

Placeholder for DOI



# SPRING FORECASTING EXPERIMENT 2024

Conducted by the

**EXPERIMENTAL FORECAST PROGRAM**

of the

**NOAA HAZARDOUS WEATHER TESTBED**

<https://hwt.nssl.noaa.gov/sfe/2024>

**Hybrid Experiment  
29 April – 31 May 2024**

## **Preliminary Findings and Results**

Adam J. Clark<sup>2,4</sup>, Israel L. Jirak<sup>1</sup>, Tim Supinie<sup>1</sup>, Kent Knopfmeier<sup>2,3</sup>, David Harrison<sup>1,3</sup>, Jake Vancil<sup>1,3</sup>, David Jahn<sup>1,3</sup>, Chris Karstens<sup>1</sup>, Eric Loken<sup>2,3</sup>, Miranda Silcott<sup>2,3</sup>, Ryan Martz<sup>2,3,4</sup>, Nathan Dahl<sup>1,3</sup>, David Imy<sup>2</sup>, Andy Wade<sup>1,3</sup>, Jeffrey Milne<sup>1,3,4</sup>, Kimberly Hoogewind<sup>2,3</sup>, Sean Ernst<sup>1,3</sup>, Joey Picca<sup>1,3</sup>, Matthew Flourney<sup>2,4</sup>, Michael Baldwin<sup>1,3</sup>, Pamela Heinselman<sup>2,4</sup>, Montgomery Flora<sup>2,3</sup>, Joshua Martin<sup>2,3</sup>, Brian Matilla<sup>2,3</sup>, Kristin Calhoun<sup>2</sup>, Thomas Galarneau<sup>2,4</sup>, Derek Stratman<sup>2,3</sup>, Corey Potvin<sup>2,4</sup>, Patrick Skinner<sup>2,3,4</sup>, and Patrick Burke<sup>2</sup>

(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

<b>List of Figures</b> .....	<b>3</b>
<b>List of Tables</b> .....	<b>8</b>
<b>Executive Summary</b> .....	<b>9</b>
<b>1. Introduction</b> .....	<b>10</b>
<b>2. Description</b> .....	<b>12</b>
<b>2.1 Experimental Models and Ensembles</b> .....	<b>12</b>
2.1.1 The Community Leveraged Unified Ensemble (CLUE).....	13
2.1.2 The High-Resolution Ensemble Forecast System version 3 (HREFv3).....	15
2.1.3 NSSL Cloud-Based Warn-on-Forecast System (cb-WoFS).....	16
2.1.4 AI-Driven NWP Emulators.....	16
<b>2.2 Daily Activities</b> .....	<b>17</b>
2.2.1 Forecast and Model Evaluations.....	17
2.2.2 Experimental Forecast Products.....	18
<b>3. Preliminary Findings and Results</b> .....	<b>19</b>
<b>3.1 Evaluation – (C)alibrated Guidance</b> .....	<b>21</b>
3.1.1 (C1-3) Day 1 & 2 4-h SPC Tornado/Hail/Wind Timing Guidance.....	21
3.1.2 (C4) 4-h ML Severe Hazard Guidance .....	24
3.1.3 (C5) Medium-Range 00Z Total Severe.....	25
<b>3.2 Evaluation – (D)eterministic CAMs</b> .....	<b>28</b>
3.2.1 (D1) CLUE: 00Z Day 1 Deterministic Flagships .....	28
3.2.2 (D2) CLUE: 12Z Day 2 Deterministic Flagships .....	32
3.2.3 (D3) CLUE: RRFS vs. HRRR.....	33
3.2.4 (D4) CLUE: RRFS vs. HRRR DA.....	37
3.2.5 (D5) CLUE: 00Z MPAS Resolution Sensitivity.....	40
3.2.6 (D6) CLUE: 1-km vs. 3-km .....	42
<b>3.3 Evaluation – CAM (E)nsembles</b> .....	<b>44</b>
3.3.1 (E1) CLUE: 00Z RRFS vs. HREF .....	44
3.3.2 (E2) CLUE: 12Z Day 1 Ensemble Flagships .....	48
3.3.3 (E3) CLUE: 12Z Day 2 Ensemble Flagships .....	51
3.3.4 (E4) CLUE: Medium-Range Lead Time/Core/Members .....	52
<b>3.4 Evaluation – (A)nalyses</b> .....	<b>56</b>
3.4.1 (A1 & A2) Mesoscale Analysis Background.....	56
3.4.2 (A3) Storm Scale Analysis .....	59
<b>3.5 (A)rtificial (I)ntelligence Evaluations</b> .....	<b>60</b>
3.5.1 (AI1) First-Guess Watch Guidance .....	60
3.5.2 (AI2) Global NWP Emulators .....	62
<b>3.6 (O)utlook Evaluations</b> .....	<b>64</b>
3.6.1 (O4 & O5) Probabilistic 0-1 & 1-2 h Outlooks .....	64
<b>4. Summary</b> .....	<b>69</b>
<b>Acknowledgements</b> .....	<b>73</b>
<b>References</b> .....	<b>74</b>
<b>APPENDIX</b> .....	<b>75</b>

# List of Figures

Figure 1. Scenes from the 2024 NOAA Hazardous Weather Testbed Spring Forecasting Experiment. .... 9

Figure 2. Number of CAMs run for SFEs since 2007. The different colored stacked bars indicate the contributing agencies. .... 13

Figure 3. Highest SPC Day 1 categorical risk for the period 29 April - 31 May. Days when the SFE was not operating are indicated by white transparent boxes. Adapted from the Iowa Environmental Mesonet (IEM; <https://mesonet.agron.iastate.edu>). .... 20

Figure 4. Running totals of annual tornado numbers for the years 1999 - 2024. The year 2024 is indicated by the bold red line. .... 20

Figure 5. Example of the website comparison page for the SPC Severe Timing Guidance during the 2024 HWT SFE. The 4-h tornado probabilities valid 23Z on 6 May to 03Z on 7 May are shown for the Day 2 HREF/GEFS (upper-left panel), Day 2 Nadocast (lower-left panel), Day 1 HREF/GEFS (upper-middle panel), and Day 1 Nadocast (lower-middle panel) Timing Guidance products. The preliminary tornado reports (as of 13Z on 7 May) during this 4-h window are shown as red triangles, and the corresponding 4-h practically perfect tornado probabilities are repeated in the right column. .... 21

Figure 6. Distributions of subjective ratings (1-10; with 10 being best) by SFE participants of the D1 and D2 4-h Tornado Timing Guidance Products based on the HREF/GEFS and Nadocast. .... 22

Figure 7. Same as Fig. 6, except for hail. .... 23

Figure 8. Same as Fig. 6, except for wind. .... 24

Figure 9. Distributions of subjective ratings (1-10; with 10 being best) by SFE participants of the 4-h tornado (left, red), hail (middle, green), and wind (right, blue) probabilities derived from the 12Z HREF (i.e., Nadocast) and 18Z and 22Z WoFS (i.e., WoFS-ML). .... 25

Figure 10. Violin plots showing the distributions of subjective ratings for the C5 00Z Medium-Range Total Severe evaluation for (a) Day 3, (b) Day 4, (c) Day 5, (d) Day 6, and (e) Day 7. The white dots represent the mean ratings for each ensemble, and the white bars indicate the 95% confidence intervals for each mean. The mean ratings are also shown at the bottom of each violin plot. .... 26

Figure 11. Severe weather probabilities valid 30 April 2024 at Day 7 lead time from the ML algorithms (a) GEFS Reforecast, (b) GEFS Operational, and (c) NCAR CAM. (d)-(f), (g)-(i), (j)-(l), and (m)-(o), same as (a)-(c), except for lead times of 6, 5, 4, and 3 days, respectively. Locations of observed storm reports are overlaid. .... 27

Figure 12. Distribution of subjective ratings received by each deterministic flagship model at Day 1 lead times during the 5-week experiment. The white dots represent the mean scores for each ensemble, and the white bars indicate the 95% confidence intervals for each mean. .... 29

Figure 13. Mean performance of five deterministic CAMs with respect to composite reflectivity above 40 dBZ. Ensemble forecasts were compared to MRMS hourly composite reflectivity over the 5-week SFE evaluation period. .... 30

Figure 14. Subjective rating distributions for 2-m temperature (upper-left panel), 2-m dewpoint (upper-right panel), SBCAPE (lower-left panel), and 6-h QPF (lower-right panel) at Day 1 lead times by SFE participants. The white dots represent the mean

scores for each ensemble, and the white bars indicate the 95% confidence intervals for each mean. ....	31
Figure 15. Same as Fig. 12, except for Day 2 lead times. ....	32
Figure 16. Same as Fig. 14, except for Day 2 lead times. ....	33
Figure 17. Example of the 2024 HWT SFE model comparison page for the RRFS control member vs. HRRR valid at 00Z on 17 May 2024. The composite reflectivity 24-hour forecasts are shown for the 00Z HRRR (left panel), the 00Z RRFS control (middle panel), and the observed MRMS composite reflectivity (right panel).....	34
Figure 18. Distributions of subjective ratings (-2 to +2) by SFE participants of the 00Z RRFS control member compared to the HRRR for composite reflectivity and UH (red), updraft speed (purple), 10-m wind speed (blue), and 6-h QPF (green). The ratings represent the RRFS control member compared to the HRRR -2: Much Worse; -1: Slightly Worse; 0: About the Same; +1: Slightly Better; +2: Much Better. ....	35
Figure 19. High-impact severe weather examples from the 2024 HWT SFE where the RRFS control member (middle column) performed much worse than the HRRR (left column) alongside MRMS composite reflectivity observations (right column). The top row shows 18-h forecasts on a High Risk day from 12Z on 6 May (valid at 06Z on 7 May 2024), and the bottom row shows 26-h forecasts on a moderate Risk day from 00Z on 8 May (valid at 02Z on 9 May 2024). ....	35
Figure 20. Performance diagram for accumulated hourly 40-km neighborhood composite reflectivity $\geq 40$ dBZ covering the 24-h convective day (i.e., 12-12Z) over the five-week period of the HWT SFE. The 00Z and 12Z HRRR (red stars) and RRFS (blue circles) performance characteristics are labeled on the diagram. The statistics are only calculated over the primary mesoscale domain used each day for evaluation activities. ....	36
Figure 21. Same as Fig. 18, except for environmental fields of SBCAPE (yellow), 2-m temperature (pink), and 2-m dewpoint (light green).....	37
Figure 22. Violin plots showing the distributions of subjective ratings for the D4 RRFS vs. HRRR DA evaluation. Mean subjective ratings are indicated by the number below each violin. ....	38
Figure 23. Simulated composite reflectivity at 9-h lead times from 21Z 6 May 2024 initializations of (a) HRRR, (b) RRFS, and (c) MRMS observations. Black ovals denote the main area of interest described in the text boxes to the right of each row. (d)-(f) same as (a)-(c), except for 1-h lead times from 00Z 20 May 2024 initializations, and (g)-(i) same as (a)-(c), except for 9-h lead times from 21Z 7 May 2024 initializations. ....	39
Figure 24. Violin plots showing the distribution of subjective ratings from 21Z and 00Z initializations of HRRR and RRFS for CAPE (left), 2-m dewpoint (middle), and 2-m temperature (right). Mean subjective ratings are indicated by the number below each violin. ....	40
Figure 25. Violin plots showing the distribution of subjective ratings for 3- and 4-km grid-spacing configurations of MPAS. Mean subjective ratings are indicated by the number below each violin.....	41
Figure 26. Simulated composite reflectivity at 24-h lead times from 0000 UTC 8 May 2024 initializations of (a) MPAS4 (4-km grid-spacing), (b) MPAS-HN (3-km grid-spacing), and (c) MRMS observations. Black ovals denote the highest impact weather	

event occurring at the time. (d) – (f) same as (a) – (c), except for 20-h lead times from 0000 UTC 28 May 2024 initializations..... 41

Figure 27. Violin plots showing the distribution of subjective ratings for HRRR and NSSL comparisons of (a) convective evolution (i.e., composite reflectivity), (b) 0-2 km AGL UH (tornadoes), and (c) maximum 10-m wind. Mean subjective ratings are indicated by the number below each violin..... 43

Figure 28. Simulated composite reflectivity at 28-h lead times from 00Z 23 May 2024 initializations of (a) HRRR, (b) NSSL1, and (c) MRMS observations. (d)-(e) same as (a)-(b), except for 0-2 km AGL UH, and (f)-(g) same as (a)-(b), except for hourly maximum 10-m wind speed..... 43

Figure 29. Example of the 2024 HWT SFE model comparison page for the REFS vs. HREF valid for the convective day of 6 May 2024. The 24-h neighborhood maximum ensemble probability (NMEP) forecasts of UH are shown for the 00Z HREF (left panel) and the 00Z REFS (right panel). The observed preliminary local storm reports (wind – blue boxes; sig wind – black boxes; hail – green circles; sig hail – black circles) are overlaid in both panels..... 44

Figure 30. Distributions of subjective ratings (-2 to +2) by SFE participants of the 00Z REFS compared to the HREF for updraft helicity (yellow), updraft speed (purple), 10-m wind speed (blue), and composite reflectivity (red). The ratings represent the REFS compared to the HREF -2: Much Worse; -1: Slightly Worse; 0: About the same; +1: Slightly Better; +2: Much Better..... 45

Figure 31. Same as bottom row of Fig. 19, except the left panel is the 26-h forecast (valid at 02Z on 9 May 2024) from REFS member 05 using the saSAS convective parameterization scheme. This member much better captures the ongoing tornadic supercells across southern Tennessee and northern Alabama than the RRFS control member..... 45

Figure 32. Same as Fig. 20, except for 00Z HREF members (stars) and 00Z REFS members (circles). Note that the REFS members using saSAS convective parameterization scheme (green and purple circles) had the best performance of any REFS members for deep convection during the SFE..... 46

Figure 33. Performance diagram (left panel) and reliability diagram (right panel) of the 00Z probabilistic forecasts of composite reflectivity  $\geq 40$  dBZ from the HREF and REFS. The probability forecasts are binned into 10% increments and accumulated hourly for each 24-h convective day during the SFE over the mesoscale domain of interest..... 47

Figure 34. Same as Fig. 30, except for environmental mean fields of 2-m temperature (pink), 2-m dewpoint (light green), and SBCAPE (yellow)..... 47

Figure 35. Distribution of subjective scores received by each ensemble at Day 1 lead times during the 5-week experiment. The white dots represent the mean scores for each ensemble, and the white bars indicate the 95% confidence intervals for each mean..... 49

Figure 36. 24-h neighborhood probabilities of updraft helicity exceeding the 99.85th percentile on 20240524. Red triangles represent tornado reports, blue squares are wind reports, and green circles are hail reports..... 50

Figure 37. (left) Performance and (right) reliability of the five ensemble configurations with respect to NMEP composite reflectivity above 40 dBZ. Ensemble forecasts were

compared to MRMS hourly composite reflectivity over the 5-week SFE evaluation period.....	51
Figure 38. As in Figure 35, but for Day 2 lead times.....	52
Figure 39. Example of multi-panel comparison webpage for the E4 Medium-Range Lead Time/Core/Members evaluation. In each panel, 24 h maximum UH (shaded) and neighborhood probability of UH $\geq$ 99.85th percentile (contours) is displayed. LSRs are also overlaid (wind – blue squares, hail – green circles, and tornado red upside-down triangles; significant reports are filled in black). All forecasts displayed are valid 12Z 21 May –12Z 22 May 2024.....	53
Figure 40. Distributions of subjective ratings for the NCAR MPAS and 5-member NCAR FV3 ensemble subset at Day 3 (left), Day 4 (middle), and Day 5 (right) lead times. Mean subjective ratings are indicated at the bottom of each violin plot. The white dots represent the mean scores for each ensemble, and the white bars indicate the 95% confidence intervals for each mean.....	54
Figure 41. Distributions of subjective ratings for the 10-member NCAR FV3 ensemble for lead times of Day 3 to Day 7. Mean subjective ratings are indicated at the bottom of each violin plot. The white dots represent the mean scores for each ensemble, and the white bars indicate the 95% confidence intervals for each mean. ....	55
Figure 42. Example of the website comparison page for the 3D-RTMA during the 2024 HWT SFE. The 3D-RTMA HRRR baseline is shown in the left panel, the 3D-RTMA RRFS is in the middle panel, and the difference plot (3D-RTMA RRFS - 3D-RTMA HRRR) is shown in the right panel. The 2-m dewpoint analysis valid at 22Z on 6 May 2024 is shaded in the left and middle panels. The difference (analysis-obs) at METAR sites is shown by the size and shading of the dots in the left and middle panels.....	57
Figure 43. Percentage of subjective ratings by SFE participants for each rating category (Much Worse, Slightly Worse, About the Same, Slightly Better, and Much Better) of the 3D-RTMA RRFS compared to the 3D-RTMA HRRR. ....	57
Figure 44. Same as Fig. 43, except for the sfcOA HRRR compared to the 3D-RTMA HRRR.....	58
Figure 45. Example of the website comparison page for the WoFS analyses during the 2024 HWT SFE. The 18Z 6 May – 03Z 7 May accumulated ensemble 90th percentile 80-m wind is shown in the upper-left panel, the ensemble maximum 2-5 km AGL UH in the upper-middle panel, and the ensemble maximum column-maximum updraft speed in the upper-right panel. The observed MRMS composite reflectivity is in the bottom-left panel, observed MRMS midlevel rotation tracks are in the bottom-middle panel, and the MRMS MESH is in the bottom-right panel. In the upper-left panel, the wind damage reports are the black circles while the measured gusts are the open squares shaded by the difference (analysis-obs) of the gust measured at that location.....	59
Figure 46. Distributions of subjective ratings (-2 to +2) by 2024 SFE participants of the WoFS storm-scale analysis for ensemble 90th percentile 80-m winds (blue), 2-5 km AGL UH (light orange), and column-maximum updraft speed (light purple), where the ratings represent how well the WoFS analyses align with the MRMS observed fields and preliminary severe wind reports: -2 - Very Poorly; -1 - Poorly; 0 - Unsure/Neutral, neither poorly nor well; 1 - Well; 2 - Very Well. ....	60

Figure 47. An example of tornado (red) and severe thunderstorm (blue) watches by county predicted by ML guidance (left) and issued by SPC (right) valid for 20240509 2300 UTC. Polygons indicate NWS tornado (red) and severe thunderstorm (blue) warnings. .... 61

Figure 48. Violin plots showing participant ratings of the watch guidance in consideration of: the recommended timing and location (blue), how appropriate the recommended type was for the observed hazards (green), and how closely the guidance matched the SPC-issued watch type (yellow). .... 62

Figure 49. 500 mb height [m] and wind speed [kts] forecasts valid for 20240516 0000 UTC. Models depicted are GraphCast (top left), Pangu-Weather (top center), FourCastNet (bottom left), and GFS (bottom right). The hourly GFS analysis (far right) is included for verification. .... 63

Figure 50. Participant ratings of the operational GFS and AI-generated global NWP 7-day forecasts of (left) 500-mb geopotential height and wind, and (right) 6-h QPF. The white dots represent the mean scores for each ensemble, and the white bars indicate the 95% confidence intervals for each mean. .... 64

Figure 51. Violin plots of subjective participant ratings pooled over (a) all hazards, (b) hail, (c) wind, and (d) tornadoes. In each panel, the first set of violins are pooled over all participant types and lead times, the second set are stratified based on participant type, and the third are stratified by lead time. Violins for forecasts made with (without) WoFS-PHI are colored red (blue). Box-and-whisker plots are shown in black, with the vertical white bar indicating the distribution mean and the white dot indicating the median. “n” indicates the number of ratings used to create each violin. .... 66

Figure 52. Histogram showing the degree to which participants indicated WoFS-PHI influenced (a) where they placed their non-zero forecast probabilities and (b) their forecast probability magnitudes. .... 67

Figure 53. Histogram showing how much trust the WoFS-PHI group participants said they had in (a) the WoFS-PHI machine learning products and (b) the WoFS non-machine-learning products for a given day. .... 67

Figure 54. Histogram showing when participants indicated they incorporated WoFS-PHI into their forecasting workflow. .... 68

Figure 55. Histogram indicating the version of WoFS-PHI participants said provided more useful information: the version trained on both local storm reports (LSRs) and warnings (i.e., LSRs-and-warnings combined) or the version trained only on LSRs (LSRs-only). .... 68

## List of Tables

Table 1. Summary of the 11 unique subsets that comprise the 2024 CLUE. ....	14
Table 2. Average subjective ratings for the GEFS Operational and GEFS Reforecast ML algorithms for 2024 and 2023. Green upward pointing arrows indicate mean ratings that increased from 2023 to 2024.....	28
Table 3. Average subjective ratings for the D4 RRFS vs. HRRR DA evaluation for 2024 and 2023. Green upward pointing arrows indicate differences in mean ratings between HRRR and RRFS (i.e., HRRR minus RRFS) that increased from 2023 to 2024.....	39
Table 4. Average subjective ratings for the NCAR MPAS, NCAR FV3 (5-member), and NCAR FV3 (10-member) ensembles for 2024 & 2023. Green upward pointing arrows indicate mean ratings that increased from 2023 to 2024, red downward arrows indicate a decrease, and the gray sideways arrows indicate little change. ....	56
Table 5. Schedule for Tuesday – Friday. On Mondays, the schedule is similar except the period 9-11am is devoted to training and introductory material. ....	75
Table 6. Schedule for Monday – Thursday evening activity. ....	75



## 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 (CAM) 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. The 2024 SFE was particularly important because the configuration of NOAA's first formally designed convection-allowing model (CAM) ensemble known as REFS (Rapid Refresh Ensemble Forecast System) was finalized in Fall 2023, so the SFE 2024 was vital to conducting a thorough subjective evaluation during the peak severe weather season before a proposed operational implementation in 2025.



Figure 1. Scenes from the 2024 NOAA Hazardous Weather Testbed Spring Forecasting Experiment.

## 1. Introduction

The 2024 Spring Forecasting Experiment (2024 SFE) was conducted from 29 April – 31 May by the Experimental Forecast Program (EFP) of the NOAA/Hazardous Weather Testbed (HWT), and was co-led by the NWS/Storm Prediction Center (SPC) and OAR/National Severe Storms Laboratory (NSSL). Additionally, important contributions of convection-allowing models (CAMs) were made by NOAA collaborators: Global Systems Laboratory (GSL), Environmental Modeling Center (EMC), and Geophysical Fluid Dynamics Laboratory (GFDL); and the National Center for Atmospheric Research (NCAR) and the National Aeronautics and Space Administration (NASA). Participants included over 160 forecasters, researchers, model developers, university faculty, and graduate students from around the world. SFE 2024 marked the second consecutive hybrid SFE, with 68 of the 160 participants contributing remotely. As in previous years, the 2024 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; Rothfus et al. 2018) and Warn-on-Forecast (WoF; Stensrud et al. 2009) visions. Below are goals from the 2024 HWT SFE for product and service improvements and applied science activities.

### Product and Service Improvements:

- Explore the ability to provide enhanced information on the conditional intensity of tornado, wind, and hail events by delineating areas expected to fall within four conditional intensity groups (CIG) defined as: no CIG, CIG 0, CIG 1, and CIG 2, for experimental outlooks covering Days 1, 2, & 3.
- Explore the ability to provide enhanced probabilistic information for Day 4 lead times by producing experimental outlooks for any type of severe hazard similar to current operational Day 3 outlooks.
- Test the utility of WoFS for updating coverage and conditional intensity forecasts in the afternoon.
- 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).
- Assess the utility of machine-learning (ML) guidance known as WoFS-PHI (Warn-on-Forecast System Probabilistic Hazard Information) that combines information from WoFS and ProbSevere by issuing 1-h time window outlooks for individual hazards (tornado, hail, and wind) with and without access to the WoFS-PHI guidance, and surveying participants on their experience using this guidance.

### Applied Science Activities:

- Calibrated Guidance:
  - Explore the skill and utility of SPC Timing Guidance extending into the Day 2 period.

- Compare new gridded WoFS-based ML guidance to the HREF-based Nadocast guidance.
- Evaluate and compare three different methods for producing calibrated severe weather guidance at 3-7 day lead times. Two methods used random forests with predictors from the Global Ensemble Forecast System (GEFS), and the other method used a neural network with environment and storm attribute predictors from an experimental, global-nest ensemble using the Finite Volume Cubed Sphere (FV3) model.
- Deterministic CAMs:
  - Scrutinize differences between the Rapid Refresh Forecast System (RRFS) and the operational High-Resolution Rapid Refresh model (HRRR).
  - Conduct direct comparisons of storm attribute and environment fields in RRFS and HRRR for short lead times in which the data assimilation strongly impacts the forecasts, and longer lead times in which the data assimilation is less important.
  - Compare and assess the skill and utility of the primary deterministic CAMs provided by each SFE 2024 collaborator for Day 1 & 2 lead times.
  - Assess horizontal resolution sensitivity in 3- and 4-km configurations of the Model for Prediction Across Scales (MPAS).
  - Examine whether decreasing horizontal grid-spacing from 3- to 1-km in Weather Research and Forecasting (WRF) model simulations provides benefits for tornado prediction and the strength of convective wind gusts.
- CAM Ensembles:
  - Compare various versions of the RRFS Ensemble Forecast System (REFS) to identify strengths and weaknesses of different configuration strategies for 1- and 2-day lead times. These comparisons were conducted within the framework of the Community Leveraged Unified Ensemble discussed below. Additional baseline comparisons were made using HREFv3.
  - Evaluate and compare the utility of global-with-nest CAM ensemble configurations using the Finite Volume Cubed Sphere (FV3) model and the Model for Prediction Across Scales (MPAS) for medium range severe weather prediction (i.e., Days 3-7).
- Analyses:
  - Compare and assess different versions of the 3D real-time mesoscale analysis (3D-RTMA) system that use different sources for the background first guess along with a surface objective analysis scheme that uses the HRRR as the first-guess background.

- Test WoFS-based analyses of 80-m maximum winds, 2-5 km AGL updraft helicity, and column-maximum updraft speed as a potential verification source for severe weather.
- Artificial Intelligence:
  - Assess three AI-driven global weather predictions driven by GFS initial conditions and compare them to GFS forecasts.
  - Evaluate ML-based, first-guess watch guidance used to predict when and where conditions will be favorable for a severe weather watch.

A suite of state-of-the-art experimental CAM guidance contributed by our large group of collaborators was critical to the 2024 SFE. For the ninth consecutive year, these contributions were formally coordinated into a single ensemble framework called the Community Leveraged Unified Ensemble (CLUE; Clark et al. 2018). The 2024 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 2024 CLUE included 33 members. Most members used 3-km grid-spacing, and the 2024 CLUE configuration allowed for several unique experiments. The 2024 SFE activities also involved testing the WoFS for the eighth consecutive year. Finally, SFE 2024 marks the first year that AI-driven NWP emulators have been assessed. More information on all of the modeling systems run for the 2024 SFE is given in subsequent sections.

This document summarizes the activities, core interests, and preliminary findings of the 2024 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 (Clark et al. 2024). The remainder of this document is organized as follows: Section 2 provides an overview of the models and ensembles examined during the 2024 SFE along with a description of the daily activities, Section 3 reviews the preliminary findings of the 2024 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 52 unique CAMs were run for the 2024 SFE, of which 34 were a part of the CLUE system. Other CAMs outside of the CLUE were contributed by NSSL (WoFS) and EMC (HREFv3). Additionally, forecasts from AI-driven NWP emulators were provided by the Cooperative Institute for Research in the Atmosphere (CIRA). Forecasting activities during the 2024 SFE emphasized the use of CAM ensembles (i.e., HREF, REFS prototypes, and WoFS) in generating experimental probabilistic forecasts of individual severe weather hazards. Additionally, the 2024 CLUE configuration enabled numerous

scientific evaluations focusing on model sensitivities and various ensemble configuration strategies.

To put the volume of CAMs run for 2024 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 since then a steady decrease. The consolidation of members into the CLUE has made this large volume of CAMs more manageable and has facilitated more controlled scientific comparisons.

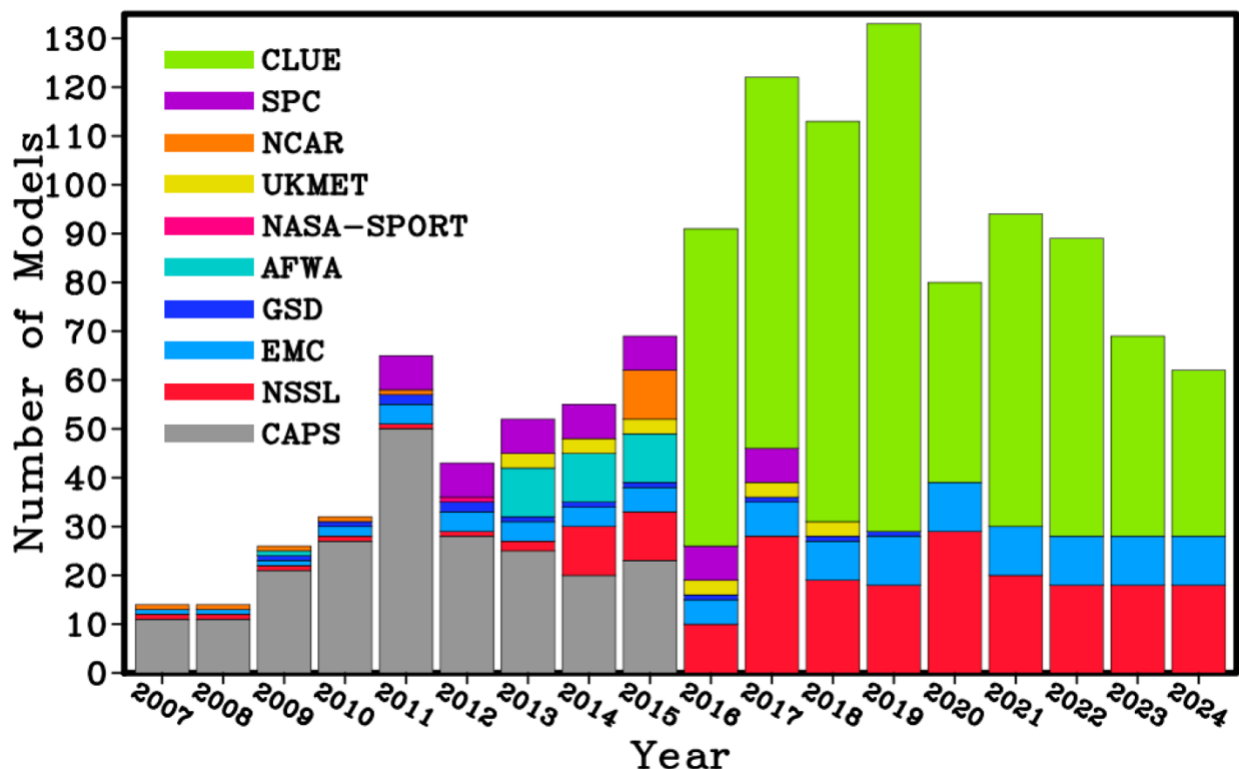


Figure 2. Number of CAMs run for SFEs since 2007. The different colored stacked bars indicate the contributing agencies.

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

The 2024 CLUE is a carefully designed ensemble with subsets of members contributed by NOAA groups at NSSL, GFDL, GSL, and EMC, and the non-NOAA groups of NCAR and NASA. With the exception of a 1-km grid-spacing member that ran over a 2/3 eastern CONUS domain, all CLUE members had a CONUS domain. Most CLUE members used 3-km grid-spacing with the following exceptions: The NASA-FV3 used 2.2-km and MPAS4 used 4-km grid-spacing. Depending on the CLUE subset, forecast lengths ranged from 18 to 192 h. To ensure consistent post-processing, visualization, and verification, CLUE contributors output all model fields to the same grid using the Unified Post Processor (UPP; Earth Prediction Innovation Center 2024). 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 2024 operations plan (Clark et al. 2024). Table 1 provides a summary of each CLUE subset.

