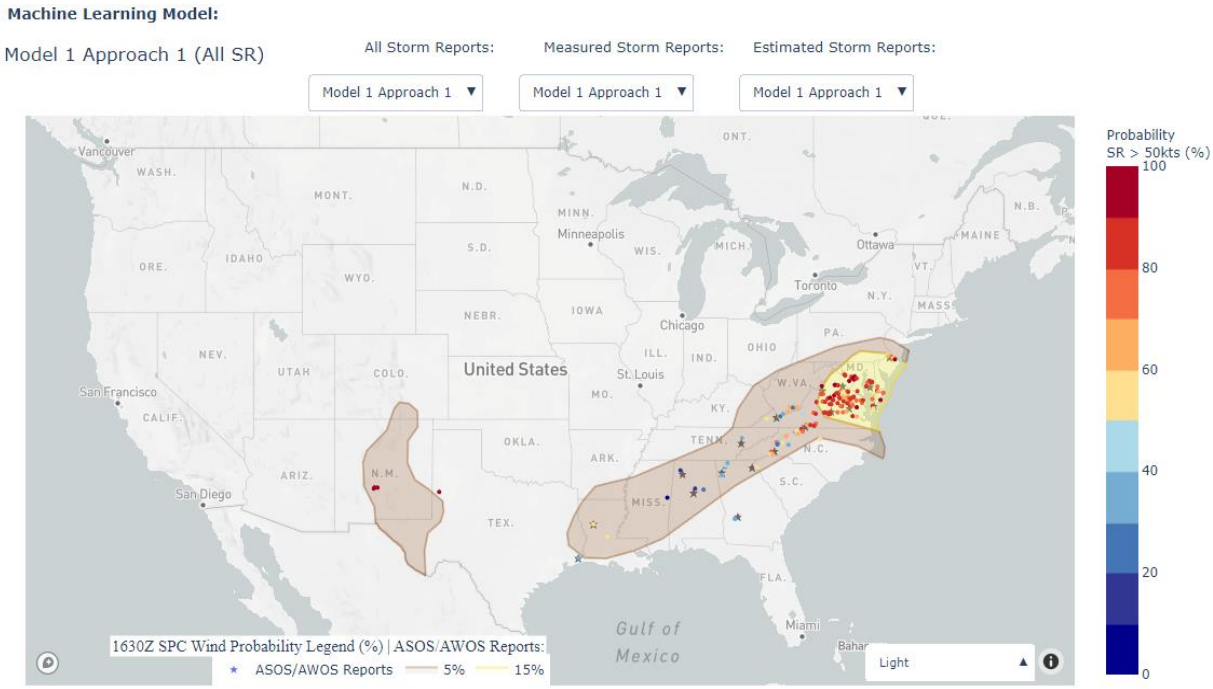


*Figure 40. Example of interactive webpage for the D1. ISU Machine-Learning Severe Wind Probability evaluation during the 2022 SFE. The preliminary wind reports are shaded with the probability that the report was associated with a wind gust of ≥50 knots from the various ML algorithms. The user has the option to zoom/road, 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; with 10 being best) 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., 5–8 out of 10) regardless of model or approach (Fig. 41). The primary findings from the subjective ratings include 1) the ML models that were trained with the additional database of sub-severe thunderstorm wind gusts generally received higher ratings than those models trained only with measured wind report, and 2) the impact of the ML model was relatively small in the subjective ratings with a very slight edge in the mean ratings to the ensemble approach (i.e., stack GLM). The SFE participants commonly noted that using the sub-severe thunderstorm wind gusts in the training often led to higher ML probabilities of severe winds, especially for wind damage reports in the eastern CONUS. For the first time, the ISU ML output was also used to calculate a practically perfect hindcast for severe wind (not shown). Overall, the participants commented that weighting the wind reports using the ML output was preferred over treating all wind reports equally in the practically perfect hindcast.

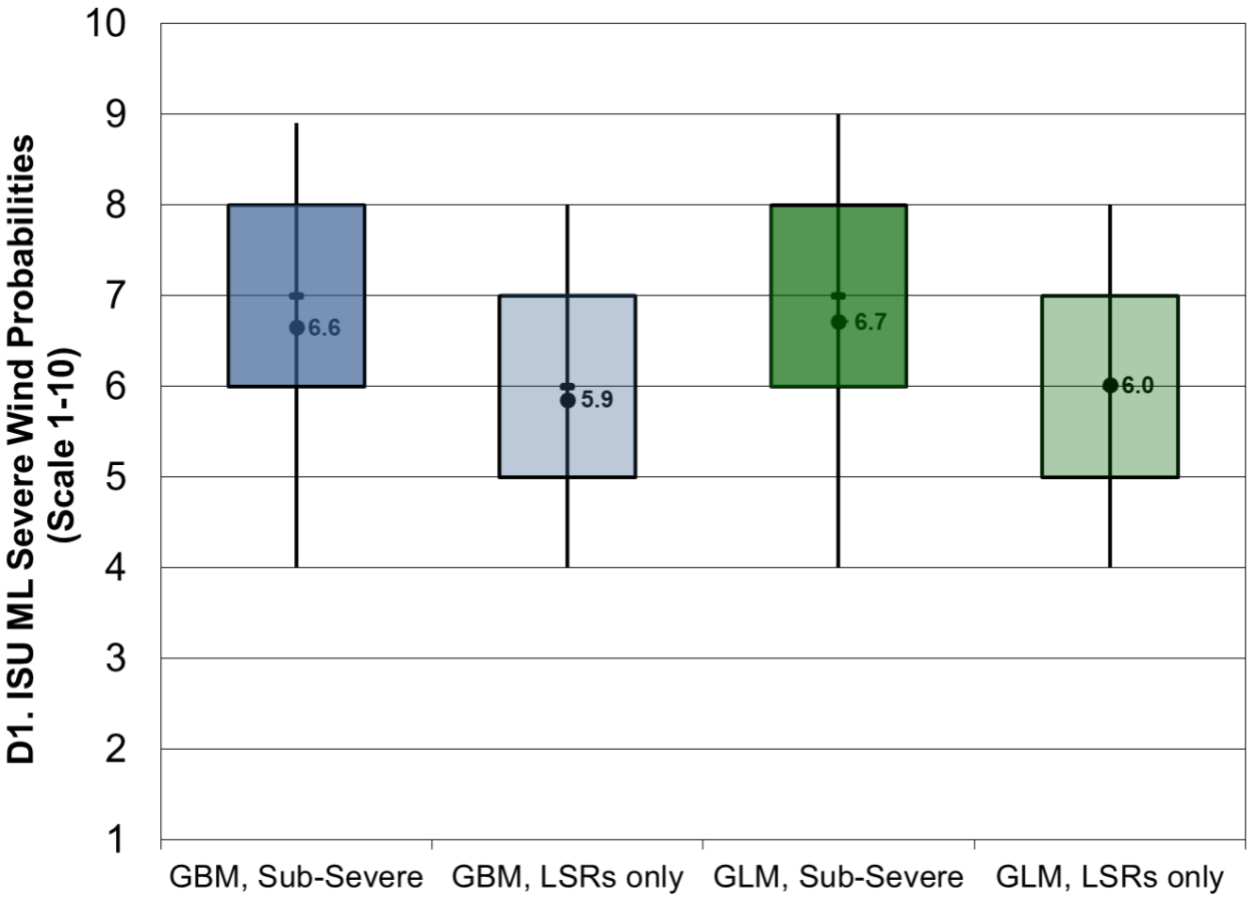


*Figure 41. Distributions of subjective ratings (1–10) by SFE participants of the ISU ML severe wind probabilities for preliminary wind reports for two models (GBM - blue; stack GLM - green) and two different approaches (trained with LSRs and sub-severe thunderstorm gusts – darkest shade; trained with only reports – lightest shade).*

### 3.4.2 NCAR ML Convective Mode Probabilities

Machine-learning algorithms were trained to provide probabilistic guidance of simulated storm mode using output from the HRRRv4. Specifically, three 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, 2) a partially-supervised CNN system, that is trained using a "proxy" field related to convective mode (i.e., object size and updraft helicity) and clustered using a Gaussian mixture model (GMM), and 3) a new deep neural network (DNN) that predicts mode based on a set of convective storm properties, such as size, area, updraft helicity, reflectivity, etc. All three systems output probabilistic predictions of supercells, quasi-linear convective systems (QLCSs), and disorganized modes for storm objects from the 00 UTC-initialized HRRR. The three ML models applied to the HRRR were evaluated based on the subjective impressions of the participants on estimating convective mode using an interactive website developed by NCAR (Fig. 42).

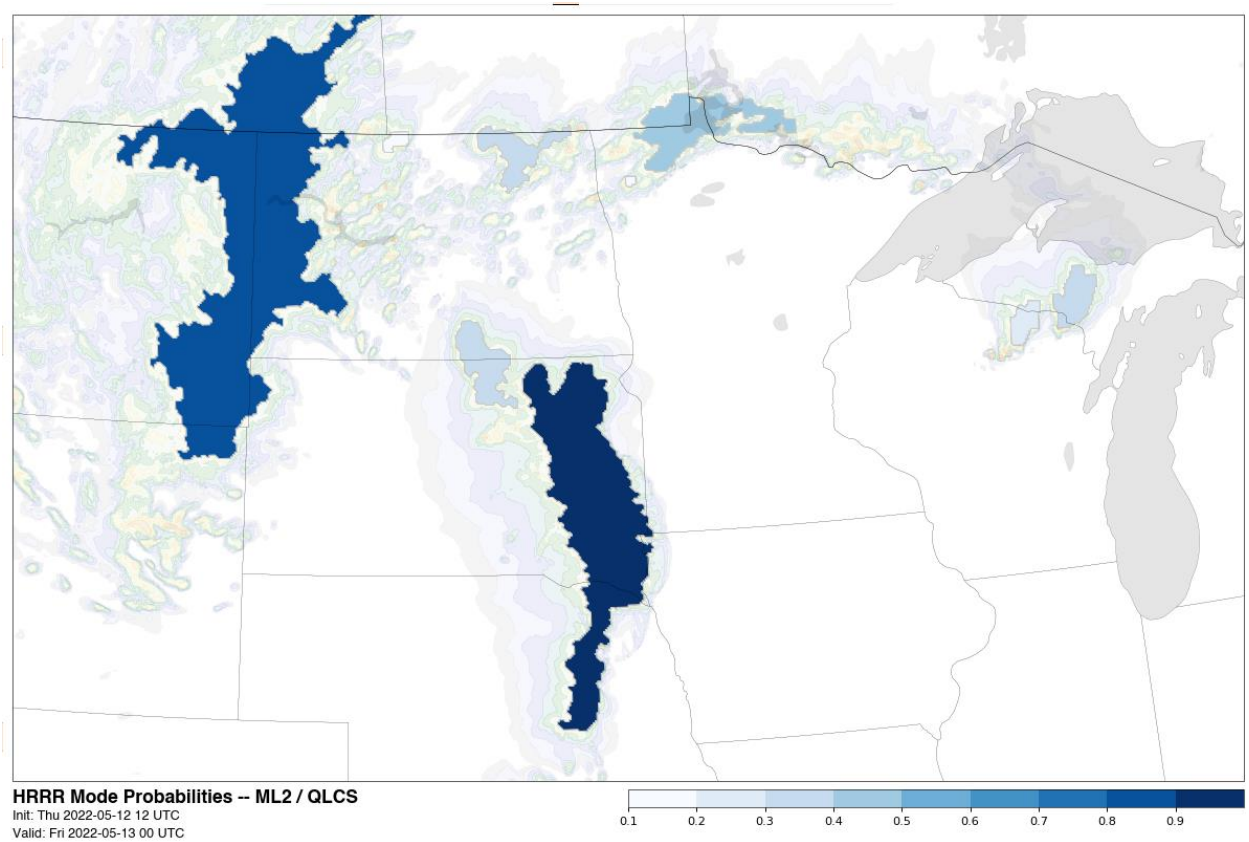


*Figure 42. Example of interactive webpage for the D2. NCAR Machine-Learning Convective Mode Probability evaluation during the 2022 SFE. Storm objects from the CAMs are shaded with the probability of being a supercell, QLCS (shown here in blue shades), or disorganized convective mode with composite reflectivity lightly shaded in the background.*

Participants were asked to evaluate (on a scale of 1 to 5; with 5 being best) 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 reveals a very similar distribution of ratings for the different ML algorithms (Fig. 43). This does not necessarily indicate that the output was similar day-to-day, but it does indicate that there was not a favored algorithm over all of the cases during the SFE. This is a positive result for the partially supervised GMM algorithm because the ratings were improved from last year and extensive hand labels are not required for training. A convective mode neighborhood probability product (not shown) generated from the time-lagged HRRR runs was also subjectively evaluated. Overall, the feedback from this product was somewhat mixed, but the participants commented about the potential utility, especially for summarizing the evolution of convective mode through the forecast period.

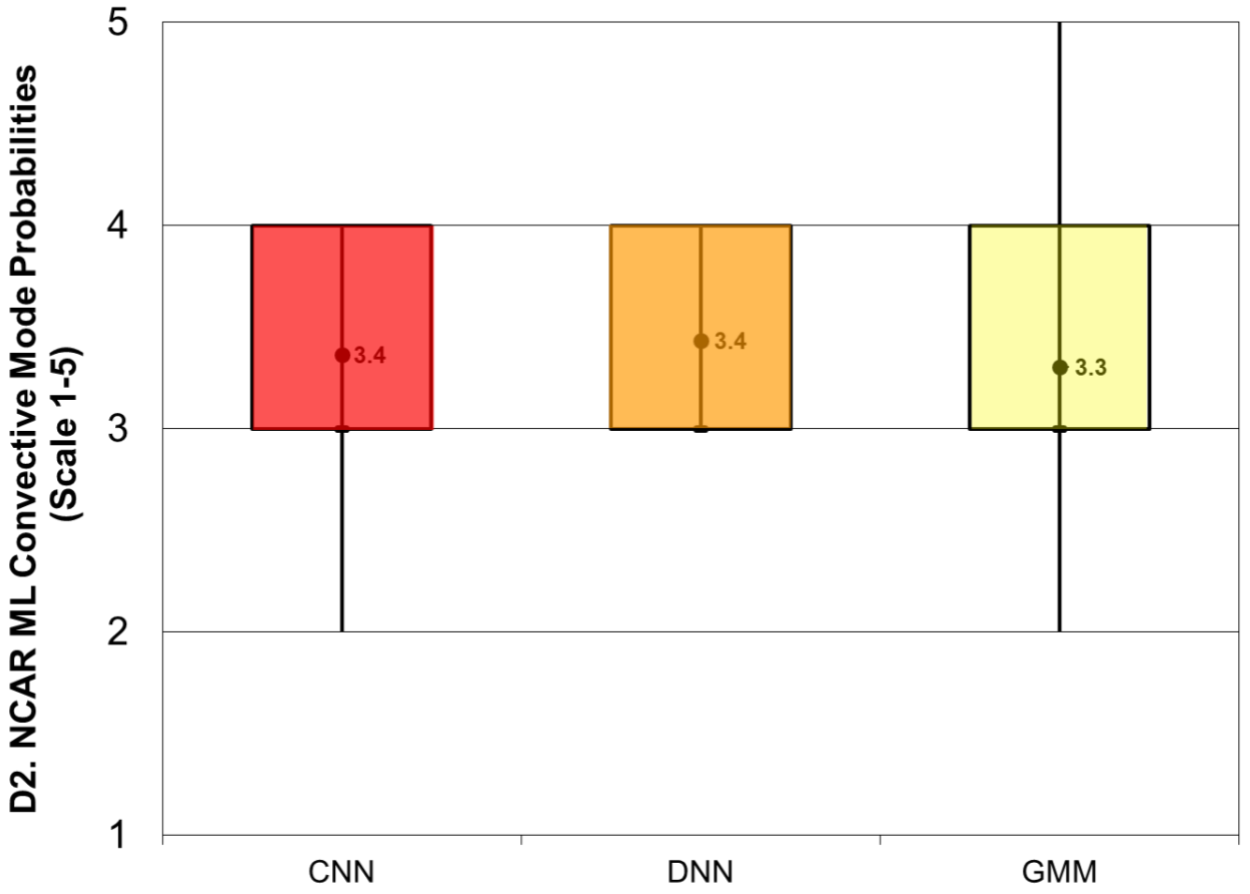


*Figure 43. Distributions of subjective ratings (1–5; where 5 is best) by SFE participants of the NCAR ML convective mode probabilities for storm objects from the HRRR and three ML algorithms (supervised CNN - red; deep neural network (DNN) - orange, and partially supervised GMM – yellow).*

### 3.4.3 Mesoscale Analysis Background

Three hourly versions of 3D-RTMA with different backgrounds were subjectively evaluated by participants during the 2022 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. Prototype 1 (i.e., 3D-RTMAp1) used the FV3-based RRFSp1 as the first-guess background information and the GDAS for background error covariances in the hybrid DA system, prototype 2 (i.e., 3D-RTMAp2) used the RRFSp2 as the first-guess background information and its own ensemble for background error covariances, and the HRRR baseline version used the operational HRRR for first-guess background and the GDAS for background error covariances. The hourly analyses for 2-m temperature, dewpoint, and SB/ML/MUCAPE were examined during the 1800–0300 UTC period on the following day (Fig. 44). Post-processing issues with the effective-layer STP and SRH precluded the evaluation of those fields.

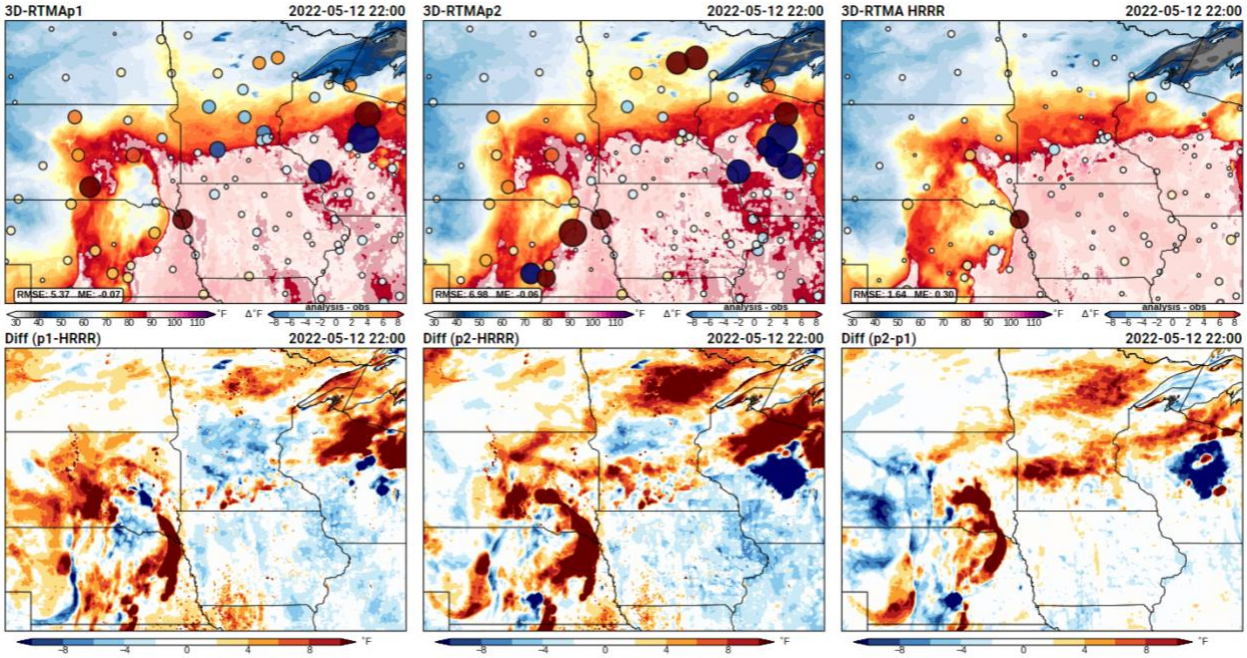


*Figure 44. Example of the website comparison page for the 3D-RTMA during the 2022 HWT SFE. The 3D-RTMAp1 is shown in the upper-left panel, the 3D-RTMAp2 is in the upper-middle panel, and the 3D-RTMA HRRR baseline is show in the upper-right panel. The difference plots are shown in the bottom row: 3D-RTMAp1 - 3D-RTMA HRRR (lower left), 3D-RTMAp2 - 3D-RTMA HRRR (bottom middle), and 3D-RTMAp2 - 3D-RTMAp1 (bottom right). The 2-m temperature analysis valid at 2200 UTC on 12 May 2022 is shaded in the upper row. The difference (analysis-obs) at METAR sites is shown by the size and shading of the dots. The corresponding 2-m temperature analysis differences are shaded in the bottom row.*

The goal of this evaluation was to assess the impact of the first-guess background on analyses for short-term severe weather forecasting applications. Overall, both of the FV3-based versions of the 3D-RTMA (i.e., p1 and p2) were rated subjectively "slightly worse" to "about the same" as the HRRR-based version (Fig. 45). Specifically,

participants commonly noted issues in the FV3-based versions in the composite reflectivity field for a high bias in both convective and stratiform regions. The FV3-based versions also tended to display a moist bias in the 2-m dewpoint field, though this moist bias was reduced from last year. Differences in the 2-m temperature field were most commonly associated with effects from convection. In general, the HRRR-based version handled the effects of convection on 2-m temperature better than the FV3-based versions through more accurate representation of the size, shape, and magnitude of cold pools and thunderstorm outflows. While the CAPE fields and soundings were also examined during the SFE, there did not appear to be any systematic biases or preferred analysis versions that stood out consistently across the domain and from day-to-day. The two FV3-based versions were also compared to one another and were more similar to each other than to the HRRR-based version with a median rating of "about the same". There were not any systematic or consistent differences that stood out during the subjective evaluation, but both versions occasionally displayed some erroneous "dry spikes" aloft when looking at the forecast soundings. This behavior was too inconsistent and isolated to identify the cause.

