ADVANCES IN ATMOSPHERIC SCIENCES, VOL. 32, NOVEMBER 2015, 1449–1459

Percentile-based Neighborhood Precipitation Verification and Its Application to a Landfalling Tropical Storm Case with Radar Data Assimilation

ZHU Kefeng[1,2], YANG Yi[3], and Ming XUE*[1,2]

[1]*Key Laboratory of Mesoscale Severe Weather/Ministry of Education and School of Atmospheric Sciences, Nanjing University, Nanjing* 210093

[2]*Center for Analysis and Prediction of Storms, University of Oklahoma, Norman Oklahoma* 73072*, USA*

[3]*College of Atmospheric Sciences, Lanzhou University, Lanzhou* 730000

(Received 1 February 2015; revised 14 April 2015; accepted 15 May 2015)

ABSTRACT

The traditional threat score based on fixed thresholds for precipitation verification is sensitive to intensity forecast bias. In this study, the neighborhood precipitation threat score is modified by defining the thresholds in terms of the percentiles of overall precipitation instead of fixed threshold values. The impact of intensity forecast bias on the calculated threat score is reduced. The method is tested with the forecasts of a tropical storm that re-intensified after making landfall and caused heavy flooding. The forecasts are produced with and without radar data assimilation. The forecast with assimilation of both radial velocity and reflectivity produce precipitation patterns that better match observations but have large positive intensity bias. When using fixed thresholds, the neighborhood threat scores fail to yield high scores for forecasts that have good pattern match with observations, due to large intensity bias. In contrast, the percentile-based neighborhood method yields the highest score for the forecast with the best pattern match and the smallest position error. The percentile-based method also yields scores that are more consistent with object-based verifications, which are less sensitive to intensity bias, demonstrating the potential value of percentile-based verification.

Key words: neighborhood precipitation threat score, percentile-based verification, radar data assimilation

Citation: Zhu, K. F., Y. Yang, and M. Xue, 2015: Percentile-based neighborhood precipitation verification and its application to a landfalling tropical storm case with radar data assimilation. *Adv. Atmos. Sci.*, **32**(11), 1449–1459, doi: 10.1007/s00376-015-5023-9.

1. Introduction

Traditional point-to-point verification scores such as the Critical Success Index (CSI), also known as the "threat" score, are often used for precipitation verification. These scores generally use a 2×2 contingency table to determine "yes" and "no" points (Wilks, 1995). Verification at high resolution, when the predicted rain storm deviates from the observations, can result in a "double penalty" in observed-but-not-forecasted and forecasted-but-not-observed cases (Ebert, 2008). Neighborhood-based methods act as if the forecast precipitation amounts on the model grid are randomly distributed in the vicinity of the correct position (Rezacova et al., 2007). Instead of a point-to-point verification of the forecasts against observations, the verification is performed with a dependence on the surrounding grid boxes (Ebert, 2008, 2009). Such methods can reduce the impact of displacement error on the calculated verification score. It has been demonstrated to be more meaningful than the traditional point-to-point methods, and it can also be used to diagnose the forecast skill as a function of spatial scale (Clark et al., 2010).

However, the displacement error is merely one form of forecast error. Errors in the intensity, size and shape of precipitation are also very common in numerical model predictions. In practice, the size and shape of the precipitation regions, or the precipitation patterns, are often more important to the forecast end users. Verification methods such as the object-based method turn the forecasts and observations into a cluster of rainfall pairs (Davis et al., 2006a, b; Brown et al., 2007). The geometric features of the object pairs, such as area, angle of axis, curvature, and centroid, are used to describe the similarity between forecasts and observations. Such a method is much closer to a subjective verification, where the precipitation pattern carries more weight, and is helpful in identifying the sources of forecast error. The problem is, as will be discussed later, the object pairs for one experiment may differ from those of another. Therefore, a fair comparison between experiments with large forecast differences is difficult because the matched pairs can differ signifi-

* Corresponding author: Ming XUE
Email: mxue@ou.edu

cantly among the experiments.

In both neighborhood and objected-based methods mentioned above, a common question is: to what extent is the forecast bias tolerable? In neighborhood verification, the forecasts require approximate agreement with observations (Ebert, 2008). A forecast with a small displacement error is still considered a "good" forecast. In object-based verification, a small intensity bias is acceptable as long as the geometric features of selected pairs are similar.

The percentile-based neighborhood method attempts to reduce the impact of intensity error as well as displacement error on the calculated verification score. In both neighborhood and object-based methods, the intensity threshold is important in determining the initial boundary of the verification pairs. However, that threshold is a fixed value. The problem is that in real forecast systems, the forecast intensity, especially for the intensity of the heavy rain area, is most likely uncertain. It can be affected by factors such as model resolution, model physics, and initial conditions. When the same fixed threshold is used across forecasts of different intensity biases, the final objective verification results may be inconsistent with the subjective assessment.

The concept of a percentile-based threshold is not entirely new. Johannes et al. (2008) used a percentile-based threshold in a traditional point-to-point verification method. In Roberts and Lean (2008), the authors used a percentile-based threshold within their neighborhood verification. Because they were comparing forecasts with different model resolutions, the use of a percentile-based threshold served to remove the impact of bias in the rainfall amounts, as the focus was placed on spatial accuracy. In their paper, newly proposed continuous statistical verification scores, such as the fraction skill score, were examined using the percentile-based threshold. Here, we apply a percentile-based threshold to the most commonly used category verification score, the CSI, and borrow the idea of the "raw threshold" from the object-based method. The latter can potentially reduce the initial size error. Details are presented in the following section.

The rest of the paper is organized as follows. In section 2, the basic verification metrics of the traditional neighborhood method and the object-based method are briefly introduced, together with our percentile-based neighborhood verification method. In section 3, the forecasts for a re-intensifying landfalling tropical storm are used as an example to examine the ability of the three verification methods in distinguishing forecasts with large intensity, size and structural differences in precipitation. These forecasts differ in whether radar data are assimilated and how they are assimilated. Finally, a summary and conclusions are given in section 4.

2. Verification methods

2.1. *Object-based verification methods*

Object-based verification methods evaluate forecasts by identifying objects in the forecast fields and comparing them with those identified in the observation fields. Their intention is to provide an evaluation of the forecasts that is more consistent with subjective assessment. They measure forecast errors in terms of the objects' properties, such as intensity, location, size and geometric differences of the objects. In this manner, the objects are no longer treated as "points". Instead, the method converts the forecasts or observations into a cluster of objects or points. Here, we introduce one typical method, proposed by Davis et al. (2006a), that was implemented in the Model Evaluation Tools (METs) (Brown et al., 2009).

There are generally two steps to finding the objects within MET: convolution and thresholding. The raw data are first convoluted using a simple filter function. Then, the convoluted field is thresholded, and the boundaries of objects are detected. Once the objects are isolated, the points within them are restored to the original values. The various attributes of an isolated object, such as intensity, area, axis angle, and aspect ratio, are calculated, and differences between pairs of objects, such as the centroid difference, are calculated as well. An index named "total interest" is then calculated, in which the attributes are weighted and summarized. The definition of total interest $T(\alpha)$ is described as (DTC, 2009)

$$T(\boldsymbol{\alpha}) = \frac{\sum_i w_i C_i(\boldsymbol{\alpha}) I_i(\alpha_i)}{\sum_i w_i C_i(\boldsymbol{\alpha})}, \tag{1}$$

where $\boldsymbol{\alpha}$ is the vector of various object attributes $(\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_n)$, C_i is the confidence map range from 0 to 1 and is a function of the entire attribute vector $(\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_n)$, $I_i(\alpha_i)$ is the interest map that depends on α_i only, and w_i is the weight assigned to each attribute. Finally, the isolated objects are merged (if they are in the same field) or matched (if they are in different fields) when they exceed a certain threshold.

