Exploring Convection-Allowing Model Evaluation Strategies for Severe Local Storms Using the Finite-Volume Cubed-Sphere (FV3) Model Core

BURKELY T. GALLO,[a,b] JAMIE K. WOLFF,[c,d] ADAM J. CLARK,[e] ISRAEL JIRAK,[b] LINDSAY R. BLANK,[c,d] BRETT ROBERTS,[a,b,e] YUNHENG WANG,[a,e] CHUNXI ZHANG,[f,g,h] MING XUE,[f] TIM SUPINIE,[f] LUCAS HARRIS,[i] LINJIONG ZHOU,[i] AND CURTIS ALEXANDER[j]

[a] *Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma*
[b] *NOAA/NWS/NCEP Storm Prediction Center, Norman, Oklahoma*
[c] *Research Applications Laboratory, National Center for Atmospheric Research, Boulder, Colorado*
[d] *Developmental Testbed Center, Boulder, Colorado*
[e] *NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma*
[f] *Center for Analysis and Prediction of Storms, Norman, Oklahoma*
[g] *NOAA/NWS/NCEP Environmental Modeling Center, College Park, Maryland*
[h] *I.M. Systems Group, College Park, Maryland*
[i] *NOAA/OAR Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey*
[j] *NOAA/Earth System Research Laboratory/Global Systems Division, Boulder, Colorado*

(Manuscript received 9 June 2020, in final form 23 September 2020)

ABSTRACT: Verification methods for convection-allowing models (CAMs) should consider the finescale spatial and temporal detail provided by CAMs, and including both neighborhood and object-based methods can account for displaced features that may still provide useful information. This work explores both contingency table–based verification techniques and object-based verification techniques as they relate to forecasts of severe convection. Two key fields in severe weather forecasting are investigated: updraft helicity (UH) and simulated composite reflectivity. UH is used to generate severe weather probabilities called surrogate severe fields, which have two tunable parameters: the UH threshold and the smoothing level. Probabilities computed using the UH threshold and smoothing level that give the best area under the receiver operating curve result in very high probabilities, while optimizing the parameters based on the Brier score reliability component results in much lower probabilities. Subjective ratings from participants in the 2018 NOAA Hazardous Weather Testbed Spring Forecasting Experiment (SFE) provide a complementary evaluation source. This work compares the verification methodologies in the context of three CAMs using the Finite-Volume Cubed-Sphere Dynamical Core (FV3), which will be the foundation of the U.S. Unified Forecast System (UFS). Three agencies ran FV3-based CAMs during the five-week 2018 SFE. These FV3-based CAMs are verified alongside a current operational CAM, the High-Resolution Rapid Refresh version 3 (HRRRv3). The HRRR is planned to eventually use the FV3 dynamical core as part of the UFS; as such evaluations relative to current HRRR configurations are imperative to maintaining high forecast quality and informing future implementation decisions.

SIGNIFICANCE STATEMENT: The United States is currently working toward unifying its numerical modeling efforts around a single dynamical core, or set of equations that serves as the model framework. We compared three models built around this new dynamical core to the current operational model, focusing on forecasts of severe convection. We also explored different verification techniques, to look at model performance from many angles. A major point discussed in this work is that subjective choices (i.e., techniques, thresholds, fields, etc. used) still play a role in objective verification. While we found that the experimental models are not yet depicting severe weather as well as the operational model according to traditional verification techniques and metrics, there may be improvements captured by newer verification techniques.

KEYWORDS: Mesoscale forecasting; Numerical weather prediction/forecasting; Operational forecasting; Model comparison; Model evaluation/performance

1. Introduction

Convection-allowing models (CAMs) are becoming more widely available and play an increasingly important role in the forecast process, particularly since the operationalization of the High-Resolution Rapid Refresh (HRRR; Benjamin et al. 2016; Alexander et al. 2017) model in 2014 and the High Resolution Ensemble Forecast system in November of 2017 (HREF; Roberts et al. 2019). CAMs can be particularly helpful in forecasting severe convective weather, since the small grid spacing (~3 km) allows for simulation of storm-scale structures such as supercells (Kain et al. 2006). As convective mode plays a key role in determining the convective hazard type (e.g., a linear convective mode is more likely to produce severe winds than severe hail; Smith et al. 2012), forecasters can use CAMs to forecast threat types prior to convective initiation. Forecasters determine expected convective mode by looking at the reflectivity structure of a storm, as well as whether or not the storm is rotating. CAMs can simulate both of these characteristics, with

Corresponding author: Burkely T. Gallo, burkely.twiest@noaa.gov

DOI: 10.1175/WAF-D-20-0090.1

the latter typically represented as updraft helicity (UH). UH is an integral of the updraft speed and the vertical vorticity taken over a specified vertical layer. Typically, CAMs output the hourly maximum UH over the 2–5-km layer such that forecasters can determine whether a simulated midlevel mesocyclone is occurring; if the hourly maximum field shows a relatively continuous swath of high UH values, the model is likely producing a simulated supercell (Kain et al. 2010).

Verifying CAMs provides a unique set of challenges, and this work attempts to provide a multifaceted verification approach to four different deterministic CAMs, three of which are experimental configurations. The purpose of this work is not only to compare and verify the performance of these specific CAMs during the spring convective season, but also to demonstrate different verification techniques and what can be gleaned from them. We hypothesize that each verification metric will provide unique insight to the model performance, and that if one model performs better than the others do across all metrics tested herein, it is more likely to be perceived as useful by forecasters.

CAMs often use postprocessing methods such as neighborhood-based techniques, since gridpoint-based statistics that use the contingency table may not reflect improvements evident in object-based methods or subjective evaluation due to small-scale displacements (see Schwartz and Sobash 2017 for a review of neighborhood approaches as applied to CAMs). These displacements result in double penalties, where a forecast displaced from observations is penalized for both a false alarm area and a missed area (Mass et al. 2002; Done et al. 2004). While a forecast from a CAM may not have a storm in the exact right location, the convective mode information, timing, and general location of convection still provide useful guidance to forecasters, adding "value" as defined by Murphy (1993). Neighborhood-based techniques look for a forecast feature of interest within a spatial and/or temporal neighborhood, and if the forecast event occurs within this neighborhood of the grid point, it is considered a "hit" on the standard 2×2 contingency table. Neighborhoods are also used in operational applications, as in the Storm Prediction Center (SPC) definition of their forecast probabilities being the probability of severe weather occurring within 25 mi of a given point.[1]

Besides neighborhood techniques, another form of CAM verification specifically focuses on forecast objects produced by the CAM. Object-based methods focus on characteristics such as size, quantity, and orientation of the forecast and observed objects (Davis et al. 2009; Wolff et al. 2014; Skinner et al. 2018). The Method for Object-Based Diagnostic Evaluation (MODE; Davis et al. 2006, 2009) is an object-based method of evaluation that attempts to mimic the process of subjectively evaluating a forecast field by assessing aspects such as object area and object size. MODE has previously been applied to CAM forecasts of precipitation accumulations (Davis et al. 2009; Gallus 2010; Wolff et al. 2014; Clark et al. 2014), cloud objects (Griffin et al. 2017), and a simulated vertically integrated liquid field to evaluate forecasts of storm characteristics such as convective mode, storm size, and number of storms (Cai and Dumais 2015). Results from MODE can complement results from traditional contingency table–based statistics and neighborhood methods, providing insight into fields that operational forecasters often consider.

In addition to leveraging verification metrics that account for the unique characteristics of CAMs, CAM output can be upscaled and smoothed via postprocessing techniques. One of these techniques generates "surrogate severe" fields by using a Gaussian kernel and UH exceedance thresholds to create probabilistic forecasts of severe weather (Sobash et al. 2011). The probabilistic surrogate severe field can then be compared to either binary yes/no reports or by applying a similar approach to reports, using the "practically perfect" technique (Hitchens et al. 2013). The practically perfect technique uses a Gaussian kernel to create smoothed probabilistic fields from local storm reports, showing what forecast a forecaster would draw with perfect prior knowledge of where the reports would occur. Taken together, the surrogate severe technique and practically perfect technique create comparable probabilistic forecast fields of a binary event: whether or not severe weather will occur in an area (Sobash et al. 2011). Surrogate severe forecasts can then be evaluated using probabilistic metrics such as reliability alongside contingency table analyses such as the area under the receiver operating curve (ROC area; Mason 1982). These postprocessed surrogate severe fields also resemble operational forecasts, allowing the forecaster to quickly assess the CAM output in familiar terms.

UH and composite reflectivity are two of the critical severe weather CAM output fields examined in annual Spring Forecasting Experiments (SFEs; Clark et al. 2012; Gallo et al. 2017) that take place at NOAA's Hazardous Weather Testbed (HWT). These experiments gather researchers and forecasters from across the meteorological community to provide feedback on cutting-edge CAM guidance and postprocessing in a real-time environment, where the participants are issuing forecasts using experimental guidance. Participants provide subjective feedback through ratings and comments, and in-depth objective evaluation typically takes place postexperiment (e.g., Gallo et al. 2016; Surcel et al. 2017; Loken et al. 2019). In recent years, experimental guidance evaluated by participants has been organized into the Community Leveraged Unified Ensemble (CLUE; Clark et al. 2018), a framework that coordinates controlled experiments of CAM ensemble configurations. While the original, 2016 CLUE was comprised solely of members using the Advanced Research version of the Weather Research and Forecasting Model (ARW; Skamarock et al. 2008) core, since 2017 members using the Finite-Volume Cubed-Sphere Dynamical Core (FV3; Putman and Lin 2007) were added to the CLUE and evaluated during the SFE. The two FV3-based members run during the 2017 SFE performed much differently than the ARW members (Potvin et al. 2019), and the three FV3-based models available during the 2018 SFE and examined herein sought to improve upon the 2017 performance.

The FV3 dynamical core is a key component of the effort in the United States toward creating a unified forecast system

[1] https://www.spc.noaa.gov/misc/SPC_probotlk_info.html.

TABLE 1. A list of the SFE 2018 cases used in this study.

