What Does a Convection-Allowing Ensemble of Opportunity Buy Us in Forecasting Thunderstorms?

BRETT ROBERTS,[a,b,c] BURKELY T. GALLO,[a,c] ISRAEL L. JIRAK,[c] ADAM J. CLARK,[b,d] DAVID C. DOWELL,[e] XUGUANG WANG,[d] AND YONGMING WANG[d]

[a] *Cooperative Institute for Mesoscale Meteorological Studies, 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] *NOAA/OAR/Earth System Research Laboratory, Boulder, Colorado*

(Manuscript received 6 May 2020, in final form 4 September 2020)

ABSTRACT: The High Resolution Ensemble Forecast v2.1 (HREFv2.1), an operational convection-allowing model (CAM) ensemble, is an ''ensemble of opportunity'' wherein forecasts from several independently designed deterministic CAMs are aggregated and postprocessed together. Multiple dimensions of diversity in the HREFv2.1 ensemble membership contribute to ensemble spread, including model core, physics parameterization schemes, initial conditions (ICs), and time lagging. In this study, HREFv2.1 forecasts are compared against the High Resolution Rapid Refresh Ensemble (HRRRE) and the Multiscale data Assimilation and Predictability (MAP) ensemble, two experimental CAM ensembles that ran during the 5-week Spring Forecasting Experiment (SFE) in spring 2018. The HRRRE and MAP are formally designed ensembles with spread achieved primarily through perturbed ICs. Verification in this study focuses on composite radar reflectivity and updraft helicity to assess ensemble performance in forecasting convective storms. The HREFv2.1 shows the highest overall skill for these forecasts, matching subjective real-time impressions from SFE participants. Analysis of the skill and variance of ensemble member forecasts suggests that the HREFv2.1 exhibits greater spread and more effectively samples model uncertainty than the HRRRE or MAP. These results imply that to optimize skill in forecasting convective storms at 1–2-day lead times, future CAM ensembles should employ either diverse membership designs or sophisticated perturbation schemes capable of representing model uncertainty with comparable efficacy.

KEYWORDS: Convection; Ensembles; Forecast verification/skill; Numerical weather prediction/forecasting; Short-range prediction

1. Introduction

During the most recent decade, convection-allowing models (CAMs) have become a staple in the operational forecasting toolbox, particularly for applications that benefit most from their combination of high spatial resolution and explicit convective structures. While deterministic CAMs have been running operationally since around 2010, the implementation of their ensemble prediction system (EPS) counterparts has lagged behind owing to a higher computational cost. Nonetheless, experimental CAM EPSs were produced by the University of Oklahoma (OU) Center for Analysis and Prediction of Storms (CAPS) in real time as early as the mid-2000s (e.g., Xue et al. 2007; Kong et al. 2007; Levit et al. 2008) for evaluation in the NOAA Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (SFE; Kain et al. 2003; Clark et al. 2012; Gallo et al. 2017). Through subsequent years, additional CAM EPSs have been run on an experimental basis during the SFE by the National Center for Atmospheric Research (NCAR; Schwartz et al. 2015), NOAA's Global Systems Laboratory (GSL) (Dowell et al. 2016, hereafter D16), the OU Multiscale data Assimilation and Predictability (MAP) group (Johnson et al. 2015; Wang and Wang 2017; Wang et al. 2018; Johnson et al. 2020), and others.

By 2016, the number of EPSs contributed to the annual SFE had grown large enough to justify instantiating the Community Leveraged Unified Ensemble (CLUE; Clark et al. 2018), a framework for scientific collaborators to set common standards for key aspects of their systems such as the model grid and data output format. The CLUE, in turn, has paved the way for more systematic, controlled comparisons of its various subsets (e.g., Potvin et al. 2019). The different CLUE subsets are distinguished from one another not only by their basic model configuration (e.g., dynamical core), but also by their membership design approach. Most of the subsets (e.g., the NCAR, MAP, and GSL systems) use a single, unified[1] model configuration with ensemble spread achieved through initial condition (IC) and lateral boundary condition (LBC) perturbations. However, some subsets (e.g., the CAPS core ensemble; Clark et al. 2018) use multiphysics configurations wherein the microphysics, planetary boundary layer (PBL), and/or land surface model (LSM) parameterization schemes also differ. Separate from the CLUE, an especially diverse class of CAM ensembles,

[1] In this paper, we use ''unified'' to describe an ensemble whose members all share the same dynamical core and model configuration; specified differences between members are limited to applied perturbations (e.g., to the initial and lateral boundary conditions of model state variables, or to variables internal to physics parameterization schemes).

Corresponding author: Brett Roberts, brett.roberts@noaa.gov

DOI: 10.1175/WAF-D-20-0069.1

"ensembles of opportunity" (EOs), have also been evaluated in SFEs since 2011. The term EO herein is the same as the "poor man's ensemble" (e.g., Ebert 2001; Arribas et al. 2005; Casanova and Ahrens 2009), which can be characterized as several independently designed, deterministic numerical weather prediction (NWP) models combined and postprocessed as an ensemble. Such ad hoc systems typically violate the ideal EPS property of equally likely member solutions (Leith 1974; Ziehmann 2000) and require extra postprocessing techniques for maximum utility (e.g., interpolation to a common grid, separate bias correction of each member, etc.). In the context of CAMs, this ensemble design strategy originated from the Storm Prediction Center (SPC) with the Storm-Scale Ensemble of Opportunity (SSEO; Jirak et al. 2012). The SSEO comprised seven deterministic CAMs from several modeling centers that used substantially different model configurations, and of which only some were even operationally supported. The SSEO was evaluated yearly in the SFE until it was supplanted by the High Resolution Ensemble Forecast system, version 2 (HREFv2; Roberts et al. 2019, hereafter R19), NOAA's first operational CAM ensemble, in late 2017. HREFv2 may be regarded as the formalization of SSEO's design philosophy, as its eight members' configurations closely mimic several of SSEO's members; albeit with finer, more uniform horizontal grid spacing of $O(3)$ km, and a closely synchronized run schedule that is operationally supported. A commonality of SSEO and HREFv2 is the inclusion of time-lagged members, which adds yet another element of membership diversity alongside member-to-member differences in dynamical core, physics, and ICs/LBCs.

Although the relatively high spatial resolution of CAM systems benefits NWP skill for a wide array of atmospheric phenomena, CAM development has been motivated in particular by the opportunity to predict convective storms with realistic structures and impacts on their surrounding environment. For example, CAM forecasts of convective mode, coverage, and severe weather hazards have been the primary focus of subjective and objective evaluations conducted in the HWT SFE, which in turn have guided CAM development pathways (e.g., D16; R19; Gallo et al. 2019). In the course of these evaluations, a consistent theme has emerged of EOs receiving among the highest scores in subjective participant ratings and limited objective verification metrics, when compared against other CAM EPSs lacking such configuration diversity (e.g., Jirak et al. 2015, 2016). Although the SSEO was largely born of necessity at a time when no formally designed CAM EPSs were available operationally, its consistent success in SFE evaluations motivated the decision to use a similar membership design for the operational HREFv2.

While SFE evaluations have highlighted the relative success of convective storm forecasts from CAM EOs, the reasons for this success have not yet been investigated rigorously. For coarser-grid EPSs, it has been demonstrated that accounting for *model uncertainty* separately from *IC uncertainty* can generate larger spread and superior forecasts for traditional synoptic fields (Du et al. 1997; Stensrud et al. 2000). Other studies have suggested that employing multiple models (Ziehmann 2000; Eckel and Mass 2005; Johnson and Swinbank 2009) or physics schemes (Jankov et al. 2005; Hacker et al. 2011) may be particularly effective ways of capturing this uncertainty. More recently, the value of representing model uncertainty within a CAM ensemble using variations in model core (Clark et al. 2008; Johnson et al. 2011; Clark 2019), PBL parameterizations (Schwartz et al. 2010; Johnson et al. 2011; Loken et al. 2019), microphysics parameterizations (Clark et al. 2008; Schwartz et al. 2010; Duda et al. 2014; Loken et al. 2019), LSM parameterizations (Duda et al. 2017), and time-lagging (Mittermaier 2007) has been shown in the context of forecasting synoptic and precipitation fields to improve ensemble skill, typically through increasing spread in underdispersive systems. Gasperoni et al. (2020) conducted experiments to compare different methods of sampling model uncertainty in the context of mutliscale ICs generated by the Gridpoint Staistical Interpolation (GSI; Wu et al. 2002; Shao et al. 2016) ensemble–variational (EnVar) system (e.g., Wang 2010; Wang et al. 2013; Wang and Wang 2017). They found that both their multimodel and multiphysics configurations were superior to their single-model single-physics configuration. It was also found that a multimodel design tends to perform best at early lead times, whereas multiphysics with stochastic physics tends to be best for later lead times.

With these findings considered, there is some a priori reason to suspect that CAM EOs—which at least attempt to represent uncertainty across several of these relevant dimensions simultaneously—may demonstrate better spread characteristics and more meaningful probability density functions (PDFs) than their fully unified EPS counterparts, which generally only represent IC/LBC uncertainty. Although this is an area where EOs offer a clear benefit over unified EPSs, the latter have their own advantages: they enable simpler, more efficient technical implementations and modularity, which in turn fosters collaborative development across the NWP community. Furthermore, a unified EPS with IC perturbations prescribed around one analysis typically only requires a single data assimilation (DA) system, yielding another important efficiency advantage with respect to intellectual investment, computational resources, and maintenance overhead. The desire to fuse the benefits of both ensemble types has spurred recent attempts to develop stochastically perturbed parameterizations (e.g., Jankov et al. 2019; Hirt et al. 2019; Wastl et al. 2019), which could obviate the need for diverse configuration choices within an EPS's membership to sample physics uncertainty. In the present study, we aim to quantify differences between EOs (represented using the HREF) and unified, formally designed CAM EPSs (represented using two CLUE subsets) with respect to their skill and spread in forecasts of convective storms.

The paper is organized as follows. Section 2 describes datasets and analysis methods. Sections 3 and 4 present analyses of composite reflectivity and surrogate severe forecasts, respectively. Section 5 summarizes our findings, draws relevant conclusions, and offers directions for future research on related topics.

2. Methodology

a. Datasets

1) NWP FORECAST DATASETS

In this study, we verify and compare CAM ensemble forecasts from the 2018 HWT Spring Forecasting Experiment (hereafter SFE2018), which ran weekdays from 30 April to 1 June. All CAM ensemble forecasts examined herein were initialized at 0000 UTC and forecast lead times of 12–36 h are verified. Owing to occasional missing NWP and/or observational data, our final verification dataset covers 21 of the 24 days SFE2018 operated: 30 April, 1–4 May, 7–8 May, 10–11 May, 14–18 May, 21 May, 24–25 May, 29–31 May, and 1 June.

The HREFv2 is an eight-member multimodel, multiphysics, multi-IC CAM EO with time lagging. Several independently developed deterministic CAMs compose the HREFv2 membership. Herein, we verify the HREFv2.1, an HREF variant with two additional members (for 10 members total): the HRRR and HRRR −6 h. The HREFv2.1 has been processed in real-time at the NOAA Storm Prediction Center since April 2019, and showed modestly improved skill over the HREFv2 in forecasting convective storms from subjective and limited objective verification during SFE2018 (Gallo et al. 2018). Further membership configuration details are given in Table 1, while a diagram of the time-lagging approach is displayed in Fig. 1. The HREFv2.1 is available as a 10-member ensemble out to a lead time of 30 h, and as a 9-member ensemble out to 36 h (with the HRRR −6 h member dropping out after 30 h). Because the native model grids differ between some members, all data are interpolated to a common 3-km grid using a nearest-neighbor approach before ensemble postprocessing. It is important to note that the HREFv2.1 verified herein uses a different membership[2] than the HREFv2 produced at NCEP and distributed via public channels, so our results will not necessarily apply to that configuration in a strict sense.

Two[3] CLUE subsets are compared to the HREFv2.1: the High Resolution Rapid Refresh Ensemble (HRRRE; D16), and the MAP Ensemble (Johnson et al. 2015; Wang and Wang 2017; Wang et al. 2018; Johnson et al. 2020). Both the HRRRE and MAP are unified, formally designed CAM EPSs. Furthermore, as both systems were designed within the CLUE framework for SFE2018, they largely shared model configuration details: the Advanced Research version of the Weather Research and Forecasting (WRF) Model (WRF-ARW; Skamarock et al. 2008) dynamical core was employed at 3-km horizontal grid spacing using the Mellor–Yamada–Nakanishi–Niino (MYNN; Nakanishi and Niino 2004) PBL and Thompson aerosol-aware (Thompson and Eidhammer 2014) microphysics parameterizations. The HRRRE comprised a 36-member ensemble DA system that was initialized at 0300 UTC daily from a Global Forecast System (GFS) background with GDAS perturbations; these members were then cycled hourly via an ensemble Kalman filter (EnKF) using conventional and radar data, until at 0000 UTC nine of the members launched 36-h forecasts. The preceding 1800 UTC GFS provided mean LBCs upon which random perturbations were added to ensemble members. Also, in the HRRRE, perturbations were introduced to the soil moisture field during the first minute of each day's DA cycling (0300–0301 UTC), representing effectively another type of IC perturbation. The MAP used a 41-member DA ensemble that ran for 6 h (1800–0000 UTC daily) prior to the initialization of its 10 forecast members at 0000 UTC. Its DA was an EnKF–3DEnVar hybrid system based on GSI that assimilated both conventional (hourly from 1800 to 0000 UTC) and radar (every 20 min from 2300 to 0000 UTC) observations. Different from HRRRE, the ensemble LBCs for MAP during DA and ensemble forecasts were provided by members of NCEP's Global Ensemble Forecast System (GEFS) and Short Range Ensemble Forecast (SREF). In the MAP forecast system, unlike the HRRRE, there is a control member (designated MAP 01 hereafter) taking its ICs from an EnVar analysis, whereas the other nine members contain specified IC perturbations from cycled and recentered GSI EnKF; a consequence is that smaller forecast error may be expected from the control member when aggregated over many cases (Johnson et al. 2020). During DA cycling, MAP perturbations were recentered around the control member prior to each cycle, while recentering was not performed during the HRRRE's DA. More complete details of the HRRRE are available in D16, and of the MAP in Johnson et al. (2015), Wang and Wang (2017), Wang et al. (2018), and Johnson et al. (2020). In summary, the HRRRE and MAP have nearly identical model configurations, but their respective approaches to DA and IC/LBC perturbation strategies differ significantly; the HRRRE also includes perturbations to soil moisture ICs, while MAP does not.

