# INITIAL DEVELOPMENT AND TESTING OF A CONVECTION-ALLOWING MODEL SCORECARD

Burkely T. Gallo, Christina P. Kalb, John Halley Gotway, Henry H. Fisher, Brett Roberts, Israel L. Jirak, Adam J. Clark, Curtis Alexander, and Tara L. Jensen

A scorecard summary diagram allows for at-a-glance visualization and comparison of convection-allowing model performance across multiple metrics and fields.

Since scientists first began modeling the Earth system, a need for verifying the subsequent forecasts has existed. Brier and Allen (1951) highlight three main reasons for forecast verification, broadly categorized under the labels of *scientific*, *administrative*, and *economic*. At its best, formal verification can identify areas for improvement in forecast models (scientific), objectively judge how changes in the models affects forecast quality (administrative), and provide the best set of metrics for different users (economic; Jolliffe and Stephenson 2011). Historical overviews of numerical weather prediction (NWP) show that while the progression of NWP is measured by objective statistics, the selection of appropriate statistics necessarily incorporates subjectivity (Shuman 1989). To restrain the impact of the subjective choices, Anthes (1983) called for a set of agreed-upon verification metrics to assess forecast quality and determine the impact of changes.

Questions of how best to evaluate forecasts continue to this day. Operational implementation of new

**AFFILIATIONS:** Gallo—Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, and NOAA/NWS/NCEP Storm Prediction Center, Norman, Oklahoma; Kalb, Halley Gotway, and Jensen—National Center for Atmospheric Research, and Developmental Testbed Center, Boulder, Colorado; Fisher—National Center for Atmospheric Research, Boulder, Colorado; Roberts—Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, and NOAA/NWS/NCEP Storm Prediction Center, and NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma; Jirak—NOAA/NWS/NCEP Storm Prediction Center, Norman, Oklahoma; Clark—NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma; Alexander—NOAA/Earth System Research Laboratory/Global Systems Division, Boulder, Colorado

**CORRESPONDING AUTHOR:** Burkely T. Gallo, burkely.twiest@noaa.gov

*The abstract for this article can be found in this issue, following the table of contents.*
DOI:10.1175/BAMS-D-18-0218.1

A supplement to this article is available online (10.1175/BAMS-D-18-0218.2)

In final form 18 July 2019

guidance occurs only after a series of tests and a thorough evaluation period, to examine the strengths and weaknesses of forecasts compared to observations and previous model iterations. This framework satisfies all reasons for forecast verification put forth by Brier and Allen (1951), but choosing which statistics to fulfill the framework remains the subject of discussion. Given the large complexity and dimensionality of most atmospheric forecast problems (Murphy 1991), care must be taken when selecting the verification information considered during the implementation of new systems.

When choosing verification metrics with the most utility and relevance, the model grid spacing and phenomena of interest are of primary importance. Global models with resolutions on the scale of tens of kilometers that are tasked with identifying the placement and magnitude of synoptic-scale features use metrics such as the anomaly correlation coefficient (ACC; Hollingsworth et al. 1980), root-mean-square error (RMSE), and equitable threat score (ETS; Gilbert 1884; Schaefer 1990). These scores summarize broad-scale, synoptic aspects of the forecast that indicate skill in short and medium ranges, evaluating forecast aspects such as the placement and intensity of high and low pressure systems. Convection-allowing models (CAMs), with typical grid spacing of ~3 km, instead primarily depict mesoscale and storm-scale features such as simulated reflectivity and convective mode. These finescale simulated features need not necessarily exactly collocate with the observed features to provide value to forecasters, and so different verification metrics allowing for some spatial and/or temporal displacement are required to determine the full value of the forecast. The neighborhood-based approach allows for displacement by recognizing model skill where forecast "yes" events may be close to the observed events (or within a "neighborhood") without necessarily overlapping. Metrics such as the critical success index (CSI; Schaefer 1990), area under the receiver operating curve (ROC area; Mason 1982), and fractions skill score (FSS; Roberts and Lean 2008) are often used in conjunction with a neighborhood-based approach during CAM verification (Schumacher and Clark 2014; Schwartz and Sobash 2017).

The difficulty with assessing a multitude of statistical scores is that often, optimizing one score will degrade another. For example, improving the ROC area can degrade the reliability of a forecast, or vice versa [as seen in Gallo et al. (2016) and Sobash et al. (2016b), respectively]. Alternately, improving the same score for one model field may reduce the same statistic in another field. For example, parallel runs performed when testing the upgrade of the Global Forecast System (GFS) to the Finite-Volume Cubed (FV3; Putman and Lin 2007; Harris and Lin 2013) GFS (FV3GFS) showed that the upgrade improved the northward QPF bias in the GFS, but worsened the low bias in instability and 2-m dewpoint fields (EMC Model Evaluation Group 2018a). Finally, improvements may occur solely at certain forecast hours, requiring metrics from multiple times and adding a dimension of needed information for a thorough forecast evaluation. Nuances and trade-offs that necessarily occur during model implementation may be inadvertently overlooked in this myriad of verification metrics, despite being relevant to one or more communities within the weather enterprise. By creating a summary visualization tool, this work hopes to show how large quantities of information can be displayed to model developers and the meteorological community as a whole, such that evidence-based decisions can be made when implementing new models.

To summarize the metrics and fields concerning model developers and end users, a scorecard is a useful visualization tool that can compare model systems at multiple field thresholds, statistics, time

## INTERPRETING THE SCORECARD

For quick and easy interpretation of the scorecard, levels of statistically significant differences on the CAM scorecard are distinguished using two primary means (Fig. SB1). First, the depth of the shading indicates the statistical significance; the darker the shading, the higher the level of statistical significance between the two models for a given field, valid time, and statistic. A square with no shading indicates no statistically significant difference. A difference at the 95% significance level has a lighter shading, and a difference at the 99% significance level has a darker shading. Second, the size of the arrow also indicates the statistical significance of the difference, with a smaller arrow indicating a difference at the 95% significance level and a larger arrow indicating a difference at the 99% significance level. The directionality of the arrow indicates which model is performing better at each square if statistical significance is reached. The scorecard has gone through multiple visualization iterations (an earlier visualization can be seen in Fig. 1) to improve visibility and comprehension for all users.

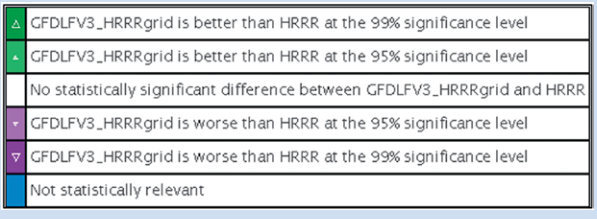


**Fig. SB1. A CAM scorecard legend indicating the degree of statistically significant difference between two model fields.**

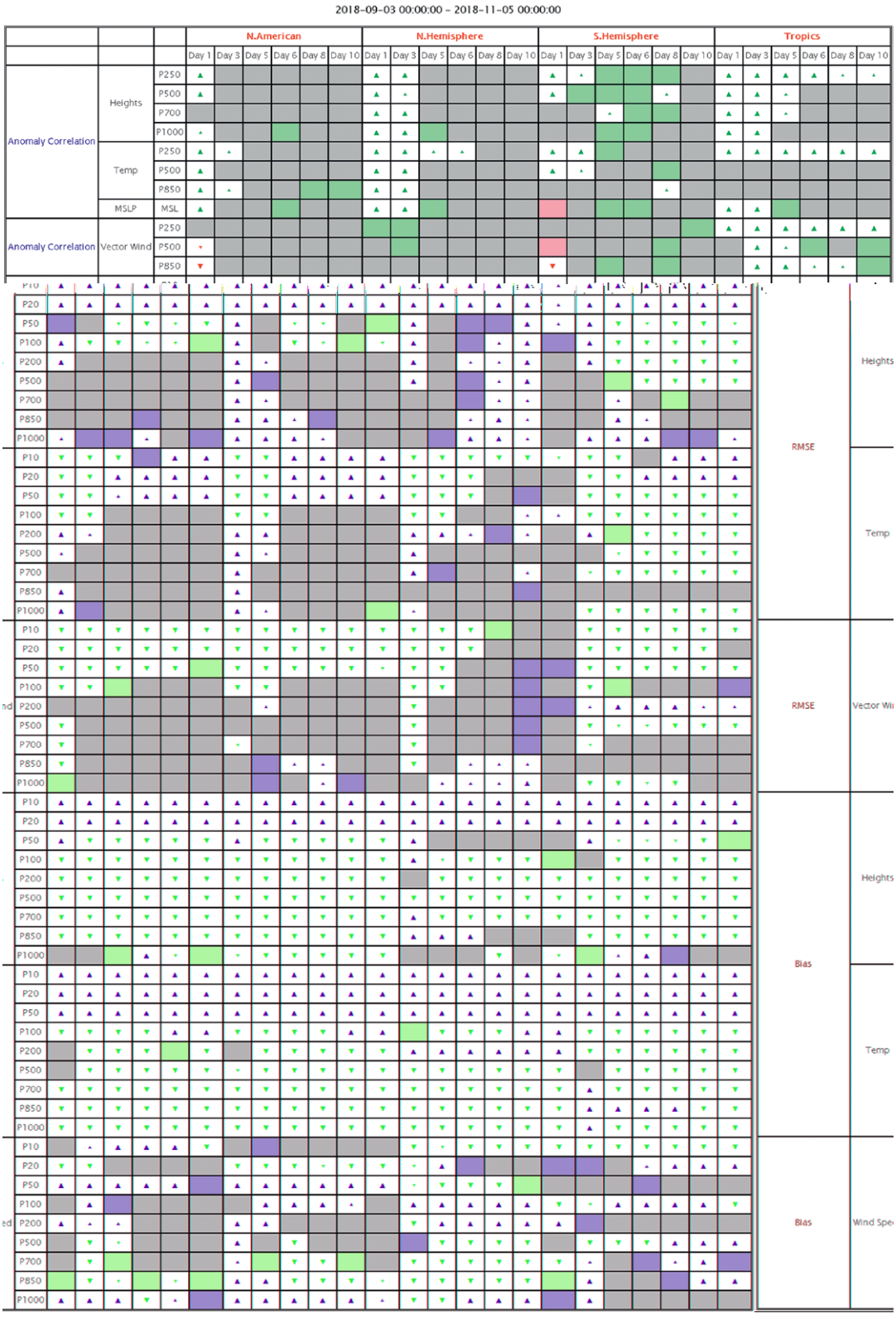


**Fig. 1. An example scorecard comparing multiple aspects of the operational GFS model as of 5 Nov 2018, and the new implementation of the GFS with the Finite Volume Cubed dynamical core (FVGFS), two convection-parameterizing models. Statistics are calculated from 3 Sep through 5 Nov 2018. Green colors and arrows indicate that the FVGFS performs better, while the red colors and arrows indicate that the GFS performs better. Fully shaded squares, small arrows, and large arrows respectively indicate a 95%, 99%, and 99.9% significance level.**

