SPRING FORECASTING EXPERIMENT 2020

Conducted by the

EXPERIMENTAL FORECAST PROGRAM

of the

NOAA HAZARDOUS WEATHER TESTBED

https://hwt.nssl.noaa.gov/sfe/2020

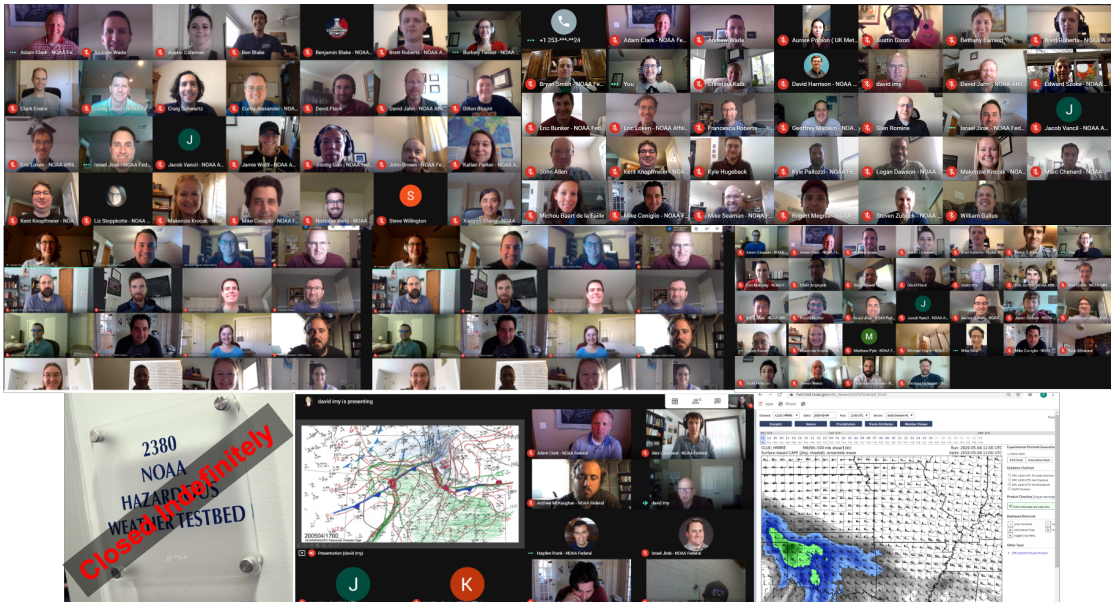
Virtual Experiment
27 April - 29 May 2020

Preliminary Findings and Results

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

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

Table of Contents



Scenes and participant screenshots from each week of the 2020 NOAA Hazardous Weather Testbed Spring Forecasting Experiment

Foreword

Because of the COVID-19 pandemic, the 2020 Spring Forecasting Experiment was truly unique, with challenges and uncertainties that were daunting and unprecedented. In terms of planning, the timing of the pandemic left very little time to adjust. As in previous years, by mid- to late-March we were close to finalizing SFE plans, but when it became clear that gathering and travel restrictions would preclude an in-person experiment, we were forced to regroup and switch gears within about a one-month time frame so that we could conduct a virtual experiment. This change in plans required creativity, innovation, and coordination like we had never undertaken before, and allowed us to continue progress in key areas of research geared toward accelerating research-to-operations for tools and concepts that improve operational severe weather forecasts and further our mission to protect life and property. Our collaborators and team members sacrificed and went well above and beyond what was expected of them to make the virtual experiment a success, all while dealing with the personal challenges and struggles associated with the pandemic. While these were challenging times, we recognize that we are fortunate to work in a field in which many of us could work from home. Thus, we also acknowledge the dedication of essential workers, those working on the frontlines to combat COVID-19, and those that have suffered personal loss from COVID-19. Ultimately, we are all proud to be a part of the SFE team and we thank NSSL and SPC management for supporting a virtual experiment and all our collaborators and participants for making the experiment a success.

1. Introduction

The 2020 Spring Forecasting Experiment (2020 SFE) was conducted from 27 April – 29 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 from collaborators including the NOAA Global Systems Laboratory (GSL), NOAA Geophysical Fluid Dynamics Laboratory (GFDL), United Kingdom Meteorological Office (Met Office), National Center for Atmospheric Research (NCAR), and NOAA/NCEP's Environmental Modeling Center (EMC). Participants included about 100 forecasters, researchers, model developers, university faculty, and graduate students from around the world (see Table 1 in the Appendix). Because of the COVID-19 pandemic, restrictions on travel and gatherings precluded an in-person experiment in the HWT facility. However, to maintain momentum in key areas of convection-allowing model development, the EFP conducted the 2020 experiment virtually. As in previous years, the 2020 SFE aimed to test emerging concepts and technologies designed to improve the prediction of hazardous convective weather, consistent with the Forecasting a Continuum of Environmental Threats (FACETs; Rothfusz et al. 2018) and Warn-on Forecast (WoF; Stensrud et al. 2009) visions:

Operational Product and Service Improvements:

- Explore the ability to generate higher temporal resolution Day 1 severe weather outlooks than those issued operationally by SPC by issuing 1- and 4-h time window outlooks for individual severe hazards (tornado, hail, and wind) using a prototype Warn-on-Forecast system (WoFS).
- Test the utility of WoFS for updating full period hazard forecasts valid 2100-1200 UTC.
- Explore the ability to provide enhanced information on the conditional intensity of tornado, wind, and hail events by delineating areas expected to follow "normal", "hatched", or "double-hatched" intensity distributions.

Applied Science Activities:

- Compare various CAM ensemble prediction systems to identify strengths and weaknesses of different configuration strategies. Most of these comparisons were conducted within the framework of the Community Leveraged Unified Ensemble discussed below. Additional baseline comparisons were made using the High-Resolution Ensemble Forecast System Version 3.0 (HREFv3), which is scheduled to become operational at the end 2020 or early 2021.
- Compare and assess different machine-learning approaches for evaluating the likelihood of wind damage reports being associated with gusts ≥ 50 knots.
- Compare and assess two machine-learning techniques for producing probabilistic hazard guidance from a deterministic 3-km grid-spacing CAM.
- Evaluate the utility of various multi-model, single-model, and time-lagged ensemble configuration strategies using HREFv3 as a baseline.
- Using the High-Resolution Rapid Refresh Ensemble (HRRRE), evaluate whether ensemble sensitivity-based subset probabilities provide improved guidance relative to probabilities produced from all 1800 and 0000 UTC HRRRE members.
- Compare and assess different verification approaches in CAM ensembles for predicting hail size.

- Evaluate configurations of the Stand-Alone-Regional Finite Volume Cubed Sphere Model (FV3-SAR) with different data assimilation, physics schemes, numbers of vertical levels, and horizontal advection settings.
- Compare and assess two different versions of the 3D real-time mesoscale analysis (3D-RTMA) system that use different sources of background error covariances.
- Focusing on forecast hours 0-12 over regions with sparse radar coverage, evaluate model configurations with and without assimilation of total lightning data from the GOES 16 Geostationary Lightning Mapper (GLM).
- Evaluate the utility of several methods for producing calibrated hazard guidance from HREFv2.1, which was the current operational version of HREF used at SPC during SFE 2020.
- Compare and assess the skill and utility of the primary deterministic CAMs provided by each SFE 2020 collaborator.
- Using deterministic CAMs provided by NCAR, NSSL, and the UK Met Office, examine forecast sensitivity to different initial conditions and model cores at convective scales.
- Evaluate WoFS for applications to short-term severe weather outlook generation, and explore the potential value provided by experimental, enhanced resolution (1.5 km grid-spacing) deterministic and ensemble WoFS configurations.

A suite of state-of-the-art experimental CAM guidance contributed by our large group of collaborators was critical to the 2020 SFE. For the fifth consecutive year, these contributions were formally coordinated into a single ensemble framework called the Community Leveraged Unified Ensemble (CLUE; Clark et al. 2018). The 2020 CLUE was constructed by having all groups coordinate as closely as possible on model specifications (e.g., grid-spacing, vertical levels, physics, etc.), domain, and post-processing so that the simulations contributed by each group could be used in controlled experiments. This design allowed us to conduct several experiments to aid in identifying optimal configuration strategies for CAM-based ensembles. The 2020 CLUE included 41 members using 3-km grid-spacing (except for the 2.2 km grid-spacing UK Met Office members) that allowed several unique experiments. The 2020 SFE activities also involved testing the WoFS for the fourth consecutive year.

This document summarizes the activities, core interests, and preliminary findings of the 2020 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. 2020; https://hwt.nssl.noaa.gov/sfe/2020/docs/HWT_SFE2020_operations_plan.pdf). The remainder of this document is organized as follows: Section 2 provides an overview of the models and ensembles examined during the 2020 SFE along with a description of the daily activities, Section 3 reviews the preliminary findings of the 2020 SFE, and Section 4 contains a summary of these findings and some directions for future work.

2. Description

a) Experimental Models and Ensembles

A total of 80 unique CAMs were run for the 2020 SFE, of which 41 were a part of the CLUE system. Other CAMs outside of the CLUE were contributed by NSSL (WoFS) and EMC (HREFv2.1 and HREFv3). Forecasting activities during the 2020 SFE emphasized the use of WoFS in generating experimental

probabilistic forecasts of individual severe weather hazards. Additionally, the 2020 CLUE configuration enabled numerous scientific evaluations focusing on model sensitivities and various ensemble configuration strategies.

To put the volume of CAMs run for 2020 SFE into context, Figure 1 shows the number of CAMs run for SFEs since 2007. The noticeable drop in 2020 relative to previous years is because three regular SFE collaborators, NCAR, the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma, and the Multi-Scale data Assimilation and Predictability Laboratory at the University of Oklahoma (OU-MAP), did not contribute ensembles for 2020. CAPS and OU-MAP are transitioning to new NOAA-funded projects that don't involve HWT experiments until 2021, while NCAR is transitioning to a new project focused on machine-learning applications for post-processing CAM output rather than optimizing CAM ensemble configurations. Aside from the abnormally low number of models in 2020, Figure 1 shows an increasing trend. The consolidation of members into the CLUE has made this increase more manageable and facilitated more controlled scientific comparisons.

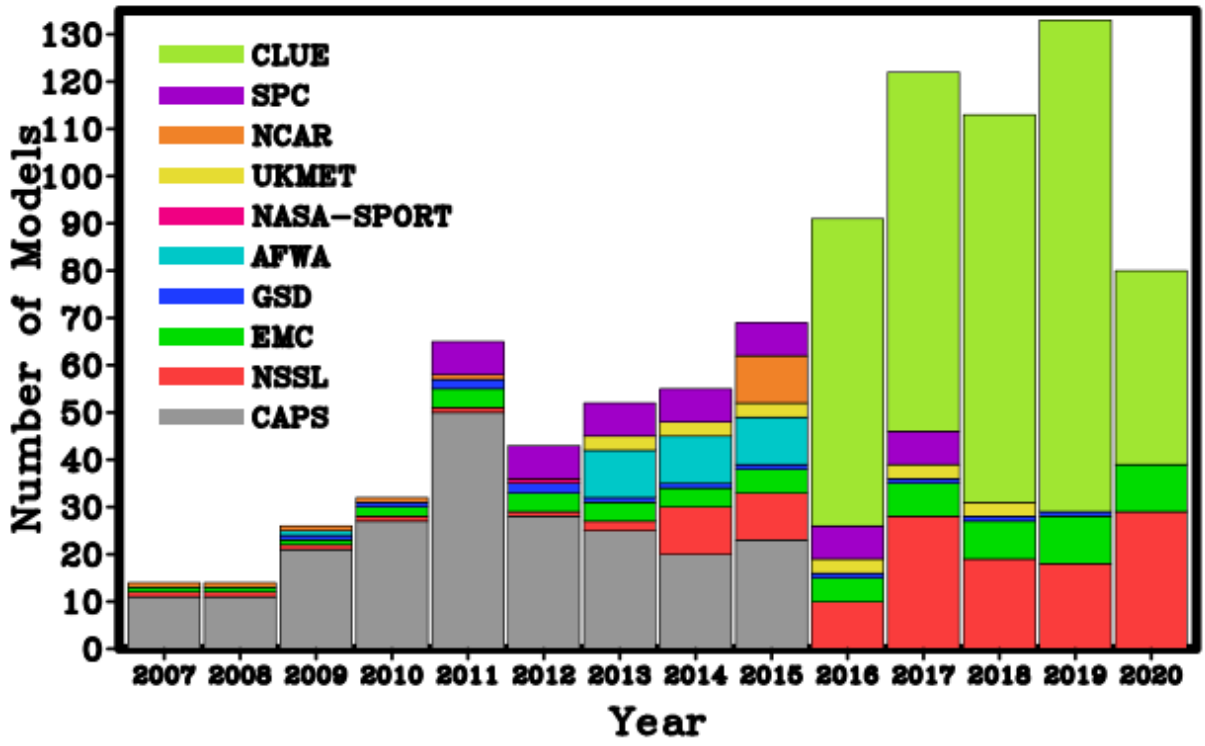


Figure 1 Number of CAMs run for SFEs since 2007. The different colored stacked bars indicate the contributing agencies.

More information on all of the modeling systems run for the 2020 SFE is given below.

1) THE COMMUNITY LEVERAGED UNIFIED ENSEMBLE (CLUE)

The 2020 CLUE is a carefully designed ensemble with subsets of members contributed by NOAA groups at NSSL, GFDL, GSL, and EMC, and non-NOAA groups at NCAR and the UK Met Office. To ensure

consistent post-processing, visualization, and verification, CLUE contributors output all model fields to the same grid. Collaborators based in the U.S. used the Unified Post Processor (UPP; available at http://www.dtcenter.org/upp/users/downloads/index.php) while the UK Met Office used their own in-house software for model post-processing. All groups output a set of storm-based, hourly-maximum diagnostics including updraft helicity over various layers, updraft speed, and hail size, as well as standard CAM diagnostics like simulated reflectivity and precipitation. While the UK Met Office output fields were somewhat limited, U.S.-based groups generally replicated the 2D fields output by the operational High Resolution Rapid Refresh (HRRR) model because of their relevance to a broad range of forecasting needs including aviation, severe weather, and precipitation. The UK Met Office runs covered a 3/4 CONUS domain, while all other CLUE members covered the full CONUS. A full list of members and further details on ensemble configurations are provided in the 2020 SFE operations plan. Table 1 provides a summary of each CLUE subset.

The design of the 2020 CLUE allowed for several unique experiments that examined issues immediately relevant to the design of a NCEP/EMC operational CAM-based ensemble. The primary groups of experiments are listed in Table 2.

Table 1 Summary of 2020 CLUE subsets.

Clue Subset	# of mems	IC/LBC perts	Mixed Physics	Data Assimilation	Model Core	Agency	Init. Time(s) UTC
HRRRv4	1	none	no	GSI-EnVar	ARW	GSL	00-23
HRRRE	9	EnKF	no	EnKF	ARW	GSL	00, 06, 12, 18
gsl-fv3sar	4	none	yes	cold start	FV3	GSL	00
arw-ICs	2	none	no	cold start	ARW	NCAR	00
ukmet	9	MOGREPS-G	no	cold start	UM	UK Met	18, 00
um-ICs	2	none	no	cold start	UM	UK Met	00
nssl-glm	1	none	no	NSSL-VAR (GLM)	ARW	NSSL	00
nssl-noglm	1	none	no	NSSL-VAR (no GLM)	ARW	NSSL	00
nssl-tl	6	none	yes	cold start	ARW	NSSL	02, 03, 05, 08, 09, 11
sarfv3-ICs	2	none	no	cold start	FV3	NSSL	00
emc-fv3sar	3	none	no	cold start	FV3	EMC	00
gfdl-fv3	1	none	no	cold start	FV3	GFDL	00

Table 2 List of CLUE experiments for the 2020 SFE. The CLUE subsets listed are from Table 1.

Experiment Name	Description	CLUE subsets
Model ICs vs. core sensitivity	NCAR, NSSL, and the UK Met office each ran two configurations of the Advanced Research Weather Research and Forecasting model (WRF), FV3-SAR, and Unified Model (UM), respectively, initialized from the Global Forecast System (GFS) and UM global models. **Goal: Examine forecast sensitivity to different initial conditions and model cores at convective scales.**	arw-ICs, um-ICs, & sarfv3-ICs
FV3-SAR Configurations	GSL, NSSL, and EMC ran various configurations of the FV3-SAR with different data assimilation, physics schemes, numbers of vertical levels, and horizontal advection settings. **Goal: Evaluate FV3-SAR sensitivities and find the optimal FV3-SAR configuration for convective weather forecasting.**	gsl-fv3sar, emc-fv3sar, & sarfv3-ICs
Single-model time-lagging	Single-model CAM ensemble configurations with and without time-lagging were compared. **Goal: Assess whether time-lagging results in improved probabilistic forecasts from single model ensembles.**	HRRRv4, HRRRE, ukmet, & nssl-tl
Model-model vs. time-lagging:	Three comparisons were conducted: (1) single-model ensembles initialized from one time, (2) single-model ensembles that are time-lagged, and (3) multi-model ensemble that are time-lagged. **Goal: Evaluate the relative impact of time-lagging and multi-model approaches in CAM ensembles for next-day severe weather forecasting.**	HRRRE & ukmet
Total Lightning Data Assimilation	WRF model configurations with and without assimilation of total lightning data from the GOES 16 GLM were examined. **Goal: Assess whether assimilation of the GLM data improves short-term forecasts (0-12 h) of thunderstorms in radar-sparse regions.**	nssl-glm & nssl-noglm

2) HIGH RESOLUTION ENSEMBLE FORECAST SYSTEM VERSIONS 2.1 & 3 (HREFv2.1 & HREFv3)

The HREFv2.1 is a 10-member CAM ensemble currently running at EMC with forecasts that can be viewed at: http://www.spc.noaa.gov/exper/href/. HREFv2.1 members use different physics, model cores [ARW and Nonhydrostatic Multiscale Model on the B-grid (NMMB)], initial and lateral boundary conditions [North American Mesoscale (NAM) and Rapid Refresh (RAP) models], and half of the members are 12-h time lagged. The design of HREFv2.1 originated from the Storm Scale Ensemble of Opportunity (SSEO), which demonstrated skill during the previous six years in the HWT and SPC prior to HREFv2.1 operational implementation. All members, except for the NAM CONUS Nest and HRRR, are initialized with a "cold-start". Forecasts to 36 h are produced at 0000 and 1200 UTC. The diversity in HREFv2.1 has proven to be a very effective configuration strategy, and it has consistently outperformed all other CAM ensembles examined in the HWT during the last few years.

HREFv3 replaced the High-Resolution Window (HRW) NMMB simulations with emc-fv3sar and HRRRv3 with HRRRv4; results from SFE 2019 found that this change did not have a large impact on the subjectively assessed performance for severe weather forecasting. Thus, EMC will move forward with implementing HREFv3 operationally late in 2020 or early in 2021.

3) NSSL EXPERIMENTAL WARN-ON-FORECAST SYSTEM

The full 36-member, 3-km grid-spacing, real-time WoFS ensemble, run from 1500 UTC Day 1 to 0300 UTC Day 2, is updated every 15 minutes by Gridpoint Statistical Interpolation Ensemble Kalman Filter (GSI-EnKF) data assimilation. Observations that are assimilated include multi-radar, multi-sensor (MRMS) reflectivity and radial velocity data, cloud water path retrievals, clear-sky radiances from the GOES-16 imager, Oklahoma Mesonet observations (when available), and conventional (i.e. prepbufr) observations. All real-time WoFS ensemble members utilize the NSSL 2-moment microphysics parameterization and the RUC land-surface model; however, the planetary boundary layer (PBL) and radiation physics options are varied amongst the ensemble members to increase ensemble spread, given the fact that the EnKF may underrepresent model physics errors. Six-hour (three-hour) 18-member ensemble forecasts are initialized from the real-time WoFS analyses hourly (half-hourly) from 1700 UTC Day 1 through 0300 UTC Day 2.

In addition to the real-time WoFS, enhanced horizontal-resolution WoFS ensemble simulations were produced for next day evaluation that had 1.5-km grid-spacing with 9 members downscaled from the 3-km real-time system and initialized hourly from 1800 to 0300 UTC. Two deterministic enhanced resolution WoFS runs were also produced with one simulation using 3-dimensional variational data assimilation and the other using dual-resolution hybrid data assimilation. All the WoFS forecast are viewable using the newly redesigned web-based WoFS Forecast Viewer (https://wof.nssl.noaa.gov/realtime). The daily WoFS domains targeted the primary region where severe weather was anticipated and covered a 900-km square region.

b) Daily Activities

SFE 2020 daily activities were centered around model evaluations, as well as limited-scope forecasting activities. A summary of evaluation activities and forecast products can be found below while a detailed schedule of daily activities is contained in the appendix (Table A2). Note, when referencing the times in this document at which experiment activities occurred, we use Central Daylight Time (CDT), which is the time zone in which the HWT facility and SFE organizers are based. However, it is worth noting that many of our virtual participants were located in different time zones as far away as the United Kingdom and Australia, so their local time was quite different.

1) FORECAST AND MODEL EVALUATIONS

Compared to previous SFEs, the model evaluation activities of the 2020 SFE consumed a much larger proportion of experiment activities, with all participants engaged in these activities for much of the morning. From 10-11am CDT, evaluations were conducted that involved comparisons of different ensemble diagnostics, CLUE ensemble subsets, HREF, and the WoFS Ensembles. Participants were split into Groups A & B, and each conducted a separate set of evaluations. Participants worked on these surveys individually, but typically stayed in the virtual meeting where SFE facilitators were available to answer any questions or troubleshoot the model evaluation webpage. Then, from 11am to noon CDT, there was a discussion period when participants could talk about interesting aspects of the model evaluations such as forecasts that performed particularly well, model behavior that seemed odd, and/or

especially large differences between different experiments, etc. Another evaluation period was focused on the previous day's experimental forecast products and occurred from 1:40 to 2pm CDT. In these evaluations, experimental forecasts were compared to observed radar reflectivity, local storm reports (LSRs), NWS warnings, and MRMS radar estimated hail sizes.

2) EXPERIMENTAL FORECAST PRODUCTS

Because of the COVID-19 pandemic, forecasting activities were limited in scope and occurred virtually from 1:30-4pm daily with a focus on adding temporal specificity to convective outlooks within the Day 1 time period using WoFS Ensemble datasets. Participation was limited to a small internal group, as well as weekly groups of NWS forecasters. 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. As in previous years, two sets of unique outlooks were generated by the R2O and Innovation groups. Both groups issued outlooks for the probability of individual hazards (tornado, wind gusts 50 knots, hail 1.0 in.) within 25 miles (40 km) of a point.

For the R2O group, participants updated the operational SPC 1630 UTC Day 1 Outlook hazard probabilities for the period 2100 – 1200 UTC. Additionally, conditional intensity forecasts were generated, for which SPC's operational probabilities of significant severe hazards (EF-2 or greater tornadoes, winds 65 kts, or hail 2 in.) could be used as a starting point. This was the second year that the R2O group has issued conditional intensity forecasts. These forecasts delineate areas that are expected to follow a "normal", "hatched", or "double-hatched" intensity distribution. In plain language, "normal" refers to a typical severe weather day, where significant severe weather is unlikely, "hatched" areas indicate where significant severe weather is possible, and "double-hatched" areas indicate where high-impact significant severe weather is expected. These forecasts could also be thought of as indicating the proportion of observed reports that are expected to be severe, where going from "normal", to "hatched", to "double-hatched" would indicate an increasing proportion of significant-severe reports (see Fig. A4 of Appendix for more detailed information on each hazard). One set of forecasts was generated 2-3pm for which WoFS data was not used, and a final set of forecasts was generated 3-4pm with available WoFS datasets.

For the Innovation group, participants generated severe hazard probabilities valid over a short time window, 4-5pm (2100-2200 UTC), and a long time window, 4-8pm (2100-0100 UTC). An initial forecast was generated during the 2-3pm period and an updated final forecast during the 3-4pm period. For both sets of initial and final forecasts, one group of forecasters used all available datasets except for WoFS, while another used all available datasets including WoFS. For the Innovation group, forecasting individual hazards was a shift from previous years when all the outlooks were focused on total severe (i.e., all hazards combined).

3. Preliminary Findings and Results

a) Model Evaluations – Group A

A1) ISU ML SEVERE WIND PROBABILITIES

Daily evaluations were performed for four machine-learning (ML) algorithms designed to diagnose the likelihood that storm reports for thunderstorm winds were due to winds 50 knots. The algorithms were created based on the perception that many wind values assigned to thunderstorm wind damage reports lacking a measurement in the Local Storm Reports (LSRs) database are overestimated. Our analysis of all ~180,000 thunderstorm wind reports during 2007-2018 found that 89% were estimates, and 40% of these were assigned a value of exactly 50 knots. Only 13% of measured reports had a value of 50 knots. The algorithms were trained using the ~18,000 thunderstorm wind damage reports that had a measured wind value during 2007-2017. From the LSRs dataset, date, time, location and episode and event narrative data were used for training along with a spatial-temporal scaled distance to the nearest measured report value. Also, 31 parameters from SPC Mesoanalyses were used as input for training over a 200 x 200 km box centered on each storm report. Based on testing on the 2018 measured thunderstorm wind reports (for which ground truth exists), three algorithms (gradient boosted model – GBM, support vector machine – SVM, and an average ensemble method) that scored best with area under the ROC curve and Brier Score were chosen, along with the well-known neural network technique, for evaluation.

Python-generated maps of the previous day's wind reports with the probability diagnosed by each of 4 algorithms shown using different levels of blue color and differentiated between estimated reports (squares) and measured ones (triangles) were shown to participants (Fig. 2) via a web site (https://sites.google.com/iastate.edu/storm-report-anonymous/yesterdays-storm-reports). Algorithm names were hidden to prevent bias. A table with details of each report and the specific probabilities for each of the four algorithms was also available on the website. The 1630 UTC outlook from SPC for probability of severe thunderstorm wind occurrence could be toggled on or off.

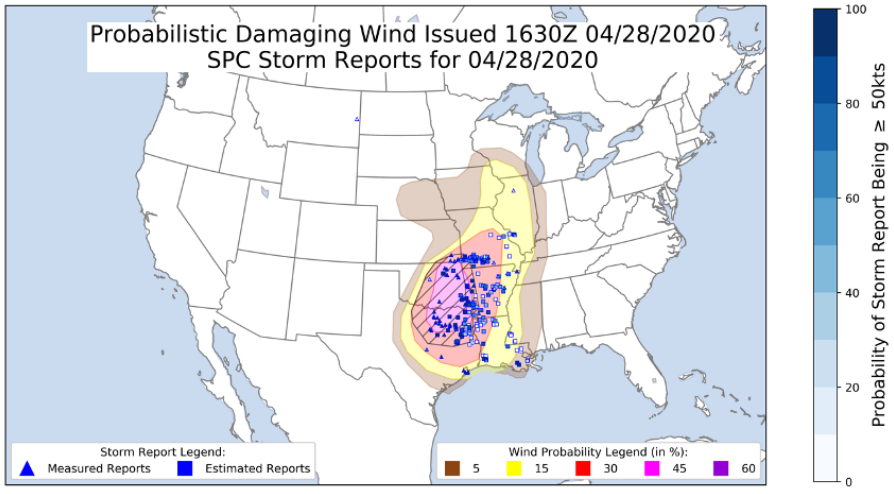


Figure 2 Storm reports color-coded for probability value (scale on right) diagnosed by ML algorithm with 1630Z SPC probabilistic damaging wind graphic overlaid for a case in late April during the 2020 SFE.

The participants were asked to evaluate the ease of use of the website, and then to subjectively assign a rating for the perceived value they felt existed in the output from each of the 4 algorithms on each day. Since there is no ground truth for any of the estimated reports (which make up the majority of all storm reports), subjective evaluations were challenging. It was felt that the best way to evaluate the algorithms would be to use two pieces of information. First, assuming SPC forecasters have some skill in their forecasts of severe wind likelihood, it would make sense that higher probabilities should be assigned by the algorithms to reports that occurred within regions where SPC forecasters had placed higher probabilities. Secondly, since the measured reports (triangles in the plots) were known to have winds above the severe threshold, the algorithms should diagnose high probabilities at those points.

Average subjective scores for the four algorithms were relatively similar (Fig. 3), near 6 on a 10-point scale, implying participants felt the techniques had value. The average ensemble scored highest, matching what objective measures showed applied to the 2018 test set (not shown). Analysis of daily results shows that the average scores of participants tended to correlate with the probabilities assigned by the algorithms (Fig. 4). Although this makes sense for measured reports where probabilities should be high, the fact it was true also for estimates (not shown) may reflect a bias in the community that any storm report should be believed. Participants generally liked the website with an average score of 6.17. Specific comments emphasized the need for (1) zooming features since it was at times difficult to read the assigned probability values when reports overlapped, (2) perhaps a different color scale using multiple colors to make it easier to know the exact probabilities, and (3) the ability to scroll over a report and see the exact value of probability assigned by each algorithm. Many users commented that they could see the value in this approach, and that the results made sense, as numerous severe weather events led to many reports involving tree damage (without any measured reports) in the eastern part of the country which were assigned low probabilities that winds were 50 knots or more, while measured reports, often in the central U.S., had much higher probabilities. Also, there was good correspondence of lower probabilities being assigned to reports that happened outside the SPC 5% risk, or within the lowest SPC probability category (5-15%), and higher probabilities where SPC had forecast a greater chance of severe winds (Fig. 5). This agreement likely is because our algorithms use environmental data in a similar way to the human forecasters at SPC.

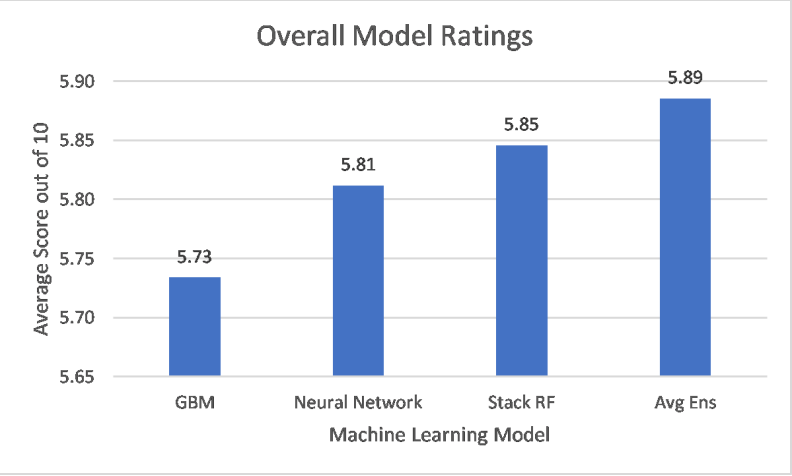


Figure 3 Average of all evaluations from 86 participants during the 2020 SFE for the four ML algorithms.

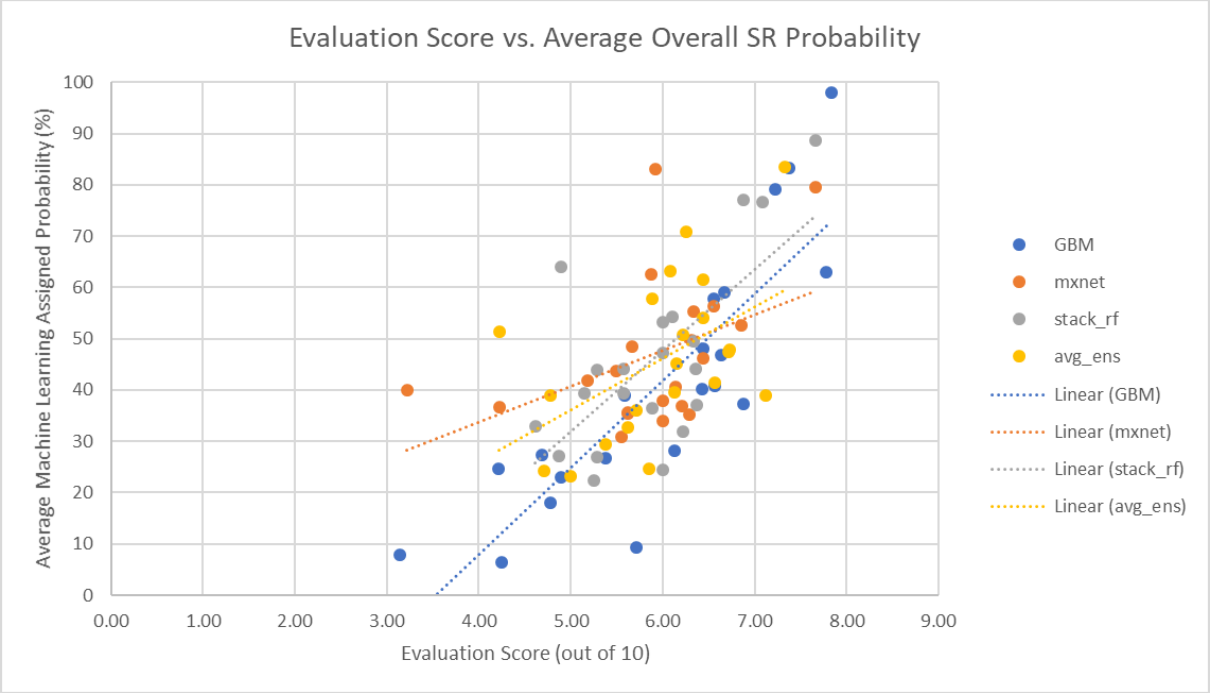


Figure 4 Comparison between the average daily evaluation score and the average daily storm report probability for each algorithm (different colors). Each point represents a different day.

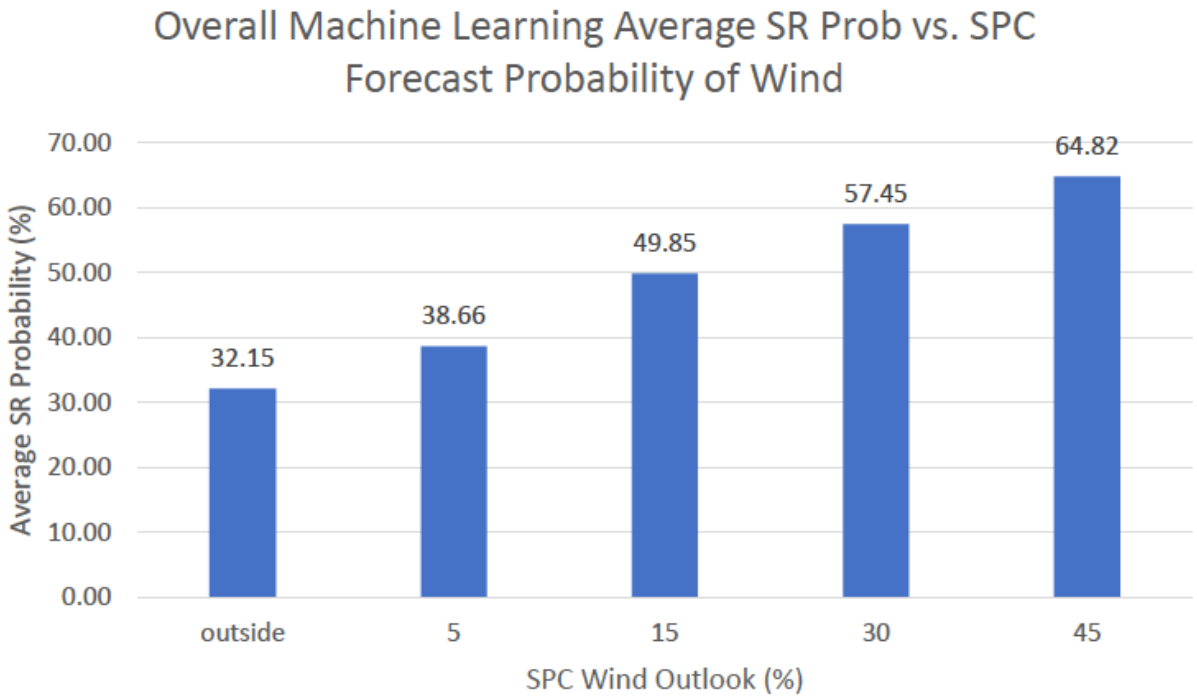


Figure 5 Comparison between the 1630 UTC SPC forecast probability of severe wind occurrence and the average storm report probability (from 4 algorithms) for all cases during the 2020 HWT SFE.

