46 AN OVERVIEW OF THE 2014 NOAA HAZARDOUS WEATHER TESTBED SPRING FORECASTING EXPERIMENT

Israel L. Jirak[1*], Michael Coniglio[2], Adam J. Clark[2,3], James Correia Jr.[1,3], Kent H. Knopfmeier[2,3], Christopher J. Melick[1,3], Steven J. Weiss[1], John S. Kain[2], M. Xue[4], F. Kong[4], K. W. Thomas[4], K. Brewster[4], Y. Wang[4], Y. Jung[4], and S. Willington[5]

[1]NOAA/NWS/NCEP/Storm Prediction Center, Norman, OK
[2]NOAA/OAR/National Severe Storms Laboratory, Norman, OK
[3]Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, OK
[4]Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, OK
[5]Met Office, Exeter, UK

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 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.

* *Corresponding author address*: Israel L. Jirak, NOAA/NWS/NCEP/Storm Prediction Center, 120 David L. Boren Blvd., Norman, OK 73072; e-mail: Israel.Jirak@noaa.gov

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

2.1 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.

2.1.1 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.

2.1.2 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.

2.1.3 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.

2.1.4 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.

2.1.5 Met Office Convection-Allowing 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.

2.2 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.

2.2.1 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).

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.

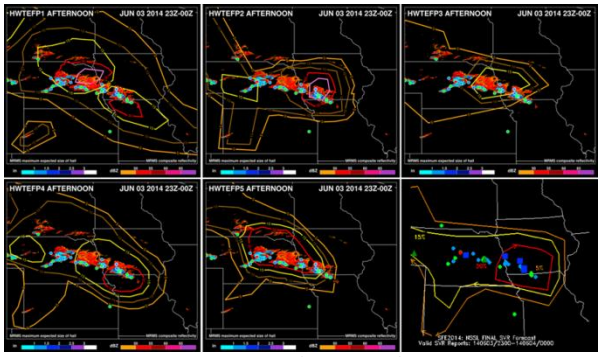

Figure 1. One-hour probability forecasts of total severe valid 23Z 3 June to 00Z 4 June. The forecasts made by the lead forecaster are shown in the lower right panel with forecasts from the participants shown in the other panels. Local storm reports are plotted along with observed composite reflectivity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

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.

2.2.2 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

3.1 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.

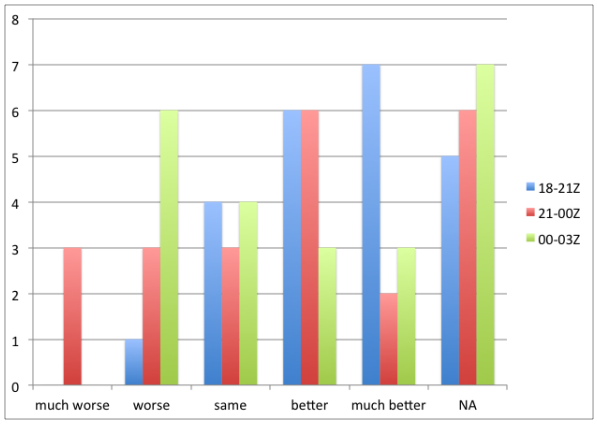

Figure 2. Number of subjective ratings of the preliminary forecast compared to the first-guess forecast.

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.

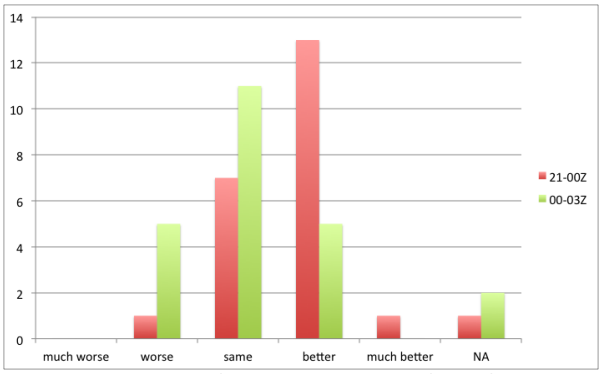

Figure 3. Number of subjective ratings of the final human forecast compared to the preliminary human forecast.

