209 SUBJECTIVE EVALUATION OF OPERATIONAL AND EXPERIMENTAL FV3-BASED CAM CONFIGURATIONS FOR SEVERE WEATHER FORECASTING

Burkely T. Gallo[1,2*], Makenzie Krocak[1,2,3], Brett J. Roberts [1,2,4], Kent H. Knopfmeier[1,4], Israel L. Jirak[2], Adam J. Clark[4,5], Yunheng Wang[1,4], Jacob Carley[6], Curtis Alexander[7], Lucas Harris[8], Kai-Yuan Cheng[8,9], and Shun Liu[6]

[1]Cooperative Institute for Severe and High-Impact Weather Research and Operations, Norman, OK
[2]NOAA/NWS/NCEP Storm Prediction Center, Norman, OK
[3]Institute for Public Policy and Research Analysis, Norman, OK
[4]NOAA/OAR National Severe Storms Laboratory, Norman, OK
[5]School of Meteorology, Norman, OK
[6]NOAA/NWS/NCEP Environmental Modeling Center, College Park, MD
[7]NOAA/OAR Global Systems Laboratory, Boulder, CO
[8]NOAA/OAR Geophysical Fluid Dynamics Laboratory, Princeton, NJ
[9]Cooperative Institute for Modeling the Earth System, Princeton, NJ

1. INTRODUCTION

As the United States transitions to a Unified Forecasting System (UFS), evaluating the resultant forecasts is crucial to ensuring that the new iterations of experimental forecasts improve upon the operational models. One venue for evaluating the forecasts specifically for severe convective weather is NOAA's Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (SFE; Gallo et al. 2017, Clark et al. 2022a). During the 5-week annual SFE, participants conduct multiple forecasting and evaluation activities during the peak of the spring convective season. Evaluation activities center on new convection-allowing model (CAM) guidance, post-processing methods, analysis techniques, and calibrated guidance. During the 2022 SFE, owing to the many experimental CAMs, ensembles, and products provided, four different evaluation groups focused on calibrated guidance, deterministic CAMs, CAM ensembles, and a medley of other guidance.

Each evaluation group conducted next-day subjective evaluations for their assigned product suite. Subjective evaluations have long been used in SFEs (Kain et al. 2003; Clark et al. 2012; Gallo et al. 2016; Miller et al. 2021) to assess the performance of experimental guidance and the impact of different configuration strategies. Subjective evaluations provide useful feedback for aspects of model guidance that may be challenging to illustrate with bulk statistics, such as storm characteristics like convective mode, convective storm size, or storm evolution. Subjective evaluations also help to provide feedback on the guidance as it would be used by forecasters; while bulk statistical metrics may tell forecasters how the model performs in aggregate over many cases, it doesn't always illustrate what those biases will look like in terms of the sensible weather or how a model is used in practice.

This work examines two of the deterministic model comparisons undertaken in SFE 2022. The first comparison, involving state-of-the-art model guidance from many different agencies, is typically performed in some form during the SFE each year. Please see Clark et al. (2022b) for full details on model configurations. The second evaluation dives more deeply into the operational CAM system (the High-Resolution Rapid Refresh forecast system or HRRRv4; Dowell et al. 2022, James et al. 2022) and its potential candidate for replacement (the Rapid Refresh Forecast System Prototype 2 Control member or RRFSp2 Control). The techniques used to evaluate these models subjectively will be described below, followed by the results of the subjective evaluation. The next section will detail objective evaluation done after

* *Corresponding author address:* Burkely T. Gallo, NOAA/NWS/NCEP/Storm Prediction Center, 120 David L. Boren Blvd., Norman, OK 73072; e-mail: burkely.twiest@noaa.gov

the conclusion of the 2022 SFE to supplement the subjective verification results, and the final section will discuss conclusions and recommendations for the 2023 SFE subjective evaluation strategies.

2. METHODS

2.1 Subjective Evaluation: Deterministic Flagships

For this evaluation, participants considered a 6-panel figure with five cutting-edge model configurations contributed by different agencies. Those models were the HRRRv4, the RRFSp1, the RRFSp2 Control, the NSSL-FV3, and the GFDL-FV3. These model configurations differed in dynamical core, parameterization schemes, data assimilation strategies, and whether the models were a global-nested or stand-alone regional configuration. Essentially, the purpose of this evaluation was to determine if any individual configuration of the experimental models could approach or exceed the skill of the currently operational model (the HRRRv4).

Participants evaluated storm-attribute fields and environmental fields. Participants were asked to consider forecast hours 13–36 in their evaluations, corresponding with 1200–1200 UTC. Storm attribute fields included 2–5 km updraft helicity (UH) and composite reflectivity (Fig. 1), which was verified with observed reflectivity from MRMS and overlaid preliminary local storm reports (LSRs). Preliminary LSRs were used due to their low latency for a next-day evaluation activity. Each participant was also randomly assigned to evaluate one of the following environmental fields: 2-m temperature, 2-m dewpoint, or surface-based convective available potential energy (SBCAPE), all of which were verified using 3D-RTMA data (i.e., the HRRR-based version).

