Evaluation of an Experimental 15-Member FV3-LAM Ensemble for the Summer 2024 Flash Flood and Intense Rainfall Experiment

ANDREW H. BERRINGTON,[a,b] KEITH A. BREWSTER,[a,b] NATHAN SNOOK,[a,b] MING XUE,[a,b] JUN PARK,[c] CHENGSI LIU,[a] SARAH TROJNIAK,[d,e] AND JAMES CORREIA[d,e]

[a] *Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma*
[b] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*
[c] *National Center for Atmospheric Research, Boulder, Colorado*
[d] *Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado*
[e] *NOAA/Weather Prediction Center, College Park, Maryland*

(Manuscript received 10 January 2026, in final form 29 June 2026, accepted 16 July 2026)

ABSTRACT: A 15-member experimental ensemble using the Finite-Volume Cubed-Sphere Limited Area Model (FV3-LAM) dynamic core is evaluated for precipitation forecasting during summer 2024. The ensemble was run once daily out to an 84-h lead, for a total of 30 days, in support of the National Oceanic and Atmospheric Administration's (NOAA) Hydrometeorological Testbed's Flash Flood and Intense Rainfall Experiment. It employs a variety of physics combinations, boundary conditions, radar data assimilation, and cumulus convective schemes. Evaluation via contingency table metrics suggests that activating cumulus convective schemes, in conjunction with Global Ensemble Forecast System (GEFS) boundary conditions, may have a detrimental effect on precipitation forecasts beyond day 1. Additionally, greater biases are noted in members using the Noah-MP land surface scheme than in others. Radar reflectivity data assimilation improves ensemble performance on day 1, while a realignment of features to a common position among members (spatial aligned mean) improves the performance of ensemble consensus products. An experimental Model for Prediction Across Scales (MPAS) member run alongside the FV3 ensemble shows competitive performance out to day 2, with slightly worse performance during days 3 and 4. The results helped inform ongoing Rapid Refresh Forecasting System (RRFS) development during and after the experiment. Recommendations based on these results are subsequently made for future model development and forecast experiments.

SIGNIFICANCE STATEMENT: Short-range prediction of heavy precipitation is important for reducing impacts from flash flooding. In this study, we evaluate an experimental model ensemble run during summer 2024 to assess its ability to predict both light and heavier precipitation. We find through some metrics that activating a cumulus convective scheme may have detrimental effects on the skill of the forecast on days 2–4 and that experimental radar data assimilation benefits forecasts on day 1. The results from this study help inform future model ensemble design and experiments focusing on precipitation forecasting.

KEYWORDS: Flood events; Precipitation; Ensembles; Forecast verification/skill; Numerical weather prediction/forecasting; Short-range prediction

1. Introduction

Flooding and excessive precipitation represent a significant threat to life and property in the United States and globally. Ashley and Ashley (2008) documented over 4500 fatalities in the United States related to flooding over the 47-yr period from 1959 to 2005, while Terti et al. (2017) documented 1079 flash flood fatalities between 1996 and 2014. Meanwhile, Smith and Katz (2013) determined that 12% of the billion dollar weather and climate disasters in the United States, evaluated in their study, were related to nontropical flooding, along with a nonnegligible percentage related to freshwater flooding from tropical cyclones. From a meteorological standpoint, the work of Maddox et al. (1979) and Doswell et al. (1996) documented some of the common conditions present on both the synoptic scale and the mesoscale during flash flooding events in the United States. The impacts from excessive precipitation and flooding in the United States prompt additional work into understanding the mechanisms leading to and improving forecasts of such events, particularly in numerical weather prediction (NWP) models. Challenges related to flash flood forecasting often result from the rare and generally localized nature of precipitation that results in flash flooding, usually defined over a 6-h time window (Hitchens et al. 2013; Barthold et al. 2015).

One ongoing effort to improve precipitation forecasts is supported by the National Oceanic and Atmospheric Administration's (NOAA) Hydrometeorological Testbed (HMT), run by the NOAA Weather Prediction Center (WPC) during the winter and summer seasons (Nelson et al. 2022). During the winter, this testbed conducts the annual Winter Weather Experiment (WWE) which focuses on winter precipitation hazards (Supinie et al. 2022), while the Flash Flood and Intense Rainfall (FFaIR) Experiment, which focuses on excessive

Corresponding author: Andrew H. Berrington, andrew.berrington@ou.edu

DOI: 10.1175/WAF-D-25-0256.1

© 2026 American Meteorological Society. This published article is licensed under the terms of the default AMS reuse license. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:44 PM UTC

precipitation (Barthold et al. 2015), is run during the summer. FFaIR is designed as a collaboration between multidisciplinary (e.g., hydrology, meteorology) centers to generate and evaluate—both subjectively through forecaster opinion and objectively through verification metrics—forecasts of excessive precipitation that may result in flash flooding (Barthold et al. 2015).

Since 2016, the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma has been a collaborator and participant in the HMT FFaIR, running experimental NWP ensembles to facilitate real-time forecasts of excessive precipitation and its drivers, in an effort to determine whether and how an ensemble of convection-allowing numerical models can be used to improve such forecasts. In recent years, the CAPS FFaIR ensembles have been designed in support of the development of the Rapid Refresh Forecasting System (RRFS) under the Unified Forecast System's framework that will replace the current operational convection-allowing models (CAMs). The RRFS, version 1, has been built around the Finite-Volume Cubed-Sphere (FV3) dynamic core (Lin 2004) and has been run with a variety of different physics combinations available through the Common Community Physics Package (CCPP) described in Heinzeller et al. (2023).

A series of studies, primarily centered around the verification of these experimental ensemble forecasts, have resulted from CAPS' collaborations with HMT and the NOAA Hazardous Weather Testbed (HWT). Zhang et al. (2019) determined during the 2018 HWT that the forecast skill for precipitation in the FV3 was comparable to that of the Weather Research and Forecasting Model and that members using Thompson microphysics (Thompson et al. 2004) slightly outperformed those using the NSSL microphysics (Mansell et al. 2010). There were generally negligible skill differences due to boundary layer scheme variance between the members. Supinie et al. (2022) performed regional verification of an FV3 ensemble run by CAPS during the 11th HMT WWE, primarily focused on the differences between combinations of model member physics. They found biases concerning snowfall in certain regions of the United States and recommended that both Thompson and NSSL microphysics (Thompson et al. 2004; Mansell et al. 2010), the Noah and Noah-MP (Ek et al. 2003; Niu et al. 2011) land surface models, and the Mellor–Yamada–Nakanishi–Niino eddy-diffusivity mass-flux (MYNN-EDMF; shortened to simply MYNN) boundary layer scheme (Nakanishi and Niino 2009; Angevine et al. 2020) be used in the operational implementation of the RRFS. The turbulent kinetic energy–based EDMF (TKE-EDMF) boundary layer scheme (Han and Bretherton 2019) was recommended for improvement.

Meanwhile, Snook et al. (2019, 2020) used techniques such as the probability-matched (PM) ensemble mean (Ebert 2001; Ma et al. 2018) and the localized PM (LPM) ensemble mean (Clark 2017) in attempts to improve the skill of ensemble consensus precipitation products. They determined that the LPM had notable benefits for precipitation forecasts and that a patchwise (Snook et al. 2019) LPM improved the computational performance for that method considerably without degrading forecast skill. A later study by Lee et al. (2024) demonstrated that a spatially aligned mean (SAM) technique can improve ensemble mean forecasts by reducing the dilutive effect on the ensemble mean attributable to small differences in the positions of precipitation features among ensemble members.

Recently, Park et al. (2023), Kong et al. (2024), and Park et al. (2025) described and evaluated experimental radar reflectivity data assimilation developed under the Joint Effort for Data Assimilation Integration (JEDI) framework. This radar reflectivity data assimilation (DA) is included in some members of the FV3 ensemble developed for the 2024 FFaIR experiment and showed notable skill improvements over nonradar-DA members in short-range forecasts out to approximately 12 h (Park et al. 2025). Johnson et al. (2025) assessed the FV3 ensemble's capabilities in predicting convective hazards during the 2022 HWT and found that updraft helicity was a more reliable predictor of severe weather, while modeled 10-m winds showed more limited prediction skill for severe wind events. Additionally, members utilizing the NSSL microphysics (Mansell et al. 2010), MYNN boundary layer (Nakanishi and Niino 2009; Angevine et al. 2020), and RUC land surface (Smirnova et al. 2016) parameterizations were more skillful with a higher frequency bias than members with Thompson microphysics, TKE-EDMF boundary layer, and Noah-MP land surface parameterizations. Finally, Snook et al. (2025) evaluated the 21-member ensemble run during the 2022 HWT, which also used the FV3 dynamic core, and found that many members exhibited underprediction for lower rainfall rates and overprediction at higher rainfall rates. Stochastic perturbations added to the ensemble improved forecasts of 2-m temperature, dewpoint, and 10-m wind, although there was less of an effect on precipitation. They also recommended further tests using the Noah-MP land surface model (Niu et al. 2011) given its tendency to yield larger frequency biases.

This study uses similar verification techniques as several previous CAPS-led studies (Snook et al. 2020; Supinie et al. 2022; Johnson et al. 2025; Snook et al. 2025) to evaluate precipitation in the CAPS ensemble produced for FFaIR 2024. Additional model adjustments added in the 2024 ensemble include the employment of cumulus convective schemes (Han et al. 2017; Freitas et al. 2021) in some members and the experimental radar data assimilation mentioned above (Park et al. 2023, 2025). The 2024 ensemble was constructed to evaluate the skill of RRFS Ensemble Forecasting System (REFS)-like configurations (Blake et al. 2025; Clark et al. 2026) and also to provide training data for machine learning applications. Stochastic physics parameter perturbations (SPPs) that were employed in previous studies (Jankov et al. 2017, 2019; Zhang et al. 2019; Snook et al. 2025) are used in certain physics schemes to slightly increase ensemble spread. Both light and higher-intensity precipitation are evaluated in the verification. Verification includes both deterministic and probabilistic (ensemble based) scores, and both pointwise and object-based verification techniques are used. While not the primary focus of this study, other ensemble consensus products [such as those of Snook et al. (2019, 2020), and Lee et al. (2024)] are peripherally discussed. Finally, as in Supinie et al. (2022) and Snook et al. (2025), recommendations based on these results are made for future high-resolution model development in subsequent testbeds.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:44 PM UTC

TABLE 1. CAPS ensemble configurations for FFaIR 2024 via Park et al. (2025). MP represents the microphysics option, PBL represents the planetary boundary layer option, SFC represents the surface layer scheme option, LSM represents the land surface model option, CU represents the cumulus scheme option, and IC/LBC represents the initial and boundary condition options. An * indicates stochastic parameter perturbations were used, while a # indicates fixed parameter perturbations were used.

Name	MP	PBL	SFC	LSM	CU	IC/lBC
CAPS01	Thompson	MYNN	MYNN	Noah	—	GFS/GFS
CAPS02	NSSL	MYNN	MYNN	Noah	—	GFS/GFS
CAPS03	Thompson	TKE-EDMF	GFS	Noah MP	—	GFS/GFS
CAPS04	Thompson	TKE-EDMF	GFS	RUC	—	GFS/GFS
CAPS05	Thompson	MYNN	MYNN	RUC	G-F deep	LETKF mean/GFS
CAPS06	Thompson*	MYNN*	MYNN*	Noah MP	G-F deep*	LETKF m03/GEFS m03
CAPS07	Thompson*	MYNN*	MYNN*	Noah MP	saSAS deep	LETKF m05/GEFS m05
CAPS08	Thompson*	MYNN*	MYNN*	RUC*	saSAS deep	LETKF m02/GEFS m02
CAPS09	Thompson*	TKE-EDMF	GFS	Noah MP	G-F dp* + sh	LETKF m04/GEFS m04
CAPS10	Thompson*	TKE-EDMF	GFS	RUC*	G-F dp* + sh	LETKF m01/GEFS m01
CAPS11	NSSL#	MYNN*	MYNN*	Noah MP	G-F deep*	LETKF m08/GEFS m08
CAPS12	NSSL#	MYNN*	MYNN*	Noah MP	saSAS deep	LETKF m10/GEFS m10
CAPS13	NSSL#	MYNN*	MYNN*	RUC*	saSAS deep	LETKF m07/GEFS m07
CAPS14	NSSL#	TKE-EDMF	GFS	Noah MP	G-F dp* + sh	LETKF m09/GEFS m09
CAPS15	NSSL#	TKE-EDMF	GFS	RUC*	G-F dp* + sh	LETKF m06/GEFS m06
MPAS01	Thompson	MYNN	MYNN	Noah	—	GFS/GFS

2. Data and methods

a. *Ensemble forecast data*

For the summer 2024 FFaIR, a 15-member ensemble centered over the contiguous United States (CONUS) was initialized at 0000 UTC for a total of 30 days during June (10–14, 18–20, and 24–28), July (8–12, 22–26, and 29–31), and August (1–2 and 18–19) 2024 using the FV3 Limited Area Model (FV3-LAM) dynamic core described in Black et al. (2021). Each ensemble member was run at a convection-allowing horizontal grid spacing of 3 km using the High-Resolution Ensemble Forecast's (HREF) CONUS grid (Roberts et al. 2019; Supinie et al. 2022; Johnson et al. 2025), with 64 vertical levels. Output was produced hourly out to 84 h for each forecast member, yielding a total sample size of $84 \times 15 \times 30 = 37\,800$ h of total forecasts. The variability in the physics within the ensemble was provided by different combinations of Thompson or NSSL microphysics (Thompson et al. 2004; Mansell et al. 2010); Noah, Noah-MP, or RUC land surface models (Ek et al. 2003; Niu et al. 2011; Smirnova et al. 2016); MYNN or TKE-EDMF boundary layer parameterizations (Nakanishi and Niino 2009; Han and Bretherton 2019); and Grell–Freitas (G-F) or scale-aware Arakawa–Schubert (saSAS) cumulus convection schemes (Han et al. 2017; Freitas et al. 2021). Some of these physics configurations mimic those found in the winter evaluation of Supinie et al. (2022), and all are available through the CCPP. Members with the Thompson/MYNN/MYNN/RUC combination (corresponding to the microphysics, boundary layer, surface scheme, and land surface model, respectively) contain the same physics as the operational High-Resolution Rapid Refresh (HRRR). The member configurations are detailed in Table 1. Here, G-F deep is the Grell–Freitas scheme with only deep convective treatments, while G-F dp + sh includes shallow cumulus treatments to account for the lack of such treatments in the TKE-EDMF boundary layer scheme versus MYNN (Angevine et al. 2020).