Week 1	Week 2	Week 3	Week 4	Week 5
30 Apr, 1, 2, 3, 4 May	7, 8, 10 May	15, 16, 17, 18 May	21, 22, 23, 24, 25 May	29, 30, 31 May, 1 Jun

(UFS in NOAA)[2] across all scales. Since this effort encompasses temporal scales ranging from minutes to seasons, and spatial scales ranging from regional to global, many stakeholders will be impacted by this endeavor and may require different types of forecast metrics for their respective applications (Gallo et al. 2019). While some evaluation has been done of forecasts using the FV3 dynamical core at large scales (see the GFS Evaluation web page)[3] work verifying specific attributes of high-resolution, convection-allowing FV3-based models has just begun. Comparisons of the precipitation forecasts between the legacy GFS and a global FV3-based model with a high-resolution nest showed that the FV3-based model was better able to capture the diurnal cycle of precipitation (Zhou et al. 2019). In addition, comparisons focused on hurricane forecasting showed that an FV3-based model performed comparably to or better than current operational hurricane models for track forecasting, and improved intensity forecasts compared to the GFS (Hazelton et al. 2018). Two early studies of precipitation forecasts during the warm season across the contiguous United States (CONUS) show that FV3-based models have skill comparable to WRF (Zhang et al. 2019; Snook et al. 2019), with some sensitivities to microphysics parameterization scheme and little sensitivity to the boundary layer parameterization scheme (Zhang et al. 2019). Harris et al. (2019) examine the performance of a high-resolution nested FV3-based CAM using large-scale metrics such as the 500-hPa anomaly correlation coefficient, but also examine several case studies from the 2017 SFE period and find that FV3-based models can successfully generate realistic convective mode outputs in simulated reflectivity, as well as well-represented hourly maximum UH tracks. This work focuses on FV3-based CAMs as they pertain to forecasting severe convection, based on metrics relevant for CAMs. For comparison, the FV3-based CAMs are evaluated alongside the version of the HRRR that was under development at the time of the 2018 SFE and became operational in July 2018 (i.e., HRRRv3). We hypothesize that the FV3-based CAMs will perform worse than the operational HRRRv3 in most metrics, simply because they are earlier in the development cycle.

Section 2a of this paper will discuss the three FV3-based CAMs examined during this experiment, as well as the HRRRv3 specifications. Section 2b will discuss specifications used in generating surrogate severe fields and selected for reflectivity values. Section 2c will cover the verification data and the specifics of computing the contingency table–based and object-based metrics evaluated. Model climatologies will be discussed in section 3a, contingency table–based statistical results will be in section 3b, and object-based statistical metrics will compose section 3c. Section 3d describes the subjective ratings given by participants during SFE 2018. A case from SFE 2018 illustrates differences in the look of a daily surrogate severe forecast depending on what metric is optimized in section 3. Finally, section 4 will provide conclusions and directions for future work.

2. Data and methodology

a. *Model configurations*

This study examines three CAMs that use the FV3 dynamical core, as well as a CAM that uses the Advanced Research version of WRF (WRF-ARW) dynamical core. All of these models were run during the 2018 SFE, which occurred on weekdays from 30 April 2018 to 1 June 2018 (excluding Memorial Day), resulting in 24 cases. The objective metrics were computed for 21 cases where a complete dataset was present for all of the models examined herein (Table 1).

All model configurations had horizontal grid spacing of ~3 km over a CONUS domain (Table 2). Each model was initialized at 0000 UTC, and had forecasts extending to 36 h. FV3-based configurations were cold start (i.e., no hydrometeors in the initial conditions). Conversely, the HRRRv3 (which became operational on 12 July 2018) uses gridpoint statistical interpolation (GSI; Wu et al. 2002; Kleist et al. 2009) hybrid data assimilation, including the latest 3D radar reflectivity. The HRRRv3 data assimilation includes conventional observations, as well as Tropospheric Airborne Meteorological Data Reporting (TAMDAR; Daniels et al. 2006) aircraft observations, and lightning flash rates (Benjamin et al. 2016). Initial conditions come from the Rapid Refresh (RAP; Alexander et al. 2017), and the lateral boundary conditions come from GFS forecasts (GFSf).

In contrast to the regional HRRRv3, all FV3-based CAMs examined in this work were globally run configurations with a high-resolution nest, though the specifics of the coarse global resolution grid differed somewhat between members. The FV3-based member run by the Geophysical Fluid Dynamics Laboratory, known from here on as the GFDL-FV3, used a combination of grid nesting (Harris and Lin 2013) and stretching (Harris et al. 2016) to transition a 13-km global grid to a 3-km nested grid over the CONUS. The National Severe Storms Laboratory (NSSL) FV3-based member, known as the NSSL-FV3, used a 25-km global grid that was refined to a 3.3-km grid over the CONUS. The third FV3-based member was provided by the Center for the Analysis and Prediction of Storms (CAPS), known as the CAPS-FV3 (Zhang et al. 2019), used an essentially uniform 13-km global grid, within which an ~3.5-km nested grid covered the CONUS. While these grid configurations differ slightly, we do not expect large differences due to horizontal grid

[2] https://ufscommunity.org/.
[3] http://www.emc.ncep.noaa.gov/users/meg/fv3gfs/.

TABLE 2. Model specifications. Microphysics schemes used include Thompson (Thompson et al. 2008) and the 6-category GFDL scheme (GFDL-6cat; Chen and Lin 2013; Zhou et al. 2019); planetary boundary layer (PBL) schemes used include the Mellor–Yamada–Nakanishi–Niino (MYNN; Nakanishi and Niino 2004, 2006), a scale-aware version of MYNN, and the Yonsei University (YSU; Hong et al. 2006). Land surface models include the Rapid Update Cycle (RUC; Smirnova et al. 2016) and the Noah (Chen and Dudhia 2001).

Model	ICs	LBCs	Microphysics	LSM	PBL	Grid spacing
HRRRv3	RAP	GFSf	Thompson	RUC	MYNN	3.0 km
NSSL-FV3	GFS	—	Thompson	Noah	MYNN	3.3 km
GFDL-FV3	GFS	—	GFDL-6cat	Noah	YSU	3.0 km
CAPS-FV3	GFS	—	Thompson	Noah	MYNN-SA	3.25–3.5 km

spacing as they are all between 3 and 3.5 km. Different microphysics and planetary boundary layer parameterization schemes, as well as land surface models, were also used in the different FV3-based models (Table 2), since each agency tested a different strategy for creating the best FV3-based CAM. Regarding the microphysics schemes, which are particularly relevant given this work's focus on convection, the 6-category GFDL scheme (GFDL-6cat; Chen and Lin 2013; Zhou et al. 2019) was used in the GFDL-FV3, the Thompson microphysics scheme in the CAPS-FV3 was a partially two-moment version, and the Thompson microphysics in the HRRRv3 was the two-moment version (Thompson et al. 2008). Finally, the code bases for each model differed slightly, with the CAPS-FV3 and GFDL-FV3 sharing a code base developed originally by GFDL but differing in the physics packages used in the code framework. The NSSL-FV3 code also originated with GFDL, but was further developed by the Environmental Modeling Center for the UFS system and implemented physics packages developed by CAPS.

b. *Surrogate severe and simulated reflectivity specifications*

To examine the differences between severe weather forecasts from the FV3-based CAMs and the WRF-based CAM, simulated reflectivity and UH are verified. Surrogate severe fields are generated following Sobash et al. (2011). To generate surrogate severe fields spanning the convective day (defined as 1200–1200 UTC the following day), the native 2–5-km UH output from each model is first regridded to an 80-km grid (specifically, the NCEP 211 grid).[4] The maximum 2–5-km UH value during the convective day and within each 80-km grid box is assigned to the grid box (i.e., a neighborhood maximum during a 24-h period). Next, the UH value at each grid point is tested to determine whether or not it exceeds a user-defined UH threshold, resulting in a binary grid of ones (did exceed the threshold) and zeroes (did not exceed the threshold). A smoother using a Gaussian kernel density weighting function is then applied to each point to create a smoothed probability field. Two tunable parameters exist in the surrogate severe fields: the UH exceedance threshold and the σ used in the Gaussian kernel density smoother. To determine the best combination of UH threshold and σ for different verification metrics, 5300 combinations of σ and UH threshold are tested, similar to the approach of Clark et al. (2018) when evaluating the 2016 CLUE ensemble subsets. These combinations use 100 different UH thresholds and 53 different σ values. During the 2017 SFE, it was found that the FV3-based CAMs generally produced higher UH values than CAMs using the WRF-ARW dynamical core when using the same horizontal grid spacing (Potvin et al. 2019). Thus, a climatology of UH over the 21 days of SFE 2018 was also computed for the models examined herein (Fig. 1a), using the maximum value of UH at each grid point throughout the day. To account for different model climatologies, UH thresholds used to construct the surrogate severe fields were based on percentile values of the UH from each model rather than from fixed thresholds. This method facilitates a fair comparison between models that may produce vastly different values of UH, particularly in the absence of direct observations of 2–5-km UH. Thus, climatologically high values (e.g., the 90th percentile) from each model are compared. Percentiles used to generate the surrogate severe fields range from the 70th percentile to the 99.7th percentile, in increments of 0.3, and σ values range from 40 to 300 km in increments of 5 km.

Simulated reflectivity was analyzed using object-based verification at three thresholds: 20, 30, and 45 dBZ. While two of these thresholds are relatively low for identifying convective storms, higher thresholds became quite noisy and patterns in the data were difficult to discern. Therefore, we only include one higher threshold, 45 dBZ. A climatology of simulated reflectivity values on each model's native grid was also generated across the cases within the 2018 SFE, using the maximum value across the day at each grid point to examine the magnitude of the strongest convection (Fig. 1b). The 45-dBZ threshold is approximately the 95th percentile for each model, although it is a higher percentile for the HRRRv3 than the FV3-based models, indicating that higher reflectivity values make up a larger part of the distribution in the FV3-based models compared to in the HRRRv3.

c. *Verification data and metrics*

Verification was performed across the eastern 2/3 of the CONUS (Fig. 2). Limiting the analysis to the eastern 2/3 CONUS helps to mitigate the effect of having many zero values of UH or other storm attributes of interest consistently across the western United States, which would decrease the climatological values at each percentile. Given that the western CONUS does not often experience severe convective storms and that radar coverage and population density is relatively

[4] https://www.nco.ncep.noaa.gov/pmb/docs/on388/tableb.html#GRID211.

