Using Object-Based Verification to Assess Improvements in Forecasts of Convective Storms between Operational HRRR Versions 3 and 4

JEFFREY D. DUDA[a,b] AND DAVID D. TURNER[b]

[a] *Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado*
[b] *National Oceanic and Atmospheric Administration/Global Systems Laboratory, Boulder, Colorado*

(Manuscript received 28 October 2022, in final form 27 July 2023, accepted 28 July 2023)

ABSTRACT: The object-based verification procedure described in a recent paper by Duda and Turner was expanded herein to compare forecasts of composite reflectivity and 6-h precipitation objects between the two most recent operational versions of the High-Resolution Rapid Refresh (HRRR) model, versions 3 and 4, over an expanded set of warm season cases in 2019 and 2020. In addition to analyzing all objects, a reduced set of forecast–observation object pairs was constructed by taking the best forecast match to a given observation object for the purposes of bias-reduction and unequivocal object comparison. Despite the apparent signal of improved scalar metrics such as the object-based threat score in HRRRv4 compared to HRRRv3, no statistically significant differences were found between the models. Nonetheless, many object attribute comparisons revealed indications of improved forecast performance in HRRRv4 compared to HRRRv3. For example, HRRRv4 had a reduced overforecasting bias for medium- and large-sized reflectivity objects, and all objects during the afternoon. HRRRv4 also better replicated the distribution of object complexity and aspect ratio. Results for 6-h precipitation also suggested superior performance in HRRRv4 over HRRRv3. However, HRRRv4 was worse with centroid displacement errors and more severely overforecast objects with a high maximum precipitation amount. Overall, this exercise revealed multiple forecast deficiencies in the HRRR, which enables developers to direct development efforts on detailed and specific endeavors to improve model forecasts.

SIGNIFICANCE STATEMENT: This work builds upon the authors' prior work in assessing model forecast quality using an alternative verification method—object-based verification. In this paper we verified two versions of the same model (one an upgrade from the other) that were making forecasts covering the same time window, using the object-based verification method. We found that the updated model was not statistically significantly better, although there were indications it performed better in certain aspects such as capturing the change in the number of storms during the daytime. We were able to identify specific problem areas in the models, which helps us direct model developers in their efforts to further improve the model.

KEYWORDS: Precipitation; Thunderstorms; Forecast verification/skill; Model comparison;
Model evaluation/performance; Numerical weather prediction/forecasting

1. Introduction

An object-based approach to forecast verification for numerical weather prediction (NWP) models is an alternative to the legacy approach in which forecast fields are verified considering a deterministic or probabilistic event at each model grid point. Object-based verification is particularly useful for feature-based fields in convection-allowing models (CAM), fields such as radar reflectivity and updraft helicity, the horizontal structure of which is typically dominated by an "empty" floor (e.g., no tangible field value, often represented using a value of 0.0) and containing discrete contiguous collections of points at which the field is nonempty. The legacy grid-to-grid verification approach, which is used heavily in the NOAA verification systems to evaluate models during their development (e.g., Turner et al. 2020), evaluates events at each grid point and, due to the typically low base rates of the events of interest, are dominated by null events.

In contrast, the object-based approach condenses the entire field into a set of objects that each encompass many grid points. A set of attributes such as location, size, shape, and structure are calculated to describe each object. In modern CAMs, there are on the order of millions or tens of millions of grid points in the horizontal. In the object-based approach, the information in such a horizontal grid space is typically reduced to tens to thousands of objects instead. Therefore, the object-based approach involves removal of data which can be a disadvantage compared to the grid-to-grid verification such that a perfect forecast could be calculated in object space when there is not a perfect correspondence between the forecast and truth field at each grid point. However, given the known difficulties with grid-to-grid verification of feature-based fields in CAMs (e.g., the "double penalty" problem[1]), the object-based approach can be useful in identifying behaviors in NWP models that grid-to-grid verification is unable to identify. For example, an object-based technique can distinguish between these two behaviors: forecast objects being larger than observation objects; the forecast containing more objects than the observations. The gridpoint-based frequency bias, on the other hand, could have the same value in

Corresponding author: Jeffrey D. Duda, Jeffrey.duda@noaa.gov

[1] A particularly helpful example of this problem is illustrated qualitatively and quantitatively as test "geom001" in Ahijevych et al. (2009).

DOI: 10.1175/WAF-D-22-0181.1

both scenarios. While this distinction may not be important when considering a simple scalar metric such as the root-mean-square error, the distinction could be meaningful to forecast developers and users since it could help isolate the underlying physical or dynamical mechanisms responsible for model forecast deficiencies.

Classifying objects in feature-specific gridded fields can offer other insights even if comparison to observation is not a goal. Guerra et al. (2022), for example, applied an object-based technique to forecasts from the National Severe Storms Laboratory's Warn-on Forecast System (WoFS) to compare convective storm existence by the age of the storm. Flora et al. (2019) performed an object-based verification of rotating thunderstorms in WoFS, even incorporating probabilistic fields to generate objects and comparing across fields. Britt et al. (2020) applied the object-based system setup in Skinner et al. (2018) for evaluating WoFS reflectivity and updraft helicity forecasts to extract inflow environment information near supercell thunderstorms.

Duda and Turner (2021, hereafter DT21) provided a demonstration of some of the insights that can be gained from object-based verification of a CAM. They presented a slate of verification analyses from forecasts of composite reflectivity from the operational High-Resolution Rapid Refresh model version 3 (HRRRv3; Benjamin et al. 2021, 2022; Dowell et al. 2022; James et al. 2022; Weygandt et al. 2022) using warm-season cases from 2019. They discovered that HRRRv3 produced too many storm objects, most of which were small. Small-sized objects were also too smooth compared to observations whereas larger objects were not, illustrating the poor resolution of individual convective storm cells in a 3-km model. Also, a strange artifact presented as a spike in object shape that led to an investigation of a storm artifact presented herein. Object-based verification has continued to grow since DT21 (for a discussion of prior research, see the references in DT21). Recently, Chen et al. (2022) published a study on object-based verification of supercell storms in a short-term convection-allowing ensemble framework similar to the modeling systems featured herein, although their emphasis was on performance differences between different radar data assimilation methods for reflectivity and updraft helicity tracks. Gallo et al. (2021) applied an object-based approach using surrogate severe probability forecasts from HRRRv3 as well as primitive versions of the upcoming Rapid-Refresh Forecast System, a Finite-Volume Cubed-based model, from the NOAA Hazardous Weather Testbed Spring Forecasting Experiment, to assess how well these models can predict the likelihood and location of severe weather. Finally, Grim et al. (2022) investigated biases in storm counts and sizes for various storm morphologies in HRRRv4 and the HRRR ensemble (HRRRE; Kalina et al. 2021).

Results from the above studies will be useful for comparison in this follow-up paper to DT21, upon which we expand in the following ways. Version 4 of the HRRR became operational in December 2020 but was running in a preoperational phase (frozen code) in the second half of 2019 and first 11 months of 2020 over a range of cases that overlapped with those covered by HRRRv3. Therefore, we compare HRRRv3 and HRRRv4 forecasts over a large sample containing parts of the warm seasons of 2019 and 2020. One goal is to identify improvements in errant aspects of HRRRv3 forecasts in HRRRv4, namely, overproduction of small objects, but also in more general character such as storm displacement and object shape. We also add the 6-h accumulated precipitation field into the verification suite to expand the investigation to more fields. Finally, we examine an expanded array of metrics and analyses to look deeper into the behaviors of the two models.

2. Methods

a. HRRR model description and updates from version 3 to 4

The HRRR is a Weather Research and Forecasting (WRF) Model (Skamarock et al. 2019) dynamical core-based convection-allowing NWP model that provides an 18- or 36-h forecast each hour, depending on time of day. All operational HRRR versions use a 1-h preforecast cycle to spin up fine-scale structures from a background, which, until HRRRv4, was provided by a 1-h Rapid Refresh forecast (Benjamin et al. 2016). Among the more significant updates from HRRRv3 to HRRRv4 was the swap to using a parallel 3-km ensemble (named the HRRR data assimilation system, HRRRDAS) to provide flow-dependent background error covariances in the data assimilation analysis, as well as using the HRRRDAS mean for the initial conditions to the 1-h preforecast. A second major update relevant to the investigation herein was the implementation of the implicit-explicit vertical advection (IEVA) scheme described in Wicker and Skamarock (2020). Previous HRRR versions implemented an ad hoc upper bound to the temperature tendency in the model microphysics to keep the model numerically stable during integration, which allowed for a longer model time step. This temperature tendency limit restricted vertical motion magnitudes, especially in explicit convective updrafts. The IEVA enabled the removal of this tendency limit; updraft speeds in WRF modeled convective storms therefore dramatically increased. Many more updates to HRRRv4 from HRRRv3 are described in Dowell et al. (2022). While any of these updates may manifest as differences in the results herein, it would be very difficult, if not impossible, to meaningfully isolate each specific update responsible for any differences between the metrics presented.

b. Case and field selection

The setup for this verification experiment is nearly identical to that in DT21. Whereas DT21 obtained cases from 1 April to 30 September 2019, herein we sample cases from 15 August to 30 September 2019 and from 1 April to 30 September 2020, a time when both models were running. This range includes as many as ~1360 forecast initializations[2] of HRRR forecasts initialized every 3 h (i.e., 0000, 0300, 0600, 0900, 1200, 1500, and 1800 UTC each day). The choice to sample forecasts every 3 h

[2] The actual number of forecasts verified (hereafter "cases") at a given forecast hour varies based on the availability of archived forecast and observation files. Also, the forecast length of HRRR varies according to initialization time. Since only half of the considered HRRR forecasts ran beyond 18 h, the number of cases at lengths longer than 18 h is about half of what it is for forecast lengths of 18 h and less.

is aimed at reducing autocorrelation/dependence of samples while also covering a wide range of times of year for improved sampling of weather regimes, and toward statistical robustness. Composite reflectivity forecasts are verified hourly through forecast hour 24 as well as at forecast hours 30 and 36 to examine object-based statistics at longer lead times. The meteorological events of interest include convective storms and associated heavy precipitation that could produce flooding. Composite reflectivity forecasts are verified against Multi-Radar Multi-Sensor (MRMS; Smith et al. 2016) observations. The 6-h accumulated precipitation forecasts over the nonoverlapping windows of 0–6, 6–12, 12–18, 18–24, 24–30, and 30–36 forecast hours are verified using Stage-IV (ST4) precipitation analyses (Du 2011; Hou et al. 2014). The verification domain includes areas within the HRRR model domain that are within the contiguous United States east of the Rocky Mountains and immediately offshore of coasts and across international borders. (See Fig. 3 for an outline of the verification domain.) Observation fields are interpolated to the HRRR grid before verifying.

c. *Object-based classification*

The Method of Object-based Diagnostic Evaluation (MODE; Davis et al. 2006) is used to identify objects and calculate single-object and forecast–observation (F-O) object pair attributes for comparison. We used MODE from version 9.0 of the Model Evaluation Tools software repository. The input fields to MODE (HRRR and MRMS composite reflectivity, and ST4 6-h precipitation) are filtered using a circular convolution smoother to remove noisy or unwanted small-scale features using a convolution radius that is controlled in a configuration file. Our interest in convective storms in this experiment results in selecting a convolution radius of one grid square to minimize the filtering of near-grid-scale structure in HRRR composite reflectivity as well as in 6-h precipitation, which already implicitly contains some filtering due to the temporal aggregation of instantaneous precipitation rate. Magnitude thresholds are then applied to the convolved fields to classify objects as contiguous sets of grid points at which the convolved field exceeds the threshold. For composite reflectivity, thresholds of 25, 30, 35, and 40 dB*Z* are used. For 6-h precipitation, thresholds of 0.254 (0.01), 2.54 (0.10), 6.35 (0.25), 12.7 (0.50), and 25.4 (1.00) mm (in.) are used to focus on a range of events including convective storms and heavy rain, both of which can lead to property damage, injury, and disruption of normal economic activity. Objects with an area of less than 16 grid squares (equivalent to $4\Delta x$, a simple threshold to exclude unresolved objects) were discarded. While an alternative approach involves selecting a single threshold to define all object sets, we choose to examine the sensitivity of calculated metrics to a range of thresholds.

After objects are identified, the original field values are restored to the grid points within the object boundaries, and object attributes are computed from the resulting collections of grid points. MODE computes several object attributes using graphical techniques (see Table 1 of DT21). One technique computes a convex hull, the smallest convex perimeter that completely encompasses the object. It is like wrapping a rubber band around the object. Using that information, object complexity is calculated as 1.0 − the ratio of the area of the object to the area of the convex hull around the object (the subtraction from 1.0 is to convert the ratio to a positively oriented measure). Another technique fits an inscribing rectangle around the object such that it is just long and wide enough to contain the object. The width and length of this rectangle are used to compute the aspect ratio of the object. The mass of an object is the sum of the gridpoint values within the object. The structure of an object can also be measured by the quantile values of the field magnitudes within an object.

MODE attempts to pair objects by calculating the interest value between all eligible F-O object pairs. Eligibility is determined using a threshold centroid distance that is user controllable. Any pair of objects separated by a distance larger than this threshold value is not evaluated by MODE (i.e., the interest value is arbitrarily set to 0.0). The threshold centroid distance for comparison is set to 500 km for composite reflectivity and 1000 km for 6-h precipitation. These thresholds represent a balance between cost saving when running MODE and a distance at which forecast objects are sufficiently close to the observations to be considered potentially useful. In a sense, this threshold represents a first-order filter to disregard forecast objects that are too far away from an observed object to be useful.

The interest value output from MODE is a weighted sum of individual attribute interest values, each calculated from user-adjustable interest maps, which are functions that convert the actual attribute difference to a normalized quantity. These interest maps allow for control over the strictness of the comparison of an object attribute between the objects in an F-O pair. Attributes that are compared include three distance measurements (centroid, object boundary, convex hull boundary), the differences in orientation angle and aspect ratio, and the ratios of the area, curvature, complexity, and a user-controllable percentile value of the two objects, as well as the consumption ratio, defined as the fraction of the smaller of the two objects that intersects with the larger of the two objects. The consumption ratio of a relatively smaller object that is fully contained within its corresponding larger object is 1.0. If there is no overlap between two objects, the consumption ratio for that pair is 0.0. The 95th percentile of composite reflectivity (p95) and the 99th percentile of 6-h precipitation (p99) are used for intensity percentile comparisons, as they both represent realistic maximum values. Weights for these attribute comparisons are user selectable. We choose to emphasize attributes describing position and size more than others for composite reflectivity, and to object shape more than object location for 6-h precipitation (Table 1). Sensitivity of interest value to the set of weights has not been rigorously tested and is left for future work.

d. *Metrics and analyses*

We evaluate several types of metrics based on MODE output. Many were detailed in DT21, including the object-based threat score (OTS), various measures of the central tendency of the distance between F-O object centroids, continuous ranked probability scores (CRPS) for object attributes, as well as a bevy of custom evaluation techniques that sample across different dimensions of object attribute distributions,

TABLE 1. Interest weight settings in MODE.

Attribute name	Composite reflectivity	6-h precipitation
Centroid distance	5.0	2.0
Boundary distance	4.0	1.0
Convex hull distance	0.0	0.0
Orientation angle difference	0.0	1.0
Aspect ratio difference	0.0	1.0
Area ratio	4.0	5.0
Consumption ratio	2.0	2.5
Curvature ratio	0.0	0.0
Complexity ratio	0.5	1.0
Intensity percentile ratio (percentile)	3.5 (95)	3.0 (99)

whether one-dimensional or joint distributions between forecasts and observations.

