DTC Visitor Final Report

The impact of radar data assimilation on short-term and next-day thunderstorm forecasts in the 2016 Community Leveraged Unified Ensemble (CLUE)

Patrick Skinner*

Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma and NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma

ABSTRACT

The impact of radar assimilation in the 2016 Community Leveraged Unified Ensemble (CLUE) is examined. Composite reflectivity forecasts are compared between two 10-member CLUE sub-ensembles that are identical except that one assimilates radar data using a three-dimensional variational technique with cloud analysis (RAD) and the other does not (NORAD). Forecasts are compared using neighborhood and object-based verification metrics. It is found that radar assimilation typically improves composite reflectivity forecasts for 3–6 hours; however, lower skill is found in next-day (6–30 hour) RAD forecasts. Differences in the next-day forecast skill are driven by thunderstorm coverage biases, with a low bias in RAD and a high bias in NORAD. These biases are consistent across cases and degrade the quality of forecasts from both ensembles. The contrasting biases in next-day RAD and NORAD composite reflectivity forecasts are influenced by variations in the extent and amplitude of predicted convective available potential energy (CAPE), with NORAD CAPE forecasts consistently predicting upwards of 10,000 more gridpoints with significant CAPE (>1000 J Kg^{-1}) than RAD forecasts. Variation in next-day CAPE forecasts is attributable to the impacts of short-term thunderstorm evolution, in particular large differences between convective cold pool representations in RAD and NORAD.

1. Introduction

Numerical weather prediction systems with horizontal grid spacing ≤4km capable of explicitly representing deep convection have profilerated over the past decade. Despite the growing availability of these convection allowing models (CAMs), many questions remain regarding their optimal design. These questions motivated mutiple institutions to conduct coordinated, controlled experiments on aspects of CAM configuration during the 2016 National Severe Storms Laboratory Hazardous Weather Testbed Spring Forecasting Experiment (SFE; Kain et al. 2003; Gallo et al. 2017). The resulting CAM superensemble, known as the Community Leveraged Unified Ensemble (CLUE; Clark et al. 2018), was used to conduct 8 experiments on CAM and CAM ensemble design. This study examines the fourth experiment: *Comparison of ensembles with and without radar assimilation.*

Assimilation of radar data provides a means of introducing energy at spatial scales typical of thunderstorms into CAMs initialized from coarser model analyses. CAM forecasts including these scales of motion in their initial conditions are often referred to as a "hot start" as opposed to "cold start" forecasts where initial conditions are interpolated from a coarser analysis and convective-scale motion develops during the forecast. As radar assimilation is expected to provide CAMs with more accurate initial conditions (Stensrud et al. 2009) it has been implemented in several real-time numerical weather prediction systems (e.g. Xue et al. 2003; Gao et al. 2004; Hu et al. 2006; Wheatley et al. 2015; Benjamin and Coauthors 2016). Although prior studies have examined the impacts of radar assimilation on short-term forecasts (Kain et al. 2010; Craig et al. 2012; Stratman et al. 2013; Keil et al. 2014; Moser et al. 2015; Surcel et al. 2016), comparitively few studies have examined the impact of radar assimilation on longer, next day, forecast periods of 6–30 hours[1].

Kain et al. (2010) first compared deterministic CAM forecasts with and without radar assimilation produced by the University of Oklahoma Center for Analysis and Prediction of Storms (CAPS) during the 2008 and 2009 SFE. They found that the benefits of radar assimilation could last as long as 15 hours for low quantitative precipitation forecast (QPF) rates. However, the forecast benefits for higher QPF thresholds or radar reflectivity values

**Corresponding author address:* Patrick Skinner Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, 120 David L. Boren Blvd. Norman, OK 7307.
E-mail: patrick.skinner@noaa.gov

[1] The terms "short term" and "next day" are used to describe the 0–6 and 6–30 hour forecast periods, respectively, throughout this manuscript.

typical of thunderstorms (i.e. >40 dBZ) only lasted 3–6 hours, which matched subjective evaluations provided by SFE participants. The longer benefits of radar assimilation for lower QPF rates were attributed to a phase lag in mesoscale convective system (MCS) forecasts initialized without assimilating radar data.

Stratman et al. (2013) compared CAPS forecasts with and without radar assimilation from the 2009 and 2010 SFE using traditional, neighborhood (Roberts and Lean 2008), and scale-separation (Casati 2010) verification techniques. Though variation is observed across different forecast quantities and verification metrics, their results largely reinforce those from Kain et al. (2010). The benefits of radar assimilation are found to last out to 12 hours for spatial scales greater than 40 km, but decrease to less than 6 hours for reflectivity values and spatial scales typical of thunderstorms. Similarly, Moser et al. (2015) used the CAPS Advanced Regional Prediction System (ARPS) three-dimensional variational (3DVAR) data assimilation system with cloud analysis (Xue et al. 2003; Gao et al. 2004; Hu et al. 2006) to initialize forecasts with and without radar data assimilation for 12 cases of heavy rainfall, finding that the benefits of radar assimilation in QPF forecasts typically lasted less than 12 hours.

These general findings that the direct benefits of radar assimilation on thunderstorm forecasts are limited to 3–6 hours are similar to intrinsic predictability limits for convective scales (Durran and Weyn 2016; Weyn and Durran 2017). However, as errors associated with deep, moist convection rapidly grow upscale to the mesoscale (Zhang et al. 2003, 2007), it is reasonable to expect that the impact of radar assimilation on short-term thunderstorm forecasts will have indirect impacts on prediction of the next-day mesoscale environment and convective evolution. For example, Carbone et al. (2002) have shown that coherent episodes of convective precipitation routinely occur over 24–48 hour periods, which is longer than the lifecycle of individual MCSs and provides evidence that antecedant convection influences subsequent convection initiation and evolution. This influence from prior convection is often manifest as features such as Mesoscale Convective Vortices and outflow boundaries along convectively-generated cold pools, which have long been known to influence next-day thunderstorm prediction (e.g. Bartels and Maddox 1991; Johns and Doswell 1992; Weckwerth and Parsons 2006). Several studies have demonstrated that improved prediction of these convectively-induced phenomena can improve subsequent forecasts of thunderstorm development and evolution (e.g. Stensrud et al. 1999; Liu and Xue 2018; Clark et al. 2010; Schumacher and Clark 2014; Thompson 2014; Nielsen and Schumacher 2016; Degelia et al. 2018).

In practice, many CAMs and CAM ensembles are initialized at 0000 UTC (including the CLUE) and produce 24–36 hour guidance that is used by operational meteorologists to generate forecasts, such as the Day 1 Convective Outlook issued by the Storm Prediction Center. As 0000 UTC occurs at the diurnal convective maximum across much of the United States, it is expected that the impact of radar assimilation on model initial conditions, short-term forecasts, and potentially next-day thunderstorm forecasts will be maximized at this time. Therefore, examination of the impacts of radar assimilation in CLUE thunderstorm forecasts will be of value to forecasters and the aim of this study is to quantify those impacts for both short-term and next-day forecasts.

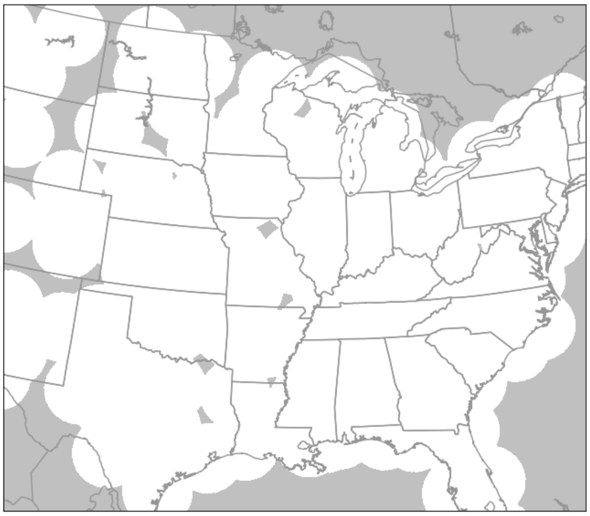

FIG. 1. The verification domain used in this study. Regions shaded gray are farther than 180 km from the nearest WSR-88D site and are not considered in verification.

Descriptions of forecast and verification datasets are provided in section 2. Comparisons between CLUE composite reflectivity and thunderstorm environment forecasts with and without radar assimilation are presented in section 3, as well as a more thorough comparison for a severe weather outbreak occurring on 24 May 2016. Discussion of differences between the two ensembles and recommendations for future research are provided in section 4 and conclusions are summarized in section 5.

2. Dataset Descriptions

a. The CLUE Forecast Dataset

This study is concerned with two 10-member subsets of the CLUE[2]: The *s-phys-rad* provided by CAPS and *s-phys-norad* provided by NSSL (the two subsets are hereafter described as RAD and NORAD, respectively). Both ensembles are run for the same CONUS domain with

[2] Readers are referred to Clark et al. (2018) for a complete description of the 2016 CLUE

3 km horizontal grid spacing using an identical model core and suite of physical parameterizations. All members use the WRF-ARW core (Skamarock et al. 2008) with Thompson microphysics (Thompson et al. 2008), the Mellor-Yamada-Janjic planetary boundary layer parameterization (Mellor and Yamada 1982), and NOAH land surface model (Ek et al. 2003). Corresponding members from each ensemble are initialized using the 0000 UTC NAM analysis with perturbations provided by NCEP's Short-Range Ensemble Forecast System. The only difference between the two ensembles is that the ARPS 3DVAR data assimilation (Xue et al. 2003; Gao et al. 2004) and cloud analysis system (Hu et al. 2006) is used to assimilate radar reflectivity, radial velocity, and traditional observations (i.e. surface and radiosonde) into the RAD ensemble[3] and a cold start with no data assimilation is used in the NORAD ensemble. The CLUE was run for 24 days between 2 May and 3 June 2016, with each member producing a 36-hour forecast for the majority of cases (Table 1).

Thunderstorms are approximated in CLUE forecasts using simulated composite reflectivity. For each hour of forecast output the instantaneous composite reflectivity field is verified over a domain covering roughly the eastern two-thirds of the continental United States, provided the grid box is less than 180 km from the nearest WSR-88D site (Fig. 1).

b. The Multi-Radar Multi-Sensor Verification Dataset

CLUE thunderstorm forecasts are verified against gridded composite reflectivity values provided by the Multi-Radar Multi-Sensor (MRMS; Smith et al. 2016) system. The MRMS composite reflectivity product quality controls WSR-88D Level 2 reflectivity observations using a neural net trained on polarimetric data (Lakshmanan et al. 2014) then merges individual radar observations onto a CONUS grid with 0.01° Latitude and Longitude spacing. The initial MRMS composite reflectivity field is interpolated onto the CLUE domain using a Cressman scheme with a 2 km radius of influence.

3. Analysis

a. Model Climatologies

One challenge with verifying thunderstorm forecasts is that corresponding observations that allow "apples-to-apples" comparisons are not available. This challenge includes composite reflectivity in CLUE forecasts and MRMS observations, where differences in sampling resolution, errors in the microphysical parameterization, and interpolation to a common grid prevent treating the quantities as equivalent. Therefore, percentile thresholds derived from CLUE and MRMS composite reflectivity climatologies are used in verification (Mittermaier and Roberts 2010; Sobash et al. 2016; Dawson et al. 2017). Composite reflectivity climatologies are constructed using the cumulative distribution function for CLUE or MRMS grid points in the forecast and verification datasets (Fig. 2) and thresholds are determined by matching extreme percentiles. Percentile thresholds allow corresponding values to be identified that minimize the frequency bias over the experimental period (i.e. the forecast and verification datasets will have the same number of gridpoints exceeding the percentile chosen over the course of the experiment).