Clue Subset	# of mems	IC/LBC perts	Mixed Physics	Data Assimilation	Model Core	Agency	Init. Times (UTC)	Forecast Length (h)	Domain
RRFS	1	none	no	Hybrid 3DEnVar	FV3	EMC/GSL	00, 06, 12, 18	60	CONUS
REFS	5	EnKF	yes	Hybrid 3DEnVar	FV3	EMC/GSL	00, 06, 12, 18	60	CONUS
NSSL1	1	none	no	HRRR ICs	ARW	NSSL	00	36	2/3 CONUS
NSSL-MPAS	3	none	no	HRRR or RRFS ICs	MPAS	NSSL	00, 12	48 or 60	CONUS
GSL-MPAS	5	EnKF	yes	RRFS ICs	MPAS	GSL	06 or 12	54 or 48	CONUS
MPAS4	1	none	no	HRRRICs	MPAS	GSL	00	36	CONUS
GFDL-FV3	1	none	no	GFS cold start	FV3	GFDL	00	126	CONUS
NASA-FV3	1	none	no	GEOS-DA	FV3	NASA	00	120	CONUS
NCAR-FV3	10	GEFS	no	GEFS cold start	FV3	NCAR	00	192	CONUS
NCAR-MPAS	5	GEFS	no	GEFS cold start	MPAS	NCAR	00	132	CONUS

Table 1. Summary of the 11 unique subsets that comprise the 2024 CLUE.

The design of the 2024 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:

#### RRFS vs. HRRR and REFS vs. HREF

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

#### REFS Configuration Strategies

- **Description:** Four different REFS versions based at 1200 UTC are compared: (1) EMC REFS uses the RRFS, all REFS members, and the HRRR initialized at 1200 and 0600 UTC; (2) SPC REFS has a larger proportion of “control” members compared to EMC REFS by only using 1200 UTC REFS perturbed members and including older ( $t-12$ ) time-lagged HRRR and RRFS control members; (3) SPC REFS ARW is similar to SPC REFS, but replaces two of the REFS perturbed

members with HRW ARW members initialized at 1200 and 0000 UTC; and (4) MPAS REFS mirrors SPC REFS, except all RRFs and REFS members are replaced by MPAS members initialized from the same respective analyses as the RRFs and REFS members.

- **Goal:** Identify a strategy within the UFS framework that performs as good as or better than HREFv3, so that it can serve as a replacement in NCEP's production suite.
- **CLUE subsets:** RRFs, REFS, NSSL MPAS, and GSL MPAS

#### Medium-Range CAM Ensembles

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

#### Model Resolution Sensitivities

- **Description:** NSSL initialized a 1-km grid-spacing WRF-ARW configuration driven by the HRRR. GSL ran a 4-km grid-spacing version of MPAS that mirrored the 3-km grid-spacing NSSL-MPAS initialized from HRRR that used Thompson microphysics.
- **Goals:** Examine grid-spacing sensitivity and assess whether enhanced resolution can provide improved severe weather guidance. Assess resolution sensitivities in MPAS. Particular attention was given to depiction of storm structure & mode, as well as low-level rotation diagnostics.
- **CLUE Subsets:** NSSL1, MPAS4, and NSSL MPAS

#### 3D-RTMA Background and Storm-Scale Analyses

- **Description:** Two hourly versions of 3D-RTMA that used different background first-guesses were compared. 15-minute WoFS forecasts of hourly maximum 80-m winds, UH, and updraft speed were compared to Multi-Radar, Multi-Sensor (MRMS) products.
- **Goals:** Assess the impact of the background first guess on the final analysis & gauge whether the 15-minute WoFS forecasts are a viable proxy for observed hazards.
- **Datasets:** 3D-RTMA HRRR, 3D-RTMA RRFs, and WoFS.

### 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/>. The design of HREFv3 originated from the SPC SSEO, which demonstrated skill for six years in the HWT and SPC prior to initial operational implementation in 2017. In HREFv3, the HRW NMMB simulations have been replaced with HRW FV3. The member configuration

diversity in HREFv3 has proven to be a very effective configuration strategy, and it has consistently outperformed all other CAM ensembles examined in the HWT during the last several years.

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

Cloud-based Warn-on-Forecast (cb-WoFS) is the next WoFS iteration, upgraded in 2022 to use current technologies in containerization and cloud computing. The entire WoFS application was rebuilt on top of multiple Platform-as-a-Service and Infrastructure-as-a-Service technologies on the Microsoft Azure platform and the WRF model itself rebuilt to run in containers optimized for HPC. With the cb-WoFS interface, administrators can easily configure the domain and dynamically create an HPC infrastructure for the run, and upon completion, tear it down, thereby reducing costs by only paying for used resources. Another benefit is that as Azure continues to add new, updated computer core types from chip manufacturers, these options are passed down to Azure customers, giving cb-WoFS operators the choice of running on the latest technologies. All parts of WoFS have been rebuilt for scalability: the containerized WRF can be executed on any node, the post-processing is built on high performance queues and containerized, so any number of post-processing jobs can run concurrently.

The cb-WoFS is a rapidly-updating 36-member, 3-km grid-spacing WRF-based ensemble data assimilation and forecast system. The cb-WoFS forecasts are initialized every 30 minutes and used to produce very short-range (0-6/0-3 h at top/bottom of the hour) probabilistic forecasts of individual thunderstorms and their associated hazardous weather phenomena such as supercell hail, high winds, flash flooding, and supercell thunderstorm rotation. The 900-km x 900-km daily cb-WoFS domain targeted the primary region where severe weather was anticipated. For SFE 2024, WoFS had the capability to run over two different domains. A second domain was implemented when there were two separate regions where severe weather was expected (e.g., Midwest and East Coast), or when there was a very large single area for which two domains were needed to cover the entire risk area.

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

### 2.1.4 AI-Driven NWP Emulators

In the last two years, fully AI-based models (i.e., AI NWP emulators) have been developed by the private sector for global weather prediction. This area of research is



advancing rapidly and has the potential to be an advancement in weather prediction since traditional skill measures of the NWP emulators commonly exceed those of the ECMWF's Integrated Forecast System (IFS; ECMWF 2020), the world's most skillful global NWP system. Furthermore, the NWP emulators can produce forecasts in minutes, orders of magnitude faster and with fewer computational resources than traditional NWP systems. The algorithms are trained using large, global, multi-year reanalysis datasets like ERA5 (Hersbach et al. 2020). Several of these algorithms have been made public, and government agencies are beginning to run and train the models themselves. While objective skill measures have been impressive, these NWP emulators have yet to be tested for real-time operational forecasting applications. Thus, during SFE 2024, several of the publicly available algorithms trained using ERA5 data were evaluated. These AI-driven NWP emulators are being run experimentally at the Cooperative Institute for Research in the Atmosphere (CIRA) by providing GFS initial conditions to these AI models. The CIRA forecasts can be viewed at: <https://aiweather.cira.colostate.edu/> and a specialized web-viewer for the model evaluation activity was implemented at: [https://hwt.nssl.noaa.gov/sfe\\_viewer/2024/ai](https://hwt.nssl.noaa.gov/sfe_viewer/2024/ai). The emulators included (1) FourCastNetv2 (Pathak et al. 2022), (2) Pangu-Weather (Bi et al. 2022), and (3) GraphCast (Lam et al. 2022).

## 2.2 Daily Activities

SFE 2024 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 (Tables A2 & A3). 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 2024 featured a period of formal evaluations from 9-11am CDT Tuesday-Friday (except for the last week which was Wednesday-Friday), for a total of 19 days of evaluation. The evaluations involved comparisons of different ensemble diagnostics, CLUE ensemble subsets, HREFv3, WoFS, and AI-driven NWP emulators. Additionally, the evaluations of yesterday's experimental forecasts products were conducted during this time, which involved comparing the experimental products to observed local storm reports (LSRs), NWS warnings, and Multi-Radar, Multi-Sensor (MRMS; Smith et al. 2016) radar reflectivity and maximum estimated size of hail (MESH). Participants were split into Groups 1, 2, and 3, and each conducted a separate set of model evaluations. These groups were hybrid meaning that they contained a mix of in-person and virtual participants. The evaluations were categorized as "(C)alibrated Guidance",

“(D)eterministic CAMs”, “CAM (E)nsembles”, “(A)nalyses”, “(A)rtificial (I)ntelligence Evaluations”, or “(O)utlooks”. The letter in parentheses combined with a number was used to label the individual evaluations in each category (e.g., E1 refers to the first CAM Ensemble evaluation). Each evaluation group conducted a mix of evaluations from each category. Participants rotated through each evaluation group at least once. Participants worked on all the surveys individually, with short discussion periods after completion of each survey. SFE facilitators were available to answer any questions, troubleshoot issues, and discuss subjective impressions of the day.

## 2.2.2 Experimental Forecast Products

The experimental forecasts covered a limited-area domain typically encompassing the primary severe threat area with a domain based on existing SPC outlooks and/or where interesting convective forecast challenges were expected. An exception was the Day 3 & 4 outlooks, which covered the entire CONUS. There were two periods of experimental forecasting activities during SFE 2024. The first occurred from 11:00am – 12:30pm CDT and focused on generating probabilistic outlooks for individual hazards for Days 1-3, as well as more precise information on the intensity of specific hazards. The Day 4 outlooks only covered total severe (i.e., no individual hazards or conditional intensity forecasts). Participants were split into three groups: (1) In-Person R2O, (2) In-Person Innovation, and (3) Virtual. As the naming convention suggests, in-person participants were in R2O and Innovation groups, while all virtual participants were in the Virtual group. The In-Person R2O group issued products for Day 1, the Virtual group issued products for Day 2, and the In-Person Innovation group issued products for Days 3 & 4.

In all groups, the morning forecasts were done collectively. The individual hazard forecasts mimicked the SPC operational Day 1 & 2 Convective Outlooks by producing individual probabilistic coverage forecasts of large hail, damaging wind, and tornadoes within 25 miles (40 km) of a point. The Day 1 outlooks covered the period 1800 UTC to 1200 UTC the next day, while the Days 2, 3, & 4 outlooks covered 1200 – 1200 UTC periods. Additionally, for experimental outlooks covering Days 1, 2, & 3, conditional intensity forecasts of tornado, wind, and hail were issued by delineating conditional intensity groups, where the reports are expected to follow pre-defined intensity distributions. These conditional intensity forecasts are similar to those issued during SFEs 2019-2023. The four possible conditional intensity groups (CIG) included: no CIG, CIG 0, CIG 1, and CIG 2. In plain language, no CIG is when no severe weather is expected, CIG 0 refers to a typical severe weather day, where significant severe weather is unlikely, CIG 1 areas indicate where significant severe weather is possible, and CIG 2 areas indicate where high impact significant severe weather is expected. All groups had access to all available operational and experimental guidance products for issuing their outlooks.

The second period of experimental forecasting activities occurred during the 2-4pm CDT time period. From 2-2:15pm CDT, a weather briefing led by Dave Imy was conducted for all participants during which an update on current weather was given. In

the In-Person R2O group, the 2:15-3:15pm CDT time period was devoted to an activity in which each participant created their own Mesoscale Discussion (MD) Product using WoFS and other available observations and CAM guidance within the SFE Drawing Tool. Then, during the 3:15-4pm CDT time period, each In-Person R2O participant used WoFS and other available guidance to update the Day 1 individual hazard coverage and conditional intensity forecasts done earlier as a group for the period 2100 – 1200 UTC.

During the 2:15-4pm CDT time period in the In-Person Innovation Group and Virtual Group, participants were split into two groups and generated experimental 0-1 and 1-2 h hazard probabilities for tornado, wind, and hail. The first set of forecasts was due at 3pm CDT and covered 3-4pm (0-1 h) and 4- 5pm (1-2 h). The second set of forecasts was due at 4pm and covered 4-5pm (0-1 h) and 5-6pm (1-2 h). One group issuing these forecasts had access to WoFS PHI (and any other guidance). The other group did not have access to WoFS PHI. In both groups, two expert forecasters were assigned whose forecasts were evaluated the next day. All other participant forecasts were combined into consensus forecasts, which were also evaluated the next day. A small group of pre-determined virtual “evening forecasters” continued this activity into the evening. They took a break from 4-5pm, then from 5-6pm issued 0-1 h and 1-2 h forecasts valid 6-7pm and 7-8pm, respectively. They repeated the activity one more time from 6-7pm with all forecasts shifted one hour later, and finally from 7-8pm finished by completing a survey on their use of WoFS and WoFS PHI.

### **3. Preliminary Findings and Results**

An important consideration for contextualizing the results of SFE 2024 is the weather regime. SFE 2024 coincided with one of the most active Mays on record, which was dominated by strongly forced severe weather events. This pattern was a stark contrast to SFE 2023, which had average levels of severe weather activity and was dominated by weakly forced events. A calendar displaying the maximum Day 1 categorical risk issued by SPC during the 2024 SFE is displayed in Figure 3. All but 5 days of the SFE had enhanced risks or higher, and the tornado activity in May largely contributed to 2024 being a top 5 year for January – May tornado numbers (Fig. 4). These strongly forced events resulted from well-defined synoptic-scale weather systems that are inherently more predictable than more subtle and less well-defined systems with weaker forcing. Thus, it would be expected that with all else being equal, experimental and operational SFE 2024 model guidance would perform better than SFE 2023. Wherever possible in the subsequent results, we will highlight performance relative to the previous year.



### Highest SPC Day 1 Convective (Categorical) Outlook for Contiguous US

Valid 01 May 2024 - 31 May 2024. Days since by threshold: ENH - 2 Days, HIGH 25 Days, MDT 5 Days, SLGT 0 Days

May 2024

SUN	MON	TUE	WED	THU	FRI	SAT
	29 MRGL	30 ENH	1 ENH	2 ENH	3 ENH	4 ENH
5 SLGT	6 HIGH	7 ENH	8 MDT	9 ENH	10 ENH	11 SLGT
12 SLGT	13 ENH	14 SLGT	15 ENH	16 ENH	17 SLGT	18 SLGT
19 MDT	20 ENH	21 MDT	22 ENH	23 ENH	24 ENH	25 MDT
26 MDT	27 SLGT	28 ENH	29 SLGT	30 ENH	31 SLGT	

Figure 3. Highest SPC Day 1 categorical risk for the period 29 April - 31 May. Days when the SFE was not operating are indicated by white transparent boxes. Adapted from the Iowa Environmental Mesonet (IEM; <https://mesonet.agron.iastate.edu>).

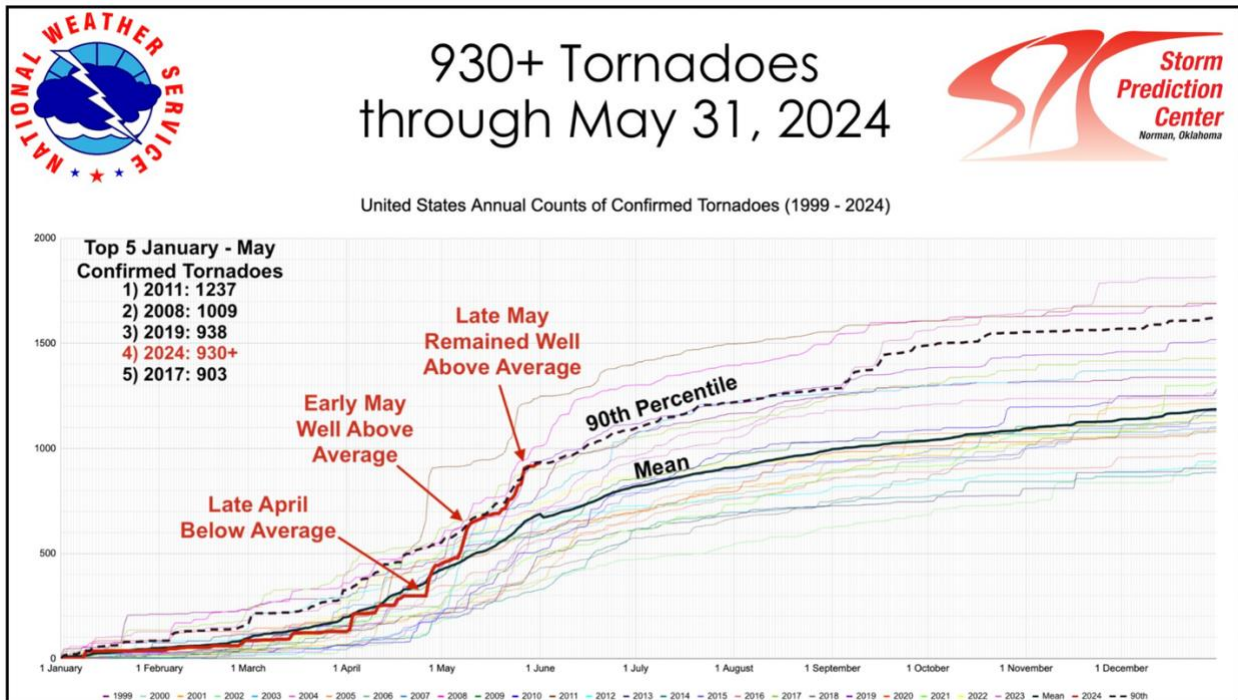


Figure 4. Running totals of annual tornado numbers for the years 1999 - 2024. The year 2024 is indicated by the bold red line.

### 3.1 Evaluation – (C)alibrated Guidance

#### 3.1.1 (C1-3) Day 1 & 2 4-h SPC Tornado/Hail/Wind Timing Guidance

In an effort to add more specific temporal information to the Day 1 Outlook, SPC has developed Severe Timing Guidance products, which are hourly 4-h severe weather probabilities through the convective day. The Severe Timing Guidance products are consistent with and constrained by the human-issued SPC Convective Outlooks and use HREF-based guidance to disaggregate the probabilities throughout the convective day. The current operational Day 1 SPC Timing Guidance probabilities leverage the operational HREF/SREF calibrated hazard probabilities, but with the planned retirement of the SREF in the coming years, other HREF-based guidance products (i.e., HREF/GEFS and Nadocast) were tested in the algorithm to determine the effect on the probabilistic output. In addition this year, the SPC Severe Timing Guidance was created for Day 2 based on the 1730Z SPC Day 2 Convective Outlook and the 12Z HREF (i.e., forecast hours 24-48) using both HREF/GEFS and Nadocast guidance as inputs to the algorithm. The respective SPC Severe Timing Guidance products for Day 1 were also evaluated alongside the Day 2 products (i.e., with greater lead time) for comparison (Fig. 5). A single rating (1-10) was given by participants for each product for the performance over the entire convective day.

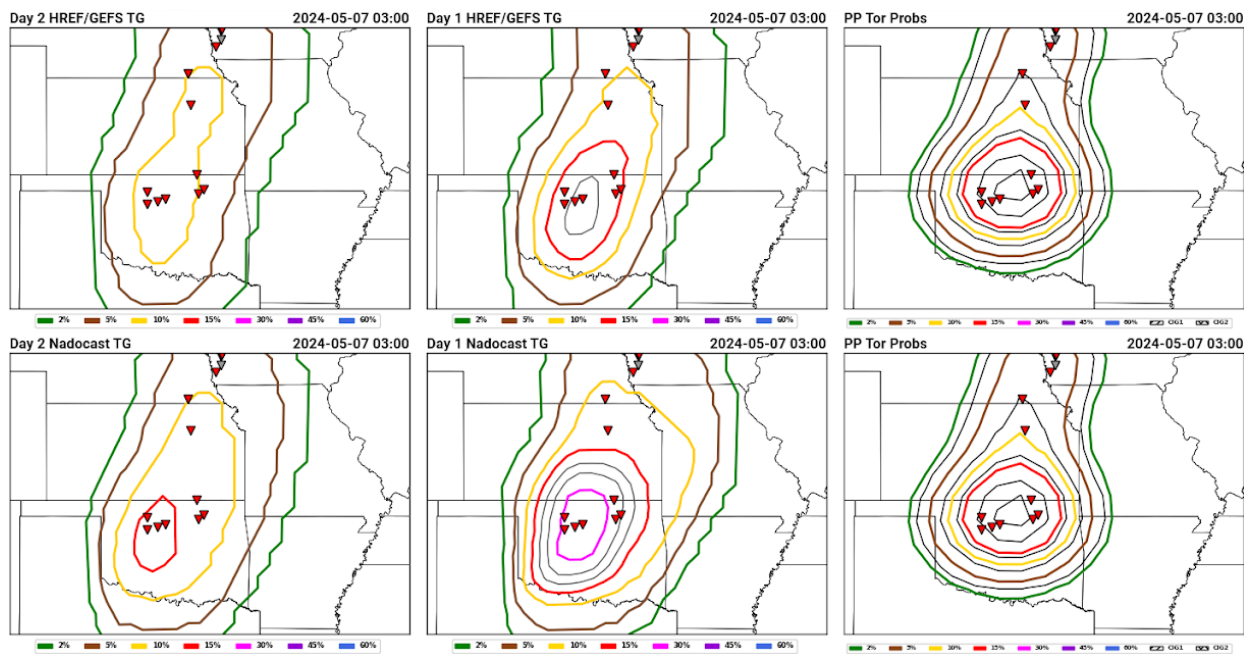


Figure 5. Example of the website comparison page for the SPC Severe Timing Guidance during the 2024 HWT SFE. The 4-h tornado probabilities valid 23Z on 6 May to 03Z on 7 May are shown for the Day 2 HREF/GEFS (upper-left panel), Day 2 Nadocast (lower-left panel), Day 1 HREF/GEFS (upper-middle panel), and Day 1 Nadocast (lower-middle panel) Timing Guidance products. The preliminary tornado reports (as of 13Z on 7 May) during this 4-h window are shown as red triangles, and the corresponding 4-h practically perfect tornado probabilities are repeated in the right column.

For the Tornado Timing Guidance, using Nadocast as the input to the algorithm produced the highest-rated timing guidance products for both Day 1 and Day 2 (Fig. 6). Based on participant comments the Nadocast versions tended to have higher probabilities than the HREF/GEFS versions and hold onto probabilities into the overnight period, which tended to help with POD and boost the subjective ratings. As expected, the rating distributions for both HREF/GEFS and Nadocast shifted higher when comparing Day 1 to the Day 2 guidance, but the Day 2 Timing Guidance generally seemed reasonable on most days.

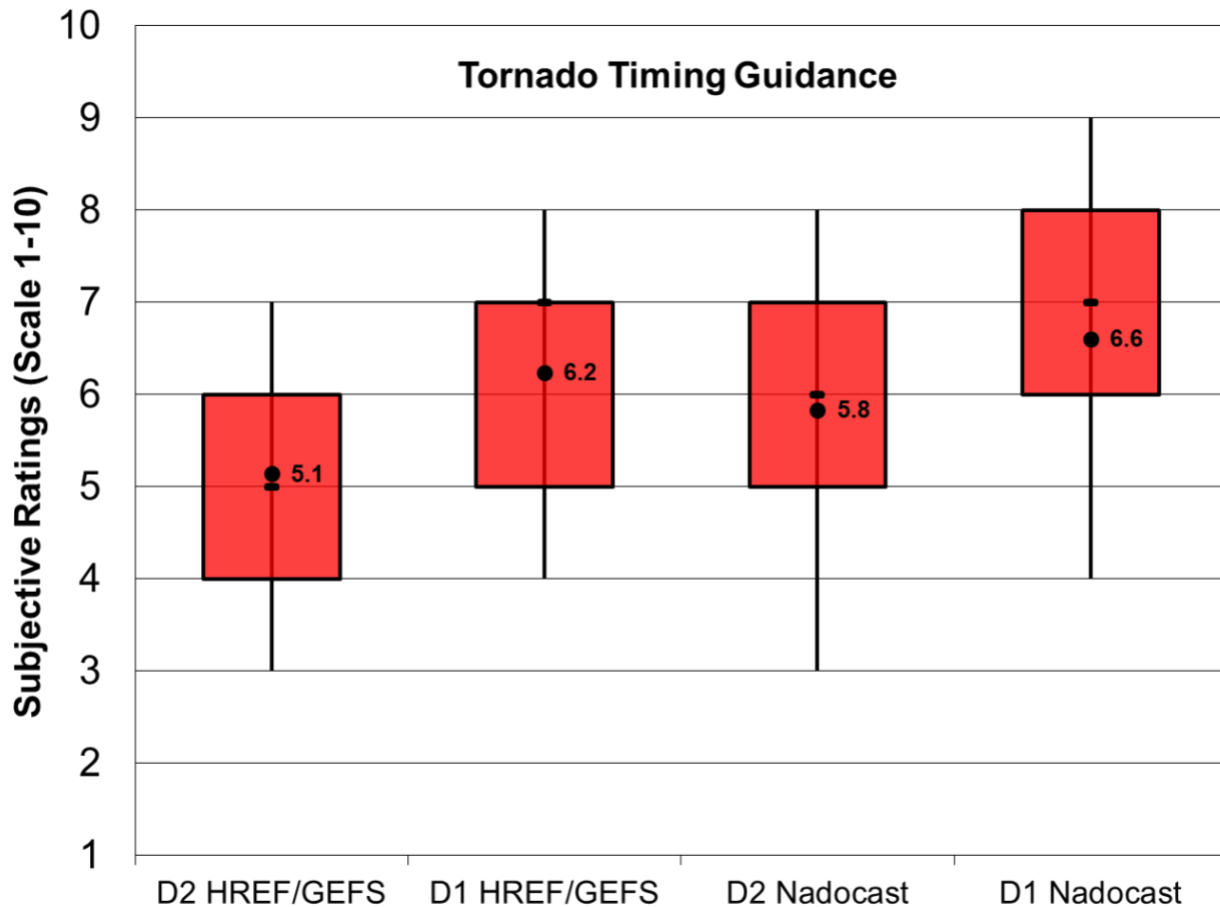


Figure 6. Distributions of subjective ratings (1-10; with 10 being best) by SFE participants of the D1 and D2 4-h Tornado Timing Guidance Products based on the HREF/GEFS and Nadocast.

For the Hail Timing Guidance, the HREF/GEFS products were rated very similarly to the Nadocast products on both Day 1 and Day 2. The subjective ratings for HREF/GEFS Timing Guidance products were higher overall for hail than for tornado while the Nadocast Timing Guidance products were rated similarly for hail and tornado (cf. Figs. 6 and 7). Overall, the Day 1 hail products received about a one-point higher rating on average over the Day 2 products during the SFE (Fig. 7).

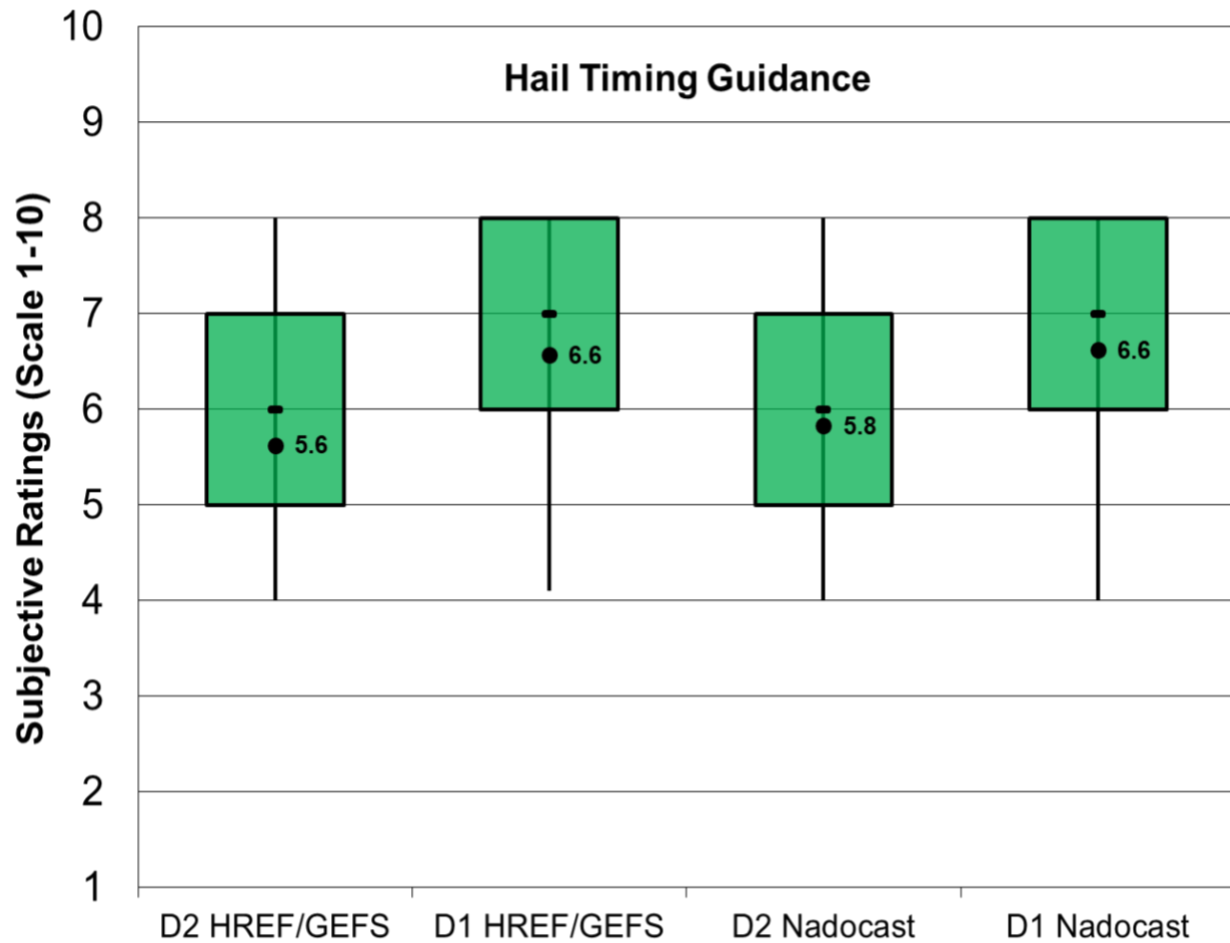


Figure 7. Same as Fig. 6, except for hail.

The subjective ratings for the Timing Guidance for wind were similar to the hail ratings, but slightly higher on Day 1 with the Nadocast version being narrowly favored (Fig. 8). Overall, the Timing Guidance products were relatively insensitive to the guidance input with Nadocast being slightly favored over HREF/GEFS, owing to higher probabilities resulting in higher POD. With the Day 2 Timing Guidance products being evaluated for the first time, the results were favorable. The probabilistic output looked reasonable and participants noted that there would be value to having that guidance be available for IDSS purposes.

