

J11.1 AN OVERVIEW OF THE 2013 NOAA HAZARDOUS WEATHER TESTBED SPRING FORECASTING EXPERIMENT

Israel L. Jirak^{1*}, Michael Coniglio², Adam J. Clark^{2,3}, James Correia Jr.^{1,3}, Kent H. Knopfmeier^{2,3}, Christopher J. Melick^{1,3}, Steven J. Weiss¹, John S. Kain², M. Xue⁴, F. Kong⁴, K. W. Thomas⁴, K. Brewster⁴, Y. Wang⁴, S. Willington⁵, and D. Suri⁵

¹NOAA/NWS/NCEP/Storm Prediction Center, Norman, OK

²NOAA/OAR/National Severe Storms Laboratory, Norman, OK

³Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, OK

⁴Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, OK

⁵Met Office, Exeter, UK

1. INTRODUCTION

The 2013 Spring Forecasting Experiment (SFE2013) was conducted from 6 May – 7 June by the Experimental Forecast Program (EFP) of the NOAA/Hazardous Weather Testbed (HWT). SFE2013 was organized by the Storm Prediction Center (SPC) and National Severe Storms Laboratory (NSSL) with participation from more than 30 forecasters, researchers, and developers to test emerging concepts and technologies designed to improve the prediction of hazardous convective weather. SFE2013 aimed to address several primary goals:

- Assess the value of convective outlooks that are updated more frequently and with higher temporal resolution than those produced operationally at SPC.
- Compare 1200 UTC-initialized convection-allowing ensembles to their 0000 UTC-initialized counterparts.
- Evaluate the NSSL Mesoscale Ensemble (NME) in diagnosing and predicting the pre-convective environment.
- Determine whether a parallel NSSL WRF-ARW initialized from the NME produces improved forecasts over the NAM-initialized version.
- Compare the performance of two Met Office Unified Model convection-allowing configurations with the NSSL WRF-ARW runs.
- Examine physics sensitivities in the convection-allowing WRF-ARW simulations.

This document summarizes the activities, core interests, and preliminary findings of SFE2013. 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_2013/HWT_SFE_2013_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 SFE2013 along with a description of the daily activities, and Section 3 reviews the preliminary findings of SFE2013. Finally, a summary can be found in Section 4.

2. DESCRIPTION

2.1 *Experimental Models and Ensembles*

Building upon successful experiments of previous years, SFE2013 focused on the generation of 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 of providing nearly continuous probabilistic hazard forecasts on increasingly fine spatial and temporal scales. As in previous experiments, a suite of new and improved experimental mesoscale and 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 Mesoscale Ensemble (NME)*

A Weather Research and Forecasting (WRF)-Advanced Research WRF core (ARW) (v3.4.1) mesoscale data assimilation system was run daily to produce three-dimensional analyses over a CONUS domain with 18-km horizontal grid spacing (278x189) and 51 vertical levels. The 36-member NME was constructed from the initial and lateral boundary conditions (ICs/LBCs) provided by the 1200 UTC Earth Systems Research Laboratory (ESRL) experimental Rapid Refresh version two (RAPv2) forecast cycle for the first three weeks of SFE2013 and the 1200 UTC 12-km North American Mesoscale (NAM) forecast cycle for the final two weeks of the SFE. The change was necessitated by the discovery of an error with the soil moisture/temperature adjustment in the RAPv2 GSI analysis that led to a moist bias during the late afternoon period (i.e., 2100-0200 UTC). Random samples of background error were generated by the WRF variational data assimilation (WRF-Var) algorithm and then added to each ensemble member, to account for

* 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

uncertainties in the ICs/LBCs of the reference analysis (Torn et al. 2006). The WRF-ARW physics options were also varied amongst the ensemble members to examine sensitivity of forecasts to variations in model physics.

Routinely available observations (of altimeter setting, temperature, dewpoint, and horizontal wind components) from land and marine stations, rawinsondes, and aircraft – as well as satellite winds – were assimilated utilizing an ensemble Kalman Filter (EnKF) (using the Data Assimilation Research Testbed (DART) software) at hourly intervals from 1300 UTC to 0300 UTC the following day. At 1400, 1600, and 1800 UTC, the resultant EnKF analyses were used to launch a full ensemble of forecasts out to 0300 UTC and were used in the experimental forecast process.

2.1.2 NSSL-WRF

SPC forecasters have used output from an experimental “cold-start” 4 km WRF-ARW produced by NSSL 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 using NAM ICs/LBCs with forecasts to 36 hours. New to the experimental numerical guidance for this year’s experiment was a parallel or “hot-start” version of the NSSL-WRF that was initialized from the “best member” of the 0000 UTC NME analysis. The best member was defined as the member with the lowest normalized RMS difference of temperature and horizontal wind components using all 0000 UTC observations.

The hot-start run was configured identically to the standard cold-start NSSL-WRF run so that the impact of the NME analyses in initializing the forecasts could be evaluated. Specifically, both runs used WRF version 3.4.1, NAM forecasts at 3 hourly intervals for LBCs, WSM6 microphysics parameterization, and MYJ turbulent-mixing (PBL) parameterization. For comparing the two NSSL-WRF runs, an interactive web display developed by NSSL called the Data Explorer utilizing Google-maps-like features and GIS was used. The web display allows zooming, overlaying of chosen fields, and side-by-side comparisons of model and observational fields.

2.1.3 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 with forecasts to 36 hrs. The 2013 0000 UTC SSEF system included 25 WRF-ARW members with 15 “core” members having IC/LBC perturbations from the NCEP operational Short-Range Ensemble Forecast (SREF) system as well as varied physics. The remaining 10 members were configured identically except for their microphysics parameterizations (six members) and turbulent-mixing (PBL) parameterizations (four members). All runs assimilated WSR-88D reflectivity and velocity data, along with available surface and upper air observations, 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 forecast process.

For the first time this year, a SSEF system initialized at 1200 UTC was 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 8 members were 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 (see below) for a more equitable comparison of the spread and skill characteristics of these sets of forecasts.

2.1.4 SPC Storm Scale Ensemble of Opportunity

The SPC Storm-Scale Ensemble of Opportunity (SSEO) is a 7-member, multi-model/physics convection-allowing ensemble consisting of deterministic CAMs available to SPC. 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 SFE2013. All members were initialized as a “cold start” from the operational NAM – i.e., no radar data assimilation or cloud model was used to produce ICs.

2.1.5 Air Force Weather Agency 4-km Ensemble

The U.S. Air Force Weather Agency (AFWA) runs a real-time 10-member, 4-km WRF-ARW ensemble, and these forecast fields were available for examination during SFE2013. Forecasts were initialized at 0000 UTC and 1200 UTC using 6 or 12 hour forecasts from three global models: an AFWA version of the Met Office Unified Model, 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.

2.1.6 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

SFE2013 with forecasts through 48 (45) hrs. The 4.4 km CONUS run took its initial and lateral boundary conditions from the 0000 UTC 25-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

SFE2013 activities were focused on forecasting severe convective weather with two separate teams generating identical forecast products with access to the same set of forecast guidance. Forecast and model evaluations also were an integral part of daily activities of SFE2013. A summary of forecast products and evaluation activities can be found below while a detailed schedule of daily activities can be found in the appendix.

