Use of Multiple Verification Methods to Evaluate Forecasts of Convection from Hot- and Cold-Start Convection-Allowing Models

Derek R. Stratman[1], Michael C. Coniglio[2], Steven E. Koch[2], and Ming Xue[1,3]

[1]*School of Meteorology, University of Oklahoma, Norman OK*
[2]*NOAA/National Severe Storms Laboratory, Norman, OK*
[3]*Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, OK*

Submitted to
Weather and Forecasting
March 2012

*Corresponding author's address:
Derek R. Stratman
National Weather Center, NSSL/FRDD
120 David L. Boren Blvd.,
Norman, OK, 73072
Email: stratman@ou.edu

Abstract

This study uses both traditional and newer verification methods to evaluate two 4-km grid-spacing WRF model forecasts; a "cold start" forecast that uses the 12-km North American Mesoscale (NAM) model analysis and forecast cycle to derive the initial and boundary conditions (C0), and a "hot start" forecast that adds radar data into the initial conditions using a 3DVAR/cloud analysis technique (CN). These forecasts were evaluated as part of 2009 and 2010 NOAA Hazardous Weather Testbed (HWT) Spring Forecasting Experiments. The Spring Forecasting Experiment participants noted that the skill of CN's explicit forecasts of convection estimated by some traditional objective metrics often seemed large compared to the subjectively-determined skill. The Gilbert Skill Score (GSS) reveals CN scores higher than C0 at lower thresholds likely due to CN having higher frequency biases than C0, but the difference is negligible at higher thresholds, where CN and C0's frequency biases are similar. This suggests that if traditional skill scores are used to quantify convective forecasts, then higher (> 35 dBZ) reflectivity thresholds should be used to be consistent with expert's subjective assessments of the lack of forecast skill for individual convective cells. The spatial verification methods show both CN and C0 generally have little to no skill at scales < 8–12Δx starting at forecast hour 1, but CN has more skill at larger spatial scales (40–320 km) than C0 for the majority of the forecasting period. This indicates that the hot start provides little to no benefit for forecasts of convective cells, but has some benefit for larger mesoscale precipitation systems.

1. Introduction

Every Spring, operational forecasters and research scientists participate in the National Oceanic and Atmospheric Administration's (NOAA) Hazardous Weather Testbed (HWT) Spring Forecasting Experiment, which is designed to improve communication and facilitate collaboration among forecasters and researchers through the generation of daily experimental convective forecasts and evaluation of experimental forecast models (Kain et al. 2010; Clark et al. 2012). For the 2009 and 2010 Spring Forecasting Experiments (SFE2009 and SFE2010, respectively), the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma produced ensemble forecasts at 4-km grid-spacing, in near-CONUS (2009) and full-CONUS (2010) domains, using the Weather Research and Forecasting (WRF) model (Xue et al. 2009; 2010). The ensemble forecasts were run once a day, starting from 0000 UTC on week days, and the length of forecasts was 30 hours. Among the ensemble members for both years, two members of interest used the WRF-ARW model; one member directly used the 0000-UTC 12-km NAM (North American Mesoscale model) analyses at the initial condition and the other member used the three-dimensional variational (3DVAR)-cloud analysis (Xue et al. 2003; Hu et al. 2006a, 2006b) initial condition that assimilated radar and other high-resolution observations (from surface stations and wind profilers). The NAM analyses were used as the background. The two runs did not include additional initial condition perturbations and are referred to as two control runs, one with radar data (called CN) and one without radar data (called C0). The comparison between CN and C0 allows the evaluation of the impact of radar and other high-resolution data in the initial condition, with radar data having a dominant effect given its relative data volume. All forecasts used NAM forecasts starting at the same initial times to provide the lateral boundary conditions. In addition to the 0000-UTC ensemble

96 forecasts, CAPS was also producing forecasts over a smaller central U.S. domain, at 1200 UTC
97 and several other times, using model configurations corresponding to those of 0000-UTC CN
98 and C0 (except for the domain size). These runs were made to support the VORTEX2 field
99 experiment (Xue et al. 2009; 2010). In this study, we will evaluate and initially compare the
100 0000- and 1200-UTC CN and C0 forecasts.

101 During 2009 and 2010, participants in the Spring Forecasting Experiment compared
102 hourly loops of CN and C0 simulated reflectivity (SR) forecasts to the observed radar
103 reflectivity (OR) for the same time periods on a large monitor. They were asked to define when
104 the cold start forecasts (C0) appeared to "catch up" with the hot start forecasts (CN) "in terms
105 of its degree of correspondence with reality." For SFE2009, nearly 60% of the participants
106 perceived 0000-UTC CN forecasts to be effectively similar to the C0 forecasts in their
107 depiction of convection after 3 to 6 hours. For example, one participant commented, "by
108 [forecast hours] 3–4 the two model runs tend to look more like each other than like the
109 obs[ervations]." This sentiment is illustrated for an individual case in Fig. 1. At the initial time,
110 CN's SR looked very similar to the OR, which is the result of the reflectivity assimilation in
111 the cloud analysis step. However, by forecast hour 3 and especially by forecast hour 6, the
112 subjective impression of the participants is that both forecasts were equally skillful/unskillful
113 in their forecasts of that convective event.

114 This study aims to complement the subjective assessment of CN's and C0's skill in
115 forecasting convection discussed above by providing a comprehensive objective assessment of
116 its skill. The skill is characterized through traditional metrics, as well as through newer
117 techniques that define model errors by spatial scales and variable thresholds. The latter
118 approach delineates the spatial scales at which CN improves over C0 and provides a more

comprehensive assessment of model skill over what can be provided by a subjective evaluation, or by traditional grid-point-by-grid-point techniques.

The HWT forecasting experiments have shown that subjective evaluations can provide valuable information on the tools that forecasters find useful, but they are limited in defining the specific error characteristics of numerical model forecasts that researchers strive to address. Simple quantitative verification techniques (that compare a forecast of some quantity to an analysis or observation of that quantity at specific points in space and time) have long been used to objectively evaluate model forecasts. As higher resolution numerical models are now used to predict highly discontinuous fields, like convection, there is an increasing need in the research community to use newer verification techniques (Casati et al. 2008; Gilleland et al. 2009; Gilleland et al. 2010). Because traditional grid-point-by-grid-point verification metrics effectively give much weight to the smallest scales allowed by the gridded model forecasts and observations, small deviations (i.e., errors) in the model forecasts from the observations can often cause misrepresentation of the useful model forecast skill (i.e., skill a forecaster would deem useful and appropriate). Newer verification techniques that attempt to better characterize model skill for discontinuous fields can be classified into one of four categories: neighborhood (or fuzzy), scale separation (or decomposition), feature-based (or object-based), and field deformation techniques (Gilleland et al. 2009). This study uses the first two types, which are different ways of evaluating model skill through a filtering of the gridded fields.

Several traditional verification metrics and two filtering techniques (described in Section 3) are used in this study to analyze the performance of CN and C0 in an individual and a comparative sense. A goal of using these spatial-scale filtering methods is to determine if the spatial verification metrics are more consistent with the SFE2009 and SFE2010 participants'

subjective evaluations since the traditional verification scores are often not appropriate measures of skill for high-resolution model forecasts of discontinuous fields (Gillelend et al. 2009), like strong convection. Another goal of this study is to assess the benefit of the CAPS 3DVAR-cloud analysis radar data assimilation technique (described in Section 2), as applied to the CN forecasts.

2. Data Sets

a) CAPS CN and C0 forecasts

As mentioned earlier, the two WRF-ARW (Advanced Research WRF) model runs examined in this study were part of the CAPS 4-km grid spacing Storm Scale Ensemble Forecast (SSEF) system run in the springs of 2009 and 2010 (see Xue et al. 2009, 2010 for specific details). The 2009 version of the SSEF system was comprised of 10 WRF-ARW members, 8 WRF-NMM (Nonhydrostatic Mesoscale Model) members, and 2 ARPS (Advanced Regional Prediction System) members, while the 2010 SSEF system contained 19 WRF-ARW members, 5 WRF-NMM members, and 2 ARPS members. For SFE2009, the 0000-UTC WRF-ARW control members (i.e., CN and C0) were integrated to 30 hours on an eastern near-CONUS size domain (Fig. 2a). For SFE2010, CN and C0 were once again integrated to 30 hours starting at 0000 UTC, but the domain increased to a full CONUS size domain (Fig. 2b). For both SFE2009 and SFE2010, the 1200-UTC CN and C0 members were integrated to 18 hours on the Central Plains domain (Fig. 2a,b).

CN assimilates radar radial velocity with a mass divergence constraint in the 3DVAR procedure to derive the wind components for the initial conditions in combination with the NAM background and additional surface and wind profiler data (Hu et al. 2006b). In addition,

CN uses a cloud analysis scheme, which adds hydrometeors and adjusts the in-cloud temperature and moisture fields through a moist-adiabatic scheme using three dimensional radar reflectivity data as well as surface cloud base and satellite cloud top observations (Xue et al. 2003; Hu et al. 2006a). Except for the initial condition, all other model configurations (i.e., boundary conditions from the NAM fields, Thompson cloud microphysics scheme, Goddard short-wave radiation physics, RRTM long-wave radiation physics, Noah land-surface model, and Mellor-Yamada-Janjic planetary boundary layer physics) are identical between CN and C0 (Xue et al. 2009; 2010).

The 0000- and 1200-UTC model runs are examined separately. Combining both years of data yielded a maximum 56 days of data for the 0000-UTC model runs and 77 days of data for the 1200-UTC model runs (the 1200-UTC forecasts were run on weekends also). These data sets are used to evaluate and compare the models' hourly SR and 1-h accumulated precipitation (APCP) fields.

b) Verification data and domain

Composite reflectivity and quantitative precipitation estimates calculated on a 1-km grid as part of the National Severe Storms Laboratory (NSSL) National 3-D Reflectivity Mosaic system are used for the verifying observations (see Vasiloff et al. 2007 for more details). Even though the 0000-UTC forecasts were performed on large CONUS-sized domains, this study focuses on the Central Great Plains region that was the focus of the VORTEX2 field experiment during the two spring seasons. Given that the wavelet scale-separation method used in this study requires domains to be $2^n \times 2^n$ grid points in size (see Section 3c) and given the 4-km grid spacing of the model forecasts, a reasonably sized verification domain was chosen to

188 be made up of 256 × 256 grid points in the horizontal (n = 8) given the smaller size of the
189 1200-UTC domain. Because the model forecasts and verification fields were on different
190 native grids, prior to verification, the fields were interpolated to a 256 × 256 portion of the
191 AWIPS #240 grid (following Schwartz et al. 2009), which has a horizontal grid-spacing of
192 4.7625 km.

193 Due to its small size relative to the model domains, the verification domain was
194 moveable[1] (Fig. 2) to follow areas where observed convection occurred. If no convection
195 occurred on a particular day, the domain was centered on Norman, OK. In addition, the
196 western edge of the domain always had a longitude of 105°W, so the domain moved north and
197 south to follow active areas of convection over the Central U.S.
198

199 **3. Verification Metrics**

200 *a) Traditional scores*

201 Using thresholds of 10, 20, 30, and 40 dBZ for SR and 0.1, 1.0, 5.0, and 10.0 mm/hr for
202 APCP, standard 2 × 2 contingency table components (i.e., hits, false alarms, misses, and
203 correct negatives) were generated hourly for forecast hours 0 through 12 for each model run
204 using the Model Evaluation Tools (MET) Version 3.0, which was developed and is currently
205 maintained by NOAA/NCAR (National Center for Atmospheric Research) through the
206 Development Testbed Center (DTC) (DTC 2011). All of the individual components of the
207 contingency tables for each model/threshold combination were summed using MET to form
208 contingency tables for each forecast hour. From these summed contingency tables, four
209 traditional metrics were computed: frequency bias (FBIAS), probability of detection (POD),

[1] It's worth noting as a caveat that the climatology varies as the verification domain is moved from location to location (Hamill and Juras 2006).

probability of false detection (POFD), and Gilbert skill score (GSS) (otherwise known as the Equitable Threat Score, or ETS). 95% confidence intervals (CI) were included for each metric for each forecast hour to assess the uncertainty in the estimates following the resampling procedure of Hamill (1999). The CIs are assigned in a comparative sense – the uncertainty in the difference in the metrics between the two model forecasts in question is assessed by computing CIs on the metric differences at each forecast hour. If the CIs on the difference estimates include the zero line for a particular forecast hour, the differences of the verification metrics are said to be not significantly different, and vice versa.

b) Neighborhood method

Using the neighborhood method based on Roberts and Lean (2008), the fractions skill score (FSS) is computed to assess the skill for different neighborhood sizes and variable thresholds. The neighborhood method allows for a "hit" to be within a certain neighborhood (radius) of the observation, which allows for forecasts that are "close enough" to be considered skillful in the objective metrics (Ebert 2008). FSS ranges from 0 for a no skill to 1 for perfect skill as given by:

$$FSS = 1 - \frac{\frac{1}{N}\sum_N \left(P_{fcst} - P_{obs}\right)^2}{\frac{1}{N}\left(\sum_N P_{fcst}^2 + \sum_N P_{obs}^2\right)}, \qquad (1)$$

where P_{fcst} and P_{obs} are the fractional forecast and observed SR (or APCP) areas in each neighborhood that exceed the specified variable threshold and N is the number of neighborhoods for each neighborhood size. (Note: larger neighborhood sizes lead to a smaller number of neighborhoods, whereas smaller sizes will result in a larger N.) In an evaluation of precipitation forecasts from convection-allowing models, Roberts and Lean (2008) estimated that forecasts have useful skill (FSS_{useful}) when $FSS_{useful} = 0.5 + f_o/2$, where f_o is the base rate,

or fraction of observed events to all grid points. They consider this value to be a reasonable "target skill" since it is halfway between random forecast skill and perfect skill. This same value for FSS_{useful} is used in this study, and forecasts for which FSS > FSS_{useful} are considered to have useful skill.

An aggregated FSS[2] was computed for each forecast hour for neighborhood widths, *n*, of 1, 3, 5, 9, 17, 33, and 65 grid points centered at the grid box in question. For each neighborhood width, FSS was calculated for different SR and APCP thresholds (i.e., 10, 15, 20, 25, 30, 35, 40, 45, and 50 dBZ and 0.1, 0.5, 1.0, 2.5, 5.0, 7.5, 10.0, 20.0, and 40.0 mm/hr, respectively). The aggregated FSS–FSS_{useful} is displayed in a matrix of neighborhood size versus variable threshold for a particular forecast hour for the individual neighborhood size and threshold combinations. Whenever FSS–FSS_{useful} is positive, the forecast is considered to have useful skill. In addition, similar plots are shown for the differences in FSS between the models along with 95% confidence intervals, which were computed again following the procedure of Hamill (1999).

c) Scale separation method

Like the neighborhood method, scale separation methods allow for non-overlapping forecasts and observations to be considered skillful in the objective metrics, but have the additional benefit of assessing the skill at individual, independent spatial scales of the errors (Casati et al. 2004). The particular intensity-scale verification (ISV) technique employed in this study is based on Casati et al. (2004), which isolates the skill at scales given by $2^l \times 2^l$ for $l = 0$,

[2] For each forecast hour, the data sets were aggregated together for each combination of neighborhood width and threshold. Knowing the FBS, FSS, and N values, the summations in the numerator and denominator of Eqn. 1 were calculated and aggregated separately for the individual data sets. The aggregated summations were then used to calculate FSS.