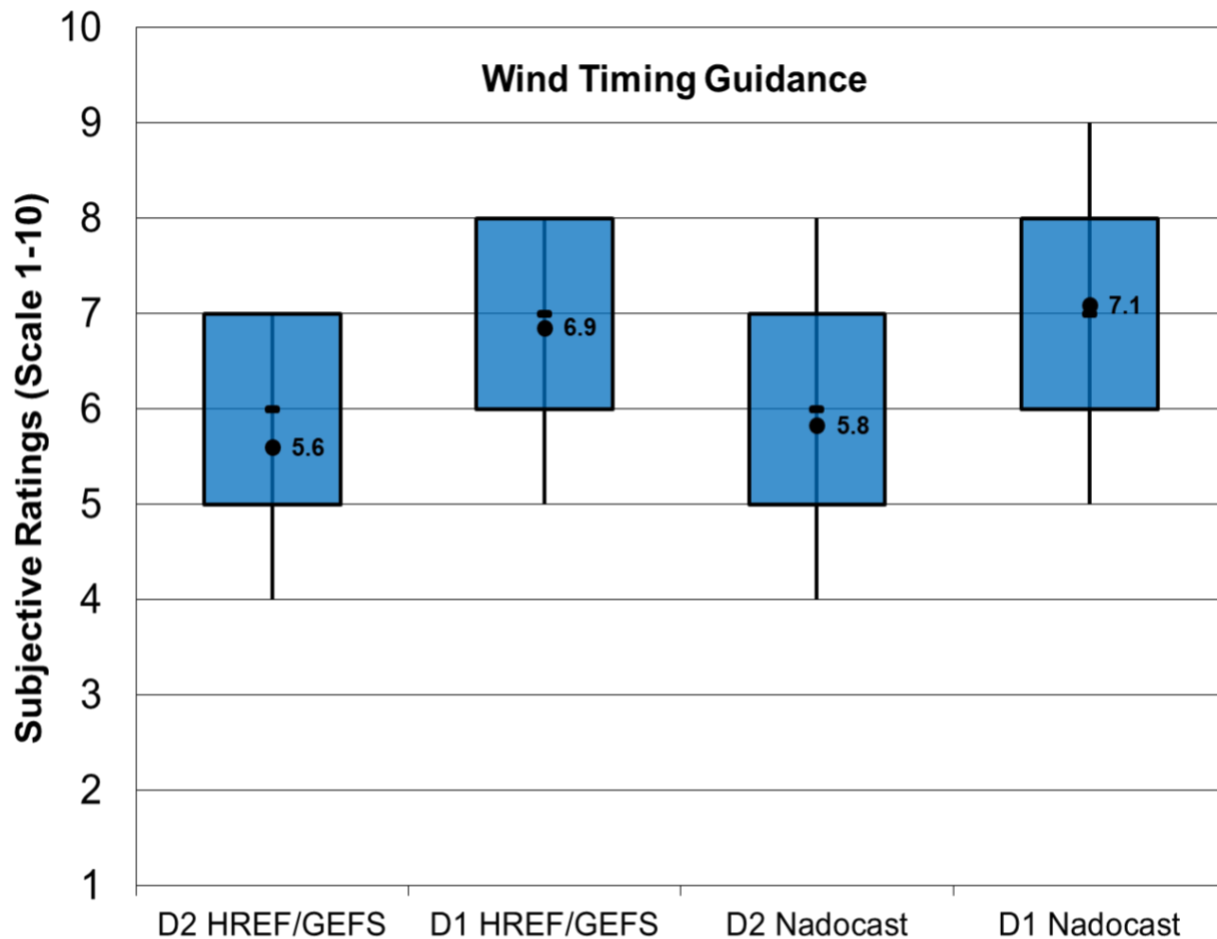


Figure 8. Same as Fig. 6, except for wind.

### 3.1.2 (C4) 4-h ML Severe Hazard Guidance

A new 4-h WoFS-based ML severe hazard probability product was assessed during the SFE. Two 4-h periods were examined: 1) 20-00Z from the 18Z WoFS run and 2) 00-04Z from the 22Z WoFS run. These products were evaluated alongside 12Z HREF-based Nadocast, which also uses an ML algorithm, over corresponding 4-h periods. This evaluation was intended to determine whether ML guidance from updated afternoon WoFS runs (18Z & 22Z) could improve upon existing, older 12Z-based HREF runs. Overall, the WoFS-ML guidance was rated slightly lower than the HREF-based Nadocast ML guidance for all hazards (Fig. 9). Participant comments noted that both forms of guidance could be hit or miss on a given day. The probabilities tended to be overconfident, but if the event was captured, then it was a major success. It should be noted that this was the first year of the WoFS 4-h ML probabilities, so additional development and refinement may help improve the product in future iterations.



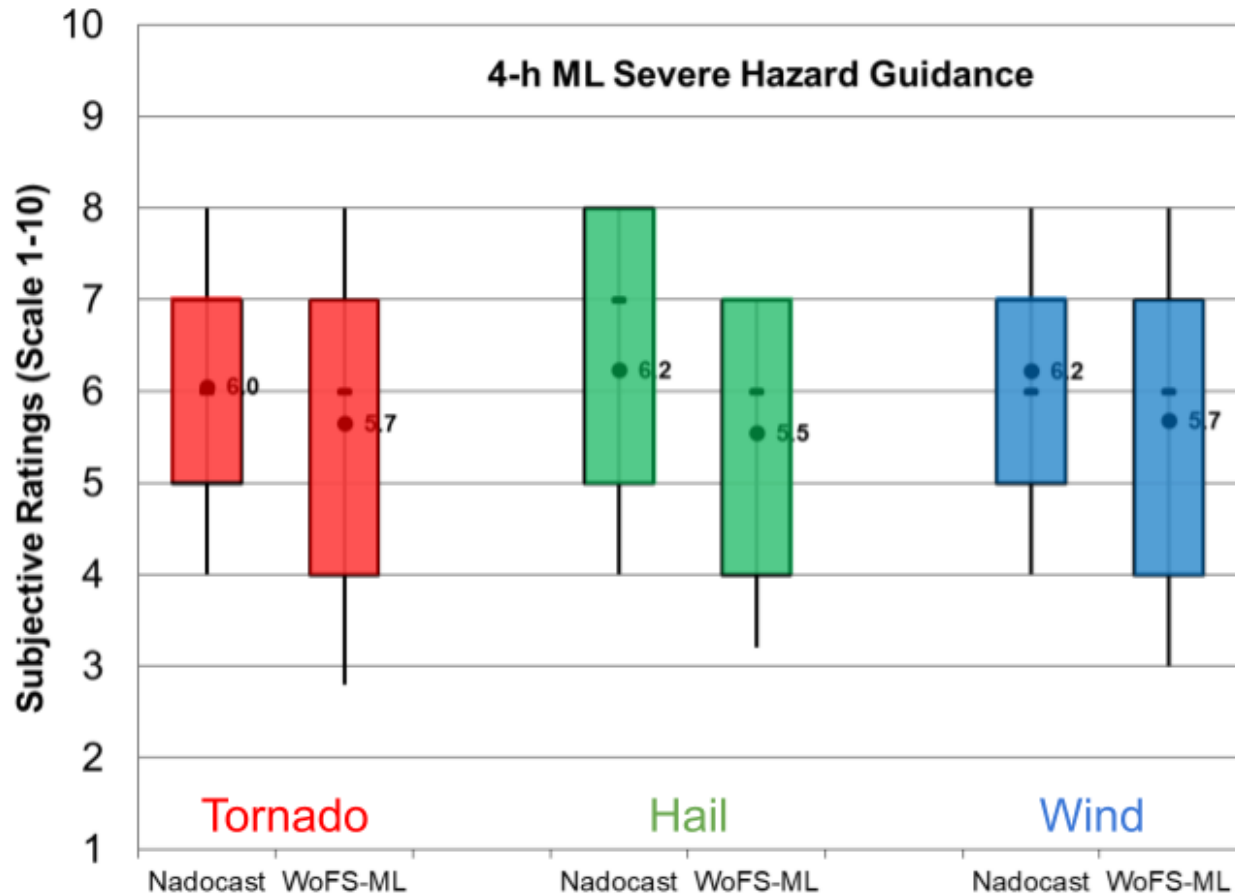


Figure 9. Distributions of subjective ratings (1-10; with 10 being best) by SFE participants of the 4-h tornado (left, red), hail (middle, green), and wind (right, blue) probabilities derived from the 12Z HREF (i.e., Nadocast) and 18Z and 22Z WoFS (i.e., WoFS-ML).

### 3.1.3 (C5) Medium-Range 00Z Total Severe

Three algorithms for producing extended-range forecasts of total severe (tornado, wind, or hail) were assigned subjective ratings for Days 3-7. GEFS Reforecast ML is a random forest algorithm from Colorado State University that uses environmental predictors from GEFS ensemble medians and is trained using 5-member GEFS reforecasts. GEFS Reforecast ML has been tested in previous SFEs with very promising results; more info can be found in Hill et al. (2023). GEFS Operational ML from NSSL is similar to GEFS Reforecast ML, but it is trained using just over 2 years of the most recent GEFS Operational forecasts, which contain 31 members (Clark et al. 2024). Finally, NCAR CAM ML uses a neural-network algorithm with environment and storm attribute predictors from NCAR’s 10-member FV3-based global-nest ensemble.

At the day 3 lead time, GEFS Operational ML and GEFS Reforecast ML performed similarly, with differences in mean subjective ratings that were not statistically significant. For day 4-7 lead times, GEFS Operational ML clearly outperformed GEFS Reforecast ML, with statistically significant differences for days 4, 5, and 7. At all lead times, NCAR

CAM ML was the worst performing algorithm with differences from the GEFS-based ML products that were significantly significant (Fig. 10). An example case illustrating typical differences between algorithms is shown in Figure 11. Common themes from survey comments were that GEFS Operational ML often generated higher probabilities at longer lead times that often corresponded quite well with observed severe weather. NCAR CAM ML consistently under-predicted probabilities at all lead times, especially the longer ranges, which was likely because the 24-h probabilities were computed by combining individual 4-h probabilities for which predictability is much lower. Future iterations of NCAR CAM ML will derive probabilities using longer time windows. Finally, it was often noted that for individual cases there was not a steady increase in skill with decreasing lead time. For example, it was not unusual for the day 4 or 5 lead times to perform better than day 3.

The mean subjective ratings in SFE 2024 were significantly higher in SFE 2024 relative to SFE 2023 (Table 3). Since the formulation of these algorithms did not change, the higher SFE 2024 ratings can be attributed to the more predictable weather regime of SFE 2024.

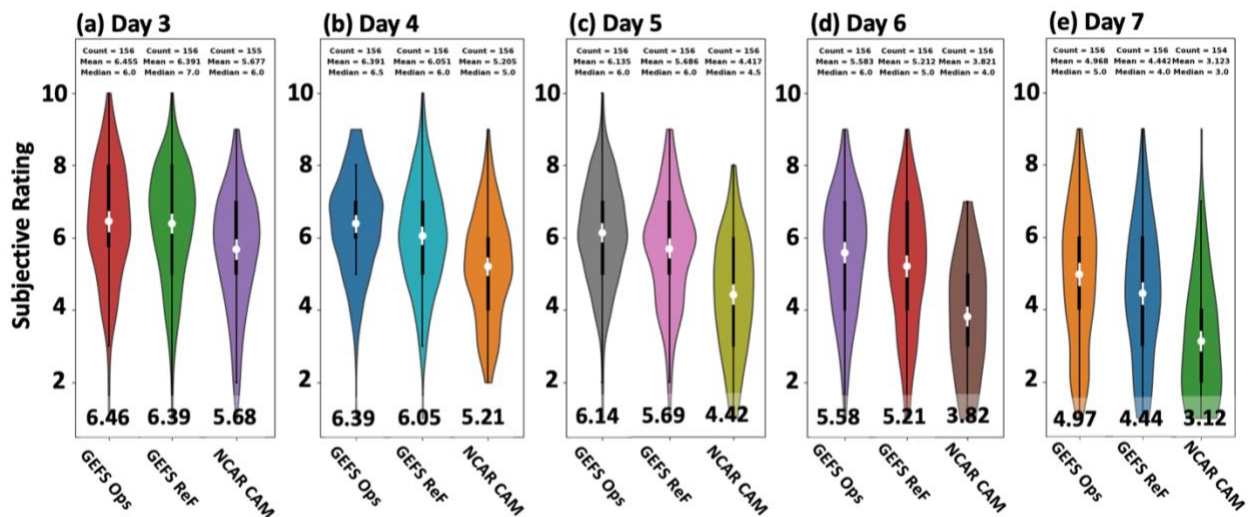


Figure 10. Violin plots showing the distributions of subjective ratings for the C5 00Z Medium-Range Total Severe evaluation for (a) Day 3, (b) Day 4, (c) Day 5, (d) Day 6, and (e) Day 7. The white dots represent the mean ratings for each ensemble, and the white bars indicate the 95% confidence intervals for each mean. The mean ratings are also shown at the bottom of each violin plot.

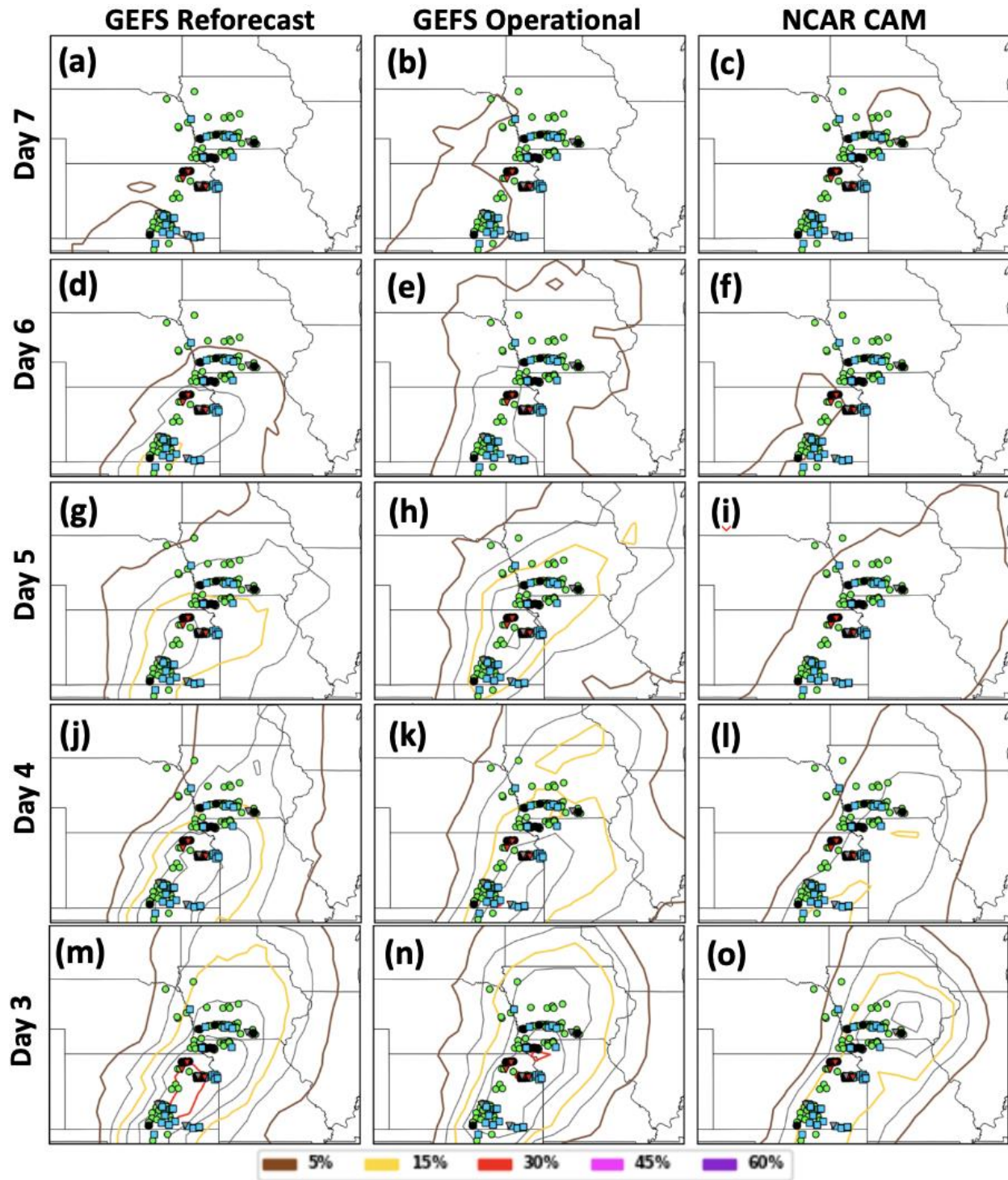


Figure 11. Severe weather probabilities valid 30 April 2024 at Day 7 lead time from the ML algorithms (a) GEFS Reforecast, (b) GEFS Operational, and (c) NCAR CAM. (d)-(f), (g)-(i), (j)-(l), and (m)-(o), same as (a)-(c), except for lead times of 6, 5, 4, and 3 days, respectively. Locations of observed storm reports are overlaid.











		2024	2023
Day 3	GEFS Operational 	6.46	6.13
	GEFS Reforecast 	6.39	5.76
Day 4	GEFS Operational 	6.39	5.66
	GEFS Reforecast 	6.05	4.84
Day 5	GEFS Operational 	6.14	5.51
	GEFS Reforecast 	5.69	4.28
Day 6	GEFS Operational 	5.58	5.21
	GEFS Reforecast 	5.21	3.43
Day 7	GEFS Operational 	4.97	4.64
	GEFS Reforecast 	4.44	3.78

Table 2. Average subjective ratings for the GEFS Operational and GEFS Reforecast ML algorithms for 2024 and 2023. Green upward pointing arrows indicate mean ratings that increased from 2023 to 2024.

### 3.2 Evaluation – (D)eterministic CAMs

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

This evaluation focused on comparing deterministic convection-allowing models which have been iterated on by their respective agencies and are relatively advanced in their development. Models included in this year’s experiment consist of the GFDL FV3, NSSL MPAS HT, RRFs, and NASA GEOS FV3, each representing unique combinations of dynamical cores, data assimilation strategies, and physics parameterizations. The operational HRRRv4 was also included as a point of comparison for the other models. Only the 00Z model initializations were assessed in this evaluation, and participants were asked to only look at forecast hours f12 - f36 when completing their surveys. This limited the evaluation to the Day 1 (1200Z – 1200Z) time period. All models were evaluated blindly such that participants were not able to see which model produced which forecast. Additionally, the order of each model was randomized daily so that participants could not anticipate a model being in the same panel day-to-day. Models were unblinded following discussion of the results and after all surveys were submitted.

Participants compared the reflectivity and 2–5 km AGL updraft helicity (UH) fields from each configuration, along with one environmental variable randomly selected from 2-m temperature, 2-m dewpoint, or surface-based convective available potential energy (SBCAPE). All participants then assessed and compared the 6-h quantitative precipitation forecast (QPF) produced by each model. Each model and field were independently rated on a scale of 1 (Very Poor) to 10 (Very Good), and participants had the option to provide additional insights via an open response box following each survey question.

The HRRRv4 received the highest ratings on average when evaluating the structure, evolution, location, and timing of simulated storm reflectivity and UH at Day 1 lead times (Fig. 12). The operational model received a mean rating of 6.7, followed by the NSSL MPAS HT at 6.3. The RRFS received a mean rating of 5.5, the NASA GEOS FV3 was given a 5.2, and the GFDL FV3 saw the lowest mean rating of 4.9. The HRRRv4 mean rating was found to be significantly higher (at the 95% confidence level) than that of the RRFS, GFDL FV3, and NASA GEOS FV3, but was not significantly different from the NSSL MPAS HT. The HRRRv4, NSSL MPAS HT, and RRFS each received a maximum score of 10 at some point during the experiment, while the NASA GEOS FV3 and GFDL FV3 had a maximum rating of 9. SFE participants were particularly critical of the RRFS reflectivity structure which frequently failed to accurately capture the intensity and mode of severe convection. Instead, the model's composite reflectivity field was often characterized by spurious low-intensity values (25-35 dBZ), which seemed to frequently preclude more realistic simulated thunderstorm development. This unusual behavior is likely related to the implementation of the Grell-Freitas (GF) deep convection parameterization scheme, which was newly added to the RRFS this year. These changes appear to have been a net detriment to the model, with this year's mean subjective score falling 0.2 points lower than last year's evaluation while the HRRR ratings were 0.3 points higher on average than last year's ratings.

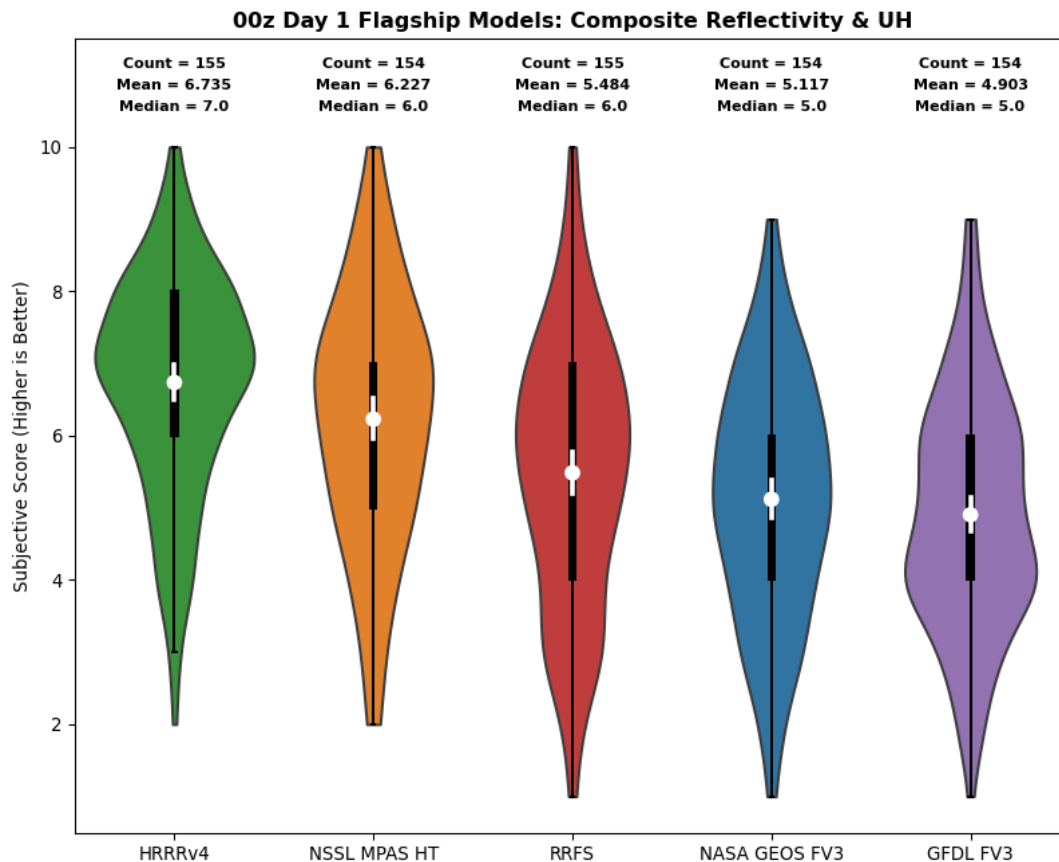


Figure 12. Distribution of subjective ratings received by each deterministic flagship model at Day 1 lead times during the 5-week experiment. The white dots represent the mean scores for each ensemble, and the white bars indicate the 95% confidence intervals for each mean.

The subjective ratings are further supported by preliminary objective verification metrics calculated over the 5-week SFE evaluation period and domains. This verification compared areas of simulated REFC  $\geq 40$  dBZ to hourly MRMS REFC to compute contingency table metrics at each forecast hour. These metrics were then accumulated across the entire experiment and plotted on a performance diagram as shown in Fig. 13. As in the subjective evaluations, the HRRR demonstrated the best forecast performance of the five models in terms of CSI and POD, followed closely by the NSSL MPAS RT. The NASA FV3 and RRFS were found to have similar CSI on average (both using the FV3 dynamic core), while the GFDL FV3 demonstrated the lowest performance during the SFE. Notably, the HRRR and NSSL MPAS runs were both found to have a high bias, whereas the RRFS and GFDL FV3 exhibited a low bias overall. Participants commonly identified this as an influential factor in their ratings of each model, and post-evaluation discussions emphasized the need for CAMs to accurately simulate convective intensity and mode in severe weather environments.

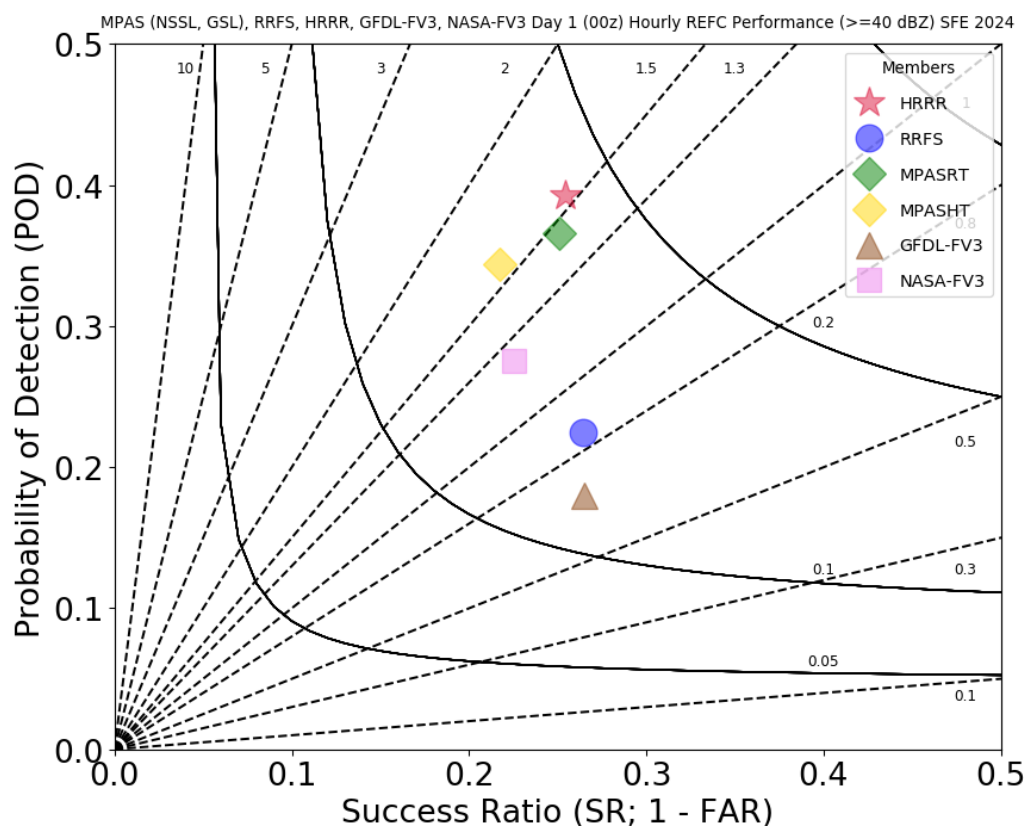


Figure 13. Mean performance of five deterministic CAMs with respect to composite reflectivity above 40 dBZ. Ensemble forecasts were compared to MRMS hourly composite reflectivity over the 5-week SFE evaluation period.

Participants rated the models much closer on average when assessing the three environmental fields and QPF (Fig. 14). The HRRRv4 again received the highest mean score for 2-m temperature, 2-m dewpoint, SBCAPE, and 6-h QPF. In general, the HRRRv4, RRFS, and NSSL MPAS HT received very similar ratings in the three environmental fields, and any differences were not significant at the 95% confidence

level. Conversely, the GFDL FV3 and NASA GEOS FV3 models consistently received the lowest mean ratings in each comparison, and these scores were found to be significant when compared to the highest-rated models. Participants cited the magnitude and location of cold pools and mesoscale boundaries as the most influential factors contributing to their ratings for each environmental field, but systematic biases in the models were also noted as points of concern. For example, respondents frequently observed a strong dry bias in the GFDL FV3's 2-m dewpoint which adversely affected its rating for that field. Greater differences were noted when evaluating the 6-h QPF, where the aforementioned behavior of the RRFS composite reflectivity and handling of deep convection had a considerable impact on the model's performance. Participants rated the RRFS fourth in the QPF evaluation, often citing the overly broad and displaced precipitation forecasts. The model's mean rating of 5.4 was found to be statistically lower than that of the HRRRv4 (6.6) and NSSL MPAS HT (6.3), and more in line with ratings given to the NASA GEOS FV3 (5.5) and GFDL FV3 (4.8). Participants primarily considered the coverage and magnitude of estimated rainfall when assessing each model's 6-h QPF, and many respondents commented that they did not place as much emphasis on location error.

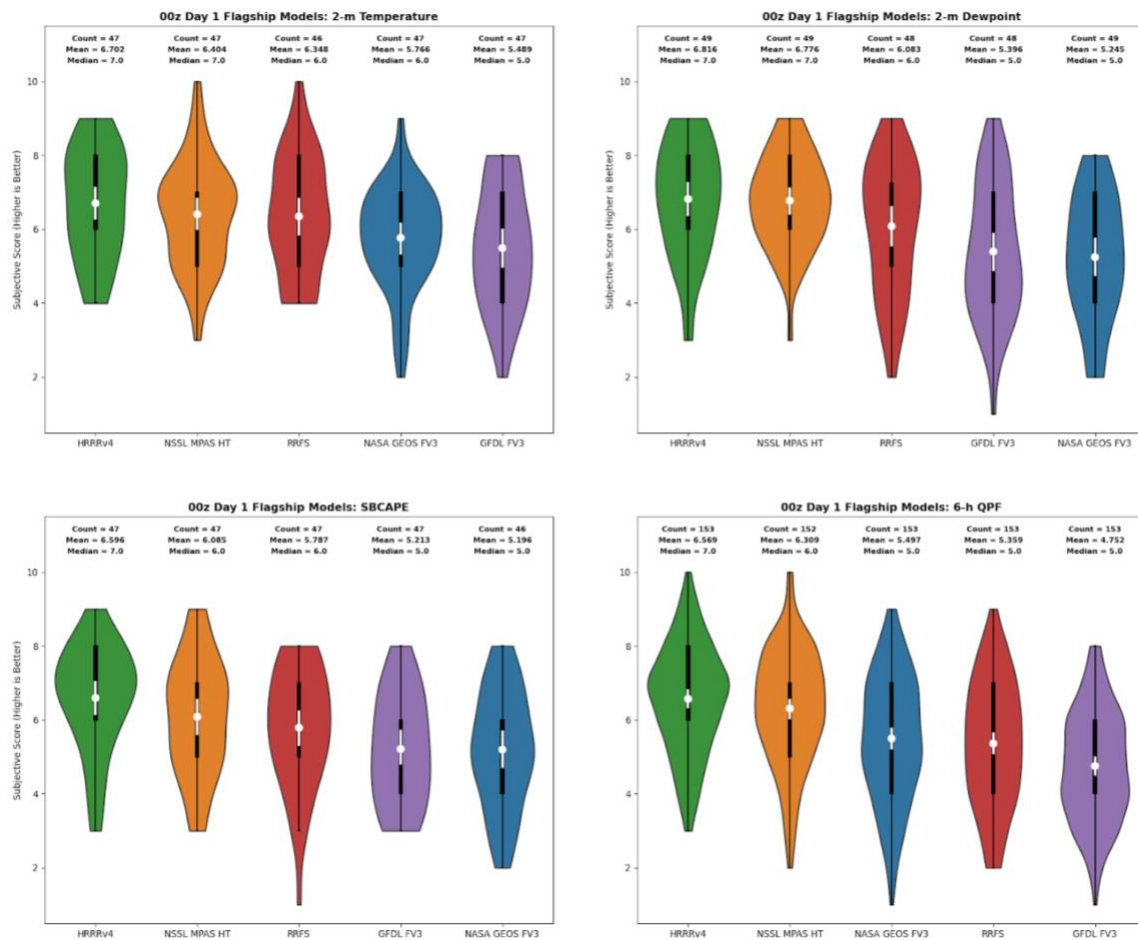


Figure 14. Subjective rating distributions for 2-m temperature (upper-left panel), 2-m dewpoint (upper-right panel), SBCAPE (lower-left panel), and 6-h QPF (lower-right panel) at Day 1 lead times by SFE participants. The white dots represent the mean scores for each ensemble, and the white bars indicate the 95% confidence intervals for each mean.

### 3.2.2 (D2) CLUE: 12Z Day 2 Deterministic Flagships

This evaluation was similar to the previous one, except participants were asked to evaluate the models at Day 2 (f25 - f48) lead times. Only the 12Z model initializations were assessed in this evaluation, and all models were blinded until after the surveys were submitted. Participants were again asked to evaluate the reflectivity and UH forecasts from each model, the same randomly selected environmental field from the Day 1 evaluation, and the 6-h QPF on a scale of 1 (Very Poor) to 10 (Very Good). The GFDL FV3 was not available from the 12Z initialization time, so only four models were compared for the Day 2 evaluation. Participants were informed of this change from the Day 1 evaluation, but were not told the identity of the missing model until after all surveys had been submitted.

