Comparing OU MAP Real-Time Convection Allowing Ensemble Forecasts Produced During 2021 & 2022 Using Neighborhood & Surrogate Verification Methods

Brett A. Castro[1], Nicholas A. Gasperoni[2,3], and Xuguang Wang [2,3]

*[1]National Weather Center Research Experiences for Undergraduates Program
Norman, Oklahoma*

*[2]School of Meteorology, University of Oklahoma,
Norman, Oklahoma*

*[3] Multiscale data Assimilation and Predictability (MAP) Laboratory
Norman, Oklahoma*

ABSTRACT

During the peak of the 2021 & 2022 Severe Weather season in the Midwest, the University of Oklahoma (OU) Multiscale Data Assimilation and Predictability (MAP) Laboratory ran two Rapid Refresh Forecast System (RRFS)-like systems within the Hazardous Weather Testbed (HWT) in an effort to test their accuracy in forecasting the development and evolution of convection. The model, referred to as the FV3-LAM consists of 10 ensembles, and was initialized at 00z between the first Monday of May and first Monday of June for both 2021 & 2022. This study will analyze the forecasting skill of the model by comparing real time observations with model simulations. Several methods are used to quantitatively examine accuracy. The standard Neighborhood Method, in which an arbitrary radius is chosen containing a set number of grid points, can be used to compare the precipitation and reflectivity within the model to observations during the same period. Additionally, a Surrogate Severe Method is also used, which maps helicity tracks generated within the model and compares them with observed storm reports. From these methods, a Fractions Skill Score (FSS) can be calculated and gives a quantitative measurement of forecast accuracy. Promising trends in model accuracy were observed between the two years, with average skill scores in 2022 outperforming 2021 across most periods. Results from the standard Neighborhood method support improvement in both placement of convection and precipitation. Conclusions based on the surrogate severe method were less concrete and require further study.

1. INTRODUCTION

Over the last decade, there has been substantial progress in short-range forecasting of convection owing to the invention and advancement of numerous convection allowing models (CAMs). Unlike global models, CAMs typically run at resolutions ≤ 4 km on regional domains that allows them to resolve small-scale features such as individual storms (e.g., Gasperoni et al. 2023). Thus, these models can simulate the development and evolution of convection over short (usually less than 48 h) time scales. This helps forecasters to not only identify where storms are most likely to develop, but also to attempt to predict storm modes and severe hazards associated with them. In fact, the ability of CAMs to provide specific information about convective properties such as storm initiation, modes, motion, and intensity has been clearly demonstrated (e.g., Schwartz et al. 2009; Sobash et al. 2011). One such operational CAM demonstrative of this progress is the High-Resolution-Rapid-Refresh model (HRRR; Dowell et al. 2022), which became operational at the National Center for Environmental Protection (NCEP) in 2014. The HRRR is a convection-allowing implementation of the Advanced Research version of the Weather Research and Forecasting (WRF-ARW) Model, which features a 3km grid covering the continental US (CONUS) designed for predicting the short-term evolution of high-impact convection and precipitating systems (Dowell et al. 2022).

Corresponding author address: Brett Castro, Department of Atmospheric and Environmental Sciences – University at Albany NY, Email: bacastro@albany.edu

Although operational CAMs such as the HRRR are a significant milestone for numerical weather prediction (NWP), these CAMs have been limited to deterministic runs. However, within the last few years, increased computational power has allowed for ensemble systems such as the High-Resolution Ensemble Forecast System (HREF e.g., Roberts et al. 2020) to become operational. The HREF contains 10 ensemble members with varying model cores and has been demonstrated to be among the best CAM ensemble systems, producing incredibly reliable probabilistic forecasts (Clark et al. 2019; Roberts et al. 2020).

Considering these advancements, CAMs will play an increasingly important role in producing more detailed, short-range forecasts for predicting severe, high-impact hazards including tornadoes, strong winds, and large hail. However, it is not efficient to continually maintain and upgrade multiple modeling systems that each contain different dynamical cores, physics schemes, grids, etc (e.g., Gasperoni et al. 2023). Further, the NOAA vision of a Unified Forecast System (UFS; https://ufscommunity.org) is attempting to unify Earth modeling systems across a variety of disciplines and scales in order to maximize the collective efforts of the scientific community to facilitate faster scientific progress. The NWP model chosen for UFS is the Finite Volume Cubed Sphere (FV3; Harris et al. 2013), which was recently implemented into the Global Forecast System (GFS) in 2019. The Rapid Refresh Forecast System (RRFS) is the next generation regional ensemble CAM system with the UFS framework that is expected to replace the HRRR within the next few years with the FV3 Limited Area Model (FV3-LAM; Black et al. 2021). However, use of the FV3-LAM at convective allowing resolutions is still in its infancy, with further development and studies required to improve the performance of predicting convective systems across the US.

Considering the need for further research in this regard, many research groups have tested FV3-based systems within the Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (SFEs) over the last several years during the May-June peak of severe weather season. The University of Oklahoma (OU) Multiscale Data Assimilation and Predictability (MAP) Laboratory ran two RRFS-like systems for 2021 and 2022 SFEs consisting of 10 ensemble members each initialized at 0000 UTC each evening for each SFE. The main objective of this study is to objectively verify and compare the accuracy of these ensemble CAM systems for 10 select high-impact cases for each year 2021 and 2022. We hypothesize that the performance of the next-day (12-36 h) forecasts from 2022 have improved compared to the 2021 system due to several upgrades in the ensemble prediction system. Finally, the overall performance impact of increasing or decreasing the number of ensembles for each year will be analyzed. This should provide valuable insight into the overall performance of the model and help to determine the influence of ensemble size on forecast accuracy. Considering the plan for the RRFS to become operational as a replacement for current CAM systems within the coming years, it is important to quantify the skill of experimental systems to mark their progress.

Measurements of model accuracy across these scenarios will be achieved using several well-established verification techniques. The first is a standard neighborhood-based verification for measuring precipitation or reflectivity coverage within both real-time model simulations and radar-based observations. As described in Schwartz et al. (2010), this method works by spatially averaging the number of points within a defined circular neighborhood whose forecast exceeds a given threshold. This process produces a neighborhood probability that can be compared between model simulations and observations, thereby limiting the influence of high-amplitude small-scale displacement errors (referred to as the “double-penalty” effect) in the metrics. A variation of the standard neighborhood technique will also be used to verify model-derived severe weather hazards. This surrogate severe method will compare model generated updraft helicity, which can be upscaled from intense storms into surrogate severe reports, to observed wind, hail, and tornado severe reports taken over the same period. The effectiveness of this method at verifying CAM forecast accuracy involving intense storms was demonstrated first in Sobash et al. (2011, 2016). Finally, Fractions Skill Score (FSS; Roberts and Lean 2008) is computed to assess model accuracy for different variables.