OTS measures the intersection area between objects in a pair and weights this quantity by the pair interest. It is formulated as

$$\mathrm{OTS} = \frac{1}{A_f + A_o}\sum_p I^p(a_f^p + a_o^p), \qquad (1)$$

where a refers to object area; A is the total area of all objects; I is the object pair interest; the subscripts f and o correspond to forecast and observation objects, respectively; and superscript p refers to an object pair.

With the addition of a second forecast dataset, we can compute the skill version of the CRPS, or the continuous ranked probability skill score (CRPSS). CRPSS is defined herein as

$$\mathrm{CRPSS} = 1 - \frac{\mathrm{CRPS}_{\mathrm{HRRRv4}}}{\mathrm{CRPS}_{\mathrm{HRRRv3}}}, \qquad (2)$$

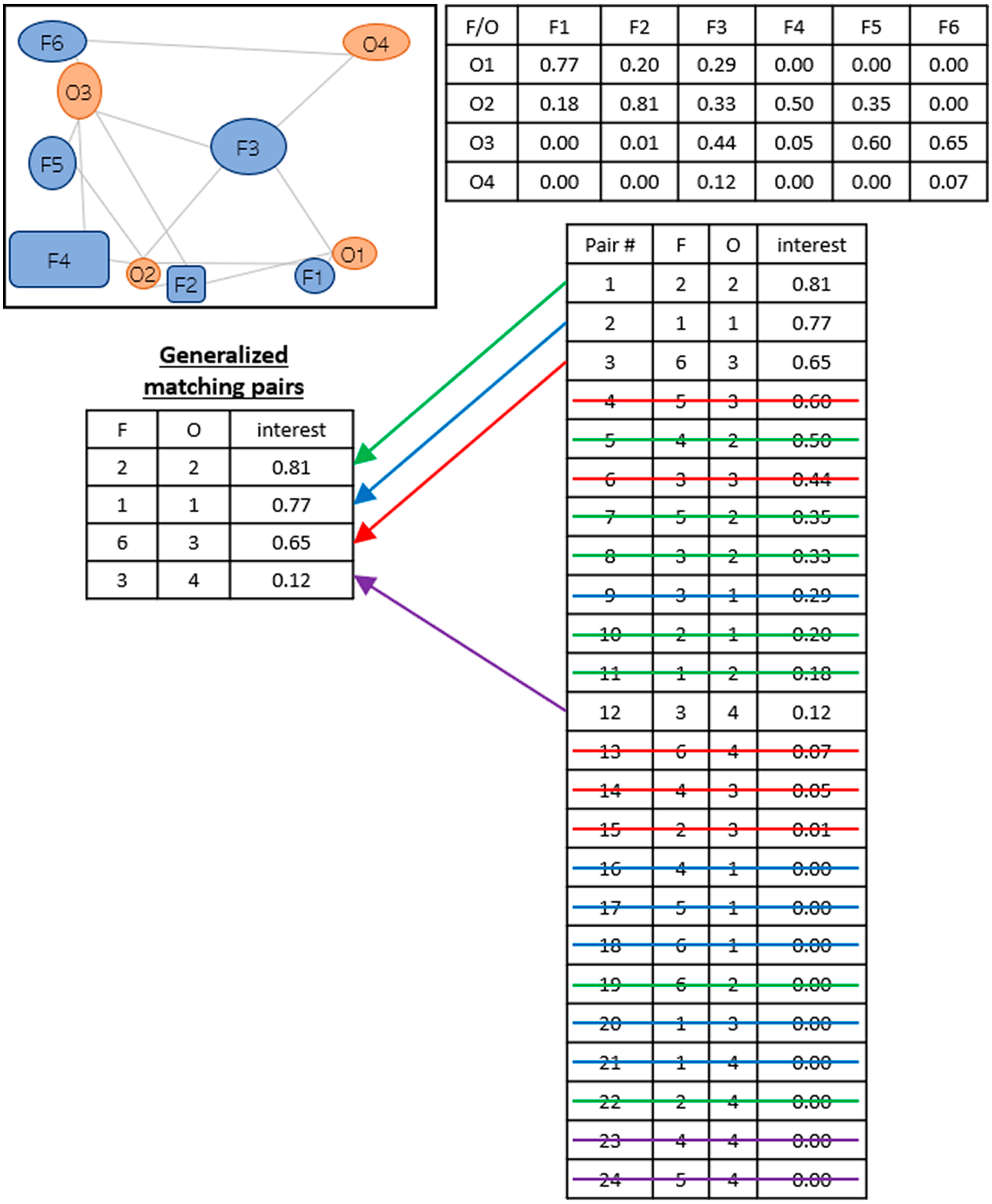
where HRRRv4 is compared to the reference HRRRv3 forecast system; CRPSS $>$ 0.0 indicates that HRRRv4 outperforms HRRRv3.

Several metrics, including OTS, require an injective matching between forecast and observation objects such that a given object in either dataset (forecast or observation) can be matched to at most one object in the other dataset. MODE performs an unassuming matching of each eligible F-O object pair by considering whether the interest value for the pair exceeds a threshold (default value is 0.70). By this construction, an object in one dataset could be matched to more than one object in the other dataset. This is problematic for computing OTS and for constructing other datasets for analysis. Therefore, as is used in the computation of OTS, we use what is hereafter referred to as generalized matching or “reduced set” to perform additional verification diagnostics (Fig. 1). In the generalized matching procedure, the set of all F-O pairs at a single forecast lead time are ranked by interest from highest to lowest. The first pair in the F-O pair list provides the first entry of the generalized matching vector of interest values; every subsequent F-O pair containing either object is precluded from contributing to the generalized matching vector. This process continues until no object pairs remain, resulting in a reduced set of F-O pairs (a subset of the full set) that are unique to each other such that they are deemed to be the closest match to each other regardless of whether their interest value exceeds the matching threshold. If the number of forecast and observation objects in each set are the same and are sufficiently close to one another, this procedure creates a bijective mapping between the two object sets. Otherwise, either there is a multiplicative object count forecast bias or some objects are poorly forecast so that some objects are not used in the subsequent evaluation. Since OTS is based on this procedure, it means that forecast object count bias will reduce the OTS value, consistent with an error-prone forecast. For other evaluations, the forecast object count bias should be considered in tandem to get a fuller picture of the forecast performance. In a sense, this process amounts to a form of object-based bias correction, and therefore can be used to determine how good the model might be if the correct number of objects were forecast.

The differences in the values of many scalar metrics between HRRRv3 and HRRRv4 are tested for statistical significance using a circular block bootstrapping technique (Gilleland 2020) with a block size of eight forecasts (roughly filling a 24-h period, even though the decorrelation time of composite reflectivity and 6-h precipitation forecasts is arguably substantially shorter). Through this procedure it was found that none of the differences between HRRRv3 and HRRRv4 for any of the metrics, thresholds, and forecast lengths tested were statistically significant, and therefore all further discussion of comparative differences between HRRRv3 and HRRRv4 are considered in the context of statistical nonsignificance.

3. Composite reflectivity results

The total number of composite reflectivity objects that were identified varied with forecast length and threshold. At the 25-dBZ threshold, the number of composite reflectivity objects ranged from about 80 000 to over 100 000 over about 1350 cases in each of the first 18 forecast hours. After 18 h the numbers approximately halved; about 50 000 objects were classified over about 650 cases each hour. For diagnostics that considered all forecast hours aggregated together, there were over 2 million objects. Around 70 000, 50 000, and 35 000 objects were classified each forecast hour (within the first 18 h) at the 30-, 35-, and 40-dBZ thresholds, respectively.



F/O	F1	F2	F3	F4	F5	F6
O1	0.77	0.20	0.29	0.00	0.00	0.00
O2	0.18	0.81	0.33	0.50	0.35	0.00
O3	0.00	0.01	0.44	0.05	0.60	0.65
O4	0.00	0.00	0.12	0.00	0.00	0.07

Generalized matching pairs

F	O	interest
2	2	0.81
1	1	0.77
6	3	0.65
3	4	0.12

Pair #	F	O	interest
1	2	2	0.81
2	1	1	0.77
3	6	3	0.65
~~4~~	~~5~~	~~3~~	~~0.60~~
~~5~~	~~4~~	~~2~~	~~0.50~~
~~6~~	~~3~~	~~3~~	~~0.44~~
~~7~~	~~5~~	~~2~~	~~0.35~~
~~8~~	~~3~~	~~2~~	~~0.33~~
~~9~~	~~3~~	~~1~~	~~0.29~~
~~10~~	~~2~~	~~1~~	~~0.20~~
~~11~~	~~1~~	~~2~~	~~0.18~~
12	3	4	0.12
~~13~~	~~6~~	~~4~~	~~0.07~~
~~14~~	~~4~~	~~3~~	~~0.05~~
~~15~~	~~2~~	~~3~~	~~0.01~~
~~16~~	~~4~~	~~1~~	~~0.00~~
~~17~~	~~5~~	~~1~~	~~0.00~~
~~18~~	~~6~~	~~1~~	~~0.00~~
~~19~~	~~6~~	~~2~~	~~0.00~~
~~20~~	~~1~~	~~3~~	~~0.00~~
~~21~~	~~1~~	~~4~~	~~0.00~~
~~22~~	~~2~~	~~4~~	~~0.00~~
~~23~~	~~4~~	~~4~~	~~0.00~~
~~24~~	~~5~~	~~4~~	~~0.00~~

FIG. 1. Ad hoc example illustrating the generalized matching procedure for a forecast (blue) with six objects and an observation (orange) dataset with four objects. The 2D interest table (at top right) is sorted by interest (at bottom right), and the vector of generalized matched pairs are built from that list by eliminating pairs that share a common object with one already selected with higher interest (eliminations due to a given pair are color coded).

a. *Object frequency bias*

DT21 found that HRRRv3 produced too many composite reflectivity objects throughout the forecast, with the overforecasting bias increasing with forecast length. James et al. (2022) also found an increasing bias with forecast length, but with a different sign during the early hours. This characteristic was evident in HRRRv4 forecasts as well (Fig. 2a), but the pattern differed somewhat from that in HRRRv3. The object-based frequency bias was slightly improved in HRRRv4 during the early forecast hours (especially at the 40-dBZ threshold) but was degraded after about forecast hour 8, similar to that seen in James et al. (2022). An exception to this behavior was at the 40-dBZ threshold, where the frequency biases of the two models were the same from forecast hours 9–17, and after which HRRRv4 retained a better frequency bias through 36 forecast hours. The

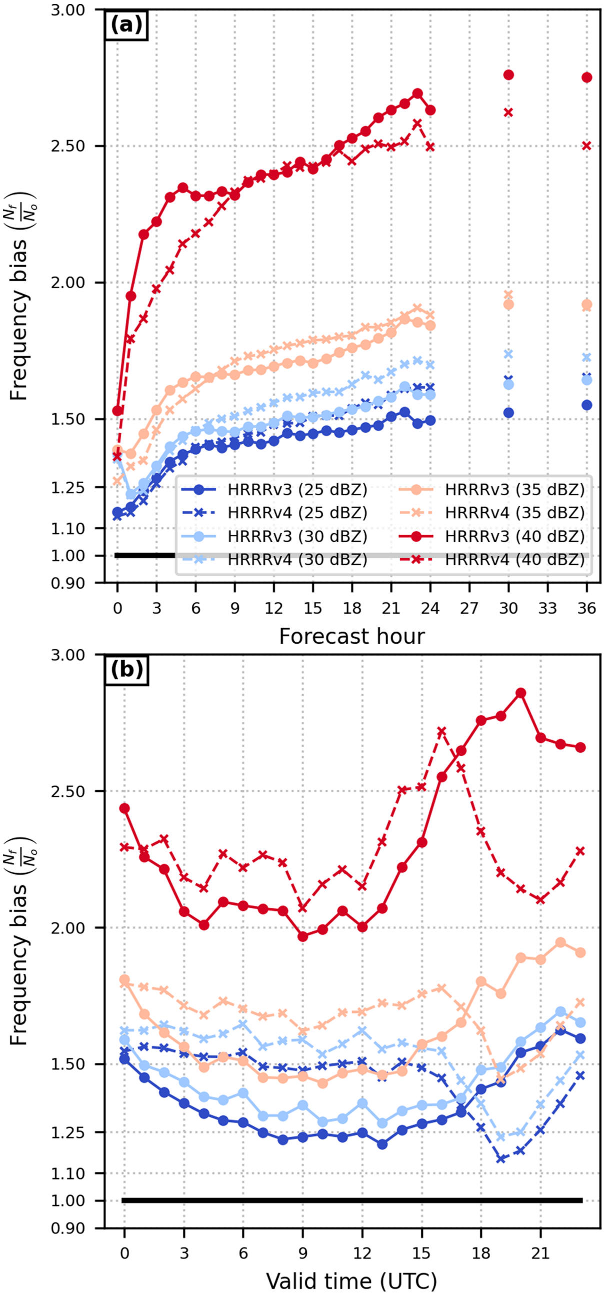


FIG. 2. Object-based frequency bias aggregated across all forecast cases as a function of (a) forecast lead time and (b) valid time of day.

change in initial condition design or model physics could be responsible for this difference in behavior, but there are too many factors to be certain. Frequency bias by valid time of day showed a different pattern (Fig. 2b). While HRRRv4 had a slightly higher bias during most of the day, the bias decreased substantially between 1500 and 0000 UTC (late morning to late afternoon), becoming lower than that of HRRRv3. This behavior was present regardless of the age of the forecast, although it was more prominent in older forecasts (not shown). Therefore, despite the apparent degradation in overall frequency bias (seen in Fig. 2a for forecast hours beyond 8 and at thresholds of 35 dBZ and lower), it was unevenly distributed in such a way that forecasts of diurnally forced convection appeared to be improved in HRRRv4.

There are many other aspects of the overforecasting issue that can be teased out of MODE output and thus provide forecast developers with better indications of which model aspects need improvement. For example, much of the increase in the storm count bias in HRRRv4 over HRRRv3 was found over the Midwest and Great Lakes regions (Fig. 3a) where the ratio of the number of HRRRv4 to HRRRv3 object centroids was greater than 1.0. On the other hand, the ratio was inverted over the High Plains, much of Texas, and the Southeast United States. This discrepancy is generally insensitive to forecast lead time but contained a diurnal signal (Figs. 3b,c). HRRRv4 produced more storms than HRRRv3 across all but the northern High Plains from the late evening through overnight and into the next morning, with the difference being domain-wide during the middle of the overnight. It was mainly during the afternoon when the reduced storm count was evident in HRRRv4 in the same regions where the diurnally aggregated ratio was also less than 1.0. The spatial pattern of object frequency bias for HRRRv4 was the same as for HRRRv3 (cf. Fig. 7 of DT21), so problems remain with overforecasting of storms across most of the eastern United States, but storms were forecast with approximately the correct frequency on the High Plains in both models. The representation of land surface interactions and boundary layer flows is suspected to play a role in these behaviors, but a deeper investigation is beyond the scope of this paper.

b. *Total forecast metrics*

OTS suggests composite reflectivity forecasts from HRRRv4 were better than those from HRRRv3 regardless of reflectivity threshold or forecast length (Fig. 4a). At all but the 40-dBZ threshold, the OTS difference between the two model versions was approximately constant with forecast length. At 40 dBZ, however, the two forecasts were essentially identical in performance during the 7–15-h forecast window. Also, consistent with the lower frequency biases during the afternoon, the OTS for HRRRv4 was higher than that for HRRRv3 during the afternoon as well (Fig. 4b). OTS was overall highest during the overnight when convective activity is low.

