AMERICAN
METEOROLOGICAL
SOCIETY

Weather and Forecasting

EARLY ONLINE RELEASE

This is a preliminary PDF of the author-produced manuscript that has been peer-reviewed and accepted for publication. Since it is being posted so soon after acceptance, it has not yet been copyedited, formatted, or processed by AMS Publications. This preliminary version of the manuscript may be downloaded, distributed, and cited, but please be aware that there will be visual differences and possibly some content differences between this version and the final published version.

The DOI for this manuscript is doi: 10.1175/2010WAF2222404.1

The final published version of this manuscript will replace the preliminary version at the above DOI once it is available.

© 2010 American Meteorological Society

Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF model simulations and the operational NAM

Adam J. Clark[*1], William A. Gallus, Jr.[2], and Morris L. Weisman[3]

[1] *NOAA/National Severe Storms Laboratory*
Norman, OK, 73071

[2] *Dept. of Geological and Atmospheric Sciences*
Iowa State University
Ames, IA, 50010

[3] *National Center for Atmospheric Research*
Boulder, CO, 80307

Revised manuscript submitted to
Weather and Forecasting
March 2010

Corresponding author's address:
Adam J. Clark
National Weather Center, NSSL/FRDD
120 David L. Boren Blvd.,
Norman, OK, 73072
email: adam.clark@noaa.gov
Phone: 405-325-6731

PRELIMINARY ACCEPTED VERSION

Abstract

Since 2003 the National Center for Atmospheric Research (NCAR) has been running various experimental convection-allowing configurations of the Weather Research and Forecasting model for domains covering a large portion of the central US during the warm season (April – July). In this study, the skill of 3-hourly accumulated precipitation forecasts from a large sample of these convection-allowing simulations conducted during 2004-2005 and 2007-2008 is compared to that from operational North American Mesoscale (NAM) model forecasts using a neighborhood-based Equitable Threat Score (ETS). Separate analyses were conducted for simulations run before and after the implementation in 2007 of positive-definite (PD) moisture transport for the NCAR-WRF simulations. The neighborhood-based ETS [denoted $\langle ETS \rangle_r$] relaxes the criteria for "hits" (i.e., correct forecasts) by considering grid-points within a specified radius r. $\langle ETS \rangle_r$ is more useful than the traditional ETS because $\langle ETS \rangle_r$ can be used to diagnose differences in precipitation forecast skill between different models as a function of spatial scale, whereas the traditional ETS only considers the spatial scale of the verification grid.

It was found that differences in $\langle ETS \rangle_r$ between NCAR-WRF and NAM generally increased with increasing r, with NCAR-WRF having higher scores. Examining time series of $\langle ETS \rangle_r$ for r = 100 and r = 0 km (which simply reduces to the "traditional" ETS), statistically significant differences between NCAR-WRF and NAM were found at many forecast lead times for $\langle ETS \rangle_{100}$ but only a few times for $\langle ETS \rangle_0$. Larger and more statistically significant differences occurred with the 2007-2008 cases relative to the 2004-2005 cases. Because of differences in model configuration and dominant large scale weather regime, a more controlled experiment would have been needed to diagnose the reason for the larger differences that occurred with the 2007-2008 cases. Finally, a compositing technique was used to diagnose differences in the spatial distribution of the forecasts. This technique implied westward displacement errors for NAM model forecasts in both sets of cases and in NCAR-

WRF model forecasts for the 2007-2008 cases. Generally, the results are encouraging because they imply that advantages in convection-allowing relative to convection-parameterizing simulations noted in recent studies are reflected in an objective neighborhood-based metric.

1. Introduction

Deficiencies in warm-season forecasts of deep moist convection have been linked to the use of cumulus parameterization (CP; e.g., Davis et al. 2003; Liu et al. 2006; Clark et al. 2007, 2009) which is necessary to depict the effects of sub-grid scale convective processes (e.g., Molinari and Dudek 1992). Thus, it is widely believed that significant improvements in warm season forecasts of convection may be obtained by using grid-spacing small enough to treat convective processes explicitly (e.g., Fritsch and Carbone 2004). However, reduction to convection-allowing grid-spacing[1] comes with considerable computational expense. For example, because of a time-step reduction and 3-D increase in number of grid-points, a decrease in grid-spacing by a factor of n requires an increase in computational expense by a factor of ~ n^3 (e.g., a decrease from 12- to 4-km would require $3^3 = 27$ times more computation time). Because of the increase in computational expense, it is very important to consider whether sufficient value is actually gained from a grid-spacing reduction (e.g., Weisman et al. 2008; Kain et al. 2008).

Further complicating decisions on whether or not to decrease grid-spacing are the increasing difficulties associated with using traditional (i.e., "point to point") metrics to evaluate forecasts that contain increasingly fine scale and high amplitude features (e.g., Baldwin et al. 2001; Gallus 2002; Mass et al. 2002; Ebert 2008; Gilleland et al. 2009) . These difficulties arise because, even when these fine scale features are realistically predicted in a model, slight displacement errors often result in "double penalties" (i.e., observed-but-not-forecast and forecast-but-not-observed errors), which occur more frequently with increasing resolution (Ebert 2008). Because of these double penalties, subjective forecast evaluations are often not consistent with objective metrics (e.g., Kain et al. 2003) and it is very difficult to assess the true quality of high resolution guidance.

1 In this study, the term "convection-allowing" is used to refer to simulations using the maximum grid-spacing (or below) at which convection can be treated explicitly and mid-latitude MCSs can be adequately resolved, which is generally thought to be ~ 4-km based on Weisman et al. (1997).

The ineffectiveness of traditional metrics has led to application of alternative verification strategies for high resolution guidance that aim to provide more useful information on spatial structure and presence of features in forecast fields. Some of these strategies have involved purely subjective approaches in which the quality of forecasts was rated based on visual inspection by human forecasters. For example, Weisman et al. (2008) ranked forecasts "good", "bad", and "okay" based on specified criteria for correspondence of observed and forecast convective events. Other strategies have involved combining subjective and objective methods (e.g., Done et al. 2004; Weisman et al. 2008) by manually categorizing possible forecast outcomes for objects (e.g., MCSs) or object attributes (e.g., MCS mode) into standard 2x2 contingency table elements (Wilks 1995) and then computing commonly used traditional metrics. Finally, numerous recent studies have developed purely objective non-traditional metrics that can generally be categorized as feature-based (e.g., Ebert and McBride 2000; Davis et al. 2006), scale-decomposition (e.g., Casati et al. 2004), or neighborhood-based approaches (Roberts and Lean 2008; Schwartz et al. 2009; Ebert 2009); see Casati et al. (2008) for a review. The goal of all these strategies is to develop measures that accurately reflect the skill and usefulness of forecasts as perceived by human forecasters.

The purpose of this study is to demonstrate use of a neighborhood-based Equitable Threat Score (ETS; Schaefer 1990) to compare precipitation forecasts from experimental convection-allowing WRF model simulations conducted April-July 2004-2008 by the National Center for Atmospheric Research (NCAR) to operational North American Mesoscale (NAM; Janjic 2003) model forecasts that use cumulus parameterization. The NAM model forecasts were used for initial and lateral boundary conditions (ICs/LBCs) for most of the convection-allowing forecasts. Neighborhood-based approaches consider values at grid-points within a specified radius (i.e., "neighborhood") of an observation. Values within the specified radius are considered equally likely estimates of the true value. The specified radius can be viewed as the amount of displacement error allowed before the forecast is

considered “wrong”. Neighborhood-based approaches have been shown to be particularly useful because varying the size of the neighborhood allows for a diagnosis of skill at different spatial scales (e.g., Ebert 2009). Previous works using subjective verification strategies have found that NCAR’s convection-allowing forecasts better predict MCS frequency and mode (Done et al. 2004) but fail to find improvements relative to convection-parameterizing forecasts in the broader characteristics of convective systems such as location, timing and relative intensity (Weisman et al. 2008). This failure is somewhat surprising given the improved model climatology of precipitation from convection-allowing relative to convection-parameterizing simulations inferred from comparisons of time-longitude diagrams (e.g., Clark et al. 2007, 2009; Weisman et al. 2008). This study examines whether an improvement in precipitation forecasts in convection-allowing relative to convection-parameterizing forecasts is reflected by a neighborhood-based ETS, and at what spatial scales any improvements are observed. This paper is organized as follows: data and methodology are provided in section 2, results are presented in section 3, and a summary and discussion is given in section 4.

2. Data and methodology

Three-hourly accumulated precipitation forecasts from convection-allowing WRF model simulations (3 to 4 km grid-spacing) conducted by NCAR using the ARW dynamics core (NCAR-WRF, hereafter) are examined. These forecasts were initialized at 0000 UTC and integrated 36 h for domains over the central US during April-July 2004-2005 and 2007-2008 (data from 2006 were not available) and compared with forecasts from NCEP’s operational NAM model. The NAM model forecasts were used as ICs/LBCs in the NCAR-WRF simulations before 2008. For the 2008 simulations, the WRF three-dimensional variational data assimilation (WRF-Var; Barker et al. 2004) system was used at 9 km grid-spacing to create a 1200 UTC analysis. Then, 3-hourly assimilation cycles were used until 0000 UTC to create the NCAR-WRF ICs, and forecasts from NCEP's Global

Forecast System (GFS; Environmental Modeling Center 2003) model were used as LBCs.

In addition to changes in initialization procedure, other aspects of the NCAR-WRF model configuration also changed from year to year based on experiences from previous years (e.g., domain, WRF version, physics parameterizations). These changes are summarized in Table 1. Microphysics parameterizations used in NCAR-WRF included the Lin [derived from Lin et al. (1983)], WRF single-moment six-class (WSM-6; Hong and Lim 2006), and Thompson et al. (2004) schemes. Boundary layer parameterizations included the Yonsei University (YSU; Noh et al. 2003) and Mellor-Yamada-Janjic (MYJ; Mellor and Yamada 1982; Janjic 2002) schemes. The Oregon State University land surface model (OSU LSM; Chen and Dudhia 2001) was used during 2004, and the NOAH land surface model, the successor to the OSU LSM, was used after 2004. For cases since 2005, the NCAR High-Resolution Land Data Assimilation System (HRLDAS; Chen et al. 2007) was employed. Physics parameterizations that were not varied in NCAR-WRF during the period included the rapid radiative transfer model (RRTM) for long-wave radiation (Mlawer et al. 1997), and the Dudhia (1989) scheme for short-wave radiation. Sensitivity tests conducted by Weisman et al. (2008) for changes in the NCAR-WRF model configurations made during the 2003-2005 period found only small changes in overall forecast accuracy. In addition, exclusion of the 18 cases from 2008 that used different ICs/LBCs did not have any significant impacts on our results (not shown).

The change in NCAR-WRF model configuration most likely to noticeably impact our results is the use of positive-definite (PD) moisture transport (Skamarock 2006) which was used for the 2007-2008 cases, but not the 2004-2005 cases. Examining some of the convection-allowing precipitation forecasts examined herein, Skamarock and Weisman (2009) found that using PD moisture transport significantly reduced large positive biases in precipitation forecasts relative to simulations that did not use PD moisture transport, especially for high precipitation thresholds. Thus, separate analyses are conducted for the cases that use and do not use PD moisture transport.