For each of the above fields, participants were asked to rank the models from the best-performing (1) to the worst-performing (5) model. They then could provide a 1-10 rating of only the best-performing model, to allow facilitators to determine whether a model was "the best of the worst", or truly performing well in a case. While the evaluations were taking place, the models were *blinded*, meaning that participants did not know which model was which. Panels were also shuffled day-to-day, so that participants who were in the same group multiple days in a row would not know which model

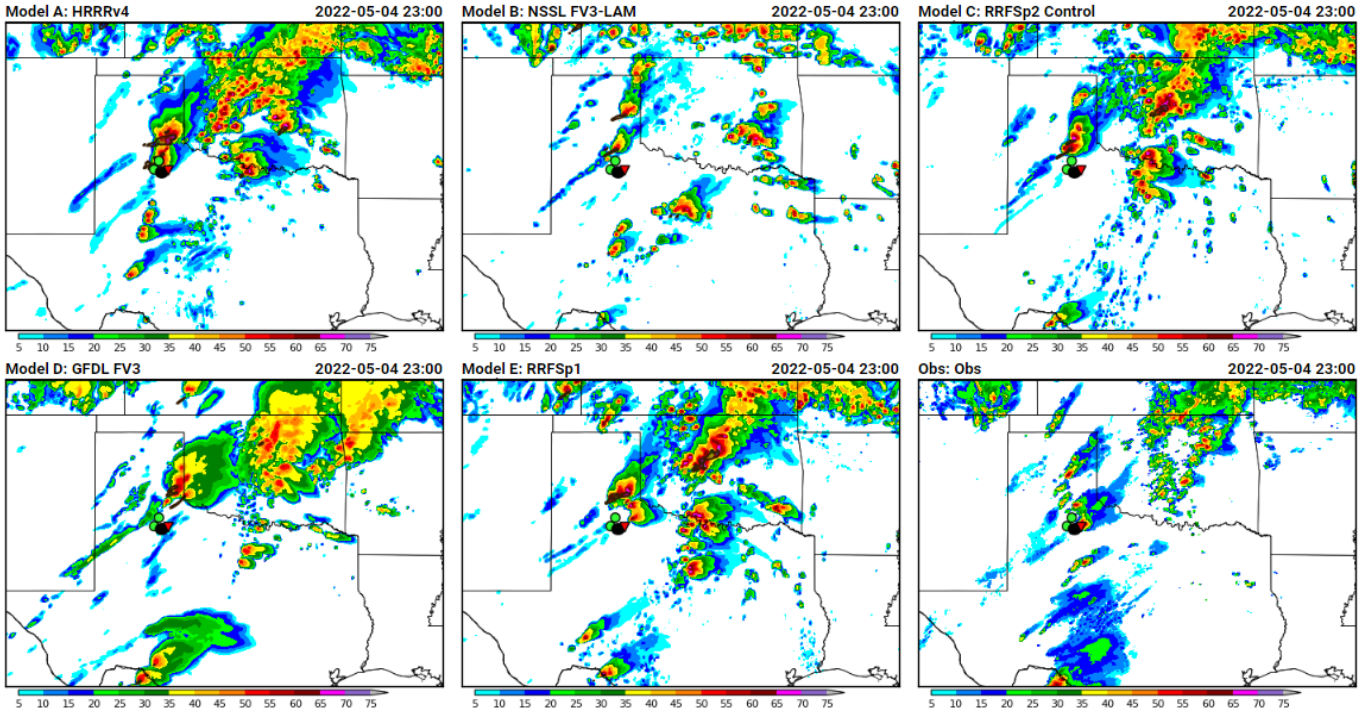

Figure 1. An example of what participants saw during the Deterministic Flagship comparison. Simulated reflectivity and 2–5 km updraft helicity exceeding the 95th percentile of model climatology from five contributed models are in the top row and the left two panels in the bottom row, while observed reflectivity is in the lower right panel. Local storm reports from the past hour are overlaid as green dots (hail), black dots (significant hail) and red inverted triangles (tornado).

was which depending on its position in the 6-panel figure (Fig. 1). After the surveys were submitted, the models were unblinded during a discussion activity, so that participants, including operational participants who may be using these models in a few years to issue forecasts, were able to examine specific model performance. Alongside their rankings, participants were also asked to provide what aspects of convection they were examining in determining their rankings in an open-ended text box.

