Generation of Ensemble Mean Precipitation Forecasts from Convection-Allowing Ensembles

ADAM J. CLARK

NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma

(Manuscript received 17 November 2016, in final form 14 June 2017)

ABSTRACT

Methods for generating ensemble mean precipitation forecasts from convection-allowing model (CAM) ensembles based on a simple average of all members at each grid point can have limited utility because of amplitude reduction and overprediction of light precipitation areas caused by averaging complex spatial fields with strong gradients and high-amplitude features. To combat these issues with the simple ensemble mean, a method known as probability matching is commonly used to replace the ensemble mean amounts with amounts sampled from the distribution of ensemble member forecasts, which results in a field that has a bias approximately equal to the average bias of the ensemble members. Thus, the probability matched mean (PM mean hereafter) is viewed as a better representation of the ensemble members compared to the mean, and previous studies find that it is more skillful than any of the individual members. Herein, using nearly a year's worth of data from a CAM-based ensemble running in real time at the National Severe Storms Laboratory, evidence is provided that the superior performance of the PM mean is at least partially an artifact of the spatial redistribution of precipitation amounts that occur when the PM mean is computed over a large domain. Specifically, the PM mean enlarges big areas of heavy precipitation and shrinks or even eliminates smaller ones. An alternative approach for the PM mean is developed that restricts the grid points used to those within a specified radius of influence. The new approach has an improved spatial representation of precipitation and is found to perform more skillfully than the PM mean at large scales when using neighborhood-based verification metrics.

1. Introduction

One of the simplest ways to utilize ensemble forecasts is by computing the ensemble mean, which is a simple arithmetic average of ensemble members at each grid point within a forecast domain. As many past studies have found (e.g., Leith 1974; Murphy 1988), when averaged over many cases, the ensemble mean will have a smaller error than the individual ensemble members. This error reduction occurs because high-predictability features that the members agree on are emphasized by the mean, while low-predictability features that the members do not agree on are filtered out or heavily dampened (e.g., Surcel et al. 2014). Although important information on forecast uncertainty from the ensemble probability distribution function (PDF) is lost when taking an ensemble mean, it is utilized in operational forecasting as a summary tool, and/or as a starting point for further analysis of ensemble PDF characteristics.

While the ensemble mean works well for fields used to diagnose large-scale weather patterns such as 500-hPa geopotential heights and winds, it has been recognized that for complex spatial fields such as precipitation that have steep gradients and high amplitudes, the mean "smears out" the important features by reducing the high amplitudes and increasing the spatial coverage of light values (e.g., Ebert 2001; Clark et al. 2008; Fang and Kuo 2013; Surcel et al. 2014). Thus, when applied to such forecasts, the ensemble mean can be a poor representation of the individual ensemble members and, thus, has limited use in forecasting, especially for extreme events. These problems with the ensemble mean are especially relevant to convection-allowing model (CAM) ensembles,[1] which are being increasingly

Corresponding author: Adam J. Clark, adam.clark@noaa.gov

[1] CAM ensembles are composed of members that explicitly depict convection (i.e., do not use convective parameterization) and typically use grid spacing of 3–4 km, which is about the coarsest resolution at which the bulk circulations in midlatitude mesoscale convective systems can be adequately resolved (e.g., Weisman et al. 1997).

DOI: 10.1175/WAF-D-16-0199.1

utilized by the operational forecasting community (e.g., Clark et al. 2012; Jirak et al. 2012, 2014; Evans et al. 2014; Barthold et al. 2015).

Recognizing the problems with the ensemble mean precipitation, Ebert (2001) proposed an alternative method for computing the mean known as probability matching, which involves replacing the ensemble mean amounts with amounts sampled from the distribution of ensemble member forecasts. The probability matched mean (PM mean hereafter) has the same spatial pattern as the ensemble mean, but the highest amplitudes from the individual members are retained, and the bias at all precipitation exceedance thresholds is nearly equal to the average bias of the ensemble members for any specified threshold. Thus, the PM mean is viewed as a better representation of the ensemble members, and several recent studies find that it is usually more skillful than the individual members and the ensemble mean (e.g., Clark et al. 2009; Kong et al. 2009; Xue et al. 2011; Berenguer et al. 2012; Schwartz et al. 2014).

However, by examining how the PM mean impacts the spectral structure and spatial distribution of precipitation forecasts derived from a CAM ensemble, Surcel et al. (2014) demonstrated that the gain in skill from the PM mean results from reduced variability in the PM mean at small scales relative to the individual ensemble members. By performing an experiment where they modify the power spectrum of precipitation forecasts from the control member of a CAM ensemble so its power spectrum is the same as that derived from the PM mean (i.e., the forecasts are modified to have the same spatial variability), Surcel et al. (2014) found that the control member has slightly higher skill than the PM mean, while before the control member was modified, the PM mean had higher skill. Indeed, it is well known that objective verification metrics will often favor a ''smoother'' forecast over a ''noisier'' one with more small-scale variability (e.g., Gilleland et al. 2009).

Herein, a different aspect of the PM mean is examined when applied to a CAM ensemble. Namely, this work examines the possible negative impacts that result when applying the PM mean over a large domain like the contiguous United States. Because of how the PM mean is computed, precipitation amounts from the individual ensemble members over a region such as Florida can be reassigned to a very different region, like the central United States. In this case, when averaged over the whole domain, the distribution of PM-mean values is representative of the individual members, but it is possible that over regional areas it is not representative, which could impact objective skill metrics. This impact is examined by formulating a new approach to computing the PM mean that only considers the distribution of ensemble member precipitation amounts within a specified radius of influence. This new approach is termed the localized PM mean, and results from this new method are compared to the PM mean. The remainder of this study is organized as follows. Section 2 contains a description of the data and methodology, section 3 presents the results, and section 4 provides a summary and discussion.

2. Data and methodology

a. Ensemble forecast system

Ensemble mean precipitation is computed from a 10-member, 4-km grid-spacing experimental CAM ensemble initialized daily at 0000 UTC with forecasts to 36 h and a domain encompassing the entire continental United States (CONUS) (Fig. 1). Each member uses version 3.4.1 of the Weather Research and Forecasting (WRF) Model with the Advanced Research WRF (ARW) dynamics core (Skamarock et al. 2008). One member uses NCEP's North American Mesoscale Forecast System (NAM; Rogers et al. 2009) for initial and lateral boundary conditions (ICs/LBCs), another member uses NCEP's Global Forecast System for ICs/LBCs, and the eight other members are initialized at 0000 UTC using 3-h forecasts from NCEP's Short-Range Ensemble Forecast (SREF; Du et al. 2014) system initialized at 2100 UTC for ICs and corresponding SREF member forecasts for LBCs. All members use the same physics parameterizations, which include the Mellor–Yamada–Janjić (MYJ; Mellor and Yamada 1982; Janjić 2002) planetary boundary layer scheme, WRF single-moment 6-class (WSM-6; Hong and Lim 2006) microphysics, the Noah (Chen and Dudhia 2001) land surface model, and the Rapid Radiative Transfer Model (RRTM; Mlawer et al. 1997) longwave radiation and Dudhia (Dudhia 1989) shortwave radiation scheme. This ensemble, which is known as the NSSL-WRF ensemble, is run by the National Severe Storms Laboratory using a NOAA-funded high-performance computing cluster known as Jet, housed in Boulder, Colorado. The NSSL-WRF ensemble is part of a permanent modeling framework established in 2006, which provides CAM-based guidance to NOAA's Storm Prediction Center forecasters and serves as a testing ground for the development of new and innovative CAM-based diagnostics (e.g., Kain et al. 2010; Gallo et al. 2016). Table 1 provides a summary of the model specifications. Precipitation forecasts are examined for the period 7 October 2014–21 October 2015. Because of sporadic missing data, 317 cases within this period were examined that contain a complete dataset from all ensemble

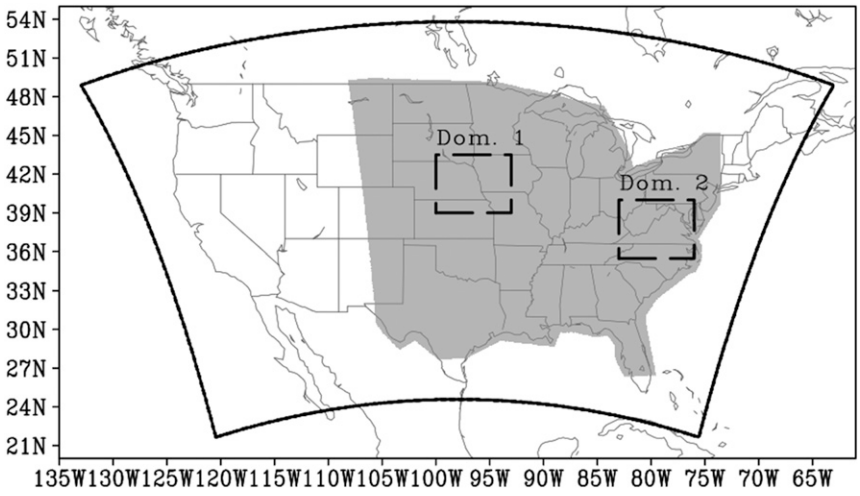

FIG. 1. The thick black line depicts the domain used for integrating the NSSL-WRF ensemble members. The shaded region is the analysis domain over which all statistics were computed. The dashed rectangles outline domains pictured in Fig. 4.