periods, and domains in one image (Fig. 1; see sidebar "Interpreting the scorecard"). Significantly better performance by one of the models compared to the other at the 95% significance level results in a shaded box. If a 99% significance level is reached, a colored arrow is displayed within the box. An abundance of one color or another across the scorecard indicates better performance by one modeling system, and displaying a square for each unique combination of domain, time period, metric, and threshold can reveal systemic differences. These systemic differences could then be examined in depth, in order to diagnose model deficiencies. For instance, if a new system has difficulty with nocturnal temperatures, that would become evident from the columns of the scorecard rather than potentially obscured by a summary metric evaluated over the entire forecast run. While the subjectivity of metric selection noted by Shuman (1989) remains, careful selection of fields, metrics, and domains of most value to key end users can optimize the scorecard to form an overall picture of model performance.

A recommendation to use scorecards for synthesizing the skill of a forecast system can be found in literature describing best practices for designing ensemble prediction systems (Sandgathe et al. 2011, 2013). Scorecards have previously compared upgrades to operational systems such as the Global Deterministic (EMC Model Evaluation Group 2018b; Buizza et al. 2018) and Ensemble Forecast System (Zhou et al. 2017), the impact of new data assimilation schemes (Kuhl et al. 2013), and aerosol impacts at the subseasonal time frame (Benedetti and Vitart 2018). These studies show the flexibility of the scorecard framework: different scorecards can be used for deterministic and ensemble forecasts, as well as encompassing metrics that concern different forecast interests (Kuhl et al. 2013). Extending the scorecard framework to determining the best operational implementations of CAMs requires consideration and planning, the first efforts toward which will be described here.

The process of determining appropriate fields, thresholds, and metrics took time and focused on problems of interest to the 2018 NOAA Hazardous Weather Testbed Spring Forecasting Experiment (SFE; Kain et al. 2003; Clark et al. 2012; Gallo et al. 2017). So, as with any verification study, we recommend that a clear scientific problem drive what the scorecard displays, allowing for a targeted approach to the decisions that go into the scorecard, which will be described in further detail below. For example, the focus of the SFE on forecasting severe convection required metrics indicating how well the model is forecasting heavy precipitation, high reflectivity cores, and a proxy for rotating storms, with the later addition of variables that determine favorable storm environments. Other applications will likely require different fields be displayed on their scorecards, and a broad community engagement can ensure that scorecards for future operational CAM implementation include relevant fields and metrics for a variety of users.

This article will discuss the initial development and implementation of a CAM scorecard specifically for the 2018 SFE, starting with the work on selecting initial CAM metrics, fields, and domains to evaluate. We will then describe aspects of the CAM scorecard and its formulation. Next, we turn to the 2018 SFE, discussing the real-time evaluation of the scorecard and lessons learned from this first implementation. Finally, discussion and future plans for the CAM scorecard will be covered, including expansion beyond the severe convective storms community.

**CAM VERIFICATION NEEDS.** To address CAM verification needs across the meteorological enterprise, two community-based working groups (established by NOAA) have combined their efforts. These are the CAM and verification and validation working groups, so the CAM scorecard lies at the intersection of their expertise. These working groups are assisting NOAA with developing a strategic implementation plan (SIP) for CAM verification as the United States transitions to a Unified Forecast System designed around the FV3 dynamical core. Developing unified metrics and verification strategies will enable critical evaluation of the Next-Generation Global Prediction System (NGGPS). Through their recommendations, modeling efforts will advance in conjunction with systemic and relevant evaluation to support evidence-based decision-making concerning the future Unified Forecast System.

To determine the most important metrics and fields for evaluating CAM performance across applications, the two working groups created a spreadsheet of 30 relevant forecast fields, which were later winnowed to 11 initial fields with applications ranging from aviation to air quality to winter weather (Table 1). Crucial details of the simulated fields such as vertical and temporal attributes, validation sources, potential stratifications, and needed statistical scores for both a deterministic and ensemble framework were considered. For each field, breakout groups at the Developmental Testbed Center (DTC) Community Unified Forecast System Test Plan and Metrics Workshop (Developmental Testbed Center 2018) assigned priority and readiness. This workshop

**Table 1. Critical CAM fields, metrics, and stratifications, as developed by the CAM and validation and verification working groups and agreed upon by participants in the DTC metrics workshop. Metrics listed here were assigned both a priority and readiness of 1 (out of 3). Note: 1 ft ≈ 0.305 m, 1 mi ≈ 1.6 km, and 1 in. = 2.54 cm.**

| Forecast field | Deterministic metrics | Deterministic stratifications | Ensemble metrics | Ensemble stratifications |
|---|---|---|---|---|
| Ceiling (column) | CSI, BIAS, FSS, POD, FAR, AUR, performance diagram | Forecast length (0–36 h), threshold (500, 1,000, 3,000, 5,000, 10,000 ft), domain (west and east CONUS) | Briar score, Briar skill score, reliability, sharpness, CRPS, CRPSS | Smoothing, probabilities (0%, 5%, 10%, …, 100%) |
| Visibility (surface) | CSI, BIAS, FSS, POD, FAR, AUR, performance diagram | Forecast length (0–36 h), threshold (0.5, 1.0, 3.0, 5.0, 10.0 mi), domain (west and east CONUS) | Briar score, Briar skill score, reliability, sharpness, CRPS, CRPSS | Smoothing, probabilities (0%, 5%, 10%, …, 100%) |
| Dewpoint (2 m) | RMSE, BIAS | Forecast length (0–36 h), diurnal (0000–2300 UTC), domain (west and east CONUS) | Spread–skill ratio, rank histogram | — |
| Specific humidity (column) | RMSE, BIAS | Forecast length (0–36 h), diurnal (0000, 1200 UTC) | Spread–skill ratio, rank histogram | — |
| Temperature (2 m) | RMSE, BIAS | Forecast length (0–36 h), diurnal (0000–2300 UTC), domain (west and east CONUS) | Spread–skill ratio, rank histogram | — |
| Temperature (column) | RMSE, BIAS | Forecast length (0–36 h), diurnal (0000, 1200 UTC) | Spread–skill ratio, rank histogram | — |
| Wind (10 m) | RMSE, BIAS | Forecast length (0–36 h), diurnal (0000–2300 UTC), domain (west and east CONUS) | Spread–skill ratio, rank histogram | — |
| Wind (column) | RMSE, BIAS | Forecast length (0–36 h), diurnal (0000, 1200 UTC) | Spread–skill ratio, rank histogram | — |
| Precipitation (surface) | CSI, BIAS, FSS, POD, FAR, AUR, performance diagram | Forecast length (0–36 h), threshold (0.01, 0.1, 0.25, 0.5, 1.0, 2.0, 3.0, 4.0, 6.0 in., percentiles), scale (3, 40 km), domain (west and east CONUS) | Briar score, Briar skill score, reliability, sharpness, CRPS, CRPSS | Neighborhoods (10, 20, 40, 80 km), smoothing, probabilities (0%, 5%, 10%, …, 100%) |
| Simulated reflectivity (composite) | CSI, BIAS, FSS, POD, FAR, AUR, performance diagram | Forecast length (0–36 h), threshold (25, 30, 35, 40, 45, 50 dBZ, percentiles), scale (3, 40 km), domain (west and east CONUS) | Briar score, Briar skill score, reliability, sharpness, CRPS, CRPSS | Neighborhoods (10, 20, 40, 80 km, ±1 h in time), smoothing, probabilities (0%, 5%, 10%, …, 100%) |
| Updraft helicity (2–5 km; 0–3 km AGL) | FSS, AUR | Forecast length (0–36 h), threshold (25, 50, 75 $m^2 s^{-2}$, percentiles), scale (3, 40 km), domain (west and east CONUS) | Briar score, Briar skill score, reliability, sharpness, CRPS, CRPSS | Time windows (4, 24 h), neighborhoods (80 km), smoothing (120 km sigma), probabilities (0%, 5%, 10%, …, 100%) |

took place in Silver Spring, Maryland, from 30 July to 1 August 2018, and included participants from different branches of NOAA, the National Aeronautics and Space Administration, universities, the U.S. Navy, the U.S. Air Force, the private sector, and international collaborators. It was through combining the priority and readiness that the initial thirty fields were narrowed.

Priority level was assigned based on the relevance of the forecast field to multiple applications; fields like temperature, precipitation, and simulated reflectivity were assigned 1 out of 3, indicating that their assessment is a key component of a future unified verification system for multiple end users. Other metrics were assigned 2 out of 3 if their importance was largely to one or two communities of interest (such as the importance of CAPE and CIN mainly being confined to forecasts of severe convective storms), indicating that those metrics are targeted for near-term implementation into a verification suite but not critical to an initial CAM scorecard effort. Finally, a field was assigned 3 out of 3 if the field had highly specific applications unrelated to most sensible weather forecasts, such as ozone.