The blinding and shuffling of panels was received well by participants, and these efforts will likely be repeated in future SFEs. However, shifting from a rating system wherein participants assign a numerical rating from 1–10 to a ranking system will likely not persist into future SFEs. While ranking is a useful practice given the difference between participant interpretation of numerical value (e.g., a "7" may not mean the same thing to all participants), ranking is optimally useful only when all available data is present. Given the experimental nature of the models contributed to the annual SFEs, data gaps throughout the experiment are nearly inevitable. For example, during SFE 2022 we had 12 cases out of 19 that met the criterion of having all models available, with 92 responses from participants, restricting our cases examined herein to a subset of the 5-week experiment.

2.2 Subjective Evaluation: HRRRv4 vs. RRFS

The second evaluation of deterministic CAMs focused more specifically on the HRRRv4 vs. the RRFSp2 Control, which was configured to resemble the HRRRv4 as closely as possible at this stage of development. Participants were first asked to provide input to at least two of the following five storm attribute fields: Composite reflectivity and 2–5 km UH, updraft speed, 10-m wind speed, 10-m wind gusts, and 0–3 km UH. For these fields, participants were asked which model performed better, with an option to select "models performed about the same". Participants were asked the same question about one of either 2-m temperature, 2-m dewpoint, or SBCAPE. Finally, participants were asked to comment on model differences between two out of five additional environmental fields (850 mb heights/winds, 700 mb heights/winds, 500 mb heights/winds, MLCAPE, and MUCAPE), which were randomly assigned. These additional environmental fields did not have verification data available, so participants were asked only to comment on differences between the fields.

Unlike the prior comparison, this evaluation was unblinded, so participants were able to see which model was the HRRRv4 and which was the RRFSp2 Control member while they were filling out their survey.

2.3 Objective Evaluation: HRRRv4 vs. RRFS

To supplement the prior subjective evaluation, objective verification was performed on the HRRRv4 and the RRFSp2 Control member after the 2022 SFE concluded. This evaluation encompassed 20 cases in which data was available for both models. For this analysis, surrogate severe fields were created by regridding the model data and reports to the NCEP 211 grid (80 km). Surrogate severe fields were created by running a Gaussian smoother with varying sigma over fields where UH exceeded a specific percentile threshold. Fields were created using 100 UH percentile thresholds ranging from the 70th to the 99.7th percentiles and 53 Gaussian smoothers with sigma ranging from 40 km to 300 km. These fields were verified using binary regridded report data. Metrics examined were the area under the Receiver Operating Curve (ROC area; Mason 1982) and the Fractions Skill Score (FSS; Roberts and Lean 2008).

3. RESULTS

3.1 Subjective Evaluation: Deterministic Flagships

Rankings for the reflectivity and UH show two groupings of model performance (Fig. 2). The HRRRv4, RRFSp1, and RRFSp2 Control were ranked relatively similarly with regards to the mean ranking, followed by the NSSL FV3-LAM, and then the GFDL FV3. The HRRRv4 was most frequently ranked first, followed by the RRFSp1 and the RRFSp2 Control. The RRFS models were most frequently rated second or third, leading the RRFSp1 to a slightly higher overall mean ranking than the HRRRv4, though these differences are likely not significant. The NSSL FV3-LAM was most frequently rated fourth or fifth, and the GFDL FV3 was most frequently rated last. When asked what characteristics of the simulated reflectivity and UH forecasts were most important to the participants when ranking the models, participants broadly cited forecasting challenges such as the convective

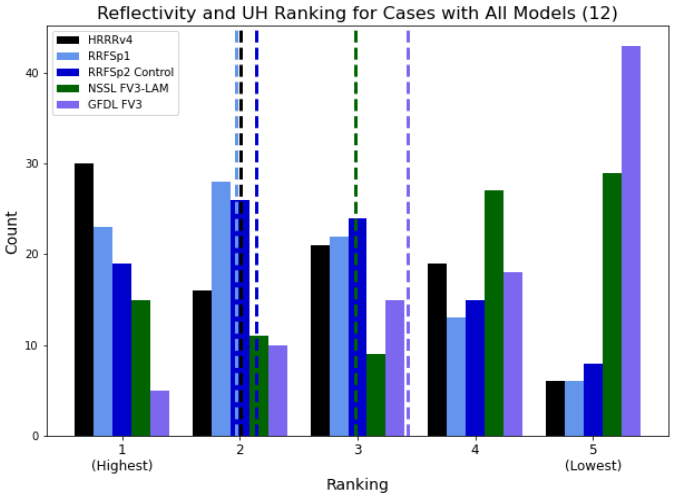

Figure 2. Reflectivity and UH rankings for models in the Deterministic Flagship comparison. Dashed lines indicate the mean ranking (lower numbers are better).

initiation, progression of storms, location of storms, intensity of storms. In a word cloud of participant responses, timing, location, and storm mode showed up frequently. Convective coverage also came up in some participant responses.

