Main Manuscript File

Generated using the official AMS LaTeX template v6.1

Evaluation of a Cumulus Scheme in the FV3-LAM and of Experimental MPAS Forecasts for the Summer 2025 Flash Flood and Intense Rainfall Experiment

Andrew H. Berrington,[a,b] Nathan Snook,[a,b] Ming Xue,[a,b] Bryan Carroll,[a] Keith A. Brewster,[a,b] Sarah Trojniak,[c,d] James Correia,[c,d]

[a] *Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma*

[b] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*

[c] *Cooperative Institute for Research in Environmental Sciences, CU Boulder, Boulder, Colorado*

[d] *NOAA/Weather Prediction Center, College Park, Maryland*

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

Early Online Release: This preliminary version has been accepted for publication in *Weather and Forecasting*, may be fully cited, and has been assigned DOI 10.1175/WAF-D-26-0084.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

© 2026 American Meteorological Society. This is an Author Accepted Manuscript distributed under the terms of the default AMS reuse license. For information regarding reuse 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 09:54 PM UTC

ABSTRACT: A convection-allowing model (CAM) ensemble was run during the NOAA Hydrometeorological Testbed's 2025 Flash Flood and Intense Rainfall Experiment. One of the goals was to evaluate the role and effects of a cumulus scheme on precipitation forecasting in CAMs, where four members of the ensemble have a scale-aware simplified Arakawa–Schubert (saSAS) cumulus scheme activated, while the other four have it turned off. The ensemble was run once daily for the 24 days of the experiment. The ensemble consisted of eight FV3-LAM members and three experimental MPAS members run on 3-km grids and employed a variety of physics combinations. Contingency table metrics and the fractions skill score indicate that activating the saSAS scheme in the FV3-LAM does not provide benefit towards summer precipitation forecasting at varying thresholds, and, in some cases, actually degrades forecast skill. The RUC land surface model interacts more strongly with the cumulus scheme than other physics schemes. Additionally, we find that MPAS members perform comparably to their FV3 counterparts on day 1, but degrade thereafter, suggesting that more evaluation of the MPAS core and optimization of the accompanying physics with MPAS is needed.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

SIGNIFICANCE STATEMENT: Short-range prediction of heavy precipitation is crucial to reduce impacts from flash flooding. In this study, we evaluate experimental model members run during summer 2025 to assess the role of a cumulus scheme and a different model dynamic core on precipitation forecasting. We find through some metrics that activating a cumulus scheme generally does not have a positive effect on precipitation verification and the alternative dynamic core shows decreased skill after day 1. The results from this study can inform future convection-allowing model design and subsequent experiments focusing on precipitation forecasting.

1. Introduction

As part of the effort to improve forecasts of precipitation and flash flooding events, which result in considerable threats to life and property in the United States (Ashley and Ashley 2008; Smith and Katz 2013), the National Oceanographic and Atmospheric Administration (NOAA) Hydrometeorological Testbed's (HMT; Nelson et al. 2022) Flash Flood and Intense Rainfall (FFaIR) Experiment (Barthold et al. 2015) was organized for its 13th year in summer 2025 by the Weather Prediction Center (WPC). Since 2016, the Center for Analysis and Prediction of Storms (CAPS) has been developing novel ensemble consensus products and evaluating ensembles and ensemble consensus products by running convection-allowing model (CAM) ensembles during FFaIR. The testing in these NOAA Testbed experiments has contributed to informing the design of the Rapid Refresh Forecast System (RRFS; Carley et al. 2023) and its accompanying RRFS Ensemble Forecasting System (REFS), which are slated to replace the current suite of CAMs under the Unified Forecast System's (UFS) framework. Additionally, most development has been dedicated to the RRFS Version 1 that has been built around the Finite-Volume Cubed-Sphere (FV3) dynamic core (Lin 2004), while the latest plan from NOAA is to use the Model for Prediction Across Scales' (MPAS; Skamarock et al. 2012) dynamic core for RRFS Version 2 (RRFSv2; Alexander et al. 2023; Potvin et al. 2025; Trojniak et al. 2025). An important reason for switching from the FV3 dynamic core to the MPAS dynamic core is the observation that FV3 tends to produce overly intense convection, especially in weakly forced environments (Trojniak and Correia 2024; Trojniak et al. 2025). The activation of a cumulus parameterization scheme, at a grid spacing of 3 km, is an effort to try and alleviate the problem. The effects of implementing a cumulus scheme have, however, not been systematically documented.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

In recent years, under the support of the NOAA UFS program, the NOAA Testbed program, and other related programs, CAPS has carried out the testing and evaluation of FV3-based CAM ensembles. These have been configured with different physics options and, in some cases, initialized using different data assimilation systems. There were also contributions to the NOAA HMT and Hazardous Weather Testbed (HWT) Experiments. For example, Zhang et al. (2019a) determined that precipitation forecast skill in the FV3 was comparable to that in the Weather Research and Forecasting (WRF) model and that Thompson microphysics (Thompson et al. 2008; Thompson and Eidhammer 2014) slightly outperformed NSSL microphysics (Mansell et al. 2010). Supinie et al. (2022) performed regional verification, generally focused on the differences among physics combinations, centered on an FV3 ensemble run by CAPS during the 11th HMT Winter Weather Experiment. They found biases concerning snowfall in certain regions of the United States and recommended that both microphysics, the Noah and Noah-MP (Ek et al. 2003; Niu et al. 2011) land surface models (LSMs), and the MYNN Eddy-Diffusivity Mass-Flux (MYNN-EDMF, shortened to MYNN; Nakanishi and Niino 2009; Angevine et al. 2020) boundary layer scheme be used in the RRFS for winter precipitation.

More recently, Johnson et al. (2025) and Snook et al. (2025) evaluated the 21-member ensemble run by CAPS during the 2022 HWT Spring Forecasting Experiment. The former found that members utilizing NSSL microphysics, MYNN boundary layer, and RUC land surface (Smirnova et al. 2016) parameterizations yielded higher skill and higher frequency bias than members with Thompson microphysics, Turbulent Kinetic Energy Eddy-Diffusivity Mass-Flux (TKE-EDMF; Han and Bretherton 2019) boundary layer, and Noah-MP land surface parameterizations. The latter determined that for many members using the FV3 dynamic core, there was under-prediction at lower rainfall rates and over-prediction at higher rainfall rates. Larger frequency biases using the Noah-MP LSM (Niu et al. 2011) were noted, and further tests were recommended.

Finally, Berrington et al. (2026) performed a broad evaluation of the 2024 FFaIR CAPS ensemble. Bias behavior with the Noah-MP LSM consistent with Snook et al. (2025) was noted. The impacts of radar reflectivity data assimilation and the activation of the Grell-Freitas (G-F; Freitas et al. 2021) or scale-aware Arakawa-Schubert (saSAS; Han et al. 2017) cumulus schemes were also examined. It was determined that the radar data assimilation benefited forecasts on day 1, while the drop-off in forecast skill of both 95th percentile (95P) and 99th percentile (99P) precipitation increased when

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

using the cumulus schemes coupled with Global Ensemble Forecast System (GEFS) boundary conditions. The 2024 CAPS ensemble, however, was mainly designed to evaluate candidate RRFSv1 configurations and to generate consistent inputs for machine learning ensemble consensus products, not to directly evaluate the impacts of any single physics configuration choice on forecast behavior. The use of different initial conditions from GEFS also prevented a clean comparison among the members in terms of physics effects. Evaluation of the 2024 ensemble (Trojniak and Correia 2024; Berrington et al. 2026) and discussions with HMT organizers highlighted the use of scale-aware cumulus parameterization as a priority for future investigation, so CAPS coordinated with HMT staff to design a 2025 ensemble to allow for direct evaluations along those lines.

The 2025 CAPS ensemble was thus designed to specifically test the saSAS cumulus scheme in a controlled experiment at convection-allowing resolutions. At 3-km horizontal grid-spacing, this amounts to running the cumulus scheme in the so-called "gray-zone" (Sakradzija et al. 2016; Kwon and Hong 2017; Zhang et al. 2021), where convective processes are only partially resolved by model dynamics. In Kwon and Hong (2017), saSAS was found to slightly improve forecasts of summer monsoonal precipitation over the Korean Peninsula at 3-km and 1-km grid-spacings. Jeworrek et al. (2019) documented two cases using WRF over the Southern Great Plains and found that the scale-aware G-F scheme performed the best at predicting the timing and intensity of precipitation. Zhang et al. (2021) also used a WRF-driven simulation to test the multi-scale Kain-Fritsch and saSAS schemes at 3-km grid-spacing. Fundamental differences in trigger functions, cloud models, and closure assumptions rendered it difficult to determine which scheme performed better, and they recommended more analysis into the parameterization of subgrid-scale-dependent processes such as entrainment and detrainment. Prior to the 2024 FFaIR Experiment, cumulus schemes had not been extensively tested with the FV3-LAM core at gray-zone resolutions. Trojniak and Correia (2024) documented that although cumulus schemes reduced the over-prediction of intense convection that was known to be an issue in FV3-based members, it also increased erroneous spread of light precipitation through suppression of deep convection. Berrington et al. (2026) found similar results in some of their members employing cumulus schemes, with percentile distributions improving relative to observations at very high thresholds, but over-predicting at lower thresholds.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

Given the results of these previous studies, along with questions raised following FFaIR 2024, this study documents a more controlled test of the saSAS cumulus scheme using the FV3-LAM core, along with a more extensive test of the MPAS dynamic core compared to that undertaken in Berrington et al. (2026). Running the MPAS members offers an opportunity to evaluate the MPAS dynamic core configured with the same physics as the FV3-LAM members. The overall 2025 CAPS FFaIR ensemble was designed not as a theoretically rigorous ensemble, but rather as a set of direct comparisons between pairs of members with the saSAS cumulus scheme turned on or off, plus three MPAS members. The initial and boundary conditions are also kept constant between all members. Thus, this work focuses on evaluation of individual member performance and comparisons intended to isolate the impacts of the saSAS cumulus scheme, with minimal focus on ensemble forecast metrics.

2. Data and Methods

The CAPS 2025 dataset for FFaIR consists of 24 forecast days, where forecasts were initialized at 0000 UTC on experiment weekdays in June (9–13 and 23–27) and July (3–4, 7–8, 14–18, and 28–31), along with 1 August 2025, using the limited-area version of the FV3 dynamic core (FV3-LAM; Black et al. 2021). Each of the eight ensemble members was run at a 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. Hourly output to 84 hours was provided in realtime for each forecast member, yielding a total sample size of 84 x 8 x 24 = 16128 hours of forecasts in total, albeit these forecasts contain inter-dependence in terms of statistical sampling. The eight members were divided into two four-member pairs which were configured identically to one another except for activation of the saSAS cumulus parameterization (Han et al. 2017). As in Berrington et al. (2026), varying physics were employed by individual members within these sets; such as the Thompson vs. NSSL microphysics (Thompson et al. 2008; Thompson and Eidhammer 2014; Mansell et al. 2010); the Noah, Noah-MP, and RUC LSMs (Ek et al. 2003; Niu et al. 2011; Smirnova et al. 2016); and the MYNN or TKE-EDMF boundary layer treatments (Nakanishi and Niino 2009; Han and Bretherton 2019). Initial and lateral boundary conditions (ICs and LBCs) in all members were provided by the National Center for Environmental

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

Name	MP	PBL	SFC	LSM	CU
CAPS01	Thompson	MYNN	MYNN	Noah	-
CAPS02	NSSL	MYNN	MYNN	Noah	-
CAPS03	Thompson	TKE-EDMF	GFS	Noah MP	-
CAPS04	Thompson	MYNN	MYNN	RUC	-
CAPS05	Thompson	MYNN	MYNN	Noah	saSAS deep
CAPS06	NSSL	MYNN	MYNN	Noah	saSAS deep
CAPS07	Thompson	TKE-EDMF	GFS	Noah MP	saSAS deep
CAPS08	Thompson	MYNN	MYNN	RUC	saSAS deep
MPAS01	Thompson	MYNN	MYNN	Noah	-
MPAS02	NSSL	MYNN	MYNN	Noah	-
MPAS04	Thompson	MYNN	MYNN	RUC	-

TABLE 1. CAPS ensemble configurations for FFaIR 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, and CU represents the cumulus scheme option. Planned MPAS03 was not run because the Noah-MP LSM was not supported by the version of MPAS used during the experiment.

