Exploring Hourly Updating Probabilistic Guidance in the 2021 Spring Forecasting Experiment with Objective and Subjective Verification

ANDREW R. WADE[a] AND ISRAEL L. JIRAK[b]

[a] *Cooperative Institute for Severe and High-Impact Weather Research and Operations, University of Oklahoma, Norman, Oklahoma*
[b] *NOAA/NCEP/Storm Prediction Center, Norman, Oklahoma*

(Manuscript received 24 November 2021, in final form 16 February 2022)

ABSTRACT: This study explored how forecasters can best use the two main forms of operational convection-allowing model guidance: the High-Resolution Ensemble Forecast (HREF) system and the hourly High-Resolution Rapid Refresh (HRRR). The former represents a wider range of possible outcomes, but the latter updates much more frequently and incorporates newer observations. HREF and time-lagged High-Resolution Rapid Refresh (HRRR-TL) probabilistic forecasts of reflectivity and updraft helicity, as well as two methods of combining HREF and HRRR into hourly updating blended guidance, were evaluated for the 2021 Spring Forecasting Experiment (SFE) period. In both objective skill and the subjective ratings of SFE participants, the 1200 UTC HREF proved difficult to outperform over this sample of events, even when incorporating HRRR initializations as late as 1800 UTC. It was usually better to use either of the experimental blending techniques than to simply discard the older HREF in favor of newer HRRR solutions. The greater model diversity and dispersion of solutions within the HREF is likely primarily responsible for this result. A possible bias in diurnal convection initiation timing and coverage in the newly upgraded HRRRv4 was also investigated, including on subdomains targeted to weakly forced diurnal initiation, and was found to have little or no systematic effect on HRRRv4's operational utility.

KEYWORDS: Forecast verification/skill; Mesoscale forecasting; Numerical weather prediction/forecasting; Short-range prediction; Model evaluation/performance

1. Introduction

The High-Resolution Rapid Refresh (HRRR; Smith et al. 2008; Benjamin et al. 2016) model and the other members of the National Centers for Environmental Prediction (NCEP) High-Resolution Ensemble Forecast (HREF; Roberts et al. 2019) compose the primary convection-allowing model (CAM) guidance for short-term severe weather forecasting in the United States, including for the Storm Prediction Center (SPC). The full membership of the HREF updates twice a day (0000 and 1200 UTC), but the HRRR runs hourly. The frequency of updates allows an informal time-lagged ensemble of the most recent HRRR runs, hereafter called HRRR-TL. Newer HRRR-TLs between the synoptic times should be expected to benefit from assimilating more recent observations than the older HREF. However, the HREF consistently outperforms more formally constructed CAM ensembles (Clark et al. 2020, 2021), and its diversity of model cores, physics, and initial and boundary conditions can be expected to capture a broader range of outcomes (Roberts et al. 2020). It has not been well established how these two types of guidance, each with a plausible advantage over the other, should be weighed in combination by human forecasters.

To address this gap, hourly updating blends of the most recent HRRR-TL and the most recent HREF output were formulated by two different methods, as detailed in the next section. Preliminary testing on supercell and MCS events from 2019 and 2020 (using HREF and HRRR versions operational at those times) suggested these blends might offer small improvements in skill by 1800–2100 UTC compared to either HREF or HRRR-TL alone. As part of the Hazardous Weather Testbed Spring Forecasting Experiment (SFE) in 2021 (Clark et al. 2021), Day 1 reflectivity and updraft helicity (UH) fields were evaluated for HREF, HRRR-TL, and the two blends. The SFE was chosen as the testing period to allow direct comparison of objective and subjective skill over the same events and domains.

This short study answers three questions:

1) How relatively skillful are the 1200 UTC HREF and the 1200, 1500, and 1800 UTC HRRR-TL for Day 1 convective forecasting?
2) Could a prescribed hourly-updating blend of the 1200 UTC HREF and more recent HRRR runs be more skillful than HREF or HRRR-TL alone?
3) How well does objective verification of this guidance agree with the subjective ratings of human forecasters?

2. Data and methods

a. HREF and HRRR-TL probabilistic fields

HREFv3 comprises the HRRRv4, the North American Mesoscale Forecast System (NAM) nest, two configurations of the Advanced Research version of the Weather Research and Forecasting (WRF-ARW) Model (Skamarock et al. 2008), and the FV3 model. Roberts et al. (2019; Table 1) details the membership of HREFv2, which remains consistent in HREFv3 except that the FV3 replaces the WRF-NMMB. [CAMs using the FV3 core were specifically evaluated in a

Corresponding author: Andrew Wade, andrew.wade@noaa.gov

DOI: 10.1175/WAF-D-21-0193.1

TABLE 1. Pairwise comparisons of HREF FSS to other guidance.

Reflectivity guidance	FSS > HREF FSS
1200 UTC HRRR-TL	5.8%
1500 UTC HRRR-TL	9.4%
1800 UTC HRRR-TL	21.7%
1800 UTC time-based blend	58.7%
1800 UTC error-based blend	55.1%

previous SFE (Gallo et al. 2021), in which they performed well by object-based verification metrics but produced somewhat less skillful surrogate severe fields than HRRRv3.] A time-lagged run of each model is also included: 6 h old for the HRRR, and 12 h old for the others, for a total of 10 members. All members are weighted equally in the HREF probabilities calculated for this work. The HRRR-TL consists of the four most recent HRRR runs, all weighted equally. HRRRv4 replaced HRRRv3 in National Weather Service operations shortly before the 2021 SFE.