1, 2,...8, where l = 0 represents the horizontal spacing of one grid cell (4.7625 km) and l = 8 represents the entire verification domain (1219 × 1219 km). This study retains the biases in the forecasts, as in Casati (2010), to also assess the bias associated with the various spatial scales and thresholds.

The first step is to transform forecast and observation fields into binary fields based on variable thresholds. The same variable thresholds used for the neighborhood method were also used for this method. A 2-D Haar wavelet decomposition is then performed on the binary difference images between the forecasts and observations (i.e., the binary difference images are decomposed into scale components in this step; see Figure 3 for an example of a binary difference field and its first five scale components). Next, a mean squared error (MSE) of the scale components of the binary field difference is calculated for each threshold and scale component. Because the sum of MSE from the individual scale components is equal to the MSE of the binary difference field image, the errors of individual scales can be examined separately (Casati et al. 2004).

To generate a baseline skill, a random MSE term is defined using the equation

$$MSE_{random} = FBIAS * BR * (1 - BR) + BR * (1 - FBIAS * BR), \quad (2)$$

where FBIAS is the frequency bias and BR is the base rate (Casati 2010). A skill score is defined for each binary forecast and observation scale component, called the intensity scale skill (ISS) score, and is given by:

$$ISS = 1 - \frac{MSE}{MSE_{random}/(L+1)}, \quad (3)$$

where L = 8 for this study. Positive ISS values are associated with skillful forecasts and negative ISS values are associated with forecasts with no skill (Casati 2010). Typically, large weather features, such as frontal convection, are fairly well forecasted by convection-allowing

models, so the larger spatial scales tend to exhibit positive skill. Conversely, small-scale weather features, such as individual convective cells, are not usually forecast well by the same convection-allowing models due to their general inability to resolve features less than ~4–8Δx and the generally faster error growth at shorter wavelengths, so the smaller spatial scales tend to exhibit little to no forecasting skill. Plots of the ISS scores were created similar to the plots of the FSS values with thresholded values on the abscissa and the spatial scale on the ordinate, corresponding to l = 0, 1, 2, 3, 4, 5, and 6 in the wavelet transform application. In addition, plots of the difference (and their statistical significance) in ISS between models were created to assess the models' differences.

Finally, the energy was assessed for both the forecast and observation for each scale component and threshold through the evaluation of the energy squared (En2) quantities (Casati 2010). For variable X, En2 is given by

$$En2(X) = \frac{1}{N}\sum_{i=1}^{N} X_i^2, \qquad (4)$$

where N is the total number of wavelet cells in the domain and X_i is the average of the grid point squared values in the i^{th} grid cell. For high-resolution NWP models, in general, high energy is associated with small thresholds because of the relatively large number of events exceeding the threshold value, and conversely, low energy is associated with large thresholds because of the relatively small number of events exceeding the threshold value (Casati 2010). The bias is then assessed by comparing the En2(F) and En2(O) values with each other for every threshold and spatial scale by computing the energy (squared) relative difference (ERD):

$$ERD = \frac{[En2(F) - En2(O)]}{[En2(F) + En2(O)]}. \qquad (5)$$

298 The ERD values range from -1 to 1. Positive ERD values indicate overforecasting (a high bias),
299 and negative ERD values indicate underforecasting (a low bias) (Casati 2010).
300
301 **4. Results**
302 *a) Traditional metrics*
303 According to the GSS, the 0000-UTC CN scored better than C0 for all forecast hours at
304 the 20-dBZ threshold at the 5% significance level (Fig. 4a). However, while significance exists
305 at the 5% level for all forecast hours, the lower bound of the 95% confidence interval is close
306 to zero beyond forecast hour (FH hereafter) 5, so no strong conclusions can be made at those
307 hours. Similar conclusions can be drawn from looking at the GSS derived from the APCP field
308 for the 1.0-mm threshold (Fig. 4c). This similarity between the SR and APCP verification
309 scores of comparable thresholds was a common theme for all of the results found in this study,
310 so only the verification scores for the SR fields are shown hereafter to eliminate redundancy.
311 At the 40-dBZ threshold (Fig. 4b), CN's GSS values remain above those of C0, but the
312 difference quickly decreases in the first 2 hours; the differences are barely significant at the 5%
313 level from FH 2–4 and lack significance beyond FH 4. This indicates a much more rapid drop
314 in relative skill between the two models for the higher threshold. In addition, the scores
315 themselves are not much better than what could be achieved at random with scores that drop
316 and remain below a GSS of 0.1 at the first forecast hour for CN. These results indicate that the
317 model is having a hard time accurately evolving, usually small-scale, high reflectivity cores
318 that are initialized from the radar reflectivity observations. Both inadequate resolution and less
319 than optimal analysis of the intense convection in the initial condition coupled with intrinsic
320 faster error growth at small scales can cause such a fast drop in skill score.

At the 20-dBZ threshold for the 1200-UTC model runs (Fig. 5a), the differences in GSS between CN and C0 are somewhat smaller than they are for the 0000-UTC runs – they become statistically insignificant at the 95% confidence level around FH 4–5. The smaller differences in scores between CN and C0 for the 1200-UTC runs may be related to the diurnal cycle of convection. Convection is typically more abundant at 0000 UTC than at 1200 UTC, so CN has an initial benefit of assimilating more radar data into the initial condition than is assimilated at 1200 UTC. Furthermore, the areal coverage of convection in the spring and summer tends to peak after 0000 UTC in the Plains and tends to be much less in the 1200–1800 UTC period (Wallace 1975; Easterling and Robinson 1985). GSS (and other traditional scores) is dependent on the base rate (i.e., higher base rate leads to larger GSS and lower base rate leads to smaller GSS) (Stephenson et al. 2008), so GSS will be larger for the forecasts with more observed convection. For the remainder of the analysis, the results for the 1200-UTC runs are qualitatively similar to those detailed for the 0000-UTC runs, so only the results for the 0000-UTC runs are shown for brevity.

The forecast hour at which the GSSs for CN and C0 converge drops from about FH 6–12 for the 20-dBZ threshold to about FH 2–3 for the 40-dBZ threshold for both the 0000- and 1200-UTC runs (Figs. 4b and 5b). This convergence of GSS for the higher thresholds generally agrees with the perceptions of the Spring Forecasting Experiment participants, who thought CN and C0 were usually equally skillful between about FH 3 and FH 6 upon an hourly visual inspection of the SR fields. A possible reason for this sentiment might have to do with color psychology. For example, humans perceive some objects that are yellow and red ($\geq$ 35-dBZ objects) to be dangerous (i.e., stop signs and red lights), while objects that are green and blue ($<$ 35-dBZ objects) are perceived as not dangerous (Elliot and Maier 2007; Litchtenfield et al.

344 2009). Hence, the Spring Forecasting Experiment participants' eyes may have focused on
345 reflectivities greater than 35 dBZ in the Spring Forecasting Experiment displays (see later Fig.
346 17 for an example of the standard reflectivity color bar), and therefore gave the more intense
347 convection greater weight than the larger areas of lighter precipitation in their subjective
348 assessment of skill.

349 The comparison of the objective scores with the subjective evaluations suggests that the
350 use of GSS at higher thresholds are preferred over the use of GSS at lower SR thresholds in an
351 evaluation of model forecasts of convection because the conclusions based on the GSS values
352 at higher thresholds are more consistent with the subjective conclusions than the conclusions
353 based on the GSS values at lower thresholds. The GSS values for CN and C0 at these higher
354 reflectivity thresholds are relatively low compared to typical GSS values of other model fields
355 on coarser grids (likely due to the dependency of the GSS on the base-rate), suggesting the
356 model forecasts have relatively little forecasting skill for stronger convection. However, a more
357 general point to be made is that the *relative* importance of CN and C0, as measured by the GSS,
358 is highly dependent on the chosen reflectivity (or precipitation) threshold. Therefore, the use of
359 a single metric like the GSS at any threshold can easily lead to a misrepresentation of model
360 performance for convection-allowing models (Doswell et al. 1990).

361 Another problem with the GSS arises due to high frequency biases[3]. In Fig. 6a, the
362 frequency bias of the 0000-UTC initialized CN approaches a value of 2 by FH 2 and remains
363 above 1.5 for the rest of the forecast period. For rare events, many traditional grid-point-by-
364 grid-point scores, such as the GSS, are maximized for frequency biases > 1 since these scores
365 are more sensitive to missed events than false alarms (Baldwin and Kain 2006). In other words,

[3] Frequency biases are highly dependent on what microphysical scheme is used in a model, so it should be noted that these bias results and their effect on verification metrics are specific only to these models (i.e., different findings might result if different cloud microphysics schemes are used).

366 CN's high bias likely resulted in the noticeable improvement in the GSS for CN versus C0 for
367 the 20-dBZ threshold. The differences in FBIAS between CN and C0 lead to higher POFD for
368 CN for several hours, particularly for the 20-dBZ threshold (Figs. 6 and 7). The fact that CN
369 has a higher FBIAS (and higher POFD) than C0 for the first few hours is not surprising since
370 C0 is spinning up convection. However, the higher 20-dBZ FBIAS for CN persists through
371 about FH 8 for the 0000-UTC runs (Fig. 6a), so the differences in FBIAS, and the effects[4] on
372 the GSS, appear to linger after C0 spins up convection. This relationship in FBIAS and POFD
373 between CN and C0 also is seen at the 40-dBZ threshold (Figs. 6b-7b), although the FBIAS
374 values and their differences are not as large.
375
376 *b) Neighborhood method results*
377 The previous section shows that a wide range of conclusions can be drawn about a
378 model's performance when using traditional grid-point-by-grid-point metrics computed from 2
379 × 2 contingency tables on explicit model forecasts of convection, depending on which
380 threshold and metric are used for evaluation. Although some scores may under penalize model
381 forecasts if model biases are not accounted for, the lack of overlay of forecasts and
382 observations at the grid scale for highly discontinuous fields can over penalize the forecasts
383 and misrepresent the skill and usefulness of the forecasts, which may be the case for the 40-
384 dBZ threshold GSS scores shown earlier. For the neighborhood metrics that attempt to account
385 for forecasts that are "close enough", not surprisingly, CN exhibits positive skill for all
386 neighborhood sizes and reflectivity thresholds at the initial analysis time for the 0000-UTC
387 initialization time (Fig. 8a). Again, this is because the hydrometeor fields are effectively

[4] The bias-adjusted GSS from Mesinger (2008) was computed (not shown) and depicted smaller differences between CN and C0 through FH 6–8.

388 inserted directly onto the native 4-km grid of the CN model through the cloud analysis scheme.
389 Not surprisingly, the base rates[5] decrease with increasing thresholds (Fig. 8b).

390 By FH 1 however, the 0000-UTC CN quickly loses useful skill at higher thresholds.
391 Forecasts at the 30-dBZ threshold lose skill up to the 20-km neighborhood (~4Δx) and the
392 forecasts at the 40-dBZ threshold lose skill up to the 60-km neighborhood (~12Δx) (Fig. 9a). In
393 effect, the 0000-UTC CN model runs lose forecasting skill for small mesoscale and
394 convective-scale neighborhoods, meaning they have little to no skill in the placement of
395 individual convective cells and small convective clusters, after one hour of integration. For
396 comparison, C0 exhibits no useful skill for all depicted scales and thresholds for both model
397 initialization times (Figs. 9b), but this is not surprising since C0 is still spinning up convection
398 for the first few forecast hours. Although the drop-off in skill for CN is rapid, CN still
399 outperforms C0 for all neighborhood and threshold combinations at FH 1 (Figs. 9c). The
400 neighborhoods and reflectivity thresholds at which the alleviation of the spin-up problem is
401 skillful are limited to neighborhoods greater than 5–10 km for the lower reflectivity thresholds
402 (i.e., 20–30 dBZ) and neighborhoods greater than 20–40 km for the higher thresholds (i.e., 35–
403 40 dBZ). Similarly, CN continues to outperform C0 at FH 2 and is, thus, not shown.

404 By FH 3, the 0000-UTC CN model runs continue to lose useful skill. The 0000-UTC
405 CN model runs have useful skill for neighborhoods greater than ~30 km at the 20-dBZ
406 threshold and for neighborhoods greater than ~140 km at the 40-dBZ threshold (Fig. 10a).
407 Conversely, C0 gains useful skill for neighborhoods greater than ~150 km at the 20-dBZ
408 threshold and for neighborhoods greater than ~310-km at the 40-dBZ threshold (Fig. 10b). CN

[5] The base rate bar graphs will be excluded from the results and discussion from this point forward, but they will be included in the figures for the reader's interest.

409 continues to have greater skill than C0 at FH 3 (Fig. 10c), but the magnitude of the differences
410 are decreasing.

411 At FH 6, the 0000-UTC CN model runs maintain useful skill for neighborhoods greater
412 than ~90 km at the 20-dBZ threshold and for neighborhoods greater than ~140 km at the 40-
413 dBZ threshold (Fig. 11a), which is similar to the results at FH 3. Useful skill exists for
414 neighborhoods greater than ~150 km at the 20-dBZ threshold and for neighborhoods greater
415 than ~320 km at the 40-dBZ threshold for the 0000-UTC C0 model runs (Fig. 11b). The
416 magnitude of the difference in skill between CN and C0 continues to decrease by FH 6, but are
417 deemed to be significant at most neighborhoods and thresholds (Fig. 11c). At forecast hours 9
418 and 12, the results are similar to FH 6 and, thus, are not shown.

419

420 *c) Scale separation method results*

421 Although the neighborhood method effectively weights small distance errors less and
422 less as the size of the neighborhood increases, it does not define the contribution of specific
423 spatial scales to the error (nor to their biases). The scale separation method is able to isolate
424 these scales. Furthermore, the scale separation results give another perspective of where in the
425 reflectivity–spatial-scale parameter space the "useful skill" exists through the ISS. This
426 additional information of useful skill is revealing because the FSS_{useful} used earlier may not be
427 the optimal measure of useful skill.

428 For the initial analysis time, CN exhibits positive skill (positive ISS values are
429 considered to define "useful skill" for the purpose of this study) for all spatial scales at
430 thresholds less than 25 dBZ and for spatial scales greater than ~10 km at the 40-dBZ threshold

(Fig. 12a[6]), so even for the initial hour, CN struggles with analyzing the amplitude of the higher reflectivity cores. This is likely due to a somewhat smoothed representation of convection produced by the cloud analysis schemes. Even though CN is largely unbiased at the initial time, it is initialized with too much coverage of higher reflectivities for large spatial scales based on ERD values greater than 0.2 (Fig. 12b).

