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)

Nathan Snook[1], Jun Park[1], Ming Xue[1,2], Keith A. Brewster[1,2], Marcus Johnson[1], Timothy Supinie[3], Xiao-Ming Hu[1], Jacob R. Carley[4], Shun Liu[4], and Ming Hu[5]

[1] *Center for Analysis and Prediction of Storms, Univ. of Oklahoma, Norman, OK*

[2] *School of Meteorology, Univ. of Oklahoma, Norman, OK*

[3] *NOAA Storm Prediction Center, Norman, OK*

[4] *NOAA Environmental Modeling Center, College Park, MD*

[5] *NOAA Global Systems Laboratory, Boulder, CO*

Corresponding author:

Nathan Snook, nsnook@ou.edu

ABSTRACT

To inform the optimization of the future operational Rapid Refresh Forecast System (RRFS), the Center for Analysis and Prediction of Storms (CAPS) performed three sets of CONUS-domain 3-km FV3-LAM ensemble forecasts with various configurations during the 2022 NOAA Hazardous Weather Testbed Spring Forecasting Experiment in real time. The first set used different physics parameterizations, the second set included additional initial and lateral boundary condition perturbations, and the third introduced additional stochastic physics perturbations. This study evaluates precipitation, temperature, dewpoint, and wind forecasts, and compares them to those of the operational High-Resolution Ensemble Forecast (HREF) and Global Ensemble Forecast System (GEFS). Precipitation forecasts are verified against NCEP Stage-IV precipitation analyses, while temperature, dewpoint, and wind forecasts are verified against URMA surface analyses and radiosonde observations.

Overall, the ensemble configurations tested are generally suitable for predicting spring-season convective rainfall. The CAPS forecasts generally outperform GEFS in terms of ETS, frequency bias, and area under the ROC curve, approaching (but not exceeding) the performance of HREF in some metrics. Including stochastic physics perturbations resulted in forecasts objectively very similar to those without such perturbations, except for a small but consistent positive impact on ensemble spread of surface variables throughout the 84-hour forecast period. The CAPS forecasts have a near-neutral to slightly negative bias in total precipitation coverage. Forecasts using the NSSL microphysics scheme have more total rainfall than forecasts using the Thompson scheme and Stage-IV analyses, while forecasts using the NOAH-MP land surface model generally have lower total precipitation than other forecasts.

1. Introduction

As part of the transition to the Unified Forecast System (UFS) framework for operational numerical weather prediction (NWP) systems, the National Oceanic and Atmospheric Administration (NOAA) is developing a rapidly-updating convection-allowing ensemble forecasting system—the Rapid Refresh Forecast System (RRFS; Banos et al 2022, Carley et al. 2023)—which is planned to facilitate the retirement of a large portion of operational high-resolution forecasting systems currently used by the National Weather Service (NWS). The RRFS will use the limited area variant (Black et al. 2021) of the Finite-Volume Cubed-Sphere (FV3) Model (FV3-LAM) (Lin 2004) and include an ensemble of forecasts optimized for short-range prediction of high-impact weather. To assist in optimizing the ensemble design and physics parameterization choices for the RRFS, the Center for the Analysis and Prediction of Storms (CAPS) at the University of Oklahoma has been running experimental FV3-LAM ensemble forecasts during the NOAA Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (SFE; Clark et al. 2020; Roberts et al. 2020) and the Hydrometerology Testbed (HMT) Flash Flood and Intense Rainfall Experiment (FFaIR) and Winter Weather Experiments (NOAA 2023) since 2018—the ensemble membership and model configuration of these forecast ensembles has varied from experiment to experiment. These experiments have demonstrated the ability of convection-allowing FV3-LAM ensembles to generate skillful forecasts, while also documenting systematic biases associated with different physics and land-surface model combinations (Zhang et al. 2019; Snook et al. 2019; Supinie et al. 2022; Hu et al. 2023).

During the 2022 HWT SFE, CAPS ran a 21-member FV3-LAM ensemble; member configurations were selected to use various combinations of land surface models (LSMs), and planetary boundary layer (PBL), surface layer, and microphysical schemes. The 2022 HWT SFE configurations evolved from those used in 2020 and 2021 (Supinie et al. 2022; Hu et al. 2023). In the 2022 ensemble, initial conditions from experimental RRFS EnVar and EnKF analyses (Carley et al. 2024) are used, instead of the GFS and Global Ensemble Forecast System (GEFS) analyses used in prior years. Ensemble analyses from the experimental RRFS EnKF run by the NOAA Global Systems Laboratory (GSL) are used to initialize perturbed-IC ensemble members; unperturbed members are initialized using 3DEnVar analysis with ensemble covariance provided by the EnKF system so that two systems are coupled. In this sense the CAPS FV3-LAM forecasts of 2022 can be considered "hot start" forecast runs. The experimental RRFS EnVar and EnKF analyses were run by GSL as hourly cycled DA systems as prototypes of the planned operational RRFS.

Multiple suites of physics parameterization schemes are chosen for the CAPS FV3-LAM forecasts, partly based on earlier performance evaluations (e.g., Supinie et al. 2022; Hu et al. 2023) and partly based on the likelihood of continued support and operational use by the NWS. Stochastic physics perturbations are also included for five ensemble members. Model configurations in this ensemble are designed to resemble current and proposed convection-allowing model configurations (these configurations are discussed in greater detail below in section 2a). The primary goal of this study is to examine the performance of these forecasts, including deterministic forecast performance of selected members and ensemble forecast performance of several sub-ensembles (i.e., subsets of the 21 members intended to investigate different ensemble design considerations), with emphasis on precipitation forecasts.

Among the 21 forecast members, five use the same initial condition (IC) and lateral boundary conditions (LBCs), with differences in physics parameterizations only, allowing for the evaluation of physics performance. These same physics combinations are also used in other sub-ensembles that introduce IC, LBC, and stochastic physics perturbations. The use of physics diversity has been shown to help improve the spread and probabilistic forecast performance of ensembles, especially at the mesoscale and convective scale (e.g., Stensrud et al. 2020; Clark et al. 2008; Berner et al. 2011), and is used in the current NWS operational High-Resolution Ensemble Forecast (HREF) system (Jirak et al. 2012; Roberts et al. 2019) which is also a multi-model ensemble. Real-time HWT ensemble forecasts produced by CAPS over the years have also used multiple physics configurations (e.g., Schwartz et al. 2008; Loken et al. 2019; Supinie et al. 2022; Hu et al. 2023).

Inclusion of stochastic physics perturbations has also been shown to improve ensemble characteristics; several perturbation approaches have been investigated in prior studies, including stochastic kinetic energy backscatter (SKEB; Berner et al. 2009), stochastic perturbations of physics tendencies (SPPT; Buizza et al. 1999, Palmer et al. 2009), stochastic perturbed humidity (SHUM; Thompkins and Berner 2008), stochastic parameter perturbations (SPP; Jankov et al., 2017, 2019), and cellular automata (Bengtsson et al. 2013). The current NCEP GEFS uses SPPT and SKEB perturbations (Zhou et al. 2022). Jankov et al. (2017, 2019) demonstrated that a single-physics ensemble with stochastic physics could perform comparably to a multi-physics ensemble for regional systems. Kalina et al. (2021) found in experimental configurations of the High-Resolution Rapid Refresh Ensemble (HRRRE) that SPP helped increase ensemble spread in surface forecast variables. They also found that SPP increased the reliability of near-term HRRRE precipitation forecasts but exacerbated a preexisting low-frequency bias in the prediction of heavy rainfall in HRRRE.

Important questions regarding stochastic physics perturbations remain to be solved. One such question is whether random perturbations introduced into physics parameters or physics tendencies cause deterioration in individual members' forecast performance. Another question is whether stochastic perturbations using a single physics suite can create sufficient spread within an ensemble—if so, the cost of having to develop and maintain multiple physics packages can be alleviated, allowing more resources to be devoted to the improvement of the best-performing physics suite. Thus far, the findings concerning the optimal ensemble configurations in terms of physics combinations and the use of stochastic perturbations are still inconclusive, and the answers are likely dependent on the specific models used, their implementation, and their applications. One subset of the CAPS ensemble examined in this study uses a combination of SKEB, SPPT and SHUM stochastic perturbations, offering us an opportunity to examine the effects of stochastic perturbations as well as to test the robustness of their implementations in combination with multiple physics parameterizations. It should be noted, however, that stochastic physics perturbations can be strongly impacted by the choice of model dynamic core, and further evaluation would be needed in the context of any specific future RRFS ensemble.

Though in this study we document and compare the performance of the CAPS ensembles to existing operational forecasts as a baseline, the primary focus is on the relative performance of the sub-ensembles and the effects of physics diversity, IC and LBC perturbations, and stochastic physics perturbations. These experiments have helped to inform future operational RRFS ensemble design; members of the RRFS development team directly consulted with CAPS, and the results presented in this study were highly influential in the construction of the initial version of RRFS.

The remainder of this study is organized as follows: the specifics of the 2022 CAPS HWT SFE ensemble forecasts, including the sub-ensemble design considerations and the verification datasets and methodologies, are discussed in section 2. Objective and subjective verifications of the ensembles using Stage-IV precipitation data, surface analyses, and sounding data are presented in section 3. Conclusions and operational implications are discussed in section 4.

2. FV3-LAM Configuration, Verification Data, and Methodology

a. Model configurations and ensemble design

The forecasts produced by CAPS during 2022 HWT SFE (referred to hereafter as the "CAPS ensemble") use a model domain shown in Fig. 1, which covers the contiguous United States (CONUS) with a horizontal grid spacing of ~3 km and 1821 × 1093 grid points together with 65 vertical layers. Forecasts were generated on weekdays from 2 May 2022 through 3 Jun. 2022, and were initialized at 0000 UTC and run for 84 hours of forecast time, providing 1200-1200 UTC daily forecasts for days 1, 2, and 3 (12-36, 36-60, and 60-84 hours of forecast time, respectively). Due to technical issues, the most common of which was unavailable or incomplete RRFS EnKF initial condition data, complete forecast ensemble data were only available for 13 of the 25 days during the 2022 HWT SFE period, as detailed below in section 2b. Post-processed model outputs were generated hourly using version 10.0.12 of the Unified Post-Processing (UPP) package developed and maintained primarily by the NOAA EMC (publicly available at https://github.com/NOAA-EMC/UPP); the UPP-processed outputs include a standard suite of 2D forecast fields which are used for the verifications and evaluations performed in this study. All computation was done on the Frontera supercomputer operated by the Texas Advanced Computing Center (TACC).

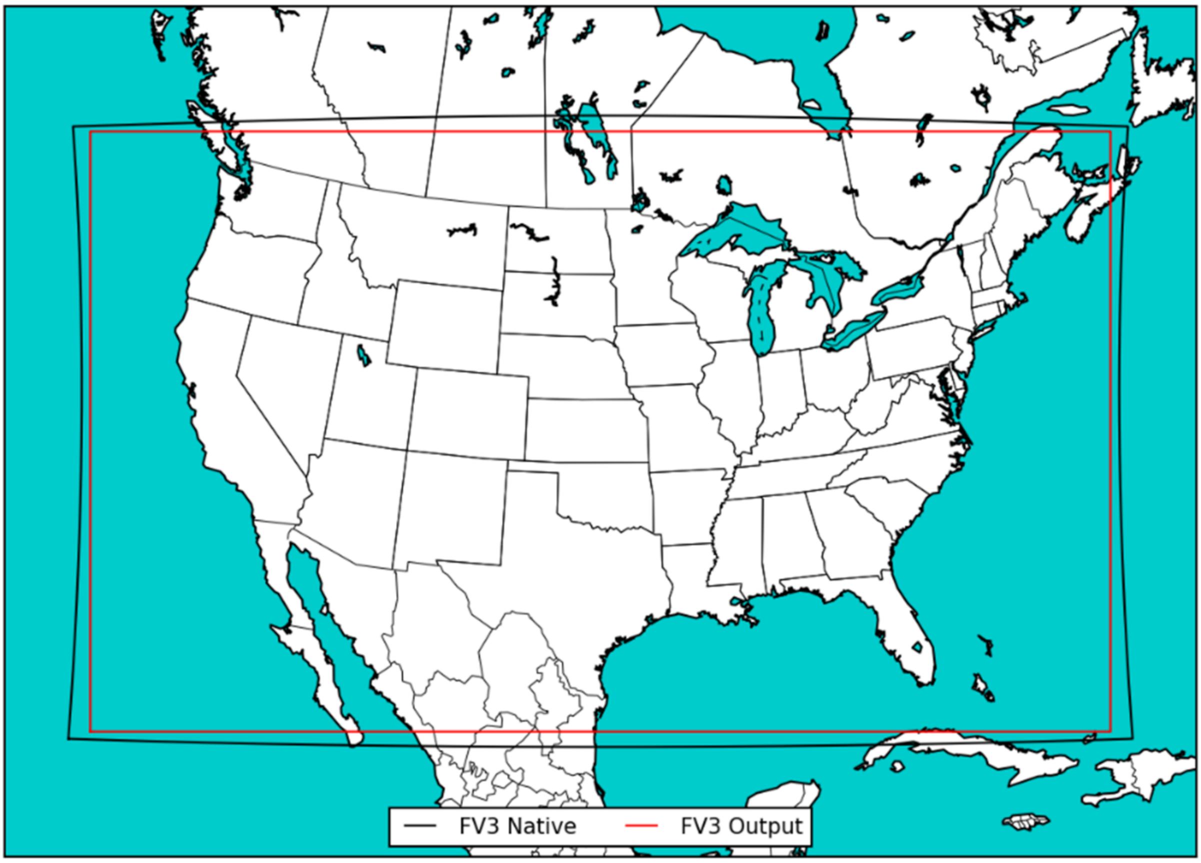

Fig. 1. Map showing the extent of the native grid (black box) and output grid (red box) used for the 2022 CAPS FV3-LAM ensemble. The output grid is interpolated from the native FV3-LAM grid which uses the Extended Schmidt Gnomonic grid (a cubed-sphere projection) to a Lambert conformal map projection.