Rankings for the environmental fields followed similar patterns to the reflectivity and UH rankings, although the HRRRv4 easily received the highest mean ranking in temperature and SBCAPE (Fig. 3a,c). The HRRRv4 was most frequently rated the highest of all of the models considered in those fields, while the RRFSp2 Control was most frequently ranked first for dewpoint (Fig. 3b). Overall, the pattern of the HRRRv4, RRFSp1, and RRFSp2 Control ranking the best continued for all environmental fields considered, followed by the GFDL FV3 and the NSSL FV3-LAM. The GFDL FV3 placed fourth in terms of highest ranking for temperature, but was most frequently rated last for dewpoint and SBCAPE. For temperature, the NSSL FV3-LAM was most frequently ranked last. When evaluating the 2-m temperature, participants looked more closely at boundaries, gradients, and mesoscale areas of bias in making their rankings. Cold pools were also considered. Similar considerations applied for the 2-m dewpoint and the SBCAPE, although the shape and orientation of boundaries were specifically cited with regards to any drylines that may have been in the SFE domain of interest. Horizontal distribution of large areas of SBCAPE (e.g., warm sectors) also played a role for some participants assigned the SBCAPE field.

Ratings results (not shown) confirm that the best performing model was frequently rated similarly

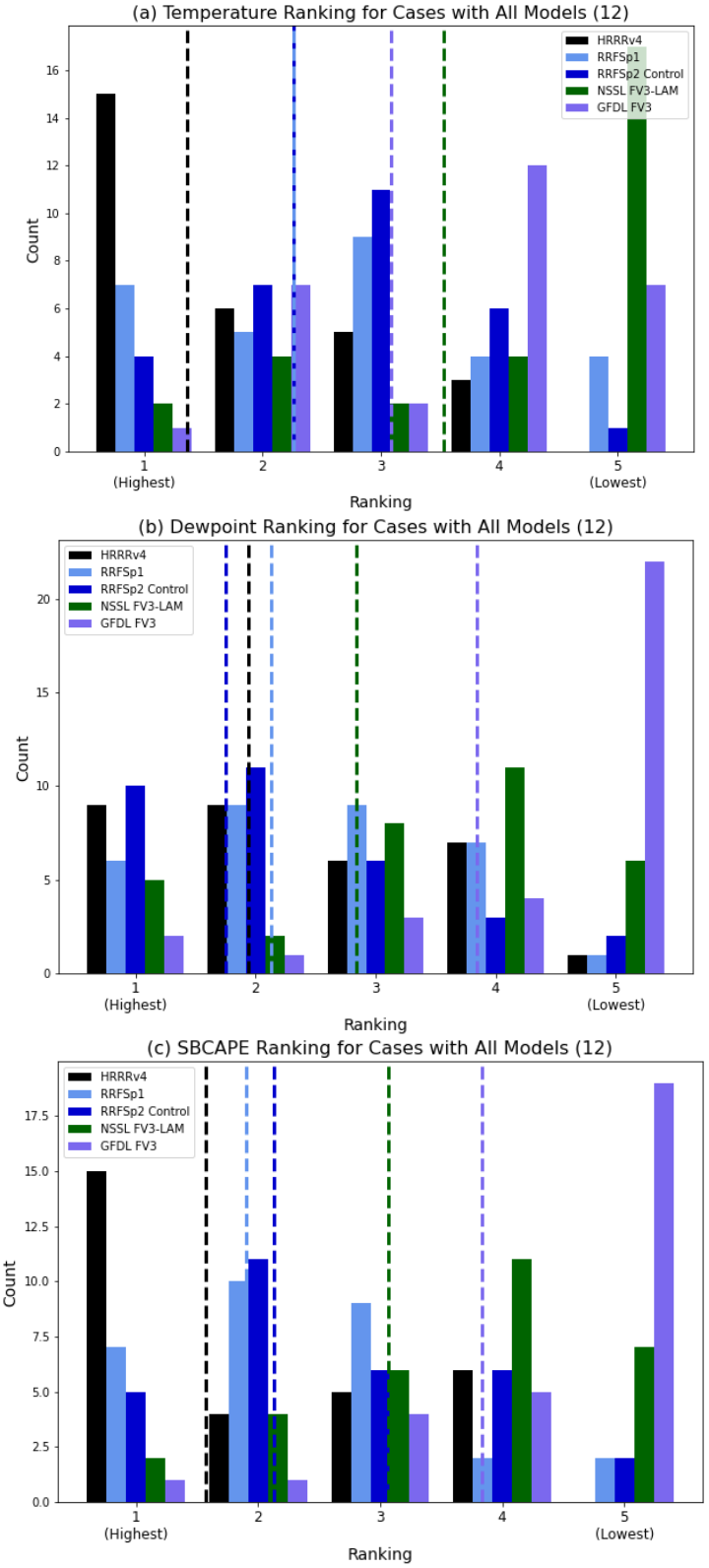

Figure 3. Rankings of environment for the Deterministic Flagship models. Rankings were completed for (a) 2-m Temperature, (b), 2-m Dewpoint, and (c) SBCAPE. Dashed lines indicate the mean ranking for the model in question (lower numbers ares better), and the dashed blue lines in (a) indicate that the RRFSp1 and the RRFSp2 Control had the same mean ranking. Note that the y-axis on these comparisons are scaled to each individual subplot.

