Assessing RRFS versus HRRR in Predicting Widespread Convective Systems over the Eastern CONUS

JOSEPH A. GRIM,[a] JAMES O. PINTO,[a] AND DAVID C. DOWELL[b]

[a] *Research Applications Laboratory, National Center for Atmospheric Research, Boulder, Colorado*
[b] *NOAA/Global Systems Laboratory, Boulder, Colorado*

(Manuscript received 30 June 2023, in final form 16 October 2023, accepted 13 November 2023)

ABSTRACT: This study provides a comparison of the operational HRRR version 4 and its eventual successor, the experimental Rapid Refresh Forecast System (RRFS) model (summer 2022 version), at predicting the evolution of convective storm characteristics during widespread convective events that occurred primarily over the eastern United States during summer 2022. In total 32 widespread convective events were selected using observations from the MRMS composite reflectivity, which includes an equal number of MCSs, quasi-linear convective systems (QLCSs), clusters, and cellular convection. Each storm system was assessed on four primary characteristics: total storm area, total storm count, storm area ratio (an indicator of mean storm size), and storm size distributions. It was found that the HRRR predictions of total storm area were comparable to MRMS, while the RRFS overpredicted total storm area by 40%–60% depending on forecast lead time. Both models tended to underpredict storm counts particularly during the storm initiation and growth period. This bias in storm counts originates early in the model runs (forecast hour 1) and propagates through the simulation in both models indicating that both miss storm initiation events and/or merge individual storm objects too quickly. Thus, both models end up with mean storm sizes that are much larger than observed (RRFS more so than HRRR). Additional analyses revealed that the storm area and individual storm biases were largest for the clusters and cellular convective modes. These results can serve as a benchmark for assessing future versions of RRFS and will aid model users in interpreting forecast guidance.

KEYWORDS: Convection; Thunderstorms; Numerical weather prediction/forecasting

1. Introduction

Widespread areas of convection are responsible for bringing beneficial rains as well as damaging winds and hail, dangerous lightning, and flash flooding. Thunderstorms also have a major impact on both ground and air transportation that can result in significant travel delays, and sometimes dangerous conditions both along highways and for aviation. Understanding the skill of numerical weather prediction models at predicting convective storms is crucial for decision makers in planning and mitigating their impact on society. In this study, we assess the skill of two convection-permitting models in predicting the characteristics of widespread convection events. Here, the term widespread is used to indicate that the size of the region impacted by convection exceeds 300 000 km^2 (about the size of Arizona), while the density of peak convection (defined at the 35-dBZ threshold) exceeds at least 1.5% of the area. These widespread events may be characterized by one or more modes of convective organization [e.g., cells, clusters, mesoscale convective systems (MCS), and quasi-linear convective systems (QLCS) but with a clear dominant mode at maturity.]

MCSs and QLCSs tend to consist of large individual convective elements. Because of their size, these convective systems produce a majority of the severe weather reports (e.g., Gallus et al. 2008; Smith et al. 2012). These larger-scale convective systems also tend to be major contributors to flight delays across the United States, with convective storms totaling over 50% of weather-related delays reported each summer (FAA 2022a). Widespread areas of storms characterized by smaller scale organization, such as cellular storms or storm clusters, can have significant impacts as well including severe characteristics of supercells that can generate large damaging hail and tornadoes (e.g., Rasmussen et al. 1994; Homeyer et al. 2023).

It has been shown that high-resolution forecast models, with horizontal grid spacing of 4 km or less, are able to adequately capture much of the mesoscale dynamics and thermodynamic processes required to effectively predict convective storm mode (e.g., Weisman et al. 1997; Done et al. 2004; Weisman et al. 2008; Schwartz et al. 2009). As such, operational modeling centers around the world are running regional NWP models at convection-permitting resolutions (e.g., Walters et al. 2019; Termonia et al. 2018; Husain et al. 2019) in order to provide short-term to next-day predictions on the timing, location, organization and severity of convective storms as well as improving the depiction of many other mesoscale weather phenomena (e.g., Benjamin et al. 2016; Roff et al. 2022). This improved guidance is critical for providing advanced notice of the potential for high impact weather. Accurate advanced prediction of the macroscale properties of convective storm systems (including the distribution of storm sizes, their coverage, shape, and orientation) are also critical for mitigating their impacts on air travel (Pinto et al. 2015). The HRRR (Benjamin et al. 2016; Dowell et al. 2022;

Denotes content that is immediately available upon publication as open access.

Corresponding author: Joseph A. Grim, grim@ucar.edu

DOI: 10.1175/WAF-D-23-0112.1

Weygandt et al. 2022) has been providing convective storm guidance to a range of users including aviation meteorologists over the past decade. James et al. (2022) summarizes the performance of the HRRR throughout the development process and provides an overview of previous studies that evaluated various aspects of the HRRR over the years.

Several studies have evaluated more recent versions (v3 and v4) of the HRRR in terms of its skill at predicting various characteristics of convective storms (e.g., Blaylock and Horel 2020; Duda and Turner 2021; Grim et al. 2022; James et al. 2022). Duda and Turner (2021) found that HRRRv3 overpredicted the number of convective storm objects, with a large contribution of this bias coming from the smallest storm objects at low reflectivity thresholds, which are not adequately resolved by the model; on the other hand, it also overforecasts large objects at high reflectivity thresholds. They found that this bias was most pronounced across much of the southern United States, and the southeastern United States in particular. This finding is consistent with that reported by Grim et al. (2022) who found that while HRRRv4 and HRRR-Ensemble were able to capture the amplitude, timing, and evolution of QLCS and MCS macroscale characteristics, neither modeling system performed well for the smaller scale convective modes (cellular and clusters), which are the dominant convective mode in the southeastern United States (Miller and Mote 2017). They found that for the cellular and cluster convective modes the storm objects tended to be 1.5–2.0 times larger than observed. James et al. (2022) found that HRRRv4 had poor skill at predicting smaller-scale diurnally forced convective storms during the climatological peak time of day for convection initiation (CI) and storm growth (i.e., between 1500 and 0300 UTC). In contrast, they found that the HRRRv4 performed best during the overnight hours which tend to be dominated by larger, long-lived convective systems. They also found that the assimilation of radar reflectivity had the largest positive impact in the runs initialized during the overnight hours with the positive impact being most pronounced during lead hours 1–4. Finally, Weygandt et al. (2022) describe how the radar latent heating initialization technique implemented in HRRRv4 shows significant improvement in short-range (0–6 h) forecasts of convective precipitation.

Over the next 1–2 years, the HRRRv4 is being transitioned to the Finite Volume Cubed (FV3)-based (Chen et al. 2013) Rapid Refresh Forecast System (RRFS, Alexander and Carley 2023). To support this transition, this study presents an initial assessment of the experimental RRFS relative to the HRRR in the prediction of convective storms. Section 2 describes the methodology used to determine the macrophysical characteristics of widespread storm events. Section 3 provides an example of how the methodology is applied to assess the relative performance of HRRR and RRFS for a single MCS case. Section 3 also discusses how the results from each case are composited and provides a statistical comparison of HRRR versus RRFS model performance across 32 cases. Finally, section 4 provides a summary of the major results of the study.

2. Methodology

The statistical object-based technique described by Grim et al. (2022) is used to assess the performance of the RRFS model relative to the HRRR at predicting the macroscale characteristics of convective storms. The HRRR data were obtained from operational runs performed at NCEP, while the experimental RRFS data were obtained from NOAA Global Systems Laboratory (GSL). Dowell et al. (2022) explains the latest features of the HRRR, while Alexander and Carley (2023) provide the latest information on the configuration and performance of RRFS. Both HRRR and RRFS use 3-km grid spacing with configurations that are similar, with the main difference being the dynamical core; HRRR uses the WRF-ARW dynamical core (Skamarock et al. 2008) while RRFS uses the FV3 dynamical core (Chen et al. 2013). Both modeling systems assimilate a range of conventional observations using hourly cycling and a hybrid ensemble-variational approach within the Gridpoint Statistical Interpolation (GSI; Benjamin et al. 2016; Dowell et al. 2022). In addition to the conventional data assimilation (DA) process, a method for producing a cloud analysis is implemented which includes cloud clearing and building based on satellite and ceilometer data and adding precipitation hydrometeors (rain, snow, and graupel) using radar reflectivity and lightning observations as discussed in Benjamin et al. (2021) and Weygandt et al. (2022). Finally, a different DA system, which uses an ensemble variational method to assimilate radar reflectivity (Wang and Wang 2017), was implemented and tested in RRFS starting on 6 July 2022. These test runs were performed in place of the original DA for forecasts issued between 1900 and 0000 UTC. In order to isolate the performance of the dynamical cores, only model runs in which the two dynamical cores used the same DA system (i.e., all runs before 6 July 2022, and only those issued between 0100 and 1800 UTC starting 6 July 2022) are evaluated in this study.

Observed areas of convection are diagnosed from a national mosaic of composite reflectivity from the operational MRMS system (see Smith et al. 2016). Based on an analysis of MRMS composite reflectivity, a set of 32 cases were selected from a 3-month period (1 June–31 August) during the 2022 convective season. Cases were chosen subjectively to fit into one of four categories, based on their predominant convective mode at maturity: cellular, cluster, QLCS, and MCS. The total number of cases chosen was primarily limited by the availability of RRFS model output that was running in an experimental environment. The availability of each dataset is summarized in Fig. 1. The union of RRFS, HRRR, and MRMS data availability indicated the time periods available from which cases were selected. Cases were only considered for periods when at most one forecast lead time or observed hour is missing. Based on this criterion, nearly half of the cases selected occurred during the second half of June when the RRFS model was most consistently available. The case start time, defined as being one hour prior to the increase in storm area coverage, is indicated in Fig. 1 with details of each case given in Table 1.

Figure 2 shows the composite reflectively at mature stage for all 32 cases, organized by convective mode and date. The

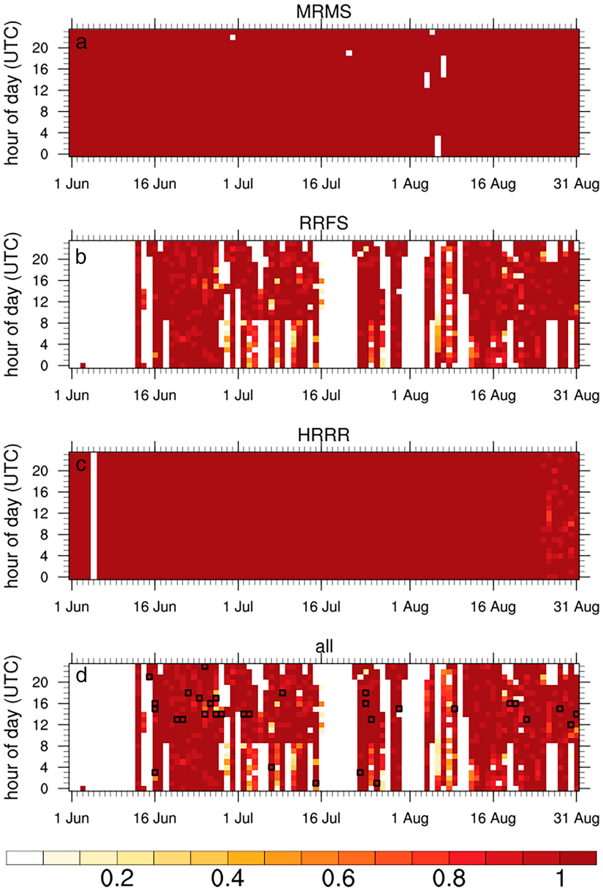

FIG. 1. Availability during summer 2022 of (a) MRMS, (b) RRFS, (c) HRRR, and (d) all products combined. MRMS availability is either zero or one, while RRFS and HRRR availability is shaded from zero to one based on the fractional availability of their respective hourly forecast lead times to 15 h. The all-product availability is the minimum of the percentage availability of the two models at their respective lead times and MRMS at its respective valid times. Overlaid on the all-product availability are the analysis times of the 32 cases (1 h before each event started.)

cellular convective mode includes thunderstorm events dominated by small-scale (i.e., <50-km diameter) mostly circular storms throughout their life cycle that primarily occur under weakly forced conditions driven by diurnal heating. The multicellular cluster mode includes cases in which storm cells grew and often merged to form clusters, and were generally a mixture of cellular convection and clusters of storms (but dominated by multicellular clusters). The QLCS convective mode includes storms that evolved into long nearly continuous lines of convection without substantial trailing or leading stratiform regions. Finally, the MCS convective mode includes only those convective systems that developed substantial stratiform regions exceeding 100 km in width. The QLCS and MCS modes were separated since the larger stratiform region of the MCSs creates a stronger cold pool, as well as to account for storm characteristic differences coming from the stratiform rain regions. For each convective mode, eight cases were selected (Table 1), providing a robust sample size for statistical comparisons between each storm category.

Cases were selected within a region bounded by the U.S. and Canadian coastlines to the south and east, 51°N latitude to the north, and the 105°W meridian on the west. The western United States and most of Canada were excluded from these analyses because there are substantial areas that lack adequate radar coverage. The start date and time of the initial evaluation periods are listed in Table 1; this marks the beginning of the observed increase in convective activity for a given convective event. The model runs used in the primary evaluation were from a single initialization time for each case, initialized one hour before the observed time of CI. The end date and time for each case was the time when the convective activity reached a minimum, or 14 h after the start date and time, whichever came first. The 14-h limit ensured that the entire case period fell entirely within a single 15-h RRFS forecast. Using these criteria, the case periods ranged from 9 to 14 h.

