Systematic Comparison of Convection-Allowing Models during the 2017 NOAA HWT Spring Forecasting Experiment📎

COREY K. POTVIN,[a,b,c] JACOB R. CARLEY,[d] ADAM J. CLARK,[b,c] LOUIS J. WICKER,[b,c]
PATRICK S. SKINNER,[a,b,c] ANTHONY E. REINHART,[a,b] BURKELY T. GALLO,[a,b] JOHN S. KAIN,[d]
GLEN S. ROMINE,[e] ERIC A. ALIGO,[f] KEITH A. BREWSTER,[g] DAVID C. DOWELL,[h] LUCAS M. HARRIS,[i]
ISRAEL L. JIRAK,[j] FANYOU KONG,[g] TIMOTHY A. SUPINIE,[g] KEVIN W. THOMAS,[g]
XUGUANG WANG,[c] YONGMING WANG,[c] AND MING XUE[c,g]

[a] *Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma*
[b] *NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma*
[c] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*
[d] *NOAA/NWS/NCEP Environmental Modeling Center, College Park, Maryland*
[e] *National Center for Atmospheric Research, Boulder, Colorado*
[f] *IMSG, NOAA/NWS/NCEP/EMC, College Park, Maryland*
[g] *Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma*
[h] *NOAA/Earth System Research Laboratory, Boulder, Colorado*
[i] *NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey*
[j] *NOAA/NWS Storm Prediction Center, Norman, Oklahoma*

(Manuscript received 18 March 2019, in final form 5 July 2019)

ABSTRACT

The 2016–18 NOAA Hazardous Weather Testbed (HWT) Spring Forecasting Experiments (SFE) featured the Community Leveraged Unified Ensemble (CLUE), a coordinated convection-allowing model (CAM) ensemble framework designed to provide empirical guidance for development of operational CAM systems. The 2017 CLUE included 81 members that all used 3-km horizontal grid spacing over the CONUS, enabling direct comparison of forecasts generated using different dynamical cores, physics schemes, and initialization procedures. This study uses forecasts from several of the 2017 CLUE members and one operational model to evaluate and compare CAM representation and next-day prediction of thunderstorms. The analysis utilizes existing techniques and novel, object-based techniques that distill important information about modeled and observed storms from many cases. The National Severe Storms Laboratory Multi-Radar Multi-Sensor product suite is used to verify model forecasts and climatologies of observed variables. Unobserved model fields are also examined to further illuminate important intermodel differences in storms and near-storm environments. No single model performed better than the others in all respects. However, there were many systematic intermodel and intercore differences in specific forecast metrics and model fields. Some of these differences can be confidently attributed to particular differences in model design. Model intercomparison studies similar to the one presented here are important to better understand the impacts of model and ensemble configurations on storm forecasts and to help optimize future operational CAM systems.

1. Introduction

Optimal design of convection-allowing models (CAMs) and CAM ensembles is a challenging and urgent topic in the U.S. weather research and operational communities. In a 2015 report on the state of the National Centers for Environmental Prediction (NCEP) modeling suite, the Model Advisory Committee of the University Corporation for Atmospheric Research Community Advisory Committee for NCEP (UCACN) stated that a CAM ensemble system is a ‘‘critical capability needed in operations to support the weather enterprise’’ and emphasized the need for ‘‘a rational, evidence-driven approach towards decision-making and modeling system environment’’ if the United States is to again be a world

📎 Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/WAF-D-19-0056.s1.

Corresponding author: Corey K. Potvin, corey.potvin@noaa.gov

DOI: 10.1175/WAF-D-19-0056.1

leader in environmental prediction.[1] This ‘‘evidence-driven approach'' to improving current convective-scale and other modeling capability is essential to the strategic vision for NOAA's Next Generation Global Prediction System (NGGPS).

An important source of empirical guidance for the design of CAMs and CAM ensembles (hereafter ‘‘CAM systems'') in the United States is generated as part of the NOAA Hazardous Weather Testbed Spring Forecasting Experiment (SFE) run annually by the National Severe Storms Laboratory (NSSL) and Storm Prediction Center (e.g., Kain et al. 2003; Clark et al. 2012; Gallo et al. 2017). The primary goal of the SFE is to leverage the combined expertise of numerous researchers and forecasters to evaluate and compare experimental CAM systems and to ultimately inform the development of operational systems. Critical to this endeavor is accurate attribution of differences in forecast performance to differences in model and ensemble design.

Before 2016, the wide range of configurations among the CAM systems contributed by the participating modeling groups had historically made systematic intercomparisons difficult. Thus, in recognition of the need for more uniformity among the CAM systems evaluated during the SFE, and further motivated by the UCACN Model Advisory Committee recommendations, the Community Leveraged Unified Ensemble (CLUE) was initiated in 2016 (Clark et al. 2018). The CLUE is a carefully designed ‘‘superensemble'' comprising CAM systems that, while still contributed by different groups, use collectively chosen settings that enable systematic evaluation of the impacts of different configuration strategies (e.g., multiphysics ensembles; initialization using radar data assimilation; employing stochastic physics perturbations). During the 2016–18 SFEs, all CLUE members were initialized daily at 0000 UTC, used similar domains encompassing the contiguous United States (CONUS), and were interpolated and postprocessed to the same 3-km grid using similar versions of the NCEP/Developmental Testbed Center Unified Post-Processor software.[2] The present study evaluates forecasts from several members of the 2017 CLUE, which ran 1 May–2 June 2017 and included 81 members contributed by NOAA laboratories: NSSL, Geophysical Fluid Dynamics Laboratory (GFDL), and Earth System Research Laboratory (ESRL); as well as academic research institutions: the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma (OU), National Center for Atmospheric Research (NCAR), and the OU Multiscale data Assimilation and Predictability Laboratory (OU-MAP). These models use the three major dynamical cores currently or very soon to be used in U.S. operational numerical weather prediction: the Advanced Research version of the Weather Research and Forecasting (ARW) Model (Skamarock et al. 2008; Powers et al. 2017), the Nonhydrostatic Multiscale Model on the B-grid (NMMB; Janjić and Gall 2012), and the Finite-Volume Cubed-Sphere Dynamical Core (FV3; Putman and Lin 2007; Harris and Lin 2013) developed at GFDL. The FV3 recently replaced the Global Spectral Model core in NCEP's Global Forecast System (GFS). Motivated by NOAA's vision of a unified modeling system, a regional FV3 system has begun to be tested against existing regional forecasting systems to facilitate the development of a unified forecasting system for all operational NWP in the United States (e.g., Clark et al. 2018).

This study leverages output from the 2017 CLUE to achieve three primary objectives:

1) Develop and demonstrate new object-based techniques for examining CAM representation and prediction of storms and near-storm environments (NSEs).
2) Use these new techniques along with existing verification methods to identify systematic forecast differences between several contemporary CAMs.
3) Based on identified intermodel differences, infer model configuration impacts on the representation and forecast accuracy of CAM storms and NSEs.

We analyze 18–26-h forecasts initialized at 0000 UTC daily during the 2017 SFE to focus on the period of peak convective activity. The NSSL Multi-Radar Multi-Sensor (MRMS; Smith et al. 2016) system is used to evaluate the realism and forecast accuracy of simulated storms. The analysis is restricted to the 17 days for which sufficient output are available from all the models and the MRMS system during the 2017 SFE: 1 May, 3–5 May, 9 May, 12 May, 15–19 May, 23 May, 25–26 May, 29 May, 31 May, and 2 June. In addition to observed, verifiable model fields, we examine several unobserved fields (e.g., vertical velocity) to further illuminate intermodel differences in representation of storms and NSEs. This enables the identification of many systematic behaviors of different model configurations.

2. CLUE and MRMS output

a. Model configurations

Six CLUE models and one operational model were used in our analyses. Highlights of each model follow; additional details can be found in Table 1 and the 2017

[1] The full report is available at www.ncep.noaa.gov/director/ucar_reports/ucacn_20151207/UMAC_Final_Report_20151207-v14.pdf.

[2] Available at https://dtcenter.org/community-code/unified-post-processor-upp.

Table 1. Configuration details for each model. The "No. levels lowest 100 mb" column lists the number of vertical levels within the lowest 100 mb assuming a 1000-mb model surface.

	Core	IC	BC	No. levels lowest 100 mb	Microphysics	LSM	PBL
fv3-gfdl	FV3	GFS	Nested within global FV3	11	GFDL 6-category	Noah	Hong and Pan
fv3-caps	FV3	GFS	Nested within global FV3	11	Thompson	Noah	Hong and Pan
ncar_01	ARW	15-km EnKF DA—only parent model that is continuously cycled	Pert GFS	7	Thompson	Noah	MYJ
hrrre_01	ARW	RAPv4 mean; GDAS perturbations; 3-km EnKF DA including radar	RAP	8	Thompson	RUC	MYNN
caps-core_01	ARW	NAM-12km; 3DVAR and cloud analysis	NAM-12km	8	Thompson	Noah	MYJ
ou-map_01	NMMB	3DEnVar direct assimilation of both radar and operational in situ observations	GFS	8	Ferrier-Aligo	Noah	MYJ
nam3km	NMMB	Hybrid 3DEnVar using GDAS EnKF members + radar DFI	NAM-12km parent domain	17	Ferrier-Aligo	Noah	MYJ

SFE Operations Plan.[3] We used both FV3-based models that were included in the 2017 CLUE. One was contributed by GFDL, and the other by CAPS; hereafter, these are labeled "fv3-gfdl" and "fv3-caps," respectively. Both versions used a 3-km domain covering CONUS that was two-way interactively nested within the simultaneously run global FV3 grid with a mean spacing of ~13 km (Snook et al. 2019). The fv3-caps was configured identically to the fv3-gfdl except that it used Thompson microphysics (Thompson et al. 2004, 2008) rather than the GFDL 6-category microphysics (Chen and Lin 2013). The remaining physics parameterizations were adopted from the GFS, except that cumulus parameterization was disabled on the 3-km grid. Both FV3-based models were initialized directly from GFS analyses. We selected three ARW-based models from the CLUE: member 1 of the "core" CAPS ensemble ("caps-core_01"), member 1 of the ESRL High-Resolution Rapid Refresh (Benjamin et al. 2016) Ensemble (HRRRE; Dowell et al. 2016; "hrrre_01"), and member 1 of the NCAR forecast ensemble (Schwartz et al. 2019; "ncar_01").[4] All three ARW-based models used Thompson microphysics, but different land surface models (LSMs) and planetary boundary layer (PBL) physics schemes: either the Rapid Update Cycle (RUC; Smirnova et al. 2016) LSM with the Mellor–Yamada–Nakanishi–Niino (MYNN; Nakanishi and Niino 2004, 2006) PBL scheme, or the Noah LSM (Mitchell et al. 2005) with the Mellor–Yamada–Janjić (MYJ; Janjić 2002) PBL scheme. The initialization procedures differed greatly among the three ARW models: ncar_01 was downscaled from the 15-km NCAR ensemble Kalman filter (EnKF; Evensen 2003) analysis assimilating conventional observations only, with no data assimilation on the 3-km grid; the caps-core_01 initial condition (IC) was generated using 3D variational assimilation of conventional observations and radial velocity data followed by cloud analysis-based assimilation of reflectivity data with the NAM analysis as the background; and the hrrre_01 IC was generated using direct EnKF assimilation of both conventional and radar reflectivity observations on the 3-km grid using the NCAR Data Assimilation Research Testbed (DART; Anderson et al. 2009) toolkit.

The only NMMB-based model output included in the 2017 CLUE was from the ensemble produced by OU-MAP (Wang and Wang 2017). To have more than one NMMB-based model in our analysis, we used both member 1 of the OU-MAP ensemble ("ou-map_01") and the operational North American Model (NAM) 3-km CONUS nest ("nam3km"). Both models used similar physics schemes, including the Ferrier–Aligo microphysics (Aligo et al. 2018), but very different initialization procedures. The nam3km features an

[3] Available at https://hwt.nssl.noaa.gov/Spring_2017/HWT_SFE2017_operations_plan_FINAL.pdf.

[4] It should be borne in mind that individual members of formal ensembles are not representative of the forecast skill of the entire ensemble (e.g., Schwartz et al. 2014).

TABLE 2. Descriptions of variables. Asterisks denote variables with corresponding MRMS products.

Variable	Description
REFLCOMP*	Composite reflectivity (dBZ)
UH_1HMAX*/ROT_1HMAX	Hourly maximum 2–5-km updraft helicity ($m^2 s^2$)/hourly maximum azimuthal shear (s^{-1})
RAIN_1H*	Hourly accumulated rainfall (in.)
W_1HMAX	Hourly maximum vertical velocity below 400 mb ($m s^{-1}$)
SBCAPE	Surface-based convective available potential energy ($J kg^{-1}$)
SRH	0–3-km storm relative helicity ($m^2 s^{-2}$)
T_2M	2-m temperature (K)
TD_2M	2-m dewpoint (K)
DIV_10M	10-m horizontal divergence (s^{-1})
WIND_10M	10-m wind speed ($m s^{-1}$)

hourly hybrid three-dimensional ensemble–variational (hybrid 3DEnVar) analysis that ingests the Global Data Assimilation System (GDAS) EnKF ensemble perturbations and assimilates conventional and satellite observations, then uses a cloud analysis and latent heating initialization to assimilate radar reflectivity (Wu et al. 2017; Gustafsson et al. 2018). There are two primary differences between ou-map_01 and nam3km. In ou-map_01, a 3DEnVar system ingests a 3-km NMMB ensemble initialized by NMMB's own GSI-based EnKF, which bypasses the use of coarser global ensemble members as in the nam3km. In addition, ou-map_01 directly assimilates both radar reflectivity and velocity within the EnVar algorithm (Johnson et al. 2015; Wang and Wang 2017).