2.2.1 Experimental Forecast Products

The experimental forecasts in SFE2013 continued to explore the ability to add temporal specificity to longer-term convective outlooks. The forecasts were made over a movable mesoscale area of interest focused on areas of expected strong/severe convection and/or regions with particular convective forecasting challenges. The forecasts provided the probability of any severe storm (large hail, damaging winds, and/or tornadoes) within 25 miles (40 km) of a point ("total severe"), as defined in the SPC operational convective outlooks. These forecasts were a simplified version of the SPC operational Day 1 Convective Outlooks, which specify separate probabilistic forecasts of severe hail, severe wind, and tornadoes. Areas of significant hail and wind (10% or greater probability of hail $\geq 2"$ in diameter or wind gusts ≥ 65 kt) were also predicted. The forecast teams first created a full-period (1600-1200 UTC) total severe outlook (where SPC forecasters have historically shown considerable skill) and then manually stratified that outlook into three periods with higher temporal resolution: 1800-2100, 2100-0000, and 0000-0300 UTC.

During SFE2012, calibrated probabilistic severe guidance from the SSEO was used to temporally disaggregate a 1600-1200 UTC period human forecast. This disaggregation procedure involved formulating a scaling factor by matching the full-period calibrated severe SSEO guidance to the human forecast, then applying this scaling factor (unique at every grid point) to the SSEO calibrated severe guidance for each individual period, and finally performing consistency checks and smoothing to arrive at the temporally disaggregated forecasts. These automated forecasts from SFE2012 fared favorably both in terms of objective metrics (e.g., CSI, FSS) and subjective impressions when compared

to manually drawn forecasts. Given the encouraging results from SFE2012, a similar technique was applied to forecasts during SFE2013. The 1600-1200 UTC human forecasts for each team were temporally disaggregated into the 3-h periods to provide a first guess for the three higher-resolution forecast periods (1800-2100, 2100-0000 and 0000-0300 UTC).

Two of the three afternoon and evening forecast periods (i.e., 2100-0000 and 0000-0300 UTC) were updated two times in the afternoon, which had not been attempted before in the SFE. In addition, the digitized probabilistic forecasts of severe convection over 3-h periods were shared with the Experimental Warning Program (EWP) and were used in preparation for their operations. This was the first such direct interaction between the forecast and warning components of the HWT and is an early manifestation of the goal of providing probabilistic hazard forecasts on multiple scales from the synoptic scale to the storm scale.

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 SFE2013 was to look back and evaluate the forecasts and model guidance from the previous day. In particular, forecasts for the 3-h periods were subjectively and objectively evaluated to assess the ability to add temporal specificity to a probabilistic severe weather forecast valid over a longer time period. The forecasts were also compared to the temporally disaggregated first guess guidance and to subsequent issuances to determine the value of updating the forecasts through the afternoon. Subjective ratings of these forecasts were recorded during these evaluations based on the overall radar evolution, watches and warnings issued, and preliminary storm reports. Additionally, objective verification statistics were calculated with respect to preliminary storm reports to assist in determining if later forecasts were more skillful.

Model evaluations for SFE2013 focused especially on new experimental guidance used in making the update forecasts throughout the day. Specifically, the NME was evaluated on its ability to accurately represent the mesoscale and synoptic-scale pre-convective environments favorable for severe weather and was compared to the RAPv2, which provides background fields for an experimental parallel version of the hourly SPC Mesoscale Analyses. The two primary foci of the evaluation were determining the fit of the 1-h forecasts of 2-m temperature and dewpoint from the NME mean and RAPv2 to observations and the ability of the 1-h forecasts of CAPE/CIN from the NME mean and RAPv2 to represent observed sounding structures from NWS radiosonde sites.

Additionally, 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.

Other model evaluations were also performed to advance our understanding of different aspects of convection-allowing models. For SFE2013, this included a comparison of forecasts from the hot-start NSSL-WRF and the Met Office convection-allowing runs to output from the current cold-start configuration of the NSSL-WRF. Additionally, an evaluation of the SSEF members varying only in microphysics schemes was performed to assess the perceived skill of simulated reflectivity and brightness temperature forecasts compared to corresponding observations.

3. PRELIMINARY FINDINGS AND RESULTS

3.1 Experimental Forecast Evaluation

With two teams making forecasts of total severe thunderstorm probabilities for four periods (1600-1200, 1800-2100, 2100-0000, and 0000-0300 UTC) including two updates to the final two periods, there were many forecasts to evaluate. Subjective ratings were assigned by the other team during the next-day evaluation period and objective forecast verification was also performed. Objective verification metrics included the critical success index (CSI) and fractions skill score (FSS). In addition, the relative skill score (Hitchens et al. 2013) was introduced to gauge the performance of the experimental forecasts against a baseline reference, namely the practically perfect hindcasts (Brooks et al. 1998). Overall, the evaluation focused on addressing these basic questions: 1) Can skillful probabilistic forecasts of total severe weather be made at higher temporal resolution?, 2) Can the temporal disaggregation of the full-period forecast using SSEO calibrated model guidance provide a reasonable first guess for the 3-h periods?, and 3) Did the forecast updates improve upon the earlier forecasts?

3.1.1 Temporal Resolution

To address the first question, the subjective ratings of the quality of the full-period forecasts were compared to the ratings of the final 3-h period forecasts (Fig. 1). The biggest difference in the distribution of forecast ratings for the full period (Fig. 1a) and the individual 3-h periods (Figs. 1b-d) was the larger number of “poor” or “very poor” forecasts in the 3-h periods when compared to the full period. This is not surprising given that an error in the expected timing of severe storms can lead to a poor forecast in a 3-h period, but would have little to no impact on the full-period forecast. Nevertheless, the 3-h period forecasts received more “good” and “very good” ratings than “poor” or “very poor” ratings. In fact, the only period for which the ratings peaked at the “good” category was the 2100-0000 UTC period, which is coincident with the occurrence of maximum afternoon instability and diurnal convection.

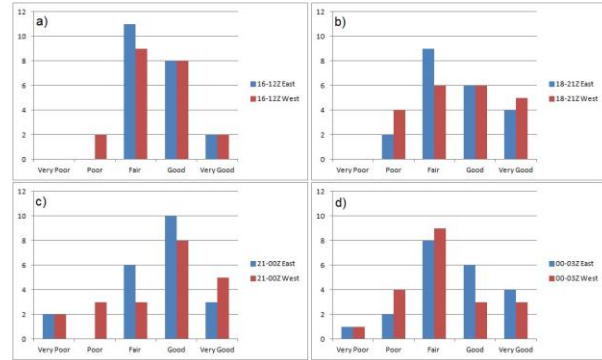


Figure 1. Subjective ratings assigned by participants to probabilistic total severe forecasts by the **east** and **west** teams valid from a) 1600-1200 UTC, b) 1800-2100 UTC, c) 2100-0000 UTC, and d) 0000-0300 UTC.