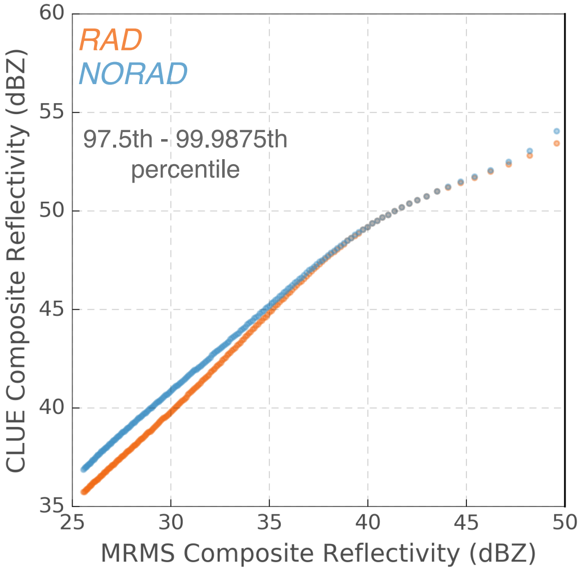

FIG. 2. Scatterplot of the 97.5th–99.875th percentile values of composite reflectivity aggregated across the CLUE dataset. The RAD (NORAD) ensembles are plotted in orange (blue) against corresponding percentiles of MRMS composite reflectivity interpolated to the model grid. The RAD and NORAD results are plotted in orange and blue, respectively through the remainder of the paper.

Comparison of climatologies for RAD, NORAD, and MRMS aggregated across all forecast times reveals composite reflectivity values in CLUE forecasts are generally higher than corresponding percentiles in MRMS data (Fig. 2). Both RAD and NORAD composite reflectivity values exceed matching percentiles in MRMS data by around 10 dBZ below approximately the 99th percentile of the climatologies, which corresponds to CLUE (MRMS) values of roughly 50 (40) dBZ in Fig. 2. This difference is attributable in part to the Cressman interpolation of MRMS values to the CLUE grid, which will smooth the highest values within the domain. A climatology created using

[3] Though multiple observation types are assimilated into the RAD ensemble, hydrometeor mixing ratios are primarily determined by radar reflectivity (Hu et al. 2006) and the broad term 'radar assimilation' is used throughout this study for simplicity.

TABLE 1. Summary of cases that make up the CLUE 2016 dataset. Case number, date, available forecast hours, and total daily Local Storm Reports (LSRs) from the Storm Prediction Center are provided in the first 4 columns. The fifth column lists missing forecasts with "NR" and "R" representing the NORAD and RAD ensembles, respectively and missing forecast hours provided parenthetically. An asterix indicates that only thermodynamic variables are missing.

Case	Date	Forecast Hours	Total LSRs	Missing Forecasts
1	2 May	36	123	NR 02(24)*
2	3 May	36	120	
3	4 May	36	16	NR 02(28)* NR 05(25)* NR 07(27)*
4	5 May	36	10	
5	6 May	24	12	
6	9 May	36	170	NR 02(23)* NR 05(36)* NR 10(33)*
7	10 May	36	220	NR 03(27)
8	11 May	36	233	
9	12 May	36	79	NR 04(23)* NR 08(35)*
10	13 May	24	69	NR 01(4,8,21)* NR 03(8,16)* NR 04(2)* NR 05(20)* NR 07(16)* NR 09(09)*
11	16 May	36	78	
12	17 May	36	54	
13	18 May	36	10	
14	19 May	36	75	
15	20 May	24	37	
16	23 May	36	72	
17	24 May	36	153	R 06(24–36)
18	25 May	36	163	
19	26 May	36	290	
20	27 May	24	66	
21	31 May	36	64	
22	1 June	36	61	R 05(15)
23	2 June	36	51	
24	3 June	24	38	R 07(4,8,12)

the uninterpolated MRMS values is generally about 6 dBZ higher than the interpolated climatology (not shown).

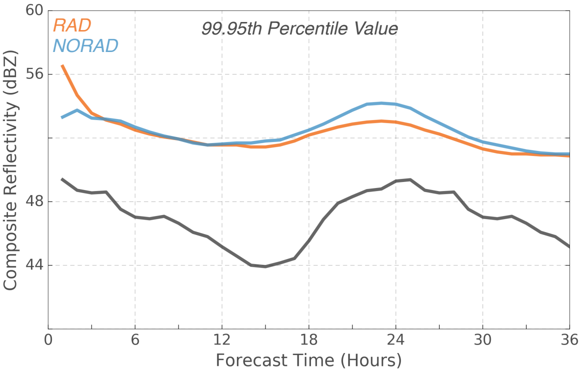

FIG. 3. Time series of the 99.95th percentile value of (orange) RAD, (blue) NORAD, and (black) MRMS composite reflectivity at each available forecast hour.

Differences are also apparent between the RAD and NORAD climatologies, with NORAD forecasts having slightly higher composite reflectivity values at corresponding percentiles. The differences between the two ensembles become more apparent when a specific percentile (99.95%) is examined for values aggregated at each forecast hour (Fig. 3). As would be expected, RAD values are higher during the initial forecast hours as convection develops within the cold start NORAD ensemble. However, RAD values during the first four hours of the forecast are higher than either ensemble during the next-day diurnal maximum (forecast hours 22–26). Higher composite reflectivity values occur at the 99.95th percentile of the NORAD ensemble at later forecast times, particularly during the next-day maximum. These differences in the next-day composite reflectivity climatology, which are exacerbated for percentile thresholds greater than 99.95%, indicate that NORAD forecasts contain more total gridpoints with high values of composite reflectivity than the RAD ensemble during this period. Calculating an hourly climatology also allows the diurnal cycle of convection in RAD and NORAD to be compared to observations (Fig. 3). Both ensembles roughly match the observed diurnal cycle in timing, but display a smaller range of values between the morning minimum and evening max.

b. Verification of Thunderstorm Forecasts

1) NEIGHBORHOOD VERIFICATION

Skill in RAD and NORAD composite reflectivity forecasts is first assessed using the Fractions Skill Score (FSS; Roberts and Lean 2008). The FSS is calculated for each member of the RAD and NORAD ensembles at each available forecast hour, allowing the distributions of scores between the ensembles over the course of the experiment to be compared (Fig. 4). A 40 dBZ composite reflectivity threshold is chosen to define thunderstorms in CLUE forecasts, 30 dBZ is used as a threshold for MRMS observations as it approximately matches the 40 dBZ percentile in the RAD and NORAD climatologies (Fig. 2).

Consistent with past studies (Kain et al. 2010; Craig et al. 2012; Stratman et al. 2013; Keil et al. 2014; Moser et al. 2015; Surcel et al. 2016), the RAD ensemble generally produces more skillful short-term forecasts. This initial period of improved skill extends until forecast hour 5 (2) for a 75 km (295 km) diameter neighborhood, and indicates improved initial conditions from radar assimilation improves short-term prediction of convective storms.

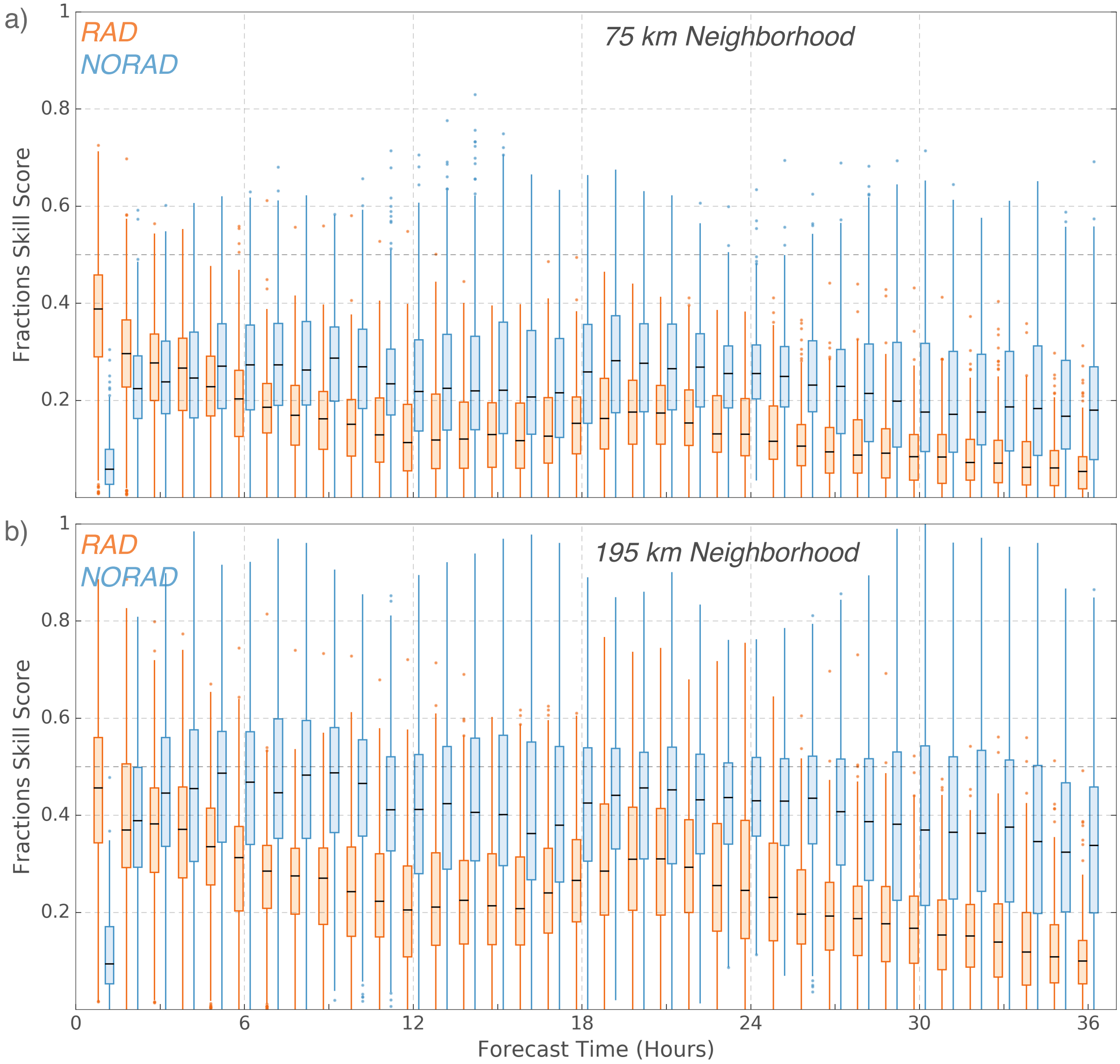

FIG. 4. Hourly box-and-whisker plots of Fractions Skill Score distributions for each composite reflectivity forecast issued by the RAD and NORAD ensembles. FSS are calculated using thresholds of 40 (30) dBZ for CLUE (MRMS) composite reflectivity and neighborhoods with a (a) 75 km and (b) 195 km diameter.

However, distributions between the two ensembles rapidly become similar and the median FSS of the NORAD ensemble surpasses RAD by forecast hour six for each reflectivity threshold and neighborhood tested[4] then remains higher through the duration of the forecast (Fig. 4). FSS differences between NORAD and RAD are particularly pronounced for larger neighborhoods and forecast hours outside the next-day diurnal maximum (i.e. hours 6–18 and 24–36), where the entire interquartile range of the NORAD FSS distribution is above the interquartile range of the RAD distribution. These large differences in FSS between the two ensembles following the first six forecast hours suggest that radar assimilation has a negative impact on thunderstorm forecast skill for the majority of the forecast period.

