Methods for Comparing Simulated and Observed Satellite Infrared Brightness Temperatures and What Do They Tell Us?

SARAH M. GRIFFIN, JASON A. OTKIN, CHRISTOPHER M. ROZOFF, JUSTIN M. SIEGLAFF, AND LEE M. CRONCE

Cooperative Institute for Meteorological Satellite Studies, University of Wisconsin–Madison, Madison, Wisconsin

CURTIS R. ALEXANDER

NOAA/Earth System Research Laboratory, Boulder, Colorado

(Manuscript received 18 May 2016, in final form 9 September 2016)

ABSTRACT

In this study, the utility of dimensioned, neighborhood-based, and object-based forecast verification metrics for cloud verification is assessed using output from the experimental High Resolution Rapid Refresh (HRRRx) model over a 1-day period containing different modes of convection. This is accomplished by comparing observed and simulated Geostationary Operational Environmental Satellite (GOES) 10.7-μm brightness temperatures (BTs). Traditional dimensioned metrics such as mean absolute error (MAE) and mean bias error (MBE) were used to assess the overall model accuracy. The MBE showed that the HRRRx BTs for forecast hours 0 and 1 are too warm compared with the observations, indicating a lack of cloud cover, but rapidly become too cold in subsequent hours because of the generation of excessive upper-level cloudiness. Neighborhood and object-based statistics were used to investigate the source of the HRRRx cloud cover errors. The neighborhood statistic fractions skill score (FSS) showed that displacement errors between cloud objects identified in the HRRRx and GOES BTs increased with time. Combined with the MBE, the FSS distinguished when changes in MAE were due to differences in the HRRRx BT bias or displacement in cloud features. The Method for Object-Based Diagnostic Evaluation (MODE) analyzed the similarity between HRRRx and GOES cloud features in shape and location. The similarity was summarized using the newly defined MODE composite score (MCS), an area-weighted calculation using the cloud feature match value from MODE. Combined with the FSS, the MCS indicated if HRRRx forecast error is the result of cloud shape, since the MCS is moderately large when forecast and observation objects are similar in size.

1. Introduction

Being able to accurately forecast cloud cover in the near term has many beneficial applications. Cloud cover and cloud properties are useful for forecasting convective initiation (Mecikalski and Bedka 2006; Sieglaff et al. 2011) and severe weather (Purdom 1993; Cintineo et al. 2013). Cloud cover has important aviation implications. Most of the air traffic delays in the United States are the result of thunderstorms (Kaplan et al. 2005; Murray 2002; Mecikalski et al. 2007), and thunderstorms are a cause of a documented in-flight aviation hazard: convectively induced turbulence (Hamilton and Proctor 2002). Furthermore, cloud cover has an implication on daily temperatures, as it is negatively correlated with the diurnal temperature range (Karl et al. 1993; Dai et al. 1999).

Different statistical techniques exist that can be used to assess the skill of a numerical weather prediction (NWP) model at predicting variables, including cloud cover. Typically, to assess the predictive skill of an NWP model, the NWP forecast and model-derived fields are compared directly to their accompanying observational fields (DelSole and Tippett 2014). However, there are many different metrics available to assess forecast skill. One type is dimensioned verification metrics, such as mean absolute error (MAE) and mean bias error (MBE) (Wilks 2006). Other metrics include neighborhood-based statistics such as

Corresponding author e-mail: Sarah M. Griffin, sarah.griffin@ssec.wisc.edu

DOI: 10.1175/WAF-D-16-0098.1

© 2017 American Meteorological Society

the fractions skill score (FSS; Roberts and Lean 2008; Roberts 2008) or object-based statistics (Davis et al. 2006, 2009). Each metric has its strengths and weaknesses. For example, dimensioned metrics are easy to implement. However, they often penalize high-resolution NWP forecasts (Mass et al. 2002; Done et al. 2004) since they require near-perfect correspondence between the forecast and observation fields, which is difficult to achieve at higher resolutions. Neighborhood-based statistics are less sensitive to spatial scale (Wolff et al. 2014) but require an arbitrary threshold to be applied to the field of interest. Object-based statistics can account for spatial displacement (Clark et al. 2014), but they can be more difficult to employ and require the use of numerous user-defined parameters to identify objects.

Numerous studies have employed the above metrics to explore the skill of high-resolution precipitation and cloud cover forecasts, including forecasts from convection-allowing models. Brown and Comrie (2002) applied the MBE to assess the accuracy of the 1 km × 1 km regression model they developed based on data from the National Climatic Data Center and used to estimate average precipitation in the southwest United States. Schwartz (2014) compared 3-km High Resolution Rapid Refresh (HRRR) precipitation forecasts to other high-resolution models using the FSS. Söhne et al. (2006) used the FSS to compare simulated brightness temperatures (BTs) from the French mesoscale model MesoNH, initialized using different analyses and horizontal grid resolutions, to observed Meteosat Second Generation Scanning Enhanced Visible Infrared Imager (SEVIRI) infrared BTs. Object-based statistics can be calculated using the Method for Object-Based Diagnostic Evaluation (MODE). MODE has been used to compare precipitation forecasts from the HRRR (Bytheway and Kummerow 2015; Cai and Dumais 2015), as well as cloud cover from a convection-permitting and near-convection-resolving Met Office Unified Model on a 2-km horizontal grid over the United Kingdom (Mittermaier and Bullock 2013).

The purpose of the paper is to demonstrate how the different verification metrics used in the above studies, as well as the introduction of a new metric, can be used to assess the accuracy of the experimental HRRR (HRRRx) cloud cover forecasts. In particular, model cloud validation will be performed with the indirect method of comparing HRRRx-simulated infrared BTs to observed infrared BTs (Morcrette 1991; Otkin and Greenwald 2008; Otkin et al. 2009; Cintineo et al. 2014, Lee et al. 2014; Thompson et al. 2016). Cloud validation will be carried out using dimensioned metrics, as well as neighborhood-based and object-based statistics, to deduce how each metric can be used to quantify the accuracy of HRRRx-simulated BTs and, therefore, the accuracy of the HRRRx cloud cover. Another objective of this manuscript is to represent each metric of forecast accuracy as a single number, since some metrics, like object-based statistics, can provide multiple methods of displaying forecast accuracy.

While not the focus of this manuscript, the methodology presented will be leveraged in future work to potentially improve operational forecasting in at least two ways. The metrics presented are computed in real time, allowing for a quick assessment of the HRRRx cloud cover accuracy and efficient determination of which HRRRx initialization best represents the observed cloud cover at a given time. Second, the statistics can be accumulated over a long time period to determine if any systematic errors exist in the HRRRx-simulated BTs, including errors associated with the diurnal cycle, specific initialization times, weather regimes, or seasons.

The manuscript is organized as follows. The datasets used in this study are described in section 2. Each verification metric will be presented in more detail in section 3, and the methodology is described in section 4. Results will be presented in section 5, and discussion and conclusions, respectively, will be shown in sections 6 and 7.

2. Data

a. *Experimental High Resolution Rapid Refresh model-simulated brightness temperatures*

The model data used in this study are generated from the HRRRx model. HRRRx was implemented at Earth System Research Laboratory on 4 May 2015 (Earth System Research Laboratory 2016). HRRRx, which covers the continental United States (CONUS), is an hourly updating model that uses 3-km horizontal grid spacing and 51 vertical levels. HRRRx uses initial conditions from the Rapid Refresh model and then applies data assimilation at 3 km including the assimilation of radar reflectivity. HRRRx is a convection-allowing model that does not include deep convective parameterization (Benjamin et al. 2016). HRRRx uses the Thompson aerosol version 3.6.1 microphysics, version 3.6 of the Mellor–Yamada–Nakanishi–Niino (MYNN) scheme with the planetary boundary layer, the RUC land surface model, and RRTMG shortwave and longwave radiation (Earth System Research Laboratory 2016).

Simulated GOES 10.7-μm BTs from the HRRRx model are used in this study. The 10.7-μm wavelength is an infrared window band that is sensitive to cloud-top properties when clouds are present and to surface skin temperature when clouds are absent. HRRRx-simulated

BTs are available hourly for forecast hours (FHs) 0–24. Simulated *GOES-13* BTs are computed for each forecast time using HRRRx model output and version 2.0.7 of the Community Radiative Transfer Model (CRTM; Han et al. 2006) in the Unified Post Processor, which incorporates the correct GOES viewing angle geometry. For clear grid points, simulated BTs are computed using several model-predicted fields, such as surface skin temperature, 10-m wind speed, pressure, and vertical profiles of temperature and water vapor. For cloudy grid points, additional information about cloud radiative properties is required to calculate the simulated BTs. Vertical profiles of mixing ratio and number concentration are used to compute the effective particle diameters for each hydrometeor species (cloud water, cloud ice, rainwater, snow, and graupel) predicted by the Thompson aerosol microphysics scheme. In the CRTM, standard lookup tables for cloud optical properties, such as extinction, single-scatter albedo, and the full scattering phase function are used to assign values to each hydrometeor species as a function of the cloud effective diameter computed using the particle size distribution assumptions for that scheme (e.g., Otkin et al. 2007). Cloud optical properties are computed for each species and model layer and then combined into an effective set of properties for each layer before computing the simulated infrared BTs.

b. *Observed brightness temperatures*

The satellite validation data used during this study are derived from the *GOES-13* imager. The 10.7-μm GOES BTs have a 4-km spatial resolution at nadir and are remapped to the 3-km HRRRx grid using a weighted average of all the observed pixels overlapping a given HRRRx model grid box. The GOES imager typically completes a scan over the CONUS every 15 min except for every 3 h (0000, 0300 UTC, etc.) when the scan at the top of the hour is skipped so the full-disk scan that started 15 min earlier can be completed. Thus, simulated HRRR BTs will be compared with the 0-min scan for most hours, but will be compared with the full-disk scan starting 15 min prior to the HRRRx forecast time for cases when the 0-min CONUS scan is skipped. This introduces some uncertainty in the analysis, but this is expected to be minor.

c. *Case study time period*

HRRRx-simulated BT imagery and GOES-observed BT observations from 1200 UTC 23 July to 1200 UTC 24 July 2015 are used during this analysis. This time frame was chosen because it contains a variety of severe weather regimes over the United States that allows us to more thoroughly assess the characteristics of each statistical method. Severe weather reports compiled by the Storm Prediction Center can be seen in Fig. 1a, with the colored boxes representing the different focus areas of this study. The red box represents the northern plains, the black box represents the central plains, and the blue box represents the southeast United States. These regions were chosen to correspond to those on the SPC mesoscale analysis web page (http://www.spc.noaa.gov/exper/mesoanalysis/). Snapshots of the GOES BTs for each sector (Figs. 1b–d) help illustrate the nature of convection in each severe weather regime. The weather in the northern plains sector is characterized by a surface low pressure in Canada and frontal passage extending into the United States, causing the broad cloud field seen in Fig. 1b. There are single-cellular cloud features in the central plains, as seen in Fig. 1c. A broad cloud field with localized colder cloud tops is observed in the southeast sector (Fig. 1d). These cloud features are primarily the product of a stationary front and sea breezes.

3. Verification measures

a. *Dimensioned metrics*

Two dimensioned metrics are used during this analysis. These metrics are called "dimensioned" since they have the same units as the variable of interest (Willmott and Matsuura 2005) and are point based, comparing a single grid point in a model field with the same grid point in the observational field. Dimensioned metrics are useful for assessing model forecast skill because they are easy to compute and provide an efficient way to compare forecasts with point or gridded observations.

The first metric used in this analysis is the MAE, defined as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{n}|F_i - O_i|, \tag{1}$$

where F represents the HRRRx-simulated BTs and O represents the GOES-observed BTs. The MAE represents the overall model error, and it is deemed to be a more appropriate measure for model comparison than the root-mean-square error (Willmott and Matsuura 2005) since the HRRRx errors do not follow a normal distribution (Chai and Draxler 2014). A perfect MAE has a value of zero. The second metric is the MBE, defined as

$$\mathrm{MBE} = \frac{1}{N}\sum_{i=1}^{n}(F_i - O_i). \tag{2}$$

The MBE indicates model bias, with a positive (negative) MBE meaning the HRRRx-simulated BTs are too warm (cold) compared with the GOES observations.