The remainder of this study is organized as follows. Section 2 describes the neighborhood probabilistic methods for verifying ensemble reflectivity, precipitation, and model-derived surrogate severe reports as well as the FSS metric. Next, section 3 evaluates the performance of the OU MAP ensemble systems from 2021 and 2022 across ten cases using these neighborhood techniques. A comparison between both years will also be made using the surrogate severe method to wrap up section 3. Finally, a summary and discussion of the important findings is given in section 4.

2. METHODS

a. OU MAP Ensemble Forecast Configurations

The basis for these comparisons will be two separate ensemble systems run by OU MAP within HWT SFEs during both 2021 and 2022. During each year an ensemble forecast system with 10 members was run each weekday, initialized at 0000 UTC. These 3 km forecasts ran out to 36 h lead times using the FV3-LAM model core. These ensembles were meant to mimic an RRFS-like system, including hourly ensemble data assimilation of conventional and radar reflectivity observations, to test its forecasting capabilities on convective scales. Additionally, several changes were made to the ensemble system between the two years in an attempt to improve accuracy in 2022. This included general upgrades to the model and physics versions, the implementation of stochastic physics to increase ensemble spread during the free forecast, and modifications to parameters controlling data assimilation.

To systematically compare ensemble forecast performance between these two years, 10 high-impact cases were chosen from each year. These cases were chosen based on overall convective activity and are listed in Table 1 along with their associated Storm Prediction Center (SPC) 1200 UTC day 1 risk levels and total 24 hour storm reports.

TABLE 1: Case dates chosen for 2021 and 2022 along with 12 UTC Day 1 SPC outlook risk level and total daily severe reports (wind, hail, & tornadoes) received for each day.

Date 2021	SPC Risk	Reports	Date 2022	SPC Risk	Reports
May 10	Slight	123	May 2	Enhanced	116
May 11	Marginal	71	May 5	Enhanced	137
May 14	Slight	66	May 9	Slight	140
May 17	Moderate	136	May 10	Slight	110
May 18	Slight	68	May 11	Enhanced	201
May 24	Slight	53	May 12	Enhanced	565
May 25	Slight	98	May 17	Enhanced	127
May 26	Moderate	643	May 18	Slight	158
May 27	Enhanced	158	May 24	Slight	174
May 28	Slight	158	May 31	Enhanced	141

b. Standard Neighborhood Verification

To achieve our goal of measuring model performance involving these cases as accurately as possible, we will use several verification techniques that have been widely adopted for use in analyzing CAM accuracy. With the advent of models with much finer grid resolutions, it quickly became apparent that standard verification methods that had worked well for course resolution, global models would no longer be sufficient. For example, a traditional method of measuring model accuracy is to quantify ensemble probability (EP) as the ensemble average of binary probability (BP) fields for each ensemble member (e.g. Schwartz et al. 2010). Typically, a particular exceedance threshold (q) is chosen for a given field such as accumulated precipitation. This process results in a BP of 1 (0) wherever the threshold is exceeded (not exceeded). The EP field is then computed by averaging at each grid point the BP of all ensemble members. This methodology works well for traditional verification measures (e.g. root mean square error, contingency table metrics) for coarser synoptic and global forecast systems. However, at the fine resolutions of CAMs these traditional metrics would result in very high errors for even small displacements of storms, known as the “double-penalty” effect. From a subjective perspective, these forecasts may be viewed as useful and skillful despite bad metrics, and in some cases an intuitively good forecast may have lower scores than a subjectively worse forecast (e.g. Roebber et al. 2004). A different approach is needed for objective verifications of simulated convection in CAMs that better correlate with subjective evaluations.

One commonly used verification approach for CAMs that has been introduced is the neighborhood method (e.g. Schwartz et al. 2010). This approach involves applying a pre-determined radius (*r*) to each grid point within the model to create a circular region known as a “neighborhood” that is much larger than the individual grid point. A neighborhood probability (NP) can be calculated as (Schwartz et al. 2010):

$$NP_{ki} = \frac{1}{N_b}\sum_{m=1}^{N_b} BP_{km}$$

where NP_{ki} for member k and central grid point i is the average of BP_{km} for the k^{th} ensemble member over each point, m, within the neighborhood of the central grid point, with N_b defining the total number of points within the neighborhood. This method can then be extended to multiple ensemble members such that:

$$NEP_i = \frac{1}{n}\sum_{k=1}^{n} NP_{ki}$$

Where NEP_i is the neighborhood ensemble probability at a grid point i when averaged over all ensemble members n.

The probabilities generated by this approach are dependent on the chosen radius, r, of the neighborhood. Choosing proper values for r is important as it will ultimately affect the results. Figure 1 shows NEP of composite reflectivity exceeding 30 dBZ, valid at 0100 UTC 27 May 2021. Notice that increasing the neighborhood radius from 12km at (a) to 48km at (c) acts as a spatial smoother, lowering the maximum values and reducing gradients.

The same neighborhood method can also be applied to observations to facilitate objective comparisons. The observed NP thus represents the proportion of observed events within a given neighborhood. By weighing ensemble probabilities against observed probabilities applied to a particular variable (such as accumulated precipitation or reflectivity), we can thereby measure the accuracy of the model.

A common quantitative metric for measuring this accuracy within the neighborhood approach is the fractions skill score (FSS; e.g. Roberts and Lean 2008, Schwartz et al. 2010). The FSS is computed as follows:

$$FSS = 1 - \frac{FBS}{FBS_{worst}}$$

where FBS is the Fraction Brier Score and FBS_{worst} defines the level of zero skill, and an FSS of 1 indicates a perfect forecast. FBS can be calculated as follows:

$$FBS = \frac{1}{N_v}\sum_{i=1}^{N_v}\left[NP_{F(i)} - NP_{O(i)}\right]^2$$

where $NP_{F(i)}$ and $NP_{O(i)}$ are the observed and forecasted neighborhood probabilities, respectively, at the same grid point i. Thus, FBS represents the mean-square-differences of these neighborhood probabilities over all grid points in the verification domain, Nv. Because FBS represents the average difference across all grid points, a lower score will indicate better model performance. The FBS is compared to $FBSWorst$, which is calculated by:

$$FBS_{Worst} = \frac{1}{N_v}\sum_{i=1}^{N_v}\left[NP_{F(i)}^2 + NP_{O(i)}^2\right]$$

where FBS_{Worst} represents the baseline value for 0 skill as the worst-case scenario where nonzero observed and forecast probabilities have no overlapping locations.