between cases and participants for both the storm-attribute fields and the environmental fields, with median ratings around 7 or 8 out of 10 in most cases. Since the environmental fields have quite small sample sizes, strong conclusions cannot be drawn from them. However, looking at the distributions of the HRRRv4, RRFSp1, and RRFSp2 Control members in the reflectivity and UH ratings show very similar distributions, indicating very similar performance across models that on days where this set of models are performing their best.

2.2 Subjective Evaluation: HRRRv4 vs. RRFS

Participants most frequently selected the reflectivity/2–5 km UH and 10-m wind speed to evaluate, although the updraft speed was a close third (Fig. 4). The 10-m wind gusts, although only evaluated fourth most often, were frequently a topic of discussion after completion of the survey. Storm-attribute field performance varied regarding which model was selected as the best performer (Fig. 5). For simulated reflectivity and updraft speed, the HRRRv4 was selected as the better-performing model more frequently than the RRFSp2 Control.

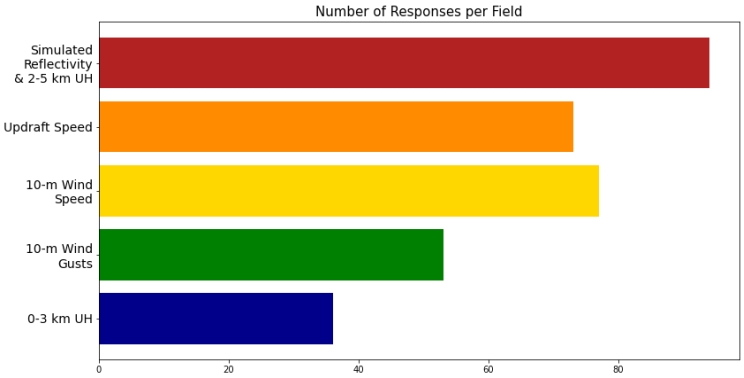

Figure 4. Number of times each storm attribute field was selected for evaluation. Note: 0-3 km UH was unavailable for the first few weeks of SFE 2022.

However, the 10-m wind speeds and the 0–3 km UH were frequently better in the RRFSp2 Control relative to the HRRRv4. The 10-m wind gust performance was similar across all categories. When commenting on the 2–5 km UH and simulated reflectivity, participants frequently noted different performance at different time periods, as exemplified by comments such as: “HRRR did better first half of the period by far, but RRFS did better with the bigger event, derecho later in

