SPRING FORECASTING EXPERIMENT 2014

Conducted by the

EXPERIMENTAL FORECAST PROGRAM

of the

NOAA HAZARDOUS WEATHER TESTBED

http://hwt.nssl.noaa.gov/Spring_2014/

HWT Facility – National Weather Center
5 May - 6 June 2014

Preliminary Findings and Results

Israel Jirak[1], Mike Coniglio[2], Adam Clark[2,3], James Correia[1,3], Kent Knopfmeier[2,3], Chris Melick[1,3], and Steve 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

1. Introduction

The 2014 Spring Forecasting Experiment (SFE2014) was conducted from 5 May – 6 June by the Experimental Forecast Program (EFP) of the NOAA/Hazardous Weather Testbed (HWT). SFE2014 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. SFE2014 aimed to address several primary goals:

- Explore the feasibility of creating 1-h convective outlooks for total severe,
- Explore the ability to generate 3-h convective outlooks for individual hazards (tornado, wind, and hail),
- Compare multiple convection-allowing ensembles and identify strengths and weaknesses of the different configurations, initializations, and perturbation strategies,
- Examine convection-allowing ensemble forecasts into Day 2 and assess their guidance for generating outlooks,
- Evaluate EMC parallel CAMs (HiResW WRF-ARW, HiResW NMMB, and NAM CONUS Nest) and compare them to operational versions,
- Investigate the use of HAILCAST (hail growth model) incorporated into WRF as a tool for predicting the size of hail,
- Test the sensitivity of WRF-ARW to new double-moment microphysics schemes: Milbrandt-Yau and Predicted Particle Properties (P3),
- Identify differences in performance between the Met Office Unified Model and WRF-ARW convection-allowing runs, and
- Explore the utility and feasibility of visualizing 3-D CAM fields in near real-time and compare to radar-observed storm structure.

This document summarizes the activities, core interests, and preliminary findings of SFE2014. 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_2014/HWT_SFE_2014_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 SFE2014 along with a description of the daily activities, and Section 3 reviews the preliminary findings of SFE2014. Finally, a summary can be found in Section 4.

2. Description

a) Experimental Models and Ensembles

Building upon successful experiments of previous years, SFE2014 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, in support of the NWS Weather-Ready Nation initiative. As in previous experiments, a suite of new and improved experimental convection-allowing model (CAM) guidance was central to the generation of these forecasts. 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.

New to the experimental numerical guidance for SFE2014 was the inclusion of 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 NCEP Short Range Ensemble Forecast (SREF) system initialized at 2100 UTC 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 multiple-physics ensemble, SPC forecasters and NSSL scientists are very familiar with the behavior of the NSSL-WRF physics, and this will allow for the isolation of spread contributed only by ICs/LBCs.

ii. CAPS Storm-Scale Ensemble Forecast System

As in previous years, the University of Oklahoma (OU) Center for Analysis and Prediction of Storms (CAPS) provided a 0000 UTC-initialized 4-km grid-spacing Storm-Scale Ensemble Forecast (SSEF) system. The 2014 SSEF system at 0000 UTC included 20 WRF-ARW members with 12 “core” members having IC/LBC perturbations from the NCEP SREF system along with varied physics. These forecasts ran out to 60 hours for the first time this year in support of the Day 2 experimental outlooks. Seven of the remaining members were configured identically, except for their microphysics parameterizations (four members) and turbulent-mixing (PBL) parameterizations (three members). All runs assimilated available surface and upper air observations along with WSR-88D reflectivity and velocity data (except for one member), using the ARPS 3DVAR/Cloud-analysis system. Hourly maximum storm-attribute fields (HMFs), such as simulated reflectivity, updraft helicity, and 10-m wind speed, were generated from the SSEF and examined as part of the experimental forecast process.

Similar to last year, a SSEF system initialized at 1200 UTC was also available for use in the forecasting activities. Computing resources for running the 1200 UTC members in real time were more limited than for the 0000 UTC ensemble, so only an 8-member subset was run at 1200 UTC. The eight members of the 1200 UTC SSEF system had the same configuration as eight members from the 0000 UTC ensemble to allow for a direct comparison of the change in skill between the two ensembles initialized 12 hours apart. Furthermore, the reduced number of members in the 1200 UTC SSEF was closer to the number of members in the other convection-allowing ensembles for a more equitable comparison of the spread and skill characteristics of these sets of forecasts.

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 in the near-term because of computational limitations in NOAA. Similar to the SSEF system, HMFs were produced from the SSEO and examined during SFE2014. All members were initialized as a “cold start” from the operational NAM – i.e., no additional data assimilation was used to produce ICs.

iv. Air Force Weather Agency 4-km Ensemble

The U.S. Air Force Weather Agency (AFWA) runs a real-time 10-member, 4-km grid spacing WRF-ARW ensemble, and these forecast fields were available for examination during SFE2014. 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. UKMET Convection-Allowing Model Runs

The Unified Model (UM) is a generalized NWP system developed by the Met Office that is run at multiple time/space scales ranging from global to storm-scale. Two fully operational, nested limited-area high-resolution 0000 (0300) UTC versions of the UM run at 4.4 (2.2) km horizontal grid spacing were supplied to SFE2014 with forecasts through 48 (45) hrs. The 4.4-km CONUS run took its ICs/LBCs from the 0000 UTC 17-km global configuration of the UM while the 2.2 km run was nested within the 4.4 km model over a slightly sub-CONUS domain. Both models had 70 vertical levels (spaced between 5 m and 40 km), and the mixing scheme used is 2D Smagorinsky in the horizontal and the boundary layer mixing scheme in the vertical with single moment microphysics. The 4.4 km model used a convective parameterization scheme that limits the convection-scheme activity, while the 2.2 km model did not utilize convective parameterization.

b) Daily Activities

SFE2014 activities were focused on forecasting severe convective weather with two separate desks, the SPC Severe Desk and the NSSL Development Desk, generating different forecast products at different temporal resolution. Forecast and model evaluations also were an integral part of daily activities of SFE2014. 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. On the SPC Severe Desk, the full-period forecast 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. This was new to SFE2014, as past experiments had only produced combined probabilities of hail, wind, and tornadoes ("total severe") over this time period. On the NSSL Development Desk, a separate Day 1 forecast was made for total severe probabilities valid over the same period.

Each desk then manually stratified their respective Day 1 forecasts into periods with higher temporal resolution. The SPC Desk generated separate probability forecasts of large hail, damaging wind, and tornadoes valid for three periods: 1800-2100 UTC, 2100-0000 UTC, and 0000-0300 UTC. As an alternative way of stratifying the Day 1 forecast, the Development Desk generated probability forecasts of total severe valid hourly from 1800-0300 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 with primarily categorical forecast products (e.g., mesoscale discussions and convective watches) and to begin to explore ways of seamlessly merging probabilistic severe weather outlooks with probabilistic severe weather warnings as part of the NOAA Warn-on-Forecast initiative (Stensrud et al. 2009).

In addition to the complete suite of observational and model data available in SPC operations, each desk also had first-guess guidance 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 temporally disaggregated into the 3-h periods (1800-2100 UTC, 2100-0000 UTC and 0000-0300 UTC) using the calibrated hazard guidance to provide a first guess for the three forecast periods. In addition, hourly probabilities of total severe were generated from the SSEO, NSSL-WRF, and CAPS SSEF ensembles to serve as first-guess fields for the human-generated forecasts at the Development Desk.

The higher temporal resolution forecasts were also generated differently at the desks. Participants at the SPC Desk jointly discussed and developed the forecast using NMAP software on the N-AWIPS workstations. Each participant at the Development Desk generated their own short-time-window forecasts (i.e., human-generated forecast ensemble) on Google Chromebooks using a web-based tool to generate their own hourly probability forecasts of total severe over the 9-hour period. The participant forecasts were also compared to a "control" forecast issued by an experienced "lead forecaster" using N-AWIPS at the Development Desk (e.g., Fig. 1).

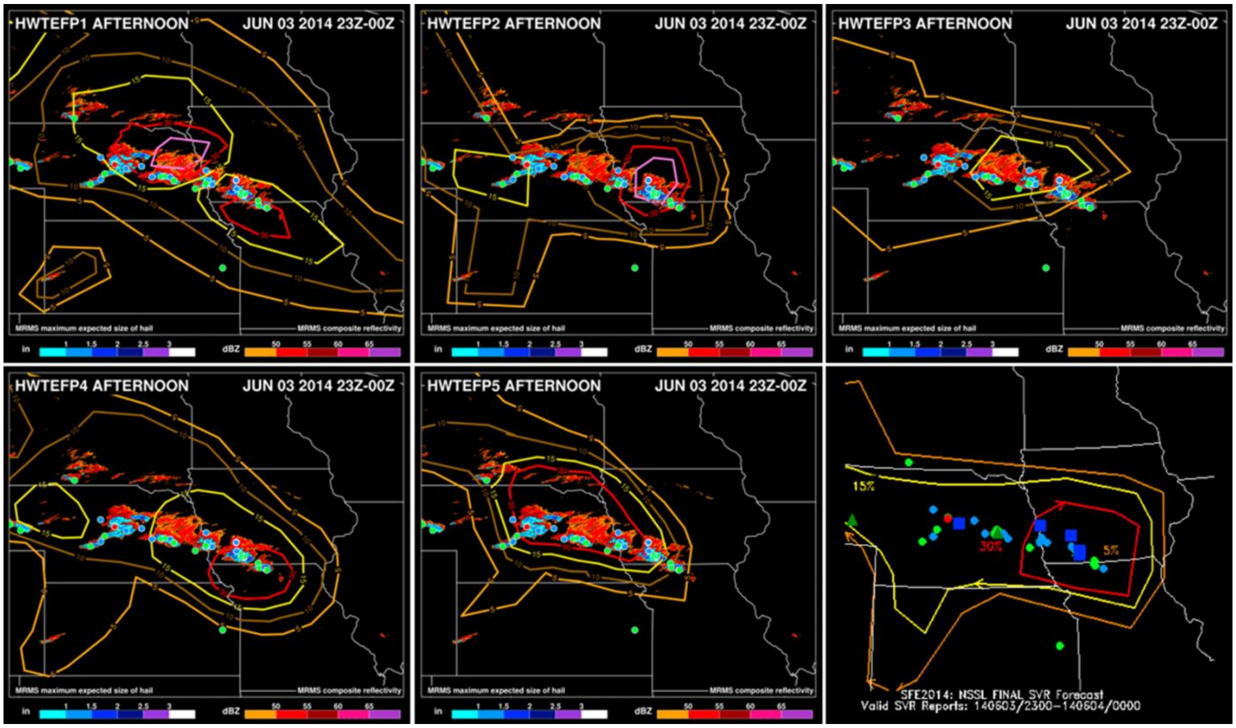


Figure 1. One-hour probability forecasts of total severe valid 23Z 3 June to 00Z 4 June. The forecasts made by the lead forecaster are shown in the lower right panel with forecasts from the participants shown in the other panels. Local storm reports are plotted along with observed composite reflectivity ≥ 45 dbZ (shades of red) and maximum estimated size of hail ≥ 0.5 in (shades of blue).

Producing any severe weather forecast into Day 2 was relatively new to the SFE2014, having not been done over the last decade. The goal was to explore the feasibility of issuing forecasts of individual severe storm hazards into Day 2, where current SPC operational forecasts for Day 2 (and beyond) only consider probabilities of total severe. If time allowed, both desks had the opportunity to examine operational guidance and experimental CAM guidance for the Day 2 period. Generally, only the SPC desk was able to generate Day 2 forecasts for large hail, damaging wind, and tornadoes on some days.

Finally, each desk examined observational trends and morning/afternoon model guidance to update their respective short-time-window forecasts made earlier in the day. Only the forecasts valid from 2100-0300 UTC were updated, as they were issued at 2100 UTC. These forecasts were digitized and shared with the Experimental Warning Program (EWP) for use in preparation for their activities.

ii. Forecast and Model Evaluations

While much can be learned from examining model guidance and creating forecasts in real time, an important component of SFE2014 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, reports of severe weather, NWS warnings, and radar-estimated hail sizes and storm rotation tracks 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 calibrated guidance, temporal disaggregation first-guess guidance, and 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 objectively evaluated in near real-time using Critical Success Index (CSI) and Fractions Skill Score (FSS) based on the local storm reports (LSRs) as the observed verification database. CSI was calculated at two fixed-probability thresholds used in SPC operational outlooks. For the first time, individual hazards of tornado, wind, and hail were also considered separately. Comparisons of results from the experimental forecasts to the first-guess automated fields were also made possible. The utility of the statistical verification metrics in assessing forecast skill for longer and shorter time periods was explored by comparing the scores to the subjective evaluations by the participants.

Model evaluations for SFE2014 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 evaluations of new microphysics schemes available in WRF-ARW and newly updated schemes provided by the developers. There were also 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. Convection-allowing ensembles initialized at 1200 UTC were utilized in making the afternoon update forecasts, and forecasts from those runs were compared to 0000 UTC-initialized ensembles on the following day. 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 preliminary storm reports of hail, wind, and tornadoes. In addition, two of the 0000 UTC ensembles (SSEF and AFWA) had forecasts extending out to 60 hours, which allowed for a first-time comparison of guidance on Day 2 versus Day 1 for these ensembles.

Finally, a new product for evaluation this year was the HAILCAST algorithm, which was used to provide explicit prediction of maximum hail size in convective storms. HAILCAST was coupled to WRF-ARW and used explicitly predicted convective cloud and updraft attributes to determine the growth of hail from initial embryos. Implementation of HAILCAST in the WRF-ARW framework is described in Adams-Selin (2013) and is based on the algorithms described by Brimelow (2002) and Jewell and Brimelow (2009). Explicit prediction of hail size from the HAILCAST model within the NSSL-WRF was evaluated against storm reports and the WSR-88D-derived maximum expected size of hail (MESH) product developed by NSSL.

3. Preliminary Findings and Results

a) Evaluation of Hourly Total Severe Forecasts

On the Development Desk, the preliminary (morning) and final (afternoon) probabilistic forecasts issued by the lead forecaster were evaluated against probabilities generated from proxy severe events in the NSSL ensemble forecasts (the "first-guess" probabilities). For the evaluation, the 1-hour forecasts were split into three periods, 1800-2100 UTC, 2100 -0000 UTC, and 0000-0300 UTC. Participants had five options for comparing the preliminary forecasts to the first-guess forecasts; much worse, worse, the same as, better, or much better. The evaluation was weighted heavily toward local storm reports, but severe weather watches and warnings, observed composite reflectivity, and tracks of the MESH were also examined for additional guidance. Given that the first-guess probabilities (when available) were initially far too high as the method for generating them was still under development, the preliminary forecast was almost always rated the same or better than the first guess in the first few weeks of SFE2014 (Fig. 2). However, as refinements to generating the first-guess probabilities were made during the experiment, there were periods when the first guess forecasts were rated better than the preliminary forecasts. During the periods spanning 1800-0000 UTC, the human preliminary forecasts were rated better than the first-guess forecasts more often than not. However, for 0000-0300 UTC, there were 6 cases when the preliminary forecast was rated worse than the first guess forecast. Figure 2 summarizes the ratings from this part of the evaluation; however, caution is advised for interpreting or generalizing these results, as much work still needs to be done in calibrating and refining the method for generating both the first guess probabilities and the human-generated probabilities valid for 1-hr forecast periods.