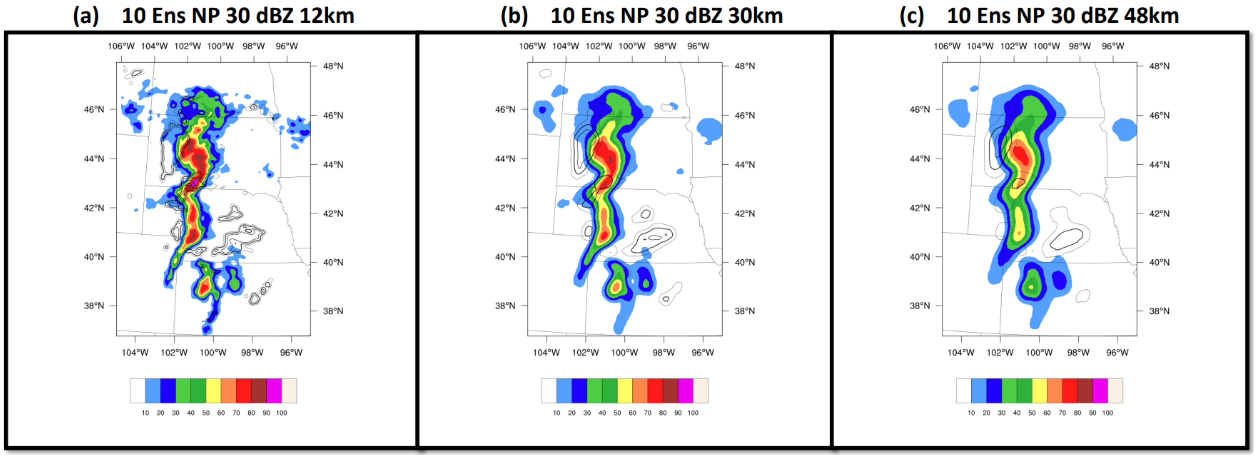
(a) 10 Ens NP 30 dBZ 12km
(b) 10 Ens NP 30 dBZ 30km
(c) 10 Ens NP 30 dBZ 48km
106°W 104°W 102°W 100°W 98°W 96°W
48°N
46°N
44°N
42°N
40°N
38°N
104°W 102°W 100°W 98°W 96°W
10 20 30 40 50 60 70 80 90 100

FIG. 1. Neighborhood ensemble probabilities (NEPs) for 10 ensembles valid 0100 UTC 27 May 2021 (in color) and observed neighborhood probabilities (black contours). 12km, 30km, & 48km neighborhood radii are used (a), (b), & (c), respectively.

By calculating FSS through the neighborhood method we can generate and plot quantitative data showing model performance over time during each case, as well as on a per case basis. FSS can also be averaged across all cases for each year, thereby allowing a direct comparison of model accuracy between separate years. In Fig. 2, we can see FSS averaged over the 10-member ensemble initialized at 0000 UTC 27 May 2021 valid for the next-day (12-36 h) forecast timeframe. FSS changes with time, indicating varying model accuracy between 12z May 27th and 12z May 28th. There is consistently higher skill at all forecast hours as neighborhood radius increases.

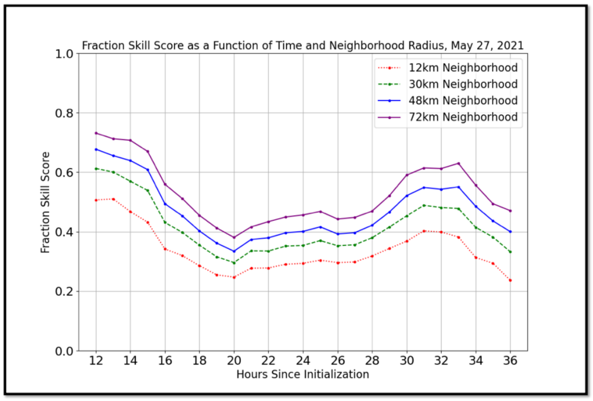

Fig. 2. Fractions Skill Score (FSS), plotted as a function of forecast time 12 – 36 h after initialization for 27 May 2021 case. Line colors correspond to FSS using 12 (red), 30 (green), 48 (blue), and 72 km (purple) neighborhood radii.

c. Surrogate severe method

While the standard neighborhood approach is useful when measuring model performance involving variables such as simulated reflectivity or precipitation, it doesn't give insight into model-indicated storm intensity associated with severe hazards. Considering one of the main objectives of CAMs is to help forecasters identify areas with increased severe weather potential such as wind, hail, and tornadoes, investigating how they perform in this regard is of high priority. To accurately verify severe potential in CAMs, a "surrogate severe" method was first developed by Sobash et al. (2011) and used in several studies since (e.g. Sobash et al. 2016; Roberts et al. 2020; Gasperoni et al. 2023). This technique utilizes model-simulated 2-5km updraft helicity (UH; Kain et al. 2008) tracks as surrogates for severe weather reports. Since UH is a measurement of mid-level rotation within thunderstorm updrafts, and updraft spin is positively correlated with storm intensity, this variable can act as a proxy, or surrogate, for severe weather within the model. These surrogate severe reports can then be compared to observed storm reports obtained via the Storm Prediction Center (SPC).

To create plots that can be verified against observations, this method upscales 24-h maximum UH onto a coarse 80-km verification grid. A threshold is then applied to each 80-km grid point to create a binary field of "surrogate severe reports" (SSRs). That is, within each 80-km grid box, any 3-km model grid point with simulated UH exceeding a given threshold labels that 80-km box with an SSR of 1. Thus, each 80-km box acts as a neighborhood "search radius", similar to the neighborhood maximum method described in Schwartz and Sobash (2017). To produce a probabilistic representation of the distribution of SSRs from the model, a Gaussian filter can be further applied as a spatial smoother with a given smoothing scale, σ. In this filtering approach, σ functions similar to the neighborhood averaging radius r in the standard neighborhood method. The smoothing of SSR produces a surrogate severe probabilistic forecast (SSPF). When working with an ensemble system, an SSPF can be produced from the SSR field of each ensemble member, and the ensemble average of these SSPFs produce an ensemble surrogate severe probabilistic forecast (E-SSPF) which has been shown to lead to better objective scores than deterministic SSPF (Sobash et al. 2016).

In Fig. 3 we can see an example of an E-SSPF compared to observed "practically perfect probabilities" generated from observed storm reports (OSRs) over the same period. Notice that applying a σ to the SSR and OSR fields results in probability distributions between 0 and 1 that paints the areas at greatest risk of severe weather. In this case on the 26 May 2021 there is significant overlap in observed probabilities and E-SSPF plots in the Great Plains showing model generated UH tracks coincided with the general area of storm reports. However, storm reports in the Northeast were numerous where the model did not generate much UH ≥ 200 $m^2 s^{-2}$, indicating a poor performance in that region. Note that practically perfect probabilities were produced with σ = 120 km, representing a "perfect" forecast which mimics probabilities of an SPC convective outlook.

