Spatial Aligned Mean: A Method to Improve Consensus Forecasts of Precipitation from Convection-Allowing Model Ensembles

CHANGJAE LEE,[a] KEITH A. BREWSTER,[b,c] NATHAN SNOOK,[b] PHILLIP SPENCER,[b] AND JUN PARK[b]

[a] *Korea Meteorological Administration, Seoul, South Korea*
[b] *Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma*
[c] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*

(Manuscript received 20 December 2023, in final form 24 July 2024, accepted 20 August 2024)

ABSTRACT: Ensembles of convection-allowing model (CAM) forecasts are increasingly being used in operational numerical weather forecasting. Several approaches have been devised to find consensus among ensemble forecast fields, including the arithmetic ensemble mean and, more recently, the patchwise localized probability-matched (LPM) mean. However, differences in spatial distribution and intensity of precipitation features among ensemble members make it difficult to construct an ensemble mean product that characterizes the consensus while preserving precipitation structures forecasted by the individual ensemble members. To overcome this problem, this study aims to develop and test a method for improving ensemble consensus precipitation forecasts by directly considering the spatial offsets among ensemble members. This study uses a multiscale spatial alignment technique to align the precipitation features of each ensemble member to a common location, and the spatial aligned mean (SAM) is obtained by averaging the realigned members. It is shown that implementing SAM and subsequently applying the LPM technique to the average of all aligned members (SAM–LPM) can significantly improve the warm season precipitation forecast scores using common metrics such as equitable threat score (ETS). Also, improvement in the structure of features of heavy rainfall is shown from summer 2023 flash-flooding cases. Thus, SAM and SAM–LPM can be excellent candidate methods for calculating an ensemble consensus and providing ensemble consensus guidance to forecasters.

SIGNIFICANCE STATEMENT: High-impact rainfall events, such as flash floods, result in many billion-dollar loss events in the United States each year. This study seeks to improve the prediction of such events when using guidance from convection-allowing model (CAM) ensemble forecasts, such as the U.S. operational High-Resolution Ensemble Forecast (HREF) and the nascent Rapid Refresh Forecast System (RRFS). The proposed method, the spatial aligned mean (SAM), directly addresses the common issue of disparity in the predicted location of convective systems among ensemble members that confounds traditional ensemble consensus methods. In this study, it is found that SAM improves ensemble consensus guidance for high-impact rainfall events in both the HREF and the Center for Analysis and Prediction of Storms (CAPS) Finite-Volume Cubed-Sphere (FV3)-limited area model (LAM) CAM ensemble forecast system, a proxy for the future RRFS.

KEYWORDS: Ensembles; Forecast verification/skill; Mesoscale forecasting; Numerical weather prediction/forecasting; Operational forecasting; Postprocessing

1. Introduction

Due to initial condition uncertainties and limited predictability inherent in an individual high-resolution numerical weather prediction (NWP) model forecast, ensembles of convection-allowing model (CAM) forecasts are increasingly being used in operational numerical weather forecasting. Such CAM ensembles include the U.S. High-Resolution Ensemble Forecast (HREF; Roberts et al. 2019) and the nascent Rapid Refresh Forecast System (RRFS; Alexander and Carley 2023). It is common to have errors in convection initiation location and timing and in storm motion in NWP models (e.g., Clark et al. 2012) which lead to variations in the predicted locations of storms among CAM ensemble member forecasts. Several approaches, such as consensus average methods and clustering (e.g., fuzzy clustering; Zheng et al. 2017), have been devised to utilize a large set of ensemble outputs and present an ensemble consensus to the forecaster or end user.

The ensemble mean, a simple point-wise arithmetic average of ensemble members, is commonly used in operational ensembles. However, due to the difference in spatial distribution and intensity of precipitation features in each ensemble member, the arithmetic ensemble mean of precipitation forecasts tends to reduce the magnitude of forecast maxima while expanding the areal coverage of light precipitation (Ebert 2001; Surcel et al. 2014). This tendency thus masks the most impactful precipitation values and alters the ensemble probability density function (PDF), while the smoothing introduced by the ensemble mean can artificially inflate certain validation metrics (e.g., Snook et al. 2019).

The probability-matched (PM; Ebert 2001) ensemble mean, pointwise localized PM (LPM) mean (Clark 2017), and patchwise

Denotes content that is immediately available upon publication as open access.

Corresponding author: ChangJae Lee, changjae.lee.3789@gmail.com

DOI: 10.1175/WAF-D-23-0229.1

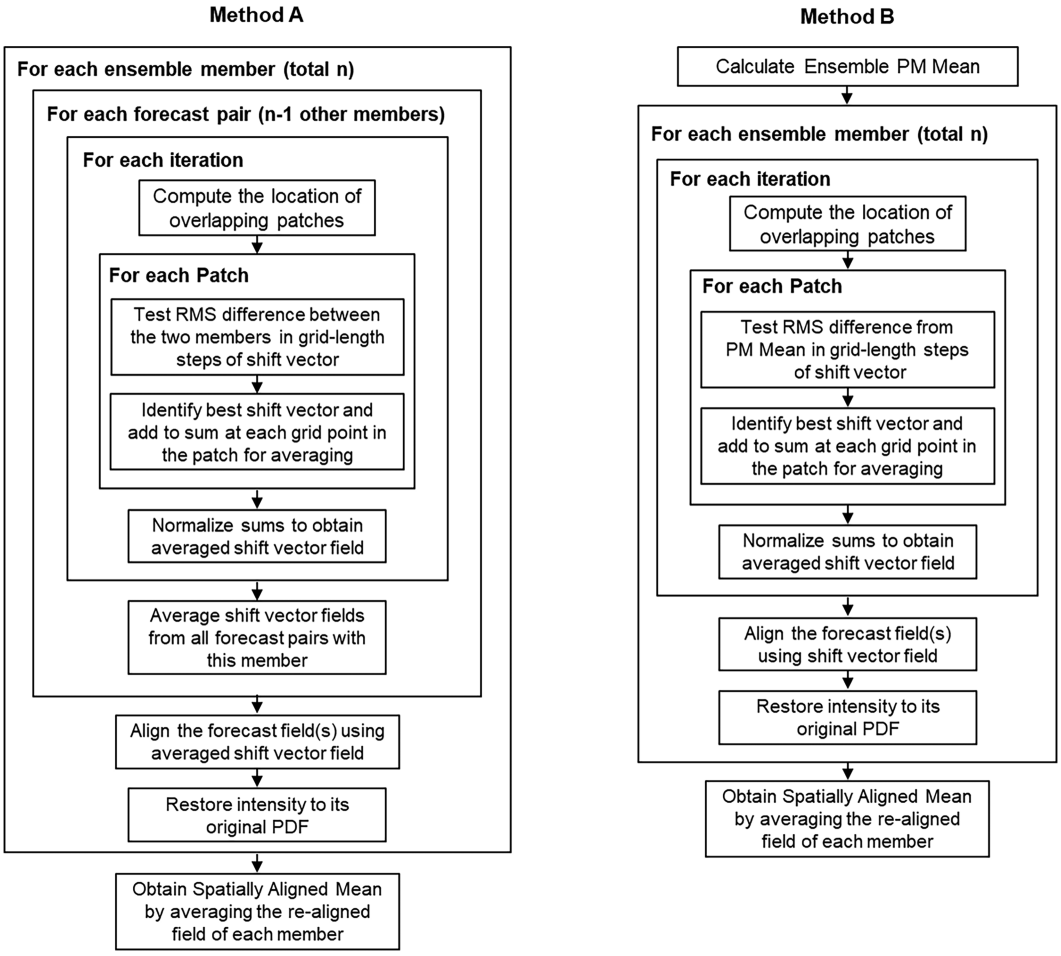

FIG. 1. Flowchart for obtaining SAM [(left) method A and (right) method B].

LPM mean (Snook et al. 2020) methods have been introduced to overcome these problems. The PM and LPM means use a PDF from all of the ensemble members to preserve the ensemble forecast's maxima. PM and LPM methods redistribute the values of each grid point of the ensemble mean to have the same distribution to entire members' PDF, with the local methods considering the PDF in local subdomains rather than over the entire forecast domain. While these methods can preserve the maximum values, they are not directly addressing the spatial differences among the members; thus, unnatural structures can be created.

Since it is known that position displacement accounts for a significant portion of the error variance in NWP (Jankov et al. 2021), this study aims to find a way to improve ensemble consensus precipitation by directly considering the spatial offsets among ensemble members. Similar approaches have been tried in the fields of data assimilation and postprocessing under different names, such as phase-error correction (Brewster 2003a,b), field coalescence (Ravela 2012), feature alignment technique (Stratman et al. 2018; Stratman and Potvin 2022), and feature-oriented mean (Feng et al. 2020), so this study adapts one of those methods to find consensus among the precipitation forecasts in operational and experimental real-time CAM ensemble forecasts.

Specifically, this study adapts algorithms of the phase-correcting data assimilation method (Brewster 2003a,b), which can apply to multiple scales of discontinuous fields, to align the fields of each ensemble member to a common location. Offsets are found for each ensemble member with respect to other members or with respect to the original PM mean. The fields of each member are shifted by its aligning offset vectors, moving the precipitation to a common location, and the spatial aligned mean (SAM) is obtained by averaging the realigned members.

In this work, the SAM method is described in section 2, and the SAM is applied to the operational HREF and a real-time experimental CAM ensemble used for the Hydrometeorology Testbed Flash Flood and Intense Rainfall (HMT-FFaIR) experiment, described in section 3. Sections 4 and 5 present the verification of the experiments, while summary and conclusions are drawn in section 6.

