A Comparison of Methods to Sample Model Errors for Convection-Allowing Ensemble Forecasts in the Setting of Multiscale Initial Conditions Produced by the GSI-Based EnVar Assimilation System

NICHOLAS A. GASPERONI, XUGUANG WANG, AND YONGMING WANG

School of Meteorology, University of Oklahoma, Norman, Oklahoma

(Manuscript received 25 April 2019, in final form 20 December 2019)

ABSTRACT

A gridpoint statistical interpolation (GSI)-based hybrid ensemble–variational (EnVar) scheme was extended for convective scales—including radar reflectivity assimilation—and implemented in real-time spring forecasting experiments. This study compares methods to address model error during the forecast under the context of multiscale initial condition error sampling provided by the EnVar system. A total of 10 retrospective cases were used to explore the optimal design of convection-allowing ensemble forecasts. In addition to single-model single-physics (SMSP) configurations, ensemble forecast experiments compared multimodel (MM) and multiphysics (MP) approaches. Stochastic physics was also applied to MP for further comparison. Neighborhood-based verification of precipitation and composite reflectivity showed each of these model error techniques to be superior to SMSP configurations. Comparisons of MM and MP approaches had mixed findings. The MM approach had better overall skill in heavy-precipitation forecasts; however, MP ensembles had better skill for light (2.54 mm) precipitation and reduced ensemble mean error of other diagnostic fields, particularly near the surface. The MM experiment had the largest spread in precipitation, and for most hours in other fields; however, rank histograms and spaghetti contours showed significant clustering of the ensemble distribution. MP plus stochastic physics was able to significantly increase spread with time to be competitive with MM by the end of the forecast. The results generally suggest that an MM approach is best for early forecast lead times up to 6–12 h, while a combination of MP and stochastic physics approaches is preferred for forecasts beyond 6–12 h.

1. Introduction

Since the 1990s, much of the research in numerical weather prediction (NWP) has focused on ensemble forecasting techniques. The use of ensembles in NWP is attractive because they can account for uncertainties in the initial conditions (ICs) and errors within the numerical model. From global and large scales down to convection-allowing scales, ensembles have been demonstrated to be superior to deterministic forecasts, even for small ensembles (e.g., Richardson 2000; Palmer 2002; Clark et al. 2009; Vié et al. 2011; Schwartz et al. 2014; Loken et al. 2017; Schwartz et al. 2017). A vital aspect in the application of ensemble techniques is the proper design of said ensembles. For global and large-scale systems, there is a large body of research that has studied optimal methods for ensemble design (e.g., Toth and Kalnay 1993; Molteni et al. 1996; Toth and Kalnay 1997; Wang and Bishop 2003; Wang et al. 2004; Eckel and Mass 2005; Candille 2009).

There is a smaller but growing body of research over the past decade to increase knowledge and understanding of best design practices at convection-allowing scales (e.g., Clark et al. 2010; Schwartz et al. 2010; Xue et al. 2010; Johnson and Wang 2012, 2017; Duda et al. 2014, 2016, 2017; Romine et al. 2014; Johnson et al. 2017). One of the avenues facilitating this research progress is the yearly NOAA Hazardous Weather Testbed (HWT) Spring Forecasting Experiments (SFEs; Kain et al. 2003; Clark et al. 2012, 2018). Through a collaborative effort between several government and university-based research agencies, the number of ensembles and ensemble designs implemented during the SFE have significantly increased over the last decade (Clark et al. 2018).

Although positive strides have been made in convection-allowing ensemble (CAE) forecasting, ensembles remain largely underdispersive, meaning the ensemble spread does not match the actual forecast uncertainty. Ensembles with

Corresponding author: Nicholas A. Gasperoni, ngaspero@ou.ed

DOI: 10.1175/MWR-D-19-0124.1

inferior spread manifest in overconfident and unreliable ensemble forecasts. Furthermore, studies have shown that despite improvements in skill from addressing IC and lateral boundary condition (LBC) uncertainties in ensemble design, CAEs remain largely underdispersive (e.g., Vié et al. 2011; Romine et al. 2014; Schwartz et al. 2014). As a result, increased focus has been put on methods to address model error within CAEs (e.g., Johnson and Wang 2012, 2017; Duda et al. 2014, 2016, 2017; Romine et al. 2014; Loken et al. 2019).

Schemes to sample model errors can be broadly categorized into four methods: 1) multimodel, 2) multiparameter, 3) multiphysics, and 4) stochastic perturbation schemes. The first multimodel, describes forecast ensembles comprised of members from two or more model dynamic cores (e.g., Ebert 2001; Wandishin et al. 2001; Eckel and Mass 2005; Candille 2009; Johnson and Wang 2012; Melhauser et al. 2017). The second, multiparameter, is a method of addressing model error from unresolved subgrid-scale physical processes by perturbing parameters of one or more physical parameterization schemes. (e.g., Gebhardt et al. 2011; Hacker et al. 2011; Yussouf and Stensrud 2012; Duda et al. 2014, 2017) The third method, multiphysics, addresses physical parameterization uncertainty by varying the physics schemes themselves within the ensemble. (e.g., Stensrud et al. 2000; Gallus and Bresch 2006; Schwartz et al. 2010; Duda et al. 2014; Johnson and Wang 2017; Loken et al. 2019). The fourth approach, stochastic physics, are schemes that represent uncertainty of subgrid-scale processes by stochastically perturbing parameterization schemes or their effects on physics and dynamical forecast tendencies during integration (e.g., Berner et al. 2011; Romine et al. 2014; Berner et al. 2015; Duda et al. 2016).

Although each of the aforementioned model error methods have been tested for CAEs, the optimal application of methods–including combinations of techniques–for CAEs still remains unknown. There are only a limited number of recent studies that have directly compared two or more methods to account for model error at convection-allowing scales (e.g., Duda et al. 2014, 2016, 2017; Melhauser et al. 2017; Jankov et al. 2019). Furthermore, the ensembles of these previous studies do not fully account for IC uncertainty at the resolution necessary for CAE forecasting; for instance, they do not account for IC uncertainty at all (e.g., initializing the ensemble from a single analysis), or they are initialized from an ensemble analysis at coarser resolution that has been downscaled or interpolated to higher CAE resolution prior to the forecast. As noted in Schwartz et al. (2019), this inconsistency of analysis model and/or resolution to the CAE forecast configuration may lead to undesirable error growth characteristics due to potential discrepancies in model biases, physics, dynamics, and scale representations. These inconsistencies imply that ensemble IC perturbations do not fully sample IC errors across all scales and variables necessary for the convection-allowing forecast model. Consequently, the use of model error techniques for CAEs with incomplete or improper IC error sampling may lead to overcompensation of error; that is, the model error techniques falsely compensate for unsampled IC errors in addition to model error. Additionally, results of a hierarchical clustering analysis of 2009 HWT SFE ensembles by Johnson et al. (2011) implied that optimal ensemble design may depend on a user's needs for a given ensemble forecast. For example, for short-term forecasts (<6 h), emphasis should be placed on properly addressing IC and microphysics uncertainty within the ensemble design; on the other hand, for next-day convection forecasts, increasing emphasis should be placed on planetary boundary layer (PBL) uncertainty.

With recent progress in realistic ensemble-based radar data assimilation (DA) and forecasting of a variety of convective system case studies (e.g., Yussouf et al. 2013, 2015; Johnson et al. 2015; Snook et al. 2015; Wang and Wang 2017; Degelia et al. 2018), it is evident that state-of-the-art ensemble-based DA can now better sample small-scale IC uncertainties. This provides a meaningful foundation to better address the question of which model error sampling techniques are preferred for CAEs. To the best of the authors' knowledge, only Johnson and Wang (2017) have examined potential benefits of model error methods—in their case, two multiphysics configurations—under the context of fully sampled IC errors by a CAE DA system. They designed a fully tuned 3-km CAE DA and forecast system for the prediction of nocturnal mesoscale convective systems (MCSs) over the Great Plains region. They found positive impacts of mixed PBL and microphysics to sample model errors in addition to the IC sampling that the convection-allowing DA provided. Specifically, improvements were found in location error, storm structure of initiating convection, and the number of members that predicted observed nocturnal convection initiation (CI).

Here, we examine impacts from different model error schemes on the prediction of convective systems over the full contiguous United States (CONUS) using ensemble ICs that already sample IC errors at convection-allowing resolution. These ICs are provided by CAE DA analyses, specifically the gridpoint statistical interpolation (GSI)-based ensemble–variational (EnVar) system modified for convective-scale assimilation of radar data by Johnson et al. (2015) and Wang and Wang (2017). This EnVar system was implemented as part of the

Community Leveraged Unified Ensemble (CLUE; Clark et al. 2018) for the 2017 and 2018 HWT SFEs (Clark et al. 2017; Gallo et al. 2018). The scope of this model error sampling study is broader than in Johnson and Wang (2017): in addition to larger CONUS domain, CAE forecasts cover a variety of retrospective cases with differing convective modes (including diurnal and nocturnal convection), initiating mechanisms, forecast evolutions, and synoptic forcing conditions. Additionally, multiple techniques to address model error are compared and contrasted.

In this study, two single-model single-physics (SMSP) forecast ensembles were initialized following their own EnVar cycling. Members from each SMSP forecast ensemble were combined to form a multimodel forecast CAE containing full-scale IC error sampling from CAE DA. Additionally, one of the cores—the Weather Research and Forecasting (WRF) Model—has many diverse physics options that can be combined to form a multiphysics ensemble. Finally, a stochastic physics scheme, specifically the Stochastic Energy Backscatter (SKEB) scheme of Berner et al. (2011), was implemented in combination with the multiphysics configuration for further study. The SKEB scheme adds random stochastic perturbations to dynamical tendencies of streamfunction and potential temperature at each forecast time step (Berner et al. 2011). This combination of SKEB with multiphysics was motivated by recent studies that have shown that combining SKEB with mixed physics or other stochastic physics schemes leads to an overall better performance in the ensemble forecast system, indicating that addressing model error may be more complex than what can be represented by one method alone (e.g., Berner et al. 2015; Duda et al. 2016; Jankov et al. 2017, 2019).

The remainder of the paper is organized as follows: section 2 describes the GSI-based EnVar system in further detail. In section 3, the 10 retrospective case studies and the DA cycling setup is described, along with descriptions of each of the forecast ensemble experiments and verification methods. Results in section 4 are split into subsections detailing precipitation and reflectivity verification, upper-level verification of other fields, surface variable verification, and further diagnosis of ensemble spread characteristics for each experiment. The results are then summarized with further discussion in section 5.

2. Data assimilation configuration

a. *Description of convective-scale GSI-based EnVar system*

The GSI-based EnVar system was operationally implemented for the Global Forecast System (GFS), showing improvements to global and tropical cyclone forecasting (Hamill et al. 2011; Wang et al. 2013; Wang and Lei 2014). The hybrid approach of EnVar leverages benefits of both ensemble-based and variational DA frameworks (e.g., summarized in Wang 2010; Wang et al. 2013). The inclusion of ensemble covariances provides more accurate, flow-dependent error covariances compared to predefined static error covariance matrices in a variational scheme. Additionally, important physical and dynamical constraints can be directly included within the variational analysis. This hybrid system was further interfaced with regional models including the Weather Research and Forecasting (WRF) Advanced Research (ARW) core used by the Rapid Refresh (RAP) system, and the Nonhydrostatic Multiscale Model on the B grid (NMMB) used by the NAM. The GSI-based ensemble Kalman filter (EnKF) system was modified by Johnson et al. (2015) to incorporate multiscale assimilation, including direct assimilation of convective-scale radar reflectivity and radial wind observations to update the full model state vector with WRF including rain, snow, and graupel hydrometeor mixing ratios. Wang and Wang (2017) further incorporated these capabilities to the GSI-based EnVar system. This GSI-based EnVar system including these convective-scale DA capabilities was implemented with High-Resolution Rapid Refresh (HRRR) and North American Mesoscale Forecast System (NAM)-like convection-allowing prediction, in terms of dynamical core, physics schemes, and similar CONUS-wide domain (Duda et al. 2019; Wang et al. 2020, manuscript submitted to *Atmopshere*). They were also demonstrated during the 2017 and 2018 HWT SFEs by the University of Oklahoma (OU) Multiscale data Assimilation and Predictability (MAP) laboratory (Johnson and Wang 2020, manuscript submitted to *Mon. Wea. Rev.*; Clark et al. 2017; Gallo et al. 2018; Potvin et al. 2019).

This study employs the same convective-scale EnVar system and configurations used for the 2017–19 HWT SFEs, with a similar configuration documented in Wang et al. 2020, manuscript submitted to *Atmopshere*). A DA ensemble of 41 members was initialized using a combination of the GFS control with 20 perturbations each from the Short-Range Ensemble Forecast (SREF) system and the Global Ensemble Forecast System (GEFS). Operational observations from the RAP data stream were assimilated hourly from 1800–0000 UTC each day, with radar reflectivity DA every 20 min for the final hour (2300–0000UTC). Observations from RAP included conventional surface and mesonet observations, flight-level aircraft observations, wind profilers, and radiosondes. The GSI performed additional quality control (QC) procedures, including observation error inflation for questionable observations and flagging observations with large

gross errors. Assumed observation errors and gross check thresholds were provided by a static table within the GSI package (see Hu et al. 2018 for more information on GSI QC procedures). Radar reflectivity observations were obtained from the Multi-Radar Multi-Sensor (MRMS; Smith et al. 2016). As with RAP data, the GSI performed additional QC gross error checks on the reflectivity observations, assuming an observation error of 5 dBZ. The Gaspari and Cohn (1999) function was applied for localization with horizontal cutoff distances of 15 and 300 km for radar reflectivity and all other observations, respectively, and vertical cutoffs in terms of natural log pressure of 1.1 and 0.55, respectively. For covariance inflation, the posterior ensemble spread was relaxed to match 95% of the prior spread via the relaxation-to-prior-spread (RTPS; Whitaker and Hamill 2012) method. These DA parameters were chosen based on results of sensitivity tests documented in Wang et al. 2020, manuscript submitted to *Atmopshere*). The control member was updated by the EnVar (Wang and Wang 2017); the remaining 40 members were updated by the EnKF and recentered around the control analysis.

The model grids for DA cycling with NMMB and ARW had 3-km horizontal grid spacing and covered the entire CONUS; both grids were similar in location and size (Fig. 1) and included a stretched vertical grid with 50 levels. A single physics configuration was applied for each model core during cycling, identical to configurations used during the 2017–19 HWT SFEs. For NMMB, the physics parameterizations included Ferrier–Aligo microphysics (Aligo et al. 2018), Mellor–Yamada–Janjić (MYJ) boundary layer physics (Janjić 1994), and the unified Noah land surface model scheme (Tewari et al. 2004). These physics settings are identical to the operational NAM Rapid Refresh (NAMRR) configuration. For WRF-ARW, the physics schemes included aerosol-aware Thompson microphysics (Thompson et al. 2008; Thompson and Eidhammer 2014), Mellor–Yamada–Nakanishi–Niino (MYNN) level 2.5 boundary layer physics (Nakanishi and Niino 2009), and the RUC land surface scheme (Benjamin et al. 2004). This physics suite is identical to the operational HRRR configuration.

b. *Retrospective case studies and DA cycling configuration*

Ten retrospective case studies were chosen to facilitate the testing of model error techniques in ensemble forecast design (Table 1). The cases included many examples of both discrete isolated storms (e.g., 22 May 2016 case) and organized MCSs (e.g., long-lived squall lines in the 16 May and 25 May 2015 cases). Diverse synoptic forcing and organizing mechanisms were also included in these cases, such as strong upper-level troughs,

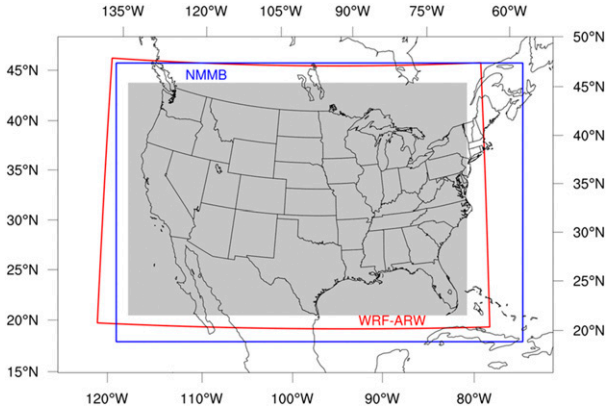

FIG. 1. Model domains used in this study. The NMMB grid (blue) was used during the 2017 HWT SFE with 1680 × 1152 grid points at 3 km horizontal grid spacing. The WRF-ARW grid (red) was used during the 2018 HWT SFE with 1620 × 1120 grid points at 3 km horizontal resolution. The gray shaded box indicates the domain used for verification.

surface cold fronts, and slow moving or stationary frontal zones with multiple storm clusters and upscale growth.