Despite large relative differences in FSS between RAD and NORAD, the median FSS for both ensembles remains below 0.5 for all forecast hours and neighborhoods con-

[4] FSS was calculated for CLUE composite reflectivity thresholds of 30, 35, 40, 45, and 50 dBZ with corresponding MRMS thresholds 10 dBZ lower and for neighborhood diameters of 75, 105, 135, 165, and 195 km. FSS values change across different thresholds and neighborhoods, but comparisons between RAD and NORAD remain similar.

sidered, indicating generally unskillful forecasts of composite reflectivity according to the criteria of Roberts and Lean (2008). A majority of reflectivity forecast FSS values fall below 0.5, even for a large neighborhood where skill would be expected to begin to asymptote (Fig. 4b), suggesting that frequency biases are potentially present in reflectivity forecasts from both the RAD and NORAD ensemble and are resulting in lower asymptote FSS (Roberts and Lean 2008).

2) Object-based Verification

The skill of RAD and NORAD composite reflectivity forecasts are also assessed using the Method for Object-based Diagnostic Evaluation (MODE; Davis et al. 2006a,b). Object-based verification techniques, such as MODE, allow a customizable total interest value to be defined that matches forecast objects to corresponding objects in a verification dataset. The configurable nature of object matching, ability to match objects across different forecast and verification fields (provided they are consistently defined [Wolff et al. 2014]) and ability to quantify specific forecast errors, such as location biases, make object-based techniques attractive for verification of thunderstorm forecasts (e.g. Burghardt et al. 2014; Pinto et al. 2015; Cai and Dumais 2015; Skinner et al. 2018).

MODE identifies objects in CLUE and MRMS composite reflectivity fields by applying a circular convolution filter with a 9 km radius to each gridpoint, then applying a threshold to the smoothed fields. Application of the convolution filter complicates the choice of a composite reflectivity threshold as values typically used for thunderstorm identification (i.e. 40 dBZ) will be smoothed in the convolved field. Given the small convolution radius applied to CLUE forecasts, which is intended to retain individual thunderstorm cells, thresholds are reduced by 5 dBZ from those used for the FSS to 35 (25) dBZ in CLUE (MRMS) composite reflectivity fields. Tests using thresholds between 25 and 45 dBZ did not significantly change interpretation of verfication scores. Following thresholding, composite reflectivity objects are retained if they encompass at least 15 contiguous grid boxes (135 km^2). The object area threshold is intended to reduce the impact of small reflectivity objects, which can result from weak or misidentified thunderstorms, on verification scores.

Object pairs between CLUE forecasts and MRMS observations are matched using a total interest score composed of measures of spatial displacement and the area ratio of the objects. The majority of the total interest is determined by spatial displacement between forecast and observed object pairs, with the object centroid displacement and boundary displacement (the minimum distance between object edges) each accounting for 40% of the overall total interest score. The final 20% of the total interest score is determined by the area ratio of the object pair, calculated with the larger area as the denominator. Fuzzy logic-based interest functions are used for assigning a value between 0 and 1 to each input to the final total interest score. Centroid displacement is assigned a value of 1 for distances $\leq$10 km that linearly decreases to 0 at 60 km displacement and boundary displacement is similar except that values decrease from 1 to 0 for all displacements up to 60 km. The area ratio interest function is assigned a value of 1 for ratios $\geq$0.8 that decreases linearly to 0 as ratios decrease. Object pairs are considered matched if the total interest score is greater than or equal to 0.6. For example, two objects of equal size with a 35 (30) km centroid (boundary) displacement would have a total interest score of 0.4*0.5 + 0.4*0.5 + 0.2*1.0 = 0.6 and be considered a match.

Object matching allows each composite reflectivity object in the CLUE and MRMS datasets to be classified as a hit (matched object pair)[5], false alarm (unmatched CLUE object), or miss (unmatched MRMS object) and object-based contingency table metrics to be calculated (e.g. Griffin et al. 2017a,b; Skinner et al. 2018). Aggregation of contingency table elements across all forecast cases and ensemble members allows the ensemble mean probability of detection (POD), false alarm ratio (FAR), frequency bias (BIAS), and critical success index (CSI) of the RAD and NORAD ensembles to be calculated for each forecast hour (Fig. 5).

Clear differences in object-based verification metrics of RAD and NORAD composite reflectivity forecasts are present. Statistically significant differences for a 99% confidence interval, calculated using a resampling technique with 1000 iterations (Hamill 1999), are present between RAD and NORAD POD, FAR, BIAS, and CSI for nearly all forecast hours (Fig. 5). Differences in POD, FAR, and CSI are influenced by contrasting biases in the number of composite reflectivity objects, with NORAD generally exhibiting a high bias and RAD a low bias. As NORAD forecasts are initialized without hydrometeors, biases are initially very low but increase rapidly as convection develops within the ensemble and remain near 1 (unbiased) for forecast hours 3–15. NORAD biases increase during the next-day convective maxima, with overforecast biases present between forecast hours 16 and 27 before dropping to near, or slightly below 1 for the final 8 hours of the forecast period. Biases in the RAD ensemble follow a similar diurnal pattern, but are considerably lower and remain below 1 throughout the forecast period. As would be expected for forecasts with a larger frequency bias, the NORAD ensemble has a higher POD than RAD for forecast hours

[5] MODE permits multiple forecast objects to be matched to the same observed object. As a result, two different measures of matched object pairs are possible: Matched forecast objects and matched observed objects. The number of matched forecast objects are chosen as the measure for ‘hits’, although small changes in contingency table metrics occur if matched observed objects are used, qualitative score comparisons between RAD and NORAD are similar (not shown).

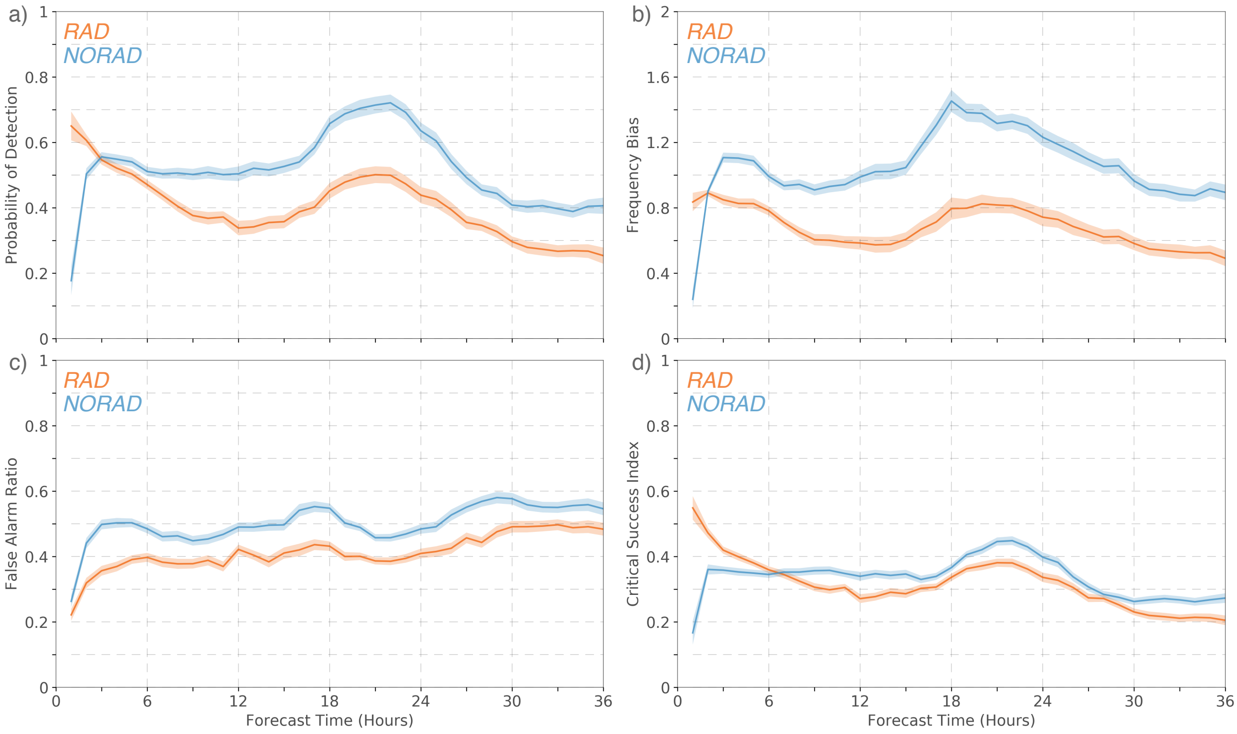

FIG. 5. Time series of MODE contingency table metrics for the RAD and NORAD ensemble forecasts of composite reflectivity, including (a) Probability of Detection, (b) frequency bias, (c) false alarm ratio, and (d) critical success index. Solid lines represent the ensemble mean for each forecast hour and shading encompasses the 99% confidence interval.

greater than 3 and a higher FAR for each forecast hour. However, the relative difference between NORAD and RAD POD is greater than FAR, resulting in higher NORAD CSI values for forecast hours 6–36. Differences between verification metrics for NORAD and RAD increase if scores are weighted by object area, with RAD scores decreasing while NORAD scores increase (not shown). This decrease in RAD scores indicates that the ensemble not only produces fewer thunderstorm objects than the NORAD ensemble or MRMS observations, but that objects are smaller in area, resulting in a more severe underforecast.

Hourly MODE contingency table scores (Fig. 5) are generally consistent with corresponding time series of FSS (Fig. 4) and the reflectivity climatology (Fig. 3). Both FSS and MODE scores support a 3–6 hour initial period of improved skill in the RAD ensemble attributable to reflectivity assimilation (e.g. Kain et al. 2010; Stratman et al. 2013) followed by improved skill in the NORAD ensemble for the remainder of the forecast period. Both FSS and MODE scores indicate a relative peak in skill for both ensembles during the next day convective maximum between forecast hours 18 and 24 (Figs. 4, 5); however, this peak is likely strongly influenced by the number of observed thunderstorms. For example, the hourly climatology of MRMS reflectivity (Fig. 3) has a greater range of values between the diurnal minimum (hour 15) and maximum (hour 25) than either ensemble, which would result in more observed gridpoints/objects to verify against during the diurnal maximum and a commensurate increase in skill scores.