members at all forecast hours. Table 2 contains a list of the cases that were examined.

b. *Description of PM mean and localized PM mean*

The PM mean and localized PM mean were computed for 24-h accumulated precipitation from forecast hours 12–36 (i.e., valid 1200–1200 UTC) over a subset of the model domain pictured in Fig. 1. This subset was chosen because it corresponds to a region with very reliable coverage provided by the observational dataset used for verification: NCEP's Stage IV dataset (Baldwin and Mitchell 1997; Lin and Mitchell 2005). The procedure followed for computing the PM mean is as follows: 1) calculate the traditional ensemble mean (i.e., simple arithmetic average of all members) at each grid point; 2) sort the values of the ensemble mean from lowest to highest, storing the rank (i.e., numbered position of the sorted values; e.g., the fourth lowest value has a rank of 4) and location of each value; 3) sort the values of all ensemble members from lowest to highest; 4) select every nth (n = number of members) value from the array of ranked ensemble member values; and 5) swap the values from step 2 with those from step 5 (i.e., the grid point with the highest precipitation amount in the ensemble mean is replaced by the highest value in the distribution of ensemble members, and so on).

The procedure for the localized PM mean is as follows: 1) calculate the traditional ensemble mean at each grid point (i.e., same as step 1 for the PM mean); 2) within a 126-km radius of each grid point, sort the values of ensemble mean from lowest to highest and record the rank R_{mean} of the center point; 3) within a 126-km radius of each grid point, sort the values of all ensemble members from lowest to highest; and 4)

TABLE 1. NSSL-WRF ensemble member specifications. ICs/LBCs in members 3–10 refer to different 2100 UTC initialized SREF members from which 3-h forecasts are used for ICs.

Member	ICs/LBCs	PBL scheme	Microphysics	Land surface	Radiation
1	0000 UTC NAM	MYJ	WSM6	Noah	RRTM/Dudhia
2	0000 UTC GFS	MYJ	WSM6	Noah	RRTM/Dudhia
3	2100 UTC EM_CTl	MYJ	WSM6	Noah	RRTM/Dudhia
4	2100 UTC NMB_N1	MYJ	WSM6	Noah	RRTM/Dudhia
5	2100 UTC NMB_P2	MYJ	WSM6	Noah	RRTM/Dudhia
6	2100 UTC NMB_CTl	MYJ	WSM6	Noah	RRTM/Dudhia
7	2100 UTC NMB_P1	MYJ	WSM6	Noah	RRTM/Dudhia
8	2100 UTC NMM_CTl	MYJ	WSM6	Noah	RRTM/Dudhia
9	2100 UTC NMM_N1	MYJ	WSM6	Noah	RRTM/Dudhia
10	2100 UTC NMM_P1	MYJ	WSM6	Noah	RRTM/Dudhia

TABLE 2. List of dates used in the analysis (317 total cases).

Oct 2014	7–15, 17–28, 30–31
Nov 2014	1, 3, 5–19, 21, 23–26, 28–30
Dec 2014	1, 3–7, 9–12, 15–19, 21–31
Jan 2015	2–14, 21–31
Feb 2015	1–4, 6–10, 12–18, 20–28
Mar 2015	3–5, 8, 10–31
Apr 2015	1–5, 8–17, 19–30
May 2015	1–3, 5–12, 14, 19–21, 23–24, 26–31
Jun 2015	1–4, 6–9, 11–13, 15–19, 22, 24–30
Jul 2015	1–9, 13–14, 16–17, 19–28, 31
Aug 2015	2–9, 11–14, 20–27, 29–31
Sep 2015	1–11, 13–27, 29–30
Oct 2015	1–7, 9–11, 13–21

replace the value of the neighborhood center point with the sorted ensemble member value ranked R_{ens}. The simplest way to compute R_{ens} is a linear calculation of rank that can be expressed as $R_{ens} = M(R_{mean})$, where $M = 10$ is the number of ensemble members. However, it was found that a linear calculation of rank resulted in precipitation biases for the lighter rainfall thresholds (i.e., 0.10, 0.25, and 0.50 in.) that were slightly lower than those of the PM mean when averaged over the entire domain. For higher rainfall thresholds (i.e., 1.00 and 2.00 in.), the precipitation biases were even lower using the linear rank calculation. Thus, a method was devised to adjust the R_{ens} ranks so that they were slightly higher than that calculated from the simple linear method, which adjusted the ranks more at the higher end of the precipitation distributions where the differences in bias compared with the PM mean were more pronounced. This method was implemented using the following formula:

$$R_{ens} = \text{nint}\left[\frac{-MN(N - R_{mean})^{\sigma}}{N^{\sigma}}\right] + MN, \qquad (1)$$

where $N = 3125$ is the number of points or ranks within each 126-km radius, and $\sigma = 1.05$ is a coefficient that eventually determines how much R_{ens} deviates from the simple linear calculation of rank. This function can essentially be thought of as a normalized exponential, where the exponential term R_{mean}^{σ} is normalized by N/N^{σ}. Then, the function is "flipped" [by taking the negative (i.e., the $-M$ term)] and "reversed" by using $(N - R_{mean})^{\sigma}$ instead of R_{mean}^{σ}.

The ideal value for σ was found through experimentation with several different values. Figure 2 illustrates how R_{ens} varies with σ. As an example, for $\sigma = 1.05$ and $R_{mean} = 2000$, $R_{ens} = 20\,560$, which is slightly greater than $R_{ens} = 10R_{mean} = 20\,000$. Essentially, the localized PM mean accomplishes the same thing as the PM mean, but over a limited area surrounding each grid point. A

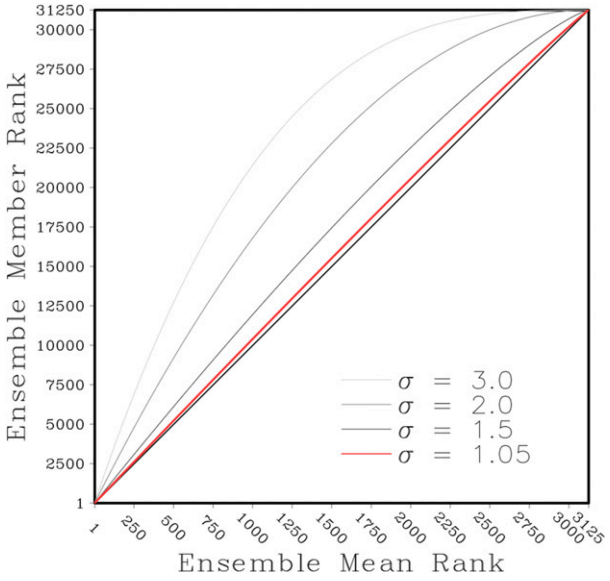

FIG. 2. Solutions for the function used to calculate the ensemble member rank in the computation of the localized PM mean. The value $\sigma = 1.05$ is used herein, while the other σ results are shown to illustrate the shape of the different function solutions. A straight diagonal line (black) is shown for reference and represents a linear calculation of rank.

126-km radius was chosen after experimenting with several radii. An example of these experiments for forecasts initialized at 0000 UTC 6 May 2015 is shown in Fig. 3. Smaller radii (Figs. 3a–c) resulted in precipitation fields that were very noisy with greater small-scale variability than the PM mean, while larger radii (Fig. 3e) were very computationally intensive. When using a 200-km radius, it took about 75 min to run the code. Subjectively, 126 km (Fig. 3d) seemed to be the radii at which the spatial variability was similar to the PM mean (Fig. 3f), and further increases in the radii had very little impact on the precipitation fields. However, there was still considerable computation expense at this radius; for a single case, it took almost 30 min to compute the localized PM mean, compared with the PM mean, which took about 30 s. Ideally, for operational applications, the code would run faster, but by taking advantage of parallelized processing, multiple forecast hours could be processed simultaneously so that all localized PM-mean fields would be present 30 min after the ensemble completes integration. Comparing the time required to generate the localized PM mean for a radius of 63 km with that for a radius of 126 km, it is found that doubling the radius results in an increase in computational expense by about a factor of 4. Note that with a large enough radius so that the area around each grid point encompassed the entire domain, the localized PM mean would converge to the PM mean.