Changes were also made to the NAM model during the period examined. The most important change was the transition from using the Eta (Janjic 1994) to the WRF-NMM (Nonhydrostatic Mesoscale Model; Janjic 2003) in June 2006, which also came with a change in data assimilation systems from Eta 3D Variational Analysis (EDAS; Parrish et al. 1996) to Gridpoint Statistical Interpolation (GSI; Wu eta al. 2002). Furthermore, although there were not any major switches in physics parameterizations accompanying the transition from Eta to WRF-NMM in June 2006, minor improvements were made to many of the individual physics schemes and additional improvements were made to the cumulus and microphysics parameterizations in Dec. 2006. Further details on NAM model updates can be found at the NCEP website http://www.emc.ncep.noaa.gov/mmb/mmbpll/eric.html#TAB4. NAM model physics include the MYJ boundary layer parameterization, NOAH land surface model, Ferrier et al. (2002) microphysics, Betts-Miller-Janjic (BMJ; Betts 1986; Betts and Miller 1986; Janjic 1994) cumulus parameterization, and GFDL shortwave (Lacis and Hansen 1975) and longwave (Fels and Schwarzkopf 1975; Schwarzkopf and Fels 1985) radiation parameterizations.

The cases were chosen based on the availability of data. NAM model forecasts were obtained from the NOAA National Operational Model Archive and Distribution System (NOMADS; http://nomads.ncdc.noaa.gov), while NCAR-WRF forecasts were obtained from NCAR's Mass Storage System. There are 199 (95) cases analyzed in the 2004-2005 (2007-2008) period without (with) PD moisture transport (Fig. 1b).

To verify the precipitation forecasts, NCEP's stage IV (Baldwin and Mitchell 1997) multi-sensor rainfall estimates are used, which are available at 1-hourly accumulation intervals on a 4-km polar stereographic grid. The stage IV data were obtained from the NCEP website http://www.emc.ncep.noaa.gov/mmb/ylin/pcpanl/stage4/. The stage IV data, as well as the NAM and NCAR-WRF model data are remapped to a common 20-km grid covering the central US (Fig. 1a)

using a neighbor-budget interpolation that conserves the total liquid volume in the domain (e.g, Accadia et al. 2003).

The traditional method for computing ETS uses a 2x2 contingency table of possible forecast outcomes at individual grid-points where the table elements are *hits* (correct forecast of an event), *misses* (observed but not forecast event), *false alarms* (forecast but not observed event), and *correct negatives* (correct forecast of non-event; e.g., Wilks 1995). Using these elements, ETS is expressed as:

$$ETS = \frac{hits - chance}{hits + misses + false\ alarms - chance} \tag{1}$$

where,

$$chance = \frac{(hits + misses)(hits + false\ alarms)}{hits + misses + correct\ negatives + false\ alarms} \tag{2}$$

The ETS can be interpreted as the fraction of correctly predicted observed events, adjusted for hits associated with random chance. A perfect ETS is 1.0, while -1/3 is the lower limit and 0.0 is the threshold for no skill. For computation of a neighborhood-based ETS, the criteria for a hit is relaxed by considering adjacent grid-points within a specified radius of each grid-point (see Ebert 2009 for a similar application of a neighborhood-based ETS). If an event is observed at a grid-point, this grid-point is a hit if the event is forecast at the grid-point or at any grid-point within a circular radius *r* of this observed event. Similarly, if an event is forecast at a grid-point, the grid-point is a hit if an event is observed at the grid-point or at any grid-point within *r* of this forecast event. A miss is assigned when an event is observed at a grid-point and none of the grid-points within *r* forecast the event, and false alarms are assigned when an event is forecast at a grid-point and not observed within *r* of the forecast.

Correct negatives are assigned the same way as for the traditional ETS computation (i.e., an event is neither forecast nor observed at a single grid-point). Average ETSs were computed by summing (i.e., aggregating) contingency table elements over all cases. The resampling methodology described in Hamill (1999) was used to determine whether differences in ETS were statistically significant (α = 0.05; resampling repeated 10000 times). For application to this study, the Hamill (1999) method involves computing a test statistic using the difference in ETS between NAM and NCAR-WRF at each rainfall threshold, forecast hour, and neighborhood radius using contingency tables elements summed over all cases. Then, a distribution of resampled test statistics is created by randomly choosing NAM or NCAR-WRF on each day and summing contingency table elements. The location of the test statistic within the distribution of resampled test statistics determines whether the differences are statistically significant.

In this study, radii of 0, 20, 40, 60, 80, 100, 130, 160, 190, 220, and 250 km are used, and $\langle ETS \rangle_r$ denotes neighborhood-based ETS computed at radius r. Note that $\langle ETS \rangle_0$ simply reduces to the traditional form of ETS. We chose to restrict our analysis to radii at or below 250 km because these are approximately the largest scales at which convective systems occur and it was very computationally expensive to examine higher radii. Note, as long as at least one observed and forecast point are present anywhere on the domain, as r approaches the domain size, $\langle ETS \rangle_r$ approaches 1.0. For reference, circles with radii of 100 and 250 km overlay the analysis domain in Fig. 1a.

3. Results

$\langle ETS \rangle_r$ (r = 0, 20, …, 250 km) is shown for NCAR-WRF and NAM models at forecast hours 18, 24, and 30 for 0.10-, 0.25-, 0.50-, and 1.00-in rainfall thresholds in Figures 2 and 3, corresponding to the 2004-2005 and 2007-2008 cases, respectively. Because ETS can reward forecasts that have higher biases relative to other forecasts (e.g., Baldwin and Kain 2006), ETSs for raw forecasts along

with bias-adjusted forecasts are shown. The bias-adjustment was implemented using a procedure based on probability matching (Ebert 2001) described by Clark et al. (2009). Basically, this procedure reassigns the distribution of forecast precipitation with that of the observed precipitation resulting in forecast precipitation fields that have the same spatial pattern as the raw forecasts, but with amplitudes exactly matching the observations. The bias-adjustment was applied to the 3-hr accumulation periods and results in a perfect bias of 1.0 for all precipitation thresholds. In addition to allowing a more equitable comparison between models, Mesinger 2008 suggest that bias-adjustment may allow position errors to be more cleanly detected, and Jenkner et al. 2008 note that bias-adjustment allows the quantiles of the forecast and observed distributions to be better compared. Because this bias-adjustment requires the verifying observations, it should not be viewed as a post-processing method. Rather, the bias-adjustment simply allows for a more equitable comparison between models. To obtain the maximum value from the forecasts in a real-time forecasting environment, post-processing methods like those described by Applequist et al. (2002), Eckel and Mass (2005), or Glahn et al. (2009) should be applied.

At the lightest thresholds examined (0.10-in; Figs. 2 and 3a-c), the bias-adjustment did not have a noticeable impact. However, at higher thresholds, especially 0.50- and 1.00-in (Figs. 2 and 3g-l), bias-adjustment resulted in improved $\langle ETS \rangle_r$ with the greatest improvements at the highest *r* examined. Furthermore, the NAM forecasts benefited the most from bias-adjustment, which was likely related to the biases in the raw forecasts. These biases were computed over the analysis domain (Fig. 1b) using the standard formulation for bias (i.e. F/O where F is the number of forecast grid-points above a specified threshold and O the number of observed) and are shown in Figs. 2-3. In NAM, the raw forecasts for the higher thresholds had biases well below 1.0 so that artificially increasing the bias required increasing the number of forecast events (i.e., grid-points with forecasts above a specified threshold). The resulting increase in $\langle ETS \rangle_r$ from NAM implies that many of the additional forecast

events obtained through bias-correction became *hits*. However, in NCAR-WRF, the raw forecasts for the higher thresholds had biases above 1.0 so that artificially decreasing the bias required decreasing the number of forecast events. Thus, the resulting increase in $\langle ETS \rangle_r$ from NCAR-WRF implies that many of the forecast events that were removed through bias correction had previously been *false alarms*.

For 2007-2008 NCAR-WRF forecasts (Fig. 3), bias-correction has a noticeably smaller impact than in 2004-2005 forecasts (Fig. 2). The smaller impact results from the use of PD moisture transport which reduces the biases in the raw 2007-2008 forecasts, especially for the highest rainfall thresholds, confirming the results of Skamarock and Weisman (2009). Hereafter, only results using bias-adjusted $\langle ETS \rangle_r$ are discussed/shown. However, from a practical perspective, it should be emphasized that the rainfall amounts from the raw forecasts are still very important. In NAM, the very low biases at the higher rainfall thresholds imply that NAM simply cannot produce heavy enough precipitation, and the differences in $\langle ETS \rangle_r$ between NAM and NCAR-WRF that occur even after bias-correction is applied suggest more severe location and/or timing errors in NAM.

Generally, values of $\langle ETS \rangle_r$ in NCAR-WRF and NAM were nearly identical at the smallest *r* examined (Figs. 2 and 3), consistent with comparisons of traditional ETS made by Done et al. (2004) for NCAR-WRF and NAM simulations during 2003. However, as *r* increased, differences in $\langle ETS \rangle_r$ between NCAR-WRF and NAM began to increase with NCAR-WRF having the higher values. These differences were larger for the 2007-2008 cases (Fig. 3) than in the 2004-2005 cases (Fig. 2). For some of the lighter rainfall thresholds examined (e.g., 0.10- and 0.25-in), the differences became smaller again at the largest radii, which was not unexpected considering that $\langle ETS \rangle_r$ from both models eventually converges to 1.0 as *r* increases. In addition to simply comparing forecast skill at different spatial scales in NCAR-WRF and NAM, Figures 2 and 3 also allow for some practical information regarding predictability to be inferred. For example, if an ETS of 0.5 is arbitrarily chosen as a

threshold for a skillful forecast, then the radius at which the ETS reaches 0.5 can be regarded as the minimum spatial scale at which skillful forecasts are obtained. For instance, in Figure 3g (forecast hour 18; 0.50-in rainfall threshold) NCAR-WRF forecasts are "skillful" (i.e., ETS ≥ 0.50) at scales down to about 130 km, but for NAM the minimum spatial scale for a skillful forecast is about 220 km.

To further illustrate the differences in $\langle ETS \rangle_r$ between NCAR-WRF and NAM, time series of $\langle ETS \rangle_0$ and $\langle ETS \rangle_{100}$ for forecast hours 12-36 at the same rainfall thresholds shown in Figures 2 and 3, are plotted in Figure 4 with times at which statistically significant differences occur highlighted. Values of $\langle ETS \rangle_0$ are much lower than $\langle ETS \rangle_{100}$ and, perhaps more importantly, the differences between $\langle ETS \rangle_r$ in NCAR-WRF and NAM with r = 100 km are much more noticeable than with r = 0 km. The more noticeable differences are reflected by the number of times at which statistically significant differences occur. For example, at the 0.50-in threshold in the 2007-2008 cases (Fig. 4g), all but one of the forecast hours examined for $\langle ETS \rangle_{100}$ has significant differences with NCAR-WRF having the higher values, while only forecast hour 18 contains significant differences for $\langle ETS \rangle_0$. Furthermore, there were larger differences and more forecast hours with significant differences for the 2007-2008 cases (Figs. 4e-h) compared to the 2004-2005 cases (Figs. 4a-d). For the 1.00-in rainfall threshold (Figs. 4d and h), the number of times at which significant differences occurred for $\langle ETS \rangle_{100}$ drops off relative to lighter thresholds because of a relatively sharp decrease in sample size (not shown).