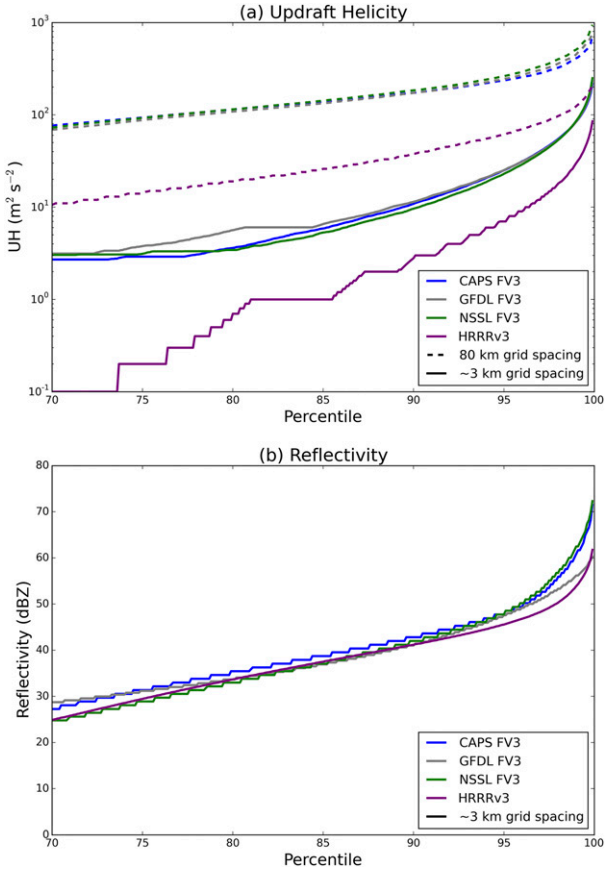

FIG. 1. Climatology of (a) daily maximum UH on the native grid (dashed lines) and 80-km grid (solid lines) and (b) reflectivity on the native grid for each model analyzed during 21 days of SFE 2018.

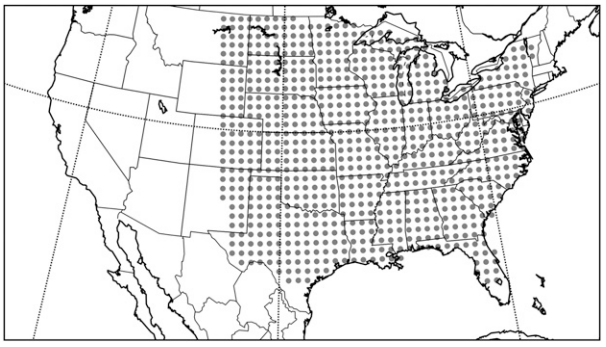

FIG. 2. The verification domain for the current study. Dots indicate grid points included in the grid-based verification statistics.

low, limiting its impact on the climatology helps ensure that the UH percentiles reflect the most likely areas to receive severe weather during the SFE. Multi-Radar Multi-Sensor (MRMS; Smith et al. 2016) composite reflectivity data were used to verify the simulated composite reflectivity at the top of each forecast hour. Surrogate severe fields were verified using local storm reports (LSRs), regridded to the same NCEP 211 80-km grid. All three types of LSRs (hail, wind, and tornadoes) were included in the regridding, as 2–5-km UH has been shown to be a good proxy for all convective hazard types (Sobash et al. 2011).

Contingency table–based statistics were calculated for the surrogate severe fields, including several metrics extracted from the traditional 2×2 contingency table. For the surrogate severe fields, the ROC area was computed using the probability of detection (POD):

$$\text{POD} = \frac{\text{hits}}{\text{hits} + \text{misses}}, \tag{1}$$

and the probability of false detection (POFD):

$$\text{POFD} = \frac{\text{false alarms}}{\text{false alarms} + \text{correct nulls}}, \tag{2}$$

computed at probability thresholds of 2%, 5%, and increments of 5%–95%, with the area being computed using the trapezoidal method (Wandishin et al. 2001). A ROC area of 1 indicates a perfect forecast, and a ROC area at or below 0.5 has no skill. However, the ROC area does not provide information about reliability. Therefore, two forecast metrics that highlight aspects of reliability are also computed. First, the reliability component of the Brier Score (Brier 1950; Murphy 1973) was calculated for each field. This score is negatively oriented, such that a value of 0 is a perfect score and indicates a perfectly reliable forecast. Second, the fractions skill score (FSS; Roberts and Lean 2008) was computed to compare the neighborhood grid coverage between forecasts and observations. The FSS is a positively oriented score, with 1 indicating a perfect forecast and 0 indicating a no skill forecast. For the surrogate severe verification, FSS was computed using binary report fields.

Finally, forecasts are examined via the critical success index (CSI; Schaefer 1990), which is calculated as follows:

$$\text{CSI} = \frac{\text{hits}}{\text{hits} + \text{misses} + \text{false alarms}}, \tag{3}$$

at the same thresholds as the ROC area. The CSI does not take into account the effect of correct nulls, which makes it a useful score for the rare event scenario, when most events fall into the correct null element of the 2×2 contingency table. The CSI is displayed using a performance diagram (Roebber 2009), which visualizes multiple verification metrics simultaneously.

Object-based verification was performed using MODE (Davis et al. 2006, 2009). MODE was applied to the simulated reflectivity forecasts at each hour, to discern diurnal trends in the number and size of storms. MODE was also applied to the 24-h surrogate severe fields at the 15% threshold, which meets the categorical definition of a slight risk of severe wind or hail according to the SPC. Slight risk areas were compared between the surrogate severe fields and practically perfect fields for surrogate severe fields with combinations of σ and UH percentile that optimized either the FSS, ROC area, or the reliability component of the Brier score. Practically perfect fields were created using $\sigma = 120$ km, following Sobash et al. (2011) and Hitchens et al. (2013).

In addition to the methods of objective verification described above, subjective evaluation by participants in the 2018

SFE took place via a survey, which asked them to rate forecasts of simulated composite reflectivity overlaid with 2–5-km UH greater than 75 $m^2 s^{-2}$ from each model on a scale from 1 (very poor) to 10 (very good). Participants were also encouraged to provide comments on the configurations, to elaborate on their scores. The comments provided information about what model characteristics captured participants' attention, highlighting forecast aspects of good performance and forecast aspects needing improvement on a case-by-case basis.

3. Results

a. Model climatologies

When calculating UH climatologies for the FV3-based models and the HRRRv3, obvious differences in the model behavior were seen at both the native grid resolution (Fig. 1a, dashed lines) and when the UH was regridded to the 80-km grid to calculate surrogate severe fields (Fig. 1a, solid lines). For example, the 70th percentile of UH in the HRRRv3 on the 80-km grid was ~10 $m^2 s^{-2}$, versus 60 $m^2 s^{-2}$ in the FV3-based models. The FV3-based models have similar UH climatologies on the 80-km grid until very high percentile values (~95th percentile and above), while at the native resolution the differences between FV3-based models get smaller as the percentile increases. In all cases, the differences between the FV3-based models and the HRRRv3 is much larger than the differences among the FV3-based models.

The differences in UH climatology reflect the need for comparing FV3-based models and WRF-ARW-based models using UH percentiles, rather than fixed UH thresholds. By using percentiles rather than thresholds, meaningful comparisons can be made between relatively high values of UH in each model. Considering that explicit measurements of UH do not exist and that there has been limited work on what a UH value should theoretically be, we instead use LSRs as a proxy for verification. Additionally, severe convective storms are relatively rare events that fall at the tails of the model distributions, so percentile exceedance values are more appropriate than fixed threshold exceedance values.

The reflectivity climatology shows smaller intramodel differences than the UH climatologies throughout most of the distribution, although the HRRRv3 has a smoother distribution (Fig. 1b). The FV3-based models have higher reflectivities at most of the larger percentiles, starting at about the 90th percentile. As simulated composite reflectivity is a diagnostic that depends on model physics and postprocessing (Koch et al. 2005), the differences could be attributable to many different sources. However, unlike UH, we have observations of composite reflectivity to compare to the model forecasts. As such, the analyses of composite reflectivity will focus on fixed thresholds of composite reflectivity rather than percentile-based thresholds.

b. Contingency table–based results

The surrogate severe fields described in section 2b were used to compare overall model performance using different verification metrics, with the idea that the best performance from each model (irrespective of the specific UH percentile/σ combination) could be compared. In terms of ROC area and FSS, the HRRRv3 outperformed all three FV3-based models (Fig. 3), with a large swath of the UH percentile/σ space having higher values than the corresponding space in the FV3-based models. Optimal smoothing levels (as indicated by σ) for the ROC area performance was similar between all four models examined, with the NSSL-FV3 achieving its highest ROC area at a slightly larger σ than the other three models. However, the HRRRv3 and the NSSL-FV3 had larger parameter space areas of higher ROC area than the CAPS-FV3 or the GFDL-FV3. The ROC area is clearly more a function of UH percentile than of σ, showing the importance of POD in the rare event scenario to the calculation of the ROC area. In the rare event scenario, the ROC area is heavily penalized for missing events, and so optimizing the ROC area often results in high probabilities and overforecasting (Gallo et al. 2018). These ROC areas are less than those found by Potvin et al. (2019; their Fig. 2). We hypothesize that this decrease is due to the underlying weather occurring during SFE 2018 compared to SFE 2017, as the lowest maximum ROC area found by Potvin et al. (2019) is higher than the highest ROC maximum area found in this work (0.876 from Potvin et al. 2019, compared to 0.875 herein). However, a full investigation of the reasons for the decrease is beyond the scope of this work. These year-to-year differences motivate longer testing periods for experimental models, to capture a large sample of meteorological events. However, the subjective evaluation data collected during these relatively shorter time frames can help capture nuance that objective metrics with a large sample size may miss, which will be discussed in section 3d.

A similar pattern emerges with the FSS results, although the FSS shows more sensitivity to σ than the ROC area does, and is optimized at lower σ values than the ROC area (Table 3). Unlike the ROC area, which had approximately the same smoothing level to achieve the highest score between models, the HRRRv3 requires much less smoothing than the FV3-based models to achieve its highest FSS (σ = 90 km). Lower smoothing values preserve information provided by the model regarding which areas are most at risk, as smoothing lowers the maximum probability value and distributes the probability over a larger area. So, the less smoothing required to achieve the maximum score, the more useful the original information from the model. Also, larger smoothers require more computational resources to implement, so smaller smoothing values are more computationally optimal. While this may be less of a concern if postprocessing a large batch of data, if postprocessing is being done on a run-by-run basis at or near–real time, reducing the computational expense is necessary. Of the FV3-based models, the CAPS-FV3 requires the least smoothing to achieve its highest score (σ = 105 km). The NSSL-FV3 scores highest of the FV3-based models in FSS, but alongside the GFDL-FV3 requires the most smoothing to maximize FSS (σ = 120 km). However, the HRRRv3 has FSS scores better than the best FV3-based model score (0.3735) across a wide variety of percentile and σ variations.