The CAPS ensemble consists of 21 FV3-LAM forecast members (summarized in Table 1); diversity among ensemble members is achieved via a combination of variation in physics configurations, IC and LBC perturbations, and the use of stochastic physics perturbations. It should be noted that, while physics, IC/LBC, and stochastic perturbations are all intended to increase ensemble spread, the goal is not merely to maximize spread; over-dispersion among ensemble members is just as undesirable as under-dispersion. In practice, however, many convection-allowing ensembles tend to be under-dispersive, hence the focus in this study on methods for increasing ensemble dispersion. All members use a version of FV3-LAM checked out from https://github.com/NOAA-GSL/ufs-weather-model on 30 March 2022 (tagged as "BaselineC-20220331"). The Rapid Radiative Transfer Model for GCMs (RRTMG; Mlawer et al. 1997) radiation parameterization is used in all members. For members using the NOAH or NOAH-MP LSM, soil temperature and moisture were initialized via interpolation from RRFS initial conditions containing RUC LSM variables. For members using the NSSL microphysical scheme, microphysical variables not present in IC data are initialized via estimation using a gamma distribution. No cumulus parameterization scheme is used.

Experiment	Microphysics	PBL	Surface	LSM	IC/BC Source
Baseline member using GFS initial conditions					
M0B0L0_PG	Thompson	MYNN	MYNN	NOAH	GFS/GFS
Multi-Physics configurations: same initial and boundary conditions (P members)					
M0B0L0_P	Thompson	MYNN	MYNN	NOAH	RRFS CNTL/GFS
M1B0L0_P	NSSL	MYNN	MYNN	NOAH	RRFS CNTL/GFS
M0B0L1_P	Thompson	MYNN	GFS	NOAH-MP	RRFS CNTL/GFS
M1B2L2_P	NSSL	TKE-EDMF	GFS	RUC	RRFS CNTL/GFS
M0B2L1_P	Thompson	TKE-EDMF	GFS	NOAH-MP	RRFS CNTL/GFS
Physics + initial condition perturbation ensemble (PI sub-ensemble)					
M0B0L0_PI	Thompson	MYNN	MYNN	NOAH	RRFS01/GEFS m1
M0B1L0_PI	Thompson	Shin-Hong	GFS	NOAH	RRFS02/GEFS m2
M0B2L1_PI	Thompson	TKE-EDMF	GFS	NOAH-MP	RRFS03/GEFS m3
M0B0L1_PI	Thompson	MYNN	GFS	NOAH-MP	RRFS04/GEFS m4
M0B2L2_PI	Thompson	TKE-EDMF	GFS	RUC	RRFS05/GEFS m5
M1B0L0_PI	NSSL	MYNN	MYNN	NOAH	RRFS06/GEFS m6
M1B1L0_PI	NSSL	Shin-Hong	GFS	NOAH	RRFS07/GEFS m7
M1B2L1_PI	NSSL	TKE-EDMF	GFS	NOAH-MP	RRFS08/GEFS m8
M1B0L1_PI	NSSL	MYNN	GFS	NOAH-MP	RRFS09/GEFS m9
M1B2L2_PI	NSSL	TKE-EDMF	GFS	RUC	RRFS10/GEFS m10
Physics + initial condition + stochastic perturbation ensemble (PSI sub-ensemble)					
M0B0L0_PSI	Thompson	MYNN	MYNN	NOAH	RRFS01/GEFS m1
M1B0L0_PSI	NSSL	MYNN	MYNN	NOAH	RRFS06/GEFS m6
M0B0L1_PSI	Thompson	MYNN	GFS	NOAH-MP	RRFS04/GEFS m4
M1B2L2_PSI	NSSL	TKE-EDMF	GFS	RUC	RRFS10/GEFS m10
M0B2L1_PSI	Thompson	TKE-EDMF	GFS	NOAH-MP	RRFS03/GEFS m3

Table 1. Member configurations for the 2022 CAPS FV3-LAM ensemble. All members were run using FV3-LAM, and use the RRTMG radiation parameterization scheme. Member names contain information regarding the configuration of each member: the number after "M" indicates the microphysical scheme, the number after "B" indicates the PBL scheme, and the number after "L" indicates the land surface model. Suffixes indicate which sub-ensemble a member belongs to: "P" for members which are part of the core configurations, "PI" for members in the initial condition perturbation ensemble which uses initial and boundary conditions derived from GEFS RRFS, and "PSI" for members in the sub-ensemble using stochastic physics perturbations in addition to the perturbations of the "PI" members. The baseline configuration member, which uses GFS initial conditions, has a "PG" suffix.

The CAPS ensemble includes one member using a baseline configuration, and three groups of members investigating different aspects of ensemble design, including five members with core physics configurations and two sub-ensembles investigating different ensemble perturbation strategies. The baseline configuration, labelled “M0B0L0_PG” in Table 1, uses the Thompson microphysical scheme, MYNN PBL and surface layer schemes, and the NOAH LSM, and uses ICs and LBCs from operational GFS. A group of five members were run using core configurations focusing on physics diversity (hereafter referred to as the “physics” or “P” members, the latter designation referring to the suffixes in the naming convention used in Table 1). Each of these members uses identical ICs and LBCs, but employs a unique combination of microphysics, PBL schemes, and LSMs (Table 1). The physics combinations examined by the P members are designed to resemble current and proposed convection-allowing model configurations, including configurations similar to those of the GFS, the High-Resolution Rapid Refresh (HRRR), the NCEP North American Model (NAM), and the Warn-on-Forecast System (WoFS). LBCs used by the P members match those of the baseline configuration, and ICs are obtained from the experimental RRFS prototype ensemble variational (EnVar) system.

Within the core configurations of the P members, two microphysics schemes are examined—the National Severe Storms Laboratory (NSSL) scheme (Mansell et al. 2010; Mansell and Ziegler 2013), and the Thompson scheme (Thompson et al. 2004, 2008; Thompson and Eidhammer 2014). Three PBL schemes are examined, including the MYNN scheme (Nakanishi and Niino 2009), the scale-aware Shin-Hong scheme (Shin and Hong 2015), and the TKE-EDMF scheme (Han and Bretherton 2019). The LSMs investigated include the NOAH (Ek et al. 2003), NOAH-MP (Niu et al. 2011), and RUC (Smirnova et al. 2016) LSMs. For further discussion of the implementation of these schemes within the FV3-LAM version used in this study, we refer the reader to Supinie et al. (2022).

One sub-ensemble adds IC and LBC diversity (hereafter referred to as the “IC/LBC perturbation” or the “PI” sub-ensemble). In the PI sub-ensemble, initial conditions for each member are provided by the first 10 members of an experimental RRFS EnKF data assimilation (DA) system produced by GSL (Liu et al. 2023), and LBCs are obtained from GEFS member forecasts. This sub-ensemble also includes physics diversity, with five of the ten members using physics combinations matching those of the five physics sub-ensemble members.

The other sub-ensemble includes five members with stochastic physics perturbations in addition to physics and IC/LBC diversity (hereafter referred to as the “stochastic” or the “PSI” sub-ensemble). Stochastic perturbations are generated using a combination of SPPT, SKEB and SHUM, using a unique random seed for each member. SPPT perturbs the physics tendency terms in the equations of the *U* and *V* wind components, temperature, and water vapor mixing ratio using a standard deviation of 0.7 in the red noise process used to generate stochastic perturbations (Buizza et al. 1999). SKEB perturbations are applied to the *U*- and *V*-component wind fields with a standard deviation of 0.5 m s^{-1}. SHUM is applied within the boundary layer to perturbations of the 3-dimensional specific humidity field using a normalized perturbation coefficient of 0.006. For all stochastic methods, perturbations have a characteristic spatial length scale of 150 km and a characteristic time scale of six hours, and are updated hourly. These parameters were inherited from defaults in the UFS Short-range Weather Application package. We believe that these settings are suitable for applications involving convective rainfall, and note that they are similar, at least in terms of time and length scales, to those used in prior studies (e.g., Jankov et al. 2017). The five PSI sub-ensemble members have the same physics combinations as the P sub-ensemble members. The physics combinations of the five PSI sub-ensemble members also match those of five corresponding PI ensemble members, allowing for a clean comparison of these PI and PSI members to evaluate the impact of stochastic physics perturbations.

b. Data and methods for forecast evaluation

In this study, we verify and evaluate 0-84 hour forecasts of the CAPS ensemble and its sub-ensembles, with a particular focus on forecasts of precipitation. The operational High Resolution Ensemble Forecast (HREF) and GEFS forecasts are used as references for comparison. The HREF is an "ensemble of opportunity" convection-allowing ensemble forecasting system made up of 5 forecasting systems running at similar 3-4 km grid spacings; through time-lagging HREF achieves 10 ensemble members (Jirak et al. 2012; Roberts et al. 2020), and is the current operational standard at convection-allowing resolution over the CONUS. GEFS is included because HREF forecasts only run for 48 hours (allowing for only 36-h forecasts after the 12-h time-lag members are accounted for); a substantial dry bias in precipitation at higher intensity thresholds is expected of the GEFS given its much coarser grid spacing of 0.25 degrees. Evaluation of GEFS and HREF forecasts is not a goal of this study; they are included purely as baselines against which to evaluate the CAPS forecast sub-ensembles.

The evaluations presented in this study were generated using the grid_stat and ensemble_stat modules of version 10.0.0 of the Model Evaluation Tools (MET) software package (Brown et al. 2021, DTC 2023). The observation datasets used as "truth" for precipitation forecast evaluation are the NCEP Stage-IV 1-h, 6-h, and 24-h precipitation analyses (Nelson et al. 2016), which combine WSR-88D radar-estimated rainfall with rain gauge observations (including manual error correction). Because Stage-IV data are only valid over the United States, the verification is constrained to the land areas of the CONUS.

Surface temperature and wind forecasts are verified against the NOAA UnRestricted Mesoscale Analysis (URMA). URMA is very similar to the Real Time Mesoscale Analysis (RTMA; de Pondeca et al. 2011; Morris et al. 2020), but is generated 6 hours later to allow inclusion of late-arriving observations not used in RTMA (de Pondeca et al. 2015). While URMA does involve interpolation of observations to a grid we note that it does so via a sophisticated, anisotropic 2DVAR analysis scheme and is designed to have high fidelity to surface observations, making it suitable for validation and verification of near-surface forecast fields (Supinie et al. 2022), and has previously been used operationally by NWS as a "truth analysis" (de Pondeca et al. 2015). Verifications of vertical profiles for temperature, dewpoint, and wind forecasts are performed against radiosonde observations (RAOB) at ~70 NWS operational sites within the CONUS. For verification, all forecast data are interpolated to the grid of the analyses—either the Stage-IV grid (which has horizontal grid spacing of approximately 4.76 km) or the 2.5-km URMA grid—using MET's mass-conserving "BUDGET" interpolation method.

