Model Configuration versus Driving Model: Influences on Next-Day Regional Convection-Allowing Model Forecasts during a Real-Time Experiment

BRETT ROBERTS,[a,b,c] ADAM J. CLARK,[b,d] ISRAEL L. JIRAK,[c] BURKELY T. GALLO,[a,c] CAROLINE BAIN,[e] DAVID L. A. FLACK,[e] JAMES WARNER,[e] CRAIG S. SCHWARTZ,[f] AND LARISSA J. REAMES[a,b]

[a] *Cooperative Institute for Severe and High-Impact Weather Research and Operations, University of Oklahoma, Norman, Oklahoma*
[b] *NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma*
[c] *NOAA/NCEP/Storm Prediction Center, Norman, Oklahoma*
[d] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*
[e] *Met Office, Exeter, United Kingdom*
[f] *National Center for Atmospheric Research, Boulder, Colorado*

(Manuscript received 22 December 2021, in final form 21 October 2022)

ABSTRACT: As part of NOAA's Hazardous Weather Testbed Spring Forecasting Experiment (SFE) in 2020, an international collaboration yielded a set of real-time convection-allowing model (CAM) forecasts over the contiguous United States in which the model configurations and initial/boundary conditions were varied in a controlled manner. Three model configurations were employed, among which the Finite Volume Cubed-Sphere (FV3), Unified Model (UM), and Advanced Research version of the Weather Research and Forecasting (WRF-ARW) Model dynamical cores were represented. Two runs were produced for each configuration: one driven by NOAA's Global Forecast System for initial and boundary conditions, and the other driven by the Met Office's operational global UM. For 32 cases during SFE2020, these runs were initialized at 0000 UTC and integrated for 36 h. Objective verification of model fields relevant to convective forecasting illuminates differences in the influence of configuration versus driving model pertinent to the ongoing problem of optimizing spread and skill in CAM ensembles. The UM and WRF configurations tend to outperform FV3 for forecasts of precipitation, thermodynamics, and simulated radar reflectivity; using a driving model with the native CAM core also tends to produce better skill in aggregate. Reflectivity and thermodynamic forecasts were found to cluster more by configuration than by driving model at lead times greater than 18 h. The two UM configuration experiments had notably similar solutions that, despite competitive aggregate skill, had large errors in the diurnal convective cycle.

KEYWORDS: Forecast verification/skill; Numerical weather prediction/forecasting; Model comparison; Model evaluation/performance

1. Introduction

Each spring since the mid-2000s the NOAA Hazardous Weather Testbed has hosted its annual Spring Forecasting Experiment (SFE; Kain et al. 2003; Clark et al. 2012; Gallo et al. 2017; Clark et al. 2020), which runs for five weeks and focuses in part on evaluating the performance of state-of-the-art convection-allowing models (CAMs) in forecasting severe convective storms. A wide array of government and academic units contribute daily real-time operational and experimental CAMs to the SFE, and since 2016, most of these CAMs have been incorporated into the Community Leveraged Unified Ensemble (CLUE; Clark et al. 2018) framework to enable systematic subjective and objective comparisons. Such comparisons reveal the relative strengths and weaknesses associated with various CAM configuration choices and facilitate productive dialog between model developers and operational users.

As described in [Roberts et al. (2020, hereafter R20)], evaluations in recent SFEs have consistently noted that most CAM ensembles are quite underdispersive with respect to forecasts of convective storms (e.g., their coverage, location, and intensity). A fundamental issue regarding the design of CAM ensembles in recent years has been the utility of diverse model configurations, initial conditions (ICs), boundary conditions (BCs), and other approaches (e.g., time-lagging and stochastically perturbed parameterizations) for increasing spread and thereby achieving a more appropriate spread–skill relationship. In 2017, the National Centers for Environmental Prediction (NCEP) implemented the High Resolution Ensemble Forecast (HREF; Roberts et al. 2019) system as its first operational CAM ensemble. The HREF is an ensemble of opportunity composed of highly diverse members with different dynamical cores, physics parameterizations, ICs/BCs, and time lagging. Although this diversity has yielded beneficial spread in forecasts of convective storms (R20), HREF's limited membership size does not include most possible combinations of its constituent ICs/BCs and model configurations, which in turn limits our ability to discern the relative influences of those attributes on ensemble spread. Formal research over the past decade has explored the problem of optimizing CAM ensemble design (e.g., Romine et al. 2014; Gasperoni et al. 2020; Johnson and Wang 2020) and the benefits of diverse CAM model configurations (e.g., Johnson et al. 2011; Clark 2019; Loken et al. 2019), providing valuable insights to guide developers. However, the CAM ensemble members in

Corresponding author: Brett Roberts, brett.roberts@alumni.ou.edu

DOI: 10.1175/WAF-D-21-0211.1

TABLE 1. Configurations for the six NWP CAM experiments in this study. PBL schemes used include the Mellor–Yamada–Nakanishi–Niino (MYNN; Nakanishi and Niino 2004), Mellor–Yamada–Janjić (MYJ; Janjić 1994), and Smagorinsky (Smagorinsky 1963; Lilly 1992) [blended, depending on the ratio of grid spacing to local PBL height, with a conventional 1D nonlocal scheme as described in Boutle et al. (2014)] formulations. Microphysics schemes include the Thompson (Thompson et al. 2008) and um-sm [based on modifications to the warm rain scheme of Wilson and Ballard (1999) and tuned for the midlatitudes]. LSMs include the Noah (Chen and Dudhia 2001) and JULES (Best et al. 2011; Clark et al. 2011) schemes. Complete details of the um experiment model configurations are consistent with the RAL2-M package used in Steptoe et al. (2021), which itself is a minor update to the RAL1-M configuration described more exhaustively in Bush et al. (2020). UM-G refers to the Met Office's operational global run of the UM.

Expt	Core	PBL	Microphysics	LSM	ICs	LBCs	Soil temperature	Soil moisture	*dx* (km)
fv3-GFS	FV3	MYNN	Thompson	Noah	GFS	GFS	GFS	GFS	3.0
fv3-UM	FV3	MYNN	Thompson	Noah	UM-G	UM-G	UM-G	UM-G	3.0
um-GFS	UM	Smag-blend	um-sm	JULES	GFS	GFS	GFS	UM-G	2.2
um-UM	UM	Smag-blend	um-sm	JULES	UM-G	UM-G	UM-G	UM-G	2.2
wrf-GFS	WRF-ARW	MYJ	Thompson	Noah	GFS	GFS	GFS	GFS	3.0
wrf-UM	WRF-ARW	MYJ	Thompson	Noah	UM-G	UM-G	UM-G	UM-G	3.0

these studies that represented IC and/or BC uncertainty used perturbations based on a single model (e.g., from an ensemble Kalman filter), rather than ICs and BCs from multiple distinct parent models like HREF. In this study, we analyze a suite of experiments with ICs and BCs from two independent analysis systems, representing the type of IC/BC diversity typically found in ensembles of opportunity.

The aforementioned barriers to discerning impacts from membership choices in the context of ensembles of opportunity like HREF motivated an internationally coordinated experiment for SFE2020 (Clark et al. 2021) which we describe in this paper. Our study leverages a more controlled suite of real-time CAM runs, which notably includes the first convection-allowing runs (to the authors' knowledge) of the Met Office Unified Model (UM; Cullen 1993) to be driven by ICs and BCs from an operational American numerical weather prediction (NWP) model. Although several facets of our experimental design are novel for the systems considered herein, the European NWP research community has explored the impact of diverse driving models for CAMs during the past decade, particularly in the context of Deutscher Wetterdienst's convective-scale Consortium for Small Scale Modeling (COSMO-DE-EPS; Gebhardt et al. 2011) system. COSMO-DE-EPS used four distinct global driving models, which were combined with perturbed physics parameters to represent IC, BC, and model uncertainty—although there was no diversity with respect to dynamical core or the parameterization schemes themselves, as there is in HREF and in this study. This "multi-analysis" ensemble design is somewhat analogous to HREF's, but otherwise uncommon in American NWP. Several studies analyzing COSMO-DE-EPS forecasts (Keil et al. 2014; Kühnlein et al. 2014; Marsigli et al. 2014), along with another study evaluating CAM ensemble forecasts from two separate driving models (Porson et al. 2019), found value in the spread added when disparate driving models were employed. Furthermore, those studies which quantified CAM ensemble spread arising from representing IC uncertainty versus other sources of uncertainty tended to find a meaningful contribution from ICs out to 6–18 h into the forecast (Kühnlein et al. 2014; Porson et al. 2019). Building on these previous findings, a key research question in the present study is: do the driving models dominate CAM forecast differences at early lead times? And, if so, at what lead time is the influence of the driving models typically superseded by that of the model configurations?

The primary goal of the present study is to investigate the impacts of model configuration (e.g., dynamical core and parameterizations) versus driving model (the global NWP model providing ICs/BCs) on CAM solution spread and skill at next-day lead times for springtime in the continental United States, where severe local storms are more commonplace than in regions covered by the aforementioned European studies. In particular, we are interested in partitioning the relative importance of model configuration versus driving model in influencing CAM forecasts and how this partitioning changes with lead time. These research questions have implications for CAM ensemble design (including ensembles of opportunity), particularly the importance of optimizing the sampling of model uncertainty versus IC/BC uncertainty at different time scales. Our experimental setup also invites a secondary goal of identifying biases and other performance characteristics of the *specific* model configurations and driving models we are testing through traditional verification metrics. For both goals, our focus is on model fields relevant to forecasts of deep moist convection—and severe local storms, in particular. The paper is organized as follows: Section 2 describes our datasets and verification techniques; section 3 details our results; and section 4 provides broad conclusions and recommendations for how future work might build on our findings.

2. Methods

a. Datasets

Our CAM experiments cover all combinations generated by three model configurations and two driving models, yielding six total experiments; full details of each experiment are presented in Table 1. Hereafter we use "driving model" to mean the global NWP model from which an experiment inherits initial and boundary conditions, including both the lateral and lower boundaries. Furthermore, we name each experiment using the convention (configuration)-(DRIVING MODEL); e.g., fv3-GFS is the experiment with an FV3-based configuration driven by ICs and BCs from the GFS. Although we do not consider our set of experiments an ensemble, we nonetheless analyze the degree of similarity between experiments as a proxy for spread in a hypothetical ensemble of opportunity with similar membership.

Each model configuration employs a unique dynamical core: the Advanced Research version of the Weather Research and Forecasting (WRF-ARW) Model (Skamarock et al. 2008, 2021); the UM; and the Finite Volume Cubed-Sphere (FV3; Putman and Lin 2007). However, additional important configuration differences exist, and the performance attributes of a model configuration in the present study should not be interpreted as necessarily intrinsic to its underlying dynamical core. Crucially, the planetary boundary layer (PBL) parameterization scheme differs between all three configurations, while the microphysics scheme and land surface model (LSM) differ between some configurations. CAM forecasts of convection and precipitation are known to be sensitive to such parameterization choices (e.g., Schwartz et al. 2010; Johnson et al. 2011; Coniglio et al. 2013; Duda et al. 2017; Loken et al. 2019). Furthermore, the um model configuration uses 2.2-km horizontal grid spacing, while the fv3 and wrf configurations use 3 km. The details of each model configuration were largely inherited from ongoing work and existing systems at each contributing research institution,[1] which were used opportunistically for this study. Our analyses herein are restricted to the set of dates for which this international collaboration for SFE2020 was planned, and also to the set of model forecast fields which could be postprocessed and transmitted by all participating institutions.

The two driving models used for each configuration are NCEP's Global Forecast System (GFS) and the Met Office's global run of the UM (UM-Global). Each driving model represents the flagship operational global deterministic NWP from its respective modeling center, and each uses its own data assimilation (DA) system, so our experiments are inheriting unperturbed analyses from independent sources. In 2020, the operational GFS used the FV3 dynamical core. As such, fv3-GFS and um-UM are the two experiments in which the CAM configuration is driven by a coarse global run with the native core. All experiments are "cold start" CAMs, meaning they make no attempt to represent convective storms or other fine-scale features not resolvable on the coarse driving model grid in their ICs, so a spinup time on the order of 6–12 h can be expected (e.g., Kain et al. 2010; Raynaud and Bouttier 2016; Wong and Skamarock 2016). There is also the potential for so-called model shock when ICs lie distant from the dynamical model's attractor[2] (Judd et al. 2008; Klocke and Rodwell 2014), a situation made more likely when a nonnative driving model is employed; the result may be a period of drift toward the model attractor during the beginning of the forecast. To the extent that the atmospheric ICs in some of our experiments produce such a shock, it is tolerated as a cost that is widely accepted for limited area models in operational NWP (e.g., HREF).

Two minor caveats in the experimental setup should be noted. First, owing to real-time data flow considerations, the um-GFS was initialized using atmospheric data with coarser vertical resolution than the um-UM. Second, regarding the soil state in the LSM for each experiment: Flack et al. (2021) describe some limitations posed on the present study when attempting to inherit soil states from nonnative driving models. In short, it was necessary to use UM-Global soil moisture conditions in the um-GFS experiment due to differences in the formulation of soil moisture parameters by the JULES and Noah LSMs used in the UM-Global and GFS, respectively. This caveat means that our experimental setup cannot be interpreted as systematically varying driving models in the strictest possible sense, although only one of the six experiments is affected in this way. Schwartz et al. (2022) found a small impact from the initial soil state (relative to atmospheric ICs) on next-day CAM forecasts (cf. their Fig. A2), increasing our confidence that this caveat should not impose major qualitative limitations on our conclusions herein. This caveat also does not apply to the soil *temperatures*, which are inherited from the appropriate driving model in all cases. The potential for some degree of model shock also exists in the LSM for nonnative driving models, and unlike atmospheric ICs, the drift resulting from such shock may occur gradually and extend throughout our forecast cycle.

All experiments were initialized daily at 0000 UTC for 32 cases in spring 2020 and integrated to a lead time of 36 h. The dates of the cases are as follows: 25 April–1 May, 3–16 May, and 18–28 May. The compute domains for each model configuration are shown in Fig. 1. For the real-time data flow, all experiments were re-gridded to a common 3-km grid for the CLUE, and those re-gridded data are used for analyses herein. For the 2.2-km um experiments, an area-weighted re-gridding approach was used to map data onto the 3-km CLUE grid. Because of the relatively small sample size of unique case days in SFE2020 ($N = 32$), we have elected not to perform verification of forecasts over regional subdomains in this study; there may be only a handful of cases with convection in some regions, and Schwartz and Sobash (2019) found some noisy regional statistics for large precipitation thresholds even using a much larger sample of cases ($N = 497$).

[1] The fv3, um, and wrf configurations were managed and executed in real time by NSSL, the Met Office, and NCAR, respectively. At each institution, the CAM runs for SFE2020 represented an annual iteration upon years of ongoing work; most configuration choices were already established in previous years.

[2] Judd et al. (2008) present evidence that "operational weather models evolve onto attracting manifolds of lower dimension than the entire state space," and that common DA techniques may produce ICs that are distant from such manifolds. These manifolds, the details of which are not generally known a priori for a given NWP model, are what we mean by "model attractors."