2.2. *Neighborhood verification methods*

A forecast bias such as position error is a common problem, especially for high-resolution models. Ebert (2008) proposed a neighborhood method to reduce the impact of displacement error. Instead of treating the point as either "yes" or "no", it turns the "point value" to "probability rain" and calculates the probability in a square box around that point. The formula is

$$\langle P \rangle_{\rm s} = \frac{1}{M} \sum_i^M I_i, \quad I_i = \begin{cases} 1, & r_i \geqslant r_{\rm thresh} \\ 0, & r_i < r_{\rm thresh} \end{cases}. \tag{2}$$

Here, M is the total number of grid points surrounding the verification point, which is determined by the neighborhood width. I_i is a Heaviside function that depends on the grid-point rain intensity value r_i and the given threshold $r_{\rm thresh}$. After the probability $\langle P \rangle_{\rm s}$ is calculated, the $\langle I \rangle_{\rm s}$ of the point is determined by giving a coverage threshold $P_{\rm thresh}$:

$$\langle I \rangle_{\rm s} = \begin{cases} 1 \langle P \rangle_s \geqslant P_{\rm thresh} \\ 0 \langle P \rangle_s < P_{\rm thresh} \end{cases}. \tag{3}$$

Using $\langle I \rangle_{\rm s}$, the calculation of various forecast skill scores is the same as in the traditional method.

Compared to the traditional point-to-point method, the neighborhood method has two other key parameters: the neighborhood width and the coverage threshold. Here, the sensitivity of those two parameters to the verification scores is demonstrated for CSI. For simplicity, the forecast is assumed to have only displacement error and has a 30 grid-point offset from the observed feature (Fig. 1a). The use of the neighborhood width increases the cross-sectional area for the forecast and observation. The coverage threshold then determines the point's properties, including "hits", "false alarms", "misses" or "correct rejections". If the same neighborhood width is used, a lower coverage threshold usually results in more "yes" points [see Eq. (3), $\langle I \rangle_s$] for both the forecast and observation. The lower threshold increases the number of "hits" points, which results in a higher CSI score (see Fig. 1b). If the same coverage threshold is used, an increase in the neighborhood width initially raises the CSI score, and then decreases it (see Fig. 1b). This occurs because the higher the neighborhood width is, the lower $\langle P \rangle_s$ is. When $\langle P \rangle_s$ is smaller than the coverage threshold, the probability of hit ratios is decreased.

2.3. *Percentile-based neighborhood verification*

In traditional neighborhood verification, r_{thresh} is fixed; hence, variations in storm intensity are not considered. When a fixed threshold is used, it is common to observe the verification scores dropping rapidly as the storm weakens. This is especially true for a high threshold. Sometimes, low forecast skill is reported during the storm's initial and dissipating stages, causing the rate of intensification or weakening to be not quite right. As such, it is difficult to distinguish between intensity and shape or size errors in the forecast.

To minimize the impact of intensity error, we propose a flexible threshold that is based on the percentile. While the fixed-value threshold attaches more importance to the intensity, the percentile-based threshold gives more weight to the size. Figure 2 represents an example of an idealized forecast. Both the predicted size and location of the storm match those of the observation, except that the maximum rain intensity is underestimated by 50 mm (see the innermost contours in Fig. 2). Here, we assume that the contours of 100 mm, 150 mm and 250 mm correspond to the 50th, 75th and 90th percentiles for the observation, respectively. For the forecast, the first two percentile values are the same as the observations, but the last one is 200 mm. If the 250 mm threshold is used to calculate the equitable threat scores (ETS), the score is zero because none of the forecast reaches the observed intensity. On the contrary, if the 90th percentile of the percentile-based threshold is used, the thresholds for the observation and forecast are set to 250 and 200 mm, respectively, and then the ETS score is 1. The percentile-based threshold can reduce the impact of the intensity error if the predicted size is the same. The formula for the percentile-based threshold is given as

$$r_{\text{thresh}} = \text{percentile}(r > r_{\text{raw}}, n) \ , \qquad (4)$$

where n represents the "nth percentile" and r_{raw} is a raw threshold. This raw threshold is necessary because the precipitation area is usually small compared to the entire verification region, which often includes too many zero points. Once the threshold is computed, the rest of the procedures follow the neighborhood verification method, described above.

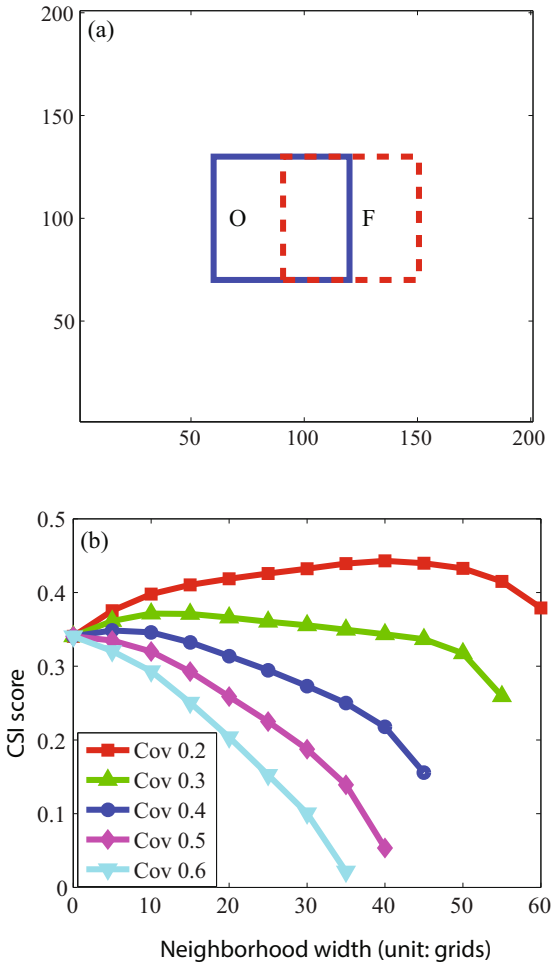

Fig. 1. An example of neighborhood CSI score as a function of neighborhood width and coverage threshold. The verification domain has 201×201 grid points. Both forecast and observed features cover 60×60 grid points but the forecast is shifted in position by 30 grid points to the east (a). The CSI scores for neighborhoods with varying widths and coverage thresholds (b). Note: "Cov" in the legend means "coverage threshold".

3. Verification results of a radar data assimilation case

In this section, a selected inland tropical storm, Erin, is used to examine the newly introduced percentile-based verification method. A subjective assessment is first made by side-to-side comparison. Next, the object-based method is used to further support the results of the subjective assessment. The traditional fixed neighborhood method and the percentile-based method are then presented and compared. The entire verification uses National Centers for Environmental Prediction (NCEP) stage IV precipitation data (Lin and Mitchell, 2005) as observations.

