An Objective Scoring Method for Evaluating the Comparative Performance of Automated Storm Identification and Tracking Algorithms

Clarice N. Satrio,[a,b] Kristin M. Calhoun,[b] P. Adrian Campbell,[a,b] Rebecca Steeves,[a,b] Travis M. Smith[a,b]

[a] *Cooperative Institute for Severe and High-Impact Weather Research and Operations, University of Oklahoma, Norman, Oklahoma*

[b] *NOAA / OAR National Severe Storms Laboratory, Norman, Oklahoma*

Corresponding author: Clarice N. Satrio, clarice.satrio@noaa.gov

Early Online Release: This preliminary version has been accepted for publication in *Weather and Forecasting*, may be fully cited, and has been assigned DOI 10.1175/WAF-D-22-0047.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT: While storm identification and tracking algorithms are used both operationally and in research, there exists no single standard technique to objectively determine performance of such algorithms. Thus, a comparative skill score is developed herein which consists of four parameters, three of which constitute the quantification of storm attributes — size consistency, linearity of tracks, and mean track duration — and the fourth which correlates performance to an optimal post-event reanalysis. The skill score is a cumulative sum of each of the parameters normalized from zero to one amongst the compared algorithms, such that a maximum skill score of four can be obtained. The skill score is intended to favor algorithms which are efficient at severe storm detection, i.e., high-scoring algorithms should detect storms that have higher current or future severe threat and minimize detection of weaker, short-lived storms with low severe potential. The skill score is shown to be capable of successfully ranking a large number of algorithms, both between varying settings within the same base algorithm and between distinct base algorithms. Through a comparison with manually-created user datasets, high-scoring algorithms are verified to match well with hand analyses, demonstrating appropriate calibration of skill score parameters.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-22-0047.1.

SIGNIFICANCE STATEMENT: With the growing number of options for storm identification and tracking techniques, it is necessary to devise an objective approach to quantify performance of different techniques. This study introduces a comparative skill score which assesses size consistency, linearity of tracks, mean track duration, and correlation to an optimal post-event reanalysis to rank diverse algorithms. This paper will show the capability of the skill score at highlighting algorithms which are efficient at detecting storms with higher severe potential, as well as those that closely resemble human-perceived storms through a comparison with manually-created user datasets. The novel methodology will be useful in improving systems which rely on such algorithms, for both operational and research purposes focusing on severe storm detection.

1. Introduction

Automated storm identification and tracking with remote sensing tools such as radar and satellite imagery has long been used in operations and research; the automated output from storm tracking algorithms provides guidance to National Weather Service forecasters about storm evolution traits like intensity, growth/decay, and motion trends, while also providing opportunities for data mining (e.g., Wilson et al. 1998; Lakshmanan and Smith 2009; Karstens et al. 2015, 2018). Its usefulness in nowcasting applications has resulted in increased attention in the research-to-operations community, such as the Probabilistic Hazard Information framework (PHI; Karstens et al. 2015), as a means of bridging the gap between watch and warning issuance. However, there does not exist a standard storm identification and tracking algorithm within the meteorological community; rather, a plethora of viable methods have been proposed in past literature and implemented operationally. Because each algorithm poses its own set of advantages and disadvantages, a method of quantification of performances between individual storm identification and tracking algorithms is necessary to identify key differences and understand how those differences relate to performance.

For storm identification and tracking algorithms, many different methodologies exist in how to determine the areal extent of a storm and how to track that storm over time. One of the first published storm identification and tracking algorithms called the Thunderstorm Identification, Tracking, and Nowcasting (TITAN) algorithm uses a single reflectivity threshold for storm identification (Dixon and Wiener 1993). Improvements to the TITAN algorithm were made by including an option for a dual-threshold, adjusting more optimally for mergers and splits, and estimating motion (Han et al.

2009) using the overlapping method (Moseley et al. 2013). Similarly, the Storm Cell Identification and Tracking (SCIT) algorithm — widely used in operations — employs the use of a reflectivity threshold. However, instead of a single reflectivity threshold, it loops through several reflectivity thresholds along radials to determine WSR-88D gates whose reflectivity values exceed that specific threshold (Johnson et al. 1998). These radial segments are then combined for each elevation to create 2D features for each reflectivity threshold which are then combined with other vertically contiguous features through multiple elevations to create a 3D snapshot of storm identification. However, SCIT produces only a centroid of the storm object and does not output a contour signifying the storm object bounds as TITAN does.

Because SCIT and TITAN are designed to be used in real-time operations, their workflow operates only with volumetric single-radar reflectivity scans. In contrast, the segmotion algorithm (w2segmotiondevll) within the Warning Decision Support System–Integrated Information suite (WDSS-II; Lakshmanan et al. 2007) uses multi-radar multi-sensor (MRMS) data (which is gridded onto a latitude-longitude-altitude grid) as input for storm identification; this algorithm employs an enhanced-watershed image segmentation rather than a predefined reflectivity threshold for storm identification (Lakshmanan et al. 2003, 2009). Beyond reflectivity-based algorithms, other tracking algorithms have been developed that use satellite variables (Schmetz et al. 1993; Raut et al. 2008; Kishtawal et al. 2009; Goswami and Bhandari 2012) or simulated meteorological data (Steiner et al. 1995; Raut et al. 2008; Heus and Seifert 2013). Furthermore, a more recent tracking algorithm called TINT Is Not TITAN (TINT) uses a tracking methodology which does not require the use of any physical storm traits (e.g., reflectivity) as part of the workflow (Raut et al. 2021).

A larger number of options arise when discussing the methodology of association of storms from one time step to the next. Most tracking algorithms use the cross-correlation (CC) method, centroid-based tracking, or the overlapping method. CC tracking attempts to match up two consecutive images (typically 2D reflectivity data) to obtain an accurate motion vector (Leese et al. 1971). Algorithms that use CC tracking (e.g., Tuttle and Foote 1990; Li et al. 1995) have the advantage of being able to track in stratiform (non-convective) precipitation — additionally, the CC method efficiently calculates the mean shift in images and thus can be used as a good first guess of mean motion (Schmetz et al. 1993; Kishtawal et al. 2009; Raut et al. 2021) as early tracking of storms is prone to larger errors (Johnson et al. 1998).

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-22-0047.1.

However, individual storms cannot be tracked purely using the CC method since it correlates entire images. Centroid-based tracking (e.g., SCIT or TITAN) calculates the center of the object based on the defined storm identification. Then, the algorithm attempts to match up storms from time t_{n-1} to t_n based on the forecasted centroid position using some method of optimization — for example, TITAN uses combinatorial optimization to find the shortest tracks using the given centroid positions. While this method bodes well for severe storm tracking with more intense storm cores, a disadvantage is potentially obtaining highly variable storm motion vectors as centroid location is dependent on storm shape, size, and strength, all of which can vary rapidly from one time step to the next. Increased stability in centroid-based tracking can be made using an “offline" approach in which only archived cases can be used. While “offline" algorithms such as the Thunderstorm Observation by Radar (ThOR) (Houston et al. 2015) have the obvious disadvantage of being operationally purposeless, the ability to use potential *future* positions to cluster objects into an identified track can improve accuracy and makes these algorithms useful for research-focused studies aiming to analyze past events. Lastly, as somewhat of a hybrid approach between centroid-based and CC tracking, tracking within segmotion correlates the current identified object backwards with the previous radar image using K-means clustering.

Thus, with many options for both storm identification and tracking, it is natural to attempt to develop a skill score to objectively compare various tracking algorithms. However, past studies have warned against developing such a “one-size-fits-all" skill score approach due to the inability of the skill score to adapt to different situations and user end goals (Lakshmanan and Smith 2010). While this sentiment is valid, it is possible to develop a skill score with the goal of extracting tracking algorithms that are optimized towards a specific use in mind. Specifically, this study aims to outline a skill score that will rank algorithms based on the ability to efficiently detect storms that are severe or have increased severe potential, i.e., high-scoring algorithms should not miss detection of longer-track severe storms while also minimizing false detection of weaker storms that are shorter-lived (less than ~15 min). The goal of extracting such algorithms is that these tracking algorithms are favorable to use in situations where potential severe objects are the main focus, i.e., within the PHI framework or within busy operational situations where forecaster burden could be reduced by objectively identifying only storms that have higher severe potential. Additionally, research focused on severe storms through data mining (e.g., obtaining storm statistics for climatologies) would benefit from tracking algorithms which are not prone to over-detection of weak, non-severe storms.

Specifically, the skill score quantifies the abilities of the algorithms to: 1) continuously identify objects, 2) correctly associate those objects between time steps, and 3) maintain consistency of the track and object geometry. Lakshmanan and Smith (2010) attempted to isolate characteristics that differentiate “good" versus “poor" tracking algorithms using specific parameters and allows the user to interpret the ranking of the tracking algorithms based on what the user deems important for the end goal of the tracking algorithm. Because this study defines the end goal of the tracking algorithm (optimized for severe storm detection), it is possible to modify and integrate parameters from Lakshmanan and Smith (2010) into one skill score.

This paper proposes an objective method to intercompare and rank the performance of a wide-variety of automated storm identification and tracking algorithms through a carefully-designed skill score formula. The skill score has two main parts: 1) quantification of three important characteristics of the track geometry and object shapes motivated from Lakshmanan and Smith (2010) and 2) comparison of the algorithm results to a post-event reanalysis which determines optimal tracks given the original output of the algorithm (best track Lakshmanan et al. 2015).