This study evaluates forecasts of composite radar reflectivity exceeding 40 dBZ and of UH in the 2–5 km AGL layer exceeding 75 $m^2 s^{-2}$, using neighborhood maximum ensemble probabilities (NMEPs) of reflectivity and UH above these thresholds, as described in Schwartz and Sobash (2017) and Roberts et al. (2019). Square neighborhoods of 40 km are used for both fields. As Roberts et al. (2019) emphasize, the NMEP at a point is the ensemble probability of threshold exceedance anywhere within the neighborhood centered on that point, rather than a fractional coverage of the neighborhood. After NMEPs are calculated, a Gaussian smoother ($\sigma = 13$) is applied. Note that while probabilistic HREF fields are available from NOAA, the NMEPs used here are calculated from the individual members to ensure the same methodology is applied to both HREF and HRRR-TL NMEPs.

b. Blend formulation

Two blends of HREF and HRRR guidance were tested in the SFE. The "time-based blend" considers the relative age of the HREF and HRRR-TL being blended. Probabilistic fields for both HREF and HRRR-TL are calculated independently. Then a weighted average of those two fields is taken with weights determined by lead time. The HREF weight is given by the ratio of the lead time of the HRRR-TL to the lead time of the HREF, and the weights add to one:

$$\text{weight}_{\text{HREF}} = \frac{t_{\text{forecast}} - t_{\text{HRRR-TL}}}{t_{\text{forecast}} - t_{\text{HREF}}};$$

$$\text{weight}_{\text{HRRR-TL}} = 1 - \text{weight}_{\text{HREF}}. \quad (1)$$

As the lead time of the HRRR-TL decreases toward zero with each hourly run, its weight increases linearly from zero to one. The 1800 UTC time-based blend is the version evaluated in the SFE. This gives greater weight to HRRR-TL for forecast times before 0000 UTC, greater weight to HREF for forecast times after 0000 UTC, and equal weights for the 0000 UTC forecast, for which the 1800 UTC HRRR-TL has exactly half the lead time of the 1200 UTC HREF.

The "error-based blend" considers short-term errors in the individual members of both HREF and HRRR-TL, such that the members with the smallest errors in observed fields at the time the blend is created are given the largest weight. This strategy followed from a preliminary finding, in the same test dataset mentioned in section 1, that such short-term errors were weakly negatively correlated with the skill of reflectivity and UH forecasts later in the period. To create the error-based blend, the 10 members of the most recent HREF and the 4 members of the most recent HRRR-TL are combined into a 14-member ensemble. Each member is compared to the NOAA Real-Time Mesoscale Analysis (RTMA; De Pondeca et al. 2011) valid at the blend's initialization time over the domain of interest. For example, for an 1800 UTC blend, the 1700 UTC HRRR-TL member's 1-h forecast and the 1200 UTC NSSL WRF's 6-h forecast would both be compared to the 1800 UTC RTMA. Root-mean-square errors are calculated over the domain for each member's 2-m temperature, 2-m dewpoint, and 10-m u and v wind components. These errors are then normalized among the 14 members. Members are scored by the total of their normalized errors in those four fields. Weights are assigned such that the largest total error score receives a weight of zero and the smallest a weight of one, with all other members' weights falling in between in proportion to their total errors.

c. Verification

The Multi-Radar Multi-Sensor (MRMS; Smith et al. 2016) merged reflectivity product is used as truth for reflectivity verification. However, UH is not operationally measurable. Because UH in CAMs is a proxy for supercell hazards, storm reports of tornadoes (of any intensity) and hail exceeding 1 in. in diameter are used as truth. This approach has several limitations. Perhaps the most serious is underreporting of severe hail in sparsely populated areas, particularly in West Texas during several SFE events. In the opposite direction, nonsupercell tornadoes penalize UH guidance that is only intended to represent supercells. However, using reports is the simplest approach and is directly tied to the intended use of UH in forecasting. Occurrences of composite reflectivity $\geq$ 40 dBZ and of tornado and severe hail reports are converted to binary neighborhood fields using the same 40-km square neighborhoods as the NMEPs. These verification fields are not smoothed.

A version of the fractions skill score (FSS; Roberts and Lean 2008) is used for objective scoring of the probabilistic forecast fields. Roberts and Lean (2008) define the FSS for neighborhood length n as

$$\text{FSS}_{(n)} = 1 - \frac{\text{MSE}_{(n)}}{\text{MSE}_{(n)\text{ref}}}. \quad (2)$$

Here $\text{MSE}_{(n)}$ is the mean square error, over an $N_x \times N_y$ grid, of forecast fractional coverage (F) versus observed fractional coverage (O) of neighborhoods of length n:

$$\text{MSE}_{(n)} = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} [O_{(n)i,j} - F_{(n)i,j}]^2. \quad (3)$$

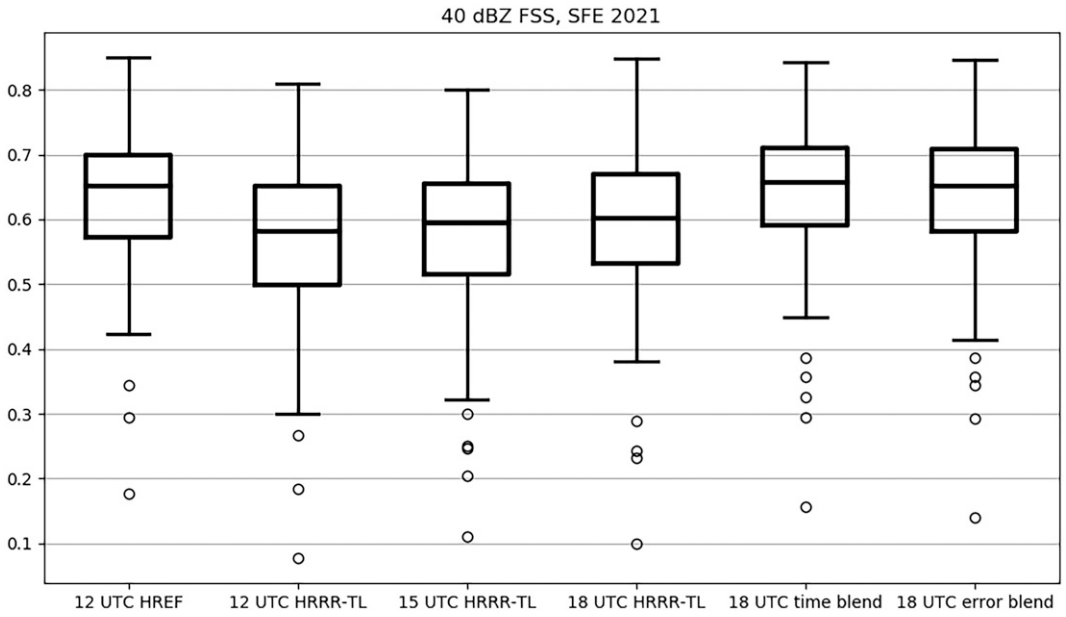

FIG. 1. Distributions of FSS for 40-dBZ reflectivity forecasts.