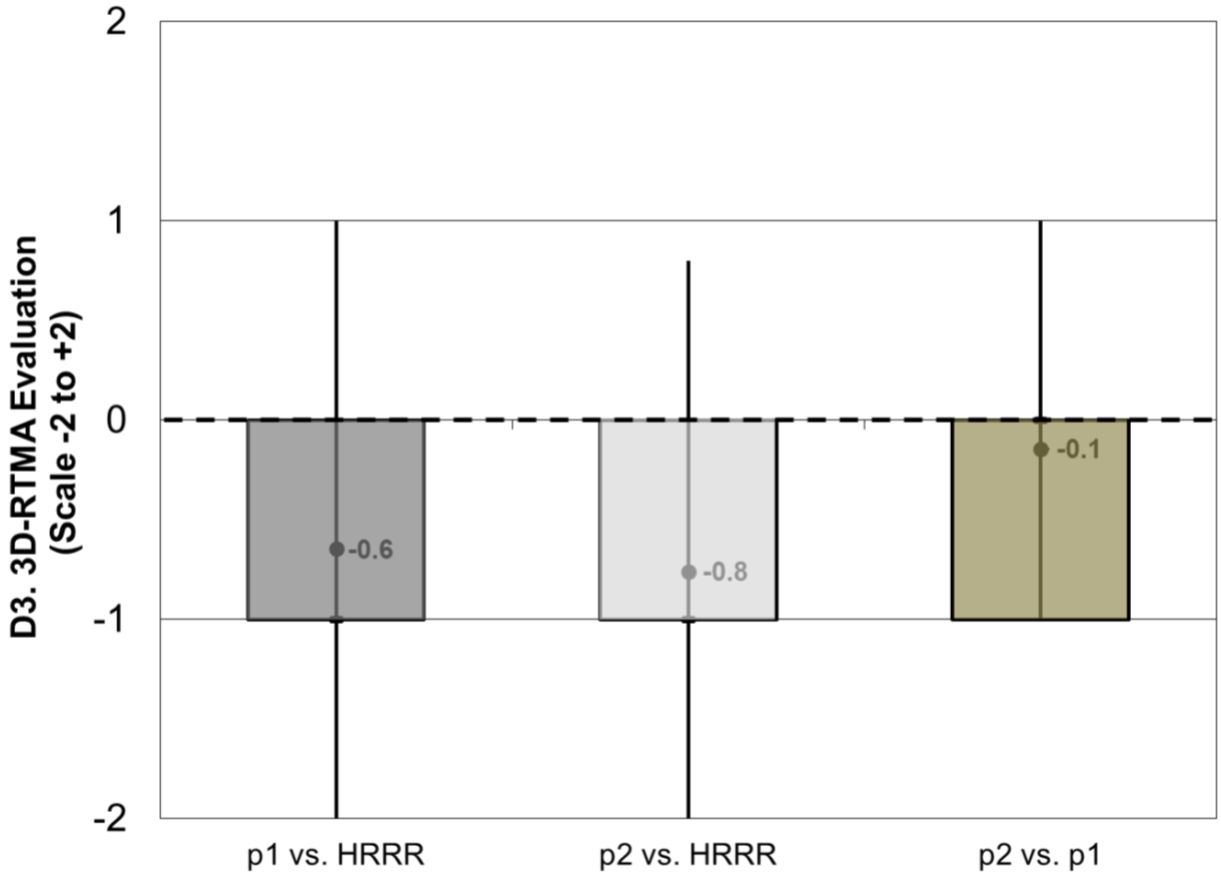


*Figure 45. Distributions of subjective ratings (-2 to +2) by SFE participants of the 3D-RTMAp1 compared to the 3D-RTMA HRRR (dark gray), 3D-RTMAp2 compared to the 3D-RTMA HRRR (light gray), and 3D-RTMAp2 compared to 3D-RTMAp1 (gold). The ratings represent how the analyses compared to one another from -2: Much Worse; -1: Slightly Worse; 0 - About the Same; +1: Slightly Better; +2: Much Better.*

### 3.4.4 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 from comparison with preliminary local storm reports, especially measured gusts (Fig. 46).

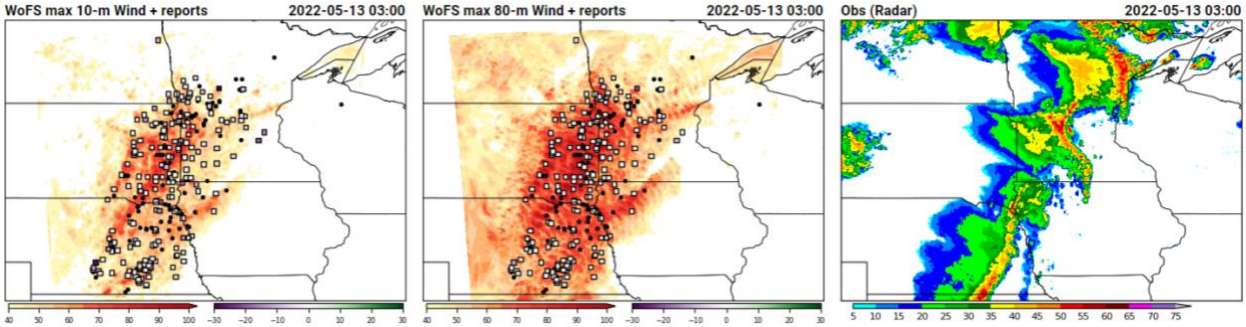


*Figure 46. Example of the website comparison page for the WoFS analyses during the 2022 HWT SFE. The 12 May 1800–0300 UTC accumulated ensemble maximum 10-m wind is shown in the left panel, the ensemble maximum 80-m wind in the middle panel, and the observed composite reflectivity in the right panel. 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 for diagnosing 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. Overall, the 80-m winds received higher subjective ratings than the 10-m winds (Fig. 47) 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). One aspect that stood out more this year than last year was that spurious convective gusts were occasionally present in the 80-m wind field even where convection did not form in reality. This appeared to often be a result of an outlier member near the domain boundary, so efforts to use observed reflectivity to filter out spurious members will be employed in the future. Overall, the participants found this to be an interesting and promising approach and use of a rapidly cycling convection-allowing ensemble system.

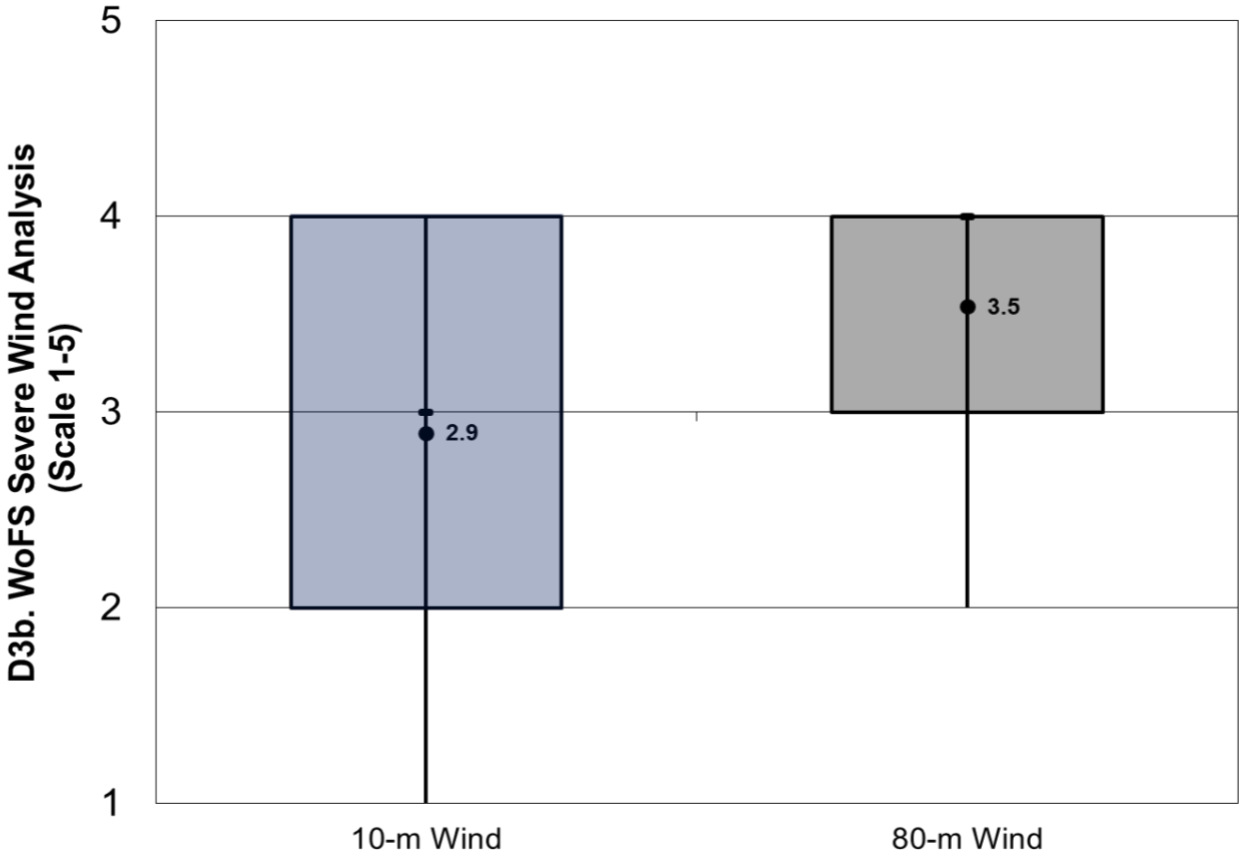


*Figure 47. Distributions of subjective ratings (1–5) by SFE participants of the WoFS storm-scale severe wind analysis for ensemble maximum 10-m winds (blue) and 80-m winds (gray), where the ratings represent how well the WoFS maximum wind analyses align with the preliminary severe wind reports and overall assessment of severe winds: 1 - Very Poorly; 2 - Poorly; 3 - Neutral, neither poorly nor well; 4 - Well; 5 - Very Well.*

### 3.4.5 Significant Severe Winds

In an effort to assess significant wind (i.e., 65+ knots) potential in CAMs from mesoscale convective systems (MCSs), a couple of variables were added to the NSSL-WRF for evaluation. The hourly maximum variables include the maximum wind in the 0–2 km AGL layer and the integrated wind in the 0–2 km AGL layer. Operational experience has highlighted that 3-km CAMs often develop intense rear-inflow jets in well-organized MCSs that can only be visualized currently in forecast soundings. The hypothesis is that these new diagnostic variables may be able to more readily highlight significant winds in simulated MCSs. This evaluation compared standard 10-m wind output with the new wind variables. Unfortunately, there was only one day during the 2022 HWT SFE with a notable MCS that produced significant severe winds (i.e., 12 May; Fig. 48), so these products were only evaluated for that single day. The maximum 0–2 km AGL wind highlighted a large swath of 65+ knot winds that verified well and even highlighted potential for 85+ knot winds, which were observed. Meanwhile, the 10-m wind output only highlighted a small area with 50+ knot winds. Thus, the utility and potential of this

new output was confirmed for this notable event, but should be evaluated over a variety of other events to learn more about its characteristics.

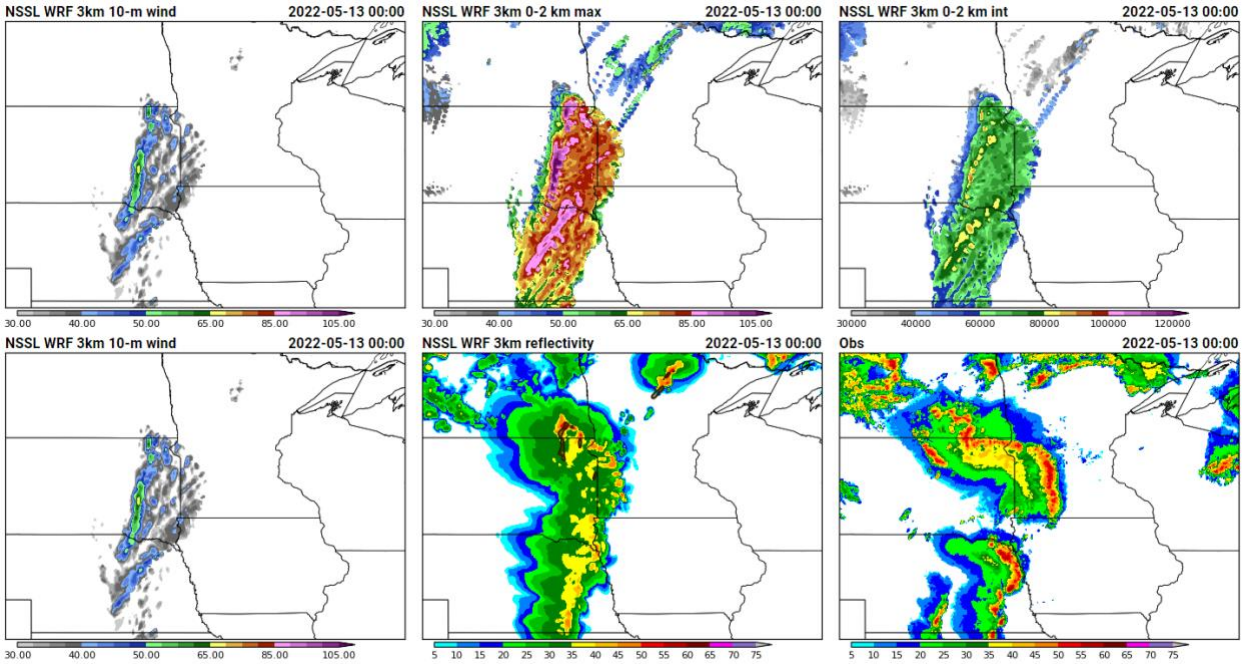


*Figure 48. Example of the website comparison page for the significant severe wind evaluation during the 2022 HWT SFE. The 24-h forecast from the 0000 UTC NSSL-WRF is shown for 12 May for 10-m wind (upper left and lower left), maximum 0–2 km AGL wind (upper middle), integrated 0–2 km AGL wind (upper right), simulated reflectivity (lower middle), and observed reflectivity (lower right).*

### 3.4.6 County-Based Watch Guidance

An HREF-based machine learning (ML) model was developed to produce automated, non-static watch products that dynamically track with the predicted severe weather threat. This guidance was derived from a gradient boosted classifier trained on HREF ensemble updraft helicity, updraft vertical velocity, 10-m wind, and sfc-500 mb shear. Estimated watch counties were inferred at each 1200 UTC HREF forecast hour from the ML probabilistic output and masked such that a county must fall within at least a 1300 UTC D1 Slight risk to qualify for a watch. Automated watches produced by the ML guidance were designed to provide at least 3 hours of lead time prior to the first occurrence of severe weather. An alternative automated watch product was also derived from the SPC Severe Timing guidance. Estimated watch counties were inferred at each forecast hour from the temporally disaggregated individual hazard probabilities provided by the 1300 UTC Severe Timing guidance and the 1200 UTC HREF. A county was considered to be in a watch at a given forecast hour if the timing guidance produced individual hazard probabilities equivalent to at least a Slight risk at that location and time. As with the ML guidance, these automated watches were designed to provide at least 3 hours of lead time prior to the first occurrence of severe weather.

The first-guess county-based watch products were presented to SFE participants via an interactive webpage with three graphic panels as shown in Figure 49. Hourly forecasts from the SPC Severe Timing Guidance were displayed in the left-most panel, the ML guidance forecasts were presented in the middle panel, and the "observed" SPC-issued Severe Thunderstorm and Tornado Watches were provided in the right-most panel. All evaluations of the ML and Severe Timing Guidance first-guess watches were performed for the previous day's severe weather. As part of the evaluation, participants were asked to complete a survey consisting of five questions. The first two questions captured metadata such as the respondents' unique participant number and the date being evaluated. Question 3 (Q3) asked respondents to subjectively rate how similar the placement and timing of the ML and Severe Timing Guidance watch products were to the operational Tornado and Severe Thunderstorm Watches issued by the NWS. Each product was assessed independently on a 5-point Likert scale with values ranging from "*Not at all similar*" to "*Extremely similar*." Respondents were instructed to consider the full 16-hour forecast period when determining their responses, and an option of "*N/A*" was provided if there were no operational watches issued for the event. Similarly, Q4 directed participants to subjectively evaluate how well the ML and Severe Timing Guidance watch products captured the location and timing of the severe weather threat during the available 16-hour forecast period. Again, the ML and non-ML products were independently assessed via a 5-point Likert scale ranging from "Terrible" to "Excellent." This evaluation was performed using preliminary local storm report (LSR) and NWS storm-based warning overlays to represent the observed location and time of severe weather occurrence. Additionally, respondents were instructed to only consider reports and warnings that fell within at least a 1300 UTC D1 SLGT to avoid penalizing the forecast products for not capturing severe hazards in locations where the guidance was systematically precluded from issuing forecasts. Finally, Q5 provided an open response field for participants to describe their thoughts about the performance of the guidance for the day.

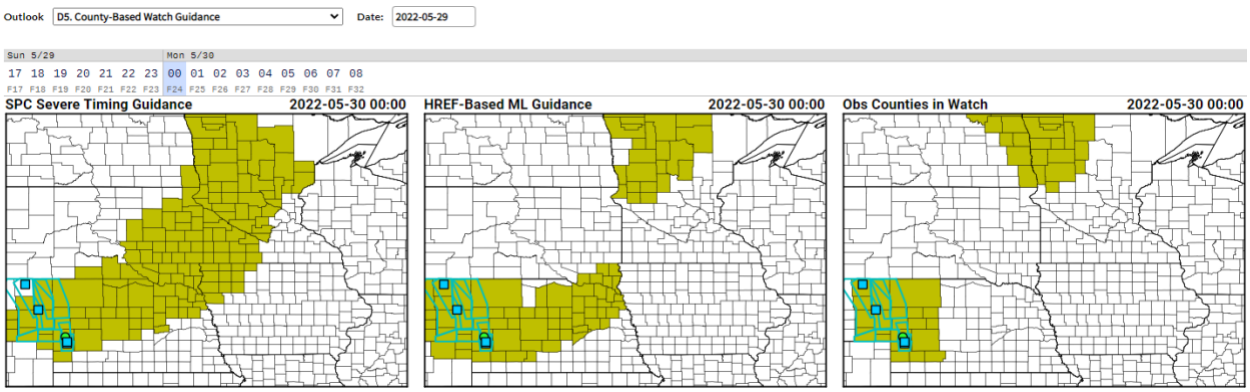


*Figure 49. Web display presented to 2022 SFE participants while evaluating the performance of the 1200 UTC HREF-based ML and 1300 UTC SPC Severe Timing Guidance first-guess watch products.*

Respondents were neutral on average when rating how similar the ML and Severe Timing Guidance first-guess watch products were to the SPC-issued Tornado and Severe

Thunderstorm watches (Fig. 50a). The ML guidance received a bootstrapped mean score of 3.13 with a standard deviation of 0.82. Similarly, the non-ML guidance was given a mean rating of 2.93 with a standard deviation of 0.93. Differences between the two products were small and ultimately not statistically significant at the 95% confidence level; however, the distribution of survey responses does at least indicate a slight trend in favor of the ML-derived first-guess watch products. Approximately 77% of survey responses indicated the ML guidance was at least "*moderately*" similar to the SPC watches, and 36% of responses found it to be "*very*" or "*extremely*" similar. Conversely, the Severe Timing Guidance was at least "*moderately*" similar in 67% of responses and "*very*" or "*extremely*" similar in only 28% of the results.