Subjective scores for each model were found to be lower at the Day 2 lead times than those discussed in the Day 1 evaluations, but the overall rankings remained the same. As before, the HRRRv4 received the highest mean rating (5.8) when assessing the structure, evolution, location, and timing of simulated storm reflectivity and UH (Fig. 15). This was followed by the NSSL MPAS HT (5.4), RRFS (5.3) and NASA GEOS FV3 (5.1). Only the NASA GEOS FV3 mean score was found to be statistically significant from the HRRRv4 at the 95% confidence level, while the other models were all statistically similar. The RRFS saw the greatest range in how it was rated during the experiment, with a maximum rating of 10 and a minimum rating of 1. The other models all received a maximum rating of 9, and the NASA GEOS FV3 also saw a minimum rating of 1 during the experiment.

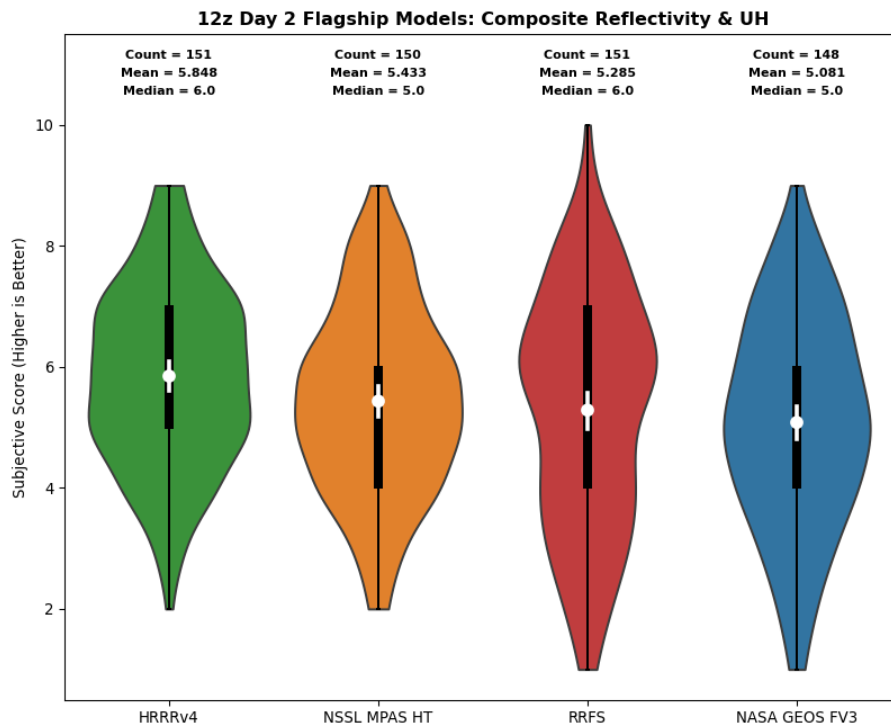


Figure 15. Same as Fig. 12, except for Day 2 lead times.



Participants gave the HRRRv4 the highest mean rating for all environmental fields at Day 2 lead times, while the NSSL MPAS HT was rated the second best – for all but 2-m temperature (Fig. 16). In general, all four models received very similar ratings in each of the environment evaluations, and any differences in the mean scores were found to not be statistically significant at the 95% confidence level. Somewhat larger differences were noted in the 6-h QPF evaluation, where the HRRRv4 was rated statistically higher than the other three models.

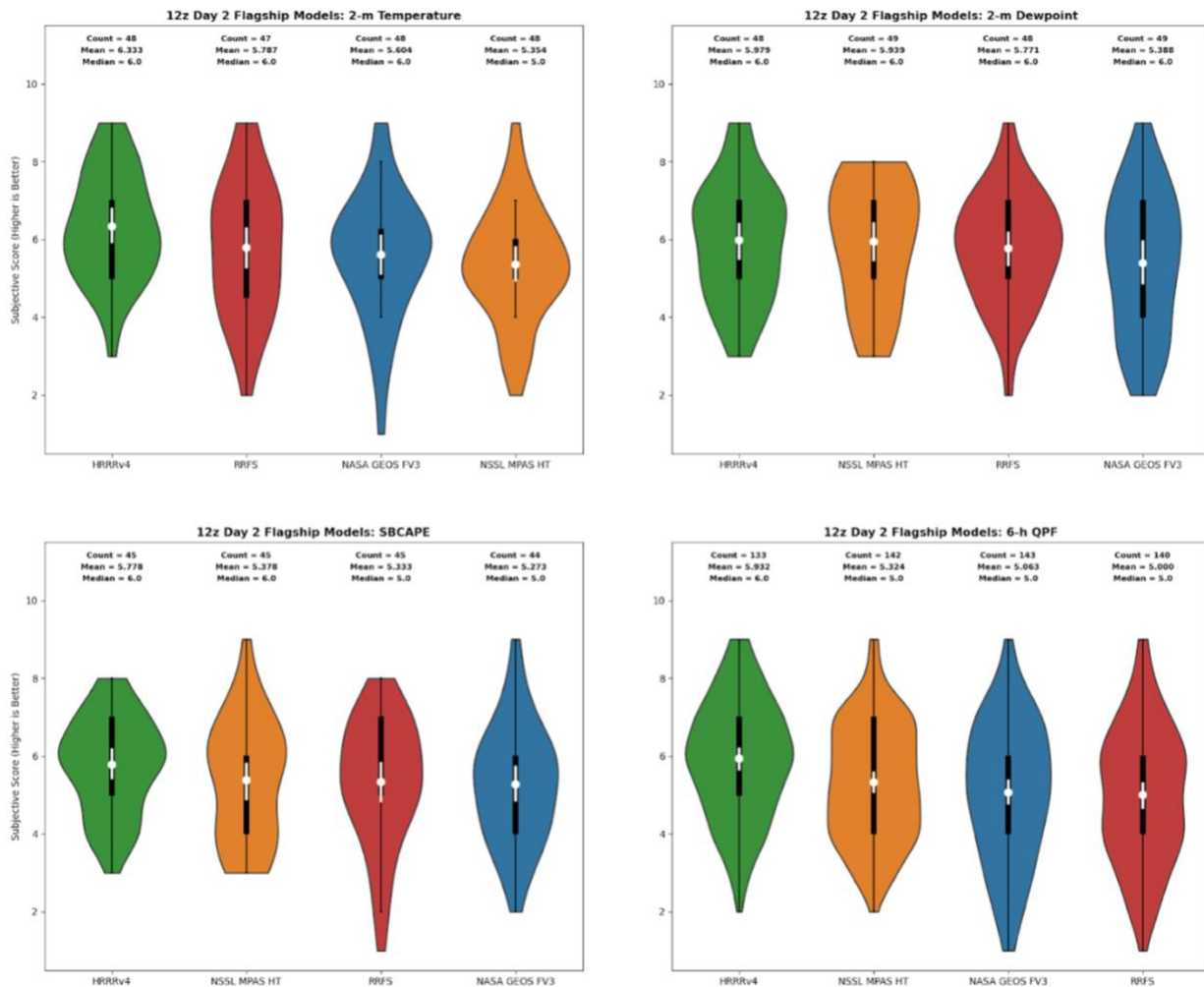


Figure 16. Same as Fig. 14, except for Day 2 lead times.

### 3.2.3 (D3) CLUE: RRFS vs. HRRR

One of the critical evaluations during the 2024 HWT SFE was directly comparing the deterministic RRFS control member to the operational HRRR. This was done for both the 00Z and 12Z runs to assess the readiness of the RRFS for operational severe weather forecasting applications on Day 1. Participants were asked to examine storm-attribute fields (e.g., Fig. 17), including composite reflectivity and UH, updraft speed, 10-

m wind speed, and 6-h QPF, and provide a single rating for the convective day (i.e., f12-f36 for the 00Z runs, and f01-f24 for the 12Z runs). For this evaluation, a five-point Likert scale was used to rate the RRFS as much worse, somewhat worse, about the same, somewhat better, or much better than the HRRR for each cycle and each field. For example, the 00Z RRFS control forecast was generally rated much worse than the 00Z HRRR forecast for the high-impact severe weather event affecting Houston on 16 May 2024 (Fig. 17).

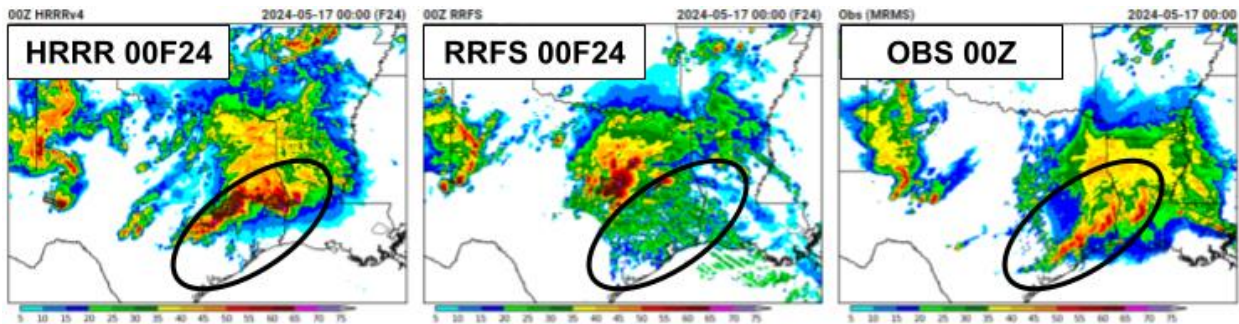


Figure 17. Example of the 2024 HWT SFE model comparison page for the RRFS control member vs. HRRR valid at 00Z on 17 May 2024. The composite reflectivity 24-hour forecasts are shown for the 00Z HRRR (left panel), the 00Z RRFS control (middle panel), and the observed MRMS composite reflectivity (right panel).

For the storm-attribute fields over the five-week SFE, the RRFS control member was most commonly rated somewhat worse to about the same as the HRRR (Fig. 18 for 00Z runs; 12Z runs not shown, but qualitatively similar results). The biggest difference in performance was for simulated reflectivity/UH and QPF, where the median rating from the SFE participants was for the RRFS control member being somewhat worse than the HRRR (Fig. 18). The most notable aspect about the RRFS control member performance was the struggles on the days with the most significant severe weather. On some of the most impactful severe weather days that were evaluated during the SFE (May 6, 8, 16, & 21), over 90% of participants rated the RRFS control member somewhat worse or much worse than the HRRR for reflectivity forecasts, and not a single participant rated the RRFS control member better than the HRRR on those days. The most common issue noted by SFE participants on these days was the absence or lack of deep convection in forecasts from the RRFS control member where significant severe weather was occurring. The Grell-Freitas (GF) deep convective parameterization scheme was implemented in the RRFS control member this year to mitigate the overly intense and spurious deep convection documented in the RRFS during the 2023 HWT SFE. Apparently, the GF deep convective parameterization scheme negatively impacted the RRFS control member forecast performance during high-impact severe weather events by overly suppressing areas of deep convection (Fig. 19).

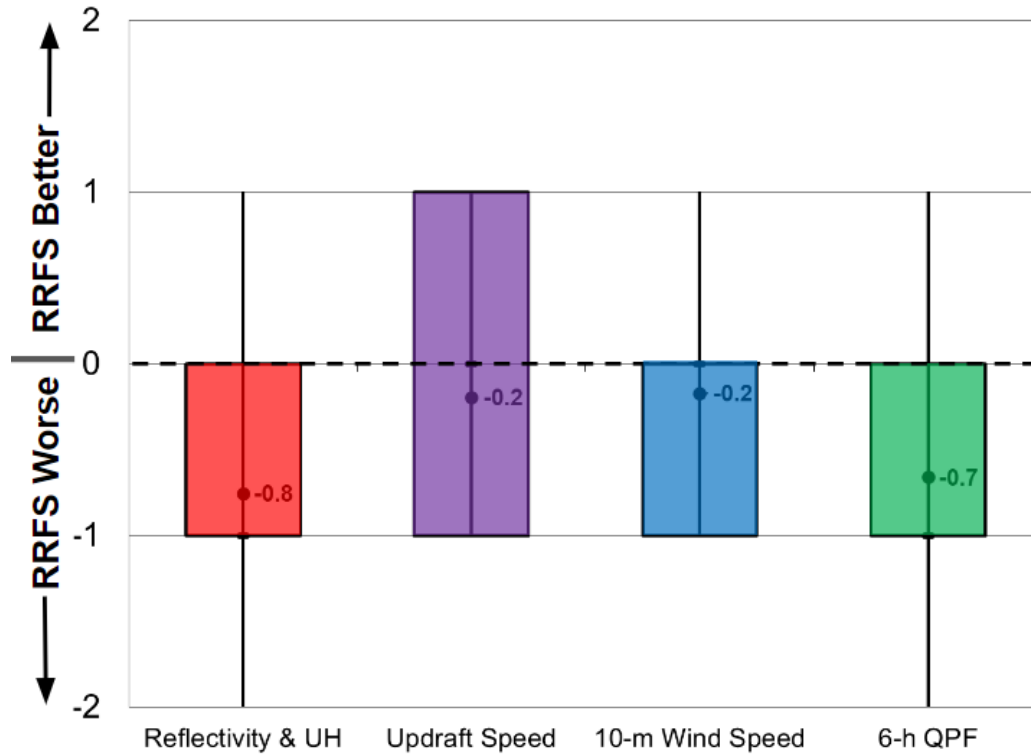


Figure 18. Distributions of subjective ratings (-2 to +2) by SFE participants of the 00Z RRFS control member compared to the HRRR for composite reflectivity and UH (red), updraft speed (purple), 10-m wind speed (blue), and 6-h QPF (green). The ratings represent the RRFS control member compared to the HRRR -2: Much Worse; -1: Slightly Worse; 0: About the Same; +1: Slightly Better; +2: Much Better.

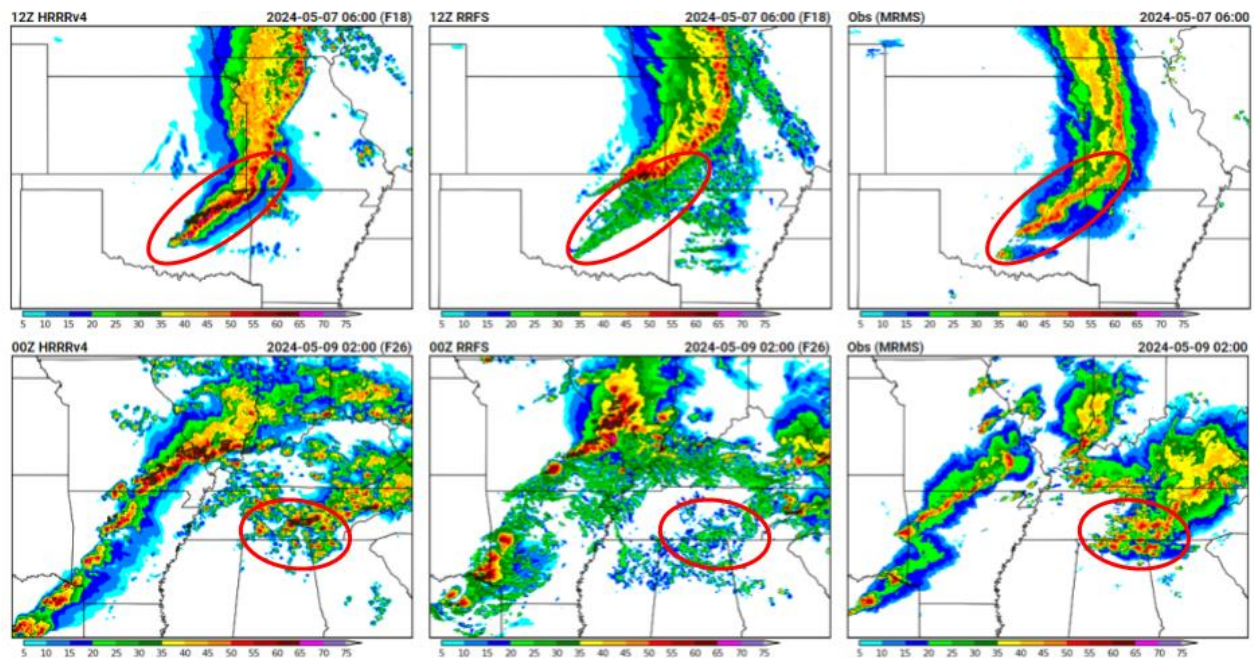


Figure 19. High-impact severe weather examples from the 2024 HWT SFE where the RRFS control member (middle column) performed much worse than the HRRR (left column) alongside MRMS composite reflectivity observations (right column). The top row shows 18-h forecasts on a High Risk day from 12Z on 6 May (valid at 06Z on 7 May 2024), and the bottom row shows 26-h forecasts on a moderate Risk day from 00Z on 8 May (valid at 02Z on 9 May 2024).

These subjective ratings and examples are supported by objective verification metrics as well. A performance diagram of 40-km neighborhood composite reflectivity  $\geq 40$  dBZ (Fig. 20) reveals that the 00Z and 12Z HRRR runs have a notably higher POD compared to the RRFS control runs while maintaining a similar FAR. This results in a higher CSI for HRRR runs in predicting deep convection during the Day 1 time period over the daily SFE mesoscale domain of interest. Note that the RRFS control runs have a low bias (i.e.,  $<1$ ), and CAMs with a low bias at 40 dBZ historically have received lower subjective ratings in previous SFEs and by SPC forecasters. In hazardous weather forecasting, there is a large penalty for missing events, so a greater importance is placed on higher POD in subjective ratings than achieving a bias close to one (which modelers often target). The results from this year's SFE are consistent with previous findings in that the HRRR with a higher POD for 40 dBZ reflectivity was rated subjectively higher than the RRFS control run, and this was confirmed by much better performance by the HRRR for several high-impact severe weather events.

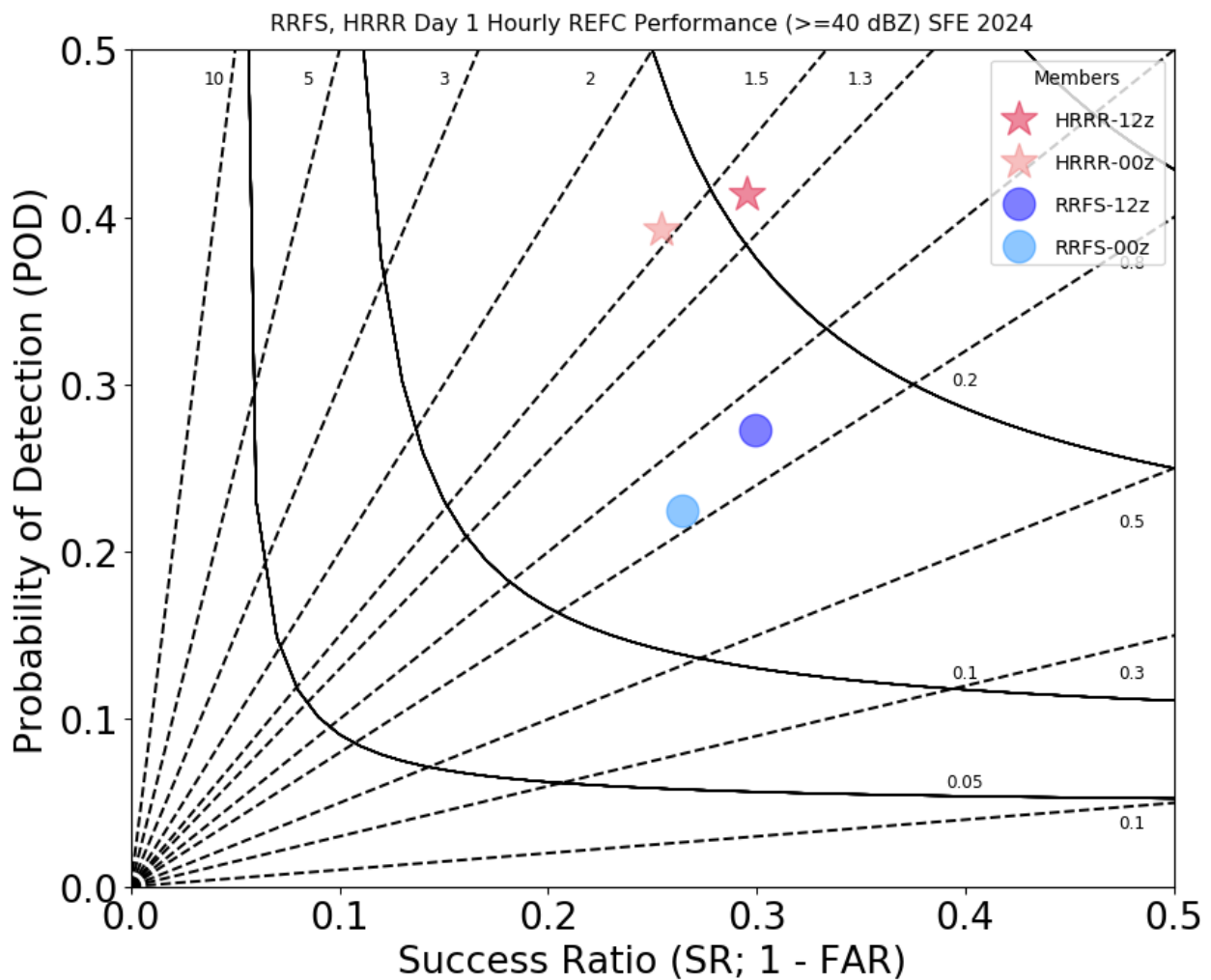


Figure 20. Performance diagram for accumulated hourly 40-km neighborhood composite reflectivity  $\geq 40$  dBZ covering the 24-h convective day (i.e., 12-12Z) over the five-week period of the HWT SFE. The 00Z and 12Z HRRR (red stars) and RRFS (blue circles) performance characteristics are labeled on the diagram. The statistics are only calculated over the primary mesoscale domain used each day for evaluation activities.

Regarding the ratings of the environment fields, SFE participants gave a slight edge to the HRRR for SBCAPE, 2-m temperature, and 2-m dewpoint over the RRFS control (Fig. 21 for 00Z; 12Z not shown, but qualitatively similar results). The most common ratings were about the same to slightly worse for the RRFS control run. Participant comments were diverse and wide ranging without any obvious systematic biases or differences between the models during the SFE. If a model had a slightly better forecast of the environment, the HRRR was more than twice as likely to be favored than the RRFS in participant ratings.

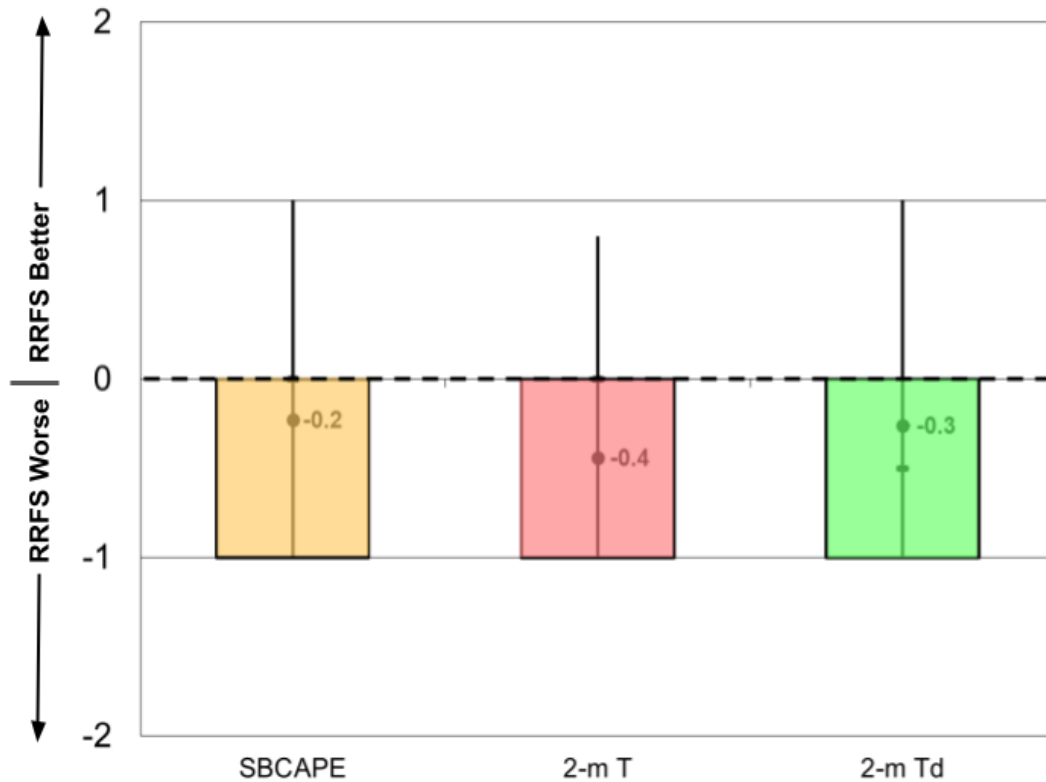


Figure 21. Same as Fig. 18, except for environmental fields of SBCAPE (yellow), 2-m temperature (pink), and 2-m dewpoint (light green).

### 3.2.4 (D4) CLUE: RRFS vs. HRRR DA

HRRRv4 and RRFS were examined during the first 12 forecast hours from the 21Z and 00Z initializations to assess forecast skill at times during which the data assimilation has a large impact. The valid times 22, 01, and 06Z were considered. Participants first compared forecast UH and simulated composite reflectivity to observed reflectivity and were asked, “Please rate on a scale of 1 (very poor) to 10 (very good) how well each model depicts storms that were ongoing at [22z, 01z, or 06z]. Consider aspects like storm retention, strength, and location in your answer.”

For all lead times at both initializations examined, the mean subjective ratings for HRRRv4 were higher than RRFS (Fig. 22) with statistically significant differences. Common themes from the comments were that RRFS often had areas of spotty, light to moderate reflectivity where there should have been storms. Also, there were several cases where RRFS immediately dissipated an intense supercell when it should have been maintained. Specific comments included, "... The HRRR appeared to perform noticeably better in storm location, coverage, and strength...", "The HRRR overall seemed to show somewhat better placement of precipitation and more realistic core strengths and relative sizes. In particular for the forecast valid at 06Z from the RRFS it did quite poorly and would have mislead a forecaster whereas the HRRR would at least get you in the ballpark...", "... RRFS seems to produce excessive light reflectivities ...", "... RRFS didn't convect in Central Oklahoma at 01z. That's a big miss ...", "... Both versions [initializations] of the HRRR performed better than the RRFS. While the intensity might be on the aggressive side, at least the HRRR captured the full picture and spread the line into southern Oklahoma unlike the RRFS...", "The RRFS at 06z seems to have some issues when it comes to its convective scheme, where it has a broad region of shallow convection which did not materialize. Also, it had shallow convection where there was observed deep convection ...". Figure 22 illustrates three example cases with the behaviors described in the comments.

Relative to SFE 2023, the mean subjective ratings in SFE 2024 from the HRRR generally increased while those from the RRFS decreased. Thus, the differences in mean ratings between HRRR and RRFS increased from SFE 2023 to SFE 2024 (Table 3).

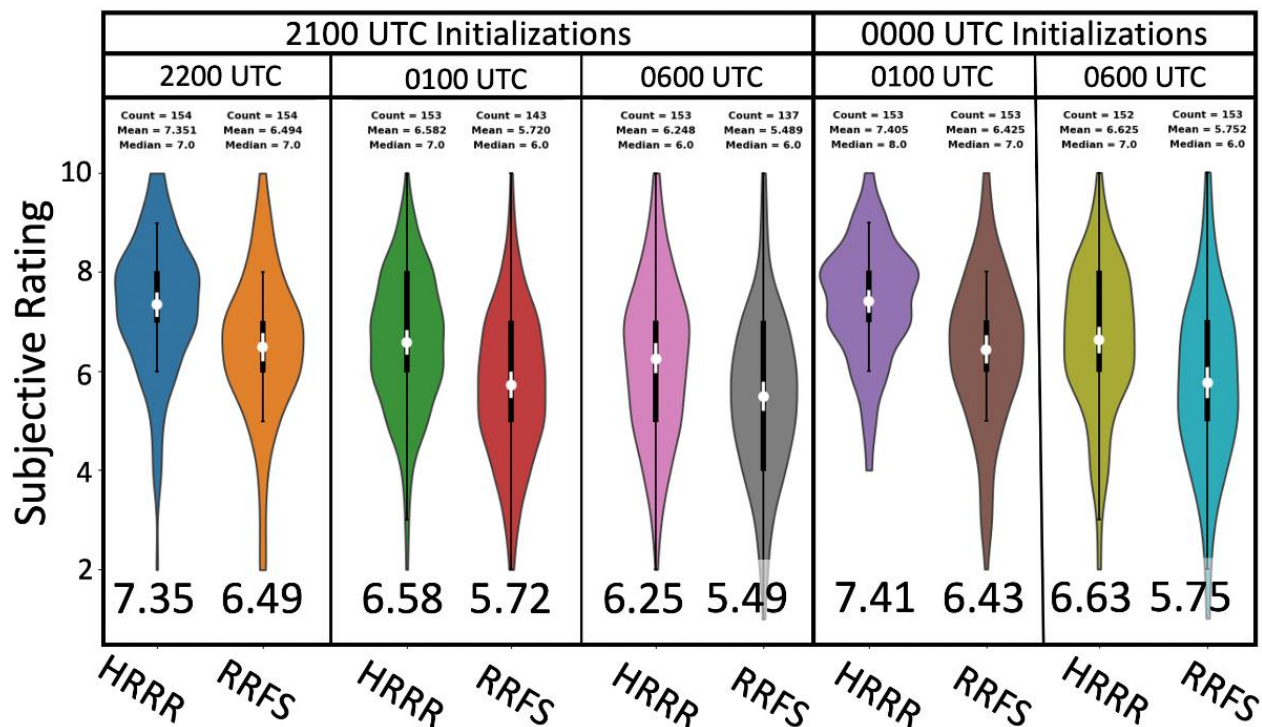


Figure 22. Violin plots showing the distributions of subjective ratings for the D4 RRFS vs. HRRR DA evaluation. Mean subjective ratings are indicated by the number below each violin.