We examine the member-by-member performance of the CAPS ensemble using equitable threat score (ETS), frequency biases, and performance diagrams for precipitation forecasts at varying accumulated precipitation intensity thresholds, calculated using a 2 × 2 contingency table computed from model precipitation forecasts and observed Stage-IV precipitation accumulations. Ensemble consensus forecasts of precipitation are also evaluated, including the probability-matched mean (PM mean; Ebert 2001), and the patchwise localized PM mean (LPM mean; Snook et al. 2019, 2020) of the CAPS ensemble and its sub-ensembles. The simple ensemble mean is not included because of well-known biases arising from the smoothing effect of ensemble averaging. Accumulation periods of 6-h and 24-h are considered. We use both fixed and percentile thresholds for evaluating precipitation forecasts. An accumulation threshold of 1 mm is used as a proxy for geographic extent of precipitation, and a threshold of 25 mm is used to focus on prediction of moderate to heavy precipitation. For percentile thresholds, the 99th percentile of accumulated precipitation is used to focus on heavy precipitation, while the 90th percentile (for 24-h accumulations) and the 95th percentile (for 6-h accumulations) are used to include regions of lighter precipitation. The use of percentile thresholds allows us to eliminate the impact of frequency bias (Carafo et al. 2021). The ensemble mean RMSEs and ensemble spread of predicted temperature, dewpoint, and zonal wind are calculated against RAOB profile data to examine the performance of various ensembles. Similar comparisons are done for 2-m temperature, 2-m dewpoint, and 10-m zonal winds against URMA analyses.

Following Supinie et al. (2022), we choose for the sake of simplicity not to use a neighborhood-based contingency table for many of our evaluations. When a neighborhood is used, we apply the neighborhood maximum ensemble probability (NMEP; Schwartz and Sobash 2017). Employing increasing neighborhood radii generally increased forecast skill, but resulted in similar relative performance among ensembles and sub-ensembles, as we discuss below in section 3b. The forecast evaluations discussed in this study use forecasts initialized at 0000 UTC on each weekday from 13 May 2022 to 3 Jun. 2022 excluding 18 May, 30 May, and 2 June (for which complete data were not available), totaling 13 days of data. While this represents a small sample size and somewhat limits the robustness of the results of this study, technical limitations precluded the generation of additional forecast runs: forecasts were not run prior to 13 May due to unavailability of experimental RRFS initial conditions from GSL, while a combination of missing RRFS initial conditions and other technical issues resulted in incomplete forecast output on later missing days.

3. Forecast evaluation results

a. Deterministic and probabilistic evaluations of precipitation forecasts

There is little difference among individual members in ETS of precipitation exceeding the 90^{th} percentile threshold (Fig. 2); all members exhibit ETS declining from between 0.4 and 0.5 on Day 1 (Fig. 2a) to between 0.3 and 0.4 on Day 3 (Fig. 2c). Within each sub-ensemble, the members do not show significant differences, but the M0B2L1 member of each sub-ensemble consistently exhibits the highest ETS for Days 1, 2, and 3. As expected, the PM and LPM outperform individual members in terms of ETS.

Individual ensemble members do, however, exhibit substantial differences in their 90^{th} percentile values (Fig. 2d-f). The 90^{th} percentile of observed Stage-IV precipitation ranges from around 2.5 mm for Day 1 (Fig. 2d) to near 3.0 mm for Day 3 (Fig. 2f)—this variation in observed 90^{th} percentile is due to day-to-day variation in total precipitation accumulation. In contrast, the 90th percentile of individual member forecasts ranges from around 1.7 to 3.1 mm—for most members, the 90th percentile of precipitation is slightly to substantially below the observed 90th percentile. In all sub-ensembles and at all lead-times, the members using the NOAH-MP LSM (“L1” in the ensemble member naming scheme; see Table 1) consistently exhibit the lowest 90th percentile precipitation value, while members using the NSSL microphysics scheme (“M1” members; see Table 1) tend to exhibit the highest. When sub-ensemble consensus measures are considered, however, the 90th percentile of sub-ensemble PM and LPM means are generally in good agreement with the observed 90th percentile value (Fig. 2d-f). Comparing the PI and PSI sub-ensembles, the inclusion of stochastic perturbations has little impact on ETS or 90th percentile value; no significant differences are noted. Furthermore, the consistent performance across sub-ensembles in terms of relative ETS and 90th percentile value among physics configurations suggests that the choice of physics configuration has a stronger impact on the precipitation forecasts at the 90th percentile threshold than either IC/LBC perturbations or stochastic perturbations.

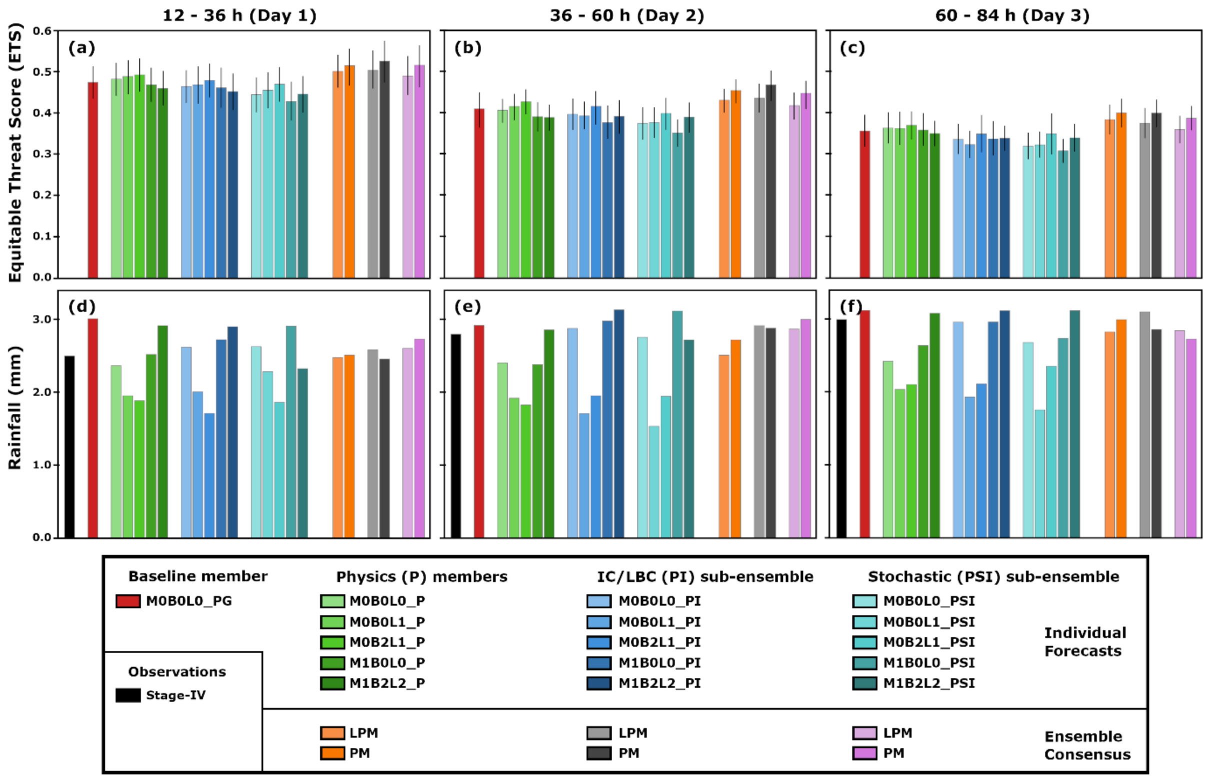

Fig. 2. (a-c) 90th percentile equitable threat score (ETS) and (d-f) 90th percentile rainfall amount for 24-h accumulated precipitation forecasts for (a, d) Day 1 (12-36 h), (b, e) Day 2 (36-60 h), and (c, f) Day 3 (60-84 h) forecasts, aggregated over all available cases from the 2022 HWT SFE period for the CAPS ensemble. Error bars for each member in panels (a-c) indicate the 5.0-95.0 percentile range using bootstrap resampling (10,000 resamples). The black bar in panels (d-f) indicates the 90th percentile of Stage-IV observed rainfall.

When we consider the 99th percentile for 24-h accumulated precipitation (Fig. 3), overall ETS values are somewhat lower than those at the 90th percentile, ranging from approximately 0.15-0.20 on Day 1 (Fig. 3a) to around 0.10 for most members on Day 3 (Fig. 3c). As at the 90th percentile threshold, differences among members are not significant, and the performance of individual members within the PI and PSI sub-ensembles is similar. Unlike at the 90th percentile, however, the 99th percentile of 24-h accumulated precipitation is higher than that of the observed Stage-IV precipitation for nearly all members on all three days (Fig. 3d-f). Taken together with the 90th percentile statistics presented in Fig. 2d-f, we find that most members have precipitation distributions which are lighter than observed for light to moderate rainfall, but which have heavier than observed accumulations in the most intense precipitation regions. These results align with the findings of prior studies, such as Zhou et al. (2019), which have indicated that FV3-based models tend to over-predict heavy precipitation.

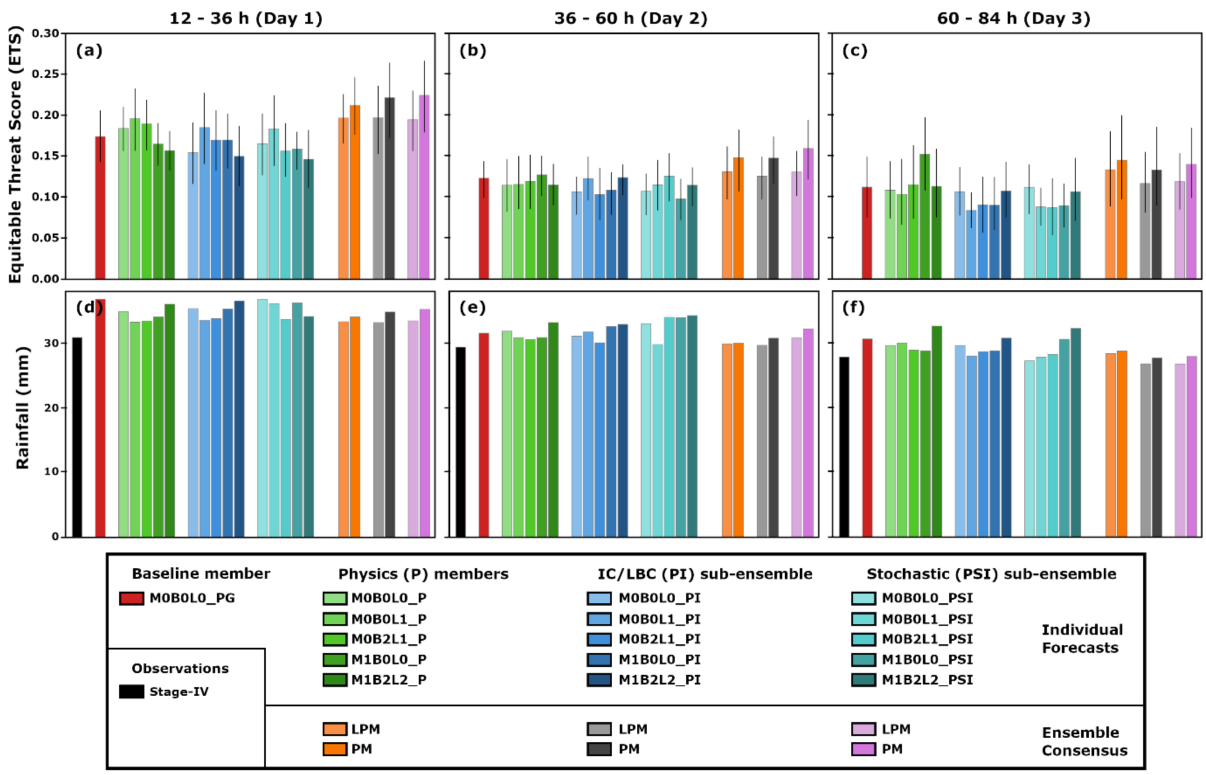

Fig. 3. As Fig. 2, but for the 99th percentile instead of the 90th.

The conclusions drawn from Figs. 2, 3 regarding precipitation distributions can be contextualized via an analysis of average total 24-h accumulated rainfall per Stage-IV grid cell (Fig. 4). In terms of total 24-h accumulated rainfall, the "M1" members, which use the NSSL microphysical scheme, stand out as having more total rainfall than the “M0” members, which use the Thompson microphysical scheme—particularly for M1B2L2. This pattern is consistent across sub-ensembles (i.e., it is present for “M1” members in the “P”, “PI”, and “PSI” sub-ensembles), and results in the “M1” members consistently producing more total rainfall than the Stage-IV analysis (Fig. 4). The “L1” members, which use the NOAH-MP land-surface model, are consistently among the lowest in terms of total rainfall.