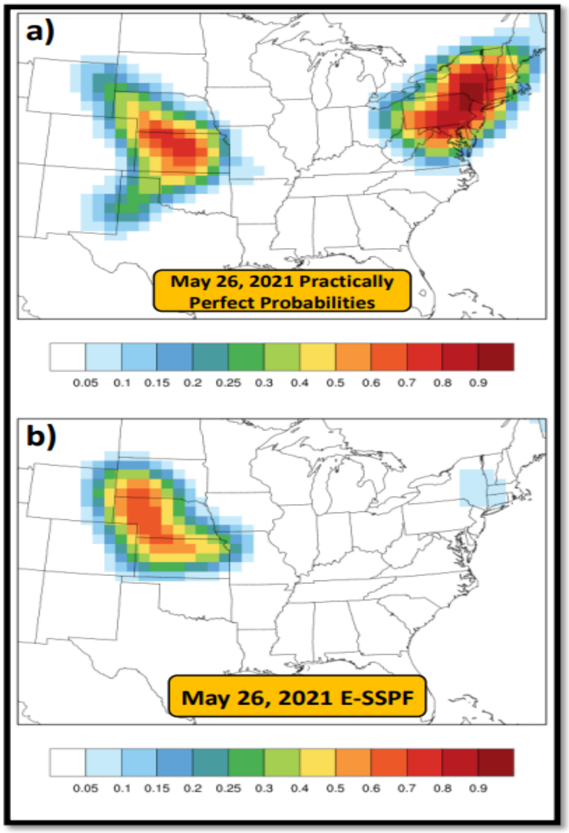

FIG. 3. Observed "practically perfect probabilities" plot created from observed storm reports (a) and E-SSPF, produced by averaging all member SSPFs (b) plotted on an 80 km grid for 26 May 2021. SSPFs were produced using UH threshold = $200\ \mathrm{m^2 s^{-2}}$ & σ = 120 km.

3. RESULTS

a. Assessing skill of composite reflectivity and 1-h precipitation

To analyze and compare model accuracy, FSS was plotted for each year across multiple thresholds involving both model simulated reflectivity (in dBZ) and 1-h accumulated precipitation (in inches). FSS was computed across four different neighborhood radii (12km, 30km, 48km & 72km). In addition to analyzing differences among averages between years, differences in performance based on ensemble size of NEP (3 vs. 10) were also measured.

For reflectivity, two thresholds were chosen (30 and 40 dBZ) with ≥ 30 dBZ generally indicative of model generated convection, while ≥ 40 dBZ is associated with heavier precipitation typically found within convective cores. All of these FSS data were plotted as a function of time between hours 12 – 36 after initialization, which corresponds to 1200 UTC – 1200 UTC the following day. Each of the following plots uses the aggregate method to combine FSS across the 10 cases in each year. The typical averaging method is produced by averaging FSS scores by case for each forecast hour. While this is a straightforward method to calculate average FSS scores, it is susceptible to sample size issues. Generally, cases that have lesser storm coverage can suffer from generally lower FSS due to reduced predictability coupled with enhanced variability

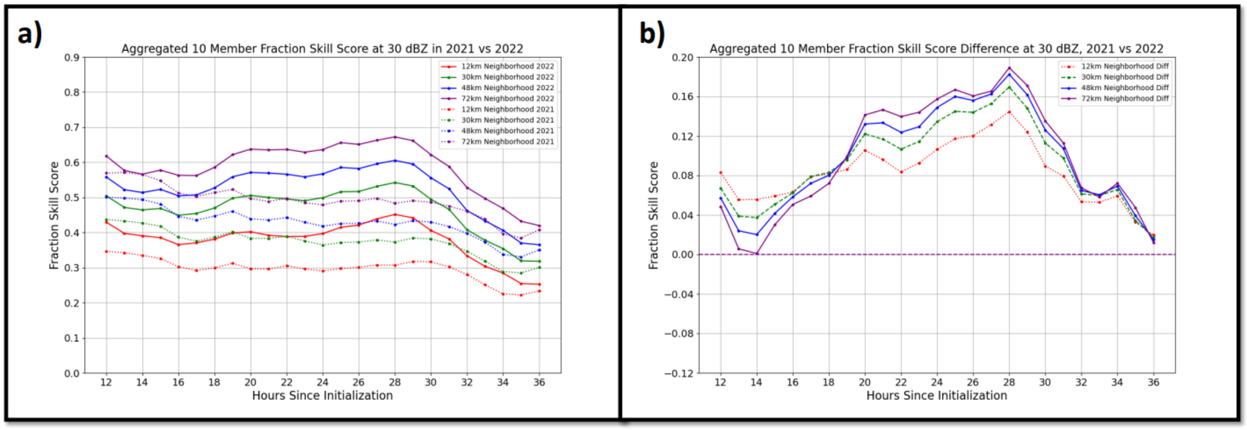

FIG. 4. (a) 10-member FSS of 30-dBZ composite reflectivity, aggregated over all 10 cases from 2021 (dashed lines) and 2022 (solid lines) with four neighborhood radii (12, 30, 48, 72 km). (b) Differences in FSS between years (FSS_{2022} – FSS_{2021}). Scores are shown for forecast hours 12-36 in each panel.

from lower sample sizes. To compensate for this, aggregated FSS is used. With this method FBS and $FBSWorst$ values for each hour are instead summed up over all 10 cases. The FBS sum and $FBSWorst$ sum for each hour are then divided to produce an aggregated FSS score.

To facilitate a direct comparison between 2021 and 2022, 10-member aggregate FSS for both years were plotted together as a function of time. Figure 4 shows 2021 vs 2022 10-member aggregate FSS plotted together (a) and their calculated differences (b). FSS in 2022 is above that of 2021 across most of the period. The greatest differences can be seen during the convective peak of day 2 where 2022 outperforms 2021 by a skill score of > 0.08 at all radii for several hours. These differences are greater at larger radii with the largest differences occurring at 72 km.

The same comparisons were also done with a 40 dBZ threshold (Fig. 5). Though the overall FSS scores are lower, similar trends can be seen in these comparisons with 2022 performing better than 2021 overall. There remains an improvement in skill in 2022 with the greatest improvements again occurring during typical periods of greater convective activity. Additionally, this improvement appears to be weakest during the earlier hours. However, the overall improvement in 2022 at this higher threshold is less than what is seen at 30 dBZ.

For precipitation, three thresholds (0.1 inch, 0.25 inch, and 0.5 inch) were analyzed. Comparisons were done between 2021 and 2022 at all of these thresholds. Ensemble sizes (3 vs 10) were also compared for both 2021 and 2022 at all thresholds.

Figure 6 shows a direct comparison between 2021 and 2022 at an accumulated 1-h precipitation threshold of 0.1 inches (a & b) and 0.25 inches (c & d). At the lower threshold, 2022 outperforms 2021 at all hours. However, the increase in accuracy is more consistent across time than both reflectivity thresholds shown earlier. Of all the thresholds tested with the standard neighborhood method, the 0.1 inch precipitation forecasts showed the greatest improvement between the two years. While improvement is also seen at the higher threshold of 0.25 inches, it's both less substantial and less consistent than the 0.1 inch threshold. This indicates that, similar to reflectivity, greater improvement seems to have occurred at lower thresholds in accumulated precipitation as well.

