THE COMMUNITY LEVERAGED UNIFIED ENSEMBLE (CLUE) IN THE 2016 NOAA/HAZARDOUS WEATHER TESTBED SPRING FORECASTING EXPERIMENT

Adam J. Clark, Israel L. Jirak, Scott R. Dembek, Gerry J. Creager, Fanyou Kong, Kevin W. Thomas, Kent H. Knopfmeier, Burkely T. Gallo, Christopher J. Melick, Ming Xue, Keith A. Brewster, Youngsun Jung, Aaron Kennedy, Xiquan Dong, Joshua Markel, Matthew Gilmore, Glen S. Romine, Kathryn R. Fossell, Ryan A. Sobash, Jacob R. Carley, Brad S. Ferrier, Matthew Pyle, Curtis R. Alexander, Steven J. Weiss, John S. Kain, Louis J. Wicker, Gregory Thompson, Rebecca D. Adams-Selin, and David A. Imy

The CLUE system represents an unprecedented effort to leverage several academic and government research institutions to help guide NOAA's operational environmental modeling efforts at the convection-allowing scale.

The National Severe Storms Laboratory (NSSL) and Storm Prediction Center (SPC) coorganize annual Spring Forecasting Experiments (SFEs), which are conducted in NOAA's Hazardous Weather Testbed (HWT) at the National Weather Center in Norman, Oklahoma, for five weeks during the climatological peak of the severe weather season. The SFEs are designed to test emerging concepts and technologies for improving the prediction of hazardous convective weather with the primary goals of accelerating the transfer of promising new tools and concepts from research to operations, inspiring new initiatives for operationally relevant research, and identifying and documenting sensitivities and performance characteristics of state-of-the-art experimental convection-allowing modeling (CAM) systems. Over the last decade, the SFEs have emerged as an international resource for developing and evaluating the performance of new CAM systems, and major advances have been made in creating, importing, processing, verifying, and extracting unique hazardous weather fields while providing analysis and visualization tools including probabilistic information, for these large and complex datasets. For example, during the 2010 experiment (Clark et al. 2012), in addition to providing a 26-member, 4-km grid-spacing CAM-based ensemble, the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma provided a 1-km contiguous U.S. (CONUS) domain forecast that required over 10,000 computing cores. In the 2015 SFE (Gallo et al. 2017), six unique and

independently designed CAM-based ensembles were contributed by CAPS, the National Center for Atmospheric Research (NCAR), NSSL, SPC, and the Air Force Weather Agency (AFWA; now called the 557th Weather Wing). Figure 1 provides a summary of CAMs examined since 2007, along with a timeline of CAM guidance milestones.

Through the SFEs, much has been learned about how to utilize and configure CAMs and CAM ensembles, and since 2007 the number of CAM systems (including ensembles) examined in the HWT has increased dramatically. Meanwhile, new technologies and physical understanding have been migrated to the SPC, enhancing the timeliness and accuracy of their severe weather forecasts. Despite these advances, progress toward identifying optimal CAM ensemble configurations has been inhibited because HWT collaborators have independently designed contributed CAM systems, which makes it difficult to attribute differences in performance characteristics. For example, during the 2015 SFE, CAPS and NSSL contributed mixed- and single-physics ensembles, respectively, but because of other differences in the configurations (e.g., initial condition perturbations, data assimilation, grid spacing, domain size, and model version), the impacts of single versus mixed-physics configurations could not be isolated. Thus, after the 2015 SFE it was clear to SFE leaders that more controlled experiments were needed. Furthermore, around the same time period, the international University Corporation for Atmospheric Research Community Advisory Committee for the National Centers for Environmental Prediction (UCACN) Model Advisory Committee, which is charged with developing recommendations for a unified NOAA modeling strategy to advance the United States to world leadership in numerical modeling, released a comprehensive set of recommendations[1] that included the following: 1) the NOAA environmental modeling community requires a rational, evidence-driven approach toward decision-making and modeling system development; 2) a unified collaborative strategy for model development across NOAA is needed; and 3) NOAA needs to better leverage the capabilities of the external community. Thus, in the spirit of these recommendations, organizers of the HWT SFEs made a major push to coordinate efforts among its large group of collaborators in 2016. Specifically, instead of each group providing a separate, independently designed CAM-based ensemble, all groups agreed on a set of model specifications so that the simulations contributed by each group could be viewed as one large, carefully designed "superensemble." This design facilitated a number of controlled experiments geared toward finding optimal configuration strategies for CAM ensembles and has been termed the Community Leveraged Unified Ensemble (CLUE, hereafter). The superensemble concept has been used in previous works for tropical cyclone, weather, climate, and seasonal prediction systems (e.g., Krishnamurti et al. 1999; Palmer et al. 2004; Krishnamurti et al. 2016 and references therein), but has yet to be applied within a CAM ensemble framework. However, the philosophy behind the CLUE design is different from these previous works on superensembles. Specifically, the CLUE has a more coordinated design with goals that are more focused on identifying impacts of different ensemble design strategies, rather than generating a single "best" forecast from independent ensemble datasets.

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

AFFILIATIONS: Clark, Gallo, Kain, Wicker, and Imy—NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma; Jirak and Weiss—NOAA/NWS/Storm Prediction Center, Norman, Oklahoma; Dembek, Creager, and Knopfmeier— NOAA/OAR/National Severe Storms Laboratory, and Cooperative Institute for Mesoscale Meteorological Studies, Norman, Oklahoma; Kong, Thomas, Brewster, and Jung—Center for Analysis and Prediction of Storms, Norman, Oklahoma; Melick— NOAA/NWS/Storm Prediction Center, and Cooperative Institute for Mesoscale Meteorological Studies, Norman, Oklahoma; Xue—School of Meteorology, University of Oklahoma, and Center for Analysis and Prediction of Storms, Norman, Oklahoma; Kennedy, Dong, Markel, and Gilmore—Department of Atmospheric Science, University of North Dakota, Grand Forks, North Dakota; Romine, Fossell, Sobash, and Thompson—National Center of Atmospheric Research, Boulder, Colorado; Carley, Ferrier, and Pyle—NOAA/Environmental Modeling Center, Camp Springs, Maryland; Alexander—NOAA/OAR/Earth System Research Laboratory/Global Systems Division, Boulder, Colorado; Adams-Selin—557th Weather Wing, Offutt Air Force Base, Nebraska

CORRESPONDING AUTHOR: Adam J. Clark, adam.clark@noaa.gov

The abstract for this article can be found in this issue, following the table of contents.

DOI:10.1175/BAMS-D-16-0309.1

A supplement to this article is available online (10.1175/BAMS-D-16-0309.2)

In final form 15 December 2017

The CLUE system represents an unprecedented effort to leverage several academic and government research institutions to help guide NOAA's operational environmental modeling efforts. In future SFEs, the CLUE will be reconfigured based on results from previous years, advances in technology, and feedback from the operational and research communities. Furthermore, the CLUE framework will help test initial convection-allowing versions of the Finite Volume Cubed Sphere Model (FV3; Putman and Lin 2007) developed at NOAA's Geophysical Fluid Dynamics Laboratory. The FV3 has been selected as the dynamic core to replace the Global Forecast System (GFS) model as part of the Next Generation Global Prediction System (NGGPS; www.weather.gov/sti/stimodeling_nggps) program

