Evaluation of Convection-Permitting Precipitation Forecast Products Using WRF, NMMB, and FV3 for the 2016–17 NOAA Hydrometeorology Testbed Flash Flood and Intense Rainfall Experiments

NATHAN SNOOK, FANYOU KONG, AND KEITH A. BREWSTER

Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma

MING XUE

Center for Analysis and Prediction of Storms, and School of Meteorology, University of Oklahoma, Norman, Oklahoma

KEVIN W. THOMAS AND TIMOTHY A. SUPINIE

Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma

SARAH PERFATER

NOAA/OAR/Office of Weather and Air Quality, Silver Spring, Maryland

BENJAMIN ALBRIGHT

Systems Research Group, Inc., Colorado Springs, Colorado

(Manuscript received 11 September 2018, in final form 25 March 2019)

ABSTRACT

During the summers of 2016 and 2017, the Center for Analysis and Prediction of Storms (CAPS) ran real-time storm-scale ensemble forecasts (SSEFs) in support of the Hydrometeorology Testbed (HMT) Flash Flood and Intense Rainfall (FFaIR) experiment. These forecasts, using WRF-ARW and Nonhydrostatic Mesoscale Model on the B-grid (NMMB) in 2016, and WRF-ARW and GFDL Finite Volume Cubed-Sphere Dynamical Core (FV3) in 2017, covered the contiguous United States at 3-km horizontal grid spacing, and supported the generation and evaluation of precipitation forecast products, including ensemble probabilistic products. Forecasts of 3-h precipitation accumulation are evaluated. Overall, the SSEF produces skillful 3-h accumulated precipitation forecasts, with ARW members generally outperforming NMMB members and the single FV3 member run in 2017 outperforming ARW members; these differences are significant at some forecast hours. Statistically significant differences exist in the performance, in terms of bias and ETS, among subensembles of members sharing common microphysics and PBL schemes. Year-to-year consistency is higher for PBL subensembles than for microphysical subensembles. Probability-matched (PM) ensemble mean forecasts outperform individual members, while the simple ensemble mean exhibits substantial bias. A newly developed localized probability-matched (LPM) ensemble mean product was produced in 2017; compared to the simple ensemble mean and the conventional PM mean, the LPM mean exhibits improved retention of small-scale structures, evident in both 2D forecast fields and variance spectra. Probabilistic forecasts of precipitation exceeding flash flood guidance (FFG) or thresholds associated with recurrence intervals (RI) ranging from 10 to 100 years show utility in predicting regions of flooding threat, but generally overpredict the occurrence of such events; however, they may still be useful in subjective flash flood risk assessment.

1. Introduction

Each summer, starting in 2012, the Hydrometeorology Testbed (HMT) at the Weather Prediction Center (WPC) of the U.S. National Weather Service has hosted

Corresponding author: Nathan Snook, nsnook@ou.edu

DOI: 10.1175/WAF-D-18-0155.1

the Flash Flood and Intense Rainfall (FFaIR) experiment. Spanning four weeks during June and July, FFaIR brings operational forecasters and members of the weather research and modeling communities together to study potential improvements to WPC's operational excessive rainfall outlook products and to investigate the skill and utility of new numerical weather prediction (NWP) tools for predicting heavy rainfall and flash flooding (WPC 2016). FFaIR provides a valuable platform for research-to-operations transition of cutting-edge NWP tools and ensemble forecast products focused on rainfall and flooding as well as providing valuable feedback to researchers—the operations to research component. Because FFaIR takes place in real time during the summer convective season considering the entire contiguous United States (CONUS), these new products are tested in a variety of synoptic environments and storm modes.

Experimental real-time severe weather forecasting experiments within the United States grew out of early collaborations between operational forecasters and research scientists during the 1980s and 1990s, culminating in Spring Program 2000—the first formal real-time forecast experiment—in which participants from the National Severe Storms Laboratory, the Cooperative Institute for Mesoscale Meteorological Studies, and the Storm Prediction Center worked together to evaluate experimental model products for severe weather forecasting (Kain et al. 2003). In subsequent years these efforts were continued in the Hazardous Weather Testbed (HWT) Spring Forecast Experiment (SFE), an annual real-time forecast experiment bringing operational and research meteorologists together to produce experimental products for severe weather events on time scales ranging from several hours to several days (Clark et al. 2012; Weiss et al. 2015) and evaluate new NWP products for use in the context of operational, real-time prediction of severe weather (e.g., Xue et al. 2007, 2009; Clark et al. 2009; Kain et al. 2010; Johnson et al. 2013; Surcel et al. 2014). The HMT experiments first occurred in 2010–11 as part of a "quantitative precipitation forecasting (QPF) component" of the HWT Spring Forecasting Experiments (Clark et al. 2012), and continued with the advent of FFaIR in summer of 2012, applying the collaborative research-to-operations framework of the HWT SFE to the prediction of high-impact short-term hydrological events.

Convection-allowing ensemble NWP model guidance [also called storm-scale ensemble forecast (SSEF)] is a major component of the HWT SFE and HMT FFaIR, both in generation of forecast products, and evaluation exercises where participants judge the skill and operational utility of experimental NWP products (Gallo et al. 2017). Recent HWT SFEs have featured the Community Leveraged Unified Ensemble (Clark et al. 2018)—a large (~60 member) ensemble forecasts at convection-allowing (or synonymously, convection permitting) 3-km horizontal grid spacing covering a common CONUS domain whose members were produced by a variety of operational and academic institutions. Such forecast ensembles have allowed for the evaluation of specific ensemble design choices (e.g., Clark et al. 2011; Duda et al. 2014; Loken et al. 2017, 2019; Clark et al. 2018), and the development and evaluation of novel forecast products (e.g., Clark et al. 2013), including probabilistic forecasts (e.g., Kain et al. 2013), and machine-learning-based products (e.g., McGovern et al. 2017; Gagne et al. 2017). The evaluation of such storm-scale NWP forecast products for use in a real-time, operational setting is vital for the development and improvement of these products and their eventual transition into operations.

The Center for Analysis and Prediction of Storms (CAPS) has produced storm-scale ensemble and deterministic forecasts at convection-allowing grid spacing of 1–4 km in support of the HWT SFE since 2005 (Kain et al. 2008) and, more recently, HMT FFaIR. CAPS produced the first ever real-time, convection-allowing, SSEFs for the 2007 HWT SFE—33-h, 10-member ensemble forecasts covering two-thirds of the CONUS in that year (Xue et al. 2007). In subsequent years, CAPS expanded the size of its forecast ensemble and forecast domain to cover the full CONUS (e.g., Xue et al. 2009; Xue et al. 2010). In addition to providing valuable datasets for investigating probabilistic forecast techniques for severe weather (e.g., Schwartz et al. 2010), the impact of radar data assimilation (e.g., Xue et al. 2009; Kain et al. 2010; Xue et al. 2013), and sensitivity of forecasts to ensemble configuration (e.g., Schwartz et al. 2010), the CAPS storm-scale ensemble forecasts were also found to substantially outperform North American Mesoscale Forecast System (NAM) 12-km forecasts (Schwartz et al. 2009; Kong et al. 2011).

Encouraged by these successes, CAPS produced a 15-member storm-scale ensemble forecast for the 2016 HMT FFaIR experiment based upon the CAPS 2016 HWT SFE ensemble design. This ensemble consisted of 13 members using the Advanced Research version of the Weather Research and Forecasting Model (WRF-ARW; Skamarock et al. 2005) and two Nonhydrostatic Mesoscale Model on the B-grid (NMMB; Janjić 2005) members, and was run at a 3-km horizontal grid spacing over the CONUS for all operational days of the 2016 FFaIR experiment. Several new heavy precipitation and flash flood forecast diagnostic products were developed and produced for FFaIR from the CAPS

SSEF, including probability of exceeding specified rainfall amounts over intervals of 1, 3, 6, 12, and 24 h, probability of exceeding flash flood guidance (FFG), and probability of extreme rainfall based on climatological recurrence intervals. New fields were also added to support ingredients-based subjective forecasting of intense rainfall, including precipitable water, 850-hPa ensemble mean wind, 850–300-hPa mean wind, and integrated water vapor transport. A similar ensemble was run for the 2017 HMT FFaIR, but with a single Geophysical Fluid Dynamics Laboratory (GFDL) Finite Volume Cubed-Sphere Dynamical Core (FV3; Lin 2004, Harris and Lin 2013) member in place of the NMMB members. Also, a new localized probability-matched mean (LPM) algorithm, which calculates a probability matched mean field (Ebert 2001) based on a mosaic of local patches in order to improve retention of local convective structures, was developed and implemented for evaluation during 2017 FFaIR.

This paper provides an overview of the CAPS SSEF and products during the 2016 and 2017 HMT FFaIR experiments, as well as verification of these ensemble forecasts and products. The 2016 and 2017 CAPS HMT FFaIR SSEF systems and their forecast products are described in detail in section 2. Objective forecast verifications are presented in section 3. Finally, section 4 contains discussions and a summary, as well as insights and plans for future development of ensemble forecasts and rainfall-specific forecast products for HMT FFaIR.

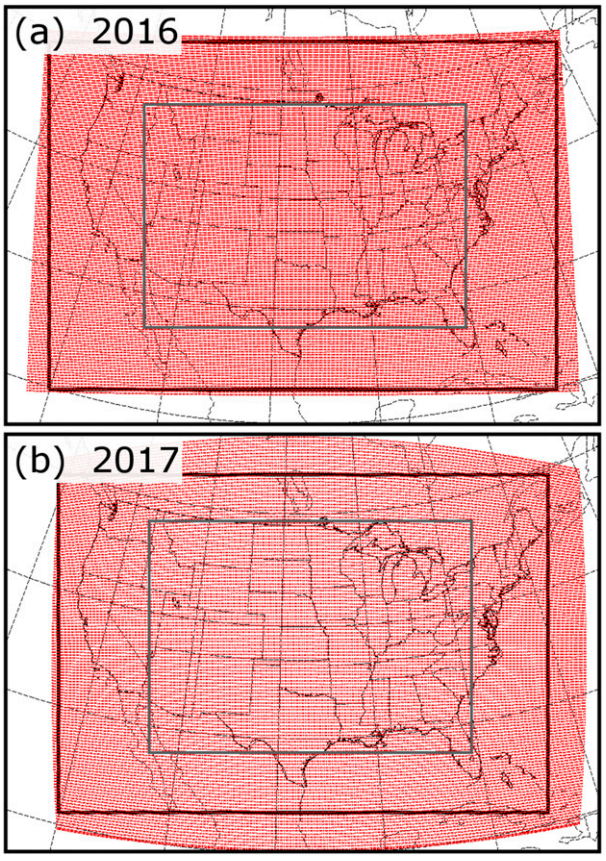

FIG. 1. The computational domains for the CAPS HMT FFaIR storm-scale ensembles in (a) 2016 and (b) 2017. The red dotted area in (a) is the 2016 NMMB domain, and the red dotted area in (b) is the 2017 FV3 domain. In both panels, the black boundary within the red dotted region denotes the domain for ARW member forecasts, while the gray (inner) boundary within the red dotted region denotes the verification domain. The outer domain (the full area shown in each panel, 1860 × 1280 grid points) is used for providing initial and boundary conditions for the NMMB, but not for FV3, which was nested within the global FV3 forecast.

2. Ensemble design and SSEF products

a. Storm-scale forecast ensemble

During the 2016 FFaIR, CAPS produced an ensemble of 15 convection-allowing model forecasts (13 using WRF-ARW and 2 using NMMB). ARW and NMMB members used similar CONUS domains with 3-km horizontal grid spacing (Fig. 1a). The ARW domain (the black outline within the red dotted region of Fig. 1a) comprised 1680 × 1152 horizontal grid points, while the NMMB domain (the red dotted area in Fig. 1a) comprised 1568 × 1120 horizontal grid points. WRF-ARW forecasts used 51 vertical levels; NMMB members used 50. Version 3.7.1 of WRF-ARW was used for ARW members. For the ARW and NMMB control members (arw_cn and nmmb_cn in Table 1), and three other ARW members without initial condition perturbations (arw_m10, arw_m11, arw_12; Table 1), forecast initial conditions came from analyses produced by ARPS 3DVAR (Xue et al. 2003; Gao et al. 2004) with a cloud analysis system (Hu et al. 2006; Brewster and Stratman 2016), assimilating full-volume observations of WSR-88D radar reflectivity and radial velocity, GOES visible and 11 μm IR, and surface and upper-air observations. The 0000 UTC operational NAM analysis was used as the analysis background, and 0000 UTC NAM forecasts were used to obtain lateral boundary conditions. Members arw_10, arw_11, and arw_12 (the physics-perturbation-only members) differ from ARW control member arw_cn only in microphysics scheme, and can therefore be used to investigate the relative performance of four microphysics schemes within WRF. For other ARW members, the initial conditions were the control initial condition plus additional perturbations derived from 3-h forecasts of the 2100 UTC cycle of the operational Short-Range Ensemble Forecast (SREF; Du et al. 2009), with boundary conditions from the corresponding SREF members. Member nmmb_m1

TABLE 1. Configurations for the 2016 CAPS HMT FFaIR storm-scale ensemble, including both ARW and NMMB members. NAMa and NAMf refer to 12-km NAM analysis and forecast, respectively. ARPSa refers to ARPS 3DVAR and cloud analysis. All members use RRTMG for parameterization of longwave and shortwave radiation and use no cumulus parameterization. The chosen initial condition perturbation for nmmb_m1 is from the first member of the Hazardous Weather Testbed NMMB ensemble.

Member	Initial conditions	Boundary conditions	Radar data	Microphysics	Land surface model	PBL
arw_cn	0000 UTC ARPSa	0000 UTC NAMf	Yes	Thompson	Noah	MYJ
arw_m3	arw_cn + arw-p1_pert	2100 UTC SREF arw-p1	Yes	P3	Noah	YSU
arw_m4	arw_cn + arw-n1_pert	2100 UTC SREF arw-n1	Yes	MY	Noah	MYNN
arw_m5	arw_cn + arw-p2_pert	2100 UTC SREF arw-p2	Yes	Morrison	Noah	MYJ
arw_m6	arw_cn + arw-n2_pert	2100 UTC SREF arw-n2	Yes	P3	Noah	YSU
arw_m7	arw_cn + nmmb-p1_pert	2100 UTC SREF nmmb-p1	Yes	MY	Noah	MYNN
arw_m8	arw_cn + nmmb-n1_pert	2100 UTC SREF nmmb-n1	Yes	Morrison	Noah	YSU
arw_m9	arw_cn + nmmb-p2_pert	2100 UTC SREF nmmb-p2	Yes	P3	Noah	MYJ
arw_m10	arw_cn + nmmb-n2_pert	2100 UTC SREF nmmb-n2	Yes	Thompson	Noah	MYNN
arw_m11	0000 UTC ARPSa	0000 UTC NAMf	Yes	P3	Noah	MYJ
arw_m12	0000 UTC ARPSa	0000 UTC NAMf	Yes	Morrison	Noah	MYJ
arw_m13	0000 UTC ARPSa	0000 UTC NAMf	Yes	MY	Noah	MYJ
arw_m14	arw_cn + arw-n2_pert	2100 UTC SREF arw-n2	Yes	Thompson	Noah	MYJ
nmmb_cn	0000 UTC ARPSa	0000 UTC NAMf	Yes	Ferrier–Aligo	Noah	MYJ
nmmb_m1	0000 UTC NAMa + arw-p3_pert	2100 UTC SREF arw-p3	No	Ferrier–Aligo	Noah	MYJ

used the 0000 UTC NAM analysis plus SREF perturbation for initial conditions, and did not include assimilation of radar observations (Table 1).