Minor evidence was found to suggest that HRRRv4 forecast storms closer to observed storms than HRRRv3, as quantified by the mean distance between the centroids of objects using generalized matching (Fig. 5). The mean distance for HRRRv4 was only a few multiples of Δx lower than that for HRRRv3 at all but the 40-dBZ threshold during the forecast, with the largest improvement occurring during the 1–8-h forecast window (Fig. 5a). At the 40-dBZ threshold, the mean distance for the two models oscillated around each other so that neither model forecast storms closer to the observations consistently. Interestingly, though, the mean distance metric showed a dependency on time of day. HRRRv4 had a lower mean centroid distance than HRRRv3 at all times of day

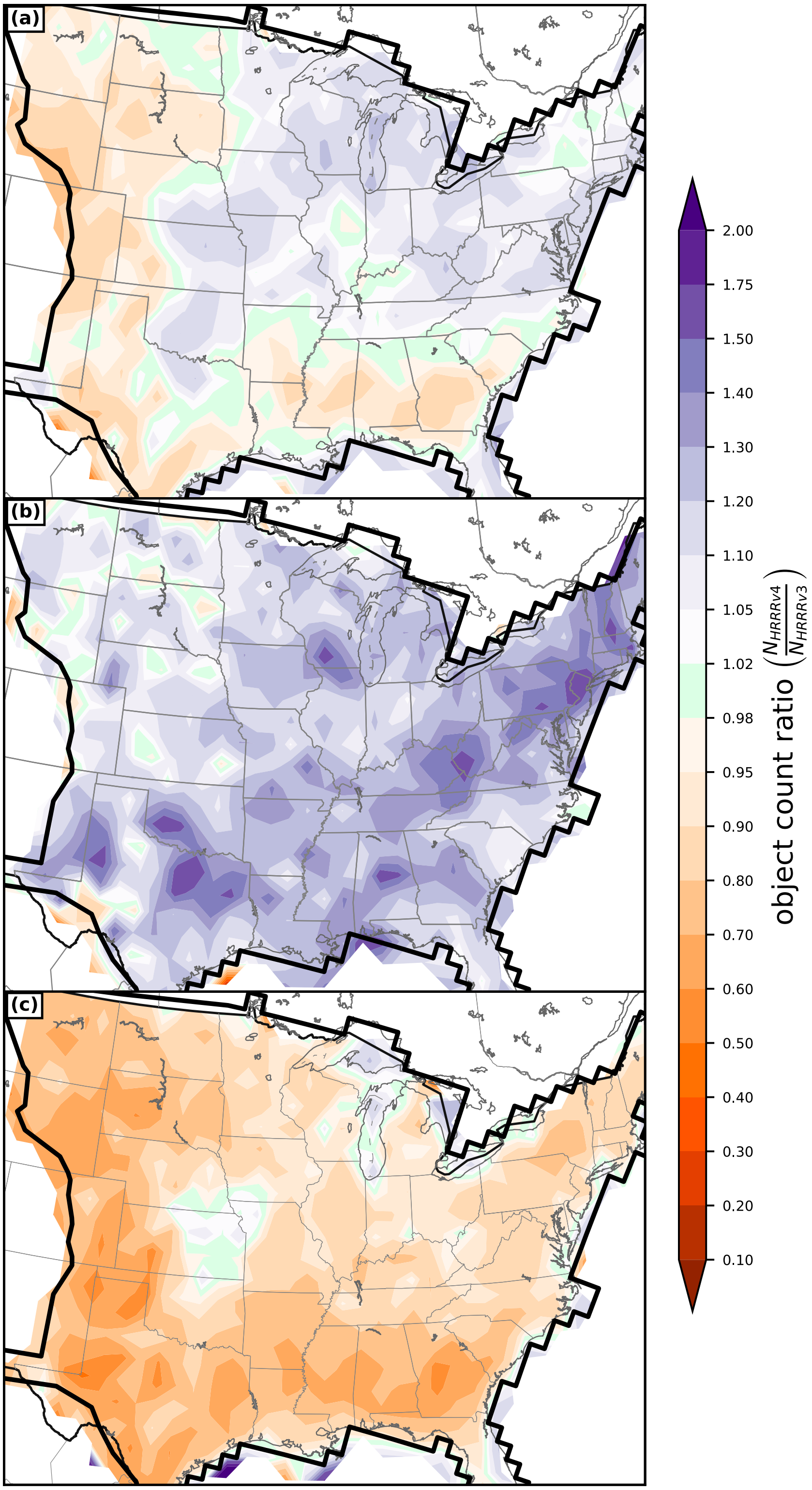


FIG. 3. Ratio of reflectivity object counts between HRRRv4 (numerator) and HRRRv3 (denominator) aggregated across all forecast hours and valid (a) at all hours of the day, (b) from 0600 to 1100 UTC, and (c) from 1800 to 2300 UTC. Objects are defined using a 25-dB*Z* reflectivity threshold. The verification domain is delineated in a thick black outline. Location bins were spaced by 1.0° latitude and longitude.

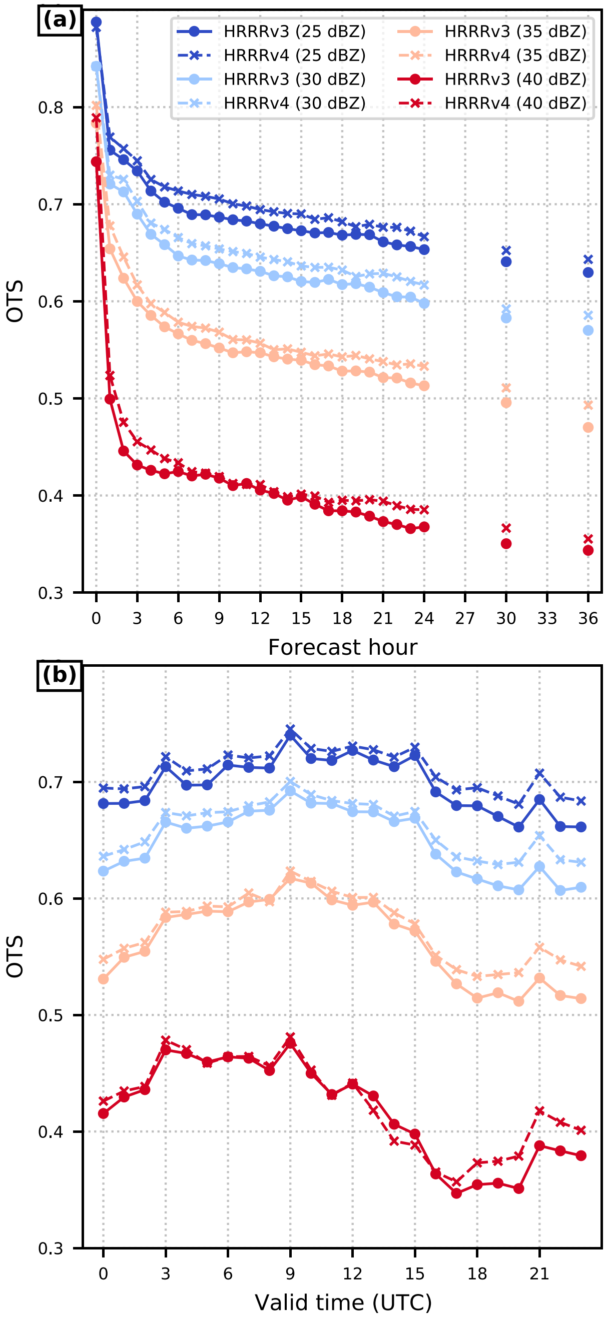


FIG. 4. As in Fig. 2, but for OTS.

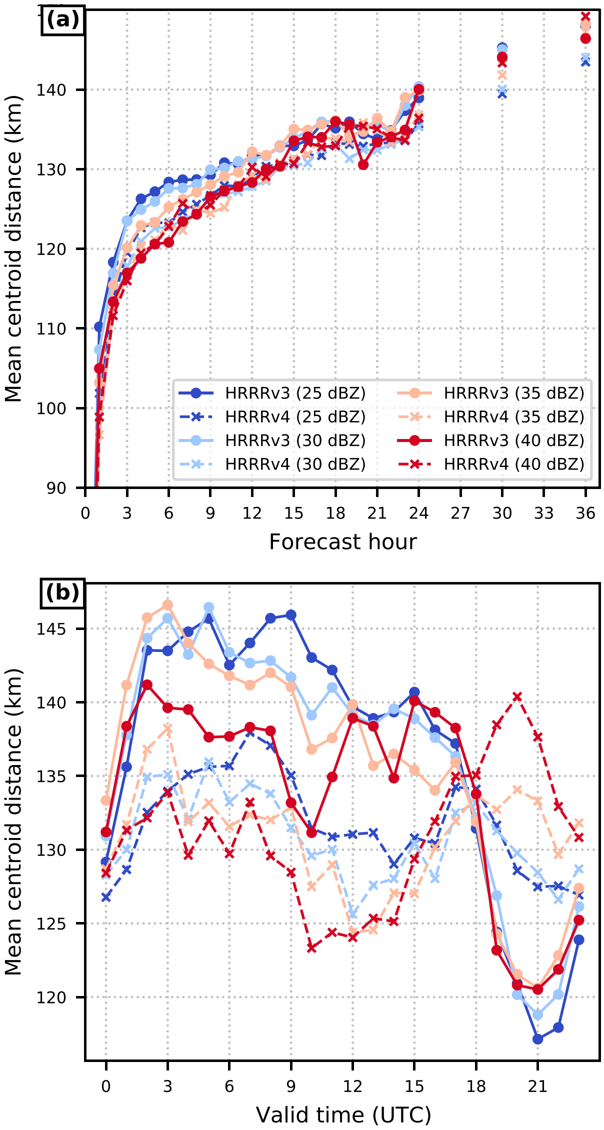


FIG. 5. Mean distance between centroids of F-O object pairs obtained from generalized matching as a function of (a) forecast lead time and (b) valid time of day but aggregated only across forecasts of length 12–18 h.

except the afternoon, when its frequency bias was lower than that of HRRRv3 (Fig. 5b). Considering that the spatial distribution of forecast object centroid displacements from their corresponding observation objects was essentially identical between the two models (not shown) and the apparent inverse correlation between frequency bias and mean centroid distance, the difference in mean distance between the two models likely includes a "paintball effect" in which the model that produced more storm objects achieved a lower mean distance by chance effects of covering the vicinity of observed storms with forecast storms, as opposed to better forecast discrimination in which forecast objects were closer to their observation counterparts only where observation objects were present.

MODE enables investigation of the spatial variance of object attributes as well, and there were discrepancies between HRRRv3 and HRRRv4 as well as between each model and the MRMS observations. Composite reflectivity object aspect ratio in the MRMS data tended to be lower in the northern and eastern halves of the United States and larger across portions of the southeastern United States and along most of the Texas and High Plains (Fig. 6a). An overall high bias is present in both models, with HRRRv3 having a higher aspect ratio

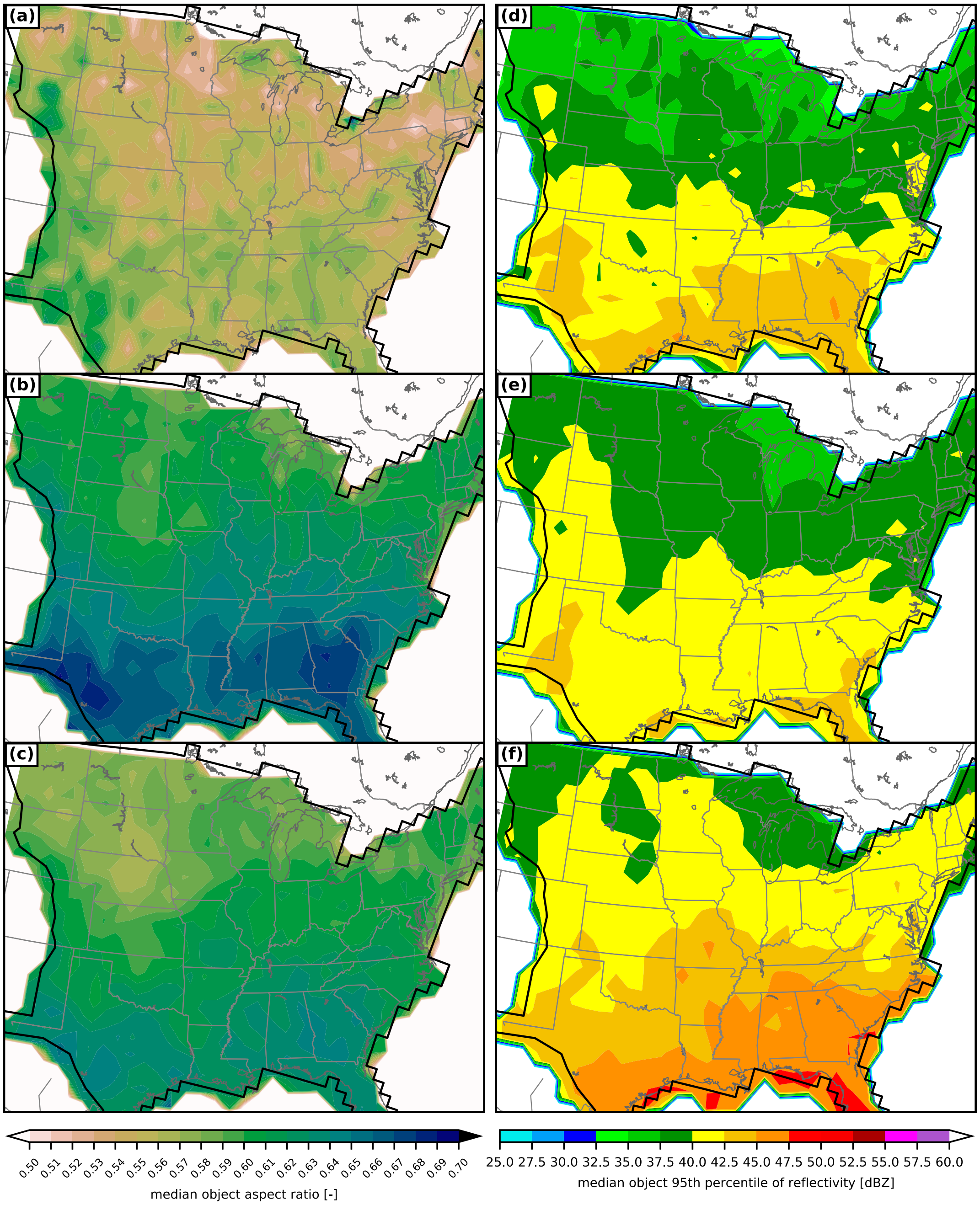


FIG. 6. Spatial distribution of median (a)–(c) object aspect ratio and (d)–(f) p95 from the 25-dBZ threshold data. (top) MRMS, (middle) HRRRv3, and (bottom) HRRRv4 data.

than HRRRv4, but the spatial pattern of aspect ratio is not quite the same in the models as in the MRMS. In particular, the median aspect ratio tended to be at a minimum from Nebraska northward through North Dakota (Figs. 6b,c). No other object attribute exhibited a similar pattern in that area, so the implication is that more storms having a somewhat more linear shape (or fewer circular storms) must have occurred there in the models, whereas observed linear storms tended to be spread more evenly across the northern United States.