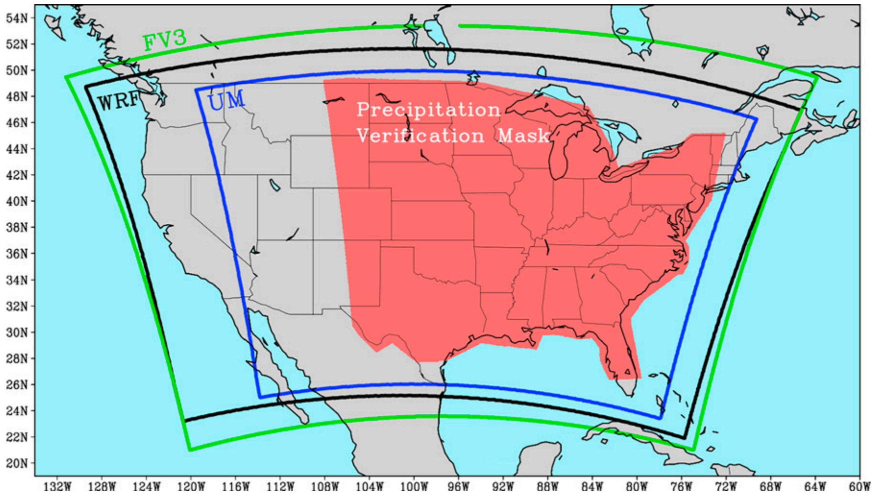

FIG. 1. Model domain for the fv3, wrf, and um configuration experiments are indicated by the green, black, and blue outlines, respectively. The red shaded region indicates the region over which precipitation and surrogate severe verification were conducted. The composite reflectivity, temperature, and dewpoint verification were conducted over the intersection of the um domain (blue outline) with the landmass of the contiguous United States.

b. Verification fields and methods

1) PRECIPITATION

We verify quantitative precipitation forecasts (QPFs) over 3- and 24-h periods to evaluate model skill in predicting the location and intensity of precipitation. NCEP's radar-derived, gauge-corrected Stage-IV quantitative precipitation estimate (QPE) dataset with 4.8-km grid spacing (Nelson et al. 2016) is used as truth. QPFs and QPEs are remapped to a common 4-km grid using a neighbor budget interpolation (Accadia et al. 2003). All QPF verification metrics are computed over the area depicted by red shading in Fig. 1, which covers most of the eastern two-thirds of the contiguous United States.

Multiplicative biases and fractions skill scores (FSSs; Roberts and Lean 2008) are computed for 3- and 24-h QPFs. The 3-h time windows cover forecast hours {0–3, 3–6, … , 33–36}, and the 24-h time window covers forecast hours 12–36. Multiplicative bias is the ratio of the number of forecast to the number of observed grid boxes exceeding a threshold; an unbiased forecast has a multiplicative bias of 1, while forecasts of too much or too little coverage of precipitation have biases greater or less than 1, respectively. FSS is based on the difference in the fraction of forecast and observed points that exceed a threshold within a specified radius of influence (ROI). Herein, FSS is directly computed using Eq. (3) in Loken et al. (2019):

$$\mathrm{FSS} = 1 - \frac{\dfrac{1}{M}\displaystyle\sum_{m=1}^{M}(F_m - O_m)^2}{\dfrac{1}{M}\left(\displaystyle\sum_{m=1}^{M}F_m^2 + \sum_{m=1}^{M}O_m^2\right)}, \qquad (1)$$

where M is the number of forecast–observation pairs, F_m is the ensemble mean forecast fraction of grid points exceeding the threshold within the ROI surrounding point m, and O_m is the equivalent fraction for the observations. Multiplicative bias and FSS are computed at QPF thresholds of 0.10, 0.25, 0.50, 0.75, and 1.00 in. Since Mittermaier and Roberts (2010) found FSSs can be sensitive to bias, a set of bias-corrected FSSs are computed by matching QPF quantiles in the forecasts and observations. For each threshold, the corresponding QPE quantile is computed, and the precipitation threshold in the QPFs matching that observed quantile is used in the FSS calculation. This process effectively removes bias, allowing a cleaner assessment of spatial placement. For bias-corrected FSSs, although the threshold used for QPFs is specified via quantile mapping, we still label the threshold by the original observational QPE value for clarity. For each of these thresholds, FSSs are computed for 12-, 24-, and 40-km ROIs.

2) COMPOSITE REFLECTIVITY

To assess skill in the forecast placement and coverage of convective storms, composite reflectivity (CREF) is verified by computing FSSs. The CREF verification domain is the intersection of the um grid (Fig. 1, blue outline) with the landmass of the contiguous United States. The verification approach is similar to that described for individual ensemble members in R20. To summarize, the instantaneous CREF field at lead times of {1, 2, … , 36} h is compared to the corresponding MRMS (Smith et al. 2016) merged reflectivity QC composite (hereafter merged reflectivity) available closest to the top of the hour (always within ±3 min). Specifically, using a 40-km ROI and 40-km Gaussian smoothing parameter (σ), neighborhood probabilities are computed for each forecast experiment-hour and compared to the equivalent smoothed probability field from merged reflectivity. A single threshold of 40 dBZ is verified, sufficient to capture the presence of most deep moist convection. A quantile mapping approach is applied by finding the percentile for each forecast experiment

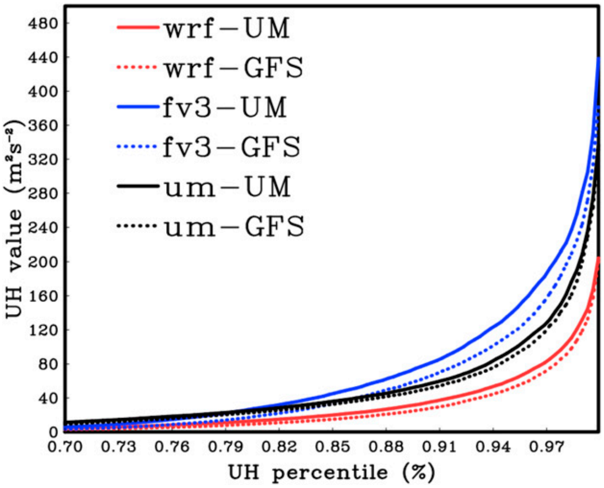

FIG. 2. UH values as a function of percentile for 24-h maximum UH (covering forecast lead times of 12–36 h) on the 81-km grid over all 32 cases for each set of model simulations.

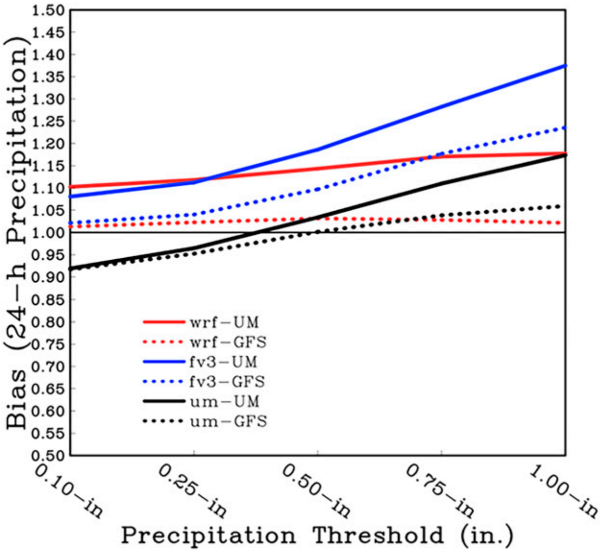

FIG. 3. Precipitation biases for 24-h accumulations (covering forecast lead times of 12–36 h) at thresholds of 0.10, 0.25, 0.50, 0.75, and 1.00 in. for each model, averaged over all 32 forecasts.

corresponding to 40 d*BZ* in the MRMS data, and that threshold is used to compute neighborhood probabilities from forecast reflectivity. As with precipitation, we simply refer to this threshold as 40 d*BZ* for clarity. We then use the bias-corrected CREF forecasts to compute FSSs with (1). To determine statistical significance, we compute 95% confidence intervals for the pairwise differences in FSSs between experiments at each lead time for all pairs that share either a common configuration or driving model. The confidence intervals are computed using the bootstrap technique of Wilks (2011) with 10 000 resamples of the 32 SFE2020 cases. The FSS difference between a pair of experiments is considered statistically significant if the 2.5th and 97.5th percentile of resampled differences both have the same sign.

Additionally, the degree of similarity between pairs of experiments is evaluated by computing the coefficient of determination (r^2) between the CREF neighborhood probability fields at each hour. There are 15 possible combinations of our six experiments, and r^2 is computed for each of these 15 pairs. Larger r^2 indicates that the two experiments being compared have more similar forecasts of storm placement and coverage. For five time bins (6–10, 12–16, 18–22, 24–28, and 30–34 h), using the same bootstrap approach described for CREF FSSs, 95% confidence intervals are computed for the difference between the mean r^2 of all pairs sharing a configuration and the mean r^2 of all pairs sharing a driving model.

3) SURROGATE SEVERE

In addition to placement and coverage of convection generally, another important role of CAM forecasts is the detection of potential severe weather [tornadoes, wind gusts ≥ 50 kt (25.7 m s^{-1}), or hail diameter ≥ 1.0 in.] specifically. To verify this aspect of our experiments, the surrogate severe approach (Sobash et al. 2011, 2016) is employed using 2–5 km above ground level (AGL) updraft helicity (UH; Kain et al. 2008) as a proxy for forecast severe weather. The approach for generating the surrogate severe forecasts is again virtually identical to that used in R20. As in R20, UH percentiles are used because the experiments have widely variable UH climatologies (Fig. 2). Surrogate severe forecasts are verified against preliminary local storm reports (LSRs) from the SPC (obtained from spc.noaa.gov/climo/reports) mapped to the same 81-km grid on which surrogate severe forecasts are generated; any grid box with one or more LSRs is assigned 1, while all others are assigned 0. One difference is that in the present study, surrogate severe forecasts are produced both for the 24-h time window covering 1200–1200 UTC (forecast lead times of 12–36 h) as in R20, and also for rolling 4-h time windows covering lead times of {4–8, 5–9, … , 32–36} h. The rolling windows afford a perspective on how skill evolves with lead time, and we use a window size of 4 h because Krocak and Brooks (2020) found that >95% of severe LSRs within 40 km of a point on any given severe weather day occur within a 4-h period. Surrogate severe forecasts are verified over the same domain as QPF (Fig. 1, red shaded area).

For surrogate severe forecasts, the relative operating characteristic curve (ROC; Mason 1982) is constructed by plotting the probability of detection (POD) versus the probability of false detection (POFD) using increasing probability thresholds. Herein, the thresholds used are 2%, 5%, 10%, 15%, … , 90%, and 95%. The area under the ROC curve (AUC) is computed, which measures the ability of the forecast to discriminate between events and nonevents. The possible range of AUC is 0–1, where values of 0.5 or below indicate no skill and 1 is a perfect forecast. The trapezoidal approximation (e.g., Wandishin et al. 2001) is used to calculate the AUC, which simply involves connecting each consecutive POD–POFD point with a straight line. This creates a series of trapezoids when considering the area directly beneath each pair of adjacent

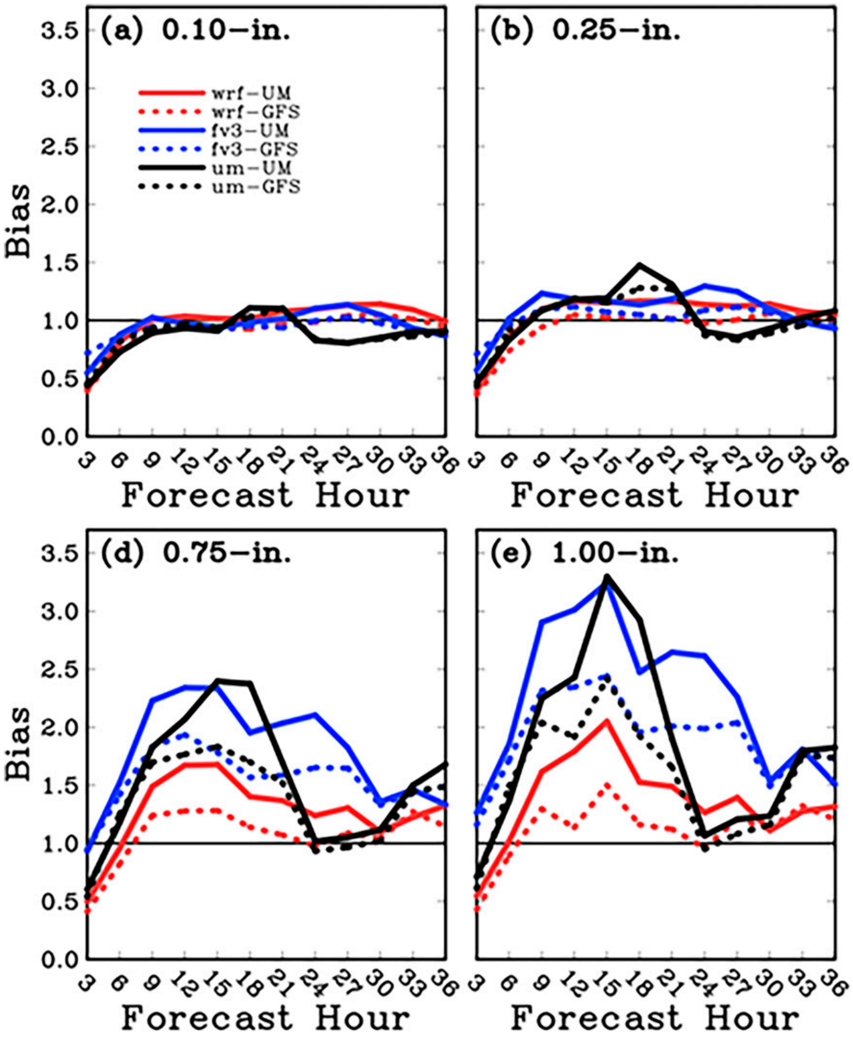

FIG. 4. Precipitation biases as a function of lead time for 3-h accumulations at thresholds of (a) 0.10, (b) 0.25, (c) 0.50, (d) 0.75, and (e) 1.00 in., averaged over all 32 forecasts. A legend is provided in (a).

POD–POFD points. The areas of all the trapezoids are summed, which gives a robust estimate of the AUC.

A version of FSS is calculated using the mean-square error of the severe weather probabilities relative to "practically perfect" observations (e.g., Hitchens et al. 2013), with the latter calculated by applying a Gaussian filter with $\sigma = 120$ km to the 81-km grid of LSRs [see Eqs. (3)–(5) in Sobash et al. (2011)]. Additionally, the Brier skill score (BSS) is computed as

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{reference}}}, \qquad (2)$$

where

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2, \qquad (3)$$

and N is the number of forecast–observation pairs, P is the severe weather probability at the ith point, and O is the observation at the ith point (1 if an event occurred and 0 if it did not occur). $\mathrm{BS}_{\mathrm{reference}}$ uses the same formula as BS, except P is the sample climatology of the event over all cases. Finally, the reliability component of the Brier score ($\mathrm{BS}_{\mathrm{rely}}$; Murphy 1973) is computed following Eq. (2) in Atger (2003): the squared differences of the probabilities (within specified bins) and their corresponding observed frequencies, weighted by the bin's frequency of forecast occurrence, are summed. $\mathrm{BS}_{\mathrm{rely}}$ measures how closely the points within a reliability diagram follow the perfect reliability line, and smaller values (closer to 0) are better.

4) THERMODYNAMIC FIELDS

We verify 2 m AGL temperature and dewpoint to assess model forecasts of near-ground thermodynamics, which are often crucial controls on the convective environment and convection initiation. These thermodynamic fields are verified over the same domain as CREF (the intersection of the um compute domain with the contiguous United States).