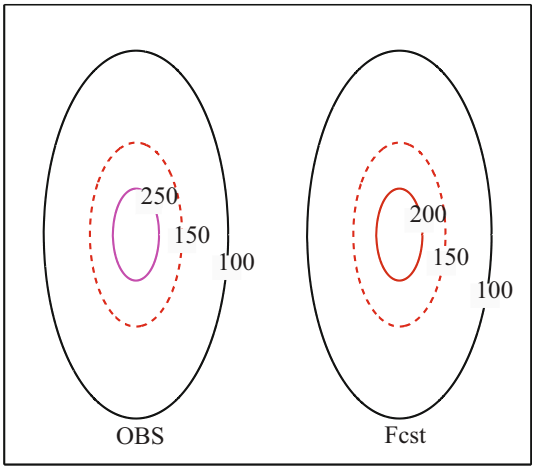

Fig. 2. The accumulated rainfall for an idealized forecast (right) as compared to observation (left). Here, we assume that the predicted rain has the same shape and location as the observation, except that it underestimates the intensity of heavy rain. The innermost observation contour is at 250 mm, while the corresponding forecast contour is at 200 mm.

3.1. *The inland tropical storm case and subjective assessment*

Erin began as Atlantic Tropical Depression Five (2007). Throughout its existence over open water, its sustained winds never exceeded 18 m s^{-1}, and its lowest reported central pressure was 1003 hPa. However, on the day immediately following Erin's landfall, it unexpectedly and dramatically re-intensified from 0000 UTC through 1500 UTC 19 August 2007 over western Oklahoma, approximately 500 miles inland from the Gulf of Mexico. It reached its peak intensity between 0600 UTC and 1200 UTC. Erin happened to move across an observation-dense area (Arndt et al., 2009) and, as such, its re-intensification process was fully captured by four Doppler radars located in the state of Oklahoma. At 0950 UTC, an eye-like feature was first observed in the radar reflectivity map and lasted for approximately three hours. This eye-like feature was the most noticeable characteristic during its re-intensification. A successful simulation should be able to reproduce this feature.

Three experiments are conducted: one without any radar data assimilation (NoDA), one with radar radial velocity data assimilated (VEL), and one with both radar reflectivity and radial velocity data assimilation (CTRRAD). Here, we use the Advanced Research Weather Research and Forecasting model (WRF-ARW) (Skamarock et al., 2005) as the forecast model and an enhanced version of the Grid-point Statistical Interpolation (GSI) three-dimensional variational (3DVAR) system (Wu et al., 2002) for data assimilation. The configuration of WRF-ARW follows that of the experimental High-Resolution Rapid Refresh system used in Hu et al. (2007). Radar data assimilation capabilities within the GSI were enhanced by the Center for Analysis and Prediction of Storms (CAPS) research group, with some of the details described in Hu et al. (2007). In this study, the radial velocity data used are the 3D radar mosaic gridded data at 1-km horizontal spacing produced by the National Severe Storm Laboratory (Zhang et al., 2005). The radial velocity data are preprocessed using a package borrowed from the Advanced Regional Prediction System (ARPS) (Brewster et al., 2005) and directly assimilated within the enhanced GSI 3DVAR system, while the reflectivity data are assimilated using a complex cloud analysis package adapted from the ARPS system and implemented within GSI after the variational analysis step. Details can be found in Hu et al. (2007) and an earlier example of applying the ARPS cloud analysis package to initialize WRF can be found in Hu and Xue (2007).

The initial analysis background and lateral boundary conditions are from NCEP operational North American Mesoscale Model analysis and forecasts, respectively. The domain has 881×881 horizontal grid points with a horizontal grid spacing of 3 km and 41 vertical levels. The experiment without radar data assimilation (NoDA) starts from 0000 UTC 19 August 2007. For the other two experiments, radar data are assimilated by GSI at 10-min intervals between 0000 and 0200 UTC using an intermittent data assimilation procedure (Hu and Xue, 2005). All forecasts end at 1800 UTC, covering the re-intensification and dissipation periods of Erin.

Figure 3 presents the observed reflectivity and the forecasts from the three experiments at 3 km MSL. From 0600 UTC to 1200 UTC, Erin was intensifying while moving northeastward. An eye feature was observed at 1200 UTC. After that, it began to weaken, and finally dissipated over northeastern Oklahoma. The NoDA experiment results fail to reach the intensity of a tropical storm. The forecasted rain band does not even rotate, not to mention the eye feature at 1200 UTC. The assimilation of radar velocity improves the circulation. The rain band structure shows closer resemblance to the observation. Especially at 1200 UTC, two fake rain bands that appear in NoDA are suppressed and organized into one narrow rain band. However, the VEL experiment also fails to reproduce the eye feature. With both radial winds and reflectivity assimilated, CTRRAD (see Fig. 3, third row) successfully simulates Erin's intensification and dissipation processes. It reproduces an eye feature at 1200 UTC, though the size is a little larger than observed. Compared to the other two experiments, the predicted rain bands during the intensification and dissipation stages are also closest to the observation. The main deficiency is that the forecast overestimates the rain area during the first few hours.

Figure 4 displays Erin's observed and forecasted tracks. For all forecasts, the predicted storm moves toward the northeast, consistent with observations. The problem is that the predicted storm moves slower than observed, with the NoDA prediction being the slowest. The assimilation of radial winds accelerates the motion. At 1200 UTC, the VEL prediction is between the NoDA prediction and the actual observation. However, the simulated storm decelerates again once the impact of radial wind disappears. With both radial wind and reflectivity assimilated, CTRRAD moves Erin fastest, resulting in the lowest track errors. At 1200 UTC, the predicted