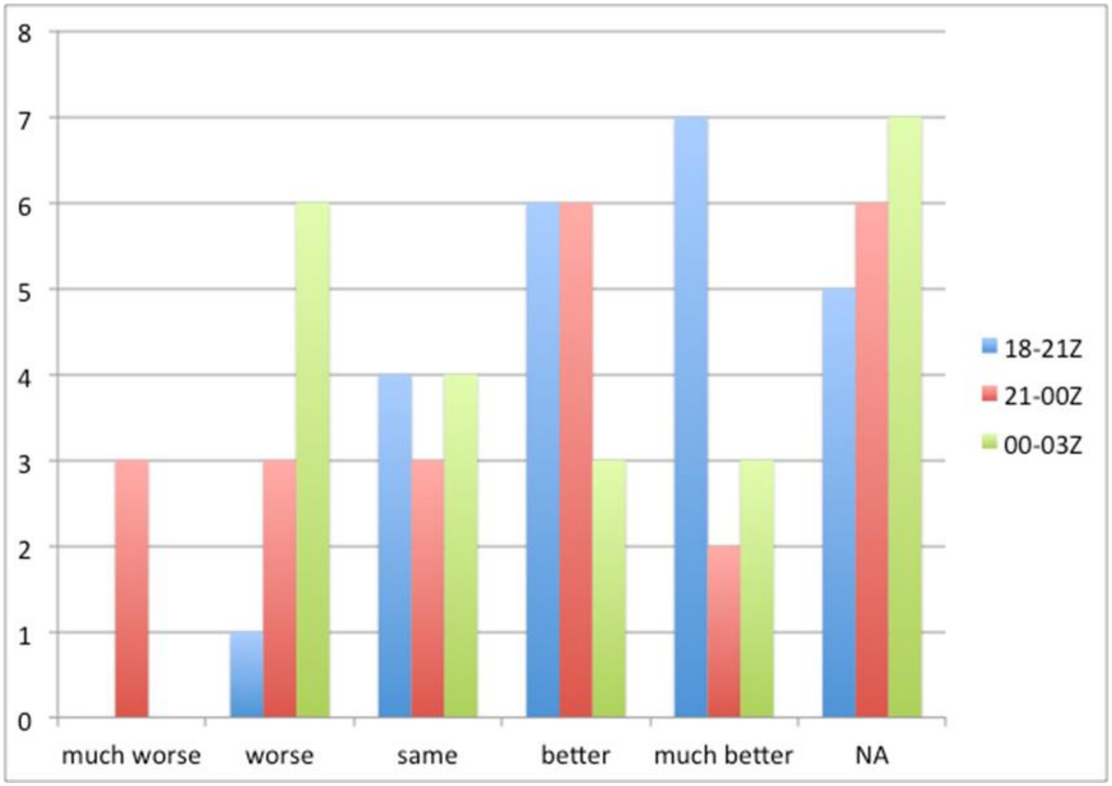


Figure 2. Number of subjective ratings of the preliminary forecast compared to the first-guess forecast.

The preliminary and final 1-hour forecasts were then compared to each other using the same rating system as described previously. Final 1-hour forecasts were only made for the 2100-0300 UTC period, as these were issued just before 2100 UTC. Not surprisingly, the final forecasts in the 2100-0000 UTC period were rated the same or better than the preliminary forecasts all but once, suggesting that the forecasters were skillful in improving their morning forecasts (Fig. 3). This was likely related to the availability of updated real-time observational data, including satellite and radar imagery, prior to the forecast issuance. Not surprisingly, the improvements made in the 1-hour forecasts became much less frequent in the 0000-0300 UTC period. In fact, there were just as many times (four) when the final forecast was rated worse than the preliminary forecast than when it was rated better than the preliminary forecast. There were 11 cases when the final and preliminary forecast skill was deemed to be the same. The fact that the lead forecaster could not consistently improve upon the preliminary forecast in the 0000-0300 UTC period suggests that the skill in predicting severe weather in these 1-hour periods did not extend beyond three to four hours.

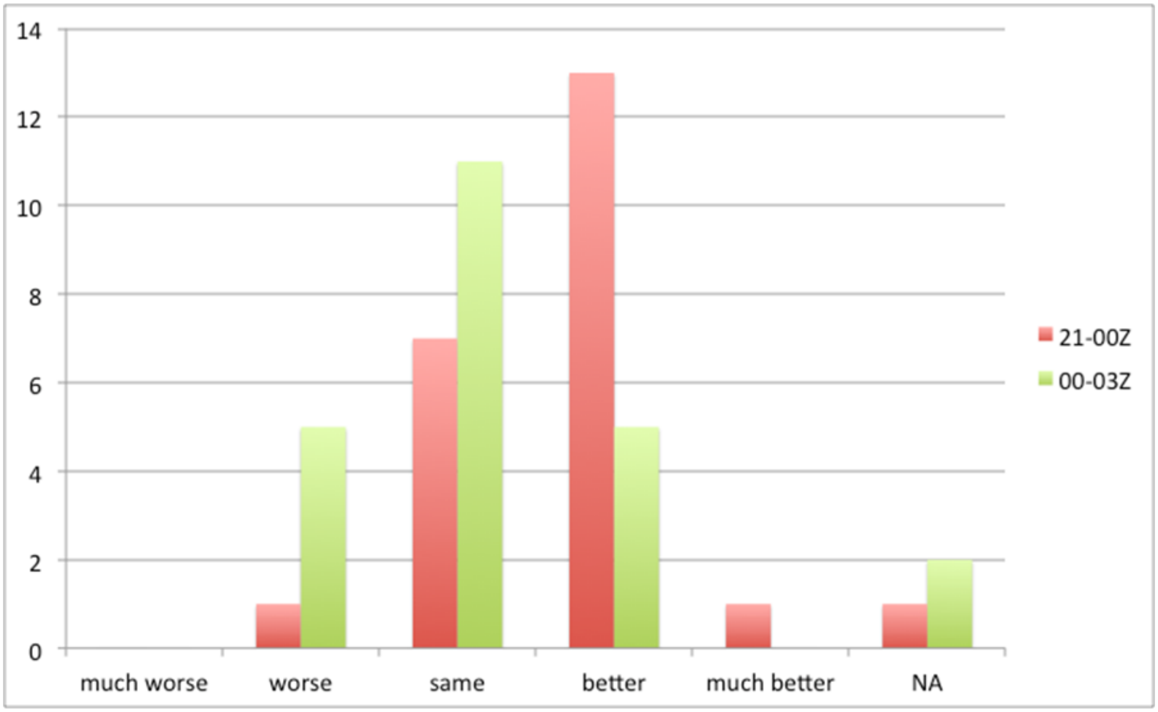


Figure 3. Number of subjective ratings of the final human forecast compared to the preliminary human forecast.

The participants also gave subjective ratings to each of the final 1-hour forecasts. The participants also used practically perfect hindcasts (Hitchens et al. 2013) that were designed for 24-h periods, but generated over 1- hour periods, as guidance. These practically perfect fields were likely too smooth, which was conveyed to the participants during the evaluation process, along with instructions to be harsher than they normally would for 24-h period forecasts. Again, not surprisingly, the forecasts for the 1800-2100 UTC period rated the highest overall, with only five cases of poor or very poor forecasts (Fig. 4). However, the 1-hour forecasts for the 0000-0300 UTC period were rated good only four times (Fig. 4). In many cases the forecaster accounted for increased uncertainty in longer lead times by drawing larger areas, but didn't lower the probabilities accordingly, leading to substantial false alarm areas. Forecasts that attempted to pinpoint locations of convective lines, clusters, etc. often (but not always) missed the area entirely. Again, this suggests that there was limited skill in predicting severe weather in 1-hour windows for these cases.

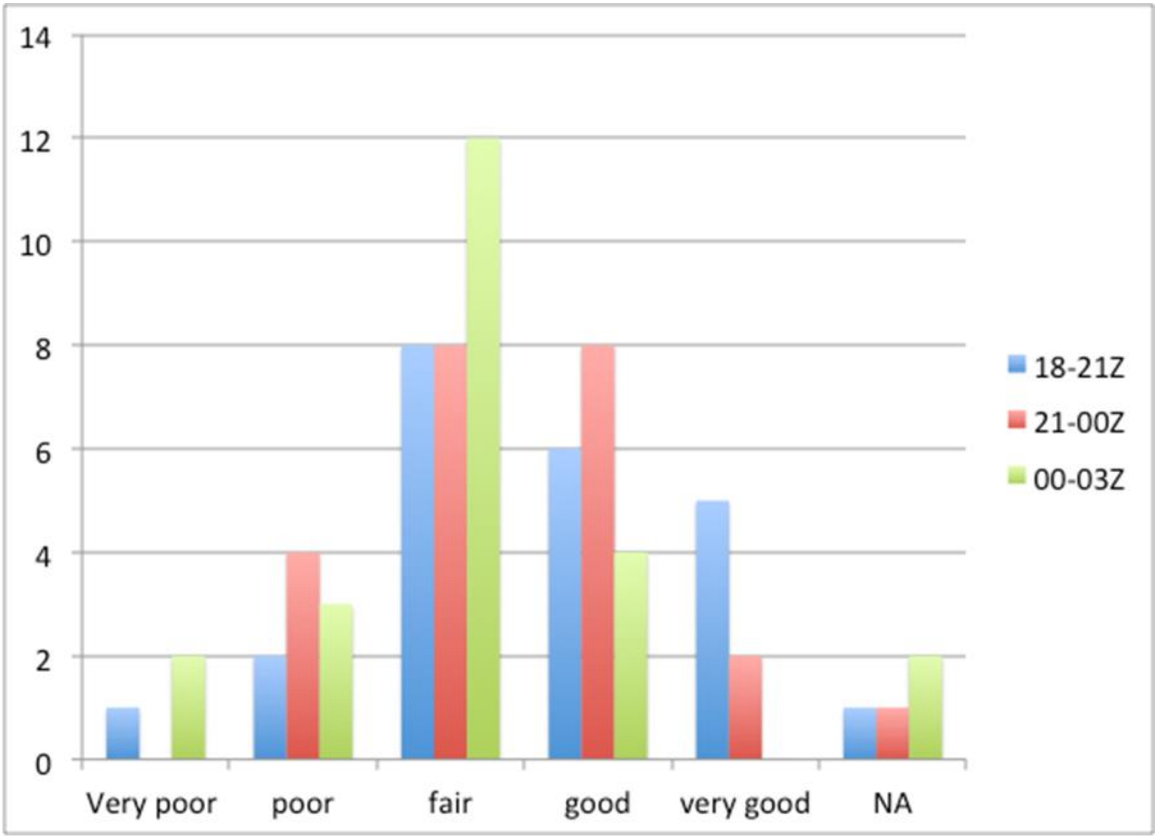


Figure 4. Number of subjective ratings of the final human forecast compared to local storm reports.

The relative skill score (Hitchens et al. 2013) of the forecasts (especially for the full period) was examined to determine if these scores agreed with the subjective impressions of the forecast performance, and whether this metric provided unique information in assessing forecast performance. The survey results were overwhelmingly positive regarding the utility of the relative skill score. Although the relative skill is positively correlated with CSI (Fig. 4), it does provide a more meaningful baseline reference (i.e., practically perfect hindcasts) against which all forecasts are measured. For example, given forecasts on two days with the same CSI, the relative skill can be quite different depending on the coverage and clustering of the reports. Thus, examination of relative skill from a long-term perspective should provide more meaningful information about forecast skill than looking at traditional metrics alone.

b) Evaluation of 3-h Forecasts of Severe Hazards

Similar to the hourly total severe forecasts, the preliminary 3-h severe hazard forecasts (i.e., tornado, hail, and wind) were compared with the first-guess guidance. The first-guess probabilities for the 3-h periods were generated using the temporal disaggregation technique (Jirak et al. 2012) by using the full-period hazard outlook to constrain the magnitude and spatial extent of the 3-h calibrated hazard probabilities (Jirak et al. 2014). The first-guess guidance was available to the participants when making the preliminary forecasts. The preliminary tornado forecasts were most commonly rated the same as the first-guess guidance, except for the 2100-0000 UTC period, when there was an equal number of "better" forecasts (Fig. 5a). The preliminary hail forecasts were more likely to be better than the first-guess guidance in the 1800-0000 UTC period than in the 0000-0300 UTC period (Fig. 5b). For wind, the preliminary forecasts had a more uniform distribution of subjective ratings from "better" to "worse" (Fig. 5c). In fact, the preliminary wind forecast was worse more often than it was better than the first-guess guidance during the 2100-0000 UTC period.

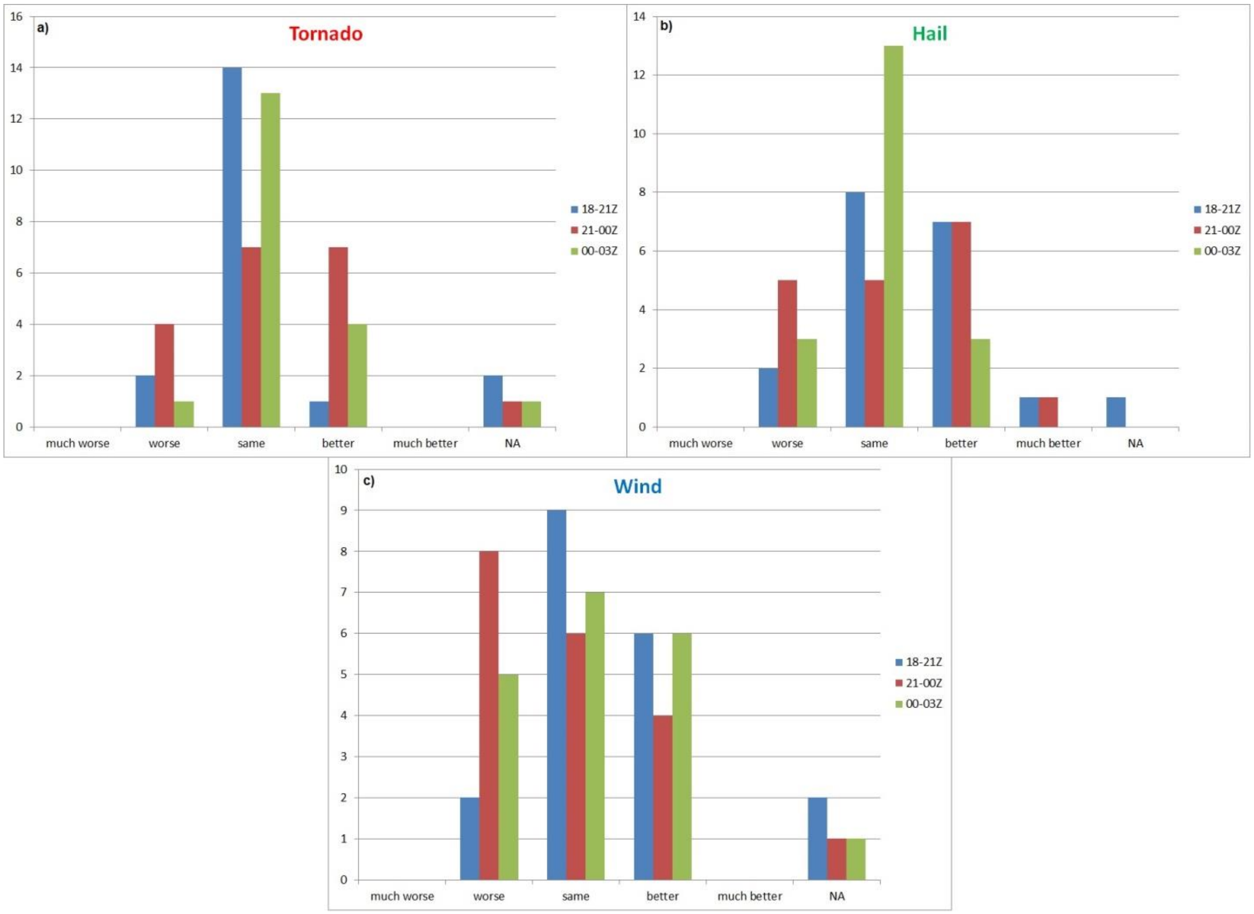


Figure 5. Number of subjective ratings of the preliminary forecast compared to the first-guess guidance for a) tornado, b) hail, and c) wind.

The preliminary and final tornado, wind, and hail forecasts were subjectively compared to determine the relative value of the afternoon forecast updates (Fig. 6). Overall, updating the forecasts in the afternoon generally resulted in similar or better forecast quality. Forecasts were most likely to be improved (i.e., "better" or higher rating) in the 2100-0000 UTC, which is not surprising given that these are shorter-range forecasts issued at 2100 UTC.