While $\langle ETS \rangle_r$ provides a method to compare the performance of NCAR-WRF and NAM at different spatial scales, it still does not provide any information on what actually causes the observed differences. However, the method for computing $\langle ETS \rangle_r$ does allow for some inference of where observations tend to occur relative to forecasts, which gives some useful diagnosis of forecast errors. For each forecast hour and rainfall threshold examined, when contingency table elements were being compiled for $\langle ETS \rangle_{250}$, the distribution of observed events relative to each forecast event (i.e., the

conditional distribution of observed events) within r = 250 km of each forecast event was derived. Then, composite conditional distributions of observed events were simply obtained by summing the conditional distributions over all cases for each forecast hour and rainfall threshold. This method is basically a simplified version of the composite verification approach proposed by Nachamkin (2004). However, in our method, single grid-points that exceed a specified rainfall threshold are treated as events, whereas Nachamkin (2004) treats contiguous regions over which a criteria is met as single events and then derives the conditional distribution of observed events relative to the geometric center of the contiguous region. Because our method uses single grid-points to define events, information such as the shape and orientation that is retained using the Nachamkin (2004) composite method is lost in our procedure. However, our method is advantageous because of its simplicity and information regarding displacement errors is still retained. For the method used herein, the observed frequency (OF_i) for each grid-point within the conditional distribution of observed events can be expressed as:

$$OF_i = \sum_{x=1}^{N} \left[\sum_{i=1}^{n} e_{i,x} \right] \tag{3}$$

where the subscript x represents the xth grid-point in the analysis domain (N = 9108), the subscript i represents the ith grid-point within r = 250 km of the xth grid-point (n = 489), and $e_{i,x}$ is 1 if an event is forecast at grid-point x and observed at grid-point i, while $e_{i,x}$ is 0 if those conditions are not met. Thus, OF_i is obtained by looping over all the grid-points in the domain (i.e., summation outside of brackets in Eq. 3), and then for each grid-point in the domain that forecasts an event, looping over all grid-points within r = 250 km (i.e., summation inside of brackets in Eq. 3). A schematic for a neighborhood with r = 100 km for grid-point x = 462 is shown in Figure 5 with the gray shaded grid-boxes indicating where $e_{i,x}$ = 1.0.

The composite conditional distributions of observed events for the 0.50-in rainfall threshold at forecast hours 12-36 for NCAR-WRF and NAM as well as their differences (i.e., NCAR-WRF minus NAM) are shown in Figures 6 and 7 for the 2004-2005 and 2007-2008 cases, respectively. To obtain an equitable comparison between both sets of cases, the observed relative frequency for 0.50-in is shown in Figures 6 and 7 (i.e., the observed frequencies are normalized by the number of cases). Thus, the units for OF_i are simply "counts"/case. Note that the strong variations in relative frequencies across forecast hours 12 to 36 reflect the diurnal cycle of rainfall over the analysis domain, which exhibits a minimum near forecast hours 18-21 and a maximum at forecast hour 30.

For the 2004-2005 cases (Fig. 6), the most noticeable differences between NCAR-WRF and NAM occur at forecast hours 27-36 (Figs. 6f-i, o-r, and x-α), times which immediately surround the diurnal peak in rainfall over the analysis domain. At these forecast hours, there is a tendency in NAM for observations to occur east of the forecasts, and at forecast hours 30 and 33 there is a slight southward component to the displacement of observations from forecasts. In other words, the NAM forecasts for areas of rainfall greater than 0.50-in at forecast hours 27-36 tend to have westward and northward displacement errors. In NCAR-WRF, the distributions of corresponding observed relative to forecast grid-points are more uniformly distributed within the composite domain than in NAM. However, the observed frequencies in NAM are larger than WRF, which likely "compensates" for the spatial errors and results in the relatively small differences in $\langle ETS \rangle_{100}$ between NAM and NCAR-WRF at these times (Fig. 4c). These spatial errors in NAM are consistent with problems depicting the zonal (west-to-east) movement of MCS-related precipitation in convection-parameterizing simulations implied by analyses of time-longitude (or Hovmoller) diagrams by Davis et al. (2003), Clark et al. (2007, 2009), and Weisman et al. (2008). Furthermore, using a modified version of the entity-based Ebert-McBride technique, Grams et al. (2006) also found north and west displacement errors for convective systems in three different 12-km grid-spacing model configurations used during IHOP

(International H2O Project; Weckwerth et al. 2004). Finally, the results are also consistent with those of Wang et al. (2009) finding westward displacement errors in NAM forecasts of precipitation areas related to mid-tropospheric perturbations over the central US.

For the 2007-2008 cases (Fig. 7), the most noticeable differences between NCAR-WRF and NAM also occur at the later forecast hours, similar to the 2004-2005 cases. However, unlike the 2004-2005 cases, *both* NCAR-WRF and NAM have a tendency for observations to occur east of the forecasts (i.e., westward displacement errors). In addition, the larger frequencies of observed relative to forecast grid-points in NCAR-WRF explain the larger $\langle ETS \rangle_{100}$ values in NCAR-WRF at these times (Fig. 4g).

Finally, to better understand how differences between $\langle ETS \rangle_r$ in NCAR-WRF and NAM varied among the 294 individual cases, the distribution of $\langle ETS \rangle_{100}$ differences (i.e., NCAR-WRF minus NAM) at forecast hour 30 for the 0.50-in rainfall threshold is shown in Figure 8. Clearly, values of $\langle ETS \rangle_{100}$ in NCAR-WRF are more often higher than NAM (mostly contributed by the 2007-2008 cases), however, there are a sizable fraction of cases in which $\langle ETS \rangle_{100}$ for NAM was higher than NCAR-WRF. Examples of precipitation forecasts representing different portions of the distribution in Figure 8 are shown in Figure 9. The purpose of showing these examples is to examine whether subjective impressions of the precipitation fields are consistent with the neighborhood-based objective measure $\langle ETS \rangle_{100}$. The "WRF better" cases (Figs. 9a-l) show the forecast and observed precipitation fields for the three cases in which NCAR-WRF had higher $\langle ETS \rangle_{100}$ than NAM by the largest amounts at forecast hour 30 for the 0.50-in threshold; the "NAM better" cases (Figs. 9m-x) are similar except for those cases in which NAM had the higher $\langle ETS \rangle_{100}$ by the largest amounts; and the "WRF and NAM about the same" cases (Figs. 9w-θ) are the cases in which NCAR-WRF and NAM $\langle ETS \rangle_{100}$ differences were smallest.

For the 06 May 2007 (Figs. 9a-d) and 23 May 2008 (Figs. 9e-h) cases, it is clear that NCAR-WRF did a much better job relative to NAM in forecasting the general regions over which 3-hourly

precipitation accumulations exceeded 0.50-in. For these two cases, the superior performance of NCAR-WRF is also reflected by values of $\langle ETS \rangle_{100}$. In contrast, for the 31 May 2005 case (Figs. 9i-l), the NCAR-WRF and NAM forecasts appear very similar, while values of $\langle ETS \rangle_{100}$ imply NCAR-WRF had a much better forecast. From the overlay in Fig. 9l, it appears NCAR-WRF scored so much higher than NAM because NCAR-WRF predicted areas of precipitation greater than 0.50-in over north-central Texas and Alabama that generally matched observed areas. Although NAM also predicted precipitation over these regions, the NAM forecast amounts did not quite exceed 0.50-in leading to the large differences.

For the "NAM better" cases (Figs. 9m-x), $\langle ETS \rangle_{100}$ accurately reflected our subjective evaluation that NAM had better forecasts than NCAR-WRF. The 16 June 2005 case (Figs. 9u-x) was notable in that NAM was able to score a high $\langle ETS \rangle_{100}$ because it correctly forecast heavy precipitation in Oklahoma that corresponded to an observed MCS, while NCAR-WRF completely missed the event.

Finally, for the "WRF and NAM about the same" cases (Figs. 9w-θ), our subjective evaluation of the forecasts was generally consistent with $\langle ETS \rangle_{100}$. However, it was clear that the similar $\langle ETS \rangle_{100}$ in NCAR-WRF and NAM did not result from similar forecasts. For these cases, it appeared that because both sets of forecasts had about the same number of forecast grid-points within the vicinity of observations, they scored very similarly.

4. Summary and discussion

A neighborhood-based ETS, $\langle ETS \rangle_r$, was used to compare precipitation forecast skill in convection-allowing simulations conducted by NCAR to convection-parameterizing NAM model forecasts. The comparison was made for the period April-July 2004-2005 (199 cases) and 2007-2008 (95 cases). $\langle ETS \rangle_r$ is computed by considering neighboring grid-points within radius, *r*, of each grid-point. Thus, by varying *r*, it is possible to examine how differences in precipitation forecast skill

between NCAR-WRF and NAM change according to the spatial scale. The most important results are summarized below.

At the smallest spatial scales examined (i.e., $\langle ETS \rangle_0$ which reduces to the traditional form of ETS in which no neighboring grid-points are considered), values of $\langle ETS \rangle_r$ in NCAR-WRF and NAM were nearly identical. However, as *r* was increased to scales of about 50 km and above, differences in $\langle ETS \rangle_r$ between NCAR-WRF and NAM became more pronounced, especially for rainfall thresholds greater than 0.25-in, with NCAR-WRF having higher values than NAM.

An examination of $\langle ETS \rangle_r$ time series for forecast hours 12-36 using *r* = 0 and *r* = 100 km revealed statistically significant differences between NCAR-WRF and NAM at many times analyzed for $\langle ETS \rangle_{100}$ and only a few times for $\langle ETS \rangle_0$. The 2007-2008 cases had larger differences and more forecast hours with statistically significant differences relative to the 2004-2005 cases. At rainfall thresholds higher than 0.50-in, the number of times with significant differences decreased relative to lighter rainfall thresholds because of a decrease in sample size.

By constructing composite distributions of observed events within *r* = 250 km of each grid-point with a forecast event it was shown that the most noticeable differences between NCAR-WRF and NAM occurred at forecast hours 27-36. At these times for the 2004-2005 cases, composite frequencies of observed relative to forecast events implied westward displacement errors in NAM forecasts, while the corresponding frequencies of observed events in NCAR-WRF were slightly smaller but much more uniformly distributed within the 250-km radius. For the 2007-2008 cases, composite observed frequencies implied that both NCAR-WRF and NAM tended to have westward and/or southward displacement errors. However, observations were more highly concentrated within *r* = 250 km in the NCAR-WRF forecasts than in NAM likely contributing to higher $\langle ETS \rangle_r$ in NCAR-WRF at these times.

Finally, it was found that a subjective comparison of forecast quality in NCAR-WRF and NAM

for nine selected cases was generally consistent with the differences in forecast quality implied by $\langle ETS \rangle_{100}$. For example, for cases in which NCAR-WRF had much higher $\langle ETS \rangle_{100}$ values than NAM, a simple visual inspection of the forecasts would also indicate that NCAR-WRF forecasts were better than those of NAM.

It is not clear what caused the differences in forecasts between the 2004-2005 and 2007-2008 cases. Although it is clear that PD moisture transport impacted the biases in the NCAR-WRF forecasts, other changes in model configuration (e.g, changes in PBL schemes or grid-spacing) may have also played a role. In addition, the two sets of cases contained slightly different portions of the warm season, and the dominant large scale weather pattern within the different years examined also varied. More controlled sensitivity tests which are beyond the scope of this study would have been necessary to attribute particular aspects of the models or types of cases to the differences between 2004-2005 and 2007-2008.

