Quantification of NSSL Warn-On-Forecast System Accuracy by Storm Age using Object-based Verification

Jorge E. Guerra, Patrick S. Skinner, Adam Clark , Montgomery Flora, Brian Matilla, Kent Knopfmeier, and Anthony E. Reinhart

Cooperative Institute for Severe and High-Impact Weather Research and Operations CIWRO, Norman, Oklahoma.

Corresponding author: Dr. Jorge E. Guerra, jorge.guerra@noaa.gov

Current affiliaton: NSSL

Current affiliation: OU/CIWRO, NSSL

Current affiliation: OU/CIWRO

Current affiliation: OU/CIWRO, NSSL

Current affiliation: OU/CIWRO

Current affiliation: OU/CIWRO

Early Online Release: This preliminary version has been accepted for publication in *Weather and Forecasting* cited, and has been assigned DOI 10.1175/WAF-D-22-0043.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT: The National Severe Storm Laboratory's Warn-on-Forecast System (WoFS) is a convection-allowing ensemble with rapidly cycled data assimilation (DA) of various satellite and radar datasets designed for prediction at 0-6 h lead time of hazardous weather. With the focus on short lead times, WoFS predictive accuracy is strongly dependent on its ability to accurately initialize and depict the evolution of ongoing storms. Since it takes multiple DA cycles to fully "spin up" ongoing storms, predictive skill is likely a function of storm age at the time of model initialization, meaning that older storms which have been through several DA cycles will be forecast with greater accuracy than newer storms which initiate just before model initialization or at any point after. To quantify this relationship, we apply an object-based spatial tracking and verification approach to map differences in the probability of detection (POD), in space-time, of predicted storm objects from WoFS with respect to Multi-Radar Multi-Sensor (MRMS) reflectivity objects. Object-tracking/matching statistics are computed for all suitable and available WoFS cases from 2017 through 2021. Our results indicate sharply increasing POD with increasing storm age for lead times within three hours. PODs were about 0.3 for storm objects that emerge 2-3 h after model initialization, while for storm objects that were at least an hour old at the time of model initialization by DA, PODs ranged from around 0.7 to 0.9 depending on the lead time. These results should aid in forecaster interpretation of WoFS, as well as guide WoFS developers on improving the model and DA system.

SIGNIFICANCE STATEMENT: The Warn-on-Forecast System (WoFS) is a collection of weather models designed to predict individual thunderstorms. Before the models can predict storms, they must ingest radar and satellite observations to put existing storms into the models. Because storms develop at different times, more observations will exist for some storms in the model domain than others, which results in WoFS forecasts with different accuracy for different storms. This paper estimates the differences in accuracy for storms that have existed for a long time and those that haven't by tracking observed and predicted storms. We find that the likelihood of WoFS accurately predicting a thunderstorm nearly doubles if the storm has existed for over an hour prior to the forecast. Understanding this relationship between storm age and forecast accuracy will help forecasters better use WoFS predictions and guide future research to improve WoFS forecasts.

1. Introduction

The National Severe Storms Laboratory's Warn-on-Forecast System (WoFS) (Stensrud et al. 2009, 2013), is an on-demand, rapidly-updating, convection-allowing ensemble designed to dramatically improve lead times for hazardous weather. WoFS targets watch to warning lead times (i.e., 0-6 h), and tentative plans are for WoFS to become operational for the National Weather Service (NWS) within the 2025-30 time frame. For several years now, NSSL has conducted successful real-time demonstrations of a prototype WoFS configuration during the spring and summer as part of forecasting activities during NOAA Hazardous Weather Testbed Spring Forecasting Experiments (Gallo et al. 2017; Clark et al. 2020, 2021) and the Weather Prediction Center's Flash Flood and Intense Rainfall Experiment (Trojniak and Albright 2019). Thus, an extensive archive of forecast cases has been amassed from 2017 through 2021. Concurrently, object-based verification methods for evaluating the quality of WoFS guidance have been developed to match predicted thunderstorm objects in WoFS to corresponding objects in gridded NEXRAD data from the Multi-Radar Multi-Sensor (MRMS; Smith et al. 2016) system. This object-based verification approach can be implemented for any diagnostic field derived from both model data and observations (e.g. Davis et al. 2006a; Gilleland et al. 2009; Wolff et al. 2014) and has been extensively applied to prediction of radar-reflectivity-based proxies for thunderstorms in WoFS (e.g. Skinner et al. 2018; Flora et al. 2019, 2021; Miller et al. 2022).

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-22-0043.1.

In this study we extend the object-based verification methodology of Skinner et al. (2018) to include thunderstorm object tracking in time of MRMS composite reflectivity objects, which we match to corresponding objects in the simulated reflectivity of WoFS members on spatial scales comparable to National Weather Service warning products. Tracking observed thunderstorm objects enables the age of storms relative to convection initiation (CI) to be estimated. Subsequent matching of MRMS objects to WoFS thunderstorm objects allows the probability of detection (POD) to be calculated as a function of storm age relative to CI.

The central objective of this research is to quantify changes in accuracy for successive WoFS analyses and forecasts produced using rapidly cycled, ensemble Kalman filter (EnKF; Houtekamer and Zhang 2016) based assimilation of remotely sensed radar and satellite observations of convective storms (e.g., Wheatley et al. 2015; Jones et al. 2016). We demonstrate a clear effect of cycled data assimilation (DA), where POD in WoFS analyses and short-lead-time forecasts increases markedly with increasing storm age. We also show that DA-based improvements in thunderstorm forecasts decrease with increasing forecast length, but are maintained through 3 hours of forecast lead time. Quantification of the impact of rapidly cycled DA on the quality of WoFS thunderstorm analysis and prediction serves two broader goals:

1. Establish expected changes in WoFS forecast quality following CI. Correlation of POD with storm age results in storm-to-storm variation in forecast quality across the WoFS domain, which complicates effective real-time interrogation of WoFS guidance. Quantification of expected changes in forecast quality with storm age will inform best practices for operational use of WoFS guidance (e.g. Wilson et al. 2021; Gallo et al. 2022), and could be useful input for machine-learning-based post-processing algorithms (e.g., Flora et al. 2021).

2. Determine the approximate number of DA cycles needed to produce accurate storm-scale analyses in WoFS. The efficiency of EnKF-based, rapidly cycled DA of radar and satellite observations has not been assessed in a quasi-operational system across a large sample of cases. Quantifying the typical ensemble "spin up" time for WoFS thunderstorm analyses will establish a baseline for future DA configurations to be tested against.

The remaining sections are organized as follows: we describe the datasets used in this study and our object-based verification software in Section 2, including details on quality control applied to the forecast and verification datasets and algorithms used for object identification, tracking, and matching. We present our results in Section 3 and provide conclusions, including potential avenues for improving the system, in Section 4.