$\mathrm{MSE}_{(n)\mathrm{ref}}$ in Eq. (2) is the $\mathrm{MSE}_{(n)}$ of a "reference forecast" with the least possible skill:

$$\mathrm{MSE}_{(n)\mathrm{ref}} = \frac{1}{N_x N_y}\left[\sum_{i=1}^{N_x}\sum_{j=1}^{N_y} O_{(n)i,j}^2 - \sum_{i=1}^{N_x}\sum_{j=1}^{N_y} F_{(n)i,j}^2\right]. \quad (4)$$

FSS ranges from 0 to 1, where 1 represents a perfect forecast. However, in the above definition, both forecast and observed fields are binary. SPC uses a slightly different formulation of FSS to accommodate probabilistic forecasts without imposing a binary threshold. This FSS replaces the binary forecast field's *fractional coverage* over the neighborhood $F_{(n)i,j}$ in Eqs. (3) and (4) with the *maximum forecast NMEP* at any gridpoint in the neighborhood. The observed field from which fractional coverage $O_{(n)i,j}$ is calculated remains binary, but in this study is also a neighborhood field itself—i.e., equal to 1 if the event occurred anywhere within the neighborhood—which may further differ from some uses of FSS. The 40-km square neighborhoods are used for these FSS calculations to match the NMEPs described above. Hereafter, "FSS" refers to this probabilistic version, so reported FSS values should not be compared directly to others calculated for binary forecasts with the traditional formulation. To match the time windows on which SFE participants were asked to focus their ratings, FSS is calculated on each hour 2200–0300 UTC, for the instantaneous reflectivity at that hour and for the maximum UH over the preceding 4 h. This results in six FSSs, one on each hour in this window, in each of 23 events, so that each ensemble or blend receives a total of 138 scores for each field.

d. SFE events and participant ratings

The 2021 SFE ran from 3 May to 4 June, as detailed in Clark et al. (2021). Over 130 forecasters, researchers, and developers throughout the meteorological community used, discussed, and evaluated a wide range of short-term tools for forecasting severe thunderstorms, including as many as 94 unique CAMs. Regional domains of interest were selected before each day's forecasting exercises and covered the U.S. Great Plains, Southeast, and mid-Atlantic at various times. Since SFE domains were defined for all convective days Monday–Friday, these are the days for which objective verification is done in this study. Weekends are omitted. Each morning, participants submitted surveys rating the previous day's CAM guidance on the chosen domain. Next-day ratings for Friday events are not available because no SFE activities were held on weekends, but since domains were still defined, Fridays are included in FSS distributions, so that the subjective scoring covers 20 of the 23 cases objectively scored. The relevant section of the survey asked participants to rate the 1200 UTC HREF, the 1200, 1500, and 1800 UTC HRRR-TL, and the two 1800 UTC blends on a subjective 1–10 scale based solely on UH and reflectivity fields compared to verification overlays. The evaluation tools that participants used can be found at https://hwt.nssl.noaa.gov/sfe_viewer/2021/model_comparisons and an example is presented in section 3 below. In all, 154 complete surveys were submitted for this group of CAM products, covering 20 events, so that there is a sample of at least several participants' ratings for each event. Participants were also given space for open-ended comments, particularly about the 1800 UTC blends.

3. Results

a. Objective verification

For the forecasts of 40-dBZ reflectivity, distributions of FSS (Fig. 1) indicate that the 1200 UTC HREF outperformed all three HRRR-TLs. The much more diverse HREF

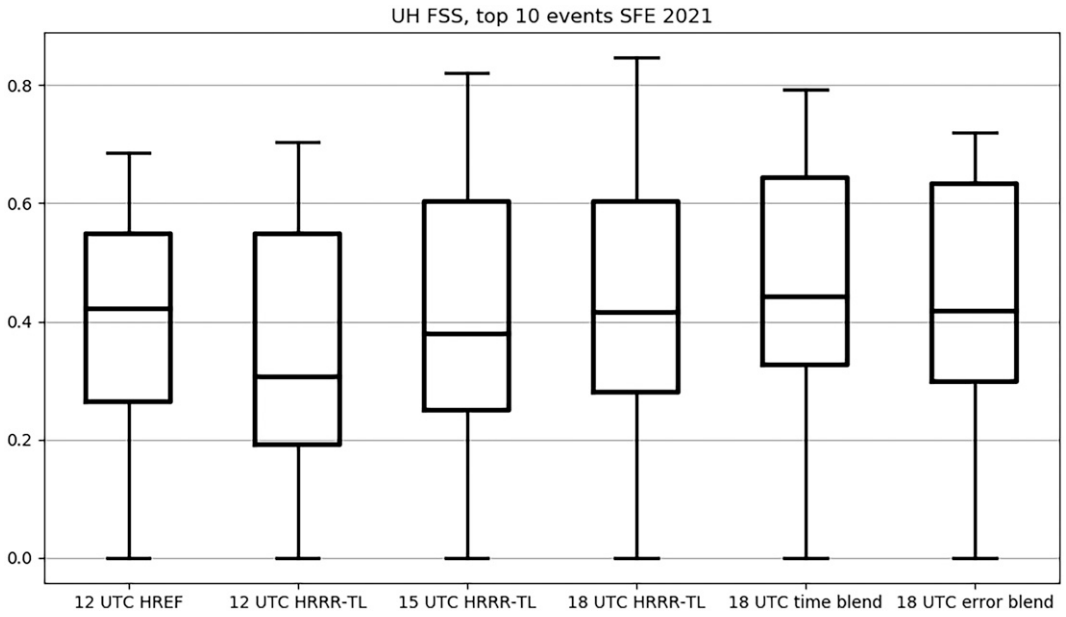

FIG. 2. Distributions of FSS for UH forecasts in the 10 SFE events with the most tornado and hail reports.

outperformed the HRRR-TL at the same 1200 UTC initialization time, as expected, but it is somewhat surprising that this remained the case as late as 1800 UTC. The two 1800 UTC blends improved on the 1800 UTC HRRR-TL, but neither was clearly preferable to the HREF. Despite such different methodologies, the two blends' score distributions were very similar.

