SPRING FORECASTING EXPERIMENT 2015

Conducted by the

EXPERIMENTAL FORECAST PROGRAM

of the

NOAA HAZARDOUS WEATHER TESTBED

http://hwt.nssl.noaa.gov/Spring_2015/

HWT Facility – National Weather Center
4 May - 5 June 2015

Preliminary Findings and Results

Israel Jirak[1], Adam Clark[2,3], James Correia[1,3], Kent Knopfmeier[2,3], Chris Melick[1,3], Burkely Twiest[2,4], Michael Coniglio[2], and Steven Weiss[1]
(1) NOAA/NWS/NCEP Storm Prediction Center, Norman, Oklahoma
(2) NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma
(3) Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma
(4) School of Meteorology, University of Oklahoma

1. Introduction

The 2015 Spring Forecasting Experiment (SFE2015) was conducted from 4 May – 5 June by the Experimental Forecast Program (EFP) of the NOAA/Hazardous Weather Testbed (HWT). SFE2015 was organized by the Storm Prediction Center (SPC) and National Severe Storms Laboratory (NSSL) with participation from numerous forecasters, researchers, and developers from around the world (see Table 1 in the Appendix) to test emerging concepts and technologies designed to improve the prediction of hazardous convective weather. SFE2015 aimed to address several primary goals that are consistent with the Forecasting a Continuum of Environmental Threats (FACETs) and Warn-on Forecast (WoF) visions:

Operational Product and Service Improvements:

- Explore the ability to generate higher temporal resolution Day 1 convective outlooks than those issued operationally by SPC.
    - 4-h periods for individual severe hazards (tornado, hail, and wind)
    - 1-h periods for all severe hazards
    - Share with and receive feedback from Experimental Warning Program participants
- Explore the ability to generate experimental Day 2 convective outlooks containing probabilistic forecasts for individual hazards (tornado, hail, wind), to provide more specific threat information compared to current operational SPC Day 2 total severe storm outlooks.

Applied Science Activities:

- Compare six different convection-allowing ensembles (SPC SSEO, USAF, NSSL, OU/CAPS SSEF, new CAPS EnKF, and new NCAR EnKF) and identify strengths and weaknesses of the different configurations and initialization/perturbation strategies.
- Examine several convection-allowing ensemble forecasts into Day 2 and assess their guidance for generating convective outlooks, including individual severe hazards.
- Evaluate EMC parallel convection-allowing models (CAMs; HiResW WRF-ARW, HiResW NMMB, and NAM CONUS Nest, HRRR) and compare them to operational versions.
- Compare and assess different approaches in CAMs for predicting hail size.
- Assess the use of environmental filters on updraft helicity for generating tornado probability guidance from the NSSL-WRF ensemble.
- Inspect differences between the SSEF 3DVar and SSEF EnKF reflectivity forecasts during the first few hours after initialization.
- Document characteristics of various microphysics schemes used with the WRF model.
- Assess the impact of horizontal resolution in 2.2-km and 1-km versions of the Met Office Unified Model and compare aspects of their forecast performance with the NSSL-WRF.
- Provide a preliminary assessment of the capability of the NCAR global Model for Prediction Across Scales (MPAS) with variable resolution (~3 km grid-spacing over the CONUS) in generating realistic and operationally useful prediction of convective storms out to Day 5.

This document summarizes the activities, core interests, and preliminary findings of SFE2015. More detailed information on the organizational structure and mission of the HWT, model and ensemble configurations, and information on various forecast tools and diagnostics can be found in the operations plan (http://hwt.nssl.noaa.gov/Spring_2015/HWT_SFE_2015_OPS_plan_final.pdf). The remainder of this document is organized as follows: Section 2 provides an overview of the models and ensembles examined during SFE2015 along with a description of the daily activities, Section 3 reviews the preliminary findings of SFE2015, and Section 4 contains a summary of the preliminary findings.

2. Description

a) Experimental Models and Ensembles

Building upon successful experiments of previous years, SFE2015 focused on the generation of experimental probabilistic forecasts of severe weather valid over shorter time periods than current operational SPC severe weather outlooks. This is an important step toward addressing a strategy within the National Weather Service (NWS) of providing nearly continuous probabilistic hazard forecasts on increasingly fine spatial and temporal scales (i.e., FACETs), in support of the NWS Weather-Ready Nation initiative. As in previous experiments, a suite of new and improved experimental CAM guidance including ensembles was central to the generation of these forecasts. For all of the models, hourly maximum fields (HMFs) of explicit storm attributes such as simulated reflectivity, updraft helicity, updraft speed, and 10-m wind speed, were generated and examined as part of the experimental forecast and evaluation process. More information on these modeling systems is given below.

i. NSSL-WRF and NSSL-WRF Ensemble

SPC forecasters have used output from an experimental 4-km grid-spacing WRF-ARW produced by NSSL (hereafter NSSL-WRF) since the fall of 2006. Currently, this WRF model is run twice daily at 0000 UTC and 1200 UTC throughout the year over a full-CONUS domain with forecasts to 36 hours.

For the second year, the NSSL-WRF ensemble was part of the experimental numerical guidance. This ensemble includes eight additional 4-km WRF-ARW runs that – along with the deterministic NSSL-WRF – comprised a nine-member NSSL-WRF-based ensemble. The additional eight members were initialized at 0000 UTC and use 3-h forecasts from the 2100 UTC NCEP Short Range Ensemble Forecast (SREF) system for initial conditions (ICs) and corresponding SREF member forecasts as lateral boundary conditions (LBCs). The physics parameterizations for each member are identical to the deterministic NSSL-WRF. Although the unvaried physics will have lower spread than a multi-physics ensemble, SPC forecasters and NSSL scientists are very familiar with the behavior of the NSSL-WRF physics, and this configuration will allow for the isolation of spread contributed only by varying the ICs/LBCs.

ii. CAPS Storm-Scale Ensemble Forecast Systems

As in previous years, CAPS provided a 0000 UTC-initialized Storm Scale Ensemble Forecast (SSEF) system, but new to the experiment from CAPS this year was an Ensemble-Kalman Filter-based system (SSEF-EnKF) that assimilated WSR-88D radar reflectivity and radial velocity into a separate ensemble of model forecasts. More details on these two ensemble systems are given below.

The legacy SSEF system had 20 members, which included 12 “core” members that were used for ensemble products. The grid-spacing of the SSEF was reduced from 4-km to 3-km for SFE2015 and, similar to SFE2014, the forecasts extended out to 60 h to support the Day 2 forecasts. As in previous years, the 0000 UTC NAM analyses available on the 12-km grid (AWIPS 218) were used for initialization of control and non-perturbed members and as first guess for the initialization of perturbed members with the IC perturbations coming directly from the NCEP SREF. WSR-88D data, along with available surface and upper air observations, were analyzed using ARPS 3DVAR/Cloud-analysis system.

A separate EnKF-based, 3-km grid-spacing, 12-member ensemble of 60-h forecasts was also produced over the same CONUS domain covered by the SSEF system. Starting at 1800 UTC, a six-hour EnKF cycling process

with 40 WRF-ARW members was performed on a 3-km grid over the CONUS domain. This ensemble was configured with initial perturbations and mixed physics options to provide input for the EnKF analysis. Each member used WSM6 microphysics with different parameter settings. All members also included random perturbations with recursive filtering of ~20 km horizontal correlations scales, with relatively small perturbations (0.5K for potential temperature and 5% for relative humidity). EnKF analysis (cycling), with radar data and other conventional data, was performed from 2300 to 0000 UTC every 15 minutes over the CONUS domain, using the 40-member ensemble as background. A 12- member ensemble forecast (out to 60-h) followed using the last EnKF analyses at 0000 UTC.

iii. SPC Storm Scale Ensemble of Opportunity

The SPC Storm-Scale Ensemble of Opportunity (SSEO) is a 7-member, multi-model and multi-physics convection-allowing ensemble consisting of deterministic CAMs with ~4-km grid spacing available to SPC year-round. This "poor man's ensemble" has been utilized in SPC operations since 2011 with forecasts to 36 hrs from 0000 and 1200 UTC and provides a practical alternative to a formal/operational storm-scale ensemble, which will not be available until 2017, owing to computational limitations in NOAA. All members were initialized as a "cold start" from the operational NAM – i.e., no additional data assimilation was used to produce ICs.

iv. United States Air Force 4-km Ensemble

The U.S. Air Force 557th Weather Wing at Offutt AFB (USAF) ran a real-time 10-member, 4-km grid spacing WRF-ARW ensemble over the CONUS, and these forecast fields were available for examination during SFE2015. Forecasts were initialized at 0000 UTC and 1200 UTC using 6 or 12 hour forecasts from three global models: the Met Office Unified Model (UM), the NCEP Global Forecast System (GFS), and the Canadian Meteorological Center Global Environmental Multiscale (GEM) Model. Diversity in the AFWA ensemble is achieved through IC/LBCs from the different global models and varied microphysics and boundary layer parameterizations. No data assimilation was performed in initializing these runs.

v. NCAR EnKF-based Ensemble

New for SFE2015, NCAR provided a 10-member, CONUS domain, 3-km grid-spacing, EnKF-based ensemble with forecasts to 48 h. This ensemble used NCAR's DART (Data Assimilation Research Testbed) software. The analysis system was comprised of 50 members (with constant physics) that were continuously cycled using the ensemble adjustment Kalman filter (EAKF). New analyses were produced every 6 h with 15-km grid-spacing using the following observational sources: MADIS ACARS, METARs, and radiosondes; NCEP MARINE; CIMMS cloud-track winds; and Oklahoma Mesonet. From this mesoscale background, 10 downscaled 3-km forecasts were initialized daily at 0000 UTC using the same physics as the data assimilation system, but without cumulus parameterization.

vi. UKMET Convection-Allowing Model Runs

Three nested, limited-area high-resolution versions of the Met Office Unified Model (UM) running once per day were provided to SFE2015: two at 2.2 km grid spacing and one at 1.1 km. The operational 2.2-km version had 70 vertical levels across a slightly sub-CONUS domain. Taking its initial and lateral boundary conditions from the 00Z 17-km horizontal grid-spacing global configuration of the UM, the 2.2-km model initialized without additional data assimilation and ran out to 48 hours. This model configuration included a 3D turbulent mixing scheme using a locally scale-dependent blending of Smagorinsky and boundary layer mixing

schemes, stochastic perturbations were made to the low-level resolved-scale temperature field in conditionally unstable regimes (to encourage the transition from subgrid to resolved scale flows) and the microphysics was single moment. Partial cloudiness was diagnosed assuming a triangular moisture distribution with a width that is a universally specified function of height only. The parallel version of the 2.2-km model was run with a new parameterization of partial cloudiness. This builds on the prognostic scheme used in the Met Office global model ("PC2") but includes an additional parameterization of subgrid moisture variability that is linked to the PBL turbulence. There was no convection parameterization in any of the high resolution UM configurations.

The 1.1-km horizontal resolution version of the UM was nested within the 2.2-km model and ran over a 1300 km by 1800 km domain centered on Oklahoma. The 1.1-km model took its initial and lateral boundary conditions from the T+3 step of the 0000 UTC 2.2-km run, thus reducing spin-up time within the 1.1-km model, and ran out to 33 hours. The 1.1-km model used the same vertical levels, planetary boundary scheme, and microphysics scheme as the 2.2-km run, and was primarily examined to assess sensitivity to horizontal resolution.

vii. NCAR Model for Prediction Across Scales (MPAS)

Another new modeling system for SFE2015 was the NCAR Model for Prediction Across Scales (MPAS; Skamarock et al. 2012). MPAS produced daily 0000 UTC initialized forecasts at 3-km grid-spacing over the CONUS with forecasts to 120 h (5 days). The MPAS horizontal mesh was based on Spherical Centriodal Voronoi Tesselations (SCVTs). These meshes allowed for both quasi-uniform discretization of the sphere and local refinement with smoothly varying mesh spacing between regions with differing resolutions. The smoothly varying mesh eliminates the major problems encountered with mesh transitions in forecast systems using traditional grid-nesting.

b) Daily Activities

SFE2015 activities were focused on forecasting severe convective weather at two separate desks, one forecasting individual hazards and the other forecasting total severe, with different experimental forecast products being generated at different temporal resolutions. Forecast and model evaluations also were an integral part of daily activities of SFE2015. A summary of forecast products and evaluation activities can be found below while a detailed schedule of daily activities is contained in the appendix.

i. Experimental Forecast Products

Similar to previous years, the experimental forecasts continued to explore the ability to add temporal specificity to longer-term convective outlooks. One desk mimicked the SPC operational Day 1 convective outlooks by producing separate probability forecasts of large hail, damaging wind, and tornadoes within 25 miles (40 km) of a point valid 1600 UTC to 1200 UTC the next day. On the other desk, a separate Day 1 forecast was made for total severe (combined hail, wind, and tornado) probabilities valid over the same period.

Each desk then manually stratified their respective Day 1 forecasts into periods with higher temporal resolution. Individual hazard probability forecasts of large hail, damaging wind, and tornadoes were generated for two four-hour periods: 1800-2200 UTC and 2200-0200 UTC. As an alternative way of stratifying the Day 1 forecast, the other desk generated hourly probability forecasts of total severe valid from 1800-0000 UTC. The goal of testing these two methods was to explore different ways of introducing probabilistic severe weather

forecasts on time scales that are currently addressed operationally with primarily non-scheduled (as needed) deterministic forecast products (e.g., mesoscale discussions and severe thunderstorm/tornado watches) and to begin to explore ways of seamlessly merging probabilistic severe weather outlooks with probabilistic severe weather warnings as part of the NOAA FACETs (Rothfusz et al., 2014) and Warn-on-Forecast initiatives (Stensrud et al. 2009).

In addition to the complete suite of observational and model data available in SPC operations, first-guess guidance for individual severe weather hazards was available to assist in generating the higher temporal resolution outlooks. Calibrated guidance for the individual hazards, as derived from the SREF (environment information) and SSEO (explicit storm attributes; Jirak et al. 2014), was available in 3-h periods. The 1600-1200 UTC human forecasts for the SPC Desk were also temporally disaggregated (Jirak et al. 2012) into the 4-h periods (1800-2200 UTC and 2200-0200 UTC) using SSEO guidance to provide additional timing information for the four-hour periods.