The ARW members included diversity in microphysics and planetary boundary layer (PBL) schemes. Microphysical schemes used included the Milbrandt and Yau two-moment (MY2; Milbrandt and Yau 2005), Morrison two-moment (Morrison et al. 2009), Thompson (Thompson et al. 2008), and the Morrison and Milbrandt Predicted Particle Properties (P3; Morrison and Milbrandt 2015) schemes. PBL schemes used included the Mellor–Yamada–Nakanishi–Niino (MYNN; Nakanishi and Niino 2009), Mellor–Yamada–Janjić (MYJ; Janjić 1994), and Yonsei University (YSU; Hong et al. 2006) schemes. The two NMMB members used the Ferrier–Aligo (Aligo et al. 2014) microphysics. All members used the Noah land surface model (Tewari et al. 2004) and the Rapid Radiative Transfer Model (RRTMG; Iacono et al. 2008) for longwave and shortwave radiation. Specific details on the forecasts of 2016 are given in Table 1.

In 2017, the two NMMB members were not run, and a single FV3 forecast was added. The FV3 dynamic core was chosen in 2016 by the National Weather Service to serve as the dynamic core of the Next-Generation Global Forecasting System (NGGPS; Zhou et al. 2019), replacing the spectral model of the operational Global Forecasting System (GFS). It is a goal of NWS to eventually use the FV3 dynamic core for all regional operational forecasting. The FV3 dynamic core and its physics packages were originally developed by NASA and NOAA GFDL for global climate simulations (e.g., Harris and Lin 2014; Xiang et al. 2015); their suitability and performance for QPF, especially at convection-allowing/resolving resolutions, had been little examined. Running FV3 in real time for HMT FFaIR thus provided a unique opportunity to examine its QPF performance relative to similarly configured WRF ARW forecasts.

The FV3 forecast used a grid with an approximately 3-km horizontal spacing covering the entire CONUS that was two-way interactively nested within a stretched global grid with a mean grid spacing of about 13 km; on the face covering CONUS (outside of the nest), the grid spacing was approximately 9 km (Fig. 1b). A scale-aware version of the simplified Arakawa–Schubert (SAS) cumulus parameterization scheme (Arakawa and Schubert 1974; Han et al. 2017) was used on the global grid only. A microphysics scheme widely used for convection-allowing forecasting, the Thompson (Thompson et al. 2008) microphysics scheme from WRF ARW, was implemented by CAPS within FV3 and used for the 2017 HMT forecasts. Other physics options included the Noah land surface model, the MRF (Hong and Pan 1996) PBL scheme, and the RRTMG for longwave and shortwave radiation. The FV3 was initialized from the native T1534 GFS analysis at 0000 UTC each day and did not require lateral boundary condition.

The 2017 WRF ARW ensemble for FFaIR included only 10 members, dropping the 3 physics-perturbation-only members of 2016, and used ARW version 3.8.1. The ARW members included four microphysics and three PBL schemes as in 2017, although the combinations were adjusted to have more P3 microphysics members,

TABLE 2. Configurations for ARW and FV3 members of the 2017 CAPS HMT FFaIR storm-scale ensemble. All members use RRTMG for parameterization of longwave and shortwave radiation. All members use no cumulus parameterization, except FV3, which uses scale-aware SAS on the global domain only.

Member	Initial conditions	Boundary conditions	Radar data	Microphysics	Land surface model	PBL
arw_cn	0000 UTC ARPSa	0000 UTC NAMf	Yes	Thompson	Noah	MYJ
arw_m2	arw_cn + arw-p1_pert	2100 UTC SREF arw-p1	Yes	P3	Noah	YSU
arw_m3	arw_cn + arw-n1_pert	2100 UTC SREF arw-n1	Yes	MY	Noah	MYNN
arw_m4	arw_cn + arw-p2_pert	2100 UTC SREF arw-p2	Yes	Morrison	Noah	MYJ
arw_m5	arw_cn + arw-n2_pert	2100 UTC SREF arw-n2	Yes	P3	Noah	MYNN
arw_m6	arw_cn + nmmb-p1_pert	2100 UTC SREF nmmb-p1	Yes	MY	Noah	MYJ
arw_m7	arw_cn + nmmb-n1_pert	2100 UTC SREF nmmb-n1	Yes	Morrison	Noah	YSU
arw_m8	arw_cn + nmmb-p2_pert	2100 UTC SREF nmmb-p2	Yes	P3	Noah	MYJ
arw_m9	arw_cn + nmmb-n2_pert	2100 UTC SREF nmmb-n2	Yes	Thompson	Noah	MYNN
arw_m10	arw_cn + arw-n3_pert	2100 UTC SREF arw-n3	Yes	Thompson	Noah	MYJ
fv3	0000 UTC GFS	—	No	Thompson	Noah	MRF

since the P3 scheme was newer and required more evaluation. Details on the member configurations are given in Table 2. The 2017 WRF ARW members again used 51 vertical levels; FV3 members used 50.

During 2016 and 2017, CAPS forecasts were run daily from 0000 UTC for the full duration of FFaIR: in 2016 on weekdays during 17 June–1 July 2016 and 11–22 July 2016, and in 2017 on weekdays during 19–30 June 2017 and 10–21 July 2017. The break between the two periods each year was to exclude the week containing the U.S. Independence Day holiday. ARW and NMMB forecasts were run to 60 h, and FV3 forecasts were run to 120 h. Forecast execution began shortly before 0200 UTC and required approximately 6 h of wall clock time, followed by data transfer and postprocessing. A standard set of two-dimensional (2D) forecast products, including specialized rainfall products developed specifically for FFaIR (described below in section 2b), was transferred to HMT in real time for use by FFaIR participants.

b. *Forecast products*

Because of bandwidth limitations for data transfer and the extremely large size of the full three-dimensional output of the CAPS SSEF, forecasts were provided to FFaIR as a suite of 2D ensemble products generated from the full 3D hourly output, including ensemble means of standard meteorological fields (e.g., zonal and meridional winds at 850, 500, and 200 hPa, sea level pressure, predicted composite radar reflectivity) and a set of specialized 2D probabilistic and ensemble products focused on extreme rainfall and flooding. Probabilistic products included probability of exceeding specific rainfall thresholds (12.5, 25, and 50 mm over 1- and 3-h periods; and 25, 50, and 75 mm over 6-, 12-, and 24-h periods) and probability of exceeding FFG (Schmidt et al. 2007, Sweeney and Baumgardner 1999) over 6-, 12-, and 24-h periods. The 2D fields for probability of exceedance of precipitation amounts corresponding to recurrence intervals (RI) of 5, 10, 25, 50, and 100 years were also calculated over each of several intervals (3-, 6-, 12-, and 24-h periods).

FFG and RI products are of particular interest for hydrological forecast applications as they take into account local hydrological conditions—RI thresholds vary spatially based on local climatology, and FFG both varies locally and, in many areas, takes into account recent precipitation and current hydrological conditions. FFG and RI data provided by WPC are interpolated to the model grid for production of and verification of these forecast products. FFG is produced by individual NWS River Forecast Centers (RFCs) in accordance with each RFC domain. WPC compiles the guidance from each RFC to create a CONUS 5-km resolution mosaic FFG grid. The CONUS mosaics are time stamped every 6 h (0000, 0600, 1200, and 1800 UTC), but are updated hourly to account for the latest guidance issued by RFCs. RI data are frequency estimates generated from a NOAA Atlas-14 climatology of USGS rain gauges (Herman and Schumacher 2016).

Probability of exceedance products were generated using a circular neighborhood algorithm considering data from within a radius of 25 km when calculating contingency table statistics, and the resulting field was smoothed using a Gaussian function defined by

$$G(x, y) = \frac{1}{2\pi\sigma^2} \exp\left[\frac{-(x^2 + y^2)}{2\sigma^2}\right], \qquad (1)$$

where σ is the standard deviation (set to 30 km). The specific neighborhood algorithm used is a neighborhood maximum ensemble probability (NMEP; e.g., Schwartz and Sobash 2017) method, which sets the probability of event occurrence at a point to 1.0 if the event occurred

anywhere in that point's neighborhood, and sets the probability to 0.0 otherwise. The neighborhood radius of 25 km and standard deviation of 30 km were chosen based on requests by HMT to match settings used in other existing experimental products. For these probabilistic products, only the subset of WRF ARW members using SREF boundary conditions were used; NMMB and FV3 members were not included.

For the 2017 HMT FFaIR, a patchwise LPM ensemble mean algorithm was developed and used to produce QPF products. The LPM is based upon the probability-matched (PM) mean (Ebert 2001), which produces a forecast field with the spatial structure of the ensemble mean, but with values sampled from the full distribution of all ensemble members. The PM mean is often more skillful than a simple ensemble mean (e.g., Clark et al. 2009; Xue et al. 2011; Schwartz et al. 2014), but exhibits a loss of small-scale structure compared to individual ensemble members (Surcel et al. 2014). Furthermore, for a large domain, the PM mean at any given point may combine precipitation information from very different mesoscale environments and/or geographic regions (Clark 2017), such as coastal sea-breeze convection and stratiform precipitation over the northern plains.

To efficiently produce an LPM product, we apply the PM mean algorithm over a series of local patches. The domain is divided into a set of rectangular local patches, each centered within a larger, rectangular LPM domain. The patches do not overlap, but the domains of adjacent or nearby patches do. A conceptual illustration of this setup is shown in Fig. 2. After the LPM mean is calculated for all points on each of the local patches, the patches are stitched together to form a single field for the full CONUS domain, and the Gaussian smoother of Eq. (1) is applied to minimize discontinuities along patch boundaries, using $\sigma = 3$ km (1 grid point).

Based on prior experimentation, the LPM implementation for the 2017 FFaIR used local patches with dimensions of 5×5 grid points, LPM domains of 60×60 grid points, and $\sigma = 3$ km for the Gaussian smoother. This configuration was found to provide a good balance between forecast quality and computational expense, though we note that the choices of these settings do have a substantial impact on the resulting LPM product. The sensitivity of the LPM to configuration settings, including patch size, LPM domain size, and smoothing parameters, as well as comparison between the performance of the patchwise LPM used in this study with the more computationally expensive point-by-point LPM algorithm of Clark (2017), will be documented in a separate paper.

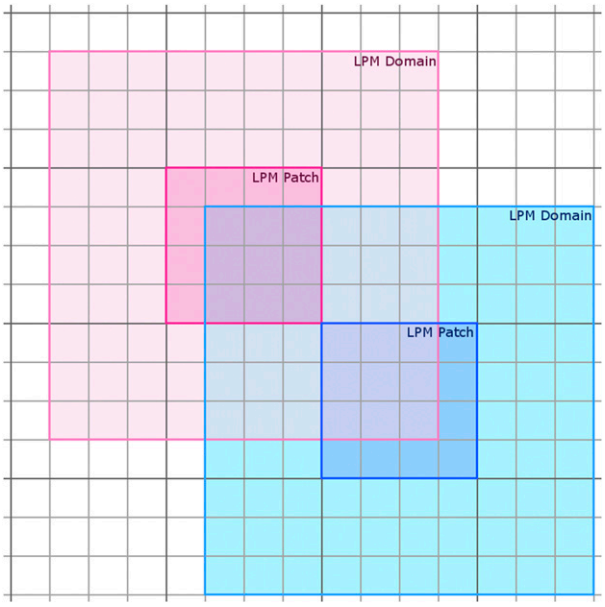

FIG. 2. Conceptual illustration of the patches used to generate localized probability matched mean products. Highlighted are two patches (the darker magenta and blue shaded regions) along with their associated calculation areas LPM domains (the lighter magenta and blue shaded regions surrounding the patches). The gray lines indicate the edges of model grid cells.

3. Verification of precipitation forecasts

In this section, we present verification of precipitation forecasts of the 2016 and 2017 CAPS SSEFs, for the full HMT FFaIR experiment periods. All verifications are performed over the verification domain, indicated by the inner gray boundary in Fig. 1. The biases and equitable threat scores [ETS, also known as the Gilbert skill score (GSS); Mason 2003] will be examined, followed by scale-dependent verifications using fractions skill scores (Roberts and Lean 2008) and examinations of the precipitation power spectra (Denis et al. 2002; Surcel et al. 2014). Forecasts of 3-h rainfall exceeding 0.01 and 0.50 in. (0.025 and 12.7 mm, respectively), as well as forecasts of rainfall exceeding flash flood guidance and rainfall exceeding 10-yr recurrence intervals will also be considered. To examine the statistical significance of differences in forecast bias and ETS performance, a bootstrap resampling method is used to generate 10 000 realizations. To perform the bootstrap, daily data are aggregated for all available forecast days for a given member or subensemble. From this pool of data, daily data are randomly sampled with replacement to generate a resampled set of daily data containing the same number of individual forecasts and days as the original dataset. This process is repeated 10 000 times to produce the full set of 10 000 realizations. This large pool of