In this study, we verify two forecast fields: instantaneous composite radar reflectivity (CREF) and hourly maximum 2–5 km above ground level updraft helicity (UH; Kain et al. 2008). UH is used to construct surrogate severe probabilistic forecasts (Sobash et al. 2011, 2016b), which are smoothed neighborhood maximum ensemble probability (NMEP; Schwartz and Sobash 2017) fields based on UH exceedence thresholds [more information is given in section 2b(1)]. For CREF, similar NMEPs are calculated to assess convective coverage, timing, and location. Thus, the bias-corrected CREF field is thresholded at 40 dBZ, above which values are typically associated with deep moist convection. Verification of UH supplements this by focusing more narrowly on *rotating* convective updrafts, which are responsible for a disproportionate share of severe weather hazards (e.g., Duda and Gallus 2010).

[2] Specific differences between SPC's HREFv2.1 and NCEP's HREFv2 are as follows: 1) HREFv2.1 adds two new HRRR members, increasing the member count from 8 to 10; 2) HREFv2.1 uses a 12-h time lag NAM Nest member, where HREFv2 uses a 6-h time lag member; and 3) HREFv2 officially assigns decreased weight to lagged members (and increased weights to nonlagged members) in computing ensemble mean fields. See Table 1 for additional information.

[3] Additional CLUE subsets, including the aforementioned NCAR ensemble, were also available in SFE2018; the MAP and HRRRE are selected for analysis herein because they received the best subjective ratings from SFE participants among the unified ensembles participating in the CLUE.

TABLE 1. Membership configuration of the HREFv2.1. HRW and NAM refer to High Resolution Window and North American Mesoscale Forecast System runs, respectively. Dynamical cores used include the Advanced Research and Forecasting version of the Weather Research and Forecasting Model (WRF-ARW; Skamarock et al. 2008) and the Nonhydrostatic Multiscale Model on the B Grid (NMMB; Janjić and Gall 2012). PBL schemes used include the Mellor–Yamada–Nakanishi–Niino (MYNN; Nakanishi and Niino 2004), Yonsei University (YSU; Hong and Lim 2006), and Mellor–Yamada–Janjić (MYJ; Janjić 1994) formulations. Microphysics schemes include the Thompson (Thompson et al. 2008), WRF single-moment 6-class (WSM6; Hong et al. 2006), Ferrier (Ferrier et al. 2011), and Ferrier–Aligo (Aligo et al. 2018) formulations. IC backgrounds and LBCs are given as the parent NWP model whose analysis or forecast state is used; "−1 h" here indicates that the parent run initialized an hour earlier than the CAM run produces the ICs/LBCs. The HRRR and NAM Nest perform cycled DA using the specified parent model as a first-guess background (Gustafsson et al. 2018), while the HRW members simply interpolate the specified parent's ICs.

Member	Core	PBL	Microphysics	Time lagging	IC background	LBCs	*dx* (km)	In NCEP HREFv2?
HRRR	WRF-ARW	MYNN	Thompson	No	RAP −1 h	RAP −1 h	3.0	No
HRRR −6 h	WRF-ARW	MYNN	Thompson	6-h	RAP −1 h	RAP −1 h	3.0	No
HRW ARW	WRF-ARW	YSU	WSM6	No	RAP	GFS −6 h	3.2	Yes
HRW ARW −12 h	WRF-ARW	YSU	WSM6	12-h	RAP	GFS −6 h	3.2	Yes
HRW NMMB	NMMB	MYJ	Ferrier	No	RAP	GFS −6 h	3.2	Yes
HRW NMMB −12 h	NMMB	MYJ	Ferrier	12-h	RAP	GFS −6 h	3.2	Yes
HRW NSSL	WRF-ARW	MYJ	WSM6	No	NAM	NAM −6 h	3.2	Yes
HRW NSSL −12 h	WRF-ARW	MYJ	WSM6	12-h	NAM	NAM −6 h	3.2	Yes
NAM Nest	NMMB	MYJ	Ferrier–Aligo	No	NAM Nest	NAM	3.0	Yes
NAM Nest −12 h	NMMB	MYJ	Ferrier–Aligo	12-h	NAM Nest	NAM	3.0	No (−6 h)

For instantaneous CREF, NMEPs are evaluated hourly for lead times of 13–30 h,[4] corresponding to the period from 1300 UTC on the initialization date to 0600 UTC on the following date. Surrogate severe forecasts are generated and evaluated for the time-maximum UH values over the convective day, which we define as the 24-h period beginning at 1200 UTC on the initialization date. For any given date, one surrogate severe field covers the entire convective day; this field represents the expected coverage of rotating storms throughout the whole diurnal cycle. Therefore, our CREF verification is much more sensitive to timing errors than our surrogate severe verification. Together, the CREF and UH verification should capture most of what an outlook forecaster at the SPC would be responsible for anticipating.[5]

Verification of CREF forecasts is performed over the CONUS, as well as the SFE daily domains. The daily domain for each date is a rectangular area of 15° longitude by 8.72° latitude manually selected to cover a relevant convective forecast challenge (typically, though not always, near the highest SPC convective outlook risk category). Surrogate severe forecasts are always verified over a domain that covers the eastern two-thirds of the CONUS.

2) OBSERVATION DATASETS FOR VERIFICATION

To verify surrogate severe forecasts, we utilize preliminary local storm reports (LSRs) from the National Weather Service (NWS). Reports of tornadoes, severe hail (exceeding an inch in diameter), and damaging wind gusts (exceeding 58 mph, if measured) are considered. LSRs are mapped onto an 80-km grid that is everywhere zero, except grid cells containing one or more LSRs are assigned a value of 1. This procedure may be interpreted as a neighborhood search that is implicit in the regridding. This field is identical to the OSR81 field used for surrogate severe verification in Sobash et al. (2011, hereafter S11). A single verification field is generated for each convective day using all LSRs that occurred between 1200 UTC on the verification date and 1200 UTC on the following date.

To verify CREF forecasts, the Merged Reflectivity Quality-Controlled Composite (MRQCC) product from the Multi-Radar Multi-Sensor (MRMS; Smith et al. 2016) system is employed. MRQCC is derived by blending data from over 140 operational WSR-88D radars across the United States and over 30 additional radars in Canada. Our largest verification domain for CREF is the continental United States (CONUS), which is covered by the MRMS mosaic. When verifying a model CREF probability field, the corresponding MRMS field is first regridded onto the 3-km model grid. Then, a binary field is computed that is everywhere 0, except it is set to 1 throughout an 80 km × 80 km neighborhood surrounding each point with MRMS reflectivity exceeding 40 dBZ. Conceptually, we are producing the same type of binary verification field for CREF forecasts as for surrogate severe, except the CREF verification field is defined on the 3-km model grid (instead of an 80-km grid).

b. Verification methods

1) COMPUTATION OF BINARY AND PROBABILISTIC FIELDS

As described above, both the forecasts and observations are thresholded and transformed into binary (and also, in

[4] The most restrictive member of HREFv2.1 (HRRR −6 h; Fig. 1) is only available out to a 30-h lead time, and we only wish to verify hourly CREF forecasts with all 10 members available.

[5] Identifying convective mode (e.g., linear versus multicellular versus supercellular) is also critically important for these types of forecasts. While no mature, practical methods exist for objective verification of mode, our surrogate severe verification should generally reward correct forecasts of rotating storms (or lack thereof).

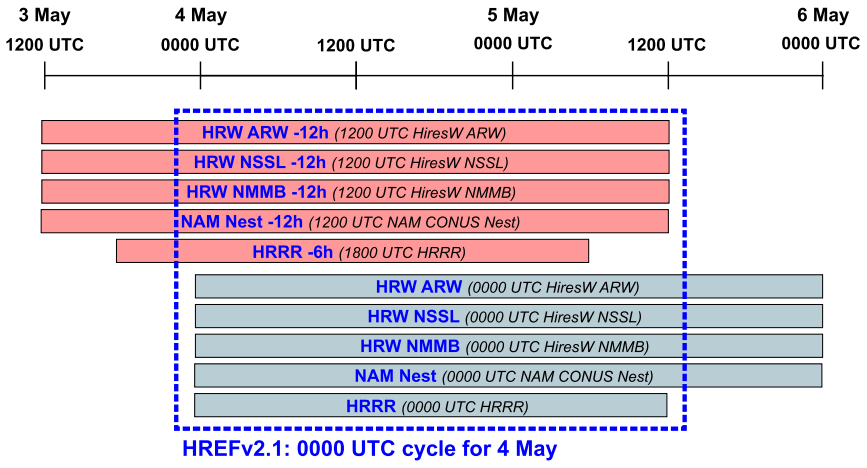

FIG. 1. Deterministic CAM membership in a hypothetical 0000 UTC run of the HREFv2.1. On each bar, the HREF member name is given in bold blue text, while the member's deterministic run name is given in italicized black. The dashed blue box contains the forecast times from each member that participates in the 36-h EO forecast. Each deterministic member runs out to a lead time of at least 48 h, except for the HRRR, whose forecast ends at 36 h.

some cases, smoothed probabilistic) fields before verification proceeds, a process illustrated in Fig. 2. In all cases, the neighborhood of grid points is searched for its *maximum* value; the remaining distribution of values in the neighborhood does not impact verification. The neighborhood maximum operation addresses the question of whether *any* storm or instance of severe weather exists nearby, which closely mimics the forecast problem for SPC convective outlook forecasters, who issue probabilities for severe weather occurrence within 40 km of a point. For observations (LSRs or MRMS), the resulting thresholded binary field (Fig. 2b) is used directly for verification. For forecasts (CREF or UH), an additional smoothing step is applied using a two-dimensional Gaussian kernel as in Eq. (1) of Hitchens et al. (2013). The resulting smoothed field (Fig. 2c) contains a continuous distribution of fractional values, which may be interpreted as probabilities of threshold exceedance in the neighborhood (as in S11). In the case of surrogate severe forecasts, numerous iterations of the field are computed using a range of σ values from 40 to 300 km. Thus, our surrogate severe verification allows us to assess how forecast skill varies with the smoothing length scale. In the case of CREF, this approach is computationally prohibitive owing to the much finer 3-km verification grid, so we only produce NMEPs with $\sigma = 40$ km; this value is commonly used for operational SPC CAM guidance (e.g., R19).

Verification of individual ensemble member forecasts involves comparing their smoothed probability field against the corresponding binary verification field. For verifying an ensemble system, we represent its forecast as the ensemble mean of smoothed member probability fields, which is equivalent to the smoothed NMEP field (e.g., R19). For the remainder of this paper, NMEP always refers to the smoothed version, as unsmoothed NMEPs are not verified herein.

When verifying threshold exceedance probabilities for an EO such as the HREF, separate bias correction of each member is desirable in order to retain only "good spread"—ensemble variance that improves forecast skill metrics and owes to diverse model attractors and/or plausible IC uncertainty, rather than disparate member biases that widen the PDF only through systematically offsetting errors—from the diverse configuration choices (Eckel and Mass 2005). Furthermore, in the case of UH, there is no truth field available whose magnitude is directly comparable to the forecast quantity (which itself can vary widely with model grid spacing and other configuration choices). To address these challenges, we compute climatologies for each member of each ensemble over the entire verification dataset for both CREF and UH. These climatologies allow us to map member CREF and UH values into climatological percentile space, where they can then be treated equitably across members (and even across different ensembles). This is fundamentally similar to the "quantile mapping" approach (e.g., Hopson and Webster 2010; Voisin et al. 2010), with the distinction that both our observed and forecast cumulative distribution functions are formed from the set of all grid points across the domain and over all 21 days we are verifying. Before computing the UH climatology, 3-km model values are first remapped to the 80-km surrogate severe grid such that each 80-km grid cell is assigned the maximum value among all 3-km grid cells inside it (this regridding is implicitly a neighborhood-maximum operation). For CREF, the 3-km model gridpoint values are used for the climatology. When computing thresholded forecast fields, all thresholds are based on percentiles from the climatology. In the case of CREF, for each ensemble member, whichever percentile matches the 40-dBZ threshold in the MRMS dataset is used as the forecast threshold. More details are given in appendix.

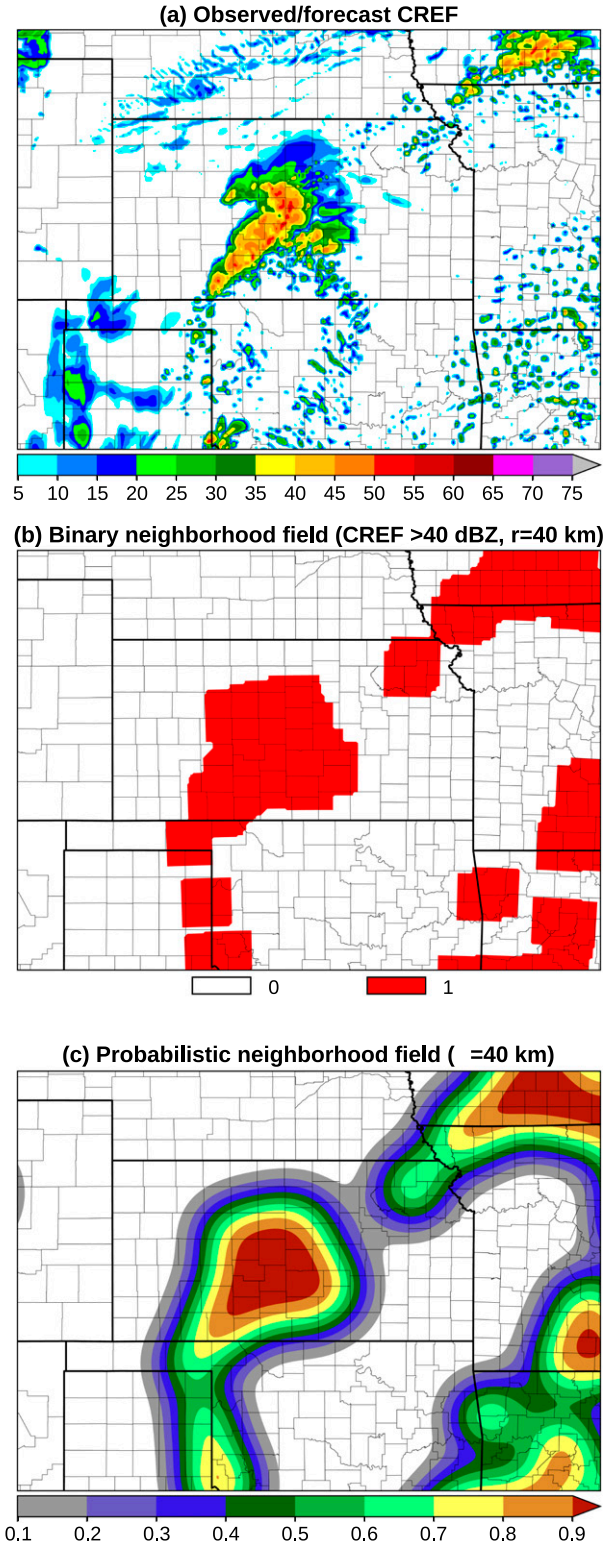

FIG. 2. Example of computing thresholded probability fields from a CREF field. (a) The observed or forecast CREF. (b) A binary field whose value at each point is 1 if the threshold is exceeded in the local neighborhood, and 0 otherwise. (c) The result of applying a Gaussian smoother to (b), which gives a continuous field with values in the range [0, 1]; these values can be interpreted as probabilities. For ensembles, NMEPs are equivalent to the ensemble mean of the member probability fields.