Participants were also able to create their own short-time-window forecasts (i.e., human-generated forecast ensemble) on Google Chromebooks using a web-based tool to draw severe weather probability lines. The participant forecasts were compared to one another and to a "control" forecast issued by the lead forecaster at each desk using N-AWIPS.

Severe weather forecasts were also generated for Day 2 to explore the feasibility of issuing forecasts of individual severe storm hazards beyond Day 1, where current SPC operational forecasts for Day 2 (and beyond) only consider probabilities of total severe. In particular, operational and experimental CAM guidance were examined to assist in the individual hazard forecasts for Day 2. Forecasts for total severe were also generated for Day 2 and/or Day 3 if time and interest allowed. This provided an opportunity to explore convection-allowing guidance from MPAS into Day 3.

Finally, each desk examined observational trends and morning/afternoon model guidance to update (or add to) their respective short-time-window forecasts made earlier in the day. The individual hazard forecasts were updated for the 2200-0200 UTC period while the hourly total severe forecasts valid from 2100-0000 UTC were updated with two additional hourly forecasts issued through 0200 UTC. These forecasts were digitized and shared with the Experimental Warning Program (EWP) for use in preparation for their daily activities.

ii. Forecast and Model Evaluations

While much can be learned from examining model guidance and utilizing it to help create experimental forecasts in real time, an important component of SFE2015 was to look back and evaluate the forecasts and model guidance from the previous day. In particular, the individual-period forecasts and the first-guess guidance were compared to observed radar reflectivity, preliminary local storm reports (LSRs) of severe weather, NWS warnings, and radar-estimated hail sizes over the same time periods. The SFE participants provided their subjective evaluations of the strengths and weaknesses of each of the forecasts. This evaluation also included examining and comparing first-guess guidance, preliminary, and final forecasts. The goal was to assess the skill of the first-guess guidance and the human-generated forecasts for all periods.

In addition, experimental forecasts were evaluated objectively in near real-time using Critical Success Index (CSI) and Fractions Skill Score (FSS) based on the LSRs as the observed verification database. CSI was calculated at two fixed-probability thresholds (5% and 15%) used in SPC operational outlooks. Comparisons of results from the experimental forecasts to the first-guess automated fields were also made. The utility of the

statistical verification metrics in assessing forecast skill for both longer and shorter time periods was explored by comparing the scores to the subjective evaluations by the participants.

Model evaluations for SFE2015 focused on the general accuracy of the forecasts in predicting severe convection explicitly, as well as the impact of various physics options on the forecasts. There were also evaluations of new hail proxies available in WRF-ARW, and comparisons of the Met Office CAMs and the NSSL-WRF using model soundings in the pre-convective environment.

Additionally, convection-allowing ensembles from 0000 UTC were compared and evaluated on their ability to provide useful severe weather guidance. The objective component of these evaluations focused on forecasts of simulated reflectivity compared to observed radar reflectivity while the subjective component examined forecasts of HMFs relative to LSRs of hail, wind, and tornadoes. In addition, some of the 0000 UTC ensembles had forecasts extending into Day 2, which allowed for a comparison of guidance on Day 2 versus Day 1 that were valid for the same time period. This was done to assess the utility of CAM ensemble guidance into Day 2 and to see if the guidance improved closer to the potential event time.

3. Preliminary Findings and Results

a) Evaluation of Hourly Total Severe Forecasts

The preliminary (issued in the morning; valid 1800-0000 UTC) and final (issued in the afternoon; valid 2100-0200 UTC) hourly probabilistic forecasts issued by the lead forecaster were subjectively rated from 1-10 (with 10 being the highest rating). The evaluation was weighted heavily toward LSRs, but severe weather watches/warnings and observed composite reflectivity were also examined to provide a more holistic evaluation process. The most notable aspect of these ratings was that the afternoon update forecasts (valid 2100-0000 UTC) were generally rated higher than the corresponding preliminary forecasts issued in the morning (Fig. 1). This was likely related to the availability of updated real-time observational data, including satellite and radar imagery, as the forecast valid time approached when making the final forecasts in the afternoon.

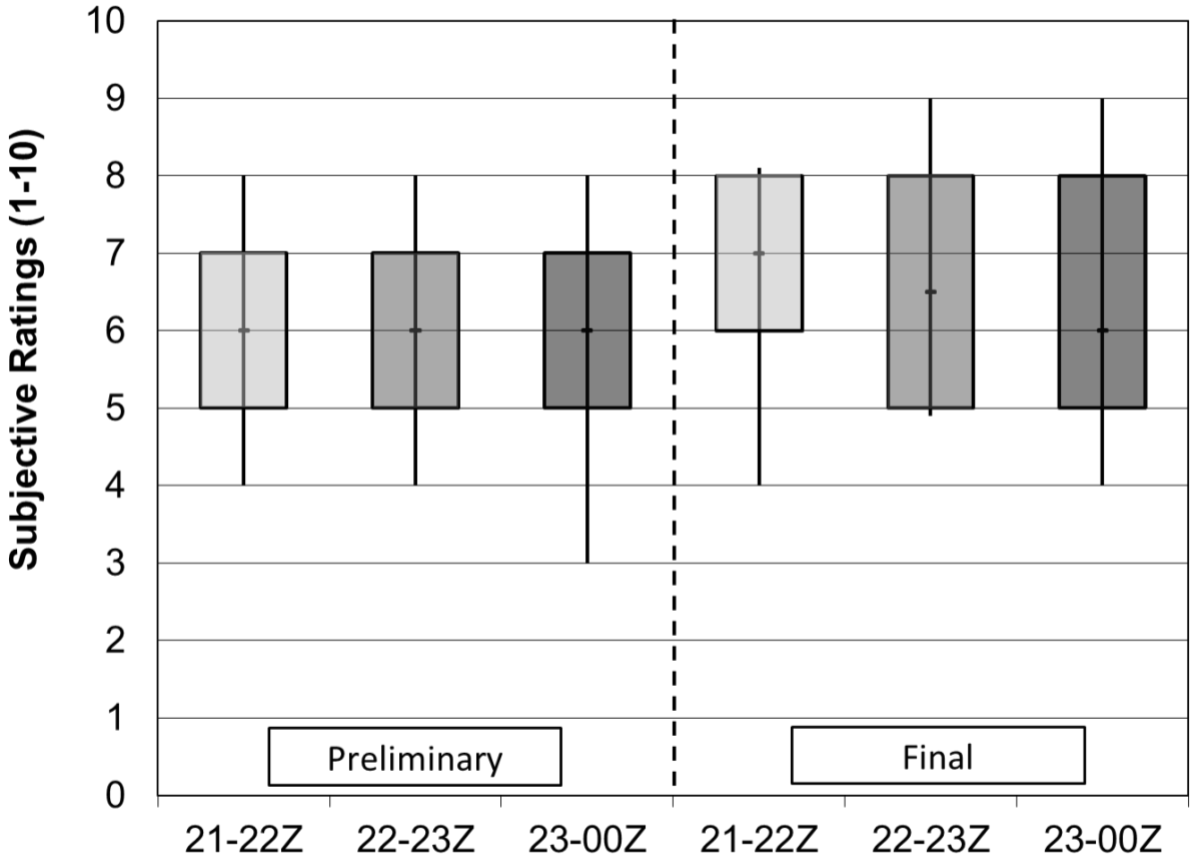

Figure 1. Distribution of subjective ratings (1 to 10) for the preliminary hourly experimental forecasts (left; 2100-0000 UTC) issued at 1600 UTC compared to the final experimental forecasts (right; valid 2100-0000 UTC) issued at 2100 UTC. The boxes comprise the interquartile range of the distributions and the tips of the whiskers extend to the 10th and 90th percentiles.

b) Evaluation of 4-h Forecasts of Severe Hazards

The preliminary 4-h severe hazard experimental forecasts (i.e., tornado, hail, and wind) were compared with the temporally disaggregated first-guess guidance. The first-guess probabilities for the 4-h periods were generated using the temporal disaggregation technique (Jirak et al. 2012) by incorporating the full-period hazard outlook to constrain and scale the magnitude and spatial extent of the 4-h SSEO neighborhood probabilities of severe proxy variables (i.e., updraft helicity for tornadoes, updraft speed for hail, and 10-m wind speed for wind). The first-guess guidance was available to the participants when making the preliminary forecasts. During the 1800-2200 UTC period, the experimental hazard forecasts were commonly rated similar to the first-guess guidance (Fig. 2; rating of 0 on a scale of -3 to +3). For the 2200-0200 UTC period, however, the experimental hazard forecasts were generally rated as an improvement over the earlier first-guess guidance, although most ratings suggested marginal improvement (i.e., 0 to +1).

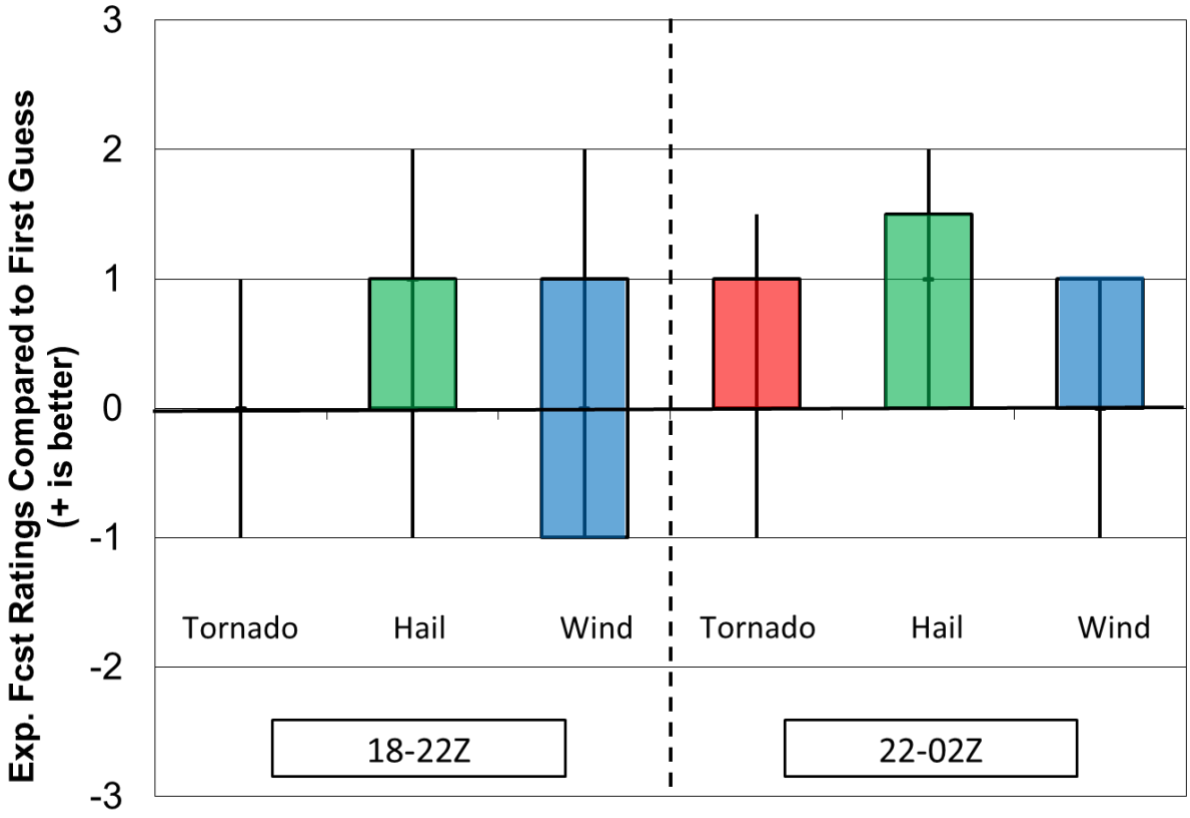

Figure 2. As in Fig. 1, except for the distribution of subjective ratings (-3 to +3) of the experimental forecasts compared to the first-guess guidance for tornado (red), hail (green), and wind (blue) during the 1800-2200 UTC (left) and 2200-0200 UTC (right) periods.

The preliminary and final tornado, wind, and hail forecasts for the 2200-0200 UTC period were subjectively compared to determine the relative value of the afternoon forecast updates (Fig. 3). Overall, updating the forecasts in the afternoon generally resulted in similar or better forecast quality. Although the improvement was marginal (i.e. 0 to +1 rating) and often provided later confirmation of the existing threat, it was rare for the afternoon updates to result in degraded forecast quality.

