An Evaluation of QPF from the WRF, NAM, and GFS Models Using Multiple Verification Methods over a Small Domain

HAIFAN YAN AND WILLIAM A. GALLUS JR.

Department of Geological and Atmospheric Sciences, Iowa State University, Ames, Iowa

(Manuscript received 28 January 2016, in final form 17 June 2016)

ABSTRACT

The ARW model was run over a small domain centered on Iowa for 9 months with 4-km grid spacing to better understand the limits of predictability of short-term (12 h) quantitative precipitation forecasts (QPFs) that might be used in hydrology models. Radar data assimilation was performed to reduce spinup problems. Three grid-to-grid verification methods, as well as two spatial techniques, neighborhood and object based, were used to compare the QPFs from the high-resolution runs with coarser operational GFS and NAM QPFs to verify QPFs for various precipitation accumulation intervals and on two grid configurations with different resolutions. In general, NAM had the worst performance not only for model skill but also for spatial feature attributes as a result of the existence of large dry bias and location errors. The finer resolution of NAM did not offer any advantage in predicting small-scale storms compared to the coarser GFS. WRF had a large advantage for high precipitation thresholds. A greater improvement in skill was noted when the accumulation time interval was increased, compared to an increase in the spatial neighborhood size. At the same neighborhood scale, the high-resolution WRF Model was less influenced by the grid on which the verification was done than the other two models. All models had the highest skill from midnight to early morning, because the least wet bias, location, and coverage errors were present then. The lowest skill was shown from late morning through afternoon. The main cause of poor skill during this period was large displacement errors.

1. Introduction

Numerical weather prediction (NWP; see the appendix for a list of key abbreviations and acronyms used in this paper) has substantially improved over the past decades because of improvements in observation datasets and computation power. Precipitation is one of the key forecast elements within NWP, as a variety of communities such as agriculture, transportation, airlines, etc. require accurate forecasts, and are especially interested in as much detail (spatial, temporal) as possible. Skillful QPFs can provide instructive information for hydrology forecasters and hydrological models; for example, skillful QPFs could be input into hydrologic models before QPE is available, thus improving the lead time for potentially hazardous flooding situations. Unfortunately, although the threat score in general for QPFs has significantly increased in the past 50 years, the skill of warm season QPF has only shown incremental improvement (Barthold et al. 2015), likely because half of the warm season precipitation is directly related to mesoscale forcing mechanisms, and over 80% of the total rainfall is directly or indirectly associated with thunderstorms (Heideman and Fritsch 1988). Thus, hydrology forecasters routinely use only quantitative precipitation estimates and not QPFs.

Numerous studies have tried to find the limitations of QPF and methods to improve it. However, the essential challenge in short- and medium-range QPFs is that numerical models are highly nonlinear, so the uncertainties of the models are still poorly understood. It is very difficult to determine which parameter is responsible for a certain deficiency (Fritsch and Heideman 1989; Cloke and Pappenberger 2009). QPFs can be largely influenced by different initializations, microphysics, and PBL schemes (Jankov et al. 2007a,b), and the impact of different physical schemes depends on initialization data as well as different cases (Jankov et al. 2007a,b). Deep, moist convection, which can result in severe weather, requires the accurate forecasts of convective initiation, which is also a known challenge for both models and humans. With grid spacings of 3–4 km, models are better

Corresponding author address: William A. Gallus Jr., Iowa State University, 3025 Agronomy Hall, Ames, IA 50011.
E-mail: wgallus@iastate.edu

DOI: 10.1175/WAF-D-16-0020.1

able to predict the timing of convective initiation, although errors are still common (Kain et al. 2013; Duda and Gallus 2013; Burghardt et al. 2014). Duda and Gallus (2013) suggest that upscale evolution is better forecasted than the initiation.

Many studies have shown that radar data assimilation has a very obvious positive effect on short-range (≤12 h) QPFs (Xiao et al. 2007; Moser et al. 2015). Although model runs with radar assimilation, often called hot starts, are generally too wet in the first 1–2 h, hot starts show much better performance in spatial attributes and skill scores than model runs without data assimilation (cold starts) (Moser et al. 2015). With higher-resolution initializations and data assimilation, the skill of QPFs can be improved up to 8–9 h (Sun et al. 2012).

As grid resolution has been refined, an increasing number of researchers have expressed concern about the verification metrics used to evaluate the performance of these models. Traditional verification methods, such as equitable threat score (ETS; also known as Gilbert skill score), critical success index (CSI), false alarm rate (FAR), probability of detection (PODY), and frequency bias (FBIAS), have been widely used in the past several decades. However, many traditional verification methods are grid-to-grid approaches, so they are sensitive to small-scale position errors. Thus, for high-resolution models, traditional methods may indicate lower skill than forecasters would expect from qualitative assessment, as the model improvements are hidden by the subtle displacement errors.

To better understand the limits on the predictability of high-resolution QPFs, a large number of new spatial verification methods have been developed in recent years. Gilleland et al. (2009) summarized the new verification methods into four categories: 1) neighborhood, 2) scale separation, 3) object based, and 4) field deformation. The first two methods both use a spatial filter on one or both of the observation and forecast fields. The last two methods both try to figure out how much the forecast field needs to be corrected in order to achieve meaningful skill.

In the present paper, in order to provide more detailed information on shortcomings that can guide the work of model developers, a matrix of verification methods including traditional, neighborhood and object-based methods is used to verify the performance of QPFs in a hot-start convection-allowing model and to compare it with QPFs from two operational models. Fractions skill score (FSS) and parameters from the Method for Object-Based Diagnostic Evaluation (MODE), which are recently proposed neighborhood and object-based methods, respectively, are the two major approaches used in this study. These two spatial techniques can provide comprehensive analysis of the performance of numerical models over different scales, and for location errors, intensity errors, structure errors, etc. Convection-allowing ARW model simulations can help us better understand the spatial and temporal limits of QPF as it is considered for hydrologic use.

The verification methods were performed over 9 months during 2013 and covered Iowa and immediately adjacent areas of other states. The QPFs of the upper Midwest including Iowa, generally have lower PODY and CSI values and higher false alarm ratios than the values in the western and northeastern parts of the continental United States (CONUS) because there is less influence from small-scale convective storms in those areas (Sukovich et al. 2014). Hence, more information about how QPF skill compares among models in the central United States, where skill can be especially poor, can assist forecasters and model developers. In this paper, section 2 describes the model configuration and verification methodology. Section 3 is the analysis of model performance via various verification methods. A discussion and conclusions follow in section 4.

2. Data and methodology

a. Model setup and data description

Version 3.5 of ARW (Skamarock et al. 2008) was run every 6 h (0000, 0600, 1200, and 1800 UTC) in order to have a better understanding about the limits of predictability of short-term (12 h) high-resolution QPFs that might be used in hydrology models. The model runs were initialized using the Advanced Regional Prediction System three-dimensional variational data assimilation system (ARPS 3DVAR) and the ARPS Data Analysis System (ADAS), which are parts of the ARPS (Xue et al. 1995, 2000, 2003), a regional to storm-scale atmospheric modeling system. Both the 12-km grid-spacing NCEP NAM (Janjić 2003) and 0.5° × 0.5° NCEP GFS (Environmental Modeling Center 2003) 0-h analyses from each model cycle archived from the NOAA's National Operational Model Archive and Distribution System (NOMADS) were used as the first-guess field, and the NAM and GFS 3-h forecasts were used as lateral boundary conditions in the ARPS. Note that radar data were only assimilated at the initialization time, as would be the case for real-time forecasts. In the present study, in order to reduce spinup problems normally encountered in model simulations that simply use output from other models for initialization, the ARPS 3DVAR and ADAS assimilated NEXRAD level II radar data from nine sites located within the domain region were used to adjust the initial NAM or GFS background fields. The radar

reflectivity data were used by a complex cloud analysis procedure, which is a component of both ADAS and ARPS 3DVAR, to adjust hydrometeors and cloud fields, and radial velocity data were analyzed via the three-dimensional variational scheme. The three-dimensional cloud and precipitation fields were constructed based on radar data (Hu et al. 2006; Moser et al. 2015). The nine sites (Fig. 1) were Aberdeen, South Dakota (KABR); Lacrosse, Wisconsin (KARX); Des Moines, Iowa (KDMX); Davenport, Iowa (KDVN); Kansas City, Missouri (KEAX); Sioux Falls, South Dakota (KFSD); St. Louis, Missouri (KLSX); Minneapolis, Minnesota (KMPX); and Omaha, Nebraska (KOAX). The input radar data covered the entire simulated domain.

The initial conditions created in the ARPS 3DVAR were then integrated into WRF (hereafter WRF and WRFGFS for NAM and GFS initializations, respectively). The model domain (Fig. 1) was centered at 41.916°N and 93.342°W with 200 × 200 horizontal grid points and 4-km cell spacing on a Lambert conformal map projection. The model top pressure was around 60 hPa. The physics parameterizations used in this study included the two-moment Thompson microphysics scheme (Thompson et al. 2008), the local MYJ PBL scheme (Janjić 1994) and the New Goddard longwave and shortwave radiation schemes (Chou and Suarez 1994).