The median reflectivity p95 values tended to increase from north to south in the MRMS (Fig. 6d) with minor longitudinal variation. The overall pattern of median p95 values in both models broadly agreed with that of the MRMS, but there were biases of this attribute as well. In particular, HRRRv3 had a high bias in the northern United States and a low bias in the southern United States (Fig. 6e). On the other hand, HRRRv4 had a high bias generally everywhere (Fig. 6f). The IEVA in HRRRv4 could be responsible for the increased p95, as stronger updrafts tend to promote higher reflectivity values.

In both models, objects tended to be slightly larger on the High Plains and lower in the southeast United States. Objects also tended to be more complex in the northern United States and less complex in the southern United States. Coupling this information with the information on spatial aspect ratio patterns suggests the storm mode was more commonly cellular in the southern and Southeast United States and more linear in the northern United States and on the plains, with embedded mesoscale convective systems also being more common there.

c. *Differences by object size*

1) Object count and size mismatches

DT21 observed that the majority of composite reflectivity objects in HRRRv3 were small, with areas of $O(100)$ km^2. These objects were found to be overforecast. This characteristic was also found in HRRRv4. Considering the effective resolution of the WRF model of $7\Delta x$ (Skamarock 2004), a fully resolved two-dimensional feature could be considered to have an area of $49\Delta x^2$ (441 km^2). On the other end of the size spectrum, a potentially useful size threshold discriminates convection having or not having mesoscale organization. Therefore, like the classification-by-area in DT21, here we analyzed object attributes from three independent size bins: small (area $\leq$ 441 km^2), medium sized (441 km^2 $<$ area $\leq$ 20 000 km^2), and large (area $>$ 20 000 km^2) in addition to analysis of the full set of objects.

Using the above size classifications, about 57% (60%) of objects were classified as small in HRRRv3 (HRRRv4), 41% (38%) as medium sized, and 1% (1%) as large at the 25-dBZ threshold. For higher reflectivity thresholds, the area distribution of composite reflectivity objects shifted increasingly toward small objects such that about 69% of all objects were classified as small at the 40-dBZ threshold in both HRRRv3 and HRRRv4, and the fraction of objects considered large was $\ll$1%. Therefore, the behavior of statistical distributions of object attributes and performance metrics was strongly linked to the behavior of the smallest objects. Also, these size classification proportions were weighted more toward smaller objects compared to the observations, which manifests as object count biases greater than 1.0 in the smaller size bins for both HRRRv3 and HRRRv4 (Figs. 7a and 8a). The high bias for small objects was pervasive throughout the forecasts (Fig. 7a).

It is clear from Fig. 7b that the decrease in frequency bias during the afternoon in HRRRv4 forecasts is mostly due to a decrease in the overforecasting of small objects, but with some contribution from medium-sized objects as well. There are differences in the behaviors of the high biases within these broader size categories, however. Specifically, HRRRv4 tended to have a worse high bias for the smallest objects compared to HRRRv3 and for objects ~7500 km^2 and larger. This result for HRRRv4 was also found in Grim et al. (2022). In size bins between 400 and 7500 km^2, however, HRRRv4 was slightly less biased than HRRRv3 at all thresholds. In fact, at the 25-dBZ threshold, HRRRv4 was overall less biased than HRRRv3 for all objects larger than 400 km^2 (Fig. 8a). However, since HRRRv3 was nearly unbiased for the large-sized objects, that means HRRRv4 had a high count bias even for the largest objects (area $>$ 20 000 km^2; Fig. 7).

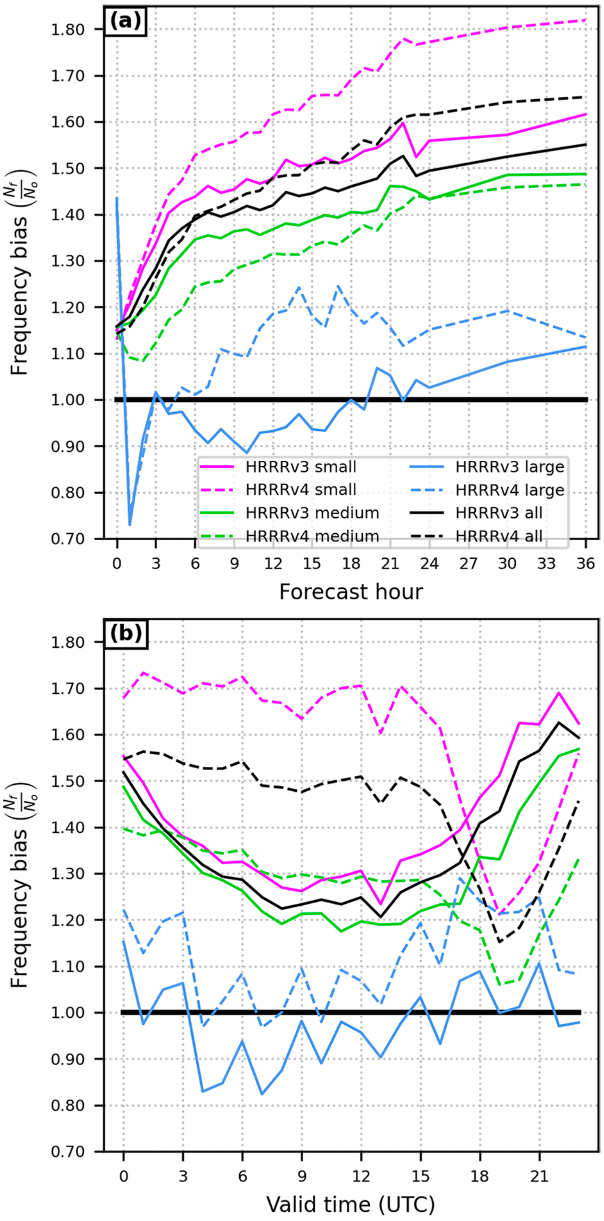


FIG. 7. Frequency bias of composite reflectivity objects subdivided by size as a function of (a) forecast length and (b) valid time of day using the 25-dBZ threshold.

DT21 contended that the particular distribution of frequency biases by object size obtained from HRRRv3 forecasts was due to the overproduction of new objects within a size bin rather than a mismatch between the sizes of comparable objects in F-O pairs. Here we computed the ratio of the areas of the objects in each F-O pair using generalized matching to help settle this debate. Figure 8b shows that for both

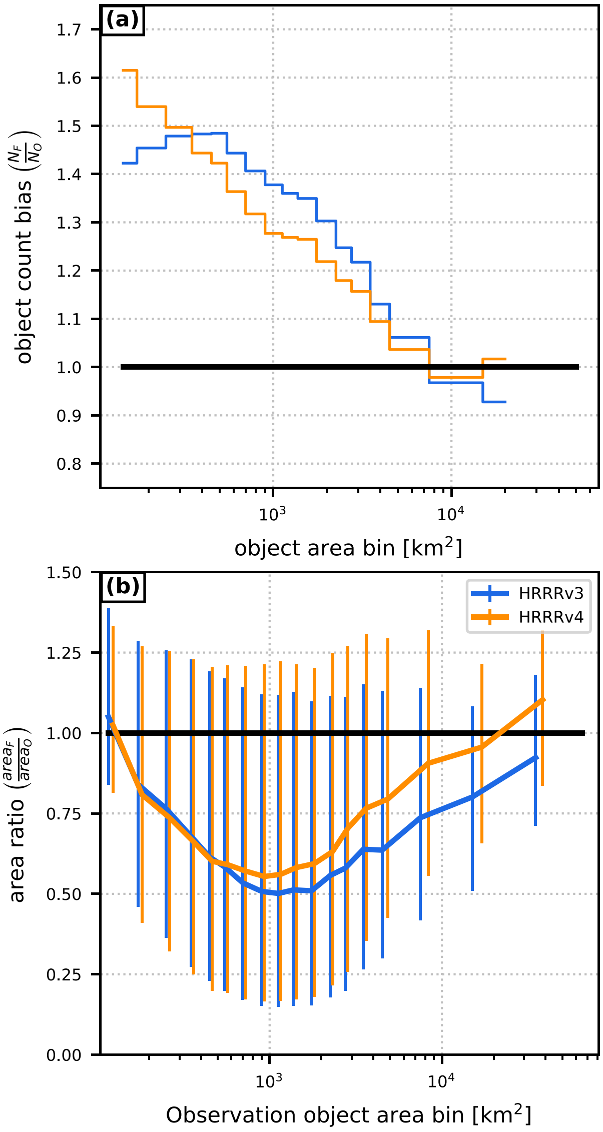


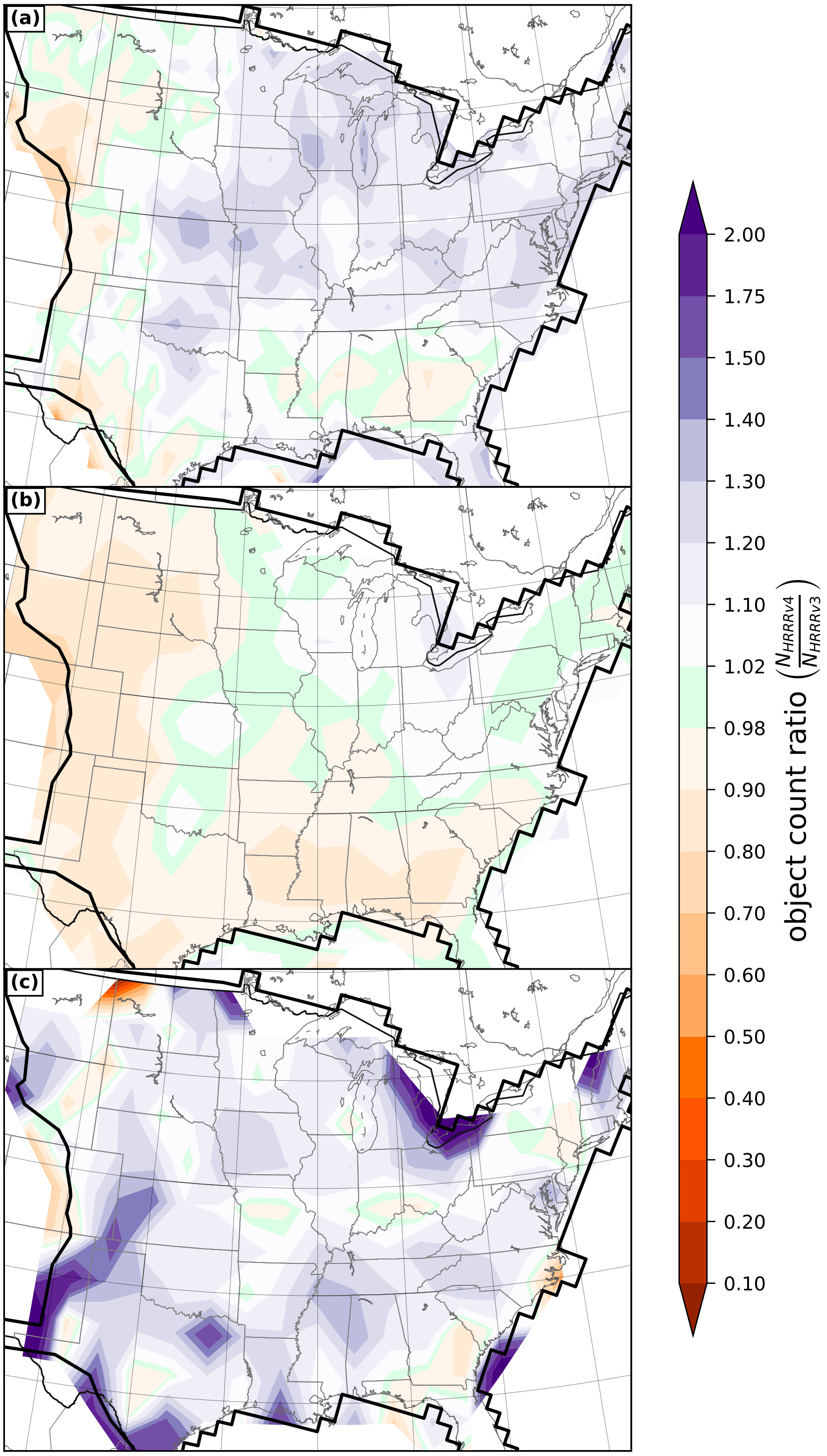
FIG. 8. (a) Unconditional (i.e., no generalized matching) object count bias, defined as N_{forecast} divided by N_{observed} within each area bin, and (b) forecast-to-observation area ratio (defined as F_{area} divided by O_{area}) from generalized F-O object pairs as a function of the observation object area. Objects are valid at the 25-dBZ reflectivity threshold. Vertical bars in (b) represent the interquartile range of the area ratio distributions within each area bin and a slight x-axis offset is used to distinguish between HRRRv3 and HRRRv4 curves. The stepping in (a) connects area bin midpoints as a means of delineating the area bins; the same bins are used for all area distributions in this paper.

the smallest and largest size bins the median ratio of forecast object area to observation object area in a given pair was close to 1.0, suggesting that the forecast object that best corresponded to each observation object in those size bins tended to be close to the correct size. Therefore, the frequency biases > 1.0 in the small size bins (Fig. 8a) suggests the models are producing too many small storms, many of which are falsely predicted (i.e., there is no observed storm to correspond to), whereas the frequency bias near 1.0 in the largest bins suggests an overall good forecast. However, for the other size bins, the median area ratio was substantially lower than 1.0, reaching as low as 0.50 for some area bins in HRRRv3. Therefore, for F-O pairs containing a medium-sized observation object, there was a tendency for the forecast object to be too small. This analysis suggests that both object count issues and incorrect size issues are contributing to the frequency bias distribution in Fig. 8a, with the magnitude of each factor likely having some dependence on object size.

There was some variability in the area ratios (Fig. 8b); the interquartile range of the area ratio distributions at each size bin span 1.0, indicating there were forecast objects that were too large as well as too small in each observation object size bin. However, since the bulk of the distribution of area ratios was less than 1.0, underforecasting of storm size was a more common issue. As an example, consider the area ratio distribution[3] corresponding to observation object sizes around 1000 km^2, where the median area ratio is about at minimum (0.50 for HRRRv3 and 0.55 for HRRRv4). Since the area ratio was <1.0, the forecast storms corresponding to observation objects in this size bin were too small; if they were the correct size there would be fewer forecast objects in this bin, and the counts for those storms would be placed into larger size bins when computing the object count bias as a function of object size, resulting in a lower object count bias in the bin centered on 900 km^2 and a higher count bias in larger size bins, overall leveling the object count bias curve in Fig. 7a. Specifically, there would be fewer forecast objects in size bins that are about 0.5 (the median area ratio) times the observation area {e.g., for the observation size bin of [800, 1000) km^2, the median size of the forecast objects would be in the range [~400, ~500) km^2}. If those forecast objects had the correct size (800–1000 km^2), then they would not be of size 400–500 km^2, which means there would be fewer counts in the [~400, ~500) km^2 size bin when computing the object count bias in Fig. 8a. Therefore, the object count bias would decrease toward 1.0 in that size bin and would increase in larger size bins. Applying this reasoning to the remaining size bins suggests the object count bias curve would flatten if all forecast objects were of the correct size, and instead a nearly constant (across size bins) bias slightly above 1.0 would remain, which is the signature of the overall overproduction of storms rather than improper storm size. The fact that both the median area ratio tended to be slightly higher and the frequency bias slightly lower in HRRRv4 compared to HRRRv3 in the medium and large size bins suggests that HRRRv4 may have done better than HRRRv3 in producing reflectivity objects of the correct size.

HRRRv4 also did better than HRRRv3 in producing fewer reflectivity objects of small and medium size spatially (Figs. 9a,b).