These FSS scores are quite a bit lower than the FSS scores in Potvin et al. (2019), which examined iterations of the

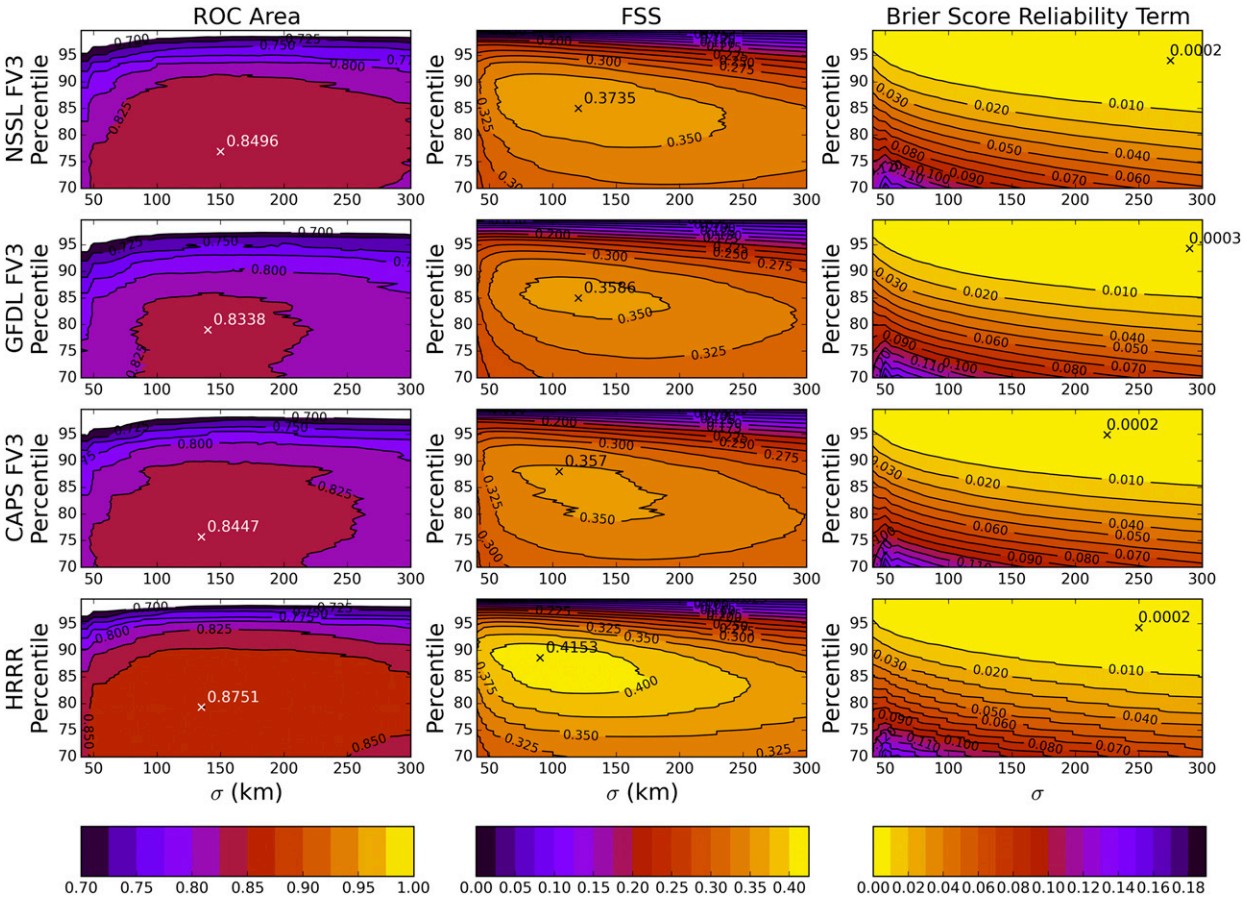

FIG. 3. Contour plots of verification metrics for each model considered (rows) for three different statistics (columns) across surrogate severe fields generated using different UH percentiles and σ in the Gaussian smoother. The combination of percentile and σ yielding the highest score for each metric and model combination is marked with an X, and the corresponding score is annotated to the upper right of the X.

GFDL-FV3 and CAPS-FV3 from the 2017 SFE. Much of this decrease was attributable to different methods of calculating the FSS; Potvin et al. (2019) use a smoothed, probabilistic field of observations in calculating their scores rather than a binary field of 1s and 0s, leading to smaller differences between the forecast and observed field. This accounts for a reduction in the FSS of ~0.2 for both models (not shown). However, besides the decrease due to the different methodology, a year-to-year decrease also occurred for these two models, with the GFDL-FV3 and CAPS-FV3 FSS scores in Fig. 3 being ~0.05 less than

TABLE 3. Percentile and σ values that maximize surrogate severe statistical scores. The probability of maximum CSI indicates the probabilistic threshold at which CSI is maximized for that UH percentile and σ combination.

Model	Score maximized	UH percentile	σ (km)	Probability of maximum CSI (%)
NSSL-FV3	ROC area	76.9	150	—
	FSS	85.0	120	—
	CSI	89.5	130	40%
CAPS-FV3	ROC area	75.7	135	—
	FSS	88.0	105	—
	CSI	93.7	130	25%
GFDL-FV3	ROC area	79.0	140	—
	FSS	85.0	120	—
	CSI	91.9	180	30%
HRRRv3	ROC area	79.3	135	—
	FSS	88.6	90	—
	CSI	82.0	130	60%

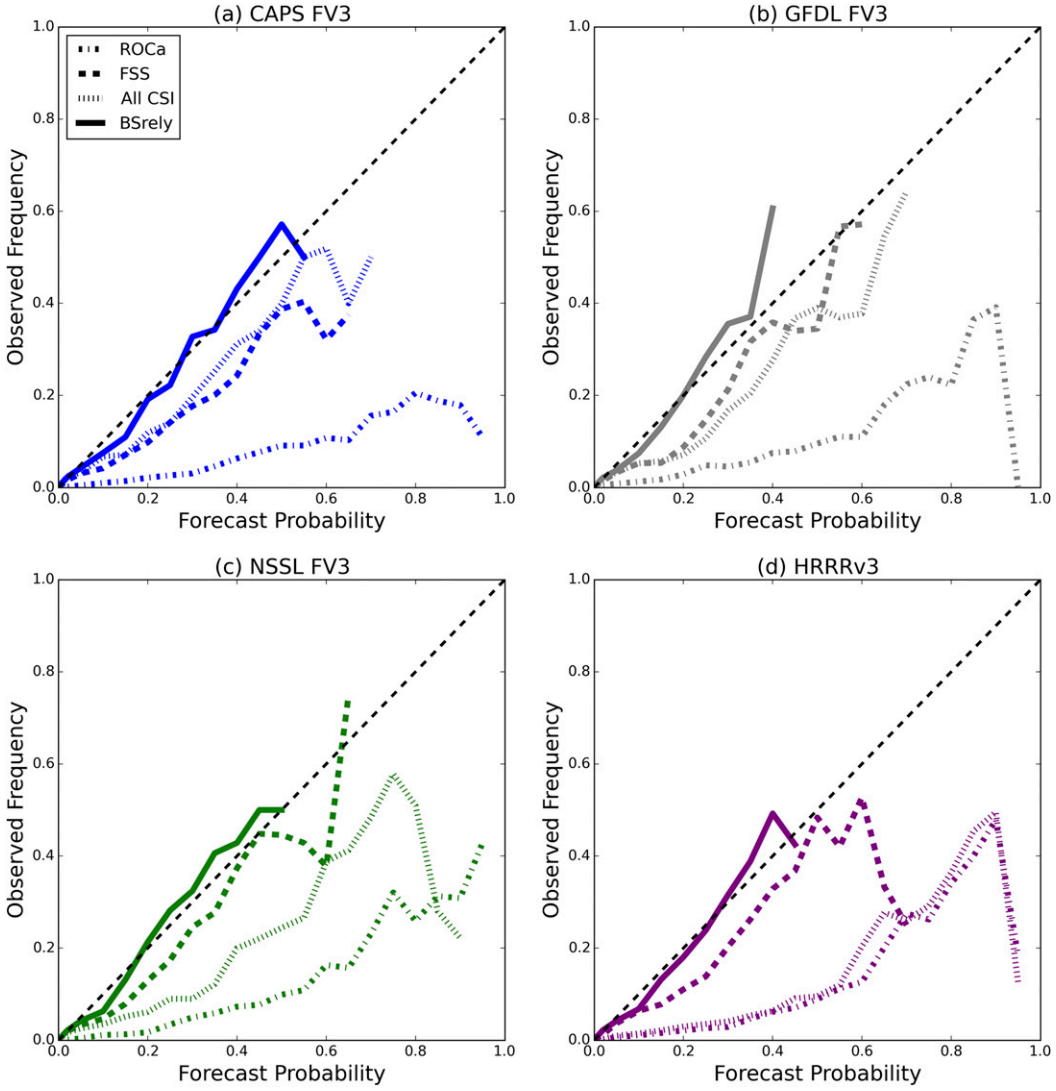

FIG. 4. A reliability diagram showing the performance of the surrogate severe fields created with the percentile and σ combination to optimize particular scores for (a) the CAPS-FV3, (b) the GFDL-FV3, (c) the NSSL-FV3, and (d) the HRRRv3. The different line styles represent different statistics optimized by the surrogate severe field. The black diagonal line indicates perfect reliability.

the Potvin et al. (2019) scores from SFE 2017 when calculated using a binary observation field (not shown). As in the case of the ROC areas, this decrease is speculated to be a function of the difference between the spring 2017 and spring 2018 seasons.

The minimum reliability component of the Brier score (Fig. 3) was extremely similar between models, indicating that a highly smoothed field generated using a high UH threshold achieved the best results for all four models. All models had a wide range of high UH values that achieved low BS_{rely} values. The reliability diagram for the surrogate severe fields generated using the optimal combinations of UH percentile and σ for each model show the high reliability of the smoothed fields that give the minimum BS_{rely} (Figs. 4a–d). However, these probabilities never get higher than 40%–55% due to the large amount of smoothing. The reliability diagram also shows the large overforecasting that goes along with maximizing ROC area, with overforecasting occurring at all probability levels for each of the models examined. The fields maximizing the FSS overforecast slightly, but fall much closer to the line of perfect reliability than the fields optimizing the ROC area. Optimizing the CSI (calculated at each potential probability threshold) led to more overforecasting than optimizing the FSS for the HRRRv3 and NSSL-FV3, but similar reliability for the GFDL-FV3 and CAPS-FV3 (Fig. 4a) between the two optimized fields. The CAPS-FV3 (Fig. 4a) and GFDL-FV3 (Fig. 4b) tended to have similar reliability when optimizing FSS and CSI, whereas the reliability of the HRRRv3 field that optimized CSI was more similar to the field that optimized the ROC area. The reliability of the NSSL-FV3

(Fig. 4c) optimized by CSI was in between the optimized FSS and optimized ROC area fields.