Beyond the distributions of scores over all events, it is also relevant how consistently the ensembles performed head-to-head in each event. Table 1 shows how frequently the other reflectivity guidance scored higher or lower than the 1200 UTC HREF. These pairwise comparisons are even less favorable for the 1200 and 1500 UTC HRRR-TL than the score distributions might suggest. This has clear implications for forecasters in the midmorning–early afternoon (central time) period when both the 1200 UTC HREF and more recent hourly HRRR runs are available.

Distributions of FSS for the UH guidance over all cases (not shown) had large variability and numerous scores of zero, many of which arose from domains and time windows in which no tornadoes or severe hail were reported or no UH above the threshold was forecast. While these marginal cases are important, the resulting scores made the distributions somewhat harder to interpret, so the UH scores presented here (Fig. 2) are for the 10 SFE events with the most tornado and hail reports. (This shifts all the distributions upward but does not change relative model performance from the full results.) The overall pattern is similar to the reflectivity scoring: the 1200 UTC HREF slightly outperformed the 1200 and 1500 UTC HRRR-TLs and was roughly matched by the 1800 UTC blends. The main difference from the reflectivity verification is that the 1800 UTC HRRR-TL UH scores fell closer to the 1200 UTC HREF scores.

b. *Subjective evaluation and selected examples*

SFE participants' ratings of the six ensembles and blends (Fig. 3) match the reflectivity FSS distributions (Fig. 1) remarkably well—so well it should be noted that participants did not have access to these products' skill scores when rating them. Table 2 is analogous to Table 1, showing how frequently the other guidance was rated higher or lower than (or equal to, since only whole numbers were allowed) the 1200 UTC HREF. As in the objective scoring, HREF outscored all three HRRR-TLs most of the time, though the preference for HREF over the 1200 and 1500 UTC HRRR-TL was not as extreme as in the FSS comparisons. Participants' slight preference for the error-based blend over the time-based blend differs from the objective scoring results, but the difference is mainly between ratings equal to HREF's and less than HREF's; the blends outperformed HREF at similarly low rates (34.4% and 35.0%).

Manual inspection of individual forecasts and SFE participant surveys suggests two main reasons for these differences in performance: the greater dispersion of HREF members and HRRRv4's possible bias related to convection initiation (CI). Per SPC's internal evaluation, HRRRv4 can be biased cool and dry near the surface, sometimes resulting in delayed or missed CI in scenarios requiring strong heating and mixing. SFE participants alluded to HRRR-TL's underdispersion and, occasionally, to delayed or underforecast CI in weakly forced diurnal settings:

- "The time-lagged HRRR options were all focused in on one solution that was only partially right. The extra options from the HREF increased the spread and better captured all the reports." (13 May; Fig. 4)
- "18Z HREF/HRRR-TL blends compared similar to the 12Z HREF but was slightly slow on initial CI." (13 May)

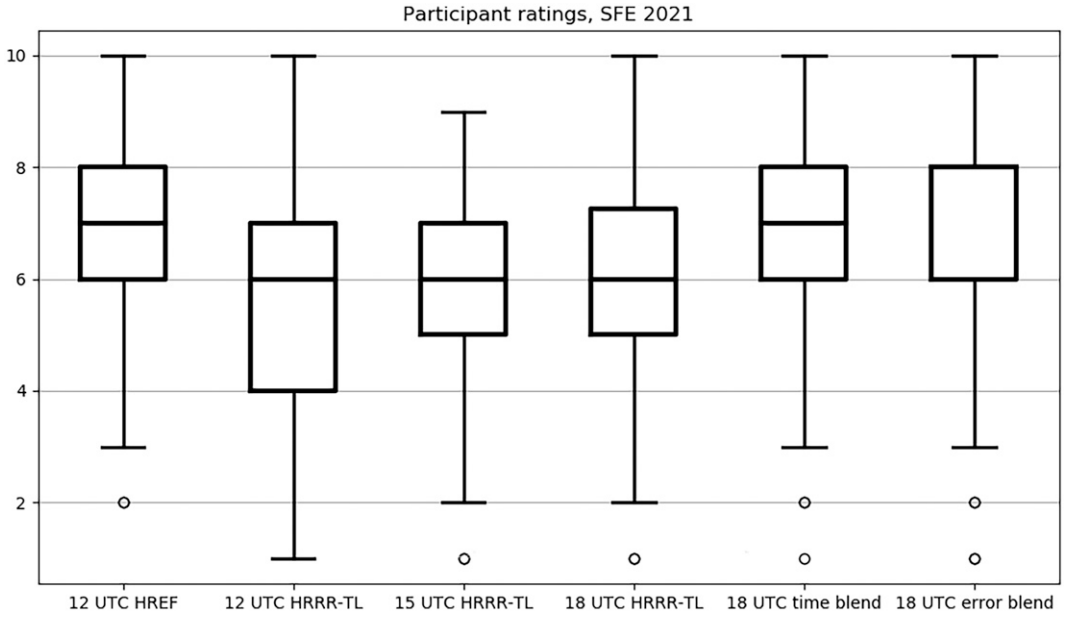

FIG. 3. Distributions of SFE participants' subjective ratings of overall forecasts. (The median has the same value as the third quartile for the 1800 UTC error blend at far right.)