Generally, the results from this study are encouraging for convection-allowing simulations, as well as neighborhood-based verification strategies. Larger $\langle ETS \rangle_r$ in NCAR-WRF relative to NAM implies that advantages in convection-allowing relative to convection-parameterizing simulations noted in previous studies using subjective verification techniques to evaluate convective system frequency and mode (e.g., Done et al. 2004; Weisman et al. 2008) or analyses of "model climatology" (e.g., Clark et al. 2007, 2009; Weisman et al. 2008) are consistent with a neighborhood-based metric. Furthermore, the results could be perceived as contradictory to those in Weisman et al. (2008) that did not find noticeable differences in the broader characteristics of convective systems (e.g., location, timing, and relative intensity) between NCAR-WRF and NAM forecasts. However, note that the Weisman et al. study examined forecast hours 24-36 for the years 2003-2005, and for most of the same forecast hours in the current study for an overlapping time period [i.e., the 199 cases from 2004 – 2005 (Figs. 4a-d)] there were not statistically significant differences in forecast skill between NCAR-WRF and NAM.

Further testing of neighborhood-based, along with other "non-traditional", verification techniques is encouraged, along with applications to convection-allowing ensembles.

Acknowledgments

The authors would like to thank Wei Wang of NCAR for producing the NCAR-WRF forecasts. In addition, three anonymous reviewers provided many helpful comments that helped improve the manuscript. The majority of this project was supported through a National Research Council Post-doctoral Award for the first author under the guidance of David Stensrud. In addition, a portion of this work was supported by NSF Grant ATM-0848200.

References

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting*, **18**, 918–932.

Applequist, S., G. E. Gahrs, R. L. Pfeffer, and X. F. Niu, 2002: Comparison of methodologies for probabilistic quantitative precipitation forecasting. *Wea. Forecasting*, **17**, 783–799.

Baldwin, M. E., and K. E. Mitchell, 1997: The NCEP hourly multisensor U. S. precipitation analysis for operations and GCIP research. Preprints, *13th Conf. On Hydrology*, Long Beach, CA, Amer. Meteor. Soc., 54-55.

——, S. Lakshmivarahan, and J. S. Kain, 2001: Verification of mesoscale features in NWP models. Preprints, *Ninth Conf. on Mesoscale Processes*, Fort Lauderdale, FL, Amer. Meteor. Soc., 255-258.

——, J. S. Kain, 2006: Sensitivity of several performance measures to displacement error, bias, and event frequency. *Wea. Forecasting*, **21**, 636-648.

Barker, D. M., W. Huang, Y. R. Guo, A. J. Bourgeois, and Q. N. Xiao, 2004: A three-dimensional variational data assimilation system for MM5: Implementation and initial results. *Mon. Wea. Rev.*, **132**, 897–914.

Betts, A. K., 1986: A new convective adjustment scheme. Part I: Observational and theoretical basis. *Quart. J. Roy. Meteor. Soc*., **112**, 677-691.

——, and M. J. Miller, 1986: A new convective adjustment scheme. Part II: Single column tests using GATE wave, BOMEX, ATEX and Arctic air-mass data sets. *Quart. J. Roy. Meteor. Soc.*, **112**, 693-709.

Casati, B., G. Ross, D. B. Stephenson, 2004: A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteor. Appl*., **11**, 141-154.

——, L. J. Wilson, D. B. Stephenson, P. Nurmi, A. Ghelli, M. Pocernich, U. Damrath, E. E. Ebert, B. G. Brown, and S. Mason, 2008: Forecast verification: Current status and future directions. *Meteor. Appl*., **15**, 3-18.

Chen, F., and J. Dudhia, 2001: Coupling and advanced land surface-hydrology model with the Penn State-NCAR MM5 modeling system. Part I: Model implementation and sensitivity. *Mon. Wea. Rev*., **129**, 569-585.

——, K. W. Manning, M. A. LeMone, S. B. Trier, J. G. Alfieri, R. Roberts, M. Tewari, D. Niyogi, T. W. Horst, S. P. Oncley, J. B. Basara, and P. D. Blanken, 2007: Description and evaluation of the characteristics of the NCAR High-Resolution Land Data Assimilation System. *J. Appl. Meteorol. Climatol*., **46**, 694-713.

Clark, A. J., W. A. Gallus, and T. C. Chen, 2007: Comparison of the diurnal precipitation cycle in convection-resolving and non-convection-resolving mesoscale models. *Mon. Wea. Rev.*, **135**, 3456–3473.

——, ——, M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Wea. Forecasting*, **24**, 1121–1140.

Davis, C. A., K. W. Manning, R. E. Carbone, S. B. Trier and J. D. Tuttle, 2003: Coherence of warm season continental rainfall in numerical weather prediction models. *Mon. Wea. Rev.*, **131**, 2667-2679.

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

Done, J., C. A. Davis, and M. L. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the Weather Research and Forecast (WRF) Model. *Atmos. Sci. Lett*., **5**, 110-117.

Dudhia, J., 1989: Numerical study of convection observed during the Winter Monsoon Experiment using a mesoscale two-dimensional model. *J. Atmos. Sci.*, **46**, 3077–3107.

Ebert, E. E. and J. L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrology*, **239**, 179-202.

——, 2008: Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. *Meteor. Appl*., **15**, 51-64.

——, 2009: Neighborhood verification – a strategy for rewarding close forecasts. *Wea. Forecasting* (In Press).

Environmental Modeling Center, 2003: The GFS atmospheric model. NCEP Office Note 442, NCEP/NWS, 14 pp. [Available online at http://www.emc.ncep.noaa.gov/officenotes/newernotes/on442.pdf.]

Fels, S. B. and M. D. Schwarzkopf, 1975: The simplified exchange approximation: A new method for radiative transfer calculations, *J. Atmos. Sci.*, **32**, 1475–1488.

Fritsch, J. M., and R.E. Carbone, 2004: Improving quantitative precipitation forecasts in the warm season: A USWRP research and development strategy. *Bull. Amer. Meteor. Soc.*, 85, 955–965.

Gallus, W. A., 2002: Impact of verification grid-Box Size on Warm-Season QPF Skill Measures. *Wea. Forecasting*, **17**, 1296–1302.

Gilleland, E., D. Ahijevych, B.G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430.

Glahn, B., M. Peroutka, J. Wiedenfeld, J. Wagner, G. Zylstra, B. Schuknecht, and B. Jackson, 2009: MOS uncertainty estimates in an ensemble framework. *Mon. Wea. Rev.*, **137**, 246–268.

Grams, J. S., W. A. Gallus, S. E. Koch, L. S. Wharton, A. Loughe, and E. E. Ebert, 2006: The use of a modified Ebert–McBride Technique to evaluate mesoscale model QPF as a function of convective system morphology during IHOP 2002. *Wea. Forecasting*, **21**, 288–306.

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

Hong, S.-Y, and J.-O. J. Lim, 2006: The WRF single-moment 6-class microphysics scheme (WSM6). *J. Korean Meteor. Soc.*, **42**, 129-151.

Janjic, Z. I., 1994: The step-mountain Eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.*, **122**, 927-945.

Jenkner, J., C. Frei, and C. Schwierz, 2008: Quantile-based short-range QPF evaluation over Switzerland. *Meteorol. Z.*, **17**, 827-848.

——, 2002: Nonsingular implementation of the Mellor-Yamada Level 2.5 Scheme in the NCEP Meso model. NCEP Office Note No. 437, NOAA/NWS, 61 pp.

——, 2003: A nonhydrostatic model based on a new approach. *Meteorol. Atmos. Phys*., 82, 271-285.

Kain, J. S., M. E. Baldwin, P. R. Janish, S. J. Weiss, M. P. Kay, and G. W. Carbin, 2003: Subjective verification of numerical models as a component of a broader interaction between research and operations. *Wea. Forecasting*, **18**, 847–860.

Kain, J. S., S. J. Weiss, D. R. Bright, M. E. Baldwin, J. J. Levit, G. W. Carbin, C. S. Schwartz, M. L. Weisman, K. K. Droegemeier, D. B. Weber, and K. W. Thomas, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931–952.

Lacis, A. A., and J. E. Hansen, 1974: A parameterization for the absorption of solar radiation in the earth's atmosphere. *J. Atmos. Sci.*, **31**, 118–133.

Lin, Y. L., R. D. Farley, and H. D. Orville, 1983: Bulk parameterization of the snow field in a cloud model. *J. Appl. Meteor.*, **22**, 1065–1092.

Liu, C., M. W. Moncrieff, J. D. Tuttle, and R. E. Carbone, 2006: Explicit and parameterized episodes of warm-season precipitation over the continental United States. *Adv. Atmos. Sci*., **23**, 91-105.

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, 83, 407–430.

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

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

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmosphere: RRTM, a validated correlated-k model for the long-wave. *J. Geophys. Res*., 102(D14), 16663-16682.

Molinari, J., and M. Dudek, 1992: Parameterization of convective precipitation in mesoscale numerical models: A Critical Review. *Mon. Wea. Rev*., **120**, 326–344.

Nachamkin, J. E., 2004: Mesoscale verification using meteorological composites. *Mon. Wea. Rev.*, **132**, 941–955.

Noh, Y., W. G. Cheon, S.-Y. Hong, and S. Raasch, 2003: Improvement of the K-profile model for the planetary boundary layer based on large eddy simulation data. *Boundary Layer Meteor*., **107**, 401-427.

Parrish, D., J. Purser, E. Rogers, and Y. Lin, 1996: The regional 3D-variational analysis for the ETA model. Preprints, *11th Conf. on Numerical Weather Prediction*, Norfolk, VA, Amer. Meteor. Soc., 454-455.

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.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570–575.

Schwartz, C. S., J. S. Kain, S. J. Weiss, M. Xue, D. R. Bright, F. Kong, K. W. Thomas, J. J. Levit, M. C. Coniglio, Matthew S. Wandishin, 2009: Toward improved convection-allowing ensembles:

Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting* (In Press).

Schwarzkopf, M. D., and S. B. Fels, 1985: Improvements to the algorithm for computing CO2 transmissivities and cooling rates. *J. Geophys. Res.*, **90** (ND6), 541–550.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang and J. G. Powers, 2005: A description of the Advanced Research WRF Version 2, NCAR Tech Note, NCAR/TN–468+STR, 88 pp. [Available from UCAR Communications, P.O. Box 3000, Boulder, CO, 80307]. Also available at: http://box.mmm.ucar.edu/wrf/users/docs/arw_v2.pdf

——, 2006: Positive-definite and monotonic limiters for unrestricted-time-step transport schemes. *Mon. Wea. Rev*., **134**, 2241-2250.

——, and M. L. Weisman, 2009: The impact of positive-definite moisture transport on NWP precipitation forecasts. *Mon. Wea. Rev.*, **137**, 488–494.

Thompson, G., R. M. Rasmussen, and K. Manning, 2004: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part I: Description and sensitivity analysis. *Mon. Wea. Rev*., **132**, 519–542.

Wang, S. Y., T. C. Chen, and S. E. Taylor, 2009: Evaluations of NAM forecasts on midtropospheric perturbation-induced convective storms over the U.S. Northern Plains. *Wea. Forecasting*, **24**, 1309–1333.