As the preceding description illustrates, the seven models examined in this study vary in a number of ways, including initialization procedure, lateral boundary conditions, dynamical core, and physics schemes. While it is therefore important to conceive of each associated set of forecasts as arising from a different model *system* comprising distinct methods for generating the initial condition, propagating this analysis forward in time, and setting the boundary conditions for this forecast solution, we refer to each forecast system examined herein as a ''model'' for brevity. Despite the multifold differences among these models, many of the systematic differences between their forecasts can be confidently attributed to particular model design choices (section 4).

b. Model and MRMS products

A variety of model output fields are analyzed; these are listed in Table 2. Model composite reflectivity (REFLCOMP) and hourly accumulated rainfall (RAIN_1H) can be directly verified with MRMS products. The MRMS REFLCOMP product we use is computed by calculating the exponential inverse-distance-weighted average of the reflectivity from each contributing radar and then taking the column maximum as described in Smith et al. (2016). MRMS RAIN_1H is obtained using the local gauge bias-corrected radar quantitative precipitation estimation method described in Zhang et al. (2016). While CLUE hourly maximum updraft helicity (UH; UH_1HMAX; Kain et al. 2010) cannot be directly verified, we follow previous studies (e.g., Carlin et al. 2017; Skinner et al. 2018) and compare to the MRMS rotation track product valid for the 2–5 km AGL layer (Miller et al. 2013; ROT_1HMAX). MRMS rotation track products are generated from hourly accumulated maximum azimuthal shear calculated using the linear least squares derivative method of Mahalik et al. (2019).

ROT_1HMAX uses a 0.5-km grid and the remaining MRMS products use a 1-km grid. To facilitate comparisons with the 3-km CLUE output, we apply a 36-point and 9-point box linear filter to the 0.5- and 1-km MRMS products, respectively.

3. Analysis methods

We combine five analysis methods targeting a wide range of storm attributes and forecast metrics to produce a comprehensive framework for evaluating and comparing CAMs. While this study focuses on next-day forecasts (more precisely, the 18–26-h forecast period valid from 1800 UTC of the current day to 0200 UTC

TABLE 3. The 99.9th percentiles of REFLCOMP and UH_1HMAX/ROT_1HMAX valid for the dates/times used in this study.

	REFLCOMP (dBZ)	UH_1HMAX or ROT_1HMAX ($m^2 s^{-2}$ or s^{-1})
mrms	44	0.003
fv3-gfdl	47	58
fv3-caps	49	58
ncar_01	53	27
hrrre_01	50	27
caps-core_01	53	30
ou-map_01	48	30
nam3km	48	26

TABLE 4. Means and (in parentheses) standard deviations of NSE-maximum RAIN_1H (in.) in the ALL_STORMS and ISOLATED_STORMS datasets.

	mrms	fv3-gfdl	fv3-caps	Ncar _01	hrrre_01	caps-core_01	ou-map_01	nam-3km
ALL STORMS	1.12 (0.46)	0.87 (0.54)	1.45 (0.82)	1.29 (0.76)	1.32 (0.68)	1.38 (0.73)	1.62 (0.81)	1.65 (0.78)
ISOLATED STORMS	0.95 (0.47)	0.52 (0.40)	0.95 (0.63)	0.83 (0.56)	0.89 (0.68)	0.86 (0.59)	1.13 (0.68)	1.35 (0.80)

of the following day), the evaluation framework is applicable to forecast lead times from $O(1)$ h to $O(1)$ week. Most thunderstorm forecast evaluation techniques are designed to measure the spatial correspondence between observed and model-predicted storms for a particular forecast lead time or period. The surrogate severe (Sobash et al. 2011, 2016; section 3a) and object-based verification (Davis et al. 2006a,b; section 3b) methods are leading examples of this approach. It can be fruitful, however, to examine and compare *model climatologies of storm characteristics* (e.g., Clark et al. 2014; Skinner et al. 2018), including characteristics that cannot be quantitatively verified (e.g., UH). This type of analysis provides an objective way to assess how realistically a particular model simulates observed storm fields and to identify systematic intermodel differences in many aspects of storm structure and behavior, including storm–environment interactions. Using an object-based framework (section 3c), we apply two traditional statistical techniques in novel ways to verify and compare the representation of storms and NSEs in our seven CAMs (sections 3d–3e).

a. Surrogate severe verification

The surrogate severe verification method (Sobash et al. 2011, 2016) measures the spatial correspondence between extreme model forecast values (typically of UH, which indicates rotating thunderstorms) and severe weather reports. We perform surrogate severe verification of our seven models in a manner consistent with previous studies. First, the UH_1HMAX from each model is regridded to an 80-km grid by assigning each 80-km grid cell the maximum 3-km UH_1HMAX within it over the forecast period (1800–0200 UTC). Then, surrogate severe fields are created by assigning a value of one (zero) to 80-km grid cells exceeding (not exceeding) the prescribed UH_1HMAX threshold. A corresponding 80-km local storm reports (LSRs) grid is assigned values of one (zero) at cells containing (not containing) at least one tornado, hail, or wind report. Next, the binary surrogate severe and LSR fields are smoothed using a Gaussian kernel with prescribed standard deviation σ to create probability fields (Theis et al. 2005). The σ used to generate the surrogate severe forecast probabilities is varied; $\sigma = 120$ km is used to generate the ‘‘practically perfect probabilities’’ from the LSR grid (Brooks et al. 2003; Hitchens et al. 2013). Finally, the surrogate severe forecast probabilities are verified against the practically perfect probabilities using spatial and contingency table verification metrics. In this study, the area under the relative operating characteristic curve (AUC) and fractions skill score (FSS; Roberts and Lean 2008) are computed for 100 UH_1HMAX percentile thresholds and 53 values of σ (section 4a). Each UH_1HMAX percentile is the value below which falls a prescribed percentage of the 80-km-grid UH_1HMAX distribution.

b. Object-based verification of reflectivity forecasts

Object-based methods allow verification to focus upon discrete features (e.g., storms) of greatest interest to the user, and avoid traditional methods’ unduly large penalty for phase errors. Skinner et al. (2018) tailored the object identification and matching technique in the Method for Object-based Diagnostic Evaluation (MODE; Davis et al. 2006a,b) to verification of model-predicted reflectivity objects. We adopt the Skinner et al. (2018) method herein to measure each model’s skill in forecasting the general spatial distribution of storms. Storm objects[5] are objectively identified by thresholding the MRMS and model REFLCOMP on their respective 99.9th percentiles (Table 3) from the full 17-day sample used in this study. Objects with area $< 100\,\mathrm{km}^2$ (12 grid cells) are discarded, and remaining objects whose boundaries lie within 10 km of each other are merged (i.e., combined into a single object).

Storm objects from each model are matched to the MRMS storm objects using a total interest function as in the MODE. Unlike in MODE, however, only one forecast object is allowed to be matched to each observed object so as not to unduly reward overprediction of storms. Total interest is calculated by averaging the boundary and centroid displacements then dividing the result by a prescribed maximum allowable displacement between forecast and observed object boundaries/centroids. This maximum allowable displacement is set to 250 km; intermodel differences in

[5] Object identification, extraction, and characterization in this study were performed using the Python Scikit-image library (Van der Walt et al. 2014).

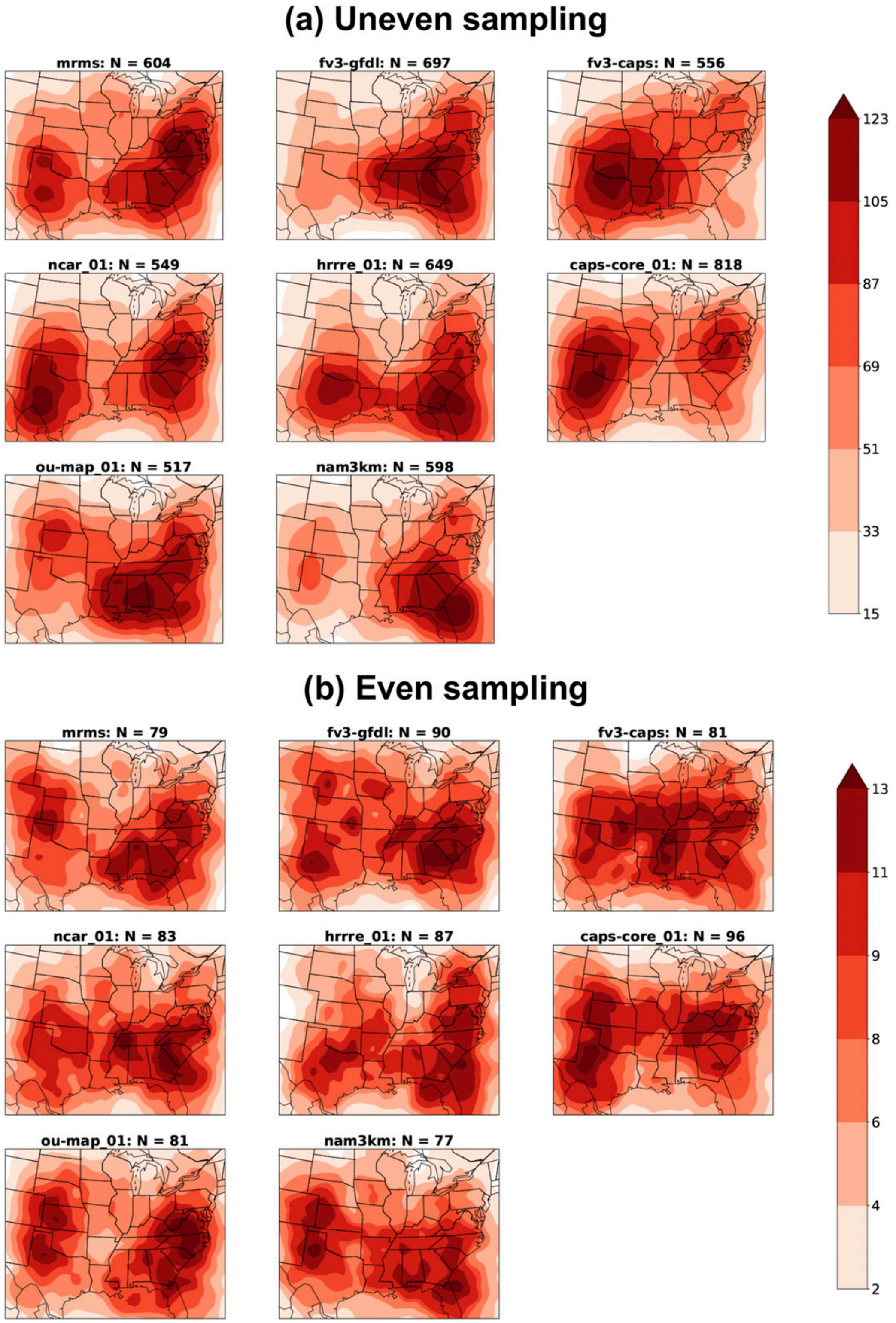

FIG. 1. Spatial distributions of storm objects (a) before and (b) after criteria 4 and 5 were imposed to reduce sampling bias. Storm object counts were computed within 500 km of each point on a 50-km grid, then filtered using a Gaussian kernel with $\sigma = 50$ km. Here and in subsequent figures, N indicates the number of storm objects for each dataset.