At FH 1, CN has no skill for both spatial scales less than about 40 km and thresholds greater than 25 dBZ (Fig. 13a). Thus, individual thunderstorms are not forecasted skillfully even after just 1 hour of integration. A source of error at these scales could be related to the initialization procedure. Because the 3DVAR and cloud analysis are only performed at one time, there likely is not a dynamical balance between the 3DVAR wind field analysis that uses the radial winds in the Doppler radar data and the hydrometeor and in-cloud temperature and moisture adjustments from the cloud analysis that uses the reflectivity data (Hu et al. 2006b). As a result, the storms that are inserted into the initial condition often undergo rapid adjustment, and new storms form along the outflow from the initial storms, or along boundaries and features that are either found in the initial condition, or are inserted into the initial condition by the 3DVAR analysis. C0 has positive skill for spatial scales greater than ~160 km for the lower reflectivity thresholds and for spatial scales greater than ~60 km for the higher reflectivity thresholds (Fig. 13b). The CN and C0 difference field reveals that CN performs better than C0 for all spatial scales and reflectivity thresholds at FH 1 (Fig. 13c), because C0 is still spinning up convection. Of significance is that CN overforecasts (high bias) for all spatial scales and reflectivity thresholds except for the highest thresholds (Fig. 13d) indicating that CN is

[6] The left ordinate in the ISS plots represents the spatial scale of the binary forecast errors and not just the neighborhood size as for the neighborhood method.

generating too much convection in the first forecast hour. The spin-up of convection in C0 is indicated by the negative ERD values for all spatial scales and reflectivity thresholds (Fig. 13e).

At FH 3, CN has no skill for both reflectivity thresholds greater than 30 dBZ and spatial scales less than ~80 km (Fig. 14a). Interestingly, a "tongue" of negative skill exists between the 40-km and 80-km spatial scales for lower reflectivity thresholds. Positive skill exists on either side of this tongue. For C0, the tongue of negative skill exists between the spatial scales of 25 km and 160 km (Fig. 14b). This region of positive skill on the small spatial scales in the plots is likely due to there being very few small-scale events with weak intensity, which causes only small errors compared to the random forecast at these scales (Casati 2010).

At FH 3, in the range of spatial scales from 40 km to 320 km, CN is noticeably better than C0 at the lower reflectivity thresholds but not at 40 dBZ (Fig. 14c). The skill for CN appears to be better than for C0 at scales less than 40 km as well, but the significance is doubtful to nonexistent. This implies that CN has noticeably better skill in the near-term forecasting of precipitation than C0 down to 40-km spatial scales (~8Δx). However, there is negative skill between the 40-km and 80-km spatial scales for CN, so CN's improvement over C0 is only useful for spatial scales greater than ~80 km (~16Δx). Once again, CN overforecasts for all spatial scales and for reflectivity thresholds less than 35 dBZ (Fig. 14d). C0 continues to underforecast for reflectivity thresholds greater than 30 dBZ, but for reflectivity thresholds less than 30 dBZ, C0 has a small positive bias (Fig. 14e) as convection has spun up by this time.

At FH 6, the positive skill at the higher reflectivity thresholds seen at FH 3 remains the same for CN, but the positive skill for the lower thresholds between 80 and 160 km is lost (i.e., the region of negative skill shift toward smaller reflectivities and larger scales) (Fig. 15a). The scales and thresholds of positive skill for C0 changed little from FH 3 to FH 6, except for the

475 slight increase in positive skill for both small spatial scales and low reflectivity thresholds (Fig.
476 15b). The difference plot reveals that C0 has nearly "caught up" with CN for spatial scales less
477 than ~40 km by FH 6 (Fig. 15c). This is consistent with the subjective impressions of the
478 SFE2009 and SFE2010 participants, who tended to say CN and C0 were of nearly equal skill
479 by FH 3 to FH 6. This suggests that the human participants were focusing not only on the
480 higher reflectivity thresholds (see section 4a), but also on relatively small scales. In other
481 words, they may have been rating CN and C0 more so based on smaller-scale convection (e.g.,
482 supercells), as opposed to larger-scale convective systems, which can have broad areas of high
483 reflectivity (> 35 dBZ).

484 However, an interesting result is that CN continues to show better skill than C0 for the
485 spatial scales between ~40 km and ~320 km. A possible reason why the spatial scales and
486 reflectivity thresholds at which CN has positive skill and C0 has significantly lesser or negative
487 skill is that the mesoscale convective systems and squall lines simulated by C0 tend to lag
488 behind the observed systems more so than for CN. A comparison of the SR and OR beginning
489 at 0000 UTC on 14 May 2009 illustrates this problem (Fig. 16). CN's simulated squall line
490 largely overlaps the observed squall line after 6 hours of integration, but C0's simulated squall
491 line lags behind. This is a characteristic that was noticed on several days by the SFE2009
492 participants and was most clearly seen for squall lines and larger convective systems at greater
493 than 30-dBZ thresholds. This shows that CN has difficulty retaining convective-scale skill
494 (scales < 40–80 km) through 3 to 6 hours, but the information on the larger scales appears to be
495 retained at and beyond 6 hours and is manifest as convective systems with a smaller lag with
496 observations compared to C0. With the initial data assimilation, CN has a better handle of the
497 current state of the atmosphere on the larger scales with respect to latent heating and

498 divergence. With that information, CN is able to maintain larger convective systems from the
499 start, while C0 has to take time to develop that same convective system. Also at FH 6, CN
500 continues to overforecast for thresholds less than 35 dBZ, but now underforecasts for
501 thresholds greater than 35 dBZ (Fig. 15d). C0 slightly overforecasts for small spatial scales and
502 low reflectivity thresholds while underforecasting for thresholds greater than 35 dBZ, similar
503 to CN (Fig. 15f), indicating that the spin-up process in C0 is nearly complete. Forecast hours 9
504 and 12 depict similar results to FH 6 and are thus not shown.
505
506 **5. Summary and conclusions**
507 During a period of several weeks in the springs of the past several years, researchers
508 and forecasters from across the country met in Norman, OK for the annual Hazardous Weather
509 Testbed (HWT) Spring Forecasting Experiment to evaluate model forecasts from experimental
510 storm-scale models. In 2009 and 2010, one of their tasks was to rate CN and C0 based on a
511 visual inspection of the simulated and observed reflectivity. Most of the time, the participants
512 noted that the skills of CN and C0 became roughly equivalent sometime between forecast
513 hours 3 and 6. However, some traditional verification metrics, like GSS at lower thresholds, do
514 not necessarily convey this message and can suggest that beneficial information from radar is
515 retained out to at least 12 hours. As such, a main goal of this study was to determine if newer
516 spatial verification techniques provide objective results that are qualitatively more similar to
517 SFE2009 and SFE2010 participants' subjective assessment of the model forecasts than the
518 traditional verification scores. Additionally, another important goal of this study was to
519 evaluate the benefit of the 3DVAR-cloud analysis radar data assimilation technique, which was
520 used in CN's forecasts.

521 To examine the reasons for the apparent discrepancy in subjective and objective metrics,
522 and as part of an objective assessment of the performance of CN and C0, several traditional
523 verification metrics were computed for the CN and C0 forecasts of convection. It was found
524 that the assessment of the relative performance of CN versus C0 depends significantly on the
525 metric of choice and on the chosen threshold of reflectivity (or similarly, 1-h accumulated
526 precipitation). According to the GSS (ETS), CN significantly outperformed C0 out to forecast
527 hour 6 (with doubtful significance out to FH 12, which does not agree well with the assessment
528 from the Spring Forecasting Experiment participants) at the 20-dBZ threshold. However, at the
529 40-dBZ threshold, CN and C0's GSS scores converged after just a few hours of integration,
530 which agrees much better with the sentiment of the SFE participants. This is likely due to the
531 fact that the participants tended to focus on the higher reflectivities in the displays (the
532 reflectivities with yellows and reds) rather than the weaker reflectivities (blues and greens).
533 Also, CN's GSS scores at the 20-dBZ threshold were likely larger than C0's GSS due to CN's
534 high frequency bias, both in an absolute sense and relative to C0, for most thresholds.
535 In addition to the computation of traditional metrics, some new spatial verification
536 techniques were used: the FSS computed using the neighborhood method was used to assess
537 the neighborhood and variable threshold combinations that yield useful forecasting skill for
538 each forecast hour (Roberts and Lean 2008; Ebert 2008). Furthermore, the ISS computed from
539 the scale separation method was used to examine the error (MSE and bias) and skill on specific
540 spatial scales (Casati 2010). These filtering methods serve to give credit to forecasts that are
541 "close enough" – grid-point-by-grid-point metrics do not do so.
542 In general, both the FSS and the ISS show that CN lost most of its useful skill at
543 neighborhood widths and spatial scales smaller than about 40 km ($8\Delta x$), and performs worse

544 the higher is the threshold, after just a few hours of integration. As discussed in Section 4c, a
545 source of additional error at these widths and scales could be related to how there is likely not a
546 dynamical balance between the 3DVAR wind field analysis and the hydrometeor and in-cloud
547 temperature and moisture adjustments from the cloud analysis in the initialization procedure
548 potentially resulting in rapid adjustments of storm coverage patterns. Even with this potential
549 source of error, CN still performed better than C0, which has negative FSS for most
550 neighborhood and threshold combinations through FH 6. Although, it is acknowledged that the
551 Roberts and Lean "target skill" may not be optimal for the more convective precipitation
552 events examined in this study possibly due to being too stringent.

553 The scale separation method applied to the forecasts revealed many similar results
554 compared to the neighborhood method, but there were some differences. For all forecast hours,
555 a benefit of the 3DVAR analysis revealed clearly by the scale-separation method is seen in the
556 larger spatial scales. The significant difference in ISS between CN and C0 for the 40–320 km
557 spatial scales and the lack of any significant differences at smaller spatial scales beyond a few
558 hours shows that convective meso-γ and meso-β scales are contributing little to nothing to the
559 improvement in skill seen for CN versus C0. A possible reason as to why this is the case was
560 discussed in Section 4c: mesoscale convective systems in C0 tend to lag behind the observed
561 systems more so than for CN. This suggests that the information assimilated through the
562 3DVAR-cloud analysis system adds little to no skill at convective and smaller meso scales (<
563 40–80 km) starting at FH 1, but adds skill compared to a cold start at larger scales, even out to
564 FH 12.

565 A goal of this work was to find objective measures of model skill that match the
566 subjective impressions of experts that evaluated the models subjectively. It is found that the

GSS for high reflectivity thresholds (> 35 dBZ) matches subjective impressions that CN performed similarly to C0 by 3–6 hours into the forecast, more so than lower thresholds (~20 dBZ). Furthermore, objective spatial verification metrics that examine model skill at scales less than 40–80 km match the subjective impressions as well. Therefore, these metrics (for these scales and thresholds) may be appropriate for use to provide an assessment consistent with experts' impressions of convection-allowing model forecast skill.

Furthermore, the two spatial filtering methods gave a more comprehensive characterization of the performance of the convection-allowing models than the traditional verification methods. The neighborhood and scale separation methods revealed where "useful skill" might exist for several forecast hours in the reflectivity–spatial-scale parameter space that was not regularly apparent in the subjective evaluations or in the objective verification using the traditional scores. It is hoped that these results encourage future use of these new spatial verification metrics rather than the continued use of traditional verification metrics at single thresholds to characterize the performance of high-resolution, convection-allowing models. This is the first known study to appear in the refereed literature to use radar reflectivity instead of accumulated precipitation as the verification field for aggregate statistics computed over multiple seasons. It was found that both fields lead to similar results for all three verification methods discussed, giving confidence in the use of hourly simulated and observed reflectivity as a robust way to measure the performance of convection-allowing models. Finally, it would be beneficial to use these spatial verification metrics on case studies of convection or, perhaps, subsets of modes of convection (e.g. supercells versus disorganized multicells versus squall lines) aggregated together in order to not only verify model forecasts of convection, but

589 also to study the different skill score structures associated with the various modes of
590 convection.
591
592 *Acknowledgements* This work was supported through NSSL and NOAA Earth System
593 Research Laboratory's Global Systems Division (GSD) director's discretionary funds and
594 through NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of
595 Oklahoma Cooperative Agreement #NA08OAR4320904, U.S. Department of Commerce. Dr.
596 Kim Elmore is thanked for his very helpful discussion of confidence intervals. Tara Jensen and
597 everyone else associated with the use of MET at the DTC are thanked for their help in using
598 MET. Drs. Jack Kain and Fred Carr are thanked for their insight into model verification. Model
599 Evaluation Tools (MET) was developed at the National Center for Atmospheric Research
600 (NCAR) through a grant from the United States Air Force Weather Agency (AFWA). NCAR
601 is sponsored by the United States National Science Foundation. The CAPS forecasts were
602 supported by the NOAA CSTAR grant to CAPS, and were produced at the National Institute of
603 Computational Science (NICS) at the University of Tennessee, and at the Oklahoma
604 Supercomputing Center for Research and Education (OSCER). Kevin Thomas and Fanyou
605 Kong were instrumental in producing the forecasts.
606
607
608
609
610
611
612

613 **References**

614 Baldwin, M. E., and J. S. Kain, 2006: Sensitivity of Several Performance Measures to
615 Displacement Error, Bias, and Event Frequency. *Wea. Forecasting*, **21**, 636–648.

616 Benjamin, S. G., S. S. Weygandt, J. M. Brown, T. Smirnova, D. Devenyi, K. Brundage, G.
617 Grell, S. Peckham, W. R. Moninger, T. W. Schlatter, T. L. Smith, and G. Manikin, 2008:
618 Implementation of the radar-enhanced RUC. 13th Conf. Aviation, Range and Aerospace
619 Meteorology, Amer. Meteor. Soc., New Orleans, LA.

620 Casati, B., G. Ross, and D. B. Stephenson, 2004: A new intensity-scale approach for the
621 verification of spatial precipitation forecasts. *Meteorological Applications*, **11**, 141–154.

622 --------, L. J. Wilson, D. B. Stephenson, P. Nurmi, A. Ghelli, M. Pocernich, U. Damrath, E. E.
623 Ebert, B. G. Brown, and S. Mason, 2008: Forecast verification: current status and future
624 directions. *Meteorological Applications*, **15**, 3–18.

625 --------, 2010: New developments of the intensity-scale technique within the Spatial
626 Verification Methods Intercomparison Project. *Wea. Forecasting*, **25**, 113–143.

627 Clark, Adam J., and Coauthors, 2012: An Overview of the 2010 Hazardous Weather Testbed
628 Experimental Forecast Program Spring Experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74.

629 Developmental Testbed Center (DTC), 2011: MET: Version 3.0 Model Evaluation Tools Users
630 Guide. Available at http://www.dtcenter.org/met/users/docs/overview.php. 209 pp.