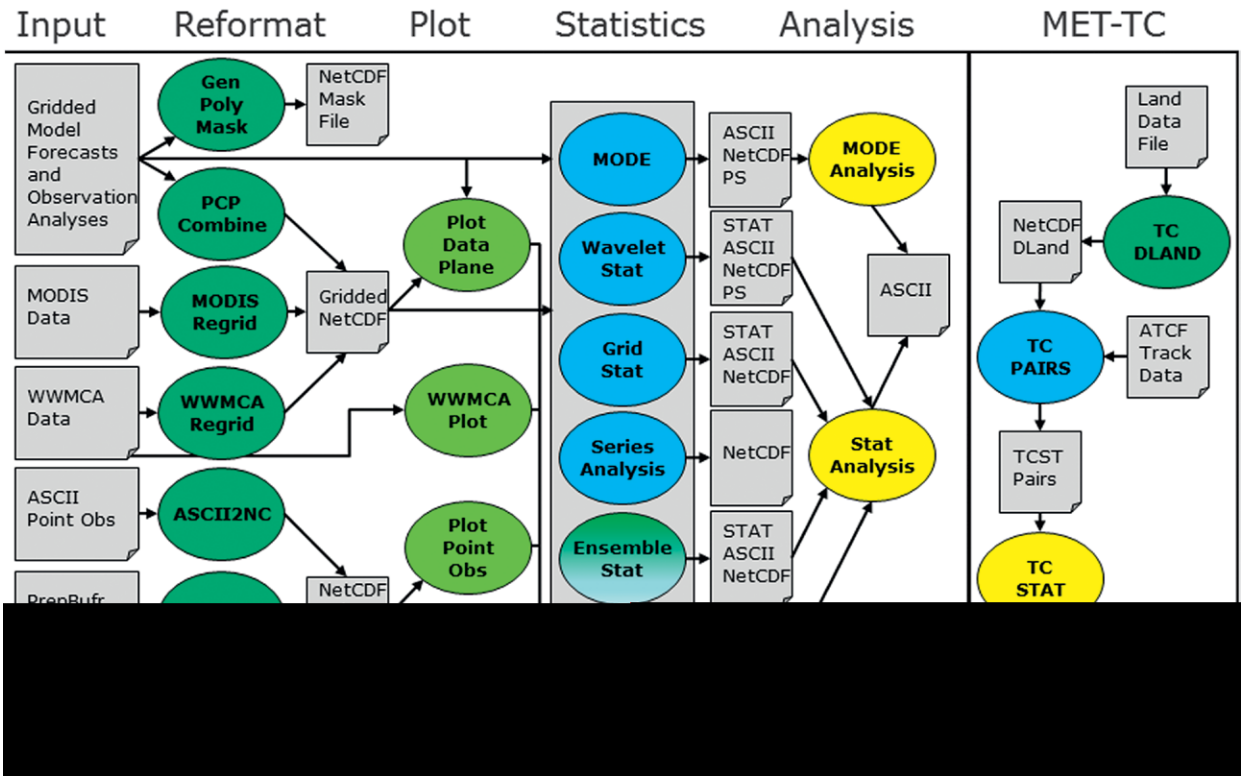


**Fig. 2. The MET verification package and suite of tools. Gray squares indicate input files or data. Dark green, light green, and blue ovals refer to reformatting, plotting, and statistics generating tools, respectively. Yellow ovals are "analysis tools," which combine results from the statistics tools. Ensemble stat is green and blue because it can generate both statistics for output and statistics (such as ensemble means) that can be inputs to other statistics tools.**

Readiness was assessed by the quality and consistency of available observations to verify the model fields, some of which do not have corresponding observations. Common fields such as accumulated precipitation and column temperature (i.e., the temperature throughout the vertical profile), which are verified using Stage IV precipitation observations (Lin 2011) and raob stations, respectively, were assigned a readiness of 1 out of 3, indicating that the observations were available and sufficient to support verification. A field had a readiness of 2 out of 3 if model or observational limitations prevented good comparisons. An example of readiness 2 would be the planetary boundary layer depth, which is not always computed consistently in models and observations. Finally, a readiness of 3 out of 3 was assigned if the workshop participants could not readily identify an observational network, such as particulate matter forecasts.

Another workshop outcome was the awareness of the myriad metrics and fields which are important to different aspects of the meteorological community. Developing a comprehensive scorecard that addresses all concerns for all applications may be impossible. As such, the workshop attendees also highlighted the need for multiple stakeholders to contribute to the selection of metrics, and raised the possibility of different scorecards for different applications.

**SCORECARD DEVELOPMENT.** The scorecard itself is generated using the Model Evaluation Tools (MET; Halley Gotway et al. 2018), a suite of statistical tools that combine to form a unified verification framework (Fig. 2; see sidebar "Sample evaluation metrics"). MET was initially developed to replicate the Environmental Modeling Center mesoscale verification system and computes over 85 different traditional statistics using both point and gridded datasets. Computation of confidence intervals is also included in the suite of tools. MET can ingest many data formats, including ASCII point and gridded observations, General Regularly-Distributed Information in Binary Form (GRIB), and Climate and Forecast-Compliant NetCDF (CF-NetCDF) files. It is designed to be flexible, and can evaluate ensembles, probabilities, and tropical cyclone tracks through different routines or combinations of routines. Object-based verification metrics are also available

in MET, complementing traditional, gridpoint-based metrics and providing a potential future direction for the CAM scorecard given the convective mode and other feature-based information provided by CAMs.

MET is at the core of METplus, a unified verification and diagnostic capability being developed for the Unified Forecast System (Adriaansen et al. 2018). METplus includes a suite of Python scripts to provide low-level automation for evaluation activities. In addition to calculating a multitude of verification metrics, METplus has a component tool, called METviewer, to visualize the output using the R statistics package (R Development Core Team 2019). METviewer is available to the community through download of the source code or a Docker container via GitHub. Within METviewer, a scorecard module generates the scorecards and calculates the *p* values for the statistical significance. The *p* value can be calculated either through a standard Student's *t* test that relaxes to a normal distribution with increasing sample size or through bootstrapping. The choice depends on whether the user wishes to compare the difference in scores to a known, theoretical distribution or to a resampled distribution. Users can specify the statistics, fields, regions, and time aggregations over which they want to compare the two modeling systems, assuming those statistics have already been calculated using the routines within the larger METplus framework. METplus provided a streamlined way to generate the CAM scorecard from a variety of model and observational data sources.

The CAM scorecard, as with its convection-parameterizing counterparts, emphasizes flexibility by allowing different users to select and examine scores relevant for their particular interests. This flexibility necessitates the ongoing discussion (begun at the DTC Metrics Workshop) of which metrics should be included.

## SAMPLE EVALUATION METRICS

**Table SB1. A sample 2 × 2 contingency table.**

| | | Observation | |
|---|---|---|---|
| | | **Yes** | **No** |
| **Forecast** | **Yes** | *a*: Hits | *b*: False alarms |
| | **No** | *c*: Misses | *d*: Correct nulls |

MET includes more than 85 different evaluation metrics. Common metrics are often based on a 2 × 2 contingency table containing four combinations of forecast and observation pairs (Table SB1). These metrics include

$$\text{Probability of detection (POD)} = \frac{\text{hits}}{\text{hits + misses}},$$

$$\text{Probability of false detection (POFD)} = \frac{\text{false alarms}}{\text{false alarms + correct nulls}},$$

$$\text{False alarm ratio (FAR)} = \frac{\text{false alarms}}{\text{hits + false alarms}},$$

$$\text{Critical success index (CSI)} = \frac{\text{hits}}{\text{hits + false alarms + misses}},$$

$$\text{Success ratio (SR)} = \frac{\text{hits}}{\text{hits + false alarms}}.$$

These metrics apply to binary forecasts and outcomes; an event is forecast or not, and occurs or does not. However, probabilistic forecasts can be evaluated using these metrics by choosing a probabilistic forecast threshold. Each value greater than that probability is then a "yes" forecast, and everything less than that probability is a "no" forecast. The receiver operating curve (ROC) is created via through such a process, by evaluating the POD and POFD at user-selected thresholds.

## TESTING THE FIRST SCORECARD IN SFE 2018.

*The 2018 Spring Forecasting Experiment.* The 2018 SFE took place from 30 April to 1 June 2018 in NOAA's Hazardous Weather Testbed (HWT). The goal of the annual SFE is to bring together researchers and forecasters from around the world to test cutting-edge numerical weather prediction and postprocessing methods in a real-time environment at the height of the spring severe convective weather season. Since 2007, SFE activities have included CAMs in their daily forecast and evaluation activities (Clark et al. 2012). Each day, participants make forecasts of severe convective weather (available at https://hwt.nssl.noaa.gov/sfe_viewer/2018/outlook_verification/) based on observations and experimental numerical weather prediction, as well as provide subjective evaluations of CAM forecast fields, postprocessing techniques, and their experimental forecasts from the previous day. Research community members attending the SFE test experimental forecast guidance and postprocessing tools, some of which they have contributed, as well as gain an understanding of the time pressures and limitations operational forecasters face on a daily basis. The operational forecasters attending the SFE learn about

innovative new numerical weather prediction tools, and see what improvements may become operational soon. They can also discuss current shortcomings of the guidance, highlighting areas for improvement to the model developers.

Given the nature of the SFE as a testing vehicle for CAMs and CAM postprocessing, it was an ideal venue to test the first CAM scorecard in real time. With most of the CAM datasets generated during the SFE, objective verification typically takes place post-experiment, when time permits a thorough examination of the large datasets generated. While a limited set of statistics have been available in previous years for some guidance (Melick et al. 2013), the CAM scorecard represented one of the largest real-time objective verification efforts in the SFE to date.