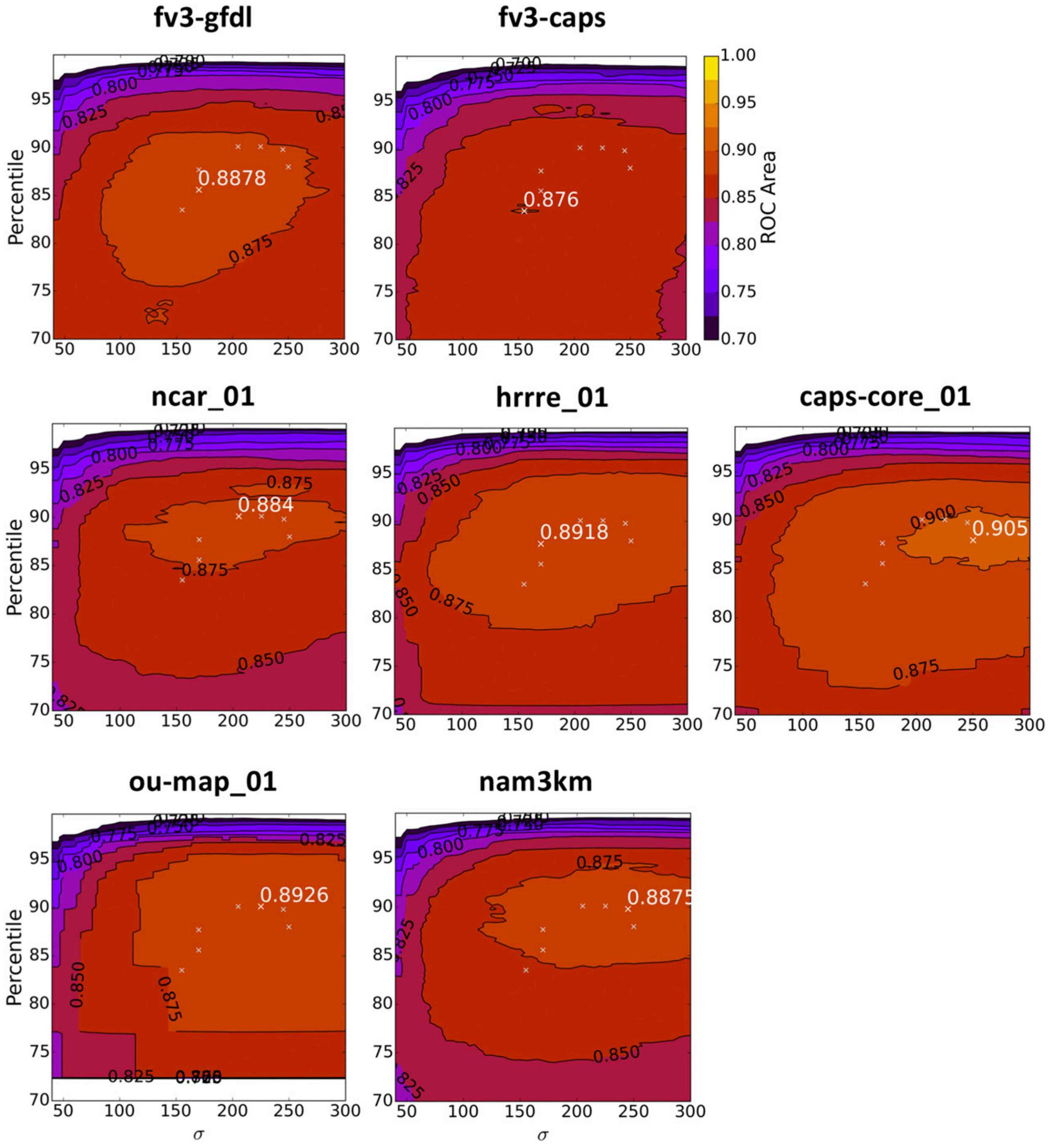

Fig. 2. AUCs of surrogate severe probability forecasts for different Gaussian kernel widths σ and UH percentile thresholds. The maximum AUC for each model is indicated in white. The location of the maximum AUC for every model is plotted as an asterisk in each panel. Note that the OU-MAP UH values were output in 10 $\mathrm{m^2\,s^{-2}}$ increments.

verification metrics were qualitatively insensitive to this threshold. Reflectivity objects are considered matched if the total interest score exceeds 0.2; for example, an object pair with a 200-km centroid displacement and 0-km boundary displacement (or vice versa) would be considered a match. If multiple forecast objects are matched to a single observed object, the pair with the highest total interest score is retained and other forecast objects classified as unmatched. Each reflectivity object is classified as a matched pair (hit), unmatched forecast object (false alarm), or unmatched observed object (miss). This permits computation of contingency table verification metrics,

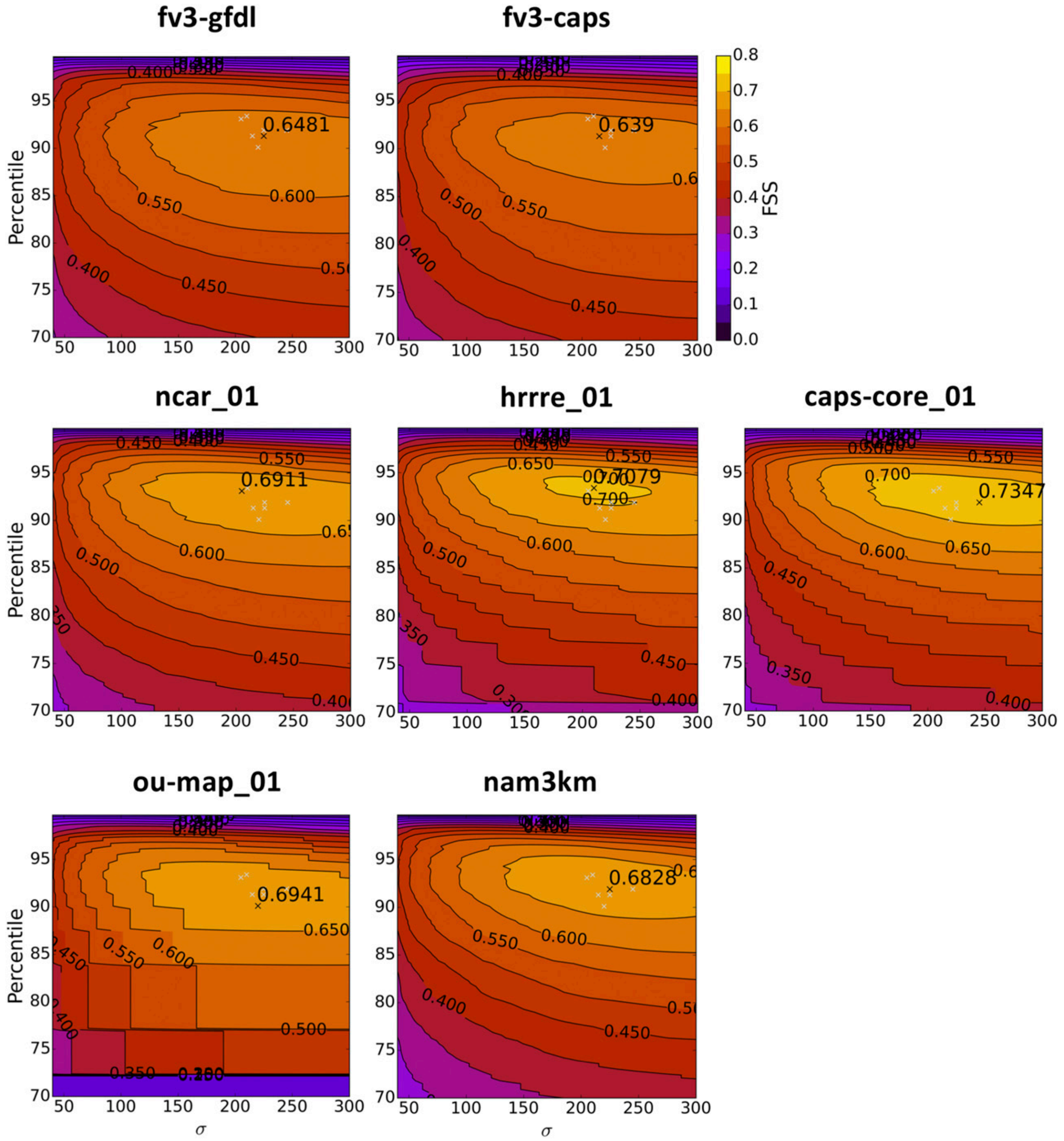

FIG. 3. FSSs of surrogate severe probability forecasts for different Gaussian kernel widths σ and UH percentile thresholds. The maximum FSS for each model is indicated in white. The location of the maximum FSS for every model is plotted as an asterisk in each panel. Note that the OU-MAP UH values were output in 10 $m^2 s^{-2}$ increments.

including probability of detection (POD), false alarm rate (FAR), critical success index (CSI), and frequency bias, all of which can be represented simultaneously on a performance diagram (Roebber 2009; section 4b).

The set of storm objects used in this verification is hereafter labeled ALL_STORMS to distinguish it from the set of storm objects described in section 3c, which is used in the remaining verification methods.

c. Storm object creation

The remaining three analysis methods are designed to illuminate intermodel differences in storm structure and intensity and in the NSE. We restrict these analyses to discrete, quasi-isolated storms to avoid conflating intermodel differences associated with different convective modes and to facilitate identification of convective feedbacks to the NSE. Storm objects are identified as follows. First, preliminary objects are created by thresholding REFLCOMP on the 99.9th percentile and imposing a minimum area threshold of 100 km^2 (criterion 1; Table 3). Then, objects are discarded if they are larger than 2500 km^2 or longer than 100 km (criterion 2). Next, objects are discarded if REFLCOMP $>$ 10 dBZ within more than half the NSE, defined herein as the 120-km-wide domain centered on the storm-maximum REFLCOMP (criterion 3). Criterion 2 filters the few MCSs that are not discarded by criterion 3. Criterion 3 filters most MCSs and cases where the NSE is heavily contaminated by surrounding convection. The use of percentile rather than fixed-value thresholds was largely motivated by the occasionally substantial differences between model climatologies, which can be readily identified by the large intermodel differences in percentiles (e.g., Table 3).

Given the brevity of our analysis period, substantial intermodel differences in analyses could arise from uneven sampling of storm environments. For example, on a given date, two models could produce numerous storms in very different regions or at very different times. If the environments sampled by the two models in this scenario were very different, any intermodel analysis differences could be due much more to the environmental differences (and resulting differences in storm characteristics) than to differences between the models themselves. To avoid conflating these two differences, we imposed an additional condition: storm objects must exist within 1000 km and 2 h of at least one storm object centroid from at least four of the six other (model and MRMS) datasets (criterion 4). Now, consider the scenario where all or most of the models produce storms in the same general region during a convective event, but one model produces many more storms than the others. This event would then be more strongly represented in analyses from the active model than from the other models, again impeding interpretation of intermodel differences. To avoid this scenario, we adopted a final condition: at each time, storm objects satisfying criterion 4 are filtered such that no two objects from the same dataset on the same date lie within 500 km of each other (criterion 5). Criterion 5 has the added advantage of leveling the influence of each convective event on the analyses. For convenience, we summarize the set of all storm object criteria below:

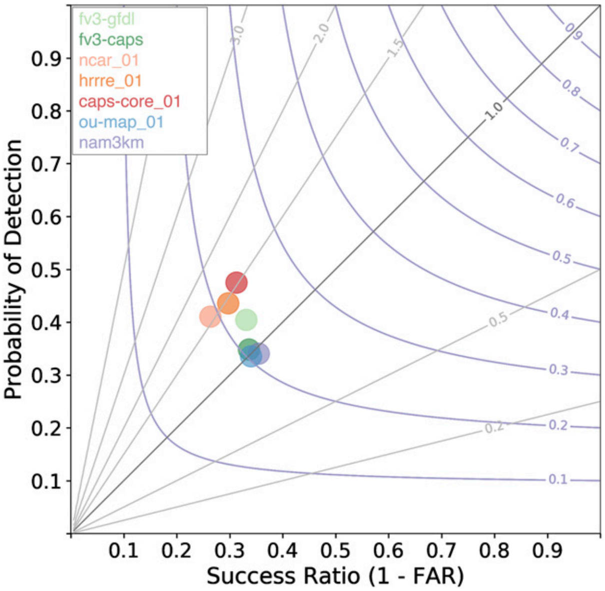

FIG. 4. Performance diagram for reflectivity object forecasts. CSI and frequency bias are contoured in blue and gray, respectively.

1) REFLCOMP exceeds the 99.9th percentile over a contiguous region with area $>$ 100 km^2;
2) This REFLCOMP (storm) object has length $<$ 100 km and area $<$ 2500 km^2;
3) REFLCOMP $>$ 10 dBZ within less than half the NSE;
4) The storm object lies within 1000 km and 2 h of at least one storm object centroid from at least four of the six other datasets;
5) The storm object does not lie within 500 km of another object from the same dataset on the same date.

We label this set of storms ISOLATED_STORMS to distinguish it from ALL_STORMS (section 3b). To provide a general idea of the intensity of the storms in both analysis datasets, we present the mean and standard deviation of NSE-maximum RAIN_1H, valid over all forecast periods used in this study, for the MRMS data and each model (Table 4).