From these metrics, differences in the surrogate severe fields seem mostly consistent across smoothing levels and percentiles. The similarity of the relative model performance at individual probability thresholds is emphasized when looking at a performance diagram (Roebber 2009), which contains a point for each probability threshold for each combination of smoothing level and UH percentile (Fig. 5). The HRRRv3 consistently has higher PODs and success ratios than any of the FV3-based runs, particularly where bias is between 5.0 and 0.5. The overall difference in CSI between the HRRRv3 and the NSSL-FV3 or CAPS-FV3 is for the most part larger than the difference between the NSSL-FV3 and the CAPS-FV3, or the CAPS-FV3 and the GFDL-FV3. The probability maximizing CSI also differed much more between the HRRRv3 and the FV3-based models than among any of the FV3-based models, with the CSI being maximized at a much higher probability for the HRRRv3. Probabilities maximizing CSI in all of the FV3 models had a bias closer to 1 than the probability maximizing the CSI of the HRRRv3. While the overall performance of the HRRRv3 is consistently better according to Fig. 5, differences in σ and UH percentile used to generate the fields can have a significant impact on the look of the forecasts day-to-day, which may be better captured using object-based metrics.

c. *Object-based results*

In addition to the contingency table–based statistics, object-based statistics provide insight to the size and location of "slight risk" equivalent areas, as indicated by surrogate severe probabilities greater than 15%. The surrogate severe fields optimizing the ROC area, FSS, and BS_{rely} were selected to understand how the characteristics of these fields differed in an object-based framework. The mean, 75th, 90th, and maximum area generated by the models was larger than the observations in all cases (Fig. 6a), suggesting that the areas are generally too large when compared to the practically perfect fields generated using the LSRs, which uses a σ of 120 km in the Gaussian smoother. While the lower end of the distribution—the smaller areas—compare relatively well to the observations, the distribution of the model fields is overall shifted toward higher values compared to the observation distributions. The statistic being maximized influences how large the areas are, with the largest 15% areas created by maximizing the ROC area. Maximizing the reliability generally resulted in areas closest to the observations because the increased smoothing decreased the amount of area covered by the 15%. The FV3-based models also tended toward larger 15% areas compared to the HRRRv3 at the higher ends of the distribution. In terms of numbers of 15% objects, maximizing by the FSS and ROC area resulted in too many areas relative to the number of observed 15% areas ($n = 45$), although the specific number of observed 15% objects is a function of the smoothing used to generate practically perfect fields from the LSRs. In comparing the models, the FV3-based models produced fewer 15% areas than the HRRRv3. Optimizing by the reliability component of the Brier score created too few areas compared to observations.

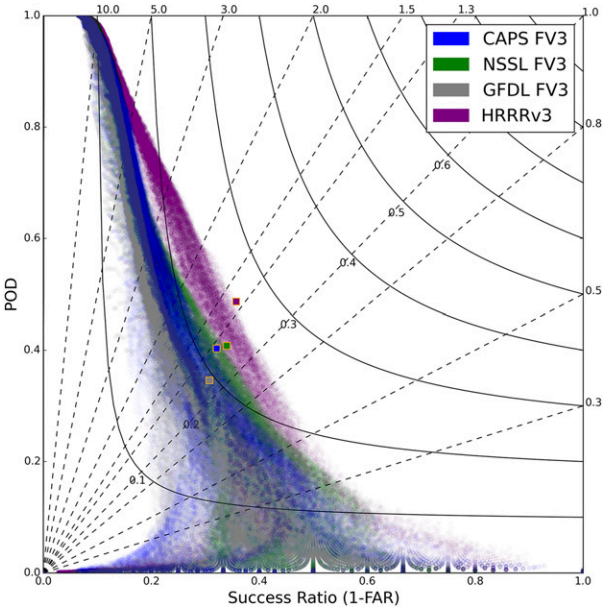

FIG. 5. A performance diagram for the surrogate severe fields generated using 100 different UH percentiles and 53 different σ values. Individual points are semitransparent; thus higher densities of color indicate more points. Percentage thresholds plotted for each of the 5300 surrogate severe fields at 2% and 5%, then in 5% intervals to 100%. The highest CSI achieved by each model for any UH percentile and σ combination is indicated by the opaque square outlined in orange. Solid black lines are lines of constant CSI, and dashed black lines are lines of constant reliability.

The only statistic showing no systemic differences between the FV3-based models and the HRRRv3 regarding the number of objects created was the BS_{rely}.

Besides wanting the correct size of 15% areas, ideally our objects also would be in the right location. Centroid distances between matched observed practically perfect objects and the model surrogate severe objects (Fig. 6b) varied from a minimum of 0.17 grid boxes (~13.8 km) to a maximum of 9.66 grid boxes (~785.1 km). Optimizing by the reliability component of the Brier score created the smallest differences between areas throughout the distribution of differences, but also the smallest number of matched pairs for a given ensemble. Maximizing by the ROC area generally created the largest differences for all models except the GFDL-FV3, and optimizing by FSS yielded the highest amount of matched objects for all given models. Within each optimized statistic, the HRRR produced more matched pairs than the FV3-based models, although this difference was much narrower for the FSS than for the other metrics.

While surrogate severe objects provide information on the daily extent and location of expected severe convection, reflectivity objects allow for examination of finer temporal scales (i.e., hourly) and of the evolution of convective systems throughout the day. Given that convective mode and evolution are two critical aspects of severe weather forecasting, reflectivity objects at thresholds somewhat lower than the typical magnitude of convective cores provide another important verification component. Reflectivity objects greater than or

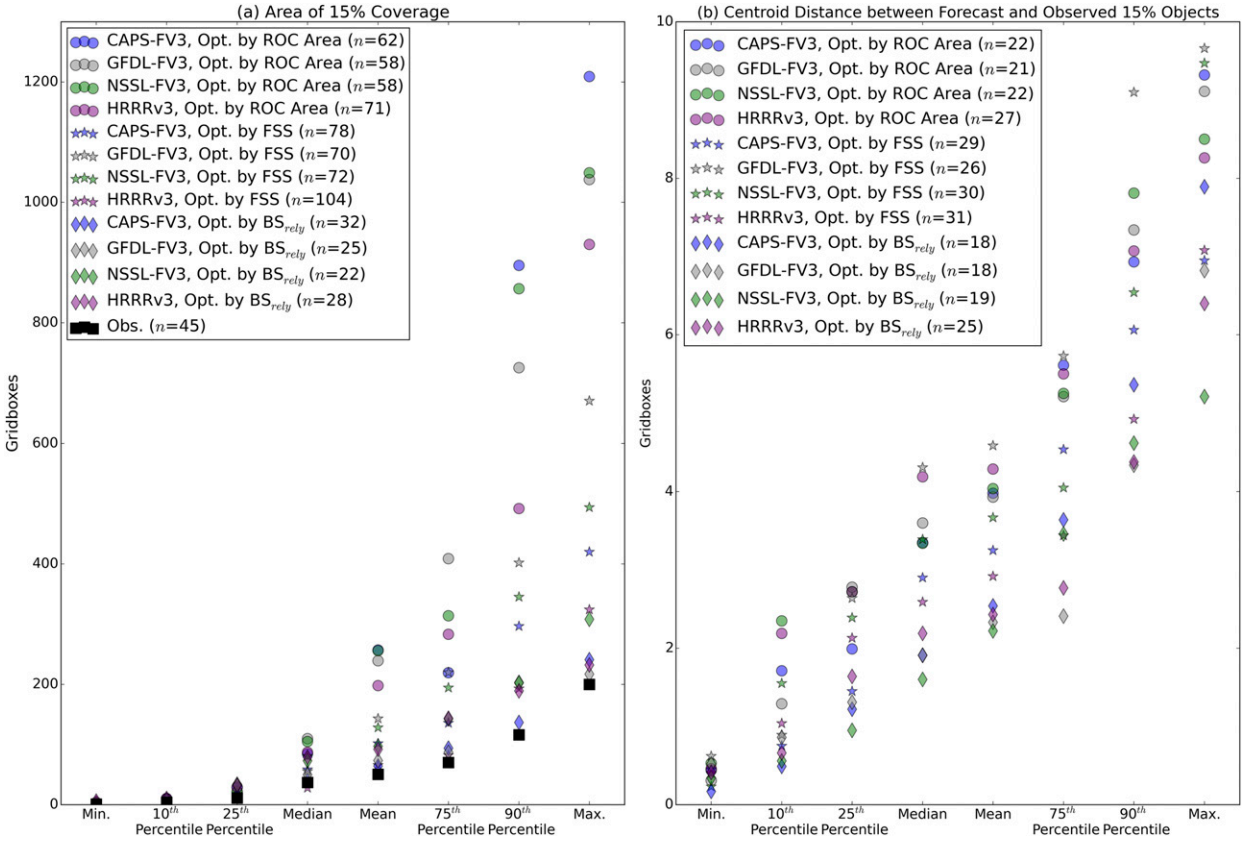

FIG. 6. Object-based statistics for the 15% surrogate severe fields generated using UH percentile and σ combinations to maximize a specific statistic. Marker shape indicates the statistic maximized and marker color indicates the model. Observations are indicated by black squares. Object-based attributes are (a) the area of 15% coverage (in terms of 80-km grid boxes), with the number of objects indicated in the legend, and (b) the centroid distance (in terms of 80-km grid boxes) between 15% objects, with the number of paired objects indicated in the legend.

equal to 30 dBZ from 2 May 2018 show how the HRRRv3 (Fig. 7d) captured the extent of the system across Kansas and Nebraska better than any of the FV3-based models (Figs. 7a–c), but also produced more areas of false alarm across Colorado and Wyoming. Additionally, the storms in Oklahoma and Texas were not captured well by any of the models, although the FV3-based models at least had reflectivity objects near the Texas and Oklahoma border.

Aggregating across the SFE 2018 cases, at the peak of the diurnal cycle the FV3-based models improve on the overall number of objects compared to the HRRRv3 at the 20- and 30-dBZ thresholds throughout the diurnal cycle. At the 20-dBZ threshold, all models overpredict the number of reflectivity objects (Fig. 8a), with the HRRRv3 and the GFDL-FV3 having the largest overprediction at the peak of the diurnal cycle, around forecast hours 21–25. For most of the time, the NSSL-FV3 is the closest to MRMS observations, although it does not maintain enough objects after the peak convective cycle. Similar patterns hold for the 30-dBZ threshold (Fig. 8b), although the GFDL-FV3 at this threshold is more similar to the other FV3-based models than it is to the HRRRv3 during the peak of the convective cycle. The peak of the convective cycle for the GFDL-FV3 is also shifted earlier than in the observations. The CAPS-FV3 overpredicts the number of objects for most forecast hours up to hour 25. The NSSL-FV3 initially has too many objects at early forecast hours, but does not have enough objects at the peak of the convective cycle and continues this underprediction through the end of the analyzed run time.