*CAM scorecard development preceding SFE 2018*. Prior to the 2018 SFE, meetings were held between the National Center for Atmospheric Research (NCAR)/DTC, the National Severe Storms Laboratory (NSSL), and the Storm Prediction Center (SPC) to determine which models would be evaluated using the CAM scorecard during the 2018 SFE. A subset of the Community Leveraged Unified Ensemble (CLUE; Clark et al. 2018), composed of three deterministic CAMs and two CAM ensembles, were chosen for evaluation. The deterministic members included the High-Resolution Rapid Refresh, version 3 (HRRRv3; Benjamin et al. 2016; Alexander et al. 2017), which became operational on 12 July 2018, as well as two experimental models that used the FV3 dynamical core and were implemented by NSSL and the Geophysical Fluid Dynamics Laboratory (GFDL). These deterministic models were chosen to reflect the U.S. commitment to moving toward a Unified Forecast System, as they included the current state-of-the-art operational CAM and two configurations of FV3 that represent preliminary tests of FV3 at convection-allowing scales. Similarly, the two CAM ensembles chosen were the High-Resolution Ensemble Forecast System, version 2 (HREFv2; Roberts et al. 2019), the current operational CAM ensemble, and the High-Resolution Rapid Refresh Ensemble system (HRRRE; Dowell et al. 2018). These ensembles have fundamentally different approaches to their configurations. One is based on an "ensemble of opportunity" (Jirak et al. 2012) and comprises members with multiple dynamical cores, initial conditions, physics parameterizations, as well as time-lagged members [HREFv2, containing the Weather Research and Forecasting Advanced Research WRF (WRF-ARW; Skamarock et al. 2008) and the Nonhydrostatic Multiscale Model on the B Grid (NMMB; Janjić and Gall 2012) cores]. The other ensemble (HRRRE) was traditionally designed, with a single dynamical core and physics parameterization suite, and includes ensemble spread generated through initial condition uncertainty from ensemble data assimilation. More detailed specifications for all of the CAMs and CAM ensembles evaluated herein can be found in the online supplementary material.

After selecting the models, the next step was to decide which model fields to compare, and what levels of statistical significance to highlight. Due to the complex nature of getting the real-time scorecard set up, a very small subset of fields was chosen for the initial scorecard, with expansion planned for the 2019 SFE. Initial fields were also focused on severe weather forecasting: simulated reflectivity (Fig. 3a), accumulated precipitation (Fig. 3b), 2–5-km updraft helicity (UH; Kain et al. 2008) (Fig. 3c), and a probabilistic surrogate severe field based on UH, following Sobash et al. (2011) (Fig. 3d). The surrogate severe field was created by gridding UH fields to a coarser, 80-km grid, and creating a binary yes–no field indicating whether a specific UH threshold is reached. Then, a Gaussian kernel was applied to the binary field to create smoothed probabilities. Simulated reflectivity and the surrogate severe field emphasized the "CAM" nature of the CAM scorecard, as a primary benefit of CAMs is their ability to simulate severe storm characteristics, such as convective mode, in ways that convection-parameterizing models cannot. For statistical significance levels, statistical significances of 95% and 99% were displayed on the scorecard, simplifying the graphic compared to some prior scorecards that had the 95%, 99%, and 99.9% statistical significance levels displayed (as in Fig. 1). The practical difference between a 99% difference in statistical significance and a 99.9% difference in statistical significance likely would be indiscernible to forecasters during a subjective evaluation, so only the 99% statistical significance threshold was retained.

Given the high resolution of the model forecasts, similarly high-resolution observations would ideally be used for verification. To verify the simulated reflectivity, Multi-Radar Multi-Sensor (MRMS; Smith et al. 2016) composite reflectivity data were used, and to verify the accumulated precipitation fields, Stage IV observations were used.[1] For the surrogate severe forecasts, local storm reports (LSRs) were smoothed

[1] Stage IV precipitation was used rather than MRMS to verify accumulated precipitation due to lower mean absolute error found for 24-h accumulated precipitation during the warm season from Stage IV than from MRMS (Zhang et al. 2016).

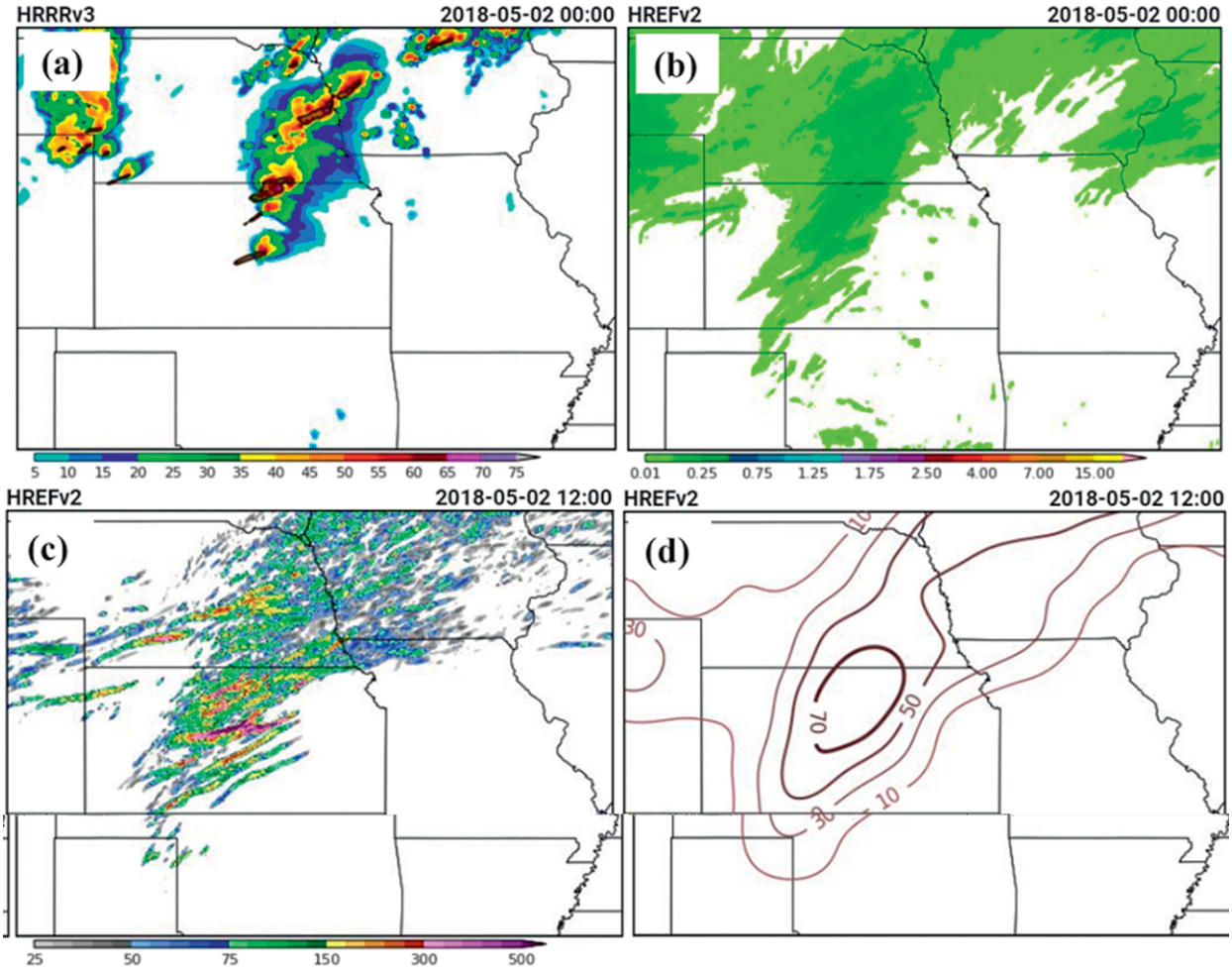


**Fig. 3. Fields evaluated using the CAM scorecard from SFE 2018. (a) Composite reflectivity (dBZ) with black contours showing areas where 2–5-km UH > 75 $m^2 s^{-2}$, (b) ensemble-mean 6-h QPF (in.), (c) ensemble-maximum 24-h UH ($m^2 s^{-2}$), and (d) probabilistic surrogate severe fields using a UH threshold of 75 $m^2 s^{-2}$, with contours occurring from 10% in intervals of 20%. Contours greater than or equal to 70% are bold.**

using a Gaussian kernel density estimation to create "practically perfect" probabilistic forecasts (Hitchens et al. 2013). When verifying the UH forecasts, a difficult problem arises. UH is calculated by integrating the updraft speed and vertical vorticity over a layer, and we do not currently have the observing capability to directly measure UH in storms. Traditionally, LSRs within a radius of a point have been used to verify UH-based forecasts (as in Sobash et al. 2011, 2016a; Loken et al. 2017), but these measurements have noted shortcomings regarding areas of low population density and overestimation of wind speeds by some types of observers (Doswell et al. 2005; Verbout et al. 2006; Trapp et al. 2006; Edwards et al. 2018). Therefore, we do not verify UH fields directly, but rely on the surrogate severe forecasts and corresponding LSRs for examining convective hazards.