[3] The shape of the area ratio distribution varied by size bin; the distribution was approximately Gaussian for the smallest and largest bins but shifted toward an inverse exponential for the other size bins.



**Note: grid spacing for contour plots differs between panels*

FIG. 9. As in Fig. 3a, but for (a) small, (b) medium-sized, and (c) large objects, each aggregated across all forecast hours.

The spatial distribution of the ratio of reflectivity objects between the two models shows that HRRRv4 produced fewer small objects overall in parts of the Southeast United States as well as most of the High Plains, including much of Texas, whereas HRRRv4 produced more small objects across the eastern plains, the Midwest, and the Appalachian region. HRRRv4 produced somewhat fewer medium-sized objects over most of the verification domain, with the biggest discrepancy running along the coast of the Gulf of Mexico and the High Plains. The two models forecast a similar number of medium-sized objects over the Great Lakes region. HRRRv4 forecasts more large objects than HRRRv3 across nearly all the verification domain, consistent with it having a higher frequency bias for the largest size bins (Fig. 9c).

2) OTHER OBJECT ATTRIBUTES

Other attributes show a dependence on object size, too. One such attribute is the aspect ratio. DT21 found an anomalous spike in the distribution of aspect ratio values in the bin of [0.80, 0.85). This spike also appears in the HRRRv4 data but is restricted to the small objects (Figs. 10a,d,g). Upon further examination, many of the objects whose aspect ratio fall into this bin were revealed to be a reflectivity signature called a "flower," a relatively small and typically high-intensity reflectivity core flanked by at least one near-grid-scale, low-reflectivity dot (petal), typically along one or both of the cartesian dimensions. The classic flower signature contains four petals, one along each of the west, north, east, and south flanks of the core. Some flowers had fewer than four petals, whereas others had as many as six. These flowers attracted the attention of NOAA/GSL scientists when evaluating HRRRv3 forecasts (e.g., Turner et al. 2020) and were not found in previous operational HRRR versions. Attempts were made to reduce or eliminate such signatures in HRRRv4, although to only limited success. Part of the cause of these signatures is believed to be related to the magnitude of the horizontal diffusion setting in the WRF, but a more substantial diagnosis is beyond the scope of this paper. Approximately 55% of nearly 1000 objects in this aspect ratio bin that were manually inspected were classified as flowers in HRRRv3 whereas 48% of about 700 manually inspected HRRRv4 objects were classified as flowers, although the flowers in HRRRv4 were typically less distinct. Flowers were found in other aspect ratio bins, too, especially bins closer to 1.0, but most were in the [0.80, 0.85) aspect ratio bin. The fact that the spike in that particular bin of aspect ratios, and the fact that a smaller fraction of objects with this description were flowers in HRRRv4 compared to HRRRv3 suggests an amelioration of this issue in HRRRv4 over HRRRv3.

In general, Fig. 10a shows that the distribution of aspect ratios of HRRRv4 shifted closer to that of the MRMS data for small objects compared to HRRRv3, although an overall bias toward higher aspect ratios (more-circular objects) remains in HRRRv4. There is a hint of a positive bias in aspect ratio for medium-sized objects as well (Fig. 10d), although to less of an extent as for the small objects. The distribution of aspect ratio of large objects (Fig. 10g), on the other hand, is similar between both HRRRs and the MRMS data, indicating a good forecast of the shape of large reflectivity objects.

The shape of the distribution of object complexity also varied with object size (Figs. 10b,e,h). Small objects had a predominately low complexity value in both models {mode value in the [0.10, 0.15) bin}, and both exhibited a low bias compared to that of the MRMS data, in which the mode of the complexity values was in the range of [0.20, 0.25). This behavior was also noted in DT21 and illustrates the under-resolved nature of objects of this size, given they are smaller than an assumed square of side $7\Delta x$. Medium-sized objects also had a low complexity bias, although not to the same extent as for the smallest objects; the distributions among the datasets have nearly the same shape with a slight negative bias and positive skew that is worse in HRRRv3 compared to HRRRv4 (Fig. 10e). It is possible that the changes in horizontal diffusion are responsible for the difference between HRRRv3 and HRRRv4. It is unlikely that the initial conditions contribute to this signature since the difference is approximately steady with forecast length (not shown). Large objects have approximately the correct complexity distribution (Fig. 10h).

The distribution of object p95 values in the MRMS data is bimodal for small and medium-sized objects (Figs. 10c,f, respectively), suggesting two separate behavioral modes were sampled. The small objects have modes in with dBZ values in the low 30s and upper 40s, whereas those for medium-sized objects are about 5 dBZ higher. This difference could be a matter of sampling: for small objects, there may only be one or two grid points at which the reflectivity is near its max, and thus the 95th percentile may undercut the one-gridpoint-maximum value; whereas medium-sized objects tend to comprise either multiple convective cores or more mesoscale organization, and thus a broader reflectivity distribution, which makes it more likely that the 95th percentile value more closely resembles the gridpoint maximum reflectivity within the object. The lower-reflectivity mode occurs chiefly during the evening through overnight and early morning when overall lower convective instability results in somewhat weaker peak storm intensities (Fig. 10c), whereas the higher-reflectivity mode appears during the afternoon when individual storms tend to be more intense in response to diurnally elevated instability. Neither HRRRv3 nor HRRRv4 adequately capture the bimodal nature of the p95 distribution, although the distribution is broader in HRRRv4 than HRRRv3, suggesting a hint that the former version may have attempted to resolve some of the detail exhibited by the MRMS data.

Both HRRRv3 and HRRRv4 exhibited patterns in frequency bias with respect to other object attributes as well. Both models tended to overforecast the number of storms that are mostly linear and nearly circular (aspect ratios close to 0.0 and 1.0, respectively; Fig. 11a). However, this behavior was less prominent in HRRRv4 over HRRRv3. The signature for large aspect ratios is consistent with the distribution of aspect ratio values for small objects indicating the presence of flowers. Since there are so few objects with aspect ratios close to 0.0, the higher frequency bias in that range of aspect ratios is most likely noise, but still indicates a tendency for the models to produce more linear objects than actually occurred.

Both models also tended to overforecast both the least-complex and most-complex reflectivity objects, again with HRRRv4

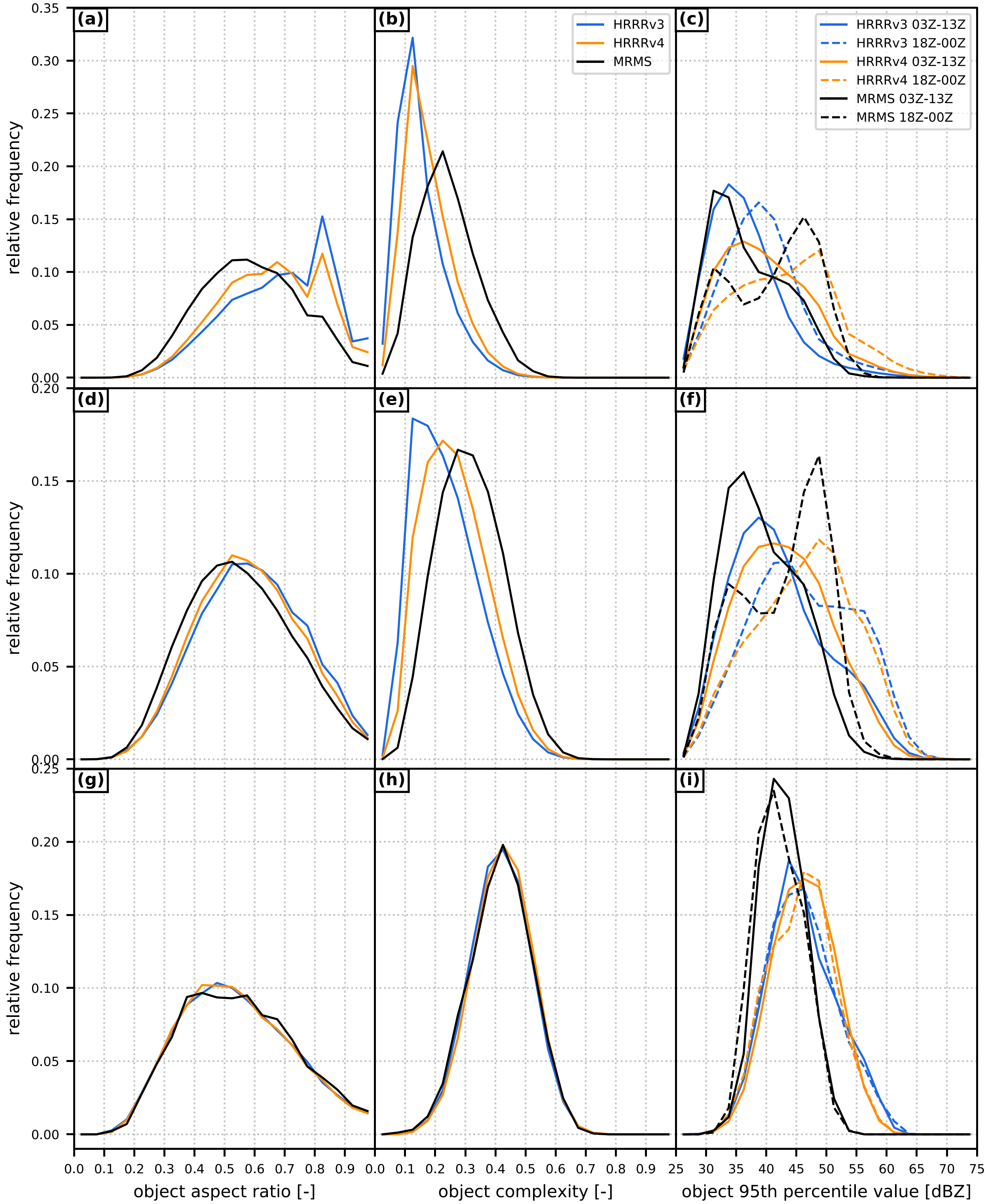


FIG. 10. Histograms of (a)–(c) small-sized, (d)–(f) medium-sized, and (g)–(i) large-sized reflectivity object (left) aspect ratio, (center) complexity, and (right) p95 attributes aggregated across all forecast hours for objects defined at 25 dBZ. For p95, the distributions have been split between diurnal and nocturnal groups to highlight differences for small- and medium-sized objects. The same *y*-axis scaling is used across the panels in each row.

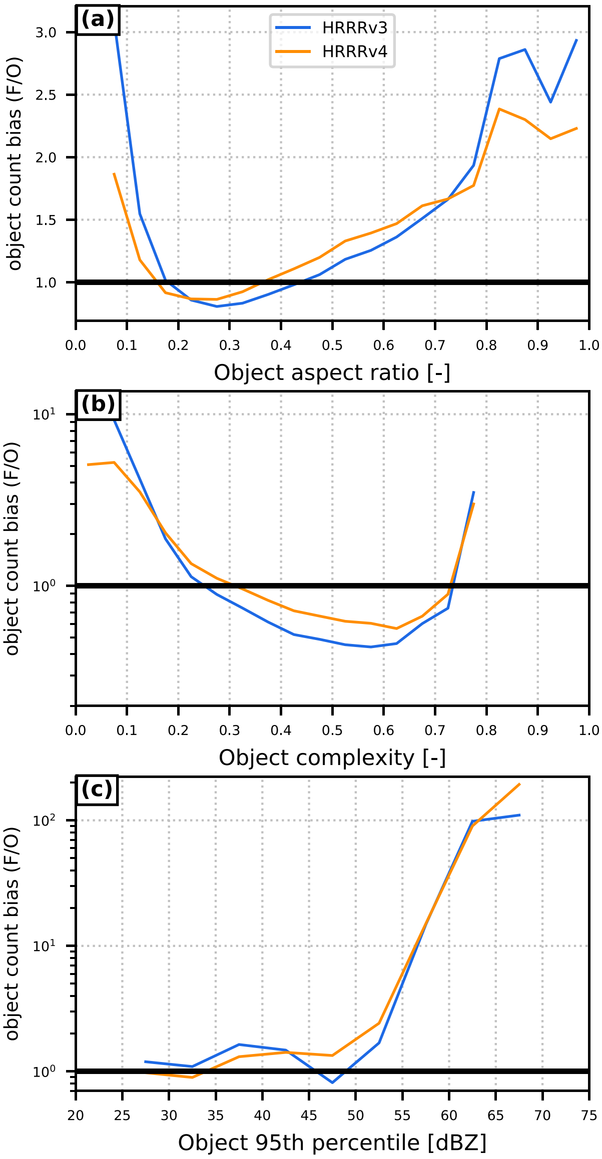


FIG. 11. Unconditional (i.e., full set of objects) forecast reflectivity object count biases as a function of object (a) aspect ratio, (b) complexity, and (c) 95th percentile of reflectivity.

exhibiting this behavior to a lesser degree (Fig. 11b). This is a prominent finding since a substantial fraction of objects have low complexity values, and this result suggests HRRRv4 produced storms with more realistic structure compared to HRRRv3 even though the grid spacing was the same between the two models. Both models produced too few storms with moderate or large complexities, especially at higher reflectivity thresholds. But again, HRRRv4 tended to not be as bad as HRRRv3.

Both models also tended to overforecast the number of storms with a high near-peak reflectivity value, especially for the strongest storms (those with p95 values of 50 dBZ or greater; Fig. 11c). We chiefly suspect this offset to be due to the models' representation of components of deep convective storms such as updraft mass flux and conversion of vapor to condensate and hydrometeors, including the distribution of various hydrometeor species within storms. Unfortunately, due to a dearth of measurements of these quantities, it is impossible to distinguish the relative impacts of these microphysics scheme errors on the reflectivity forecasts. Higher updraft speeds permitted by the IEVA scheme may have worsened this for HRRRv4 given the slight uptick in bias in the 65+-dBZ bin. The role of the reflectivity diagnostic in the Thompson et al. (2008) microphysics scheme within the HRRR must also be considered; it includes an amplification factor at grid points containing "wet" snow (snow with liquid water on its surface), which is more likely to occur in storms with some mesoscale organization (i.e., medium-sized to large-sized objects), and there is some indication of a worsening high bias for larger objects (Fig. 10i).

d. *The impact of generalized matching*

Generalized matching produces a lesser-biased dataset from which to perform forecast verification. A first-order evaluation is to check how many objects in the datasets were paired. About 1.4 million, 980 000, 645 000, and 300 000 F-O pairs were classified across all forecast hours at the 25-, 30-, 35-, and 40-dBZ thresholds, respectively, which amount to about 55%–60% of the number of forecast and observation objects in the full datasets, except at 40 dBZ where the fraction is ~37%. Consistent with the high bias in the total number of objects, a substantially larger fraction of forecast objects was unpaired compared to the number of observation objects (cf. Figs. 12a,b). The fraction of unpaired forecast objects was smallest at forecast initialization and increased dramatically during the first few hours, then slowly thereafter. The fraction of unpaired forecast objects also increased with increasing reflectivity threshold, also consistent with the frequency bias trend. In contrast to the forecast objects, far fewer observation objects were unpaired. Other than at forecast initialization in which only about 2%–7% of observation objects were unpaired, about 10%–15% of observation objects were unpaired during the forecast. These values are consistent with the imperfect PODs (based on pair interest threshold of 0.70) that were around 0.95 at forecast initialization and decreased slowly from 0.70 to 0.60 throughout the forecast (not shown). Ultimately, the fact that the number of unpaired observation objects was anything but nearly zero suggests the HRRR missed some objects, even despite its overall high bias.