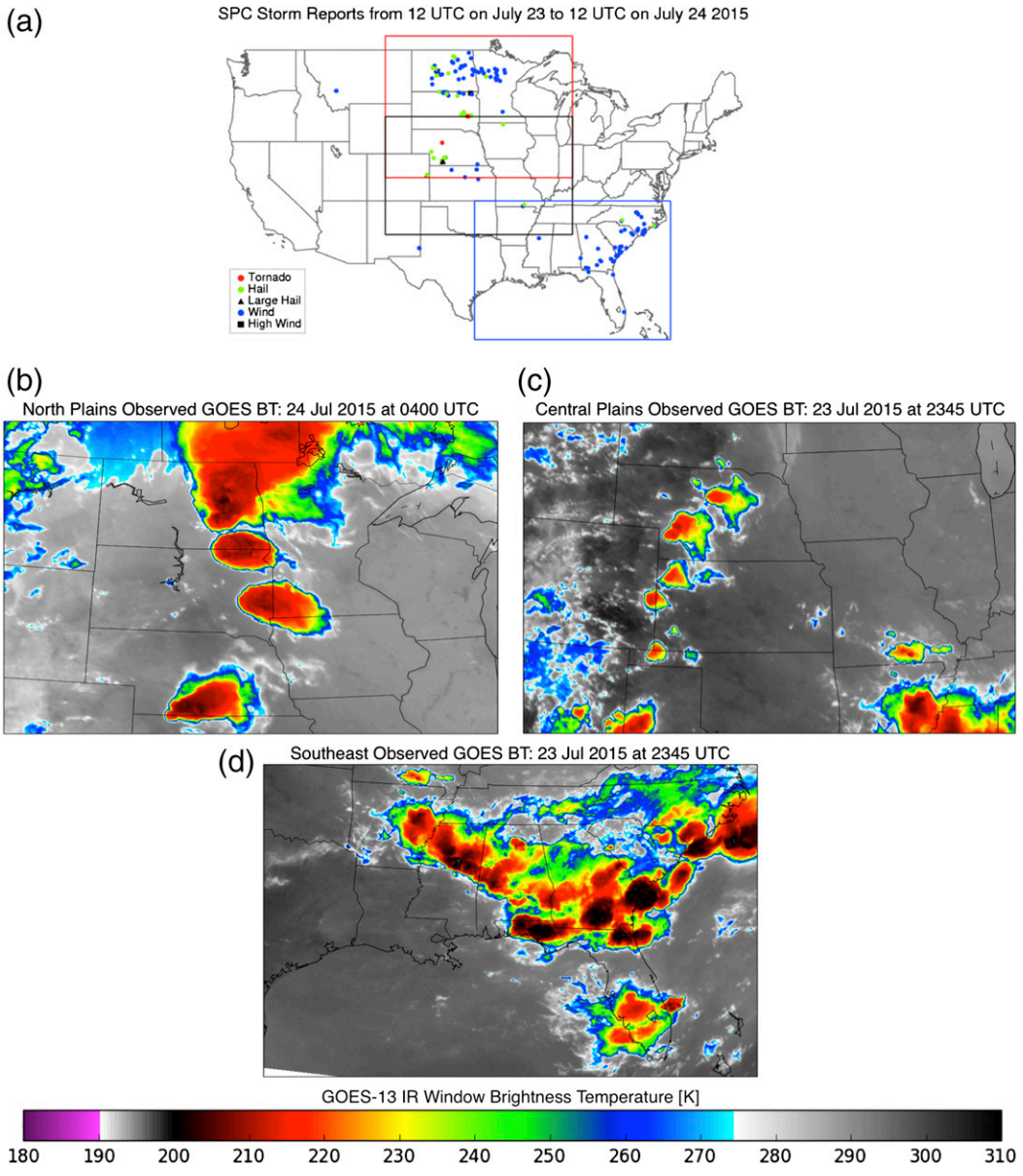

FIG. 1. (a) Storm Prediction Center storm reports from 1200 UTC 23 Jul to 1200 UTC 24 Jul 2015. Colored boxes represent the different sectors focused on in this analysis and are characterized by different causes of severe weather. IR window BT image for (b) the north plains sector at 0400 UTC 24 Jul 2015 and (c) the central plains and (d) southeast sectors at 2345 UTC 23 Jul 2015.

b. *Fractions skill score*

The FSS is a neighborhood-based statistic that is less sensitive to spatial errors than traditional gridpoint statistics (Mittermaier and Roberts 2010). While the FSS can be characterized as a point-based statistic like the dimensioned metrics, its value is determined by the grid points located within a specified region (defined as $n \times n$ grid points) surrounding each grid point. Therefore, the FSS is an objective measure of how the forecast skill varies with spatial scale (Wolff et al. 2014) and can also provide an assessment of displacement errors (Mittermaier and Roberts 2010). Benefits of the FSS include it being less sensitive to small-scale or displacement errors compared to dimensioned metrics while still remaining easy to implement.

The FSS is fully described in Roberts and Lean (2008), but a shorter description is provided here for context. First, binary yes/no forecast and observation fields of ones (zeros) exceeding (not exceeding) a

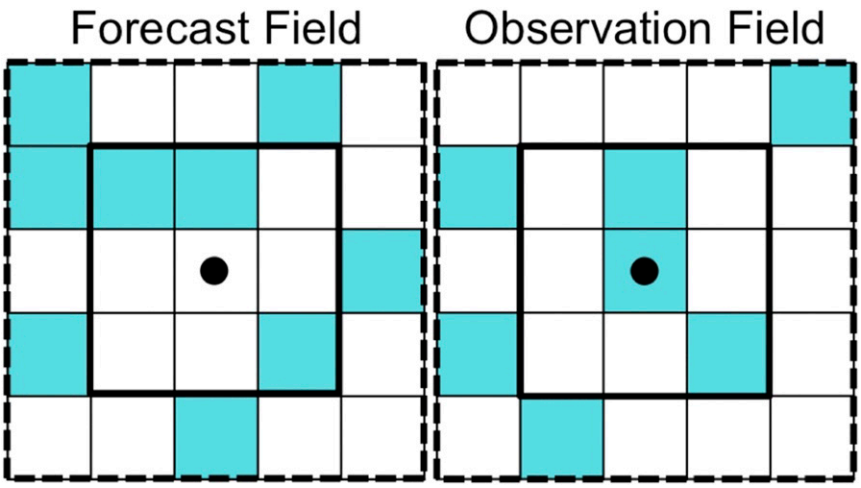

FIG. 2. An example of FSS neighborhood calculations. Blue (white) squares represent grid points exceeding (not exceeding) the threshold and thus having a value of 1 (0). At the grid point identified with the circle, the forecast (observation) value for square length of 3 (solid square) is 3/9 (3/9). The forecast (observation) value for square length of 5 (dashed square) is 9/25 (7/25). [Adapted from Roberts and Lean (2008) see their Fig. 2.]

given threshold are created. Then, fractions are computed for each forecast (observation) grid point based on the surrounding points within a square length n grid points in the forecast (observation) binary field. The number of grid points in the square surrounding a given point is therefore $n \times n$. An example of calculating the fraction for a single grid point can be seen in Fig. 2, adapted from Roberts and Lean (2008). Blue (white) squares represent grid points exceeding (not exceeding) the chosen threshold. At the center grid point, identified with a circle, the forecast (observation) fraction for a square length n of 3 points (solid square) is 3/9 (3/9). The forecast (observation) fraction for a square length n of 5 points (dashed square) is 9/25 (7/25). Finally, the FSS is calculated using the equation

$$\mathrm{FSS} = 1 - \frac{\mathrm{MSE}}{\mathrm{MSE}_{\mathrm{ref}}}. \tag{3}$$

The mean-squared error (MSE) is calculated using the equation

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\mathrm{O\_fraction}_i - \mathrm{F\_fraction}_i)^2,$$

and $\mathrm{MSE}_{\mathrm{ref}}$ is calculated using the equation

$$\mathrm{MSE}_{\mathrm{ref}} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{O\_fraction}_i^2 - \frac{1}{N}\sum_{i=1}^{N}\mathrm{F\_fraction}_i^2,$$

where $\mathrm{O\_fraction}_i$ denotes the observation fractions and $\mathrm{F\_fraction}_i$ denotes the forecast fractions. The low-skill reference forecast is represented by $\mathrm{MSE}_{\mathrm{ref}}$, the largest possible MSE that can be obtained if no overlap exists between the forecast and observation binary fields (Wolff et al. 2014). A disadvantage of the FSS is it is only valid for values above or below a user-provided threshold, unlike the dimensioned metrics.

The FSS has a range from 0 to 1, with an FSS of zero indicating that a forecast has no skill at the square length n it was assessed while an FSS equal to one represents a perfect forecast at the assessed square length of n. When errors are present, the lowest FSS is observed with a square length $n = 1$, and FSS scores increase as the neighborhood square length increases (Roberts and Lean 2008). Forecasts contain useful information when the FSS equals or exceeds the uniform FSS (Wolff et al. 2014). The uniform FSS is defined as

$$\mathrm{FSS}_{\mathrm{uniform}} = 0.5 + \frac{\mathrm{FSS}_{\mathrm{random}}}{2}, \tag{4}$$

where $\mathrm{FSS}_{\mathrm{random}}$ is the FSS that would be obtained from a random forecast with the same fractional coverage over the observation domain. An FSS greater (less) than the $\mathrm{FSS}_{\mathrm{uniform}}$ indicates the displacement error is smaller (larger) than the square length n divided by 2 (Roberts and Lean 2008). For Fig. 2, $\mathrm{FSS}_{\mathrm{random}}$ equals 0.28 (7/25), assuming the 25 grid-square box represents the full domain. Thus, forecast accuracy can be compared using the FSS at a given scale as well as the spatial scale at which a forecast becomes useful.

c. Method for Object-Based Diagnostic Evaluation

MODE is a technique for identifying and matching objects in two different fields (Davis et al. 2006, 2009). Unlike point-based statistics, which are unable to account for spatial errors when assessing model accuracy, and the FSS, which can assess object displacement but not shape, object-based statistics allow for comparison of features characteristics even if they are spatially separated. Objects are meant to represent ''regions of interest'' (Developmental Testbed Center 2014), which for this analysis are upper-level cloud systems containing cold infrared BTs. The BT data used in the analysis are most sensitive to cold clouds. Therefore, the two MODE fields are the GOES-observed 10.7-μm BTs and the HRRRx-simulated 10.7-μm BTs. The MODE process is described in Davis et al. (2006), but a short outline as applied to cloud systems is provided here for context:

1) smooth forecast and observed BT fields using a process called convolution thresholding to identify objects;
2) calculate various object attributes, for each observed and forecast cloud object;
3) match forecast and observed cloud objects using a fuzzy logic algorithm and calculate attributes of paired objects, such as intersection area and distance; and
4) output attributes for individual objects and matched object pairs for assessment.

The MODE process is highly configurable. For the convolution thresholding process, users determine the convolution radius, which defines the radius of the circular convolution applied to smooth the raw data fields, and the convolution threshold, the value applied to the smoothed field to define discrete objects. In addition, weights for the object pair attributes, used for object matching and merging, also need to be set.

The settings used in this study are tuned to best identify individual forecast and observation objects, as well as object matches between the forecast and observation fields. These settings were chosen after testing multiple combinations of convolution radii and attribute weights (not shown) using 5-h HRRRx forecasts from 1400 and 2000 UTC 23 July. The 5-h HRRRx forecasts are used to avoid potential errors associated with model spinup. Based on the objects identified from the 5-h HRRRx forecasts, a convolution radius of five grid points (15 km) is used for both the observed and forecast fields to allow for the analysis of small-scale storms. This convolution radius is consistent with a range from two to eight grid spaces stated by Cai and Dumais (2015) as identifying convective storm objects in ~4-km resolution radar imagery. The convolution threshold for this study will be the 10th percentile of the BTs and will vary based on the valid time to account for the diurnal cycles.

TABLE 1. User-defined weights and brief descriptions of the object pair attributes used in this analysis.