- "They [the blends] did a decent job overall. The HRRR missed the Colorado border cell completely however." (13 May)
- "12Z HREF was a hard bar to beat. 18Z HRRR-TL just didn't capture the activity in KS at all. Very low probabilities. Although maybe it was a low probability event." (26 May)
- "Blends slightly outperformed HRRR-TL by capturing convection along the SE TX coast where HRRR didn't have reflectivity probs … " (11 May)
- "HRRR-TL missed some of the later development in W TX/TX PH, but blending in HRRR probs slightly reduced some false alarm areas compared to HREF." (27 May)
- "This is kind of splitting hairs in such a marginal event, but I think the HRRR-TL ensembles all suffered from some degree of over-forecast in southwestern GA later in the day. HREF remained the best guidance." (12 May)

While the above responses are reasonably representative of the distributions of FSS and ratings, other comments also illustrate the case-to-case variability. For example, most respondents preferred at least one version of the HRRR-TL to the HREF for the 17 May 2021 event centered in Texas:

- "All of the time lags performed better than the 12z HREF. The blend simulation performed better than the 12z HREF by having the area of severe coverage further west. However, the 18z HRRR-TL was able to have higher confidence in the correct spots for severe reports than the blends." (17 May)
- "The blends were better than the HREF alone, particularly for the supercell near Lubbock. The 18Z HRRR-TL was better than the blends. Overall, 15Z HRRR-TL was best because it nailed the Lubbock supercell, extended higher probs toward Wichita Falls, and nailed the region west of Amarillo." (17 May)

Finally, while successive runs of the HRRR-TL did tend to become more skillful and more highly rated as lead time decreased, this was not true in all individual events. It was somewhat common for forecasters to rate a 1500 or 1800 UTC HRRR-TL lower than an earlier initialization. Some commented on this:

- "Interesting to see that the HRRR-TL does better for the older runs, though we sometimes do see this in HRRR runs." (19 May)
- "15Z HRRR-TL has very good coverage of storms in SE WY and NE CO, better than even the 18Z HRRR-TL." (20 May)

TABLE 2. Pairwise comparisons of HREF ratings to other guidance.

Guidance	Rating < HREF	Rating = HREF	Rating > HREF
1200 UTC HRRR-TL	62.6%	19.6%	17.8%
1500 UTC HRRR-TL	63.2%	16.6%	20.2%
1800 UTC HRRR-TL	57.7%	18.4%	23.9%
1800 UTC time-based blend	35.6%	29.4%	35.0%
1800 UTC error-based blend	27.9%	37.7%	34.4%

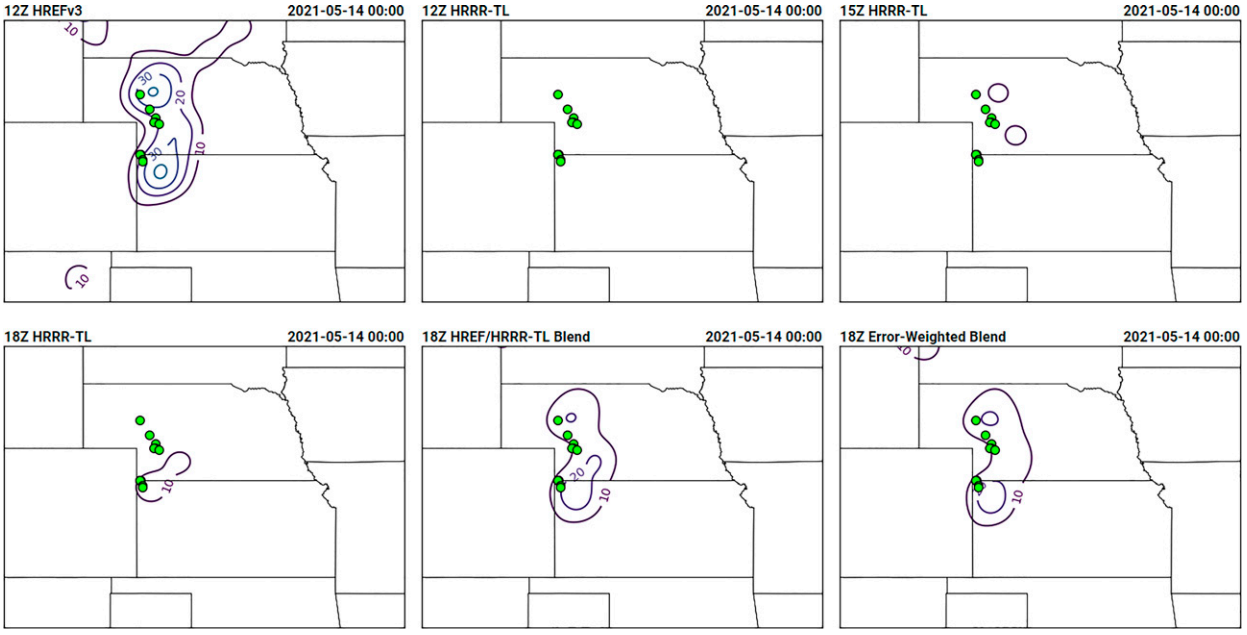

FIG. 4. Model evaluation viewer used by SFE participants displaying the 13 May 2021 event. Contours are NMEPs of UH, labeled in percentages, and green circles are severe hail reports, both over the 4-h period ending 0000 UTC 14 May.

- "18Z HRRR-TL was by far the poorest performer of the set." (24 May)