Prediction's (NCEP) operational Global Forecast System (GFS) version 15 at hourly intervals (Zhou et al. 2022). Member configurations are detailed in Table 1.

In addition to the eight FV3-LAM members described above, CAPS also ran three members using the Model for Prediction Across Scales (MPAS) version 8.0.1 dynamic core (Skamarock et al. 2012); the configuration of these members is also described in Table 1. This was an earlier version of MPAS than is currently in development for RRFSv2, and was adapted from the 2024 Hazardous Weather Testbed due to its ease of transference to the Texas Advanced Computing Center (TACC) computers utilized for this project. Therefore, the results concerning MPAS detail the comparisons between less-mature MPAS configurations (with fewer vertical levels) and more-developed FV3-LAM members with greater tuning for convection-allowing resolutions, and therefore, should be treated with some degree of caution. Nevertheless, three MPAS members were generated for the same dates as the CAPS FV3-LAM ensemble, using the same underlying physics as members CAPS01, CAPS02, and CAPS04. An MPAS member equivalent to CAPS03 was not created as the relevant physics combination was not yet available for MPAS. Horizontal grid spacing and ICs/LBCs were the same as the CAPS FV3-LAM ensemble, while the MPAS had 55 vertical levels compared to the FV3-LAM's 64.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

Observations for precipitation verification were extracted from Multi-Radar Multi-Sensor (MRMS) QPE (Zhang et al. 2016). Compared to available alternatives, such as NOAA Stage IV (Nelson et al. 2016) data, MRMS is more suitable over a full CONUS domain that includes the western United States. MRMS QPE was evaluated by White and Nelson (2024), who documented increased error due to complex terrain affecting radar estimates of precipitation over Colorado, which can be applicable to the remainder of the CONUS in mountainous regions. Additionally, areas of poor radar coverage where beam heights are considerably high off the ground may impact radar-derived QPE. However, MRMS QPE did not suffer from the same rain gauge and river forecast center concerns that were documented in Nelson et al. (2016) over larger regions of the western CONUS. Thus, the full CONUS domain was used for verification in this study. FV3-LAM and MPAS model data was re-gridded using the budget approach (Mesinger 1996; Accadia et al. 2003) and masked onto the MRMS QPE 0.01^o CONUS grid using the Model Evaluation Tools (MET) version 11.0 grid-stat tool (Brown et al. 2021). Additional tests (not shown) regridding to the 3 km CAPS forecast grid yielded minimal differences both qualitatively and quantitatively.

Forecast evaluation metrics used in this study were similar to those used in Berrington et al. (2026). Both 6-hour and 1-hour accumulated precipitation were examined. 6-hour scores were also aggregated into daily windows where days 1, 2, 3, and 4 represent forecast leads of 0–24 hours, 24–48 hours, 48–72 hours, and 72–84 hours, respectively. Most verification statistics were derived from binary contingency tables and, rather than fixed-value thresholds, use the 95P and 99P from across the full dataset as thresholds. Again following Berrington et al. (2026), where frequency bias was desired as a metric such as in performance diagrams (Roebber 2009), fixed-value thresholds were set at 3 mm and 15 mm (unless otherwise specified). The ETS, success ratio (SR), and POD were subsequently calculated from the contingency tables (Clark et al. 2010; Joliffe and Stephenson 2012). The latter two metrics then comprise the components of the performance diagram. In addition to aggregated CONUS-wide metrics, this study also used spatial plots of ETS and composite differences to assess regional variation in skill. To do this, contingency table statistics and model fields across the full set of forecasts were binned into 50 x 50 grid point (0.5 x 0.5 degree) boxes. In order to increase sample sizes for skill scores within each of these boxes, we adjusted the thresholds to 1 mm and 10 mm 6-hour precipitation in this spatial analysis. The relevant metrics were then computed within each of these boxes and mapped onto the CONUS domain.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

Another spatial verification technique, the fractions skill score (FSS), was used to determine the minimum skillful scale (MSS) where the forecasts resemble the observations (Roberts and Lean 2008). Calculating the useful FSS for percentile x (e.g. x = 0.95 for 95P) follows the simple calculation of FSS_{useful} = 0.5 + f_o/2, where f_o = 1 - x. Since member-to-member comparisons are the focus, namely CU OFF (CAPS01–04) vs. CU ON (CAPS05–08) and MPAS vs. FV3-LAM, we did not use probabilistic measures of forecast quality in this work as in Johnson et al. (2025), Snook et al. (2025), and Berrington et al. (2026). When relevant, significance testing was applied using bootstrapping, with 10000 resamples.

3. FV3-LAM Verification

a. Individual Member Verification

Performance diagrams evaluating 6-hourly precipitation at both the 3 mm and 15 mm thresholds are presented in Figure 1. For the first three pairs of CU OFF and CU ON members — CAPS01 and CAPS05, CAPS02 and CAPS06, and CAPS03 and CAPS07 — there is generally negligible difference in critical success index (CSI). For 3 mm precipitation, all six members have a dry bias that is not noticeably affected by the activation of saSAS, although there are very slight improvements to underforecasting biases on day 3 for CAPS05 and CAPS06. CAPS03 and CAPS07 have stronger dry biases that have been linked to the Noah-MP LSM and TKE-EDMF boundary layer treatments in previous work (Supinie et al. 2022; Snook et al. 2025; Berrington et al. 2026). For the 15 mm threshold, CAPS01 and CAPS02 are relatively unbiased on each day of the forecast, along with their CU ON counterparts. CAPS03 and CAPS07 start off unbiased on day 1, but become dry-biased on subsequent days of the forecast. More notable differences between CU OFF and CU ON become apparent with the comparison of CAPS04 and CAPS08, which utilizes the RUC LSM. While 3 mm results are generally similar to the other comparisons, there is a slight decline in CSI on day 2 and a slight increase on day 3 when activating saSAS, which does not rise to the level of statistical significance, along with small bias reductions for each forecast day. However, for 15 mm precipitation, there is a notable dry bias introduced when activating saSAS in the comparison between CAPS04 and CAPS08. Overall, CAPS08 has a slightly higher SR than CAPS04 but a lower POD and a dry bias; meaning that when saSAS is activated, SR increases at the expense of POD and bias. The difference in SR between CAPS04 and CAPS08

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

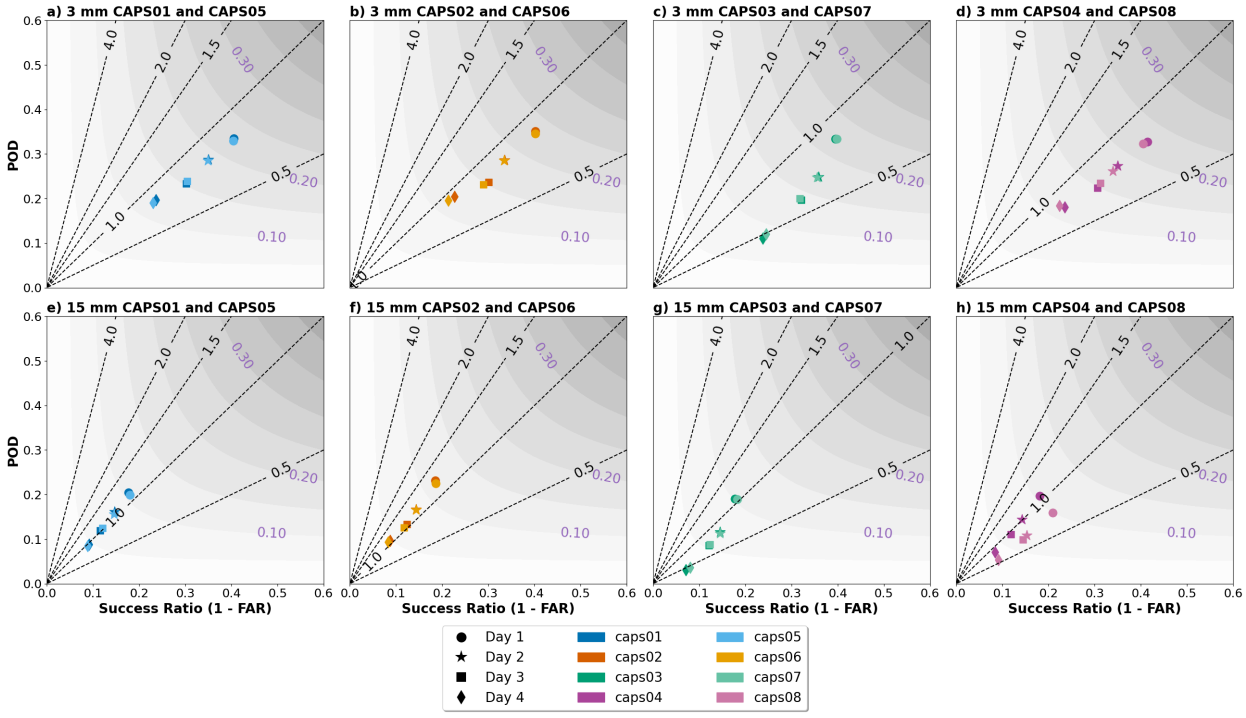


FIG. 1. Performance diagrams for 6-hour precipitation on each forecast day (different markers) for the CAPS ensemble members aggregated for the 3 mm threshold for a) CAPS01 and CAPS05, b) CAPS02 and CAPS06, c) CAPS03 and CAPS07, and d) CAPS04 and CAPS08 and the 15 mm threshold for e) CAPS01 and CAPS05, f) CAPS02 and CAPS06, g) CAPS03 and CAPS07, and h) CAPS04 and CAPS08. The grayscale color-shading represents contours of CSI, while the black dashed lines are frequency bias contours.

is typically smaller than their differences in POD and bias, thus suggesting a small degradation in overall performance of this configuration when saSAS is on. The difference between CU OFF/ON metrics for this group decrease as the forecast lead times increase. Overall, we find no evidence that activating saSAS meaningfully improves precipitation forecasts.