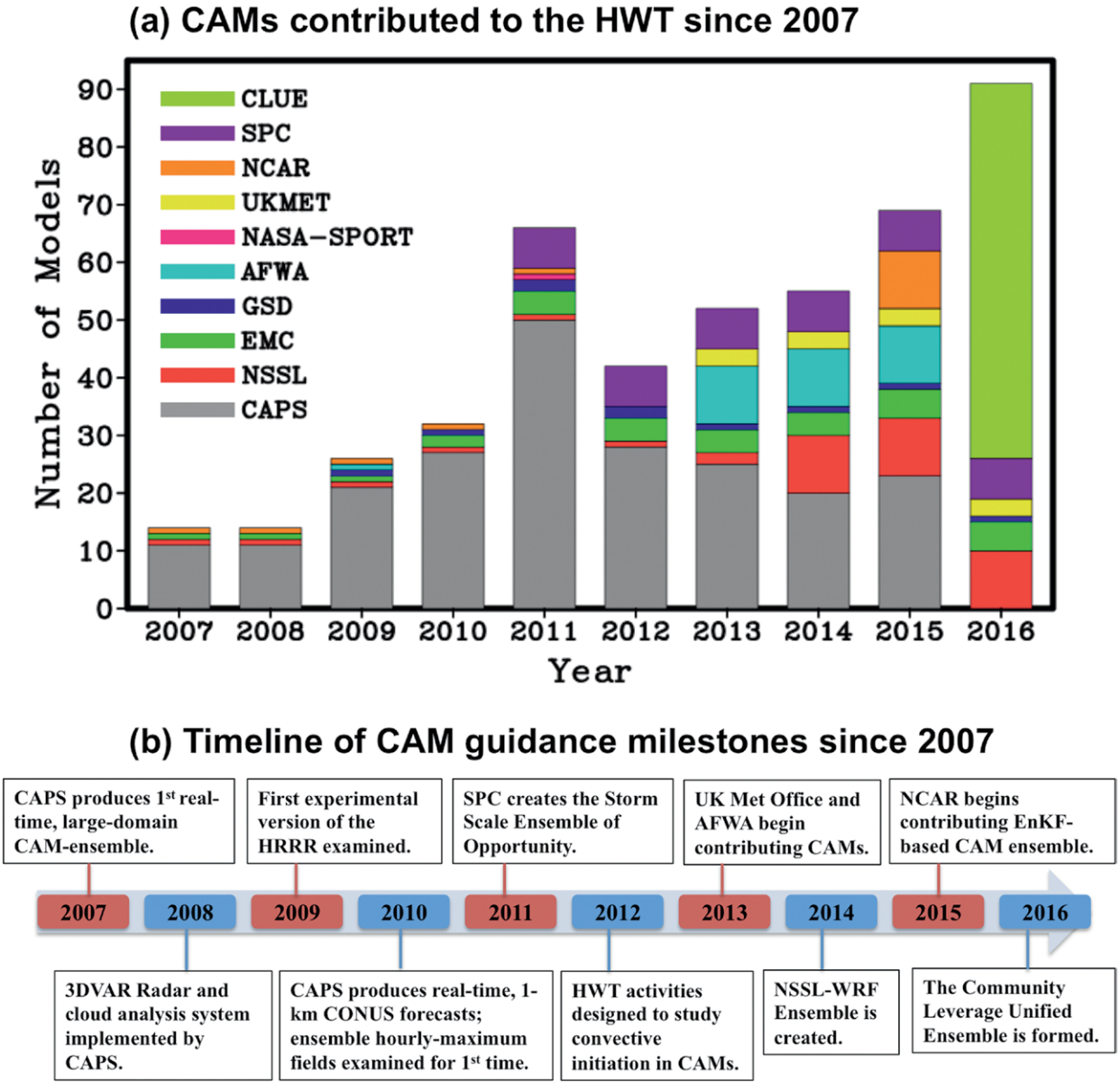

Fig. 1. (a) Stacked bar graph indicating the number of unique CAMs used each year since 2007 in the HWT SFEs. The different colors denote the number of models contributed by the different agencies. A legend is provided at the top left. Abbreviations are defined as follows: CLUE, Community Leveraged Unified Ensemble; SPC, Storm Prediction Center; NCAR, National Center for Atmospheric Research; UKMET, Met Office; NASA SPORT, National Aeronautics and Space Administration Short-term Prediction Research and Transition Center; AFWA, Air Force Weather Agency; GSD, Global Systems Division of NOAA's Earth System Research Laboratory; EMC, NOAA's Environmental Modeling Center; NSSL, National Severe Storms Laboratory; and CAPS, Center for Analysis and Prediction of Storms. (b) Timeline of CAM guidance milestones at the HWT since 2007.

and FV3 is envisioned as the eventual foundation for NOAA's regional models and ensemble systems. This will require much research, development, and testing to ensure that FV3 performs equal to or better than existing regional short-term forecasting systems.

This article describes the design of the 2016 CLUE system and the eight specific experiments that were conducted within the CLUE framework. Additionally, as an example of the research enabled by the CLUE framework, results are presented from one of the experiments that examined the impact of using single versus multicore CAM ensemble configurations.

CLUE CONFIGURATION. The idea for the CLUE system was formulated in fall of 2015. At this time, plans were already in place for several groups of collaborators to contribute model data to the 2016 SFE through NOAA-funded research-to-operations projects. For example, NCAR and CAPS had projects funded by NOAA's Oceanic and Atmospheric Research (OAR) Office of Water and Air Quality (OWAQ), and the University of North Dakota (UND) had a project funded by the National Weather Service Research to Operations Initiative. While the model runs NSSL contributed to the CLUE were not supported by a specific grant, the Texas Advanced Computing Center (TACC) provided generous computing resources for their contribution. Since participation in the CLUE would require work beyond that outlined in their already-existing projects, leaders from each group of collaborators were approached individually to gauge whether they had the resources and willingness to participate. Fortunately, because of the mutually beneficial research that the CLUE system would enable, along with the potential to provide evidence to help optimize NOAA's first operational CAM-ensemble configuration, all collaborators were eager and willing to participate.

The CLUE configuration was formulated by considering some basic research questions, such as how to optimize CAM-ensemble configurations and how to build around each collaborator's already existing plans for model data contributions. Ultimately, the CLUE was designed to have 66 members: 35 contributed by CAPS, 15 by NSSL, 10 by NCAR, 5 by the UND, and 1 from the Earth System Research Laboratory/Global Systems Division (ESRL/GSD). The runs were conducted on several different high-performance computing systems. CAPS used TACC's Stampede system and the University of Tennessee's National Institute for Computation Science's (NICS) Darter system, NSSL used TACC's Lonestar5 system, NCAR used the Yellowstone supercomputer, UND used TACC's Stampede system, and ESRL/GSD used NOAA's Jet system.

All members were initialized at 0000 UTC on weekdays with forecasts to 36 h using 3-km grid spacing over a CONUS domain. Members included the Advanced Research version of the Weather Research and Forecasting (WRF-ARW) Model (Skamarock et al. 2008), as well as the Nonhydrostatic Multiscale Model on the B grid (NMMB; Janjić and Gall 2012). The CAPS, UND, and NSSL members all shared a set of common model versions, domain specifications (including vertical levels), physics parameterizations, and postprocessing methods. The ESRL/GSD member was a developmental version of the High Resolution Rapid Refresh (HRRR) model (Benjamin et al. 2016) run to 36 h, which had a slightly different domain than the other members. The NCAR members also had a slightly different domain and used a 1-yr-older version of WRF, which was necessary because their members were from an already established ensemble system whose configuration was based on extensive testing and verification (Schwartz et al. 2015a). The NCAR group did not want to risk introducing changes to their system by adhering exactly to the CLUE specifications, since it could introduce unwanted systematic biases. Despite some minor differences in the NCAR and ESRL/GSD members, postprocessing was standardized across all ensemble subsets (described later).

The basic strategy in designing the CLUE was to formulate several subsets of up to 20 members that could be used to test specific configuration strategies in controlled experiments. Ten unique subsets were formulated, with CAPS contributing five subsets, NSSL two, and ESRL/GSD, NCAR, and UND each contributing one. Some experiments utilized combinations of these subsets. These subsets are described as follows:

1) core (CAPS)—Nine WRF-ARW members were designed to account for as many error sources as possible. The control member used initial conditions (ICs) and lateral boundary conditions (LBCs; 3-h updates) from 12-km grid-spacing North American Mesoscale Forecast System (NAM) analyses and forecasts, respectively. Radar reflectivity and velocity data and other traditional data, including surface observations and rawinsondes, were assimilated into the ICs using the Advanced Regional Prediction System (ARPS) three-dimensional variational data assimilation (3DVAR; Xue et al. 2003; Gao et al. 2004) and cloud analysis (Xue et al. 2003; Hu et al. 2006) system. The other core subset members also used ARPS-3DVAR, but IC perturbations

were derived from evolved (through 3 h) perturbations of 2100 UTC initialized members of the National Centers for Environmental Prediction (NCEP) Short-Range Ensemble Forecast (SREF) system (Du et al. 2006) and added to the control member ICs, with corresponding SREF forecasts used for LBCs. Mixed physics were implemented in the core subset using various combinations of microphysics and planetary boundary layer (PBL)/turbulence schemes.