Each dataset is mapped onto a common grid defined using a regular latitude–longitude projection and 0.05° grid spacing applying bilinear interpolation between grid points. The influence of each pixel was then expanded by setting the value at each grid point to the maximum value occurring within a three-gridpoint radius. This was done so that small gaps between convective elements that are part of a larger convective system were considered as a single storm object following Davis et al. (2006). The TITAN (Dixon and Wiener 1993) software was used to identify storms using a technique similar to that described by Pinto et al. (2015) and Grim et al. (2022). A composite reflectivity threshold of 35 dB*Z* was initially used to identify storm objects in both the MRMS and model data. This threshold is comparable to that used previously to identify areas of impactful convective rainfall (e.g., Roberts and Rutledge 2003; Davis et al. 2006). While this threshold is lower than that used in other studies to detect the most intense convective storms (e.g., Skinner et al. 2018; Potvin et al. 2019), the 35-dB*Z* threshold used here better captures the full range of intensities associated with convective storms.

Figure 3 compares probability density functions observed with MRMS with those obtained for all forecast lead times from both RRFS and HRRR during the period when both models used the same DA system. The area used to produce this plot was defined as all land areas of the continental United States and far southeastern Canada east of 105°W and south of 51°N. The mean area for a given reflectivity bin (0–70 dB*Z* × 1 dB*Z*) was found by summing areas for a given bin using concurrently available data for all 92 days of JJA 2022. Peak observed mean area occurs around 12 dB*Z*. The observed reflectivity decreases for values below 12 dB*Z* due to the suppression of these values by radar clutter removal techniques. Thus, comparisons between modeled and observed values should be constrained to be for reflectivity values greater than ~15 dB*Z*. It is evident that (excluding analyses—i.e., forecast hour 0) both models tend to have nearly equal or lesser area coverage at thresholds between 15 and 32 dB*Z* (more so HRRR) and overpredict the area coverage at thresholds greater than 40 dB*Z*. It is also interesting to note that in both models, the reflectivity area increases with forecast lead time across all

TABLE 1. Case start, peak, and end dates, period, dominant mature storm organization type, center latitude–longitude, and if it is included in the lag subset (indicated by an asterisk).

Start time/date	Peak time/date	End time/date	Case period (h)	Dominant storm type	Center lat (°)	Center lon (°)	Lag subset
1600 UTC 16 Jun	2300 UTC 16 Jun	0200 UTC 17 Jun	10	MCS	34.9	−80.5	
1800 UTC 24 Jun	0800 UTC 25 Jun	0800 UTC 25 Jun	14	MCS	47.0	−98.4	*
0000 UTC 26 Jun	0900 UTC 26 Jun	1400 UTC 26 Jun	14	MCS	38.5	−94.2	*
0500 UTC 7 Jul	1200 UTC 7 Jul	1900 UTC 7 Jul	14	MCS	38.9	−92.8	
1900 UTC 9 Jul	0200 UTC 10 Jul	0500 UTC 10 Jul	10	MCS	32.1	−88.2	*
0200 UTC 15 Jul	1200 UTC 15 Jul	1600 UTC 15 Jul	14	MCS	41.9	−94.1	
0400 UTC 23 Jul	1500 UTC 23 Jul	1800 UTC 23 Jul	14	MCS	41.3	−87.5	*
0200 UTC 26 Jul	1300 UTC 26 Jul	1500 UTC 26 Jul	13	MCS	38.4	−89.6	
2200 UTC 15 Jun	0400 UTC 16 Jun	0900 UTC 16 Jun	11	QLCS	43.5	−90.1	
1700 UTC 16 Jun	2000 UTC 16 Jun	0300 UTC 17 Jun	10	QLCS	43.2	−77.0	
1400 UTC 21 Jun	0300 UTC 22 Jun	0400 UTC 22 Jun	14	QLCS	41.0	−95.4	*
1500 UTC 27 Jun	2000 UTC 27 Jun	0100 UTC 28 Jun	10	QLCS	36.6	−79.8	
1900 UTC 24 Jul	0100 UTC 25 Jul	0900 UTC 25 Jul	14	QLCS	42.1	−80.1	
1700 UTC 20 Aug	0400 UTC 21 Aug	0700 UTC 21 Aug	14	QLCS	42.0	−86.4	
1600 UTC 28 Aug	0400 UTC 29 Aug	0600 UTC 29 Aug	14	QLCS	41.3	−89.2	*
1300 UTC 30 Aug	2300 UTC 30 Aug	0100 UTC 31 Aug	12	QLCS	36.4	−82.8	
0400 UTC 16 Jun	1100 UTC 16 Jun	1700 UTC 16 Jun	13	Clusters	39.7	−76.5	
1400 UTC 20 Jun	0200 UTC 21 Jun	0400 UTC 21 Jun	14	Clusters	47.3	−98.9	*
1900 UTC 22 Jun	0100 UTC 23 Jun	0500 UTC 23 Jun	10	Clusters	39.5	−79.2	*
1700 UTC 26 Jun	0000 UTC 27 Jun	0500 UTC 27 Jun	12	Clusters	32.8	−89.4	*
1800 UTC 27 Jun	0100 UTC 28 Jun	0700 UTC 28 Jun	13	Clusters	30.4	−97.6	
1500 UTC 3 Jul	2000 UTC 3 Jul	0500 UTC 4 Jul	14	Clusters	31.0	−84.9	*
1600 UTC 30 Jul	0000 UTC 31 Jul	0300 UTC 31 Jul	11	Clusters	33.0	−89.6	
1500 UTC 31 Aug	2100 UTC 31 Aug	0200 UTC 1 Sep	11	Clusters	32.7	−95.8	
1500 UTC 25 Jun	2100 UTC 25 Jun	0200 UTC 26 Jun	11	Cellular	32.2	−85.0	
1500 UTC 28 Jun	2300 UTC 28 Jun	0400 UTC 29 Jun	13	Cellular	29.9	−85.3	
1500 UTC 2 Jul	2000 UTC 2 Jul	0000 UTC 3 Jul	9	Cellular	38.5	−81.1	*
1700 UTC 24 Jul	2200 UTC 24 Jul	0300 UTC 25 Jul	10	Cellular	31.0	−86.2	
1400 UTC 25 Jul	1900 UTC 25 Jul	0100 UTC 26 Jul	11	Cellular	32.5	−87.6	*
1600 UTC 9 Aug	2100 UTC 9 Aug	0400 UTC 10 Aug	12	Cellular	33.8	−84.7	
1700 UTC 19 Aug	2300 UTC 19 Aug	0200 UTC 20 Aug	9	Cellular	43.2	−94.2	
1400 UTC 22 Aug	1900 UTC 22 Aug	0300 UTC 23 Aug	13	Cellular	39.8	−74.7	*

reflectivity area bins. As such, both models show the best correspondence with the MRMS at the longest lead times for reflectivities less than ~35 dB*Z*; it is also evident that RRFS tends to have less bias than the HRRR at these lower thresholds. In contrast, for reflectivities greater than 40 dB*Z*, both models overpredict the mean area with RRFS having the larger overprediction bias.

Finally, it is seen that there is an unrealistic spike in the distribution of mean reflectivity area for the analyses (forecast hour 0) of both models. The spike in the HRRR data has a sawtooth-like pattern between 18 and 32 dB*Z* while RRFS has a more Gaussian-looking spike between 22 and 42 dB*Z*. Both spikes lie well above the observed mean area within the respective ranges, while the assimilated coverage of lower reflectivity values is significantly underestimated. These spikes in reflectivity result from the interplay between the dynamical core of each model and the cloud analysis methodology which utilizes simplified equations to compute the hydrometeor characteristics (i.e., rain and snow mixing ratios) from radar reflectivity (Weygandt et al. 2022). Interestingly, these spikes in the 0-h reflectivity are entirely gone by forecast hour 1.

An example of how objects are identified by TITAN when using a 35-dB*Z* threshold is shown in Fig. 4. Any 0.05° × 0.05° pixel or collection of adjacent pixels exceeding 35 dB*Z* is classified as a convective storm. Note the allowance of gaps between small scale storm elements within a given polygon. Each TITAN storm object is stored as a polygon with up to 72 vertices that surround groupings of related pixels including the small gaps between convective elements. The area of each storm was determined by summing the area of each pixel (colored blue in Fig. 4) within each TITAN polygon (red outlines in Fig. 4). (The few tiny green areas are where the interpolated radar data used by TITAN slightly smooths out the reflectivity, thus its interpolated value that was previously barely exceeding 35 dB*Z* no longer exceeds this threshold; this was done so as to not over-/underestimate storm area as much as possible between the different products with varying source resolutions.) As evident from Fig. 4, this technique produces far more accurate estimates of individual storm area than the TITAN polygon. This is especially important for the larger storms because the 72-vertex limit of TITAN does not allow it to tightly encompass larger storms with complex shapes (e.g., see storm object indicated by the black arrow in Fig. 4). The number of storms at each time for each case is given as the number of TITAN storm polygons, while the

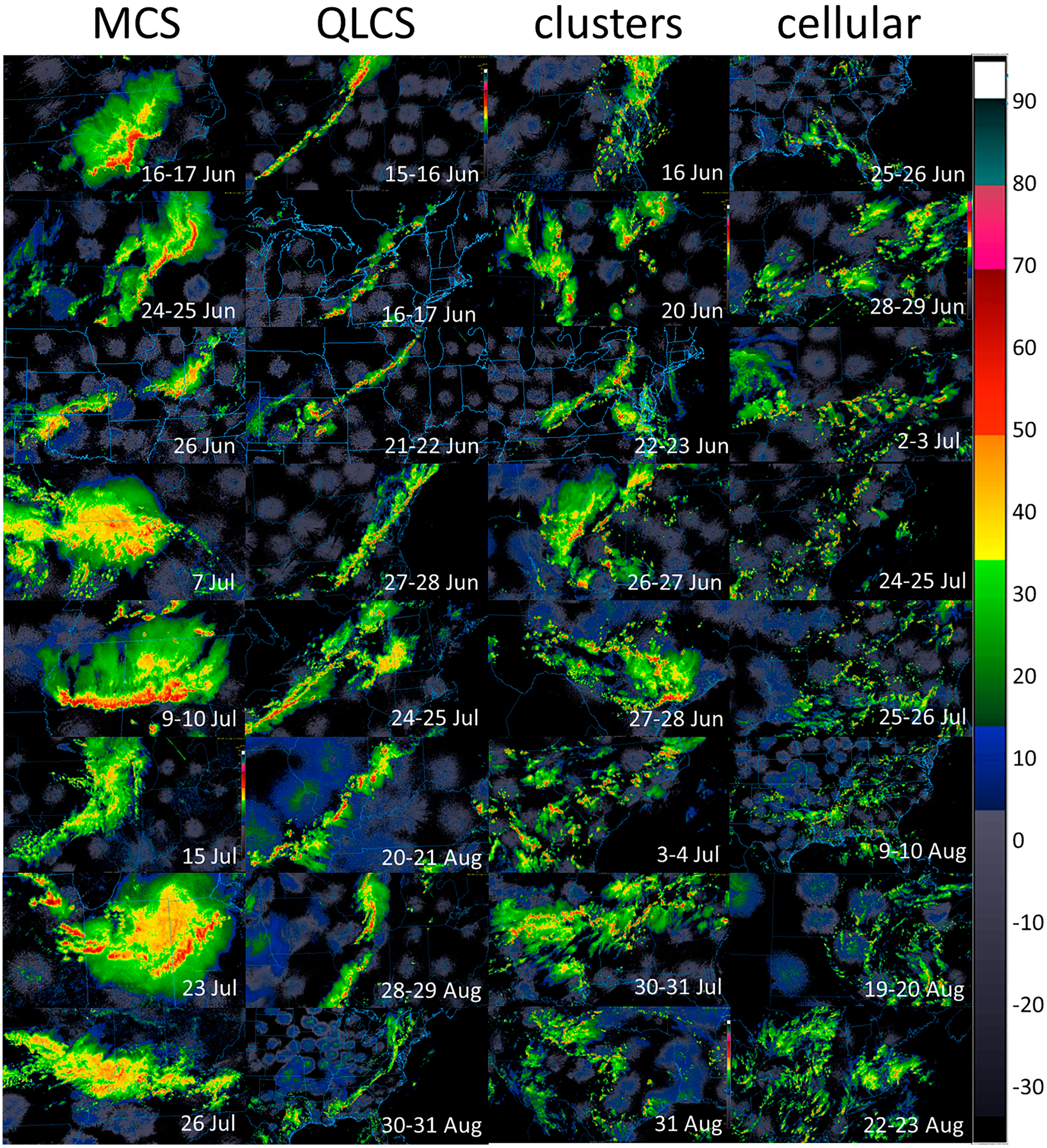

FIG. 2. Radar depictions from the College of DuPage (https://weather.cod.edu/satrad/index.php) of the 32 convective cases during the mature stage of each event from the summer of 2022. The cases are ordered by date and event type.

mean individual storm area is computed by dividing the total storm area by the number of TITAN polygons.