Aggregation of MODE contingency table elements by forecast hour provides a bulk measure of forecast skill but does not provide insight into case-to-case variations in skill, which have been identified in past studies examining the impact of radar assimilation (Craig et al. 2012; Keil et al. 2014; Surcel et al. 2016). Therefore, contingency table elements are additionally aggregated for 3-hour periods during the expected cross-over time for skill in the RAD and NORAD ensembles (forecast hours 4–7) and during the next-day convective maximum (forecast hours 21–24) for each day during the experiment. Performance diagrams (Roebber 2009) for ensemble mean scores from each case show that the contrasting biases between NORAD and RAD were generally consistent across cases (Fig. 6). During the cross-over period (Fig. 6a) NORAD cases exhibit small biases, although a slim majority of cases have a slight overforecast between 1 and 1.5. The case-by-case distribution of RAD biases covers a larger range than NORAD (roughly <0.1 to 1.5) with most cases

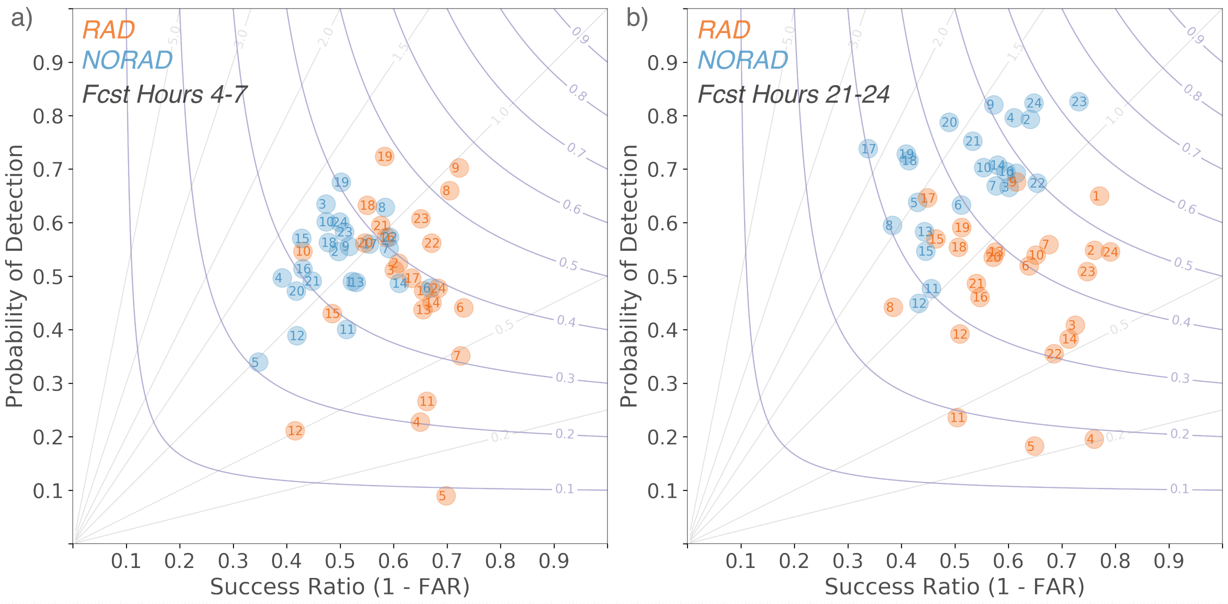

FIG. 6. Performance diagrams (Roebber 2009) of ensemble mean scores for each RAD and NORAD case in the 2016 CLUE dataset for composite reflectivity forecasts between (a) 4–7 hours and (b) 21–24 hours. Numbers within each marker correspond to the case numbers in Table 1.

underforecasting composite reflectivity objects. Case-by-case variation is also larger for RAD CSI scores, which vary between <0.1 and nearly 0.6 compared to roughly 0.2 to 0.45 for NORAD cases. The subset of RAD cases with higher CSI scores than any NORAD case indicates that, for some cases, improved forecast initial conditions provided by radar assimilation are resulting in improvements in composite reflectivity forecasts out to 6 hours or beyond. However, a separate subset of RAD cases with low frequency biases <0.5 and lower CSI scores than any NORAD case suggest that radar assimilation can also degrade short-term forecasts.

Consistent case-by-case differences in biases of NORAD and RAD forecasts are present during the next-day convective maximum (Fig. 6b). A frequency bias >1 is present for each NORAD case during this period while the majority of RAD cases exhibit a bias <1. Though ranges in CSI scores are similar between the two ensembles during this period, the range of NORAD scores is roughly 0.1 higher than RAD scores (i.e. roughly 0.3–0.6 for NORAD and 0.2–0.5 for RAD). This relative improvement in NORAD forecasts is consistent with bulk FSS and contingency table metrics (Figs. 4, 5). Despite the overall higher scores for NORAD forecasts, there are several cases where NORAD and RAD CSI scores are similar but NORAD biases are higher. These cases include the two periods producing the most widespread severe weather, as measured by the count of local storm reports (Table 1) on 9, 10 May (cases 6, 7) and 24–26 May (cases 17–19).

3) SUBJECTIVE VERIFICATION: 24 MAY 2016

In order to provide a sense of how differences in the bulk verification measures of RAD and NORAD manifest in a single composite reflectivity forecast, the 24 May 2016 case is examined in greater detail (Fig. 7). The 24 May case is selected as it was one of the most active days in the experiment for severe thunderstorms, producing one of the highest totals of local storm reports during the experiment (Table 1) and multiple tornadic supercells in eastern Colorado and western Kansas (Weinhoff et al. 2018).

Comparison of RAD and NORAD 1-hour forecasts reveals the impacts of 3DVAR radar assimilation and cloud analysis on short-term forecasts (Figs. 7a, d). Storms have just begun to develop in the NORAD forecast and the ensemble mean number of gridpoints exceeding 40 dBZ is roughly one third of the number of MRMS gridpoints exceeding 30 dBZ. In contrast, the RAD forecast shows most members predicting thunderstorms coincident with observations; however, the spatial extent of these thunderstorms is too large, resulting in an ensemble mean gridpoint count of composite reflectivity more than twice as large as the count of MRMS values. By forecast hour six (Figs. 7b, e) more storms have developed in NORAD members and the overforecast bias in RAD has decreased, with both ensembles providing a generally accurate prediction of thunderstorms extending from western Iowa southwestward to western Kansas. Although these storms are accurately predicted in both ensembles, most RAD members, and a few NORAD members, have erroneously predicted

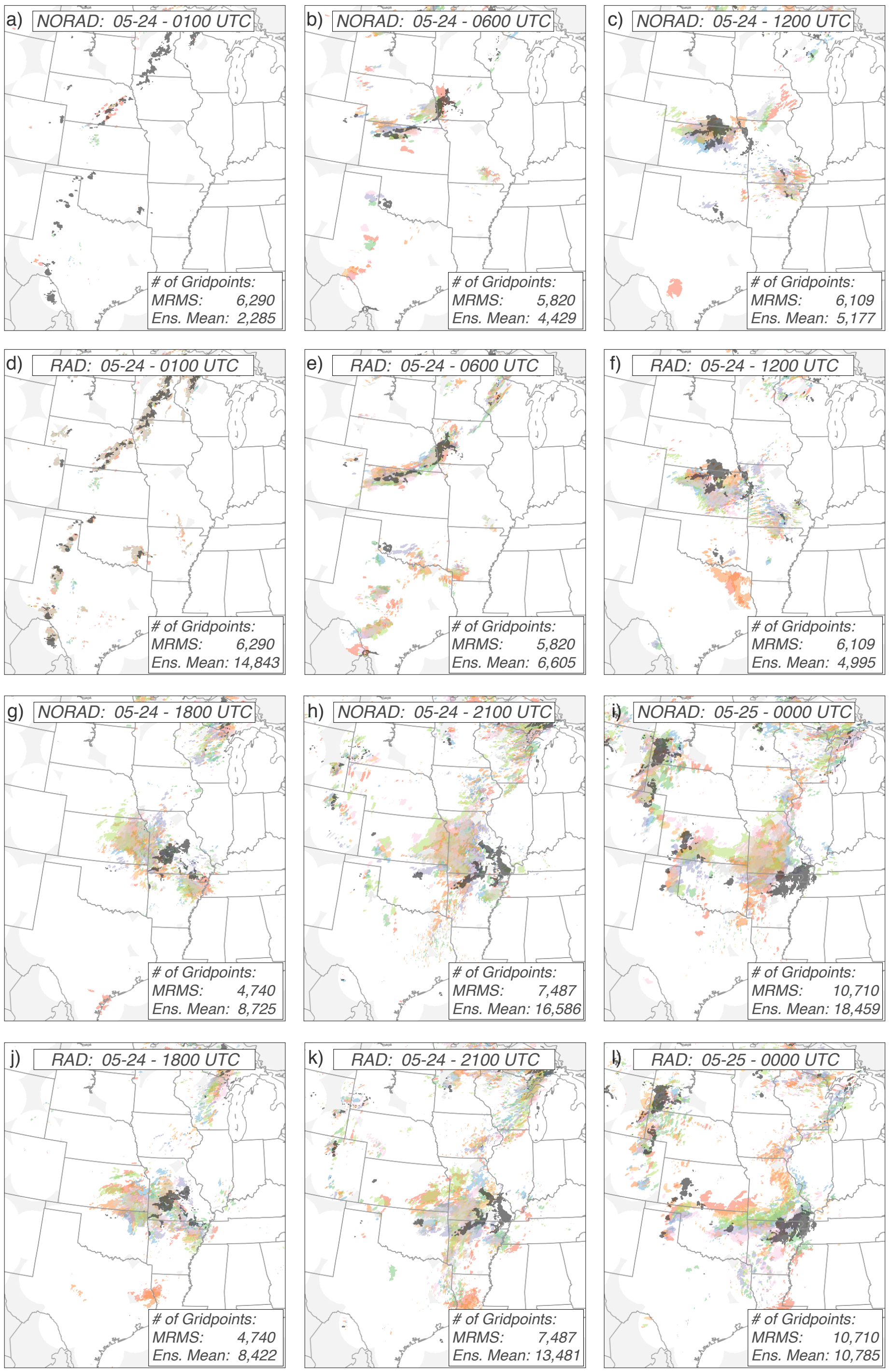

FIG. 7. Paintball plots of (a, b, c, g, h, i) NORAD and (d, e, f, j, k, l) RAD forecasts initialized at 0000 UTC on 24 May 2016 and valid at (a, c) 0100, (b, e) 0600, (c, f) 1200, (g, j) 1800, (h, k) 2100, and (i, l) 0000 UTC on 24 and 25 May. Gridpoints with composite reflectivity exceeding 40 dBZ are plotted with a unique color for each member and corresponding MRMS composite reflectivity values exceeding 30 dBZ are shaded gray. Light gray shading indicates masked regions in the verification domain and the ensemble mean and total MRMS gridpoints exceeding the prescribed thresholds are annotated at the lower right of each panel. Note that only a portion of the verification domain is plotted to improve clarity.

dryline convection in western Texas to grow upscale and move into central Texas. In general, the coverage of spurious convection is larger in the RAD ensemble, and remains larger at forecast hour 12 (Figs. 7c, f) despite both ensembles having a small low bias in gridpoint reflectivity counts. The RAD 12-hour forecast additionally predicts the decaying MCS in central Kansas to cover a larger area and propogate further southward than either the NORAD ensemble or MRMS observations.

Despite the large differences in RAD and NORAD 12-hour forecasts, the coverage of composite reflectivity objects in the two ensembles are remarkably similar at forecast hour 18 (Figs. 7g, j). Both ensembles have a strong overprediction bias, with nearly double the ensemble mean gridpoints associated with convection as MRMS observations. This overprediction is consistent with large differences between the CLUE and MRMS climatologies in the 12–18 hour period (Fig. 3). With the exception of one outlier member in RAD predicting an MCS along the Texas and Louisiana border, the overprediction results from a greater spatial extent of thunderstorms in Missouri and the Great Lakes Region. An overprediction of thunderstorm coverage is maintained in both ensembles at 21 UTC (Figs. 7h, k) but is stronger in the NORAD ensemble[6]. One reason for the larger bias in the NORAD ensemble is that more expansive convection initiation is predicted along the front range of the Rocky Mountains in Colorado and Wyoming and along the dryline extending southward from a triple point with the outflow boundary of the nocturnal MCS in southwestern Kansas. NORAD overprediction of thunderstorms is maintained in the 24-hour forecast (Fig. 7i), with dramatic overpredictions of thunderstorm coverage in Missouri extending northeastward into the Great Lakes Region, in the western portion of the domain from roughly the Black Hills of South Dakota southward into northeastern Colorado, and along the dryline from western Kansas into the Texas Panhandle. This positive bias results in nearly twice as many ensemble mean NORAD gridpoints associated with thunderstorms as MRMS observations, which cover a total area of >70,000 km^2. The 24-hour RAD forecast is nearly unbiased, which is atypical of RAD forecasts during this period (Figs. 5, 6). Despite the improved bias, significant location errors are present in the RAD forecast, including an overprediction of convection along the eastern extent of an outflow boundary from the nocturnal MCS through southeastern Kansas, northeastern Oklahoma, and into Missouri and a southward shift in the locations of dryline thunderstorms. This southward shift results in minimal thunderstorm coverage over southwestern Kansas, where the majority of the severe weather, including 28 tornado reports, occurred.