Examining bias further for statistical significance at the 90% confidence level in Figure 2, we observe two notable trends. The first is that, while all members have a dry bias, members CAPS03 and CAPS07 (teal data points in Fig. 2) are especially low biased, and that bias gets worse with lead time. The second is that, for 15 mm precipitation, CAPS08 is substantially drier than its CU OFF counterpart, CAPS04, on every day of the forecast. While confidence intervals overlap in the CAPS04–CAPS08 comparison, the actual bias of CAPS08 lies below the 5th percentile of the CAPS04 bootstrapped distribution. Meanwhile, the remaining CU ON members barely show

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

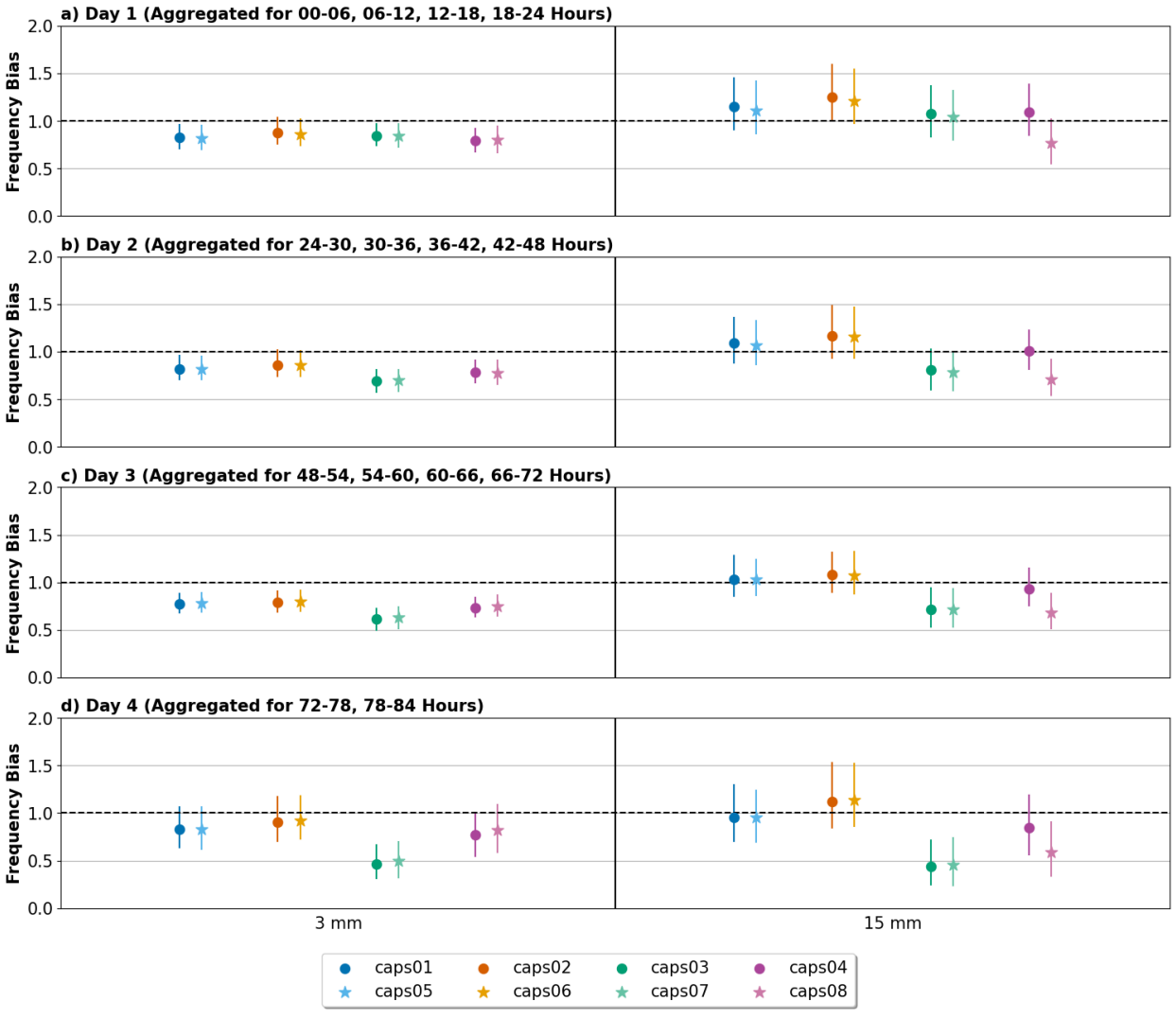


FIG. 2. Frequency bias for CU OFF (circular markers) and CU ON (star markers) CAPS ensemble members for both 3 mm (left of black vertical line) and 15 mm 6-hour precipitation (right of black vertical line) aggregated during days a) 1, b) 2, c) 3, and d) 4. Whiskers represent the 90% confidence interval from the bootstrapping significance test.

any deviation from their CU OFF counterparts. A similar test for CSI (not shown) was conducted and showed no substantial differences, consistent with Fig. 1, although CAPS03 and CAPS07 had slightly lower scores than the remaining members.

We additionally examine ETS of 1-hour precipitation at the 99P threshold in Figure 3. Differences between the CU OFF and CU ON pairs at all forecast lead times are generally negligible and the

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

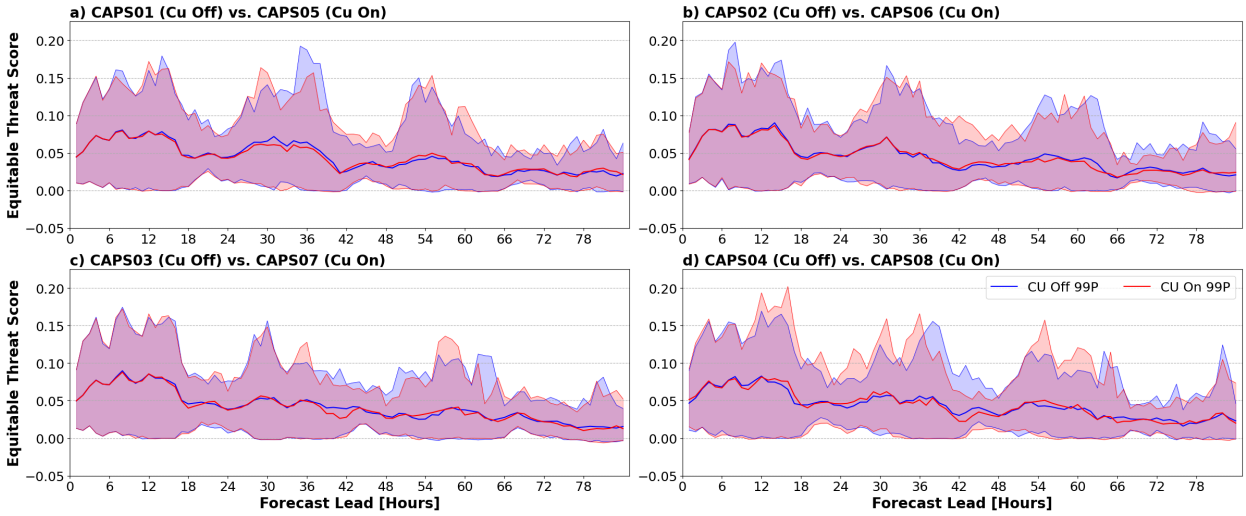


FIG. 3. ETS for CONUS 1-hour precipitation at the 99P threshold as a function of forecast lead (a–d). Blue lines indicate CU OFF members, while red lines indicate CU ON members. The shaded region represents the 90 percent confidence interval from bootstrapping.

diurnal cycle is evident by increased variance in skill during the afternoon and evening hours. Statistical significance at the 90% confidence interval is not indicated for any of the member comparisons. Absolute values of ETS are quite low, below 0.10, for all members at all lead-times for 1-hour 99P precipitation, and values for days 3 and 4 decline below 0.05. This is generally to be expected with the short time scale and high precipitation threshold (Zhang et al. 2019b), particularly with no additional data assimilation on day 1 (Sun et al. 2014; Park et al. 2025). A separate comparison for 95P precipitation (not shown) indicates that activating saSAS results in a small decrease in ETS during the overnight and morning hours for the RUC LSM member pair (CAPS04 and CAPS08), while having no discernible effect for the remaining member pairs. The conclusion from the aggregated contingency table metrics suggest that configurations using the RUC LSM are more sensitive to saSAS activation than the other physics configurations when it comes to precipitation verification. There is a minimal change in skill scores overall for both precipitation thresholds when activating saSAS for the other member pairs.

We also show FSS for 99P precipitation in Figure 4. With the lone exception of CAPS01 vs. CAPS05 on day 3, no member shows improvement in 99P FSS at any lead-time when the saSAS scheme is turned on. Differences are generally small for the first three sets of members (Fig. 4a–c

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

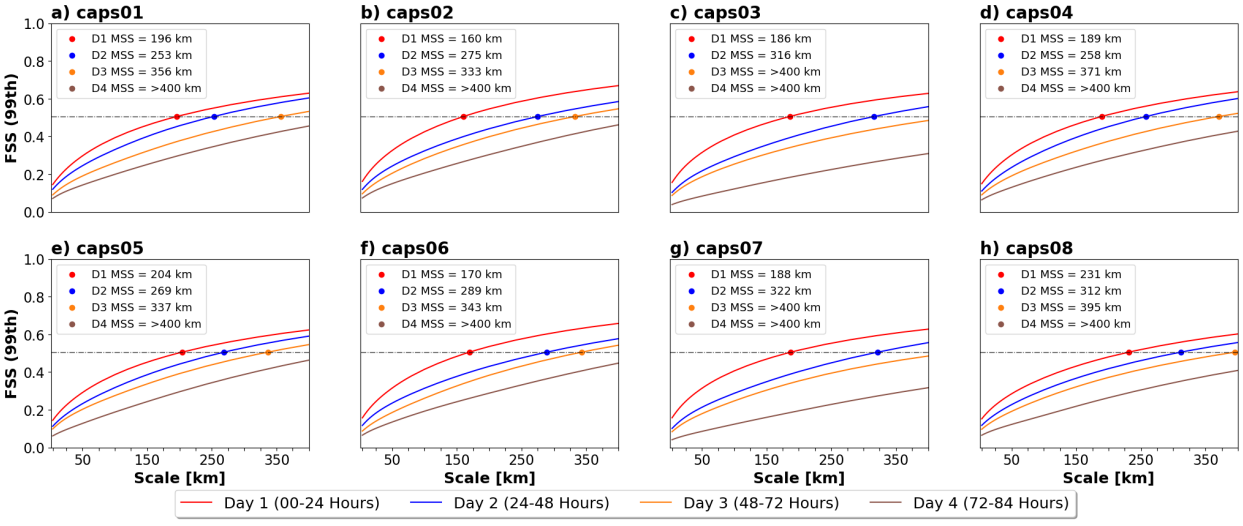


FIG. 4. FSS for 99P 6-hour precipitation for the 2025 CAPS ensemble members averaged during days 1 (red), 2 (blue), 3 (orange), and 4 (brown). Dashed-dotted line represents FSS_{useful} for the 99P. Colored dots represent where each curve crosses the FSS_{useful} threshold for each day.

and e–g), similar to the other contingency scores, with day 1 MSS differences of 2–10 km, day 2 MSS differences of 5–20 km, and day 3 MSS differences of 10–20 km. No member shows a MSS of less than 400 km on day 4. As with the other metrics, there are larger differences between CAPS04 and CAPS08. Day 1 MSS increases by over 40 km, day 2 MSS increases by over 50 km, and day 3 MSS increases by 24 km in the CU ON member versus the CU OFF member. This FSS degradation, particularly on day 2, is closer to that found in Berrington et al. (2026) for members that employed cumulus schemes with GEFS ICs and LBCs. Unlike in Berrington et al. (2026), where multiple configuration differences existed between members, however, the only difference between CAPS04 and CAPS08 is the activation of the saSAS cumulus scheme. The smaller differences in FSS and bias between the pairs of CAPS01 and CAPS05 vs. CAPS04 and CAPS08 are notable, given that the choice of LSM is the only difference between these two pairs of members. This again points to larger effects (potentially detrimental) when employing saSAS with the RUC LSM, that are not as large when using the Noah LSM, which is investigated more thoroughly below in Section 3b.

Given the construction of the 2025 CAPS ensemble, it is possible to conduct direct member comparisons on a spatial/regional basis, which were not possible in Snook et al. (2025) or Berrington

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

et al. (2026). This additionally assists in the more intensive regional examination in Section 3b, as it helps identify regions worthy of more scrutiny. Figures 5 and 6 show the spatial composite differences in ETS (aggregated across all forecast hours) for 1 mm and 10 mm 6-hour precipitation, respectively, along with histograms of these differences. In these figures, a positive difference corresponds to an improvement, and a negative difference corresponds to a degradation in ETS when the cumulus scheme is turned on. For 1 mm precipitation, CAPS05 shows notable improvements over east Texas compared to CAPS01, while displaying a slight degradation over the northern Plains. Less notable differences are displayed between CAPS06 and CAPS02, along with CAPS07 and CAPS03, although there is improvement for CU ON over far south Texas. Histograms in Fig. 5b, d, and f are generally symmetric, although there are slightly more improvements with CU ON members in the +0.02–0.04 range than degradations of the same magnitude. Finally, the CAPS08 versus CAPS04 comparison shows more substantial differences, often in opposite locations as observed in the other pairs. Here, the central Plains and Upper Midwest show more notable improvements, with relatively coherent spatial structure (less noise) than the other pairs' ETS difference maps. On the other hand, there is a broad pattern of degradation in ETS across the south-central and southeastern United States. Meanwhile, for the 10 mm threshold in Figure 6, the results are noisier. That said, there is a slight decrease in ETS noted over the Upper Midwest in CAPS05 - CAPS01, CAPS06 - CAPS02, and CAPS07 - CAPS03. There also is degradation in ETS over portions of central and southern Texas in all four comparisons. A broader pattern of degradation again occurs in the CAPS08 - CAPS04 comparison over the south-central and southeastern United States. Histograms suggest that the cumulus scheme is generally not having a beneficial influence on ETS across the CONUS, similar to the findings in Fig. 3, although the first three pairs do have some improvements at the 1 mm threshold in flash food prone regions of east and southeast Texas.