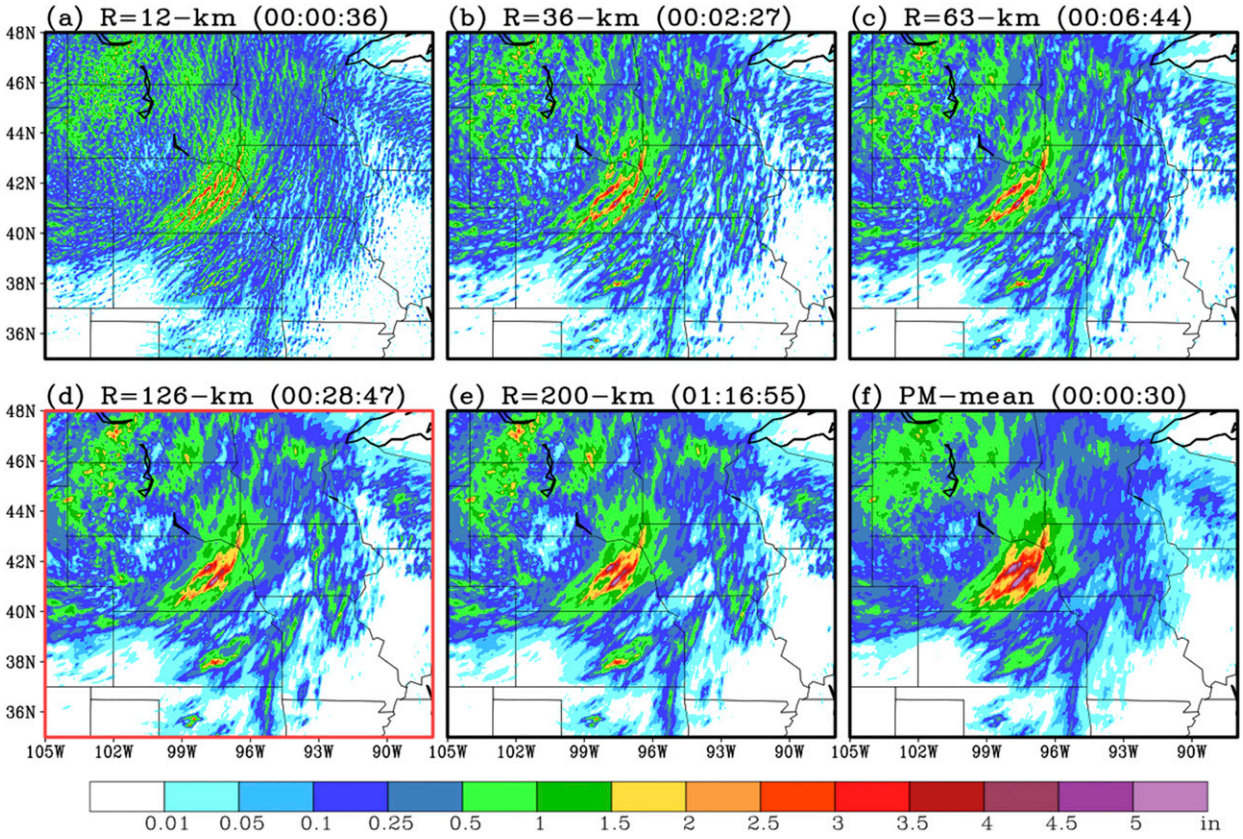

FIG. 3. Forecast 24-h accumulated precipitation valid 1200–1200 UTC 6–7 May 2015 from the localized PM mean computed using (a) a radius of 12 km. Results similar to those in (a) are shown but for radii of (b) 36, (c) 63, (d) 126, and (e) 200 km. (f) As in (a), but for the PM mean. The values in parentheses at the top of each panel denote the computational time needed to generate each field over the analysis domain (hh:mm:ss). Red highlighting is used in (d) to show the localized PM mean computed using the radius used for the rest of the study.

3. Results

a. Example case: 6 May 2015

To illustrate the application of the methods for computing ensemble mean precipitation, 24-h accumulated precipitation forecasts valid 1200–1200 UTC 6–7 May 2015 over two different domains are shown in Fig. 4. During this time period the large-scale weather regime was characterized by a deep, broad 500-hPa trough that covered the western half of the United States. Embedded within this trough, a short-wave trough axis progressed from west to east through Nebraska, Kansas, and Oklahoma during the afternoon and evening of 6 May. An expansive moist and unstable warm sector was present along and to the east of the trough axis, creating an environment favorable for organized convection and heavy rainfall. Indeed, widespread severe weather occurred during the afternoon and evening of 6 May, and a mesoscale convective system (MCS) eventually evolved that moved from northern Kansas to southeastern Nebraska with training convective cells that were associated with observed rainfall amounts greater than 5.0 in. (Fig. 4j).

Over the eastern United States, a ridge axis was progressing through the Great Lakes and midtropospheric winds over the mid-Atlantic states were very weak. However, insolation and weak low- to midlevel lapse rates led to enough instability so that unorganized, widely scattered convection formed over the mountains of West Virginia, Virginia, and western Maryland. A few of these storms grew upscale into a small, loosely organized MCS that produced rainfall between 0.5 and 1.0 in. over a small area of northern Virginia during the afternoon and early evening of 6 May (Fig. 4t).

For the first domain, the NSSL-WRF ensemble members predicted areas of heavy rainfall in eastern Nebraska, although there were considerable differences in the exact location and magnitude of the heavy rainfall area among the members. For example, the NAM-initialized member (Fig. 4d) had a clear rainfall maximum centered over east-central Nebraska where amounts exceeded 5.0 in. In contrast, although the

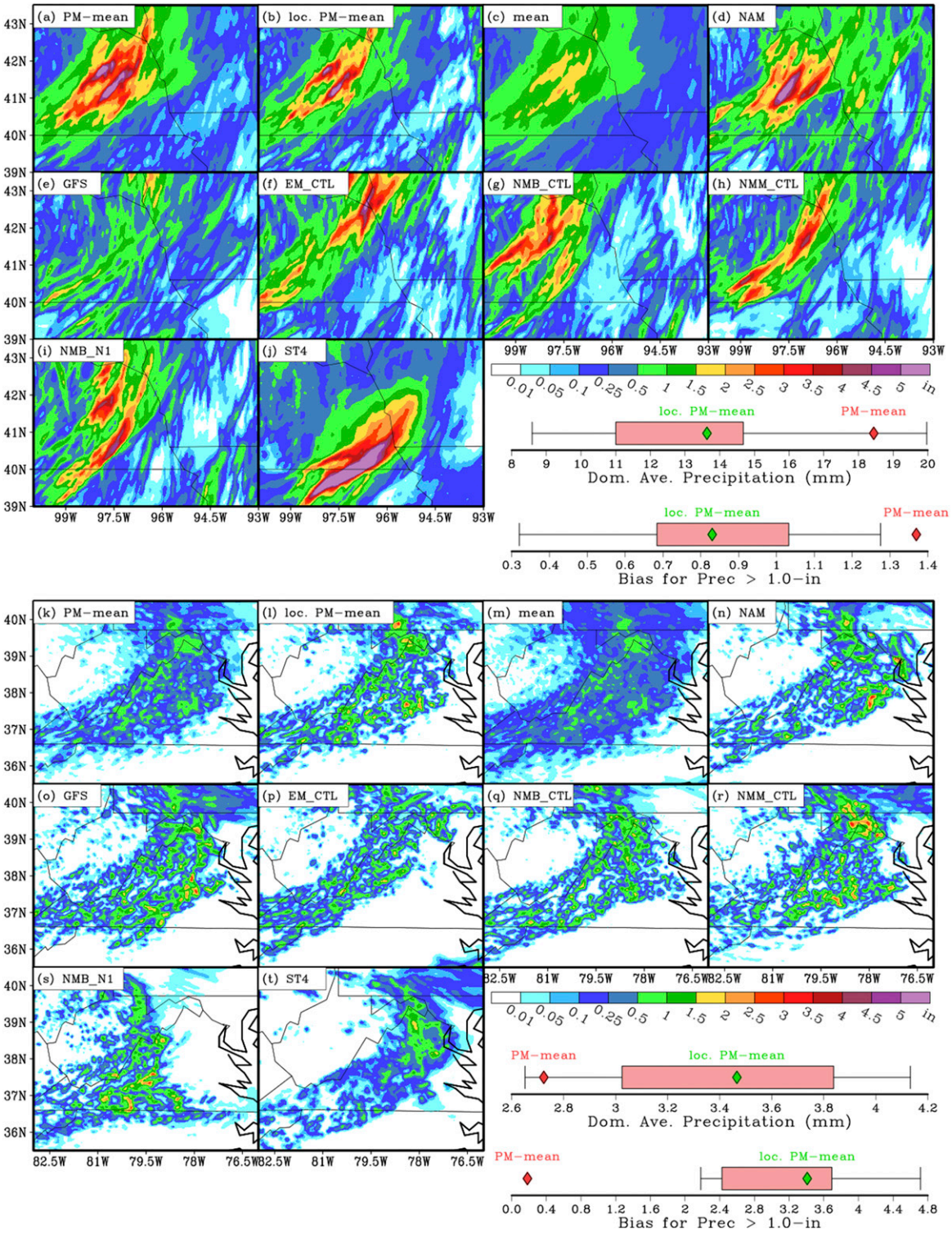

FIG. 4. Forecast 24-h accumulated precipitation valid 1200–1200 UTC 6–7 May 2015 over Dom 1 pictured in Fig. 1 from (a) the PM mean, (b) the localized PM mean, and (c) the traditional mean. Results similar to those in (a) are shown but for ensemble members initialized from (d) the NAM, (e) the GFS, (f) SREF member EM_CTL, (g) SREF member NMB_CTL, (h) SREF member NMM_CTL, and (i) SREF member NMB_N1. (j) Stage IV precipitation corresponding to (a)–(i). The top box plot shows the distribution from all NSSL-WRF ensemble members of precipitation amounts averaged over Dom 1. The red (green) diamond indicates where the corresponding amount from the PM mean (localized PM mean) falls within this distribution. The bottom box plot is the same as the top, but it shows the coverage bias for the 1.0-in. precipitation threshold. (k)–(t) As in (a)–(j), but over Dom 2 pictured in Fig. 1.