For each storm event, the evaluation area was determined using the entire history of modeled and observed storm objects (e.g., Fig. 5). A large case-specific polygon was then manually drawn around each event that encompassed the entire area where MRMS, HRRR, and RRFS had substantial convection. In this way, the observed (MRMS) and modeled (both HRRR and RRFS) convective storm systems are matched for each storm event. Only those individual TITAN storm objects whose center points are within the large manually drawn case-specific polygons are used in the calculation of storm statistics. This methodology allows for focusing the assessment on the statistical properties of the convective

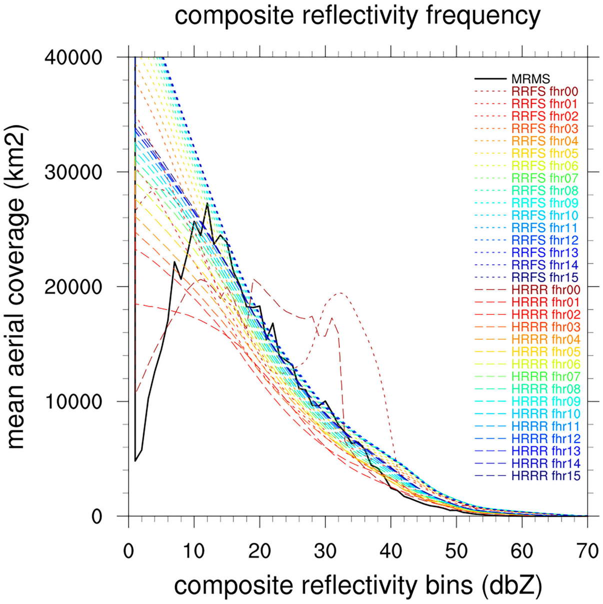

FIG. 3. Hourly mean frequency of pixels within 1-dBZ-wide bins for all times in JJA 2022, except those when RRFS used ensemble variational radar-reflectivity data assimilation, for MRMS (thick, solid black), HRRR (rainbow-colored long dashes), and RRFS (rainbow-colored short dashes) over the continental United States and far southeastern Canada east of 105°W and south of 51°N for all forecast lead times.

storm areas while not requiring any of the storm areas to overlap. However, as can been seen in Fig. 5, there tends to be both significant overlap of the storm areas and also notable differences in the overall extent of convection in each dataset. Figure 6 shows the polygons for all 32 cases. The evaluation regions nearly cover the entire United States east of the Rocky Mountains and southern Canada, with the greatest number of cases included in this study occurring over the southeastern United States.

To assess the representativeness of the macroscale properties of the 32 convective storm systems selected for this study, the distribution of observed storm sizes obtained from the cases is compared with that obtained for the entire 92-day period: 1 June–31 August 2022 (Fig. 7). The area limits (km^2) of the 20 bins used to define the probability density function were varied with size to account for much more numerous small convective storm cells, according to the equation: $\mathrm{bin}_i = 100 \times 10^{0.2i}$, where i is the bin limit number, ranging from 0 to 20. The shape of the observed storm size distribution obtained for the 32 cases selected for this study is very similar to that obtained for the entire summer, indicating the general applicability of the findings. It is noted that the 32 cases had a higher frequency of occurrence for nearly all storm sizes larger than 400 km^2, with the difference being most notable for storms exceeding 20 000 km^2. This indicates that the performance statistics presented below are slightly weighted toward convective systems that include larger storms, as expected since this study focuses more on widespread convective cases.

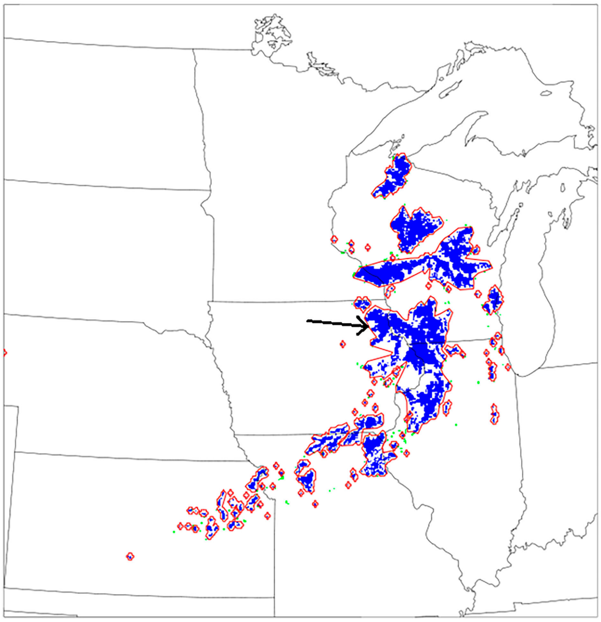

FIG. 4. MRMS depiction of raw composite reflectivity greater than 35 dBZ (blue and green shades) at 1200 UTC 15 Jul 2022. The TITAN-identified objects are outlined in red, while the area of each storm within a TITAN object is indicated by blue, and tiny sections of storms outside of TITAN objects are shaded green. A black arrow points to the TITAN storm object mentioned in the text.

3. Results

The object-based storm identification technique discussed above was used to identify the macroscale properties of the convective systems (total storm area, total storm count, and individual storm area) for 32 total cases. Each case was selected based on the detection of a widespread storm system in the MRMS reflectivity data and subsequent matching of the

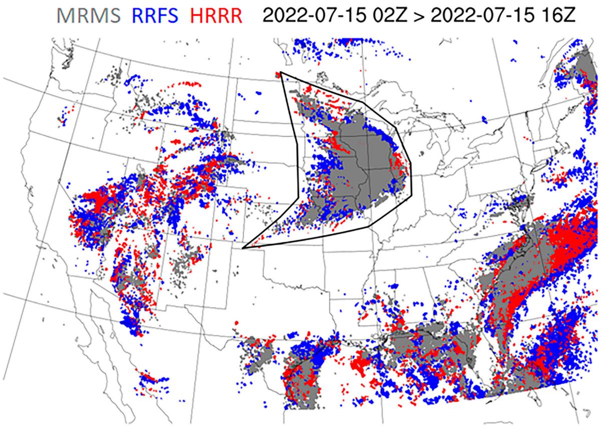

FIG. 5. All areas of MRMS (gray), HRRR (red), and RRFS (blue) composite reflectivity exceeding 35 dBZ during the duration of the MCS case on 15 Jul 2022. The black outline indicates the subjectively defined event area encompassing the case.

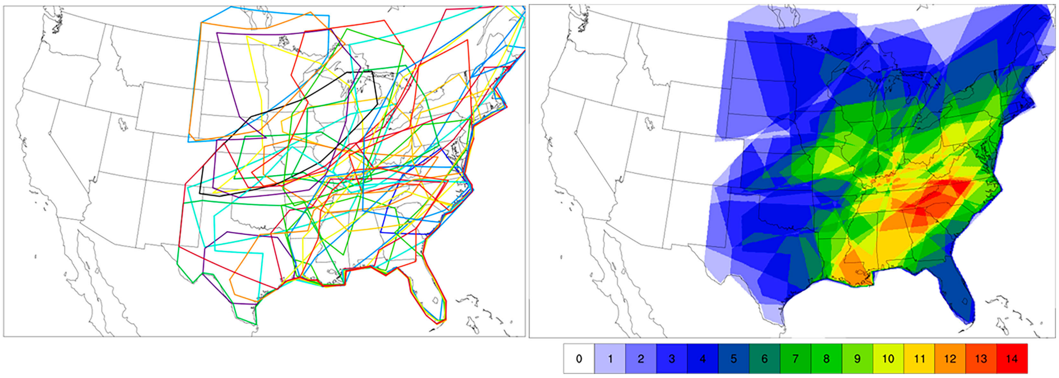

FIG. 6. (left) Polygons outlining the maximum extent of each of the 32 cases used in this study. Lines are colored by date from 1 Jun to 31 Aug 2022: blue–green–yellow–orange–red–maroon. (right) The number of polygons encircling each pixel over the domain.

model predicted storm areas. An important caveat of this matching approach is that the cases were conditionally sampled based on the observed composite reflectivity, so it does not consider model storm systems that were substantially more developed than in the observations.

a. Case study

The verification technique is demonstrated using an MCS event that occurred over the Midwestern United States on 15 July 2022 (Fig. 8). At 0000 UTC, daytime convection was weakening near the North Dakota–Minnesota border (not shown.) At 0200 UTC, MRMS observations indicate that new convection initiated on the northern end of an area of otherwise decaying precipitation. Based on this storm initiation time, the forecasts issued at 0100 UTC were selected for model evaluation for this case. By 0700 UTC, the system started to quickly grow upscale resulting in a 700-km-long northwest–southeast-oriented broken line of storms (falling into the QLCS category at that time) that extended from near Fargo, North Dakota, to central Wisconsin (Fig. 8a). Both models are delayed in the development of convective area by 0700 UTC with RRFS only capturing the convection over the North Dakota–Minnesota border while HRRR captured the QLCS over Minnesota–Iowa. Both models have little of the weaker convective area observed over Iowa and northwestern Illinois (Figs. 8a–c). Over the next few hours, the QLCS continued to strengthen such that by 1200 UTC, several individual storm elements had converged into a large moderately well-organized MCS (Fig. 8d), stretching from northern Wisconsin into eastern Kansas. Both models also have a good representation of the MCS at this stage with the RRFS model capturing the southern extent better and the HRRR better capturing the structure of convection over Wisconsin (Figs. 8d–f).

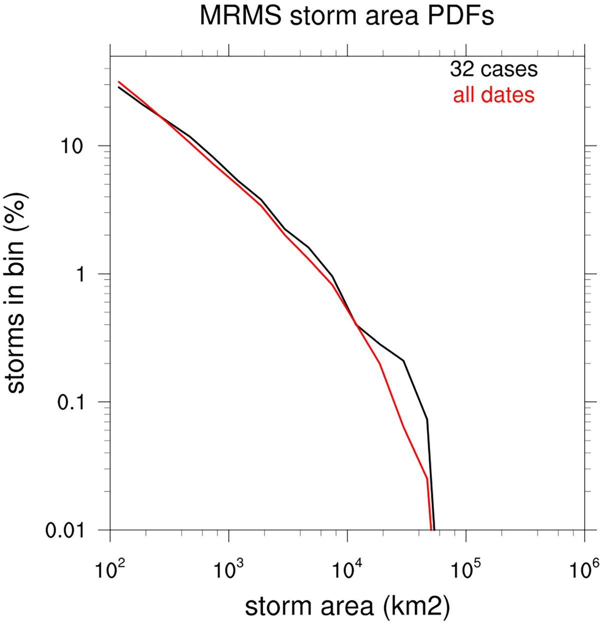

FIG. 7. PDF of MRMS storm area, as defined by reflectivity > 35 d*BZ* over the United States and Canada east of 105°W and south of 51°N. The black line is the PDF for all dates and times encompassing the 32 cases evaluated in this study, while the red line is the PDF obtained using MRMS data from the entire summer: 1 Jun–31 Aug 2022 during the period that both models used the same DA scheme.

Figure 9 shows a time series of modeled and observed storm characteristics for this case, using the validation region encompassing its entire evolution obtained from both models and the observations (see Fig. 5). The observations and both models had minimal convective area at 0200 UTC (Fig. 9) with very limited storm development between 0200 and 0600 UTC. Between 0600 and 0700 UTC, the observations indicated a burst of rapid storm development that both models were late to capture with rapid growth (as indicated by sudden change in the slope of the lines for total storm area) starting one hour late in the RRFS and two hours late in the HRRR (Fig. 9a). MRMS reached peak aerial coverage at 1200 UTC, while RRFS and HRRR maximized an hour later, after which they all decreased in coverage through the end of the 15-h forecast.

The MRMS total storm count increased from 9 discrete storm objects at 0200 UTC to a maximum of 86 storm objects at 1000 UTC (Fig. 9b); RRFS and HRRR started with similar

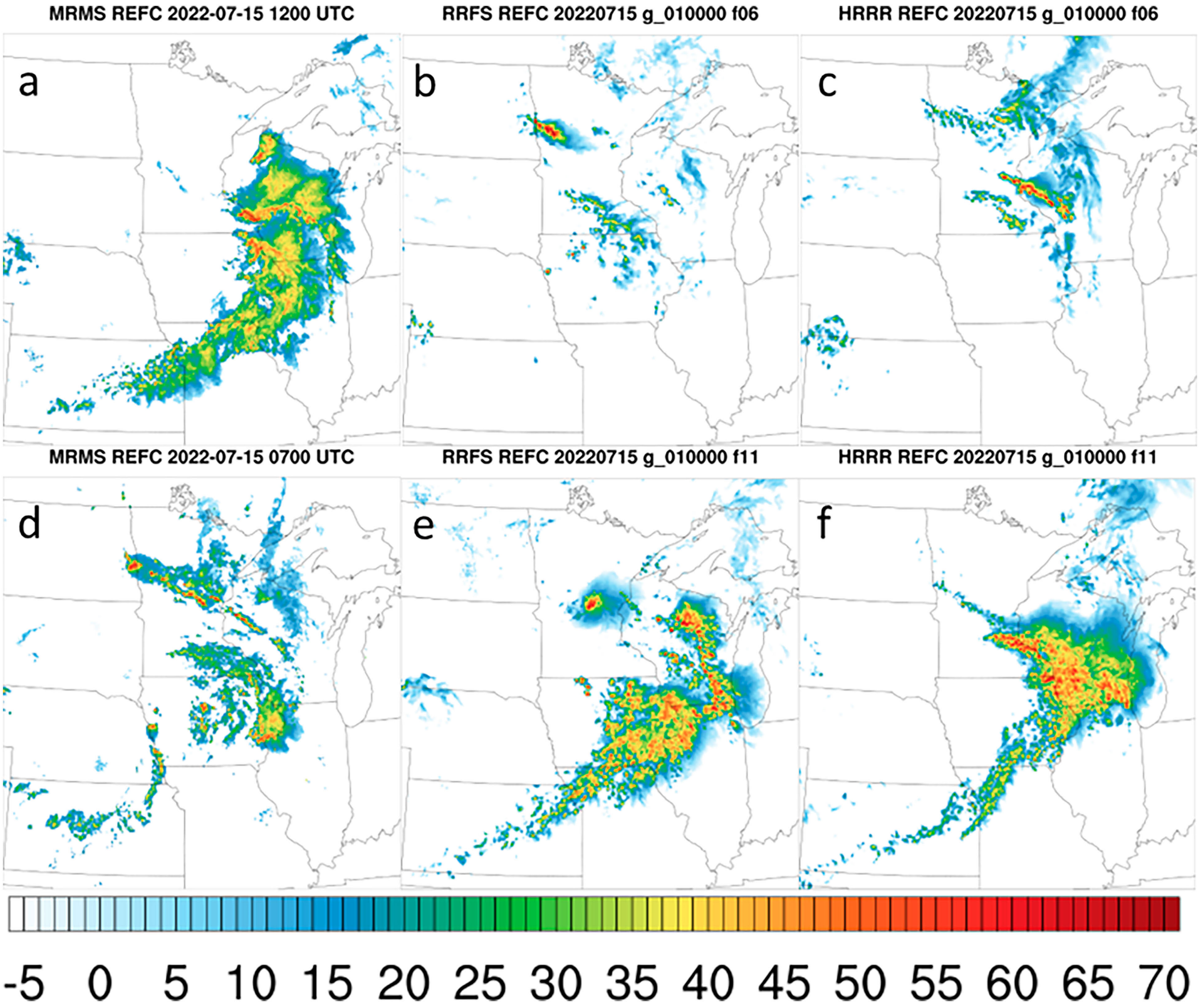
MRMS REFC 2022-07-15 1200 UTC
RRFS REFC 20220715 g_010000 f06
HRRR REFC 20220715 g_010000 f06
a
b
c
MRMS REFC 2022-07-15 0700 UTC
RRFS REFC 20220715 g_010000 f11
HRRR REFC 20220715 g_010000 f11
d
e
f
-5 0 5 10 15 20 25 30 35 40 45 50 55 60 65 70

FIG. 8. Composite reflectivity from (a),(d) MRMS; (b),(e) RRFS; and (c),(f) HRRR for the 15 Jul 2022 MCS event. Composite reflectivity is valid at (top) 0700 UTC 15 Jul and (bottom) 1200 UTC 15 Jul. Forecasts were issued at 0100 UTC 15 Jul 2022 with forecast hour 6 shown in (b) and (e) and forecast hour 11 shown in (c) and (f).

numbers of storm objects, but both models were delayed at initiating a substantial number of new storms with RRFS being delayed until 0700 UTC and HRRR being delayed until 0800 UTC. In fact, the HRRR significantly underpredicted the storm count throughout the simulation whereas RRFS storm counts more closely match the observed counts from 0800 UTC onward.