Figure 1 illustrates the impact of criteria 4 and 5 on the spatial distributions of analyzed storms. The geographical distributions of storms are substantially more similar across the datasets upon imposing these criteria (cf. Figs. 1a,b) . Imposing these criteria produced some substantial changes in intermodel differences (not shown). This underscores the importance of minimizing sampling errors when performing intermodel comparisons of potentially disparate sets of storms. This is especially crucial when the number of convective episodes in the analysis is relatively small since undesirable differences arising

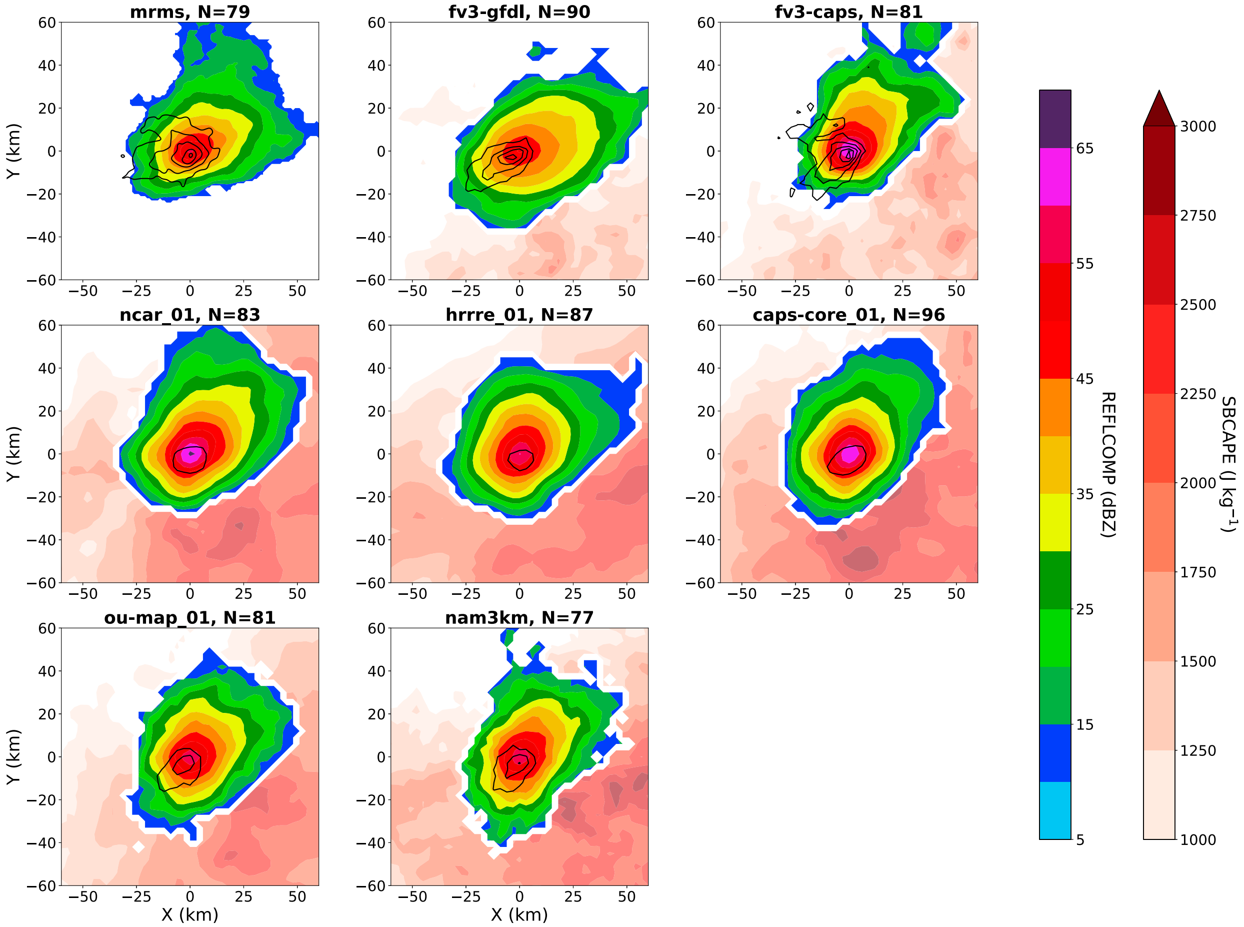

FIG. 5. PMMs of REFLCOMP (dBZ; multicolor shading), SBCAPE (J kg^{-1}; red shading; only shown where REFLCOMP $<$ 10 dBZ), and ROT_1HMAX (contoured every 0.001 s^{-1} starting at 0.001 s^{-1}) or UH_1HMAX (contoured every 25 m^2 s^2 starting at 25 m^2 s^2).

from sampling different storm populations could dominate differences arising from systematic intermodel differences.

d. Probability-matched means

Fully understanding differences in CAM representation of storms and NSEs requires detailed analysis of the spatial structure of both observed and unobserved model fields. While subjective inspection of individual cases is indispensable to this process, the identification and communication of intermodel differences can be greatly facilitated by objective compositing methods. The most obvious compositing option is to average prescribed model fields over many cases, and then compare these averages between models and, when possible, observation-derived products like MRMS. Simple averaging, however, excessively smooths important structural characteristics of storms and NSEs due to the considerable case-to-case variability among storms.

We therefore calculate probability-matched means (PMMs; Ebert 2001), which preserve the average probability density function (PDF) of the individual cases and thereby exhibit much more realistic storm amplitudes and gradients than simple means do. The PMM of a given variable for a given dataset is generated as follows:

1) Average over all cases to obtain the 2D mean field.
2) For each case, create an ordered 1D array of all field values, then average over all cases.
3) Assign the highest value of the mean ordered array (from step 2) to the location of the highest value in the 2D mean (from step 1), and so on for all elements of the mean ordered array.

Of course, given the large variability of storms, any deterministic spatial representation (PMM or otherwise) will poorly represent many of the storms and will be overly spatially smooth. Despite this and other potential limitations (Surcel et al. 2014; Clark 2017), the PMMs prove effective in encapsulating important intermodel

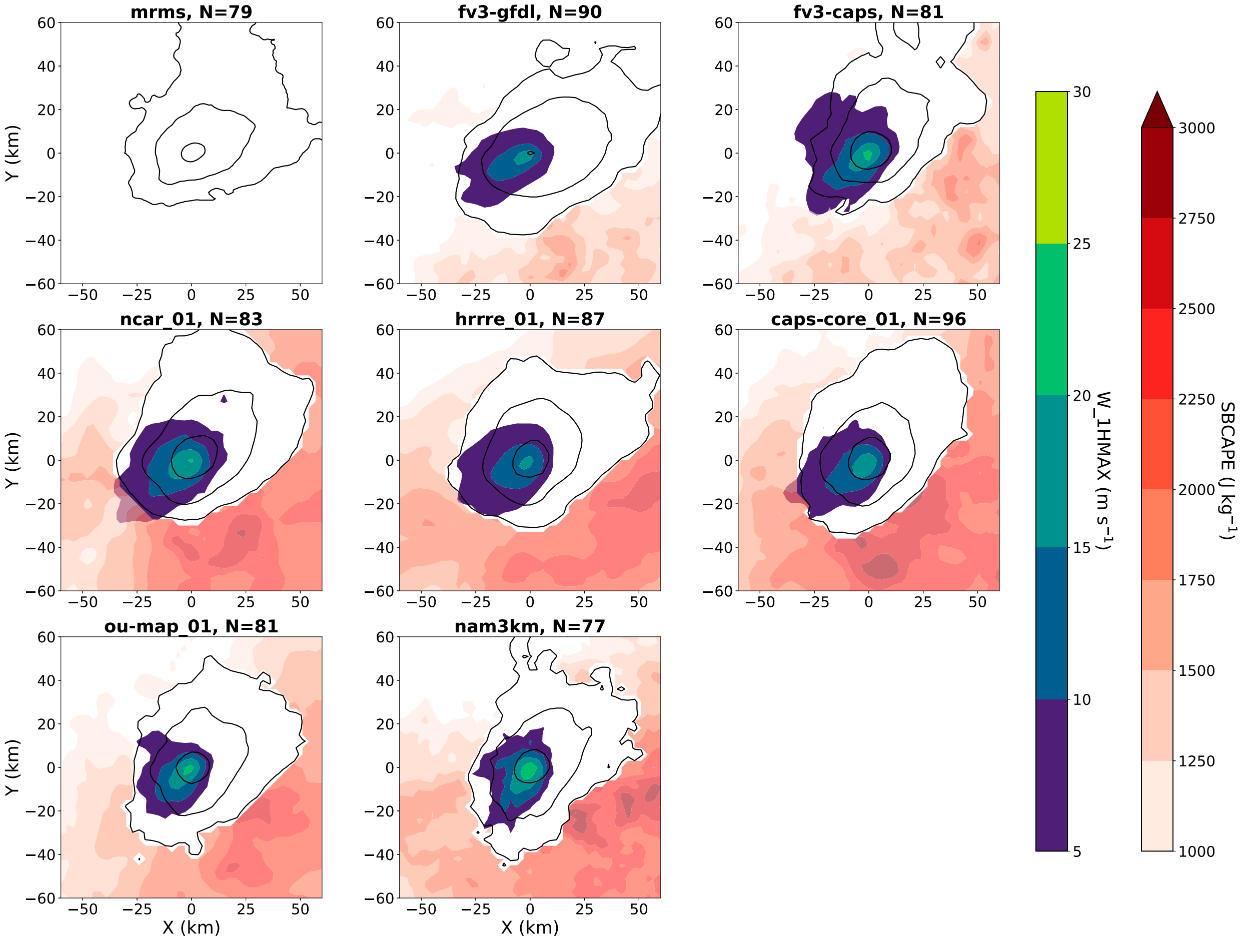

FIG. 6. PMMs of WMAX_1H (m s^{-1}; blue-green shading), REFLCOMP (contoured at 10, 30, 50 dBZ), and SBCAPE (J kg^{-1}; red shading; only shown where REFLCOMP < 10 dBZ).

differences in storms and NSEs (section 4c) that are plainly seen in individual cases.

e. Kernel density estimation

The PMMs alone provide a limited view of the distributions of observed and model-predicted storms, particularly for nonnormal variables. To gain a more complete view of how storms and NSEs vary among the models, we compare PDFs of case-maximum field (e.g., UH, CAPE) values (recall that for each case, the analysis domain is 120 km wide and centered on the storm-maximum REFLCOMP; section 3c). The PDFs are computed using kernel density estimation (KDE), a simple, nonparametric method that more accurately estimates underlying distributions than do histograms, which are sensitive to the prescribed bin intervals and endpoints and exhibit higher variance for small samples. Univariate and bivariate KDE are performed using a Gaussian kernel with bandwidth computed by Scott's Rule. The univariate KDE plots reveal important intermodel differences that cannot be discerned from the PMMs in cases of highly nonnormal distributions, and the bivariate KDE plots reveal important intermodel differences in relationships between storm and/or NSE fields (section 4d).

4. Results

We now present results of each analysis method in the order in which they were described in section 3.

a. Surrogate severe probability forecast verification

Major intermodel differences did not occur in surrogate severe forecast performance. The maximum AUCs were comparable, as were the AUCs at UH–σ points near the maximum AUC (Fig. 2). The fv3-caps (caps-core_01) AUCs were lower (higher) than for the other models, but the absolute differences were small (~0.01–0.02). Larger intermodel differences occurred in FSS (Fig. 3), with both FV3 models lagging the other

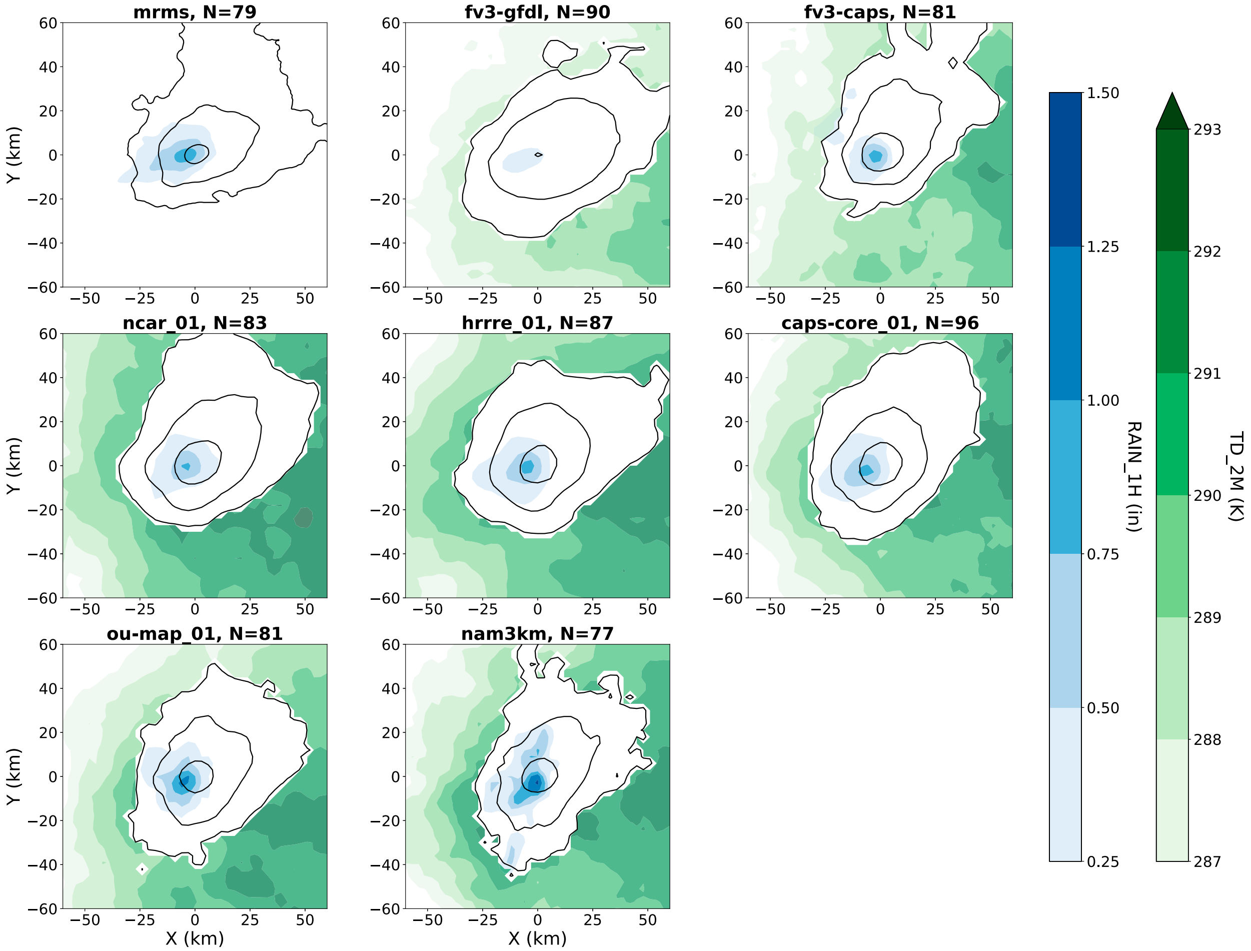

FIG. 7. PMMs of RAIN_1H (in.; blue shading), REFLCOMP (contoured at 10, 30, 50 dBZ), and TD_2M (K; green shading; only shown where REFLCOMP < 10 dBZ).

models by ~0.05–0.1. In summary, the FV3 model UH forecasts corresponded slightly less well with severe storm reports than did the other CLUE models, and the caps-core_01 performed the best. The NMMB and ARW models performed similarly to one another, underscoring their years of development at CAM scales. Given that these verification statistics are valid for only 17 days spanning a period of just a month, however, quantitative intermodel differences in scores should not be interpreted too generally.

b. Object-based verification of reflectivity forecasts

Applying object matching verification to ALL_STORMS (section 3b) produced similar CSIs for all seven models (~0.19–0.22; Fig. 4). The ARW models, however, had substantially larger PODs and FARs. We attribute these differences to the larger frequency bias in these models (~1.5 versus ~1.0). These frequency biases cannot be attributed solely to the ARW models' use of Thompson microphysics since the fv3-caps (which also used the Thompson scheme) had almost no frequency bias.