Object pair attribute	Weight (%)	Description
centroid_dist	4 (25.0)	Distance between objects' ''center of mass''
boundary_dist	3 (18.75)	Min distance between the objects
convex_hull_dist	1 (6.25)	Min distance between the polygons surrounding the objects
angle_diff	1 (6.25)	Orientation angle difference
area_ratio	4 (25.0)	Ratio of the forecast and observation objects' areas (whichever yields a lower value)
int_area_ratio	3 (18.75)	Ratio of observation (forecast) object to the objects' intersection area (whichever yields a higher value)

For matched objects, MODE computes an interest value that is a single number portraying the correspondence between two objects. The interest value is a weighted calculation of the object pair attributes, with individual parameter weights assigned by the user. Interest values range from 0 to 1, with a perfect match having an interest value of 1. The user-defined attribute weights used during this analysis can be seen in Table 1. Overall, this analysis prioritizes the distance and size comparison between the objects, though other options could also be used. The minimum distance (boundary_dist) between objects has a lower user-defined weight than the centroid distance (centroid_dist) to put more emphasis on the displacement between the objects. However, MODE-assigned centroid distance weight is the user-defined weight multiplied by the ratio of the objects' areas. Therefore, the boundary distance between objects has greater weight when the ratio between the observation and forecast area is less than 0.75. The ratio of the intersection area of objects to the observation/forecast object's area (int_area_ratio) has a lower weight than the ratio of the objects' areas (area_ratio) because the int_area_ratio value can be artificially high when a small object is fully consumed by a large object. As is default in MODE, no object merging in the individual observation and forecast fields is performed.

MODE COMPOSITE SCORE

MODE output provides multiple ways to interpret and display forecast accuracy. For example, users can

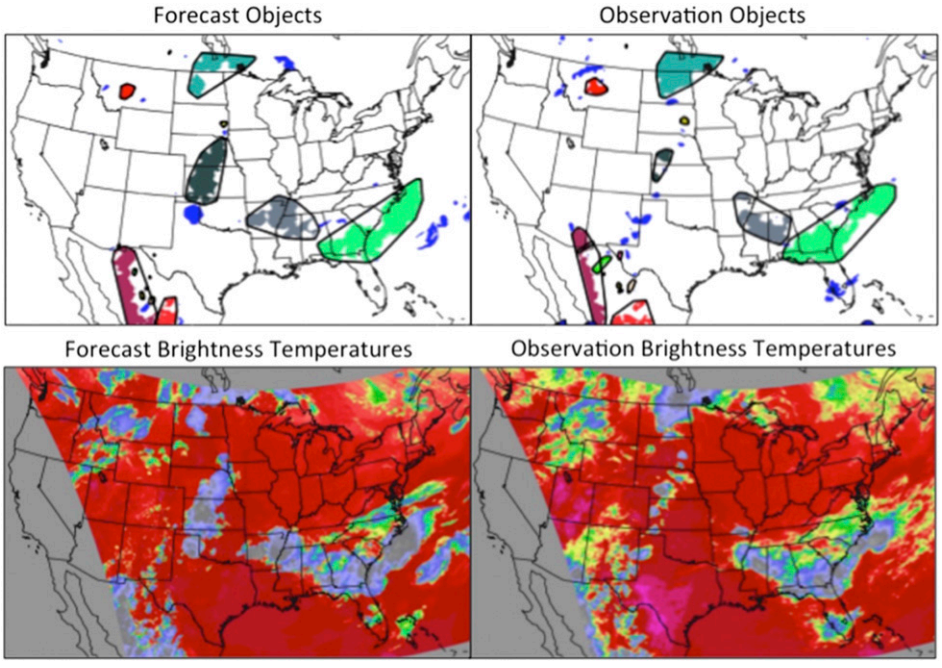

FIG. 3. Example of (top) MODE objects created from (bottom) BT imagery from (left) the HRRRx forecast at 2000 UTC 23 Jul 2015 and (right) GOES observations valid at 0100 UTC 24 Jul 2015. Black contours around the objects represent clusters defined using a total interest threshold of 0.65.

compare the number and area of objects (Wolff et al. 2014) or distance between objects (Bytheway and Kummerow 2015). For this study, however, a single number to express both object size and distance between objects is desired. Thus, we developed a new analysis metric called the MODE composite score (MCS) to summarize this information. The MCS is an area-weighted calculation using the interest values from the MODE output:

$$\mathrm{MCS} = \sum_{i=1}^{N_C} \frac{\mathrm{Area}_{\text{Observed Cluster}}(i)}{\text{Total Area}} \times \text{Interest Value}(i) + \sum_{j=1}^{N_O} \frac{\mathrm{Area}_{\text{Observed Object}}(j)}{\text{Total Area}} \times \text{Interest Value}(j), \tag{5}$$

where the total area is the area of the objects in the observation field plus the area of the objects in the forecast field that are unmatched to observation objects. We use N_c and N_O to represent the number of observation clusters and objects, respectively. Area weighting is used so that large objects are given more weight than small objects. Since the maximum centroid distance in this analysis is 200 km, any object in the observation field will be matched and, therefore, have an interest value, with a forecast object that is within 200 km. However, each observation object and corresponding forecast object can only be used once when calculating the MCS. To calculate the highest possible MCS, object matches are analyzed from the highest interest value to the lowest. Object matches with an area ratio less than 5% are not included. Like the FSS, the MCS has a range from 0 to 1. An MCS of zero indicates a forecast that has no skill while an MCS equal to one represents a perfect forecast.

The MCS calculation first examines clusters to account for multiple objects in the observation (forecast) field that may correspond to a single object in the forecast (observation) field. A cluster is defined as any set of one or more objects in one field that matches any one or more objects in the other field (Developmental Testbed Center 2014). Clusters are defined as object matches exceeding the interest value threshold, which is set to 0.65 in this analysis. Objects can be components of the same cluster if one or more of the observation (forecast) objects match the same forecast (observation) object. An example of clusters can be seen in Fig. 3. Clusters are identified by thick black lines surrounding the same-colored objects. In Nebraska and Kansas, the HRRRx forecast (Fig. 3, top left) has a much larger object area than the GOES observations (Fig. 3, top right). Since the two circled observation

objects and the forecast object are identified as a cluster (black line circling the objects), both observation objects are equated to the single forecast object in the MCS calculation. If clusters are not used when calculating the MCS, the circled observation object in Kansas would not match a forecast object, and its interest value in the MCS calculation would be zero.

4. Methodology

Calculating the FSS and MCS requires the forecast and observation data to be masked based on a given BT threshold. The FSS needs a mask to create the binary yes/no field, and MODE needs a mask for the convolution thresholding process. For this study, we employ a percentile-based BT threshold when computing these values. This is advantageous because it does not constrain the FSS and MODE to a specific time of day and year based on differences in the cloud field. The 10th percentile of the BTs will be used as the threshold for each field so that the analysis focuses on the coldest cloud tops. For the GOES-observed BTs, the 10th percentile of the BT threshold ranges from 253.5 to 238.5 K. Convective clouds account for 7.6%–9.3% of the absolute cloud amount, depending on the season (Chang and Li 2005). The observed (forecast) BT threshold used during this study is obtained using observed (simulated) BTs during the 10-day period prior to and including the valid (forecast) time. The threshold is computed for each hour of the day to account for different cloud characteristics due to the diurnal cycle. The threshold applied to the HRRRx-simulated BT from a given HRRRx forecast must have the same initialization time and forecast hour as the given HRRRx forecast to account for any potential variations between different HRRRx initialization times. The BT percentile thresholds for FSS and MODE are computed over a domain covering 25°–55°N and 110°–74°W.

When calculating the MCS, MODE will search over the full domain seen in the bottom images of Fig. 3, except for the areas containing no data (shaded gray). The full domain is used to identify the objects in order to avoid complications associated with cloud objects overlapping the boundaries of the smaller focus areas. Clusters and objects are also matched in the full domain. When calculating the MCS for the sectors seen in Fig. 1a, the sector-specific MCS is calculated using the observation objects and clusters that are over half contained within the sector's domain. Forecast objects that do not have a matching observation object in the full MODE domain and are over half contained within the sector's domain will also be included in the sector-specific MCS calculation. This would result in a lower MCS because the total area in Eq. (5) is increased with the unmatched forecast objects.

5. Results

a. *Dimensioned metrics*

1) Mean absolute error

The overall error in the simulated BTs is analyzed using the MAE. An example for the central plains sector can be seen in the "quilt" plot shown in Fig. 4, with each square showing the MAE for a specific HRRRx forecast hour and valid observation time. Forecast hours increase upward along the columns, while valid times increase to the right along a row. This means that forecast hours from an individual HRRRx forecast move upward and to the right along a diagonal line. Therefore, this display method allows for quick comparison of the forecast errors between different HRRRx forecasts valid at the same time as well as providing a comparison between HRRRx forecasts based on time of day.

Intuition would suggest that the forecast skill would decrease for longer forecast lead times due to error growth, and therefore the MAE would increase upward along a given column in Fig. 4. However, for this single day, this is not always the case. For example, for a valid time of 0000 UTC 24 July, the 12-h forecast from the HRRRx forecast initialized at 1200 UTC 23 July has a lower MAE (12.89 K) than the 5-h HRRRx forecast from the HRRRx forecast initialized at 1900 UTC (MAE equal to 18.62 K). Thus, the 12-h forecast has less overall error than the 5-h forecast. Qualitative comparison of these forecasts in Fig. 5 agrees with the MAE analysis. The 12-h forecast better represents the GOES BTs than the 5-h forecast because the cold BTs are less expansive along the line from Nebraska to Texas. Another example is the MAE for the HRRRx forecasts initialized at 1900 UTC 23 July, where the 6-h forecast valid at 0100 UTC 24 July is the less accurate than the 17-h forecast valid at 1200 UTC 24 July.

Comparing the MAE from different HRRRx initializations valid at the same time can indicate which HRRRx initialization will better represent a later GOES observation. For the example in Fig. 5, the 1-h forecasts from 23 July 2015 at 1200 and 1900 UTC have MAEs of 6.00 and 9.52 K, respectively. The 5-h forecast from 1200 UTC 23 July 2015 also has a lower MAE (9.19 K) than the 5-h forecast from 1900 UTC 23 July 2015 (18.62 K). Overall, the MAE for 6-h or earlier HRRRx forecasts is moderately correlated (subjectively defined as a correlation coefficient greater than or equal to 0.5) with the MAE for a HRRRx forecast valid 5 h in

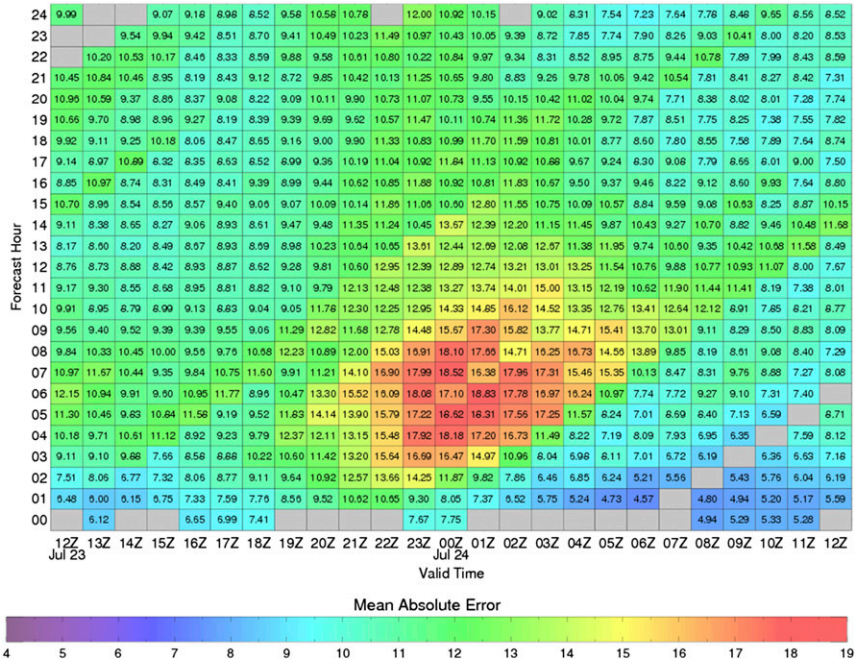

FIG. 4. Quilt plot of MAE as a function of HRRRx forecast hour and valid time for the central plains sector. Each square indicates the MAE associated with a HRRRx forecast hour (listed along the *y* axis), valid at a given time (listed along the *x* axis).

the future. Therefore, it can be assumed that, when comparing multiple HRRRx initializations valid at the same time, the initialization with the lowest MAE will better represent future GOES observations.

2) MEAN BIAS ERROR

The overall bias between the HRRRx-simulated BTs compared with the GOES-observed BTs can be evaluated using the MBE. The biases in the north plains sector can be seen in Fig. 6. The MBE demonstrates a consistent bias among all HRRRx initialization times. Specifically, the MBE for 0- and 1-h HRRRx forecasts is positive by 1 K or more, indicating a warm bias initially exists; however, as the forecast hour increases, the MBE decreases. By forecast hour 3, a negative MBE, or a cold bias, is observed in over half of the HRRRx initializations. This indicates that the first 1–2 forecast hours do not have sufficient areal extent of the cold BTs (less cloud cover than the observations). This is consistent with Bytheway and Kummerow (2015), whose work indicated the HRRRx requires an additional 1–2 h of spinup beyond the 1 h currently built into the HRRRx.