Mean individual storm area (MISA, Fig. 9c) was calculated by dividing the total storm area at a given time by the total storm count. In this case, both models were similar to the observed evolution of MISA for the first nine hours indicating compensation of biases in storm area (too small) and storm counts (too few) but that the models were able to capture the general trend of upscale storm growth. Starting around 1200 UTC, the modeled MISA were generally much larger than observed, with the HRRR having larger storms than the RRFS.

Identical analyses were performed for all 31 other cases (not shown) the results of which were then composited (using the technique described below) to assess the RRFS model performance relative to the HRRR as a function of convective mode.

b. Calculation of composite results

A normalization approach was used for compositing, whereby the maximum of the observed storm characteristic (total area or total count) from MRMS for a given case (spanning all case times) was used to normalize the results for that metric across all event times t. For a given case, the normalized MRMS-derived storm characteristic $O(t)$ is defined as

$$O(t) = \frac{O(t)}{O_{\max}},$$

where $O(t)$ is the observed storm characteristic at a given time for a given case, and $O_{\max}$ is the maximum observed storm characteristic value for a given case. Therefore, $O(t)$ varies between 0 and 1. Similarly, the normalized model

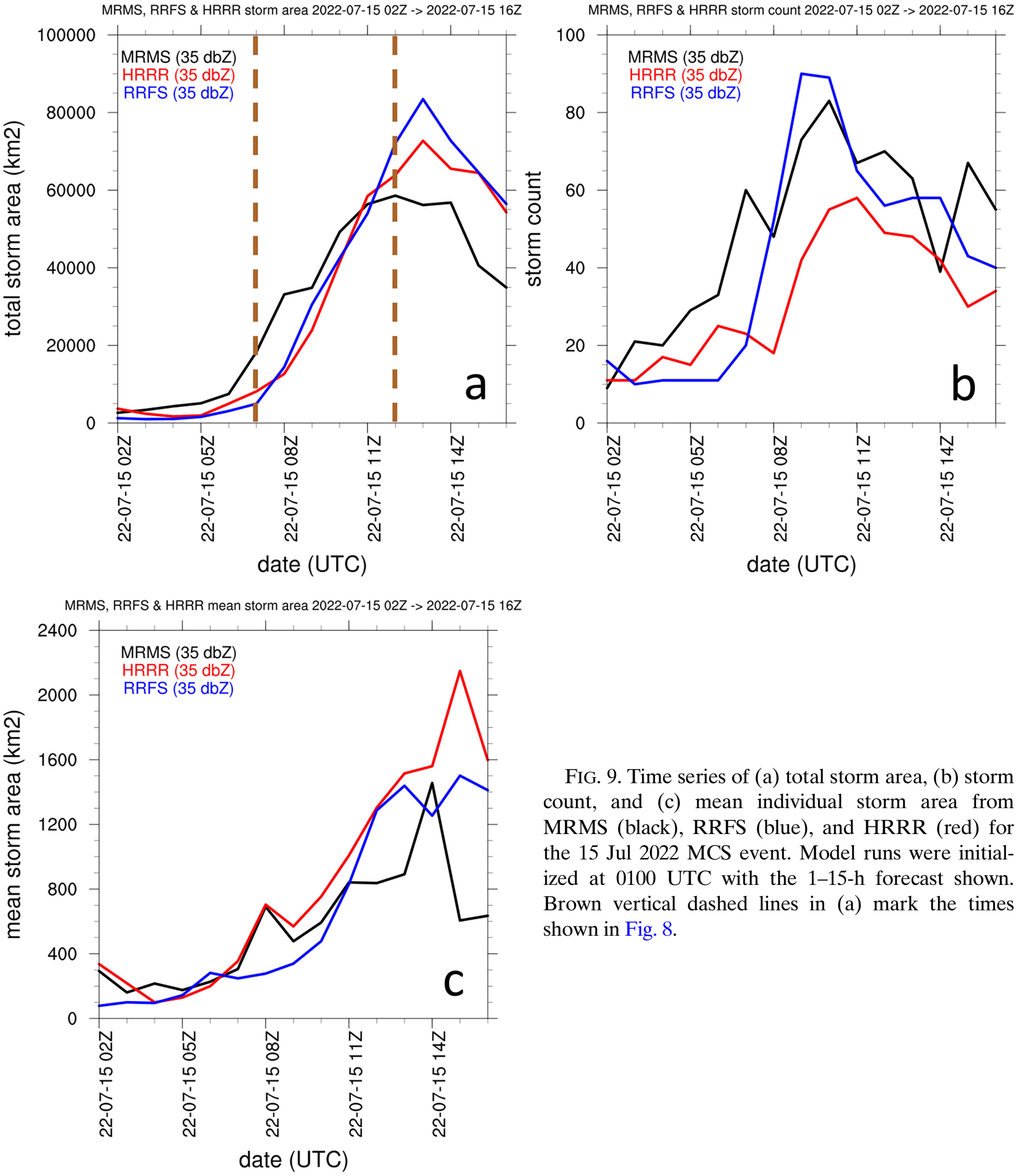

FIG. 9. Time series of (a) total storm area, (b) storm count, and (c) mean individual storm area from MRMS (black), RRFS (blue), and HRRR (red) for the 15 Jul 2022 MCS event. Model runs were initialized at 0100 UTC with the 1–15-h forecast shown. Brown vertical dashed lines in (a) mark the times shown in Fig. 8.

values for each storm characteristic $M(t)$ are related to the maximum observed value:

$$M(t) = \frac{M(t)}{O_{\max}},$$

where $M(t)$ is the model storm characteristic value for a given case and time. It is possible for $M(t)$ to be greater than one if the modeled storm characteristic exceeds $O_{\max}$. Composite values were found by averaging all normalized values available for a given event hour following:

$$\overline{O(t)} = \frac{1}{N}\sum_{j=1}^{N} O_j(t),$$

$$\overline{M(t)} = \frac{1}{N}\sum_{j=1}^{N} M_j(t),$$

where N is the number of cases. Thus, despite using verification polygons that vary between each case, this normalization approach weighs each case's contribution equally to the composite time evolution for each storm characteristic.

To calculate the normalized MISA [referred to hereafter as "individual storm area ratio" (ISAR)], the normalized total storm area at each time for each case was divided by the normalized storm count for the corresponding case and time; this was then averaged over all cases to create the ISAR. This was done since early and late times in an event can have a very small number of storms (the denominator in the ratio), so

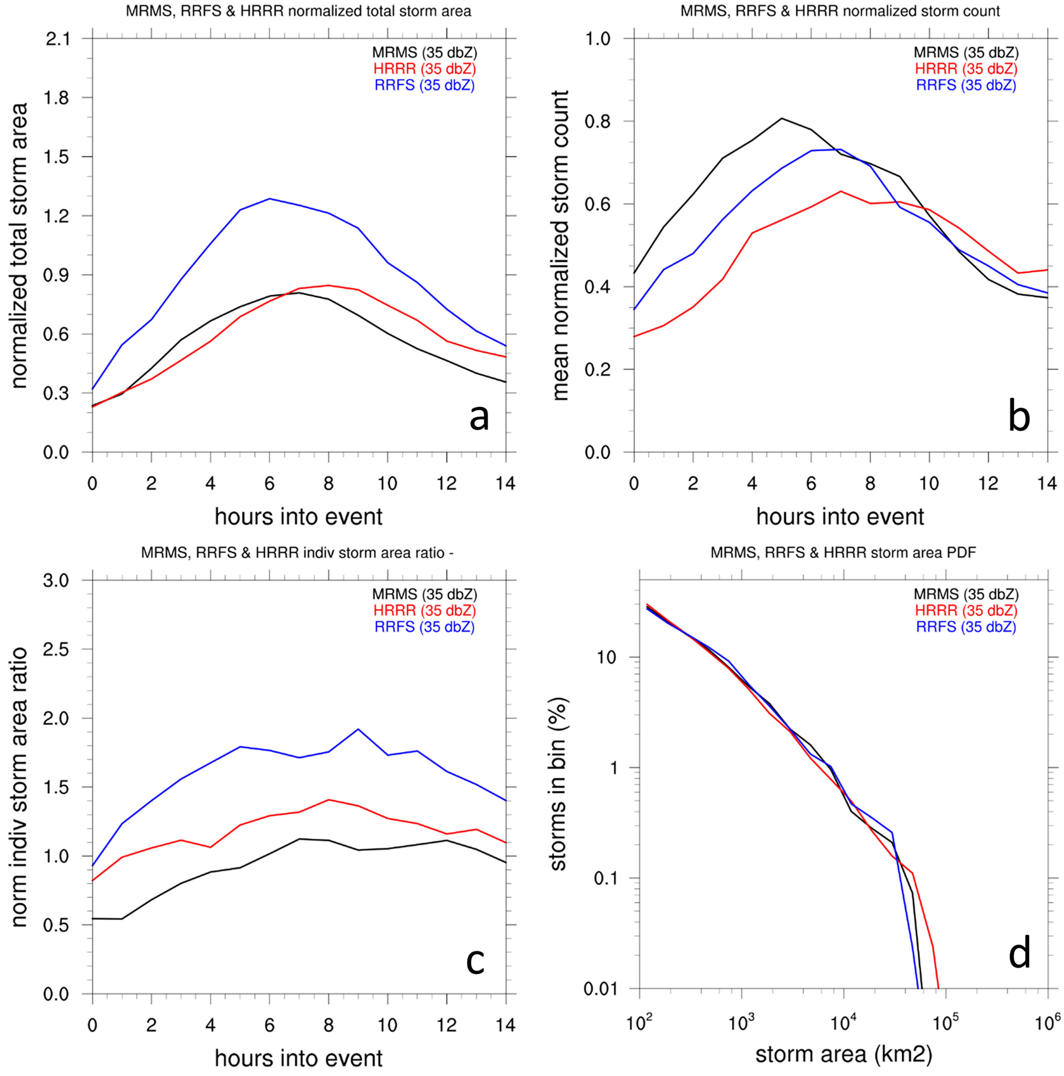

FIG. 10. Normalized storm metrics for all 32 cases for MRMS (black), HRRR (red), and RRFS (blue). (a) Total storm area, (b) storm count, (c) individual storm area ratio, and (d) storm area PDF, using a threshold of 35 dB*Z* for each product.

that splits or mergers can result in wild fluctuations in mean storm area from one hour to the next. It is noted that since the ISAR is a ratio of two other normalized values, the observed ISAR can be greater than one at times.

c. Composite results

1) OVERALL ASSESSMENT AT 35 dB*Z*

Figure 10 compares the time evolution of modeled and observed normalized total storm areas, total counts, and ISARs composited across all 32 cases. The normalized total storm area can be used to assess model skill at predicting the life cycle of widespread convective storm events. The composited MRMS data indicate that averaged across all 32 cases, the storm initiation and growth phase occurs through event hour 5, a mature phase lasts from hours 6 to 8, and the decay phase occurring thereafter (Fig. 10a). (Reminder: the event hour is one hour less than the forecast hour, since the model analysis time is one hour before the event starts. That is, event hours 0 and 1 correspond with forecast hours 1 and 2 and so on). The most striking difference between the two models is that RRFS total storm area is ~40%–60% higher than MRMS throughout the life cycle while the magnitude of HRRR total storm area is fairly similar to MRMS. The other notable difference is that the RRFS better captures the overall timing of the entire life cycle, while in HRRR the life cycle appears to be out of phase by 1 h late. It is also evident that the growth and decay rates in RRFS are both faster than observed resulting in a much larger amplitude of normalized storm area than is observed or simulated by the HRRR.