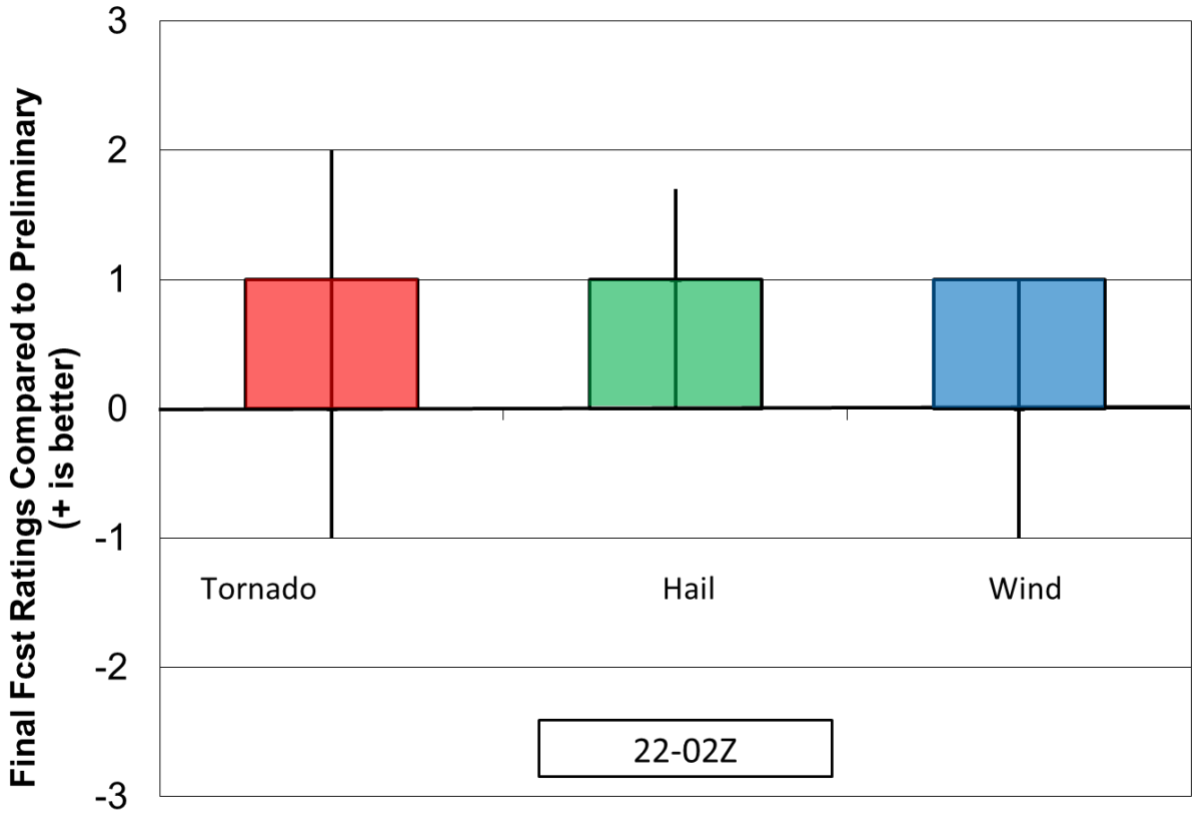

Figure 3. As in Fig. 1, except for the distribution of subjective ratings (-3 to +3) for the final forecast compared to the preliminary forecast for tornado (red), hail (green), and wind (blue) during the 2200-0200 UTC period.

c) Comparison of Convection-Allowing Ensembles

Forecasts from six different 0000 UTC-initialized ensembles were available for examination in SFE2015, providing an opportunity for comparisons among multiple convection-allowing ensemble designs with varying degrees of complexity and diversity. There were two primary components to this comparison of the convection-allowing ensembles: 1) objective evaluation of neighborhood probabilities of reflectivity ≥40 dBZ and 2) subjective verification of ensemble HMFs relative to LSRs.

The fractions skill score (FSS; Roberts and Lean 2008; Schwartz et al., 2010) was calculated for the ensemble neighborhood probability of 1-km AGL simulated reflectivity ≥40 dBZ using observed radar reflectivity for verification. The ensembles had a similar distribution of daily FSS over the five-week SFE2015 (Fig. 4) with the SSEF EnKF showing the lowest skill overall. While the AFWA and NCAR ensembles tended to produce more forecasts of lower skill than the SSEO, NSSL, and SSEF, the median and upper quartile values were similar among the five best-performing ensembles.

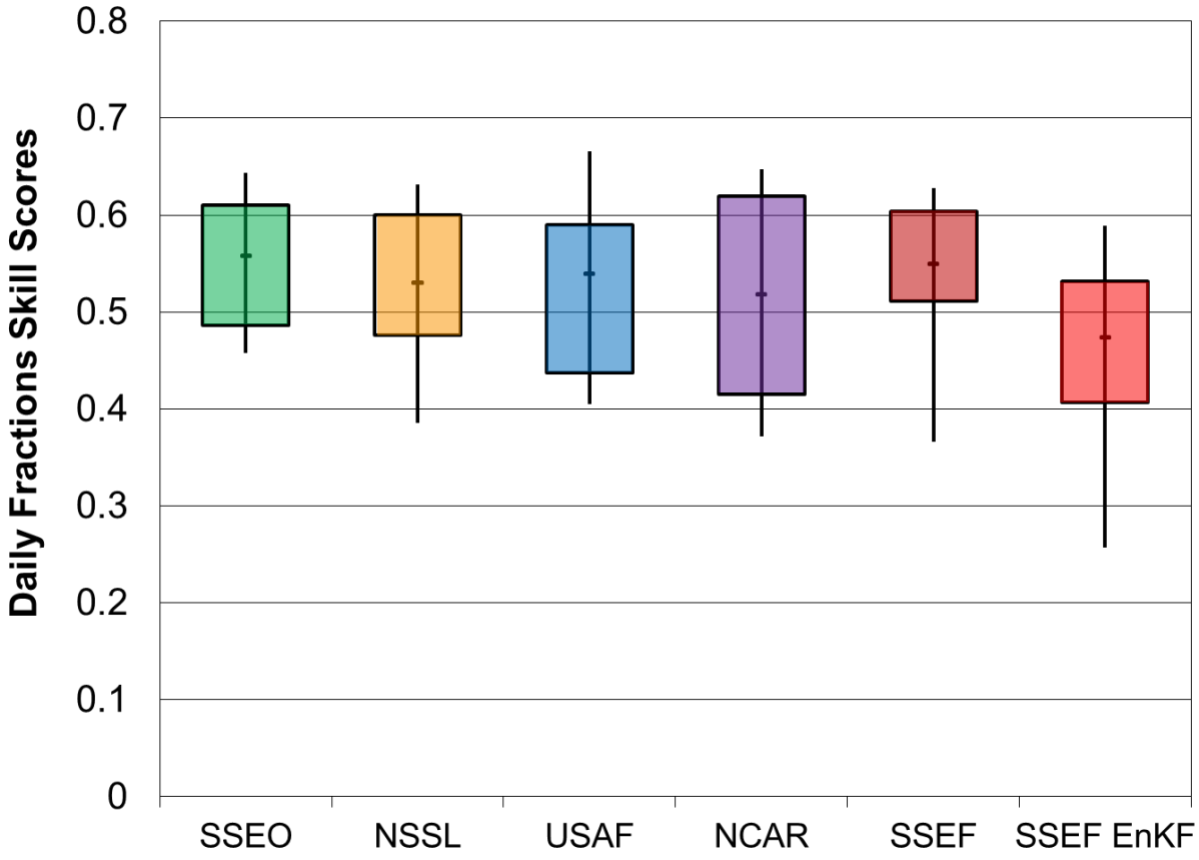

Figure 4. As in Fig. 1, except for the distribution of daily FSS for ensemble neighborhood reflectivity forecasts from the six different convection-allowing ensembles.

When looking at the FSS for reflectivity by forecast hour (Fig. 5), some additional characteristics are apparent regarding the ensembles. The SSEF EnKF generally had the lowest FSS throughout the forecast cycle. Although the SSEO had the highest cumulative FSS from 13 to 22 hours into the forecast, it finished the forecast at 36 hours with the lowest FSS. Aside from the SSEF EnKF, the other five ensembles generally had similar performance during the peak convective period of 2000-0200 UTC. Interestingly, even with very different configurations and methods of initialization, the ensembles appeared to perform statistically similarly during the spring, suggesting that optimizing CAM ensemble design strategies require further research.

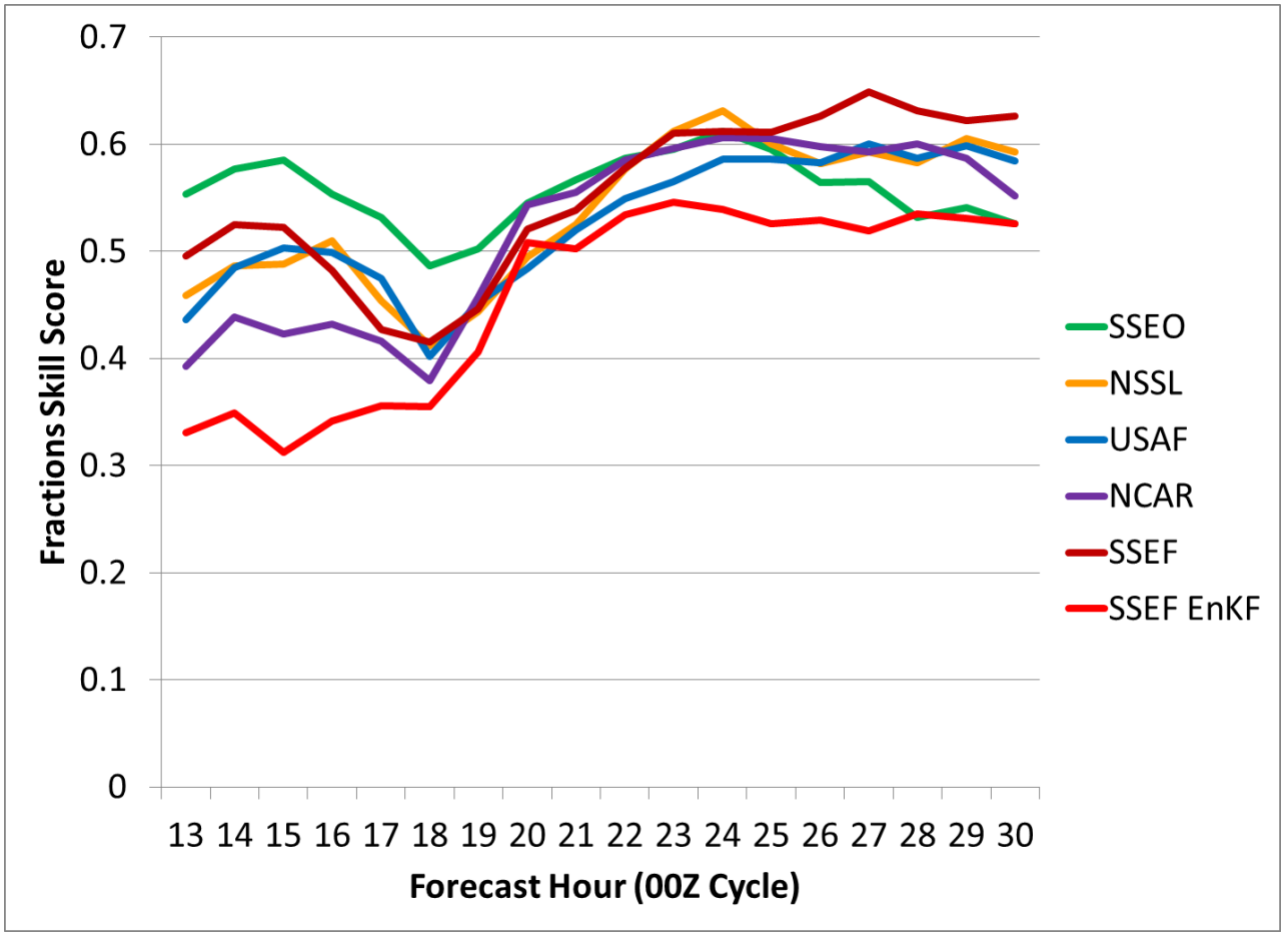

Figure 5. FSS by forecast hour for ensemble neighborhood reflectivity forecasts from the six different convection-allowing ensembles

In terms of the subjective ratings of the ensemble hourly-maximum field (HMF) forecasts in providing guidance for severe weather forecasts, the distribution of ratings among the ensembles was again rather similar (Fig. 6), except for the SSEF EnKF, which was clearly the lowest-rated ensemble. For the top-performing ensembles, they more often than not provided useful severe weather guidance (i.e. mean rating above 5). The NSSL ensemble had a slightly higher mean/median rating than the other ensembles while the NCAR and AFWA ensembles had slightly lower mean ratings than the SSEO, NSSL, and SSEF. The similar ratings among the ensembles highlight the fact that the complexity of convection-allowing ensemble design does not appear to strongly correspond to the ability of an ensemble to provide forecasters with useful guidance for severe weather outlooks.

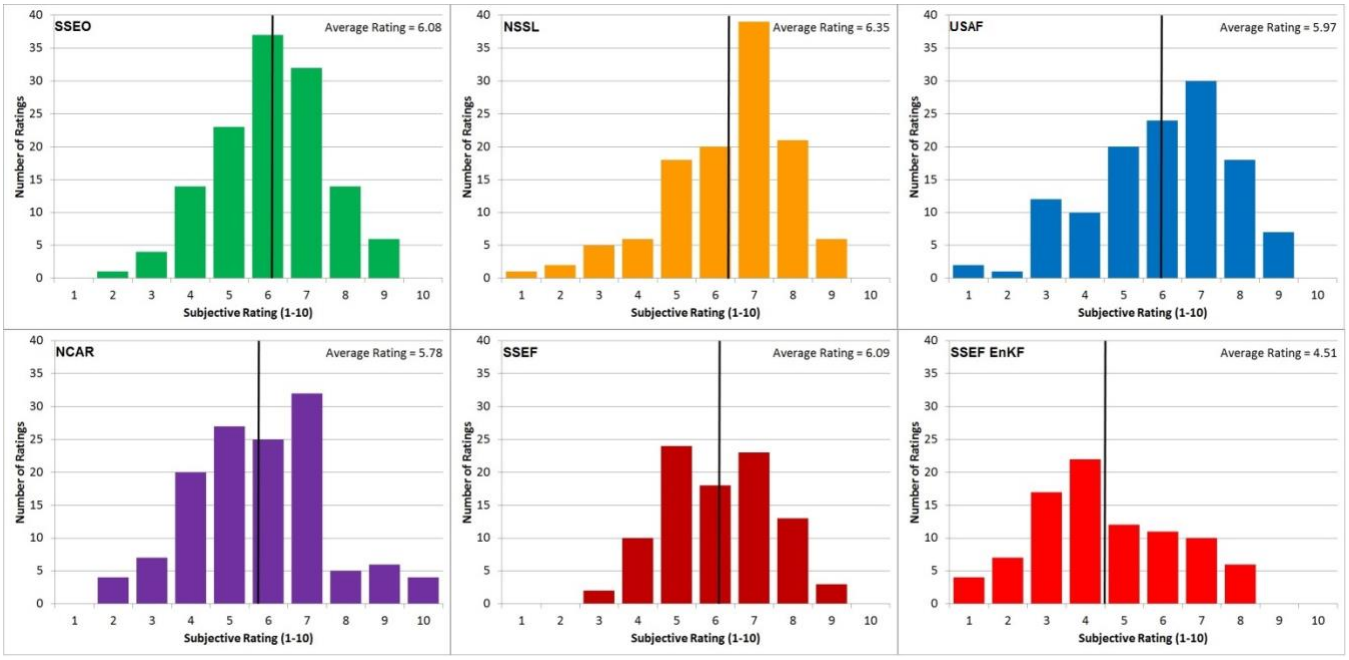

Figure 6. Distribution of subjective ratings (1 to 10) for the ensemble HMF forecasts compared to local storm reports for the six different convection-allowing ensembles.

d) Convection-Allowing Ensembles for Day 2

Convection-allowing ensembles were also examined into the Day 2 period (i.e., f36-f60 from 0000 UTC-initialized runs). The evaluation of Day 2 ensemble output was done less frequently than the Day 1 evaluation, owing to computing/data issues. Nevertheless, the preliminary results from the spring period provided some insights. The USAF and NCAR ensembles were more likely to have Day 2 forecasts rated similar to or better than (i.e., ≥0 ratings) their Day 1 forecasts compared to the SSEF or SSEF EnKF (Fig. 7). Figure 8 shows an example where the Day 2 USAF ensemble forecast was rated higher than the Day 1 USAF ensemble forecast. Even though the sample size was very limited, the Day 2 forecasts from convection-allowing ensembles often provided useful severe weather guidance during the five-week period in the spring.