The results presented above suggest that RUC LSM combined with saSAS may have detrimental effects on precipitation forecast skill over the south-central and southeastern United States during the warm season. In the summer, this is a region generally characterized by warm and humid conditions, weak synoptic forcing and temperature gradients, and so-called "weakly-forced thunderstorms" (WFTs; Miller and Mote 2017). These WFTs, which account for a significant portion of precipitation in this region during the warm season, are generally difficult to predict since they

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

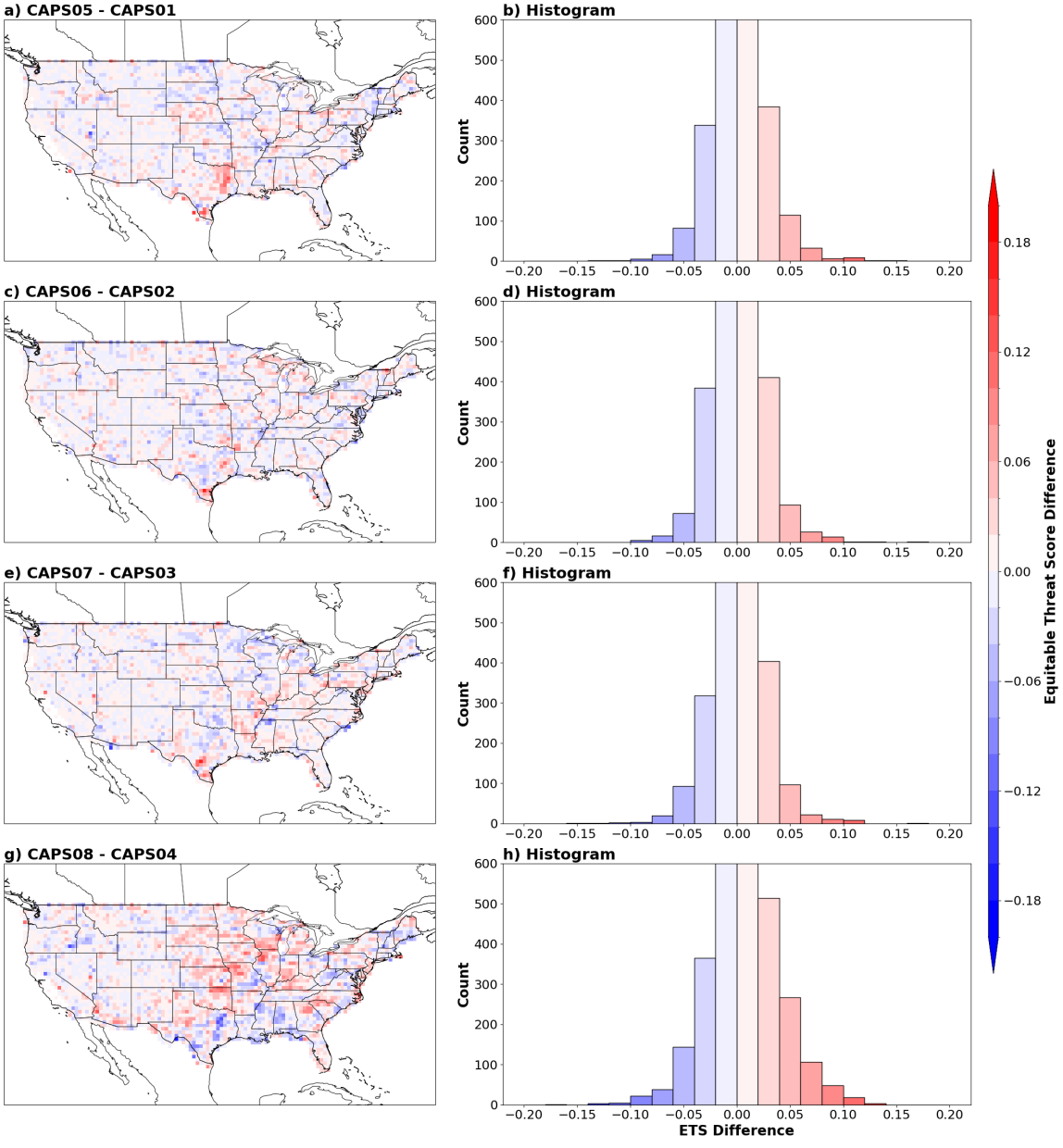


FIG. 5. Spatial composite difference of ETS (color-shaded) calculated across all forecast hours between CU ON and CU OFF members for 1 mm 6-hour precipitation (left column). Histograms of ETS improvement (resolution of 0.02), with color shading that corresponds to the spatial plot to its immediate left (right column).

lack synoptic scale or even organized mesoscale forcing, and are affected by the diurnal cycle. They also can be affected by land-use (Miller and Debbage 2023), suggesting that the LSM may have a disproportionate effect on precipitation, as well as any potential feedbacks that exist with other physics. Another region that shows some larger differences when saSAS is activated with the

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

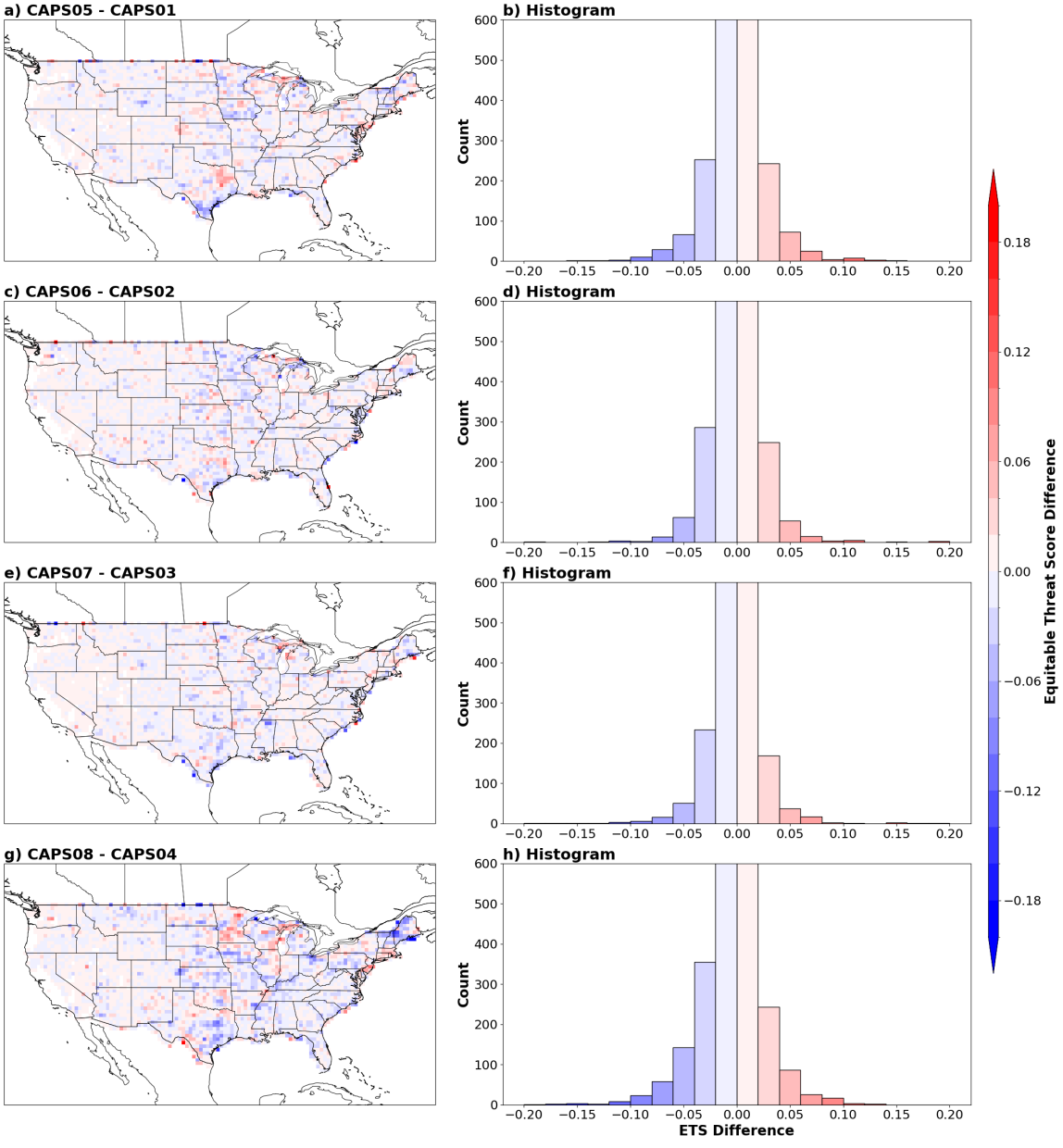


FIG. 6. As in Fig. 5, but for 10 mm 6-hour precipitation.

RUC LSM is in the central Plains and Midwest, which is strongly affected by evapotranspiration from mature crops in the warm season (another potential LSM interaction). Section 3b attempts to investigate this potential connection between LSM and saSAS further, particularly across the southern United States.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

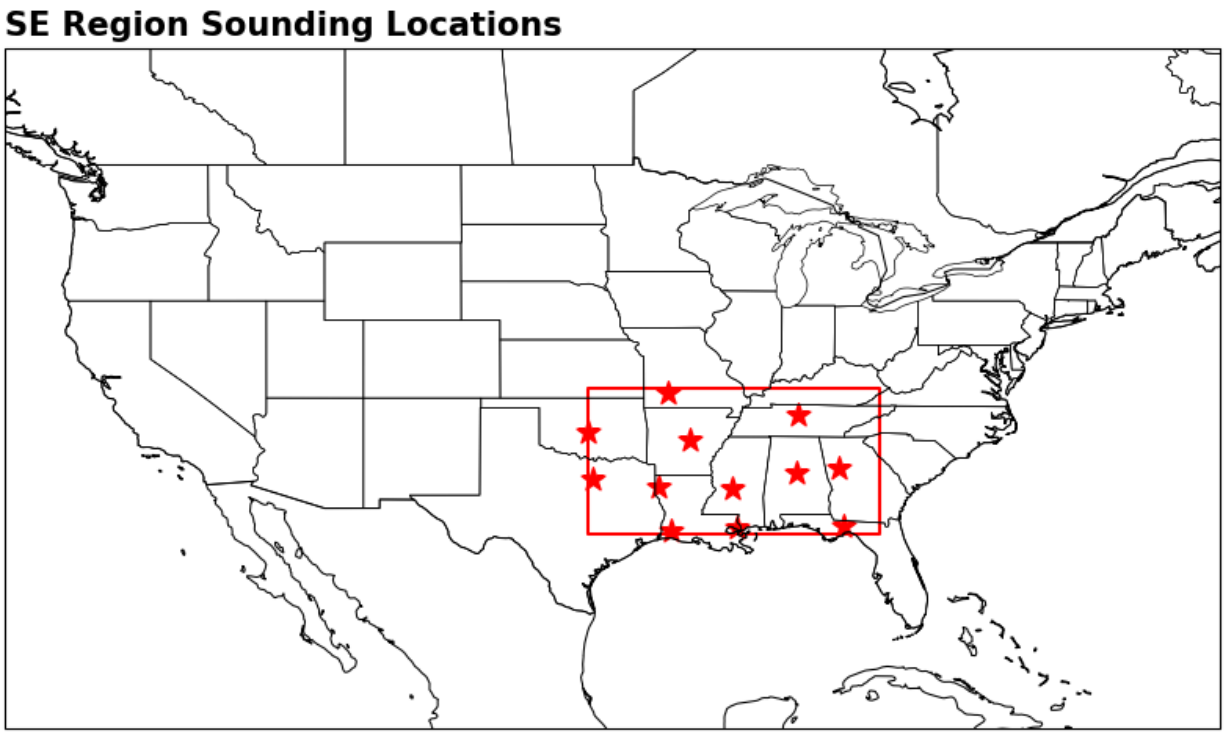


FIG. 7. Locations used in the extraction of forecast vertical profiles across the southern United States. The red box also represents the region used for frequency bias calculations in Fig. 10.

b. Physical Reasoning

To determine some factors for the greater loss of precipitation forecast skill between CAPS08 and CAPS04 (using RUC LSM) vs. CAPS05 and CAPS01 (using Noah LSM) in the southern United States when using the saSAS scheme, we extract vertical profiles of several variables every three hours[1] from the model for all forecast days within the dataset, as in Supinie et al. (2022). The sounding locations used are presented in Figure 7, and comprise 12 standard rawinsonde launch sites in the south-central and southeastern United States. We specifically focus on the late afternoon and evening hours (2100 UTC to 0300 UTC). Differences between averaged variables are then calculated between the CU ON and CU OFF members in each LSM pair.