We use two separate truth datasets to verify the 2 m AGL thermodynamic fields. The first is the NCEP Real Time

Mesoscale Analysis (RTMA; e.g., Morris et al. 2020), which provides an hourly gridded analysis. For 2 m AGL temperature and dewpoint, the RTMA uses a blend of the most recent available High Resolution Rapid Refresh (HRRR; Benjamin et al. 2016) and North American Mesoscale (NAM; e.g., Aligo et al. 2018) CONUS nest model runs as its first-guess field. The weights assigned by RTMA to each of these model forecasts are inversely related to the forecast's age, with HRRR forecasts initialized hourly and NAM CONUS nest forecasts initialized every 6 h; thus, the HRRR always receives a larger weight, and is effectively dominant at some hours. The HRRR itself may be subject to systematic biases; e.g., Lee et al. (2019) found consistently warm 2-m temperature biases relative to micrometeorological tower measurements in Alabama during an 8-month study period. These HRRR biases may pass through to the RTMA to some degree, especially in areas with poor observational coverage or quality. To address this issue, we also verify 2 m AGL temperature and dewpoint against all aviation routine weather report (METARs) available within the verification domain within ±30 min of the forecast valid time. Although the aggregate METAR verification statistics may be somewhat spatially biased by clustering of sites, they provide a separate perspective unaffected by any systematic RTMA biases.

For each experiment, using both the RTMA and METARs as truth, additive bias and root-mean-square error (RMSE) are computed hourly at forecast lead times of 1–36 h. Additive bias is simply the domain-wide mean deviation of the forecast field from the observation field; an unbiased forecast (relative to the RTMA truth) has a bias of 0.

Analogously to r^2 for CREF, we also compute root-mean-square difference (RMSD) between pairs of experiment forecasts for 2-m temperature and dewpoint. The orientation of this metric is opposite r^2, as larger RMSD indicates *less* similarity between the two forecasts being considered. For five time bins (6–10, 12–16, 18–22, 24–28, and 30–34 h), using the same bootstrap approach described for CREF FSSs, 95% confidence intervals are computed for the difference between the mean RMSD of all pairs sharing a configuration and the mean RMSD of all pairs sharing a driving model.

Although parcel-based diagnostics like convective available potential energy were not available for all forecast datasets owing to data flow and postprocessing constraints, we verify 700–500-mb (1 mb = 1 hPa) lapse rate and 850-mb temperature forecasts in order to assess thermodynamic forecasts aloft. For these fields, we evaluate model forecasts at 24-h lead time, and use rawinsonde observations taken at 0000 UTC daily from 54 sites dispersed across the intersection of all three model configuration domains and the contiguous United States (Fig. 1) as truth. We compute additive bias and RMSE for these fields.

3. Results

a. Precipitation

Multiplicative biases for 24-h QPF (Fig. 3) reveal generally larger high biases for larger precipitation thresholds across experiments, except for the wrf-GFS. Experiments driven by the UM-Global (solid curves) exhibit modestly higher biases than those driven by the GFS (dashed curves) for all three configurations. Additionally, um experiments (black curves) have the smallest biases overall, except at high thresholds. Figure 4 shows multiplicative QPF biases for 3-h windows within the forecast cycle. Both um experiments, but particularly the um-UM, display a peak in bias around 18-h lead time for low and moderate thresholds. This is followed by a rapid decrease at later lead times, when the um experiments have notably smaller biases than the other experiments. It can thus be inferred that the smaller 24-h QPF biases for the um configuration (Fig. 3) are mainly a result of less forecast precipitation during lead times of 21–30 h, corresponding to the late afternoon and early evening over the United States. This characteristic of the um QPFs is related to diurnal precipitation cycle errors they suffer from, which will be explored in more detail later in this subsection.

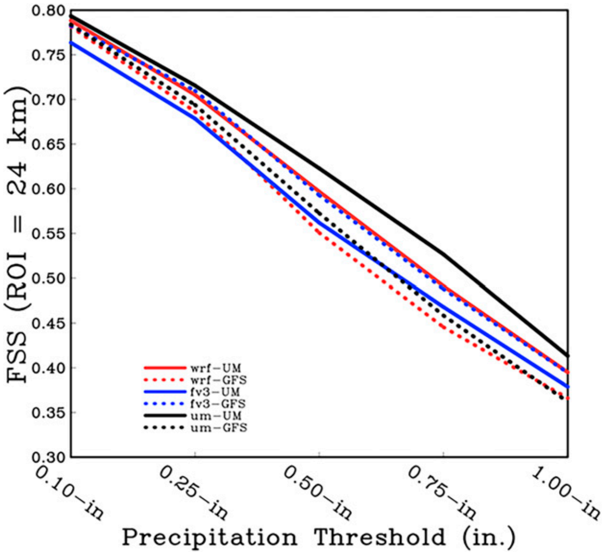

FIG. 5. Precipitation FSSs (ROI = 24 km; bias corrected) for 24-h accumulations (covering forecast lead times of 12–36 h), as a function of precipitation threshold in the observations, for each set of simulations averaged over all 32 forecasts. The threshold used for QPF is actually the percentile from that experiment's QPF distribution corresponding to the percentile of the labeled threshold in the QPE distribution.

Figure 5 shows FSSs for bias-corrected 24-h QPF as a function of threshold. The um-UM has the best skill across all thresholds, and its advantage is considerable at thresholds ≥ 0.5 in. UM-Global-driven experiments (solid curves) outperform GFS-driven experiments (dashed curves) for um and wrf, but the opposite is true for fv3. In the case of the fv3 and um configurations, experiments using the native driving model (fv3-GFS and um-UM) achieve notably higher FSSs than those using the nonnative driving model (fv3-UM and um-GFS). Non-bias-corrected FSSs and FSSs for other ROIs were computed and, aside from differences in overall FSS magnitude, gave very similar results for the 24-h period (not shown).

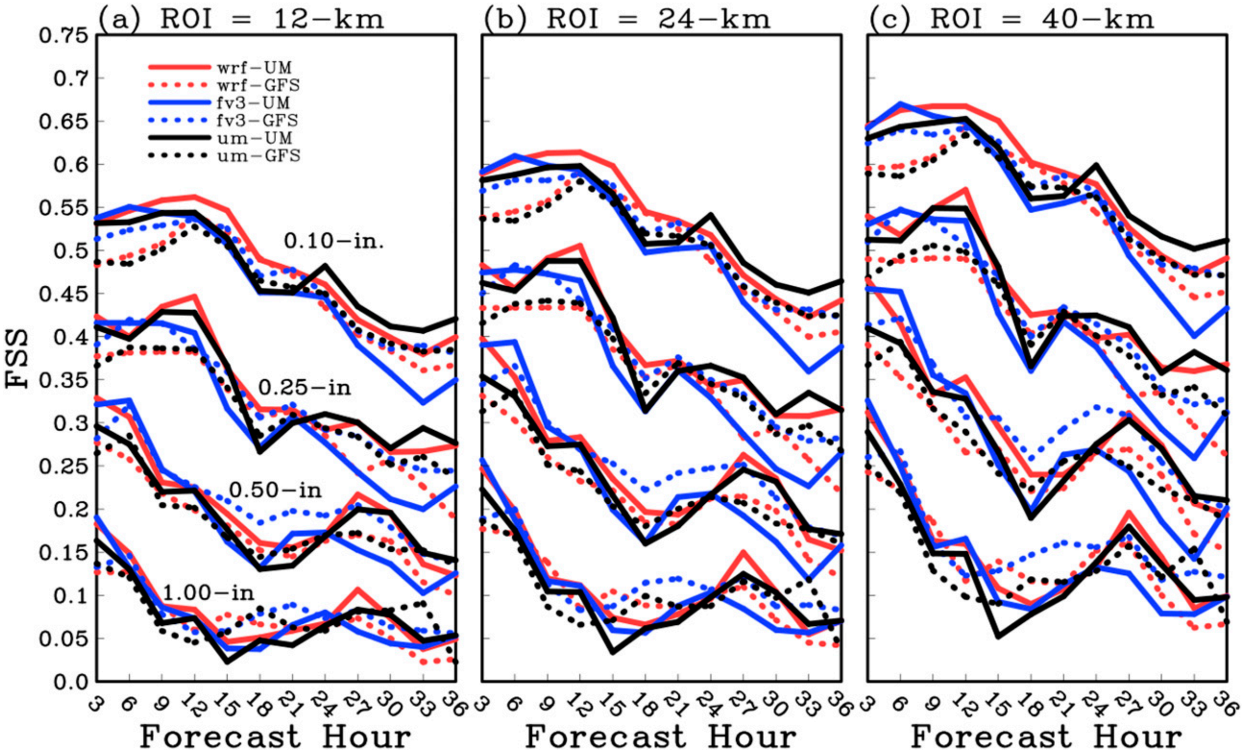

FIG. 6. Precipitation FSSs (bias corrected) for 3-h accumulations as a function of lead time for the observed thresholds 0.10, 0.25, 0.50, and 1.00 in., using ROIs of (a) 12, (b) 24, and (c) 40 km. FSSs are averaged over all 32 forecasts. In (a), the thresholds corresponding to each set of FSSs are given in the legend. The threshold used for QPF is actually the percentile from that experiment's QPF distribution corresponding to the percentile of the labeled threshold in the QPE distribution.

FSSs for bias-corrected 3-h QPF (Fig. 6), however, reveal that skill differences are conditional on lead time. The um-UM, which performs best at all thresholds for aggregate 24-h QPF, is only among the most skillful experiments after 24–30-h lead time at most thresholds. At higher thresholds, the um-UM also exhibits a pronounced minimum in FSS for the 15–21-h period, corresponding temporally to its sharp peak in bias (Fig. 4). The wrf-UM is consistently among the most skillful experiments throughout the forecast, particularly early (note that spinup is likely incomplete during the first 6–12 h, but the wrf-UM still tends to outperform the um-UM until 18-h lead time). UM-Global-driven experiments (solid curves) generally outperform their GFS-driven equivalents (dashed curves) for the first 12–15 h at lower thresholds, especially in the case of um and wrf configurations, for which the magnitude of the UM-Global-driven FSS advantage is typically near 0.05 during this period. This suggests the UM ICs may provide an advantage over GFS ICs with respect to precipitation features. Overall, however, there is not a particular configuration or driving model that consistently performs best across QPF thresholds and lead times for 3-h QPF, even after applying bias correction. This suggests that, when forecasting across lead times and weather regimes, the diversity of our experiments' configurations and driving models could be useful in a CAM ensemble of opportunity.

To investigate the substantial discrepancies in experiment rank order for bias and FSS with lead time, we examine time–longitude plots of hourly QPF for each experiment (Figs. 7b–g), averaged across all cases; for reference, the observational Stage-IV data are also provided (Fig. 7a). It is apparent that both um experiments exhibit a diurnal QPF maximum in the eastern United States around 18–24 h (Figs. 7f,g), earlier in the forecast period than seen in the other experiments or observations; this shift is especially pronounced in the um-UM. Accordingly, the spatial correlation coefficients with Stage-IV are nearly 10% smaller for the um experiments than the fv3 and wrf. Domain-averaged 1-h QPF (Fig. 7h) confirms that um experiments (black curves) display a marked offset of the diurnal precipitation peak: it occurs at 19- or 20-h lead time, whereas the other configurations match Stage-IV's peak at 23-h lead time. Overall, these results suggest a rather pronounced systematic early bias in the diurnal convective cycle for the um configuration used in the present study; yet, when QPF is summed over the entire diurnal cycle, the um-UM still performs remarkably well with a heightened skill advantage in the second overnight period, particularly at the 0.1- and 0.25-in. thresholds (Fig. 4).

Subjective examination of QPF and composite reflectivity forecasts for all 32 cases (not shown) reveals a persistent tendency for early initiation (often between 1700 and 1900 UTC) of cellular convection in both um experiments, particularly the um-UM; this behavior was most pronounced from 23 to 28 May, when a weak upper-level low pressure system lingered over the south-central United States. Figure 8 displays

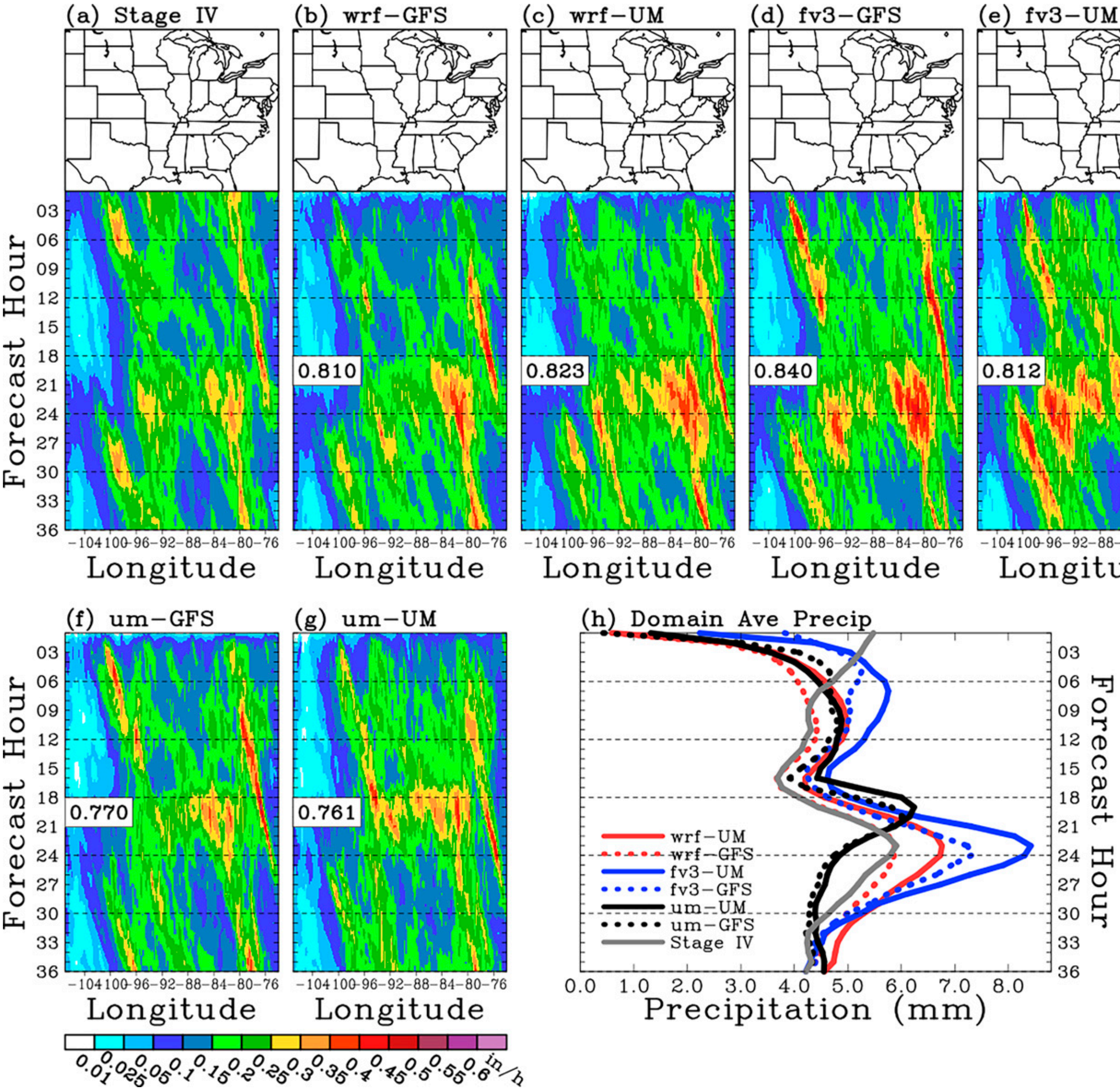

FIG. 7. Time–longitude diagrams of diurnally averaged precipitation for (a) Stage-IV, (b) wrf-GFS, (c) wrf-UM, (d) fv3-GFS, (e) fv3-UM, (f) um-GFS, and (g) um-UM over all 32 forecasts. In (a)–(e), the map at the top indicates the domain over which the time–longitude diagrams are constructed. In, (b)–(g), the spatial correlation between the forecast and Stage-IV observations at forecast hours 3–36 is denoted at the middle left. (h) Hourly domain averaged precipitation for Stage-IV and each set of forecasts.

composite reflectivity for all experiments initialized at 0000 UTC 24 May 2020 valid at 18-h lead time, along with the corresponding MRMS merged reflectivity valid at 1800 UTC 24 May. In this illustrative case, widespread cellular convection developed by early afternoon over a large region around the periphery of the upper-level low in both um experiments (Figs. 8e,f). Although the fv3 and wrf configurations depict too little convection in eastern Texas (Figs. 8a–d), observed reflectivity (Fig. 8g) nonetheless reveals that both um experiments—and particularly the um-UM—are far too aggressive with the coverage of discrete, cellular storms elsewhere. Although they ran over a U.K. domain, Clark et al. (2016) found a resolution dependence in the UM wherein their 4-km configuration had storms “too few, too intense, and too organized” (the opposite problem we see in Figs. 8e,f) relative to their 1.5-km configuration (cf. their Figs. 1, 3, and 4). Keat et al. (2019) found typically a 2-h early bias in convection initiation for a set of cases run over South Africa with a 1.5-km UM configuration, and the early bias was even worse for a corresponding 300-m configuration. Thus, there is some precedent in the literature for premature convection initiation by UM CAM runs, particularly as grid spacing decreases from 4 km. The factor(s) responsible for the early, aggressive initiation of convection in the um configuration used herein constitute an important future research question, and may also be explored further during future testbed experiments.

