Large-Sample Application of Radar Reflectivity Object-Based Verification to Evaluate HRRR Warm-Season Forecasts

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] *NOAA/Global Systems Laboratory, Boulder, Colorado*

(Manuscript received 28 October 2020, in final form 4 March 2021)

ABSTRACT: The Method of Object-based Diagnostic Evaluation (MODE) is used to perform an object-based verification of approximately 1400 forecasts of composite reflectivity from the operational HRRR during April–September 2019. In this study, MODE is configured to prioritize deep, moist convective storm cells typical of those that produce severe weather across the central and eastern United States during the warm season. In particular, attributes related to distance and size are given the greatest attribute weights for computing interest in MODE. HRRR tends to overforecast all objects, but substantially overforecasts both small objects at low-reflectivity thresholds and large objects at high-reflectivity thresholds. HRRR tends to either underforecast objects in the southern and central plains or has a correct frequency bias there, whereas it overforecasts objects across the southern and eastern United States. Attribute comparisons reveal the inability of the HRRR to fully resolve convective-scale features and the impact of data assimilation and loss of skill during the initial hours of the forecasts. Scalar metrics are defined and computed based on MODE output, chiefly relying on the interest value. The object-based threat score (OTS), in particular, reveals similar performance of HRRR forecasts as does the Heidke skill score, but with differing magnitudes, suggesting value in adopting an object-based approach to forecast verification. The typical distance between centroids of objects is also analyzed and shows gradual degradation with increasing forecast length.

SIGNIFICANCE STATEMENT: Improving weather forecast models requires determining where the model does well and where it does not. Gridpoint-based methods for assessing model forecasts have known shortfalls when applied to high-resolution models that can forecast individual thunderstorms. We present an object-based verification procedure that focuses on identifying actual meteorological features such as thunderstorms instead of gridpoint-by-gridpoint comparison between forecasts and verifying truth. This article reveals some of the information ascertained from this assessment and illustrates the enhancement of information obtained from object-based verification to gridpoint-based assessment.

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

1. Background

Object-based verification is seeing increasing use in the convection-allowing model (CAM) forecast community, especially for fields dominated by contiguous spatial features that are small compared to the size of the domain—hereafter referred to as feature-based fields, and akin to the nomenclature used in Ahijevych et al. (2009) for verification methods designed to be applied to such fields. One major benefit to using object-based methods is to avoid the double penalty problem common in gridpoint-based verification metrics for CAM-scale forecasts. Another benefit is to isolate the performance of specific object attributes, such as the size of convective storms or biases in storm location. Herein we apply object-based techniques to 3-km grid spacing High-Resolution Rapid Refresh (HRRR; Alexander et al. 2010; Benjamin et al. 2016) forecasts of composite reflectivity.

Done et al. (2004) used a primitive form of object-based verification to assess the performance of 4-km forecasts of mesoscale convective systems (MCS) produced by the Weather Research and Forecasting (WRF) Model (Skamarock et al. 2008), in particular, whether the WRF Model could even capture the observed MCS in a real-world case. Pinto et al. (2015) applied a more formalized method to assess attributes such as size, location, and orientation of MCSs forecast by the HRRR. Duda and Gallus (2013) determined the location accuracy of initiating convection and upscale development of MCSs in 3-km WRF forecasts. Hartung et al. (2011) evaluated the impact of new boundary layer thermodynamic and kinematic profilers in an observation system simulation experiment. Johnson and Wang (2012) generated object-based probabilistic forecasts of 1-h accumulated precipitation, later refined through the OBPROB technique (Johnson et al. 2020). Recently, verification of convective storms from 3-km CAM forecasts using various advanced radar data assimilation (DA) methods were compared to both traditional and neighborhood-based verification metrics (Duda et al. 2019), where it was found that object-based metrics can provide a comparable, but unique, assessment of model performance in object-based space compared to in gridpoint space.

A few specific object-based verification methods have been developed. These include the structure-amplitude-location technique (Wernli et al. 2008, 2009), the Contiguous Rain

Corresponding author: Jeffrey D. Duda, jeffduda319@gmail.com

DOI: 10.1175/WAF-D-20-0203.1

Area (Ebert and Gallus 2009) method, and other more refined and/or eclectic approaches used to verify only the specific fields deemed necessary (e.g., Skinner et al. 2016, 2018; Stratman and Brewster 2017; Jones et al. 2018; Flora et al. 2019; Potvin et al. 2019). Davis et al. (2006) introduced the Method of Object-based Diagnostic Evaluation (MODE), which is used herein.

MODE is a complex method of object-based verification, but the added detail enables output of a broad set of object attributes as well as more complex metrics. MODE applies fuzzy logic to objects classified from the gridded input datasets (usually a forecast and an observation set) and attempts to match them. The most unique output from MODE is a scalar interest value between each pair of forecast and observation objects. Ranging between 0.0 and 1.0, it is a normalized value that can be understood as the probability that the compared objects in the two datasets are the same object, where 1.0 implies an effectively perfect correspondence between the two objects. MODE has seen increasing use in verification of high-resolution forecasts (e.g., Duda and Gallus 2013; Johnson and Wang 2013; Cai and Dumais 2015; Moser et al. 2015; Bytheway et al. 2017; Griffin et al. 2017, 2020; Schwartz et al. 2017; Adams-Selin et al. 2019; Duda et al. 2019; Gallus et al. 2019; Johnson et al. 2020; Squitieri and Gallus 2020). MODE is used herein to assess a variety of characteristics of the reflectivity field output from the HRRR model. The preference for MODE is its connection with the operational forecasting community; MODE is a component of the Model Evaluation Tools software package (METplus; Halley Gotway et al. 2018; Brown et al. 2021) developed by the Developmental Testbed Center at the National Center for Atmospheric Research for use both by the research and operational forecasting community. The open-source and continuously supported nature of METplus makes it an attractive option for performing object-based verification. Despite the extensive output from MODE, there is presently no published procedure in METplus for combining the raw outputs into useful metrics. Some metrics not previously reported in MODE literature are introduced herein and evaluated for their utility in assessing forecast performance.

The HRRR model is a 3-km grid spacing forecast system that uses the Advanced Research version of the WRF (ARW) dynamical core. It is informally nested within the larger-domain 13-km Rapid Refresh model (Benjamin et al. 2016), which provides the initial and lateral boundary forcing. The HRRR is unique among operational CAMs in that new forecasts are initialized every hour; the rapidly updating nature of the HRRR makes it attractive to forecasters for short-term weather forecasting, especially for severe convective storms and extreme rainfall, among other hazards. Furthermore, the HRRR is often judged favorably among other operational and experimental deterministic CAM forecasts by researchers, model developers, and forecasters alike perennially at NOAA/Hazardous Weather Testbed Spring Forecasting Experiments [e.g., Fig. 36 of Clark et al. (2016); cf. Figs. 28, 30, 32, 34, 36, and 38 of Clark et al. (2017); Fig. 36 of Clark et al. (2018); Fig. 27 of Clark et al. (2019); Figs. 42–44 of Clark et al. (2020); Roberts et al. (2020)]. The HRRR became formally operational on 30 September 2014 and has since undergone three major operational upgrades: version 2 on 23 August 2016, version 3 on 12 July 2018, and version 4 on 2 December 2020. Operational forecasts from HRRRv3 (hereafter HRRR) are verified herein.

The two goals of this paper are as follows: 1) illustrate the underutilized abilities of MODE, in particular, to provide information that standard verification practices do not and that are helpful in assessing model performance; and 2) analyze the ability of HRRR to simulate convective storms (as represented via computed reflectivity fields) so that it can be improved.

2. Methodology

a. Data selection

One goal of this paper is to assess the ability of HRRR forecasts to capture the object-based details of convective storms. Therefore, the cases selected for verification occurred during the 2019 warm season (1 April–30 September). Only forecasts initialized every three hours (i.e., 0000, 0300, . . ., 1800, 2100 UTC) are selected, both to save computational storage space and as a compromise between large-sample collection and reduction of sample dependence. This selection criteria resulted in over 1400 cases that are verified. HRRR forecasts run to 18 h except for at 0000, 0600, 1200, and 1800 UTC, at which they run to 36 h. Only the first 24 forecast hours (at 1-h frequency) are verified regardless of total forecast length.

The verification domain includes the eastern two-thirds of the CONUS, roughly all land east of the Interstate 25 corridor and including parts of Montana (e.g., Fig. 7 includes an outline of the verification domain). There are limited areas east of Interstate 25 that have poor radar coverage and/or complex terrain, but those areas are nonetheless included in the verification domain. Areas generally 100 km or less offshore of the Gulf of Mexico and Atlantic Ocean coasts are also included where the radar coverage is adequate.

The composite radar reflectivity field—calculated as the column maximum of simulated reflectivity from the HRRR—is verified. The composite reflectivity product from the multiradar/multisensor project (MRMS; Smith et al. 2016; Zhang et al. 2016) is used as the observations. MRMS data are constructed from a regional mosaic of reflectivity values from the WSR-88D radars around the United States and Canada and therefore are dependent on the quality and coverage of the radar data collected at each radar site, which varies by site and time. MRMS data from the closest available time to the top of each hour are used and assumed to be uniformly valid at exactly the top of the hour. This choice can introduce additional artificial error into the verification, but such error should be negligible compared to the time scale at which the HRRR forecasts are verified. Since the MRMS data are on a 0.01° latitude/longitude grid, which is substantially finer than the HRRR grid, MRMS data are interpolated bilinearly to the HRRR grid prior to verification with the understanding that the interpolation does not necessarily remove all features on the finer scale of the MRMS grid, which can introduce artifacts in the results.

TABLE 1. Description of object attributes in MODE.

Attribute name	Description
Area	Object area
Length/width/aspect ratio	Dimensions of the smallest inscribing rectangle for the object
Axis angle	Angle of long axis of smallest inscribing rectangle relative to longest grid dimension (e.g., 0 corresponds to west–east orientation for common grids)
Volume (mass)	Sum of field values within object
Complexity	Ratio of the area of the convex hull *not* in the object to the area of the convex hull. The convex hull is the smallest convex polygon that inscribes the object. It can be thought of as placing a rubber band around the object.
Curvature	Mathematical definition of radius of curvature
Intensity percentiles	10th, 25th, 50th, 75th, 90th, plus one additional custom percentile value (95th used)
Centroid	Center of mass

b. *MODE settings*

MODE is configured to capture details relevant to convective storms. Therefore, the following settings are used. The convolution radius is set to 1 grid square; little filtering/smoothing is selected to maintain as many near-gridscale details as possible in the composite reflectivity field. Reflectivity values of 25, 30, 35, and 40 dBZ are used as the thresholds; these values are typical within convective storms while also accounting for the slight reduction in peak values within local reflectivity maxima due to the convolution smoothing. The sample size at 40 dBZ is small relative to the lower reflectivity thresholds. It is noted that the imperfect correspondence between the HRRR microphysics and reflectivity diagnostic and the measured reflectivity in the MRMS data can influence the classification of objects, especially the count and size. The methodology herein differs from that in, e.g., Skinner et al. (2018) and Potvin et al. (2019), who used percentile thresholds to eliminate bias. However, there is no perfect strategy to account for a such a biased diagnostic field computed from forecasted microphysical fields when performing object-based verification; using a percentile thresholding technique could introduce other undesirable artifacts into the dataset to hamper analysis. Therefore, this caveat underlies the analysis performed herein.