2. Object-based Tracking and Matching

Object-based verification techniques provide a robust and flexible means to quantify the skill of an NWP system and are often used to assess the forecast quality of guidance for discrete events, such as areas of heavy precipitation (e.g. Davis et al. 2006b, 2009; Ebert and Gallus 2009; Johnson et al. 2013; Clark et al. 2014; Wolff et al. 2014; Bytheway and Kummerow 2015), jet streaks (Hewson and Titley 2010; Mittermaier et al. 2016), or upper-level clouds (Mittermaier and Bullock 2013; Griffin et al. 2017; Jones et al. 2018; Griffin et al. 2021). Object-based techniques have proven to be particularly well suited to verification of convection-allowing model (CAM) thunderstorm forecasts (e.g. Kain et al. 2013; Cai and Dumais 2015; Sobash et al. 2016; Schwartz et al. 2017; Potvin et al. 2019; Duda and Turner 2021), including predictions of CI (Burghardt et al. 2014; Burlingame et al. 2017), storm mode (Pinto et al. 2015; Johnson et al. 2020), and mesocyclone occurrence (e.g. Clark et al. 2013, 2014; Skinner et al. 2016; Stratman and Brewster 2017). As WoFS is designed to predict hazards within individual thunderstorms at short lead times (0–6 hr), object-based verification is a natural fit for evaluation of WoFS forecast quality and has been used to establish baselines of WoFS skill for thunderstorm and mesocyclone prediction (Skinner et al. 2018; Flora et al. 2019) and quantify changes in forecast skill across different system configurations (Jones et al. 2018, 2020; Flora et al. 2021; Kerr et al. 2021; Miller et al. 2022).

a. Forecast and Verification Datasets

WoFS composite reflectivity forecasts for 176 cases between the years of 2017 and 2021 are evaluated. There have been several minor changes to the WoFS configuration during this period, most notably a switch from the Data Assimilation Research Testbed (DART; Anderson and Collins 2007; Anderson et al. 2009) to the Community Gridpoint Statistical Interpolation (GSI; Kleist et al. 2009) data assimilation system in the summer of 2018. Additionally, changes to the initial and boundary conditions provided by the High-Resolution Rapid Refresh Ensemble (HRRRE; Dowell et al. 2016) and pre-processing of radar and satellite observations assimilated have been made.

Please refer to Jones et al. (2018); Skinner et al. (2018); Yussouf and Knopfmeier (2019); Jones et al. (2020) for detailed descriptions of WoFS configurations during the period. Despite these changes, the system has used the National Severe Storms Laboratory 2-moment microphysical parameterization (NSSL 2-moment; Mansell et al. 2010) to calculate simulated reflectivity during the full period and subjective and object-based comparisons of reflectivity forecasts from year-to-year have not revealed large changes in system performance; therefore, we treat WoFS configurations as approximately consistent through the period.

To ensure consistency across the dataset, we consider 18-member WoFS forecasts initialized hourly between 20 and 02 UTC with up to 3 hours of forecast lead time for each case. MRMS gridded composite reflectivity observations are produced for each possible valid time of WoFS forecasts to serve as a verification dataset. MRMS reflectivities are interpolated to the WoFS grid using a Cressman filter with a 3-km radius of influence to match the WoFS model native horizontal resolution of 3 km. Lastly, both WoFS and interpolated MRMS datasets contain output every 5 minutes.

b. Object Identification and Quality Control

A schematic for our tracking and matching implementation is shown in Figure 1, including the various parameters needed to perform quality control on the inputs and achieve a physically consistent set of object matches. The parameters shown in Figure 1 are described therein.

The primary task of the object identification software is to determine the physical area within the composite reflectivity field that comprises a thunderstorm. A storm object is first identified as the closed contour of values exceeding a prescribed reflectivity threshold, dBZ_1 . For our purposes, dBZ_1 is set to 40 in MRMS and 43 in WoFS. WoFS uses a slightly higher threshold because it has a slight high bias and 43 dBZ is approximately equivalent to the same percentile as 40 dBZ in the MRMS data. The objects are then labeled using a consecutive integer for a given instant in time. Two filters are then applied: The first filter aims to eliminate objects at scales less than the effective model resolution, which is about 4-6 times the horizontal grid-spacing ($A_f = 144\ km^2$ corresponds to a 12 x 12 km or 4 x 4 grid-length box) and the second discards objects that have a maximum intensity less than a second reflectivity threshold (dBZ_2 , set to 48, 45 dBZ in WoFS and MRMS data, respectively). In the first case we seek to eliminate object signals that are regarded as

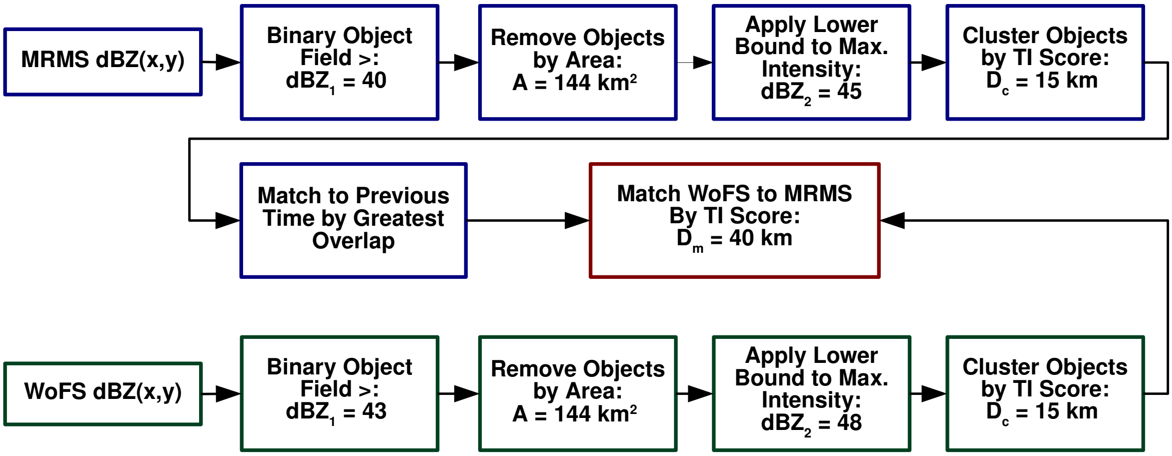
MRMS dBZ(x,y)
Binary Object Field >: dBZ$_1$ = 40
Remove Objects by Area: A = 144 km²
Apply Lower Bound to Max. Intensity: dBZ$_2$ = 45
Cluster Objects by TI Score: D$_c$ = 15 km
Match to Previous Time by Greatest Overlap
Match WoFS to MRMS By TI Score: D$_m$ = 40 km
WoFS dBZ(x,y)
Binary Object Field >: dBZ$_1$ = 43
Remove Objects by Area: A = 144 km²
Apply Lower Bound to Max. Intensity: dBZ$_2$ = 48
Cluster Objects by TI Score: D$_c$ = 15 km

FIG. 1. Tracking and matching algorithm sequence for a single coincident valid time. Parameters with their respective units are as follows: dBZ_1 (dBZ); Initial object field built from values greater than this. dBZ_2 (dBZ); Objects with a lower maximum intensity are filtered out. A (km^2); Objects smaller in area than A are filtered out. D_c (km); Reference distance for TI matching of object clusters. D_m (km); Reference distance for TI matching of model to observation objects.

spurious and can contribute to an over-prediction bias of convective initiation, taking into account a judgement of the effective resolution of both model and radar fields. The second filter aims to disregard precipitation modes that are not associated with convective storms, in particular, intense stratiform regions trailing mesoscale convective systems that have been found to exceed 45 dBZ in WoFS forecasts. The matching thresholds/parameters specified here and shown in Fig. 1 are derived from Skinner et al. (2018), which also provides a sensitivity analysis for small changes in threshold values.