2. Methodology

There are two main parts of the skill score: first, quantification of track and object characteristics and second, a post-event reanalysis to determine optimal tracks. The quantification of object and track characteristics is implemented through an adaptation of a method developed in Lakshmanan and Smith (2010) using three characteristics that best represent the basic storm morphology: 1) consistency of the shape and size of the objects, 2) object duration, and 3) linearity of the track. The post-event reanalysis for the second part of the skill score will use *best track* as developed in Lakshmanan et al. (2015). The combination of these four parameters yields an objective skill score for determining comparative performances between algorithms and extracts information pertinent to understanding the performance score of an algorithm. The parameters yield a skill score that is calibrated to identify algorithms which are efficient at detecting objects associated with primarily long-lived, potentially severe storms (hereafter, “object" refers to the contour determined by a

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-22-0047.1.

particular tracking algorithm while a “storm" refers to a radar-detected echo whether or not it is identified by a tracking algorithm).

a. Object Consistency

The first and perhaps most intuitive algorithm characteristic to be quantified is the consistency of the identified objects' geometry. Object consistency is often based upon the preservation of a storm attribute. In Lakshmanan and Smith (2010), the standard deviation of maximum vertically integrated liquid (VIL) within objects identified for a particular track greater than the median duration was calculated and then averaged through all tracks. The method of Lakshmanan and Smith (2010) deemed an object to have high consistency when the standard deviation of VIL is low. However, in addition to the fact that maximum VIL through a storm's lifetime can vary significantly, maximum VIL inherently contains information from only one grid point within the object. Therefore, to capture more information about the consistency of the entirety of the object, we choose to calculate the absolute value percent change of the object size from one time step to the next, which is mentioned as a potential consistency attribute in Lakshmanan and Smith (2010). The average absolute value percent change in area for a track is calculated by

$$\Delta Area(\%) = \frac{\sum_i^{N-1} |A_{i+1} - A_i| / A_i}{N-1} \tag{1}$$

where A_i is the area of the object at the ith time step and N is the number of points for that track. Percent change in area is used over total change in area since algorithms whose objects tend to be larger would be penalized more for the same percentage growth — thus, total change was found to inherently score smaller-object-algorithms as better performing (not shown). This equation is an improvement from Lakshmanan and Smith (2010) for two reasons. First, the object size inherently contains information about the variability of the entire object which is previously not taken into account when using VIL. This ties into the second improvement in that while a storm attribute is expected to change rather drastically through its lifetime (especially for longer-lived storms), unlike standard deviation, this equation only correlates two consecutive time steps such that changes in storm attributes should be much smaller. Therefore, if a large absolute value percent change in area occurs, this can be attributed to object identification inconsistencies rather than storm evolution.

b. Duration

The second algorithm characteristic to quantify is the duration of the objects. In general, if the results for a particular algorithm contain higher durations, this implies that the algorithm was skilled in capturing longer-lived storms and was less susceptible to broken tracks. Lakshmanan and Smith (2010) utilized the median duration of all the identified objects as a way to quantify the duration; the median was chosen such that outliers of either very short and/or long tracks would not skew the data one way or another. However, in order to create separation between algorithms, very short-lived objects (15 min or less) and very long-lived objects (2 h or more) *should* weigh negatively and positively on the overall skill score, respectively. For example, if two algorithms had the same exact results except that the first algorithm correctly identified one long track and the second algorithm split the track in two, the mean duration would then rank the first algorithm above the second algorithm whereas the median would likely rank them equally. Thus, we modified the methodology from Lakshmanan and Smith (2010) to use the mean duration of all of the identified objects in order to quantify the second algorithm characteristic used in our skill score.

c. Linearity

The last algorithm characteristic to quantify is the linearity of the tracks with higher linearity correlating to better algorithm performance. The linearity of the track implicitly contains important information about the performance of object identification. First, a track with high linearity indicates that the object is not shifting around from one part of a storm to another — when a storm has a larger area of higher reflectivities, a poor algorithm may be prone to capturing only a portion of the storm at one time step and then jumping to another part of the storm at the subsequent time step, which would lead to a low linearity. A relevant example could be in an MCS case where an algorithm may capture a larger part of the line then jump to focus on an embedded core or vice versa, leading to an inconsistent track and low linearity.

In order to quantify the linearity, the root mean square error (RMSE) of each track within a particular algorithm is calculated, with RMSE simply defined as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} d_i}{N}} \tag{2}$$

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-22-0047.1.

where d_i is the distance between the object point and the best-fit line for that particular track and N is the number of points for that track — note that the “object point" is simply the unweighted centroid of the object. The RMSEs are calculated for all the tracks *longer than the mean duration* for that particular case, and then averaged to acquire an overall mean RMSE for each algorithm. Only the tracks longer than the mean duration are used to prevent calculating RMSE for very short tracks which typically influence RMSE to be much lower but do not have enough points to be statistically meaningful. RMSE calculations were not found to be significantly influenced by choice of reflectivity-weighted versus non-weighted centroids — in fact, RMSE performance was calculated to be slightly higher for reflectivity-weighted centroids. Thus, we are confident that using non-weighted centroids is sufficient and use of a weighted centroid would not improve upon the skill score.

d. Best Track Optimization

In addition to assessing properties of the individual tracks, it is also necessary to know how the algorithms perform compared to an optimal analysis. The most accurate optimal analysis would be one done by hand. However, hand analyses are extremely tedious and time-consuming. Because the objective of the skill score is to construct a way to *quickly* compare different tracking algorithms for a multitude of cases, including subsets of storm modes and environments, requiring the existence of a hand analysis to compare the tracking algorithm results for each case would detract from the purpose of the skill score. Therefore, we opt to use an automated method to extract the optimal tracks post-event, developed by Lakshmanan et al. (2015).

Lakshmanan et al. (2015) utilizes the initial object tracks outputted by an algorithm and computes the Theil-Sen (TS) slope for each uniquely identified track. It then analyzes each identified storm point using its centroid and compares its distance to all calculated TS slopes; if, for a storm object, there is a closer TS line than the one it is originally associated with, the object is then grouped with the track that has the closer TS line (Fig. 1a; see also Fig. 2 in Lakshmanan et al. 2015). Identical trajectories are combined, and new tracks with fewer than three time steps are either moved to the nearest trajectory within a specified spatiotemporal bounding box or removed if one does not exist (for more details, see Lakshmanan et al. 2015). Fig. 1b demonstrates the association of a candidate point to the nearest trajectory which, in the example, better aligns with that of hand analyses. A

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-22-0047.1.

couple modifications to the original best track steps are made here for better performance including: 1) defining the distance bounding box to be 0.1° latitude / longitude for which an object can be associated with a new track, and 2) automatically re-associating tracks with only one time step to the nearest track within the bounding box prior to pruning.

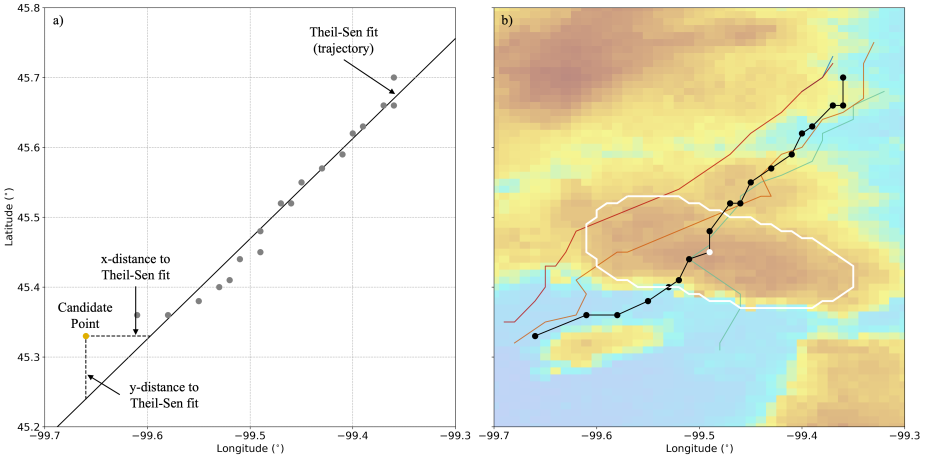

FIG. 1. Example of best track optimization using a Theil-Sen fit selected from the 7 June 2020 case: a) demonstrating a candidate point (yellow dot) being considered for inclusion in existing cluster / track (gray dots) and b) optimized track (black line / dots) with candidate point determined to be re-associated overlaid on hand analyses (red, orange, green, and blue lines; see Section 2f). Both the optimized track and the user tracks are valid from 232838 to 000637 UTC, with the white contour (dot) representative of the storm object outline (centroid) overlaid onto merged reflectivity QC composite at 234442 UTC.