We compute the interest value for each forecast–observation object pair by weighting the attributes such that location is most important with various shape attributes having moderate importance (descriptions of object attributes are in Table 1). Attributes such as curvature ratio, orientation angle difference, and aspect ratio difference are considered irrelevant in this context and are given zero weight. MODE calculates an interest value between all pairs of forecast and observation objects that are within 500 km (user-controllable) of each other; while this setting is primarily used to save computational time, it is an appropriate decision since objects separated by a significant distance are almost certainly not characteristic of a good forecast. Therefore, any object pairs that have a larger centroid distance than that are automatically assigned an interest value of 0.0. An interest value of 0.0 means the two objects have no meaningful correspondence, whereas an interest of 1.0 means the two objects have as much correspondence with each other as is deemed necessary to consider the forecast effectively perfect. These settings represent an open-ended verification philosophy, since the setting of interest maps and attribute weights can be chosen to represent either a more lenient validation in which large errors in object location, shape, size, or intensity can still result in very high (or perfect) interest values, or a stricter state in which object attributes must be approximately perfectly matched to obtain an interest value of 1.0. As forecast systems improve, object-based verification from MODE can be recomputed using stricter interest maps to enforce a higher standard for declaring a forecast "good."

The selection of weights (Table 2) is somewhat arbitrary as the weights represent no explicit physical meaning but instead are purely statistical. It cannot be claimed that this is necessarily the optimal set of weights to use, as even the meaning of "optimal" is ill-defined, and each forecast user will prioritize object attributes individually. It is possible to optimize the weights based on a specific metric, but this work is designed to be broad reaching and introductory, and therefore no single metric is deemed better than all others. For uses of MODE in which different forecasting systems are being compared, the selection of weights is arguably unimportant, as all systems would be compared equally. It is also important to note that the settings for MODE indicate a high level of dimensionality regarding object-based verification; in fact, there are innumerable ways in which forecasts and observations can be compared. Therefore, there is some innate uncertainty in the interest values. However, reasonable choices of interest maps and attribute weights should minimize uncertainty. Quantifying the uncertainty associated with the choice of the interest values is beyond the scope of this paper.

c. *Types of verification graphics/metrics and description of each*

The possible set of verification diagrams or metrics that can be obtained from MODE output is staggeringly large, too large for each to be identified and compared. Those invoking object-based verification should approach the exercise with specific attributes in mind to assess and compare to limit information overload.

Here, we present four broad categories of verification metrics using MODE output. They include comparison of object attribute distributions in multiple dimensions, object counts,

TABLE 2. Interest weights for object pair attribute comparisons.

Attribute	Weight	Description
Centroid distance	5.0	Distance between centroids of forecast and observation object
Boundary distance	4.0	Closest distance between object boundaries (not using convex hull)
Convex hull distance	0.0	Closest distance between object convex hulls
Angle difference	0.0	Difference between object orientation angles
Aspect ratio difference	0.0	Linear difference in object aspect ratio
Area ratio	4.0	Ratio of area of forecast object to that of observation object
Consumption ratio	2.0	Fraction of smaller object overlapping with (consumed by) larger object
Curvature ratio	0.0	Ratio of curvature of forecast object to that of observation object
Complexity ratio	0.5	Ratio of complexity of forecast object to that of observation object
Intensity percentile ratio	3.5	Ratio of the 95th percentile value of the forecast object to that of the observation object

single valued metrics, and performance diagrams. In particular, area, complexity, aspect ratio, and intensity percentile (for composite reflectivity, the 95th percentile value within each object was selected to represent a near-max value within each object; Table 1) attribute distributions are directly compared. The continuous ranked probability score (CRPS; Hersbach 2000) offers a convenient scalar metric by which to compress probability distribution data into a single-valued measure of the accuracy with which object attribute distributions are forecast. The CRPS is formulated as

$$\mathrm{CRPS} = \int_{-\infty}^{\infty} [\mathrm{CDF}_f(x) - \mathrm{CDF}_o(x)]^2 \, dx, \tag{1}$$

where CDF_f and CDF_o are the cumulative distribution functions of the given object attribute from the HRRR and MRMS data, respectively, along the range of values of the attribute x. The frequency bias diagnostic (F/O where F represents the number of forecast objects identified, and similar for the observations O) is heavily relied upon to compare object counts.

Three additional scalar metrics rely on either the interest value or one specific object attribute (e.g., object area or centroid location). They include the object-based threat score (OTS; Johnson and Wang 2013), the median of maximum interest (MMI; Davis et al. 2006), and the mean distance between object centroids. OTS is a normalized threat score that is a function only of object-pair interest and area, formulated as

$$\mathrm{OTS} = \frac{1}{A_f + A_o} \sum_p I^p (a_p^f + a_p^o), \tag{2}$$

where A_f and A_o represent the total area of all forecast and observation objects, respectively; I^p is the interest value between forecast-observation object pair p; and a_p^f and a_p^o are the areas of the forecast and observation objects in pair p, respectively. The sum is calculated over interest-ranked unique object pairs (hereafter, the ''generalized'' method), meaning that the interest value between all pairs in a dataset are first ranked in order of decreasing interest value. Starting from the pair with the highest interest, all subsequent pairs that include either that forecast or observation object are removed from further consideration in the sum. The sum continues until no pairs remain. This computation methodology ensures that object frequency count bias will not artificially inflate the OTS; in a forecast where there is a severe discrepancy between the number of forecast and observation objects, most of the objects in the dataset with the larger plurality will not be included in the sum, which will decrease the numerator (but not the denominator) and result in a lower OTS. Since the interest value is limited to the range [0.0, 1.0], OTS is also limited to this range and it is a positively oriented score. An OTS of 1.0 can only be achieved if there is both a 1:1 correspondence between the number of forecast and observation objects as well as an effectively perfect correspondence between the forecast and observation object in each pair. Effectively, OTS = 1.0 represents a perfect forecast.

The classical computation of MMI is described as follows (also illustrated in Davis et al. 2009, their Fig. 1). A two-dimensional table with size $N_f \times N_o$ (the number of forecast and observation objects, respectively) is constructed. The value in each cell is the interest value for that object pair. A one-dimensional vector consisting of the maximum interest value in each row and column of the interest table is then constructed; the MMI is the median of that vector. This calculation is rather arbitrary, and a small number of object pairs can exert undue influence on the final value. Therefore, an alternative calculation is tested here. It uses the generalized method to create the one-dimensional vector of interest values. Since this strategy tends to remove the many 0.0 interest values, as 0.0 interest values with a multiplicity equal to the difference in the number of objects between the forecast and observations are appended to the interest vector to account for object count biases. The alternative MMI is then the median of that vector. Therefore, this alternate MMI cannot be artificially inflated by an object count bias.

Performance diagrams (Roebber 2009) are constructed using the matching feature of MODE. Object pairs with an interest value exceeding 0.70 (default, but user-controllable) are considered ''matched'' (note: the matching behavior of MODE allows for an object in one dataset to be matched to more than one object in the other dataset). In the context of a 2×2 contingency table like those used for verification of dichotomous events, a match is equivalent to a ''hit,'' whereas any forecast object that is not matched to an observation object is considered a ''false alarm,'' and any observation object that is not matched to a forecast object is considered a ''miss.'' It is

impossible to define "correct null" in this context, but the performance diagram does not use this contingency.

Finally, various measures of the central tendency of the distribution of centroid displacement errors are calculated. Several sets of objects, classified morphologically or whether matched or not, are considered. The generalized method is also included for comparison. For this object pair attribute there is no finite maximum distance corresponding to a useless forecast. Therefore, no artificial values are appended to the vector of centroid distances from which the mean or median is computed, which means the generalized centroid distance formula is sensitive to object count bias. Centroid distances are also decomposed into west–east and south–north components to determine any directional bias in displacement errors.

3. Results

a. Object attributes

First, we examine how well HRRR forecasts replicated specific object attributes of the composite reflectivity field. The overall shape of the object distributions was invariant with respect to forecast hour aside from a shift between forecast hours 0 and 1 (representing the shift associated with dynamically imbalanced fields following DA and forecast initialization). Therefore, only the 24-h aggregated distributions are discussed. HRRR forecasts mimic the overall shape of the observation distribution of object area (Fig. 1a). However, there are discrepancies between HRRR forecasts and MRMS observations in the proportion of small and large objects, with medium-sized objects being particularly well replicated. This tendency is more prominent at the lower reflectivity thresholds, but not at the higher reflectivity thresholds (not shown). Regarding shape-specific attributes, HRRR forecasts contained objects that were too circular—there was a noticeable shift toward higher aspect ratio (more circular/square shaped objects than oblong objects) relative to the MRMS observations (Fig. 1b). There is an anomalous number of objects in the [0.80, 0.85) bin for the HRRR forecasts, which could be an artifact unique to the HRRR system (perhaps a resolution dependency) since it is present at all reflectivity thresholds but not at all present in the MRMS distributions. Similarly, there is a noticeable shift in the distribution of object complexity toward lower values in the HRRR forecasts (Fig. 1c). This shift likely reflects underresolved convective-scale features that survive the convolution filtering performed in MODE but are captured in the relatively finer scales obtained by S-band radars that comprise the MRMS data (and that remain even after horizontal interpolation). Indeed, the complexity distribution errors are more significant with the smaller objects (those covering 2250 km^2 or less; Fig. 2a) compared to those from medium and larger sized objects (larger than 20 000 km^2; Fig. 2c). But there are approximately one and two orders of magnitude more small objects than medium and large objects, respectively, so the small objects dominate the total distribution.

The impact of the DA is dramatically visible from comparison of f00 and f01 distributions (Fig. 3). This change almost

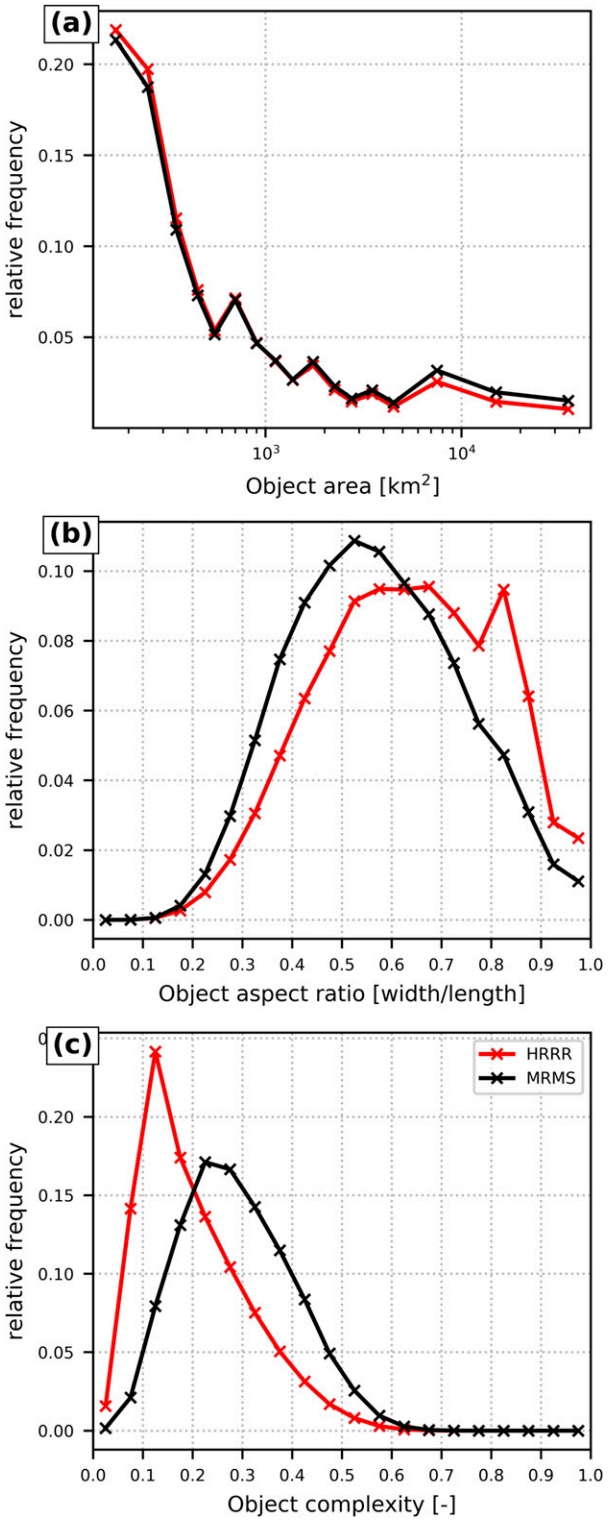

FIG. 1. Distributions of HRRR and MRMS composite reflectivity object (a) area, (b) aspect ratio, and (c) complexity, aggregated over all forecast hours for the 25-dBZ threshold.