The 24 May composite reflectivity forecast provides an example of how the evolution of nocturnal convection, in this case the MCS in the central Plains, can influence the location and timing of thunderstorm development the following day (Stensrud et al. 1999). Biases in overnight thunderstorm coverage impacted the location and extent of the cold pool produced by the MCS and an associated outflow boundary responsible for convection initiation the following afternoon. Additionally, more widespread thunderstorm coverage in the RAD ensemble overnight followed by less coverage the following day suggest that radar assimilation is impacting the extent and magnitude of the next-day potential instability.

c. Verification of the Thunderstorm Environment

Differences in the next-day convective environment of the RAD and NORAD ensembles are first examined using ensemble mean surface-based Convective Available Potential Energy (CAPE; Fig. 8). For 24 May, relatively large ensemble mean CAPE differences are already present 1 hour into the forecast (Figs. 8a, d). Though the coverage of positive CAPE is consistent between the two ensembles, large differences in magnitude, locally greater than 1000 J Kg^{-1}, are present from the Gulf Coast through central Minnesota. Furthermore, the axis of high CAPE values (>3600 J Kg^{-1}) ahead of the dryline in western Texas is smaller in the RAD ensemble with CAPE minima, indicative of convective cold pools, coincident with the locations of predicted thunderstorms (Fig. 7d).

Differences between the 6-hour RAD and NORAD CAPE forecasts are dramatic (Figs. 8b, e). Convective cold pools with ensemble mean CAPE values near zero have spread across much of western Texas in the RAD ensemble whereas the NORAD CAPE field in west Texas is largely unchanged from the 1-hour forecast. The extent of cold pools in the RAD forecast is likely exacerbated by the spurious maintenance of dryline convection into central Texas (Fig. 7e). The near-zero CAPE values within cold pools in the RAD ensemble have diminished by forecast hour 12 (Fig. 8f) but large differences (>3000 J Kg^{-1}) persist throughout west Texas. CAPE values in Texas and the southern Plains have largely recovered by 18 UTC in the RAD forecast but generally remain lower than those in the NORAD forecast (Figs. 8g, j). Additionally, differences in the location and spatial extent of the cold pool associated with the nocturnal MCS in Nebraska and Kansas are apparent at 18 UTC, with a larger cold pool in the RAD ensemble that has propagated into northern Oklahoma. The additional southward propagation of this outflow boundary in the RAD ensemble results in weaker CAPE in southwestern Kansas and eastern Colorado during the early afternoon and a dearth of storms in the regions of widespread severe weather on 24 May (Figs. 8h, 8k, 7i, 7j). Where the RAD ensemble

[6] MODE object-based frequency biases during this period are the highest of any case during the experiment for both ensembles (Fig. 6b; Case 17).

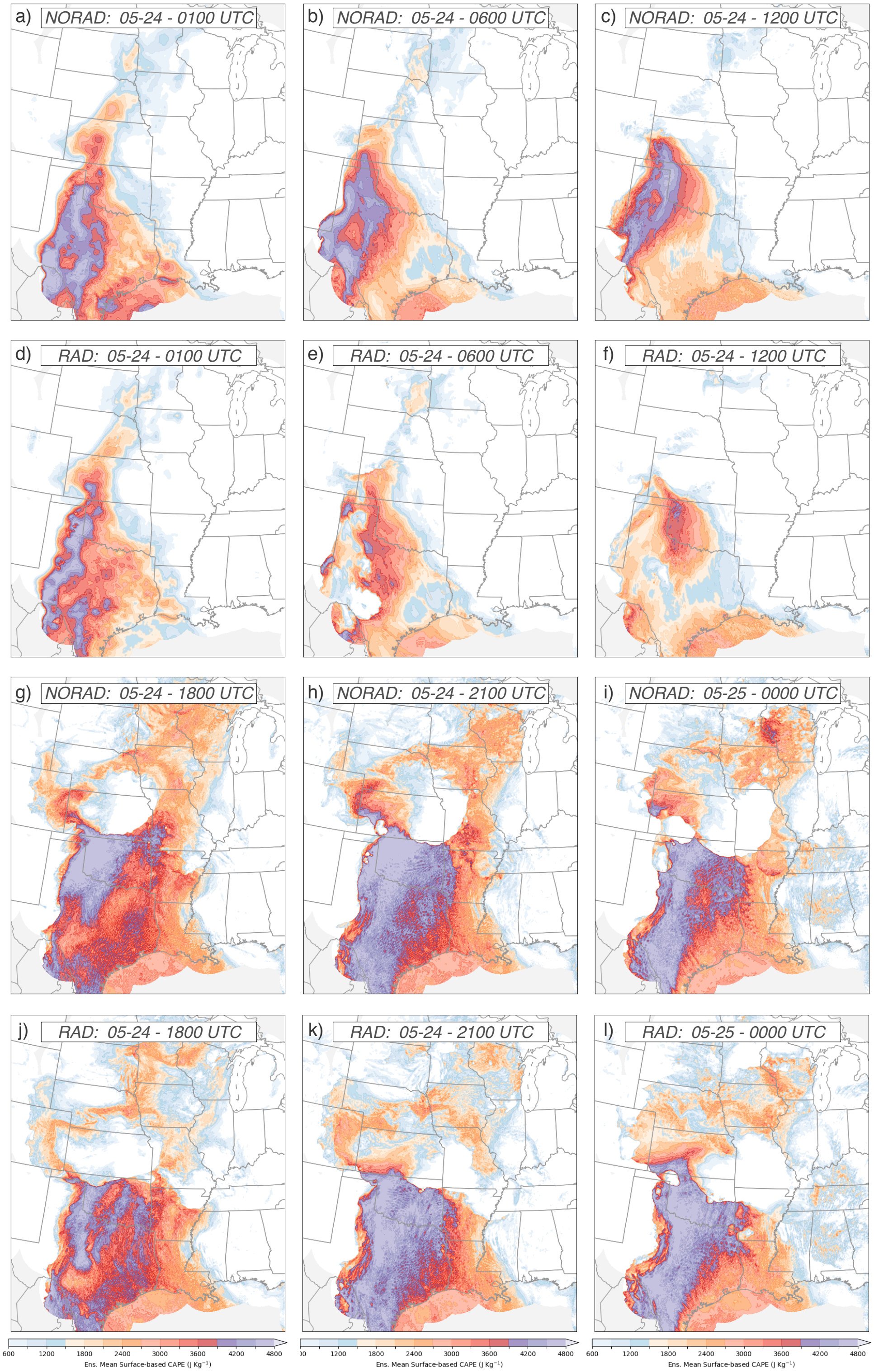
a) NORAD: 05-24 - 0100 UTC
b) NORAD: 05-24 - 0600 UTC
c) NORAD: 05-24 - 1200 UTC
d) RAD: 05-24 - 0100 UTC
e) RAD: 05-24 - 0600 UTC
f) RAD: 05-24 - 1200 UTC
g) NORAD: 05-24 - 1800 UTC
h) NORAD: 05-24 - 2100 UTC
i) NORAD: 05-25 - 0000 UTC
j) RAD: 05-24 - 1800 UTC
k) RAD: 05-24 - 2100 UTC
l) RAD: 05-25 - 0000 UTC
600 1200 1800 2400 3000 3600 4200 4800
Ens. Mean Surface-based CAPE (J Kg^{-1})
00 1200 1800 2400 3000 3600 4200 4800
Ens. Mean Surface-based CAPE (J Kg^{-1})
600 1200 1800 2400 3000 3600 4200 4800
Ens. Mean Surface-based CAPE (J Kg^{-1})

FIG. 8. As in Fig. 7 except for ensemble mean surface-based CAPE (J Kg^{-1})

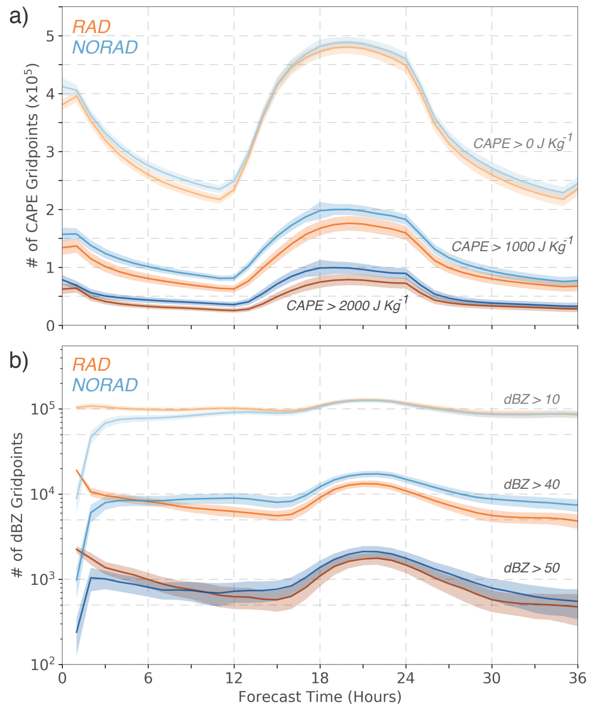

FIG. 9. Hourly ensemble mean counts of gridpoints exceeding thresholds of (a) 0, 1000, and 2000 J Kg^{-1} of surface-based CAPE and (b) 10, 40, and 50 dBZ in composite reflectivity. Shading indicates the region within one standard deviation of the mean, calculated as the mean of standard deviations for each individual case. A logorithmic y-axis is used for composite reflectivity counts to improve clarity.

has limited to moderate potential instability in the corridor from southwest Kansas to eastern Colorado where most severe weather reports occurred, the NORAD ensemble mean CAPE often exceeds 3000 J Kg^{-1}. In general, the CAPE field in the NORAD forecast is higher than RAD throughout the domain at 18, 21, and 00 UTC. The greater extent and magnitude of potential instability in the NORAD next-day forecasts suggests that the high frequency bias in NORAD composite reflectivity objects (Fig. 7) is at least partially attributable to greater potential instability.

The tendency for the NORAD ensemble to predict more gridpoints with high composite reflectivity and surface-based CAPE is consistent throughout the experiment (Fig. 9). The ensemble mean count of gridpoints exceeding 40 dBZ in the NORAD ensemble is several thousand gridpoints greater than the RAD count beyond forecast hour 9. Similarly, counts of ensemble mean surface-based CAPE $>$1000 J Kg^{-1} in NORAD are more than 10,000 points higher than RAD for the majority of the forecast period. Although differences are smaller, NORAD also predicts more gridpoints exceeding higher composite reflectivity (CAPE) thresholds of 50 dBZ (2000 J Kg^{-1}). In contrast, a smaller increase in NORAD gridpoint counts of CAPE greather than 0 J Kg^{-1} is present and counts of reflectivity exceeding 10 dBZ are similar for both ensembles during much of the forecast. Furthermore, higher gridpoint counts of composite reflectivity greather than 10 dBZ occur in the RAD ensemble during the first 12 hours. These higher counts suggest that radar assimilation is affecting forecasts for a longer period of time for weak reflectivity thresholds, which are typical of stratiform precipitation and larger spatial scales than thunderstorms and consistent with the results of Kain et al. (2010), Stratman et al. (2013), and Surcel et al. (2016).