A second level of processing is used to identify clusters of storms that may be regarded as a single object in the subsequent object matching. In this study we process MRMS and WoFS object fields independently to identify clusters of storms based on a simplified total interest (TI) score based on (Davis et al. 2009; Skinner et al. 2018) given by:

$$TI_{\text{cluster}} = \frac{(D_c - D_{min})}{D_c}. \tag{1}$$

where D_c is the distance threshold for clustering and D_{min} is the minimum distance between a pair of storm objects computed as the minimum distance between the boundaries of two objects. Here, D_{min} is set to 15 km and objects with a TI_{cluster} greater than 0.2 are considered a cluster. This threshold was set heuristically by examination of matched fields over a small sample of cases.

The search algorithm finds clusters of storm objects, relabels them to a common number, and computes and stores aggregate diagnostic properties for the cluster. It is important to note that this type of merging occurs prior to matching and is only applied to objects in a spatial field at a single time.

c. MRMS-object-tracking to estimate storm age

A novel aspect of our object-based approach is to include tracking or self-matching in time for MRMS objects (Fig. 2a). A simple tracking strategy allows for direct measurements of observed storm age relative to convection initiation. The high temporal resolution (5-min) of WoFS and MRMS output simplifies object-tracking by limiting the distance objects move between times so that a simple greatest-overlap search is implemented. Object labels in matched objects from the prior time are reassigned to current objects, and object "age" is the aggregate time from the first appearance of the object in the observational data through the observed lifetime of that object. As shown in Figure 2a, given the sufficiently small time increment (5 min. here), this method works robustly without the need for sophisticated reconstruction of trajectories or other assumptions.

d. Treatment of merging or splitting

The intent of this study is to isolate the relationship between storm prediction and its age measured from CI. As such, the tracking/matching algorithm does the following:

1. Mergers are accounted for in the clustering procedure described in section 2b so that an object that is incorporated into a cluster assumes the label of the cluster. Thus, all objects trace back to their respective CI.

2. Splitting objects are not relabeled in our algorithm in order to prevent the splitting from creating multiple new labels (e.g. Van der Walt et al. 2014) that would present a false signal of CI in objects that have potentially undergone several assimilation cycles.

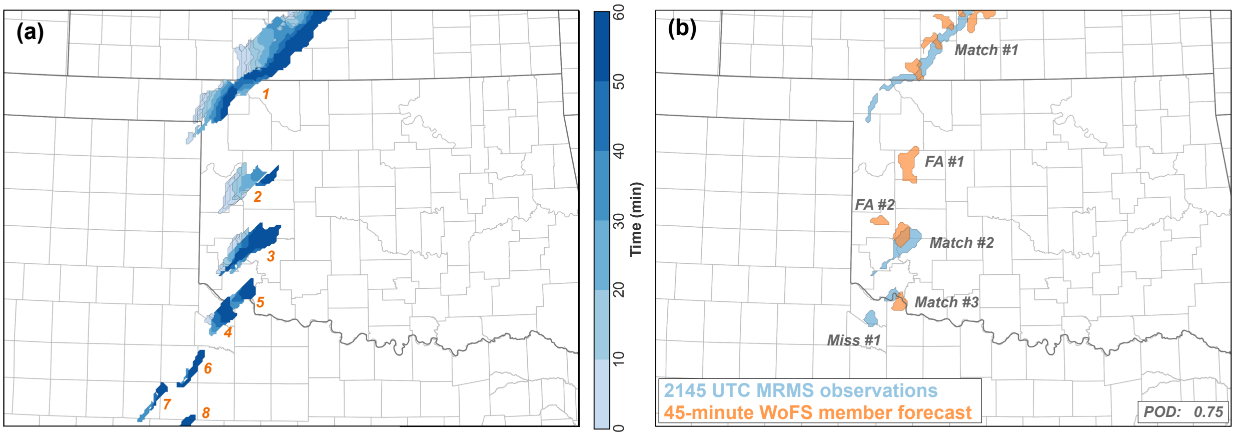

FIG. 2. (a): Diagnostic field of time-tracked MRMS objects on 2 May 2018 between 21Z and 22Z for storms in western Oklahoma. Forecast initialization is at 21Z. Increasing labels indicate younger storms. (b): Diagnostic field of tracked and matched objects on May 2, 2018 valid at 2145Z for storms in western Oklahoma. The Probability of Detection (POD) for the example is provided in the lower-right corner.

e. Object Matching of Model to Observations

Following quality control, spatial fields of WoFS and MRMS objects at a coincident time are matched based on a Total Interest score defined as,

$$TI_{\text{match}} = 0.5\left[\frac{(D_m - D_{min})}{D_m} + \frac{(D_m - D_{cnt})}{D_m}\right]. \tag{2}$$

where D_m is the constant distance threshold for matching, while D_{min} is the minimum boundary distance between a pair of storms, and D_{cnt} is the distance between object centroids. Both D_{min} and D_{cnt} are computed from each object pair being interrogated.

We note that the definition in equation 2 differs from Skinner et al. (2018) in that "time" proximity matching is not included. We use a single spatial distance threshold ($D_m = 40\ km$) for both minimum and centroid distance measures and a TI threshold of 0.2 to define a matched object pair. These thresholds are the same as Skinner et al. (2018) and were chosen to approximate the typical scale of an NWS warning product.

In this study we consider probability of detection (POD) as a simplified definition of skill. With reference to Figure 2b, the instantaneous POD is,

$$POD = \frac{MC}{OC}, \tag{3}$$

where MC is the number of identified matched objects and OC is the total number of observed objects. As such, POD is directly dependent on observed objects and thus their age, which provides the foundation for subsequent analysis. We note that our analysis is limited to POD by necessity since we need to be able to relate the age of an object to its existence in both observations and simulated fields. As a result, we do not consider false-alarm ratio (FAR) or other 2x2 contingency-table-based verification metrics as we are unable to associate false alarms with specific ages of observed storms, which, by definition of a false alarm, do not exist.

Lastly, we note that we performed several visual inspections of track/matching performance by generating lengthy animations over several cases from each year of labeled MRMS and WoFS objects superimposed on a spatial grid where Figure 2b represents and example of such a time snapshot. From these development studies, we were able to verify the time tracking of MRMS objects for age as well as the relabeling for matches when compared to forecast fields. In both cases we found the algorithm to have robust and physically consistent performance with the given parameters i.e. clustering and matching of objects was as expected for a given field and time without evidence of misplaced objects. Samples of such animations, as generated by the tracking/matching software, are made available through our code repository: `https://github.com/WarnOnForecast/WoFS_Verif2020.git`.