Comparisons between 3 and 10 member ensemble sizes were also done within each year for each precipitation threshold (Fig. 7). As expected, in both years and at all thresholds, 10 ensembles show greater accuracy than 3 ensembles at almost all time periods. For both years the advantage in forecast skill of the larger member size also appears to grow as the precipitation threshold increases. For example, the difference in accuracy between 3 and 10 ensembles is the greatest at the 0.5 inch threshold for both years seen in plots (c and f). However, the most interesting observation is the fact that this increase in skill difference at larger thresholds is far more substantial in 2022 than 2021. This can be seen in plots (e and f), which show 0.25 inch and 0.5 inch thresholds respectively for 2022. The largest increase in skill for both of these thresholds occurs between hour 16 and 24.

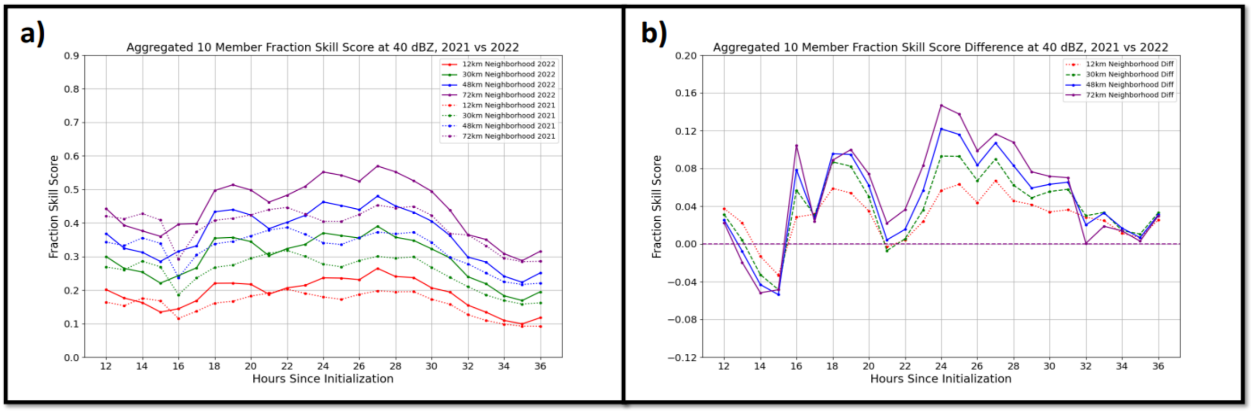

Fig. 5. As in Fig. 4, but for 40-dBZ composite reflectivity.

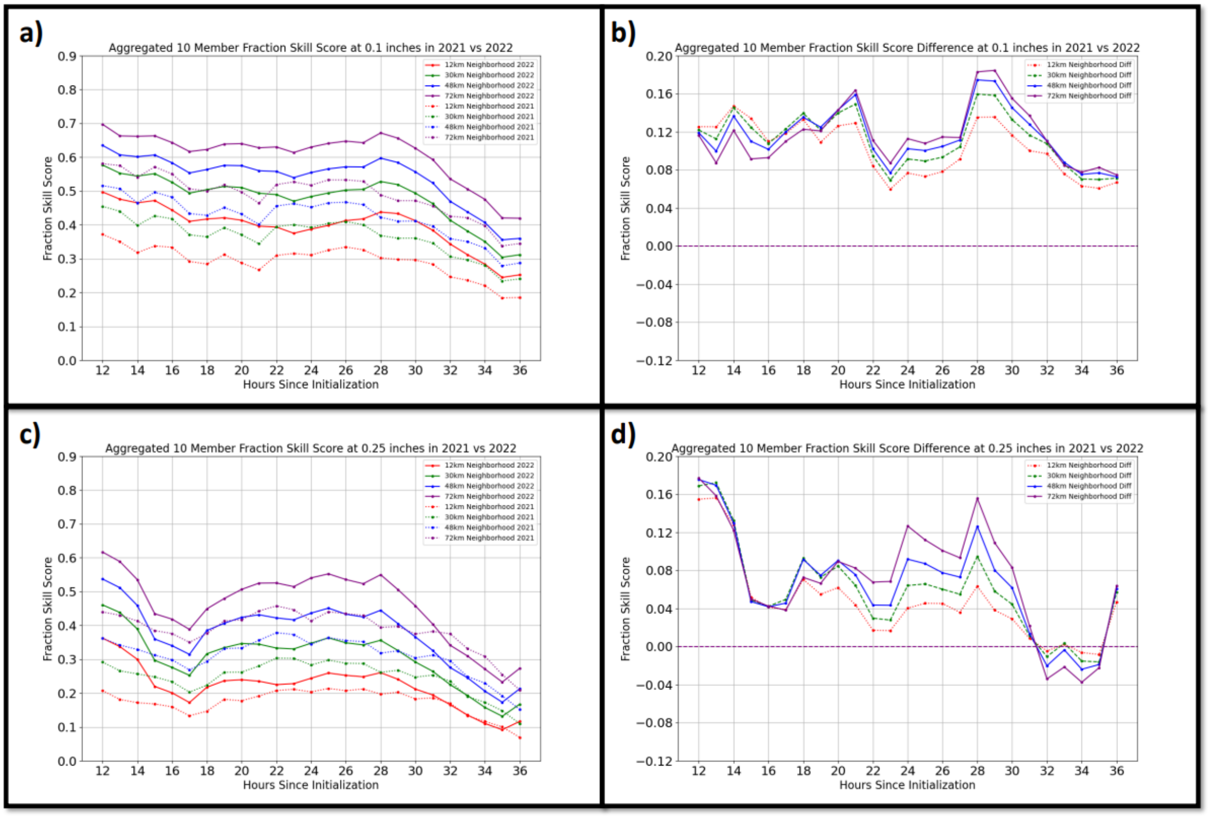

FIG. 6. As in Fig. 4, but for 1-h accumulated precipitation at (a,b) 0.1 inch and (c,d) 0.25 inch thresholds (2.54 mm & 6.35 mm, respectively)

