

An Objective Method of Evaluating and Devising Storm Tracking Algorithms

Valliappa Lakshmanan^{1,2*}; Travis Smith^{1,2}

*Corresponding author: V Lakshmanan, 120 David L. Boren Blvd, Norman OK 73072; lakshman@ou.edu

¹Cooperative Institute of Mesoscale Meteorological Studies, University of Oklahoma; ²National Oceanic and Atmospheric Administration / National Severe Storms Laboratory

ABSTRACT

Although storm tracking algorithms are a key ingredient of nowcasting systems, evaluation of storm tracking algorithms has been indirect, labor intensive or non-specific. In this paper, we introduce a set of easily computable bulk statistics that can be used to directly evaluate the performance of tracking algorithms on specific characteristics. We apply the evaluation method to a diverse set of radar reflectivity data cases and note the characteristic behavior of five different storm tracking algorithms proposed in the literature and now employed in widely used nowcasting systems. Based on this objective evaluation, we devise a storm tracking algorithm that performs consistently and better than any of the previously suggested techniques.

1. Introduction

Algorithms that can extract properties of storm cells¹ and track those properties over time provide information that is important to forecasters in assessing storm intensity, growth and decay (Wilson et al. 1998). However, associating storm cells across frames of remotely sensed images poses a difficult problem because storms evolve, split and merge. Because storm tracking algorithms are a key component of nowcasting systems, the problem of how to track storms has received a lot of attention by the research community. Several criteria for associating storm cells across time have been suggested in the literature: using extent of overlap (Morel et al. 1997), using projected centroid location (Johnson et al. 1998), minimizing a global cost function (Dixon and Wiener 1993), greedy optimization of position error and longevity (Lakshmanan et al. 2009) and checking overlap followed by a global cost function (Han et al. 2009). Preprocessing operations such as median filters (Stumpf et al. 2005), quality control (Lakshmanan et al. 2007) and morphological operations (Han et al. 2009) have also been suggested as possibly improving the trackability of storm cells. It is important to be able to objectively evaluate these suggested techniques in order to determine which criterion or set of criteria provide the best skill.

a. Evaluating Storm Tracking Algorithms

One approach to evaluating storm tracking algorithms is to use the tracking algorithm to create a short-term forecast and then compare the short-term forecast with actual data (Lak-

¹For the purposes of this study, storm cells are defined on the basis of their radar reflectivity values as reflectivity peaks greater than 30 *dBZ* and having a size greater than 20 *km*².

shmanan et al. 2003). However, this is an *indirect* measure of storm tracking effectiveness since there is no way to separate out the effects of storm tracking from that of storm evolution. As pointed out by Wilson et al. (1998), the key reason for poor extrapolation forecasts is not errors in forecast displacement, but the growth and decay of storms in the forecast period. Indeed, because Han et al. (2009) employed extrapolation to compare the skill of their enhanced TITAN² tracking technique to the original TITAN technique, any improvement in forecast displacement was swamped by errors due to growth and decay. Consequently, Han et al. (2009) were able to demonstrate only a limited improvement provided by their morphological preprocessing and enhanced association algorithm. In this paper, we will use a better way of evaluating storm tracking to show that the enhancements suggested by Han et al. (2009) do make a significant improvement to the association algorithm of TITAN.

A more direct way of measuring the performance of the storm tracking component of storm identification and tracking algorithms was carried out by Johnson et al. (1998). A "percent correct" of time associations was computed by comparing the automated association of cells with a human association. This method suffers from three serious flaws:

1. Human association of storm cells is extremely *labor intensive* and time consuming. Indeed, even though the storm cell identification algorithm of Johnson et al. (1998) was evaluated on 17 cases, the associated storm tracking algorithm was evaluated on just 4 of those cases. Yet, even these four cases involved human truthing of 750 time associations. Also, the effect of fatigue and day-to-day variability on the quality of human associations can not be easily disregarded.

²Thunderstorm Identification, Tracking, Analysis and Nowcasting, a key component of the nowcasting system developed at the National Center for Atmospheric Research

2. The resulting skill is a gross *overestimate*. For example, on the four cases that the projected centroid-based storm tracking method of Johnson et al. (1998) was evaluated, the time association was correct 96% of the time. That these percent correct numbers overestimate the skill can be easily seen by considering the single storm cell track in Figure 1a. If the algorithm misses the association between the 5th and 6th time instance, then the end result is that there are two tracks instead of one leading to incomplete trend information for the second half of the sequence. Yet, this grossly incorrect cell track gets a skill of 90% since 9 out of 10 of the time associations are correct. A truer estimate would be close to 50% because the trend information in the second half of the sequence would be compromised by the association error.
3. The skill measure is *non-specific*. Consider the time associations shown in Figures 1b,c,d. In all three cases, the percent correct is 67% since 2 out of 3 associations are correct but the problem that causes the incorrect association is different: Figure 1b shows a dropped association, whereas Figure 1c shows a mismatch and Figure 1d shows an incorrect "jump" from a decayed cell to a new storm cell. Yet, in all three cases, the percent correct score is 67%.