Initial and lateral boundary conditions (ICs and LBCs) were provided by the National Centers for Environmental Prediction's (NCEP) Global Forecast System (GFS), version 15, at hourly intervals or the Global Ensemble Forecast System (GEFS), version 12 (Zhou et al. 2022), members at 3-hourly intervals. CAPS01–04 were run from the 0000 UTC GFS analyses as initial conditions without supplemental data assimilation, and the subgrid cumulus convection schemes were turned off. CAPS05–15 members used experimental JEDI radar data assimilation (Park et al. 2023) and had the convective scheme turned on. When comparisons are made between these two sets of members, they will be referred to as "DA-CU OFF" and "DA-CU ON," respectively. CAPS06–15 also had SPP turned on for certain physics options (Jankov et al. 2017, 2019), as noted in Table 1. As with the perturbations in Snook et al. (2025), SPP options have a characteristic length scale of 150 km and a temporal scale of 6 h. In addition, foreshadowing possible future RRFS development (Carley et al. 2023), an experimental member (MPAS01 in Table 1) using the Model for Prediction Across Scales (MPAS) dynamic core (Skamarock et al. 2012) was run separately from the main ensemble. This experimental member was run over a similar 3-km-resolution CONUS domain, with 55 vertical levels, and used the same physics configuration as CAPS01. GFS ICs and LBCs were used for this member, and it was run without supplemental data assimilation (DA-CU OFF).

To complement the individual ensemble members, a variety of ensemble consensus products were also produced for each of the 84-h forecasts. These consist of the standard ensemble mean, the PM ensemble mean (Ebert 2001; Ma et al. 2018; Snook et al. 2025), the LPM ensemble mean (Clark 2017; Snook et al. 2020, 2025), and 39-km neighborhood probabilities of exceeding thresholds (Clark et al. 2010). For the different ensemble means, the spatial-alignment method described in Lee et al. (2024) is also applied to attempt to reduce the effects of spatial errors resulting from phase/timing differences

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:44 PM UTC

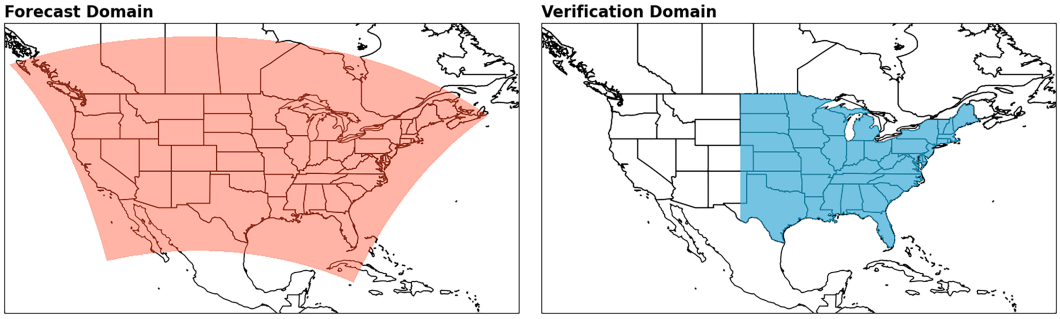


FIG. 1. Comparison of (left) forecast (red shading) and (right) verification domains (blue shading) for this study.

in precipitation features (e.g., convection areas and cells) among the ensemble members. Referred to as the SAM, the SAM-PM, and the SAM-LPM, these add three separate ensemble means for analysis. The SAM method has been shown to improve the spatial structure of precipitation features and modestly improve the verification scores of the ensemble means compared to their standard counterparts (Lee et al. 2024).

b. *Forecast evaluation metrics*

The primary product being evaluated in subsequent tests is quantitative precipitation forecasts (QPFs) for 6-hourly accumulations. While Supinie et al. (2022) primarily examined 24-hourly accumulations, 6-hourly accumulations can provide additional insight such as the fidelity of the diurnal cycle of precipitation and shorter-duration/high-intensity precipitation events implicated in flash flooding. The 6-hourly verification statistics can also be aggregated over daily periods. Verifying observations were provided by NOAA Stage IV (Nelson et al. 2016) quantitative precipitation estimates, and forecast data were regridded using the nearest-neighbor approach to the Stage IV 4-km CONUS grid. A mask was applied to remove areas over the ocean that are poorly observed as in Supinie et al. (2022), as well as areas of Canada and Mexico, restricting verification to within the CONUS land area where the Stage IV data are valid. All regridding and masking were handled by Model Evaluation Tools (MET), version 11.0 (Brown et al. 2021), and the package's Grid-Stat tool was used to pair observations with forecasts so that verification metrics can be applied accordingly. Due to known issues with Stage IV data over the western CONUS such as the lack of hourly data from the Northwest River Forecast Center and faulty gauge data (Nelson et al. 2016), the verification domain was further restricted to include only the portion of the CONUS land area east of 104°W longitude. This also removes large areas where radar coverage (another source of Stage IV observations) is complicated by complex terrain. A graphic showing the difference between the forecast and verification domains is given in Fig. 1.

Two-dimensional contingency tables are calculated for both the 95th (95P) and 99th (99P) percentile 6-hourly precipitation, differing from some previous work (e.g., Supinie et al. 2022) that used uniform thresholds such as 1 or 25 mm. The choice to use percentile thresholds based on the ensemble member instead of uniform thresholds was primarily motivated by reducing the frequency bias that may be introduced when using a uniform threshold across many different ensemble members, especially for higher-end amounts (Snook et al. 2025). However, since performance diagrams (Roebber 2009) for deterministic forecasts show frequency bias as one of the metrics, uniform thresholds of 3 and 15 mm were used for this metric alone, which roughly correspond to the observed 93–94P and 98–99P precipitation in Fig. 2 (Snook et al. 2025). The 95P precipitation generally represents light-to-moderate 6-h precipitation accumulations associated with stratiform cloud cover or weaker convection, while the 99P represents heavier precipitation accumulations often resulting from deep convection. To give an idea of the sample size for both percentiles, for each member, when 6-hourly precipitation is aggregated into daily windows for Stage IV, a total of 1 892 217 values on day 1; 1 779 026 values on day 2; 1 670 932 values on day 3; and 716 900 values on day 4 exceed the 95P. Meanwhile, a total of 384 244 values on day 1; 347 053 values on day 2; 317 550 values on day 3; and 145 226 values on day 4 exceed the 99P. The lower amounts for day 4 are due to there being only 12 h of forecasts during day 4. In 6-h windows, as would be expected in accordance with the diurnal cycle, precipitation exceeds the thresholds most often during the 1800–0000 and 0000–0600 UTC time frames. The forecasts generally follow a similar pattern, albeit data assimilation results in increased exceedances during the first 6–12 h of forecasts in DA-CU ON members (not shown).

Contingency table components—hits, misses, false alarms, and true negatives—are used to calculate the equitable threat score (ETS), success ratio [SR; also known as the false alarm ratio (FAR) subtracted from 1], and the probability of detection (POD). The ETS (Clark et al. 2010; Joliffe and Stephenson 2012) measures the skill of a forecast relative to chance and rewards hits while not distinguishing the source of forecast errors and is commonly used in verifying precipitation forecasts (Clark et al. 2010; Baxter et al. 2014; Iyer et al. 2016; Supinie et al. 2022). Meanwhile, the SR and POD are plotted against each other to generate performance diagrams (Roebber 2009) that visualize four separate measures of forecast quality: the SR, POD, critical success index (CSI), and frequency bias. Performance diagrams can either be plotted for individual deterministic forecasts or using a series of probabilities—a set of 10 bins from 0% to 100% in this study—for ensemble forecasts. The

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:44 PM UTC

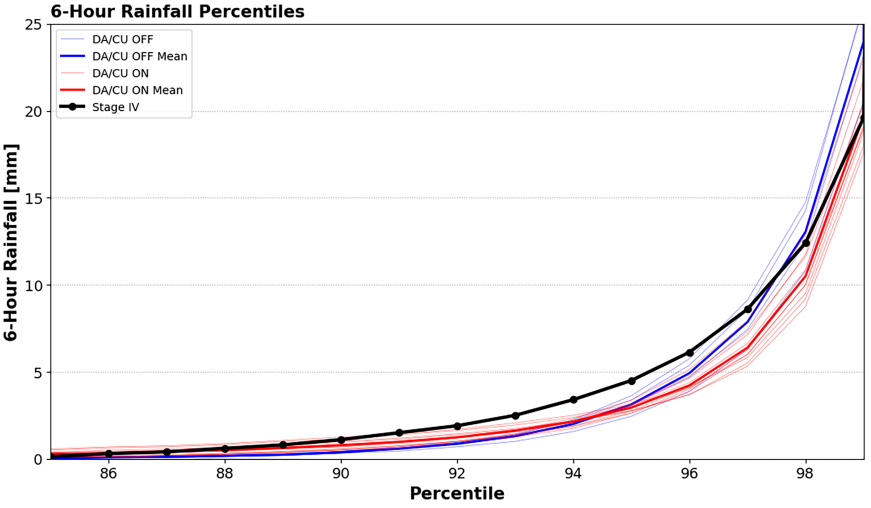


FIG. 2. Percentiles of 6-h precipitation (mm) from the CAPS ensemble and Stage IV observations over all forecasts and the verification domain. Blue curves represent the DA-CU OFF members, red curves represent the DA-CU ON members, and the black dotted curve represents observed percentiles.

latter method enables neighborhood verification (Clark et al. 2010; Dahl and Xue 2016; Schwartz and Sobash 2017; Ma et al. 2018) techniques such as the neighborhood maximum ensemble probability (NMEP) that can help reduce the "double penalty" where spatial and temporal displacements can lead to decreased hits and increased false alarms (Schwartz and Sobash 2017; Supinie et al. 2022; Johnson et al. 2025). For this study, a neighborhood search radius of 39 km (13×3 km) was used to correspond to the 40-km radius used in WPC's Excessive Rainfall Outlooks (EROs) for determining areas likely to see precipitation exceeding flash flood guidance (Burke et al. 2023). We did not apply additional smoothing to the NMEP fields, similar to the unsmoothed NMEP presented in Schwartz and Sobash (2017).

In addition to the ETS and performance diagrams, the fractions skill score (FSS) was used for object-based spatial-scale verification (Roberts and Lean 2008; Johnson et al. 2025). The FSS determines the minimum skillful spatial scale (MSS) at which the forecasts of the relevant variable resemble those of the observations and is particularly useful for rare events, such as 99P precipitation in this study. Calculating the useful FSS for percentile x (e.g., $x = 0.95$ for 95P) follows the simple calculation of $\mathrm{FSS}_{\mathrm{useful}} = 0.5 + f_o/2$, where $f_o = 1 - x$. Ensemble-based techniques are employed to assess the performance of the CAPS ensemble as a whole as opposed to deterministically evaluating individual members. The area under the relative operating characteristic (ROC) curve (AUC; Mason 1982; Joliffe and Stephenson 2012) was used to gauge the ensemble's ability to distinguish between events and nonevents and is conditioned on the observations, with a score of 1 being perfect and a score of 0.5 indicating no skill. Its counterpart, the reliability diagram (Hamill 1997), represents whether the predicted probability of an event (e.g., exceeding the 95P threshold for precipitation) is proportional to the event's true frequency. For this metric, as probabilities increase, an ideal ensemble will closely follow the 1:1 line, which represents perfect reliability. Forecast curves to the left of the 1:1 line indicate underforecasting for the given observed frequency (i.e., the forecast is underconfident), while curves to the right indicate overforecasting (i.e., the forecast probability is higher than the observed frequency). For both metrics, neighborhood, as well as pointwise, verification was employed to assess whether the former improves skill, particularly for rare events (Hitchens et al. 2013). Significance testing, when applicable, was provided by bootstrapping (using 10 000 samples with replacement).

Finally, while assessing the CAPS ensemble in isolation is useful, it is helpful to have a reference forecast for comparison. In this case, the reference forecast is provided by the HREF, which consists of the 0000 UTC (matching the initialization of the CAPS ensemble) and 1200 UTC initializations from the previous day and consists of eight members overall, with the FV3-based members being omitted. Since the HREF is only run for 48 h and the prior 1200 UTC initializations are used in a time-lagged ensemble, only the first 36 h (short range) of the CAPS forecasts will be compared to the HREF in subsequent verification metrics.