c. *The role of diurnal CI in HRRRv4 performance*

It is operationally important to clarify both skill scores and participant comments involving the recent upgrade to HRRRv4. Did the HRRR-TL generally fall short of HREFv3 because of its lack of model diversity, or because of a pervasive bias in handling certain CI scenarios? If HRRRv4 forecasts of diurnal CI were systematically poor, this should appear as decreased skill in the afternoon and early evening hours as probability of detection (POD) of deep convection is temporarily reduced. This would have considerable implications for HRRRv4's utility in short-term convective forecasting. Daily time series of contingency table–based statistics were calculated on all primary SFE domains (i.e., the same set of 23 domains used in FSS calculations above) at each hour for the instantaneous 40-dBZ reflectivity forecasts of all HREF members initialized at 1200 UTC, including HRRR, as well as the 1800 UTC HRRR. To emphasize presence or absence of convection rather than its precise placement, a large 120-km neighborhood was used. MRMS was again used as truth. This method reveals no systematic drop in POD (Fig. 5a) during peak diurnal CI hours that would be consistent with late or underforecast CI, nor in the critical success index (CSI; Schaefer 1990; Fig. 5c). These time series also show that the HRRR's best performance relative to other models fell outside the 2200–0300 UTC evaluation window. While the 1200 UTC guidance was tightly clustered in CSI during the day, the 1200 UTC HRRR gained higher mean CSI than the others by around 0600 UTC. The gap in CSI between the 1800 UTC HRRR and the 1200 UTC non-HRRR models also widened on average around 0300 UTC. The benefit of assimilating newer observations is evident in the 1800 UTC HRRR's higher mean CSI than all 1200 UTC models over the entire 2200–0300 UTC evaluation window. That the probabilistic 1800 UTC HRRR-TL still did not consistently outperform the 1200 UTC HREF despite this advantage, as shown in the previous sections, reinforces the role of the HRRR-TL's underdispersion in that finding.

Finally, to eliminate the possibility that a HRRRv4 bias in a few weakly forced, strongly heated CI regimes was simply outweighed by more common scenarios, subdomains were manually defined for SFE events qualitatively fitting that description. Weather Prediction Center surface analyses were used to identify regions in or near an SFE daily domain where initiation of severe storms occurred between 1800 and 0000 UTC in the absence of an analyzed cold front or warm front. This yielded a sample of 12 subdomains with primary CI mechanisms that included drylines, lee troughs, stationary fronts, remnant outflow boundaries, and orography. All of these subdomains were located mostly or entirely within the Great Plains and contained multiple reports of severe weather. The resulting time series (Fig. 6) still are not consistent with systematically late or underforecast diurnal CI in HRRRv4. While the 1200 and 1800 UTC HRRR did have the lowest mean POD (Fig. 6a) around 2200 UTC, the margin was very small, and both HRRRs' mean CSI (Fig. 6c) remained tightly clustered with multiple other models' CSI from 1900 to 0000 UTC. A series of performance diagrams (Roebber 2009) summarizes the progression of models' skill through the relevant 1800–0000 UTC period and illustrates the variability in model performance across events (Fig. 7).

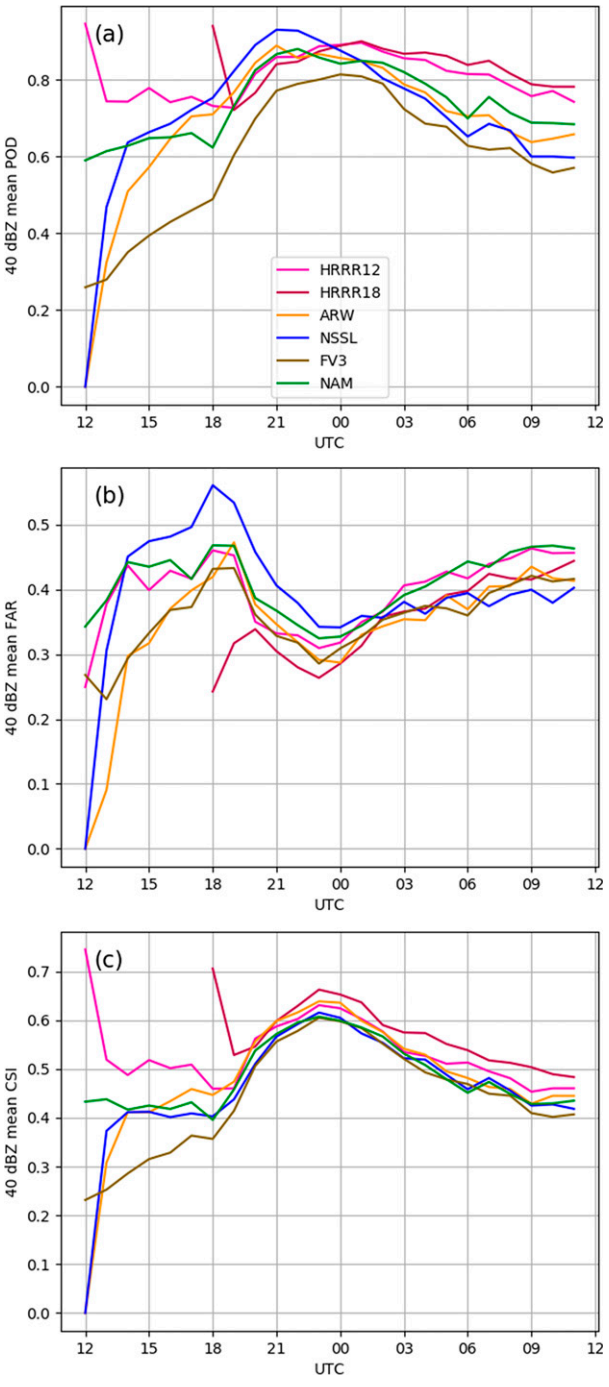

FIG. 5. Hourly time series of mean (a) POD, (b) false alarm ratio (FAR), and (c) CSI for all 1200 UTC CAM initializations and the 1800 UTC HRRR, over all SFE events.