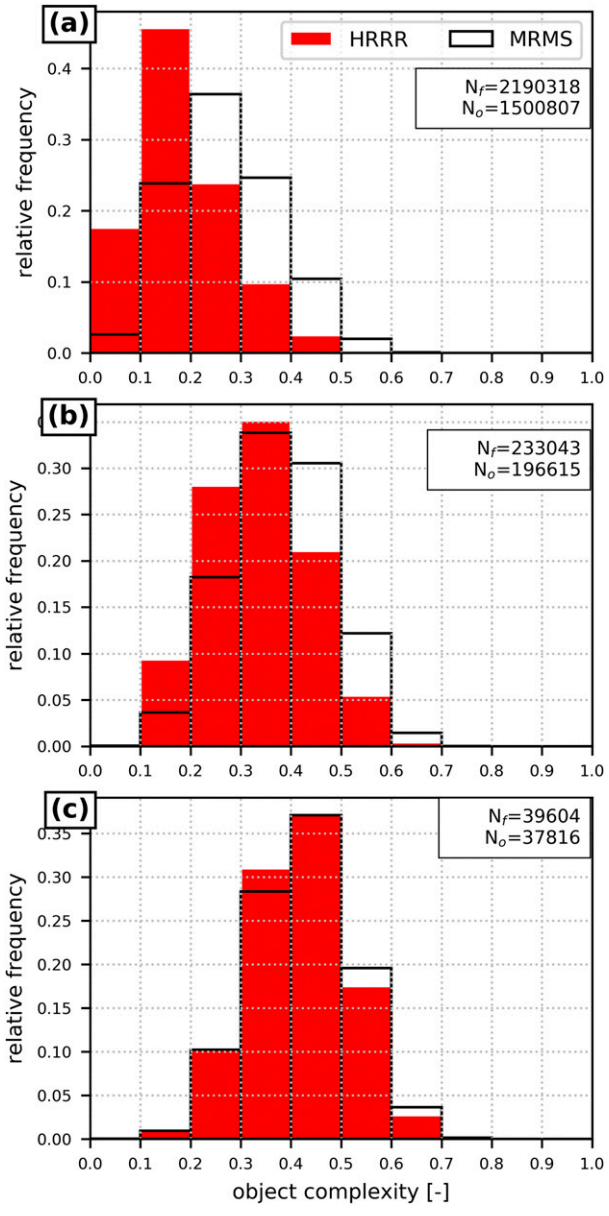

FIG. 2. Object complexity at 25 dBZ for (a) small (area $\leq 2250\,\text{km}^2$), (b) medium, and (c) large (area $> 20\,000\,\text{km}^2$) objects. Data aggregated across all forecast hours. Solid black lines are used to denote the MRMS distribution, and red filled bars are used for the HRRR data. The number of objects in the HRRR forecasts (f) and MRMS observations (o) are provided in the insets of each panel.

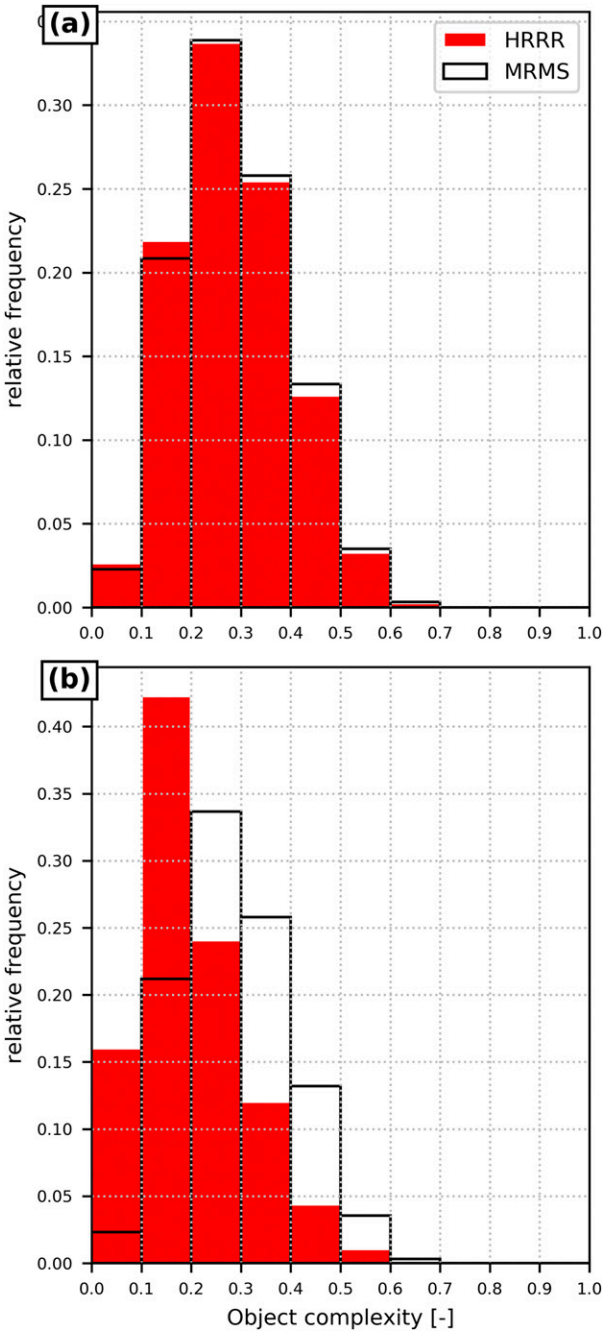

FIG. 3. Object complexity distribution at 25 dBZ at forecast hour (a) 0 and (b) 1 for HRRR (red bars) forecasts and MRMS (open bars with black edges) observations.

certainly illustrates the impact of the cloud analysis (Hu et al. 2006)—at forecast initialization the reflectivity field very closely resembles that of the observation reflectivity field. Because the cloud analysis does not update related fields such as temperature or wind, the initial condition of the forecast contains a dynamically unbalanced state. This imbalance includes a lack of buoyancy or updraft velocity to sustain newly assimilated convective storms. Therefore, some of the newly assimilated reflectivity dissipates due to hydrometeors falling out. As Fig. 4a illustrates, there is little additional degradation in the quality of the object attribute distributions for longer forecast hours. This conclusion is further supported by the scalar object-based verification metrics in section 3c.

In contrast to object attribute distributions by forecast hour, the shape and position of most attribute distributions did shift with respect to time of day. The time series of CRPS for the

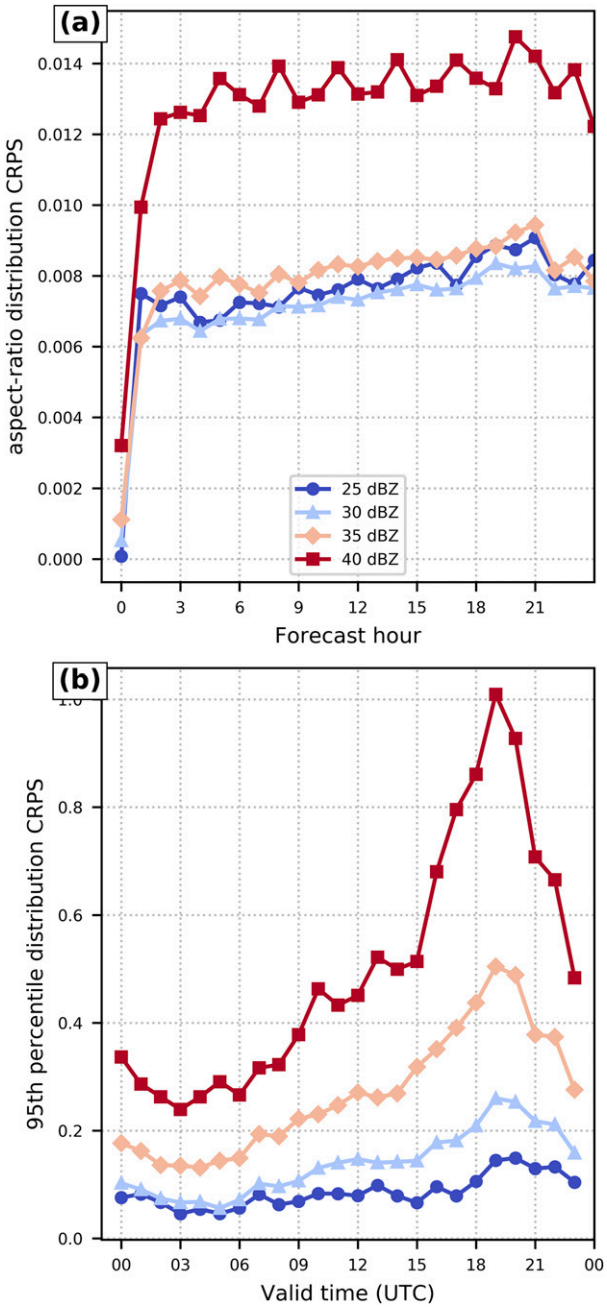

FIG. 4. CRPS at various reflectivity thresholds of (a) object aspect ratio as a function of forecast hour and (b) object 95th percentile of composite reflectivity as a function of the time of day, both for objects defined at the 40-dBZ threshold. Note: sunrise and sunset are approximately at 1200 and 0100 UTC, respectively, over the analysis domain for this time period.