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.

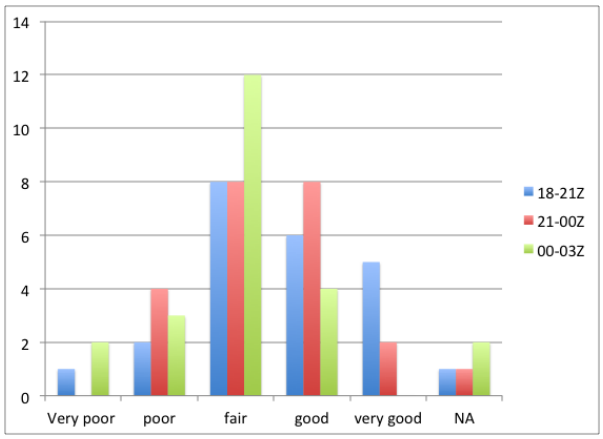

Figure 4. Number of subjective ratings of the final human forecast compared to local storm reports.

Efforts to objectively evaluate the 1-hour forecasts are ongoing. Although the main goal of having the participants generate their forecasts was to immerse them into the activities and forecast process more than in previous SFEs, their forecasts will also be evaluated with objective metrics so that some measure of variability to the metrics can be found. Preliminary verification efforts indicate that the hourly forecasts appear to be most reliable when verified with LSRs on a 40-km grid (not shown). Further effort is needed to identify best methods and approaches for verifying hourly forecasts of severe weather.

3.2 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. 5). 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. 6). For wind, the preliminary forecasts had a more uniform distribution of subjective ratings from "better" to "worse" (Fig. 7). 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.

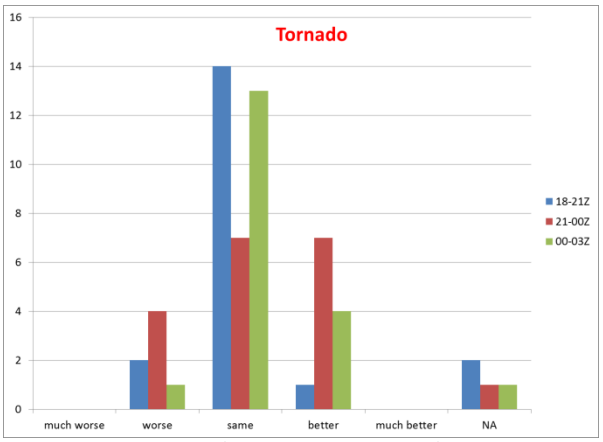

Figure 5. Number of subjective ratings of the preliminary tornado forecast compared to the first-guess tornado forecast.

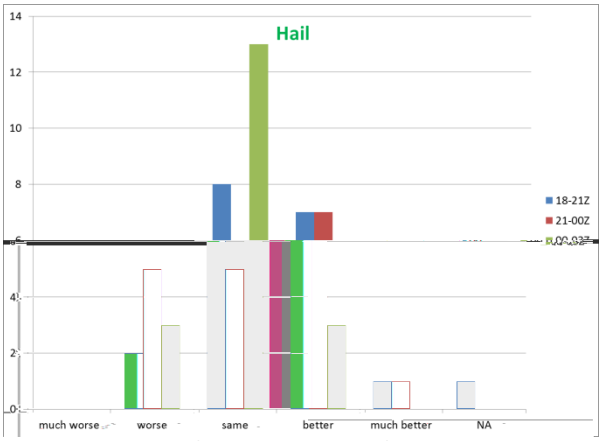

Figure 6. Number of subjective ratings of the preliminary hail forecast compared to the first-guess hail forecast.

