12B.3 EVALUATION OF THE RRFS ENSEMBLE FORECAST SYSTEM DURING THE 2024 NOAA HAZARDOUS WEATHER TESTBED SPRING FORECASTING EXPERIMENT

Jacob T. Vancil*
Cooperative Institute for Severe and High-Impact Weather Research and Operations (CIWRO), Norman, OK

Israel L. Jirak
Storm Prediction Center, Norman, OK

1. INTRODUCTION

Every spring, the Experimental Forecast Program (EFP) of the NOAA/Hazardous Weather Testbed (HWT) conducts the Spring Forecasting Experiment (SFE). This collaborative experiment is organized by the Storm Prediction Center (SPC) and the National Severe Storms Laboratory (NSSL). The 2024 edition of the HWT SFE (in-person and virtual) took place in the National Weather Center in Norman, Oklahoma. The 2024 SFE occurred on weekdays, excluding Memorial Day, between 29 April and 31 May 2024 (24 days).

SFE activities occurred over a seven-hour schedule each day. The first two-hour block each Tuesday through Friday consisted of the prior day (deterministic and ensemble) convective allowing model (CAM) and forecast evaluations. A suite of new and improved experimental CAM products was contributed by a large group of SFE collaborators. As in prior SFEs, all contributed CAMs where a part of an ensemble framework called the Community Leveraged Unified Ensemble (CLUE; Clark et al. 2018). Each year the CLUE is constructed by using common model specifications (e.g., grid-spacing, domain size, post-processing, etc.) so that CAM output from each contributor can be used in various controlled experiments. Additionally, the High-Resolution Ensemble Forecast system version 3 (HREFv3) and High-Resolution Rapid Refresh Version 4 (HRRRv4) were evaluated as operational modeling baselines.

Annual goals of the SFE include to accelerate research-to-operation (R2O) activities, promote operationally relevant research, and explore the the performance of CAM systems. Due to the proposed timeline of the Unified Forecast System (UFS) for future forecast model implementations and the retirement of legacy systems, the 2024 SFE had specific research objectives. A few of these objectives included, (1) evaluate the Rapid Refresh Forecast System (RRFS) deterministic run against the HRRRv4 and (2) evaluate the RRFS Ensemble Forecast System (REFS) against the HREFv3. An emphasis was also placed on accessing applications of severe weather forecasting for the REFS/RRFS and HREFv3/HRRRv4 systems.

Daily participant subjective ratings (1-10) of the various ensembles and deterministic models (00Z and 12Z cycles) were accumulated over all 24 days of the 2024 SFE. This study looks to add quantitative verification metrics, in addition to the subjective ratings, to evaluate each of the listed CAM's performance over the entire 2024 SFE.

2. DATA AND METHODS

2.1 RRFS AND REFS

The REFS, an updated version of the RRFS ensemble (SFE 2023), was created as part of the NOAA UFS initiative. The UFS is a community that includes researchers, developers, and users from NOAA, federal agencies, academia, and the private sector. This community was tasked with developing, improving, and implementing a new simplified suite of weather prediction systems for NOAA. NOAA's current suite of forecasting models consists of many independent forecast systems, each of which must be maintained and improved separately. The simplification of the model suite to a single system could increase the efficiency of future system maintenance and development.

*Corresponding author: Jacob T. Vancil, jacob.vancil@noaa.gov

Members: RRFS	ICs	LBCs	Micro-physics	PBL/SFC	LSM	Radiation	Cumulus	Dynamical Core
RRFS	RRFS hybrid 3DEnVar	GFS	Thompson	MYNN/MYNN	RUC	RRTMG	GF-deep	FV3

Members: REFS	ICs	LBCs	Micro-physics	PBL/SFC	LSM	Radiation	Cumulus	Dynamical Core
REFS01	RRFS enkf1	GEFS m1	Thompson*	TKE-EDMF/GFS	RUC*	RRTMG*	GF-deep*+sh	FV3
REFS02	RRFS enkf2	GEFS m2	Thompson*	MYNN*/MYNN*	RUC*	RRTMG*	saSAS deep	FV3
REFS03	RRFS enkf3	GEFS m3	NSSL#	MYNN*/MYNN*	RUC*	RRTMG*	GF deep*	FV3
REFS04	RRFS enkf4	GEFS m4	NSSL#	TKE-EDMF/GFS	RUC*	RRTMG*	GF-deep*+sh	FV3
REFS05	RRFS enkf5	GEFS m5	NSSL#	MYNN*/MYNN*	RUC*	RRTMG*	saSAS deep	FV3

Members: REFS	ICs	LBCs	Microphysics	PBL	dx (km)	Vertical Levels	HREF hours
HRRRv4	HRRRDAS	RAP -1h	Thompson	MYNN	3.0	50	0 – 48

Fig. 1. SFE 2024 version of the REFS ensemble member configurations.