These flaws have direct impacts on the design of a good tracking algorithm. Because human association is time consuming, the tracking algorithm would have to be developed on a very small dataset. This affects how robust the tracking code is because real world data is much more diverse than the training dataset. Secondly, because the skill when expressed in percent correct is an overestimate, it is difficult to demonstrate true improvements. There is just 4% of possible improvement between a 96% score that is not very good and the

theoretical maximum of 100%. Thirdly, the non-specificity makes it difficult to understand how to improve the tracking algorithm. If the percent correct score is 67%, should the improvement be in the form of increasing the search radius (to limit dropouts), reducing the search radius (to reduce the number of mismatches) or incorporating checks on changes in storm attributes (to reduce the number of jumps)?

b. Storm Tracking Algorithms

The basic unit of a storm tracking algorithm is the method by which storms identified in one time frame are associated with the already labeled storms in the previous time frame – a storm that is associated with a storm in the previous time frame inherits its label (usually termed its cell ID) and its time history. A "track" consists of the locations of a storm from the time it was first assigned a cell ID to the last time at which that ID was observed.

Many heuristics have been proposed to associate storms identified at the current time frame, t_n , with storms identified at the previous time frame t_{n-1} :

1. *PRJ* (Johnson et al. 1998): Cell centroid locations at t_{n-1} are projected (PRJ³) to where they would be at t_n based on the position of the cell centroid at times $t_{n-k} | k > 1$. Then, each cell at t_n is assigned to the closest unassigned centroid within a certain search radius. If no centroid is close by, then the cell is given a new ID.

³This mnemonic was not used by Johnson et al. (1998). They referred to their entire algorithm as SCIT, an acronym for Storm Cell Identification and Tracking. Because we wish to emphasize that the comparisons in this paper are carried out using a common identification algorithm (not the one in SCIT), and changing only the association algorithm, we assigned mnemonics that refer to just the association algorithm used in the various studies.

2. *CST* (Dixon and Wiener 1993): A global cost (CST) function, formulated as the sum of the Euclidean distance between matched centroids and a distance metric based on some property that should be relatively consistent, is minimized. Dixon and Wiener (1993) employ the volume of the cells as this consistent property; in this paper, we'll use the area of the cells since our comparison of tracking algorithms will be on two-dimensional images.
3. *AGE* (Lakshmanan et al. 2009): All projected cells within a size-based radius (given by $\sqrt{A/\pi}$ where A is the area of the storm) are considered "tied" in terms of position error, and such ties are resolved in favor of the longer-lived storm, i.e. based on age.
4. *OV* (Morel et al. 1997): A storm at t_n gets the ID of the cell at t_{n-1} with which it has maximum overlap (OV) and whose ID has already not been assigned. Cells are considered in order of size, with the largest cells assigned first.
5. *OC* (Han et al. 2009): This is a combination of the *OV* and *CST* methods carried out in sequence. Cells at t_n that have 50% or greater overlap with cells from t_{n-1} are first matched. Unmatched cells are then associated using a global cost function or assigned a new ID.

We will employ the objective evaluation of storm tracking introduced in this paper to compare these heuristics on different cases and use that analysis to devise a hybrid association technique that builds on the strengths of each of the component techniques.

2. Evaluation Method

Looking at Figure 1a again, the dropped association results in two tracks with duration half of that what an ideal algorithm would have produced, thus yielding a skill (based on just duration) of 50%. This skill score is intuitively more pleasing than the percent correct skill score of 90% since the trend information for the second half of the sequence is wrong. Similarly, in Figure 1b, the dropped association results in a lower duration. If one were to track the variability of some property of the cell such as the Vertically Integrated Liquid (VIL; Greene and Clark (1972)) across time, a mismatch such as in Figure 1c would result in the VIL being less consistent than that from an ideal algorithm that did not suffer from the mismatch problem. Finally, the jump in Figure 1d would result in the track not being as linear as it would have been if the tracking algorithm had been ideal.

Therefore, rather than evaluate a tracking method by counting the number of correct associations, it would be better to evaluate the tracks themselves in terms of three factors: the duration (length) of the track, the linearity of the track and the preservation of a storm attribute. In general, longer, more linear tracks where the storm attribute is relatively constant between frames are better. These criteria need to be balanced because a "jump" will lead to a longer track, but the track will be less linear than if the association had led to two, less long-lived tracks (Roberts et al. 2009).

It should be noted that we do not postulate that every track should be linear or that every track should be long-lived or that all cells should have constant VIL. Instead, we postulate that if a large enough data set is considered, a better association algorithm will produce longer tracks than a technique that frequently drops associations. On the other

hand, jumps will cause the tracks to be less linear than they would be had the jumps not occurred. In other words, we claim that correct associations will (in general) result in more linear tracks, not that all tracks are linear. Similarly, mismatches will lead to less consistent VIL than correct associations. A skill score balanced between these three factors is useful when considered in bulk (a large enough dataset of tracks) and when used to compare two methods. The absolute numbers will vary from dataset to dataset, but the difference in skill between two techniques on the same dataset can be used to evaluate the techniques relative to each other.