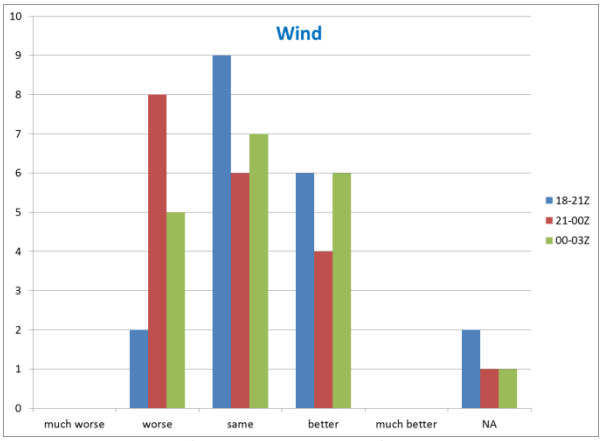

Figure 7. Number of subjective ratings of the preliminary wind forecast compared to the first-guess wind forecast.

The preliminary and final tornado, wind, and hail forecasts were subjectively compared to determine the relative value of the afternoon forecast updates (Figs. 8-10). 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.

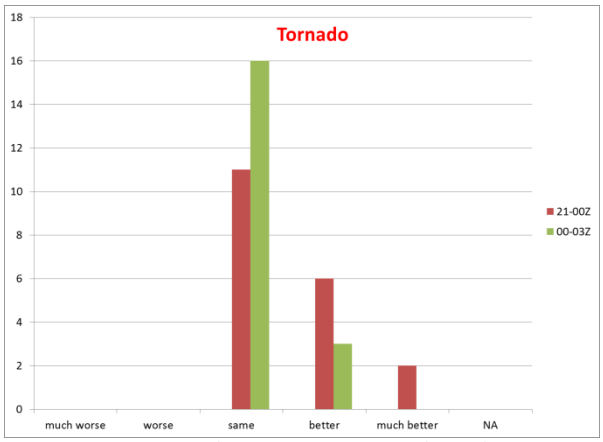

Figure 8. Number of subjective ratings of the final tornado forecast compared to the preliminary tornado forecast.

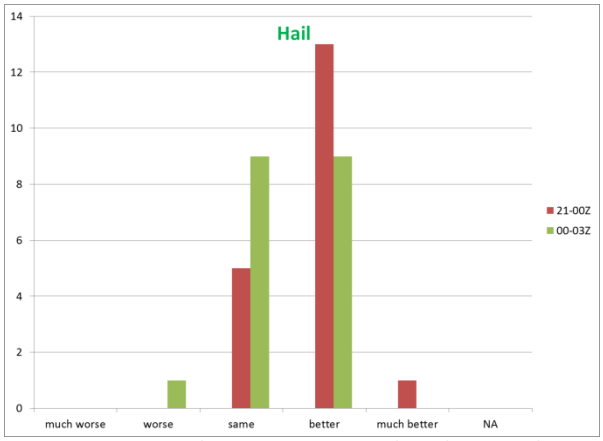

Figure 9. Number of subjective ratings of the final hail forecast compared to the preliminary hail forecast.

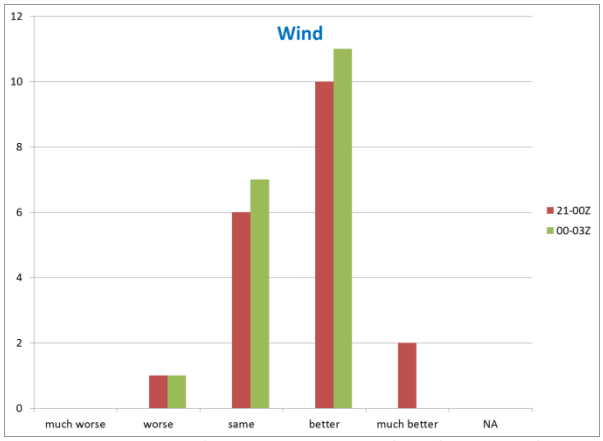

Figure 10. Number of subjective ratings of the final wind forecast compared to the preliminary wind forecast.