2. Methods

The algorithm used in SAM consists of three parts: 1) The shift vector field that aligns each member's precipitation field

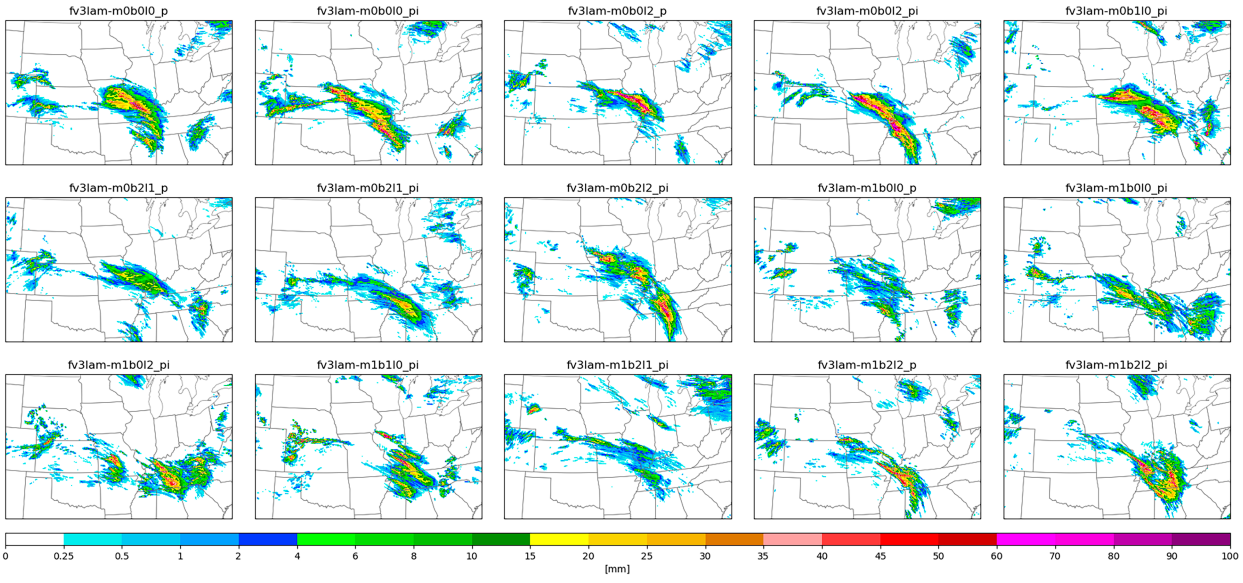

FIG. 2. The 6-h rainfall (mm) of CAPS FV3-LAM ensemble result for 2023 midsouth U.S. flood event (84-h forecast valid at 1200 UTC 3 Aug 2023).

is found, 2) each member is thus repositioned, and 3) finally, ensemble consensus is found from the realigned members.

The method used for finding the shift vector field is based on the alignment algorithm in the phase-correcting data assimilation method (Brewster 2003a) because it has demonstrated success in adjusting storm-scale fields across multiple scales in the data assimilation context and (Brewster 2003b) is simple to apply and can easily be parallelized. Applied in the context of aligning ensemble members, the algorithm aims to minimize a squared difference sum of the output from a pair of forecasts by translating the targeted variables incrementally in the model x and y directions. Since spatial offsets can vary across the model domain, the algorithm proceeds by dividing the domain into overlapping patches. The patch size is flexible and can be set while considering the horizontal scale of errors to account for spatial offsets in, say, synoptic-, mesoalpha-, and mesobeta-scale features.

For each test patch, shift vectors are determined by finding the offset of grid points in the x/y directions, which minimizes the root-mean-square (RMS) differences in fields from two forecasts, either two individual ensemble members or one member and a mean, including a penalty for a large offset distance:

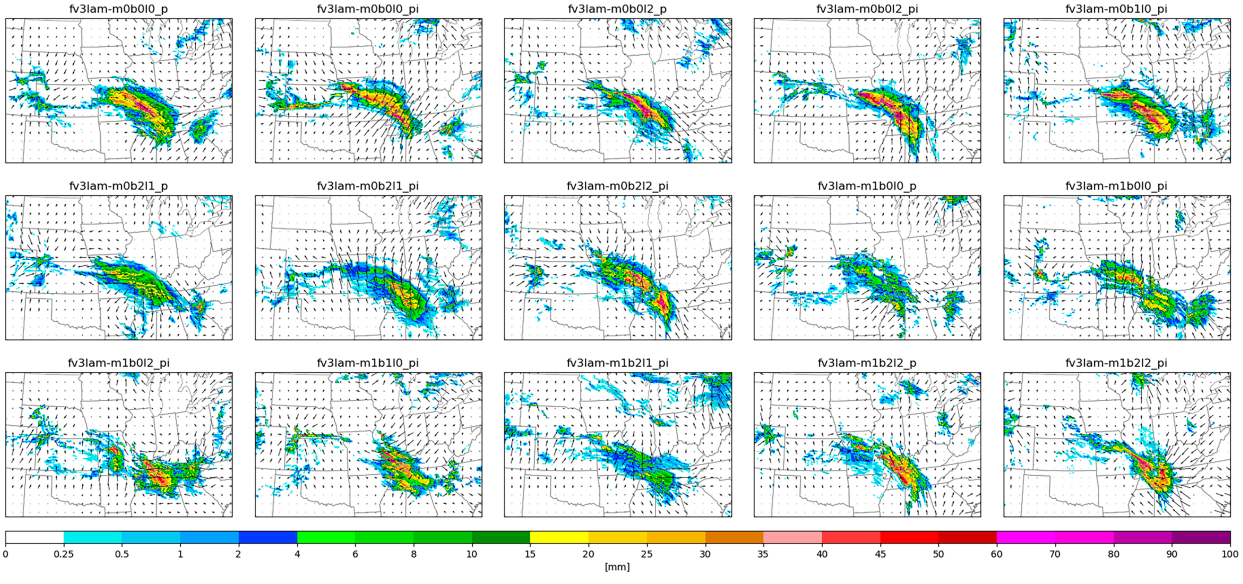

FIG. 3. The 6-h rainfall (mm) of phase-shifted member-to-mean aligned fields of CAPS FV3-LAM ensemble result with shift vectors for 2023 midsouth U.S. flood event (84-h forecast valid at 1200 UTC 3 Aug 2023).

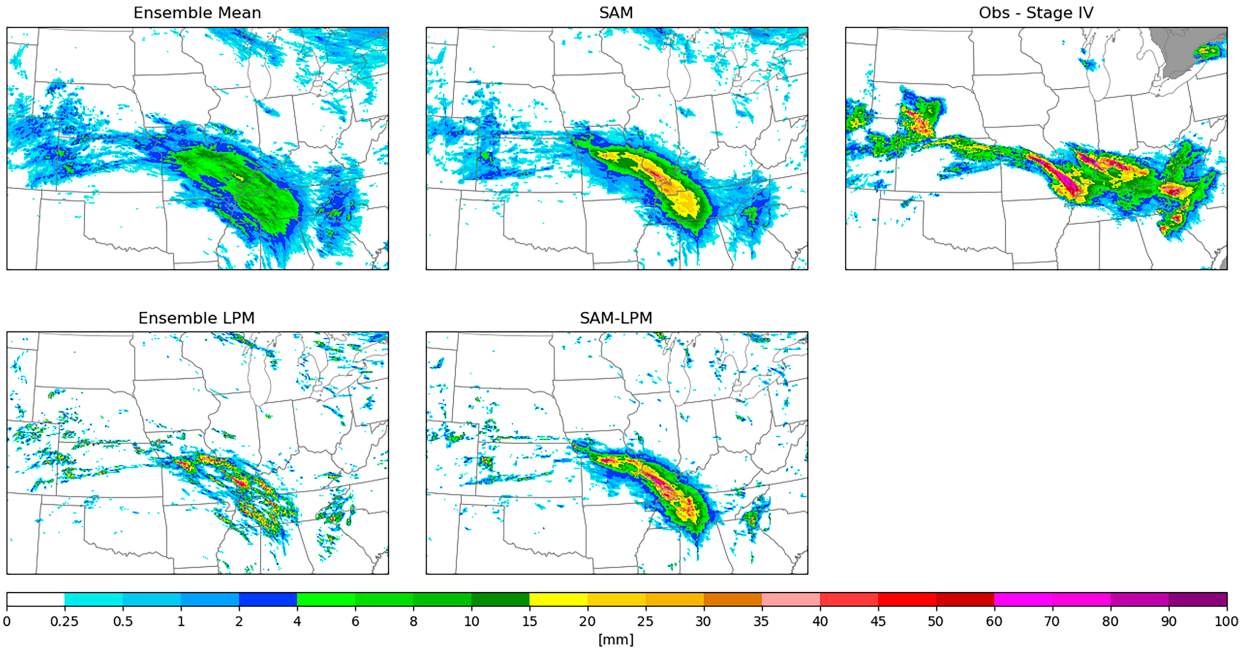

FIG. 4. The 6-h rainfall (mm) of ensemble mean, member-to-mean aligned SAM, LPM, and member-to-mean aligned SAM–LPM fields for CAPS FV3-LAM ensemble and observation (stage IV) for 2023 midsouth U.S. flood event (84-h forecast valid at 1200 UTC 3 Aug 2023).

$$J(\delta\mathbf{x}) = \frac{s(|\delta\mathbf{x}|l^{-1})}{m}\sum_{i=1}^{m}[F_a(\mathbf{x}_i + \delta\mathbf{x}) - F_b(\mathbf{x}_i)]^2, \quad (1)$$

where F_a and F_b are the pair forecasts, $\mathbf{x}$ is a grid point in the patch with a total m number of grid points, and $\delta\mathbf{x}$ is the horizontal displacement vector.