However, the previous thresholds are below what typically indicates convective storms. At the highest reflectivity threshold (45 dBZ), the HRRRv3 does much better than the other models, generating roughly the same number of objects as reality. All FV3-based models at the 45-dBZ threshold overpredict the number of objects, particularly early in the forecast cycle (e.g., forecast hours 2–12), perhaps indicating too intense of reflectivity values being produced by nonconvective storms. Additionally, the data assimilation used by the HRRRv3 likely contributes to the more realistic number of storms, since the FV3-based models are ''cold start.'' Early overprediction could occur in this scenario if, for example, the cold-start model attempts to develop a line of storms but instead gets multiple smaller cells. Future work will focus on incorporating data assimilation to FV3-based CAMs, to create a more direct comparison with the HRRR. There is only slight overprediction by the FV3-based models in forecast hours 13–19, but then the overprediction is amplified

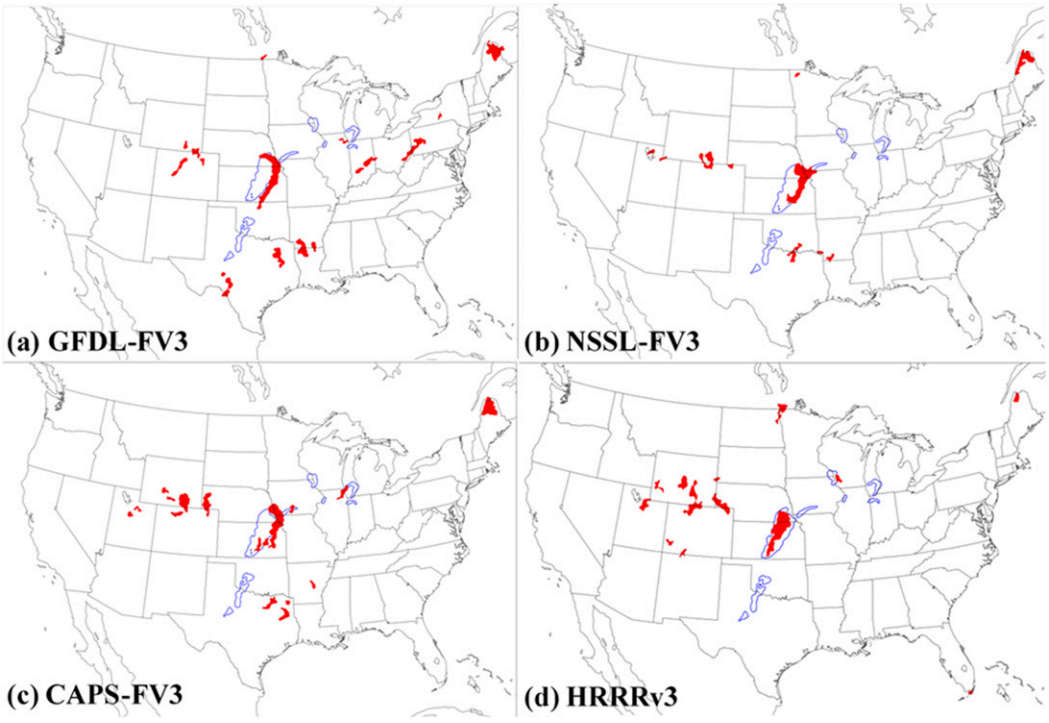

FIG. 7. Reflectivity objects greater than or equal to 30 dBZ at forecast hour 23 for (a) the GFDL-FV3, (b) the NSSL-FV3, (c) the CAPS-FV3, and (d) the HRRRv3 at 2300 UTC 2 May 2018. Model-simulated reflectivity objects are shown by the red filled contours, and observed reflectivity objects are shown by the blue contours.

again during forecast hours 21–36: the peak time for severe convection in the diurnal cycle. At this point in the forecast cycle, spinup issues from cold-start initialization are not expected. Therefore, for convective reflectivity objects, the HRRRv3 is still performing better than the FV3-based models. One potential concern with reflectivity objects is that a model with more small-scale detail may produce more but smaller continuous regions over a certain threshold, which may help explain the behavior of the FV3-based models.

In addition to object count, we examine the object area to determine if the models are producing storms that are covering approximately the same amount of area as the observed MRMS storms (Fig. 9). The area statistics are more similar to observations than the object count statistics. Therefore, since the models were generally overforecasting the number of objects, the storms produced by the models would necessarily be smaller than the observed storms for most individual storms. All models underforecast the area of 20-dBZ objects from forecast hours 19–30 (Fig. 9a), covering the peak convective coverage of the day. The HRRRv3 has the smallest fluctuation in area covered over the course of the day, relative to the rest of the models. At higher reflectivity thresholds, differences between the models and the observations lessen.

d. Subjective analyses

SFE 2018 participants assigned ratings to each of the models examined here, based on the composite reflectivity and hourly maximum UH. They were told to consider factors such as convective initiation, mode, evolution and timing in their ratings. Participant ratings of the models showed a similar pattern to the contingency table–based metrics of the surrogate severe fields, with the HRRRv3 ratings distribution having the highest median score (6/10; Fig. 10a). The CAPS-FV3 and NSSL-FV3 were often rated similarly, and the GFDL-FV3 was typically rated lowest of the four models in question. This pattern differs from the 2017 SFE, where an earlier version of the CAPS-FV3 was rated lower than the GFDL-FV3. Overall, the participant ratings of the HRRR and the FV3-GFDL decreased by 0.3 and 0.8 points compared to 2017, while the participant ratings of the CAPS-FV3 increased by 0.5 points.

Participant comments often focused on overforecasting the coverage and intensity of convective storms compared to the observations. Overforecasting was an issue common to all of the FV3-based models (Figs. 10b–d), but participant comments most often mentioned the GFDL-FV3 (Fig. 10c). This overforecasting likely led to a lower subjective score for the GFDL-FV3 compared to the other FV3-based models. Subjective scores help us understand how forecasters and other end-users of the model output interrogate the model output, which can help target areas for improvement that could be less evident from the bulk statistics. For example, the rapid and intense upscale growth mentioned by participants in their survey responses may also influence objective verification metrics, giving model developers a mechanism to investigate when trying to improve the overall objective verification statistics.

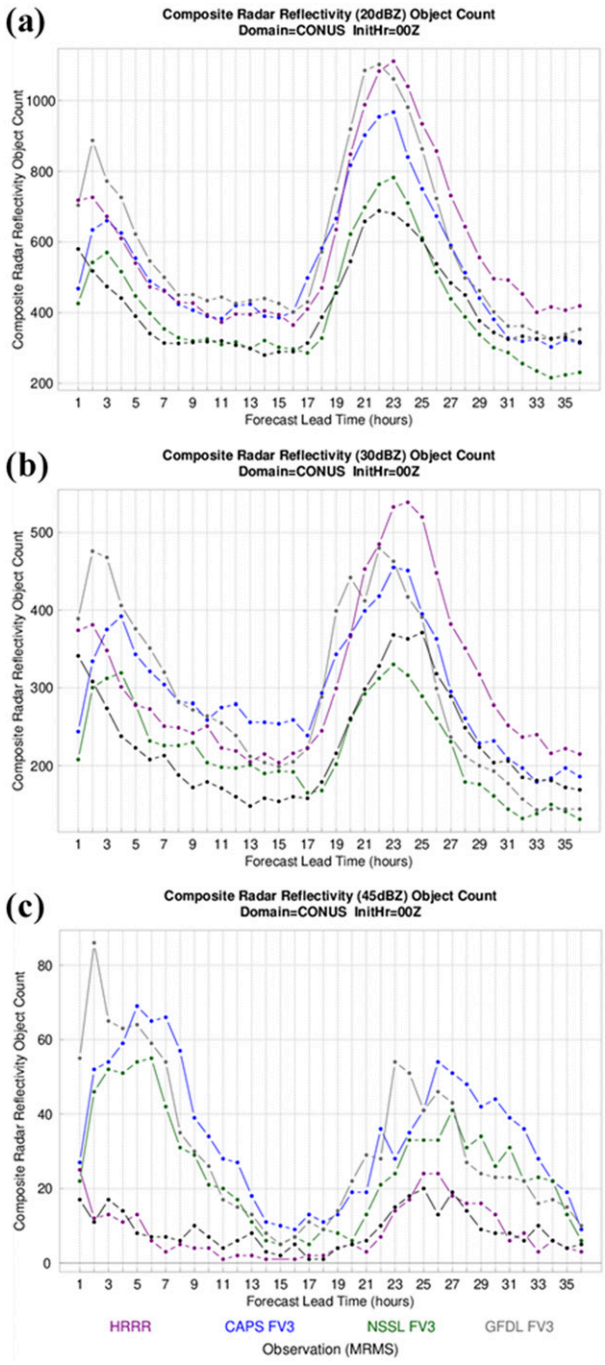

FIG. 8. Reflectivity object counts as a function of lead time at (a) 20-, b) 30-, and (c) 45-dBZ thresholds. Note that the y axis differs between subplots, scaling to show the diurnal cycle in object counts.