2) SKILL SCORES AND METRICS

Standard metrics for probabilistic forecast verification are employed herein. The first is the relative operating characteristic (ROC; Mason 1982) area under the curve (AUC). For computing the AUC, trapezoidal integration is employed within the Model Evaluation Tools (MET; Fowler et al. 2017) 7.0 software suite. AUC measures the ability to discriminate between events and nonevents; it is conditioned on the observations (i.e., whether or not the event occurred) and is insensitive to forecast bias. AUC ranges from 0 to 1: a value of 1 is a perfect forecast, while 0.5 indicates no skill, and 0.7 is sometimes used as the minimum score for a useful forecast.

The fractions skill score (FSS; Roberts and Lean 2008, hereafter R08) is computed for both the surrogate severe forecasts and CREF probabilities. However, a notable departure from the R08 definition is that at each grid point, the fractional probability value from our forecast (either a smoothed member probability field or ensemble NMEP) is substituted for the true neighborhood fractional coverage; this is similar to the approach of Schwartz et al. (2010). Our formulation of FSS is as follows:

$$\mathrm{FSS} = 1 - \frac{\sum_{i=1}^{N} (P_{F(i)} - B_{O(i)})^2}{\sum_{i=1}^{N} (P_{F(i)}^2 + B_{O(i)}^2)}, \tag{1}$$

where $P_{F(i)}$ and $B_{O(i)}$ are the forecast probability and observed binary value, respectively, at the ith grid point. Note that we use the observed binary field, rather than a smoothed (e.g., practically perfect; Schwartz et al. 2010) field, as truth. This avoids potential penalization of spatially precise forecasts due to ad hoc smoothing of observations whose location is actually known with certainty (an issue discussed at length in section 4). Our version of FSS is closely related to the Brier skill score (Brier 1950); the two are differentiated chiefly by the reference forecast for FSS considering both the truth *and* forecast fields. Specifically, the reference forecast represents one in which the mean squared values of both the forecast and truth fields are held constant, but they are redistributed in space to be maximally nonoverlapping; in other words, the worst possible forecast that could be made while retaining the existing distribution of fractional values in the probability forecasts and observed binary fields. FSS ranges from 0 to 1: a value of 1 is a perfect forecast, and 0 is the worst possible forecast containing the same distribution of forecast and observed fractions. Although 0.5 has been suggested as the lower limit for a useful forecast in the literature, we caution that this does not apply to our formulation of FSS, given: 1) the neighborhood-maximum operation in our verification, which departs from the neighborhood-coverage-based definition in R08; and 2) our use of a binary truth field. Because of (2), we

←

expect substantially lower FSS scores[6] than in studies that use a fractional truth field, so our FSS scores should not be compared directly with such values. For CREF FSSs averaged over all cases in the verification dataset, we compute 90% confidence intervals (CIs) for each ensemble mean and member forecast using the bootstrapping technique described in Wilks (2011) with 10 000 resamples.

To evaluate the reliability of our CREF probability forecasts, we create attributes diagrams (Hsu and Murphy 1986). Additionally, we compute the so-called reliability component of the Brier score as follows:

$$\mathrm{BS}_{\mathrm{REL}} = \frac{1}{n}\sum_{i=1}^{I} N_i(f_i - \overline{x_i})^2, \tag{2}$$

where I is the number of probability bins; n is the total number of grid points in the dataset; N_i is the number of grid points with probabilities in bin i; f_i is the forecast probability value associated with bin i (e.g., $f_i = 0.2$ for the bin $0.15 < P < 0.25$); and $\overline{x_i}$ is the base rate of the observed binary field in bin i (i.e., the frequency of event occurrence when probabilities in that bin were forecast). $\mathrm{BS}_{\mathrm{REL}}$ is 0 for a perfectly reliable forecast and increases as reliability (weighted by bin forecast frequency) becomes worse. In the same manner as described for CREF FSSs, we compute 90% CIs for $\mathrm{BS}_{\mathrm{REL}}$ and for the base rate within each bin.

In addition to verifying NWP forecast skill, we also apply spread–skill metrics traditionally used for continuous variables (e.g., temperature) to our CREF probability forecasts. Specifically, we compute the mean squared error (MSE) and mean ensemble variance (MEV) for each of the 378 hourly CREF snapshots. In this context, ''mean'' refers to the spatial average over the verification grid (e.g., the mean statistic for all grid points in the CONUS at a particular verification time); ''error'' refers to the difference between the ensemble NMEP value and practically perfect[7] value at a grid point; ''variance'' is computed for the set of all member probability values at the grid point; and terms similar to Bessel's correction are included when calculating MSE and MEV, following appendix B of Eckel and Mass (2005). Through the neighborhood-maximum operation we are transforming a field of discontinuous, sparse features (storms with CREF > 40 dBZ) into a more continuous field (viz., the probability that a storm exists in the general area). After this transformation, we then examine MSE and MEV under the implicit assumption that the traditional spread–skill relationship can be expected to hold; this assumption may be explored more rigorously in future work to elucidate precisely what the MSE–MEV relationship signifies under different conditions for NMEPs. We also compute the consistency ratio (CR), defined as the ratio of the MEV to MSE for the aggregate of all forecast cases in our dataset: a system with perfect statistical consistency has a CR of 1, while CR less than 1 indicates aggregate ensemble underdispersion, and greater than 1 indicates overdispersion.

[6] Testing multiple formulations of the neighborhood-maximum-based FSS for the same set of cases revealed that an FSS of ~0.65 using a smoothed truth field is equivalent to an FSS of ~0.4 using a binary truth field. This is valid for an 80 km × 80 km neighborhood and a smoothed field with σ = 40 km.

[7] Strictly for this spread–skill analysis, we apply a Gaussian smoother to the observed binary field (using the same σ = 40 km as the forecast fields), yielding a ''practically perfect'' truth field. This is necessary in order to ensure MSE and MEV are directly comparable in magnitude.

3. Verification of composite reflectivity forecasts

a. Forecast skill

Figure 3a presents FSSs for CREF forecasts aggregated over all 378 snapshots for the CONUS domain. For each of the three ensembles (bottom), the mean FSS of its member probability fields is displayed as a color-coded bar outlined in black, while the FSS of the ensemble NMEPs is displayed as a red bar. The difference between the member mean (color-coded) and NMEP (red extension) FSS, which we will call $\mathrm{FSS}_{\mathrm{gained}}$ (i.e., the skill gained by the ensemble relative to its constituent member solutions), is annotated to the right of the bars. In some sense, $\mathrm{FSS}_{\mathrm{gained}}$ should indicate how effectively an ensemble is utilizing its members to fill out a realistic forecast PDF. Unsurprisingly, there is more variability in skill among the probability fields of HREFv2.1 members than HRRRE or MAP members, confirming that equal likelihood of member solutions cannot reasonably be expected for this type of EO. For the ensemble NMEPs, HREFv2.1 performs best (0.49; 90% CI nonoverlapping with the other two ensembles), while HRRRE (0.42) and MAP (0.45) lag behind. However, it is striking how much smaller the gap in skill is between the three systems with respect to the mean of their member forecasts: indeed, HREFv2.1 ensemble NMEPs show roughly double the $\mathrm{FSS}_{\mathrm{gained}}$ (+0.07), compared to HRRRE (+0.04) or MAP (+0.03). When focusing on the SFE daily domains (Fig. 3b), performance differences between systems and members are broadly similar to those over the CONUS. One difference over the daily domains is that MAP members are actually more skillful overall than HREFv2.1 members, resulting in statistically similar NMEP FSSs for those two systems. Nonetheless, HREFv2.1 again exhibits substantially larger $\mathrm{FSS}_{\mathrm{gained}}$. Within MAP, MAP 01 shows consistently better skill than other members, suggesting the analysis produced by the EnVar control is more skillful than the recentered EnKF analyses.

Figures 4a and 4b show ROC AUC values for the CONUS and SFE daily domains, respectively. The relative differences between members and ensembles are once again similar: MAP members outperform HREFv2.1 and HRRRE members over the SFE daily domains (with comparable skill over the CONUS), but the ensemble NMEPs are most skillful from HREFv2.1. $\mathrm{AUC}_{\mathrm{gained}}$ is larger for HREFv2.1 than for HRRRE or MAP by nearly a factor of 2.

In summary, the skill scores for CREF probability fields have the following properties:

1) CAM skill is modestly better when computed over the entire CONUS than when limited to the SFE daily domains;

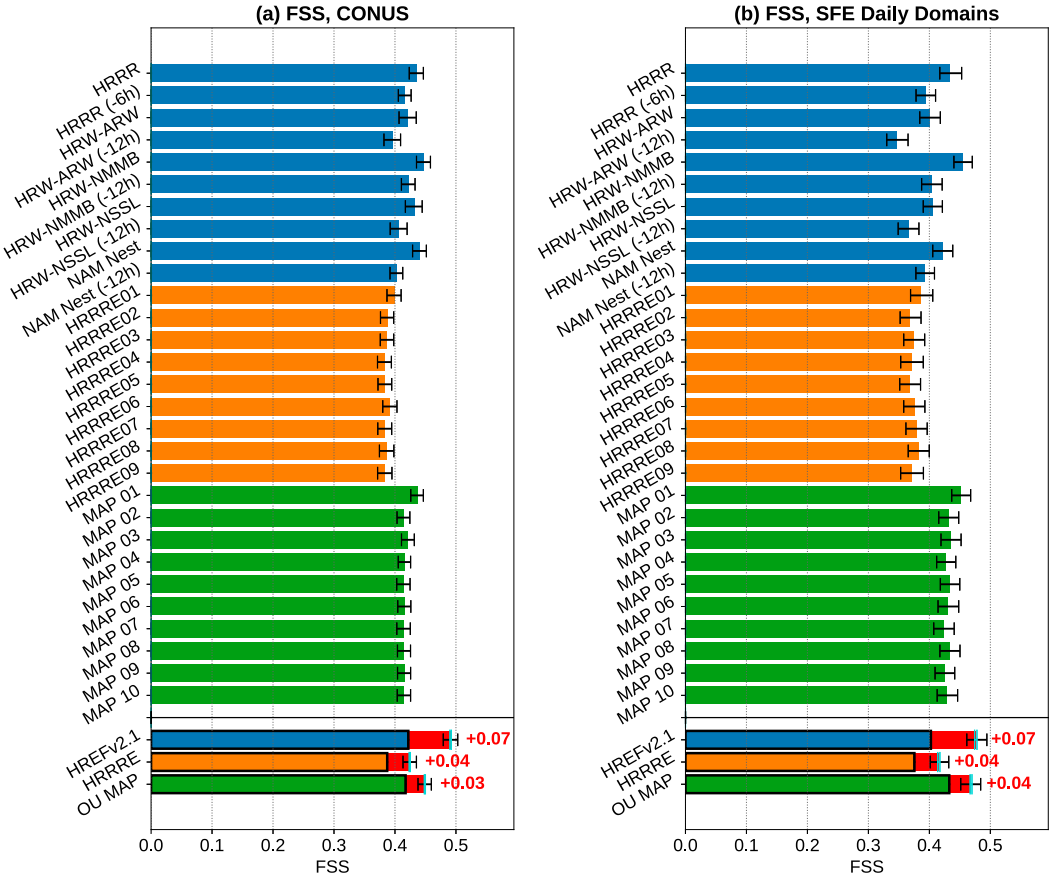

FIG. 3. For CREF > 40 dBZ smoothed probability fields (for members) and NMEPs (for ensembles), mean FSS over the 378 hourly snapshots for (a) CONUS and (b) SFE daily domains. For the three ensembles at bottom, the bolded, color-coded bar shows the mean FSS of its individual member forecasts (each of which appears as its own bar above), while the red bar extends to the FSS for the ensemble mean forecast. The length of the red extension thus represents FSS_{gained}, which is also annotated as red text to the right. CIs at the 90% level for FSS values are shown as black error bars, and separately for FSS_{gained} values as cyan error bars (bottom three bars only; these CIs are quite small and in some cases appear as a single cyan line).

this is presumably due to the abundance of "easy nulls" over the CONUS.

2) Individual member skill, in the aggregate, is best for MAP members, followed closely by HREFv2.1 members, and then HRRRE members.
3) Ensemble skill, in the aggregate, is best for HREFv2.1, followed by MAP, and then HRRRE.
4) The skill added by the ensemble NMEPs over their constituent member probability fields is *substantially* larger for HREFv2.1 than for MAP or HRRRE, suggesting HREFv2.1 contains more useful ensemble spread with respect to convective storm coverage and placement.
5) The variation in skill among individual HREFv2.1 members is larger than that among MAP or HRRRE members.

b. *Forecast spread*

To evaluate how each member is contributing spread to its parent ensemble, the coefficient of determination (r^2; the square of the Pearson correlation coefficient) is computed for the CREF probability forecasts of every possible pair of members within each system. For an N-member ensemble, there are $N(N-1)/2$ such pairs. Correlation matrices for the CONUS domain are presented in Figs. 5a–c. In terms of the mean r^2 among all member pairs, HREFv2.1 is least correlated, followed by HRRRE, and then MAP. For HRRRE and MAP, the strength of correlation between one pair of members is quite similar to any other possible pair; an exception is that MAP 01 (the MAP control member) is more similar to its sibling members than they are to one another, as expected for a control member. For HREFv2.1, however, substantial differences in r^2 values exist among member pairs. The most similar pair, HRRR and HRRR −6 h (two identically configured runs initialized 6 h apart), have an r^2 value comparable to pairs of HRRRE members. Otherwise, HREFv2.1 correlations are relatively low, with some clustering evident by model core (WRF-ARW versus NMMB) and ICs (NAM versus RAP). When member probability forecasts are compared over only the SFE daily domains (Figs. 5d–f), r^2 values decrease modestly