2) s-phys-rad (CAPS)—Ten WRF-ARW members (including the control member of core) were configured the same as core but used a single set of physics.
3) caps-enkf (CAPS)—Ten WRF-ARW members used the same set of physics and LBCs as core, but with ICs that were derived from an ensemble Kalman filter (EnKF) system.
4) caps-nmmb-rad (CAPS)—A single NMMB run used the same ICs/LBCs as the core control member.
5) caps-nmmb (CAPS)—Five NMMB members had the same ICs/LBCs as five of the s-phys-rad members but did not use ARPS-3DVAR (i.e., a "cold start" was used).
6) s-phys-norad (NSSL)—Ten WRF members were the same as s-phys-rad, but without ARPS-3DVAR (i.e., cold start).
7) nssl-nmmb (NSSL)—Five NMMB members were configured the same as the caps-nmmb members, except they shared a different set of the s-phys-rad ICs/LBCs.
8) HRRR36 (ESRL/GSD)—A development version of the HRRR was configured to provide 36-h forecasts. The HRRR is a 3-km grid-spacing, ARW-based model that is initialized hourly and provides 18-h forecasts.
9) ncar-enkf (NCAR)—Ten WRF members used single physics and ICs/LBCs derived from NCAR's Data Assimilation Research Testbed (DART; Anderson et al. 2009) software (Schwartz et al. 2015a).
10) mp (UND)—Five WRF members had the same ICs/LBCs as the core control member, but with different microphysics parameterizations in each member.

Table 1 provides a summary of the specifications for each CLUE subset, and further details including specifications for every member can be found in the online supplement (https://doi.org/10.1175/BAMS-D-16-0309.2).

CLUE EXPERIMENTS. The design of CLUE allowed for eight unique experiments, which are described as follows:

1) ARW versus NMMB—A direct comparison of the subjective and objective skill of ARW and NMMB dynamic cores was conducted. These

Table 1. Summary of CLUE subsets. IC/LBC perturbations labeled SREF indicate that IC perturbations were extracted from members of NCEP's SREF system and added to 0000 UTC NAM analyses. In subsets with "yes" indicated for mixed physics, the microphysics and turbulence parameterizations were varied, except for subset mp, which only varied the microphysics. Note that the control member of the core ensemble was also used as the control member in the mp and s-phys-rad ensembles. Thus, although the total number of members adds to 67, there were 66 unique members. Further, one member planned for the core subset was not ready for real-time implementation; thus, only nine core members were actually run. The HPC column provides the names of the high-performance computers used for each set of simulations. The agencies that maintain each system are given in the text.

CLUE subset	No. of members	IC/LBC perturbations	Mixed physics?	Data assimilation	Model core	Agency	HPC
core	10 (9)	SREF	Yes	ARPS-3DVAR	ARW	CAPS	Stampede
s-phys-rad	10	SREF	No	ARPS-3DVAR	ARW	CAPS	Stampede
caps-enkf	10	EnKF (CAPS)	Yes	EnKF (CAPS)	ARW	CAPS	Darter
caps-nmmb-rad	1	None	No	ARPS-3DVAR	NMMB	CAPS	Stampede
caps-nmmb	5	SREF	No	Cold start	NMMB	CAPS	Stampede
s-phys-norad	10	SREF	No	Cold start	ARW	NSSL	Lonestar5
nssl-nmmb	5	SREF	No	Cold start	NMMB	NSSL	Lonestar5
HRRR36	1	None	No	RAP-GSI/DFI	ARW	ESRL/GSD	Jet
ncar-enkf	10	EnKF (DART)	No	EnKF (DART)	ARW	NCAR	Yellowstone
mp	5	None	Yes	ARPS-3DVAR	ARW	UND	Stampede

direct comparisons were possible because 10 pairs of NMMB and ARW members within the caps-nmmb and nssl-nmmb, and s-phys-norad, subsets had different model cores but shared the same ICs/LBCs. The optimal dynamic core for CAM applications is still an open question. NMMB is known to be more computationally efficient than ARW, but ARW has been preferred by severe weather forecasters and the severe weather research community because of its more realistic depiction of storm structure and evolution.

2) Multicore versus single-core ensemble design—Three ensembles were compared to test the effectiveness of single-core versus multicore configurations. The first ensemble used five ARW and five NMMB members from the s-phys-norad and nssl-nmmb subsets, respectively; the second used the 10 ARW members from the s-phys-norad subset; and the third used 10 NMMB members from the caps-nmmb and nssl-nmmb subsets. The effectiveness of multicore (or multimodel) ensemble configuration strategies has been demonstrated for seasonal [e.g., the North American Multimodel Ensemble (NMME); Kirtman et al. (2014)], medium-range [e.g., the North American Ensemble Forecast System (NAEFS); Candille (2009)], and short-range (e.g., the SREF; Du et al. 2006) forecasting applications. The use of multiple models with different but equally valid methods for initialization and integration helps to better sample the range of future states than a single modeling system. Although the Storm Scale Ensemble of Opportunity (SSEO; Jirak et al. 2012) has been shown to be a skillful multimodel CAM ensemble, the multimodel strategy has not been tested for CAM applications in controlled experiments. Furthermore, given the push toward model core unification that will better focus model development efforts (e.g., UCAR 2015), it is preferred that a future operational CAM ensemble will be single core. Thus, it is important to quantify how much skill (if any) is sacrificed from a single-core configuration within the context of a controlled experiment.
3) Single physics versus multiphysics—Two ensembles with the same set of perturbed ICs/LBCs were compared to test the impact of single versus multiphysics. One ensemble, core, used varied turbulence and microphysics schemes, while another, s-phys-rad, used a common set of physics. Although past SFEs have quantified the error growth from varied physics within a perfect analysis framework (i.e., nonperturbed ICs/LBCs; e.g., Clark et al. 2010b), there has not been an experiment designed in the SFE to examine the impact of varied physics with perturbed ICs/LBCs in a CAM ensemble. Furthermore, while multiple physics schemes have been shown to increase spread, leading to improved forecast skill (e.g., Stensrud et al. 2000; Hacker et al. 2011; Berner et al. 2011, 2015), there are theoretical and practical disadvantages to multiphysics approaches, including the resource-intensive need to develop and maintain multiple parameterizations, as well as the introduction of systematic biases (e.g., Jankov et al. 2017). Thus, it is important to quantify the gain in skill (if any) from using multiphysics. Future SFEs will explore whether stochastic physics perturbations (e.g., Jankov et al. 2017 and references therein) in a single-physics ensemble can match or exceed the spread and skill from the multiphysics approach.
4) Comparison of ensembles with and without radar data assimilation—Two single-physics ensembles with perturbed ICs/LBCs were identically configured, except one, s-phys-rad, used ARPS-3DVAR to assimilate radar data and other observations in all members, while another, s-phys-norad, used a cold start in all members. Previous studies have documented the impact of radar data assimilation by comparing deterministic models with and without radar data assimilation (e.g., Kain et al. 2010; Stratman et al. 2013), finding that the positive impact of the assimilation is strongest within the first 3–6 h of the forecast but can last up to 12 h. However, these comparisons have not been conducted within an ensemble framework to determine the time length and magnitude of the positive impact of radar assimilation.
5) 3DVAR versus EnKF data assimilation strategies—The core, caps-enkf, and ncar-enkf subsets were compared. Although it is much more computationally expensive than 3DVAR, the EnKF data assimilation method is advantageous because it provides flow-dependent background error covariances that result in higher correlations between the model state and observed variables (e.g., Johnson et al. 2015). Despite the theoretical advantages to EnKF, subjective and objective comparisons of CAM ensembles from past SFEs did not find that those using EnKF performed any better than other data assimilation methods (Jirak et al. 2015). Thus, more work is needed to optimize EnKF for CAM ensemble applications. However, this experiment was not as controlled as the others, because aspects of the subset of configurations other than the data assimilation methods also differed.

6) GSD radar versus CAPS radar assimilation—Two methods for assimilating radar data were compared. One used ARPS-3DVAR and the other used the Digital Diabatic Filter Initialization (DDFI; Benjamin et al. 2016) system used in the HRRR. It was planned to include a core member configured the same as HRRR36, but using the ARPS-3DVAR system to generate the ICs. Because of time constraints, the core member planned for this experiment was not ready for implementation in real time. Thus, this experiment was not conducted.
7) Microphysics sensitivities—Using the five members of the mp subset, the impact of different microphysics parameterizations on forecast storm structure and evolution was examined. This experiment has been conducted in SFEs since 2010 (e.g., Clark et al. 2012, 2014), and through the participation of microphysics scheme developers each year, parameterizations have been improved and valuable interactions have occurred with forecasters and modelers.
8) Ensemble size experiment—A comparison of ensembles with equal contributions of NMMB and ARW members using 2, 4, 6, 10, and 20 members was conducted to examine the impact of ensemble size. The ensembles used combinations of members from the caps-nmmb, nssl-nmmb, and s-phys-norad subsets. While very large ensembles (e.g., hundreds of members) would be ideal if computational expense were not an issue, the "optimal" ensemble size is generally considered one in which only relatively small gains are achieved by adding additional members. Using a 50-member CAM-ensemble, Schwartz et al. (2014) found that only small gains in precipitation forecast skill are attained after about 20 members and argue that in an operational setting with limited computational resources, sizes greater than 10 would be difficult to justify. Clark et al. (2011) also found that a CAM ensemble of around 10 members has similar quantitative precipitation forecast (QPF) skill to larger ensembles and point out that the optimal number of members varies as a function of forecast length and spatial scale.