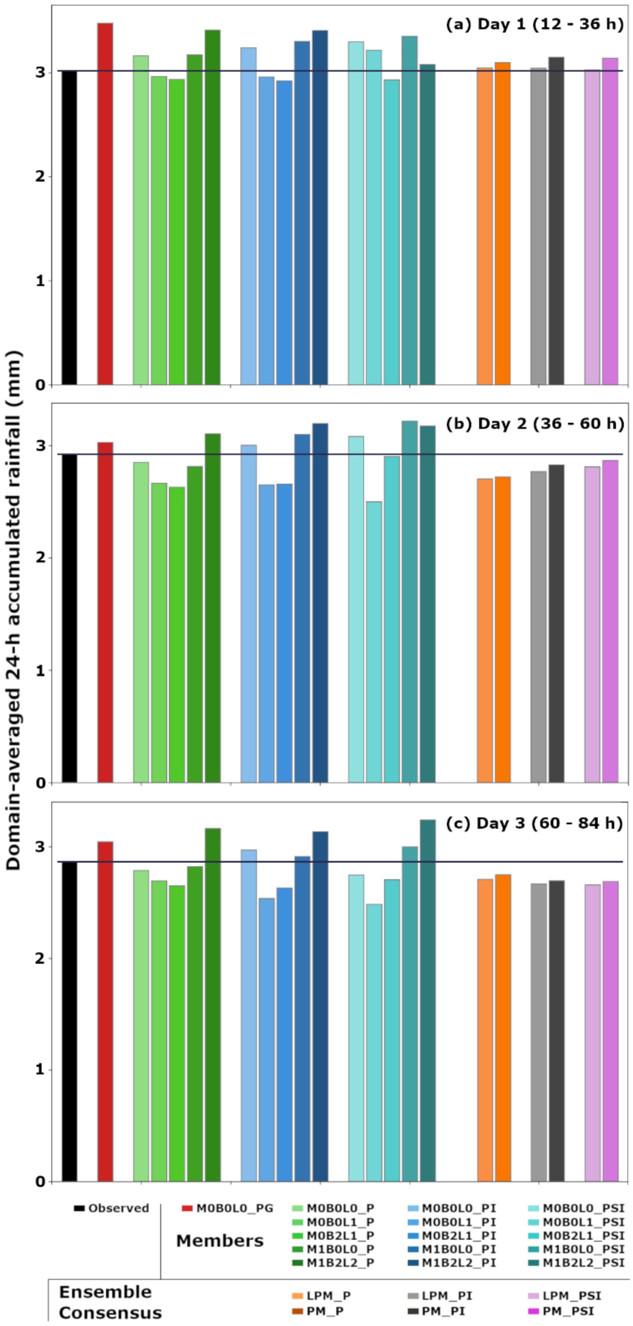

Fig. 4. Domain-averaged total 24-h accumulated precipitation for (a) 12-36 h, (b) 36-60 h, and (c) 60-84 h of forecast time for selected CAPS FV3 members and ensemble/sub-ensemble consensus forecasts (PM and LPM mean), compared to that of the Stage-IV analysis ("observed"). Color-coding of the plotted bars matches that used in Figs. 2 and 3. The horizontal black line in each panel is plotted at the observed value (to facilitate visual comparison of forecasts with observations).

Performance of 24-h rainfall accumulation forecasts is also examined in Fig. 5 via performance diagrams (Roebber 2009) using thresholds of 1.0 mm and 25.4 mm. Forecast performance improves toward the upper right of the diagram; forecasts along the one-to-one diagonal are unbiased. The performance diagrams shown in Fig. 5 use fixed rather than percentile-based precipitation thresholds, allowing forecast bias assessment. At the 1.0 mm threshold (Fig. 5a, c, e), GEFS members exhibit a high bias, while the CAPS FV3 and HREF members and CAPS sub-ensemble PM and LPMs exhibit near-neutral or slightly low biases.

At the 25.4 mm (1.0 inch) threshold (Fig. 5b, d, f), forecast accuracy is generally quite similar between the PI and PSI ensembles, with neither clearly outperforming the other. At this higher threshold, HREF members perform similarly to or slightly better than the PI and PSI members; both the CAPS sub-ensembles and the HREF exhibit a slight over-prediction bias on day 1 (Fig. 5b), and close-to-unbiased performance on days 2 and 3 (Fig. 5d, f). GEFS performs noticeably worse in both CSI and bias compared to the CAPS sub-ensembles and HREF, which is unsurprising given its much lower resolution. The overall performance of the CAPS sub-ensembles is similar to that noted by Johnson et al. (2023) for CAPS forecasts produced during the 2018 HWT SFE.

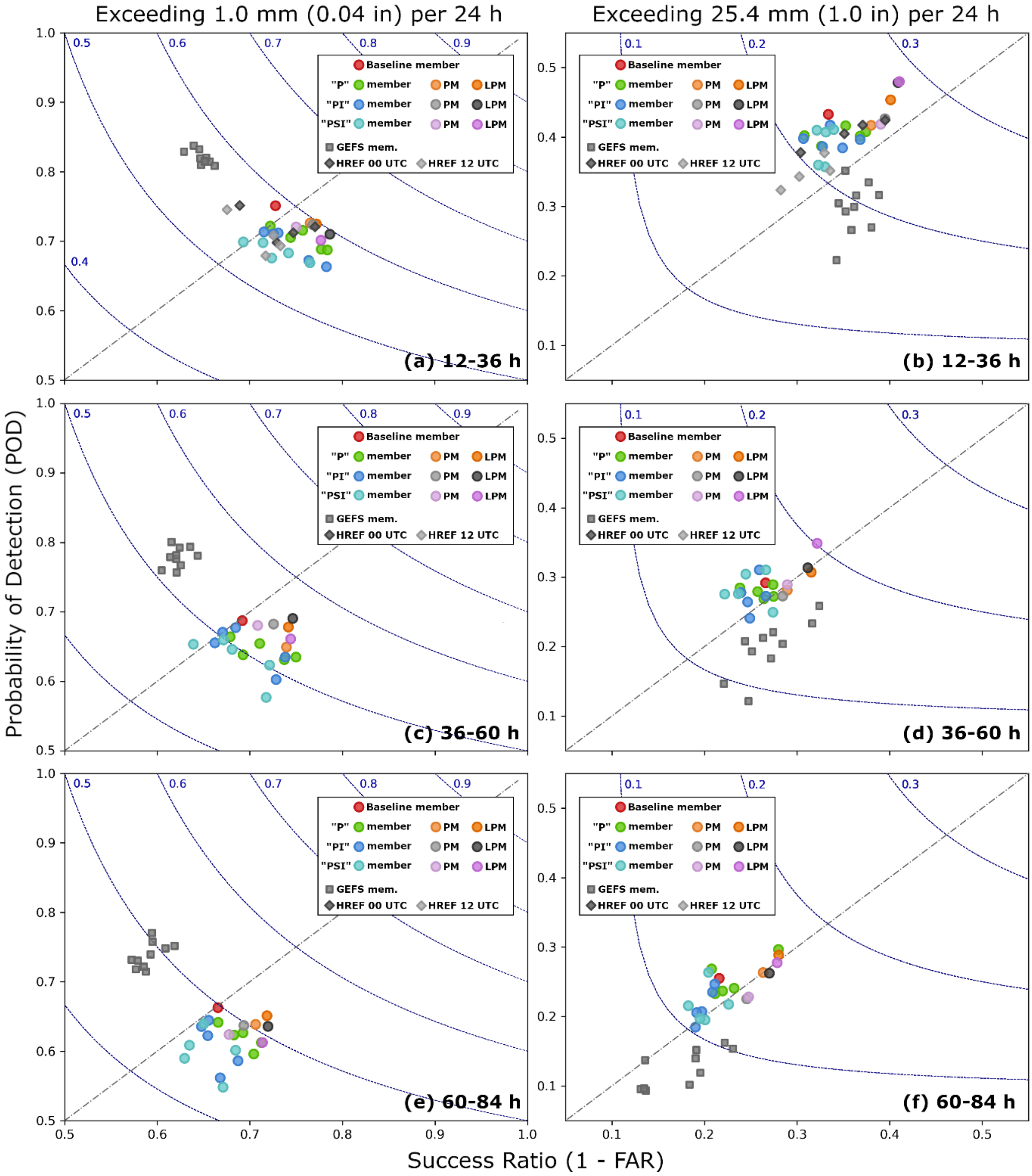

Fig. 5. Performance diagrams (zoomed-in) for forecasts of 24-h accumulated precipitation exceeding (a, c, e) 0.04 inches or (b, d, f) 1.00 inches at forecast lead times of (a, b) 12-36 h, (c, d) 36-60 h, and (e, f) 60-84 h using all available days from the 2022 HWT SFE period. The dark blue labelled contours indicate lines of constant critical success index (CSI), while the diagonal dotted line indicates a constant bias of 1.0. Color-coding of ensemble members is indicated by the legends in panel (a). Due to the shorter run-length of HREF, HREF data are shown only in panels (a) and (b). To facilitate readability, the left and right columns are zoomed in on different portions of the performance diagram.

The relative performances of the five-member PI and PSI sub-ensembles are compared to one another and to those of the five members of HREF without time-lagging and the five GEFS members which provided boundary conditions for the PSI sub-ensemble (see Table 1) in terms of area under the receiver operating characteristic (ROC) curve (AUC) (Mason 1982) for accumulated precipitation in Fig. 6. Shown are verification for 1-h (Fig. 6a, b) and 6-h (Fig. 6c, d) accumulation at 95th and 99th percentile thresholds. Due to limitations in the temporal resolution and format of available data, 1-h verification is shown only for the PI and PSI sub-ensembles. AUC measures resolution (i.e., the ability to distinguish between events and non-events), and is constructed by calculating the area beneath the curve when plotting probability of detection (POD) versus probability of false detection (POFD) (e.g., Clark et al. 2021).

At both percentile thresholds, the CAPS PI and PSI sub-ensembles exhibit qualitatively and quantitatively similar performance in terms of AUC, with only very minor differences between the two (Fig. 6). Overall patterns of forecast skill are similar for 1-h and 6-h accumulations, though the 1-h accumulations exhibit additional evidence of diurnal variation in forecast skill in the PI and PSI sub-ensembles, with higher forecast skill between 0600 and 1200 UTC and lower forecast skill between 1800 and 0000 UTC; this pattern is particularly notable at the 95th percentile threshold (Fig. 6a). When the CAPS PI and PSI sub-ensembles are compared to the operational HREF and GEFS ensembles, HREF exhibits superior performance in terms of AUC up to 12 hours of forecast time, and then performs similarly to or better than the CAPS sub-ensembles through the remainder of its forecast period (ending at 48 h of forecast time). The CAPS sub-ensembles do outperform GEFS, particularly at the 99th percentile threshold (Fig. 6d). The superior performance of the CAPS sub-ensembles and HREF compared to GEFS is not surprising, given the large resolution difference (~3 km vs. ~25 km). The inferior performance of coarser-resolution global models in predicting precipitation (especially heavy precipitation) compared to convection-allowing models has been well-documented in previous studies (e.g., Zhu et al. 2018).