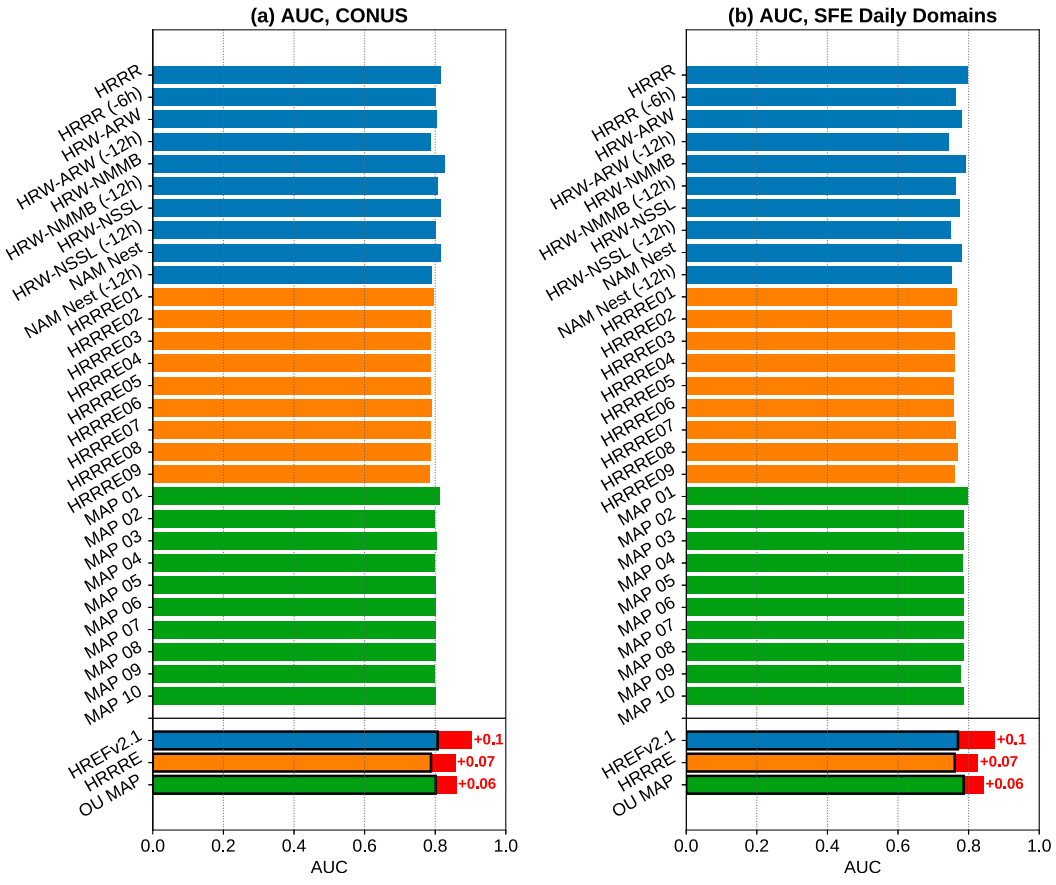

FIG. 4. As in Fig. 3, but for ROC AUC, and CIs are omitted.

for all pairs compared to the CONUS domain. Over the daily domains, the mean correlation magnitude for MAP pairs ($\overline{r^2} = 0.46$) is fully twice as large as for HREFv2.1 pairs ($\overline{r^2} = 0.23$), suggesting that meaningful spread in HREFv2.1 is substantially larger than in MAP (which includes no sampling of model uncertainty).

Another perspective on the spread contributed by each ensemble member can be gained by identifying grid points at which the member's smoothed probability field differs from its parent ensemble's NMEP by a value exceeding some threshold. We will call such points "outlier points:" here, the member is either 1) predicting storms in an area where most other members do not, or 2) failing to predict storms in an area where most other members do. Gridpoint frequencies for outlier points in each ensemble member are given in Figs. 6a–c for the CONUS, and in Figs. 6d–f for the SFE daily domains. For the most stringent threshold of 0.7 (Figs. 6c,f), over both domains, the typical HREF member has about three times as many outlier grid points in the verification dataset as does the typical HRRRE or MAP member. This discrepancy is somewhat less pronounced for thresholds of 0.6 (Figs. 6b,e) and 0.5 (Figs. 6a,d). Nonetheless, it is clear that a typical HREF member departs sharply from its parent ensemble on forecasting storm occurrence or nonoccurrence more frequently than a typical HRRRE or MAP member. In terms of the ensemble PDF, this means long tails are more often present in HREFv2.1.

Figure 7 presents an attributes diagram for NMEPs from the three ensembles over the CONUS. HREFv2.1 exhibits remarkably good reliability, while HRRRE and MAP are both somewhat overconfident (i.e., low NMEPs are underforecasts and high NMEPs are overforecasts). BS_{REL} (for which zero represents perfect reliability) is 5 times larger for HRRRE, and 7 times larger for MAP, than for HREFv2.1. For the SFE daily domains (Fig. 8), qualitatively similar results hold. However, there is some notable degradation of reliability for NMEP $>$ 0.6 in HREFv2.1, where it becomes similarly overconfident to HRRRE and MAP. Nonetheless, differences in BS_{REL} favor HREFv2.1 by about an order of magnitude over the HRRRE and MAP. Also noteworthy in the bin histograms is the underrepresentation of HREFv2.1 in both the smallest and largest probability bins ($P \leq 0.05$ and $P > 0.95$, respectively), a reflection of its more frequent forecasts falling into intermediate bins associated with meaningful member disagreement.

Figure 9a presents CR as a function of lead time over the full verification period and full CONUS. The CR tends to increase during the diurnal convective maximum (lead times of 18–26 h, corresponding to 1800–0200 UTC daily), but generally does not vary in time by more than about 30% for a given ensemble. HREFv2.1 ($\overline{CR} = 1.01$) demonstrates remarkably

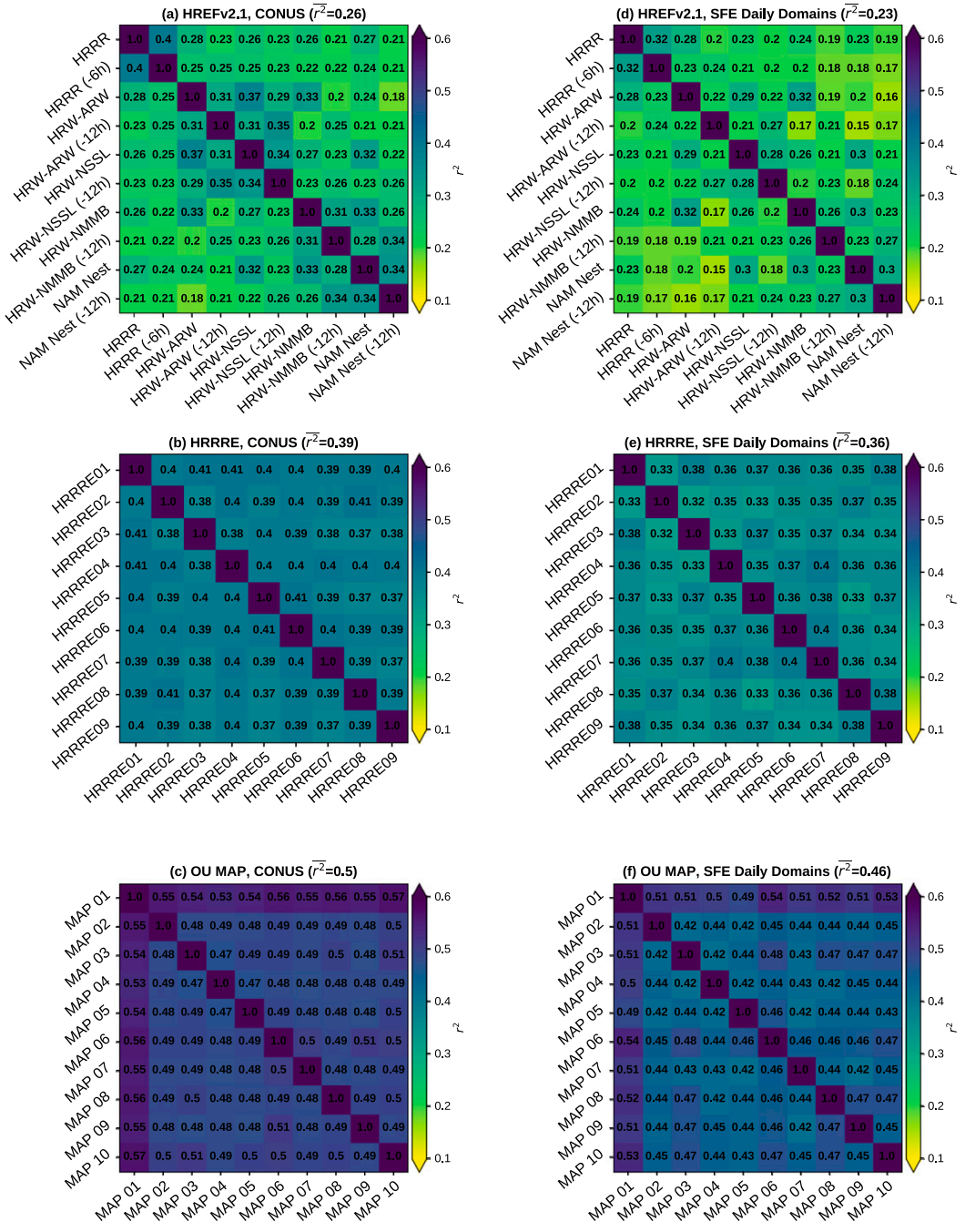

FIG. 5. Matrices of the coefficient of determination (r^2) between ensemble member CREF > 40 dBZ probability fields for (a) HREFv2.1, (b) HRRRE, and (c) MAP over the CONUS domain and across all 378 snapshots. (d)–(f) As in (a)–(c), but over the SFE daily domains. In each panel, the mean r^2 of all unique pairs of members is given above the matrix.

good statistical consistency, while HRRRE ($\overline{\text{CR}} = 0.46$) and MAP ($\overline{\text{CR}} = 0.37$) are quite underdispersive. For the SFE daily domains (Fig. 9b), the results are very similar. While these CR values reveal much about the total ensemble spread aggregated over all cases, they do not address whether MEV for a single hourly snapshot is a good predictor of MSE for that same snapshot, as is true of an ideal ensemble. To evaluate this, we also compute r^2 between MSE and MEV for the set of all 378 snapshots. Figure 10 presents scatterplots of MEV versus MSE for all three systems and both verification domains. For an ideal ensemble in which MEV = MSE, all points would lie along the red line. However, even for highly underdispersive

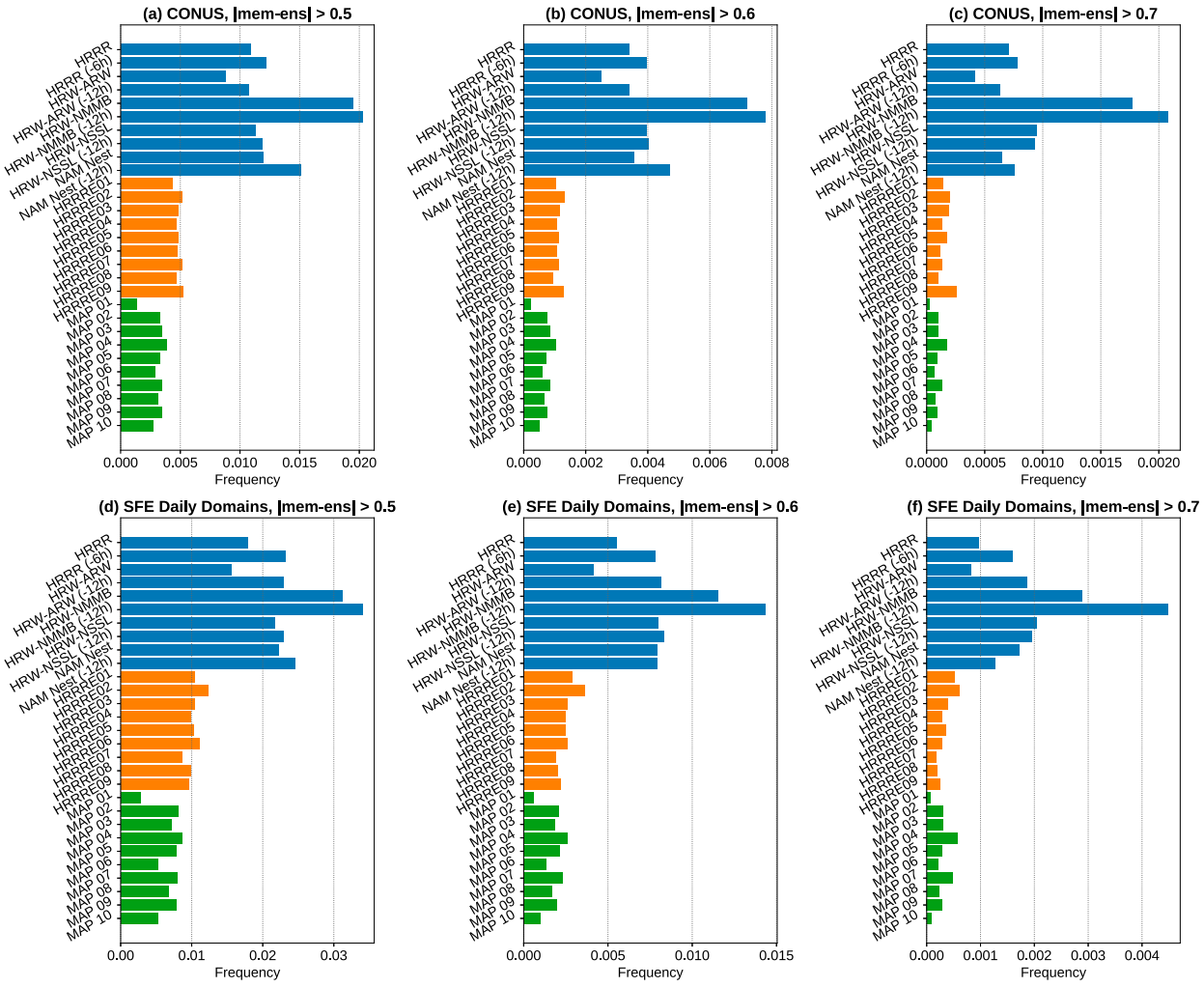

FIG. 6. Over the CONUS domain, the fractional gridpoint frequency of each ensemble member's CREF $>$ 40 dBZ probability field differing in magnitude from its parent ensemble's NMEP by at least (a) 0.5, (b) 0.6, and (c) 0.7. (d),(e),(f) Gridpoint frequencies for the same respective thresholds are shown for the SFE daily domains. Note that the range of the abscissa changes between panels.

ensembles, a strong correlation between MEV and MSE (i.e., points lying along a line of lesser slope passing through the origin) is still desirable. Over the CONUS, r^2 is quite high for all three systems; however, r^2 drops into the 0.5–0.65 range for the SFE daily domains. This is a reflection of the relative abundance of easy correct nulls over the CONUS domain, as storms can be expected not to exist in most areas most of the time. To focus more narrowly on grid points with meaningful forecast challenges, Fig. 11 presents the same statistics when all correct nulls (i.e., grid points where all ensemble members have a zero probability *and* no storm is observed nearby in reality) are removed. As expected, r^2 is reduced substantially in this dataset; however, the reduction is much less severe for HREFv2.1 than for HRRRE or MAP. In fact, r^2 is 40%–70% larger for HREFv2.1 than for the other two systems over both domains. This implies that ensemble disagreement regarding the presence of storms within HREFv2.1's membership predicts MSE better than it does in HRRRE's or MAP's.