CLUE POSTPROCESSING. In past SFEs, the members from each of the unique CAM ensembles contributed to the HWT were postprocessed[2] by each collaborator using their own software. Furthermore, some collaborators, such as CAPS, provided separate sets of postprocessed files containing ensemble-derived fields (e.g., probabilities, ensemble maximum, and ensemble mean). Thus, ingesting the datasets into the HWT workstations required different procedures to account for different file formats, fields, and grids. Furthermore, combining ensemble members from different contributors was cumbersome and rarely done, since it required an extra regridding step before computing any ensemble-derived field. Thus, standardizing the postprocessing procedure was one of the most important aspects of the CLUE since it streamlined the workflow and allowed for consistent postprocessed fields, visualization, and verification.

To standardize the postprocessing, NSSL worked closely with scientists at the Developmental Testbed Center (DTC) and NCEP's Environmental Modeling Center (EMC) to modify the most recent version of the Unified Post-Processor (UPP) software, which is maintained by the DTC (information on the most recent version is available at www.dtcenter.org/upp/users/index.php). The UPP was modified to output a set of 107 fields from each CLUE member in gridded binary (grib2) format over a 3-km grid-spacing CONUS domain. The fields match the two-dimensional fields output by the operational HRRR and were chosen because of their relevance to a broad range of forecasting needs, including aviation, severe weather, and precipitation. Additional output fields, which were requested by NCEP's Weather Prediction Center (WPC), SPC, and Aviation Weather Center (AWC), were also included. This special version of the UPP was distributed by NSSL to collaborators in February 2016 to allow time for testing and implementation. All contributors were asked to supply all 107 fields but were also allowed to add additional diagnostics based on their own research interests. The online supplement contains a table listing all postprocessed fields.

CLUE RESULTS. Given the sheer volume of data composing the CLUE, it is impossible to present results in this article from each of the experiments. Additionally, active research is still being conducted to examine several of the CLUE datasets. For example, DTC Visitor Program projects (www.dtcenter.org/visitors/) are currently under way, examining the value of radar data assimilation using object-based verification methods, as well as the impact of mixed physics in the CLUE. Preliminary findings and results from the 2016 SFE, including some preliminary CLUE results, can be found in Clark et al. (2016). It is important to recognize that annual HWT assessment activities typically include

[2] Postprocessing refers to the procedure used to convert raw model output to standard grids and pressure levels, as well as to compute diagnostic quantities (e.g., convective available potential energy and storm relative helicity).

a combination of subjective and objective evaluation methods (Kain et al. 2003), which together provide a more complete picture of the potential utility of new forecast techniques in an operational environment. However, given the space limitations, this section will focus on results from the single versus multicore CLUE experiments as an example of the research enabled within the CLUE framework.

Single versus multicore experiment: Severe weather verification. Objective verification of four ensemble subsets was conducted for severe weather occurrence, which included 1) NMMB, a 10-member, single-physics NMMB ensemble with perturbed ICs/LBCs; 2) ARW, a 10-member, single-physics ARW ensemble with the same perturbed ICs/LBC as NMMB; 3) $MIX_{10\text{-mem}}$, a combination of five of the NMMB and ARW members; and 4) $MIX_{20\text{-mem}}$, a combination of all 10 NMMB and ARW members. Complete datasets from these ensembles were analyzed for each day the SFE operated (24 days; 2 May–3 June, excluding weekends/holidays).

To verify severe weather (defined as a tornado, damaging winds, or large hail), ensemble-derived severe weather probabilities were computed by considering extreme values of hourly maximum updraft helicity (UH; e.g., Kain et al. 2010) as severe storm proxies following the "surrogate severe" approach outlined by Sobash et al. (2011, 2016b). This approach has been increasingly utilized for verifying CAM-based forecasts of severe weather (e.g., Schwartz et al. 2015a,b; Sobash et al. 2016a; Gallo et al. 2016; Loken et al. 2017; Dawson et al. 2017). The basic idea behind the surrogate severe approach is that "extremes" in simulated storm diagnostics are strongly correlated with observed severe weather. However, given the inherent uncertainty associated with convection forecasts at 12–36-h lead times, coarsened grids and spatial smoothing must be applied to account for timing and displacement errors. Furthermore, the skill and reliability of surrogate severe forecasts are heavily dependent on the threshold or percentile chosen to represent extremes, as well as the amount of smoothing applied. Thus, in the methods described below, a range of UH thresholds and smoothing levels are chosen, which are known to produce reliable forecasts based on previous work.

For application of the surrogate severe approach, the maximum UH at each grid point was computed

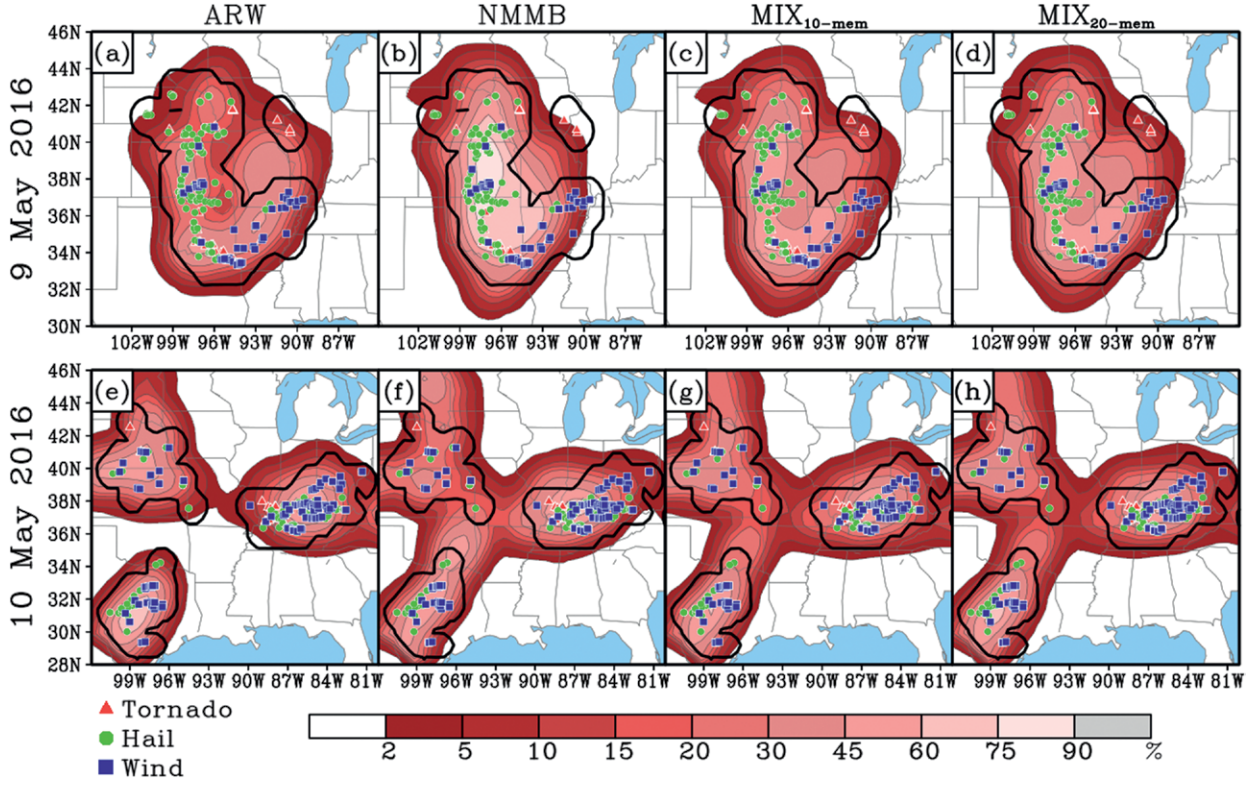

Fig. 2. SSPFs (shaded) using σ = 80 km and p = 0.99 for ensemble forecasts initialized at 0000 UTC 9 May 2016 and valid over forecast hours 13–36 for the ensemble subsets (a) ARW, (b) NMMB, (c) $MIX_{10\text{-mem}}$, and (d) $MIX_{20\text{-mem}}$. (e)–(h) As in (a)–(d), respectively, but for 10 May 2016. Locations of storm reports are overlaid with a legend indicating the type of report at the bottom left. The thick black contour indicates the area within 40 km of any storm report.

over the 24-h period 1200–1200 UTC (forecast hours 13–36) for each ensemble member. Then, for each member, these maximum UH values were remapped onto the 81-km NCEP 211 grid by assigning each 81-km grid box the maximum value of UH out of all 3-km grid points within the 81-km boxes. This methodology is consistent with the SPC operational day 1 convective outlook, which provides categorical and probabilistic forecasts for the 1200–1200 UTC time period and represents severe weather threats within 25 mi (~40 km) of a point. Next, severe weather probabilities [hereafter, surrogate severe probabilistic forecasts (SSPFs)] were computed by finding the ratio of members with UH greater than or equal to a specified percentile p and then applying a two-dimensional Gaussian filter to these ratios. The UH percentiles were computed separately for the set of members in each ensemble subset with the same model core using the distribution of UH values from the 81-km grids over all 24 cases. The percentiles, rather than thresholds, were used to avoid giving more weight to ensemble members with climatologically higher values of UH in the computation of SSPFs. In this dataset, the NMMB tended to have slightly higher UH than ARW (e.g., at $p = 0.99$, the UH values in NMMB and ARW were 152 and 141 $m^2 s^{-2}$, respectively).