Brier scores were also computed for the 2020 SFE cases (Table 3) to determine if subjective ratings matched an objective measure. Unfortunately, there were not enough events to compute Areas Under the ROC Curve, which had been done for the 2018 test set of data. Unlike with the archived Storm Reports dataset where episode narrative or event narrative text (or both) exists for all reports, for the real time reports available during SFE, some measured values lack any text information. Therefore, a separate version of the algorithms had to be trained to be applied to reports lacking text information. The Brier Scores in Table 3 are an average of the output that used the original algorithms and that which had to use versions for which no text data were used in training. These values are not as skillful as those obtained using the 2018 test set, where Brier Scores were as low as 0.145. Of note, when only the algorithms not using text data were used for the 2020 SFE cases, BSs improved and were as low as 0.167 for GBM and 0.17 for Ave Ens. Unlike in 2018, where the use of text data improved skill for some algorithms, during 2020 it diminished the skill for all algorithms. This result implies a challenge for the development efforts, as real time storm reports end up with a different distribution of text than the archived reports which must be used for training. It is possible the lower BSs during the 2020 SFE were also due to an unusual severe weather season, with an almost complete absence of storm reports in the north-central US (which would typically be relatively active in May) and a relative abundance of reports in the Northwest and the East. Over the long period used for training, a substantial majority of measured reports occurred in the central US with far fewer in the western and eastern parts of the country. The BSs in Table 3 agree with the subjective scores in showing the Average Ensemble to be best, and for the Neural Network to perform worse. However, the objective and subjective scores differ greatly for GBM, which has the second-best average BS, but the lowest subjective score.

Table 3 Brier Scores averaged over all 2020 SFE cases for the four algorithms tested.

algorithm	GBM	Neural Network	Stack RF	Ave Ens
BS	0.203	0.243	0.236	0.190

Future work will include adding composite radar reflectivity and azimuthal shear data to the training, along with population density, land use, and elevation information. In addition, although to some participants it seemed reasonable that the algorithms assigned low probabilities to the majority of estimated storm reports that involved tree damage in the eastern United States, further investigation should be performed to be sure the lower probabilities are not an artifact of less measured reports typically occurring in this part of the country, which might influence the training process to not recognize environments that are relatively humid and that lead to wet microbursts with truly severe winds. Although it is possible the new data sources will reduce any problems that might be present, further study will include the development of algorithms specific to three regions of the country – the East, Central, and West, to see how sensitive the assigned probabilities are to the distribution of the training data. In addition, further study will be performed for textual analysis techniques, as it would seem the textual information should improve objective skill scores and not harm them.

A2) NCAR ML HAZARD GUIDANCE

For the 2020 virtual HWT SFE, NCAR tested a ML-based system to produce convective hazard guidance using output from a deterministic CAM forecast. The goals for the HWT were to assess the added value of the ML hazard guidance compared to contemporary UH-based hazard guidance derived from CAMs, as well as optimize the presentation of the forecast guidance.

Hazard guidance was generated from a neural network (NN) and random forest (RF) that were trained with an upscaled (80-km grid) set of diagnostics from ~400 deterministic 3-km WRF forecasts from events in 2010-2015 and corresponding storm reports. Output from a real-time deterministic 3-km WRF forecast configured identically to the model used for training was input into the trained ML models to make the predictions. The ML models were configured to produce guidance for wind, hail, tornado, significant wind, significant hail, and any severe report at 80-km grid point locations within the contiguous United States. The models were trained to predict reports within 2-hr, and 40 km and 120 km of each forecast hour and grid point for each of the six hazard types. The ML any-hazard predictions were compared to a smoothed UH forecast produced by thresholding the UH field. The UH threshold varied with latitude, longitude, day of the year, and time of day. The NN, RF, and UH output for each forecast hour were visualized and compared in a web-based tool (Fig. 6). The visualization interface and archive of all forecasts from Spring 2020 is available at: https://www2.mmm.ucar.edu/projects/ncar_ensemble/camviewer/.

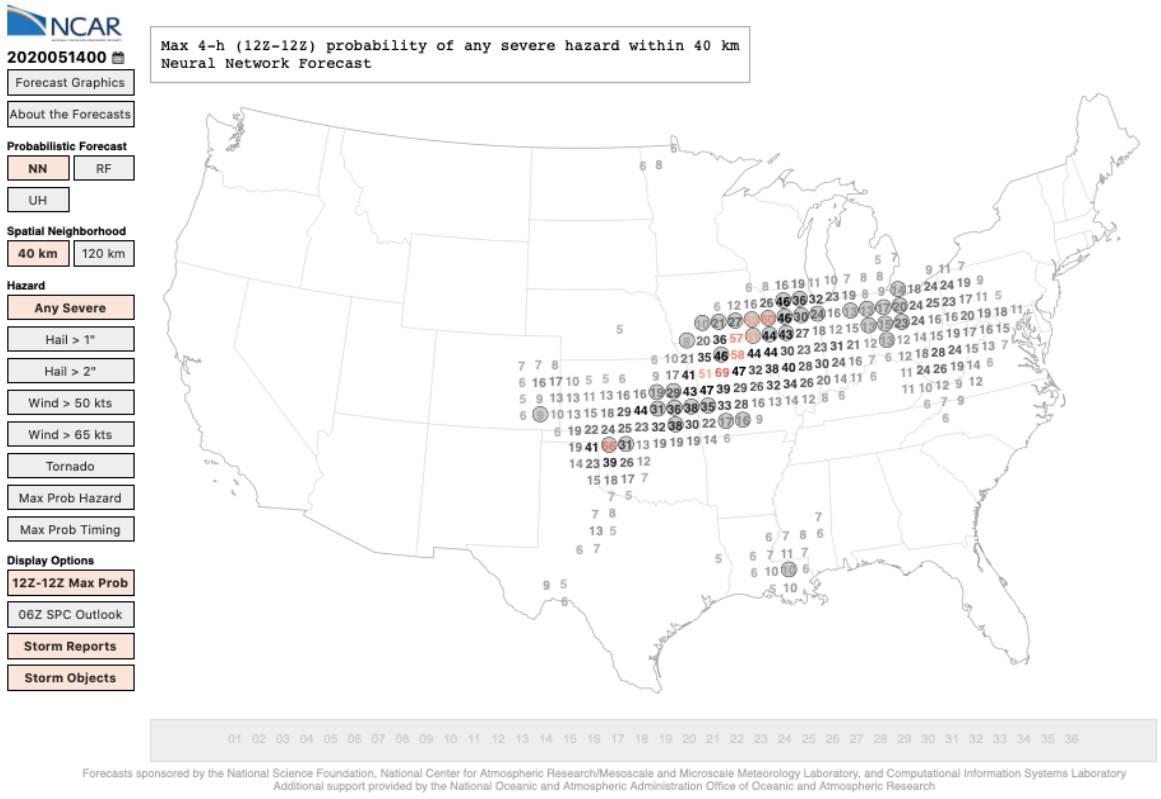


Figure 6 Screenshot of visualization interface displaying 40-km any-hazard NN forecasts for 14 May 2020. Grid boxes with at least one storm report are highlighted in gray circles.

Participants were asked to provide subjective feedback on various aspects of the ML forecasts, including rating the NN, RF, and UH any-hazard forecasts at both the 40-km and 120-km length-scales, and rating the individual hazard forecasts at the 40-km length scale. Subjective skill was evaluated by comparing forecast guidance with the locations of storm reports. Open-ended questions were provided for general feedback on the utility of the ML guidance and the web-based visualization interface. The survey was administered using Google Forms. There were 242 unique responses during the HWT period. A summary of the responses related to the four main evaluations are provided below.

i) Evaluation of any-hazard NN, RF, and UH forecasts at 40-km spatial scale

Participants were asked to "*rate the 12Z-12Z Maximum 4-hr, 40-km any-hazard probabilistic forecasts on a scale of 1-5.*" for the NN, RF, and UH 40-km forecasts. Additionally, they were asked to rate the 120-km NN any-hazard forecasts. The average ratings and most commonly assigned rating (i.e., mode; in parentheses) for the 40-km NN, RF, and UH forecasts were 3.53 (4) , 3.30 (3), and 2.69 (3), respectively (Fig. 7). While the most commonly assigned rating for the 40-km RF and UH forecasts was the same (3), the number of ratings of 1 or 2 was twice as large for the UH forecasts, indicating many more poor forecasts relative to both the RF and NN forecasts. ***Conclusion:* while the NN any-hazard guidance was subjectively more skillful than the RF, the differences were modest. More importantly, the NN and RF were deemed superior to the UH guidance at the 40-km spatial scale.**

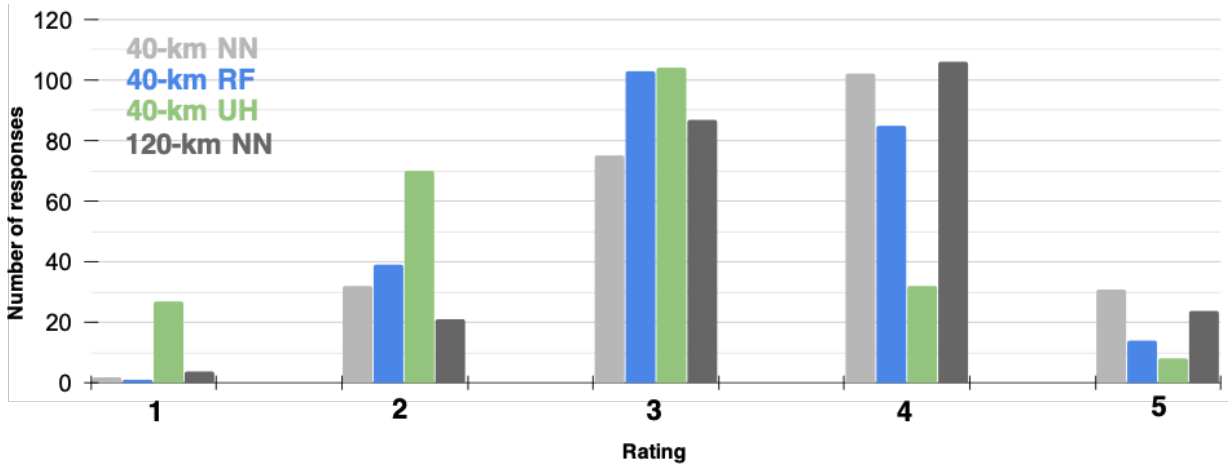


Figure 7 Histogram of ratings for 40-km any-hazard NN, RF, and UH forecasts, and the 120-km any-hazard NN forecasts.

ii) Evaluation of hail, wind, and tornado 40-km NN and RF forecasts

Participants were asked to "*rate the individual hazard probabilities for hail 1", wind 50kts, and tornado.*" for the 40-km NN and RF forecasts. The average ratings for the 40-km NN wind (3.21) and hail (3.35) probabilities were higher for the NN compared to the RF (3.12 and 3.05, respectively), consistent with the results of the any-hazard evaluation (Fig. 8). The opposite was true for the tornado probabilities, as the average rating was higher for the RF (3.41) compared to the NN (3.24), although the distribution of

scores for the tornado ratings was larger, with more 1 point and 5 point ratings compared to the hail or wind forecasts (Fig. 8). Given the tendency for the NN probabilities to be higher than the RF (see objective verification below) and the relative rarity of tornadoes during the 2020 HWT, the NN probabilities were likely too aggressive, although participants noted both the NN and RF tornado probability magnitudes were in line with typical SPC outlook probabilities. Given the lack of tornado events during 2020, the NN and RF tornado forecast ratings were likely heavily weighted by non-events (e.g., high subjective ratings for having near zero probabilities when tornadoes did not occur), boosting the mean rating. The wind probabilities were rated the lowest for both the NN and RF, suggesting the ML algorithms have slightly more difficulty anticipating the locations of convectively generated wind reports relative to hail.

Open-ended feedback was also solicited related to the ability of the guidance to distinguish between hazards on a given day. The majority of respondents noted that the guidance did an excellent job of distinguishing between days and regions where wind or hail was the primary observed hazard, and often kept tornado probabilities near-zero when tornadoes did not occur. Additionally, it was noted during many forecasts that the ML guidance often correctly captured the transition from a hail threat with initial convection, to a severe wind threat as convection became more linear and grew upscale. ***Conclusion: Similar to the any-hazard predictions, the NN individual hazard forecasts for hail and wind were superior to the RF forecasts, although the differences were modest. The RF and NN tornado predictions appear credible, although the lack of events in 2020 biases their ratings toward non-events. Both the NN and RF were able to distinguish between likely hazard types on a given day, especially between hail and wind and the transition between those two hazards.***

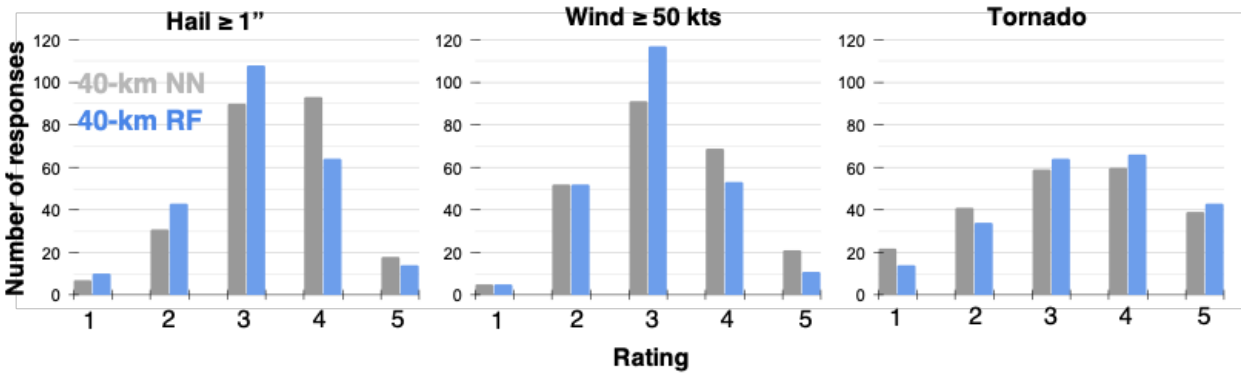


Figure 8 Histogram of ratings for the 40-km NN and RT forecasts for (left) hail 1.0-in, (center) wind 50 knots, and (c) tornadoes.

iii) Evaluation of 40-km vs. 120-km ML guidance

Participants rated the 120-km NN guidance (for comparison to the 40-km NN guidance) and were also asked "*did you prefer either the 40-km or 120-km forecast guidance for this forecast? If so, why?*" The average rating for the 120-km NN forecasts was 3.51, almost identical to the mean forecast rating for the 40-km NN forecast of 3.53 (Fig. 7). Even though the subjective ratings were the same when evaluated with storm reports, most participants preferred the 40-km guidance. Of the 230 responses to the open-ended feedback question, 131 (57%) indicated a preference exclusively for the 40-km guidance and 60 (26%)

indicated a preference exclusively for the 120-km guidance. 30 responses (13%) indicated that both the 40-km and 120-km guidance were both useful, with several participants noting that interrogating both spatial scales was beneficial. 9 responses (3%) indicated no preference.

Participants that preferred the 40-km guidance often disliked the large areas of false alarm that occurred within the 120-km guidance, and preferred the spatial detail and granularity present when using the 40-km spatial scale. Those that preferred the 120-km guidance often mentioned the enhanced probability of detection and that the 120-km spatial scale was useful to define the overall threat area on a given day, acknowledging that the spatial details in the 40-km forecasts were more often inaccurate than the smoother 120-km probabilities. Additionally, many participants noted that the 40-km probabilities were aligned better with SPC probabilistic products, and that the 120-km probabilities were too large. ***Conclusion: the subjective feedback and ratings support the presentation of the ML guidance products on both spatial scales, with a preference for the 40-km probabilities due to their detail and equivalence to SPC outlook probabilities.***

iv) Feedback on visualization system and products

The feedback was overwhelmingly positive regarding the visualization interface. Feedback generally focused on the intuitive design of the interface, although many suggested the inclusion of additional products and features (e.g., overlay of NWS warnings, ability to zoom in to regions, and providing the option to overlay probability contours instead of numeric probabilities). Several participants noted that the color choices for the probabilities should be clarified (some participants wondered if the shading was meant to convey some property of the forecast) and that a better description of some of the products should be added to the site. ***Conclusion: the web interface was a useful tool for visualizing the ML and UH hazard probabilities. Additional features should be added in the future to enhance the interface, with clarification of existing products and color choices.***

In addition to information compiled from the HWT surveys, objective verification was also performed. The 40-km and 120-km NN, RF, and UH any-hazard forecasts produced between 1 March 2020 through 30 June 2020 (122 events) were verified with SPC preliminary storm reports for all CONUS 80-km grid points. Verification statistics were computed with all 4-hr forecasts valid between 12Z - 12Z. While the verification includes a larger subset of cases than those subjectively evaluated during the HWT, inclusion of several months of forecasts allows for a more robust objective evaluation of forecast system performance. Verification metrics included the Brier skill score (BSS), area under the relative operating characteristic curve (AUC), and attributes diagrams. The BSS was computed with a 30-year severe weather climatology that varies based on day of year, time of day, and grid box.

Both the NN and RF 4-h any-hazard forecasts possessed excellent forecast resolution at both the 40-km and 120-km spatial scales, with AUCs of ~0.96, while the UH forecasts had smaller AUCs of ~0.85 (the AUCs were similar for both the 40-km and 120-km length scales; Fig. 9). The AUC differences reflect the ability of the ML forecasts to provide non-zero probabilities for a much larger fraction of events, in cases where the UH forecasts were zero (e.g., in situations where the UH threshold was not met). The AUC values and AUC differences between the ML and UH forecasts for the 2020 forecasts are very similar

to the AUC computed using previous seasons during initial testing (e.g., 2016). The differences in AUC between the two ML techniques were much smaller, e.g., ~0.02 at both length-scales.

While the ML forecasts possessed large AUCs, the reliability characteristics varied between the NN, RF, and UH forecasts (Fig. 9). At the 40-km length-scale, the NN forecasts had a pronounced overforecasting bias, while the RF forecasts were better calibrated. UH forecast reliability tended to be between the RF and NN forecast reliability curves. All forecasts were better calibrated at the 120-km spatial scale, but still tended to overforecast, with the overforecasting issue largest for the NN compared to the RF. Finally, the ML forecasts produced larger BSSs than the UH forecasts at both 40-km and 120-km, with the BSS differences being largest at 120-km (Fig. 9). For example, the BSS gained by using the RF over the UH forecasts was ~0.04 at 40-km and ~0.09 at 120-km. The addition of non-zero probabilities by the ML techniques in areas where the UH forecasts were zero likely led to these differences, although differences in reliability may have played a role as well.

These results confirm subjective impressions of the guidance received in the surveys and through informal discussions during the HWT period. After the conclusion of the experiment, the PIs were able to reduce the NN overforecasting bias by modifying the training parameters of the NNs, as well as adding more training data. Future versions of the ML algorithms will incorporate this large training dataset and modified training parameters. ***Conclusion**: the 4-h ML forecasts better discriminated between severe and non-severe storms than the 4-h UH forecasts during Spring 2020 (i.e., had larger AUCs and BSSs) at both the 40-km and 120-km length scales. The 4-h NN forecasts tended to be poorly calibrated more so than the RF and UH 4-h forecasts, especially at 40-km. Overforecasting was mitigated in the 2020 forecasts in post-HWT experiments by using a larger training dataset and modified training hyperparameters.*

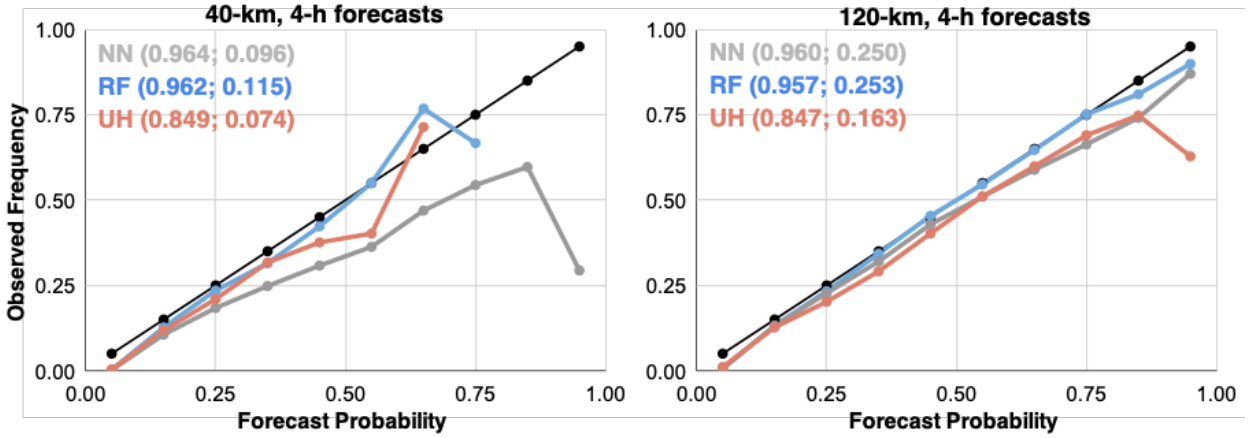


Figure 9 Reliability diagrams for (left) 40-km and (right) 120-km, 4-h, NN, RF, and UH any-hazard forecasts issued between March 2020 and 30 June 2020. AUC and BSS are provided in parentheses.

A3) CLUE 0000 UTC CAM TL-Ensemble

Three separate experiments were conducted during the 2020 SFE to evaluate time-lagging as a formal strategy for CAM ensemble design. These evaluations were focused over a mesoscale area of interest with the greatest potential for severe weather over the CONUS during the convective day (i.e., 1200-1200 UTC). The forecast field most commonly examined during this severe weather evaluation was the 24-h summary of 2-5 km AGL hourly maximum UH. The ensemble maximum UH and neighborhood UH probabilities (>99.85th percentile) were displayed along with preliminary local storm reports (e.g., Fig. 10), and participants rated the forecasts (on a scale of 1-10) based on the quality of guidance provided to a severe weather forecaster. One of the CAM ensemble evaluations (i.e., A3: CLUE 00Z CAM TL-Ensemble) compared two single-model ensembles (HRRRE and UM) initialized at 0000 UTC to their respective time-lagged ensembles (i.e., half of the members each from 0000 UTC and 1800 UTC). These single-model ensembles were compared to the HREFv2.1 and HREFv3 (Fig. 10), which serve as the operational baseline for CAM ensemble performance.

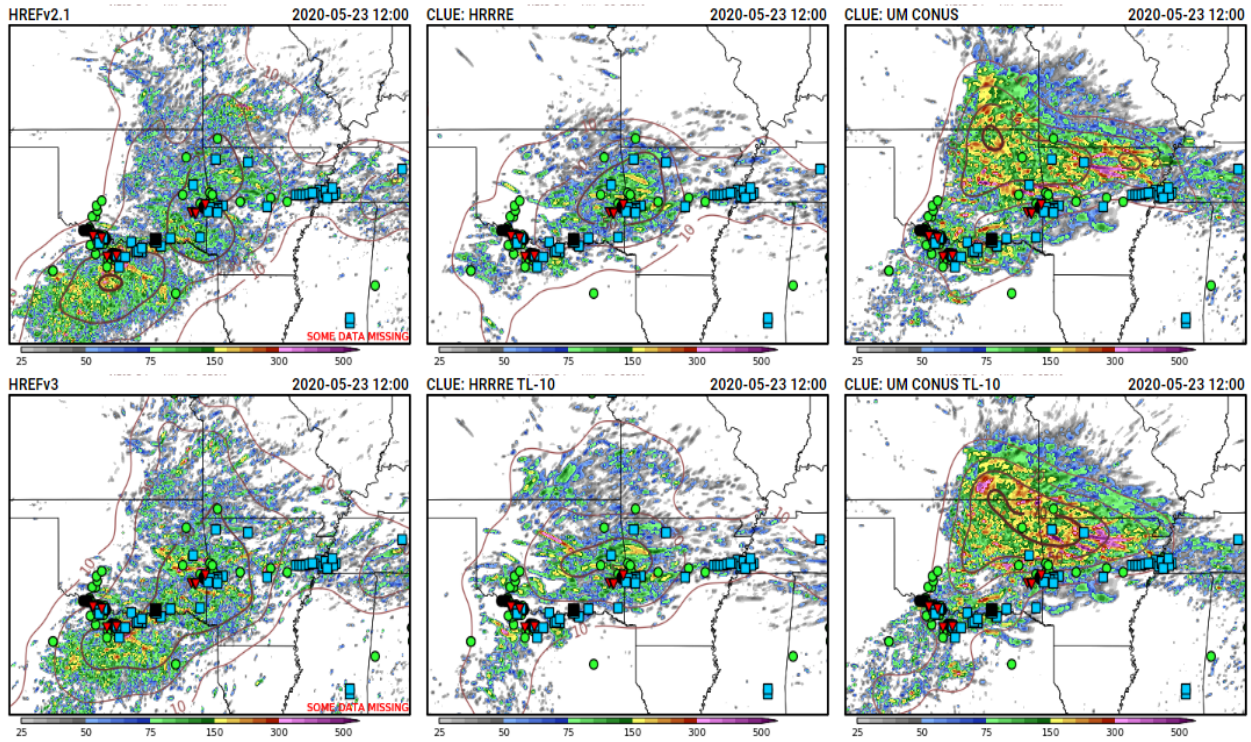


Figure 10 Example of multi-panel comparison webpage for the 0000 UTC CAM ensemble A3 evaluation during the 2020 SFE. The 24-h ensemble maximum UH (shaded) and neighborhood probability of UH>99.85th percentile (contoured) is displayed for HREFv2.1 (upper left), HRRRE (upper middle), UM (upper right), HREFv3 (lower left), HRRRE TL-10 (lower middle), and UM TL-10 (lower right) for 22 May 2020. Preliminary severe storm reports are also overlaid (wind - blue squares, hail - green circles, and tornado - red upside-down triangles).

The distribution of subjective ratings by participants during the five-week SFE for the time-lagged ensembles in the A3 evaluation were similar to their respective 0000 UTC ensembles (i.e., HRRRE TL-10 vs. HRRRE and UM TL-10 vs. UM; Fig. 11). The SFE participants commonly noted that daily forecasts from the time-lagged ensembles looked very similar to the corresponding non-time-lagged forecasts. While the time-lagged ensembles did not often improve probabilistic forecasts of severe weather, they also did not notably degrade the forecasts. Thus, more rigorous and quantitative investigation should be conducted to determine the optimal number of ensemble members to run at a single time for CAM ensemble applications, given the expense to run additional forecast members. As has been demonstrated in previous SFEs, the HREF continues to stand as a formidable baseline for experimental CAM ensembles, with the HREFv3 receiving the highest ratings overall of the CAM ensembles (Fig. 11).

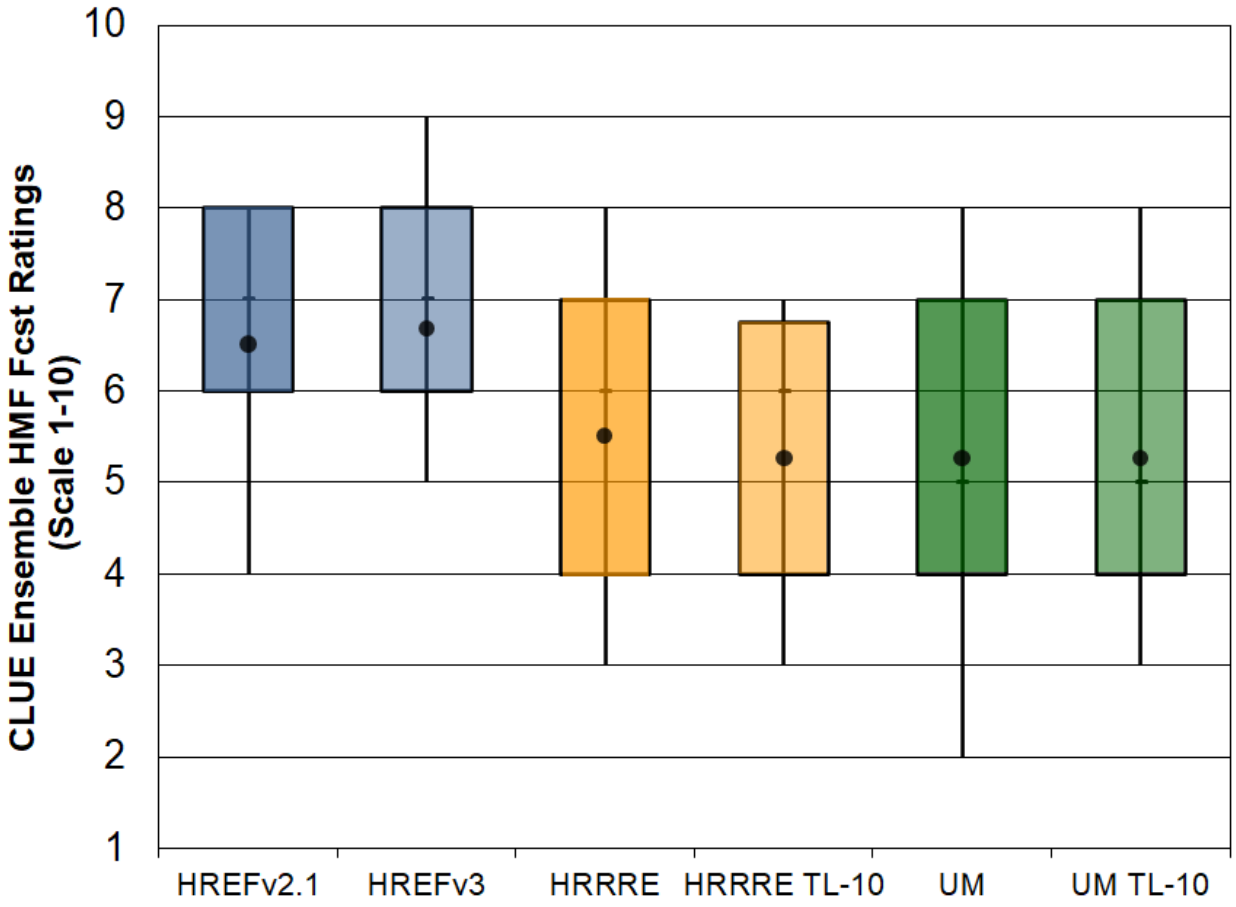


Figure 11 Distributions of subjective ratings (1-10) by SFE participants of hourly maximum fields for severe weather forecasting over a mesoscale area of interest for the forecast hours 13-36 for the A3: CLUE 00Z CAM TL-Ensemble evaluation (HRRRE - orange; HRRRE TL-10 - light orange; UM - green; UM TL-10 - light green) compared to the HREFv2.1 (blue) and HREFv3 (light blue).

A4) CLUE TTU ENSEMBLE SUBSETTING

A daily evaluation of probabilities from a time-lagged 18-member HRRRE ensemble and those based on 6-member ensemble subsets chosen objectively through the sensitivity-based subsetting technique (Ancell 2016) was performed at the 2020 SFE. Ensemble sensitivity was calculated within the real-time Texas Tech University (TTU) 42-member DART/WRF ensemble assimilation and forecasting system, and HRRRE member subsets were chosen based on the smallest errors in sensitive regions measured against TTU EnKF analyses. The main goal of this evaluation was to assess the generality of the subsetting technique, and whether it can be used across ensemble systems as a viable operational tool.

Each day, a response function location was chosen based on input from SFE participants through a web-based graphical user interface (GUI). The GUI was used to identify locations in which uncertain probability signals (typically 30-70% probability hourly maximum 2-5km UH exceeded 25 m^2/s^2 within 20 miles of a point) existed in both the TTU and HRRRE ensemble forecasts. The Day 1 response function area was selected at a forecast hour between 1800 UTC (the 18hr forecast) and 1200 UTC (the 36hr forecast) in areas that exhibited high uncertainty over the prior 6hr period. This response selection process was performed by viewing the full TTU 42-member probabilities of exceeding 25 and 100 m^2/s^2 2-5km UH valid over the 18-36hr forecast period, as well as the HRRRE 9-member probabilities of exceeding 75 m^2/s^2 2-5km UH over the 24-42hr and 18-36hr periods for the 1800 UTC and 0000 UTC initializations, respectively.

Once the response function time and location were chosen, the sensitivity for a preset response function (the number of grid points exceeding 50 m^2/s^2 2-5km UH) was calculated. These sensitivities were calculated with respect to 250-, 500-, 700-, 850-hPa temperature, winds, and geopotential heights, 2-meter dewpoint and temperature, 10-meter U- and V-wind components, and SLP on the 12-km TTU CONUS domain from the 7-hr forecast state (valid 0700 UTC). The 6 ensemble members that possessed the smallest sensitivity-weighted errors (chosen using the sum resulting from the projection of the ensemble differences with the analysis onto the ensemble sensitivity field over the greatest 50% of sensitivity magnitudes) were chosen. The analysis used to determine the errors was the 1hr forecast ensemble mean (valid at 0700 UTC) from the 0600 UTC Texas Tech ensemble initial conditions determined through the DART EAKF data assimilation procedure. The 1hr forecast at 0700 UTC was used in lieu of the analysis valid at 0600 UTC due to significant imbalance present after the assimilation procedure. Probability fields of 75 m^2/s^2 2-5km UH and 40 dBZ simulated near-surface reflectivity of Day 1 convection were generated for the 6-member ensemble subset and the best member of the subset, which were subsequently compared against probabilities from the full ensemble the following day after the severe event occurred.