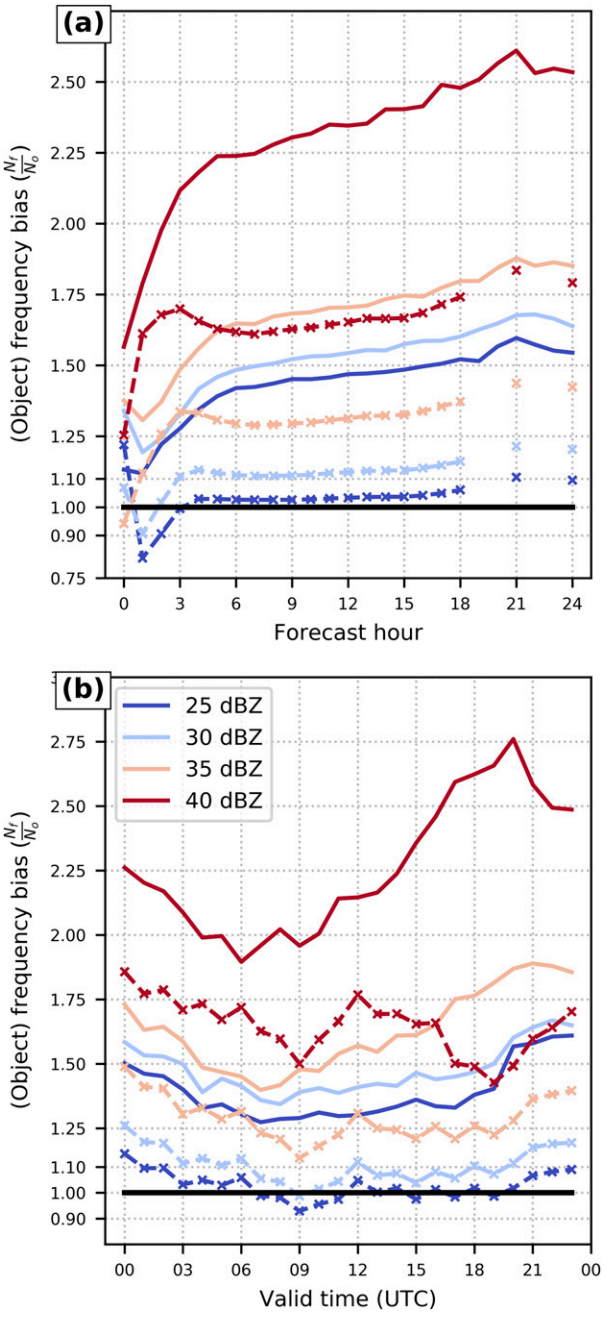

FIG. 5. Object (solid) and gridpoint (dashed with "X" marks) frequency bias for different reflectivity thresholds as a function of (a) forecast lead time and (b) time of day.

95th percentile of reflectivity within each object (Fig. 4b) illustrates a trend exhibited by all attributes (except for object area, not shown) in that the distributions were predicted with poorer quality (higher CRPS) during the afternoon through early evening. As discussed in section 3b, an overprediction of new, small, and high-magnitude reflectivity objects (i.e., convective storms) seems likely behind this change in performance.

b. *Object frequency bias*

HRRR forecasts exhibit a noticeable overforecasting bias in terms of the number of reflectivity objects present which increases with forecast lead time (Fig. 5a), and which is only somewhat consistent with the gridpoint-based frequency biases

[computed using the Model Analysis Tool Suite (MATS); Turner et al. 2020]. At all but the 40-dBZ threshold, object-based biases at initialization are close to 1.25 and steadily increase into the 1.5–1.9 range by forecast hour 24; at 40 dBZ, however, a bias of 1.55 at initialization is followed by a short, rapid increase and then a slower increase after forecast hour 3 or so, to an object count bias well in excess of 2.0 (exceeding 2.5 by forecast hour 20 or so). This result implies the HRRR tends to significantly over forecast the number of high-reflectivity objects.

MODE enables a more detailed examination of what aspects of HRRR forecasts contribute to the overforecasting of objects. At the low reflectivity thresholds the smallest objects contribute the most to the object count bias. Whereas objects smaller than 2000 km^2 have biases of generally 1.4–1.5 (Fig. 6a), the largest objects (10 000 km^2 and larger) are either forecast in appropriate numbers or have an underforecast bias. At 35 and 40 dBZ, the distributions shift from being small-object weighted to large-object weighted. While the smaller objects maintain object count biases of 1.5–2.0 (Fig. 6b), larger objects attain biases well above 2.0 to around 10.0 at 40 dBZ. The reason these larger objects do not contribute overall to a high total object count bias is the very small number of objects of that size (Fig. 1a). However, since the object count bias value is merely the ratio of the HRRR object count in that size bin to the MRMS object count in that same size bin, it is possible that otherwise-properly forecasted storms that are the wrong size due to the reflectivity diagnostic are included in a given size bin, causing the bias value to not strictly be dependent on the number of meteorological entities in that size bin (this behavior would also manifest as low biases in other bins).

Gridpoint frequency biases, on the other hand, are closer to 1.0 at a given threshold than the corresponding object frequency bias. Also, at the 25- and 30-dBZ reflectivity thresholds, the decrease in gridpoint frequency bias between f00 and f01 is more apparent than that in the object frequency biases. Finally, the gridpoint frequency biases increase more slowly with forecast hour than do the object frequency biases. The addition of the object area distribution to this analysis provides information about why the gridpoint frequency bias is generally lower than its object-based counterpart: the HRRR forecasts too many small storms, but each individual storm contributes very little to the gridpoint count, so the gridpoint frequency bias is not as large as the object frequency bias. This comparison provides a powerful illustration of the information that can be obtained from supplementing traditional verification metrics with object-based verification techniques.

The overall overforecasting of reflectivity objects is not spatially homogeneous. At low reflectivity thresholds, objects tend to be forecast at appropriate frequencies over the plains region, especially the southern plains and portions of western South Dakota and Nebraska (Fig. 7). Consistent with Fig. 6, the largest objects are substantially underforecast in this area as well as in the Mississippi River Valley (Fig. 8). On the other hand, objects are more substantially overforecast over the southern United States and parts of the eastern U.S. coast as well as on the High Plains of Montana and Wyoming. Large objects are overforecast over the northern

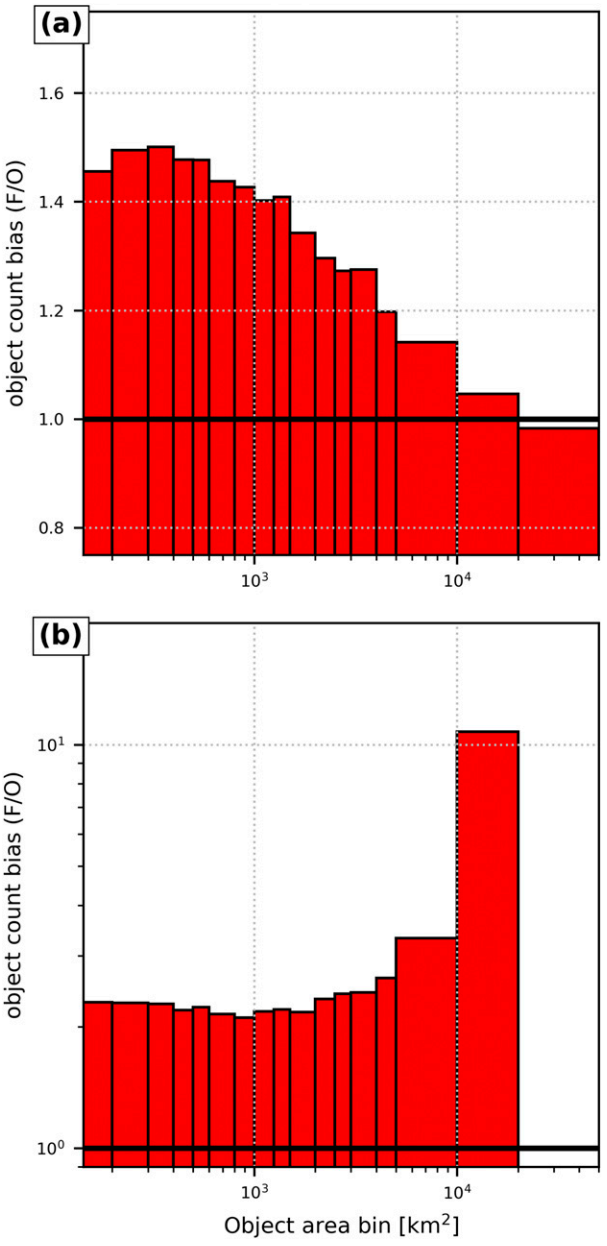

FIG. 6. Object frequency bias as a function of object area aggregated over all forecast hours for objects classified at (a) 25 and (b) 40 dBZ.

High Plains as well, and over the Appalachian Mountains. Radar coverage—especially in the vertical—is poor in some of these areas, especially along the Montana–North Dakota border, eastern Wyoming, and over a portion of Lake Superior. Although the MRMS project incorporates data from most Canadian radars, some of them, especially from eastern Saskatchewan through southern Ontario, are C-band and may suffer from limited range, especially over north-central Lake Superior (Zhang et al. 2016). Despite the limited low-level coverage, convective storms exceeding roughly 10 km in height, which includes vigorous thunderstorms with strong updrafts,

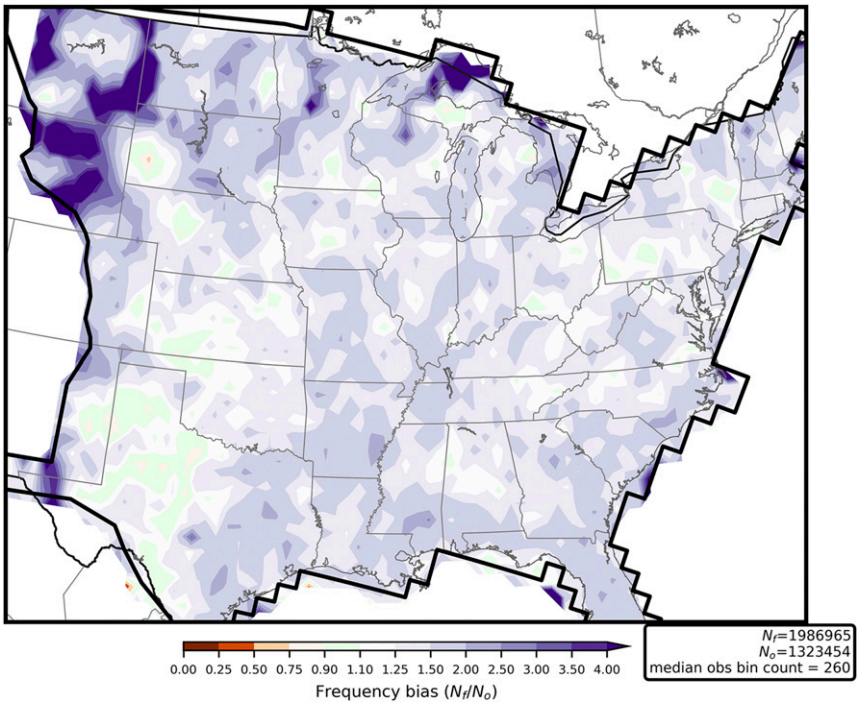

FIG. 7. Horizontal distribution of object frequency bias at 30 dBZ aggregated over all 24 forecast hours. The verification domain is outlined in thick black.

should still be detected; however, vertically limited precipitating systems will likely be missed. Thus, while there is insufficient information to conclusively determine that objects are truly overforecast in these areas, the biases from nearby regions suggest there may still be a legitimate signal. The increase in object count bias with increasing threshold is universal; the high-reflectivity objects are severely overforecast at nearly all locations in the verification domain (not shown).