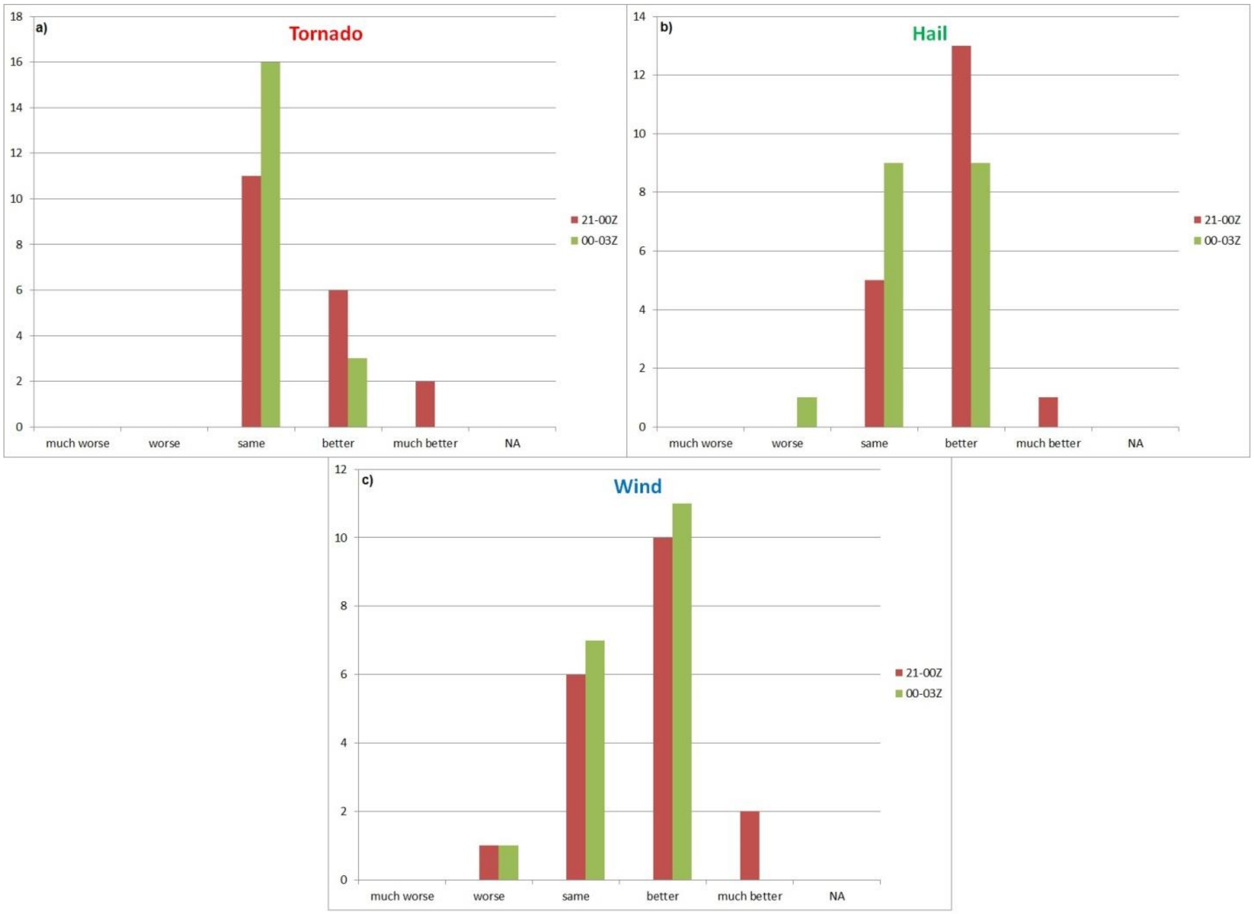


Figure 6. Number of subjective ratings of the final forecast compared to the preliminary forecast for a) tornado, b) hail, and c) wind.

c) Comparison of Convection-Allowing Ensembles

Forecasts from the 0000 UTC NSSL-WRF ensemble were available for examination for the first time in SFE2014, providing an opportunity for comparisons among multiple convection-allowing ensemble designs with varying degrees and types of ensemble diversity. There were two primary components to this comparison of the convection-allowing ensembles: 1) evaluation of neighborhood probabilities of reflectivity ≥40 dBZ and 2) subjective verification of ensemble HMFs relative to preliminary storm reports.

When subjectively comparing the characteristics (timing, location, orientation, magnitude, etc.) of ensemble probabilities to radar reflectivity observations during the 1300-0600 UTC forecast period, the NSSL-WRF ensemble fared very well in terms of ratings (Fig. 7). The NSSL-WRF ensemble had as many “good” ratings as the SSEF, but also had fewer “poor” ratings than the SSEF. For comparison, most of the SSEO and AFWA reflectivity probability forecasts were rated as “fair”.

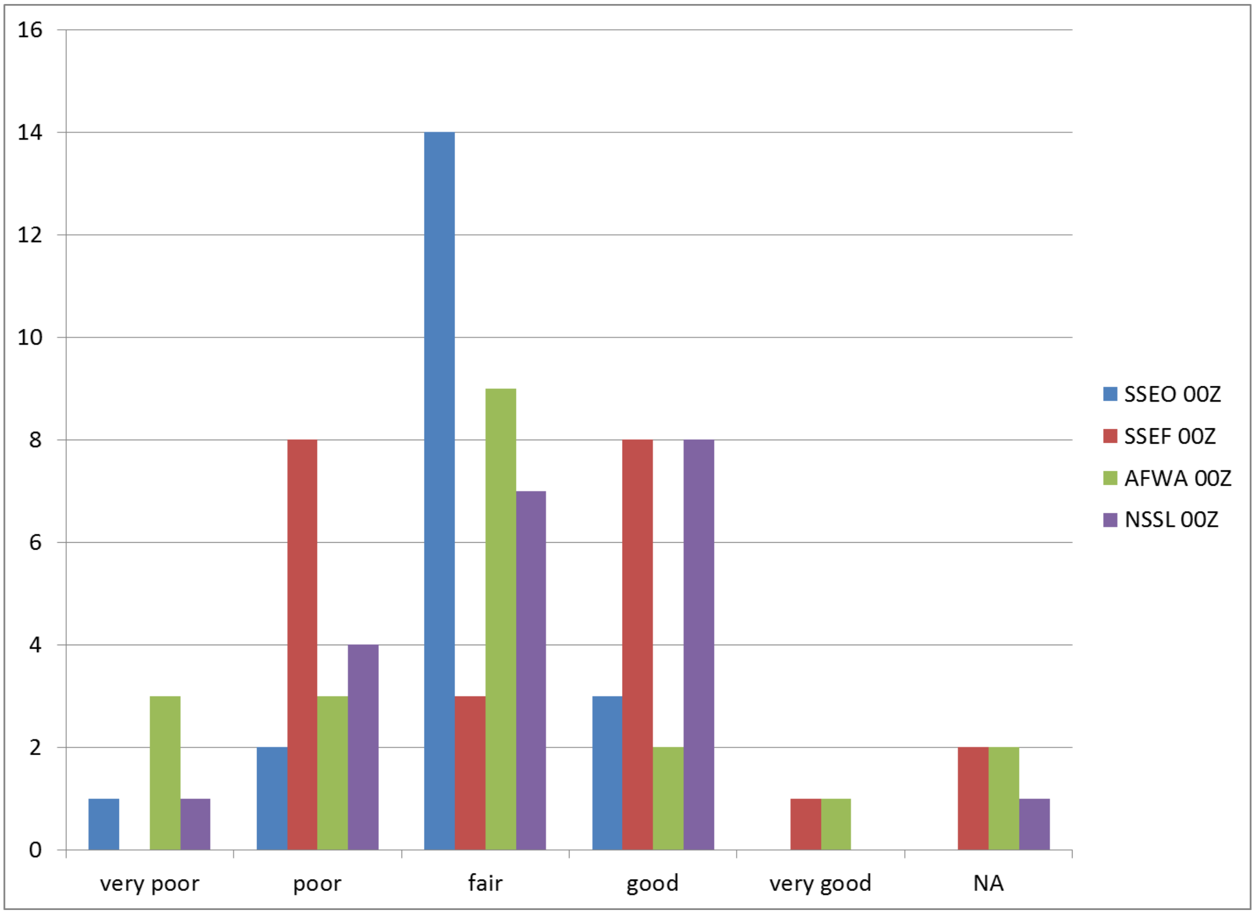


Figure 7. Number of subjective ratings for the ensemble neighborhood reflectivity forecasts compared to observed radar reflectivity.

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 rather similar (Fig. 8). In fact, it is difficult to identify any features that stand out in comparing the subjective ratings of the convection-allowing ensembles other than that all of them more often than not provided useful severe weather guidance (i.e. rating of “fair” or better), including the NSSL-WRF ensemble. This highlights the fact that the complexity of convection-allowing ensemble design does not appear to strongly correspond to the ability of an ensemble to provide useful guidance for severe weather outlooks.

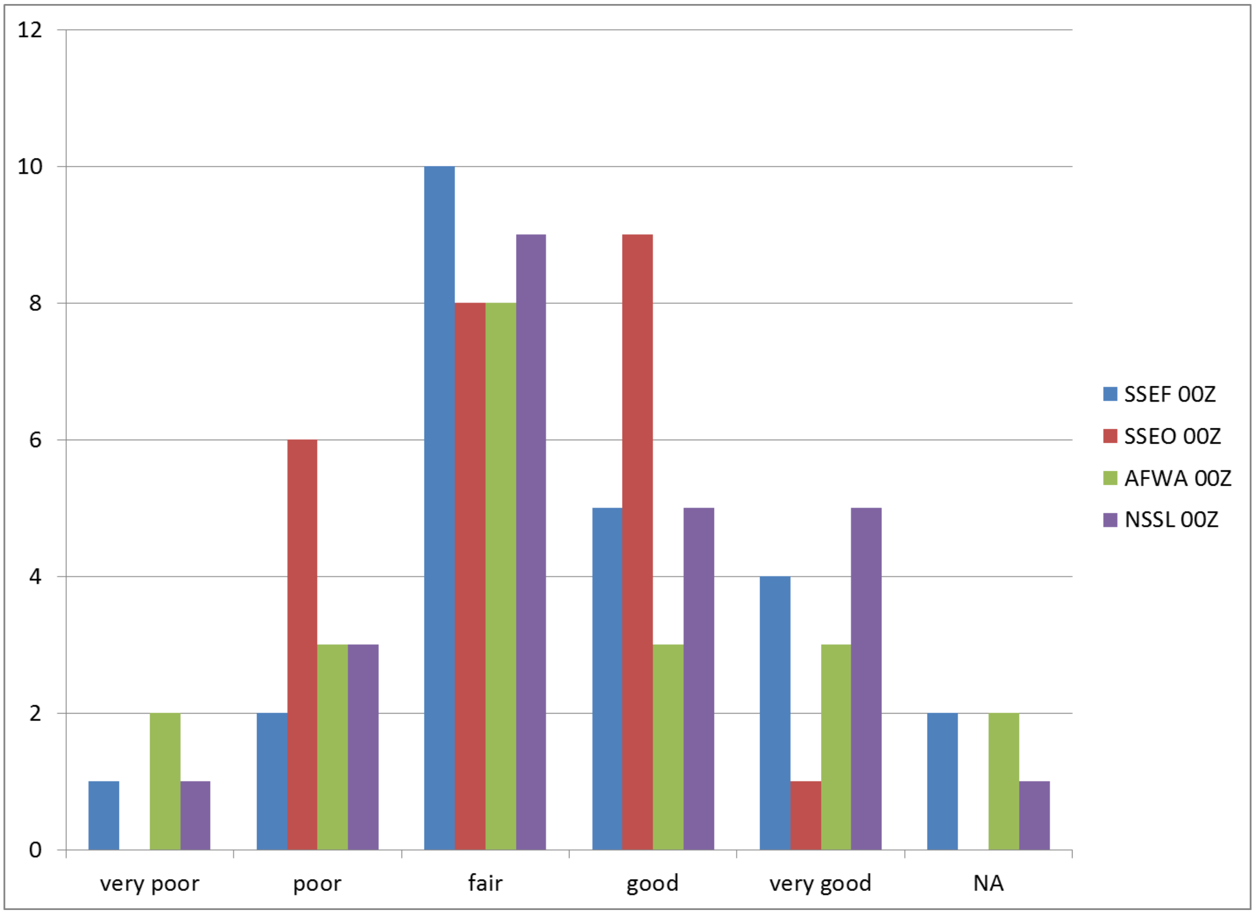


Figure 8. Number of subjective ratings for the ensemble HMF forecasts compared to local storm reports.

d) Convection-Allowing Ensembles for Day 2

Convection-allowing ensembles were examined into the Day 2 period (i.e., f36-f60 from 0000-UTC initialized runs) for the first time during SFE2014. The ensemble output was available on only a limited number of days, owing to computing resource limitations/issues. Nevertheless, the preliminary results from this spring period provided some initial insights. The Day 2 forecasts from the 0000 UTC SSEF were rated the same as or better than the Day 1 forecasts on 9 out of 14 days (Fig. 9). Figure 10 shows the SSEF Day 2 forecast of updraft helicity (UH) valid 0000-0300 UTC (i.e., f48-f51) on 4 June (bottom row) compared to the SSEF Day 1 forecast of UH valid at the same time (i.e., f24-f27). Even though this Day 2 forecast (bottom row) was rated worse than the Day 1 forecast (top row), owing to a slight displacement error in UH tracks and probabilities, there is still value in this Day 2 forecast, and it was rated “good” overall. The Day 2 AFWA ensemble forecasts also fared well in the evaluation with 5 out of 10 rated better than the Day 1 forecasts (Fig. 9). Even though the sample size was very limited, the overall quality of the forecasts on Day 2 from convection-allowing ensembles was better than expected for severe weather guidance during this five-week period in the spring.