3. CAPS ensemble verification

a. Individual member evaluation

Since precipitation percentiles are used as exceedance thresholds as opposed to raw precipitation totals, it follows that the distribution of precipitation across the different ensemble members should be examined by percentile. These distributions were calculated across all forecast hours for the verification domain. Shown in Fig. 2, the 6-h precipitation percentiles are color coded according to whether they belong to the DA-CU OFF or DA-CU ON members, as these were the

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:44 PM UTC

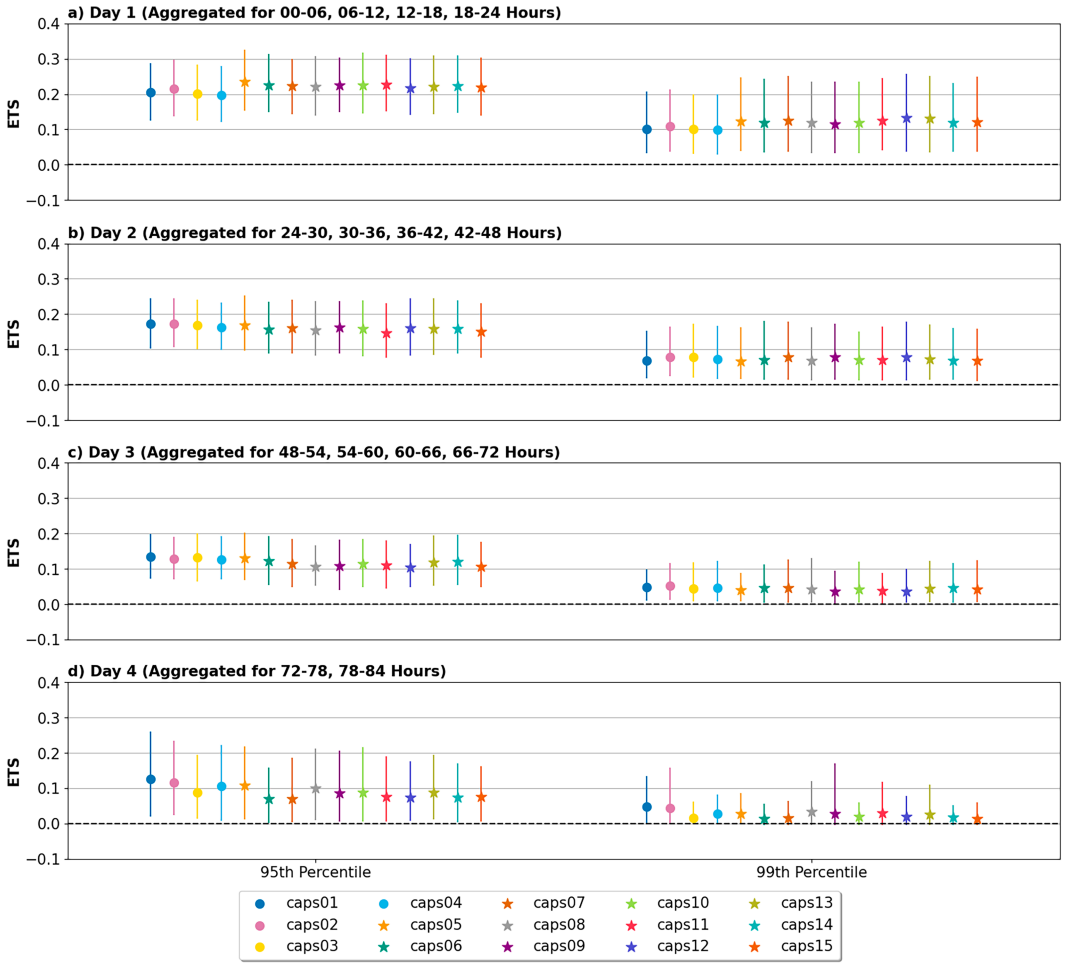


FIG. 3. ETS for DA-CU OFF (circular markers) and DA-CU ON (star markers) CAPS ensemble members for both 95th percentile 6-h precipitation (left) and 99th percentile precipitation (right) aggregated during days (a) 1, (b) 2, (c) 3, and (d) 4. Whiskers represent the 90% confidence interval from the bootstrapping significance test.

primary additions to the 2024 CAPS ensemble compared to previous iterations, while the observed thresholds from Stage IV are highlighted in black. The observed values for the 99P precipitation are just under 20 mm (0.8 in.) for the 99P and 5 mm (0.2 in.) for the 95P. At percentiles greater than the 95th, DA-CU OFF members have higher precipitation than DA-CU ON members, while the reverse is true below the 94–95P. DA-CU ON members increase precipitation values for lighter percentiles while dampening heavier precipitation percentiles. The DA-CU ON members do appear to help correct erroneously heavy precipitation around the 99P compared to DA-CU OFF members but are overly dry at percentiles immediately below that threshold compared to both DA-CU OFF members and observations. Around the 95P, the two sets of members are nearly identical, and both underestimate the observed 95P.

ETS for both 95P and 99P precipitation is shown for each CAPS member in Fig. 3. For day 1, with aggregated scores between forecast hours 0 and 24, ETSs generally are between 0.2 and 0.3 for 95P precipitation and 0.1–0.15 for 99P precipitation. These scores are considerably lower than those found in Supinie et al. (2022), largely due to the shorter time window used for accumulation (6 vs 24 h) and the higher precipitation thresholds. ETS in Supinie et al. (2022) for 50 mm of precipitation in 24 h is similar to this study's 99P precipitation for the first 2 days of their forecasts, defined as 12–36 and 36–60 h, respectively, with day 3 defined as 60–84 h. The differing aggregation windows are chosen to assess the possible benefits of data assimilation (Park et al. 2025). ETS is generally slightly higher, and the 90% confidence interval is shifted upward on day 1 forecasts for the DA-CU ON members in both 95P and 99P precipitation, consistent with the results of Park et al. (2025), wherein ETS for those members decreased between 12 and 18 h into the forecast.

For days 2–4, ETS decreases with increasing lead time. There are generally few differences between DA-CU ON and DA-CU OFF members by day 2, indicating that the increased skill imparted by data assimilation largely does not persist beyond 24 h (Park et al. 2025). Again, DA-CU ON NSSL microphysics members perform slightly worse than their Thompson

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:44 PM UTC

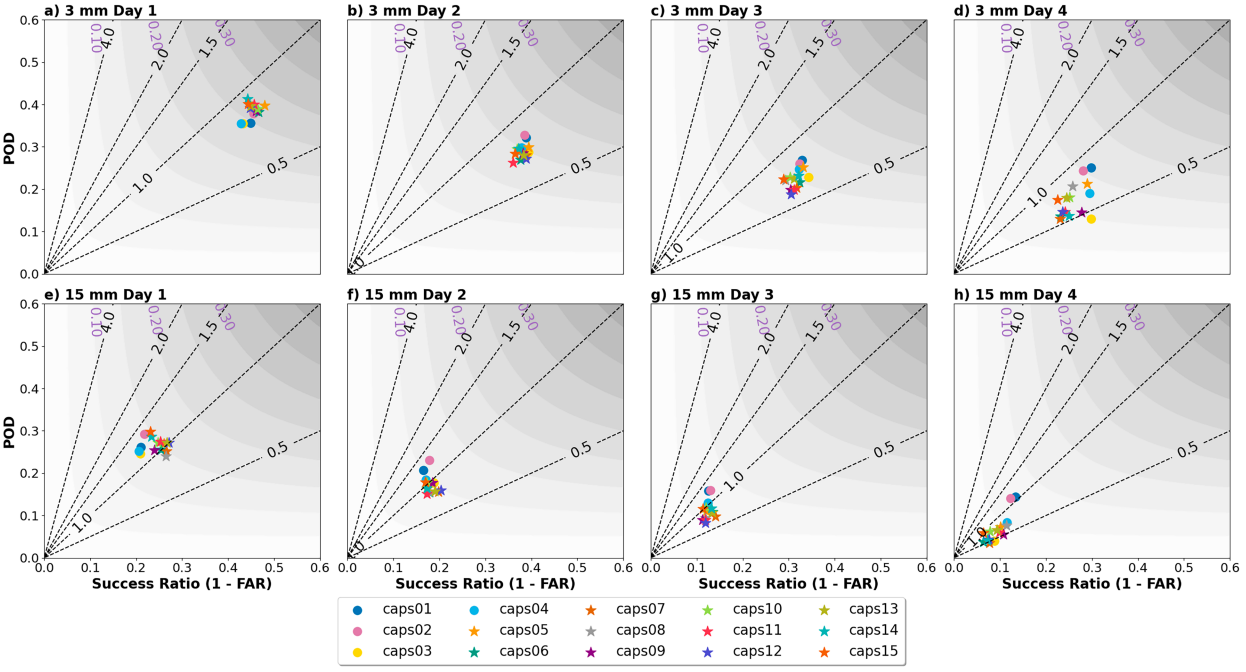


FIG. 4. Performance diagrams for DA-CU OFF (circular markers) and DA-CU ON (starred markers) 6-h precipitation for the CAPS ensemble members aggregated for 3-mm precipitation during days (a) 1, (b) 2, (c) 3, and (d) 4 and 15-mm precipitation during days (e) 1, (f) 2, (g) 3, and (h) 4. The grayscale color shading represents contours of CSI, while the black dashed lines are frequency bias contours.

counterparts for 95P precipitation, possibly owing to the drop size distribution and evaporative differences documented in Supinie et al. (2022). By day 4 (which corresponds to only half a day from 72 to 84 h), most skill via ETS is lost for 99P precipitation, with values barely above zero for most members. This is to be expected as rarer events will show a sharper decrease in predictive skill with increasing lead time (Hitchens et al. 2013), especially for convective features that become difficult to predict beyond 24–48 h, as shown by the intrinsic limits of predictability calculated in Zhang et al. (2019). The predictability limit of features less than 200 km in horizontal scale in their study was less than 3 days, which has ramifications for the 99P precipitation results in this study as discussed hereafter. However, CAPS03, CAPS06, CAPS07, CAPS12, CAPS14, and CAPS15 show slightly lower skill for 99P precipitation at this range. All but one of these members (CAPS15) employ the Noah-MP land surface model, which is discussed in the next paragraph. CAPS01 and CAPS02 (Figs. 3a,b) show comparatively better results than the remainder of the members on day 4, nearly matching their day 3 results. Both members use the MYNN boundary layer scheme and Noah land surface model, the latter of which is not used in any of the other members aside from MPAS01. It should be emphasized that no members are statistically significantly better or worse than others at either precipitation threshold at the 90% level.

Examining the performance diagrams of individual members in Fig. 4, some differences in biases become apparent. Starting with the DA-CU OFF members on day 1, CAPS03 has a larger low-frequency bias for both 3- and 15-mm precipitation than the remaining three DA-CU OFF members. All DA-CU OFF members have a low bias for 3-mm precipitation, while CAPS01 and CAPS02 are high biased for 15-mm precipitation, and CAPS04 is nearly unbiased for 15-mm precipitation on days 2 and 3. These more amplified biases for CAPS03 were also documented in Supinie et al. (2022) and Snook et al. (2025) for lower precipitation thresholds and were attributed to enhanced latent heat fluxes using the TKE-EDMF boundary layer, GFS surface scheme, and Noah-MP land surface model near the Great Lakes and Gulf Coasts in the former study. CAPS03 also has larger low-frequency biases during days 3 and 4, to the point of statistical significance (computed separately, not shown) compared to CAPS01 and CAPS02, although all four DA-CU OFF members tend to contain these longer-range low biases, especially for 3-mm precipitation.

Introducing radar data assimilation generally results in increased day 1 CSI for DA-CU ON members versus DA-CU OFF members, particularly CAPS05, which yields the highest CSI on day 1 of all members in the ensemble. CAPS10, which uses the TKE-EDMF boundary layer scheme in place of the MYNN scheme, performs similarly well to CAPS05. DA-CU ON members running NSSL microphysics (CAPS11–15) seem to perform slightly better, but not to the point of statistical significance, in terms of bias on day 1 versus those running Thompson microphysics. CAPS06–07, CAPS09, and CAPS11–12 all have low-frequency biases during all days of the forecast at the 3-mm threshold and for days 2–4 at the 15-mm threshold. All of these members use the Noah-MP land surface model as with CAPS03 (Snook et al. 2025). Most skill for 15-mm precipitation generally decays by day 3 in all members although DA-CU OFF members are notably higher, while 3-mm precipitation shows skill via CSI

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:44 PM UTC

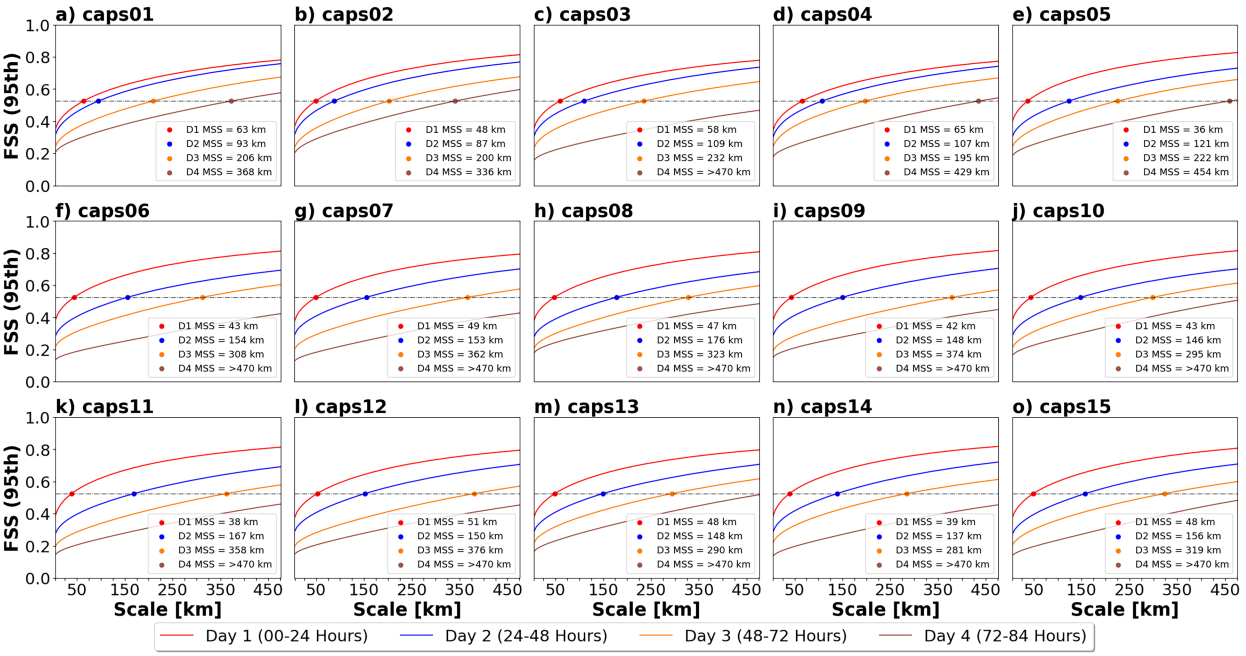


FIG. 5. FSS for 95P 6-h precipitation for the CAPS ensemble members aggregated during days 1 (red), 2 (blue), 3 (orange), and 4 (brown). The dashed–dotted line represents FSS_{useful} for the 95P. Markers indicate where each line crosses the FSS_{useful} threshold, if applicable. MSS for each day is indicated in each subplot legend.

through the full range of forecasts. The results for contingency metrics on an individual member basis generally suggest that a mixture of radar data assimilation, the MYNN boundary layer, and the RUC land surface model using the FV3 core yields optimal results in terms of forecasting precipitation, while the choice of microphysics (e.g., CAPS10 and CAPS15 performing similarly) or cumulus schemes is less influential. Members employing subgrid cumulus schemes do exhibit a slightly faster drop-off in CSI with increasing lead time than those without, while CAPS01 and CAPS02 perform slightly better relative to the remainder of the members on days 2–4. These differences are generally not to the point of statistical significance with either CSI or bias, however.