4. Verification of surrogate severe forecasts

As described in section 2, surrogate severe forecasts are computed on an 80-km grid, making verification computationally cheaper than CREF NMEPs. This affords us the opportunity to verify surrogate severe forecasts over a range of UH percentile thresholds and Gaussian σ values, giving insight into the dependence of CAM ensemble UH forecast skill on intensity and smoothing length scale. Figure 12 presents ensemble surrogate severe AUC (left) and FSS (right) as a function of these parameters for HREFv2.1, HRRRE, and MAP. In each panel, the ensemble's maximum score within the percentile-σ parameter space is represented by a white square and annotated with the AUC or FSS value. White circles, which are shaded by score using the main color scale, represent the maximum scores achieved by each ensemble member's surrogate severe forecast (e.g., on the HRRRE FSS panel, the maximum score achieved by member HRRRE01's forecasts will be plotted as a circle at the σ-percentile coordinate where that score occurs).

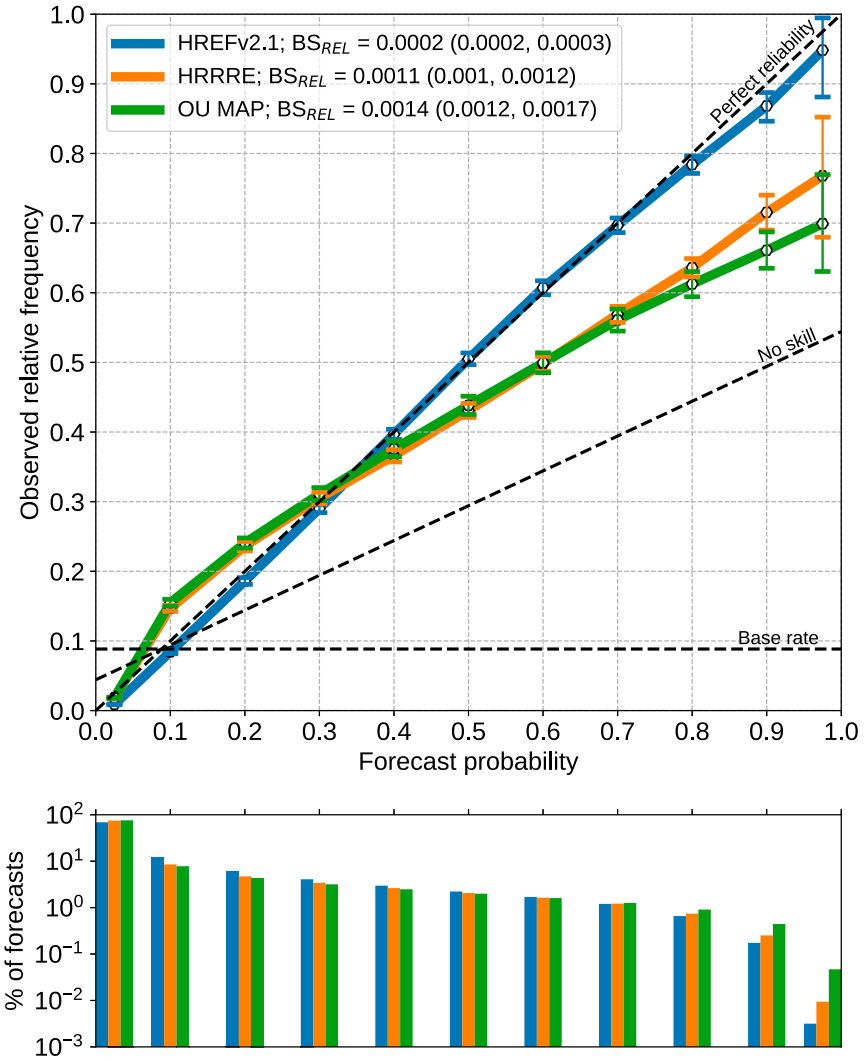

FIG. 7. Attributes diagram for CREF NMEPs over the CONUS. Binned probabilities are used; each bin is plotted as a white dot on the diagram, with a colored curve connecting all the dots for each ensemble. The 90% CIs for the base rate in each bin are plotted as error bars, also colored by ensemble. Below the attributes diagram, bars give the frequency of occurrence of probabilities within each bin (as a percentage of all grid points in the dataset). The reliability component of the Brier score (BS_{REL}) for each ensemble is given in the legend, with its 90% CI in parentheses.

For AUC, maximum scores for each ensemble are attained at relatively low intensities in the 75th–85th percentile range. The performance ranking of the three ensembles by maximum AUC score, with HREFv2.1 first and HRRRE last, matches our CREF verification results for ensemble NMEPs. Interestingly, the σ value associated with the maximum AUC also varies considerably between ensembles: MAP AUC values are highest with relatively strong smoothing ($\sigma \sim 110$ km), whereas HREFv2.1 achieves its highest AUC value with less smoothing ($\sigma \sim 75$ km). For FSS, the ranking of the three ensembles remains the same as for AUC, though the performance gap between HREFv2.1 and MAP is larger than for AUC. Higher intensities in the 85th–90th percentile range, and weaker smoothing, are required to maximize FSS than AUC. The former is likely true because AUC tends to reward the overforecasts resulting from choosing a low UH threshold (e.g., Gallo et al. 2016), whereas FSS has a more balanced response to the trade-off between POD and FAR. As with AUC, HREFv2.1 maximizes FSS at a smaller σ than does HRRRE or MAP.

For both AUC and FSS, the individual member surrogate severe forecasts (white circles) consistently require stronger smoothing to optimize skill than the ensemble surrogate severe forecasts (solid white squares). This result, combined with the notable difference in score-maximizing σ values between three ensembles' surrogate severe forecasts, motivates us to revisit

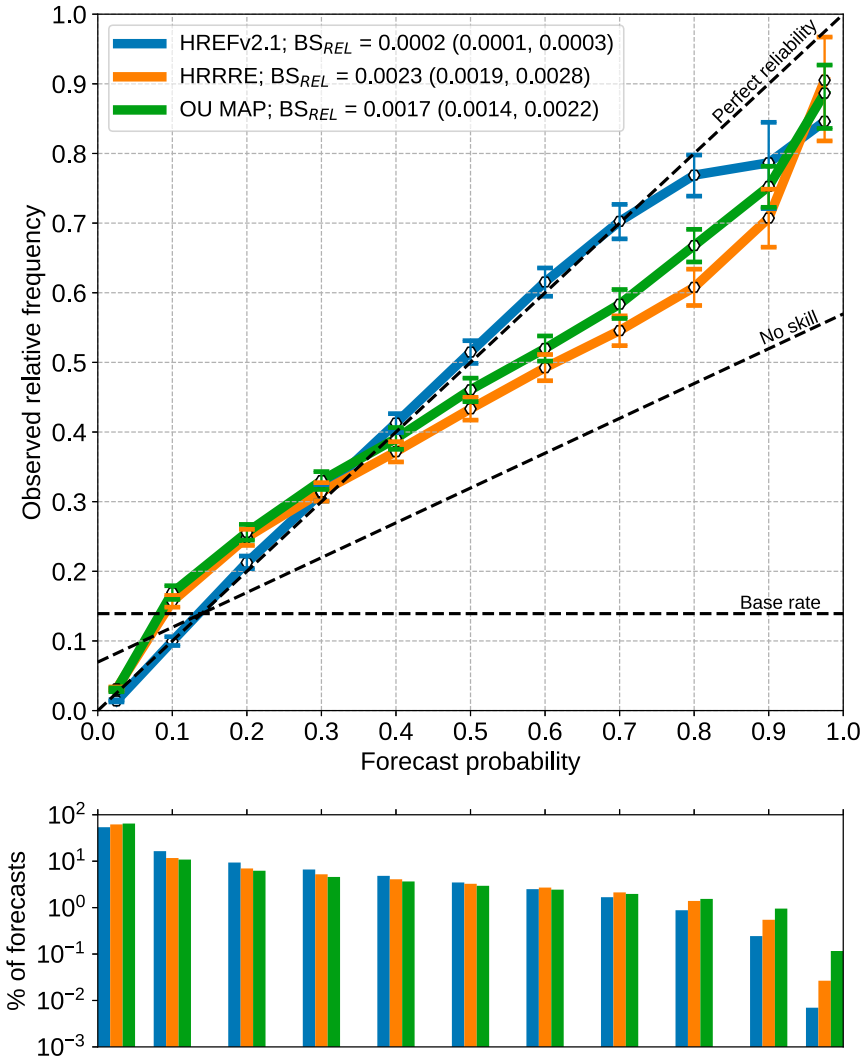

FIG. 8. As in Fig. 7, but for the SFE daily domains.

the history of and best practices for spatially filtering CAM NMEPs. When computing practically perfect truth fields, Hitchens et al. (2013) chose a two-dimensional Gaussian kernel that effectively used $\sigma = 120$ km, in part because this degree of smoothing "better represent[ed] the outlooks issued by the SPC" than other values they tested. Separately, studies verifying surrogate severe forecasts from deterministic CAMs (S11) and CAM ensembles (Sobash et al. 2016b,a; Loken et al. 2017; Sobash et al. 2019) have typically found skill (e.g., FSS and reliability) maximized at $\sigma \geq 120$ km; due to these findings, in some cases, $\sigma = 120$ km is simply chosen as the default value (e.g., Sobash and Kain 2017). Some trade-offs entailed in varying σ are explored in S11: larger values tend to improve reliability (to a point), but reduce sharpness and virtually eliminate coverage of high probabilities. As cautioned by Schwartz and Sobash (2017), defining the neighborhood and smoothing length scales separately can complicate interpretation and poses a risk of conflating scales. For example, when FSS is computed across smoothing length scales using smoothed continuous (instead of binary) truth fields (as in S11 and others), smoothing applied to the practically perfect fields typically varies to match what is applied to the surrogate severe forecasts. When such FSSs are found to be maximized at a large σ value, it does not necessarily imply that strong smoothing produces the most skillful forecasts at the true *neighborhood* length scale (which is almost always defined by a radius of 40 km, in line with the SPC's convective outlook definition). Because our FSSs herein use a binary truth field, this caveat does not apply, and our FSSs should directly reflect forecast skill in answering the question: "What is the probability of a severe weather event within 40 km of this point?" When we assign our smoothing length scale σ to exceed the neighborhood size, then, it is simply a postprocessing technique that can potentially improve skill in predicting neighborhood-scale probabilities by accounting for uncertainty in storm placement. Given similar skill of the resulting fields, using smaller σ should be preferred operationally in order to retain smaller-scale spatial detail in the forecast. Thus, HREFv2.1 demonstrates

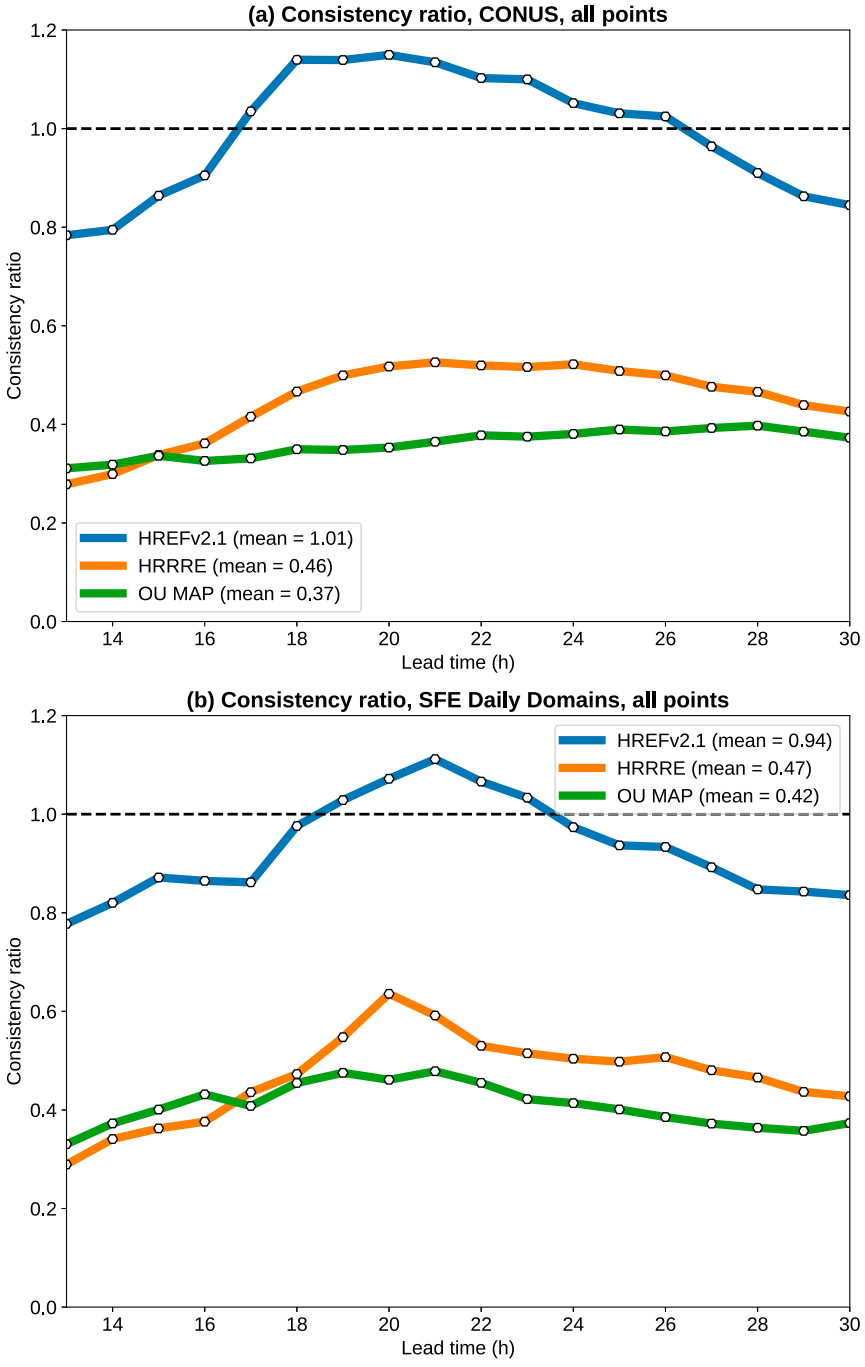

FIG. 9. The consistency ratio of ensemble CREF > 40 dBZ NMEPs from each system as a function of forecast lead time over the (a) CONUS and (b) SFE daily domains. The mean consistency ratio over all lead times is given in the legend on each panel. The dashed line denotes the ideal consistency ratio of unity.

added value over HRRRE and MAP not only in terms of maximum skill scores, but also by achieving those scores with smaller σ (Fig. 12). Note that the individual members of all three ensembles tend to maximize AUC and FSS in approximately the same σ range, implying that HREFv2.1 performs best at smaller σ primarily because of complementary information from its diverse members; not because its member surrogate severe forecasts individually need less smoothing.