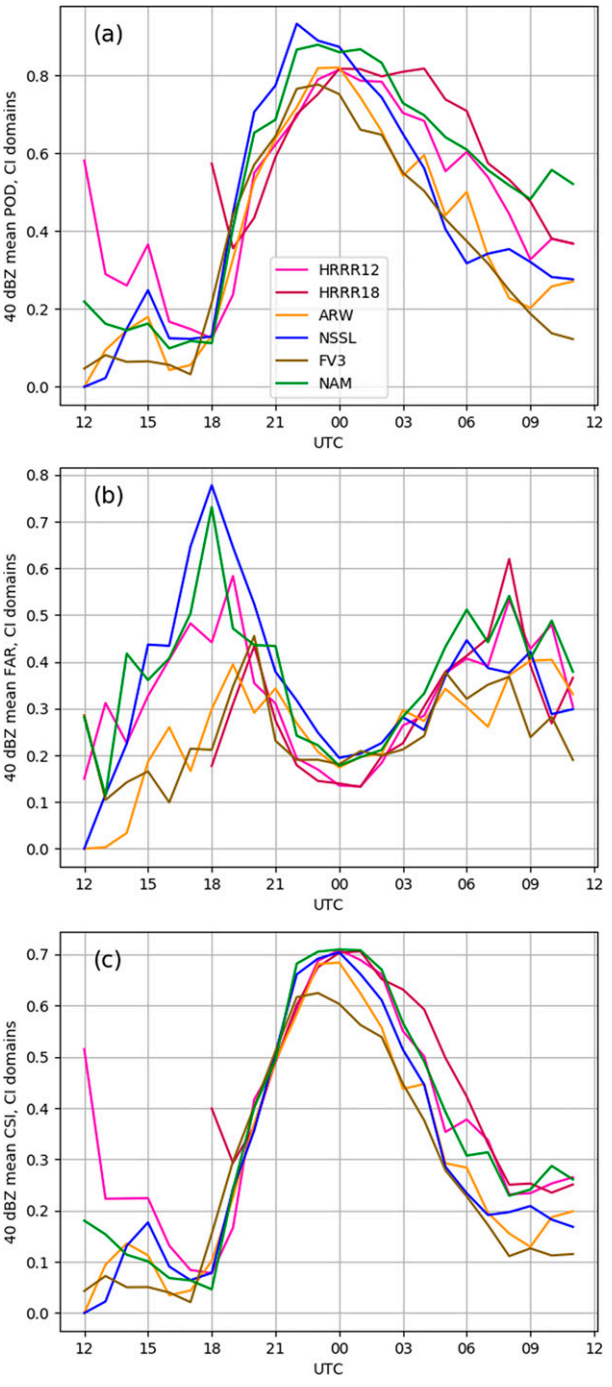

FIG. 6. As in Fig. 5, but over the 12 CI subdomains detailed in section 3c.

Low skill on the diurnal CI subdomains at 1800 UTC (Fig. 7b) owes to minimal deep convection ongoing at that time. As newly initiated convection became more prevalent at 2100 UTC (Fig. 7d), HRRRv4 POD and FAR were not practically different from those of the WRF-ARW and FV3 members. All models attained high skill with similar POD and FAR around 0000 UTC, over both the full SFE dataset and the 12 CI subdomains (Figs. 7e,f).

Overall, this analysis targeted to the HRRRv4's hypothesized weakness still places it well within the range of other HREF members' performance. This does not definitively rule out any bias at all in forecasting the most subtly forced individual storms. There are well-established small biases in

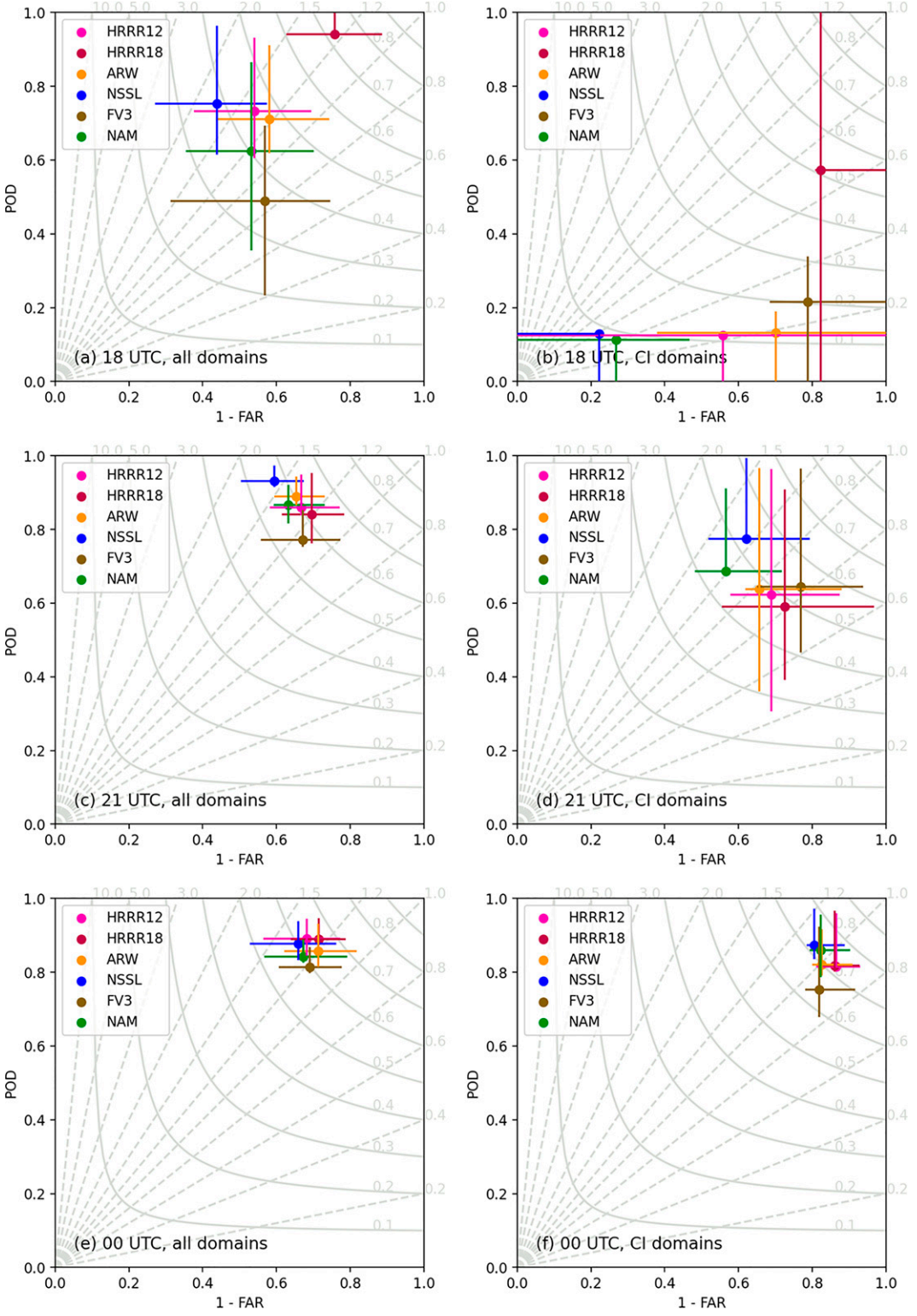

FIG. 7. Performance diagrams for neighborhood 40-dBZ reflectivity forecasts from the 1200 UTC CAM initializations and the 1800 UTC HRRR. Points represent model means over the indicated set of events and domains; error bars extend to the first and third quartiles in both dimensions.