The observed normalized storm count increases rapidly during the initiation and growth phase (i.e., first 5 h) of the

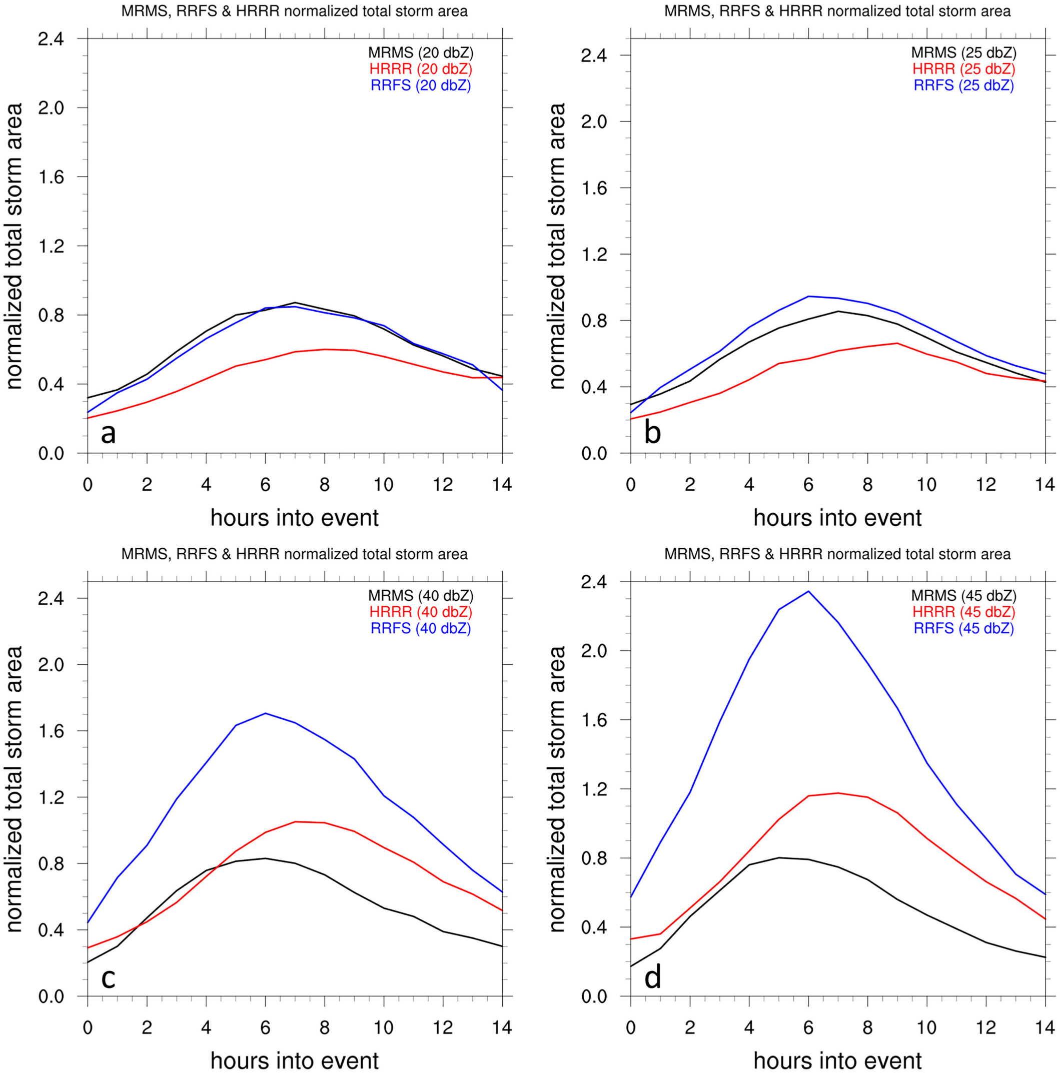

FIG. 11. As in Fig. 10a, but using 20-, 25-, 40-, and 45-dBZ thresholds.

event (Fig. 10b). This is followed by a period of slightly decreasing counts as storms grow up scale during the mature phase of the storm until hour 9 when a more rapid decrease with time indicates storms are dissipating. Both models have fewer storm objects than observed during the initiation and growth phase of the storm system life cycle indicating an underprediction and delay in the number of storms initiated. This bias is evident at forecast hour 1 (i.e., 0 h into event) and carries through several hours into the simulation as both an early undercount and a delay in the maximum count (Fig. 10b). It is seen that the modeled peak storm count occurs 2 h later than observed in both models, but it is better represented by the RRFS than the HRRR.

Given that the RRFS dramatically overpredicts storm area and also tends to underpredict storm counts (especially during the initiation and growth period), it is found that the RRFS ISAR (an indicator of storm size) is up to 90% greater than observed (Fig. 10c). The HRRR's bias in ISAR tends to be around half that found for RRFS throughout the convective system life cycle. The ISAR reveals that the total area bias in RRFS was due to storms being too big (as indicated by area ratios greater than 1.5) and slightly too few, whereas the limited bias in total area in HRRR can be attributed to compensating errors with storms being both modestly too big and somewhat too few. These results are consistent with those from Wicker (2023), who also found that the FV3 dynamic core used in RRFS has a higher frequency of elevated CAPE, producing storms with deeper and stronger updrafts, larger convective cores, and more intense precipitation rates in idealized simulations. Both models generally get the proportion of small storms and large storms correct with a slight tendency for HRRR to overpredict the number of large storms (Fig. 10d).

2) ASSESSMENT AT OTHER REFLECTIVITY THRESHOLDS

Separate analyses were performed using additional composite reflectivity thresholds (20, 25, 40, and 45 dBZ) to assess model performance for different aspects of the widespread

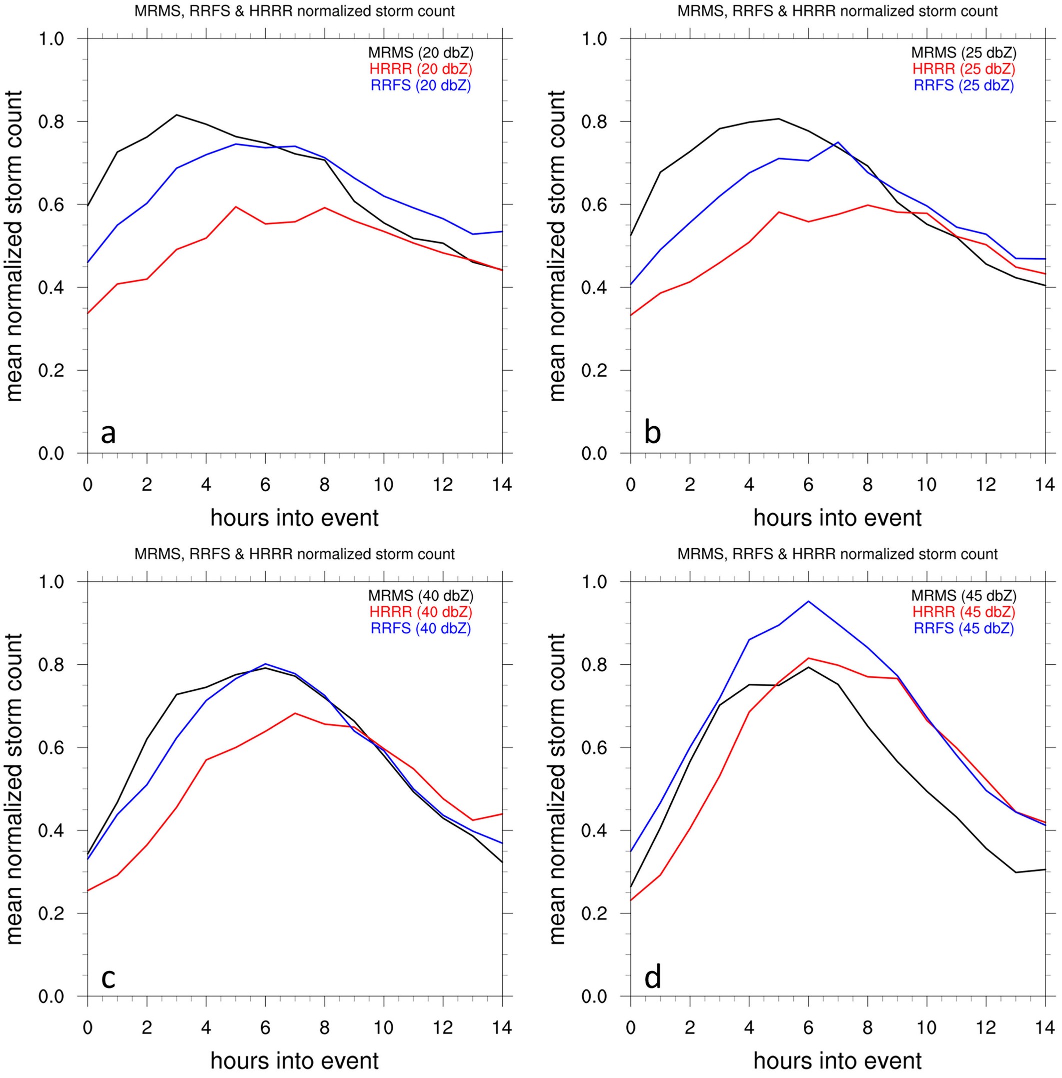

FIG. 12. As in Fig. 10b, but using 20-, 25-, 40-, and 45-dBZ thresholds.

convective storm areas (Figs. 11–13). The lower thresholds are used to focus more on the evaluation of the stratiform regions of precipitation within the widespread event, while the higher thresholds focus on the representation of the convective cores in the models. For the lower stratiform precipitation-encompassing thresholds RRFS nearly matches the observed normalized total storm area (Figs. 11a,b) in both timing and amplitude. In contrast, large biases are evident in the RRFS at thresholds more indicative of the convective cores with biases increasing as a function of increasing threshold and exceeding 100% at thresholds of 40 dBZ and greater (Figs. 11c,d). On the other hand, HRRR underpredicts the total area coverage at lower dBZ thresholds by 25% (Figs. 11a,b) pointing to a tendency to underpredict the coverage of stratiform rainfall associated with convection. At the same time HRRR also overpredicts the area associated with the convective cores, but at a much lower magnitude (e.g., 50% at 45 dBZ) than RRFS (Figs. 11c,d). It is interesting to note that the growth of both convective and stratiform rain areas are delayed in the HRRR with peak values occurring 1–2 h late depending on the threshold used. These bias increases with increasing reflectivity threshold are to be expected when considering the reflectivity frequencies shown in Fig. 3, as the model frequencies change from being equally frequent or slightly less at 20 dBZ (depending on the forecast hour) to universally more frequent at 40 dBZ and above.

The skill of the model at predicting the evolution of storm counts was also explored as a function of threshold (Fig. 12). At the lower thresholds that encompass both stratiform precipitation and convective cores (i.e., 20 and 25 dBZ), both models tend to underpredict the peak storm counts indicating that a notable fraction of the observed CI events were missed. This finding is similar to that shown in Fig. 10b with both models underpredicting storm counts during the storm initiation and growth stage starting at event hour 0. However, at the 40-dBZ threshold, the RRFS predictions of storm counts (indicating the initiation of new convective cores) are notably

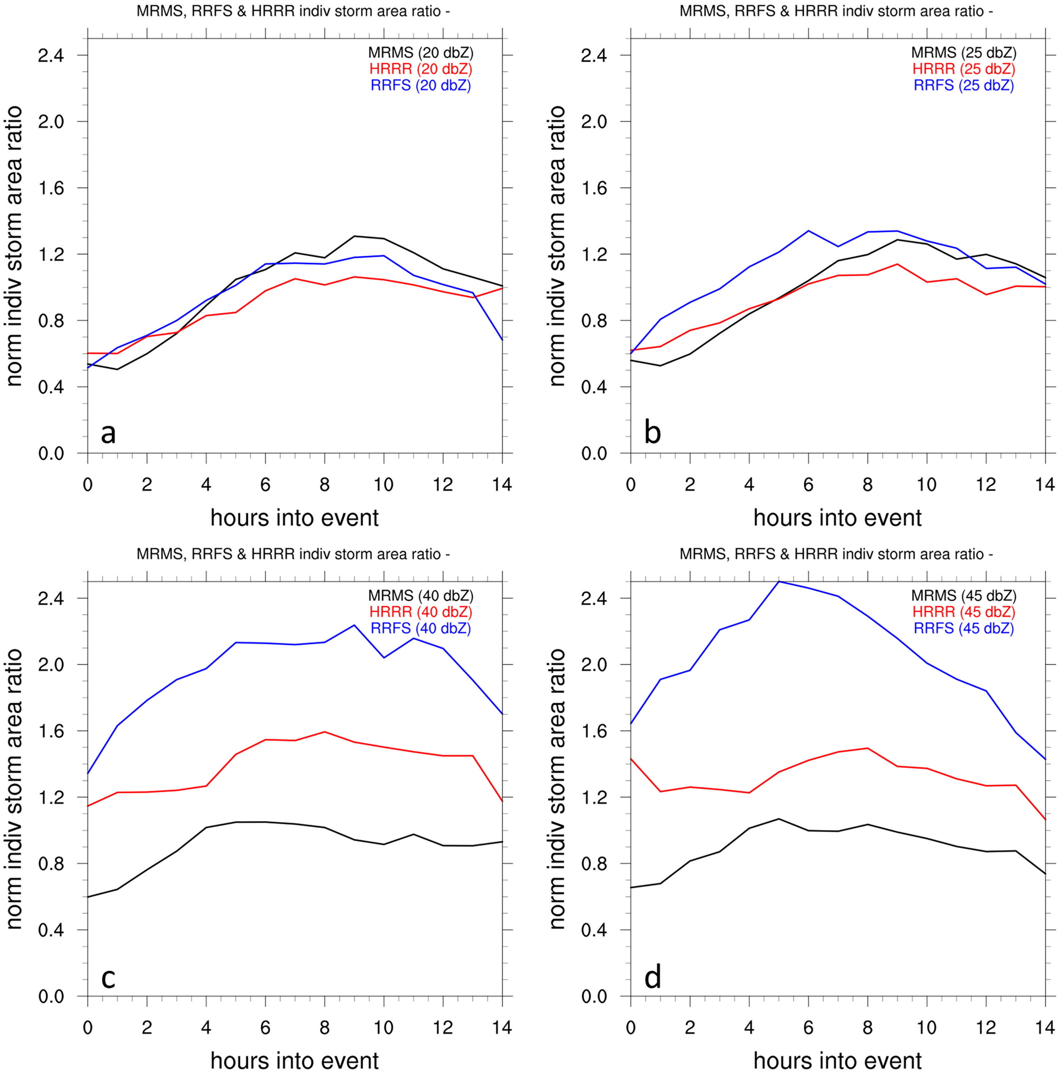

FIG. 13. As in Fig. 10c, but using 20-, 25-, 40-, and 45-dBZ thresholds.

better than that found for the HRRR, which is still delayed and underpredicts the peak count (Fig. 12b). At 45 dBZ, the RRFS has too many convective cores due to a longer period of CI as compared to MRMS while the HRRR has the correct number of CI events albeit the center of the peak appears to be 2 h later than observed.