Respondents generally rated the ML and Severe Timing Guidance first-guess watch products more favorably in regard to how well they captured the spatial and temporal domains of the true severe weather hazards, with bootstrapped mean scores of 3.58 and 3.37 respectively (Fig. 50b). Additionally, respondent agreement was nearly identical for both products as indicated by a standard deviation of 0.82 for the ML and 0.81 for the non-ML products. As before, these minute differences between the product ratings were not found to be statistically significant at the 95% confidence level, but the response distribution of the ML guidance again trended towards somewhat higher ratings than that of the Severe Timing Guidance. About 86% of responses stated that the ML first-guess watches captured the timing and spatial coverage of the observed NWS warnings and LSRs with at least "*average*" skill, and 61% said the model performance was "*good*" or "*excellent*." In comparison, the Severe Timing Guidance performance was rated as "*average*" or better in 83% of responses and "*good*" or "*excellent*" in 48% of the results.

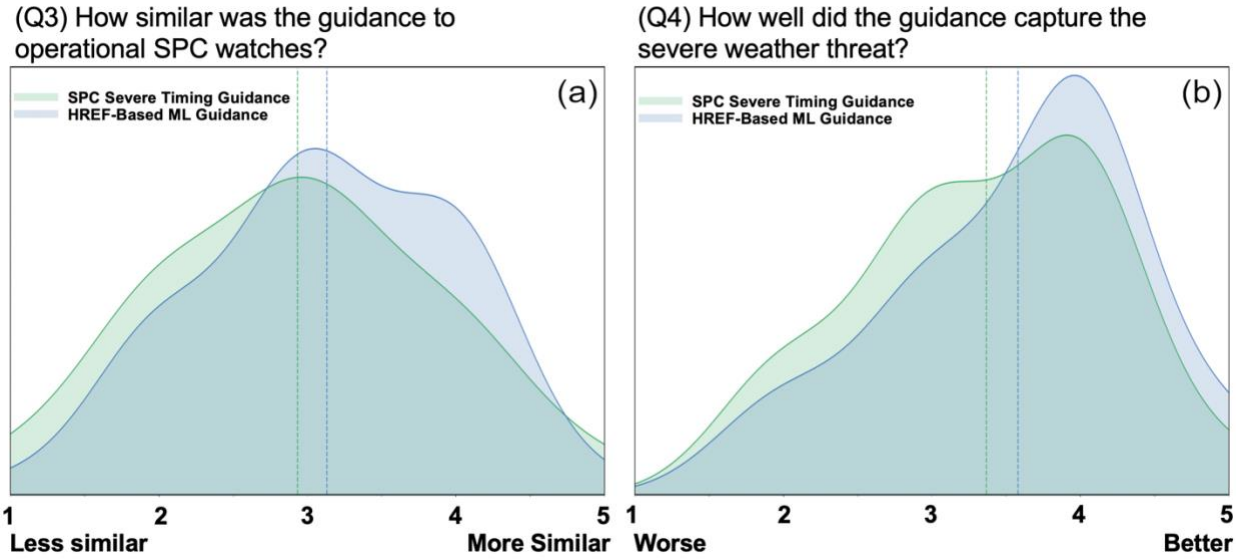


*Figure 50. (a) Survey Q3 and (b) Q4 responses approximated as KDE curves. Dashed vertical lines represent the mean score for each guidance product.*

### 3.4.7 GEFS vs. SREF – Severe Weather Forecasting

As part of the 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 in the Day 2 to Day 3 time-frame, an evaluation was performed during the 2022 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 2 and Day 3 periods using a multi-panel webpage with the SPC mesoanalysis as the verification standard (Fig. 51).

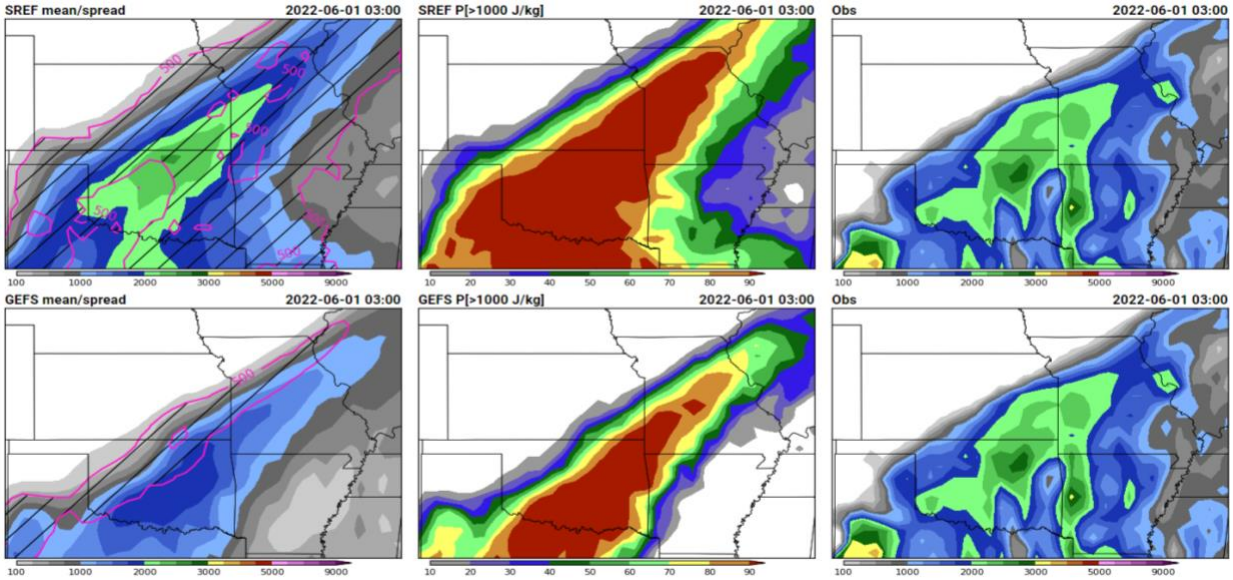


*Figure 51. Example of the website comparison page for the GEFS comparison to the SREF during the 2022 HWT SFE. The SREF forecasts are shown in the top row with the GEFS forecasts in the bottom row. The Day 3 forecasts of MLCAPE mean/spread (left column) and probability of exceeding 1000 J/kg (middle column) are shown for comparison with the SPC Mesoanalysis (right column) as the "observation" valid at 0300 UTC on 1 June 2022.*

For the Day 3 environment forecasts, the GEFS severe weather fields were subjectively rated similar to SREF overall (Fig. 52). There are some days/locations where the GEFS does better than the SREF and vice versa for the environmental fields, with the median and mean ratings centered on "about the same". Some of the common concerns with the GEFS environment forecasts noted by participants were the low instability bias and the overconfident solution at times. The Day 3 calibrated guidance offers a different perspective on the GEFS performance relative to the SREF (Fig. 53). The GEFS calibrated thunder and severe guidance was more frequently rated better than the SREF than for the environmental fields. In fact, the median rating for the GEFS calibrated fields was "slightly better" than the SREF calibrated fields. 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 one year of data.

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. 54), with the GEFS forecasts slightly favored for 2-m dewpoint and the SREF forecasts slightly favored for MLCAPE. The Day 2 calibrated guidance evaluation reveals slightly higher mean subjective ratings for the GEFS calibrated thunder and severe guidance compared to the SREF (Fig. 55); however, the Day 2 forecast improvement is less than that seen in the Day 3 forecasts, as indicated by a median rating of "about the same" between the calibrated guidance products.

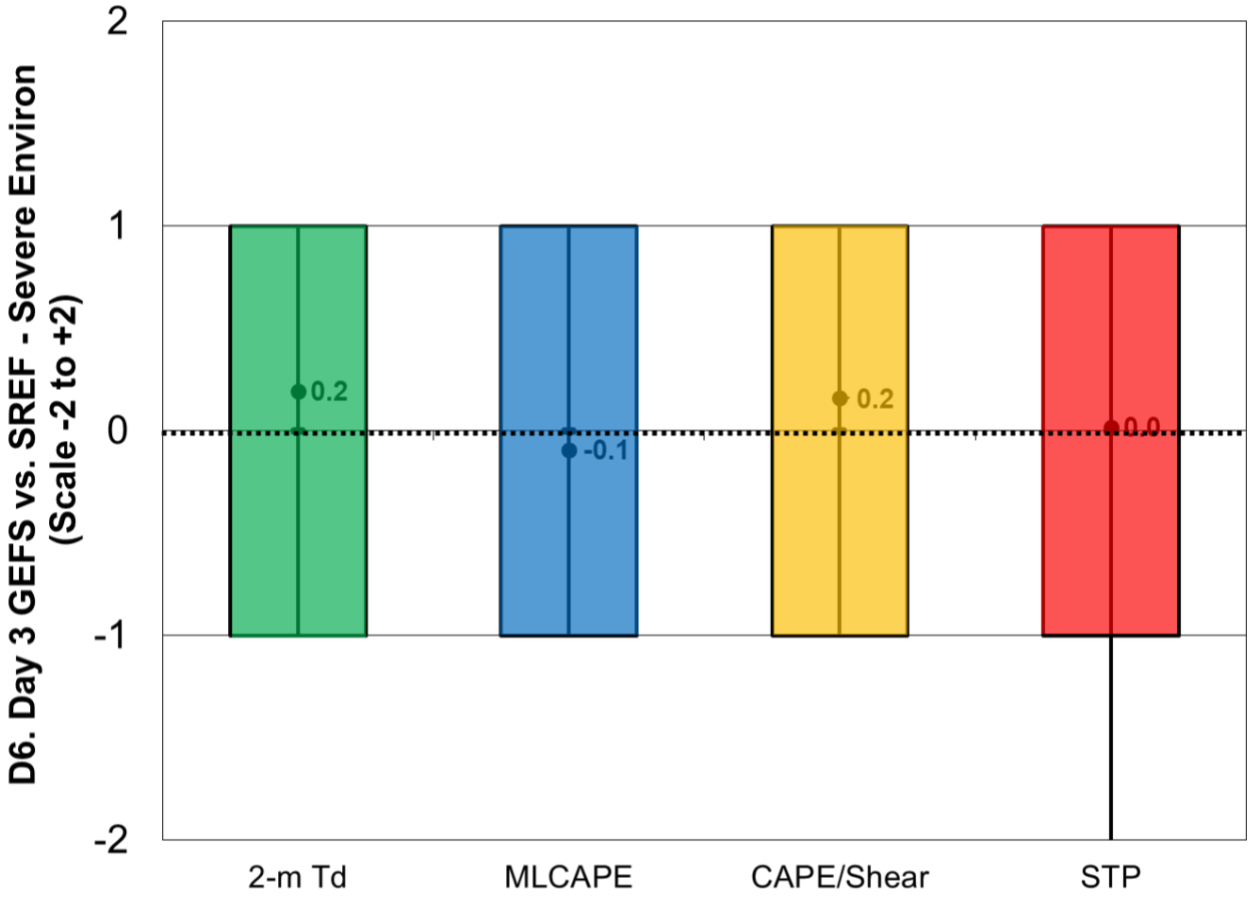


*Figure 52. Distributions of Day 3 subjective ratings (-2 to +2) by SFE participants of the GEFS environment forecasts compared to the SREF forecasts for ensemble fields of 2-m dewpoint (green), MLCAPE (blue), CAPE & shear (orange), STP (red). The ratings represent how the GEFS compared to the SREF for these environment fields from -2: Much Worse; -1: Slightly Worse; 0 - About the Same; +1: Slightly Better; +2: Much Better.*

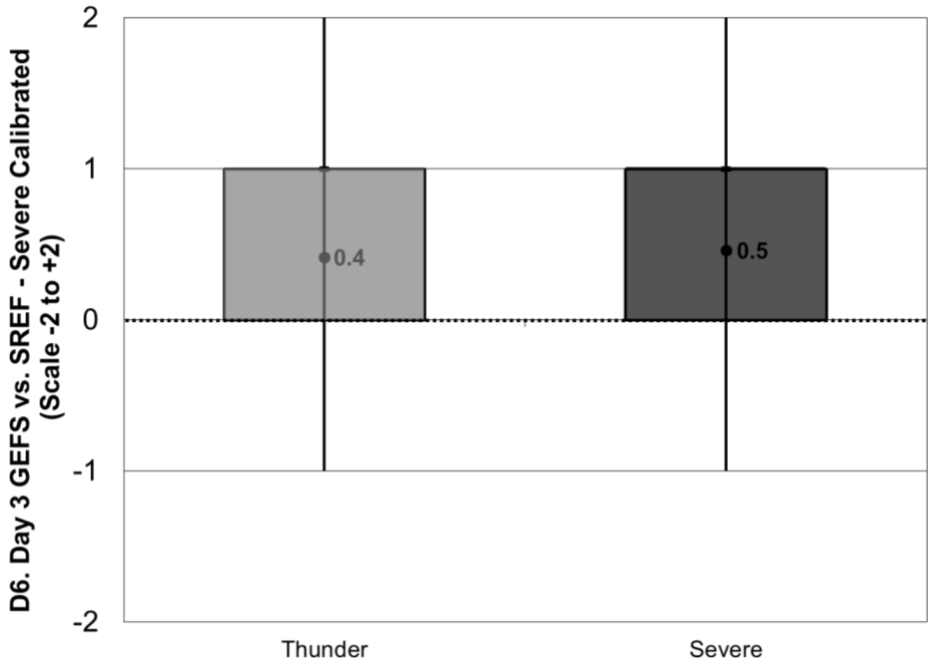


*Figure 53. Distributions of Day 3 subjective ratings (-2 to +2) by SFE participants of the GEFS calibrated forecasts compared to the SREF calibrated forecasts for thunder (light gray) and severe (dark gray). The ratings represent how the GEFS calibrated guidance compared to the SREF calibrated guidance from -2: Much Worse; -1: Slightly Worse; 0 - About the Same; +1: Slightly Better; +2: Much Better.*

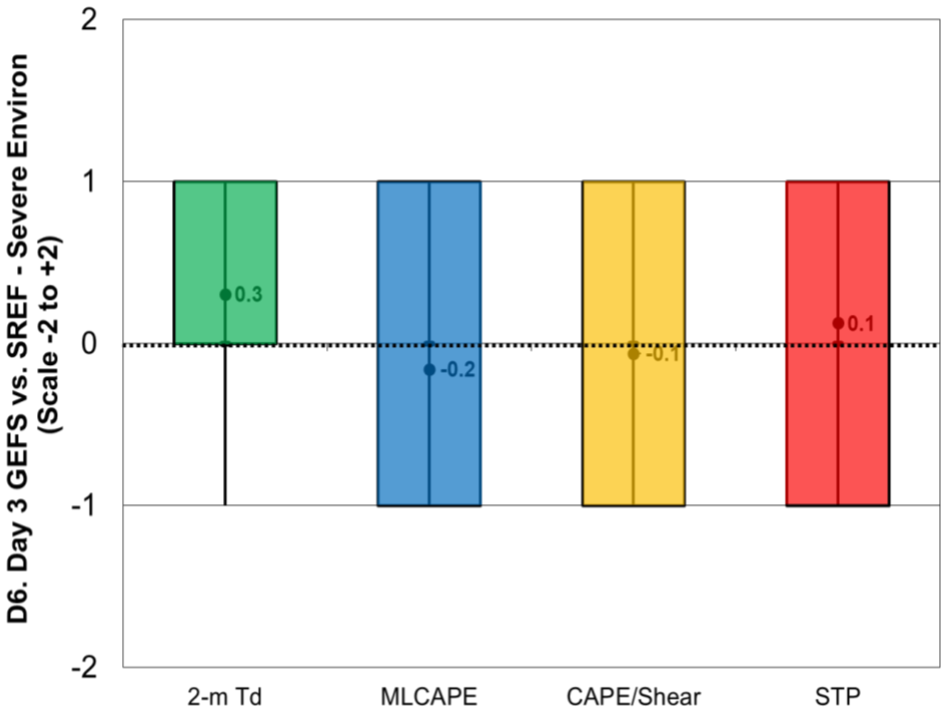


*Figure 54. Distributions of Day 2 subjective ratings (-2 to +2) by SFE participants of the GEFS environment forecasts compared to the SREF forecasts for ensemble fields of 2-m dewpoint (green), MLCAPE (blue), CAPE & shear (orange), STP (red). The ratings represent how the GEFS compared to the SREF for these environment fields from -2: Much Worse; -1: Slightly Worse; 0 - About the Same; +1: Slightly Better; +2: Much Better.*

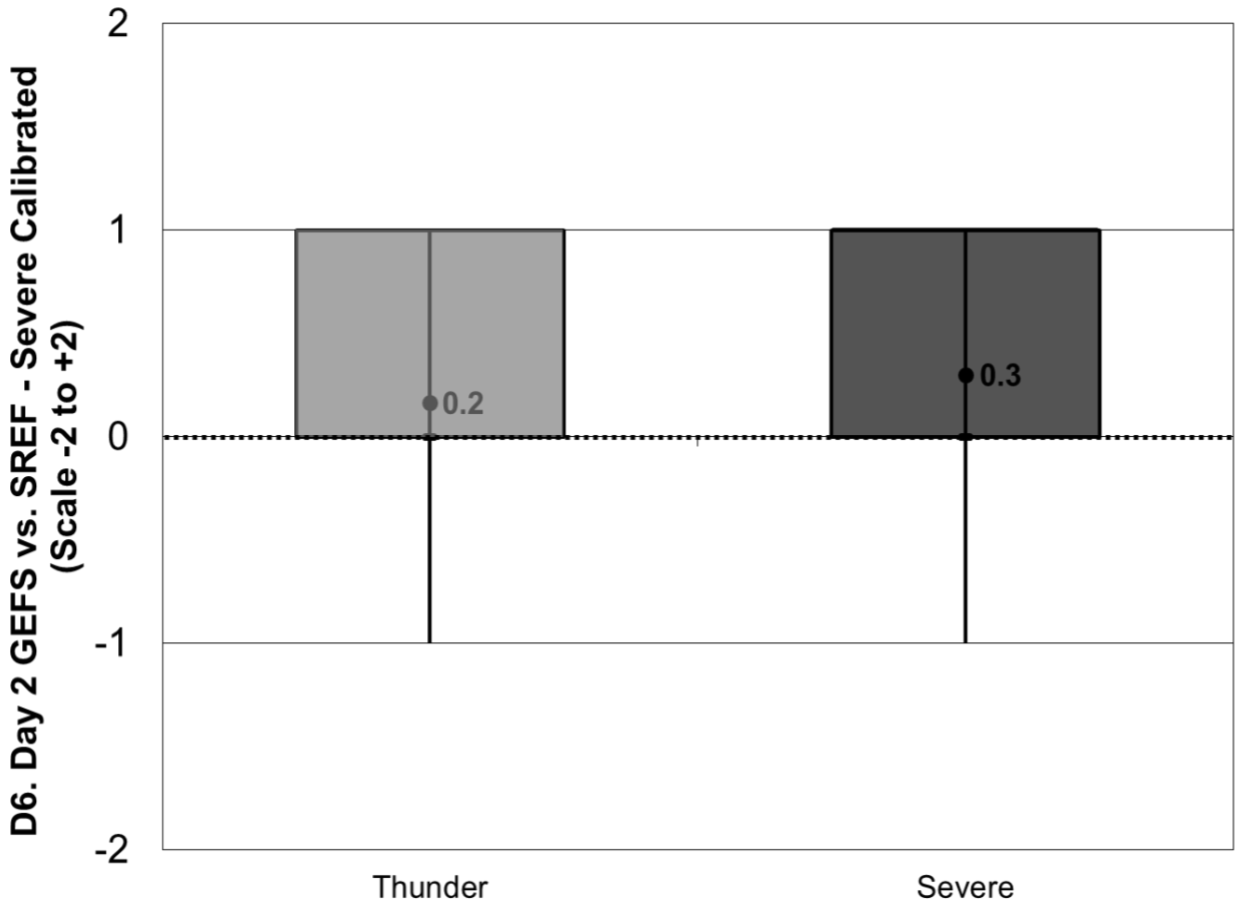


*Figure 55. Distributions of Day 2 subjective ratings (-2 to +2) by SFE participants of the GEFS calibrated forecasts compared to the SREF calibrated forecasts for thunder (light gray) and severe (dark gray). The ratings represent how the GEFS calibrated guidance compared to the SREF calibrated guidance from -2: Much Worse; -1: Slightly Worse; 0 - About the Same; +1: Slightly Better; +2: Much Better.*

## 3.5 Evaluation of Experimental Forecast Products

### 3.5.1 Days 1, 2, & 3 Hazards Coverage & Conditional Intensity Forecasts

Each morning, participants contributed to a group outlook activity. These outlooks were led by pairs of facilitators, who went over experimental guidance covering the period of interest and then drew coverage probabilities and conditional intensity forecasts for tornadoes, wind, and hail separately. An hour was allotted for these activities. One group drew forecasts for the Day 3 time period, one group covered the Day 2 period, and two groups covered the Day 1 period. Of the Day 1 groups, one used experimental calibrated guidance to draw its forecasts, while the other did not use the experimental calibrated guidance. The next day, all participants provided subjective ratings of all four group forecasts for each hazard. Coverage and conditional intensity forecasts were evaluated separately. Conditional intensity forecasts were easier to evaluate for wind and hail forecasts relative to tornado forecasts, since significant wind and hail reports are more likely to be available as next-day observations compared to significant tornado reports,

as tornado ratings are assigned following NWS damage surveys. Due to the nature of the Day 2 and Day 3 forecasts and evaluation activities (e.g., Day 3 and Day 2 forecasts were not issued over the weekend, and Friday forecasts were not subjectively verified the next day), Day 3 forecasts were only available to subjectively evaluate on Thursday and Friday (forecasts valid Wednesday and Thursday), while Day 2 forecasts were available to evaluate on Wednesday–Friday (forecasts valid Tuesday–Thursday).