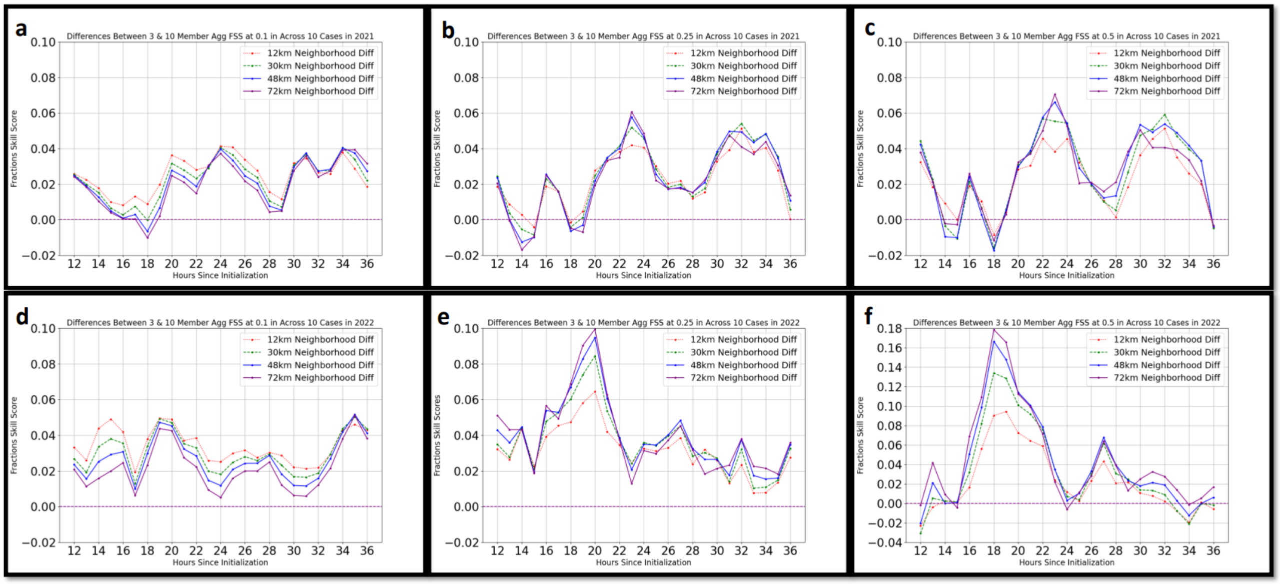

FIG. 7. 3 vs 10 member aggregated FSS differences (10mem – 3mem) at three precipitation thresholds for both 2021 (a,b,c) and 2022 (d,e,f) as a function of forecast hour (12-36) for four neighborhood radii. Thresholds increase from left to right with (a,d) 0.1 inches, (b,e) 0.25 inches, and (c,f) 0.5 inches (2.54, 6.35, & 12.7 mm, respectively).

b. Surrogate severe comparisons

Analysis of FSS differences between SSPF's from 2021 and 2022 is more mixed. Figure 8a shows the result of aggregating FSS among all cases in 2021 and 2022. FSS scores are plotted as a function of both neighborhood smoothing (σ) on the x-axis and UH thresholds on the y-axis. The highest skill for 2022 shifts toward lower σ values (120-140 km) compared to 2021 (160-180 km), which is more consistent with the 120-km scale of observed practically perfect probabilities. Both years peak in skill around a UH threshold of 150 $m^2\, s^{-2}$. The differences between these two averages are also shown in Fig. 8b. Aggregated FSS from 2021 has a wider range of higher skill at greater UH thresholds and σ; however, the aggregated skill from 2022 is improved at lower σ more consistent with observations and with UH thresholds where FSS is maximized (150 $m2s{-}2$).

Figure 9 shows FSS plotted as a function of the same variables, but on a case-by-case basis. There is greater case-to-case consistency in FSS in 2022, most notably with the locations of FSS maxima. Though some 2021 cases have a similar location of FSS maxima as 2022 cases, there are several cases where the maxima are shifted towards higher UH thresholds and/or higher smoothing scales. On the other hand, despite less consistent results, there are a few 2021 cases with substantially higher FSS than all 2022 cases (17, 27, and 28 May 2021).

4. DISCUSSION

FSS data between 2021 and 2022 showed some promising results. When comparing aggregated FSS across 10 separate cases in 2021 and 2022, the latter shows superior skill across most of the 12-36 h forecast timeframe. When focusing on reflectivity with a threshold of 30 dBZ, aggregate FSS in 2022 is greater throughout the 12-36 h period. However, this increase in skill was not consistent and appears to manifest most strongly between hours 20 and 30 of day 2. This is likely due to much greater forecasting skill in 2022 during periods of greater convective coverage, which are typically seen in the late afternoon and evenings of most of the case days. While 2021 also shows an increase in accuracy during more active convective periods, this increase is more substantial in 2022. Considering the largest skill differences occur during periods of widespread convection, this is evidence that the 2022 ensembles handled

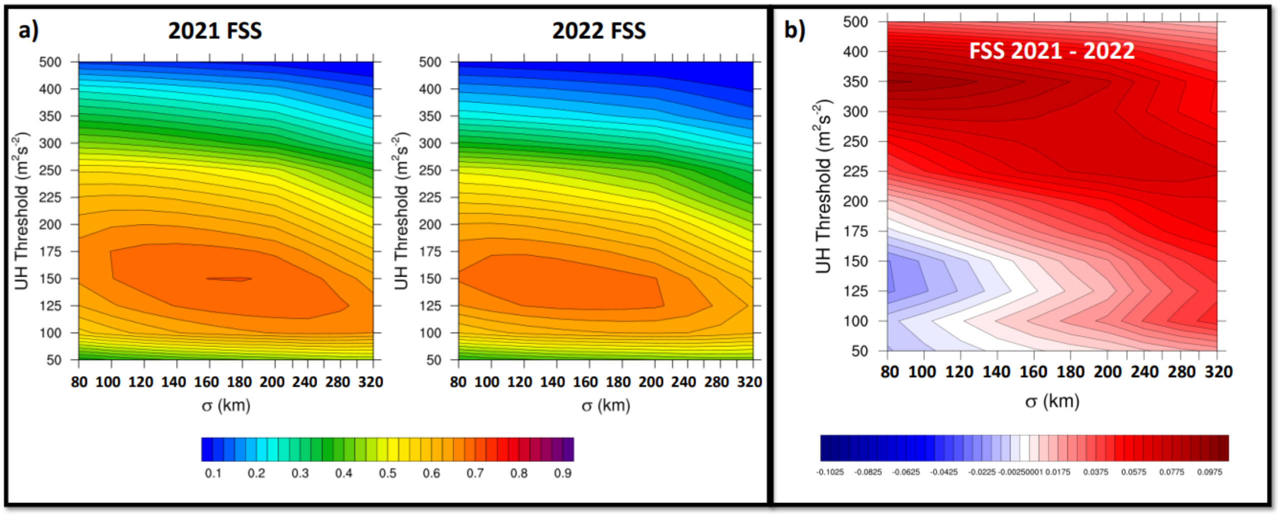

FIG. 8. (a) FSS of 24-h E-SSPF aggregated over all 10 cases from 2021 (left) and 2022 (right), plotted as a function of UH thresholds (y-axis) and σ (x-axis) in two separate graphs for 2021 & 2022. Color contours show FSS ranging between values of 0 & 1. (b) Differences in aggregated FSS (2022 – 2021) plotted as a function of UH thresholds (y-axis) and σ (x-axis). Blue colors show increased FSS for 2022 while red colors show increased FSS for 2021.