To further compare the impact of radar assimilation on the next-day environment, predicted 2-m temperature and dewpoint temperature fields in the first member of the RAD and NORAD ensembles are verified against Automated Surface Observing System (ASOS) observations on 24 May (Figs. 10, 11). Consistent, and sometimes large, biases are found compared to ASOS observations in both members across much of the domain. For example, a cool and moist bias is present in both the RAD and NORAD members along and east of the Mississippi River Valley throughout the forecast, with a broad region of large, positive dewpoint errors in excess of 5 K in the 18-hour forecast (Figs. 11c, f, 12). The generally similar errors between the two members result in similar distributions of point verification metrics such as median error (Fig. 12) and root mean square error (not shown). However, locally large differences in near-surface temperature and dewpoint are present between the RAD and NORAD members near thunderstorms (Figs. 10g–i, 11g–i).

The influence of convective cold pools on the 1-hour RAD 2-m temperature forecast is obvious (Fig. 10g) and colocated with observed regions of high reflectivity (Fig. 7a). These cold pools result in smaller positive temperature biases in the RAD member in northeastern Colorado and Minnesota; however, there is some evidence a cold bias is being introduced by radar assimilation in the northeast Texas Panhandle (although most predicted cold pools in West Texas lie between ASOS observations). A Conflicting temperature bias in the RAD and NORAD ensembles is present near convection in central Nebraska, with errors of up to -5 K in the RAD ensemble coincident with errors of 5 K in the NORAD ensemble. These conflicting biases provide an example of how radar assimilation can lead to overprediction of cold pool intensity and result in an overcorrection of temperature errors in a cold start forecast.

Changes associated with radar assimilation are more widespread in the 1-hour near-surface dewpoint forecast (Fig. 11g) and largely result in higher values in the RAD member than NORAD member (Fig. 12). Higher dewpoints in the RAD member reduce negative biases in the Southeast US and a small area of lower dewpoints in RAD results in a smaller positive bias in central Nebraska. However, larger positive biases compared to ASOS observations are present in the majority of regions where radar as-

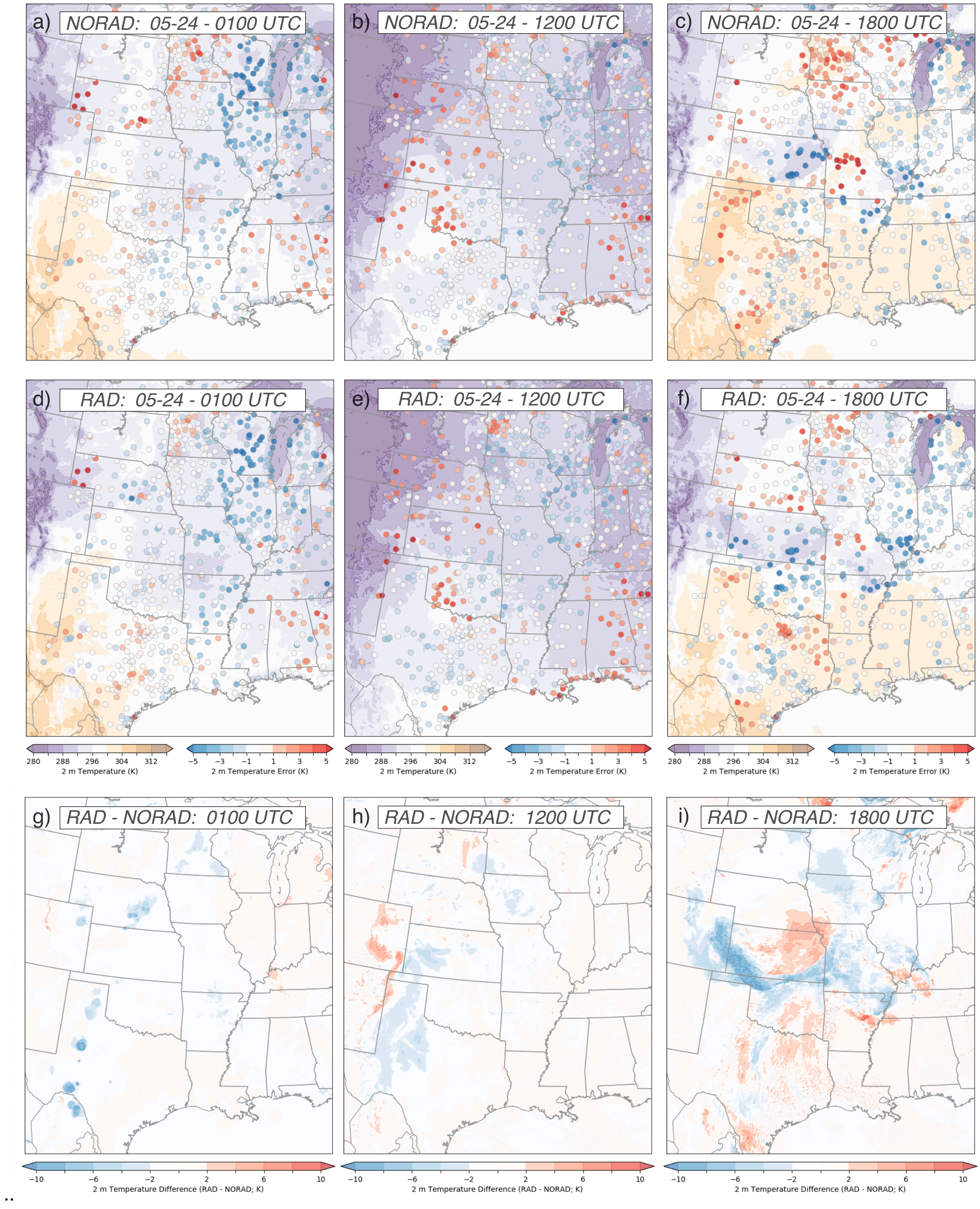

FIG. 10. Plots of 2 m temperature (K) for the first member of the (a, b, c) NORAD and (d, e, f) RAD ensembles intitialized at 0000 UTC 24 May and valid at (a, d) 0100, (b, e) 1200, and (c, f) 1800 UTC. ASOS observations are overlain on each plot and color coded according to the difference resulting from subtracting the observed value from the predicted value. The difference between the RAD and NORAD forecasts is plotted in panels g, h, and i.

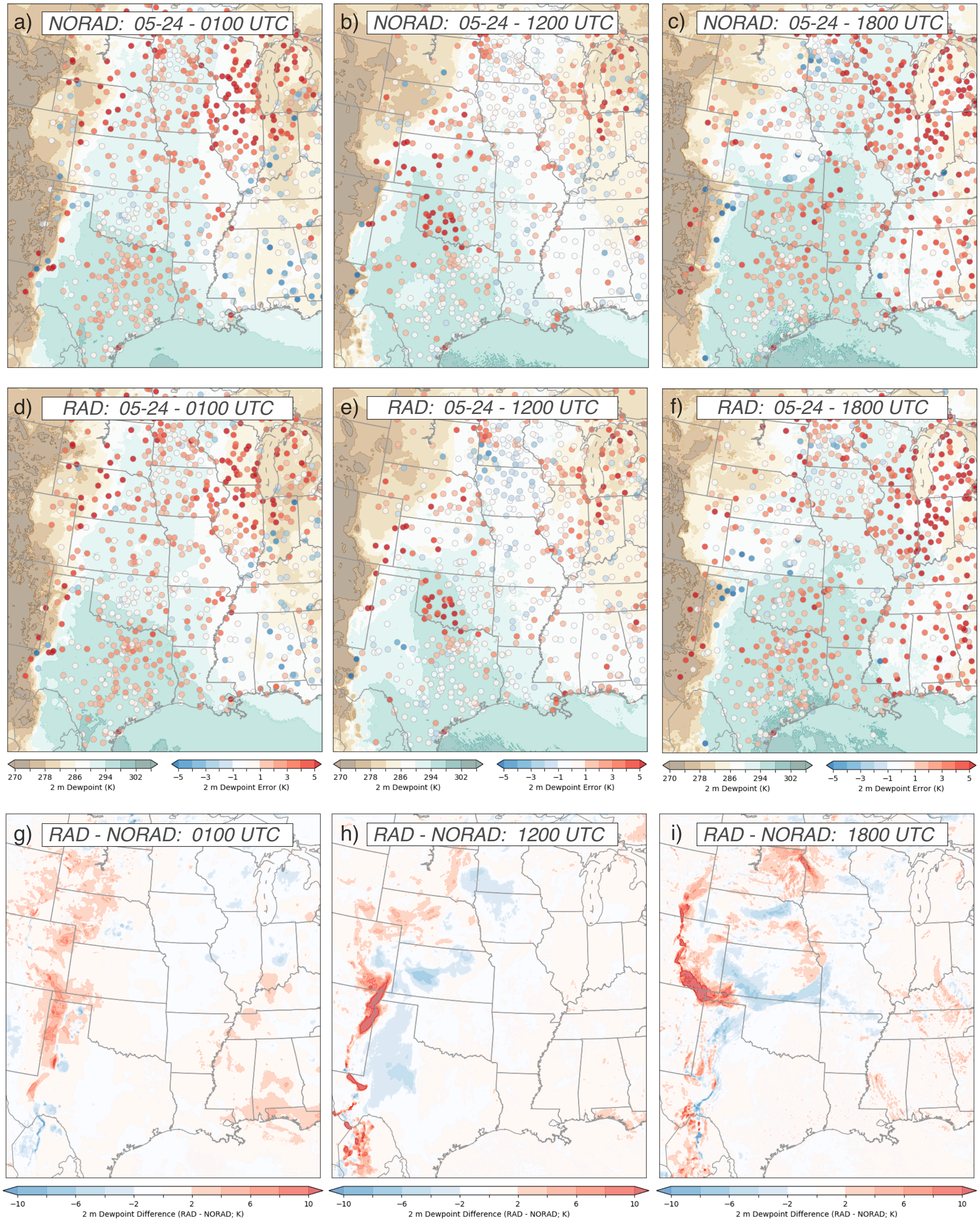

FIG. 11. As in Fig. 10 except for dewpoint temperature.

similation modified the low-level water vapor field (Figs. 11a, d, g). In particular, regions trailing ongoing convection from West Texas northward through the Oklahoma Panhandle, western Kansas, eastern Colorado, and western, Nebraska switch from mixed to modest negative errors in the NORAD member to consistently positive errors in the RAD member, suggesting that 3DVAR radar assimilation and cloud analysis has introduced a positive dewpoint bias.

The temperature and dewpoint differences associated with radar assimilation combine with differences in the forecasted evolution of the nocturnal MCS in Nebraska and Kansas (Fig. 7) to influence near-surface thermodynamic errors later in the forecast period. For example, higher dewpoint values are present in the RAD member along and behind the dryline through the forecast period (Fig. 11). While the strongest differences (>10 K higher in RAD) are attributable to differences in dryline position, broader areas of higher dewpoints in the RAD member, particularly in eastern Colorado, appear partially attributable to moisture introduced by data assimilation. Additionally, generally cooler temperatures persist in the RAD member near, and downstream, of locations of convective cold pools in the 1-hour forecast (Fig. 10).

Differences between the RAD and NORAD members later in the forecast period are clearly influenced by the evolution of the nocturnal MCS. The propagation of the MCS cold pool farther to the southwest in the RAD member results in a band of cooler temperatures (in places more than 5 K less than NORAD) and lower dewpoints from northwest Oklahoma through western Kansas and into eastern Colorado in the 18-hour forecast (Figs. 10c, f, 11c, f). This region represents the pre-convective environment of most severe thunderstorms later in the afternoon and the cool and dry biases in the RAD member contribute to lower CAPE and a lack of predicted storms (Figs. 7, 8). On the other hand, the more expansive outflow in the RAD member results in cooler, and generally dryer, forecasts from Missouri through the upper Mississippi River Valley. These differences produce generally smaller errors compared to ASOS observations through the region, particularly in Missouri, than the NORAD member and a smaller overprediction of thunderstorm development than NORAD (Fig. 7).