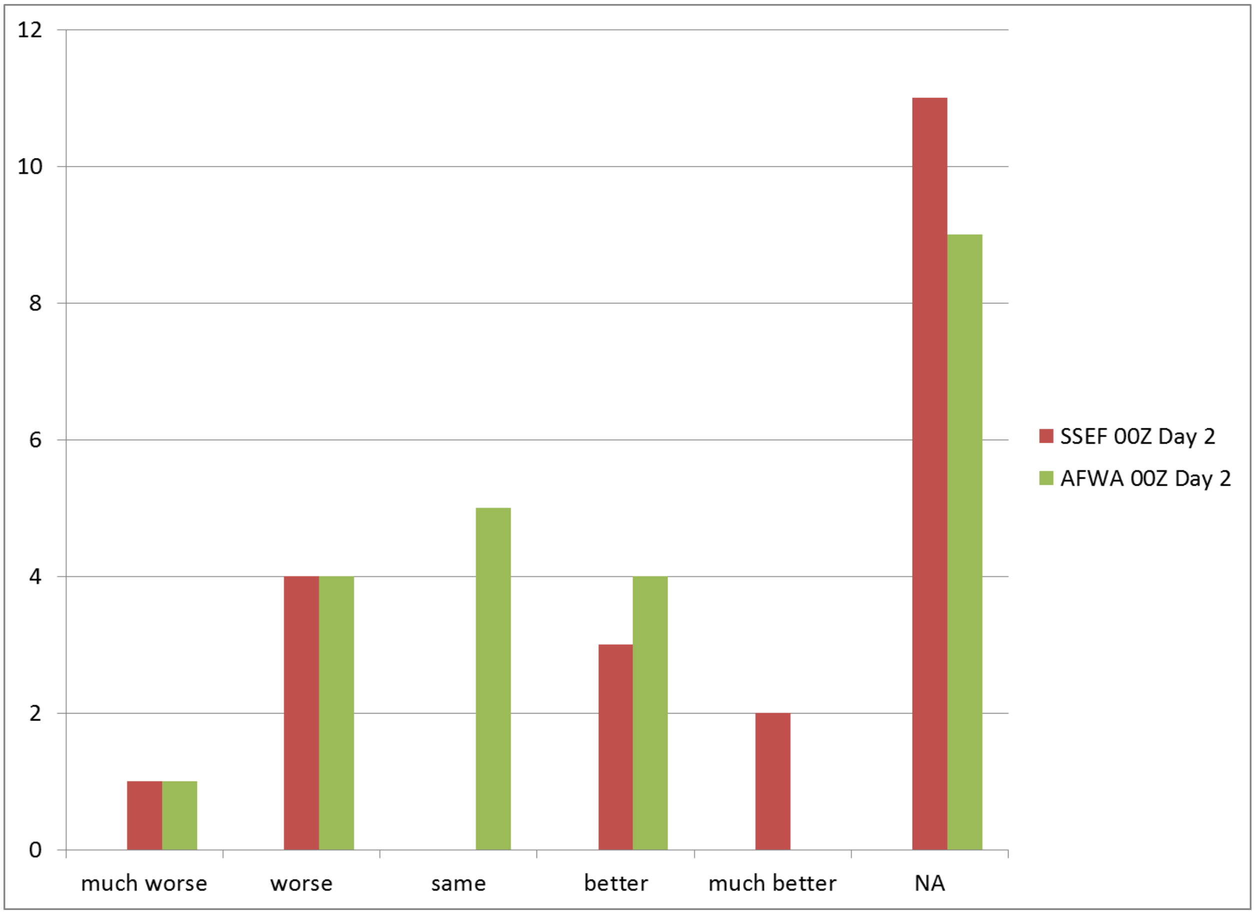
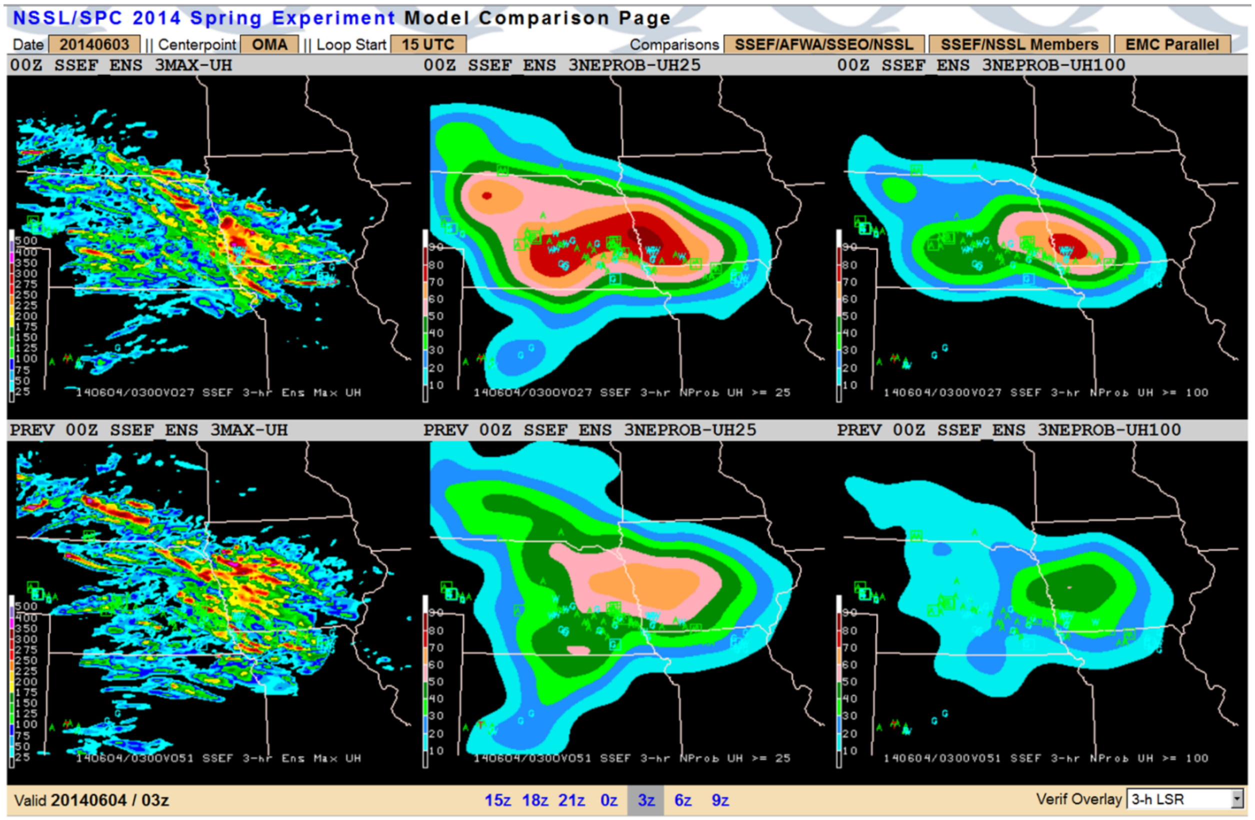


Figure 9. Number of subjective ratings for the Day 2 ensemble forecasts from the SSEF (red) and AFWA (green) compared to the Day 1 forecasts.

e) Evaluation of EMC Parallel CAMs

During SFE2014, the SPC had access to parallel CAMs from EMC for comparison to the operational versions of the CAMs. The parallel versions contained improvements over their operational counterparts and following formal evaluations, they were intended to be implemented operationally by EMC during the summer. Specifically, the parallel HiResW ARW was expanded to full CONUS with increased resolution (4.2-km horizontal grid spacing and 40 vertical levels) compared to the operational version (5.15-km grid spacing and 35 vertical levels). Some other changes to the HiResW ARW included an upgrade to the microphysics scheme from WSM3 to WSM6 and a change in initialization from the NAM to the RAP ICs. The parallel HiResW ARW was subjectively rated the same as or better than the operational HiResW on 18 of 23 days during SFE2014 (Fig. 11) for convective-storm guidance. Figure 12 illustrates an example where the parallel HiResW ARW (upper middle) was rated better than the operational HiResW ARW (upper left).



Date 20140603 || Centerpoint OMA || Loop Start 15 UTC

Valid 20140604 / 03z

Verif Overlay 3-h LSR

Figure 10. SSEF Day 1 (top row) and Day 2 (bottom row) forecasts of 3-h ensemble maximum UH (left column), ensemble neighborhood probability of UH ≥25 m2s-2 (middle column), and ensemble neighborhood probability of UH ≥100 m2s-2 (right column) valid 0000-0300 UTC on 4 June 2014. The severe reports during this 3-h period are plotted as letters in each panel.

The parallel HiResW NMMB was also evaluated during SFE2014. This CONUS upgrade included a change in physics and model core (i.e., from WRF-NMM to NMMB), an increase in resolution (i.e., from 4-km grid spacing and 35 vertical levels to 3.6-km grid spacing and 40 vertical levels), and initialization from the RAP ICs rather than the NAM. Very positive results were also seen for the parallel HiResW NMMB, as it was rated the same as or better than the operational HiResW WRF-NMM on 21 of 23 days during the SFE2014 (Fig 11). Both of the parallel HiResW versions (i.e. ARW and NMMB) were implemented operationally after SFE2014 on 11 June 2014.

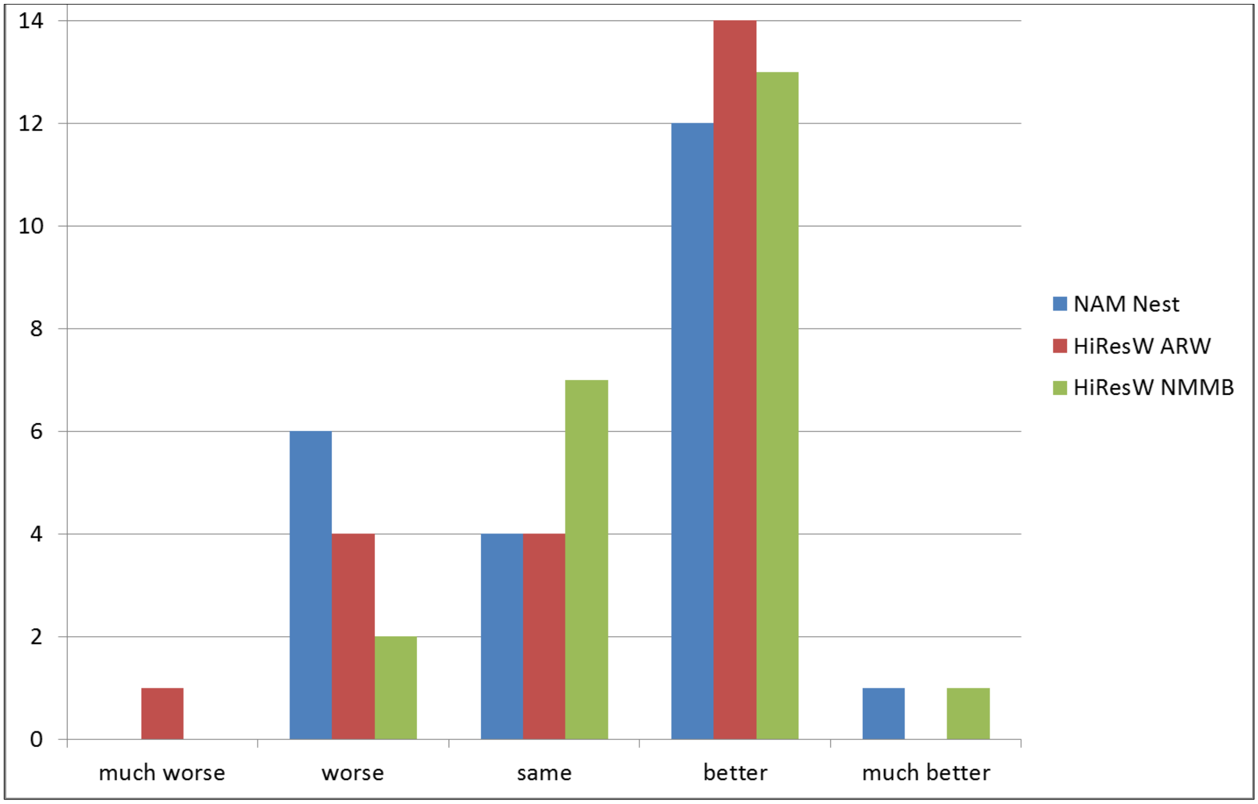


Figure 11. Subjective ratings of the parallel versions of the EMC CAMs (NAM Nest – blue, HiResW ARW – red, and HiResW NMMB – green) compared to the operational versions.

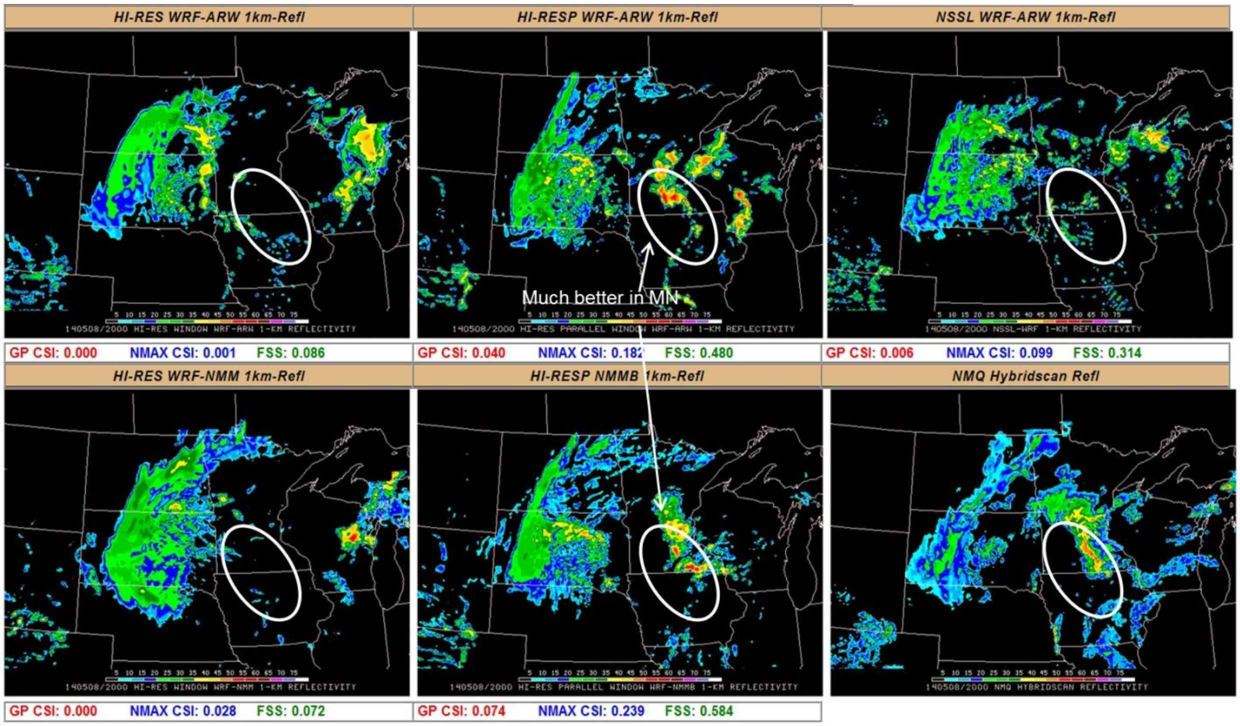


Figure 12. Simulated reflectivity forecasts valid at 2000 UTC on 8 May 2014 for the operational HiResW ARW (upper left), parallel HiResW ARW (upper middle), NSSL-WRF (upper right), HiResW WRF-NMM (lower left), HiResW NMMB (lower middle) and observed reflectivity (lower right) at that time.

A parallel NAM CONUS Nest was also available for evaluation during SFE2014. This parallel version was nested (at 4-km grid spacing) inside an upgraded 12-km parent NAM with an improved microphysics scheme and no convective parameterization. The subjective results were generally positive for the parallel NAM Nest, as participants noted improved structure and intensity of simulated storms over the operational version. In fact, the parallel NAM Nest was rated the same as or better than the operational NAM Nest on 17 of 23 days during SFE2014 (Fig. 11). The NAM was upgraded operationally with this parallel version on 12 August 2014.

f) Investigation of HAILCAST

For the first time during SFE2014, a maximum hail- size diagnostic was output from the various convection-allowing models produced by CAPS and NSSL, which was based on the HAILCAST model coupled to WRF-ARW. The implementation of HAILCAST into WRF-ARW is described by Adams-Selin (2013). 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 are used to determine the initial embryo size. For the formal evaluation activity, explicit predictions of hail size from the HAILCAST model within the NSSL-WRF ensemble were evaluated against storm reports and the WSR-88D-derived MESH product developed by NSSL as part of the Warning Decision Support System – Integrated Information (WDSS-II) suite of algorithms.

Each day, SFE2014 participants were asked the following two questions:

"Using the PHI tool, and focusing on areas of interesting weather, evaluate the HAILCAST forecasts of maximum hail size. First, focus on spatial correspondence. How well do areas of forecast hail correspond to observed hail? Here, we are looking for general spatial agreement, not point-to-point matches."

"Using the PHI tool, and focusing on areas of interesting weather, evaluate the amplitude of the HAILCAST forecasts. How well do the distributions of forecast hail size match the MESH product?"

For each question, participants used ratings of "Excellent", "Good", "Fair", "Poor", or "Extremely Poor". After the first two weeks of the experiment, it became very apparent that HAILCAST substantially over-predicted hail sizes. An example HAILCAST forecast is illustrated in Fig. 13. In this particular case, the NSSL-WRF provided a very skillful forecast of an MCS over central Kansas, but the hail size output from HAILCAST was grossly over-forecast (Fig. 13). Practically every storm contained greater than 1-inch hail. Thus, the feedback was very negative and, although we continued to view the HAILCAST forecasts, we stopped doing the evaluation activity after the third week of SFE2014. As a result, changes were made to HAILCAST after the experiment concluded that resulted in more realistic hail size forecasts. Specifically, rime soaking and variable density options were added, and the dependency on microphysics scheme was removed by using five constant initial embryo sizes, as opposed to those predicted in the schemes themselves. The changes to HAILCAST were implemented in the NSSL-WRF and NSSL-WRF ensemble on 9 July 2014.