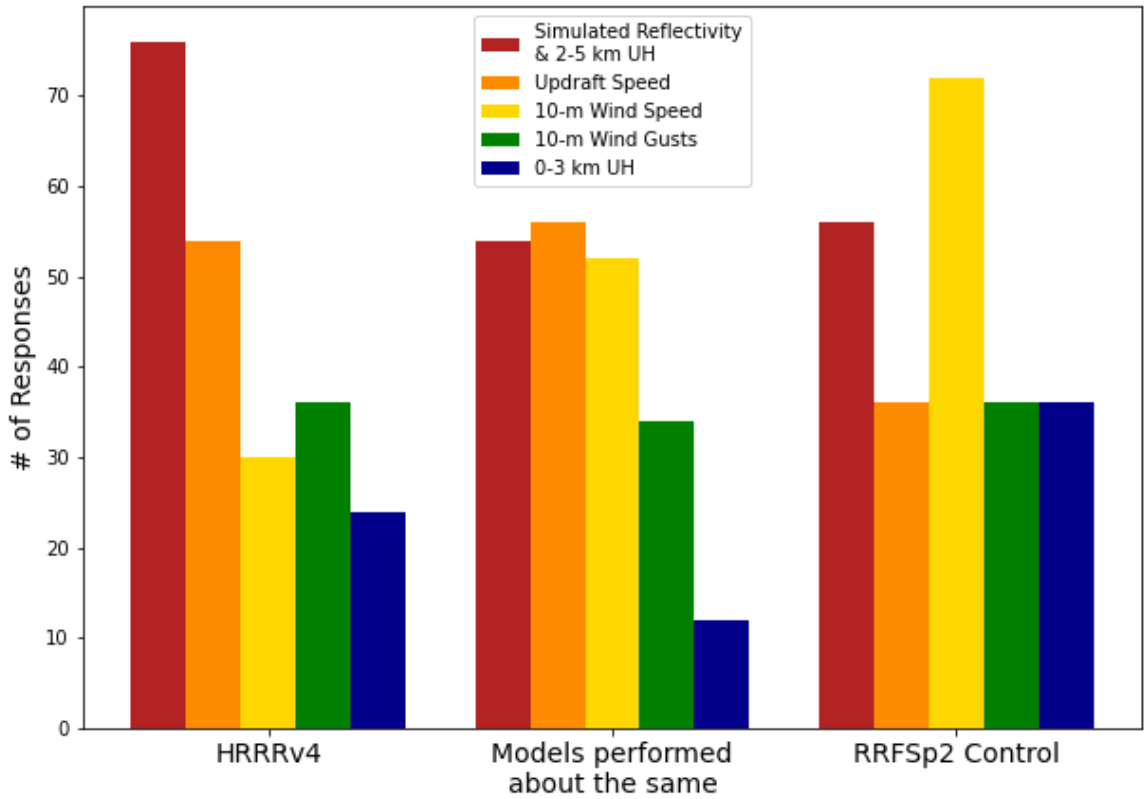

Figure 5. Answers to the question, “Which model performed best for this field?”, in which participants were asked to select at least two of the five fields presented to evaluate.

period". Comments such as these highlight the necessity of objective verification across the entire convective day, which is time-prohibitive to do subjectively in the context of the SFE. Participants frequently commented that the 10-m wind speed was too low, particularly in the HRRRv4. The 10-m wind gust product, which is not constrained by having to meet a reflectivity criteria, was noted by the participants to show swaths of strong wind gusts in the RRFSp2 Control that appeared to be synoptically driven rather than associated with convection. However, during one discussion session, a WFO forecaster mentioned that it wouldn't necessarily be bad for the model to show high synoptic gusts, as they currently faced a forecast challenge in getting good guidance for gusty winds that were not associated with convection.

Participants next evaluated a randomly selected environmental field. Fields were evenly assigned between 2-m temperature, 2-m dewpoint, and SBCAPE. For temperature and CAPE, the most frequent response from participants was that the HRRRv4 and the RRFSp2 Control performed about the same (Fig. 6). For dewpoint, however, the RRFSp2 Control being better was the most frequent response. Overall, the HRRRv4 appears to perform better with regards to the 2-m temperature and the SBCAPE, but the RRFSp2 Control forecasts the 2-m dewpoints better than the HRRRv4. Participant comments surrounding the temperature spoke to the placement and intensity of boundaries and cold pools, and some participants focused in on the reasoning why specific biases may be preferred: "*The RRFSp2 appeared slightly closer to true values but given it was cool biased compared to HRRRv4 warm bias, I preferred the warmer solution given the impacts of the day may be made more significant with a warmer boundary layer.*" Participants frequently commented on a dry bias in the HRRRv4's 2-m dewpoints, and the comments surrounding the CAPE showed no clear trends.

Finally, participants were asked to evaluate fields that were new to formal subjective evaluation, and were asked to comment on differences between three of the following fields that were randomly assigned: 500 mb height/wind, 700 mb height/wind, 850 mb height/wind, MUCAPE, and MLCAPE. Comments on the 500 mb fields were frequently that the models were similar, but for some cases

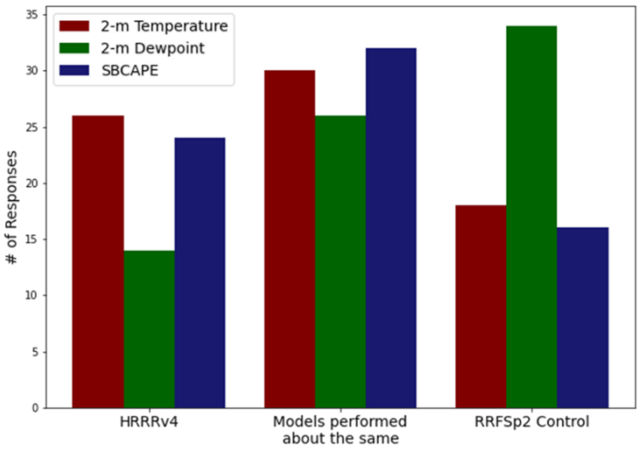

Figure 6. As in Fig. 5, but with environmental fields that were randomly assigned.