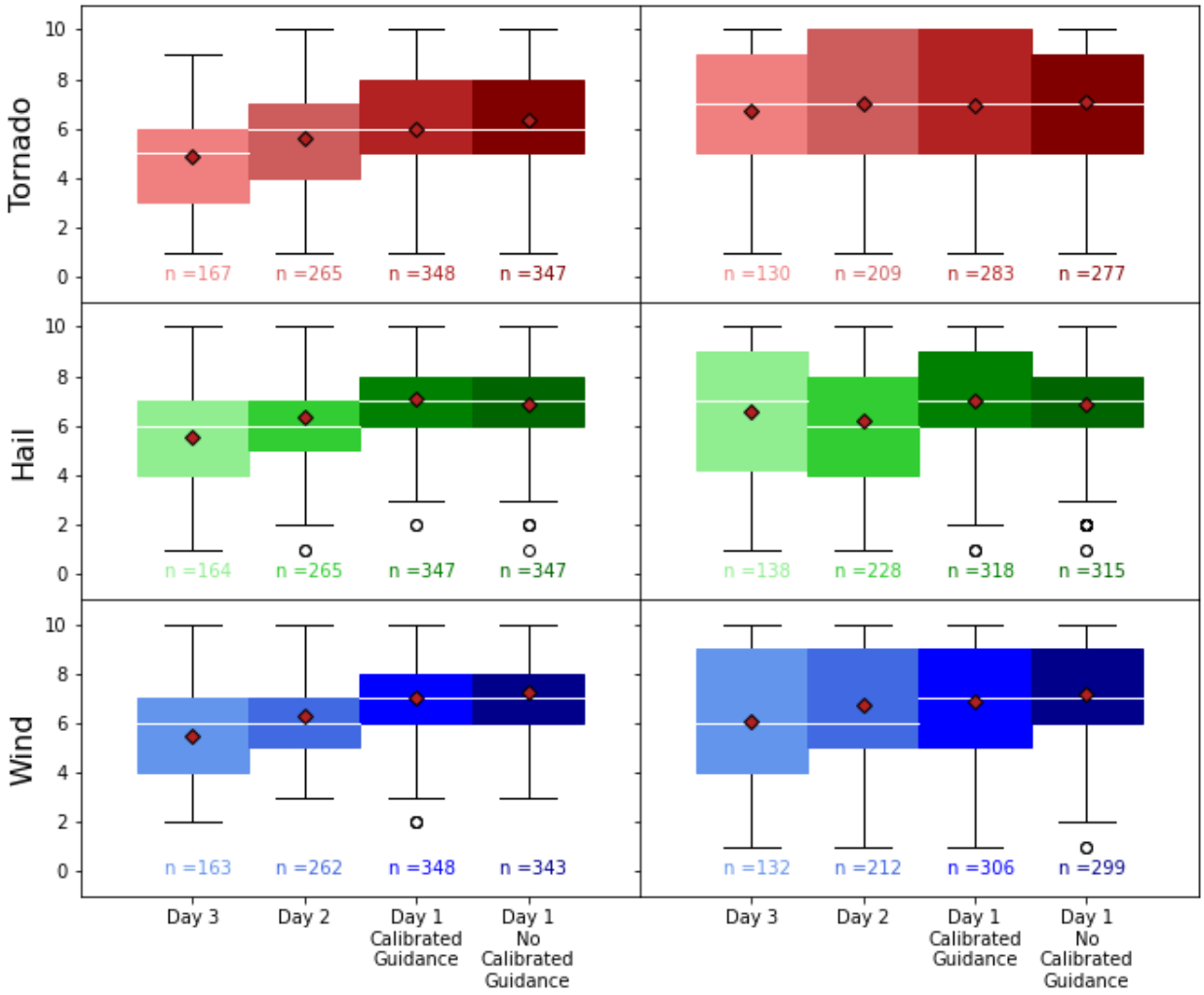


*Figure 56. Participant subjective ratings of tornado (top row), hail (middle row), and wind (bottom row) forecasts of coverage probabilities (left column) and conditional intensity forecasts (right column). Sample size for each distribution is annotated below the box of interest.*

Subjective scores for each hazard generally increased as the lead time decreased, particularly for the coverage probabilities (Fig. 56), with the Day 3 probabilities scoring the lowest and the Day 1 groups scoring the highest. The Calibrated Guidance and No Calibrated Guidance groups performed similarly for all hazards except tornado, where the No Calibrated Guidance group scores were significantly higher for the tornado coverage probabilities at the $p<.05$ level (Table 4). Mann-Whitney U Rank tests were used to

determine statistical significance due to their independence from assumptions about the underlying distributions. Trends in the conditional intensity guidance differed between hazards, unlike the general improvement seen in the coverage probabilities. The wind hazard showed improvements in the forecasts with shorter lead times, but the hail conditional intensity forecasts showed a decrease in the skill of the Day 2 forecasts relative to the Day 3 forecasts, before the scores increased again for the two Day 1 groups. Finally, the tornado conditional intensity guidance was very similar between groups, likely due to the lag in significant tornado information and the relatively limited tornado season CONUS-wide during spring 2022.

All differences between the tornado coverage probabilities were significant at least at the $p<.05$ level, while none of the differences between the tornado conditional intensity ratings were significant (Table 3). The difference in hail and wind coverage probability ratings were also typically statistically significant, with the exception of the comparison between the two Day 1 forecasting groups. For all hazards evaluated, the differences in the conditional intensity forecasts were less likely to be statistically significant than coverage forecasts.

| | Tornado Coverage | Tornado Conditional Intensity | Hail Coverage | Hail Conditional Intensity | Wind Coverage | Wind Conditional Intensity |
|---|---|---|---|---|---|---|
| Day 3/Day 2 | **.0005** | .1495 | **$2.20\times10^{-5}$** | .1063 | **$8.32\times10^{-6}$** | *.025* |
| Day 2/Day 1 Calibrated Guidance | **.0025** | .4390 | **$2.55\times10^{-8}$** | **$7.93\times10^{-5}$** | **$1.08\times10^{-11}$** | .116 |
| Day 2/Day 1 No Calibrated Guidance | **$1.02\times10^{-7}$** | .3879 | **$1.68\times10^{-5}$** | **.0006** | **$4.33\times10^{-16}$** | **.0095** |
| Day 1 Calibrated Guidance/No Calibrated Guidance | *.0101* | .3411 | *.0372* | .1340 | .219 | .2078 |
| Day 3/Day 1 Calibrated Guidance | **$7.53\times10^{-9}$** | .1575 | **$7.27\times10^{-18}$** | *.0213* | **$5.32\times10^{-18}$** | **.0027** |
| Day 3/Day 1 No Calibrated Guidance | **$2.41\times10^{-14}$** | .0812 | **$1.75\times10^{-14}$** | .0722 | **$3.23\times10^{-22}$** | **.0001** |

*Table 3. p-values from the Mann-Whitney significance test between subjective ratings of different outlook combinations. Green boxes with bold text show statistically significant values at p<.01, orange boxes show statistically insignificant differences, and yellow boxes with italicized text show differences that are significant at the p<.05 level but not at the p<.01 level.*

Participants were also asked about the relative importance of the coverage vs. the conditional intensity information if they were describing the forecast to someone for the previous day's forecast, and slid a sliderbar between two extremes to reflect their views. The slider was initially positioned in the middle, to indicate that Intensity and Coverage were Equally important. Generally, participants weighted the coverage information higher in importance than the conditional intensity information (Fig. 57), with some even stating that only coverage of the severe hazard was important. Future work will break down these responses by case, as in cases when no significant severe weather was anticipated the conditional intensity forecasts may be judged as less important by participants.

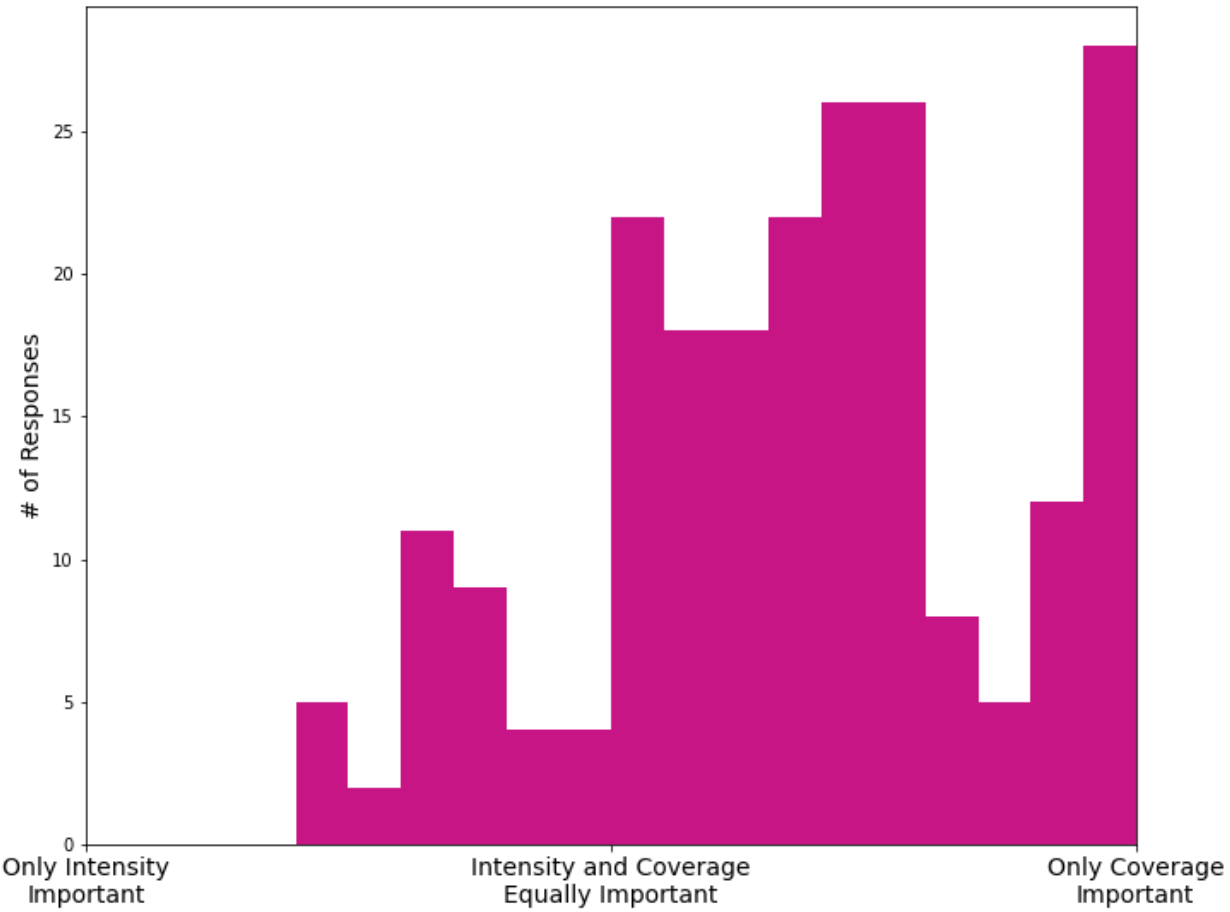


*Figure 57. Participant responses to the question, "If you were explaining the forecast to someone, how would you weigh the importance of coverage vs. intensity for yesterday's forecast?"*

### 3.5.2 WoFS-Focused Mesoscale Discussion Activity

As part of the afternoon forecasting activities on the R2O Desk, experimental mesoscale discussions (MDs) were generated during the 2022 HWT SFE. These MDs were generated daily in Google Slides (example provided in Fig. 58) by all R2O Group participants from 2:15–3:00 pm 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 MD forecasting activity was commonly cited by participants as their favorite activity. Participants noted that these forecasting 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.

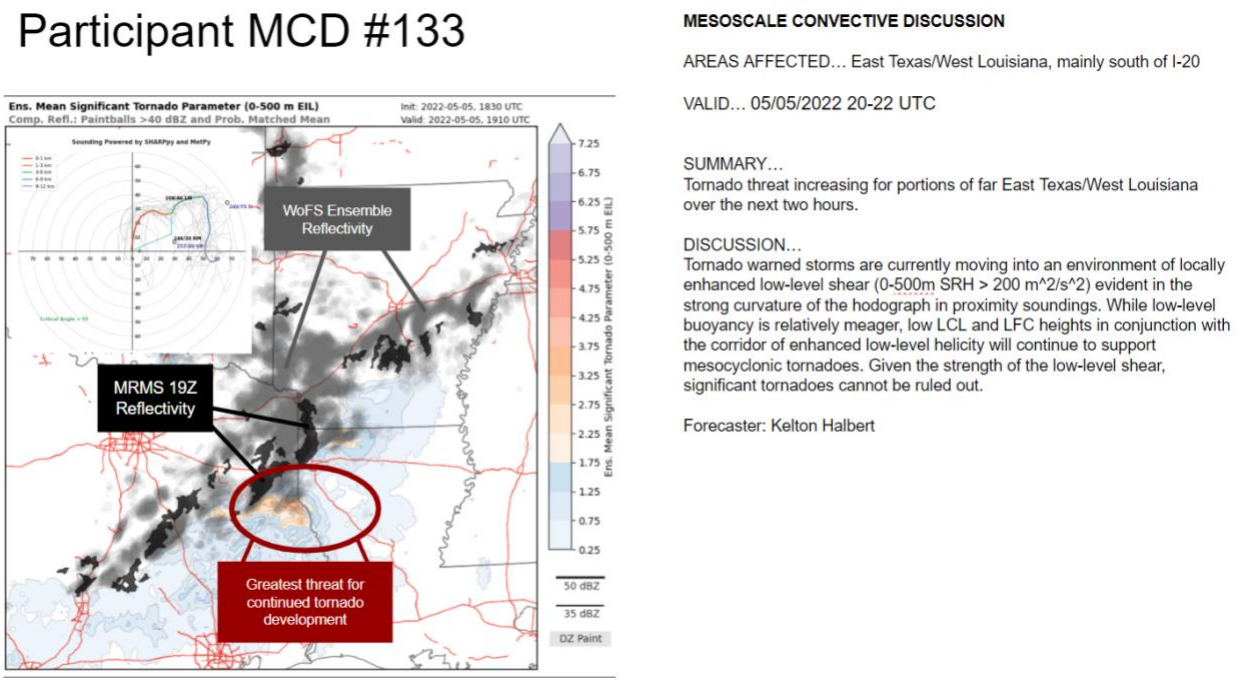


*Figure 58. Example of an experimental MD created on 5 May 2022 using WoFS output.*

### 3.5.3 Day 1 Outlook Updates Using WoFS

In another afternoon activity in the R2O group, slightly less than half of participants had the opportunity to update the Day 1 forecast using the observations and updated model guidance, including guidance from WoFS. Participants could utilize the morning forecasts from either Day 1 group (calibrated guidance or no calibrated guidance), or start drawing their outlooks from scratch. As with the morning Day 1 forecasts, participants were asked to issue coverage probability and conditional intensity forecasts of tornadoes, hail, and wind. They had approximately one hour to complete these forecasts. Two operational NWS forecasters were assigned individual usernames and had their forecasts displayed as issued, while all other participants had their forecasts grouped into consensus forecasts. Consensus coverage probabilities were calculated by simply taking the average of all participant forecasts. Consensus conditional intensity contours were issued by determining if at least 50% of participants issued a conditional intensity contour

at a given point. During the next-day subjective evaluation, participants evaluated the forecasts that they contributed to; so expert operational forecasters evaluated their own individual forecasts, while the other participants evaluated the consensus outlooks.

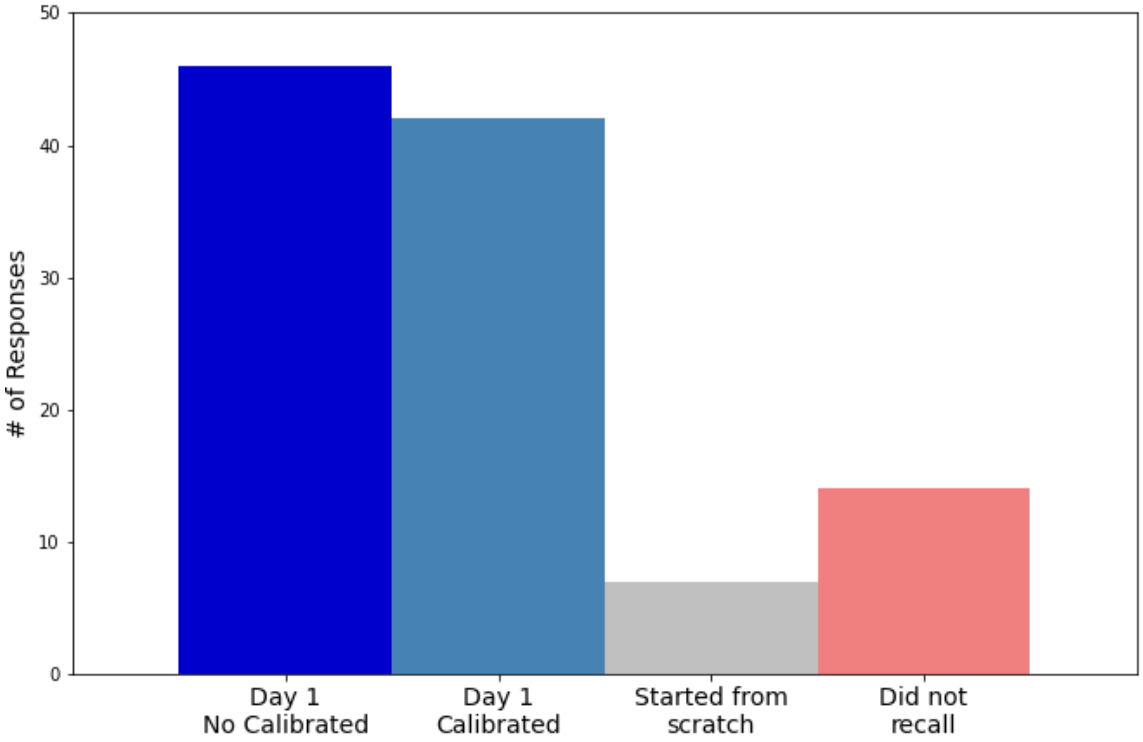


*Figure 59. Self-reported starting points for participants generating updated Day 1 convective outlook forecasts.*

Generally, participants chose to utilize one of the previously issued Day 1 forecasts as a starting point for their outlook updates (Fig. 59). However, a handful of participants started from scratch across the experiment, two of which were NWS expert forecasters. The Day 1 Calibrated and No Calibrated forecasts were each used a similar number of times, suggesting perhaps that the outlook that participants were involved in issuing each morning served as their starting point in the afternoon. Participant ratings of their updated forecasts were similar to the ratings for the initial Day 1 forecasts issued in the morning, except that the expert forecasters tended to give their forecasts lower ratings than the consensus forecast contributors assigned their forecasts (Fig. 60). Further objective verification should be undertaken to determine whether the expert updated forecasts truly were less skillful than the morning forecasts or the consensus, or whether the lower ratings are an effect of the operational forecasters having different mindsets when evaluating the forecasts relative to the rest of the participants. Preliminary stratification of the morning outlook's ratings based on whether participants issued expert or consensus forecasts indicated that distributions were similar (not shown), so these differences are not solely due to forecasters generally scoring forecasts lower subjectively than other participants.

In addition to rating their updated forecasts, participants provided feedback on the difficulty in creating consensus outlooks generally and the utility of WoFS in generating their updated Day 1 outlooks (Fig. 61). Both of these questions were hazard agnostic, so participants should have considered all hazards in their responses, but may have put more weight on the most challenging or impactful hazard to forecast for a given day. Most frequently, participants found the conditional forecasts "somewhat difficult" to create, although more participants responded "extremely easy" relative to "extremely difficult". The second most common response was that the forecasts were "Neither easy nor difficult" to create. These results indicate that the conditional intensity forecasts remain somewhat challenging to conceptualize, emphasizing the need for training and associated materials for the best grasp of what is being communicated by the conditional intensity forecasts and in what situations they provide the most benefit.