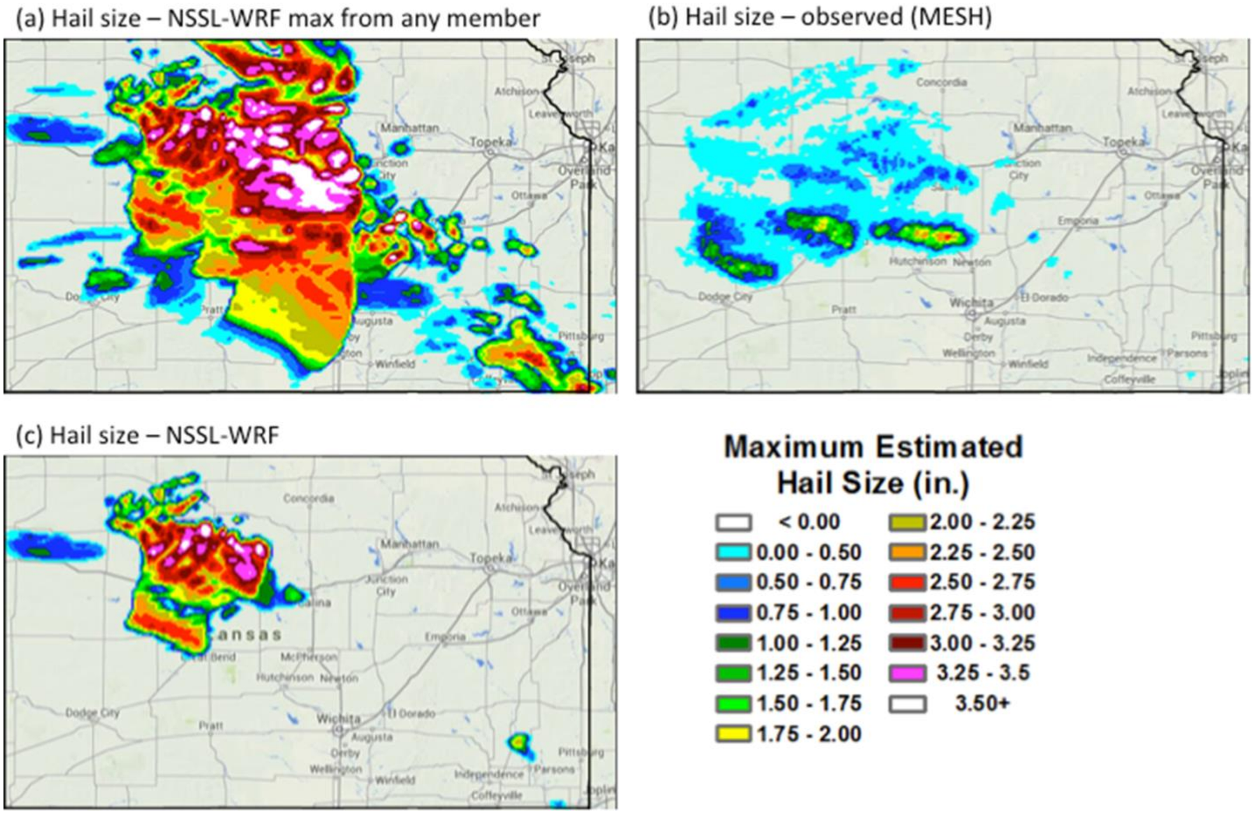


Figure 13. Maximum hail size over the previous hour valid 1000 UTC on 5 June 2014. (a) NSSL-WRF ensemble maximum from any member HAILCAST forecast (0000 UTC 5 June initialization), (b) observed maximum hail size from MESH, and (c) NSSL-WRF control member HAILCAST forecast.

g) 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 SFE2014, the following double-moment microphysics parameterizations were systemically examined: Thompson, Milbrandt and Yau (MY), the Predicted Particle Properties (P3) scheme, Morrison, and a new version of MY that had not yet been made publicly available in a WRF release (MY2). MY2 included an adjustment to the ice-snow balance, which favored snow and significantly reduced the excessive quantities of high ice and broad anvil shields. Also, the graupel-hail balance was adjusted to allow for more hail, and the rate of rain drop break-up was increased, which produces more evaporation and stronger cold pools. The P3 scheme, developed by Hugh Morrison and Jason Milbrandt, was also new to the WRF model and SFE this year. The P3 scheme is unique in that it predicts particle properties (mean density, size, rime fraction, etc.) for a single ice category, unlike other current WRF schemes that partition different types of ice using pre-defined categories like cloud ice, snow, and graupel.

Each day participants were asked the following:

"Comment on any differences and perceived level of skill in forecasts of composite reflectivity, MTR (minus 10 reflectivity), and simulated satellite for the control member CN (Thompson), m17 (MY2), m18 (MY), m19 (P3), and m20 (Morrison) during the 18z-12z period, based on comparisons with corresponding observations."

One general theme among the participant responses (see Table 2 in the Appendix) was that MY2 was an obvious improvement over MY, with convective cloud shields that were more realistic (i.e., warmer and smaller areal coverage) than MY. P3, the only newly developed scheme examined for SFE2014, performed at about the same level as the other schemes, which was encouraging because it is more computational efficient than the other schemes (approximately 9% faster than Thompson and Morrison, and 25% faster than MY and MY2). The general conclusion among participants from previous years was that it was becoming harder to discern systematic differences between the various schemes. However, for this year, Thompson was mentioned most often as being the most realistic. Finally, all the schemes often had a tendency to over-predict CAPE, which has been noticed in previous years, and in a few cases P3 had noticeably higher values of CAPE than the other schemes.

An example case is illustrated in Figures 14-16. In this case, an MCS had developed the night before and moved across southern Missouri during the morning of 5 June. At around 1700 UTC two areas of severe wind reports were observed – one in southern Missouri associated with the convective line and another near Kansas City, which was associated with a wake low that had formed within the stratiform precipitation region of the MCS. Figure 14 shows that all schemes had generally similar depictions of the MCS, but Thompson arguably had the most realistic depiction of the most intense convection associated with the leading convective line, as well as the stratiform precipitation that extended into central and northern Missouri. Interestingly, as can be seen in Figure 15, Thompson was the only scheme that was able to depict the high winds associated with the wake low near Kansas City. Finally, Figure 16 clearly shows the improvement in MY2 relative to MY, with the extent of colder cloud tops reduced and overall temperatures warmed.

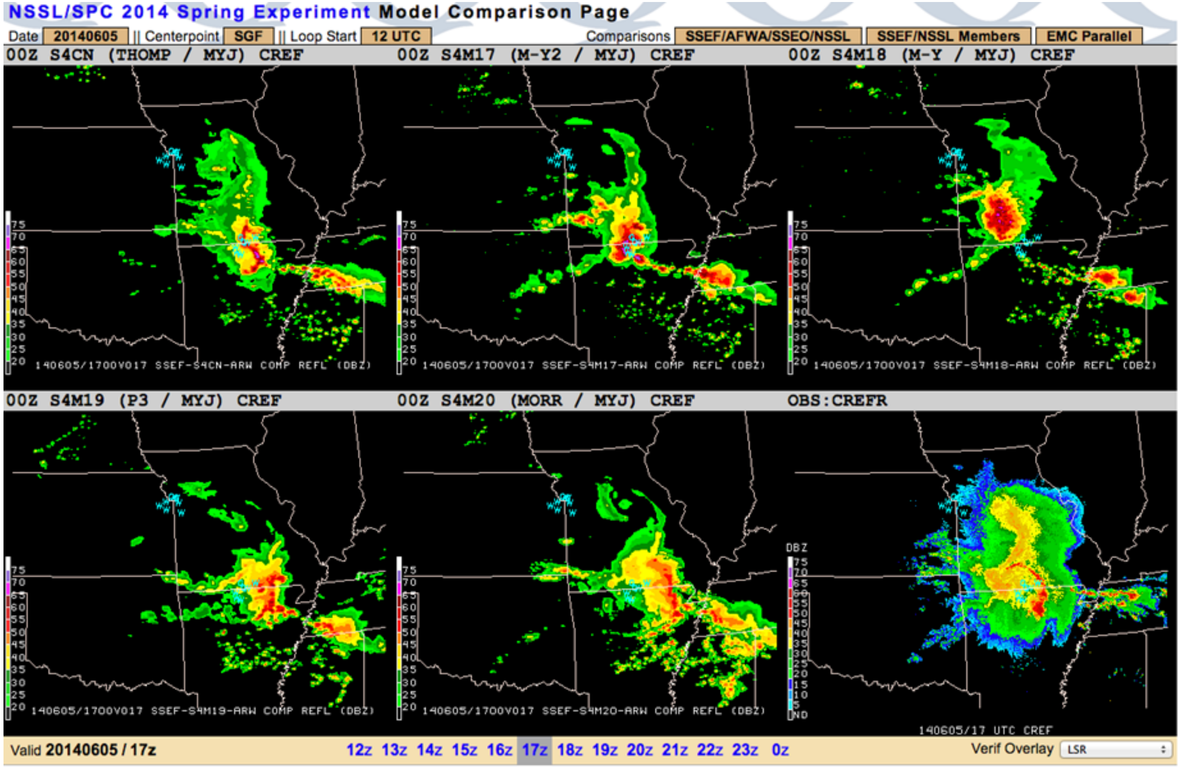


Figure 14. Forecasts and observations of composite reflectivity valid 1700 UTC 5 June 2014. The forecasts were initialized 0000 UTC 5 June and are from the members of the SSEF system configured identically except for their microphysics schemes. The panels include Thompson (upper-left), MY2 (upper-middle), MY (upper-right), P3 (lower-left), Morrison (lower-middle) and observations (lower-right). Locations of observed severe wind reports are indicated by small blue "W"s.

h) Comparison of Met Office CAMs with NSSL-WRF

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 SFE2014 participants were asked the following:

"Using the NSSL Interactive Data Explorer, and focusing on areas of interesting weather, compare the UKMET forecasts to the operational NSSL-WRF. Please provide explanation/description/reasoning for the answer."

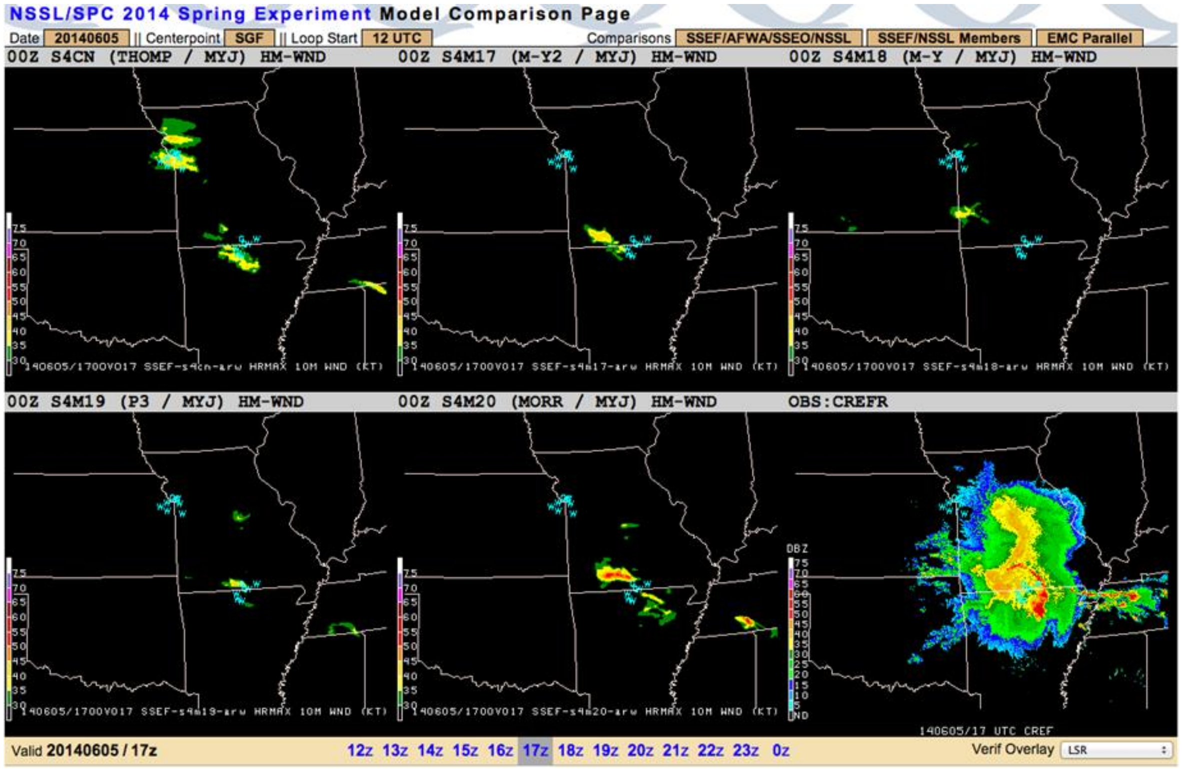


Figure 15. Same as Fig. 14, except forecasts of hourly maximum 10-m wind speed (lower-right panel still includes observed composite reflectivity).

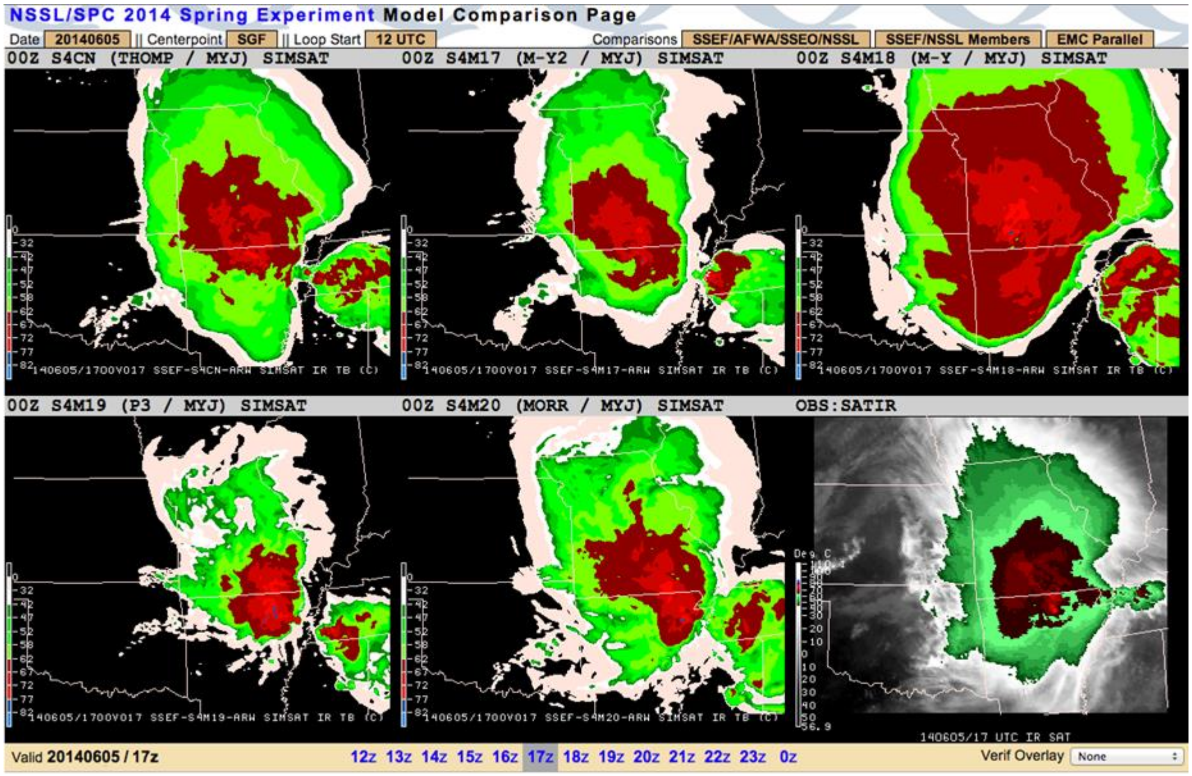


Figure 16. Same as Figure 14, except for simulated IR brightness temperatures.