Object attributes remain imperfectly forecast when considering the lesser-biased set of objects obtained through generalized matching. Consider object frequency bias as a function of area (Figs. 13a,c). While some degree of high bias remains, the values are much lower than for the full dataset, with the bias for small objects reduced by about 90% compared to using the full set of objects (cf. Fig. 8a). Some low biases appeared. Objects of size of $O(100)$ km^2 (HRRRv4) and about 2000 km^2 (HRRRv3) and larger were reduced to having a low bias, and the low bias for the largest objects was degraded slightly. The relationship between frequency biases for HRRRv3

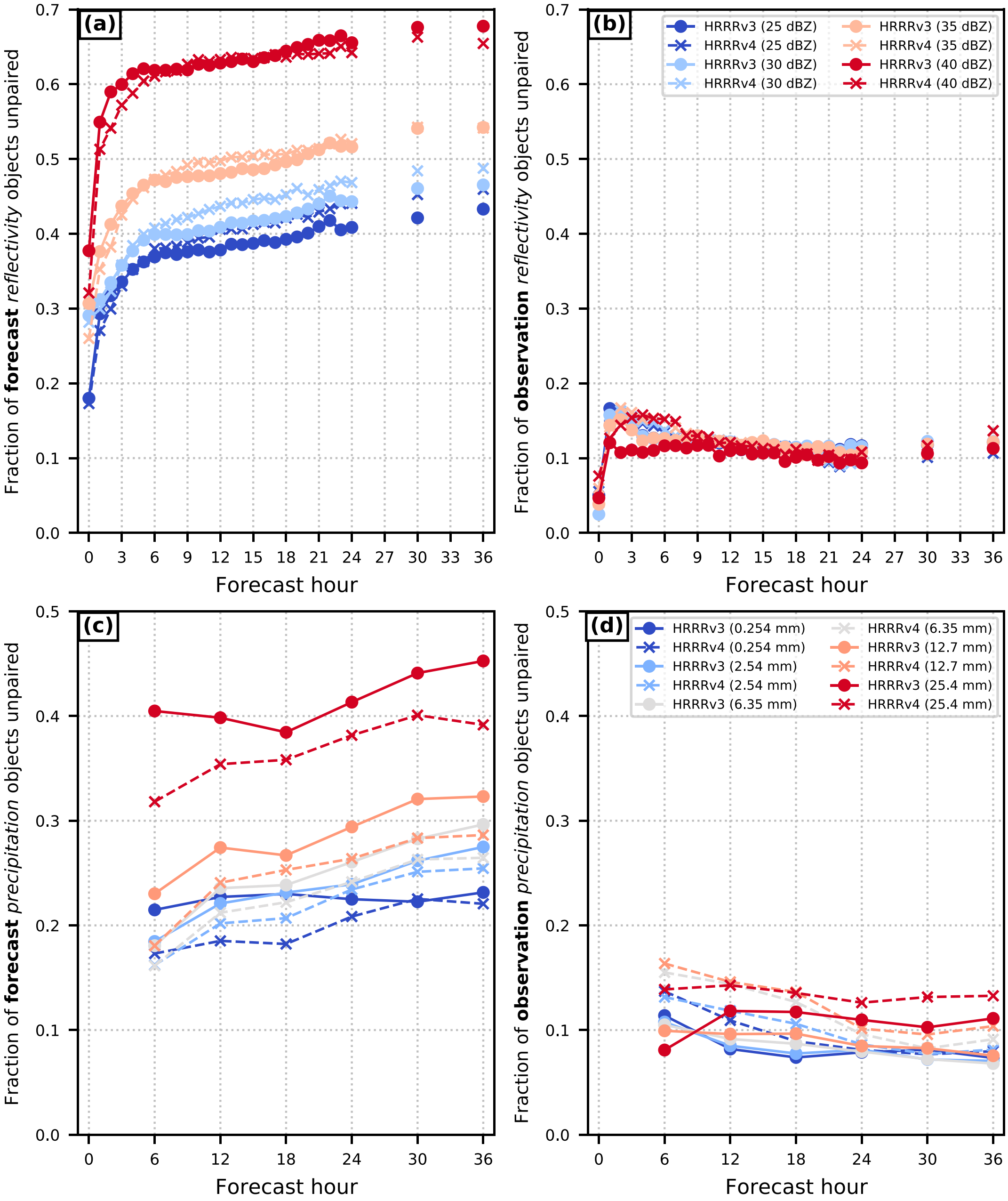


FIG. 12. Fraction of all (a),(b) reflectivity and (c),(d) 6-h precipitation objects unpaired after generalized matching as a function of forecast lead time. Values for (left) forecast objects and (right) observation objects.

and HRRRv4 was the same as for the full dataset (cf. Figs. 8a and 13a), and the pattern for HRRRv4 is nearly the same as that for HRRRv3 except shifted toward smaller object sizes; the lower bias values for HRRRv4 compared to HRRRv3 for small- and medium-sized objects makes it a poorer forecast than HRRRv3. At higher thresholds, both models were approximately unbiased for all but the large objects (Fig. 13c). The largest objects were substantially overforecast by both models, with HRRRv4 forecasting more such objects than HRRRv3. The similarity in bias ratios for the largest objects between the full set (right side of Fig. 8a) and the reduced set obtained from generalized matching implies either that few objects were removed from either the numerator (forecast object counts) or denominator (observation object counts) of the ratio, or that a similar number were removed to maintain the ratio of object counts. The former explanation seems more likely considering that centroid location and area ratio were emphasized in the interest calculation, so objects with similar locations and sizes are more likely to have higher interest values and thus survive the generalized matching. Also, the nature of a large object is that it

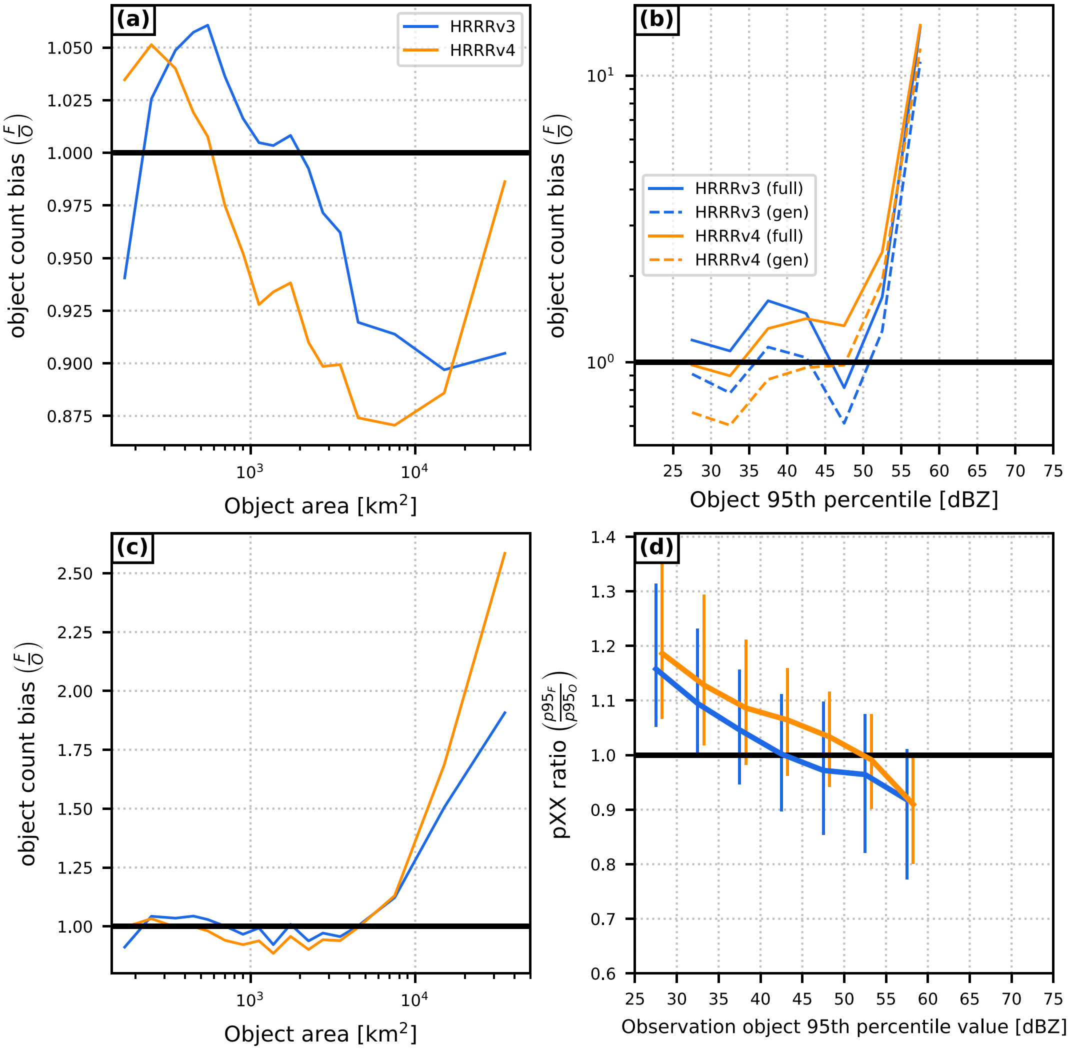


FIG. 13. Conditional reflectivity object frequency bias from generalized matching as a function of observation object area for objects defined at the (a) 25- and (c) 35-dB*Z* thresholds. (b) Reflectivity object frequency bias as a function of (observation) object 95th percentile of reflectivity for both the conditional ("gen") and unconditional ("full") object sets. (d) Forecast-to-observation ratio of object 95th percentile of reflectivity from the generalized matching set. Panels (b) and (d) are also valid at the 25-dB*Z* threshold. All data are aggregated across all forecast hours. Vertical bars in (d) illustrate variance as in Fig. 7b. Due to poor sample size, data were omitted in bins of 55 dB*Z* and greater in (b) and (d).

occupies more areal space, which limits the number of objects that could be located close to the object's own centroid, and therefore limit the interest value with nearby objects in the other dataset, which suggests little change should occur when performing generalized matching.

While the object count bias as a function of most other attributes decreased by a similar magnitude as to object area when comparing the full set of objects to the reduced set obtained through generalized matching, that as a function of p95 value changed little with generalized matching (Fig. 13b). The fact that the median ratio of forecast to observation p95 values in the generalized matching set was less than 1.0 for reflectivity values of 55 dB*Z* and higher (Fig. 13d) is intriguing. However, since the weight assigned to the intensity percentile ratio for computing object pair interest was 3.5 compared to 5.0 for centroid distance and 4.0 for minimum boundary and convex hull distance, as well as to area ratio, it is likely that the forecast object that best matched a given observation object did so due to a better resemblance of those location and size attributes rather than resembling the near-max reflectivity. It must have been the case that this difference in weighting caused the matching algorithm to favor closer storms that happened to be weaker rather than matching a stronger forecast storm to the observed storm. This observation also may unmask some of the sensitivity to the results that comes from the choice of attribute weights for the interest calculation.

4. 6-h precipitation results

Stage IV precipitation observations were less readily available than MRMS composite reflectivity observations, and therefore an average of about 1250 cases were verified in each of the first 18 forecast hours for 6-h precipitation. About 100 000, 70 000, 50 000, 35 000, and 17 500 forecast precipitation objects were classified at the 0.254-, 2.54-, 6.35-, 12.7-, and 25.4-mm thresholds, respectively. The distribution of sizes of 6-h precipitation objects was nearly identical to that of the composite reflectivity objects except the smallest objects constituted a slightly smaller fraction of all objects and the largest objects constituted a slightly larger fraction (not shown). Considering the causal relationship between composite reflectivity and precipitation and factoring in the temporal aggregation, this result is sensible.

a. *Object frequency bias*

Scalar metrics indicate mixed performance between HRRRv3 and HRRRv4 for 6-h precipitation. The object-based frequency bias for 6-h precipitation looks overall similar to that for composite reflectivity in that biases are all generally above 1.0, indicating overforecasting of precipitation regions (Fig. 14a). Also similar to composite reflectivity, the bias tended to increase with forecast lead time. HRRRv4 was less biased than HRRRv3 throughout the forecast except after forecast hour 24 at the 0.254-mm threshold, where the two models had similar biases. HRRRv4 was nearly unbiased at forecast hours 0–6. As with composite reflectivity, most of the total high frequency bias in the object counts came from the smallest objects, reflecting the precipitation produced by the excessive number of small reflectivity objects (not shown). However, the medium-sized objects also had a high bias, but to less of an extent than the small objects. The largest precipitation objects actually had a substantial low bias.

Frequency biases tended to be highest around 0000 UTC (early evening) and lowest around 1200 UTC (morning; Fig. 14b). Both models exhibited the same diurnal swapping of order seen in the composite reflectivity frequency biases around 0000 UTC, with HRRRv4 having an overall lower bias during the afternoon and early evening and a higher bias otherwise. While HRRRv3 was nearly unbiased during the morning and early afternoon at the 0.254-mm threshold, it had such a high bias at the 25.4-mm threshold that HRRRv4 only had a higher bias at 1200 and 1500 UTC at that threshold and otherwise had a lower bias.

The overall lower biases in 6-h precipitation compared to composite reflectivity are consistent with the fraction of unpaired objects (Figs. 12c,d). A substantially lower fraction of forecast precipitation objects was unpaired in the generalized matching compared to composite reflectivity objects. However, the fraction of unpaired forecast 6-h precipitation objects was higher than that of the unpaired observation objects. There was a decrease in the fraction of unpaired observation objects during the early part of the forecast, which contrasts with the increase seen in this quantity for composite reflectivity. However, the fraction decreases after forecast hour 1 in composite reflectivity, so this may also be happening in the precipitation field but is being smoothed out by the temporal aggregation.