The percentiles from 0.80 to 0.998 in increments of 0.02 (100 unique percentiles) were examined, and for each percentile, a range of standard deviations σ in the Gaussian filter from 40 to 300 km in increments of 5 km were tested (i.e., 53 unique σ values). Physically, 1σ can be thought of as the radius containing 68% of the Gaussian kernel weights. Thus, for each case and ensemble subset, there were $100 \times 53 = 5{,}300$ sets of SSPFs. Examples of these SSPFs using $\sigma = 80$ km and $p = 0.99$ for 9 and 10 May 2016 along with the verifying storm reports are shown in Fig. 2. To verify the SSPFs, preliminary observed storm reports from SPC (accessible at www.spc.noaa.gov/climo/reports/) were mapped onto the same 81-km grid as the SSPFs. Any grid box with one of more reports over the 1200–1200 UTC time period was assigned 1 while boxes with zero reports were assigned 0. Verification metrics were computed over the masked area displayed in Fig. 3, which was chosen to limit verification to land and near-coastal areas, as well as to eliminate the Intermountain West, where storm reports and precipitation estimates are not as reliable.

Three metrics are used for objective verification. 1) Area under the relative operating characteristic curve (AUC; Mason 1982) is computed by plotting the probability of detection (POD) versus the

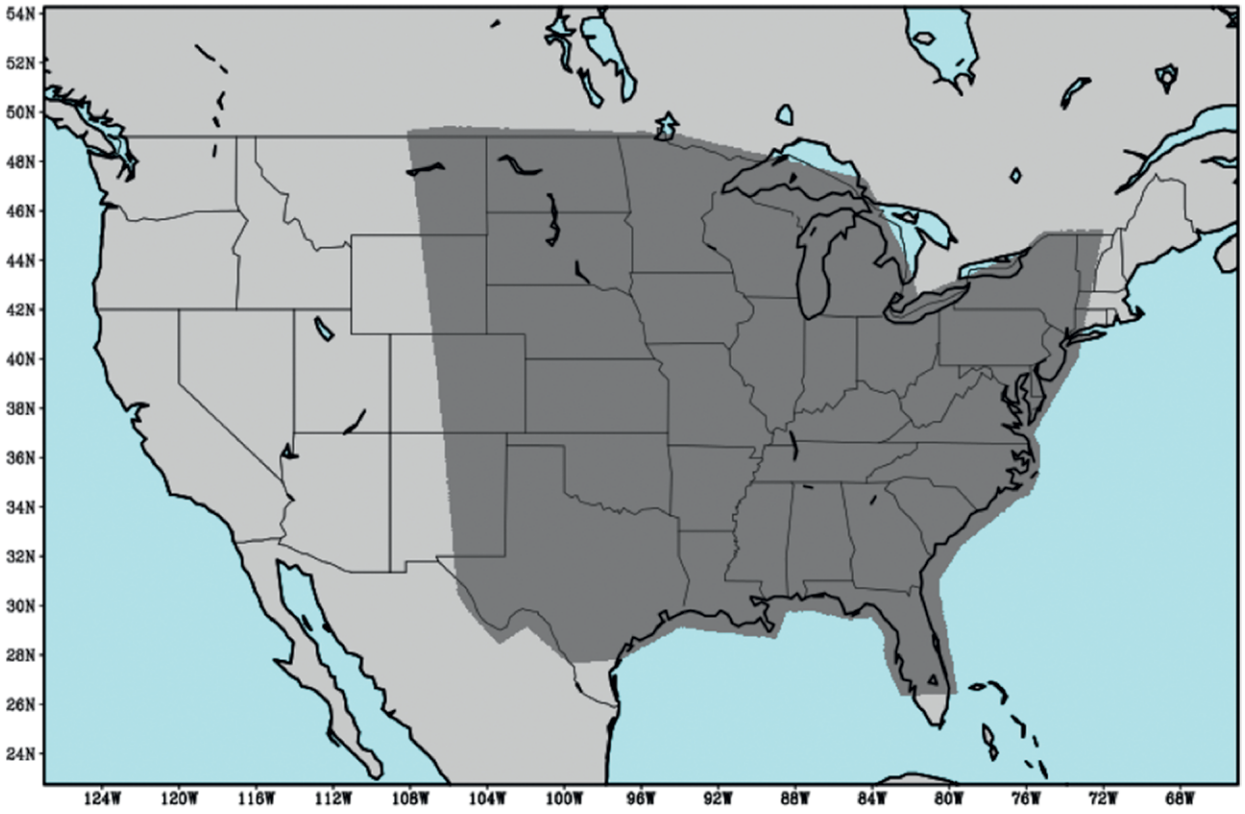

Fig. 3. Area over which verification metrics were computed.

probability of false detection (POFD) for a range of probabilistic thresholds (herein, 2% and 5%–95% in increments of 5% are used). The area under the curve connecting each POD–POFD pair is computed using a trapezoidal approximation (e.g., Wandishin et al. 2001). The AUC measures the ability of the forecast system to discriminate between events and nonevents. A value of 1.0 is considered a perfect AUC, while 0.5 and below is considered to have no skill. 2) Fractions skill score (FSS; Roberts and Lean 2008) is calculated by computing the mean-square error (MSE) of the SSPFs relative to "practically perfect" observations (e.g., Hitchens et al. 2013), which are constructed by applying a Gaussian filter with σ = 120 km to the 81-km grid of storm reports. The MSEs of the SSPFs are normalized by a worst-case reference forecast and subtracted from 1.0 to get the FSS [see Eqs. (3)–(5) in Sobash et al. (2011)]. The FSS ranges from 0 (no skill) to 1 (perfect forecast). 3) The reliability component of the Brier score (BS_{rely}; Brier 1950; Murphy 1973) is computed by taking the squared difference of the probabilities within specified bins and their corresponding observed frequencies [see Eq. (2) in Atger (2003)]. The BS_{rely} essentially measures how closely the points within a reliability diagram follow the perfect reliability line, where the squared error for each point is weighted according to the number of forecasts within each probability bin. Lower BS_{rely} indicates increasing reliability, with BS_{rely} = 0 indicating perfect reliability. These three metrics were chosen because they are very well known and provide complementary information on discriminating ability (AUC), forecast accuracy (FSS), and reliability (BS_{rely}).

Each skill metric for each ensemble is presented as a function of σ and the UH percentile in Fig. 4. The metrics behave quite differently in terms of where the best scores fall within the σ–UH percentile phase space. AUC has the highest scores at relatively low σ values (60–100 km) and UH percentiles (0.82–0.86), FSS maximizes at higher σ (150–180 km) and UH percentiles (0.92–0.94), and BS_{rely} is best at the highest σ (240–300 km) and UH percentiles (0.95–0.96). For reference, in each panel in Fig. 4, the UH percentile at which the number of surrogate severe storm reports is approximately equal to the number of observed severe reports over all cases is shown by the turquoise dashed line (p = 0.974; i.e., bias = 1.0). Thus, AUCs maximize at UH percentiles associated with biases well above 1.0. In fact, the biases at these lower ranges of UH percentiles range from 6.0 to 7.8 (not shown). The high biases associated with the maximum AUCs are not surprising because AUC does not account for bias or reliability. Furthermore, for rare-event forecasts, increasing the number of forecast events almost always acts to increase the POD more than the POFD, thereby increasing the AUC, because correct negatives so heavily weight the POFD. For FSS and BS_{rely}, the scores maximize at UH percentiles closer to bias = 1.0 than AUC. For BS_{rely}, it may seem intuitive that the best reliability would occur when bias = 1.0; however, underdispersion causes probabilities to be too high, and the additional spatial uncertainty provided by a bias slightly higher than 1.0 along with very strong smoothing apparently achieves the best reliability.