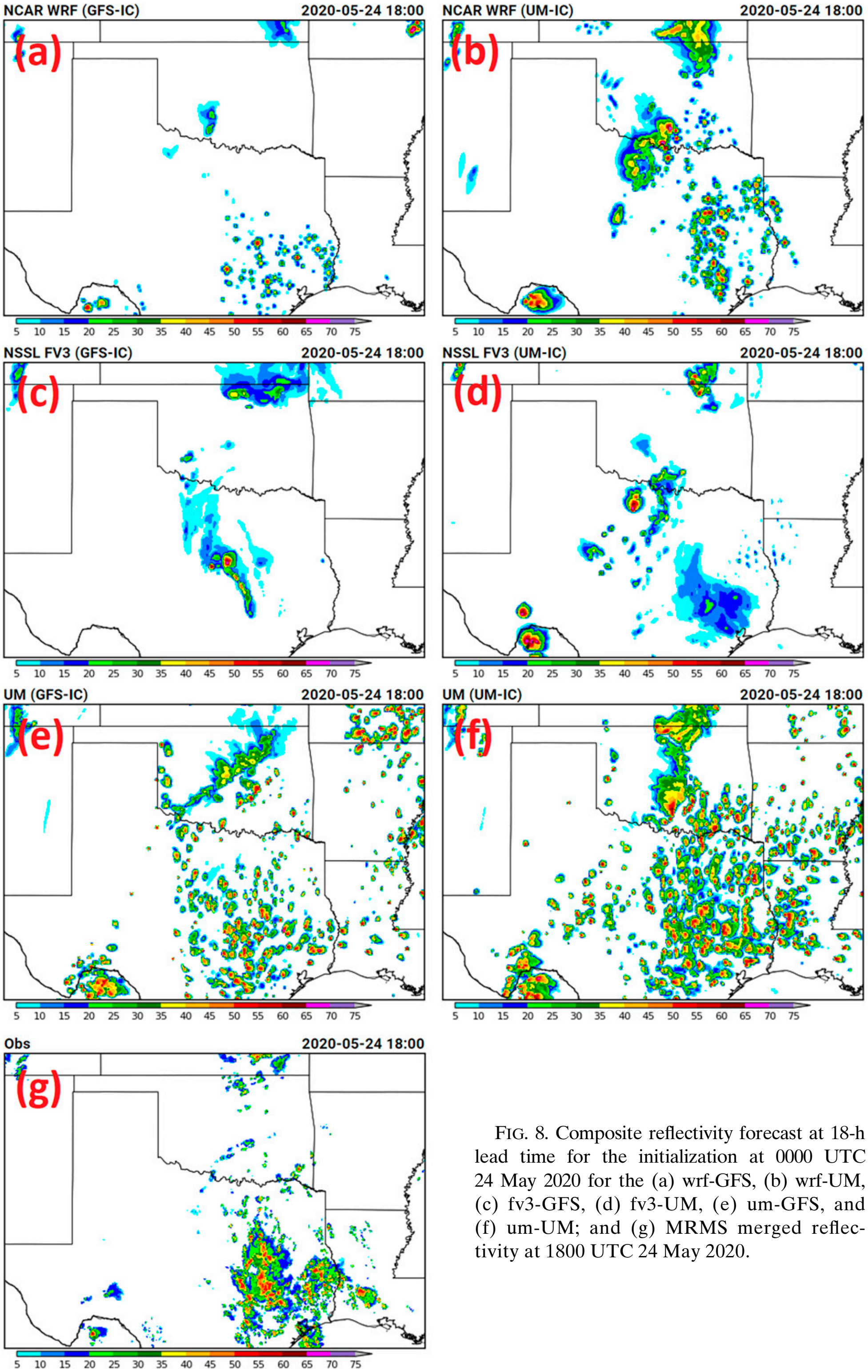

FIG. 8. Composite reflectivity forecast at 18-h lead time for the initialization at 0000 UTC 24 May 2020 for the (a) wrf-GFS, (b) wrf-UM, (c) fv3-GFS, (d) fv3-UM, (e) um-GFS, and (f) um-UM; and (g) MRMS merged reflectivity at 1800 UTC 24 May 2020.

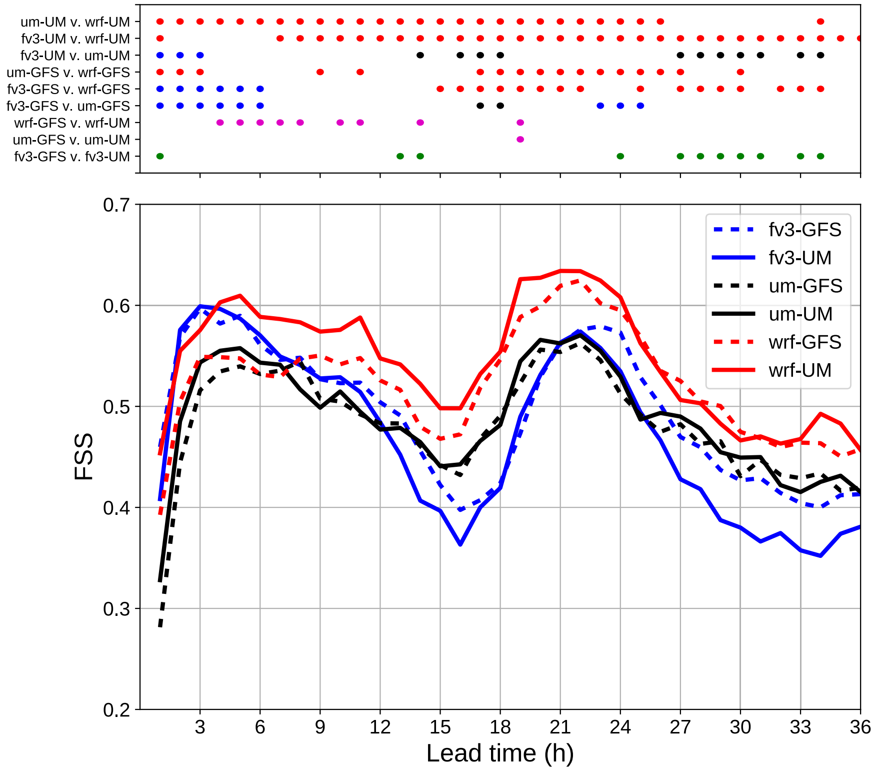

FIG. 9. FSSs for CREF neighborhood probabilities ($\geq$40 dBZ, ROI = 40 km) for each experiment as a function of lead time, averaged over all 32 cases. Instantaneous reflectivity forecasts are verified hourly at the top of the hour. The threshold used for forecast reflectivity is actually the percentile from that experiment's reflectivity distribution corresponding to the percentile of 40 dBZ in the MRMS reflectivity distribution. Above the time series, statistical significance is indicated for differences between some pairs of experiments. For pairs sharing a driving model (top six rows), a dot indicates the configuration whose color matches the time series legend has a better FSS at the 95% confidence level, while the absence of a dot indicates the difference is not statistically significant. For pairs sharing a configuration (bottom three rows), a magenta dot indicates that the UM-driven run has a better FSS at the 95% confidence level, a green dot indicates that the GFS-driven run is better at the 95% confidence level, and the absence of a dot indicates that the difference is not statistically significant.

b. *Composite reflectivity*

Figure 9 shows CREF $\geq$ 40 dBZ FSSs as a function of lead time for each experiment. From 12-h lead time onward, with few exceptions, wrf experiments (red curves) perform best, followed by um (black curves), then fv3 (blue curves). The advantage for wrf experiments over their fv3 and um equivalents is statistically significant at most lead times, particularly for the UM-driven experiments. There is also some indication of higher FSSs for experiments driven by their native model. The um-UM outperforms the fv3-UM by a statistically significant margin at numerous time steps late in the forecast cycle, whereas the fv3-GFS generally outperforms the um-GFS (albeit only with statistical significance for a short period in the afternoon). Also, the native fv3-GFS and um-UM outperform the nonnative fv3-UM and um-GFS, respectively, with statistical significance at some lead times. Overall, the CREF FSSs suggest a skill advantage for the wrf configuration, and also for native-driven experiments broadly, in predicting the location and coverage of storms. The native-driven model advantage corroborates our findings for QPF skill from section 3a; however, the pronounced wrf advantage for CREF is not seen for QPF. This suggests a difference exists in the model configurations' relative performance when focusing specifically on deep moist convection (e.g., CREF $\geq$ 40 dBZ) versus precipitation broadly.

Figure 10 shows the mean coefficient of determination (r^2) computed between the neighborhood probability fields for pairs of experiments. For lead times of 6–10 and 12–16 h, experiments sharing a driving model are more similar than those sharing a configuration, though only with a statistically significant difference in the first period. By lead times of 18–22 h, however, this pattern has reversed, and configuration pairs remain more similar than driving model pairs by a statistically significant margin through 30–34 h, the final period analyzed. The relative excess of $\overline{r^2}$ for configuration pairs over driving model pairs is largest at 18–22 h, the period during which um

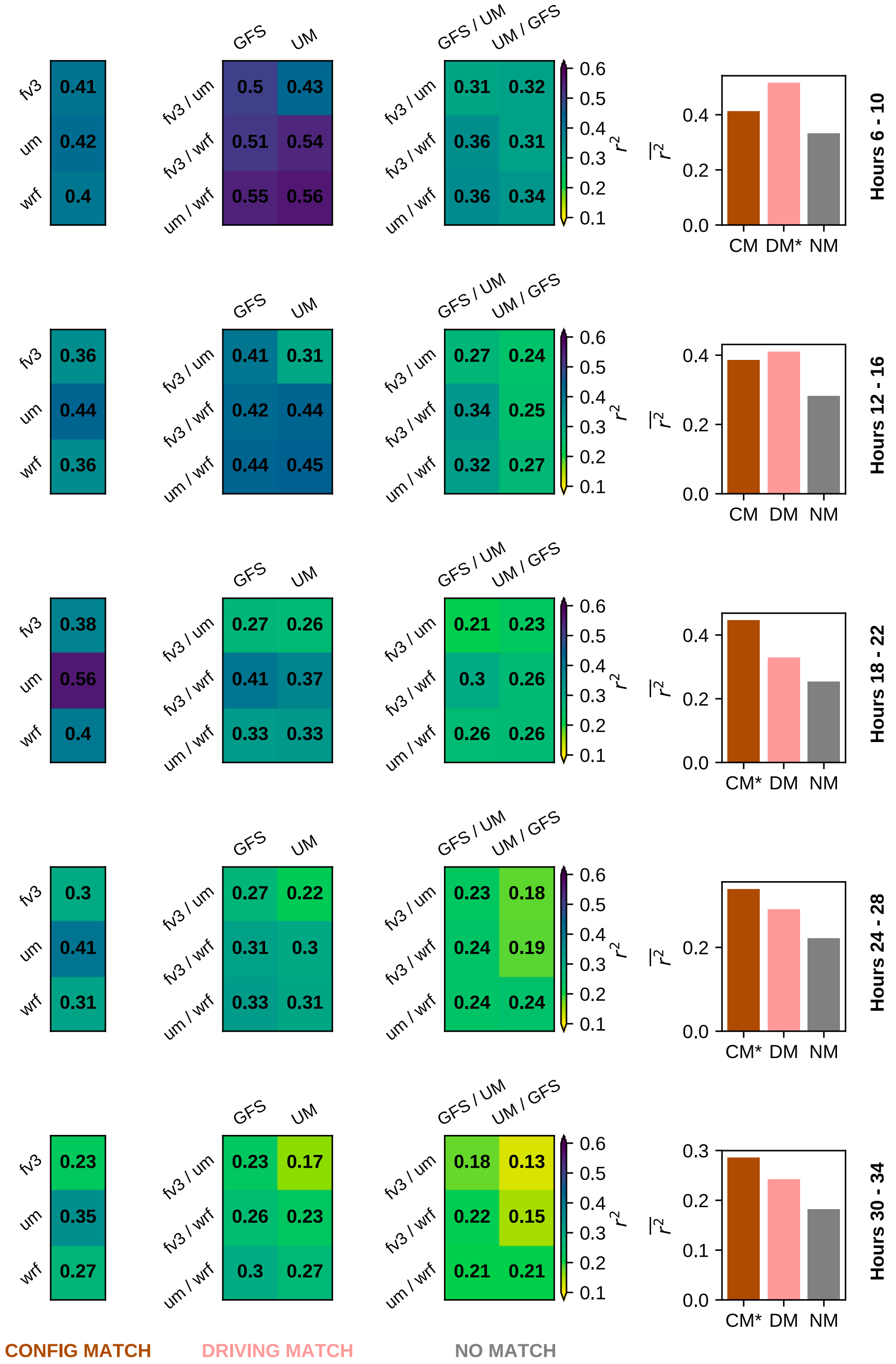

Fig. 10. Mean coefficient of determination (r^2), over all 32 cases, between CREF $>$ 40-dBZ neighborhood probabilities for pairs of experiments. Data are binned by forecast lead time, and each row represents a 4-h bin. Each bin covers probabilities for the instantaneous CREF field at the top of each hour in the bin, inclusive. The first, second, and third column of panels show $\overline{r^2}$ for experiment pairs sharing a configuration [“CONFIG MATCH” (CM)], driving model [“DRIVING MATCH” (DM)], and neither attribute [“NO MATCH” (NM)], respectively. For the leftmost column (CM), the left labels indicate the configuration shared by the experiment pair. For the second column from left (DM), the top labels indicate the shared driving model, and the left labels indicate the combination of configurations. For the third column from left, the top labels indicate the combination of driving models, and the left labels give the corresponding configurations (e.g., column “GFS/UM” and row “fv3/um” gives $\overline{r^2}$ for the fv3-GFS and um-UM). The rightmost column displays bar charts with $\overline{r^2}$ averaged over *all* experiment pairs within each of these three categories. An asterisk after “CM” or “DM” indicates that $\overline{r^2}$ for that group is statistically significantly larger than $\overline{r^2}$ for the other two groups at the 95% confidence level.

experiments also display a diurnal shift toward earlier QPF maxima (cf. Fig. 7h). Indeed, the pair of um experiments attains the largest $\overline{r^2}$ among all experiment pairs for every time bin after 18-h lead time, suggesting this configuration may have relatively unique attractors for state variables directly controlling convective evolution. Also notable is the particularly small $\overline{r^2}$ for pairings of the fv3 and um experiments. Experiment pairs driven by the GFS consistently show modestly larger $\overline{r^2}$ than those driven by the UM-Global. Across all time bins, experiment pairs which share neither a configuration nor driving model exhibit the lowest $\overline{r^2}$ on average, as expected. Broadly speaking, the clustering trends for CREF suggest a transition from stronger driving model influence to stronger model configuration influence somewhere around 16–18-h lead time (although the diurnal cycle could also be important, in which case this transition may preferentially fall near the time of local noon), providing one perspective on the answer to our key research question.

c. *Surrogate severe*

Figure 11 presents AUC, FSS, and $\mathrm{BS}_{\mathrm{rely}}$ for the 24-h surrogate severe forecasts (treating the entire period as a single field). Differences between all experiments in AUC are relatively small. For FSS, GFS-driven experiments using wrf and fv3 perform relatively well, but the equivalent um experiment is the worst of the six. Although the absolute magnitudes of $\mathrm{BS}_{\mathrm{rely}}$ differences among experiments are small, the wrf experiments perform best, with notably reliable forecasts of midrange probabilities. Overall, the only clear trend across these metrics is for um experiments to perform somewhat worse than wrf and fv3 for the 24-h surrogate severe forecasts; performance of the driving models is quite inconsistent across configurations.