We evaluate an algorithm by computing the following statistics on each track produced by that algorithm:

1. dur is the duration of the track. The duration is longer if there are fewer dropped associations.
2. σ_V is the standard deviation of the VIL of the cell in time (i.e. along a track). The σ_V is lower if there are fewer mismatches.
3. e_{xy} is the Root Mean Square Error (RMSE) of centroid positions from their optimal line fit. The e_{xy} is lower for more linear tracks.

Central tendencies of the above statistics are computed on a large dataset of tracks:

1. \widetilde{dur} is the median duration of tracks in the dataset. The better the association technique, the fewer the number of short-lived tracks that result from the technique and the greater \widetilde{dur} is since the distribution of track lengths will be skewed towards longer-lived tracks. It is better to use the median rather than the mean so that it is not as effected

by the presence of outliers (tracks of duration less than 2 frames or a few extremely long-lived tracks).

2. The mismatch error ($\overline{\sigma_V}$) is the mean σ_V on tracks with duration greater than \widetilde{dur} . Fewer mismatches are indicated by more consistent VIL values and, thus, by a lower $\overline{\sigma_V}$. Because the standard deviation is highly sensitive when computed on small sample sizes, this statistic is computed only on tracks with duration greater than the median duration.
3. The linearity error ($\overline{e_{xy}}$) is the mean e_{xy} on all tracks with duration greater than \widetilde{dur} . Such tracks should have enough points to meaningfully compute the error of a line fit if the association technique is reasonably good. A technique where the median track has less than three centroids is not worth evaluating!

While it is possible to create a composite skill score as a weighted sum of all the above parameters, determining the appropriate weights is subjective. Instead, the parameters can be used to compare different techniques or to tune algorithm parameters. For example, one way to tune the search radius would be to increase it as long as \widetilde{dur} keeps increasing and $\overline{\sigma_V}$ and $\overline{e_{xy}}$ remain below some threshold determined by the performance of a simple method such as PRJ.

In order to perform a fair comparison of different storm tracking techniques, they were run against cells identified using the same storm identification technique with the same parameters. Storms were identified using the extended watershed approach of Lakshmanan et al. (2009) on median-filtered (in a 9x9-pixel neighborhood) reflectivity composite images and searching for cells above 30 *dBZ* with a minimum size of 20 *km*². Motion was estimated

over the entire field by cross-correlation of storm cells at t_n against the image at t_{n-1} (i.e. the correlation is between cells in the current frame against pixel values in the previous frame), interpolating between the cells and smoothing over time using a Kalman filter (See Lakshmanan and Smith (2008) for details). The identified cells were then associated using each of the techniques listed in Section 1b. The various techniques were implemented by us from their description in the literature, i.e. we did not use the operational SCIT⁴ or TITAN (since they are tied to their own storm identification techniques). Instead, we implemented centroid projection and a global cost reduction approach based on the descriptions in Johnson et al. (1998) and Dixon and Wiener (1993) respectively.

The techniques were evaluated on a common dataset consisting of the following WSR-88D radar data (from 18:00 UTC to 23:59 UTC on each of the days): KBIS, Bismark, ND on May 21, 1995; KCBX, Boise, ID on May 1, 1995; KIWA, Phoenix, AZ on Aug. 6, 1993, Aug. 20, 1993 and Aug. 6, 2003; KLSX, St. Louis, MO on June 8, 1993 and July 2, 1993; KLWX, Sterling, VA on Apr. 14, 1993, May 1, 1994, Oct. 6, 1995 and Oct. 6, 2005; KMLB, Melbourne, FL on Mar. 25, 1992, June 9, 1992 and June 12, 1992; and KTLX, Oklahoma City, OK on June 18, 1992 and Feb. 21, 1994. These cases are diverse geographically and in terms of the storm types. For example, they include a mesoscale convective system (KMLB, Melbourne, FL on Mar 25, 1992), a convective line (KLSX, St. Louis, MO on June 8, 1993), a stratiform event (KTLX, Oklahoma City, OK on Feb 21, 1994), isolated storms (KIWA, Phoenix, AZ on Aug 6, 1993) and a minisupercell (KLWX, Sterling, VA on Oct 6, 1995).

The above cases were chosen because they were verified by hand in Johnson et al. (1998).⁵

⁴Storm Cell Identification and Tracking, the algorithm used within the National Weather Service for radar-based storm tracking

⁵The KIWA 2003 and KLWX 2005 cases were used in lieu of the cases considered by Johnson et al. (1998)

Therefore, we know approximately how many cells we should expect to find in each of these cases and we could ensure that the storm identification algorithm was finding a similar number of cells. This is important because tracking algorithms can not be compared fairly if the identification algorithm is so lax as to identify too many "cells" or is so strict that it does not detect many true cells.