The REFS is a fourteen-member, time-lagged ensemble. The RRFS is the control member of the REFS where an additional five members are created from perturbations of the RRFS initial conditions (ICs). The HRRRv4 is also included in the REFS ensemble. The remaining seven members consist of 6-h time-lagged duplicates of the previously mentioned members. Further REFS member configurations (ICs, LBCs, etc.) are shown (Fig 1). A major difference between the REFS and HREFv3 ensembles is the inclusion of deep convection parameterization schemes in all REFS members (excluding the HRRRv4). The RRFS and three perturbed REFS members use the Grel-Freitas (GF) deep convection parameterization scheme while the remaining two perturbed REFS members use an updated scale-aware Simplified Arakawa-Schubert (saSAS) scheme. This is the first instance of a CAM ensemble being evaluated in the SFE with parameterized deep convection. With the UFS's intent for the REFS to replace the currently operational HREFv3, evaluation of the REFS during the 2024 SFE was a major objective.

2.2 HREF

The HREFv3 is a ten-member, multi-dynamical core, mixed-physics, and time-lagged ensemble. The five non-time-lagged members consist of the High-Resolution Window Advanced Research version of the Weather Research and Forecast Model (HRW ARW), the HRW NSSL model, the North American Mesoscale (NAM) CONUS Nest, the HRW Finite Volume Cubed Sphere (FV3) model, and the HRRRv4. The remaining five members consist of 12-h time-lagged duplicates of the HRW members and NAM CONUS Nest, and the 6-h time-lagged initialization of the HRRRv4. Further details on each member's configuration is also provided (Fig. 3). This study performed verification of the composite reflectivity (REFC) variable from the HREFv3.

HREFv3	ICs	LBCs	Microphysics	PBL	dx (km)	Vertical Levels	HREF hours
HRRRv4	HRRRDAS	RAP -1h	Thompson	MYNN	3.0	50	0 – 48
HRRRv4 -6h	HRRRDAS	RAP -1h	Thompson	MYNN	3.0	50	0 – 42
HRW ARW	RAP	GFS -6h	WSM6	YSU	3.2	50	0 – 48
HRW ARW -12h	RAP	GFS -6h	WSM6	YSU	3.2	50	0 – 36
HRW FV3	GFS	GFS -6h	GFDL	EDMF	3	50	0 – 60
HRW FV3 -12h	GFS	GFS-6h	GFDL	EDMF	3	50	0 – 48
HRW NSSL	NAM	NAM -6h	WSM6	MYJ	3.2	40	0 – 48
HRW NSSL -12h	NAM	NAM -6h	WSM6	MYJ	3.2	40	0 – 36
NAM CONUS Nest	NAM	NAM	Ferrier-Aligo	MYJ	3.0	60	0 – 60
NAM CONUS Nest -12h	NAM	NAM	Ferrier-Aligo	MYJ	3.0	60	0 – 48

Fig. 3. SFE 2024 version of the HREFv3 ensemble member configurations.

2.3 MULTI-RADAR/MULTI-SENSOR SYSTEM

The Multi-Radar/Multi-Sensor System (MRMS), developed at the NSSL, is a fully automated system that quickly integrates data from multiple radars, surface and upper air observations, lightning detection systems, satellite observations, and forecast models. MRMS products aid in the detection of severe weather hazards (tornado, wind, and hail), precipitation estimations, convection, and several other products (Zhang et al. 2011). MRMS hourly merged composite reflectivity (quality controlled; REFC) was used as the observational dataset for the REFC verification. A 40 dBZ REFC threshold is frequently associated with convective storms, making it suitable to be used in this study for evaluating ensemble skill in severe weather forecasting applications.