Interestingly, this ARW model frequency bias does not manifest in counts of the discrete, quasi-isolated storm objects (ISOLATED_STORMS) identified using criteria 1–3 in section 3c (Fig. 1a). Relaxing criteria 2 and 3 to better match the ALL_STORMS criteria results in frequency biases of ~1.3–1.5 for the three ARW models and biases very close to unity for the remaining models (not shown). These frequency biases are much more consistent with those for ALL_STORMS (Fig. 4), confirming that applying criteria 2 and 3 substantially reduces the ARW frequency bias in ISOLATED_STORMS. We remind the reader that the ISOLATED_STORMS database is used in the remaining analyses (sections 4c–4d).

To determine how much of the ARW frequency bias in ALL_STORMS is associated with MCSs versus cases where the NSE is contaminated by other storms, we recomputed the biases for all models upon discarding

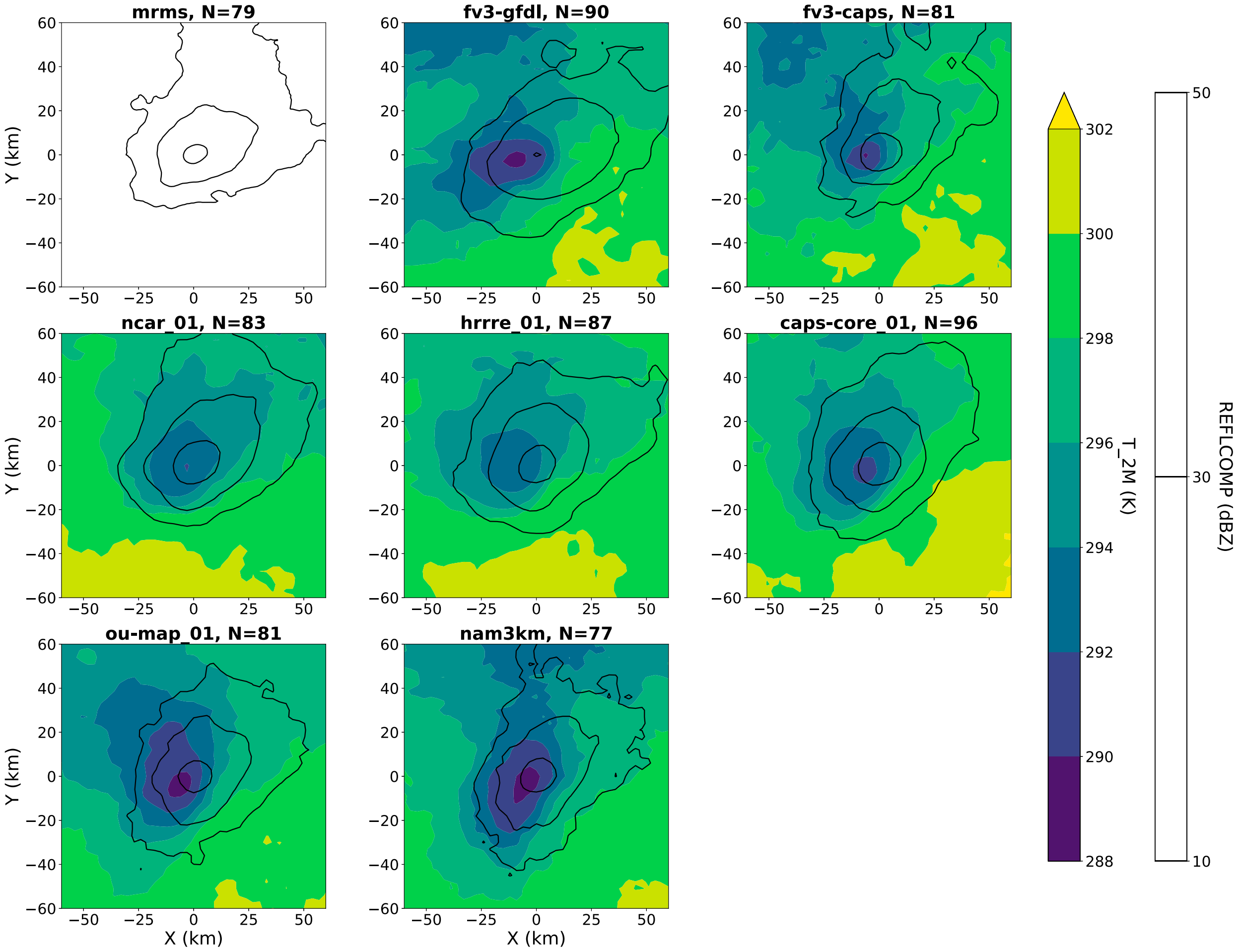

FIG. 8. PMMs of T_2M (K; shading) and REFLCOMP (contoured at 10, 30, 50 dBZ).

objects lying within regions where REFLCOMP exceeds the 95th percentile—around 25 dBZ—over a contiguous area of ≥ 2500 km^2. Case-by-case inspection confirmed this new condition achieved the intended effect of discarding MCS cases while retaining multistorm cases. Removing the MCS cases caused the frequency biases to increase by ~0.3 (relative to satisfying criterion 1 alone) for all models. It follows that all of the models, but especially the ARW models, tended to produce storm clusters with unrealistically large numbers of storms.

c. *Probability-matched means*

The PMMs (Figs. 5–10) revealed many important intermodel differences in storm and NSE characteristics for ISOLATED_STORMS that were confirmed through inspection of individual cases (plots of several model fields for each of the hrrre_01 and fv3-gfdl cases are available in the online supplemental material). Storm precipitation regions were more zonally oriented in the fv3-gfdl than in the other models, in better agreement with observations (Fig. 5). Storms in the fv3-caps were substantially more meridionally oriented than in the fv3-gfdl, which likely arose from its use of Thompson microphysics (rather than the GFDL 6-category scheme) since that was the only difference between the two models. The similarity of PMMs of 500-mb (1 mb = 1 hPa) wind vectors for the two models (not shown) confirms that uneven sampling of storm environments is not the primary source of this difference. Two fv3-gfdl REFLCOMP biases that were subjectively discerned by participants of the 2017 SFE—a low bias in high-percentile values, and a high bias in medium-percentile values—are also apparent in the PMMs. Positive REFLCOMP biases at higher values are apparent in the fv3-caps, ncar_01, and caps-core_01 PMMs. Model-dependent biases like these are a major motivation for the use of percentile (rather than fixed-value) storm object thresholds in this study.

UH_1HMAX was substantially lower in the ARW models than in the NMMB and FV3 models (Fig. 5). To compute $w\zeta$ (the integrand of the UH formula) at

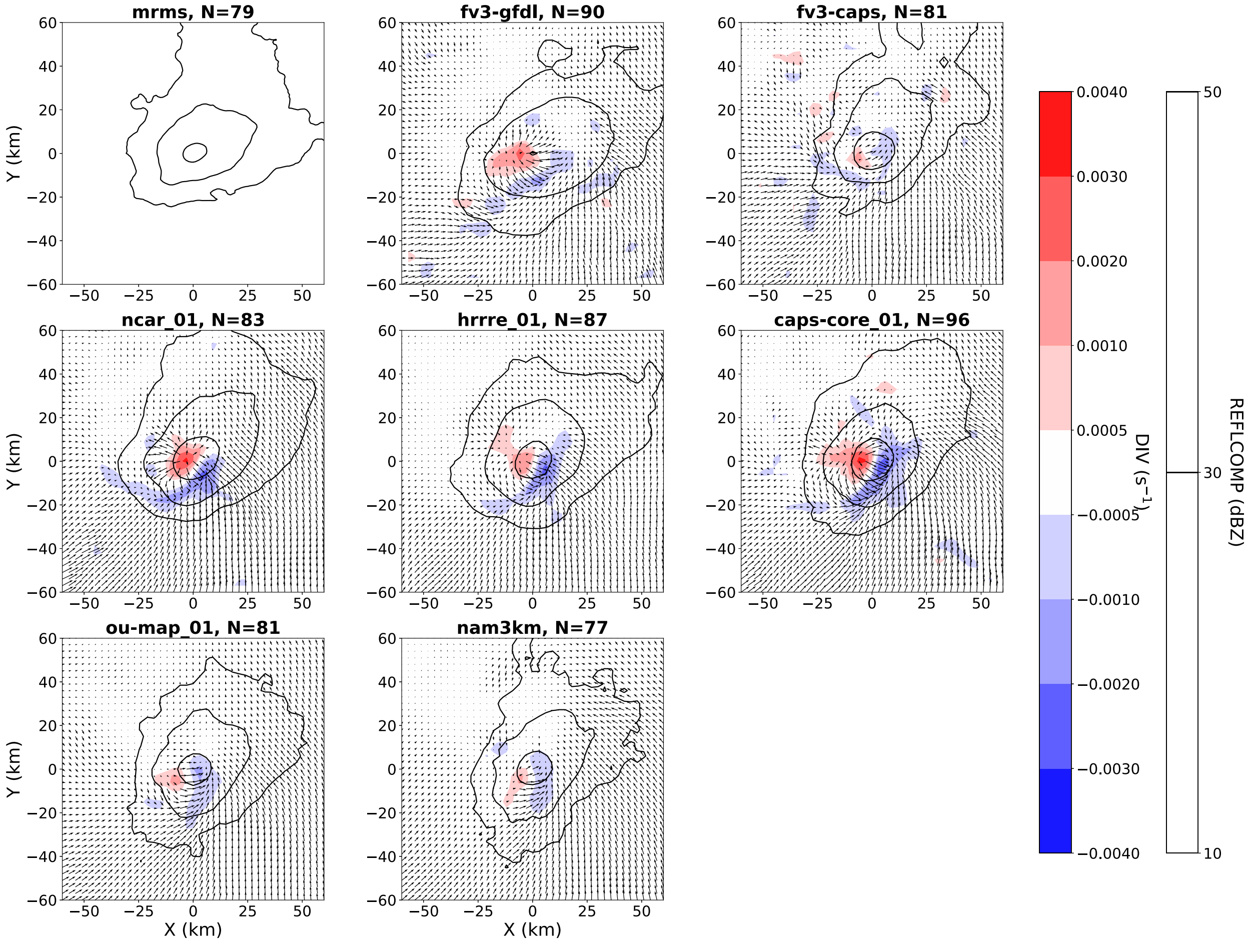

FIG. 9. PMMs of DIV_10M (s^{-1}; shading), REFLCOMP (contoured at 10, 30, 50 dB*Z*), and 10-m wind vectors (m s^{-1}; arrows).

the center of each grid box, both the Arakawa-C grid (ARW models) and Arakawa-B grid (NMMB models) require a four-point (horizontal) average for ζ and a two-point (vertical) average for w. The Arakawa-D grid used by the FV3 models requires no averaging in the calculation of UH. The ARW models, but not the NMMB and FV3 models, additionally apply a nine-point smoother before outputting the final UH field. It is therefore not surprising that PMM UH_1HMAX is so much lower for the ARW models. The fact that the ARW UH_1HMAX is so much lower than the NMMB UH_1HMAX despite both dynamical cores requiring similar degrees of w and ζ averaging likely owes to the nine-point smoother in the ARW models that damps UH maxima. The large intermodel differences in UH_1HMAX underscore the importance of accounting for different models' climatologies when comparing forecast output.

The ncar_01 produces storms with larger W_1HMAX than the other two ARW models (Fig. 6). This presumably is partly a consequence of the aggressive vertical velocity damping used in the HRRRE and CAPS core ensembles in 2017. In both ensembles, the maximum latent heating tendency limit was set to 0.07 K s^{-1} and Rayleigh damping was used to reduce the local w whenever the vertical Courant number exceeded 1.2. These features are included in the RAP, HRRR, and HRRRE (Benjamin et al. 2016) so that the forecasts are stable with relatively large time steps, aiding forecasts to finish on time within their computing allocations. Larger numerical diffusion also likely contributed to the weaker W_1HMAX in hrrre_01 and caps-core_01; the sixth-order numerical diffusion rate was set to 0.25 in those models and to 0.12 in ncar_01.