Figure 13 presents $\text{AUC}_{\text{gained}}$ and $\text{FSS}_{\text{gained}}$ for the three ensembles (i.e., the score of the ensemble surrogate severe forecasts minus the mean score of the member surrogate severe forecasts). $\text{FSS}_{\text{gained}}$, in particular, highlights HREFv2.1's

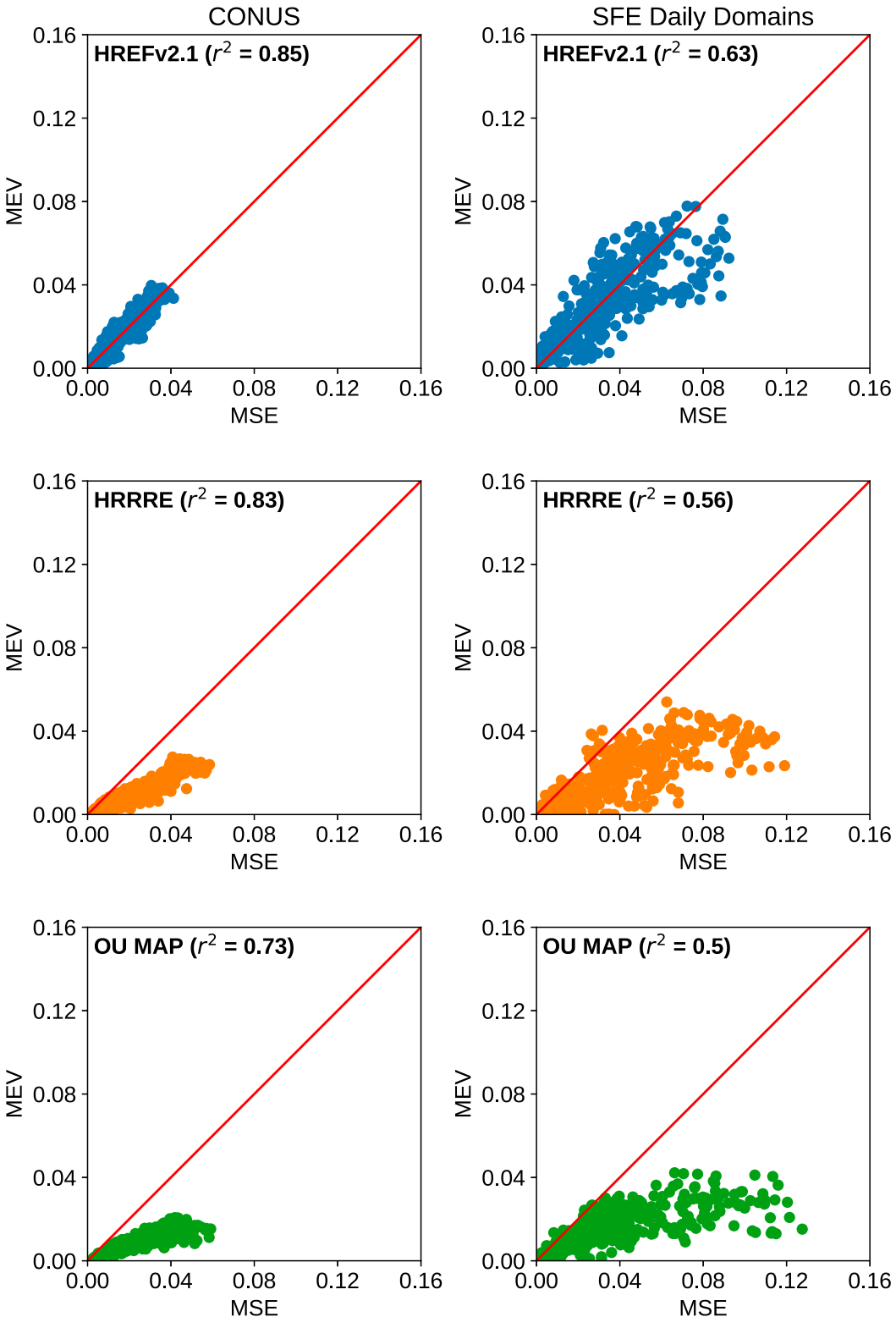

FIG. 10. Scatterplots of mean ensemble variance (MEV) vs mean squared error (MSE) of CREF > 40 dBZ NMEPs for the three ensemble systems over the CONUS and SFE daily domains. Each point on the scatterplot represents the MEV and MSE for one of the 378 snapshots. In each panel, the coefficient of determination r^2 is given in the label at top left. The red diagonal line denotes perfect correspondence between the MEV and MSE; a snapshot with perfect statistical consistency will lie along this line.

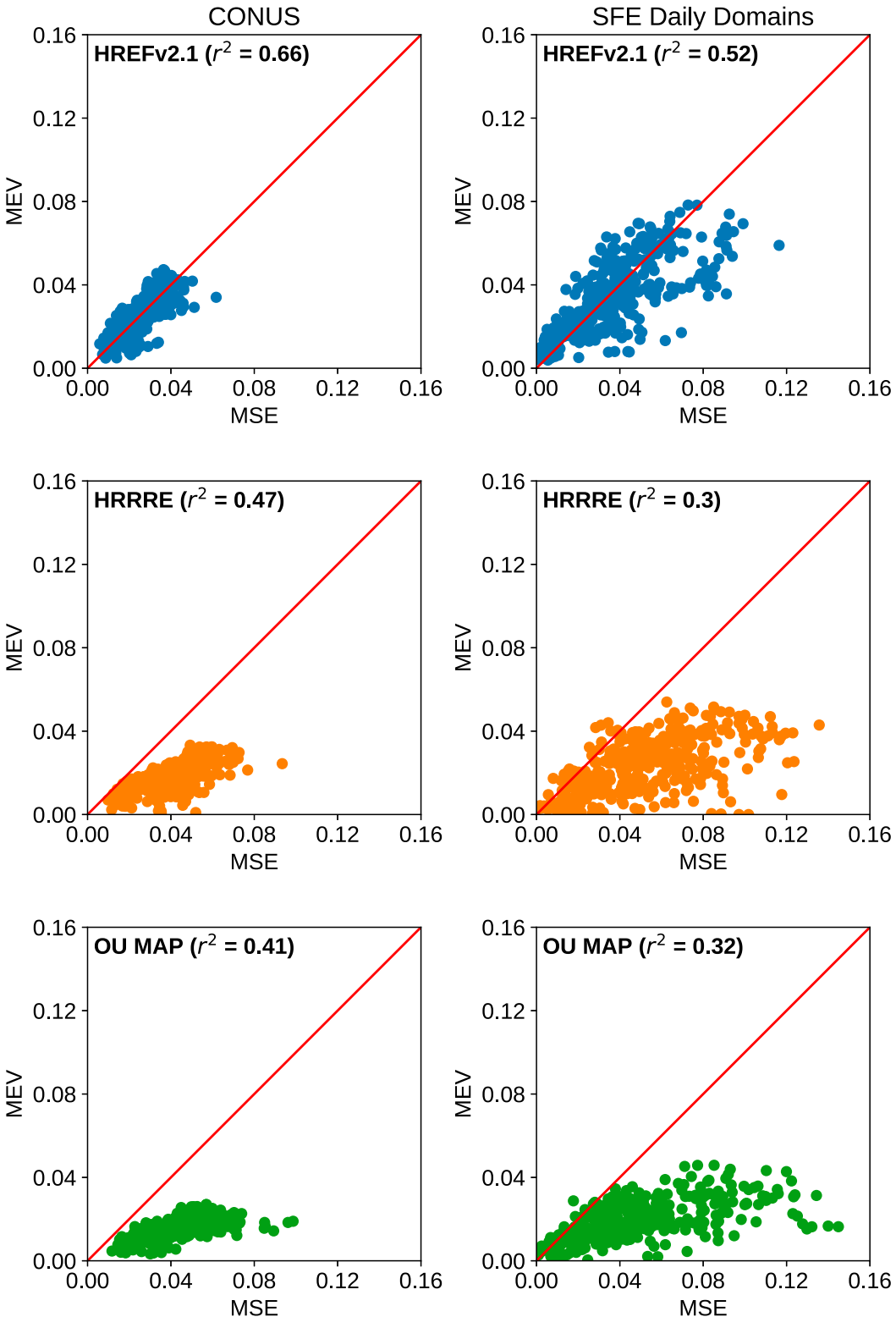

FIG. 11. As in Fig. 10, but grid points with values of zero in both the forecast and observed fields (correct nulls) are excluded from the calculation. For any single snapshot, the total squared error and total variance summed over the domain remains as in Fig. 10 (since the excluded correct null points are zero in all fields), but the MEV and MSE may change due to fewer grid points being considered in the average.

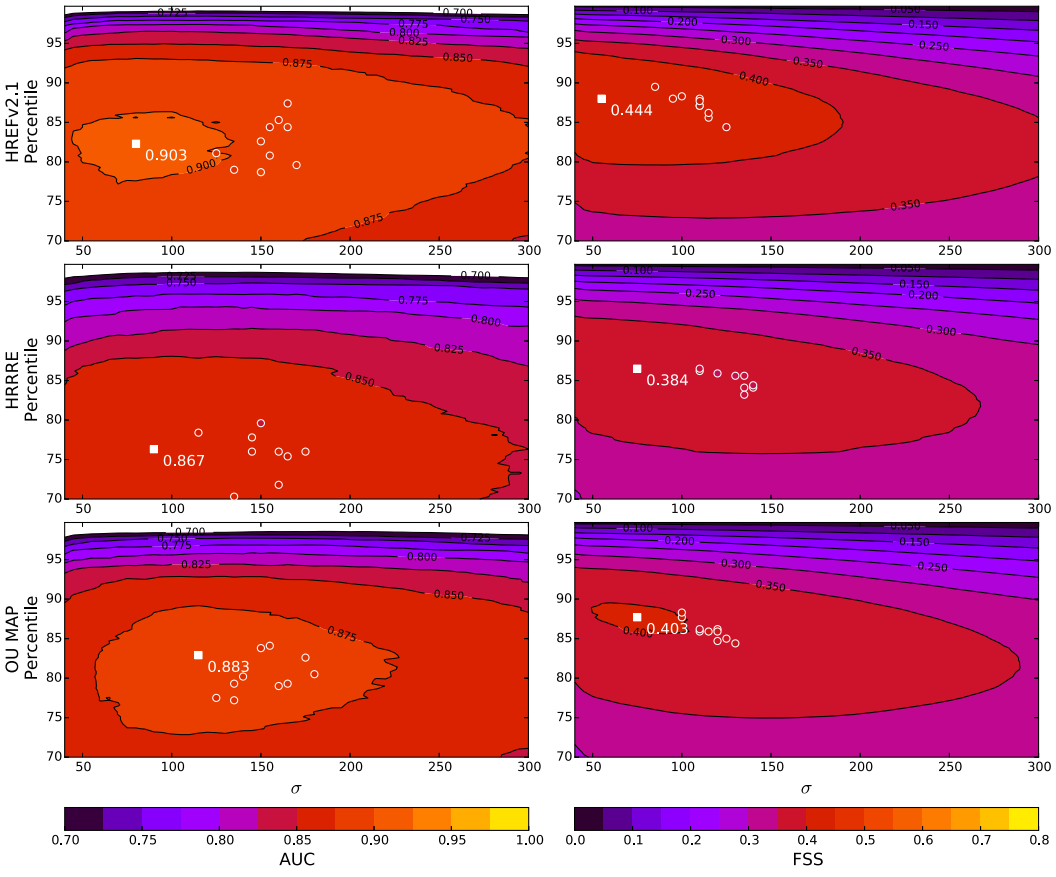

FIG. 12. For the ensemble mean surrogate severe forecasts, the aggregate (left) AUC and (right) FSS for all 21 days in the verification dataset over the eastern 2/3 CONUS domain. Scores are displayed as a function of the Gaussian smoothing σ (abscissa) and the member UH climatology percentile (ordinate). On each panel, the σ and percentile where the maximum score is achieved for the ensemble mean surrogate severe forecasts is indicated by a solid white square and annotated. Additionally, the σ and percentile where each ensemble member's surrogate severe forecast achieves its maximum score is denoted as a white circle, with color fill inside the circle corresponding to the member's score.

augmented advantage over its members at $\sigma \leq 80$ km. It is clear from Fig. 13 that when aggressive smoothing ($\sigma \geq 120$ km) is used to produce ensemble surrogate severe forecasts, the resulting field is only marginally more skillful than applying the same smoother to a typical ensemble member's binary field, especially in the case of HRRRE and MAP. A corollary is that the added computational expense of running a full CAM ensemble for the purpose of producing skillful surrogate severe forecasts generally yields diminishing returns as the choice of σ increases, since an optimized deterministic CAM could provide a comparable product.

To assess the smoothing scale dependence of surrogate severe forecast reliability, Fig. 14 presents attributes diagrams for surrogate severe forecasts produced using three different σ values. At each σ and for each ensemble, the UH percentile that minimizes BS_{REL} is selected to plot. At $\sigma = 60$ km (Fig. 14a), HREFv2.1 demonstrates better reliability than HRRRE and particularly MAP, with the latter two showing more overconfidence. Increasing σ to 120 km reduces the disparity between HREFv2.1 and HRRRE, although MAP remains notably more overconfident (Fig. 14b). Finally, at $\sigma =$ 180 km (Fig. 14c), meaningful differences in reliability between the three ensembles have been greatly minimized (indeed, HREFv2.1 has the worst reliability, although none of the ensembles display obvious overconfidence). Given these results, larger σ may tend to mask underlying skill differences between ensembles at the true neighborhood length scale.