631 Doswell, Charles A., Robert Davies-Jones, David L. Keller, 1990: On Summary
632 Measures of Skill in Rare Event Forecasting Based on Contingency Tables. *Wea.*
633 *Forecasting*, **5**, 576–585.

634 Easterling, David R., Peter J. Robinson, 1985: The Diurnal Variation of Thunderstorm Activity
635 in the United States. *J. Climate Appl. Meteor.*, **24**, 1048–1058.

636 Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: a review and
637 proposed framework. *Meteorological Applications*, **15**, 51–64.
638 --------, 2009: Neighborhood Verification: A Strategy for Rewarding Close Forecasts. *Wea.*
639 *Forecasting*, **24**, 1498–1510.
640 Elliot, A. J., and M. A. Maier, 2007: Color and psychological functioning. *Curr. Dir. Psychol.*
641 *Sci.*, **16**, 250-254.
642 Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of
643 Spatial Forecast Verification Methods. *Wea. Forecasting*, **24**, 1416–1430.
644 --------, --------, --------, and E. E. Ebert, 2010: Verifying Forecasts Spatially. *Bull. Amer.*
645 *Meteor. Soc.*, **91**, 1365–1373.
646 Hamill, T. M., 1999: Hypothesis Tests for Evaluating Numerical Precipitation Forecasts. *Wea.*
647 *Forecasting*, **14**, 155–167.
648 Hamill, T. M. and Juras, J. (2006), Measuring forecast skill: is it real skill or is it the varying
649 climatology?. *Q.J.R. Meteorol. Soc.*, **132**: 2905–2923.
650 Hu, M., M. Xue, and K. Brewster, 2006: 3DVAR and Cloud Analysis with WSR-88D Level-II
651 Data for the Prediction of the Fort Worth, Texas, Tornadic Thunderstorms. Part I: Cloud
652 Analysis and Its Impact. *Mon. Wea. Rev.*, **134**, 675–698.
653 --------, --------, J. Gao, and K. Brewster, 2006: 3DVAR and Cloud Analysis with WSR-88D
654 Level-II Data for the Prediction of the Fort Worth, Texas, Tornadic Thunderstorms. Part II:
655 Impact of Radial Velocity Analysis via 3DVAR. *Mon. Wea. Rev.*, **134**, 699–721.
656 Kain, J. S., M. C. Coniglio, S. J. Weiss, T. L. Jensen, B. G. Brown, M. Xue, F. Kong, K. W.
657 Thomas, C. S. Schwartz, and J. J. Levit, 2010: Assessing Advances in the Assimilation of

Radar Data and Other Mesoscale Observations within a Collaborative Forecasting–Research Environment. *Wea. Forecasting*, **25**, 1510–1521.

Lichtenfeld, S., Maier, M. A., Elliot, A. J., Pekrun, R., 2009: The semantic red effect: processing the word red undermines intellectual performance. *J. Exp. Soc. Psychol.*, **45**, 1273–1276.

Mesinger, F., 2008: Bias adjusted precipitation threat scores. *Adv. Geosciences* **16**, 137–142.

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.

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.

Schwartz, Craig S., and Coauthors, 2009: Next-Day Convection-Allowing WRF Model Guidance: A Second Look at 2-km versus 4-km Grid Spacing. *Mon. Wea. Rev.*, **137**, 3351–3372.

Stephenson, D. B., B. Casati, C. A. T. Ferro, and C. A. Wilson, 2008: The extreme dependency score: a non-vanishing measure for forecasts of rare events. *Meteor. Appl.*, **15**: 41–50.

Vasiloff, Steven V., and Coauthors, 2007: Improving QPE and Very Short Term QPF: An Initiative for a Community-Wide Integrated Approach. *Bull. Amer. Meteor. Soc.*, **88**, 1899–1911

Wallace, J. M., 1975: Diurnal Variations in Precipitation and Thunderstorm Frequency over the Conterminous United States. *Mon. Wea. Rev.*, **103**, 406–419.

679 Wurman, J. D. Dowell, Y. Richardson, P. Markowski, D. Burgess, L. Wicker, and H. Bluestein,
680 2012: Verification of the Origin of Rotation in Tornadoes Experiment 2: VORTEX 2. *Bull.*
681 *Amer. Meteor. Soc.*, Accepted with minor revision.
682 Xue, M., D.-H. Wang, J.-D. Gao, K. Brewster, and K. K. Droegemeier (2003), The Advanced
683 Regional Prediction System (ARPS), storm-scale numerical weather prediction and data
684 assimilation, *Meteor. Atmos. Physics*, **82**, 139-170.
685 --------, and Coauthors, 2009: CAPS realtime multi-model convection-allowing ensemble and
686 1-km convection-resolving forecasts for the NOAA Hazardous Weather Testbed 2009
687 Spring Experiment. 23rd Conference on Weather Analysis and Forecasting/19th Conference
688 on Numerical Weather Prediction, Omaha, NE, Amer. Meteor. Soc., 16A.2.
689 --------, and Coauthors, 2010: CAPS realtime storm scale ensemble and high resolution
690 forecasts for the NOAA Hazardous Weather Testbed 2010 Spring Experiment. 25th Conf.
691 On Severe Local Storms, Denver, CO, Amer. Meteor. Soc., 7B.3.
692
693
694
695
696
697
698
699
700
701
702

List of Figures

Fig. 1. Observed composite (column maximum) reflectivity (left column) and 0000-UTC initialized simulated composite reflectivity from CAPS CN (middle column) and CAPS C0 (right column) from 5 June 2008 for hours 0000, 0100, 0300, and 0600 UTC (from Kain et al. 2010).

Fig. 2. The inner thick box is the domain for 0000-UTC WRF-ARW with 900x672 grid points in 2009 in (a) and 999x790 grid points in 2010 in (b), and the outer thick box is the somewhat larger ARPS 3DVAR analysis domain. The inner dashed box is the 1200-UTC model domain for both years (444x480 grid points). The grey shaded polygon is an example "VORTEX2" moveable domain with 256x256 horizontal grid points used for verification (see Xue et al. 2009; Xue et al. 2010).

Fig. 3. (a) Example of a binary difference field, where the blue shading represents misses, the red shading represents false alarms, and the white areas represent hits and correct negatives, of simulated reflectivity ≥ 20 dBZ, (b) scale 1 component, (c) scale 2 component, (d) scale 3 component, (e) scale 4 component, and (f) scale 5 component for 10 May 2010 1500 UTC CAPS CN at forecast hour 8. Major tornado outbreak occurred from the afternoon into evening in central-eastern Oklahoma and Kansas.

Fig. 4. GSS for 2009/2010 0000-UTC CAPS CN (long-dash line) and CAPS C0 (short-dash line) at the (a) 20-dBZ threshold, (b) 40-dBZ threshold, and (c)1.0-mm/hr threshold. Solid black line is the difference between CN and C0, and vertical black lines on the difference line represent the 95% confidence intervals.

Fig. 5. Same as Figs. 4a and 4b, except for 1200-UTC CAPS CN and CAPS C0.

Fig. 6. Same as Figs. 4a and 4b, except for frequency bias (FBIAS).

726 Fig. 7. Same as Figs. 4a and 4b, except for POFD.

727 Fig. 8. (a) FSS–FSS_{useful} for 2009/2010 0000-UTC CAPS CN at FH 0 for reflectivity thresholds
728 every 5 dBZ from 20 dBZ and 40 dBZ and for neighborhood sizes from 5 km to 320
729 km. Grey shading with solid contours represent useful skill, and grey shading with
730 dashed contours (not depicted here) represent non-useful skill. Values along the right-
731 ordinate represent multiples of grid-spacing. (b) Base rates of observed reflectivity for
732 each threshold.

733 Fig. 9. FSS–FSS_{useful} for 2009/2010 0000-UTC (a) CAPS CN and (b) CAPS C0 at FH 1 for
734 reflectivity thresholds every 5 dBZ from 20 dBZ and 40 dBZ and for spatial scales from
735 5 km to 320 km. Grey shading with solid contours represent useful skill, and grey
736 shading with dashed contours represent non-useful skill. Also, the differences between
737 CN and C0 are shown in (c), where grey shading with solid contours represent $FSS_{CN} >$
738 FSS_{C0}, grey shading with dashed contours (not depicted in these plots) represent FSS_{CN}
739 $< FSS_{C0}$, and stippling depicts the 95% confidence interval (note, significance exists for
740 all sizes and thresholds). Values along the right-ordinate represent multiples of grid-
741 spacing. (d) Base rates of observed reflectivity for each threshold.

742 Fig. 10. Same as Fig. 9, except for FH 3.

743 Fig. 11. Same as Fig. 9, except for FH 6.

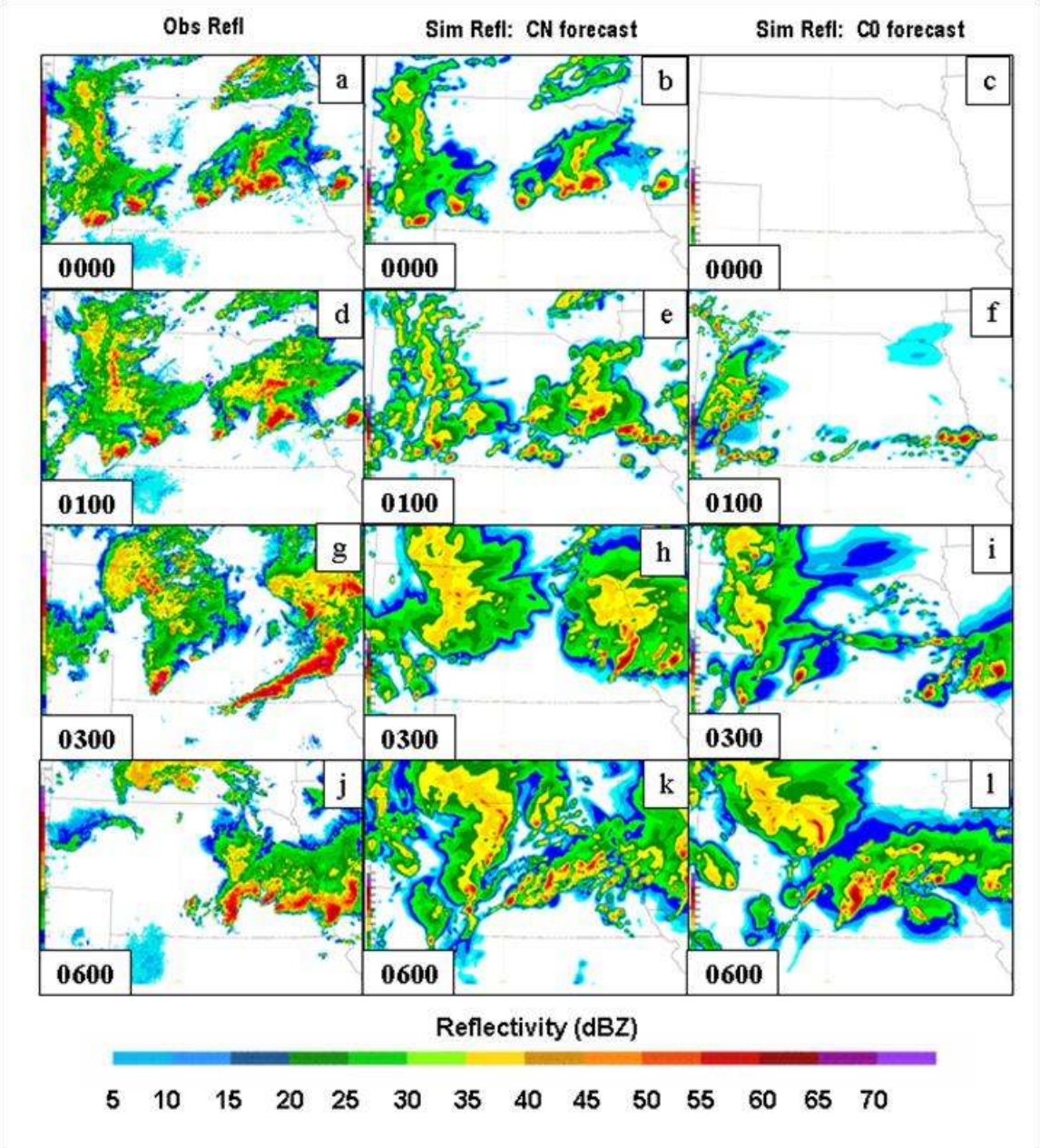
744 Fig. 12. (a) ISS and (b) ERD values for 2009/2010 0000-UTC CAPS CN at FH 0 for
745 reflectivity thresholds every 5 dBZ from 20 dBZ to 40 dBZ and spatial scales from 5
746 km to 320 km. Grey shading with solid contours in (a) represent positive skill, and grey
747 shading with dashed contours in (a) represent negative skill. Grey shading with solid
748 contours in (b) represent overforecasting, and grey shading with negative contours in

(b) represent underforecasting. Values along the right-ordinate represent multiples of grid-spacing.

Fig. 13. ISS for 2009/2010 0000-UTC (a) CAPS CN and (b) CAPS C0 at FH 1 for reflectivity thresholds every 5 dBZ from 20 dBZ to 40 dBZ and spatial scales from 5 km to 320 km. Grey shading with solid contours in (a) and (b) represent positive skill, and grey shading with negative contours in (a) and (b) represent negative skill. ISS differences between CN and C0 are shown in (c), where grey shading with solid contours represent $ISS_{CN} > ISS_{C0}$, grey shading with dashed contours (not depicted in these plots) represent $ISS_{CN} < ISS_{C0}$, and stippling represents 95% statistical significance. ERD values for (d) CN and (e) C0, where grey shading with solid contours in (d) and (e) represent overforecasting, and grey shading with dashed contours in (d) and (e) represent underforecasting. Values along the right-ordinate represent multiples of grid-spacing. (f) Base rates of observed reflectivity for each threshold.

Fig. 15. Same as Fig. 13, except for FH 6.

Fig. 16. First row is observed composite reflectivity from 14 May 2009 for (a) 0000 UTC, (f) 0200 UTC, (k) 0400 UTC, and (p) 0600 UTC. Second and third rows are simulated reflectivity forecasts from 0000-UTC CAPS CN (b, g, l, and q) and CAPS C0 (c, h, m, and r) for the same times. In the bottom two rows, 30-dBZ thresholded observed reflectivity is marked by the thin blue line (d/e, i/j, n/o, and s/t). Red shading represents 30-dBZ thresholded simulated reflectivity for CN (d, i, n, and s) and C0 (e, j, o, and t). From Kain et al (2010).