2.4 METPLUS

The Model Evaluation Tools (MET) software was developed by the Developmental Testbed Center (DTC) with support from the 557th Weather Wind of the United States Air Force, NOAA, and the National Center for Atmospheric Research (NCAR). MET is designed to be a customizable suite of verification tools. MET's goal is to provide a framework for reproducible verification methods and results. METplus, contains the core framework and tools of MET with additional python wrappers to improve automation and functionality. METplus will also be integrated into the UFS framework as an important tool for verification. This study specifically used MET V10.0 and METplus V5.0.

For this study, deterministic models were evaluated using the METplus GridStat tool and ensemble systems were evaluated using the METplus EnsembleStat tool. All verification metrics were calculated for each hour of the day by convective day (12z – 12z) throughout the SFE. Verification metrics were then accumulated over each hour to produce statistics over the entire SFE period. The full period statistics were calculated by using the raw hourly 2x2 contingency table results.

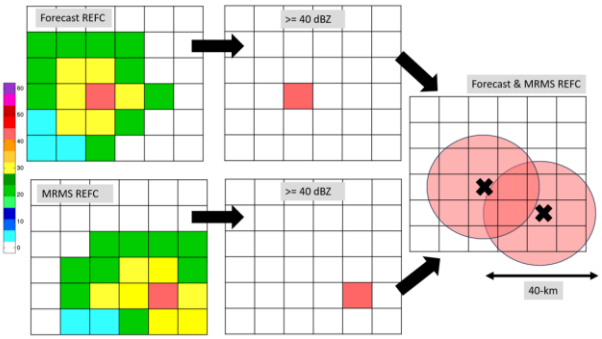

Fig. 4. Simplified diagram of how REFC forecast and MRMS observation fields are processed before verification.

Deterministic models were evaluated using a binary 40-km circular neighborhood with a >=40 dBZ REFC threshold. Ensemble systems were also evaluated using a 40-km circular neighborhood with the same >=40 dBZ REFC threshold. A neighborhood maximum ensemble probability (NMEP) of REFC was used for the verification of each ensemble system. A simplified example of how forecast and observation data were processed is shown (Fig. 4). Both deterministic and ensemble forecast systems used the MRMS merged composite reflectivity as the observational dataset.

3. RESULTS

3.1 DETERMINISITC CAM VERIFICATION

Examining the verification metrics of each REFS and HREFv3 member (00z cycle) reveals interesting REFC performance characteristics. The HRRRv4, an ensemble member of both the HREFv3 and the REFS, is the clear top performer compared to all other ensemble members. Even the 6-hr time lagged HRRRv4 was found to outperform all other ensemble members (time lagged and non-time lagged members) in terms of CSI. Other members of the HREFv3 and the REFS (excluding the HRRRv4) were shown to have similar skill metrics (CSI). However, the bias characteristics of each ensemble's members were found to differ. HREFv3 members were shown to have a bias of >1, while REFS

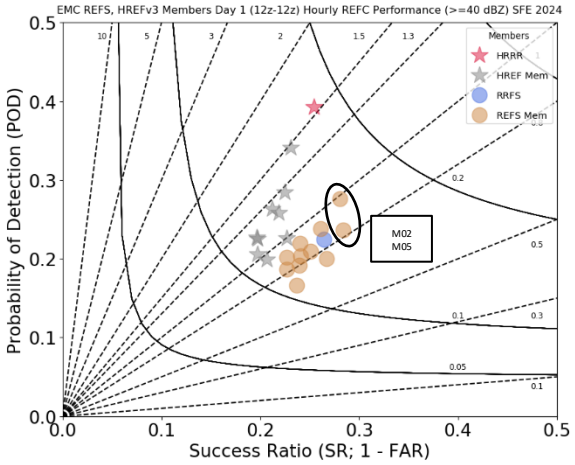

Fig. 5. RRFS and HREFv3 performance diagram over the entire 2024 SFE period. REFS members are shown by the circles. HREFv3 members are shown by the stars. HRRRv4 is the red star. RRFS is the blue star. REFS M02, M05 are circled.

members (excluding the HRRRv4) were shown to have a bias of <1 (Fig. 5). Similar bias metrics for ensemble members of the HREFv3 and the REFS were found for the 12z cycles of all members (not shown).