Throughout SFE2013, the full-period forecasts objectively verified better than the 3-h forecast periods (Fig. 2). A close inspection reveals that this improvement (i.e., in CSI) is mostly a result of lower FAR in the full-period forecast when compared to the 3-h period forecasts. The 1800-2100 UTC period had the fewest number of severe weather reports (Fig. 3) and the lowest CSI of the 3-h periods, which is likely related to larger uncertainty in both the timing of convective initiation and the transition of storms to severe levels during the early-to-mid afternoon. The CSI increased during the 2100-0000 UTC period and then dropped off slightly during the 0000-0300 UTC period. Overall, the subjective ratings and verification metrics indicate that satisfactory probabilistic forecasts of severe weather were made for 3-h periods during SFE2013 with the highest ratings/scores for the 2100-0000 UTC period.

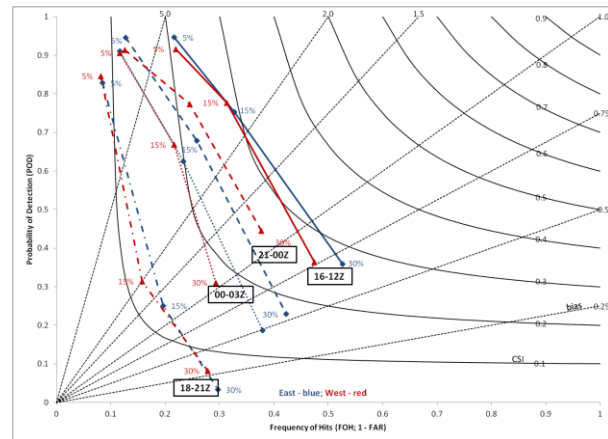


Figure 2. Performance diagram showing the accumulated statistics during SFE2013 for the final forecasts from the **east** and **west** teams. The full period (1600-1200 UTC – solid lines) and 3-h periods (1800-2100 UTC – dot/dash lines; 2100-0000 UTC – dash lines; 0000-0300 UTC – dotted lines) are shown. Data points denote forecast performance for 5%, 15%, and 30% probability thresholds.

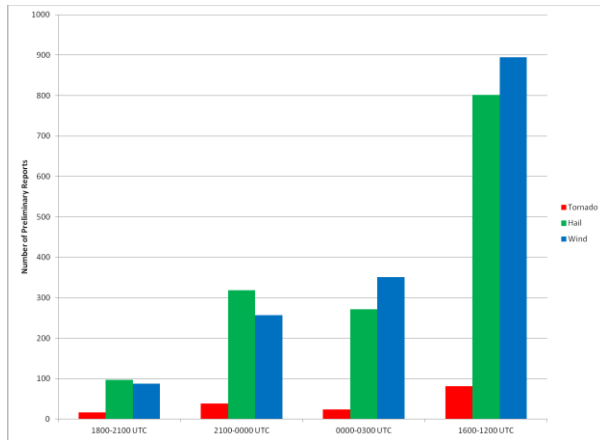


Figure 3. Total number of preliminary severe reports of **tornadoes**, **hail**, and **wind** within the daily mesoscale area of interest during SFE2013 for each of the forecast periods: 1800-2100, 2100-0000, 0000-0300, and 1600-1200 UTC.

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.

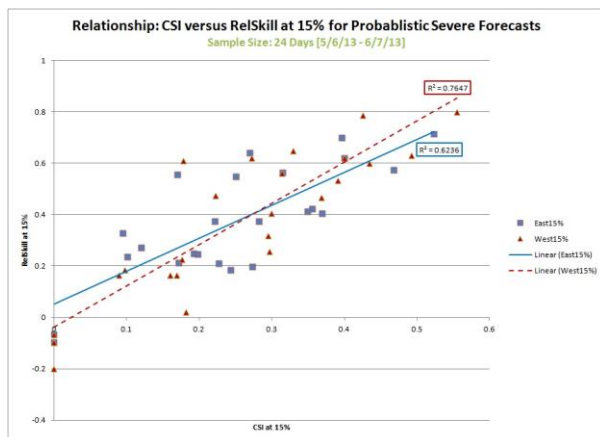


Figure 4. Scatter plot of CSI versus relative skill for the full-period (i.e., 1600-1200 UTC) probabilistic severe forecasts at 15% by the **east** and **west** teams during SFE2013.

3.1.2 Temporal Disaggregation

With the greatest forecaster skill typically occurring in longer-period outlooks (e.g., Fig. 2), a method to temporally disaggregate those forecasts into 3-h periods was applied during SFE2013. During the next-day evaluations, the initial human 3-h forecasts were subjectively compared to the temporally disaggregated 3-h forecasts. The results of this survey revealed that the manually drawn 3-h forecasts were generally “better” to “about the same” as the temporally disaggregated automated forecasts (Fig. 5). The manual forecast was rated “worse” than the temporally disaggregated forecast only a small number of times over the five-week period.

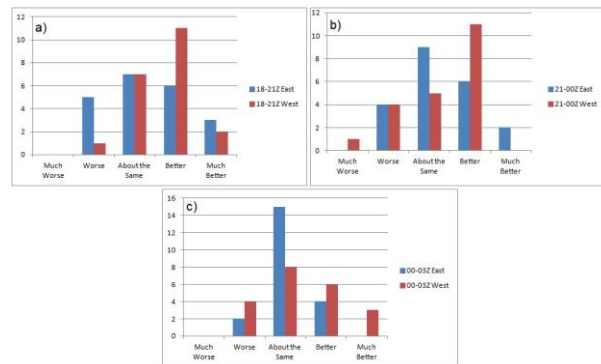


Figure 5. Subjective ratings assigned by participants to the initial forecasts relative to the temporally disaggregated first guess valid from a) 1800-2100 UTC, b) 2100-0000 UTC, and c) 0000-0300 UTC.

The objective verification statistics for the temporally disaggregated forecasts (Figs. 6 and 7) are in good agreement with the subjective ratings shown in Fig. 5. For the east team, the temporally disaggregated forecasts were statistically very similar to the initial manual forecast for the 2100-0000 UTC and 0000-0300 UTC periods (Fig. 6). For the west team, the manual forecasts were generally a little better statistically for all periods and thresholds (Fig. 7). This is consistent with the subjective impressions where the east team forecast was more likely to be rated “about the same” as the temporally disaggregated forecast while the west team forecast was more likely to be rated “better” than the temporally disaggregated forecast. The difference in results between the teams is likely related to the east forecast team more closely following the SSEO calibrated severe guidance, which is used in the temporal disaggregation procedure.

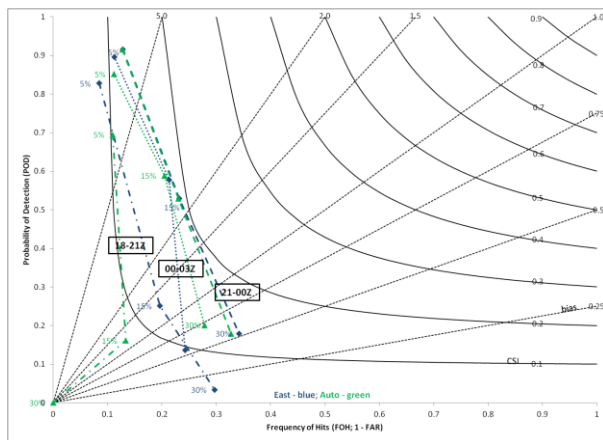


Figure 6. Performance diagram showing the accumulated statistics during SFE2013 for the initial forecasts from the **east** team and the **temporally disaggregated** forecasts from the east team full-period forecast. The individual 3-h periods (1800-2100 UTC – dot/dash lines; 2100-0000 UTC – dash lines; 0000-0300 UTC – dotted lines) are shown.