Based on these analyses, we suggest verification and postprocessing of CAM ensemble rare-event NMEPs (including, but not limited to, surrogate severe forecasts) that employ a Gaussian smoother should, when possible, be tested across a range of σ values that extend well below the traditional surrogate severe forecast default of $\sigma = 120$ km. For example, our surrogate severe forecast results suggest near-maximum skill can now be extracted from a diverse CAM ensemble such as HREFv2.1 using a smaller smoothing length scale ($\sigma \sim 60$ km). Furthermore, during verification, ensemble NMEP skill at a given σ value should ideally be contextualized through comparison with individual member probabilities produced using the same Gaussian parameter, as in Fig. 13. This assesses

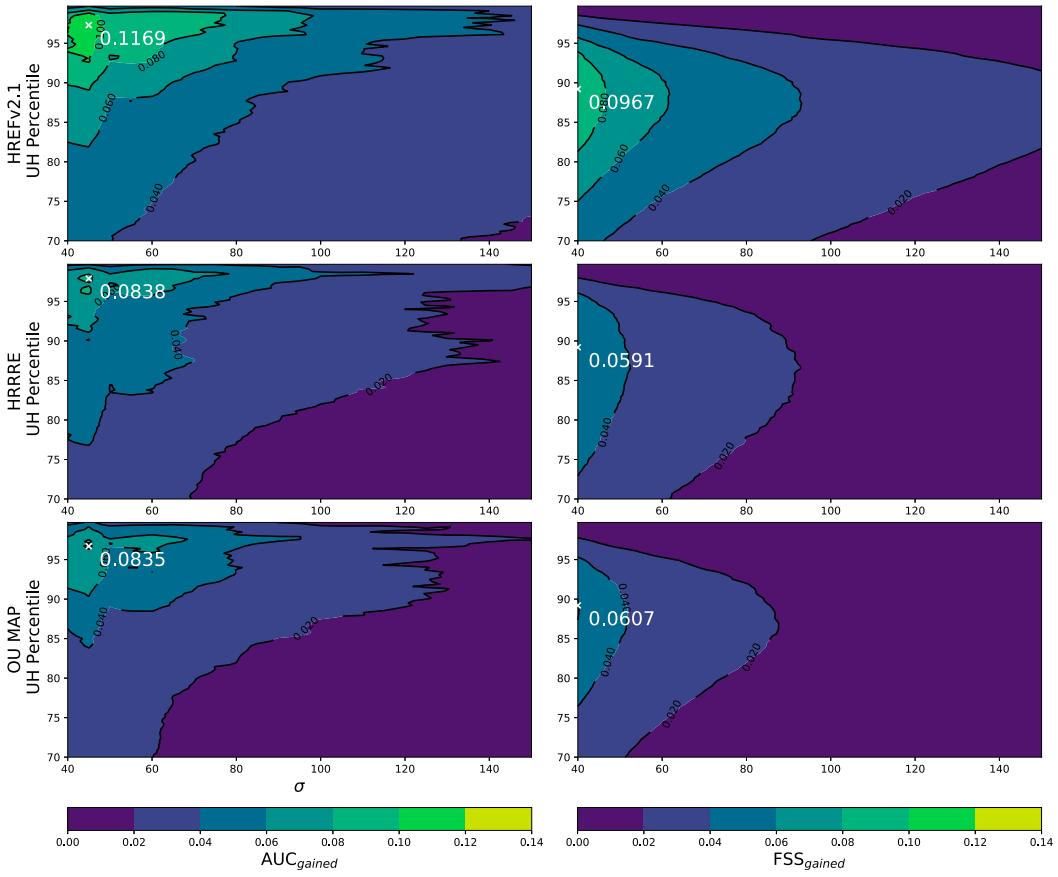

FIG. 13. As in Fig. 12, but the plotted quantity is the difference between the score of the ensemble mean surrogate severe forecasts and the mean score of the member surrogate severe forecasts (i.e., AUC_{gained} and FSS_{gained}), and dots for ensemble and member score maxima are omitted. Note that the range of the abscissa differs from Fig. 12 to highlight the portion of the parameter space with nonnegligible AUC_{gained} and FSS_{gained}.

whether smoothing a single member's binary field in the same way as the ensemble mean could provide nearly equivalent skill, in which case the ensemble is not adding substantial value.

5. Summary and conclusions

In this study, we compared the ability of three CAM ensemble systems to produce skillful probabilistic forecasts of convective storms and severe weather hazards within the context of the next-day forecast problem. The first ensemble, HREFv2.1, is an ensemble of opportunity (EO) comprising highly diverse deterministic CAMs processed together as an ad hoc ensemble. The other two ensembles, HRRRE and MAP, are formally designed ensemble prediction systems (EPSs) with unified model configurations across their members. Owing to their membership designs, HREFv2.1 samples both model and IC uncertainty, whereas HRRRE and MAP only sample IC uncertainty.

Verification of bias-corrected composite reflectivity (CREF) exceeding 40 dBZ within an 80 km × 80 km neighborhood was performed on hourly snapshots over the 21-day dataset in the spring of 2018 for lead times of 13–30 h. Intended to evaluate ensemble skill in the overall placement and coverage of convective storms, this analysis revealed that HREFv2.1 produced the most skillful forecasts, followed by MAP, and then HRRRE. When forecasts from individual ensemble members were verified, member skill between HREFv2.1 and MAP was generally quite similar, with MAP members actually outperforming HREFv2.1 members in some metrics. However, HREFv2.1 NMEPs showed a substantially larger improvement over its constituent member forecasts than did MAP NMEPs. This suggests HREFv2.1 members are more effectively filling out a realistic PDF, whereas MAP members are more duplicative of one another (a weakness shared, to a somewhat lesser extent, by HRRRE). This finding motivated further quantitative evaluation of ensemble spread in the three systems. Correlations between member NMEPs, along with the grid-point frequency of member "outlier points," were computed, and both indicated substantially more ensemble spread exists in HREFv2.1 than HRRRE and MAP. Attributes diagrams also showed much better reliability for HREFv2.1, whereas HRRRE and MAP exhibited overconfident probabilities. Spread–skill metrics computed for NMEPs indicated very good statistical consistency for HREFv2.1; by contrast, HRRRE and

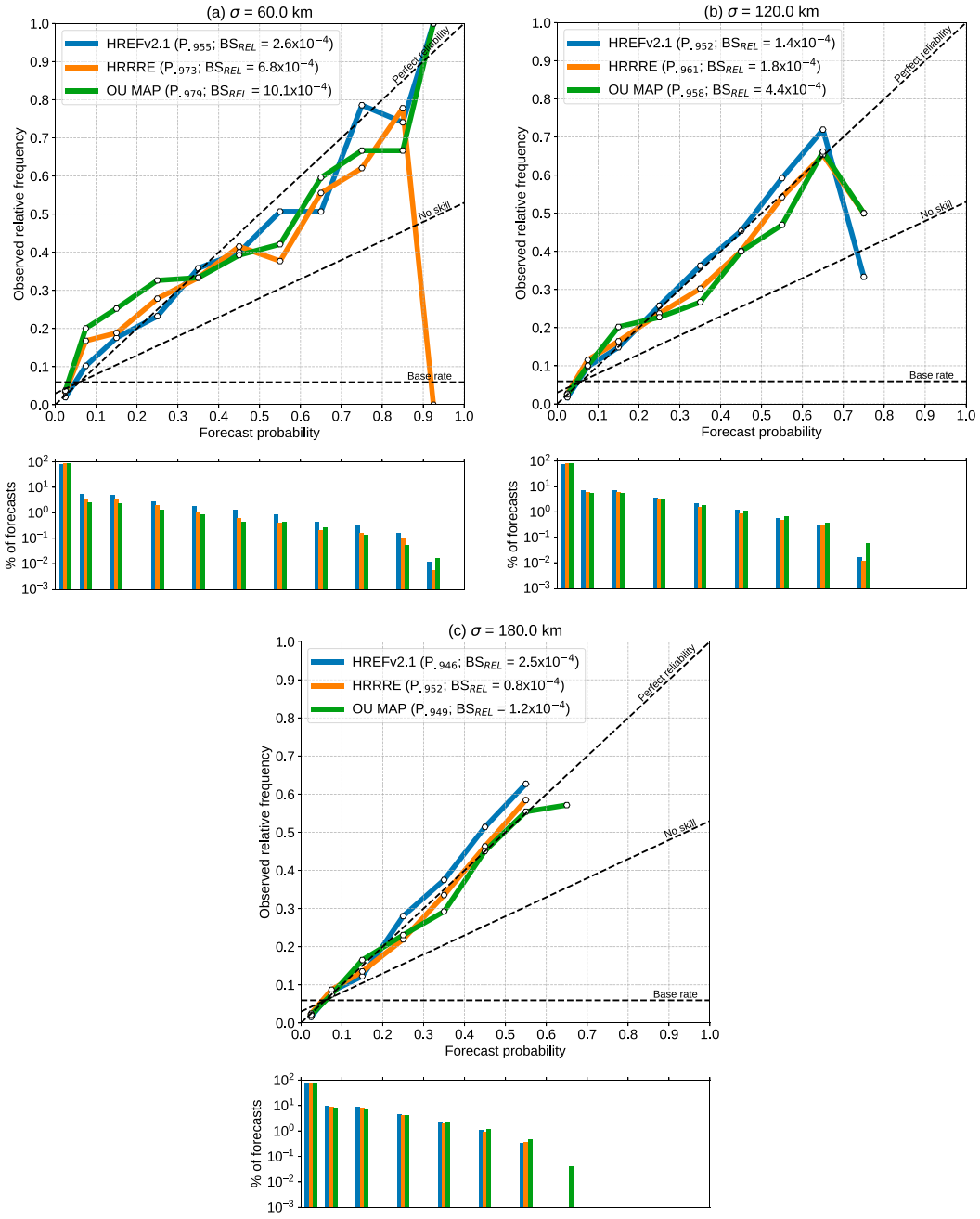

FIG. 14. Attributes diagram for surrogate severe forecasts produced with a Gaussian filter using (a) $\sigma = 60$ km, (b) $\sigma = 120$ km, and (c) $\sigma = 180$ km. At each σ value and for each ensemble, the UH percentile, which minimizes the reliability component of the Brier score (BS_{REL}) is used to produce the surrogate severe forecasts; this percentile and the resulting BS_{REL} are reported in the legend for each panel.

MAP had approximately half of the spread needed for good consistency. This inferior consistency was overwhelmingly due to smaller spread in HRRRE and MAP, with their slightly larger NMEP errors playing only a minor role.

To complement CREF verification, we also verified daily surrogate severe forecasts with the goal of assessing ensemble skill in predicting intense storms associated with severe convective hazards (tornadoes, hail, and wind gusts). By aggregating the surrogate severe forecasts over a 24-h period, this verification was largely insensitive to timing errors. Additionally, surrogate severe forecasts were produced and verified on a relatively coarse 80-km grid, decreasing the computational cost and allowing us to test a wide range of UH thresholds and Gaussian σ values. At their respective performance maxima

within the UH–σ parameter space, ensemble skill differences between the three systems largely mirrored our CREF results: HREFv2.1 performed best, followed by MAP, and then HRRRE. Additionally, HREFv2.1 maximized AUC and FSS with a smaller σ than HRRRE or MAP, implying that its advantage over the two other systems is especially pronounced when less smoothing is applied to produce the surrogate severe forecasts. We computed skill differences between the ensemble mean and member surrogate severe forecasts throughout the parameter space and found the ensembles (particularly HREFv2.1) offered their greatest added value at small σ, whereas strong smoothing (e.g., $\sigma > 120$ km) washed out much of the meaningful skill difference between ensemble and deterministic surrogate severe forecasts.

These results have potential implications for postprocessing and verifying neighborhood-based CAM ensemble products. First, diverse ensembles such as HREFv2.1 with relatively good spread characteristics may be capable of forecasting convective storms with near-optimal skill at smaller spatial scales than is assumed a priori when a Gaussian smoother is applied using the traditional $\sigma = 120$ km. To the extent this is true, applying weaker smoothing to real-time NMEPs benefits operational users by retaining more spatial detail from the skillful model solutions. Additionally, when verifying CAM ensemble NMEPs (including surrogate severe forecasts) over a range of σ values, the value that maximizes a skill score does not necessarily highlight where the ensemble is adding the most value over deterministic CAMs. As more aggressive smoothing is applied to gridpoint NMEPs, they increasingly verify similarly to the equivalently smoothed versions of their underlying member binary fields. This is another reason that an ensemble that achieves maximal skill using less smoothing (HREFv2.1, in the present study) is preferable to one that attains comparable skill only after applying more smoothing (particularly true of MAP in the present study).

Given that HRRRE and MAP do not include sampling of model uncertainty in any fashion, the superior spread and skill of HREFv2.1 suggests the critical need to sample model errors optimally in CAM ensemble design. Our results corroborate the advantage of sampling model uncertainty previously shown in the context of controlled CAM ensemble experiments (Romine et al. 2014; Gasperoni et al. 2020); in the present study, this was shown for ensembles implemented successfully for real-time applications, and also for verification focused on convective storms. Note that although HREFv2.1 is the only ensemble herein to sample model uncertainty meaningfully, it also typically contains more diverse ICs than HRRRE or MAP, so further work is needed to isolate and quantify the specific contribution of HREFv2.1's model uncertainty sampling.[8]

[8] Analysis of the member pair CREF NMEP correlations in Fig. 5a reveals that a shared dynamical core or PBL scheme between two members is more strongly associated with increased r^2 than shared parent model backgrounds or initialization times. This suggests model uncertainty may contribute more spread than IC uncertainty within HREFv2.1 at the lead times we verified, but further analyses or experiments are needed to confirm this rigorously.

Nonetheless, the superior spread of HREFv2.1 reported in this study illustrates the compelling benefits of processing CAM EOs with multidimensional member configuration diversity. As we focused our verification on lead times of 12–36 h, often described in convective forecasting as the ''next-day problem,'' we cannot yet address whether accounting for complex model uncertainty is similarly crucial at shorter lead times. Also, vigorous development of CAM ensembles has only accelerated in earnest over the most recent decade, and operational implementations remain very limited. As additional research is performed to bring stochastic physics schemes to maturity, model uncertainty is likely to become more adequately represented in future unified CAM ensembles, potentially adding spread and improving their skill in forecasting convective storms. Nonetheless, our results highlight the impressive potential of dynamical core, physics, IC analysis, and time-lagging diversity working in tandem to represent the highly nonlinear forecast uncertainties that modulate convective initiation and evolution, suggesting they should be given due consideration in future CAM ensemble design and implementation decisions.

Acknowledgments. The first and second authors were provided support by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA16OAR4320115, U.S. Department of Commerce. Authors ILJ and AJC completed this work as part of regular duties at the federally funded NOAA Storm Prediction Center and National Severe Storms Laboratory, respectively. Authors XW and YW were supported by NA16OAR4590236. We 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 also thank all SFE2018 participants for their real-time subjective ratings and feedback on CAM ensemble performance, which helped to focus this work. 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. We thank three anonymous reviewers and Dr. Gary Lackmann for their thorough critiques of the original manuscript, which yielded numerous improvements in both content and presentation.

Data availability statement. HREFv2.1 model data used in this study were provided to the Storm Prediction Center (SPC) by NCEP's Environmental Modeling Center (EMC), and are archived internally at the National Severe Storms Laboratory (NSSL). HRRRE and OU-MAP model data were provided by NOAA's Earth System Research Laboratory (ESRL) and the OU School of Meteorology, respectively. These datasets were transferred to NSSL as part of the Community Leveraged Unified Ensemble (CLUE) during the 2018 HWT Spring Forecasting Experiment, and are archived internally. The Multi-Radar Multi-Sensor (MRMS) data used for composite reflectivity verification were obtained in real time from the NCEP FTP service (https://mrms.ncep.noaa.gov/data); an archive that includes the period used in this study is maintained

TABLE A1. Unbiased CREF thresholds T_{unbiased} (dBZ) corresponding to the MRMS MRQCC 40-dBZ threshold for each model configuration. Thresholds were computed for the CONUS domain based on 378 hourly snapshots of CREF.