The two operational models used for WRF initialization, NAM and GFS, were also examined using QPF verification to establish a benchmark to which the WRF runs could be compared. The NAM differs from the explicit 4-km WRF simulations in that it includes the Nonhydrostatic Multiscale Model as the major dynamic component and also includes the Betts–Miller–Janjić (BMJ) shallow–deep convection parameterization. Hourly observed precipitation over the CONUS is assimilated in NAM (Rogers et al. 2009). The GFS simulates the shallow/deep convection based on the simplified Arakawa–Schubert scheme. The GFS also uses a hybrid variational ensemble assimilation system. NCEP stage IV precipitation data (Lin and Mitchell 2005) were used to represent ground truth in the verification process. To be consistent with the WRF simulations, only the first 12 h of the NAM and GFS output were considered in the present study and compared with Stage IV data at the corresponding times. The QPFs in all of the verified models and the Stage IV data were interpolated into the same domain configuration as the WRF (Hres) through the Unified Postprocessor using the budget method, which is able to more accurately conserve the total precipitation magnitude. In addition, in order to study the possible effects of interpolation on various verification metrics, all four types of data were also interpolated using the budget method as well to a

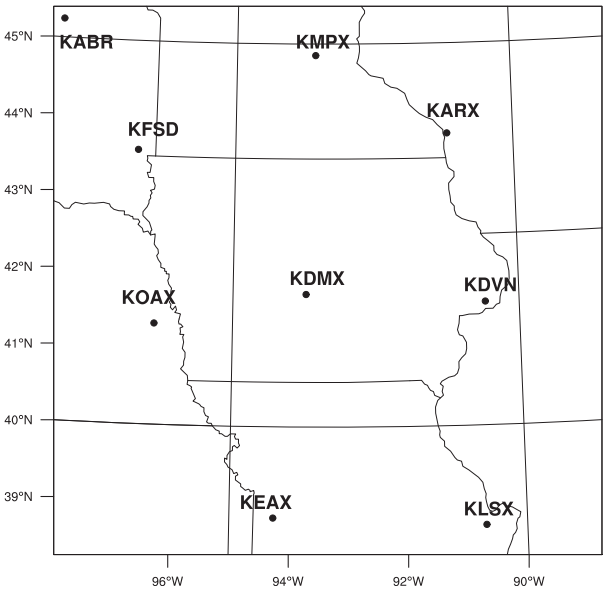

FIG. 1. Domain configuration and the location of nine radar sites used for data assimilation.

latitude–longitude map projection with 0.5° × 0.5° GFS (Lres) resolution, which is roughly around 55 km in the meridional direction and 42.5 km in the zonal direction. The domain region used for the Lres verification was the portion of the GFS grid for which data were also available from the WRF simulations. Note that this 0.5° × 0.5° GFS grid had already been regridded before dissemination to the public. The native resolution of GFS is T574 (~27 km), which is finer than 0.5° × 0.5, but it is common to use these gridded data for research purposes.

b. *Verification methods*

In this study, five approaches were used to verify the models including three traditional metrics: ETS, FAR, and FBIAS, as well as two spatial methods: FSS and MODE. All of these methods are included in a NWP verification software package developed by the Developmental Testbed Center (DTC; http://www.dtcenter.org/), known collectively as Model Evaluation Tools (MET). The two spatial methods will be particularly emphasized in the present study. Hourly WRF and stage IV data, 3-hourly NAM and GFS data, and 6-hourly Stage IV data were summed to 3-, 6-, and 12-h periods as necessary using the Pcp-combine tool in MET. The verification techniques were applied to the precipitation accumulation intervals of 1, 3, 6, and 12 h.

Traditional grid-to-grid verification methods such as ETS, FAR, and FBIAS are calculated based on a contingency table (Table 1). Of the total T forecast–observation pairs, whether or not accumulated precipitation (APCP)

TABLE 1. The 2 × 2 contingency table of four possible outcomes of a forecast of accumulated precipitation.

Total events $T = N_H + N_{FA} + N_M + N_{CN}$		Observation	
		Yes	No
Forecast	Yes	Hit (N_H)	False alarm (N_{FA})
	No	Miss (N_M)	Correct negative (N_{CN})

exceeds a specified threshold is used to determine if an event is a hit, false alarm, miss, or correct negative. The ETS is calculated based on the number of points where the events are correctly forecasted to occur relative to the total number of points where they are either forecasted or observed. ETS is further corrected by the chance forecasts (ref), which are the product of forecasted events and observed events, divided by the total counts. The value of ETS ranges from $-1/3$ to 1. The FAR represents the fraction of the forecasted events that were not observed. The FBIAS compares the total number of forecasts and the number of observations. A perfect forecast would have an ETS of 1, FAR of 0, and FBIAS of 1. The formulas of ETS, FAR, and FBIAS are defined as

$$\mathrm{ETS} = \frac{N_H - \mathrm{ref}}{N_H + N_{\mathrm{FA}} + N_M - \mathrm{ref}}, \tag{1}$$

$$\mathrm{ref} = \frac{(N_H + N_{\mathrm{FA}})(N_H + N_M)}{T}, \tag{2}$$

$$\mathrm{FAR} = \frac{N_{\mathrm{FA}}}{N_H + N_{\mathrm{FA}}}, \quad \text{and} \tag{3}$$

$$\mathrm{FBIAS} = \frac{N_H + N_{\mathrm{FA}}}{N_H + N_M}, \tag{4}$$

where the different subscripts for N represent the counts of hits N_H, false alarms N_{FA}, or misses N_M (Table 1). For FBIAS, it is possible that the count of a forecast event is hundreds of times larger than the number of occurrences that may be a very small value, yielding an enormous FBIAS that would inflate the mean FBIAS in a misleading way. Hence, the counts of events used in the formula above are the total counts of the 9 months rather than the counts of each 3-h run. To maintain consistency with FBIAS, ETS and FAR are both calculated by the total counts of hits, false alarms, and misses during the 9-month period. The traditional scores (shown later; see Fig. 5) were calculated using these summed contingency tables based on Eqs. (1)–(4). Note that the ETS, FAR, and FBIAS results are related to each other based on the contingency table.

The thresholds used to generate binary fields in traditional methods as well as to define events in FSS and MODE were 0.254 mm (0.01 in.), 2.54 mm (0.1 in.), 6.35 mm (0.25 in.), and 12.7 mm (0.5 in.), so the verification methods cover a range from light to relatively heavy intensity.

FSS is a neighborhood verification method developed by Roberts and Lean (2008) and further discussed by Roberts (2008) and Mittermaier and Roberts (2010). It is normalized based on fractions Brier score and is able to show how forecast skill varies with different spatial scales and thresholds. FSS is calculated in the following three steps. First, both forecast F and observation O fields are transformed into binary fields. A grid box will have the value of 1 if APCP exceeds a specified threshold; otherwise, it will have a value of 0. Although APCP is the only variable that will be verified for QPF in this research, other variables such as wind speed and radar reflectivity can also be verified using FSS. Second, the fraction of each grid point (i, j) in the binary observation field $O(i, j)$ [or forecast field $F(i, j)$] is generated from the neighborhood square centered in (i, j). The fraction ($P_{O_{(L)}}$ or $P_{F_{(L)}}$) is calculated by the number of grid boxes having the value of 1 over the number of all grid boxes within the neighborhood square. For example, as shown in Fig. 2, the fraction of (i, j) in the observed field is $P_{O(5)}(i, j) = 5/25$, and the fraction of (i, j) in the forecast field is $P_{F(5)}(i, j) = 6/25$. Third, FSS is calculated using the following formula:

$$\mathrm{FSS}_{(L)} = 1 - \frac{\dfrac{1}{N_{(L)}} \sum_{N_{(L)}} (P_{O_{(L)}} - P_{F_{(L)}})^2}{\dfrac{1}{N_{(L)}} \left[\sum_{N_{(L)}} (P_{O_{(L)}})^2 + \sum_{N_{(L)}} (P_{F_{(L)}})^2 \right]}, \tag{5}$$

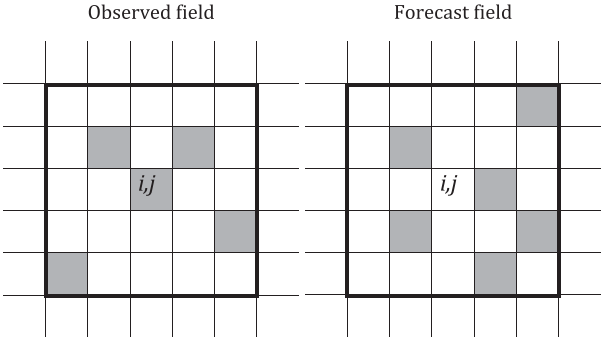

FIG. 2. A visual illustration of how the fraction is computed at the neighborhood scale of five grid lengths (see text). The grid boxes shaded in gray are those where the APCP of the grid box exceeds the specified threshold.

where $N_{(L)}$ is the number of valid neighborhoods at the neighborhood scale of L. The forecasts can be regarded as reasonably skillful when FSS reaches up to $0.5 + f_0/2$ according to Roberts and Lean (2008). The f_0 is a sample climatology variable known as base rate (BR), which represents the fraction of event occurrences over the whole domain in the binary raw observation field without smoothing; in other words, f_0 is the climatological chance of precipitation happening, so it is also used to represent random skill. Because FSS is calculated through a fuzzy box, some displacement errors considered as misses or false alarms in a traditional contingency table can be considered hits as long as the displacement happened within the neighborhood square.

In this study, in order to show how skill varies with scale, an arithmetic sequence of neighborhood sizes, 5, 9, 13, . . . , 101, was used for smoothing. Fractions were not calculated if part of a neighborhood square was outside of the domain boundaries. It is acknowledged that hydrology applications are more concerned with smaller scales rather than a fuzzy box containing 101 × 101 pixels, which was around a quarter of the whole domain. However, the spatial scale is an essential parameter determining the variation of FSS, so the extension of neighborhoods to these larger sizes can provide more information about the trend of FSS curves and give some guidance about the choice of a reasonable scale interval for hydrology applications.

MODE is a feature-based verification methodology based on Davis et al. (2006a,b) and Davis et al. (2009). Many features of matched pairs between model simulations and observations can be investigated using MODE, such as centroid distance (CD), boundary distance, intensity sum (IS; total rain volume), angle orientation, areal coverage, etc. The raw forecast and observation data are convolved using a circular filter with a specified radius. Because a five-gridpoint radius was recommended by Davis et al. (2006a) for practical use, in order to avoid too much smoothing, this specified radius was applied to generate convolved fields in the present study. Then, the APCP falling within the circular region is averaged to get the convolved field. The filtered regions used for feature comparisons can be obtained after the threshold of 2.54 mm, which is a moderate threshold for 3-h QPFs, is applied on the convolved field. The raw data within filtered regions is restored to get simple objects that are individual objects without matching or merging into cluster objects. Figure 3 provides an illustration of the MODE output and how MODE generates simple objects.