3. Results

a. Condensed tracked/matched database

Figure 2b shows a typical matched field for a single member at a particular valid time. This type graphical diagnostic output is instrumental in verifying the performance of the matching software given the various user-defined parameters. The results are stored in a Python dictionary file so that each object (model and observational) is accounted for with labels, times, areas, maximum intensity and other information. The purpose is to generate a condensed archive that is much smaller than the gridded fields processed. From this discrete database, analysis can then be performed by means

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-22-0043.1.

of slicing into the respective dictionary objects and applying discretionary filters as needed. An added benefit is that the resulting database is highly portable (on the order of a few Mb of memory).

b. Initial characterization of results

We begin our analysis with a basic characterization of the results database as specified above. There are 20 cases for 2017, 44 in 2018, 45 in 2019, 39 in 2020, and 29 in 2021. Each case is comprised of 8 hourly forecasts from 20UTC to 03UTC. This gives a total of 1416 forecasts considered in our study. Figure 3a shows that POD behavior with respect to age is a general feature of the WoFS and insensitive to changes that may have been applied to the system over the previous five years. Similarly, grouping the results by season, where cases driven by collaboration with the Storm Prediction Center (SPC) are in April and May while cases driven by collaboration with the Weather Prediction Center (WPC) are in June through September, also display similar behavior. The latter suggests that our results are also relatively insensitive to changes in storm mode or synoptic configuration. We do note a drop in maximum POD for WPC cases in figure 3a, but the behavior remains consistent. Here we present aggregate results for all lead times less than 30 minutes in order to demonstrate the trends more clearly within a critical time frame for WoFS guidance (0 to 30 minutes of lead time).

Our results also allow us to confirm an intuitive assumption that older storms objects are often larger than shorter-lived storms. Figure 3b demonstrates a strong positive and linear correlation between estimated object area and age. Evident in this result is the effect of the minimum area filter implemented in our tracking algorithm as objects of zero age have a finite area of approximately 250 km^2. We note the presence of 2 cases from 2020 where a large convective system was advected into the WoFS domain resulting in "young" but otherwise anomalously large storm objects.

c. Accuracy by object age and lead time

To analyze the impact of object age on forecast accuracy, ensemble-mean POD (computed as POD from eq. 3 arithmetically averaged over ensemble members at a given time) is shown as a function of object age and lead time in Figure 4. Here, the object age is computed as the true estimate of age for the object as measured from observations (which we refer to as “absolute object age”). For example, the orange line in Fig. 4b shows POD for a range of object ages 90 minutes into

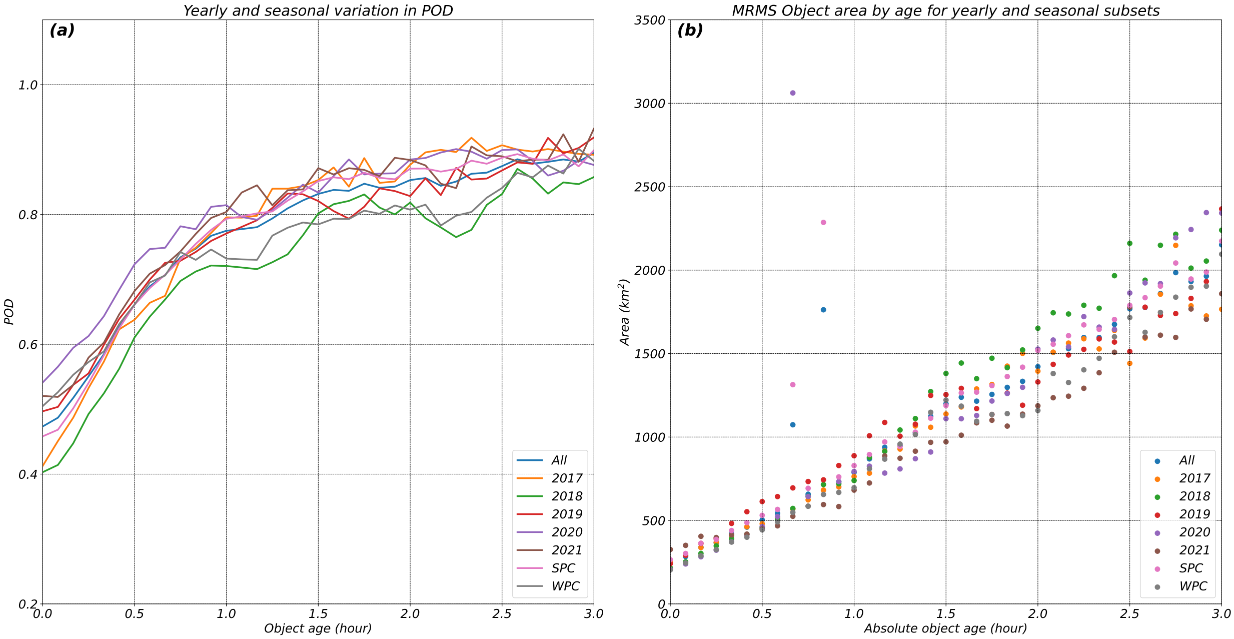

FIG. 3. (a): Ensemble mean probability of detection (POD) as a function of object age further averaged over all lead times less than 30 minutes and for various subsets (yearly and seasonal) of the forecast database. (b): Area and absolute age correlation in MRMS reflectivity objects over various subsets (yearly and seasonal) of the reference database. Seasonal subsets are denoted by SPC refer to cases that predominantly belong to the months of April and May while WPC corresponds to cases belonging to the months of June through September.

the forecast. For this orange line, an object age of 1 hour means that the object started 30 minutes after model initialization and is thus 1 hour old at 90 minutes into the forecast. Similarly, for the green line in Fig. 4a, which corresponds to 30-minute lead time, an object age of 1 hour means that the object started at 30 minutes before model initialization and is thus 1 hour old at 30-minute lead time. Because the dashed black curve is valid for the analysis time, it indicates approximately how many DA cycles are needed to skillfully spin up a storm. For a storm that first appears at the analysis time (i.e., object age of 0.0), the POD is very low at just above 0.2 [1]. However, the POD quickly rises with increasing object age, going from about 0.65 to 0.75 to just above 0.8 for object ages of 0.5, 1.0, and 1.5 hours, respectively, and then leveling off. Since there are four DA cycles per hour, we can say that it takes 4-6 DA cycles to reach the maximum skill in the analysis, which

[1] The WoFS reflectivity analysis is determined by the EnKF and will more closely match observed reflectivity values than reflectivity in forecasts, which are solely determined by the microphysical parameterization and are known to exhibit a high bias of CI. The very youngest objects that emerge within the first time step of a forecast cannot be accounted for in the analysis and will also be poorly located leading to the very low POD for objects of 0 min. of age.

is similar to prior estimates in idealized EnKF-based radar assimilation experiments (e.g. Tong and Xue 2005). This also illustrates a clear avenue for improving WoFS: for example, if any aspect of WoFS DA (e.g., more frequent cycling) could raise POD faster for younger objects, it would very likely improve the subsequent forecasts. The other lead times pictured in Fig. 4 show similar patterns of increasing PODs with increasing object ages. Furthermore, there are fairly uniform decreases in POD with later lead times, as would be expected because of increasing error growth.