Storms in the fv3-gfdl were generally weaker than in the fv3-caps. This is evident in the PMMs of UH_1HMAX (Fig. 5), hourly maximum vertical velocity below 400 mb (W_1HMAX; Fig. 6), RAIN_1H (Fig. 7), and other variables not presented herein. It is difficult to judge which of the two FV3 models better represents storm intensity given the lack of verifying observations. It is noteworthy, however, that the fv3-caps

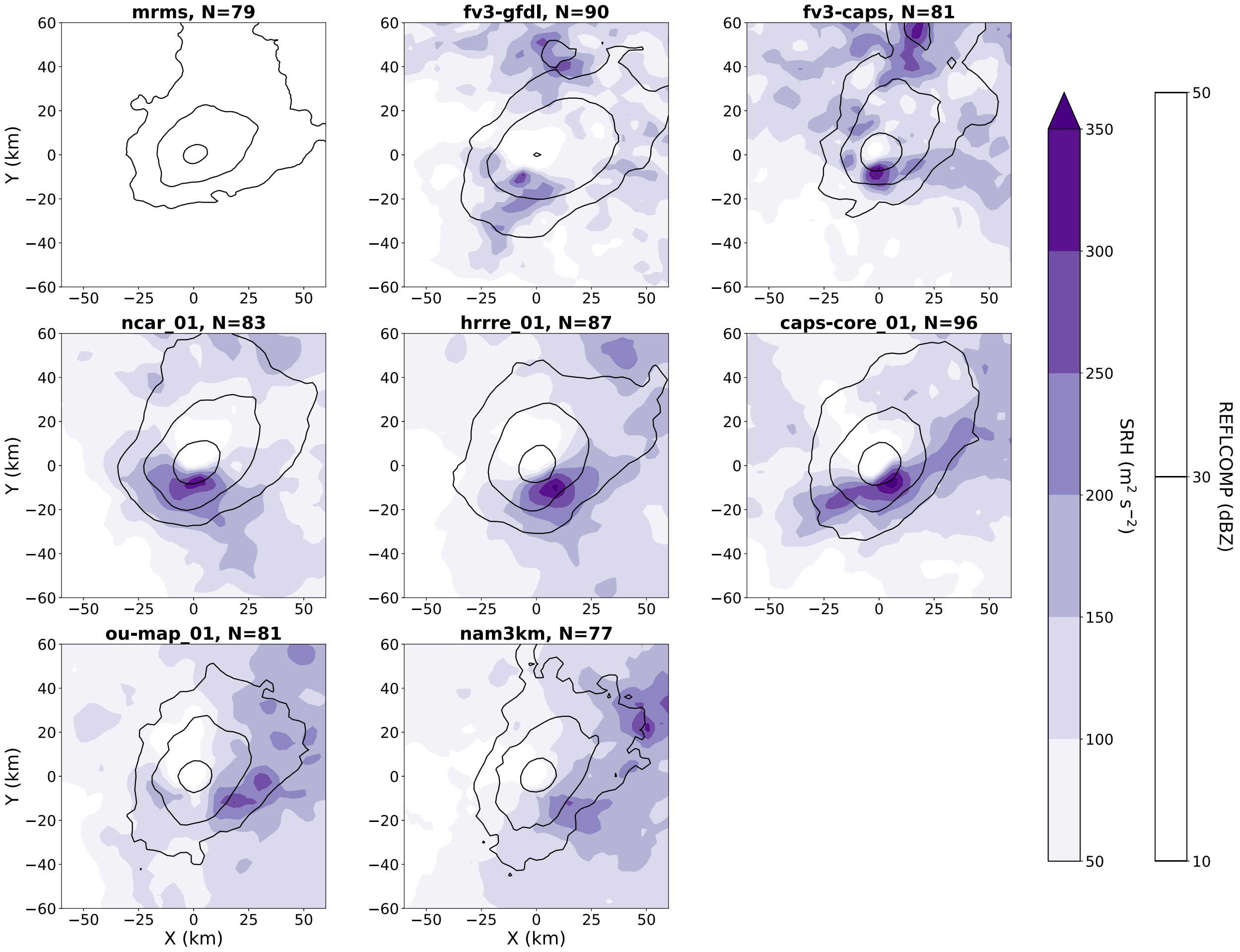

FIG. 10. PMMs of SRH (m^2 s^{-2}; shading) and REFLCOMP (contoured at 10, 30, 50 dBZ).

PMM RAIN_1H much better matches the mrms PMM RAIN_1H (Fig. 7).

Storm inflow generally had lower surface-based convective available potential energy (SBCAPE) in the FV3 models than in the other models (Fig. 5). This is due in part to lower storm-inflow 2-m dewpoints (TD_2M) in the FV3 models (Fig. 7). With regard to 2-m temperature (T_2M), all the models that use the Thompson microphysics, including the fv3-caps, produced warmer cold pools relative to NSE surface temperatures than the other models (Fig. 8).

Low-level convergence was substantially larger in the ARW models than in the other models (Fig. 9), despite the ARW models having the warmest cold pools (Fig. 8). The larger low-level convergence in the ARW models is associated with stronger flow perpendicular to the gust front, which is much more distinct than in the FV3 and NMMB models (Fig. 9). The PMM SRH in all of the models is locally maximized very near the mean location of the storm low-level updraft (Figs. 6 and 10). The models that use the Thompson microphysics exhibit the largest PMM SRH maxima.

d. Kernel density estimates

The univariate KDEs (Fig. 11) are sufficiently normal that intermodel displacements between the modes of KDEs correspond well with intermodel differences between peak (i.e., case-maximum; this modifier is implicit in the remainder of this section) PMM values. For example, the UH_1HMAX KDE modes are smallest for the ARW models and largest for the FV3 models, consistent with the UH_1HMAX PMMs (cf. Figs. 5 and 11a). Likewise, the larger W_1HMAX KDE mode in the ncar_01 than in the other two ARW models is consistent with the W_1HMAX PMMs (cf. Figs. 6 and 11b), and the smaller SBCAPE KDE modes in the FV3 models than in the ARW and NMMB models is consistent with the SBCAPE PMMs (cf. Figs. 5 and 11c). In the cases of more nonnormally distributed variables, however, the KDEs provide information beyond what

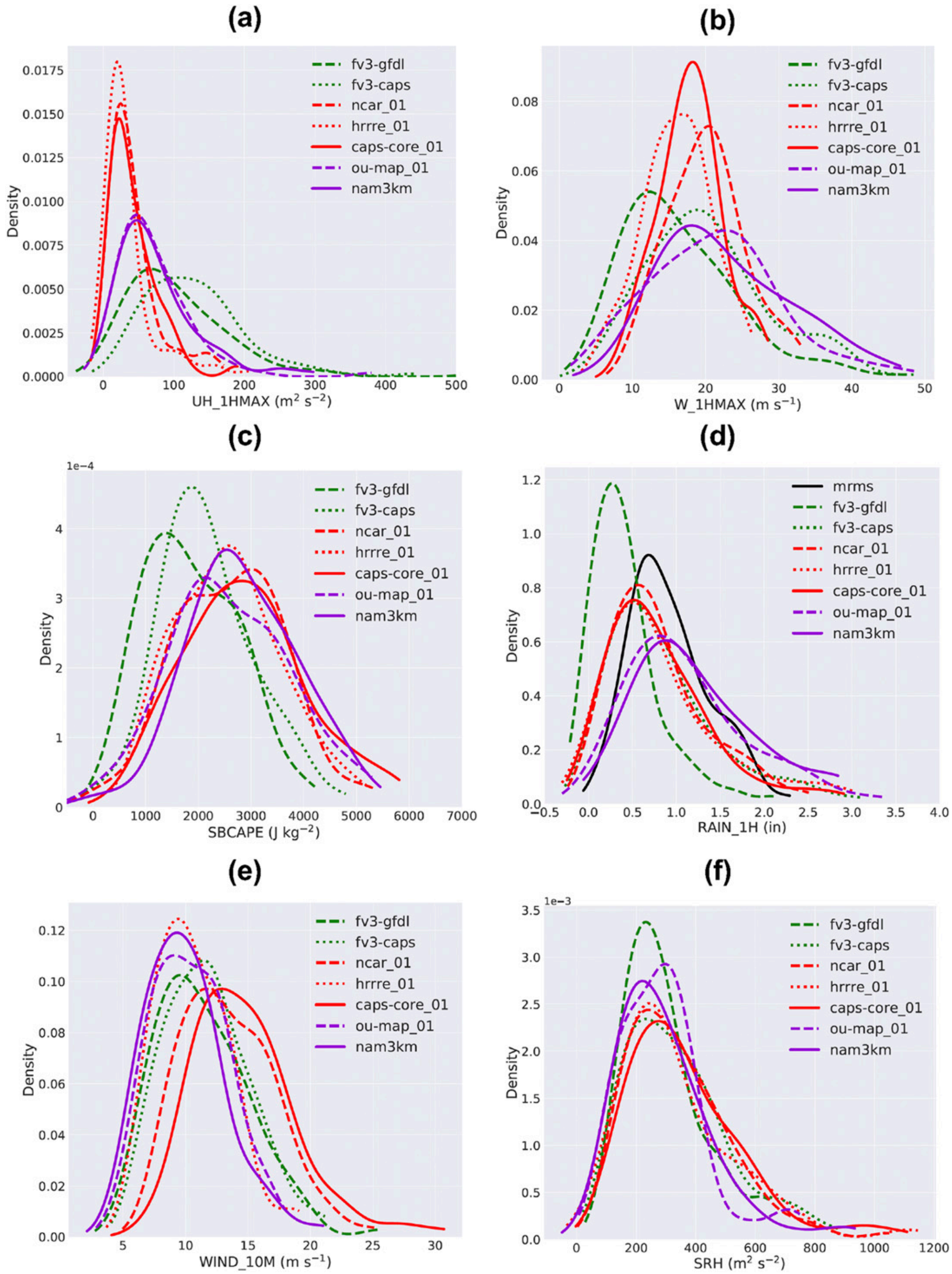

FIG. 11. KDEs of maximum (a) UH_1HMAX (m^2 s^{-2}), (b) W_1HMAX (m s^{-1}), (c) SBCAPE (J kg^{-1}), (d) RAIN_1H (in.), (e) WIND_10M (m s^{-1}), and (f) SRH (m^2 s^{-2}).

the PMMs can. For example, while the PMM RAIN_1H values for the fv3-caps are much larger than for the fv3-gfdl (Fig. 7), the modes of the RAIN_1H distributions for each model are very similar since the fv3-caps RAIN_1H PDF is much more positively skewed than the fv3-gfdl PDF (Fig. 11d). In other words, rather than the fv3-caps tending to produce storms with moderately more rainfall than the fv3-gfdl, the heavier-rainfall storms in the fv3-caps produced much more rainfall than the heavier-rainfall storms in the fv3-gfdl.

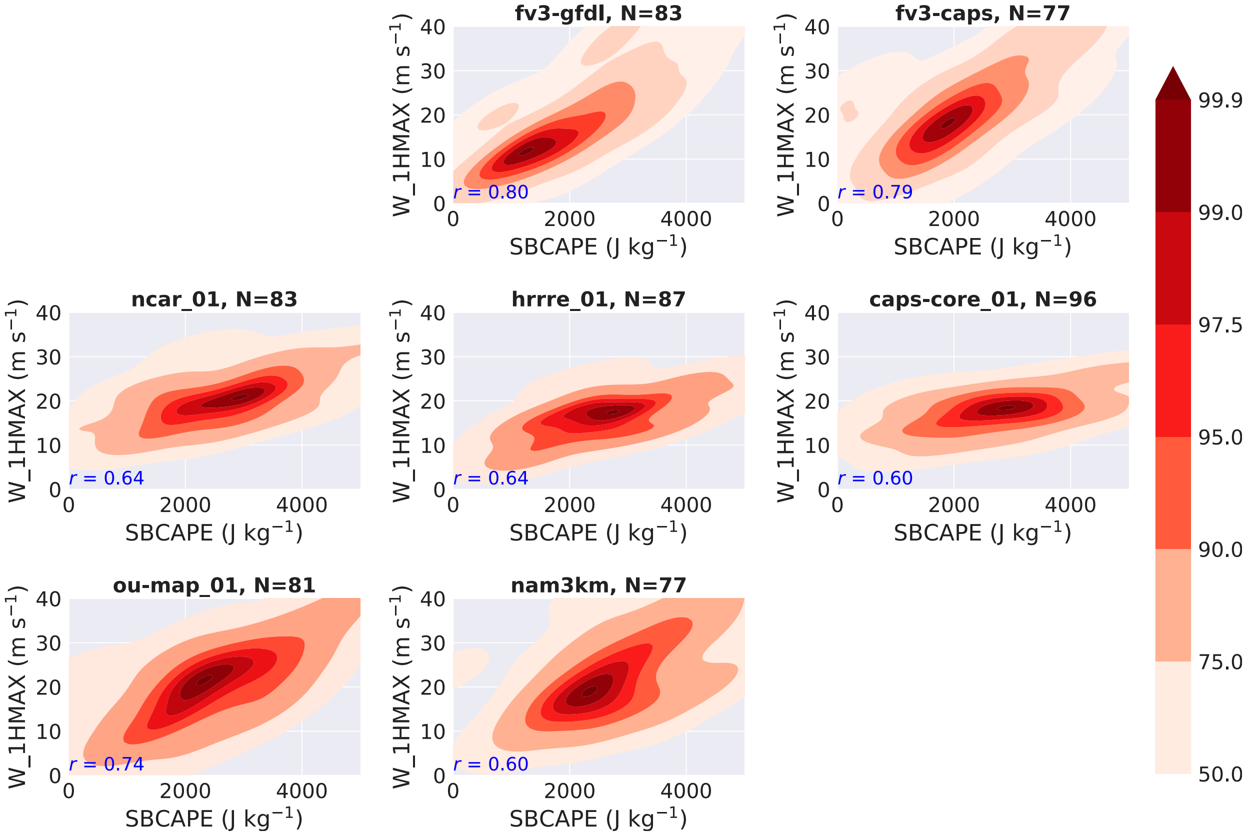

FIG. 12. KDEs of maximum W_1HMAX (m s^{-1}) and SBCAPE (J kg^{-1}). The Pearson correlation coefficient r for the two variables is shown on each panel.