Figure 12a shows maximum FSSs for 4-h rolling windows of surrogate severe forecasts. As with QPF, this finer temporal verification reveals marked changes in the relative performance of experiments across different periods of the forecast. Between lead times of 12–24 h, representing the morning and afternoon of the first diurnal cycle, the wrf-UM exhibits a substantial lead in skill among all experiments, and both um experiments (black curves) perform far worse than the other configurations. Between 21- and 25-h lead time, the um experiments make a rapid recovery after which they perform relatively well, alongside the fv3-GFS and wrf-GFS. Figures 12b–g show FSSs over the full computed σ-percentile space for each experiment for the 4-h period ending at 30-h lead time. By this second evening time frame, the UM-Global-driven (solid curves) experiments perform worse than the GFS-driven experiments (dashed curves), except for the native-driven um-UM. Figure 13 presents an equivalent plot for AUC, generally reflecting results similar to those found for FSS; however, relative differences between experiments are smaller, and the um-GFS shows skill similar to the fv3 and wrf configurations throughout the forecast. Overall, the 4-h surrogate severe forecasts reflect the same pronounced um deficiency seen for QPF in capturing the diurnal convective cycle from about 12–24-h lead time, and also suggest some advantage for GFS-driven fv3 and wrf experiments over their UM-Global-driven counterparts in capturing severe weather events.

To corroborate the um configuration's diurnal cycle deficiencies in the context of severe storms specifically, we plot a time–longitude diagram of normalized hourly average LSR density (Fig. 14a) to compare against equivalent plots of average UH for each experiment (Figs. 14b–g). Although the vastly disparate UH climatologies across configurations are apparent, the overall character of the UH distribution matches LSRs reasonably well for the fv3 and wrf experiments, as evidenced by $r > 0.775$ (Figs. 14b–e). However, the um experiments again display a notable early bias in UH. When considering domain-average LSR density and average UH (Fig. 14h), both the fv3 and wrf configurations peak about 1 h too early, whereas um peaks 3 h too early; this discrepancy, while substantial, is not quite as extreme as the discrepancy found for QPF. This is likely because much of the LSR and UH density is contributed from convective storms over the Great Plains region west of longitude 96°W (Figs. 14a,f,g), where the um experiments exhibit a less pronounced tendency for premature QPF than areas farther east (Figs. 7a,f,g).

d. *Thermodynamic fields*

Figures 15a and 15b present 2 m AGL temperature additive bias as a function of lead time. Except for a couple brief periods in the um-UM, all six experiments exhibit a cool bias throughout the forecast period. During the first 8 h, the cool bias is larger for GFS-driven experiments (dashed curves) than UM-Global-driven (solid curves). This trend reverses for the wrf (red curves) and fv3 (blue curves) configurations by 10-h lead time, with their GFS-driven experiments becoming warmer for the remainder of the forecast cycle; such a reversal is not present in the um configuration (black curves). At lead times of 18 h onward, there is distinct separation of mean bias by configuration: fv3 is coolest, followed by wrf, and um has the least pronounced cool biases. RMSE for 2 m AGL temperature (Figs. 15c,d) also exhibits clustering by configuration, with fv3 incurring the largest errors and um the smallest for most of the period. UM-Global-driven experiments uniformly outperform GFS-driven experiments for the first 10 h, suggesting a possible advantage for UM-Global atmospheric temperature ICs over the GFS. The fact that the fv3-UM has smaller RMSE than the fv3-GFS early in the forecast cycle (Figs. 15c,g), despite the potential for model shock in the fv3-UM related to the nonnative soil model, is further evidence of superior UM ICs. RMSE is maximized for all experiments at 20–26-h lead times, and during this period of the forecast, fv3 experiments show RMSEs up to 1–1.5 K larger than their equivalent wrf and um experiments. The rank order of experiments by RMSE closely resembles the rank order of bias magnitude throughout the forecast, suggesting systematic biases are a large driver of total error. For the fv3 and um configurations, the native-core driving model (GFS and UM-Global, respectively) affords smaller RMSE. For both bias and RMSE, results are qualitatively similar using either the RTMA or METARs as truth.

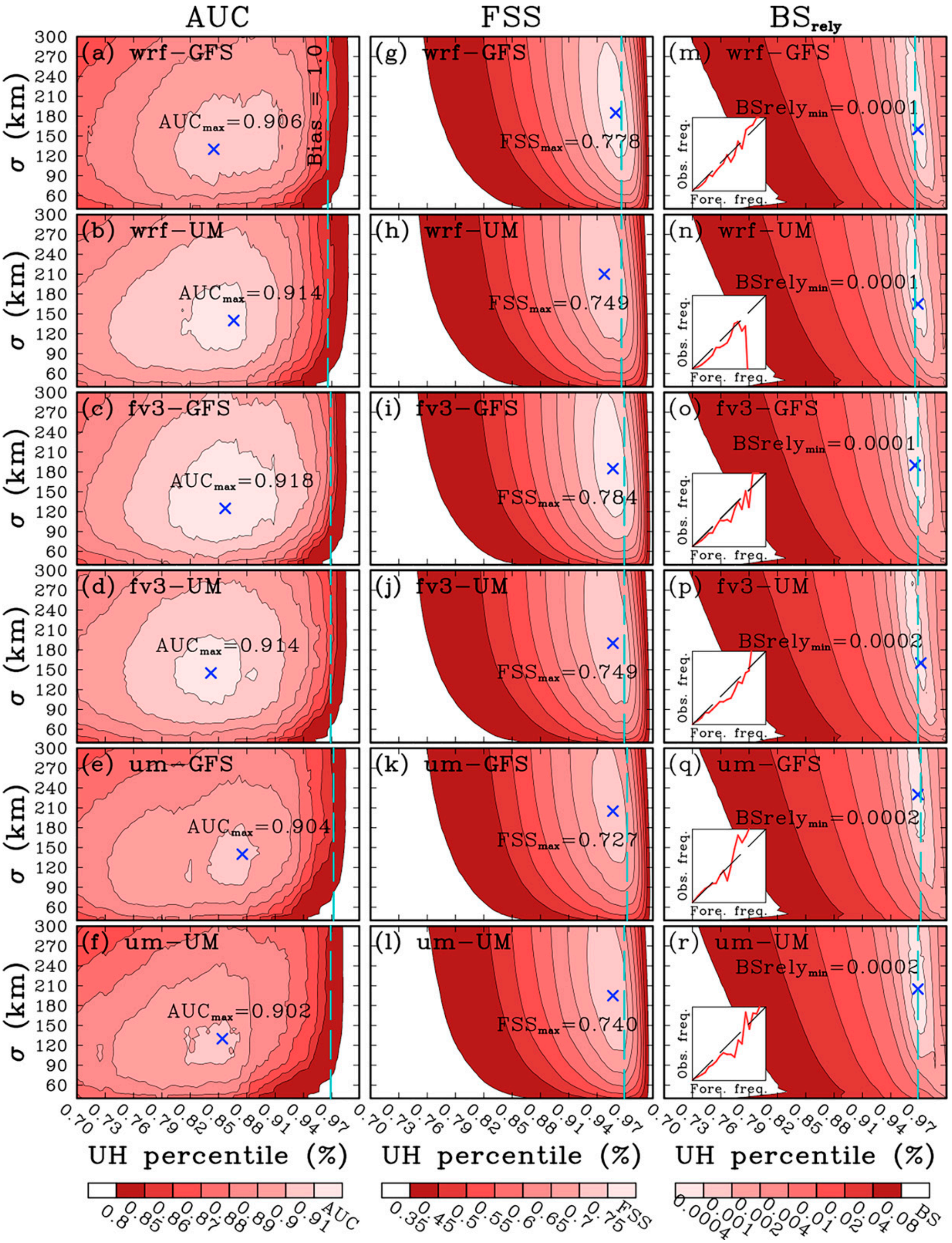

FIG. 11. AUC for 24-h surrogate severe forecasts as a function of σ and UH percentile for (a) wrf-GFS, (b) wrf-UM, (c) fv3-GFS, (d) fv3-UM, (e) um-GFS, and (f) um-UM averaged over all 32 SFE2020 cases. (g)–(l) As in (a)–(f), respectively, but for FSS. (m)–(r) As in (a)–(f), respectively, but for BS$_{rely}$. In each panel, a blue "x" marks the best score, which is indicated by nearby text. The vertical dashed line marks the UH percentile at which the number of surrogate severe LSRs approximately matches the number of observed reports (i.e., bias = 1.0). In (m)–(r), the reliability diagrams are shown corresponding to the σ and UH percentile at which BS$_{rely}$ is minimized.

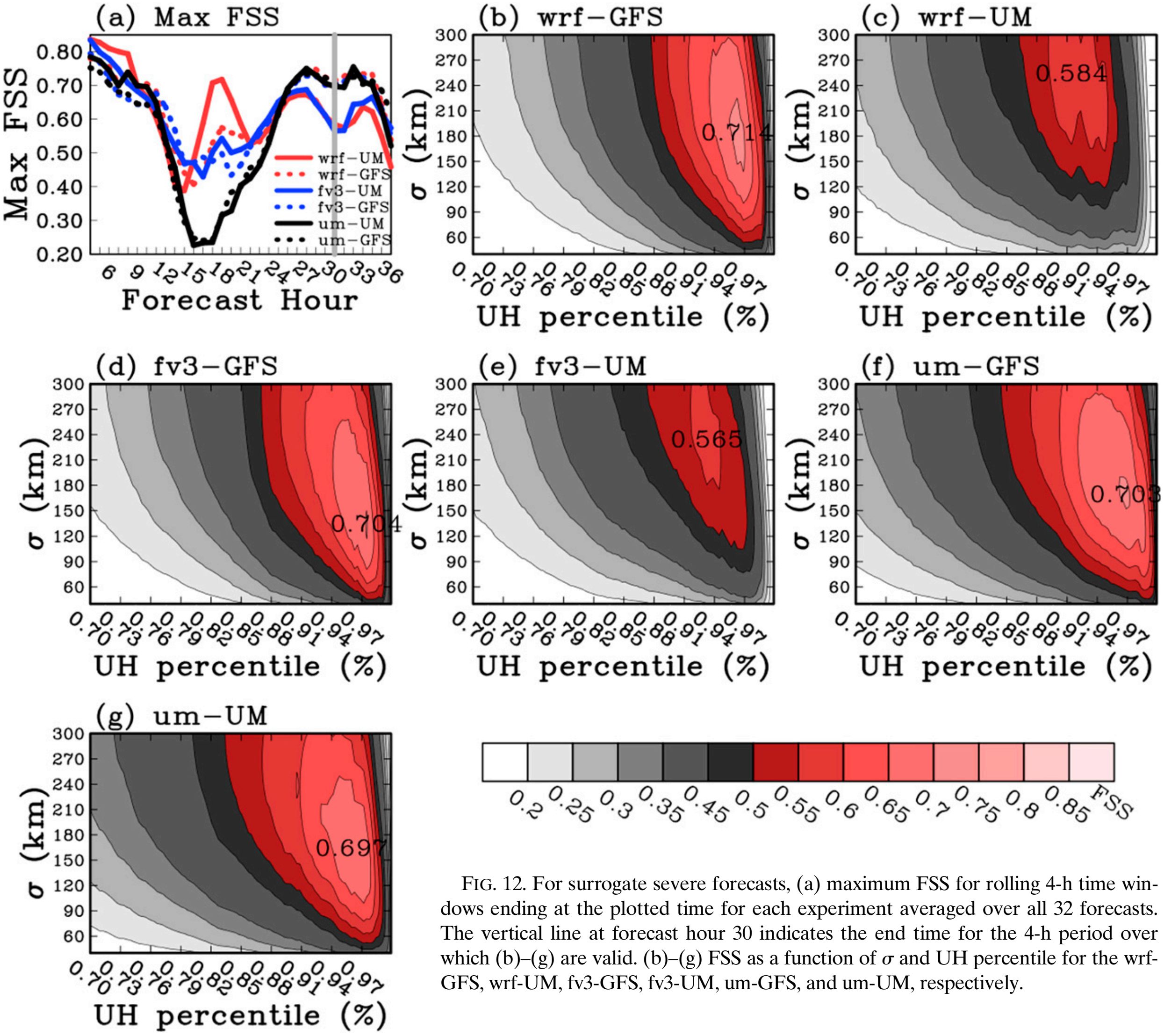

FIG. 12. For surrogate severe forecasts, (a) maximum FSS for rolling 4-h time windows ending at the plotted time for each experiment averaged over all 32 forecasts. The vertical line at forecast hour 30 indicates the end time for the 4-h period over which (b)–(g) are valid. (b)–(g) FSS as a function of σ and UH percentile for the wrf-GFS, wrf-UM, fv3-GFS, fv3-UM, um-GFS, and um-UM, respectively.