Participants could select from, “UKMET better than NSSL-WRF”, “UKMET worse than NSSL-WRF”, or “Same”. The responses (20 cases) are summarized in Fig. 17 and Table 3 (Appendix). The majority of the responses (50%) rated the Met Office UM as better, while 30% were “Same” and only 20% (4 cases) rated the Met Office UM as worse than NSSL-WRF. These results were very similar to those from SFE2013 when the Met Office UM was rated better than NSSL-WRF in 50% of the cases, the same in 37.5%, and worse in only 12.5% of the cases. For the cases in which Met Office UM was rated as performing better than NSSL-WRF, there were a variety of reasons. One common theme was that the Met Office UM seemed to often spin up convection much better than NSSL-WRF, which resulted in much improved forecasts within the first 6 to 12 hours for the UKMET. Also, although it was not part of the formal evaluation, the Met Office 2.2-km run was generally perceived as performing even better than the 4.4-km run, especially at longer lead times.

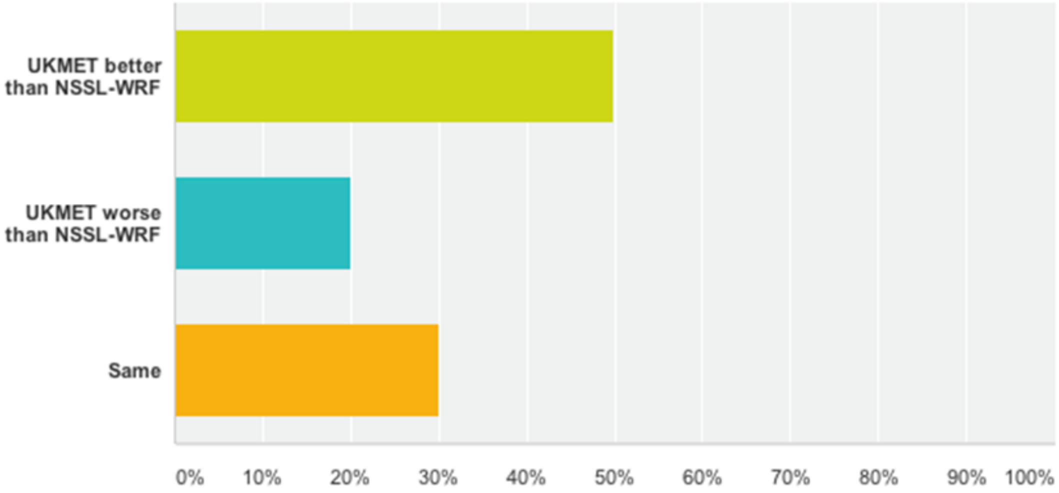


Figure 17. Summary of responses for the NSSL-WRF and Met Office CAM comparisons.

In addition, a striking difference between the NSSL-WRF and UM was noticed for forecast vertical profiles of temperature and moisture when capping inversions were present. The UM oftentimes very accurately depicted the sharp gradients in temperature and moisture at the interface of the boundary layer and elevated mixed layer, while the NSSL-WRF and high resolution WRF model simulations in general had very smoothed out temperature/moisture gradients at this interface (e.g., Figures 18 and 19).

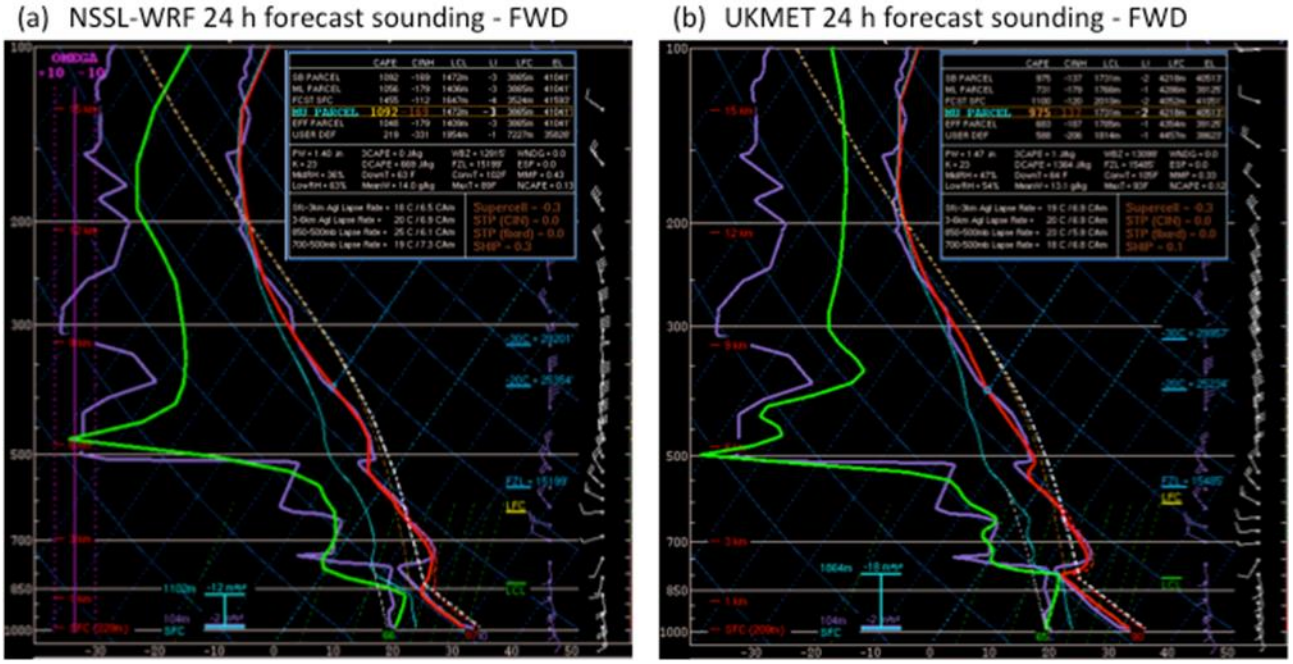


Figure 18. **Forecast soundings valid 3 June 2014 for FWD from 24 h forecasts of the (a) NSSL-WRF, and (b) the UKMET. In both panels, the corresponding observed sounding is overlaid in purple.**

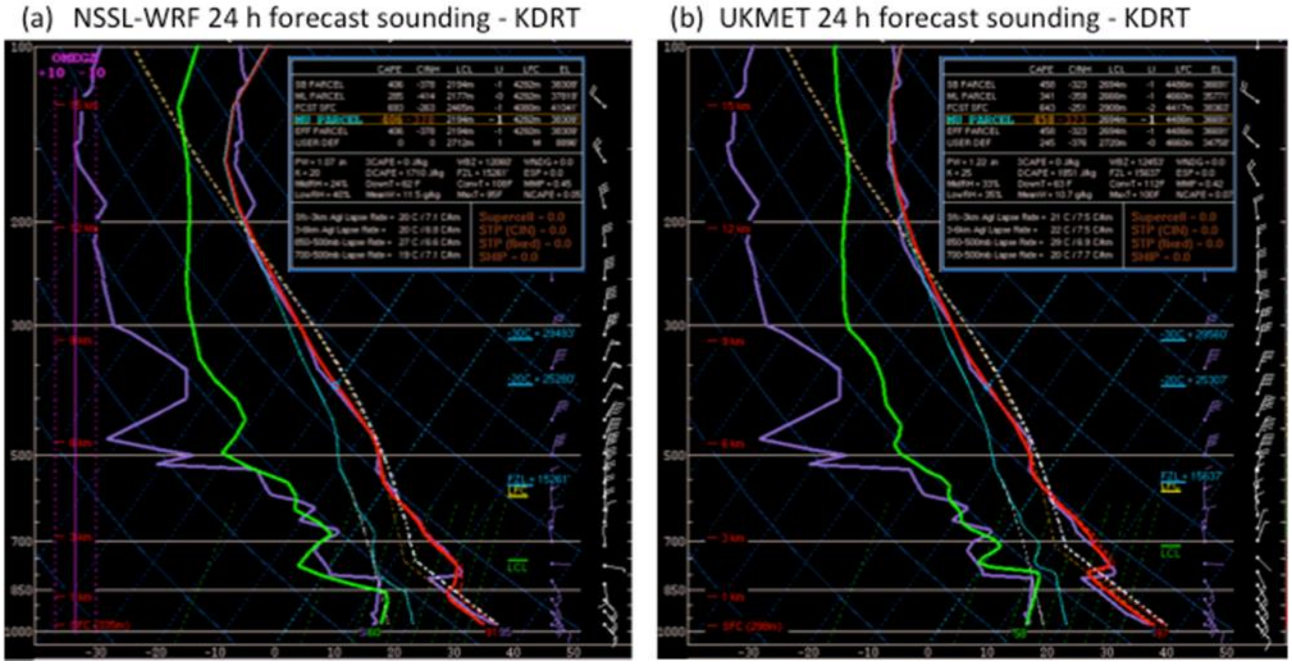


Figure 19. **Same as Fig. 18, except for KDRT.**

i) Exploration of 3-D Visualization

For the first time in the HWT SFE, CAM output was viewed in three-dimensional (3D) displays in near-real time as part of the Development Desk activities. Selected 3D model fields over a mesoscale region at 10-minute output frequency for 18 – 30 h forecasts was interrogated on several days using the WDSS-II display system. The goal was to explore CAM storm characteristics like vertical vorticity, graupel mixing ratio, simulated reflectivity, and cold pools in 3D to learn more about how simulated storms are structured on WRF-ARW convection-allowing grids. Although this 3D output wasn't used in the forecast process, it was surmised that this type of output may give confidence to forecasters in their expectation for convective modes for the day in a similar manner to how simulated reflectivity gave forecasters confidence when it was introduced over ten years ago.

An example of how this output might give confidence to forecasters on the mode and severity components of their forecasts was seen for the 3 June High-Risk day in Nebraska and Iowa. The prominent convective mode, along with the timing of the changes in the prominent convective modes, was a key forecast problem on this day- "Would storms consolidate quickly into lines and bows, or would supercell modes be persistent? Would supercells transition quickly to HP, limiting the strong, long-lived tornado threat, or would classic supercells with strong, long-lived tornadoes occur?" Interrogation of the SSEF control member in 3D suggested there would be a very intense storm along the warm front over eastern Nebraska. The 3D fields showed a persistent hybrid HP-supercell/bowing structure with a vorticity column tilted by a very strong cold pool that extended well south of the main updraft (Figs. 20a and 20b). It also suggested the storm would produce very strong winds near the ground and large hail associated with a very strong and persistent mesocyclone (Fig. 20c). The values of mid-level vertical vorticity seen for this storm were the largest observed for any model storm that was interrogated. Although the model storms developed later than the actual storms, this scenario is very similar to what occurred on this day, with a very damaging wind/hail storm north of Omaha (Fig. 20d).

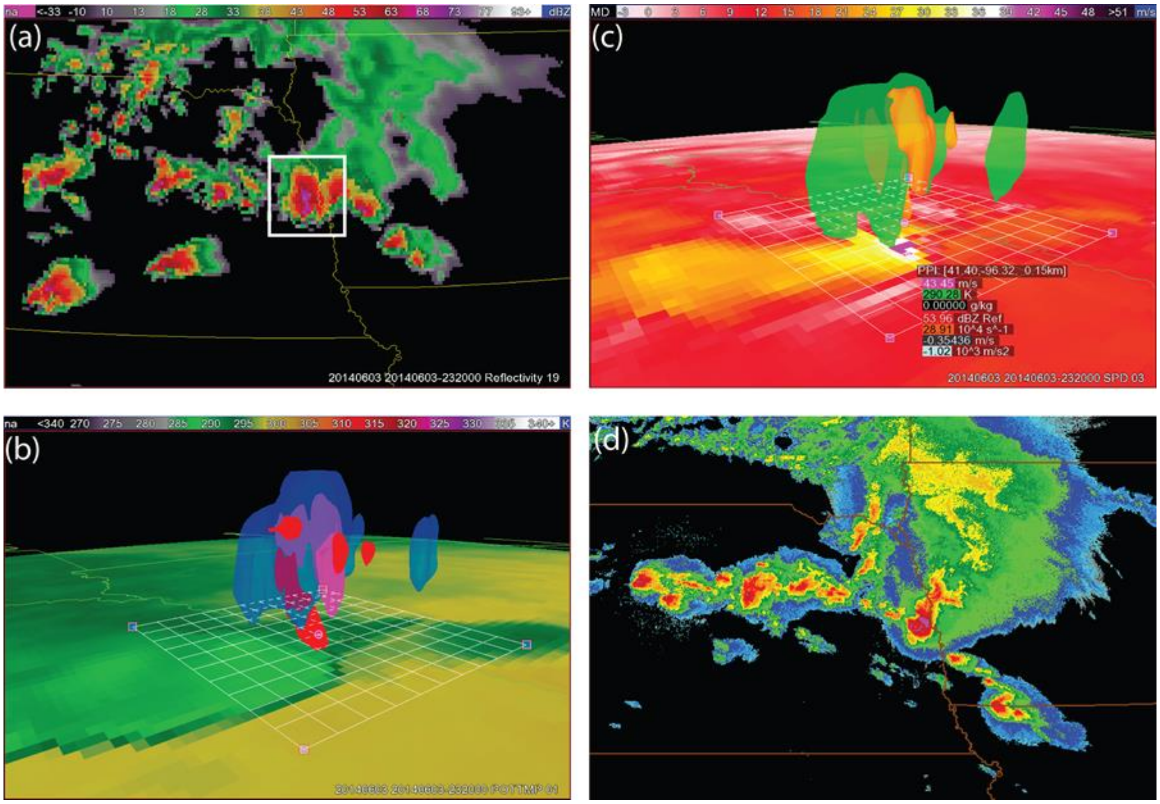


Figure 20. Example of viewing a model storm in 3D using WDSS-II display software. (a) Plan view of simulated reflectivity on model level 19 valid 2320 UTC for the 00Z SSEF control member. The white box encloses the storm interrogated in 3D in panels (b) and (c). In (b), isosurfaces of vertical velocity > 21 m s-1 (light purple), vertical vorticity > 45 x 10-3 (red), and graupel mixing ratio > 4 g kg-1 (blue) are shown from a perspective from the southwest of the storm. The underlying color fill shows potential temperature (K) on the lowest model level. In (c), isosurfaces of the product of vertical velocity and vertical vorticity (m s-2) > 53 x 10-3 (orange) and simulated reflectivity > 54 dBZ (green) are shown from the same perspective as in (b). The underlying color fill in (c) is the wind speed on model level 3 (about 150 m AGL). The observed composite reflectivity valid ar 2110 UTC from the NSSL multi-radar multi-sensor analysis is shown in (d).

There were also interesting spatial variations in the structures of the storms on 3 June. The 0000 UTC SSEF control member suggested that many additional storms would develop farther west along the warm front and take on more discrete supercell structures. There were very strong UH tracks with these storms. However, despite a near colocation of the main updraft and vertical vorticity, and weaker cold pools compared to the storm farther east, the 3D fields suggested that these storms would have only transient low-level circulations because of undercutting by the front and rapid transitions to HP supercell structures. Again, this is very much like what occurred on this day, with HP supercells along the front with only transient tornadoes observed in central Nebraska in the afternoon and evening.

In a few other cases, the evolution of the model storms was consistent with expected storm behavior gleaned from storm-environment relationships. For example, there were several days when supercell modes were expected to be prominent, but high LCLs or weak winds in lower levels expected to keep the tornado threat low. Viewing the storms in 3D in these cases showed columns of vertical vorticity reaching the ground occasionally, but strong cold pools quickly undercut the vorticity. For one case, it was found that the strongest UH was found in the lowest 3 km AGL, with only weak UH above that level, so that the tracks in the traditional 2-5-km integrated UH displays did not reveal the main areas of rotation in the storms. A few storms did develop on this day that showed the strongest rotation in low levels along a QLCS-type system.