For AUC and FSS, the $MIXED_{10\text{-mem}}$ and $MIXED_{20\text{-mem}}$ have slightly higher maximum scores than ARW and NMMB, which are very similar to each other. The best BS_{rely} values are nearly identical among the four ensembles. To evaluate whether any of the differences in maximum scores were significant, the resampling approach of Hamill (1999) was utilized and it was found that none of the differences between ensembles were significant at α = 0.05. Thus, although the multicore approach has slightly higher scores than the single-model approach for severe weather forecasting, a larger sample is necessary to determine whether these differences can be attributed to more than just randomness. Continuation of CLUE-related experiments in subsequent years will contribute to larger data samples and lead to more robust statistical results.

Single versus multicore experiment: QPF verification. Similar to severe weather, accurate precipitation forecasting is notoriously difficult for numerical weather prediction (NWP) models (e.g., Carbone et al. 2002; Roebber et al. 2004), but CAM-based systems have led to major improvements in QPFs over convection-parameterizing models (e.g., Clark et al. 2009, 2010a, 2012; Weisman et al. 2008; Iyer et al. 2016). In many ways, QPF verification is simpler than severe weather, because more reliable and higher-resolution observational precipitation datasets exist (e.g., NCEP's Stage IV dataset), and QPFs are directly output from models and thus do not require surrogates like severe weather. Although these differences allow QPF verification to be reliably performed at higher spatial and temporal resolution than severe weather, to compare QPF and severe weather performance, verification is performed at the same scale as severe weather. Follow-up work will perform QPF verification at higher spatial and temporal resolution.

To perform QPF verification analogously to severe weather, 24-h accumulated precipitation from

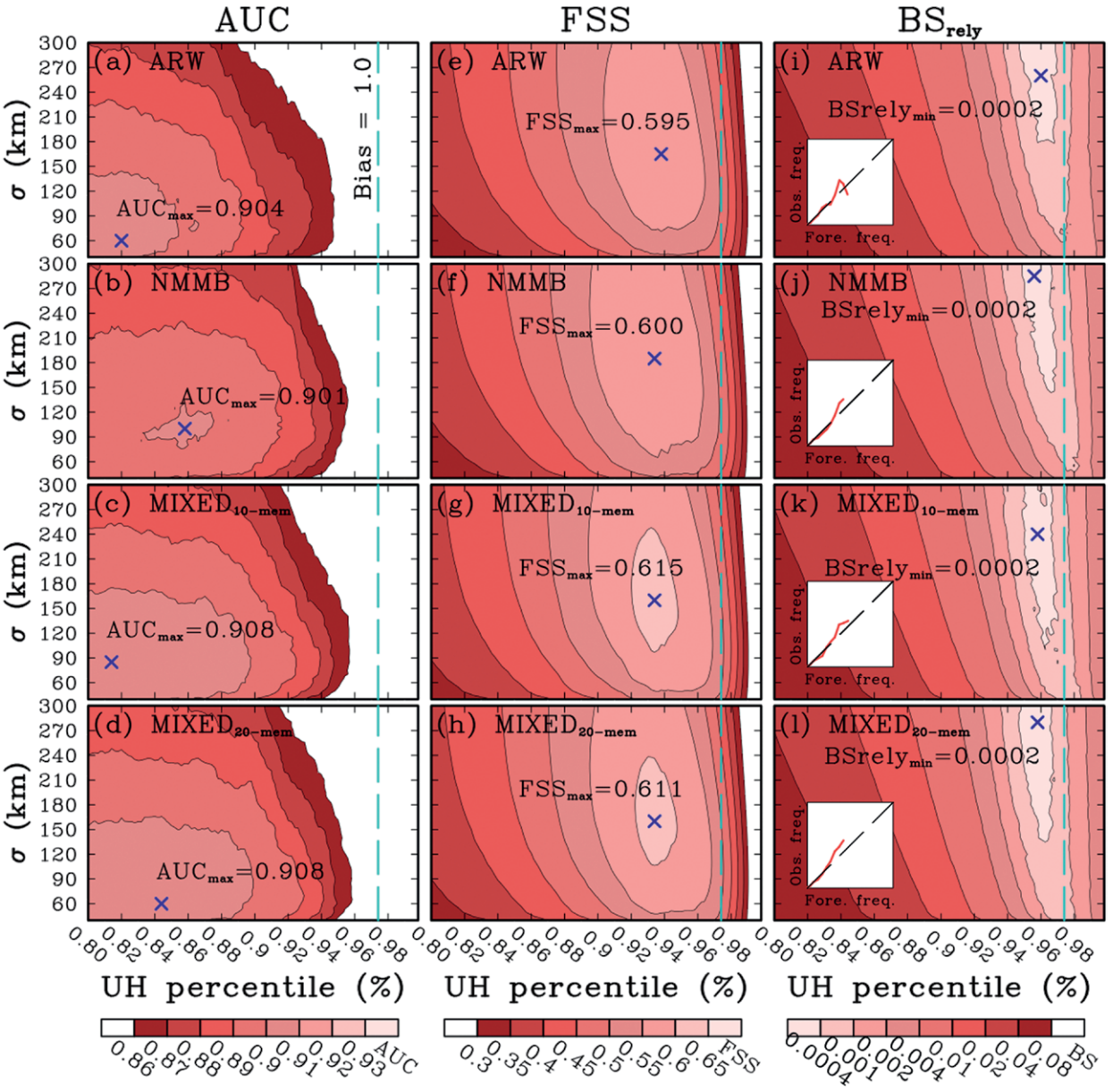

FIG. 4. AUC as a function of σ and UH percentile for the ensembles (a) ARW, (b) NMMB, (c) MIXED$_{10\text{-mem}}$, and (d) MIXED$_{20\text{-mem}}$. (e)–(h) As in (a)–(d), respectively, but for FSS. (i)–(l) As in (a)–(d), respectively, but for BS$_{rely}$. In each panel, a blue × marks the best score, which is indicated in the text; the vertical dashed turquoise line marks the UH percentile at which bias = 1 (i.e., the number of surrogate severe reports approximately matches the number of observed reports). In (i)–(l), the reliability diagrams are shown corresponding to σ and UH percentile at which BS$_{rely}$ is minimized.

NCEP's 4-km grid-spacing Stage IV dataset (Lin and Mitchell 2005; Nelson et al. 2016) was remapped onto the NCEP 211 grid using the maximum 24-h precipitation amount from all 4-km Stage IV grid points within each 81-km NCEP 211 grid box. Then, the precipitation amount that resulted in the same number of observed severe weather events was found, which was 2.69 in. (1 in. = 2.54 cm). In other words, the total number of 81-km grid boxes with an observed severe weather report was equal to the total number of 81-km grid boxes in which the maximum observed precipitation was 2.69 in. or greater. Then, maximum 24-h accumulated QPFs from each ensemble member were remapped onto the 81-km grid in the same manner as UH, and heavy rainfall probabilities were also computed similarly to UH.