When comparing the HRRRv4 to the RRFS (i.e., REFS control member) the HRRRv4 outperformed the RRFS in terms of CSI. Looking closer, much of the advantage of the HRRRv4 comes from an increased probability of detection (POD) compared to the RRFS (Fig. 5). Both deterministic models were found to have a similar false alarm ratio (FAR). Similar results for the HRRRv4 versus the RRFS were also found for the 12z cycle of each model (not shown).

It was also found that the RRFS control member was not the best performing member of the REFS, which is not a desirable characteristic of the REFS design. When excluding the HRRRv4, the best performing members of the REFS were members two (M02) and five (M05) (Fig. 5). These two members (M02 and M05) were the

only REFS members that used the saSAS deep convection parameterization scheme (other members used the GF convection scheme).

SFE participants were tasked with daily rating (1-10) deterministic models based on each model's REFC field compared to observations (MRMS REFC). Participant's subjective ratings were accumulated across all SFE days to rank each model in terms of subjective rating. The HRRRv4 received the highest mean rating from participants during the 2024 SFE. When compared to the mean rating of the RRFS, the HRRRv4 had statistically significantly higher mean ratings. Additionally, both the 25th and 75th percentile of subjective ratings for the HRRRv4 were higher than the RRFS (Fig. 6).

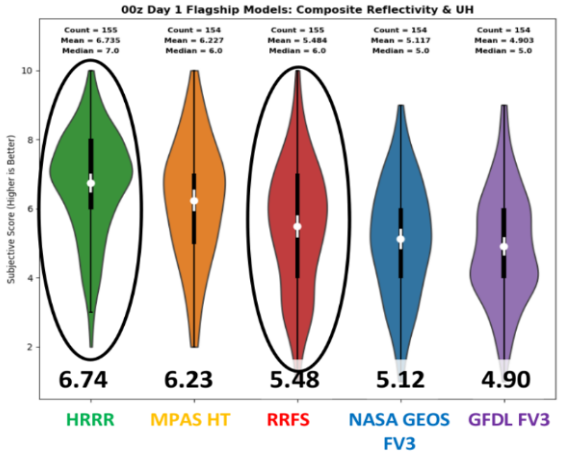

Fig. 6. The subjective ratings of the Day 1 00z cycle deterministic flagship models from 2024 SFE. The RRFS is shown as the red violin. The HRRRv4 is shown as the green violin.

3.2 CAM ENSEMBLE VERIFICATION

The 2024 SFE verification metrics of the HREFv3 and REFS showed intriguing results, especially following the deterministic RRFS control member results. Performance metrics for each ensemble were calculated for 10% bins (i.e.,10% - 100%) of the NMEP REFC field (>=40 dBZ). Performance metrics for both the HREFv3 and REFS were shown to lie along the same performance curve. The primary differences in performance metrics occurred in the POD and FAR space, particularly for lower NMEP intervals (<50%) (Fig. 7). In terms of forecast reliability, the REFS actually had better reliability than the HREFv3 at the lower REFC NMEPs (<50%). Above the 50% forecast NMEP, the REFS and HREFv3 were very similar in their forecast reliability (Fig. 8).

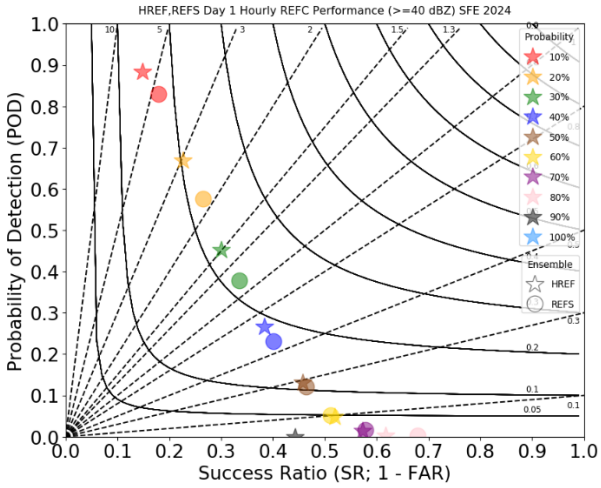

Fig. 7. REFS and HREFv3 performance diagram over the entire 2024 SFE period. The REFS is shown in the circle shapes. The HREFv3 is shown in the star shapes. Each color represents a NMEP ranging from 10% - 100%.