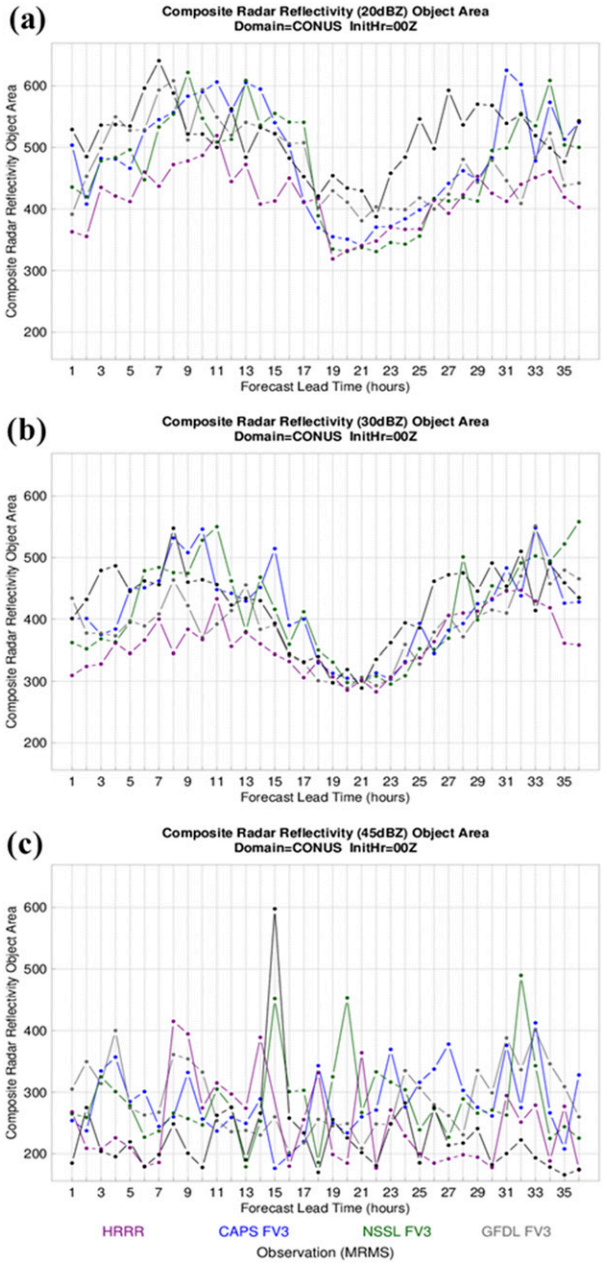

FIG. 9. Reflectivity object cumulative area in native resolution grid boxes as a function of lead time at (a) 20-, (b) 30-, and (c) 45-dBZ thresholds.

e. *Illustration of methods: 30 May 2018*

Accumulating aggregated statistics for surrogate severe fields calculated using multiple UH percentiles and smoothing values determines what combination is most skillful for a given metric. However, given that we have previously seen how the ''most skillful'' parameter combination can vary depending on what metric is maximized (Fig. 3), an example case is presented here to show how the practical appearance of the forecast can also subsequently vary. This case demonstrates why care must be taken when determining how to create surrogate severe fields for comparison across models.

On 30 May 2018, multiple areas of the CONUS faced a threat of severe weather. A negatively tilted shortwave trough moving toward the upper Mississippi valley, southern-stream perturbation affecting the central Rockies and High

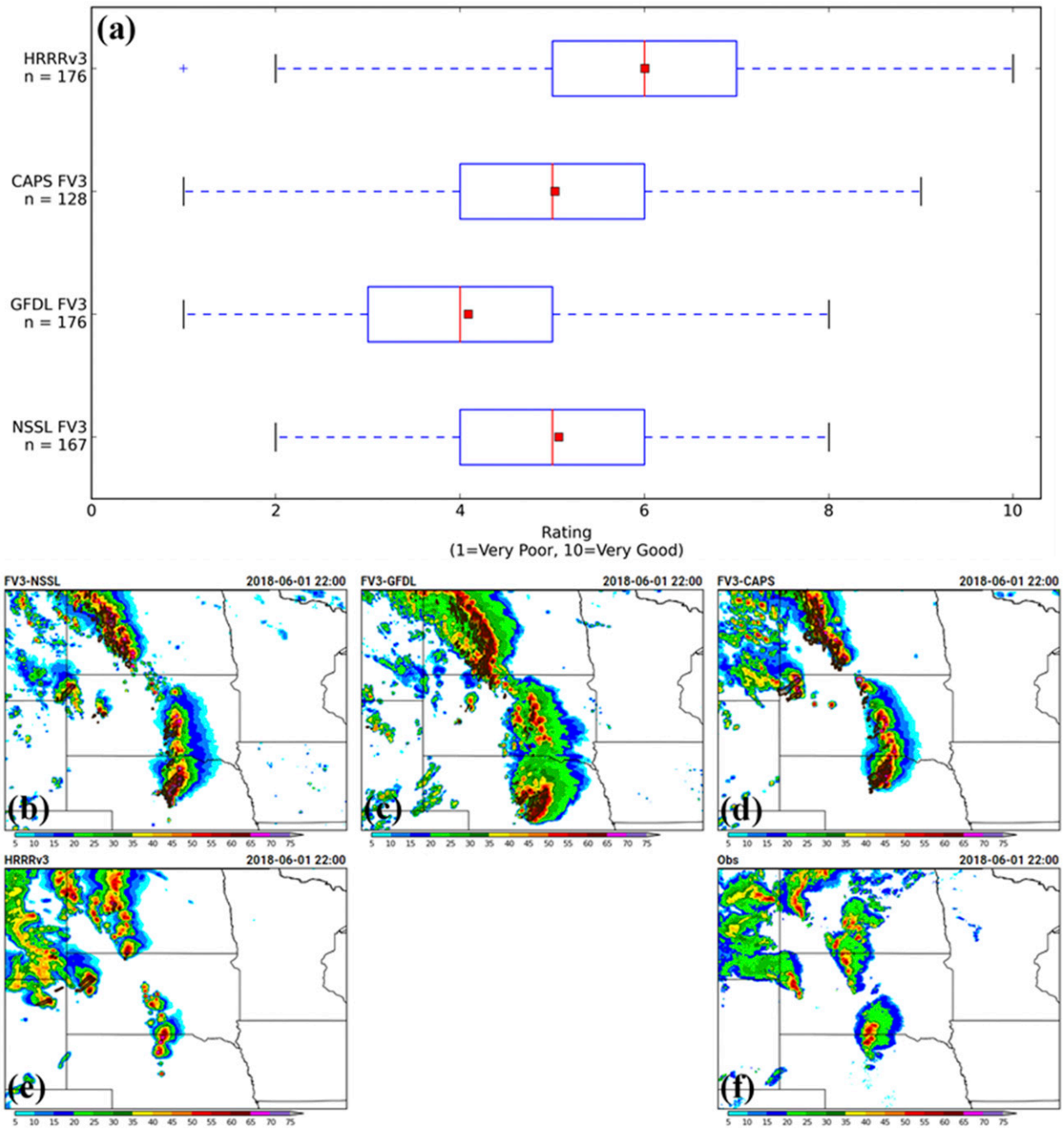

FIG. 10. (a) Subjective evaluations from SFE 2018 participants throughout the experiment, which were assigned based on looking at 24-h loops of simulated composite reflectivity and UH from (b) the NSSL-FV3, (c) GFDL-FV3, (d) CAPS-FV3, and (e) HRRRv3 compared to (f) observed composite reflectivity; plots for a sample case (1 Jun 2018) are shown. Local storm reports could also be overlaid for verification purposes (not shown).

Plains, and the remnants of a tropical system were all affecting different regions of the CONUS. Resultant mesoscale boundaries were also abundant from convection the previous day. Subjective ratings of model performance within the daily domain of interest were mixed, with the HRRRv3 performing best according to the nine survey respondents (not shown).

Surrogate severe fields designed to optimize the ROC area, FSS, and the reliability component of the Brier score show how different a forecast from the same model can look, simply by adjusting the UH threshold and σ value used to generate the surrogate severe fields (Fig. 11). In this case, we will examine forecasts from the NSSL-FV3, although similar results occurred for all models examined in this study. Maximizing the

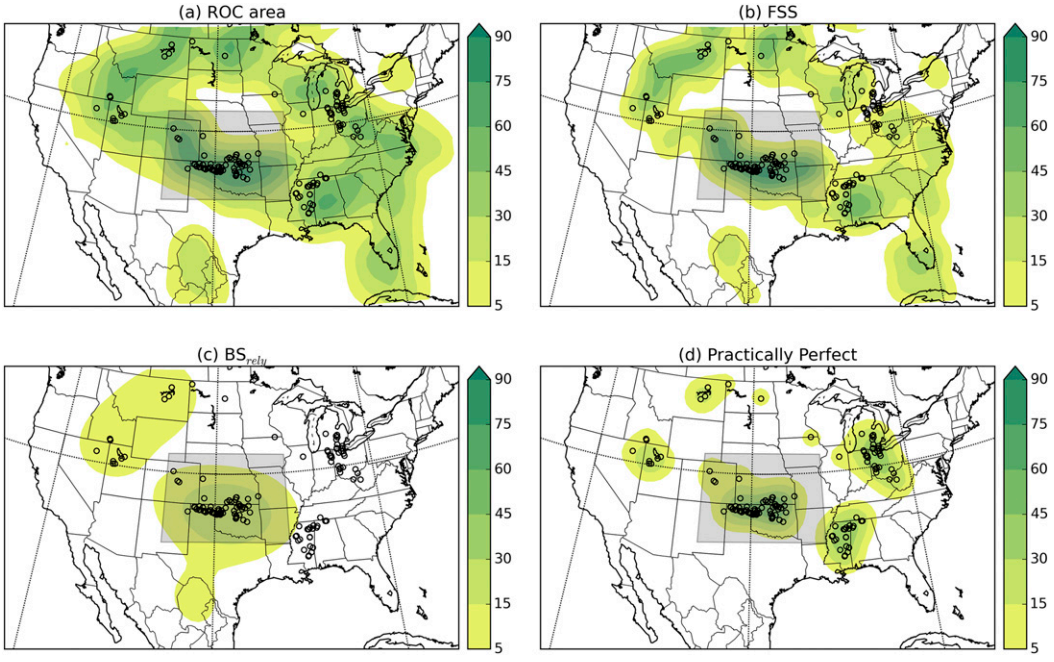

FIG. 11. Surrogate severe fields on 30 May 2018 generated from the NSSL-FV3 using the UH percentile and σ that optimizes (a) ROC area, (b) FSS, and (c) the reliability component of the Brier score throughout the entire SFE. The black circles indicate local storm reports for the day, which are used to construct (d) the practically perfect field. The shaded box indicates the domain of the day during the SFE, over which the subjective evaluations were performed.

ROC area (Fig. 11a) results in broad swaths of high probability and an abundance of false alarm area. While all of the reports fall within areas of probability $> 5\%$, the forecaster would be hard-pressed from this depiction to determine which area of the CONUS is most at threat for severe weather. Optimizing by FSS maintains low probabilities across much of the CONUS, but decreases probabilities in the eastern part of the domain, where less observed severe weather occurred. However, a relative maximum remains in the southeast, shifted just east of where the reports occurred. High probability areas across the Oklahoma and Texas panhandles are also maintained. This area, which was the area of focus during the SFE that day and thus had the highest anticipated threat for severe weather, maintains high probabilities when optimizing by either the ROC area or the FSS. Finally, optimizing by the reliability component of the Brier score (Fig. 11c) leads to large swaths of relatively low probability across the southern Plains and the northern Rockies, completely eliminating the area across the southeast and the area with many reports in the Midwest. While the area of most concentrated reports is encompassed within the area of highest probability, the probability is only between 15% and 30% and lacks specificity compared to the other optimization methods, covering all of Oklahoma, half of Kansas, and all of the Texas Panhandle.