Differences between the full ensemble and subset probabilities as well as the best member forecasts were subjectively evaluated by participants, relative to SPC storm reports, NWS warnings, and practically-perfect probability fields. Figures 12 and 13 show two examples of the subsetting product during the 5-week experiment that participants evaluated. Figure 12 depicts a successful case for convection in southern Texas valid between 2000 and 0000 UTC (forecast hours 20-24) on May 27. Probabilities of UH exceeding 75 m^2/s^2 are shown from the full ensemble (left), the 6-member subset (middle), and the best member (right), with storm reports and NWS warnings overlaid. The subset increased probabilities near the epicenter of the most active weather and the best member better highlighted the axis of the severe activity. This case is particularly interesting given that the subset

improves the forecast even though a wide-range of different hazards are produced within a relatively small spatiotemporal window; the primary threat in the western half of the probability signal appears to be hail, while wind and possible embedded tornadoes dominate in the eastern half of the signal. In contrast, Figure 13 shows a failure case for convection in southwest Kansas on 21 May 2020. In this case the UH coverage subset produced reduced probabilities near the area of most active severe weather and increased probabilities to the south, where far fewer severe reports occurred. The best member performed even more poorly, with the highest probabilities (smoothed from a single deterministic solution) between two clusters of active weather, and zero probabilities in the area with the most reports.

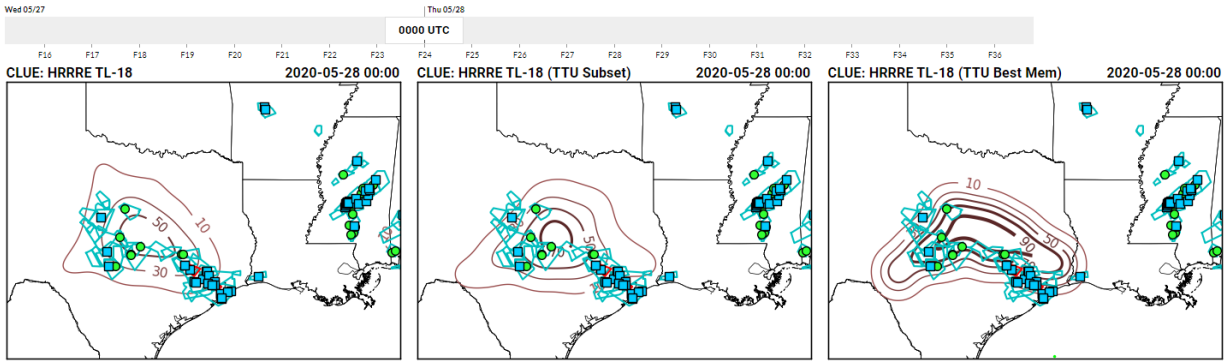


Figure 12 Smoothed 40-km neighborhood probabilities of UH > 75 m^2/s^2 from full time-lagged 18-member HRRRE (left), 6-member HRRRE subset (middle), and best deterministic subset member (right) valid from 2000 UTC to 0000 UTC (forecast hours 20-24) on May 27-28. LSRs and NWS warning polygons are overlaid for subjective evaluation.

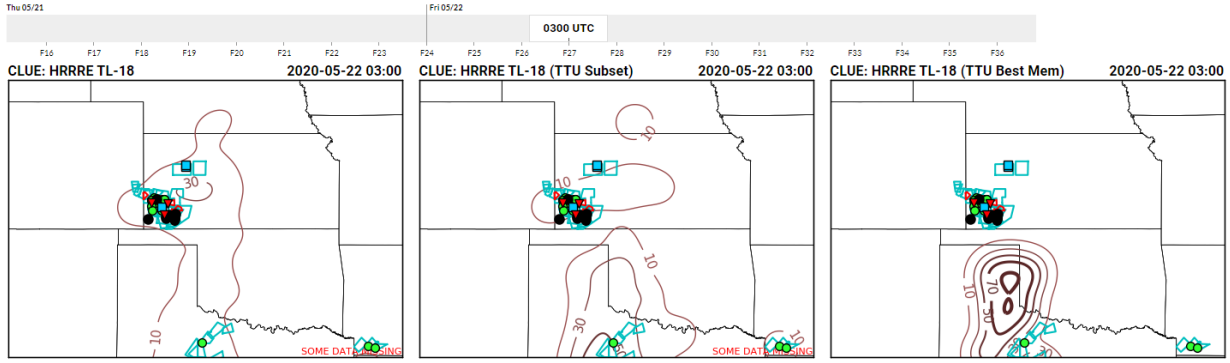


Figure 13 As in Figure 12, except valid from 1200 to 0300 UTC (forecast hours 23-27) on 21-22 May.

In general, failure cases outnumbered success cases. This result was apparent through participant feedback (Figures 14 and 15) that indicated that most of the time, the subset skill remained approximately the same as that of the full ensemble, and that the technique degraded probability fields more often than it improved them. Further, the best member rarely added value in the opinion of participants, and in fact detracted from the overall forecast guidance most of the time. While further objective analysis of the five weeks of cases is planned, an initial assessment reveals that the cross-system application of the subsetting process likely suffers from biases in the different systems. In particular, ensemble sensitivity fields become inflated in areas of under-dispersiveness, substantially overweighting subsetting projections there. This

tended to occur here as TTU surface fields (e.g. 2-meter temperature) with too little spread caused the subsetting procedure to choose HRRRE members with the smallest surface field errors (which can have their own bias issues). Addressing this issue will be a primary focus in the near future.

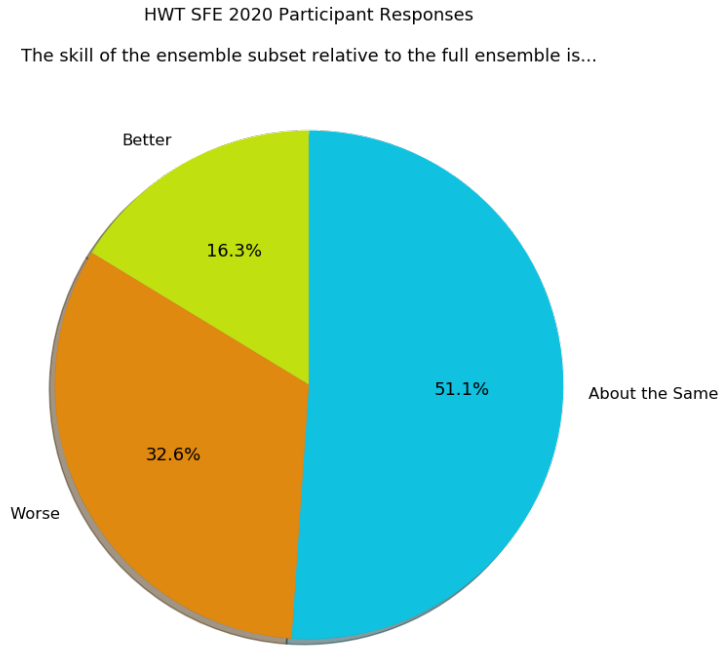


Figure 14 Breakdown of participant responses to the survey question: "The skill of the ensemble subset relative to the full ensemble is…" [a] better, [b] worse, or [c] about the same.

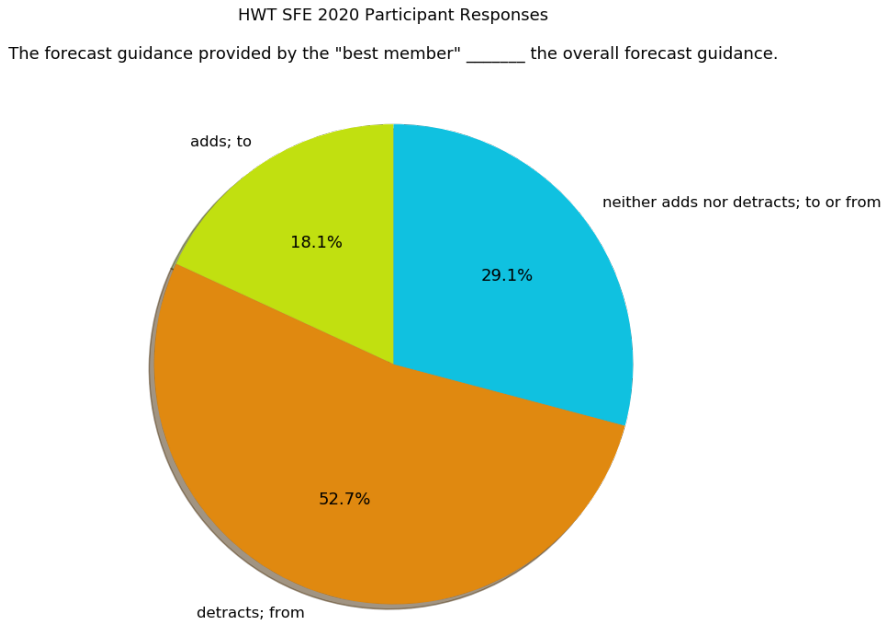


Figure 15 Breakdown of participant responses to the survey question: "The forecast guidance provided by the 'best member' _____ the overall forecast guidance" with choices including [a] adds; to, [b] detracts; from, or [c] neither adds nor detracts; to or from.

A5) CLUE ENSEMBLE HAIL GUIDANCE

During the 2020 SFE, participants evaluated multiple methods for verifying hail forecasts. Hail forecasting – much like forecasting of all convective hazards – requires forecasting not only hazard occurrence, spatial location, and timing, but also hazard intensity (in this case, hail size). Hail forecast verification can choose to focus on each aspect individually, attempt to synthesize them all, or fit somewhere in the middle. For this experiment, the focus was on learning which of those aspects of a forecast HWT participants found most critical to skill when determining a "good" hail forecast.

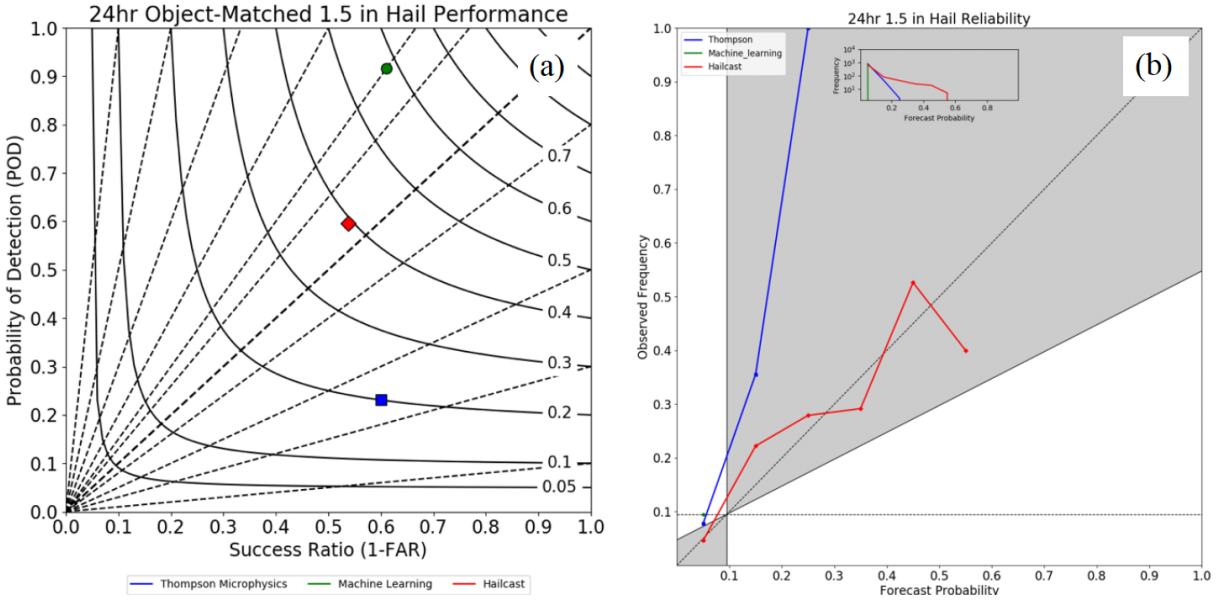


Figure 16 Verification of 1.5" hail forecasts from each of the three models, Thompson, Machine Learning, and CAM HAILCAST, from the week of 24-28 May 2020. (a) Performance diagram validating peak sizes from identified and matched forecast and observed objects. (b) Reliability diagram validating probabilistic ensemble forecasts of 1.5" hail within a 40-km radius of each grid point.

Participants were presented daily figures of ensemble maximum hail size, overlaid with probabilities of 1" or 2" hail, from three different hail forecasting methods: the CAM HAILCAST hail model (Adams-Selin and Ziegler 2016; Adams-Selin et al. 2019), the Thompson hail forecasting method that extracts sizes directly from the microphysical parameterization, and a machine learning method (Gagne et al. 2017; Burke et al. 2020). The ensemble used for these methods was the High-Resolution Rapid Refresh Ensemble, or HRRRE. Multi-Radar Multi-Sensor Maximum Estimated Size of Hail (MRMS MESH; Smith et al. 2016) estimates of 1" and 2" hail occurrence was available as an overlay and validation source. At the end of the week, participants were presented with six diagrams evaluating different aspects of each model's week of hail forecasts. An example of two of these diagrams is pictured in Figure 16. Each diagram used different verification methods as well as varying spatial and/or temporal scales over which the verification was performed.

Half of the diagrams presented results from new updates added to MODE (Method for Object-based Diagnostic Evaluation) within METplus that identifies forecasted and MRMS MESH hail swaths,

matches them based on swath distances and internal characteristics, and then compares peak hail sizes within swath matches (Fig. 16a). For example, if the peak sizes in both swaths are larger than 1.5", that is considered a "hit" in a typical confusion matrix resulting in a Probability of Detection (POD) of 1 and a False Alarm Rate (FAR) of 0. That result would be plotted as a perfect forecast in the upper right corner of a performance diagram (Roebber 2009). Only matched swaths were evaluated. In other words, this verification method focused on the forecast of hail size, not its location, in order to separate the skill of the underlying HRRRE in producing convection with the skill of the hail models in specifically forecasting hail size.

Naturally, spatial location is also likely to be of concern when presented with a hail forecast. To verify those aspects, reliability diagrams (Fig. 16b) were created to verify the location and timing of the 1.5" hail forecast via a grid-based method. The ensemble probability of occurrence of 1.5" hail within a 40 km radius of each model grid point was calculated, smoothed with a Gaussian smoother, and then compared to the gridded MRMS MESH. The reliability diagrams plot the observed frequency against the forecast probability of 1.5" in hail; a perfectly reliable forecast would follow the diagonal. Thus, while hail size is a factor in this verification method, the location of the convection is also included.

Finally, to verify model forecasts temporally, verification was performed over three different temporal and spatial scales. The first pair of performance and reliability diagrams (also shown in Fig. 16) verified forecast and observed hail sizes that had been aggregated over a 24-h period before being compared. The MODE object-matching configuration was designed to match hail swaths produced by supercell/multicell families or a single MCS; the neighborhood verification was conducted as described above with a 40-km radius. The goal of this configuration was to evaluate forecasts on the same spatial and temporal scale as a convective outlook. The second pair of diagrams verified hail sizes aggregated over a 6-h period with the same object-matching configuration; this configuration goal was verification on the scale of a watch. Finally, the third pair of diagrams verified hail sizes aggregated over a 1-h period, used an object-matching configuration designed to match single storm hail swaths, with a 10-km neighborhood radius. The goal of this configuration was warning-scale verification.

In order to evaluate participants' opinions on the different verification methods, they were asked the following questions at the end of each week: (1) What do you mean when you say a 1.5" hail forecast is "good"? Do you think any of these figures successful capture your opinion of the skill of the two different 1.5" hail forecasting methods over the course of the week? Why or why not? (2) Do you think validating hail forecasts over different time/spatial scales is helpful? How effective at capturing hail forecast performance over the different time/spatial scales do you feel the three pairs of figures are?

For Question 1, participants expressed a range of opinions about the contents of a "good" hail forecast, as shown below in Figure 17. The total number of responses received over the course of the experiment to the first question was 44, although it should be noted that not all responses answered each of the specific questions above. An emphasis on "correct location" was noted most frequently – in 30 of 44 responses (Fig. 17a). Half as many responses (15) noted the importance of size, and only 6 responses found timing to be important. Of the participants concerned with hail size, several noted they would consider a hail size forecast of within 0.5" of the observed reports as "good".

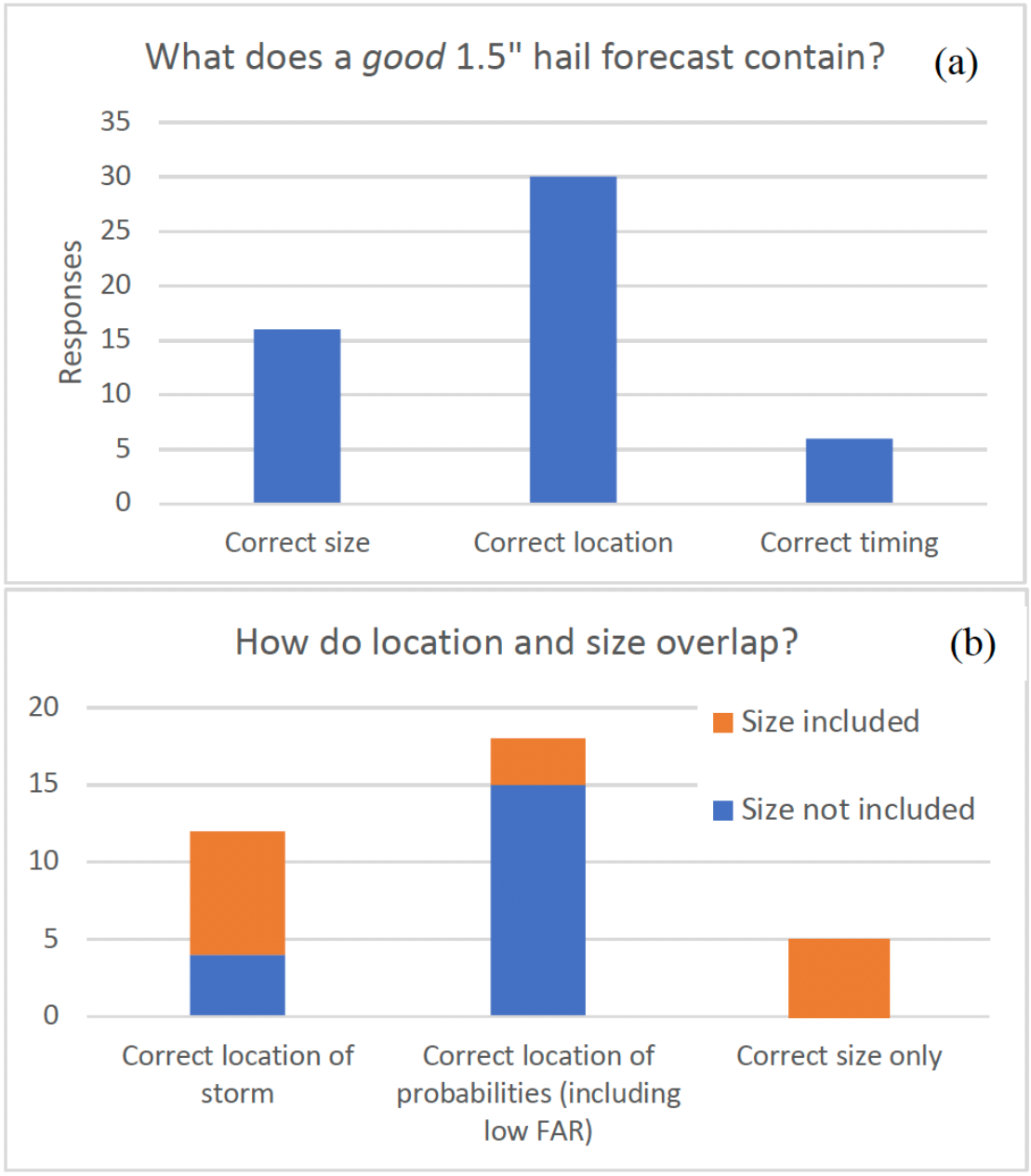


Figure 17 (a) Table showing number of responses to question (1) within each category. Total number of responses was 44. Participants could provide multiple categories listed here within one response. (b) Specific breakdown of responses within the "correct size" and "correct location" categories in (a). In the first (second) column all answers were concerned with correct location of the forecasted storm (probabilities). The blue portion of the columns represent responses that did not include hail size as important; the orange portion did include hail size as important.

The correct size and location responses were further analyzed for overlap among responses (Fig. 17b). "Correct location" could be divided into two groups of emphasis: correct forecast of individual storm location, and correct forecast placement of the ensemble probabilities of 1.5" hail (Fig. 17b). Responses focusing on individual storm location often also provided what they considered to be a reasonable spatial

error threshold: for example, “within 2 or 3 counties” or “within 25-50 miles”, although it was noted that forecasting within urban areas requires more precision. Distinguishing hail-producing ability among multiple CAM-forecasted convective cells was also desired. Responses concerned with accuracy of storm location were mostly also concerned with accuracy of forecasted hail size (8 of 12 responses).

Conversely, responses focusing on the ensemble probabilities of 1.5” hail wanted to see a high POD with a small area of FAR, with a goal of focusing attention on the regions with the highest probability of hail. This group of responses was largely concerned with model-predicted regions of high forecasted probability of hail, on the order of 100-200 km, in which hail did not occur. Only 3 of 18 “correct location of probability” responses also mentioned accuracy of size in their response.

Not all responses mentioned the performance diagram and reliability figures or conveyed the participants’ opinions on their usefulness; additionally, during the first week the figures were created after the scheduled response period. However, of the responses that mentioned the figures (13) all stated they found the figures helpful, and by the last week of the experiment participants were expressing their opinion of a good hail forecast solely in the context of the figures! Several (5) participants found the performance diagrams conveyed skill more clearly, mentioning ease at determining over- and under-forecasting; a few requested displays of additional size thresholds. The responses noted a “lack of signal” from the reliability diagrams: one partial explanation for this response was that the gridded machine learning probabilities were not available for display in the reliability diagrams throughout the SFE. Interestingly, the responses favoring the performance diagrams were not limited to those who considered either correct storm location or correct probabilities more important; participants with different ideas of what constituted a “good” hail forecast nevertheless found the performance diagrams helpful.

The results from Question 2 were overwhelmingly in favor of verification statistics calculated over a range of spatial and temporal resolutions with no responses opposed. Participants liked having verification conducted over 24-h time periods to understand the full storm system as an event, as well as periods smaller than 24 h to understand the model’s effectiveness at forecasting the evolution of the storm system. Many responses (8) suggested 4 h as a preferred resolution as opposed to the 6 and 1 h shown here; a few commented that expecting accuracy on a 1-h timescale is too unrealistic for 24-36 h forecasts.

All responses that expressed an opinion on the figures (23, save 1 concerned with sample size) found them helpful for understanding model performance over the different spatial and temporal scales. Again, a few respondents (4) expressed preference for the performance diagrams citing faster interpretation; none expressed preference for the reliability diagrams.

The verification statistics discussed above were calculated for the full extent of the SFE and are displayed in Figure 18. The performance diagrams (Figs. 18a-c), show the machine learning model with largest Critical Success Index (CSI) but with a high size bias. (Note that because only matched objects are being evaluated in these diagrams, a high bias is indicative of a high size bias, not an overforecast in occurrence.) CAM-HAILCAST had the next largest CSI, with a bias closer to 1. Finally, the Thompson method showed a lower CSI, again with a bias closer to 1. Interestingly, the machine learning model showed an improvement in bias as the temporal interval shrank to 1 h (Fig. 18c) and the Thompson model improved in CSI. These results suggest that the machine learning model struggled with total event hail sizes but more successfully captured the temporal evolution of the storm systems. Conversely, CAM HAILCAST more successfully captured total event hail sizes but struggled with the temporal evolution.

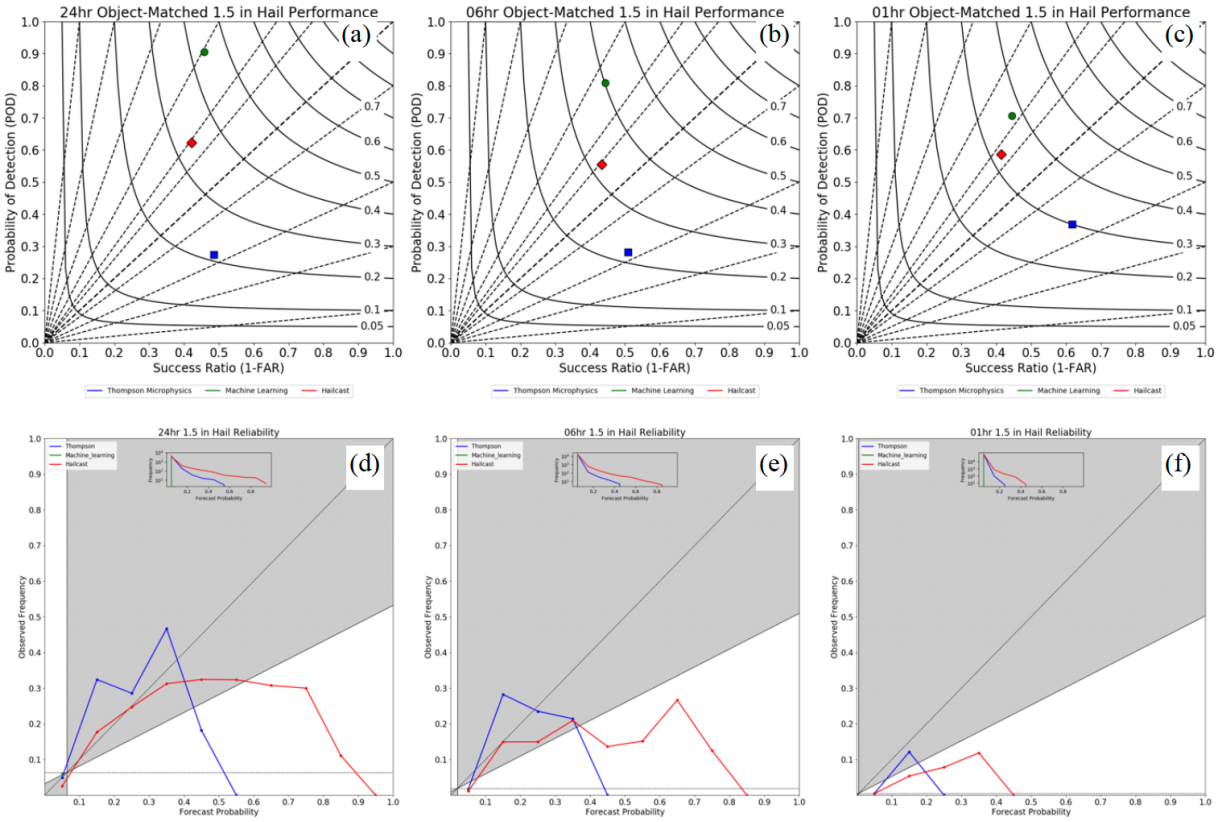


Figure 18 (a-c) Performance diagrams displaying skill of forecast hail swath objects aggregated over (a) 24-, (b) 6-, and (c) 1-h intervals as described above. (d-f) Reliability diagrams for gridded forecast hail probabilities aggregated over (d) 24-, (d) 6-, and (f) 1-h intervals and neighborhoods of (d) 40-, (e) 40-, and (f) 10 km.

During the 2020 SFE, participants were asked to evaluate not only hail forecasts, but also the methods for verifying the hail forecasts themselves. Participants found location, size, and timing all important constituents of a "good" hail forecast, in that order. Location, however, could refer to specific storm location, in which case hail size was also an important part of the forecast, or the location of the ensemble probabilities, in which case hail size was not as frequently included. These results suggest that participants were considering two different forecast use cases when responding about a "good" hail forecast: a probabilistic "convective outlook" style forecast, and a deterministic convective cell-based forecast. The cell-based forecast use case appeared warning-like in its desire for size accuracy (e.g., hail size within 0.5"), but was more forgiving in spatial accuracy (e.g., 25-50 miles or 2-3 counties). These two different theoretical use cases individually developed by participants point to the need for verification to take place over different spatial and temporal scales. To that end, participants were particularly excited about additional verification by varying spatial and temporal resolution. All participants responding about the figures found them helpful, with partiality shown to the performance diagrams due to ease in interpretation. By the end of the experiment some of the participants even used the figures to discuss the success of the hail forecasts being evaluated.

A6) CLUE FV3-SAR PHYSICS, DATA ASSIMILATION, & VERTICAL LEVELS

With the adoption of the FV3 as the dynamical core for the Unified Forecast System (UFS) framework, the SFE has been involved in testing FV3-based CAMs to determine the optimal configuration of an FV3-based CAM for forecasting severe convective storms, as well as testing the sensitivity at CAM scales of the FV3 core to different parameter adjustments. The A6 and A7 comparisons tested the impact of multiple factors on forecasts from regional FV3-based CAMs provided by GSL, NSSL, and EMC. The first comparison focuses on the impact of advanced physics packages, data assimilation, and vertical levels using models provided by NSSL and EMC. Pairs of models could be compared, including configurations that were identical except for the physics packages used (FV3-EMC SARX used more advanced microphysics and PBL parameterizations then FV3-EMC SAR), configurations that differed only in the number of vertical levels (FV3-EMC SARX had 50 vertical levels, while the FV3-NSSL SAR had 80 vertical levels), and a pair that differed in the addition of hourly DA over the 6-h prior to forecast launch (FV3-EMC SAR DA compared to FV3-EMC SAR).

Unlike in previous years, participants were asked about the reflectivity and 2-5 km UH fields at three specific times during the convective life cycle: 1800, 2300, and 0400 UTC. At all times, but particularly later in the convective cycle, participants preferred the EMC FV3-SARX to the EMC FV3-SAR, which has more advanced physics packages (Fig. 19a). Mean and median values were higher in the SARX at 2300 UTC and 0400 UTC. The EMC FV3-SARX has the same physics as the NSSL FV3-SAR, but the NSSL FV3-SAR has more vertical levels than the EMC FV3-SARX. These increased vertical levels did not have much of an impact on the subjective ratings of simulated composite reflectivity and UH, as the distributions were very similar between the two models (Fig. 19b). However, a preliminary look at sounding structure in the two models suggests that the increased vertical levels in the NSSL FV3-SAR may better depict the vertical structure of the atmosphere, including aspects such as inversion structure. Further sounding analysis is planned for future SFEs. Data assimilation appeared to have some impact on the forecasts, with lower scores for the EMC FV3-SAR DA reflectivity and UH at 1800 UTC and 2300 UTC (Fig. 19c). At 0400 UTC, the EMC FV3-SAR DA generally scored higher than the EMC FV3-SAR. However, due to the limited availability of the EMC FV3-SAR DA and subsequently smaller case sample size, these results may not be as certain as the other comparisons herein. In their comments, participants noted higher reflectivity and more convective coverage in the NSSL FV3-SAR and the EMC FV3-SARX, which was mentioned as both a positive (capturing intensity of storms better) and a negative (too many intense storms with high simulated reflectivity values when storms in reality were weaker and not widespread). Participants also noted that storms in the EMC FV3-SAR were overly smoothed, and often commented that it did not do as well as the other models in this comparison.

When asked overall which model best depicted the convective evolution of the day, the NSSL FV3-SAR and EMC FV3-SARX were the most frequently chosen options (Fig. 20), indicating that the advanced physics parameterizations resulted in better forecasts of severe convection. In a few cases, participant comments indicated that the NSSL FV3-SAR and the EMC FV3-SARX tended toward too discrete and too cellular storm modes compared to observations.

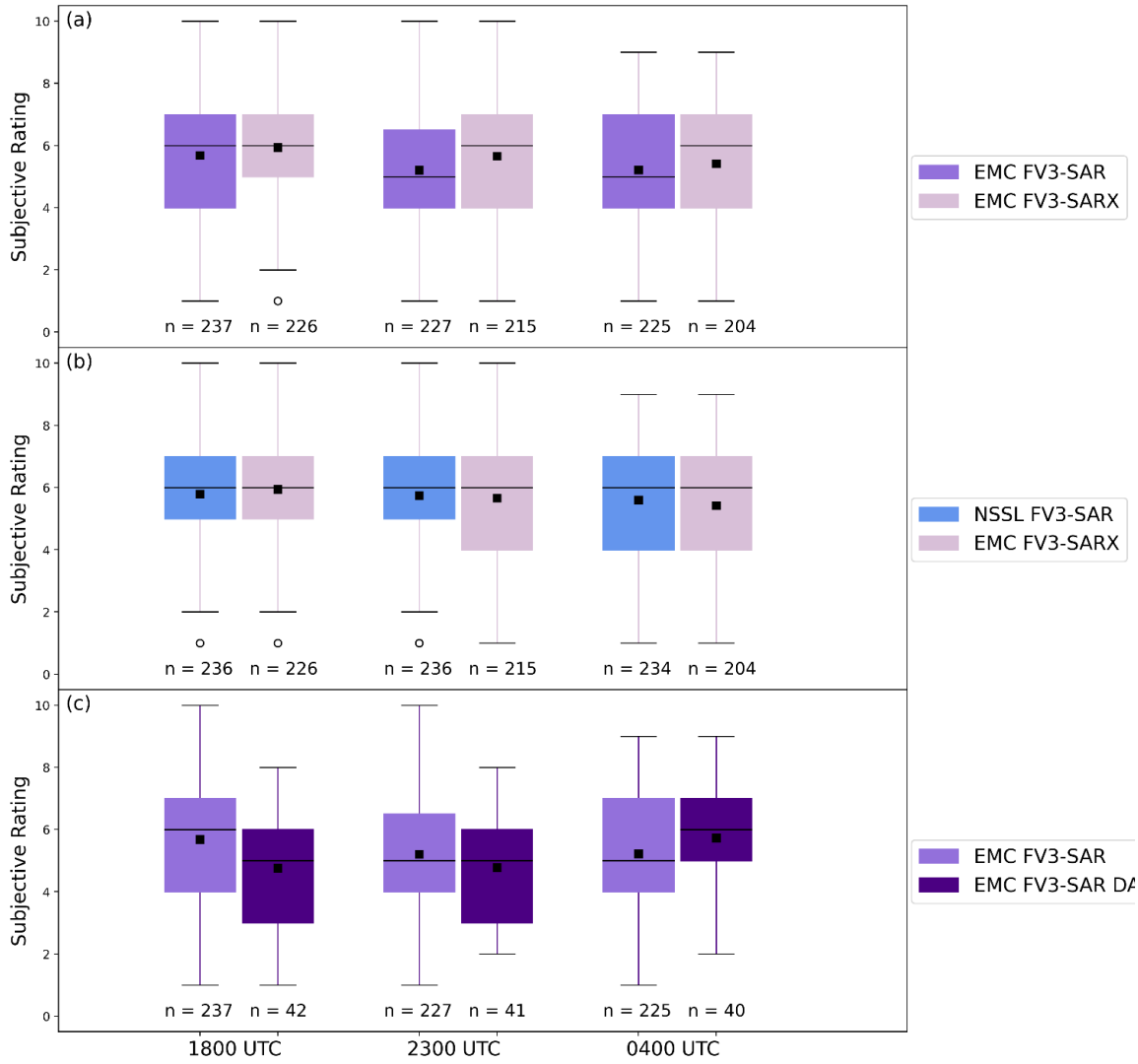


Figure 19 Participant ratings of the composite reflectivity and UH at three different times comparing models with (a) different physics parameterizations, (b) different numbers of vertical levels, and (c) with and without a data assimilation cycle.