772
773 *Fig. 1. Observed composite (column maximum) reflectivity (left column) and 0000-UTC*
774 *initialized simulated composite reflectivity from CAPS CN (middle column) and CAPS C0*
775 *(right column) from 5 June 2008 for hours 0000, 0100, 0300, and 0600 UTC (from Kain et al.*
776 *2010).*
777

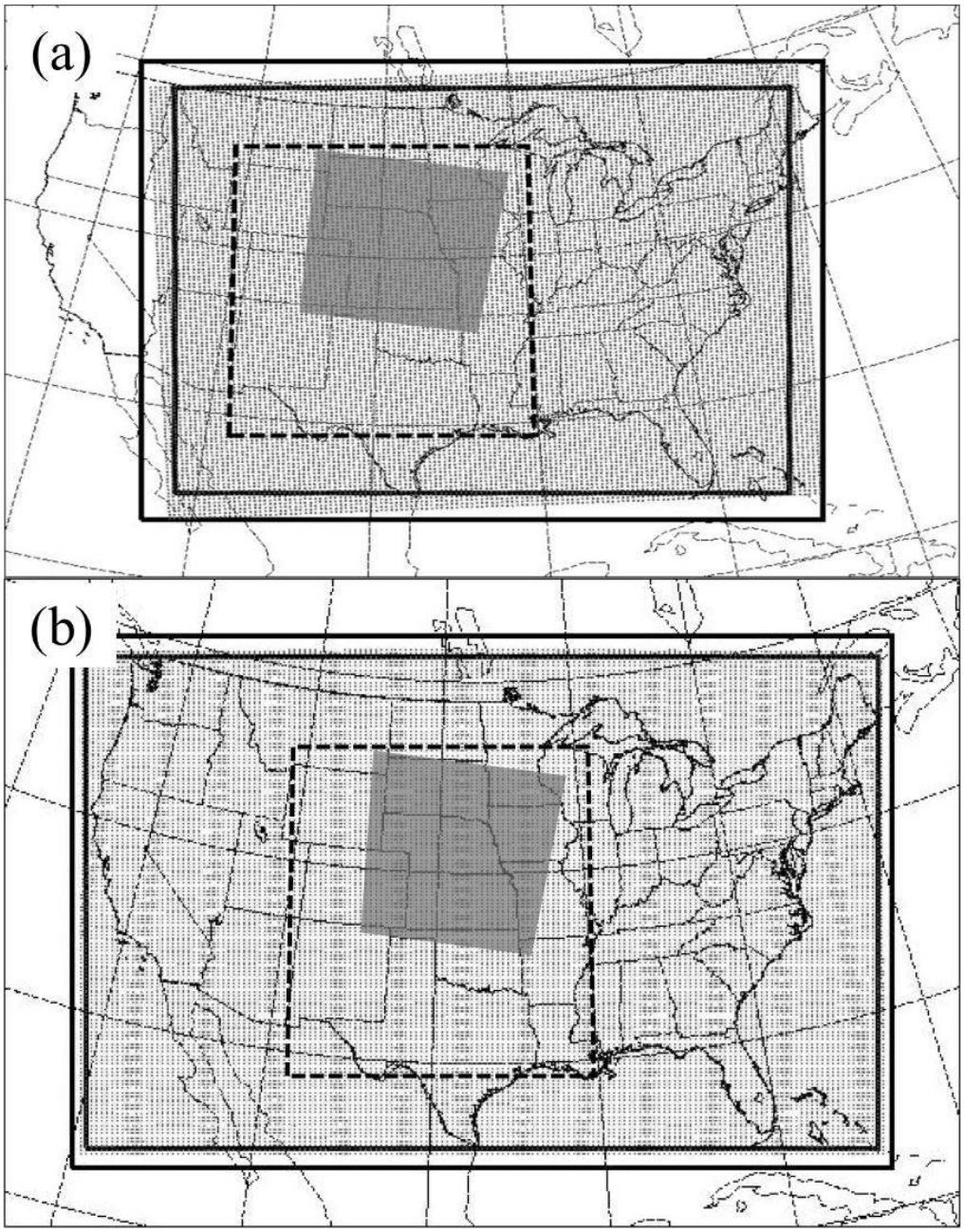


Fig. 2. The inner thick box is the domain for 0000-UTC WRF-ARW with 900x672 grid points in 2009 in (a) and 999x790 grid points in 2010 in (b), and the outer thick box is the somewhat larger ARPS 3DVAR analysis domain. The inner dashed box is the 1200-UTC model domain for both years (444x480 grid points). The grey shaded polygon is an example "VORTEX2" moveable domain with 256x256 horizontal grid points used for verification (see Xue et al. 2009; Xue et al. 2010).

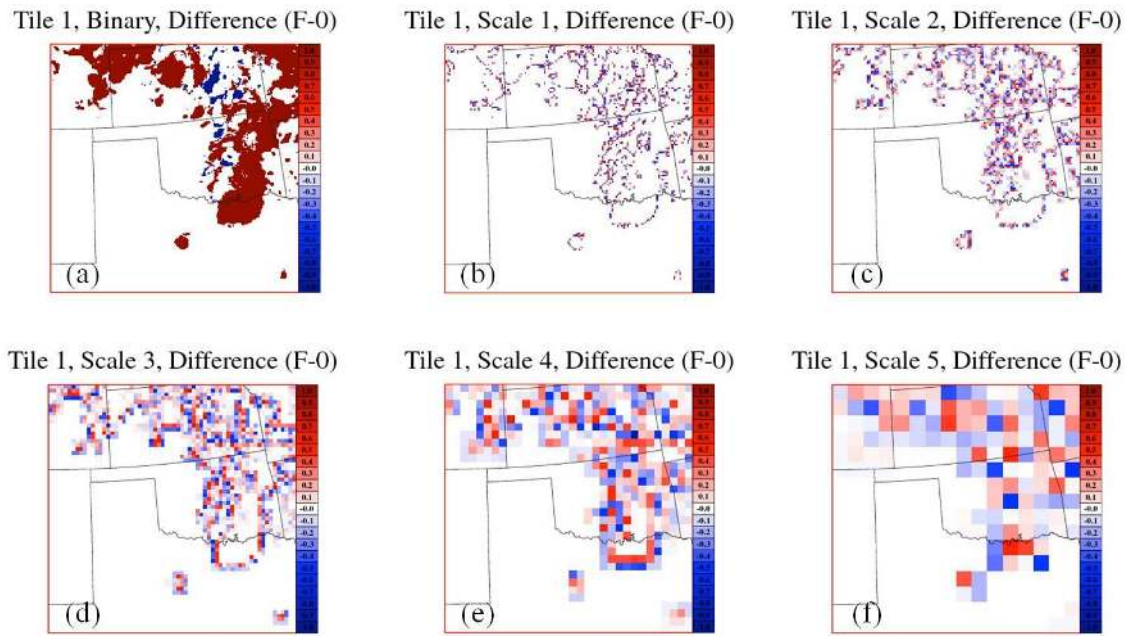


Fig. 3. (a) Example of a binary difference field, where the blue shading represents misses, the red shading represents false alarms, and the white areas represent hits and correct negatives, of simulated reflectivity ≥ 20 dBZ, (b) scale 1 component, (c) scale 2 component, (d) scale 3 component, (e) scale 4 component, and (f) scale 5 component for 10 May 2010 1500 UTC CAPS CN at forecast hour 8. Major tornado outbreak occurred from the afternoon into evening in central-eastern Oklahoma and Kansas.

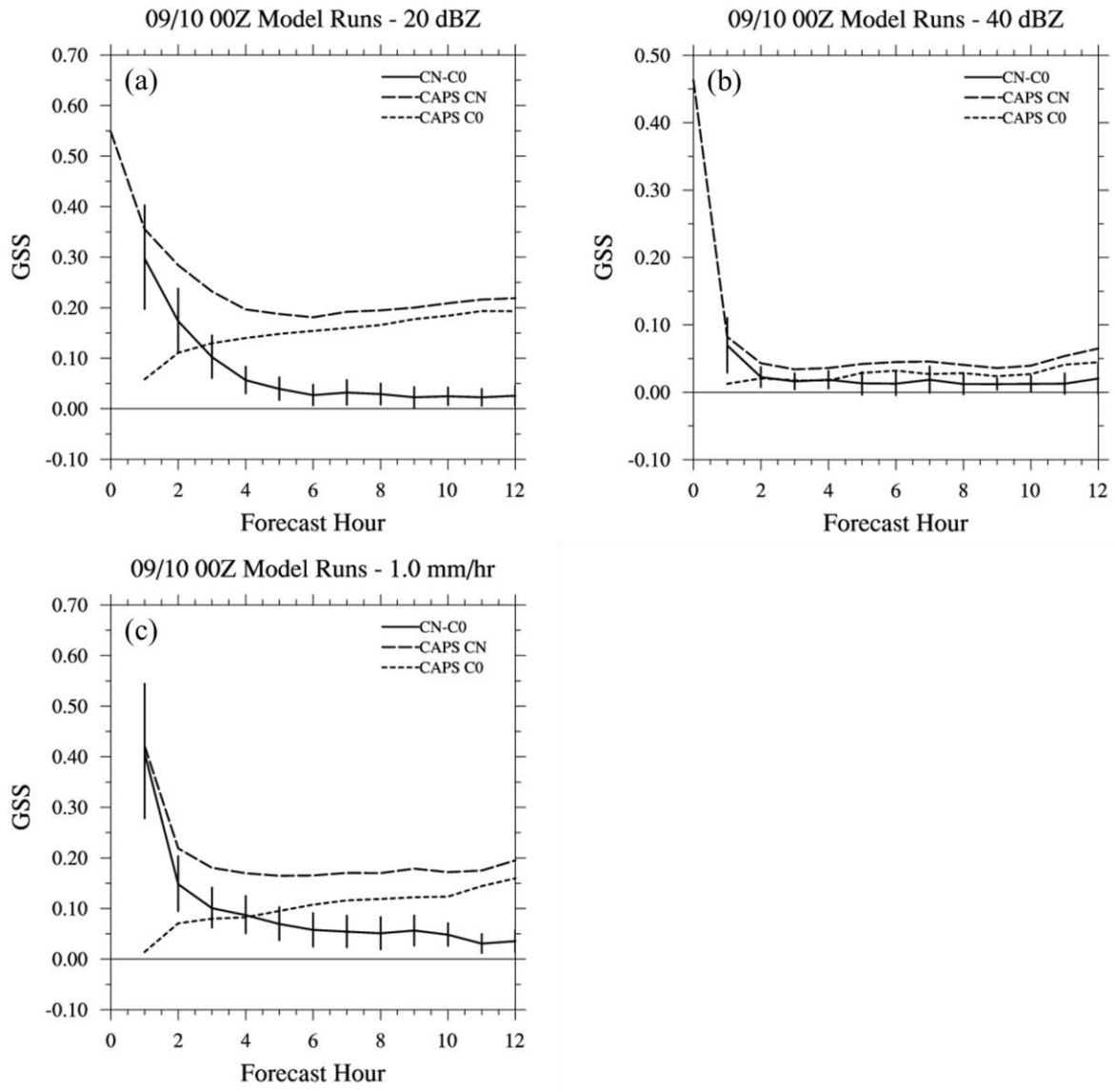


Fig. 4. GSS for 2009/2010 0000-UTC CAPS CN (long-dash line) and CAPS C0 (short-dash line) at the (a) 20-dBZ threshold, (b) 40-dBZ threshold, and (c)1.0-mm/hr threshold. Solid black line is the difference between CN and C0, and vertical black lines on the difference line represent the 95% confidence intervals.

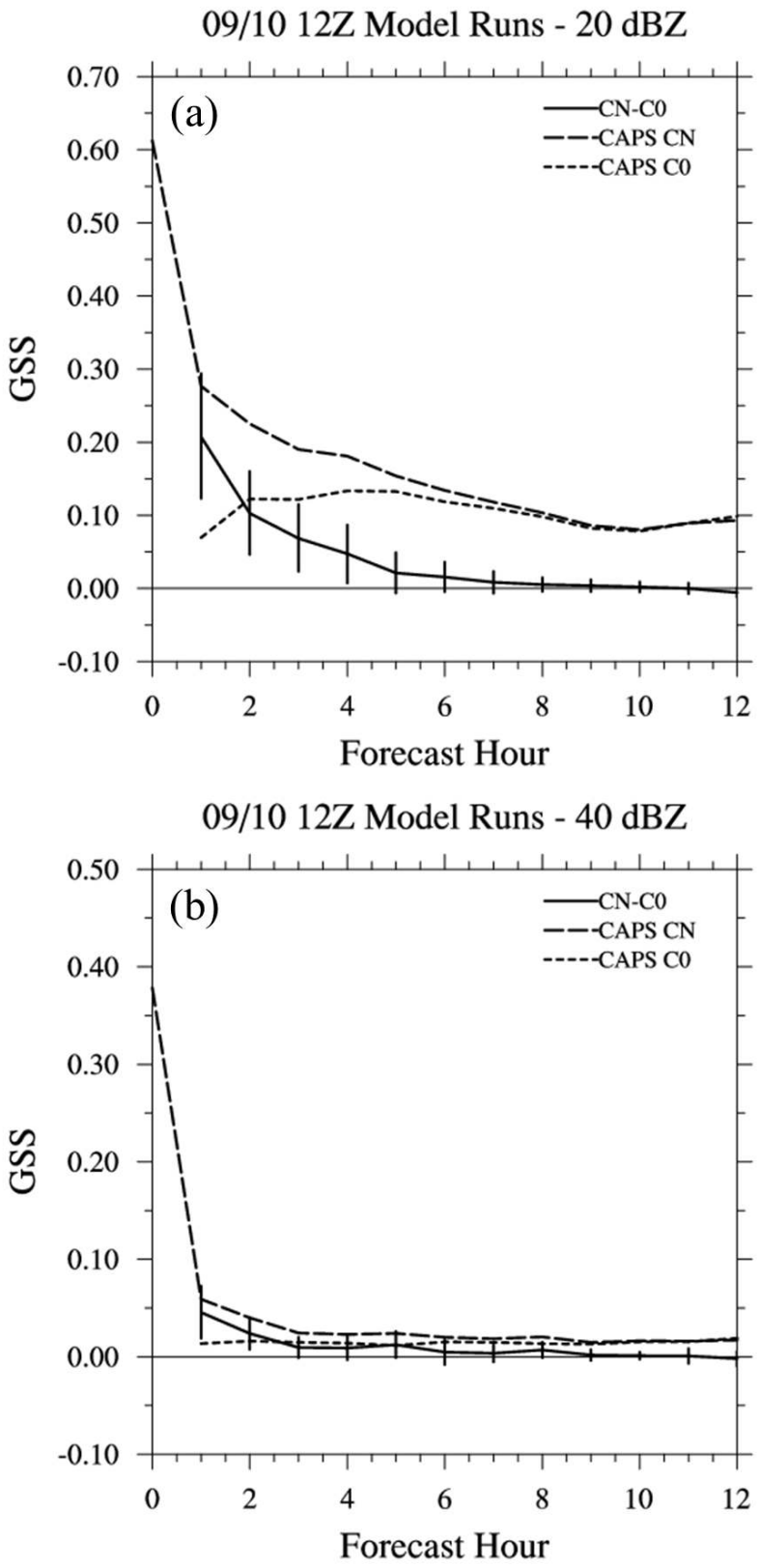


Fig. 5. Same as Figs. 4a and 4b, except for 1200-UTC CAPS CN and CAPS C0.