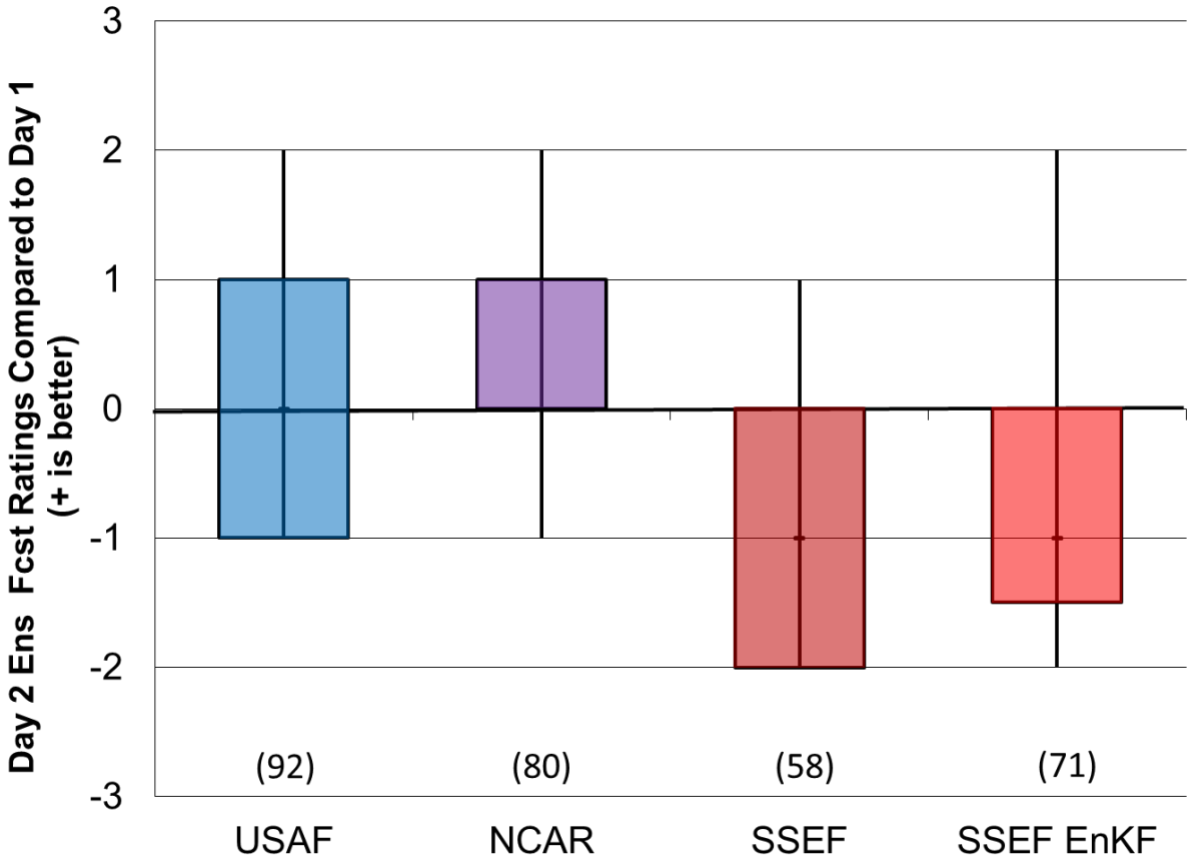

Figure 7. As in Fig. 1, except for the distribution of subjective ratings (-3 to +3) for the Day 2 ensemble forecasts from the USAF (blue), NCAR (purple), SSEF (dark red) and SSEF EnKF (red) compared to the Day 1 forecasts, valid for the same time period. The number of ratings for each ensemble is listed in parentheses above the ensemble name.

e) Evaluation of Parallel CAMs

During SFE2015, SPC had access to parallel CAMs from EMC and GSD for comparison to the operational versions of the CAMs. The parallel versions contained changes/improvements over their operational counterparts, and following formal evaluations, they were/are intended to be implemented operationally by NCEP. Most of the changes to the Hi-Res Window (HRW) ARW and NMMB runs were relatively minor (e.g., increase in vertical levels from 40 to 50), and the similar subjective ratings support the limited overall changes to the forecast performance (Fig. 9). The parallel HRW runs were implemented operationally at NCEP on 8 September 2015.

The parallel NAM Nest and parallel HRRR both showed improvements over their operational counterparts. The parallel NAM Nest was run at 3-km grid spacing, as opposed to the 4-km operational NAM Nest, in addition to being nested within an upgraded parent NAM. The parallel HRRR (run by GSD) included many physics changes to improve the afternoon warm, dry bias (and subsequent convective initiation issues) of the operational HRRR. Figure 10 reveals examples of improved convective initiation forecasts from the parallel HRRR (HRRRP) when compared to the operational HRRR.

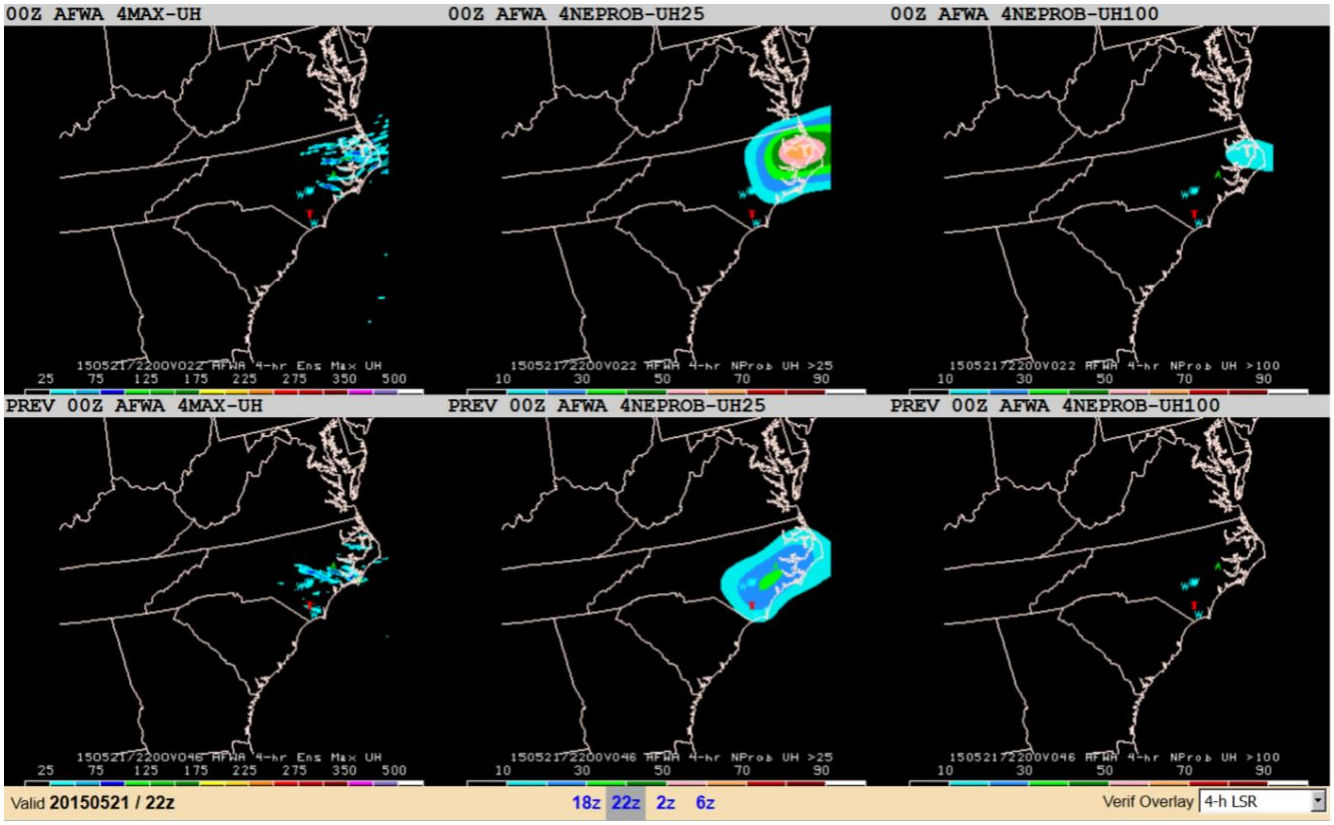

Figure 8. AFWA Day 1 (top row) and Day 2 (bottom row) forecasts of 4-h ensemble maximum UH (left column), ensemble neighborhood probability of UH ≥25 m^2s^{-2} (middle column), and ensemble neighborhood probability of UH ≥100 m^2s^{-2} (right column) valid 1800-2200 UTC on 21 May 2015. The severe reports during this 4-h period are plotted as letters in each panel.

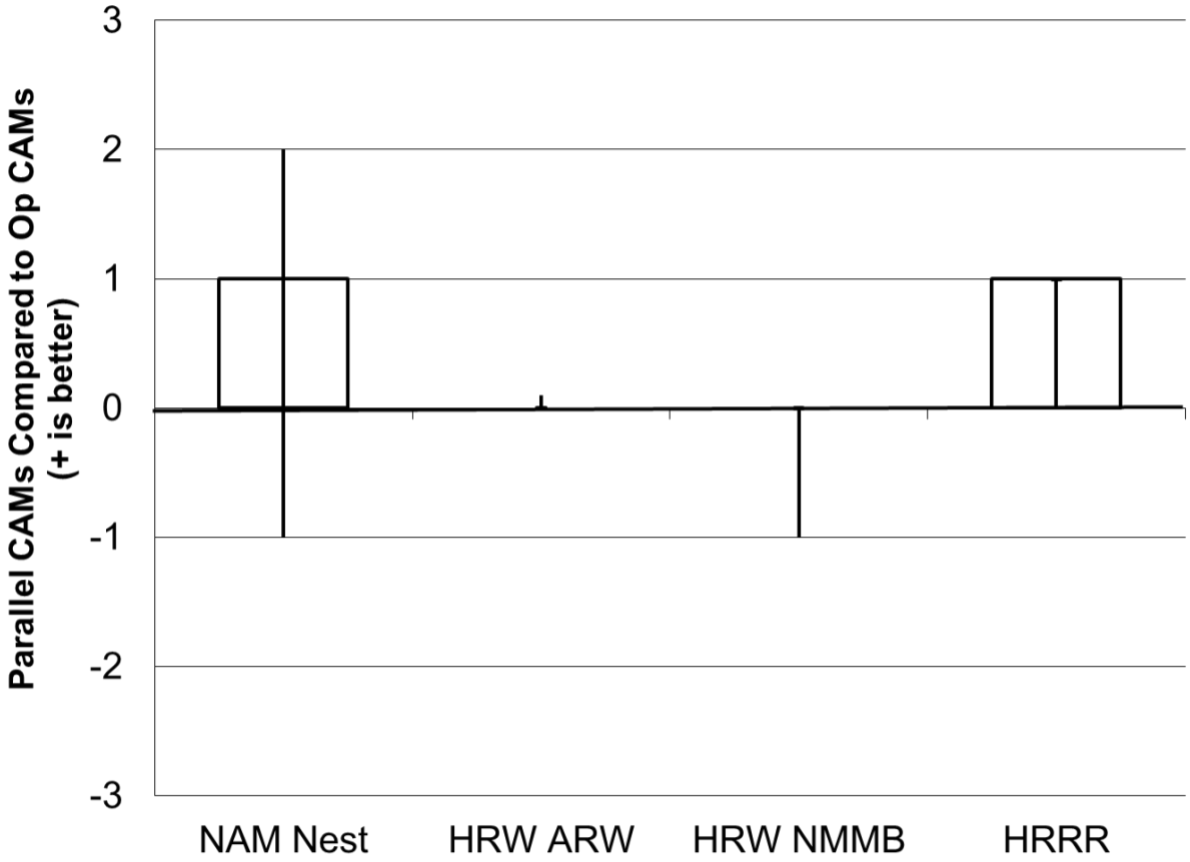

Figure 9. As in Fig. 1, except for the distribution of subjective ratings (-3 to +3) of the parallel versions of the CAMs compared to the operational versions. Boxes do not show up for HRW runs because most ratings (within interquartile range) were 0 (i.e., no change).