3.3 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. 11). 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".

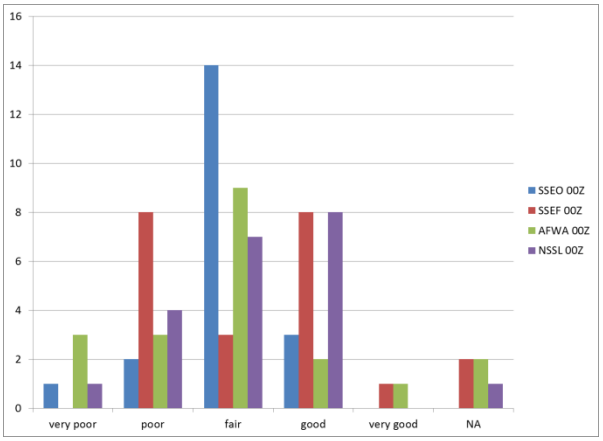

Figure 11. Number of subjective ratings for the ensemble neighborhood reflectivity forecasts compared to observed radar reflectivity.

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. 12). 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.

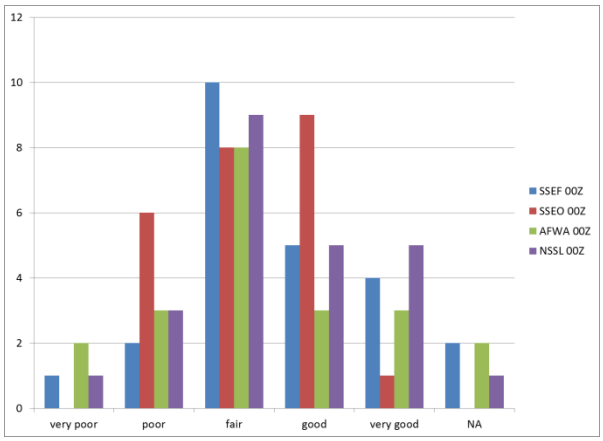

Figure 12. Number of subjective ratings for the ensemble HMF forecasts compared to local storm reports.

3.4 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. 13). Figure 14 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. 13). 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.

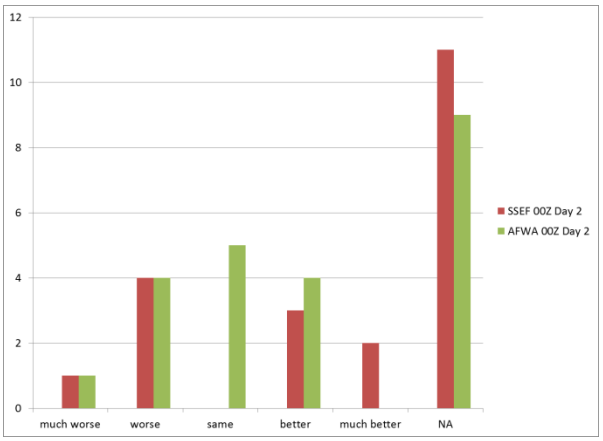

Figure 13. Number of subjective ratings for the Day 2 ensemble forecasts from the SSEF (red) and AFWA (green) compared to the Day 1 forecasts.