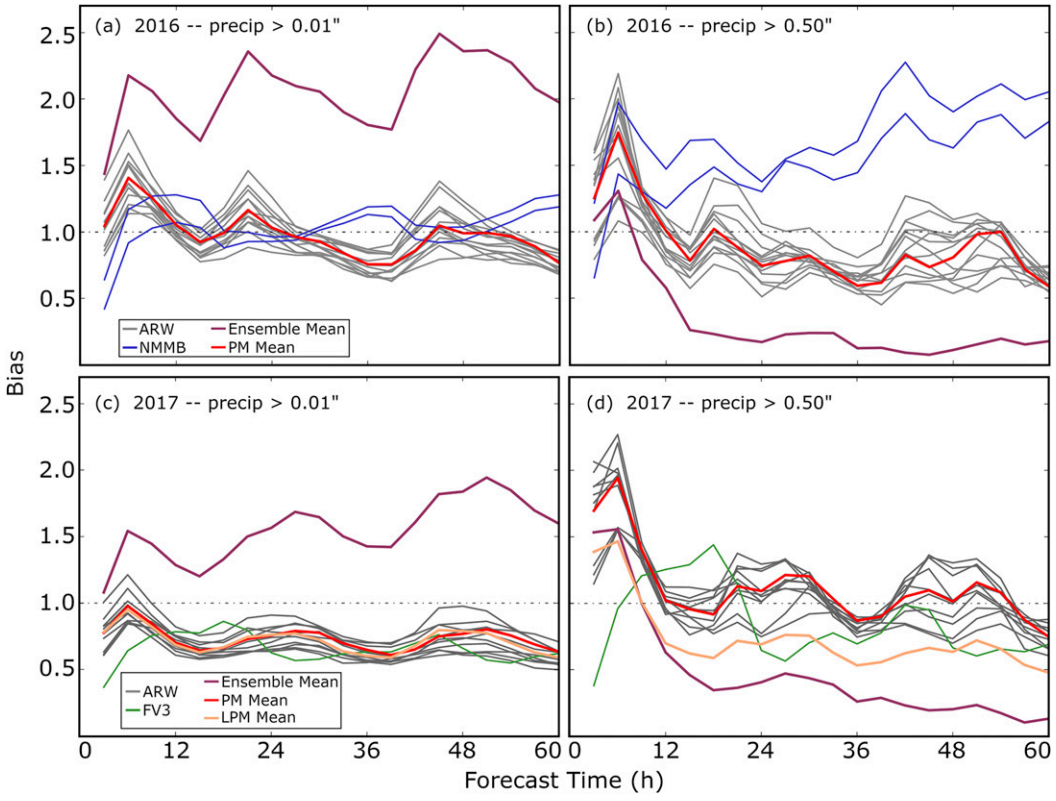

FIG. 3. Frequency bias of 3-h accumulated precipitation, verified against 3-h accumulations calculated from observed hourly MRMS precipitation data, for all operational days of the CAPS (a),(b) 2016 and (c),(d) 2017 HMT ensemble forecasts for regions with precipitation (left) exceeding 0.01 in. (0.254 mm) and (right) exceeding 0.50 in. (12.7 mm). The horizontal dash-dotted line in each panel indicates a bias value of 1.0 (unbiased). Biases are plotted at 3-h intervals between 3 and 60 h of forecast time.

resampled forecasts allows us to examine the statistical significance of differences among ensemble members or subensembles using, for example, the 5th–95th percentile range of the resampled forecasts (two forecasts can be considered to differ significantly when the value of a verification metric (e.g., bias, ETS) for one member falls outside the 5th–95th percentile range of that metric in the other member). For probabilistic forecasts, area under the ROC curve (Mason 1982) and reliability diagrams will be used for objective verification.

a. Grid-based verification of 2016 and 2017 precipitation forecasts

Rainfall accumulations from NOAA multiradar/multisensor (MRMS; Zhang et al. 2016) precipitation estimates are used to verify precipitation accumulation forecasts. For use in verification, MRMS data were regridded from their native grid with 1-km horizontal grid spacing to the HMT forecast grid (which uses 3-km horizontal grid spacing) via bilinear interpolation. No additional quality control or filtering was applied to MRMS data after regridding, and only days for which full MRMS data were available were used for verification. Verification of forecasts against MRMS data are performed on the model grid over a verification subdomain encompassing most of the CONUS [indicated by the gray (innermost) box in Fig. 1a for 2016 and Fig. 1b for 2017]. Data from all days for which complete datasets are available for both MRMS precipitation and the CAPS SSEF are used for verification.

Precipitation frequency biases for 2016 and 2017 individual member forecasts and for simple, PM, and LPM means, are plotted in Fig. 3 for areas with 3-h precipitation accumulation exceeding 0.01 in. (0.254 mm; Figs. 3a,c) and 0.50 in. (12.7 mm; Figs. 3b,d). When considering areas of precipitation exceeding 0.01 in. (Figs. 3a,c), the simple mean has the greatest bias (1.5–2.5), compared to bias values of around 0.5–1.5 for individual members and the PM and LPM means. The large biases of the simple mean result from the presence of large areas of light precipitation created by the smoothing effect of the mean. In general, bias is higher during afternoon and evening hours, around 18–24 and 42–48 h of forecast time (Figs. 3a,c), for the ARW

members, consistent with the peak in convective activity in the late afternoon. The NMMB members of the 2016 ensemble (Fig. 3a) and the FV3 member of the 2017 ensemble (Fig. 3c) exhibit low biases initially due to lack of radar data assimilation, but exhibit biases at the 0.01-in. threshold similar to those of ARW members by 6–9 h of forecast time. Bias at the 0.01-in. threshold is lower in 2017 (Fig. 3c) than in 2016 (Fig. 3a) by approximately 0.2–0.5 for both individual ensemble members and the ensemble mean; given that there were no major changes to the dynamic cores of the models used or to the implementations of PBL or microphysical schemes within the ARW model between the running of the 2016 and 2017 CAPS SSEF forecasts, we speculate that the lower overall bias in 2017 may result in part from differences in large-scale flow on seasonal or subseasonal time scales and differences in the overall prevalence of convection between the 2016 and 2017 HMT operational periods.

Bias in areas of heavy precipitation (3-h accumulations exceeding 0.50 in.; Figs. 3b,d) is similar in magnitude for individual ARW ensemble members, with values ranging from 1.3–2.3 before 9 h and 0.5–1.3 thereafter, with similar behavior in 2016 and 2017 (Figs. 3b,d). The larger high biases in the first 9 h are attributable to assimilation of radar reflectivity data using cloud analysis, which does improve ETS scores (see Fig. 7); nmmb_m1 and FV3, neither of which assimilate radar data, exhibit lower biases than most or all ARW members during the initial hours at the 0.50-in. threshold (Figs. 3b,d).

In 2016, three ARW ensemble members—arw_cn, arw_m11, arw_m12, and arw_m13 in Table 1—differed only in microphysical scheme; together with the control member arw_cn, these comprise a physics-perturbation-only subensemble that can be used to investigate the impact of the microphysical scheme on forecast performance. Among this physics-perturbation-only subensemble (Fig. 4a), the Thompson, MY2, and P3 members exhibit good performance in terms of bias in accumulated precipitation at the 0.01-in. threshold, with bias values ranging from 1.0 to 1.2 throughout the forecast period. The Morrison member exhibits higher biases at the 0.01-in. threshold than to the other three members (Fig. 4a); at many times these differences are statistically significant, as the bias of the Morrison member falls outside of the 5th–95th percentile range of the other (P3, Thompson, MY2) members (based on the bootstrap resampling utilized here). At the 0.50-in. threshold (Fig. 4b), during the first 9 h, the control (Thompson) member (arw_cn) has lower biases than most other members, possibly because the unperturbed initial conditions, produced by the cloud analysis and used by all

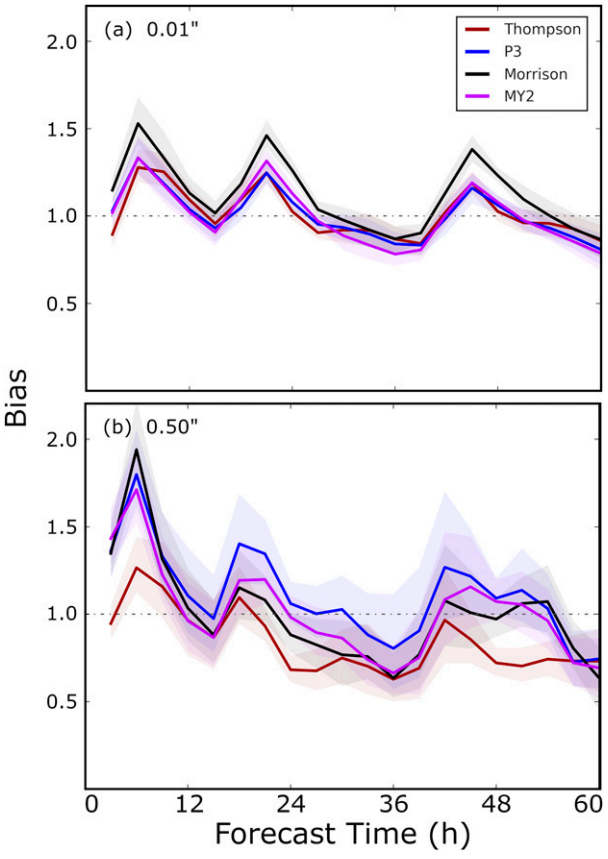

FIG. 4. Frequency bias of 3-h accumulated precipitation, verified against 3-h accumulations calculated from observed hourly MRMS precipitation data for individual members of the ARW physics-only subensemble from 2016 using the Thompson (red), P3 (blue), Morrison (black), and Milbrandt and Yau two-moment (purple) microphysics schemes, for precipitation exceeding (a) 0.01 in. (0.25 mm) and (b) 0.50 in. (12.7 mm). The light colored shading indicates the 5th–95th percentile range for each subensemble based upon bootstrap resampling of cases using 10 000 samples. The horizontal dash-dotted line in each panel indicates a bias value of 1.0 (unbiased).

members of the physics-perturbation-only subensemble, are more consistent with the microphysical scheme used in the control member than in the other members. The Thompson microphysical scheme used by the control member is double moment for rain and cloud ice and single moment for all other hydrometeor species, which is more consistent with the forward operator of the cloud analysis system (which, as run for this experiment, initialized microphysics consistent with the single moment Lin scheme). The other members of the physics-perturbation-only subensemble use fully double-moment schemes (MY2, Morrison) or more complex schemes (P3). As all members of the physics-perturbation-only subensemble share the control initial conditions, the substantially higher biases of the

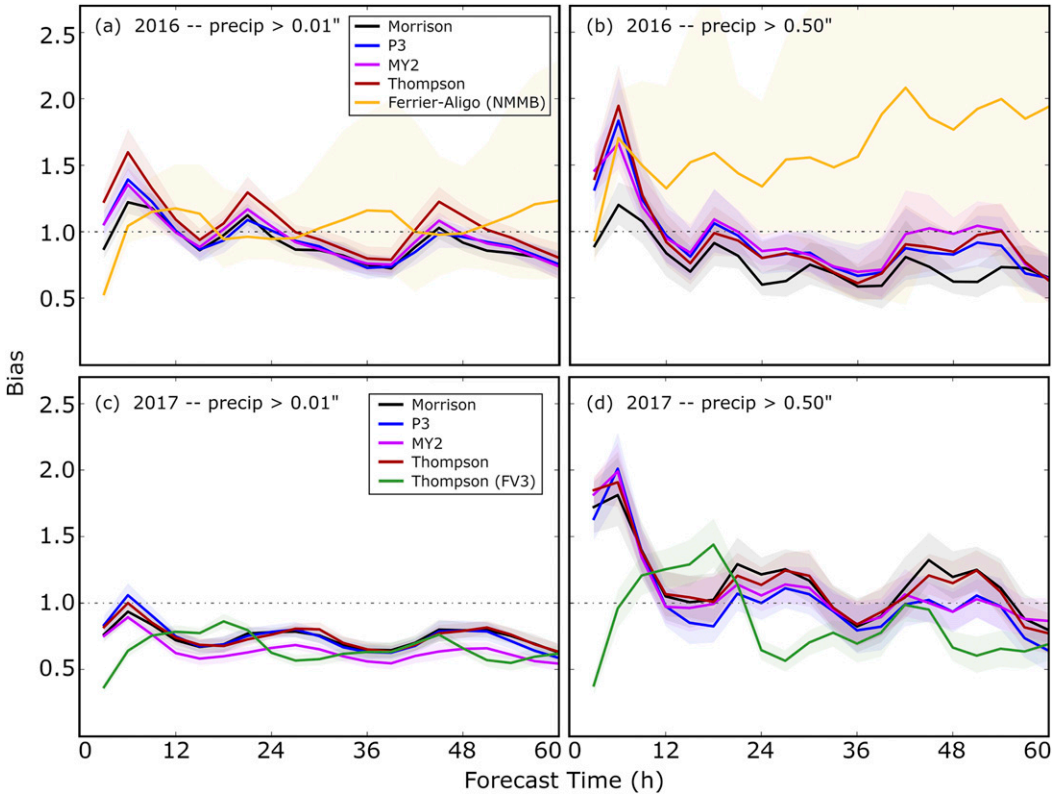

FIG. 5. Frequency bias of 3-h accumulated precipitation, verified against 3-h accumulations calculated from observed hourly MRMS precipitation data, for all operational days of the CAPS (a),(b) 2016 and (c),(d) 2017 HMT ensemble forecasts for regions with precipitation (left) exceeding 0.01 in. (0.25 mm) and (right) exceeding 0.50 in. (12.7 mm). Forecasts are grouped as subensembles containing members with the same microphysical parameterization. The light colored shading indicates the 5th–95th percentile range for each subensemble based upon bootstrap resampling of subensemble members and cases using 10 000 samples. The horizontal dash-dotted line in each panel indicates a bias value of 1.0 (unbiased). Biases are plotted at 3-h intervals between 3 and 60 h of forecast time.

Morrison, MY2, and P3 members during the initial 9-h period (Fig. 4b) suggests a strong sensitivity to the choice of microphysical scheme in the initial hours after the cloud analysis.

Sensitivity of forecast bias to the choice of microphysical and PBL schemes can also be considered by dividing the 2016 and 2017 CAPS SSEF forecasts into subensembles that share a common microphysical or PBL scheme. Bias is plotted for subensembles of CAPS SSEF members sharing a common microphysical scheme in Fig. 5, and for subensembles of members sharing a common PBL scheme in Fig. 6. The NMMB members of the 2016 CAPS SSEF used the MYNN PBL scheme, and the FV3 member of the 2017 SSEF used the Thompson microphysics scheme, in both of these cases sharing a scheme with some of the ARW members run that year. Despite this, the FV3 and NMMB members are considered as separate subensembles to account for difference in the model core (e.g., ARW versus FV3) and differences in implementation of the schemes between the models.