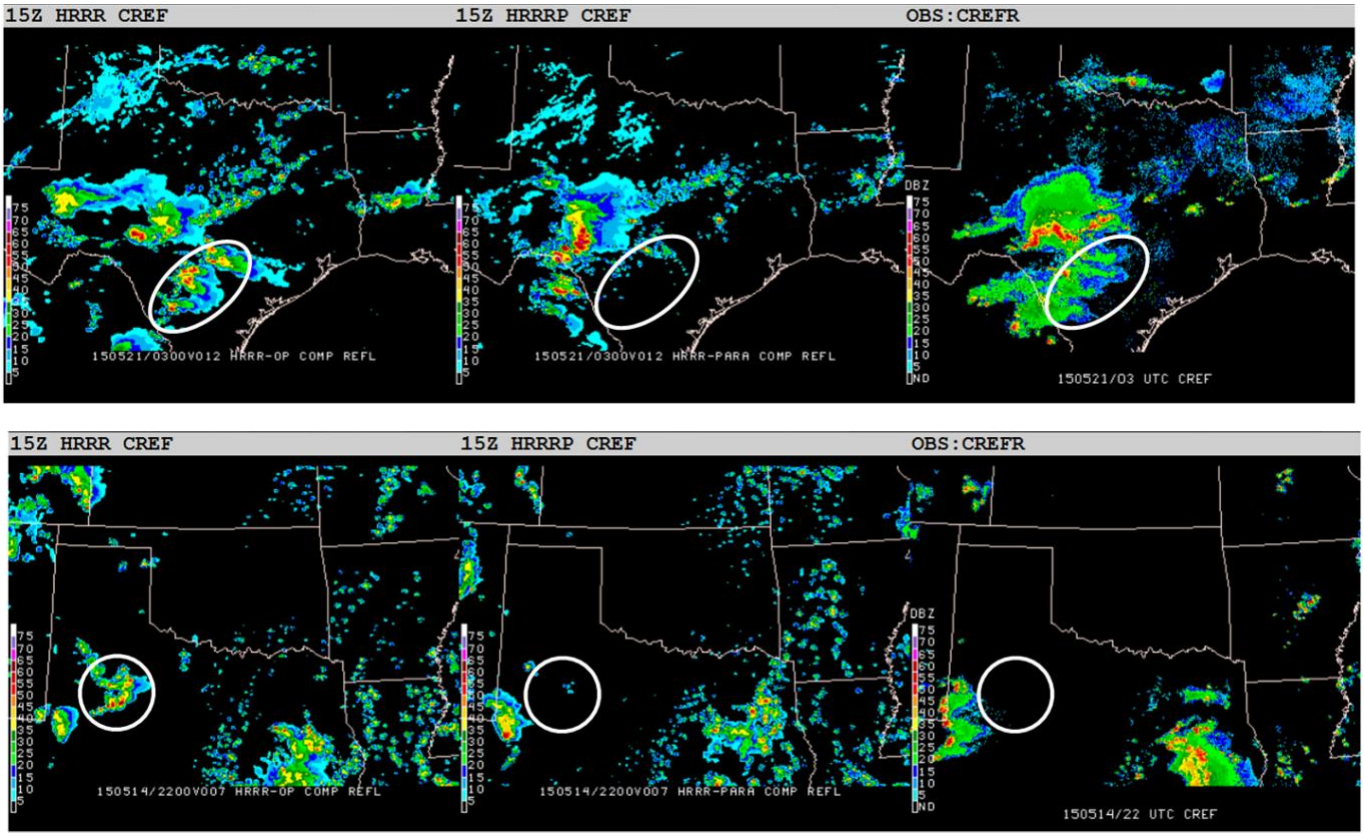

Figure 10. Simulated reflectivity forecasts valid at 0300 UTC on 21 May 2015 from the 15 UTC operational HRRR (upper left), parallel HRRR (upper middle), and observed reflectivity (upper right). Simulated reflectivity forecasts valid at 2200 UTC on 14 May 2015 from the 15 UTC operational HRRR (lower left), parallel HRRR (lower middle), and observed reflectivity (lower right).

f) Evaluation of Hail Diagnostics

For the second year, the HAILCAST algorithm implemented in WRF-ARW was used to predict hail size (Adams-Selin 2013). This is a modified version of the algorithm in Brimelow et al. (2002) and Jewell and Brimelow (2009) that was applied to coarser resolution regional models that include parameterized convection. Rather than predict hail size explicitly, the HAILCAST model uses convective cloud and updraft attributes to determine the growth of hail from initial embryos. The cloud attributes for the model are those predicted explicitly in the WRF-ARW forecasts and the snow, ice, and graupel mixing ratios at the first level above the freezing level at which they exist are used to determine the initial embryo size. During SFE2014, it was very evident that HAILCAST routinely over-predicted hail sizes, as nearly every convective storm contained greater than 1-inch hail. As a result, changes were made to HAILCAST after SFE2014 that resulted in more realistic hail-size forecasts. Specifically, rime soaking and variable density options were added, and the dependency on the microphysics scheme was removed by using five constant initial-embryo sizes, as opposed to those predicted explicitly in the schemes themselves. The changes were implemented in the NSSL-WRF and NSSL-WRF ensemble on 9 July 2014. Additionally, the updated HAILCAST algorithm was available in both CAPS ensembles for SFE2015.

New to SFE2015 was a hail size diagnostic derived directly from the microphysics parameterizations, which was implemented by Greg Thompson of NCAR. There was a compatibility issue with the code used by

CAPS, so this hail diagnostic was not available in the CAPS SSEF ensembles. Thus, we were not able to compare the Thompson method directly to HAILCAST within the same modeling system. The Thompson hail-size diagnostic was available in the NCAR EnKF-based ensemble.

Comparing the two methods (keeping in mind that they were run in two different ensemble systems), it was noted that HAILCAST in the NSSL-WRF and CAPS ensembles produced hail sizes that were generally larger than those produced by the Thompson method in the NCAR ensemble. Compared to the WSR-88D-derived maximum expected size of hail (MESH), the general feeling among participants was that HAILCAST slightly overestimated hail size while the Thompson method slightly under-estimated hail size. Additionally, it was noted that the Thompson method produced hail of any size over more widespread regions than HAILCAST, which is very likely due to the Thompson method not having any updraft speed and longevity criteria for producing hail. The most common areas where the Thompson method produced hail (of any size) while the HAILCAST output did not produce hail were over high elevations of the West.

For the formal evaluations, participants were asked a series of questions:

1. "Using 4-hourly forecasts from the CAPS control member, evaluate HAILCAST and GT (Greg Thompson) methods for diagnosing hail size. Use a rating scale from -3 to 3, where -3 is severe under-prediction and 3 is severe over-prediction of hail size compared to MESH." (note: because the GT method was not available in the CAPS runs, only the HAILCAST forecasts were evaluated during this activity).

Results from this evaluation over the three cases (4, 11, and 19 May 2015) that were examined included 16 individual responses from which the average was 0.35, indicating very slight over-prediction. The comments also reflected slight over-prediction, but many participants also noted large displacement errors.

2. "Subjectively rate the usefulness of hail size ≥ 1-in probability forecasts from the SSEF 3DVAR ensemble derived using HAILCAST and GT methods for diagnosing hail size, using a rating scale of Not Useful (1) to Very Useful (10)." (note: GT method was not available)

The average rating was 4.56. Many of the comments noted large areas of false alarm, but that areas of 1-in hail were, in general, captured by the probabilities.

3. "Subjectively rate the usefulness of hail size ≥ 2-in probability forecasts from the SSEF 3DVAR ensemble derived using HAILCAST and GT methods for diagnosing hail size, using a rating scale of Not Useful (1) to Very Useful (10)." (note: GT method was not available)

The average rating was 5.9, but there was a much larger variance in the range of ratings that were assigned relative to those for the 1-in hail probabilities. From the comments, it was clear that some of the higher ratings were for cases of "correct nulls" where no 2-in hail occurred and none was forecast. The lower ratings were assigned for a case in which very large hail occurred in southwest Texas and the probabilities were very low to zero.

4. "Do the hail size forecasts provide additional useful information relative to traditionally used HMFs like UH?"

There were 16 responses and they were all "yes".

g) Tornado Probabilities from NSSL-WRF Ensemble

During SFE2015, probabilistic forecasts for tornadoes were generated from the NSSL-WRF ensemble using updraft helicity (UH) ≥ 75m^2s^{-2} as a proxy for tornadoes with a variety of environmental filters: LCL < 1500m, SBCAPE/MUCAPE ≥ 0.75, and STP ≥ 1. Each participant was asked on a daily basis to rate probabilistic forecasts of tornadoes generated from the NSSL-WRF ensemble on a scale from one (Very Poor) to ten (Very Good). Tornado reports from the LSR database were overlaid on the forecast probabilities for verification purposes, as illustrated in Fig. 11.

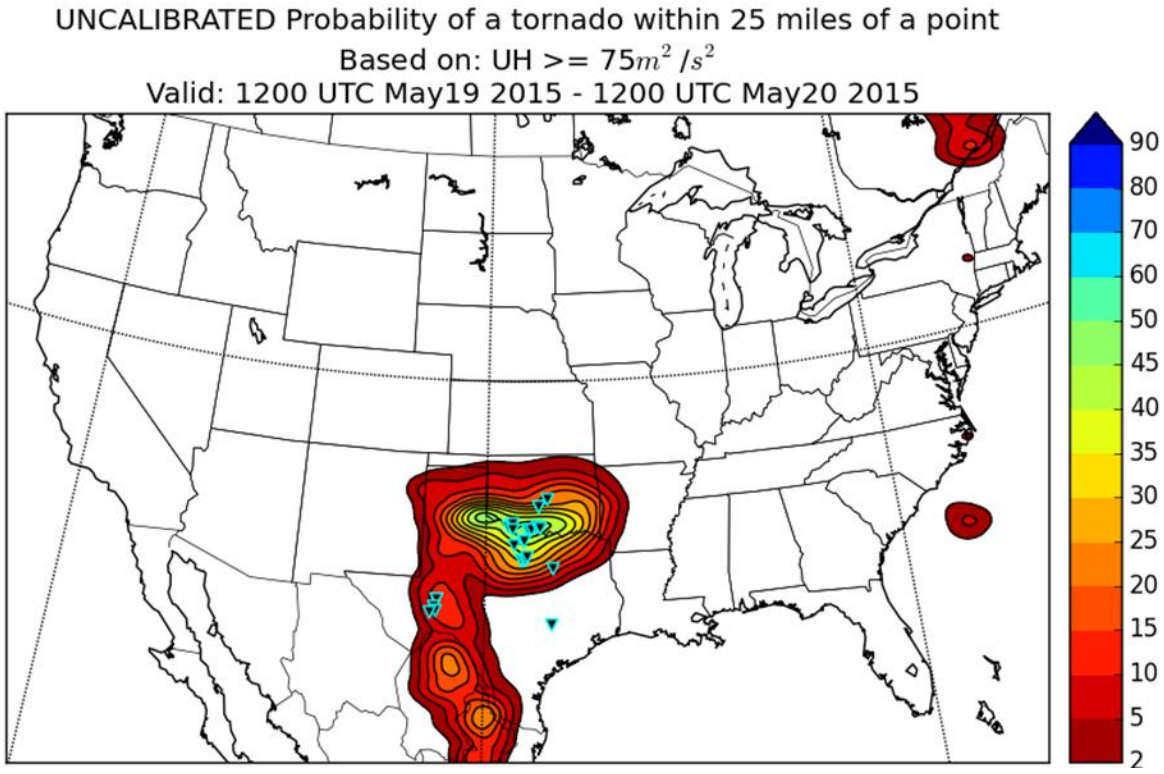

Figure 11. Probability of UH ≥75 m^2s^{-2} within 25 miles of a point from the NSSL-WRF ensemble valid from 1200 UTC 19 May 2015 to 1200 UTC 20 May 2015. Inverted triangles are tornado reports.

The distributions of subjective ratings assigned to the 24h probabilities by the individual participants suggest that incorporating environmental information results in an improved forecast over solely using UH (Fig. 12). None of the environmental filters (LCL, CAPE, STP, or combined) clearly stood out as the best method; however, they all generally improved the UH guidance. Participants often noted that the incorporation of environmental information helped focus the area of interest and reduce the false alarm area.

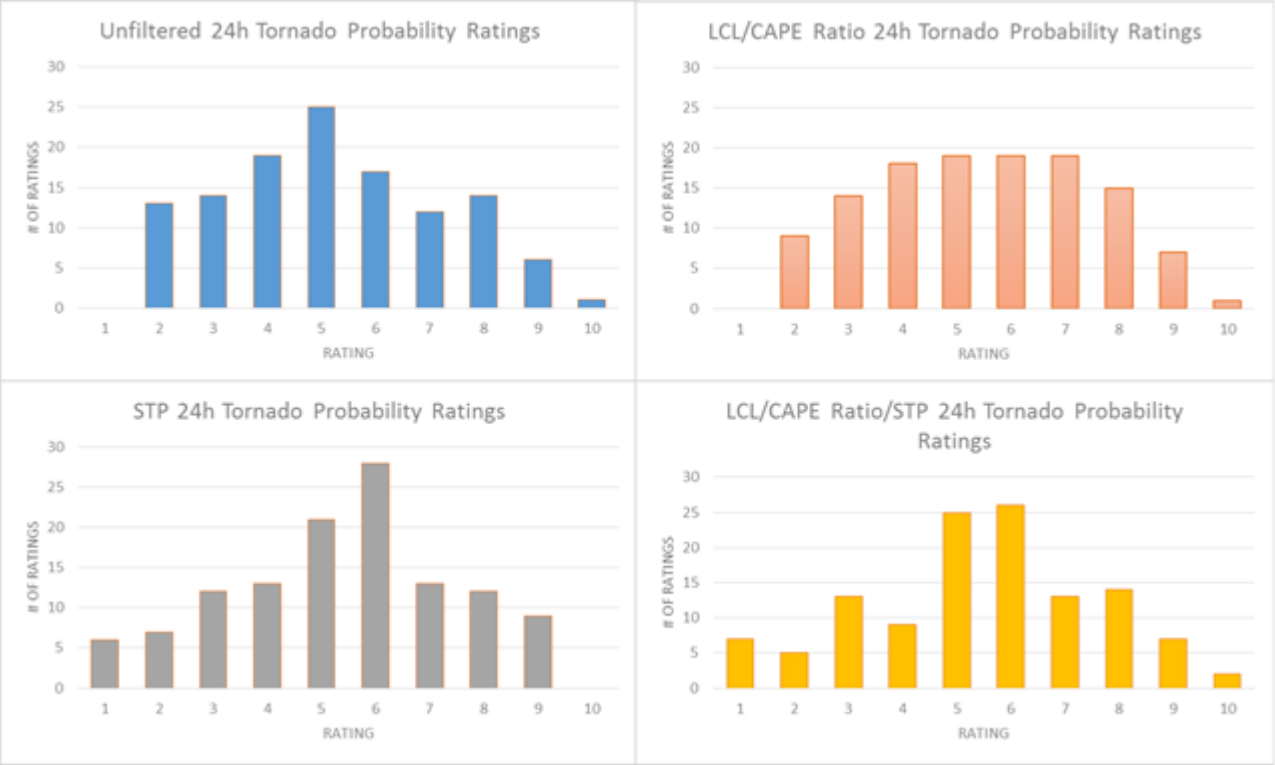

Figure 12. Distribution of subjective ratings of 24-h tornado probabilities as generated from the NSSL-WRF ensemble.

h) Microphysics Sensitivity Tests

Since 2010 one component of model evaluation activities during annual SFEs has involved subjectively examining sensitivity to microphysics parameterizations used in the WRF model. This has been done by comparing various forecast fields including simulated reflectivity, simulated brightness temperature, low-level temperature and moisture, and instability for the set of SSEF ensemble members with identical configurations except for their microphysical parameterization. During SFE2015, the following microphysics parameterizations were tested: Thompson, Morrison, Milbrandt and Yau (MY), and the Predicted Particle Properties (P3) scheme. Additionally, a newer version of P3 was tested that uses two-category ice, as opposed to one-category ice, which is used in the older version. Finally, another version of Thompson was tested that accounts for sub-grid scale clouds in the RRTMG radiation scheme based on research by Greg Thompson (NCAR). Unlike previous years, there was not a formal evaluation activity to document the microphysics sensitivities. This was decided because there was a general feeling among participants from previous years that it was becoming harder to discern systematic differences between the different schemes. However, comments were recorded on two days when microphysics experts were participating (listed in Table 2 in the Appendix) and further post-experiment analyses of these members are planned. An example case is shown in Fig. 13. In the example case, there were some very noticeable differences in convective organization and placement among the microphysics schemes.