Figure 5 shows that all the precipitation skill metrics computed using the 2.69-in. threshold were noticeably higher than those for UH; thus, these CAM ensembles provide more skillful forecasts of extreme

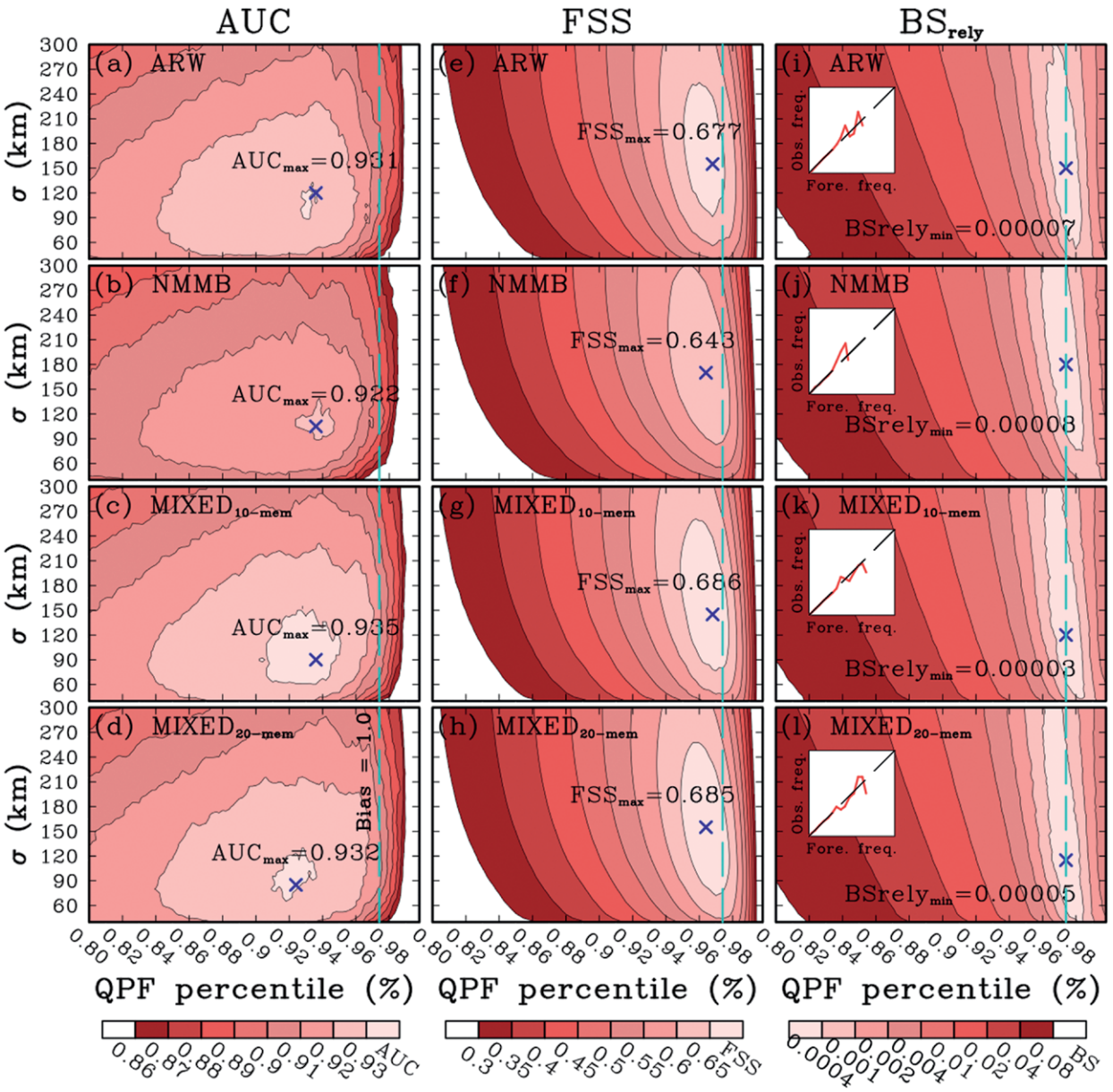

Fig. 5. As in Fig. 4, but for 24-h precipitation forecasts.

rainfall than severe weather. Also, the multicore systems (MIXED$_{10\text{-mem}}$ and MIXED$_{20\text{-mem}}$) had better AUC, FSS, and BS$_{rely}$ results than the single-core systems, but as with UH, none of the differences between the single- and multicore subsets were statistically significant. There are also noticeable differences in σ and the rainfall percentiles at which the metrics are maximized. Namely, relative to UH, there is a shift toward higher percentiles (biases closer to 1.0) and smaller σ (i.e., less smoothing) for QPF. The reason for this shift is not clear, but it is speculated that it is because of differences in the spatial characteristics of heavy rainfall and severe weather. However, further work is needed for substantiation.

CONCLUDING REMARKS. The 2016 SFE marks year 17 of annual SFEs organized by the SPC and NSSL, which aim to accelerate the advancement of new technologies and concepts from research to operations for improving hazardous convective weather prediction. Since 2004, a main focus of SFEs has been on evaluating performance characteristics of CAMs, as well as making advances in creating, importing, processing, verifying, and extracting unique hazardous weather fields, as well as providing analysis and visualization tools for CAMs. With increasing numbers of CAMs contributed to SFEs every year, and the strong community call for evidence-driven decision-making as EMC and the modeling

community configure the first generation of operational CAM-based ensemble prediction systems, a major initiative was started during the 2016 SFE to coordinate and standardize the CAM contributions from each of our external collaborators, so that each group of CAMs could be considered part of one large ensemble termed the Community Leveraged Unified Ensemble (CLUE).

The CLUE was designed to enable up to eight different controlled experiments focused on optimizing CAM ensemble configurations. Results from one of these experiments—single versus multicore ensemble design—were reported upon herein, while research is in progress for several other CLUE experiments. For the single versus multicore results, objective metrics for severe weather forecast skill indicated small differences in forecast skill, with multicore systems having slightly higher scores than those for a single core. Additionally, a 20-member mixed-core ensemble performed almost identically to a 10-member mixed-core ensemble. None of the differences were statistically significant, but with only 24 cases, significance would likely require a larger sample size.

For precipitation verification, probabilistic QPFs were found to be more skillful than those for severe weather when the verification was performed similarly. Additionally, the mixed-core ensembles had slightly better objective metrics for QPF than the single-core ensembles. Future work is planned to perform the precipitation verification from the multicore versus single-core ensemble design experiment at higher spatial and temporal resolution.

HWT has a long and productive history of bringing together different parts of the research, operational, and academic meteorological communities to work collaboratively in a real-time simulated forecasting environment, focusing on severe weather forecasting problems. We envision continuing the CLUE system in subsequent experiments, and there is ample reason to believe that it can further enhance effective engagement between the modeling and operational communities, as well as provide important scientific evidence necessary for informed decision-making, so that future U.S. hazardous weather prediction capabilities are the best possible.

ACKNOWLEDGMENTS. Special thanks for the NWS and NSSL for visitor travel support. We recognize the full support of SPC and NSSL management, the leadership of SPC forecasters in guiding the severe weather component each year, and contributions from many participants who clearly demonstrate the value of collaborative experiments, and whose talents and enthusiasm resulted in a positive learning experience for everyone. SRD, KHK, GJC, and CJM, were provided support by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce. AJC, JSK, LJW, and DAI completed this work as part of their regular duties at the federally funded NOAA/National Severe Storms Laboratory. ILJ and SJW completed this work as part of their regular duties at the federally funded NOAA/Storm Prediction Center. BSF and JRC completed this work as part of their regular duties at I. M. Systems Group, Inc., and the federally funded NOAA/Environmental Modeling Center. BTG was supported by the National Science Foundation (NSF) Graduate Research Fellowship under Grant DGE-1102691, Project A00-4125. CAPS scientists MX, YJ, KWT, KAB, and FK received support from NOAA Collaborative Science and Technology Applied Research (CSTAR) Grant DOC-NOAA NA16NWS4680002, the Warn-on-Forecast project under Grant NA16OAR4320115, NOAA HWT Grant NA15OAR4590186, and NOAA HMT Grant NA15OAR4590159. AK, XD, MG, and JM were supported by NOAA R2O Project NA15NWS4680004. GSR, RAS, and KRF received support from NOAA HWT Award NA15OAR4590191 and the National Center for Atmospheric Research, which is sponsored by the NSF. GT contributed as part of regular duties at NCAR. Work by RDA was performed as part of the Systems Engineering Management and Sustainment contract with the Air Force Life Cycle Management Center, and the Cooperative Research Data Agreement between the 557th Weather Wing and Atmospheric and Environmental Research, Inc. (AER), and internal research funding was provided by AER. CLUE simulations from CAPS used the Texas Advanced Computing Center's (TACC) Stampede System and the University of Tennessee's National Institute for Computational Science's (NICS) Darter System NSSL simulations used TACC's Lonestar5 system. UND used the San Diego Computing Center's (SDCS) Comet system. The Stampede, Darter, Lonestar5, and Comet systems are all part of the Extreme Science and Engineering Discovery Environment (XSEDE; Towns et al. 2014), which is supported by National Science Foundation Grant ACI-1548562. NCAR used the Yellowstone (ark:/85065/d7wd3xhc) supercomputer provided by NCAR's Computational Information System Laboratory, sponsored by the NSF, and ESRL/GSD used NOAA's Jet system.

REFERENCES

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

Atger, F., 2003: Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Mon. Wea. Rev.*, **131**, 1509–1523, https://doi.org/10.1175//1520-0493(2003)131<1509:SAIVOT>2.0.CO;2.

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.

Berner, J., S.-Y. Ha, J. P. Hacker, A. Fournier, and C. Snyder, 2011: Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations. *Mon. Wea. Rev.*, **139**, 1972–1995, https://doi.org/10.1175/2010MWR3595.1.