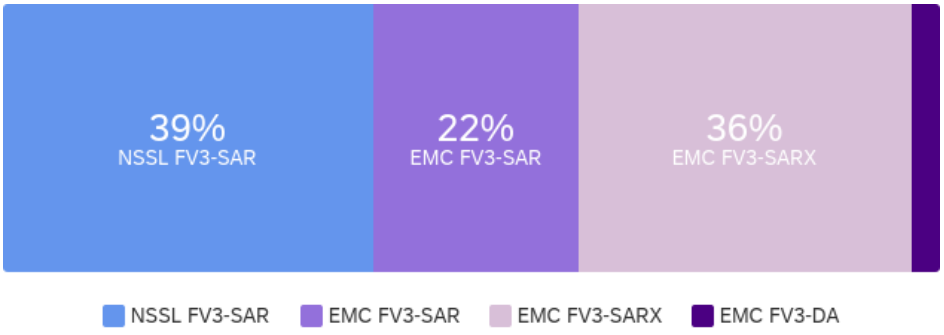


Figure 20 Participant responses to the question: "Which deterministic CAM(s) best captured the convective evolution (i.e., timing, CI, convective mode) through the entire forecast run?"

Environmental variables (2-m temperature, 2-m dewpoint, and SBCAPE) were evaluated at two times: 1800 and 0000 UTC. For environmental variables, the more advanced physics schemes seem to have less of an impact on the subjective skill of the forecasts, with the EMC FV3-SAR and the EMC FV3-SARX having very similar means (Fig. 21a). However, the EMC FV3-SARX does have higher 25th and 75th percentile values at 0000 UTC, perhaps indicating the influence of convection. As in the reflectivity and UH evaluation, increased vertical levels do not have much of an influence on the 2-m temperature, 2-m dew point, or the SBCAPE (Fig. 21b), though the NSSL FV3-SAR does have a slightly higher mean than the EMC FV3-SARX at both 1800 UTC and 0000 UTC. Relative to the reflectivity and UH results, DA appears to either have little effect (at 1800 UTC) or provide a slight improvement (at 0000 UTC) in the environmental variables (Fig. 21c). The sample size issue discussed previously regarding the DA comparison applies here as well. Participant comments frequently noted biases in all of the models, with cool temperatures, low instability, and high moisture biases being mentioned for all available models. Participants also noticed that on occasion, the 2-m temperature and 2-m dewpoint were not too far off, but the SBCAPE differed greatly from observations, leading to the conclusion that differences aloft were having a big impact. When differences were mentioned between the models, they were similar to the following comment: "*EMC SARX and NSSL SAR appear to do better with environmental conditions (not storm altered), while EMC SAR does better with temperature and dewpoint inside of cold pools (especially with the original MCS.)*"

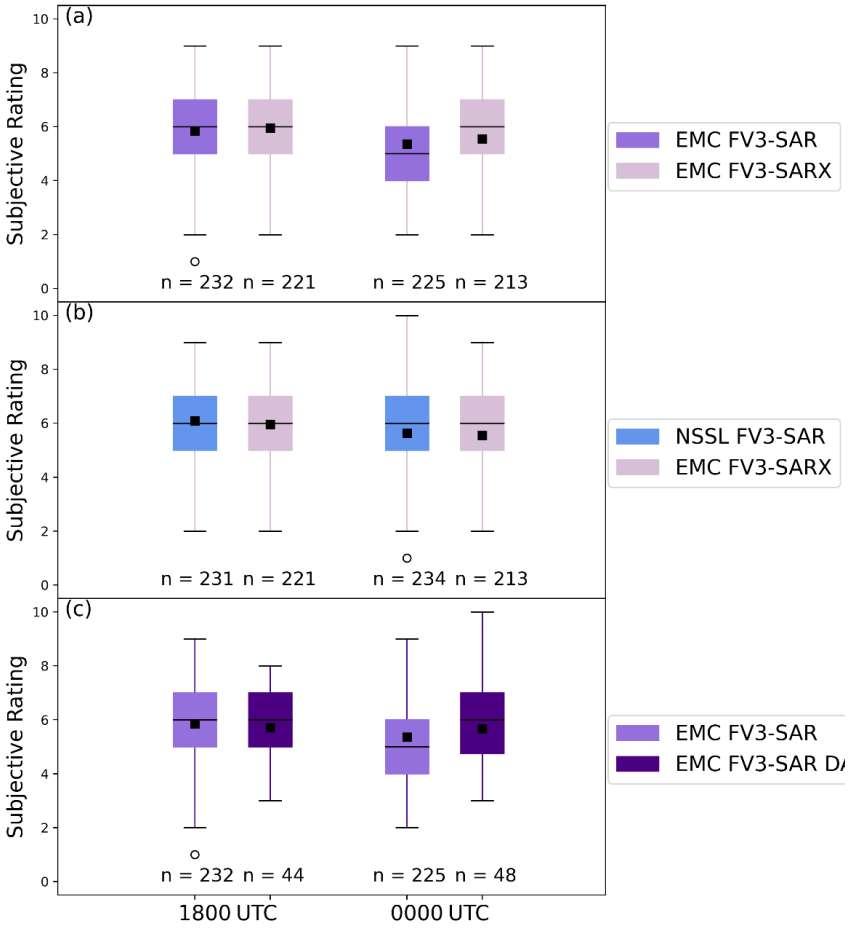


Figure 21 As in Figure 19, except for 2-m temperature, 2-m dewpoint, and SBCAPE

A7) CLUE FV3-SAR IC, HORIZONTAL ADVECTION SCHEME, & LAND SURFACE MODEL

The second comparison to look at the effects of different configuration details on a regional FV3-based CAM tested the impacts of initial conditions, diffusivity settings, and the land-surface model (LSM) using models provided by EMC and GSL. These models were configured such that two pairs of models identical but for the diffusivity settings and two pairs of models identical but for the ICs used could be compared, as well as one pair of models that used different LSMs (Fig. 22). As in prior comparisons, participants were asked to evaluate the composite reflectivity and UH at three times, and the 2-m temperature, 2-m dewpoint, and SBCAPE at two times during the forecast. Since one of the models involved in this comparison, the EMC FV3-SARX, was repeated from A6, participants were reminded of their ratings of the EMC FV3-SARX when rating the remaining four models.

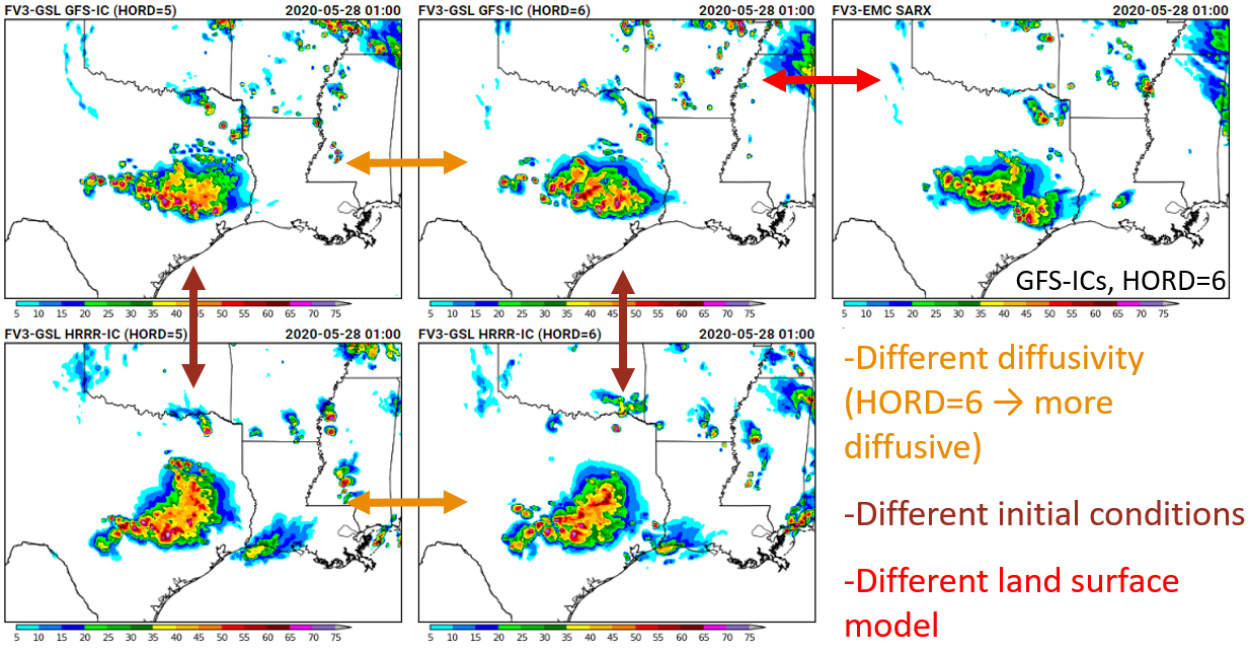


Figure 22 An annotated example of the panels participants evaluated. Colored arrows indicate the differences between pairs of models. Annotations cover a sixth panel, which showed radar observations during participant evaluations.

Differences between composite reflectivity and UH forecasts were relatively minimal (Fig. 23), with the exception of the EMC FV3-SARX generally performing the best. The 0400 UTC ratings differed the most between forecasts. The HORD=5 option scored higher than the HORD=6 option, indicating that less diffusivity scored higher for both sets of ICs. For both diffusivity settings, the GFS ICs scored higher than the HRRR ICs, with the difference in mean score being approximately as large as the difference between the mean scores of different diffusivity options. Therefore, from these results, it seems that the diffusivity option and the ICs selected have about the same impact on the subsequent reflectivity and UH forecasts. In terms of the LSM comparison (the EMC FV3-SARX and the GSL FV3-SAR with GFS-ICs and HORD=6), the NOAH LSM in the EMC FV3-SARX performed better than the RUC in the GSL FV3-SAR versions.

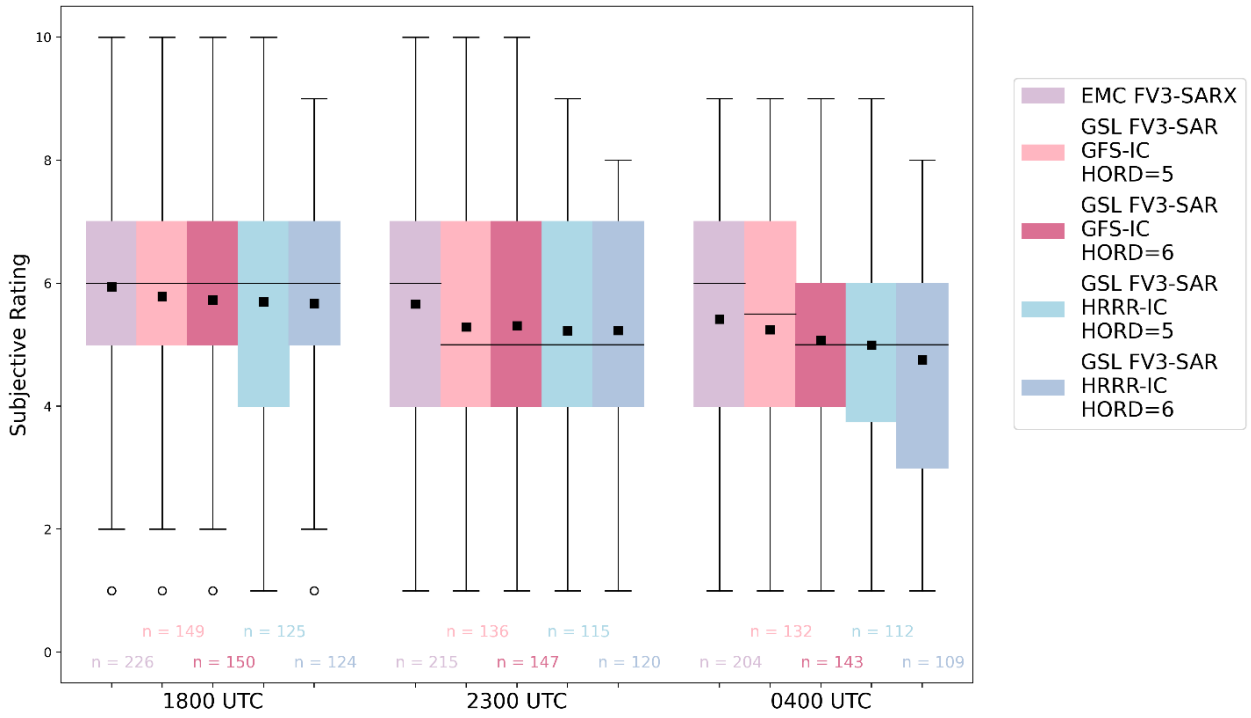


Figure 23 Participant ratings of the composite reflectivity and UH at three different times comparing models with different diffusivity settings, different initial conditions, and different LSMs.

When asked which model performed the best in terms of overall convective evolution, interestingly the GSL FV3-SAR with HORD=5 and GFS-ICs was selected most frequently, despite its lower ratings compared to the EMC FV3-SARX (Fig. 24). Comments suggest that the participants may have been focusing mainly on the GSL models for this comparison, which may in part be due to the labels clearly describing the differences between these configurations. Data availability issues also frequently were mentioned in the comments. Initial conditions seemed to be where participants noted the largest differences, but in the discussions the large differences were not always favoring the HRRR-ICs or the GFS-ICs in particular.

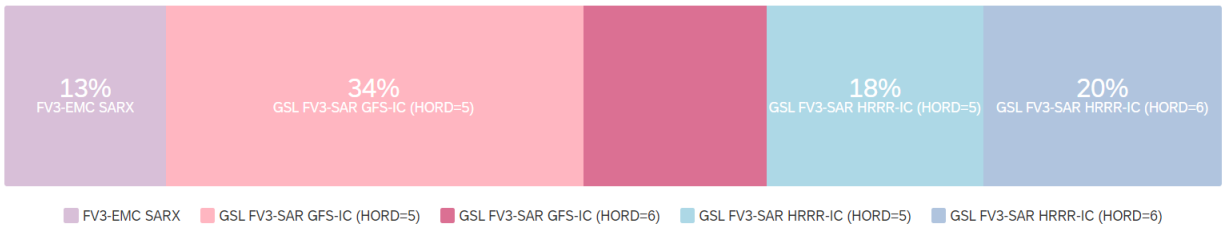


Figure 24 Participant responses to the question: "Which deterministic CAM(s) best captured the convective evolution (i.e., timing, CI, convective mode) through the entire forecast run?"

For environmental variables, larger differences were seen at 0000 UTC compared to 1800 UTC (Fig. 25), perhaps due to the effects of convection. In this case, the HRRR-ICs performed better than the

GFS-ICs, opposite to what was found for the composite reflectivity and UH (Fig. 23). However, the means were very close, so this difference is likely insignificant. Similarly, there was almost no difference in the ratings for the different diffusivity settings. Once again, the NOAH LSM in the EMC FV3-SARX outperformed the RUC LSM in the GSL FV3-SAR with GFS-ICs and HORD=6. Participant comments support the HRRR-ICs scoring slightly higher; a few comments point to the ability of the HRRR-ICs to help negate the known cool and moist bias of the FV3 core.

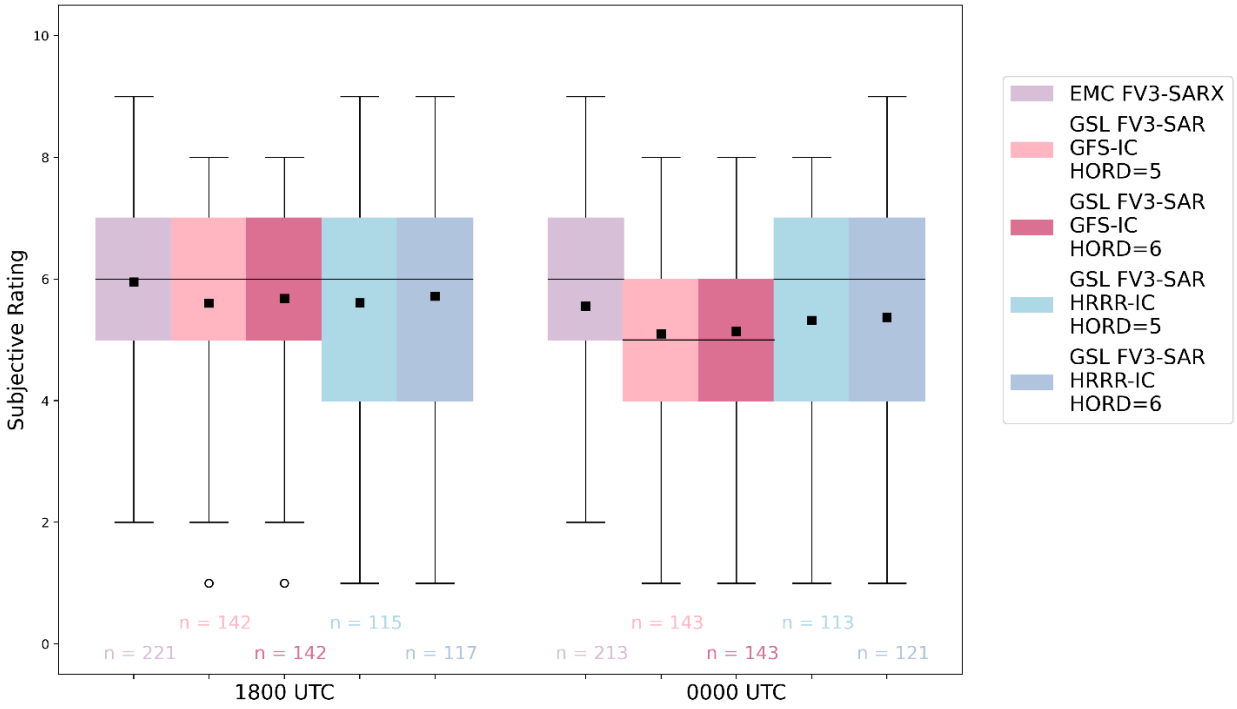


Figure 25 As in Figure 23, but for 2-m temperature, 2-m dewpoint, and SBCAPE.

Overall, the summary impacts of model configuration choices that impact the performance of FV3-based CAMs from A6 and A7 are, in order:

- Advanced physics suites – large improvement
- NOAH LSM – improvement
- Increased vertical levels – small improvement (mostly at earlier times)
- Less diffusivity – small improvement
- Initial conditions – large impact on subsequent forecasts, but no set of ICs performed consistently better

A8) MESOSCALE ANALYSIS

Two different versions of 3D-RTMA were subjectively evaluated by participants during the 2020 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. One version was provided by GSL and used the GDAS for background error covariance information in the hybrid DA system, and the other version provided by EMC used the HRRRDAS (i.e., convection-allowing data assimilation system) for background error covariance information. The 15-minute output data were examined during the 18-03 UTC period on the next day (Fig. 26). The goal was to assess whether information from a CAM ensemble (i.e., HRRRDAS) can improve the analysis for short-term weather forecasting applications. Overall, both versions of the 3D-RTMA were subjectively similar to one another, with participants most commonly rating them "about the same" (Fig. 27). Not surprisingly, the largest differences were often in or around areas of convection and/or in areas with limited surface observations. In terms of overall performance for situational awareness in short-term convective forecasting, both systems performed well. The primary issues/artifacts noted in the comments from participants were 1) local maxima/"bullseyes" at some locations/times in different surface-based fields and 2) a discontinuity in the temporal evolution of the analysis fields going from the 15-minute updates at 15, 30, and 45 minutes past to the top-of-the-hour analysis (i.e., drift and reset).

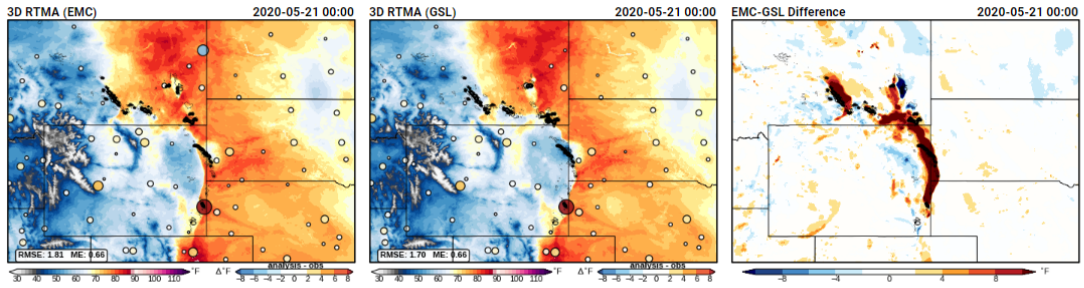


Figure 26 Website comparison example for the 3D-RTMA. The EMC version is in the left panel, the GSL in the middle, and the difference (EMC-GSL) in the right. The 2-m temperature analysis valid at 0000 UTC 21 May 2020 is shaded (left two panels - 40 dBZ reflectivity contours). The difference (analysis-obs) at METAR sites is shown by the size and shading of the dots. Corresponding 2-m temperature difference (shaded) on the right panel.

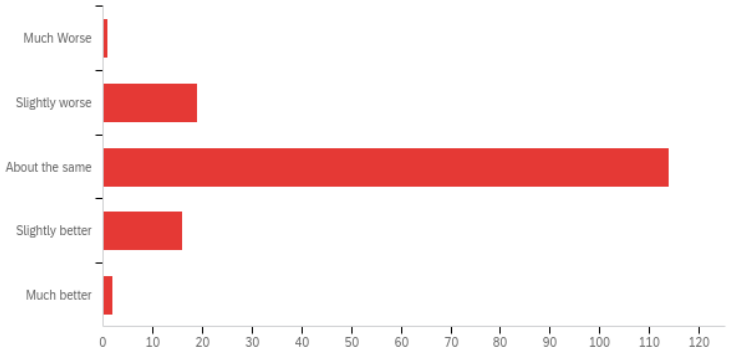


Figure 27 Participant subjective ratings of whether the EMC version of 3D-RTMA was "much better", "slightly better", "about the same", "slightly worse", or "much worse" than the GSL version of 3D-RTMA.

A9) GLM LIGHTNING DATA ASSIMILATION

This experiment focused on the assimilation of recently available total lightning data from the Geostationary Lightning Mapper (GLM) aboard the Geostationary Operational Environmental Satellites (GOES) 16 (Goodman et al. 2013). The assimilation exercise employed the GLM "flashes", a variable more closely related to flash initiation locations and, thus, provides an estimated measure of flash origin densities (per unit time) when tallied and projected onto the model grid. Additional GLM variables related to the horizontal extent of flashes such as "events", will be considered in follow-on work in addition to GLM data from GOES-17, from which lightning coverage was shown to be superior to GOES-16 in some portions of the western US, particularly the northwest. One of the advantages of the GLM lies in its ability to detect the presence of lightning-active (mixed-phase) convection over vast geographical areas with limited observations from ground-based platforms (e.g., radars, METARS, soundings). Thus, for this evaluation, the DA exercise were focused over areas known to suffer from paucity in radar coverage, such as the mountainous west and oceanic regions within the CLUE domain. To gauge the potential benefit of GLM lightning DA over routinely available level II 88D radar data (reflectivity factor + Doppler radial winds), this evaluation primarily focused on two DA experiments: the first labeled "GLM" assimilating both GLM and level II radar data and a second labeled "noGLM" assimilating only level II radar data (e.g., Fig. 28).

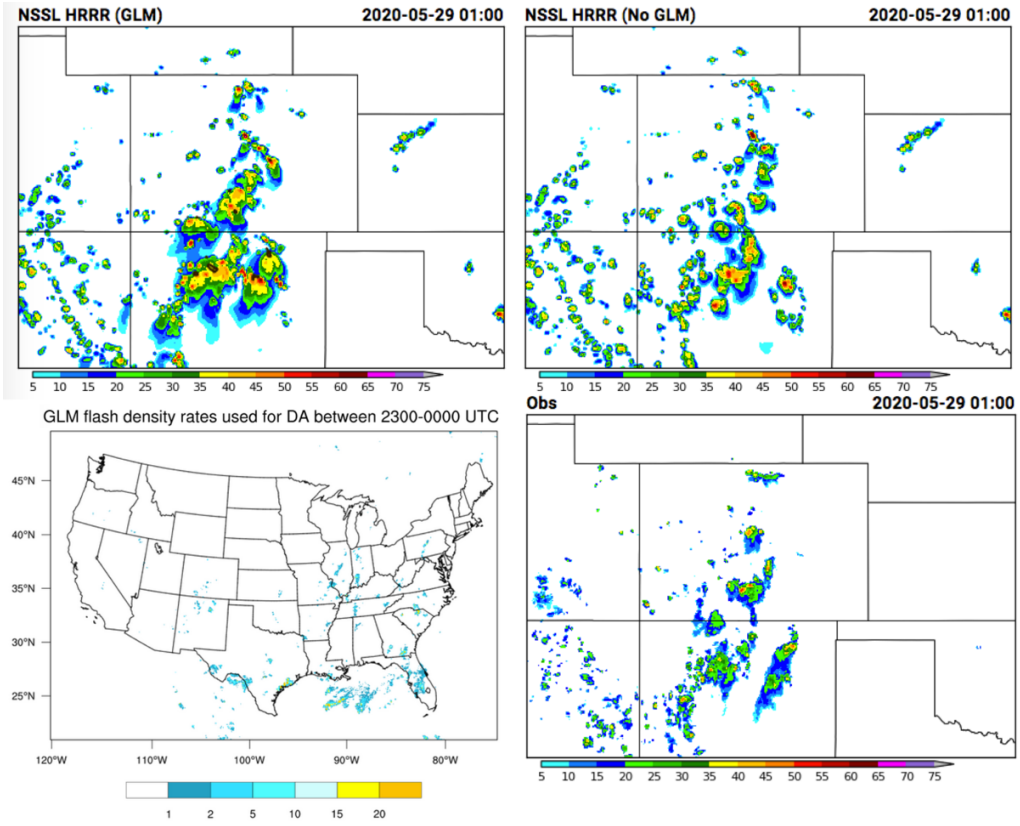


Figure 28 Horizontal cross sections of composite reflectivity fields (dBZ) at 1-h forecast for the GLM (top left), noGLM (top right) DA experiments, 2300 – 0000 UTC accumulated GLM flash density used during the DA (bottom left, shown here on a pseudo GLM 10x10 km^2 grid for better visualization of the contours) and observed composite reflectivity fields from NSSL's MRMS product (bottom right).

For subjective evaluation of the forecasts during the SFE, the experiments were complemented by radar observations. The observed GLM flash densities used during the 1-h assimilation window prior to the forecast (i.e., 3DVAR analysis), were also provided as a guidance to users and forecasters for defining the location of the evaluation domain. If the GLM indicated the presence of lightning over either the mountainous west or over oceanic regions, the primary domain used for this evaluation (labelled domain #3) was placed over these locations. The domains used for the other model evaluation and forecasting activities, which typically included GLM-active areas over the CONUS, were also available for further guidance.

The modeling vehicle used for these experiments was the quasi-operational HRRRv4 code based on WRF-ARW V3.9.1.1 kindly provided by Ming Hu and colleagues at GSL. These SFE simulations mimicked the real time HRRR settings by downscaling RAPv5 input data for the initial conditions and by employing GSL's physics settings and CLUE grid specifications.

The 3DVAR code that was coupled with the HRRRv4 model is NSSL's Experimental Warn-on-Forecast System for 3DVAR (NEWS3DVAR, Wang et al. 2019), which is based on a 3DVAR DA system initially developed for the Advanced Regional Prediction System (ARPS) (Gao et al. 1999; Xue et al. 2001). The lightning DA procedure follows those described in Fierro et al. (2019) wherein modeled water vapor mass mixing ratios are adjusted (increased) towards near saturation values within each column characterized by nonzero GLM flash densities. To curtail imbalances resulting from the mass increase incurred during the 3DVAR analysis, the water vapor adjustments are confined within a 3-km deep layer above cloud base (surrogated by the lifting condensation level). The single deterministic simulations (i.e., no ensembles) used successive 3DVAR analysis cycles at a 15-min frequency between 2300 UTC on the previous day and 0000 UTC (see 1-h accumulated GLM flash density fields in Fig. 28). To further reduce the impact of Qv-induced mass-wind imbalances and potential wet biases in the forecasts, the initially coarser ~10x10 km^2 GLM densities (shown in Fig. 28) were thinned down to that of the grid spacing of the simulation domain, namely 3x3 km^2.

Survey analysis from SFE participants performed between 27 April and 28 May are summarized herein (Fig. 29) for the three following questions labelled Q1, Q2 and Q3:

- Q1: Focusing on the simulated composite reflectivity field, forecasts of the number, location, intensity, and mode of convective storms are _____ between the runs with and without GLM DA.

- Q2: Although observations may be limited in this region to perform a full assessment, please complete the following statement: The short-term forecasts of thunderstorms are _____ when assimilating GLM data.

- Q3: If the forecasts of thunderstorms are different between these DA runs, how long into the forecasts do those differences last?

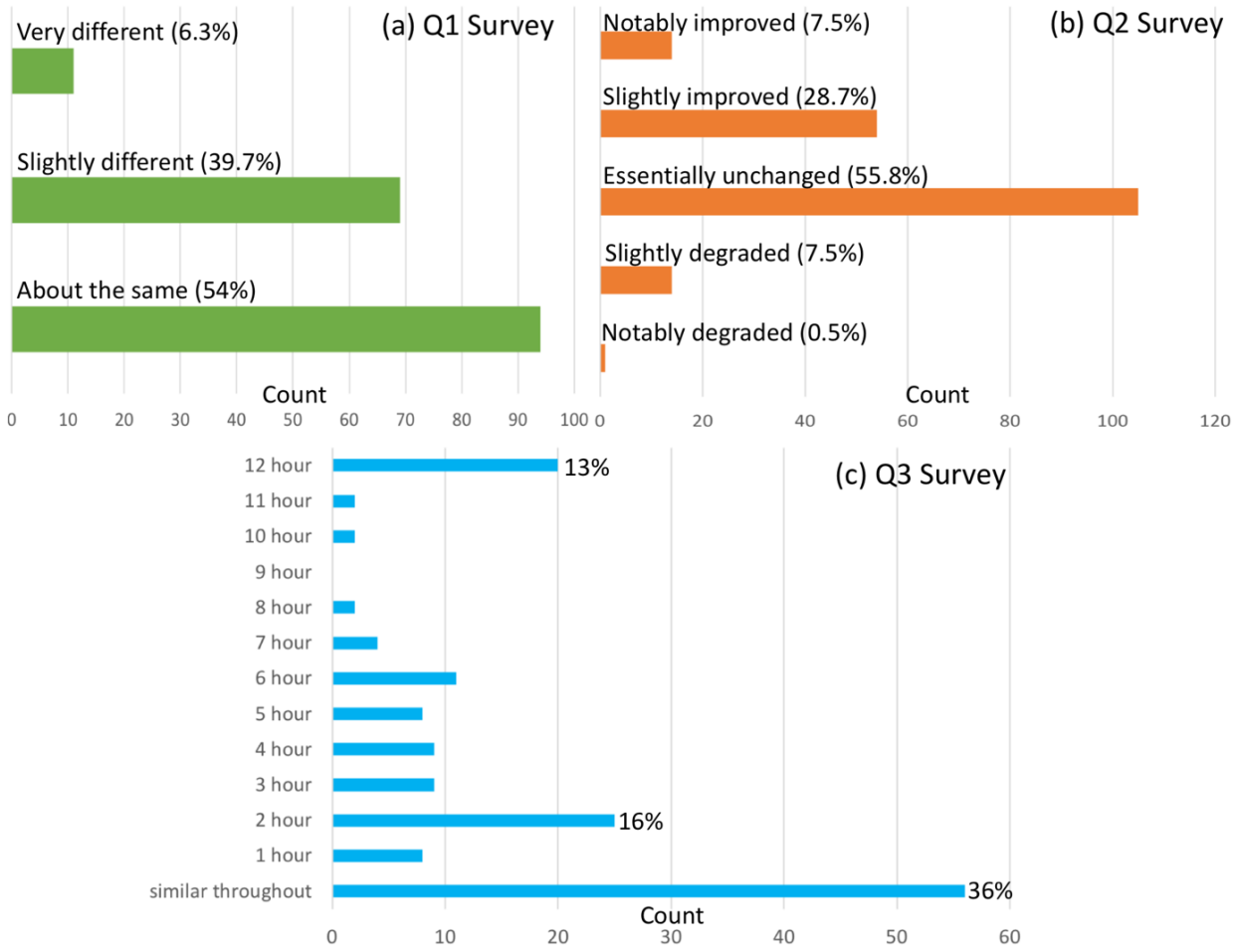


Figure 29 Bar charts and percentiles of the responses given by the SFE participants to questions (a) Q1, (b) Q2 and (c) Q3 for the GLM lightning data assimilation experiment.

Before interpreting the participant's responses in this survey, it is relevant to re-iterate and underline that the main goal of this modeling exercise was to evaluate the degree of added benefit incurred by the assimilation of two-dimensional GLM flash density rates over volumetric level II radar data in areas of the country suffering from poor radar coverage. Given that vast amount of storm-scale kinematic and microphysical information contained in 3D radar sweeps compared to the information provided by 2D flash density fields, improving storm-scale radar DA forecasts is, by design, a challenging task to achieve. For Q1, the survey in Figure 29a reveals that while 54% did not report any noteworthy differences in forecast composite reflectivity fields between the two experiments (i.e., GLM vs noGLM), a sizeable portion (46%) did notice some differences. The survey from Q2 in Figure 29b is intended to reveal if any of the differences in composite reflectivity fields reported in Q1 were either positive or negative with a viewpoint/theme centered on overall forecast skill. Positive improvements were noted 36.2% of the times compared to 8% for negative ones. Not surprisingly, the fraction of participants not reporting any changes in the forecast in the survey for Q2 were about the same as in Q1 (Figs 29a, b). Focusing the participants' attention on the forecast improvements they documented in Q2, a noteworthy fraction indicated that these were generally maintained up to either 2-h (16%) or as far as 12-h forecast (13%).

This result favoring the two near opposite ends of the forecast length spectrum is generally consistent with earlier proof-of-concept work with this GLM DA algorithm: (i) Short-term forecast improvements for most convective regimes were generally lost after 2-3-h forecasts due to inherent biases and errors contained in the initial conditions downscaled from larger-scale models (here the RAPv5 data) and (ii) forecast improvements beyond 3-h were generally achieved for more organized convective systems such as squall lines, QLCSs, MCSs or MCCs, which are of common occurrence during the Spring over the eastern CONUS. Rationales for (ii) is that: 1) MCSs evolution is primarily governed by the location and timing of the incipient cold pool, which the GLM DA is able to capture during the 3DVAR analysis and 2) the large scale conditions favoring organized convective systems such as: stationary boundary, unidirectional shear, large areas of warm, humid, unstable air mass, low level jet etc. are generally well captured by larger-scale models.

To place the subjective evaluations in Fig. 29 into context, preliminary analyses of the SFE runs are briefly described here focusing on: (i) 00-06UTC accumulated precipitation summed over all the 29 forecast days covering the duration of the SFE (Fig. 30) and (ii) Roebber performance diagrams for aggregate contingency statistics (Fig. 31).