In Figure 8a, temperatures are warmer both near the surface and between 900 and 600 hPa when turning saSAS on when the RUC LSM is used compared to the Noah LSM, in which the

[1]Temporal resolution of the soundings is three hours due to previous software implementation and the infeasibility to re-run the entire experiment for higher temporal resolutions because of computational restraints following the end of the testbed experiment.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

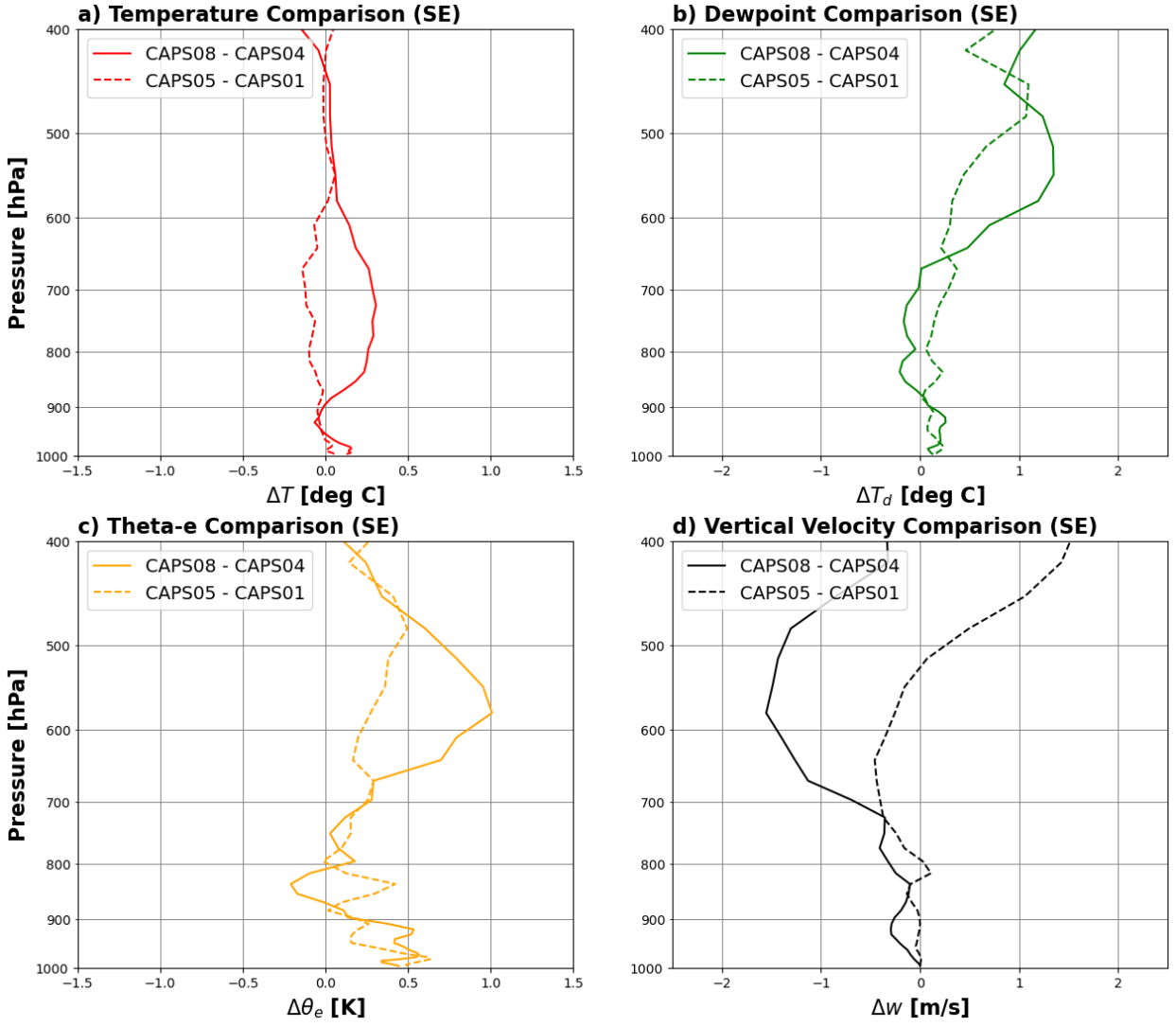


FIG. 8. Late afternoon and evening-averaged (2100 UTC to 0300 UTC) vertical profiles using pressure coordinates of a) averaged temperature, b) averaged dewpoint, c) averaged theta-e, and d) 99th percentile vertical velocity across the locations shown in Fig. 7.

differences are negligible. Moisture (Fig. 8b) is also increased in the middle troposphere (700–500 hPa) and near the surface, and slightly reduced in the lower troposphere (900–700 hPa) with the RUC LSM active when saSAS is enabled, suggestive of more aggressive triggering of the cumulus scheme and subsequent vertical moisture transport (Han et al. 2017; Freitas et al. 2021). The combined effects of the two yield higher theta-e (Fig. 8c) between 700 and 500 hPa in the RUC LSM comparison. However, stronger vertical velocities at or above the 99th percentile (Fig. 8d)

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

in the profiles are reduced considerably more in the middle troposphere when the RUC LSM and saSAS are concurrently employed compared to the Noah LSM/saSAS pair. This indicates stronger convection is being suppressed more with saSAS (or, hypothetically, lapse rates are being neutralized and leading to decreased instability for later convection) when the RUC LSM is used, and may contribute to the dry biases in 15 mm precipitation seen in Fig. 2 along with the slight decreases in ETS seen in 10 mm precipitation across the southern United States in Fig. 6.

Given that point profiles can often under-sample their environments, afternoon (18 UTC – 00 UTC) percentiles of several two-dimensional fields over the box in Fig. 7 are presented for CAPS04 and CAPS08 in Figure 9. Here, there is a decrease in all percentile values for 700 hPa vertical velocity (Fig. 9a) when saSAS and the RUC LSM are used, augmenting the vertical profile analysis. Supporting the idea of more light precipitation and a decrease in heavy precipitation (stronger convection) with saSAS turned on, composite reflectivity (Fig. 9b) has a notable increase in values below the 98th percentile relative to CAPS04, while the values decrease relative to CAPS04 above the 98th percentile. In Fig. 9c and d, MUCAPE percentile values increase when saSAS is active, while MUCIN values remain relatively unchanged. The former is likely related to a slight increase in moisture near the surface as shown in Fig. 8b and potentially the decrease of stronger convection (and corresponding overturning) on average suggested by Fig. 9b. This also somewhat refutes the hypothesis suggested above that a decrease in overall instability is the reason behind the precipitation biases. Instead, this may imply that convective triggering from saSAS is altered when using the RUC LSM and the convection itself is therefore being modified rather than the broader thermodynamics.

When considering the differences in the interactions of saSAS between the RUC LSM and Noah LSM, it follows to evaluate near-surface processes that may differ between LSMs. Particularly when concerning interactions with a cumulus scheme, this concerns vertical moisture flux from the surface. Thus, the differences in latent heat flux (LHF) between CU ON and CU OFF members, along with the differences between LSMs, are shown in Figure 10. Firstly, regardless of whether saSAS is employed or not (Fig. 10c), there are substantial differences between the RUC LSM and Noah LSM across most of the CONUS. LHF is substantially less across the Plains and Midwest in the RUC LSM, and notably higher over the Southeast, Pacific Northwest, and Northeast. This suggests that variations in evapotranspiration and soil moisture handling (Ek et al. 2003; Smirnova

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

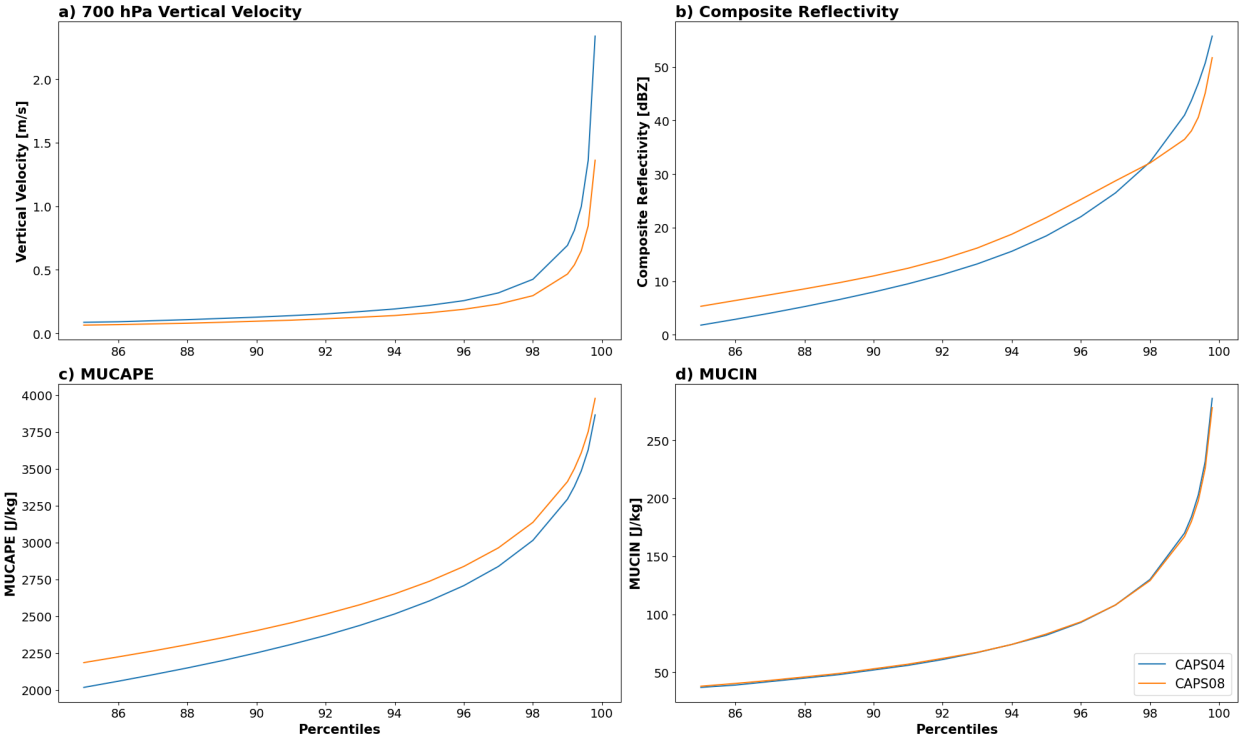


FIG. 9. CAPS04 and CAPS08 percentile values of a) 700 hPa vertical velocity (m s^{-1}), b) composite reflectivity (dBZ), c) MUCAPE (J kg^{-1}), and d) MUCIN (J kg^{-1}) at 18 UTC and 00 UTC calculated over the box in Fig. 7 and for the full dataset.