The fv3-caps rainfall is very similar to that in the ARW models, which we attribute to their use of the same (Thompson) microphysics scheme. Wind speeds at 10 m AGL (WIND_10M) were substantially lower in the hrrre_01 than in the other ARW models (Fig. 11e). This could be due in part to the use of a different LSM and PBL scheme in the hrrre_01 than in the ncar_01 and caps-core_01 (Table 1).

It was observed in section 4c that, for most of the models, the PMM SRH maximum occurs within the storm inflow very near the updraft rather than in the far field (Fig. 10). Thus, the intermodel differences in the SRH KDEs (Fig. 11f) largely reflect differences in convective feedbacks to the environment and not necessarily differences in the storm environments themselves (though such differences could have led to the differences in convective feedbacks). This comparison of the SRH PMMs and KDEs illustrates the importance of spatial field analysis for informed interpretation of low-order model climatologies of storms.

The *bivariate* KDEs reveal intermodel differences in intervariable relationships that the PMMs and univariate KDEs cannot provide. For example, relative to the other models, the ARW models exhibit less sensitivity of W_1HMAX to SBCAPE (Fig. 12). This result might be expected for the hrrre_01 and caps-core_01 due to their use of more aggressive vertical velocity damping and numerical diffusion, but this would not explain the ncar_01's insensitivity of W_1HMAX to SBCAPE. Comparing the two FV3 models, W_1HMAX is much less sensitive to SBCAPE in the fv3-gfdl than in the fv3-caps at SBCAPE < 2000 J kg^{-1} (Fig. 12); similar behavior is seen in SBCAPE-UH_1HMAX KDEs (not shown). The greater W_1HMAX-SBCAPE sensitivity in the fv3-caps than in the fv3-gfdl is presumably due to the fv3-caps' use of Thompson microphysics rather than the GFDL 6-category microphysics, since this is the only difference between these two models. Given that the ARW models also used Thompson microphysics but exhibited much less W_1HMAX-SBCAPE sensitivity than the fv3-caps and the remaining models, some other factor appears to strongly contribute to the relative insensitivity of W_1HMAX to SBCAPE in the ARW models. While identifying this factor is outside the scope of this paper, the preceding analysis and discussion illustrates how bivariate KDEs

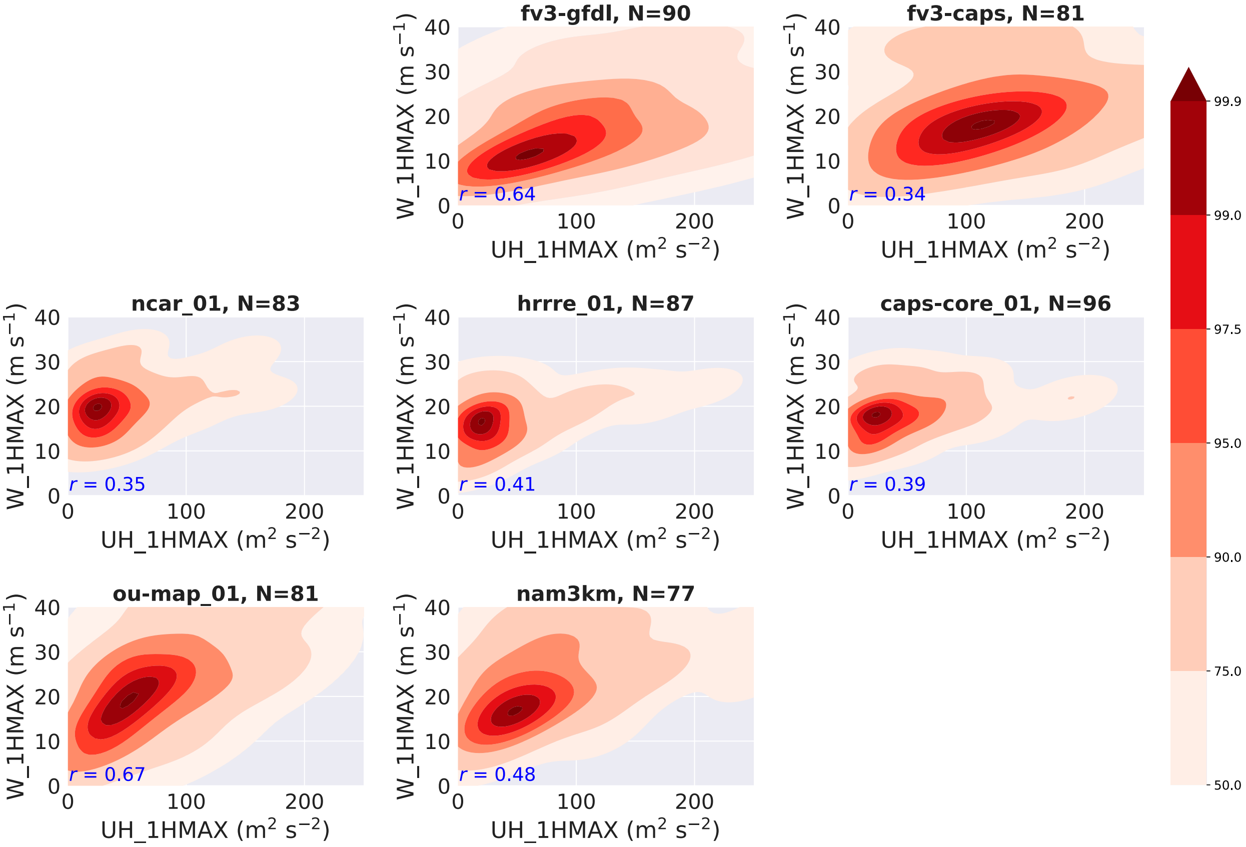

FIG. 13. KDEs of maximum W_1HMAX (m s^{-1}) and UH_1HMAX ($m^2 s^{-2}$). The Pearson correlation coefficient r for the two variables is shown on each panel.

can be used to identify possible systematic intermodel differences and to inspire and guide investigation into their sources.

The ARW model KDEs are noticeably compressed in the UH_1HMAX dimension (Figs. 13 and 14), but this is expected given the post hoc UH smoothing in those models (section 4c). Such differences in processing must be considered when attempting to attribute differences in bivariate relationships to modeled storm dynamics. The fv3-gfdl exhibits much less interdependence of UH_1MAX and SRH than does the fv3-caps (Fig. 14), which may be related to the generally weaker storms in the fv3-gfdl.

Additional analysis with different members of the NCAR, HRRR, and OU-MAP ensembles suggests that all of the intermodel differences highlighted in this section are qualitatively robust and not unduly contaminated by sampling error (see the appendix).

5. Conclusions

Optimization of future operational CAMs and CAM ensembles is a challenging and urgent endeavor that requires systematic comparisons of existing CAM systems. The Community Leveraged Unified Ensemble (CLUE) that has been run since 2016 as part of the annual HWT Spring Forecasting Experiment (SFE) provides an excellent opportunity to perform such intermodel comparisons and so illuminate the impacts of different dynamical cores, physics parameterization schemes, initialization procedures, and so forth on the realism of simulated storms and the accuracy of storm forecasts. The primary goals of this study are to leverage 2017 CLUE output to 1) develop and demonstrate new object-based techniques for verifying and characterizing CAM forecasts of storms and near-storm environments (NSEs), 2) use these new techniques along with existing verification methods to identify systematic forecast differences between several contemporary CAMs, and 3) infer impacts of model configuration choices on the representation and prediction of storms in CAMs.

We conclude the following about the novel methods developed herein:

1) Spatial probability-matched means (PMMs) of model storms and NSEs facilitate identification of systematic intermodel differences in storm structure, intensity,

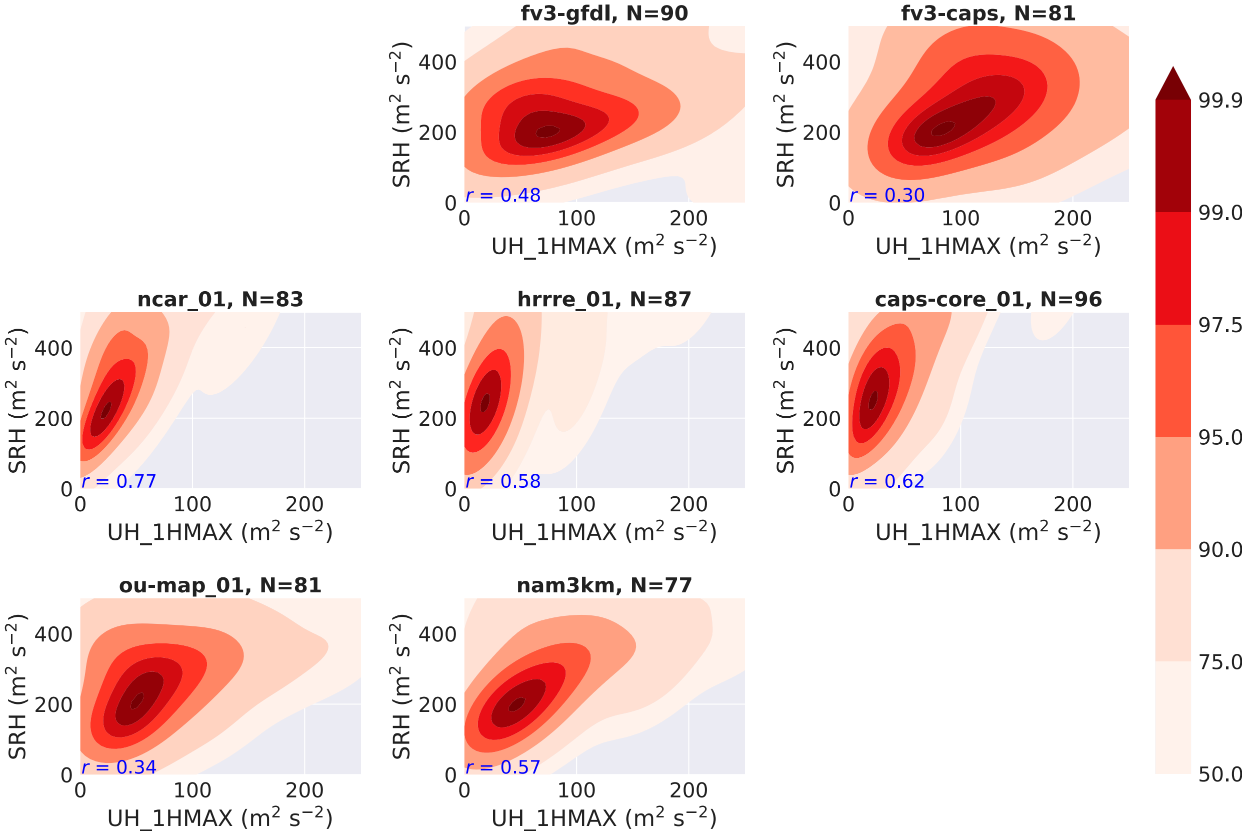

FIG. 14. KDEs of maximum SRH (m^2 s^{-2}) and UH_1HMAX (m^2 s^{-2}). The Pearson correlation coefficient *r* for the two variables is shown on each panel.

and environmental interactions that could be missed by manual inspection of individual cases.

2) Kernel density estimation (KDE) distills much of the important information in the PMMs, provides additional information in cases of nonnormally distributed variables, and when applied to bivariate distributions, illuminates intervariable relationships.

We conclude the following about the performance and behavior of the CAMs examined herein:

1) No single model or group of models sharing the same dynamical core (FV3, ARW, or NMMB) appeared to substantially outperform the others.
2) Both the GFDL and CAPS FV3 models slightly lagged the other models in surrogate severe forecast skill. This perhaps is not surprising given that the FV3 was still early in its development as a CAM.
3) All models were similarly skillful (in terms of object-based CSI) in forecasting storm locations. The ARW models, however, exhibited a positive frequency bias and therefore higher POD and FAR than the other models. This frequency bias was not exhibited by the CAPS FV3 model and therefore cannot be attributed solely to the Thompson microphysics. All the models, especially the ARW models, tended to produce too many storms in cases of storm clusters.

The remaining conclusions concern quasi-isolated, discrete storms only, which were the subject of the novel, object-based methods demonstrated in this study.

4) The updraft helicity fields output by current ARW models are strongly damped by the nine-point spatial smoother that is applied. This factor alone may account for much of the difference in updraft helicity climatology between the ARW models and the NMMB and FV3 models.
5) In terms of composite reflectivity structure, the GFDL FV3 model best matched the observed (MRMS-derived) storm orientation, while the remaining models produced storms that were too meridionally tilted.
6) The use of Thompson microphysics in place of the GFDL 6-category scheme in the CAPS FV3 model produced a variety of important changes in storm and NSE characteristics, including stronger updrafts, heavier rainfall (in much better agreement with observations), and a meridional bias in storm orientation.