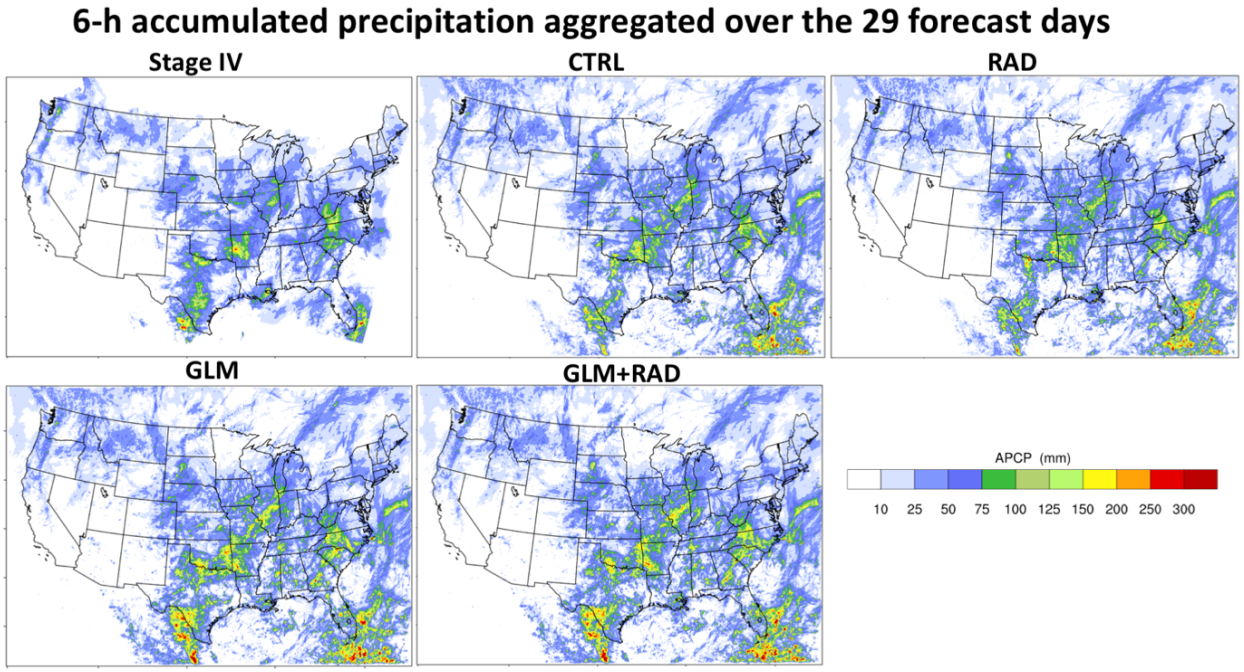


Figure 30 00-06UTC accumulated precipitation aggregated over all the 29 forecast days during the SFE wherein; CTRL=control (no DA), RAD (only level II radar data were assimilated = noGLM experiment on the SFE page), RAD+GLM (both radar and GLM were assimilated = GLM experiment on the SFE page) and GLM (GLM DA only).

When examining Fig. 30, it becomes evident that, despite an overall relatively good performance of the CTRL runs (no DA), the assimilation of GLM data had a more noticeable aggregate impact over the southeastern US (including its oceanic regions) and the eastern portions of the Sierra Madre in Mexico. Aggregate rainfall differences over the mountainous west remain overall comparatively small and, thus,

will be analyzed on a case-to-case basis. Despite an overall broadly consistent topology of rainfall contours in all experiments, Fig. 31 indicates that, in the aggregate, hourly precipitation forecast skill for the GLM-based runs at 1, 3 and 6-h remain overall superior to that of CTRL, illustrating the benefit of assimilation GLM and/or radar data. Consistent with previous work (Fierro et al. 2019, Hu et al. 2020), the best skill was achieved when both datasets were employed during the DA.

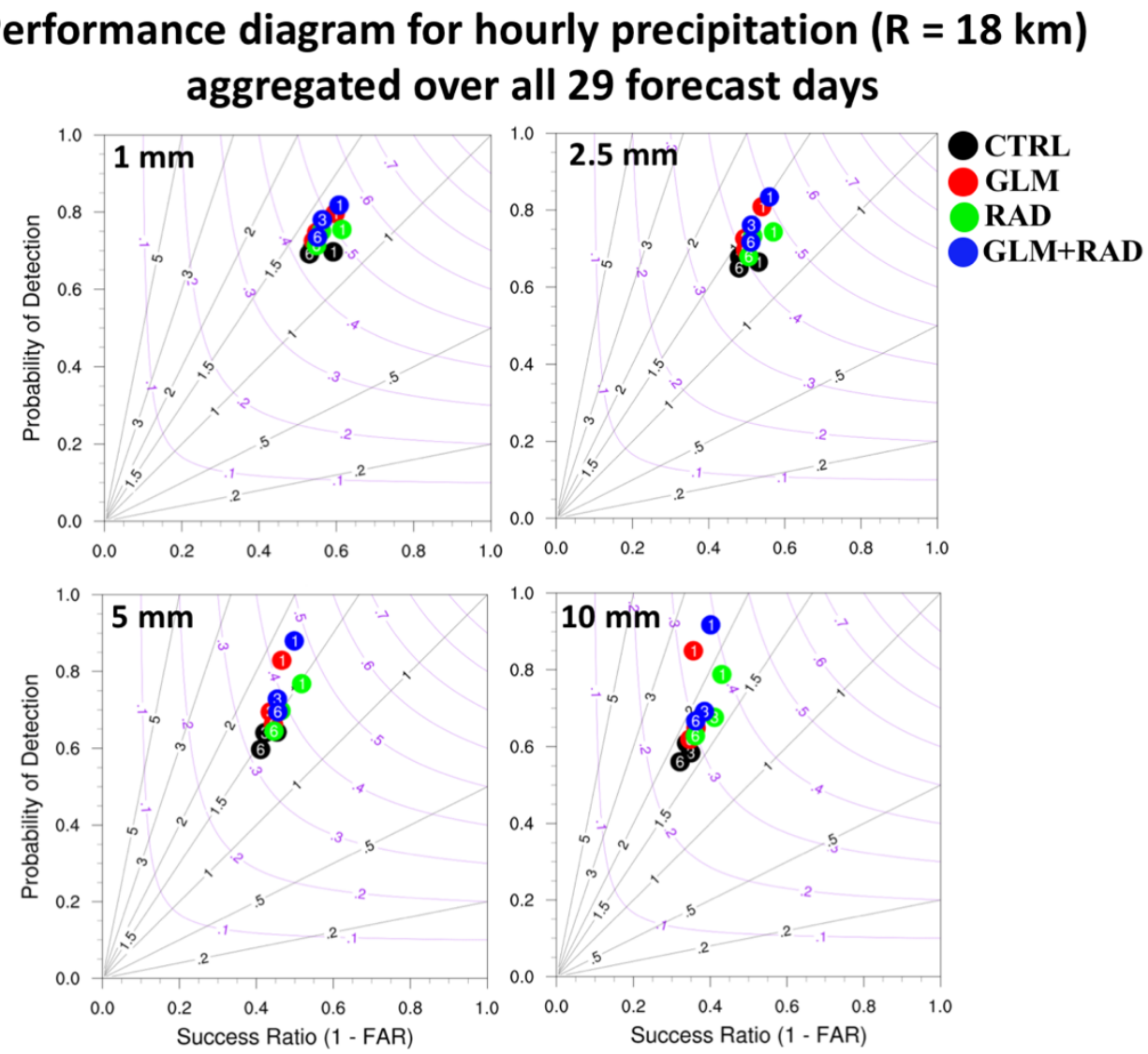


Figure 31 Roebber performance diagrams for selected hourly precipitation thresholds at 1, 3-h and 6-h forecast aggregated over all 29 forecast days.

b) Model Evaluations – Group B

B1) CALIBRATED, MACHINE-LEARNING, & SPC TIMING GUIDANCE

Participants were asked to evaluate a suite of 24-hour and 4-hour guidance forecasts for severe weather hazards including tornadoes, wind (50 kts), and hail (1.0-in). All guidance products were based on forecast output from the operational HREFv2.1. Using storm reports and WFO warnings from the time window coincident with the forecast period, participants scored forecasts on a scale of 1 to 10 (with 10 indicating an excellent forecast).

i) 24 h Tornado Forecast Guidance

Five tornado guidance products were evaluated: HREF/SREF-calibrated ("HREF/SREF Cal", Jirak et al. 2014), two calibrated methods using an STP distribution within a 40-km radius domain defined either over a circle ("STP Circle", Gallo et al. 2018) or the inflow quadrant ("STP Inflow", Jahn et al. 2020), the STP-calibrated inflow method for which regions associated with an MCS are filtered ("STP Inflow MCS Filter"), and a ML model based on a random forest (RF) method ("ML RF", Loken et al. 2020).

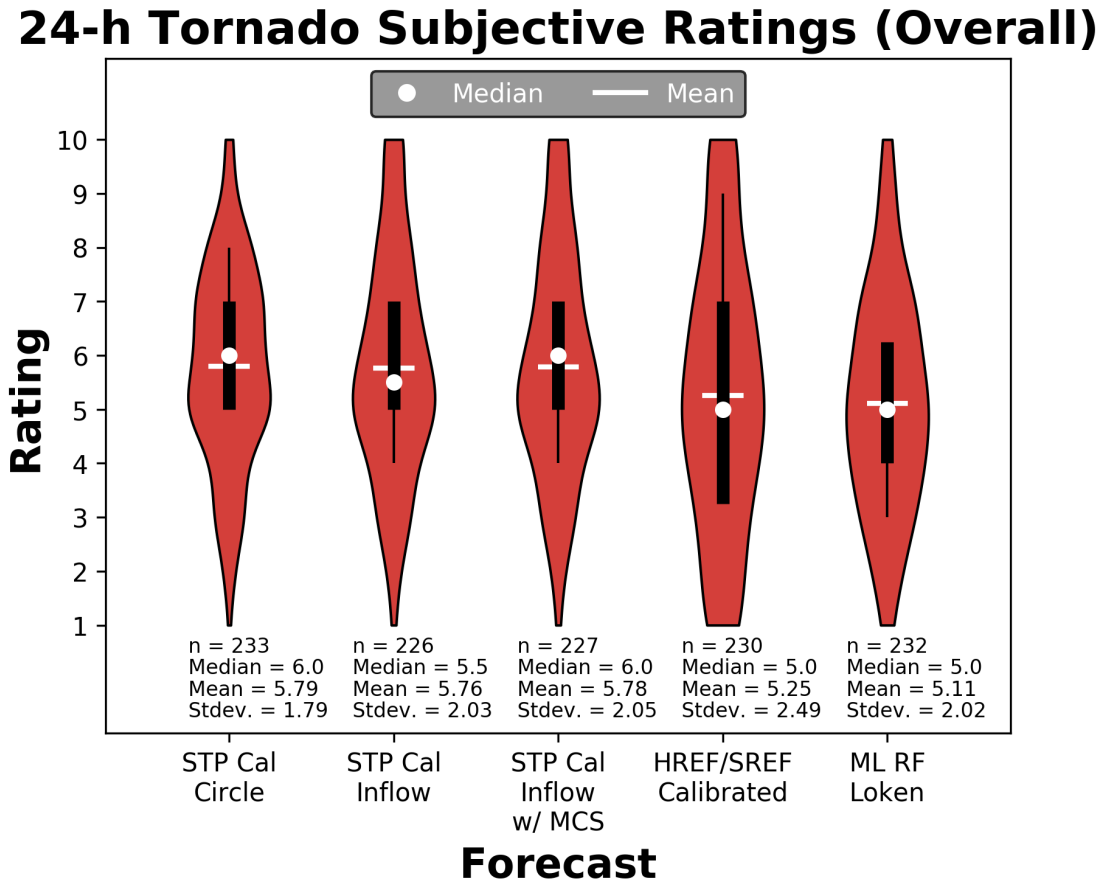


Figure 32 Violin plots showing distributions of subjective ratings for 5 tornado guidance products (see text), median (white dot), mean (white line), interquartile range (wide black vertical line) and 1.5 interquartile range (thin black line).

Subjective evaluations of the five products are summarized in Figure 32 using violin plots, which are based on 227 to 233 responses per product. The mean and median values across all products are between 5 and 6, which suggests that no one product stands out as convincingly superior or inferior. The STP-calibrated products (left 3 plots of Figure 32) all register a mean value of 5.8, which is slightly higher than the HREF/SREF and ML products. A reduced performance by "ML RF" may be due to several cases for which it generated probability contours significantly higher than the other methods and inconsistent with observed tornado frequency. "STP Circle" and "STP Inflow" register near identical means; however, evaluator comments often noted that "STP Circle" produced a larger false alarm area than "STP Inflow". This may be the reason for more responses at 9 and above for the latter than the former. A relatively high standard deviation of 2.49 suggests a lower consistency in the perceived value of the HREF/SREF forecast guidance; certain cases it performed the best and certain cases the worst. The "STP Inflow MCS Filter" were near identical to "STP Inflow" results and provided negligible improvement in forecast guidance.

ii) 4 h Tornado Forecast Guidance

Two of the calibrated methods, "STP Cal Circle" and "HREF/SREF Cal", were used to generate rolling 4-h probabilistic tornado guidance and were evaluated with SPC Timing Guidance products based at 0600 and 1300 UTC (Fig. 33). SPC Timing Guidance products are generated using a temporal disaggregation method using HREF/SREF calibrated guidance as applied to the operationally issued convective outlooks at 0600 and 1300 UTC (Jirak et al. 2012, 2020). Thus, they are a blend of the human forecast and the first-guess guidance. The SPC Timing Guidance products were rated higher (mean scores greater than 5.98 and standard deviations less than 2.43) than the calibrated product based on HREF/SREF alone (mean score of 5.37 and standard deviation of 2.65). The "STP Cal Circle" product standard deviation of 1.93 indicates its performance was the most consistent among the 4 products. Its subjective rating (mean of 5.67) is lower than the SPC Guidance products and greater than HREF/SREF Calibrated.

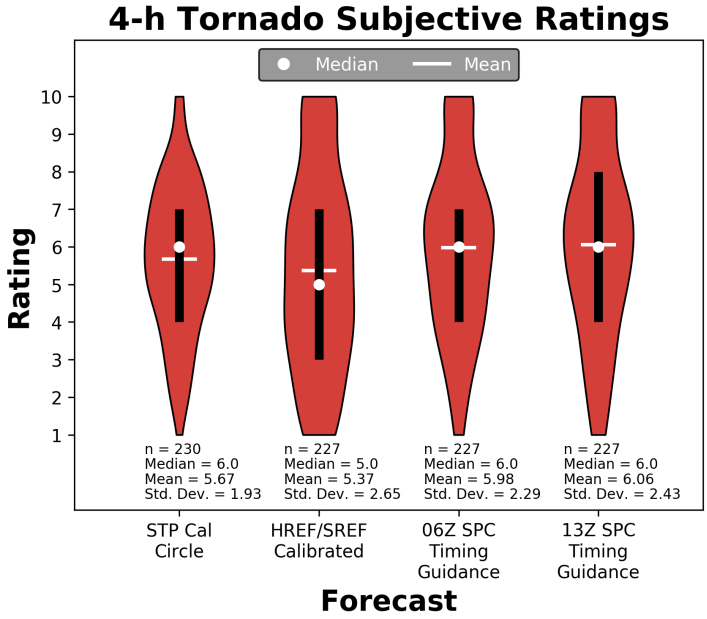


Figure 33 As in Figure 32, but for 4-h tornado guidance products.

iii) 24 h Hail Forecast Guidance

Participants evaluated three methods for producing calibrated 24-h severe hail forecasts. These included two ML RF-based methods ("ML Burke", Burke et al. 2020; and "ML Loken", Loken et al. 2020), and an approach that considers data from both the 0000 UTC HREF and 2100 UTC SREF ("HREF/SREF Calibrated", Jirak et al. 2014).

ML Loken received the highest mean (6.68) and median (7.0) subjective ratings with the lowest standard deviation (1.77) and smallest proportion of low-end (i.e., 1-3) ratings (Fig. 34), suggesting consistently strong performance. Ratings for ML Burke (mean 5.72, median 6.0) and HREF/SREF Calibrated (mean 5.68, median 6.0) tended to be slightly lower than ML Loken but were still favorable.

Participants noted that the three methods could provide substantially different forecasts on a given day. ML Loken tended to produce broader areas of low-end probabilities, frequently giving it a high probability of detection (POD) and making it look most like the practically perfect guidance. Meanwhile, ML Burke tended to give sharper probabilities over smaller regions, which helped forecasters identify areas of greatest threat, but generally resulted in lower POD compared to ML Loken. In general, participants had high praise for ML Loken, and preferred ML Burke to HREF/SREF Calibrated.

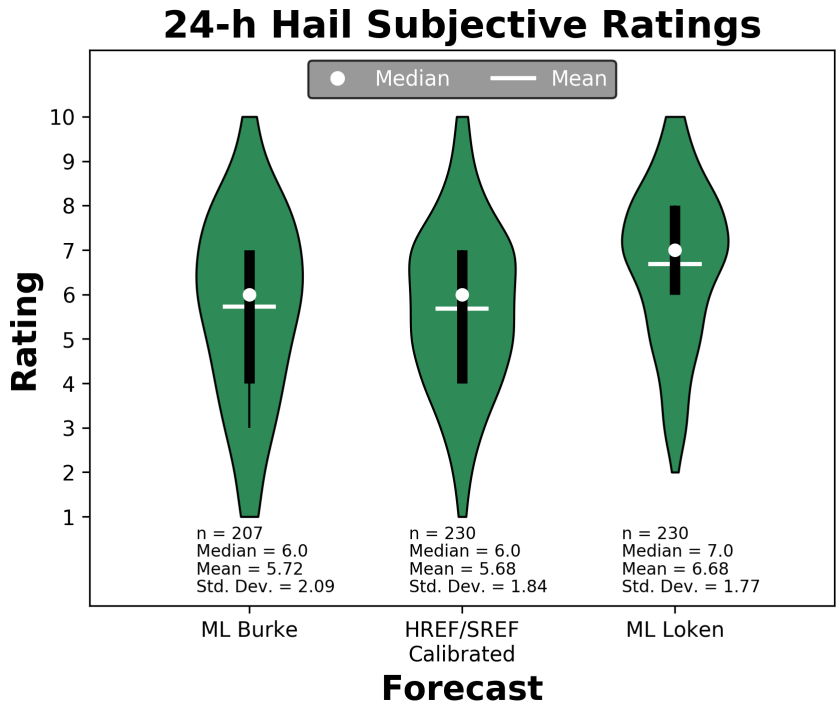


Figure 34 As in Figure 32, but for 24 h hail guidance products (see text).

iv) 4 h Hail Forecast Guidance

Participants evaluated four methods for producing 4-h calibrated hail guidance. These included the "ML RF Burke" (Burke et al. 2020) and "HREF/SREF Calibrated" (Jirak et al. 2014) methods as well as

SPC Timing Guidance produced at 0600 UTC ("06Z SPC Timing Guidance") and 1300 UTC ("13Z SPC Timing Guidance").

The two SPC Timing Guidance products received the highest mean subjective ratings (6.59 and 6.31 for the 13Z and 06Z SPC Timing Guidance, respectively; Fig. 35), as well as the smallest rating standard deviations (1.66 and 1.70 for the 13Z and 06Z guidance, respectively). Indeed, both SPC Timing Guidance Forecasts seldom received ratings below 4 (Fig. 14), indicating consistently strong performance. ML RF Burke had mean (5.90) and median (6.0) ratings that were similar to those from the 06Z SPC timing guidance but with a larger standard deviation (2.12). Figure 14 reveals that ML RF Burke received many ratings at or above 7 but also had a (smaller) local maximum in the rating distribution around 3, suggesting mostly strong performance with a few instances of relatively poor performance. Meanwhile, the HREF/SREF Calibrated method received the lowest mean (5.19) and median (5.0) subjective ratings and had the highest proportion of ratings below 5 (Fig. 35).

Participants noted that all methods provided useful timing guidance, but the SPC Timing Guidance products were the consensus favorite, followed by the ML RF Burke. Participants felt that, overall, the SPC Timing Guidance had a larger spatial distribution of non-zero probabilities, which better aligned with observed local storm reports (LSRs) and resulted in a higher POD. Participants also felt that the SPC Timing Guidance tended to be more accurate than the ML RF Burke and HREF/SREF Calibrated methods during both the early and late stages of convective development, with the ML RF Burke and HREF/SREF Calibrated methods sometimes introducing (removing) non-zero probabilities too late (early). Some participants praised the ML RF Burke method for its tighter probability gradients and higher precision compared to the SPC Timing Guidance; they felt the ML RF Burke method could better highlight more localized regions of increased threat. However, they noted that ML RF Burke sometimes produced relatively high probabilities and did not always highlight a large enough area to capture all LSRs.

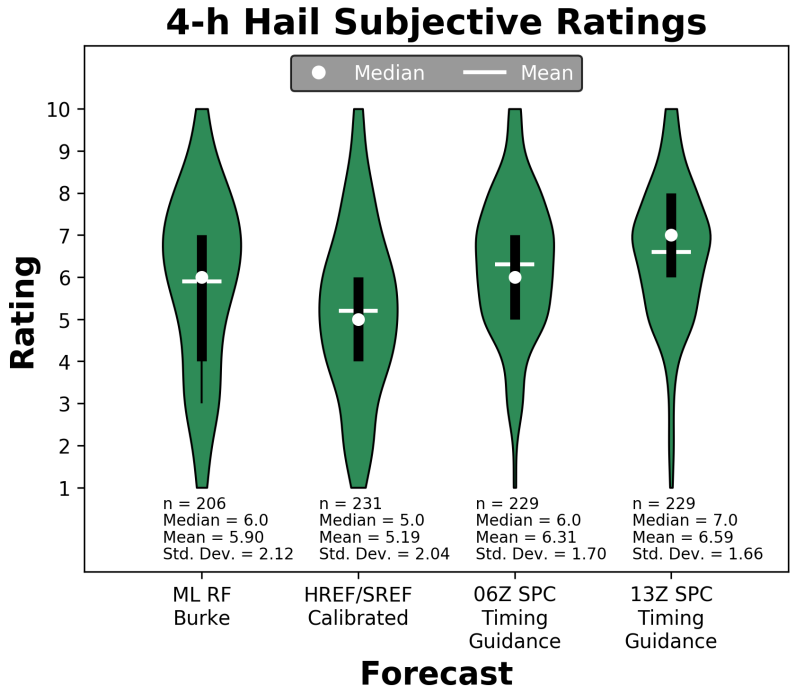


Figure 35 As in Figure 32, but for 4 h hail guidance products.

v) 24 h Wind Forecast Guidance

24-h severe wind probabilities were evaluated from the "HREF/SREF Calibrated" (Jirak et al. 2014) and "ML RF Loken" (Loken et al. 2020) methods. While both sets of forecasts had a median subjective rating of 6.0, mean ratings were slightly higher for ML RF Loken (6.00 vs. 5.50) with slightly less variance (standard deviation of 1.79 vs. 1.90). Moreover, ML RF Loken received a lower proportion of ratings less than 4 (Fig. 36), suggesting the ML RF Loken method produced more medium-to-good (i.e., rating 5-8) forecasts and fewer low-quality (i.e., rating 1-3) forecasts compared to HREF/SREF Calibrated.

Interestingly, participants frequently stated that they gave the two methods similar subjective ratings but for different reasons. The ML RF Loken method tended to give broader areas of non-zero probabilities, which gave it a better POD but also a higher FAR. Forecasters also commented that, many times, the wind probabilities from ML RF Loken seemed too high, although, spatially, the probabilities tended to capture the correct axis of LSRs. In contrast, the HREF/SREF probabilities tended to be lower and focused over a smaller area, helping forecasters identify areas of the highest concern. However, the areas highlighted by the HREF/SREF were sometimes displaced from the main axis of LSRs. Overall, participants slightly favored the ML RF Loken method, but many suggested that a blend of the two approaches may give the most useful guidance since the methods tended to have different (and complementary) strengths and weaknesses.

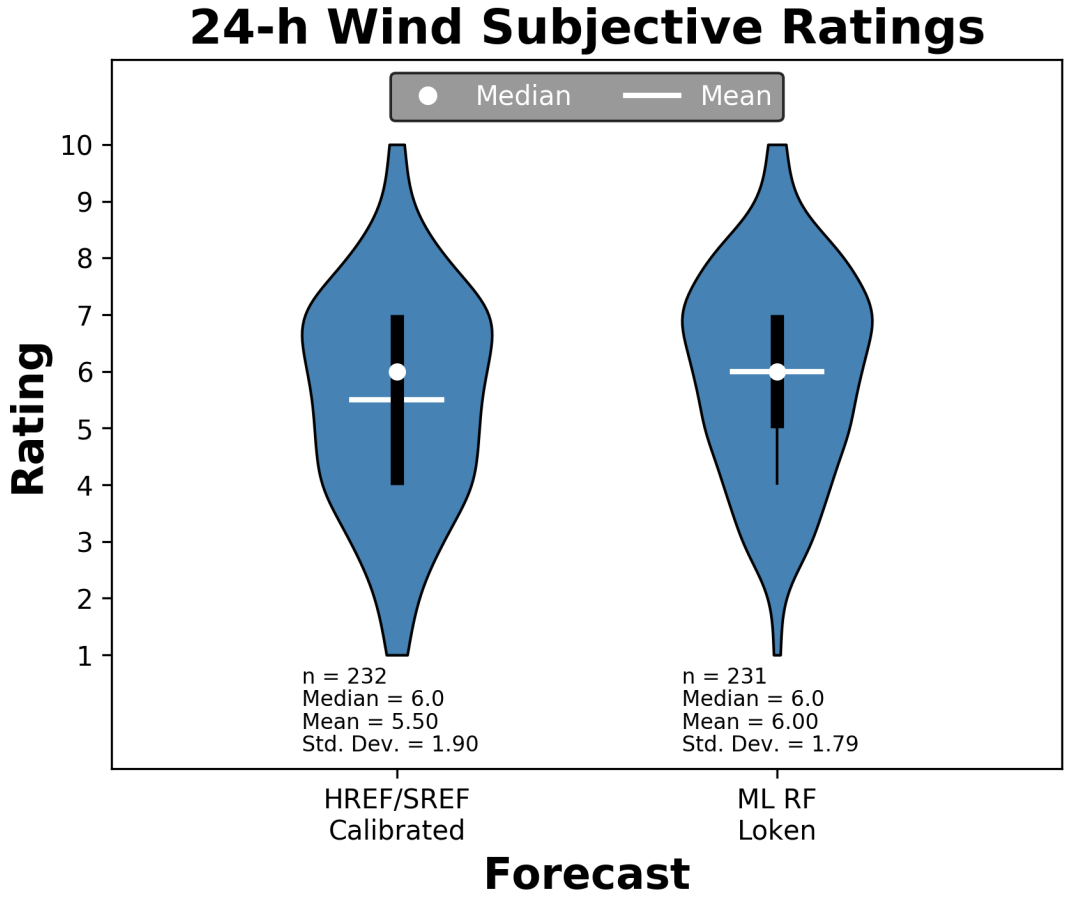


Figure 36 As in Figure 32, but for 24 h wind guidance products (see text).

vi) 4 h Wind Forecast Guidance

4-h calibrated severe wind guidance was evaluated from the HREF/SREF Calibrated (Jirak et al. 2014) method as well as the SPC Timing Guidance product produced at 0600 UTC ("06Z SPC Timing Guidance") and 1300 UTC ("13Z SPC Timing Guidance").

Overall, both SPC Timing Guidance products had similar mean (6.10 for the 06Z and 5.97 for the 13Z guidance; Fig. 37) and median (6.0 for both) subjective ratings, while the HREF/SREF Calibrated received lower mean (5.00) and median (5.0) ratings. Additionally, the HREF/SREF Calibrated guidance received proportionally more low-end (1-3) ratings and fewer high-end (7-10) ratings compared to the SPC products (Fig. 16), suggesting participants generally preferred the SPC Timing Guidance products to the HREF/SREF Calibrated timing guidance.

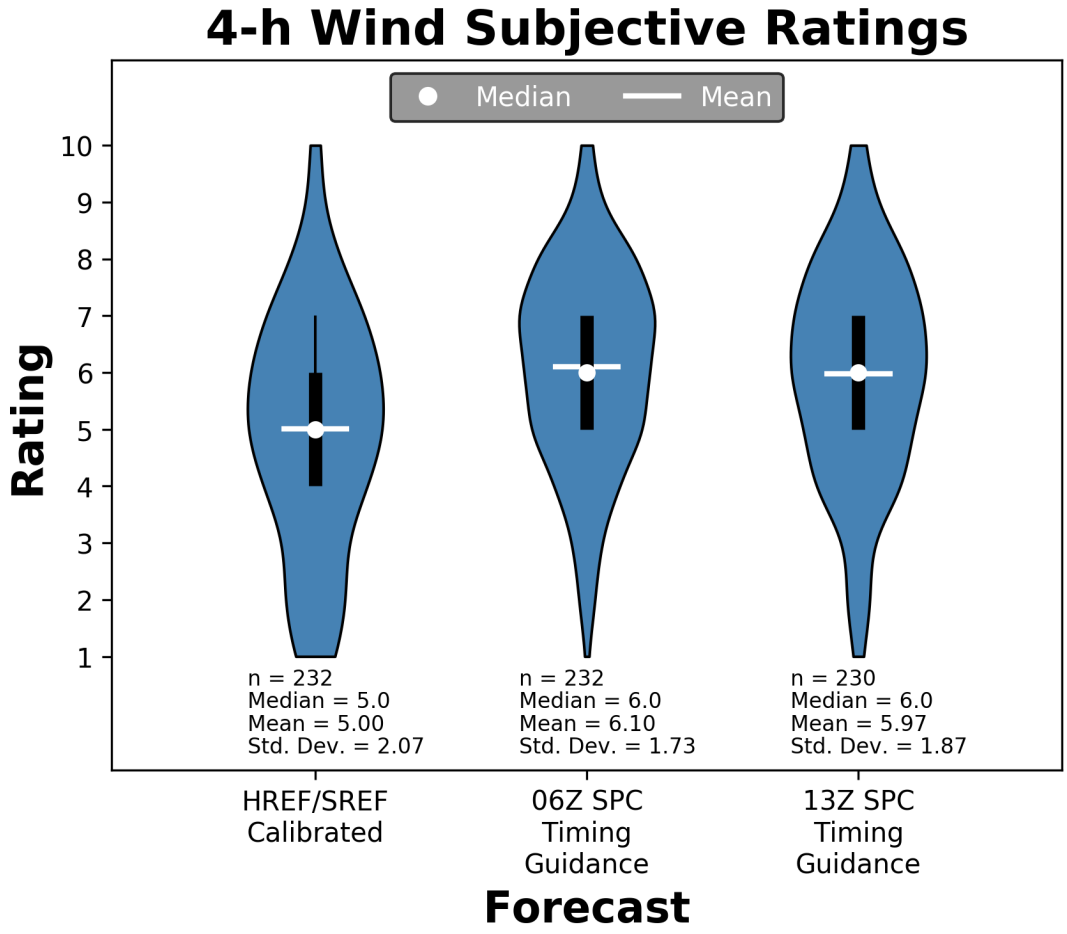


Figure 37 As in Figure 32, but for 4 h wind guidance products.

B2) CLUE 00Z Multi-Model Ensemble

The B2: CLUE 00Z Multi-Model Ensemble evaluation was another CAM ensemble evaluation similar to the A3 evaluation. In this evaluation, two single-model, time-lagged ensembles (HRRRE and UM; 1800 and 0000 UTC members) were compared to a 0000 UTC multi-model (HRRRE and UM) ensemble and a 36-member, time-lagged, multi-model ensemble (Fig. 38). HREF (v2.1 and v3) was used as a baseline for performance, as in the A3 evaluation. The B3 evaluation sought to identify the relative importance of time-lagging versus multi-model strategies in CAM ensemble performance for severe weather forecasting.

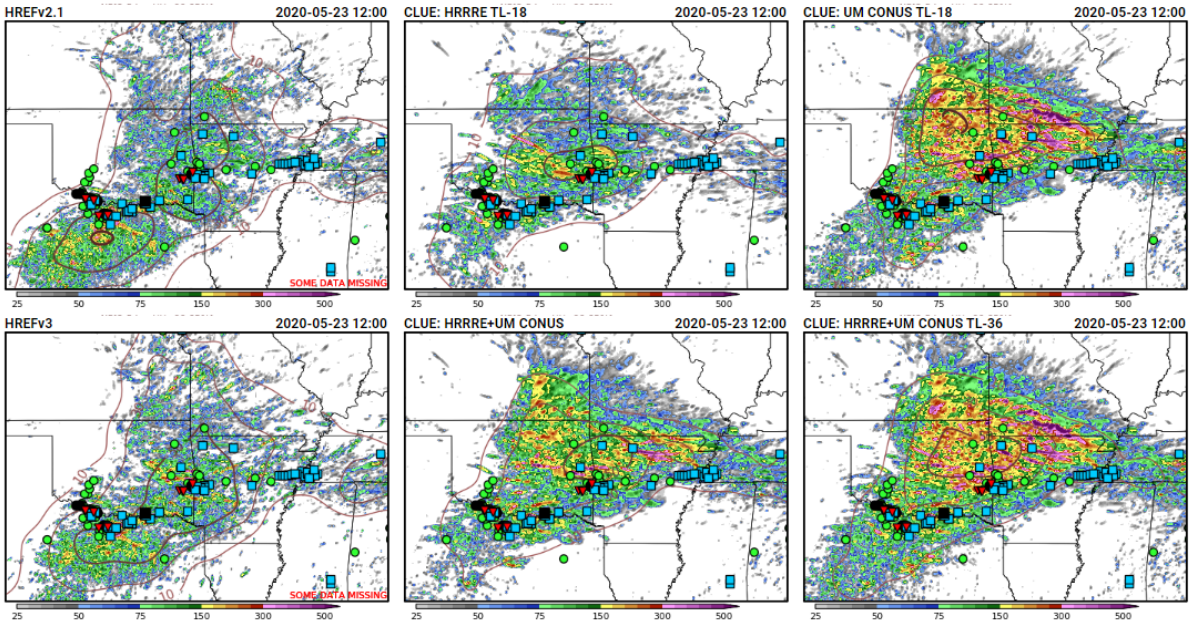


Figure 38 Example of multi-panel comparison webpage for the 0000 UTC CAM ensemble B2 evaluation during the 2020 SFE. The 24-h ensemble maximum UH (shaded) and neighborhood probability of UH>99.85th percentile (contoured) is displayed for HREFv2.1 (upper left), HRRRE TL-18 (upper middle), UM TL-18 (upper right), HREFv3 (lower left), HRRRE+UM (lower middle), and HRRRE+UM TL-36 (lower right) for 22 May 2020. Preliminary severe storm reports are also overlaid (wind - blue squares, hail - green circles, and tornado - red upside-down triangles).