The ISARs shown in Fig. 13 demonstrate two performance regimes. For the lower thresholds (which encompasses both stratiform regions and convective cores), the modeled evolution of this individual storm size property matches the observed values fairly well with HRRR storm objects being generally smaller than RRFS. The sizes of the more convective storm portion of the storm objects tend to be larger than observed, similar to that shown in Fig. 10c. However, it is found that this bias tends to change less for the HRRR than for the RRFS (Figs. 13c,d). That is, the most intense portion of the storms tend to be increasingly too large in the RRFS model. In fact, at a threshold of 45 dBZ, the RRFS storm objects are up to 140% larger than observed (Fig. 13d).

3) ASSESSMENT BY MATURE STORM ORGANIZATION

Model performance is also evaluated as a function of mature convective mode using the 35-dBZ threshold. Results are obtained by compositing across the eight cases selected for each convective mode listed in Table 1. (See section 3b as a reminder for how the storms were normalized and composited.) The most organized and largest storm type, MCSs, (mean peak area of 45 717 km^2), with a substantial trailing stratiform region and stronger cold pool, showed both models were late in reaching the peak area and underestimated it by 15% (Fig. 14a). While the peak values were similar for RRFS and HRRR, differences in the growth period and growth rates are evident with the HRRR growth rate being too slow and delayed by ~2 h. Note that a percentage difference of 15% corresponds with an absolute error in storm area of 5450 km^2 for the mean case at peak

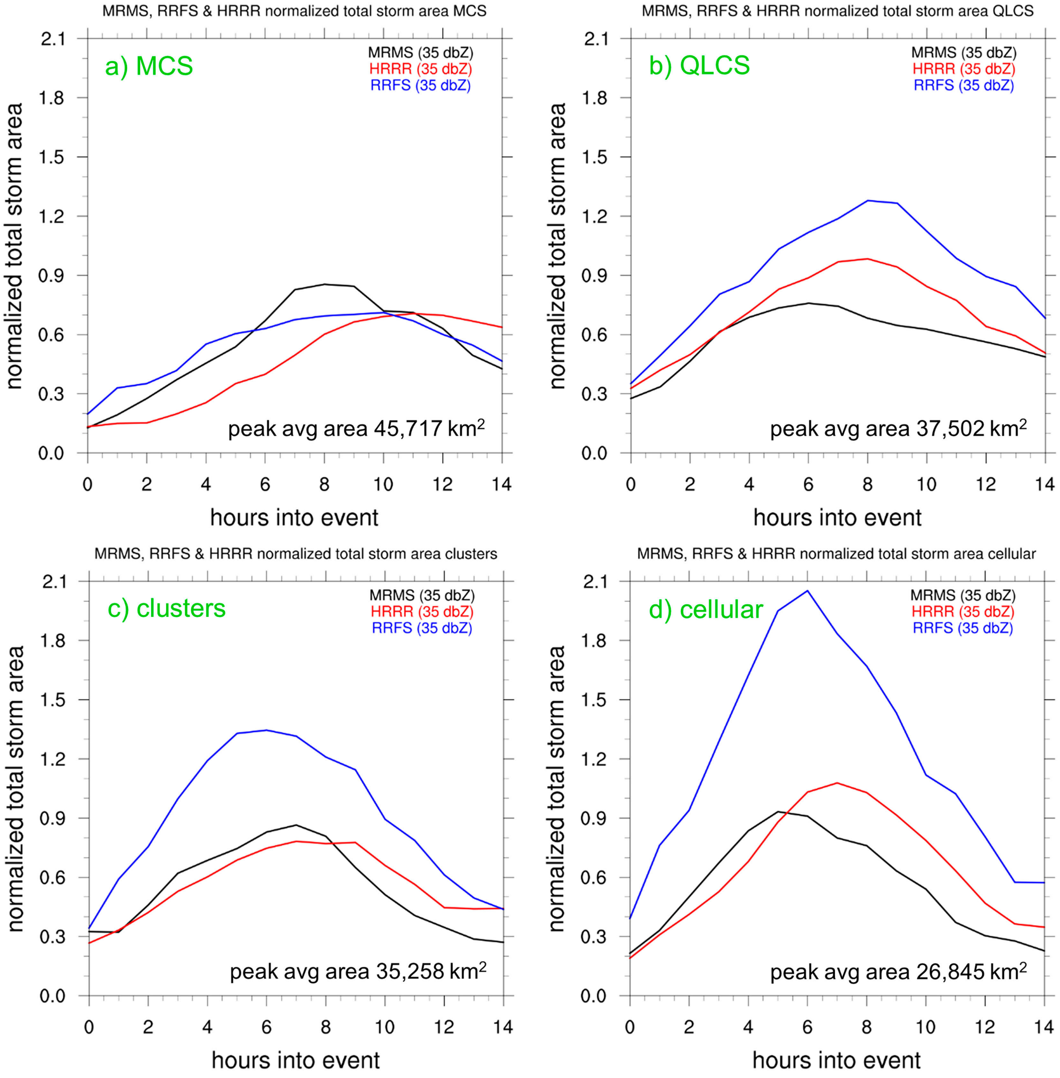

FIG. 14. Normalized total storm area using a 35-dBZ threshold, subset by mature convective mode: (a) MCSs, (b) QLCSs, (c) clusters, and (d) cellular. Mean peak storm area for each mode from MRMS is given near the bottom of each panel.

aerial coverage. Differences between the models and MRMS are much larger for the QLCS convective mode (MRMS mean peak area of 37 502 km^2) with HRRR and RRFS peaking 40% and 85% higher than MRMS, respectively (Fig. 14b). These overestimates in total storm area can be attributed, in part, to the modeled growth period being 2 h longer than observed. Also of note in the QLCS mode is that the model storm dissipation rate is faster than observed. For the clusters convective mode, the HRRR captured the growth rate and amplitude of storm life cycle fairly well. In contrast, the RRFS had a much more rapid storm growth rate, overpredicting the observed peak area by nearly 60% (Fig. 14c), compared with the MRMS peak mean area of 35 258 km^2. Finally, there are dramatic differences in the representation of the evolution of cellular storms between the two models. The RRFS growth rate is far too large for this convective mode such that, despite capturing the timing of the peak better than the HRRR, the RRFS peak area was 120% larger than observed (Fig. 14d). Meanwhile, the amplitude of the cellular storm life cycle (MRMS mean peak area of 26 845 km^2) is much better captured by the HRRR, albeit with a 2-h delay in the peak area as compared with MRMS.

Exploring the model skill at predicting storm counts reveals that both models exhibit a similar underprediction bias in the number of storms during the initiation and growth phase regardless of convective mode (Fig. 15). This negative bias is most evident in the first 4–6 h of the convective system life cycle and tends to be larger for the HRRR than RRFS. Comparing Figs. 14 and 15, it is evident that the HRRR bias in total storm area is due, in part, to its underprediction of new storm initiations (too few storms) consistent with that reported by James et al. (2022). Moreover, the largest discrepancy between the models in predicting storm counts is evident for the

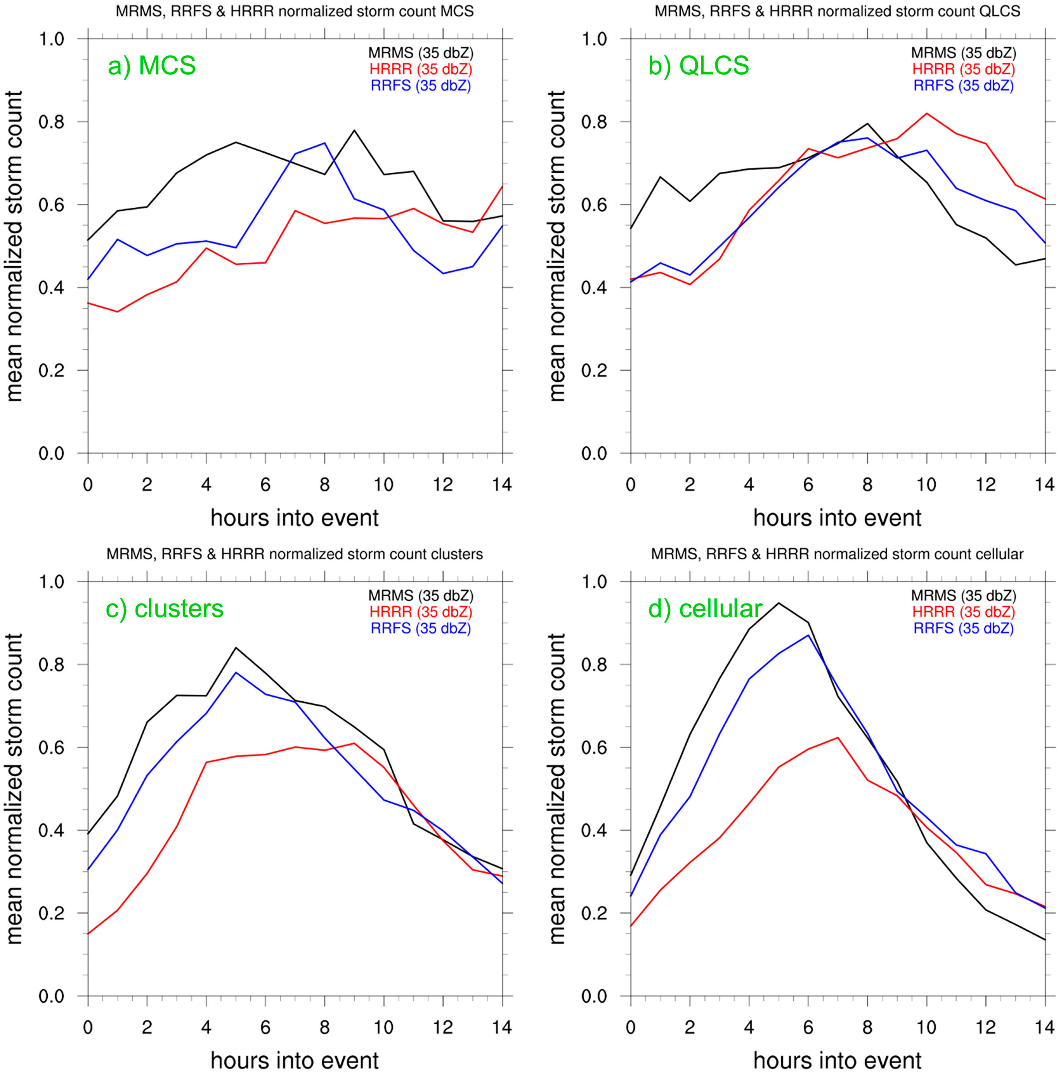

FIG. 15. Normalized total storm count using a 35-dB*Z* threshold, subset by mature convective mode.

clusters and cellular convective modes (Fig. 15d). The RRFS better captures the observed peak in counts for these two convective modes, while the HRRR underpredicts the peak counts for both convective modes reaching an underprediction bias of 40% for the cellular convective mode. However, the fact that RRFS captures the counts fairly well while drastically overpredicting total storm area (Fig. 14d) indicates that the individual storms predicted by RRFS are far too large.

Another interesting difference between the two models is that RRFS captures the timing of the peak counts and overall amplitude of evolution in storm counts throughout the storm system life cycle much better than the HRRR for the clusters and cellular storm categories. The counts predicted by the HRRR model tend to be delayed by 1–2 h for these two convective modes while the timing of peak values obtained with RRFS are generally much closer to that observed. Thus, while HRRR generally performed better at predicting total storm area (Fig. 14), RRFS performed better in storm counts (Fig. 15).

Finally, the ISAR is shown in Fig. 16. This metric provides an indication of model skill at predicting mean storm size. It is seen that both models handle the evolution of mean storm size for the MCS category fairly well compared with the other modes, with mean storm size peaking later in the models than observed. For all other storm convective modes, both models tend to predict individual storms that are much larger than observed with this bias generally increasing as a function of decreasing storm organization, with the RRFS ISAR up to 140% greater than observed for the cellular mode.