After the best track reanalysis is complete, it is compared with the original algorithm output and scored based on a novel point-system:

- +1 points if the original object exists in the best track reanalysis and is correctly associated with the best track trajectory
- +0.5 points if the original object exists in the best track reanalysis but is re-associated to a different best track trajectory
- +0 points if the original object does not exist in the best track reanalysis (i.e., it is dropped)

The total points are then divided by the number of original identified storm objects such that a “perfect" algorithm in which best track reanalysis matches that of the original analysis would receive a score of one. It is important to understand that best track is only intended to improve the tracking of the original algorithm by fixing track breaks and pruning false detection of short tracks. While best track is intended to better align with a manual hand analysis, it is not in and of itself a tracking algorithm and cannot be used as “truth" due to the limitation that it cannot detect any new storms that the algorithm originally missed. Therefore, although best track outputs improved tracking results, it is limited by the original algorithm detection — nevertheless, best track is useful in that it allows for quantification of how well the algorithm performed in regards to track breaks as well as “false" detection of weaker short-lived storms. The quantification of these two properties through best track is unique from Lakshmanan and Smith (2010) and is important when discussing algorithms that are optimized in severe storm detection, as track breaks and false detections should be minimized.

e. Scoring System

All four of these parameters — object consistency, duration, linearity, and best track — are then combined to create an overall skill score for each of the tracking algorithms examined. For the first three characteristics, the algorithms are ranked from best to worst, i.e., from highest to lowest duration, lowest to highest linearity error, and lowest to highest percent change in area. For each of the three characteristics, the algorithms are then normalized in score from one to zero, with one being the best performing and zero being the worst. Thus, if an algorithm scored the best comparatively in all three categories, it would receive a one for object consistency, duration, and linearity, receiving a total score of three. Lastly, the best track reanalysis scores are also normalized and added to the three characteristic scores, giving a maximum score of four; or

$$\begin{aligned} Skill\ Score =& norm(area\ parameter) + norm(duration\ parameter) + \\ & norm(linearity\ parameter) + norm(best\ track\ parameter) \end{aligned} \tag{3}$$

where $norm(x)$ means that the score for that particular parameter, x, is normalized to one by the maximum score across all tracking algorithms. Because the scores are normalized to the set of algorithms, the final skill scores are always *relative* to the algorithms being inputted into the skill score and will change if different algorithms are compared. While an absolute skill score was considered, this would require the establishment of an absolute metric of each of the parameters

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-22-0047.1.

which is left to the user's discretion if desired. It is argued that the use of relative versus absolute scores would not vary results much — thus, we opt to use a algorithm-based normalized approach in order to prevent the need to establish set baselines for each parameter.

f. User Datasets

Though this skill score consists of four appropriate parameters that work to obtain the most efficient severe storm tracking and identification algorithms, we verify proper calibration of the skill score through the use of manually-created user datasets, discussed further in Section 3c. Proper calibration would entail user datasets scoring highly within the skill score, indicating similarity between high-scoring algorithms and hand analyses. The user datasets are created through a dynamic web-based mapping tool, where an individual can manually draw storm objects and track them from time step to time step with additional object evolution options such as merging and splitting (see Steeves et al. (2021) for more details). In addition to object control, the user also has access to multiple MRMS fields (reflectivity at -10°C, merged reflectivity QC composite, reflectivity at lowest altitude, etc.) which can be used jointly to aid in decision-making. Visualization of tracking for a particular storm for four different users is shown in Fig. 1b, illustrating that each user, though similar, have unique interpretations of storm tracking.

g. Limitations

While the skill score is undoubtedly a useful tool, Lakshmanan and Smith (2010) actually warns against the creation of such a score due to possible ambiguities in defining such a score across all situations, so there are limitations that the user must be aware of. The first limitation is that the parameters are being combined into a single skill score with equal weight. While assigning weights to the individual parameters was considered, establishing set parameter weights, perhaps based on the error corresponding to that particular parameter, would take a significant amount of trial-and-error and is beyond the scope of this study. Additionally, the skill score is easily amendable to a user's end goal — for example, if a user wanted to prioritize determining algorithms with the longest tracks, the weight for the duration parameter can be easily increased relative to the other three parameters. A second limitation which has been briefly mentioned above is that the skill score is calibrated to determine algorithms which are most efficient at identifying storms that have (or have already realized) severe potential — the skill score is not intended to be used in situations where detection of stratiform precipitation is desired, or in situations where the user aims to track all storms from first radar echo to decay (e.g., SCIT).

Thirdly, there is no definitive check within the skill score to determine whether an algorithm is over-identifying or under-identifying the number of tracks. Therefore, if an algorithm was to identify only the most intense, long-lived tracks, this algorithm would likely perform very well within the skill score; it would not be penalized for missing weaker, but still robust storms which may pose an eventual severe threat. Thus, subjective confirmation of the top-scoring algorithms by observation is recommended. Reflectivity thresholds could be utilized for a rough object count. For example, the total object count for an algorithm could be required to be less than the number of storms which have a reflectivity greater than 35 dBZ indicative of deep convective initiation (preventing over-identification) but smaller than the number of storms which have a reflectivity greater than 50 dBZ (preventing under-identification). However, the addition of such a check is left for future work. While the number of objects could also be assessed manually similar to Lakshmanan and Smith (2010), this would be time-prohibitive and would detract from the efficiency of the skill score. Lastly, there may be some concern in the way that the skill score is defined that algorithms that explicitly consider size or duration through a cost function (e.g., Morel et al. 1997; Lakshmanan et al. 2009; Han et al. 2009) will hold an unfair advantage against those that do not — it is argued, however, that these algorithms should be appropriately rewarded for these attributes inherent to their design and that these strengths in the algorithm should be reflected within the overall skill score.

Despite the limitations presented, the skill score serves an important purpose in that it has the ability to easily and objectively rank an abundance of algorithms and output those that are efficient in severe storm detection. While the method from Lakshmanan and Smith (2010) of assessing each parameter individually provides insights into strengths and weaknesses of each algorithm, it is impractical when assessing a large number of algorithms as it would be difficult for a human to discern which algorithm(s) perform the best overall (in Lakshmanan and Smith (2010), only six algorithms were compared). To illustrate this point, we have plotted each parameter as defined in Lakshmanan and Smith (2010) for 55 algorithms (to be detailed in Section 3a) including four user datasets for a single case (7 June 2020; Fig. 2) — however, for a better one-to-one comparison between the Lakshmanan and Smith (2010) and the skill score presented herein, we opt to use normalized standard deviation of size rather than VIL for object consistency as size was suggested as an alternative in Lakshmanan and Smith (2010). While Fig. 2 shows a good breakdown of performance by parameter of the 55 algorithms, it is unfeasible to objectively rank the algorithms shown just based on the parameter values; this becomes even more unfeasible when comparing hundreds or potentially thousands of different tracking algorithms. Thus, an example workflow could be as such: use the skill score defined here to narrow down many algorithms into the top-performing ones, then assess those top-performing algorithms by each parameter to assess trade-offs within those algorithms and choose one that is most optimal for user needs.

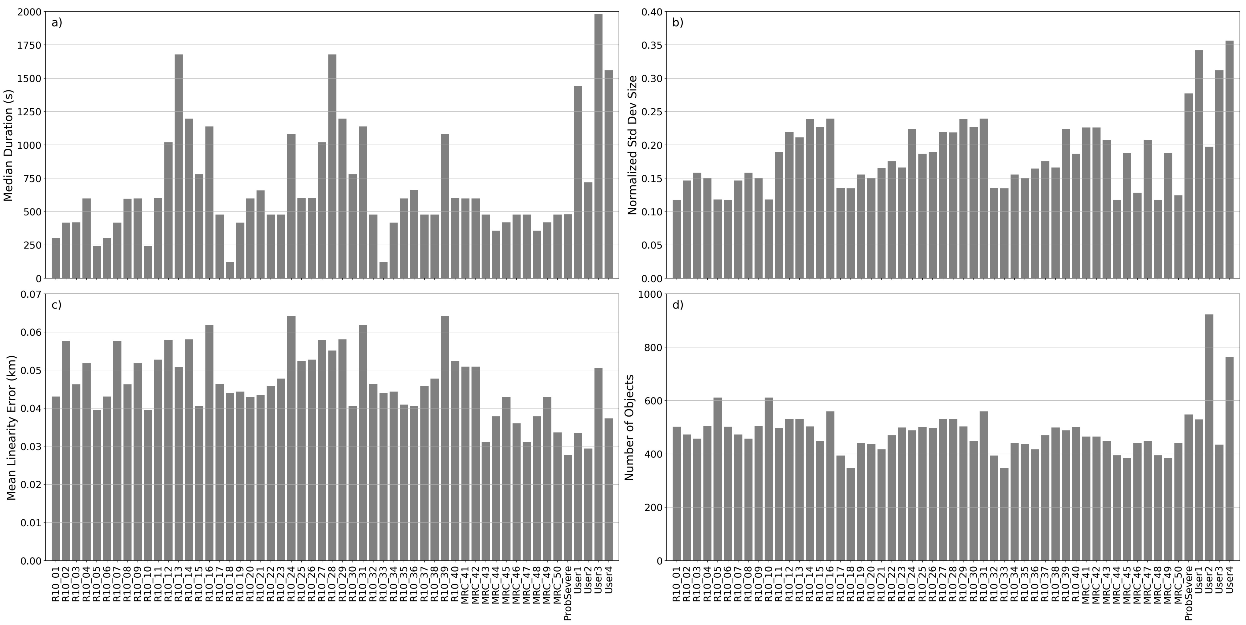

FIG. 2. Evaluation of tracking algorithms by parameter as described in Lakshmanan and Smith (2010) for the 7 June 2020 case: a) median duration of all tracks, b) normalized standard deviation of size for tracks longer than the median duration, c) mean linearity error for tracks longer than the median duration, and d) total number of identified objects. See Fig. 3 for nomenclature of tracking algorithms.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-22-0047.1.