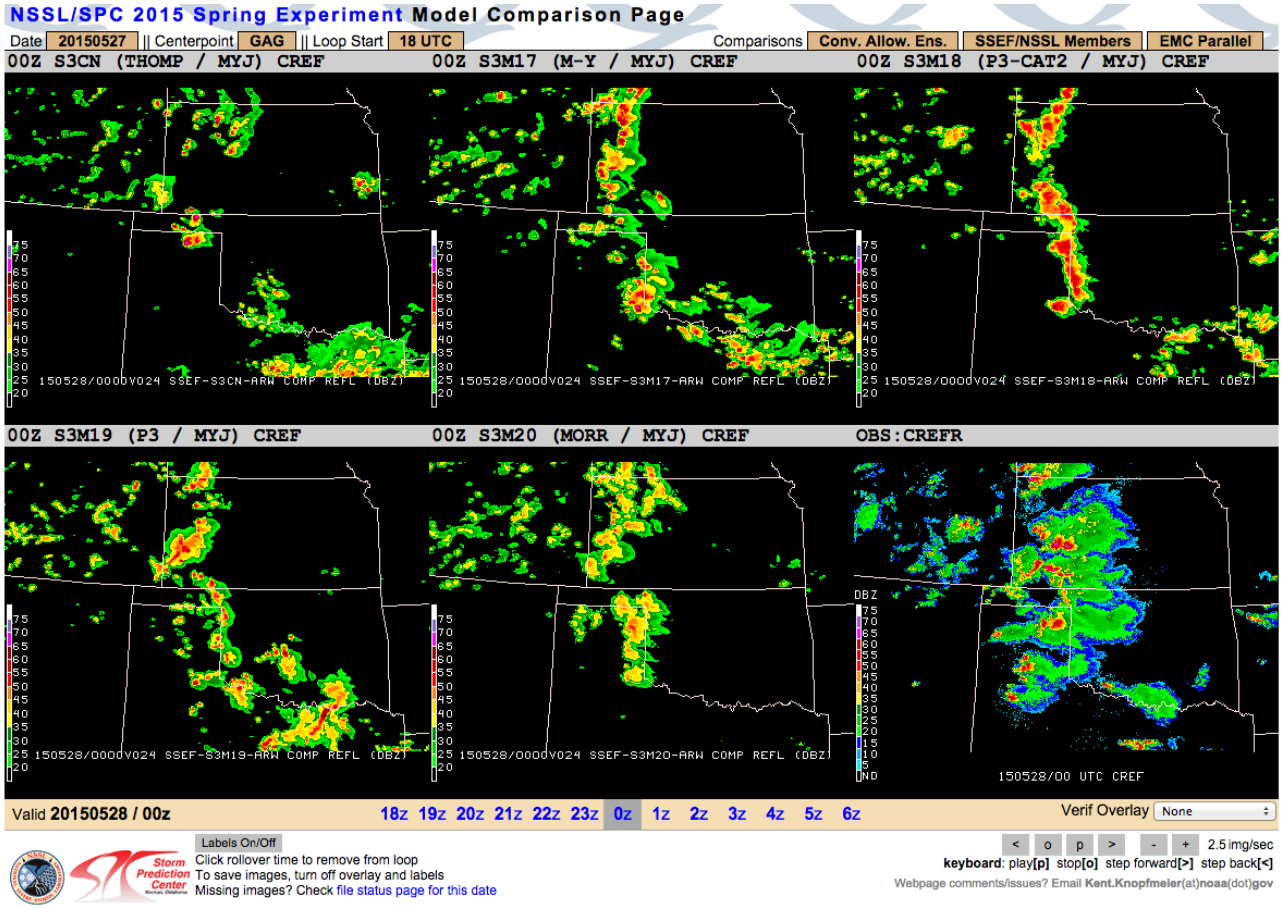

Figure 13. Forecasts and observations of composite reflectivity valid 0000 UTC 28 May 2015. The forecasts were initialized 0000 UTC 27 May and are from the members of the SSEF system configured identically except for their microphysics schemes. The panels include Thompson (upper-left), MY (upper-middle), P3-CAT2 (upper-right), P3 (lower-left), Morrison (lower-middle) and observations (lower-right).

i) Comparison of Met Office CAMs with NSSL-WRF

Several Met Office researchers and forecasters participated in the SFE2015, continuing this beneficial collaboration that permits an examination of forecast quality over a much more geographically diverse region than the United Kingdom. To facilitate this, the Met Office has implemented some of the unique storm-scale diagnostics developed at NSSL/SPC like simulated reflectivity and updraft helicity into the UM CAMs. This has enabled NSSL and SPC to examine forecasts of convection from a high-resolution modeling system completely independent of the WRF model and other US modeling systems. Also, because the Met Office has devoted a very large effort to accurately depicting the boundary layer due to its importance in the UK, NSSL and SPC were particularly interested in the quality of forecast low-level vertical profiles from the convection-allowing versions of the UM since this is a well-known weakness in many US models.

Similar to previous SFEs, to gauge the quality of the convection-allowing UM forecasts, daily subjective comparisons of simulated reflectivity were made to the 4-km grid-spacing NSSL-WRF and corresponding observations. The NSSL-WRF has been used to provide storm-scale guidance to SPC forecasters since 2006 and

is generally highly regarded. Thus, it served as a well-known baseline against which to compare the UM forecasts. Each day SFE2015 participants were asked a series of questions:

1. "Using the Probabilistic Hazard Information (PHI) tool, and focusing on areas of interesting weather, compare the NSSL-WRF and operational version of the 2.2 km UM (first 12 h)." Choices were "UM better than NSSL-WRF", "UM worse than NSSL-WRF", "Same", or "n/a".

Out of 133 total responses, 73 (55%) were that the UM was better than the NSSL-WRF, 31 (23%) were than the UM was worse than the NSSL-WRF, and 29 (22%) were that they were the same.

2. "Using the Probabilistic Hazard Information (PHI) tool, and focusing on areas of interesting weather, compare the NSSL-WRF and operational version of the 2.2 km UM (12 to 36 h)." Choices were "UM better than NSSL-WRF", "UM worse than NSSL-WRF", "Same", or "n/a".

Out of 132 total responses, 69 (52%) were that the UM was better than the NSSL-WRF, 29 (22%) were than the UM was worse than the NSSL-WRF, and 34 (26%) were that they were the same.

3. "Using the PHI tool, compare the 2.2 km Operational and Parallel UM (0-48 h)."

Out of 122 responses, 29 (24 %) were that the Parallel was better, 56 (46%) were than the Parallel was worse, and 37 (30%) were than they were the same.

4. "Using the PHI tool, compare the 1.1 km to the 2.2 km Operational UM."

Out of 104 total responses, 26 (25%) were that the 1.1 km was better than the 2.2 km, 33 (32%) were that the 1.1 km was worse than the 2.2 km, and 45 (43%) were that they were the same.

5. "Compare forecast soundings in regions with EMLs from the NSSL-WRF and 2.2 km Operational UM at sites where observed raob data is available. With a focus on sounding structure in the PBL and depiction of any capping inversions, which model has the best forecast sounding?"

Out of 89 total responses, 60 (67%) were that the UM was better than the NSSL-WRF, 9 (10%) were than the UM was worse than the NSSL-WRF, and 19 (21%) were that they were the same. Many of the comments noted that the UM was much better at depicting the sharpness and structure of strong capping inversions compared to the NSSL-WRF.

j) Reflectivity forecast comparison following initialization of CAPS 3DVAR- and EnKF-based ensembles and HRRR

This evaluation activity focused on the first 6 hours of the CAPS 3DVAR and EnKF-based ensembles, as well as the HRRR. The focus was on a regional area of severe weather interest and how well the specific members of each ensemble depicted storms in the initial conditions and their subsequent evolution during the first 6 h of the forecast. The member of the EnKF-based ensemble that used the mean EnKF analysis as the initial condition was compared to the control member of the 3DVAR-based ensemble. Both members had the same physics configuration. Participants were asked to rate the forecasts as follows:

"Subjectively rate 15-min interval composite reflectivity forecasts during the first 6 h of each forecast. Using a rating scale of Very Poor (1) to Very Good (10) consider the quality of the initial correspondence with observed radar. Then, consider how well model storm maintain a stable transition from image to image during the first 6 h, and how well the forecasts correspond to observations."

The average rating for the 3DVAR control member was 6.6, the EnKF-based member was 5.8, and the HRRR was 6.3. In the comments, it was clear that the EnKF-based member received lower ratings because it had a tendency to weaken convection too quickly in its short-range forecasts. Figure 2 displays an example of one such case that occurred for the 4-h forecast initialized 0000 UTC 26 May 2015. In this case, the EnKF-based run erroneously dissipated a large portion of a southern end of a squall line that was moving southeast across eastern Texas. The other runs examined did a much better job of maintaining the southern end of the squall line.

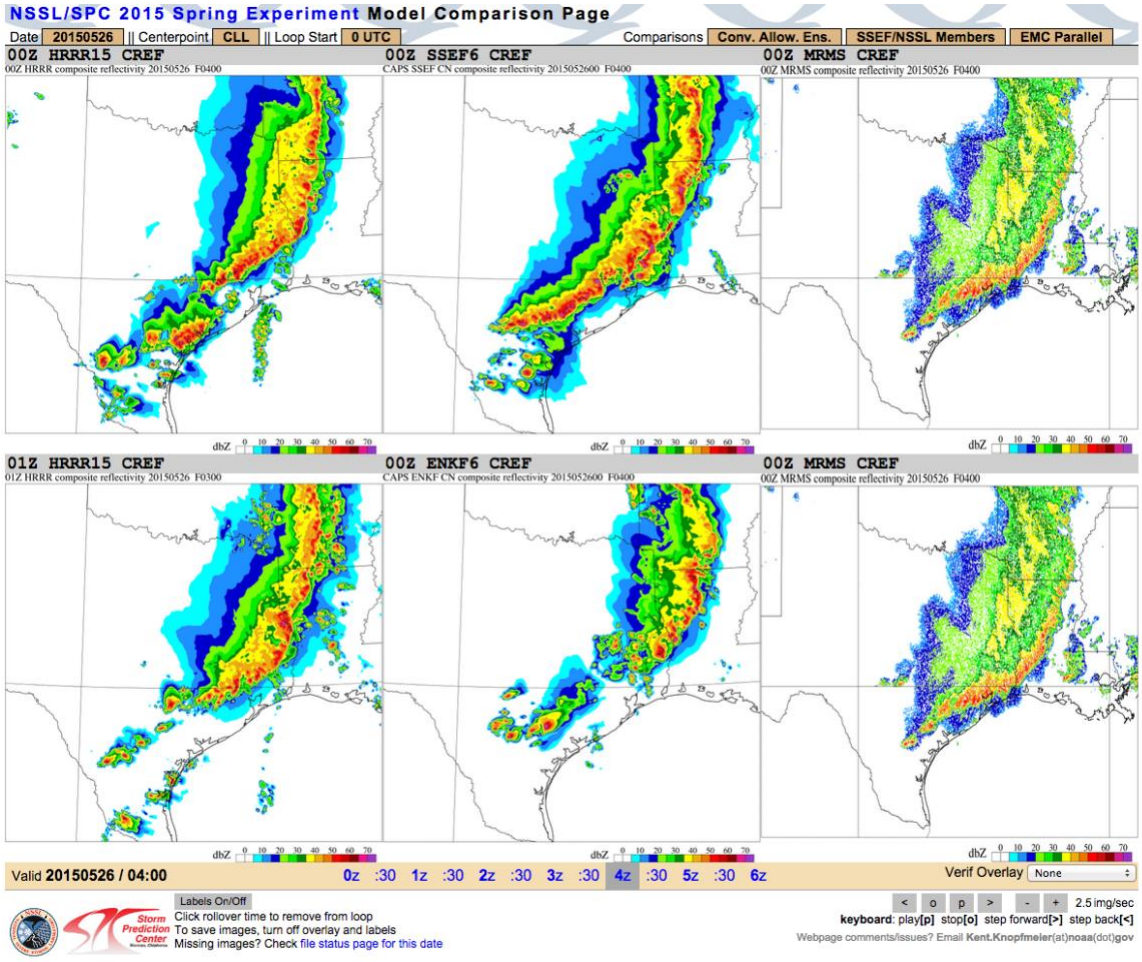

Figure 14. Forecasts and observations of composite reflectivity valid 0400 UTC 26 May 2015. The forecasts were initialized 0000 UTC 26 May, except for the 0100 UTC initialized HRRR in the bottom left panel. The other panels include the HRRR (top-left), the control member of the SSEF system (top-middle), the observed composite reflectivity (both right panels), and the member initialized from the mean EnKF-based analysis (bottom-middle).

k) Exploration of the NCAR Model for Prediction Across Scales (MPAS)

A new model examined in SFE2015 was the NCAR Model for Prediction Across Scales (Skamarock et al. 2012). MPAS produced forecasts initialized daily at 0000 UTC with 3-km grid-spacing over the CONUS and forecasts extending out to 120 h (5 days). Idealized convective-scale tests, in addition to real-data hindcast tests on 3-km global meshes, show that the MPAS produces convective realizations similar to that of the ARW model. MPAS uses "scale-aware" physics, which allows for variable resolution across the globe, with some regions at convection-allowing resolution, which allows for the output of explicit storm attributes (e.g., UH; Fig 15). The main science question to be addressed with MPAS is whether running a unified modeling system (i.e., no grid-nesting/downscaling) is better than current methodologies where regional non-hydrostatic modeling systems are "nested" within operational global models.