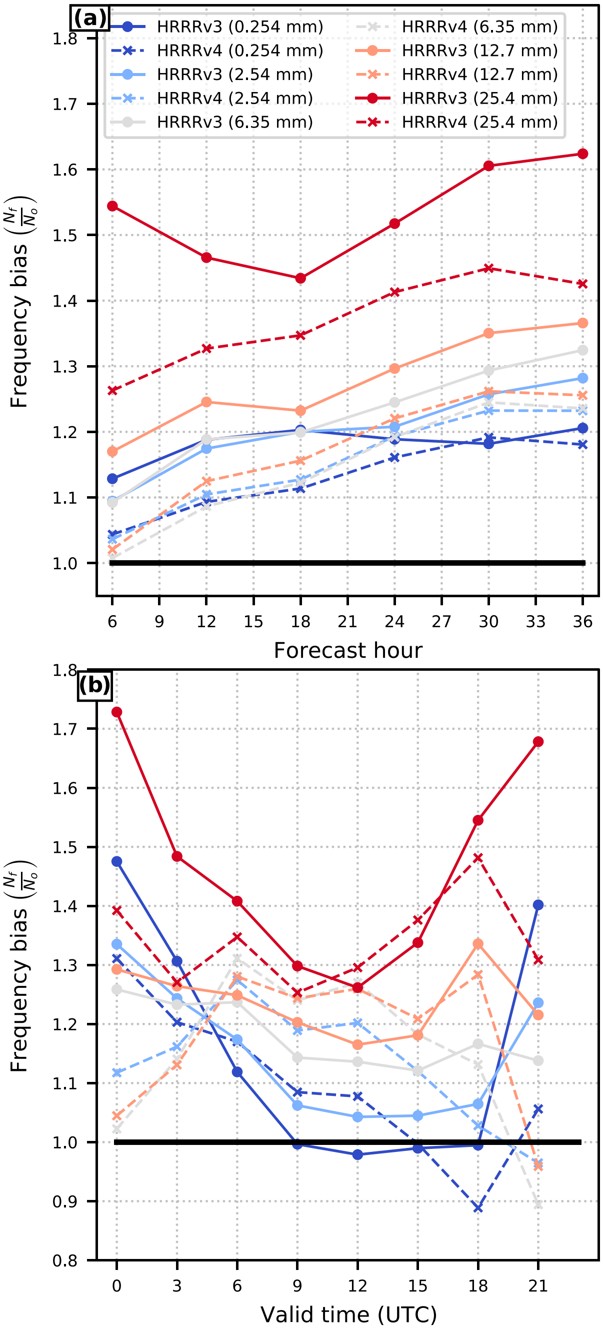


FIG. 14. As in Fig. 2, but for 6-h precipitation.

b. *Total forecast metrics*

As a result of its higher bias, HRRRv3 tended to have a higher POD, but at the expense of a slightly higher FAR as well (not shown). The bias was enough to penalize HRRRv3 in terms of OTS, which was slightly lower than for HRRRv4 (Fig. 15a). OTS

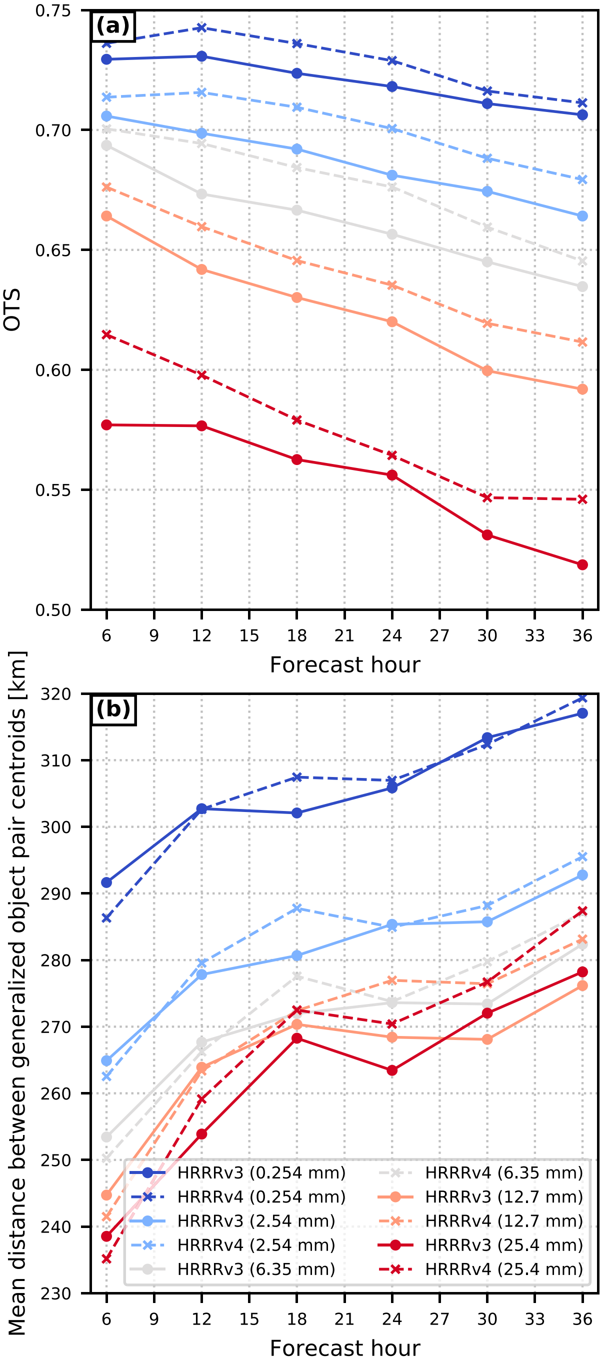


FIG. 15. (a) OTS and (b) mean centroid distance between generalized 6-h precipitation F-O object pairs, both as a function of forecast lead time.

decreased slowly throughout the forecast, similar to the trend in OTS values for composite reflectivity. Except for the first six forecast hours, HRRRv3 generally had lower mean centroid distances than HRRRv4 (Fig. 15b). Overall, mean centroid distances for 6-h precipitation increased steadily throughout the forecast and were much larger than those for composite reflectivity, ranging from 230 to 320 km. The increase in distance with lead time is steeper than for composite reflectivity. The overall broader scale of 6-h precipitation objects and degrees of freedom of distribution of precipitation within the objects is likely the main factor in this difference. As with composite reflectivity, the two-dimensional pattern of centroid errors was broadly isotropic, although there was a slight preference toward an error in the southeastward direction (not shown). This preference was slightly stronger for HRRRv4 than HRRRv3, but the difference in mean distance is a matter of a few kilometers. This contrasts with the slight northwesterly centroid error bias in composite reflectivity.

CRPSS values generally favored HRRRv4 for quality of precipitation forecast object attributes. The distributions of forecast object aspect ratio and complexity, in particular, were much better forecast by HRRRv4 than by HRRRv3. This is clearly visible from visual inspection of the distribution of these attributes at any given lead time and precipitation threshold (not shown). The HRRRv4 clearly alleviated some noticeable deficiencies evident in the distributions of these attributes in HRRRv3, including a low bias for complexity in HRRRv3 that was alleviated, although not completely eliminated, in HRRRv4 (Fig. 16a), and a spike in the same aspect ratio bin as was seen with flowers in composite reflectivity. The removal of this statistical mass from that bin in HRRRv4 was sufficient to result in CRPSS values above 0.5 (Fig. 16b). The distribution of object areas was better forecast by HRRRv4 than by HRRRv3 except for during the early parts of the forecast (and more of the forecast at lower precipitation thresholds; Fig. 16c). Since all datasets herein are dominated by small objects and the frequency bias of the small and medium-sized objects is lower in HRRRv4 than in HRRRv3, similar to composite reflectivity, that is the source of improvement that manifests in the positive CRPSS values. This improvement in frequency bias is more pronounced at higher thresholds; hence the increased CRPSS values at higher thresholds. The CRPSSs for p99 values, on the other hand, showed poorer performance in HRRRv4 compared to HRRRv3 (Fig. 16d); CRPSS < 0.0 at all thresholds and forecast lead times. These negative CRPSSs likely manifest from the degraded high bias in the number of 6-h precipitation objects with high p99 values as discussed in connection with Fig. 18b below.

c. *Analysis from generalized matching*

One of the most important aspects of spatial precipitation verification is how much overlap occurs between forecast and observed areas of rainfall, otherwise known as a “hit” in a 2 × 2 contingency table. The number of hits strongly impacts a given forecast performance metric but is almost always aggregated over a large spatial domain, which inhibits analysis of precipitation coverage overlap on local scales. Object-based verification, on the other hand, enables such analysis by examining various statistics describing the intersection area in an F-O object pair. Here we chose to use the consumption ratio. However, this metric can belie forecast quality if there is a high area bias, which is a problem for the HRRR. Therefore, consumption ratio must be analyzed in conjunction with other size-dependent metrics to get a more complete sense of the forecast quality. The median

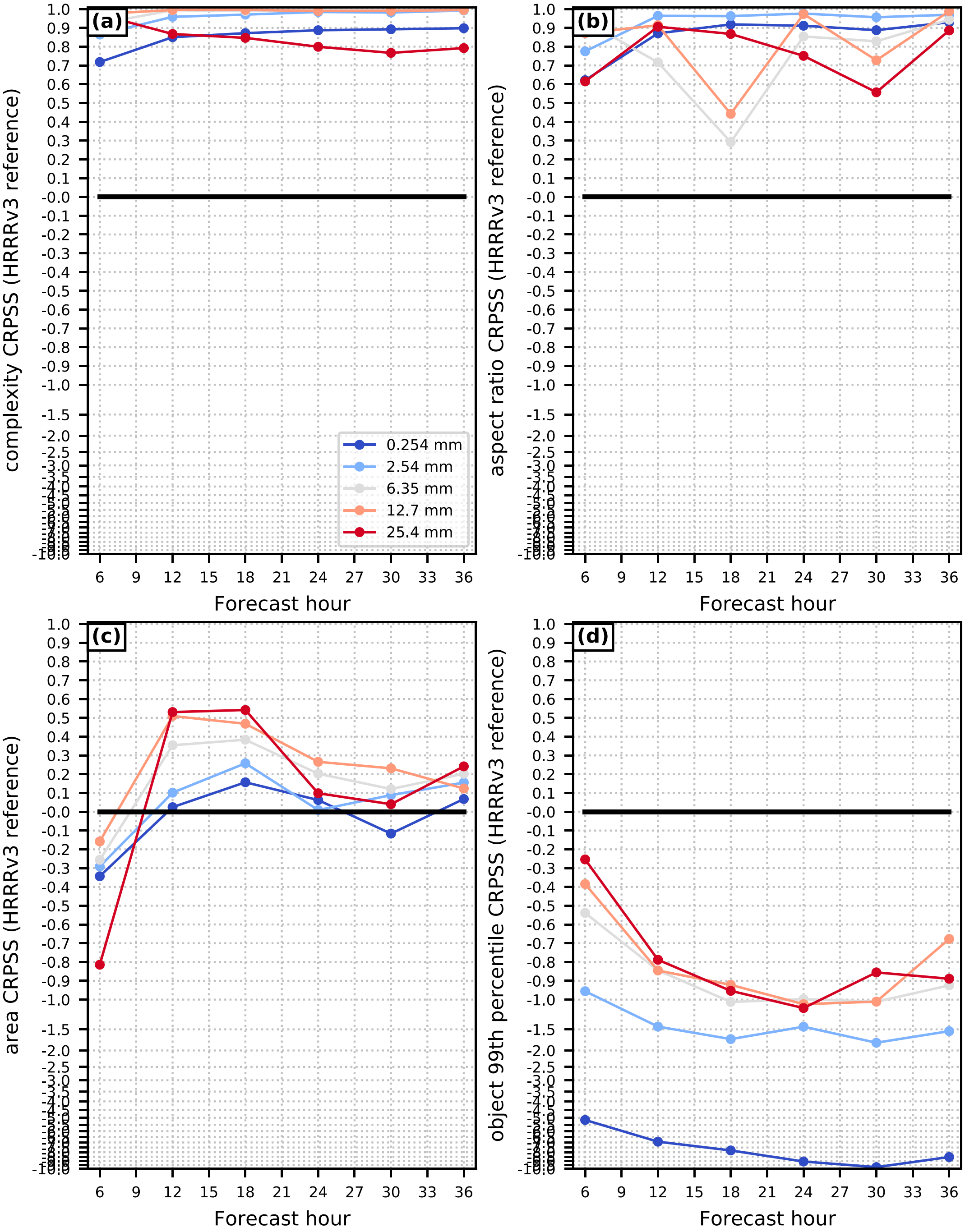


FIG. 16. HRRRv4 CRPSSs (using HRRRv3 as the reference forecast) for the following 6-h precipitation object attributes: (a) complexity, (b) aspect ratio, (c) area, and (d) p99. All scores plotted as a function of forecast hour.

consumption ratio (Fig. 17a) was 0.0 for the smallest observation objects, consistent with object pairs having no spatial overlap (in essence, a "miss"). At sizes larger than 500–700 km^2 consumption ratios rapidly increased with increasing observation area up to almost 10000 km^2, above which the rate of increase slowed. The largest objects and those defined at the lightest precipitation threshold had the highest consumption ratios (median of about 0.8). Considering both that the 75th percentile of the consumption ratio distribution was below 0.9 and the median area ratio for the largest objects was about 0.8–0.9 (Fig. 17b), it is apparent that there was significant displacement between large observation objects and their corresponding forecast object such that the forecast objects in the F-O object pairs, which were usually smaller than the

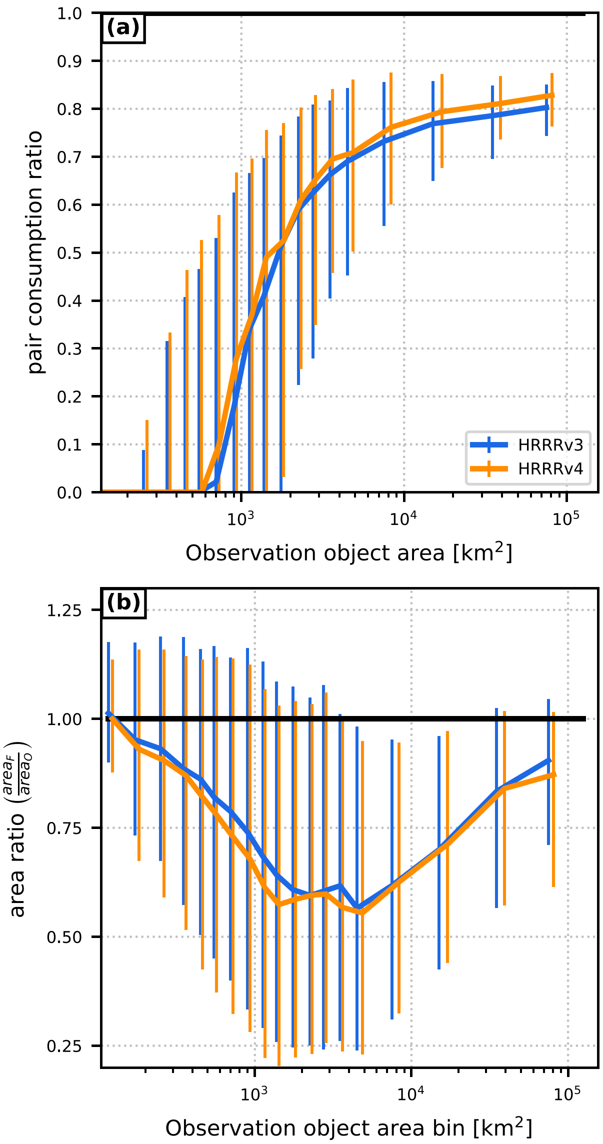


FIG. 17. (a) Consumption ratio as a function of observation object area, and (b) as in Fig. 8b, but for 6-h precipitation objects (obtained from generalized matching). Data are valid at the 0.254-mm threshold. Vertical bars and horizontal staggering are as in Fig. 8b.

observation object, were not fully encompassed by the observation object; In fact only about 80%–90% of the smaller of the two was contained within the bounds of its corresponding larger object. The median value of the consumption ratio tended to be slightly higher in HRRRv4 than in HRRRv3, but the ordering of the area ratios was inverted from that. HRRRv4 precipitation objects were overall smaller than HRRRv3 precipitation objects (not shown), so the increased consumption ratio in HRRRv4 suggests objects in the latter either were closer to or fit better the observation objects than those in HRRRv3.

Another of the most important aspects of precipitation forecasting is the amount, either in an areal sense or as a point maximum. The models' ability to forecast the maximum precipitation amount in each object (p99) varied by rainfall amount (Fig. 18a). Both models forecast objects with light and moderate precipitation amounts with approximately the correct frequency. However, they also substantially overforecast the number of objects with heavy peak rainfall amounts centered around 100 mm, followed by an adequate forecast of the number of objects with the most extreme precipitation maxima. This overforecasting was extreme, with object count ratios up to 4.0, and was worse in HRRRv4. The objects associated with this spike in bias are mostly large objects with higher complexity than objects with p99 values further from 100 mm (Fig. 18b). Total precipitation object mass was generally underforecast by both models, although both models slightly overforecast the mass of precipitation objects with low total rain mass (Fig. 18c). Neither model performed clearly better than the other with respect to object precipitation mass. The IEVA scheme could be contributing to the increased precipitation object count through more intense convective updrafts producing more precipitation.