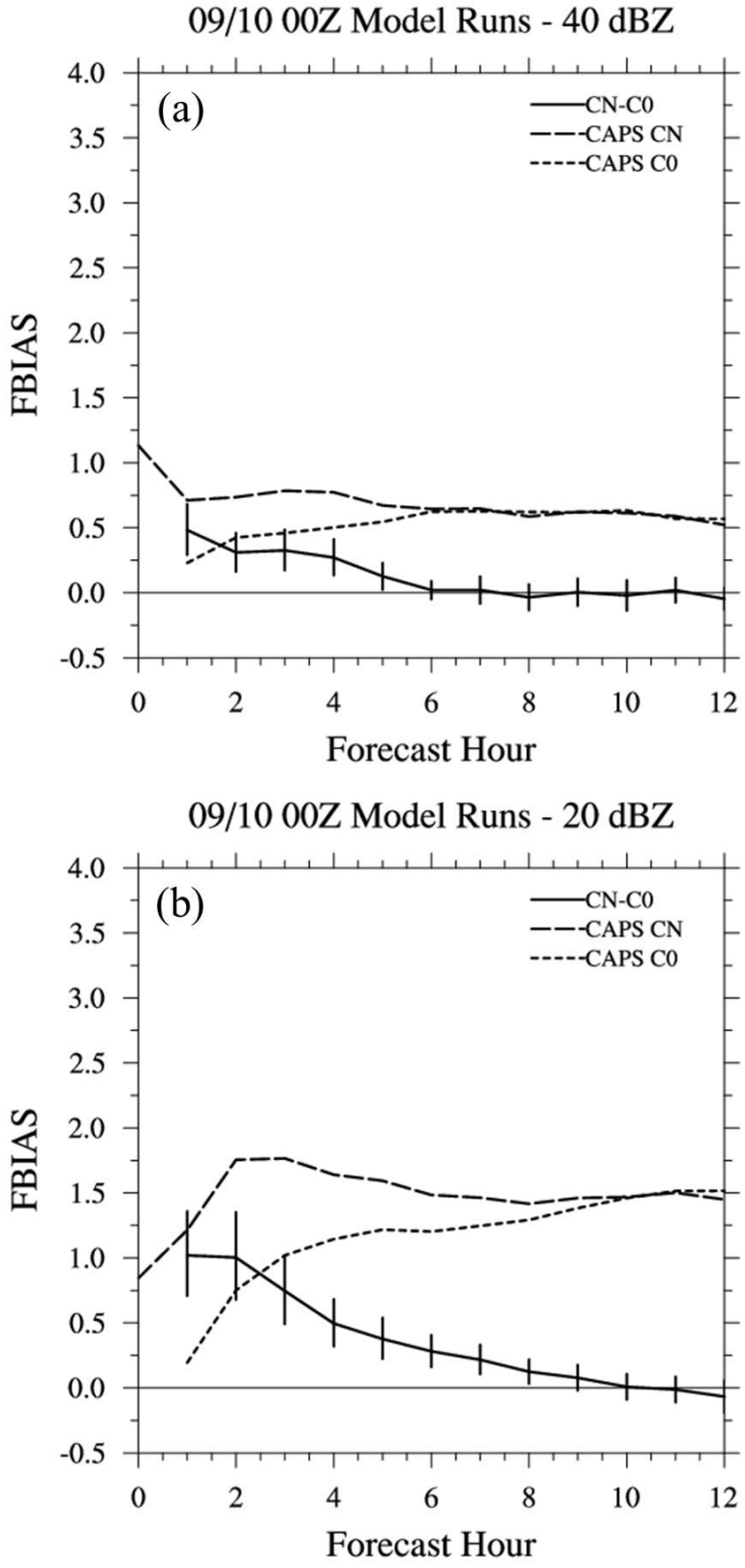


Fig. 6. Same as Figs. 4a and 4b, except for frequency bias (FBIAS).

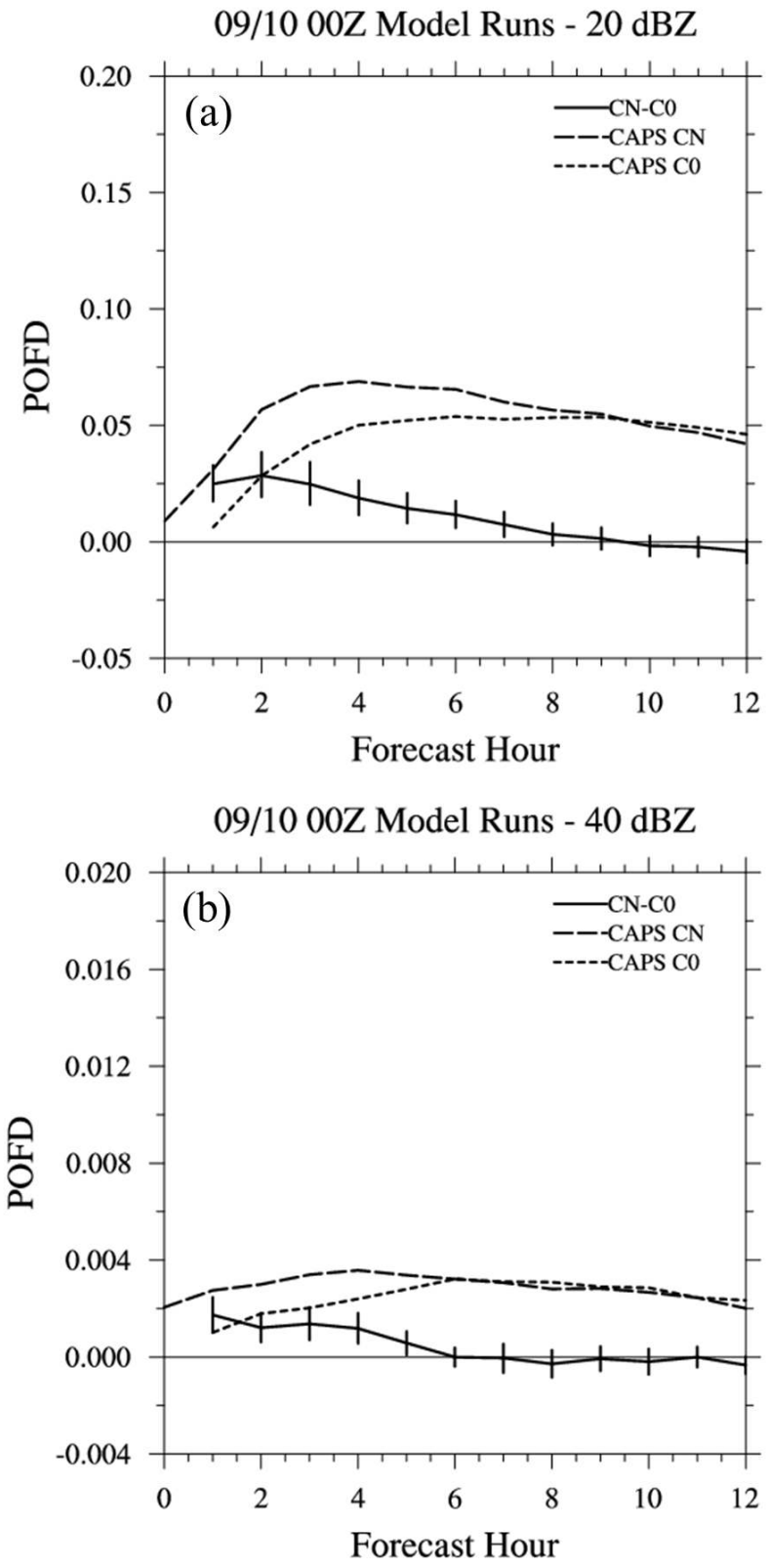


Fig. 7. Same as Figs. 4a and 4b, except for POFD.

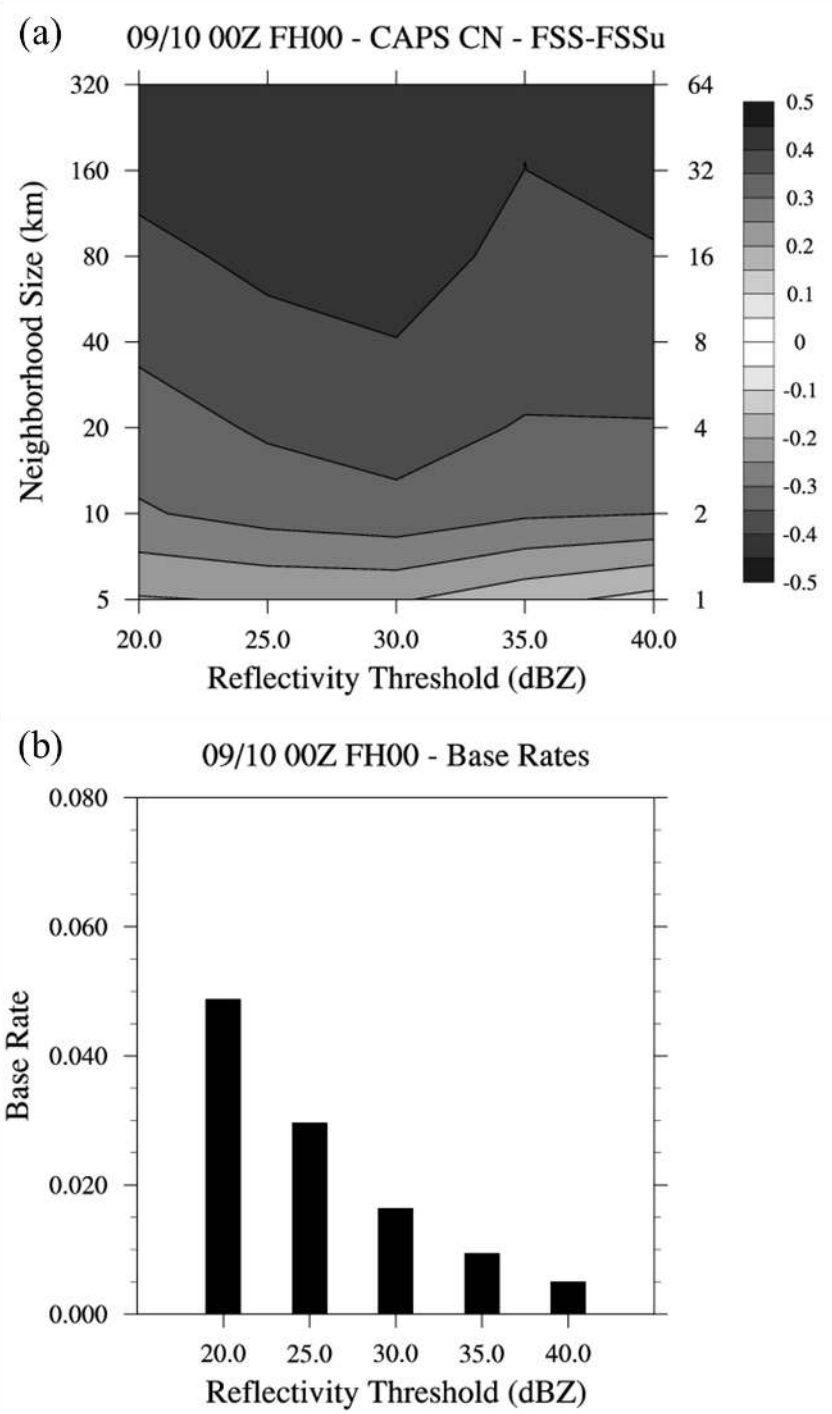


Fig. 8. (a) FSS–FSS_{useful} for 2009/2010 0000-UTC CAPS CN at FH 0 for reflectivity thresholds every 5 dBZ from 20 dBZ and 40 dBZ and for neighborhood sizes from 5 km to 320 km. Grey shading with solid contours represent useful skill, and grey shading with dashed contours (not depicted here) represent non-useful skill. Values along the right-ordinate represent multiples of grid-spacing. (b) Base rates of observed reflectivity for each threshold.

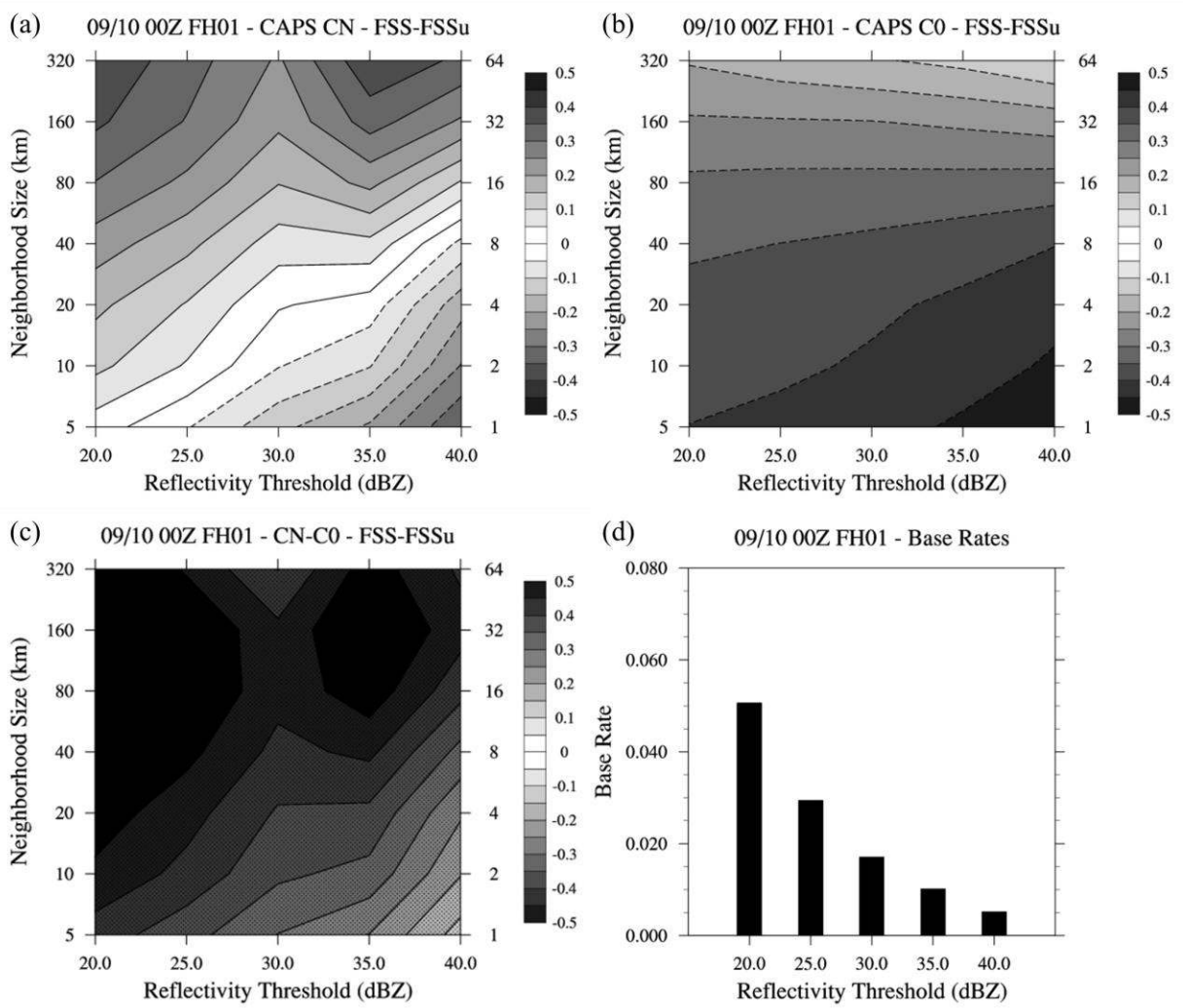


Fig. 9. FSS–FSS_{useful} for 2009/2010 0000-UTC (a) CAPS CN and (b) CAPS C0 at FH 1 for reflectivity thresholds every 5 dBZ from 20 dBZ and 40 dBZ and for spatial scales from 5 km to 320 km. Grey shading with solid contours represent useful skill, and grey shading with dashed contours represent non-useful skill. Also, the differences between CN and C0 are shown in (c), where grey shading with solid contours represent $FSS_{CN} > FSS_{C0}$, grey shading with dashed contours (not depicted in these plots) represent $FSS_{CN} < FSS_{C0}$, and stippling depicts the 95% confidence interval (note, significance exists for all sizes and thresholds). Values along the right-ordinate represent multiples of grid-spacing. (d) Base rates of observed reflectivity for each threshold.