4. Discussion

Several differences in the short-term and next-day thunderstorm forecasts produced by the RAD and NORAD ensembles in 2016 have been identified. Many aspects of these differences are consistent with past studies. Specifically, more accurate model initial conditions provided by radar assimilation produce improvements in thunderstorm forecasts that typically persist for 3-6 hours (Figs. 4–6; Kain et al. 2010; Craig et al. 2012; Stratman et al. 2013; Keil et al. 2014; Surcel et al. 2016). Additionally, the length of improved skill in short-term forecasts that assimilate radar data varies from case-to-case (Fig. 6) similarly to the findings of Craig et al. (2012) and Keil et al. (2014). However, in contrast to some past studies (Kain et al. 2010; Stratman et al. 2013), large differences are found between RAD and NORAD next-day thunderstorm forecast skill (Figs. 4–6), thunderstorm coverage (Figs. 7, 9), and convective environment (Figs. 8-11).

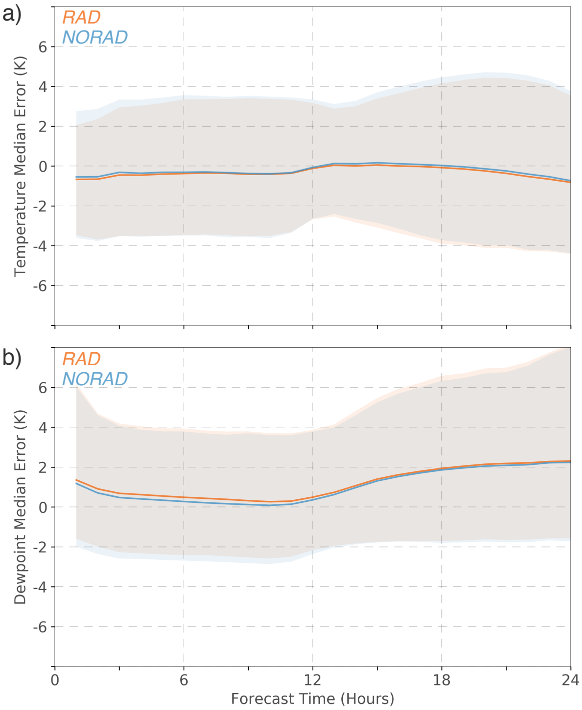

FIG. 12. Hourly median median error of the (a) 2 m temperature (K) and (b) 2 m dewpoint temperature (K) for the first member of the RAD and NORAD ensembles and corresponding ASOS observations. Shading encompasses the 10th–90th percentiles of the error distribution.

Though large differences in next-day thunderstorm forecast skill are present, with NORAD producing higher FSS and object-based CSI following forecast hour 6, differences in the skill scores are likely partially attributable to contrasting frequency biases in the RAD and NORAD ensembles. For example, the majority of next-day thunderstorm FSS from both ensembles are considered unskillful according to the criteria of Roberts and Lean (2008) (Fig. 4). Furthermore, higher scores in the NORAD ensemble may be influenced by more random hits to observations produced by larger coverage and, for FSS, potentially by a higher asymptote score resulting from a smaller frequency bias than the RAD ensemble. Similarly, a higher frequency bias in NORAD thunderstorm objects identified by MODE results in both a higher POD and FAR for much of the forecast period (Fig. 5). Finally, clear differences in 24-hour forecast skill were not identified in subjective evaluations by 2016 SFE participants (Clark et al. 2017), which suggests differences in bulk next-day verification metrics may not translate to practical forecast usefulness. Despite these ambiguity in interpreting relative differences in next-day thunderstorm forecast skill, large, contrasting, and consistent biases in thunderstorm coverage are present in both ensembles that degrade the overall accuracy of forecasts.

The cold start NORAD ensemble overpredicts next-day thunderstorm coverage, with object-based frequency biases greater than 1 through much of the 6–30 hour forecast period and the highest biases during the afternoon convective maximum (Fig. 5b). The bias is remarkably consistent between cases, with an overforecast of composite reflectivity objects during forecast hours 21 through 24 present for all 24 cases examined (Fig. 6b). The high thunderstorm coverage bias appears to stem from an overforecast of potential instability in the next-day convective environment, as evidenced by experiment-long mean counts of NORAD CAPE gridpoints exceeding 1000 J Kg^{-1} that are several thousand gridpoints higher than corresponding RAD counts (Fig. 9).

The overprediction biases in the NORAD ensemble are similar to biases produced by model spinup of convective scales. Wong and Skamarock (2017) found that over 24 hours of forecast time was required to spin up convective scales in convection-allowing Model for Prediction Across Scales (MPAS) forecasts during the spring of 2016. The spin up period was characterized by an overprediction of next-day hourly rainfall rates compared to MRMS observations. Additionally, the bias was mitigated if forecasts were initialized using prior day 24-hour forecasts that contained convective-scale motion and hydrometeors. The similarity of the findings of Wong and Skamarock (2017) and the present study suggest that inclusion of convective scales of motion and hydrometeor fields are important for producing unbiased next-day thunderstorm forecasts in CAMs. As upscale error growth from convection influences the mesoscale environment (Zhang et al. 2003, 2007), particularly in regions of widespread convection (Nielsen and Schumacher 2016), inclusion of convective scales in forecast initial conditions have the potential to improve predictions of convectively-induced features such as MCVs and cold pools. Improved prediction of these features would then be expected to improve next-day forecasts of the convective environment and thunderstorm evolution (Stensrud et al. 1999; Liu and Xue 2018; Clark et al. 2010; Schumacher and Clark 2014; Thompson 2014; Degelia et al. 2018).

Despite the apparent potential for improving next-day CAM forecasts by including convective scales, RAD forecasts illustrate difficulties in realizing this potential. The RAD ensemble produces a low bias of thunderstorm coverage, consistent across most individual cases, through the entire forecast period (Figs. 5b, 6b). As with the overprediction bias of thunderstorm coverage in NORAD, this low bias appears attributable in part to the available potential instability (Figs. 8, 9). In the CLUE, 3DVAR radar assimilation and cloud analysis introduces convective cold pools and hydrometeor fields to the initial conditions, which largely results in negative (positive) increments to the near-surface temperature (dewpoint) fields (Figs. 10, 11). Errors in these data assimilation increments can introduce biases into model initial conditions, which then degrade short-term forecasts of thunderstorm evolution (Fig. 7) and subsequent development of features important to the next-day convective environment and evolution. The potential for short-term errors introduced by radar assimilation to degrade the next-day forecast was identified by 2016 SFE participants (Gallo et al. 2017) and the RAD forecast from 24 May provides one example. On 24 May, a short-term overforecast of convective coverage and introduction of errors related to cold pools leads to a more expansive nocturnal MCS. Outflow from this MCS propagates farther southwestward than in the NORAD ensemble and results in reduced CAPE and limited convection initiation in the region with the highest concentration of severe weather reports in the 18–24 hour forecast period.

Realizing potential gains in next-day thunderstorm forecasts from assimilation of radar data likely requires improvements to two aspects of convective-allowing ensemble forecast systems: 1) Improving the accuracy of thunderstorm scales in the initial condition and resulting short-term forecasts and 2) reduction of errors associated with model physics. Convective-scale radar data assimilation is an active area of research and indirect (Benjamin and Coauthors 2016), variational (Gao et al. 2004), ensemble-based (Wheatley et al. 2015), or hybrid (Gao and Stensrud 2014) assimilation techniques have been developed for real-time forecast systems. Comparison of different techniques is beyond the scope of this paper and is being addressed in other CLUE experiments (Clark et al. 2018); however, examination of RAD forecasts in the 2016 CLUE suggest that assimilation techniques that provide both accurate initializations of convective scales and reduce biases in the mesoscale environment have the highest potential for improving both short-term and next-day thunderstorm prediction. Similarly, extensive research has documented next-day forecast sensitivity to physical parameterizations in CAMS, particularly the planetary boundary layer (e.g. Clark et al. 2015; Cohen et al. 2017) and microphysical (e.g. (Yussouf et al. 2013; Clark et al. 2014; Wheatley et al. 2014) parameterizations. Improved representations of physical processes important to thunderstorm evolution, particularly those related to cold pool development and propagation (Wheatley et al. 2014), offer similar potential for improving short-term predictions of convection and resulting impacts on the next-day convective environment.

5. Conclusions

This study has examined the impact of 3DVAR radar assimilation and cloud analysis (Xue et al. 2003; Gao et al. 2004; Hu et al. 2006) on short-term and next-day ensemble forecasts of thunderstorms produced by the Community Leveraged Unified Ensemble over 24 days during the spring of 2016. The skill of CLUE composite reflectivity forecasts in two 10-member sub-ensembles, one which assimilates radar data and one that does not, is assessed using neighborhood and object-based verification against MRMS composite reflectivity observations. Additionally, the convective environment in the two ensembles is compared and a case that produced extensive severe weather on 24 May 2016 is subjectively interpreted in order to examine differences in forecast evolution between the two ensembles.

It is found that large differences in thunderstorm coverage are present between the two ensembles during the entirety of the 36-hour forecast period (Figs. 4–7, 9). These differences result in contrasting frequency biases in the next-day RAD and NORAD composite reflectivity forecasts compared to a matched-percentile threshold (Figs. 2, 3) in MRMS observations, with NORAD producing an overforecast and RAD an underforecast. These biases are influenced by differences in the thunderstorm environment, as larger areas of significant CAPE (i.e. >1000 J Kg^{-1}) are present in NORAD forecasts (Figs. 8, 9). Because RAD and NORAD forecasts are identical except for changes in the initial conditions provided by radar assimilation, the consistent, contrasting biases between the two ensembles demonstrate that inclusion of convective-scale motion in model initial conditions (e.g. hot vs. cold start) can alter resulting forecasts beyond the intrinsic predictability limit of convective scales. The forecast evolution on 24 May provides a specific example of this process, as radar assimilation results in local biases to near-surface temerature and dewpoint fields (Figs. 10, 11) and a short-term overforecast in convective coverage (Fig. 7). These short-term differences in the RAD ensemble result in more extensive convective cold pools and a southwestward displacement of an outflow boundary responsible for next-day convection initiation, which then limits thunderstorm coverage in the primary region of severe weather (Figs. 7, 8).

Comparison of Fractions Skill Score and MODE-based contingency table metrics between RAD and NORAD indicate the duration of improvements in composite reflectivity forecasts from radar assimilation is generally between 3–6 hours (Figs. 4–6), which is similar to prior studies (Kain et al. 2010; Craig et al. 2012; Stratman et al. 2013; Keil et al. 2014; Moser et al. 2015; Surcel et al. 2016). Both verification methods then indicate more skillful thunderstorm in the NORAD ensemble for all forecast times beyond 6 hours. However, this result should be interpreted with caution as most next-day forecasts from both ensembles produce unskillful FSSs and high frequency biases in NORAD composite reflectivity forecasts may result in more random matches with observed values. Therefore, the relative values of next-day verification scores are believed to be of lower relevance for improving operational CAM forecasts than the large biases present in both ensembles. In other words, both ensembles produce biased, inaccurate forecasts of next-day thunderstorm coverage.

The relationship between short-term thunderstorm evolution and the next-day convective environment and thunderstorm evolution suggest potential for improving CAM forecasts. As an idealized example, radar assimilation which both accurately initializes thunderstorms within the CAM and improves analyses of the convective environment can improve short-term forecasts of convection. Reduction of errors in short-term forecasts, coupled with accurate parameterization of physical processes in the CAM, would then limit biases in the next-day convective environment and improve prediction of features such as MCVs and cold pools that influence next-day thunderstorm development.