There is a diurnal cycle to the object count bias featuring a broad maximum between 1800 and 2100 UTC and at a minimum between 0600 and 0900 UTC (Fig. 5b). The former is around the peak of convective activity overall (not shown) in the United States in the late afternoon and early evening as convective instability and PBL vertical motions are at their maxima. The latter is during the overnight period when convective instability is substantially reduced, although the convective minimum tends to occur slightly later (between 1200 and 1500 UTC, but the bias values do not change much between about 0300 and 1200 UTC). The spatial pattern of object count biases is also heterogeneous with some diurnal signal. The lowest count biases are found over the High Plains regardless of reflectivity threshold, where some underforecasting is evident during the 0600–1800 UTC period of reduced convective activity at all but the 40-dBZ threshold (not shown). There is a small-to-negligible frequency bias over parts of the mid-Atlantic region and the upper Midwest during that time at the 25- and 30-dBZ thresholds, but objects are overforecast generally everywhere else (not shown). The overforecasting is more serious at the 35- and 40-dBZ thresholds everywhere except for over the southern plains, where object counts are accurately forecast for the most part (Fig. 9). The overforecasting is particularly severe over portions of the eastern CONUS as well as the High Plains of Colorado and Wyoming.

There is some correlation between the spatial pattern of overforecasting and the more complex/elevated terrain of the Appalachian Mountains in the east, suggesting a potential for the HRRR to be too aggressive with terrain-following low-level wind flow to force deep convective storms. The object frequency bias for small objects at 40 dBZ is also quite high in these same areas, as well as across the upper Midwest (Fig. 10). There is spatial correlation in the western areas of the verification domain as well; the sloping terrain of the Great Plains may be smoother than that in the Appalachians, but the gradual ascent from east to west is notorious for channeling orographically forced upslope flow to force deep convective storms in the spring and summer, so the same issue could still be at play. There are also locally higher frequency biases for small objects over the Raton Mesa area of southeast Colorado and northeast New Mexico, as well as near the Guadalupe Mountains in southeast New Mexico and far western Texas, all areas of locally complex terrain. This investigation reveals that a deeper look at how the HRRR handles low-level flow in the presence of complex terrain is warranted.

The overall significant count bias for large objects at high reflectivity thresholds is intriguing. It could be a diagnostics issue related to the computation of composite reflectivity output from the Thompson microphysics scheme, or an issue within the Thompson MP scheme itself. The version of the scheme used in the HRRR includes a term for wet snow which

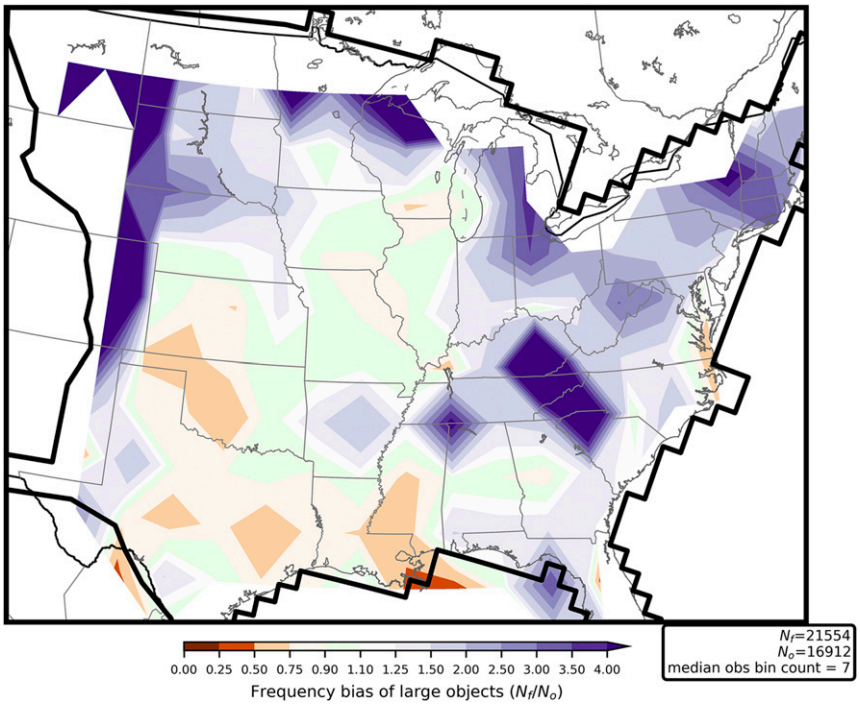

FIG. 8. As in Fig. 7, but only objects larger than 20 000 km^2 are considered.

contributes to higher reflectivity values when present. While the HRRR can struggle to adequately depict stratiform precipitation in regions associated with squall lines and mesoscale convective systems and complexes, if it does produce a stratiform region with wet snow falling through the melting level, it could result in an expansive area of increased reflectivity compared to what might be seen from observed radar in areas where bright banding occurs.

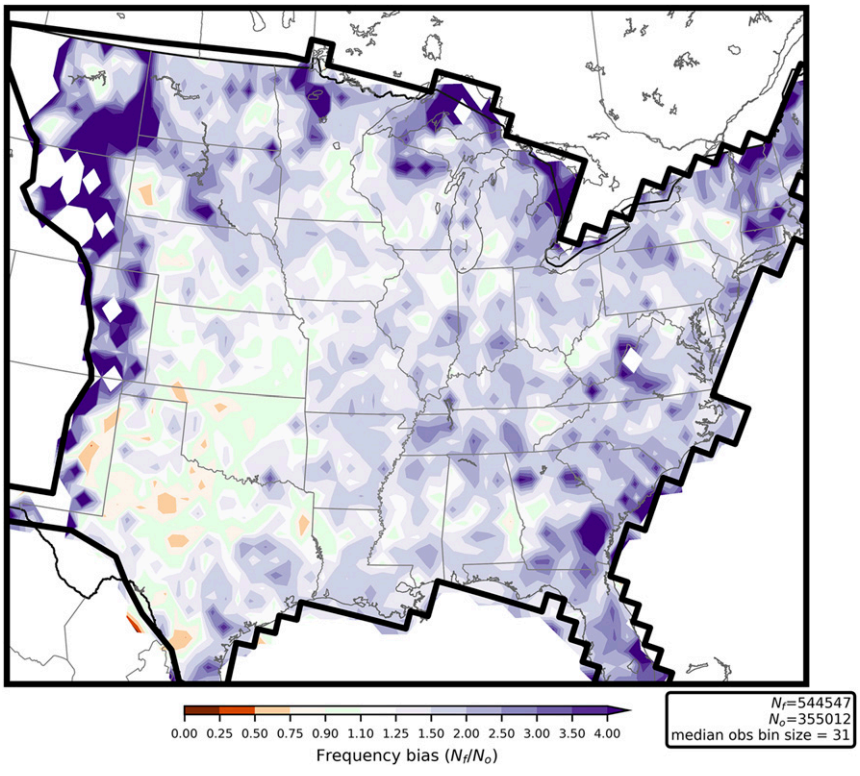

FIG. 9. As in Fig. 7, but at 35 dBZ and for forecasts valid between 0600 and 1700 UTC.

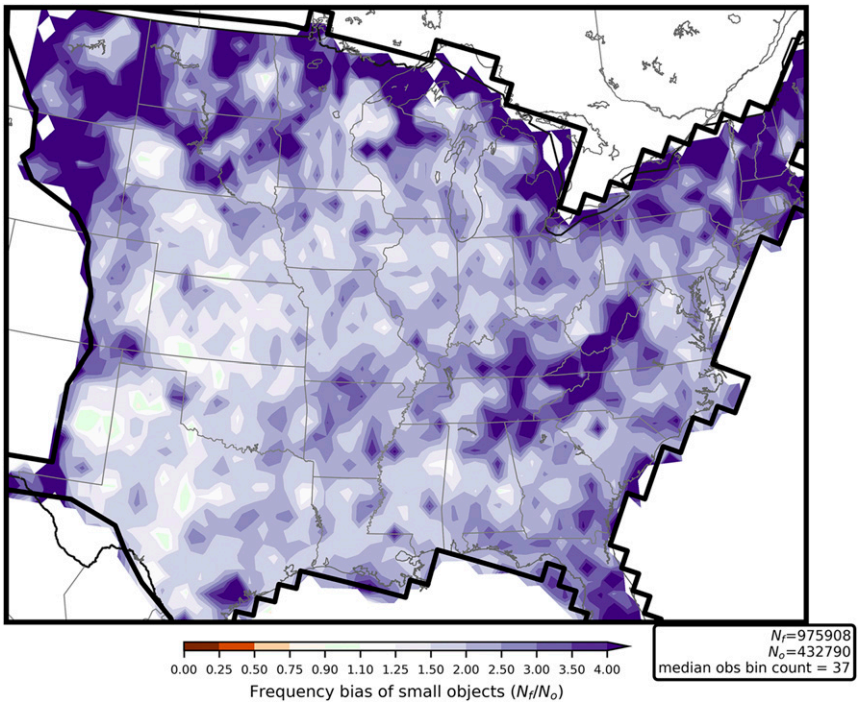

FIG. 10. As in Fig. 7, but at 40 dBZ and only objects smaller than 2250 km^2 are considered.

c. Scalar metrics

Examination of OTS as a function of forecast lead time [Eq. (2), Fig. 11a] shows that this metric obeys common trends in gridpoint verification metrics for reflectivity (among other features-based fields like precipitation): 1) the score is highest at initialization and decreases steadily with increasing forecast length; and 2) scores decrease monotonically with increasing reflectivity threshold. In the case of HRRR forecasts, the impacts of model error appear quickly with OTS decreasing dramatically during the first forecast hour. Since the ICs for HRRR forecasts incorporate radar reflectivity via the cloud analysis routine, which adjusts hydrometeor content and temperature but not vertical motion, reflectivity at initialization tends to closely resemble the observation data source (three-dimensional MRMS reflectivity were used for radar DA while the two-dimensional composite was used for verification). However, any newly assimilated reflectivity structures within the model quickly dissipate due to the lack of accompanying necessary buoyancy or upward motion to sustain convective updrafts (not shown). Therefore, the large decrease in OTS with forecast hour is not surprising. This evolution of OTS is consistent with that of Heidke skill scores (HSS; HSS ranges from negative infinity to 1.0 and so does not have the same range as OTS) for gridpoint forecasts valid over the same period (Fig. 11a), showing that it is appropriate for object-based verification. There are differences in absolute magnitude between the two metrics, and the HSS curves drop more significantly during the first forecast hour, but otherwise the two metrics show similar evolutions. Speculatively, the OTS curves appear to converge with decreasing reflectivity threshold, as the difference between consecutive curves decreases with decreasing threshold, whereas for HSS the difference between consecutive curves appears constant.

OTSs as a function of valid time (Fig. 11b) suggest that the best reflectivity forecasts occur during the 0600–1200 UTC period (overnight/early morning in the United States) with the worst forecasts happening during the 1800–0000 UTC (mid-to-late afternoon) period. This behavior corresponds to the trend in object frequency biases; namely, the overproduction of objects by the HRRR causes the numerator in the OTS formula to decrease relative to the denominator. The apparent 3-hourly OTS signal (Fig. 11b) is an artifact resulting from only using forecasts initialized every three hours. OTSs are substantially higher at initialization compared to later forecast hours, so the averaging incorporates this increase only during times of day containing forecast initializations.