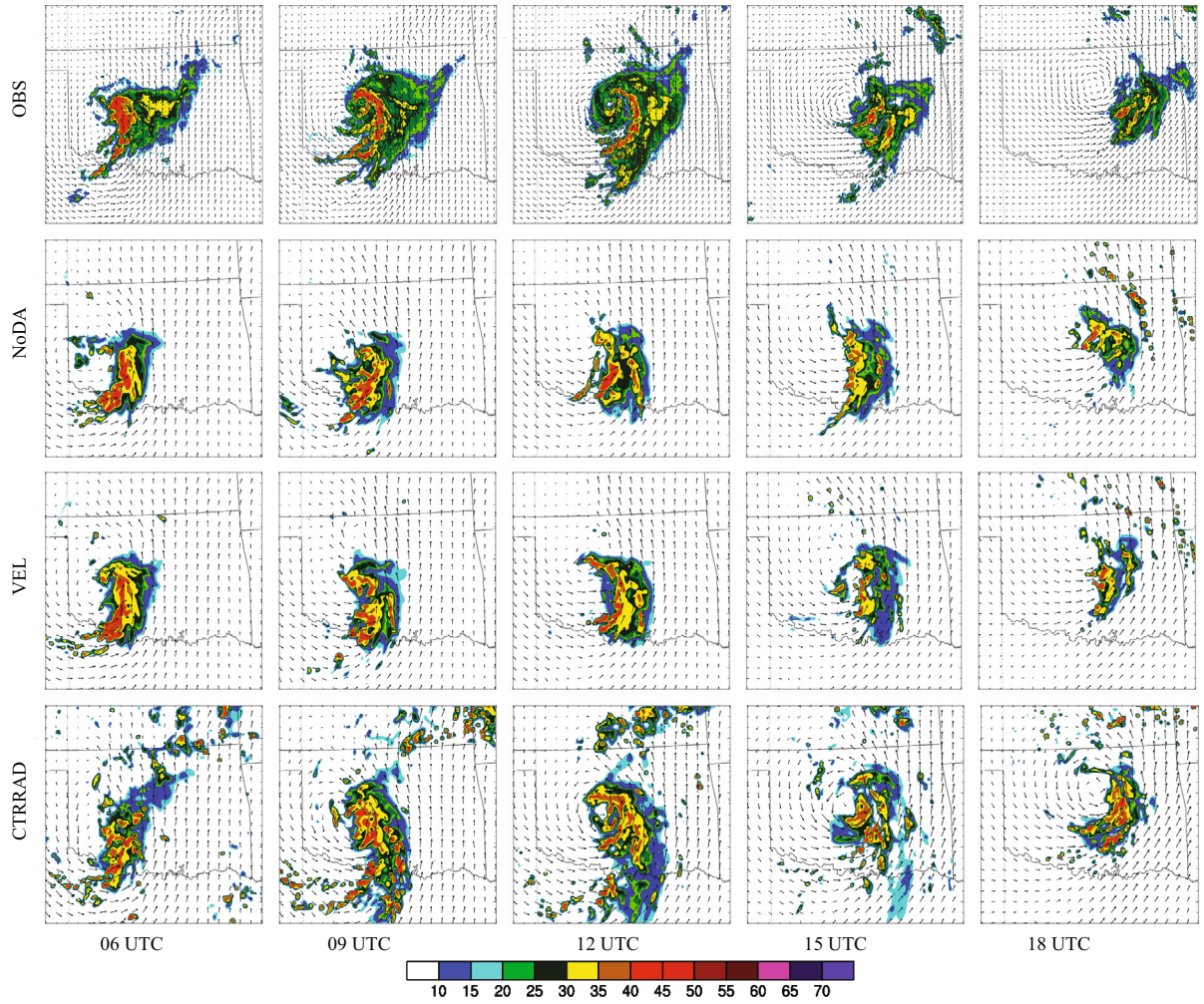

Fig. 3. The observed 3 km-height radar reflectivity (top row) and the corresponding forecasted reflectivity of NoDA (second row), VEL (third row) and CTRRAD (fourth row). The columns from left to right are times from 0600 to 1800 UTC.

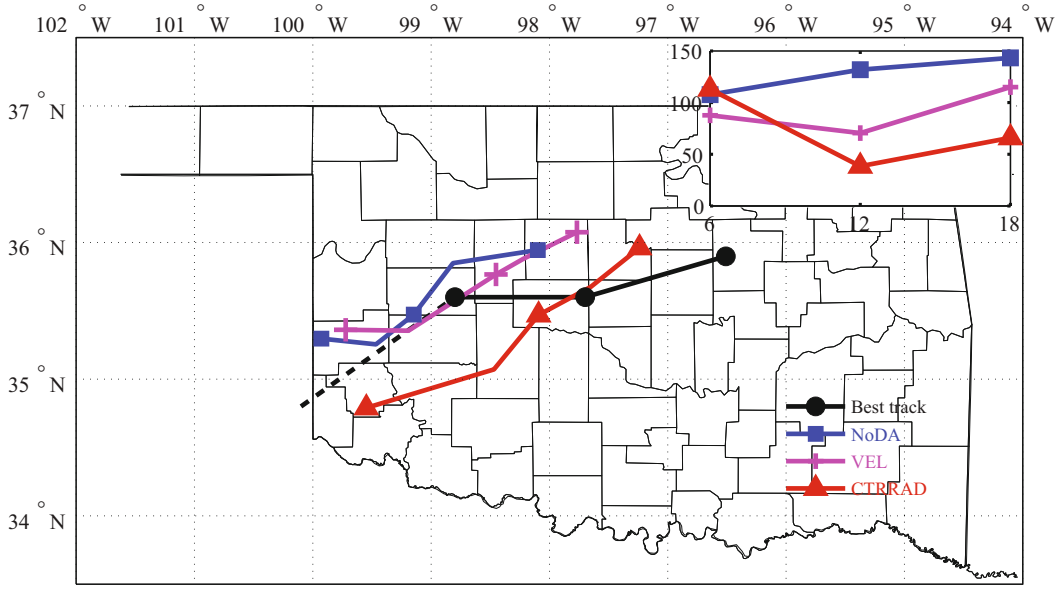

Fig. 4. National Hurricane Center best-track data and predicted tracks of Erin. The inner small panel in the upper-right corner is the track errors (km). The best-track is plotted every six hours, starting at 0000 UTC, while the predicted track is plotted every three hours, starting at 0600 UTC. All tracks end at 1800 UTC 21 August 2007.

storm center is merely 38 km away from the observed center.

Overall, the subjective assessment suggests that the CTRRAD forecast is the best in terms of both storm structure and track, followed by VEL. The NoDA prediction is the worst. This is especially true at 1200 UTC. At that time, both VEL and NoDA fail to reproduce the eye feature of Erin.

3.2. *Results of the object-based verification method*

In this subsection, to objectively evaluate the forecasts, an object-based method is employed. Here, only a high threshold (coverage threshold of 15 mm) is examined because the prediction of the heavy rainstorm is the main concern. Figure 5 presents an example of isolated objects from CTRRAD and NCEP Stage IV data at 1200 UTC. There are four objects in CTRRAD but only two of them match with the observation (see Figs. 5e and f). The isolated objects in the observation field merged together (Fig. 5d), while the forecast objects remain separate. The main objects in CTRRAD and the observations do match. Unfortunately, the forecast still has three objects that are not matched by any observations (see the blue