MODE can show location and structure errors of precipitation objects. The same thresholds were applied to MODE as was done with FSS to convolved fields to determine the boundaries of filtered regions. Many spatial features were collected to define a single number called the total interest. The value of total interest ranges from 0 (least agreement) to 1 (perfect agreement) to compare the similarity of two objects, and it was a weighted average of the following attributes: the centroid distance, the boundary distance, the convex hull distance, the orientation angle difference, the object area ratio, the intersection divided by the union area ratio, the complexity ratio, and the intensity ratio. In the present study, the relative weight of each attribute used the default setting in MET (Halley-Gotway et al. 2014). The displacement errors including centroid distance and boundary distance were weighted the greatest in the calculation of total interest. Two simple objects of forecast and observation fields had the chance to be defined as matched pairs only when CD was smaller than 50 grid points. Furthermore, object pairs would be matched when their total interest values were above 0.7.

The normalized differences of IS and areal coverage are used to link the feature attributes in forecast and observation fields for the MODE analysis. The IS difference of the whole domain (ISD) is presented in the following as an example to show the form of normalization:

$$\mathrm{ISD} = \frac{\mathrm{IS}_F - \mathrm{IS}_O}{\frac{1}{2}(\mathrm{IS}_F + \mathrm{IS}_O)}, \tag{6}$$

where IS_F and IS_O represent the total IS (in mm) over the whole domain in the forecast and observation fields, respectively. Besides the ISD, IS differences for matched pairs (ISDPs), areal coverage difference (in grid squares) of the whole domain (AD), and areal coverage difference for matched pairs (ADP) are also normalized using the form of the formula above.

Statistical significance t tests of pairwise differencing were performed at the 95% confidence level for all the means in the FSS and MODE analyses. The means of two models can be regarded as statistically significantly different if 0 is not included in the confidence interval (CI) of pairwise differencing (i.e., p value ≤ 0.05).

3. Analysis and results

a. *Climatology distribution*

Before presenting results from various skill metrics, some general rainfall characteristics of the forecasts will be discussed. A climatological frequency distribution of domain-averaged 12-h APCP (Fig. 4) suggests that WRF underpredicted and NAM overpredicted the number of null precipitation cases; these errors reduce the skill scores. The underprediction and overprediction of null precipitation can partly explain the wet bias of

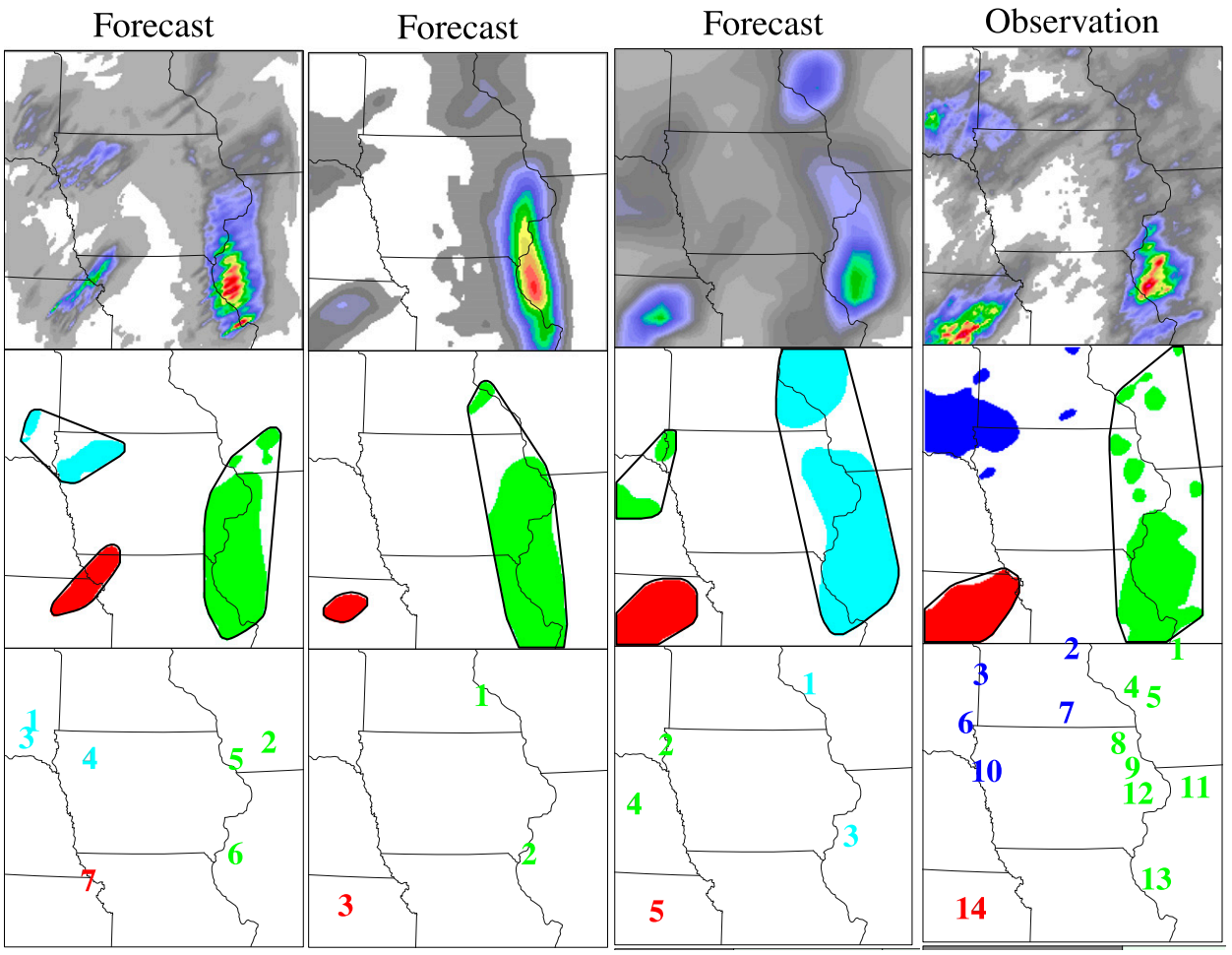

FIG. 3. An example of MODE output of (from left to right) WRF, NAM, GFS, and Stage IV valid for the period 1200–1500 UTC 9 Mar 2013, with model runs initialized at 1200 UTC 9 Mar 2013. (top) The forecasted and observed raw precipitation fields. (middle) The identified simple objects and (bottom) the denoted object index. The objects and index in the same color between different model fields indicate that these objects are matched, while the objects that are colored royal blue are unmatched.

WRF and dry bias of NAM, respectively. The overprediction of NAM may be caused by the triggering function of the BMJ cumulus scheme. Studies have shown that the BMJ scheme may trigger insufficiently often during the warm season (Xue et al. 2001). For heavy precipitation cases, WRF was the only model to suggest the true magnitude of heavy rain even though it still underpredicted the frequency of the heavy rainfall cases; NAM and GFS largely underestimated the rainfall amount and especially greatly underestimated the potential for more substantial rainfall amounts ranging from moderate to sufficient to cause severe flash floods [defined here to be the rightmost part (>6.35 mm) of each plot in Fig. 4]. The underestimation of the precipitation amount for heavy precipitation cases could be the result of the coarse effective resolution of NAM and GFS, because the coarse-resolution models cannot resolve well small-scale features and that fact, combined with deficiencies in the convective parameterizations, might prevent the production of higher precipitation amounts. The dry bias of NAM, which likely resulted in the low skill at moderate and high thresholds, was the most outstanding issue seen in the climatology.

b. *Traditional verification methods*

Traditional point-to-point verification methods are widely used to determine whether simulations can be regarded as "good" forecasts. Although traditional methods are sensitive to subtle displacements and deformations, they are applied on the raw fields without smoothing or convolving, so fewer tunable parameters influence the results. Because a 3-h accumulation interval is the minimum common temporal resolution for the three models, it is the primary accumulation interval that will be used in the following analysis. Diurnal variations of ETS [Eq. (1)], FAR [Eq. (3)], and FBIAS

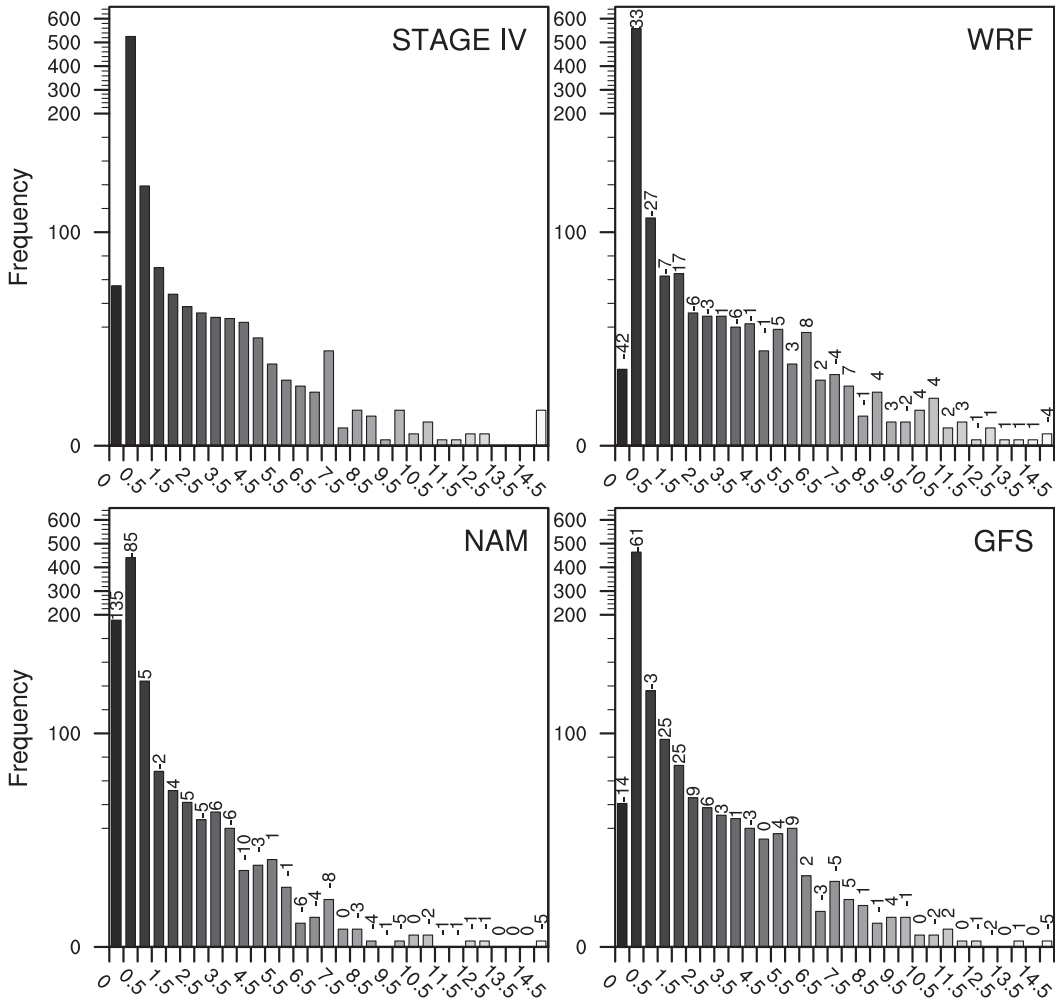

FIG. 4. Frequency distribution of domain-averaged 12-h accumulated precipitation (mm) from March to November. The leftmost bin represents cases with no precipitation. The numbers above each bar in the plots of WRF, NAM, and GFS refer to the differences in bin counts from that for the Stage IV data.