GFS-initialized member (Fig. 4e) did have relatively high rainfall amounts over east-central Nebraska, the maximum amounts of just over 2 in. were much lighter than the other members. Compared to the observed precipitation (Fig. 4j), the heaviest forecast amounts were slightly underpredicted by the ensemble members with a noticeable northward displacement, which is clearly reflected in the PM mean and localized PM mean (Figs. 4a,b). However, in the traditional mean (Fig. 4c), the maximum rainfall amounts are much lower and do not reflect the higher amounts present in the individual members. Although the localized PM mean and PM mean appear to adequately reflect the average location of the rainfall maxima from the ensemble members, there are clear differences in the rainfall magnitudes, with the PM mean producing considerably higher amounts. To show how well the two methods for computing ensemble means represent precipitation amounts from the ensemble members over the area pictured, the distribution of domain-averaged precipitation over this area from each ensemble member is shown by the top box plot associated with Figs. 4a–j. The localized PM-mean domain-averaged rainfall falls within the center of the interquartile range (IQR; green diamond), while the PM mean falls within the upper tail (red diamond). The bottom box plot associated with Figs. 4a–j shows the distribution of biases for amounts greater than 1.0 in. Here, bias is defined as the ratio of forecast to observed grid points with precipitation greater than 1.0 in. Thus, a bias of 1.0 is a perfect or unbiased forecast. Similar to the analysis of domain-averaged precipitation amounts, the localized PM-mean bias falls within the IQR, while the PM-mean bias falls in the upper tail. In fact, the PM mean had a bias higher than every ensemble member, which implies that some of the rainfall amounts greater than 1.0 in. must have been reassigned from other areas of the domain.

For the second domain, NSSL-WRF ensemble members generally predicted areas of light-to-moderate rainfall with small areas of amounts greater than 1.0 in. across western and northern Virginia, eastern West Virginia, and western Maryland. Compared with the observed precipitation (Fig. 4r), most members accurately predicted the general area of rainfall, but the amounts—particularly the high amounts—were generally overpredicted. While the highest observed amounts were just over 1.5 in. in northern Virginia, many of the ensemble members had small areas with rainfall greater than 2.5 in. in various parts of this subdomain. As in the first domain, the traditional mean (Fig. 4m) does not contain any of the higher precipitation amounts present in the individual ensemble members. Additionally, the traditional mean has much broader coverage of lighter rainfall amounts compared with any of the individual members. The PM mean (Fig. 4k) and localized PM mean (Fig. 4l) appear to adequately reflect the location of forecast rainfall from the ensemble members, but like the first domain, there are very noticeable differences in the rainfall amounts. The PM mean appears to have greater areal coverage of light rainfall amounts than the localized PM mean, while the localized PM mean has embedded maxima with much higher amounts than the PM mean. The box plot depicting the distribution of domain-averaged rainfall for this area shows that the localized PM mean falls within the IQR of the members, while the PM mean falls at the lower tail. For biases of rainfall greater than 1.0 in., the localized PM mean falls within the middle of the IQR of the members, but the PM mean falls well below the ensemble member with the lowest bias. This implies that the PM mean is reassigning heavier rainfall amounts that occur within this area to other areas of the domain (e.g., midwestern United States). Furthermore, the spatial variability in the localized PM mean looks more similar to the individual members compared with the PM mean.

In summary, for the first domain where a widespread area of heavy precipitation occurred, the PM mean inflated rainfall amounts relative to the individual ensemble members. For the second domain, where isolated heavy rainfall occurred, the PM mean reduced the higher rainfall amounts relative to those forecast by the individual members. In both cases, the localized PM mean was a better representation of precipitation forecasts from the individual ensemble members. Analyses in subsequent sections will examine how this type of behavior impacts aggregate forecast skill metrics and other statistics measuring the spatial character of forecast rainfall.

b. Aggregate precipitation amounts

To examine the observed and model climatology of precipitation, amounts are summed at each grid point over the analysis domain over all 317 cases (Fig. 5). Generally, the heaviest observed rainfall amounts over this period occurred over the southeast United States with local areas of heavier rainfall over eastern Kentucky and Tennessee (Fig. 5k). The spatial distribution of observed precipitation generally matches the well-documented observed climatology (e.g., Durre et al. 2013). The amounts and spatial distribution of precipitation in the localized PM mean and PM mean are quite similar (Figs. 5a,b). Both have a relative maximum over Kentucky and Tennessee, as well as maxima over coastal areas along the Gulf of Mexico and mid-Atlantic states like Virginia and South Carolina. Compared with observations, the maxima along the coastal areas are too

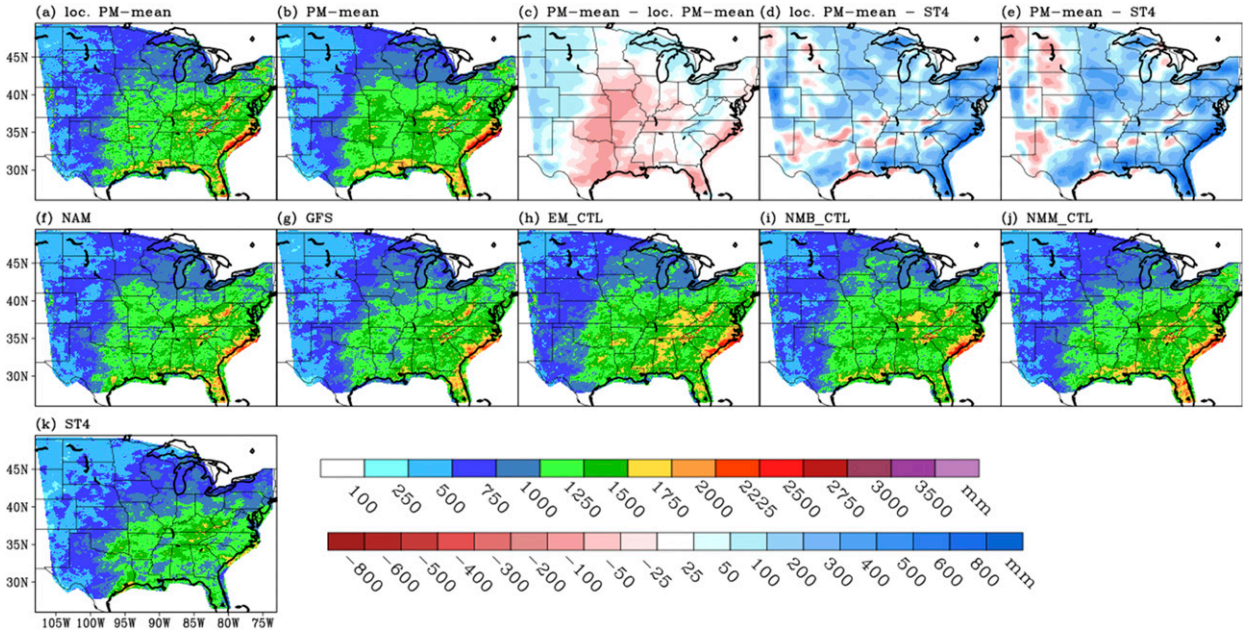

FIG. 5. Total 24-h precipitation summed over all 317 cases for the (a) localized PM mean and (b) PM mean. Results similar to those in (a) and (b) are shown but for ensemble members initialized from (f) the NAM, (g) the GFS, (h) SREF member EM_CTL, (i) SREF member NMB_CTL, and (j) SREF member NMM_CTL. (k) As in (a),(b), but for the Stage IV observations. (c) Difference between PM mean and localized PM mean [i.e., (b) − (a)]. (d) Difference between localized PM mean and Stage IV [(i.e., (a) − (k)]. (e) Difference between PM mean and Stage IV [i.e., (b) − (k)]. In (c)–(e), a Gaussian smoother ($\sigma = 40$ km) has been applied to the difference fields to reduce the noise and better highlight the areas with larger differences.

large, but the overprediction is more dramatic in the PM mean (Fig. 5e). Furthermore, the PM mean has a clear maximum across eastern Oklahoma, western Arkansas, and Missouri. This maximum is notable because it does not appear in the localized PM mean, and this area has the largest differences in precipitation amounts between the localized PM-mean and PM-mean fields (Fig. 5c). Examining the differences with the Stage IV observations (Figs. 5d,e), the localized PM mean is too wet in most areas, except for parts of the high plains and parts of the southern United States stretching from sections of Texas and Oklahoma to the lower Mississippi valley. The PM mean is also too wet, but compared with the localized PM mean, a larger area of the high plains is too dry in the PM mean. Total precipitation from five representative ensemble members has similar patterns, but there is considerable variability in the magnitude regarding some of the maxima (Figs. 5f–j).