Gauging whether the storm cells identified by the algorithm were the same as that which a human would call a storm is approximate because the list of actual human verified cells is no longer available (Pam Heinselman, personal communication). For the isolated storms event observed by the WSR-88D at Phoenix, AZ on Aug 6, 1993, it was reported that 867 cells were identified by hand. Unfortunately, the time period considered on that day is not reported, so this number is not very useful. Therefore, we computed an order-of-magnitude estimate from the statistics reported in the paper as follows. Using the isolated storms event observed by the WSR-88D at Phoenix, AZ on Aug 6, 1993 as an example:

1. For isolated storms, the reported probability of detection was 27% for cells between 30-39 dBZ, 70% for cells between 40-49 dBZ and 96% for cells above 50 dBZ.
2. The algorithm of Johnson et al. (1998) when run on the Phoenix AZ data from 18:00 UTC and 23:59 UTC identified 2377 cells of which 227 cells were in the 30-39 dBZ range, 1174 cells were in the 40-49 dBZ range and 976 cells had a peak reflectivity above 50 dBZ.

3. Based on the above data, and assuming that the probability of detection (POD) is

– KFDR, Fredrick, OK on Apr. 20, 1992 and KOUN, Norman, OK on Sep. 2, 1992 – that are not available in the National Climatic Data Center archives. The 2003 and 2005 cases were not verified by hand.

similar to that for the time period that was actually human verified, then there should have been a total of $227/0.27 + 1174/0.7 + 976/0.96$ cells i.e. 3535 cells.

4. The storm identification technique used in this paper to evaluate the different tracking techniques identified 2309 cells, which is of the same order of magnitude as 3535 (the number of cells one expects to see in the dataset based on the human verification).

It should be emphasized that this exercise of estimating the number of human-truthed cells in the dataset was carried out only to ensure that the tracking algorithms are not presented with an unrealistic number of cells. The number of "true" cells was estimated rather than laboriously counted by hand because the actual number (3535 in the example) is not important, but only its order of magnitude. We would have had a problem if a human thinks that there are really 3000 cells in the dataset, but our storm identification algorithm only detected 300.

a. Analysis

The evaluation of each of the techniques is shown in Figure 2. Each row of graphs consists of the evaluation of a case using the three criteria described in Section 2. Each column of graphs corresponds to one of the metrics. The techniques being evaluated and their mnemonics are described in Section 1b. In each graph, the best two methods are shown in black. If several techniques tied for second place (within the bounds of statistical significance shown by the error bars) as in the case of mismatches for the first case, there may be more than two black bars in a graph. Similarly, the worst two methods (with a rank of 5 or 6) are shown with white bars. Gray bars indicate middling (rank of 3 or 4)

performance.

In the case of the mismatch error ($\overline{\sigma_V}$) and linearity error ($\overline{e_{xy}}$), the confidence intervals are computed from the standard deviation σ of these errors on the N tracks for which these errors are computed (recall that these statistics are computed only on the longest 50% of tracks in the dataset) as $\mu + / - \alpha\sigma/\sqrt{N}$ where μ is the mean error and α is obtained from statistical tables of a two-tailed Student's T distribution. As \widetilde{dur} is a median, its confidence interval is given by the durations of the $(N/2 + / - \alpha\sqrt{N/4})$ th longest tracks: the number of observations less than the q^{th} quantile is an observation from a Binomial distribution with parameters N and q and therefore has mean Nq and standard deviation $\sqrt{Nq(1-q)}$ (Conover 1980). A 50% confidence interval ($\alpha \approx 0.67$ for $N > 30$) was employed because the purpose of the error bars is to gauge visually whether one technique is *likely* to be better than another for the purposes of creating a reasonable ranking of the techniques. A 95% confidence interval would have been used if our purpose had been to *prove* that one technique was better than another. For example, consider the graph in Figure 2 for the mismatch error on the KLSX line case (second row, first column). The CST method is likely to have more mismatches than the OC method on the KLSX case because the 50% confidence intervals do not overlap. Thus, CST is the worst performing method while the AGE, OV and NEW (whose confidence bars do overlap) are all ranked first. The error bars corresponding to the PRJ and NEW methods do not overlap. Therefore, the PRJ method is ranked 4.

It can be noted from the first column of graphs in Figure 2 that the mismatch error ($\overline{\sigma_V}$) is lowest when using the overlap (OV) method. No other method has black bars (good performance) in all five of the cases considered. This is not surprising because the overlap

function is the most conservative form of storm association. The drawback of using such a conservative approach to associating cells is that the median duration of tracks is bad (white bars) in four of the five cases – only for isolated cells does the OV method have good performance on all three measures.

Similarly, it can be noted from the second column of graphs that the linearity error ($\overline{e_{xy}}$) is lowest when using the projected centroid (PRJ) method of Johnson et al. (1998). Again, this is not surprising because the centroid projection method explicitly minimizes position error after accounting for storm movement, thus emphasizing linearity at the cost of duration. Indeed, the PRJ method has bad performance in two of the five cases on the length metric.