The impact of using the generalized method to compute MMI manifests as reduced scores, both as a function of lead time and as a function of valid time of day (Fig. 12). This result is sensible considering the stricter criteria used to include a given forecast–observation object pair in the calculation: since any given forecast or observation object is removed from further consideration upon its first appearance in the ranked-interest list, the final interest array contains more zeros than the corresponding array in the standard calculation. It remains to be determined, however, which formulation is more useful or meaningful in assessing forecast quality. MMI exhibits similar behavior to OTS, including a substantial decrease between forecast hours 0 and 1 and a small decrease (or no change) beyond that. MMI tends not to vary substantially as a function of time of day, however, unlike the behavior of OTS. Additionally, there is

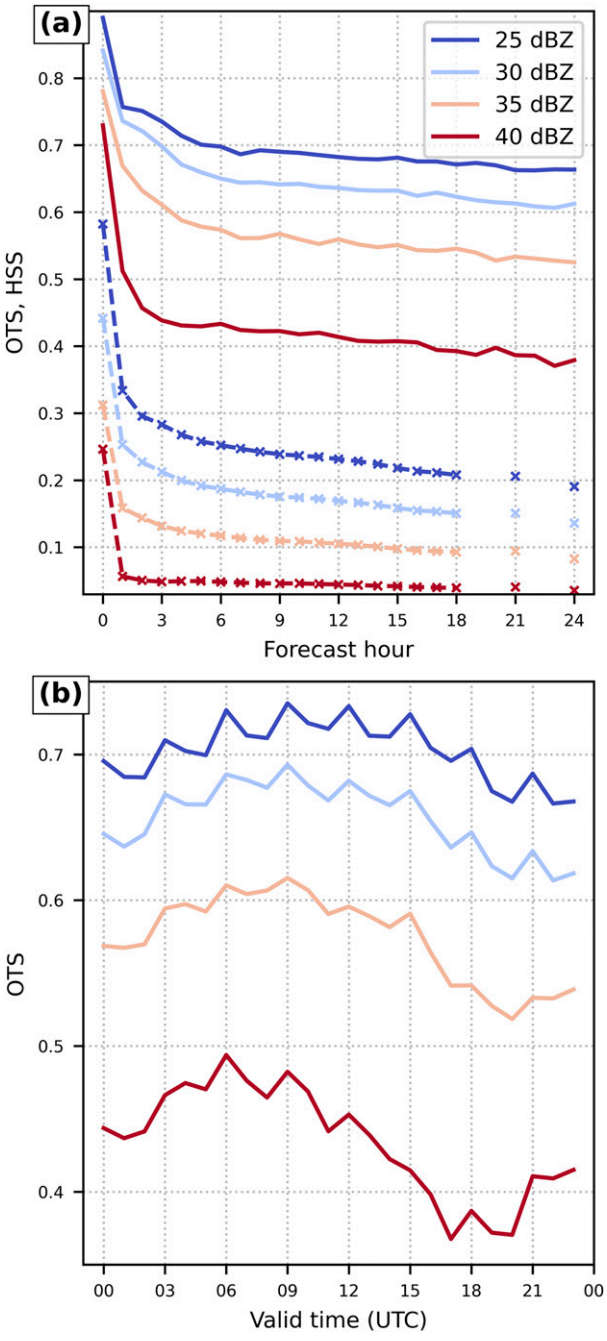

FIG. 11. OTS (solid lines) as a function of (a) forecast hour and (b) time of day. HSS values derived from gridpoint comparisons are plotted in (a) in dashed lines with "X" marks.

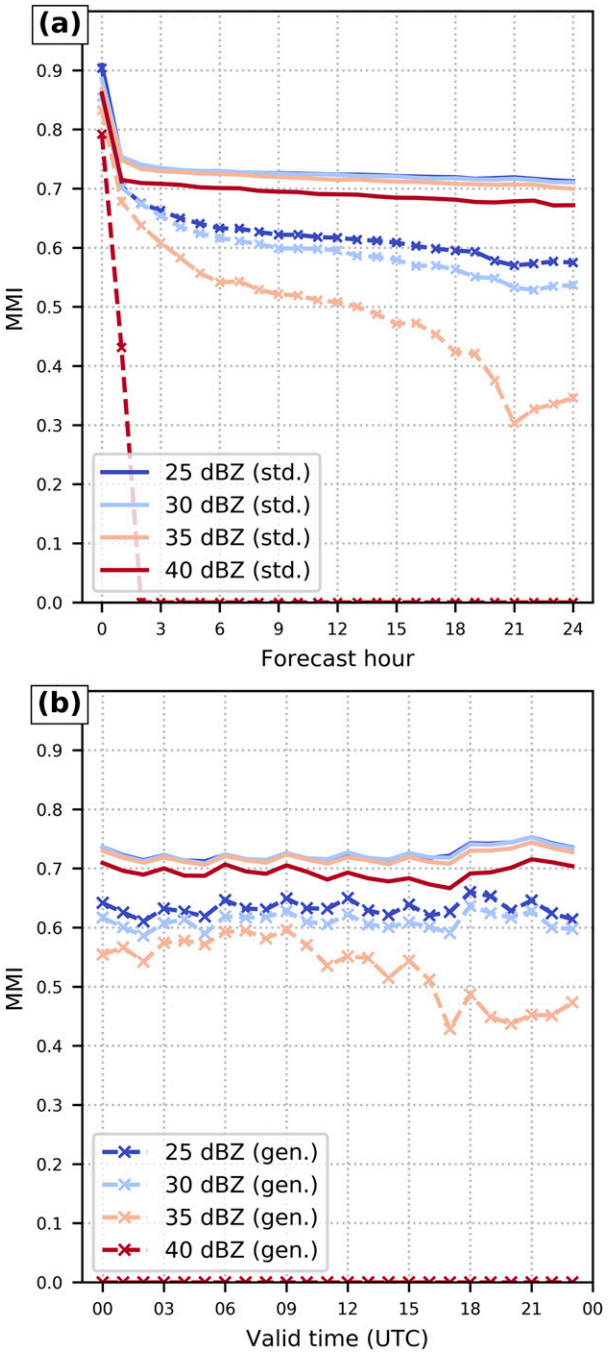

FIG. 12. MMI as a function of (a) forecast lead time and (b) valid time of day for the indicated reflectivity thresholds. Solid lines represent the standard calculation of MMI, whereas dashed lines with "X" marks denote alternative MMIs calculated using the generalized method.

minimal variance of MMI values among the reflectivity thresholds for the standard formulation, whereas OTS values decrease noticeably and steadily with increasing reflectivity threshold. The alternative MMI (calculated using the generalized method), however, exhibits a strong sensitivity to reflectivity threshold; at the 25- and 30-dBZ thresholds, MMI is nearly the same. However, alternative MMI values decrease more between the 30- and 35-dBZ thresholds, and then precipitously (generally going to 0.0) at the 40-dBZ threshold. This sensitivity reveals a characteristic to the alternative formulation that does not describe the standard formulation, and it is speculated that this formulation reveals more about the reflectivity values at which the forecasts become substantially

less accurate than does the standard formulation. But more testing is needed.

Compared to OTS and MMI, which are normalized statistical quantities, mean centroid distance has an inherent physical meaning, and therefore individual values are meaningful for assessing skill of HRRR forecasts. Regardless of whether only storm-shaped objects (small area, high 95th percentile value) or all matched objects are considered, and whether the generalized method is used, the mean centroid displacement is essentially fixed among thresholds (Fig. 13a). This behavior is at least somewhat expected since the set of objects defined at a higher threshold is entirely contained within the set of objects defined at all lower thresholds. However, this logic does not explain all behavior of mean centroid displacements since, e.g., there are approximately 3–4 times as many object pairs at 25 dBZ as at 40 dBZ. Distributions of pair centroid distances (not shown) are effectively identical among thresholds. These results imply that displacement errors in reflectivity objects defined at one threshold but not at higher thresholds have essentially the same distributions. The other major finding is that the generalized mean centroid displacement error is lowest (~40 km) at forecast initialization and increases dramatically in the first forecast hour, then slowly and steadily beyond that, similar to the signal of OTS and MMI. This behavior lends credence to the use of mean centroid displacement as a representative measure of forecast accuracy. It is also noteworthy that the mean displacements between all matched objects (Fig. 13c) is approximately the same as the mean displacements between objects classified as discrete cells (Fig. 13b), whereas the generalized mean displacement values are about 40 km larger (Fig. 13a). The difference between the mean and median values also suggests there are differences in the skew of the centroid displacement distributions; for the generalized method, there must necessarily be more extremely high centroid displacement errors compared to the other object sets. Since pair interest values are not considered when calculating mean centroid distances using the generalized method, it makes sense that these larger distances come from object pairs with lower interest values. Evidently, these pairs also do not involve discrete storm objects. These centroid displacement, OTS, and MMI results are broadly consistent with those of Blaylock and Horel (2020) who discerned centroid displacement errors of around 60 km, increased errors in the late afternoon, and a quick drop in forecast quality shortly after initialization when using the fractions skill score (FSS).

The two-dimensional object displacement distributions reveal a nearly anisotropic and unbiased error direction tendency among HRRR forecasts, especially at the 25- and 30-dBZ reflectivity thresholds (not shown). At 35 and 40 dBZ, however, and especially during the 0600–1700 UTC time, a northwesterly bias appears in the centroid displacement distribution; the mean displacement is about 26 km in a north-northwest direction, although the mode of the joint (west–east/south–north) probability distribution remains close to (0, 0) km (Fig. 14).

d. Performance diagrams

The performance diagram introduced in Roebber (2009) can also be used in an object-based sense by using the interest value threshold for matching forecast-observation object pairs to determine the elements of a 2 × 2 contingency table. The results (Fig. 15) are consistent with examination of object frequency bias (cf. Figs. 5 and 7 with 15) with all points laying in the upper-left half of the diagram (where frequency bias > 1.0). The results are also consistent with OTS (cf. Figs. 11 and 15) in

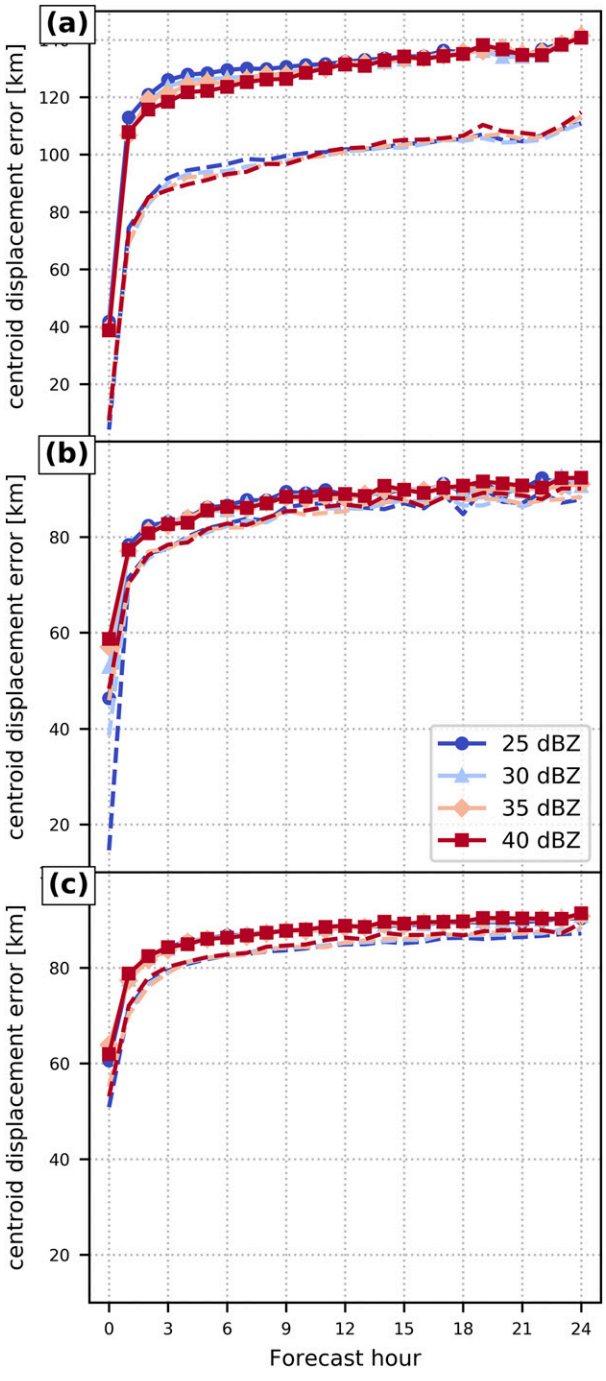

FIG. 13. Mean (solid) and median (dashed) centroid displacement error for (a) generalized method, (b) only objects representing a discrete convective storm, and (c) all matched objects.