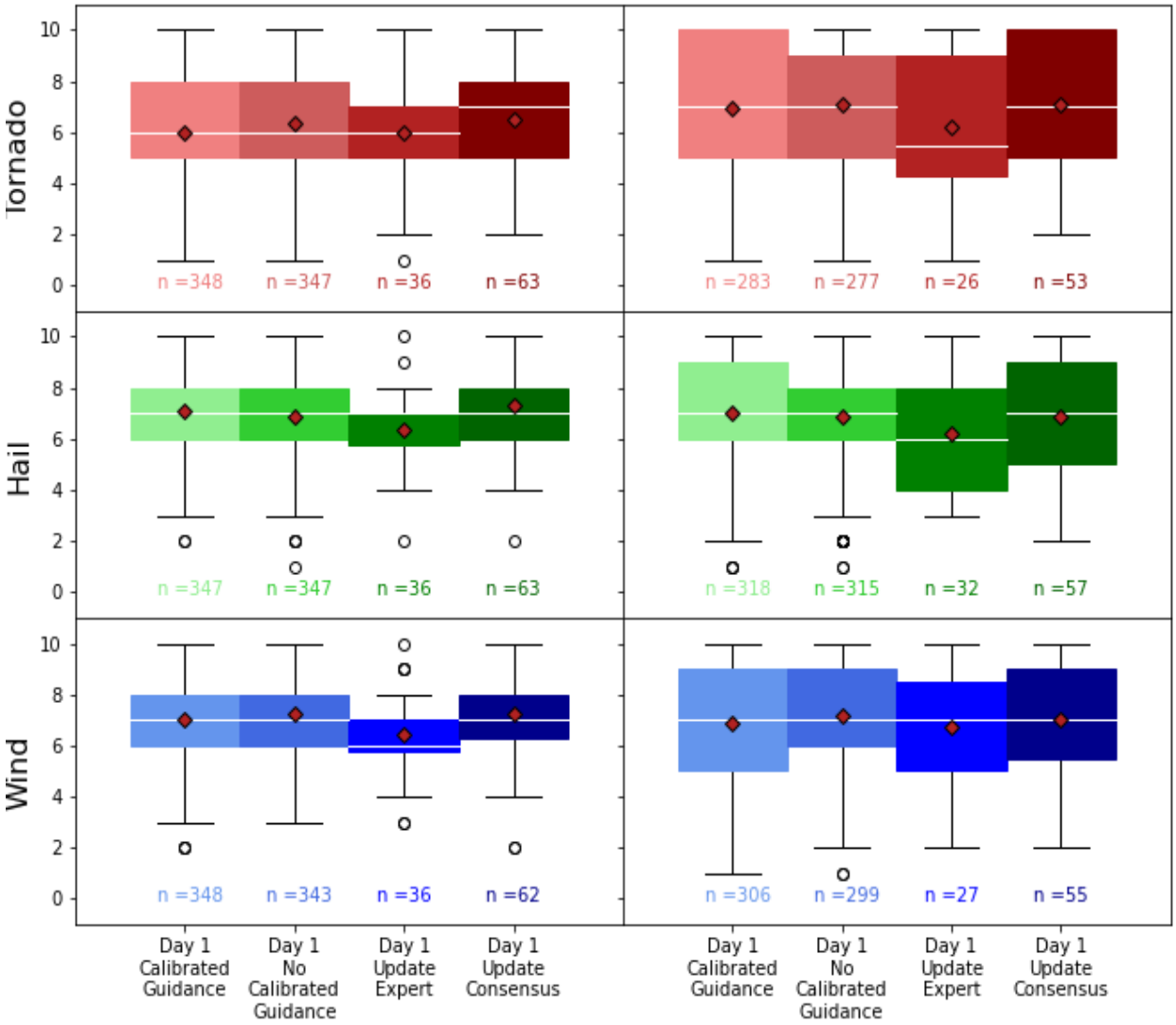


*Figure 60. As in Fig. 56, but showing the Day 1 Calibrated and No Calibrated forecasts, and the expert and consensus forecast updates for Day 1.*

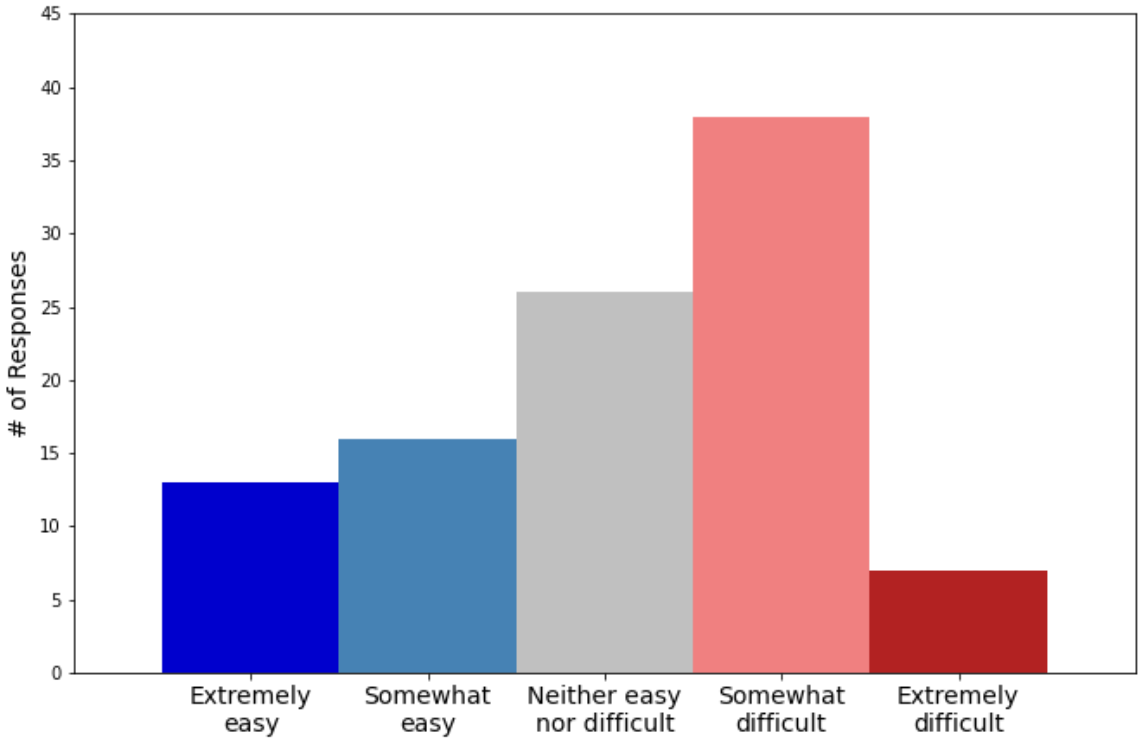


*Figure 61. Participant responses to the question, "How difficult was it to create the conditional intensity forecasts yesterday?"*

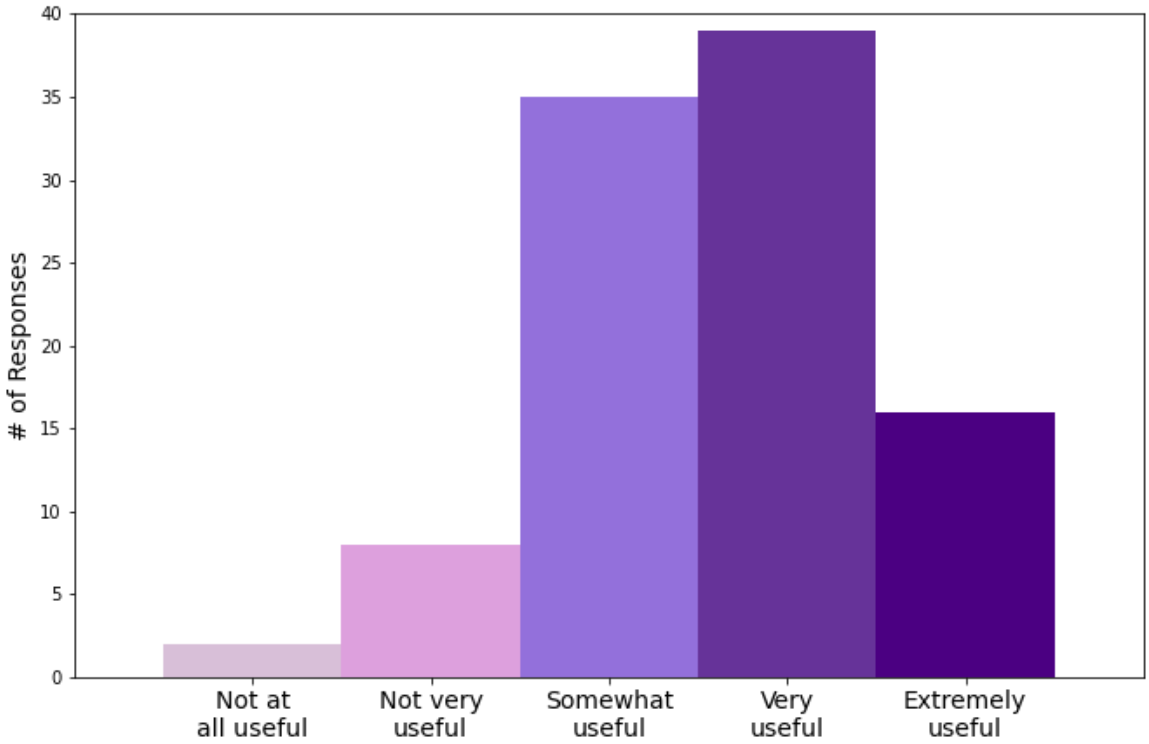


*Figure 62. Participant responses to the question, "How useful was the Warn-on-Forecast System (WoFS) to you yesterday in issuing your forecasts?"*

Aligning with results from many prior SFEs, participants typically found WoFS to be "*somewhat useful*" or "*very useful*" (Fig. 62). Given the traction WoFS has gained with the forecasting community, WoFS developers may utilize these responses to investigate cases in which WoFS was either not useful (e.g., garnered "*Not at all useful*" or "*Not very useful*" responses) or WoFS was especially useful (e.g., garnered "*Extremely useful*" ratings). Three days had notable numbers of participants indicating that WoFS wasn't very useful: 9 May (4 responses), 16 May (2 responses), and 17 May (2 responses). Four days had multiple "Extremely useful" responses as well, including one case that also had two "*Not very useful*" responses: 16 May (3 responses), 26 May (3 responses), 18 May (2 responses), and 1 June (2 responses). From these results, 16 May seems to warrant further investigation, due to the polarizing nature of the responses surrounding WoFS on this date.

### 3.5.4 Focus Group on Conditional Intensity Guidance

As part of the afternoon activities of the 2022 SFE, participants took part in a focus group that furthered research on a two-year project titled "Enhancing the Storm Prediction Center's Convective Outlook with Continuous Probabilities and Conditional Intensity Forecasts". Participants were asked to take part in a 30–35-minute conversation about the use of conditional intensity information in SPC products. This activity aimed to understand how these innovations would or would not benefit weather forecasters and their ability to communicate with partners and members of the public. To assess the views of the weather forecasters, participants were asked the following questions:

1. What are your first impressions of a convective outlook product that more clearly delineates between coverage and intensity? What are some benefits and/or challenges to splitting these two pieces apart?
2. How would you weigh coverage and intensity? Is one more important than the other? Do partners and members of the public ask for one more than the other? Should they always be displayed together, or can you separate the two?
3. Right now, the difference between levels of conditional intensity is more qualitative in nature, do you like this, or do you think the levels should have more strict guidelines?

These questions were asked of every focus group participant, though the facilitator did ask more probing and follow up questions based on the flow of the discussion. Many of these questions saw split reactions among the participants. A recurring benefit discussed during the "first impressions" question was that it would make it easier to visualize and communicate high intensity, low probability days. There was debate about whether or not conditional intensity information would be beneficial to the public, though many agreed it was information partners could both understand and benefit from. Some example quotes from this discussion are displayed below:

- "*From a DSS perspective, this information good for EMs. They'll get trained to look for things in certain way and make decisions based on the information*."
- "*I like that you can communicate if you are not sure if the event is going to happen, and you can highlight that if it does it will be intense. But I am worried about how the public will respond because I don't think they have a good grasp on what "hatched" means. I think it could be useful if it is communicated well."*
- "*As far as the public goes, I'm not even sure people know the difference between what intensity means and what coverage means*."

When asked how they weighed the importance of coverage vs. intensity information, participants were well split, with some forecasters indicating that timing information may actually be more important than either coverage or intensity. This question was originally asked to try to determine which piece of information should be prioritized in graphics. With this framing in mind, there were also many instances when the idea of toggling layers on and off maps was suggested by the participants.

- "*In my experience, I weight more towards intensity. Even at the EM level, coverage is less well-understood. The question tends to be "how bad will it be? How bad could this get*?"
- "*People might assume they'll be affected, so coverage info is more assumed by the public based on forecast*."
- "*People will tell you they care more about intensity, but real thing they care about is coverage ('Will I get hit?')"*
- *"From a meteorology perspective, it is better to have one map with a toggle on and off perhaps - instead of comparing two images side by side."*

Finally, when asked about more nebulous vs. specific definitions for the different levels of hatching used in SPC products — no hatch, single hatch, and double hatch — there was also disagreement over whether flexible vs. distinct definitions were better, though the challenges in communicating hatching were often brought up. Many participants liked the idea of SPC forecasters having flexibility when assigning a hatch, but also noted the challenges that nebulous definitions bring in terms of consistency.

- *"Having hard and fast definitions makes it easier for younger forecasters to make decisions, but with experience gut definitions kick in, making flexibility is better."*
- *"I think it's a tricky road to navigate, but regardless of the definitions, there needs to be a lot of thought before putting the hatch (or double hatch) out - it needs to mean something if you are going that high."*
- *"I like as a forecaster having a bit more freedom — I like the idea of it being more nebulous because there is more room for interpretation. As a communicator, if a*

*message is not consistent then we have a problem. One person might deem something worthy of a slight but another doesn't. We can get into problems where the nebulous bits lead to us not speaking the same language and not communicating effectively to the public."*

More thorough analysis of themes and comments will be conducted this fall. While this analysis may reveal more consensus in the themes and sub-themes within the answers given by the participants, it is clear there was debate among the forecasting community over the ways in which conditional intensity should be used, visualized and communicated.

### 3.5.5 WoFS-Based, 1-h Outlooks With and Without Machine-Learning Guidance

During the 2:15–4 pm CDT time period in the Innovation Group, participants generated severe hazard probabilities valid over 1-h time windows covering 2100–2200 UTC and 2200–2300 UTC. Initial forecasts were generated during the 2:15–3:15 pm period and final forecasts were generated during the 3:15–3:45 pm period. After the final forecasts were issued, from approximately 3:45–4 pm, participants completed a survey to gain insight on the use of ML-based forecast products from WoFS. All of the Innovation Group afternoon forecasting activities were conducted in two sub-groups. One group had access to calibrated, WoFS-based ML guidance when issuing their forecasts, while the other only used the uncalibrated WoFS products. For both sets of initial and final forecasts, two forecasters were in the group that included ML guidance, while two other forecasters were in the group without ML guidance. Additionally, other participants in each group issued forecasts with and without the ML guidance similarly to the expert forecasters, which were combined into consensus forecasts.

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. Additionally, participants were presented training material to familiarize them with the ML guidance products. An example set of forecasts from 1 June 2022 is shown in Figure 63.

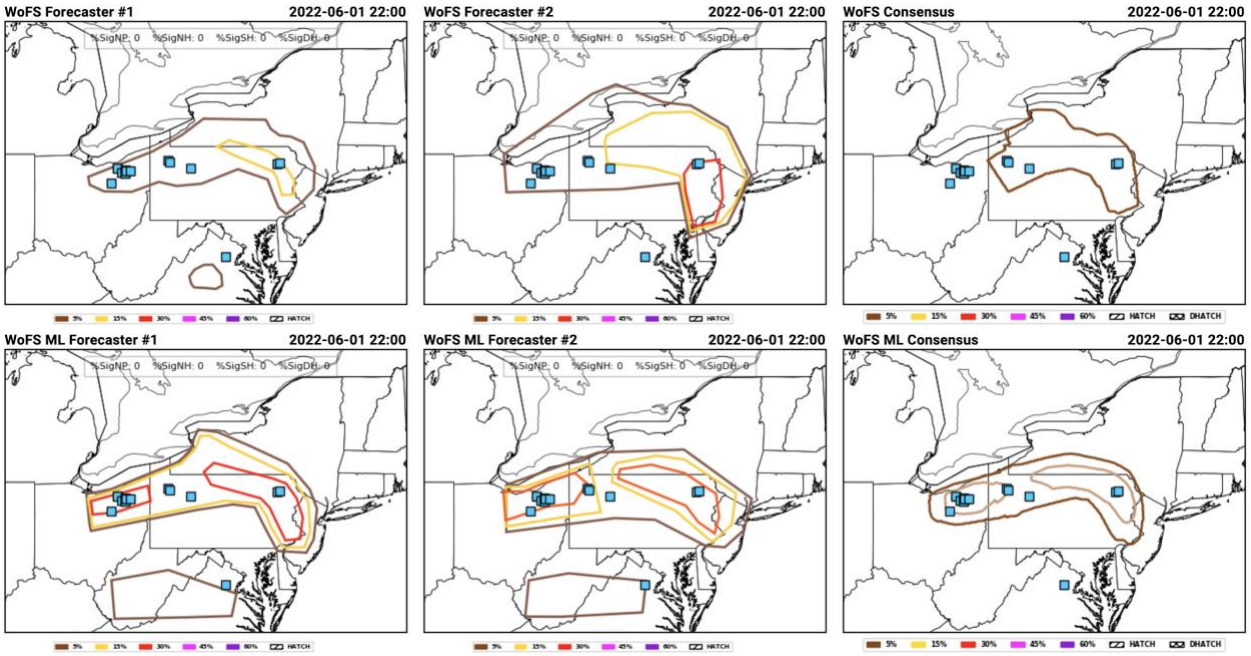


*Figure 63. Innovation Group outlooks generated as part of the afternoon forecasting activity highlighting the probability of severe wind gusts covering the 1-h period 2100–2200 UTC on 1 June 2022: WoFS Forecaster #1 (upper left), WoFS Forecaster #2 (upper middle), WoFS Consensus (upper right), WoFS ML Forecaster #1 (lower left), WoFS ML Forecaster #2 (lower middle), and WoFS ML 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 72 outlooks that were evaluated each day (3 hazards x 4 times x 6 forecasts = 72). The primary goal of this exercise was to quantify the value of machine-learning to the experimental outlooks by comparing those made with and without the machine-learning guidance. 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 WoFS ML Forecasters #1 and #2. Furthermore, initial forecasts (issued 2:15–3:15 pm) and final forecasts (issued 3:15–4 pm) were averaged together to create the composite forecast ratings in Figure 64. Plots showing the results separated by the initial and final outlooks, as well as 2100–2200 and 2200–2300 UTC time periods, are available in the appendix.

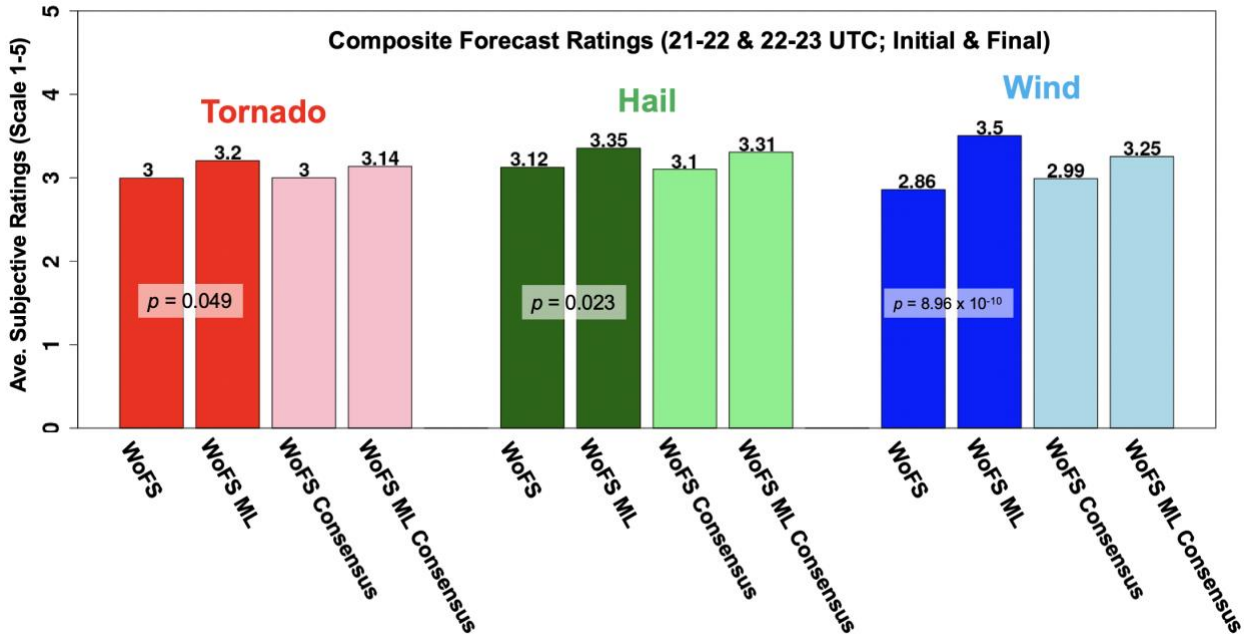


*Figure 64. Average subjective ratings for WoFS, WoFS ML, WoFS Consensus, and WoFS ML Consensus for all three hazards averaged for the 2100–2200 and 2200–2300 UTC time periods, as well as the initial and final outlooks. p-values from a Welch's t-test comparing the WoFS and WoFS ML outlooks are overlaid on the histogram bars for each hazard.*

Overall, the WoFS ML forecasts were rated higher than WoFS, with the most dramatic differences for wind. For the aggregate ratings, differences between WoFS and WoFS ML were statistically significant for all three hazards. In addition, the average ratings for WoFS ML Consensus were higher than WoFS Consensus, but none of the differences reached the threshold for statistical significance. These results indicate that ML can provide significant value on top of that provided by the raw WoFS products.