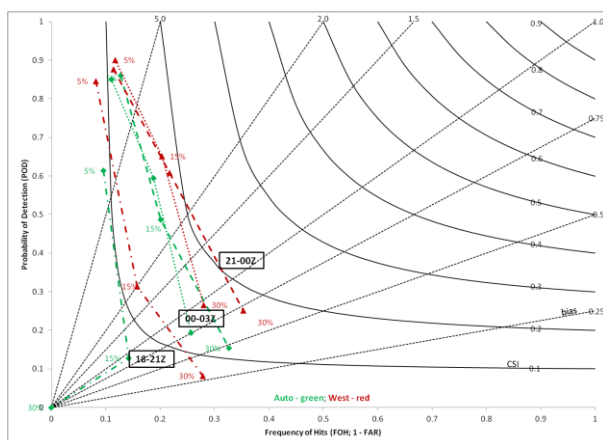


Figure 7. Performance diagram showing the accumulated statistics during SFE2013 for the initial forecasts from the **west** team and the **temporally disaggregated** forecasts from the west team full-period forecast. The individual 3-h periods (1800-2100 UTC – dot/dash lines; 2100-0000 UTC – dash lines; 0000-0300 UTC – dotted lines) are shown.

3.1.3 Forecast Updates

The last question regarding forecast evaluation focused on assessing whether improvement was made in the forecast updates as new guidance and observations became available. The subjective ratings from the participants indicated that the forecast updates were usually “about the same” as or “better” than the previous forecast (Fig. 8). The updates rarely resulted in a degraded forecast, nor did they often result in a “much better” forecast. The other key point to note is that the final update forecast (Figs. 8b and 8d) was more likely to be “about the same” as the previous forecast than the earlier update. Anecdotally, the participants often felt that the updated hourly guidance [e.g., NME and High-Resolution Rapid Refresh (HRRR)] wasn’t compelling and/or different enough to make significant changes in the final update – only small

adjustments were typically made and were based on observational trends, especially to the 2100-0000 UTC period if storms had already formed.

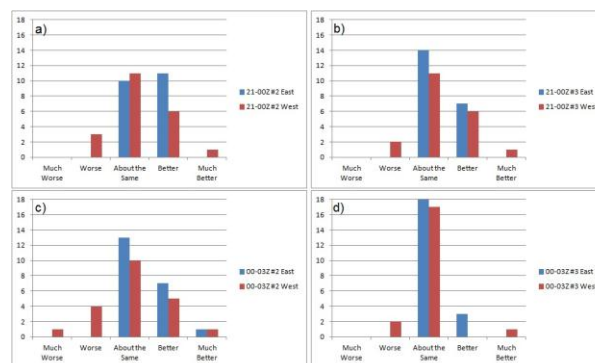


Figure 8. Subjective ratings assigned by participants to the forecast updates relative to the previous forecasts: a) 2100-0000 UTC Update, b) 2100-0000 UTC Final, c) 0000-0300 UTC Update, and d) 0000-0300 UTC Final.

The update forecasts generally showed a modest, steady statistical improvement from the initial forecast, to the update forecast, and ultimately the final forecast (Figs. 9 and 10). The east team generally showed the most improvement at higher thresholds for the update forecast (Fig. 9). There was less statistical improvement for the final forecast, which was consistent with the subjective results when the majority of final forecasts were rated “about the same” as the previous forecasts. The statistical results were a little different for the west team, as the 0000-0300 UTC forecasts with longer lead time did not vary much statistically with each update (Fig. 10) while the 2100-0000 UTC forecasts showed a steady improvement by update with the largest increase occurring with the final update. A difference in forecast update philosophy between the east and west teams is evident when comparing the 30% threshold for the 2100-0000 UTC forecasts. The east team tended to maintain a constant bias with each update (i.e., reduce FAR while barely increasing POD) while the west team had an increasing bias with each update (i.e., increase POD without much decrease in FAR). Overall, these results from SFE2013 show that there is some value in updating the forecasts both from a subjective and objective perspective; however, the frequency of useful updates would likely depend on the new guidance available (i.e., observational and model) and the specific weather scenario.

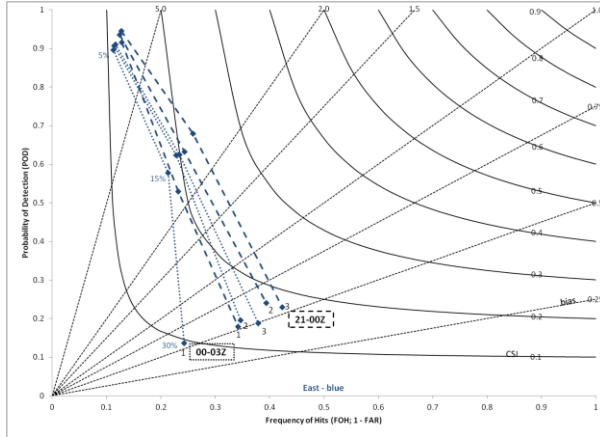


Figure 9. Performance diagram showing the accumulated statistics during SFE2013 for the initial (1), update (2), and final (3) forecasts from the east team for the 2100-0000 UTC (dashed lines) and 0000-0300 UTC (dotted lines) periods.

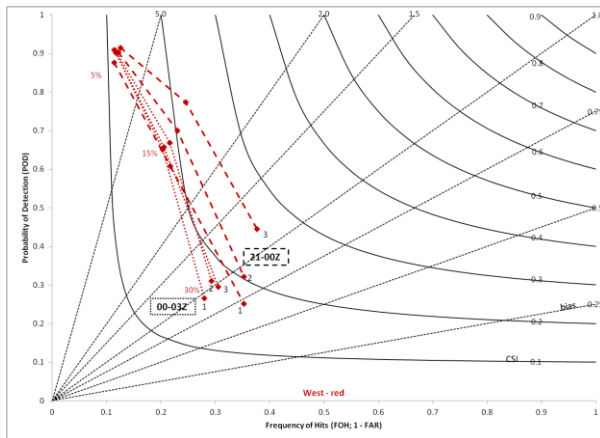


Figure 10. Performance diagram showing the accumulated statistics during SFE2013 for the initial (1), update (2), and final (3) forecasts from the west team for the 2100-0000 UTC (dashed lines) and 0000-0300 UTC (dotted lines) periods.

3.2 NME Evaluation

The fit of the 1-h forecasts of 2-m temperature from the NME mean and RAPv2 to surface observations was subjectively rated about the same for a majority (~65%) of the SFE 5-week period, while ~25% of the time, the NME mean fit was better. A consistent signal was shown when comparing the NME mean to RAPv2 2-m temperatures throughout the SFE. During the 1400 – 1800 UTC timeframe, the RAPv2 was generally much warmer, relative to the observations, when compared to the NME (e.g., one 1800 UTC comparison is shown in Fig. 11). The RAPv2 1-h forecast had a $\geq 2^\circ\text{F}$ warm bias during this time period over the daily mesoscale area of interest (Fig. 12). This signal in the RAPv2 1-h forecast lessens in the 1800 UTC – 0000 UTC period and in fact switches to a cool bias by 0000 UTC. In the 0000 UTC – 0300 UTC timeframe, the RMSE of the 1-h forecasts of the NME and RAPv2 are very similar with both showing a cool bias at 0300 UTC. The evolution in the RAPv2 1-h forecast of 2-m temperature is consistent

with the MYNN PBL scheme used in the model. The 2-m temperature field was also useful for identifying convectively-generated cold pools present in the models. The NME produced smoother cold pool structures, as expected from an ensemble mean, when compared to the RAPv2, but the NME mean cold pools were generally too warm when compared against observations, making the RAPv2 a better fit.