[Eq. (4)] using a low threshold of 0.254 mm and a high threshold of 6.35 mm are documented in Fig. 5. The oscillation of ETS for WRF indicates model skill changing with lead time, because all peaks occurred during the 0–3- and 6–9-h periods of each simulation, with lower scores in the 3–6- and 9–12-h periods. Moser et al. (2015) noted that the skill of hot runs decreased from a high value during the first 3 h and became steady during 6–12 h, a result that explains the periodic oscillation evident every 6 h in the 3-h verification methods (since WRF was run every 6 h in the present study). Ignoring the peak values in the first 3 h when the data assimilation in the WRF runs substantially increased scores, WRF did not show large advantages over the two operational models.

ETSs for NAM did not shown any advantage compared with GFS, even though the NAM output is from a much finer grid, suggesting that the interpolation from the NAM grid to the 4-km WRF grid might not be the major reason for the low skill. For GFS, the interpolation to Hres results in the extension of light precipitation near objects' boundaries, potentially explaining the much higher values of FAR (Fig. 5) and hit rate (not shown here) at the lowest threshold. In general, models had the higher ETSs during the early morning (0300–1500 UTC) and the lower skill during the late afternoon (1500–0000 UTC), but the portion of the false alarm became larger from morning to afternoon, resulting in the decrease of model skill. At the low threshold, WRF had the lowest value of FAR, but the FAR of WRF was the greatest at the high threshold. FBIAS of WRF at the high threshold also showed a similar result with a value much higher than the two other models and also much higher than 1.0. Thus, for large thresholds, WRF

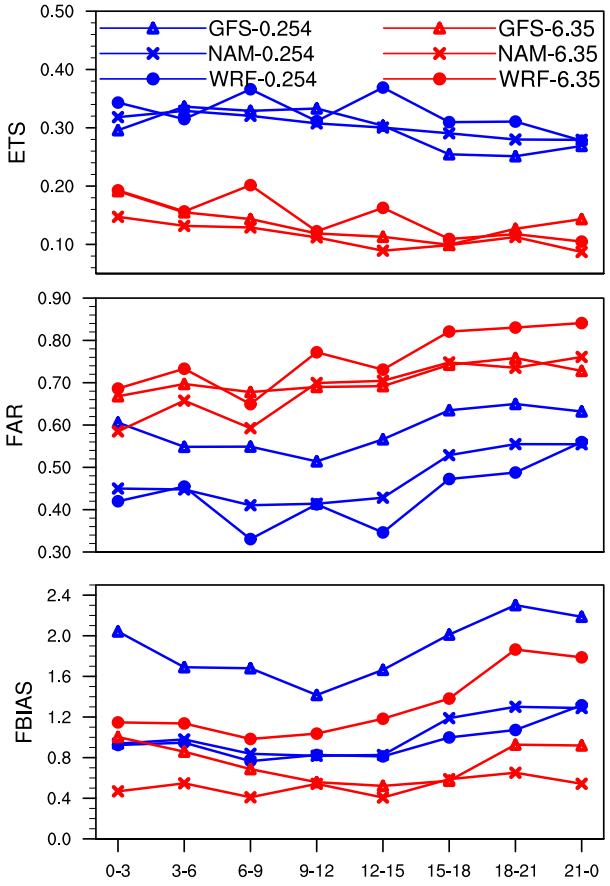

FIG. 5. Diurnal variations (UTC, along the *x* axis) of 3-h ETS, FAR, and FBIAS for the thresholds of 0.254 and 6.35 mm.

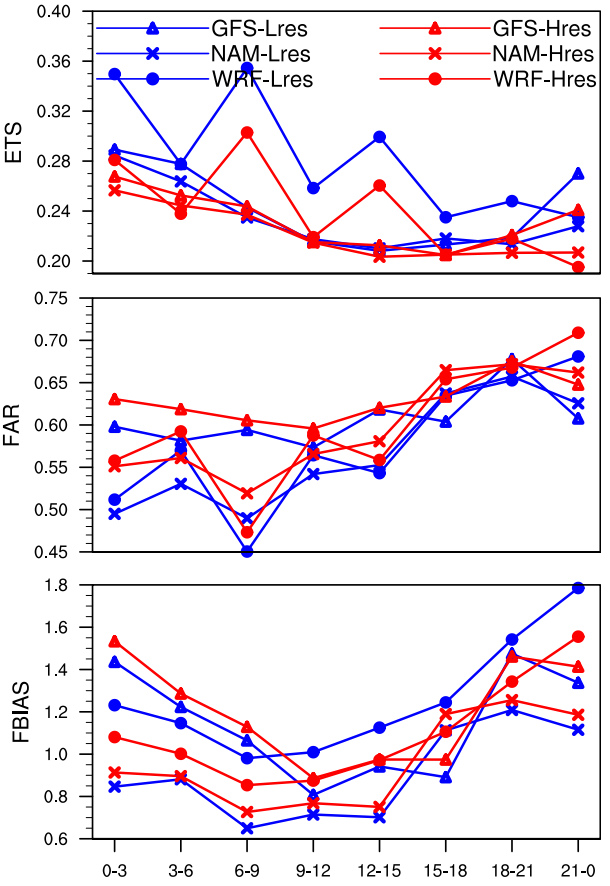

FIG. 6. Diurnal variation (UTC, along the *x* axis) of ETS, FAR, and FBIAS for the threshold of 2.54 mm on the WRF (Hres) and GFS (Lres) grids.

forecast precipitation too frequently, especially in the afternoon. In addition, the difference in FBIAS between WRF and the operational models was even larger than the difference in FAR, indicating that besides false alarms, WRF also might have a higher proportion of hits and a lower proportion of misses at the large threshold. The much larger FBIAS of GFS at the low threshold indicated that the GFS forecasted too frequently at low thresholds, which can also be explained by the large area of light precipitation resulting from the coarse resolution. However, for large thresholds, precipitating grid points in NAM and GFS were not forecasted frequently enough (FBIAS < 1). The ETS and FBIAS computed in the present study are comparable with the values found in Wolff et al. (2014), and several studies (Yang 2012a,b; Wolff et al. 2014) also found that the NAM had a significantly lower ETS and higher FBIAS than GFS.

Because of the likelihood that interpolation impacted the model skill scores, ETS, FAR, and FBIAS of 3-hourly aggregated QPFs were also computed on Lres using the threshold of 2.54 mm (Fig. 6), which is a moderate threshold for 3-h QPFs. In general, when all models were verified on the coarser-resolution grid, WRF showed a larger advantage for ETS even though GFS was now being verified on its own grid. This result is likely because the small-scale systems simulated by WRF were more realistic than those shown in the coarser-resolution operational models, and is also partly because the parameterized NAM and GFS could not resolve meteorological features explicitly. NAM and GFS differed little no matter which verification grid was used. For FAR and FBIAS, both NAM and GFS showed slightly higher values on Hres than Lres, but WRF had higher FARs on Hres than Lres, while the opposite was true for FBIAS. This suggests that the portion of hits increased while the portion of misses decreased on Lres. In general, models evaluated on Lres showed higher skill than on Hres. With traditional grid-to-grid metrics, it is often more difficult for high-resolution simulations to reach the same level of accuracy as low-resolution model runs because smooth features tend to be rewarded, and finescale details are penalized if spatial or temporal errors exist. Thus, it