Additive biases for 2 m AGL dewpoint (Figs. 15e,f) reveal a more persistent influence of the driving models wherein UM-Global-driven experiments are drier than GFS-driven experiments, although the difference is of negligible magnitude in the case of the um experiments. Unlike for temperature, there are qualitative differences between the RTMA and METAR verification statistics: using METARs as truth results in systematically larger moist biases by around 1 K across experiments, compared to using the RTMA. Model configuration differences dominate the diurnal cycle of bias: um experiments tend toward a moist bias overnight and near-zero bias (against METARs) during peak diurnal heating; wrf experiments display the opposite diurnal cycle; and fv3 experiments are closest to unbiased for most of the forecast cycle outside the afternoon and early evening hours (20–28-h lead time), where they exhibit a pronounced moist bias. RMSE for 2 m AGL dewpoint (Figs. 15g,h) reveals similar diurnal trends in error magnitude for all experiments as for temperature, albeit with more uniform performance across the six experiments: RMSE never differs by more than 1 K between any pair of experiments at a given lead time, except briefly from 22 to 25 h. um experiments exhibit the largest errors for lead times earlier than 6 h, while RMSE is consistently largest for fv3-UM after 14-h lead time. Although fv3 experiments perform worst during the second half of the forecast cycle, their RMSE excess for dewpoint is not as large as for temperature. Furthermore, the rank order of experiments by dewpoint RMSE late in the forecast does *not* follow the rank order by bias, and small bias does not imply small RMSE.

At 24-h lead time, additive biases for 700–500-mb lapse rate (Fig. 16a) suggest that fv3 and um configurations are nearly unbiased, while wrf configurations exhibit a modest high bias of 0.14–0.30 K km^{-1}. For 850-mb temperature (Fig. 16b), cool biases are present across experiments, but um experiments are the least biased overall. The rankings of 850-mb temperature biases by configuration match those seen at 24-h lead time for 2-m temperature (cf. Fig. 15a), with the fv3 coolest and um warmest at both vertical levels. RMSEs for 700–500-mb

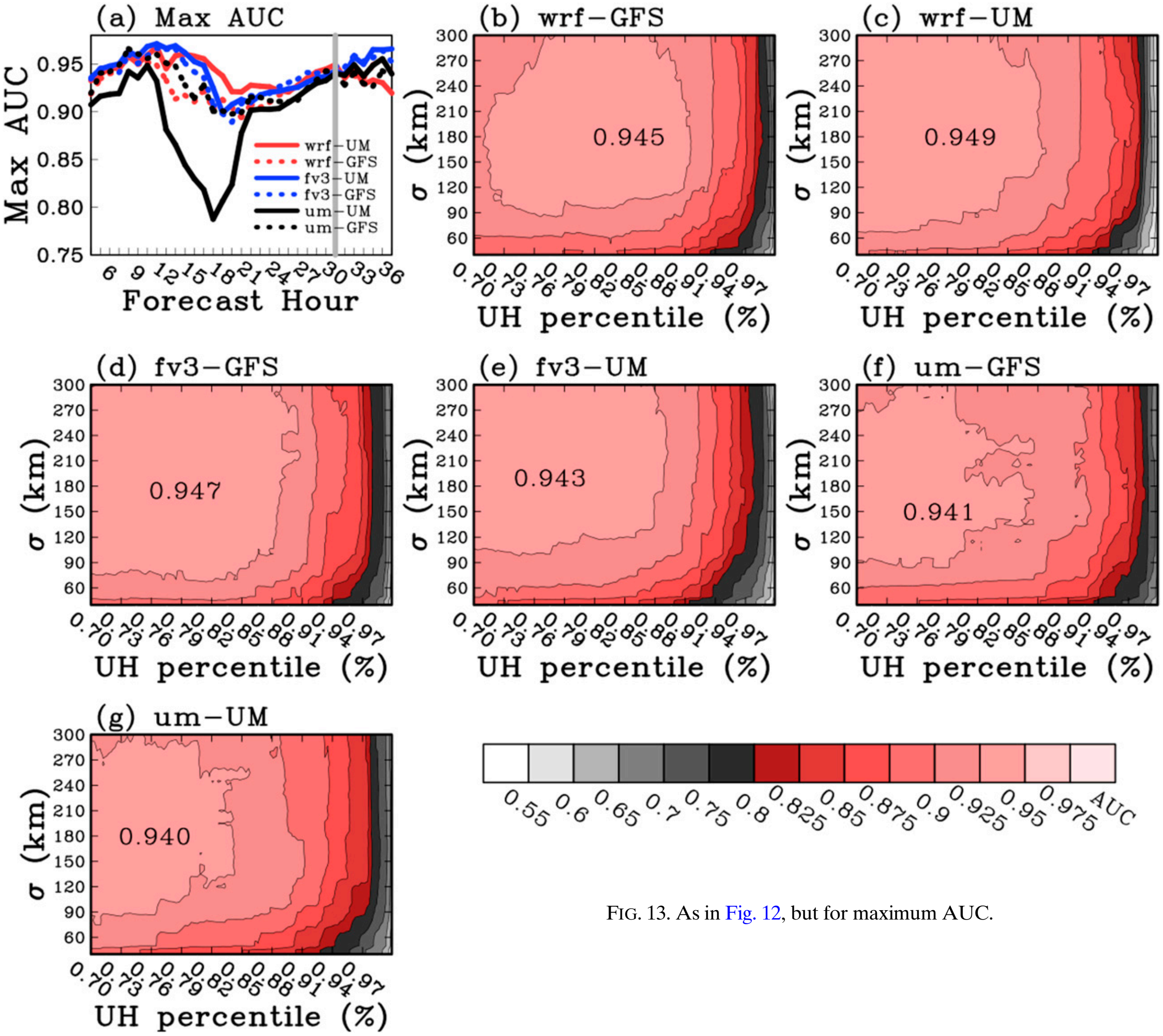

FIG. 13. As in Fig. 12, but for maximum AUC.

lapse rate and 850-mb temperature (Figs. 16c,d) are qualitatively similar among all six experiments. Although it is possible that the um configuration's smaller afternoon cool bias at 2 m AGL compared to fv3 and wrf reflects a tendency to reach the convective temperature more easily, its corresponding tendency to be warmer at 850 mb would also suggest typically stronger boundary layer capping. As such, our analysis does not provide a straightforward thermodynamic explanation for the um's tendency for premature convection initiation discussed in section 3a.

Figure 17 displays 2 m AGL temperature RMSD between pairs of experiments. Among shared configurations, um and wrf experiments have smaller differences than fv3 experiments. Among shared driving models, GFS-driven runs are modestly more similar to one another than are UM-Global-driven runs, in the aggregate. Figure 18 shows equivalent RMSD for 2 m AGL dewpoint. The same trends seen for temperature RMSD hold for both the model configurations and driving models, and the separation between groups is larger overall for dewpoint. These thermodynamic fields follow some trends seen with CREF r^2 regarding the clustering of experiments sharing a configuration or driving model. One potentially important difference is that clustering by driving model is less dominant at 6–10- and 12–16-h lead times for 2 m AGL thermodynamics than for CREF; in fact, clustering is statistically significantly stronger by configuration even at these early lead times for temperature. This points to an even more immediate influence of configuration details on temperature than on CREF. We speculate that this may be modulated to a large degree by differences in the PBL schemes of each configuration, but this is a question for future work in which individual parameterization schemes are varied systematically. After 18-h lead time, clustering by model configuration is substantially stronger than by driving model, as also seen for CREF. Additionally, the two experiments sharing the um configuration are more similar than any other pair of experiments.

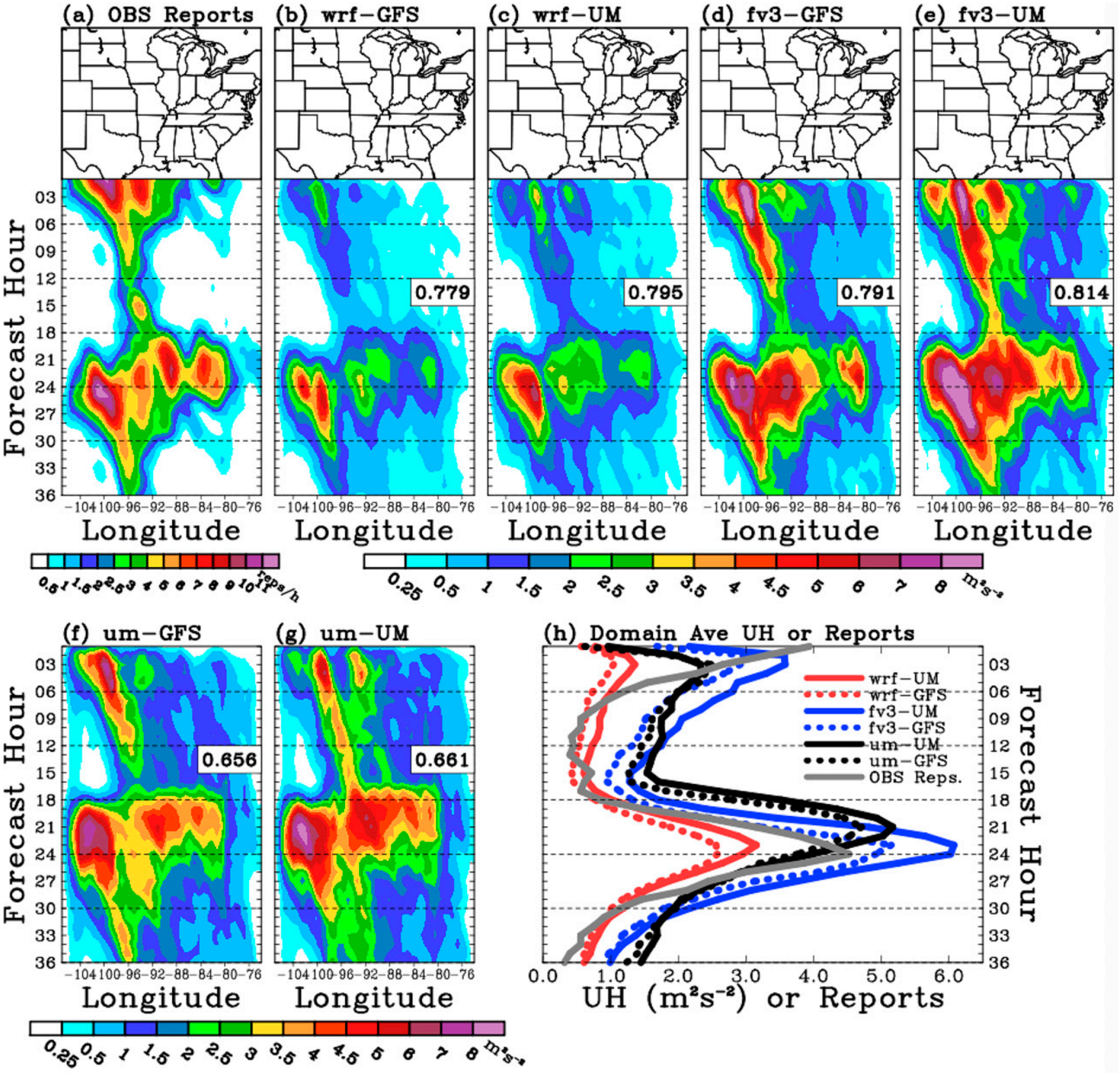

FIG. 14. (a) Time–longitude diagram of diurnally averaged storm report frequencies, and (b)–(g) as in (a), but for UH for wrf-GFS, wrf-UM, fv3-GFS, fv3-UM, um-GFS, and um-UM, respectively. In (a)–(e), the map at the top indicates the domain over which the time–longitude diagrams are constructed. In (b)–(g), the spatial correlation between the forecast and observed storm report frequencies at forecast hours 3–36 is denoted at the middle left. (h) Hourly domain-averaged storm report frequencies and UH for each set of forecasts. The storm report frequencies are scaled by a factor of 700.

In aggregate, influences from both the driving model and model configuration are evident in thermodynamic fields at 2 m AGL. Model configuration tends to play the larger role in bias after diurnal heating commences at 16–18-h lead times, which addresses our research question regarding the lead times at which driving model influence loses primary importance in CAM solutions. During the crucial afternoon period when the background environment modulating diurnally enhanced convection is established, um experiments tend substantially warmer and drier than wrf or fv3 experiments, with fv3 exhibiting the coolest and moistest fields by a small margin over wrf. RMSE for both fields suggests the fv3 configuration has notably worse skill than um and wrf during the afternoon and evening, similar to the result found for CREF FSSs.

4. Summary and conclusions

A suite of six deterministic CAM experiments was run in real time over the central and eastern contiguous United States for 32 cases in spring 2020. The experiments covered

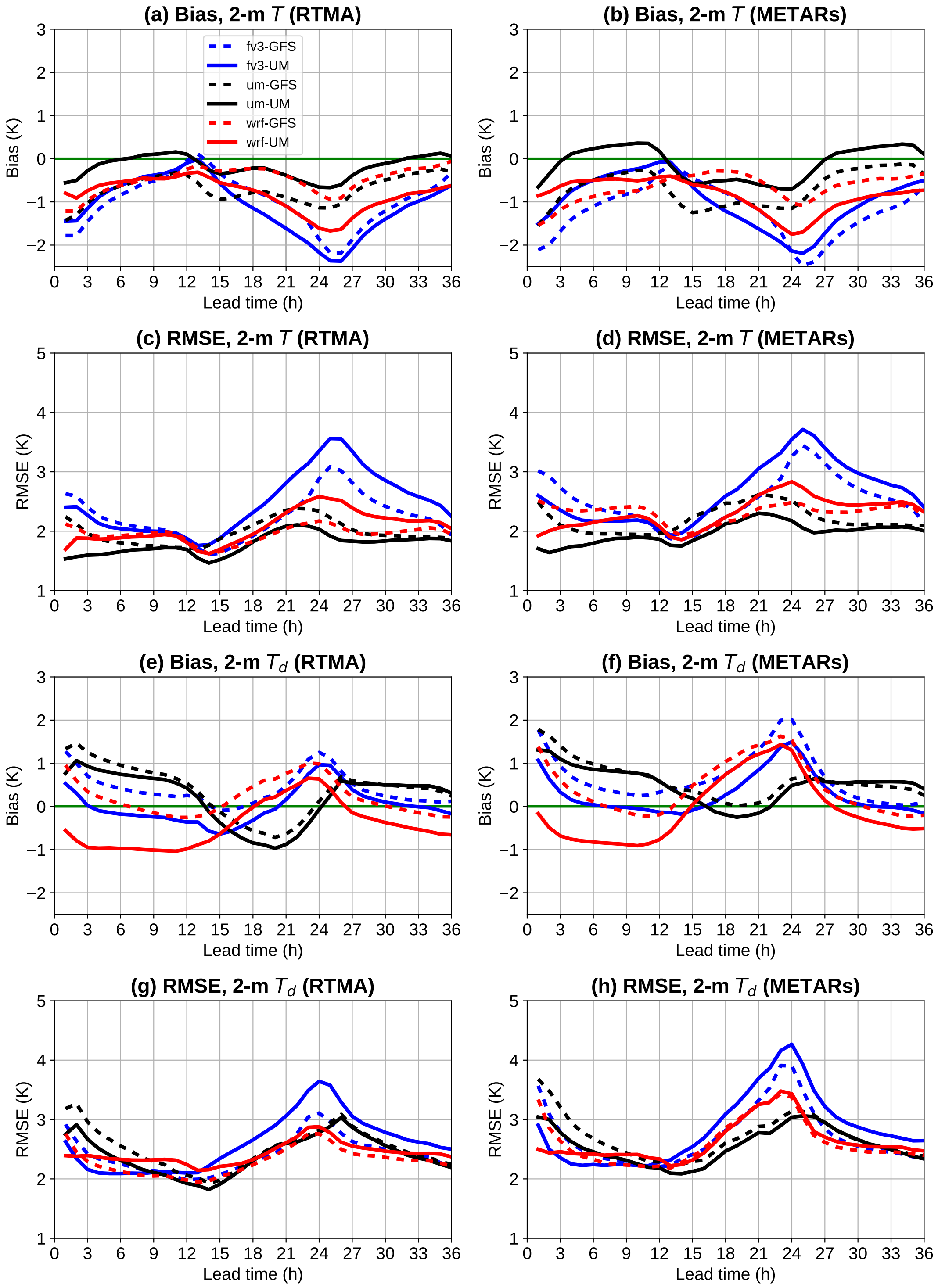

FIG. 15. Mean additive bias for 2 m AGL temperature as a function of lead time, averaged over all 32 cases, using (a) RTMA and (b) METARs as truth. (c),(d) As in (a) and (b), but for 2 m AGL temperature RMSE. (e),(f) As in (a) and (b), but for 2 m AGL dewpoint additive bias. (g),(h) As in (a) and (b), but for 2 m AGL dewpoint RMSE. Instantaneous forecasts are verified hourly at the top of the hour.