Among subensembles of members sharing a common microphysical scheme (Fig. 5), bias at the 0.01-in. threshold in 2016 (Fig. 5a) is highest (among the ARW subensembles) for members using the Thompson microphysical scheme at all hours, and except between approximately 30 and 39 h of forecast time, these differences are statistically significant in the 5th–95th percentile range calculated using bootstrap resampling. During later forecast hours, bias is closer to 1.0 in the Thompson subensemble than in the P3, MY2, and Morrison subensembles (differences between these other three ARW subensembles were not statistically significant). In 2017 (Fig. 5c), when bias at the 0.01-in. threshold was lower overall, the Thompson subensemble had bias similar to the Morrison and P3 subensembles, while the MY2 subensemble had the most substantial low bias among ARW subensembles

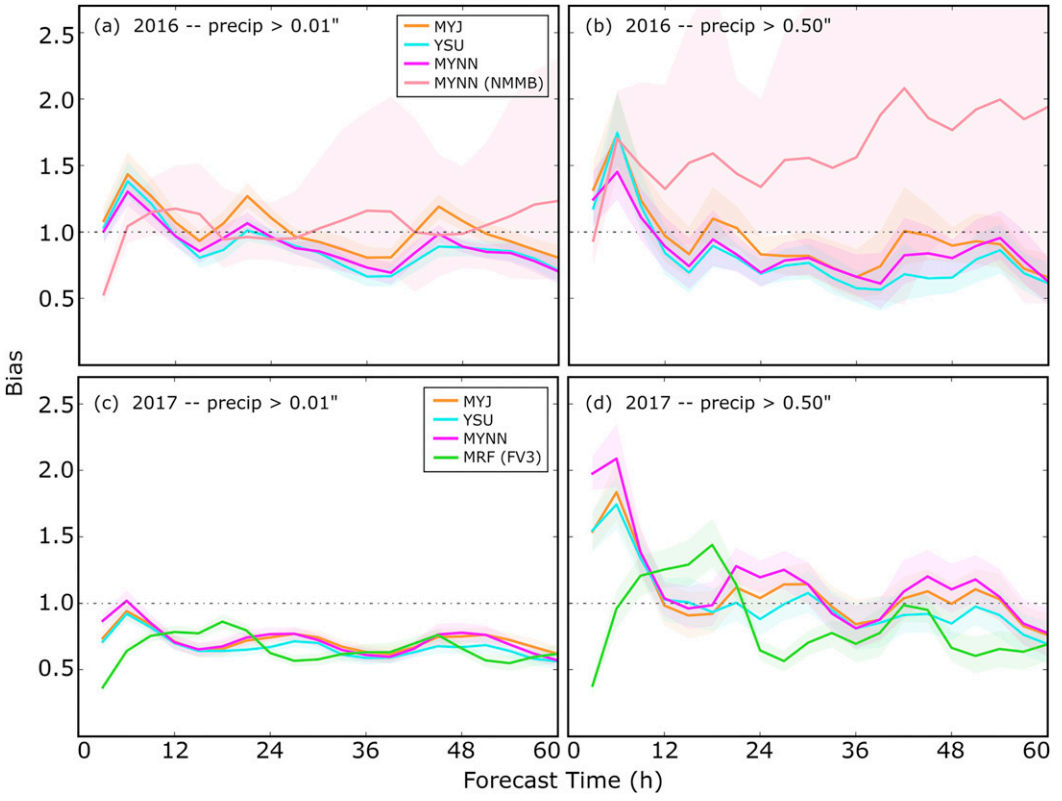

FIG. 6. As in Fig. 5, but for subensembles of members sharing a common PBL parameterization scheme rather than a common microphysical scheme.

(this difference was statistically significant between 3 and 57 h of forecast time). At the 0.50-in. threshold, the Morrison subensemble exhibits the largest negative bias during 2016 (Fig. 5b)—this difference is significant for some but not all forecast hours—and the P3 subensemble exhibits the smallest high bias in 2017 (even exhibiting a low bias at later forecast hours), though the difference is not significant at most forecast hours.

The bias behavior of the subensemble composed of the two NMMB members, which use the Ferrier-Aligo microphysics scheme differs substantially from the ARW subensembles (Figs. 5a,b), showing a somewhat different diurnal cycle at the 0.01-in. threshold (Fig. 5a) and much higher bias overall at the 0.50-in. threshold (Fig. 5b). The 5th–95th percentile range, determined using bootstrap resampling, is also much larger for the NMMB subensemble than for any of the ARW subensembles—a result of large day-to-day variability in the biases of the NMMB members (though both members tended to overpredict the prevalence of heavy precipitation, as evidenced by the biases of this subensemble approaching 2.0 at the 0.50-in. threshold). The ARW members and the FV3 member of the 2017 SSEF (Figs. 5c,d) do not exhibit such large day-to-day variations in bias.

As noted earlier, differences in overall bias behavior between the 2016 and 2017 CAPS SSEFs may result in part from seasonal differences in large-scale flow and differences in the overall prevalence of convection during these two years, as well as the relative frequency of different modes of convection. The lack of a consistent pattern in bias among the ARW microphysical subensembles between 2016 and 2017 suggests that the sensitivity to the microphysical scheme is also affected by the factors such as those mentioned above; though examining data from a larger set of years (including future CAPS SSEFs) would likely be necessary to further clarify such impacts.

Among subensembles of members sharing a common PBL scheme (Fig. 6), bias in the 2016 ensemble is largest in magnitude among ARW members for those using the MYJ scheme at most forecast hours (Figs. 6a,b); this difference is statistically significant from other ARW subensembles after 12 h of forecast time at the 0.01-in. threshold (Fig. 6a). Later in the forecast period (24–60 h), the higher bias in MYJ members results in superior forecast performance, as the YSU and MYNN

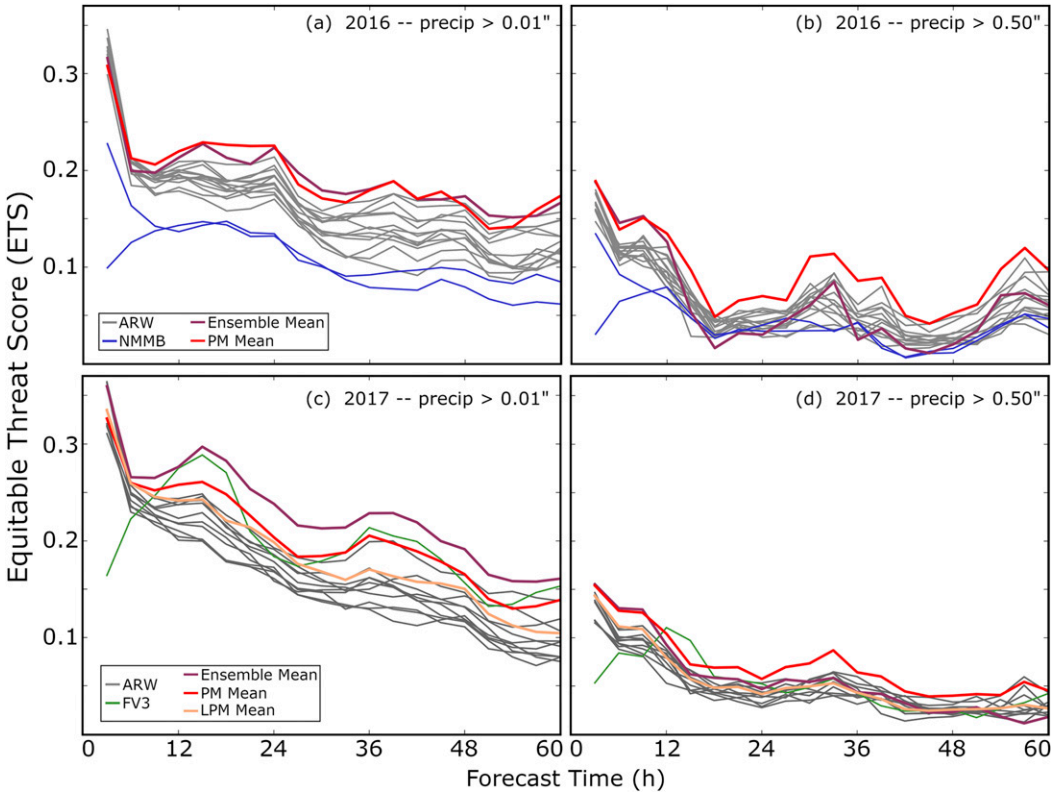

FIG. 7. As in Fig. 3, but ETS of 3-h accumulated precipitation verified against 3-h accumulated precipitation from MRMS data.

subensembles exhibit negative biases at both the 0.01- and 0.50-in. thresholds (Figs. 6a,b). In the 2017 ensemble, the YSU subensemble exhibits a significantly larger low bias than the MYJ and MYNN subensembles during the afternoon and evening hours (18–30 and 42–54 h of forecast time) at both the 0.01- (Fig. 6c) and 0.50-in. (Fig. 6d) thresholds.

While there is not a clear pattern in bias for subensembles sharing a common microphysical scheme, in both 2016 and 2017, the subensemble of ARW members with YSU PBL scheme consistently exhibits the most substantial low bias (among ARW subensembles) at both 0.01- and 0.50-in. thresholds (Fig. 6). Also of note, in both years the MYJ subensemble exhibits either the least or close to the least bias (i.e., remains closest to a bias of 1.0).

In Fig. 7, ETS is plotted for 2016 and 2017 SSEF forecasts of precipitation exceeding 0.01 in. (Figs. 7a,c) and 0.50 in. (Figs. 7b,d). ETS can range from −0.33 to 1.0, with higher scores indicating more skillful forecasts. Except for the two ensemble members that do not assimilate radar observations (nmmb_m1 in 2016 and FV3 in 2017), ETS generally decreases with increasing forecast time. This decrease is most rapid during the initial 3–6 h of forecast, as the immediate impact of radar data assimilation decreases. In the two members that do not assimilate radar observations, the ETS is initially much lower than for the radar-assimilating members due to the necessary precipitation spinup, but increases during the first 12 h of the forecast, after which performance is generally indistinguishable from radar-assimilating members. These behaviors of radar data impact are similar to those documented for earlier CAPS SSEF forecasts performed during prior HWT SFEs (e.g., Kain et al. 2010, Xue et al. 2013).

At the 0.01-in. threshold, ETS declines from between 0.30 and 0.35 at 3 h of forecast time to between 0.10 and 0.20 by 60 h of forecast time; the range and evolution of ETS scores are similar between the 2016 (Fig. 7a) and 2017 (Fig. 7c) ensembles. In 2016, almost all ARW members substantially outperform the NMMB members throughout the forecast period (Fig. 7a). In 2017, the FV3 member performs slightly better than most ARW members at the 0.01-in. threshold after the first 9 h (Fig. 7c). For this low threshold, the simple ensemble mean has the highest ETS, especially in 2017, as a result of widespread light precipitation in the simple mean. The PM mean slightly outperforms the best

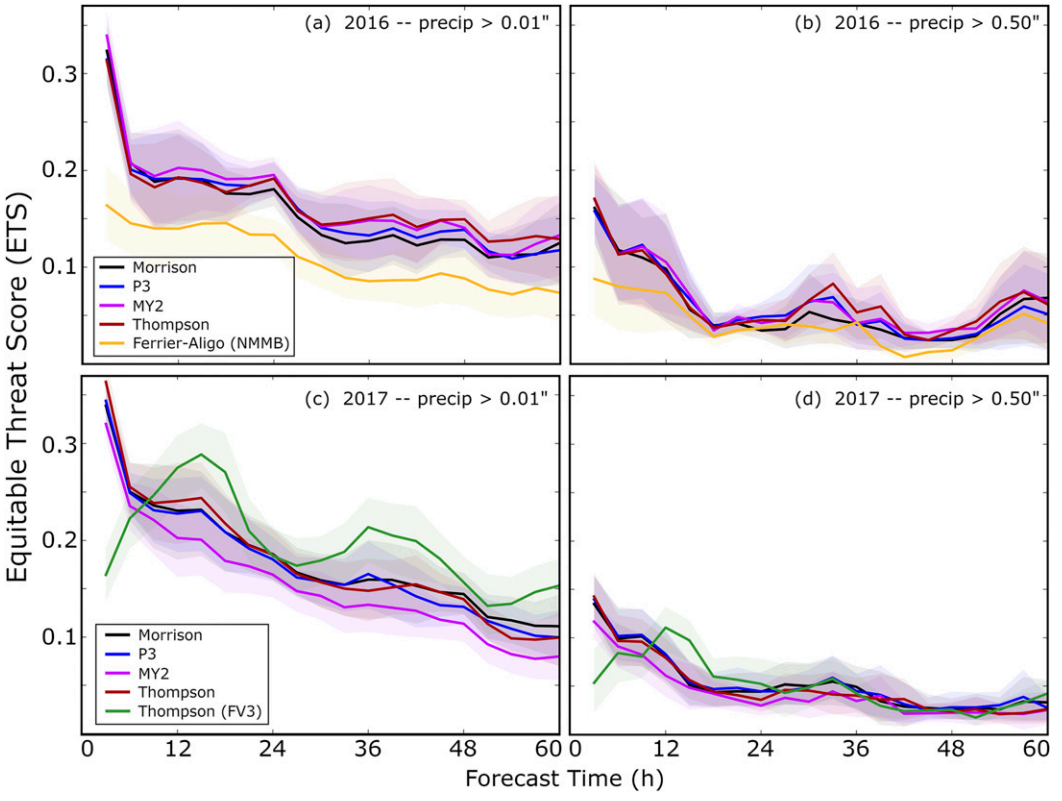

FIG. 8. As in Fig. 5, but for ETS of 3-h accumulated precipitation verified against 3-h accumulated precipitation from MRMS data.

ARW members, and the LPM mean outperforms all ARW members except the control member.

At the 0.50-in. threshold, ETS scores are lower throughout the forecast period, falling from between 0.1 and 0.2 at the start of the forecast to generally below 0.1 by 60 h. As at the 0.01-in. threshold, ARW members generally outperform NMMB members in 2016 (Fig. 7b). The 2017 FV3 member performs similarly or slightly better than the ARW members (Fig. 7d). At the 0.50-in. threshold, the PM mean outperforms the simple mean after the first 12 h, and outperforms the LPM mean throughout the forecast period (Figs. 7b,d). Clark (2017) found that skill scores for PM mean forecasts using a large domain can be inflated as a result of redistribution of precipitation; this is one possible reason for the superior performance of the PM compared to the LPM in terms of ETS. The LPM algorithm is designed to retain local precipitation structures and magnitude distributions, and thus would not exhibit the type of skill score inflation noted for the PM mean by Clark (2017).

ETS is considered for subensembles of the 2016 and 2017 sharing a common microphysical scheme and the statistical significance of the differences is evaluated using bootstrapping in Fig. 8. Among the ARW subensembles (divided into subensembles of members using the Morrison, P3, MY2, and Thompson microphysical schemes), there are some differences in ETS, but these differences are not generally statistically significant, either in 2016 or 2017 (Fig. 8). As with the bias (Fig. 4), there is not a consistent pattern in ETS among microphysical subensembles between 2016 and 2017; in 2016 the Thompson and MY2 subensembles exhibit slightly better ETS scores, particularly later in the forecast period (Figs. 8a,b), while in 2017 the MY2 members exhibit the lowest ETS score overall and the Thompson members are still at or near the top, particularly at the 0.01-in. threshold (Fig. 8c), though the differences are not statistically significant at most forecast hours.