3. Applications

a. Small-scale Modifications within a Base Tracking Algorithm

1) Experiment Setup

One application of the skill score is the ability to objectively, and quickly, differentiate performance between small-scale modifications within the same underlying tracking algorithm. For example, modifications to minimum size or reflectivity threshold results in similar, but unique, outputs. The skill score provides an objective way to rank performance between similar outputs. In this study, the base “adjustable" algorithm that is subjected to small-scale modifications is segmotion, an algorithm within the WDSS-II suite which uses K-means clustering and an enhanced-watershed method to identify storm objects at multiple scales (Lakshmanan et al. 2009) — while segmotion is used to demonstrate the capabilities of the skill score, this is not intended as a study for determining optimal segmotion settings. The segmotion algorithm consists of adjustable settings that can be altered to output distinct storm object datasets for a given case. The primary settings are tracking variable, data binning, pruner size, and smoothing filter. The segmotion algorithm also outputs objects at multiple spatial scales — while it is possible to combine multiple scales into one via a post-processing step with the intention of capturing processes outside of the storm core (as in Cintineo et al. 2020), this option was deemed unnecessary as 1) the skill score is focused on efficiency in severe storm detection and 2) segmotion is used simply as a proof-of-concept in how the skill score operates.

Initial experiments of 576 different segmotion setting combinations at the “0^{th}" and “1^{st}" spatial scale were run to determine how the settings modified the output. Due to the impracticality of showing all 576 settings throughout the study, this base understanding of setting behavior was then used to narrow the 576 setting combinations down to 50 segmotion settings within segmotion that performed acceptably. These 50 settings have variations in tracking variable, data binning, pruning size, and smoothing filter — full description of modifiable settings within segmotion can be found in Lakshmanan and Smith (2009); Lakshmanan et al. (2009); Lakshmanan and Smith (2010) and Appendix A in Cintineo et al. (2020). These settings will be denoted as text numbered 01 to 50, with the tracking variable specified prior. For example, R10_01 (MRC_41) would correspond to the 1^{st} (41^{st}) unique setting combination which had a tracking variable of reflectivity at -10°C

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-22-0047.1.

(merged reflectivity QC composite). The reflectivity data are sourced from the reprocessed Multi-Year Reanalysis of Remotely Sensed Storms project (Williams et al. 2021). This study evaluates these 50 settings in addition to the current segmotion configuration that is implemented within ProbSevere (hereafter denoted as “ProbSevere"; Cintineo et al. 2014, 2018, 2020) within the skill score.

The algorithms are run through four separate cases — 7 June 2020 (Complex), 24 February 2011 (QLCS), 23 March 2011 (Multicell), and 24 May 2011 (Supercell) — and are objectively scored based on the four skill score parameters detailed in Section 2. Each of the cases are restricted spatially and temporally such that convection within the domain is selectively representative of the convective system of interest or particular storm mode except for 7 June 2020 which is meant to be a multi-mode event (Table 1); for instance, the temporal window for 24 May 2011 ends before the supercells grow upscale to represent more QLCS organization. The 7 June 2020 case also contains the user datasets, so the algorithms will be verified against the user datasets for this case.

	Min Latitude (°)	Max Latitude (°)	Min Longitude (°)	Max Longitude (°)	Time (UTC)	Mode
20200607	42.6	46.9	-102.1	-98.2	2200 - 0008	Complex
20110224	29.5	38.2	-97	-83.1	2100 - 0600	QLCS
20110323	34.3	41.5	-89.5	-74	2200 - 0200	Multicell
20110524	32	38.93	-101.77	-95.52	1900 - 2230	Supercell

TABLE 1. Spatial and temporal domain specifications for four cases, each representing a specific storm mode.

2) 7 JUNE 2020 RESULTS

The case chosen to demonstrate the skill score application in detail is from 7 June 2020 from 2200 UTC to 0008 UTC, with the domain restricted to northern Nebraska into North Dakota (Table 1). The storms produced nearly 200 severe weather reports, including 12 tornado reports, primarily across central South Dakota into eastern North Dakota. Initial storms started as discrete supercells which eventually grew upscale by ~2300 UTC, characterized by clustered supercell organization. Fig. 3 shows ProbSevere scores poorly relative to the top 50 settings and user datasets, indicating that the skill score is successful in finding settings that perform better at identifying and tracking storms for this particular case. Specifically, ProbSevere tends to break tracks and over-identify stratiform precipitation regions, which are inconsistent, leading to an overall lower mean duration than the better-performing settings. Objects within ProbSevere are also susceptible to abrupt shape

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-22-0047.1.

changes between consecutive time steps, and thus, a larger mean size change. Both track and shape inconsistencies also yield a relatively low best track score, as many short-lived objects within ProbSevere are dropped in the post-event reanalysis.

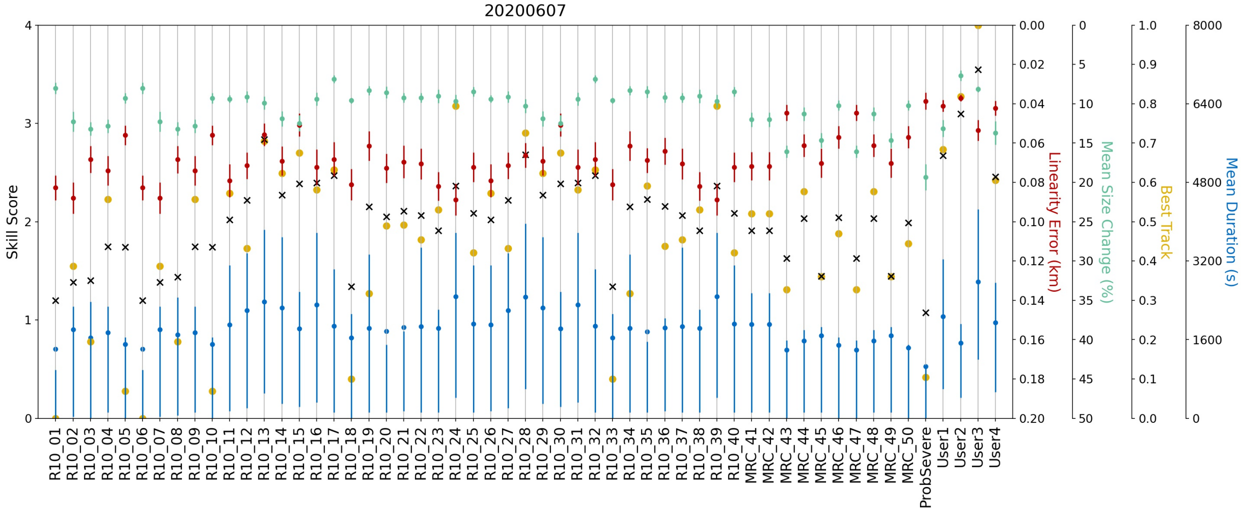

FIG. 3. Skill score and its four parameters for the top 50 settings, ProbSevere, and user datasets from the 7 June 2020 case. R10 and MRC are abbreviations which correspond to settings that use reflectivity at -10°C and merged reflectivity QC composite as their tracking variable, respectively. The skill score is marked by a black “X" with the value corresponding to the left y-axis. The four parameters — linearity error, mean percent size change, best track, and mean duration — are given by the red, green, yellow, and blue dots and associated error bars, respectively, with values indicated on the right y-axes. Error bars for linearity error and mean size change are given by $\mu \pm \alpha\sigma/N$, where μ is the mean linearity error / percent size change, the value of α comes from statistical tables of a two-tailed Student’s t distribution, the standard deviation given by σ, and N is the number of tracks longer than the mean duration. The error bars for duration are given by the interquartile range.

Fig. 4 visualizes differences in performance between the highest-scoring setting (R10_13) and ProbSevere for four times within the 7 June 2020 case. The first apparent difference between ProbSevere and R10_13 is the size of the objects; ProbSevere tends to capture more lower reflectivities surrounding the core of a storm and also identifies multiple reflectivity maxima within one object while R10_13 objects focus in on the core itself, with one reflectivity maxima (and updraft) being reflected by one object (e.g., Fig. 4d,h). There are notable differences in maximum reflectivity distributions between ProbSevere and R10_13 objects, with R10_13 eliminating objects with lower maximum reflectivity at -10°C (Fig. 5a). This disparity becomes more significant when including objects for the entirety of CONUS during a 24-hour period (7 June 2020 at 1200 UTC to 8 June 2020 at 1200 UTC, Fig. 5b); the maximum reflectivity distribution for ProbSevere has a secondary maxima at reflectivity at -10°C equal to 35 dBZ, indicating a tendency of ProbSevere to over-identify stratiform precipitation or storms which have not undergone deep convective initiation (35 dBZ, Roberts and Rutledge 2003; Mecikalski et al. 2008; Walker et al. 2012). These short-lived tracks not only reduce the mean age of tracks for ProbSevere, but also reduce its best track score as many of these tracks end up being dropped within the post-event reanalysis. Meanwhile, R10_13 successfully eliminates those weaker storms during this 24-hour period while still retaining storms with maximum reflectivity at -10°C and merged reflectivity QC composite greater than 50 dBZ (Fig. 5b).