It can be noted that even techniques that do not try to minimize mismatch error (such as PRJ, AGE or OV) attain quite good scores on that measure. The cost function in CST and OC explicitly include size preservation in an attempt to reduce mismatches. However, size preservation does not seem to induce a corresponding reduction in the variability of VIL. The CST and OC techniques have the worst performance in terms of mismatches and jumps, but also (as a tradeoff) the best performance in terms of duration. The centroid location-based methods (PRJ and AGE) do not consider preservation of the value of any attribute. Yet, they consistently do well as far as mismatch error is concerned. This indicates that $\overline{\sigma_V}$, for the most part, is maintained along a track even if the tracking algorithm only minimizes location error. This also indicates that the CST and OC methods may be overemphasizing size preservation and, therefore, producing more non-linear tracks.

It is also apparent that, in every situation, the enhancements proposed by Han et al. (2009) do improve the tracking. The OC method has longer tracks than the CST method

for every case. It achieves this longer tracking with fewer jumps and a similar mismatch error. It should be noted that Han et al. (2009) were unable to demonstrate this improvement because they tried to compare the two techniques using forecast error. Our use of appropriate metrics has enabled us to verify that the enhanced TITAN tracking method of Han et al. (2009) is indeed better than the original TITAN method of Dixon and Wiener (1993).

The AGE method that was introduced "for simplicity" in Lakshmanan et al. (2009) performs surprisingly well for all cases. That method finds reasonable candidates in terms of location error and then chooses among these candidates first in terms of longevity and then (if there is a tie in terms of age) on size and finally in terms of intensity. Later experimentation determined that longevity alone was enough and it is that even simpler version that was used in this paper. The good performance of AGE indicates that the key parameters for a tracking algorithm are location error and longevity.

3. Devising a New Tracking Algorithm

Our aim was to devise a technique whose performance is consistently good i.e. on all cases on all metrics, the technique should be among the best performers. The fact that Han et al. (2009) were able to combine two poorly performing methods (OV and CST) in sequence and create a pretty good technique (OC) gave us an idea of how to proceed in devising a good, consistent tracking algorithm by combining the best aspects of all the previously considered techniques.

The tracking technique marked as "NEW" in Figure 2 was carried out as follows. At each step, only those storms that have not yet been associated are considered.

1. Project storm cells identified at t_{n-1} to their expected location at t_n .
2. Sort the storm cells at t_{n-1} by track length, so that longer-lived tracks are considered first in Step 3.
3. For each (unassociated) projected centroid, identify all centroids at t_n that are within d_{n-1} kms of the projected centroid. d_{n-1} is given by $\sqrt{A/\pi}$ where A is the area of the projected storm cell at t_{n-1} .
4. If there is only one centroid within the search radius in Step 3, and if the distance between it and the projected centroid is within 5 km, then associate the two storms.
5. Repeat steps 3 and 4 until no changes happen. At this point, all unique centroid matches have been performed.
6. Define a cost function c_{ij} for the association of candidate cell i at t_n and cell j projected forward from t_{n-1} as:

$$c_{ij} = (x_i - x_j)^2 + (y_i - y_j)^2 + \frac{A_j}{\pi} \left(\frac{|A_i - A_j|}{A_i \wedge A_j} + \frac{|d_i - d_j|}{d_i \wedge d_j} \right) \quad (1)$$

where x_i, y_i is the location, A_i the area and d_i the peak pixel value of cell i (in the spatial field in which cells are being detected). $|a|$ refers to the magnitude of a and $a \wedge b$ refers to the maximum of a and b .

7. For each unassociated centroid at t_n , identify all projected centroids within d_n kms where d_n is expressed in terms of the area of the cell at t_n as $\sqrt{A/\pi}$.
8. Associate each unassociated centroid at t_n with the unassociated, projected centroid within d_n for which the cost function c is minimum. If there are no centroids within

the search radius, mark it as a new cell.

As can be seen, the tracking technique is a judicious combination of PRJ (Step 1), AGE (Step 2), OC (size-based search radius in Step 3) and CST (Step 6). The uniqueness check of Step 4 is novel, and allows the algorithm to distinguish between line storms (where there will be multiple storms within the search radius, leading to the use of the cost function) and isolated cells (where centroid matching performs well). The cost function here incorporates both size and peak intensity.

a. Results

The new tracking technique introduced here exhibits consistently good performance as evidenced by the black and gray bars in Figure 2 on all cases and metrics. The NEW technique has slightly better performance than AGE in terms of mismatch error on the minisupercell case but this difference is not statistically significant. All the other techniques (PRJ, CST, OV and OC) have poor performance (white bars) in at least one case or measure. The performance of the algorithm on those cases where it has only middling (gray bars in Figure 2) performance has to be examined in depth to further improve the algorithm.

b. Directions for further exploration

Because storm identification methods are based on thresholding input fields, whether with a single threshold or by multiple thresholds, storm identification can be inconsistent from frame-to-frame due to the natural evolution of storms. Consequently, several preprocessing

operations have been proposed to improve the trackability of identified storms (Stumpf et al. 2005; Lakshmanan et al. 2007; Han et al. 2009). The performance of the preprocessing filter and association technique may depend heavily on an appropriate choice of parameters. The search radius within which to conduct a search for best association (and beyond which to call it a new cell) can be chosen based on size (as done by Lakshmanan and Smith (2008) and Han et al. (2009)) or based on a directional constraint as done by Johnson et al. (1998). Similarly, the number of frames to "coast" an unmatched cell before it is finally dropped can affect the longevity of the tracks. Johnson et al. (1998) did not coast at all, while Lakshmanan and Smith (2008) coast for three frames. It is possible to use the objective evaluation of storm tracking introduced in this paper to make an appropriate choice of such constraints also.