The single-model time-lagged ensembles (HRRRE TL-18 and UM TL-18) performed similarly to one another in terms of overall subjective ratings (Fig. 39) with a slight edge in mean rating to the HRRRE TL-18. Interestingly, a multi-model combination of HRRRE+UM did not improve the subjective ratings over the single-model time-lagged ensembles. Likewise, combining all of the HRRRE and UM runs together in a multi-model, time-lagged ensemble did not improve the probabilistic forecasts, based on the subjective ratings (Fig. 39). The hypothesis going into the SFE was that the multi-model, time-lagged HRRRE+UM would produce the best probabilistic forecasts of the CLUE CAM ensembles and approach the skill of the HREF; however, that was not the case and indicates the HREF configuration (i.e., multi-model, time-lagged, multi-physics, and multi-initial conditions) is unique in optimizing probabilistic forecasts for severe weather.

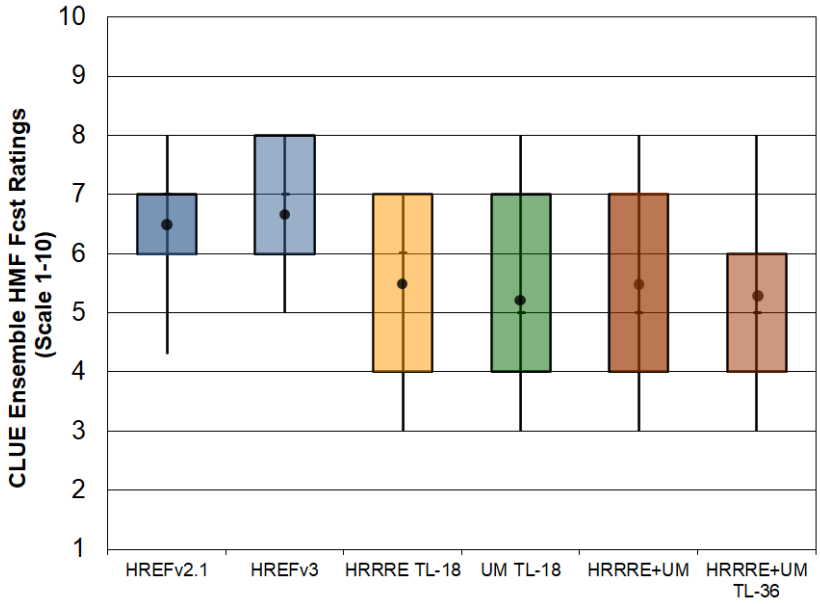


Figure 39 Distributions of subjective ratings (1-10) by SFE participants of hourly maximum fields for severe weather forecasting over a mesoscale area of interest for the forecast hours 13-36 for the B2: CLUE 00Z CAM Multi-Model Ensemble evaluation (HRRRE TL-18 - light orange; UM TL-18 - green; HRRRE+UM - red; HRRRE+UM TL-36 - light red) compared to the HREFv2.1 (blue) and HREFv3 (light blue).

B3) CLUE 12Z CAM TL-ENSEMBLE

The third and final CAM ensemble evaluation (B3: CLUE 12Z CAM TL-Ensemble) compared three nine-member, single-model ensembles using different time-lagging strategies based at 1200 UTC. These strategies included no time-lagged members (HRRRE), three time-lagged members from each of the 0000, 0600, and 1200 UTC HRRRE initializations (HRRRE TL-9); and a mix of five deterministic HRRR and four deterministic NSSL-WRF configurations each initialized with HRRRv4 initial conditions at different times between 0000 and 1200 UTC (HRRR/NSSL WRF-TL9; Fig. 40).

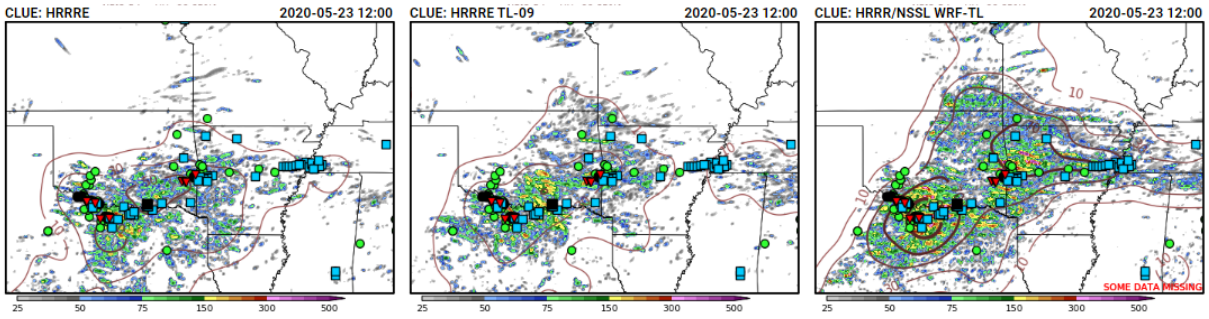


Figure 40 Example of multi-panel comparison webpage for the 1200 UTC CAM ensemble B3 evaluation during the 2020 SFE. The 24-h ensemble maximum UH (shaded) and neighborhood probability of UH>99.85th percentile (contoured) is displayed for HRRRE (left), HRRRE TL-9 (middle), and HRRR/NSSL WRF-TL9 (right) for 22 May 2020. Preliminary severe storm reports are also overlaid (wind - blue squares, hail - green circles, and tornado - red upside-down triangles).

As is evident in Fig. 40, the HRRR/NSSL WRF-TL9 ensemble typically had higher UH probabilities in daily forecasts compared to the HRRRE or HRRRE TL-9. Another comment often noted from participants was higher POD of severe weather events from the HRRR/NSSL WRF-TL ensemble versus the HRRRE or HRRRE TL-9. The time-lagged HRRRE performed similarly to the 12Z HRRRE according to the subjective ratings (Fig. 41), supporting the need to quantify the optimal number of CAM ensemble forecast members required at a single initialization time. The HRRR/NSSL WRF-TL ensemble received notably higher subjective ratings overall compared to the HRRRE or HRRRE TL-9, indicating a potentially useful single-model strategy for configuring a CAM ensemble.

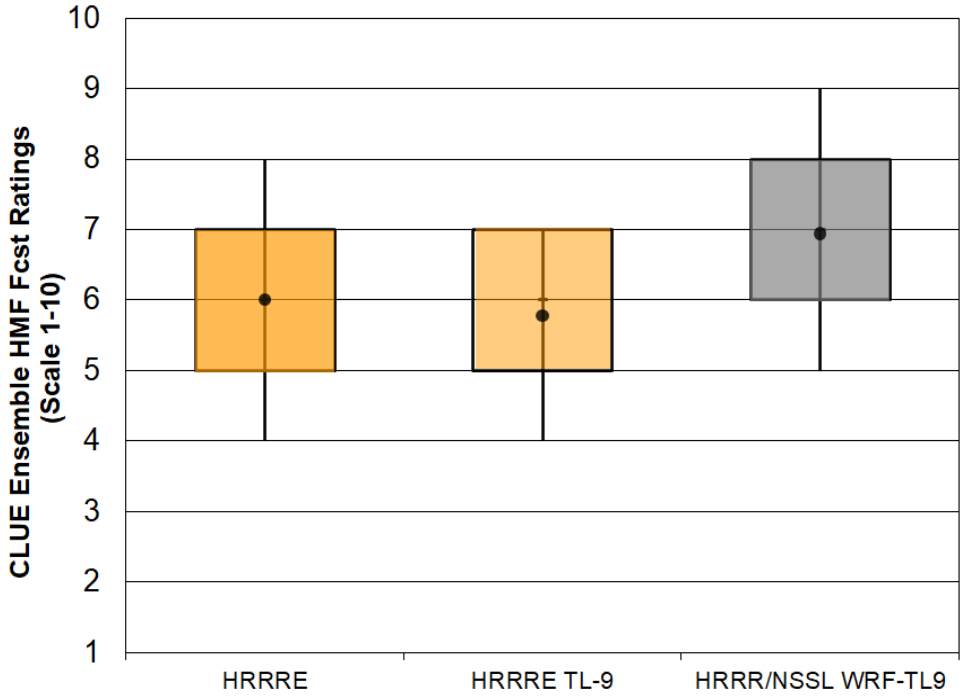


Figure 41 Distributions of subjective ratings (1-10) by SFE participants of hourly maximum fields for severe weather forecasting over a mesoscale area of interest for the forecast hours 1-24 for the B3: CLUE 12Z CAM TL-Ensemble evaluation (HRRRE - orange; HRRRE TL-9 - light orange; and HRRR/NSSL WRF-TL9 - gray).

B4) Deterministic Flagships

As in previous years, various agencies contributed state-of-the-art deterministic model guidance to SFE 2020. These runs were evaluated against the soon-to-be operational HRRRv4 provided by GSL, and included a UM core run provided by the Met Office, a global FV3 run provided by GFDL, and two FV3-SAR runs provided by EMC and NSSL. Following feedback from prior SFEs, participants were asked to evaluate the composite reflectivity and UH guidance at three specific times: 1800 UTC, 2300 UTC, and 0400 UTC. These were chosen to represent times near convective initiation, times at the peak of the convective life cycle, and times when convection would have either dissipated or grown upscale. Environmental fields (2-m temperature, 2-m dewpoint, and SBCAPE) were evaluated at two times: 1800 UTC and 0000 UTC, to similarly capture the pre-convective environment and the behavior of the environment influenced by convection (e.g., cold pools).

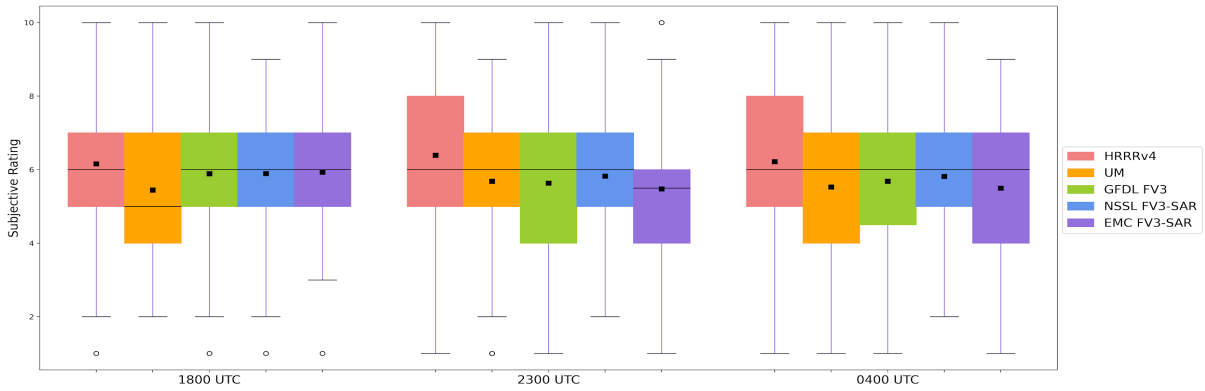


Figure 42 Subjective evaluation scores of composite reflectivity and UH from the models in the Deterministic Flagship comparison. Black squares indicate the mean score.

Overall, scores were similar across most of the models at the earliest time (1800 UTC; Fig. 42), except for the UM, which scored slightly lower than others. However, at later times (2300 and 0400 UTC), the HRRRv4 has much higher ratings than the UM or any of the FV3-based models. Amongst the FV3-based models, the NSSL FV3-SAR mean ratings were highest at the later times. Many comments indicated that the FV3-based models often had too big and too intense of storms, as depicted by composite reflectivity. Despite this overprediction, participants sometimes ranked the NSSL FV3-SAR as slightly more realistic than the GFDL FV3 or EMC FV3-SAR. However, some participants mentioned the NSSL FV3-SAR as having too much or too intense convection. Though one of the questions on the survey asked about differences between models with different cores, participants often noted that differences within the three FV3-based models were as large as the differences between cores. The HRRRv4 and UM were also noted as being too slow in a couple instances. Other participants noted that the UM produced too much cellular convection, which at times organized upscale where convection did not actually occur. There was also a lack of stratiform precipitation in the UM compared to the HRRRv4 and FV3-based forecasts. These comments support the findings of which CAM best captured the full convective evolution over the entire forecast run, as ~40% of the time participants chose the HRRRv4 (Fig. 43). Following the HRRRv4 in frequency of performing best were, in order, the NSSL FV3-SAR, GFDL FV3, EMC FV3-SAR, and UM. Participants could choose more than one model, so if two or more models were performing similarly, both could be selected.

The environmental fields told a similar story (Fig. 44), with the mean HRRRv4 score highest. The UM was rated lower than nearly all of the other models at 1800 UTC, but by 0000 UTC had the second-highest mean rating. This was likely partially influenced by ongoing convection; given the overextended coverage of convection in the FV3-based models noted by participants, we would expect more cold pools that would lead to larger differences from the observations compared to the HRRRv4 and UM forecasts. Also, of note, the UM did not have SBCAPE available for ratings, so its ratings only considered 2-m temperature and dewpoint. Participants noted that the GFDL FV3 cold pools and CAPE often closely matched observations, although an overall cool and moist bias was once more seen in the FV3-based models. These biases led to an overall low instability relative to observations, which often led participants to wonder about how the overextended convection seen in the reflectivity and UH panel was supported. Conversely, participants also mentioned a warm and dry bias in the HRRRv4 relative to the observations.

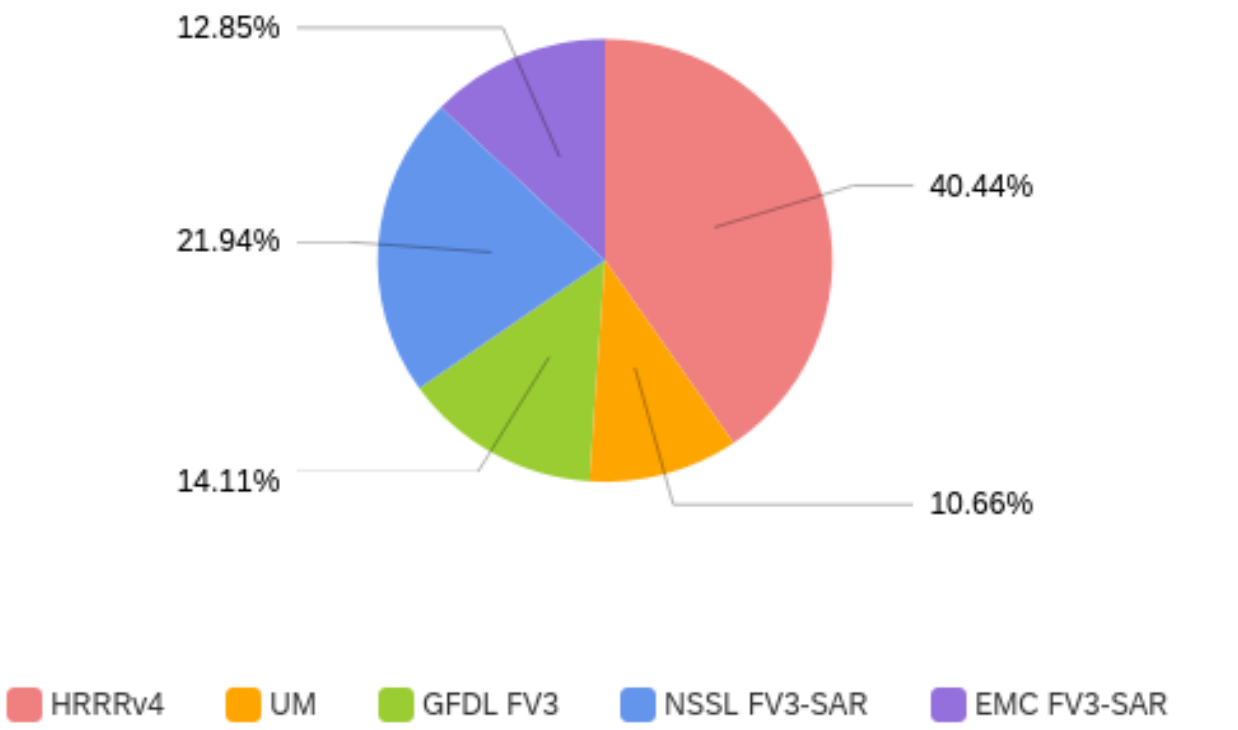


Figure 43 Participant responses to the question, "Which deterministic CAM(s) best captured the convective evolution (i.e., timing, CI, convective mode) through the entire forecast run?"

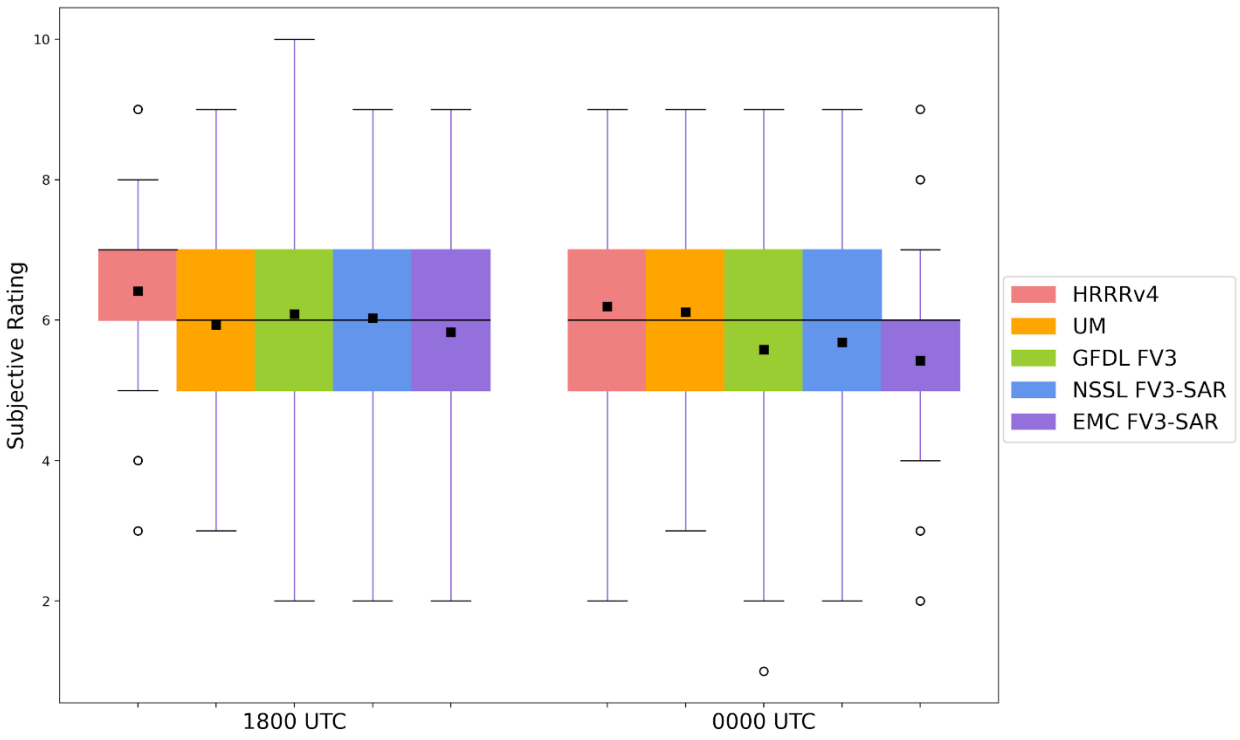


Figure 44 As in Figure 42, but for 2-m temperature, 2-m dewpoint, and SBCAPE.

B5) CLUE CORE AND ICS

A new comparison for SFE 2020 examined the impact of different initial conditions and model dynamical cores by comparing six different model runs. Two models each using the WRF, FV3, and UM dynamical cores were run using initial conditions from either the GFS or the UM global models. Similar to the other deterministic model evaluations (A6, A7, and B4), participants were asked to evaluate the simulated composite reflectivity and UH at 1800 UTC, 2300 UTC, and 0400 UTC to capture a given day's full convective life cycle. They were also asked to evaluate the 2-m temperature, 2-m dewpoint, and surface-based CAPE at 1800 UTC and 0000 UTC. Two models from the B4 comparison were also displayed in this comparison: the NSSL FV3-SAR (FV3 GFS-ICs) and the UM (UM UM-ICs). Participants were reminded of their ratings from the B4 evaluation of these two models and asked only to assign numerical ratings to the four new model forecasts.

Participants were also asked to evaluate whether differences were larger between models with different dynamical cores but the same initial conditions or models with different initial conditions but the same dynamical cores. A slider positioned initially between those two options could be dragged to one side or the other, with the default indicating that the ICs and dynamical cores appeared to have the same impact on the forecasts. Participants differed on what had the most impact on the forecasts, with the composite reflectivity and UH fields having larger differences between models with different dynamical cores (Fig. 45a) and the environmental fields having larger differences between models with different initial conditions (Fig. 45b). This was demonstrated both in the means, but also in the frequency of answers where participants slid the slider fully to the left or to the right, indicating that they clearly saw larger differences in one aspect of the model configurations compared to the other.

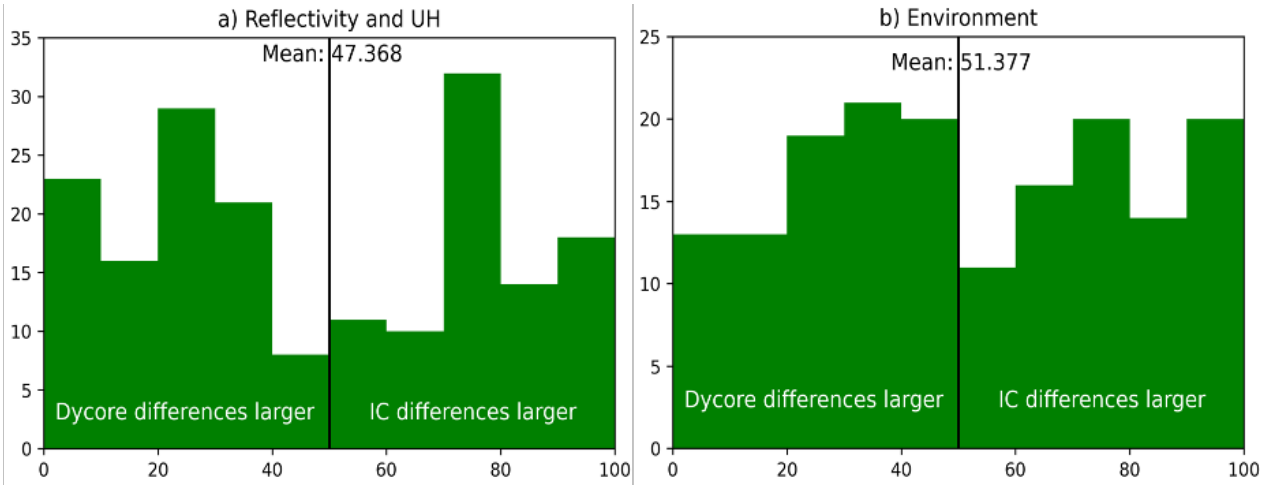


Figure 45 Participant answers to the question, "Did you see more differences between <variable> from models with the same ICs and different dynamical cores (i.e., between forecasts in the same column), or between from models with the same dynamical core and the same ICs (i.e., between forecasts in the same row)?"

These impressions of relative importance between the dynamical cores and the ICs were also reflected in the distributions of the ratings, with larger differences between model cores than between initial conditions for the composite reflectivity and UH ratings at all times (Fig. 46). Note that model availability for this comparison shifted throughout the SFE, with only 6 days having all 6 models available.

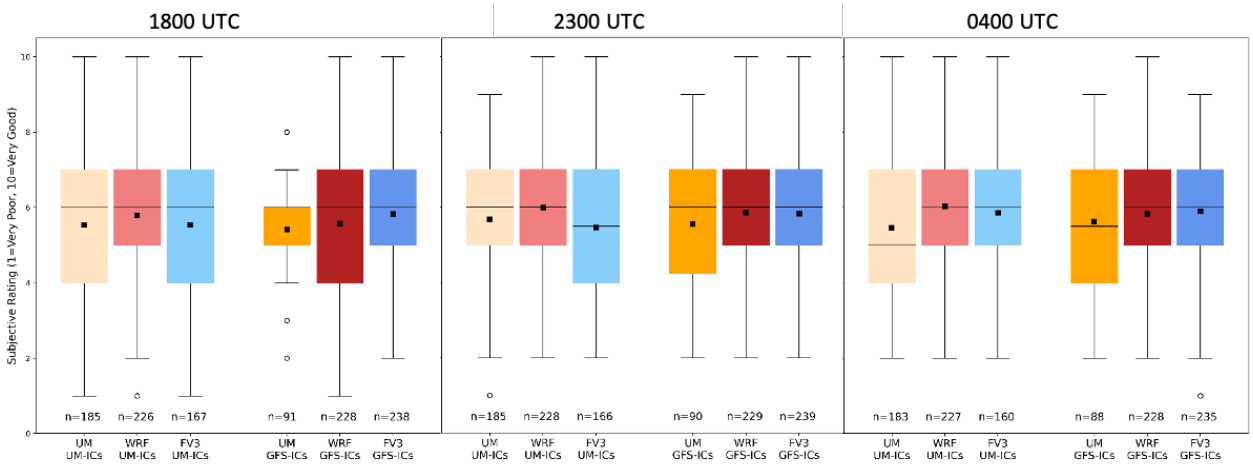


Figure 46 Participant ratings of the composite reflectivity and UH forecasts from models with different ICs and dynamical cores at (a) 1800 UTC, (b) 2300 UTC, and (c) 0400 UTC.

Generally, medians were the same for all models, although the UM with UM-ICs and the FV3 with UM-ICs each had a time with a lower median score than the other two dynamical cores using the same ICs. Across the times, the WRF and FV3 dynamical cores had the highest ratings, with a WRF-based model having the highest mean at 2300 UTC (Fig. 46b) and 0400 UTC (Fig. 46c). However, the FV3 with GFS-ICs also performed quite well at 1800 UTC (Fig. 46a), with a mean that was approximately equivalent to the WRF with UM-ICs. The WRF dynamical core performed better with UM ICs than with GFS ICs, while the UM and GFS dynamical cores showed similar subjective performance at 1800 UTC and 0400 UTC. At 2300 UTC, however, the UM and FV3 both perform best when using ICs from a parent model that shares their dynamical core. Among the models using UM ICs, the WRF dynamical core had the highest mean score at all times. Among the models using GFS ICs, the FV3 alone (1800 UTC) or the FV3 and the WRF dynamical cores (2300 UTC and 0400 UTC) had the highest mean scores. In their comments, participants often mentioned differences in timing, coverage, and simulated storm structure as being the aspects of the forecasts that differed most. When asked what model overall performed best for convective evolution throughout the entire forecast, WRF dynamical cores were more frequently selected than the other cores (56% of responses; Fig. 47), followed by the FV3-based models (36% of responses) and the UM-based models (9% of responses). These differences were much larger than the differences between ICs, with the models using UM-ICs being selected as performing best 54% of the time and models using GFS-ICs being selected 47% of the time (Note: percentages do not add to 100% due to rounding).

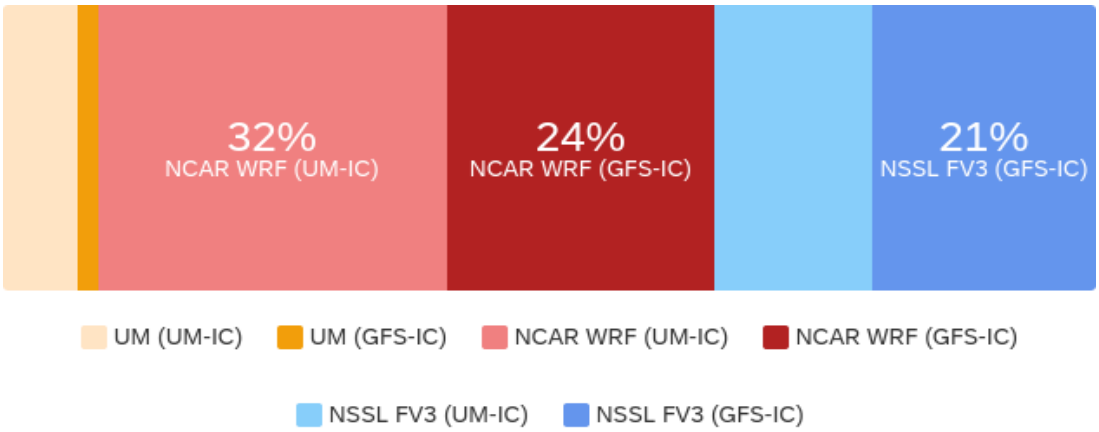


Figure 47 Participant responses to the question: "Which deterministic CAM(s) best captured the convective evolution (i.e., timing, CI, convective mode) through the entire forecast run?"

For the environmental fields, the importance of ICs was perceived to have slightly more of an effect than the dynamical cores throughout the entirety of the model forecast. This was particularly evident during the daily discussions with participants, as they would often mention the clear differences in the environmental fields at early forecast hours (e.g., f01-f03), prior to the hours asked about in the survey, which occurred much later in the forecast period. For the environmental fields, the dynamical core performed best with ICs from the same parent model (i.e., UM with UM-ICs and FV3 with GFS-ICs; Fig. 48). This trend stood out much more in the environmental fields relative to the composite reflectivity and UH fields, with an evident shift in the distributions of the UM and the FV3 with the different ICs used. The WRF model showed much less of a dependence on the ICs in the overall subjective evaluation distributions, with a slightly higher mean using UM ICs at 1800 UTC (Fig. 48a) and a higher mean using GFS-ICs at 0000 UTC (Fig. 48b). Overall, the WRF models had the highest mean rating at 1800 UTC, and the UM with UM ICs performed best at 0000 UTC. When asked about what differed between dynamical cores, comments highlighted the cool and moist tendencies of the FV3 dynamical core, the warm and dry tendencies of the WRF dynamical core, appearance of cold pools linked to small-scale convection in the UM, and varying cool pool strength. When asked about what differed between the ICs, participants noted that boundary locations differed more as a function of ICs, that GFS-ICs were more often too cool and moist, and that the SBCAPE amplitudes varied more as a function of IC rather than dynamical core. Participants also noted cases where one aspect of the forecasts seemed to make a large difference earlier (typically ICs), but by the end of the forecast the other aspect of the forecasts made a larger difference (typically cores).

Though the subjective evaluations from these models show small differences in the distributions of numerical ratings, large day-to-day variability was evident during the experiment as to the relative importance of ICs vs. dynamical core, with no clear sense of what had the highest impact overall by the end of the five weeks. This day-to-day variability can be seen by the relatively frequent answering of participants at the extreme left or right of the histograms in Fig. 45, as well as the mean values lying close to the midpoint of the histograms. Thus, ongoing research will investigate the model performance using objective statistics in both the aggregate and on a case-by-case basis, as well as the subjective responses in a case-by-case framework.

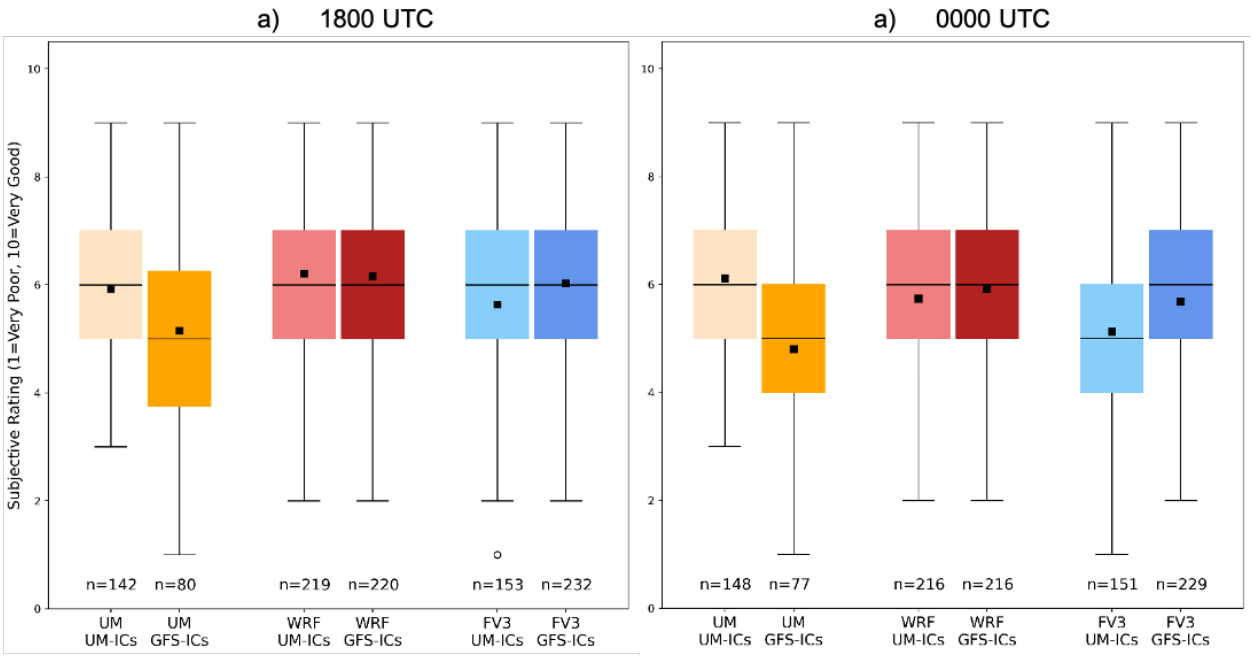


Figure 48 As in Figure 55, but for 2-m temperature, 2-m dewpoint, and SBCAPE at (a) 1800 UTC and (b) 0000 UTC.

B6) WoFS CONFIGURATONS

Multiple configurations of the WoFS were tested during the 2020 SFE, including ensemble and deterministic configurations. First, participants were asked questions about two WoFS ensembles with horizontal grid spacings of 3 km and 1.5 km. Three different initialization times were evaluated to see how WoFS performance changed as initialization times grew later, in addition to comparing the performance of the 3 km and 1.5 km resolution ensembles. The 1.5 km ensemble ran for three hours and had nine members, compared to eighteen members in the 3 km ensemble, which ran for six hours.

Participants were instructed to only consider the first three hours of the forecasts in their ratings. For the most part, the 3 km and 1.5 km ensembles performed similarly to one another (Fig. 49), although the mean rating for the 1.5 km ensemble was less than the mean rating of the 3 km ensemble at the same initialization time. The initialization times also didn't differ much within each ensemble, although mean ratings tended to increase as the ensemble initialization time grew later. This is likely at least partially due to the higher likelihood of convection for the WoFS ensembles to assimilate later in the convective day. Though the overall ratings don't show much difference in the distributions between the initialization times, looking at differences in the individual participant ratings (Fig. 50) for each case between initialization times can help show whether participants increased or decreased their ratings for each ensemble with later initializations. The mean change in rating between even the earliest initialization (20z) and latest initialization (00z) was zero, but each distribution has some extremes, with later initializations performing up to four points better or worse than earlier initializations for both the 3 km and the 1.5 km ensembles. However, most of the differences were within a one-point range, indicating that participants saw mostly small changes between initializations on a given day. Outliers occurred when the WoFS began picking up on convection, or missed ongoing convection entirely.