As a result of the tendency of ProbSevere to over-identify weaker storms and lower-reflectivity stratiform regions, the mean duration of ProbSevere's objects are lowered as these storms tend to be shorter-lived and less steady-state (e.g., Fig. 4a,c,d). R10_13 significantly cuts down on identified objects that last one or two time steps (300 or 600 s) compared to ProbSevere, within the domain for the 2-h period (Fig. 5c) and within CONUS for the 24-h period (Fig. 5d). Shorter-lived storms are unlikely to be pose a threat to life and property, and thus would typically not need to be isolated for forecaster use in a system like PHI or uniquely monitored by forecasters for trends. For the 2-h period within the restricted domain, R10_13 reduces the percentage of tracks that last only one time step by nearly half and contains more tracks that last over 30 min, or 1800 s (Fig. 5c). Lastly, R10_13 focuses in on higher reflectivity regions which tend to be more consistent over time; this leads to less variation in object area and centroid placements between consecutive time steps resulting in higher track linearity, both of which weigh positively within the skill score.

3) Storm Mode

We can visualize changes in performance of each of the 50 settings plus ProbSevere by running the settings through the skill score for the first three cases in Table 1, each of which represent a singular storm mode (Fig. 6). It is apparent that the cases have drastically different performance based on storm mode; for example, merged reflectivity QC composite and reflectivity at lowest altitude settings perform equally to reflectivity at -10°C settings for the supercell case (Fig.6a) but suffer in performance for the QLCS case (Fig. 6b). This decline in performance is attributed to: 1) large

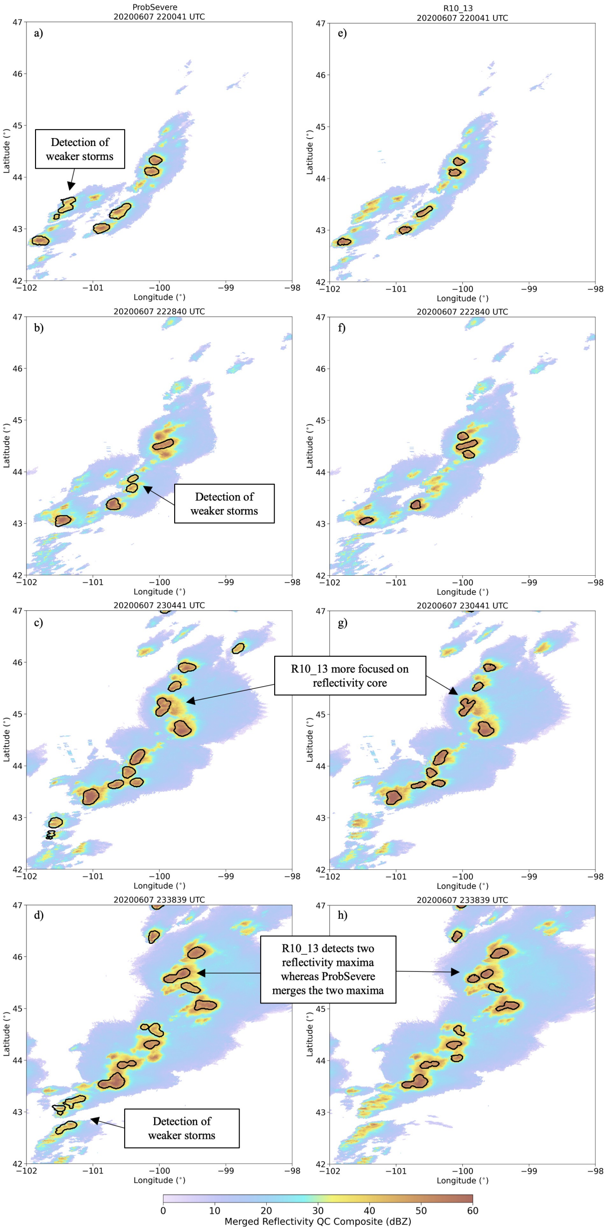
ProbSevere
20200607 220041 UTC
R10_13
20200607 220041 UTC
a)
e)
Detection of weaker storms
20200607 222840 UTC
b)
f)
Detection of weaker storms
20200607 230441 UTC
c)
g)
R10_13 more focused on reflectivity core
20200607 233839 UTC
d)
h)
R10_13 detects two reflectivity maxima whereas ProbSevere merges the two maxima
Detection of weaker storms
Latitude (°)
Longitude (°)
47
46
45
44
43
42
−102
−101
−100
−99
−98
0
10
20
30
40
50
60
Merged Reflectivity QC Composite (dBZ)

FIG. 4. Comparison of storm object selections (black contours) from ProbSevere (a-d) and R10_13 (e-h) overlaid onto merged reflectivity QC composite within the restricted temporal and spatial domain for the 7 June 2020 case for 220041, 222840, 230441, and 233839 UTC.

mean percent size changes and 2) worse best track performance comparative to reflectivity at -10°C settings, and 3) slightly shorter mean durations. This highlights that for optimal identification and tracking, there may not necessarily be a "one-size-fits-all" algorithm; rather, different algorithms

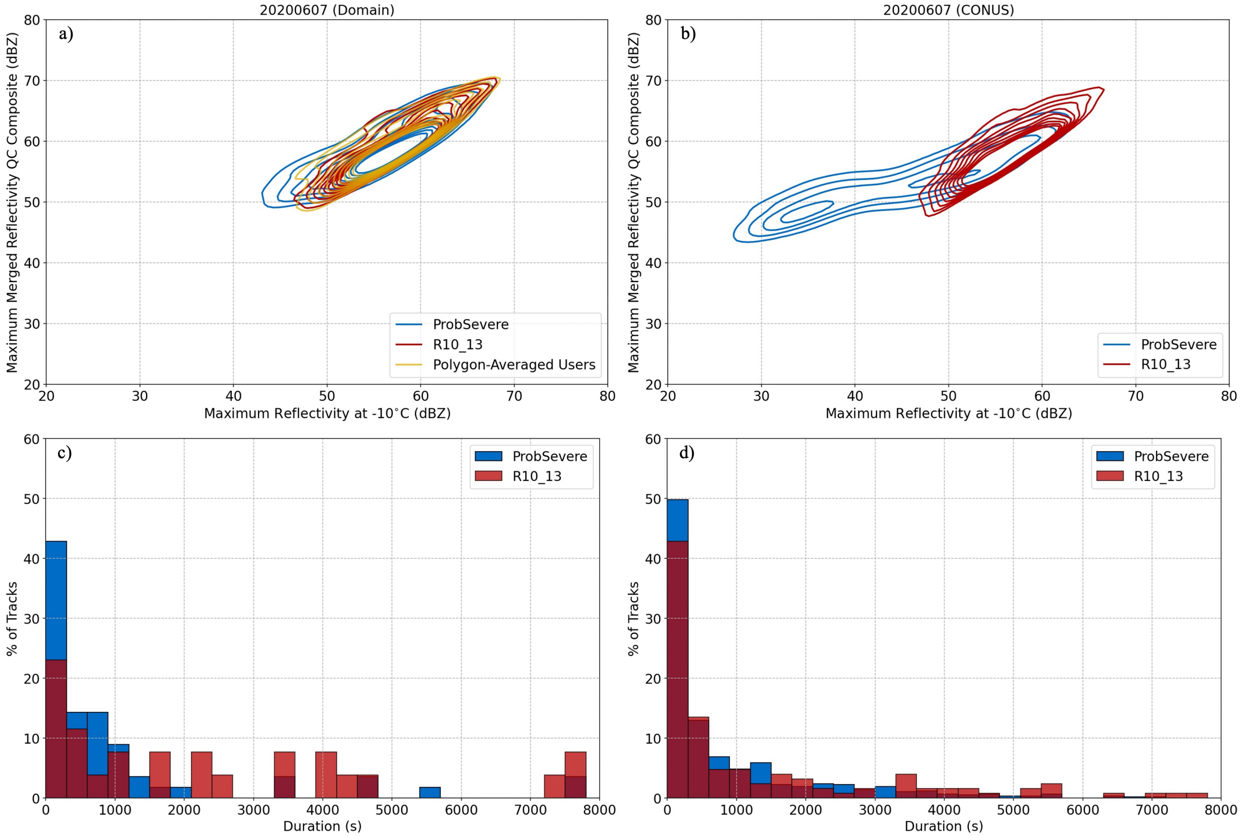

FIG. 5. Kernel Density Estimation (KDE) contour plots for the 7 June 2020 case for maximum reflectivity distributions between reflectivity at -10°C and merged reflectivity QC composite for a) restricted grid for ProbSevere, R10_13, and polygon-averaged user datasets and b) CONUS-wide grid for ProbSevere and R10_13 — contours start at 0.0015 and are plotted every 0.001. Histograms quantifying percent of tracks by duration for ProbSevere and R10_13 for the c) restricted grid and d) CONUS-wide grid.

may be more advantageous for certain situations than others, which the flexibility of the skill score can quickly deduce.