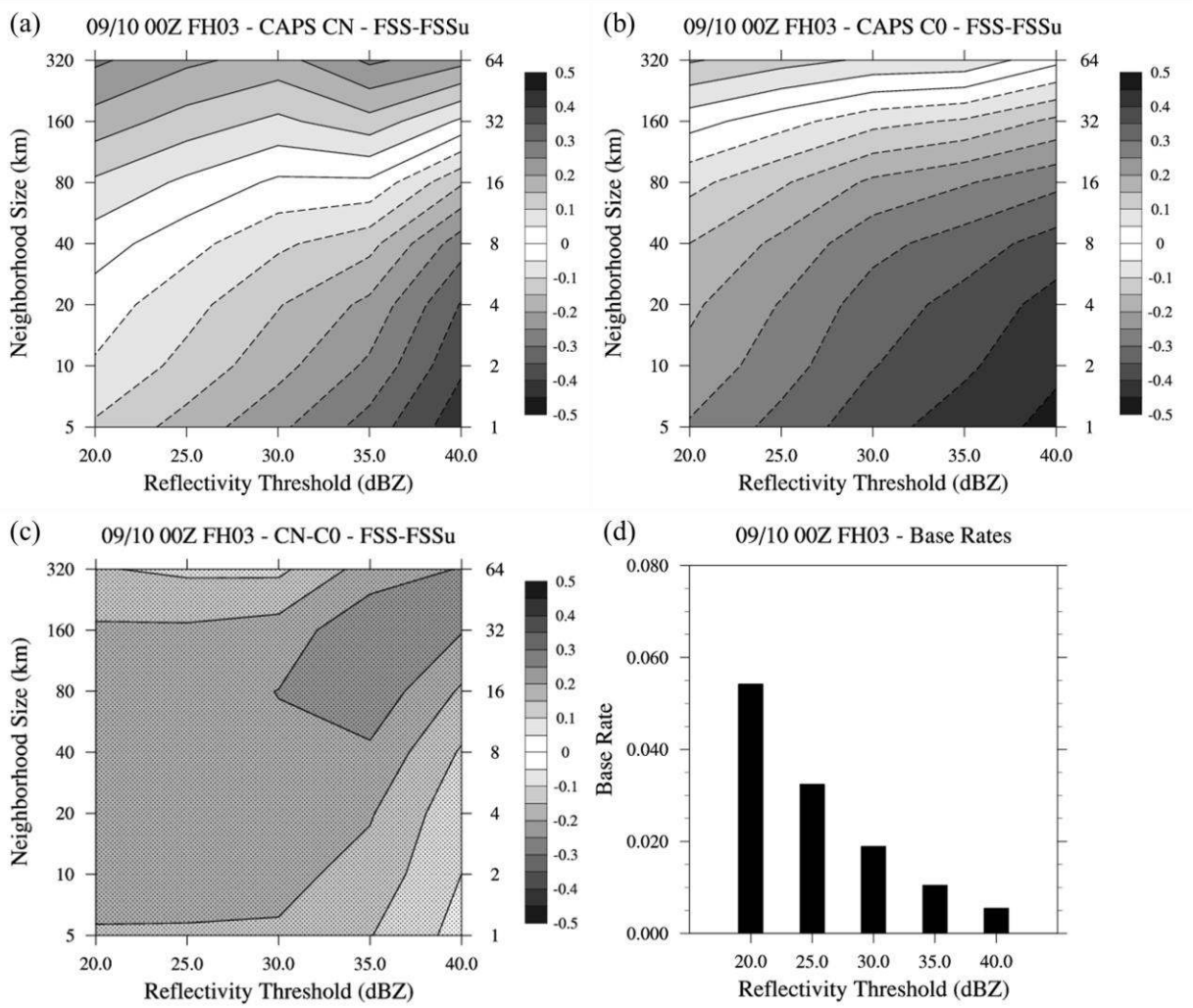


Fig. 10. Same as Fig. 9, except for FH 3.

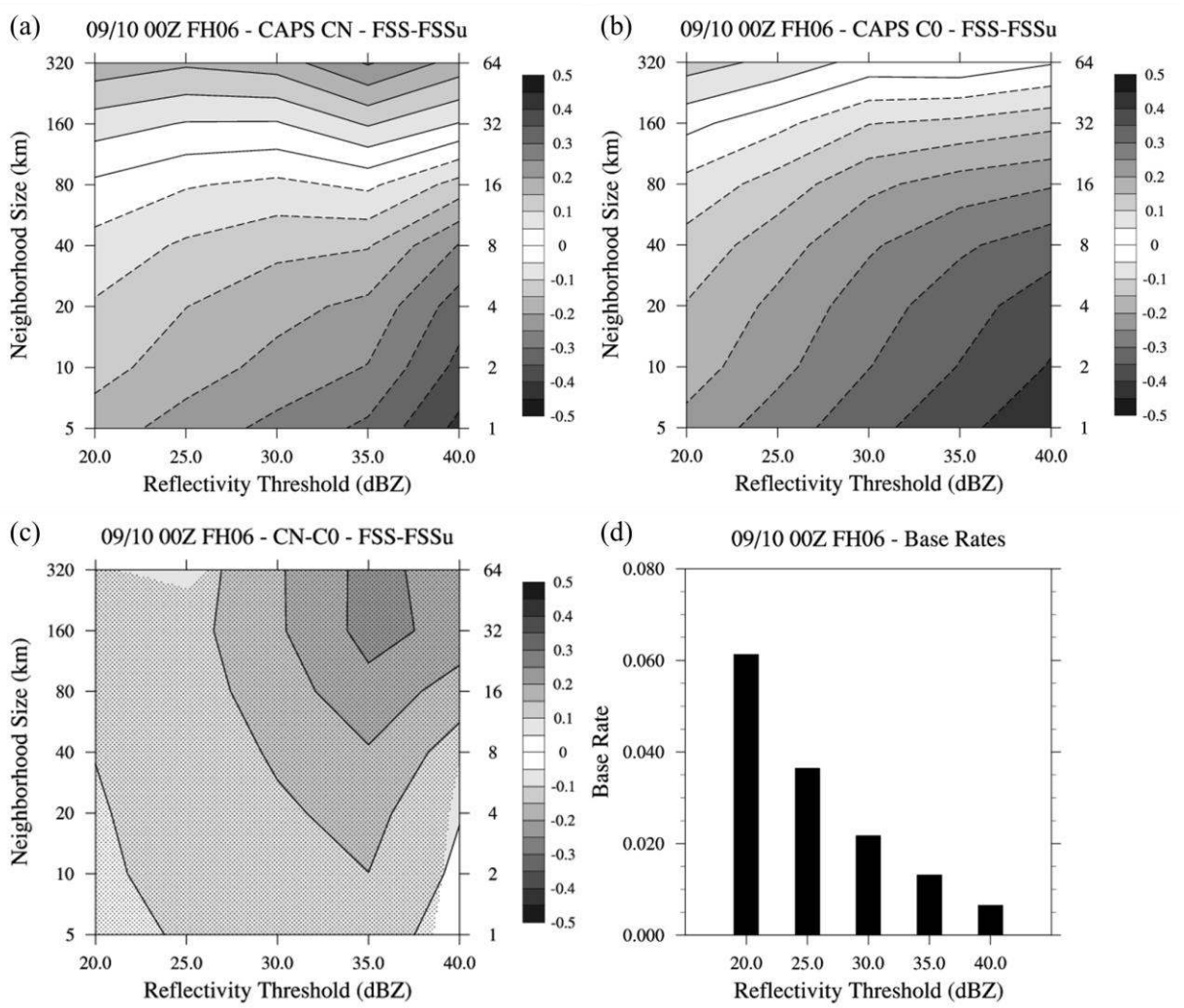


Fig. 11. Same as Fig. 9, except for FH 6.

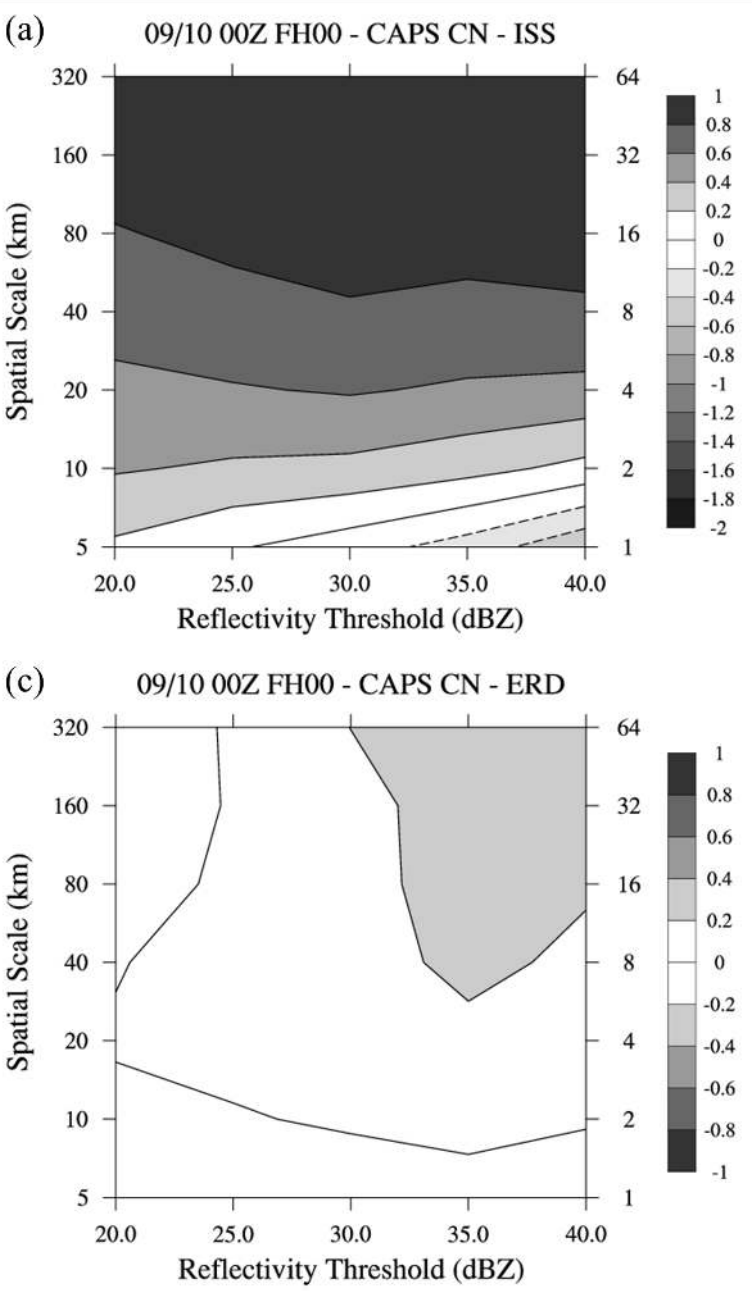


Fig. 12. (a) ISS and (b) ERD values for 2009/2010 0000-UTC CAPS CN at FH 0 for reflectivity thresholds every 5 dBZ from 20 dBZ to 40 dBZ and spatial scales from 5 km to 320 km. Grey shading with solid contours in (a) represent positive skill, and grey shading with dashed contours in (a) represent negative skill. Grey shading with solid contours in (b) represent overforecasting, and grey shading with negative contours in (b) represent underforecasting. Values along the right-ordinate represent multiples of grid-spacing.