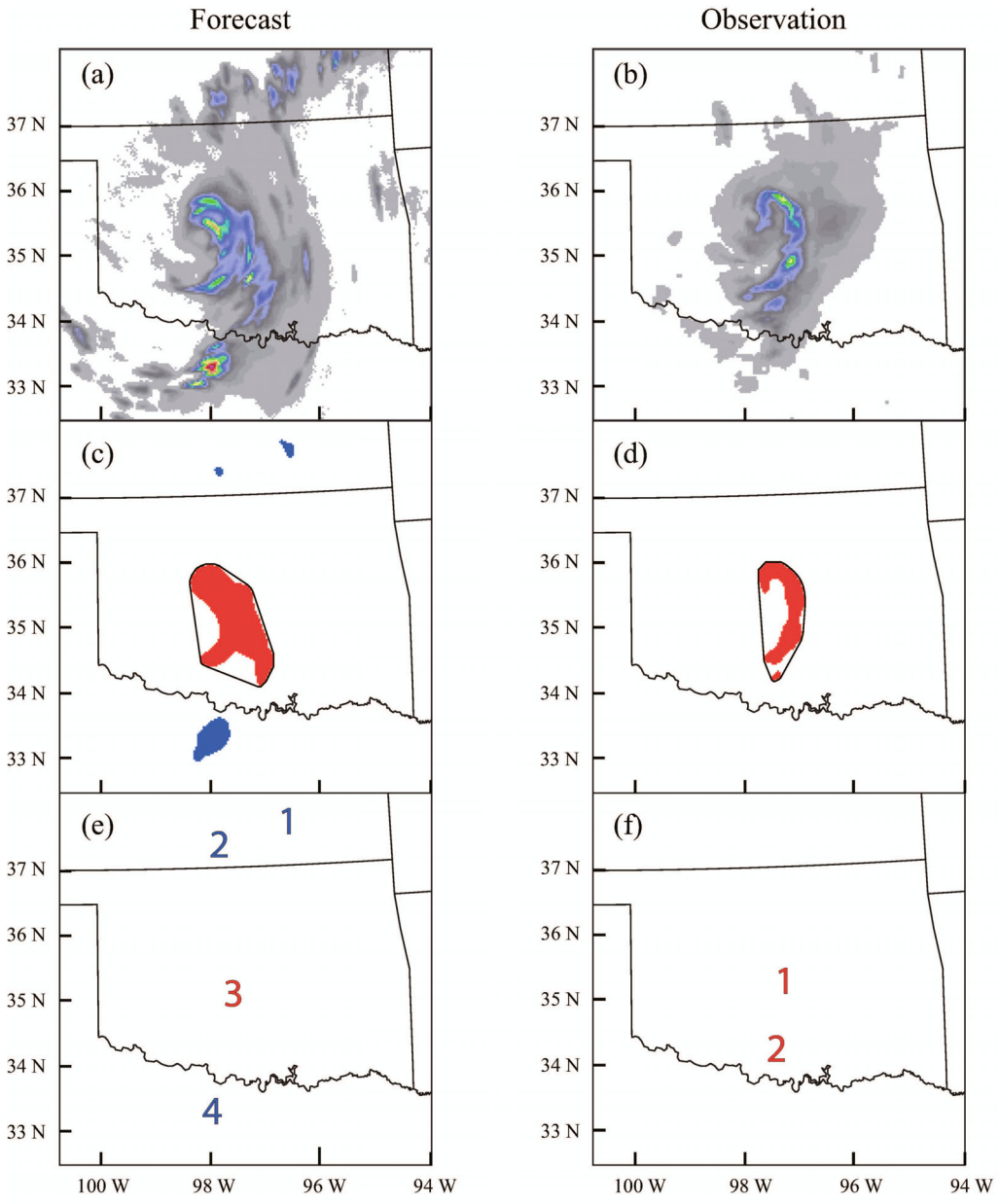

Fig. 5. METs isolated objects from CTRRAD and Stage IV precipitation valid at 1200 UTC 19 August 2007. Panels (a) and (b) are hourly accumulated rainfall. Here, the raw data threshold and the convolution threshold are 10 mm [light blue in (a)] and 15 mm [blue in (a)], respectively. Panels (c) and (d) are the isolated and merged objects; (e) and (f) are the serial numbers of the objects. Panels (a), (c) and (e) are from the forecast, while (b), (d) and (f) are from the observation.

patches in Fig. 5c). This means that CTRRAD produces an unrealistic rain storm. For the matched objects, the main feature is similar, but CTRRAD over-predicts the size and has a fake tail on the east side.

Figure 6 shows an example of using the isolated property centroid to draw the path of the moving storm and to calculate the position error. Clearly, NoDA has the largest position error. With the radial wind assimilated, VEL is able to correct its direction of motion. It can be observed that the position error is greatly reduced in the first few hours. However, the impact of radial wind only lasts for a few hours. As the impact of radial wind vanishes, the direction of motion is the same as in NoDA. The position error therefore increases again. With both reflectivity and radial wind assimilated, the storm in CTRRAD moves along the direction of the observed one, and, especially in the later hours, the direction of motion is almost the same as observed, except that the speed is a little slower. Its position error is smallest among all three experiments. At 1500 UTC, CTRRAD has only a 20 km position error (see inset panel in Fig. 6). The impact of assimilating full radar datasets seems to last longer. This is not surprising because the assimilation of full radar datasets has the potential to improve both dynamic and thermodynamic structures, while most of the direct adjustment is dynamic when only radial velocity is assimilated.

Figure 7 illustrates the index of total interest. The properties used for this plot include area (1), intersection area (2), centroid (2), boundary (4), intensity (1), angle (2), and convex hull (1), where the numbers in parentheses are the weights for each property. Most settings follow the MET default settings. The cross-sectional area between the forecast and observation and its relative position are still considered to be important elements. Shape properties such as the angle are given a higher weight, while the weight of intensity is lower. CTRRAD obtains a higher total interest than VEL for most of the forecast times, while both are better than NoDA. At 1200 UTC, the total interest indicates that CTRRAD has the best performance, followed by VEL, and subsequently NoDA. Compared to the neighborhood verification, this result is much more consistent with the earlier subjective assessment. Other settings, such as turning off the property intensity have also been tested. Problems occur when the parameters are changed. That is, the matched object pairs may change because some isolated objects are sensitive to a certain property while others are not. Therefore, manual procedures are necessary to ensure that the same matched object pairs are compared. Except for this deficiency, the general conclusion remains unchanged, as long as the settings are in the appropriate range.

All in all, the object-based verification also indicates that the CTRRAD prediction is the best, followed by VEL, and then NoDA. This result is consistent with the subjective assessment.

3.3. *Results of the neighborhood verification method*

Normally, if the forecast only has a displacement error, the neighborhood method can help distinguish between a bad and good forecast by using proper neighborhood width and coverage threshold. However, in our case, the position error is not the main issue: the size and intensity are.

Figure 8 presents the results of neighborhood CSI scores. For the small threshold of 1.25 mm h^{-1}, the neighborhood method has no problem distinguishing good from bad forecasts. The CSI scores indicate the VEL has better performance than NoDA for most of the forecast times (see Figs. 8a and b), while CTRRAD is the worst. Here, CTRRAD is not expected to be the best because it clearly over-forecasts the size. The controversy lies in the results of the high threshold of 15 mm h^{-1}. For that threshold, the over-forecast issue is not as serious as the lower threshold (1.25 mm h^{-1}).

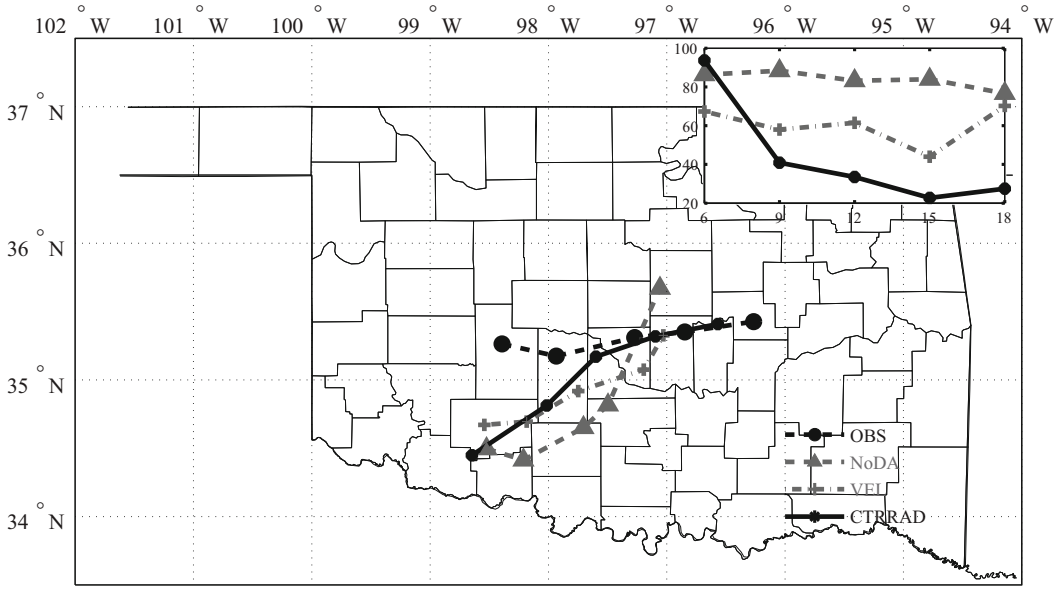

Fig. 6. Similar to Fig. 4, except for the geometrical center of the rain storm. The inset panel shows the distance between two rain storms.