For subensembles of members sharing a common PBL scheme, the MYJ subensemble consistently exhibits the highest ETS among subensembles of ARW members in both 2016 and 2017 (Fig. 9), particularly at the 0.01-in. threshold (Figs. 9a,c), although this difference is marginally significant, at best. The MYNN subensemble consistently exhibits the lowest ETS. The consistency in ETS behavior between 2016

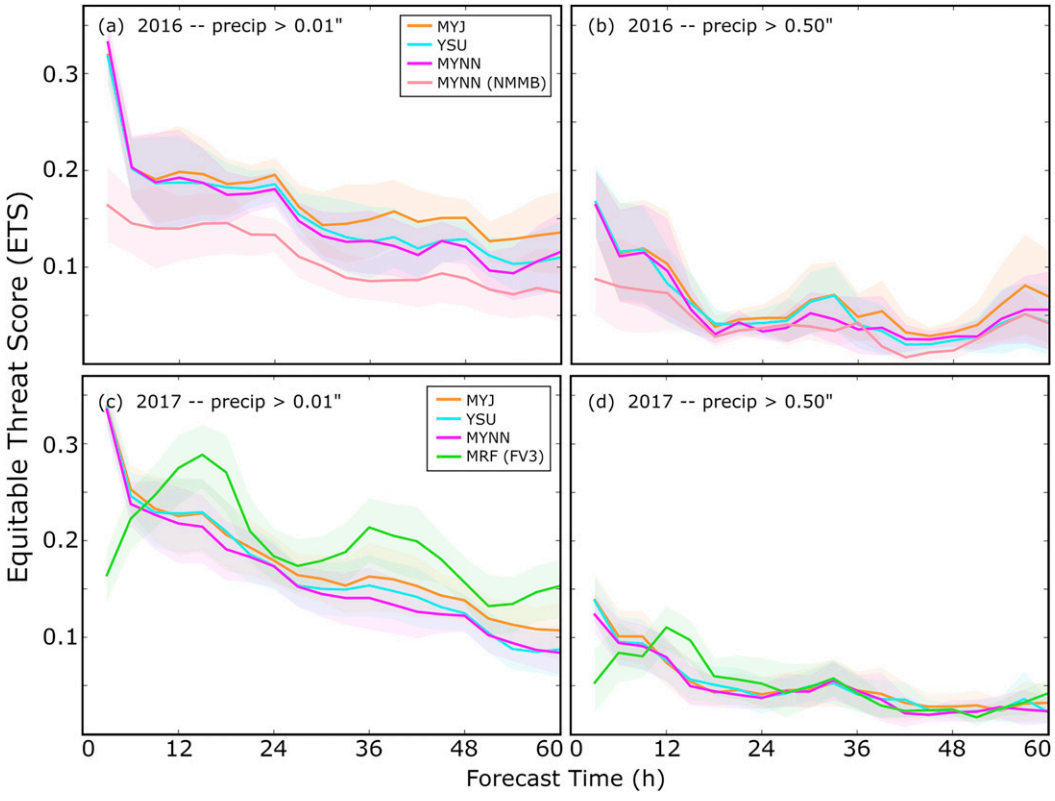

FIG. 9. As in Fig. 6, but for ETS of 3-h accumulated precipitation verified against 3-h accumulated precipitation from MRMS data.

and 2017 for subensembles sharing a common PBL scheme (and lack of similar consistency for subensembles sharing a common microphysical scheme), similar to the pattern noted in precipitation bias (Figs. 4 and 5), further suggests that sensitivity of model performance to PBL scheme may be more consistent year-to-year than sensitivity of model performance to microphysical scheme, although additional years of data would be needed to confirm this tendency.

From Figs. 8 and 9 we can also note that the relatively poor performance of the NMMB members of the 2016 ensemble in terms of ETS is statistically significant at most hours at the 0.01-in. threshold (Figs. 8a and 9a), while the FV3 member of the 2017 ensemble exhibits significantly better ETS performance at the 0.01-in. threshold at many forecast hours (Figs. 8c and 9c). The better performance of the FV3 member is particularly pronounced at the 0.01-in. threshold during early morning hours (12–21 and 36–45 h of forecast time), suggesting that the FV3 member better handles the spatial distribution of light precipitation during the diurnal convective minimum. During these forecast hours, the FV3 member also exhibits a bias closer to 1.0 compared to the ARW subensembles (Figs. 3c and 5c).

b. *Scale-dependent evaluations of precipitation forecast skill*

The skill scores presented in the previous subsection are point based, and thus do not account for any position/displacement errors, nor provide information about the scales at which forecasts are skillful. The fractions skill score (FSS; Roberts and Lean 2008), however, provides information on both. The FSSs for the CAPS 2017 HMT forecasts are calculated for 3-h accumulated precipitation exceeding 0.01 (Figs. 10a,c) and 0.50 in. (Figs. 10b,d), for forecasts valid at 24 (Figs. 10a,b) and 36 h (Figs. 10c,d), over a range of spatial scales (neighborhood radii) from 3 to 200 km. FSS can vary from 0 to 1, with higher scores indicating greater correspondence between the forecast and observations (i.e., higher skill). The minimum FSS considered to indicate useful skill is dependent on the fraction of the domain for which the forecast event is observed; this threshold is indicated by the horizontal dotted lines in the panels of Fig. 10.

For 24-h forecasts, the FSS suggests individual members and ensemble mean products exhibit useful skill for

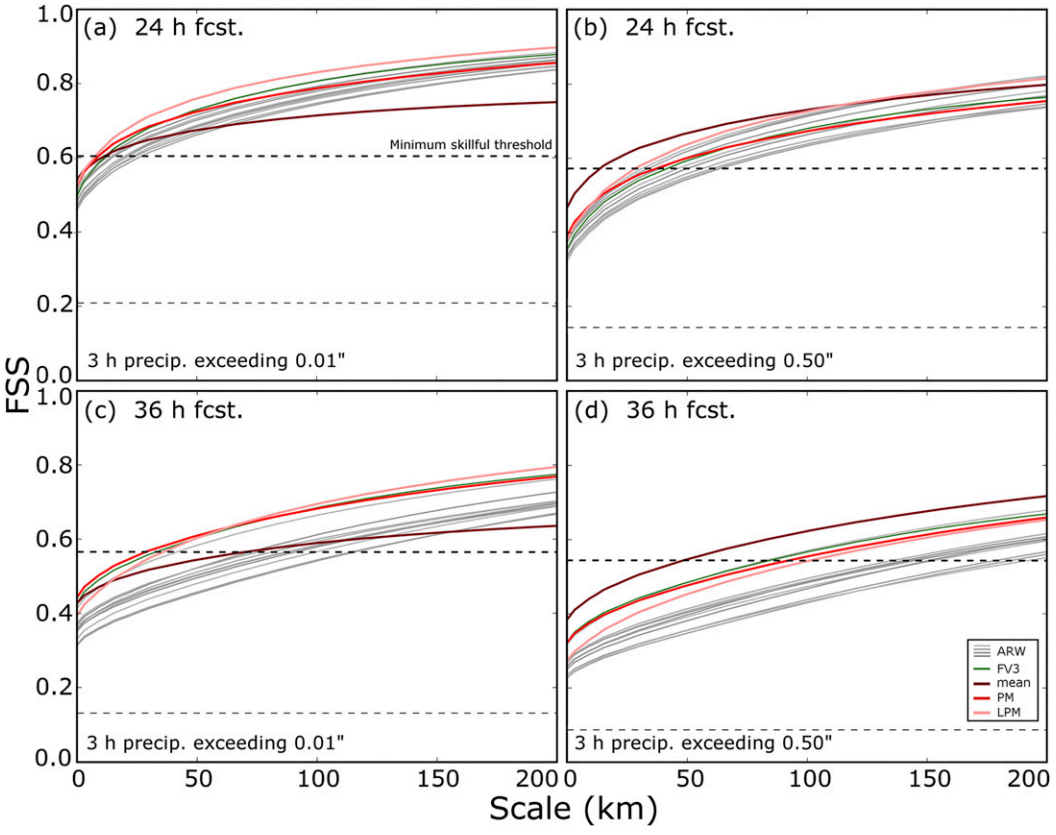

FIG. 10. Fractions skill score of 3-h accumulated precipitation forecasts for scales up to 200 km, averaged over all days for which the CAPS SSEF was run during the 2017 HMT FFaIR, for 3-h accumulated precipitation forecasts valid at (a),(b) 24 and (c),(d) 36 h of forecast time, for accumulated precipitation exceeding (left) 0.01 and (right) 0.50 in. Shown are individual ARW members (gray), the single FV3 member (green), and three different variants of the ensemble mean (varying shades of red) including the simple ensemble mean, the probability-matched mean (PM), and the localized probability matched mean (LPM).

scales larger than approximately 40 km at the 0.01-in. threshold (Fig. 10a) and 70 km at the 0.50-in. threshold (Fig. 10b). The minimum skillful scale is slightly higher (worse) for 36-h forecasts: as high as 120 km at the 0.01-in. threshold (Fig. 10c) and 200 km at the 0.50-in. threshold for some members (Fig. 10d). The minimum skillful scale is generally lower (better) for ensemble mean products than for most or all individual members at both thresholds, particularly for 36-h forecasts. At the 0.01-in. threshold, FV3 outperforms most or all ARW members for scales less than approximately 300 km (Figs. 10a,c), while at the 0.50-in. threshold it outperforms most ARW members at scales less than about 150 km (Figs. 10b,d).

Among the ensemble mean products, the simple mean has the smallest minimum skillful scale at the 0.50-in. threshold, but is substantially outperformed by the PM and LPM means at the 0.01-in. threshold, particularly at larger scales (Fig. 10). The poor performance of the simple ensemble mean at this low threshold is largely due to the extremely high bias in the coverage of 0.01-in. precipitation (see Fig. 3c) resulting from smoothing of the precipitation field when averaging members with spatially disparate precipitation. This smoothing leads to overly large precipitation fractions, causing a direct negative impact on the FSS. At the 0.50-in. threshold, the simple ensemble mean outperforms the PM and LPM at all scales in the 36-h forecast (Fig. 10d). The LPM generally outperforms the PM and simple mean at scales larger than around 20 km at the 0.01-in. threshold (Figs. 10a,c), and outperforms them at all scales in the 24-h forecast at the 0.50-in. threshold (Fig. 10b). The generally good performance of the LPM compared to the PM suggests that the localization used by the LPM algorithm is succeeding in improving the spatial structure of this forecast product (in terms of matching the spatial structure of precipitation in the observations).

When the microphysical scheme is considered (Fig. 11), two notable features become apparent. First, FV3, which uses an implementation of the Thompson microphysics scheme, generally has among the highest

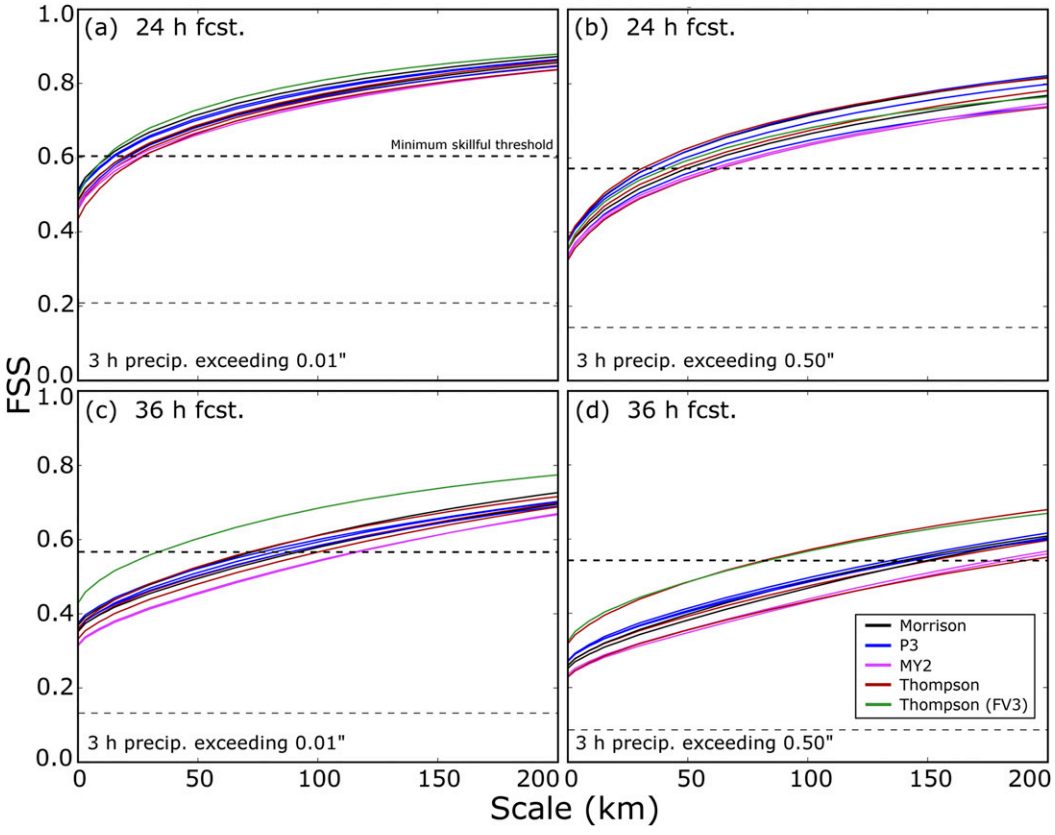

FIG. 11. As in Fig. 10, but for individual WRF-ARW members and the FV3 member of the 2017 ensemble, color-coded by microphysical scheme used.

FSS of all the members at the 0.01-in. threshold for all scales less than 200 km (Figs. 11a,c). The good performance of FV3 at small scales will be discussed in more detail below. Second, among the ARW members, those using MY2 generally perform relatively poorly in terms of FSS, having among the largest minimum skillful scale of all members at both the 0.01- and 0.50-in. thresholds in 24- and 36-h forecasts (Fig. 11).

Another way of examining the scale-dependent properties of forecasts is examination of variance spectra. Spectra of forecast fields contain information on whether the model is correctly simulating and distributing the variance (colloquially ''power'') of a forecast field across its resolvable scales (e.g., Skamarock 2004; Surcel et al. 2014). We consider spectra of 3-h accumulated precipitation (Fig. 12) over the verification domain (see Fig. 13), calculated using a 2D discrete cosine transform method (Denis et al. 2002; Surcel et al. 2014) and verified against observed 3-h MRMS accumulated precipitation (Fig. 12). Spectra are calculated for accumulated 3-h precipitation between 0900 and 01200 UTC (12 h of forecast time; Fig. 12a) and between 2100 and 0000 UTC (24 h of forecast time; Fig. 12b), averaged over all days of the 2017 HMT FFaIR period.