et al. 2016; Constantinidou et al. 2020), particularly the former given the dichotomy between forested (e.g. the Southeast and Pacific Northwest) and grassland/cropland (e.g. the Plains and Midwest) land-use types, may impact forecasts of moisture-related fields between the two LSMs. However, there is additional enhancement of LHF across the Southeast with CU ON in the RUC LSM comparison (Fig. 10b) as opposed to the Noah LSM comparison (Fig. 10a), where the difference is muted. As with Fig. 8, this indicates a potential land-atmosphere feedback between the RUC LSM and cumulus scheme, particularly in the warm/moist and forested southern United States, that is potentially responsible for some of the skill differences seen in earlier results. A similar test (not shown) was conducted for sensible heat flux, and while differences between the LSMs themselves were notable as expected, the differences when saSAS was turned on or off were of lesser magnitude than LHF, particularly when using the RUC LSM. We hypothesize that the feedback may result from aggressive triggering of the saSAS cumulus scheme and spreading of

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

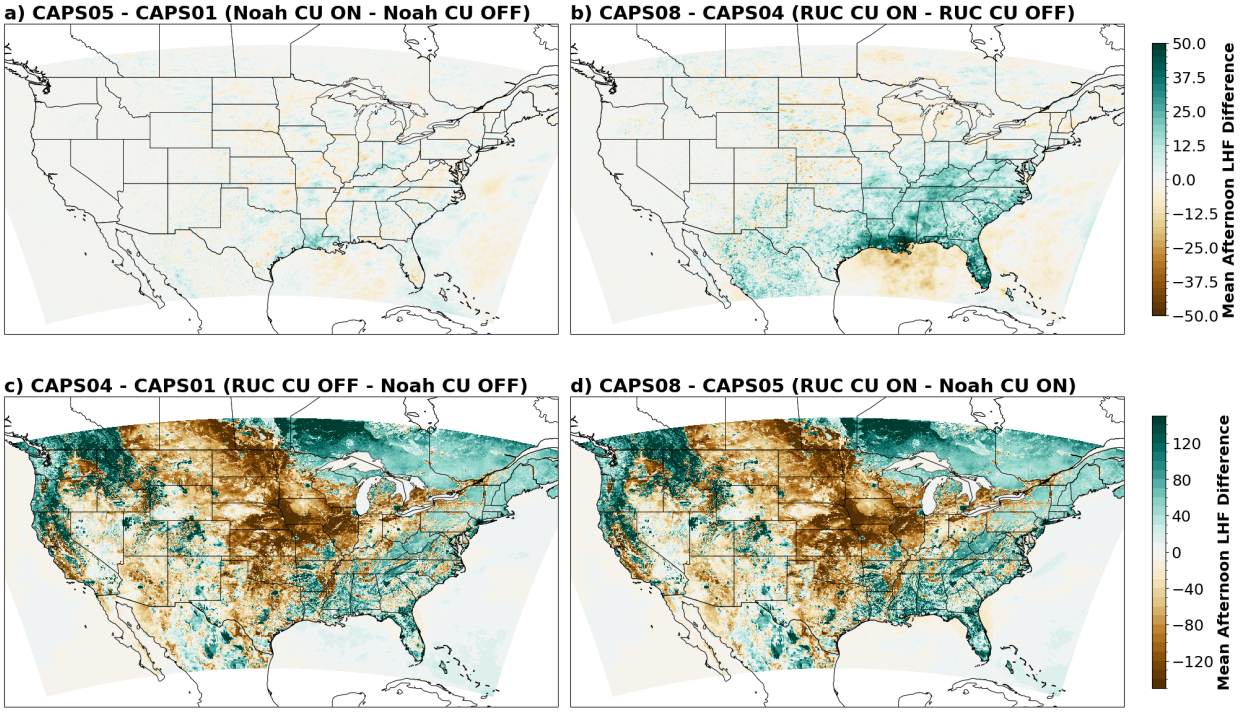


FIG. 10. Dataset- and afternoon-averaged (1800 UTC to 0000 UTC) latent heat flux composite differences (shaded, in W m^{-2}) between a) CAPS05 and CAPS01 (Noah LSM/CU ON vs. CU OFF), b) CAPS08 and CAPS04 (RUC LSM/CU ON vs. CU OFF), c) CAPS04 and CAPS01 (RUC LSM vs. Noah LSM/CU OFF), and d) CAPS08 and CAPS05 (RUC LSM vs. Noah LSM/CU ON). Color-bar in the upper right corresponds to panels a) and b), while the color-bar in the lower right corresponds to panels c) and d).

light precipitation over larger areas. The light precipitation also enhances LHF from subsequent evaporation from the surface (Fig. 10b), resulting in further triggering of saSAS, and further enhancing LHF. Ultimately, this results in a more gradual release of instability through the day and a suppression of convection later in the day (an alteration of the convective time scale), resulting in some of the dry biases (Figs. 1, 2, and 11) and weakened vertical velocities (Figs. 8d and 9a) when comparing CAPS08 to CAPS04.

To conclude this section, an indirect test of the feedback hypothesized above is presented in Figure 11. Here, time series of 1-hour precipitation frequency biases, calculated over the region delineated by the box in Fig. 7, are compared between 0.25 mm (94–95th percentile) and 3 mm (98–99th percentile) precipitation with CAPS05/01 on the left and CAPS08/04 on the right. The results show that using saSAS with the RUC LSM dramatically amplifies the diurnal cycle of

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

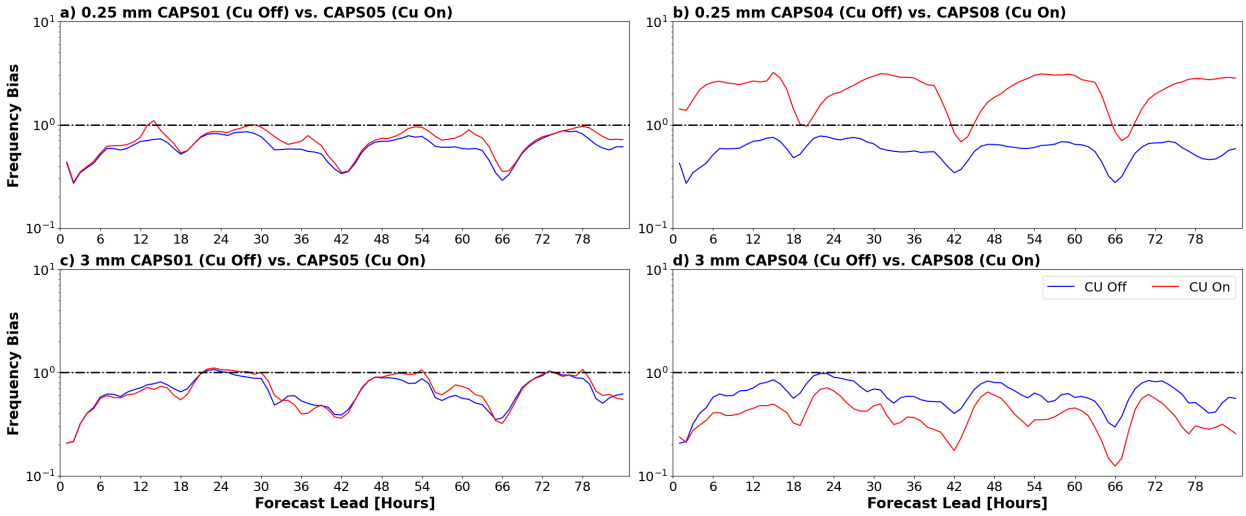


FIG. 11. 1-hour averaged frequency biases calculated over the box in Fig. 7 between a) 0.25 mm CAPS05 and CAPS01, b) 0.25 mm CAPS08 and CAPS04, c) 3 mm CAPS05 and CAPS01, and d) 3 mm CAPS08 and CAPS04. Bias is plotted using a logarithmic scale.

precipitation and yields high frequency biases for 0.25 mm precipitation (Fig. 11b), while the differences are much smaller in the Noah LSM comparison. This likely also leads to the dry biases at all forecast leads for 3 mm 1-hour precipitation for CAPS08 in Fig. 11d, while CAPS05 actually sees slight improvement in frequency biases for 3 mm precipitation over CAPS01 for some lead times. While the physical explanations for some of the verification scores here are certainly not exhaustive, it does indicate that utilizing saSAS seems to have negative effects when coupled with the RUC LSM over the Southeast region at least — it creates high frequency bias for very light precipitation and amplifies the low frequency bias for somewhat heavier precipitation (also supported by Fig. 9b). Such impacts should be further examined in future work. This future work could perhaps include a deeper analysis of how the cumulus scheme interacts with different land surface models, especially at convection-allowing resolutions, and especially with the use of the MPAS dynamic core.

4. MPAS Verification

Concerning MPAS verification, Potvin et al. (2025) evaluated experimental forecasts generated during the 2023 HWT and determined that MPAS composite simulated reflectivity storm objects

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

best matched corresponding MRMS observations regarding size, shape, and convective mode compared to other models evaluated in the HWT that year. FV3-based models in their study showed low biases in temperature and instability, while MPAS variants were less biased. Additionally, Trojniak et al. (2025) evaluated updated versions of the experimental MPAS and FV3-LAM forecasts from those in Potvin et al. (2025) for the 2025 FFaIR Experiment, both objectively as well as subjectively through participant input. Their MPAS members were judged to have been better than the FV3-based RRFS at capturing precipitation timing and location, but had dry biases and struggled to yield high precipitation totals. However, peaks in precipitation intensity for the MPAS members did a good job at matching the observations. Finally, Berrington et al. (2026) compared an MPAS member using Thompson microphysics, Noah LSM, and MYNN PBL schemes to its FV3-LAM counterpart and determined that MPAS generally matched FV3-LAM in performance on day 1 at different precipitation thresholds, while declining thereafter.

Performance diagrams comparing CAPS01, CAPS02, and CAPS04 with their MPAS counterparts are shown in Figure 12. For 3 mm precipitation, although not statistically significant, all CAPS FV3 members slightly outperform their MPAS counterparts in terms of CSI, while biases are generally unchanged. However, CSI notably drops off, and is notably worse, for days 3 and 4 when using MPAS. For 15 mm precipitation, MPAS members are noticeably drier (Trojniak et al. 2025) than their FV3 counterparts, along with some decrease in CSI, even on day 1. The bias differences are statistically significant when bootstrapping is applied (not shown). Similar to Fig. 3, ETS for 1-hour 99P precipitation is presented in Figure 13. While ETS between FV3 and MPAS generally remains consistent through day 1 of the forecast, there are periods thereafter where the FV3 member outperforms the MPAS variant. In Fig. 13a (CAPS01 vs. MPAS01), this occurs during days 2 and 3. Statistical significance is nearly reached on day 3, just prior to 54 hours lead time. In Fig. 13b (CAPS02 vs. MPAS02), the FV3 member outperforms MPAS slightly during day 2 and more substantially during days 3 and 4. Statistical significance is also nearly reached on day 3 between 60 and 66 hours lead time. Finally, in Fig. 13c (CAPS04 vs. MPAS04), ETSs remain consistent through 36 hours lead time between FV3 and MPAS. However, the FV3 is notably better between 36 and 42 hours, and nearly reaches statistical significance again between 60 and 66 hours.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