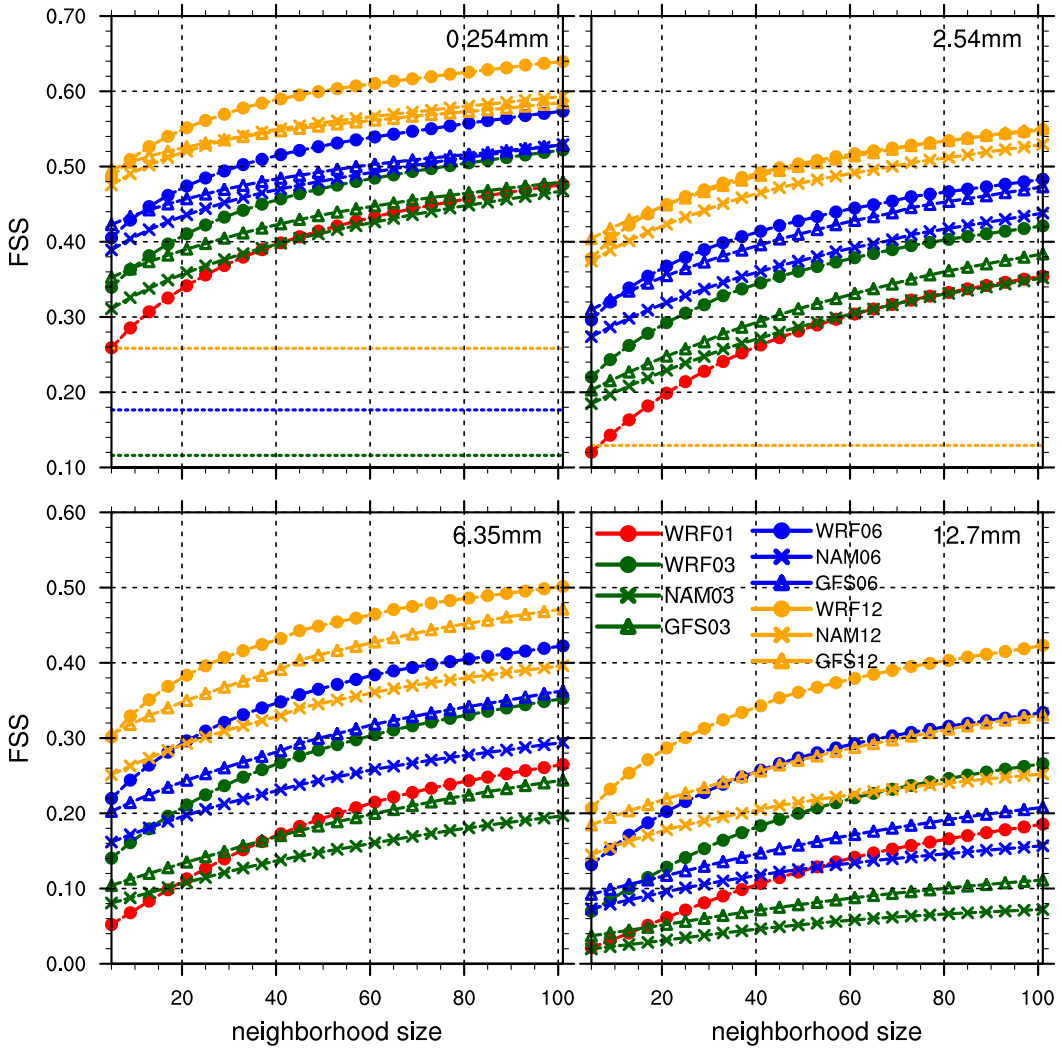

FIG. 7. Mean FSSs of 1-, 3-, 6-, and 12-h (01, 03, 06 and 12 in the legend) accumulation intervals for the three models (colored curves) as a function of neighborhood size (in grid units) for four rainfall thresholds. The yellow, blue, and green dotted lines represent the BRs of 12-, 6-, and 3-h accumulation intervals.

makes sense that Lres generally was more skillful than Hres considering the combined analysis of ETS and FAR, especially for WRF.

c. *FSS analysis*

To examine model performance with the increased horizontal and temporal scales, the mean FSS was computed, aggregated to various accumulation intervals over the whole 12-h simulation period (Fig. 7), using different thresholds. The mean BRs of 3-, 6-, and 12-h QPFs at 0.254 mm and 12-h QPFs at 2.54 mm are also shown in Fig. 7. The useful skill is given by $0.5 + f_0/2$. For other temporal accumulations and thresholds, useful skill can be approximated to be 0.5 because of the low mean BR over the 9 months. However, for larger thresholds such as 6.35 and 12.7 mm, almost none of the accumulation intervals and scales were as high as 0.5, and for moderate thresholds such as 2.54 mm, only 12-hourly QPFs of WRF at scales over 80 grid spacings could reach this useful skill value.

The FSS curves spanning 9 months (Fig. 7) show that, in general, the high-resolution WRF Model performed better than NAM and GFS, but the coarser GFS had better performance than NAM, partly as a result of the dry bias of the NAM. For low and moderate thresholds such as 0.254 and 2.54 mm, the superiority of WRF was not obvious and the skill of GFS was comparable with WRF for the threshold of 2.54 mm and the 12-h time accumulation, and this phenomenon was true particularly for the smaller neighborhoods. However, WRF

showed an advantage for higher thresholds, and the improvements of the scores compared with other models were as large as 0.05–0.1. Because of the better performance of GFS, WRFGFS was also evaluated in the experiment in order to check whether a better initialization applied to WRF would improve the QPF skill. However, skill scores for WRFGFS did not differ much compared with WRF (not shown here), so these different initializations did not have large effects on the high-resolution model QPF of this 9-month period making use of radar data assimilation.

Compared with NAM and GFS, WRF showed a larger improvement when increasing the horizontal scales, suggesting that the main issue for high-resolution models is that they are challenged at small spatial scales, especially for larger thresholds. Moreover, the improvement of FSS from 5 to 101 fuzzy lengths did not have much difference for 3-, 6-, and 12-h accumulation intervals. For example, the increase in FSS with spatial scales at a 3-h interval was similar to the increase with scales at a 12-h interval. In addition, doubling the time intervals led to a larger skill improvement than doubling the neighborhood scales regardless of the model examined. With the increased neighborhood sizes, more and more grid cells affect the calculation so that each cell has a smaller impact in the calculation than it would in smaller neighborhoods. However, with the increase in time intervals, no such diminishing of the importance of a cell occurs, since the threshold is fixed. Thus, the increase of FSS with increasing spatial scales at the same time interval was smaller than the increase of FSS with increasing time intervals at the same neighborhood size. For the purpose of increasing simulation QPF skill, an increased accumulation time interval is more important than increased spatial scales.

Because 3-h mean FSS failed to meet the threshold for useful skill, an appropriate criterion is needed to select a reasonable neighborhood scale for further analysis. For the threshold of 2.54 mm, FSS had a higher rate of increase within 25 smoothing scales than it did for larger scales, and this higher rate also existed for other time intervals and thresholds. This behavior is consistent with Wolff et al. (2014) and Mittermaier et al. (2013). Furthermore, at the scale of around 25 grid lengths, FSS reached half of the total FSS augmentation within the neighborhood scales used in this study [i.e., $FSS_{(25)} \approx 0.5(FSS_{(5)} + FSS_{(101)})$]. Moreover, this neighborhood scale did not cause too much smoothing and, thus, was used in the QPF skill analysis to be discussed next.

Diurnal cycles of QPF skill using a 3-h time interval and a 2.54-mm threshold are shown in Fig. 8. Hourly FSS of WRF is also shown in order to provide additional detail on variations with lead time. Similar to the

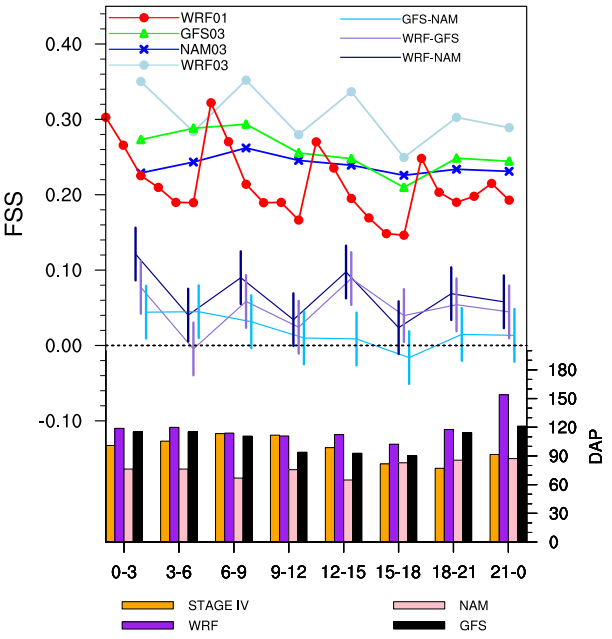

FIG. 8. Diurnal variation (UTC, along the *x* axis) of DAP (mm), and 3-h FSSs of three models as well as 1-h FSS of WRF at a threshold of 2.54 mm and a 25-grid-space neighborhood size. The *x*–*y* curves in the middle plot show the results of statistic tests of pairwise differencing between WRF, NAM, and GFS.

point-to-point verification methods, FSS for NAM and GFS did not exhibit statistically significantly obvious variations with lead time. The FSS results of NAM and GFS in the present study are about 0.05–0.1 lower than the FSS for the CONUS suggested by Wolff et al. (2014). This difference seems reasonable because many factors can influence FSS; for example, it is challenging to accurately predict strong convection in the central plains, and Wolff et al. (2014) only examined 0000 UTC initializations. The dry bias of NAM existed all day long except for a short period in the afternoon, while the wet bias of WRF and GFS existed for the entire day except during the early morning according to the 9-month accumulated domain-averaged APCP (DAP). The lowest skill happened when the rain volume did not show large bias errors (late morning, 1500–1800 UTC), so displacement or area/shape errors are likely the main cause. In the afternoon (1800–2400 UTC), the largest diurnal wet biases were found in WRF and GFS, but the FSS results for GFS and NAM had increased compared to the previous 3 h. Compared with night and morning, both the 3-h FSS of all models and the 1-h FSS of WRF showed that WRF did not lose skill during 2100–0000 UTC as might be expected because of the variation with lead time indicated by hourly WRF FSS. However, ETS (Fig. 5) suggested a lower level of skill compared with late morning, so whether the intensity error was the

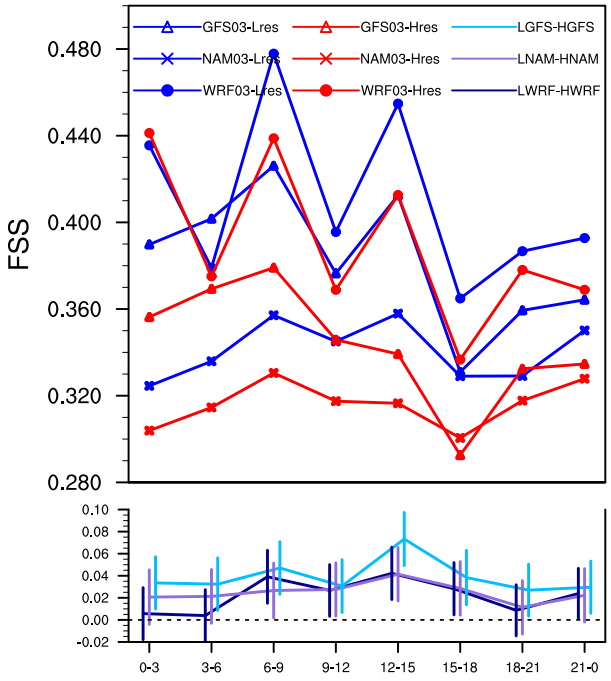

FIG. 9. Diurnal variation (UTC, along the *x* axis) of 3-h mean FSSs for Hres (red) and Lres (blue) verification grids at the smoothing sizes of 69 (Hres) and 5 (Lres) grid spacings, respectively, for the threshold of 2.54 mm. The bottom plot shows the statistic tests of pairwise differing between the skill scores on the Hres and Lres.