Weckwerth, T. M., D. B. Parsons, S. E. Koch, J. A. Moore, M. A. LeMone, B. B. Demoz, C. Flamant, B. Geerts, J. Wang, and W.F. Feltz, 2004: An overview of the International H_2O Project (IHOP_2002) and some preliminary highlights. *Bull. Amer. Meteor. Soc.*, **85**, 253–277.

Weisman, M. L., W. C. Skamarock, J. B. Klemp, 1997: The resolution dependence of explicitly modeled convective systems. *Mon. Wea. Rev*., **125**, 527–548.

——, C. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with 0–36-h explicit

convective forecasts with the WRF-ARW model. *Wea. Forecasting*, **23**, 407–437.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction.* Academic Press, 467 pp.

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

List of Tables

Table 1 NCAR-WRF model specifications. See text for additional information.

List of Figures

Figure 1 (a) Outlined domains were used for the experimental NCAR-WRF simulations during 2004, 2005, 2007, and 2008 (legend provided at top-right) and the gray shaded domain was used for the analyses in this study. The outer and inner white circles within the analysis domain have radii of 250 and 100 km, respectively (discussed in text). (b) Gray shaded dates indicate cases used in this study (294 total cases).

Figure 2 $\langle ETS \rangle_r$ during 2004 and 2005 at increasing *r* for 3-hourly accumulated precipitation forecasts from NCAR-WRF (black lines) and NAM (gray lines) at the 0.10-in precipitation threshold for forecast hours (a) 18, (b) 24, and (c) 30. (d) – (f), (g) –(i), and (j) – (l) same as (a) – (c) except for the 0.25-, 0.50, and 1.00-in precipitation thresholds, respectively. The thin lines with circles are for the raw forecasts and the thick lines are for bias-corrected (bc) forecasts [legend provided in panel (a)]. The biases for the raw forecasts corresponding to each rainfall threshold and forecast hour pictured are provided in each panel. The horizontal lines drawn through the middle of each panel mark where ETS = 0.50.

Figure 3 Same as Figure 2, except for 2007 and 2008.

Figure 4 Time series of $\langle ETS \rangle_r$ for 2004 and 2005 from NCAR-WRF (black line) and NAM (gray line) using *r* = 100 km and *r* = 0 km for increasing forecast lead times at precipitation thresholds (a) 0.10-, (b) 0.25-, (c) 0.50-, and (d) 1.00-in. (e) – (h) same as (a) – (d) except for 2007 and 2008. Gray shaded regions indicate time periods at which differences in ETS between NCAR-WRF and NAM were statistically significant (α = 0.05).

Figure 5 Example of a neighborhood with radius *r* = 100 km used for the compositing method. Each grid-box is 20 km x 20 km. x marks a forecast event, o marks an observed event, and the variable i is the i*th* grid-point within the neighborhood around the x*th* grid-point of the domain [x = 462 (center-point of the neighborhood marked in bold) for this particular neighborhood].

For this neighborhood, $e_{462,i}$ = 1.0 for the gray shaded grid-boxes. For all other grid-boxes $e_{462,i}$ = 0.0. Note that for the composites shown in this study (Figs. 6 and 7), r = 250 km which results in a total number of grid-boxes for each neighborhood of n = 489, rather than the values of r = 100 and n = 81 shown in this example.

Figure 6 Composite frequency of observed rainfall above 0.50-in relative to grid-points forecasting rainfall above 0.50-in. from NCAR-WRF forecasts from 2004 and 2005 for forecast hours (a) 12, (b) 15, (c) 18, (d) 21, (e) 24, (f) 27, (g) 30, (h) 33, and (i) 36. (j) – (r) and (s) – (α) same as (a) – (i) except for NAM forecasts, and differences between NCAR-WRF and NAM forecasts (i.e., NCAR-WRF – NAM), respectively. The bold dot in each panel marks the center of the composite domain.

Figure 7 Same as Figure 6, except for 2007 and 2008.

Figure 8 Frequency histogram of $\langle ETS \rangle_{100}$ differences (NCAR-WRF minus NAM) at forecast hour 30 and the 0.50-in precipitation threshold.

Figure 9 Accumulated precipitation forecasts (3-hourly; bias-corrected) at forecast hour 30 for simulations initialized 06 May 2007 from (a) NAM and (b) NCAR-WRF, and (c) verifying Stage IV analyses. (d) Overlay of NCAR-WRF (hatched blue) and NAM forecasts (hatched green) for precipitation greater than 0.50-in. from (a) and (b) along with the verifying Stage IV analyses (shaded red). (e) – (h), (i) – (l), (m) – (p), (q) – (t), (u) – (x), (w) – (z), (α) – (δ), and (ε) – (θ) same as (a) – (d) except for simulations initialized at the dates indicated to the left of each row of panels.

Table 1 NCAR-WRF model specifications. See text for additional information.

Year	**Domain size (km)**	**WRF Version**	**ICs**	**LBCs**	**Grid-spacing**	**Vertical levels**	**Boundary Layer**	**Micro-physics**	**PD moisture transport**
2004	2000x2000	1.3	NAM	NAM	4-km	35	YSU	Lin	No
2005	3900x3000	2.0.3.1	NAM	NAM	4-km	35	YSU	WSM6	No
2007	3300x2700	2.2	NAM	NAM	3-km	35	MYJ	Thompson	Yes
2008	2900x2700	3.0	WRF-Var	GFS	3-km	40	MYJ	Thompson	Yes

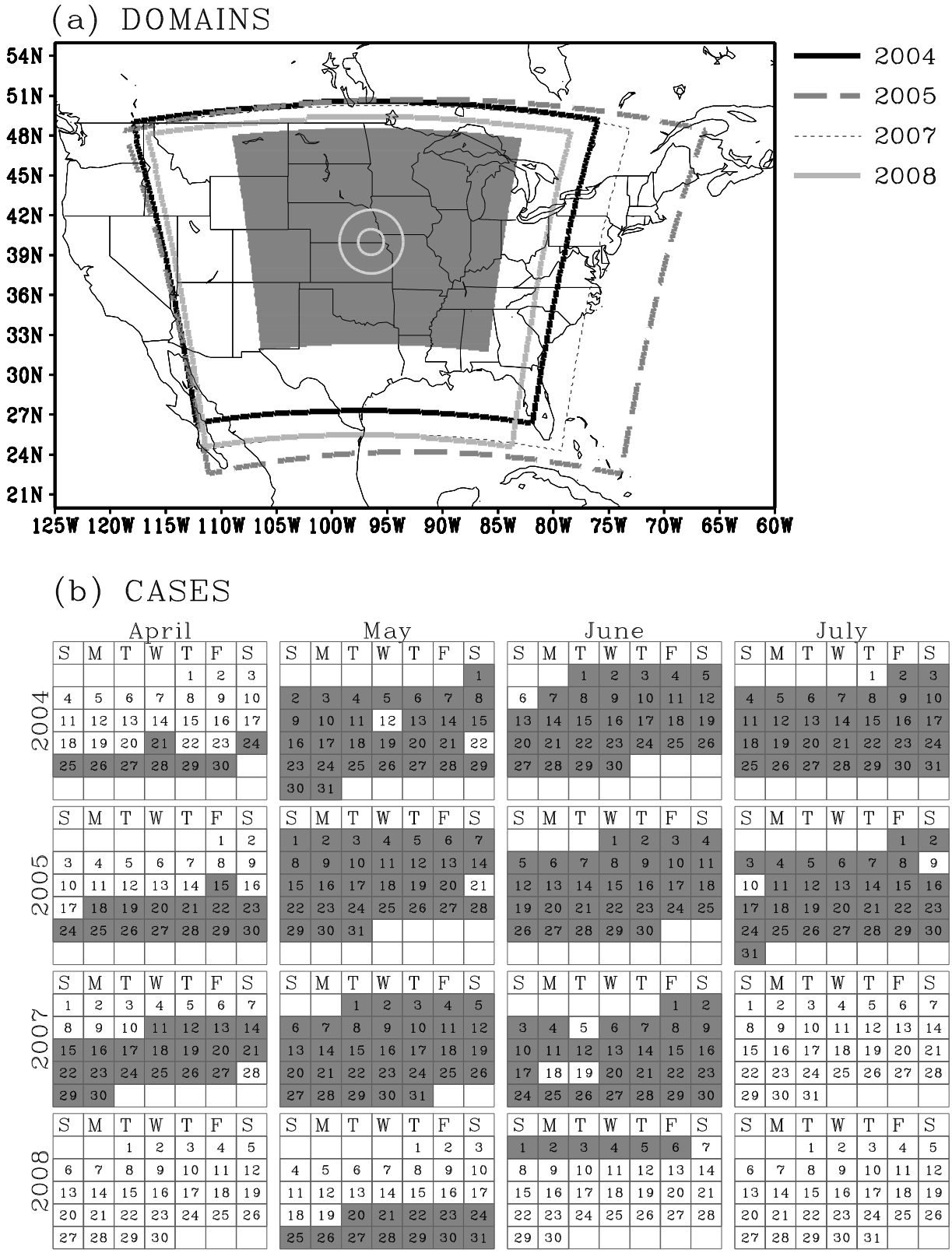

Figure 1 (a) Outlined domains were used for the experimental NCAR-WRF simulations during 2004, 2005, 2007, and 2008 (legend provided at top-right) and the gray shaded domain was used for the analyses in this study. The outer and inner white circles within the analysis domain have radii of 250 and 100 km, respectively (discussed in text). (b) Gray shaded dates indicate cases used in this study (294 total cases).