The DA cycling setup is shown in Fig. 2. This cycling is equivalent to that used during the 2017 and 2018 HWT SFEs (discussed in section 2a), except for each retrospective case different 6-h time windows are chosen for cycling (Table 1). These times were chosen based on various features of interest for each case (e.g., near the start of an MCS event) in order to have a sample that reflects important features to be captured by the EnVar convective-scale radar DA. For each retrospective case, two DA cyclings were performed: one using the NMMB dynamical core configuration and one using the WRF-ARW dynamical core configuration. In other words, there were two full EnVar analyses available at each case's final analysis time, which were then used to initialize 18-h free forecast experiments for different model error techniques. The settings for these free forecast experiments are discussed in the next section.

3. Experiment setup

a. *Experiment description*

Five experiments were conducted for this study, as shown in Table 2. For each experiment, a 10-member subset ensemble was initialized using the EnVar control analysis and first 9 recentered EnKF perturbations. The 10 members were chosen based on practical recommendations of previous studies that found diminishing returns for ensemble forecast sizes above 10, which may not justify the significant added computational cost of running more than 10-member forecasts. (e.g., Clark et al. 2011; Schwartz et al. 2014), The first two

TABLE 1. Retrospective cases used for this study, including synoptic forcing and dominant convective features. Synoptic forcing for each case was determined subjectively.

Case date	Final analysis time	Synoptic forcing	Dominant features
16 May 2015	2300 UTC	Strong	Dryline convection, upscale growth to long-lived squall line, TX to MO
25 May 2015	1300 UTC	Strong	Bowing, long-lived squall line, southeast TX
26 Jun 2015	0400 UTC	Weak	Nocturnal MCSs
14 Jul 2015	1900 UTC	Strong	Cold front advancing southward across MS and OH valley, squall line and numerous high wind reports
11 Sep 2015	0100 UTC	Moderate	Numerous severe hail-producing supercells across KS, upscale growth to MCS with advancing cold front
22 May 2016	2300 UTC	Moderate	Strong discrete isolated convection ahead of dryline (west TX) and weak quasi-stationary fronts (north plains)
17 Jun 2016	2000 UTC	Weak	Bowing squall line in southeast United States; two MCSs in north and central plains, development along outflow boundaries
6 Jul 2016	0100 UTC	Weak	Convective clusters in KS and NE with modest upscale growth; severe squall line in southern MN
7 Jul 2016	0000 UTC	Weak	Long-lived bowing MCS originating from supercellular convection
10 Jul 2016	0400 UTC	Weak	Nocturnal MCS across Dakotas with MCV

experiments, NMMB and ARW-SP, were SMSP configurations that extended the same respective model configurations from DA through the 18-h free forecast. The third experiment, MM, is a multimodel configuration that for each case randomly combined five members each from the ARW-SP and NMMB experiments' ensemble forecasts into a 10-member forecast ensemble. Experiment ARW-MP used various combinations of microphysics, PBL, and land surface model (LSM) schemes available in the WRF-ARW, equivalent to mixed-physics combinations used by other groups during the 2018 HWT SFE (see Table 1 of Gallo et al. 2018). In total there were four microphysics schemes, three PBL schemes, and two LSM schemes in various combinations for the ARW-MP experiment.[1]

An additional experiment, ARW-MPSKEB, applied the SKEB scheme (Berner et al. 2011) during forecast integration in combination with the multiphysics of ARW-MP. Stochastic schemes, such as SKEB, function to represent the variability of subgrid-scale processes during the forecast integration by adding random stochastic noise perturbations to tendency terms. In the case of SKEB, these random perturbations are added to the streamfunction and potential temperature dynamical tendency terms at each forecast time step. Duda et al. (2016) showed for a CAE that the combination of SKEB with a mixed physics ensemble leads to slightly improved Brier scores in 1-h precipitation and significant spread increases of several variables when compared to ensembles applying just one of those methods. Here we applied the SKEB scheme in WRF-ARW following the parameter tuning of Duda et al. (2016) for experiment ARW-MPSKEB. Because of significant changes to SKEB within WRF and the shorter 18-h free forecast duration examined in this study, some additional tuning of the SKEB parameters was performed (not shown). The final parameters used in ARW-MPSKEB are shown in Table 3.

b. *Verification methods*

Neighborhood-based ensemble probability (NEP) was used to perform objective verifications in terms of 1-h accumulated precipitation and composite reflectivity fields:

$$\mathrm{NEP}_{ij} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{NP}_{k,ij}. \tag{1}$$

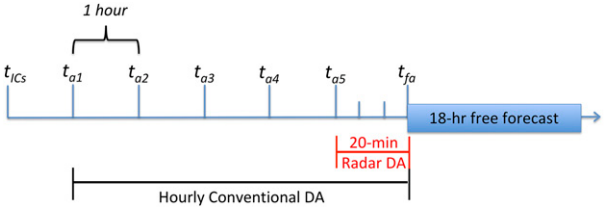

FIG. 2. Cycling setup for DA. The ensemble was initialized at t_{ICs}, 6 h prior to the time of final analysis, t_{fa}, for each retrospective case (see Table 1). EnVar analyses using conventional observations from RAP were produced hourly ($t_{a1,\ldots,}$ t_{a5}; t_{fa}), with 20-min radar reflectivity DA cycling for the last hour. Note that analyses at t_{a5} and t_{fa} included assimilation of both conventional and radar reflectivity observations.

[1] The microphysics schemes include aerosol-aware Thompson, National Severe Storms Laboratory (NSSL) bulk two-moment (Mansell et al. 2010), Morrison two-moment (Morrison et al. 2009), and the newly implemented P3 scheme (Morrison and Milbrandt 2015). The PBL schemes include MYJ, MYNN, and the Yonsei University scheme (YSU; Hong et al. 2006). The LSM schemes include Noah (with four soil levels) and RUC (with nine soil levels).

TABLE 2. Experiment names and configurations.

Experiment name	Model core	Member number	Microphysics	PBL	LSM
NMMB	NMMB	0–9	Ferrier–Aligo	MYJ	Noah
ARW-SP (single physics)	ARW	0–9	Thompson	MYNN	RUC
MM (multimodel)	NMMB + ARW	Members 0–9 randomly split and taken from NMMB and ARW-SP experiments			
ARW-MP (multiphysics)	ARW	0 (control)	Thompson	MYNN	RUC
		1	Thompson	MYJ	Noah
		2	NSSL	YSU	Noah
		3	NSSL	MYNN	Noah
		4	Morrison	MYJ	Noah
		5	P3	YSU	Noah
		6	NSSL	MYJ	Noah
		7	Morrison	YSU	Noah
		8	P3	MYNN	Noah
		9	Thompson	MYNN	Noah
ARW-MPSKEB	ARW	As in ARW-MP, but including application of SKEB during forecast (see Table 3)			

NEP at grid point *ij* in the verification is the *K*-member ensemble average of neighborhood probability (NP)–the percentage of grid points that exceed a chosen threshold within a prescribed averaging radius. Additionally, grid point-based ensemble mean error, spread, and bias was calculated for evaluation of surface and upper-air variables. All WRF-ARW forecast and observation verification fields were bilinearly interpolated to the NMMB grid prior to computing verification scores. The domain used for verification was the gray-shaded box in Fig. 1, where both WRF-ARW and NMMB grids overlap.

Neighborhood-based metrics included fractions skill score (FSS), relative operating characteristics (ROC) area under the curve, and reliability diagrams. Each of these metrics is defined as they are encountered in the results. The chosen neighborhood radius for each metric was 48 km (16 grid points), similar to the size used in several CAE studies (e.g., Johnson and Wang 2012; Duda et al. 2014, 2016; Romine et al. 2014). Tuning tests of skill versus a range of neighborhood radius size confirmed 48 km to be a good compromise between higher skill with increasing radius and the loss of detail due to increased smoothing of larger radii (not shown). These metrics were applied to 1-h accumulated precipitation fields at 2.54, 6.35, and 12.7 mm thresholds (0.1, 0.25, and 0.5 in., respectively), as well as 30, 40, and 50 dBZ thresholds in composite reflectivity fields. Verification fields were obtained from the MRMS, which in addition to radar mosaics includes a suite of quantitative precipitation estimation (QPE) products (Zhang et al. 2016). Here the local gauge bias-corrected radar-based 1-h QPE mosaic field was used for precipitation verification, which has been similarly used in other CAE studies (e.g., Duda et al. 2016; Johnson et al. 2017; Jankov et al. 2019).

Hourly RAP analyses were used as verification fields for gridpoint evaluation of upper-air variables. These variables included temperature, wind, geopotential height, and dewpoint temperature. Additionally, surface variables 2-m temperature, 2-m dewpoint, and 10-m winds were evaluated against hourly 2.5-km Real-Time Mesoscale Analysis (RTMA; De Pondeca et al. 2011) fields. This allowed verification of higher-resolution details that the 3-km CAE forecast can resolve, since the RTMA is of comparable resolution. Further, the RTMA has important downscaling procedures for the background field plus rigorous QC methods for mesoscale surface observations to ensure accurate and physically consistent 2.5-km analyses.

Statistical significance tests were performed on differences of verification metrics (FSS, ROC, RMSE, spread) among the experiments using a one-sided paired permutation-resampling test with daily contingency tables from each case as separate samples to ensure independence (Hamill 1999). Significance is noted at 90%

TABLE 3. SKEB scheme parameter values chosen for experiment ARW-MPSKEB.

SKEB parameter	Value chosen
Backscatter rate for streamfunction	1×10^{-5}
Backscatter rate for potential temperature	5×10^{-6}
Decorrelation time (both variables)	3240 s
Spectral slope (both variables)	−2.567
Random number seed	$2 + k$ (for given member k from 0–9)

and 99% confidence levels in the figures, with markers color coded by the color of the experiment, which is significantly "better" depending on what verification metric is used (e.g., higher values for both FSS and ROC scores)

4. Results

a. *Verification of QPE and composite reflectivity fields*

1) EVALUATION OF SINGLE-MODEL SINGLE-PHYSICS AND MULTIMODEL ENSEMBLES

Objective verification of 1-h QPE accuracy was performed using the FSS (e.g., Schwartz et al. 2010), the skill score extension of the neighborhood-based fractions Brier score (FBS):

$$\mathrm{FSS} = 1 - \frac{\mathrm{FBS}}{\mathrm{FBS}_{\mathrm{worst}}} = 1 - \frac{\frac{1}{N_v}\sum_{ij=1}^{N_v}(\mathrm{NEP}_{F(ij)} - \mathrm{NP}_{O(ij)})^2}{\frac{1}{N_v}\sum_{ij=1}^{N_v}(\mathrm{NEP}^2_{F(ij)} + \mathrm{NP}^2_{O(ij)})}, \tag{2}$$

where N_v is the number of grid points in the verification domain, $\mathrm{NEP}_{F(ij)}$ is the neighborhood ensemble forecast probability at grid point ij, and $\mathrm{NP}_{O(ij)}$ is the observed neighborhood probability at grid point ij–the percentage of grid points within a predefined radius of ij where observations exceed a given threshold. FBS is the domainwide mean-squared-differences of neighborhood probabilities, and $\mathrm{FBS}_{\mathrm{worst}}$ represents the maximum value of FBS assuming no overlap in nonzero probability grid points. FSS is positively oriented, ranging from 0 (no skill forecast) to 1 (perfect forecast).

The FSS of 1-h QPE is shown in the left column of Fig. 3, calculated using contingency table components aggregated (summed) over all 10 retrospective cases. Comparing SMSP experiments, the FSS of NMMB and ARW-SP show similar patterns at all three precipitation thresholds, where the ARW-SP has higher skill than NMMB at early forecast lead times (statistically significant from hours 1–12 at the light 2.54 mm threshold) followed by the NMMB having higher skill at later lead times, particularly for the heavier thresholds (statistically significant for 3–4 of the final 5 h at heavier thresholds). The MM experiment shows FSS almost universally higher than both SMSP experiments, with significantly higher scores compared to the lowest performing single model at all forecast times. MM also has significantly higher scores than the best SMSP experiment for 3, 5, and 3 out of 18 forecast hours for the 2.54, 6.35, and 12.7 mm thresholds, respectively. Although the higher skill scores are not always statistically significant compared to the best single model experiment at each time, the best single model experiment changes depending on which forecast hour and threshold is in question. Thus, MM offers consistency in higher FSS at all times whereas each SMSP experiment has time periods of worse scores.

The right column of Fig. 3 shows the ROC (Mason 1982; Mason and Graham 2002) area under the curve for composite reflectivity NEP at 30, 40, and 50-dBZ thresholds. ROC area measures the skill of a probabilistic forecast system in discriminating between observed events and nonevents over a series of probabilistic decision thresholds. A ROC area of 1.0 is considered perfect discrimination, and 0.5 and below is no better than a random forecast and considered unskillful. Statistically significant difference markers of Figs. 3b,3d, and 3f show the significantly higher experiment. ARW-SP has higher ROC area than NMMB at each threshold and forecast time, significant for 12, 15, and 10 of the 18 h for low to high thresholds, respectively (though for the 50-dBZ threshold all skill is lost after approximately 12 h). At 30 dBZ, MM is higher than NMMB for all hours (statistically significant for hours 1–15); however, MM is only slightly higher (~0.01–0.02), than ARW-SP with just three statistically significant times (Fig. 3b). At 40 and 50 dBZ, MM is again consistently, significantly higher than NMMB, but approximately equivalent or lower than ARW-SP, with 5 h significantly lower at 50 dBZ (Figs. 3d,f). So in terms of ROC area where one single-model (ARW-SP) is clearly superior to the other (NMMB), the comparative performance of MM is more mixed.

A neighborhood reliability diagram (Fig. 4) plots observed frequency against various NEP forecast bins and displays forecast reliability (proximity to diagonal), resolution (distance from horizontal climatology base rate line), and an accompanying sharpness plot showing forecast counts within each probability bin. Note that ARW-MP from Fig. 4 will be discussed in the next section. At the lighter 2.54 mm threshold (Fig. 4a), both SMSP experiments have a general overforecasting bias, meaning the NEP is overestimating the probability of observed events (falls below the diagonal). Conversely, MM has an underforecasting bias from 20%–70% NEP; however, at higher probabilities above 80%, MM is more reliable than NMMB and ARW-SP.

At the higher 12.7 mm threshold, forecast sharpness is not sufficient to construct a full reliability diagram given extreme events have smaller sample sizes. In this case, Schwartz and Sobash (2017) recommends the use of neighborhood maximum ensemble probability (NMEP) for extreme events:

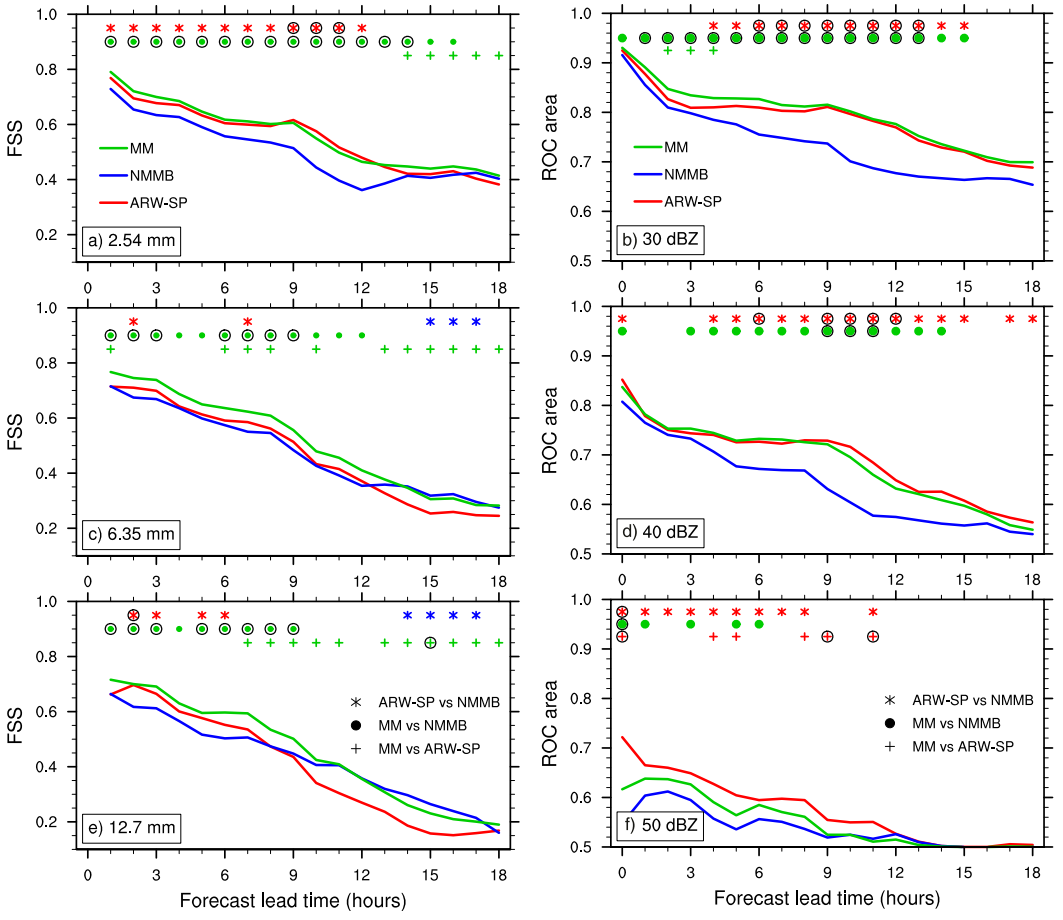

FIG. 3. (left) FSS for NEP of 1-h accumulated precipitation at (a) 2.54, (c) 6.35, and (e) 12.7 mm thresholds. (right) ROC area under the curve for NEP of composite reflectivity at (b) 30, (d) 40, and (f) 50 dBZ thresholds. A 48-km radius was used to calculate neighborhood-based fields. Markers across the top of each panel indicate statistically significant differences (90% confidence) for the three difference tests indicated in (e),(f), color coded by the significantly higher experiment in FSS or ROC area. For instance, a red asterisk on an FSS panel indicates that experiment ARW-SP was significantly higher than experiment NMMB at a given time. Markers outlined with a black circle indicate statistically significant differences at the 99% confidence level.

$$\mathrm{NMEP}_{ij} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{BNP}_{k,ij}. \tag{3}$$

NMEP is the K-member ensemble average of *binary* neighborhood probability (BNP) at grid point ij, where if an event occurs at least once within a search radius r of the grid point, the probability at the central grid point is equal to 1 for that ensemble member (0 if no event within r occurs). This is different from NEP, which is an ensemble average of *fractional* NPs. In other words, with NMEP the neighborhood radius is a *search* radius, checking whether an event occurs within the neighborhood; in contrast, with NEP the neighborhood radius is an *averaging* radius, where the fraction of events within the neighborhood is recorded. For a more in depth explanation see Schwartz and Sobash (2017). To maintain consistency, the search radius r used here for NMEP calculations is identical to the averaging radius chosen for NEP (48 km). Additionally, verification observations for NMEP-based plots follow the construction of NMEP (i.e., observations are converted to a field of BNPs for each event threshold).