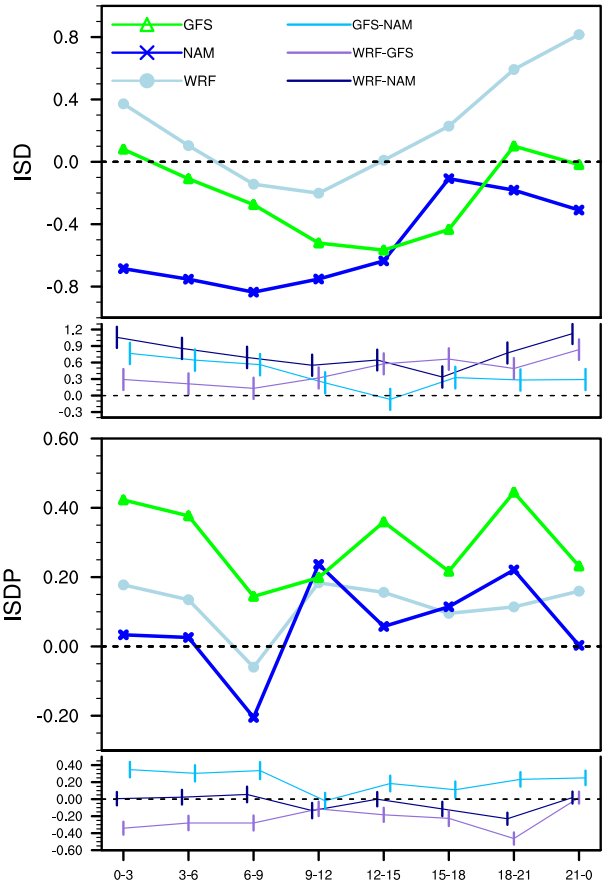

FIG. 10. Diurnal variation (UTC, along the *x* axis) of normalized ISD, and ISDP of 3- hourly QPFs. The bottom of each plot shows statistical tests of pairwise differencing.

main barrier for skill improvement and why FSS and ETS showed conflicting results needs to be studied further.

Higher scores for Lres (Fig. 6) may be due to the larger amount of smoothing, which increases model skill as indicated by Fig. 7. Hence, FSS of Lres at the neighborhood scale of 5 grid lengths, which is 275 km in the meridional direction and 212.5 km in the zonal direction, was compared with Hres at the scale of 69 grid spacings, which is 276 km in both directions. The FSS of models on the Lres grid (Fig. 9) was about 0.02–0.04 higher than the FSS on the Lres grid. The neighborhood scale of Lres is slightly smaller than that of Hres, so the higher skill of Lres than Hres is even more noteworthy than if the two neighborhoods had exactly the same scale. In addition, the model interpolated from the finest resolution, the WRF, showed a smaller difference between Lres and Hres, implying it was less influenced by the choice of the interpolation grid. FSS of GFS on the Lres was statistically significantly different from the FSS on the Hres. WRF and NAM had comparable time periods when the FSS were not statistically significantly different between Hres and Lres. However, at 0000–0300 and 0300–0600 UTC, WRF FSS differences were 0.005 while NAM was 0.03 and GFS was 0.025. The FSS of GFS was much increased on its own coarse grid, but the NAM still had the worst performance on both of the two grids. However, for WRF at the same neighborhood scale, FSS was not influenced as much as traditional methods were by the grid on which verification was done.

d. MODE

1) INTENSITY SUM

Intensity is often the item of most interest related to QPF, particularly for potential flood events. The 3-hourly QPF periods were also used to obtain feature attributes from MODE. More than 4000 forecasts of each model were used to study the attributes of IS difference, location errors, and areal coverage. The diurnal variations of mean normalized 3-h ISD [Eq. (6); Fig. 10] of WRF and NAM showed the same characteristics as DAP (Fig. 8), with WRF having the smallest wet bias during 0600–1500 UTC and NAM having the smallest dry bias during 1500–0000 UTC. WRF had an increasing wet bias during 1500–0000 UTC and reached a peak

value during 2100–0000 UTC. DAP of GFS also showed a wet bias for most times, which was even comparable to WRF, but the ISD of GFS was statistically significantly smaller than the ISD of WRF all day. Even though the wet bias of DAP became larger during 1500–0000 UTC, ISD still showed a negative value (or close to 0) because GFS had largely overpredicted the number of light and moderate cases indicated by the rainfall frequency distribution (Fig. 4). Hence, in general, GFS underpredicted the IS for 3-hourly QPF, despite the wet bias shown in DAP.

MODE produced a large number of attributes for matched pairs linking the model and observation fields, and these attributes can be used for more detailed comparisons of single storms. The ISDP [Eq. (6)] curves (Fig. 10) showed the normalized bias for each matched pair in the forecast–observation fields. The statistically significantly largest ISDP for GFS and the positive value for NAM suggest the coarse-resolution models did not have the capability to simulate localized storms, so the smaller objects in the observation field were matched with large forecast rain regions, which will be more comprehensively discussed in the section 3d(3). Even though the ISD of WRF kept increasing during the 1500–0000 UTC period, the ISDP curve did not show the same increasing trend and even decreased during this period. In other words, during the afternoon (1800–0000 UTC), WRF still performed well for the matched objects. However, those unmatched objects could contribute to the wet bias shown in ISD and DAP. The unmatched objects may be caused by the overprediction/underprediction of storms, which is also supported by Burghardt et al. (2014), and the possible existence of substantial location errors, because CD is the precondition and a necessary parameter for matching. It is also possible that other factors may also contribute to the low skill. During 0600–0900 UTC, GFS and WRF had the least ISDP bias, consistent with the high model skill shown for both models in FSS and ETS at this time.

2) Location errors

The feature attributes from MODE are analyzed based on simple objects from the convolved fields. WRF had the largest number of objects (Fig. 11), likely because the finer resolution is able to simulate objects of smaller scales. Compared with the Stage IV data, WRF performed very well before 1300 central daylight time (CDT; 1800 UTC). However, in the afternoon, the number of objects predicted by WRF was almost double that of the observations. It should be noted that the GFS produced a few more objects than NAM. Although NAM was run at a finer resolution, it did not show an advantage in producing small-scale storms.

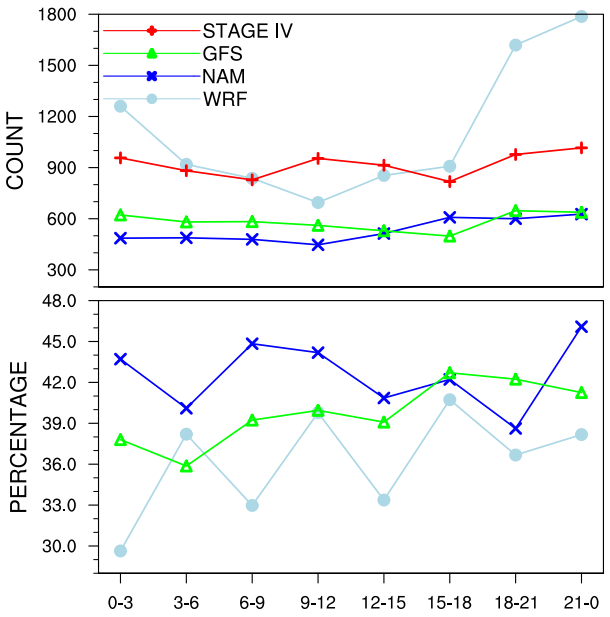

FIG. 11. Diurnal variation (UTC, along the *x* axis) of the number of simple objects of Stage IV, WRF, NAM, and GFS from MODE and the percentage of unmatched objects. The percentage is calculated as the sum of simple unmatched objects in the forecast and observation fields over the total number of objects in the forecast and observation fields.

The maximum CD, used to determine whether the objects in the model and observation fields could be matched, was set to be 50 grid spacings, which was a quarter of the length of the entire domain, so the potential for substantial location errors may be one possible reason for the unmatched objects. The bottom plot in Fig. 11 is the percentage of unmatched forecast and observed simple objects. Both the WRF and GFS reached a peak value between 1500 and 1800 UTC, indicating that many forecasted and observed objects might be unable to be matched because of location errors; however, the NAM simulations behaved differently. The NAM had the highest percentage of unmatched objects, but the percentage reached a relative minimum during 1800–2100 UTC.

Even though the percentage of unmatched objects in WRF kept decreasing from 1800 to 0000 UTC, there were still a large number of objects that could not be matched because of the large counts of simple objects. An examination of the magnitude of matched pairs shows that the magnitude of IS of an individual object predicted by WRF was highly accurate, but the overprediction of storms, especially for those storms that were far away from the observations, resulted in the wet bias of DAP and ISD.