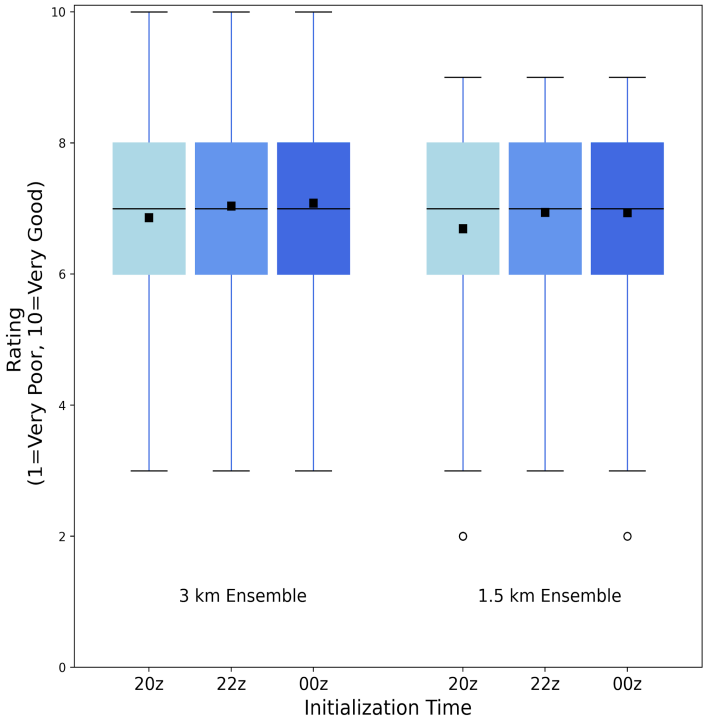


Figure 49 Subjective evaluation of the hourly maximum UH neighborhood probability fields from the 3 km and 1.5 km WoFS ensemble forecasts, initialized at three different times. The first three forecast hours of each initialization was evaluated.

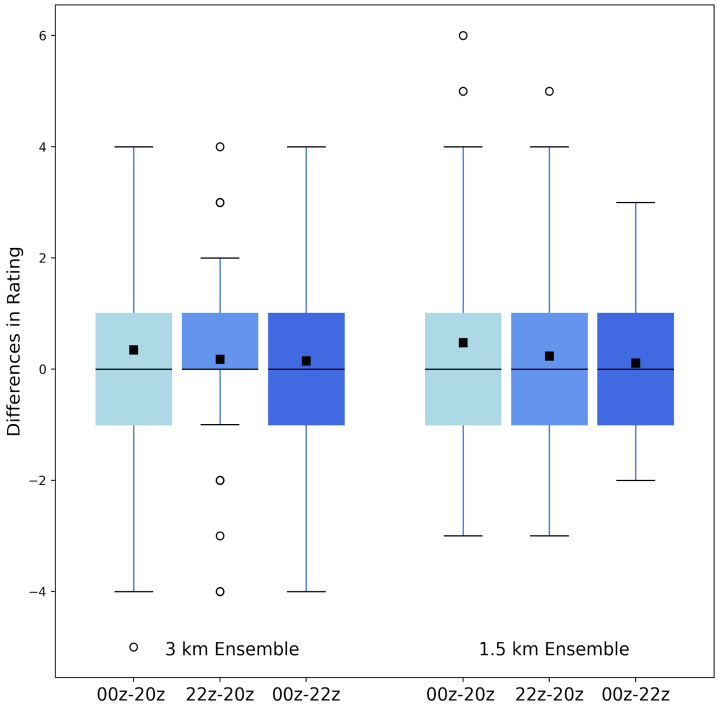


Figure 50 Participant rating differences for various initializations of WoFS forecast. A '0' difference indicates that the two initializations were rated the same by a participant on a given day.

In addition to asking participants to provide a 1-10 rating of the usefulness of the WoFS ensemble outputs, a few qualitative questions were asked regarding the convective mode and convective initiation timing, to help determine whether the higher horizontal resolution of the 1.5 km WoFS ensemble was providing better mode information or capturing convective initiation better than the 3 km WoFS ensemble (CI is a noted challenge for the WoFS, as found in previous SFEs). Since the 1.5 km ensemble was available less frequently than the 3 km ensemble, the following results are presented in terms of frequency of response, rather than in raw response counts.

When asked how the various ensembles and forecasts were depicting convective mode, nearly half of the responses for any ensemble forecast were “Very accurately” (Fig. 51). Later initializations were more likely to have a response of “Extremely accurately”, and the 1.5 km 00z initialization of the WoFS had the highest frequency of this response. Also, while the 3 km had a few responses of “Not accurately at all”, none of the 1.5 km WoFS runs received this answer.

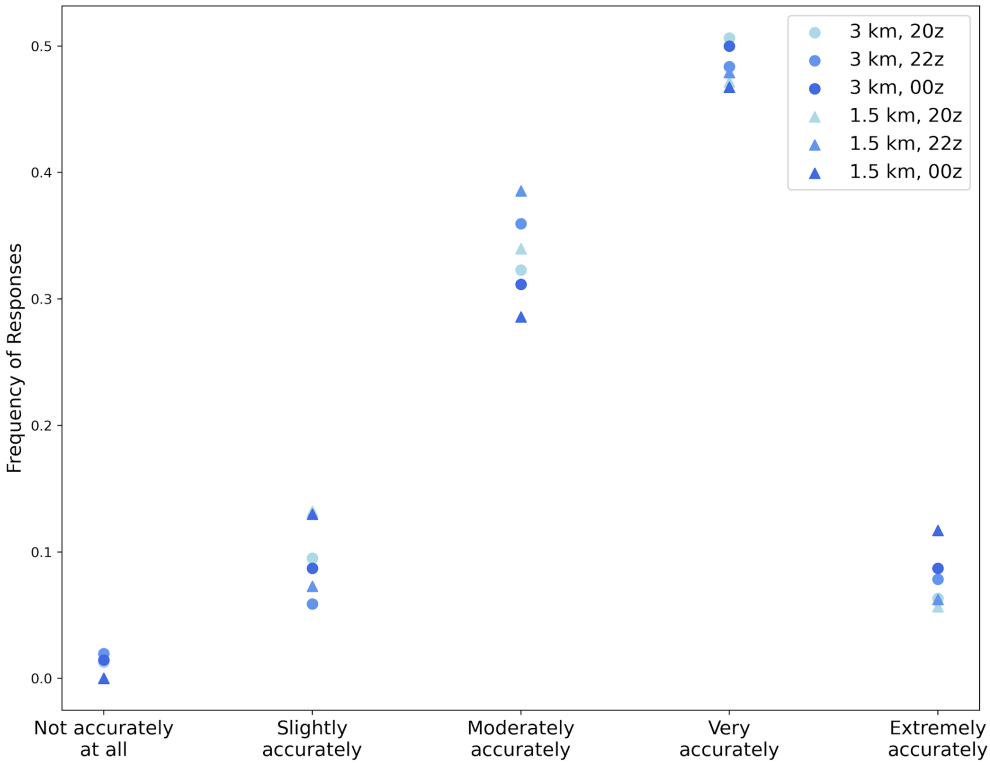


Figure 51 Participant responses to the question "If appropriate, how do the following ensembles initialized at XXXX UTC depict convective initiation?"

Perhaps contrary to our expectations, at every initialization time, the 3 km WoFS was more frequently rated “About right” for convective initiation time if convection was not already ongoing at the start of the forecast (Fig. 52). Convection was ongoing for about half of the forecasts and more frequently was ongoing at later initialization times, as expected given the typical diurnal convective cycle. Convective initiation was more frequently noted as occurring “Too fast” than “Too slow”, particularly at early initialization times.

When asked explicitly if they thought that the 1.5 km ensemble provided value above the 3 km ensemble, participants most frequently responded “Might or might not”, followed by “Probably not” (Fig. 53). Of all forecast initialization times, the 20z forecast seemed to be where the largest differences were evident to participants, since this time received the fewest “Might or might not” responses and the most “Definitely yes” and “Definitely no” responses from participants. Results are more mixed for the 22z forecast, but the 00z forecast shows the most “Might or might not” responses. These results suggest that, at least subjectively, the most information to be gained by the higher resolutions may be at the earlier forecast times. However, it is not yet clear that the additional information is from better depiction of convective mode or convective initiation. One additional caveat to these results is that neighborhood probabilities of UH, updraft speed, 10m wind, and composite reflectivity were being compared between these ensembles. Given that higher horizontal resolution leads to different expected values in many of these variables, different thresholds were chosen for generating the neighborhood probabilities for the 1.5 km ensemble (e.g., UH ≥ 400m^2s^{-2}) vs. the 3 km ensemble (e.g., UH ≥ 75m^2s^{-2}). These thresholds values were calculated using a subset of data and relationships derived from previous studies of other convection-allowing models, but it is possible that these threshold values were not calibrated correctly. Comments from participants reflect this uncertainty, given that very high values of UH from individual 1.5 km WoFS members were visible using the maximum UH underlay.

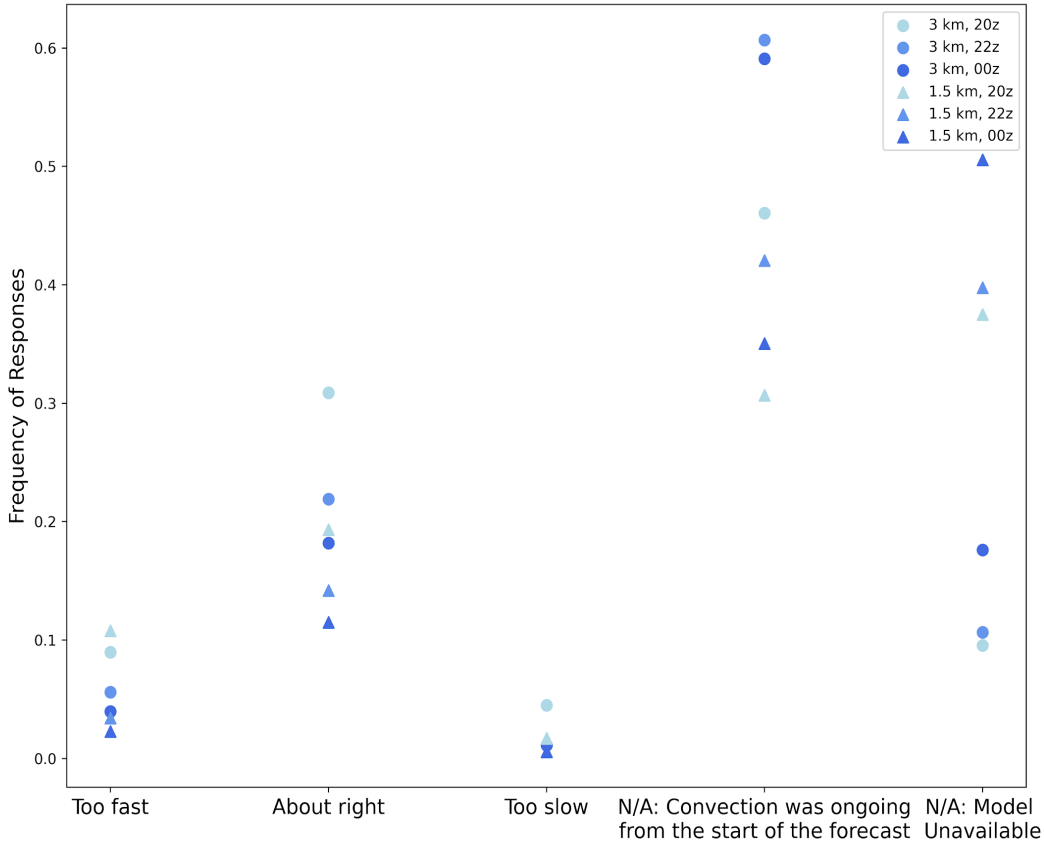


Figure 52 Participant responses to the question "If appropriate, how do the following ensembles initialized at XXXX UTC depict convective initiation?"

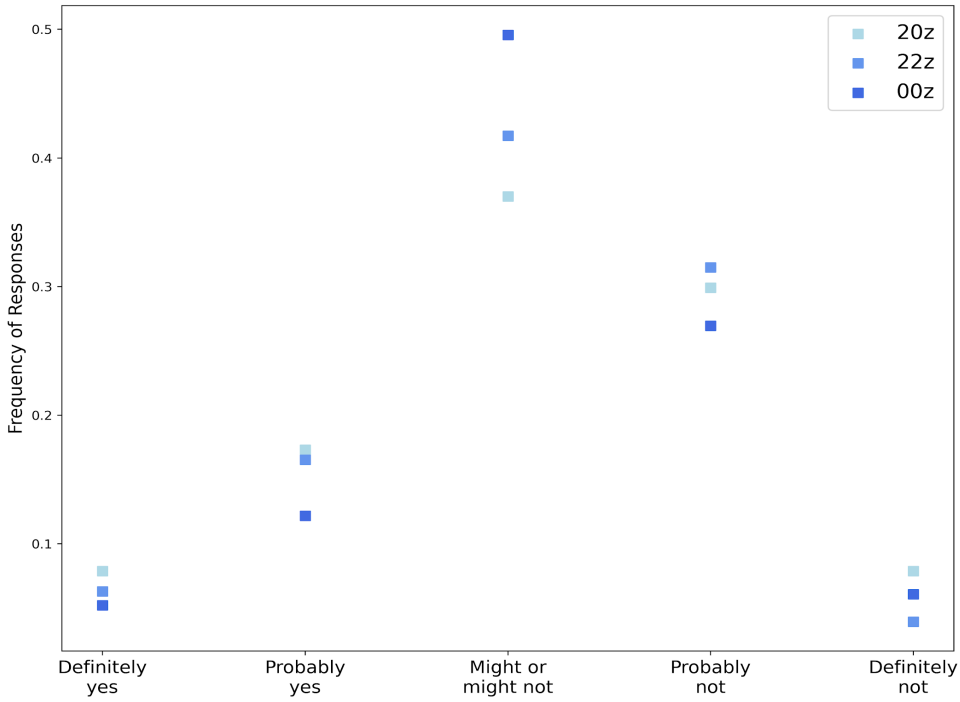


Figure 53 Participant responses to the question "At XXXX UTC, does the experimental 1.5-km ensemble provide additional useful information compared to the RT 3.0-km ensemble?

The 6-h forecast length of the 3 km ensemble allowed for comparison of 1-h, 2-h, 3-h, 4-h, 5-h, and 6-h forecasts on one six panel screen (Fig. 54). Participants were asked to look at this six-panel figure and determine whether WoFS was performing better or worse at shorter lead times.

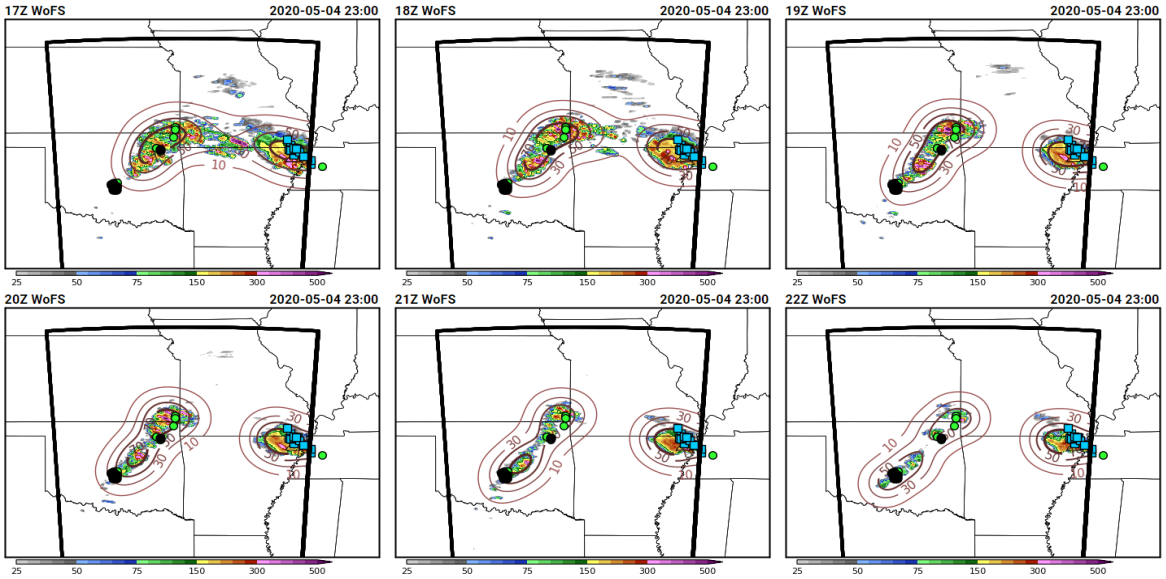


Figure 54 Forecasts of max UH and neighborhood probability of UH > 99th percentile from six subsequent initialization times of the 3 km WoFS, valid from 2300 - 0000 UTC on 4 May 2020. Storm reports of wind (blue squares) and hail (green circles). Significant reports have the same shapes, but are filled black.

As expected, given the WoFS system's advanced data assimilation, the most common responses were that the WoFS was either "Much" or "Slightly" better at shorter lead times (Fig. 55). On only a few occasions were the forecasts with shorter lead times worse, which according to participant feedback occurred when the WoFS decreased the intensity close to the event compared to previous runs but severe weather was reported.

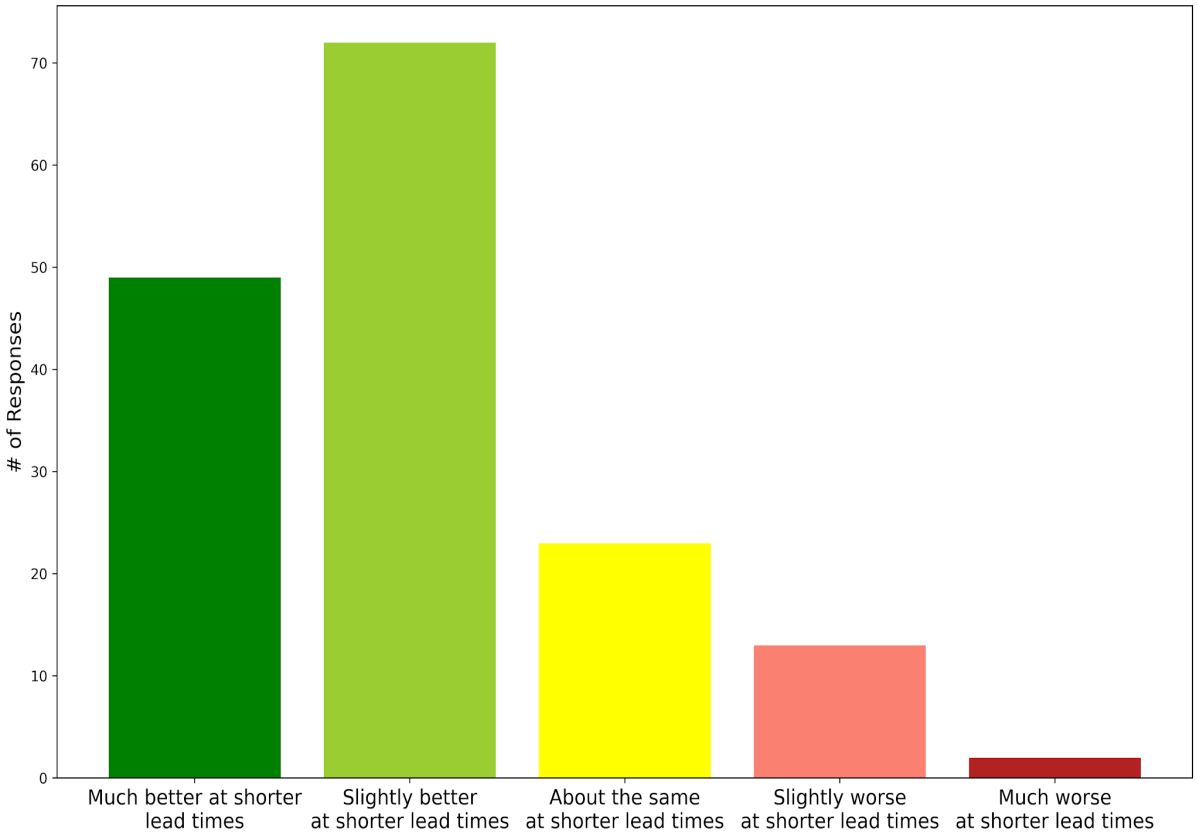


Figure 55 Participant responses to the question "How does WoFS forecast performance change with decreasing lead-time?"

The final WoFS evaluation involved two deterministic 1.5 km runs that used different data assimilation strategies. Participants were asked to rate composite reflectivity output from these runs on a scale of 1-10, to investigate the value of a deterministic high-resolution (1.5 km) run compared to an ensemble of relatively coarse-resolution runs (3 km). Generally, scores were lower for the deterministic runs (Figure 56) than for the 3 km ensemble (Fig. 49), with the median score for all of the deterministic runs except the 22z Hybrid run being one point lower than the corresponding ensemble scores. Since participants were reminded of the score that they had given to the corresponding 3 km ensemble while providing these 1-10 ratings, the ensemble appears to provide about a point of value on the 1-10 scoring scale relative to a single deterministic run. Overall, the Hybrid DA run scored higher than the Var DA for both runs examined by participants. Scores for the 22z runs of each respective model were also higher than the 20z runs. Mean differences were relatively small between runs, but the overall distribution of the Hybrid DA run had higher scores.

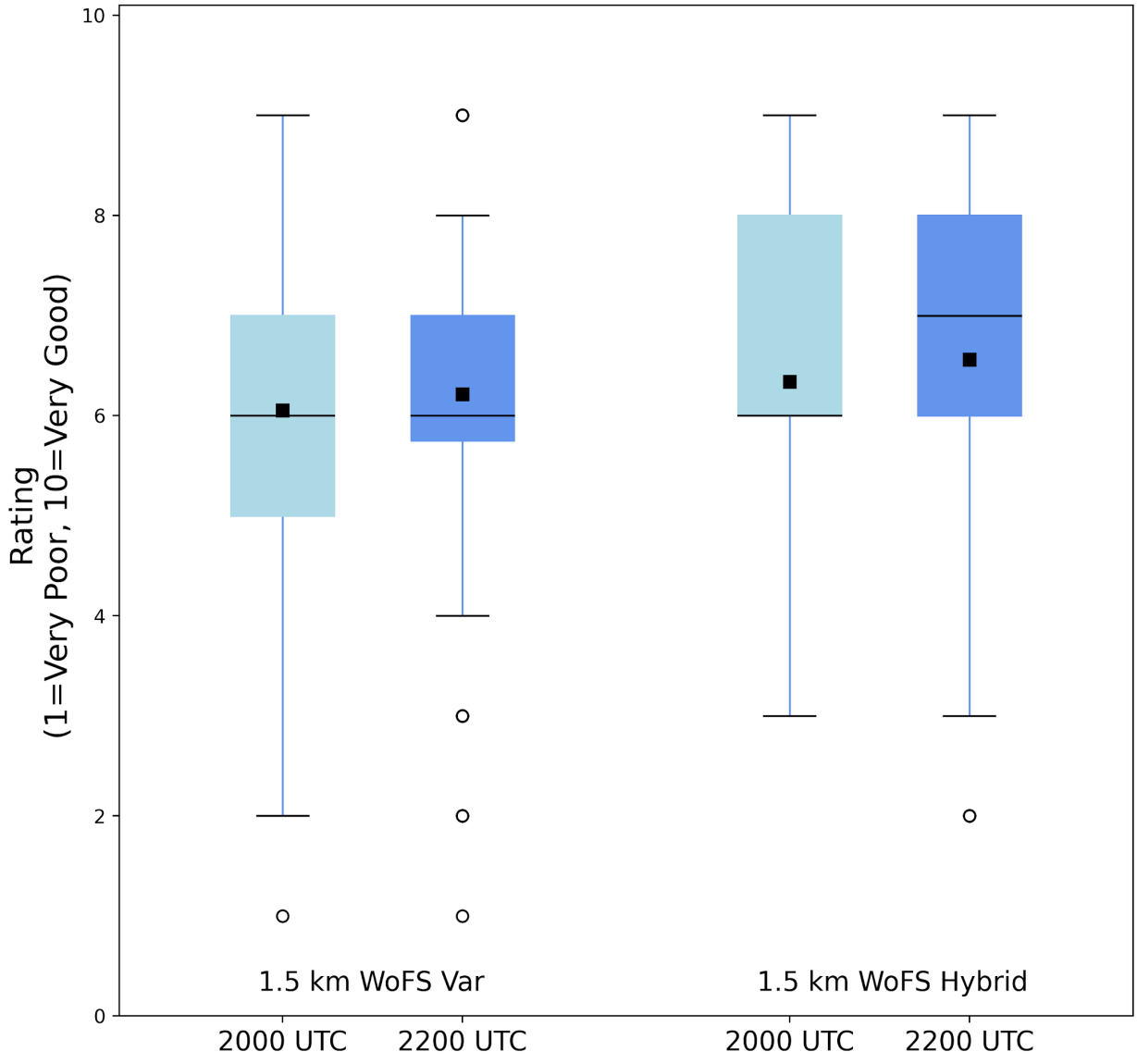


Figure 56 Subjective ratings of the composite reflectivity and UH of two deterministic WoFS forecasts using different data assimilation strategies, at two initialization times.

c) Evaluation of experimental forecast products – Innovation Group

For the Innovation group forecasting activity, participants generated individual severe hazard probabilities for a 1-h time window valid 4-5pm (2100-2200 UTC), and a 4-h time window, valid 4-8pm (2100-0100 UTC). An initial forecast was generated during the 2-3pm period and an updated forecast during the 3-4pm period. One group of participants used WoFS guidance to generate these outlooks, while WoFS guidance was withheld for the other group. For both groups, participants were able to use all experimental and operational guidance that was available to them through HWT and other public websites. In addition to the two or three NWS forecasters that participated in this activity each week, two other forecasters – Dave Imy (retired SPC) and Mike Coniglio (NSSL researcher/part time SPC

forecaster) – participated all five weeks, as well as SFE facilitator Adam Clark (NSSL researcher). Dave and Mike were always in opposite groups (i.e., when Dave used WoFS, Mike did not use WoFS and vice versa), while Adam always used WoFS. Figure 57 is an example of 4-h hail outlooks produced on 5 May 2020.

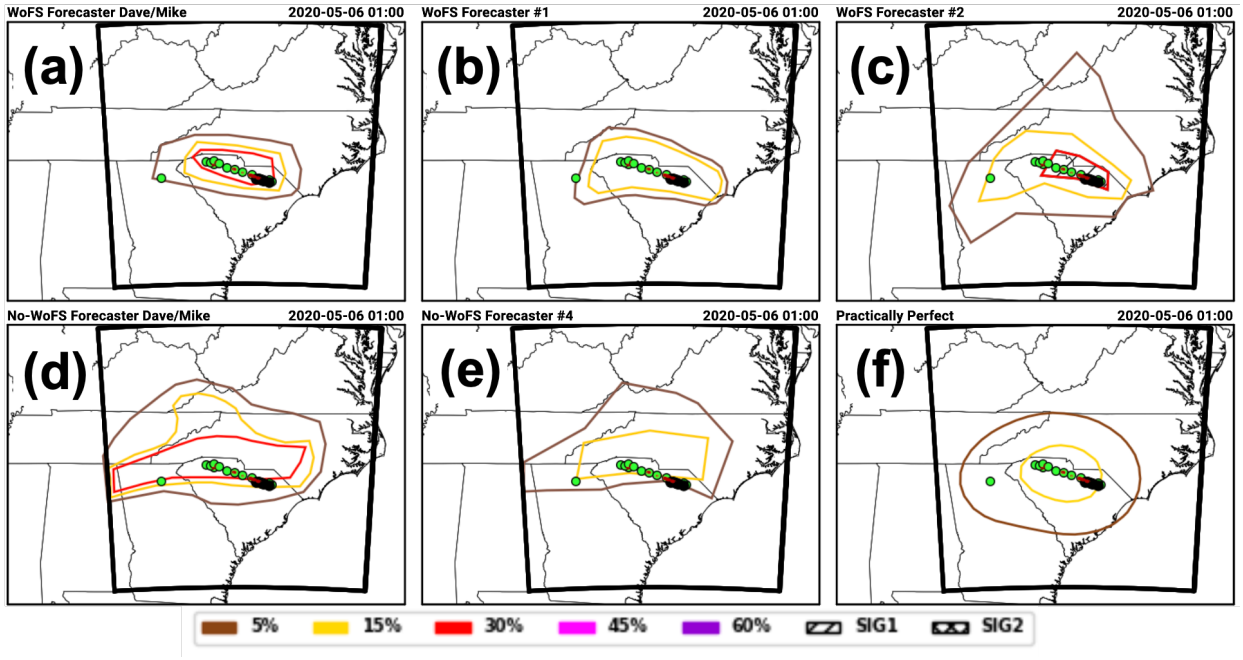


Figure 57 Experimental, initial 4-h hail outlooks (contours) issued by 3pm (2000 UTC) and valid 4-8pm (2100-0100 UTC). (a) and (d) are the "WoFS" and "No-WoFS" outlooks, respectively, generated by David and Michael. (b) and (c) are "WoFS" outlooks generated by NWS forecasters and Adam, (e) is the "No-WoFS" outlook generated by an NWS forecaster, and (f) is the practically perfect (Hitchens et al. 2013) outlook used as a verification tool. In each panel, the thick black outline indicates the WoFS domain, green dots indicate hail 1.0-in, black dots hail 2.0-in., and contour colors refer to specific probabilities, which are indicated by the legend at the bottom.

For consistency and to make the next-day evaluations manageable, ratings were only assigned to the "Dave and Mike" forecasts, which were referred to as "WoFS" and "No-WoFS". A total of 24 unique outlooks were evaluated each day (preliminary and updated 1- and 4-h WoFS and No-WoFS outlooks for each hazard). Results for the 1-h outlooks are shown with box plots in Figure 58. Generally, tornado outlooks received higher average ratings than wind and hail. Differences between the initial and updated outlooks were very small. Finally, comparisons of the WoFS and No-WoFS ratings for each hazard in both the initial and update outlooks revealed very small differences. In the updates (Fig. 58; right panel), WoFS had slightly higher scores than No-WoFS, but a simple Welch's t-test indicated that the differences were not significant ().

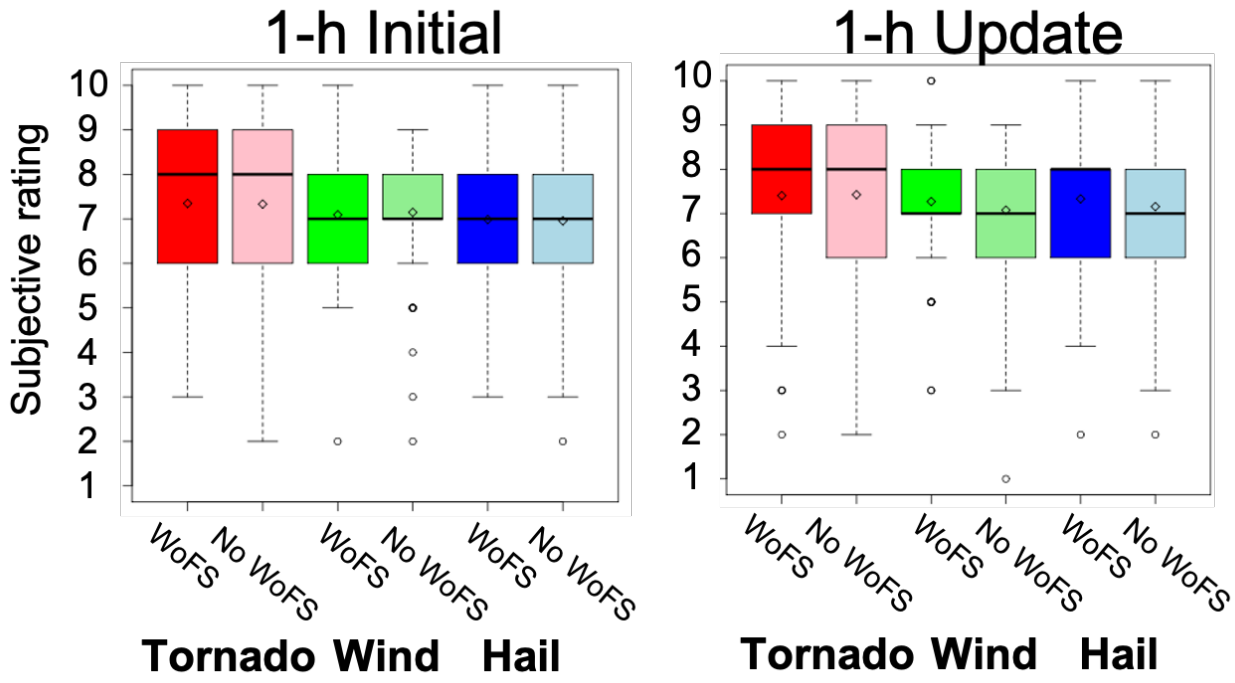


Figure 58 Box plots showing the distributions of subjective ratings for WoFS and No-WoFS forecasts of individual hazards. Left panel is the initial outlook and right is the update. The thick black line within the shaded regions indicates the median, while the diamond is the mean.

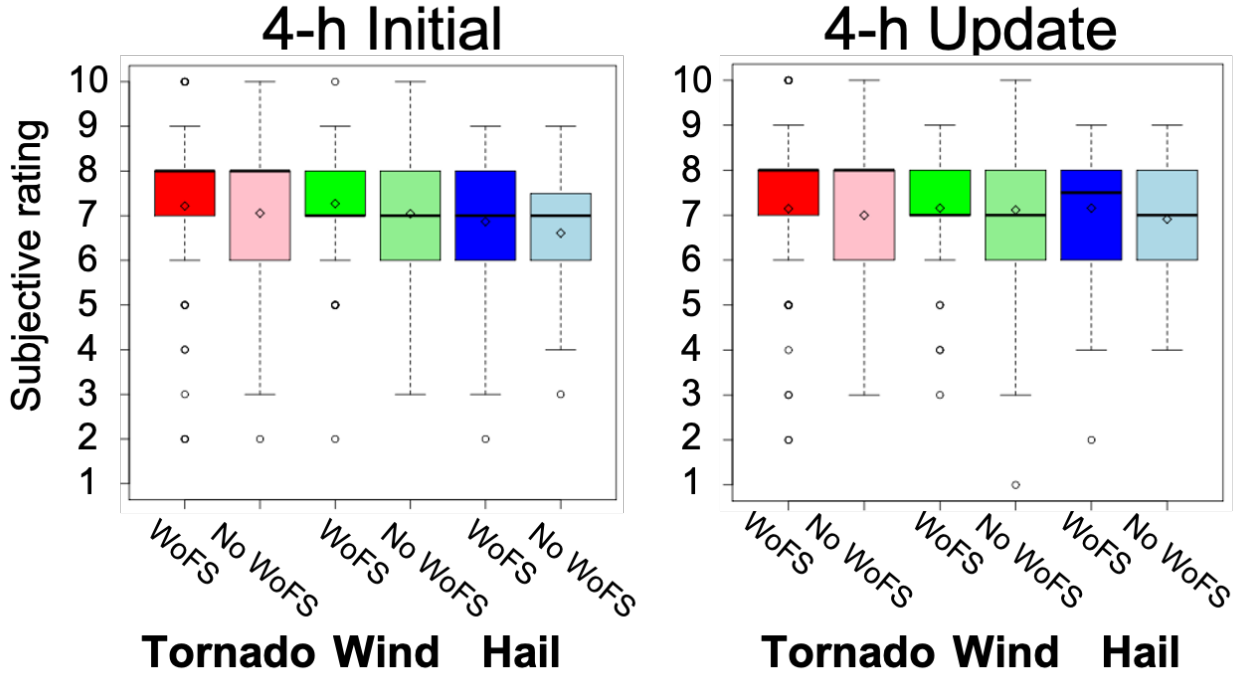


Figure 59 As in Figure 58, except for the 4-h time window outlooks.