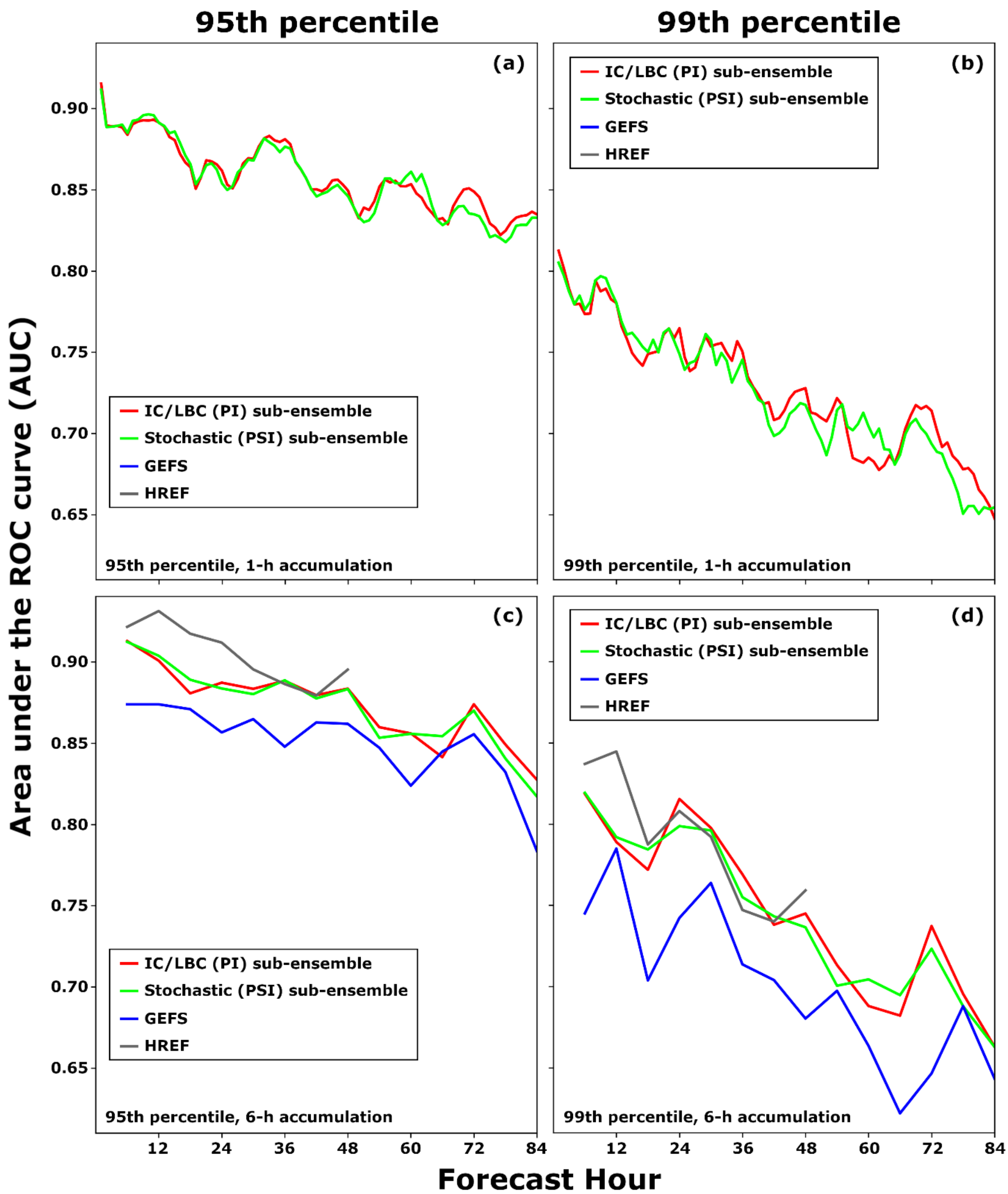

Fig. 6. Area under the receiver operating characteristic curve (ROC AUC) for forecasts of (a, b) 1-h accumulated precipitation and (c, d) 6-h accumulated precipitation exceeding (a, c) the 95th percentile or (b, d) the 99th percentile at various forecast lead times. Shown are the CAPS PI and PSI 5-member sub-ensembles compared to the five members of GEFS which provided boundary conditions for the CAPS FV3 PSI sub-ensemble and the 0000 UTC HREF. Values shown here are calculated using all days during the 2022 HWT SFE period with valid CAPS ensemble data. Due to the shorter run-length of HREF, HREF is shown only for forecast times of 6 to 36 hours.

The impact of using a neighborhood during probabilistic forecast generation is examined in Fig. 7, which compares the AUC of forecasts generated using a pixelwise (no neighborhood) ensemble approach to those generated using ensemble NMEP with neighborhood radii of 20 or 40 km. Only 0 and 40 km radii are used for GEFS, as a neighborhood of 20 km is smaller than the grid spacing of GEFS. Across neighborhood radii, HREF consistently exhibits the highest AUC, followed by the CAPS PI and PSI sub-ensembles which exhibit similar AUC performance to one another, slightly worse than HREF. GEFS consistently exhibits the lowest AUC of the ensembles examined. The impact of using a neighborhood via NMEP is uniformly positive in terms of AUC, but the increase in AUC is greater for the CAPS sub-ensembles and the HREF than it is for GEFS (Fig. 7c, g), likely because the relatively coarse GEFS grid results in fewer localized heavy rainfall areas for which forecasts would be greatly improved by using a 40 km neighborhood. The CAPS sub-ensembles and the HREF, which both have native horizontal grid spacing of around 3 km, exhibit similar magnitudes of increase in AUC going from a 0 km to 20 km to 40 km neighborhood.

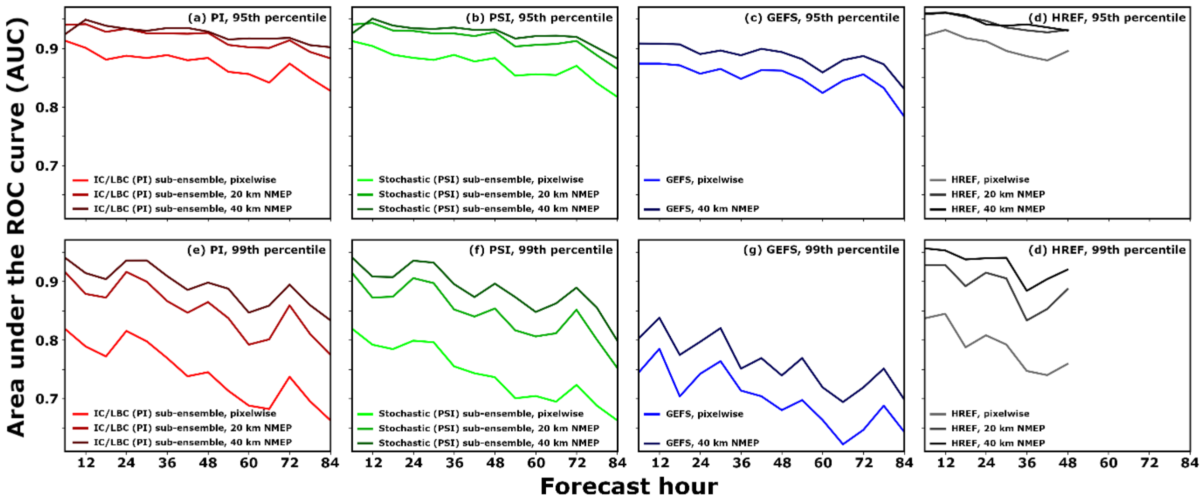

Fig. 7. Area under the receiver operating characteristic curve (ROC AUC) for forecasts of 6-h precipitation accumulation exceeding (a-d) the 95th percentile and (e-h) the 99th percentile at forecast lead times ranging from 6 hours to 84 hours for (a, e) the 5-member CAPS PI and (b, f) PSI sub-ensembles, as well as (c, g) the five members of GEFS which provided boundary conditions for the CAPS FV3 PSI sub-ensemble and (d, h) the 0000 UTC HREF. Curves are plotted using NMEP with neighborhood radii of 0, 20, and 40 km for PI, PSI, and HREF, and with radii of 0 and 40 km for GEFS. As in Fig. 6, values are calculated using all days during the 2022 HWT SFE period with valid CAPS ensemble data.

Fractions Skill Score (FSS; Roberts and Lean 2008) is another neighborhood-focused forecast skill metric which evaluates the degree of similarity between a forecast and observations at varying spatial scales. FSS varies between 0.0 and 1.0, with higher values indicating higher forecast skill. A forecast can be considered useful at a given scale if the FSS at that scale exceeds that which would be obtained at the grid scale from a forecast with probability at each point equal to the base rate of the event being evaluated (Roberts and Lean 2008). At the 90th percentile (Fig. 8), all CAPS forecast members exhibit skillful forecasts of 24-h accumulated precipitation in terms of FSS, with modest variation among members. In terms of ensemble consensus measures, the PM exhibits slightly higher skill than the LPM in terms of FSS at the 90th percentile, with the PI sub-ensemble slightly outperforming the PSI sub-ensemble. Individual HREF members perform similarly to the best-performing CAPS members on Day 1 (Fig. 8a), while GEFS members consistently exhibit lower FSS than either CAPS or HREF members, although even GEFS is skillful at all scales on Day 1 and 2, and at all but the smallest scales on Day 3.

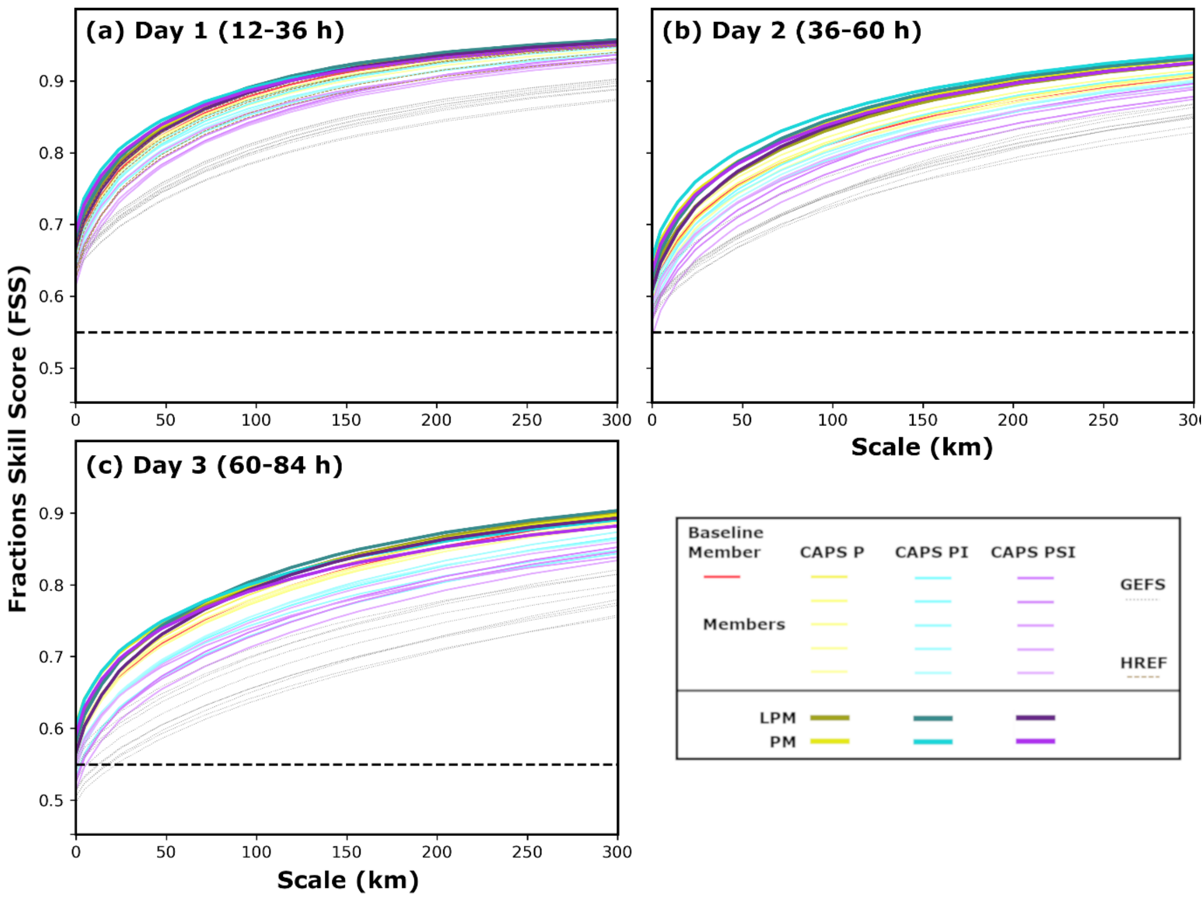

Fig. 8. Fractions Skill Score (FSS) for forecasts of 24-h precipitation accumulation exceeding the 90th percentile for (a) Day 1, (b) Day 2, and (c) Day 3 forecasts for individual CAPS ensemble members, the PM and LPM of the CAPS P, PI and PSI sub-ensembles, and corresponding members of the HREF and GEFS ensembles for comparison. The horizontal dashed line indicates the minimum useful value of FSS.

At the 99th percentile (Fig. 9), there is more substantial variation in FSS across the CAPS members, although neither PI nor PSI members consistently outperform one another, while HREF members perform comparably to better-performing members of the CAPS sub-ensembles (Fig. 9a). GEFS performs notably worse than either the CAPS or HREF members. By Day 2, no GEFS member exhibits a skillful FSS at any scale up to 300 km (Fig. 9b, c).