The skill score rankings also yield notable behaviors for the settings overall based on storm mode. Firstly, out of the three storm modes, tracks within the supercell case tend to have the highest overall duration by a significant amount while QLCS tracks have the lowest durations; multicell durations fall somewhere in the middle. This indicates that while the base algorithm is performing sufficiently at capturing long supercell tracks, it tends to break tracks associated with stronger segments within the QLCS as these are more transient and less steady-state compared to supercells. This is also reflected in the large size errors especially with merged reflectivity QC

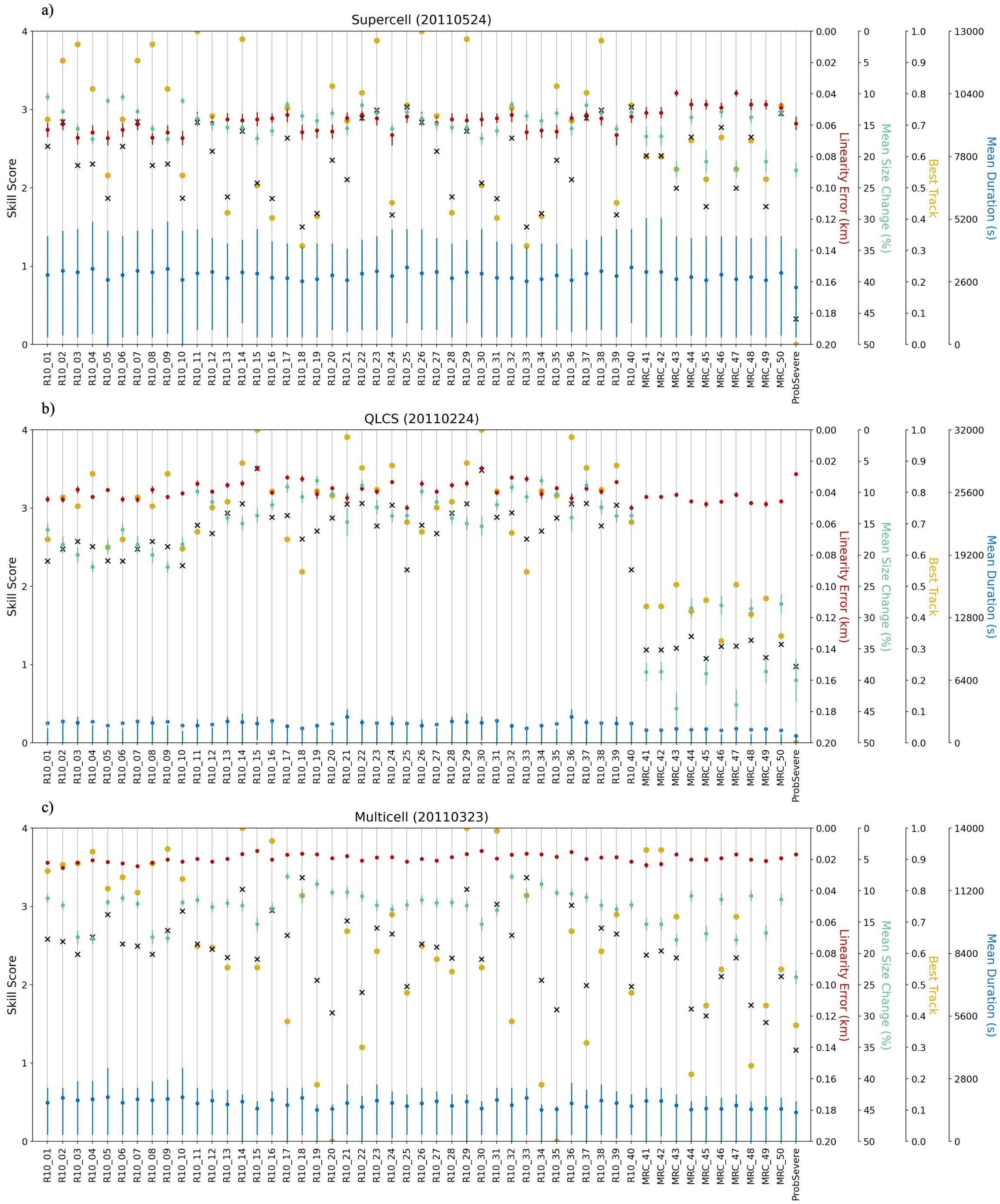

FIG. 6. Similar to Fig. 3 but for a) 20110524 (supercell), b) 20110224 (QLCS), and c) 20110323 (multicell) for the top 50 settings and ProbSevere.

composite QLCS objects; these settings are transitioning between capturing larger segments within the QLCS and smaller embedded cores, leading to large percent mean size changes. Overall, the

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-22-0047.1.

variation between the skill scores of the 51 settings is much smaller for the supercell case, as the interpretation of a storm “object" is more apparent compared to a QLCS or multicell case where different settings have more varying object interpretations.

b. Comparison between Tracking Algorithms

With the growing number of algorithms, some of which are operationally in-use, it is necessary to have the ability to identify advantages and disadvantages between different base tracking algorithms. As an example, this study tests the SCIT algorithm (Johnson et al. 1998) for the 24 May 2011 case. This case was chosen due to limitations of SCIT which is reliant on sufficient WSR-88D radar coverage for optimal performance; because the supercell event occurred in close proximity to KTLX, this case is ideal for a SCIT comparison. However, while SCIT uses full volume scans coupled with multiple reflectivity thresholds to obtain an accurate storm centroid, it does not calculate the boundaries of storm objects and therefore, storm size and mean size percent change cannot be calculated. Thus, two different comparisons against SCIT are presented to showcase the potential flexibility of the skill score to address user needs. In the first method, SCIT is assumed to have a perfect size consistency, since one can potentially assume that the object is simply a constant circle around the centroid which is representative of the storm object. In this method, the skill score remains the same as previously defined with a maximum score of four (Fig. 7a). However, because leaving the size consistency parameter within the skill score is not a like-for-like comparison with other algorithms (which must interpret the object boundaries), the second method eliminates the size parameter for all algorithms such that the maximum skill score becomes three and all algorithms are scored based only on their age, linearity, and post-event reanalysis, the latter two of which are dependent only on centroid placement (Fig. 7b).

Despite the different approaches, both methods highlight that SCIT does not perform well. In fact, SCIT ranks 51st out of 52 algorithms (including ProbSevere and SCIT) within the first method and last per the second method due to its comparatively *very* low mean duration versus segmotion settings. Investigating further, SCIT tends to over-identify weaker, unorganized storms which have a higher likelihood of dissipation. However, this should not come as a surprise as SCIT’s intended operational use is to identify *all* storms that have a potential of becoming organized — meanwhile, the skill score developed herein is intended to favor algorithms which identify only storms that have

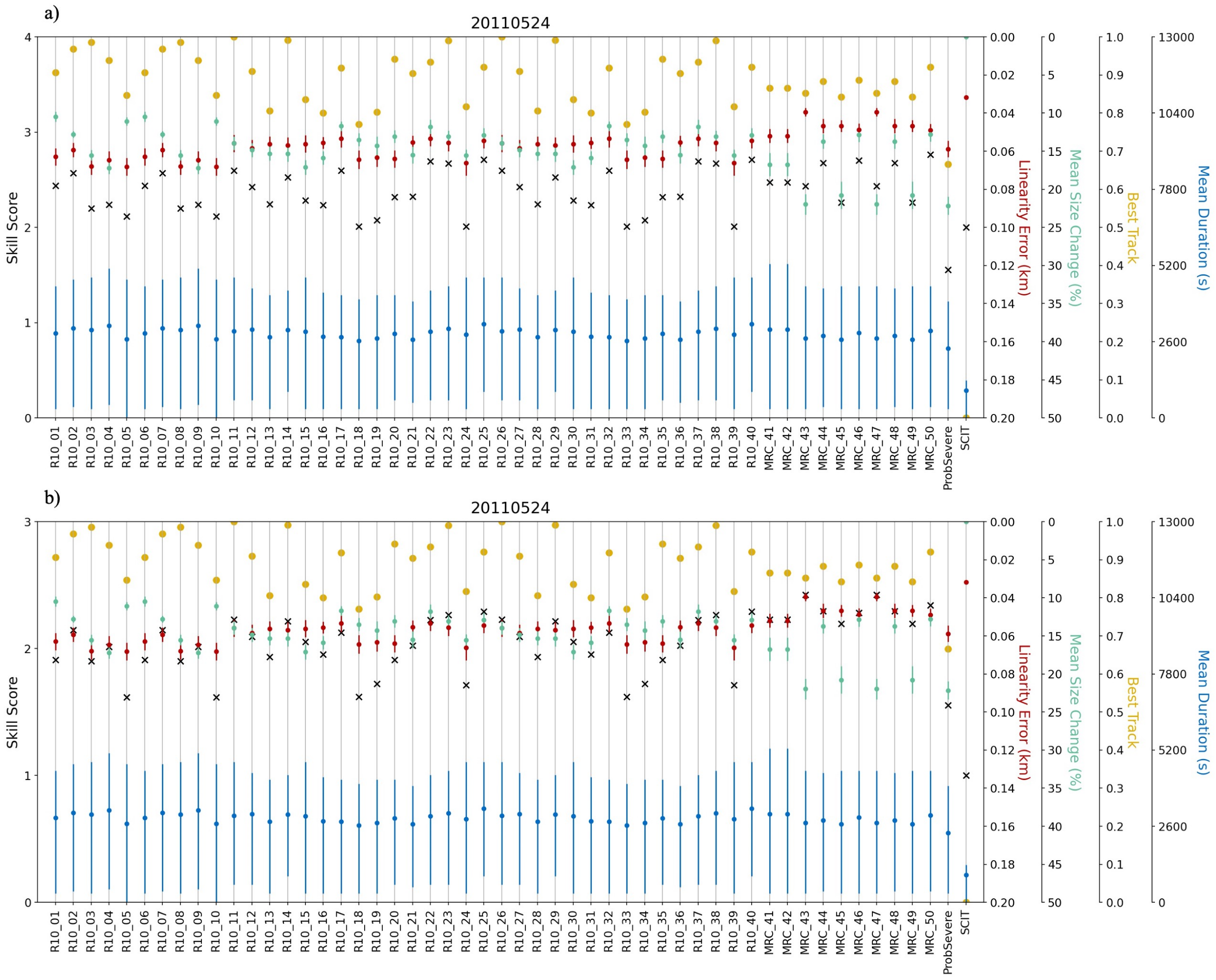

FIG. 7. Similar to Fig. 3 for the top 50 settings, ProbSevere, and SCIT for 20110524 using a) normalized mean percent size change of one for SCIT and b) eliminating the mean percent size change parameter giving a maximum skill score of three.