It may be tempting to boil down the three measures – property consistency, linearity and duration – into a single measure of skill. However, the temptation ought to be resisted. The advantage of using all three specific characteristics is that they provide insight into the tradeoffs that can be made in changing a tracking algorithm. For example, if it is felt that the duration of tracks produced by the NEW algorithm is not sufficient, then the remedy would be to increase the search radius. Also, the values of these measures will vary from case to case. The duration of tracks that can be expected when tracking continental-scale frontal systems will be much longer than when tracking pulse storms in desert regions. Reducing a measure like duration to a number in the range $[0,1]$ might be difficult.

Similarly, it is tempting to compute these metrics on all the tracks in all the cases considered (see Figure 3). However, that graph should be read with caution because these cases are not climatologically representative. For example, isolated storms may be several times more likely than line storms, and so performance on the isolated storms case should

be weighted more heavily. Therefore, if the interest is in gauging the likely performance of a technique in operations over a wide variety of cases, these measures should be computed on several years' worth of data, not averaged over a few, possibly unrepresentative, cases.

In this paper, we defined the mismatch error based on the consistency of VIL along a track. If the tracking algorithms are being carried out on other fields such as satellite infrared imagery or total lightning, the consistency criterion should be chosen appropriately. Choosing a criterion such as size could be a general-purpose choice in that it would work on all fields. However, it would greatly bias the mismatch error towards the CST and OC techniques since those techniques use size in their cost functions.

An important aspect of storm tracking algorithms is how they handle splitting or merging of storms. The statistics introduced in this paper reward good handling of splits and merges by the algorithm. For example, the way to handle a split might be to either (a) choose one of the storm cells after the split to carry on the old ID and assign a new ID to the other cell or (b) assign the history of positions to both the cells after the split. The second method will result in tracks with longer durations but possibly higher mismatch and linearity errors especially if one of the cells executes a turn or is slower moving than the combined entity was. Thus, an algorithm that considers the trajectory and morphology of storms to choose between the two options will exhibit longer duration and lower mismatch and linearity errors. Similarly, when two cells merge, there is a choice of whether to propagate one of the cell IDs to the combined entity. Propagating the longer-lived cell will always increase the duration of the track, but at the potential cost of mismatch and linearity errors. None of the storm cell tracking algorithms in the literature perform this sort of sophisticated analysis to handle splits and merges, mainly because there was no way to evaluate the efficacy of such analysis.

We hope that the introduction of these objective criteria for evaluating tracks will prompt new research into this topic.

This paper presented a framework that allows for the comparison of tracking algorithms and the design of a composite tracking algorithm. The actual comparisons will turn out differently if carried out on storms identified at different scales. For example, the overlap-based methods may perform better if the storms had been identified at a 200 km^2 scale rather than a 20 km^2 scale while the centroid-based methods may have performed more poorly if the sizes of the storms were more variable. Therefore, the relative performance of the techniques indicated in Figure 3 should not be extrapolated to other types of imagery or other scales of storms. Instead, the criteria introduced in this paper ought to be employed when choosing the tracking algorithm that will perform best on the imagery and storm scale of interest.

4. Summary

Although storm tracking algorithms are a key ingredient of nowcasting systems, evaluation of storm tracking algorithms has been indirect, labor intensive or non-specific. In this paper, we introduced a set of easily computable bulk statistics that can be used to directly evaluate the performance of tracking algorithms on specific characteristics. We applied the evaluation method to a diverse set of radar reflectivity data cases and noted the characteristic behavior of five different storm tracking algorithms proposed in the literature and now employed in widely used nowcasting systems. Based on this objective evaluation, we devised a storm tracking algorithm that performs consistently and better than any of the previously

suggested techniques.

Acknowledgements

Funding for the authors was provided under NOAA-OU Cooperative Agreement NA17RJ1227.

We gratefully acknowledge useful discussions with Benjamin Root and Madison Burnett.

Brett Roberts carried out some of the initial studies on which this paper is based.

REFERENCES

- Conover, W., (Ed.) , 1980: *Practical Nonparametric Statistics*. John Wiley and Sons, New York, 493 pp.
- Dixon, M. and G. Wiener, 1993: TITAN: Thunderstorm identification, tracking, analysis and nowcasting – a radar-based methodology. *J. Atmos. Ocean. Tech.*, **10**, 785–797.
- Greene, D. R. and R. A. Clark, 1972: Vertically integrated liquid water – A new analysis tool. *Mon. Wea. Rev.*, **100**, 548–552.
- 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. Ocean. Tech.*, **26**, 719–732.