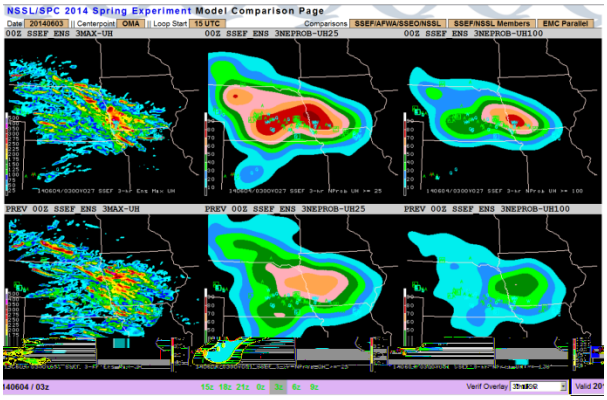

Figure 14. SSEF Day 1 (top row) and Day 2 (bottom row) forecasts of 3-h ensemble maximum UH (left column), ensemble neighborhood probability of UH ≥25 m2s-2 (middle column), and ensemble neighborhood probability of UH ≥100 m2s-2 (right column) valid 0000-0300 UTC on 4 June 2014. The severe reports during this 3-h period are plotted as letters in each panel.

3.5 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. 15) for convective-storm guidance. Figure 16 illustrates an example where the parallel HiResW ARW (upper middle) was rated better than the operational HiResW ARW (upper left).

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 15). Both of the parallel HiResW versions (i.e. ARW and NMMB) were implemented operationally after SFE2014 on 11 June 2014.

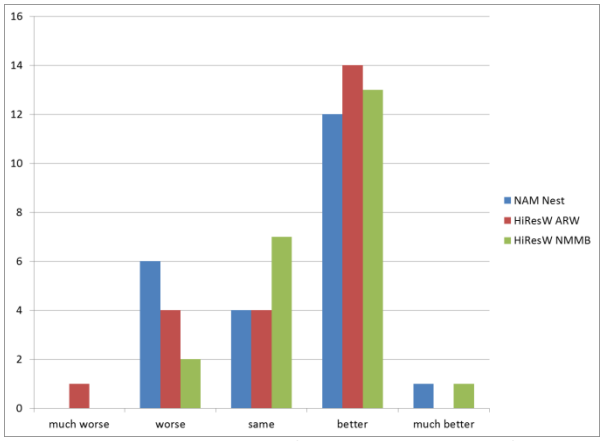

Figure 15. Subjective ratings of the parallel versions of the EMC CAMs (NAM Nest – blue, HiResW ARW – red, and HiResW NMMB – green) compared to the operational versions.

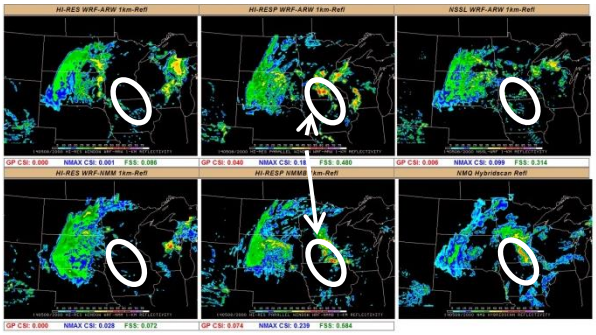

Figure 16. Simulated reflectivity forecasts valid at 2000 UTC on 8 May 2014 for the operational HiResW ARW (upper left), parallel HiResW ARW (upper middle), NSSL-WRF (upper right), HiResW WRF-NMM (lower left), HiResW NMMB (lower middle) and observed reflectivity (lower right) at that time.

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. 15). The NAM was upgraded

operationally with this parallel version on 12 August 2014.

3.6 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. 17. 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. 17). 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.

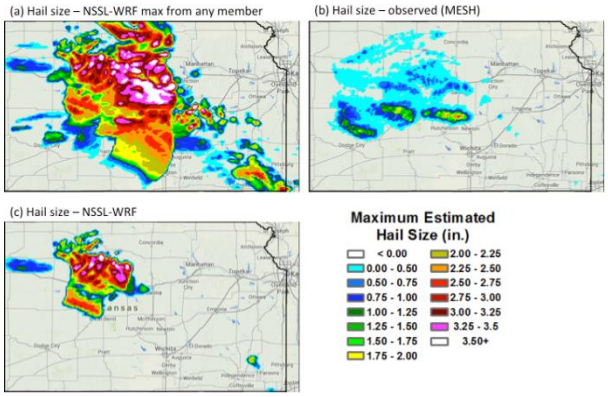

Figure 17. Maximum hail size over the previous hour valid 1000 UTC on 5 June 2014. (a) NSSL-WRF ensemble maximum from any member HAILCAST forecast (0000 UTC 5 June initialization), (b) observed maximum hail size from MESH, and (c) NSSL-WRF control member HAILCAST forecast.