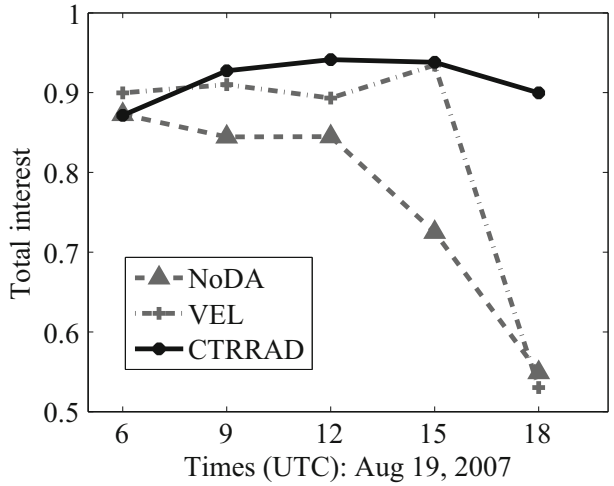

Fig. 7. The total interest index of CTRRAD (solid black line), VEL (dashed-dotted line) and NoDA (dashed line).

CTRRAD has the smallest position error and the best shape prediction. At 1200 UTC, it successfully reproduces the eye feature. However, when using a small neighborhood width of two grid points, CTRRAD is no better than VEL (see Fig. 8c). This is because the assimilation of radar reflectivity results in a larger area of fake rain, which is probably due to the over-adjustment of water vapor content (Zhao and Xue, 2009; Schenkman et al., 2011). When using a large neighborhood width of eight grid points, CTRRAD is better than VEL. The problem is that in the later hours, VEL is worse than NoDA (see Fig. 8d). The fixed neighborhood fails to give a result that is consistent with the subjective assessment.

3.4. *Results of percentile-based neighborhood verification*

Verification scores such as the CSI use a category method. The threshold determines the boundary of the verification objects. The larger the cross-sectional area between forecast and observation, the higher the verification score. Because the neighborhood method can reduce the impact of displacement error by using proper neighborhood configurations, the remaining question is how to obtain reasonable object pairs between forecast and observation. The traditional method uses a fixed threshold. However, in reality, quantitative precipitation forecasting remains a great challenge for numerical models. Even with a similar pattern, the intensity may differ greatly for different experiments. The use of the same threshold will contain not only the intensity error but also the size error. In this case study, when the 15 mm threshold is used, although the main rain band of CTRRAD is similar to the observation, the identified object is approximately twice as large in terms of width (see Fig. 5). NoDA and VEL have

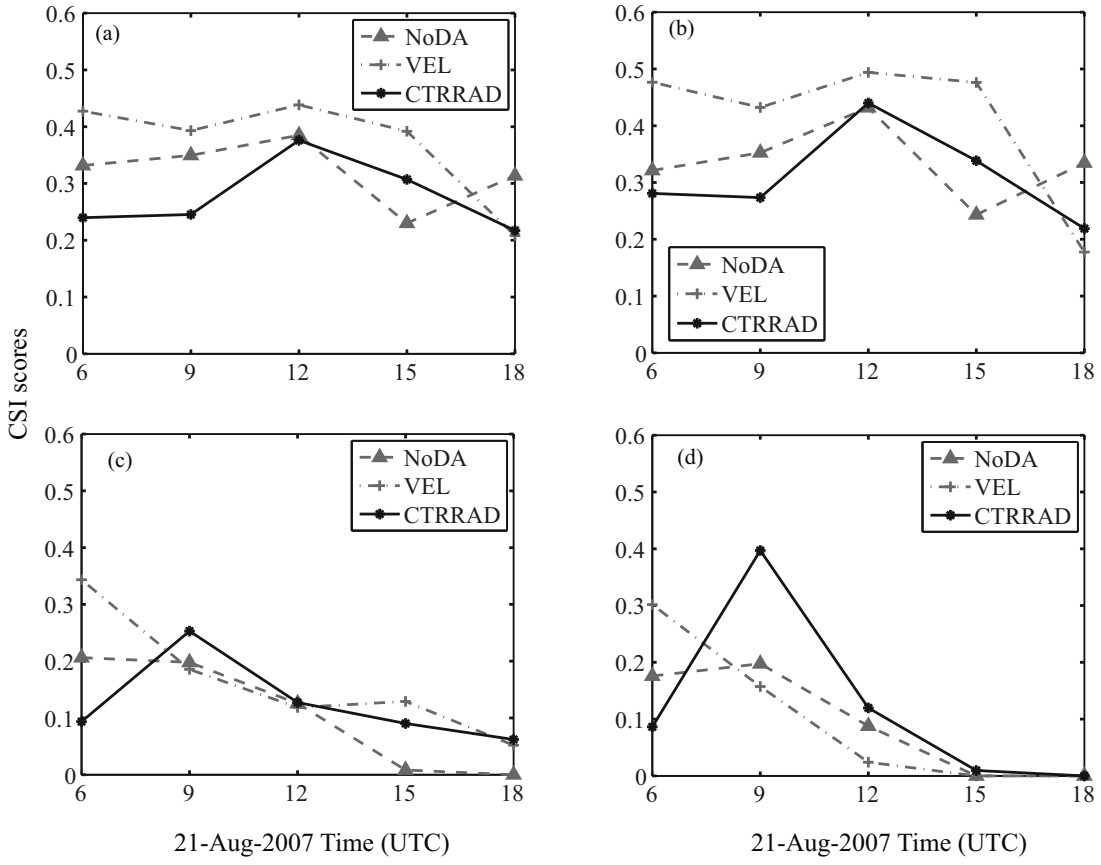

Fig. 8. Hourly neighborhood CSI scores from Fig. 3, with the coverage threshold set to 50%: (a) threshold = 1.25 mm h^{-1} and neighborhood width r is two grid intervals; (b) threshold = 1.25 mm h^{-1}, r = eight grid intervals; (c) threshold = 15 mm h^{-1}, r = two grid intervals; and (d) threshold = 15 mm h^{-1}, r = eight grid intervals.

similar issues, though the size is not as large as with CTRRAD (not shown). Thus, in this study, the 90th percentile value is used to replace the fixed threshold. Compared to the 15 mm threshold, the size of the observation does not change much; the sizes of VEL and NoDA are much closer to the observation, while the size of CTRRAD is greatly reduced (not shown).

Figure 9 displays the percentile-based neighborhood CSI scores. The neighborhood width is the same as that in Fig. 8, except that the coverage threshold is reduced by 30%. When a neighborhood width of two grid points is used, CTRRAD is the best for almost all forecast times except 0600 UTC. At 1200 UTC, CTRRAD has better performance than VEL, while NoDA is the worst. When a neighborhood width of 8 grid points is used, the advantage of CTRRAD over VEL and NoDA is more obvious. The CSI scores are much more consistent with the subjective assessment when compared to the fixed threshold.