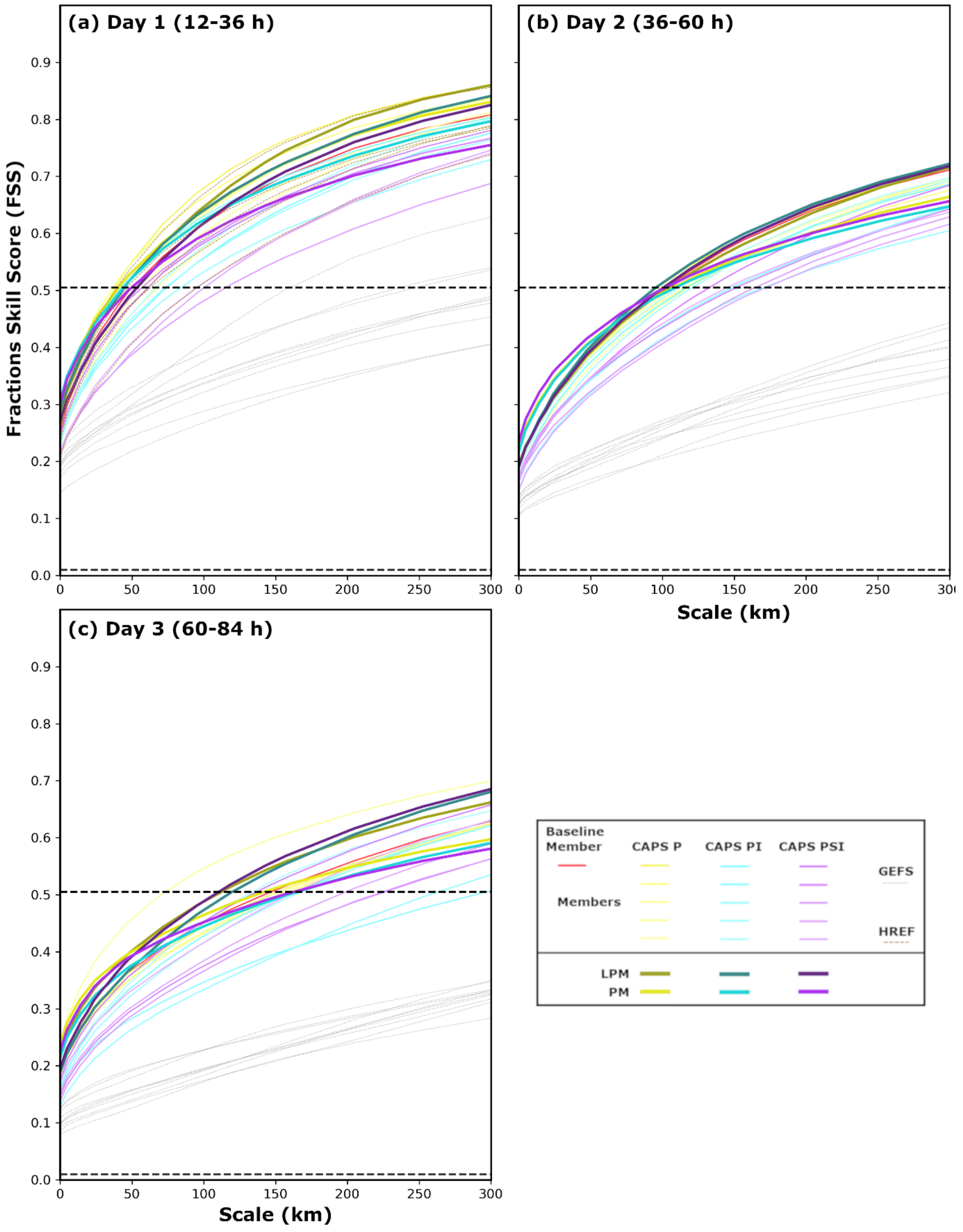

Fig. 9. As Fig. 8, but for the 99th percentile of 24-h accumulated rainfall instead of the 90th percentile. The lower horizontal dashed line in each panel indicates the base rate, while the upper horizontal dashed line indicates the minimum useful value of FSS.

In Fig. 10, rank histograms are presented for forecasts of 24-h accumulated precipitation. Rank histograms measure ensemble reliability. Ideally, the fraction of points within each rank should be similar, leading to a flat rank histogram (Hamill 2001). The PI and PSI sub-ensembles (Fig. 10a-f) exhibit nearly flat rank histograms for 24-h accumulated precipitation for days 1-3, with a slight downward slope towards higher ranks, indicating slight over-prediction of precipitation. When the CAPS P members are treated as a sub-ensemble, having no initial condition perturbations and relying entirely upon physics diversity to generate ensemble spread, under-dispersion is evident in the form of a U-shaped rank histogram (Fig. 10g-i), underscoring the importance of the initial condition perturbations applied in the PI and PSI sub-ensembles. Given the similarity between the PI and PSI sub-ensembles, it is likely that the initial condition perturbations impart greater impact than the stochastic perturbations. In contrast to the CAPS sub-ensembles, the GEFS 5-member sub-ensemble (Fig. 10j-l) shows a substantial over-forecasting bias on all days, with the strongest bias on day 1 (Fig. 10j). The HREF 0000 UTC members (Fig. 10m) exhibit better rank histogram behavior than GEFS, but with slightly greater over-forecasting bias than the CAPS sub-ensembles (evidenced by the decreasing occurrence of grid cells at increasing ranks in HREF, compared to the flatter rank histogram of the CAPS PI and PSI sub-ensembles).

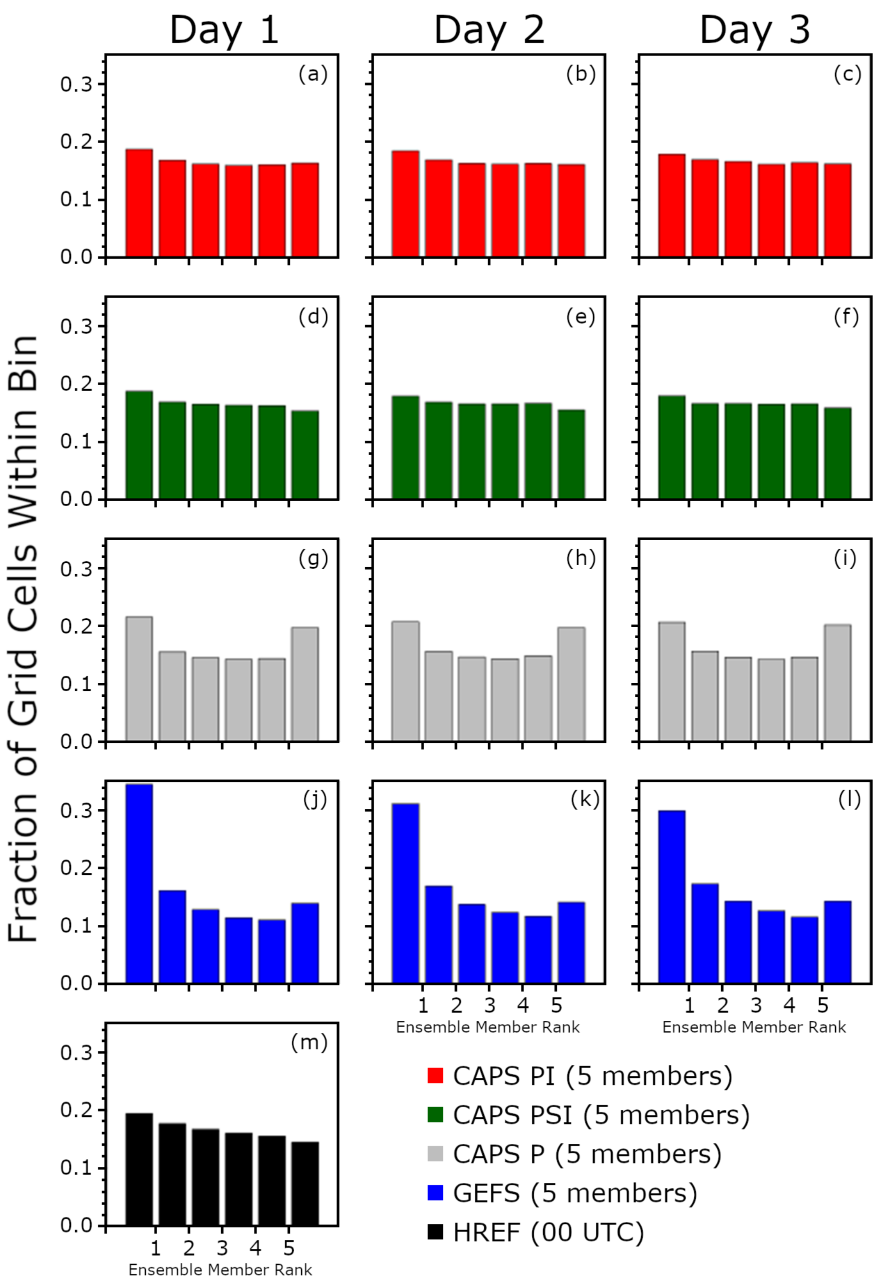

Fig. 10. Rank histograms of 24-h accumulated precipitation for (a-c) the CAPS PI sub-ensemble, (d-f) the CAPS PSI sub-ensemble, (g-i) the CAPS P sub-ensemble, (j-l) the five members of the GEFS which provided boundary conditions for the CAPS PSI 5-member sub-ensemble, and (m) the 00 UTC members of the HREF. Data are shown for forecast lead times of 12-36 h (left column), 36-60 h (center column), and 60-84 h (right column).

b. Evaluations of wind, temperature, and dewpoint forecasts

For 2 meter temperature (T), 2 meter dewpoint (T_d), and 10 meter zonal wind (U), the 5-member CAPS PI and PSI sub-ensembles consistently exhibit lower RMSEs and greater ensemble spread than GEFS, while HREF has the lowest RMSEs and highest spread throughout that the period for which HREF forecasts are available (Fig. 11). RMSE and spread curves exhibit a strong diurnal variation for T, T_d, and U, with maxima typically occurring around 0000 UTC (during the afternoon and evening hours over the CONUS when convective activity is typically near its peak). RMSEs of T_d and U are very similar between the PI and PSI sub-ensembles, while RMSE of T is slightly higher in the PSI sub-ensemble during the diurnal maximum. While statistics are not shown for the meridional wind component, performance is qualitatively similar to that of U (not shown). For all three variables ensemble spread is consistently slightly higher for the PSI sub-ensemble than for the PI sub-ensemble, indicating that stochastic perturbations do help improve the ensemble spread, with only small increases in RMSE at a few times. A small improvement in spread to RMSE ratio in the PSI sub-ensemble compared to the PI sub-ensemble is present throughout the forecast period out to the maximum forecast lead-time of 84 h.

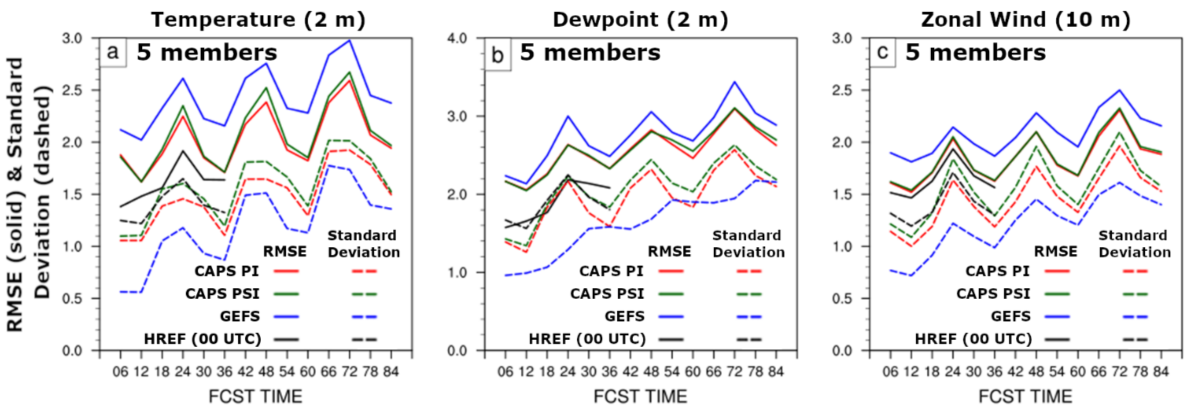

Fig. 11. RMSE and ensemble spread for forecasts of (a) 2-meter temperature (K), (b) 2-meter dewpoint (K), and (c) 10-meter zonal (u) wind component (ms^{-1}). Shown are the 5-member CAPS PI and PSI 5-member sub-ensembles compared to the five members of GEFS which provided boundary conditions for the CAPS FV3 PSI sub-ensemble and the 0000 UTC HREF.

Vertical profiles of RMSE and spread calculated against sounding data are plotted for T, T_d, and U at 36, 60, and 84 hours of forecast time in Fig. 12. While some differences between the performance of the CAPS PI and PSI sub-ensembles were noted near the surface (Fig. 11), in Fig. 12 the CAPS PI and PSI sub-ensembles generally exhibit similar performance, both performing slightly better than the GEFS. The most notable difference between the PI and PSI sub-ensembles is a slightly improved spread to RMSE ratio in the PSI sub-ensemble, primarily below 500 hPa. At 36 hours, HREF exhibits slightly smaller RMSEs for T and U (Figs. 12a,g), slightly smaller spread for U (Fig. 12g) and larger spread for T_d (Fig. 12d) than other ensembles.

Overall, HREF has better error-spread consistency, especially for T_d. Spread in temperature is slightly higher in GEFS than in the HREF or CAPS sub-ensembles at 60 and 84 hours, albeit with larger RMSE; this may reflect better tuning in the stochastic physics used by GEFS than in the CAPS PSI sub-ensemble.