- Johnson, J., P. MacKeen, A. Witt, E. Mitchell, G. Stumpf, M. Eilts, and K. Thomas, 1998: The storm cell identification and tracking algorithm: An enhanced WSR-88D algorithm. *Weather and Forecasting*, **13** (6), 263–276.
- Lakshmanan, V., A. Fritz, T. Smith, K. Hondl, and G. J. Stumpf, 2007: An automated technique to quality control radar reflectivity data. *J. Applied Meteorology*, **46** (3), 288–305.
- Lakshmanan, V., K. Hondl, and R. Rabin, 2009: An efficient, general-purpose technique for identifying storm cells in geospatial images. *J. Ocean. Atmos. Tech.*, **26** (3), 523–37.
- Lakshmanan, V., R. Rabin, and V. DeBrunner, 2003: Multiscale storm identification and forecast. *J. Atm. Res.*, **67**, 367–380.
- Lakshmanan, V. and T. Smith, 2008: Data mining storm attributes from spatial grids. *J. Ocea. and Atmos. Tech.*, In Press.
- Morel, C., F. Orain, and S. Senesi, 1997: Automated detection and characterization of MCS using the meteosat infrared channel. *Proc. Meteor. Satellite Data Users Conf.*, Eumetsat, Brussels, 213–220.
- Roberts, B., V. Lakshmanan, and T. Smith, 2009: Evaluation of a multi-scale storm-tracking technique. *25th Int’l Conf. on Inter. Inf. Proc. Sys. (IIPS) for Meteor., Ocean., and Hydr.*, Phoenix, AZ, Amer. Meteor. Soc., P2.10.
- Stumpf, G., S. Smith, and K. Kelleher, 2005: Collaborative activities of the NWS MDL and NSSL to improve and develop new multiple-sensor severe weather warning guidance ap-

plications. *Preprints, 21st Int'l Conf. on Inter. Inf. Proc. Sys. (IIPS) for Meteor., Ocean., and Hydr.*, Amer. Meteor. Soc., San Diego, CA, P2.13.

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.

List of Figures

1	Using the "percent correct" of time associations is flawed as a way of evaluating the performance of tracking algorithms because it is an overestimate as discussed in the text and shown in (a) and is non-specific as discussed in the text and shown in (b,c,d). Dashed lines as in (a,b) indicate a "dropped" association while arrows indicate a wrong association which could be due to a mismatch as in (c) or due to a "jump" as in (d). Solid lines indicate a correct time association.	25
2	Evaluation of different tracking techniques. Black bars denote good performance while white bars indicate poor performance.	26
3	Evaluation of tracking techniques on all 16 cases. Black bars denote good performance while white bars indicate poor performance.	27

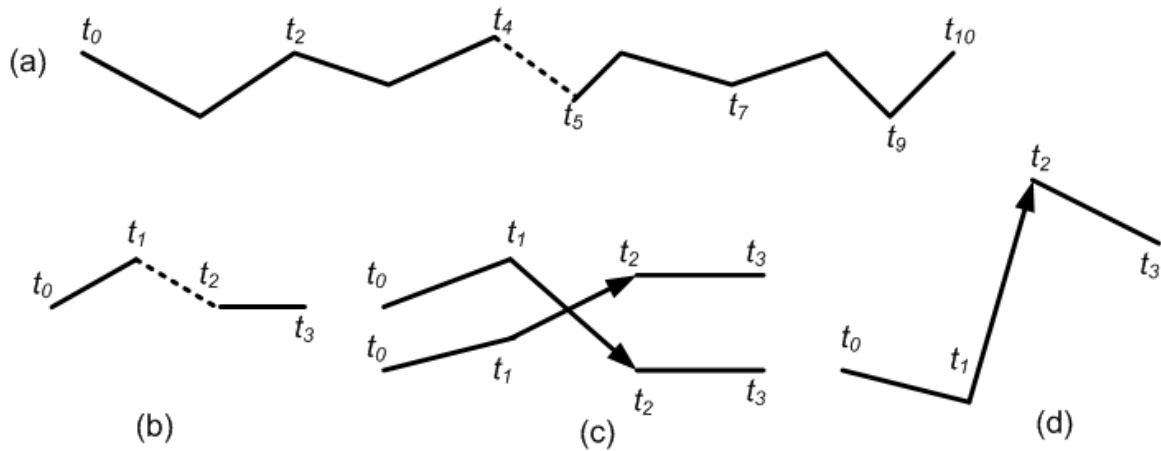


FIG. 1. Using the "percent correct" of time associations is flawed as a way of evaluating the performance of tracking algorithms because it is an overestimate as discussed in the text and shown in (a) and is non-specific as discussed in the text and shown in (b,c,d). Dashed lines as in (a,b) indicate a "dropped" association while arrows indicate a wrong association which could be due to a mismatch as in (c) or due to a "jump" as in (d). Solid lines indicate a correct time association.