——, K. R. Fossell, S.-Y. Ha, J. P. Hacker, and C. Snyder, 2015: Increasing the skill of probabilistic forecasts: Understanding performance improvements from model-error representations. *Mon. Wea. Rev.*, **143**, 1295–1320, https://doi.org/10.1175/MWR-D-14-00091.1.

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

Candille, G., 2009: The multiensemble approach: The NAEFS example. *Mon. Wea. Rev.*, **137**, 1655–1665, https://doi.org/10.1175/2008MWR2682.1.

Carbone, R. E., J. D. Tuttle, D. A. Ahijevych, and S. B. Trier, 2002: Inferences of predictability associated with warm season precipitation episodes. *J. Atmos. Sci.*, **59**, 2033–2056, https://doi.org/10.1175/1520-0469(2002)059<2033:IOPAWW>2.0.CO;2.

Clark, A. J., W. A. Gallus Jr., M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Wea. Forecasting*, **24**, 1121–1140, https://doi.org/10.1175/2009WAF2222222.1.

——, ——, and M. L. Weisman, 2010a: Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF Model simulations and the operational NAM. *Wea. Forecasting*, **25**, 1495–1509, https://doi.org/10.1175/2010WAF2222404.1.

——, ——, M. Xue, and F. Kong, 2010b: Growth of spread in convection-allowing and convection-parameterizing ensembles. *Wea. Forecasting*, **25**, 594–612, https://doi.org/10.1175/2009WAF2222318.1.

——, and Coauthors, 2011: Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Mon. Wea. Rev.*, **139**, 1410–1418, https://doi.org/10.1175/2010MWR3624.1.

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

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

——, and Coauthors, 2016: Spring Forecasting Experiment 2016: Preliminary findings and results. NOAA/NSSL/SPC, 50 pp., https://hwt.nssl.noaa.gov/Spring_2016/HWT_SFE_2016_preliminary_findings_final.pdf.

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.

Du, J., J. McQueen, G. DiMego, Z. Toth, D. Jovic, B. Zhou, and H.-Y. Chuang, 2006: New dimension of NCEP Short-Range Ensemble Forecasting (SREF) system: Inclusion of WRF members. Preprints, *WMO Expert Team Meeting on Ensemble Prediction Systems*, Exeter, United Kingdom, WMO, www.wcrp-climate.org/WGNE/BlueBook/2006/individual-articles/05_Du_Jun_WMO06.pdf.

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

——, and Coauthors, 2017: Breaking new ground in severe weather prediction: The 2015 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Wea. Forecasting*, **32**, 1541–1568, https://doi.org/10.1175/WAF-D-16-0178.1.

Gao, J., M. Xue, K. Brewster, and K. K. Droegemeier, 2004: A three-dimensional variational data analysis method with recursive filter for Doppler radars. *J. Atmos. Oceanic Technol.*, **21**, 457–469, https://doi.org/10.1175/1520-0426(2004)021<0457:ATVDAM>2.0.CO;2.

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

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

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.

Hu, M., M. Xue, and K. Brewster, 2006: 3DVAR and cloud analysis with WSR-88D level-II data for the prediction of the Fort Worth, Texas, tornadic thunderstorms. Part I: Cloud analysis and its impact. *Mon. Wea. Rev.*, **134**, 675–698, https://doi.org/10.1175/MWR3092.1.

Iyer, E. R., A. J. Clark, M. Xue, and F. Kong, 2016: A comparison of 36–60-h precipitation forecasts from convection-allowing and convection-parameterizing ensembles. *Wea. Forecasting*, **31**, 647–661, https://doi.org/10.1175/WAF-D-15-0143.1.

Janjić, Z. I., and R. Gall, 2012: Scientific documentation of the NCEP Nonhydrostatic Multiscale Model on the B Grid (NMMB). Part 1: Dynamics. NCAR/TN-4891STR, 75 pp., http://nldr.library.ucar.edu/repository/assets/technotes/TECH-NOTE-000-000-000-857.pdf.

Jankov, I., and Coauthors, 2017: A performance comparison between multiphysics and stochastic approaches within a North American RAP ensemble. *Mon. Wea. Rev.*, **145**, 1161–1179, https://doi.org/10.1175/MWR-D-16-0160.1.

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

——, A. J. Clark, J. Correia Jr., K. Knopfmeier, C. Melick, B. T. Gallo, M. Coniglio, and S. J. Weiss, 2015: Spring Forecasting Experiment 2015: Preliminary findings and results. NOAA/NSSL/SPC Rep., 32 pp., https://hwt.nssl.noaa.gov/Spring_2015/HWT_SFE_2015_Prelim_Findings_Final.pdf.

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

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

——, and Coauthors, 2010: Assessing advances in the assimilation of radar data and other mesoscale observations within a collaborative forecasting–research environment. *Wea. Forecasting*, **25**, 1510–1521, https://doi.org/10.1175/2010WAF2222405.1.

Kirtman, B. P., and Coauthors, 2014: The North American Multimodel Ensemble: Phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bull. Amer. Meteor. Soc.*, **95**, 585–601, https://doi.org/10.1175/BAMS-D-12-00050.1.

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, **285**, 1548–1550, https://doi.org/10.1126/science.285.5433.1548.

——, V. Kumar, A. Simon, A. Bhardwaj, T. Ghosh, and R. Ross, 2016: A review of multimodel superensemble forecasting for weather, season climate, and hurricanes. *Rev. Geophys.*, **54**, 336–377, https://doi.org/10.1002/2015RG000513.

Lin, Y., and K. E. Mitchell, 2005: The NCEP stage II/IV hourly precipitation analyses: Development and applications. *19th Conf. on Hydrology*, San Diego, CA, Amer. Meteor. Soc., 1.2, https://ams.confex.com/ams/Annual2005/techprogram/paper_83847.htm.

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.

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

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

Nelson, B. R., O. P. Prat, D.-J. Seo, and E. Habib, 2016: Assessment and implications of NCEP Stage IV quantitative precipitation estimates for product intercomparisons. *Wea. Forecasting*, **31**, 371–394, https://doi.org/10.1175/WAF-D-14-00112.1.

Palmer, T. N., and Coauthors, 2004: Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER). *Bull. Amer. Meteor. Soc.*, **85**, 853–872, https://doi.org/10.1175/BAMS-85-6-853.

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

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1.

Roebber, P. J., D. M. Schultz, B. A. Colle, and D. J. Stensrud, 2004: Toward improved prediction: High-resolution and ensemble modeling systems in

operations. *Wea. Forecasting*, **19**, 936–949, https://doi.org/10.1175/1520-0434(2004)019<0936:TIPHAE>2.0.CO;2.

Schwartz, C. S., G. S. Romine, K. R. Smith, and M. L. Weisman, 2014: Characterizing and optimizing precipitation forecasts from a convection-permitting ensemble initialized by a mesoscale ensemble Kalman filter. *Wea. Forecasting*, **29**, 1295–1318, https://doi.org/10.1175/WAF-D-13-00145.1.

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

——, ——, M. L. Weisman, R. A. Sobash, K. R. Fossell, K. W. Manning, and S. B. Trier, 2015b: A real-time convection-allowing ensemble prediction system initialized by mesoscale ensemble Kalman filter analyses. *Wea. Forecasting*, **30**, 1158–1181, https://doi.org/10.1175/WAF-D-15-0013.1.

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.

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

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

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

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

Stratman, D. R., M. C. Coniglio, S. E. Koch, and M. Xue, 2013: Use of multiple verification methods to evaluate forecasts of convection from hot- and cold-start convection-allowing models. *Wea. Forecasting*, **28**, 119–138, https://doi.org/10.1175/WAF-D-12-00022.1.

Towns, J., and Coauthors, 2014: XSEDE: Accelerating scientific discovery. *Comput. Sci. Eng.*, **16**, 62–74, https://doi.org/10.1109/MCSE.2014.80.

UCAR, 2015: Report of the UCACN Model Advisory Committee. University Corporation for Atmospheric Research, Boulder, CO, 72 pp., www.ncep.noaa.gov/director/ucar_reports/ucacn_20151207/UMAC_Final_Report_20151207-v14.pdf.

Xue, M., D. Wang, J. Gao, K. Brewster, and K. K. Droegemeier, 2003: The Advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation. *Meteor. Atmos. Phys.*, **82**, 139–170.

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

Weisman, M. L., C. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with 0–36-h explicit convective forecasts with the WRF-ARW Model. *Wea. Forecasting*, **23**, 407–437, https://doi.org/10.1175/2007WAF2007005.1.