As noted in Schwartz and Sobash (2017), NMEP-based reliability plots tend to show an overconfidence bias, and that is true here as well in Fig. 4b.[2] Comparing SMSP experiments, NMMB is more reliable at 0% and

[2] Note the base rate is *higher* than in Fig. 4a because, despite the more extreme event threshold, NMEP uses *binary* NPs while NEP uses *fractional* NPs, the latter which will always remain less than or equal to BNPs at all grid points for a given threshold (assuming the same neighborhood size).

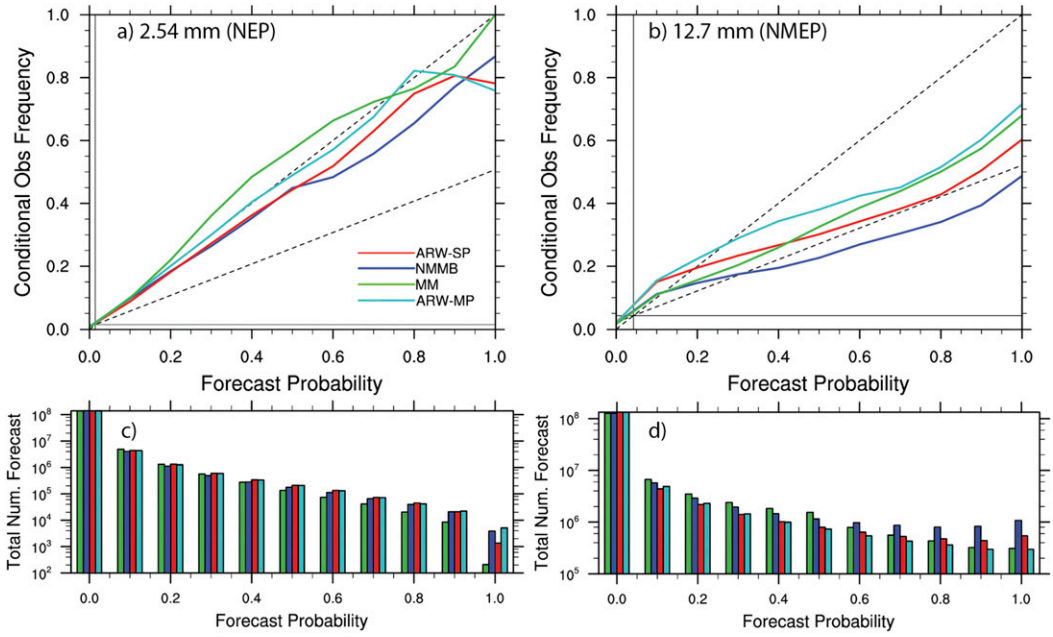

FIG. 4. Neighborhood-based (a),(b) reliability and (c),(d) sharpness diagrams for experiments ARW-SP, NMMB, ARW-MP, and MM, aggregated over forecast hours 7–18 and 10 cases for 1-h accumulated precipitation at indicated thresholds. In (a) and (b), the diagonal dashed ($y = x$) line indicates perfect reliability, while the dashed line below it is the no skill line. Solid vertical and horizontal lines represent sample base rate climatology. Note that neighborhood maximum ensemble probability is used for the 12.7 mm threshold plots (right column; see text for further details).

10% than ARW-SP; however, ARW-SP is more reliable at higher probabilities while NMMB falls below the no skill line. Experiment MM has the best reliability, compared to SMSP experiments, at higher NMEP between 50% and 100%. There is an accompanying reduction in sharpness at both QPE thresholds for MM (Figs. 4c,d).

Subjective evaluation of fields of NEP $\geq$ 30 dBZ showed two main benefits to MM over SMSP experiments, as demonstrated in Fig. 5. The first benefit of MM occurs where NMMB and ARW-SP have similar magnitude of errors, but in opposite directions or differing locations within the domain. For example, the 16 May 2015 case featured a widespread outbreak of severe weather from Texas through Oklahoma and into Missouri, with upscale growth of convection into a line as it progressed eastward. The NMMB (Fig. 5a) and ARW-SP (Fig. 5b) show mixed simulations of the evolution of this convection; the NMMB simulates the southernmost portion of convection in Texas well, but is missing convection in Missouri and has a slight eastward bias in between, while the ARW-SP has much stronger probabilities through Arkansas and Missouri but has a slow propagation bias and misses the southern and eastern extent of observations. MM (Fig. 5c), on the other hand, has probability coverage of the full length and width of the line from Missouri down into Texas, as the combined effects of the two single model simulations cancel out the relative location errors of each model.

The second benefit to MM is the resistance to forecast skill drop-offs that are present in the NMMB and ARW-SP experiments. This is illustrated with two times from the 14 July 2015 case (Figs. 5d–i). Early in the forecast, the ARW-SP had much lower skill than the NMMB because it missed the bulk of convection that was occurring in the southeast (Fig. 5e), whereas the NMMB had good coverage (Fig. 5d). But 9 h later, the skill of ARW-SP increased substantially while the NMMB dropped out. This can be seen by the MCS that developed over the Great Plains–the NMMB has just a small area of storms with weak probabilities too far south (Fig. 5g), while the ARW-SP model has stronger probabilities in the correct locations (Fig. 5h). The MM at each of these times (Figs. 5f,i) has probability coverage at all of these locations. While the MM is not as skillful as the best single model at each time due to reductions in probabilities, the MM is skillful at both times whereas the NMMB and ARW-SP were only skillful at one time each.

2) EVALUATION OF MULTIPHYSICS

The ARW-MP experiment significantly improves upon the FSS, at 99% confidence, of ARW-SP at all

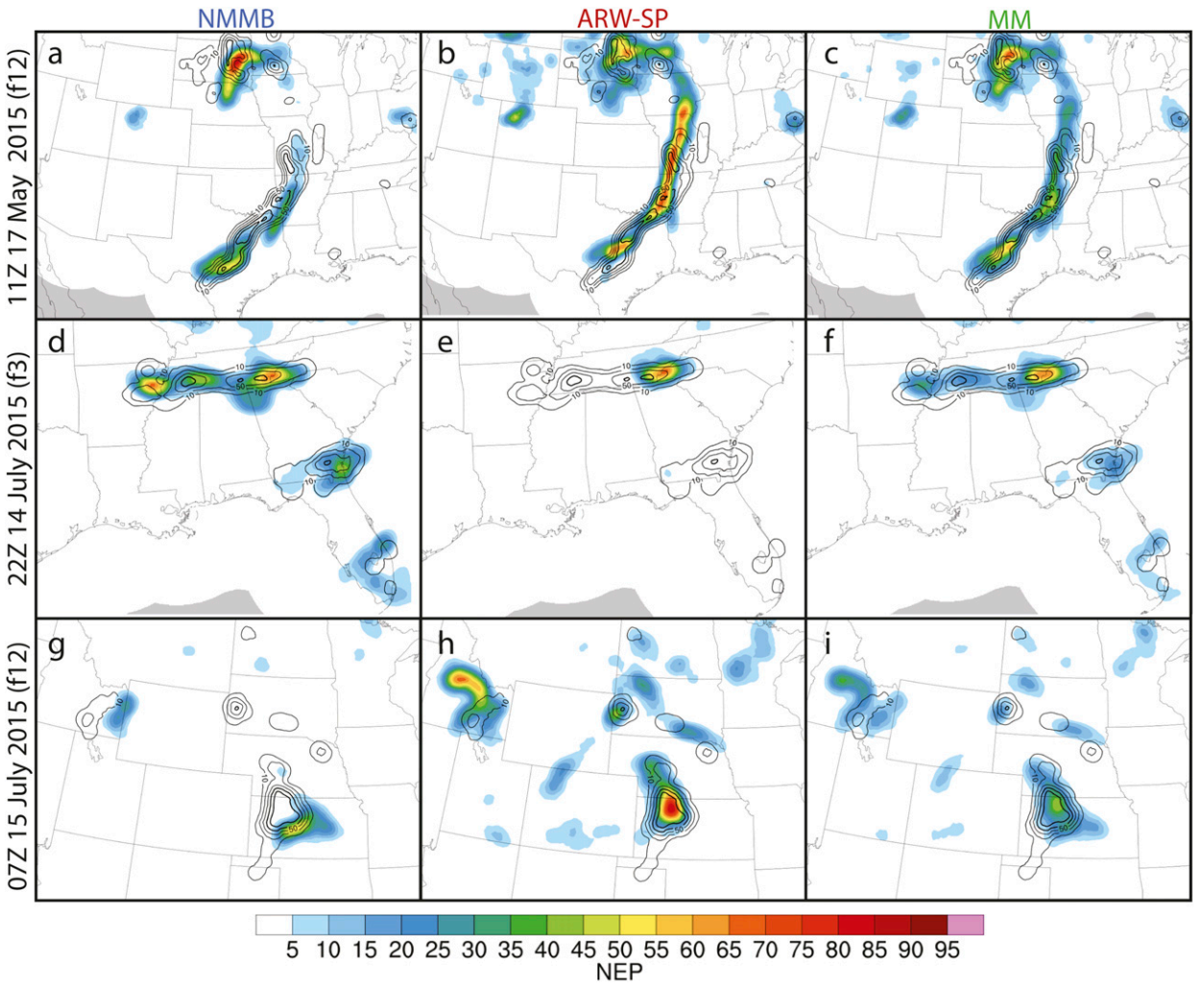

FIG. 5. NEP (color fill) of composite reflectivity greater than 30 dBZ for (left) NMMB, (middle) ARW-SP, and (right) MM experiments. Valid times for each row are shown at the left with forecast hour in parentheses. Neighborhood probability of MRMS observed composite reflectivity greater than 30 dBZ is plotted in black contours at 20% intervals, starting at 10%.

but two forecast hours at the low 2.54 mm precipitation threshold (Fig. 6a). Additionally, ARW-MP is about equal to or higher in FSS than MM at all forecast hours, with 7 of the final 10 h showing statistical significance. At higher thresholds (Figs. 6c,e), ARW-MP showed worse FSS for early forecast lead times compared to ARW-SP, then significantly improved FSS in the second half of the forecast, the times that ARW-SP had poorer results. ARW-MP is approximately equal to MM in skill at the 6.35 mm threshold for hours 9–18. MM otherwise has more times of statistically significantly better skill at higher thresholds (for 6 and 11 out of 18 forecast hours, respectively) than ARW-MP despite the improvement of ARW-MP over ARW-SP over the latter half of the forecast. In terms of ROC area of composite reflectivity, ARW-MP showed significant improvement over each NMMB, ARW-SP, and MM experiments at both 30 and 40 dBZ thresholds, with a majority of times showing statistical significance (Figs. 6b,d). In terms of 1-h QPF, the ARW-MP experiment had nearly perfect reliability from 0%–80% NEP at 2.54 mm with a slight overforecast bias at 90% and 100% (Fig. 4a). At the 12.7 mm threshold, ARW-MP had improved reliability over ARW-SP, NMMB, and MM at NMEP above 30%, with the largest improvements over MM in the 30%–60% range (Fig. 4b).

Figures 7 and 8 show subjective examples of cases and forecast hours when ARW-MP was more skillful than NMMB, ARW-SP, and MM. These cases tended to have strong sensitivity to near-surface physics, with the diversity of sampled physics from ARW-MP leading to improved forecasts. The first example is from the

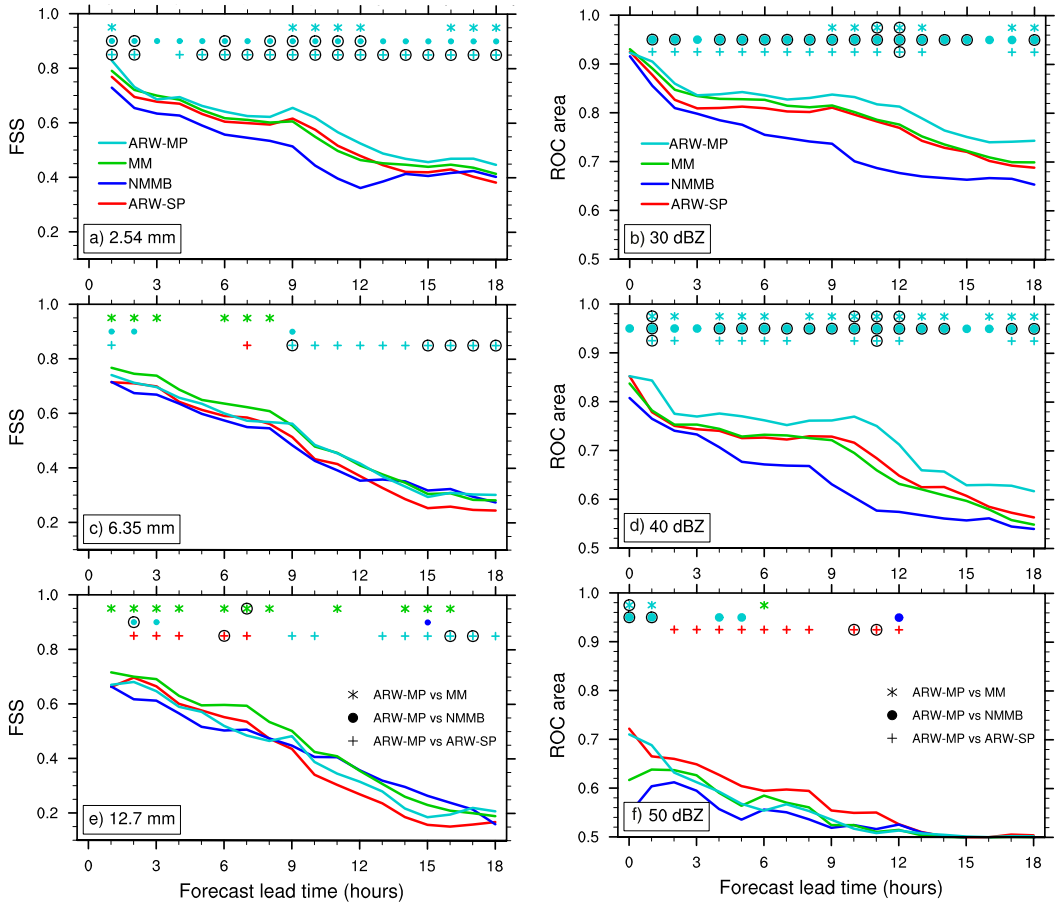

FIG. 6. As in Fig. 3, but with the addition of experiment ARW-MP and statistical significance tests of ARW-MP against MM, NMMB, and ARW-SP experiments.