Considering FSS using 95P precipitation in Fig. 5, the MSS on day 1 is approximately 45–65 km for the DA-CU OFF members, with the smallest skillful scale (best performance) from CAPS02. DA-CU ON members do show an improvement on day 1, with the MSS shifted closer to 40 km. CAPS05 has the best performance of all members with an MSS of 36 km, as it did for day 1 CSI in the performance diagrams. FSS supports the contingency table results that radar data assimilation is providing benefits for forecast skill during day 1. However, for days 2–4, differences in FSS behavior begin to appear between DA-CU OFF and DA-CU ON members. The MSS for 95P precipitation on day 2 for DA-CU OFF members is roughly 90–110 km. For DA-CU ON members, this value rises to 120–175 km, with the only member near 120 km being CAPS05. For day 3, the MSS increases to 200–240 km for DA-CU OFF members and 225–400 km for DA-CU ON members. Skill is notably low for day 3 in CAPS07, CAPS09, CAPS11, and CAPS12, with MSSs of over 350 km. All of these members use the Noah-MP land surface model that was found to yield degradation in skill scores by Snook et al. (2025). CAPS03, which also uses Noah-MP, shows notably poor performance on day 4, with an MSS of greater than 475 km. The other DA-CU OFF members have day 4 MSSs of 325–430 km. Many of the DA-CU ON members do not reach the FSS_{useful} threshold on day 4. Aside from the poor performance of the Noah-MP members, a consistent result is the more substantial decline in forecast skill via FSS after day 1 for DA-CU ON members. This result is replicated for 99P precipitation in Fig. 6. Here, no DA-CU ON members reach FSS_{useful} within the given bounds on days 3 or 4. Again, CAPS06–7, CAPS09, and CAPS11–12 are notably poor performers on day 2, while day 1 results slightly favor the DA-CU OFF members.

As described above, the physics differences in these DA-CU ON versus DA-CU OFF members stem from three main factors: initial and boundary conditions (GEFS vs GFS), radar data assimilation, and the activation of cumulus convective schemes. CAPS05 performs the best among the DA-CU ON members and uses GFS initial conditions augmented by data assimilation. However, for 99P precipitation, CAPS05 suffers a similar sharp decrease in skill as the other DA-CU ON members between day 1 and day 2, which somewhat weakens the hypothesis that boundary conditions are behind the differences. That said, boundary condition differences should still be factored in, especially considering that CAPS05 uses the DA-augmented GEFS mean and that no DA-CU ON members use GFS boundary conditions. Stochastic physics perturbations (see Table 1) are also used in some physics schemes, which could act as a source of variance. However, Snook et al. (2025) found that there was little impact on precipitation

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:44 PM UTC

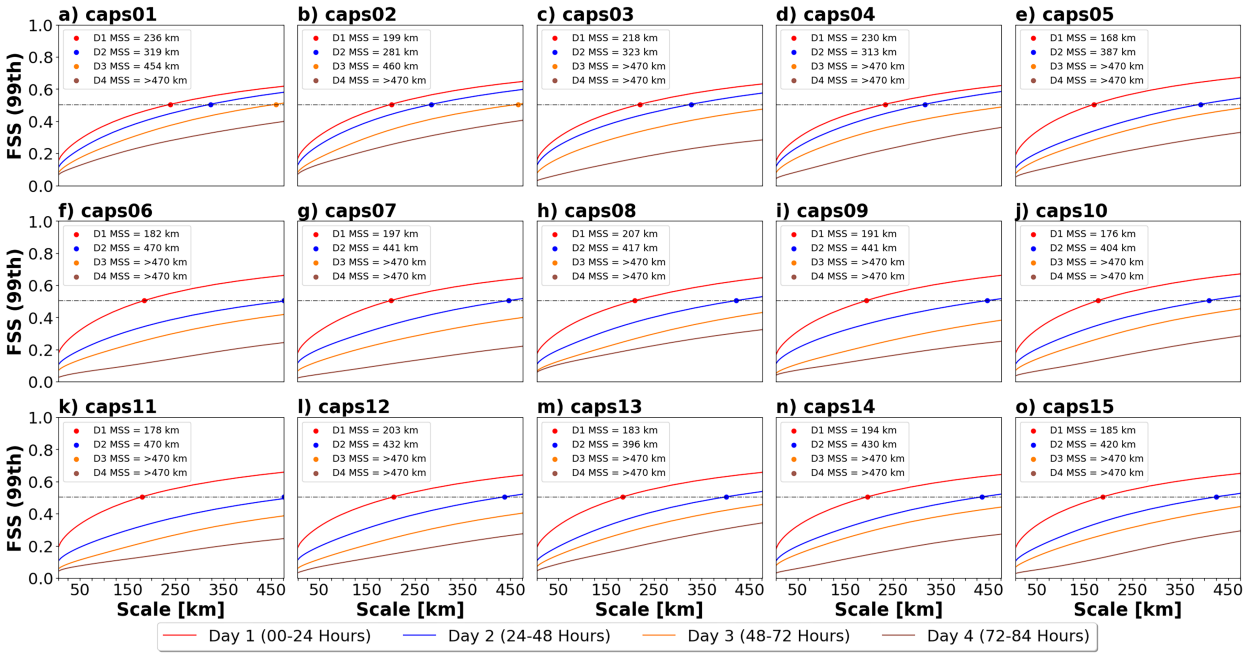


FIG. 6. As in Fig. 5, but for 99P 6-h precipitation.

forecasts from SPP in an earlier version of the CAPS ensemble. Improved performance using data assimilation diminishes by day 2, which agrees with earlier results and with Park et al. (2025). Given the lack of influence from data assimilation beyond 12–18 h, the final remaining factor is cumulus convective schemes. Skill as measured by FSS and CSI decreases more quickly with increasing lead time in members with cumulus schemes turned on than in members which do not use a cumulus scheme. Results of previous studies (Wang and Seaman 1997; Holloway et al. 2012; Han et al. 2016; Sun et al. 2016; Qiao and Liang 2017; Budakoti et al. 2019; Freitas et al. 2021) describe some of the uncertainties associated with the impacts of cumulus schemes on precipitation forecasts, including difficulties with the distribution of precipitation intensities such as overprediction of light precipitation. It should be emphasized that direct attributions of skill changes to any one physics change are not inherently possible given the design of this experiment. Further elaboration is deferred to this paper's conclusion section.

To conclude the analysis of individual members, we briefly investigate the performance of the experimental MPAS member (MPAS01) using performance diagrams and ETS (Fig. 7). Given that MPAS01 and CAPS01 use the same model physics configuration and the same boundary conditions, it follows to compare these two members directly. MPAS01 performs comparably to CAPS01 during day 1 for both thresholds of precipitation, with a slight low-frequency bias, compared to a slight high bias for 15-mm precipitation in CAPS01. For both 3- and 15-mm precipitation accumulation thresholds, there is degraded skill for MPAS01 on days 2, 3, and particularly day 4 compared to CAPS01. The degradation does not exceed the uncertainty limits estimated from bootstrapping, however. Given the earlier stage of tuning of physics packages with the MPAS dynamic core relative to the FV3, it does make sense that there is some degraded skill, particularly later in the forecast. However, the competitiveness through at least the first day agrees with Park et al. (2025) and suggests promise in future developments using the MPAS dynamic core. The MPAS dynamic core has already been suggested by some to have benefits relative to the FV3 core for certain situations, particularly from HWT experiments concerning severe convection (Carley et al. 2023). It would be beneficial to compare additional MPAS members with different physics configurations to their FV3 counterparts in future HMT experiments to further evaluate the performance of the MPAS dynamic core for RRFS development and precipitation forecasts.

b. *Ensemble consensus products*

The 6-h precipitation percentiles for CAPS ensemble consensus products are presented in Fig. 8. The standard ensemble mean generally overestimates light–moderate precipitation while underestimating heavy precipitation, this is expected due to the smoothing and smearing effects of the standard mean. The spatially aligned variant corrects these errors somewhat, although the SAM retains some of the biases of the standard mean. For the PM and LPM and their SAM counterparts, light precipitation (below the 90th percentile) is represented quite well relative to observations along with higher-end precipitation (above the 98th percentile). Precipitation in the moderate percentiles (92nd–97th percentiles) is generally underestimated by these metrics. For the SAM, the results align with Lee et al. (2024), who found improvements in higher-end precipitation estimates across all means when employing the SAM technique. However, Lee et al. (2024) did not generally evaluate

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:44 PM UTC

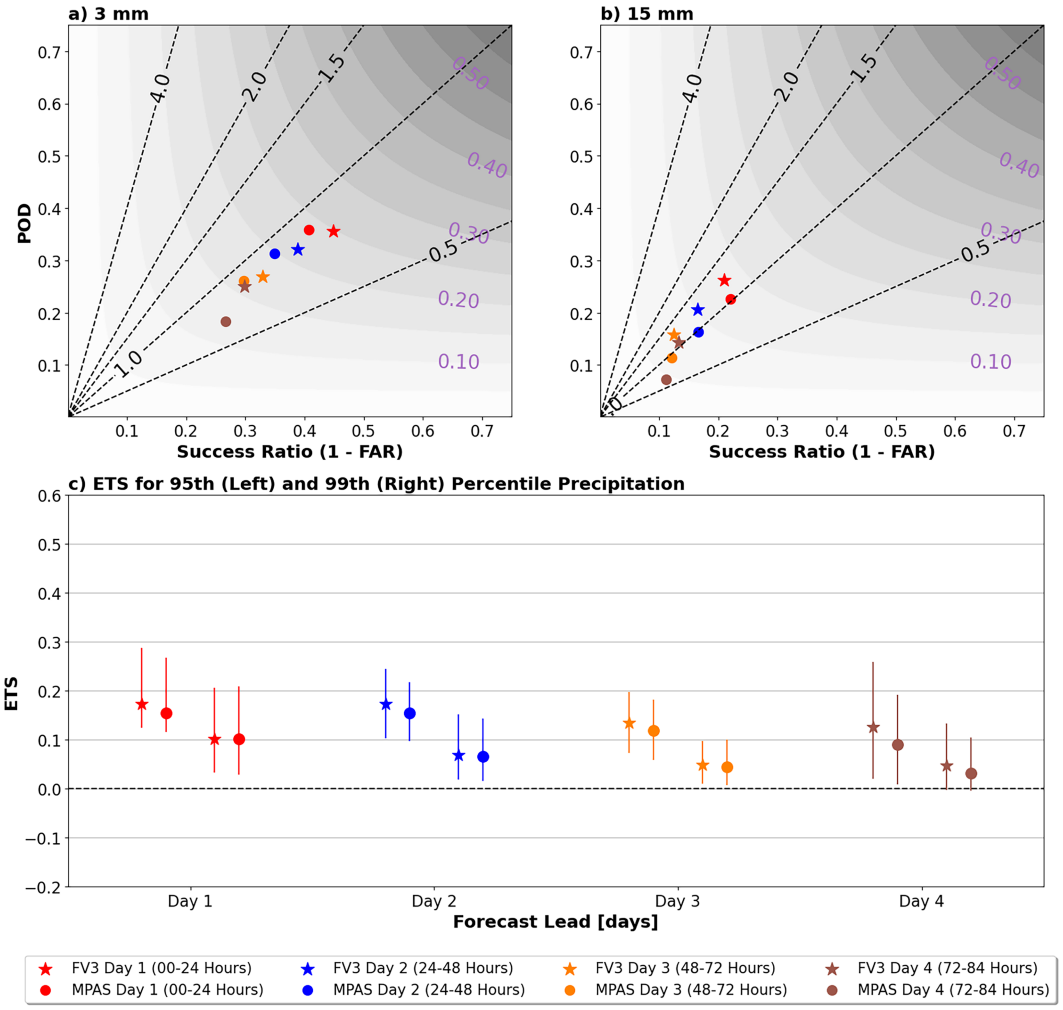


FIG. 7. Performance diagrams for (a) 3- and (b) 15-mm 6-h precipitation and (c) ETS for the CAPS01 member (star markers) and MPAS01 member (circular markers). Color codes for forecast leads follow conventions in Fig. 5. In (c), 95P results are positioned left of each forecast lead, while 99P results are positioned right.

precipitation amounts at lower thresholds in their study. Snook et al. (2019, 2020) suggested that their LPM (pointwise or patchwise) underperformed the PM at lower precipitation thresholds but improved for higher precipitation thresholds and yielded improved spatial structure.