The diurnal curves of the distance of the highest intensity (HID) and the mean of CD of matched simple

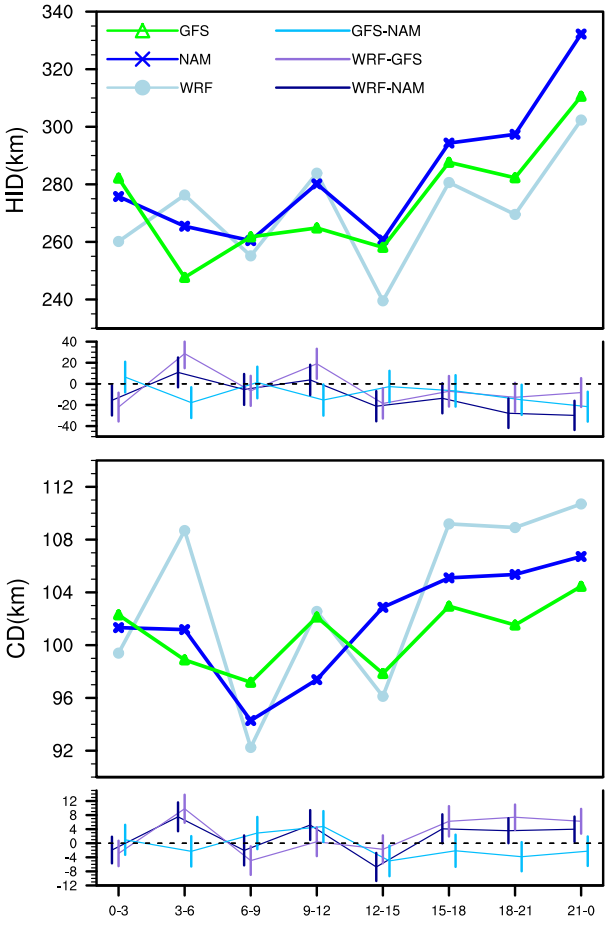

FIG. 12. Diurnal variation (UTC, along the *x* axis) of HID and CD (km). The bottom of each plot shows statistical tests of pairwise differencing.

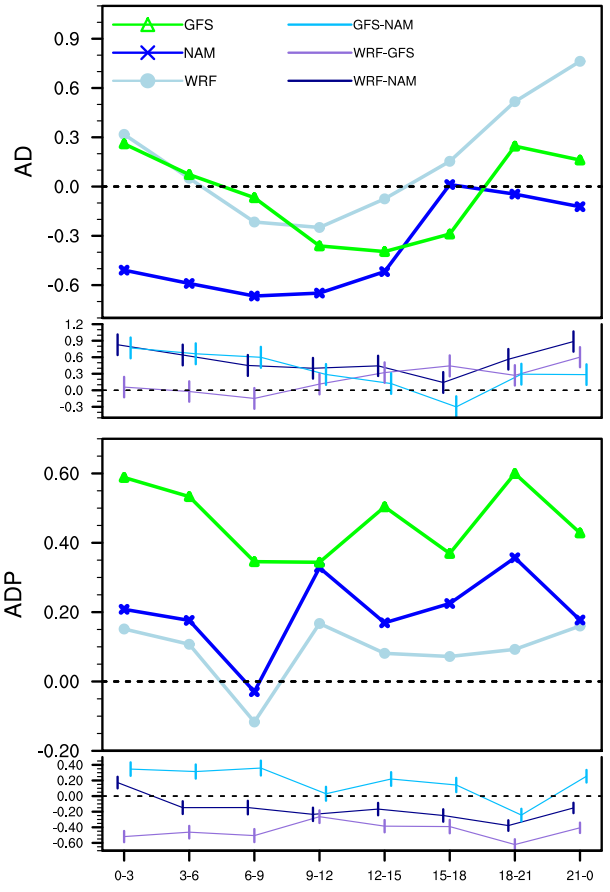

FIG. 13. Diurnal variation (UTC, along the *x* axis) of AD and ADP (grid squares). The bottom of each plot shows statistical tests of pairwise differencing.

objects are also shown in Fig. 12. HID was calculated using the locations of the grid points that had the highest intensity over the whole model and observation domains. If there was more than one grid point sharing the same highest intensity, the distance used was the mean distance of all the combinations of model–observation grid pairs. The highest intensity of QPF is an indicator of the location of the most intense part of the convective systems, which would be the region with an increased probability of flash floods or severe thunderstorms. While CD can be greatly influenced by a large region of light precipitation, HID was not restricted within 50 grid spacings, so it is necessary to compare HID to CD. In Fig. 7, the increased rate of FSS began to plateau around 25 grid spacings (~100 km). The CD also suggested a baseline of predictive skill around the scale of 100 km. Although WRF had a large mean CD, the HID showed that the location errors of WRF were not statistically significantly larger than NAM and GFS. From the curves of HID and CD, the location errors of all three models increased during 1500–0000 UTC, which was the main reason for the low ETS during that period, and the FSS had not kept decreasing with lead time during 1800–0000 UTC. The displacement can be corrected when the verified box is upscaled but the displacement reduces the skill scores for grid-to-grid verification methods. Both plots in Fig. 12 showed that displacement errors of WRF had an obvious variation with lead time, which was possibly an artifact of the radar data assimilation. In general, from midnight to early morning (0600–1500 UTC), the models tended to have smaller displacement errors.

3) AREAL COVERAGE

The diurnal mean normalized AD and ADP [Eq. (6); Fig. 13] showed a strong correlation with ISD (Fig. 10), with higher AD during 1500–0000 UTC and lower AD during 0600–1500 UTC. Though the WRF generally overpredicted ISDP, ADP showed that the

WRF had a very small coverage bias, suggesting that objects from WRF were much more intense. As with IS, especially in the local afternoon hours, the phenomenon of high AD with low ADP for WRF was contributed to by the existence of unmatched model objects and the overpredicted storm counts. The statistically significantly higher ADP for the NAM and GFS simulations was a result of the coarser resolution not being able to produce small-scale objects, which is consistent with the higher ISDP for NAM and GFS.

4. Conclusions and discussion

Multiple verification metrics were applied in this study to examine the skill of QPF obtained from convection-allowing ARW runs and to compare it with the skill of two coarser-grid operational NWP models for a small domain centered over Iowa. The ARW was run from March through November 2013 with 4-km grid spacing to better understand the limits of predictability of short-term (12 h) QPFs that might be used in hydrology models. WRF runs used both NAM and GFS output as the first-guess fields in the ARPS 3DVAR system, and then radar data were assimilated. Several verification methods were used to compare the QPFs from the three models. NCEP Stage IV precipitation data were used to represent ground truth in the verification process. WRF, NAM, GFS, and Stage IV output were interpolated using a water budget preservation approach to both the 4-km WRF grid and the roughly 55-km GFS grid. Additional diagnostic information was obtained from the relatively newly developed neighborhood and object-based techniques of FSS and MODE, respectively. These two spatial methods provided some additional guidance on specific issues of interest such as horizontal and temporal scales, intensity and location errors, coverage errors, and hit rates, among others, for the precipitation systems.

QPF skill was rather poor, using standard definitions for FSS, in all three models tested. Only the 12-h QPF of WRF at or smaller than the threshold of 2.54 mm was able to reach the uniform skill threshold. For the threshold of 2.54 mm, 12-h QPFs could be reliable at the scale of 320 km. However, the threshold of 2.54 mm is too light for hydrology concerns for 12-h QPFs, and the scale of 320 km causes more smoothing than desired. At the scale of 100 km (25 grid squares), FSS began to plateau, and the FSS here was roughly half of the total FSS augmentation, so this scale was used for more detailed skill evaluation. It was found that QPF skill increased more as the accumulation time interval increased than for increased spatial scale.

In general, NAM performed the worst among the three models evaluated in this study, not only for model skill over the full domain but also for characteristics of spatial features. A large DAP dry bias and substantial location errors existed for almost the entire forecast period. Besides the insufficient trigging of the BMJ scheme (Xue et al. 2001), Wang et al. (2009) also found the NAM to underpredict the rainfall amount and indicated that the NAM is unable to generate atmospheric moisture sufficiently over the central CONUS, which results in too weak convergence of the water vapor flux. In addition, the finer resolution of NAM did not show any advantages in predicting small-scale storms than the GFS. The high-resolution WRF model had a much higher skill for larger thresholds, and this was not only indicated by the neighborhood method but also was suggested by traditional techniques, which usually favor the smoother forecast fields of coarser-resolution models. In addition, WRF had the smallest displacement errors and was able to most correctly forecast the intensity magnitude of simple objects. The better performance of WRF in these aspects may show the importance of running convection-allowing models to obtain the most accurate QPFs. WRF was able to simulate localized storms, but the WRF was generally too widespread with precipitation in the afternoon, resulting from an overprediction of storm counts. Besides better skill scores, WRF also performed better with object intensity magnitude, areal coverage, and the location of most intense part of the systems. Considering the possibility that the high skill of GFS was an artifact related to the large amount of smoothing to get its output onto the WRF grid, the verification was also performed on a low-resolution grid. However, the NAM still showed the lowest skill on the Lres grid. The scores for the high-resolution WRF model were less influenced by the grid on which the verification was done.

Overall, the models had the highest skill from midnight to early morning. Because this period had the smallest bias, location, and coverage errors, all three models were able to correctly forecast the frequency of events and had fewer false alarms, resulting in the most reliable QPF over the entire day during this period. One possible reason is that convective systems are larger scale and more organized at night, while initiation is a known difficulty in models, and that is more likely in the afternoon. The lowest skill occurred from late morning to afternoon, but at the same time, the NAM and GFS had the least dry bias and areal coverage errors while the WRF had small intensity and coverage errors in the afternoon. For hydrological use, in order to obtain skillful QPFs during this period, besides the

overprediction/underprediction of storm numbers, more attention should be paid to the large location errors. The displacement errors started to grow in late morning and reached a peak value during the late afternoon. Because the displacement errors can be partly corrected with the increasing of scales, FSS did not keep decreasing in the late afternoon.

The present study is a preliminary exploration of the evaluation of QPF from models using multiple verification methods, and additional work is needed. Future work should be performed using a much larger domain. Additional analysis is needed to determine why all of the models have large displacement errors in late morning and afternoon over the Iowa region and how to fix the errors. Are these predicted storms displaced behind (possibly because they formed too slowly) or ahead of (formed too rapidly) the observations? Moreover, approaches that would reduce the overprediction of the number of convective systems in WRF should be investigated. These approaches could also help to fix the overestimation of DAP.

Acknowledgments. This research was supported primarily by the Iowa Flood Center with some additional support from National Science Foundation Grant AGS1222383. The rainfall observations [Precipitation NCEP/EMC 4KM Gridded Data (GRIB) Stage IV Data] are provided by NCAR/EOL under sponsorship of the National Science Foundation and are available online (http://data.eol.ucar.edu/).