Once the target fields were selected, we next selected verification thresholds. While the scorecard visualizes multiple thresholds of interest, having a row for each potential rainfall threshold or surrogate severe probability increment would likely be overwhelming without adding value for most users. Owing to the SFE's interest in severe convective weather, simulated reflectivity at the thresholds of 25–50 dBZ, in 5-dBZ increments, were chosen to evaluate the model performance at depicting features related to convection. Similarly, high thresholds of accumulated precipitation over both 3- and 1-h time windows were selected to examine the most intense storms. Accumulated precipitation ≥0.25, ≥0.50, ≥1.00, and ≥2.00 in. were evaluated for the 1- and 3-h time periods, similar to prior work defining extreme values of accumulated precipitation on the order of ≥1.00 in. for a 6-h period (Marsh et al. 2012). These precipitation and reflectivity fields were evaluated for the deterministic models, with expansion to the ensembles planned for later implementation. The

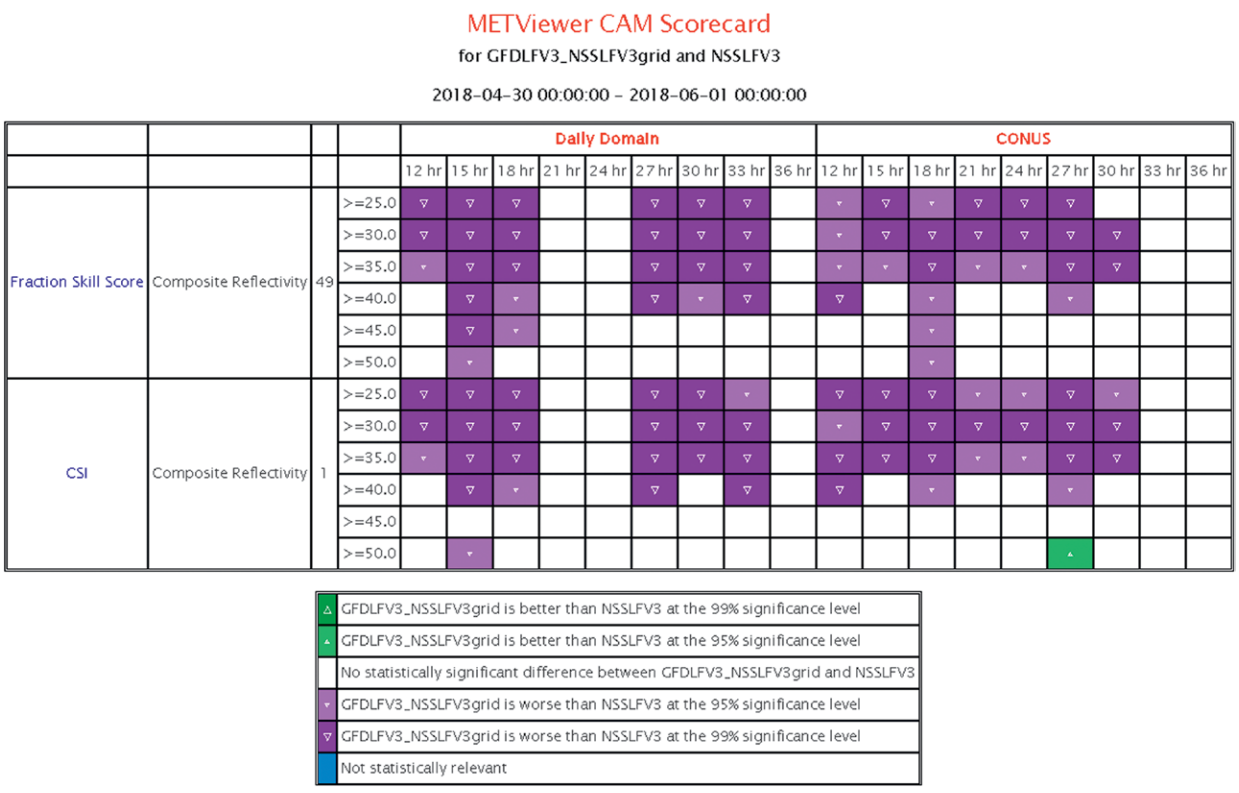


**Fig. 4. A scorecard comparing the FSS and CSI for the composite reflectivity every 3 h from forecast hours 12–36 in the GFDL-FV3 and the NSSL-FV3. The third column indicates the gridpoint neighborhood being tested. Numbers 1 and 49 refer to a one- and seven-gridpoint radius, respectively. Purple colors indicate that the NSSL-FV3 is performing better, while green colors indicate that the GFDL-FV3 is performing better.**

surrogate severe fields calculated for the deterministic (Sobash et al. 2011) and ensemble (Sobash et al. 2016a) guidance used four different UH thresholds to generate the probabilities; the lower the UH threshold, the more area covered by the probabilities for a given case. UH thresholds chosen were 50, 75, 100, and 125 $m^2 s^{-2}$, based on previous studies of UH and severe convective weather (Kain et al. 2008; Sobash et al. 2011, 2016b; Gallo et al. 2016; Loken et al. 2017). These thresholds were changed to percentiles post-SFE (the 75th–95th percentiles in increments of five percentiles), after it was determined that the different model climatologies prevented a useful comparison at specific thresholds. Once the probabilities were generated, they were evaluated at thresholds that the SPC currently uses in their operational convective outlooks: 2%, 5%, 10%, 15%, 30%, 45%, and 60%.

As the main problem of interest during the SFE was NWP performance in predicting fields relevant to severe convection, two domains were selected to verify each model; the full CONUS and a movable daily domain (8.72° latitude × 15° longitude) centered on the location where the most severe convective weather was expected. One final set of choices remained once the fields and thresholds were selected. Which verification metrics should be included? Again, a small initial set of metrics were chosen based on prior usage in the severe convective forecasting community (Sobash et al. 2011; Gallo et al. 2016; Sobash and Kain 2017; Dawson et al. 2017; Gallo et al. 2018; Adams-Selin et al. 2019). For the categorical fields, such as reflectivity, accumulated precipitation, and updraft helicity, FSS and CSI were calculated at each threshold. The FSS used three different circular neighborhoods to account for spatial displacement of features of interest, and test whether statistically significant results were dependent on the radius. The three radii chosen for initial testing were 3, 7, and 13 grid points, corresponding to 9, 21, and 39 km, respectively, corresponding to a quarter, half, and the distance defined by the SPC's probabilistic definition of severe weather occurring within 25 mi (~40 km) of a point. For the probabilistic surrogate severe metrics, the CSI was again calculated for each forecast threshold on the scorecard.

As mentioned previously, a limitation of the scorecard method is that only a certain number of rows is feasible for simultaneous display. If too many rows are

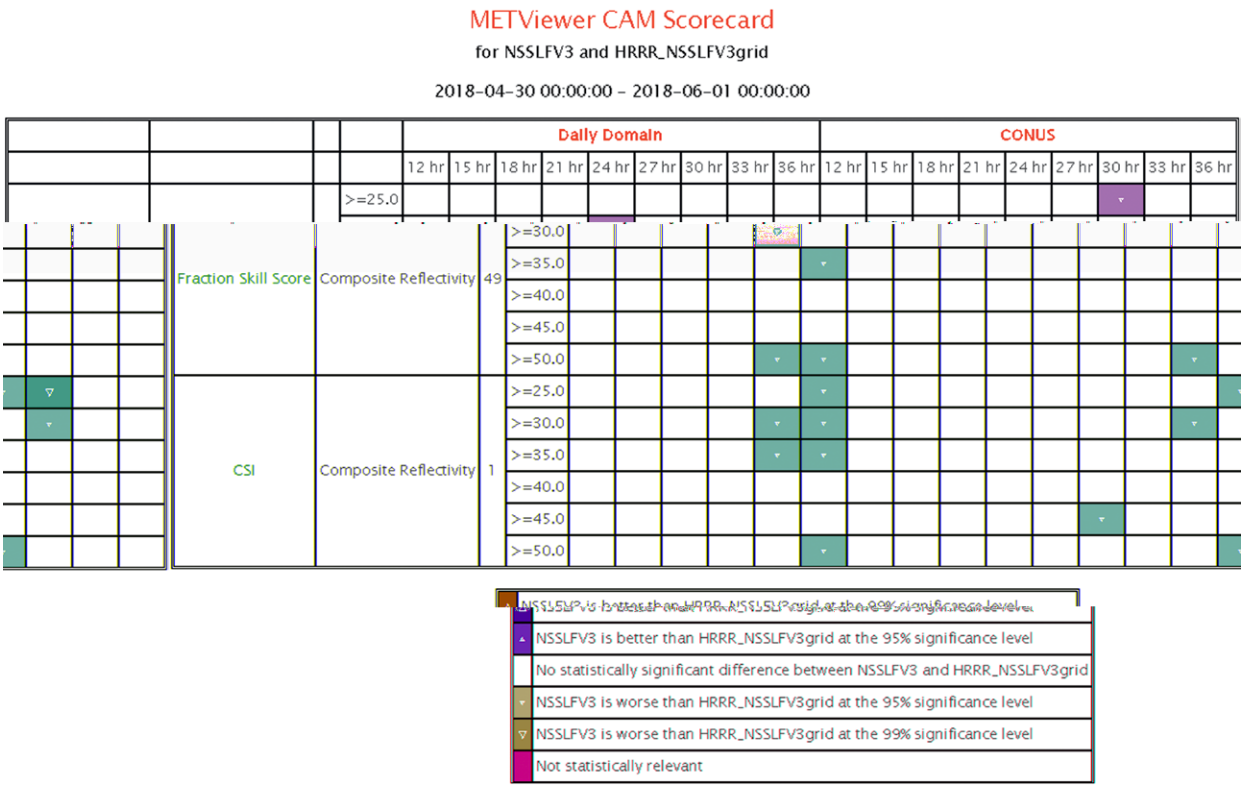


**Fig. 5. A scorecard comparing the FSS and CSI for the composite reflectivity every 3 h from forecast hours 12–36 in the NSSL-FV3 and the HRRRv3. The third column indicates the gridpoint neighborhood being tested. Numbers 1 and 49 refer to a one- and seven-gridpoint radius, respectively. Purple colors indicate that the HRRRv3 is performing better, while green colors indicate that the NSSL-FV3 is performing better.**

included in the scorecard, it becomes unwieldy. The selection process described previously demonstrates the large number of subjective choices that still go into objective evaluation; choosing what to evaluate, how to evaluate it, and at what thresholds is rife with subjectivity.

*The CAM scorecard within SFE 2018.* During the 2018 SFE, the scorecard was presented during the morning forecast discussion (open to all residents of the National Weather Center in addition to SFE participants) on Fridays. The Friday presentation allowed participants to match their subjective impressions of the models formed throughout the week with the objective verification provided by the scorecard. Another advantage of the Friday presentation was that the sample size for each week was largest on Fridays—while the first week of the experiment only had statistics spanning four days (Monday–Thursday of the first week), the scorecard shown on the final Friday of the experiment contained information from the entire experiment, except for that day.