SFE participants also were asked to rate (1-10) the REFC NMEP field of both the HREFv3 and the REFS compared to MRMS observations (>=40 dBZ contour). Participant ensemble subjective ratings were accumulated over the entire SFE. Participant REFC ratings for each ensemble were found to be similar with a slight advantage to the HREFv3 (Fig. 10; red box).

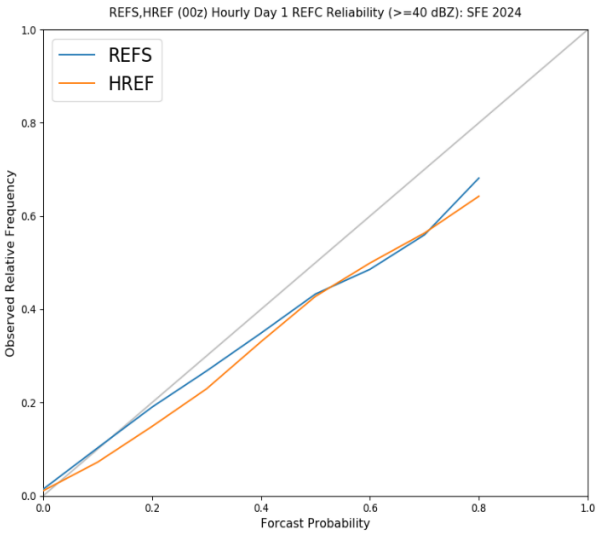

Fig. 8. REFS and HREFv3 reliability diagram. The REFS is shown with the blue line. The HREFv3 is shown with the orange line.

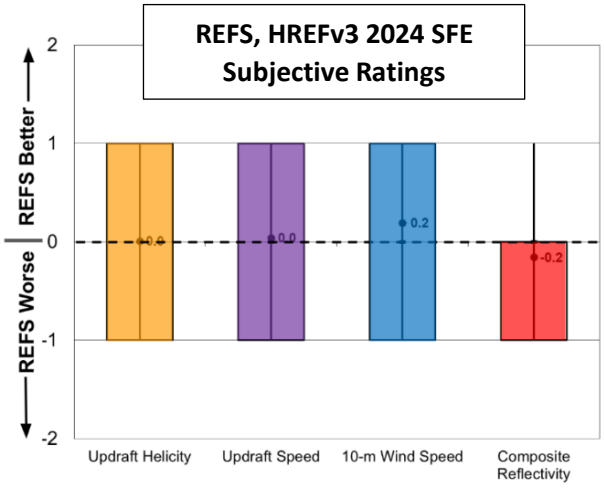

Fig. 9. The subjective ratings of the day 1 00z REFS and HREFv3 from the 2024 SFE. Ensemble variables are highlighted by color. Subjective REFC ratings are shown with the red bar. Values below 0 indicate the HREFv3 was rated higher. Values above 0 indicate the REFS was rated higher.

4. DISCUSSION

The REFC verification of the HREFv3 and REFS, and their respective members, provided interesting results. While individual member REFC performance between these two ensembles was similar in terms of CSI, the HRRRv4 was a clear top performer among all members. Also of note was the individual bias characteristics of each ensemble's members. REFS members (except the HRRRv4) were shown to have a bias of <1.0, while HREF members had a bias of >1.0 (Fig. 5). The low REFC bias found with REFS members is likely due to the use of deep convection parameterization schemes for each member.

Throughout the SFE, participants and facilitators noted several cases, particularly with the RRFS control member, where the deep convection scheme negatively impacted the REFC forecast. The RRFS missed several high REFC (>40 dBZ) areas on High or Moderate risk (SPC categorical outlooks) days during the SFE period. These misses occurred across all types of storm modes (linear, supercell, single cell, etc.) and typically showed areas of light reflectivity instead of the high (>40 dBZ) REFC values that occurred with observations (Fig. 10). Suppression of the deep convection in the RRFS also seemed to occur across differing forecast hours. In comparison, the HRRRv4, while not producing perfect REFC forecasts, at least showed the possibility of deep convection over the areas where severe hazards occurred (Fig. 10).