The multiplier on the right-hand side s is a penalty function for the large displacement, using the inverse of the second-order autoregressive (SOAR) function (Thiebaux et al. 1990):

$$s(|\delta\mathbf{x}|l^{-1}) = \frac{\exp(|\delta\mathbf{x}|l^{-1})}{(1 + |\delta\mathbf{x}|l^{-1})}, \quad (2)$$

where l is a length scale parameter, which is a function of patch size in the x and y directions L_x and L_y and the scale parameter α.

$$l = \alpha\sqrt{L_x^2 + L_y^2}, \quad (3)$$

where $\alpha > 0$, and smaller α penalizes more for larger spatial displacements. This method can also be applied to multiple variables with weighted sums depending on the expected error of each variable. Here, we evaluate Eq. (1) using the forecasted precipitation fields only.

The entire domain's shift vector field is obtained by averaging all the overlapping test patches' shift vectors $\delta\mathbf{x}$. An

TABLE 1. Details of the membership of the CAPS CAM ensemble used in the 2023 FFaIR experiment.

Experiment	Microphysics	PBL	Surface	LSM	IC/LBC (like system)	AI member
GFS IC for baseline configuration						
M0B0L0_PG	Thompson	MYNN	MYNN	NOAH	GFS/GFS	AI-1
M1B0L0_PG	NSSL	MYNN	MYNN	NOAH	GFS/GFS (WoF)	AI-2
M0B0L2_PG	Thompson	MYNN	MYNN	RUC	GFS/GFS (RRFSm1)	AI-3
M1B2L2_PG	NSSL	TKE-EDMF	GFS	RUC	GFS/GFS (RRFSmphys8)	
M0B2L1_PG	Thompson	TKE-EDMF	GFS	NOAHMP	GFS/GFS (GFSv16)	AI-4
Physics + IC perturbation ensemble						
M0B0L2_PI	Thompson	MYNN	MYNN	RUC	GEFS_m1	
M0B1L0_PI	Thompson	Shin-Hong	GFS	NOAH	GEFS_m2	
M0B2L1_PI	Thompson	TKE-EDMF	GFS	NOAHMP	GEFS_m3	
M0B0L0_PI	Thompson	MYNN	MYNN	NOAH	GEFS_m4	
M0B2L2_PI	Thompson	TKE-EDMF	GFS	RUC	GEFS_m5	
M1B0L2_PI	NSSL	MYNN	MYNN	RUC	GEFS_m6	
M1B1L0_PI	NSSL	Shin-Hong	GFS	NOAH	GEFS_m7	
M1B2L1_PI	NSSL	TKE-EDMF	GFS	NOAHMP	GEFS_m8	
M1B0L0_PI	NSSL	MYNN	MYNN	NOAH	GEFS_m9	
M1B2L2_PI	NSSL	TKE-EDMF	GFS	RUC	GEFS_m10	

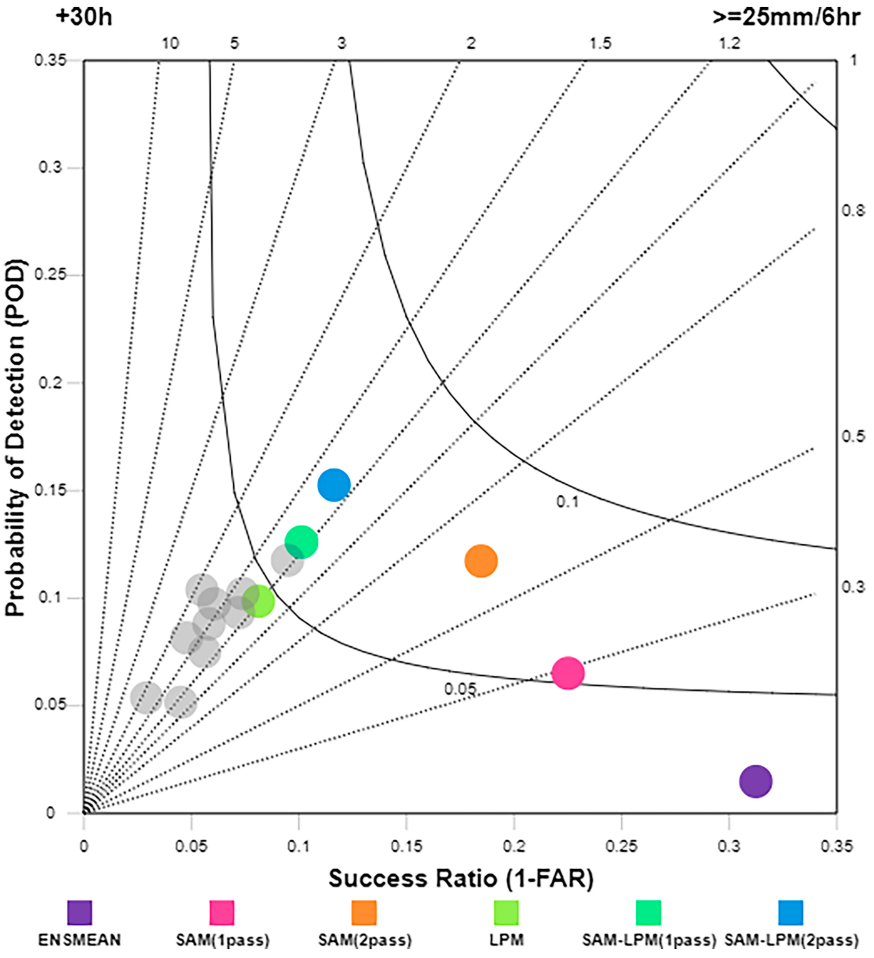

FIG. 5. Performance diagram for SAM and SAM–LPM (both using member-to-mean alignment) at the 25 mm $(6\ \text{h})^{-1}$ rainfall threshold at the 24–30-h forecast lead time for HREF over the 2023 FFaIR period (gray dots are individual members).

iterative approach using a cascade of test patch sizes can be applied so that, first, large-scale and then small-scale position differences can be resolved.

This phase-correcting technique can be applied in either of two ways: 1) aligning all ensemble member-to-member pairs, hereafter called member-to-member alignment, $n \times (n-1)$ pairs with a total of n members and 2) aligning ensemble member-to-ensemble-mean pairs, hereafter called member-to-mean alignment, n pairs with a total of n members. Both methods are illustrated in Fig. 1 and described herein.

a. Member-to-member alignment (with a total of n members)

For each member, shift vectors are determined for all the other members in pairs, and the vectors are then averaged to find the offset to bring the fields to a central location. The jth member's shift vector field $\mathbf{D}_j$ is an averaged sum of obtained shift vector fields with all the other pairs:

$$\mathbf{D}_j = \frac{1}{n}\sum_{i=1}^{n} \mathbf{D}_{j,i}, \tag{4}$$

where $\mathbf{D}_{j,i}$ is the shift vector field of ith and jth member pairs. The total number of computations for calculating all the pairs is $O(n^2)$, which could become expensive for large n, although all pairings can be handled independently; thus, the process is naturally parallel.

b. Member-to-mean alignment (with a total of n members)

Shift vectors can be determined by comparing each member's field with an ensemble mean field. However, since RMS differences between the two fields are affected by not only the spatial offset but also the intensity differences, it is preferred to use the PM mean field instead of the arithmetic mean field as the arithmetic mean will commonly dilute field maxima. This method thus brings the spatial position of features in all the members close to the PM mean. The number of pairs to align is the same as the number of ensemble members, so the number of computations is $O(n)$, potentially offering significant computation time savings for large n, and is also naturally parallel after the computation of the PM mean.

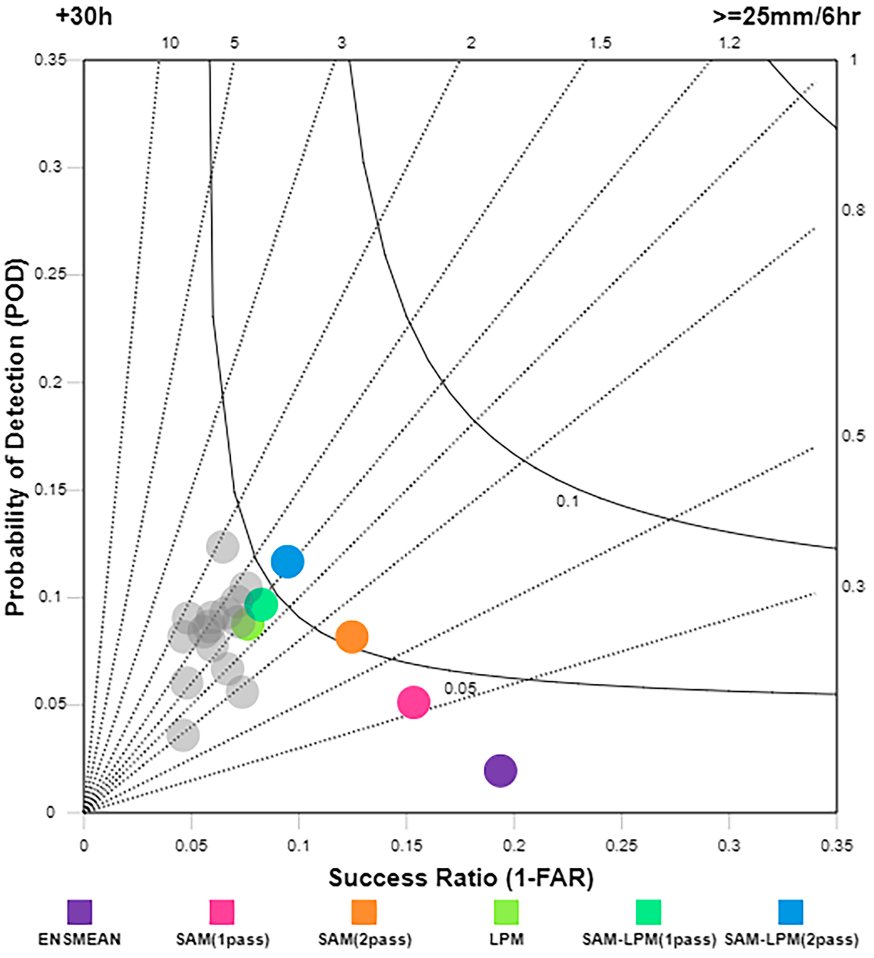

FIG. 6. Performance diagram for SAM and SAM–LPM (both using member-to-mean alignment) at the 25 mm (6 h)$^{-1}$ rainfall threshold at the 24–30-h forecast lead time of CAPS FV3-LAM ensemble over the 2023 FFaIR period (gray dots are individual members).