Configuration	T_{unbiased} (dBZ) for 40 dBZ
HRRRv3	44.8
HRW ARW	43.0
HRW NSSL	43.4
HRW NMMB	47.7
NAM Nest	42.4
HRRRE	44.2
MAP	44.5

internally at NSSL. Local storm reports (LSRs) used for surrogate severe verification were obtained from SPC's public logs (https://www.spc.noaa.gov/climo/online). Datasets stored internally at NSSL may be shared upon request (pending the consent of the original dataset creators, in the case of MRMS and OU-MAP).

APPENDIX

CREF Bias Correction

Because we evaluate neighborhood probability forecasts for CREF ≥ 40 dBZ in this study, we are concerned with the frequency bias for each ensemble member in exceeding that threshold. As mentioned in section 2, we choose an approach conceptually similar to "quantile mapping" (Hopson and Webster 2010; Voisin et al. 2010). The dataset used for bias correction is the same as the verification dataset. Our procedure for computing bias-corrected CREF ≥ 40 dBZ probabilities is as follows:

1) Compute the gridpoint frequency of CREF ≥ T for T = [35, 36, 37, … , 50] dBZ for each ensemble member over the CONUS for all 378 hourly snapshots.
2) Compute the gridpoint frequency of CREF ≥ 40 dBZ for the MRMS MRQCC over the CONUS for all 378 hourly snapshots.
3) For each ensemble member, compute the bias $B_{T,O}$ for T = [35, 36, … , 50] dBZ and O = 40 dBZ, where T is the forecast threshold and O is the observed threshold.
4) For each ensemble member, choose an unbiased threshold T_{unbiased} by linearly interpolating between the computed $B_{T,O}$ values (available at 1-dBZ increments) to estimate the T value at which $B_{T,40} = 1$.

As an example, if $B_{43,40} = 1.05$ and $B_{44,40} = 0.95$ are computed for a member, we estimate $T_{\text{unbiased}} = 43.5$ dBZ. Bias-corrected NMEPs for ensembles are then computed with respect to a member-dependent exceedance threshold of T_{unbiased}, rather than a fixed 40-dBZ threshold. Note that our procedure guarantees forecast exceedance probabilities will be approximately unbiased with respect to MRQCC for any ensemble *at the grid scale*, but does not strictly guarantee neighborhood probability fields will be unbiased: it is possible for N grid points exceeding T_{unbiased} in a forecast to be systematically more or less spatially clustered than N grid points exceeding 40 dBZ in the MRMS verification dataset, which in turn would yield biased coverage of NMEP > 0 when the neighborhood size is much larger than one grid point.

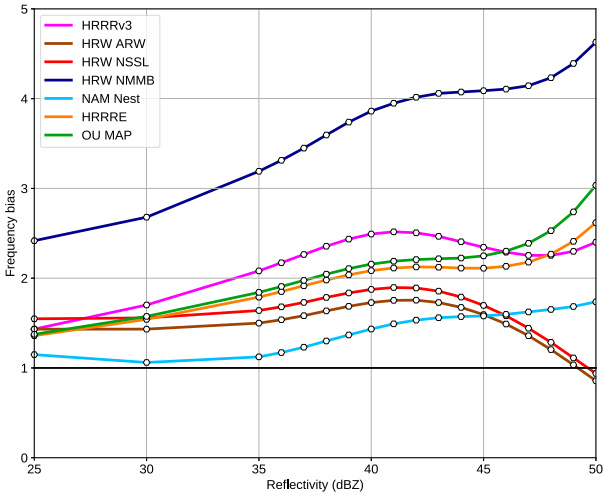

FIG. A1. CREF gridpoint exceedance frequency biases for each model configuration in the verification dataset. Bias is computed over the entire 21-day dataset (378 hourly snapshots), and over the CONUS domain, at 1-dBZ intervals for 35 ≤ CREF ≤ 50 dBZ, as well as at 25 and 30 dBZ.

After computing the biases for each member of all three ensembles, members are assigned into groups sharing identical model configurations: MAP (N = 10), HRRRE (N = 9), HRRR (N = 2), HRW ARW (N = 2), HRW NMMB (N = 2), HRW NSSL (N = 2), and NAM Nest (N = 2). For each of these groups, the mean T_{unbiased} of its members is used to compute NMEPs. These values are displayed in Table A1. Our theoretical goal in performing this correction is to ignore discrepancies in how storms are *depicted* in different configurations' CREF fields; in particular, discrepancies owing strictly to idiosyncrasies of the microphysics scheme, numerical diffusion, etc. Put another way, if a particular storm with a particular structure and intensity is "correctly" predicted in all of the model configurations, we hope to treat the simulated manifestation of that storm the same in each configuration during verification. In practice, however, it is possible that the configurations exhibit different biases in the actual coverage of convective storms, which could lead us to assign more aggressive configurations a T_{unbiased} corresponding to more intense storms than less aggressive configurations.

In the course of performing the bias correction, gridpoint frequency biases were computed for numerous CREF thresholds for each model configuration over the CONUS. These biases are presented in Fig. A1. Regarding the aforementioned possibility of discrepancies in real storm coverage unduly influencing our thresholds: it is encouraging that configurations sharing a common microphysics parameterization scheme generally exhibit similar bias curves (e.g., the MAP, HRRRE, and HRRR, all using Thompson microphysics; or the HRW ARW and HRW NSSL, both using WSM6 microphysics). At the 40-dBZ exceedance threshold, frequency biases range from 1.5 to 3.9 across the configurations.

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.

Arribas, A., K. B. Robertson, and K. R. Mylne, 2005: Test of a poor man's ensemble prediction system for short-range probability forecasting. *Mon. Wea. Rev.*, **133**, 1825–1839, https://doi.org/10.1175/MWR2911.1.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Casanova, S., and B. Ahrens, 2009: On the weighting of multimodel ensembles in seasonal and short-range weather forecasting. *Mon. Wea. Rev.*, **137**, 3811–3822, https://doi.org/10.1175/2009MWR2893.1.

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.

——, W. A. Gallus, and T.-C. Chen, 2008: Contributions of mixed physics versus perturbed initial/lateral boundary conditions to ensemble-based precipitation forecast skill. *Mon. Wea. Rev.*, **136**, 2140–2156, https://doi.org/10.1175/2007MWR2029.1.

——, 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.

Dowell, D. C., 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.

Du, J., S. L. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.*, **125**, 2427–2459, https://doi.org/10.1175/1520-0493(1997)125<2427:SREFOQ>2.0.CO;2.

Duda, J. D., and W. A. Gallus, 2010: Spring and summer Midwestern severe weather reports in supercells compared to other morphologies. *Wea. Forecasting*, **25**, 190–206, https://doi.org/10.1175/2009WAF2222338.1.

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

——, ——, and 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.

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

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

Ferrier, B. S., W. Wang, and E. Colon, 2011: Evaluating cloud microphysics schemes in nested NMMB forecasts. *24th Conf. on Weather and Forecasting/20th Conf. on Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc., 14B.1, https://ams.confex.com/ams/91Annual/webprogram/Paper179488.html.

Fowler, T., J. Halley Gotway, K. Newman, T. L. Jensen, B. G. Brown, and R. G. Bullock, 2017: The Model Evaluation Tools v7.0 (METv7.0) user's guide. Tech. Rep., Developmental Testbed Center, 408 pp., https://dtcenter.org/sites/default/files/community-code/met/docs/user-guide/MET_Users_Guide_v7.0.pdf.

Gallo, B. T., A. J. Clark, and S. R. Dembek, 2016: Forecasting tornadoes using convection-permitting ensembles. *Wea. Forecasting*, **31**, 273–295, https://doi.org/10.1175/WAF-D-15-0134.1.

——, 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.

——, B. Roberts, I. L. Jirak, A. J. Clark, C. P. Kalb, and T. L. Jensen, 2018: Evaluating potential future configurations of the High Resolution Ensemble Forecast system. *29th Conf. on Severe Local Storms*, Stowe, VT, Amer. Meteor. Soc., 76, https://ams.confex.com/ams/29SLS/webprogram/Paper348791.html.

——, and Coauthors, 2019: Initial development and testing of a convection-allowing model scorecard. *Bull. Amer. Meteor. Soc.*, **100**, ES367–ES384, https://doi.org/10.1175/BAMS-D-18-0218.1.

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.

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.

Hacker, J. P., and Coauthors, 2011: The U.S. Air Force Weather Agency's mesoscale ensemble: Scientific description and performance results. *Tellus*, **63A**, 625–641, https://doi.org/10.1111/j.1600-0870.2010.00497.x.

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.

Hong, S.-Y., and J.-O. J. Lim, 2006: The WRF single-moment 6-class microphysics scheme (WSM6). *J. Korean Meteor. Soc.*, **42**, 129–151.

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

Hopson, T. M., and P. J. Webster, 2010: A 1–10-day ensemble forecasting scheme for the major river basins of Bangladesh: Forecasting severe floods of 2003–07. *J. Hydrometeor.*, **11**, 618–641, https://doi.org/10.1175/2009JHM1006.1.

Hsu, W., and A. H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting*, **2**, 285–293, https://doi.org/10.1016/0169-2070(86)90048-8.

Janjić, Z., 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.

——, and R. L. 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.

Jankov, I., W. A. Gallus, M. Segal, B. Shaw, and S. E. Koch, 2005: The impact of different WRF model physical parameterizations and their interactions on warm season MCS rainfall. *Wea. Forecasting*, **20**, 1048–1060, https://doi.org/10.1175/WAF888.1.

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

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

——, C. J. Melick, and S. J. Weiss, 2015: Comparison of convection-allowing ensembles during the 2015 NOAA Hazardous Weather Testbed Spring Forecasting Experiment. *40th Natl. Wea. Assoc., Annual Meeting*, Oklahoma City, OK, AP-36, https://www.spc.noaa.gov/publications/jirak/camcomp.pdf.

——, ——, and ——, 2016: Comparison of the SPC storm-scale ensemble of opportunity to other convection-allowing ensembles for severe weather forecasting. *28th Conf. on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., 102, https://ams.confex.com/ams/28SLS/webprogram/Paper300910.html.

Johnson, A., X. Wang, F. Kong, and M. Xue, 2011: Hierarchical cluster analysis of a convection-allowing ensemble during the Hazardous Weather Testbed 2009 Spring Experiment. Part I: Development of the object-oriented cluster analysis method for precipitation fields. *Mon. Wea. Rev.*, **139**, 3673–3693, https://doi.org/10.1175/MWR-D-11-00015.1.

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

——, ——, Y. Wang, A. Reinhart, A. J. Clark, and I. L. Jirak, 2020: Neighborhood- and object-based probabilistic verification of the OU MAP ensemble forecasts during 2017 and 2018 Hazardous Weather Testbeds. *Wea. Forecasting*, **35**, 169–191, https://doi.org/10.1175/WAF-D-19-0060.1.

Johnson, C., and R. Swinbank, 2009: Medium-range multimodel ensemble combination and calibration. *Quart. J. Roy. Meteor. Soc.*, **135**, 777–794, https://doi.org/10.1002/qj.383.

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.

Kong, F., and Coauthors, 2007: Preliminary analysis on the real-time Storm-Scale Ensemble Forecasts produced as a part of the NOAA Hazardous Weather Testbed 2007 Spring Experiment. *22nd Conf. on Weather Analysis and Forecasting/18th Conf. on Numerical Weather Prediction*, Park City, UT, Amer. Meteor. Soc., 3B.2, https://ams.confex.com/ams/pdfpapers/124667.pdf.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418, https://doi.org/10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2.

Levit, J. J., and Coauthors, 2008: The NOAA Hazardous Weather Testbed 2008 Spring Experiment: Technical and scientific challenges of creating a data visualization environment for storm-scale deterministic and ensemble forecasts. *24th Conf. on Severe Local Storms*, Savannah, GA, Amer. Meteor. Soc., P10.5, https://ams.confex.com/ams/pdfpapers/141785.pdf.

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

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

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

Mittermaier, M. P., 2007: Improving short-range high-resolution model precipitation forecast skill using time-lagged ensembles. *Quart. J. Roy. Meteor. Soc.*, **133**, 1487–1500, https://doi.org/10.1002/qj.135.

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.

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

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.

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

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

——, G. S. Romine, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2015: NCAR's experimental real-time convection-allowing ensemble prediction system. *Wea. Forecasting*, **30**, 1645–1654, https://doi.org/10.1175/WAF-D-15-0103.1.

Shao, H., and Coauthors, 2016: Bridging research to operations transitions: Status and plans of community GSI. *Bull. Amer.*

Meteor. Soc., **97**, 1427–1440, https://doi.org/10.1175/BAMS-D-13-00245.1.

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 .

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., and J. S. Kain, 2017: Seasonal variations in severe weather forecast skill in an experimental convection-allowing model. *Wea. Forecasting*, **32**, 1885–1902, https://doi.org/10.1175/WAF-D-17-0043.1.

——, ——, 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, 2016a: 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.

——, C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016b: 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.

——, ——, ——, and M. L. Weisman, 2019: Next-day prediction of tornadoes using convection-allowing models with 1-km horizontal grid spacing. *Wea. Forecasting*, **34**, 1117–1135, https://doi.org/10.1175/WAF-D-19-0044.1.

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

Thompson, G., and T. Eidhammer, 2014: A study of aerosol impacts on clouds and precipitation development in a large winter cyclone. *J. Atmos. Sci.*, **71**, 3636–3658, https://doi.org/10.1175/JAS-D-13-0305.1.

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

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

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

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

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

——, ——, J. R. Carley, and D. C. Dowell, 2018: Development and research of GSI based EnVar with direct radar data assimilation for the US NWS operational regional convection allowing modeling systems to improve convection-allowing hazardous weather forecast and results during the HWT Spring Experiments. *29th Conf. on Weather Analysis and Forecasting*, Denver, CO, Amer. Meteor. Soc., 7A.4, https://ams.confex.com/ams/29WAF25NWP/webprogram/Paper345833.html.

Wastl, C., Y. Wang, A. Atencia, and C. Wittmann, 2019: A hybrid stochastically perturbed parametrization scheme in a convection-permitting ensemble. *Mon. Wea. Rev.*, **147**, 2217–2230, https://doi.org/10.1175/MWR-D-18-0415.1.

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

Wu, W.-S., R. J. Purser, and D. F. Parrish, 2002: Three-dimensional variational analysis with spatially inhomogeneous covariances. *Mon. Wea. Rev.*, **130**, 2905–2916, https://doi.org/10.1175/1520-0493(2002)130<2905:TDVAWS>2.0.CO;2.

Xue, M., and Coauthors, 2007: CAPS realtime storm-scale ensemble and high-resolution forecasts as part of the NOAA Hazardous Weather Testbed 2007 spring experiment. *22nd Conf. on Weather Analysis and Forecasting/18th Conf. on Numerical Weather Prediction*. Park City, UT, Amer. Meteor. Soc., 3B.1, https://ams.confex.com/ams/pdfpapers/124587.pdf.

Ziehmann, C., 2000: Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models. *Tellus*, **52A**, 280–299, https://doi.org/10.3402/tellusa.v52i3.12266.