22 May 2016 case, which featured initially severe discrete convection ahead of a dryline in western Texas northward into Kansas, as well as ahead of a cold front extending northward. This convection grew upscale into MCSs during the overnight hours, fueled by moisture from the south advected from a strong low-level jet. Figures 7a–d shows that ARW-MP tended to capture the northern plains convection with highest probabilities in the correct locations, as well as coverage of isolated cells in west-central Texas. Eight hours later (Figs. 7e–h), ARW-MP has managed to maintain the highest probabilities and largest spread of coverage over all the MCS activity in the northern plains as well as the MCS in southern Oklahoma. The second example is from the 17 June 2016 case, which featured two MCS systems and weak large-scale ascent (Fig. 8). Outflow from a midday MCS in Minnesota helped fuel additional storm development and upscale MCS growth to the south along the cold front, while an MCS initiated ahead of a cold front over Nebraska and Kansas due to strong instability from daytime heating. The NMMB does a particularly poor job simulating these MCSs (Figs. 8a,e), while the ARW-SP does markedly better (Figs. 8b,f). ARW-MP clearly improved upon ARW-SP with stronger probabilities and better southward propagation in both MCSs (Figs. 8c,g). MM does have probability coverage of both MCSs, but is significantly hindered by the poor NMMB simulation.

3) IMPACT OF STOCHASTIC PHYSICS

With SKEB applied to ARW-MP, small but statistically significant improvements were found in FSS at the highest QPE thresholds compared to ARW-MP, particularly for the 12.7 mm threshold within the last 8 forecast hours, 5 of those being significant (Fig. 9). On the other hand, small magnitude reductions in FSS were found at the light threshold for the last 8 h, 4 of which were significant. These small differences are statistically significant because, compared to other experiments, there is less variation in the case-by-case differences due to the persistent, larger-scale effects of SKEB.

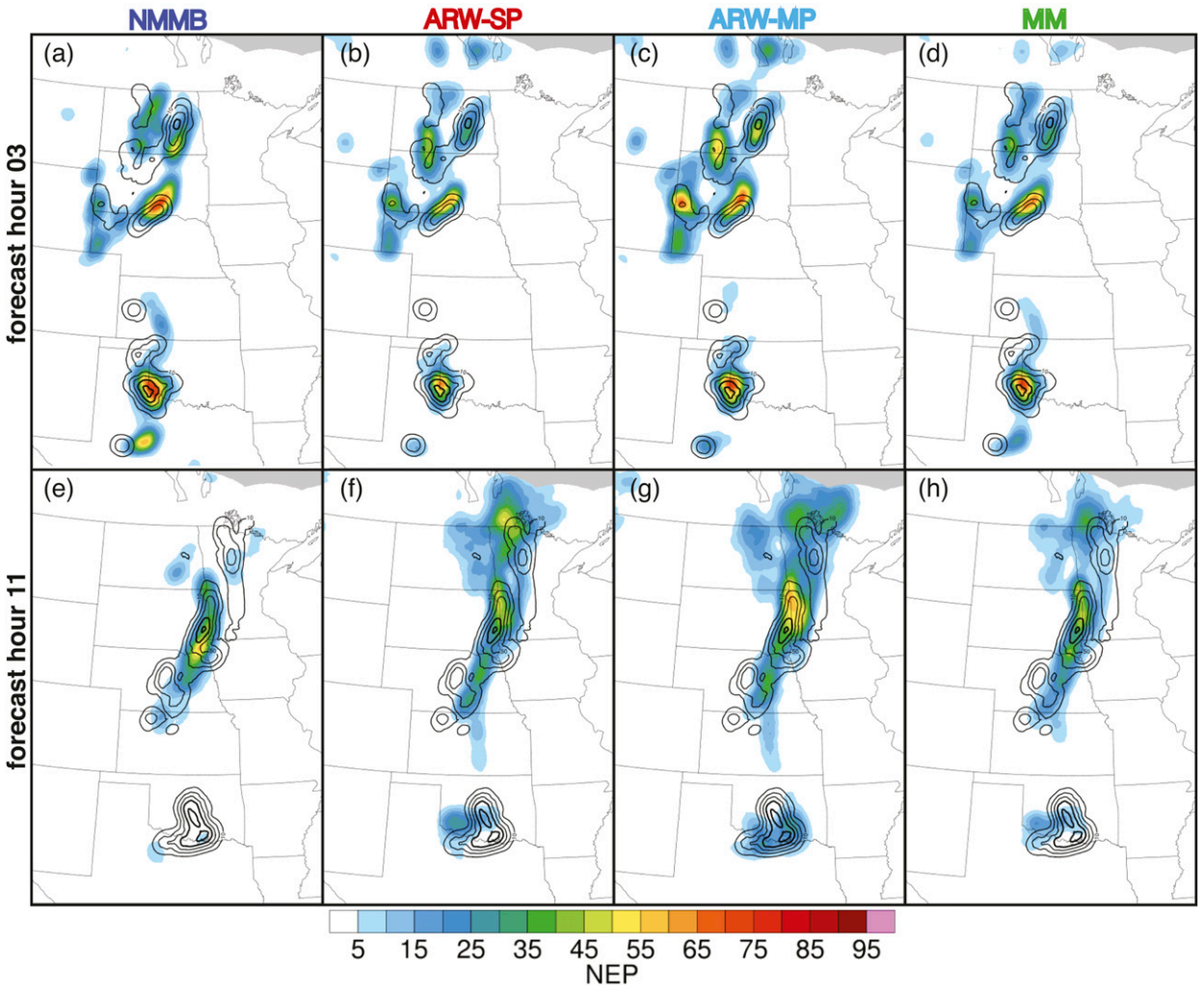

FIG. 7. NEP of CREF > 30 dBZ as in Fig. 5, but for the 22 May 2016 case at forecast hours (a)–(d) 3 and (e)–(h) 11. Respective experiments are indicated at the top of each column.

This result is similar, though less pronounced, to results of Duda et al. (2016), who found consistent improvements in BS with the application of SKEB in combination with a multiphysics ensemble. The less pronounced results may be related to IC uncertainty being sampled by the convective-scale DA system, whereas Duda et al. (2016) did not address IC uncertainty causing the SKEB to also partially compensate for the missing IC uncertainty. By the final approximately 3 h of the forecast, the FSS of ARW-MPSKEB surpasses the skill of MM at 6.35 and 12.7 mm, with statistical significance at hours 17 and 18, a notable improvement compared to the lower FSS of ARW-MP (without SKEB). In terms of ROC area, the differences between ARW-MPSKEB and ARW-MP were minor for all thresholds, though like FSS, some differences were also statistically significant favoring SKEB (not shown).

The reliability diagram in Fig. 10 shows results aggregated over the final 6 h of the forecast, when SKEB has the largest impact within the forecast. At the lighter 2.54 mm threshold, there is a noticeable improvement in reliability of the mid 50%–70% probability ranges comparing ARW-MPSKEB to ARW-MP. For the NMEP-based reliability diagram at 12.7 mm, the ARW-MPSKEB experiment improves the reliability of the 70%–100% range, despite a reduction in sharpness at these probabilities (Figs. 10b,d). In each case, ARW-MPSKEB has the best reliability of all experiments.

Although skill is an important consideration in the application of SKEB, the primary goal of SKEB is to increase the spread of generally underdispersive CAE forecasts. A simple way to quantify the ensemble spread of precipitation is by the application of the correspondence ratio (CR; Stensrud and Wandishin 2000). CR was

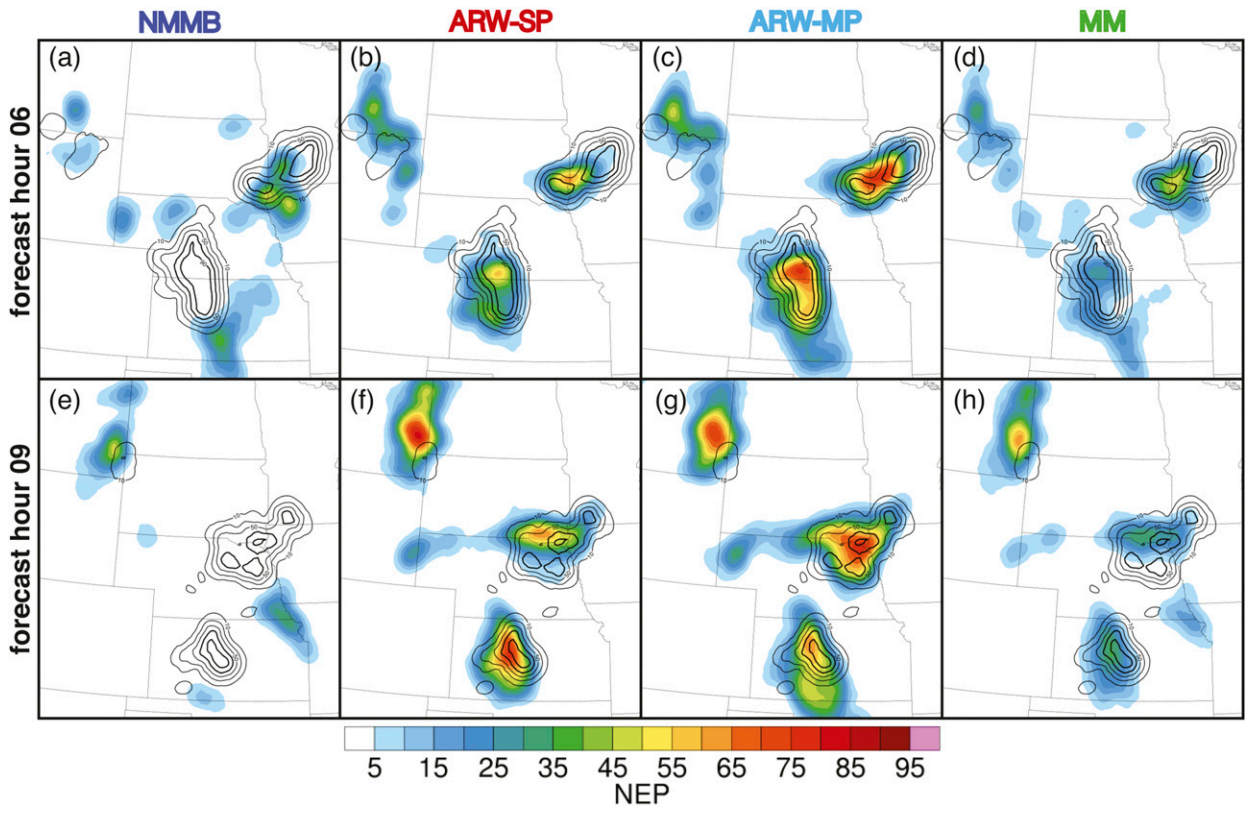

FIG. 8. As in Fig. 7, but for the 17 Jun 2016 case at forecast hours (a)–(d) 6 and (e)–(h) 9.

applied for nocturnal MCS simulations by Johnson and Wang (2017) and Johnson et al. (2017). Briefly, CR is the ratio of the number of grid points where ensemble members agree that a threshold was exceeded (intersection area), to the number of grid points where any ensemble member exceeded the threshold (union area), each here summed over all grid points in the 3-km verification domain. Any predetermined number of ''ensemble agreement'' can be used, with higher numbers being stricter applications (the strictest being a requirement that all ensemble members agree). Here we chose a more relaxed number of 4 out of 10 members as was also used in Johnson and Wang (2017).[3] A CR of 0.0 (1.0) indicates all members disagree (agree) on where precipitation is forecast, thus maximizing (minimizing) spread. In Fig. 11, MM has the largest amount of spread (lowest ensemble agreement) in QPE among all experiments at all times, while ARW-SP and NMMB tend to have the least amount of spread. Experiment ARW-MP improves the spread of ARW-SP after about 6 h into the forecast, and the addition of SKEB in ARW-MPSKEB further increases the spread (decreases CR) within the last half of the forecast. By hour 18, ARW-MPSKEB has as much spread in precipitation as MM.

4) IMPACT OF BIAS CORRECTION

Systematic biases may be playing a role in the precipitation and reflectivity verification, possibly affecting some of the comparisons. These biases are important to consider given that they only artificially increase forecast uncertainty within an ensemble (e.g., Eckel and Mass 2005). Further, Loken et al. (2019) found that their bias-corrected single-physics ensemble was as skillful and reliable as their multiphysics configuration for precipitation verification. This leads to an important consideration—how are the SMSP and model error experiments each affected by systematic biases?

A cumulative density function (CDF)-based bias correction procedure was applied to both composite reflectivity and 1-h precipitation forecasts. CDF-based bias correction works by replacing forecast data with the observed value at an equivalent percentile (for more

[3] Johnson and Wang (2017) examined CR with 4- and 8-member agreement, out of 10 total members. CR with 4-member agreement had larger magnitude values and differences among experiments than 8-member agreement; however, the relative positioning between experiments remained the same for both agreement values.

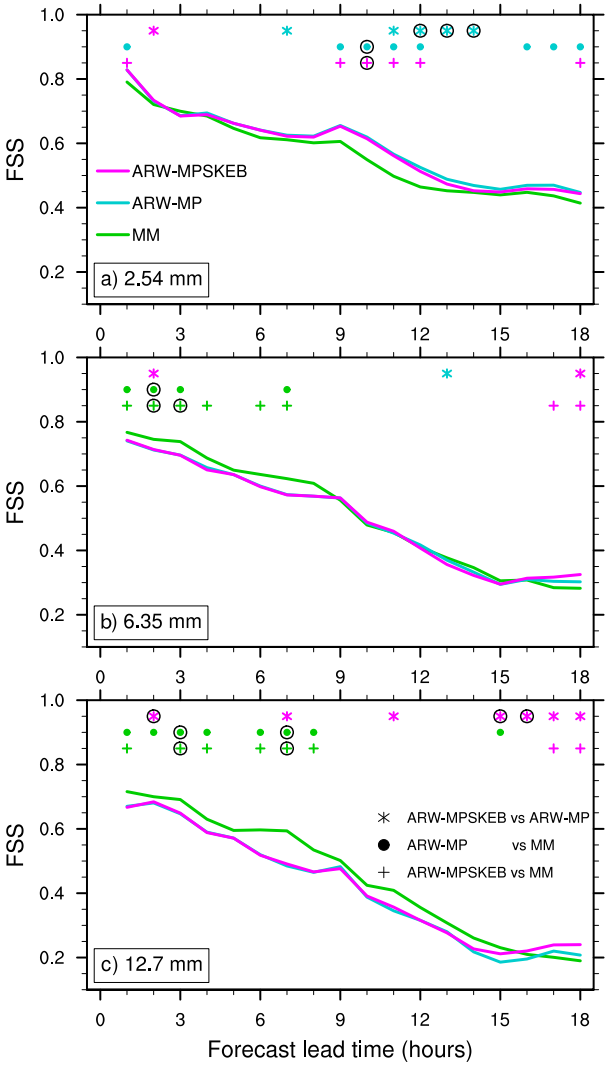

FIG. 9. As in the left column of Fig. 3, but for experiments MM (green), ARW-MP (cyan), and ARW-MPSKEB (purple). Markers indicate statistically significant differences between pairs of experiments indicated in (c), color coded by the experiment with the higher FSS at a given time.

details see, e.g., Johnson and Wang 2012; Voisin et al. 2010). This CDF-based correction was implemented hour-by-hour and ensemble member-by-member for each of the experiments with forecast and observed CDFs aggregated over all 10 retrospective cases. Typically the CDFs would be built with an independent training dataset; however, here the CDFs were built with all 10 cases to ensure a robust distribution.

Figure 12 presents the results of bias-corrected verification FSS in 1-h precipitation and ROC area of composite reflectivity. The relative position of experiments remains nearly identical to Fig. 3. The differences in SMSP experiments are increased at early lead times, notably due to a slight reduction in skill of NMMB after bias correction coupled with an increase for ARW-SP within the first 6 h of the forecast. In terms of ROC area, on the other hand, the differences among SMSP experiments were reduced due to the increased score for experiment NMMB. There are less statistically significant differences in SMSP experiments than shown in Fig. 3. particularly for higher thresholds of ROC area and FSS. The MM experiment is also notably improved with highest ROC area at 30 dBZ, significantly higher than ARW-MP in 5 of the first 7 forecast hours. Although FSS did not improve much, MM had notably reduced FBS after bias correction that was significantly lower than each of the other experiments at most forecast hours (not shown). Experiment ARW-MP also has some improved FSS for 1-h precipitation, but the relative differences compared to ARW-SP—including significance—remains largely unchanged after bias correction. On the other hand, ARW-MP has notably lower scores in ROC area, indicating that systematic biases were artificially inflating ROC area for reflectivity previously. The noted advantages of ARW-MP in ROC area have been significantly reduced overall, though many forecast hours remain significantly higher than either SMSP experiment, especially at the lowest threshold.