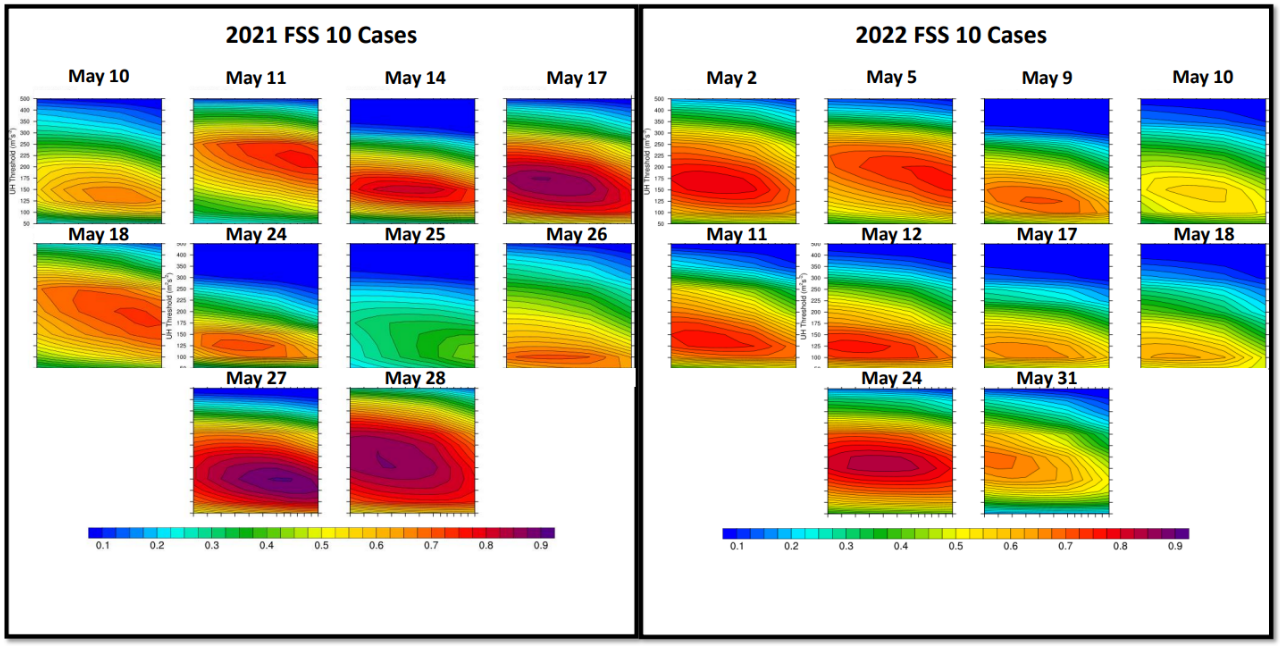

Fig. 9. FSS plotted as a function of both UH thresholds (y-axis) and σ (x-axis) from each of the 10 cases in both 2021 (left) & 2022 (right). Note axis and color bar ranges for individual plots are the same as in Fig. 8a.

thunderstorm coverage and intensity better than in 2021. This increase in skill shows up at the 40 dBZ threshold as well, further supporting this hypothesis. Observing these improvements at values of 40 dBZ is especially convincing since these higher reflectivity values typically show up in convective cores, which are more extreme, rare events with lower predictability. Thus, a higher skill score at these values is indicative of measurable improvement in the placement of intense convection within the model in relation to observations.

It should be noted that when *averaging* FSS by cases (not shown), 2021 did slightly outperform 2022 at 30 and 40 dBZ during hours 12-15. However, this advantage disappears when the aggregate method is used. The fact that this method smooths out these differences could be due to sample size issues between cases, with early time periods during 2022 cases displaying less storm activity. If that is indeed the case, the apparent early skill advantage in the average for 2021 could be more strongly influenced by sample size issues rather than actual forecasting skill of the model. Across the rest of the 12 – 36 h period, 2022 held a substantial lead in skill across both methods.

Another interesting observation is that the score differences between 2021 and 2022 are greatest among larger radii. Though the reason for this is not certain, a potential hypothesis is that the updated dynamics and physics in 2022 allowed ensembles to improve most during strongly-forced scenarios where convection is more widespread. However, there was less improvement at smaller scale details in the exact placement of convective cores. Presumably, larger neighborhoods would smooth out these small-scale imperfections and put more weight on overall reflectivity coverage rather than precise placement of individual storms.

When comparing FSS across precipitation thresholds similar conclusions can be drawn. At all thresholds (0.1 inch, 0.25 inch, 0.5 inch) 2022 shows superior skill on average. However, the advantage is much larger and far more consistent at the lightest threshold 0.1 inches, with progressively less improvement at 0.25 inches and 0.5 inches. Lower thresholds such as 0.1 inch hourly accumulated precipitation reflect where the model produces precipitation in general while higher thresholds (0.25 inch and 0.5 inch) would correlate more closely with regions of convective activity. Reflectivity values chosen also correlate more closely with convection, with 30 dBZ indicative of broad regions of convection, with 40 dBZ would is rare outside of convective cores. The fact that the 0.1 inch threshold saw a greater increase in skill than either the reflectivity thresholds or the higher precipitation thresholds is evidence that the greatest skill improvement was on synoptic scales due to improvements in the model dynamics from 2021 to 2022.

While the main objective of this study was to compare data between years, an analysis of skill differences between ensemble sizes can provide further insight into model accuracy. In both 2021 and 2022 the forecast accuracy was greater among 10 ensembles when compared to a smaller set of 3 ensembles. This is not surprising considering greater ensemble size typically results in greater forecasting potential, as the ensemble spread better reflects actual forecast uncertainty. The more interesting result focuses on how large the skill increase is between ensemble sizes in 2021 vs 2022. While 2021 did show an increase in skill with the larger ensemble size, 2022 shows a *greater* skill jump when ensemble members are increased from 3 to 10, especially during the next-day period where new convection is developing (16-22 h). When looking at these comparisons across all three precipitation thresholds this large skill jump is most apparent at the 0.25 inches and 0.5 inches. The skill increase at 0.5 inches peaks at about 0.18 at hour 18, nearly twice as much as the increase seen at 0.25 inches. These larger skill improvements of greater ensemble sizes in 2022 could be reflective of improved ensemble spread due to the implementation of stochastic physics for that year. This was meant to increase ensemble spread through time, which should increase the diversity and thus enhance the advantages of greater ensemble sizes.

The results derived from the surrogate severe method were less conclusive. When the differences of aggregate FSS between the two years are plotted, 2021 shows higher skill across a greater range of UH thresholds and σ. However, 2022 does show higher skill at lower UH thresholds, (50 to 175 m^2s^{-2}) and lower σ (≤ 120 km), near where FSS is maximized. Since 2022 performs better with less smoothing at lower UH values this may be evidence that model generated UH tracks within these ranges were generally closer to observed storm reports. How well these results reflect actual skill differences is uncertain. When observing FSS on a case-by-case basis for each year it's apparent that there was greater case-to-case variability in 2021, with three cases (May 17th, 27th, 28th) producing substantially higher FSS. Meanwhile, 2022 shows greater consistency in the location and magnitude of the FSS maximum. This is evidence that 2022 may have given more reliable forecasts and holds an advantage from a predictability standpoint, though other verification scores would be needed to confirm. The variation in scores in 2021 may also be an indication that the aggregated scores are more strongly influenced by case variability rather than purely reflecting model accuracy. Overall, this motivates the need for further study involving the surrogate severe method before solid conclusions can be drawn.