Then, later forecast hours have too much cloud cover. An example can be seen in Fig. 7 for the GOES observation at 1100 UTC 24 July 2015. The top-left image in Fig. 7 shows the 1-h forecast at 1000 UTC 24 July 2015, and the lack of cloud cover extent in Wisconsin and southeastern Nebraska is apparent. The bottom-left image in Fig. 7 presents the 4-h forecast at 0700 UTC 24 July 2015, and too much cloud cover is produced over Minnesota and Wisconsin.

In addition to the HRRRx warm bias in the early forecast hours due to model spinup, a secondary warm bias can be observed during the night in association with nocturnal convection. For forecast hours greater than nine valid between 0500 and 1200 UTC 24 July 2015 (midnight to 0700 local time), the MBE is exclusively positive. This indicates a lack of cloud cover compared to the GOES observations and is due to the HRRRx model dissipating the cloud cover too quickly during the night. An example is shown in Fig. 8. The left and center columns correspond to HRRRx forecasts initialized at 1700 and 1900 UTC, respectively, with individual panels valid at the GOES observation time in the right column. Both sets of HRRRx forecasts exhibit a cold bias at 0000 UTC 24 July 2015 (top row in Fig. 8); however, by 0500 UTC (center row), both forecasts exhibit a warm bias that continues to grow during the next 4 h. Comparison of the images shows that the warm brightness temperature bias is primarily due to insufficient coverage of thin cirrus clouds surrounding the thunderstorms and insufficient storm coverage in the southern half of the region.

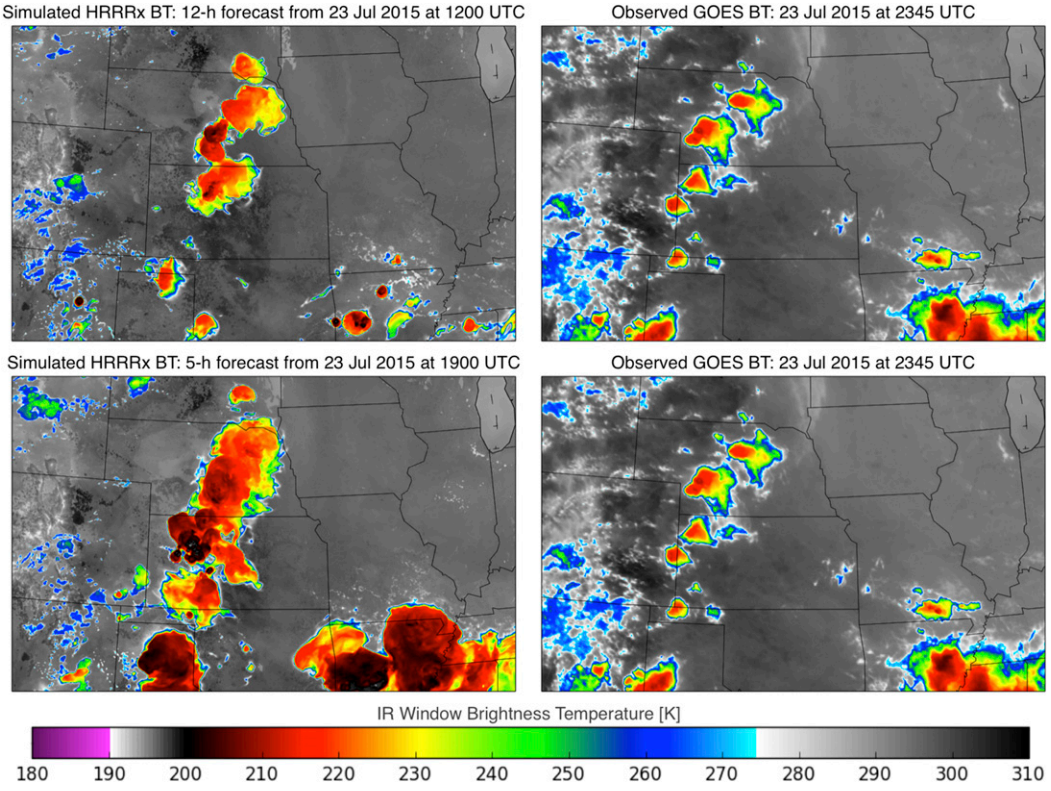

FIG. 5. HRRRx-simulated BT imagery from (top left) a 12-h forecast at 1200 UTC 23 Jul 2015 and (bottom left) a 5-h forecast at 1900 UTC 23 Jul 2015, valid at 0000 UTC 24 Jul 2015 for the central plains sector. (right) GOES observations at 2345 UTC 23 July 2015.

b. HRRRx-simulated brightness temperature cold bias

A different measure of the simulated cold BT bias can also be observed in Fig. 7. Overall, the simulated BTs for deep convective clouds are colder than the observations. This is even apparent for the cloud features in the 1-h forecast (Fig. 7, top) despite the overall warm bias, as indicated by the MBE of 3.06 K, when assessed over the entire region (Fig. 6). For example, for the cloud system near the Nebraska–Kansas border, the 10th percentile of simulated BTs is approximately 207.5 K while the 10th percentile of the observed BTs is about 211.3 K. The simulated BTs along the Minnesota–Wisconsin border are also colder than the observations, with the 10th percentile of the simulated (observed) BTs approximately 208.9 K (214.1 K).

Cumulative distribution functions (CDFs) provide an alternative way of identifying the simulated cold BT bias. Figure 9 displays the HRRRx-simulated BT CDF in red for 1- and 5-h forecasts valid between 1200 UTC 23 July and 1200 UTC 24 July 2015, with the corresponding GOES BT CDF plotted in blue. As seen in Fig. 9a, the simulated CDFs for the 1-h forecasts are larger for colder temperatures (BT $<$ 230 K) compared with the GOES CDFs. The MBE is positive for 1-h forecasts because the HRRRx CDF is smaller than the GOES CDF for temperatures warmer than 260 K, indicating more HRRRx pixels have a BT greater than 260 K (fewer cloudy pixels) compared with GOES. This again demonstrates the HRRRx produces convective cloud tops that are too cold and the areal extent is too small in the 1-h forecast. For the 5-h forecasts, seen in Fig. 9b, the simulated CDFs for BTs less than 270 K are larger compared with the GOES CDFs. This presumably larger extent of cloud pixels results in an MBE that is negative.

This cold bias needs to be accounted for when applying a masking BT threshold for FSS and MODE. Therefore, separate thresholds are defined for the GOES and HRRRx data to account for the bias. Each threshold is the 10th percentile of BTs over the prior 10-day period described previously. The values of this BT threshold based on the HRRRx forecast hour can be observed in the box-and-whisker plot in Fig. 10. The colored boxes extend the range of the middle 50% of the 10th percentile BTs, and the black line represents the median FSS. Each HRRRx box for a given forecast hour indicates the 10th percentile BTs from HRRRx initializations whose forecast for that given forecast hour

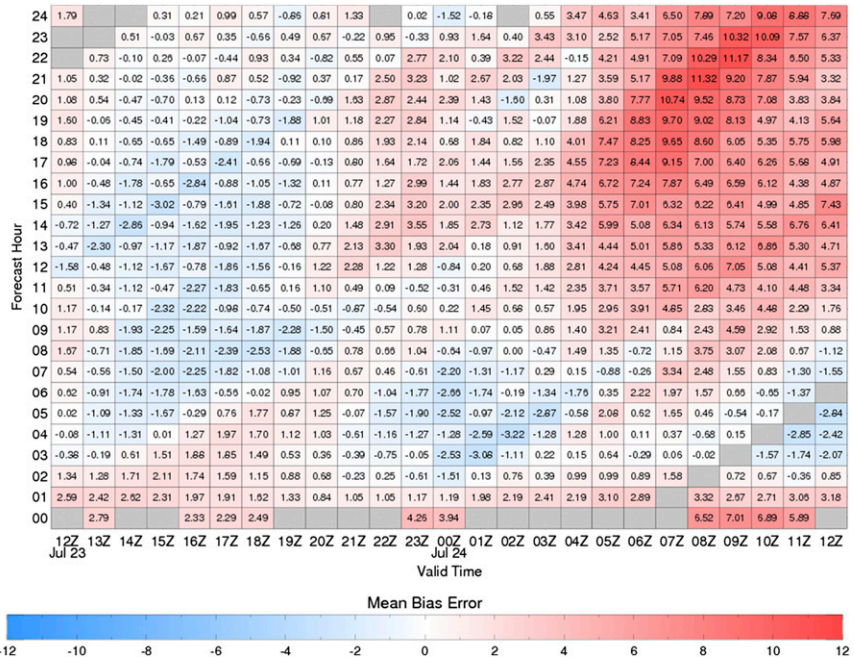

FIG. 6. As in Fig. 4, but for MBE for the north plains sector. Blue (red) represent HRRRx-simulated BT imagery that is overall colder (warmer) than the corresponding GOES BT imagery.

is valid between 1200 UTC 23 July and 1200 UTC 24 July. Each GOES box represents the 10th percentile BTs for GOES observations corresponding to the HRRRx forecasts. The small difference in GOES BT for 0 and 24 forecast hours is due to fewer 0- and 24-h HRRRx forecasts being valid in this time frame. Figure 10 indicates that the area of HRRRx cold, convective clouds at forecast hours 0–1 is too small and then too large at forecast hour 3 and greater. While the HRRRx BT threshold for FH 0 is warmer than most BTs associated with deep convection (Konduru et al. 2013), it will still be used in this analysis as it is provided and any potential errors should be assessed.

c. *Neighborhood-based statistic*

The neighborhood-based statistic used in this study is the fractions skill score. The FSS can be used in two separate ways to assess forecast accuracy. The first method compares the FSS values for a given square length or lengths of n grid points surrounding a grid point. FSS values for varying square lengths can be seen in the box-and-whisker plot in Fig. 11. Each box and whisker represents 304 HRRRx forecasts with 0–12-h lead times that are valid between 1200 UTC July 23 and 1200 UTC July 24. As expected, the FSS increases with increasing square length (Roberts and Lean 2008), since lower skill is associated with finer-resolution NWP (Ebert 2009). For each square length, the median FSS is lower for the central plains sector than the other two sectors, presumably because the localized cellular cloud features are more difficult to accurately forecast in location.

FSS can also be used to identify the most skillful forecast for features of a given scale L. This is accomplished by computing the FSS at a square length n, where n equals L times 2 divided by the horizontal grid spacing. To assess the HRRRx forecast accuracy for features with L equal to 100 km, FSS would be calculated at a square length n of 66 pixels (~200 km) for the 3-km horizontal grid spacing HRRRx. An FSS of 0.5 at a square length n of 200 km indicates the 100-km forecast feature is displaced by 100 km, with smaller (larger) displacements having a higher (lower) FSS (Mittermaier and Roberts 2010). Figure 12 displays the central plain sector FSS at a square length of 66 grid points, double the approximate size of cloud cover associated with single cellular convection (Heymsfield and Blackmer 1988), for forecast hours 0–12. The dashed line represents the uniform FSS described in section 3b and notches in the box display a confidence interval around the median. If notches from two boxes do not overlap, there is 95% confidence the medians differ (Chambers et al. 1983). The 1-h forecast has the highest FSSs, with a median FSS of 0.932, and thus exhibits the smallest displacement in features. Since the notches for the 1-h forecast do not overlap with the