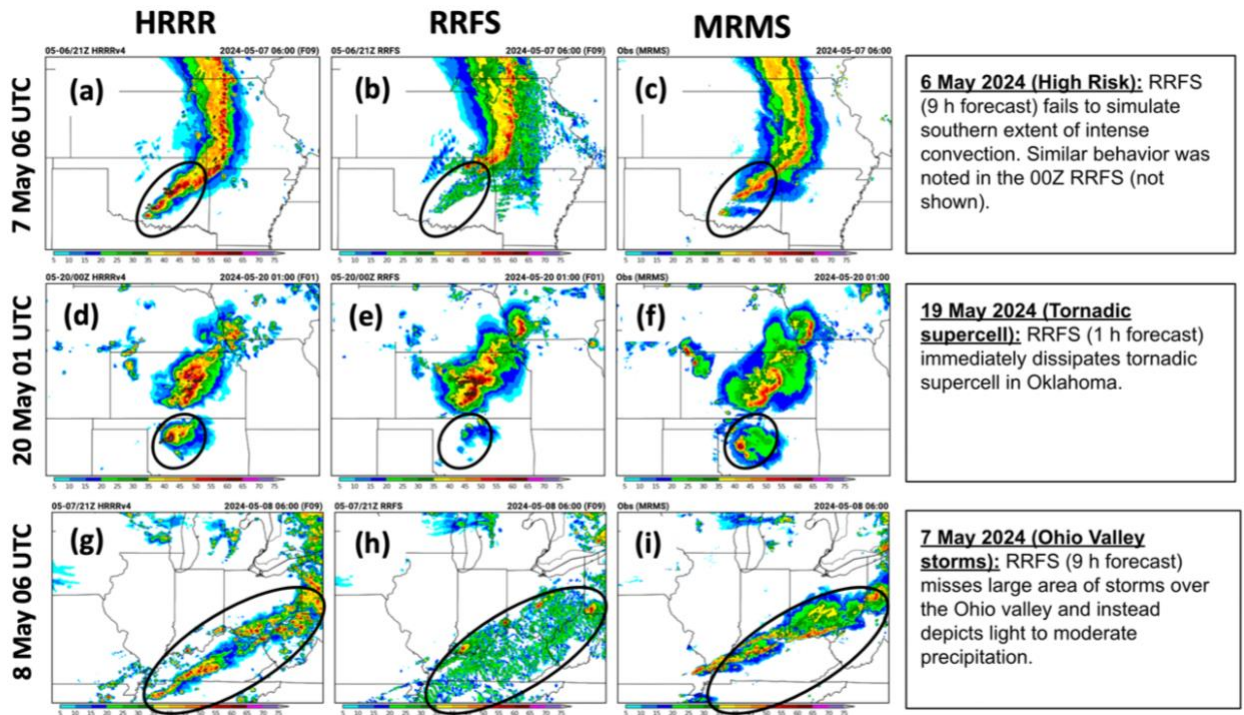


Figure 23. Simulated composite reflectivity at 9-h lead times from 21Z 6 May 2024 initializations of (a) HRRR, (b) RRFS, and (c) MRMS observations. Black ovals denote the main area of interest described in the text boxes to the right of each row. (d)-(f) same as (a)-(c), except for 1-h lead times from 00Z 20 May 2024 initializations, and (g)-(i) same as (a)-(c), except for 9-h lead times from 21Z 7 May 2024 initializations.

		2024	HRRR-RRFS	2023	HRRR-RRFS
2100 UTC Initialization	HRRR @ 22 UTC	7.35	0.86 ↑	7.08	0.35
	RRFS @ 22 UTC	6.49		6.73	
	HRRR @ 01 UTC	6.58	0.86 ↑	6.62	0.24
	RRFS @ 01 UTC	5.72		6.38	
	HRRR @ 06 UTC	6.25	0.76 ↑	6.13	0.62
	RRFS @ 06 UTC	5.49		5.51	
0000 UTC Initialization	HRRR @ 01 UTC	7.41	0.98 ↑	7.27	0.41
	RRFS @ 01 UTC	6.43		6.86	
	HRRR @ 06 UTC	6.63	0.88 ↑	6.41	0.27
	RRFS @ 06 UTC	5.75		6.14	

Table 3. Average subjective ratings for the D4 RRFS vs. HRRR DA evaluation for 2024 and 2023. Green upward pointing arrows indicate differences in mean ratings between HRRR and RRFS (i.e., HRRR minus RRFS) that increased from 2023 to 2024.

Next, participants were asked to evaluate one of three randomly selected environment fields, and assign a subjective rating on a scale of 1-10 reflecting the overall

skill during the first 12 hours of the forecast. Again, for each field and at both initialization times, the mean subjective ratings in HRRR were higher than RRFS. The largest differences were with surface-based CAPE. Only the differences for CAPE were statistically significant (Fig. 24).

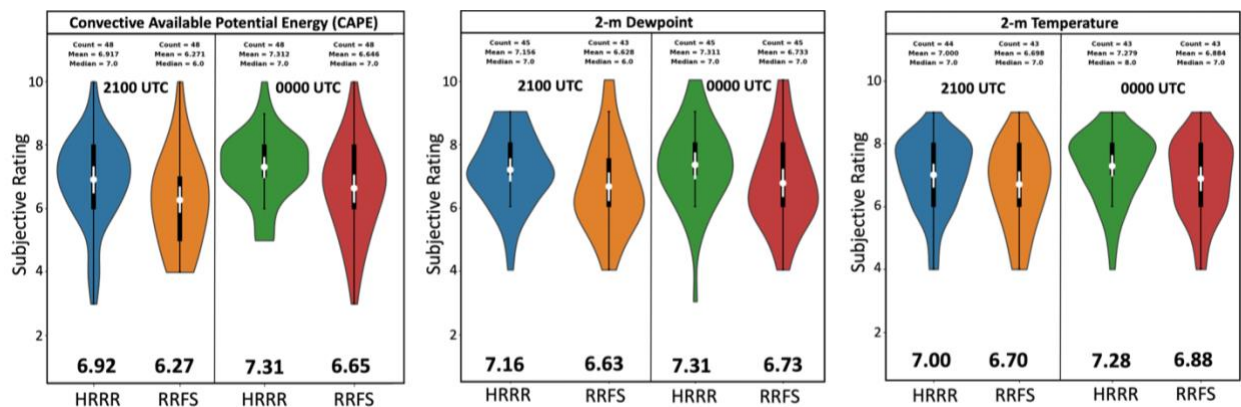


Figure 24. Violin plots showing the distribution of subjective ratings from 21Z and 00Z initializations of HRRR and RRFS for CAPE (left), 2-m dewpoint (middle), and 2-m temperature (right). Mean subjective ratings are indicated by the number below each violin.

### 3.2.5 (D5) CLUE: 00Z MPAS Resolution Sensitivity

In this evaluation, participants were asked to subjectively rate on a scale of 1-10 similarly configured 2-, 3-, and 4-km horizontal grid-spacing versions of MPAS to assess resolution sensitivities. Specifically, the survey asked, “... please consider how the MPAS2 (2-km grid-spacing) and MPAS4 (4-km grid-spacing) models differ from the baseline NSSL-MPAS-HN (3-km grid-spacing) model. In particular, focus on any differences in how the models depict the timing, location, and mode of thunderstorms within the domain and how those forecasts compare to observations.” Unfortunately, the 2-km grid-spacing MPAS configuration was not available in time for SFE 2024, so the comparison only involved the 3- and 4-km configurations.

The distributions of subjective ratings were very similar with the mean ratings from NSSL-MPAS-HN (3-km) slightly higher (Fig. 25). However, differences were not statistically significant. Survey comments mainly reflected the perceived similarity between the two sets of forecasts, but at times indicated that the 3-km grid-spacing forecasts were better than the 4-km ones. For example, “... HN [3-km grid-spacing] was a lot better with positioning of the MCS than the MPAS4 [4-km grid-spacing], and both models overestimated the strength of the MCS, especially as it entered into the Atlantic ...”, “... Both models were in general very similar, especially early and late in the event. Both models had similar biases (e.g., moved convection out of Texas and Oklahoma too quickly and in a more organized MCS than in observations). The one notable difference is in the supercellular convection in Tennessee - the 3-km version better indicated the potential for robust rotating thunderstorms in that region...”, and “... Both NSSL-MPAS-HN and MPAS4 performed similarly, with finer storm-scale features more evident in



NSSL-MPAS-HN ...”. Two example cases are shown in Figure 26. For the 8 May 2024 case (Fig. 26a-c), the 3-km configuration better depicted strong supercells in central TN, and for the 28 May case (Fig. 26d-f), both forecasts were very similar with timing errors for an MCS that affected the TX gulf coast.

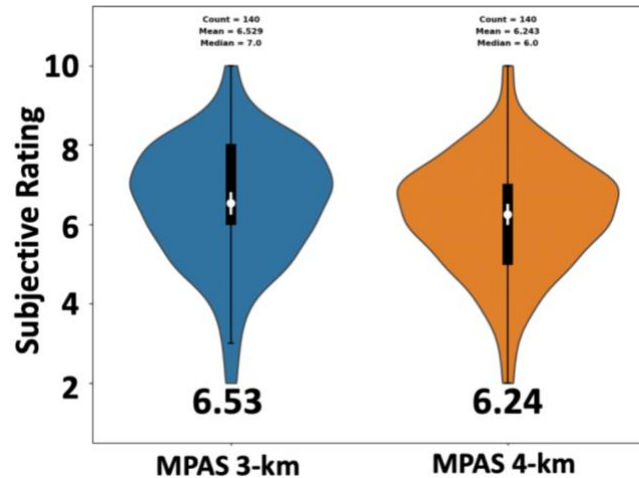


Figure 25. Violin plots showing the distribution of subjective ratings for 3- and 4-km grid-spacing configurations of MPAS. Mean subjective ratings are indicated by the number below each violin.

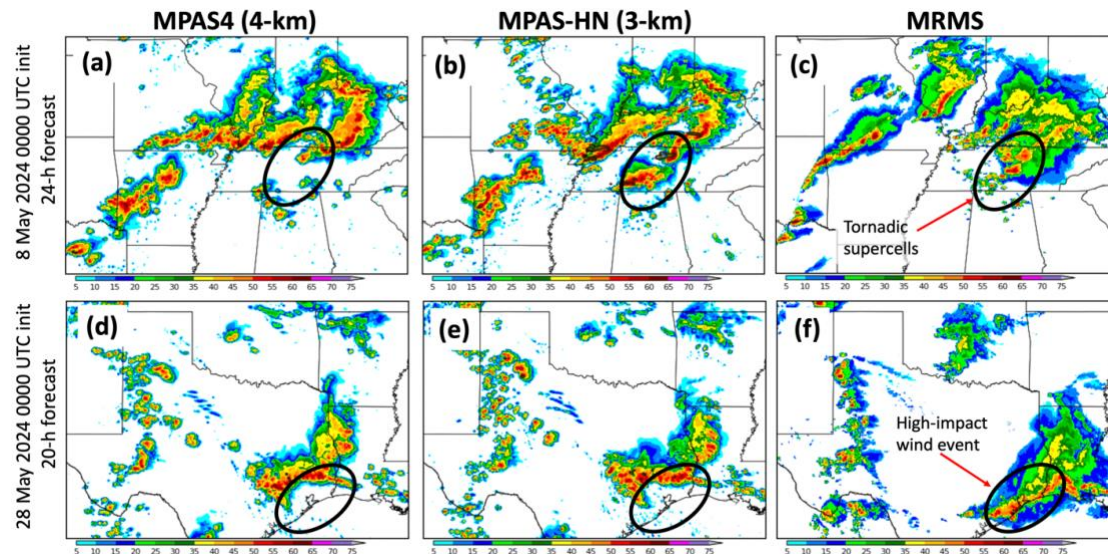


Figure 26. Simulated composite reflectivity at 24-h lead times from 0000 UTC 8 May 2024 initializations of (a) MPAS4 (4-km grid-spacing), (b) MPAS-HN (3-km grid-spacing), and (c) MRMS observations. Black ovals denote the highest impact weather event occurring at the time. (d) – (f) same as (a) – (c), except for 20-h lead times from 0000 UTC 28 May 2024 initializations.

Next, participants were asked to comment on any notable differences in how the MPAS configurations depicted one of three randomly assigned environment fields: surface-based CAPE, 2-m temperature, or 2-m dewpoint. For 2-m temperature, most responses indicated that there were not any notable differences, however, a couple times

it was noted that the cold pools in the 3-km runs were slightly colder and better represented reality as indicated by 3DRTMA. Comments included, “... *As with the reflectivity forecasts, the differences between the two models are within expected noise level and appear to be related to how convection is resolved ...*”, and “... *There were some notable differences in the temperature early in the day in Tennessee. The 3-km version of MPAS produced a colder cold pool behind the early day MCS. These colder temperatures seemed to better reflect observations ...*”. For surface-based CAPE, once again the majority of comments indicated the runs were very similar. A couple comments indicated slight improvements in the 3-km runs and there were also several comments indicating that both runs over-predicted CAPE. Example comments included, “... *All MPAS models were slightly aggressive with SBCAPE, but the MPAS-HN did fair better showing the more stable airmass in the wake of the MCS ...*”, “... *Only very slight differences that weren't meaningful ...*”, and, “... *Once again, basically the same between the 4km and 3km MPAS. Both models overdid SBCAPE quite strongly ...*”. Finally, for dewpoint most responses indicated that the forecasts were very similar. One participant noticed that the 3-km MPAS depicted a heat burst that the 4-km MPAS did not depict. Example comments included, “... *Interestingly, there was a heat burst signature at 0900 UTC in MPAS3 that was not in MPAS4 ...*”, “... *Both models are biased high on 2m dewpoint ahead of the dry line ...*”, “... *Very minimal difference in 2-m DPT from 3km to 4km ...*”, “... *Little difference between two models. Both had higher dewpoints spreading further north than in observations ...*”.

### 3.2.6 (D6) CLUE: 1-km vs. 3-km

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

For all three comparisons, differences in mean subjective ratings were quite similar, and although HRRR had slightly higher mean ratings, the differences were not significant (Fig. 27). For this same comparison in SFE 2023, it was found that NSSL1 had a significant advantage over the HRRR for maximum 10-m winds and several cases were noted for high-end, wind-producing MCSs that had a better signal in the 1-km grid-spacing NSSL1. Interestingly, although there were a few high-end, wind producing

MCSs during SFE 2024, there was not an advantage noted for NSSL1. An example case is shown in Figure 28 and a few survey comments are highlighted as follows: “... Didn't seem like there was any distinct advantage in 1-km ...”, “... The HRRR 10-m max winds did a pretty amazing job capturing the cluster of wind reports that moved NE across west Texas from 20 to 00Z ...”, “... NSSL1 overall did better in capturing the different aspects of severe weather. HRRRv4 actually had more spurious convection than NSSL1 ...”, and “... 1km WRF completely missed the tornadic convection in central TN around 00Z also did not handle the morning MCS across MO well ...”.

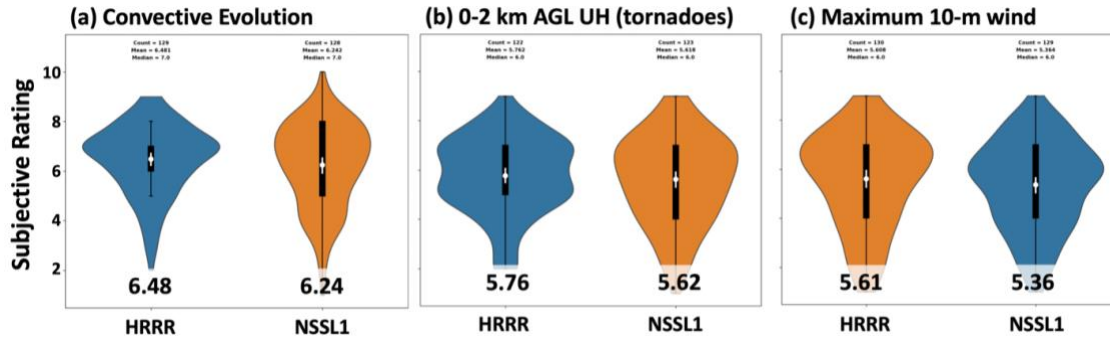


Figure 27. Violin plots showing the distribution of subjective ratings for HRRR and NSSL comparisons of (a) convective evolution (i.e., composite reflectivity), (b) 0-2 km AGL UH (tornadoes), and (c) maximum 10-m wind. Mean subjective ratings are indicated by the number below each violin.

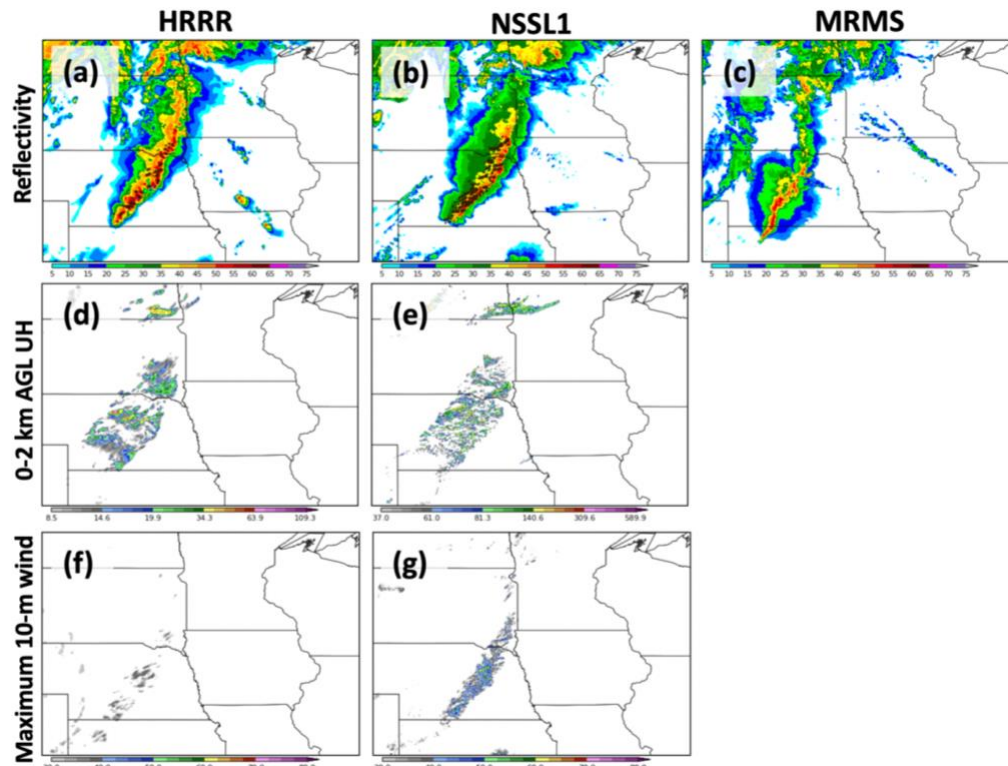


Figure 28. Simulated composite reflectivity at 28-h lead times from 00Z 23 May 2024 initializations of (a) HRRR, (b) NSSL1, and (c) MRMS observations. (d)-(e) same as (a)-(b), except for 0-2 km AGL UH, and (f)-(g) same as (a)-(b), except for hourly maximum 10-m wind speed.

### 3.3 Evaluation – CAM (E)nsembles

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

Similar to the deterministic evaluation of the RRFS control member to the operational HRRR (section D3), the RRFS Ensemble Forecast System (REFS) was compared directly to the HREF, which is the operational CAM ensemble in the NWS. This evaluation was done for the 00Z run to assess the readiness of the REFS to replace the HREF for operational convective forecasting applications on Day 1. Participants were asked to examine probabilistic storm-attribute fields, including updraft helicity, updraft speed, 10-m wind speed, and composite reflectivity, and provide a single rating for the convective day (i.e., f12-f36). A five-point Likert scale was also used in this evaluation to rate the REFS as much worse, slightly worse, about the same, slightly better, or much better than the HREF for each field. For example, the 00Z REFS forecast was generally rated slightly worse than the 00Z HREF forecast for the High Risk on 6 May 2024, owing to the HREF having higher probabilities across northeastern Oklahoma where the strongest tornado occurred on that day (Fig. 29). The REFS highest probabilities are generally offset to the north and west of the most significant severe weather for this event.

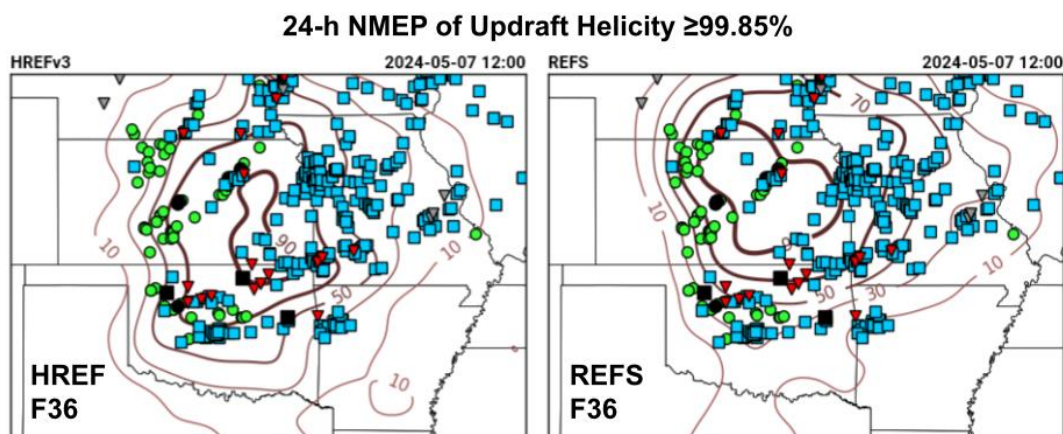


Figure 29. Example of the 2024 HWT SFE model comparison page for the REFS vs. HREF valid for the convective day of 6 May 2024. The 24-h neighborhood maximum ensemble probability (NMEP) forecasts of UH are shown for the 00Z HREF (left panel) and the 00Z REFS (right panel). The observed preliminary local storm reports (wind – blue boxes; sig wind – black boxes; hail – green circles; sig hail – black circles) are overlaid in both panels.

For the storm-attribute fields, the distribution of ratings is centered on the REFS being about the same as the HREF for updraft helicity, updraft speed, and 10-m wind speed (Fig. 30). The only field in which the HREF had a slight edge in subjective ratings was for composite reflectivity. The subjective comments from SFE participants were varied and wide ranging from case-to-case with no clear systematic issues, errors, or biases. This was rather surprising given the results from the HRRR and RRFS control member comparison (see Section D3). Typically, in an ensemble, the control member is the best-performing member and is representative of the ensemble performance and characteristics. This was not the case for the REFS during the SFE, as the REFS performed comparably better than its control member. It took a couple of weeks of

scrutinizing the ensemble forecasts and individual member forecasts to understand this apparent discrepancy. As noted in Section D3, the RRFS control member with the GF convective parameterization scheme overly suppressed deep convection in several high-impact severe weather events. This was not true for some of the members of the REFS that used different convective parameterization schemes. In fact, the REFS members using the scale-aware SAS (saSAS) convective parameterization scheme performed much better for those events (e.g., Fig. 31) and overall for deep convection (Fig. 32) during the SFE than those using the GF convective parameterization scheme. All of the REFS members have a frequency bias below one for neighborhood reflectivity  $\geq 40$  dBZ (Fig. 32), which is a consequence of employing deep convective parameterization schemes in these runs to temper the high bias noted last year. As mentioned previously, a low bias often lowers POD, which is a difficult compromise for high-impact convective forecasting.

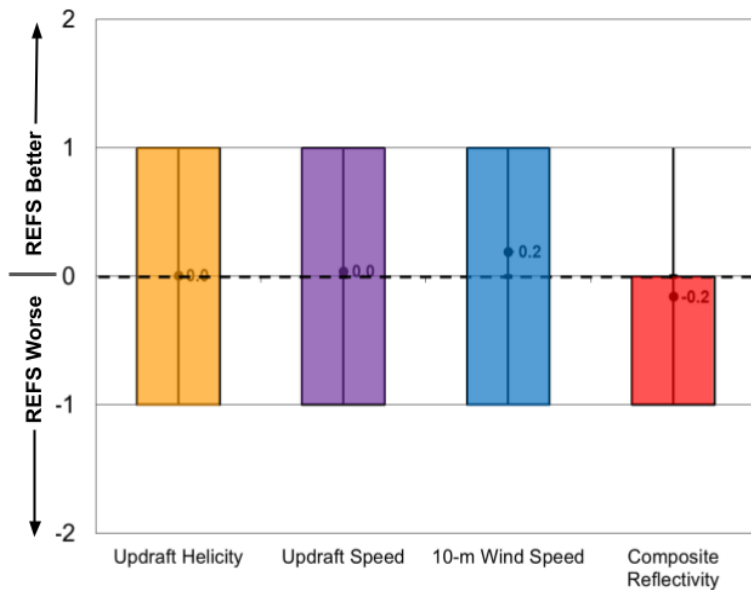


Figure 30. Distributions of subjective ratings (-2 to +2) by SFE participants of the 00Z REFS compared to the HREF for updraft helicity (yellow), updraft speed (purple), 10-m wind speed (blue), and composite reflectivity (red). The ratings represent the REFS compared to the HREF -2: Much Worse; -1: Slightly Worse; 0: About the same; +1: Slightly Better; +2: Much Better.



Figure 31. Same as bottom row of Fig. 19, except the left panel is the 26-h forecast (valid at 02Z on 9 May 2024) from REFS member 05 using the saSAS convective parameterization scheme. This member much better captures the ongoing tornadic supercells across southern Tennessee and northern Alabama than the RRFS control member.

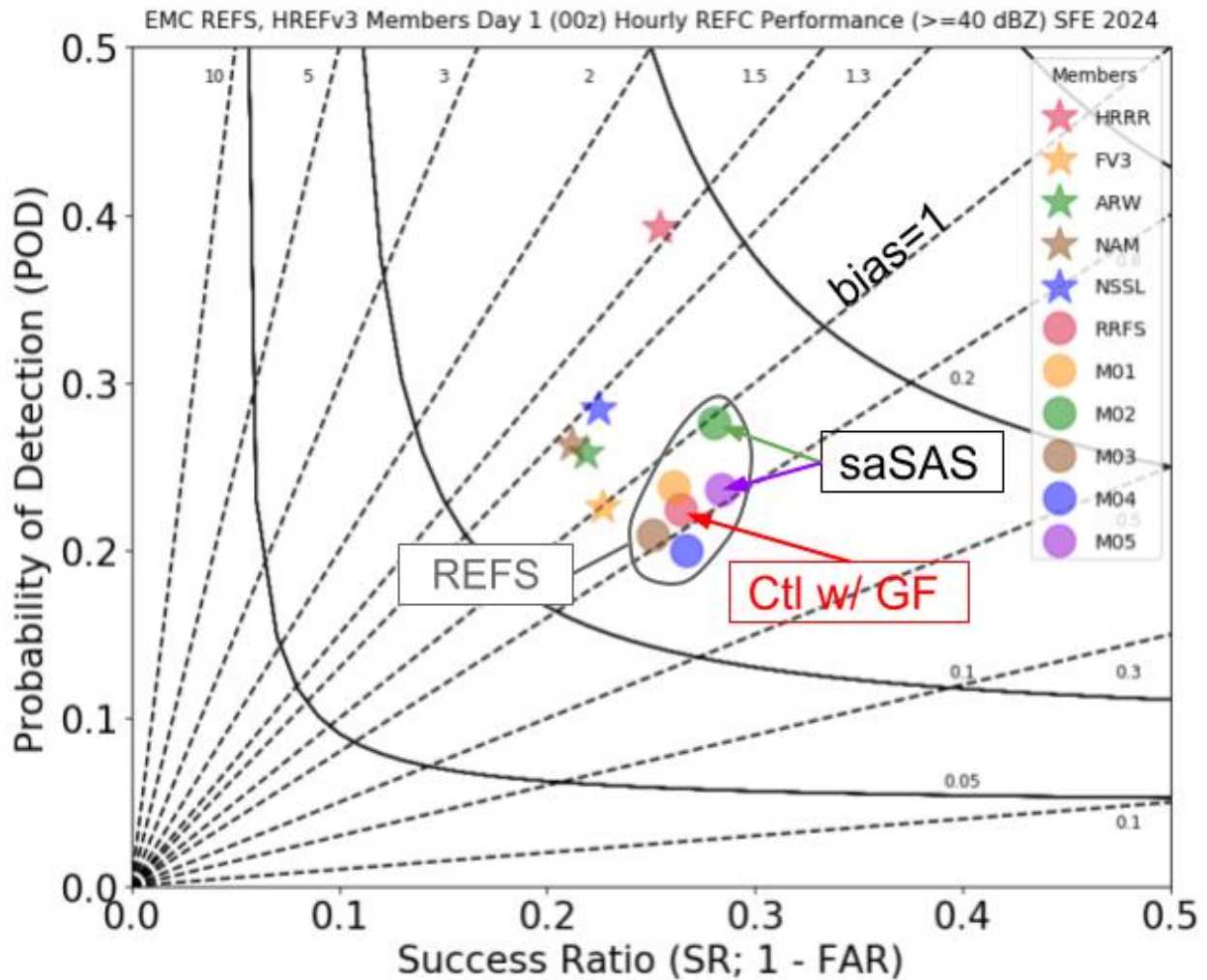


Figure 32. Same as Fig. 20, except for 00Z HREF members (stars) and 00Z REFS members (circles). Note that the REFS members using saSAS convective parameterization scheme (green and purple circles) had the best performance of any REFS members for deep convection during the SFE.

Despite the disparate performance characteristics of the REFS members depending on convective parameterization scheme utilized, the overall ensemble objective performance of the REFS was on par with the HREF in forecasting deep convection during the SFE over the daily mesoscale domains (Fig. 33). For given probabilistic thresholds, the HREF tended to have higher POD while the REFS tended to have lower FAR, resulting in similar CSI values for most probability bins. Additionally, the REFS showed slightly better reliability than the HREF for probability forecasts below 40%, which is an impressive result from a single-model ensemble. Historically, single-model ensembles tend to be underdispersive, resulting in probabilistic overforecasts, but the multi-physics REFS with stochastic physics perturbations this year produced more statistically reliable convective forecasts than the multi-model HREF. The REFS improved reliability, however, may come at the expense of a low frequency bias and lower POD than the HREF.

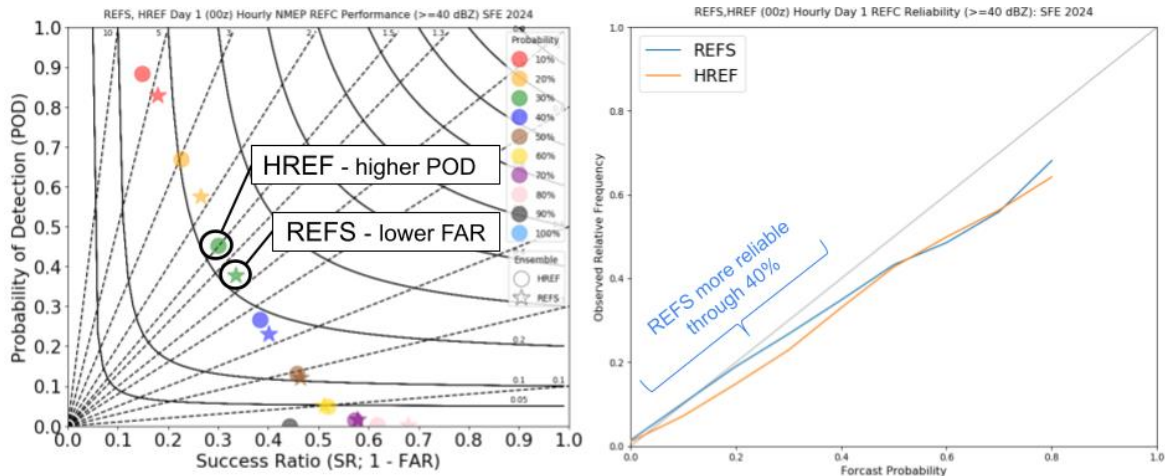


Figure 33. Performance diagram (left panel) and reliability diagram (right panel) of the 00Z probabilistic forecasts of composite reflectivity  $\geq 40$  dBZ from the HREF and REFS. The probability forecasts are binned into 10% increments and accumulated hourly for each 24-h convective day during the SFE over the mesoscale domain of interest.

For the ensemble mean environmental fields, the REFS was typically rated about the same to slightly better than the HREF (Fig. 34). Most of the comments from SFE participants discuss the environmental forecasts being similar between the REFS and HREF with some times/locations slightly favoring one ensemble over the other. The 2-m dewpoint was perhaps slightly more favored for the REFS because the HREF tended to be biased (either too moist or dry depending on location with respect to the warm/moist sector) more commonly. Another frequent comment was that the REFS mean environmental fields displayed more detailed structure and gradients than the HREF mean environmental fields, which is not surprising given the greater diversity (i.e., resulting in washing out of features) in the HREF in terms of model core and physics.