4. Conclusions

This work examined numerical weather prediction forecasts made as part of NOAA's Hazardous Weather Testbed SFE in 2018, including three experimental convection-allowing models using the FV3 dynamical core. Since FV3 will serve as the dynamical core of the U.S. Unified Forecast System, evaluating its performance for forecasting severe convection is critical. As a baseline, the models are evaluated alongside an operational deterministic CAM, the HRRRv3. These early iterations of FV3-based CAMs show promise, particularly in the realm of object-based verification measures. The NSSL-FV3 and CAPS-FV3 create more realistic diurnal cycles of numbers of storms at lower reflectivity thresholds than the current operational HRRRv3, although those storms may be smaller than the storms in the observed MRMS dataset. However, work remains to improve the performance of the FV3-based systems at the higher reflectivity thresholds that are more indicative of convective processes. The surrogate severe field areas also tended to be larger than corresponding areas of practically perfect probability generated using LSRs.

Contingency table–based verification metrics, however, show that the FV3-based models still need work to achieve the bulk statistical skill of the HRRRv3. Even when testing surrogate severe fields generated using a variety of UH percentiles and σ values, the highest scores achieved by FV3-based models were lower than those produced by the HRRRv3. Objective verification scores such as the POD, success ratio, FSS, and ROC area mirrored the subjective evaluations carried out by participants in showing the HRRRv3 scoring best, followed in order by the NSSL-FV3 and CAPS-FV3, with the GFDL-FV3 generally scoring the lowest. While reflectivity climatologies were similar between all of the models, the UH of FV3-based

systems tended to contain much higher values than those of the HRRRv3, mirroring what Potvin et al. (2019) found with models run during the 2017 SFE. However, scores overall were lower than those found by Potvin et al. (2019) for SFE 2017 for prior iterations of the GFDL-FV3 and CAPS-FV3.

This work also demonstrates the impact of a critical facet of designing verification of any system destined for widespread adaptation: score selection. Computing the surrogate severe fields based on what metric was maximized resulted in drastically different forecasts, despite the fact that each of these forecasts was ''best'' in one way or another. Integrating subjective evaluation may help make the decision of what score to maximize, but it may be that the tradeoffs made to optimize a given metric could degrade other aspects of the forecast to the extent that the forecast is no longer useful. For example, optimizing based on ROC area leads to large overforecasting, but optimizing based on the reliability component of the Brier score leads to broad swaths of relatively low probability and a decrease in sharpness of the forecasts. Future work will examine how to best compute surrogate severe fields, and whether it is possible to design an optimized metric that combines aspects of multiple verification scores in such a way that minimizes trade-offs between important forecast aspects and matches subjective end-user impressions.

Acknowledgments. The authors thank the participants of the 2018 SFE for contributing their input and perspective regarding the relative performance of these models and discussion of metrics. BTG, BR, and YW were provided support by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA16OAR4320115, U.S. Department of Commerce. Author AJC completed this work as part of regular duties at the federally funded NOAA National Severe Storms Laboratory. Author ILJ completed this work as part of regular duties at the federally funded NOAA Storm Prediction Center. CAPS's FV3 runs were supported by NOAA Grants NA19OAR4590141, NA17OAR4590186, and NA16NWS4680002. The Developmental Testbed Center (DTC) is funded by the National Oceanic and Atmospheric Administration (NOAA), the U.S. Air Force, the National Center for Atmospheric Research (NCAR), and the National Science Foundation (NSF). NCAR is a major facility sponsored by the National Science Foundation under Cooperative Agreement 1852977.

Data availability statement. HRRRv3 model data used in this study were provided to the Storm Prediction Center (SPC) by NCEP's Environmental Modeling Center and are archived internally at the National Severe Storms Laboratory (NSSL). GFDL-FV3 and CAPS-FV3 model data were provided by the Geophysical Fluid Dynamics Laboratory (GFDL) and Center for the Analysis and Prediction of Storms (CAPS) at the University of Oklahoma, respectively. GFDL-FV3, CAPS-FV3, and NSSL-FV3 model data were transferred to NSSL as part of the Community Leveraged Unified Ensemble (CLUE; Clark et al. 2018) during the 2018 Spring Forecasting Experiment. These datasets are archived internally at NSSL for the period used in this study. 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 internally at the Developmental Testbed Center (DTC). 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, GFDL-FV3, and CAPS-FV3 datasets).

REFERENCES

Alexander, C., and Coauthors, 2017: WRF-ARW research to operations update: The Rapid-Refresh (RAP) version 4, High-Resolution Rapid Refresh (HRRR) version 3 and convection-allowing ensemble prediction. *18th WRF User's Workshop*, Boulder, CO, UCAR–NCAR, 2.5, https://ruc.noaa.gov/ruc/ppt_pres/Alexander_WRFworkshop_2017_Final.pdf.

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

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.

Cai, H., and R. E. Dumais, 2015: Object-based evaluation of a numerical weather prediction model's performance through forecast storm characteristic analysis. *Wea. Forecasting*, **30**, 1451–1468, https://doi.org/10.1175/WAF-D-15-0008.1.

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

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

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

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

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

Daniels, T. S., W. R. Moninger, and R. D. Mamrosh, 2006: Tropospheric Airborne Meteorological Data Reporting (TAMDAR) overview. *10th Symp. on Integrated Observing and Assimilation Systems for Atmosphere, Oceans, and Land Surface,* Atlanta, GA, Amer. Meteor. Soc., 9.1, http://ams.confex.com/ams/pdfpapers/104773.pdf.

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

——, ——, ——, and J. Halley-Gotway, 2009: The Method for Object-Based Diagnostic Evaluation (MODE) applied to

numerical forecasts from the 2005 NSSL/SPC Spring Program. *Wea. Forecasting*, **24**, 1252–1267, https://doi.org/10.1175/2009WAF2222241.1.

Done, J., C. A. Davis, and M. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the Weather Research and Forecasting (WRF) model. *Atmos. Sci. Lett.*, **5**, 110–117, https://doi.org/10.1002/asl.72.

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.

——, A. J. Clark, B. T. Smith, R. L. Thompson, I. Jirak, and S. R. Dembek, 2018: Blended probabilistic tornado forecasts: Combining climatological frequencies with NSSL-WRF ensemble forecasts. *Wea. Forecasting*, **33**, 443–460, https://doi.org/10.1175/WAF-D-17-0132.1.

——, 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.

Gallus, W. A., 2010: Application of object-based verification techniques to ensemble precipitation forecasts. *Wea. Forecasting*, **25**, 144–158, https://doi.org/10.1175/2009WAF2222274.1.

Griffin, S. M., J. A. Otkin, C. M. Rozoff, J. M. Sieglaff, L. M. Cronce, C. R. Alexander, T. L. Jensen, and J. K. Wolff, 2017: Seasonal analysis of cloud objects in the High-Resolution Rapid Refresh (HRRR) model using object-based verification. *J. Appl. Meteor. Climatol.*, **56**, 2317–2334, https://doi.org/10.1175/JAMC-D-17-0004.1.

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

——, ——, and C.-Y. Tu, 2016: High-resolution climate simulations using GFDL HiRAM with a stretched global grid. *J. Climate*, **29**, 4293–4314, https://doi.org/10.1175/JCLI-D-15-0389.1.

——, S. L. Rees, M. Morin, L. Zhou, and W. F. Stern, 2019: Explicit prediction of continental convection in a skillful variable-resolution global model. *J. Adv. Model. Earth Syst.*, **11**, 1847–1869, https://doi.org/10.1029/2018MS001542.

Hazelton, A. T., M. Bender, M. Morin, L. Harris, and S. Lin, 2018: 2017 Atlantic hurricane forecasts from a high-resolution version of the GFDL fvGFS model: Evaluation of track, intensity, and structure. *Wea. Forecasting*, **33**, 1317–1337, https://doi.org/10.1175/WAF-D-18-0056.1.

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., 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.

Kain, J. S., S. J. Weiss, J. J. Levit, M. E. Baldwin, and D. R. Bright, 2006: Examination of convection-allowing configurations of the WRF model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. *Wea. Forecasting*, **21**, 167–181, https://doi.org/10.1175/WAF906.1.

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

Kleist, D. T., D. F. Parrish, J. C. Derber, R. Treadon, W.-S. Wu, and S. Lord, 2009: Introduction of the GSI into the NCEP Global Data Assimilation System. *Wea. Forecasting*, **24**, 1691–1705, https://doi.org/10.1175/2009WAF2222201.1.

Koch, S. E., B. S. Ferrier, M. T. Stoelinga, E. J. Szoke, S. J. Weiss, and J. S. Kain, 2005: The use of simulated radar reflectivity fields in diagnosis of mesoscale phenomena from high-resolution WRF model forecasts. *11th Conf. on Mesoscale Processes/32nd Conf. on Radar Meteorology*, Albuquerque, NM, Amer. Meteor. Soc., J4J.7, http://ams.confex.com/ams/pdfpapers/97032.pdf.

Loken, E. D., A. J. Clark, M. Xue, and F. Kong, 2019: Spread and skill in mixed- and single-physics convection-allowing ensembles. *Wea. Forecasting*, **34**, 305–330, https://doi.org/10.1175/WAF-D-18-0078.1.

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

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83**, 407–430, https://doi.org/10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2.

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

——, 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

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

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

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.

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

Roberts, B., I. Jirak, A. Clark, S. Weiss, and J. 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.

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

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570–575, https://doi.org/10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2.

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.

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

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

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

Smith, B. T., R. L. Thompson, J. S. Grams, C. Broyles, and H. E. Brooks, 2012: Convective modes for significant severe thunderstorms in the contiguous United States. Part I: Storm classification and climatology. *Wea. Forecasting*, **27**, 1114–1135, https://doi.org/10.1175/WAF-D-11-00115.1.

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

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

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

Surcel, M., I. Zawadzki, M. K. Yau, M. Xue, and F. Kong, 2017: More on the scale dependence of the predictability of precipitation patterns: Extension to the 2009–13 CAPS Spring Experiment ensemble forecasts. *Mon. Wea. Rev.*, **145**, 3625–3646, https://doi.org/10.1175/MWR-D-16-0362.1.

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

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

Wolff, J. K., M. Harrold, T. Fowler, J. H. Gotway, L. Nance, and B. G. Brown, 2014: Beyond the basics: Evaluating model-based precipitation forecasts using traditional, spatial, and object-based methods. *Wea. Forecasting*, **29**, 1451–1472, https://doi.org/10.1175/WAF-D-13-00135.1.

Wu, W., 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.

Zhang, C., and Coauthors, 2019: How well does the FV3-based model predict precipitation at a convection-allowing resolution? Results from CAPS forecasts for the 2018 NOAA Hazardous Weather Testbed with different physics combinations. *Geophys. Res. Lett.*, **46**, 3523–3531, https://doi.org/10.1029/2018GL081702.

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