To gauge how representative the total precipitation amounts from the PM mean and localized PM mean are of the total amounts from the ensemble members, the distance in units of standard deviations from the total mean precipitation of all ensemble members over all 317 cases is computed for each grid point and displayed in Fig. 6. Essentially, this analysis shows where, within the distribution of ensemble members and averaged over all cases, the PM-mean and localized PM-mean precipitation amounts fall. Areas enclosed by contours are those in which the total precipitation from the PM mean or localized PM mean fell outside of the distribution of amounts from all members. For the PM mean (Fig. 6a), there is a clear pattern; over an area including Missouri, eastern Oklahoma, and western Arkansas, the PM mean falls in the upper part of the distribution from all members with many points in which the PM-mean amounts are larger than any of the individual members. There are also many coastal areas, as well as an area of Kentucky and Tennessee, where the PM mean tends to fall in the upper part of the ensemble member distribution. On the other hand, over a large area of the high plains and around the Great Lakes, the PM mean falls in the lower part of the distribution from all members with many points where the PM-mean amounts are lower than any of the individual members. Interpretation of this pattern reveals that the computation of the PM mean results in the reassignment of heavier precipitation amounts that occurs preferentially over the areas with red/pink shading. These heavier amounts are being reassigned from the blue-shaded areas. This likely occurs because the precipitation over the pink/red-shaded areas tends to come from long-lasting, widespread moderate-to-heavy precipitation events in

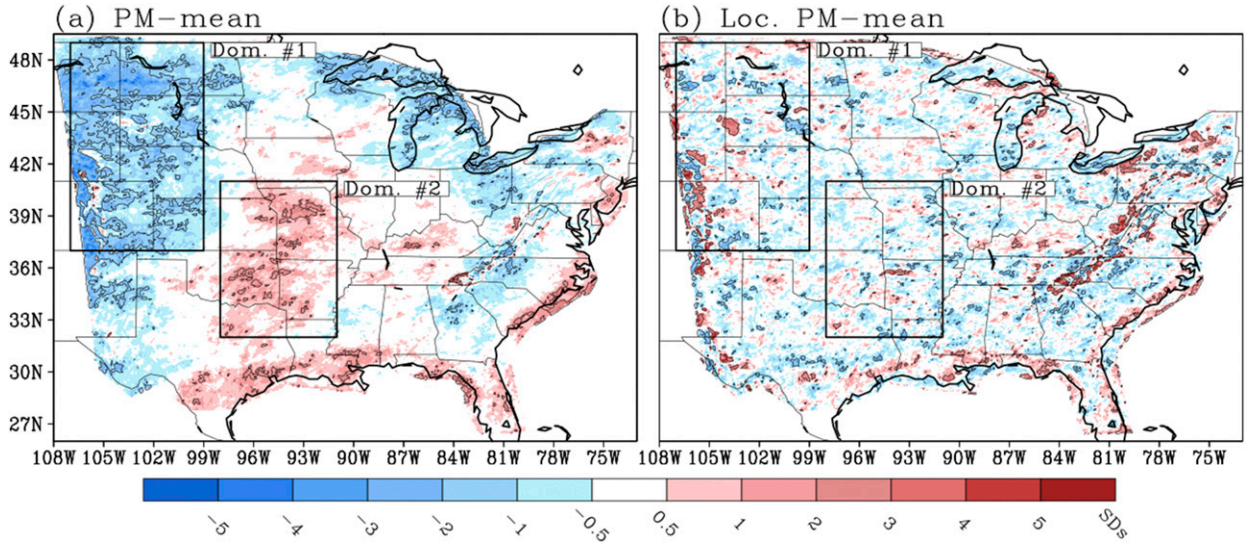

FIG. 6. Distance in units of standard deviation from the total mean precipitation of all ensemble members for (a) the PM mean and (b) the localized PM mean. The contours denote areas in which precipitation amounts from the PM mean or localized PM mean fell outside the distribution of amounts from all members. The boxes marked Dom 1 and Dom 2 show areas over which the bias is calculated in Fig. 7.

which there is considerable overlap in the ensemble member forecasts of precipitation. With this level of overlap, the mean retains larger values, and thus the PM mean assigns the larger values from the individual members over these areas of overlap. Over areas with more sporadic and less widespread heavy precipitation, the ensemble members tend to have less overlap. Thus, the mean is dampened more than in the cases with considerable overlap, and, if there is an area of the domain with heavy rainfall and more overlap, the heavier amounts from the ''sporadic'' area are likely to be reassigned to the ''widespread'' area. The differences from the mean for the localized PM mean (Fig. 6b) are more random compared with the PM mean, but there does appear to be some higher-amplitude small-scale variability tied to terrain features like the Appalachian and Rocky Mountains and the Black Hills of South Dakota.

c. *Forecast performance: Biases*

First, to gauge how well the areal coverage of precipitation at various thresholds matches the observations, bias is calculated for all ensemble members, PM mean, localized PM mean, and the traditional mean (Fig. 7). The bias is computed over the full analysis domain as well as domains 1 (Dom 1) and 2 (Dom 2), marked in Fig. 6. Again, bias is defined as the ratio of forecast to observed grid points that predict precipitation greater than the specified threshold. For the analysis domain, for all datasets except the traditional mean, bias increases with increasing precipitation threshold (Fig. 7a). The problem with using the traditional mean is immediately clear: light precipitation amounts are overpredicted, whereas heavy precipitation amounts are underpredicted, which is an artifact of averaging across a spatial field with high variability and sharp gradients, and the traditional mean is not at all representative of the individual ensemble members. As expected, the PM-mean bias falls right along the middle of the envelope of ensemble members. Because of how the PM mean is computed, its bias should be approximately equal to the average bias of the ensemble members. The localized PM mean has a slightly higher (lower) bias than the PM mean at lower (higher) precipitation thresholds.

Over domain 1 (Fig. 7b), the same pattern in ensemble member biases is present, with bias increasing with increasing threshold. The localized PM-mean biases generally fall within the middle of the ensemble member biases, but the PM-mean biases are noticeably lower than all ensemble member biases up to the 0.50-in. threshold and then fall within the lower bounds of the ensemble member distribution for the higher thresholds. These results are consistent with Fig. 6a, which shows that over domain 1, the PM-mean precipitation amounts fall just within or beyond the lower bounds of the ensemble member distributions.

Over domain 2 (Fig. 7c), ensemble member biases are more flat with increasing threshold, with the exception of one or two members. The localized PM-mean biases generally fall within the middle of the ensemble member

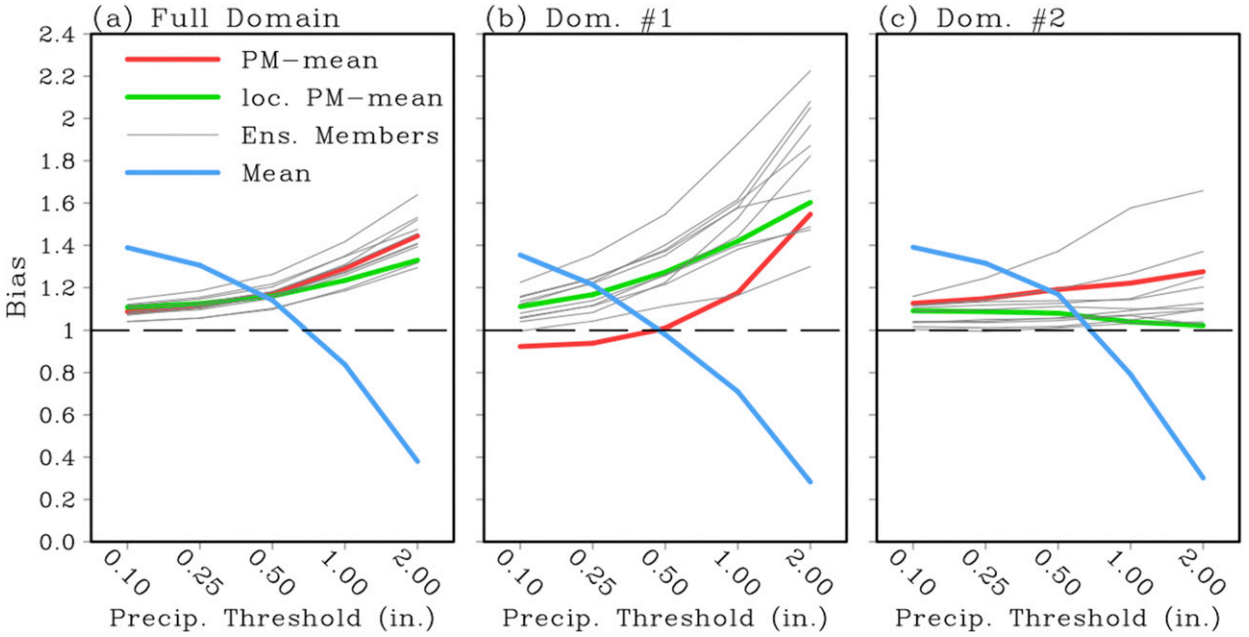

FIG. 7. Precipitation bias as a function of threshold for the PM mean, localized PM mean, individual ensemble members, and traditional mean for (a) the entire masked areas, (b) the subdomain marked Dom 1 in Fig. 6, and (c) the subdomain marked Dom 2 in Fig. 6.