The ETSs for the 99P precipitation for the ensemble consensus products are presented in Fig. 9. The results suggest that the LPM mean (Fig. 9c) underperforms the PM and standard means slightly in terms of ETS at this percentile threshold. Differences between the standard mean and the PM mean appear relatively negligible for ETS. While this may seem to contradict Snook et al. (2020, 2025) to a degree, it is important to highlight that the 99P precipitation in this study is still near the lower end of the precipitation thresholds used in that work, and that, unlike in those prior studies, the conclusion is drawn from data from the summer rather than the spring season. Those studies also used fixed thresholds rather than percentiles. The LPM also earned higher subjective ratings from forecasters due to retained structural details (Snook et al. 2020). The results for precipitation thresholds in the 10–20-mm range in Snook et al. (2020) were less favorable to the LPM mean relative to the PM mean. Additionally, the characteristics and physics of the 2024 CAPS FFaIR ensemble differ from those of the CAPS ensemble in those years. The SAM variants tend to match or slightly outperform the standard means, especially the LPM, consistent with the results of Lee et al. (2024) and the subjective verification done in the 2024 FFaIR experiment (Trojniak and Correia 2024). The 95P ETS (not shown) generally follows the same patterns, except with the ETSs and confidence intervals shifted upward by 0.1–0.15 at all lead times. None of these results for ETS are statistically significant, however.

Performance diagrams for ensemble consensus products in Fig. 10 are similar in terms of bias and skill to those of the individual members in Fig. 4 at the 15-mm threshold. A substantial low-frequency bias for day 1 is noted in the standard ensemble mean, while the SAM counterpart mitigates this bias and slightly increases CSI, resulting in a forecast with

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:44 PM UTC

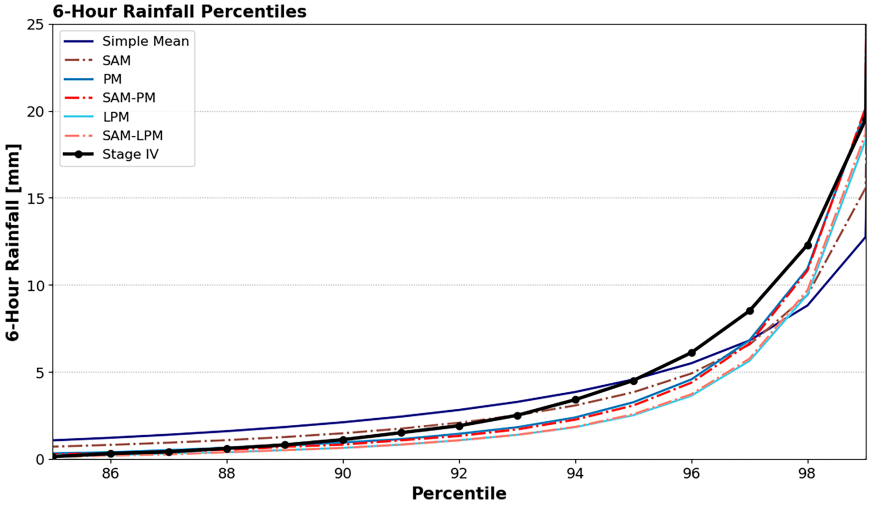


FIG. 8. As in Fig. 2, but for the ensemble consensus products and for each of the 6-h periods of the forecasts. Blue shades represent the standard mean, PM, and LPM, while red shades represent their SAM counterparts.

greater overall skill. The SAM-PM and SAM-LPM show increased CSI over the versions of the PM and LPM without SAM, and the LPM overall exhibits notably lower CSI compared to the other ensemble mean products evaluated. By day 2, a pronounced low-frequency bias is present for the standard mean, and more moderate biases are present for the PM and LPM. As on day 1, the SAM-PM and SAM-LPM exhibit similar (PM) or higher (LPM) CSI and slightly improved frequency bias compared to the variants without SAM. For days 3 and 4, low-frequency biases are evident for all

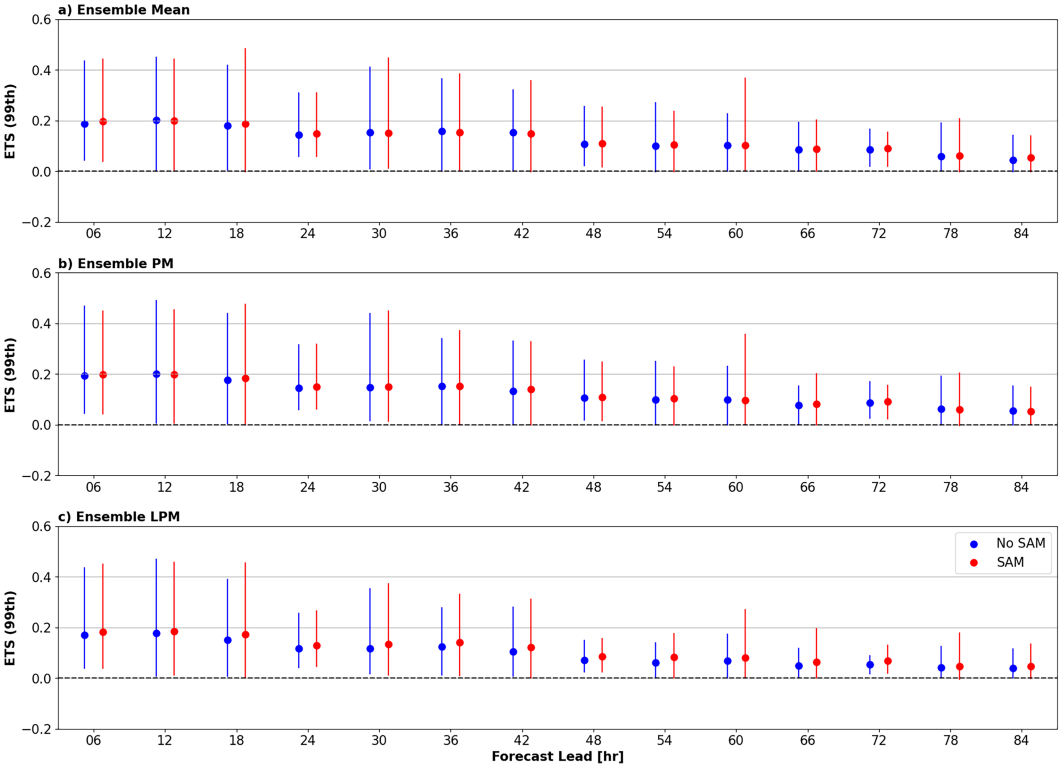


FIG. 9. As in Fig. 3, but only for 99P 6-h precipitation of the ensemble consensus products and for each of the 6-h periods of the forecasts instead of daily aggregations. Blue represents the standard mean, PM, and LPM, while red represents their SAM counterparts.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:44 PM UTC

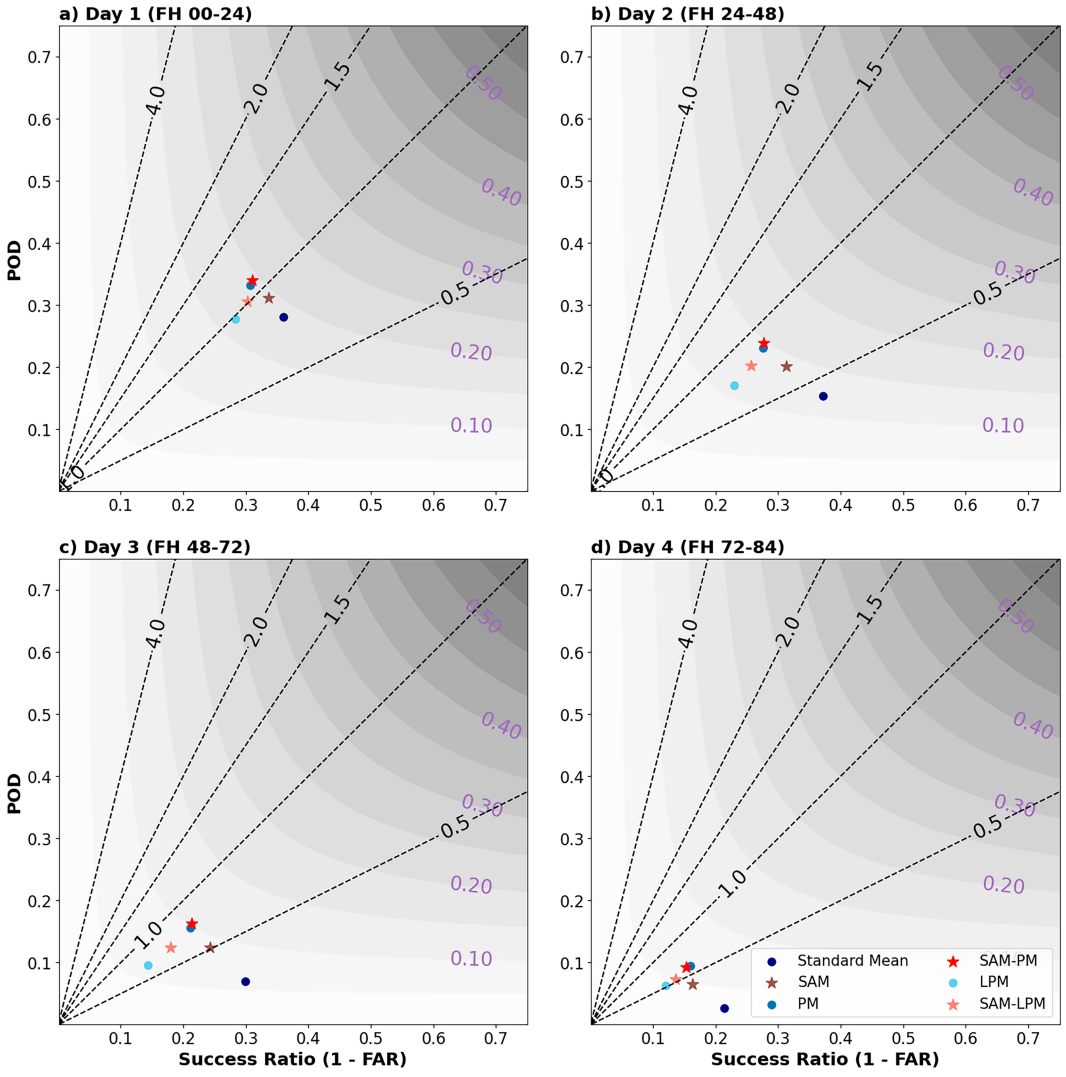


FIG. 10. Performance diagrams for 15-mm 6-h precipitation in the ensemble consensus products at (a) day 1, (b) day 2, (c) day 3, and (d) day 4. Blue shades (with circular markers) represent the standard mean, PM, and LPM, while red shades (with star markers) represent their SAM counterparts.

products, as with the majority of individual ensemble members (Fig. 4); this low bias is particularly notable for the standard ensemble mean. However, the SAM continues to improve upon the standard means for all lead times. Similar patterns regarding the impact of SAM are also present for CSI at the 3-mm precipitation threshold (not shown), although the impact is greater at the 15-mm precipitation threshold, highlighting the benefits of the SAM's improved spatial structures, which are also depicted in the case studies evaluated by Lee et al. (2024). We refer the reader to this prior study for a more detailed analysis of the SAM.

The results of this study and those of Lee et al. (2024) suggest that it might be beneficial to test the ensemble means for FFaIR 2024 at higher precipitation thresholds (i.e., for rare, intense events). However, the SAM still seems to add forecast value, particularly when evaluating the performance diagrams, even at lower thresholds. This suggests that it remains a useful product for further investigation and potential implementation in future operational forecast ensembles.

c. *Probability-based pointwise and NMEP metrics and short-range comparison with the HREF*

In this section, we focus on probabilistic metrics such as AUC, reliability, and performance diagrams, along with neighborhood verification as outlined in section 2. Because the full HREF ensemble is unavailable beyond 36 h of forecast time, and because the general impacts of applying neighborhood methods are similar at all forecast lead times, this section focuses exclusively on short-range forecasts (0–36-h lead time). Results from Supinie et al. (2022) and Snook et al. (2025) generally indicate that the CAPS ensemble outperforms the GEFS forecasts that extend to 84 h, in large part due to its finer grid resolution. Here, we compare the CAPS ensemble to the HREF, with similar grid resolution, as the reference for shorter-term forecasts. We note that