7) Low-level convergence associated with storm updrafts was substantially stronger in the ARW models than in the NMMB and FV3 models.
8) The NCAR ensemble member produced the strongest storms of the ARW models. The temperature-tendency limiter in the HRRR ensemble member and the CAPS core member presumably weakened the storms somewhat in these ARW models.
9) Storm cold pools were warmer in models using the Thompson microphysics.

This study and the novel analysis techniques developed herein could be extended in many ways. Repeating this investigation for consecutive SFEs would facilitate tracking of CAM system performance over time and evaluation of the impacts of model configuration changes. Strategic changes in CLUE membership will occasionally enable particularly timely intermodel comparisons; for example, the 2018 CLUE included 14 FV3 models to help accelerate the development of the FV3 as a CAM dynamical core. Fully leveraging these opportunities to inform the design of future operational CAM systems requires sophisticated analysis techniques like those presented here. These techniques could also be applied to multiphysics CAM ensembles to identify systematic impacts of different parameterization schemes, or to any CAM or CAM ensemble to illuminate systematic differences in storms, NSEs, and storm–environment interactions within different mesoscale regimes (e.g., low CAPE, high shear versus high CAPE, high shear) or between different storm modes (e.g., tornadic versus nontornadic). Finally, applying these analyses to longer datasets would enable more restrictive criteria for mitigating uneven spatiotemporal sampling, thereby further ensuring intermodel differences primarily reflect differences in model behavior.

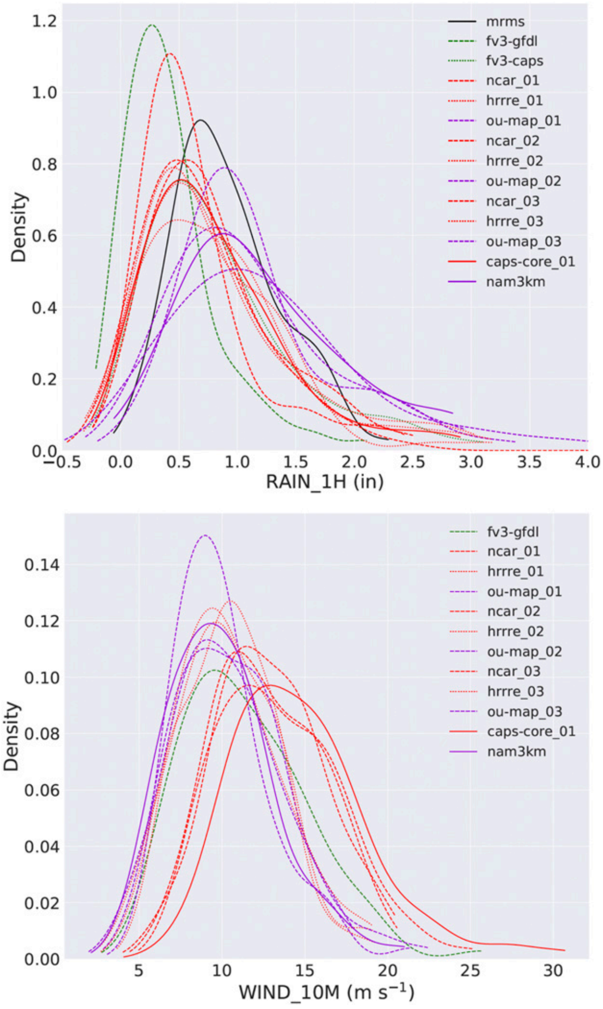

FIG. A1. KDEs of maximum (top) RAIN_1H and (bottom) WIND_10M with members 2 and 3 of the NCAR, HRRR, and OU-MAP ensembles included.

Acknowledgments. This work was prepared by the authors with funding provided by National Science Foundation Grant AGS-1062932 and NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce, and NOAA Awards NA15OAR4590193, NA17OAR4590114, and NA16OAR4590236. The CAPS forecasts were produced under the support of NOAA Grants NA16NWS4680002 and NA17OAR4590186, using NSF Extreme Science and Engineering Discovery Environment (XSEDE) supercomputing resources at the University of Texas Advanced Computing Center (TACC), and those of the University of Oklahoma Supercomputer for Education and Research (OSCER). OU MAP forecasts were produced using NSF XSEDE supercomputers. We thank the three formal reviewers whose comments substantially improved the manuscript. We are also grateful to Brett Roberts for informally reviewing an early version of the manuscript, and to Curtis Alexander, Cameron Homeyer, Derek Stratman, and Ted Mansell for valuable discussions. Valuable local computing assistance was provided by Gerry Creager, Karen Cooper, Jesse Butler, and Jeff Horn. All analyses and visualizations were produced using the freely provided Anaconda Python distribution.

APPENDIX

Sampling Error Impacts

The small samples used in this study call into question the generality of the results. To address this concern, we repeated all of our analyses with either ncar_01, hrrre_01, or ou-map_01 swapped with member 2 or member 3 of their respective ensemble (a total of 3 × 2 = 6 new analyses). Figure A1 presents univariate KDEs for the two variables in Fig. 11 that varied most with the choice of ensemble member. Major intermodel differences in all of the original KDEs are qualitatively preserved in these new analyses, including differences between groups of models with the same dynamic core. The intermodel differences in the bivariate KDEs are similarly insensitive to the choice of ensemble member (not shown). Consistent with the small changes in the univariate KDE modes (Fig. A1), the maximum value within each PMM varies little with ensemble member, though smaller-scale details of the spatial patterns are more sensitive (not shown). As with the KDEs, all of the intermodel PMM differences highlighted in section 4 are retained in the new PMMs. The results of these sensitivity tests (along with the application of criteria 4 and 5 in generating the ISOLATED_STORMS database; section 3c) strongly imply that the intermodel forecast differences noted throughout this study are not merely artifacts of sampling errors.

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.

Anderson, J. L., T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn, and A. Avellano, 2009: The Data Assimilation Research Testbed: A community data assimilation facility. *Bull. Amer. Meteor. Soc.*, **90**, 1283–1296, https://doi.org/10.1175/2009BAMS2618.1.

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.

Brooks, H., C. A. Doswell III, and M. P. Kay, 2003: Climatological estimates of local daily tornado probability for the United States. *Wea. Forecasting*, **18**, 626–640, https://doi.org/10.1175/1520-0434(2003)018<0626:CEOLDT>2.0.CO;2.

Carlin, J. T., J. Gao, J. C. Snyder, and A. V. Ryzhkov, 2017: Assimilation of Z_{DR} columns for improving the spinup and forecast of convective storms in storm-scale models: Proof-of-concept experiments. *Mon. Wea. Rev.*, **145**, 5033–5057, https://doi.org/10.1175/MWR-D-17-0103.1.

Chen, J.-H., and S.-J. Lin, 2013: Seasonal predictions of tropical cyclones using a 25-km-resolution general circulation model. *J. Climate*, **26**, 380–398, https://doi.org/10.1175/JCLI-D-12-00061.1.

Clark, A. J., 2017: Generation of ensemble mean precipitation forecasts from convection-allowing ensembles. *Wea. Forecasting*, **32**, 1569–1583, https://doi.org/10.1175/WAF-D-16-0199.1.

——, and Coauthors, 2012: An overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74, https://doi.org/10.1175/BAMS-D-11-00040.1.

——, R. G. Bullock, T. L. Jensen, M. Xue, and F. Kong, 2014: Application of object-based time-domain diagnostics for tracking precipitation systems in convection-allowing models. *Wea. Forecasting*, **29**, 517–542, https://doi.org/10.1175/WAF-D-13-00098.1.

——, 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.

Davis, C. A., B. G. Brown, and R. G. Bullock, 2006a: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772–1784, https://doi.org/10.1175/MWR3145.1.

——, ——, and ——, 2006b: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Mon. Wea. Rev.*, **134**, 1785–1795, https://doi.org/10.1175/MWR3146.1.

Dowell, D., and Coauthors, 2016: Development of a High-Resolution Rapid Refresh Ensemble (HRRRE) for severe weather forecasting. *28th Conf. on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., 8B.2, https://ams.confex.com/ams/28SLS/webprogram/Paper301555.html.

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.

Evensen, G., 2003: The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dyn.*, **53**, 343–367, https://doi.org/10.1007/s10236-003-0036-9.

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.

Gustafsson, N., and Coauthors, 2018: Survey of data assimilation methods for convective-scale numerical weather prediction at operational centres. *Quart. J. Roy. Meteor. Soc.*, **144**, 1218–1256, https://doi.org/10.1002/qj.3179.

Harris, L. M., and S.-J. Lin, 2013: A two-way nested global-regional dynamical core on the cubed sphere grid. *Mon. Wea. Rev.*, **141**, 283–306, https://doi.org/10.1175/MWR-D-11-00201.1.

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., 2002: Nonsingular implementation of the Mellor–Yamada level 2.5 scheme in the NCEP Meso model. NCEP Office Note 437, 61 pp., http://www.emc.ncep.noaa.gov/officenotes/newernotes/on437.pdf.

——, and R. Gall, 2012: Scientific documentation of the NCEP Nonhydrostatic Multiscale Model on the B Grid (NMMB). Part 1: Dynamics. NCAR Tech. Note NCAR/TN-489+STR, 75 pp., https://doi.org/10.5065/D6WH2MZX.

Johnson, A., X. Wang, 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.

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.

——, S. R. Dembek, S. J. Weiss, J. L. Case, J. J. Levit, and R. A. Sobash, 2010: Extracting unique information from high-resolution forecast models: Monitoring selected fields and phenomena every time step. *Wea. Forecasting*, **25**, 1536–1542, https://doi.org/10.1175/2010WAF2222430.1.

Mahalik, M. C., B. R. Smith, K. L. Elmore, D. M. Kingfield, K. L. Ortega, and T. M. Smith, 2019: Estimates of gradients in radar moments using a linear least squares derivative technique. *Wea. Forecasting*, **34**, 415–434, https://doi.org/10.1175/WAF-D-18-0095.1.

Miller, M. L., V. Lakshmanan, and T. Smith, 2013: An automated method for depicting mesocyclone paths and intensities. *Wea. Forecasting*, **28**, 570–585, https://doi.org/10.1175/WAF-D-12-00065.1.

Mitchell, K. E., and Coauthors, 2005: The Community Noah Land Surface Model (LSM)—User's guide (v2.7.1). NOAA/NCEP, 26 pp., ftp://ftp.emc.ncep.noaa.gov/mmb/gcp/ldas/noahlsm/ver_2.7.1.

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.

——, and ——, 2006: An improved Mellor–Yamada level-3 model: Its numerical stability and application to a regional prediction of advection fog. *Bound.-Layer Meteor.*, **119**, 397–407, https://doi.org/10.1007/s10546-005-9030-8.

Powers, J., and Coauthors, 2017: The Weather Research and Forecasting (WRF) Model: Overview, system efforts, and future directions. *Bull. Amer. Meteor. Soc.*, **98**, 1717–1737, https://doi.org/10.1175/BAMS-D-15-00308.1.

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.

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.

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

Schwartz, C. S., 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.

——, ——, 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.

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://doi.org/10.5065/D68S4MVH.

Skinner, P. S., and Coauthors, 2018: Object-based verification of a prototype Warn-on-Forecast System. *Wea. Forecasting*, **33**, 1225–1250, https://doi.org/10.1175/WAF-D-18-0020.1.

Smirnova, T. G., J. M. Brown, S. G. Benjamin, and J. S. Kenyon, 2016: Modifications to the rapid update cycle land surface model (RUC LSM) available in the Weather Research and Forecasting (WRF) Model. *Mon. Wea. Rev.*, **144**, 1851–1865, https://doi.org/10.1175/MWR-D-15-0198.1.

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., F. Kong, K. Brewster, M. Xue, K. W. Thomas, T. A. Supinie, S. Perfater, and B. Albright, 2019: Evaluation of convection-permitting precipitation forecast products using WRF, NMMB, and FV3 for the 2016–17 NOAA Hydrometeorology Testbed Flash Flood and Intense Rainfall Experiments. *Wea. Forecasting*, **34**, 781–804, https://doi.org/10.1175/WAF-D-18-0155.1.

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.

——, C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Wea. Forecasting*, **31**, 255–271, https://doi.org/10.1175/WAF-D-15-0138.1.

Surcel, M., I. Zawadzki, and M. K. Yau, 2014: On the filtering properties of ensemble averaging for storm-scale precipitation forecasts. *Mon. Wea. Rev.*, **142**, 1093–1105, https://doi.org/10.1175/MWR-D-13-00134.1.

Theis, S. E., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: A pragmatic approach. *Meteor. Appl.*, **12**, 257–268, https://doi.org/10.1017/S1350482705001763.

Thompson, G., R. M. Rasmussen, and K. Manning, 2004: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part I: Description and sensitivity analysis. *Mon. Wea. Rev.*, **132**, 519–542, https://doi.org/10.1175/1520-0493(2004)132<0519:EFOWPU>2.0.CO;2.

——, 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.

Van der Walt, S., J. L. Schonberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, 2014: Scikit-image: Image processing in Python. *PeerJ*, **2**, e453, https://doi.org/10.7717/peerj.453.

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.

Wu, W., D. F. Parrish, E. Rogers, and Y. Lin, 2017: Regional ensemble–variational data assimilation using global ensemble forecasts. *Wea. Forecasting*, **32**, 83–96, https://doi.org/10.1175/WAF-D-16-0045.1.

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.