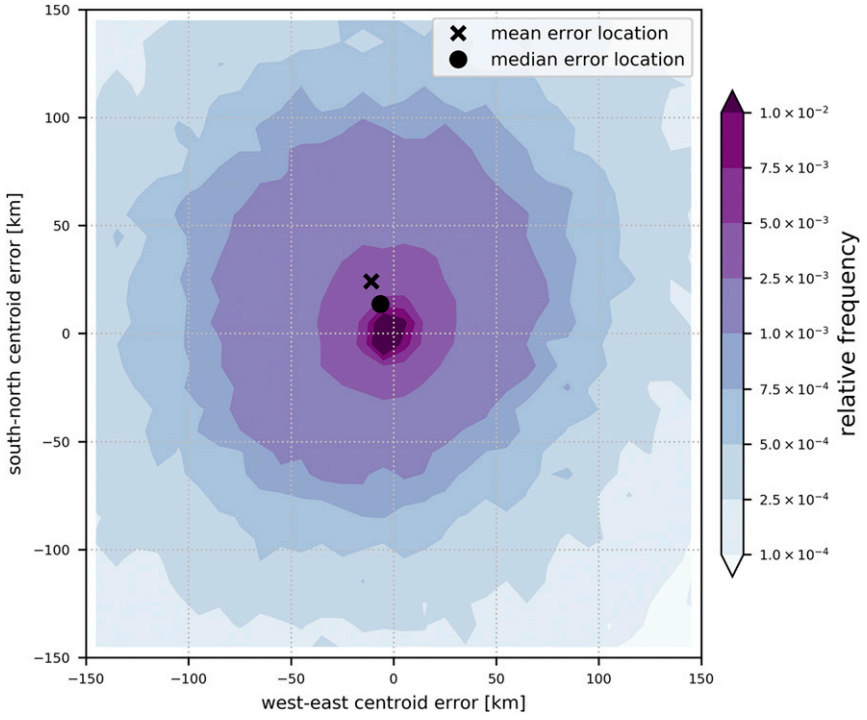

FIG. 14. Probability density function of two-dimensional centroid displacement error at 40 dBZ aggregated between 0600 and 1700 UTC.

that forecasts at initialization have high accuracy (CSI ≈ 0.8 for 25-, 30-, and 35-dBZ thresholds and 0.7 for 40 dBZ) followed by a rapid decrease in forecast quality during the first two or three forecast hours and then a slow decrease beyond that. Using this phase space, it is clear that the forecast quality decrease is coincident with a decrease in the success ratio (an increasing number of false alarms) whereas the probability of detection remains either constant or decreases only slightly with increasing forecast length. This is consistent with prior results that the HRRR produces too many objects late in the forecast.

With respect to time of day, the best forecasts tend to occur between 2100 and 0000 UTC, during the late afternoon, with forecast quality at a minimum centered around 1200 UTC, consistent with results from Blaylock and Horel (2020). This outcome is inconsistent with OTSs as a function of time of day. However, the inconsistency is not problematic, as these two verification metrics assess different forecast components; OTS is impacted by object frequency bias, which is at a minimum (closest to 1.0) around 1200 UTC and is higher during the late afternoon. Object-based CSI, on the other hand, does not explicitly account for object count bias, so it is likely that the higher forecast object counts in the late afternoon result in multiple forecast objects being matched to a single observation object, which contributes to increased hit counts in the numerator of the CSI formula.

4. Conclusions

The Method of Object-based Diagnostic Evaluation (MODE) was used to conduct an object-based verification analysis of approximately 1400 operational HRRR forecasts from the 2019 U.S. warm season. Composite reflectivity forecasts initialized every three hours were verified hourly out to 24 h. MODE was configured to classify primarily convective storm objects. The attribute weights, which enable the specification of a single statistic to encompass a number of physical features of the objects, were subjectively set to be largest for distance, near-max reflectivity, and area comparisons between forecast and observation objects. A variety of verification metrics and assessment methods were used to ascertain aspects of HRRR reflectivity forecasts against MRMS radar reflectivity observations that traditional gridpoint metrics are not equipped to provide. A summary of the findings from object-based verification of the HRRR forecasts is below:

- HRRR overpredicts the number of reflectivity objects in general, especially small-sized objects (i.e., discrete convective storms). This overprediction is particularly high during the afternoon and early evening when convective activity is at its diurnal maximum. However, unknown bias errors in the reflectivity diagnostic in the model may have contributed to the overprediction.
- The overprediction of storms is worse across the south-central, southeast, and northeastern CONUS. Reflectivity objects are predicted with better frequency across the southern plains. Furthermore, there is some spatial correlation between overforecasting of small storms and complex or elevated terrain, which may suggest model physics errors associated with the near-surface flow.

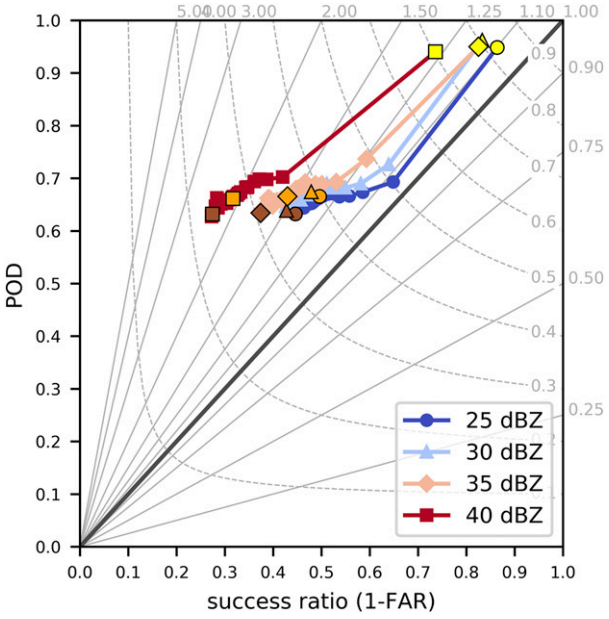

FIG. 15. Performance diagram for HRRR composite reflectivity forecasts. Each symbol represents a forecast hour. Alternately colored symbols emphasize the following key forecast hours: yellow (0), orange (12), and brown (24).

- Resolution issues associated with the 3-km grid spacing of the HRRR impact some shape parameters of reflectivity objects. In particular, small objects do not have the proper spatial structure. This result illustrates an unavoidable outcome from verifying a forecast field with an observation field derived on a different scale; the MRMS data contain features on a scale that are underresolved on the HRRR grid, and the binary interpolation of the MRMS data to the HRRR grid are not guaranteed to fully remove all of the finer-scale features. Therefore, this particular result should be viewed with that caveat.
- There is a rapid and substantial drop in forecast quality with lead time, especially in the first 1–3 h. Adjustments from a dynamically unbalanced initial state plus burgeoning model error likely contribute to this behavior.
- There is a slight northwest bias in object displacement at high thresholds and during the less convectively active times of day. At lower thresholds and busier times of day, no systematic bias in location is observed.

An important result from this work is the illustration that an object-based approach can add information to gridpoint-based assessment of the quality of NWP forecasts. Additionally, object-based verification can provide information on specific attributes of the forecast that either excel or need improvement that cannot be ascertained from gridpoint verification. Finally, gridpoint- and object-based verification can be related to each other via certain degrees of similarity in certain metrics (e.g., frequency bias and OTS versus HSS). This should help aid researchers and developers in transitioning from a verification framework dominated by traditional gridpoint metrics to one that incorporates object-based assessments into the traditional approach.

This work demonstrated how object-based verification can be used to evaluate HRRR reflectivity forecasts, in particular. Future work will use MODE to evaluate the accuracy of 1- and 6-h accumulated precipitation forecasts. Additionally, we will use these object-based verification techniques to evaluate the relative difference between HRRR v3 (analyzed here) and v4 (implemented operationally at the National Centers for Environmental Prediction on 2 December 2020). Evaluating multiple forecast systems can also shed light on which MMI formulation is preferrable.

It is important to stress that the verification herein was conducted in a fixed-time framework, meaning temporal errors were ignored; forecasts were compared to observations only at the same valid time. It is well known that features in NWP model output can contain temporal errors without spatial errors, such as a thunderstorm predicted in the correct location but one hour late compared to observations. Ignoring such errors in this work has implications on the effectiveness of this verification strategy, albeit not sufficiently substantial to invalidate our conclusions. There are options for incorporating the temporal aspects of forecast error. One such option is the time-domain expansion to MODE, called MODE-TD (Clark et al. 2014). However, preliminary experimentation with MODE-TD revealed additional computational requirements that currently make it infeasible for verifying HRRR forecasts beyond a few hours. Once the MODE-TD software matures, this work can be revisited by applying the time dimension to 15-min reflectivity output (for better resolved temporal forecasts) to enable assessment of metrics such as timing errors of storms. Additionally, MODE-TD is ideal for verifying hourly max fields such as updraft helicity (UH) due to occasions when overlapping UH swaths are caused by storms occurring in the same location but at different times within an integration period. Prior unreported work applying MODE to UH forecasts resulted in many objects considered unrepresentative of the actual forecast evolution due to overlapping tracks that could not be separated using MODE.

This use of object-based verification provides one way to evaluate the physical processes and shortcomings in the DA system that lead to disagreements between model forecasts and observations. We anticipate integrating this object-based method into the Model Analysis Tool Suite, which is being increasingly used by model developers within the National Weather Service and NOAA.