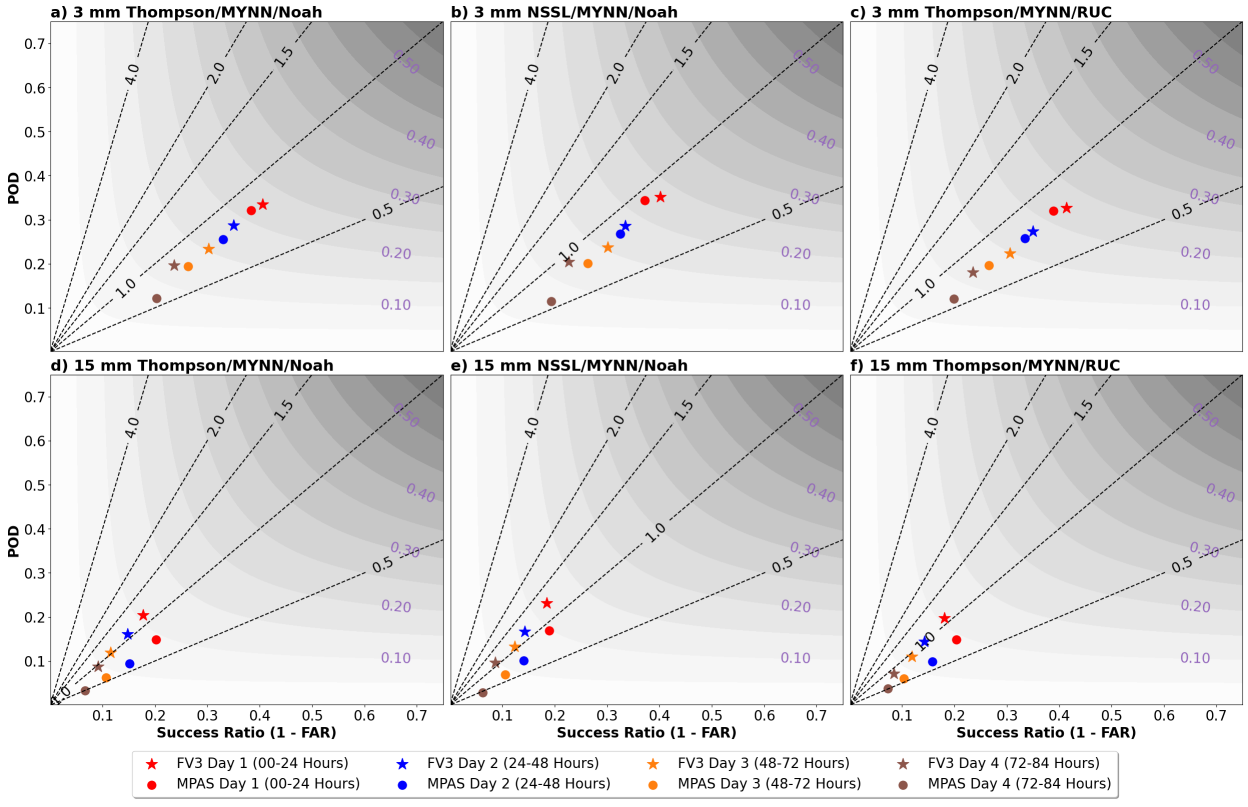


FIG. 12. Daily-aggregated performance diagrams for comparisons of MPAS (circular markers) and FV3-LAM (star markers) for a-c) 3 mm and d-f) 15 mm 6-hour precipitation. Physics configurations are listed in the sub-plot titles. Color codes for forecast leads are red for day 1, blue for day 2, orange for day 3, and brown for day 4.

Figure 14 shows the FSS for 99P precipitation for the CAPS FV3 and MPAS member pairs. Comparing CAPS01 (Fig. 14a) to MPAS01 (Fig. 14d), the MSS on day 1 is quite similar between FV3 and MPAS. Meanwhile on day 2, MSS increases by over 100 km when using MPAS01 instead of CAPS01. This is a notably poorer result, exceeding those found in the saSAS comparisons in Fig. 4. Day 3 MSS is 356 km for CAPS01, while MPAS01 has MSS greater than 400 km on days 3 or 4. In Fig. 14b and e, there is substantial degradation on day 2 when using MPAS with NSSL microphysics compared to the FV3 core, while day 1 remains less affected. Similar results to CAPS/MPAS01 are found for days 3 and 4. For CAPS04 and MPAS04 in Fig. 14c and f, day-2 MSS again increases by nearly 75 km when MPAS is used, along with similar day 3 and 4 degradations. These results suggest that using the MPAS with similar physics and boundary conditions to the FV3 employed in the 2025 CAPS FFaIR ensemble yields lower skill for precipitation on days 2, 3, and

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

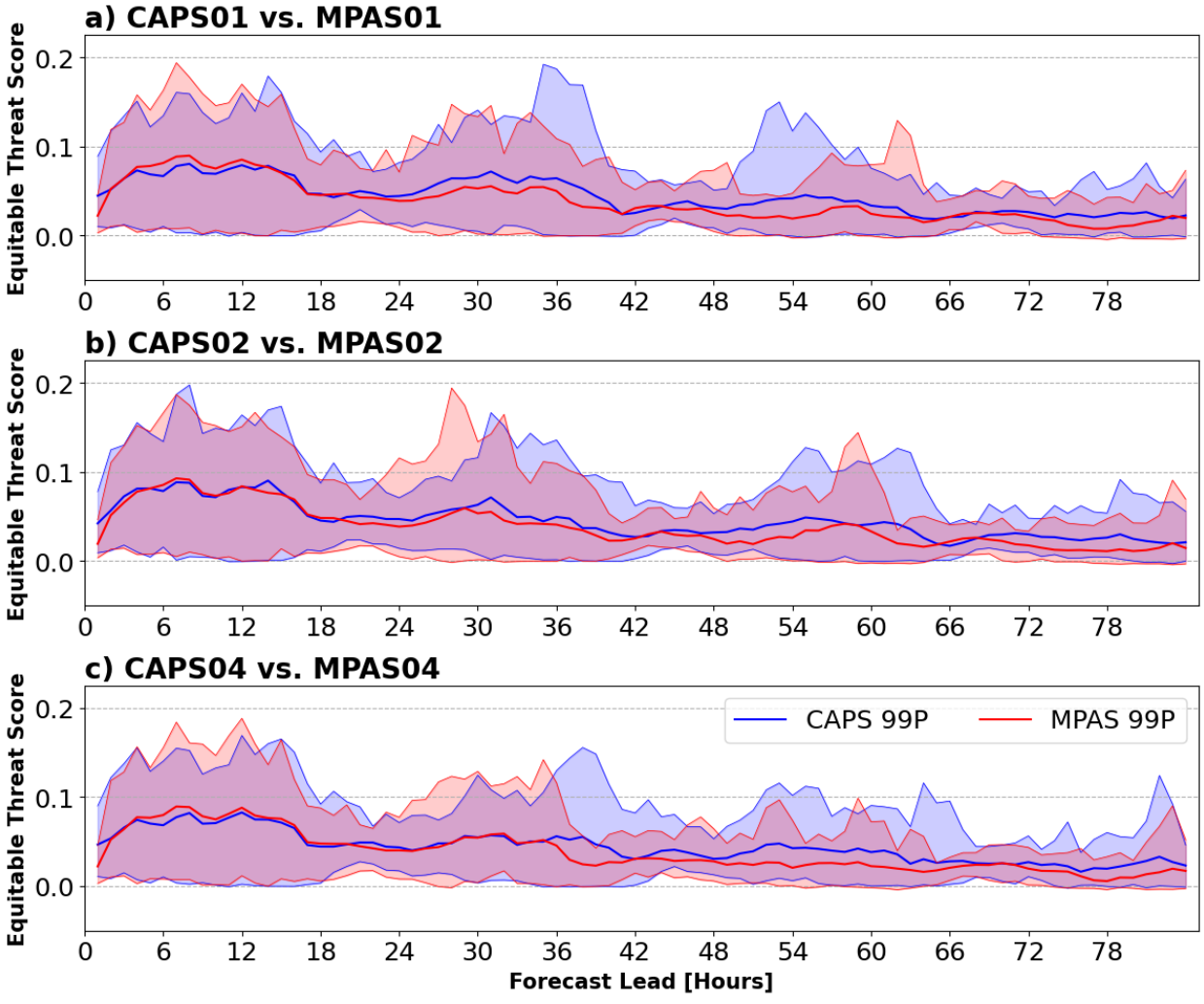


FIG. 13. As in Fig. 3, but comparing CAPS FV3 (blue) and MPAS (red) members.

4. Day 1 results are generally comparable between the two, similar to the findings of Berrington et al. (2026).

It should be emphasized again that the MPAS-based CAM forecast system is still very much in development (Alexander et al. 2023; Potvin et al. 2025) and that this study uses an earlier MPAS version than that evaluated in Trojniak et al. (2025). The latter study's MPAS members also used differing ICs and LBCs, such as the RAP analysis, than MPAS01, MPAS02, and MPAS04 in this study, and were evaluated for more than just the 0000 UTC cycle (such as 1200 UTC). Still, these results suggest that further optimization and tuning with the MPAS-based configurations,

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

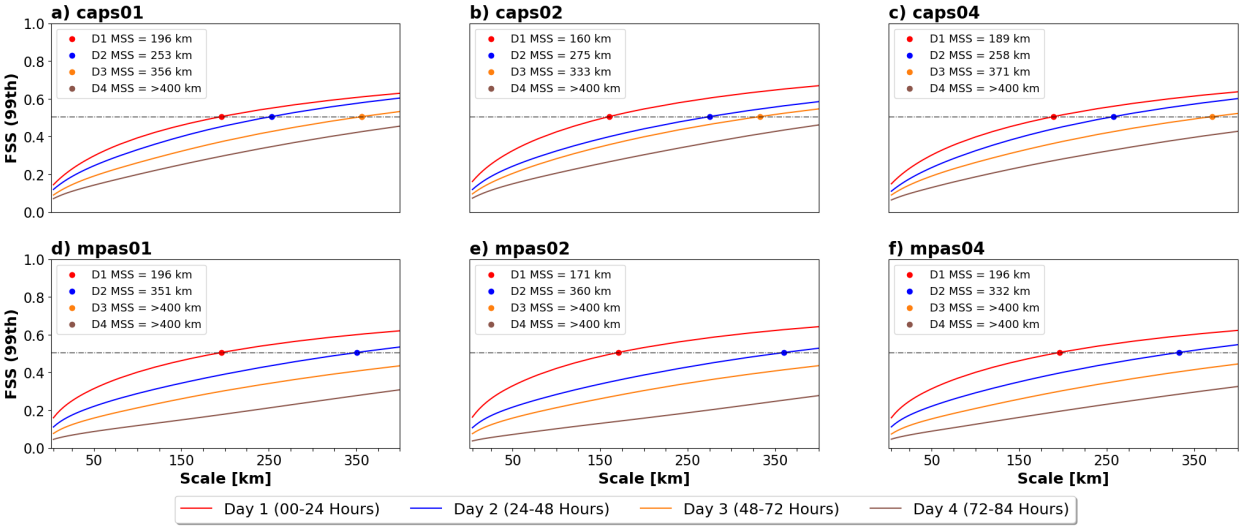


FIG. 14. As in Fig. 4, but comparing CAPS FV3 and MPAS members.

at least for warm season precipitation, and further tests will be necessary before MPAS is used operationally for RRFSv2.

5. Conclusion and Discussion

This study presents an evaluation of the 2025 CAPS FFaIR model suite, which is composed of 8 FV3-LAM members and 3 MPAS members. The experiment was designed, in part, to address two questions that were raised by previous studies (Trojniak and Correia 2024; Berrington et al. 2026) that evaluated the 2024 CAPS FFaIR ensemble. The first is to determine the impact of activating the saSAS cumulus scheme (CU ON) on precipitation verification in the FV3 members. The second question is to evaluate the MPAS dynamic core against the FV3-LAM dynamic core in members with similar physics between them.

Results indicate that when the saSAS cumulus scheme is employed in FV3-based members, models typically perform comparably or worse than if no convective scheme was used, when using CONUS-aggregated verification metrics for ETS, performance diagrams, and FSS. While differences are minimal between CAPS01–03 and their CU ON counterparts, there are more substantial differences between CAPS04 and CAPS08, which employ the RUC LSM. CAPS08 is statistically significantly drier than CAPS04 at the 15 mm precipitation threshold, and also shows greater degradation in FSS. Spatial ETS verifications suggest that some of this degradation in skill

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

between CAPS04 and CAPS08 can be attributed to errors in the southern United States, where WFTs are common during the warm season. A deeper investigation into the reasons behind the RUC LSM-saSAS interaction reveals that stronger vertical velocities in the afternoon and evening are being suppressed more when the RUC LSM is employed with saSAS utilized compared to when it is not. Additionally, it is found that surface latent heat flux is enhanced across the forested terrain of the southeastern United States with the RUC LSM compared to the Noah LSM. The aggressive activation of saSAS leads to high frequency biases in light precipitation (Trojniak and Correia 2024) and possible feedback into latent heat flux in the RUC LSM through evapotranspiration. Since stronger vertical velocities (convection) are suppressed, this results in greater dry biases at high precipitation thresholds in CAPS08 than in CAPS04. These findings suggest that the interaction between cumulus schemes and land surface models should be investigated more in future model (e.g. RRFS) development, especially at convection-allowing or "gray-zone" resolutions. More broadly, physics implementations should be thoroughly assessed and fine-tuned prior to implementation in operational products.