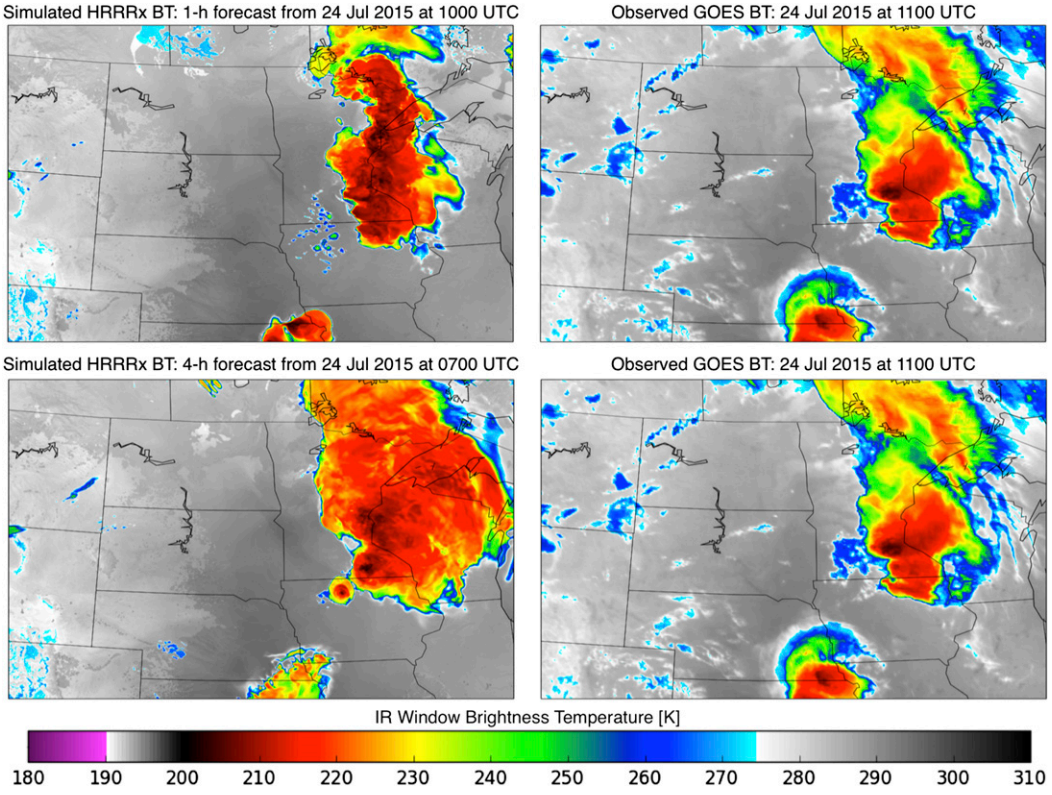

FIG. 7. As in Fig. 5, but for (top left) a 1-h forecast at 1000 UTC 24 Jul 2015 and (bottom left) a 4-h forecast at 0700 UTC 24 Jul 2015 for the north plains sector.

0-h forecast, it can be stated with 95% confidence that the 1-h forecast is more skillful at identifying the location of the single cellular cloud features than the 0-h forecast. Forecast skill then decreases from forecast hour 1 and over half the forecasts are no longer skillful by forecast hour 6, indicating the displacement error is greater than the scale of the given features (100 km). Because of the larger size of the cloud features in the north plains and southeast sector (see Fig. 1b), FSS is calculated at a square length n of 100 for an L equal to 150 km. The 1-h forecast has the highest FSSs, however not at the 95% confidence level for the north plains sector, and the median forecast is still skillful even at a 12-h forecast (not shown).

Another method for assessing HRRRx forecast accuracy is by finding the spatial scale for which an HRRRx forecast is deemed to be skillful ($\text{scale}_{\text{min}}$). This occurs when the FSS at a square length n equals or exceeds the uniform FSS described in section 3b. A smaller value of $\text{scale}_{\text{min}}$ indicates a forecast that is more skillful, as the displacement errors are smaller (Roberts and Lean 2008). The average $\text{scale}_{\text{min}}$ for each sector for FHs 0–12 can be seen in Fig. 13, with the error bars representing the 95% confidence interval around the mean. The value of $\text{scale}_{\text{min}}$ is smallest for early forecast hours, indicating smaller displacement errors, and then increases for later forecast hours for all sectors except the northern plains. The northern plains $\text{scale}_{\text{min}}$ increases before slightly decreasing and remaining constant after FH 5. We believe this is due to the nature of the convection in the northern plains sector. In this sector, the convection is large (Fig. 1b), and thus displacement errors in the HRRRx cloud features still result in overlapping with the GOES cloud feature. This convection is also more strongly forced, thereby making it more predictable. It is important to note that $\text{scale}_{\text{min}}$ will not identify the most accurate forecast if multiple forecasts have the same $\text{scale}_{\text{min}}$. Note that $\text{scale}_{\text{min}}$ only indicates the FSS for these forecasts are greater than the uniform FSS, not which forecast has the highest FSS at the square length of $\text{scale}_{\text{min}}$ grid points. In these instances, FSS should be calculated at a square length n that is equal to $\text{scale}_{\text{min}}$ to identify the most accurate forecast.

d. Method for Object-Based Diagnostic Evaluation

HRRRx forecast accuracy is assessed in this section using MODE and the MCS from Eq. (5). A quilt plot of

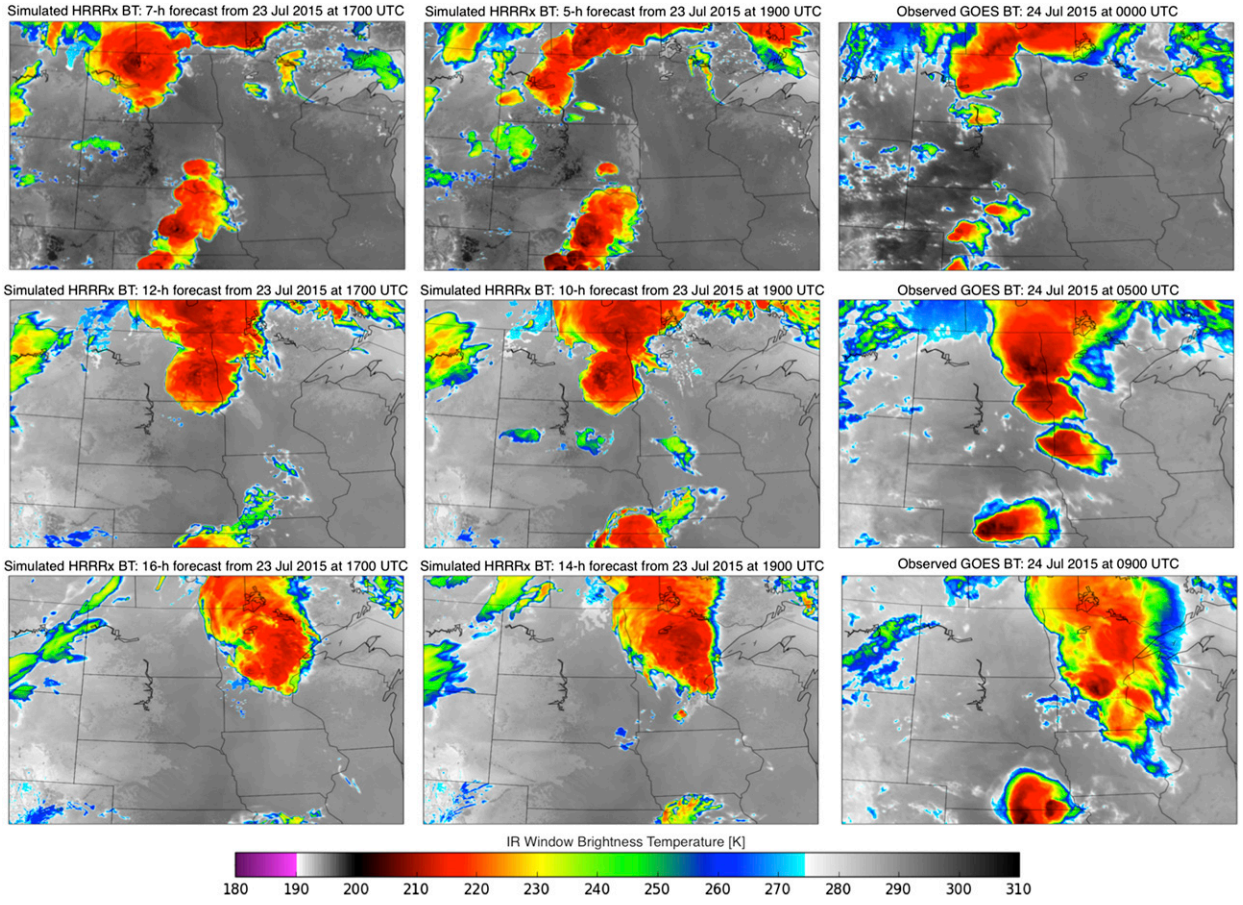

FIG. 8. HRRRx-simulated BT imagery at (left) 1700 UTC 23 Jul 2015 and (center) 1900 UTC 23 Jul 2015 for (right) valid GOES observations. Valid times are at (top) 0000 UTC 24 Jul 2015, (middle) 0500 UTC 24 Jul 2015, and (bottom) 0900 UTC 24 Jul 2015.

MCS values from the southeast sector is shown in Fig. 14. Figure 14 shows there is no distinct pattern evident in the MCS values. Unlike the MAE and $\text{scale}_{\text{min}}$, which usually increase with increasing forecast hour (thus indicating decreasing forecast skill, particularly at smaller scales), the MCS varies as the forecast hour increases. This is especially noticeable for HRRRx forecasts valid at 0800 UTC 24 July 2015, where the MCS equals 0.67 for a 6-h forecast, 0.39 for a 7-h forecast, and 0.56 for an 8-h forecast.

This variation of the MCS is considered a strength compared to dimensioned and neighborhood statistics like the MAE and $\text{scale}_{\text{min}}$. Unlike the dimensioned and neighborhood statistics, which mainly just identify object displacement, the MCS can account for both object displacement between forecast and observations as well as consistency between the object shapes and sizes. The MCS is calculated using the interest values from forecast and observation objects that are matched. As may be observed in the histogram plot in Fig. 15, these interest values are correlated to both the displacement and size comparison between the forecast and observation objects. Each colored bar represents the number of occasions the correlation value between the interest value and the attribute is observed. The interest values are negatively correlated with the distance attributes, indicating that larger displacements between forecast and observation objects result in smaller interest values for matched objects. The area attributes are positively correlated with the interest value, indicating forecast objects that have the same size or overlap observation objects produce larger interest values. Therefore, as displacement between objects increases with increasing forecast hour, as observed with $\text{scale}_{\text{min}}$, the MCS can still be moderately large if the forecast and observation objects are similar in size. In these cases, HRRRx had appropriately developed the extent and timing of the deep convection, but it is displaced from the observation.

Given that the MODE configuration in this study results in interest values correlated with both object distance and area, two HRRRx forecasts can have a similar

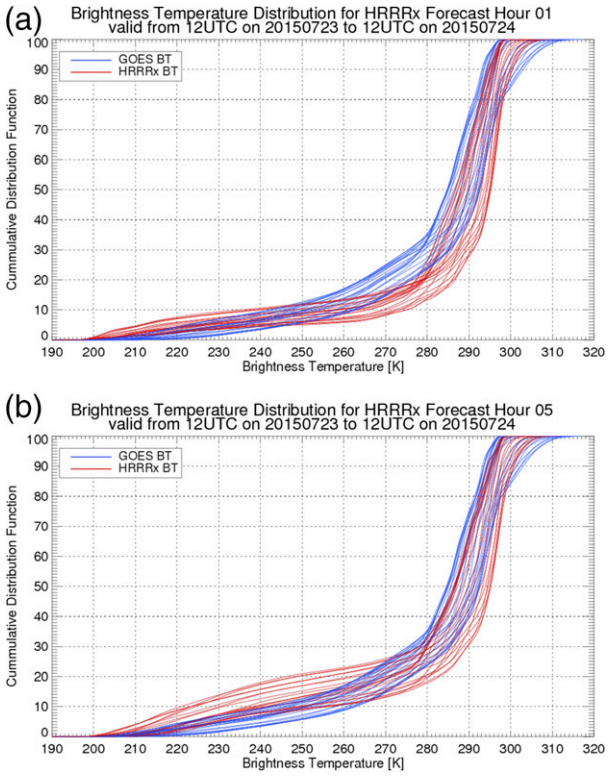

FIG. 9. (a) BT CDFs for 1-h HRRRx forecasts (red) and corresponding GOES observations (blue) valid between 1200 UTC 23 Jul 2015 and 1200 UTC 24 Jul 2015. (b) As in (a), but for 5-h HRRRx forecasts.

MCS for different reasons. An example can be observed in Fig. 16. Valid at 2300 UTC 23 July 2015, the left image is an 5-h forecast initialized at 1800 UTC 23 July 2015 and has an MCS of 0.79 while the right image is an 24-h forecast initialized at 2300 UTC 22 July 2015 with an MCS equal to 0.82. However, the forecast objects (red) from the two HRRRx initializations differ compared with the GOES observation objects (blue). The sizes of the 5-h forecast objects appear more similar to the observed objects; however, almost every object is displaced spatially. More overlapping objects are observed with the 24-h HRRRx forecast; however, the forecast and observation object sizes are less similar. However, this is not a weakness of the MCS; it is a result of the emphasis placed on both the area and distance between objects in the MODE configuration. As will be shown in the discussion, combining the MCS with metrics like the FSS can help indicate if forecast error is the result of the distances between objects.