Acknowledgments. This work was supported by the NOAA Cooperative Agreement with CIRES, NA17OAR4320101. The author thanks the support and discussion with members of the Developmental Testbed Center team that developed and maintained the Model Evaluation Tools suite and the MODE software, including John Halley Gotway, Tara Jensen, Randy Bullock, and Julie Prestopnik. This work was partially inspired by discussions with, and presentations given by Michael Erickson at CIRES/NOAA Weather Prediction Center. Jeff Hamilton at CIRES/GSL provided helpful feedback and suggestions of an early version of this manuscript. The constructive comments from John Horel and two additional reviewers further improved this manuscript.

Data availability statement. The HRRR forecast data used in this study are archived in a number of places, both public and private. Although the following resource was not used herein, the data can be obtained from the operational HRRR output archive from the Google Cloud Platform at https://console.cloud.google.com/marketplace/product/noaa-public/hrrr. While MRMS observation data for the particular cases herein were obtained from NOAA's Research and Development High-Performance System's (RDHPCS) High-Performance Storage System (HPSS) archive and are not available publicly, a publicly available archive of MRMS files beginning in September 2019 is available at http://mesonet.agron.iastate.edu/archive. Instructions for reproducing the work herein have been uploaded to Mendeley Data (doi:10.17632/h4248b6gcc.1).

REFERENCES

Adams-Selin, R. D., A. J. Clark, C. J. Melick, S. R. Dembeck, I. L. Jirak, and C. L. Ziegler, 2019: Evolution of WRF-HAILCAST during the 2014–16 NOAA/Hazardous Weather Testbed Spring Forecasting Experiments. *Wea. Forecasting*, **34**, 61–79, https://doi.org/10.1175/WAF-D-18-0024.1.

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.

Alexander, C. R., S. S. Weygandt, T. G. Smirnova, S. Benjamin, P. Hofmann, E. P. James, and D. A. Koch, 2010: High Resolution Rapid Refresh (HRRR): Recent enhancements and evaluation during the 2010 convective season. *25th Conf. on Severe Local Storms*, Denver, CO, Amer. Meteor. Soc., 9.2, https://ams.confex.com/ams/25SLS/techprogram/paper_175722.htm.

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.

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.

Brown, B., and Coauthors, 2021: The Model Evaluation Tools (MET): More than a decade of community-supported forecast verification. *Bull. Amer. Meteor. Soc.*, **102**, E782–E807, https://doi.org/10.1175/BAMS-D-19-0093.1.

Bytheway, J. L., C. D. Kummerow, and C. Alexander, 2017: A features-based assessment of the evolution of warm season precipitation forecasts from the HRRR model over three years of development. *Wea. Forecasting*, **32**, 1841–1856, https://doi.org/10.1175/WAF-D-17-0050.1.

Cai, H., and R. E. Dumais Jr., 2015: Object-based evaluation of a numerical weather prediction model's performance through forecast storm characteristic analysis. *Wea. Forecasting*, **30**, 1451–1468, https://doi.org/10.1175/WAF-D-15-0008.1.

Clark, A. J., R. G. Bullock, T. A. Jensen, M. Xue, and F. Kong, 2014: Application of object-based time-domain diagnostics for tracking precipitation systems in convection-allowing models. *Wea. Forecasting*, **29**, 517–542, https://doi.org/10.1175/WAF-D-13-00098.1.

——, and Coauthors, 2016: Spring Forecasting Experiment 2016: Preliminary findings and results. NOAA/NSSL/SPC, 50 pp., https://hwt.nssl.noaa.gov/Spring_2016/HWT_SFE_2016_preliminary_findings_final.pdf.

——, and Coauthors, 2017: Spring Forecasting Experiment 2017: Preliminary findings and results. NOAA/NSSL/SPC, 50 pp.

——, and Coauthors, 2018: Spring Forecasting Experiment 2018: Preliminary findings and results. NOAA/NSSL/SPC, 69 pp., https://hwt.nssl.noaa.gov/sfe/2018/docs/HWT_SFE_2018_Prelim_Findings_v1.pdf.

——, and Coauthors, 2019: Spring Forecasting Experiment 2019: Preliminary findings and results. NOAA/NSSL/SPC, 77 pp., https://hwt.nssl.noaa.gov/sfe/2019/docs/HWT_SFE_2019_Prelim_Findings_FINAL.pdf.

——, and Coauthors, 2020: Spring Forecasting Experiment 2020: Preliminary findings and results. NOAA/NSSL/SPC, 77 pp., https://hwt.nssl.noaa.gov/sfe/2020/docs/HWT_SFE_2020_Prelim_Findings_FINAL.pdf.

Davis, C. A., B. G. Brown, and R. G. 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.

——, ——, ——, and J. Halley-Gotway, 2009: The Method for Object-Based Diagnostic Evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC spring program. *Wea. Forecasting*, **24**, 1252–1267, https://doi.org/10.1175/2009WAF2222241.1.

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

Duda, J. D., and W. A. Gallus, 2013: The impact of large-scale forcing on skill of simulated convective initiation and upscale evolution with convection-allowing grid spacings in the WRF. *Wea. Forecasting*, **28**, 994–1018, https://doi.org/10.1175/WAF-D-13-00005.1.

——, X. Wang, Y. Wang, and J. R. Carley, 2019: Comparing the assimilation of radar reflectivity using the direct GSI-based Ensemble–Variational (EnVar) and indirect cloud analysis methods in convection-allowing forecasts over the continental United States. *Mon. Wea. Rev.*, **147**, 1655–1678, https://doi.org/10.1175/MWR-D-18-0171.1.

Ebert, E. E., and W. A. Gallus, 2009: Toward better understanding of the contiguous rain area (CRA) method for spatial forecast verification. *Wea. Forecasting*, **24**, 1401–1415, https://doi.org/10.1175/2009WAF2222252.1.

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.

Gallus, W. A., J. Wolff, J. Halley Gotway, M. Harrold, L. Blank, and J. Beck, 2019: The impacts of using mixed physics in the Community Leveraged Unified Ensemble. *Wea. Forecasting*, **34**, 849–867, https://doi.org/10.1175/WAF-D-18-0197.1.

Griffin, S. M., J. A. Otkin, C. M. Rozoff, J. M. Sieglaff, L. M. Cronce, and C. R. Alexander, 2017: Methods for comparing simulated and observed satellite infrared brightness temperatures and what do they tell us? *Wea. Forecasting*, **32**, 5–25, https://doi.org/10.1175/WAF-D-16-0098.1.

——, ——, G. Thompson, M. Frediani, J. Berner, and F. Kong, 2020: Assessing the impact of stochastic perturbations in cloud microphysics using *GOES-16* infrared brightness temperatures. *Mon. Wea. Rev.*, **148**, 3111–3137, https://doi.org/10.1175/MWR-D-20-0078.1.

Halley Gotway, J., K. Newman, T. Jensen, B. Brown, R. Bullock, and T. Fowler, 2018: Model Evaluation Tools version 8.0

(METv8.0) user's guide. Developmental Testbed Center, 432 pp., https://dtcenter.org/sites/default/files/community-code/met/docs/user-guide/MET_Users_Guide_v8.0.pdf .

Hartung, D. C., J. A. Otkin, R. A. Petersen, D. D. Turner, and W. F. Feltz, 2011: Assimilation of surface-based boundary layer profiler observations during a cool-season weather event using an observing system simulation experiment. Part II: Forecast assessment. *Mon. Wea. Rev.*, **139**, 2327–2346, https://doi.org/10.1175/2011MWR3623.1.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hu, M., M. Xue, and K. Brewster, 2006: 3DVAR and cloud analysis with WSR-88D level-II data for the prediction of the Fort Worth, Texas, tornadic thunderstorms. Part I: Cloud analysis and its impact. *Mon. Wea. Rev.*, **134**, 675–698, https://doi.org/10.1175/MWR3092.1.

Johnson, A., and X. Wang, 2012: Verification and calibration of neighborhood and object-based probabilistic precipitation forecasts from a multimodel convection-allowing ensemble. *Mon. Wea. Rev.*, **140**, 3054–3077, https://doi.org/10.1175/MWR-D-11-00356.1.

——, and ——, 2013: Object-based evaluation of a storm-scale ensemble during the 2009 NOAA Hazardous Weather Testbed Spring Experiment. *Mon. Wea. Rev.*, **141**, 1079–1098, https://doi.org/10.1175/MWR-D-12-00140.1.

——, ——, Y. Wang, A. Reinhart, A. J. Clark, and I. L. Jirak, 2020: Neighborhood- and object-based probabilistic verification of the OU MAP ensemble forecasts during 2017 and 2018 Hazardous Weather Testbeds. *Wea. Forecasting*, **35**, 169–191, https://doi.org/10.1175/WAF-D-19-0060.1.

Jones, T. A., P. Skinner, K. Knopfmeier, E. Mansell, P. Minnis, R. Palikonda, and W. Smith Jr., 2018: Comparison of cloud microphysics schemes in a warn-on-forecast system using synthetic satellite objects. *Wea. Forecasting*, **33**, 1681–1708, https://doi.org/10.1175/WAF-D-18-0112.1.

Moser, B. A., W. A. Gallus, and R. Mantilla, 2015: An initial assessment of radar data assimilation on warm season rainfall forecasts for use in hydrologic models. *Wea. Forecasting*, **30**, 1491–1520, https://doi.org/10.1175/WAF-D-14-00125.1.

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.

Roberts, B., B. T. Gallo, I. L. Jirak, A. J. Clark, D. C. Dowell, X. Wang, and Y. Wang, 2020: What does a convection-allowing ensemble of opportunity buy us in forecasting thunderstorms? UFS Users' Workshop, UFS, 12 pp., https://dtcenter.org/sites/default/files/events/2020/3-roberts-brett.pdf.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, https://doi.org/10.1175/2008WAF2222159.1.

Schwartz, C. S., G. S. Romine, K. R. Fossell, R. A. Sobash, and M. A. Weisman, 2017: Toward 1-km ensemble forecasts over large domains. *Mon. Wea. Rev.*, **145**, 2943–2969, https://doi.org/10.1175/MWR-D-16-0410.1.

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., L. J. Wicker, D. M. Wheatley, and K. H. Knopfmeier, 2016: Application of two spatial verification methods to ensemble forecasts of low-level rotation. *Wea. Forecasting*, **31**, 713–735, https://doi.org/10.1175/WAF-D-15-0129.1.

——, 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.

Squitieri, B. J., and W. A. Gallus, 2020: On the forecast sensitivity of MCS cold pools and related features to horizontal grid spacing in convection-allowing WRF simulations. *Wea. Forecasting*, **35**, 325–346, https://doi.org/10.1175/WAF-D-19-0016.1.

Stratman, D. R., and K. A. Brewster, 2017: Sensitivities of 1-km forecasts of 24 May 2011 tornadic supercells to microphysics parameterizations. *Mon. Wea. Rev.*, **145**, 2697–2721, https://doi.org/10.1175/MWR-D-16-0282.1.

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.

Wernli, H., M. Paulat, M. Hagen, and C. Frei, 2008: SAL—A novel quality measure for the verification of quantitative precipitation forecasts. *Mon. Wea. Rev.*, **136**, 4470–4487, https://doi.org/10.1175/2008MWR2415.1.

——, C. Hofmann, and M. Zimmer, 2009: Spatial forecast verification methods intercomparison project: Application of the SAL technique. *Wea. Forecasting*, **24**, 1472–1484, https://doi.org/10.1175/2009WAF2222271.1.

Zhang, J., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 621–638, https://doi.org/10.1175/BAMS-D-14-00174.1.