After evaluating a set of initial or final forecasts, participants who completed the hourly outlooks were asked questions about how confident they would be using WoFS overall, as well as the machine learning guidance specifically. Both groups were also asked how useful WoFS was for each individual hazard. Overall, participants were typically at least "moderately confident" in using WoFS and the machine learning guidance after seeing the verification for the previous day (Fig. 65). These results held whether participants were in the group that had the ML guidance or in the group that did not have access to the ML guidance while issuing their forecast. Overall, the group that had the WoFS ML guidance was slightly more confident in utilizing WoFS for future events, though the differences in the distributions were not large. For the WoFS ML guidance, the group that utilized the guidance had slightly more polarized feelings of confidence in the ML guidance after seeing the verification: the ML group had more responses than the no-ML group in every category except "Moderately confident", which was the middle response. However, more forecasters in the ML group indicated that they would be "extremely confident" using the ML guidance going forward relative to those with ML guidance.

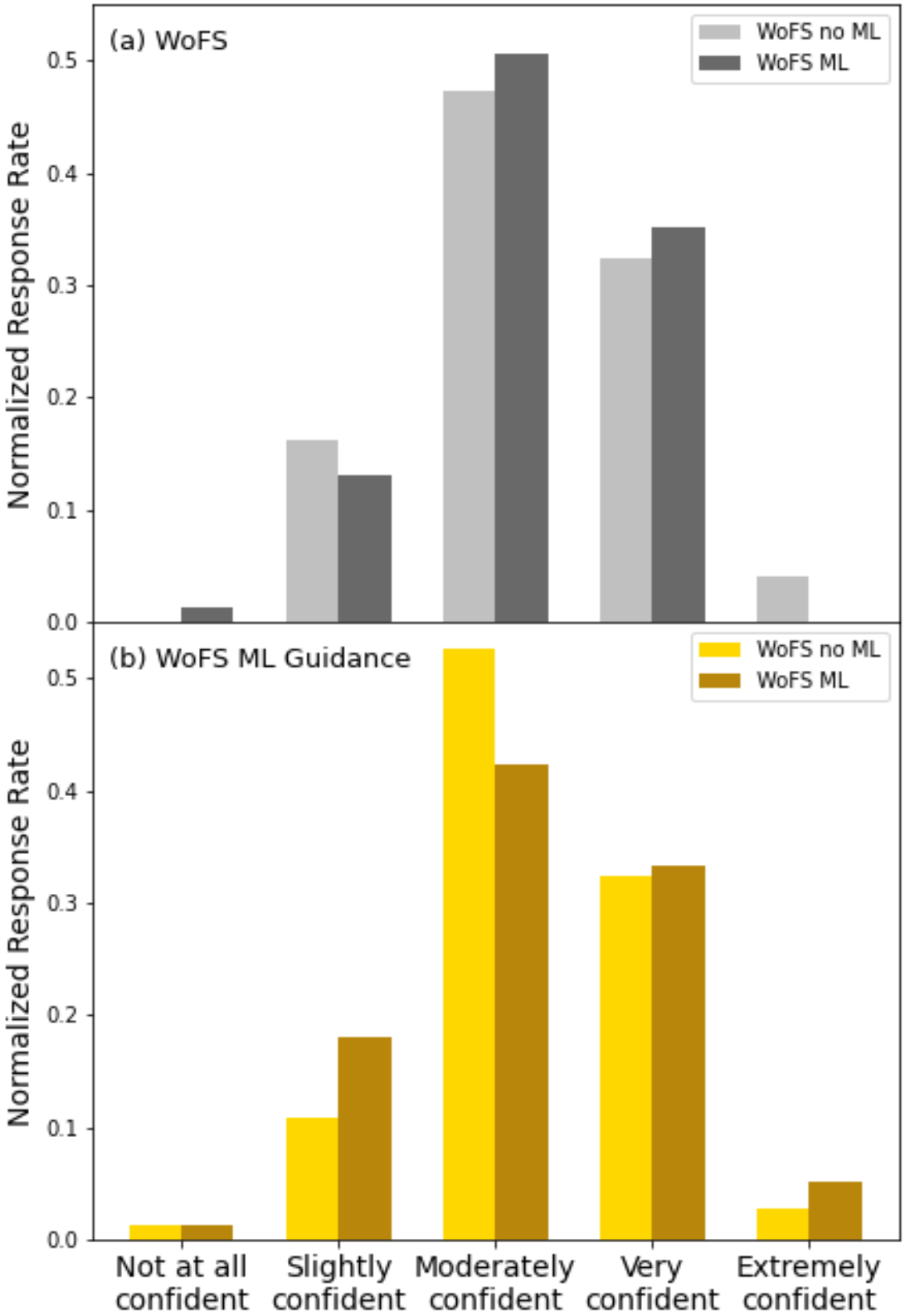


*Figure 65. Participant responses to the questions, (a) "After seeing the forecast verification, how confident would you be in using the WoFS while issuing a future forecast?" and (b) "After seeing the forecast verification, how confident would you be in using the WoFS machine learning guidance while issuing a future forecast?"*

When asked about the usefulness of WoFS for the individual hazards, generally participants found the WoFS to be most useful for the hail and wind hazards, although many participants found utility for the tornado guidance as well (Fig. 66). The group utilizing the machine learning guidance found more utility from the WoFS for all hazards relative to those without machine learning guidance (i.e., more responses of “Very useful” or “Extremely useful”, suggesting that the machine learning tornado guidance is a useful hazard forecasting tool that benefits participants. Overall, participants found the WoFS to be quite useful for all hazards forecasted during the afternoon forecasting activity, demonstrating once more the utility of WoFS for short-term hazard forecasting.

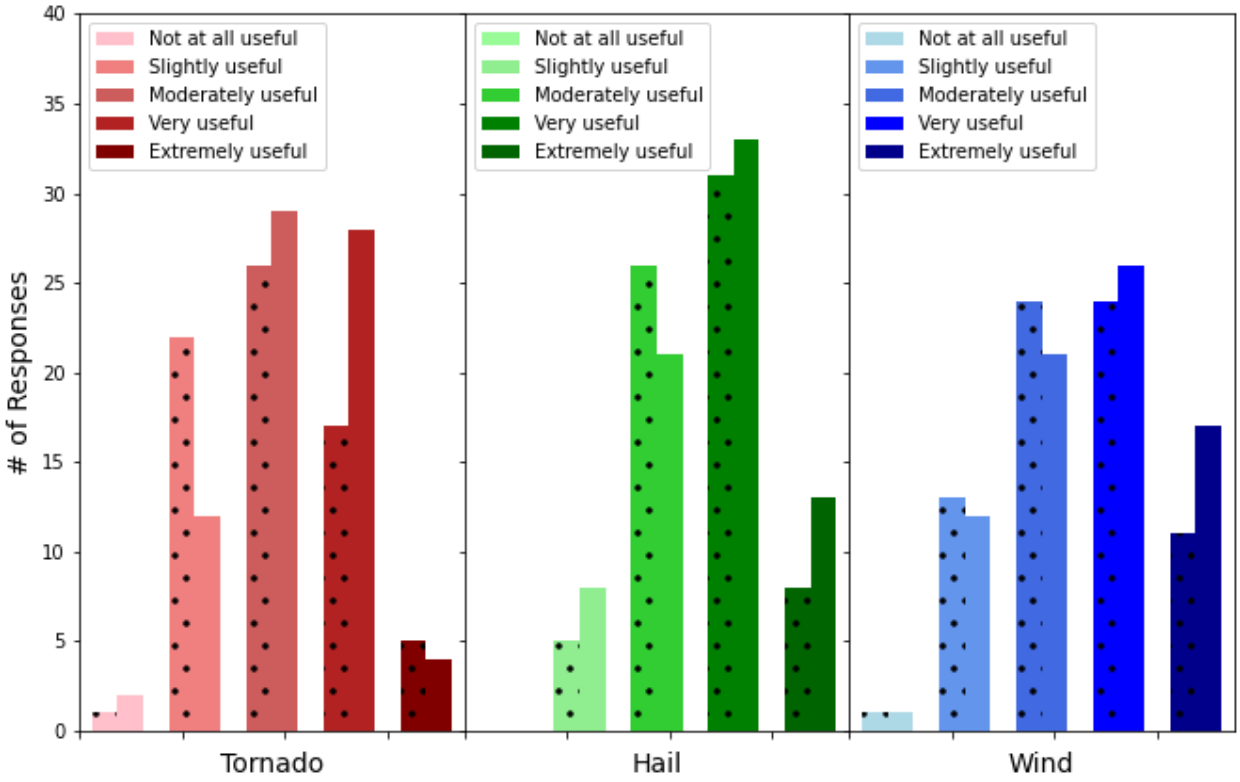


*Figure 66. Participant responses to the question, “Please indicate the usefulness of WoFS for the following hazards today.” Dotted bars are from the group without access to ML guidance, while solid bars are from the group that had access to the ML guidance.*

Finally, the machine learning group was asked a question about the explainability graphics, in which they were presented with a static image (Fig. 67) and asked their opinion on which sets of predictors would be preferred. Participants had access to the explainability graphics during the forecasting exercise, and a global perspective (e.g., consistent fields) was used for this initial attempt at incorporating explainability graphics into the forecasting process.

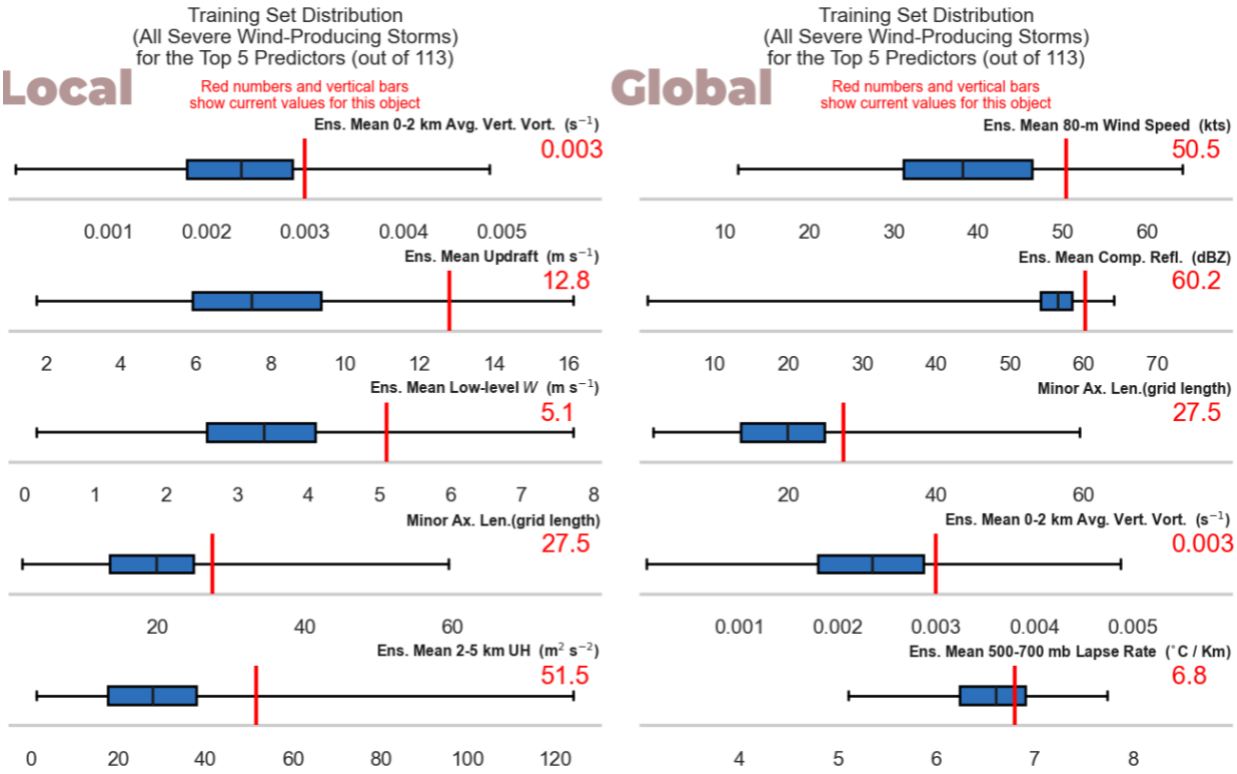


*Figure 67. An example of local (left) and global (right) sets of predictors for the explainability graphics. Local predictor fields would change depending on the storm object, while global predictor fields would remain the same between objects. Participants were shown this image before asking which set of predictors they preferred.*

All response options were selected by at least some participants, but the storm-specific fields were most commonly preferred, followed by "Do not prefer one or the other" (Fig. 68). Participants also had the option to write in a suggestion under the "other" response. Some participants used this response to indicate that they wanted more training with the product, to see the product demonstrated during an event, or to see the product demonstrated for an event with more storms. before making their selections. Others indicated a nuanced take, where different preferences would match to different scenarios. As one participant said, "*I would prefer the global attributes if I were forecasting near the maximum in severe wind. If I were forecasting in an area that does not see stronger winds often, I may prefer the local variables*". Another requested a mixture, incorporating 2–3 parameters from each option onto the plot. This feedback will allow for iteration of the explainability products, and expanded utility of the machine learning guidance.

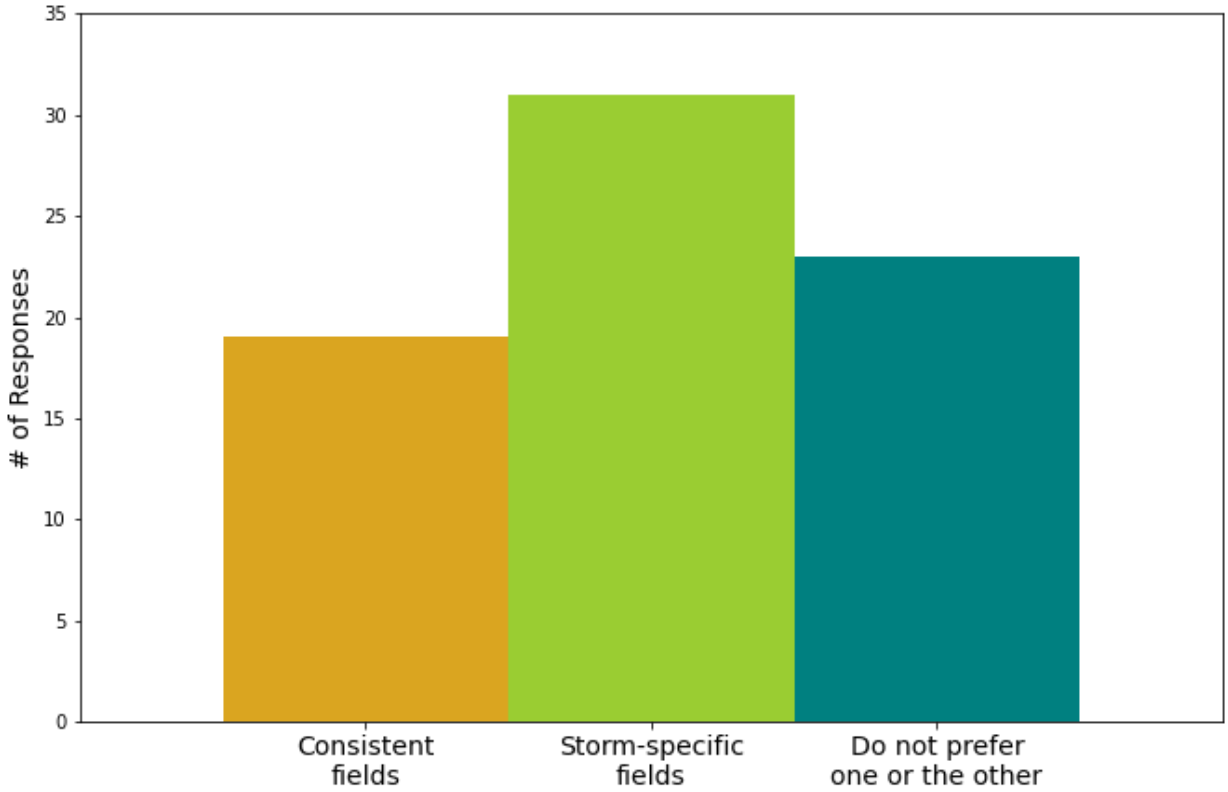


*Figure 68. Participant responses to the question, "Would you prefer consistent fields (explaining how the same set of predictors contribute to the prediction regardless of the storm) or storm-specific fields (using a different set of predictors contribute to each storm's prediction)?" An "other" response with a write-in was also available, and responses in this category are discussed in the text.*

SFE 2022 participants were also surveyed on their use of the WoFS ML guidance after they issued their forecasts each afternoon. After completing both of their outlooks, participants were asked to complete a survey asking them about the number of products they used, and the confidence in their forecasts of each hazard. The group that used the machine learning guidance were asked a few additional questions, including aspects of the guidance that worked or did not work for them, and perceived utility of the guidance. Explainability graphics were also available to allow participants to see the underlying values of inputs to the machine learning guidance, and participants were also asked for feedback on that guidance. The analysis herein focuses mainly on the Likert Scale questions, and parsing of the open-ended question data collected is still underway. The analyses plotted here encompass 21 cases, with 260 participant responses spanning those cases.

Both the group that used WoFS and machine learning (WoFS ML) and the group solely using WoFS (WoFS no ML) answered questions about the confidence they had in each of their forecasts and the number of products used, to determine if the ML guidance influenced participant behavior. Overall, the people using the ML guidance tended to self-report using more products (Fig. 69). For participants who didn't use the ML guidance, they most frequently looked at 6–10 different products, while the participants using the ML guidance more frequently looked at 11–15 products. The group using the ML

guidance also more frequently responded that they used 16+ different products on the WoFS page. Initial hypotheses were that the WoFS ML guidance may lead people to look at fewer products since the ML guidance ingests and accounts for multiple WoFS fields; this hypothesis does not seem to be supported by these findings. Instead, it may be that the participants simply add the ML guidance to the usual suite of products that they look at, increasing the overall number of products considered. Increased familiarity with and exposure to WoFS could change this initial behavior.