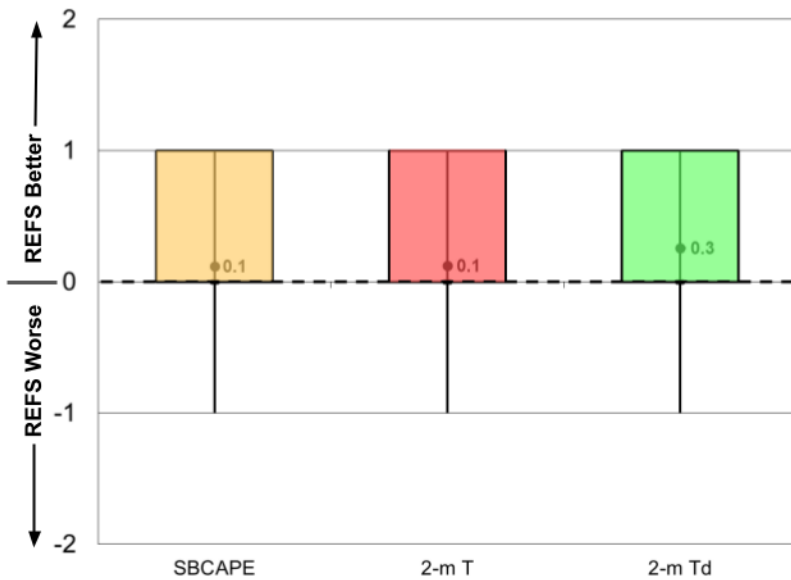


Figure 34. Same as Fig. 30, except for environmental mean fields of 2-m temperature (pink), 2-m dewpoint (light green), and SBCAPE (yellow).

### 3.3.2 (E2) CLUE: 12Z Day 1 Ensemble Flagships

This evaluation assessed the skill of four different REFS configuration strategies at Day 1 lead times to determine whether they could meet or exceed the skill of the operational baseline HREF. The four REFS configuration strategies varied the number and proportion of “control” and time-lagged membership in each ensemble to assess what impact these changes might have on the overall spread and skill of ensemble-derived probabilistic forecasts. The EMC-REFS ensemble configuration was composed of the RRFS, five perturbed REFS members, and the HRRR, each initialized at 12Z and 06Z for 14 members in total. REFS-SPC represented a greater proportion of control and time-lagged members, with only four perturbed REFS members initialized at 12Z and six RRFS and HRRR control members initialized at 12, 06, and 00Z (10 members total). The REFS-SPC-ARW configuration contained the same membership as the REFS-SPC, except two of the REFS perturbed members were replaced with HRW ARW members initialized at 12 and 00Z. Finally, the MPAS REFS mirrored the REFS-SPC but with all RRFS and REFS members replaced by MPAS members (without physics diversity) at the same respective initialization times.

SFE participants were provided 4-h and 24-h updraft helicity, updraft speed, and 10-m wind speed neighborhood probabilities, as well as 1-h composite reflectivity and 6-h QPF fields with which to base their assessment of the ensembles. All ensembles were evaluated blindly such that participants were not able to see which ensemble produced which forecast. Additionally, the order of each ensemble was randomized daily so that participants could not anticipate an ensemble being in the same panel day-to-day. Ensembles were unblinded following discussion of the results and after all surveys were submitted. MRMS MESH, local storm reports, NWS warnings, and NLDN lightning flashes were provided as ground truth observations.

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

The HREF received the highest rating on average during the 5-week experiment with a mean subjective score of 6.82 (Fig. 35). The REFS-SPC ranked second at 6.80, followed by the EMC-REFS (6.68), REFS-SPC-ARW (6.66), and MPAS REFS (6.41). Differences in the mean ratings of each ensemble fell well within the 95% confidence interval, and the subjective performance of the five ensembles was found to be



statistically similar across the board. This is further supported by the response distributions shown in Fig. 35, which demonstrate similar characteristics across all configurations although the mode rating for the EMC-REFS is lower than that of the HREF.

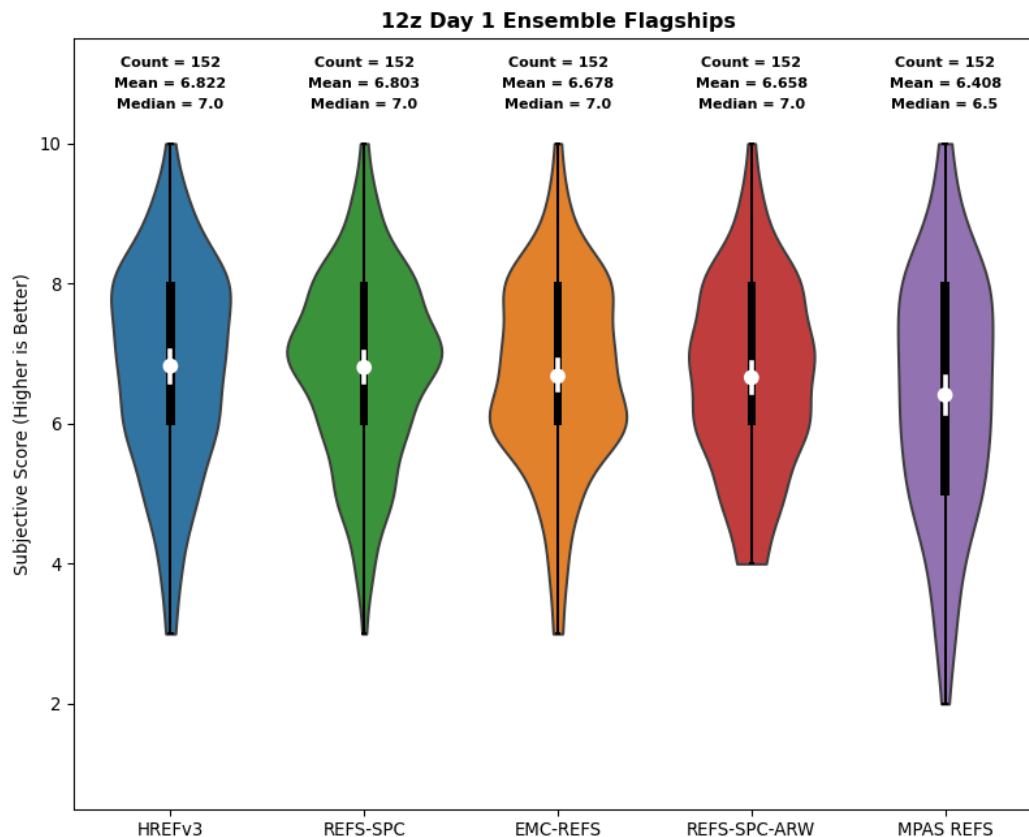


Figure 35. Distribution of subjective scores received by each ensemble at Day 1 lead times during the 5-week experiment. The white dots represent the mean scores for each ensemble, and the white bars indicate the 95% confidence intervals for each mean.

In the open response question and post-survey discussion, participants commented on how similar the ensembles were on average at Day 1 lead times. When differences in the forecasts were observed, respondents noted that the EMC-REFS, REFS-SPC, and REFS-SPC-ARW configurations tended to produce neighborhood probabilities that were higher in magnitude and more spatially focused than the HREF. These “bullseye” forecasts were often praised by respondents for a perceived reduction in false alarm area, but the smaller probability fields also occasionally missed or underforecast severe reports near the periphery of the events (e.g., Fig. 36). Participant opinions of these differences varied greatly each day, and post-survey discussion frequently revealed that ratings were closely tied to whether the respondent placed greater value on POD or FAR when making their assessments. In contrast, the MPAS REFS was found to frequently produce UH neighborhood probabilities that were lower in magnitude than the other ensembles. Participants often commented that the MPAS

REFS appeared to underforecast severe weather events, and this was a major point of discussion during the evaluations. However, it was discovered late in the SFE that the wrong model climatology was applied when calculating the MPAS neighborhood probability of UH > 99.85%, and this artificially reduced the UH neighborhood probability magnitudes that were presented to participants. Other fields like the updraft speed and 10-m wind speed neighborhood probabilities were not impacted by the error, and those fields anecdotally appeared much more similar to the other ensembles on average. Unfortunately, this issue impacted the MPAS REFS’s subjective scores, and results related to the MPAS REFS’s UH forecast may not be completely representative of the ensemble’s true performance. Thus, subjective ratings from MPAS REFS should not be leveraged to inform decisions on further development.

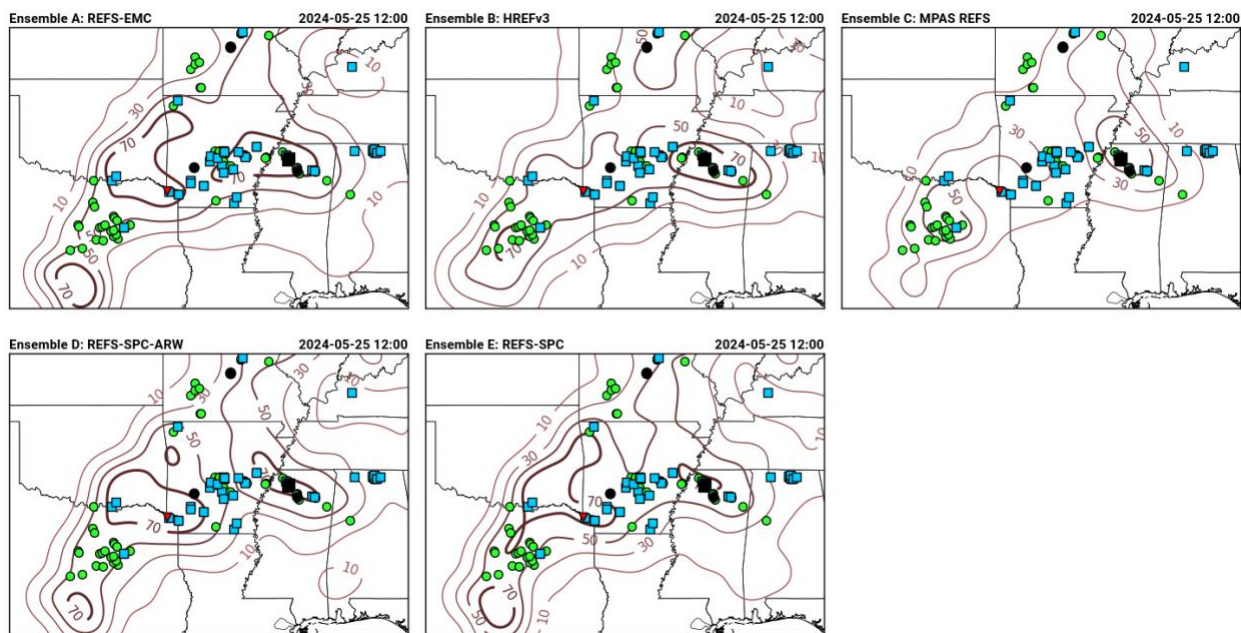


Figure 36. 24-h neighborhood probabilities of updraft helicity exceeding the 99.85th percentile on 20240524. Red triangles represent tornado reports, blue squares are wind reports, and green circles are hail reports.

Preliminary objective verification of the five ensembles generally aligned well with participants’ subjective evaluations. This verification specifically focused on the hourly neighborhood maximum ensemble probability (NMEP) of composite reflectivity (REFC)  $\geq 40$  dBZ for each ensemble and how those forecasts compared to hourly MRMS composite reflectivity during the SFE. Initial results show that the REFS and MPAS ensembles demonstrated similar performance metrics to the HREF during the SFE, with all ensembles achieving a maximum CSI around 0.26 (Fig. 37). The EMC-REFS generally exhibited a decreased FAR compared to the HREF, but this was paired with a similar reduction in POD as well. Conversely, the REFS-SPC and REFS-SPC-ARW configurations were found to closely match the POD of the HREF on average but with a reduced FAR more in line with that of the EMC-REFS. These results suggest that the REFS-SPC and REFS-SPC-ARW configurations may best combine the benefits of both the HREF and REFS memberships and could be viable options for future REFS

development. All ensembles were found to overforecast REFC at the 40 dBZ threshold as shown in the reliability diagram (Fig. 37). However, the EMC-REFS demonstrated greater reliability than the other ensemble configurations through an NMEP threshold of about 40%. The REFS-SPC and REFS-SPC-ARW also demonstrated improved reliability over the HREF, but the MPAS REFS was found to overforecast REFC to a much greater extent than the other ensembles. Note that the aforementioned issue in the MPAS REFS UH probabilities did not impact the composite reflectivity field and thus should not be a factor in the objective evaluation.

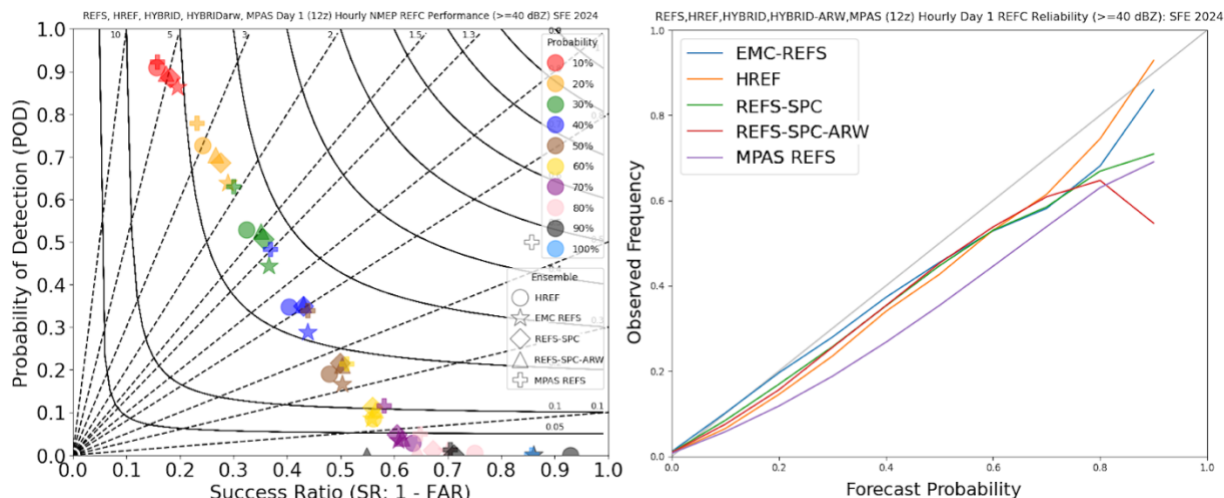


Figure 37. (left) Performance and (right) reliability of the five ensemble configurations with respect to NMEP composite reflectivity above 40 dBZ. Ensemble forecasts were compared to MRMS hourly composite reflectivity over the 5-week SFE evaluation period.

### 3.3.3 (E3) CLUE: 12Z Day 2 Ensemble Flagships

The E3 evaluation was identical to E2, except participants evaluated the five ensembles at Day 2 lead times. Specifically, respondents were asked to “Subjectively rate on a scale of 1 (Very Poor) to 10 (Very Good) the 24-h ensemble storm-attribute products during the Day 2 12-12Z period with regard to the quality of guidance for severe weather forecasting. Focus primarily on Updraft Helicity, but Updraft Speed & 10-m Winds can be used to supplement the ranking, especially on days without supercells.” Participants were given access to the same fields and observations as before, but the forecast products were derived from the previous day’s 12z ensemble runs. As such, this evaluation focused on ensemble forecast quality at forecast hours 24 - 48.

The EMC-REFS configuration received the highest mean score of 6.29 at Day 2 lead times (Fig. 38). The REFS-SPC-ARW ranked second at 6.24, followed by the HREF (6.22), REFS-SPC (6.15), and MPAS REFS (5.50). As before, most differences in the mean scores fell within the 95% confidence interval and were not found to be statistically significant. Notably, most ensemble configurations only saw a mean rating decrease of 0.4 – 0.6 when compared to their ratings at Day 1 lead times, suggesting impressive

consistency in forecast quality at longer lead times. The one exception to this was the MPAS REFS which saw a decrease from 6.4 in the Day 1 evaluation to 5.5 at Day 2. As discussed in the E2 results, an issue in how the MPAS REFS UH neighborhood probabilities were calculated may have artificially reduced the magnitude of those probabilistic forecasts and may have impacted participants' ratings of that ensemble. However, this issue was consistent at both Day 1 and Day 2 lead times, so it is not clear if or to what extent the issue may be related to this drop in ratings. All ensembles received a minimum rating of 2 at some point during the SFE, and all but the EMC-REFS had a maximum rating of 10, despite it having the highest average rating of the ensembles.

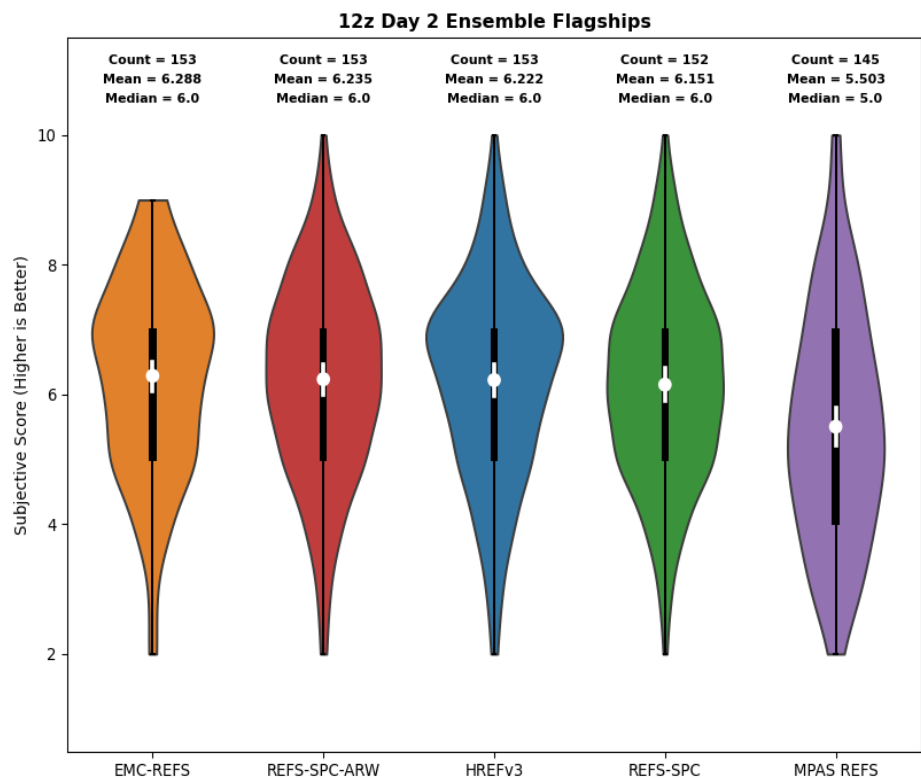


Figure 38. As in Figure 35, but for Day 2 lead times.

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

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

probabilities can be utilized to adjust or fine-tune the overall ratings.” An example of the forecasts is shown in Figure 39 for 21 May 2024 when a significant severe weather outbreak occurred over the upper Midwest that included an EF4 tornado that impacted Greenfield, IA. For this case, at all lead times examined NCAR FV3 failed to highlight the area of highest impact severe weather that occurred in southwest Iowa, while NCAR MPAS highlighted this area quite well, and at Day 3 even included a bullseye of 90%+ probabilities co-located with where the EF4 Greenfield, IA tornado occurred.

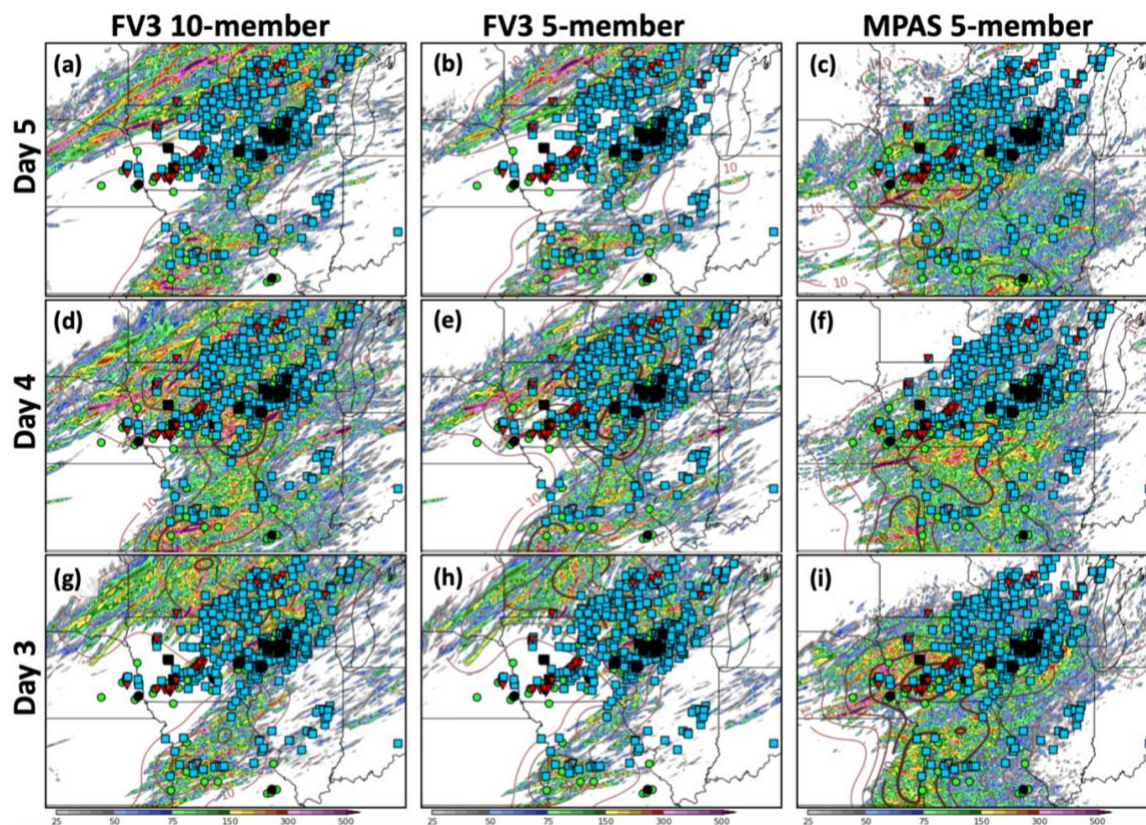


Figure 39. Example of multi-panel comparison webpage for the E4 Medium-Range Lead Time/Core/Members evaluation. In each panel, 24 h maximum UH (shaded) and neighborhood probability of UH  $\geq$  99.85th percentile (contours) is displayed. LSRs are also overlaid (wind – blue squares, hail – green circles, and tornado red upside-down triangles; significant reports are filled in black). All forecasts displayed are valid 12Z 21 May – 12Z 22 May 2024.

In comparisons of aggregate subjective ratings between the NCAR MPAS and 5-member NCAR FV3 ensemble subset (Fig. 40), the NCAR MPAS ensemble performed best at every lead time it was available, and the differences in mean subjective ratings were statistically significant for 3- and 4-day lead times, but not 5-day (Student’s t-test with  $\alpha = 0.05$ ). By far, the biggest advantage in MPAS occurred at the 3-day lead time and many participant comments noted MPAS having the biggest advantage at these shorter lead times. A couple of these comments included: “... NCAR-MPAS locked on as early as Day 4, which is very notable! The FV3 struggled the entire time to grasp the location and timing of the severe weather potential ...”, and “... It seemed like the FV3

may have done better in the longer range but then the MPAS became better in the short range. That seemed like a robust signal ...”.

Many comments also noted eastward and northward biases in the FV3 forecasts, for example, “... The FV3 based model has too much activity north and east of where it actually occurred ...”, “... The MPAS-based forecasts were clearly superior in placement and magnitude of the threat. FV3-based products had the threat displaced too far north and east, even at shorter lead times ...”, “... NCAR FV3 on day 6/7 continued to be too far north with its predictions of UH as much of the weather was further south. UH forecasts became progressively better in the day 4 and day 3 outlooks ...”, and “... Everything in the FV3 was fairly far east, which is not great. MPAS did a better job shifting everything west, but not really until Day 3 ...”.

Similar to SFE 2023, it was noted in SFE 2024 that on some days there was not a consistent run-to-run improvement in the forecasts. For example, although day 3 forecasts were on average rated highest, sometimes the day 5 forecasts would outperform the day 3 and 4 ones. A couple comments highlighting this behavior included, “... Day 5 is my favorite forecast, with a sudden degradation at Day 4, and models are not returning to their Day 5 performance even by day 3. MPAS has an edge over FV3 better capturing the hazard reports to the west ...” and “... The FV3 ensembles were pretty good on day 5, but seemed to actually lose value on day 4. MPAS had a good read on the highest risk areas with observed severe hazards by day 4. On day 3, MPAS had the best handle on the exact location of the most reports ...”.

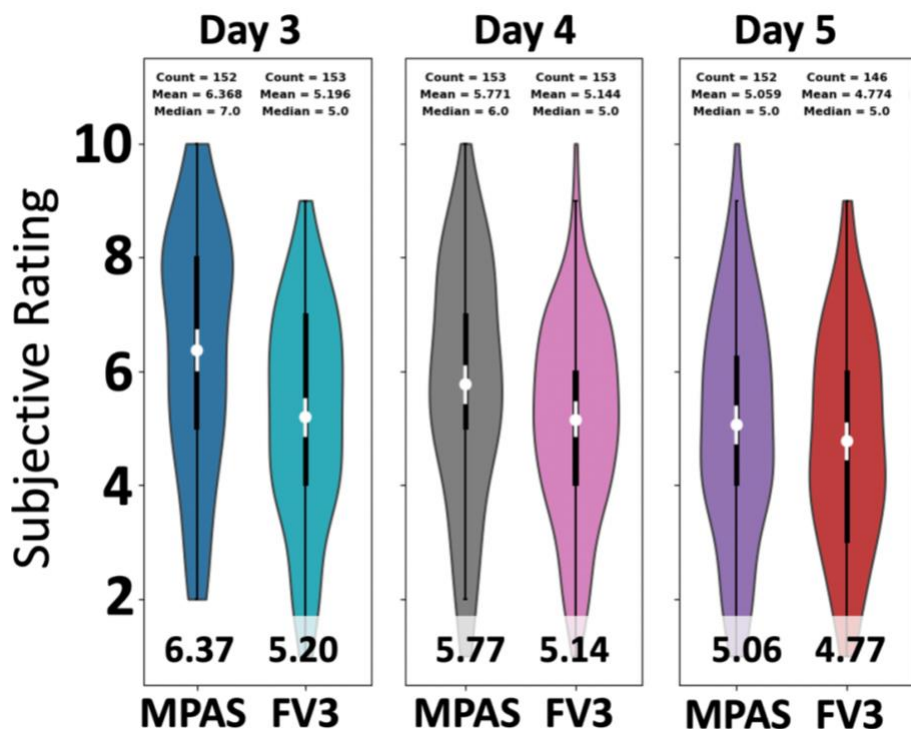


Figure 40. Distributions of subjective ratings for the NCAR MPAS and 5-member NCAR FV3 ensemble subset at Day 3 (left), Day 4 (middle), and Day 5 (right) lead times. Mean subjective ratings are indicated at the bottom of each violin plot. The white dots represent the mean scores for each ensemble, and the white bars indicate the 95% confidence intervals for each mean.

For the 10-member NCAR FV3, there was degradation in mean subjective ratings with increasing lead time, which seemed to accelerate at day 5 and beyond (Fig. 41). Relative to SFE 2023, the day 6 and 7 NCAR FV3 forecasts performed worse, but the day 3-5 forecasts were rated slightly higher (Table 4). Also, the 10-member NCAR FV3 forecasts were rated very similar to the 5-member NCAR FV3 forecasts. For NCAR MPAS, the day 3 and 4 forecasts were an improvement in SFE 2024, with an especially big jump in mean ratings at day 3. However, the day 5 NCAR MPAS forecasts were rated slightly lower in SFE 2024 (Table 4).

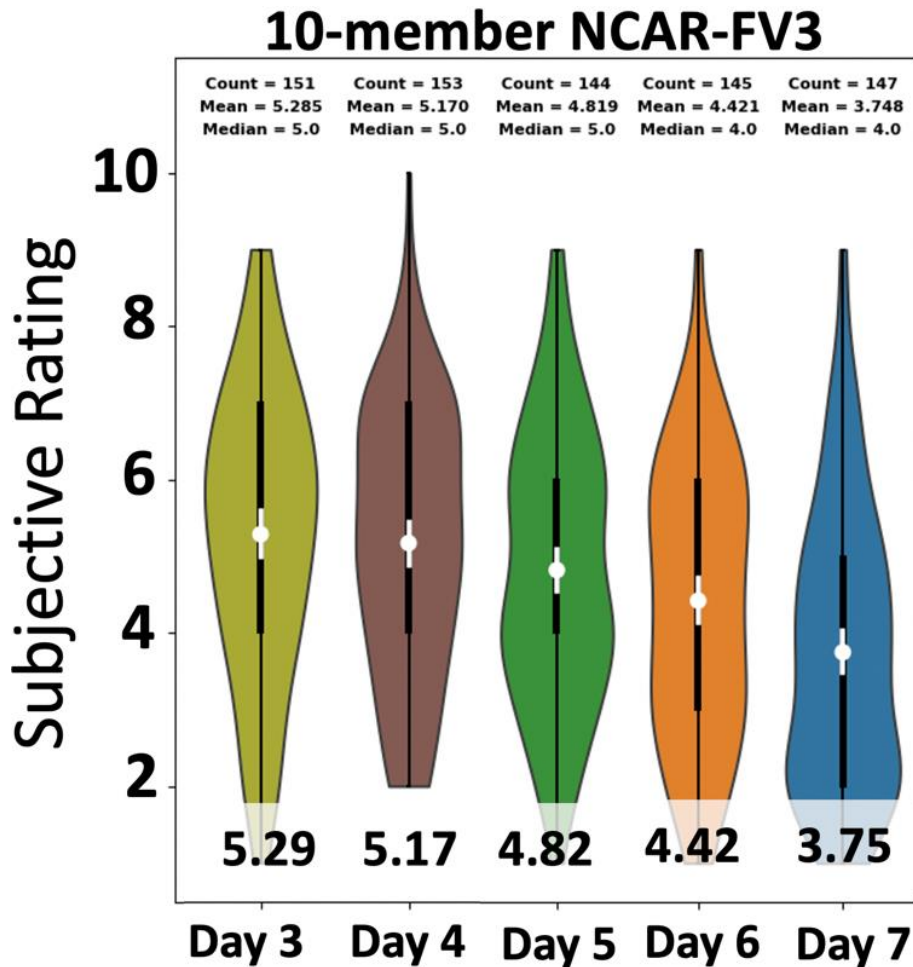


Figure 41. Distributions of subjective ratings for the 10-member NCAR FV3 ensemble for lead times of Day 3 to Day 7. Mean subjective ratings are indicated at the bottom of each violin plot. The white dots represent the mean scores for each ensemble, and the white bars indicate the 95% confidence intervals for each mean.

			2024	2023
Day 3	MPAS	↑	6.37	5.54
	FV3 (5-member)	↑	5.20	5.00
	FV3 (10-member)	↑	5.29	5.10
Day 4	MPAS	↑	5.77	5.46
	FV3 (5-member)	→	5.14	5.16
	FV3 (10-member)	↑	5.17	5.07
Day 5	MPAS	↓	5.06	5.13
	FV3 (5-member)	↑	4.77	4.61
	FV3 (10-member)	↑	4.82	4.64
Day 6	FV3 (10-member)	↓	4.42	4.50
Day 7	FV3 (10-member)	↓	3.75	4.50

Table 4. Average subjective ratings for the NCAR MPAS, NCAR FV3 (5-member), and NCAR FV3 (10-member) ensembles for 2024 & 2023. Green upward pointing arrows indicate mean ratings that increased from 2023 to 2024, red downward arrows indicate a decrease, and the gray sideways arrows indicate little change.