These results suggest that precipitation and reflectivity forecasts with MM configurations may be helped by bias correction in cases where one model has significantly larger error than the other. Although the MM avoids the "skill drop-off" issue of SMSP experiments, when one model core is inferior to the other the MM will not have the best skill. Since bias correction helps correct the error gaps between models (as seen in ROC area, for example), the skill of MM increases. The caveat here is that this bias correction is "perfect"–as in bias corrections were not performed with a training set of data due to the limited number of cases. Thus, the differences from raw to bias-corrected verification might be overestimating the effect of bias correction based on independent training data.

b. *Upper-level verification against RAP analyses*

Time series of ensemble mean RMSE and ensemble spread are shown for selected upper-air variables in Fig. 13, with vertical profiles in Fig. 14 at forecast hour 18, each averaged over all 10 cases. In terms of error, ARW-SP is lower than NMMB by about 1 m s^{-1} for 250 hPa zonal wind and 0.2 K in 500 hPA temperature (Figs. 13a,b). In fact, the NMMB has the highest error at nearly all levels of wind, temperature, and dewpoint; only with geopotential height does the NMMB have lower errors competitive with other experiments (Fig. 14)

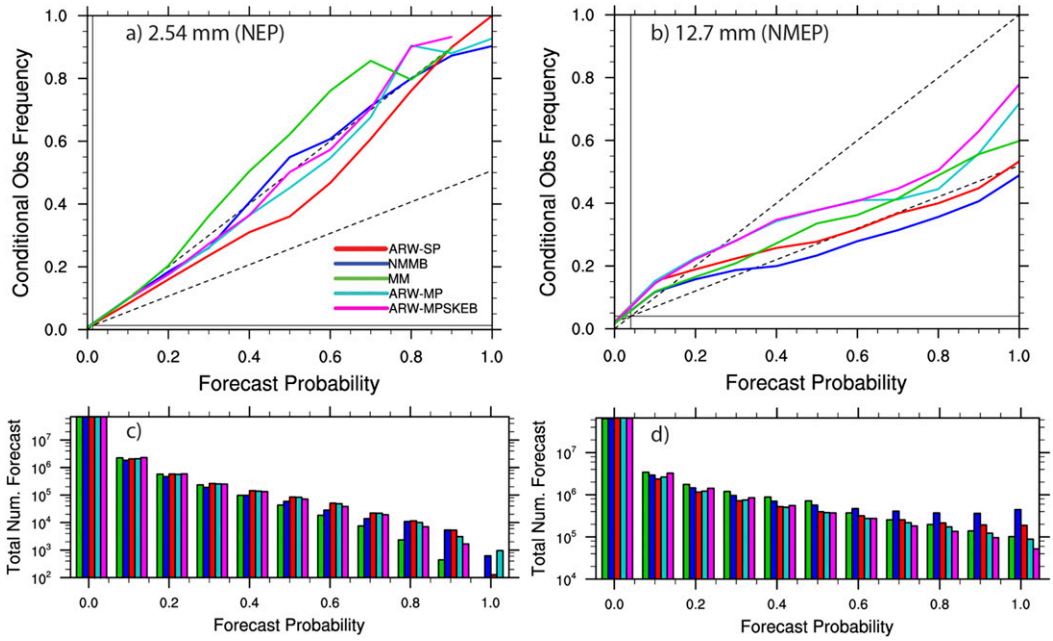

FIG. 10. As in Fig. 4, but aggregated over forecast hours 13–18 and with the addition of experiment ARW-MPSKEB (purple).

The other experiments (ARW-SP, ARW-MP, MM, ARW-MPSKEB) are clustered more closely together for upper-level fields above 500 hPa. At and below 700 hPa, in particular for thermodynamic variables, the experiments have larger error differences with ARW-MP and ARW-MPSKEB the lowest in error followed by slightly higher MM (Figs. 13 and 14). The exception is MM with the lowest error for early forecast lead times of 850 hPa zonal wind (Fig. 13d) and the lowest geopotential height error across all levels and forecast times (Figs. 13c and 14c), whereas the ARW-based experiments have large errors due in part to biases (not shown). Both MM and ARW-MP show clearer error reduction for lower-level variables. For instance, MM has the lowest error in 700 hPa geopotential height and 850 hPa zonal wind for many of the forecast hours (Figs. 13c,d). ARW-MP and ARW-MPSKEB have the lowest magnitude error in 850 hPa temperature and 925 hPa dewpoint. SKEB has practically no impact on the error of any of these variables, as ARW-MPSKEB and ARW-MP have nearly identical error at all levels. Past studies comparing experiments with and without SKEB also showed small changes in error of most upper-air fields (e.g., Berner et al. 2015; Duda et al. 2016; Jankov et al. 2017, 2019).

In terms of spread, both SMSP experiments tended to have the lowest amounts among the experiments, as expected. MM displays a large amount of spread at all levels and variables (Figs. 13 and 14); furthermore, this increased spread occurs at the beginning of the forecast and is maintained throughout (Fig. 13). ARW-MP adds only a marginal amount of spread at upper levels; however, closer to the surface the amount of spread added increases significantly, particularly in thermodynamic variables around 850 hPa and below (Fig. 14). Additionally, the spread is increased over time and it takes approximately 4–6 h to show noticeable increases over ARW-SP (Fig. 13). The addition of SKEB in ARW-MPSKEB has a large impact in ensemble spread growth throughout the forecast at all levels, most notably at the highest levels and for wind variables. By the end of the forecast, ARW-MPSKEB has larger spread than MM for all wind levels, and in temperature and geopotential height for levels 700 hPa and below (Figs. 14a–d); in dewpoint the spread is still slightly lower for upper levels, with differences decreasing closer to the surface (Fig. 14e). In general SKEB has less impact on spread in thermodynamic variables, particularly in dewpoint. Conversely, MM shows the most spread in moisture above the surface.

c. *Surface field verification against RTMA analyses*

Verification against 2.5-km RTMA analyses for surface variables is shown in Figs. 15 and 16. Comparing SMSP experiments, the error of ARW-SP wind is about 0.3–0.4 m s^{-1} lower than NMMB—significantly lower at 99% confidence across all hours—as well as significantly lower temperature error for middle forecast hours 5–14, with differences between 0.1 and 0.3 K. On the other

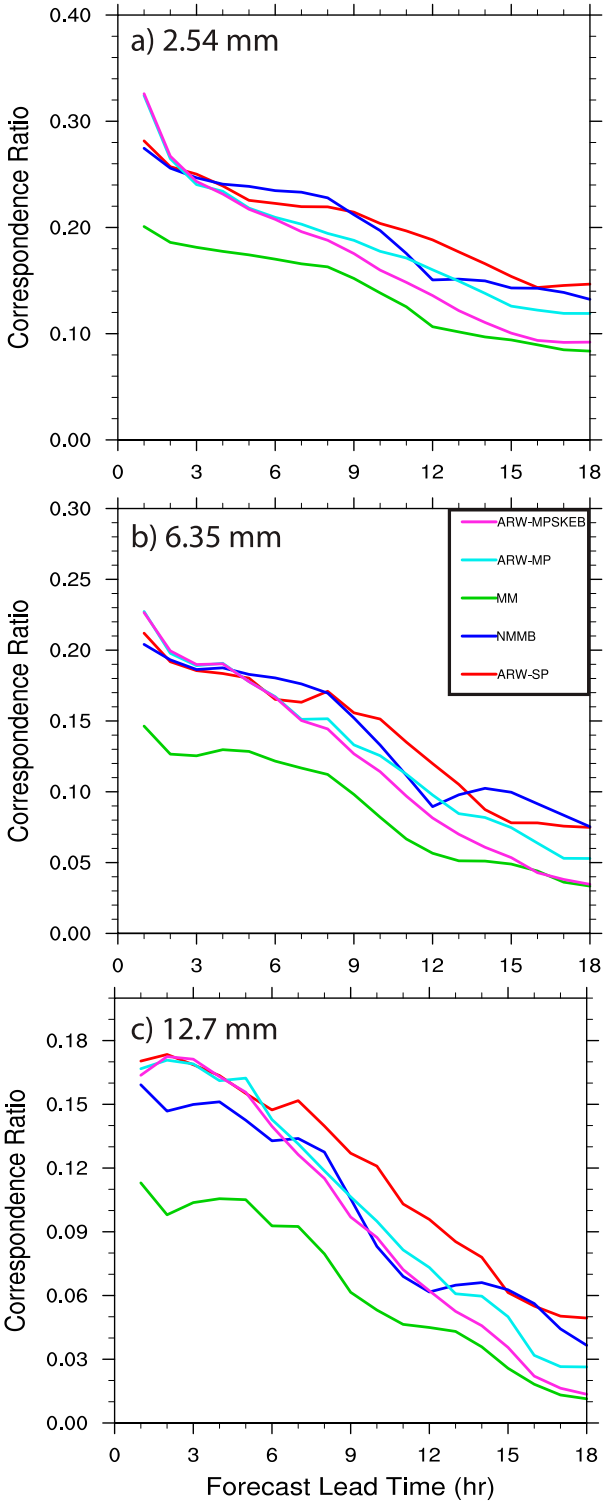

FIG. 11. Correspondence ratio at indicated thresholds in 1-h precipitation for 4-member agreement.

hand, NMMB has significantly lower dewpoint error across the forecast, around 1 K less than ARW-SP (significant at 99% confidence). MM has significantly lower temperature error (averaging ~0.3 K) across nearly all hours than both SMSP experiments. However, MM is only significantly lower in error than the worst SMSP experiment for dewpoint and wind, and against the best SMSP experiment is at times significantly worse (e.g., last 6 h in wind versus ARW-SP; first 8 h in dewpoint versus NMMB). This is consistent with contingency-based verifications that showed when one SMSP is clearly inferior to the other, MM can do no better than the best SMSP experiment. The ARW-MP experiment has statistically significant reductions in temperature (0.1–0.3 K) and dewpoint error (~1 K) at most forecast hours compared to ARW-SP.

Comparing model error experiments MM to ARW-MP, the differences in wind are generally small (no larger than 0.1 m s^{-1}), with MM lower in early hours 1–4 and ARW-MP lower at later forecast hours 13–18. In temperature, MM is significantly lower (0.1–0.2 K) than ARW-MP for early hours 1–6, after which the differences become negligible. However, in terms of dewpoint ARW-MP is significantly lower in error by about 0.1–0.3 K, significant over forecast hours 2–18. Statistically significant reductions in error are found with ARW-MPSKEB compared against ARW-MP for many forecast hours of dewpoint and wind, although the reduction is nearly negligible in terms of magnitude (~0.01–0.02 K and ~0.01–0.02 m s^{-1}, respectively). Note in Figs. 15 and 16 that because of the small magnitude differences, ARW-MP (cyan) is visually overlapped by ARW-MPSKEB (pink) at most times in terms of RMSE.

The spread for surface variables follows a similar pattern of spread in upper-level variables below 700 hPa. MM tended to present the most spread, particularly for early forecast hours, averaging around 1.1–1.2 K and m s^{-1} in temperature and wind, respectively, and 1.3 K in dewpoint. NMMB had roughly 0.2 K larger spread in both thermodynamic variables than ARW-SP, but about 0.1 m s^{-1} less spread for wind. The increase in spread of ARW-MP is more immediately apparent for surface variables with statistically significant (not shown) increases compared to ARW-SP within the first 3 h, whereas at upper levels it took longer for the forecast to noticeably increase spread. The spread increases throughout the forecast and matches the spread of MM within later forecast hours. The addition of SKEB increases the spread significantly (not shown) in both thermodynamic and dynamic fields; however, the effect is larger in magnitude and present earlier for 10-m wind fields.

The most notable of the surface biases shown in Fig. 16 is in 2-m dewpoint. ARW-SP has a significant

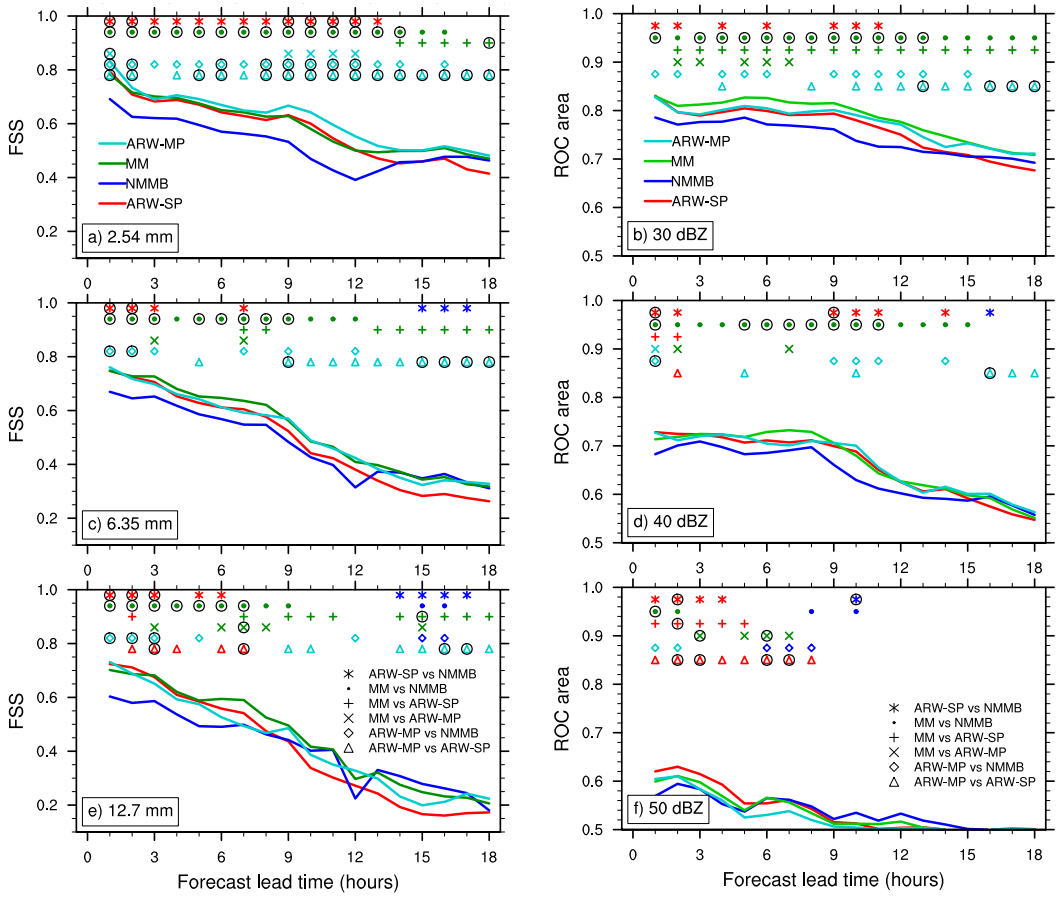

FIG. 12. As in Fig. 6, but using bias corrected (left) precipitation and (right) composite reflectivity.

moist bias around 1.1 K across the forecast. NMMB also shows a moist bias, though it averages closer to 0.5 K. Because both biases are of the same sign, MM thus also has a significant moist bias, roughly split in between the biases of ARW-SP and NMMB (~0.8 K). Experiments ARW-MP and ARW-MPSKEB show the least amount of bias in dewpoint (slight dry bias ~0.2 K). This reduced bias is partially responsible for the significant thermodynamic error reductions of ARW-MP compared to ARW-SP (Figs. 15a,b). A region-based (western versus eastern CONUS) bias correction for surface variables was attempted, but the reduction in RMSE was relatively small after correction and had little impact on the relative significant differences among experiments as seen in Fig. 15 (not shown).

d. Ensemble spread diagnostics

The increased ensemble spread within MM and ARW-MPSKEB can be illustrated and further compared in two-dimensional atmospheric fields. For example, Fig. 17 shows the 500 hPa geopotential height spread for the final forecast hour of the 16 May 2015 case over the eastern CONUS. The 5760-m contour is also shown here for the ensemble members and RAP analysis extending to the northeast through Texas, Oklahoma, and Missouri. This contour was part of the trough that helped set up a significant severe weather outbreak for this case. The NMMB model (Fig. 17a) has a slight eastward bias of the ensemble, though the RAP analysis remains within the ensemble envelope across most of the central CONUS. On the other hand, ARW-SP (Fig. 17b) has a large westward bias in ensemble members, with the RAP analysis contour lying mostly outside of the envelope. The MM experiment (Fig. 17c) displays greatly increased spread across the domain, with the RAP analysis contour falling within the ensemble envelope at all locations. It should be noted, however, that clustering appears in the ensemble, most substantially in the northeast region. The ARW-MPSKEB experiment (Fig. 17d) displays a notable increase in the spread compared to the ARW-SP experiment. Although a slightly westward bias still exists, the RAP contour lies within or very near the ensemble envelope at most locations. Another example of spread is shown in Fig. 18