The total power present in the observed MRMS accumulated precipitation field is greater, particularly at small scales, at 0000 UTC (Fig. 12b) than at 1200 UTC (Fig. 12a); this is expected, as the 0000 UTC forecasts correspond to late afternoon, when convective storms are often more abundant over the CONUS region, while the 1200 UTC forecasts correspond to early morning hours when convection is less prevalent. In general, ARW members exhibit slightly higher power in the precipitation spectrum than MRMS observations for both 0000 and 1200 UTC forecasts at scales larger than 20 km (Fig. 12). At scales less than 20 km, power quickly drops off both in the ARW members and in the ensemble mean products, as the model fails to capture power in the precipitation field at the smallest scales (at wavelengths less than about $6\Delta x$). Interestingly, FV3 does not exhibit as much drop-off in power at the small scales; compared to MRMS, the power starts to drop off at about 10-km wavelength that is close to $4\Delta x$ (Fig. 12). While this could be in part due to the tendency for this

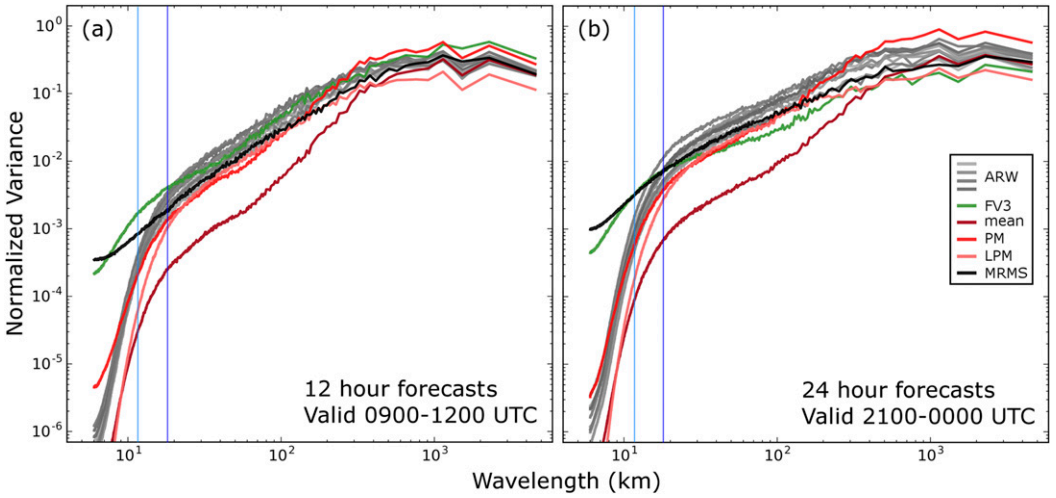

FIG. 12. Normalized variance spectra, averaged over all days for which the CAPS SSEF was run during the 2017 HMT FFaIR, for (a) 12-h forecasts of 3-h precipitation accumulation between 0900 and 01200 UTC and (b) 24-h forecasts of 3-h precipitation accumulation between 2100 and 0000 UTC. Individual ensemble members are plotted in gray (ARW) and green (FV3), while ensemble mean products (including PM and LPM) are plotted in red. Observed 3-h precipitation accumulation from MRMS is plotted in black. For reference, blue vertical lines are plotted at wavelengths equivalent to 4 (lighter blue) and 6 (darker blue) times the grid spacing.

implementation of FV3 to produce relatively small convective cells (Potvin et al. 2018), it also suggests that FV3 better replicates the types of structures seen at small scales in the precipitation field compared to ARW. The difference in the behavior of the FV3 member and the ARW members is at least partially due to the presence of relatively aggressive damping and diffusion used in the ARW members [set to be consistent with those used in the operational High-Resolution Rapid Refresh (HRRR) model, made necessary by its relatively small vertical spacings near the tropopause]. FV3 also exhibits slightly lower overall power than both the ARW members and the MRMS observations in the 2100–0000 UTC forecasts, again resulting in part from the low bias in heavy convective precipitation in the FV3 forecasts (see Fig. 3d).

Among the ensemble mean products, the simple ensemble mean exhibits a very substantial low bias in terms of power in the accumulated precipitation field at scales up to around 300 km in the 12-h forecast (Fig. 12a) and 500 km in the 24-h forecast (Fig. 12b). This result is consistent with Surcel et al. (2014), who found that spatial smoothing resulted in a loss of power in the ensemble mean for small scales, and furthermore at the largest scale at which the ensemble mean lost power due to spatial averaging that increased with increasing forecast lead time. In contrast, the PM and LPM means exhibit similar accumulated precipitation spectra to MRMS for scales greater than 20 km. Below 20 km, the PM mean spectra drop off slightly slower than those of the individual ARW members, presumably because the largest values among all ensemble members are concentrated near common grid points the value of the ensemble mean is largest, creating extra power near the grid scale. For the LPM mean, the drop-off rate is somewhat faster than that of individual ARW members, reaching similar levels to the simple mean at the smallest scales, suggesting that the recovery of small-scale structures is limited by the use of local patches, and the smoothing effect of the ensemble mean at the smallest scales remains.

The PM, however, exhibits excess power in the precipitation spectrum at large scales (>200 km), particularly during the convectively active evening hours (Fig. 12b). This excess can result from the reassignment of the values of the PM based upon the distribution of values in the individual ensemble members (Ebert 2001), which tend to overpredict heavy precipitation, while retaining the smoothed spatial structure of the ensemble mean. The LPM, in contrast, does not exhibit excess power at large scales, but produces power that is slightly below that of MRMS (Fig. 12b). This could be due to the use of neighborhood regions in the LPM calculations, resulting in less power on scales larger than that of the LPM domain size (180 km × 180 km). These results suggest that the LPM mean generally preserves the power spectra of the individual ensemble members, while overcoming the significant deficiencies of the simple ensemble mean.

Differences in spatial structure among the ensemble mean forecast products can also be seen directly in 2D

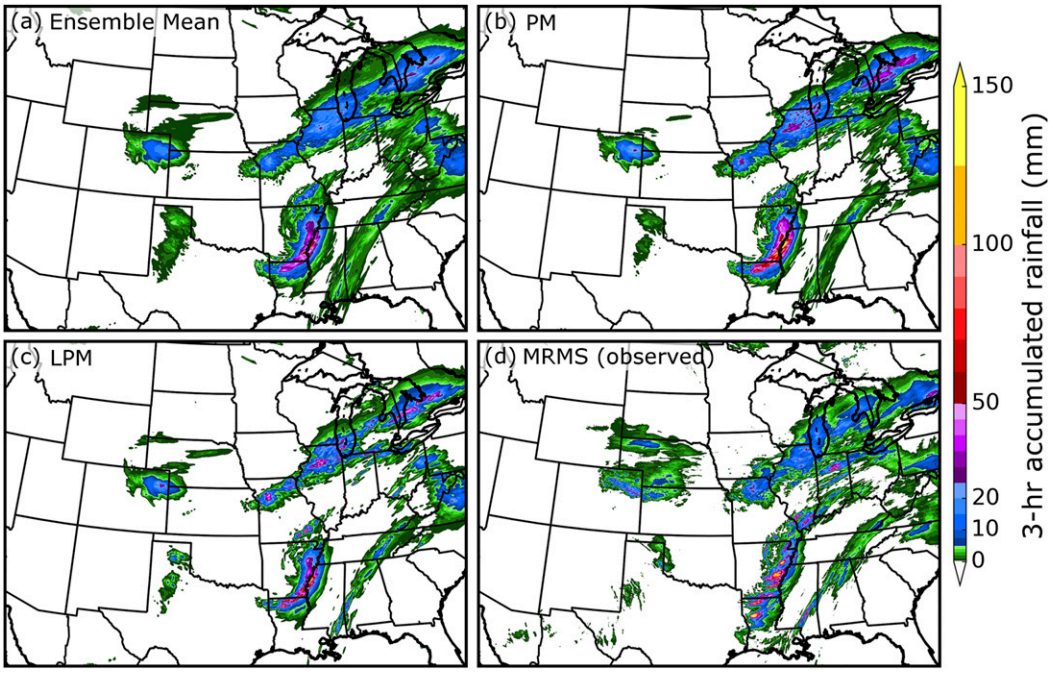

FIG. 13. The 12-h forecasts, valid at 1200 UTC 23 Jun 2017, of 3-h accumulated precipitation. Shown are (a) the simple ensemble mean, (b) the probability-matched mean, and (c) the localized probability-matched mean. Also shown is (d) observed 3-h accumulated precipitation from MRMS, also valid at 1200 UTC 23 Jun 2017. The plotted region corresponds to the subdomain used to calculate verification statistics (including those shown in Figs. 3–11).

spatial QPF fields. In Fig. 13, 12-h forecasts of 3-h accumulated precipitation valid at 1200 UTC 23 June 2017 are plotted for the simple (Fig. 13a), PM (Fig. 13b), and LPM (Fig. 13c) means, along with MRMS estimated precipitation accumulation (Fig. 13d). In general, all of the ensemble mean products (Figs. 13a–c) accurately capture the large-scale precipitation features present in the MRMS observations (Fig. 13d), including heavy rainfall associated with thunderstorms over eastern Arkansas and northern Louisiana, and a broad swath of more moderate precipitation extending from northern Missouri across the Great Lakes. The ensemble mean products do, however, exhibit notable differences in smaller-scale features and in the intensity of precipitation. In the MRMS observations (Fig. 13d), for example, heavy rainfall over Arkansas and Louisiana exhibits local maxima following individual convective storm tracks with maximum 3-h accumulated rainfall values of over 150 mm embedded in a larger region of 20–30-mm accumulations. In the simple mean (Fig. 13a), some smaller-scale features are evident within the precipitation maximum over Arkansas and Louisiana, though they manifest as 40–60-mm accumulations in a larger region of 25–35-mm accumulations—small-scale, local precipitation gradients are much less pronounced than in the observations (Fig. 13d). The PM mean (Fig. 13b) and LPM mean (Fig. 13c) exhibit much stronger local maxima, with rainfall accumulations exceeding 100 mm, which more closely match the maxima of the observed rainfall field. The PM mean (Fig. 13b), however, overestimates the precipitation falling outside of the local maxima, predicting a large swath of precipitation exceeding 30 mm, while the observed rainfall in this region, outside of the local maxima (Fig. 13f) does not typically exceed 25 mm. The LPM mean (Fig. 13c), does not suffer as severely from this overprediction.

Differences in behavior between the LPM mean and the other ensemble mean products are even more pronounced in the narrow band of precipitation extending from near New Orleans, Louisiana, to West Virginia. In the observations (Fig. 13d), this band is quite thin, with maximum rainfall accumulations of around 50 mm over Mississippi. The LPM mean (Fig. 13c) predicts a thin band of precipitation relatively consistent with observations, with maximum accumulations of around 50 mm located in Mississippi. In contrast, the PM mean (Fig. 13b) and simple mean (Fig. 13a) predict a wider swath of precipitation with maximum accumulation of only about 10 mm. This is because in the PM calculation, similar values of lower percentiles among predicted values from the entire model domain of all ensemble members are used to assign to this region of relatively smooth band of precipitation, resulting in a band of

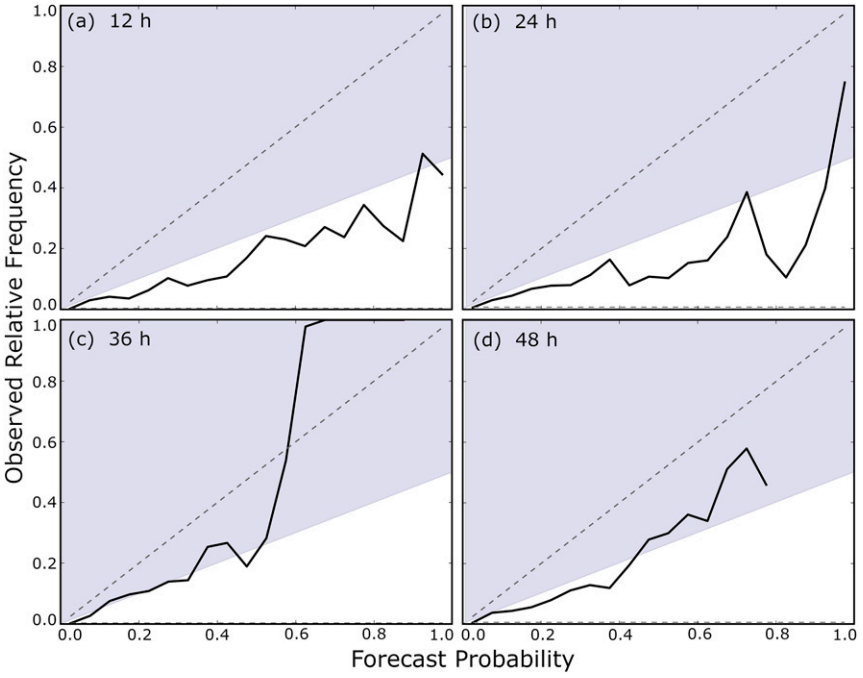

FIG. 14. Reliability diagrams for (a) 12-, (b) 24-, (c) 36-, and (d) 48-h forecasts of the probability of 3-h accumulated precipitation exceeding FFG verified over the verification subdomain (see Fig. 1) for the duration of the 2017 HMT FFaIR operational period.

smooth light precipitation. This is also why the PM mean significantly over predicts power at the larger scales. In contrast, the LPM mean more accurately represents the width and intensity of this precipitation band.

c. *Verification of probabilistic forecasts for precipitation exceeding FFG and RI*

For the 2017 SSEF, forecast products were produced for the first time predicting probability of precipitation exceeding flash flood guidance and probability of precipitation exceeding thresholds associated with recurrence intervals ranging from 5 to 100 years. These products are of operational interest as they take into account geographic differences in rainfall climatology, local infrastructure, and, in the case of FFG, antecedent hydrological conditions. Products were calculated for all ARW SSEF members, and are verified over the verification subdomain (see Fig. 1).

Reliability diagrams are presented for 12-, 24-, 36-, and 48-h forecasts of 3-h accumulated precipitation exceeding FFG in Fig. 14, using data from the 2017 SSEF operational period. In these diagrams, the forecast probability of rainfall exceeding FFG is plotted against the observed frequency of such events (verified against 3-h accumulated MRMS QPE interpolated to the model grid); the dashed diagonal line indicates perfect reliability, while the shaded region indicates a skillful forecast. Overall, the SSEF forecasts are marginally skillful at most hours; the SSEF substantially overpredicts the occurrence of rainfall exceeding FFG (Fig. 14). The anomalous behavior of the 36-h forecasts at high forecast probabilities (Fig. 14c), where precipitation exceeding FFG is actually underpredicted, is an artifact of the small sample size of high forecast probabilities (i.e., probabilities exceeding 0.5).

Precipitation exceeding FFG is quite a rare event; sample climatology for QPE exceeding FFG within 25 km of a location ranges between 0.2% and 0.8% of total grid points. Occurrences of QPE exceeding FFG, and of areas where model QPF exceeds FFG are also often quite localized, many times consisting of just one or two model grid points. To illustrate this, MRMS QPE, regions where QPE exceeds FFG, and probability of SSEF QPF exceeding FFG are plotted for one case—3-h accumulated precipitation valid at 0000 UTC 15 July 2017—in Fig. 15. In this case, isolated convection occurred over the western United States, with heavy rainfall from thunderstorms exceeding FFG over localized regions of Arizona, New Mexico, and Colorado. More widespread convection also occurred over the southeastern United States, with QPF exceeding FFG at a few isolated points (Figs. 15a,b).