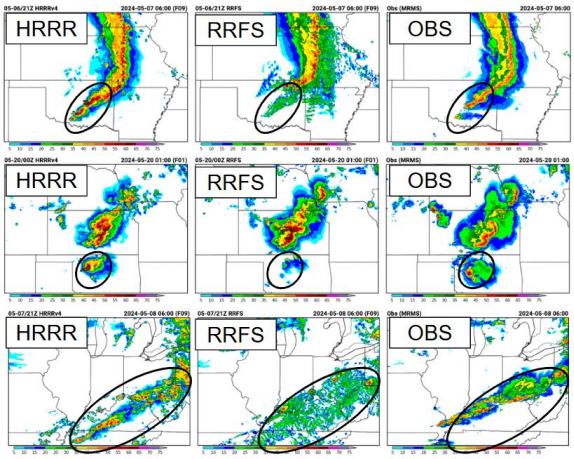

Fig. 10. REFC from the SFE 2024. Row one – HRRRv4, RRFS (21Z) valid 7 May 2024 06Z (F09). Row two- HRRRv4, RRFS (00Z) valid 20 May 2024 01Z (F01). Row three – HRRRv4, RRFS valid 8 May 2024 06Z (F09).

The REFS members using the GF deep convection schemes were shown to perform worse than REFS members using the saSAS convection scheme. During the SFE several high-end severe weather events highlighted this performance discrepancy between REFS members. The 7 May 2024 (02Z) Barnsdall tornado (EF4) was very well forecast with REFS M02, where the RRFS was not suggesting any discrete cells ahead of the linear system. The RRFS was suppressing the deep convection over northeastern Oklahoma only showing low REFC (<25 dBZ) was being forecast (Fig. 14). Similarly, REFS M05 forecast the tornadic outbreak on 8 May 2024 better than the RRFS. REFS M05 was forecasting higher (>=40 dBZ) REFC values over the correct area showing the possibility of severe hazards for this day (Fig. 12). The RRFS again failed to develop deep convection over the area of interest, entirely missing the tornadic event on this day.

The use of deep convection parameterization schemes in REFS members seemed to lessen the over forecast bias of previous CAMs. However, this negatively impacted the

probability of detection (POD) of REFC values >=40 dBZ. For severe weather forecasting a higher POD is typically desired as the penalty for over forecasting a severe event is not as impactful as missing the event entirely. A comparison of the HRRRv4 and the RRFS also shows a nearly identical false alarm ratio (FAR) between the two models, but the HRRRv4 greatly outperforms in terms of POD. These full-period objective statistics were also observed during individual forecast days within the SFE and were reflected in SFE participant subjective ratings.

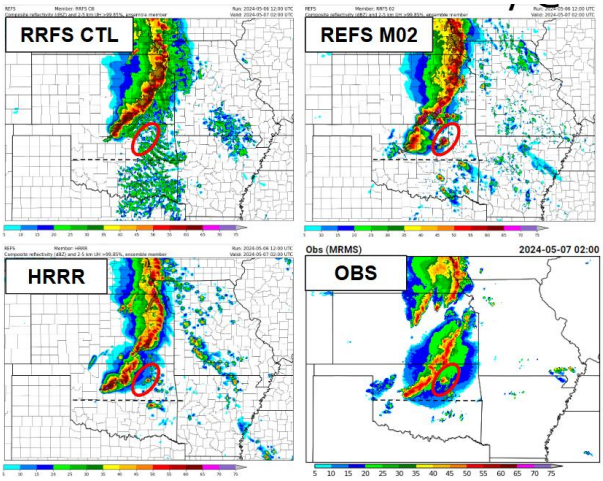

Fig. 11. REFC from RRFS, REFS M02, HRRRv4, and MRMS observations. All model cycles are 12Z (6 May) valid for 02Z 7 May 2024 (F14).

Even with the REFS members' use of deep convection schemes and the superior performance of the HRRRv4, the overall performance of the HREFv3 and REFS ensembles were similar. The impact of using parameterized deep convection may be seen on the REFC NMEP performance diagram, showing lower POD from the REFS at <50% NMEP. However, all NMEPs lie on the same performance curve further highlighting the two ensembles similar performance. The 2024 SFE version of the REFS was also the first ensemble to outperform the HREFv3 in terms of reliability at any NMEP interval (Fig. 8). However, it should be noted that the HRRRv4 is an ensemble member of both the HREFv3 and the REFS. Ongoing work (not shown), has shown the HRRRv4 to be a large contributor to ensemble performance metrics (CSI, POD) for both the REFS and HREFv3, potentially leading to the similar ensemble performance that was found.