For SFE2015, there was not a formal evaluation activity for the MPAS forecasts, however, the forecasts were examined on a daily basis and used during the forecasting process, especially for the construction of Day 2 and Day 3 experimental outlooks. MPAS provided useful convective-scale guidance out to Day 3 and beyond for several severe weather events, and two of those cases are described here.

Several days in advance of 9 May 2015, operational models were indicating a synoptic pattern very favorable for a severe weather outbreak on this day across the southern plains, and the SPC Day 3 convective outlook outlined an area across Oklahoma and Kansas as having a moderate risk for severe storms. In reality, during the late morning of 9 May, strong forcing for ascent combined with a weak capping inversion led to widespread convection and associated cloud cover across much of western Oklahoma and Kansas with that inhibited heating and destabilization during the afternoon. In Fig. 16a, which illustrates the surface-based CAPE and CIN from the SPC mesoanalyses valid at 2100 UTC 9 May, the impact of the early convection can be seen from the minimal CAPE (<1000 J/kg) across much of Oklahoma and Kansas. Although severe storms did occur from Texas into western Kansas, because of the early storms, the event as a whole ended up being less significant than what some of the earlier model guidance and convective outlooks had suggested (although these outlooks did mention the forecast uncertainty and factors that could possibly reduce the severe potential). While forecasting a synoptic scale pattern favorable for widespread severe weather 3 days in advance of 9 May similar to the operational models, the MPAS forecasts also were indicating that widespread convection would develop early in the day on 9 May. From Fig. 16c, which displays the 69 h forecast of CAPE valid at 2100 UTC 9 May, the impact of this early convection (Fig. 16e) is very apparent from the reduced CAPE in Oklahoma and Kansas. The scenario depicted by MPAS 3 days in advance was consistent with what actually occurred.

Another case in which MPAS provided useful extended range convective guidance was on 16 May 2015. This was another situation in which operational models were indicating a synoptic pattern very favorable for a severe weather outbreak several days in advance. However, similar to 9 May, the extent and intensity of the severe weather threat was quite uncertain because it was not clear how much morning/early afternoon convection would inhibit heating and destabilization in the warm sector. Although a shallow layer of clouds inhibited heating to some extent across the warm sector, especially across central Oklahoma, a lack of widespread early convection allowed enough destabilization to occur to support a significant severe weather event and several long lived supercells that produced tornadoes across the Texas panhandle, Oklahoma, and into Missouri. Figure 16b shows the surface-based CAPE and CIN from the SPC mesoanalysis valid at 2100 UTC 16 May. Note the large area of CAPE ranging from 2000 to 3000 J/kg across much of Texas, Oklahoma, and Missouri, which, combined with sufficient low-level and deep-layer shear, resulted in widespread severe

storms. The forecasts from MPAS 3 days in advance were consistent with this scenario. Figure 16d shows the 69 h forecast of CAPE, which did not exhibit large impacts from early convection and matched quite well the observed range of values across Oklahoma and Texas. Furthermore, the MPAS forecasts depicted intense supercells forming in the warm sector where significant destabilization occurred in central Oklahoma. Although there were notable timing and geographic displacement errors with these forecast storms (in reality, supercells formed much further west during the early afternoon and did not move across central Oklahoma until evening), the overall forecast scenario corresponded to the observations reasonably well and, again, would have provided useful convective scale guidance to forecasters in the extended range.

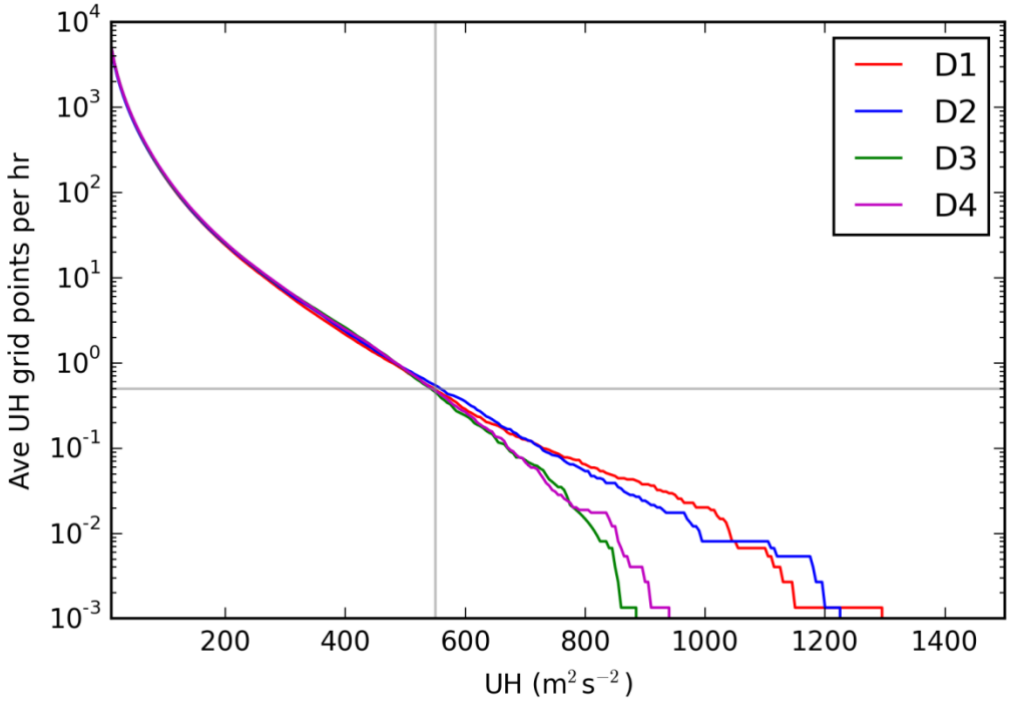

Figure 15. Plot of average UH grid-point counts per hour from 0000 UTC-initialized MPAS forecasts by UH threshold during SFE2015 for each valid convective day (1200 – 1200 UTC): D1 (forecast hours 12-36), D2 (forecast hours 36-60), D3 (forecast hours 60-84), and D4 (forecast hours 84-108).

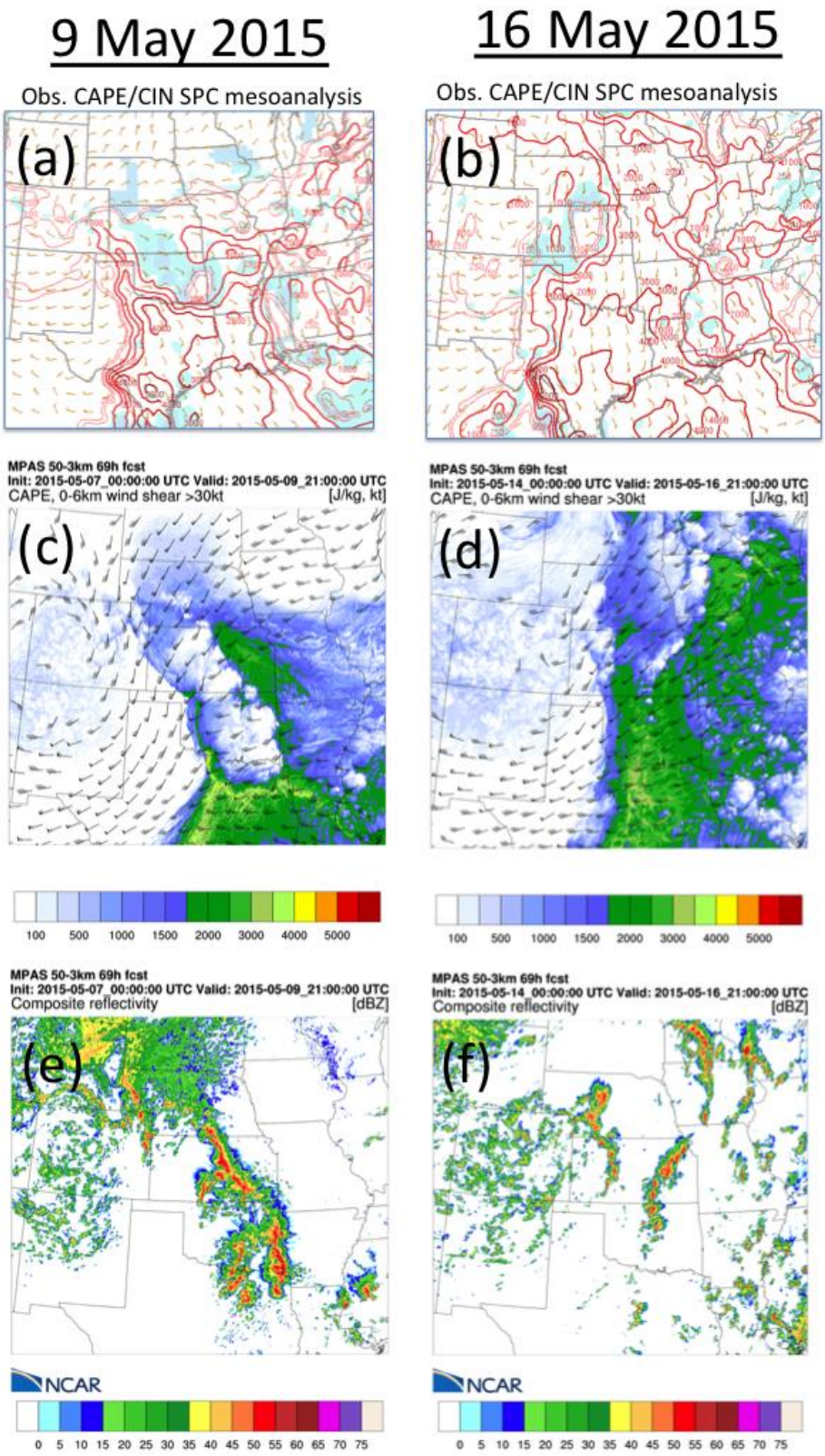

Figure 16. (a) CAPE and CIN from SPC's mesoanalysis valid at 2100 UTC 9 May 2015. 69 h MPAS forecasts of (c) CAPE and 0-6 km shear vectors, and (e) simulated composite reflectivity valid 2100 UTC 9 May 2015. (b), (d), and (f) same as (a), (c), and (e) except for 16 May 2015.

4. Summary

The 2015 Spring Forecasting Experiment (SFE2015) was conducted at the NOAA Hazardous Weather Testbed from 4 May – 5 June by the SPC and NSSL with participation from forecasters, researchers, and developers from around the world. The primary theme of SFE2015 was to utilize convection-allowing model and ensemble guidance in creating high-temporal resolution probabilistic forecasts of severe weather hazards, including extension into the Day 2 period. Several preliminary findings from SFE2015 are listed below:

- Generated high temporal resolution outlooks for individual severe hazards using temporally disaggregated full-period outlook from a convection-allowing ensemble as first-guess guidance.
- Examined six different convection-allowing ensemble systems and found that, regardless of design and complexity, all of the ensembles provided similar, useful guidance for Day 1 severe weather forecasting.
- Utilized several convection-allowing ensembles for creating Day 2 forecasts for individual severe hazards, noting ensemble utility beyond the Day 1 period.
- Recommended operational implementation (which occurred in September 2015) of HRW ARW and NMMB parallel CAM runs and identified improved guidance from the parallel HRRR and NAM Nest for convective storms compared to operational versions.
- Examined a modified version of HAILCAST, noting more realistic forecasts of hail size compared to the previous version, and explored a hail size diagnostic (from Greg Thompson of NCAR) based on microphysics scheme.
- Determined that applying environmental filters to explicit updraft helicity (UH) in the NSSL-WRF ensemble resulted in improved guidance for probabilistic tornado forecasting compared to using UH only.
- Documented a better representation of strong vertical gradients in temperature and moisture near capping inversions in Met Office CAMs compared to the NSSL-WRF. Also noted that the 1.1-km UM version did not generally outperform the 2.2-km run.
- Explored the variable-resolution MPAS run at convection-allowing scale over the CONUS and documented its capability in generating realistic simulated storm structures out to Day 5.

Overall, SFE2015 was successful in testing new forecast products and modeling systems to address relevant issues related to the prediction of hazardous convective weather. The findings and questions exposed during SFE2015 directly promote continued progress to improve forecasting of severe weather in support of the NWS Weather-Ready Nation initiative.

Acknowledgements

SFE2015 would not have been possible without dedicated participants and the support and assistance of numerous individuals at SPC and NSSL. In addition, collaborations with OU CAPS, USAF, NCAR and the Met Office were vital to the success of SFE2015. In particular, Ming Xue (OU CAPS), Fanyou Kong (OU CAPS), Kevin Thomas (OU CAPS), Keith Brewster (OU CAPS), Yunheng Wang (OU CAPS), Evan Kuchera (USAF), Scott Rentschler (USAF), Glen Romine (NCAR), Craig Schwartz (NCAR), Ryan Sobash (NCAR), and Steve Willington (Met Office) were essential in generating and providing access to model forecasts examined on a daily basis.

References

Adams-Selin, R. 2013: In-line 1D WRF hail diagnostic. AFWA Internal Tech. Memo, SEMSD.21495.

Brimelow, J.C., 1999: Modeling maximum hail size in Alberta thunderstorms. *Wea. Forecasting*, **17**, 1048-1062.

Hitchens, N.M., H.E. Brooks, and M.P. Kay, 2013: Objective limits on forecasting skill of rare events. *Wea. Forecasting*, **28**, 525–534.

Jewell, R., and J. Brimelow, 2009: Evaluation of Alberta hail growth model using severe hail proximity soundings from the United States. *Wea. Forecasting*, **24**, 1592-1609.

Jirak, I. L., C. J. Melick, A. R. Dean, S. J. Weiss, and J. Correia, Jr., 2012: Investigation of an automated temporal disaggregation technique for convective outlooks during the 2012 Hazardous Weather Testbed Spring Forecasting Experiment. Preprints, *26th Conf. on Severe Local Storms*, Nashville, TN, Amer. Meteor. Soc., 10.2.