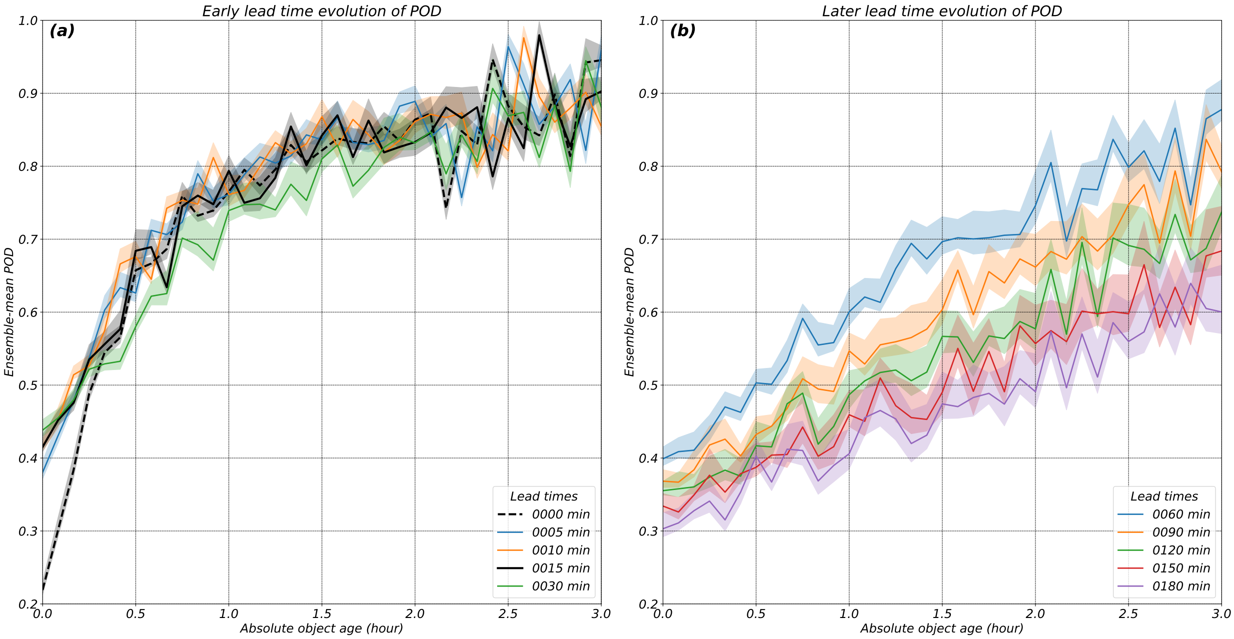

FIG. 4. Ensemble mean probability of detection (POD) as a function of object age relative to model lead time for lead times of (a): 0, 5, 10, 15, and 30 minutes, and (b): 60, 90, 120, 150, and 180 minutes. Shaded regions around each curve show the 95th percentile confidence level for each based on a resampled ensemble (18 member) distribution.

Another way to examine the impact of object age on model performance is by analyzing accuracy for object ages relative to WoFS initialization time (which we refer to as "relative object age") instead of lead time, which is done in Figure 5. Similar to Fig. 4b, the orange line in Fig. 5b shows ensemble mean POD for a range of object ages at 90 minutes into the forecast. However, with object age given at the initialization time ($ROA = OA - LT$), an object age of 1 hour now means that object started 1 hour before model initialization and would thus be 2.5 hours old at 90 minutes into the forecast. Relative object ages that are negative mean that the objects started after model initialization (e.g., -1 hour means that the object started 1 h after initialization). Conducting the analysis in this way allows us to better isolate the impact of analysis quality on subsequent forecasts. The main result, again, is that POD increases with increasing relative object age with less dependence on lead time as a consequence of re-scaling the object age. In other words, the object age at the time of model initialization appears to be the main driver of forecast quality rather than lead time. For example, for object ages 1.5 hours or greater, the PODs maintain a high level of skill in the range of about 0.7 to 0.9 all the way out to lead times of 180 minutes (3 hours).

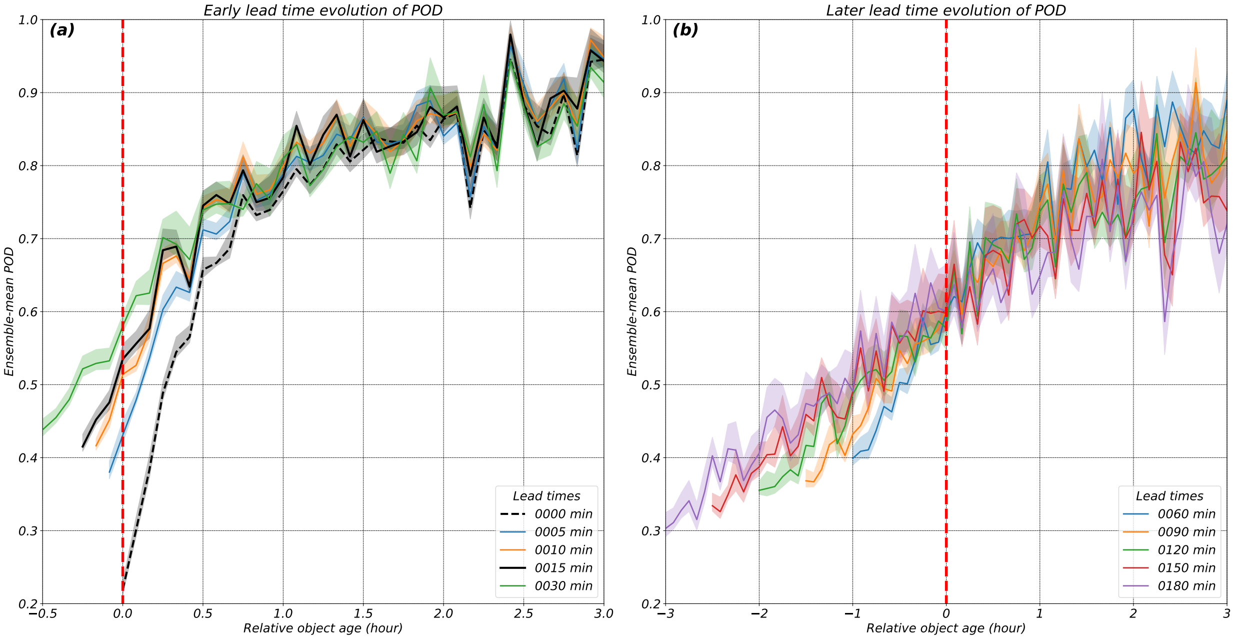

FIG. 5. Same as Figure 4, but for object age at the time of model initialization time $ROA = OA - LT$. The vertical dashed red dashed line marks a relative object age of 0.0; negative ROA corresponds to objects that are not assimilated while positive ROA is otherwise. Shaded regions around each curve show the 95th percentile confidence level for each based on a resampled ensemble (18 member) distribution.

We may interpret the data shown in Figure 5 as a surface in object age and lead time coordinates and further isolate the effect of model analysis on forecast performance, as in Figure 6a and b. These results show stratification with increasing POD curves for increasing relative object age bins i.e., longer-lived, larger, better-resolved convection as confirmed in Fig. 3. As a baseline of performance, objects that do not pass through DA are subject to a POD ≈ 0.45 lower bound (i.e., black dashed line in Figure 6a) and the averaged accuracy for all objects is POD ≈ 0.7 at a lead time of 1 hour. We provide the 95th percentile confidence intervals for each curve in Figure 6a and b where we observe that uncertainty is bounded and within 10% for each ROA bin, but grows with lead time for unassimilated objects (Figure 6a). Also, uncertainty in the total POD curve is relatively large due to the effect of including the non-analyzed contributions. It is evident from these results that a storm object that is at least 1 hour old at the time of initialization can be expected to be simulated with a POD > 0.7 through 3 hours of lead time; this is nearly double the skill for an equivalent object that is not assimilated and may very well be part of the same field of storms. Figure 6 clearly demonstrates the utility of the object-based approach for characterizing a forecast as constituent regions that are subject to independent skill, primarily controlled by the age of an object/storm as it is assimilated into the forecast. Furthermore, it provides a practical reference for what a user may expect in predictive skill when the current age of a storm is known.