Overall, the model captured the general pattern of regions where precipitation exceeded FFG, correctly

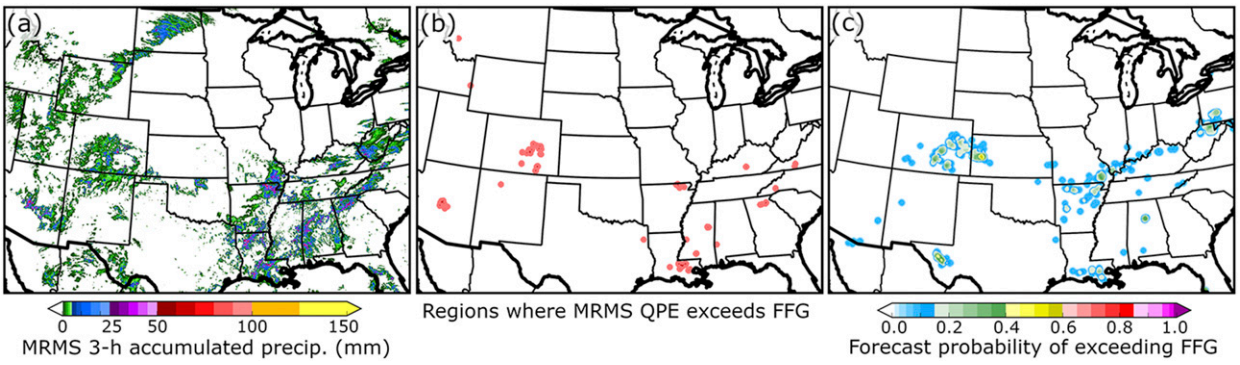

FIG. 15. (a) MRMS 3-h accumulated rainfall for the period of 2100 UTC 14 Jul 2017–0000 UTC 15 Jul 2017. (b) Shaded regions indicate areas where MRMS 3-h accumulated during this period exceeded flash flood guidance within 25 km of that location. (c) Probability of exceeding flash flood guidance within 25 km of a point for the same period calculated from 3-h accumulated rainfall in the 24-h forecasts of the 2017 CAPS SSEF ARW members valid 0000 UTC 15 Jul 2017.

identifying the regions of heavy precipitation over Colorado, and indicating areas of precipitation exceeding flash flood guidance in the southeastern United States (Fig. 15c), although these regions were in many cases somewhat displaced (such as over southern Louisiana, and over eastern Colorado and western Kansas). Maximum QPF from the SSEF was also higher than observed QPE in the areas of heaviest precipitation, resulting in the geographic extent of predicted probabilities of precipitation exceeding FFG being much greater than the extent of areas where QPE actually exceeded FFG. Combined, these two tendencies (overprediction and displacement) result in many false alarms and misses, and contribute to the overall poor objective skill of these forecasts, particularly for measures such as area under the ROC curve (AUC; Mason 1982). Despite the subjective usefulness of the forecasts; AUC for forecasts of the probability of 3-h accumulated precipitation exceeding FFG ranged between 0.00 and 0.03.

For forecasts of rainfall exceeding thresholds associated with 10–100-yr recurrence intervals, overprediction by the SSEF is even more pronounced, again resulting in very low AUC values (generally below 0.02). For 6-h accumulated precipitation exceeding the threshold associated with the 10-yr recurrence interval, the sample climatology in MRMS QPE data ranges between 0.2% and 1.1%. The 10-yr recurrence intervals for 6-h accumulated precipitation vary greatly geographically, and are much lower over the Rocky Mountains and western United States than over the southeastern United States (FFG also exhibits geographic variations, but these variations are less extreme and also take into account antecedent precipitation). The impacts of this geographic variation can be seen for the 15 July 2017 case in Fig. 16; 6-h accumulated precipitation exceeds the 10-yr RI in MRMS QPE at a number of isolated locations over the Rocky Mountains for accumulations of 20–50 mm, while widespread heavier accumulations of up to 100 mm only exceed the 10-yr RI at four isolated points over the southeastern United States (Figs. 16a,b). The SSEF correctly predicts that the 10-yr RI will be exceeded over Colorado, but misses many of the other occurrences of precipitation exceeding the 10-yr RI over the Rockies, while erroneously predicting a wide swath of low probability of RI exceedance over the Ohio valley (Fig. 16c). For these reasons, SSEF predictions of precipitation exceeding FFG were generally more skillful than those of precipitation exceeding RI thresholds.

4. Summary and discussion

During 2016 and 2017, CAPS produced real-time, CONUS-scale storm-scale ensemble forecasts (SSEFs) for the HMT FFaIR experiments, using WRF ARW and NNMB in 2016, and WRF ARW and FV3 in 2017, allowing for generation and verification of new forecast products focused on QPF and flash flood potential. A wide range of precipitation-focused 2D forecast fields were provided for evaluation by HMT participants, including traditional ensemble and probability-matched mean forecasts of precipitation accumulation, probability of exceedance of flash flood guidance, and probability of exceedance of precipitation values corresponding to recurrence intervals ranging from 5 to 100 years. During 2017, a new localized probability matched mean (LPM) algorithm that employs a series of overlapping patches to constrain the probability-matched mean to use only spatially nearby information, was developed, and LPM precipitation forecasts were generated in real-time for the 2017 CAPS HMT ensemble.

Overall, the 2016 and 2017 CAPS SSEFs generally produce skillful forecasts of 3-h accumulated rainfall.

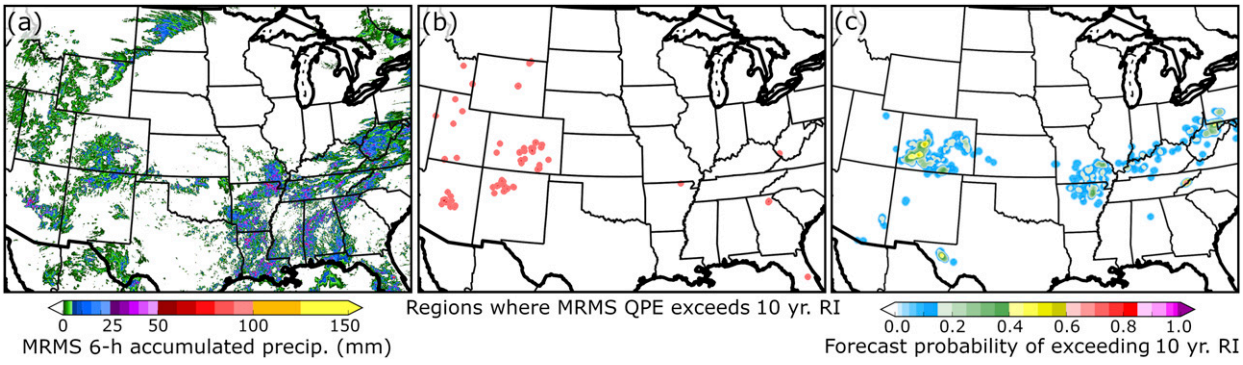

FIG. 16. (a) MRMS 6-h accumulated rainfall during the period of 1800 UTC 14 Jul 2017–0000 UTC 15 Jul 2017. (b) Shaded regions indicate areas where 6-h accumulated MRMS rainfall during this period exceeded the 10-yr recurrence interval for 6-h rainfall accumulation within 25 km of that location. (c) Probability of exceeding the 10-yr recurrence interval for 6-h accumulated rainfall within 25 km of a point for the same period calculated from 6-h accumulated rainfall in the 24-h forecasts of the 2017 CAPS SSEF ARW members valid 0000 UTC 15 Jul. 2017.

PM and LPM mean forecasts exhibit little bias for precipitation exceeding 0.01 in.; at a higher threshold of 0.50 in., the LPM mean exhibits a modest low bias after about 12 h of forecast time, while the regular PM mean exhibits very little bias, particularly in the 2017 ensemble. Skill measured by the equitable threat score (ETS) is positive throughout the forecast period, with ensemble mean products generally outperforming most (if not all) individual ensemble members. We note that for ETS skill scores and related statistical significance testing, we did not apply bias correction before calculating the scores. Without bias correction, larger positive bias may reward the forecast in terms of ETS (Hamill 1999), but bias correction can unfairly reward forecasts with large biases, particularly those with large negative biases (Zhu et al. 2018). Therefore we choose to present ETS scores without bias correction while presenting and discussing frequency biases corresponding to all the ETS scores.

In 2016, the NMMB members generally perform significantly less skillfully than the WRF-ARW members, exhibiting a large high bias in heavy precipitation and significantly lower ETS. In 2017, the single FV3 member generally outperforms most WRF-ARW members, exhibiting higher ETS, and relatively little bias. For a subensemble of 2016 WRF-ARW members differing only in their microphysical scheme, the Morrison member exhibits a significantly higher bias in light precipitation, and the Thompson member exhibits a significantly lower bias in heavy precipitation, though ETS does not differ significantly among these members.

When the full SSEF is broken into subensembles using the same microphysical or PBL scheme, greater consistency between 2016 and 2017 is noted for the subensembles sharing a common PBL scheme than for those sharing a common microphysics scheme. We speculate that this greater year-to-year consistency may be at least in part because the PBL scheme within the model (which is active over the full extent of the domain) has more widespread influence than the microphysical scheme (which is active primarily where clouds and precipitation are occurring), and because the PBL scheme could lead to consistent biases in, for example, near-surface temperature or moisture that would consistently impact the initiation and extent of convection, while sensitivities of model performance to the microphysical scheme would likely depend on factors such as convective mode, distribution and extent of precipitation, and large-scale flow pattern that are more likely to vary from year to year.

The skill of 3-h accumulated rainfall forecasts from the 2017 CAPS HMT ensemble is also evaluated in terms of ability to accurately represent features of varying length scales using fractions skill score (FSS) and variance spectra. In terms of FSS, the forecasts exhibit skill on scales larger than approximately 20–70 km at the 0.01-in. threshold and 30–120 km at the 0.50-in. threshold; the minimum skillful scale is smaller (better) for ensemble mean products and the FV3 member than for WRF-ARW members. The improved performance of the ensemble mean products over individual members is likely due at least in part to reduced overall spatial/displacement error when averaging members. In terms of variance spectra, FV3 exhibits variance much closer to that of MRMS observations at small scales compared to the WRF-ARW members—unlike the MRMS observations or the FV3 member, the WRF-ARW members exhibited a substantial drop in variance at scales less than about 15 km. The variance spectra of individual members are otherwise qualitatively similar to those of MRMS observations, both for early morning

(12 h; 1200 UTC) and evening (24 h; 0000 UTC) forecasts. The better spectral properties of FV3 at small scales are due in large part to the absence of aggressive damping, which was used in the ARW members utilizing the operational configuration of HRRR as prescribed for CLUE.

Compared to individual members and MRMS, the simple ensemble mean exhibits substantial loss of variance at scales less than 400 km, largely due to smoothing of precipitation resulting from averaging spatially disparate members. The variance of the PM mean is similar to that of MRMS at scales between 20 and 150 km, but larger than that of MRMS at larger scales. The variance spectra of LPM mean forecasts best agree with those of MRMS observations, producing similar values at all scales above 20 km, with slight underprediction at the largest scales. The variance spectra of LPM mean forecasts do, however, exhibit a rapid drop off in variance below 20 km, which may be partially due to the Gaussian smoother applied during LPM forecast generation.

Probabilistic forecasts of precipitation exceeding FFG or thresholds associated with recurrence intervals ranging from 10 to 100 years generally overpredict the occurrence of such events; this tendency is particularly pronounced for RI exceedance forecasts. Precipitation exceeding FFG or RI thresholds is a relatively rare event, and generally occurs in isolated regions only a few model grid volumes in size, resulting in displacement errors in the forecasts and overall low objective skill scores, although examining individual cases indicates that the SSEF does exhibit some subjective skill in identifying areas where FFG or RI thresholds are likely to be exceeded. These results suggest that traditional objective skill scores, such as AUC, may not be well suited to measure the subjective utility of forecasts of very rare events. That said, taking into account the rarity of such events and the likely uncertainties and displacement errors for such rare events being predicted many hours or even days in advance, it may be beneficial to apply greater smoothing or increase the neighborhood associated with forecasts of these events in the future.

Producing skillful ensemble-based QPF products remains an active area of research, and further investigation into how to best represent ensemble output is needed. In future CAPS HMT forecast ensembles, we will continue to develop and verify LPM and other products. Possible avenues for improvement to the LPM include the use of differently shaped or flow-dependent patches, and further evaluation of LPM parameters (e.g., size of patches and LPM domains, smoother settings). Work is also ongoing to develop and tune a new spatially aligned mean method that reduces spatial offsets among ensemble members before averaging using methods developed for phase correcting data assimilation (Brewster 2003). This approach could also improve retention of small-scale details in the ensemble mean and improve forecast skill. Finally, we note that it will be important to integrate these products into new and emerging ensemble forecast systems, such as convection-allowing ensembles composed of FV3 members, and to evaluate such forecasts in terms of sensitivity to model physics (Zhang et al. 2019).

Acknowledgments. This work was primarily supported by NOAA Grant NA17OAR4590120 for the Hydrometeorology Testbed Program; supplementary support was provided by NOAA CSTAR Grant NA16NWS4680002. Supercomputing resources used to generate the storm-scale forecast ensemble were provided primarily through XSEDE; systems used include the Stampede, Stampede2, and Lonestar supercomputers at the University of Texas Advanced Computing Center (TACC). Some post-processing was performed on University of Oklahoma Supercomputing Center for Education and Research (OSCER) systems.

REFERENCES

Aligo, E., B. Ferrier, J. Carley, E. Rodgers, M. Pyle, S. J. Weiss, and I. L. Jirak, 2014: Modified microphysics for use in high-resolution NAM forecasts. *27th Conf. on Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., 16A.1, https://ams.confex.com/ams/27SLS/webprogram/Paper255732.html.

Arakawa, A., and W. H. Schubert, 1974: Interaction of a cumulus cloud ensemble with the large-scale environment, Part I. *J. Atmos. Sci.*, **31**, 674–701, https://doi.org/10.1175/1520-0469(1974)031<0674:IOACCE>2.0.CO;2.

Brewster, K. A., 2003: Phase-correction data assimilation and application to storm-scale numerical weather prediction. Part I: Method description and simulation testing. *Mon. Wea. Rev.*, **131**, 480–492, https://doi.org/10.1175/1520-0493(2003)131<0480:PCDAAA>2.0.CO;2.

——, and D. R. Stratman, 2016: Tuning an analysis and incremental analysis updating assimilation for an efficient high-resolution forecast system. *20th Conf. on Integrated Observing and Assimilation Systems for the Atmosphere, Oceans, and Land Surface (IAOS-AOLS),* New Orleans, LA, Amer. Meteor. Soc., 10.6, https://ams.confex.com/ams/96Annual/webprogram/Paper289235.html.