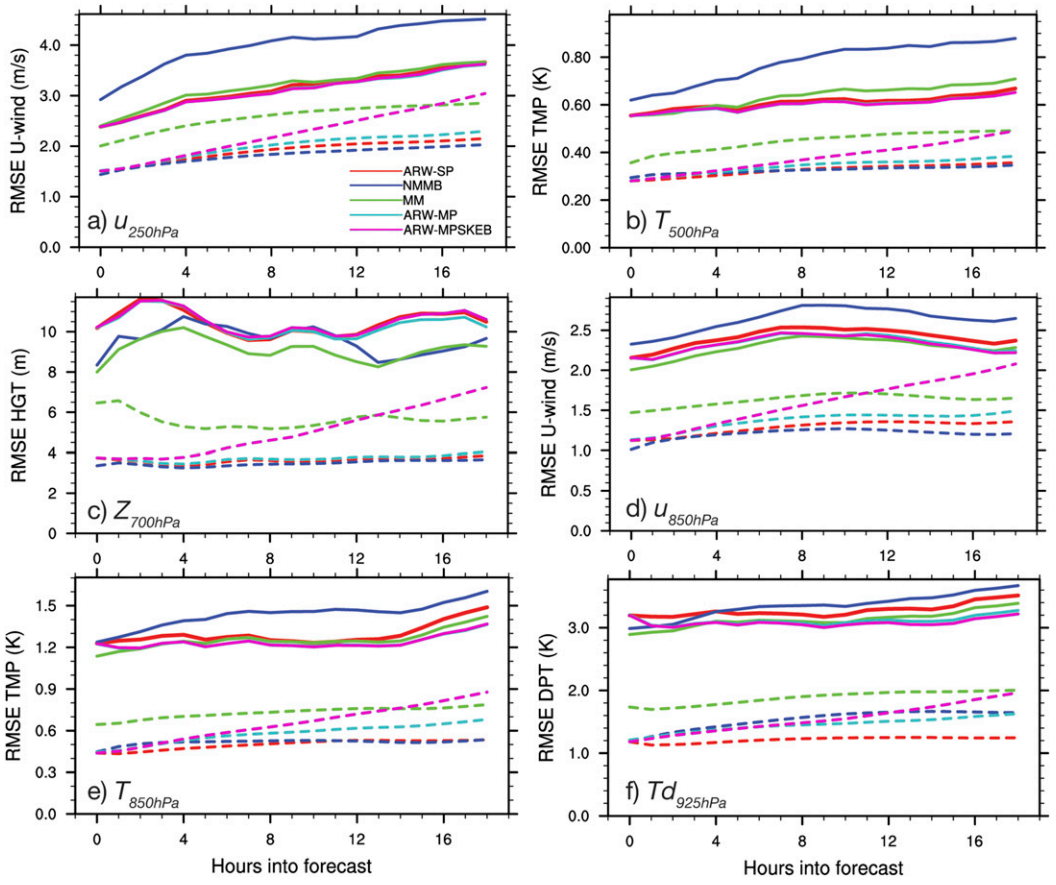

FIG. 13. Ensemble-mean RMSE (solid) verified against RAP observations, and ensemble spread (dashed) for all experiments in (a) 250-hPa u wind ($m\,s^{-1}$), (b) 500-hPa temperature (K), (c) 700-hPa geopotential height (m), (d) 850-hPa u wind ($m\,s^{-1}$), (e) 850-hPa temperature (K), and (f) 925-hPa dewpoint (K).

for 850-hPa dewpoint, with the 6°C contour plotted for the 22 May 2016 case. This contour outlines the moisture plume extending from the gulf to the northern Great Plains, with a sharp highly variable dryline extending northward through west Texas. Both the NMMB and ARW-SP struggle maintaining spread to capture all of the details of the east–west variations in the RAP moisture contour, while MM and ARW-MPSKEB have substantial improvements to the spread in order to capture these moisture variations.

Another method to diagnose ensemble spread is via rank histograms, where observations are counted according to rank relative to ensemble forecasts (e.g., Hamill 2001). With ''perfect spread,'' we would expect an equal chance for a given observation to fall within any of the 11 bins created by sorted ensemble members. Most often for CAE, significant underdispersion exists where most observations fall in the lowest and highest bins and lower amounts exist in middle bins, as seen most dramatically for experiments NMMB and ARW-SP in Fig. 19. The MM experiment has a ''w'' shape in all fields examined, where there are an excessive number of observations in the middle bin in addition to the lowest and highest bins, and lower amounts in other bins. This is a clear indication of clustering within the ensemble, as we subjectively saw in Fig. 17. While the spread is certainly increased compared to either SMSP experiment, clustering is an undesirable quality of the resulting ensemble distribution. Because of this high middle bin, the actual closeness of spread to an ideal uniform rank distribution is substantially worse for MM than for ARW-MPSKEB, in terms of median-absolute-differences of bin frequencies (shown in legends of Fig. 19). The impact of SKEB is again seen most strongly at upper-level fields (Figs. 19a,b). For surface fields, biases can be seen within the rank histograms. In temperature, NMMB and ARW-SP (and MM to a lesser extent) show cold biases in addition to underdispersion, while experiments ARW-MP and ARW-MPSKEB have minimal bias and further flatten the histogram. In dewpoint, each of NMMB, ARW-SP,

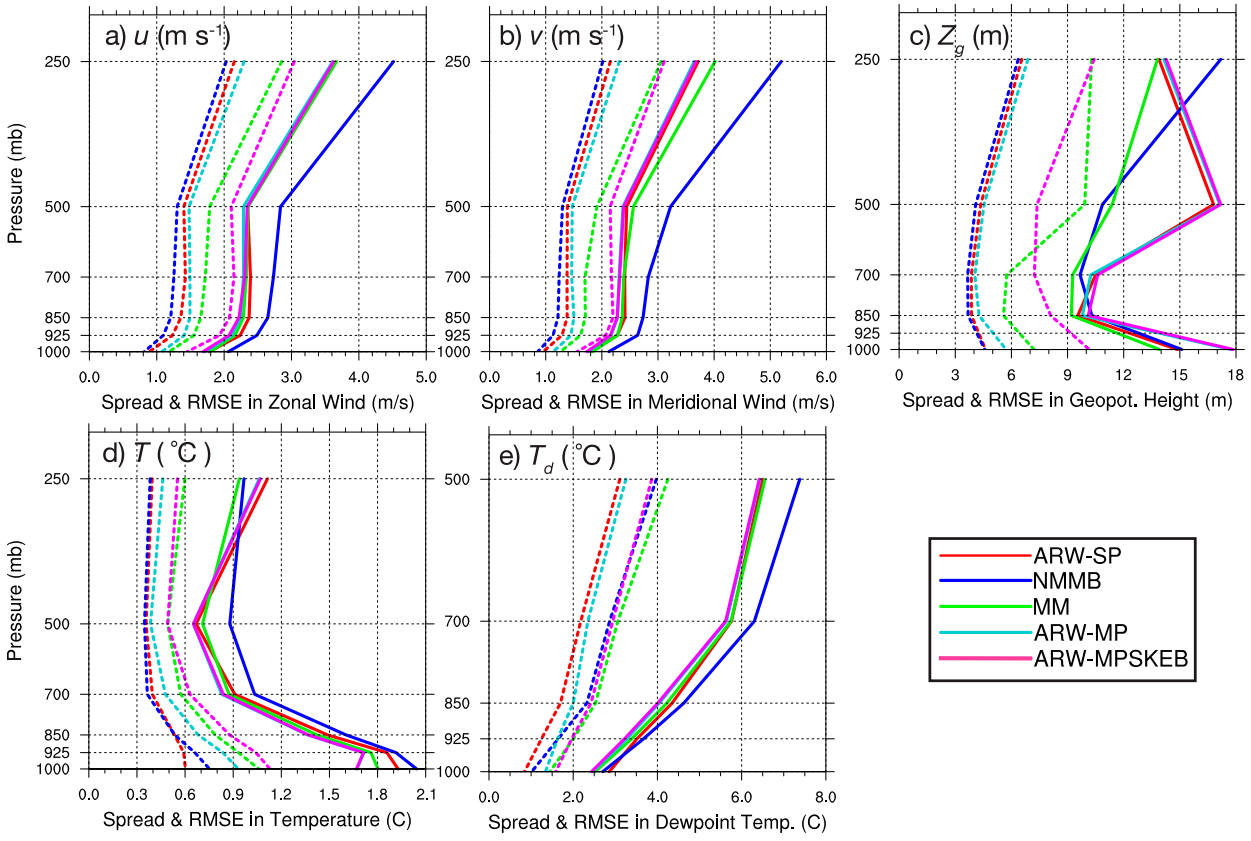

FIG. 14. Vertical profiles of ensemble-mean RMSE (solid) and ensemble spread (dashed) for the indicated fields in each plot, averaged over all cases for forecast hour 18. Note the different vertical scale in (e) for dewpoint.

and MM have a large moist bias as most observations fall in the first rank, while ARW-MP and ARW-MPSKEB have flatter, mostly unbiased histograms.

5. Summary and discussion

The optimal design of convection-allowing ensemble (CAE) forecasting systems necessitates the use of techniques to address model error uncertainty within the ensemble. Some studies have demonstrated the usefulness of various model error techniques for CAE, but usually in systems where IC uncertainty is either neglected or only crudely addressed. Here, the GSI-based EnVar system extended for convective-scale DA by Johnson et al. (2015) and Wang and Wang (2017) was applied to provide a full multiscale sampling of IC uncertainties, which provides a more meaningful basis for a comparative study on the effects of model error techniques. Two configurations of this EnVar were used in this study: EnVar with the NMMB model during DA cycling, and EnVar with WRF-ARW model during DA cycling. These configurations were applied to 10 retrospective case studies with a variety of different synoptic forcing and severe weather morphologies. The 10-member CAE forecasts out to 18 h were launched from both of these samples of multiscale analyses, covering a total of five CAE forecast experiments. Two experiments, NMMB and ARW-SP, applied the same single-model single-physics (SMSP) model settings to the forecast. Additionally, a multimodel (MM) experiment combined five random ensemble members each from NMMB and ARW-SP. Another experiment applied model error via different combinations of PBL, microphysics, and land surface physics schemes (ARW-MP), as well as the combination of multiphysics with the SKEB stochastic physics scheme to address the effect of subgrid-scale physics uncertainty on the large-scale dynamical tendencies (ARW-MPSKEB).

The results were compared objectively in terms of 1-h QPE and composite reflectivity using neighborhood-based verifications metrics including FSS, ROC area, reliability and sharpness diagrams, and correspondence ratio. Subjective analysis of NEP $\geq$ 30 dBZ highlighted scenarios that MM and ARW-MP performed the best. Additional verification was performed on ensemble

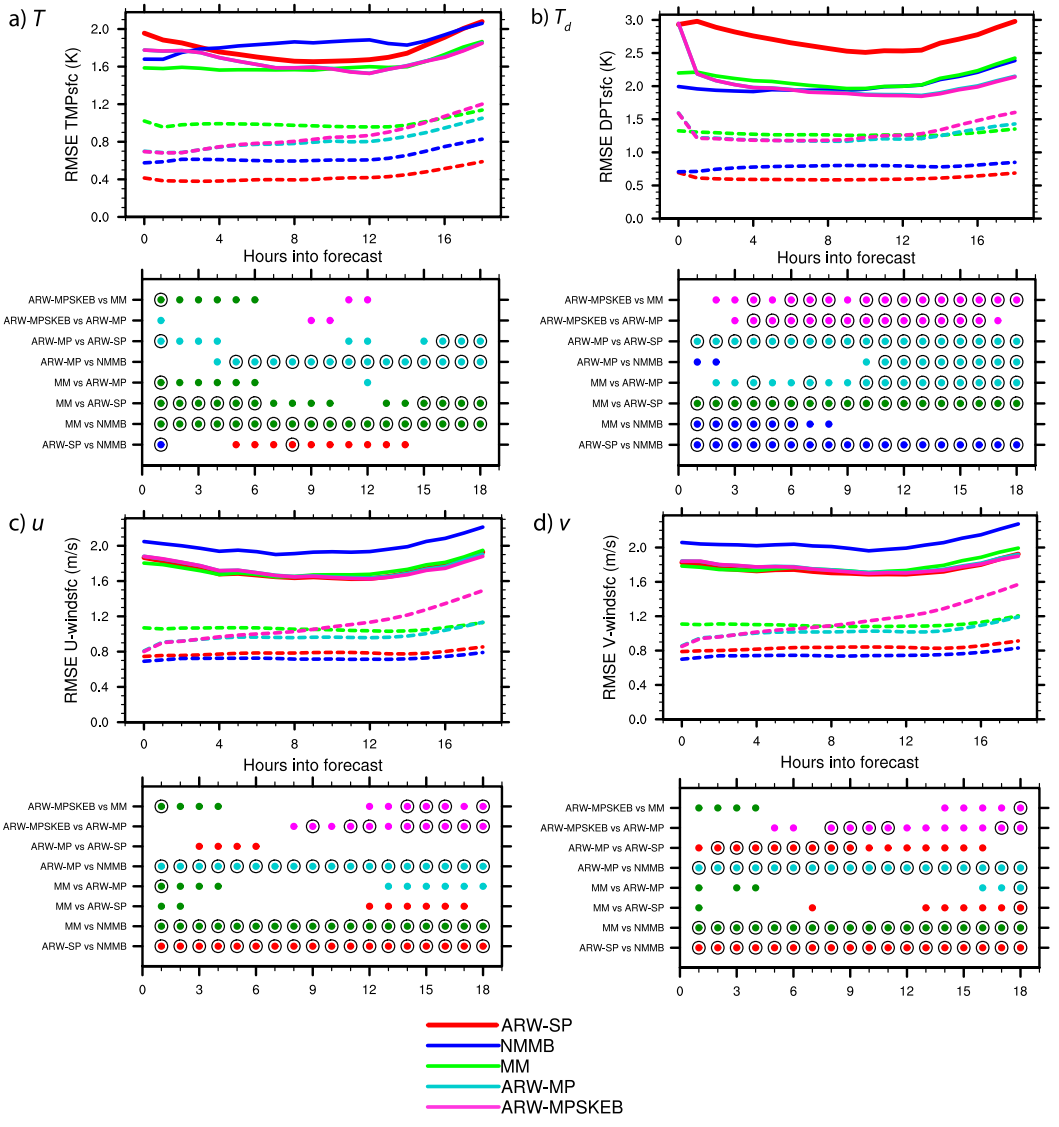

FIG. 15. Verification (RMSE, solid lines) and ensemble spread (dashed lines) of surface variables, with a panel of markers below indicating statistically significant differences (90% confidence) in RMSE for the difference tests indicated, color coded by the significantly lower experiment in RMSE. Markers outlined with black indicate significant differences at 99% confidence. (a) 2-m temperature (K), (b) 2-m dewpoint (K), (c) 10-m u wind (m s^{-1}), and (d) 10-m v wind (m s^{-1}). The ensemble-mean fields were verified against RTMA 2.5-km analyses.

mean upper-level fields (against hourly RAP analyses) and surface fields (against 2.5-km RTMA analyses) and compared to average ensemble spread. The differences in spread among the experiments were further diagnosed using rank histograms and representative 2D examples. The main results of this study are summarized in the following bullets:

- Among SMSP experiments, ARW-SP had superior performance to the NMMB for lighter precipitation fields and earlier forecast times, as well as much of the mean RMSE verification. The NMMB decayed MCSs too early in cases where decaying MCSs occurred in reality (most often between midforecast hours 6–12), relative to the ARW-SP. However, the NMMB had superior performance in the final 6 forecast hours, particularly with heavy precipitation (6.35 and 12.7 mm). This was primarily due to enhanced accuracy in the extent of new CI and redeveloping MCS events, while the ARW-SP often had enhanced spurious activity and overforecasted some MCS areas.
- Each of the model error experiments MM, ARW-MP, and ARW-MPSKEB compared favorably to NMMB and ARW-SP in many of objective verification scores and

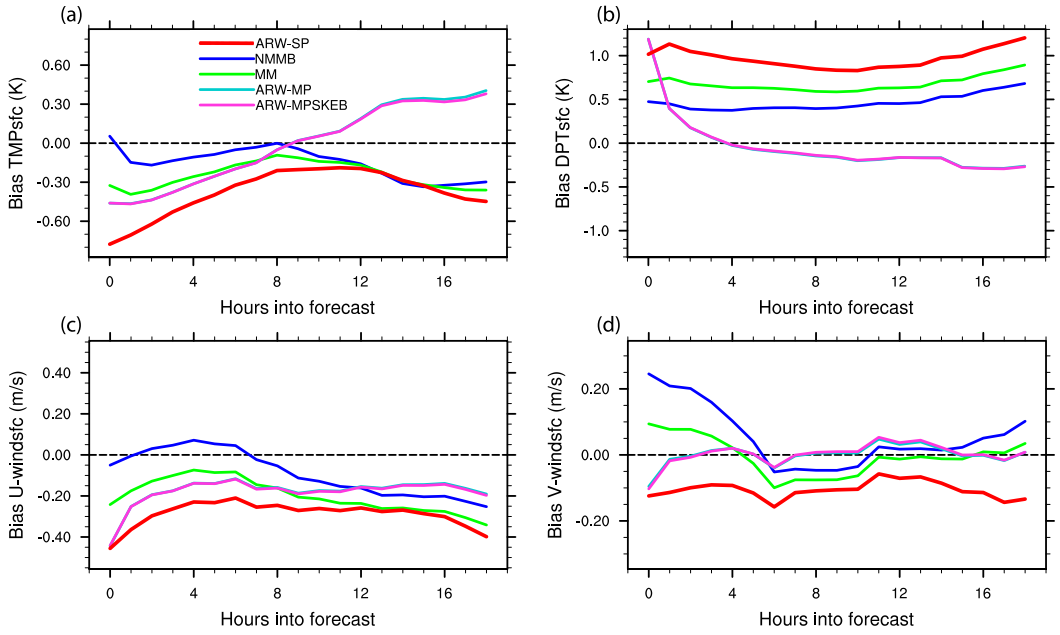

FIG. 16. Biases of the ensemble mean for surface fields (a) 2-m temperature (K), (b) 2-m dewpoint (K), (c) 10-m u wind ($m\,s^{-1}$), and (d) 10-m v wind ($m\,s^{-1}$).

all of the ensemble spread diagnostics. This is largely because each model error experiment successfully avoided the various differing deficiencies of the SMSP ensembles at any particular time, such as the low FSS of ARW-SP for later hours in heavy precipitation, or the low ROC area in reflectivity of NMMB. While the results are consistent with other CAE model error studies, this is the first to compare two or more techniques to address model error uncertainty for CAE forecasting in the context of already-sampled multiscale IC uncertainty. Therefore, it can be reasonably inferred that the sampling methods used here are primarily sampling the deficiencies in model error uncertainty rather than the full error coming from both the IC and model errors.