a higher probability of posing a future severe weather threat. Therefore, it is reasonable to expect that SCIT would perform poorly within the skill score, and its poor performance is actually a metric that the skill score is correctly calibrated to penalize algorithms that over-detect weaker storms. It is also worth noting that the parameter that SCIT scores very highly in is linearity error; because SCIT uses numerous reflectivity thresholds to obtain what is essentially a reflectivity-weighted centroid, the centroid placement along the track is less variable than the centroid calculated from the segmotion settings, which are obtained from the object shape itself and can be highly variable.

These subtle differences that the skill score detects can help identify strengths and weaknesses of different base tracking algorithms.

c. Verification with User Datasets

In order to ensure that the skill score is calibrated correctly, we opt to use hand analyses via user datasets to be scored against the 51 settings in Section 3a. To reiterate, it is expected that the user datasets will score highly within the skill score; if so, this indicates that high-scoring algorithms are more similar to storm objects that a human would interpret compared to those that rank lower. Four user datasets are completed for 7 June 2020, which is a case that was specifically chosen to represent a more complex event with evolving storm modes (supercells growing upscale into a QLCS).

The user datasets scored highly as expected, with User 3 and 2 scoring 1^{st} and 2^{nd} overall, respectively (Fig. 3). These datasets scored favorably because, in addition to longer mean durations, user datasets tracks have high linearity and a smaller mean size change as human perception of a storm object does not vary drastically in between consecutive time steps (120 s). To better consolidate the user datasets for comparison, a polygon-averaging technique is used; the four user objects are superimposed at each time step, and for a particular grid point, if the number of users that drew an object on that grid point is equal to or exceeds a subjectively-defined count, then that grid point is deemed to be within a “polygon-averaged" object (Fig. 8). The maximum reflectivity distribution for the users as an aggregate dataset (or ”Polygon-Averaged Users") is extracted to assess with R10_13 and ProbSevere (Fig. 5a). The Polygon-Averaged Users distribution closely resembles that of R10_13, which further confirms that 1) the top setting matches closely with that of user datasets and 2) R10_13 is correctly eliminating objects with lower maximum reflectivity at -10°C. Thus, the performance of the user datasets combined with nearly identical reflectivity distributions gives confidence that the skill score can distinguish high-scoring algorithms that are similar to what a human would perceive.

4. Conclusions

While there is an increasing number of storm identification and tracking techniques, there exists a lack of objective approaches to comparatively assess those differences in performance. Thus,

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-22-0047.1.

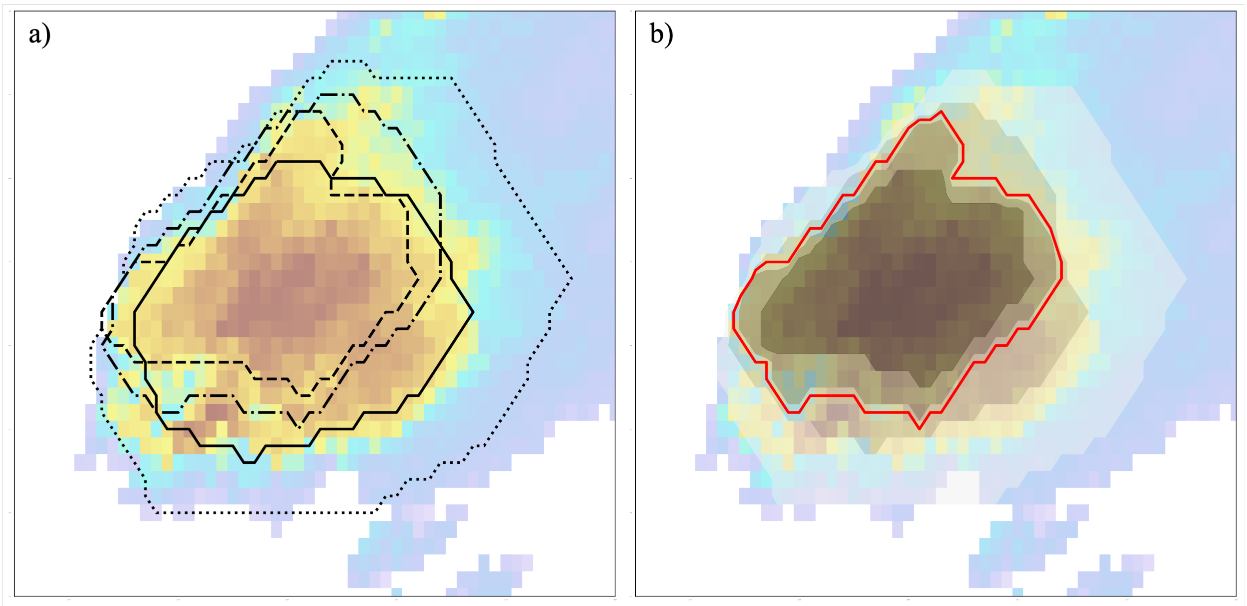

FIG. 8. Schematic of a) four individual user-drawn objects with each linestyle representative of a different user and b) gray-shaded contours illustrating overlapping grid points of user-drawn objects. A polygon-averaged object using a count of >2 is outlined in red.

this study introduces an objective skill score which comparatively ranks algorithm performance. Because building a one-size-fits-all skill score that works amongst all variations of situations and user needs is too broad, the skill score designed herein is intended to identify tracking algorithms that are efficient in severe storm detection. Specifically, the skill score is built as a means of extracting algorithms that detect the highest percentage of severe (or potentially future severe) storms while detecting the lowest percentage of those that are not. The development of such a skill score could be beneficial both operationally (e.g., within the PHI framework) and within severe storm research (e.g., for data mining severe storm statistics).

To quantify algorithm properties that encase efficient severe storm detection, the skill score consists of two main parts — the first is quantification of the object characteristics modified from Lakshmanan and Smith (2010) and the second part is a comparison to an optimal post-event reanalysis using best track (Lakshmanan et al. 2015). Quantification of object characteristics are further comprised of three parameters: 1) percent change of object size between consecutive time steps, 2) mean duration of tracks, and 3) average linearity error of tracks longer than the mean duration. Additionally, the novel best track score developed within the study aims to directly quantify the algorithm's ability to avoid track breaks and false detection of weaker short-lived storms with low severe potential.

Applications of the skill score are flexible to the user's focus and are shown to be successful at highlighting optimal components within a particular base tracking algorithm (e.g., segmotion) as well as comparing different base algorithms (e.g., segmotion versus SCIT). Algorithms that score highly are successfully shown to avoid detection of weaker storms which tend to be short-lived and do not pose a future severe threat, be less prone to breaking of tracks, and are less variable in terms of object size and shape. Additionally, algorithms which are geared towards identifying stratiform precipitation or those that track all storms from first radar echo (SCIT) are shown to score poorly. For further verification of skill score performance, reflectivity distributions of high-scoring algorithms match closely with (also high-scoring) manually-drawn user datasets for the 7 June 2020 multi-mode case, indicating that the methodology is correctly calibrated to identify algorithms which are comparable to what a human would perceive.

This study also highlights the effect storm mode has on algorithm performance through analysis of three various cases, each of which are spatially and temporally domained to represent a particular storm mode (supercell, QLCS, and multicell). Through the use of the skill score, it is found that there is not necessarily an optimal algorithm for all cases and storm modes. Rather, some algorithms may be more advantageous than others in certain situations, and the skill score is able to quickly deduce the appropriate algorithm for each case.

While this study mainly focuses on the methodology of the skill score and demonstrating the use / flexibility of it through four cases, future work includes conducting a more comprehensive study with additional cases and storm identification and tracking algorithms. This would allow for a detailed investigation on the advantages and disadvantages of different techniques which may have operational implications. Additionally, further work should be done using the polygon-averaging technique to create a (potentially) improved ensemble output as well as explore uses for the user datasets which provide invaluable verification. Last, improvements to the skill score itself, including establishing set objective weights to the four parameters as well as defining an option for an *absolute* normalization technique to provide an immutable score for each algorithm is left for future work.

Acknowledgments. Funding for this research was supported by NOAA / Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement #NA21OAR4320204, U.S. Department of Commerce, NOAA Weather Program Office Grant #NA18OAR4590386, and Joint Technology Transfer Initiative Grant "FACETs: Advancing physical and social science concepts toward operational implementation of Probabilistic Hazard Information." The authors thank Pat Hyland for his contribution to the user dataset. We also would like to thank John Krause for providing SCIT data as well as his and Jeff Snyder's helpful correspondence. Last, we thank Pat Skinner for reviewing a draft of the manuscript and providing insightful feedback in addition to John Cintineo and two other anonymous reviewers for their invaluable contributions. Some of the analyses in this study were conducted through OU Supercomputing Center for Education & Research (OSCER) at the University of Oklahoma.