### 3.4 Evaluation – (A)nalyses

#### 3.4.1 (A1 & A2) Mesoscale Analysis Background

Two hourly versions of 3D-RTMA with different backgrounds were subjectively evaluated by participants during the 2023 HWT SFE. The evaluation was performed to assess the quality and utility of these analysis systems for situational awareness and short-term forecasting of convective-weather scenarios. The 3D-RTMA RRFS used the FV3-based RRFS control member as the first-guess background while the 3D-RTMA HRRR used the operational HRRR for first-guess background and serves as the baseline for this evaluation. The hourly analyses for composite reflectivity, 2-m temperature and dewpoint (Fig. 42), SB/ML/MUCAPE, and the significant tornado parameter (STP) were examined during the 18-03Z period on the following day. The SFE participants were tasked with looking through all of these fields during this period and arrive at a single rating of the quality of the 3D-RTMA RRFS compared to the 3D-RTMA HRRR using a five-point Likert scale (i.e., much worse, slightly worse, about the same, slightly better, or much better).



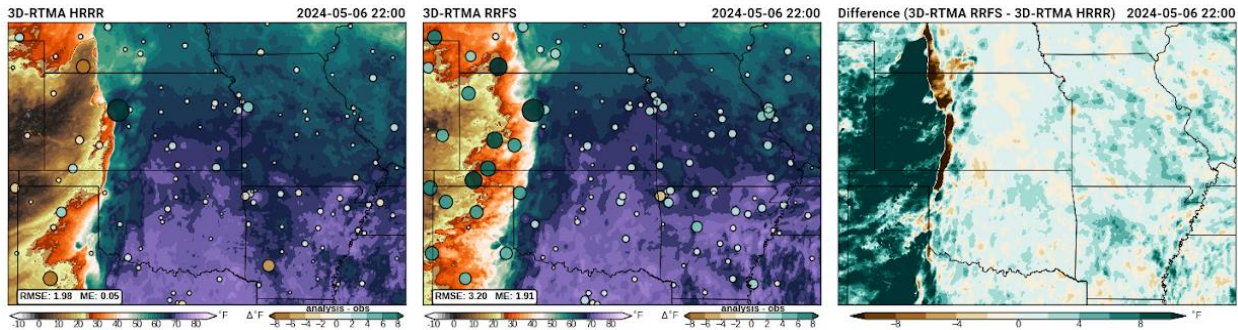


Figure 42. Example of the website comparison page for the 3D-RTMA during the 2024 HWT SFE. The 3D-RTMA HRRR baseline is shown in the left panel, the 3D-RTMA RRFS is in the middle panel, and the difference plot (3D-RTMA RRFS - 3D-RTMA HRRR) is shown in the right panel. The 2-m dewpoint analysis valid at 22Z on 6 May 2024 is shaded in the left and middle panels. The difference (analysis-obs) at METAR sites is shown by the size and shading of the dots in the left and middle panels.

In general, the two versions of 3D-RTMA were typically similar to one another with the 3D-RTMA RRFS having slightly larger errors over the domains in 2-m temperature and dewpoint. As seen in prior years, the biggest differences in the 2-m temperature field were most commonly associated with effects from convection although those differences were much reduced compared to previous years, owing to fewer spurious storms in this year’s RRFS-based version (with the GF convective parameterization scheme). In terms of the overall subjective ratings from SFE participants, the majority of responses indicated the 3D-RTMA RRFS was slightly worse to about the same as the HRRR-based version (Fig. 43). The participants noted some common issues in the RRFS-based version: overall 2-m moist bias across the domain (including the warm sector; Fig. 42), too moist in dry air (e.g., behind dryline; Fig. 42), and a low CAPE bias overall.

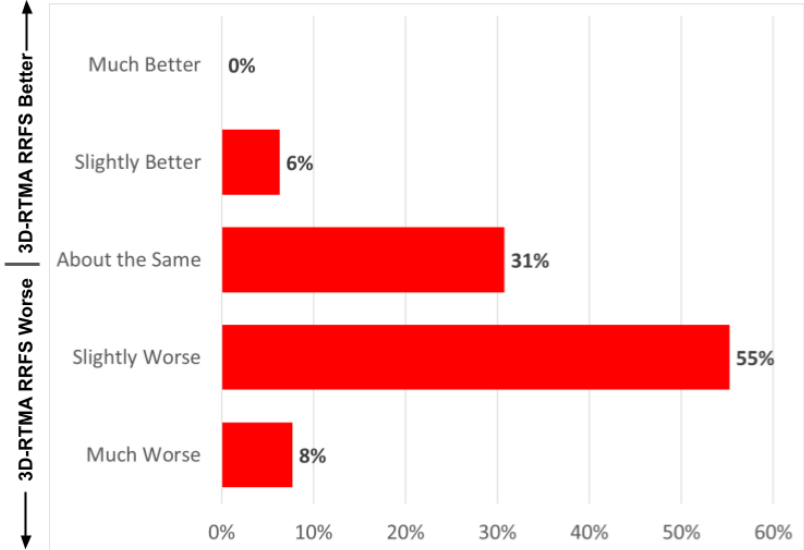


Figure 43. Percentage of subjective ratings by SFE participants for each rating category (Much Worse, Slightly Worse, About the Same, Slightly Better, and Much Better) of the 3D-RTMA RRFS compared to the 3D-RTMA HRRR.

New this year to the SFE was another 3-km analysis system based on the HRRR. This system, called surface objective analysis (sfcOA) HRRR, performed a simple 2-pass Barnes analysis on surface observations with the HRRR analysis as the first-guess field, then directly uses the HRRR analysis for the atmospheric state above the surface. This system provides a risk-reduction option given uncertainties on 3D-RTMA implementation (timeline, version, etc.) and offers a low-latency option when paired with a 1-h HRRR forecast. The sfcOA HRRR performance was similar to the 3D-RTMA HRRR as indicated by the majority of participants rating the systems about the same (Fig. 44). This is a very nice result and validates the viability of the simplistic sfcOA HRRR approach for real-time mesoscale analysis for short-term severe weather applications.

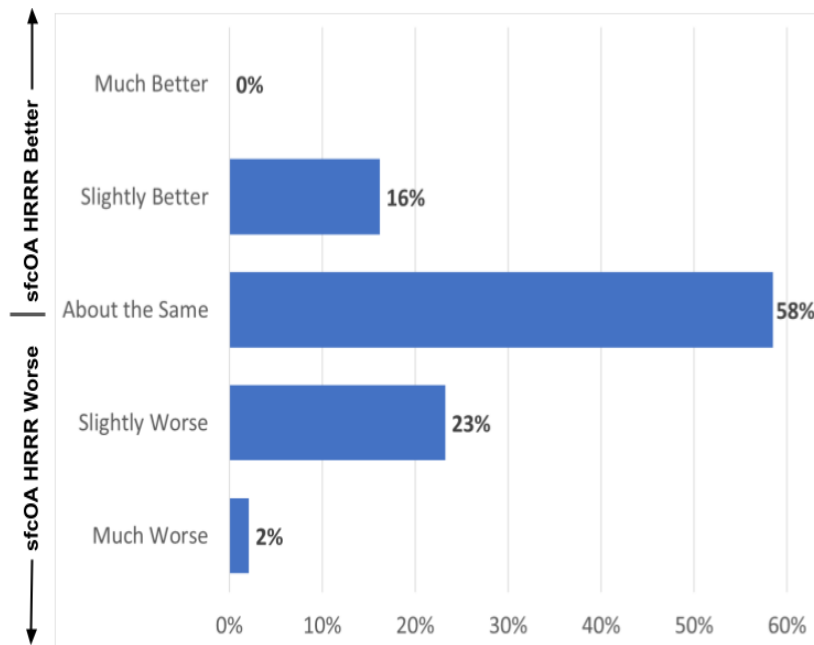


Figure 44. Same as Fig. 43, except for the sfcOA HRRR compared to the 3D-RTMA HRRR.

In addition to examining the 3D-RTMA systems at their native resolution (i.e., 3-km grid spacing), participants also examined upscaled analysis versions (i.e., 40-km grid) for comparison to the widely used SPC RAP-based mesoanalysis. This was a practical exercise to determine the readiness of the 3D-RTMA systems and sfcOA HRRR to replace the functionality and capacity currently served by the SPC mesoanalysis. SFE participants were asked to rank (from best to worst; i.e., 1 to 4) the overall quality of the analysis for situational awareness and short-term forecasting, despite the limitation of not having observational truths for all of the fields. The sfcOA HRRR and 3D-RTMA HRRR had the lowest (i.e., best) mean rankings of 1.9, while the RAP-based SPC mesoanalysis came in second with a mean ranking of 2.9, and the 3D-RTMA RRFS had the lowest mean ranking of 3.3 and the largest number of fourth-place rankings. This provides additional supporting evidence that a HRRR-based analysis system can replace and likely improve upon the current role served by the RAP-based SPC mesoanalysis system.

### 3.4.2 (A3) Storm Scale Analysis

The WoFS was used to explore whether a high resolution, rapidly updating ensemble DA system can serve as a verification source for severe weather. Specifically, the 15-minute maximum forecasts of 80-m winds, 2-5 km AGL UH, and column-maximum updraft speed from WoFS (cycled every 15 minutes) were used as a proxy for the analysis (i.e., ground truth) of storm attributes associated with severe weather, as opposed to relying on point-based storm reports. The WoFS ensemble analysis fields were accumulated from 18Z through 03Z for comparison with MRMS-derived products [composite reflectivity, midlevel rotation tracks, and maximum estimated size of hail (MESH)] and preliminary local storm reports, (Fig. 45).

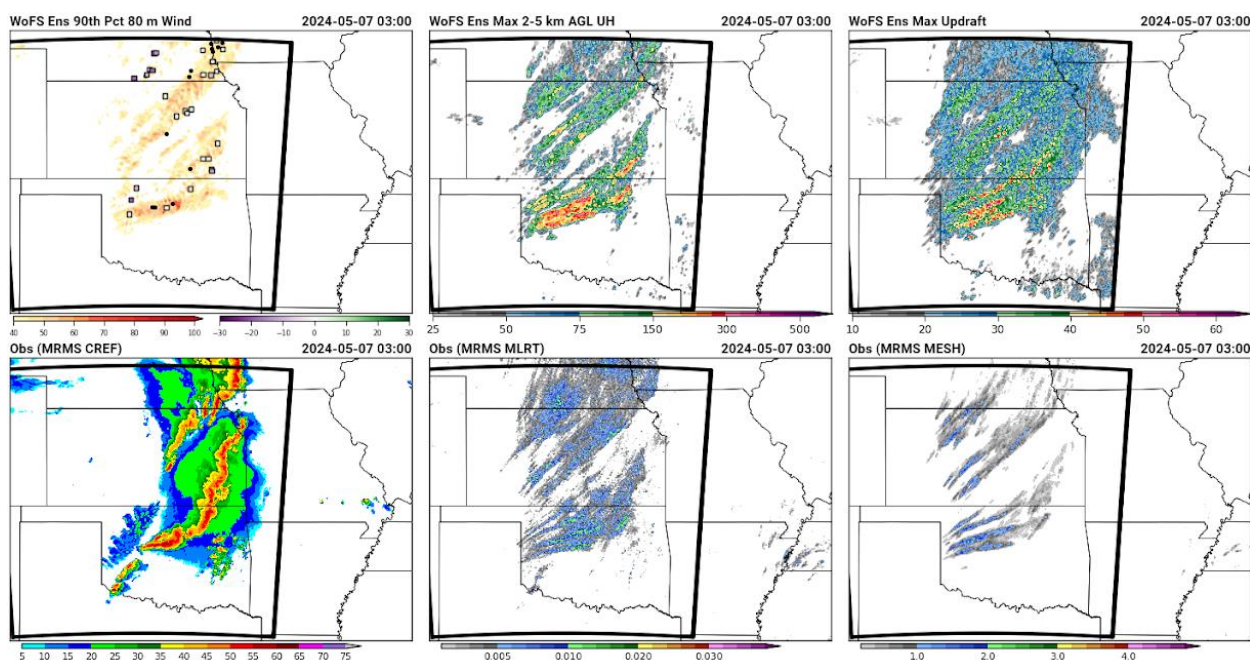


Figure 45. Example of the website comparison page for the WoFS analyses during the 2024 HWT SFE. The 18Z 6 May – 03Z 7 May accumulated ensemble 90th percentile 80-m wind is shown in the upper-left panel, the ensemble maximum 2-5 km AGL UH in the upper-middle panel, and the ensemble maximum column-maximum updraft speed in the upper-right panel. The observed MRMS composite reflectivity is in the bottom-left panel, observed MRMS midlevel rotation tracks are in the bottom-middle panel, and the MRMS MESH is in the bottom-right panel. In the upper-left panel, the wind damage reports are the black circles while the measured gusts are the open squares shaded by the difference (analysis-obs) of the gust measured at that location.

The goal of the evaluation was to assess the current capability of WoFS to produce output for diagnosing severe weather. Overall, the WoFS ensemble analysis fields were positively viewed in terms of lining up with radar-derived proxies of severe weather, preliminary local storm reports, and a subjective assessment of severe weather based on the environment. Overall, the WoFS analyses of column-maximum updraft speed received lower subjective ratings than 2-5 km AGL UH and 80-m winds in terms of alignment with severe-weather occurrence (Fig. 46), but the ratings were still neutral to positive. The lower participant ratings for updraft speed were largely owing to the

broader-than-observed footprint of stronger storms, so there is room for improvement in terms of optimizing the analysis product, which simply uses the ensemble maximum. Overall, the participants found this to be an interesting and promising approach for using a rapidly cycling convection-allowing ensemble system.

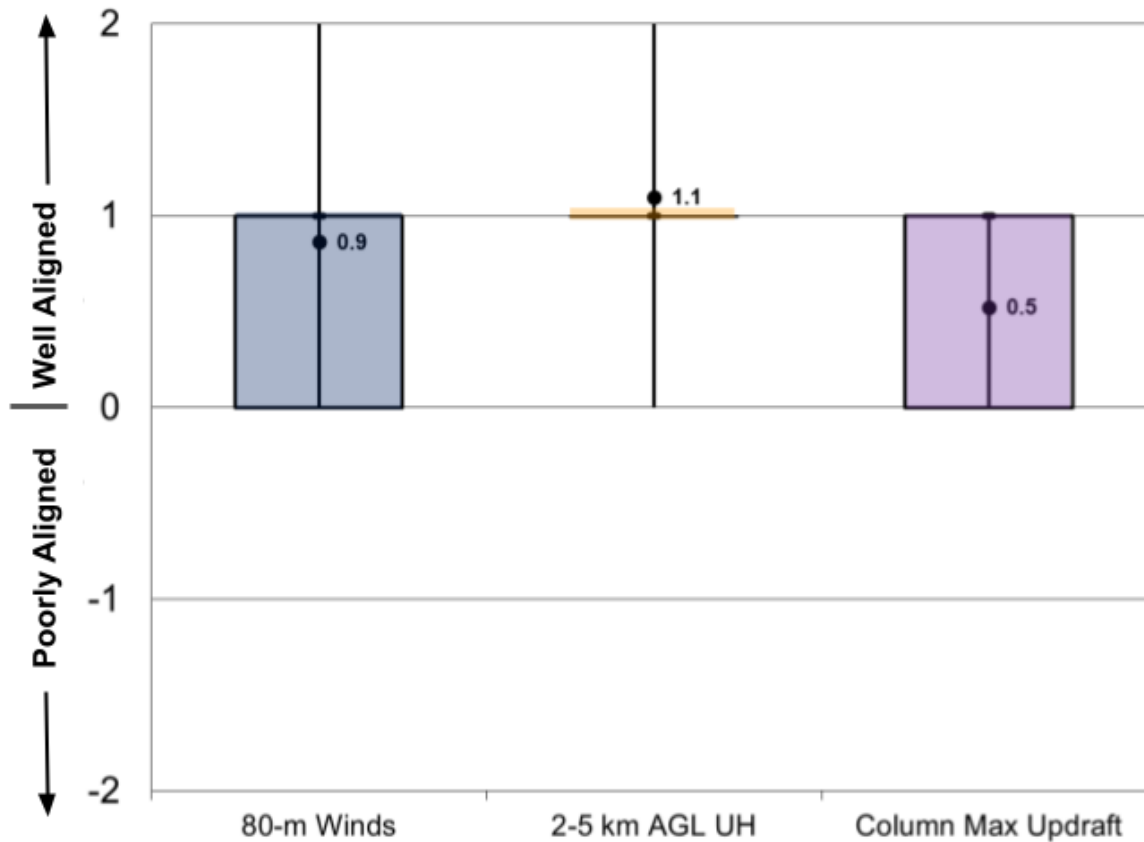


Figure 46. Distributions of subjective ratings (-2 to +2) by 2024 SFE participants of the WoFS storm-scale analysis for ensemble 90th percentile 80-m winds (blue), 2-5 km AGL UH (light orange), and column-maximum updraft speed (light purple), where the ratings represent how well the WoFS analyses align with the MRMS observed fields and preliminary severe wind reports: -2 - Very Poorly; -1 - Poorly; 0 - Unsure/Neutral, neither poorly nor well; 1 - Well; 2 - Very Well.

### 3.5 (A)rtificial (I)ntelligence Evaluations

#### 3.5.1 (AI1) First-Guess Watch Guidance

This HREF-based machine-learning product provides first-guess guidance as to where and when environmental conditions might warrant the issuance of a severe thunderstorm or tornado watch with up to three hours of lead time prior to severe weather occurrence. SFE participants evaluated the guidance for the previous day and over the primary SFE forecast domain for that day. Recommended watches based on the 12Z HREF were compared to operational, SPC-issued watches in addition to any NWS-issued warnings and LSRs over the period. An example of the watch guidance product is given in Figure 47.

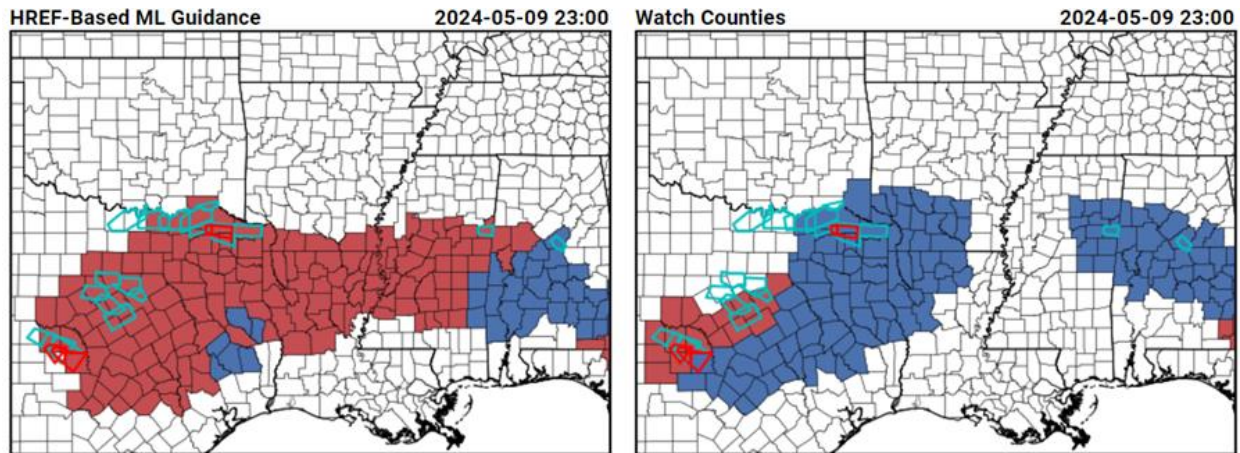


Figure 47. An example of tornado (red) and severe thunderstorm (blue) watches by county predicted by ML guidance (left) and issued by SPC (right) valid for 20240509 2300 UTC. Polygons indicate NWS tornado (red) and severe thunderstorm (blue) warnings.

Overall comments made by participants were favorable. Operational forecasters, in particular, indicated that the watch guidance was useful; it provided a quick initial synopsis of HREF data as to where watches may be needed within the current 24-hour forecast cycle. The median rating of 7 for placement and timing (Fig. 48) indicates a general positive consensus that this product corresponded with the location and progression of severe weather. Comments were given, however, that the overall spatial coverage of the guidance was often too large. This effect is seen in Figure 47 where the watch guidance predicted a continuous area of watches from central Texas through central Alabama. In reality, SPC forecasters initially left a gap in the operational watches that excluded the lower Mississippi River valley. However, SPC did eventually issue a watch for the area several hours later. The sometimes early and broad nature of the watch guidance may be in part an artifact of its design to provide up to 3 hours of lead time prior to the first severe weather occurrence. This is generally a longer lead time goal than what is employed for operational SPC-issued watches, and so the guidance will often recommend watches earlier by design.

One detracting factor often identified by participants was an apparent bias toward tornado watches over severe thunderstorm watches. For the case in Figure 47, the guidance primarily predicted tornado watches while SPC mostly issued thunderstorm watches. The vast majority of storms were non-tornadic during this event, but one tornado warning was issued in far northeast Texas where a tornado watch was recommended. Participants evaluated two aspects of the guidance's ability to predict watch type. The first aspect, how well the predicted watch type aligned with the SPC-issued watches, received a median score of 5. The second aspect asked participants to rate how appropriate the predicted watch type was for the observed hazards, regardless of what watch types were operationally issued. Participants provided slightly higher ratings for this consideration, with a median score of 6. These ratings, combined with comments made during an open discussion period, revealed mixed opinions on the guidance's ability to appropriately recommend watch type.

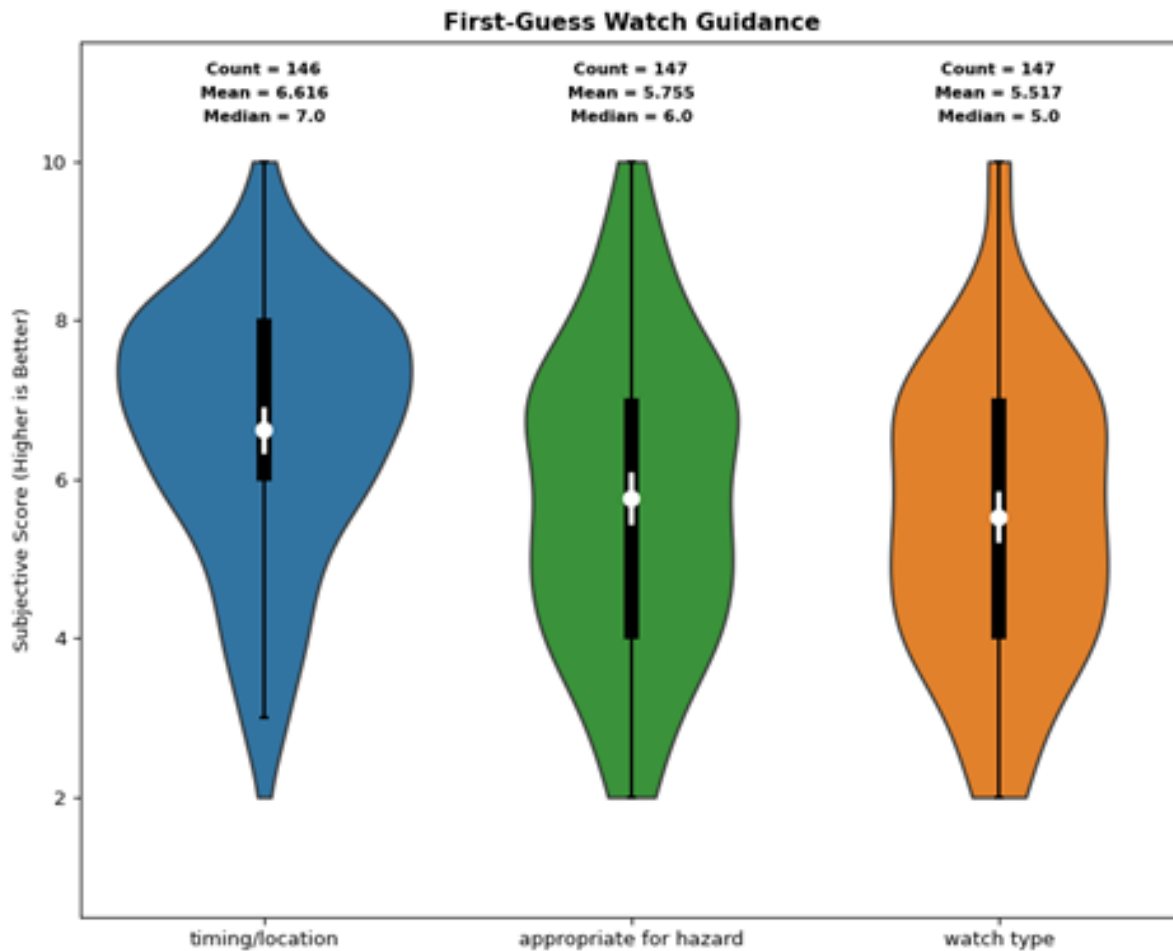


Figure 48. Violin plots showing participant ratings of the watch guidance in consideration of: the recommended timing and location (blue), how appropriate the recommended type was for the observed hazards (green), and how closely the guidance matched the SPC-issued watch type (yellow).

### 3.5.2 (AI2) Global NWP Emulators

This evaluation asked participants to assess the skill of three AI-generated global NWP emulators at 7-day lead times. These forecasts were compared to the operational GFS, and the hourly GFS analysis was provided as an “observational” dataset. The three AI models were comprised of publicly available versions of Google’s GraphCast, Huawei Cloud’s Pangu-Weather, and NVIDIA’s FourCastNet v2. Evaluations focused primarily on forecasts of 500 mb geopotential height and winds (Fig. 49), but 850 mb geopotential heights, 2-m temperature, 10-m winds, and MSLP were also considered.

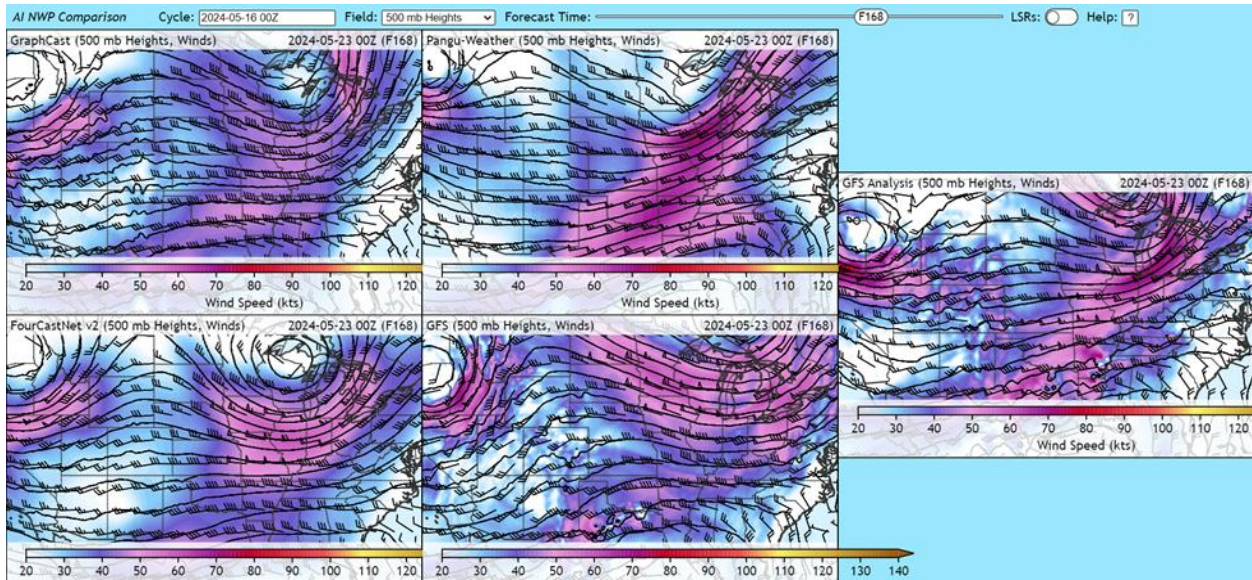


Figure 49. 500 mb height [m] and wind speed [kts] forecasts valid for 20240516 0000 UTC. Models depicted are GraphCast (top left), Pangu-Weather (top center), FourCastNet (bottom left), and GFS (bottom right). The hourly GFS analysis (far right) is included for verification.

SFE participants noted that the AI models often demonstrated skill in predicting the synoptic-scale pattern out to seven days, sometimes to an impressive degree. For example, the case in Figure 49 shows the operational GFS and the three AI models all accurately depicted a relatively strong trough over Wisconsin and an associated wind max at 500 mb. Even so, participants generally favored the GFS forecast which received a mean score of 6.5 over the course of the experiment (Fig. 50). GraphCast was rated the highest of the AI models with a mean score of 5.8, followed by Pangu-Weather (5.4), and FourCastNet v2 (5.1). GraphCast was the only AI model found to be statistically similar to the GFS, where Pangu-Weather and FourCastNet v2 were rated statistically worse on average. Participants expressed some concern that the AI-generated forecasts were overly smoothed and often failed to resolve more subtle shortwave troughs that were present in the GFS and GFS analysis. This was particularly noted during the second half of the SFE when the pattern over North America transitioned from a strongly forced, strongly sheared environment, to one characterized by weaker forcing and a split-flow regime. In general, the AI models were found to provide little guidance for severe weather events that were triggered by more subtle forcing mechanisms, and the GFS was largely favored in these situations.

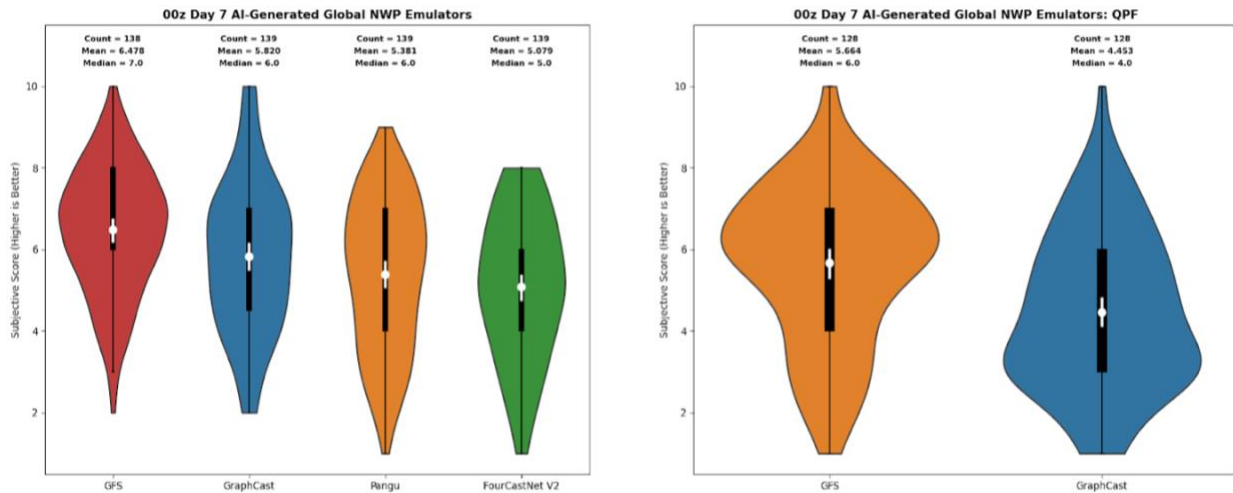


Figure 50. Participant ratings of the operational GFS and AI-generated global NWP 7-day forecasts of (left) 500-mb geopotential height and wind, and (right) 6-h QPF. The white dots represent the mean scores for each ensemble, and the white bars indicate the 95% confidence intervals for each mean.

GraphCast was the only AI model to provide a 6-h QPF forecast at the time of this experiment. When compared to the operational GFS and MRMS rainfall observations, GraphCast was rated significantly worse (at the 95% confidence level) when compared to the GFS (Fig. 50). Participants again noted the QPF forecasts tended to be too broad and overly smoothed, generally causing the model to miss areas of observed heavy precipitation. While the GFS often had large spatial errors at 7-day lead times, it did at least provide some indication of heavier rainfall potential and thus received higher ratings each day.