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:44 PM UTC

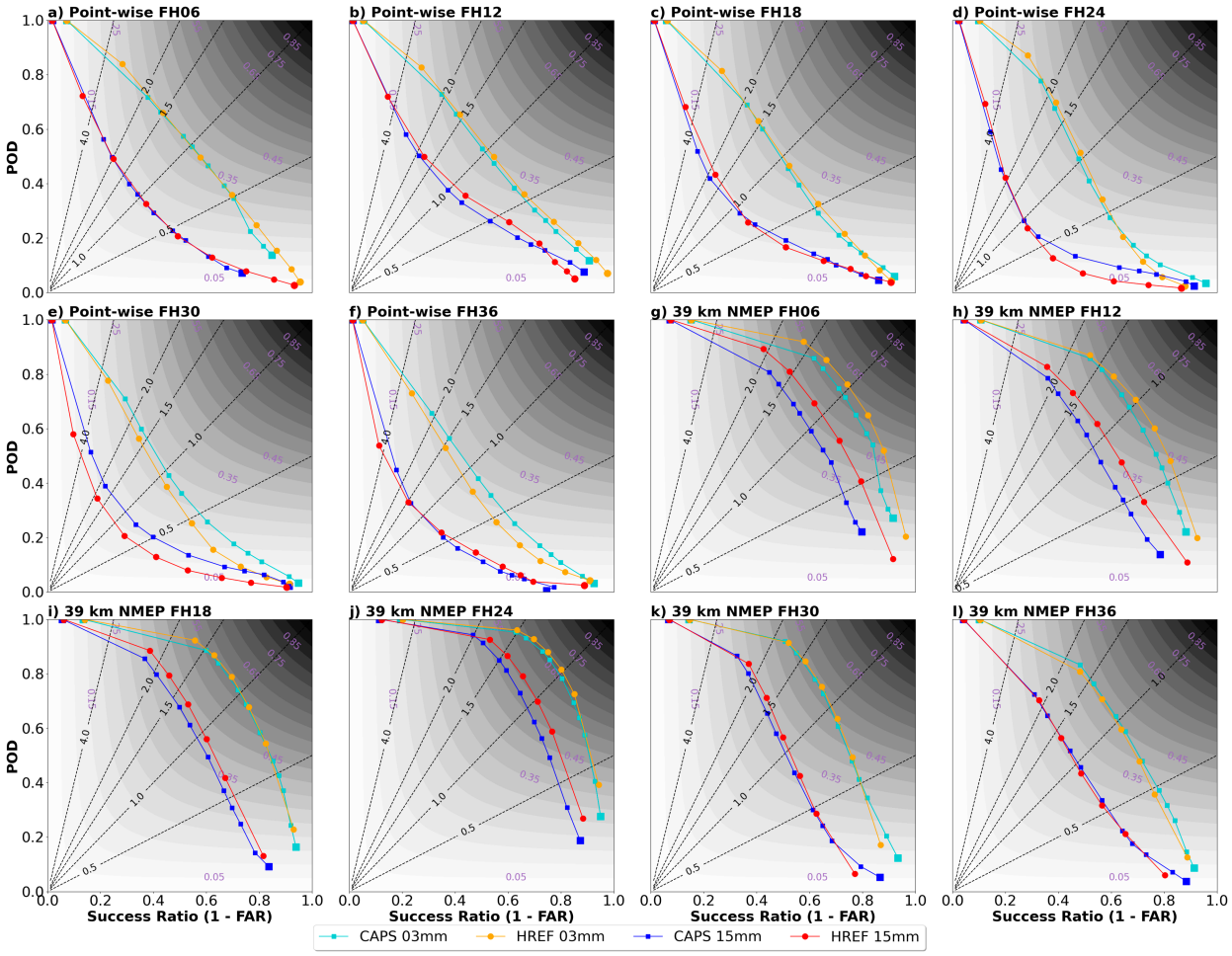


FIG. 11. (a)–(f) Probabilistic pointwise and (g)–(l) 39-km NMEP performance diagrams for 6-h precipitation at forecast hours (a),(g) 6, (b),(h) 12, (c),(i) 18, (d),(j) 24, (e),(k) 30, and (f),(l) 36. Light blue (3 mm) and blue (15 mm) lines represent the CAPS ensemble, while orange (3 mm) and red (15 mm) lines represent the HREF.

the CAPS ensemble does not contain the same number of members and does not contain a time-lagged component as the HREF does, and thus the results comparing the two should be taken with some degree of caution. Performance diagrams for probabilities between 0% and 100% of exceeding the given threshold on a pointwise basis are shown in Figs. 11a–f. At a 6-h precipitation accumulation threshold of 3 mm, the CAPS ensemble exhibits higher CSI than the HREF at hours 30 and 36 for probabilities of 30%–60%, while exhibiting comparable or slightly lower CSI at other forecast hours. The CAPS ensemble exhibits higher-frequency bias at lower probabilities compared to the HREF, particularly during the morning hours.

Notable differences at the 15-mm threshold include higher CSI compared to the HREF at forecast hours 24 and 30, along with a similar improvement in frequency bias as for the 3-mm threshold. The results are not very consistent across all forecast hours suggesting that the CAPS ensemble and the HREF likely differ in terms of forecast skill at different points during the diurnal cycle of precipitation. The influences of cumulus schemes and data assimilation on the diurnal precipitation cycle should be further investigated in future work. Applying a 39-km NMEP to the ensemble data dramatically improves CSI performance for both ensembles (Figs. 11g–l). The HREF outperforms the CAPS ensemble for 15-mm precipitation when using neighborhood verification, while the two generally perform comparably at the lower 6-h accumulation threshold of 3 mm, aside from forecast hours 6 and 12, where the HREF is superior. The performance of the CAPS ensemble relative to the HREF is promising since the HREF (Roberts et al. 2019) is known to be a very capable ensemble having benefited from extensive development and optimization of its members (Lee et al. 2024; Snook et al. 2025).

We also consider pointwise and 39-km NMEP ROC-AUC scores calculated for the CAPS ensemble and the HREF, which are presented in Fig. 12. Since AUC is plotted as a time series, we additionally calculate AUC for 1-h precipitation at the same percentile thresholds for this comparison to highlight diurnal variation and metrics at a higher temporal resolution. All 1-h postprocessing and verification scoring are calculated in a similar fashion to those for 6-h precipitation.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:44 PM UTC

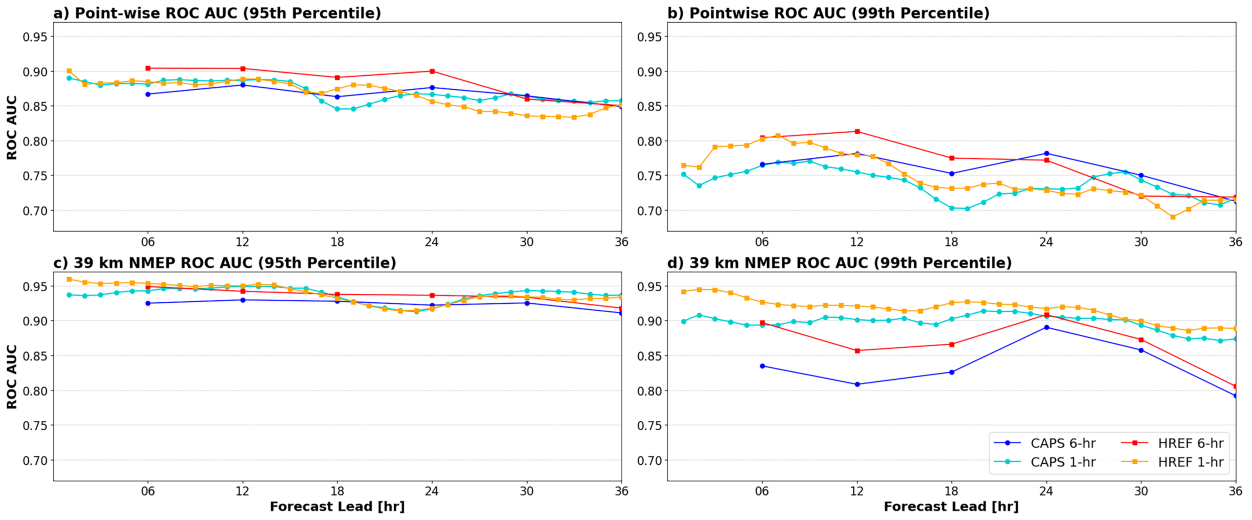


FIG. 12. AUC for (a) pointwise 95th, (b) pointwise 99th, (c) 39-km NMEP 95th, and (d) 39-km NMEP 99th percentile verification for both 6-h (blue and red) and 1-h (cyan and orange) precipitation in eight selected members of the CAPS ensemble (cool colors; circular markers) and the HREF (warm colors; square markers).

Additionally, since AUC can be affected by differing ensemble sizes, especially for rare events (Ben Bouallègue and Richardson 2022), we select eight of the best-performing members (via ETS and performance diagrams in Figs. 3 and 4) in the individual member evaluations—CAPS01, CAPS02, CAPS04, CAPS05, CAPS07, CAPS09, CAPS12, and CAPS13—from the CAPS ensemble to provide a fairer comparison to the HREF for this test. For pointwise verification, the CAPS ensemble AUCs are slightly lower than those of the HREF for 95P 6-h accumulated precipitation (Fig. 12a). Generally, pointwise 1-h AUC scores are slightly lower than their 6-h counterparts. At forecast hours 30 and 36, the CAPS ensemble slightly outperforms the HREF forecasts. AUC scores at the 95P threshold generally remain around 0.85–0.9 out to 36 h of forecast lead time, indicating good performance. For 99P 6-h precipitation (Fig. 12b), AUC scores decrease to around 0.75–0.8 by 24 h of forecast lead time and decrease further for forecast hours 30 and 36. The HREF slightly outperforms the CAPS ensemble on the first day, with the CAPS ensemble improving relative to the HREF through 24–30 h. Similar results are reflected in the 1-h verification. The impacts of diurnal variation are evident in the CAPS ensemble for 99P precipitation, with a relative minimum at 18 h (early afternoon) and a relative maximum at 3–9 and 27–30 h (late evening).

Applying the 39-km NMEP yields increased skill via AUC for both ensembles at both thresholds. The CAPS ensemble nearly matches the HREF at all lead times for 95P precipitation (Fig. 12c) at both temporal resolutions, with an AUC of roughly 0.95. The HREF slightly outperforms the CAPS ensemble for 99P 6-h precipitation during day 1, while the two ensembles perform similarly beyond 24 h of forecast lead time. For 99P 6-h NMEP verification, there is a distinct rise in AUC during the evening hours, which again may reflect the diurnal cycle favoring heavy precipitation (convection) during these hours at a higher frequency. The decay of morning convection producing heavier precipitation, often in the form of mesoscale convective systems (Schumacher and Johnson 2008; Hu et al. 2021; Cui et al. 2024), may be less accurately depicted in the models (Kiel et al. 2022). Using 1-h precipitation with NMEP, the diurnal cycle is muted, while AUC rises to 0.9 or higher for both ensembles. The CAPS ensemble and the HREF are generally comparable in performance, although the HREF tends to slightly outperform the CAPS ensemble for 99P precipitation.

The CAPS ensemble is evaluated out to day 4 to examine the ensemble's calibration at longer lead times. In terms of forecast reliability, for pointwise verification (Figs. 13a–d), the ensemble remains skillful in this metric for 95P precipitation through the duration of the forecast. However, it tends to overpredict with higher forecast probabilities than observed. At the 99P threshold, there is overprediction to the point of losing skill at lower probabilities during day 1 but improved skill at higher probabilities to the point of nearly matching that for 95P precipitation. For days 2 and even 3, skill is yielded via reliability at all probabilities (although less so for low probabilities) for 99P precipitation. Day 4 sees a degradation compared to shorter lead times below the no-skill line for 99P precipitation, consistent with earlier performance diagram and ETS results for individual members. Introducing NMEP to the reliability calculations (Figs. 13e–h) results in overprediction at the 95P to the point of losing skill for the CAPS ensemble. Using a neighborhood for 95P precipitation typically is not preferred given that regions of lighter precipitation are generally widespread enough to be captured by pointwise verification. However, using NMEP for 99P precipitation yields skill via reliability for all forecast lead times, although some degree of overprediction remains. It follows that for higher-intensity precipitation (leaning toward smaller-scale convective features), NMEP provides additional benefits consistent with previous work (Clark et al. 2010; Schwartz and Sobash 2017).

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:44 PM UTC

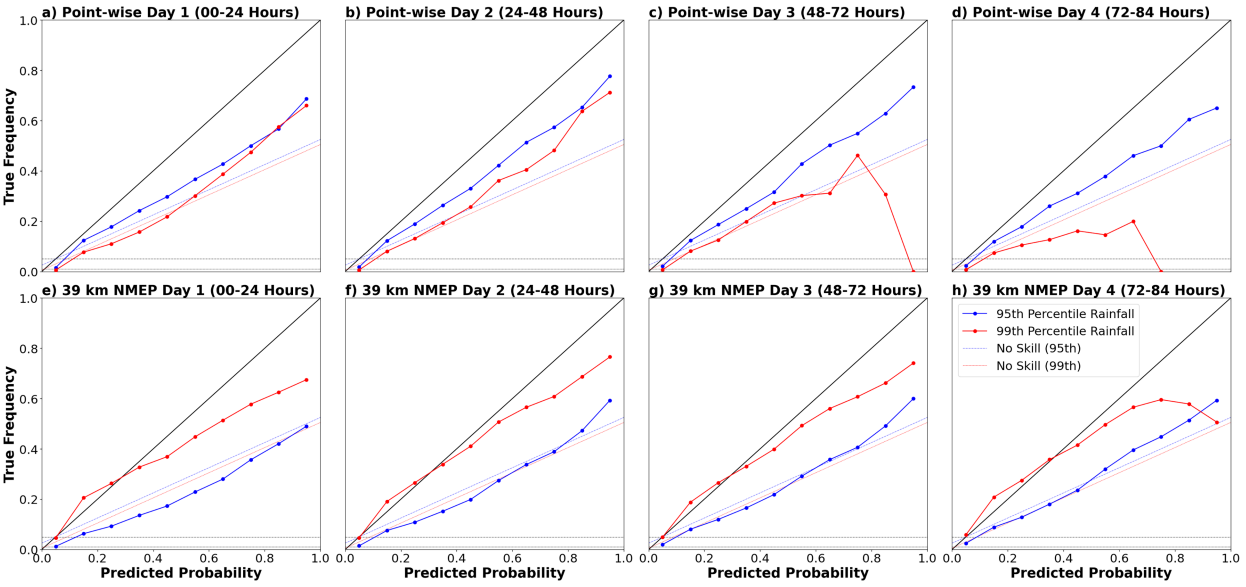


FIG. 13. Reliability diagrams for (top) pointwise and (bottom) 39-km NMEP verification for both 95th (blue) and 99th (red) percentile 6-h precipitation averaged over (a),(e) day 1, (b),(f) day 2, (c),(g) day 3, and (d),(h) day 4. The black solid line represents perfect reliability, while the black dashed and dotted lines represent climatological true frequency (0.05 and 0.01 for the 95th and 99P, respectively). The blue dashed line represents the line of no skill for 95P precipitation, while the red dashed line is the line of no skill for 99P precipitation.