Clark, A. J., 2017: Generation of ensemble mean precipitation forecasts from convection-allowing ensembles. *Wea. Forecasting*, **32**, 1569–1583, https://doi.org/10.1175/WAF-D-16-0199.1.

——, W. A. Gallus Jr., 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, https://doi.org/10.1175/2009WAF2222222.1.

——, and Coauthors, 2011: Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a

convection-allowing ensemble. *Mon. Wea. Rev.*, **139**, 1410–1418, https://doi.org/10.1175/2010MWR3624.1.

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

——, J. Gao, P. T. Marsh, T. Smith, J. S. Kain, J. Correia, M. Xue, and F. Kong, 2013: Tornado pathlength forecasts from 2010 to 2011 using ensemble updraft helicity. *Wea. Forecasting*, **28**, 387–407, https://doi.org/10.1175/WAF-D-12-00038.1.

——, and Coauthors, 2018: The Community Leveraged Unified Ensemble (CLUE) in the 2016 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Bull. Amer. Meteor. Soc.*, **99**, 1433–1448, https://doi.org/10.1175/BAMS-D-16-0309.1.

Denis, B., J. Cote, and R. Laprise, 2002: Spectral decomposition of two-dimensional atmospheric fields on limited-area domains using the discrete cosine transform (DCT). *Mon. Wea. Rev.*, **130**, 1812–1829, https://doi.org/10.1175/1520-0493(2002)130<1812:SDOTDA>2.0.CO;2.

Du, J., G. Dimego, Z. Toth, D. Jovic, B. Zhou, J. Zhu, J. Wang, and H. Juang, 2009: Recent upgrade of NCEP Short-Range Ensemble Forecast (SREF) system. *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., 4A.4, http://ams.confex.com/ams/23WAF19NWP/techprogram/paper_153264.htm.

Duda, J. D., X. Wang, F. Kong, and M. Xue, 2014: Using varied microphysics to account for uncertainty in warm-season QPF in a convection-allowing ensemble. *Mon. Wea. Rev.*, **142**, 2198–2219, https://doi.org/10.1175/MWR-D-13-00297.1.

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, https://doi.org/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

Gagne, D. J., II, A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Wea. Forecasting*, **32**, 1819–1840, https://doi.org/10.1175/WAF-D-17-0010.1.

Gallo, B. T., and Coauthors, 2017: Breaking new ground in severe weather prediction: The 2015 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Wea. Forecasting*, **32**, 1541–1568, https://doi.org/10.1175/WAF-D-16-0178.1.

Gao, J.-D., M. Xue, K. Brewster, and K. K. Droegemeier, 2004: A three-dimensional variational data analysis method with recursive filter for Doppler radars. *J. Atmos. Oceanic Technol.*, **21**, 457–469, https://doi.org/10.1175/1520-0426(2004)021<0457:ATVDAM>2.0.CO;2.

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

Han, J., W. Wang, Y. C. Kwan, S.-Y. Hong, V. Tallapragada, and F. Yang, 2017: Updates in the NCEP GFS cumulus convection schemes with scale and aerosol awareness. *Wea. Forecasting*, **32**, 2005–2017, https://doi.org/10.1175/WAF-D-17-0046.1.

Harris, L. M., and S.-J. Lin, 2013: A two-way nested global-regional dynamical core on the cubed-sphere grid. *Mon. Wea. Rev.*, **141**, 283–306, https://doi.org/10.1175/MWR-D-11-00201.1.

——, and S. Lin, 2014: Global-to-regional nested grid climate simulations in the GFDL high resolution atmospheric model. *J. Climate*, **27**, 4890–4910, https://doi.org/10.1175/JCLI-D-13-00596.1.

Herman, G. R., and R. S. Schumacher, 2016: Extreme precipitation in models: An evaluation. *Wea. Forecasting*, **31**, 1853–1879, https://doi.org/10.1175/WAF-D-16-0093.1.

Hong, S.-Y., and H.-L. Pan, 1996: Nonlocal boundary layer vertical diffusion in a Medium-Range Forecast model. *Mon. Wea. Rev.*, **124**, 2322–2339, https://doi.org/10.1175/1520-0493(1996)124<2322:NBLVDI>2.0.CO;2.

——, Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon. Wea. Rev.*, **134**, 2318–2341, https://doi.org/10.1175/MWR3199.1.

Hu, M., M. Xue, and K. Brewster, 2006: 3DVAR and cloud analysis with WSR-88D Level-II data for the prediction of the Fort Worth, Texas, tornadic thunderstorms. Part I: Cloud analysis and its impact. *Mon. Wea. Rev.*, **134**, 675–698, https://doi.org/10.1175/MWR3092.1.

Iacono, M. J., J. S. Delamere, E. J. Mlawer, M. W. Shephard, S. A. Clough, and W. D. Collins, 2008: Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *J. Geophys. Res.*, **113**, D13103, https://doi.org/10.1029/2008JD009944.

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

——, 2005: A unified model approach from meso to global scales. *Geophys. Research Abstracts*, Vol. 7, 05582, SRef-ID: 1607-7962/gra/EGU05-A-05582, https://www.cosis.net/abstracts/EGU05/05582/EGU05-J-05582.pdf.

Johnson, A., X. Wang, F. Kong, and M. Xue, 2013: Object-based evaluation of the impact of horizontal grid spacing on convection-allowing forecasts. *Mon. Wea. Rev.*, **141**, 3413–3425, https://doi.org/10.1175/MWR-D-13-00027.1.

Kain, J. S., P. R. Janish, S. J. Weiss, M. E. Baldwin, R. S. Schneider, and H. E. Brooks, 2003: Collaboration between forecasters and research scientists at the NSSL and SPC: The Spring Program. *Bull. Amer. Meteor. Soc.*, **84**, 1797–1806, https://doi.org/10.1175/BAMS-84-12-1797.

——, and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931–952, https://doi.org/10.1175/WAF2007106.1.

——, and Coauthors, 2010: Assessing advances in the assimilation of radar data within a collaborative forecasting-research environment. *Wea. Forecasting*, **25**, 1510–1521, https://doi.org/10.1175/2010WAF2222405.1.

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

Kong, F., and Coauthors, 2011: Storm-scale ensemble forecasting for the NOAA Hazardous Weather Testbed. *Sixth European Conf. on Severe Storms*, Palma de Mallorca, Balaric Islands, Spain, ECSS, https://www.essl.org/ECSS/2011/programme/abstracts/171.pdf.

Lin, S.-J., 2004: A vertically Lagrangian finite-volume dynamical core for global models. *Mon. Wea. Rev.*, **132**, 2293–2307, https://doi.org/10.1175/1520-0493(2004)132<2293:AVLFDC>2.0.CO;2.

Loken, E., A. Clark, M. Xue, and F. Kong, 2017: Impact of horizontal resolution on CAM-derived next-day probabilistic severe weather forecasts. *Wea. Forecasting*, **32**, 1403–1421, https://doi.org/10.1175/WAF-D-16-0200.1.

Loken, E. D., A. J. Clark, M. Xue, and F. Kong, 2019: Spread and skill in mixed- and single-physics convection-allowing ensembles. *Wea. Forecasting*, **34**, 305–330, https://doi.org/10.1175/WAF-D-18-0078.1.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.

——, 2003: Binary events. *Forecast Verification—A Practitioner's Guide in Atmospheric Science*, I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, 37–76.

McGovern, A., K. L. Elmore, D. J. Gagne II, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith, and J. K. Williams, 2017: Using artificial intelligence to improve real-time decision making for high-impact weather. *Bull. Amer. Meteor. Soc.*, **98**, 2073–2090, https://doi.org/10.1175/BAMS-D-16-0123.1.

Milbrandt, J. A., and M. K. Yau, 2005: A multimoment bulk microphysics parameterization. Part II: A proposed three-moment closure and scheme description. *J. Atmos. Sci.*, **62**, 3065–3081, https://doi.org/10.1175/JAS3535.1.

Morrison, H., and J. A. Milbrandt, 2015: Parameterization of cloud microphysics based on the prediction of bulk ice particle properties. Part I: Scheme description and idealized tests. *J. Atmos. Sci.*, **72**, 287–311, https://doi.org/10.1175/JAS-D-14-0065.1.

——, G. Thompson, and V. Tatarskii, 2009: Impact of cloud microphysics on the development of trailing stratiform precipitation in a simulated squall line: Comparison of one- and two-moment schemes. *Mon. Wea. Rev.*, **137**, 991–1007, https://doi.org/10.1175/2008MWR2556.1.

Nakanishi, M., and H. Niino, 2009: Development of an improved turbulence closure model for the atmospheric boundary layer. *J. Meteor. Soc. Japan*, **87**, 895–912, https://doi.org/10.2151/jmsj.87.895.

Potvin, C. K., J. R. Carley, A. J. Clark, L. J. Wicker, P. S. Skinner, A. E. Reinhart, and J. S. Kain, 2018: Inter-model storm-scale comparisons from the 2017 HWT spring forecasting experiment. *25th Conf. on Numerical Weather Prediction*, Denver, CO, Amer. Meteor. Soc., 14B.1, https://ams.confex.com/ams/29WAF25NWP/webprogram/Paper345657.html.

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, https://doi.org/10.1175/2007MWR2123.1.

Schmidt, J., A. Anderson, and J. Paul, 2007: Spatially-variable, physically-derived, flash flood guidance. *21st Conf. on Hydrology*, San Antonio, TX, Amer. Meteor. Soc., 6B.2, https://ams.confex.com/ams/87ANNUAL/techprogram/paper_120022.htm.

Schwartz, C. S., and R. A. Sobash, 2017: Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: A review and recommendations. *Mon. Wea. Rev.*, **145**, 3397–3418, https://doi.org/10.1175/MWR-D-16-0400.1.

——, and Coauthors, 2009: Next-day convection-allowing WRF Model guidance: A second look at 2-km versus 4-km grid spacing. *Mon. Wea. Rev.*, **137**, 3351–3372, https://doi.org/10.1175/2009MWR2924.1.

——, and Coauthors, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263–280, https://doi.org/10.1175/2009WAF2222267.1.

——, 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, https://doi.org/10.1175/WAF-D-13-00145.1.

Skamarock, W. C., 2004: Evaluating mesoscale NWP models using kinetic energy spectra. *Mon. Wea. Rev.*, **132**, 3019–3032, https://doi.org/10.1175/MWR2830.1.

——, J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, 2005: A description of the Advanced Research WRF version 2. NCAR Tech. Note NCAR/TN-468+STR, 88 pp., https://doi.org/10.5065/D6DZ069T.

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, https://doi.org/10.1175/MWR-D-13-00134.1.

Sweeney, T., and T. Baumgardner, 1999: Modernized flash flood guidance. Rep. to NWS Hydrology Laboratory, 11 pp., http://www.nws.noaa.gov/oh/hrl/ffg/modflash.htm.

Tewari, M., and Coauthors, 2004: Implementation and verification of the unified Noah land surface model in the WRF model. *20th Conf. on Weather Analysis and Forecasting/16th Conf. on Numerical Weather Prediction*, Amer. Meteor. Soc., Seattle, WA, 14.2a, https://ams.confex.com/ams/84Annual/techprogram/paper_69061.htm .

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, https://doi.org/10.1175/2008MWR2387.1.

Weiss, S. J., and Coauthors, 2015: Progress over the last decade in the development and use of convection-allowing models in operational severe weather prediction. *27th Conf. on Weather Analysis and Forecasting/23rd Conf. on Numerical Weather Prediction*, Chicago, IL, Amer. Meteor. Soc., 7B.4, https://ams.confex.com/ams/27WAF23NWP/webprogram/Paper273802.html.

WPC, 2016: 2016 flash flood and intense rainfall experiment final report. NOAA, 62 pp., accessed 14 August 2017, http://www.wpc.ncep.noaa.gov/hmt/2016_FFaIR_Final_Report.pdf.

Xiang, B., M. Zhao, X. Jiang, S.-J. Lin, T. Li, X. Fu, and G. Vecchi, 2015: The 3–4 week MJO prediction skill in a GFDL coupled model. *J. Climate*, **28**, 5351–5364, https://doi.org/10.1175/JCLI-D-15-0102.1.

Xue, M., 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, https://doi.org/10.1007/s00703-001-0595-6.

——, and Coauthors, 2007: CAPS realtime storm-scale ensemble and high-resolution forecasts as part of the NOAA Hazardous Weather Testbed 2007 spring experiment. *22nd Conf. on Weather Analysis and Forecasting/18th Conf. on Numerical Weather Prediction*, Park City, UT, Amer. Meteor. Soc., 3B.1, https://ams.confex.com/ams/22WAF18NWP/techprogram/paper_124587.htm.

——, and Coauthors, 2009: CAPS realtime 4-km multi-model convection-allowing ensemble and 1-km convection-resolving forecasts for the NOAA Hazardous Weather Testbed 2009 Spring Experiment. *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., 16A.2, https://ams.confex.com/ams/23WAF19NWP/techprogram/paper_154323.htm.

——, and Coauthors, 2010: CAPS realtime storm-scale ensemble and convection-resolving high-resolution forecasts for the NOAA Hazardous Weather Testbed 2010 spring experiment. *25th Conf. on Severe Local Storms*, Denver, CO, Amer. Meteor. Soc., 7B.3.

——, and Coauthors, 2011: Realtime convection-permitting ensemble and convection-resolving deterministic forecasts of CAPS for the Hazardous Weather Testbed 2010 Spring Experiment. *24th Conf. on Weather and Forecasting/20th Conf. on Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc., 9A.2, https://ams.confex.com/ams/91Annual/webprogram/Paper183227.html.

——, F. Kong, K. W. Thomas, J. Gao, Y. Wang, K. Brewster, and K. K. Droegemeier, 2013: Prediction of convective storms at convection-resolving 1-km resolution over continental United States with radar data assimilation: An example case of 26 May 2008 and precipitation forecasts from spring 2009. *Adv. Meteor.*, **2013**, 259052, https://doi.org/10.1155/2013/259052.

Zhang, C., and Coauthors, 2019: How well does the FV3-based model predict precipitation at a convection-allowing resolution? Results from CAPS forecasts for the 2018 NOAA Hazardous Weather Test bed with different physics combinations. *Geophys. Res. Lett.*, **46**, 3523–3531, https://doi.org/10.1029/2018GL081702.

Zhang, J., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 621–638, https://doi.org/10.1175/BAMS-D-14-00174.1.

Zhou, L., S. Lin, J. Chen, L. M. Harris, X. Chen, and S. L. Rees, 2019: Toward convective-scale prediction within the next generation global prediction system. *Bull. Amer. Meteor. Soc.*, https://doi.org/10.1175/BAMS-D-17-0246.1, in press.

Zhu, K., and Coauthors, 2018: Evaluation of real-time convection-permitting precipitation forecasts in China during the 2013–2014 summer season. *J. Geophys. Res. Atmos.*, **123**, 1037–1064, https://doi.org/10.1002/2017JD027445.