biases for the lower rainfall thresholds up to about 0.50 in., and then fall within the lower bounds of the ensemble member distributions. On the other hand, for all thresholds, the PM-mean biases fall within the upper bound of the ensemble member distributions. Once again, these results are consistent with Fig. 6a, which shows that over domain 2, the PM-mean precipitation amounts tend to fall just within or above the upper bounds of the ensemble member distributions.

In general, it can be concluded that when computed over the entire analysis domain, the localized PM mean and PM mean have similar biases; however, when considering more localized regions, the localized PM mean better represents the ensemble member bias distributions.

d. *Forecast performance: Neighborhood-based equitable threat scores*

To gauge the overall skill of the precipitation forecasts, the neighborhood-based equitable threat score (ETS) is used (Clark et al. 2010). As many researchers have noted, when verifying high-resolution models, it is extremely important to consider spatial scales larger than the native model grid spacing (e.g., Casati et al. 2004; Roberts and Lean 2008; Gilleland et al. 2009), because inherent predictability limitations severely restrict the time scales at which high-resolution forecasts are accurate at the grid scale (e.g., Lorenz 1969). Neighborhood approaches are one of several methods that can be used to examine spatial scales larger than the native model grid.

The formulation of a neighborhood-based ETS, which is a method used to relax the criteria for successful forecasts and is based on contingency table elements (e.g., Wilks 2011), is described in Clark et al. (2010). Basically, if an event is observed (forecast) at a grid point, it is considered a *hit* if the event is forecast (observed) at the grid point or at any grid point within a circular radius r. A *miss* (*false alarm*) is assigned when an event is observed (forecast) at a grid point and no grid points within r forecast (observe) the event. Finally, *correct negatives* are assigned when an event is neither forecast nor observed at a single point. Then, neighborhood-based ETS can be computed as

$$\text{ETS} = \frac{\text{hits} - \text{chance}}{\text{hits} + \text{misses} + \text{false alarms} - \text{chance}}, \quad (2)$$

where

$$\text{chance} = \frac{(\text{hits} + \text{misses})(\text{hits} + \text{false alarms})}{\text{hits} + \text{misses} + \text{correct negative} + \text{false alarms}}. \quad (3)$$

For $r = 0$, the neighborhood-based ETS is the same as the traditional version of ETS that requires gridpoint-to-gridpoint matches. To assess whether differences in the PM-mean and localized PM-mean ETSs were statistically significant, the resampling procedure described by

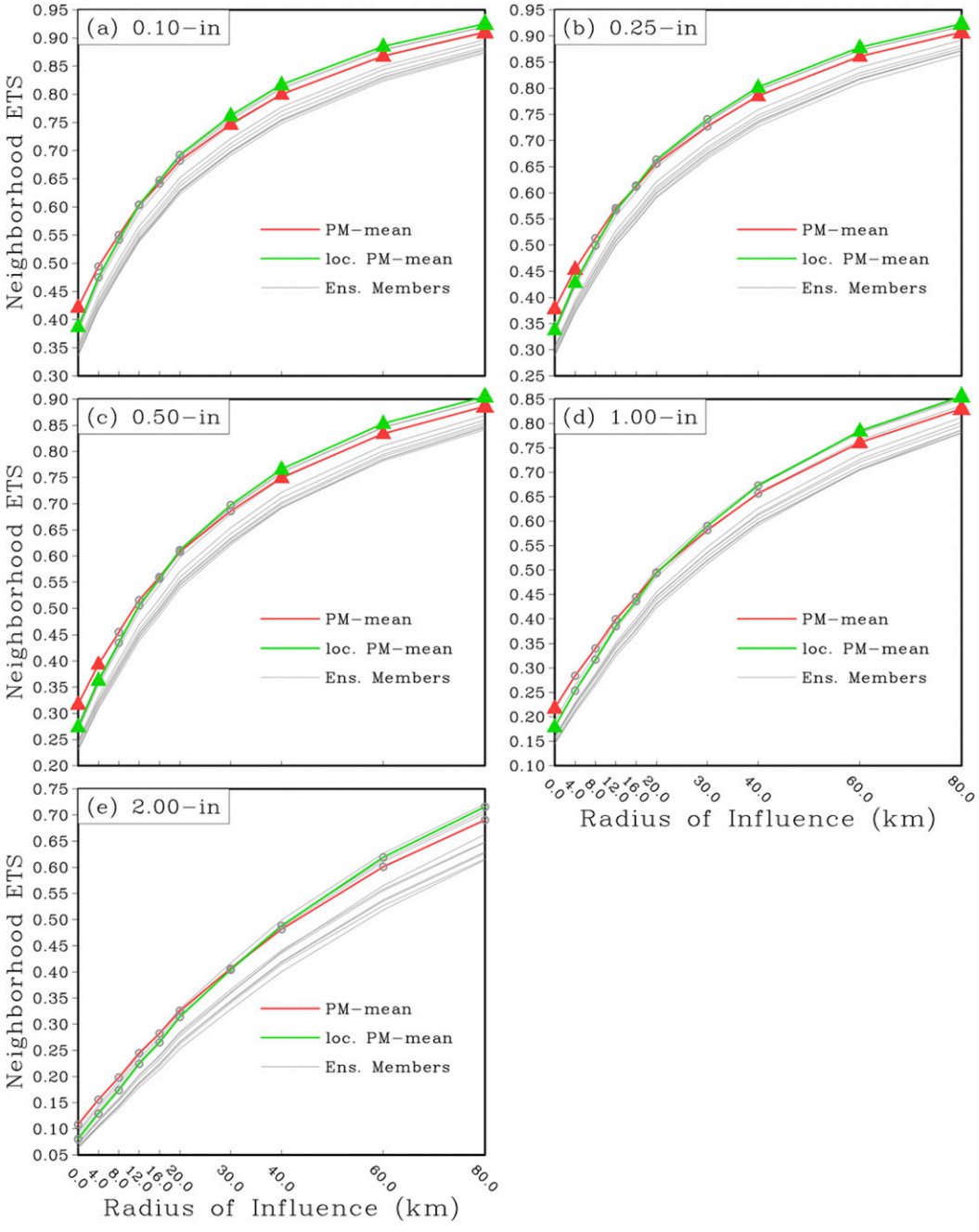

FIG. 8. Neighborhood-based ETSs as a function of radius of influence for the PM mean, localized PM mean, and NSSL-WRF ensemble members at rainfall thresholds of (a) 0.10, (b) 0.25, (c) 0.50, (d) 1.00, and (e) 2.00 in. Points marked with triangles (open circles) indicate that differences between the PM mean and localized PM mean were statistically significant (not significant).

Hamill (1999) was utilized, with resampling repeated 1000 times and $\alpha = 0.05$.