While not shown for the sake of brevity, comparing the CAPS ensemble to the HREF for reliability suggests comparable performance for the two ensembles for 95P precipitation in the first 36 h of the forecast. There is, however, notable overprediction during the first 6–12 h of the forecast at both thresholds in the CAPS ensemble. Adding neighborhood verification degrades forecasts of 95P precipitation while improving those of 99P precipitation, as before. The overprediction in the early hours of the forecast indicates that adding data assimilation (Park et al. 2025) may result in increased skill via deterministic metrics through increasing the number of hits but degrading reliability through overprediction (Fig. 4), or it may be due to the influence of the Noah-MP land surface model (Supinie et al. 2022; Snook et al. 2025). Further investigation of these sensitivities is beyond the scope of this paper and will require ensemble members with direct comparisons between different physics options being possible.

To summarize, the CAPS ensemble compares reasonably to the HREF in terms of probabilistic performance and AUC. There is some degradation in reliability for the first 6–12 h of the forecast, while subsequent forecast hours compare favorably. By and large, this suggests that the CAPS ensemble is relatively well calibrated and performs reasonably compared to the current standard high-resolution ensemble for warm-season precipitation. However, further work is needed to both investigate the reasons behind differences between the two ensembles and verify the 2024 FFaIR CAPS ensemble against a reference forecast that goes out to 84 h as in Supinie et al. (2022) and Snook et al. (2025).

4. Conclusions and discussion

A 15-member CAM ensemble using the FV3-LAM (Lin 2004; Black et al. 2021) dynamic core was run by CAPS during the summer 2024 HMT FFaIR (Barthold et al. 2015; Trojniak and Correia 2024) Experiment in support of the development of the RRFS. The ensemble used a variety of different physics configurations among its members, varied its boundary conditions, and used experimental radar data assimilation (Park et al. 2023, 2025). In this study, a broad evaluation of this ensemble through objective verification metrics is presented, with the goal of informing future ensemble development and assessing the ensemble's ability to predict summer precipitation at different thresholds.

Perhaps the most notable result above is the difference in forecast skill, especially in FSS, between members that employ cumulus convective schemes (Han et al. 2017; Freitas et al. 2021) and GEFS initial/boundary conditions and those that do not. The difference between DA-CU OFF and DA-CU ON members for FSS warrants some further deliberation. First, some consideration should be given to the findings of Baldwin et al. (2026), who determined that FSS for high event frequencies may behave in a less useful manner, although 95P is by definition a moderately "rare" event. A discussion by Molinari and Dudek (1992) highlighted the necessity of understanding convective precipitation parameterization in mesoscale models and the resulting uncertainties. Wang and Seaman (1997) compared several cumulus schemes in a mesoscale model and found that the model underpredicted heavy precipitation and overpredicted lighter precipitation. This result is generally consistent with the percentile distribution in Fig. 2. Meanwhile, tropical precipitation—somewhat analogous to warm-season precipitation in the midlatitudes—was analyzed by Holloway et al. (2012) using both explicit and convective parameterization, and they also found a preferred scale at lighter rain rates using convective parameterization.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:44 PM UTC

Sun et al. (2016), in their study of Plains summer precipitation, explicitly mentioned avoiding cumulus parameterization given the uncertainty it introduces in precipitation forecasts. A subsequent investigation by Qiao and Liang (2017) found that cumulus closures had a significant effect on the fidelity of precipitation diurnal cycles and precipitation magnitude during the summer over the CONUS, particularly in the Southeast. Han et al. (2016) determined, using an earlier version of one of the cumulus schemes employed in the CAPS ensemble during a 2013 rainfall case in South Korea, that heavy precipitation was reduced and recommended additional tests regarding the interaction of microphysical processes and cumulus schemes. The reason behind the lesser precipitation efficiency was tied to poor conversion of cloud condensate to convective precipitation at colder temperatures (Han et al. 2016). Using the other cumulus scheme in this study, Freitas et al. (2021) found an underestimate of heavy precipitation and an overestimate of light precipitation over the Indo-Pacific. Studies have also examined the convective gray zone (the transition between parameterized and resolved convective scales) and cumulus schemes at higher resolutions such as in the Weather Research and Forecasting (WRF) Model. Jeworrek et al. (2019) found that the choice of cumulus parameterization was very important in affecting precipitation patterns in their WRF simulations at 3-km grid spacing and that the scale-aware G-F scheme was the best performer. A later work by Zhang et al. (2021) at the same resolution using WRF found that differences between two cumulus schemes in terms of trigger functions, cloud models, and closure assumptions made it difficult to determine which was more useful at the subgrid scale. Finally, Park et al. (2024) directly addressed whether scale-aware cumulus schemes were improving heavy rainfall simulations over the Korean Peninsula. They found that the Kain–Fritsch scheme tended to overpredict light precipitation, aligning with earlier findings by Freitas et al. (2021). However, they did not use a convection-permitting resolution.

The results in Figs. 2, 5, and 6 and previous studies suggest that the role of cumulus parameterizations (if employed) should be more closely examined in future precipitation experiments using the CAPS ensemble, along with the performance of the Noah-MP land surface model (Snook et al. 2025). Since the inclusion of members using cumulus schemes is a new addition to the CAPS ensemble in 2024, and given that members in the 2024 FFaIR CAPS ensemble differ in more than just their use of cumulus schemes (i.e., data assimilation, boundary conditions, stochastic physics perturbations), a possible direction for future research would be to isolate the impacts of the cumulus scheme by running one or more pairs of members with identical physics configurations and boundary conditions, one with the cumulus scheme and the other without. We thus emphasize that a clear attribution of changes in skill between physics options, such as data assimilation, boundary conditions, or cumulus schemes, cannot be definitively determined given the configuration of the 2024 ensemble. Complementary to the idea of addressing this is the design of the 2025 CAPS FFaIR ensemble, which will be specifically designed to test the saSAS cumulus scheme directly. The ensemble will consist of a set of members with saSAS turned on and a set of members with saSAS turned off, with no other differences in physics or boundary conditions. Therefore, a more intensive evaluation of the effects of saSAS can be undertaken with apples-to-apples comparisons possible. The evaluation of this newer CAPS ensemble will be the topic of a forthcoming manuscript.

Greater frequency biases and poorer FSS were noted in members using the Noah-MP land surface scheme (Supinie et al. 2022; Snook et al. 2025), and improvement in CSI for members using experimental radar data assimilation (Park et al. 2025) was generally limited to day 1. The Noah-MP finding may be tied to GFS boundary layer and surface treatments in general (Noah-MP, TKE-EDMF boundary layer, GFS surface layer). Overall, in terms of precipitation forecast skill, the MYNN boundary layer scheme (Nakanishi and Niino 2009) and the RUC land surface model (Smirnova et al. 2016) seem to perform best with the FV3 core in this study. Both microphysical schemes tested in the 2024 CAPS ensemble (Thompson et al. 2004; Mansell et al. 2010) produce reasonable forecasts; of the two, the Thompson scheme performs slightly better (Zhang et al. 2019), albeit not statistically significantly so. The experimental MPAS (Skamarock et al. 2012) member performed well, particularly for heavier precipitation, through the first day of the forecast, with declining performance compared to the corresponding CAPS-FV3 member afterward. Subsequent RRFS development may incorporate the MPAS dynamic core (Carley et al. 2023), potentially providing future opportunities for evaluating the performance of MPAS and providing core diversity that may improve ensemble performance.

In terms of probabilistic and ensemble consensus metrics, the CAPS ensemble generally performs well. When compared to the HREF (Roberts et al. 2019), the CAPS ensemble exhibits competitive performance, AUC scores, and reliability for precipitation forecasts. Employing neighborhood verification methods showed that the HREF does tend to outperform the CAPS ensemble. As expected, neighborhood verification substantially improves skill scores for both ensembles with 99P precipitation. Producing ensemble consensus forecasts using the SAM method (Lee et al. 2024) improves CSI and reduces bias, a result consistent with prior studies, which suggests it might be suitable for use with future convection-allowing ensembles such as the REFS, which contains some similar member configurations to the CAPS ensemble (Blake et al. 2025). Overall, the probabilistic metrics (such as AUC) indicate that the 2024 CAPS configuration performs fairly well as an ensemble, but individual membership could be fine-tuned to improve individual member skill (as indicated by the deterministic metrics such as FSS and performance diagrams), while still providing desired amounts of spread. We also emphasize that the probabilistic metrics are primarily assessing the first 36 h of the forecast to enable comparison with the HREF. With that said, AUC for 99P precipitation certainly degrades in the 30–36-h period (Fig. 12b), and pointwise reliability becomes less than skillful on days 3–4 (Figs. 13c,d). Further regional-based verification should be performed over a broader range of cases to highlight the challenges faced by FV3 and MPAS ensembles in predicting precipitation and its relevant covariances. This is especially true given that this study removes much of the western CONUS due

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:44 PM UTC

to issues with Stage IV data (Nelson et al. 2016) over this area. Such proposed work could also focus on the mechanisms behind cumulus schemes affecting model precipitation under different conditions and on various methods of data assimilation (Park et al. 2023; Kong et al. 2024) affecting skill in the ensemble, which will aid in the future development of convection-allowing ensembles, including RRFS and its counterpart ensemble, the REFS.

Acknowledgments. This work was primarily supported by the NOAA Testbed Grant NA22OAR4590522 and NOAA UFS R2O Program Grant NA21OAR4320204. The CAPS real-time forecasts were run on the Frontera supercomputer at the Texas Advanced Computing Center (TACC) at the University of Texas at Austin, a National Science Foundation–supported high-performance computing center. We thank the three anonymous reviewers and the editor, whose helpful comments greatly improved the manuscript from its original version.

Data availability statement. Datasets used in this manuscript are publicly available. GFS forecasts used in this study are available from UCAR's Research Data Archive (RDA) repository (https://rda.ucar.edu/datasets/ds084.1/), while GEFS forecasts can be downloaded from NOAA's registry of open data on AWS at https://registry.opendata.aws/noaa-gefs/. Stage IV precipitation analyses can be downloaded from UCAR's Earth Observing Laboratory archive (https://data.eol.ucar.edu/dataset/21.093). HREF, CAPS FV3-LAM, and CAPS MPAS precipitation forecasts are available via OSF at https://osf.io/4zbxw/.

REFERENCES

Angevine, W. M., J. Olson, J. J. Gristey, I. Glenn, G. Feingold, and D. D. Turner, 2020: Scale awareness, resolved circulations, and practical limits in the MYNN–EDMF boundary layer and shallow cumulus scheme. *Mon. Wea. Rev.*, **148**, 4629–4639, https://doi.org/10.1175/MWR-D-20-0066.1.

Ashley, S. T., and W. S. Ashley, 2008: Flood fatalities in the United States. *J. Appl. Meteor. Climatol.*, **47**, 805–818, https://doi.org/10.1175/2007JAMC1611.1.

Baldwin, M. E., I. L. Jirak, and H. E. Brooks, 2026: Issues with measuring skill using the fractions skill score. *29th Conf. on Probability and Statistics*, Houston, TX, Amer. Meteor. Soc., 4.6, https://ams.confex.com/ams/106ANNUAL/meetingapp.cgi/Paper/477669.

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

Baxter, M. A., G. M. Lackmann, K. M. Mahoney, T. E. Workoff, and T. M. Hamill, 2014: Verification of quantitative precipitation reforecasts over the southeastern United States. *Wea. Forecasting*, **29**, 1199–1207, https://doi.org/10.1175/WAF-D-14-00055.1.

Ben Bouallègue, Z., and D. S. Richardson, 2022: On the ROC area of ensemble forecasts for rare events. *Wea. Forecasting*, **37**, 787–796, https://doi.org/10.1175/WAF-D-21-0195.1.

Black, T. L., and Coauthors, 2021: A limited area modeling capability for the Finite-Volume Cubed-Sphere (FV3) dynamical core and comparison with a global two-way nest. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002483, https://doi.org/10.1029/2021MS002483.

Blake, B., S. Liu, L. Gan, M. Caron, D. Lippi, A. Benjamin, and M. Pyle, 2025: RRFS parallels configuration files. NCEP NOAA, https://www.emc.ncep.noaa.gov/users/emc.campara/rrfs/namelist.html.

Brown, B., and Coauthors, 2021: The Model Evaluation Tools (MET): More than a decade of community-supported forecast verification. *Bull. Amer. Meteor. Soc.*, **102**, E782–E807, https://doi.org/10.1175/BAMS-D-19-0093.1.

Budakoti, S., C. Singh, and P. K. Pal, 2019: Assessment of various cumulus parameterization schemes for the simulation of very heavy rainfall event based on optimal ensemble approach. *Atmos. Res.*, **218**, 195–206, https://doi.org/10.1016/j.atmosres.2018.12.005.

Burke, P. C., A. Lamers, G. Carbin, M. J. Erickson, M. Klein, M. Chenard, J. McNatt, and L. Wood, 2023: The excessive rainfall outlook at the Weather Prediction Center: Operational definition, construction, and real-time collaboration. *Bull. Amer. Meteor. Soc.*, **104**, E542–E562, https://doi.org/10.1175/BAMS-D-21-0281.1.

Carley, J. R., C. R. Alexander, L. J. Wicker, C. Jablonowski, A. J. Clark, J. A. Nelson, I. L. Jirak, and K. Viner, 2023: Mitigation efforts to address Rapid Refresh Forecast System (RRFS) v1 dynamical core performance issues and recommendations for RRFS v2. NCEP Office Note 516, 88 pp., https://doi.org/10.25923/ccgj-7140.

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

——, and Coauthors, 2026: Results and insights on next-generation models from a real-time severe weather forecasting experiment. *Bull. Amer. Meteor. Soc.*, **107**, E880–E883, https://doi.org/10.1175/BAMS-D-25-0323.1.

Cui, W., T. J. Galarneau Jr., and K. A. Hoogewind, 2024: Changes in mesoscale convective system precipitation structures in response to a warming climate. *J. Geophys. Res. Atmos.*, **129**, e2023JD039920, https://doi.org/10.1029/2023JD039920.

Dahl, N., and M. Xue, 2016: Prediction of the 14 June 2010 Oklahoma City extreme precipitation and flooding event in a multiphysics multi-initial-conditions storm-scale ensemble forecasting system. *Wea. Forecasting*, **31**, 1215–1246, https://doi.org/10.1175/WAF-D-15-0116.1.