Another noticeable occurrence in Fig. 14 is the reduction in the MCS between valid times at 0400 and 0600 UTC 24 July for each model initialization time (e.g., column). This abrupt shift is primarily a result of the size of the GOES observation objects. The MCS is an area-weighted calculation and, therefore, can be dominated by large objects. The GOES observation objects, colored by interest values, for 0400 and 0600 UTC 24 July 2015 can be seen in Fig. 17. The 0400 UTC GOES observation (left) has 41 748 object pixels, with one object with a near-perfect interest value accounting for 34 434 pixels. The 0600 UTC GOES observation (Fig. 17, right) only has 2966 object pixels, since over half of the large object seen at 0400 UTC is now outside the sector domain. Therefore, caution should be used when employing the MCS to compare forecast hours from a single model initialization that are valid at different times, especially for sectors smaller than the full MODE region.

6. Discussion

An example of using the four forecast accuracy metrics jointly is shown in Fig. 18. Figure 18 displays the value of each forecast metric for different forecast hours and initialization times valid at 0100 UTC 24 July 2015 for the central plains sector. The forecast hours range from 1 to 11.

According to the MAE, the smallest errors occur in the 1-h forecast, with the forecast accuracy decreasing thereafter. However, it is unclear, based on the MAE alone, whether the diminished accuracy is due to cloud BT errors or the displacement of the coldest BTs between the forecast and observation. This can be observed by inspecting the MBE and $\text{scale}_{\text{min}}$ metrics. A small (large) MBE magnitude represents small (large) bias in the BTs, while a small (large) $\text{scale}_{\text{min}}$ indicates a small (large) displacement in the coldest BTs. The 2-h forecast has a higher-magnitude MBE compared with the 1-h forecast. This indicates the larger MAE is probably due to the colder HRRRx-simulated BTs in the 2-h forecast compared with the observations, since the MBE for FH 2 is negative. This is further corroborated by $\text{scale}_{\text{min}}$, which remains unchanged between FH 1 and FH 2 and therefore both the 1- and 2-h forecasts are skillful at the same scale. In addition, the MCS increases slightly between FH 1 and FH 2. Since the MCS accounts for the BT bias when defining objects, a slightly higher MCS indicates the 2-h HRRRx forecast is more accurate once the BT bias is also considered.

Overall, combining MBE, $\text{scale}_{\text{min}}$, and MCS can explain differences in the overall model error indicated by the MAE between different forecast hours. For example, the MAE for FH 7 is smaller than for FH 8, indicating the 7-h HRRRx forecast is more accurate when assessed using traditional dimensioned

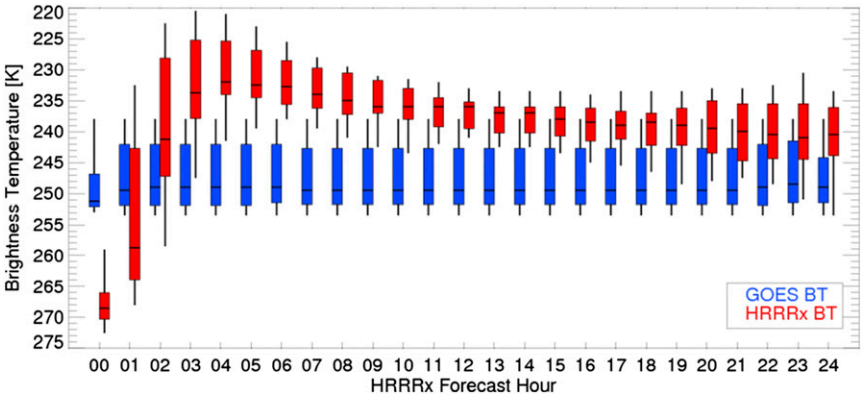

FIG. 10. Tenth percentile BT thresholds for GOES and HRRRx based on HRRRx forecast hour. Each HRRRx box for a given forecast hour indicates the 10th percentile BTs from HRRRx initializations whose forecast for that given forecast hour is valid between 1200 UTC 23 Jul and 1200 UTC 24 Jul. Each GOES box represents the 10th percentile BTs for GOES observations corresponding to the HRRRx forecasts. The small difference in GOES BTs for 0 and 24 forecast hours is due to fewer 0- and 24-h HRRRx forecasts valid in this time frame.

verification metrics. At FH 7, the MBE is greater in magnitude and the $\text{scale}_{\text{min}}$ is lower compared with FH 8. Therefore, the differences in MBE and $\text{scale}_{\text{min}}$ indicate that the lower MAE for the 7-h HRRRx forecast is the result of a smaller displacement in the cloud features. The consistent MCS but larger $\text{scale}_{\text{min}}$ between FH 7 and FH 8 indicates the forecast objects from FH 8 better represent the observation objects; however, greater displacement is experienced. An example of the MAE becoming smaller with increasing forecast hours occurs between FHs 10 and 11. While this is partially due to smaller BT bias and displacement errors, the MCS also indicates there is a large difference in the MODE objects between these two forecasts.

The MCS increases from 0.18 at FH 10 to 0.61 at FH 11 because a forecast object of 10 384 pixels is unmatched in the 10-h forecast. This feature over Oklahoma and Arkansas in Fig. 19 (left) does not appear in the 11-h forecast (right), and thus the total area for Eq. (5) is lower and the MCS is higher. In general, a large increase in MCS can be indicative of two things. If a large increase in MCS occurs between two forecast hours with the same valid time and therefore different HRRRx initializations, unmatched forecast objects that appear in the forecast with the lower MCS are not present in the forecast with the higher MCSs. If a large increase in MCS occurs between two forecast hours from the same HRRRx initialization and thus different valid times, then large observation objects are

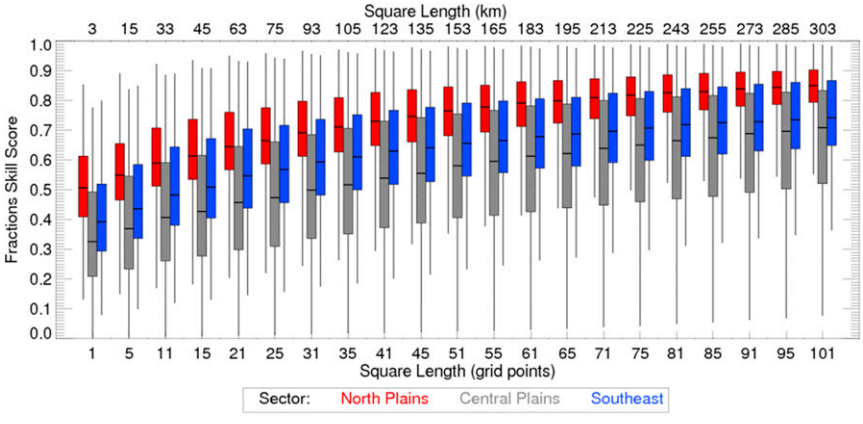

FIG. 11. Box-and-whisker plot of FSS based on square lengths surrounding each pixel. Each box and whisker represents 304 HRRRx forecasts with 0- to 12-h lead times valid between 1200 UTC 23 Jul and 1200 UTC 24 Jul.

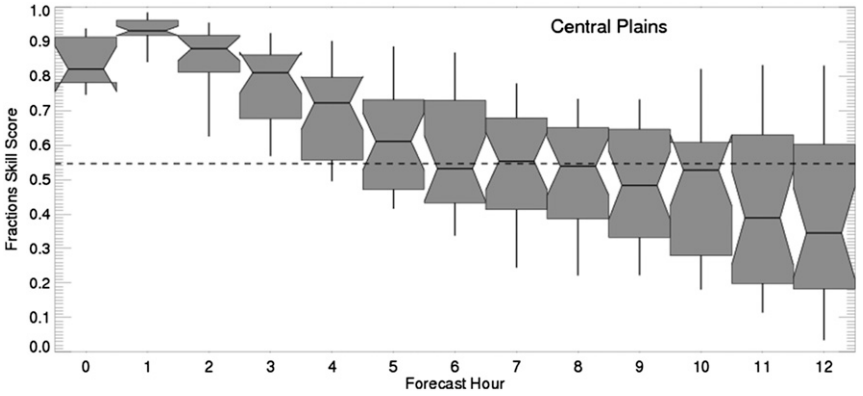

FIG. 12. Box-and-whisker plot of FSS for a 33-pixel square length surrounding each pixel for the central plains sector. The dashed line represents a skillful forecast. Notches in the box display a confidence interval around the median. If the notches from two boxes do not overlap, there is 95% confidence the medians differ (Chambers et al. 1983).

being matched with forecast objects at the forecast hour with larger MCSs. Large decreases in MCS between forecast hours indicate large forecast or observation objects that are no longer being matched at the forecast hour with smaller MCSs.

7. Conclusions

In this study, dimensioned, neighborhood, and object-based statistical metrics are used to assess the accuracy of experimental HRRRx cloud cover forecasts. This is accomplished through comparison of observed and simulated GOES 10.7-μm BTs for a 1-day time period (23–24 July 2015) containing different modes of convection across parts of the United States. Results are used to investigate how each statistic conveys information about the model forecast accuracy. Overall, during 23–24 July 2015, dimensioned statistics indicate a warm bias existed in the HRRRx-simulated BTs for forecast hours 0 and 1. This bias then shifts to a cold bias for the next few forecast hours. Therefore, the HRRRx initially does not have enough cloud cover before rapidly having too many upper-level, cold clouds in this case study. This forecast bias behavior is consistent with current limitations in the HRRRx data assimilation. The HRRRx is not yet fully cycling forecasts in successive initializations and restarts with a 13-km analysis each hour, which, when combined with the current 3-km radar reflectivity data assimilation, results in a spinup and transition from slightly under- to overforecasting cloud mass (convection) in the first few hours. Additionally, the HRRRx does not currently assimilate any clear or all-sky satellite radiances. Instead, the HRRRx assimilates retrieved cloud-top pressure estimates, but is currently limited in their application to the building (addition) of clouds only at low levels below 1200 m

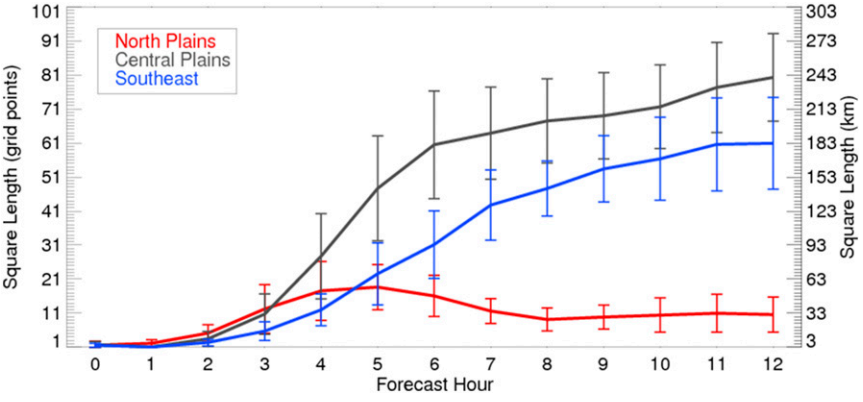

FIG. 13. Square length a forecast contains useful information based on forecast hour for the north plains (red), central plains (gray), and southeast (blue) sectors. Error bars represent the 95% confidence interval around the mean.