In Fig. 8, neighborhood-based ETSs computed over the full analysis domain for each dataset at different thresholds are displayed as a function of r, where values of r include 0, 4, 8, 12, 16, 20, 30, 40, 60, and 80 km. As expected, ETS increases with increasing r in each dataset. Also, depending on the threshold, the PM mean and localized PM mean have higher ETS values than all or most of the ensemble members. Interestingly, at the smallest scales and all thresholds examined, the PM mean performed best by quite a large margin relative to

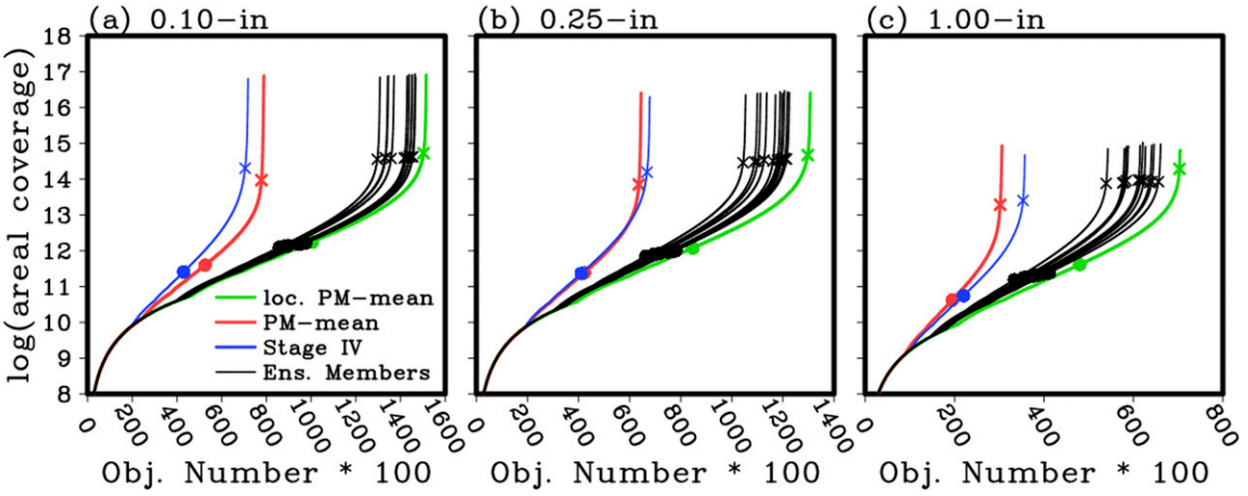

FIG. 9. Running sum of total object areal coverage for object sizes sorted from smallest to largest for objects defined using the thresholds of (a) 0.10, (b) 0.25, and (c) 1.00 in. The rightward extent of each line represents the total number of objects, while the height of each line is the total areal coverage. The filled circles (crosses) mark the point at which object sizes of 5 (1000) grid points or greater begin contributing to the sums.

the other datasets. However, as r increased, the differences became smaller, and in general for $r \geq 20$ km, the localized PM mean performed best. For every threshold except 2.0 in., the PM mean had significantly higher ETS values for the smallest scales, while the localized PM mean had significantly higher ETS values for the largest scales.

This behavior very likely reflects how the PM mean redistributes precipitation amounts across a large domain. As the example case in Fig. 4 illustrated, widespread precipitation areas with large overlap between ensemble members will be preferentially assigned the largest precipitation amounts from *anywhere* over the domain. Thus, if there are other areas of the domain where heavy precipitation is being forecast, but there is less overlap/agreement between ensemble members, these precipitation values will be preferentially reassigned to the area where there is more overlap. This causes an artificial enlargement of heavy precipitation where there is low forecast uncertainty and artificial shrinking of heavy precipitation where there is high forecast uncertainty. This regional artificial inflation in the size of precipitation areas is likely aiding the ETS, as forecast bias can artificially inflate ETS (e.g., Hamill 1999). In areas with greater forecast uncertainty where precipitation areas may shrink, the ETS may not be impacted much because it is already very low. However, as neighborhood size increases, the areas with less forecast agreement will begin to have more impact on the ETS. If these areas can maintain large values of precipitation, as the localized PM mean is meant to do, then it is possible for many of the grid points to become *hits* that were not hits at smaller neighborhoods. However, if the large precipitation values are reassigned to other regions of the domain, as occurs with the PM mean, then there may not be points that exceed the specified thresholds that can become hits.

e. Object size distributions

To show the impact of the PM mean and localized PM mean on characteristics of the spatial precipitation distributions, an analysis of object size and frequency was conducted. Objects are defined as contiguous regions of grid points exceeding a specified threshold. An object identification algorithm was applied to each precipitation dataset and every object over the analysis domain for the thresholds of 0.10, 0.25, and 1.00 in. was identified with its size recorded, where the size is simply defined by the number of grid points. Then, in each dataset, object sizes are sorted from smallest to largest, and running sums of areal coverage contributed by object sizes equal to or less than the ith sorted object are displayed in Fig. 9. For example, object number $i = 1200$ corresponds to the object with the 1200th smallest size, and its corresponding y-axis value (logarithmic scale) is the sum of object sizes less than or equal to the size of the 1200th smallest object. Thus, the rightward extent of each line in Fig. 9 represents the total number of objects, while the height of each line is the total areal coverage.

Several things can be said about the precipitation distributions from Fig. 9. First, differences in the running total areal coverage (i.e., highest value along the y axis for each curve) among the ensemble members and means for each threshold appear fairly small, which is

consistent with the biases displayed in Fig. 7a (note that the differences appear smaller than they actually are because of the *y*-axis log scale). Second, there are large differences in the numbers of objects. Many of the ensemble members contain more than twice as many objects as the Stage IV observations. At each threshold, the dataset with the largest number of objects is the localized PM mean, which is followed by all the ensemble members. The localized PM mean has a 3.4%, 6.7%, and 12.9% larger number of objects than the highest ensemble member for the 0.10-, 0.25-, and 1.00-in. precipitation thresholds, respectively. The most likely explanation for the increase in object numbers in the localized PM mean is that, when the traditional ensemble mean is computed before values are reassigned from the ensemble members, additional objects can be ''created'' in areas with little overlap between members. In these situations, the PM mean eliminates these objects because they are low amplitude in the mean, and the larger values from the individual members get reassigned to other areas of the domain with more overlap. However, in the localized PM mean, the larger values from individual members remain within the 126-km radius of influence.

The PM mean object numbers are, on average, about half those of the ensemble members. This net reduction in object numbers is consistent with the PM mean shrinking or eliminating objects in areas where there is not a lot of overlap among ensemble members, and enlarging objects where there is a lot of overlap. Ironically, although the PM-mean object numbers do not depict those of the ensemble members very well, they match very closely with the Stage IV object numbers. This close match is serendipitous, as one can imagine that if the ensemble member object numbers matched Stage IV observations well (i.e., lines in Fig. 9 were shifted to the left), the PM-mean object numbers would be much too small. Third, although the total areal coverage of objects is quite similar, the sizes of the objects that make up the total areal object coverage are quite different. As previously discussed, the PM mean shrinks or eliminates small objects and enlarges larger objects. This effect can be quantified from the data used to construct Fig. 9. For example, the filled circles in each Fig. 9 panel mark the point at which object sizes of five grid points or greater begin contributing to the sums. For the 1.0-in. threshold, objects with five or fewer grid points comprise 3.5% of the total areal coverage greater than 1.0 in. for the localized PM mean, but only 1.35% for the PM mean. Thus, for the PM mean, there are fewer smaller objects and they contribute less to the total areal coverage. The crosses in each Fig. 9 panel mark the points at which object sizes of 1000 grid points or greater begin contributing to the sums. For the 1.0-in. threshold, object sizes greater than 1000 grid points comprise 56% of the total areal coverage greater than 1.0 in. for the localized PM mean, but 82% for the PM mean. Furthermore, there are 418 objects greater than 1000 grid points in the localized PM mean, and 361 for the PM mean. Thus, for the PM mean, there are a greater number of larger objects (i.e., object sizes $\gg$ 1000 points), and they contribute more to the total areal coverage.

4. Summary and conclusions

This paper examined characteristics of ensemble mean 24-h accumulated precipitation forecasts generated from the 4-km, CONUS domain, convection-allowing NSSL-WRF ensemble using various methods. An approximately 1-yr period of forecasts was examined from October 2014 to October 2015, which included a sample of 317 cases. Problems with the traditional ensemble mean (i.e., simple average of all ensemble members) are well known, and include the overprediction of light precipitation and the amplitude reduction of precipitation maxima. To address the issues with the traditional ensemble mean, a method known as probability matching (PM mean) has commonly been used to replace the ensemble mean amounts with amounts sampled from the distribution of ensemble member forecasts, which results in a field where the higher amplitudes are retained, and the bias is approximately equal to the average bias of the ensemble members. Previous studies find that the PM mean is more skillful than any of the individual ensemble members. However, this work provides evidence that the superior performance of the PM mean is at least partially an artifact of the spatial redistribution of precipitation amounts that occur when the PM mean is computed over a relatively large domain. Specifically, the PM mean increases the sizes of the large precipitation areas and shrinks or eliminates smaller ones, which was demonstrated for an example case, as well as aggregate statistics examining characteristics of object sizes. An alternative approach for the PM mean is developed, known as the localized PM mean, which restricts the grid points used to those within a specified radius of influence. For an example case, the localized PM mean better represented precipitation amounts from the individual ensemble members. In terms of aggregate forecast skill, once larger neighborhoods were considered in the computation of a neighborhood-based equitable threat score, the localized PM mean performed best. It is likely that higher scores in the PM mean for the smaller neighborhoods occur because regional inflation of the size of large precipitation areas inflates the ETS, as forecast bias is known to artificially inflate ETS. It should be stressed that problems

with the PM mean "reassigning" precipitation values to different areas of the domain are most relevant when large domains are used. Users of regional ensembles with domains only encompassing a few states may find that the reassignment is not as much a concern as for CONUS-wide domains.