Once the shift vectors of each member are calculated for each patch and locally averaged, each member forecast is repositioned by using those vectors and linear interpolation since averaged vectors are generally noninteger grid offsets. After moving the field using the shift vectors, the forecast values of each member are reassigned using the PDF of that original forecast field to restore maxima. This is necessary because moving the grid involves interpolation, which has some smoothing effect, generally resulting in a reduction in maxima in the field. Either PM or LPM methods can be utilized for rescaling the distribution of each member forecast at that point. After the PDF restoration process, the spatial aligned mean is obtained by a simple average of the realigned and rescaled members.

To illustrate this, we present an example of member-to-mean alignment applied to CAM ensemble models. Figure 2 is the 6-h precipitation accumulation with 84-h lead time from the individual members of a real-time experimental CAM ensemble [Center for Analysis and Prediction of Storms (CAPS) Finite-Volume Cubed-Sphere (FV3)-limited area model (LAM) ensemble, detailed in section 5] for the 2023 midsouth U.S. flood event, while Fig. 3 shows the result after the phase-shift algorithm has been applied using the shift vectors. Figure 4 shows the forecast consensus precipitation field from SAM (middle top) and from SAM with the patchwise LPM technique applied (SAM–LPM; middle bottom). Figures 2–4 are plotted based on the results from the experiment described in section 3.

Due to the difference in the location of storms in every member, the arithmetic ensemble mean (top left in Fig. 4) field has spread the precipitation area widely, and LPM (bottom left in Fig. 4) shows scattered precipitation features. On the other hand, SAM and SAM–LPM results have a more cohesive structure of the mesoscale convective complex path, which more closely matches the corresponding observed rainfall from stage IV analysis (rightmost panel in Fig. 4).

3. Experimental design

In this experiment, the SAM technique is applied to the 6-h accumulated precipitation output from two high-resolution

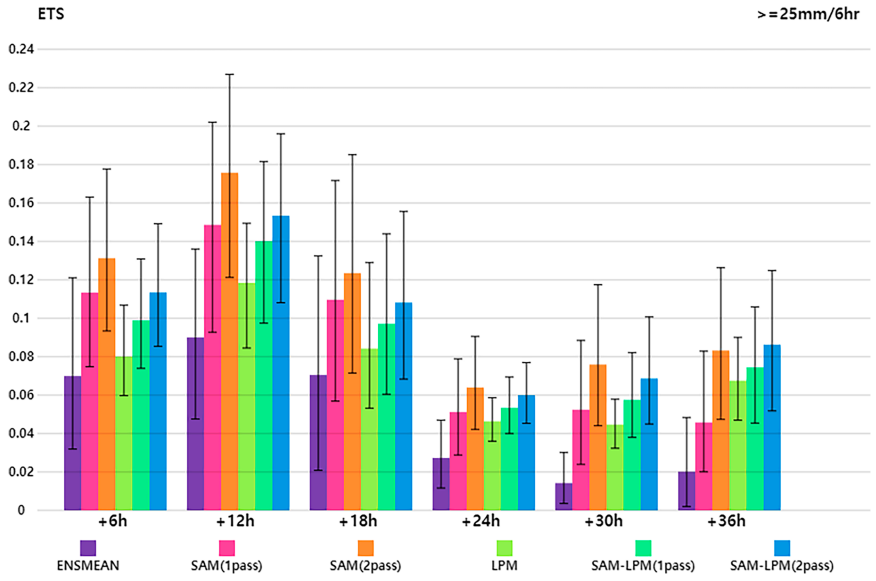

FIG. 7. ETS results of HREF at the 25 mm (6 h)$^{-1}$ rainfall threshold with 95% confidence interval using bootstrapping (10 000 samples) in variation with forecast lead time (from 0–6 to 30–36 h) over the 2023 FFaIR.

(~3 km) CAM ensembles: 1) the operational HREF, which has 10 members including the 12-h time-lagged members, and 2) a 15-member ensemble using the FV3-LAM (Black et al. 2021) designed by the CAPS. The FV3 serves as the dynamic core for the current version of the Unified Forecast System Medium-Range Weather Application or GFSv16 (Lin 2004; Yang et al. 2021). A convection-allowing ensemble configured using FV3-LAM members (i.e., RRFSv1) is in the latter stages of development and is scheduled to replace HREFv3 in 2025 as part of NOAA's Unified Forecast System vision; here, the CAPS CAM ensemble serves as a surrogate RRFS since RRFS is not yet operationally available. The membership of the CAPS CAM ensemble is described in Table 1, consisting of members with variations in physics options, including microphysics, planetary boundary layer physics, and land surface physics as well as initial and lateral boundary perturbations. For the period examined here, the CAPS CAM ensemble was run with initial conditions provided by the operational GFS and initial and lateral boundary condition perturbations from the operational GEFS.

In addition to the SAM results, the SAM with the patch-wise LPM technique subsequently applied (SAM–LPM) is examined to see if that combination can better preserve forecast maxima. The proposed SAM and SAM–LPM techniques are applied for lead times of 6–84 h (up to 36 h for HREF, which is the maximum available ensemble forecast extent due to shorter forecasts and use of time-lag members) of 0000 UTC runs over the contiguous United States (CONUS). Verification is performed against NOAA stage IV 4-km resolution precipitation data (Nelson et al. 2016). Testing is done for 29 days spanning 10 weeks (5–10 June, 12–13 June, 26–30 June, 10–14 July, 17–18 July, 31 July–4 August, and 7–10 August) in the summer of 2023 corresponding to the period of the HMT FFaIR experiment (Trojniak et al. 2024).

In this experiment, two cascading test patch sizes are used in the alignment algorithm. For the first pass, the patch size is 600 km (aiming to address synoptic scale offsets), and 225 km (for mesoalpha scale offsets) is used for the patch size in the second pass. Results of both SAM and SAM–LPM are evaluated after each of the iterations.

Pointwise verification and spatial feature verification are performed with several precipitation thresholds using standard algorithms within the Model Evaluation Tools (MET) program (Brown et al. 2021). The Method for Object-Based Diagnostic Evaluation (MODE; Davis et al. 2009) is used for spatial verification. For MODE verification, a 20-km convolution radius is used, and the weights used for MODE interest, a normalized weighted mean of MODE metrics ranging from 0 to 1, are as follows: centroid distance (distance between two object centroids, 20%), boundary distance (minimum distance between the boundaries of two objects, 40%), angle difference (difference between the axis angles of two objects, 10%), area ratio (the forecast object area divided by the observation object area, 10%), and intersection area ratio (ratio of intersection area to the lesser of the forecast and observation object areas, 20%).

4. Verification results

The verification is performed with several 6-h precipitation thresholds (1, 5, 10, 15, 20, and 25 mm) in order to evaluate performance for various rainfall values ranging from minimal to intense rainfall. The verification results discussed in this section are at the 25 mm (6 h)$^{-1}$ threshold to focus on the warm season's high-threshold rainfall events.

In section 4a, the overall tendency for SAM and SAM–LPM, in terms of common forecast skill metrics such as probability of detection (POD), success ratio, critical success index

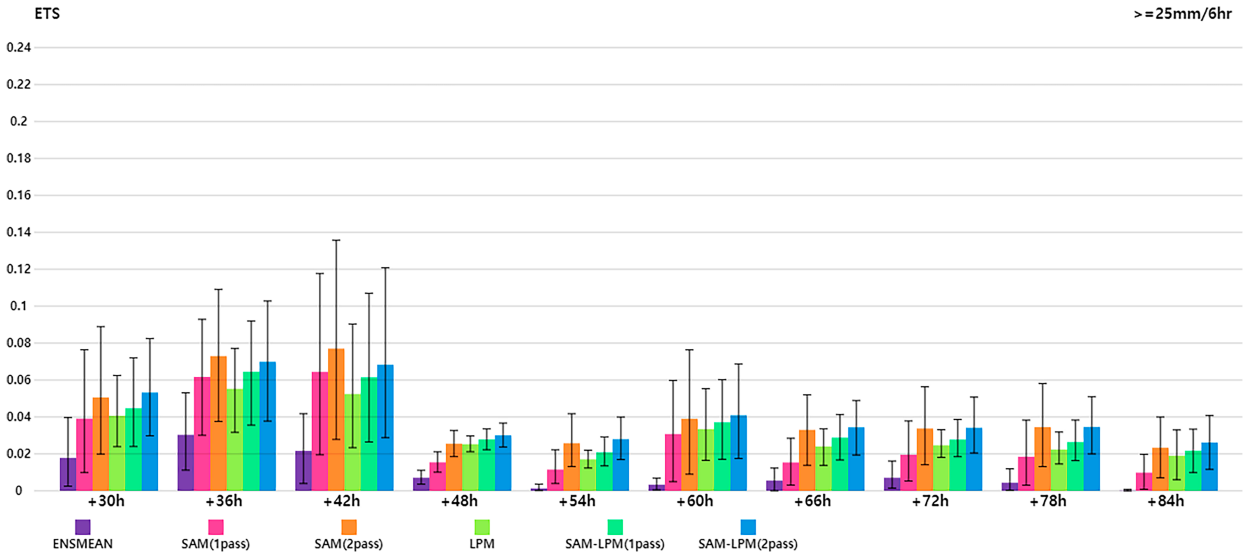

FIG. 8. ETS results of CAPS FV3-LAM ensemble at the 25 mm (6 h)$^{-1}$ rainfall threshold with 95% confidence interval using bootstrapping (10 000 samples) in variation with forecast lead time (from 24–30 to 78–84 h) over the 2023 FFaIR.