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

2.3 Objective Evaluation: HRRRv4 vs. RRFS

When looking at the HRRRv4 and RRFSp2 Control members objectively via the surrogate severe fields (Fig. 7), maximum scores among all smoothing and percentile thresholds are similar. For the ROC area, the maximum score achieved by the HRRRv4 is only 0.012 higher than the maximum score of the RRFSp2 Control. Similarly, for the FSS the difference is only 0.0193 between the two models. However, it should be noted that the HRRRv4 also

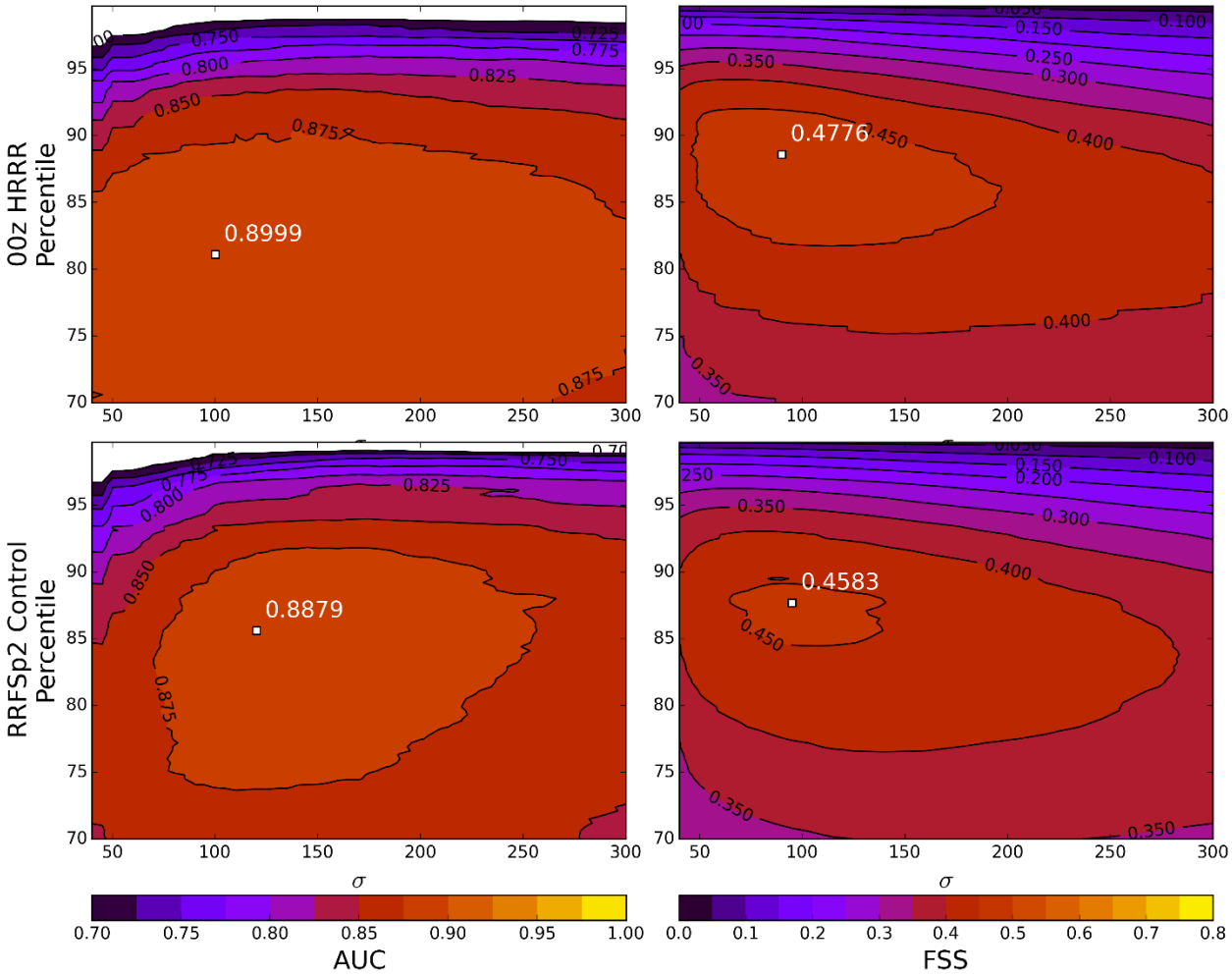

Figure 7. Objective verification of surrogate severe fields generated using the HRRRv4 and the RRFSp2 Control. Maximum scores are indicated with white squares, and the maximum score is annotated in white.

has a broader area of high performance relative to the RRFSp2 Control, showing that it can provide skillful forecasts for a range of surrogate severe fields and may not be as sensitive to the specifications of the surrogate severe formulation. The FSS of the RRFSp2 Control is maximized at a slightly lower percentile threshold than the HRRRv4, but the ROC area is maximized at a higher percentile threshold. Smoothing levels maximizing skill are similar for the FSS, but the HRRRv4 optimizes ROC area at a smaller smoothing radius than the RRFSp2 Control.