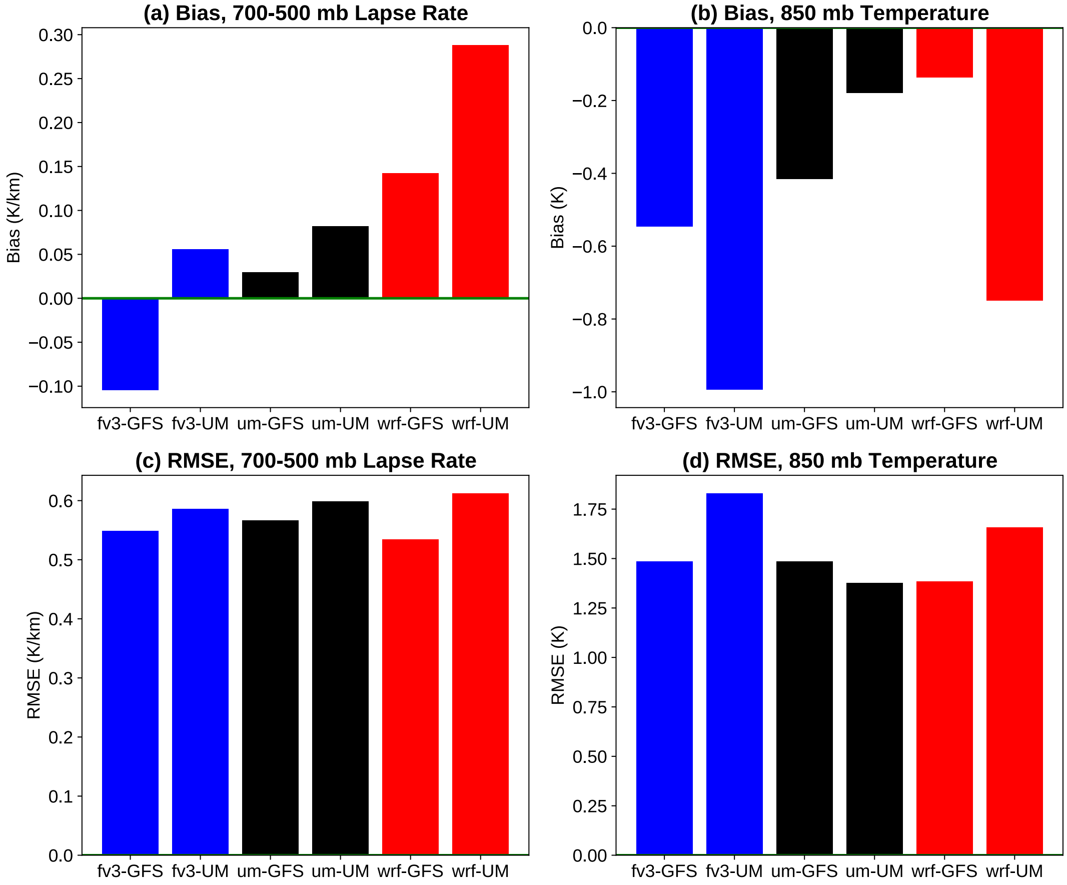

FIG. 16. Mean additive bias for (a) 700–500-mb lapse rate and (b) 850-mb temperature, averaged over all 32 cases at 54 rawinsonde sites within the contiguous United States, at 24-h lead time (all valid at 0000 UTC). (c),(d) As in (a) and (b), but for RMSE.

three model configurations (using the FV3, UM, and WRF dynamical cores), each of which was driven by ICs, LBCs, and soil states from two global models (GFS and UM-Global). Forecasts were initialized daily at 0000 UTC and ran out to a lead time of 36 h. To evaluate the skill of and differences between experiments for convective forecasting, we analyzed model QPF, CREF, UH, and thermodynamic forecast fields. Our key research question involved the relative influence of driving models versus model configurations upon these fields, with a particular interest in how those influences may change with lead time. We were also interested in documenting the biases, skill, and character of these fields for the model configurations we used; these model configurations represent three distinct modeling systems used internationally in operational NWP, and our dataset provides a unique opportunity for a relatively clean comparison.

a. Influence of model configuration versus driving model with lead time

We analyzed the similarity between our six experiments in terms of their forecasts for CREF (using correlations between neighborhood probability forecasts) and 2 m AGL thermodynamic fields (using RMSDs). Our results showed that these fields reliably clustered more by configuration than driving model after about 18-h lead time, whereas the clustering by configuration and driving model was comparable at 6–16-h lead times. Furthermore, clustering by configuration was consistently stronger for the um than for fv3 and wrf. It seems probable that the um configuration used herein has substantially different model attractors in its solution space for state variables directly impacting the evolution of deep moist convection, relative to the fv3 and wrf configurations whose dynamical cores are more common in American NWP. This suggests the um configuration could potentially add especially useful diversity to a multimodel CAM ensemble, considering that our um experiments—despite having solutions quite different from the others—performed competitively overall. In terms of our key research question, these results suggest the driving model's influence may be superseded by the model configurations for cold-start CAM forecasts around 16–18-h lead time. Since our forecasts were always initialized at 0000 UTC, these lead times corresponded to 1600–1800 UTC daily (early afternoon in the central and eastern United States), so there may also be a diurnal component to this handoff.

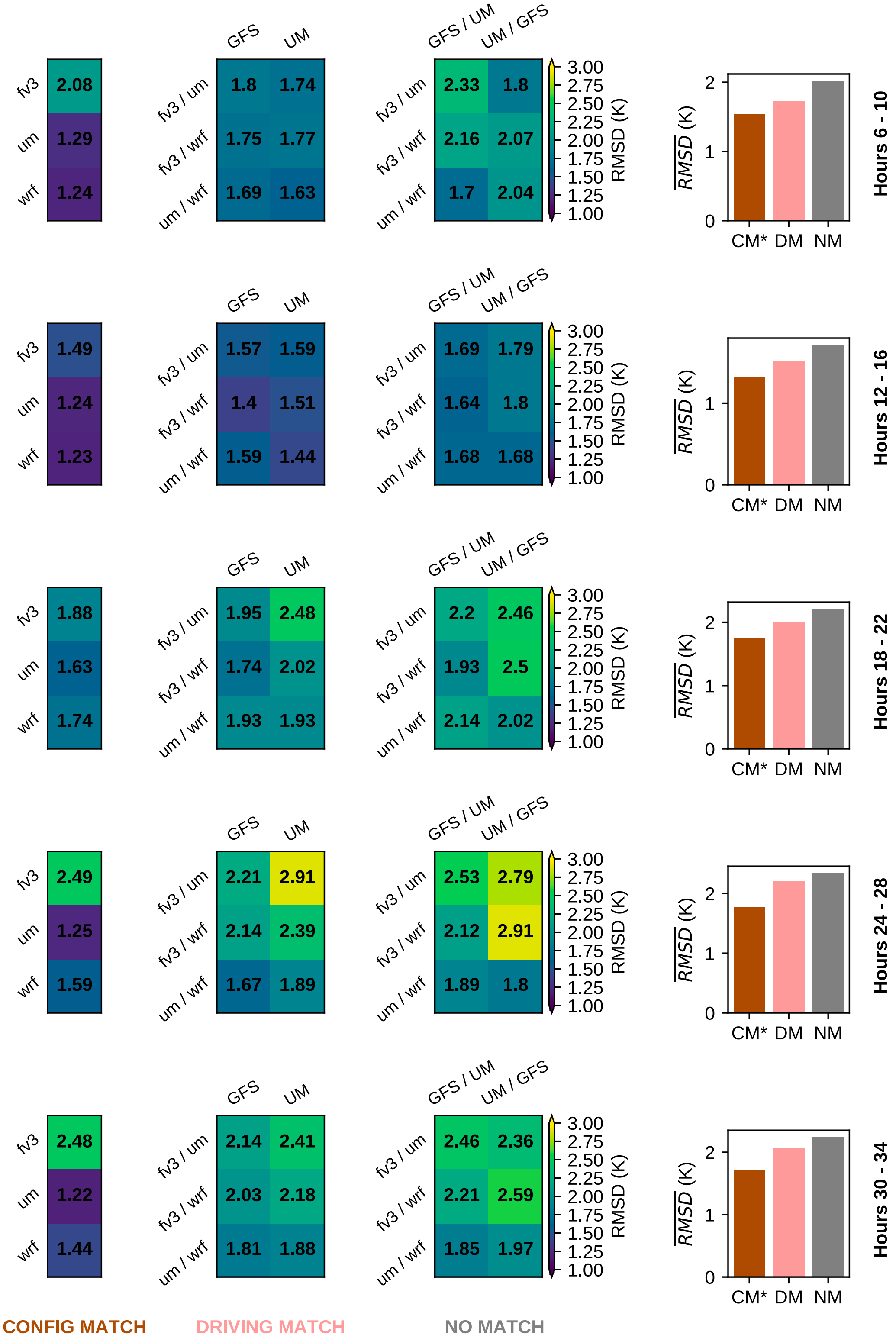

FIG. 17. As in Fig. 10, but for 2 m AGL temperature RMSD instead of CREF r^2. Larger RMSD implies less similarity between experiments (opposite of r^2 in Fig. 10). An asterisk after “CM” or “DM” indicates $\overline{\text{RMSD}}$ for that group is smaller than $\overline{\text{RMSD}}$ for the other two groups at the 95% confidence level.

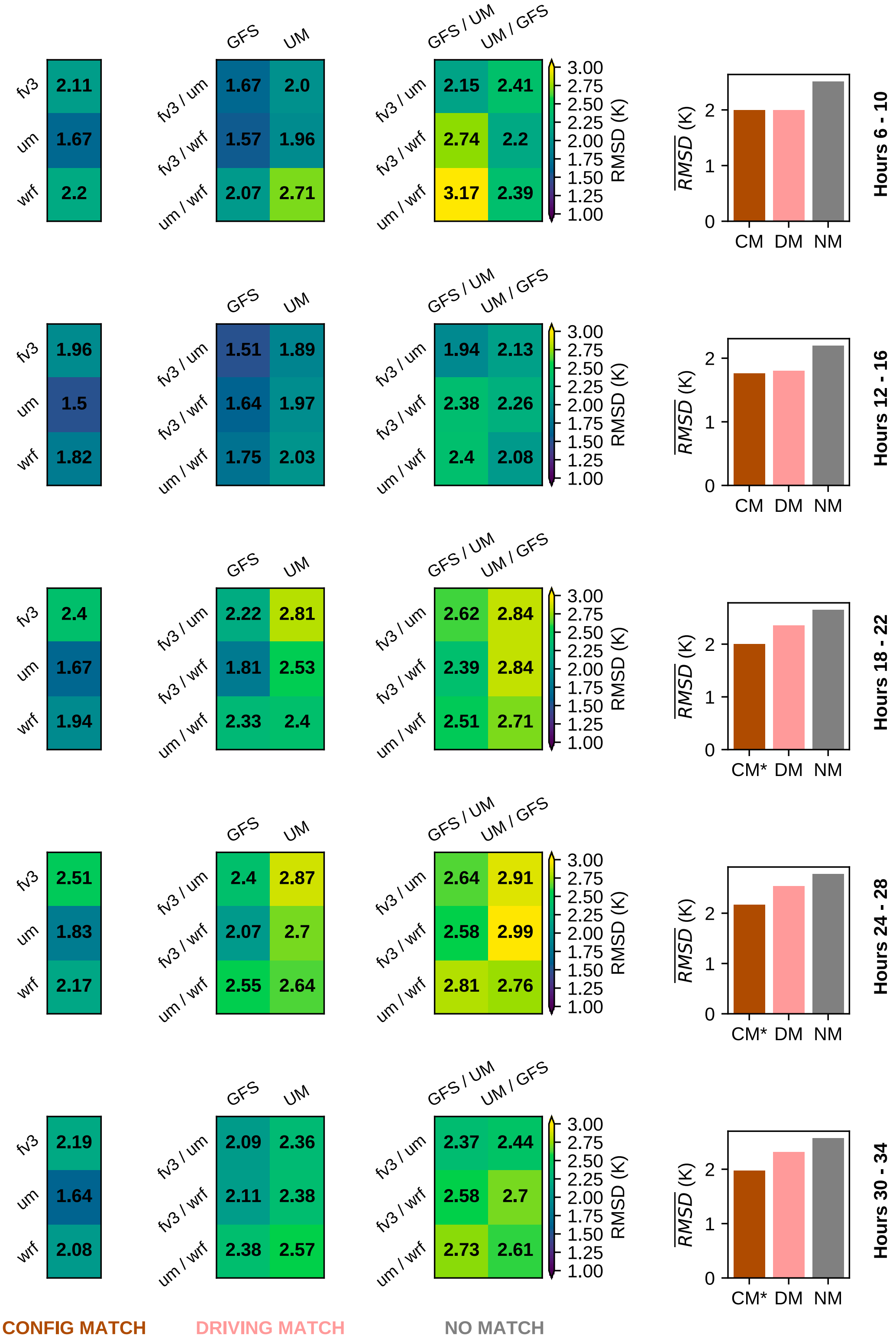

FIG. 18. As in Fig. 17, but for 2 m AGL dewpoint.

b. Biases and skill of our specific model configurations and driving models

Key results for the specific model configurations and driving models represented among our experiments include:

- Verification of 24-h QPF accumulated over 12–36-h lead times found the um-UM experiment performed best, but examination of 3-h QPF for subintervals of this period revealed that its superior skill is dominated by better forecasts at lead times after 24 h; both um experiments displayed a sharp drop out in skill for 15–24-h lead times.
- Further analysis of experiments' diurnal precipitation cycles showed a markedly early peak in QPF around 18–19-h lead time for the um experiments, whereas fv3 and wrf experiments matched observations with a peak around 23-h lead time.
- Skill for hourly CREF forecasts at the 40-dBZ threshold was best for wrf experiments at most lead times, followed by um and fv3.
- Skill for surrogate severe forecasts constructed from model UH favored the fv3 and wrf configurations over the um, with the um again exhibiting a large skill dropout during the morning and early afternoon period coinciding with its premature diurnal precipitation maximum.
- For both QPF and reflectivity, um and wrf experiments driven by the UM-Global outperformed those driven by the GFS; the opposite was true for fv3, suggesting a possible advantage to using a native-core driving model.
- Forecasts of 2-m AGL temperature and dewpoint were notably less skillful for fv3 than um and wrf; all configurations had a cold bias, while the sign of biases for moisture varied with lead time and had smaller magnitudes.
- Biases in forecasts of 850-mb temperature at 24-h lead time mirrored those found at 2 m AGL: all configurations were cool-biased at both levels, with um least so and fv3 most.