(CSI), and frequency bias, is shown with the performance diagram. Section 4b shows the verification and significance test results in variation with forecast lead time in terms of equitable threat score (ETS; Wilks 1995) metrics. Member-to-mean alignment using the original ensemble PM mean showed better results than member-to-member alignment, which is not shown. Therefore, the verification results described in this section are obtained from member-to-mean alignment.

a. Overall tendency

Figures 5 and 6 show performance diagrams (Roebber 2009) at the 25 mm (6 h)$^{-1}$ rainfall threshold and 24-to-30-h

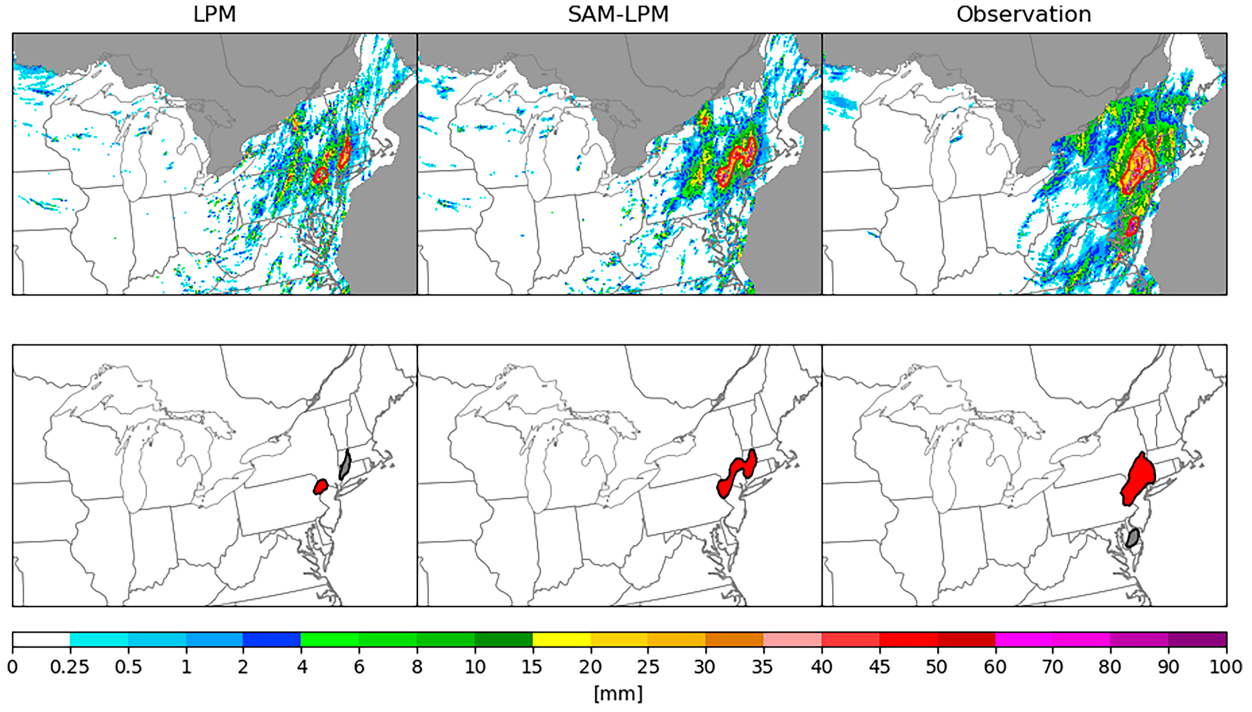

FIG. 9. (left) LPM, (center) SAM–LPM, and (right) stage IV observed (top) 6-h rainfall (mm) and (bottom) MODE objects [≥20 mm (6 h)$^{-1}$] for CAPS FV3-LAM ensemble 60-h forecast valid at 1200 UTC 16 Jul 2023.

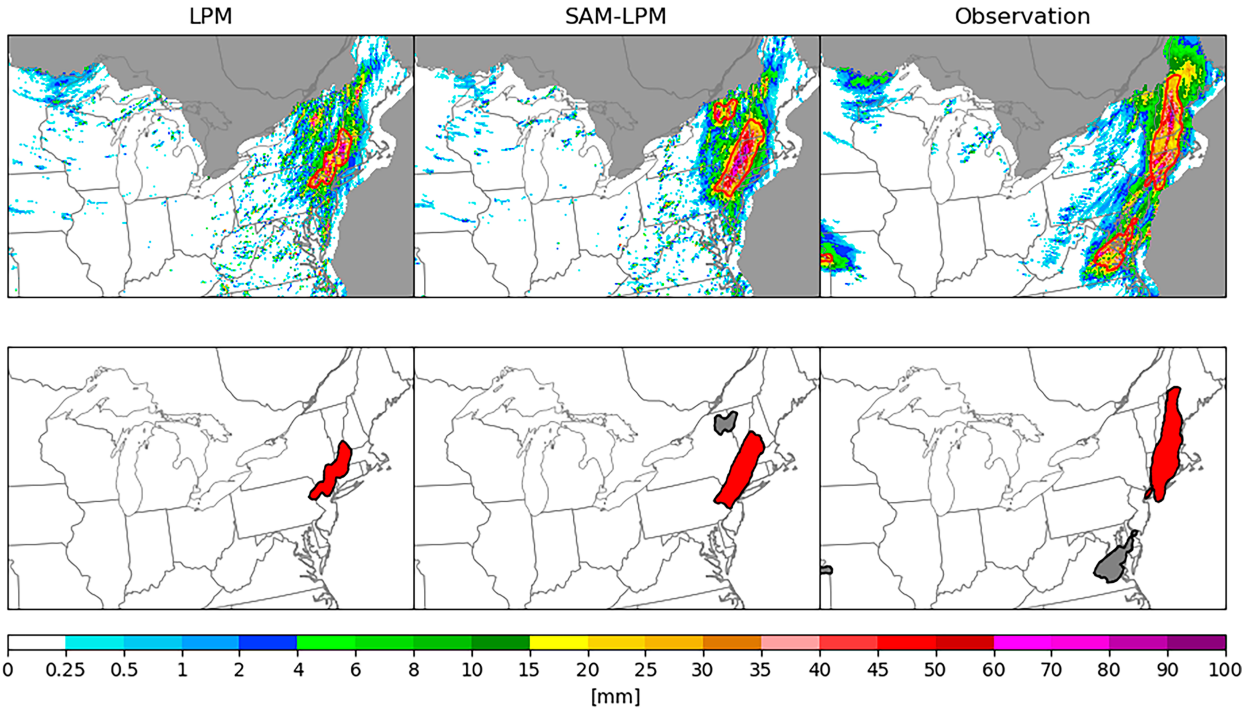

FIG. 10. (left) LPM, (center) SAM–LPM, and (right) stage IV observed (top) 6-h rainfall (mm) and (bottom) MODE objects [≥20 mm $(6\ h)^{-1}$] for the 2023 New Hampshire flood event for CAPS FV3-LAM ensemble 66-h forecast valid at 1800 UTC 16 Jul 2023.

forecast lead time over the 2023 FFaIR period for the HREF and CAPS FV3-LAM ensemble, respectively. Member-to-mean alignment is used for the results in these figures, but the overall tendencies in the results for the member-to-member method, not shown, are similar. The results of the arithmetic ensemble mean, SAM, standard LPM, SAM–LPM, and all individual members are plotted in this diagram. For SAM and SAM–LPM, both 1-pass, aligned for synoptic scale, and iterative 2-pass, aligned for both synoptic and mesoscale, results are included.

The use of SAM or SAM–LPM increases the POD (the *y* axis in the figure) compared to the arithmetic mean and standard LPM methods. The score is further improved in the 2-pass method, aligned for both synoptic and mesoscale offsets.

The success ratio, the *x* axis in the figure, for SAM decreases compared to the arithmetic mean, but in light of its improved POD, the net result has a better CSI (curved lines in the figure). However, the success ratio for SAM–LPM increases compared to the standard LPM, and this result, along with increased POD, leads to improved CSI.

Frequency bias, dotted lines in the figure, is higher for SAM (and close to unity, ideal) than that of the arithmetic mean, which significantly underforecasts rainfall at this threshold. The frequency bias for SAM–LPM is similar to the standard LPM because both of them use the same PDF.

It is notable that SAM–LPM (2 pass) improves POD, CSI, and success ratio compared to all individual members and frequency bias is improved relative to a majority of the members. Overall tendencies shown in the figures are the same in other forecast lead times, not shown, while the scores naturally drop with increased lead time. The overall scores for HREF are higher than those for the CAPS FV3-LAM ensemble for this experiment; it is beyond the scope of this paper to delve into the causes of that, but mature, better-tuned member models and diversity in the dynamic cores of the membership are likely contributors to that result.