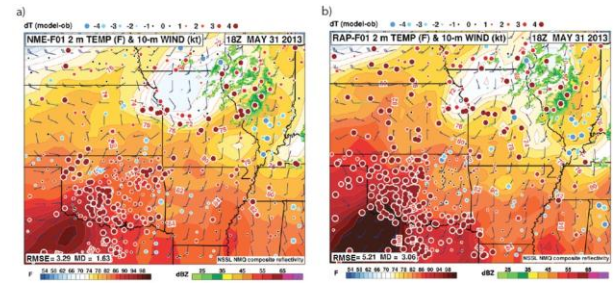


Figure 11. 1-h forecast of 2-m temperature (\square F) and 10-m winds (kts) valid at 1800 UTC 31 May 2013 from the a) NME and b) RAPv2. Red (blue) dots indicate points where the model temperature is warmer (cooler) than the observations. Domain-averaged root-mean-squared error (RMSE) and mean difference (MD; bias) between the model and observations are shown in the bottom-left corner. NSSL NMQ composite reflectivity (dBZ; see label bar) at the valid time is also shown.

The fit of the NME mean and RAPv2 1-h forecasts of 2-m dewpoint temperature to observations exhibited a similar pattern to the 2-m temperatures. For a majority (~59%) of the SFE, their performance was similar, while for ~32% of the time, the NME showed a better fit to the observations. Overall, the NME mean 1-h forecasts of 2-m dewpoint showed lower RMSE than the RAPv2 during SFE2013, especially during the 1500 UTC – 0000 UTC period (Fig. 13). The most substantial differences between the NME and the RAPv2 were in the vicinity of drylines. The RAPv2 1-h forecast generally placed the dryline too far east too quickly during the day, indicating that the NME 1-h forecast had a better location of the dryline, which has significant implications for convective initiation forecasts.

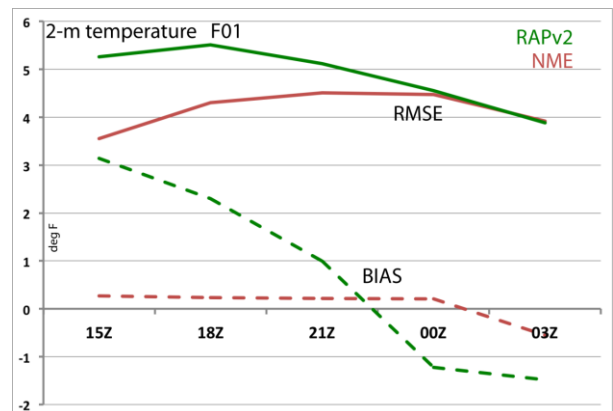


Figure 12. Cumulative RMSE (solid lines) and bias (MD; dashed lines) by valid time for one-hour forecasts of 2-m temperature for the RAPv2 (green) and the NME mean (red) over the mesoscale area of interest during SFE2013.

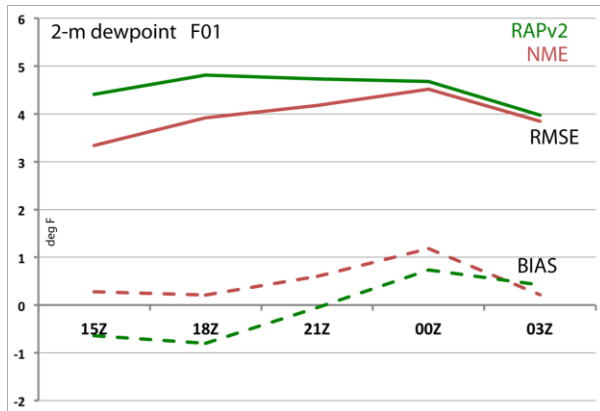


Figure 13. Same as Fig. 12, except for one-hour forecasts of 2-m dewpoint temperature.

Comparison of the 1-h forecasts of surface-based (SB) CAPE/CIN between the NME mean and RAPv2 showed a similar pattern as the 1-h forecasts of 2-m temperature and dewpoint. Both performed similarly from a subjective perspective for ~61% of the time, while the NME mean performed better for ~28% of the SFE period. Each had trouble capturing the strength of strong inversions on a few days, which has implications for the likelihood of convective development. On May 15th, a high-impact severe weather day with a few strong tornadoes in the Dallas-Fort Worth (DFW) Metropolitan area, the NME 1-h forecast of SBCAPE/SBCIN provided a much better representation of the pre-convective environment around 0000 UTC 16 May 2013 (Fig. 14). Observed SBCAPE from the Fort Worth (FWD) sounding was ~2700 J kg⁻¹, which was better represented by the NME 1-h forecast with maximum values between 2000-2500 J kg⁻¹. The RAPv2 1-h forecast indicated lower values of SBCAPE of 1500 – 2000 J kg⁻¹. The high SBCAPE present across the DFW metropolitan area was a primary factor in the intense supercellular development in this region. The NME 1-h forecast of SBCAPE thus provided better guidance of this convective potential, which was ultimately realized.

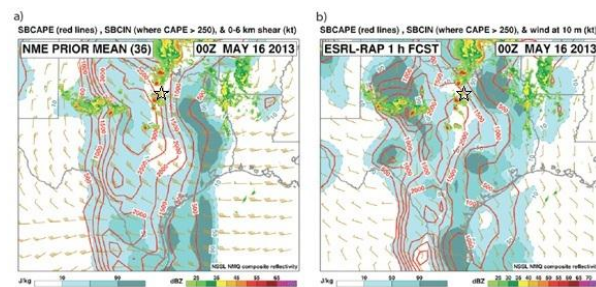


Figure 14. 1-h forecast of surface-based convective potential energy (SBCAPE; red lines), surface-based convective inhibition (SBCIN; blue shading) where SBCAPE > 250 J kg⁻¹ valid at 0000 UTC 16 May 2013 from the a) NME and b) RAPv2. The left panel shows the 0-6 km shear (kts), while the right panel shows the winds at 10 m above ground level. NSSL NMQ composite reflectivity (dBZ; see label bar) at the valid time is also shown with approximate location of FWD indicated by the star.