Final scorecards generated at the end of SFE 2018 compared the NSSL-FV3 and the GFDL-FV3 (Fig. 4), the HRRRv3 and the NSSL-FV3 (Fig. 5), and the HRRRv3 and the GFDL-FV3 (see supplementary material). Prior to evaluation, each pair of models was regridded to a common grid matching the coarser of the two models—the NSSL-FV3 grid in comparisons involving the NSSL-FV3, and the HRRRv3 grid for the GFDL-FV3/HRRRv3 comparison. Only the HRRRv3 and GFDL-FV3 comparison included accumulated precipitation. In terms of composite reflectivity, the NSSL-FV3 outperformed the GFDL-FV3 for most hours (Fig. 4), particularly at lower d*BZ* thresholds. The daily domain showed statistically similar performance around forecast hours 21–24 (often near the time of convective initiation), but the CONUS-wide domain showed larger model differences throughout the forecast day. Conversely, the NSSL-FV3 and HRRRv3 scorecard (Fig. 5) showed relatively similar performance in reflectivity, with most of the significant differences occurring at the 95% significance level. In those comparisons, the HRRRv3 outperformed the NSSL-FV3. These slight differences within the daily domain were during forecast hours 24 and 27, which often had initiating or ongoing convection.

These differences may in part be due to differences in microphysics schemes between the different models; small changes is assumed particle size distributions can contribute to large differences in the reflectivity fields (Koch et al. 2005). Since the GFDL-FV3 used the GFDL-6 category microphysics scheme (Chen and Lin 2013) and the NSSL-FV3 and HRRRv3 used different versions of the Thompson microphysics (Thompson et al. 2008), composite reflectivity differences may reflect differences in the hydrometeor distributions of these schemes. However, given that the evaluation of FV3 at CAM scales is relatively recent, comparing the simulated reflectivity values using different microphysics schemes may provide guidance as to which microphysics scheme is performing best with the FV3 dynamical core for warm-season convection. Additionally, simulated reflectivity is best described as a surrogate for observed reflectivity, given that observed reflectivity values can come from multiple combinations of hydrometeors (Kain et al. 2008). However, systemic biases and information regarding features such as the diurnal cycle of convection can still be demonstrated by comparing the observed and simulated reflectivity fields, as in Kain et al. (2008).

As would be expected from the previous two scorecards, when comparing the composite reflectivity of the HRRRv3 and the GFDL-FV3 the HRRRv3 outperforms the GFDL-FV3 for the metrics shown and where a statistically significant difference between the two models exists. This scorecard can be found in the online supplementary material. The accumulated precipitation shows the same results, with statistical significance occurring at even more forecast hours for the 1-h accumulated precipitation than for the composite reflectivity, although there were some hours where the statistical significance decreased going from 1- to 3-h accumulated precipitation. The 3-h accumulated precipitation also tends to have more statistically significant differences between the GFDL-FV3 and the HRRRv3 than the 1-h accumulated precipitation. Across both accumulated precipitation variables, model differences are more statistically significant across the CONUS than across the daily domain, which may be a function of the sample size. There are fewer grid points within the daily domain than within the CONUS, although the daily domain is positioned to capture the most convectively interesting features within the CONUS each day. Therefore, we would expect the most relevant features to a CAM scorecard for the SFE to be within the daily domain. Surrogate severe forecasts from the deterministic models showed little statistically significant difference (not shown).

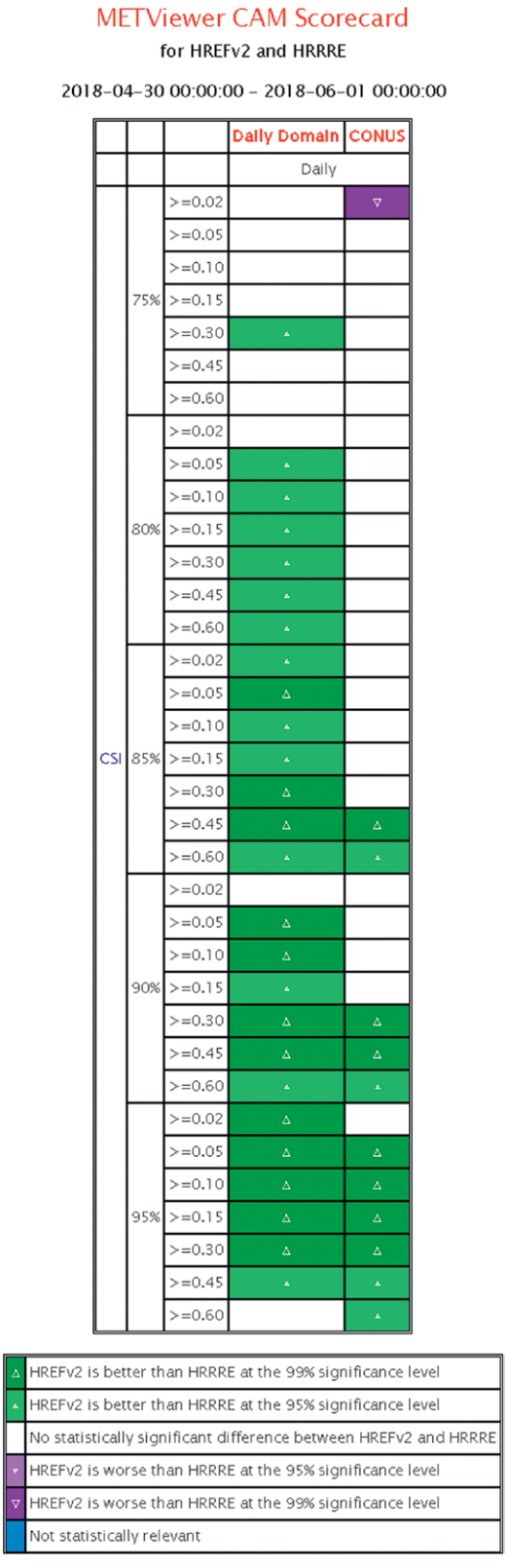


**Fig. 6. A scorecard comparing surrogate severe fields between the HREFv2 and the HRRRE for different percentiles of UH used to generate the field and probability thresholds. Green (purple) colors indicate that the HREFv2 scored higher (lower) than the HRRRE.**

Cumulative stats for SFE 2018: FSS for REFC [40]

| Model | F13 | F14 | F15 | F16 | F17 | F18 | F19 | F20 | F21 | F22 | F23 | F24 | F25 | F26 | F27 | F28 | F29 | F30 | F31 | F32 | F33 | F34 | F35 | F36 | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FV3-GFDL | 0.040 | 0.034 | 0.031 | 0.020 | 0.020 | 0.029 | 0.044 | 0.057 | 0.071 | 0.082 | 0.081 | 0.085 | 0.071 | 0.076 | 0.081 | 0.079 | 0.063 | 0.082 | 0.081 | 0.095 | 0.076 | 0.056 | 0.067 | 0.071 | **0.101** |
| FV3-NSSL | 0.129 | 0.114 | 0.067 | 0.071 | 0.053 | 0.069 | 0.076 | 0.087 | 0.095 | 0.107 | 0.098 | 0.100 | 0.097 | 0.122 | 0.123 | 0.125 | 0.118 | 0.113 | 0.131 | 0.139 | 0.134 | 0.149 | 0.103 | 0.067 | **0.124** |
| HRRRv3 | 0.135 | 0.099 | 0.075 | 0.066 | 0.053 | 0.055 | 0.073 | 0.078 | 0.100 | 0.102 | 0.114 | 0.140 | 0.138 | 0.158 | 0.162 | 0.207 | 0.198 | 0.163 | 0.139 | 0.160 | 0.172 | 0.186 | 0.121 | 0.095 | **0.167** |

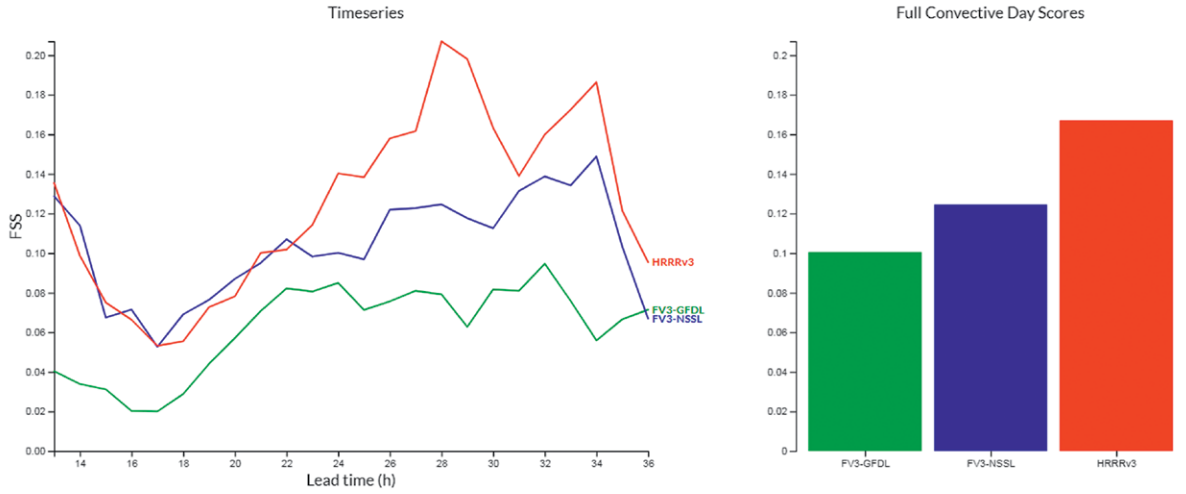


**Fig. 7. An example of objective statistics available on the SFE 2018 website. Shown is the FSS for composite reflectivity at 35 dBZ for the deterministic models (left) at hourly intervals over forecast hours 12–36 and (right) aggregated over the full 24-h convective day. (top) The table shows the exact values of the FSS, with the best-performing model highlighted in green.**

A scorecard comparing the surrogate severe fields from the HRRRE and the HREFv2 (Fig. 6) shows statistically significant differences between the two ensembles, particularly over the daily domain, at thresholds higher than 2%, and at the 80th percentile of UH and above. In these cases, the HREFv2 performed better than the HRRRE, which matched the subjective impressions of participants within the SFE. At the 80th- and 85th-percentile thresholds, these differences were focused in the daily domains, with little statistically significant difference occurring across the entire CONUS. At higher percentiles, however, the results of the daily domain and the CONUS domain are more similar, particularly at higher probability thresholds like 45%. This result likely shows that the daily domain successfully encompassed the high surrogate severe probabilities, as indicated by the presence of model UH tracks and observed local storm reports.