4) ASSESSMENT BY MODEL FORECAST ANALYSIS TIME AND FORECAST LENGTH

Other factors that can affect model performance are the analysis time and forecast length relative to the observed storm initiation time. To investigate these factors, the model

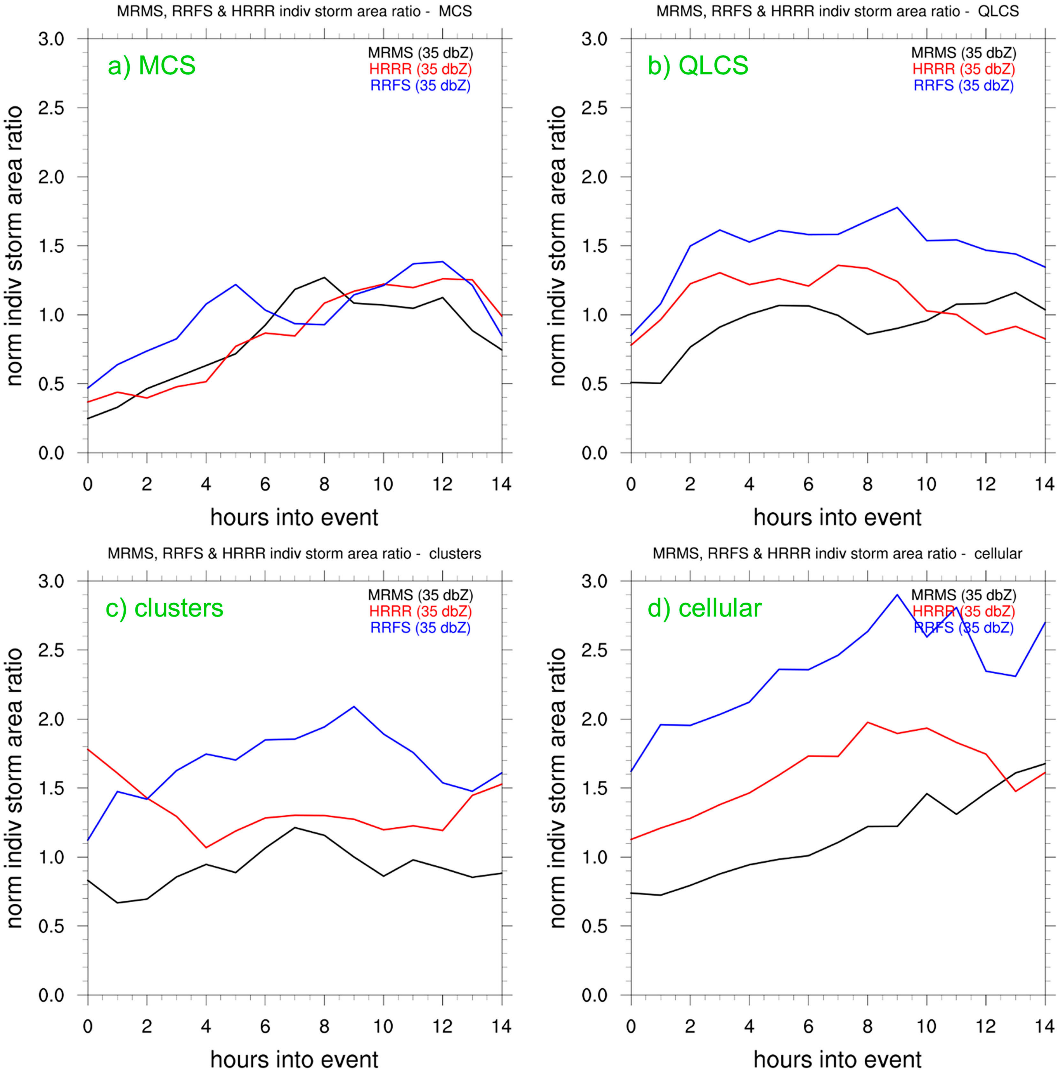

FIG. 16. Individual storm area ratio using a 35-dBZ threshold, subset by mature convective mode.

analysis time was varied from 6 h before to 4 h after the observed storm initiation time (±5 h relative to the primary model analysis time used earlier in this study). Only cases with at most one missing model forecast time out to 15 h were selected, which limited this subset to 13 cases: 4 MCSs, 2 QLCSs, 4 clusters, and 3 cellular (Table 1).

Figure 17 shows the normalized total storm area for MRMS, HRRR and RRFS as a function of hours into the event with hour 0 indicating the observed storm initiation time. Lines for the forecasts initialized after the observed storm initiation time (green-to-purple colored lines) start at x values greater than or equal to zero. As previously described from Fig. 10, the RRFS model predicts larger storm areas than HRRR at all lead times for the forecast issued 1 h before initiation. This plot indicates that bias in RRFS relative to the HRRR is consistently high regardless of whether the convective storm system was present in the observations (i.e., positive lags) or had not yet formed (negative lags). It is noted that the normalized total storm area values for the model analysis time (i.e., forecast hour 0) are not shown in the interest of clarity because they tend to be dramatically higher than all of the forecast hours due to the very high biases found at a threshold of 35 dBZ (Fig. 3).

As seen in Fig. 17, the MRMS observed total storm area increases from event hour 0 to 7, and then gradually decreases thereafter. The HRRR total storm area increases faster than observed during the storm initiation and growth period and are nearly all higher than observed throughout the period, with peak times ranging from 6 to 9 h. The RRFS total storm area is too high and increases much faster than the HRRR for analysis lags from −5 to +1 h, while at lags of +2 h and greater the peak storm area is not biased quite as high. These trends are also inspected as a function of model analysis time for a given valid time of 7 h into the event (Fig. 18). Here it is evident that the best performing analysis lag times are from 1 h before to 3 h after CI with the HRRR, while for the RRFS

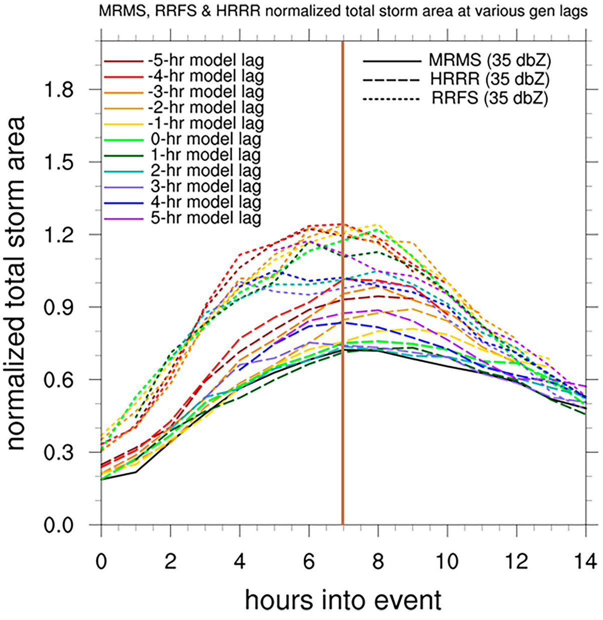

FIG. 17. Normalized total storm area using model analysis times varying from −5 to +5 h relative to the primary model analysis time (−1 h) used in the rest of the article, and only using the subset of 13 cases with sufficient availability across model runs. Lines are colored by model analysis lag and are dashed to differentiate between the products. The vertical brown line at event hour 7 indicates the event hour depicted in Fig. 18.

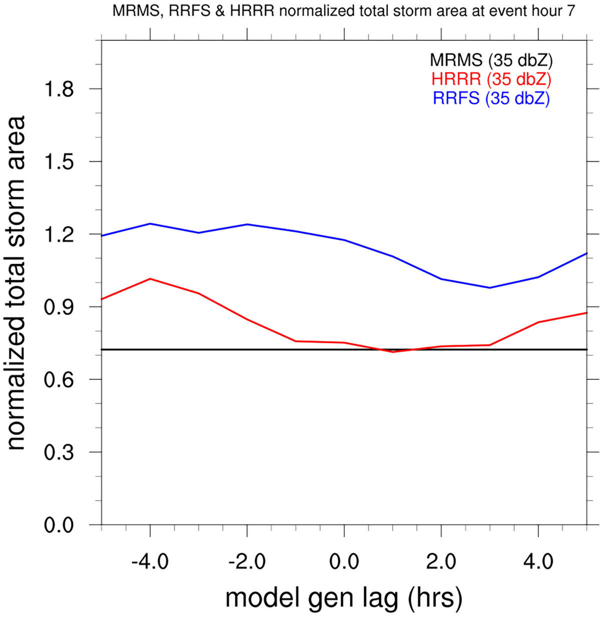

FIG. 18. Normalized total storm area by model analysis lag at event hour 7 (indicated by the brown vertical line in Fig. 17.) A horizontal black line shows the single MRMS value for reference.

the best performing analysis times are 2–4 h after CI. Also of note is that the best performing RRFS runs have biases similar to the worst performing HRRR run (that issued 4 h before observed CI). These results suggest that there may be an optimal stage of storm maturity early in the convective system's life cycle for which the DA and cloud analysis is most effective.

Normalized storm counts were mostly underestimated regardless of the model analysis times, with HRRR generally underestimating the counts more than RRFS, particularly during the storm initiation and growth phase (Fig. 19). Also evident is that the RRFS simulations issued after CI more closely matched the observed counts than the HRRR simulations issued at CI. A similar pattern is evident in the magnitude of the storm counts with the HRRR model having higher peak values at earlier (from −5 to −2 h) analysis lags, and RRFS having higher peak values at earlier (from −5 to 0 h) and very late (+5 h) analysis lags. The biases in modeled ISAR do not appear to have any relationship with analysis time relative to the storm initiation event with all lags showing that both models have much larger area ratios (larger storm sizes) than observed, with RRFS having larger storms than the HRRR (Fig. 20).

4. Discussion and conclusions

This study assesses the skill of two convection-permitting models at predicting the macroscale characteristics of widespread convective events observed during JJA 2022. The two models evaluated were the operational HRRR (version 4) and the summer 2022 version of the experimental RRFS. The convective events were manually selected based on visual inspection of the MRMS composite reflectivity during periods when all three products had nearly complete availability. A total of 32 cases in total were selected, 8 from 4 different mature convective modes: MCSs, QLCSs, clusters, and cellular. These cases had similar storm object size distributions to the summer as a whole, indicating the general applicability of the

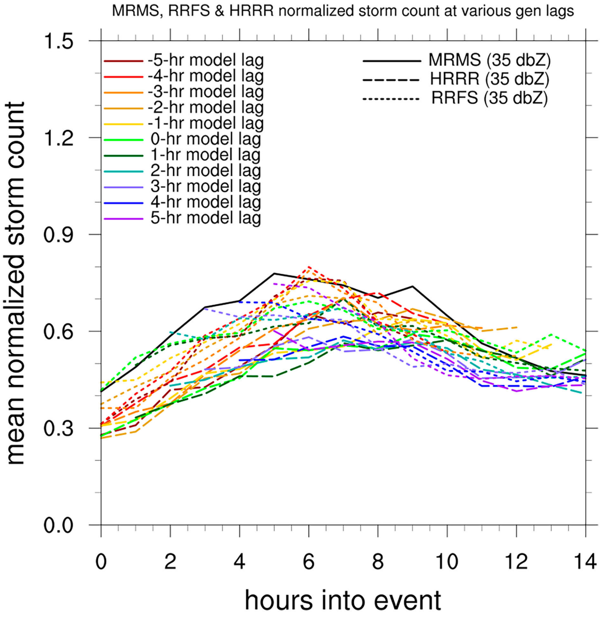

FIG. 19. As in Fig. 17, but for normalized total storm count.

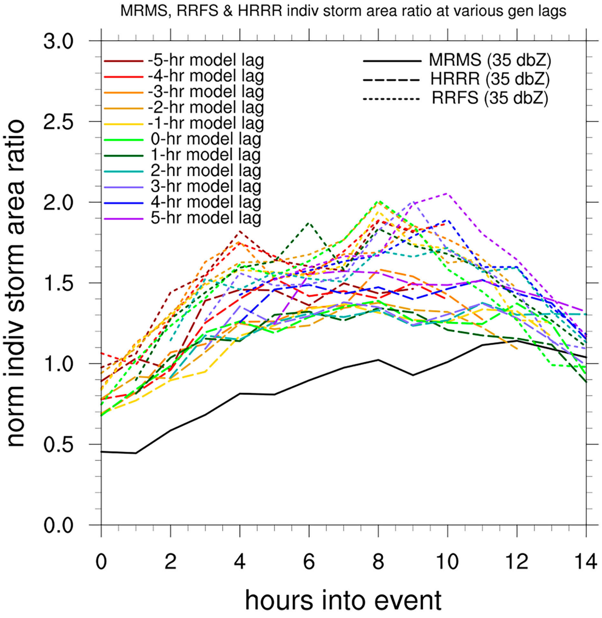

FIG. 20. As in Fig. 17, but for individual storm area ratio.

findings. However, since cases were conditionally sampled based on the observations, the study does not address the possibility of either model predicting a widespread convective event that did not occur.

Instead of trying to match individual storm objects obtained with a given model to those found in the observations as discussed by Davis et al. (2006), validation areas were manually drawn to ensure that the distribution of storm objects being compared were from the same widespread storm event in each dataset. This allowed for statistical comparisons of the macroscale properties of convective elements contained within the validation area without requiring individual storm elements within the convective storm area to overlap. The modeled evolution of storm macrophysical properties (total storm area, total storm count, and individual storm area ratio were evaluated relative to that obtained from MRMS observations using a compositing technique.

Several notable differences in the statistical performance of HRRR and RRFS were found. The distribution of intensities at the analysis times differed dramatically between the two forecast models despite both using the same method for assimilating radar reflectivity and for performing the cloud analyses. This difference at analysis time provides an indication of the different nature of the coupling between the DA system and the two dynamical cores. Using an object-identifying threshold of 35 d*BZ* revealed that convective storm areas predicted by RRFS were 60% larger than those predicted by HRRR (Table 2). It was also found that this bias in RRFS increased with decreasing storm organization, with the largest positive bias being for total storm area for widespread areas of predominantly cellular storms (Table 2). On the other hand, the HRRR bias was less a function of storm organization.