3.3 Convection-Allowing Ensemble Evaluation

Forecasts from the 1200 UTC-initialized ensembles were available for examination for the first time in SFE2013. Given model spin-up time in developing convection and the climatological difference in convective activity between 0000 UTC and 1200 UTC, the 1200 UTC guidance provided an opportunity for an interesting comparison of ensembles at different initialization times and with different initialization strategies. There were two primary components to this comparison between 1200 UTC and 0000 UTC 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 timing, location, orientation, magnitude, etc. of ensemble probabilities to radar reflectivity observations during the 1300-0600 UTC forecast period, the 1200 UTC ensembles were generally rated “about the same” as or “better” than their corresponding 0000 UTC ensemble probabilities (Fig. 15). It is worth noting that forecasts could be rated “about the same” without actually being similar to one another during much of the evaluation period (i.e., positive and negative aspects cancelling each other). The 1200 UTC SSEO was most frequently rated “about the same” as the 0000 UTC SSEO while the 1200 UTC SSEF received a “better” rating than the 0000 UTC SSEF more often than any other rating. Much of the benefit in the 1200 UTC SSEF reflectivity probabilities occurs in the first several hours, as the assimilation of radar data in this ensemble provides information about the location, intensity, and orientation of ongoing storms. The AFWA probabilistic reflectivity fields could often not be cleanly evaluated owing to a processing error in which reflectivity fields were accumulated over time.

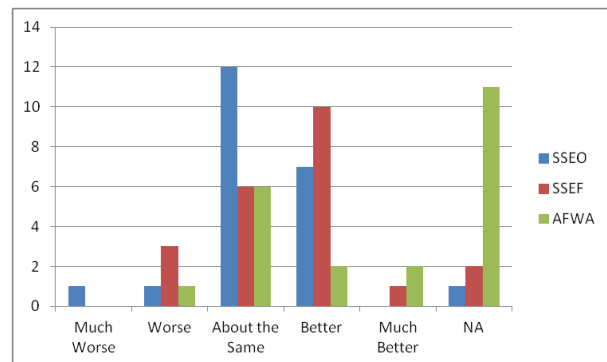


Figure 15. Subjective ratings assigned by participants to the 1200 UTC ensemble neighborhood probability forecasts of 1-km AGL reflectivity ≥40 dBZ relative to the same forecasts from the 0000 UTC ensembles (SSEO, 8-member SSEF, and AFWA) valid during the 1300-0600 UTC forecast period.

Examination of the fractions skill score (FSS) by forecast hour during the experiment over the daily movable mesoscale area of interest reveals more information about the ensemble characteristics (Fig. 16). Forecasts of reflectivity from the 1200 UTC SSEF were much better than the 0000 UTC SSEF during the first

four hours of the 1200 UTC cycle, illustrating the near-term benefits of the assimilation of radar data. The 0000 UTC and 1200 UTC SSEF forecasts were comparable from that time on through the end of the forecast cycle (i.e., 0600 UTC). On the other hand, the 1200 UTC SSEO, which does not assimilate radar data, had much lower FSS than the 0000 UTC SSEO for the first two hours of the forecast cycle. After the initial spin-up time, the 1200 UTC SSEO held a narrow advantage over the 0000 UTC SSEO and the SSEF forecasts during the period of peak convective activity (i.e., 2200–0600 UTC).

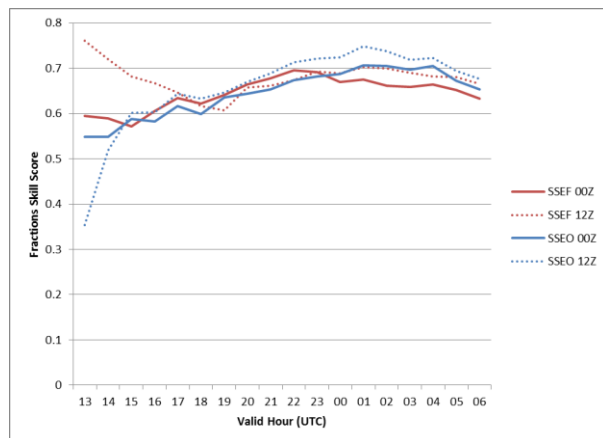


Figure 16. Accumulated fractions skill score by forecast hour for neighborhood probabilities of reflectivity ≥ 40 dBZ for the SSEF and SSEO.

Interestingly, the subjective ratings of the ensemble forecasts of hourly maximum storm-attribute fields (HMFs; Fig. 17) for severe weather forecasting purposes are distributed quite a bit differently than the ratings for the reflectivity forecasts. Compared to the reflectivity evaluation, there were more instances when the 1200 UTC ensemble forecasts were rated “worse” than the 0000 UTC ensemble forecasts (cf., Figs. 15 and 17). In fact, there were more HMF forecasts from the 1200 UTC SSEF rated “worse” than those rated “better” when compared to the 0000 UTC SSEF even though the 1200 UTC reflectivity forecasts were generally considered better (cf., Fig. 15). In general, there was a nearly even distribution of ratings among “worse”, “about the same”, and “better” for each of the ensembles, which suggests that the more recently initialized 1200 UTC ensembles need to be carefully scrutinized to determine whether they are an improvement over the 0000 UTC ensembles from a severe storm-attribute field perspective.

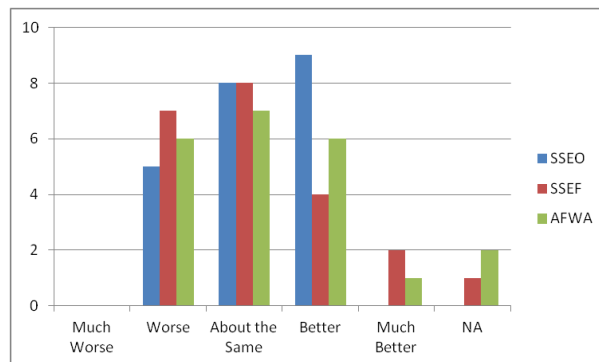


Figure 17. Subjective ratings assigned by participants to the 1200 UTC ensemble forecasts of hourly maximum storm-attribute fields relative to the same forecasts from the 0000 UTC ensembles (SSEO, 8-member SSEF, and AFWA).

When comparing the subjective ratings of the overall usefulness of the ensemble HMFs for severe weather forecasting, several features stand out (Fig. 18). For the 0000 UTC ensembles, the distribution of rankings was fairly narrow with the majority of forecasts being rated as “fair”. The 0000 UTC SSEO and SSEF forecasts were skewed toward the “good” rating while the 0000 UTC AFWA was slightly skewed toward the “poor” rating. The 1200 UTC ensembles had a much broader distribution of ratings (Fig. 18b) than the 0000 UTC ensembles (Fig. 18a). The peak in ratings was no longer pronounced at the “fair” rating, as more “good”, “very good”, and “poor” ratings were given to all of the ensembles. Thus, even though the 1200 UTC ensembles are initialized closer to the time of the event, the distribution of the perceived quality of the forecasts is broader than those forecasts initialized 12 hours earlier at 0000 UTC. Overall, the three ensembles were subjectively ranked similarly during SFE2013 for severe weather forecasting, indicating that current formal approaches with more advanced physics and data assimilation to storm-scale ensembles do not necessarily result in an obvious performance advantage at this stage of development. These results indicate that continuing research is needed to improve the configuration of storm-scale ensembles, including the testing of scale-appropriate perturbation strategies.

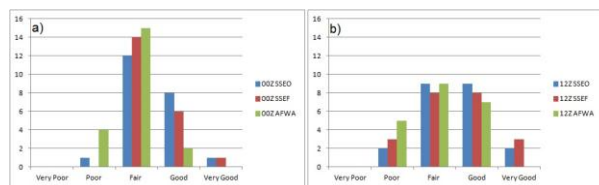


Figure 18. Subjective ratings assigned by participants to the a) 0000 UTC and b) 1200 UTC ensemble forecasts of hourly maximum storm-attribute fields on the usefulness to a severe weather forecaster (SSEO, 8-member SSEF, and AFWA).