Participant impressions of the scorecard were generally favorable, with participants stating that they would like to see more verification work like this undertaken as part of the SFE's daily activities. In addition to the scorecard, MET output was plotted each day for select forecast system comparisons (Fig. 7) and available on the SFE's website (https://hwt.nssl.noaa.gov/sfe_viewer/2018/verification/; Roberts et al. 2019), so participants were able to see how the scores changed as the experiment progressed. These graphical outputs presented a complementary display to the scorecard by showing the actual values of the statistics. Similar graphics can be generated on demand by using the online METviewer tool. This ability could allow participants to query particular metrics, fields, and thresholds that may have been excluded from the scorecard, as well as view multiple models simultaneously.

*Challenges in development and implementation.* A few major challenges were faced while developing the CAM scorecard for the 2018 SFE. Ensuring proper data flow and processing delayed implementation to the later weeks of the experiment. As such, we recommend that attempts to implement the scorecard for real-time use leave a development period sufficient to ensure timely data availability for scorecard generation. The process of determining appropriate fields, thresholds, and metrics also took time and focused on problems of interest to the 2018 SFE. In addition, technical challenges may arise while determining how to best verify CAMs, hindering a useful intercomparison. These challenges may also provide information about the CAMs that could be useful to the forecasters and model developers. For instance, initially thresholds of UH (e.g., 75 $m^2 s^{-2}$) were used to generate the surrogate severe fields. However, the

**Table 2. UH thresholds ($m^2 s^{-2}$) corresponding to percentiles selected for generating the probabilistic surrogate severe field. Mean values from the ensemble were used for the HRRRE due to the consistent composition of the ensemble members.**

| Model | 75th percentile | 80th percentile | 85th percentile | 90th percentile | 95th percentile |
|---|---|---|---|---|---|
| HRRRv3 | 14.2 | 19.0 | 26.0 | 38.0 | 61.0 |
| NSSL-FV3 | 92.4 | 114.7 | 142.4 | 184.1 | 262.2 |
| GFDL-FV3 | 87.7 | 108.6 | 133.3 | 173.0 | 244.7 |
| HRRRE | 12.3 | 16.0 | 21.9 | 32.4 | 54.6 |
| EMC HRW ARW | 9.2 | 13.5 | 19.7 | 29.7 | 51.3 |
| EMC HRW ARW2 | 12.4 | 17.4 | 24.1 | 35.1 | 59.4 |
| EMC HRW NMMB | 27.8 | 31.2 | 49.6 | 68.5 | 104.8 |
| EMC NAM CONUS NEST | 23.1 | 33.4 | 46.0 | 65.0 | 99.4 |

UH climatologies differ greatly between dynamical cores; FV3-based models tend to have higher UH values than WRF-based models, in part due to differences in how UH is calculated between dynamical cores (Potvin et al. 2019). Therefore, a change from UH thresholds to selected percentiles of UH (Table 2) was implemented after the 2018 SFE to ensure a fair comparison between all model cores, particularly at high percentiles where climatological differences can be exacerbated. These lessons will be applied in SFE 2019, when a daily real-time scorecard is planned.

## THE FUTURE OF THE CAM SCORECARD.

After the 2018 SFE, planned upgrades to METplus include the addition of surrogate severe and percentile capabilities, so that METplus can incorporate preprocessing of these data and eliminate steps that users currently have to complete. Working with the datasets generated during SFE 2018, statistics for additional environmental fields such as 2-m temperature and 10-m zonal ($U$) and meridional ($V$) wind components were included in the scorecard (Fig. 8), and often showed more mixed results of which model was performing better than the storm attribute and precipitation fields did. While these fields are critical to forecasting severe convective weather, they are also fundamental environmental fields and therefore of interest to a wider meteorological community. The use of categorical statistics for the 2-m temperature and winds demonstrates the utility of using scores beyond traditional continuous measures. For example, it demonstrates that during the 2018 SFE, HRRRv3 tends to perform better in cold temperatures within the domain, but NSSL-FV3 tends to have higher skill at warmer temperatures, which were a larger part of this dataset. Additionally, it appears that the NSSL-FV3 performs better at lower wind speed thresholds and HRRRv3 at higher ones.

Mixed results such as the ones found on the environmental field scorecard can be commonplace if enough different fields, metrics, and times are evaluated—it is exceptionally challenging to develop a new implementation of a model that exceeds the performance of the prior model across in all ways. This is especially true looking from a broader perspective, across applications beyond severe weather. For example, when the NWS implements changes to their numerical models, they must be concerned about forecast problems ranging from air quality to winter weather to tropical systems. A scorecard for any single of these applications could have a plethora of rows and mixed results, let alone an enterprise-wide scorecard. It is therefore imperative to consider practical significance as well as statistical significance in determining the difference between the two modeling systems. However, it is likely that the scorecard will rarely provide a clear "correct answer" across all aspects being evaluated.

The expansion of the CAM scorecard for the SFE into environmental information is our initial effort toward having the scorecard encompass other meteorological scales and processes, and demonstrates how the CAM scorecard can distinguish between models that may be quite similar in aggregate statistics or for a smaller selection of metrics. During the expansion process, we hope to involve multiple stakeholders as was done in the DTC Metrics Workshop. Combining perspectives from groups throughout the meteorological community can provide consistent judgment of new model implementations from upgrade to upgrade, and interested parties can hone in on metrics, fields, and thresholds important to them. By strengthening and fostering these partnerships during development, input from across the weather enterprise can be incorporated and the scorecard can be developed to best serve the community. It is

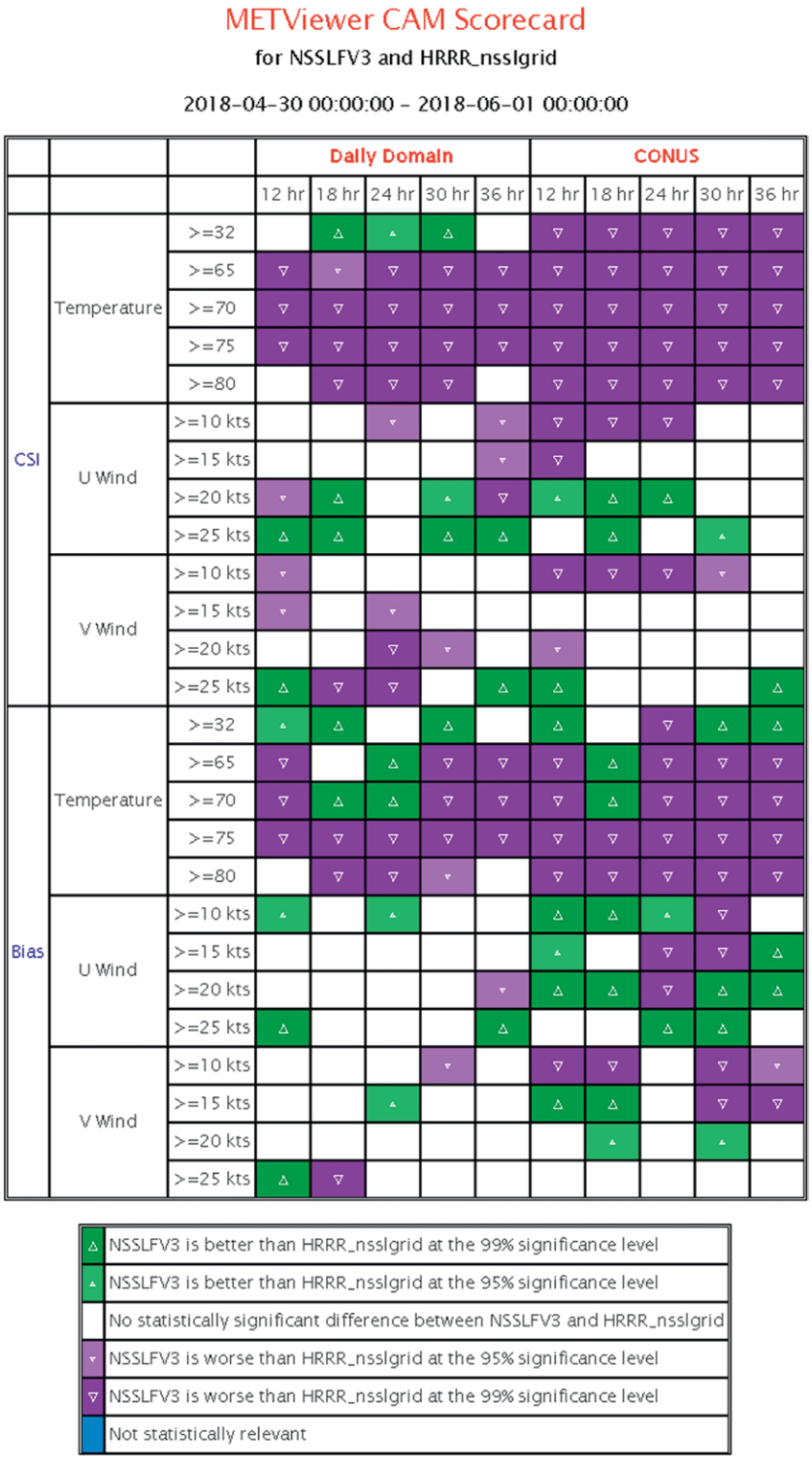


**Fig. 8. A scorecard comparing the bias and CSI of the temperature and 10-m *U* and *V* wind components between the NSSL-FV3 and the HRRRv3. Green (purple) colors indicate that the NSSL-FV3 scored higher (lower) than the HRRRv3.**

our hope that the scorecard can provide a visualization tool for a unified framework that includes aspects of model performance important to both model developers and end users such as operational forecasters.