4. Summary

The 2014 Spring Forecasting Experiment (SFE2014) was conducted at the NOAA Hazardous Weather Testbed from 5 May – 6 June by the SPC and NSSL with participation from forecasters, researchers, and developers from around the world. The primary theme of SFE2014 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 SFE2014 are listed below:

- Creating hourly probabilistic forecasts of total severe was challenging and time-consuming. Even though preliminary results were promising, additional work is needed to refine and improve the high temporal resolution convective-storm guidance and short-term forecasting methodology.

- Forecasts of severe weather hazards (i.e., tornado, wind, and hail) in 3-h periods were reasonably good with temporally disaggregated output providing useful first-guess guidance, especially at longer lead times.

- Regardless of design, all convection-allowing ensembles examined, including the new NSSL-WRF ensemble, were often able to produce useful guidance for severe weather forecasting.

- Although the number of cases examined was very limited, the Day 2 output from the SSEF and AFWA convection-allowing ensembles was as good as the Day 1 forecasts on several days.

- Improvements were made to the EMC parallel CAMs, especially in terms of simulated storm structure and intensity, and these models have since been implemented operationally.

- The explicit forecast hail-size output from HAILCAST produced a consistent overforecast of hail size in most instances. Modifications to account for rime soaking and variable density have already been made to the algorithm to improve on this bias.

- The updated double-moment microphysics schemes were notably improved, and the new P3 scheme proved to be promising given its computational efficiency.

- Met Office CAMs again performed very well relative to NSSL-WRF runs and were better able to reproduce strong vertical gradients in temperature and moisture near capping inversions.

- The use of 3D visualization software provided useful insight into simulated storm structure and intensity.

Overall, SFE2014 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 SFE2014 are certain to lead to continued progress in the forecasting of severe weather in support of the NWS Weather-Ready Nation initiative.

Acknowledgements

SFE2014 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, AFWA, and the Met Office were vital to the success of SFE2014. In particular, Ming Xue (OU CAPS), Fanyou Kong (OU CAPS), Kevin Thomas (OU CAPS), Keith Brewster (OU CAPS), Yunheng Wang (OU CAPS), Evan Kuchera (AFWA), Scott Rentschler (AFWA), and Steve Willington (UKMET) 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.

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.

SPC/Severe Desk | NSSL/Development Desk

0800 – 0845: **Evaluation of Previous Day's Experimental Forecasts**

- Subjective rating relative to radar evolution/characteristics, warnings, and preliminary reports and objective verification using preliminary reports and MESH:
    - SPC/Severe Desk:
        - Day 1 & 2 full-period probabilistic forecasts of tornado, wind, and hail
        - Day 1 3-h period forecasts and guidance for tornado, wind, and hail
    - NSSL/Development Desk:
        - Day 1 & 2 full-period probabilistic forecasts of total severe
        - Day 1 1-h period forecasts and guidance of total severe

0845 – 1100: **Day 1 Convective Outlook Generation**

- After hand analyses of 12Z upper-air maps and surface charts and discussion:
    - SPC/Severe Desk:
        - Prepare probability forecasts for tornado, wind, and hail valid 16-12Z over mesoscale area of interest
        - Adjust temporally disaggregated first guess for tornado, wind, and hail forecasts valid for 3-h periods: 18-21, 21-00, and 00-03Z; make these available to EWP
    - NSSL/Development Desk:
        - Prepare probability forecasts for total severe valid 16-12Z over mesoscale area of interest
        - Adjust first guess for total severe forecasts valid for 1-h periods: 18-03Z; make these available to EWP

1100 – 1200: **Day 2 Convective Outlook Generation**

- SPC/Severe Desk:
    - Prepare probability forecasts for tornado, wind, and hail valid 12-12Z on Day 2 over mesoscale area of interest
- NSSL/Development Desk:
    - If time allows, prepare probability forecasts for total severe valid 12-12Z on Day 2 over mesoscale area of interest

1200 – 1300: **Lunch**

1300 – 1330: **Briefing**

- Overview and discussion of today's forecast challenges and products
- Highlight interesting features/findings from yesterday including 3-D visualization

1330 – 1430: **Scientific Evaluations**

- SPC/Severe Desk:
    - Examine convection-allowing ensemble guidance: Day 2 vs Day 1
    - Compare convection-allowing guidance (SSEO, SSEF, AFWA, and NSSL; 00Z and 12Z)
- NSSL/Development Desk:
    - UKMET convection-allowing runs
    - Model guidance for hailPBL & Microphysics Comparison

1430 – 1600: **Short-term Outlook Update**

- SPC/Severe Desk:
    - Update probability forecasts for tornado, wind, and hail valid 21-00 and 00-03Z; make these available to EWP
- NSSL/Development Desk:
    - Update hourly probability forecasts for total severe valid 21-03Z; make these available to EWP

Table 1. List of weekly participants (with affiliation) during SFE2014.

5-9 May	12-16 May	19-23 May	27 May – 30 May	2-6 June
J. Blaes (NWS)	D. Dawson (CAPS)	R. Adams-Selin (AFWA)	J. Barnwell (NWS)	T. Alcott (NWS)
M. Chenard (NWS)	G. Dial (SPC)	C. Alexander (GSD)	J. Carley (EMC)	E. Aligo (EMC)
G. Creighton (AFWA)	D. Gagne III (OU)	L. Bosart (SUNYA)	B. Entwistle (AWC)	B. Ancell (TTU)
J. Guyer (SPC)	B. Gallus (ISU)	A. Cohen (SPC)	B. Etherton (GSD)	B. Burghardt (TTU)
S. Hitchcock (OU)	N. Grahame (UKMO)	M. Evans (NWS)	L. Gilchrist (UKMO)	J. Brown (GSD)
A. Kahraman (PSU)	A. Hill (TTU)	L. Gilchrist (UKMO)	J. Grams (SPC)	C. Franks (NWS)
G. Romine (NCAR)	T. Hultquist (NWS)	C. Guastini (SUNYA)	D. Harris (UKMO)	M. Hirsch (NWS)
C. Schultz (NWS)	J. Lawson (ISU)	D. Harris (UKMO)	S. Lack (NWS)	B. Rubin-Oster (WPC)
M. Seltzer (UKMO)	P. Manousos (FE)	T. Hultquist (NWS)	A. Lese (NWS)	B. Smith (SPC)
Y.-G. Skabar (SMN)	J. Milbrandt (EC)	H. Lean (UKMO)	H. Richter (CAWAR)	C. Tubbs (UKMO)
E. Szoke (GSD)	H. Morrison (NCAR)	H. Richter (CAWAR)	S. Roberts (NWS)	B. Twiest (PSU)
	C. Nowotarski (TAMU)	M. Vaughn (SUNYA)	M. Wandishin (GSD)	C. Weiss (TTU)
	N. Roberts (UKMO)	C. Ziegler (NSSL)		
	S. Rogowski (NWS)			
	G. Thompson (NCAR)			
	M. Weeks (UKMO)			

Table 2. Daily responses collected from the microphysics evaluations conducted during SFE2014. The date refers to the model initialization time.

28 April: M-Y2 reflectivities are too high and too large in size. MYJ transitions supercells too fast into a line. Thompson reflectivity looks the most realistic. Thompson also good with cellular nature of storms , though some issues with location of linear storms that evolved later on. I km similar to CR Thompson has less colder tops than other schemes. Morrison cold pool looks more organized and stronger than what happened and other models. Morrsion instability decreases moreso than other models and reality behind cold pool/outflow boundary.
7 May: all models were about 3 hours too fast with convective development and ewd motion. Fewer hot cores with Morrison. m-y2 is the hottest with reflectivity cores. MM has largest hail sizes, followed by Thompson and the rest are the same. sbcape - all models overdo cape.
8 May: Simulated satellite - None of the member depicted the early convection that moved from north-central Texas into southern, central, and eastern Oklahoma which stabilized the region and precluded later convective development. MY-2 is an obvious improvement over MY. All schemes handle low-level clouds across the Dakotas well. Composite reflectivity - All schemes had trouble with the initiation of convection across IA. Convective mode issues across southeastern MN, models tended to develop more cellular convection. Over forecast of convection across northeastern MO late in period. CAPE- All members significantly overdue CAPE values across the region. MY and MY-2 schemes have higher CAPE maxima when compared to the other members. Hourly Max 10-m winds - Consistent signal across northern MO into southeastern IA, but does not match the observed.
11 May: thompson scheme correctly predicted very little convection in AR, while M-Y, MILB-Morr and M-Y2 had too much convection in Arkansas. Larger than typical differences. Big cloud shields will old M-Y schemes seem to be alleviated in M-Y2 and Milbrandt/Morrison. Hailcast fields are very high in Milbrandt-Morrison....probably just not using the right embryo sizes for the particular scheme. G. Thompson: really should use the PSD information for the initial embryos. Great discussion of how to extract hail size from the microphysics scheme. A lot of it was discussed in the long email change, but they suggested using a set concentration number for the graupel hail category and getting the max size from that (You can't just pick a max size from the gamma or exponential distriubtions). We also noticed the Milbrandt/Morrison scheme was heating up too much and had too much SBCAPE...Jason thinks it might be a cloud cover issue...too few clouds and too much heating...not sure what level the cloud issue is...looks like too few low-to-mid level clouds. But the surface dewpoints were also too high. All the cold pools from the storms actually look pretty similar and comparable to what we can tell from the obs. Noticed none of the models cool off the cloud free areas fast enough over the northern part of the domain. Another case of the models being too bullish with redevelopment behind the morning MCS (this day in Arkansas, the day before in Oklahoma).
12 May: We looked at TODAY's forecasts through 18Z since there weren't any available yesterday. m17 and m18 not available today. Jason M. said Fanyou told him to take the M&M (P3) simulated IR results with a grain of salt as the assumptions don't necessarily gel with the assumptions in the microphysics scheme. The P3 scheme MAY have fewer high (ice) clouds than the others and may dissipate those clouds around convection as fast as the others, but we'll need to make sure that isn't just an artifact of the simulated IR algorithm. For composite reflectivity the P3 seemed to handle the CI along the front better. The others were too disorganized and never really developed a line or line segments like what happened. We also noticed P3 had higher CAPE again into the thumb of Michigan, which did

seem to be close to reality. Then looked at 10.7 micron (longwave infrared) simulated channel and it was pretty clear that the Thompson scheme had more low-level clouds in lower Michigan, whereas P3 had very little. Greg T. said there's an accumulated radiation variable you can turn on in WRF that would be helpful for looking at effects of clouds in different schemes. We saw P3 has warmer/moister cold pools than Thompson and Morrison. Implies not as much evaporative cooling in P3.
13 May: No M-Y or M-Y2 available. Morrison seems to have more light-moderate composite reflectivity values than Thompson and P3. All of the models had trouble with the timing of the various clusters/lines of convection across the area. They had way too much over Ohio in the 00Z time frame and thus couldn't capture the late night supercell that happened over central Ohio. P3 seemed to do best with the coverage of the greens, but also seemed to have too little coverage of any mid-high level clouds. There were abundant mid-high level clouds in the obs. P3 was again warmest overall at 2m in pre-convective areas and had the warmest cold pools, but today the SBCAPE seemed to be smallest overall in the P3. No major differences were found in the 2m dewpoints. P3 was maybe a little wetter toward 01-03Z.
14 May: M-Y has too much coverage of the coldest cloud tops, followed by the Morrison scheme. It's hard to determine any biases in cloud top coolness in the other three schemes. If anything, P3 might be under doing the coldness of the strongest storms, but it's a small bias if anything. There are small areas where the P3, M-Y, and Thompson overdo the coldness of the cloud tops like over the terrain-induced storms early in the period over WV and western PA. M-Y2 and Morrison look pretty good with the coldness of the tops of the convective cores. In the area of our failure from yesterday's forecast in the TN/MS/AL area, the models didn't capture morning convection in AL that spread a cloud shield/light stratiform precip into northern AL/TN, that ended up shutting down the severe chances with the squall line along the front further west. The models developed storms in this area that were too strong and didn't develop anvils like in observations. All but P3 had a good weakening trend with the squall line coming in but too much severe with the cells that developed over AL/TN. M-Y2 seems to have convective cores that are too hot overall. Looking at SBCAPE, the models erode the stable stratus layer over VA too aggressively, especially M-Y. Thompson seems to represent the SFCOA analysis the best. P3 seemed to have a high SBCAPE in the strip from MS into Ohio. It was too warm at 2 m again overall and had warm cold pools. There was an interesting area of drying over the higher terrain in far western VA/southern WV and eastern KY, probably associated with mixing down of dry air just above the surface . The HRRR was showing this, but dried too much. All of the runs here did not show nearly enough of this drying (MYJ).
15 May: Similar biases to what we've seen in previous days on the simulated IR: M-Y has too many cold cloud tops, followed by Morrison, which is better, but still too cold. The other three look pretty similar in terms of coverage of greens (collard). P3 may have too cold tops in the convective cores but this could just be a calibration issue with the colors. Dusty: They all suck. Composite reflectivity and 1km reflectivity: The convective cores look too splotchy in most of the runs, especially in P3, M-Y and M-Y2. Morrison and Thompson look the best but still have splotchy areas and too much coverage of higher dBZ. In terms of skill, all the runs were too slow with the corridor of convection and the redevelopment of the squall line and cells out ahead of the line in VA/PA/MD.
19 May: all models created much stronger storms than developed across portions of AR/OK. models did pretty good with strati-form areas of pcpn. P3 reflectivities were a little too great compared to reality and other models. All schemes overdid SBCAPE compared to reality. cloud tops too cold due to too much instability.
20 May: All models over forecast convection in SD along the cold front. Shame on them. We went 10% in that area based on that (and spending more time with the main area of interest over SE WY and the NE panhandle). All the models are way overdoing the number of storms in both the cold front area and in the upslope region in WY and NE. Looking at 1 km reflectivity, we see that there's just too much in the composite reflectivity. There's a huge range in the 1-h max hail fields from HAILCAST, with P3 having huge hail and M-Y2 having smaller hail in the hail streaks. Becky brought up that she didn't have access to P3 when working with Scott, so she's not sure what Scott used for assumptions- they might not be a good match for the scheme. Conrad brought up that we need to be careful in attributing everything to the scheme as the storms may not have the same updraft structures because of accumulation of errors. For the NE panhandle storms, M-Y actually seemed to handle the cold cloud tops the best. Ughhhh. M-Y tended to have the coldest tops in previous days. Someone brought up that this was a case of localized instability, so maybe that's an important difference? Morrison in particular was way too cold and expansive for these storms but looked better as the storms were decaying. Unlike previous days, that wasn't much difference in SBCAPE or surface temps/dewpoints among the runs. Not surprising as this was a clean slate case.
21 May: Lots of comments on how we do the evaluation. Are there too many errors accumulating that prevent an accurate assessment of the schemes themselves? Lance mentioned that a few years ago the differences in microphysical fields a few years ago seemed much bigger than they do now, so many it's getting harder to discern systematic biases. Thompson and M-Y don't develop stratiform areas like those observed. Also, Thompson and P3 seemed to have a very similar solution, whereas M-Y and M-Y2 were quite similar. Morrison seemed to be a mixture of the two. The cores in M-Y and M-Y2 look too big (and are very similar). P3 had a good signal of the main bowing MCS. All the models seemed to develop convection too far west into Indiana/Illinois early and therefore didn't capture the late night initiation and MCS from Indiana into central KY. For the simulated satellite, the mid-range reds don't