Jirak, I. L. C. J. Melick, and S. J. Weiss, 2014: Combining probabilistic ensemble information from the environment with simulated storm attributes to generate calibrated probabilities of severe weather hazards. Preprints, *27th Conf. on Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., 2.5.

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.

Rothfusz, L. P., P. T. Schlatter, E. Jacks, and T. M. Smith, 2014: A future warning concept: Forecasting A Continuum of Environmental Threats (FACETs). *2nd Symposium on Building a Weather-Ready Nation: Enhancing Our Nation's Readiness, Responsiveness, and Resilience to High Impact Weather Events,* Atlanta, GA, Amer. Meteor. Soc., 2.1.

Schwartz, C. S., J. S. Kain, S. J. Weiss, M. Xue, D. R. Bright, F. Kong, K. W. Thomas, J. J. Levit, M. C. Coniglio, and M. S. Wandishin, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263–280.Skamarock, W. C., Klemp, J. B., Duda, M., Fowler, L. D., Park, S.-H., and T. Ringler, 2012: A multiscale nonhydrostatic atmospheric model using centroidal Voronoi tesselations and c-grid staggering. *Mon. Wea. Rev*, ***140***, 3090–3105.

Stensrud, D. J., and Co-authors, 2009: Convective-scale warn-on-forecast system. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1499.

APPENDIX

Daily activities schedule in local (CDT) time

0800 – 0845: **Evaluation of Experimental Forecasts & Guidance**
Subjective rating relative to radar evolution/characteristics, warnings, and preliminary reports and objective verification using preliminary reports and MESH

- Day 1 & 2 full-period probabilistic forecasts of tornado, wind, and hail
- Day 1 4-h period forecasts and guidance for tornado, wind, and hail
- Days 1, 2, & 3 full-period probabilistic forecast of total severe
- Day 1 1-h period forecasts and guidance for total severe

0845 – 1115: **Day 1 Convective Outlook Generation**
Hand analysis of 12Z upper-air maps and surface charts

- Day 1 full-period probabilistic forecasts of tornado, wind, and hail valid 16-12Z over mesoscale area of interest (Fig. 17)
- Day 1 4-h probabilistic forecasts of tornado, wind, and hail valid 18-22 and 22-02Z*
- Day 1 full-period probabilistic forecast of total severe valid 16-12Z over mesoscale area of interest (Fig. 17)
- Day 1 1-h probabilistic forecasts of total severe valid 18-00Z*

1115 – 1130: **Break**
Prepare for map discussion and discuss relationship/translation from probabilities to watch

1130 – 1200: **Map Discussion**
Overview and discussion of today's forecast challenges and products
Highlight interesting findings from previous days

1200 – 1300: **Lunch**
Brief EWP participants at 1245

1300 – 1400: **Day 2 Convective Outlook Generation**

- Day 2 full-period probabilistic forecasts of tornado, wind, and hail valid 12-12Z over mesoscale area of interest
- Day 2 or Day 3 full-period probabilistic forecasts of total severe valid 12-12Z over mesoscale area of interest

1400 – 1500: **Scientific Evaluations**

- Convection-allowing ensemble comparison (reflectivity and HMFs): SSEO, AFWA, NSSL, SSEF, SSEF EnKF, NCAR EnKF.
- EMC parallel CAM comparison (reflectivity): NAM Nest, HiResW, HRRR
- Met Office CAMs: vertical resolution
- SSEF 3DVar vs. EnKF Comparison: impact on first few hours of control forecast
- Model forecasts of explicit hail size: HAILCAST, Thompson
- MPAS

1500 – 1600: **Short-term Outlook**

- Update 4-h probabilistic forecasts of tornado, wind, and hail valid 22-02Z*
- Generate 1-h probabilistic forecasts of tornado valid 22-02Z
- Update and generate 1-h probabilistic forecasts of total severe valid 21-02Z*

* Denotes forecasts also made by participants using the PHI tool on Chromebooks.

Table 1. List of weekly participants (with affiliation) during SFE2015.

Week 1	Week 2	Week 3	Week 4	Week 5
May 4-8	**May 11-15**	**May 18-22**	**May 26-29**	**June 1-5**
Nick Grahame (Met Office)	Nick Grahame (Met Office)	Mark Seltzer (Met Office)	Brent Walker (Met Office)	Brent Walker (Met Office) M-W
Jason Otkin (CIMSS) M-Th	Mark Seltzer (Met Office)	Kirsty Hanley (Met Office)	Michael Fowle (WFO ABR)	Steve Ramsdale (Met Office)
Michael Dutter (WFO MQT)	Jacob Carley (EMC)	Rob Hepper (AFWA)	Brad Ferrier (EMC)	Eric Aligo (EMC)
Jun Du (EMC)	Curtis Alexander (GSD)	Lance Bosart (SUNYA)	Isidora Jankov (GSD)	Brian Kolts (FirstEnergy)
David Dowell (GSD)	Eric James (GSD)	Matt Vaughan (SUNYA)	Jaymes Kenyon (GSD)	Ed Szoke (GSD)
Terra Ladwig (GSD)	Brock Burghardt (TTU)	Kyle Pallozzi (SUNYA)	Mike Watts (FedEx)	TJ Turnage (WFO GRR)
Becky Adams-Selin (AFWA)	Pat Spoden (WFO PAH)	Jeff Beck (GSD)	Mike Lawson (WFO AFC)	Tom Lonka (WFO MHX)
Brian Montgomery (WFO ALY)	Glen Romine (NCAR)	John Brown (GSD)	Harald Richter (BOM)	Steven Cavallo (OU)
Bill Skamarock (NCAR)	Bruce Entwistle (AWC)	Harald Richter (BOM)	Ryan Torn (SUNYA)	Dan Zacharias (AWC)
Casey Crosbie (CWSU ZID)	Gail Hartfield (WFO RAH)	Jeremy Berman (SUNYA)	Junella Tam (Hong Kong)	Stephen Konarik (WFO MFL)
Ryan Sobash (NCAR)	Brad Mickelson (WFO GGW)	Lou Wicker (NSSL)	Hugh Morrison (NCAR) Th-F	Junella Tam (Hong Kong) M-Th
Mark Loeffelbein (WRHQ)	Sarah Perfater (WPC) T-Th	Mark Klein (WPC)	Clark Evans (UWM)	Aaron Kennedy (UND)
David Imy (SPC Ret.) M-Th	James Thomas (WFO SGX)	David Gagne (OU)	Bryan Burlingame (UWM)	David Goines (UND)
	Kate-Lynn Walsh (OU student)	Bill Lapenta (NCEP) Th-F	Todd Chambers (WFO BYZ)	Ron Stenz (UND)
		Rich Bann (WPC)		

Table 2. Daily responses collected from the microphysics evaluations conducted during SFE2015. The date refers to the model initialization time.

5/27/2015: The members outside of the control seemed to struggle with spin up after the initialization hour initially overdoing reflectivity but then tapered back closer to reality in subsequent hours. Morrison overall performed the best in showing coverage, orientation and intensity of the convection as it initiated and evolved.
5/27/2015: Composite reflectivity: All relatively similar in the first twelve hours of each forecast, with some tendency for the M-Y, P3-2Cat, and P3 parameterizations to produce hotter reflectivity than the Thompson and Morrison members. All but the Morrison produce (presumably) elevated convection in southern Oklahoma, with the Thompson member being most aggressive in doing so, that verified reasonably well. Conversely, the Thompson member was much slower and less robust with CI and convective evolution across the High Plains during the daytime hours. 1 km AGL reflectivity: stratiform precipitation underdone by all parameterizations, with the Thompson member perhaps least underdone. CAPE and 2-m dewpoint: the Morrison member had lower values of each, better in line with the surface OA fields as compared to the other members. Unclear as to why.
5/27/2015: M-Y and P3-CAT2 to look the most realistic really become hot quickly. P3 seems the most realistic at first. After that first timeframe, the solutions all look more realistic. P3 and Thompson maintain convection in OK where there is none at the end of the time period. Morrison scheme is the only one that does not attempt to fire convection in eastern OK where there is none. Re-initiation of storms in the second half of the day is way overdone in the Thompson - it initiates an MCS and then is late in initiating the supercells. The P2 gets the initialization time well.

Morrison misses the convection in OK and is late in initiating the TX panhandle storm. The P3-CAT2 seems to do best in location and timing of supercells. P3 also doesn't seem to move the storms off so quickly.. Overall, I think it's the best of these models in this time frame. In the 18Z-6Z, Morrison seemed to do the best. CAPE field: MYJ and P3 seem to get the magnitude of the CAPE best. Overall shape of the high CAPE axis is well-captured by Morrison. Morrison doesn't return the CAPE as quickly, which is the main reason for its difference. Mesoanalysis makes for a difficult comparison, since the smoothed field can't contain the details that the CAMs have. Morrison seems to represent the dewpoints more accurately, but may be a bit too cool.
5/28/2015: MY & P3-Cat2 had a strange high bias that formed at hour 1-2. All of them seemed to struggle with simulating the reflectivity PDFs associated with the stratiform region; the Thompson may have done the best? They all missed the conv over TX panhandle near the end of the forecast. Morrison seemed drier Ismaller area of echos) with lower low-level Tds that led to lower CAPEs, but the lower dewpoints compared better against the obs.
5/27/2015: P3-2 seems a little too hot early. Thompson way too hot later with a big MCS in OK that wasn't observed. MY and Morrison schemes did a good job on TX Panhandle storms. P3 not bad, P3-2 develops Panhandle storms into an MCS too early. MY/P3/P3-2 too little coverage in later period. Morrison shows lower surface Td's and lower SBCAPE for some reason.
5/27/2015: FIrst 12hrs - P3 CAT 2 excessive devloping convection over SE OK late in the period. All others seem reasonable but have underdone elevated convection over SW Kansas. 12hrs+ - large differences emerge. Thomp develops MCS over OK which is not there in reality. MORR late to initiate but then does catch up to provide useful guidance later and overall better than the others. Others reasonable but activity generally looks to be underdone short-lived compared to obs. MORR slower to recover CAPE with less moisture advection from the south, perhaps explaining delayed convection initiation.
5/27/2015: p3 nyj: good on Canadian and western Kansas;, p3catz a little slow compared to obs; s3m17 not bad for the high plains, similar to p3catz; morr is slow but gets the areas of convection reasonably well;
5/27/2015: Thompson produces a large MCS in SE Oklahoma, there is some convection there in the other members but is weaker. Morrison is the most different from other members. Reality is between Thompson and the other members. It has less convection from 12-24 Z, but more after that especially from western Oklahoma to the Texas border. All schemes are missing storms in Texa panhandle. Thermodynamic fields are distinctly different in Morrison (e.g., lower CAPE around 24 Z).
5/28/2015: MY and P3 cat2 both suffered from same initial enhancement of simulated reflectivity after the initial conditions. Reflectivity was handled well by all members initially however spread increased with convective evolution through the overnight hours.
5/28/2015: All similar in the first 6 hrs. Reflectivity looks to be too high with the developing MCS over NW Texas in the MY version; others look better. MY and P3 two tends to be overdone when upscaling convection.
5/28/2015: Overall placement and timing of storms is similar. MY2 and 2-cat P3 tend to produce initially spurious high reflectivity within the first couple hours after initializing, and then it quickly subsides. Perhaps some issue with how initialization is being coupled with the schemes?
5/28/2015: m-y and p3cat2 get really hot with reflectivity especially overnight. Thompson seemed most realistic compared to actual.
5/28/2015: MY scheme had very high reflectivity in a complex in SW OK early, relative to the others. P3 and P3-2 were similar in structure. All schemes were too aggressive with squall line moving into TX Panhandle and western OK. Thompson and Morrison schemes exhibited lower CAPE overall. Otherwise schemes appeared similar to each other in overall evolution.
5/28/2015: The Thompson microphysics scheme had a larger area of rain in western KS than what was observed at 12z. Also, the storms in eastern KS were too weak. The other schemes captured what happened well through 12Z,

especially the Morrison scheme.
12-0Z: Again, the Thompson scheme was too aggressive with the anvil reflectivity as well as having a larger rain area in eastern KS/western MO than what was observed. A common theme among the other members was that they were too wide with the area of reflectivities yellow and above.
CAPE: Morrison and Thompson CAPEs are lower, but still higher than what was actually observed.

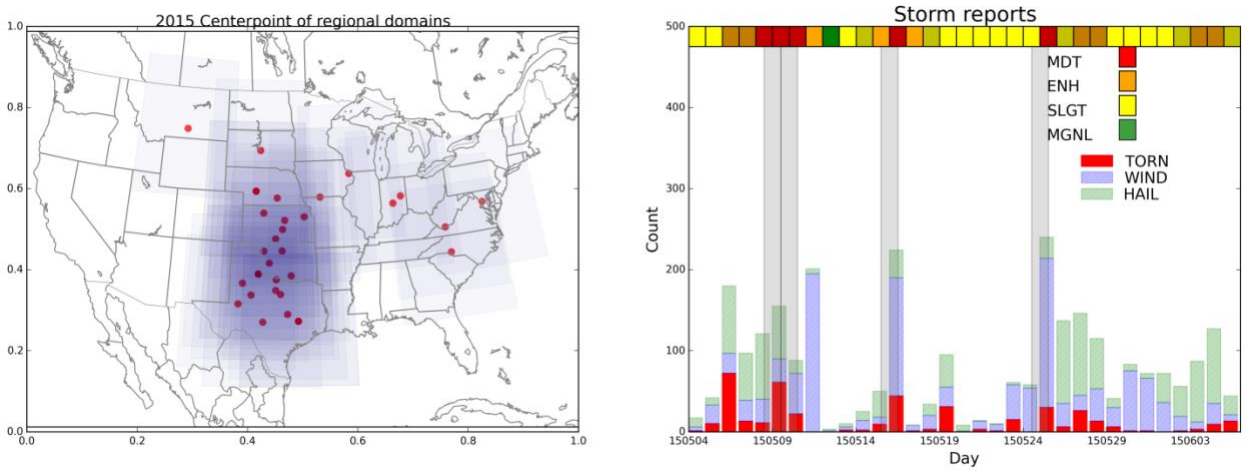

Figure 17. Locations of mesoscale areas of interest during SFE2015 (left panel), where the red dots indicate the daily centerpoints, and the blue boxes highlight the domain. Storm report counts along with SPC operational maximum categorical risk for each day during SFE2015 are shown in the right panel.