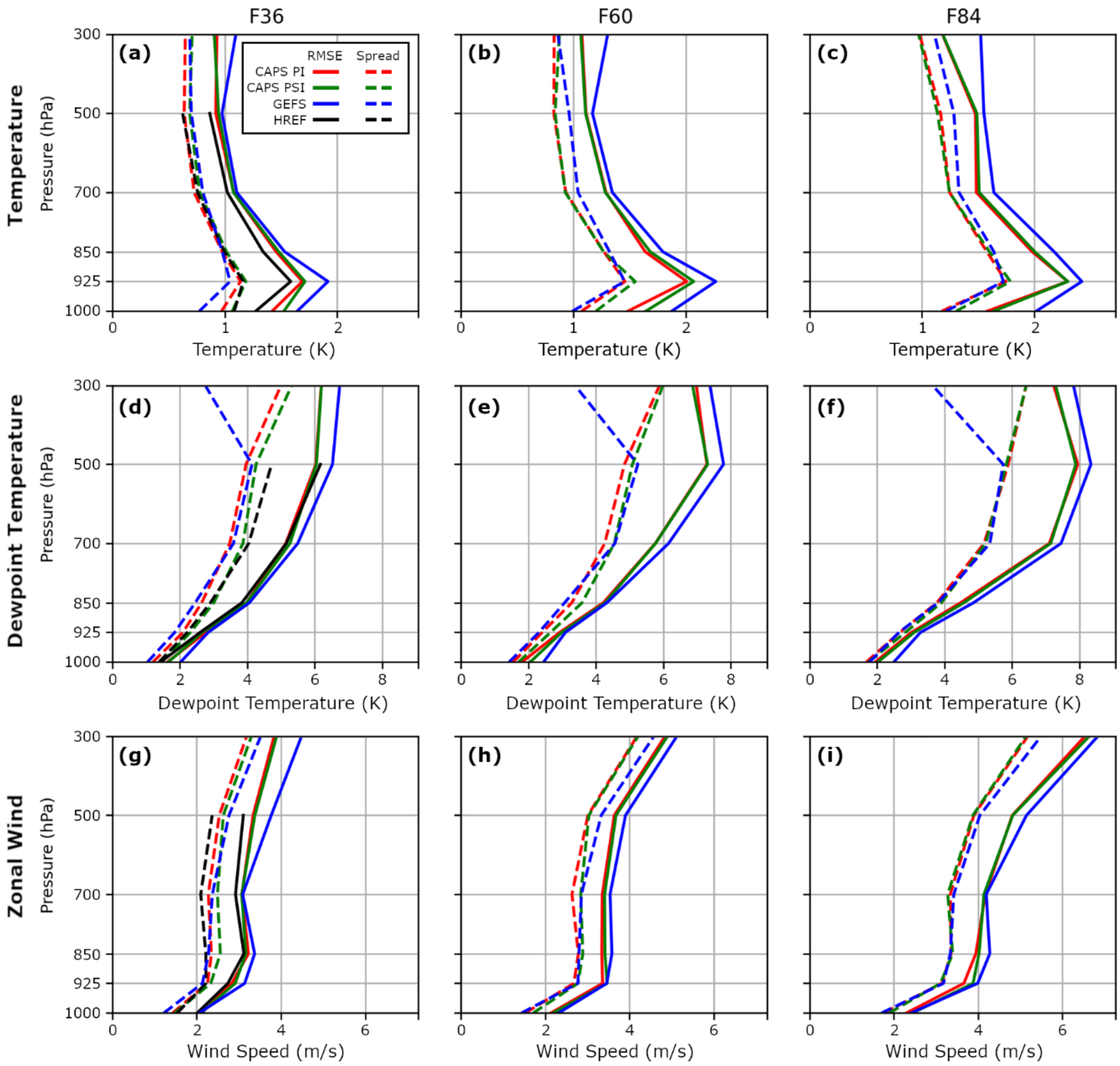

Fig. 12. Vertical RMSE and spread profiles of (a-c) temperature, (d-f) dewpoint, and (g-i) zonal wind component at forecast lead times of 36 h (left column), 60 h (center column), and 84 h (right column) for the 5-member CAPS PI and PSI sub-ensembles, as well as for the 0000 UTC HREF and five members of GEFS. HREF data are available only up to 500 hPa.

4. Summary and discussion

During the 2022 NOAA HWT SFE, CAPS ran an ensemble of 21 FV3-LAM forecasts in real-time using a variety of configurations with different combinations of microphysical schemes, LSMs, and PBL parameterizations. These forecasts used 3-km grid spacing over the contiguous United States and were subdivided into three sub-ensembles: a "P" sub-ensemble with physics diversity only, a "PI" sub-ensemble that included perturbations to initial and lateral boundary conditions in addition to physics diversity, and a "PSI" sub-ensemble that included stochastic physics perturbations in addition to IC/LBC and physics diversity. Precipitation, temperature, dewpoint, and wind forecasts from this ensemble and its sub-ensembles are evaluated using the operational HREF and GEFS as references for comparison.

For 6-h accumulated precipitation, the CAPS FV3-LAM forecasts exhibited 90th percentile values ranging from around 60% to 120% of the observed 90th percentile in Stage-IV rainfall, with more relative impact from model physics configuration than from the inclusion of initial and lateral boundary condition perturbations or stochastic perturbations Members using the NOAH-MP LSM exhibited the lowest 90th percentile rainfall values, while members using the NSSL microphysics scheme exhibited among the highest 90th percentile values. Similar patterns in the relative precipitation intensity of NOAH-MP and NSSL members were also noted in terms of domain-averaged total rainfall. At the 99th percentile threshold, almost all members had 99th percentile values slightly above that of Stage-IV observations, with relatively little variation among members. In general, many members exhibited underprediction at low rainfall rates and overprediction at relatively high rainfall rates. Difference in ETS for individual forecast members, as well as among members in the PI and PSI sub-ensembles, were generally small at both percentile thresholds. Sub-ensemble consensus, measured using the PM and LPM, showed good agreement with observed Stage-IV rainfall both in terms of percentile thresholds and domain-averaged total rainfall accumulation.

Overall, the greatest positive impact of including stochastic perturbations (which were present in the CAPS PSI sub-ensemble but not in the PI sub-ensemble) was the improved spread-error ratio noted in forecasts of 2-m temperature and dewpoint and 10-m zonal wind. Some minor improvement in spread-error ratio is also noted for temperature, dewpoint, and zonal wind verified against soundings, extending up to around 500 hPa on forecast days 1 and 2, with dewpoint exhibiting the greatest improvement. When the PI and PSI sub-ensembles were compared using other metrics, including AUC for 6-h precipitation forecasts, performance diagrams, rank histograms, and FSS for 24-h accumulated rainfall, and domain-averaged total rainfall, there was very little noticeable difference in the performance of the PI and PSI sub-ensembles. These sub-ensemble evaluations do, however, highlight other important aspects of CAPS forecast performance, such as the improved performance of the LPM mean over the PM mean in terms of FSS at higher precipitation thresholds, suggesting that the patchwise LPM algorithm being used is successful in its goal of retaining localized structures in the precipitation field. We note that some prior studies have found greater benefit when applying stochastic perturbations. For example, Jankov et al. (2019) found that including stochastic perturbations in a multi-physics CAM ensemble did improve the RMSE of 10-m wind forecasts, and Zhang (2019) reported a positive impact of stochastic perturbations for CAM forecasts of precipitation. In light of these prior studies, it is possible that, with additional tuning of the stochastic perturbation methods used, greater benefit might be obtainable from inclusion of stochastic perturbations.

Compared to the HREF and GEFS forecasts used as baseline comparisons for evaluation, the CAPS FV3-LAM forecasts typically performed slightly worse than (or at best no better than) HREF over the 0-48 hour forecast timeframe when HREF forecasts were available, though the CAPS FV3-LAM forecasts did consistently outperform GEFS throughout the 84 hour forecast period. The better performance of the CAPS FV3-LAM forecasts compared to GEFS is unsurprising given the much higher spatial resolution of the FV3-LAM forecasts.

As was noted in the introduction, one of the motivations for this study is to inform the design of future convection-allowing NWP ensembles, in particular the planned RRFS ensemble using FV3-LAM. Overall, the configurations of FV3-LAM examined in this study appear to be generally suitable for predicting spring season convective rainfall, outperforming GEFS across a variety of forecast quality metrics, and approaching the performance of HREF in some metrics. HREF has been tuned and optimized over many years of operational use, and is widely used for prediction of severe weather and high-impact precipitation events, so the ability of an experimental FV3-based ensemble with limited tuning to approach the performance of HREF is encouraging. We recommend that future studies continue investigation into the bias behavior of the NOAH-MP LSM; reducing the biases inherent in this scheme will help improve the performance of future operational ensembles. The use of stochastic perturbations enhanced ensemble spread, but further optimization of their configurations is needed for them to contribute more significantly towards improving the ensemble spread without increasing biases of the ensembles. Future investigations of stochastic physics perturbations could also examine the impacts of varying stochastic perturbation settings, or applying stochastic perturbation methods (such as SKEB, SPPT, and

SHUM) individually or in different combinations. Ideally, if stochastic perturbations alone are able to outperform a multi-physics ensemble, efforts can be focused on developing and maintaining a single well-performing physics suite. Finally, we note that while this study generally focused on features aggregated over 6 or 24 hours, future studies will examine FV3-LAM ensemble forecast performance in terms of more localized, shorter-duration features and severe weather hazards.

Acknowledgments

This work was primarily supported by the NOAA UFS R2O program under grant NA21OAR4320204, with additional support from NOAA Testbed grants NA19OAR4590141 and NA22OAR4590522. The CAPS FV3-LAM forecasts were run in real-time on Frontera at the Texas Advanced Computing Center (TACC) at the University of Texas-Austin, a National Science Foundation supported high performance computing center. The authors would like to thank the NOAA staff involved in organizing and leading the HWT SFE, including Adam Clark and Israel Jirak. The authors would also like to thank Philip Pegion for assistance with the configurations of stochastic perturbations. The authors also thank Dr. Curtis Alexander for facilitating the use of GSL-produced experimental RRFS analysis initial conditions.

Data Availability Statement.

The FV3-LAM and HREF model output evaluated in this study, has been archived by the Center for Analysis and Prediction of Storms (CAPS) and can be downloaded via the Open Science Framework at https://osf.io/u38yp/. GFS forecasts are available from UCAR's Research Data Archive (RDA) repository (https://rda.ucar.edu/datasets/ds084.1/), while GEFS forecasts can be downloaded from NOAA's registry of open data on AWS at https://registry.opendata.aws/noaa-gefs/. The Unrestricted Mesoscale Analysis (URMA) data provided by NOAA EMC which used for forecast evaluation can be accessed via a NOAA open data registry at https://registry.opendata.aws/noaa-rtma/. Stage-IV precipitation analyses can be downloaded from UCAR's Earth Observing Laboratory archive (https://data.eol.ucar.edu/dataset/21.093).

REFERENCES

Banos, I. H., W. D. Mayfield, G. Q. Ge, L. F. Sapucci, J. R. Carley, and L. Nance, 2022: Assessment of the data assimilation framework for the Rapid Refresh Forecast System v0.1 and impacts on forecasts of a convective storm case study. *Geoscientific Model Development*, **15**, 6891-6917, https://doi.org/10.5194/gmd-15-6891-2022.

Bengtsson, L., M. Steinheimer, P. Bechtold, and J.F. Geleyn, 2013: A stochastic parametrization for deep convection using cellular automata. *Quart. J. Roy. Meteor. Soc.*, **139(675)**, 1533-1543, doi:10.1002/qj.2108

Berner, J., G. Shutts, M. Leutbecher, and T. Palmer, 2009: A spectral stochastic kinetic energy backscatter scheme and its impact on flow- dependent predictability in the ECMWF ensemble prediction system. *J. Atmos. Sci.*, **66**, 603–626, doi:10.1175/2008JAS2677.1.

Berner, J., S. Y. Ha, J. P. Hacker, A. Fournier, and C. Snyder, 2011: Model Uncertainty in a Mesoscale Ensemble Prediction System: Stochastic versus Multiphysics Representations. Mon. Wea. Rev., 139, 1972-1995, 10.1175/2010mwr3595.1.

Black, T. L., Abeles, J. A., Blake, B. T., Jovic, D., Rogers, E., Zhang, X., et al., 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**, e2021MS002483, 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. *Bull. Amer. Meteor. Soc.*, **102**, E782–E807, https://doi.org/10.1175/BAMS-D-19-0093.1.

Buizza, R., M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887–2908, doi:10.1002/qj.49712556006.

Cafaro, C., and Coauthors, 2021: Do Convection-Permitting Ensembles Lead to More Skillful Short-Range Probabilistic Rainfall Forecasts over Tropical East Africa?. *Wea. Forecasting*, **36**, 697–716, https://doi.org/10.1175/WAF-D-20-0172.1.