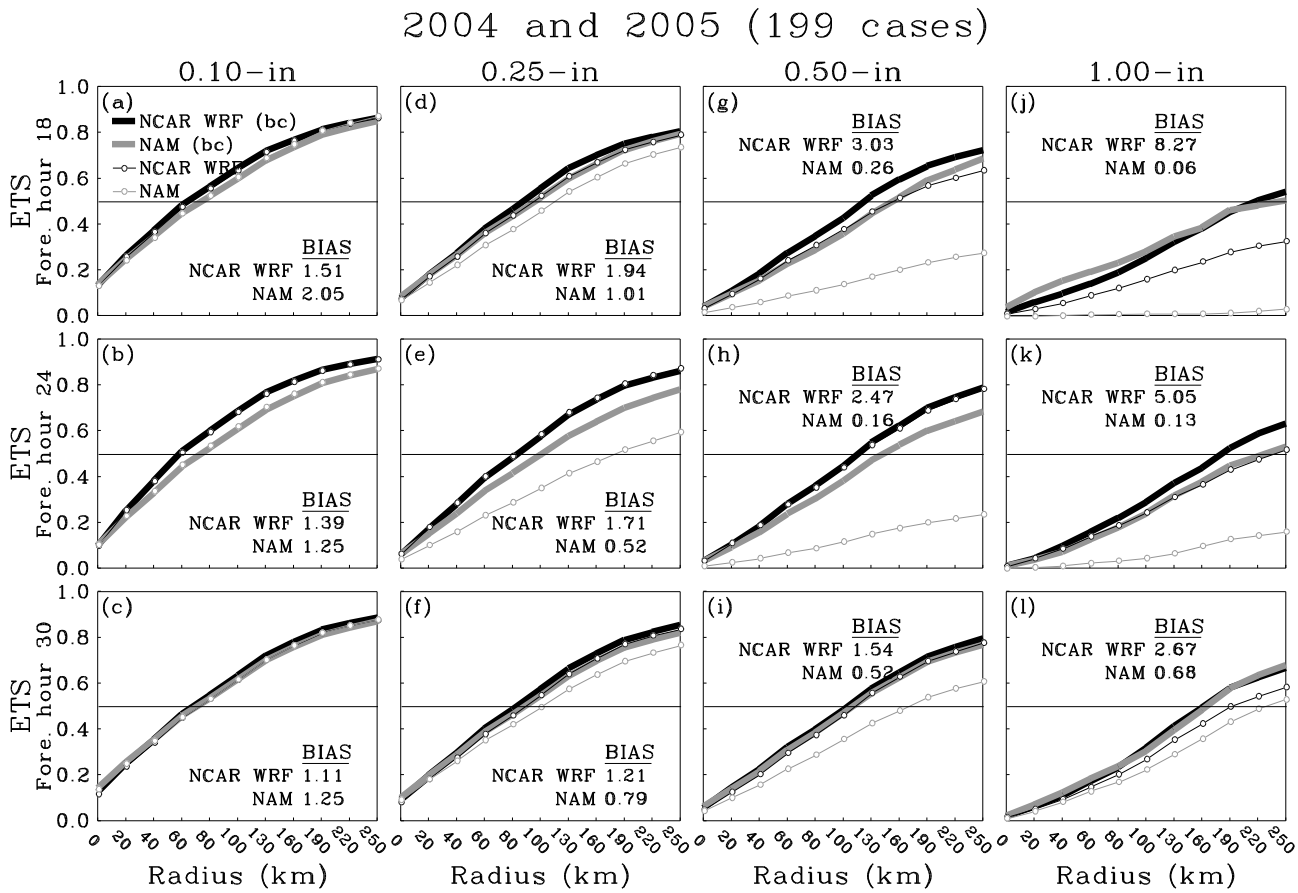

Figure 2 $\langle ETS \rangle_r$ during 2004 and 2005 at increasing *r* for 3-hourly accumulated precipitation forecasts from NCAR-WRF (black lines) and NAM (gray lines) at the 0.10-in precipitation threshold for forecast hours (a) 18, (b) 24, and (c) 30. (d) – (f), (g) –(i), and (j) – (l) same as (a) – (c) except for the 0.25-, 0.50, and 1.00-in precipitation thresholds, respectively. The thin lines with circles are for the raw forecasts and the thick lines are for bias-corrected (bc) forecasts [legend provided in panel (a)]. The biases for the raw forecasts corresponding to each rainfall threshold and forecast hour pictured are provided in each panel. The horizontal lines drawn through the middle of each panel mark where ETS = 0.50.

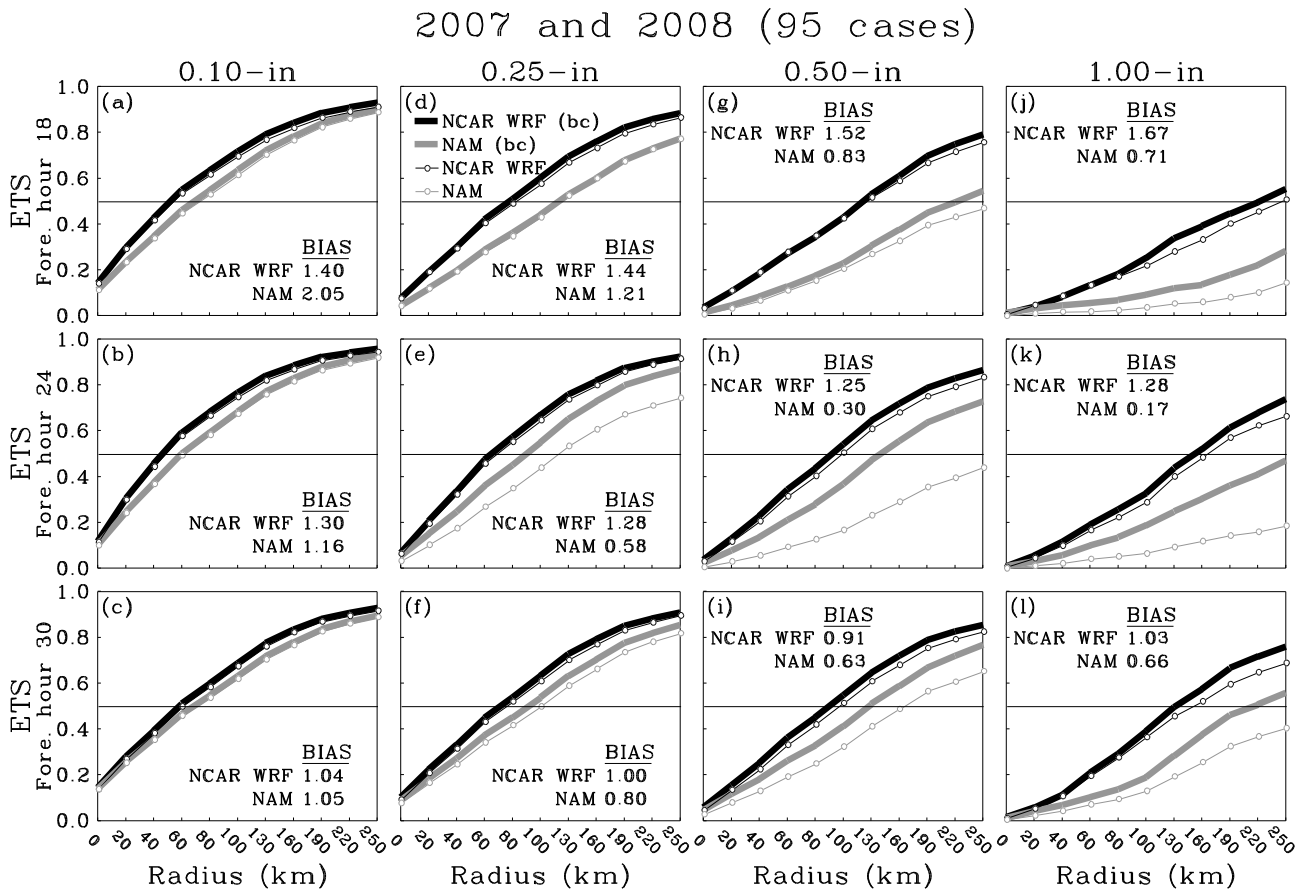

Figure 3 Same as Figure 2, except for 2007 and 2008.

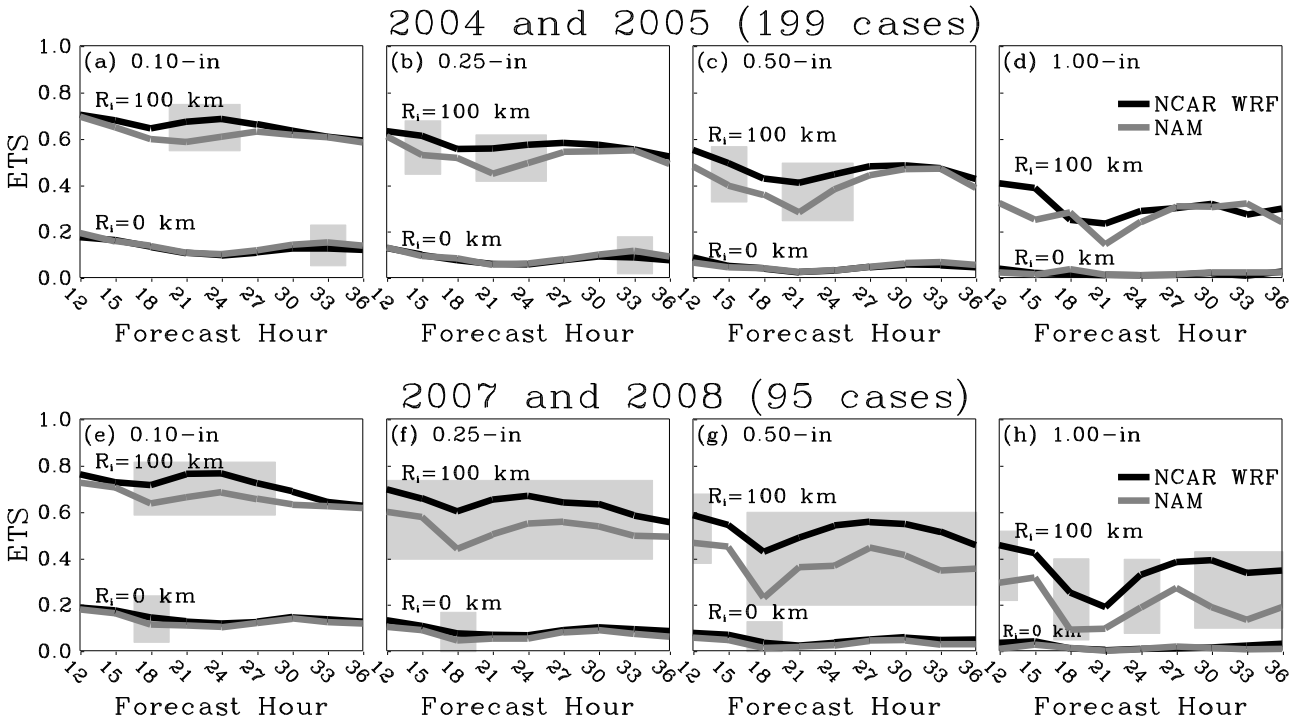

Figure 4 Time series of $\langle ETS \rangle_r$ for 2004 and 2005 from NCAR-WRF (black line) and NAM (gray line) using *r* = 100 km and *r* = 0 km for increasing forecast lead times at precipitation thresholds (a) 0.10-, (b) 0.25-, (c) 0.50-, and (d) 1.00-in. (e) – (h) same as (a) – (d) except for 2007 and 2008. Gray shaded regions indicate time periods at which differences in ETS between NCAR-WRF and NAM were statistically significant (α = 0.05).