- The MM experiment had the highest FSS for heavy precipitation thresholds, as well as the most amount of spread added consistently throughout the entire forecast. Subjectively, the MM avoided single-model skill drop-offs and performed better in cases where each single model had errors of opposite signs or in differing locations. Additionally, bias correction in precipitation and reflectivity slightly improved the verification scores of MM relative to other experiments. This is consistent with Johnson and Wang (2012), whose bias-calibrated MM ensemble showed more times with greater skill compared to SMSP ensembles. However, there was a cost in forecast sharpness, and undesirable ensemble clustering was found in rank histogram distributions of spread in all variables. Additionally, the MM showed minimal benefit when one single-model experiment had consistently superior performance to another, for instance where ARW-SP had superior ROC area and ensemble mean error of upper-level fields above 500 hPa.
- The ARW-MP experiment significantly improved upon FSS precipitation verification over ARW-SP for lighter precipitation thresholds and the final 9–12 h of heavier precipitation thresholds. This was accompanied with large increases in forecast discrimination and reliability of precipitation, as well as forecast spread of near-surface fields. However, the increase in spread was limited to lower-level fields below 850 hPa and took a ''spinup'' period of at least 6 h for noticeable increases to appear. Bias correction reduced the differences between ARW-MP and the SMSP experiments, particularly in terms of ROC area for composite reflectivity. However, many of the differences remained statistically significant—particularly for lighter thresholds (2.54 mm in 1-h QPF; 30 dBZ in composite reflectivity). This is somewhat in contrast to Loken et al. (2019), who showed little statistically significant differences among bias-corrected single- and mixed-physics ensembles for 6-h QPF verifications.
- Adding SKEB on top of ARW-MP showed small but significant improvements to precipitation verification in the latter half of the forecast for heavy (6.35 and 12.7 mm) precipitation. This is consistent with past results (e.g., Duda et al. 2016; Jankov et al. 2019), but noteworthy because this study already incorporates a full multiscale sampling of IC uncertainty from the

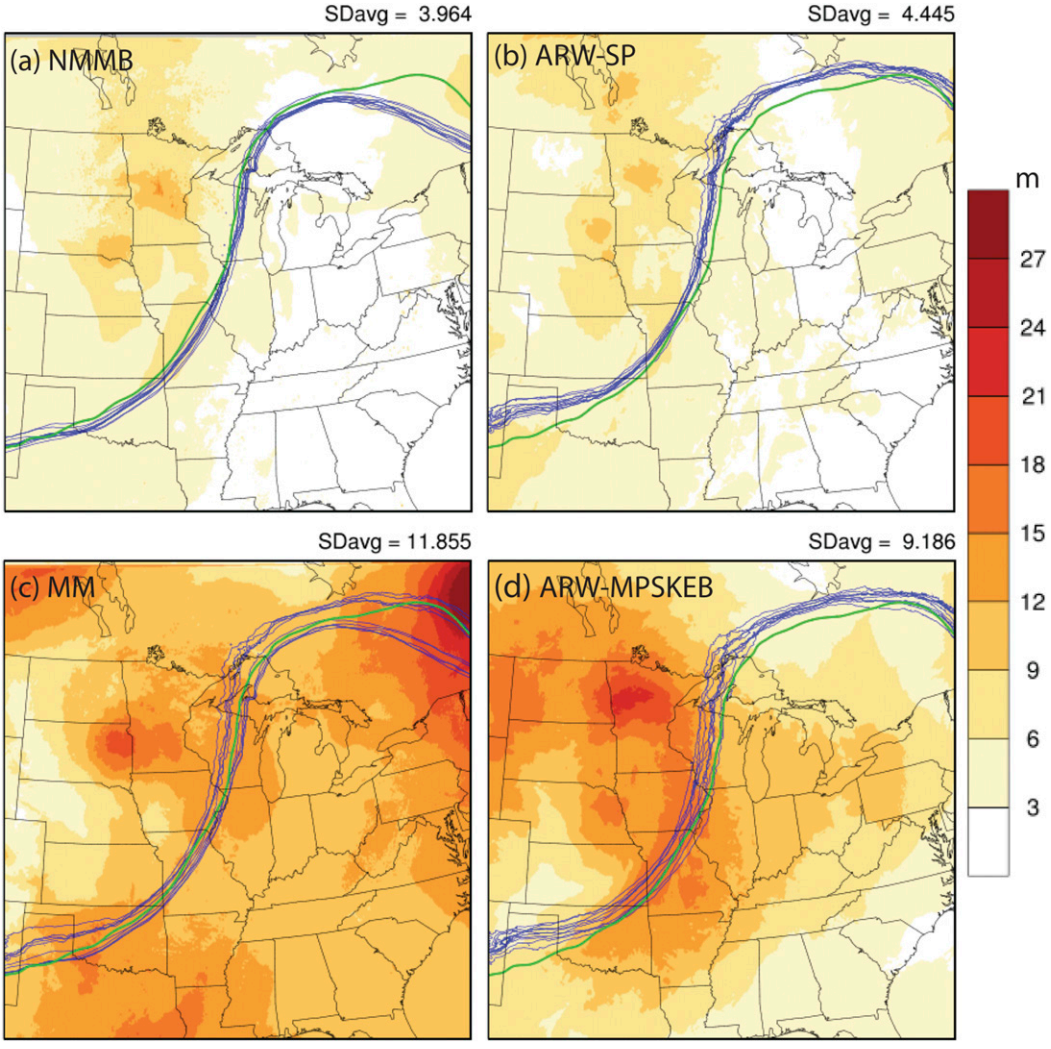

FIG. 17. Ensemble standard deviation (color fill) of 500-hPa geopotential height (m) valid at 1700 UTC 16 May 2015 (forecast hour 18). Individual member 5760-m height contours (blue) are shown with the analyzed RAP 5760-m contour (green). Values on the upper right of each panel indicate standard deviation averaged over each plotting domain.

EnVar DA system. While there was an accompanying drop in forecast sharpness; however, there was an increase in the reliability of midprobabilities and overall spread of precipitation systems. Among other fields, there was practically no change in error but substantial increases in ensemble spread over time, particularly for wind and geopotential height fields.

The comparison of MM to ARW-MP and ARW-MPSKEB led to some mixed results and depends upon what aspect of the forecast is examined. ARW-MP and ARW-MPSKEB compared favorably to MM with lighter precipitation thresholds, particularly for weakly forced cases that are more sensitive to physics uncertainties. However, MM had significantly better skill in heavy precipitation across all forecast times, with noticeably increased spread of precipitation systems throughout the forecast that was only matched by ARW-MPSKEB by the end of the forecast. Similarly in upper-level fields, MM has consistently the highest amount of spread at all forecast times, including early lead times, while ARW-MPSKEB only matched or exceeded this spread within the last 3–6 h of the forecast. In terms of error, ARW-MPSKEB was slightly lower than MM across most upper-level variables, and most noticeably lower for 2-m dewpoint due to the removal of moist biases present in all other experiments. Additionally, clustering of ensemble spread was found across all variables in MM–an undesirable statistical quality of the ensemble distributions. While we cannot rule out clustering as a significant issue in ARW-MP, the increased variety of physics options led to a smoother

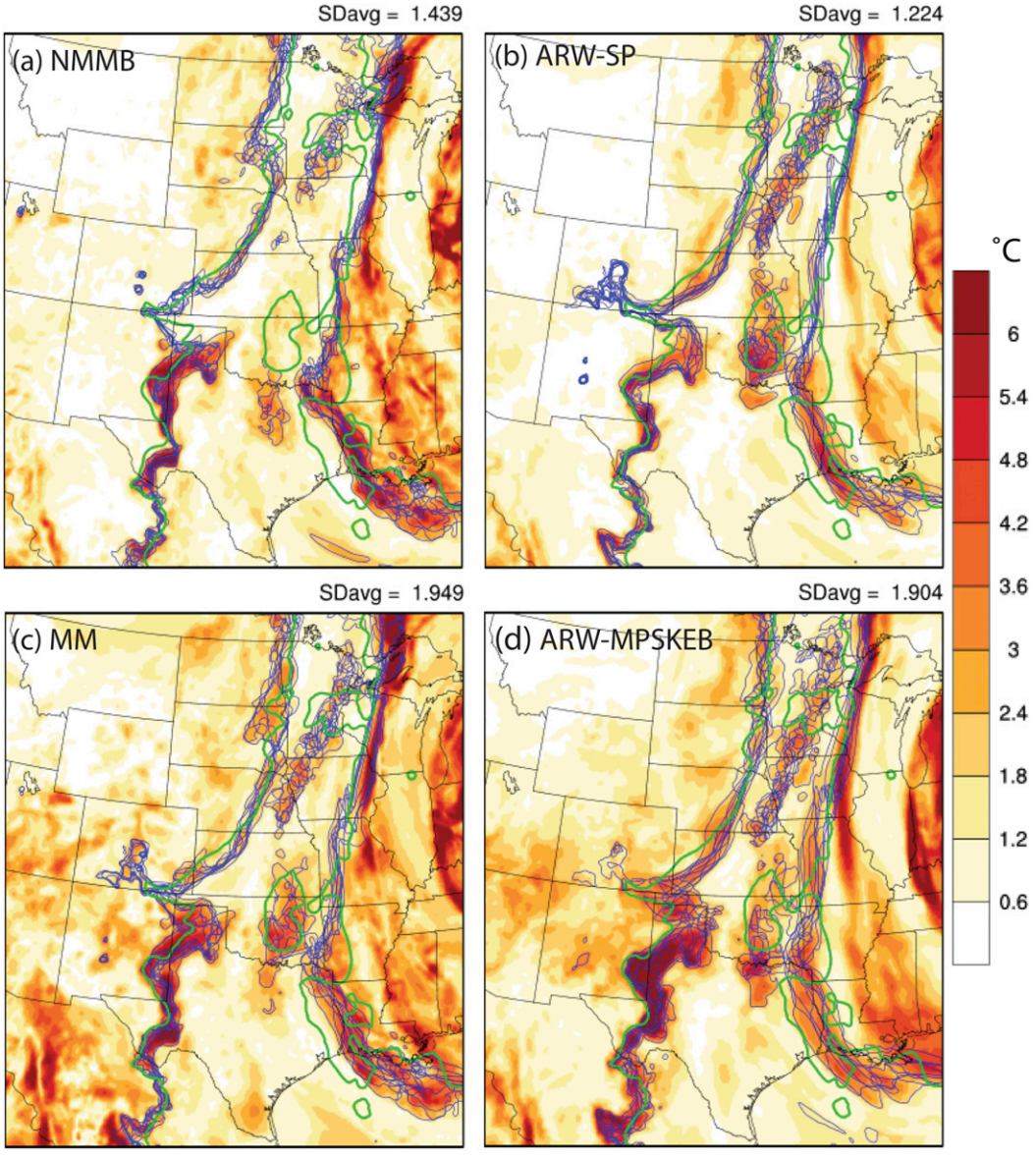

FIG. 18. As in Fig. 17, but for 850-hPa dewpoint (with 6°C spaghetti contour) valid 1500 UTC 24 May 2016 (forecast hour 16).

distribution when considered over the whole range of cases as compared to MM.

The ARW-MPSKEB experiment had the overall best spread distribution by the end of the forecast, with SKEB and multiphysics approaches complimenting each other. SKEB increased the spread of upper-level fields, particularly in wind, while the multiphysics increased the spread of near-surface fields, particularly for thermodynamic variables. However, the main drawback in declaring ARW-MPSKEB superior to MM is that this ensemble spread has a ''spinup'' period, for both the multiphysics and SKEB effects. When looking at forecast times less than about 12 h, the MM is still the best approach to immediately address model error and increase ensemble spread despite the noted clustering effects. These results suggest that single-model approaches to address model error may feasibly compete with a multimodel approach; however, more research on model error techniques is needed to find a configuration that outperforms a multimodel approach over all forecast aspects, particularly within earlier hours of the forecast.

This study did not attempt to evaluate every possible model error technique. Many different kinds of stochastic physics techniques are being studied across the literature, with just one chosen for testing in this work.

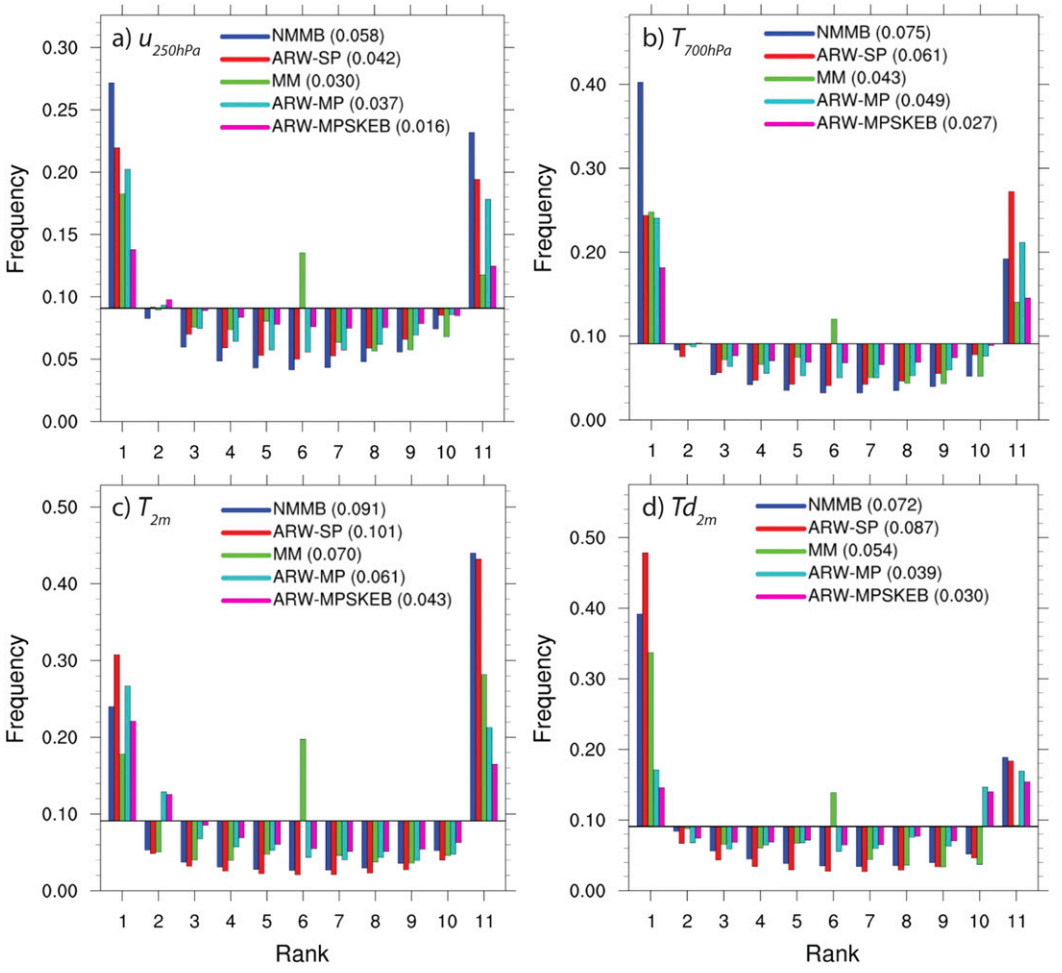

FIG. 19. Rank histogram plots in (a) 250-hPa *u* wind, (b) 700-hPa temperature, (c) 2-m temperature, and (d) 2-m dewpoint, shown relative to the "ideal" ensemble spread having uniform distribution (horizontal line at frequency 0.091 for 11 bins). Numbers in parentheses of panel legends indicate the median of absolute differences in bin frequency for each experiment relative to the uniform distribution frequency of 0.091.