Both models had difficulty capturing the storm initiation as evidenced by underprediction of the storm counts during the storm initiation and growth phase (i.e., first 4–6 h of storm evolution) regardless of convective mode (Table 2). This underprediction of storm initiation and growth is consistent with that found by James et al. (2022) for the HRRR. However, the RRFS better captured CI than the HRRR in terms of timing with a reduced lag in the period of increasing storm counts (Table 2). This improvement in the timing of CI shows up most prominently in the clusters and cellular convective modes. At the same time, the HRRR better captured the amplitude of the evolution of storm area and storm size particularly for the clusters and cellular convective storm modes. Details of the comparison of relative performance of the RRFS and HRRR for several key metrics (including those discussed above) are given in Table 2.

Analyses using other reflectivity thresholds focused more on the models' ability to evaluate storms surrounded by stratiform rain regions (20 and 25 d*BZ*), or to focus primarily on stronger convective cores (40 and 40 d*BZ*). For the lower stratiform precipitation-encompassing thresholds, RRFS nearly matches the observed normalized total storm area in both timing and amplitude, but predicts too few storms. In contrast, even though storm counts are close to reality, RRFS has large total storm area biases at thresholds more indicative of the convective cores, with biases increasing as a function of increasing threshold. On the other hand, HRRR underpredicts the total area coverage and counts at lower d*BZ* thresholds, pointing to a tendency to underpredict the number and coverage of stratiform rainfall associated with convection; at higher convective core

TABLE 2. Relative performance of HRRR and RRFS for key characteristics of convection.

Storm characteristic	Figure	HRRR	RRFS
Bias in total storm area at peak maturity	Fig. 10a	+5%	+60%
Bias in total storm area at peak maturity—clusters	Fig. 14c	−10%	+60%
Bias in total storm area at peak maturity—cellular storms	Fig. 14d	+10%	+120%
Bias in storm counts at peak maturity	Fig. 10b	−20%	−10%
Bias in peak ISAR at 20 d*BZ*	Fig. 13a	−20%	−10%
Bias in peak ISAR at 45 d*BZ*	Fig. 13d	+40%	+140%
Bias in storm counts during CI/growth phase of storm evolution	Fig. 10b	−40%	−20%
Bias in storm counts during CI/growth phase of storm evolution–clusters	Fig. 15c	−60%	−20%
Bias in storm counts during CI/growth phase of storm evolution–cellular storms	Fig. 15d	−60%	−20%
Bias in timing of CI based on storm counts during first few hours of storm evolution	Fig. 10b	−3 h	−1.5 h

thresholds, HRRR overpredicts the area but at a much lower magnitude than RRFS, while the storm counts are nearer to observations.

It is noted that not all aspects of the relative performance of RRFS and HRRR have been evaluated in this study. For example, the 32 cases used in this study were conditionally selected based on observed widespread storm events. While the models generated too much storm area in a few cases, it is not possible to determine whether or not either model falsely predicted widespread events that were not present in the observations. Another limitation of this study is that most of the model evaluation focused on runs initialized 1 h before the beginning of storm initiation for a given event [exception to this is the lag analyses presented from the 13 cases in section 3c(4) that had sufficient model availability]. By evaluating only forecasts issued 1 h before CI, it is not possible to evaluate whether the models tended to initiate storms before they were observed. As such, most of the analyses can only uncover delays in the modeled timing of CI. Another limitation of this study is that only 8 of the cases were observed to initiate between 0000 and 1300 UTC; therefore, the statistics are dominated by storms that initiated during daytime.

While the differences in skill between HRRR and RRFS are expected to change as RRFS research and development efforts continue, there is recent evidence suggesting that the biases reported herein may not be completely alleviated prior to the RRFS becoming operational (Alexander et al. 2023). In particular, this is true for the oversized nature of the convective cores identified using thresholds of 35 dB*Z* or greater. As such, use of the operational RRFS predictions of convective weather should take into account both the improved characterization of the timing and evolution of convection of the smaller scale storms (clusters and cellular) while at the same time understanding that the size of these smaller scale storms (and overall area coverage) will tend to be overpredicted. Given that the area coverage is likely to be overdone for these storm types, users of future releases of the RRFS model might need to take this into account when issuing forecast products that include area coverage such as the Traffic Flow Management (TFM) Convective Forecast (TFC) which is produced by the Aviation Weather Center. Future work will be needed to determine whether the biases reported herein are manifested in the operational version that will be released in early 2025 (Alexander et al. 2023). Moreover, the impact of such biases on downstream decision support tools and applications like TFC and the NextGen Weather Processor (FAA 2022b) will need to be determined once Version 1 of the RRFS modeling system has been finalized.

Acknowledgments. This research is in response to requirements and funding by the Federal Aviation Administration (FAA). The views expressed are those of the authors and do not necessarily represent the official policy or position of the FAA. We greatly appreciate the reviews from Barry Bowers, Patrick Skinner, and an anonymous reviewer.

Data availability statement. MRMS composite reflectivity data were obtained from https://mrms.nssl.noaa.gov. HRRR data were obtained from operational runs performed at NCEP, while the experimental RRFS data were obtained from NOAA Global Systems Laboratory (NOAA/GSL).

REFERENCES

Alexander, C. R., and J. R. Carley, 2023: Progression towards the first operational implementation of the UFS-based Rapid Refresh forecast system as a convection allowing model application. *13th Conf. on Transition of Research to Operations*, Denver, CO, Amer. Meteor. Soc., 1B.4, https://ams.confex.com/ams/103ANNUAL/meetingapp.cgi/Paper/420286.

——, ——, and M. Pyle, 2023: The Rapid Refresh Forecast System: Looking beyond the first operational version. *Unifying Innovations in Forecasting Capabilities Workshop*, Boulder, CO, NOAA, 29 pp., https://epic.noaa.gov/wp-content/uploads/2023/08/UIFCW-2023-Tue-9.-Alexander_UFS_UIFCW_2023_Final-2.pdf.

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, https://doi.org/10.1175/MWR-D-15-0242.1.

——, and Coauthors, 2021: Stratiform cloud-hydrometeor assimilation for HRRR and RAP model short-range weather prediction. *Mon. Wea. Rev.*, **149**, 2673–2694, https://doi.org/10.1175/MWR-D-20-0319.1.

Blaylock, B. K., and J. D. Horel, 2020: Comparison of lightning forecasts from the High-Resolution Rapid Refresh Model to Geostationary Lightning Mapper observations. *Wea. Forecasting*, **35**, 401–416, https://doi.org/10.1175/WAF-D-19-0141.1.

Chen, X., N. Andronova, B. Van Leer, J. E. Penner, J. P. Boyd, C. Jablonowski, and S.-J. Lin, 2013: A control-volume model of the compressible Euler equations with a vertical Lagrangian coordinate. *Mon. Wea. Rev.*, **141**, 2526–2544, https://doi.org/10.1175/MWR-D-12-00129.1.

Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772–1784, https://doi.org/10.1175/MWR3145.1.

Dixon, M., and G. Wiener, 1993: TITAN: Thunderstorm Identification, Tracking, Analysis, and Nowcasting—A radar-based methodology. *J. Atmos. Oceanic Technol.*, **10**, 785–797, https://doi.org/10.1175/1520-0426(1993)010<0785:TTITAA>2.0.CO;2.

Done, J., C. A. Davis, and M. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the Weather Research and Forecast (WRF) Model. *Atmos. Sci. Lett.*, **5**, 110–117, https://doi.org/10.1002/asl.72.

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

Duda, J. D., and D. D. Turner, 2021: Large-sample application of radar reflectivity object-based verification to evaluate HRRR warm-season forecasts. *Wea. Forecasting*, **36**, 805–821, https://doi.org/10.1175/WAF-D-20-0203.1.

FAA, 2022a: FAQ: Weather delay. Accessed 27 June 2023, https://www.faa.gov/nextgen/programs/weather/faq.

——, 2022b: NextGen Weather Processor (NWP). Accessed 27 June 2023, https://www.faa.gov/nextgen/programs/weather/nwp.

Gallus, W. A., Jr., N. A. Snook, and E. V. Johnson, 2008: Spring and summer severe weather reports over the Midwest as a function of convective mode: A preliminary study. *Wea. Forecasting*, **23**, 101–113, https://doi.org/10.1175/2007WAF2006120.1.

Grim, J. A., J. O. Pinto, T. Blitz, K. Stone, and D. C. Dowell, 2022: Biases in the prediction of convective storm characteristics with a convection allowing ensemble. *Wea. Forecasting*, **37**, 65–83, https://doi.org/10.1175/WAF-D-21-0106.1.

Homeyer, C. R., E. M. Murillo, and M. R. Kumjian, 2023: Relationships between 10 years of radar-observed supercell characteristics and hail potential. *Mon. Wea. Rev.*, **151**, 2609–2632, https://doi.org/10.1175/MWR-D-23-0019.1.

Husain, S. Z., C. Girard, A. Qaddouri, and A. Plante, 2019: A new dynamical core of the Global Environmental Multiscale (GEM) model with a height-based terrain-following vertical coordinate. *Mon. Wea. Rev.*, **147**, 2555–2578, https://doi.org/10.1175/MWR-D-18-0438.1.

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

Miller, P. W., and T. L. Mote, 2017: A climatology of weakly forced and pulse thunderstorms in the Southeast United States. *J. Appl. Meteor. Climatol.*, **56**, 3017–3033, https://doi.org/10.1175/JAMC-D-17-0005.1.

Pinto, J. O., J. A. Grim, and M. Steiner, 2015: Assessment of the High-Resolution Rapid Refresh Model's ability to predict mesoscale convective systems using object-based evaluation. *Wea. Forecasting*, **30**, 892–913, https://doi.org/10.1175/WAF-D-14-00118.1.

Potvin, C. K., and Coauthors, 2019: Systematic comparison of convection-allowing models during the 2017 NOAA HWT Spring Forecasting Experiment. *Wea. Forecasting*, **34**, 1395–1416, https://doi.org/10.1175/WAF-D-19-0056.1.

Rasmussen, E. N., J. M. Straka, R. Davies-Jones, C. A. Doswell III, F. H. Carr, M. D. Eilts, and D. R. MacGorman, 1994: Verifications of the origins of rotation in tornadoes experiment: VORTEX. *Bull. Amer. Meteor. Soc.*, **75**, 995–1006, https://doi.org/10.1175/1520-0477(1994)075<0995:VOTOOR>2.0.CO;2.

Roberts, R. D., and S. Rutledge, 2003: Nowcasting storm initiation and growth using *GOES-8* and WSR-88D data. *Wea. Forecasting*, **18**, 562–584, https://doi.org/10.1175/1520-0434(2003)018<0562:NSIAGU>2.0.CO;2.

Roff, G., and Coauthors, 2022: APS2-ACCESS-C2: The first Australian operational NWP convection-permitting model. *J. South. Hemisphere Earth Syst. Sci.*, **72**, 1–18, https://doi.org/10.1071/ES21013.

Schwartz, C. S., and Coauthors, 2009: Next-day convection-allowing WRF Model guidance: A second look at 2-km versus 4-km grid spacing. *Mon. Wea. Rev.*, **137**, 3351–3372, https://doi.org/10.1175/2009MWR2924.1.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., https://doi.org/10.5065/D68S4MVH.

Skinner, P. S., and Coauthors, 2018: Object-based verification of a prototype Warn-on-Forecast System. *Wea. Forecasting*, **33**, 1225–1250, https://doi.org/10.1175/WAF-D-18-0020.1.

Smith, B. T., R. L. Thompson, J. S. Grams, C. Broyles, and H. E. Brooks, 2012: Convective modes for significant severe thunderstorms in the contiguous United States. Part I: Storm classification and climatology. *Wea. Forecasting*, **27**, 1114–1135, https://doi.org/10.1175/WAF-D-11-00115.1.

Smith, T. M., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) severe weather and aviation products: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 1617–1630, https://doi.org/10.1175/BAMS-D-14-00173.1.

Termonia, P., and Coauthors, 2018: The ALADIN system and its canonical model configurations AROME CY41T1 and ALARO CY40T1. *Geosci. Model Dev.*, **11**, 257–281, https://doi.org/10.5194/gmd-11-257-2018.

Walters, D., and Coauthors, 2019: The Met Office Unified Model global atmosphere 7.0/7.1 and JULES global land 7.0 configurations. *Geosci. Model Dev.*, **12**, 1909–1963, https://doi.org/10.5194/gmd-12-1909-2019.

Wang, Y., and X. Wang, 2017: Direct assimilation of radar reflectivity without tangent linear and adjoint of the nonlinear observation operator in the GSI-based EnVar system: Methodology and experiment with the 8 May 2003 Oklahoma City tornadic supercell. *Mon. Wea. Rev.*, **145**, 1447–1471, https://doi.org/10.1175/MWR-D-16-0231.1.

Weisman, M. L., W. C. Skamarock, and J. B. Klemp, 1997: The resolution dependence of explicitly modeled convective systems. *Mon. Wea. Rev.*, **125**, 527–548, https://doi.org/10.1175/1520-0493(1997)125<0527:TRDOEM>2.0.CO;2.

——, C. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with 0–36-h explicit convective forecasts with the WRF-ARW Model. *Wea. Forecasting*, **23**, 407–437, https://doi.org/10.1175/2007WAF2007005.1.

Weygandt, S. S., S. G. Benjamin, M. Hu, C. R. Alexander, T. G. Smirnova, and E. P. James, 2022: Radar reflectivity-based model initialization using specified latent heating (Radar-LHI) within a diabatic digital filter or pre-forecast integration. *Wea. Forecasting*, **37**, 1419–1434, https://doi.org/10.1175/WAF-D-21-0142.1.

Wicker, L., 2023: Assessment of convective-scale attributes of the FV3 Dycore using idealized simulations. *Unifying Innovations in Forecasting Capabilities Workshop*, Boulder, CO, NOAA, 18 pp., https://epic.noaa.gov/wp-content/uploads/2023/08/UIFCW-2023-Tue-4.-Wicker_FV3_UFS_Talk-1.pdf.