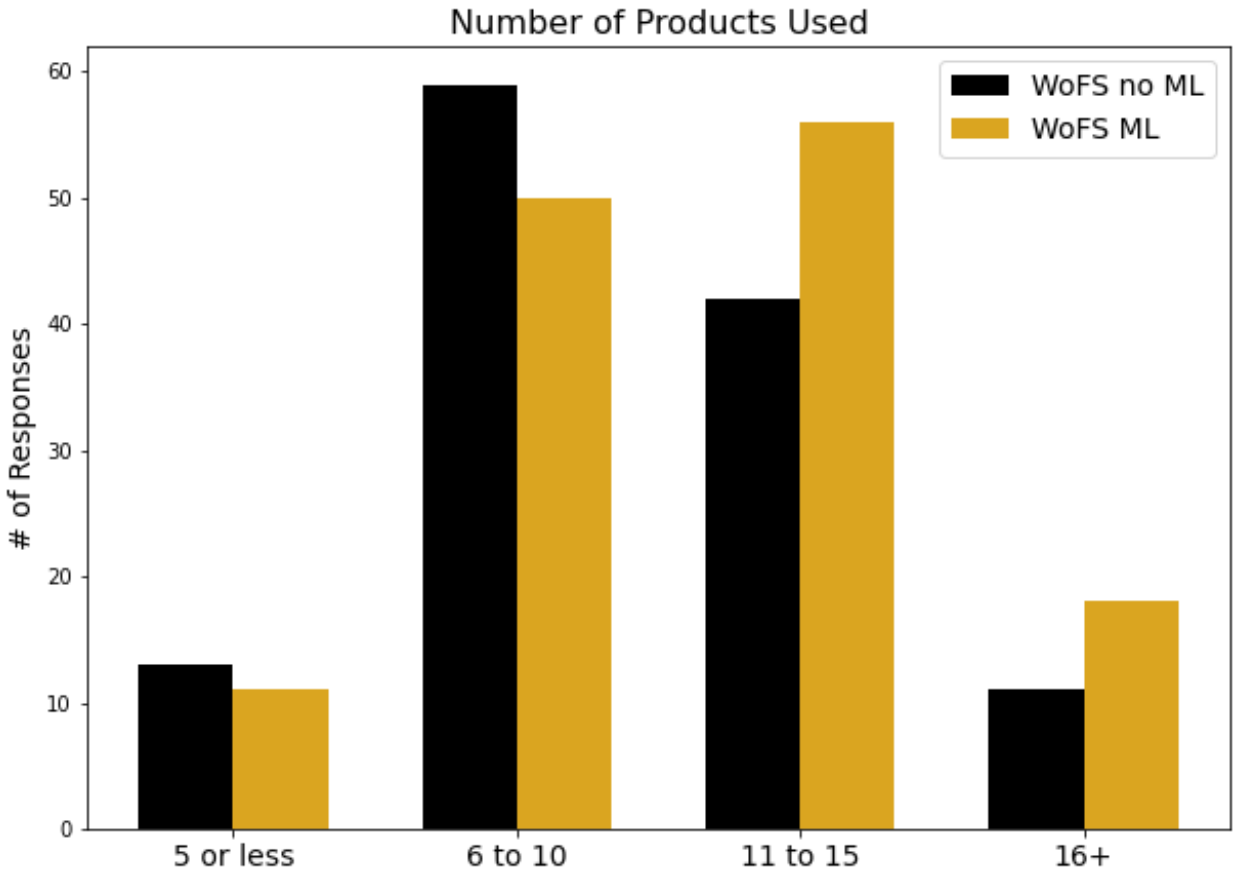


*Figure 69. Participant responses to the question, "Approximately how many different WoFS products did you look at today when formulating your forecasts?" Participants were given the response options shown here, i.e., the number of products was pre-binned in the responses.*

The afternoon survey also asked participants how confident they were in their forecasts of each individual hazard (Fig. 70). While differences were relatively small between the groups, participants using the ML guidance did have more responses of being "Very" or "Extremely" confident for their tornado and hail forecasts. For the wind forecasts, the WoFS ML group was more likely to say that they were "moderately" or "very" confident in their forecasts. Conversely, the WoFS no ML group was more likely for all hazards to say that they were slightly confident in their forecasts. Overall, participants showed a Gaussian distribution of responses around the middle response,

with the typical response from participants of both groups as “moderately” confident for all three hazards.

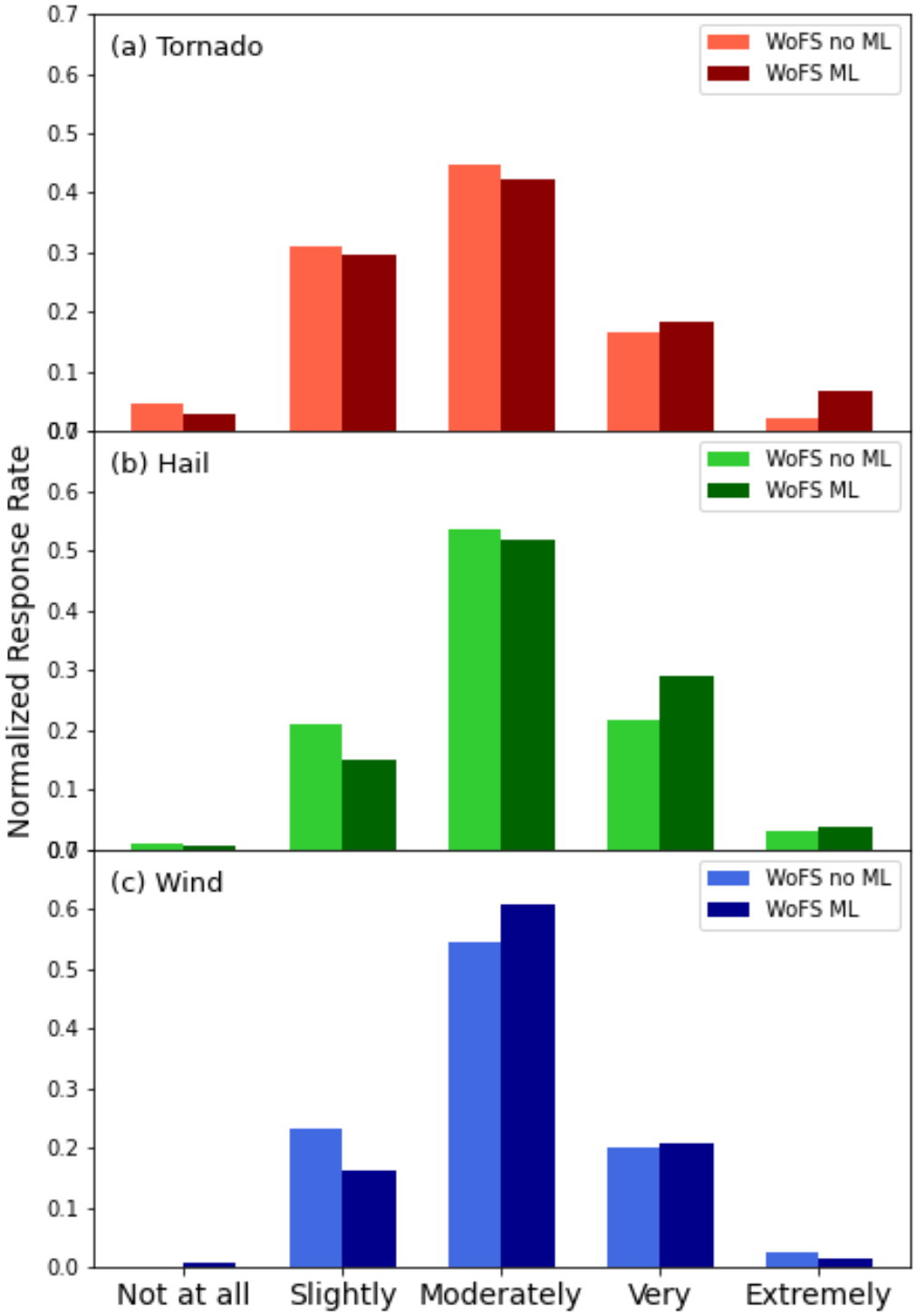


*Figure 70. Participant responses to the question “How confident are you in your forecasts of the following hazards today (considering both the 2100–2200 and the 2200–2300 UTC time periods)?”. Participants responded separately for the (a) tornado, (b) hail, and (c) wind hazards.*

For the question asking about the utility of the WoFS ML guidance, participants found the guidance to be somewhat or very useful most of the time (Fig. 71). The guidance was most useful for the wind threat, where a majority of participants rated the guidance as “very” useful after using it to make their forecasts. Very useful was the most common answer for the hail threat as well, while the tornado threat was most frequently rated as “somewhat” useful. Please note that these questions were asked prior to the verification of the forecasts, so participants did not yet know how their forecasts would turn out. When asked about the utility of the supplemental explainability graphics associated with each forecast object, participants found slightly less utility in them relative to the machine learning guidance itself (Fig. 72). However, the explainability graphics were still found to be useful, with the majority of participants indicating that they were somewhat useful for all hazards. Participants may need more time to explore the explainability graphics, since this is their first utilization in the SFE, and additional questions in the Evaluation of Yesterday’s Forecasts will provide feedback to refine these graphics for future usage. Some participant comments reflected a need to further understand the explainability graphics, but many participants commented that they liked the ideas behind these graphics.

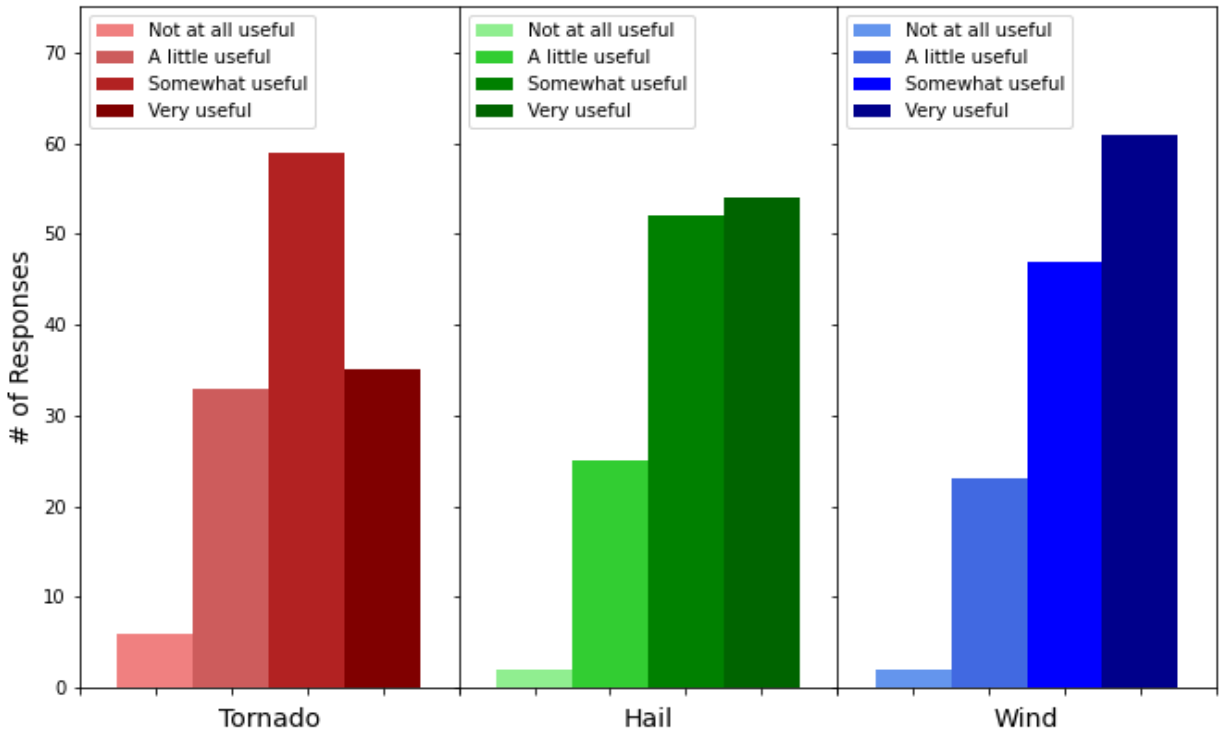


*Figure 71. Participant responses to the question, “How useful was the machine learning guidance when creating forecasts of the following hazards today (considering both the 2100–2200 and the 2200–2300 UTC time periods)?” for each hazard.*

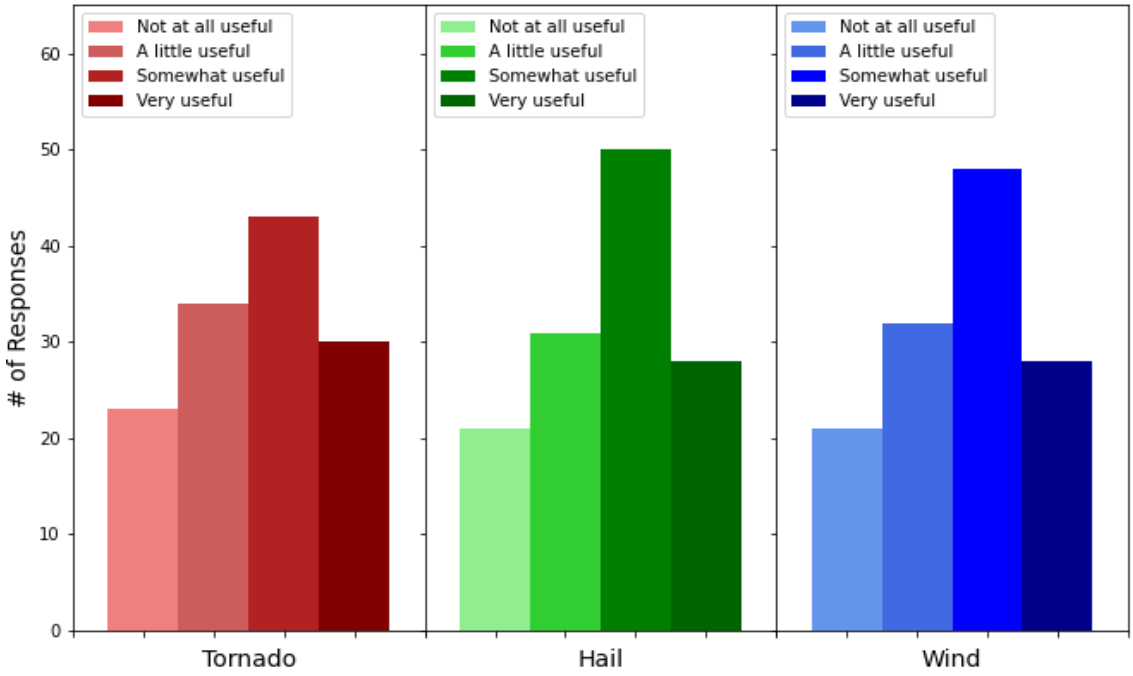


*Figure 72. Participant responses to the question, "How useful were the explainability graphic when creating forecasts of the following hazards today (considering both the 2100–2200 and the 2200–2300 UTC time periods)?" for each hazard.*

## 4. Summary

The 2022 NOAA HWT Spring Forecasting Experiment (2022 SFE) was conducted virtually from 2 May – 3 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 2022 SFE were to, (1) evaluate convection-allowing model and ensemble guidance for identifying optimal configurations of convection-allowing versions of FV3 and CAM ensembles, including several carefully designed and controlled experiments as part of the Community Leveraged Unified Ensemble (CLUE), (2) study how forecasters and meteorologists utilize CAMs and CAM ensembles, such as WoFS, and evaluate various experimental severe weather outlooks generated using WoFS and other CAM ensembles for lead times from one hour to 3 days, and (3) evaluate different CAM ensemble post-processed guidance with an emphasis on those using machine-learning algorithms.

Several preliminary findings/accomplishments from the 2022 SFE are listed below:

- Experimental short-term individual hazard outlooks were generated using WoFS with and without machine-learning guidance. 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 using WoFS with machine-learning provided a statistically significant advantage relative to outlooks produced using WoFS without machine learning. The biggest advantage provided by the machine learning products was for the 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, an ML algorithm using HREF and SREF predictors known as “Nadocast” performed best overall for Day 1 lead times. However, for days defined as “active” (i.e., SPC outlook tornado probabilities of 5% or greater) Nadocast and the STP-cal methods exhibited similar performance.
    - For hail, an HREF-based ML random forest algorithm performed particularly well for both Day 1 and Day 2 lead times. Furthermore, the GEFS-based

ML hail forecasts, which were evaluated separately, also performed notably well at Day 1–3 lead times.

- For wind, the HREF-based ML random forest algorithm performed best for Day 1 and Day 2 lead times.

- Examined various **deterministic** CAM systems within the CLUE using HRRRv4 as a baseline.
  - In blinded evaluations, the RRFSp1 and RRFSp2 Control show skill approaching the HRRRv4 for simulated reflectivity and 2–5 km UH, 2-m dewpoint, and SBCAPE.
  - The HRRRv4 performs best in terms of 2-m temperature forecasts relative to the other Deterministic Flagship models.
  - The HRRRv4 was most frequently ranked as the best model for simulated reflectivity and 2–5 km UH, 2-m dewpoint, and SBCAPE.
  - For simulated reflectivity and updraft speed, the HRRRv4 performed subjectively better more frequently than the RRFSp2 Control. However, the 10-m wind speeds and the 0–3 km UH were frequently better in the RRFSp2 Control relative to the HRRRv4.
  - The RadVTS control seemed to perform better in the first forecast hour relative to the RRFS BothVTS Control, but the runs performed similarly at forecast hour six.
  - SBCAPE values from FV3-based models continue to be too low in the first 18 hours of the forecasts.
  - Of the physics suites examined in B4, the NSSL microphysics, MYNN PBL scheme, and NOAH LSM perform best overall. For simulated reflectivity and 2–5 km UH, the NSSL microphysics performed better than the Thompson microphysics, but the SBCAPE forecast may be degraded relative to the fields produced by the Thompson microphysics.
  - Initial responses show better indication of severe weather threat from a 1-km horizontal grid spacing version of the NSSL-WRF, but further work remains to evaluate the full impact of 1-km horizontal grid resolution relative to 3-km horizontal grid resolution.

- Examined various **ensemble** CAM systems within the CLUE using HREFv3 as a baseline.
  - In blinded evaluations, HREFv3 and RRFSp2e were consistently ranked highest and performed quite similarly among five unique CAM ensembles that were evaluated. This is a noteworthy result since it marks the first time that a CAM ensemble has approached the skill of HREFv3.
  - In direct comparisons between HREFv3 and RRFSp2e, RRFSp2e had better 2-m dewpoint and SBCAPE forecasts and HREFv3 had better 2-m temperature forecasts. The two performed similarly for 2-5 km AGL UH.

- Comparing ensembles at 0–12 h lead times that used valid-time-shifting (VTS) approach with different observations ingested, a radar-only VTS approach performed slightly better than VTS using both radar and conventional observations during the 0–4 and 5–8 h forecast periods for 2–5 km AGL UH and composite reflectivity. However, by the 9–12 h forecast period, the performance was similar.
- In an experiment using ensemble sensitivity analysis, it was found that the subset ensemble with the smallest errors early in the forecast usually had no impact or slightly improved the probabilistic 2–5 km AGL UH forecasts.

- Examined utility of WoFS for short-term severe weather forecasting application in the watch-to-warning timeframe.
  - Comparing 2100 and 2300 UTC WoFS initializations with different numbers of members revealed that forecast probabilities derived from 9, 13, and 18 WoFS forecasts performed very similarly. This suggests that gains in skill may be achieved from reducing membership but using more advanced physics, data assimilation, and/or enhancing the resolution, while using the same amount of computational resources.
  - Comparing different time-lagging strategies for 2100 and 2300 UTC WoFS initializations revealed that time-lagging slightly degraded the forecasts, especially for the earlier initializations. Thus, time-lagging is likely not a viable strategy for WoFS.
- 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 $\geq$ 50 knots. The primary results were: (1) the ML models that were trained with the additional database of sub-severe thunderstorm wind gusts generally received higher ratings than those models trained only with measured wind reports, (2) the impact of which specific ML model was used was relatively small in the subjective ratings, and (3) in construction of practically perfect hindcasts for severe wind, participants generally agreed that weighting the wind reports using the ML output was preferred over treating all wind reports equally.
  - Three unique ML algorithms were trained to provide probabilistic guidance on storm mode using output from the HRRR. Distributions of subjective ratings revealed there was not a preferred algorithm, which is a favorable result for the partially supervised GMM approach that does not require extensive hand labeling. Feedback from a new neighborhood probability product for convective mode was somewhat mixed, but participants commented about the potential utility, especially for summarizing convective mode evolution.

- Three versions of 3D-RTMA with different backgrounds were evaluated. The version that used HRRR performed best, while two FV3-based versions usually performed slightly worse or about the same as the HRRR version. Generally, the HRRR-based version handled the effects of convection on 2-m temperature better than the FV3-based versions through more accurate representation of the size, shape, and magnitude of thunderstorm outflows.
- 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, and the 80-m winds received higher subjective ratings and were found to better match the magnitudes of measured gusts than the 10-m winds. However, spurious convective gusts were occasionally present in the 80-m wind field even where convection did not form in reality. Thus, it may be useful to use observed reflectivity to filter out spurious members in the future.
- To assess significant severe wind potential, maximum wind in the 0–2 km AGL layer and the integrated wind in the 0–2 km AGL layer were added as hourly maximum fields in the NSSL-WRF. These new variables performed very well for the only significant severe wind event that occurred, but more cases are needed to learn more about performance characteristics.
- Automated, county-based watch guidance was generated using an ML algorithm with 1200 UTC HREFv3 guidance as predictors. The ML guidance performed similarly to SPC Severe Timing Guidance and both were favorably rated in capturing the severe weather evolution of the day.
- To assess the readiness of the Global Ensemble Forecast System (GEFS) to replace the SREF, an evaluation was performed during the 2022 HWT SFE. For severe weather applications at Day 2 & 3 lead times, GEFS generally performed as well as the SREF for environment fields, except for MLCAPE, and better than the SREF for calibrated thunder and severe products.