The conclusion that the high skill in the PM mean is somewhat of an artifact was also argued by Surcel et al. (2014). However, they took a different approach, examining how the spectral structure of precipitation forecasts impacted the skill of the PM mean and ensemble members, and finding that the gain in skill from the PM mean results from reduced variability in the PM mean at small scales relative to the individual members. I generally agree with this argument. However, this paper provides evidence that the gain in skill, at least for contingency table–based metrics like ETS, is also partially caused by regional biases (i.e., changes in the distribution of precipitation amounts across a large domain). Nonetheless, the fact that the localized PM mean has higher ETS values than all the ensemble members and the PM mean for the larger neighborhood sizes does support the notion that the localized PM mean provides value. In general, this work shows that users should exercise caution when interpreting the PM mean and that use of a localized PM mean is preferred because it better represents the ensemble members. Future work should explore ways to further improve characteristics of the localized PM mean, such as its overprediction of object numbers and its tendency to have a bias below that of the individual ensemble members for very high rainfall thresholds (i.e., over 2.0 in.).

Acknowledgments. Scott Dembek of CIMMS/NSSL maintains the real-time NSSL-WRF modeling system and data archival. Suggestions and feedback from three anonymous reviewers and Craig Schwartz of NCAR were much appreciated and helped improve the manuscript. The NSSL-WRF ensemble is run on the NOAA supercomputer, Jet, in Boulder, Colorado. This work was completed as part of regular work duties at the federally funded NOAA National Severe Storms Laboratory.

REFERENCES

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

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.

Berenguer, M., M. Surcel, I. Zawadzki, M. Xue, and F. Kong, 2012: The diurnal cycle of precipitation from continental radar mosaics and numerical weather prediction models. Part II: Intercomparison among numerical models and with nowcasting. *Mon. Wea. Rev.*, **140**, 2689–2705, doi:10.1175/MWR-D-11-00181.1.

Casati, B., G. Ross, and D. B. Stephenson, 2004: A new intensity scale approach for the verification of spatial precipitation forecasts. *Meteor. Appl.*, **11**, 141–154, doi:10.1017/S1350482704001239.

Chen, F., and J. Dudhia, 2001: Coupling an advanced land surface–hydrology model with the Penn State–NCAR MM5 modeling system. Part I: Model description and implementation. *Mon. Wea. Rev.*, **129**, 569–585, doi:10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2.

Clark, A. J., W. A. Gallus Jr., and T.-C. Chen, 2008: Contributions of mixed physics versus perturbed initial/lateral boundary conditions to ensemble-based precipitation forecast skill. *Mon. Wea. Rev.*, **136**, 2140–2156, doi:10.1175/2007MWR2029.1.

——, ——, 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, doi:10.1175/2009WAF2222222.1.

——, ——, and M. L. Weisman, 2010: Neighborhood-based verification of precipitation forecast from convection-allowing WRF Model simulations and the operational NAM. *Wea. Forecasting*, **25**, 1495–1509, doi:10.1175/2010WAF2222404.1.

——, and Coauthors, 2012: An overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74, doi:10.1175/BAMS-D-11-00040.1.

Du, J., and Coauthors, 2014: NCEP regional ensemble update: Current systems and planned storm-scale ensembles. *26th Conf. on Weather Analysis and Forecasting/22nd Conf. on Numerical Weather Prediction*, Atlanta, GA, Amer. Meteor. Soc., J1.4. [Available online at https://ams.confex.com/ams/94Annual/webprogram/Paper239030.html.]

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, doi:10.1175/1520-0469(1989)046<3077:NSOCOD>2.0.CO;2.

Durre, I., M. F. Squires, and R. S. Vose, 2013: NOAA's 1981–2010 U.S. climate normals: Monthly precipitation, snowfall, and snow depth. *J. Appl. Meteor. Climatol.*, **52**, 2377–2395, doi:10.1175/JAMC-D-13-051.1.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, doi:10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

Evans, C., D. F. Van Dyke, and T. Lericos, 2014: How do forecasters utilize output from a convection-permitting ensemble forecast system? Case study of a high-impact precipitation event. *Wea. Forecasting*, **29**, 466–486, doi:10.1175/WAF-D-13-00064.1.

Fang, X., and Y.-H. Kuo, 2013: Improving ensemble-based quantitative precipitation forecasts for topography-enhanced typhoon heavy rainfall over Taiwan with a modified probability-matching technique. *Mon. Wea. Rev.*, **141**, 3908–3932, doi:10.1175/MWR-D-13-00012.1.

Gallo, B. T., A. J. Clark, and S. R. Dembek, 2016: Forecasting tornadoes using convection-permitting ensembles. *Wea. Forecasting*, **31**, 273–295, doi:10.1175/WAF-D-15-0134.1.

Gilleland, E., D. Ahijevych, and B. G. Brown, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430, doi:10.1175/2009WAF2222269.1.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, doi:10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

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.

Janjić, Z. I., 2002: Nonsingular implementation of the Mellor–Yamada level 2.5 scheme in the NCEP Meso model. NCEP Office Note 437, 61 pp. [Available online at http://www.emc.ncep.noaa.gov/officenotes/newernotes/on437.pdf.]

Jirak, I. L., S. J. Weiss, and C. J. Melick, 2012: The SPC storm-scale ensemble of opportunity: Overview and results from the 2012 Hazardous Weather Testbed Spring Forecasting Experiment. *27th Conf. on Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., P9.137. [Available online at https://ams.confex.com/ams/26SLS/webprogram/Paper211729.html.]

——, C. J. Melick, and S. J. Weiss, 2014: Combining probabilistic ensemble information from the environment with simulated storm attributes to generate calibrated probabilities of severe weather hazards. *27th Conf. on Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., 2.5. [Available online at https://ams.confex.com/ams/27SLS/webprogram/Paper254649.html.]

Kain, J. S., S. R. Dembek, S. J. Weiss, J. L. Case, J. J. Levit, and R. A. Sobash, 2010: Extracting unique information from high-resolution forecast models: Monitoring selected fields and phenomena every time step. *Wea. Forecasting*, **25**, 1536–1542, doi:10.1175/2010WAF2222430.1.

Kong, F. M., and Coauthors, 2009: A real-time storm-scale ensemble forecast system: 2009 Spring Experiment. *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., 16A.3. [Available online at https://ams.confex.com/ams/23WAF19NWP/techprogram/paper_154118.htm.]

Leith, C. D., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418, doi:10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2.

Lin, Y., and K. Mitchell, 2005: The NCEP Stage II/IV hourly precipitation analyses: Development and applications. *19th Conf. on Hydrology,* San Diego, CA, Amer. Meteor. Soc., 1.2. [Available online at https://ams.confex.com/ams/Annual2005/techprogram/paper_83847.htm.]

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307, doi:10.3402/tellusa.v21i3.10086.

Mellor, G. L., and T. Yamada, 1982: Development of a turbulence closure model for geophysical fluid problems. *Rev. Geophys.*, **20**, 851–875, doi:10.1029/RG020i004p00851.

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 longwave. *J. Geophys. Res.*, **102**, 16 663–16 682, doi:10.1029/97JD00237.

Murphy, A. H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.*, **116**, 2417–2424, doi:10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2.

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, doi:10.1175/2007MWR2123.1.

Rogers, E., and Coauthors, 2009: The NCEP North American Mesoscale modeling system: Recent changes and future plans. *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., 2A.4. [Available online at https://ams.confex.com/ams/23WAF19NWP/techprogram/paper_154114.htm.]

Schwartz, C. S., G. S. Romine, K. R. Smith, and M. L. Weisman, 2014: Characterizing and optimizing precipitation forecasts from a convection-permitting ensemble initialized by a mesoscale ensemble Kalman filter. *Wea. Forecasting*, **29**, 1295–1318, doi:10.1175/WAF-D-13-00145.1.

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.

Surcel, M., I. Zawadzki, and M. K. Yau, 2014: On the filtering properties of ensemble averaging for storm-scale precipitation forecasts. *Mon. Wea. Rev.*, **142**, 1093–1105, doi:10.1175/MWR-D-13-00134.1.

Weisman, M. L., W. C. Skamarock, and J. B. Klemp, 1997: The resolution dependence of explicitly modeled convective systems. *Mon. Wea. Rev.*, **125**, 527–548, doi:10.1175/1520-0493(1997)125<0527:TRDOEM>2.0.CO;2.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences: An Introduction*. 3rd ed. Elsevier, 676 pp.

Xue, M., and Coauthors, 2011: Realtime convection-permitting ensemble and convection-resolving deterministic forecasts of CAPS for the Hazardous Weather Testbed 2010 Spring Experiment. *25th Conf. on Weather and Forecasting/20th Conf. on Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc., 9A.2.