For the 4-h time window outlooks (Fig. 59), the differences between mean subjective ratings of the WoFS and No-WoFS outlooks for both the initial and updated outlooks were larger than those of the 1-h outlooks, but still not statistically significant.

Part of the forecasting activity at the Innovation Desk was meant to serve as a proof-of-concept for a future evening activity, similar to what was done in SFE 2019 when a small group of NWS forecasters issued a series of WoFS-based outlooks until 8pm. However, prior to committing participants to a full day of forecasting with or without WoFS, we wanted to ensure that participants with WoFS had enough time in an hour to draw probabilistic individual hazard forecasts for two time periods (six forecasts total). We also wanted to ensure that the forecaster without WoFS had enough data to look at to update their forecasts based on operational CAMs such as the HRRR, and current observational trends. To that end, forecasters were asked whether or not they felt they had enough time to generate their forecasts. Most forecasters indicated that they had "Neither too much nor too little time" (Fig. 60), indicating that this activity is suitable for future SFE evening activities.

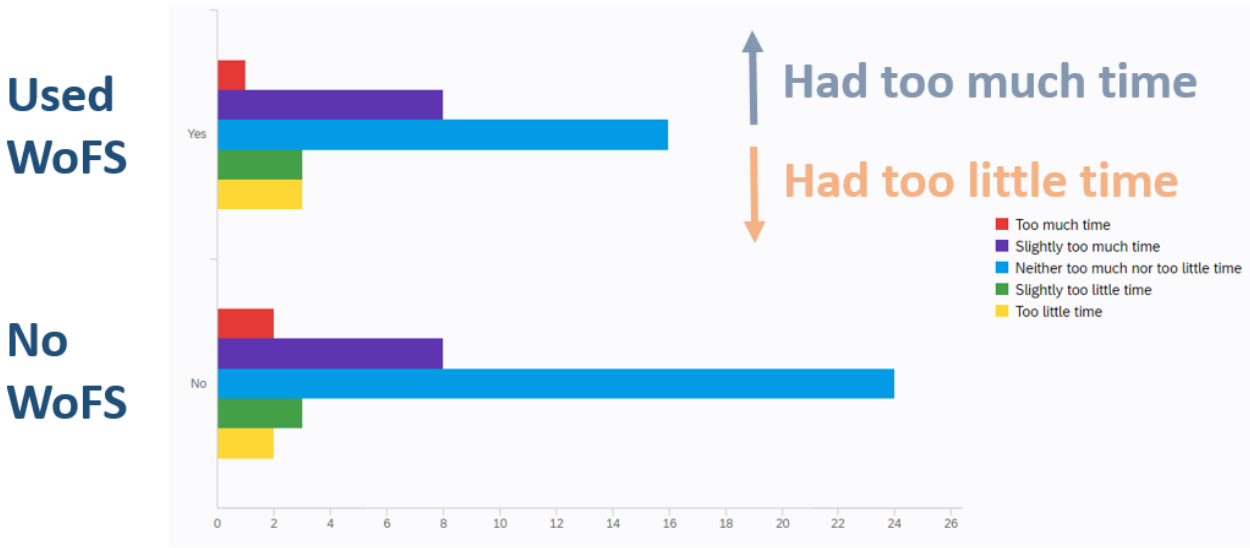


Figure 60 Participant responses to the question "Do you feel as though you had enough, too much, or too little time to complete your outlooks yesterday?", sorted by whether or not they used WoFS during their forecast process.

The participants that used WoFS found the WoFS guidance to be moderately or very useful most of the time, and they indicated that they relied more on model guidance than observations when generating their experimental forecasts. Participants were also asked a set of questions immediately after they were finished issuing their products for the day (*pre*-verification) and the next day, after they had seen how their outlooks performed (*post*-verification). In both cases, participants found the WoFS guidance to be most useful for the hail, which is most likely a function of hail being a more frequent hazard than tornadoes and wind during SFE 2020. Participants were also asked what products they found to be most useful pre- and post-verification. In both cases, hourly maximum 2-5 km above ground level updraft helicity (UH; hereafter) probabilities were recalled to be the most useful product (Fig. 61). However, after that, there were larger differences in what was most useful pre- and post-verification.

Probability products were generally the most popular, but hail products (including probabilities) were seen as more useful pre-verification. Paintball products were seen as more useful post-verification, perhaps because participants had a mental model of how the reflectivity and LSRs looked in reality, and could match it to a paintball depiction in WoFS. These preliminary results can be used to guide future evening activities, and further explore product usage in WoFS.

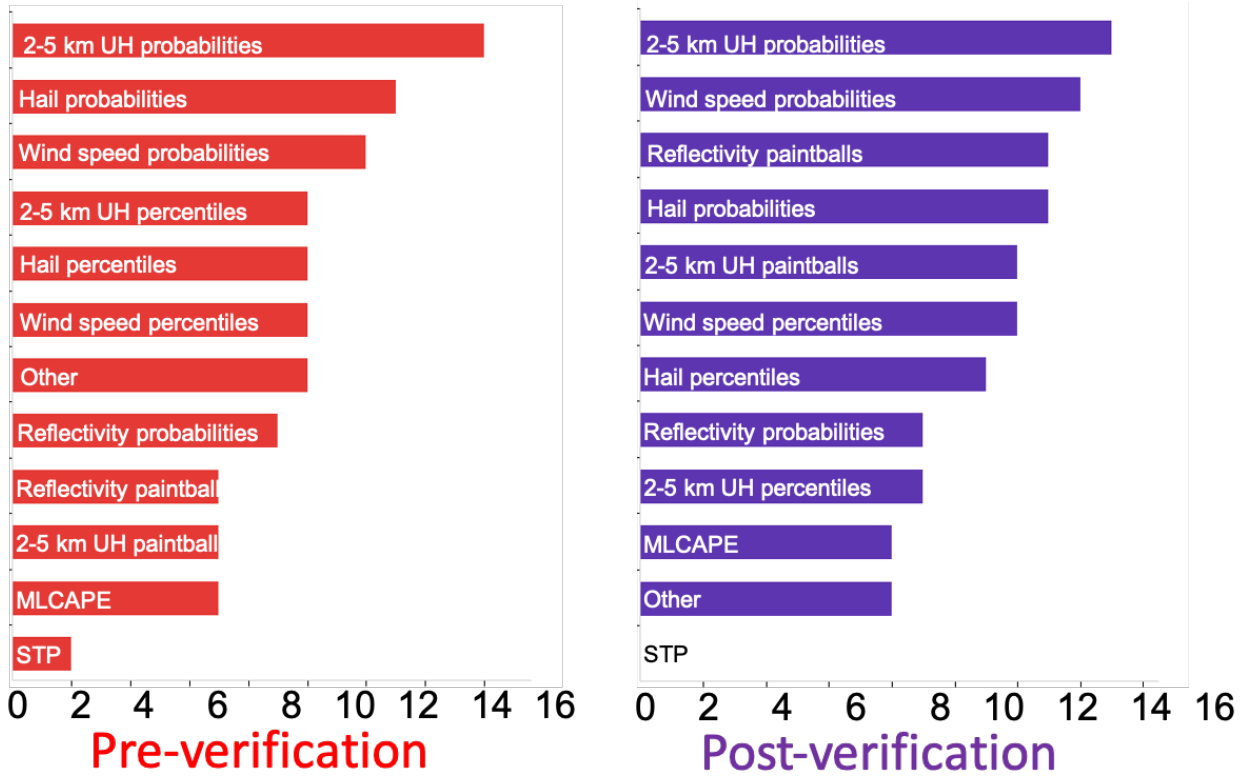


Figure 61 Participant answers to the question "Please check which WoFS product(s) you found to be most useful yesterday/today". This question was asked solely of participants taking place in the afternoon forecasting activity who used WoFS.

d) Evaluation of experimental forecast products – R2O group

At the R2O desk during the 1:30-4pm time period, participants consisting of NWS (including SPC) forecasters generated hazard coverage probabilities (tornado, hail, and wind within 25 miles) for the remainder of the Day 1 period. Additionally, conditional intensity forecasts were generated as a separate layer of the outlook for the second year. An initial set of outlooks were generated from 2-3pm without the use of WoFS data, and a final set of outlooks were updated from 3-4pm with all available data, including WoFS. The outlooks were created online using the SFE Forecast Tool. An example of a set of update forecasts (i.e., using WoFS) are shown in Figure 62 below.

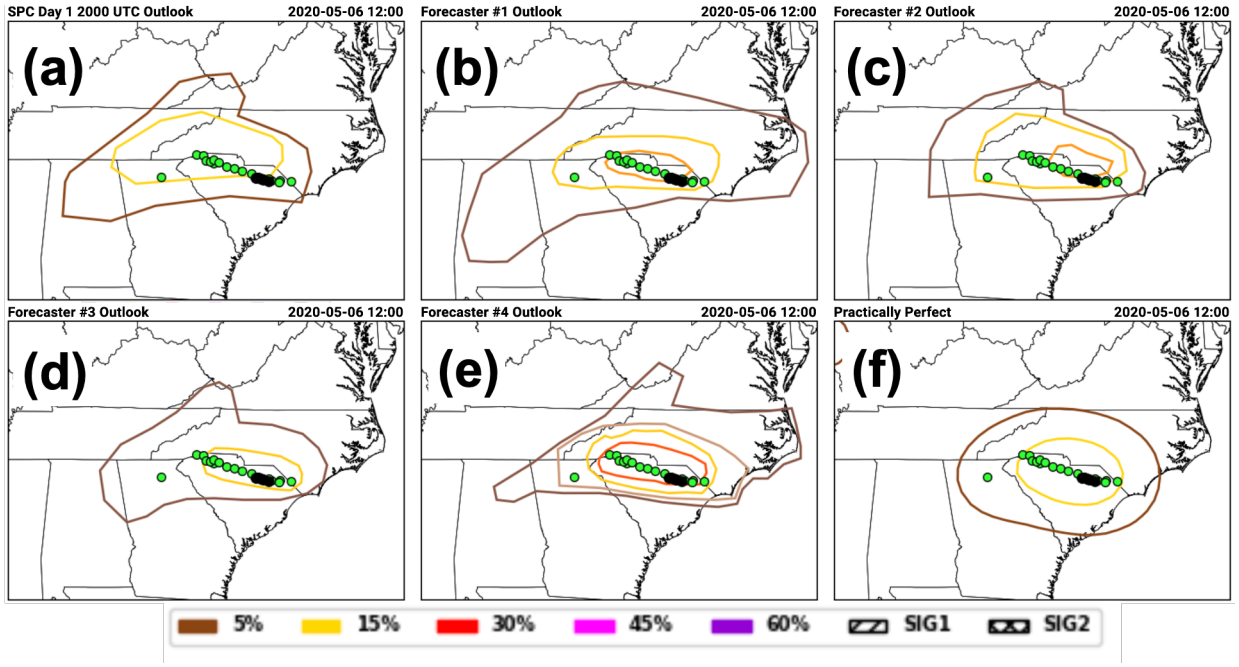


Figure 62 (a) Operational SPC Day 1 hail outlook issued 5 May 2020 at 2000 UTC. (b)-(e) Corresponding experimental Day 1 hail probability updates generated by SFE participants. (f) Practically perfect hail outlook used as a verification tool.

Each experimental outlook was rated the following day by the forecaster who generated the outlook. The primary purpose of the subjective ratings was to determine if the final, updated outlook that included use of WoFS data was an improvement over the initial outlook (i.e., without WoFS data). For most of the hazard (tornado, hail, wind) outlooks (coverage & conditional intensity), the final outlook (issued just an hour after the initial outlook) was a slight improvement (Fig. 63). The largest subjective improvements generally occurred with the hail outlooks. Overall, the forecaster comments regarding the role of WoFS in the outlook updates was to increase confidence in various aspects of the forecast: CI, location, coverage, and intensity.

The forecasters generally quickly grasped the concept of conditional intensity outlooks, though most of them were being introduced to the concept for the first time during the 2020 SFE. The forecasters most commonly cited the conditional intensity outlooks as being “neither difficult nor easy” to generate (Fig. 64). This general sentiment along with the conditional intensity outlooks often receiving higher ratings than the coverage outlooks (Fig. 63) is a promising sign of forecasters being able to successfully grasp the concept and execute the conditional intensity forecasts. Overall, the wind outlooks were deemed the most difficult to generate (Fig. 65), which is confirmed by the wind outlooks also having the lowest subjective ratings of the hazards (Fig. 63).

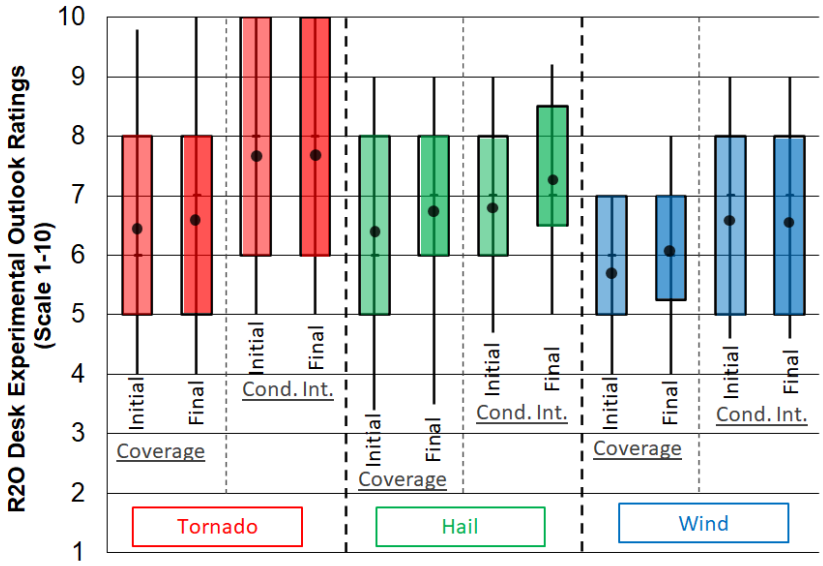


Figure 63 Distribution of subjective ratings of experimental initial/final coverage and conditional intensity outlooks issued by SFE participants for tornado (red), hail (green), and wind (blue).

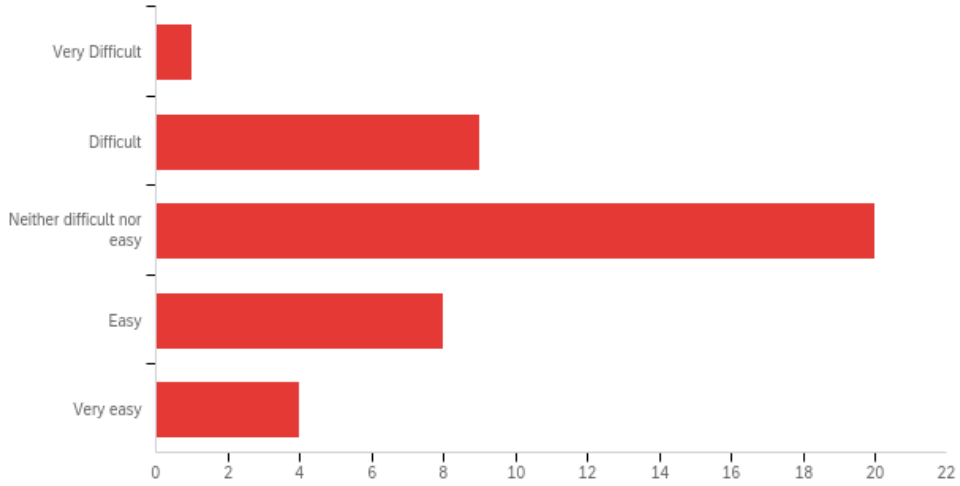


Figure 64 Subjective daily rating counts by 2020 SFE participants regarding the difficulty of drawing conditional intensity outlooks: "Very easy", "Easy", "Neither difficult nor easy", "Difficult", or "Very difficult".

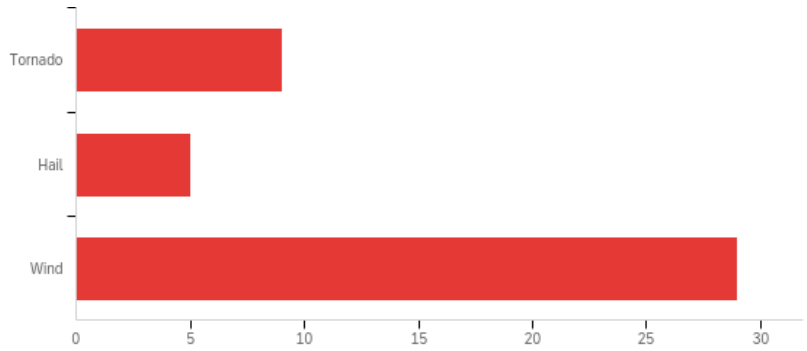


Figure 65 Subjective daily rating counts by 2020 SFE participants regarding which hazard was most difficult in drawing the outlooks.

4. Summary

The 2020 Spring Forecasting Experiment (2020 SFE) was conducted virtually from 27 April – 29 May by the SPC and NSSL with participation from forecasters, researchers, model developers, university faculty, and graduate students from around the world. The primary theme of the 2020 SFE was to 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). Furthermore, NSSL's prototype Warn-on-Forecast System was utilized in creating experimental high-temporal resolution probabilistic forecasts of severe weather hazards.

Several preliminary findings/accomplishments from the 2020 SFE are listed below:

- Used a prototype Warn-on-Forecast System (WoFS) to generate short-term individual hazard guidance, and for updating full-period hazard forecasts valid 2100-1200 UTC and corresponding conditional intensity guidance. Corresponding outlooks were also generated without using WoFS.
    - In the Innovation Group, differences between subjective ratings of WoFS and No-WoFS outlooks were most apparent for the 4-h time windows, with mean ratings slightly higher for WoFS. Participants generally found that they had enough time to issue 1- and 4-h individual hazards outlooks, motivating expansion of this activity in future experiments.
    - In the R2O group, updated Day 1 outlooks that used WoFS were given slightly higher subjective ratings than initial outlooks issued without WoFS. The majority of participants (75%) stated that generating the conditional intensity forecasts was "neither difficulty nor easy", "easy", or "very easy", and wind was the hazard most often cited as being the most difficult to generate conditional intensity forecasts.
- Examined and assessed various methods to produce first guess hazard guidance based on forecast output from HREFv2.1.
    - For tornadoes, methods combining UH with forecast STP and associated tornado climatologies ("STP Cal") performed best for 24-h periods, while the Loken et al. (2020) machine-learning-based forecast performed worst. For the 4-h time window tornado outlooks, the SPC Timing Guidance performed best.
    - For 24-h hail outlooks, the Loken et al. (2020) machine-learning-based guidance was notable for its exceptional performance, exceeding the next closest method by a full point in its mean ratings. For 4-h hail outlooks, the SPC Timing Guidance had the highest mean ratings followed closely by the Burke et al. (2020) machine-learning-based guidance.
    - For 24-h wind outlooks, mean ratings for the Loken et al. (2020) method was slightly better than HREF/SREF calibrated forecasts, and the SPC Timing Guidance performed best for the 4-h wind outlooks.
- Examined various deterministic and ensemble CAM systems within the CLUE using HREFv2.1 and HREFv3 as a baseline.

- Time-lagged ensembles based at 0000 UTC neither improved nor degraded their non-time-lagged counterparts, which motivates further investigation to determine the optimal number of members to run at a single time. HREF continues to stand as a formidable baseline for CAM ensembles, with HREFv3 receiving the highest overall ratings of all CAM ensembles.
- Ensemble sub-setting using ensemble sensitivity analysis applied to 1800 and 0000 UTC initialized HRRRE members found that subset skill remained about the same as that of the full ensemble, and that severe weather probabilities were degraded by the subset more often than they were improved. These suboptimal results may be related to inconsistencies between the Texas Tech ensemble, which was used for calculating ensemble sensitivity, and the HRRRE, which was used for the sub-setting.
- Using MODE to objectively measure skill of three hail forecasting methods applied to HRRRE found that machine-learning had the highest CSI (but a high size bias), followed in order by HAILCAST and the Thompson method. In surveys that queried participants on hail forecast verification methods, participants found location, size, and timing all important aspects of "good" hail forecasts, in that order.
- EMC FV3-SARX, which includes an updated physics package, was found to be an improvement relative to EMC FV3-SAR. NSSL FV3-SAR, which is configured with more vertical levels, had a slight improvement at earlier forecast hours relative to EMC FV3-SARX.
- Comparing pairs of FV3-SAR runs with different land-surface models found that NOAH LSM performed best. For pairs with different initial conditions, there was a large impact on forecasts, but no set of initial conditions performed best. The HORD=5 option (less diffusivity) received slightly higher ratings than HORD=6.
- In comparisons of several single- and multi-model time-lagged ensembles based at 0000 UTC, subjective ratings were very similar, demonstrating that for these configurations, multiple models did not provide an advantage. The baseline HREF configurations were superior to all of the various single- and multi-model ensemble configurations, even those that included all possible combinations of 36 members.
- In comparisons between two different time-lagging strategies for ensembles based at 1200 UTC, it was found that an ensemble configured with two sets of physics and members initialized from different HRRR initial conditions was notably superior to the HRRRE and a time-lagged HRRRE. Thus, this could be a potentially useful strategy for configuring a single-model CAM ensemble.
- Evaluations of deterministic CAMs provided by each SFE collaborator (EMC, NSSL, GFDL, UK Met, and GSL), found that GSL's HRRRv4 displayed superior performance for both convective evolution and environment fields.
- In assessments involving a matrix of CAMs with different model cores and driving models, although there was significant day-to-day variability, it was generally found that forecasts of reflectivity and UH were more sensitive to the model core, while environmental fields were more sensitive to the driving model. WRF-ARW was most often cited as the best

performing model core, while the better performing driving model was dependent on which model core the driving model was coupled with.

- Machine-learning-based algorithms were used to diagnose the likelihood that severe wind reports were actually associated with winds 50 knots. In the subjective assessments of three different approaches, the ensemble average of the three methods was rated highest and the "Stack RF" approach was rated second highest. While Brier Scores also indicated that the ensemble average performed best, the second best Brier Score was associated with a gradient boosted model (GBM).
- A neural network and random forest algorithm were trained on deterministic 3-km WRF forecasts from 2010-15 to produce hazard guidance for tornadoes, wind, and hail, with UH-based probabilities used as a baseline comparison. In general, the neural network forecasts were rated most favorably, but objective verification results revealed that the neural networks suffered from over-prediction, with the random forest forecasts having better reliability. Both sets of guidance were rated more highly than the UH baseline.
- Two versions of 3D-RTMA were subjectively evaluated to assess quality and utility for situational awareness and short-term severe weather forecasting. Both systems performed well with only slight differences in and around areas of convection or areas with limited surface observations.
- WRF simulations were examined that assimilated radar data with and without assimilation of total lightning data. Subjective evaluations revealed that most of the time there was little to no impact on forecast skill from the lightning DA, but when there was an impact, there were improvements more often than when there were degradations. Objective verification showed that the lightning DA increased precipitation biases, but also resulted in higher CSI relative to the runs without lightning DA.
- Subjective comparisons between the 3-km real-time WoFS and a 1.5-km enhanced resolution WoFS revealed very small differences in subjective ratings. Also, a 1.5-km deterministic WoFS run using a hybrid DA system performed better than a similarly configured WoFS run that used 3DVAR.

Overall, the 2020 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 2020 SFE directly promote continued progress to improve forecasting of severe weather in support of the NWS Weather-Ready Nation initiative. In subsequent years, we plan to continue exploring the potential forecasting applications of Warn-on-Forecast, continue examining strategies for CAM ensemble design, accelerate work with our partners to optimize FV3-SAR for CAM forecasting applications, and explore new ways to leverage AI-based strategies for calibrating and post-processing CAM output to aid forecasters. Additionally, we expect that this work will take on particular importance and aid with evidence-based decision making as NOAA moves forward with its plans for a Unified Forecasting System.

Acknowledgements

The 2020 SFE would not have been possible without dedicated participants and the support and assistance of numerous individuals at SPC and NSSL. In addition, collaborations with NCAR, the United Kingdom Met Office, ESRL/GSL, GFDL, and EMC were vital to the success of the 2020 SFE. In particular, Ryan Sobash (NCAR), Craig Schwartz (NCAR), David Ahijevych (NCAR), Glen Romine (NCAR), Aurore Porson (UK Met), Caroline Bain (UK Met), Steve Willington (UK Met), David Flack (UK Met), James Warner (UK Met), Mike Bush (UK Met), Nigel Roberts (UK Met), David Walters (UK Met), Mark Weeks (UK Met), Curtis Alexander (GSL), David Dowell (GSL), Jacob Carley (EMC), Eric Aligo (EMC), Lucas Harris (GFDL), Matthew Morin (GFDL), Kai-Yuan Cheng (GFDL), Becky Adams-Selin (AER), John Henderson (AER), Austin Coleman (TTU), Brian Ancell (TTU), and Christina Kalb (NCAR) were essential in generating and providing access to model forecasts or products examined on a daily basis.

References

Adams-Selin, R. D., A. J. Clark, C. J. Melick, S. R. Dembek, and I. L. Jirak, 2019: Evolution of WRF-HAILCAST during the 2014-16 NOAA/Hazardous Weather Testbed Spring Forecasting Experiments. *Wea. Forecasting*, **34**, 61-79.

Adams-Selin, R. D., and C. L. Ziegler, 2016: Forecasting Hail using a One-Dimensional Hail Growth Model within WRF. *Mon. Wea. Rev*., **144**, 4919-4939.

Ancell, B.C., 2016: Improving High-Impact Forecasts through Sensitivity-Based Ensemble Subsets: Demonstration and Initial Tests. *Wea. Forecasting*, **31**, 1019-1036.

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

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

Fierro, A. O., Y. Wang, J. Gao, and E. R. Mansell, 2019: Variational Assimilation of Radar Data and GLM Lightning-Derived Water Vapor for the Short-Term Forecasts of High-Impact Convective Events. *Mon. Wea. Rev*., **147**, 4045-4069.

Gagne, D.J., A. McGovern, S.E. Haupt, R.A. Sobash, J.K. Williams, and M. Xue, 2017: Storm-Based Probabilistic Hail Forecasting with Machine Learning Applied to Convection-Allowing Ensembles. *Wea. Forecasting*, **32**, 1819–1840.

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

Gao, J., M. Xue, A. Shapiro, and K. K. Droegemeier, 1999: A variational method for the retrieval of three-dimesnsional wind fields from dual-Dopler radars. *Mon. Wea. Rev*., **127**, 2128-2142.

Goodman, S. J., and Coauthors, 2013: The GOES-R Geostationary Lightning Mapper (GLM). *Atmos. Res*., **125-126**, 34-39.

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

Hu, J., A. O. Fierro, Y. Wang, J. Gao, and E. R. Mansell, 2020: Exploring the Assimilation of GLM-Derived Water Vapro Mass in a Cycled 3DVAR Framework for the Short-Term Forecasts of High-Impact Convective Events. *Mon. Wea. Rev*., **148**, 1005-1028.

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

Jirak, I. L., C. J. Melick, A. R. Dean, S. J. Weiss, and J. Correia, Jr., 2012: Investigation of an automated temporal disaggregation technique for convective outlooks during the 2012 Hazardous Weather Testbed Spring Forecasting Experiment. Preprints, 26th Conf. on Severe Local Storms, Nashville, TN, Amer. Meteor. Soc., 10.2.

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

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

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

Roebber, P. J., 2009: Visualizing Multiple Measures of Forecast Quality. *Wea. Forecasting*, **24**, 601-608.

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

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

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

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

Xue, M., and Coauthors, 2001: The Advanced Regional Prediction System (ARPS) – A multiscale nonhydrostatic atmospheric simulation and prediction tool. Part II: Model physics and applications. *Meteor. Atmos. Phys*., **76**, 143-165.

APPENDIX

Table A1 Weekly participants during the 2020 SFE. Facilitators/leaders for the 2020 SFE included Adam Clark (NSSL), Israel Jirak (SPC), David Imy (retired SPC), Mike Coniglio (NSSL/SPC), Burkely Gallo (CIMMS/SPC), Kenzie Krocak (CIMMS/NSSL/OU), Brett Roberts (CIMMS/SPC/NSSL), Kent Knopfmeier (CIMMS/NSSL), and Andy Dean (SPC).

Week 1 **April 27-May 1**	**Week 2** **May 4-8**	**Week 3** **May 11-15**	**Week 4** **May 18-22**	**Week 5** **May 26-29**
Lizzie Tirone (ISU)	Lizzie Tirone (ISU)	Lizzie Tirone (ISU)	Lizzie Tirone (ISU)	Lizzie Tirone (ISU)
Jeremiah Pyle (AWC)	Brice Coffer (NCState)	Bill Gallus (ISU)	Lance Bosart (Suny-Albany)	Kallan Parker (PSU; Hollings)
Victor Gensini (NIU)	Lucia Scaff (U. Sask)	Kyle Hugeback (ISU)	Bruno Ribeiro (Suny-Albany)	Clark Evans (UWM)
Ryan Sobash (NCAR)	Corey Potvin (NSSL)	Michou Baart de la Faille (KNMI)	Scott Feldman (Suny-Albany)	Dillon Blount (UWM)
Yongming Wong (OU/MAP)	Becky Adams-Selin (AER)	Tina Kalb (DTC)	Steve Weiss (SPC Ret.)	Craig Schwartz (NCAR)
Amanda Burke (OU)	Alicia Bentley (EMC)	John Allen (CMU)	Harald Richter (BoM)	Ben Blake (EMC)
Jacob Carley (EMC)	Aaron Johnson (MAP)	Glen Romine (NCAR)	Tom Galarneau (CIMMS/NSSL)	Xiaoyan Zhang (EMC)
Brett Borchardt (WFO LOT)	Andrew McKaughan (WFO PIH)	Paige Crafter (USAF)	Tony Oakley (USAF)	Austin Coleman (TTU)
Matt Anderson (WFO MRX)	Alex Lukinbeal (WFO MSO)	Logan Dawson (EMC)	Gang Zhou (EMC)	Jidong Gao (NSSL)
Alex Krull (WFO DMX)	Hayden Frank (WFO BOX)	Austin Dixon (OU)	Matt Pyle (EMC)	Jamie Wolff (DTC)
David Harrison (CIMMS/SPC)	Patrick Skinner (CIMMS/NSSL)	Austin Coleman (TTU)	Austin Coleman (TTU)	Corey Mead (WFO OAX)
Derek Stratman (CIMMS/NSSL)	Yibing Su (Princeton)	Mike Seaman (WFO SLC)	Jason Godwin (WFO FWD)	Nick Vertz (WFO BYZ)
Joe Pollina (WFO OKX)	Jeff Beck (GSL)	Eric Bunker (WFO TAE)	Tom Hultquist (WFO MPX)	Curtis Alexander (GSL)
Jeff Duda (GSL)	Terra Ladwig (GSL)	Robert Megnia (WFO LCH)	Dan McKemy (WFO LMK)	John Brown (GSL)
Dave Turner (GSL)	Nigel Roberts (UK Met)	Steve Zubrick (WFO LWX)	Mike Evans (WFO ALY)	
Aurore Porson (UK Met)		Geoff Manikin (EMC)	David Dowell (GSL)	
		John Brown (GSL)	Eric James (GSL)	
		Ed Szoke (GSL)	Mike Bush (UK Met)	
		Aurore Porson (UK Met)	Dave Ahijevych (NCAR)	
		Nigel Roberts (UK Met)		
		Bethany Earnest (CIMMS/SPC)		

Table A2 Model evaluations schedule.

Model Evaluations: Monday		
Time (CDT)	*Topic*	*Moderator*
10:00 a.m.	**Welcome and Introductions**	Israel
10:20 a.m.	**Overview of SFE Model Contributions and Scientific Goals**	Israel and PIs
11:00 a.m.	**Preview of the Evaluations** (Science Questions, Examples)	Group A: Israel & David J. Group B: Burkely & Adam
Model Evaluations: Tuesday-Friday		
9:45 a.m.	**Overview of Yesterday's Severe Weather** (David Imy) Break into Virtual Groups (A & B)	
	Group A (Israel & David J.)	***Group B (Burkely & Adam)***
10:00 a.m.	**Independent Evaluations** (with moderators available for questions)	**Independent Evaluations** (with moderators available for questions)
11:00 a.m.	**Discussion of Evaluations:** A1. ISU ML Severe Wind Probs A2. NCAR ML Hazard Guidance A3. CLUE: 00Z CAM TL-Ensemble A4. CLUE: TTU Ensemble Subsetting A5. CLUE: Ens. Hail Guidance *(Fri)* A6. CLUE: FV3-SAR Physics/DA/VL A7. CLUE: FV3-SAR IC/Hord/LSM A8. Mesoscale Analyses A9. CLUE: Lightning DA	**Discussion of Evaluations:** B1(a-f). HREF Calibrated Guidance B2. CLUE: 00Z CAM Multi-Model Ens. B3. CLUE: 12Z CAM TL-Ensemble B4. CLUE: Deterministic Flagships B5. CLUE: Core and ICs B6(a-f). WoFS Configurations

Table A3 Short-term forecasting schedule.

Short-Term Forecasting: Monday-Friday		
1:30 p.m.	**Overview of Today's Severe Weather Threat** (David Imy) Break into Virtual Groups (R2O & Innovation)	
	R2O (Israel & Mike)	***Innovation (David Imy & Adam)***
1:40 p.m.	**Overview of SFE Drawing Tool (M); Evaluation of Yesterday's Forecasts (T-F)**	**Overview of WoFS Drawing Tool (M); Evaluation of Yesterday's Forecasts (T-F)**
2:00 p.m.	**Day 1 Outlook Generation*** Full period (20-12Z) coverage and conditional intensity forecasts of tornado, hail, and wind using available 12Z CAM ensemble guidance (not WoFS) and observations.	**Short-Term Outlook Generation*^** 1-h (21-22Z) and 4-h (21-01Z) probabilistic forecasts of tornado, hail, and wind. Some forecasters with access to WoFS^ and some without*.
3:00 p.m.	**Day 1 Outlook Update*** Update full period (21-12Z) coverage and conditional intensity forecasts of tornado, hail, and wind using WoFS and observations.	**Short-Term Outlook Update*^** 1-h (21-22Z) and 4-h (21-01Z) probabilistic forecasts of tornado, hail, and wind. Same forecasters with access to WoFS^ and same without*.

* Using SFE Drawing Tool

^ Using WoFS Drawing Tool

Table A4 Description of "non-hatched" (normal), "hatched", and "double-hatch" conditional intensity forecasts for wind, hail, and tornadoes.

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