Finally, participants evaluated the vertical consistency of the AI models by assessing derived sounding profiles. These profiles were not evaluated using numerical ratings, but participant comments suggested that the vertical structure of temperature and dewpoint in the models were physically realistic and reasonable when compared to the operational GFS. The AI models did each provide fewer vertical levels than the GFS, and the lower vertical resolution resulted in rather smooth profiles that often struggled to resolve capping inversions or the EML. Even so, participants tended to somewhat favor the sounding profiles of Pangu-Weather over the other AI models.

### 3.6 (O)utlook Evaluations<sup>1</sup>

#### 3.6.1 (O4 & O5) Probabilistic 0-1 & 1-2 h Outlooks

In this forecasting activity, participants were divided into two groups: one with access to Warn-On-Forecast System-Probabilistic Hazard Information (WoFS-PHI) machine learning severe weather probabilities and one without. All participants created 0-1- and 1-2-h probability forecasts for each severe weather hazard. Most participants

<sup>1</sup> Note, Results for O1-3 are not included in this report because there were too few responses.



completed this activity over two hours in the afternoon each day. However, 2-5 National Weather Service (NWS) forecasters (referred to as “lead forecasters”) per week received additional training on WoFS-PHI and extended the activity for 4 hours into the evening on Monday-Thursday. Forecasts from lead forecasters and consensus forecasts from all other participants were subjectively evaluated on a scale from 1 (worst) to 10 (best) by SFE participants the next day.

Preliminary results suggest that those with access to WoFS-PHI produced higher-rated forecasts than those without access to WoFS-PHI. This was true when ratings from all hazards were pooled together (Fig. 51a) as well as for individual hazards (Fig. 51b-d). This was also true regardless of participant type (i.e., lead forecaster or participant consensus) or forecast lead time (0-1- or 1-2-h). In general, the effect was modest but noticeable, with the WoFS-PHI forecasts tending to score around 0.5 points higher than the non-WoFS-PHI forecasts (Fig. 51). For example, when ratings for all forecasts, hazards, and lead times were combined, the WoFS-PHI forecasts had mean ratings 0.45 points better than the non-WoFS-PHI forecasts (Fig. 51a). For hail, wind, and tornadoes, the overall mean rating differences were 0.3, 0.41, and 0.74 in favor of the WoFS-PHI forecasts, respectively (Fig. 51b-d).

Overall, having access to WoFS-PHI benefited the participant consensus forecasts slightly more than those made by the lead forecasters (mean subjective rating difference of 0.63 vs. 0.41; Fig. 51a). For wind, the participant consensus forecasts benefited from WoFS-PHI for wind much more than those from the lead forecasters (mean rating difference of 1.02 vs. 0.27; Fig. 51c), while for tornadoes the reverse was true (mean rating difference of 0.53 for participant consensus vs. 0.81 for lead forecasters; Fig. 51d). For hail, both the consensus forecasts and those from the lead forecasters benefited about equally from WoFS-PHI, with mean rating differences of 0.31 and 0.29, respectively; Fig. 51b).

Pooling ratings and stratifying by lead time showed that the 1-2-h forecasts made with WoFS-PHI tended to receive similar ratings as the 0-1-h forecasts made without WoFS-PHI (Fig. 51). These results, though preliminary, are exciting, and suggest that WoFS-PHI could help forecasters gain an extra hour of lead time, which could potentially help extend the lead times of severe weather warnings. Although a bit of an outlier compared to the other hazards, the tornado ratings (Fig. 51d) were particularly striking, as the ratings of the 1-2h WoFS-PHI forecasts were actually greater than those from the 0-1h forecasts made without WoFS-PHI. Note, all differences between pairs of “ML” and “No ML” forecasts in Figure 51 were statistically significant except for the consensus wind (Fig. 51c) and tornado (Fig. 51d).

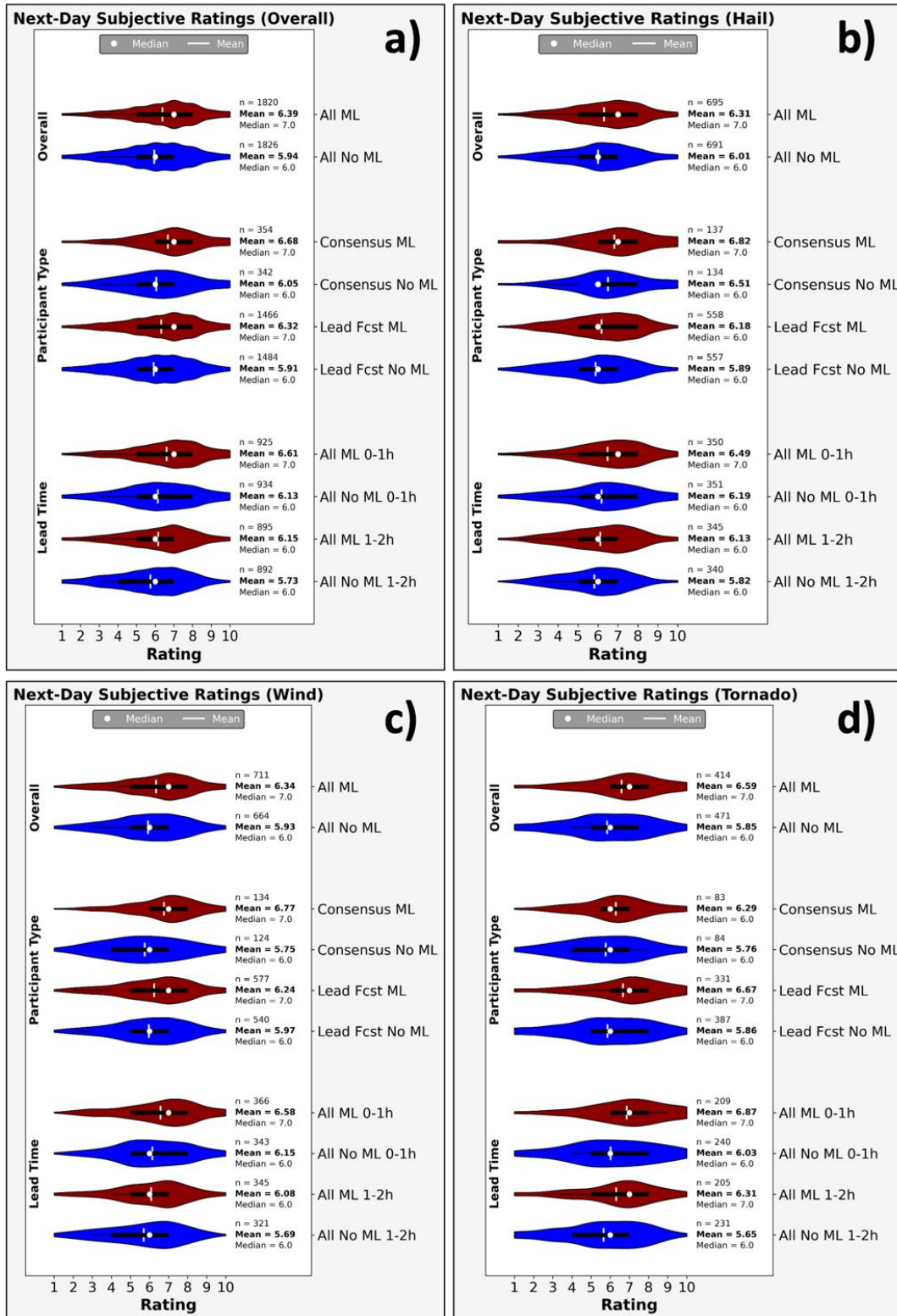


Figure 51. Violin plots of subjective participant ratings pooled over (a) all hazards, (b) hail, (c) wind, and (d) tornadoes. In each panel, the first set of violins are pooled over all participant types and lead times, the second set are stratified based on participant type, and the third are stratified by lead time. Violins for forecasts made with (without) WoFS-PHI are colored red (blue). Box-and-whisker plots are shown in black, with the vertical white bar indicating the distribution mean and the white dot indicating the median. “n” indicates the number of ratings used to create each violin.

In daily post-activity surveys, most participants indicated that WoFS-PHI had at least a moderate influence on both where they placed non-zero forecast probabilities (Fig. 52a) as well as their probability magnitudes (Fig. 52b). Moreover, most participants said they felt a moderate degree of trust in the WoFS-PHI products for a given day (Fig. 53a), with their trust in the WoFS-PHI products nearly as high as trust in the more established non-machine learning products on the WoFS Viewer (Fig. 53b). Although participants incorporated WoFS-PHI throughout the forecasting process, more participants used it at the beginning than in the middle or end (Fig. 54). Finally, survey responses indicated a clear preference for the version of WoFS-PHI trained on LSRs and warnings over the version trained only on LSRs (Fig. 55). This feedback will guide future WoFS-PHI development.

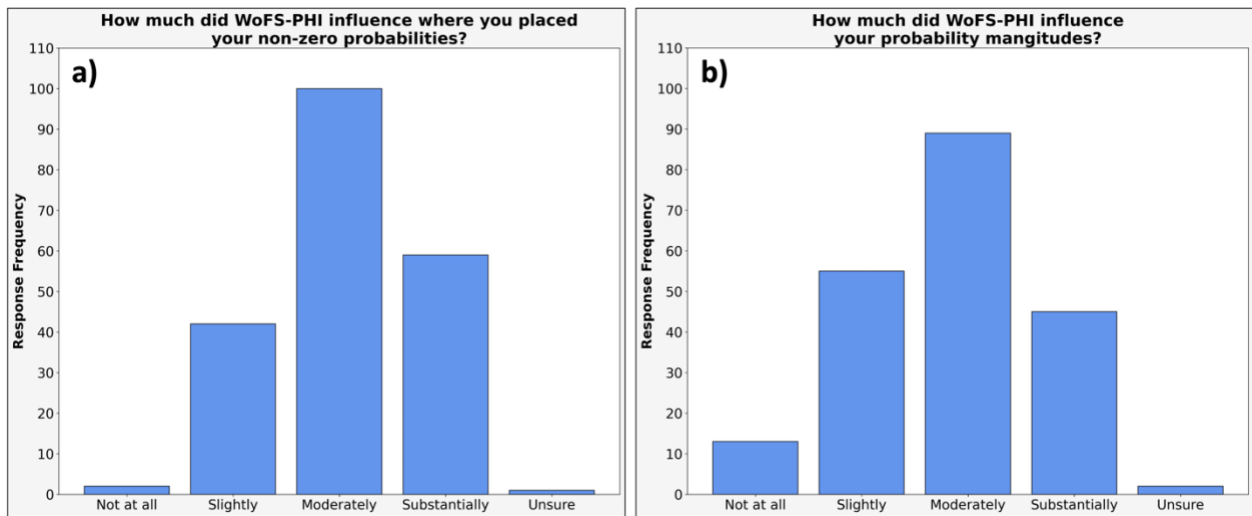


Figure 52. Histogram showing the degree to which participants indicated WoFS-PHI influenced (a) where they placed their non-zero forecast probabilities and (b) their forecast probability magnitudes.

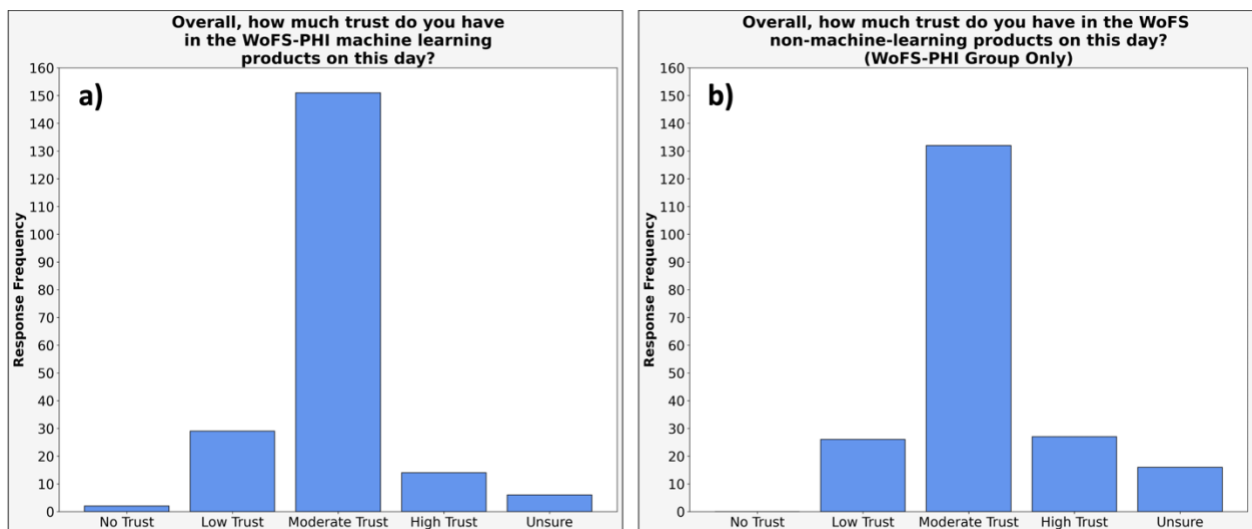


Figure 53. Histogram showing how much trust the WoFS-PHI group participants said they had in (a) the WoFS-PHI machine learning products and (b) the WoFS non-machine-learning products for a given day.

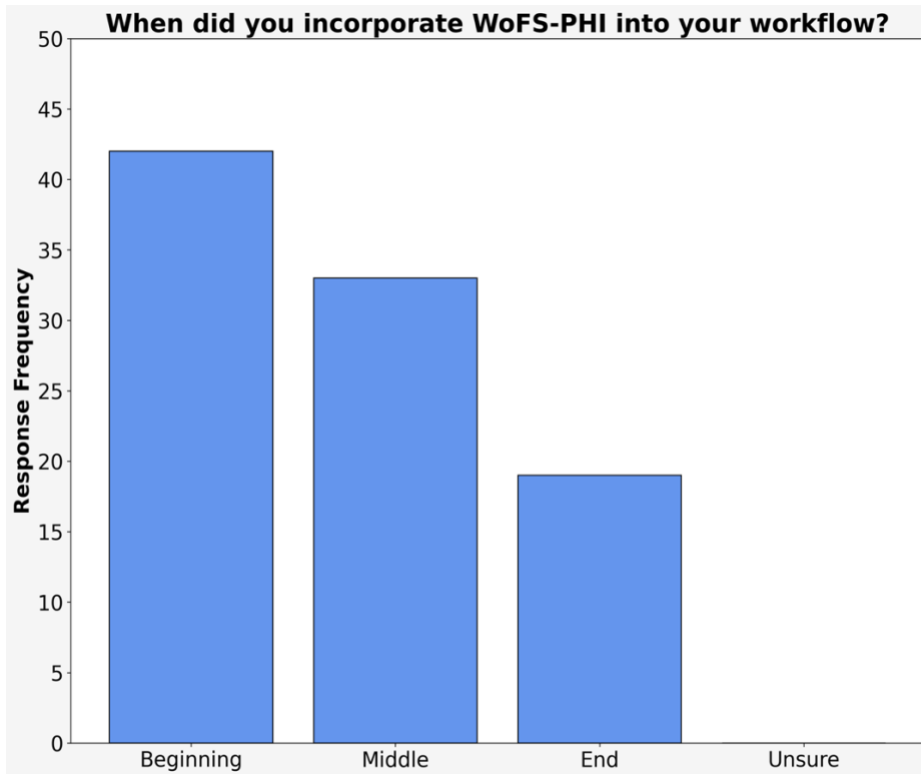


Figure 54. Histogram showing when participants indicated they incorporated WoFS-PHI into their forecasting workflow.

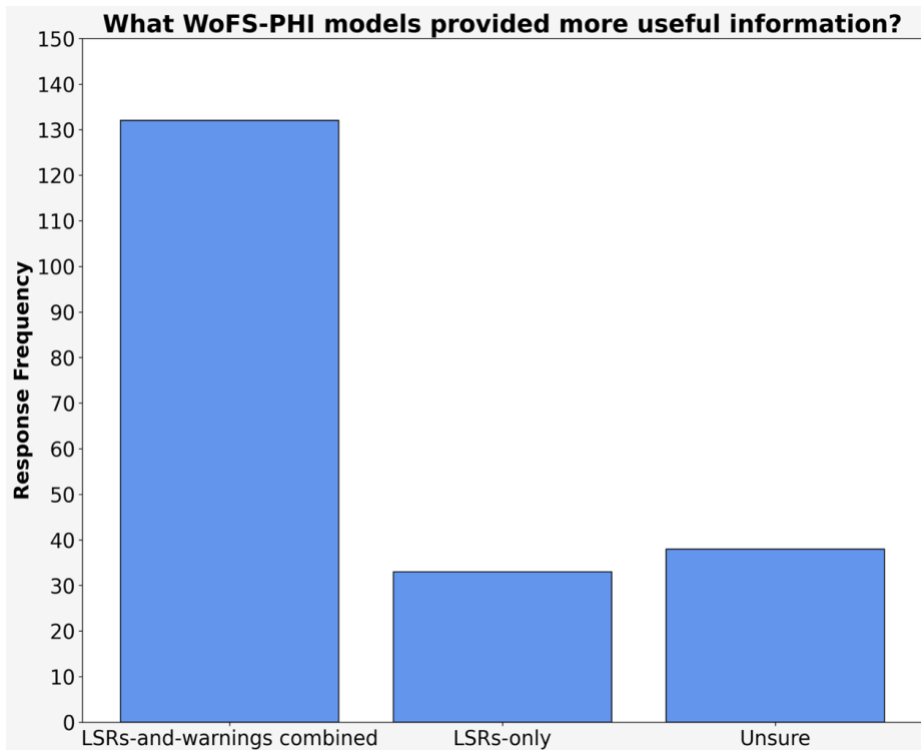


Figure 55. Histogram indicating the version of WoFS-PHI participants said provided more useful information: the version trained on both local storm reports (LSRs) and warnings (i.e., LSRs-and-warnings combined) or the version trained only on LSRs (LSRs-only).

## 4. Summary

The 2024 NOAA HWT Spring Forecasting Experiment (2024 SFE) was conducted from 29 April – 31 May by the SPC and NSSL with participation from forecasters, researchers, model developers, university faculty, and graduate students from around the world. The primary goals of the 2024 SFE were to (1) evaluate convection-allowing model and ensemble guidance for identifying optimal configurations of convection-allowing versions of FV3 and CAM ensembles, including several carefully designed and controlled experiments as part of the Community Leveraged Unified Ensemble (CLUE), (2) study how forecasters and meteorologists utilize CAMs and CAM ensembles, such as WoFS, and evaluate various experimental severe weather outlooks generated using WoFS and other CAM ensembles for lead times from one hour to 4 days, (3) evaluate different CAM ensemble post-processed guidance with an emphasis on those using machine-learning algorithms, and (4) conduct a preliminary assessment of AI-driven NWP emulators.

Several preliminary findings/accomplishments from the 2024 SFE are listed below:

- Examined and assessed various methods to produce first-guess calibrated probabilistic hazard guidance based on forecast output from HREFv3, GEFS, and the experimental NCAR-FV3 ensemble.
  - For timing guidance products at Day 1 and 2 lead times, the Nadocast-based guidance performed better than the HREF/GEFS-based guidance for tornadoes and wind, but both guidance versions performed similarly for hail.
  - In comparisons between a new 4-h (2-6 h lead time) WoFS-based ML hazard probability product and 12Z HREF-based Nadocast probabilities for the periods 20-00Z and 00-04Z, Nadocast received higher subjective ratings for each hazard.
  - Three algorithms for producing extended-range total severe forecasts for Days 3-7 were examined. Subjective ratings indicated that the random forest trained with operational GEFS performed best, while the neural network approach from NCAR performed significantly worse than both GEFS-based algorithms.
- Examined various **deterministic** CAM systems within the CLUE using HRRRv4 as a baseline.
  - In blinded 00Z Day 1 evaluations, HRRRv4 was the clear top performer for simulated reflectivity and UH as measured by average subjective ratings and objective metrics. HRRRv4 was also the top performer for environmental fields and QPF. The NSSL MPAS-HT was a close runner-up in all fields examined, while the three FV3-based CAMs (RRFS, NASA GEOS, and GFDL FV3) were rated noticeably lower.

- The RRFS control member performance in the blinded 00Z Day 1 evaluations was markedly worse than in the 2023 SFE, so that the performance gap between HRRR/MPAS and RRFS actually widened relative to the previous year.
- In blinded 12Z Day 2 evaluations, HRRRv4 received the highest average subjective ratings for all fields examined, but scores were more tightly clustered relative to Day 1. Similar to Day 1, NSSL MPAS HT was the runner up for all fields examined, except for 2-m temperature for which RRFS was runner up.
- In direct comparisons between 00Z-initialized RRFS and HRRR, RRFS was on average rated worse than HRRR for all fields examined which included reflectivity and UH, updraft speed, 10-m wind speed, 6-h QPF, SBCAPE, 2-m temperature, and 2-m dewpoint. Most notably, RRFS was found to struggle on days with the most significant severe weather, owing to erroneous suppression of deep convection by the GF convective parameterization scheme.
- In direct comparisons between 21 and 00Z initializations of RRFS and HRRR focused on 0-12 h lead times, the HRRR received the highest ratings at all lead times for each field examined. RRFS performance was markedly worse relative to SFE 2023, so once again the performance gap between HRRR and RRFS widened relative to the previous year.
- In comparisons between similarly configured 3- and 4-km versions of MPAS, subjective ratings from the 3-km MPAS were only slightly higher than the 4-km version and did not reach the threshold for statistical significance.
- In comparisons between a 1-km grid-spacing WRF-ARW configuration and the HRRR, the HRRR received higher subjective ratings for all fields, which included reflectivity and UH, 0-2 km AGL UH, and maximum 10-m wind.
- Examined various **ensemble** CAM systems within the CLUE using HREFv3 as a baseline.
  - In direct comparisons between 00Z initializations of REFS and HREF, subjective ratings were about the same for UH, updraft speed, and 10-m wind speed, while HREF had a slight edge for composite reflectivity. REFS received slightly higher subjective ratings than HREF for SBCAPE, 2-m temperature, and 2-m dewpoint.
  - The discrepancy in deterministic and ensemble results comparing HRRR/HREF and RRFS/REFS was related to the convective parameterization schemes utilized in the REFS. The REFS control member had a GF configuration that performed especially poor, which hurt the deterministic comparisons, while several of the perturbed REFS ensemble members had convective parameterization configurations that performed quite well, which helped the ensemble comparisons.

- For Day 1 and 2 comparisons between HREF and four different REFS configuration strategies, distributions in subjective ratings were very tightly clustered, with the exception of MPAS REFS, which had noticeably lower ratings. However, an improper UH percentile value artificially lowered the MPAS REFS probabilities, which negatively impacted the subjective scores. Objective scores for reflectivity forecasts confirmed that MPAS REFS performance was similar to the other ensembles, although it was more underdispersive than the other ensembles.
- In comparisons of NCAR FV3 and NCAR MPAS ensembles at Day 3-5 lead times, the MPAS ensemble received significantly higher mean subjective ratings for severe weather forecasting applications.
- Various other projects and products were assessed and evaluated related to severe weather prediction, including mesoscale and storm-scale analyses, ML-based tools for producing watch guidance, AI global NWP emulators, and an ML product that combined WoFS and ProbSevere to produce severe weather guidance at 0-3 h lead times.
  - Two versions of 3D-RTMA with HRRR and RRFS backgrounds were evaluated. Generally, the two versions were similar with the 3D-RTMA RRFS having slightly larger errors over the domains in 2-m temperature and dewpoint. Common issues in the RRFS-based version included a moist bias and low CAPE bias.
  - A surface objective analysis based on the HRRR performed similarly to 3D-RTMA HRRR.
  - In subjective rankings of upscaled mesoanalyses (i.e., 40-km grid), the sfcOA HRRR and 3D-RTMA HRRR were ranked best, while the RAP-based SPC mesoanalysis was third, followed by the 3D-RTMA RRFS.
  - Fifteen-minute maximum forecasts of 80-m winds, 2-5 km AGL UH, and column maximum updraft speed from WoFS were used as a proxy for the analysis of severe weather. Overall, the WoFS ensemble analysis fields were positively viewed in terms of lining up with radar-derived proxies of severe weather, preliminary local storm reports, and a subjective assessment of severe weather based on the environment.
  - An HREF-based ML product was used to provide first-guess guidance on severe thunderstorm and tornado watches at 0-3 h lead times with comparisons made to SPC-issued watches as well as warnings and storm reports. Overall, participant comments were favorable but spatial coverage was often too large and the watch-type guidance was biased toward tornado watches.
  - Forecasts with 7-day lead time of large-scale fields were examined from three AI-generated global NWP emulators and the GFS. The AI models often demonstrated skill in predicting the synoptic-scale pattern, but overall participants favored the GFS. Concern was expressed that AI forecasts were too smooth and often failed to resolve more subtle shortwave troughs

as depicted by 500 mb heights and winds. For 6-h QPF, GraphCast was rated significantly worse than GFS because of QPF forecasts that were overly smooth, which led to many instances of missed heavy precipitation (note, GraphCast was the only AI model that produced QPF). Finally, examinations of derived vertical profiles found that temperature and dewpoint vertical structures were physically realistic though resolution was very coarse.

- A random-forest ML algorithm called WoFS-PHI was used to combine information from ProbSevere Version 2 and WoFS to produce spatial hazard probabilities at 0-3 h lead times. Two groups of SFE participants produced 0-1 and 1-2 h severe weather probabilities with and without the WoFS-PHI forecasts. Overall, participants with access to WoFS-PHI produced forecasts rated significantly better than participants that did not have access.

Overall, the 2024 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 2024 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 emulating NWP models as well as calibrating and post-processing CAM output to aid forecasters. Additionally, we expect that this work will continue to take on particular importance and assist with evidence-based decision making as NOAA moves forward with its plans for a Unified Forecasting System. SFE 2024 marked the second hybrid experiment (i.e., both in-person and virtual participation), a format we plan to continue since having in-person participation is much more conducive to science-based discussions and establishing new collaborations, while virtual participation enables people to participate that are unable to attend in-person, which expands the accessibility and scope of the SFE.



## Acknowledgements

The 2024 SFE would not have been possible without dedicated participants and the support and assistance of numerous individuals at SPC, NSSL, and CIWRO. In addition, collaborations with NCAR, GSL, CIRA, GFDL, NASA, and EMC were vital to the success of the 2024 SFE. In particular, Ryan Sobash (NCAR), Craig Schwartz (NCAR), Dave Ahijevych (NCAR), Amanda Back (GSL), Anders Jensen (GSL), Chunhua Zhou (CIRES/GSL), Craig Hartsough (CIRES/GSL), Curtis Alexander (GSL), David Dowell (GSL), Eric James (GSL), Georg Grell (GSL), Guoqing Ge (CIRES/GSL), Haidao Lin (GSL), Haiqin Li (CIRES/GSL), Hongli Wang (CIRES/GSL), Janet Derrico (CIRES/GSL), Jaymes Kenyon (CIRES/GSL), Jeff Beck (GSL), Jeff Hamilton (GSL), Joe Olson (GSL), Liaofan Lin (CIRA/GSL), Ligia Bernardet (GSL), Ming Hu (GSL), Molly Smith (CIRES/GSL), Ruifang Li (CIRES/GSL), Stan Benjamin (CIRES/GSL), Steve Weygandt (GSL), Tanya Smirnova (CIRES/GSL), Terra Ladwig (GSL), Trevor Alcott (GSL), Jacob Radford (CIRA/GSL), Ben Blake (Lynker/EMC), Donnie Lippi (Lynker/EMC), Matt Morris (SAIC/EMC), Shun Liu (EMC), Nick Esposito (SAIC/EMC), Eric Aligo (EMC), Marcel Caron (SAIC/EMC), Gang Zhao (SAIC/EMC), Jili Dong (SAIC/EMC), Ed Colon (Lynker/EMC), Matthew Pyle (EMC), Kai-Yuan Cheng (GFDL), Lucas Harris (GFDL), Matthew Morin (GFDL), Linjiong Zhou (GFDL), William Putman (NASA), and Scott Rabenhorst (NASA) were essential in generating and providing access to model forecasts or products examined on a daily basis.

## References

- Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, 2022: Pangu-Weather: A 3D High-resolution Model for Fast and Accurate Global Weather Forecast. <https://doi.org/10.48550/arXiv.2211.02556>.
- 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.
- Clark, A. J. and Coauthors, 2024: Spring Forecasting Experiment 2024: Program Overview and Operations Plan. Accessed 2 July 2024, [https://hwt.nssl.noaa.gov/sfe/2024/docs/HWT\\_SFE2024\\_operations\\_v1.pdf](https://hwt.nssl.noaa.gov/sfe/2024/docs/HWT_SFE2024_operations_v1.pdf).
- Clark, A. J., K. A. Hoogewind, A. Hill, and E. D. Loken, 2024: Extended range machine-learning severe weather guidance based on the operational GEFS. *Wea. Forecasting* (submitted).
- Earth Prediction Innovation Center, 2024: Unified Post-Processor (UPP), Accessed 2 July 2024, <https://epic.noaa.gov/unified-post-processor/>.
- Hill, A. J., R. S. Schumacher, and I. L. Jirak, 2023: A new paradigm for medium-range severe weather forecasts: Probabilistic random-forest-based predictions. *Wea. Forecasting*, **38**, 251-272.
- Karstens, C. D., R. Clark III, I. L. Jirak, P. T. Marsh, R. Schneider, and S. J. Weiss, 2019: Enhancements to Storm Prediction Center convective outlooks. Ninth Conf. on Transition of Research to Operations, Phoenix, AZ, Amer. Meteor. Soc., J7.3, <https://ams.confex.com/ams/2019Annual/webprogram/Paper355037.html>.
- Lam, R., and Coauthors, 2022: GraphCast: Learning skillful medium-range global weather forecasting. <https://doi.org/10.48550/arXiv.2212.12794>.
- Pathak, J., and Coauthors, 2022: FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. <https://doi.org/10.48550/arXiv.2202.11214>.
- 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. Meteor. Soc.*, **99**, 2025-2043.
- Smith, T. M., V. Lakshmanan, G. J. Stumpf, K. L. Ortega, K. Hondl, K. Cooper, K. M. Calhoun, D. M. Kingfield, K. L. Manross, R. Toomey, and J. Brogden, 2016: Multi-Radar Multi-Sensor (MRMS) Severe Weather and Aviation Products: Initial Operating Capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 1617-1630.
- Stensrud, D. J., and Co-authors, 2009: Convective-scale warn-on-forecast system. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1499.

## APPENDIX

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

Table 5. Schedule for Tuesday – Friday. On Mondays, the schedule is similar except the period 9-11am is devoted to training and introductory material.

Time (CDT)	
12:00 PM – 2:00 PM	<b>WoFS-PHI Introduction and Training (Mondays only)</b> <i>Evening Forecasters</i>
2:00 PM – 2:15 PM	<b>Update on Today’s Weather</b> <i>Hybrid All &amp; Evening Forecasters</i> (David Imy)
2:15 PM – 3:15 PM	<b>Individual Forecasting Activity</b> (WoFS PHI) <i>In-Person Innovation</i> (WoFS PHI); <i>Virtual &amp; Evening Forecasters</i> (WoFS PHI)
3:15 PM – 4:00 PM	<b>Individual Forecasting Activity</b> (WoFS PHI) <i>In-Person Innovation</i> (WoFS PHI); <i>Virtual &amp; Evening Forecasters</i> (WoFS PHI)
4:00 PM – 8:00 PM	<b>Individual Forecasting Activity</b> <i>Evening Forecasters</i> (WoFS PHI)

Table 6. Schedule for Monday – Thursday evening activity.