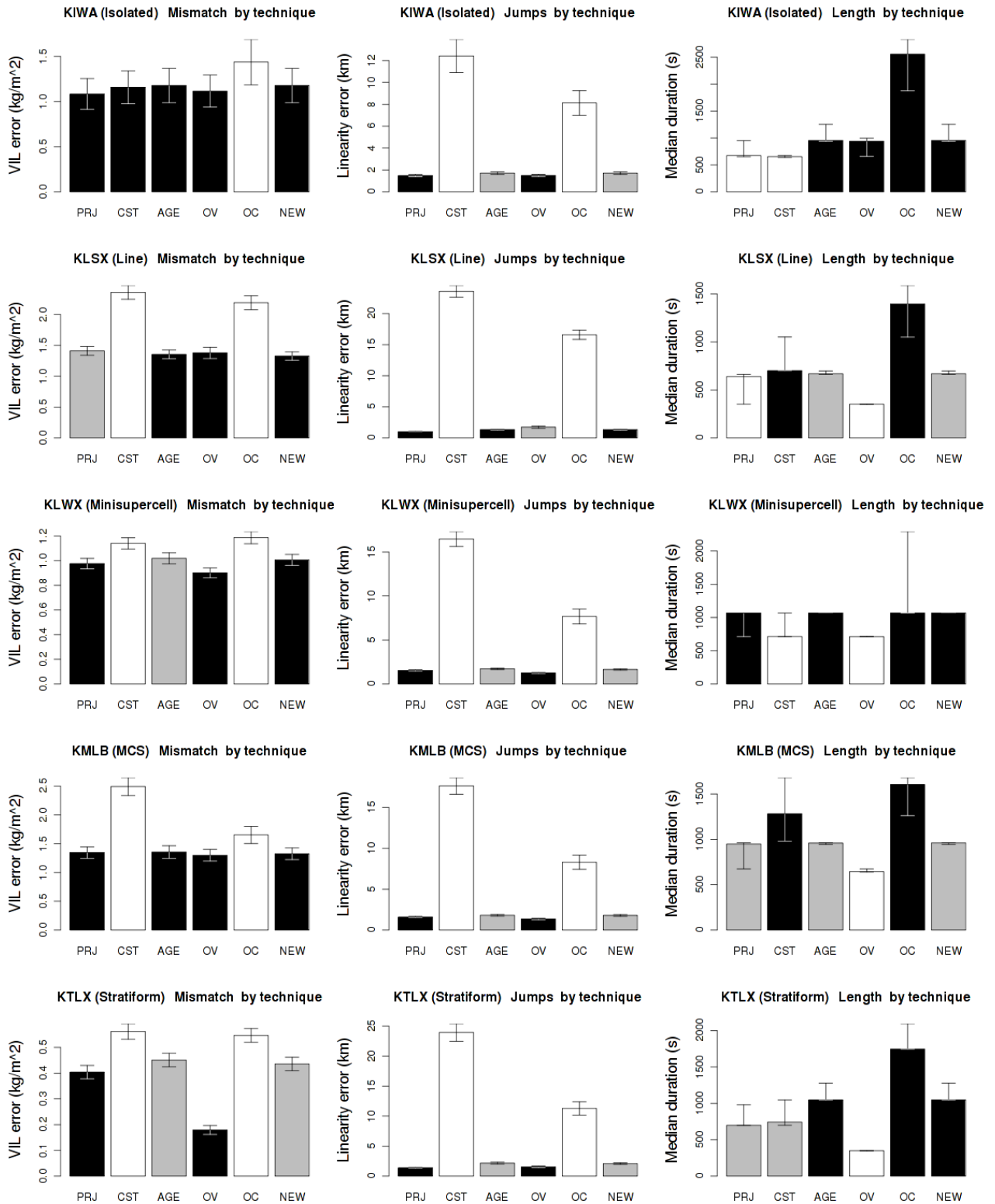


FIG. 2. Evaluation of different tracking techniques. Black bars denote good performance while white bars indicate poor performance. 26

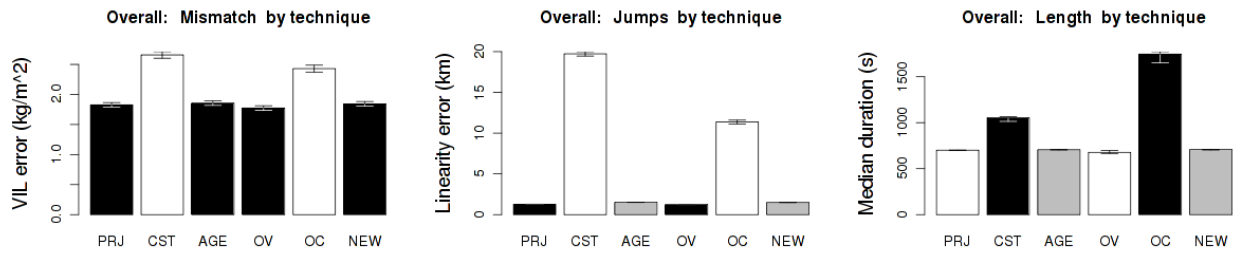


FIG. 3. Evaluation of tracking techniques on all 16 cases. Black bars denote good performance while white bars indicate poor performance.