To explain the behavior of POD for all objects, which gradually decreases with increasing lead time, Fig. 6c shows cumulative MRMS object counts at each lead time while Fig. 6d shows cumulative matched object counts. The total MRMS object counts are the denominator in the POD calculation, while the matched counts are the numerator. While the matched object counts for relative object ages greater than zero gradually decrease with lead time, the object counts for relative object ages less than zero increase until 1 h lead time and then remain fairly constant. Thus, as lead time increases the PODs for relative object ages less than zero have more weight, which gradually causes total POD to trend downward.

The relationship between object age and POD in Figures 3–5 is consistent across various subsets of the dataset, with variation of POD with object age much larger than variation between different years (and associated system configuration differences), different seasons (spring vs. summer), or different forecast initialization times (not shown). We note that the rapid increase in POD from 0 to 5 minutes of lead time in Figures 4 through 6b is primarily associated with the change from reflectivity values in the ensemble analysis produced by the EnKF to reflectivity produced by the microphysical parameterization in forecasts. Finally, we note that the decrease in total object counts with lead time evident in Figure 6b, c is a result of the majority of WoFS forecasts occurring after the diurnal convective maximum. Since we consider 3-hr forecasts initialized between 20 and 02 UTC, we include more forecast time after 00 UTC than before it. The additional forecast time after 00 UTC will more likely occur during a period of diminishing storm coverage owing to

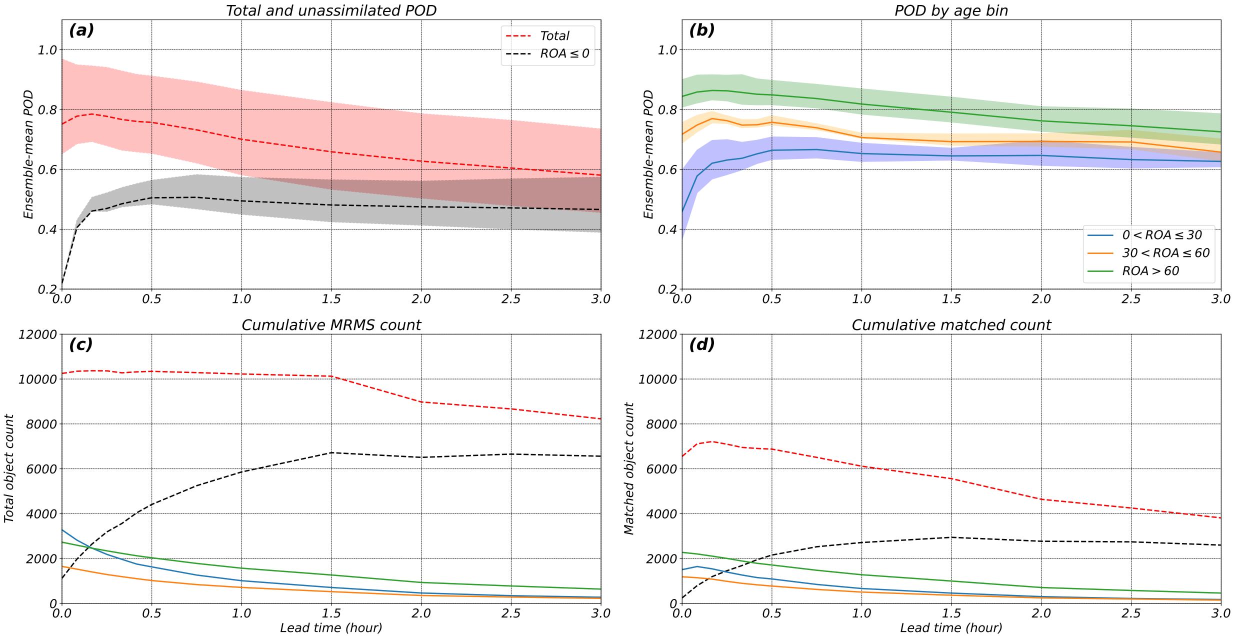

FIG. 6. (a): Probability of detection (POD) broken up into bins of relative object age with respect to lead time. The dashed black line corresponds to all objects of any age that started after model initialization (i.e., did not "pass through" the analysis). The dashed red line indicates the total POD for all objects at the specified lead time. (b): POD broken out by relative object age bins. Standard deviation shown as transparent filled areas for (a) and (b): (c): Total MRMS object counts at each lead time (i.e., denominator of POD), and (d): total matched MRMS object counts at each lead time (i.e., numerator of POD). ROA is age at the time of initialization. Shaded regions around each curve show the 95th percentile confidence level for each based on a resampled distribution with each age bin.

nocturnal stabilization or during upscale transition to a linear storm mode, which will generally produce fewer, larger reflectivity objects.

4. Conclusions

In this study we present a characterization of WoFS accuracy using an object-based verification approach encompassing hundreds of cases over 5 years of operation. We introduce a novel tracking/matching algorithm that exploits the high temporal resolution of the system, calculates storm age based on MRMS histories, and matches storm objects to forecast fields. Based on these data we are able to directly evaluate the skill, in terms of Probability of Detection, of WoFS as a function of the age of a storm object as it is assimilated into the system for a given forecast cycle. Our central finding is that storm objects that are at least one hour old at the time of assimilation enjoy POD > 0.7 through 3 hours of lead time with the greatest accuracy shown within 1 hour of lead time. This fits precisely with the design intent of WoFS in providing guidance on rapidly evolving severe weather. In contrast, objects of any age that are not assimilated into a forecast are likely to be simulated correctly at a rate of POD $\approx$ 0.45.

Our findings indicate that the age/maturity of a storm at the time of forecast initialization is the dominant factor that determines the performance of the system for that storm object as we are unable to find other variables that have a similar influence. In other words, the accuracy of short-term WoFS thunderstorm forecasts is primarily driven by the accuracy of the initial condition, consistent with prior idealized studies (Flora et al. 2018). A practical ramification of this finding is that that the skill of WoFS thunderstorm forecasts will generally improve dramatically following CI, resulting in variable forecast accuracy between different thunderstorms within the system domain in any given forecast.

The current study is limited to examining reflectivity objects as proxies for convective storms and we have not explored alternative proxies when applying our tracking/matching software. Nevertheless, because matching is a spatial proximity search, we are confident that reflectivity objects provide a robust and generalizable analysis. Thus, given a sufficiently robust dynamical core and physics architecture, the greatest benefit to short-term thunderstorm forecast quality is derived from improvements to the data assimilation system.

The current study does not examine WoFS skill relative to alternative short-term thunderstorm prediction methods. Specifically, future work should compare WoFS guidance to extrapolation-based prediction of observed thunderstorms. Such a comparison will enable quantification of potential gains in forecast accuracy from numeric prediction of storm processes over extrapolation (Hwang et al. 2015). Similarly, storm-age based analyses of WoFS proxies for thunderstorm-related hazards may be used as inputs, or as a baseline comparison for explicit thunderstorm hazard prediction by machine learning models trained on WoFS guidance (Flora et al. 2021; Clark and Loken 2022).