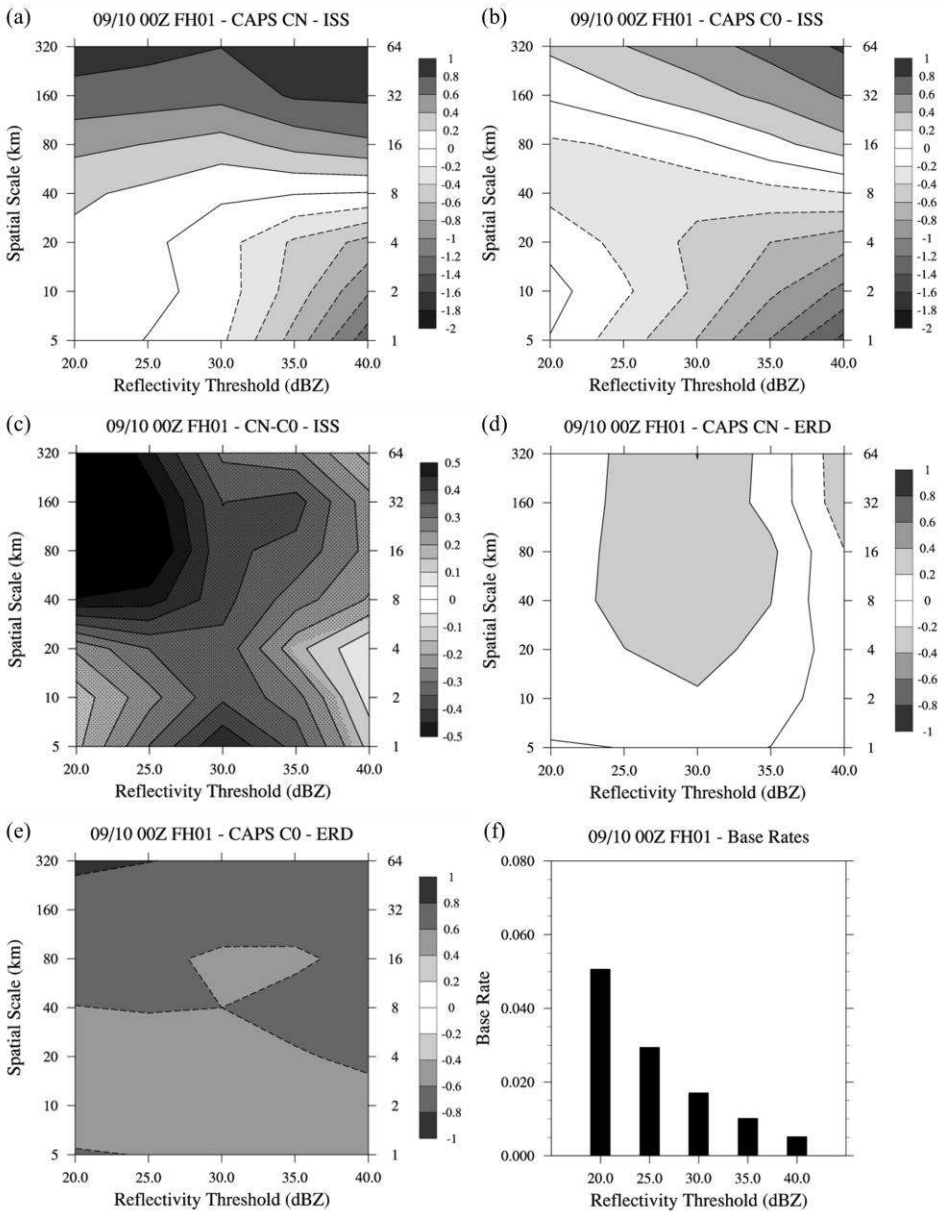
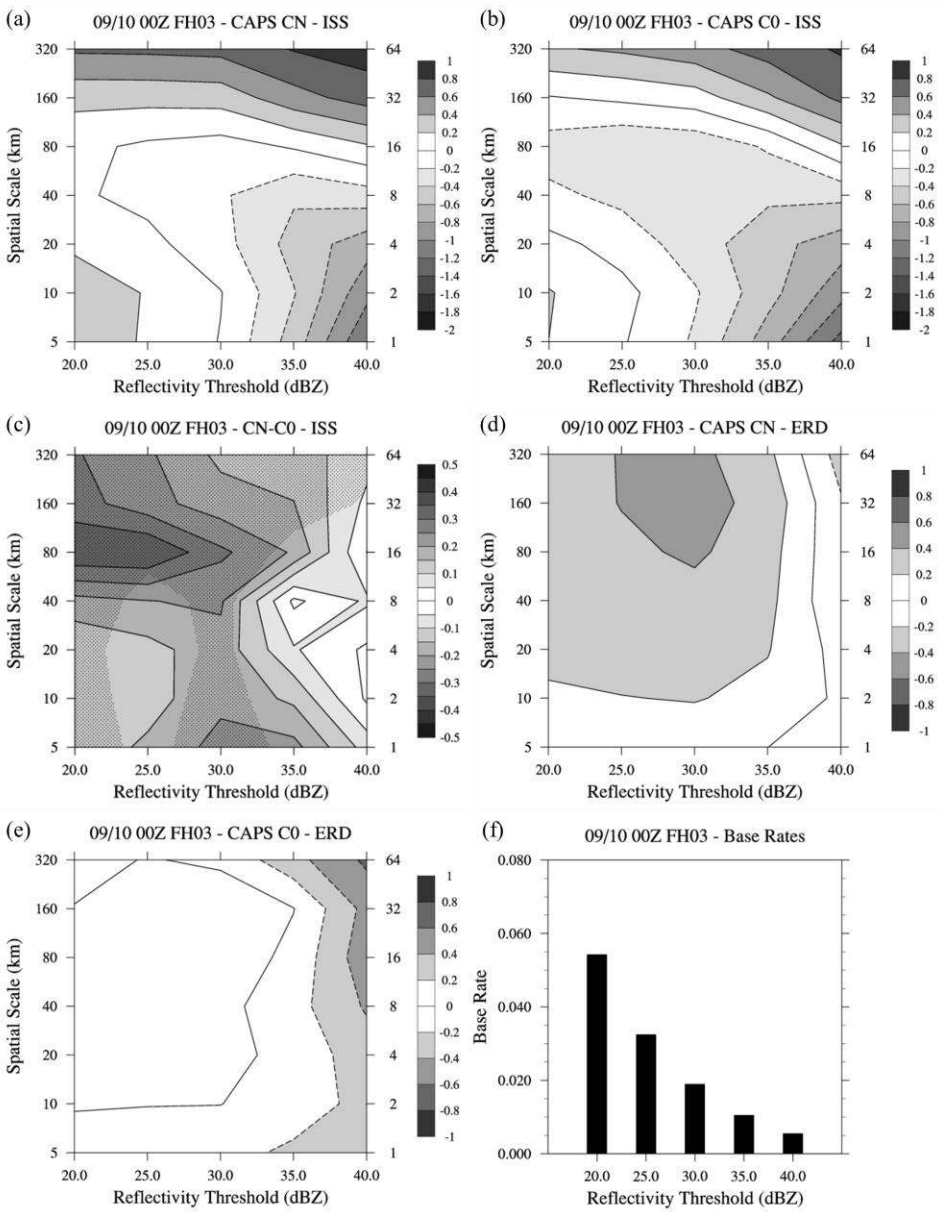


Fig. 13. ISS for 2009/2010 0000-UTC (a) CAPS CN and (b) CAPS C0 at FH 1 for reflectivity thresholds every 5 dBZ from 20 dBZ to 40 dBZ and spatial scales from 5 km to 320 km. Grey shading with solid contours in (a) and (b) represent positive skill, and grey shading with negative contours in (a) and (b) represent negative skill. ISS differences between CN and C0 are shown in (c), where grey shading with solid contours represent $ISS_{CN} > ISS_{C0}$, grey shading with dashed contours (not depicted in these plots) represent $ISS_{CN} < ISS_{C0}$, and stippling represents 95% statistical significance. ERD values for (d) CN and (e) C0, where grey shading with solid contours in (d) and (e) represent overforecasting, and grey shading with dashed contours in (d) and (e) represent underforecasting. Values along the right-ordinate represent multiples of grid-spacing. (f) Base rates of observed reflectivity for each threshold.



867
868
869
870 *Fig. 14. Same as Fig. 13, except for FH 3.*
871
872

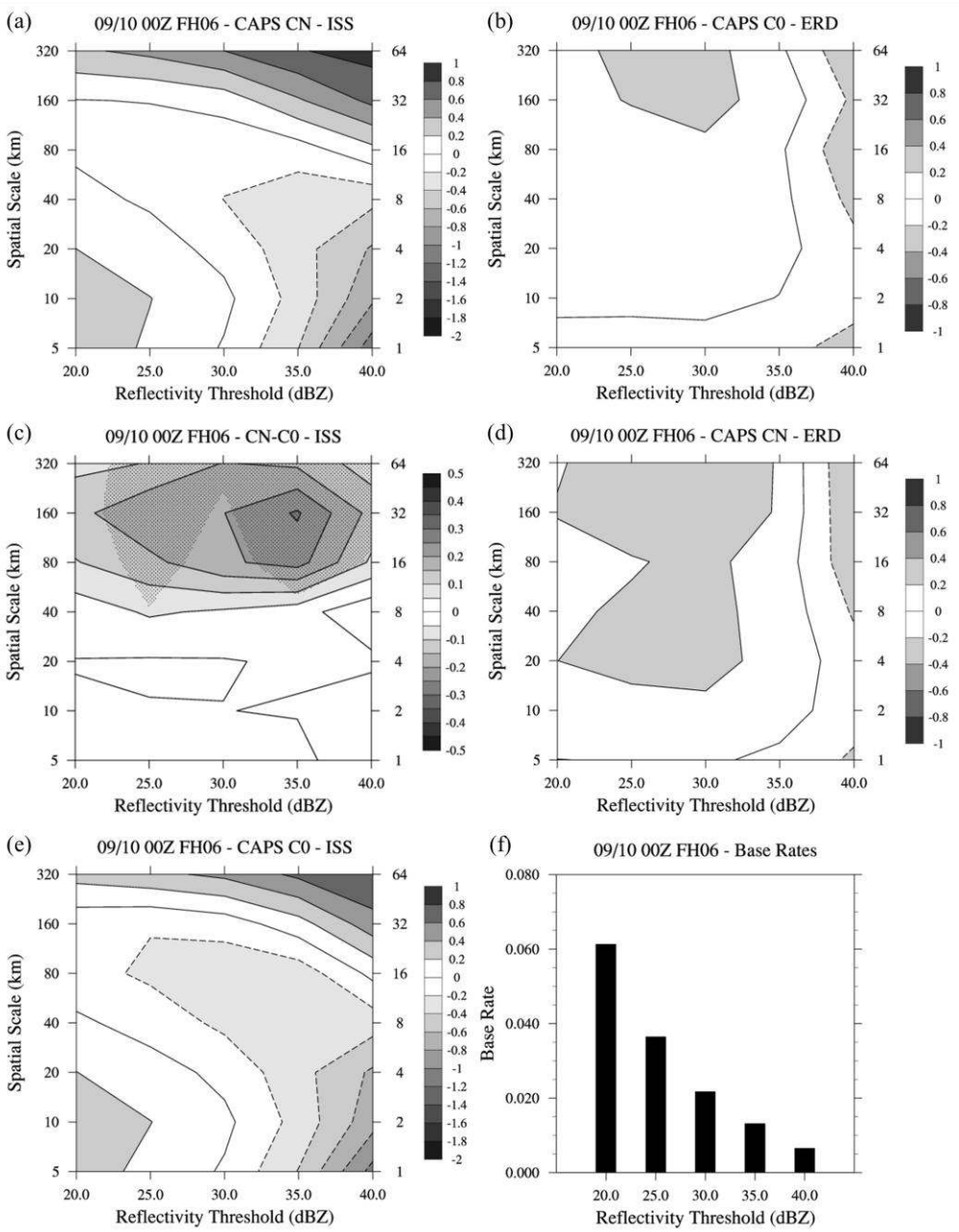


Fig. 15. Same as Fig. 13, except for FH 6.

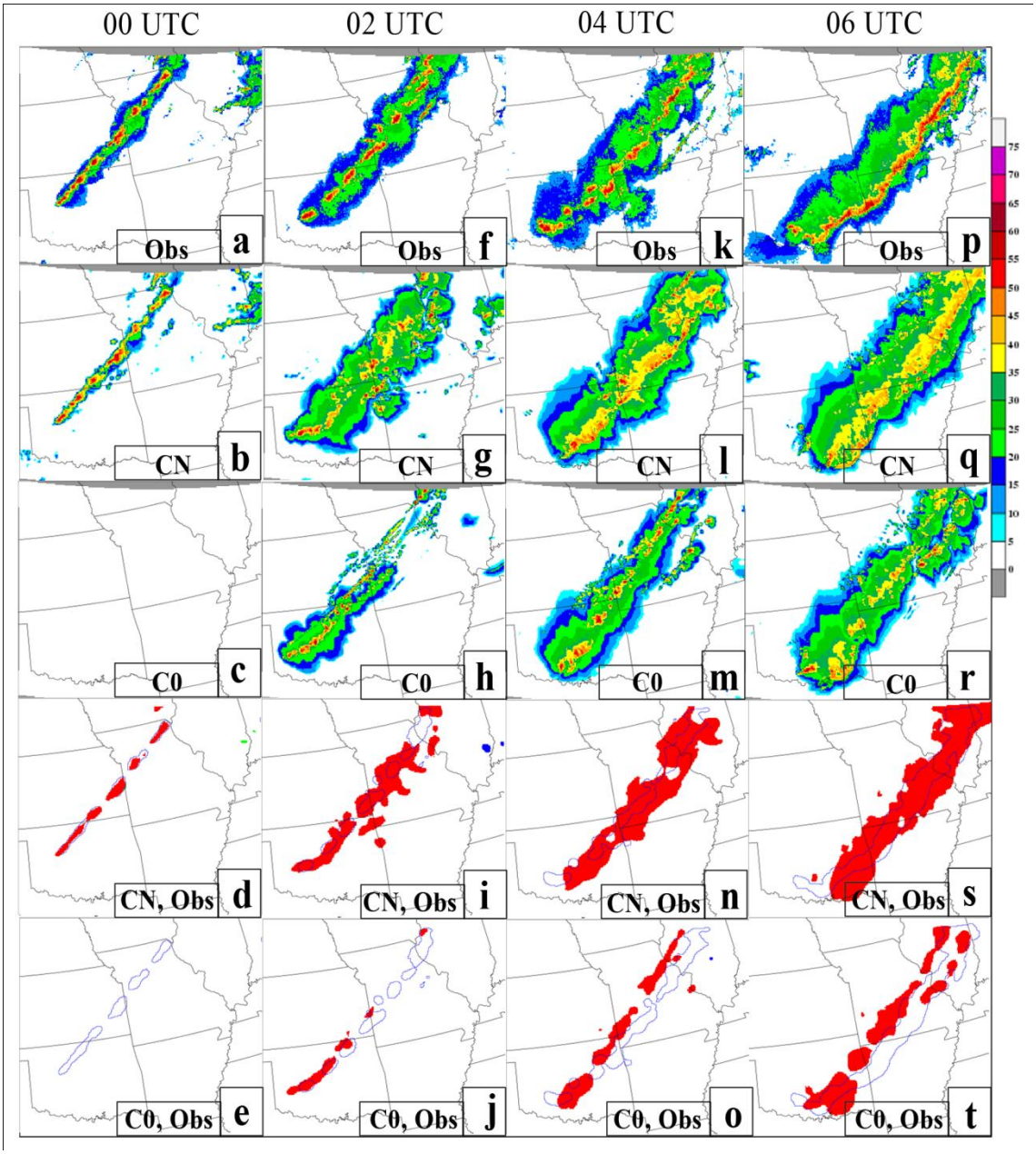


Fig. 16. First row is observed composite reflectivity from 14 May 2009 for (a) 0000 UTC, (f) 0200 UTC, (k) 0400 UTC, and (p) 0600 UTC. Second and third rows are simulated reflectivity forecasts from 0000-UTC CAPS CN (b, g, l, and q) and CAPS C0 (c, h, m, and r) for the same times. In the bottom two rows, 30-dBZ thresholded observed reflectivity is marked by the thin blue line (d/e, i/j, n/o, and s/t). Red shading represents 30-dBZ thresholded simulated reflectivity for CN (d, i, n, and s) and C0 (e, j, o, and t). From Kain et al (2010).