The main takeaway here is that the objective skill of the experimental RRFSp2 Control member is approaching the skill of the HRRRv4, similar to what was seen in the subjective evaluations. Since SFEs in prior years that tested FV3-based CAM configurations have not shown those configurations to possess a similar level of skill, these results should be encouraging to the UFS community and the model developers who have spent many years working to improve these forecasts.

4. CONCLUSIONS AND KEY TAKEAWAYS

During the 2022 SFE, subjective verification took place for several deterministic and ensemble comparisons. For a description of all preliminary results for SFE 2022, please see Clark et al. (2022a). This work examined two subjective evaluations focused on deterministic guidance, specifically the cutting-edge deterministic guidance contributed by many different agencies to the SFE that are candidates to replace the operational deterministic CAM configuration. The primary takeaways are as follows:

(1) For the first time, the subjective and objective skill of the RRFS prototypes is

approaching the HRRRv4 baseline for severe convective storms forecasting.

(2) Remaining areas for targeted improvement include mitigating a low SBCAPE bias in the RRFS.

(3) Data assimilation in the HRRRv4, RRFSp1, and RRFSp2 Control improve forecasts drastically relative to the cold-start NSSL FV3-LAM and the GFDL-FV3.

(4) Potentially impactful differences between the HRRRv4 and the RRFSp2 Control were seen in fields like surface wind gusts and upper-air winds, motivating their objective verification prior to implementation or retirement.

For SFE 2023, we anticipate maintaining blinded model evaluations, although it is likely that we will return to rating models rather than ranking them. For the ratings, however, we will likely iterate from a 1–10 rating system to a Likert scale to provide more distinction between rating categories. We look forward to the next iteration of experimental CAMs to be evaluated in the SFE, and continued research-to-operations, operations-to-research efforts surrounding severe convective storms.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the efforts made by the development teams in generating the experimental dataflows examined herein during SFE 2022. We would also like to thank all of the participants that took time to provide input during SFE 2022 via the survey and discussion exercises. This extended abstract was prepared by Burkely T. Gallo with funding provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement #NA21OAR4320204, U.S. Department of Commerce. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of NOAA or the U.S. Department of Commerce.

REFERENCES

Clark, A. J., S.J. Weiss, J.S. Kain, I.L. Jirak, M. Coniglio, C.J. Melick, C. Siewert, R. A. Sobash, P.T. Marsh, A.R. Dean, M. Xue, F.Y. Kong, K.W. Thomas, Y.H. Wang, K. Brewster, J.D. Gao, X.G. Wang, J. Du, D.R. Novak, F.E. Barthold, M.J. Bodner, J.J. Levit, C.B. Entwistle, T.L. Jensen, and J. Correia, 2012: An overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Amer. Meteor. Soc*., **93**, 55-74.

Clark, A. J., and Coauthors 2022a: Spring Forecasting Experiment 2022 Preliminary Findings and Results. Available online at: https://hwt.nssl.noaa.gov/sfe/2022/docs/HWT_SFE_2022_Prelim_Findings_FINAL.pdf

Clark, A. J., and Coauthors 2022b: Spring Forecasting Experiment 2022 Program Overview and Operations Plan. Available online at: https://hwt.nssl.noaa.gov/sfe/2022/docs/HWT_SFE2022_operations_plan.pdf

Dowell, D. C., and Coauthors, 2022: The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part I: Motivation and system description. *Wea. Forecasting,* **37**, 1371–1395, https://doi.org/10.1175/WAF-D-21-0151.1.

James, E. P., and Coauthors, 2022: The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part II: Forecast performance. *Wea. Forecasting*, **37**, 1397–1417, https://doi.org/10.1175/WAF-D-21-0130.1.

Gallo, B. T., A. J. Clark, and S. R. Dembek, 2016: Forecasting Tornadoes Using Convection-Permitting Ensembles. *Wea. Forecasting*, **31**, 273–295. doi: http://dx.doi.org/10.1175/WAF-D-15-0134.1.

Gallo, B.T., and Coauthors, 2017: Breaking New Ground in Severe Weather Prediction: The 2015 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Wea. Forecasting*, **32**, 1541–1568, https://doi.org/10.1175/WAF-D-16-0178.1

Kain, J. S., P. R. Janish, S. J. Weiss, M. E. Baldwin, R. S. Schneider, and H. E. Brooks, 2003: Collaboration between forecasters and research scientists at the NSSL and SPC: The spring program. *Bull. Amer. Meteor. Soc.*, **84**, 1797-1806.

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

Miller, W. J. S., C. K. Potvin, M. L. Flora, B. T. Gallo, L. J. Wicker, T. A. Jones, P. S. Skinner, B. Matilla, and K. H. Knopfmeier, 2021: Exploring the Usefulness of Downscaling Free Forecasts from the Warn-on-Forecast System. *Wea. Forecasting*, **37**, 181–203. https://doi.org/10.1175/WAF-D-21-0079.1.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1