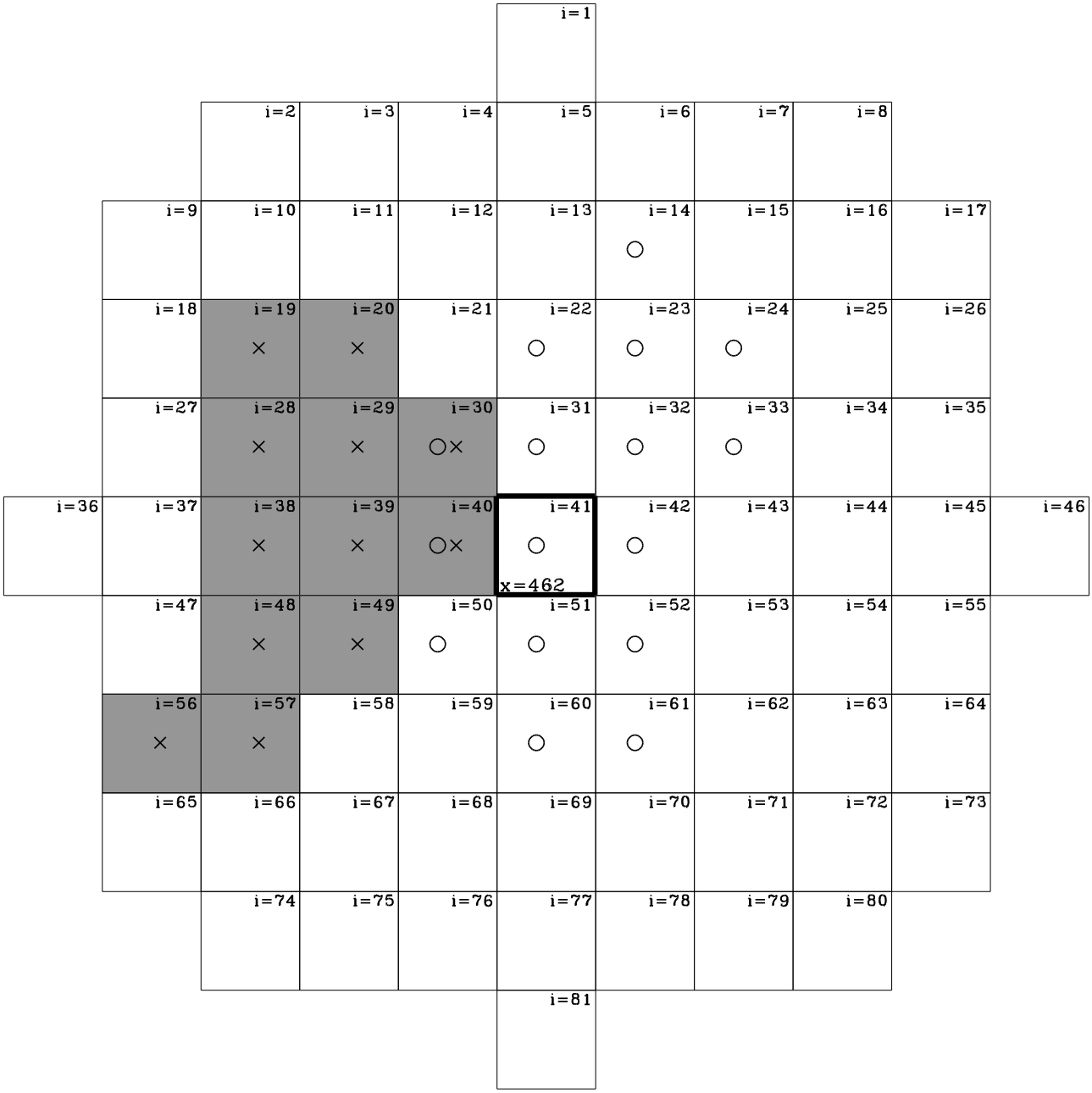

Figure 5 Example of a neighborhood with radius r = 100 km used for the compositing method. Each grid-box is 20 km x 20 km. x marks a forecast event, o marks an observed event, and the variable i is the i*th* grid-point within the neighborhood around the x*th* grid-point of the domain [x = 462 (center-point of the neighborhood marked in bold) for this particular neighborhood]. For this neighborhood, $e_{462,i}$ = 1.0 for the gray shaded grid-boxes. For all other grid-boxes $e_{462,i}$ = 0.0. Note that for the composites shown in this study (Figs. 6 and 7), r = 250 km which results in a total number of grid-boxes for each neighborhood of n = 489, rather than the values of r = 100 and n = 81 shown in this example.

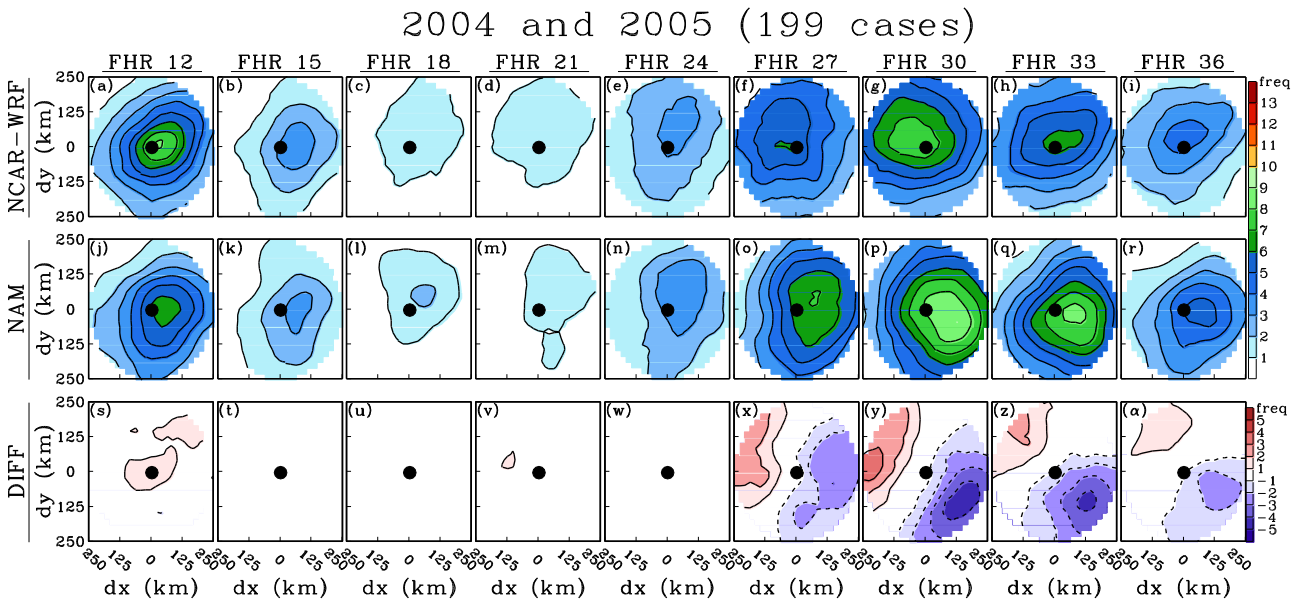

Figure 6 Composite frequency of observed rainfall above 0.50-in relative to grid-points forecasting rainfall above 0.50-in from NCAR-WRF forecasts from 2004 and 2005 for forecast hours (a) 12, (b) 15, (c) 18, (d) 21, (e) 24, (f) 27, (g) 30, (h) 33, and (i) 36. (j) – (r) and (s) – (α) same as (a) – (i) except for NAM forecasts, and differences between NCAR-WRF and NAM forecasts (i.e., NCAR-WRF – NAM), respectively. The bold dot in each panel marks the center of the composite domain.

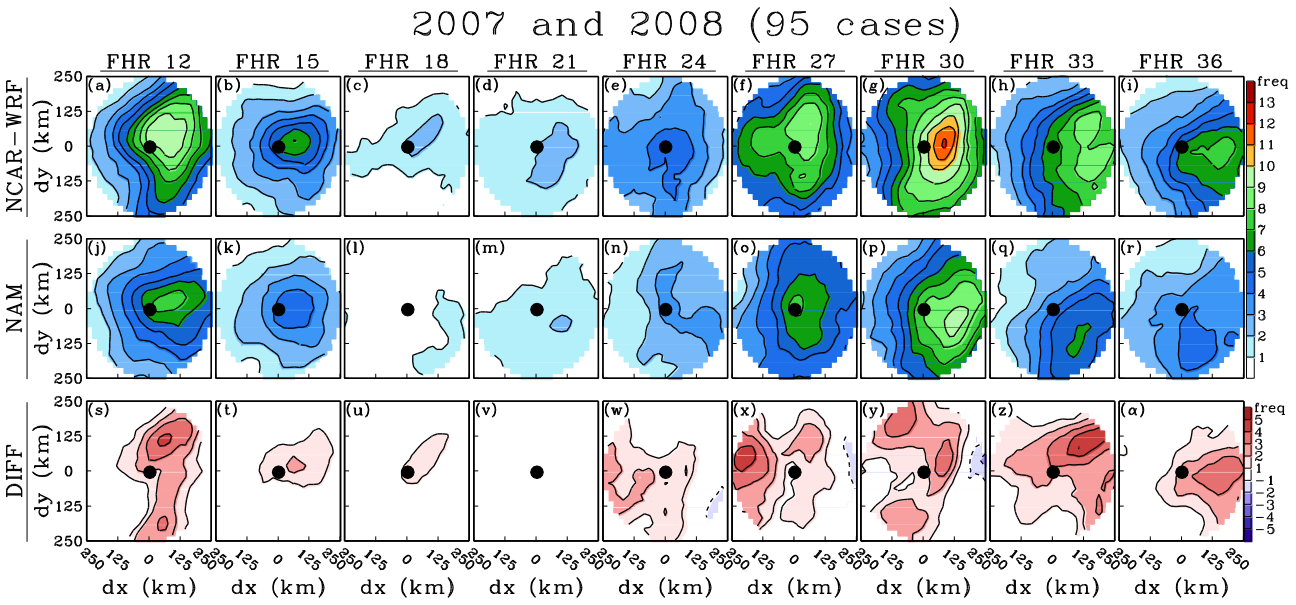

Figure 7 Same as Figure 6, except for 2007 and 2008.

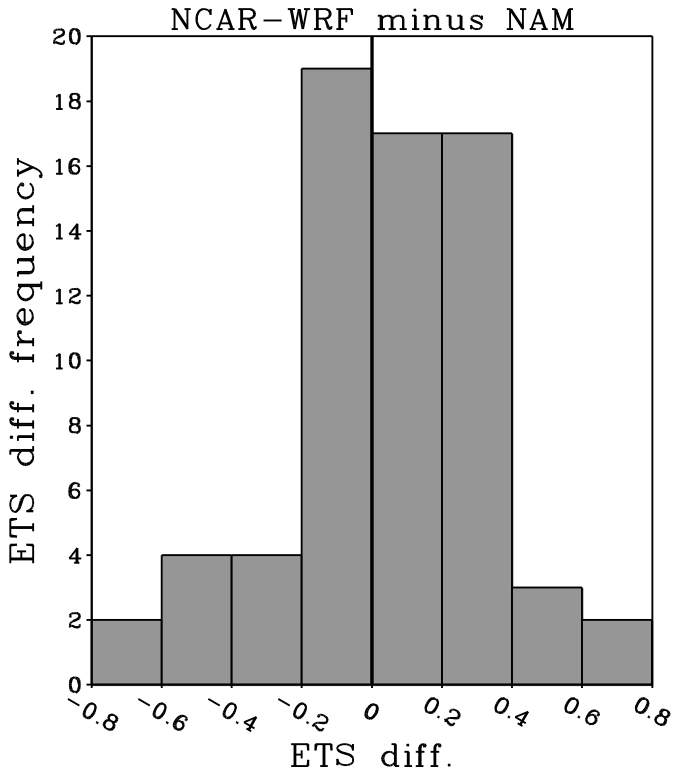

Figure 8 Frequency histogram of $\langle ETS \rangle_{100}$ differences (NCAR-WRF minus NAM) at forecast hour 30 and the 0.50-in precipitation threshold.