Doswell, C. A., H. E. Brooks, and R. A. Maddox, 1996: Flash flood forecasting: An ingredients-based methodology. *Wea. Forecasting*, **11**, 560–581, https://doi.org/10.1175/1520-0434(1996)011<0560:FFFAIB>2.0.CO;2.

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.

Ek, M. B., K. E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, and J. D. Tarpley, 2003: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:44 PM UTC

model. *J. Geophys. Res.*, **108**, 8851, https://doi.org/10.1029/2002JD003296.

Freitas, S. R., G. A. Grell, and H. Li, 2021: The Grell–Freitas (GF) convection parameterization: Recent developments, extensions, and applications. *Geosci. Model Dev.*, **14**, 5393–5411, https://doi.org/10.5194/gmd-14-5393-2021.

Hamill, T. M., 1997: Reliability diagrams for multicategory probabilistic forecasts. *Wea. Forecasting*, **12**, 736–741, https://doi.org/10.1175/1520-0434(1997)012<0736:RDFMPF>2.0.CO;2.

Han, J., and C. S. Bretherton, 2019: TKE-based moist eddy-diffusivity mass-flux (EDMF) parameterization for vertical turbulent mixing. *Wea. Forecasting*, **34**, 869–886, https://doi.org/10.1175/WAF-D-18-0146.1.

——, W. Wang, Y. C. Kwon, 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.

Han, J. Y., S. Y. Hong, K. S. S. Lim, and J. Han, 2016: Sensitivity of a cumulus parameterization scheme to precipitation production representation and its impact on a heavy rain event over Korea. *Mon. Wea. Rev.*, **144**, 2125–2135, https://doi.org/10.1175/MWR-D-15-0255.1.

Heinzeller, D., L. Bernardet, G. Firl, M. Zhang, X. Sun, and M. Ek, 2023: The Common Community Physics Package (CCPP) framework v6. *Geosci. Model Dev.*, **16**, 2235–2259, https://doi.org/10.5194/gmd-16-2235-2023.

Hitchens, N. M., H. E. Brooks, and M. P. Kay, 2013: Objective limits on forecasting skill of rare events. *Wea. Forecasting*, **28**, 525–534, https://doi.org/10.1175/WAF-D-12-00113.1.

Holloway, C. E., S. J. Woolnough, and G. M. Lister, 2012: Precipitation distributions for explicit versus parametrized convection in a large-domain high-resolution tropical case study. *Quart. J. Roy. Meteor. Soc.*, **138**, 1692–1708, https://doi.org/10.1002/qj.1903.

Hu, H., Z. Feng, and L. Y. R. Leung, 2021: Linking flood frequency with mesoscale convective systems in the US. *Geophys. Res. Lett.*, **48**, e2021GL092546, https://doi.org/10.1029/2021GL092546.

Iyer, E. R., A. J. Clark, M. Xue, and F. Kong, 2016: A comparison of 36–60-h precipitation forecasts from convection-allowing and convection-parameterizing ensembles. *Wea. Forecasting*, **31**, 647–661, https://doi.org/10.1175/WAF-D-15-0143.1.

Jankov, I., and Coauthors, 2017: A performance comparison between multiphysics and stochastic approaches within a North American RAP ensemble. *Mon. Wea. Rev.*, **145**, 1161–1179, https://doi.org/10.1175/MWR-D-16-0160.1.

——, J. Beck, J. Wolff, M. Harrold, J. B. Olson, T. Smirnova, C. Alexander, and J. Berner, 2019: Stochastically perturbed parameterizations in an HRRR-based ensemble. *Mon. Wea. Rev.*, **147**, 153–173, https://doi.org/10.1175/MWR-D-18-0092.1.

Jeworrek, J., G. West, and R. Stull, 2019: Evaluation of cumulus and microphysics parameterizations in WRF across the convective gray zone. *Wea. Forecasting*, **34**, 1097–1115, https://doi.org/10.1175/WAF-D-18-0178.1.

Johnson, M., N. Snook, J. Park, M. Xue, K. A. Brewster, T. Supinie, and X.-M. Hu, 2025: Severe weather verification of an FV3-LAM regional ensemble during the 2022 NOAA Hazardous Weather Testbed Spring Forecasting Experiment. *Wea. Forecasting*, **40**, 577–592, https://doi.org/10.1175/WAF-D-24-0034.1.

Joliffe, I., and D. E. Stephenson, 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* 2nd ed. John Wiley and Sons, 304 pp.

Kiel, B. M., W. A. Gallus, K. J. Franz, and N. Erickson, 2022: A preliminary examination of warm season precipitation displacement errors in the Upper Midwest in the HRRRE and HREF ensembles. *J. Hydrometeor.*, **23**, 1007–1024, https://doi.org/10.1175/JHM-D-21-0076.1.

Kong, R., M. Xue, E. R. Mansell, C. Liu, and A. O. Fierro, 2024: Assimilation of GOES-R Geostationary Lightning Mapper flash extent density data in GSI 3DVar, EnKF, and hybrid En3DVar for the analysis and short-term forecast of a supercell storm case. *Adv. Atmos. Sci.*, **41**, 263–277, https://doi.org/10.1007/s00376-023-2340-2.

Lee, C., K. A. Brewster, N. Snook, P. Spencer, and J. Park, 2024: Spatial aligned mean: A method to improve consensus forecasts of precipitation from convection-allowing model ensembles. *Wea. Forecasting*, **39**, 1545–1558, https://doi.org/10.1175/WAF-D-23-0229.1.

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.

Ma, S., C. Chen, H. He, D. Wu, and C. Zhang, 2018: Assessing the skill of convection-allowing ensemble forecasts of precipitation by optimization of spatial-temporal neighborhoods. *Atmosphere*, **9**, 43, https://doi.org/10.3390/atmos9020043.

Maddox, R. A., C. F. Chappell, and L. R. Hoxit, 1979: Synoptic and meso-α scale aspects of flash flood events. *Bull. Amer. Meteor. Soc.*, **60**, 115–123, https://doi.org/10.1175/1520-0477-60.2.115.

Mansell, E. R., C. L. Ziegler, and E. C. Bruning, 2010: Simulated electrification of a small thunderstorm with two-moment bulk microphysics. *J. Atmos. Sci.*, **67**, 171–194, https://doi.org/10.1175/2009JAS2965.1.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303, https://doi.org/10.1071/ES82036.

Molinari, J., and M. Dudek, 1992: Parameterization of convective precipitation in mesoscale numerical models: A critical review. *Mon. Wea. Rev.*, **120**, 326–344, https://doi.org/10.1175/1520-0493(1992)120<0326:POCPIM>2.0.CO;2.

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.

Nelson, B. R., O. P. Prat, D. J. Seo, and E. Habib, 2016: Assessment and implications of NCEP Stage IV quantitative precipitation estimates for product intercomparisons. *Wea. Forecasting*, **31**, 371–394, https://doi.org/10.1175/WAF-D-14-00112.1.

Nelson, J. A., A. J. Ray, J. Correia, K. Harnos, and S. Trojniak, 2022: Successes of the NOAA Hydrometeorological Testbed: Transitioning research to operations for predicting hydrometeorological extremes. *2020 Fall Meeting*, Chicago, IL, Amer. Geophys. Union, Abstract H42E-1349.

Niu, G.-Y., and Coauthors, 2011: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *J. Geophys. Res.*, **116**, D12109, https://doi.org/10.1029/2010JD015139.

Park, H., J. Hwang, D.-H. Cha, M.-I. Lee, C.-K. Song, J. Kim, S.-H. Park, and D.-K. Lee, 2024: Does a scale-aware convective parameterization scheme improve the simulation of heavy rainfall events? *J. Geophys. Res. Atmos.*, **129**, e2023JD039407, https://doi.org/10.1029/2023JD039407.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:44 PM UTC

Park, J., M. Xue, and C. Liu, 2023: Implementation and testing of radar data assimilation capabilities within the joint effort for data assimilation integration framework with ensemble transformation Kalman filter coupled with FV3-LAM model. *Geophys. Res. Lett.*, **50**, e2022GL102709, https://doi.org/10.1029/2022GL102709.

——, ——, ——, R. Kong, K. A. Brewster, N. A. Snook, and A. H. Berrington, 2025: Testing and evaluation of radar reflectivity and GOES-R lightning data assimilation capabilities in JEDI LETKF for RRFS-like configurations during the 2024 Summer Hydrometeorology Testbed Forecast Experiment. *15th Conf. on Transition of Research to Operations (15R2O)*, New Orleans, LA, Amer. Meteor. Soc., 9A.6, https://ams.confex.com/ams/105ANNUAL/meetingapp.cgi/Paper/456334.

Qiao, F., and X. Z. Liang, 2017: Effects of cumulus parameterization closures on simulations of summer precipitation over the continental United States. *Climate Dyn.*, **49**, 225–247, https://doi.org/10.1007/s00382-016-3338-6.

Roberts, B., B. Gallo, I. Jirak, and A. Clark, 2019: The High Resolution Ensemble Forecast (HREF) system: Applications and performance for forecasting convective storms. *2019 Fall Meeting*, San Francisco, CA, Amer. Geophys. Union, Abstract A31O-2797.

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.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, https://doi.org/10.1175/2008WAF2222159.1.

Schumacher, R. S., and R. H. Johnson, 2008: Mesoscale processes contributing to extreme rainfall in a midlatitude warm-season flash flood. *Mon. Wea. Rev.*, **136**, 3964–3986, https://doi.org/10.1175/2008MWR2471.1.

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.

Skamarock, W. C., J. B. Klemp, M. G. Duda, L. D. Fowler, S. H. Park, and T. D. Ringler, 2012: A multiscale nonhydrostatic atmospheric model using centroidal Voronoi tesselations and C-grid staggering. *Mon. Wea. Rev.*, **140**, 3090–3105, https://doi.org/10.1175/MWR-D-11-00215.1.

Smirnova, T. G., J. M. Brown, S. G. Benjamin, and J. S. Kenyon, 2016: Modifications to the Rapid Update Cycle land surface model (RUC LSM) available in the Weather Research and Forecasting (WRF) Model. *Mon. Wea. Rev.*, **144**, 1851–1865, https://doi.org/10.1175/MWR-D-15-0198.1.

Smith, A. B., and R. W. Katz, 2013: US billion-dollar weather and climate disasters: Data sources, trends, accuracy and biases. *Nat. Hazards*, **67**, 387–410, https://doi.org/10.1007/s11069-013-0566-5.

Snook, N., F. Kong, K. A. Brewster, M. Xue, K. W. Thomas, T. A. Supinie, S. Perfater, and B. Albright, 2019: 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. *Wea. Forecasting*, **34**, 781–804, https://doi.org/10.1175/WAF-D-18-0155.1.

——, ——, A. Clark, B. Roberts, K. A. Brewster, and M. Xue, 2020: Comparison and verification of point-wise and patch-wise localized probability-matched mean algorithms for ensemble consensus precipitation forecasts. *Geophys. Res. Lett.*, **47**, e2020GL087839, https://doi.org/10.1029/2020GL087839.

——, and Coauthors, 2025: Evaluation of CAPS convection-allowing FV3-LAM ensembles during the 2022 HWT Spring Forecasting Experiment to inform the design of the Rapid Refresh Forecast System (RRFS). *Wea. Forecasting*, **40**, 857–874, https://doi.org/10.1175/WAF-D-23-0217.1.

Sun, X., M. Xue, J. Brotzge, R. A. McPherson, X. M. Hu, and X. Q. Yang, 2016: An evaluation of dynamical downscaling of central plains summer precipitation using a WRF-based regional climate model at a convection-permitting 4 km resolution. *J. Geophys. Res. Atmos.*, **121**, 13 801–13 825, https://doi.org/10.1002/2016JD024796.

Supinie, T. A., J. Park, N. Snook, X. M. Hu, K. A. Brewster, M. Xue, and J. R. Carley, 2022: Cool-season evaluation of FV3-LAM-based CONUS-scale forecasts with physics configurations of experimental RRFS ensembles. *Mon. Wea. Rev.*, **150**, 2379–2398, https://doi.org/10.1175/MWR-D-21-0331.1.

Terti, G., I. Ruin, S. Anquetin, and J. J. Gourley, 2017: A situation-based analysis of flash flood fatalities in the United States. *Bull. Amer. Meteor. Soc.*, **98**, 333–345, https://doi.org/10.1175/BAMS-D-15-00276.1.

Thompson, G., R. M. Rasmussen, and K. Manning, 2004: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part I: Description and sensitivity analysis. *Mon. Wea. Rev.*, **132**, 519–542, https://doi.org/10.1175/1520-0493(2004)132<0519:EFOWPU>2.0.CO;2.

Trojniak, S., and J. Correia, 2024: 2024 Flash Flood and Intense Rainfall (FFaIR) final report. NOAA Weather Prediction Center Tech. Rep., 114 pp., https://www.wpc.ncep.noaa.gov/hmt/Reports/FFaIR/2024_FFaIR_Final_Report.pdf.

Wang, W., and N. L. Seaman, 1997: A comparison study of convective parameterization schemes in a mesoscale model. *Mon. Wea. Rev.*, **125**, 252–278, https://doi.org/10.1175/1520-0493(1997)125<0252:ACSOCP>2.0.CO;2.

Zhang, C., and Coauthors, 2019: How well does an 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, Y., J. K. Roundy, and J. A. Santanello, 2021: Evaluating the impact of model resolutions and cumulus parameterization on precipitation in NU-WRF: A case study in the Central Great Plains. *Environ. Modell. Software*, **145**, 105184, https://doi.org/10.1016/j.envsoft.2021.105184.

Zhou, X., and Coauthors, 2022: The development of the NCEP Global Ensemble Forecast System version 12. *Wea. Forecasting*, **37**, 1069–1084, https://doi.org/10.1175/WAF-D-21-0112.1.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:44 PM UTC