As a further attempt, we use a series of other combinations of configurations. Herein, the fixed threshold of 15 mm h^{-1} is presented as a comparison. For the fixed threshold, when a large coverage threshold of 50% is adopted, the CSI scores of NoDA and VEL decrease as the neighborhood width increases. However, for CTRRAD, the CSI first increases then decreases (Fig. 10a). For a small coverage threshold of 30%, the CSI scores of all the experiments first increase then decrease (Fig. 10c). The results of the different experiments indicate different behaviors as the neighborhood configurations change. However, the general conclusion remains the same: VEL is the best, followed by CTTRAD, and NoDA is the worst. Note that CTRAD does surpass the other two when the neighborhood width is larger than eight grid points (Fig. 10a). The problem is that at that range, the VEL and NoDA obtain "NaN" values in the later forecast hours. Therefore, the results indicated by a larger neighborhood width (e.g., in Fig. 10a when the neighborhood width is larger than eight) becomes meaningless. Figures 10b and d show the percentile-based CSI scores. All CSI scores are between 0 and 1, and "NaNs" are avoided. Moreover, for almost all settings, CTRRAD outperforms VEL, especially for the larger neighborhood width, while both are consistently better than NoDA. Therefore, the result of percentile-based neighborhood verification is more consistent with the subjective assessment. Compared to the fixed threshold method, the percentile-based CSI scores reduce the impact of intensity error.

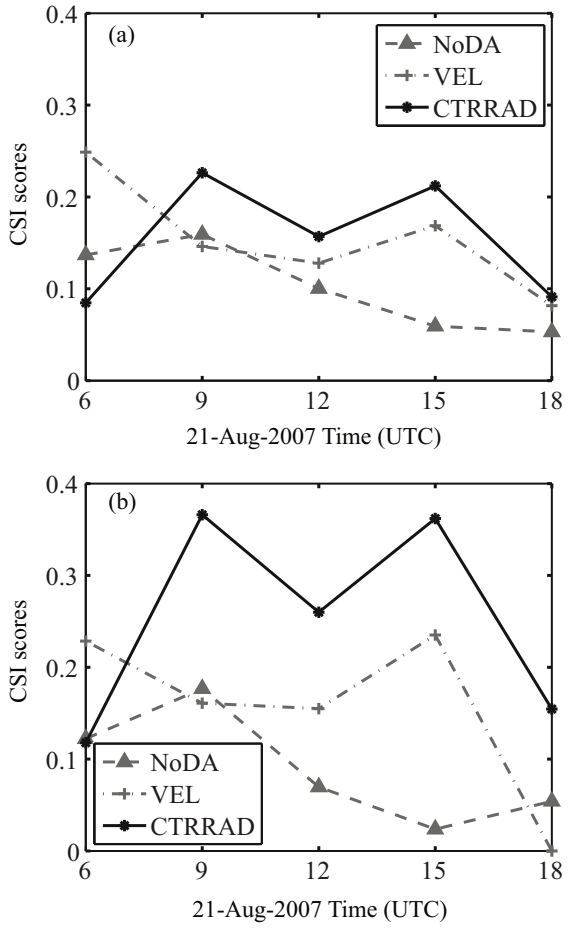

Fig. 9. Percentile-based neighborhood CSI scores for a threshold of 15 mm h^{-1}, with a coverage threshold of 30%: (a) neighborhood width r = two grid intervals; (b) r = eight grid intervals.

4. Summary and discussion

In this paper, a percentile-based neighborhood method is proposed and used to calculate a category verification score, i.e., the CSI. The purpose of using a percentile-based instead of a fixed threshold is to reduce the impact of forecast intensity bias on the calculation of verification scores. A case of a re-intensifying tropical storm after landfall is selected for the purpose of examining radar data assimilation impact. A key feature of this tropical storm is the eye feature during its re-intensification. The forecast without radar data assimilation (experiment NoDA) fails to reach the intensity of a tropical storm, and the predicted rain band is the worst based on the side-by-side subjective comparisons with observations. With the radial velocity data assimilated (experiment VEL), the rain band structure is improved and the track error is reduced, but the experiment still fails to reproduce the eye feature. When the reflectivity data are assimilated together with the radial velocity data (experiment CTRRAD), the eye is successfully reproduced, although the size of the eye is somewhat larger. The track error of CTRRAD is the smallest among the three experiments. However, CTRRAD over-forecasts the rain intensity and size.

To objectively demonstrate that the forecast with the rain band structure is overall the best, an object-based evaluation method within MET is employed. The object-based method calculates the geometric properties such as area, centroid, curvature, angle, and so on. The evaluation results are close to the subjective assessment. The index of total interest, which weights various properties, is used to distinguish the three forecasts. CTRRAD is found to out-perform VEL, while both of them are better than NoDA. This result is con-