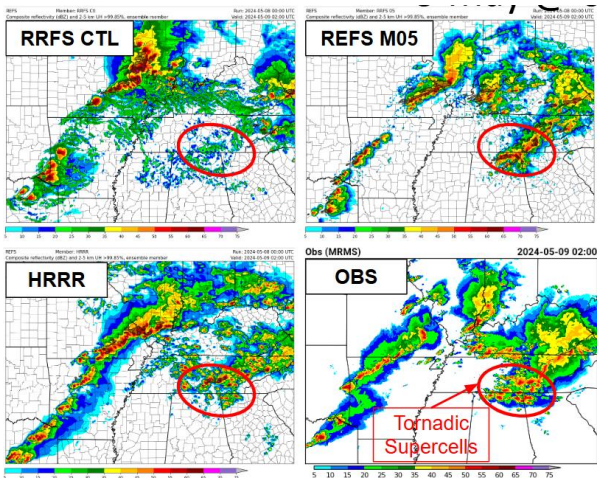

Fig. 12. REFC from RRFS, REFS M05, HRRRv4, and MRMS observations. All model cycles are 00Z (8 May) valid for 02Z 9 May 2024 (F26).

5. SUMMARY

The 2024 SFE successfully provided subjective verification of the HREFv3 and REFS ensembles, as well as key deterministic models. This study objectively assessed these CAM and CAM ensembles on applications for severe weather forecasting. Composite reflectivity verification metrics of each ensemble member (00Z) were provided. While the HRRRv4 was found to greatly outperform the RRFS, the performance of the REFS was shown to be comparable to that of the HREFv3. The use of deep convection parameterization schemes in each of the REFS members (except the HRRRv4) was found to negatively affect REFC (>=40 dBZ) forecasts, especially those using the GF scheme. REFS members using the saSAS deep convection scheme performed better than those using the GF scheme. Switching the RRFS (REFS control member) deep convection scheme to the saSAS scheme has been shown to provide a boost in REFC forecast performance in retrospective testing following the SFE. Given the REFC forecast performance of the RRFS, compared to the HRRRv4, model configuration adjustments might be necessary before use in operational severe weather forecasting. However, the current performance

of the REFS appears to be comparable to the HREFv3, at least in terms of REFC (>=40 dBZ) forecasts, especially because it includes two HRRRv4 members.

The objective verification of these forecast systems is essential for the advancement of the UFS goals. The annual SFE also plays an important role in assessing experimental ensembles in real-world severe hazard forecasting applications. The combination of the SFE subjective ratings and this study's objective verification metrics aim to provide feedback and potential concerns regarding potential model implementations and retirements, including implementation timelines. Further evaluation of these forecast systems will be necessary to advance and improve future model development.

REFERENCES

Clark, A. J., et al., 2018: The Community Leveraged Unified Ensemble (CLUE) in the 2016 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. Bull. Amer. Meteor. Soc., 99, 1433–1448, https://doi.org/10.1175/BAMS-D-16-0309.1.

Zhang, J., and Coauthors, 2011: National Mosaic and Multi-Sensor QPE (NMQ) system: Description, results, and future plans. Bull. Amer. Meteor. Soc., 92, 1321–1338, https://doi.org/10.1175/2011BAMS-D-11-00047.1.

Prestopnik, J., J. Opatz, J. Halley Gotway, T. Jensen, J. Vigh, M. Row, C. Kalb, H. Fisher, L. Goodrich, D. Adriaansen, M. Win-Gildenmeister, G. McCabe, J. Frimel, L. Blank, T. Arbetter, 2022: The METplus Version 5.0.2 User's Guide. Developmental-Testbed Center, https://github.com/dtcenter/METplus/releases.

Clark, A., Jirak, I., et. al., 2024: Spring Forecasting Experiment 2024 conducted by the Experimental Forecast Program of the NOAA/Hazardous Weather Testbed – Program Overview and Operations Plan. NOAA/NWS/NCEP Storm Prediction Center, NOAA/OAR National Severe Storms Laboratory, Cooperative Institute for Severe and High-Impact Weather Research and Operations, University of Oklahoma School of Meteorology, Institute for Public Policy Research and Analysis.