The results are not included here, but 10, 15, and 20 mm over 6-h threshold results show the same trends, while lower thresholds, 1 and 5 mm over 6 h, show that SAM has lower POD and higher success ratio results compared to those for ensemble mean. As mentioned in the introduction, these results stem from the ensemble mean broadening the areal coverage of light precipitation due to the spatial offsets among members.

b. *Evaluation results in variation with forecast lead time and significance tests*

Figures 7 and 8 show the variation of ETS with forecast lead time at the 25 mm $(6\ h)^{-1}$ rainfall threshold over the 2023 FFaIR period, for both the operational HREF and CAPS FV3-LAM ensembles, respectively. Member-to-mean alignment toward the original ensemble PM mean is used for the results in these figures. SAM and SAM–LPM show improvement in ETS for both ensembles compared to arithmetic ensemble mean and standard LPM for all forecast lead times. Also, the results show that the iterative 2-pass method, aligned for both synoptic and mesoscale, improves ETS more than the 1-pass method, aligned only for synoptic scale.

Beyond the 48-h forecast lead time, the score of the arithmetic mean drops (Fig. 8). This occurs because the spatial offsets among members grow in part due to variations in outflow

TABLE 2. The scores for selected MODE metrics [≥20 mm (6 h)$^{-1}$; red object pairs in Figs. 9 and 10] and ETS metrics for LPM and SAM–LPM in the New Hampshire Flood event for the CAPS FV3-LAM ensemble 60- and 66-h forecast valid at 1200 and 1800 UTC 16 Jul 2023.

	60-h lead time		66-h lead time	
	LPM	SAM–LPM	LPM	SAM–LPM
MODE interest	0.8952	0.9828	0.9060	0.9464
Angle difference (°)	12.80	16.45	24.05	16.20
Area ratio	0.14	0.68	0.47	0.81
Intersection area (km^2)	210	705	237	405
ETS	0.084	0.124	0.068	0.086

from storms on day 1, adding uncertainty to storm initiation details in the afternoon and evening of day 2. SAM and SAM–LPM results beyond 48 h show that the spatial alignment technique can improve the ETS score and recover some of the lost skills.

Some SAM results have better ETS scores than SAM–LPM, especially at the beginning of the forecasts. As shown in section 4a, SAM–LPM has better POD and frequency biases but a lower success ratio than SAM, which leads to higher ETS for some of the SAM results.

For the significance tests (Hamill 1999), 95% confidence intervals obtained from bootstrapping with 10 000 samples are plotted in the figures. It is notable that the 2-pass results of SAM and SAM–LPM improve ETS scores significantly compared to the arithmetic ensemble mean for all forecast times. Also, some of the SAM–LPM (2-pass) results show significant improvement in ETS scores compared to LPM, even though these results have relatively large variances considering the small sample size (29 cases) and relatively rare high-threshold rainfall events.

5. Spatial verification

One goal of aligning ensemble members prior to averaging is to try to preserve the structure of the precipitation features observed in the individual members in the consensus products. Preserving the structure of precipitation features can better inform forecasters about the nature of the precipitation in the model. To determine if SAM improves the spatial features of the ensemble consensus, spatial verification using MODE, a part of the MET toolkit, is performed on notable flooding cases during the 2023 FFaIR experiment. Three of the most impactful cases covered by the 2023 FFaIR forecasts are examined in order to compare MODE verification metrics of the standard LPM mean to the SAM–LPM with two iterations of member-to-mean alignment applied to the CAPS FV3-LAM ensemble using the original ensemble PM mean field.

a. *Flooding in New Hampshire of 16 July 2023*

The first case is a flooding event in New Hampshire on 16 July 2023, which caused considerable damage in and around Manchester including sinkholes, flooded basements, and road closures (see, e.g., New Hampshire Public Radio 2023). Figure 9 shows an ensemble consensus for the CAPS FV3-LAM ensemble

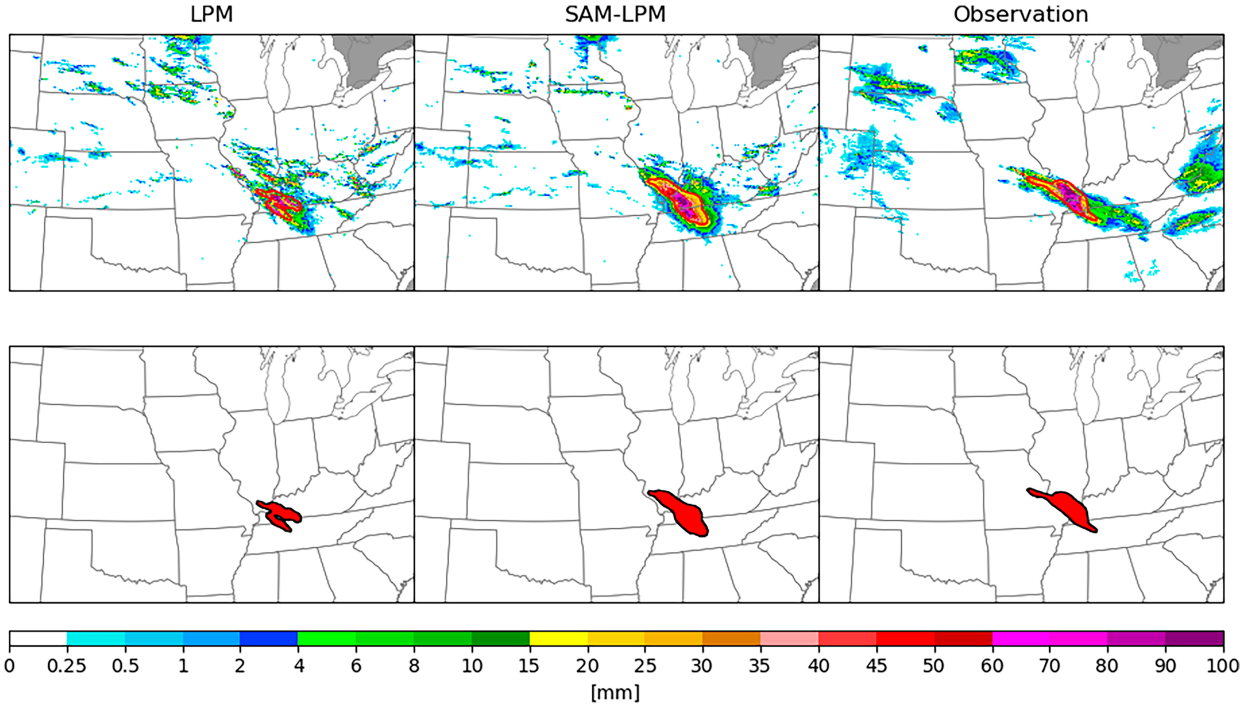

FIG. 11. (left) SAM, (center) SAM–LPM, and (right) stage IV observed (top) 6-h rainfall (mm) and (bottom) MODE objects [≥20 mm (6 h)$^{-1}$] for the 2023 western Kentucky flood event from the CAPS FV3-LAM ensemble 60-h forecast valid at 1200 UTC 19 Jul 2023.

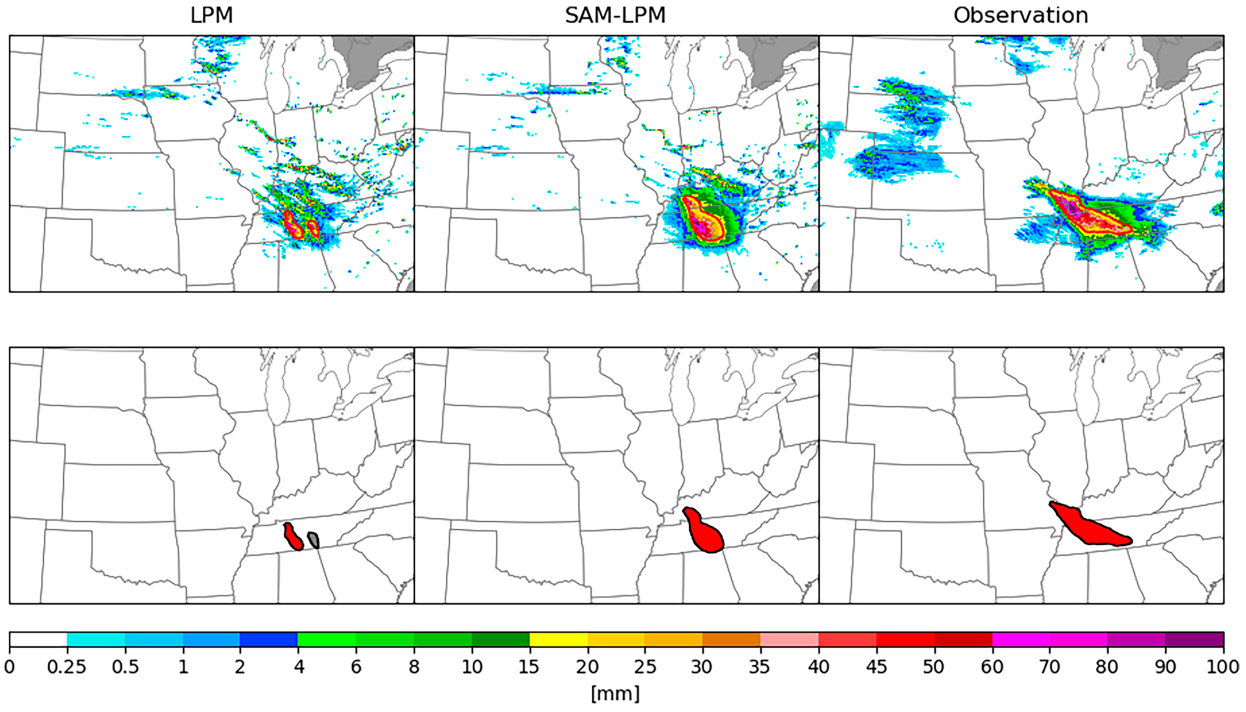

FIG. 12. (left) SAM, (center) SAM–LPM, and (right) stage IV observed (top) 6-h rainfall (mm) and (bottom) MODE objects [≥20 mm (6 h)$^{-1}$] for the 2023 western Kentucky flood event from the CAPS FV3-LAM ensemble 66-h forecast valid at 1800 UTC 19 Jul 2023.