3.7 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 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 18-20. 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 18 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 19, Thompson was the only scheme that was able to depict the high winds associated with the wake low near Kansas City. Finally, Figure 20 clearly shows the improvement in MY2 relative to MY, with the extent of colder cloud tops reduced and overall temperatures warmed.

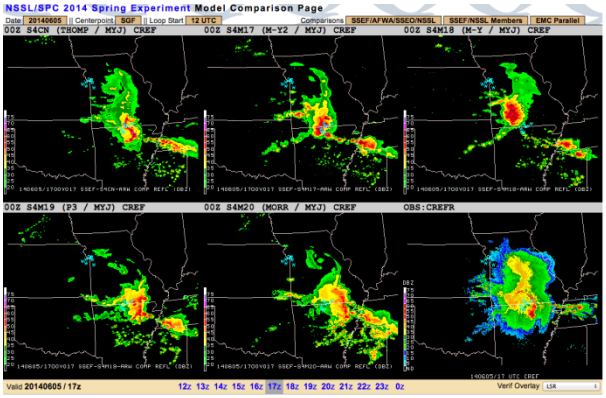

Figure 18. Forecasts and observations of composite reflectivity valid 1700 UTC 5 June 2014. The forecasts were initialized 0000 UTC 5 June and are from the members of the SSEF system configured identically except for their microphysics schemes. The panels include Thompson (upper-left), MY2 (upper-middle), 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.

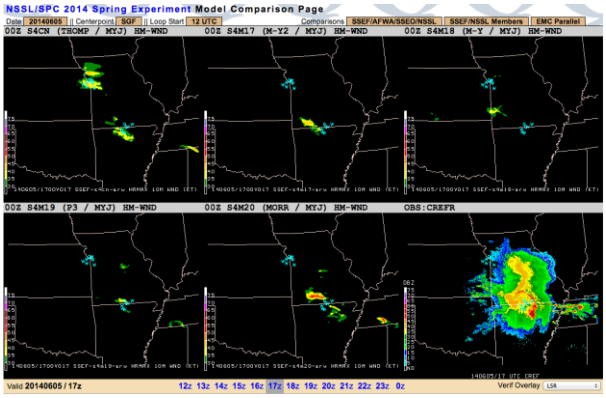

Figure 19. Same as Fig. 18, except forecasts of hourly maximum 10-m wind speed (lower-right panel still includes observed composite reflectivity).

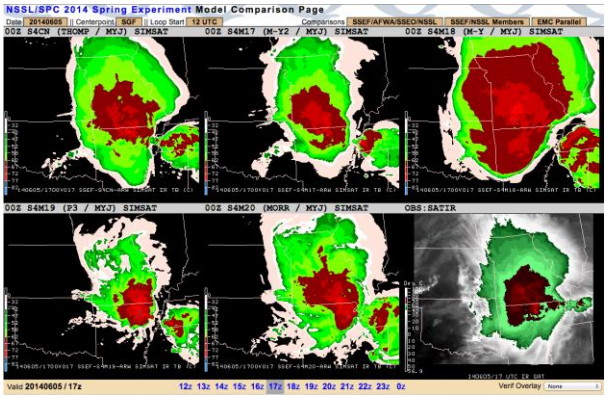

Figure 20. Same as Figure 18, except for simulated IR brightness temperatures.

3.8 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."

Participants could select from, "UKMET better than NSSL-WRF", "UKMET worse than NSSL-WRF", or "Same". The responses (20 cases) are summarized in Fig. 21. 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.