5. CONCLUSIONS

This study sought to quantify skill differences in two RRFS-like ensemble systems run by OU MAP Lab in 2021 and 2022. Accuracy was measured over 10 separate cases for each year using the standard neighborhood method and surrogate severe method. Overall, results were promising with skill being generally higher in 2022 at several reflectivity and precipitation thresholds with the largest and most consistent improvement seen at lower thresholds. This is evidence for improvement in the placement of both convection and accumulated precipitation between the two years, which demonstrates clear progress in the forecasting accuracy of these new ensemble systems. However, results from the surrogate severe method were mixed with 2022 showing greater consistency among cases while 2021 performed better across a greater range of updraft helicity thresholds. Ultimately this enforces the need for further research and analysis before confident conclusions can be drawn. While evidence shows these FV3 ensemble systems are indeed improving at convective allowing scales, further improvement and testing will likely be needed before they are ready to be widely used in forecasting.

There are plenty of areas involving verification of these systems that would benefit from future research as well. While skill differences between 2021 and 2022 were measured in this study, further verifications (e.g. reliability, bias) are necessary for a more comprehensive understanding, including statistical significance testing of score differences. Additionally, data produced by OU MAP during HWT from 2017-2019 using the HRRR and NMMB models could be used to contextualize the progress seen in these next-gen RRFS-like ensemble systems that use the FV3-LAM.

ACKNOWLEDGMENTS

This work was supported by Grant No. AGS-2050267 from the National Science Foundation. This study was made possible thanks to the SFEs run by OU MAP Lab and data generated using the OU Supercomputing Center for Education & Research (OSCER) computer system as well as the work of Dr. Yongming Wang.

A special thanks to my mentors Dr. Nicholas Gasperoni and Dr. Xuguang Wang for providing guidance and insight throughout this research process. I would also like to thank Daphne LaDue and Alex Marmo for coordinating the National Weather Center Research Experience for Undergraduates program that provided the opportunity to conduct this study and gain valuable research experience.

REFERENCES

Black, T. L., and Coauthors, 2021: A Limited Area Modeling Capability for the Finite-Volume Cubed-Sphere (FV3) Dynamical Core and Comparison With a Global Two-Way Nest. *J Adv Model Earth Syst*, **13**, https://doi.org/10.1029/2021MS002483.

Clark, and Coauthors, 2019: Spring Forecasting Experiment 2019 conducted by the Experimental Forecast Program of the NOAA Hazardous Weather Testbed. NOAA Preliminary Findings and Results Rep., 77 pp., https://hwt.nssl.noaa.gov/sfe/2019/docs/HWT_SFE_2019_Prelim_Findings_FINAL.pdf.

Dowell, D. C., and Coauthors, 2022: The High-Resolution Rapid Refresh (HRRR): An Hourly Updating Convection-Allowing Forecast Model. Part I: Motivation and System Description. *Weather and Forecasting*, **37**, 1371–1395, https://doi.org/10.1175/WAF-D-21-0151.1.

Gasperoni, N. A., X. Wang, and Y. Wang, 2023: Valid Time Shifting for an Experimental RRFS Convection-Allowing EnVar Data Assimilation and Forecast System: Description and Systematic Evaluation in Real Time. *Monthly Weather Review*, **151**, 1229–1245, https://doi.org/10.1175/MWR-D-22-0089.1.

Harris, L. M., and S.-J. Lin, 2013: A Two-Way Nested Global-Regional Dynamical Core on the Cubed-Sphere Grid. *Monthly Weather Review*, **141**, 283–306, https://doi.org/10.1175/MWR-D-11-00201.1.

James, E. P., and Coauthors, 2022: The High-Resolution Rapid Refresh (HRRR): An Hourly Updating Convection-Allowing Forecast Model. Part II: Forecast Performance. *Weather and Forecasting*, **37**, 1397–1417, https://doi.org/10.1175/WAF-D-21-0130.1.

Kain, J. S., and Coauthors, 2008: Some Practical Considerations Regarding Horizontal Resolution in the First Generation of Operational Convection-Allowing NWP. *Weather and Forecasting*, **23**, 931–952, https://doi.org/10.1175/WAF2007106.1.

Roberts, B., I. L. Jirak, A. J. Clark, S. J. Weiss, and J. S. Kain, 2019: PostProcessing and Visualization Techniques for Convection-Allowing Ensembles. *Bulletin of the American Meteorological Society*, **100**, 1245–1258, https://doi.org/10.1175/BAMS-D-18-0041.1.

——, B. T. Gallo, I. L. Jirak, A. J. Clark, D. C. Dowell, X. Wang, and Y. Wang, 2020: What Does a Convection-Allowing Ensemble of Opportunity Buy Us in Forecasting Thunderstorms? *Weather and Forecasting*, **35**, 2293–2316, https://doi.org/10.1175/WAF-D-20-0069.1.

Roberts, N. M., and H. W. Lean, 2008: Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events. *Monthly Weather Review*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1.

Roebber, P. J., D. M. Schultz, B. A. Colle, and D. J. Stensrud, 2004: Toward Improved Prediction: High-Resolution and Ensemble Modeling Systems in Operations. *Wea. Forecasting*, **19**, 936–949, https://doi.org/10.1175/1520-0434(2004)019<0936:TIPHAE>2.0.CO;2.

Schwartz, C. S., and R. A. Sobash, 2017: Generating Probabilistic Forecasts from Convection-Allowing Ensembles Using Neighborhood Approaches: A Review and Recommendations. *Mon. Wea. Rev.*, **145**, 3397–3418, https://doi.org/10.1175/MWR-D-16-0400.1.

——, and Coauthors, 2009: Next-Day Convection-Allowing WRF Model Guidance: A Second Look at 2-km versus 4-km Grid Spacing. *Monthly Weather Review*, **137**, 3351–3372, https://doi.org/10.1175/2009MWR2924.1.

——, and Coauthors, 2010: Toward Improved Convection-Allowing Ensembles: Model Physics Sensitivities and Optimizing Probabilistic Guidance with Small Ensemble Membership. *Weather and Forecasting*, **25**, 263–280, https://doi.org/10.1175/2009WAF2222267.1.

Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic Forecast Guidance for Severe Thunderstorms Based on the Identification of Extreme Phenomena in Convection-Allowing Model Forecasts. *Weather and Forecasting*, **26**, 714–728, https://doi.org/10.1175/WAF-D-10-05046.1.

——, C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016: Severe Weather Prediction Using Storm Surrogates from an Ensemble Forecasting System. *Weather and Forecasting*, **31**, 255–271, https://doi.org/10.1175/WAF-D-15-0138.1.