Overall, the skill metrics we examined for these forecast fields and thresholds yielded highly inconsistent rankings for the three model configurations. Furthermore, although we found some indication of UM-Global ICs providing superior forecasts at very short (0–10 h) lead times relative to GFS ICs, there was not evidence of a consistently better driving model across the entire 36-h forecast cycle.

c. Conclusions and future work

Rather than pointing to a particular configuration or driving model as convincingly superior, our verification results instead provide an argument in favor of model configuration diversity in CAM ensembles of opportunity like NCEP's HREF. Based on the skill metrics presented herein, limiting such an ensemble to any single configuration among the three we analyzed would imply sacrificing skill for at least one model field crucial to forecasting convective storms. Furthermore, three of the fields we analyzed (QPF, CREF, and UH) are quite sensitive to the highly nonlinear processes of deep moist convection initiation and evolution, so straightforward postprocessing techniques may not easily compensate for configuration-specific skill deficits represented in these fields. Stochastically perturbed parameterizations (e.g., Jankov et al. 2019; Hirt et al. 2019), however, may soon offer a viable alternative solution for sampling model error without the developmental and operational inefficiencies of maintaining multiple siloed modeling systems.

We reiterate that this study grew opportunistically out of the SFE2020 testbed experiment—future work could build upon our testbed data by conducting even broader, more controlled NWP experiments analogous to ours in a formal research setting where a single institution controls all model configurations, postprocessing, and data flows. Our dataset nonetheless provided a unique side-by-side perspective on three important modeling systems currently used for operational CAMs. An unexpected but valuable result we documented was the um's erroneous diurnal convective cycle over the CONUS, warranting further investigation into which aspect(s) of the configuration might be tunable to improve the timing of convection initiation. Most crucially, our results suggest a particular importance beyond 16–18-h lead times of model configuration diversity in CAM ensembles and ensembles of opportunity: while IC uncertainty may be of primary importance for a watch-to-warning scale application like Warn-on-Forecast (Stensrud et al. 2009, 2013), the next-day convective forecast problem currently demands a robust focus on model uncertainty.

Acknowledgments. Authors BR, BTG, and LJR were provided support by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA21OAR4320204, U.S. Department of Commerce. Authors AJC and ILJ completed this work as part of regular duties at the federally funded National Severe Storms Laboratory and NOAA Storm Prediction Center, respectively. We thank three anonymous reviewers for their detailed feedback and suggestions, which considerably strengthened this manuscript. We also thank the Multi-Radar Multi-Sensor (MRMS) team at the National Severe Storms Laboratory for producing the high-quality dataset we used for verifying forecast radar reflectivity. We thank all SFE2020 participants for their real-time subjective ratings and feedback on CAM ensemble performance, including the model runs analyzed herein. The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the authors and do not necessarily reflect the views of NOAA or the Department of Commerce.

Data availability statement. FV3 model data used in this study were generated by NSSL and may be shared upon request. WRF and UM model data used in this study were generated by NCAR and the Met Office, respectively, and may be shared upon request pending the consent of the dataset creators. SPC local storm reports used to verify surrogate severe forecasts are available publicly at spc.noaa.gov/climo/reports. MRMS data used to verify composite reflectivity forecasts were obtained in real time from NCEP's public FTP server (mrms.ncep.noaa.gov/data). Stage-IV precipitation data were retrieved from NCAR's public archive at data.eol.ucar.edu/dataset/21.087. RTMA data used to verify 2-m AGL temperature and dewpoint forecasts were generated by NCEP/EMC, are archived at NSSL, and may be shared upon request. Rawinsonde data used to verify upper-air temperatures and lapse rates were obtained from the Iowa Environmental Mesonet website (mesonet.agron.iastate.edu/archive/raob).

REFERENCES

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting*, **18**, 918–932, https://doi.org/10.1175/1520-0434(2003)018<0918:SOPFSS>2.0.CO;2.

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.

Atger, F., 2003: Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Mon. Wea. Rev.*, **131**, 1509–1523, https://doi.org/10.1175//1520-0493(2003)131<1509:SAIVOT>2.0.CO;2.

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, https://doi.org/10.1175/MWR-D-15-0242.1.

Best, M. J., and Coauthors, 2011: The joint UK land environment simulator (JULES), model description—Part 1: Energy and water fluxes. *Geosci. Model Dev.*, **4**, 677–699, https://doi.org/10.5194/gmd-4-677-2011.

Boutle, I. A., J. E. J. Eyre, and A. P. Lock, 2014: Seamless stratocumulus simulation across the turbulent gray zone. *Mon. Wea. Rev.*, **142**, 1655–1668, https://doi.org/10.1175/MWR-D-13-00229.1.

Bush, M., and Coauthors, 2020: The first Met Office Unified Model–JULES regional atmosphere and land configuration, RAL1. *Geosci. Model Dev.*, **13**, 1999–2029, https://doi.org/10.5194/gmd-13-1999-2020.

Chen, F., and J. Dudhia, 2001: Coupling an advanced land surface-hydrology model with the Penn State–NCAR MM5 modeling system. Part I: Model implementation and sensitivity. *Mon. Wea. Rev.*, **129**, 569–585, https://doi.org/10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2.

Clark, A. J., 2019: Comparisons of QPFs derived from single- and multi-core convection-allowing ensembles. *Wea. Forecasting*, **34**, 1955–1964, https://doi.org/10.1175/WAF-D-19-0128.1.

——, 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, 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.

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

——, and Coauthors, 2021: A real-time, virtual spring forecasting experiment to advance severe weather prediction. *Bull. Amer. Meteor. Soc.*, **102**, E814–E816, https://doi.org/10.1175/BAMS-D-20-0268.1.

Clark, D. B., and Coauthors, 2011: The joint UK land environment simulator (JULES), model description—Part 2: Carbon fluxes and vegetation dynamics. *Geosci. Model Dev.*, **4**, 701–722, https://doi.org/10.5194/gmd-4-701-2011.

Clark, P., N. Roberts, H. Lean, S. P. Ballard, and C. Charlton-Perez, 2016: Convection-permitting models: A step-change in rainfall forecasting. *Meteor. Appl.*, **23**, 165–181, https://doi.org/10.1002/met.1538.

Coniglio, M. C., J. Correia, P. T. Marsh, and F. Kong, 2013: Verification of convection-allowing WRF Model forecasts of the planetary boundary layer using sounding observations. *Wea. Forecasting*, **28**, 842–862, https://doi.org/10.1175/WAF-D-12-00103.1.

Cullen, M. J. P., 1993: The unified forecast/climate model. *Meteor. Mag.*, **122**, 81–94.

Duda, J. D., X. Wang, 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.

Flack, D. L. A., C. Bain, and J. Warner, 2021: Sensitivity of convective forecasts to driving and regional models during the 2020 Hazardous Weather Testbed. Met Office Forecasting Research Tech. Rep. 649, Met Office, 56 pp., https://www.metoffice.gov.uk/binaries/content/assets/metofficegovuk/pdf/research/weather-science/frtr_649_2021p.pdf.

Gallo, B. T., and Coauthors, 2017: Breaking new ground in severe weather prediction: The 2015 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Wea. Forecasting*, **32**, 1541–1568, https://doi.org/10.1175/WAF-D-16-0178.1.

Gasperoni, N. A., X. Wang, and Y. Wang, 2020: 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. *Mon. Wea. Rev.*, **148**, 1177–1203, https://doi.org/10.1175/MWR-D-19-0124.1.

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.

Hirt, M., S. Rasp, U. Blahak, and G. C. Craig, 2019: Stochastic parameterization of processes leading to convective initiation in kilometer-scale models. *Mon. Wea. Rev.*, **147**, 3917–3934, https://doi.org/10.1175/MWR-D-19-0060.1.

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

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., 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, 2020: Interactions between physics diversity and multiscale initial condition perturbations for storm-scale ensemble forecasting. *Mon. Wea. Rev.*, **148**, 3549–3565, https://doi.org/10.1175/MWR-D-20-0112.1.

——, ——, 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.

Judd, K., C. A. Reynolds, T. E. Rosmond, and L. A. Smith, 2008: The geometry of model error. *J. Atmos. Sci.*, **65**, 1749–1772, https://doi.org/10.1175/2007JAS2327.1.

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.

——, and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931–952, https://doi.org/10.1175/WAF2007106.1.

——, and Coauthors, 2010: Assessing advances in the assimilation of radar data and other mesoscale observations within a collaborative forecasting–research environment. *Wea. Forecasting*, **25**, 1510–1521, https://doi.org/10.1175/2010WAF2222405.1.

Keat, W. J., and Coauthors, 2019: Convective initiation and storm life cycles in convection-permitting simulations of the Met Office unified model over South Africa. *Quart. J. Roy. Meteor. Soc.*, **145**, 1323–1336, https://doi.org/10.1002/qj.3487.

Keil, C., F. Heinlein, and G. C. Craig, 2014: The convective adjustment time-scale as indicator of predictability of convective precipitation. *Quart. J. Roy. Meteor. Soc.*, **140**, 480–490, https://doi.org/10.1002/qj.2143.

Klocke, D., and M. J. Rodwell, 2014: A comparison of two numerical weather prediction methods for diagnosing fast-physics errors in climate models. *Quart. J. Roy. Meteor. Soc.*, **140**, 517–524, https://doi.org/10.1002/qj.2172.

Krocak, M. J., and H. E. Brooks, 2020: An analysis of subdaily severe thunderstorm probabilities for the United States. *Wea. Forecasting*, **35**, 107–112, https://doi.org/10.1175/WAF-D-19-0145.1.

Kühnlein, C., C. Keil, G. C. Craig, and C. Gebhardt, 2014: The impact of downscaled initial condition perturbations on convective-scale ensemble forecasts of precipitation. *Quart. J. Roy. Meteor. Soc.*, **140**, 1552–1562, https://doi.org/10.1002/qj.2238.

Lee, T. R., M. Buban, D. D. Turner, T. P. Meyers, and C. B. Baker, 2019: Evaluation of the High-Resolution Rapid Refresh (HRRR) model using near-surface meteorological and flux observations from northern Alabama. *Wea. Forecasting*, **34**, 635–663, https://doi.org/10.1175/WAF-D-18-0184.1.

Lilly, D. K., 1992: A proposed modification of the Germano subgrid-scale closure method. *Phys. Fluids*, **A4**, 633, https://doi.org/10.1063/1.858280.

Loken, E. D., A. J. Clark, M. Xue, and F. Kong, 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.

Marsigli, C., A. Montani, and T. Paccagnella, 2014: Perturbation of initial and boundary conditions for a limited-area ensemble: Multi-model versus single-model approach. *Quart. J. Roy. Meteor. Soc.*, **140**, 197–208, https://doi.org/10.1002/qj.2128.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.

Mittermaier, M., and N. Roberts, 2010: Intercomparison of spatial forecast verification methods: Identifying skillful spatial scales using the fractions skill score. *Wea. Forecasting*, **25**, 343–354, https://doi.org/10.1175/2009WAF2222260.1.

Morris, M. T., J. R. Carley, E. Colón, A. Gibbs, M. S. F. V. De Pondeca, and S. Levine, 2020: A quality assessment of the Real-Time Mesoscale Analysis (RTMA) for aviation. *Wea. Forecasting*, **35**, 977–996, https://doi.org/10.1175/WAF-D-19-0201.1.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor. Climatol.*, **12**, 595–600, https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Nakanishi, M., and H. Niino, 2004: An improved Mellor–Yamada level-3 model with condensation physics: Its design and verification. *Bound.-Layer Meteor.*, **112**, 1–31, https://doi.org/10.1023/B:BOUN.0000020164.04146.98.

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

Porson, A. N., S. Hagelin, D. F. A. Boyd, N. M. Roberts, R. North, S. Webster, and J. C.-F. Lo, 2019: Extreme rainfall sensitivity in convective-scale ensemble modelling over Singapore. *Quart. J. Roy. Meteor. Soc.*, **145**, 3004–3022, https://doi.org/10.1002/qj.3601.

Putman, W. M., and S.-J. Lin, 2007: Finite-volume transport on various cubed-sphere grids. *J. Comput. Phys.*, **227**, 55–78, https://doi.org/10.1016/j.jcp.2007.07.022.

Raynaud, L., and F. Bouttier, 2016: Comparison of initial perturbation methods for ensemble prediction at convective scale. *Quart. J. Roy. Meteor. Soc.*, **142**, 854–866, https://doi.org/10.1002/qj.2686.

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

——, B. T. Gallo, I. L. Jirak, A. J. Clark, D. C. Dowell, X. Wang, and Y. Wang, 2020: What does a convection-allowing ensemble of opportunity buy us in forecasting thunderstorms? *Wea. Forecasting*, **35**, 2293–2316, https://doi.org/10.1175/WAF-D-20-0069.1.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1.

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, 2019: Revisiting sensitivity to horizontal grid spacing in convection-allowing models over the central and eastern United States. *Mon. Wea. Rev.*, **147**, 4411–4435, https://doi.org/10.1175/MWR-D-19-0115.1.

——, 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.

——, J. Poterjoy, J. R. Carley, D. C. Dowell, G. S. Romine, and K. Ide, 2022: Comparing partial and continuously cycling ensemble Kalman filter data assimilation systems for convection-allowing ensemble forecast initialization. *Wea. Forecasting*, **37**, 85–112, https://doi.org/10.1175/WAF-D-21-0069.1.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., https://dx.doi.org/10.5065/D68S4MVH.

——, and Coauthors, 2021: A description of the Advanced Research WRF Model version 4.3. NCAR Tech. Note NCAR/TN-556+STR, 165 pp., https://doi.org/10.5065/1dfh-6p97.

Smagorinsky, J., 1963: General circulation experiments with the primitive equations: I. The basic experiment. *Mon. Wea. Rev.*, **91**, 99–164, https://doi.org/10.1175/1520-0493(1963)091<0099:GCEWTP>2.3.CO;2.

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.

Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme phenomena in convection-allowing model forecasts. *Wea. Forecasting*, **26**, 714–728, https://doi.org/10.1175/WAF-D-10-05046.1.

——, G. S. Romine, C. S. Schwartz, D. J. Gagne, and M. L. Weisman, 2016: Explicit forecasts of low-level rotation from convection-allowing models for next-day tornado prediction. *Wea. Forecasting*, **31**, 1591–1614, https://doi.org/10.1175/WAF-D-16-0073.1.

Stensrud, D. J., and Coauthors, 2009: Convective-scale Warn-on-Forecast System: A vision for 2020. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1500, https://doi.org/10.1175/2009BAMS2795.1.

——, and Coauthors, 2013: Progress and challenges with Warn-on-Forecast. *Atmos. Res.*, **123**, 2–16, https://doi.org/10.1016/j.atmosres.2012.04.004.

Steptoe, H., N. H. Savage, S. Sadri, K. Salmon, Z. Maalick, and S. Webster, 2021: Tropical cyclone simulations over Bangladesh at convection permitting 4.4 km & 1.5 km resolution. *Sci. Data*, **8**, 62, https://doi.org/10.1038/s41597-021-00847-5.

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

Wandishin, M. S., S. L. Mullen, D. J. Stensrud, 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.

Wilks, D., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

Wilson, D. R., and S. P. Ballard, 1999: A microphysically based precipitation scheme for the UK Meteorological Office Unified Model. *Quart. J. Roy. Meteor. Soc.*, **125**, 1607–1636, https://doi.org/10.1002/qj.49712555707.

Wong, M., and W. C. Skamarock, 2016: Spectral characteristics of convective-scale precipitation observations and forecasts. *Mon. Wea. Rev.*, **144**, 4183–4196, https://doi.org/10.1175/MWR-D-16-0183.1.