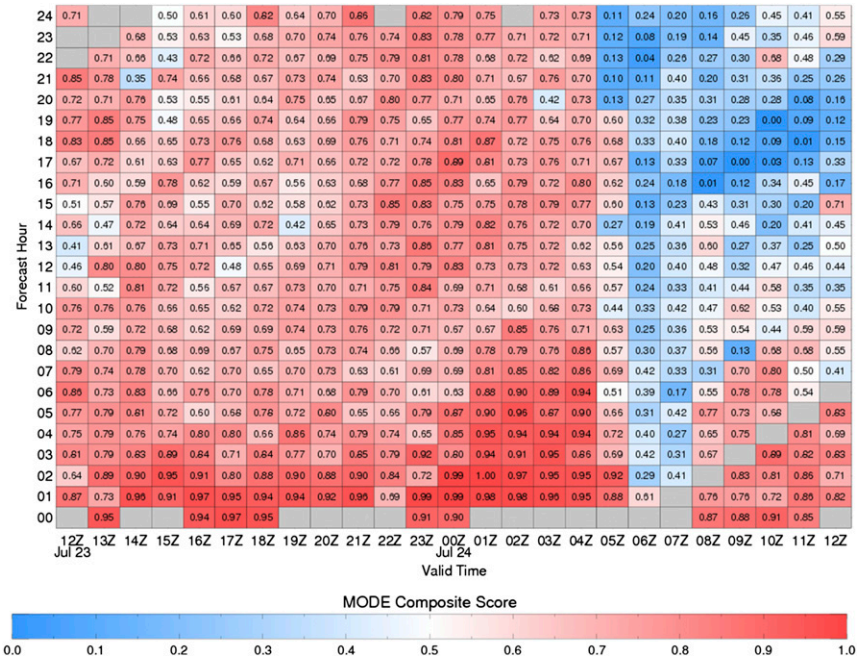

FIG. 14. As in Fig. 4, but for MCSs for the southeast sector.

AGL in an attempt to avoid an overall moist (relative humidity and precipitation) bias during the model forecasts. Clearing (removal) of clouds is applied at all levels in the model (Benjamin et al. 2016). Finally, a HRRRx model bias exists, most notably at longer forecast lengths, which results in forecasts with too much upper-level relative humidity and insufficient conversion to upper-level cloud mass. Neighborhood statistics indicate the displacement error between HRRRx-simulated BTs and GOES-observed BTs is the smallest for early forecast hours. Displacement between the forecast and observations increases with increasing forecast hour, and smaller-scale cloud features are more difficult to accurately forecast.

The MAE is a useful method for assessing overall model accuracy. However, it does not identify if existing model error is the result of displacement between forecast and observation features or a BT bias

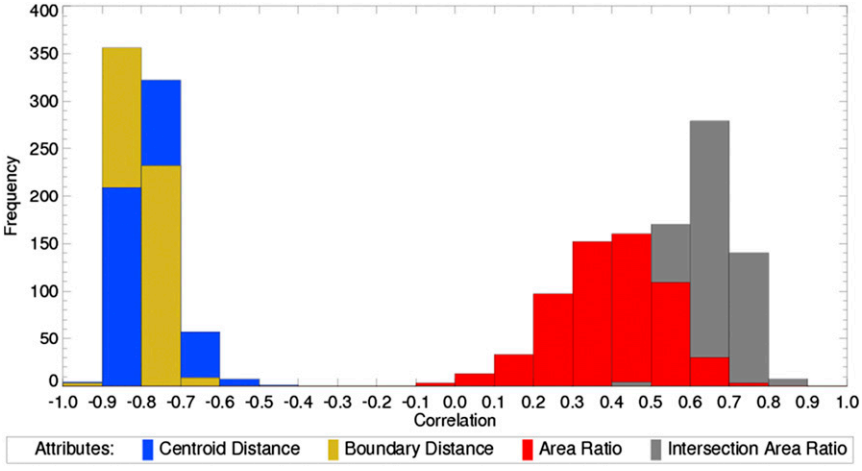

FIG. 15. Histogram of the correlation between attributes and interest values. Each colored bar represents the number of occasions the correlation value between the interest value and the attribute is observed.

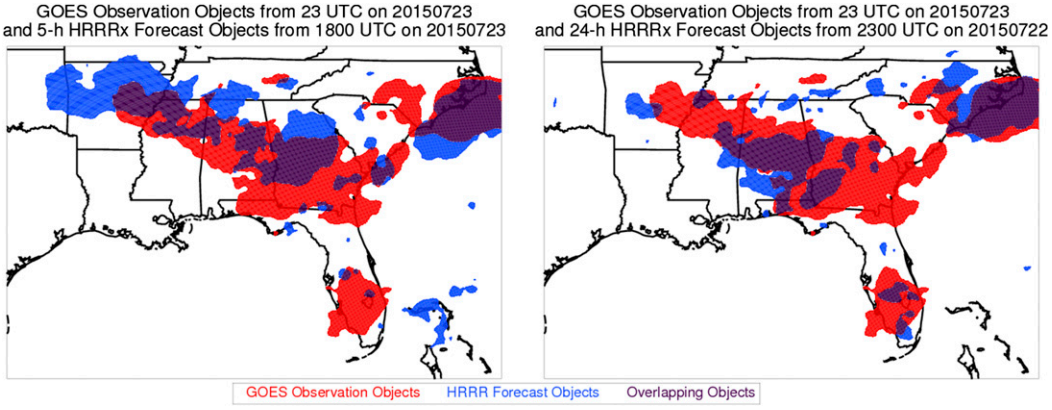

FIG. 16. Comparison of MODE HRRRx forecast objects and GOES observation objects for (left) a 5-h forecast initialized at 1800 UTC 23 Jul 2015 and (right) a 24-h forecast initialized at 2300 UTC 22 Jul 2015 valid at 2300 UTC 23 Jul 2015.

between the forecast and observation fields. The MBE can be used to assess the overall model bias by indicating how the forecast field represents the observation field.

A neighborhood-based statistic, the FSS, is useful for identifying displacement errors between forecast and observation features. By using the 10th percentile of HRRRx-simulated BTs and GOES-observed BTs for thresholding, respectively, when calculating the FSS, the FSS can account for the bias in the HRRRx-simulated BTs. The FSS can be utilized in two ways. First, the displacement of a feature of a given scale L can be evaluated by calculating the FSS using a square with a length of n grid points, where n is L divided by the grid resolution. However, this method requires knowledge of feature sizes or a scale where displacement is acceptable. Second, the FSS can indicate the scale for which a forecast contains useful information ($\text{scale}_{\text{min}}$). However, this method does not indicate the best forecast if two forecasts have the same value for $\text{scale}_{\text{min}}$.

A new object-based statistic called the MODE composite score (MCS) was introduced in this manuscript. The MCS is calculated using interest value output from the Method for Object-Based Diagnostic Evaluation (MODE). In this study, the weight of the area ratio between objects is given will have a similar but slightly higher weight than the displacement between objects when calculating the MODE interest values. Therefore, the MCS accounts for object size while still assessing the object displacement in its calculation, making it overall a better metric for accessing forecast accuracy. However, it is the most difficult metric to implement, as it requires MODE to calculate its results.

While each individual metric is useful for determining accuracy, comparing statistical metrics can best assess

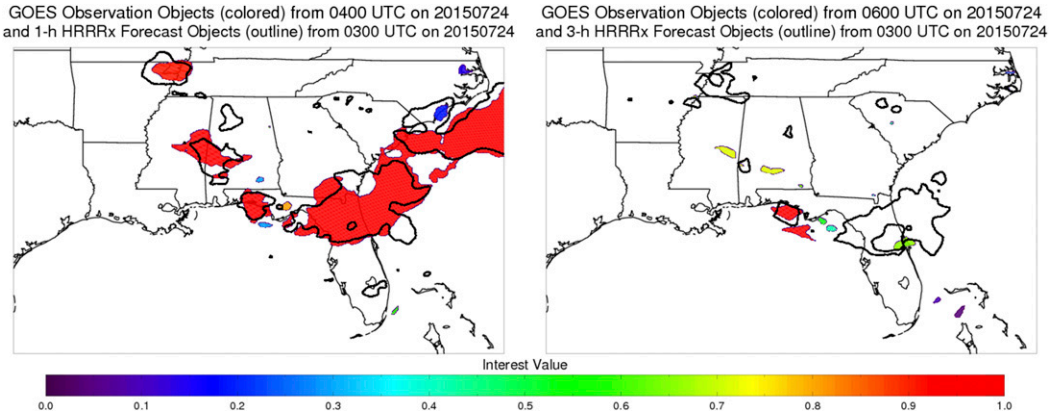

FIG. 17. (left) GOES observation objects (colored) from 0400 UTC 24 Jul 2015 and 1-h HRRRx forecast objects (outlined) from 0300 UTC 24 Jul 2015. (right) GOES observation objects (colored) from 0600 UTC 24 Jul 2015 and 3-h HRRRx forecast objects (outlined) from 0300 UTC 24 Jul 2015.

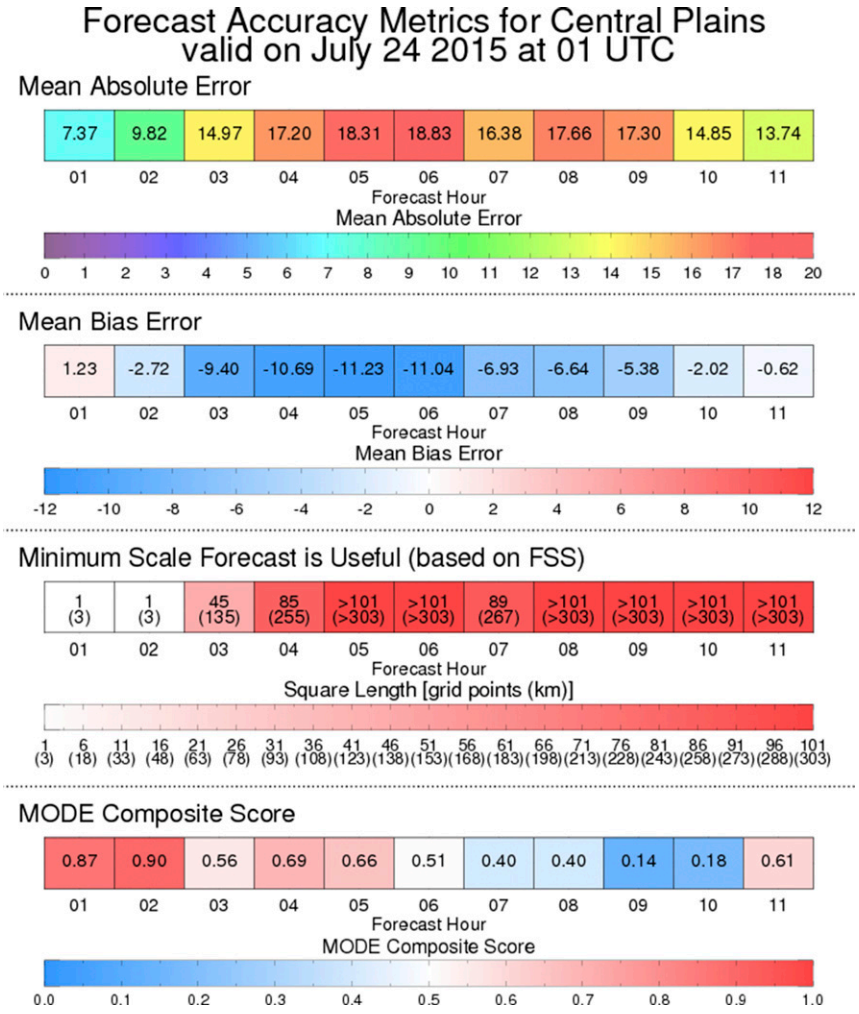

FIG. 18. Central plains MAE, MBE, minimum scale forecast is considered useful based on FSS, and MCS for forecast hours 1–12 valid at 0100 UTC 24 Jul 2015.

forecast accuracy. By comparing MAE, MBE, MCS, and $\text{scale}_{\text{min}}$, it can be inferred whether a change in forecast accuracy is due to changes in the bias between the forecast and observation fields or a change in the displacement between the forecast and observation objects. An unchanged $\text{scale}_{\text{min}}$ and an increasing (decreasing) MCS indicate a forecast with objects that have a better (worse) representation of the observation objects.

This study serves as a basis for identifying how model accuracy can be quantified using different statistical metrics. These metrics can be used by operational forecasters in real time to determine which HRRR model run may be the most accurate at depicting the current cloud features and to improve short-term weather forecasting. Real-time satellite-based verification metrics can be found online (http://cimss.ssec.wisc.edu/hrrrval/). Future work includes expanding the verification system to also support the operational version of the HRRR model and then using these statistical metrics to assess the accuracy of both models over a long time period to determine if any systematic errors exist in the model forecasts. The accuracy of the HRRR model will be investigated during the summer and winter seasons to provide a comparison between different weather regimes. In addition, MODE will be used more extensively to investigate object properties using the attribute values identified in this study. Finally, work is also under way to diagnose the cause of the cold bias that is often seen in the simulated brightness temperatures in regions containing ice clouds in the upper troposphere. Some preliminary tests using other forward radiative transfer models suggest that this bias is primarily due to the cloud property lookup tables that are being used by the CRTM. This issue will be addressed more thoroughly in future studies.