The MPAS, using identical physics to several FV3 members in the CAPS ensemble, does not perform as well at either precipitation threshold. Trojniak et al. (2025) also documented dry precipitation biases using MPAS in their evaluation. Therefore, fine-tuning both of the physics suites (Heinzeller et al. 2023), and the MPAS dynamic core as a whole, will be necessary in future work when it replaces the FV3 core in RRFSv2, keeping in mind that optimal configurations for FV3 may not be the same for MPAS.

Ultimately, a deeper investigation into the causes of precipitation forecast skill differences with physics changes is recommended, especially if cumulus schemes continue to be used in operational CAM products. As seen here, feedbacks with the choice of land surface model may play a role in determining forecast skill. Given the differences in latent heat flux between the Noah and RUC land surface models themselves, a separate (or concurrent) study could determine which processes are driving it, especially over crops and forested terrain over North America (Constantinidou et al. 2020). Such further studies could further inform CAM development, and hopefully result in a final configuration that performs well in all seasons.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

Acknowledgments. This work was primarily supported by the NOAA Testbed grant NA22OAR4590522 and NOAA UFS R2O Program grant NA21OAR4320204. The CAPS real-time ensemble forecast members were run on the Frontera supercomputer at the Texas Advanced Computing Center (TACC) at the University of Texas-Austin, a National Science Foundation-supported high performance computing center. We thank the three anonymous reviewers and editor, who's helpful comments improved the manuscript greatly from its initial 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/). MRMS QPE analyses can be downloaded from Amazon Web Services (https://registry.opendata.aws/noaa-mrms-pds/). CAPS FV3-LAM and CAPS MPAS precipitation forecasts are available via OSF at https://osf.io/4zbxw/.

References

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Weather and Forecasting*, **18 (5)**, https://doi.org/10.1175/1520-0434(2003)018⟨0918:SOPFSS⟩2.0.CO;2.

Alexander, C. R., J. R. Carley, and M. Pyle, 2023: The Rapid Refresh Forecast System: Looking Beyond the First Operational Version. *28th Conference on Numerical Weather Prediction*, Wisconsin, Madison.

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. *Monthly Weather Review*, **148 (11)**, https://doi.org/10.1175/MWR-D-20-0066.1.

Ashley, S. T., and W. S. Ashley, 2008: Flood fatalities in the United States. *Journal of Applied Meteorology and Climatology*, **47 (3)**, https://doi.org/10.1175/2007JAMC1611.1.

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. *Bulletin of the American Meteorological Society*, **96 (11)**, https://doi.org/10.1175/BAMS-D-14-00201.1.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

Berrington, A. H., K. A. Brewster, N. Snook, M. Xue, J. Park, C. Liu, S. Trojniak, and J. Correia, 2026: Evaluation of an Experimental 15-Member FV3-LAM Ensemble for the Summer 2024 Flash Flood and Intense Rainfall Experiment. *Weather and Forecasting*, **41**, 1737–1755, https://doi.org/10.1175/WAF-D-25-0256.1.

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. *Journal of Advances in Modeling Earth Systems*, **13 (6)**, https://doi.org/10.1029/2021MS002483.

Brown, B., and Coauthors, 2021: The model evaluation tools (MET): More than a decade of community-supported forecast verification. *Bulletin of the American Meteorological Society*, **102 (4)**, https://doi.org/10.1175/BAMS-D-19-0093.1.

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. Tech. rep., National Centers for Environmental Prediction, 88 pp.

Clark, A. J., 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. *Weather and Forecasting*, **25 (5)**, https://doi.org/10.1175/2010WAF2222404.1.

Constantinidou, K., P. Hadjinicolaou, G. Zittis, and J. Lelieveld, 2020: Sensitivity of simulated climate over the MENA region related to different land surface schemes in the WRF model. *Theoretical and Applied Climatology*, **141 (3-4)**, https://doi.org/10.1007/s00704-020-03258-5.

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 model. *Journal of Geophysical Research: Atmospheres*, **108 (22)**, 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. *Geoscientific Model Development*, **14 (9)**, https://doi.org/10.5194/gmd-14-5393-2021.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

Han, J., and C. S. Bretherton, 2019: TKE-based moist eddy-diffusivity mass-flux (EDMF) parameterization for vertical turbulent mixing. *Weather and Forecasting*, **34 (4)**, https://doi.org/10.1175/WAF-D-18-0146.1.

Han, J., 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. *Weather and Forecasting*, **32 (5)**, https://doi.org/10.1175/WAF-D-17-0046.1.

Heinzeller, D., L. Bernardet, G. Firl, M. Zhang, X. Sun, and M. Ek, 2023: The Common Community Physics Package (CCPP) Framework v6. *Geoscientific Model Development*, **16 (8)**, https://doi.org/10.5194/gmd-16-2235-2023.

Jeworrek, J., G. West, and R. Stull, 2019: Evaluation of cumulus and microphysics parameterizations in WRF across the convective gray zone. *Weather and Forecasting*, **34 (4)**, 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 a FV3-LAM regional ensemble during the 2022 NOAA Hazardous Weather Testbed Spring Forecasting Experiment. *Weather and Forecasting*, https://doi.org/10.1175/WAF-D-24-0034.1, URL https://journals.ametsoc.org/view/journals/wefo/aop/WAF-D-24-0034.1/WAF-D-24-0034.1.xml.

Joliffe, I., and D. e. Stephenson, 2012: Forecast Verification: A Practitioner's Guide. *Atmospheric Science, Second Edition*.

Kwon, Y. C., and S. Y. Hong, 2017: A mass-flux cumulus parameterization scheme across gray-zone resolutions. *Monthly Weather Review*, **145 (2)**, https://doi.org/10.1175/MWR-D-16-0034.1.

Lin, S. J., 2004: A "vertically Lagrangian" finite-volume dynamical core for global models. *Monthly Weather Review*, **132 (10)**, https://doi.org/10.1175/1520-0493(2004)132⟨2293:AVLFDC⟩2.0.CO;2.

Mansell, E. R., C. L. Ziegler, and E. C. Bruning, 2010: Simulated electrification of a small thunderstorm with two-moment bulk microphysics. *Journal of the Atmospheric Sciences*, **67 (1)**, https://doi.org/10.1175/2009JAS2965.1.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

Mesinger, F., 1996: Improvements in quantitative precipitation forecasts with the eta regional model at the national centers for environmental prediction: The 48-km upgrade. *Bulletin of the American Meteorological Society*, **77 (11)**, https://doi.org/10.1175/1520-0477(1996)077⟨2637:IIQPFW⟩2.0.CO;2.

Miller, P., and N. Debbage, 2023: Weakly Forced Thunderstorms in the Southeast US Are Stronger Near Urban Areas. *Geophysical Research Letters*, **50 (18)**, https://doi.org/10.1029/2023GL105081.

Miller, P. W., and T. L. Mote, 2017: Standardizing the definition of a "pulse" thunderstorm. https://doi.org/10.1175/BAMS-D-16-0064.1.

Nakanishi, M., and H. Niino, 2009: Development of an improved turbulence closure model for the atmospheric boundary layer. *Journal of the Meteorological Society of Japan*, **87 (5)**, 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. *Weather and Forecasting*, **31 (2)**, 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. *AGU Fall Meeting Abstracts*, Vol. 2022, 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. *Journal of Geophysical Research Atmospheres*, **116 (12)**, https://doi.org/10.1029/2010JD015139.

Park, J., M. Xue, C. Liu, 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 HMT Summer FFaIR Forecast Experiment. *105th AMS Annual Meeting*, New Orleans, LA, URL https://ams.confex.com/ams/105ANNUAL/meetingapp.cgi/Session/69613.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

Potvin, C. K., and Coauthors, 2025: Storm-Based Verification and Intercomparison of Forecasts during May and June 2023 from the HRRR, RRFS, NSSL MPAS, and C-SHiELD Models. *Weather and Forecasting*, **41 (1)**, https://doi.org/10.1175/waf-d-25-0019.1.

Roberts, B., B. Gallo, I. Jirak, and A. Clark, 2019: The High Resolution Ensemble Forecast (HREF) system: Applications and Performance for Forecasting Convective Storms. *AGU Fall Meeting Abstracts*, Vol. 2019, A31O–2797.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, **136 (1)**, https://doi.org/10.1175/2007MWR2123.1.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Weather and Forecasting*, **24 (2)**, https://doi.org/10.1175/2008WAF2222159.1.

Sakradzija, M., A. Seifert, and A. Dipankar, 2016: A stochastic scale-aware parameterization of shallow cumulus convection across the convective gray zone. *Journal of Advances in Modeling Earth Systems*, **8 (2)**, https://doi.org/10.1002/2016MS000634.

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. *Monthly Weather Review*, **140 (9)**, 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. *Monthly Weather Review*, **144 (5)**, 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. *Natural Hazards*, **67 (2)**, https://doi.org/10.1007/s11069-013-0566-5.

Snook, N. A., 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). *Weather and Forecasting*, **in press**.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

Sun, J., and Coauthors, 2014: Use of nwp for nowcasting convective precipitation: Recent progress and challenges. *Bulletin of the American Meteorological Society*, **95 (3)**, https://doi.org/10.1175/BAMS-D-11-00263.1.

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. *Monthly Weather Review*, **150 (9)**, https://doi.org/10.1175/MWR-D-21-0331.1.

Thompson, G., and T. Eidhammer, 2014: A study of aerosol impacts on clouds and precipitation development in a large winter cyclone. *Journal of the Atmospheric Sciences*, **71 (10)**, https://doi.org/10.1175/JAS-D-13-0305.1.

Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Monthly Weather Review*, **136 (12)**, https://doi.org/10.1175/2008MWR2387.1.

Trojniak, S., and J. Correia, 2024: 2024 Flash Flood and Intense Rainfall (FFaIR) Final Report. Tech. rep., NOAA Weather Prediction Center, 113 pp. URL https://www.wpc.ncep.noaa.gov/hmt/Reports/FFaIR/2024{_}FFaIR{_}Final{_}Report.pdf.

Trojniak, S., J. Correia, and A. Coleman, 2025: 2025 Flash Flood and Intense Rainfall (FFaIR) Final Report. Tech. rep., NOAA Weather Prediction Center, 83 pp. URL https://www.wpc.ncep.noaa.gov/hmt/Reports/FFaIR/2025{_}FFaIR{_}Final{_}Report.pdf.

White, P., and P. A. Nelson, 2024: Evaluation of Sub-Hourly MRMS Quantitative Precipitation Estimates in Mountainous Terrain Using Machine Learning. *Water Resources Research*, **60 (12)**, https://doi.org/10.1029/2024WR037437.

Zhang, C., and Coauthors, 2019a: 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. *Geophysical Research Letters*, **46 (6)**, https://doi.org/10.1029/2018GL081702.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.

Zhang, F., Y. Qiang Sun, L. Magnusson, R. Buizza, S. J. Lin, J. H. Chen, and K. Emanuel, 2019b: What is the predictability limit of midlatitude weather? *Journal of the Atmospheric Sciences*, **76 (4)**, 1077–1091, https://doi.org/10.1175/JAS-D-18-0269.1.

Zhang, J., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. *Bulletin of the American Meteorological Society*, **97 (4)**, https://doi.org/10.1175/BAMS-D-14-00174.1.

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. *Environmental Modelling and Software*, **145**, 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. *Weather and Forecasting*, **37 (6)**, https://doi.org/10.1175/waf-d-21-0112.1.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 10/12/26 09:54 PM UTC

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-26-0084.1.