HRRRv4's surface fields (Alexander et al. 2020) that likely do influence CI in some way. But this subset of cases rules out any CI bias large enough to affect relative skill, leaving HREF's model diversity as its apparent advantage over the HRRR-TL in both objective and subjective verification. It also implies that HRRRv4 can be used operationally in these settings with just as much confidence as any other single CAM.

4. Conclusions

Regarding the questions in section 1:

1) The 1200 UTC HREFv3's Day 1 forecasts of 40-dBZ composite reflectivity and 75 $m^2 s^{-2}$ 2–5-km UH are more skillful than the 1200 and 1500 UTC HRRR-TL; the 1800 UTC HRRR-TL's skill is similar to that of the 1200 UTC HREF for UH, but lower than HREF for reflectivity.
2) Blended HREF and HRRR-TL output around 1800 UTC does not consistently outperform the 1200 UTC HREF, though it is more skillful than the 1800 UTC HRRR-TL alone.
3) SFE participants' subjective ratings of these forecasts largely agree with FSS, especially that of the 40-dBZ reflectivity field, and support the same conclusions.

The 1200 UTC HREFv3 proved very difficult to outperform in springtime convective events. This remained true even with up to 6 h of additional information in the case of the 1800 UTC HRRR-TL, and even when weighting all available guidance by how well it represented observed conditions at 1800 UTC. HREF's superior diversity (rather than a suggested bias in HRRRv4 CI forecasts, which was not found) seems to be the dominant reason. Practically, this cautions against quickly discarding the morning's HREF guidance in favor of newer HRRR solutions. Forecasters might reasonably expect the 1800 UTC HRRR-TL to handle the evening maximum in severe convective storms far more accurately than the 1200 UTC HREF, which has twice the lead time, but this is not generally the case for springtime thunderstorms in the United States. The high skill of the HREFv3 is also consistent with findings from previous SFEs in which multiple experimental ensemble configurations fell short of HREFv2.

The primary limitation in generalizing these results is the nature of the SFE sample (20 events subjectively rated and 23 events objectively scored). Although the daily domains covered most of the central and eastern United States at various times in the SFE period, other seasonal regimes could not be sampled, particularly cool-season severe weather in the Southeast. It is uncertain whether relative model performance would be the same in such events. Second, the 2200–0300 UTC evaluation period presented to SFE participants ended before the overnight climatological peak of mesoscale convective systems in the central United States, for which relative model performance could plausibly be somewhat different than for other convective modes. However, the length of hourly HRRR forecasts (18 h for all initializations except 0000, 0600, 1200, and 1800 UTC) makes this limitation unavoidable for the HRRR-TL. Future work should clarify CAM performance outside of the narrow season and diurnal period sampled by the SFE.

Acknowledgments. The lead author was supported by NOAA–University of Oklahoma Cooperative Agreement NA16OAR4320115. The efforts of participants and forecasters from across the weather enterprise made the 2021 SFE possible in a virtual setting. The authors also appreciate the help of Adam Clark (NSSL), Brett Roberts (CIWRO), and Burkely Gallo (CIWRO) in preparing these CAM products for evaluation during the SFE.

Data availability statement. The model data (HREFv3 and HRRRv4) and MRMS radar data for the SFE period are archived internally at the National Severe Storms Laboratory (NSSL) and available upon request. Plotted HREF products are also viewable at https://www.spc.noaa.gov/exper/href/. Storm reports used for UH verification are publicly available at https://www.spc.noaa.gov/climo/online.

REFERENCES

Alexander, C., and Coauthors, 2020: Rapid Refresh (RAP) and High-Resolution Rapid Refresh (HRRR) model development. *30th Conf. on Weather Analysis and Forecasting (WAF)/26th Conf. on Numerical Weather Prediction*, Boston, MA, Amer. Meteor. Soc., 8A.1, https://ams.confex.com/ams/2020Annual/webprogram/Paper370205.html.

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.

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

——, and Coauthors, 2021: Spring Forecasting Experiment 2021: Preliminary findings and results. NOAA Hazardous Weather Testbed, NOAA, 86 pp., https://hwt.nssl.noaa.gov/sfe/2021/docs/HWT_SFE_2021_Prelim_Findings_FINAL.pdf.

De Pondeca, M. S. F. V., and Coauthors, 2011: The real-time mesoscale analysis at NOAA's National Centers for Environmental Prediction: Current status and development. *Wea. Forecasting*, **26**, 593–612, https://doi.org/10.1175/WAF-D-10-05037.1.

Gallo, B. T., and Coauthors, 2021: Exploring convection-allowing model evaluation strategies for severe local storms using the Finite-Volume Cubed-Sphere (FV3) model core. *Wea. Forecasting*, **36**, 3–19, https://doi.org/10.1175/WAF-D-20-0090.1.

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.

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

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of

convective events. *Mon. Wea. Rev.*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1.

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.

Smith, T. L., S. G. Benjamin, J. M. Brown, S. Weygandt, T. Smirnova, and B. Schwartz, 2008: Convection forecasts from the hourly updated, 3-km High Resolution Rapid Refresh (HRRR) model. *24th Conf. on Severe Local Storms*, Savannah, GA, Amer. Meteor. Soc., 11.1., https://ams.confex.com/ams/24SLS/techprogram/paper_142055.htm.

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.