seem to have enough coverage in Thompson and P3. Overall, P3 seemed to handle the cloud tops best. M-Y was very cold again, whereas M-Y2 was much better. Really not much difference in SBCAPE among the runs. Thompson and P3 seem to have weaker cold pools again. M-Y, M-Y2 and Morrison have very expansive cold pools. Morrison had very high 2 m wind speeds behind the bowing line in the cold pool. Looks like larger scale rear inflow getting to the ground. We have the 3D fields for this case so I can look at this later.
22 May: Composite reflectivity - All members captured the convective mode correctly. Differences difficult to determine. All members do develop too much convection over North Carolina. M-Y and M-Y2 have higher reflectivities in North Carolina. Presence of convection across Virginia and North Carolina could result in placement of higher severe probabilities across these areas, which would result in a high FAR because no convection actually occurred. All members have convection across eastern North Carolina around 0000 UTC, which was not observed. Simulated satellite - Convective development across North Carolina produces very cold cloud top temperatures. These were not present in the observations. CAPE - Model CAPE is significantly+ overdone across all ensemble members. CAPEs in excess of 2500 J/kg were present in the models, whereas observed CAPE is only around 1000 J/kg.
26 May: Composite reflectivity - Models overpredict convection across central and eastern Tennessee.at 2100 UTC. Morrison has weaker cores, while the M-Y schemes are much stronger. All members miss the dying MCS across western North Carolina. This results in all members producing too much convection across the Carolinas later in the period. Terrain may be impacting convection initiation across the Carolinas, similar to what has been seen across the Front Range of the Rockies. Simulated satellite - Significant improvement in the depiction of cloud-top brightness between the M-Y and M-Y2 schemes. All members move the convection off the coast too quickly. M-Y2 scheme looks to perform the best late in the period. CAPE - Models do not have higher CAPE across eastern North Carolina. However, the models do sustain intense convection across the area, which seems counterintuitive.
27 May: Composite reflectivity - All schemes were too slow to evolve into a linear MCS early in the period. Morrison gets a handle on the convection, but it is about 5 or 6 hours too late. No schemes really capture the bowing out of the MCS, and they all lack to capture the areal extent of the stratiform precipitation area. Thompson has the least spatial extent of the high cores, but this is closest to reality. No schemes capture the supercell across southern TX. Simulated satellite - All schemes have a good handle on the spatial extent of the cloud shield at 1500 UTC. Improvement between M-Y and M-Y2 also noted. CAPE - Thompson has the least amount of CAPE and has the correct placement of the high CAPE axis around 1800 UTC, which matches well with observations. M-Y and M-Y2 develop too much CAPE
28 May: All models had too much convection in eastern Montana with too much of a linear organization, but Thompson had the best depiction of the reflectivity structures being fairly weak in this area. Later, P3 had a very good depiction of the main convection near Great Falls and northeast of there, with a small bowing system expanding with time, along with two seperate cells behind it. Morrison was pretty good with this too. M-Y2 seemed to have cores that were way too strong. In the simulated IR, Thompson again appeared to have the best depiction of the cold cloud tops, but all of them seemed to be pretty good (except for M-Y which was too hot again). All the models were very similar with SBCAPE forecasts and with surface T/Tds.
29 May: All forecasts were very similar today. There were just minor differences in the details of the simulated reflectivity hotness, with M-Y2 having the biggest, hottest cores. Thompson again appeared to have the smallest bias errors in reflectivity and simulated brightness temps in the IR. Morrison had the biggest, fattest cold pools. SBCAPE was very similar.
31 May: The Thompson scheme had too few high reflectivity areas in NE CO. We looked at SBCAPE to see if that could explain it. Not really. Dave thought the P3 scheme looked very good in this case with its depiction of the line in the Dakotas. Looking at simulated IR M-Y again was too hot. Thompson and Morrison actually looked reasonable in the IR, but Thompson had too much cold cloud tops even though the cells didn't look very strong in the composite reflectivity. P3 looked to have a small overall bias but seemed to have too many individual hot cells.
3 June: Composite reflectivity - M-Y2 has very high cores in the cells across northern NE. Morrison seems to have the best depiction of convection there. Thompson has multiple supercells across NE, which did not actually occur. Thompson looks to have a weaker cold pool, which mitigated upscale growth into an MCS.
5 June: We looked at the evolution of the KS MCS...P3 and Morrison had a really nice depiction of the structure of the MCS. Thompson and the two M-Ys were too disorganized. BUT, Thompson captured the strong winds with the wake low that went through NE KS/NW MO very well. A though- maybe sublimation of snow, which is favored a lot in Thompson, was important for the wake low in this case. P3 and Morrison did as well but were spottier (which may have actually been better...we don't really know). P3 and Morrison in particular continued to have a really nice evolution to the MCS into AR/TN with Thompson catching on to the structure later. Back west there was far too much development into OK/MO along the old OFB. They destabilized too much behind the MCS, but maybe this was a day where the smoothing of the inversion really hurt the WRF-ARW forecasts? We looked at forecast soundings in NE OK and it looked like the WRF-ARW soundings did away with the MLCIN whereas the UKMET held on to it with a sharper cap.

Table 3. Daily ratings and responses for the comparison of the Met Office CAMs with the NSSL-WRF. The date refers to the model initialization time.

7 May – Same: UK system better early and NSSL was better later
9 May - UKMET worse than NSSL-WRF: This is the forecast for 20140511 - the day of the big HP storm in SE Nebraska. The forecasts are generally the same before 12Z, with the NSSL WRF pulling ahead with a great depiction of an MCS across SE NE into SC Iowa- the UKMET had the system but was more disorganized. It also killed the system too early. Both models were too late with CI along the dryline and warm front/OFB in Nebraska. The NSSL WRF continued to be marginally better throughout the evolution of the line of supercells and the upscale growth of the system into Iowa. Both models missed the redevelopment of storms further south along the dryline/front at night...***BUT*** the 4 km UKMET eventually got the redvelopment and the 2.2 km model had the CI of it handled really well. A great example of the difference that higher resolution can make. The 2.2 km (along with the other 4 km runs) has problems with the elevated MCS over NW MO.
12 May – Same: 4.4 km UKmet and NSSL WRF both spin up a convective system over IA quite well over the first 6 h of the forecast - UKmet is more realistic especially with widespread convective/stratiform precipitation in MN. Also, a small MCS develops in OH/IN during the first 6 h of the forecast - the UKmet nails this system. 2.2 km UKmet much worse during first 6 h. IA system moves too fast to the east and convection is too widespread over Wisconsin into OH (not an obvious small MCS in IN/OH in 2.2 km UKmet). At 24 h, NSSL-WRF replicates a long MCS stretching from E IA to E OK very well. Placement is almost perfect. UKmet 2.2 km has too many bowing segments and it too far east. 4.4 km UKmet has the linear character of the convection depicted well but it is too far west (slow) with the system. Summary: UKmet better during first 15 h, but the next diurnal cycle NSSL-WRF is better with timing placement of convection. 2.2 km UKmet was worse than the 4.4 km.
13 May – Same: UKMET is able to spin up a weak squall line over IL/MO in the first 6 hours of the forecast, whereas NSSL WRF does not. UKMET overdoes the spun up squall line over Lake Michigan. Both of them handle the elevated MCS over western New York fairly well. The 2.2km run looks even better than the 4.4km run. It has the structure of the squall line a lot better, although it also overdoes the northern part of the squall line over Lake Michigan. Both 4 km models are too slow with the main development over south central Michigan into eastern Indiana. The 2.2 km wasn't really any better. Later (21Z) the NSSL WRF hints that there'd be a lot of pop-up storms over the Appalachians and Ohio valley, but it over did it. The UKMET was even slower than NSSL WRF with the main line in Michigan.
14 May - UKMET worse than NSSL-WRF: 1 km reflectivity continues to be too hot in the convective cores in the UKMET 4 and 2 km runs. There were a few times when the NSSL WRF and UKMET 2 km were remarkably good (~15-17Z). I thought the NSSL WRF reflectivity was on both sides of the display at first but the obs were indeed up there. There were several areas of developing convection at this time. But later things changed quickly as they had a hard time with the timing of the activity in Ohio and with way too much intensity to the convection in TN/AL. They struggled with the stratiform could shield in this area earlier in the day. This was a tough comparison. The vote was 4-2 in favor of NSSL WRF better than the 4 km UKMET, with the 2 thinking they were the same. If we were comparing the 2.2 km UKMET to NSSL WRF, it would have been a wash.
15 May – Same: UKMet models better at the beginning, especially at 1200 UTC, but NSSL-WRF does better through the rest of the forecast period. NSSL-WRF emphasizes development on the east side of the cloud shield. Both models are slow with the evolution of the line and had too much development out ahead of the line.
18 May (not sure of correct forecast date on this one) - UKMET better than NSSL-WRF: NSSL WRF has higher dewpoints than UKMET. NSSL had a couple discrete supercells where UKMET did not.
19 May - UKMET better than NSSL-WRF: Both models evolved the initial storms at 21-22Z too much, whereas in reality there was a 3-4 hour period of attempted CI/SI in SE WY before two storms became dominant. The UKMET was slightly better in the 6-12 h time frame as it didn't produce the few weak spurious storms that were in the NSSL WRF. They were a wash before the main CI. CI timing in SE WY was about the same, but UKMET developed one dominant supercell and took it southeast (2-3 hours too early) whereas NSSL WRF didn't develop any long-lived rotating storms. Some people thought we need to penalize the NSSL WRF compared to the UKMET whereas others thought it was a wash (including Dave I. and our NWS forecaster Mike Evans). The NWS forecasters were considering timing, location, coverage, speed, etc. and said it really wasn't any better. "If there's a discrete storm in that environment it's gonna rotate regardless of what the explicit NSSL WRF fields show". The 2 km was even faster than the 4 km.
20 May – Same: UKMET 4- and 2-km models do a good job with timing of initiation, but lacks the correct intensity. NSSL-WRF does a better job with the intensity of the initiating convection. UKMET 2-km model does a much better job than the UKMET 4-km model with later convection after 0000 UTC. The UKMET 4-km model totally misses the intense convection across the Chicago metro area. UKMET 4- and 2-km models has better depiction of convection across southeastern MI, northwest OH when compared to the NSSL-WRF near the end of the forecast period.
21 May - UKMET better than NSSL-WRF: The UKMET again spun up existing convection better with the system

over Northern Illinois and was deemed better from 0-12Z. It was a wash between 12-00Z between the two, but then UKMET did marginally better with the two areas of convection in Ohio and Indiana/Kentucky, but it didn't organize the systems into a bow echo. The NSSL WRF was very misleading with a sagging bowing system to the south and southwest.
23 May – Same: UKMet has a better depiction of convection across the southeastern portion of North Carolina, although it does produce too much convection across the NC/SC border. UKMet and NSSL-WRF both bad in different eays at 0000 UTC. NSSL-WRF recovers late in the period, and correctly places the convection just off the South Carolina coast.
28 May (not sure of correct date) - UKMET better than NSSL-WRF: 6 h into the forecast, UKmet depicts an ongoing MCS much more skillfully than the NSSL-WRF. Later, UKmet has shape of MCS better simulated but has spurious convection elsewhere in the domain. Overall, UKmet better.
28 May (Not sure of correct date) - UKMET better than NSSL-WRF: UKmet depicted MCS in northern Montana better than NSSL. Also, better info on general convective evolution in UKmet.
29 May - UKMET worse than NSSL-WRF: UKMET usually spins up better but not this time. NSSLWRF had a better handle on spinnnig up the strong system near Great Falls and evolving it into Canada. UKMET also had a spurious storm in the soin up period behind the main one. Nothing much happened from 06-18Z, then the NSSLWRF did a pretty good job with the areas and timing of CI on the terrain and along the front. UKMET struggled with CI along the front. We spent time trying to figure out why. It wasn't obvious because UKMET seemed to handle the 2m Temps and Dews better along the front, as well as in the warm sector. We looked at several UKMET soundings and they were impressively good, although the Bismark sounding was a bit too dry. Maybe this was enough for it to not convect
1 June - UKMET better than NSSL-WRF: UKMET did a slightly better job spinning up the storms along the front in the Dakotas in the 00-12Z period. NSSL WRF does better with the early diurnal period of convection slightly. Both models are a bit late with CI along the front but the UKMET had a better orientation of the convection along the front and with the widespread convection over NE CO. They both didn't get the evolution well in the evening and overnight, with both models too far east with the frontal convection early, then were too far west with the convection toward 12Z.
2 June - UKMET better than NSSL-WRF: UKMET spun up the squall line in Kansas VERY fast, but after that it was considered a wash through 06Z. The following two periods showed a clear advantage to the UKMET model. The 18-00Z period was dominated by the handling of the two weak storms over northern OK and the weak storms over eastern AR. The NSSL WRF had the weak storms in northern OK but also had spurious storms in Kansas and misplaced the storms in AR. The UKMET displaced the OK storms a bit in OK but the 2.2 km was even better than the 4.4km UKMET model. For the 00-06Z period the NSSL WRF is better because the UKMET initiated storm in western NE too early but the UKMET was better in the last period as it handled the elevated storms in western NE much better toward morning. The sounding comparisons were interesting again. UKMET beat the pants off NSSL WRF in the nearby soundings (DRT, DFW, DDC). OUN and LMN were contaminated in NSSL WRF.
3 June - UKMET better than NSSL-WRF: NOTE: THE NSSL WRF Lowest level dewpoint was messed up in the initialization....but it didn't mess much up since storms in the early period were elevated. Both NSSL WRF and UKMET were pretty good in the 00-12Z period as they spun up elevated storms over SD and NE. In 12-18Z period NSSL WRF was late developing the elevated severe storm over northern NE. UKMET was better at this although it was too far north but had a good depiction of the intensity of the storm at 18Z. The 2.2 km UKMET TOP 00Z raob: UKMET better PBL depth, better BL moisture, slightly too warm (NSSL was better here), but more correct MLCIN.
4 June - UKMET better than NSSL-WRF: Both models did a good job spinning up the bowing line segments and general MCS in the first 6 hours. Both models have too much precip in the warm-advection wing. UKMET has too much convection in the comma head of the MCS, but NSSL WRF weakened the line too quickly. The structure of the system in the UKMET was deemed better. NSSL WRF developed a spurious squall line behind the decaying MCS by 18Z that was not there in the UKMET. The UKMET has a very good depiction of the decaying system (but with reflectivity too hot). NSSL WRF continues to overdo the squall line through 00Z. The UKMET was clearly better showing the somewhat disorganized nature of the line at 00Z. The NSSL WRF organized the line too much and ended up behind faster. UKMET got some "wows" with this forecast. Back west the NSSL WRF had an MCS coming through northern OK much earlier than the MCS in KS. The 4.4 UKMET was far too late with the MCS coming out of CO/KS. The 2.2 km was amazingly good with the CO/KS MCS.
5 June - UKMET worse than NSSL-WRF: The UKMET again spun things up better, especially with the line in Tennessee. The NSSL WRF had better placement with the monster storm in E CO- the UKMET was a bit too fast. But it was a wash overall in the first 6 hours. Both models were too far north with the MCS in KS. The NSSL WRF was slightly better. However, the 2.2km UKMET improved quite a bit over the 4.4km in getting the bow structure on the southern end, giving the NSSL WRF +0.5. This trend continued in the 12-18Z period- the NSSL WRF had better structure to the MCS than the 4.4 km UKMET, which showed more of a parallel stratiform MCS, so it got another +0.5. However again the 2.2km UKMET was much better than the 4.4km with the structure of the MCS. The slight edge was given to UKMET in the 18-00Z period but there wasn't much going on. The UKMET did capture the four general aeras of CI in MN/NW IA and CO better than the NSSL WRF. The 2.2km was a bit worse than the 4.4km in

this period, except for a bit better structure to the weakening MCS over AL/GA. After 00Z both models had very wrong evolution to the convection through 12Z. NSSL WRF came out on top by +0.5. Steve W. showed that the 2.2km parallel run captured the OK MCS extremely well in 30-36 h forecasts. Should be a very interesting case to unravel to see how they evolved differently. Brain Ancell said the 12Z TTU WRF initialized with the GFS did very well too. Adam said he can go back to see if the NSSL WRF member initialized with the GFS got it as well. Also noted that the 12Z NSSL WRF was really bad with a spurious MCS in Missouri. Soundings were mixed this time, with UKMET overmixing at OUN and DFW with a dry bias. It also had a dry bias at AMA but under mixed a bit.