3.4 NSSL-WRF Comparisons

The cold-start NSSL-WRF was compared subjectively to three other convection-allowing runs during SFE2013: the hot-start NSSL-WRF, the 4.4 km Met Office UM, and the 2.2 km Met Office UM. The hot-start NSSL-WRF was configured identically to the cold-start NSSL-WRF run so that the impact of the NME analyses in initializing the forecasts could be evaluated. In addition, the NSSL-WRF was also compared to the two deterministic convection-allowing models run by the Met Office. To compare these runs, a new interactive web display called the NSSL Interactive Data Explorer was developed to allow zooming, overlaying of chosen fields, and side-by-side comparisons of model and observational fields. The evaluations focused especially on simulated reflectivity during the Day 1 period, and participants were asked the following:

1. "Using the NSSL Interactive Experimental Data Explorer, and focusing on areas of interesting weather, evaluate whether the "hot-start" NSSL-WRF forecasts improved upon the cold-start NSSL-WRF. Please provide explanation/description/reasoning for answer."
2. "Using the NSSL Interactive Experimental Data Explorer, and focusing on areas of interesting weather, compare the 4.4-km Met Office forecasts to the cold-start NSSL-WRF. Please provide explanation/description/reasoning for answer."
3. "Please comment on the utility of the NSSL Interactive Data Explorer in conducting these evaluations. How does this tool compare to other methods for forecast evaluation? Do you have suggestions for improvements?"

For item 1, there were a total of 20 responses, which are summarized in Fig. 19. Slightly more often (40% of the time), it was determined that the hot-start run was worse than the cold start. However, 30% of responses rated the hot start runs as better and another 30% rated the runs as not being better or worse. Given the small sample size, the results should be used with caution. Some general themes from the comparisons were that there were often very large differences in the forecasts. In many cases, for one particular time period either the hot or the cold start would perform better, but then at other time periods the best performing model would switch. Thus, in many of the cases in which the models were rated as performing the same, it was not because they had similar forecasts, rather the relative good or bad skill during particular periods cancelled out in the overall rating. Finally, on many occasions it was obvious that the quality of the forecast during the afternoon was strongly tied to how well overnight/early morning convection was depicted earlier in the model integration.




		Response Percent	Response Count
Hot-start NSSL-WRF better than cold-start		30.0%	6
Hot-start NSSL-WRF worse than cold-start		40.0%	8
Neither model better/worse than the other		30.0%	6

Figure 19. Summary of responses for the hot versus cold start comparisons.

For item 2 there were a total of 16 responses. The majority of responses indicated that the 4.4-km Met Office UM forecast was better (50%) or the same (37.5%) relative to the cold-start NSSL-WRF, with only two cases (12.5%) in which the NSSL-WRF was rated as better than the Met Office UM (Fig. 20). For the cases in which Met Office UM performed better than NSSL-WRF there was a wide variety of reasons. These reasons include the following: 1) the Met Office UM better depicting an MCV and related convection, 2) the Met Office UM suppressing convection in the correct locations, and 3) the Met Office UM better depicting timing and placement of convection. Perhaps one flaw that was noticed in the Met Office UM was that it did not appear to handle the upscale growth and transition of storms into linear systems very well. Oftentimes, when a well-defined linear convective system existed in reality, the Met Office UM would depict large clusters of intense storms that never organized into coherent lines. It was speculated that the Met Office UM was not simulating cold pools very well, but this is an avenue for more thorough analysis. In the comments, participants were encouraged to identify differences between the 2.2 and 4.4 km versions of the Met Office UM. For these comparisons, there were a couple cases in which it was noted that the 2.2 km version did better with convective mode and evolution of storms, but for the most part participants described the 2.2 and 4.4 km forecasts as being very similar.




		Response Percent	Response Count
UKMET better than NSSL-WRF		50.0%	8
UKMET worse than NSSL-WRF		12.5%	2
Same		37.5%	6

Figure 20. Summary of responses for the Met Office UM (UKMET) versus NSSL-WRF comparisons.

Item 3 asked participants to comment on the utility of the NSSL Interactive Data Explorer (e.g., Fig. 21). Some of the comments, such as expressing the need for multi-panel displays and clearer plot labels, were incorporated into the Explorer during the experiment. In general, the Data Explorer was received very positively by participants, supporting the further utilization of this visualization and analysis tool in future SFEs.

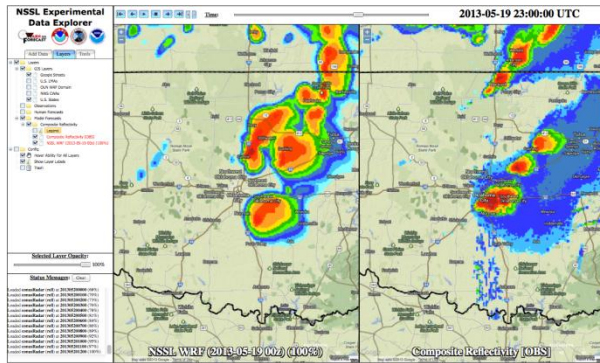


Figure 21. Example of side-by-side zoomed-in Data Explorer display of cold-start 0000 UTC NSSL-WRF forecasts of simulated reflectivity valid at 2300 UTC on 19 May (left) and corresponding observations of composite reflectivity (right).

3.5 Microphysics Comparisons

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 SFEF ensemble members with identical configurations except for their microphysical parameterization. During SFE2013, the following microphysics parameterizations were systematically examined: Thompson, Milbrandt-Yau (MY), Morrison, NSSL, WDM6, and a modified version of Thompson in which the coupling to the RRTMG short-wave radiation scheme was improved (Thompson-mod). In Thompson-mod, the effective radii of cloud water, ice, and snow is passed from the microphysics to RRTMG, unlike Thompson in which internal assumptions within RRTMG about the size of cloud droplets, ice, and snow are used. SFE2013 also marked the first time that the NSSL microphysics scheme was examined. The NSSL scheme is also known as the Ziegler Variable Density (ZVD) scheme and is double-moment with respect to cloud droplets, rain drops, ice crystals, snow, graupel, and hail.

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), m20 (Milbrandt-Yau), m21 (Morrison), m22 (WDM6), and m23 (NSSL) during the 18z-12z period, based on comparisons with corresponding observations. Also, comment on CN (Thompson) versus m25 (Thompson with coupled radiation).”

Some of the general themes from the responses were that, Morrison, Thompson, and NSSL generally had the most realistic depiction of convection in terms of simulated reflectivity and brightness temperatures. MY had a tendency to simulate storms that were too intense and too large. One of the most striking characteristics of MY was its tendency to produce regions of cold cloud tops associated with convection that were significantly larger than the other microphysics scheme and

observations. In contrast, WDM6 tended to produce regions of cold cloud tops associated with convection that were much smaller than the other schemes and observations, with storms that often dissipated too quickly. In addition, WDM6 oftentimes produced the most intense cold pools associated with convection that would expand and eliminate convective instability (e.g., SBCAPE) much faster than the other schemes (i.e., a characteristic of outflow-dominant storms). In general, all the schemes over-predict convective instability, with the NSSL scheme associated with the largest instability. It was not clear what was causing the larger values of CAPE in NSSL because examination of low-level temperature and dewpoint fields did not reveal noticeable differences relative to the other schemes. Figures 22 and 23 illustrate examples of forecast simulated brightness temperatures and composite reflectivity, respectively, which were the main fields examined on a daily basis for comparing the microphysics. Future work is planned to conduct more objective and systematic comparisons of these members. The initial findings from these subjective evaluations should provide a starting point for future studies.