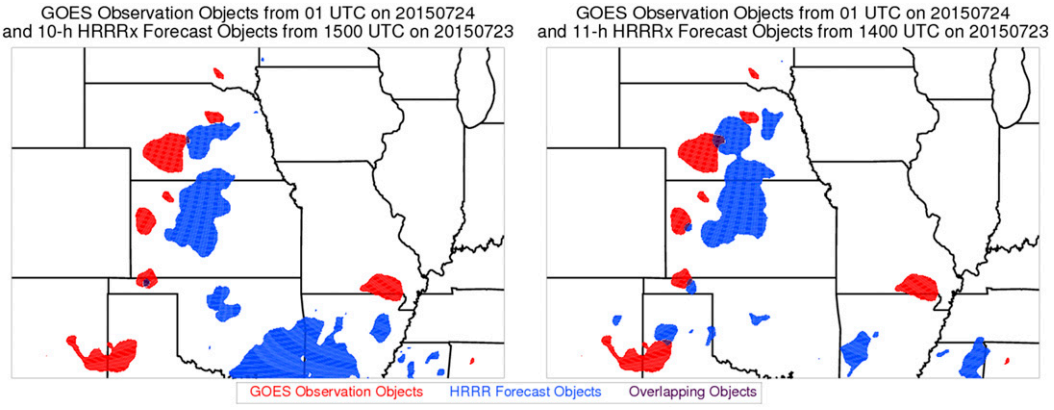

FIG. 19. As in Fig. 16, but for (left) a 10-h forecast initialized at 1500 UTC 23 Jul 2015 and (right) an 11-h forecast initialized at 1400 UTC 23 Jul 2015 valid at 0100 UTC 24 Jul 2015.

Acknowledgments. The authors thank John Halley Gotway at NCAR/RAL for his help installing and troubleshooting MODE and Randy Bullock at NCAR/RAL for his explanation of the MODE interest values. We also thank Jamie Wolff, Tara Jensen, Michelle Harrold, and Marybeth Zarlingo for their assistance during our trips to the Developmental Testbed Center. The authors thank Louis Grasso and two anonymous reviewers for their contributions to this manuscript. This project was funded by the GOES-R Risk Reduction Program via NOAA Cooperative Agreement NA15NES4320001. Support for this project was provided by the Developmental Testbed Center (DTC). The DTC Visitor Program is funded by the National Oceanic and Atmospheric Administration, the National Center for Atmospheric Research, and the National Science Foundation.

REFERENCES

Benjamin, S., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, doi:10.1175/MWR-D-15-0242.1.

Brown, D. P., and A. C. Comrie, 2002: Spatial modeling of winter temperature and precipitation in Arizona and New Mexico, USA. *Climate Res.*, **22**, 115–128, doi:10.3354/cr022115.

Bytheway, J. L., and C. D. Kummerow, 2015: Toward an object-based assessment of high-resolution forecasts of long-lived convective precipitation in the central U.S. *J. Adv. Model. Earth Syst.*, **7**, 1248–1264, doi:10.1002/2015MS000497.

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

Chai, T., and R. R. Draxler, 2014: Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.*, **7**, 1247–1250, doi:10.5194/gmd-7-1247-2014.

Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey, 1983: *Graphical Methods for Data Analysis*. Wadsworth International Group, 395 pp.

Chang, F.-L., and Z. Li, 2005: A near-global climatology of single-layer and overlapped clouds and their optical properties retrieved from *Terra*/MODIS data using a new algorithm. *J. Climate*, **18**, 4752–4771, doi:10.1175/JCLI3553.1.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, and A. K. Heidinger, 2013: Evolution of severe and nonsevere convection inferred from GOES-derived cloud properties. *J. Appl. Meteor. Climatol.*, **52**, 2009–2023, doi:10.1175/JAMC-D-12-0330.1.

Cintineo, R., J. A. Otkin, M. Xue, and F. Kong, 2014: Evaluating the performance of planetary boundary layer and cloud microphysical parameterization schemes in convection-permitting ensemble forecasts using synthetic *GOES-13* satellite observations. *Mon. Wea. Rev.*, **142**, 163–182, doi:10.1175/MWR-D-13-00143.1.

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

Dai, A., K. E. Trenberth, and T. R. Karl, 1999: Effects of clouds, soil moisture, precipitation, and water vapor on diurnal temperature range. *J. Climate*, **12**, 2451–2473, doi:10.1175/1520-0442(1999)012<2451:EOCSMP>2.0.CO;2.

Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772–1784, doi:10.1175/MWR3145.1.

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

DelSole, T., and M. K. Tippett, 2014: Comparing forecast skill. *Mon. Wea. Rev.*, **142**, 4658–4678, doi:10.1175/MWR-D-14-00045.1.

Developmental Testbed Center, 2014: Model Evaluation Tools version 5.0 (METv5.0) user's guide 5.0. DTC Tech. Rep., 241 pp. [Available online at http://www.dtcenter.org//met/users/docs/users_guide/MET_Users_Guide_v5.0.pdf.]

Done, J., C. A. Davis, and M. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the Weather

Research and Forecasting (WRF) Model. *Atmos. Sci. Lett.*, **5**, 110–117, doi:10.1002/asl.72.

Earth System Research Laboratory, 2016: Rapid Refresh. [Available online at http://rapidrefresh.noaa.gov/.]

Ebert, E. E., 2009: Neighborhood verification: A strategy for rewarding close forecasts. *Wea. Forecasting*, **24**, 1498–1510, doi:10.1175/2009WAF2222251.1.

Hamilton, D. W., and F. H. Proctor, 2002: Convectively induced turbulence encountered during NASA'S fall-2000 flight experiments. Preprints, *10th Conf. on Aviation, Range, and Aerospace Meteorology*, Portland, OR, Amer. Meteor. Soc., 10.8. [Available online at https://ams.confex.com/ams/pdfpapers/40038.pdf.]

Han, Y., P. van Delst, Q. Liu, F. Weng, B. Yan, R. Treadon, and J. Derber, 2006: JCSDA Community Radiative Transfer Model (CRTM)—version 1. NOAA Tech. Rep. NESDIS 122, 40 pp. [Available online at http://docs.lib.noaa.gov/noaa_documents/NESDIS/TR_NESDIS/TR_NESDIS_122.pdf.]

Heymsfield, G. M., and R. H. Blackmer Jr., 1988: Satellite-observed characteristics of Midwest severe thunderstorm anvils. *Mon. Wea. Rev.*, **116**, 2200–2224, doi:10.1175/1520-0493(1988)116<2200:SOCOMS>2.0.CO;2.

Kaplan, M. L., A. W. Huffman, K. M. Lux, J. J. Charney, A. J. Riordan, and Y.-L. Lin, 2005: Characterizing the severe turbulence environments associated with commercial aviation accidents. Part 1: A 44-case study synoptic observational analyses. *Meteor. Atmos. Phys.*, **88**, 129–153, doi:10.1007/s00703-004-0080-0.

Karl, T. R., and Coauthors, 1993: A new perspective on recent global warming: Asymmetric trends of daily maximum and minimum temperature. *Bull. Amer. Meteor. Soc.*, **74**, 1007–1023, doi:10.1175/1520-0477(1993)074<1007:ANPORG>2.0.CO;2.

Konduru, R. T., C. M. Kishtawal, and S. Shah, 2013: A new perspective on the infrared brightness temperature distribution of the deep convective clouds. *J. Earth Syst. Sci.*, **122**, 1195–1206, doi:10.1007/s12040-013-0345-4.

Lee, Y.-K., J. A. Otkin, and T. J. Greenwald, 2014: Evaluating the accuracy of a high-resolution model simulation through comparison with MODIS observations. *J. Appl. Meteor. Climatol.*, **53**, 1046–1058, doi:10.1175/JAMC-D-13-0140.1.

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83**, 407–430, doi:10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2.

Mecikalski, J. R., and K. Bedka, 2006: Forecasting convective initiation by monitoring the evolution of moving cumulus in daytime GOES imagery. *Mon. Wea. Rev.*, **134**, 49–78, doi:10.1175/MWR3062.1.

——, and Coauthors, 2007: Aviation applications for satellite-based observations of cloud properties, convective initiation, in-flight icing, turbulence, and volcanic ash. *Bull. Amer. Meteor. Soc.*, **88**, 1589–1607, doi:10.1175/BAMS-88-10-1589.

Mittermaier, M., and N. Roberts, 2010: Intercomparison of spatial forecast verification methods: Identifying skillful spatial scales using the fractions skill score. *Wea. Forecasting*, **25**, 343–354, doi:10.1175/2009WAF2222260.1.

——, and R. Bullock, 2013: Using MODE to explore the spatial and temporal characteristics of cloud cover forecasts from high-resolution NWP models. *Meteor. Appl.*, **20**, 187–196, doi:10.1002/met.1393.

Morcrette, J. J., 1991: Evaluation of model-generated cloudiness: Satellite-observed and model-generated diurnal variability of brightness temperature. *Mon. Wea. Rev.*, **119**, 1205–1224, doi:10.1175/1520-0493(1991)119<1205:EOMGCS>2.0.CO;2.

Murray, J. J., 2002: Aviation weather applications of Earth Science Enterprise data. *Earth Obs. Mag.*, **11**, 26–30.

Otkin, J. A., and T. J. Greenwald, 2008: Comparison of WRF model-simulated and MODIS-derived cloud data. *Mon. Wea. Rev.*, **136**, 1957–1970, doi:10.1175/2007MWR2293.1.

——, D. J. Posselt, E. R. Olson, H.-L. Huang, J. E. Davies, J. Li, and C. S. Velden, 2007: Mesoscale numerical weather prediction models used in support of infrared hyperspectral measurements simulation and product algorithm development. *J. Atmos. Oceanic Technol.*, **24**, 585–601, doi:10.1175/JTECH1994.1.

——, T. J. Greenwald, J. Sieglaff, and H.-L. Huang, 2009: Validation of a large-scale simulated brightness temperature dataset using SEVIRI satellite observations. *J. Appl. Meteor. Climatol.*, **48**, 1613–1626, doi:10.1175/2009JAMC2142.1.

Purdom, J. F. W., 1993: Satellite observations of tornadic thunderstorms. *The Tornado: Its Structure, Dynamics, Prediction, and Hazards*, *Geophys. Monogr.*, Vol. 79, Amer. Geophys. Union, 265–274.

Roberts, N. M., 2008: Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. *Meteor. Appl.*, **15**, 163–169, doi:10.1002/met.57.

——, and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, doi:10.1175/2007MWR2123.1.

Schwartz, C. S., 2014: Reproducing the September 2013 record-breaking rainfall over the Colorado Front Range with high-resolution WRF forecasts. *Wea. Forecasting*, **29**, 393–402, doi:10.1175/WAF-D-13-00136.1.

Sieglaff, J. M., L. M. Cronce, W. F. Feltz, K. M. Bedka, M. J. Pavolonis, and A. K. Heidinger, 2011: Nowcasting convective storm initiation using satellite-based box-averaged cloud-top cooling and cloud-type trends. *J. Appl. Meteor. Climatol.*, **50**, 110–126, doi:10.1175/2010JAMC2496.1.

Söhne, N., J.-P. Chaboureau, S. Argence, D. Lambert, and E. Richard, 2006: Objective evaluation of mesoscale simulations of the Algiers 2001 flash flood by the model-to-satellite approach. *Adv. Geosci.*, **7**, 247–250, doi:10.5194/adgeo-7-247-2006.

Thompson, G., M. Tewari, K. Ikeda, S. Tessendorf, C. Weeks, J. A. Otkin, and F. Kong, 2016: Explicitly-coupled cloud physics and radiation parameterizations and subsequent evaluation in WRF high-resolution convective forecasts. *Atmos. Res.*, **168**, 92–104, doi:10.1016/j.atmosres.2015.09.005.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Academic Press, 627 pp.

Willmot, C. J., and K. Matsuura, 2005: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Res.*, **30**, 79–82, doi:10.3354/cr030079.

Wolff, J. K., M. Harrold, T. Fowler, J. H. Gotway, L. Nance, and B. G. Brown, 2014: Beyond the basics: Evaluating model-based precipitation forecasts using traditional, spatial, and object-based methods. *Wea. Forecasting*, **29**, 1451–1472, doi:10.1175/WAF-D-13-00135.1.