6-h rainfall at 60-h lead time using standard LPM and SAM–LPM compared to observed stage IV precipitation valid at 1200 UTC 16 July 2023, 6 h before the flooding event, and Fig. 10 shows the same utilizing consensus method at 66-h lead time valid at 1800 UTC 16 July 2023.

Considering the precipitation features greater than 20 mm (6 h)$^{-1}$, the SAM–LPM has a precipitation feature which is closer in shape to the observed event, although the location of the feature still has an offset to the west. This similarity of shape is measured quantitatively by MODE (Table 2), the axis angle, area ratio, and intersection area compared to the feature from the observed precipitation. These results, combined with two other metrics (centroid distance and boundary distance, not shown), lead to a better interest score, a normalized weighted mean of MODE metrics ranging from 0 to 1, in MODE score and ETS score for SAM–LPM.

b. *Western Kentucky flood of 19 July 2023*

The next case is a historic flash-flooding event on 19 July 2023 across western Kentucky and southern Illinois. A total of 150–300 mm (6–12 in.) of rain was recorded for the 24-h ending at 1500 UTC 19 July, with most falling between 0200 and 1500 UTC during this event. The Kentucky Mesonet site near Mayfield recorded 11.28 in., breaking the state 24-h rainfall record. Several roads, many homes, and businesses were inundated, and numerous people were rescued in flood water (NOAA/NWS Paducah, KY 2023a). Figures 11 and 12 show 6-h rainfall from ensemble forecast, 60 and 66 h, respectively, using standard LPM and SAM–LPM compared to stage IV precipitation observations, valid at 1200 and 1800 UTC 19 July 2023. The standard LPM consensus method exhibits a mottled appearance in the main rainfall feature centered in western Kentucky due to position differences among storm cells within the individual members, whereas the SAM–LPM consensus shows a larger, more cohesive, precipitation area similar to the observation.

The improvement in 6-h rainfall forecast structure for SAM–LPM versus LPM is measured by MODE using the 20-mm precipitation threshold (Table 3). The interest score for this event is higher for SAM–LPM, due especially to an area ratio closer to unity and a larger intersection area. ETS scores are also improved for SAM–LPM.

c. *Midsouth U.S. floods on 2–4 August 2023*

The last case examined is significant flooding across northwestern Tennessee, southwestern Kentucky, and southeastern

TABLE 3. Selected MODE metrics [≥20 mm (6 h)$^{-1}$; red object pairs in Figs. 11 and 12] and ETS metrics of LPM and SAM–LPM for the 2023 western Kentucky Flood event for the CAPS FV3-LAM ensemble 60- and 66-h forecasts valid at 1200 and 1800 UTC 19 Jul 2023, respectively.

	60-h lead time		66-h lead time	
	LPM	SAM–LPM	LPM	SAM–LPM
MODE interest	0.9580	0.9813	0.8986	0.9654
Angle difference (°)	7.43	6.38	32.42	24.34
Area ratio	0.63	1.19	0.19	0.66
Intersection area (km^2)	473	780	390	906
ETS	0.086	0.155	0.093	0.158

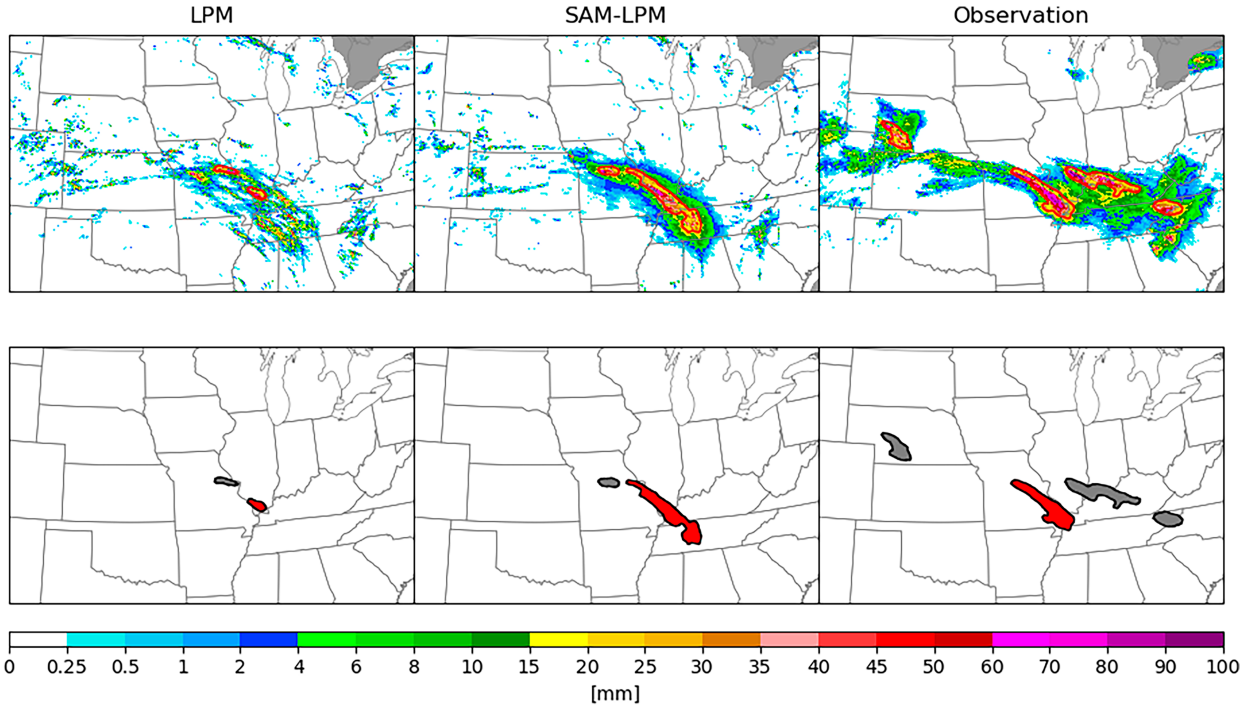

FIG. 13. (left) LPM, (center) SAM–LPM, and (right) stage IV observed (top) 6-h rainfall (mm) and (bottom) MODE objects [≥20 mm (6 h)$^{-1}$] for 2023 midsouth U.S. flood event from CAPS FV3-LAM ensemble 84-h forecast valid at 1200 UTC 3 Aug 2023.

Missouri on 2–4 August 2023. A rare flash flood emergency was issued during this event with numerous road closures, several evacuations of homes, and a few water rescues (NOAA/NWS Paducah, KY 2023b). Figure 13 shows LPM and SAM–LPM consensus 6-h rainfall forecast at 84 h and stage IV rainfall observations valid at 1200 UTC 3 August 2023. The standard LPM shows the possibility of storms between Missouri and Tennessee, but the long forecast lead time resulted in considerable spatial offsets among members, resulting in this spotty pattern to the forecast rainfall in the area. SAM–LPM, however, shows a large contiguous rainfall feature closely resembling the observations, with a small offset to the east. None of the members predicted precipitation in central Kentucky at this long lead time, so the SAM–LPM result could not recover the missed event in that area.

Using the 20-mm threshold, the 6-h SAM–LPM forecast consensus has a better matching precipitation feature with an angle and aspect close to those of the observed event, and the MODE scores (Table 4) of SAM–LPM are improved over LPM in several categories, leading to a much better overall MODE interest score and ETS scores.

To assess differences in spatial verification beyond these significant cases, the MODE interest scores for all cases in the 2023 FFaIR period were aggregated by averaging MODE interest scores among 20 mm (6 h)$^{-1}$ threshold objects for each 6-h time period between 30 and 84 h. MODE interest is obtained from pairs of objects in the observation and forecast fields. An object from the observation field can pair with multiple objects in the forecast field, so in cases where there are multiple matches to a given observed object, the highest scoring match is utilized in the aggregation. If there is no match for a given observed object, an interest score of zero is used in the average. Figure 14 shows such aggregated MODE interest scores from the CAPS FV3-LAM ensemble consensus products at the 20 mm 6 h^{-1} rainfall threshold. Both the member-to-mean aligned SAM and corresponding SAM–LPM improve the average MODE interest metric compared to the arithmetic mean and standard LPM methods.

6. Summary and conclusions

The spatial aligned mean (SAM) is applied to operational HREF and experimental real-time (CAPS FV3-LAM) CAM ensemble forecasts for 29 days spanning 10 weeks of the summer of 2023. From the pointwise verification of 6-h precipitation forecasts in both ensembles, the SAM ensemble consensus technique outperforms the simple ensemble mean at all lead times, and the spatial aligned LPM (SAM–LPM) shows improved results than

TABLE 4. Selected MODE metrics [≥20 mm (6 h)$^{-1}$; red object pairs in Fig. 13] and ETS metrics of LPM and SAM–LPM for the 2023 midsouth U.S. flood event for the CAPS FV3-LAM ensemble 84-h forecast valid at 1200 UTC 3 Aug 2023.