Overall, the 2022 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 2022 SFE directly promote continued progress to improve forecasting of severe weather in support of the NWS Weather-Ready Nation initiative. In subsequent years, we plan to continue exploring the potential forecasting applications of Warn-on-Forecast, continue examining strategies for CAM ensemble design, accelerate work with our partners to optimize the UFS for CAM forecasting applications, and explore new ways to leverage AI/ML-based strategies for calibrating and post-processing CAM output to aid forecasters. Additionally, we expect that this work will take on particular importance and assist with evidence-based decision making as NOAA moves forward with its plans for a Unified Forecasting System. In the third year of a virtual experiment, we emphasize once again 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 2022 SFE would not have been possible without dedicated participants and the support and assistance of numerous individuals at SPC and NSSL. In addition, collaborations with NCAR, GSL, GFDL, OU MAP, OU CAPS, and EMC were vital to the success of the 2022 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), Jili Dong (IMSG/EMC), Matt Pyle (EMC), Ben Blake (IMSG/EMC), Eric Rogers (EMC), Eric Aligo (IMSG/EMC), Xiaoyan Zhang (IMSG/EMC), Ting Lei (IMSG/EMC), Shun Liu (EMC), Logan Dawson (EMC), Perry Shafran (IMSG/EMC), Manuel Pondeca (IMSG/EMC), Edward Colon (IMSG/EMC), Matthew Morris (SRG/EMC), Gang Zhao (IMSG/EMC), Annette Gibbs (IMSG/EMC), Marcel Caron (IMSG/EMC), Andrew Benjamin (EMC), Kai-Yuan Cheng (GFDL), Lucas Harris (GFDL), Matthew Morin (GFDL), Linjiong Zhou (GFDL), Xuguang Wang (OU MAP), Yongming Wang (OU MAP), Nick Gasperoni (OU MAP), Tsunghan Li (OU MAP), Aaron Johnson (OU MAP), Ming Xue (OU CAPS), Tim Supinie (OU CAPS), Keith Brewster (OU CAPS), Jun Park (OU CAPS), and Xiaoming Hu (OU CAPS) were essential in generating and providing access to model forecasts or products examined on a daily basis.

## References


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

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

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

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

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

# APPENDIX

<table>
<tr><td>Time (CDT)</td><td colspan="2">R2O Group</td><td colspan="2">Innovation Group</td></tr>
<tr><td>9:00 AM – 9:15 AM</td><td colspan="4">Overview of Yesterday’s Severe Weather<br>David Imy</td></tr>
<tr><td rowspan="2">9:15 AM – 11:00 AM</td><td colspan="4">Evaluation Orientation, Individual Working Time, and Discussion</td></tr>
<tr><td>Group A: Calibrated Guidance</td><td>Group B: Deterministic CAMs</td><td>Group C: CAM Ensembles</td><td>Group D: Medley</td></tr>
<tr><td>11:00 AM - 11:15 AM</td><td colspan="4">Break</td></tr>
<tr><td>11:15 AM – 11:30 AM</td><td colspan="4">Weather Briefing<br>David Imy</td></tr>
<tr><td rowspan="2">11:30 AM – 12:30 PM</td><td colspan="2">Issue Day 1 Hazards Coverage and Conditional Intensity Forecasts (2 groups)</td><td colspan="2">Issue Day 2 and Day 3 Hazards Coverage and Conditional Intensity Forecasts (2 groups)</td></tr>
<tr><td>No Cal. Guidance</td><td>Cal. Guidance</td><td>Day 2</td><td>Day 3</td></tr>
<tr><td>12:30 PM – 2:00 PM</td><td colspan="4">Lunch/Break (Tues., Thurs., Fri.)<br>Lunch/Science Brown Bag (Wed.)</td></tr>
<tr><td>2:00 PM – 2:15 PM</td><td colspan="4">Update on Today’s Weather<br>David Imy</td></tr>
<tr><td rowspan="2">2:15 PM – 3:00 PM</td><td colspan="2">Issue MD Product</td><td colspan="2">Issue 1-h outlooks (21-22, 22-23Z)</td></tr>
<tr><td colspan="2">WoFS & obs</td><td>WoFS ML</td><td>WoFS No ML</td></tr>
<tr><td rowspan="2">3:00 PM – 4:00 PM</td><td>Update Day 1 Outlook</td><td>Focus Group Activity</td><td colspan="2">Issue 1-h outlooks (21-22, 22-23Z), End-of-Day WoFS ML Survey</td></tr>
<tr><td>WoFS & other guidance</td><td>Conditional Intensity Discussion</td><td>WoFS ML</td><td>WoFS No ML</td></tr>
</table>

*Table 4. Schedule for Tuesday – Friday. On Mondays, the schedule is similar except the period 9-11:15 am is devoted to training and introductory material.*

| | 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><br>Wind: Supercells, organized clusters, or squall line with bowing segments | Tornadoes and hail: Discrete supercells<br><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 |

*Table 5. Description of "non-hatched" (normal), "hatched", and "double-hatch" conditional intensity forecasts for wind, hail, and tornadoes.*

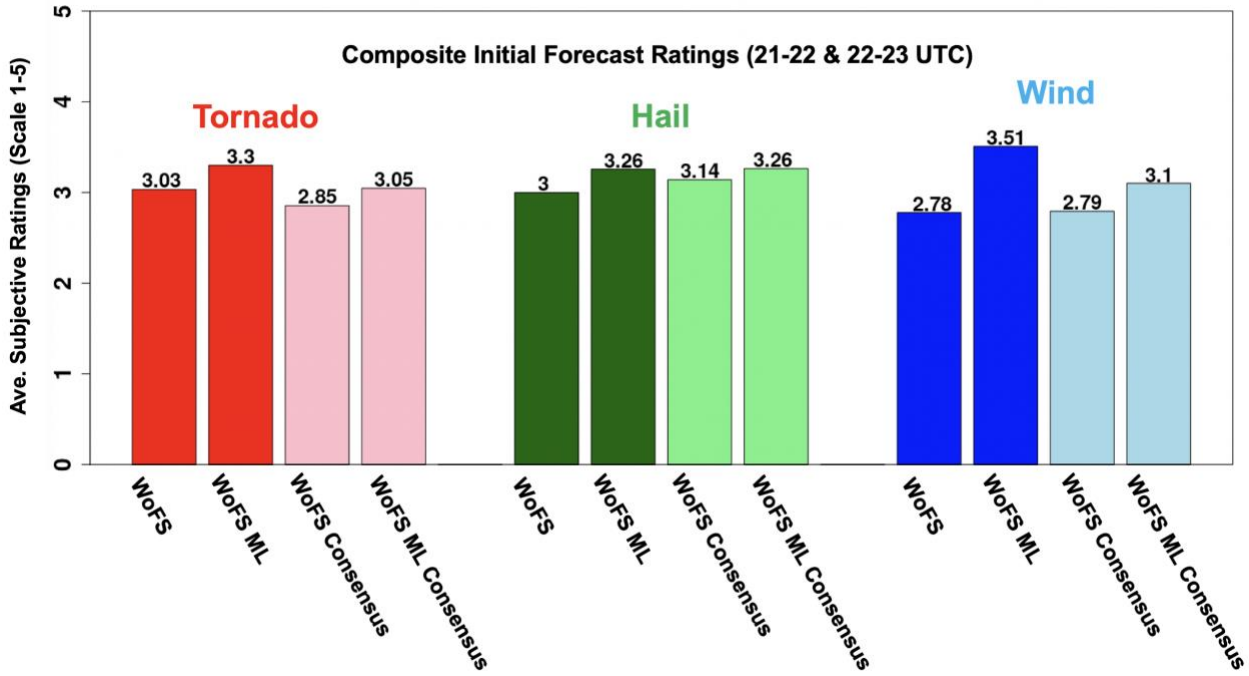


*Figure 73. Average subjective ratings for WoFS, WoFS ML, WoFS Consensus, and WoFS ML Consensus for all three hazards averaged for the 2100–2200 and 2200–2300 UTC time periods for the initial forecasts.*

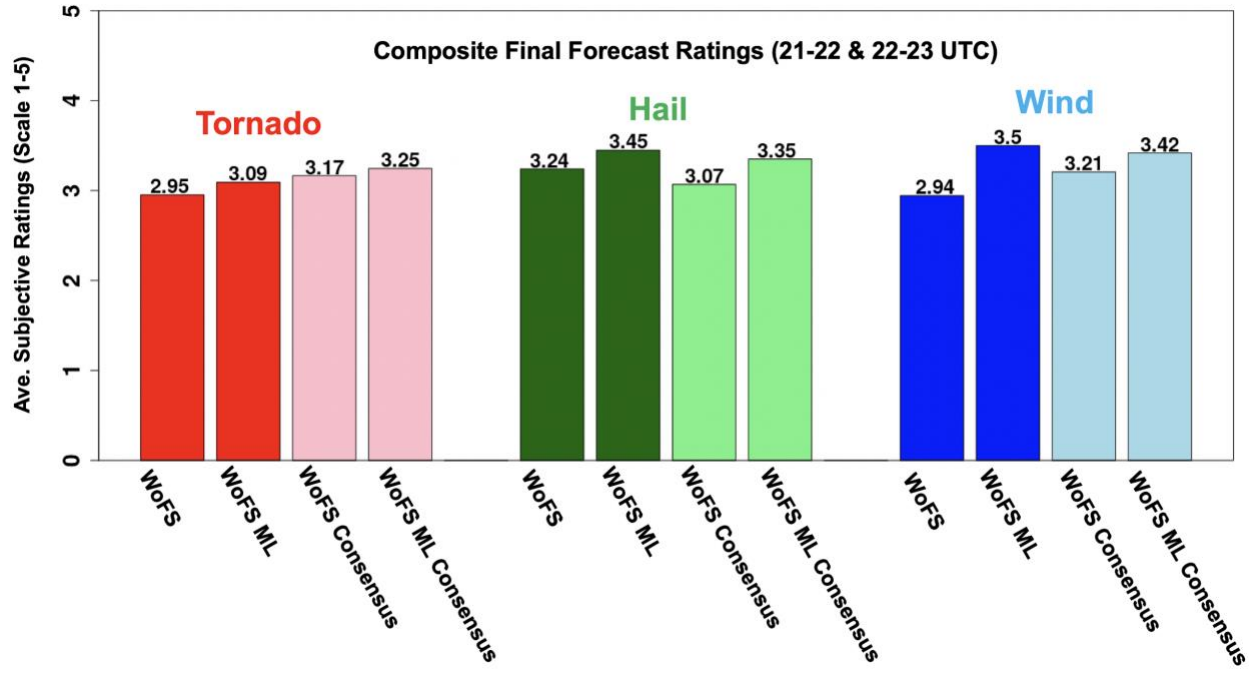


*Figure 74. Same as Fig. A1, except for the final forecasts.*

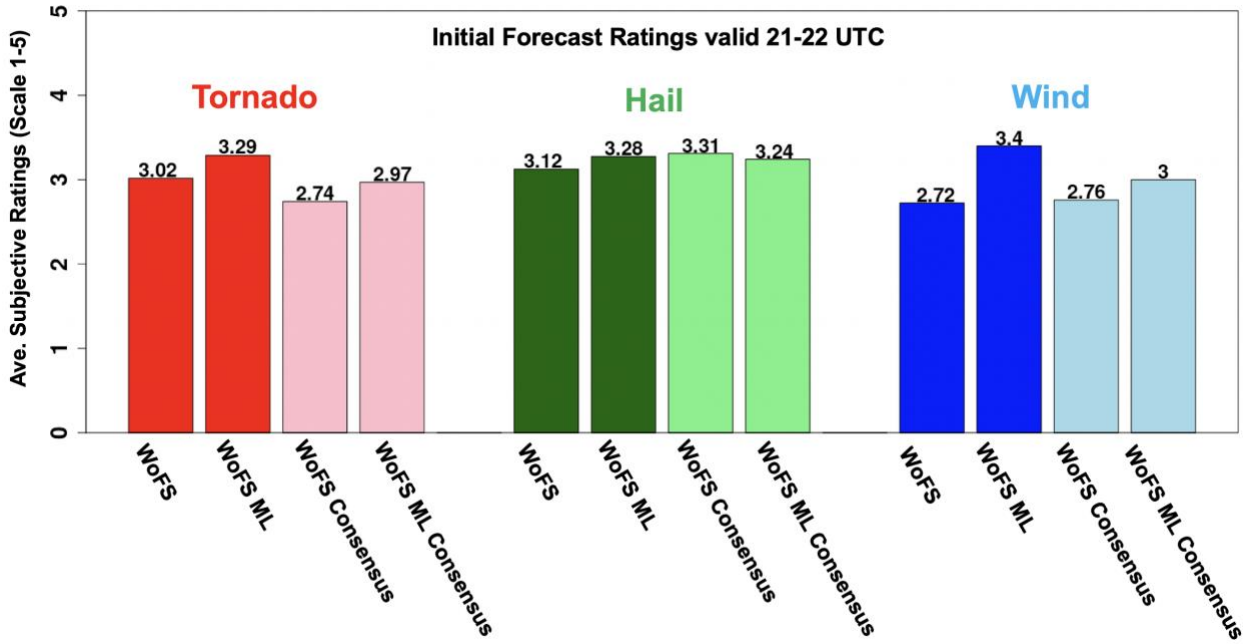


*Figure 75. Same as A1, except for initial forecast ratings valid 2100–2200 UTC.*

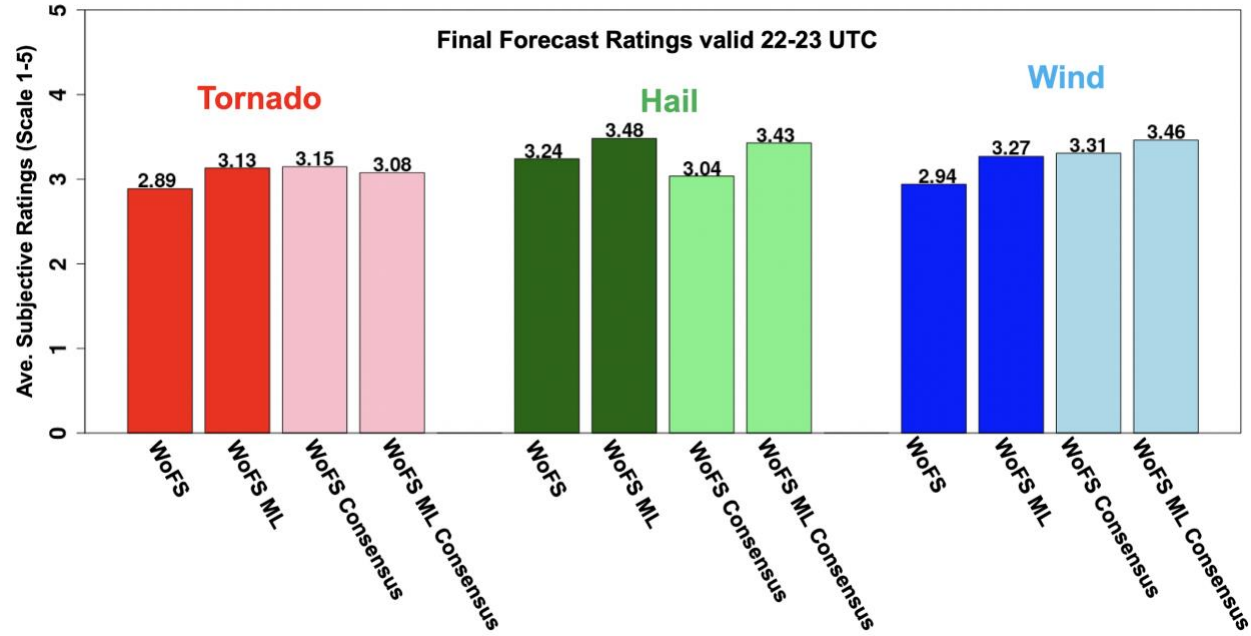


*Figure 76. Same as A1, except for final forecast ratings valid 2100–2200 UTC.*

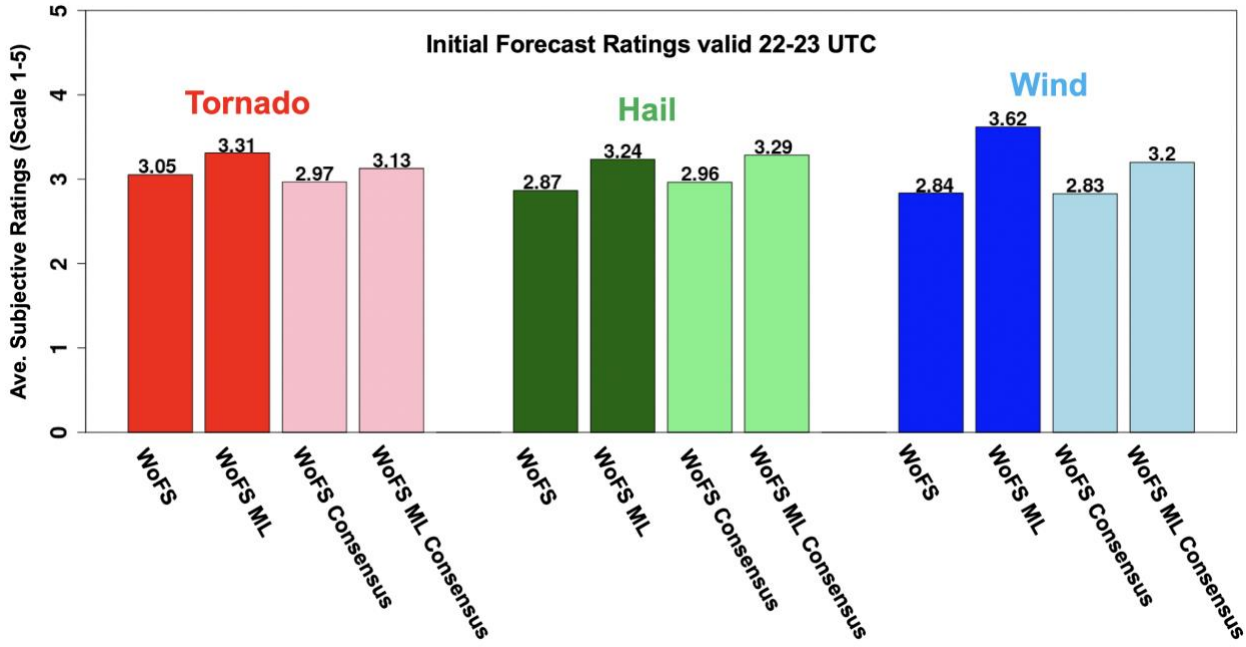


*Figure 77. Same as A1, except for initial forecast ratings valid 2200–2300 UTC.*

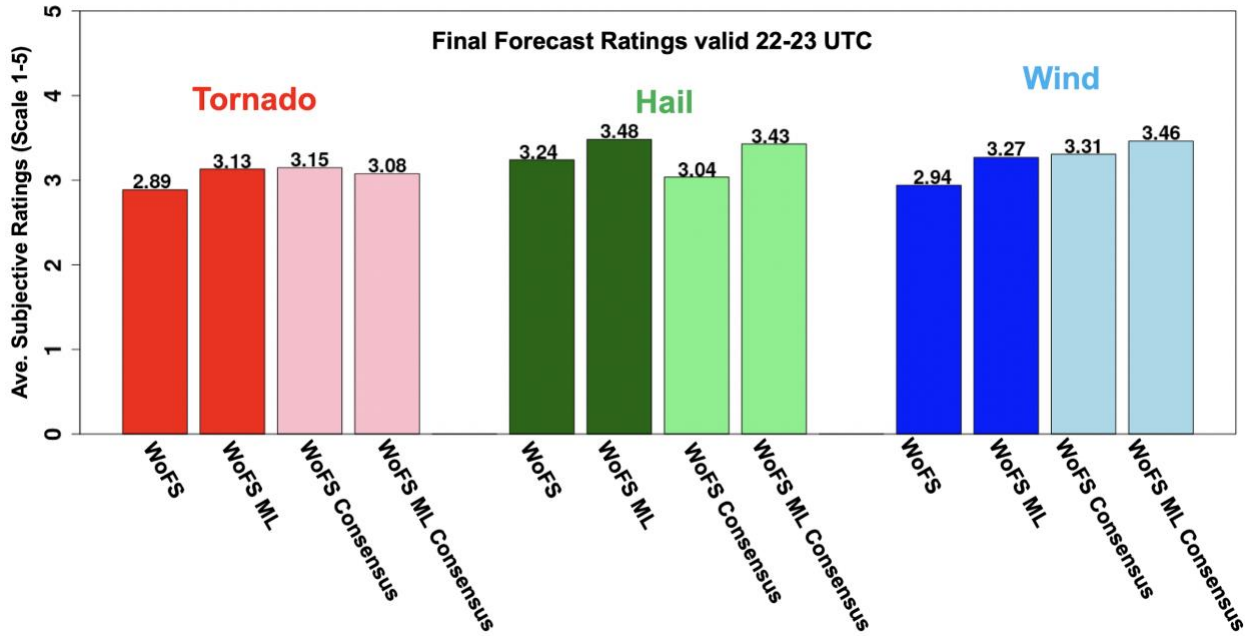


*Figure 78. Same as A1, except for final forecast ratings valid 2200–2300 UTC.*