5. Summary and conclusions

The High-Resolution Rapid Refresh model, a flagship convection-allowing forecast system useful for short-term high-impact weather, has been running operationally since 2014. The composite reflectivity and 6-h precipitation fields from versions 3 and 4, which ran in parallel during a portion of the 2019 and 2020 warm season months, were evaluated against each other using an object-based approach introduced in a companion paper (Duda and Turner 2021) and expanded herein. The purpose of this object-based verification study was to compare the performance of the two model versions to determine if the updates from version 3 to 4 resulted in improved forecasts of reflectivity and precipitation, and to elucidate specific aspects of where each forecast succeeded and where each forecast needed improvement in a way that traditional grid-to-grid verification would not be able to provide information. This study generated a large sample size, and statistical significance tests for the differences in forecast performance failed to provide a clear distinction between HRRRv3 and HRRRv4 in terms of forecast quality. Nonetheless, many overall scalar metrics at least hinted that HRRRv4 may have outperformed HRRRv3 in many ways, but may have degraded performance compared to HRRRv3 in others.

Substantial high storm count biases were confirmed in the 2020 forecasts from HRRRv3 to corroborate the high biases also found in 2019 in DT21. While the storm count bias in HRRRv4 was somewhat reduced compared to HRRRv3 during most of the forecast, HRRRv4 presented worse problems with overforecasting the smallest storms compared to HRRRv3. However, during the afternoon, HRRRv4 had a lower high bias in storm counts compared to HRRRv3. Since the lower bias was most strongly present in older forecasts (not shown), this change is unlikely to be related to improved initial conditions from the changes to the data assimilation, which suggests that an improvement in the model physics led to more realistic prediction of the diurnal cycle of storm counts. These differences were

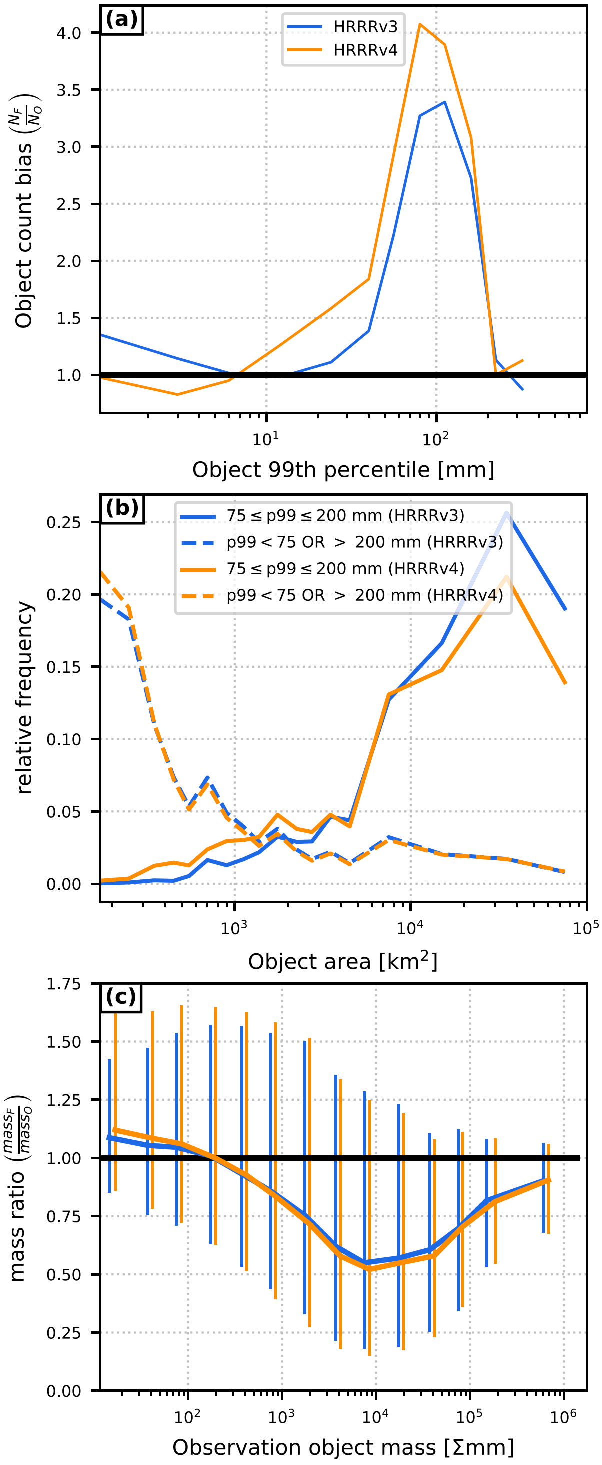


FIG. 18. (a) Unconditional forecast 6-h precipitation object count bias as a function of object 99th percentile of precipitation amount; (b) conditional distribution of object areas based on p99; (c) forecast-to-observation ratio of 6-h precipitation object masses as a function of observation object mass for generalized matching object pairs.

not spatially uniform, however; HRRRv4 tended to produce more storms than HRRRv3 in the Midwest and fewer storms in an arc across the Southeast United States through the southern plains and into the central High Plains. A detailed analysis of storm size comparisons revealed that the high count bias was due to two factors: 1) overproduction of new storms and 2) existing storms being too small.

Considering individual object attributes, DT21 revealed an anomalous spike in the distribution of reflectivity object aspect ratios that was identified as a computational/dynamical artifact called a flower. A larger fraction of inspected suspect objects in HRRRv3 were classified as flowers compared to HRRRv4, so the reduction in the spike in the aspect ratio distribution in HRRRv4 compared to HRRRv3 reflects the modest improvement in reducing the presence of flowers in HRRRv4. Although both models struggled with handling the reflectivity object complexity attribute, HRRRv4 tended to handle it better. The disparity in complexity between the models and MRMS reflects the discrepancy in horizontal length scales resolved in each dataset, even though both were interpolated to the same grid before running MODE. Finally, it was found that both HRRR versions did not capture the observed distribution of near-max reflectivity value within objects. In particular, the models produce too many objects with a high near-peak reflectivity value, suggesting the HRRR intensifies convective storms too much and that the distribution of hydrometeors within modeled convective storms does not match that in the observations. But there may be other issues as well.

In an attempt to account for object count bias, a generalized matching procedure was used that selected F-O object pairs consisting of forecast and observation objects that are uniquely linked to each other (i.e., each object is paired to at most one object in the other dataset). The generalized matching procedure created a new set of object pairs in which the large objects were substantially overforecast in number, the aspect ratio and complexity distribution shapes were largely unchanged, and objects with low near-max reflectivity values tended to be predicted to have a larger reflectivity value and vice versa, in spite of the overwhelming overforecasting of the high near-max reflectivity objects. This generalized matching technique therefore revealed that high storm count biases are not the only source of deficiencies in reflectivity object forecasting from the HRRR. Notably, percentile-based thresholding (98th and 99th percentiles) was also examined. While object count biases decreased up to 25%, they remained above 1.0 throughout the forecast. The biases between HRRRv3 and HRRRv4 were reduced except for during the middle of the forecast when HRRRv4 retained a lower bias than HRRRv4. Beyond bias, percentile-based thresholding resulted in a dramatic reduction in differences between HRRRv3 and HRRRv4 in most metrics, in particular the OTS. Peculiarly, HRRRv4 object attribute distributions shifted toward those of HRRRv3 whereas the HRRRv3 distributions did not shift closer to the MRMS observations. These results raised questions that are beyond the scope of this paper and are left for future work.

Most scalar metrics evaluated herein suggest HRRRv4 forecast 6-h precipitation better than HRRRv3, although there were exceptions. While both versions overforecast the total number of precipitation objects, HRRRv4 had a less severe bias than HRRRv3, and HRRRv4 better forecast most object attribute distributions than HRRRv3. However, HRRRv4 precipitation objects had a higher mean centroid displacement than HRRRv3 and HRRRv4 had a more severe problem overforecasting objects with a high maximum precipitation amount. Subsequent analysis pointed to this problem arising from the largest precipitation objects, which directs model developers toward examination of where high precipitation amounts are coming from in large precipitation objects.

In general, it can be concluded that the changes made to the model physics, dynamics, and data assimilation led to distinct differences of behavior in HRRRv4 compared to HRRRv3, but the extent to which these changes led to improvements in these storm-specific fields is both statistically insignificant and varied depending on a variety of conditions (field, field threshold, time of day, forecast hour, location within the United States). However, this paper introduced topics and metrics for ways to compare two forecasting systems using an object-based approach. This work presents the next step in developing a rigorous approach to using object-based verification to assess NWP model performance. We will continue to expand this object-based approach into more areas of CAM-scale NWP that are not yet regularly assessed. This includes applying object-based verification to an ensemble such as HRRRE and implementing the time-dimension to assess not only spatial, but temporal errors and biases in CAM forecasts as well. Future work should also consider a morphological breakdown of reflectivity object-based verification by considering storm types such as cellular, linear, and bowing. We have taken initial strides to examine the morphology aspect, but the inherent subjectivity has so far hampered efforts in this venture. A possible way to compress this kind of object-based verification is to perform the analysis only on objects defined at the lowest threshold (a minimum to optimally capture individual objects), but classify the objects based on whether the *maximum value* within the object exceeds the given thresholds rather than classifying objects based on the area in which the field exceeds the fixed threshold (e.g., 35 dB*Z*) exceeding a minimum size (16 grid squares), as was done here. Finally, we anticipate applying this verification exercise to the upcoming replacement for HRRRv4, the Rapid-Refresh Forecast System (version 1), which is tentatively scheduled to replace HRRRv4 in late 2024 or early 2025.

Acknowledgments. Terra Ladwig of NOAA/GSL provided helpful support and suggestions on research trails to follow to improve the quality of this research. Curtis Alexander at NOAA/GSL provided an internal review of this manuscript. HRRR forecasts made use of several NOAA Research Development and High-Performance Computing System resources. We acknowledge the work done by the software developers and scientists at the Developmental Testbed Center for continuing to improve the MET software for use in this research. This work was supported by the CIRES–NOAA collaborative agreement Award NA17OAR4320101. The helpful comments from four anonymous reviewers improved the quality of this manuscript. We also acknowledge the comments from one particular reviewer who urged us to apply percentile-based thresholding for this work. That opened an avenue of research that we felt was beyond the scope of this paper, but is worthy of further consideration nonetheless.

Data availability statement. All forecast output verified herein were pulled from the NOAA RDHPCS HPSS, which itself is not publicly available (information available at https://rdhpcs.noaa.gov/). However, the data used herein are also located on Amazon Web Services (AWS) within various buckets maintained by NOAA. GRIB2 HRRR output, in particular, can be obtained publicly from the AWS HRRR bucket at https://noaa-hrrr-bdp-pds.s3.amazonaws.com/index.html. MRMS composite reflectivity fields are located on the AWS MRMS bucket at https://noaa-mrms-pds.s3.amazonaws.com/index.html#CONUS/MergedReflectivityQCComposite_00.50. Stage-IV precipitation observations were obtained from the Earth Observing System archive at https://data.eol.ucar.edu/dataset/21.093. The MODE code used here can be found at https://github.com/dtcenter/MET/tree/main_v9.0.

REFERENCES

Ahijevych, D., E. Gilleland, B. G. Brown, and E. E. Ebert, 2009: Application of spatial verification methods to idealized and NWP-gridded precipitation forecasts. *Wea. Forecasting*, **24**, 1485–1497, https://doi.org/10.1175/2009WAF2222298.1.

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.

——, T. G. Smirnova, E. P. James, L.-F. Lin, M. Hu, D. D. Turner, and S. He, 2022: Land-snow data assimilation including a moderately coupled initialization method applied to NWP. *J. Hydrometeor.*, **23**, 825–845, https://doi.org/10.1175/JHM-D-21-0198.1.

Britt, K. C., P. S. Skinner, P. L. Heinselman, and K. H. Knopfmeier, 2020: Effects of horizontal grid spacing and inflow environment on forecasts of cyclic mesocyclogenesis in NSSL's Warn-on-Forecast System (WoFS). *Wea. Forecasting*, **35**, 2423–2444, https://doi.org/10.1175/WAF-D-20-0094.1.

Chen, L., C. Liu, Y. Jung, P. Skinner, M. Xue, and R. Kong, 2022: Object-based verification of GSI EnKF and hybrid En3DVar radar data assimilation and convection-allowing forecasts within a Warn-on-Forecast framework. *Wea. Forecasting*, **37**, 639–658, https://doi.org/10.1175/WAF-D-20-0180.1.

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

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

Du, J., 2011: NCEP/EMC 4KM gridded data (GRIB) stage IV data, version 1.0. UCAR/NCAR–Earth Observing Laboratory, accessed 4 September 2020, https://doi.org/10.5065/D6PG1QDD.

Duda, J. D., and D. D. Turner, 2021: Large-scale 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.

Flora, M. L., P. S. Skinner, C. K. Potvin, A. E. Reinhart, T. A. Jones, N. Yussouf, and K. H. Knopfmeier, 2019: Object-based verification of short-term, storm-scale probabilistic mesocyclone guidance from an experimental Warn-on-Forecast System. *Wea. Forecasting*, **34**, 1721–1739, https://doi.org/10.1175/WAF-D-19-0094.1.

Gallo, B. T., and Coauthors, 2021: Exploring convection-allowing model evaluation strategies for severe local storms using the finite-volume cubed-sphere (FV3) model core. *Wea. Forecasting*, **36**, 3–19, https://doi.org/10.1175/WAF-D-20-0090.1.

Gilleland, E., 2020: Bootstrap methods for statistical inference. Part I: Comparative forecast verification for continuous variables. *J. Atmos. Oceanic Technol.*, **37**, 2117–2134, https://doi.org/10.1175/JTECH-D-20-0069.1.

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.

Guerra, J. E., P. S. Skinner, A. Clark, M. Flora, B. Matilla, K. Knopfmeier, and A. E. Reinhart, 2022: Quantification of NSSL Warn-on-Forecast System accuracy by storm age using object-based verification. *Wea. Forecasting*, **37**, 1973–1983, https://doi.org/10.1175/WAF-D-22-0043.1.

Hou, D., and Coauthors, 2014: Climatology-calibrated precipitation analysis at fine scales: Statistical adjustment of stage IV toward CPC gauge-based analysis. *J. Hydrometeor.*, **15**, 2542–2557, https://doi.org/10.1175/JHM-D-11-0140.1.

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.

Kalina, E. A., I. Jankov, T. Alcott, J. Olson, J. Beck, J. Berner, D. Dowell, and C. Alexander, 2021: A progress report on the development of the High-Resolution Rapid Refresh ensemble. *Wea. Forecasting*, **36**, 791–804, https://doi.org/10.1175/WAF-D-20-0098.1.

Skamarock, W. C., 2004: Evaluating mesoscale NWP models using kinetic energy spectra. *Mon. Wea. Rev.*, **132**, 3019–3032, https://doi.org/10.1175/MWR2830.1.

——, and Coauthors, 2019: A description of the Advanced Research WRF Model version 4. NCAR Tech. Note NCAR/TN-556+STR, 145 pp., https://doi.org/10.5065/1dfh-6p97.

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, 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.

Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, https://doi.org/10.1175/2008MWR2387.1.

Turner, D. D., and Coauthors, 2020: A verification approach used in developing the Rapid Refresh and other numerical weather prediction models. *J. Oper. Meteor.*, **8**, 39–53, https://doi.org/10.15191/nwajom.2020.0803.

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. J., and W. C. Skamarock, 2020: An implicit-explicit vertical transport scheme for convection allowing models. *Mon. Wea. Rev.*, **148**, 3893–3910, https://doi.org/10.1175/MWR-D-20-0055.1.