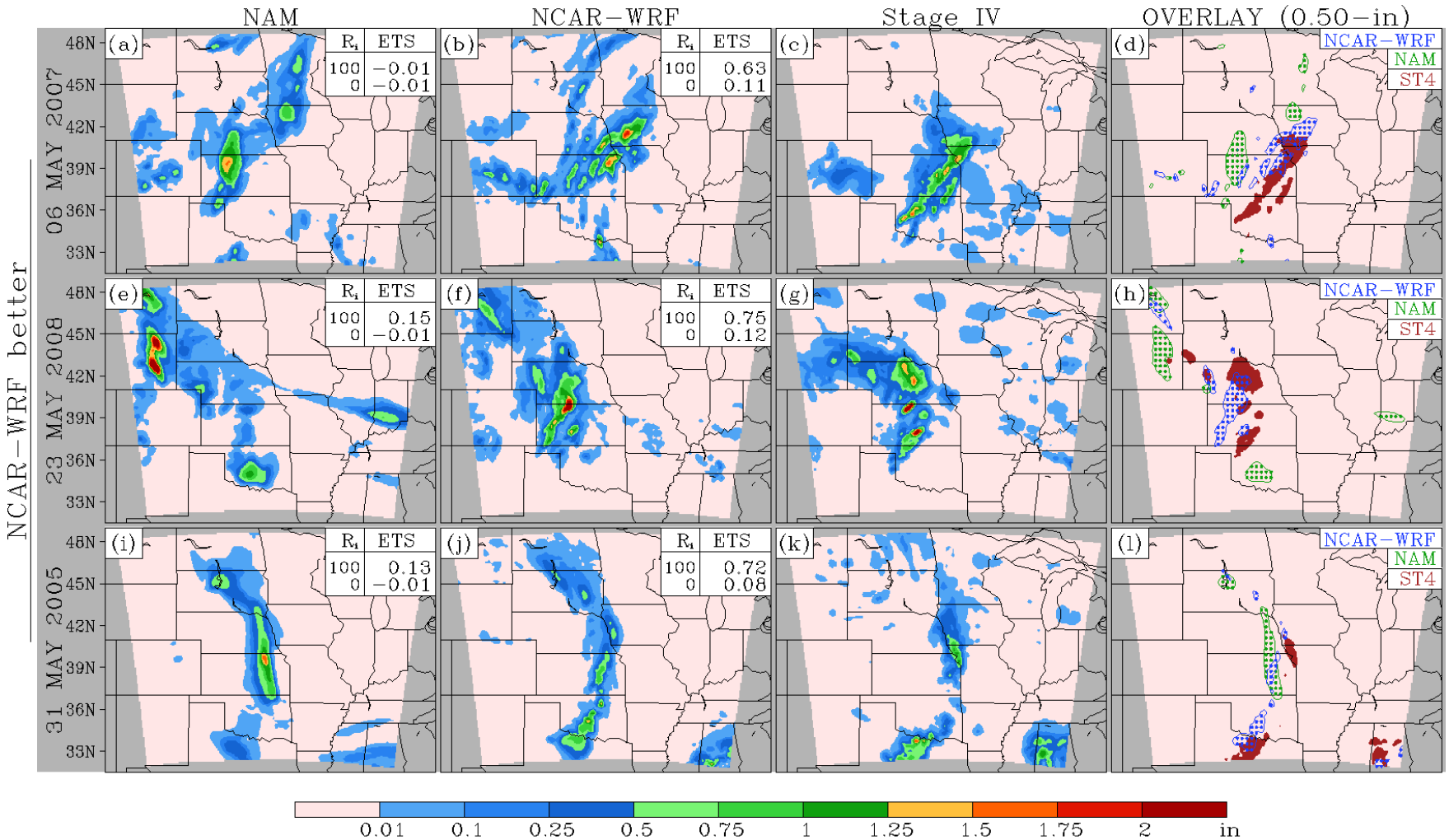

Figure 9 Accumulated precipitation forecasts (3-hourly; bias-corrected) at forecast hour 30 for simulations initialized 06 May 2007 from (a) NAM and (b) NCAR-WRF, and (c) verifying Stage IV analyses. (d) Overlay of NCAR-WRF (hatched blue) and NAM forecasts (hatched green) for precipitation greater than 0.50-in. from (a) and (b) along with the verifying Stage IV analyses (shaded red). (e) – (h), (i) – (l), (m) – (p), (q) – (t), (u) – (x), (w) – (z), (α) – (δ), and (ε) – (θ) same as (a) – (d) except for simulations initialized at the dates indicated to the left of each row of panels.

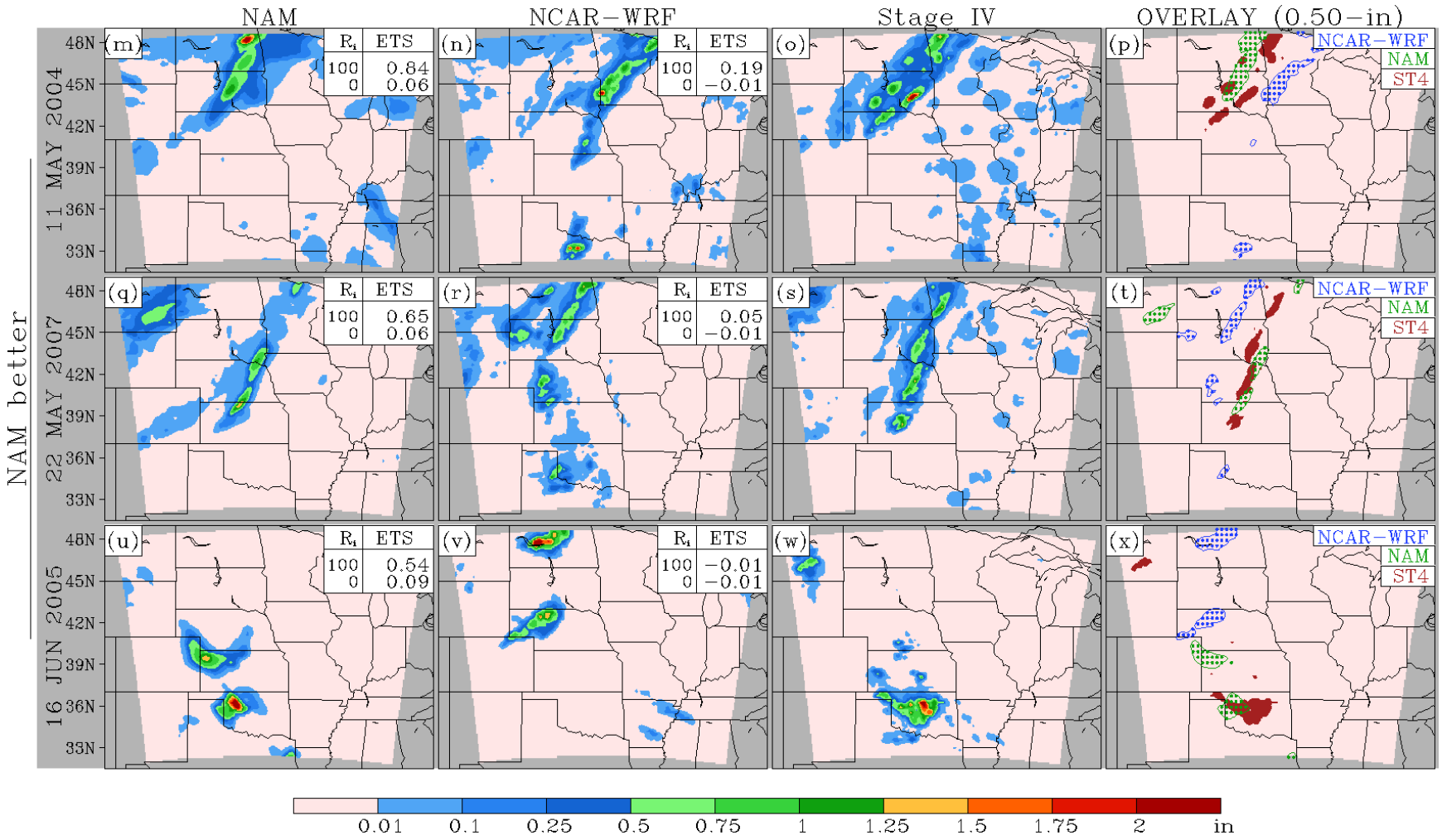

Figure 9 (cont.)

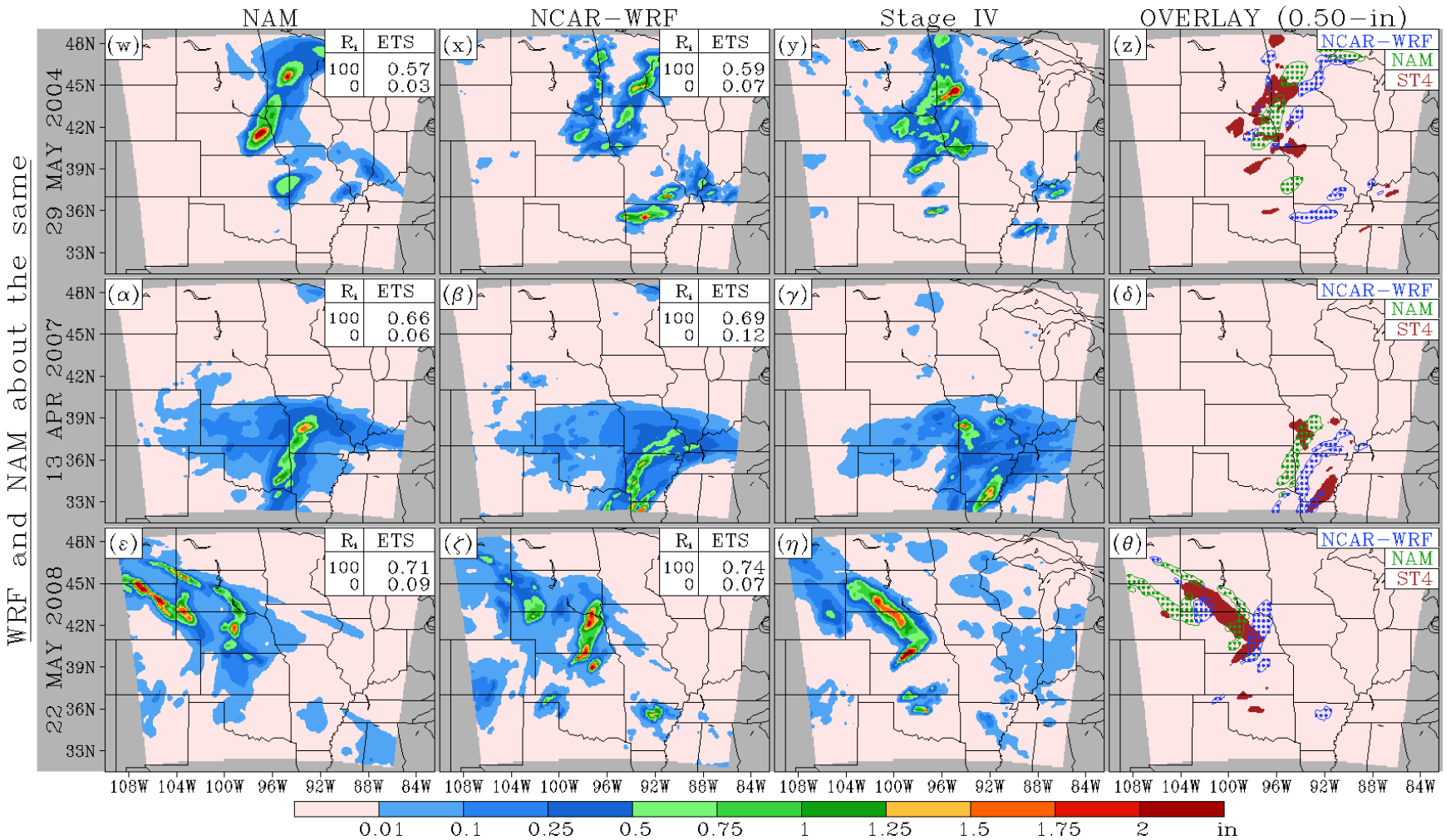

Figure 9 (cont.)