	LPM (84 h)	SAM–LPM (84 h)
MODE interest	0.6434	0.8176
Angle difference (°)	13.03	0.79
Area ratio	0.14	1.17
Intersection area (km^2)	0	105
ETS	0.012	0.042

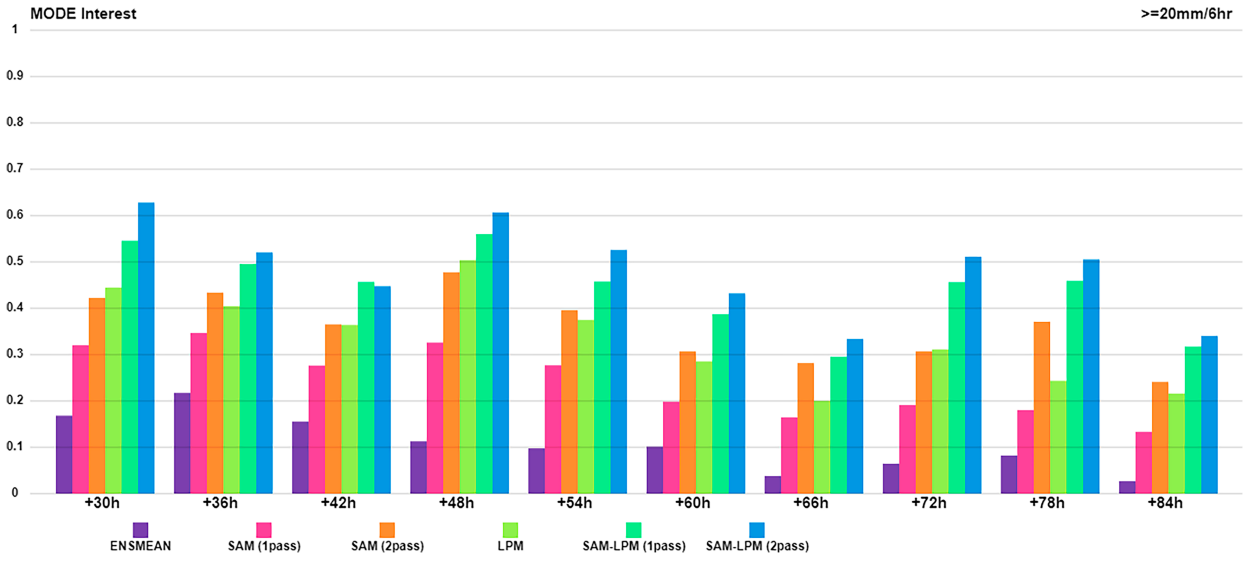

FIG. 14. Average MODE interest scores from the CAPS FV3-LAM ensemble at the 20 mm $(6\ \text{h})^{-1}$ rainfall threshold by forecast lead time (from 24–30 to 78–84 h) over the 2023 FFaIR period.

the standard LPM method. The results show that the spatial alignment technique improves the ensemble consensus significantly in common metrics such as ETS in high-threshold rainfall events. In particular, SAM–LPM improves both POD and the success ratio of high-threshold rainfall events compared to the standard LPM, which is an ideal improvement.

Also, from the spatial verification, the SAM–LPM shows improvement in the structure of the mean fields, as demonstrated in MODE spatial characteristics compared to verifying precipitation fields, while preserving the ensemble forecast maxima. Thus, SAM–LPM seems to be the best performing method for calculating an ensemble consensus for these fields. The flood cases in section 5 show that SAM–LPM is worthwhile to add value to the ensemble forecast in terms of improving precipitation features, even though the increase in the ETS metrics is modest.

Although the conclusions drawn for the individual consensus methods are the same, overall scores for HREF are better than for the CAPS FV3-LAM ensemble as configured for this experiment, showing some improvement is needed in FV3-LAM to meet the current operational standard. There are ongoing tuning and improved initialization schemes being designed to upgrade FV3-LAM before operational implementation as the RRFS, and a second dynamic core is envisioned for operational version 2 of the RRFS.

As a consensus method applied to each time step of ensemble forecasts, the spatial aligned mean has some limitations: SAM may cause discontinuities between output intervals and does not represent forecast uncertainty directly.

Though the spatial alignment technique is applied only to the precipitation field in this study, this technique can be extended to other variables as well. The spatial aligned mean of other fields can be provided as ensemble consensus products, which may also help forecasters optimally utilize a large set of ensemble forecasts. Also, it may prove useful to extend the alignment search algorithm to include the time dimension, something for future study.

Acknowledgments. This work is funded by NOAA/OAR/OWAQ Grants to the University of Oklahoma, NA19OAR4590141 and NA22OAR4590522. The first author was supported during his visit to OU/CAPS by a Korean Government Long-Term Fellowship for Overseas Studies. The CAPS FV3-LAM CAM ensemble was run in real time on the Frontera supercomputer (Stanzione et al. 2020) by generous arrangement with the staff at the Texas Advanced Computing Center (TACC) at the University of Texas, a high-performance computing center supported by the U.S. National Science Foundation.

Data availability statement. Files containing the precipitation fields for all member forecasts, along with resultant alignment vectors, aligned fields, and means, are publicly available in an Open Science Foundation database at https://osf.io/6bame/.

REFERENCES

Alexander, C. R., and J. R. Carley, 2023: Progression towards the first operational implementation of the UFS-based Rapid Refresh Forecast System as a convection allowing model application. *13th Conf. on Transition of Research to Operations*, Denver, CO, Amer. Meteor. Soc., 1B.4, https://ams.confex.com/ams/103ANNUAL/meetingapp.cgi/Paper/420286.

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.

Brewster, K. A., 2003a: Phase-correcting 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.

——, 2003b: Phase-correcting data assimilation and application to storm-scale numerical weather prediction. Part II: Application to a severe storm outbreak. *Mon. Wea. Rev.*, **131**, 493–507, https://doi.org/10.1175/1520-0493(2003)131<0493:PCDAAA>2.0.CO;2.

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.

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.

——, 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.

Davis, C. A., B. G. Brown, R. Bullock, and J. Halley-Gotway, 2009: The method for object-based diagnostic evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC Spring Program. *Wea. Forecasting*, **24**, 1252–1267, https://doi.org/10.1175/2009WAF2222241.1.

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.

Feng, J., J. Zhang, Z. Toth, M. Peña, and S. Ravela, 2020: A new measure of ensemble central tendency. *Wea. Forecasting*, **35**, 879–889, https://doi.org/10.1175/WAF-D-19-0213.1.

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.

Jankov, I., S. Gregory, S. Ravela, Z. Toth, and M. Peña, 2021: Partition of forecast error into positional and structural components. *Adv. Atmos. Sci.*, **38**, 1012–1019, https://doi.org/10.1007/s00376-021-0251-7.

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.

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.

New Hampshire Public Radio, 2023: Sinkholes, flooded basements, road closures: More rain brings more damage across NH. *New Hampshire Public Radio*, 16 July, https://www.nhpr.org/nh-news/2023-07-16/sinkholes-flooded-basements-road-closures-more-rain-brings-damage-across-nh.

NOAA/NWS Paducah, KY, 2023a: Summary of historic flash flooding on July 19, 2023. Accessed 6 December 2023, https://www.weather.gov/pah/FloodingJuly19_2023.

——, 2023b: Flash flooding August 3-4, 2023. Accessed 6 December 2023, https://www.weather.gov/pah/August2023Flooding.

Ravela, S., 2012: Quantifying uncertainty for coherent structures. *Procedia Comput. Sci.*, **9**, 1187–1196, https://doi.org/10.1016/j.procs.2012.04.128.

Roberts, B., I. L. Jirak, A. J. Clark, S. J. Weiss, and J. S. Kain, 2019: Postprocessing and visualization techniques for convection-allowing ensembles. *Bull. Amer. Meteor. Soc.*, **100**, 1245–1258, https://doi.org/10.1175/BAMS-D-18-0041.1.

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

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.

Stanzione, D., J. West, R. T. Evans, T. Minyard, O. Ghattas, and D. K. Panda, 2020: Frontera: The evolution of leadership computing at the National Science Foundation. *PEARC'20: Practice and Experience in Advanced Research Computing 2020: Catch the Wave*, Portland, OR, Association for Computing Machinery, 106–111, https://doi.org/10.1145/3311790.3396656.

Stratman, D. R., and C. K. Potvin, 2022: Testing the feature alignment technique (FAT) in an ensemble-based data assimilation and forecast system with multiple-storm scenarios. *Mon. Wea. Rev.*, **150**, 2033–2054, https://doi.org/10.1175/MWR-D-21-0289.1.

——, ——, and L. J. Wicker, 2018: Correcting storm displacement errors in ensembles using the feature alignment technique (FAT). *Mon. Wea. Rev.*, **146**, 2125–2145, https://doi.org/10.1175/MWR-D-17-0357.1.

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.

Thiebaux, H. J., P. R. Julian, and G. J. DiMego, 1990: Areal versus collocation data quality control. *Int. Symp. on Assimilation of Observations in Meteorology and Oceanography*, Clermont-Ferrand, France, WMO, 255–260.

Trojniak, S., J. Correia Jr., and W. M. Bartolini, 2024: 2023 Flash Flood and Intense Rainfall (FFaIR) final report: Part 1—RRFS related results and findings. NOAA/NWS/WPC/HMT Rep., 92 pp., https://www.wpc.ncep.noaa.gov/hmt/Reports/FFaIR/2023_FFaIR_Final_Report_Part1.pdf.

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

Yang, F., V. S. Tallapragada, D. T. Kleist, A. Chawla, J. Wang, R. Treadon, and J. Whitaker, 2021: On the development and evaluation of NWS Global Forecast Systems version 16. *Special Symp. on Global and Mesoscale Models: Updates and Center*, Online, Amer. Meteor. Soc., 12.2, https://ams.confex.com/ams/101ANNUAL/meetingapp.cgi/Paper/378135.

Zheng, M., E. K. M. Chang, B. A. Colle, Y. Luo, and Y. Zhu, 2017: Applying fuzzy clustering to a multimodel ensemble for U.S. East Coast winter storms: Scenario identification and forecast verification. *Wea. Forecasting*, **32**, 881–903, https://doi.org/10.1175/WAF-D-16-0112.1.