**ACKNOWLEDGMENTS.** The authors thank the participants and facilitators of the 2018 SFE, as well as the many collaborators who contribute significant work to ensure the success of the experiment each year. Particular thanks go to Dr. Lucas Harris and Dr. Yunheng Wang for their work in implementing the GFDL-FV3 and NSSL-FV3, respectively, during the 2018 SFE. We would also like to thank the participants in the DTC Metrics Workshop for their thoughtful contributions to the workshop and efforts at determining the high-priority targets for verification efforts. BTG and BR 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. Authors TLJ, CPK, JHG, and HHF completed this work as part of duties associated with NOAA OAR OWAQ Project NA17OAR4590119 entitled "Developing an Objective Evaluation Scorecard for Storm Scale Prediction." We would also like to thank three anonymous reviewers of the manuscript for their constructive and helpful comments on earlier drafts of this work.

## REFERENCES


Adams-Selin, R. D., A. J. Clark, C. J. Melick, S. R. Dembek, I. L. Jirak, and C. L. Ziegler, 2019: Evolution of WRF-HAILCAST during the 2014–16 NOAA/Hazardous Weather Testbed Spring Forecasting Experiments. *Wea. Forecasting*, **34**, 61–79, https://doi.org/10.1175/WAF-D-18-0024.1.

Adriaansen, D., and Coauthors, 2018: The METplus version 2.0 user's guide. Developmental Testbed Center, 85 pp., https://github.com/NCAR/METplus/releases.

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.

Anthes, R. A., 1983: Regional models of the atmosphere in middle latitudes. *Mon. Wea. Rev.*, **111**, 1306–1335, https://doi.org/10.1175/1520-0493(1983)111<1306:RMOTAI>2.0.CO;2.

Benedetti, A., and F. Vitart, 2018: Can the direct effect of aerosols improve subseasonal predictability? *Mon. Wea. Rev.*, **146**, 3481–3498, https://doi.org/10.1175/MWR-D-17-0282.1.

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

Brier, G. W., and R. A. Allen, 1951: Verification of weather forecasts. *Compendium of Meteorology*, T. Malone, Ed., Amer. Meteor. Soc., 841–848.

Buizza, R., and Coauthors, 2018: The development and evaluation process followed at ECMWF to upgrade the Integrated Forecasting System (IFS). ECMWF Tech. Memo. 829, 47 pp., https://doi.org/10.21957/xzopnhty9.

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.

——, 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.

Dawson, L.C., G.S. Romine, R.J. Trapp, and M.E. Baldwin, 2017: Verifying supercellular rotation in a convection-permitting ensemble forecasting system with radar-derived rotation track data. *Wea. Forecasting*, **32**, 781–795, https://doi.org/10.1175/WAF-D-16-0121.1.

Developmental Testbed Center, 2018: 2018 DTC Community Unified Forecast System Test Plan and Metrics Workshop, 8 pp., https://dtcenter.org/sites/default/files/events/2018/dtc-testplans-metrics-workshop-report.pdf.

Doswell, C. A., III, H. E. Brooks, and M. P. Kay, 2005: Climatological estimates of daily local nontornadic severe thunderstorm probability for the United States. *Wea. Forecasting*, **20**, 577–595, https://doi.org/10.1175/WAF866.1.

Dowell, D., C. A. Alexander, T. Alcott, and T. T. Ladwig, 2018: HRRR Ensemble (HRRRE) Guidance 2018 HWT Spring Experiment. 6 pp., https://rapidrefresh.noaa.gov/internal/pdfs/2018_Spring_Experiment_HRRRE_Documentation.pdf.

Edwards, R., J. T. Allen, and G. W. Carbin, 2018: Reliability and climatological impacts of convective wind estimations. *J. Climate Appl. Meteor.*, **57**, 1825–1845, https://doi.org/10.1175/JAMC-D-17-0306.1.

EMC Model Evaluation Group, 2018a: Conclusion of the FV3GFS evaluation. 93 pp., www.emc.ncep.noaa.gov/users/Alicia.Bentley/fv3gfs/updates/MEG_9-27-18_FV3GFS_EVAL.pptx.

EMC Model Evaluation Group, 2018b: Retrospective forecast performance statistics [full period (June 2015-September 2018)]. www.emc.ncep.noaa.gov/users/meg/fv3gfs.

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.

Gilbert, G. K., 1884: Finley's tornado predictions. *Amer. Meteor. J.*, **1**, 166–172.

Halley Gotway, J., and Coauthors, 2018: The Model Evaluation Tools v8.0 (METv8.0) user's guide. Developmental Testbed Center, 407 pp., www.dtcenter.org/met/users/docs/users_guide/MET_Users_Guide_v8.0.pdf.

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

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

Hollingsworth, A., K. Arpe, M. Tiedtke, M. Capaldo, and H. Savijärvi, 1980: The performance of a medium-range forecast model in winter–impact of physical parameterizations. *Mon. Wea. Rev.*, **108**, 1736–1773, https://doi.org/10.1175/1520-0493(1980)108<1736:TPOAMR>2.0.CO;2.

Janjić, Z. I., and R. Gall, 2012: Scientific documentation of the NCEP nonhydrostatic multiscale model on the B grid (NMMB). Part 1 Dynamics. NCAR Tech. Note NCAR/TN-489+STR, 75 pp., https://doi.org/10.5065/D6WH2MZX.

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

Jolliffe, I. T., and D. Stephenson, 2011: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons, 292 pp.

Kain, J. S., P. R. Janish, S. J. Weiss, M. E. Baldwin, R. S. Schneider, and H. E. Brooks, 2003: Collaboration between forecasters and research scientists at the NSSL and SPC: The spring program. *Bull. Amer. Meteor. Soc.*, **84**, 1797–1806, https://doi.org/10.1175/BAMS-84-12-1797.

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

Koch, S. E., B. Ferrier, M. Stolinga, E. Szoke, S. J. Weiss, and J. S. Kain, 2005: The use of simulated radar reflectivity fields in the diagnosis of mesoscale phenomena from high-resolution WRF model forecasts. Preprints, *12th Conf. on Mesoscale Processes*, Albuquerque, NM, Amer. Meteor. Soc., J4J.7, http://ams.confex.com/ams/pdfpapers/97032.pdf.

Kuhl, D. D., T. E. Rosmond, C. H. Bishop, J. McLay, and N. L. Baker, 2013: Comparison of hybrid ensemble/4DVar and 4DVar within the NAVDAS-AR data assimilation framework. *Mon. Wea. Rev.*, **141**, 2740–2758, https://doi.org/10.1175/MWR-D-12-00182.1.

Lin, Y., 2011. GCIP/EOP Surface: Precipitation NCEP/EMC 4KM Gridded Data (GRIB) Stage IV Data. Version 1.0, UCAR/NCAR Earth Observing Laboratory, https://doi.org/10.5065/D6PG1QDD.

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

Marsh, P. T., J. S. Kain, V. Lakshmanan, A. J. Clark, N. M. Hitchens, and J. Hardy, 2012: A method for calibrating deterministic forecasts of rare events. *Wea. Forecasting*, **27**, 531–538, https://doi.org/10.1175/WAF-D-11-00074.1.

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

Melick, C. J., I. L. Jirak, J. Correia Jr., A. R. Dean, and S. J. Weiss, 2013: Utility of objective verification metrics during the 2013 HWT Spring Forecasting Experiment. Preprints, *38th NWA Annual Meeting*, Charleston, SC, National Weather Association, P.1.27.

Murphy, A. H., 1991: Forecast verification: Its complexity and dimensionality. *Mon. Wea. Rev.*, **119**, 1590–1601, https://doi.org/10.1175/1520-0493(1991)119<1590:FVICAD>2.0.CO;2.

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.

R Development Core Team, 2019: R: A language and environment for statistical computing. R Foundation for Statistical Computing, www.R-project.org/.

Roberts, B., I. Jirak, A. Clark, S. Weiss, and J. Kain, 2019: Post-processing 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.

Sandgathe, S., W. O'Connor, N. Lett, and D. McCarren, 2011: National Unified Operational Prediction Capability Initiative. *Bull. Amer. Meteor. Soc.*, **92**, 1347–1351, https://doi.org/10.1175/2011BAMS3212.1.

——, B. Brown, B. Etherton, and E. Tollerud, 2013: Designing multimodel ensembles requires meaningful

methodologies. *Bull. Amer. Meteor. Soc.*, **94**, ES183–ES185, https://doi.org/10.1175/BAMS-D-12-00234.1.

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.

Schumacher, R. S., and A. J. Clark, 2014: Evaluation of ensemble configurations for the analysis and prediction of heavy-rain-producing mesoscale convective systems. *Mon. Wea. Rev.*, **142**, 4108–4138, https://doi.org/10.1175/MWR-D-13-00357.1.

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

Shuman, F. G., 1989: History of numerical weather prediction at the National Meteorological Center. *Wea. Forecasting*, **4**, 286–296, https://doi.org/10.1175/1520-0434(1989)004<0286:HONWPA>2.0.CO;2.

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

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

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

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

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

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

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.

Trapp, R. J., D. M. Wheatley, N. T. Atkins, R. W. Przybylinski, and R. Wolf, 2006: Buyer beware: Some words of caution on the use of severe wind reports in postevent assessment and research. *Wea. Forecasting*, **21**, 408–415, https://doi.org/10.1175/WAF925.1.

Verbout, S. M., H. E. Brooks, L. M. Leslie, and D. M. Schultz, 2006: Evolution of the U.S. tornado database: 1954–2003. *Wea. Forecasting*, **21**, 86–93, https://doi.org/10.1175/WAF910.1.

Zhang, J., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 621–638, https://doi.org/10.1175/BAMS-D-14-00174.1.

Zhou, X., Y. Zhu, D. Hou, Y. Luo, J. Peng, and R. Wobus, 2017: Performance of the new NCEP Global Ensemble Forecast System in a parallel experiment. *Wea. Forecasting*, **32**, 1989–2004, https://doi.org/10.1175/WAF-D-17-0023.1.