Carley, J. R., M. E. Pyle, C. R. Alexander, and S. Weygandt, 2023: On the Development of NOAA's Rapid Refresh Forecast System. *Research Activities in Earth System Modelling. Working Group on Numerical Experimentation*. Report No. 53. WCRP Report

No.6/2023, 5.07-05.08. https://wgne.net/bluebook/uploads/2023/docs/05_Carley_Jacob_RapidRefreshModel.pdf

Carley, J. R., M. E. Pyle, C. R. Alexander, and S. Weygandt, 2024: Development and Status of NOAA’s Rapid Refresh Forecast System. *14th Conference on Transition of Research to Operations*. Baltimore, MD, 23 Jan. 2024, Amer. Meteor. Soc., 3B.1, https://ams.confex.com/ams/104ANNUAL/meetingapp.cgi/Paper/430385

Clark, A. J., W. A. Gallus, and T.-C. Chen, 2008: Contributions of Mixed Physics versus Perturbed Initial/Lateral Boundary Conditions to Ensemble-Based Precipitation Forecast Skill. Mon. Wea. Rev., 136, 2140-2156, 10.1175/2007mwr2029.1.

Clark, A. J., and Coauthors, 2011: Probabilistic Precipitation Forecast Skill as a Function of Ensemble Size and Spatial Scale in a Convection-Allowing Ensemble. *Mon. Wea. Rev.*, **139**, 1410–1418, https://doi.org/10.1175/2010MWR3624.1.

Clark, A. J., and Coauthors, 2020: A real-time, simulated forecasting experiment for advancing the prediction of hazardous convective weather. *Bull. Am. Meteorol. Soc.*, **101**, E2022–E2024, https://doi.org/10.1175/BAMS-D-19-0298.1.

De Pondeca, M. S. F. V., and Coauthors, 2011: The real-time mesoscale analysis at NOAA’s National Centers for Environmental Prediction: Current status and development. *Wea. Forecasting*, **26**, 593–612, https://doi.org/10.1175/WAF-D-10-05037.1.

DTC 2023: Developmental Testbed Center: Model Evaluation Tools (MET). Accessed 7 Nov. 2023. https://dtcenter.org/community-code/model-evaluation-tools-met.

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.

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. *J. Geophys. Res. Atmos.*, **108**, 1–16, https://doi.org/10.1029/2002jd003296.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. Mon. Wea. Rev., 129, 550, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

Hu, X.-M., Park, J., Supinie, T., Snook, N. A., Xue, M., Brewster, K. A., Brotzge, J., & Carley, J. R. (2023). Diagnosing Near-Surface Model Errors with Candidate Physics Parameterization Schemes for the Multiphysics Rapid Refresh Forecast System (RRFS) Ensemble during Winter over the Northeastern United States and Southern Great Plains, *Monthly Weather Review*, **151(1)**, 39-61, https://doi.org/10.1175/MWR-D-22-0085.1

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

Jankov, I., and et al., 2017: A performance comparison between multiphysics and stochastic approaches within a North American RAP ensemble. *Monthly Weather Review*, **145**, 1161–1179, doi:10.1175/MWR-D-16-0160.1

Jankov, I., J. Beck, J. Wolff, M. Harrold, J. B. Olson, T. Smirnova, C. Alexander, and J. Berner, 2019: Stochastically perturbed parameterizations in an HRRR-based ensemble. *Monthly Weather Review*, **147**, 153–173, doi:10.1175/MWR-D-18-0092.1

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

Johnson, M., M. Xue, and Y. Jung, 2023: Biases and skill of four two-moment bulk microphysics schemes in convection-allowing forecasts for the 2018 Hazardous Weather Testbed Spring Forecasting Experiment period. *Wea. Forecasting*., **38**, 1621–1642,https://doi.org/10.1175/WAF-D-22-0171.1.

Kalina, E. A., and et al. 2021: A Progress Report on the Development of the High-Resolution Rapid Refresh Ensemble. *Weather and Forecasting*, **36(3)**, 791-804. doi:10.1175/WAF-D-20-0098.1

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

Liu, Shun, M. Hu, T. Lei, X. Zhang, S. Yokota, B. Blake, E. Rogers, C. R. Martin, T. T. Ladwig, D. Dowell, C. Zhou, J. R. Carley, C. R. Alexander, and D. T. Kleist, 2023: Data

Assimilation Testing and Development Toward Implementation of the Rapid Refresh Forecast System. *27th Conference on Integrated Observing and Assimilation Systems for the Atmosphere, Oceans, and Land Surface (IOAS-AOLS),* Denver, CO, Amer. Meteor. Soc., 86, https://ams.confex.com/ams/103ANNUAL/meetingapp.cgi/Paper/414100.

Loken, E. D., J. C. Adam, X. Ming, and K. Fanyou, 2019: Spread and skill in mixed- and single-physics convection-allowing ensembles. Wea. Forecasting, 34, 305-330, https://doi-org.ezproxy.lib.ou.edu/10.1175/WAF-D-18-0078.1.

Mansell, E. R., and C. L. Ziegler, 2013: Aerosol effects on simulated storm electrification and precipitation in a two-moment bulk microphysics model. *J. Atmos. Sci.*, **70**, 2032–2050, https://doi.org/10.1175/JAS-D-12-0264.1.

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

Mason, I. B., 1982: A model for assessment of weather forecasts. Aust. Meteor. Mag., 30, 291-303, https://doi.org/10.1175/1520-0493(1979)107<0207:ORPFTY>2.0.CO;2.

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res. Atmos.*, **102**, 16663–16682, https://doi.org/10.1029/97JD00237.

Morris, M. T., J. R. Carley, E. Colon, A. Gibbs, M. S. F. V. De Pondeca, and S. Levine, 2020: A Quality Assessment of the Real-Time Mesoscale Analysis (RTMA) for Aviation. *Weather and Forecasting*, **35,** 977-996, https://doi.org/10.1175/Waf-D-19-0201.1.

Nakanishi, M., and H. Niino, 2009: Development of an Improved Turbulence Closure Model for the Atmospheric Boundary Layer. *J. Meteorol. Soc. Japan*, **87**, 895–912, https://doi.org/10.2151/jmsj.87.895.

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

Niu, G. Y., and Coauthors, 2011: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *J. Geophys. Res. Atmos.*, **116**, 1–19, https://doi.org/10.1029/2010JD015139.

NOAA, 2023: WPC Hydrometeorology Testbed. Accessed 6 Oct. 2023, https://www.wpc.ncep.noaa.gov/hmt/

Palmer, T. N., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. J. Shutts,M. Steinheimer, and A.Weisheimer, 2009: Stochastic parametrization and model uncertainty. *ECMWF Tech. Memo.*, **598**, 42 pp, doi:10.21957/ps8gbwbdv

Pondeca, M., S. Levine, J. Carley, Y. Lin, Y. Zhu, and J. Purser, 2015: Ongoing improvements to the NCEP Real Time Mesoscale Analysis (RTMA) and UnRestricted Mesoscale Analysis (URMA) and NCEP/EMC. NOAA, 2 pp., https://www.wcrp-climate.org/WGNE/BlueBook/2015/individual-articles/01_Pondeca_Manuel_etal_RTMA.pdf.

Roberts, B., B. T. Gallo, I. L. Jirak, A. J. Clark, D. C. Dowell, X. Wang, and Y. Wang, 2020: What Does a Convection-Allowing Ensemble of Opportunity Buy Us in Forecasting Thunderstorms? *Weather and Forecasting*, **35**, 2293-2316, https://doi.org/10.1175/waf-d-20-0069.1.

Roberts, B., I. L. Jirak, A. J. Clark, S. J. Weiss, and J. S. Kain, 2019: Post-Processing and Visualization Techniques for Convection-Allowing Ensembles. Bull. Amer. Meteor. Soc., 100, 1245–1258, https://doi.org/10.1175/BAMS-D-18-0041.1Roberts, B., B. T. Gallo, I. L. Jirak, A. J. Clark, D. C. Dowell, X. Wang, and Y. Wang, 2020: What does a convection-allowing ensemble of opportunity buy us in forecasting thunderstorms? *Weather and Forecasting*, **35**, 2293–2316, https://doi.org/10.1175/WAF-D-20-0069.1.

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**, 78-97.

Roebber, P. J., 2009: Visualizing Multiple Measures of Forecast Quality. *Weather and Forecasting*, **24,** 601-608, https://doi.org/10.1175/2008waf2222159.1.

Shin, H. H., and S.-Y. Hong, 2015: Representation of the subgrid-scale turbulent transport in convective boundary layers at gray-zone resolutions. *Monthly Weather Review*, 143, 250–271, https://doi.org/10.1175/MWR-D-14-00116.1.

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**, 1851–1865, https://doi.org/10.1175/MWR-D-15-0198.1.

Snook, N., F. Kong, A. Clark, B. Roberts, K. A. Brewster, and M. Xue, 2020: Comparison and Verification of Pointwise and Patchwise Localized Probability Matched Mean Algorithms for Convective Ensemble Forecasting. *Geophysical Research Letters*, **47**, https://doi.org/2020GL087839.

Snook, N., F. Kong, K. Brewster, M. Xue, B. Albright, S. Perfater, K. Thomas, and T. Supinie, 2019: Evaluation of Convection-Permitting Precipitation Forecast Products using WRF, NMMB, and FV3 for the 2016-2017 NOAA Hydrometeorology Testbed Flash Flood and Intense Rainfall Experiments. *Weather and Forecasting*, **34,** 781–804, https://doi.org/10.1175/WAF-D-18-0155.1.

Stensrud, D. J., J. W. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. Mon. Wea. Rev., 128, 2077-2107, Doi 10.1175/1520-0493(2000)128<2077:Uicamp>2.0.Co;2.

Supinie, T. A., Park, J., Snook, N., Hu, X.-M., Brewster, K. A., Xue, M., & Carley, J. R., 2022: Cool-Season Evaluation of FV3-LAM-Based CONUS-Scale Forecasts with Physics Configurations of Experimental RRFS Ensembles. *Monthly Weather Review*, **150(9)**, 2379-2398, https://doi.org/10.1175/MWR-D-21-0331.1

Schwartz, C. S., J. S. Kain, S. J. Weiss, M. Xue, D. R. Bright, F. Kong, K. W. Thomas, J. J. Levit, M. C. Coniglio, and M. S. Wandishin, 2010: Toward Improved Convection-Allowing Ensembles: Model Physics Sensitivities and Optimizing Probabilistic Guidance with Small Ensemble Membership. Wea. Forecasting, 25, 263-280, 10.1175/2009waf2222267.1.

Tompkins, A. M., and J. Berner, 2008: A stochastic convective approach to account for model uncertainty due to unresolved humidity variability. *J. Geophys. Res.*, **113**, D18101, doi:10.1029/2007JD009284

Thompson, G., and T. Eidhammer, 2014: A Study of Aerosol Impacts on Clouds and Precipitation Development in a Large Winter Cyclone. *J. Atmos. Sci.*, **71**, 3636–3658, https://doi.org/10.1175/JAS-D-13-0305.1.

Thompson, G., R. M. Rasmussen, and K. Manning, 2004: Explicit Forecasts of Winter Precipitation Using an Improved Bulk Microphysics Scheme. Part I: Description and Sensitivity Analysis. *Monthly Weather Review*, **132**, 519–542, https://doi.org/10.1175/1520-0493(2004)132<0519:EFOWPU>2.0.CO;2.

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**, 5095–5115, https://doi.org/10.1175/2008MWR2387.1.

Zhang, C., M. Xue, T. A. Supinie, F. Kong, N. Snook, K. W. Thomas, K. Brewster, Y. Jung, L. M. Harris, and S.-J. Lin, 2019: How Well Does the FV3 Model Predict Precipitation at a Convection-Allowing Resolution? Results from CAPS Forecasts for the 2018 NOAA Hazardous Weather Testbed with Different Physics Combinations. *Geophysical Research Letters*, **46**, 3523-3531.

Zhang, X., 2019: Multiscale Characteristics of Different-Source Perturbations and Their Interactions for Convection-Permitting Ensemble Forecasting during SCMREX. *Mon. Wea. Rev.*, **147**, 291–310, https://doi.org/10.1175/MWR-D-18-0218.1.

Zhou, L., S. Lin, J. Chen, L. M. Harris, X. Chen, and S. L. Rees, 2019: Toward Convective-Scale Prediction within the Next Generation Global Prediction System. *Bull. Amer. Meteor. Soc.*, **100**, 1225–1243, https://doi.org/10.1175/BAMS-D-17-0246.1.

Zhou, X., Y. Zhu, D. Hou, B. Fu, W. Li, H. Guan, E. Sinsky, W. Kolczynski, X. Xue, Y. Luo, J. Peng, B. Yang, V. Tallapragada, and P. Pegion, 2022: The Development of the NCEP Global Ensemble Forecast System Version 12. Wea. Forecasting, 37, 1069-1084, 10.1175/WAF-D-21-0112.1.

Zhu, K., M. Xue, B. Zhou, K. Zhao, Z. Sun, P. Fu, Y. Zheng, X. Zhang, and Q. Meng, 2018: Evaluation of real-time precipitation forecasts during 2013-2014 summer seasons over China at a convection-permitting resolution: spatial distribution, propagation, diurnal cycles and skill scores. J. Geophy. Res., 123, 1037-1064.