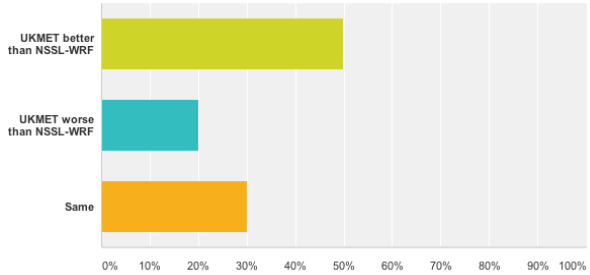

Figure 21. Summary of responses for the NSSL-WRF and Met Office CAM comparisons.

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 22 and 23).

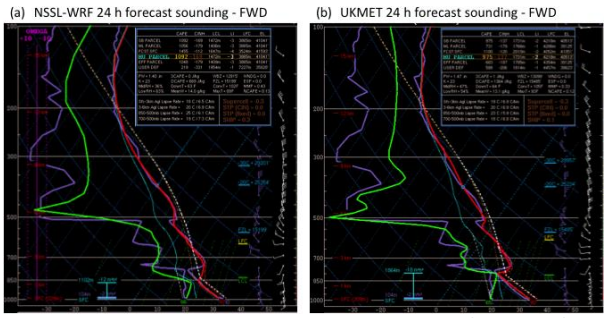

Figure 22. Forecast soundings valid 3 June 2014 for FWD from 24 h forecasts of the (a) NSSL-WRF, and (b) the UKMET. In both panels, the corresponding observed sounding is overlaid in purple.

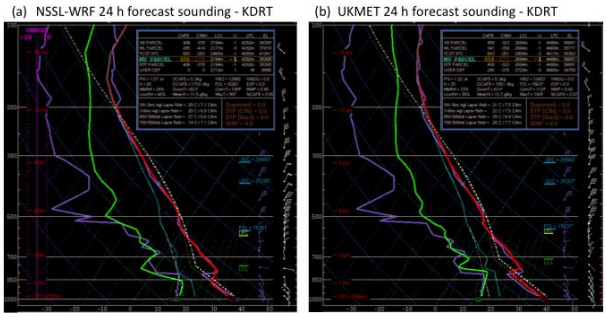

Figure 23. Same as Fig. 22, except for KDRT.

3.9 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. 24a and 24b). 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. 24c). 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. 24d).

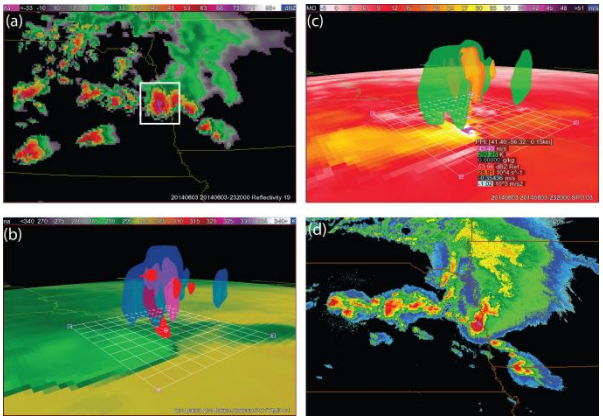

Figure 24. Example of viewing a model storm in 3D using WDSS-II display software. (a) Plan view of simulated reflectivity on model level 19 valid 2320 UTC for the 00Z SSEF control member. The white box encloses the storm interrogated in 3D in panels (b) and (c). In (b), isosurfaces of vertical velocity > 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. Evan Kuchera and Scott Rentschler of AFWA generously provided AFWA data during SFE2014.

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:
  - 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
  - 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:
  - 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
  - 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**

- Prepare probability forecasts for tornado, wind, and hail valid 12-12Z on Day 2 over mesoscale area of interest
- 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**

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

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

- Update probability forecasts for tornado, wind, and hail valid 21-00 and 00-03Z; make these available to EWP
- Update hourly probability forecasts for total severe valid 21-03Z; make these available to EWP