The results herein point towards several paths for improving the quality and potential value of WoFS thunderstorm guidance. The most straightforward path is accelerating the spin up of accurate thunderstorm analyses in WoFS through increased data assimilation frequency, improved data assimilation methods, or better use of available observations. Additionally, improved prediction of CI in WoFS will raise the POD for objects not present at model initialization. To that end, our expectation is that higher spatial resolution in WoFS will lead to a better representation of the conditions leading to CI. Therefore, improved observation and prediction of (near) storm environments, particularly of air mass boundaries often responsible for CI, will raise the overall quality of WoFS thunderstorm guidance. Finally, the dependence of WoFS forecast quality on thunderstorm age results in a unique challenge for end user interpretation, as each WoFS forecast can be considered multiple independent predictions of varying quality for thunderstorms within the system domain. Further research is needed to understand how end users assess confidence in WoFS predictions of individual storms and to develop guidance that can quantify the expected confidence in the accuracy of an ensemble thunderstorm forecast.

Acknowledgments. This research was funded by the NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement #NA21OAR4320204, U.S. Department of Commerce and the Warn-on-Forecast project. We thank Gerry Creager for managing the NSSL high-performance computing system used to run WoFS and David Dowell, Terra Ladwig, and Curtis Alexander for provdiding HRRRE initial and boundary conditions. This manuscript benefited from a thoughtful review by Eric Loken. Lastly, we would like to thank Noah Carpenter of the University of Oklahoma School of Meteorology for performing and independent review of our tracking software and providing fixes. All figures and analyses in this paper were produced using freely provided Anaconda python distribution and the Scikit-image, matplotlib, and basemap libraries.

Data availability statement. All datasets and software used for this study are stored on the NSSL high-performance computing server and the data are available upon request. Processing software for tracking and matching is hosted in the following: `https://github.com/WarnOnForecast/WoFS_Verif2020.git`

References

Anderson, J. L., and N. Collins, 2007: Scalable implementations of ensemble filter algorithms for data assimilation. *J. Atmos. Oceanic Technol.*, **59**, 1452–1463.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-22-0043.1.

Anderson, J. L., T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn, and A. Avellano, 2009: The Data Assimilation Research Testbed: A community facility. *Bull. Amer. Meteor. Soc.*, **90**, 1283–1296.

Burghardt, B. J., C. Evans, and P. J. Roebber, 2014: Assessing the predictability of convection initiation in the High Plains using an object-based approach. *Wea. Forecasting*, **29**, 403–418.

Burlingame, B. M., C. Evans, and P. J. Roebber, 2017: The influence of PBL parameterization on the practical predictability of convection initiation during the Mesoscale Predictability Experiment (MPEX). *Wea. Forecasting*, **32**, 1161–1183.

Bytheway, J. L., and C. D. Kummerow, 2015: Toward an object-based assessment of high-resolution forecasts of long-lived convective precipitation in the central U.S. *Journal of Advances in Modeling Earth Systems*, **7 (3)**, 1248–1264, https://doi.org/10.1002/2015MS000497.

Cai, H., and R. E. Dumais, 2015: Object-based evaluation of a numerical weather prediction model's performance through storm characteristic analysis. *Wea. and Forecasting*, **31**, 1451–1468.

Clark, A. J., R. G. Bullock, T. L. 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.

Clark, A. J., J. Gao, P. T. Marsh, T. Smith, J. S. Kain, J. C. Jr., M. Xue, and F. Kong, 2013: Tornado pathlength forecasts from 2010 to 2011 using ensemble updraft helicity. *Wea. Forecasting*, **28**, 387–407.

Clark, A. J., and E. D. Loken, 2022: Machine-learning-derived severe weather probabilities from a warn-on-forecast system. *IN REVIEW*.

Clark, A. J., and Coauthors, 2021: A real-time, virtual spring forecasting experiment to advance severe weather prediction. *Bulletin of the American Meteorological Society*, **102 (4)**, E814 – E816, https://doi.org/10.1175/BAMS-D-20-0268.1.

Clark, A. J. L. J., and Coauthors, 2020: A real-time, simulated forecasting experiment for advancing the prediction of hazardous convective weather. *Bulletin of the American Meteorological Society*, **101 (11)**, E2022 – E2024, https://doi.org/10.1175/BAMS-D-19-0298.1.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-22-0043.1.

Davis, C. A., B. G. Brown, R. Bullock, and J. Halley-Gotway, 2009: The method for object-based diagnostic evaluation (mode) applied to numerical forecasts from the 2005 nssl/spc spring program. *Weather and Forecasting*, **24 (5)**, 1252 – 1267, https://doi.org/10.1175/2009WAF2222241.1.

Davis, C. A., B. G. Brown, and R. G. Bullock, 2006a: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772–1784.

Davis, C. A., B. G. Brown, and R. G. Bullock, 2006b: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Mon. Wea. Rev.*, **134**, 1785–1795.

Dowell, D., and Coauthors, 2016: Development of a High-Resolution Rapid Refresh Ensemble (HRRRE) for severe weather forecasting. *Preprints, 28th Conf. Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., 8B.2.

Duda, J. D., and D. D. Turner, 2021: Large-sample application of radar reflectivity object-based verification to evaluate hrrr warm-season forecasts. *Weather and Forecasting*, **36 (3)**, 805 – 821, https://doi.org/10.1175/WAF-D-20-0203.1.

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.

Flora, M. L., C. K. Potvin, P. S. Skinner, S. Handler, and A. McGovern, 2021: Using machine learning to generate storm-scale probabilistic guidance of severe weather hazards in the warn-on-forecast system. *Monthly Weather Review*, **149 (5)**, 1535 – 1557, https://doi.org/10.1175/MWR-D-20-0194.1.

Flora, M. L., C. K. Potvin, and L. J. Wicker, 2018: Practical predictability of supercells: Exploring ensemble forecast sensitivity to initial condition spread. *Monthly Weather Review*, **146 (8)**, 2361 – 2379, https://doi.org/10.1175/MWR-D-17-0374.1, URL https://journals.ametsoc.org/view/journals/mwre/146/8/mwr-d-17-0374.1.xml.

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 (6)**, 1721 – 1739, https://doi.org/10.1175/WAF-D-19-0094.1.

Gallo, B. T., and Coauthors, 2017: Breaking new ground in severe weather prediction: The 2015 noaa/hazardous weather testbed spring forecasting experiment. *Weather and Forecasting*, **32 (4)**, 1541 – 1568, https://doi.org/10.1175/WAF-D-16-0178.1.

Gallo, B. T., and Coauthors, 2022: Exploring the watch-to-warning space: Experimental outlook performance during the 2019 spring forecasting experiment in noaa's hazardous weather testbed. *Weather and Forecasting*.

Gilleland, E., D. Ahijevych, B. Brown, and E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430.

Griffin, S. M., J. A. Otkin, S. E. Nebuda, T. L. Jensen, P. S. Skinner, E. Gilleland, T. A. Supinie, and M. Xue, 2021: Evaluating the impact of planetary boundary layer, land surface model, and microphysics parameterization schemes on cold cloud objects in simulated goes-16 brightness temperatures. *Journal of Geophysical Research: Atmospheres*, **126 (15)**, e2021JD034 709, https://doi.org/https://doi.org/10.1029/2021JD034709.

Griffin, S. M., J. A. Otkin, C. M. Rozoff, J. M. Sieglaff, L. M. Cronce, C. R. Alexander, T. L. Jensen, and J. K. Wolff, 2017: Seasonal analysis of cloud objects in the high-resolution rapid refresh (hrrr) model using object-based verification. *Journal of Applied Meteorology and Climatology*, **56 (8)**, 2317 – 2334, https://doi.org/10.1175/JAMC-D-17-0004.1.