Acknowledgments. Support for this study was provided by the Developmental Testbed Center Visitor Program, which is funded by the National Atmospheric and Oceanic Administration, National Center for Atmospheric Research, and National Science Foundation. Jamie Wolff, Tara Jensen, John Halley Gotway, Randy Bullock, and the staff of the DTC are thanked for their assistance and guidance during the course of this research. Additionally, Adam Clark and Ming Xue are thanked for running and providing CLUE output and Anthony Reinhart for providing MRMS observations. Finally, this research benefitted from many helpful conversations with Derek Stratman and Corey Potvin. All verification statistics in this study were produced using NCAR's Model Evaluation Tools and all visualizations with the freely-provided Anaconda Python distribution.

References

Bartels, D. L., and R. A. Maddox, 1991: Midlevel cyclonic vortices generated by mesoscale convective systems. *Mon. Wea. Rev.*, **119**, 104–118.

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694.

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

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

Carbone, R. E., J. D. Tuttle, D. A. Ahijevych, and S. B. Trier, 2002: Inferences of predictability associated with warm season precipitation episodes. *J. Atmos. Sci.*, **59**, 2033–2056.

Casati, B., 2010: New developments of the intensity-scale technique within the Spatial Verification Methods Intercomparison Project. *Wea. Forecasting*, **25**, 113–143.

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

Clark, A. J., M. C. Coniglio, B. E. Coffer, G. Thompson, M. Xue, and F. Kong, 2015: Sensitivity of 24-h forecast dryline position and structure to boundary layer parameterizations in convection-allowing WRF model simulations. *Wea. Forecasting*, **30**, 613–638.

Clark, A. J., W. A. Gallus, M. Xue, and F. Kong, 2010: Convection-allowing and convection-parameterizing ensemble forecasts of a Mesoscale Convective Vortex and associated severe weather environment. *Wea. Forecasting*, **25**, 1052–1081.

Clark, A. J., and Coauthors, 2017: Spring forecasting experiment 2016. Tech. rep., National Severe Storms Laboratory, 50 pp.

Clark, A. J., and Coauthors, 2018: The Community Leveraged Unified Ensemble (CLUE) in the 2016 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Bull. Amer. Meteor. Soc.*, **99**, 1433–1448.

Cohen, A. E., S. M. Cavallo, M. C. Coniglio, H. E. Brooks, and I. L. Jirak, 2017: Evaluation of multiple planetary boundary layer parameterization schemes in Southeast U.S. cold season severe thunderstorm environments. *Wea. Forecasting*, **32**, 1857–1884.

Craig, G. C., C. Keil, and D. Leuenberger, 2012: Constraints on the impact of radar rainfall data assimilation on forecasts of cumulus convection. *Quart. J. Roy. Meteor. Soc.*, **138**, 340–352.

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

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

Dawson, L. C., G. S. Romine, R. J. Trapp, and M. E. Baldwin, 2017: Verifying supercellular rotation in a convection-permitting ensemble forecasting system with radar-derived rotation track data. *Wea. Forecasting*, **32**, 781–795.

Degelia, S. K., X. Wang, D. J. Stensrud, and A. Johnson, 2018: Understanding the impact of radar and in situ observations on the prediction of a nocturnal convection initiation event on 25 June 2013 using an ensemble-based multiscale data assimilation system. *Mon. Wea. Rev.*, **146**, 1837–1859.

Durran, D. R., and J. A. Weyn, 2016: Thunderstorms do not get butterflies. *Bull. Amer. Meteor. Soc.*, **97**, 237–243.

Ek, M., K. E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, and J. D. Tarpley, 2003: Implementation of Noah land-surface model advances in the NCEP operational mesoscale eta model. *J. Geophys. Res.*, **108**, 8851.

Gallo, B. T., and Coauthors, 2017: Breaking new ground in severe weather prediction: The 2015 NOAA/Hazardous Weather Testbed spring forecasting experiment. *Wea. Forecasting*, **32**, 1541–1568.

Gao, J., and D. J. Stensrud, 2014: Some observing system simulation experiments with a hybrid 3DEnVAR system for storm-scale radar data assimilation. *Mon. Wea. Rev.*, **142**, 3326–3346.

Gao, J., M. Xue, K. Brewster, and K. K. Droegemeier, 2004: A three-dimensional variational data analysis method with recursive filter for Doppler radars. *J. Atmos. Oceanic Technol.*, **21**, 457–469.

Griffin, S. M., J. A. Otkin, C. M. Rozoff, J. M. Sieglaff, L. M. Cronce, and C. R. Alexander, 2017a: Methods for comparing simulated and observed satellite infrared brightness temperatures and what do they tell us? *Wea. Forecasting*, **32**, 5–25.

Griffin, S. M., J. A. Otkin, C. M. Rozoff, J. M. Sieglaff, L. M. Cronce, C. R. Alexander, T. L. Jensen, and J. K. Wolff, 2017b: Seasonal analysis of cloud objects in the High-Resolution Rapid Refresh (HRRR model using object-based verification. *J. Appl. Meteor. and Climatology*, **56**, 2317–2334.

Hamill, T. M., 1999: Hypothesis test for evaluating numerical prediction forecasts. *Wea. Forecasting*, **14**, 155–167.

Hu, M., M. Xue, and K. Brewster, 2006: 3DVAR and cloud analysis with WSR-88D Level II data for the prediction of the Fort Worth, Texas, tornadic thunderstorms. Part I: Cloud analysis and its impact. *Mon. Wea. Rev.*, **134**, 675–698.

Johns, R. H., and C. A. Doswell, 1992: Severe local storms forecasting. *Wea. Forecasting*, **7**, 588–612.

Kain, J. S., P. R. Janish, S. J. Weiss, R. S. Schneider, M. E. Baldwin, and H. E. Brooks, 2003: Collaboration betweeen forecasters and research scientists at the NSSL and SPC: The spring program. *Bull. Amer. Meteor. Soc.*, **84**, 1797–1806.

Kain, J. S., and Coauthors, 2010: Assessing advances in the assimilation of radar data within a collaborative forecasting-research environment. *Wea. Forecasting*, **25**, 1510–1521.

Keil, C., F. Heinlein, and G. C. Craig, 2014: The convective adjustment time-scale as indicator of predictability of convective precipitation. *Quart. J. Roy. Meteor. Soc.*, **140**, 480–490.

Lakshmanan, V., C. Karstens, J. Krause, and L. Tang, 2014: Quality control of weather radar data using polarimetric variables. *J. Atmos. Oceanic Technol.*, **31**, 1234–1249.

Liu, H., and M. Xue, 2018: Prediction of convective initiation and storm evolution on 12 June 2002 during IHOP_2002. Part I: Control simulation and sensitivity experiments. *Mon. Wea. Rev.*, **136**, 2261–2282.

Mellor, G. L., and T. Yamada, 1982: Development of a turbulence closure model for geophysical fluid problems. *Rev. Geophys.*, **20**, 851–875.

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

Moser, B. A., W. A. Gallus, and R. Mantilla, 2015: An initial assessment of radar data assimilation on warm season rainfall forecasts for use in hydrologic models. *Wea. Forecasting*, **30**, 1491–1520.

Nielsen, E. R., and R. S. Schumacher, 2016: Using convection-allowing ensembles to understand the predictability of an extreme rainfall event. *Mon. Wea. Rev.*, **144**, 3651–3676.

Pinto, J. O., J. A. Grim, and M. Steiner, 2015: Assessment of the High-Resolution Rapid Refresh model's ability to predict mesoscale convective systems using object-based evaluation. *Wea. Forecasting.*, **30**, 892–913.

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.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting.*, **24**, 601–608.

Schumacher, R. S., and A. J. Clark, 2014: Evaluation of ensemble configurations for the analysis and prediction of heavy-rain-producing mesoscale convective systems. *Mon. Wea. Rev.*, **142**, 4108–4138.

Skamarock, W. C., and Coauthors, 2008: A description of the advanced research wrf version 3. ATC TN-475+STR, National Center for Atmospheric Research, 113 pp.

Skinner, P. S., and Coauthors, 2018: Object-based verification of a prototype Warn-on-Forecast system. *Wea. Forecasting*, **33**, In Press.

Smith, T. M., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS severe weather and aviation products: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 1617–1630.

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

Stensrud, D. J., G. S. Manikin, E. Rogers, and K. E. Mitchell, 1999: Importance of cold pools to NCEP mesoscale Eta Model forecasts. *Wea. Forecasting*, **14**, 650–670.

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

Stratman, D. R., M. C. Coniglio, S. E. Koch, and M. Xue, 2013: Use of multiple verification methods to evaluate forecasts of convection from hot- and cold-start convection-allowing models. *Wea. Forecasting*, **28**, 119–138.

Surcel, M., I. Zawadzki, and M. K. Yau, 2016: The case-to-case variability of the predictability of precipitation by a storm-scale ensemble forecasting system. *Mon. Wea. Rev.*, **144**, 193–212.

Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115.

Thompson, T., 2014: Ensemble Kalman filter methods for convective-scale radar data assimilation and multi-scale data assimilation of the 13 June 2010 tornadic supercell environment. Ph.D. thesis, University of Oklahoma, 233 pp.

Weckwerth, T. M., and D. B. Parsons, 2006: A review of convection initiation and motivation for IHOP_2002. *Mon. Wea. Rev.*, **134**, 5–22.

Weinhoff, Z. B., H. B. Bluestein, L. J. Wicker, J. C. Snyder, A. Shapiro, C. K. Potvin, J. B. Houser, and D. W. Reif, 2018: Applications of a spatially variable advection correction technique for temporal correction of dual-doppler analyses of tornadic supercells. *Mon. Wea. Rev.*, **146**, In Press.

Weyn, J. A., and D. R. Durran, 2017: The dependence of the predictability of mesoscale convective systems on the horizontal scale and amplitude of initial errors in idealized simulations. *J. Atmos. Sci.*, **74**, 2191–2210.

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

Wheatley, D. M., N. Yussouf, and D. J. Stensrud, 2014: Ensemble Kalman filter analyses and forecasts of a severe mesoscale convective system using different choices of microphysics schemes. *Mon. Wea. Rev.*, **142**, 3243–3263.

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

Wong, M., and W. Skamarock, 2017: Precipitation spin-up processes in a global model with a cloud-permitting-scale mesh refinement. *Preprints, 17th Conf. on Mesoscale Processes*, San Diego, CA, Amer. Meteor. Soc., 10.4.

Xue, M., D. Wang, J. Gao, K. Brewster, and K. K. Droegemeier, 2003: The Advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation. *Meteor. Atmos. Phys.*, **82**, 139–170.

Yussouf, N., E. R. Mansell, L. J. Wicker, D. M. Wheatley, and D. J. Stensrud, 2013: The ensemble Kalman filter analyses and forecasts of the 8 May 2003 Oklahoma City tornadic supercell storm using single- and double-moment microphysics schemes. *Mon. Wea. Rev.*, **141**, 3388–3412.

Zhang, F. Q., N. F. Bei, R. Rotunno, C. Snyder, and C. C. Epifanio, 2007: Mesoscale predictability of moist baroclinic waves: Convection-permitting experiments and multistage error growth dynamics. *J. Atmos. Sci.*, **64**, 3579–3594.

Zhang, F. Q., C. Snyder, and R. Rotunno, 2003: Effects of moist convection on mesoscale predictability. *J. Atmos. Sci.*, **60**, 1173–1185.