Data availability statement. The MYRORSS data used in this study are provided by the University of Oklahoma and CIWRO and are available online at https://doi.org/10.15763/DBS.CIMMS.MYRORSS. The WDSS-II software is available online at http://www.wdssii.org/.

References

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, L. Cronce, and J. Brunner, 2020: NOAA ProbSevere v2.0 — ProbHail, ProbWind, and ProbTor. *Wea. Forecasting*, **35**, 1523–1543, https://doi.org/10.1175/WAF-D-19-0242.1.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, and D. T. Lindsey, 2014: An empirical model for assessing the severe weather potential of developing convection. *Wea. Forecasting*, **29**, 639–653, https://doi.org/10.1175/WAF-D-13-00113.1.

Cintineo, J. L., and Coauthors, 2018: The NOAA/CIMSS ProbSevere model: Incorporation of total lightning and validation. *Wea. Forecasting*, **33**, 331–345, https://doi.org/10.1175/WAF-D-17-0099.1.

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

Goswami, B., and G. Bhandari, 2012: Convective cloud detection and tracking from series of infrared images. *J. Indian Society of Remote Sensing*, **41**, 291–299, https://doi.org/10.1007/s12524-012-0234-3.

Han, L., S. Fu, L. Zhao, Y. Zheng, H. Wang, and Y. Lin, 2009: 3D convective storm identification, tracking, and forecasting—An enhanced TITAN algorithm. *J. Atmos. Oceanic Technol.*, **26**, 719–732, https://doi.org/10.1175/2008JTECHA1084.1.

Heus, T., and A. Seifert, 2013: Automated tracking of shallow cumulus clouds in large domain, long duration large eddy simulations. *Geosci. Model Dev.*, **6**, 1261–1273, https://doi.org/10.1175/1520-0450(1971)010<0118:AATFOC>2.0.CO;2.

Houston, A. L., N. A. Lock, J. Lahowetz, B. L. Barjenbruch, G. Limpert, and C. Oppermann, 2015: Thunderstorm observation by radar (ThOR): An algorithm to develop a climatology of thunderstorms. *J. Atmos. Oceanic Technol.*, **5**, 961–981, https://doi.org/10.1175/JTECH-D-14-00118.1.

Johnson, J. T., P. L. MacKeen, A. Witt, E. D. Mitchell, G. J. Stumpf, M. D. Eilts, and K. W. Thomas, 1998: The Storm Cell Identification and Tracking algorithm: An enchanced WSR-88D algorithm. *Wea. Forecasting*, **13**, 263–276, https://doi.org/10.1175/1520-0434(1998)013<0263:TSCIAT>2.0.CO;2.

Karstens, C. D., and Coauthors, 2015: Evaluation of a probabilistic forecasting methodology for severe convective weather in the 2014 Hazardous Weather Testbed. *Wea. Forecasting*, **30**, 1551–1570, https://doi.org/10.1175/WAF-D-14-00163.1.

Karstens, C. D., and Coauthors, 2018: Development of a human–machine mix for forecasting severe convective events. *Wea. Forecasting*, **33**, 715–737, https://doi.org/10.1175/WAF-D-17-0188.1.

Kishtawal, C., S. Deb, P. Pal, and P. Joshi, 2009: Estimation of atmospheric motion vectors from Kalpana-1 imagers. *J. Appl. Meteor. Climatol.*, **48**, 2410–2421, https://doi.org/10.1175/2009JAMC2159.1.

Lakshmanan, V., B. Herzog, and D. Kingfield, 2015: A method for extracting postevent storm tracks. *J. Appl. Meteor.*, **54**, 451–462, https://doi.org/10.1175/JAMC-D-14-0132.1.

Lakshmanan, V., K. Hundl, and R. Rabin, 2009: An efficient, general-purpose technique for identifying storm cells in geospatial images. *J. Atmos. Oceanic Technol.*, **26**, 523–537, https://doi.org/10.1175/2008JTECHA1153.1.

Lakshmanan, V., R. Rabin, and V. DeBrunner, 2003: Multiscale storm identification and forecast. *J. Atm. Res.*, **67**, 367–380, https://doi.org/10.1016/S0169-8095(03)00068-1.

Lakshmanan, V., and T. Smith, 2009: Data mining storm attributes from spatial grids. *J. Atmos. Oceanic Technol.*, **26**, 2353–2365, https://doi.org/10.1175/2009JTECHA1257.1.

Lakshmanan, V., and T. Smith, 2010: An objective method of evaluating and devising storm-tracking algorithms. *Wea. Forecasting*, **25**, 701–709, https://doi.org/10.1175/2009WAF2222330.1.

Lakshmanan, V., T. Smith, G. Stumpf, and K. Hondl, 2007: The Warning Decision Support System-Integrated Information. *Wea. Forecasting*, **22**, 596–612, https://doi.org/10.1175/WAF1009.1.

Leese, J. A., C. S. Novak, and B. B. Clark, 1971: An automated technique for obtaining cloud motion from geosynchronous satellite data using cross correlation. *J. Appl. Meteor.*, **10**, 118–132, https://doi.org/10.1175/1520-0450(1971)010<0118:AATFOC>2.0.CO;2.

Li, L., W. Schmid, and J. Joss, 1995: Nowcasting of motion and growth of precipitation with radar over a complex orography. *J. Appl. Meteor.*, **34**, 1286–1300, https://doi.org/10.1175/1520-0450(1995)034<1286:NOMAGO>2.0.CO;2.

Mecikalski, J. R., K. M. Bedka, S. J. Paech, and L. A. Litten, 2008: A statistical evaluation of goes cloud-top properties for nowcasting convective initiation. *Mon. Wea. Rev.*, **136**, 4899–4914, https://doi.org/10.1175/2008MWR2352.1.

Morel, C., F. Orain, and S. Senesi, 1997: Automated detection and characterization of MCS using the meteosat infrared channel. *Proc. Meteorological Satellite Data Users Conf.*, EUMETSAT, Ed., Brussels, Belgium, 213–220.

Moseley, C., P. Berg, and J. O. Haerter, 2013: Probing the precipitation life cycle by iterative rain cell tracking. *J. Geophysical Research*, **118**, 13 361–13 370, https://doi.org/10.1002/2013JD020868.

Raut, B. A., R. Jackson, M. Picel, S. M. Collis, M. Bergemann, and C. Jakob, 2021: An adaptive tracking algorithm for convection in simulated and remote sensing data. *J. Appl. Meteor. Climatol.*, **4**, 513–526, https://doi.org/10.1175/JAMC-D-20-0119.1.

Raut, B. A., R. N. Karekar, and D. M. Puranik, 2008: Wavelet-based technique to extract convective clouds from infrared satellite images. *IEEE*, **5**, 328–330, https://doi.org/10.1109/LGRS.2008.916072.

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

Schmetz, J., K. Holmlund, J. Hoffman, B. Strauss, B. Mason, A. K. V. Gaertner, and L. V. D. Berg, 1993: Operational cloud-motion winds from meteosat infrared images. *J. Appl. Meteor.*, **32**, 1206–1225, https://doi.org/10.1175/1520-0450(1993)032<1206:OCMWFM>2.0.CO;2.

Steeves, R. B., P. A. Campbell, K. M. Calhoun, and T. M. Smith, 2021: Establishing a truth dataset of storm objects using a web-based mapping tool. *37th Conference on Environmental Information Processing Technologies*, A. M. Soc., Ed., Virtual Conf.

Steiner, M., R. Houze, and S. E. Yuter, 1995: Climatological characterization of three-dimensional storm structure from operational radar and rain gauge data. *J. Appl. Meteor.*, **34**, 1978–2007, https://doi.org/10.1007/s12524-012-0234-3.

Tuttle, J. D., and G. B. Foote, 1990: Determination of the boundary layer airflow from a single Doppler radar. *J. Atmos. Oceanic Technol.*, **7**, 218–232, https://doi.org/10.1175/1520-0426(1990)007<0218:DOTBLA>2.0.CO;2.

Walker, J. R., W. M. MacKenzie, J. R. Mecikalski, and C. P. Jewett, 2012: An enhanced geostationary satellite-based convective initiation algorithm for 0–2-h nowcasting with object tracking. *J. Appl. Meteor. Climatol.*, **51**, 1931–1949, https://doi.org/10.1175/JAMC-D-11-0246.1.

Williams, S. S., K. L. Ortega, T. M. Smith, and A. E. Reinhart, 2021: Comprehensive radar data for the contiguous United States: Multi-Year Reanalysis of Remotely Sensed Storms. *Bull. Amer. Meteor. Soc.*, **EOR**, https://doi.org/10.1175/BAMS-D-20-0316.1.

Wilson, J., N. A. Crook, C. K. Mueller, J. Z. Sun, and M. Dixon, 1998: Nowcasting thunderstorms: A status report. *Bull. Amer. Meteor. Soc.*, **79**, 2079–2099, https://doi.org/10.1175/1520-0477(1998)079<2079:NTASR>2.0.CO;2.