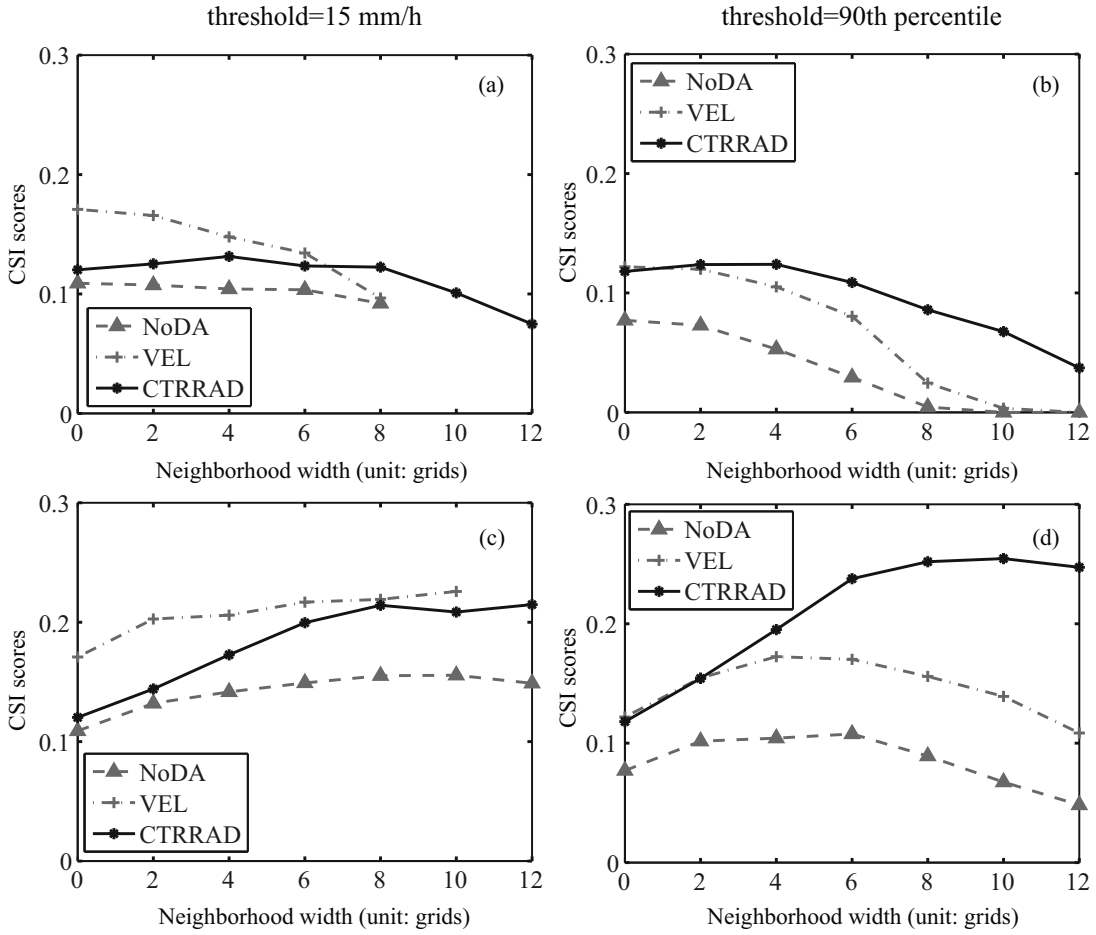

Fig. 10. The mean CSI scores of five verified moments from 0600 UTC to 1800 UTC for (a) the threshold 15 mm h^{-1} and (b) the 90th percentile. The coverage threshold is 50%. Panels (c) and (d) are the same as (a) and (b) but with the coverage threshold set to 30%. Here, the x-axis is the neighborhood width.

sistent with the subjective assessment.

However, the traditional fixed threshold neighborhood fails to indicate that CTRRAD is the best because of the over-forecasting problems. Instead, for most of the settings, VEL, which has a relatively clean forecast, and is better than CTRRAD. With the percentile threshold, the neighborhood CSI score indicates that CTRRAD consistently outperforms VEL, while the latter is always better than NoDA. This result is consistent with the object-based verification as well as the subjective assessment. The percentile-based method is therefore better at handling the forecast intensity than the fixed threshold.

Finally, we note that the percentile threshold could also be combined with object-based methods in which the boundaries of the isolated objects are determined by a convolution threshold. As shown in Fig. 5, the use of a fixed threshold results in over-sized objects, which impacts the calculation of various properties. If the percentile threshold is used, it may match the observations much better. At this point, we leave this work for future investigation. We also point out that the results of this paper are based on a single case only. The methods should be tested in the future with more cases to obtain more robust results.

Acknowledgements. This work was primarily supported by the National 973 Fundamental Research Program of China (Grant No. 2013CB430103) and the Department of Transportation Federal Aviation Administration (Grant No. NA17RJ1227) through the National Oceanic and Atmospheric Administration. The work was also supported by the National Science Foundation of China (Grant No. 41405100) and the Fundamental Research Funds for the Central Universities (Grant No. 20620140343).

REFERENCES

Arndt, D. S., J. B. Basara, R. A. McPherson, B. G. Illston, G. D. McManus, and D. B. Demko, 2009: Observations of the overland reintensification of tropical storm Erin (2007). *Bull. Amer. Meteor. Soc.*, **90**, 1079–1093.

Brewster, K., M. Hu, M. Xue, and J. Gao, 2005: Efficient assimilation of radar data at high resolution for short-range numerical weather prediction. *WWRP Int. Symp. Nowcasting Very Short Range Forecasting*, Tolouse, France, WMO, Symposium CD, Paper 3.06.

Brown, B. G., J. H. Gotway, R. Bullock, E. Gilleland, and D. Ahijevych, 2009: The Model Evaluation Tools (MET): Community tools for forecast evaluation. 25*th Conf. Int. Interactive Information and Processing Systems (IIPS) for Meteorology*,

Oceanography, *and Hydrology*, Paper 9A. 6.

Brown, B. G., R. Bullock, J. H. Gotway, D. Ahijevych, C. A. Davis, E. Gilleland, and L. Holland, 2007: Application of the MODE object-based verification tool for the evaluation of model precipitation fields. 22*nd Conf. Wea. Anal. Forecasting/*18*th Conf. Num. Wea. Pred.*, Park City, Utah, Amer. Meteor. Soc., Paper 10A. 2.

Clark, A. J., W. A. Gallus Jr., and M. L. Weisman, 2010: Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF model simulations and the operational NAM. *Wea. Forecasting*, **25**, 1495–1509.

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

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

DTC, 2009: *Model Evaluation Tools Version 2.0 (METv2.0) User's Guide*. Boulder, Colorado, USA. [Available online at http://www.dtcenter.org/met/users/docs/users_guide/MET_Users_Guide_v2.0_rev2.pdf.]

Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteorological Applications*, **15**, 51–64.

Ebert, E. E., 2009: Neighborhood verification: A strategy for rewarding close forecasts. *Wea. Forecasting*, **24**, 1498–1510.

Hu, M., and M. Xue, 2005: Impact of configurations of rapid intermittent assimilation of WSR-88D radar data for the 8 May 2003 Oklahoma City tornadic thunderstorm case. *Mon. Wea. Rev.*, **135**, 507–525.

Hu, M., and M. Xue, 2007: Implementation and evaluation of cloud analysis with WSR-88D reflectivity data for GSI and WRF-ARW. *Geophys. Res. Lett.*, **34**, L07808, doi: 07810.01029/02006GL028847.

Hu, M., S. Weygandt, M. Xue, and S. Benjamin, 2007: Development and testing of a new cloud analysis package using radar, satellite, and surface cloud observations within GSI for initializing rapid refresh. 22*nd Conf. Wea. Anal. Forecasting/*18*th Conf. Num. Wea. Pred.*, Salt Lake City, Utah, Amer. Meteor. Soc., CDROM P2.5.

Johannes, J., F. Christoph, and S. Cornelia, 2008: Quantile-based short-range QPF evaluation over Switzerland. *Meteor. Z.*, **17**, 827–848.

Lin, Y., and K. E. Mitchell, 2005: The NCEP Stage II/IV hourly precipitation analyses: Development and applications. 19*th Conf. Hydrology*, San Diego, CA, Amer. Meteor. Soc., Paper 1.2.

Rezacova, D., Z. Sokol, and P. Pesice, 2007: A radar-based verification of precipitation forecast for local convective storms. *Atmospheric Research*, **83**, 211–224.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev*, **136**, 78–97.

Schenkman, A. D., M. Xue, A. Shapiro, K. Brewster, and J. D. Gao, 2011: The analysis and prediction of the 8–9 May 2007 Oklahoma tornadic mesoscale convective system by assimilating WSR-88D and CASA radar data using 3DVAR. *Mon. Wea. Rev.*, **139**, 224–246.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. D. Powers, 2005: A description of the advanced research WRF Version 2. National Center for Atmospheric Research, Boulder, Colorado, USA, 88 pp.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 464 pp.

Wu, W.-S., R. J. Purser, and D. F. Parrish, 2002: Three-dimensional variational analysis with spatially inhomogeneous covariances. *Mon. Wea. Rev.*, **130**, 2905–2916.

Zhang, J., K. Howard, and J. J. Gourley, 2005: Constructing three-dimensional multiple-radar reflectivity mosaics: Examples of convective storms and stratiform rain echoes. *J. Atmos. Oceanic Technol.*, **22**, 30–42.

Zhao, K., and M. Xue, 2009: Assimilation of coastal Doppler radar data with the ARPS 3DVAR and cloud analysis for the prediction of Hurricane Ike (2008). *Geophys. Res. Lett.*, **36**, L12803, doi: 10.1029/2009GL038658.