Hewson, T. D., and H. A. Titley, 2010: Objective identification, typing and tracking of the complete life-cycles of cyclonic features at high spatial resolution. *Meteorological Applications*, **17 (3)**, 355–381, https://doi.org/https://doi.org/10.1002/met.204.

Houtekamer, P. L., and F. Zhang, 2016: Review of the ensemble kalman filter for atmospheric data assimilation. *Monthly Weather Review*, **144 (12)**, 4489 – 4532, https://doi.org/10.1175/MWR-D-15-0440.1.

Hwang, Y., A. J. Clark, V. Lakshmanan, and S. E. Koch, 2015: Improved nowcasts by blending extrapolation and model forecasts. *Weather and Forecasting*, **30 (5)**, 1201 – 1217, https://doi.org/10.1175/WAF-D-15-0057.1, URL https://journals.ametsoc.org/view/journals/wefo/30/5/waf-d-15-0057_1.xml.

Johnson, A., X. Wang, F. Kong, and M. Xue, 2013: Object-based evaluation of the impact of horizontal grid spacing on convection-allowing forecasts. *Mon. Wea. Rev.*, **141**, 3413–3425.

Johnson, A., X. Wang, 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. *Weather and Forecasting*, **35 (1)**, 169 – 191, https://doi.org/10.1175/WAF-D-19-0060.1.35.1.test.

Jones, T. A., K. Knopfmeier, D. Wheatley, G. Creager, P. Minnis, and R. Palikondo, 2016: Storm-scale data assimilation and ensemble forecasting with the NSSL Experimental Warn-on-Forecast System. Part 2: Combined radar and satellite data experiments. *Wea. Forecasting*, **31**, 297–327.

Jones, T. A., X. Wang, P. S. Skinner, A. Johnson, and Y. Wang, 2018: Assimilation of GOES-13 imager clear-sky water vapor (6.5 μm) radiances into a Warn-on-Forecast system. *Mon. Wea. Rev.*, **145**, In Review.

Jones, T. A., and Coauthors, 2020: Assimilation of goes-16 radiances and retrievals into the warn-on-forecast system. *Monthly Weather Review*, **148 (5)**, 1829 – 1859, https://doi.org/10.1175/MWR-D-19-0379.1.

Kain, J. S., and Coauthors, 2013: A feasibility study for probabilistic convection initiation forecasts based on explicit numerical guidance. *Bulletin of the American Meteorological Society*, **94 (8)**, 1213 – 1225, https://doi.org/10.1175/BAMS-D-11-00264.1.

Kerr, C. A., L. J. Wicker, and P. S. Skinner, 2021: Updraft-based adaptive assimilation of radial velocity observations in a warn-on-forecast system. *Weather and Forecasting*, **36 (1)**, 21 – 37, https://doi.org/10.1175/WAF-D-19-0251.1.

Kleist, D. T., D. F. Parrish, J. C. Derber, R. Treadon, W.-S. Wu, and S. Lord, 2009: Introduction of the gsi into the ncep global data assimilation system. *Weather and Forecasting*, **24 (6)**, 1691 – 1705, https://doi.org/10.1175/2009WAF2222201.1.

Mansell, E. R., C. L. Ziegler, and E. C. Bruning, 2010: Simulated electrification of a small thunderstorm with two-moment bulk microphysics. *J. Atmos. Sci.*, **67**, 171–194.

Miller, W. J. S., and Coauthors, 2022: Exploring the usefulness of downscaling free forecasts from the warn-on-forecast system. *Weather and Forecasting*.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-22-0043.1.

Mittermaier, M., R. North, A. Semple, and R. Bullock, 2016: Feature-based diagnostic evaluation of global nwp forecasts. *Monthly Weather Review*, **144 (10)**, 3871 – 3893, https://doi.org/10.1175/MWR-D-15-0167.1.

Mittermaier, M. P., and R. Bullock, 2013: Using mode to explore the spatial and temporal characteristics of cloud cover forecasts from high-resolution nwp models. *Meteorological Applications*, **20 (2)**, 187–196, https://doi.org/https://doi.org/10.1002/met.1393.

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.

Potvin, C. K., and Coauthors, 2019: Systematic comparison of convection-allowing models during the 2017 noaa hwt spring forecasting experiment. *Weather and Forecasting*, **34 (5)**, 1395 – 1416, https://doi.org/10.1175/WAF-D-19-0056.1.

Schwartz, C. S., G. S. Romine, K. R. Fossell, R. A. Sobash, and M. L. Weisman, 2017: Toward 1-km ensemble forecasts over large domains. *Mon. Wea. Rev.*, **145**, 2943–2969.

Skinner, P. S., L. 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.

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.

Sobash, R. A., C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Wea. Forecasting*, **31**, 255–271.

Stensrud, D. J., and Coauthors, 2009: Convective-scale Warn-On-Forecast system: A vision for 2020. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1499.

Stensrud, D. J., and Coauthors, 2013: Progress and challenges with Warn-On-Forecast. *Atmos. Res.*, **123**, 2–16.

Stratman, D. R., and K. A. Brewster, 2017: Sensitivities of 1-km forecasts of 24 may 2011 tornadic supercells to microphysics parameterizations. *Monthly Weather Review*, **145 (7)**, 2697 – 2721, https://doi.org/10.1175/MWR-D-16-0282.1.

Tong, M., and M. Xue, 2005: Ensemble kalman filter assimilation of doppler radar data with a compressible nonhydrostatic model: Oss experiments. *Monthly Weather Review*, **133 (7)**, 1789 – 1807, https://doi.org/10.1175/MWR2898.1.

Trojniak, S., and B. Albright, 2019: 2019 flash flood and intense rainfall experiment: Findings and results. Tech. rep., Weather Prediction Center. URL https://www.wpc.ncep.noaa.gov/hmt/Final_Report_2019_FFaIR.pdf.

Van der Walt, S., J. L. Schonberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, 2014: Scikit-image: Image processing in Python. *PeerJ*, **2**, e453.

Wheatley, D. M., K. H. Knopfmeier, T. A. Jones, and G. J. Creager, 2015: Storm-scale data assimilation and ensemble forecasting with the NSSL Experimental Warn-on-Forecast System. Part 1: Radar data experiments. *Wea. Forecasting*, **30**, 1795–1817.

Wilson, K. A., B. T. Gallo, P. Skinner, A. Clark, P. Heinselman, and J. J. Choate, 2021: Analysis of end user access of warn-on-forecast guidance products during an experimental forecasting task. *Weather, Climate, and Society*, **13 (4)**, 859 – 874, https://doi.org/10.1175/WCAS-D-20-0175.1.

Wolff, J. K., M. Harrold, T. Fowler, J. H. Gotway, L. Nance, and B. G. Brown, 2014: Beyond the basics: Evaluating model-based precipitation forecasts using traditional, spatial, and object-based methods. *Wea. and Forecasting*, **29**, 1451–1472.

Yussouf, N., and K. H. Knopfmeier, 2019: Application of the warn-on-forecast system for flash-flood-producing heavy convective rainfall events. *Quarterly Journal of the Royal Meteorological Society*, **145 (723)**, 2385–2403, https://doi.org/https://doi.org/10.1002/qj.3568.