Additionally, different combinations of dynamical models or different combinations of multiphysics or multiparameter schemes may lead to significantly different results than those shown here. However, this study applied a practical approach to the problem, by using readily available model dynamical cores and multiphysics configurations already implemented in WRF and known to test well for CAE due to experiences within the HWT SFE. In the future, additional combinations of stochastic and/or multiphysics techniques can be studied given the results here and in other works (e.g., Berner at el. 2015; Duda et al. 2016; Jankov et al. 2019) suggest that optimally addressing model error may rely on combinations of techniques to address uncertainties. An optimal combination for single-model approach that is competitive with a multimodel approach is ideally preferred, since multimodel ensembles rely on costly development and maintenance of multiple dynamical cores rather than just one. Furthermore, since multiphysics ensembles require development and updates to a variety of physics schemes, a practical goal within CAE forecasting may be to address model error uncertainty using solely stochastic physics approaches.

Acknowledgments. This research was supported by NOAA Grants NA16OAR4590236 and NA15OAR4590193. The authors would like to acknowledge and thank Dr. Judith Berner for help troubleshooting various issues related to the SKEB scheme. The authors also thank the three anonymous peer reviewers for their useful comments that have helped to improve the manuscript overall. All DA and forecast experiments were performed on the Stampede 2 supercomputer at the Texas Advanced Supercomputer Center as part of

NSF's Extreme Science and Engineering Discovery Environment (XSEDE) program. Computational verification was conducted on resources from the University of Oklahoma (OU) Supercomputing Center for Education and Research (OSCER).

REFERENCES

Aligo, E. A., B. Ferrier, and J. R. Carley, 2018: Modified NAM microphysics for forecasts of deep convective storms. *Mon. Wea. Rev.*, **146**, 4115–4153, https://doi.org/10.1175/MWR-D-17-0277.1.

Benjamin, S. G., G. A. Grell, J. M. Brown, T. G. Smirnova, and R. Bleck, 2004: Mesoscale weather prediction with the RUC hybrid isentropic–terrain-following coordinate model. *Mon. Wea. Rev.*, **132**, 473–494, https://doi.org/10.1175/1520-0493(2004)132<0473:MWPWTR>2.0.CO;2.

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, https://doi.org/10.1175/2010MWR3595.1.

——, K. R. Fossell, S. Y. Ha, J. P. Hacker, and C. Snyder, 2015: Increasing the skill of probabilistic forecasts: Understanding performance improvements from model-error representations. *Mon. Wea. Rev.*, **143**, 1295–1320, https://doi.org/10.1175/MWR-D-14-00091.1.

Candille, G., 2009: The multiensemble approach: The NAEFS example. *Mon. Wea. Rev.*, **137**, 1655–1665, https://doi.org/10.1175/2008MWR2682.1.

Clark, A. J., W. A. Gallus Jr., M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-permitting and large convection-parameterizing ensembles. *Wea. Forecasting*, **24**, 1121–1140, https://doi.org/10.1175/2009WAF2222222.1.

——, ——, ——, and ——, 2010: Growth of spread in convection-allowing and convection-parameterizing ensembles. *Wea. Forecasting*, **25**, 594–612, https://doi.org/10.1175/2009WAF2222318.1.

——, and Coauthors, 2011: Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Mon. Wea. Rev.*, **139**, 1410–1418, https://doi.org/10.1175/2010MWR3624.1.

——, and Coauthors, 2012: An overview of the 2010 Hazardous Weather Testbed experimental forecast program spring experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74, https://doi.org/10.1175/BAMS-D-11-00040.1.

——, and Coauthors, 2017: Spring forecasting experiment 2017 conducted by the experimental forecast program of the NOAA/Hazardous Weather Testbed: Program overview and operations plan. NOAA, 34 pp., https://hwt.nssl.noaa.gov/Spring_2017/HWT_SFE2017_operations_plan_FINAL.pdf.

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

Degelia, S. K., X. Wang, D. J. Stensrud, and A. Johnson, 2018: Understanding the impact of radar and in situ observations on the prediction of a nocturnal convection initiation event on 25 June 2013 using an ensemble-based multiscale data assimilation system. *Mon. Wea. Rev.*, **146**, 1837–1859, https://doi.org/10.1175/MWR-D-17-0128.1.

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.

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

——, ——, ——, ——, and J. Berner, 2016: Impact of a stochastic kinetic energy backscatter scheme on warm season convection-allowing ensemble forecasts. *Mon. Wea. Rev.*, **144**, 1887–1908, https://doi.org/10.1175/MWR-D-15-0092.1.

——, ——, and M. Xue, 2017: Sensitivity of convection-allowing forecasts to land surface model perturbations and implications for ensemble design. *Mon. Wea. Rev.*, **145**, 2001–2025, https://doi.org/10.1175/MWR-D-16-0349.1.

——, ——, Y. Wang, and J. Carley, 2019: Comparing the assimilation of radar reflectivity using the direct GSI based ensemble-variational (EnVar) and indirect cloud analysis methods in convection-allowing forecasts over the continental United States. *Mon. Wea. Rev.*, **147**, 1655–1678, https://doi.org/10.1175/MWR-D-18-0171.1.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, https://doi.org/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

Eckel, F. A., and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350, https://doi.org/10.1175/WAF843.1.

Gallo, B. T., and Coauthors, 2018: Spring forecasting experiment 2018 conducted by the experimental forecast program of the NOAA/Hazardous Weather Testbed: Program overview and operations plan. NOAA, 46 pp., https://hwt.nssl.noaa.gov/sfe/2018/docs/HWT_SFE2018_operations_plan.pdf.

Gallus, W. A., and J. F. Bresch, 2006: Comparison of impacts of WRF dynamic core, physics package, and initial conditions on warm season rainfall forecasts. *Mon. Wea. Rev.*, **134**, 2632–2641, https://doi.org/10.1175/MWR3198.1.

Gaspari, G., and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, **125**, 723–757, https://doi.org/10.1002/qj.49712555417.

Gebhardt, C., S. E. Theis, M. Paulat, and Z. Ben Bouallègue, 2011: Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmos. Res.*, **100**, 168–177, https://doi.org/10.1016/j.atmosres.2010.12.008.

Hacker, J. P., C. Snyder, S. Y. Ha, and M. Pocernich, 2011: Linear and non-linear response to parameter variations in a mesoscale model. *Tellus*, **63A**, 429–444, https://doi.org/10.1111/j.1600-0870.2010.00505.x.

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

——, 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

——, J. S. Whitaker, D. T. Kleist, M. Fiorino, and S. G. Benjamin, 2011: Predictions of 2010's tropical cyclones using the GFS and ensemble-based data assimilation methods. *Mon. Wea. Rev.*, **139**, 3243–3247, https://doi.org/10.1175/MWR-D-11-00079.1.

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

Hu, M., and Coauthors, 2018: Grid-point statistical interpolation (GSI) user's guide version 3.7. Developmental Testbed Center, 147 pp., http://www.dtcenter.org/com-GSI/users/docs/index.php.

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

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

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

Johnson, A., and X. Wang, 2012: Verification and calibration of neighborhood and object-based probabilistic precipitation forecasts from a multimodel convection-allowing ensemble. *Mon. Wea. Rev.*, **140**, 3054–3077, https://doi.org/10.1175/MWR-D-11-00356.1.

——, and ——, 2017: Design and implementation of a GSI-based convection-allowing ensemble data assimilation and forecast system for the PECAN field experiment. Part I: Optimal configurations for nocturnal convection prediction using retrospective cases. *Wea. Forecasting*, **32**, 289–315, https://doi.org/10.1175/WAF-D-16-0102.1.

——, ——, M. Xue, and F. Kong, 2011: Hierarchical cluster analysis of a convection-allowing ensemble during the Hazardous Weather Testbed 2009 spring experiment. Part II: Ensemble clustering over the whole experiment period. *Mon. Wea. Rev.*, **139**, 3694–3710, https://doi.org/10.1175/MWR-D-11-00016.1.

——, ——, J. R. Carley, L. J. Wicker, and C. Karstens, 2015: A comparison of multiscale GSI-based EnKF and 3dvar data assimilation using radar and conventional observations for midlatitude convective-scale precipitation forecasts. *Mon. Wea. Rev.*, **143**, 3087–3108, https://doi.org/10.1175/MWR-D-14-00345.1.

——, ——, and S. Degelia, 2017: Design and implementation of a GSI-based convection-allowing ensemble-based data assimilation and forecast system for the PECAN field experiment. Part II: Overview and evaluation of a real-time system. *Wea. Forecasting*, **32**, 1227–1251, https://doi.org/10.1175/WAF-D-16-0201.1.

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

Loken, E. D., A. J. Clark, M. Xue, and F. Kong, 2017: Comparison of next-day probabilistic severe weather forecasts from coarse- and fine-resolution CAMs and a convection-allowing ensemble. *Wea. Forecasting*, **32**, 1403–1421, https://doi.org/10.1175/WAF-D-16-0200.1.

——, ——, ——, and ——, 2019: Spread and skill in mixed- and single-physics convection-allowing ensembles. *Wea. Forecasting*, **34**, 305–330, https://doi.org/10.1175/WAF-D-18-0078.1.

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 the assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.

Mason, S. J., and N. E. Graham, 2002: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quart. J. Roy. Meteor. Soc.*, **128**, 2145–2166, https://doi.org/10.1256/003590002320603584.

Melhauser, C., F. Zhang, Y. Weng, Y. Jin, H. Jin, and Q. Zhao, 2017: A multiple-model convection-permitting ensemble examination of the probabilistic prediction of tropical cyclones: Hurricanes Sandy (2012) and Edouard (2014). *Wea. Forecasting*, **32**, 665–688, https://doi.org/10.1175/WAF-D-16-0082.1.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119, https://doi.org/10.1002/qj.49712252905.

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

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

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

Palmer, T. N., 2002: The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quart. J. Roy. Meteor. Soc.*, **128**, 747–774, https://doi.org/10.1256/0035900021643593.

Potvin, C., and Coauthors, 2019: Systematic comparison of convection-allowing models during the 2017 NOAA HWT Spring Forecasting Experiment. *Wea. Forecasting*, **34**, 1395–1416, https://doi.org/10.1175/WAF-D-19-0056.1.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667, https://doi.org/10.1002/qj.49712656313.

Romine, G. S., C. S. Schwartz, J. Berner, K. R. Fossell, C. Snyder, J. L. Anderson, and M. L. Weisman, 2014: Representing forecast error in a convection-permitting ensemble system. *Mon. Wea. Rev.*, **142**, 4519–4541, https://doi.org/10.1175/MWR-D-14-00100.1.

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

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

——, G. S. Romine, K. R. Smith, and M. L. Weisman, 2014: Characterizing and optimizing precipitation forecasts from a convection-permitting ensemble initialized by a mesoscale

ensemble Kalman filter. *Wea. Forecasting*, **29**, 1295–1318, https://doi.org/10.1175/WAF-D-13-00145.1.

——, ——, K. R. Fossell, R. A. Sobash, and M. L. Weisman, 2017: Toward 1-km ensemble forecasts over large domains. *Mon. Wea. Rev.*, **145**, 2943–2969, https://doi.org/10.1175/MWR-D-16-0410.1.

——, ——, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2019: NCAR's real-time convection-allowing ensemble project. *Bull. Amer. Meteor. Soc.*, **100**, 321–343, https://doi.org/10.1175/BAMS-D-17-0297.1.

Smith, T. M., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) severe weather and aviation products: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 1617–1630, https://doi.org/10.1175/BAMS-D-14-00173.1.

Snook, N., M. Xue, and Y. Jung, 2015: Multiscale EnKF assimilation of radar and conventional observations and ensemble forecasting for a tornadic mesoscale convective system. *Mon. Wea. Rev.*, **143**, 1035–1057, https://doi.org/10.1175/MWR-D-13-00262.1.

Stensrud, D. J., and M. S. Wandishin, 2000: The correspondence ratio in forecast evaluation. *Wea. Forecasting*, **15**, 593–602, https://doi.org/10.1175/1520-0434(2000)015<0593:TCRIFE>2.0.CO;2.

——, 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, https://doi.org/10.1175/1520-0493(2000)128<2077:UICAMP>2.0.CO;2.

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

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.

——, P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, https://doi.org/10.1175/2008MWR2387.1.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330, https://doi.org/10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2.

——, and ——, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297–3319, https://doi.org/10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2.

Vié, B., O. Nuissier, and V. Ducrocq, 2011: Cloud-resolving ensemble simulations of Mediterranean heavy precipitating events: Uncertainty on initial conditions and lateral boundary conditions. *Mon. Wea. Rev.*, **139**, 403–423, https://doi.org/10.1175/2010MWR3487.1.

Voisin, N., J. C. Schaake, and D. P. Lettenmaier, 2010: Calibration and downscaling methods for quantitative ensemble precipitation forecasts. *Wea. Forecasting*, **25**, 1603–1627, https://doi.org/10.1175/2010WAF2222367.1.

Wandishin, M. S., S. L. Mullen, and H. E. Brooks, 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev.*, **129**, 729–747, https://doi.org/10.1175/1520-0493(2001)129<0729:EOASRM>2.0.CO;2.

Wang, X., 2010: Incorporating ensemble covariance in the gridpoint statistical interpolation variational minimization: A mathematical framework. *Mon. Wea. Rev.*, **138**, 2990–2995, https://doi.org/10.1175/2010MWR3245.1.

——, and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.*, **60**, 1140–1158, https://doi.org/10.1175/1520-0469(2003)060<1140:ACOBAE>2.0.CO;2.

——, and T. Lei, 2014: GSI-based four-dimensional ensemble–variational (4DEnsVar) data assimilation: Formulation and single-resolution experiments with real data for NCEP Global Forecast System. *Mon. Wea. Rev.*, **142**, 3303–3325, https://doi.org/10.1175/MWR-D-13-00303.1.

——, C. H. Bishop, and S. J. Julier, 2004: Which is better, an ensemble of positive–negative pairs or a centered spherical simplex ensemble? *Mon. Wea. Rev.*, **132**, 1590–1605, https://doi.org/10.1175/1520-0493(2004)132<1590:WIBAEO>2.0.CO;2.

——, D. Parrish, D. Kleist, and J. Whitaker, 2013: GSI 3DVar-based ensemble–variational hybrid data assimilation for NCEP Global Forecast System: Single-resolution experiments. *Mon. Wea. Rev.*, **141**, 4098–4117, https://doi.org/10.1175/MWR-D-12-00141.1.

Wang, Y., and X. Wang, 2017: Direct assimilation of radar reflectivity without tangent linear and adjoint of the nonlinear observation operator in the GSI-based envar system: Methodology and experiment with the 8 May 2003 Oklahoma City tornadic supercell. *Mon. Wea. Rev.*, **145**, 1447–1471, https://doi.org/10.1175/MWR-D-16-0231.1.

Whitaker, J. S., and T. M. Hamill, 2012: Evaluating methods to account for system errors in ensemble data assimilation. *Mon. Wea. Rev.*, **140**, 3078–3089, https://doi.org/10.1175/MWR-D-11-00276.1.

Xue, M., Y. Jung, and G. Zhang, 2010: State estimation of convective storms with a two-moment microphysics scheme and an ensemble Kalman filter: Experiments with simulated radar data. *Quart. J. Roy. Meteor. Soc.*, **136**, 685–700, https://doi.org/10.1002/QJ.593.

Yussouf, N., and D. J. Stensrud, 2012: Comparison of single-parameter and multiparameter ensembles for assimilation of radar observations using the ensemble Kalman filter. *Mon. Wea. Rev.*, **140**, 562–586, https://doi.org/10.1175/MWR-D-10-05074.1.

——, E. R. Mansell, L. J. Wicker, D. M. Wheatley, and D. J. Stensrud, 2013: The ensemble Kalman filter analyses and forecasts of the 8 May 2003 Oklahoma City tornadic supercell storm using single- and double-moment microphysics schemes. *Mon. Wea. Rev.*, **141**, 3388–3412, https://doi.org/10.1175/MWR-D-12-00237.1.

——, D. C. Dowell, L. J. Wicker, K. H. Knopfmeier, and D. M. Wheatley, 2015: Storm-scale data assimilation and ensemble forecasts for the 27 April 2011 severe weather outbreak in Alabama. *Mon. Wea. Rev.*, **143**, 3044–3066, https://doi.org/10.1175/MWR-D-14-00268.1.

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