APPENDIX

Key Abbreviations and Acronyms Used in This Paper

AD	The normalized areal coverage difference (in grid squares) of the whole domain between the observed field and the forecasted field
ADP	The normalized areal coverage difference of the two simple objects had been matched between the observed field and the forecasted field
APCP	Accumulated precipitation (mm)
BR	Base rate
CD	Centroid distance (km) of two matched simple objects
CI	Confidence interval
CSI	Critical success index
DAP	Domain-averaged APCP accumulated over 9 months
ETS	Equitable threat score (also known as Gilbert skill score)
FAR	False alarm rate
FBIAS	Frequency bias
FSS	Fractions skill score
HID	Distance between the highest intensity point in the observed and forecasted fields
Hres	High-resolution model simulation, using the same domain configuration as in WRF
IS	Intensity sum, also known as total rain volume (mm)
ISD	Normalized IS difference of the whole domain between the observed field and the forecasted field
ISDP	Normalized IS difference of the two simple objects had been matched between the observed field and the forecasted field
Lres	Low-resolution model simulation, using the same domain configuration as in GFS
MET	Model Evaluation Tools
MODE	Method for Object-Based Diagnostic Evaluation
NWP	Numerical weather prediction
PODY	Probability of detection
QPF	Quantitative precipitation forecast
WRFGFS	QPFs using the initializations and lateral boundary conditions as in GFS

REFERENCES

Barthold, F. E., T. E. Workoff, B. A. Cosgrove, J. J. Gourley, D. R. Novak, and K. M. Mahoney, 2015: Improving flash flood forecasts: The HMT-WPC Flash Flood and Intense Rainfall Experiment. *Bull. Amer. Meteor. Soc.*, **96**, 1859–1866, doi:10.1175/BAMS-D-14-00201.1.

Burghardt, B. J., C. Evans, and P. J. Roebber, 2014: Assessing the predictability of convection initiation in the high plains using an object-based approach. *Wea. Forecasting*, **29**, 403–418, doi:10.1175/WAF-D-13-00089.1.

Chou, M. D., and M. J. Suarez, 1994: An efficient thermal infrared radiation parameterization for use in general circulation models. NASA Tech. Memo. 104606, Vol. 3, 84 pp. [Available online at http://gmao.gsfc.nasa.gov/pubs/docs/Chou128.pdf.]

Cloke, H. L., and F. Pappenberger, 2009: Ensemble flood forecasting: A review. *J. Hydrol.*, **375**, 613–626, doi:10.1016/j.jhydrol.2009.06.005.

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

——, ——, and ——, 2006b: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Mon. Wea. Rev.*, **134**, 1785–1795, doi:10.1175/MWR3146.1.

——, ——, ——, and J. Halley-Gotway, 2009: The method for object-based diagnostic evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC Spring Program. *Wea. Forecasting*, **24**, 1252–1267, doi:10.1175/2009WAF2222241.1.

Duda, J. D., and W. A. Gallus Jr., 2013: The impact of large-scale forcing on skill of simulated convective initiation and upscale evolution with convection-allowing grid spacings in the WRF. *Wea. Forecasting*, **28**, 994–1018, doi:10.1175/WAF-D-13-00005.1.

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

Fritsch, J. M., and K. F. Heideman, 1989: Some characteristics of the Limited-Area Fine-Mesh (LFM) model quantitative precipitation forecasts (QPF) during the 1982 and 1983 warm seasons. *Wea. Forecasting*, **4**, 173–185, doi:10.1175/1520-0434(1989)004<0173:SCOTLA>2.0.CO;2.

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, doi:10.1175/2009WAF2222269.1.

Halley-Gotway, J., and Coauthors, 2014: Model Evaluation Tools version 5.0 (METv5.0): User's guide 5.0. Developmental Testbed Center Rep., 241 pp. [Available online at http://www.dtcenter.org/met/users/docs/users_guide/MET_Users_Guide_v5.0.pdf.]

Heideman, K. F., and J. M. Fritsch, 1988: Forcing mechanisms and other characteristics of significant summertime precipitation. *Wea. Forecasting*, **3**, 115–130, doi:10.1175/1520-0434(1988)003<0115:FMAOCO>2.0.CO;2.

Hu, M., M. Xue, J. Gao, and K. Brewster, 2006: 3DVAR and cloud analysis with WSR-88D level-II data for the prediction of the Fort Worth, Texas, tornadic thunderstorms. Part II: Impact of radial velocity analysis via 3DVAR. *Mon. Wea. Rev.*, **134**, 699–721, doi:10.1175/MWR3093.1.

Janjić, Z. I., 1994: The step-mountain Eta coordinate model: Further development of the convection, viscous sublayer, and turbulent closure schemes. *Mon. Wea. Rev.*, **122**, 927–945, doi:10.1175/1520-0493(1994)122<0927:TSMECM>2.0.CO;2.

——, 2003: A nonhydrostatic model based on a new approach. *Meteor. Atmos. Phys.*, **82**, 271–285, doi:10.1007/s00703-001-0587-6.

Jankov, I., W. A. Gallus Jr., M. Segal, and S. E. Koch, 2007a: Influence of initial conditions on the WRF-ARW model QPF response to physical parameterization changes. *Wea. Forecasting*, **22**, 501–519, doi:10.1175/WAF998.1.

——, P. J. Schultz, C. J. Anderson, and S. E. Koch, 2007b: The impact of different physical parameterizations and their interactions on cold season QPF in the American River basin. *J. Hydrometeor.*, **8**, 1141–1151, doi:10.1175/JHM630.1.

Kain, J. S., and Coauthors, 2013: A feasibility study for probabilistic convection initiation forecasts based on explicit numerical guidance. *Bull. Amer. Meteor. Soc.*, **94**, 1213–1225, doi:10.1175/BAMS-D-11-00264.1.

Lin, Y., and K. Mitchell, 2005: The NCEP stage II/IV hourly precipitation analyses: development and applications. Preprints, *19th Conf. on Hydrology*, San Diego, CA, Amer. Meteor. Soc., 1.2. [Available online at https://ams.confex.com/ams/pdfpapers/83847.pdf.]

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

——, ——, and S. A. Thompson, 2013: A long-term assessment of precipitation forecast skill using the fractions skill score. *Meteor. Appl.*, **20**, 176–186, doi:10.1002/met.296.

Moser, B., W. A. Gallus Jr., and R. Mantilla, 2015: An initial assessment of radar data assimilation on warm season rainfall forecasts for use in hydrologic models. *Wea. Forecasting*, **30**, 1491–1520, doi:10.1175/WAF-D-14-00125.1.

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

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

Rogers, E., and Coauthors, 2009: The NCEP North American Mesoscale modeling system: Recent changes and future plans. Preprints, *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., 2A.4. [Available online at http://ams.confex.com/ams/pdfpapers/154114.pdf.]

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., doi:10.5065/D68S4MVH. [Available online at http://www2.mmm.ucar.edu/wrf/users/docs/arw_v3.pdf.]

Sukovich, E. M., F. M. Ralph, F. E. Barthold, D. W. Reynolds, and D. R. Novak, 2014: Extreme quantitative precipitation forecast performance at the Weather Prediction Center from 2001 to 2011. *Wea. Forecasting*, **29**, 894–911, doi:10.1175/WAF-D-13-00061.1.

Sun, J., S. B. Trier, Q. Xiao, M. L. Weisman, H. Wang, Z. Ying, M. Xu, and Y. Zhang, 2012: Sensitivity of 0–12-h warm-season precipitation forecasts over the central United States to model initialization. *Wea. Forecasting*, **27**, 832–855, doi:10.1175/WAF-D-11-00075.1.

Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, doi:10.1175/2008MWR2387.1.

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, doi:10.1175/2009WAF2222185.1.

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

Xiao, Q., Y. H. Kuo, J. Sun, W. C. Lee, D. M. Barker, and E. Lim, 2007: An approach of radar reflectivity data assimilation and its assessment with the inland QPF of Typhoon Rusa (2002) at landfall. *J. Appl. Meteor. Climatol.*, **46**, 14–22, doi:10.1175/JAM2439.1.

Xue, M., K. K. Droegemeier, V. Wong, A. Shapiro, and K. Brewster, 1995: ARPS Version 4.0 User's Guide. Center for Analysis and Prediction of Storms, 380 pp. [Available online at http://www.caps.ou.edu/ARPS/arpsdoc.html.]

——, ——, and ——, 2000: The Advanced Regional Prediction System (ARPS)—A multiscale nonhydrostatic atmospheric simulation and prediction tool. Part I: Model dynamics and

verification. *Meteor. Atmos. Phys.*, **75**, 161–193, doi:10.1007/s007030070003.

——, D.-H. Wang, J.-D. Gao, K. Brewster, and K. K. Droegemeier, 2003: The Advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation. *Meteor. Atmos. Phys.*, **82**, 139–170, doi:10.1007/s00703-001-0595-6.

Xue, Y., F. J. Zeng, K. E. Mitchell, Z. Janjić, and E. Rogers, 2001: The impact of land surface processes on simulations of the U.S. hydrological cycle: A case study of the 1993 flood using the SSiB land surface model in the NCEP Eta regional model. *Mon. Wea. Rev.*, **129**, 2833–2860, doi:10.1175/1520-0493(2001)129<2833:TIOLSP>2.0.CO;2.

Yang, F., 2012a: Review of GFS forecast skills in 2012. IMSG–Environmental Modeling Center, National Centers for Environmental Prediction. [Available online at http://www.atmos.albany.edu/daes/atmclasses/atm401/PPTs-PDFs_files/GFS.performance.review.2012.pdf.]

——, 2012b: Review of NCEP GFS forecast skills in 2011 and beyond. *46th CMOS Congress/21st Conf. on Numerical Weather Prediction/25th Conf. on Weather Analysis and Forecasting*, Montréal, QC, Canada, Canadian Meteorological and Oceanographic Society–Amer. Meteor. Soc. [Available online at http://web2.sca.uqam.ca/~wgne/CMOS/PRESENTATIONS/5330_2b1.5_yang_fanglin.pdf.]