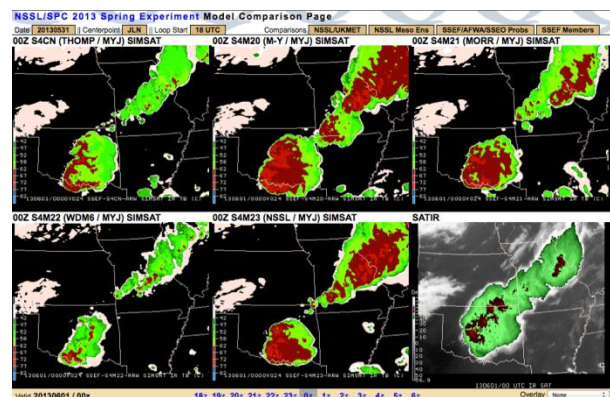


Figure 22. 24-hour forecasts of simulated brightness temperatures valid 0000 UTC 1 June 2013 from the Thompson, M-Y, Morrison, WDM6, and NSSL microphysics members. Corresponding observations are in the lower-right panel. The member labels are at the top of each plot.

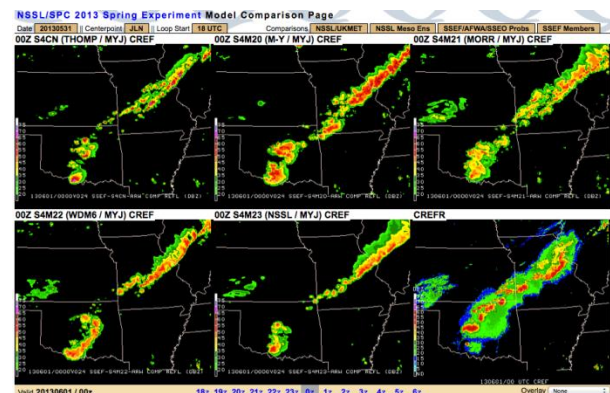


Figure 23. Same as Fig. 22, except for composite reflectivity.

4. SUMMARY

The 2013 Spring Forecasting Experiment (SFE2013) was conducted at the NOAA Hazardous Weather Testbed from May 6 – June 7 by the SPC and NSSL with participation from more than 30 forecasters, researchers, and developers from around the world. The primary theme of SFE2013 was to explore the utility of short-term convection-allowing and mesoscale ensemble model guidance in creating frequently updated, high-temporal resolution probabilistic forecasts of severe weather. Several preliminary findings from SFE2013 are listed below:

- Next-day verification metrics provided a useful tool for objectively evaluating experimental forecasts and model performance and offered a standard reference against which subjective impressions could be compared.
- The full-period forecasts generally verified better than 3-h periods owing primarily to lower FAR with the most skillful 3-h probabilistic forecasts of severe weather occurring from 2100-0000 UTC.
- Updates typically improved the forecasts from both a subjective and objective perspective though improvements for the final update, especially with more lead time (i.e., 0000-0300 UTC period), were usually small.
- The NME generally performed better than the deterministic RAPv2 for short-term forecasts of the pre-convective environment. With more development work, this promising ensemble approach should improve analyses and short-term forecasts of the environment relevant to convective forecasting.
- Forecasts from 1200 UTC convection-allowing ensembles displayed a broader distribution of forecast ratings than the 0000 UTC ensembles for severe weather guidance.
- More work is needed in the perturbation strategy and design of formal convection-allowing ensembles to improve the overall forecast performance for severe weather events.
- The initial conditions had a noticeable impact on 0000 UTC NSSL WRF convective forecasts with the quality of the forecast during the afternoon often strongly tied to how well overnight and early morning convection was depicted.
- An effective collaboration with the Met Office was established through five-week participation and examination of their convection-allowing model runs, which proved to be very competitive with WRF-ARW based models.

Overall, SFE2013 was successful in testing new tools and modeling systems to address relevant issues related to the prediction of hazardous convective weather. The findings and questions exposed during SFE2013 are certain to lead to continued progress in the forecasting of severe weather.

Acknowledgements. SFE2013 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 SFE2013. Evan Kuchera and Scott Rentschler of AFWA generously provided AFWA data during SFE2013.

A full version (with comments from participants) of this report can be found on the SFE2013 website: http://hwt.nssl.noaa.gov/Spring_2013/HWT_SFE_2013_Prelim_Findings_final.pdf

REFERENCES

- Brooks, H.E., M.P. Kay, and J.A. Hart, 1998: Objective limits on forecasting skill of rare events. *Preprints*, 19th Conf. Severe Local Storms. Minneapolis, MN, Amer. Meteor. Soc., 552-555.
- Hitchens, N. M., H. E. Brooks, and M. P. Kay, 2013: Objective limits on forecasting skill of rare events. *Wea. Forecasting*, **28**, 525–534.
- Torn, R., D., G. J. Hakim, and C. Snyder, 2006: Boundary conditions for limited-area ensemble Kalman filters. *Mon. Wea. Rev.*, **134**, 2490-2502.

APPENDIX

Daily Activities Schedule

Scheduled activities are in local (CDT) time and conducted as one large group unless otherwise indicated. Two separate groups will be generating identical forecast products.

Pre-0800: “Teaser”. Because we will not immediately begin evaluating the previous day’s forecast, relevant loops (radar, water vapor, visible imagery, storm reports, etc.) will be displayed as participants arrive so they can get a quick look at how the previous day’s forecasts verified.

0800 – 0930: Full-period forecast. Begin activities with hand analyses of 1200 UTC upper-air data and surface charts. Then, large-scale overview and group forecast discussion with consensus selection of a forecast domain. Break into two forecast groups and issue probabilistic forecasts of total severe valid 1600 UTC to 1200 UTC the next day.

0930 – 0945: Break

0945 – 1015: Evaluation of previous day’s human forecasts. As two groups, each forecast will be subjectively rated. Each group will rate the forecasts generated by the other group. Also, it will be decided whether the updates continuously improved the forecasts.

1015 – 1100: Model evaluations. Participants will remain in two separate groups. Group 1 will perform evaluations comparing the 0000 UTC initialized storm-scale ensembles to their 1200 UTC initialized counterparts (SSEO, AFWA, and SSEF systems). Group 1 will also compare analyses generated from the NSSL Mesoscale Ensemble (NME) to those generated from the ESRL RAPv2-based SFC-Objective Analyses (SFCOA). Group 2 will examine the impact of microphysics schemes by comparing forecasts from the 5 SSEF system members that differ only by their microphysics parameterizations. Emphases will be placed on comparing two versions of the Thompson scheme as well as the new NSSL double-moment scheme. Group 2 will also conduct comparisons of the operational NSSL-WRF to a parallel version initialized from the 0000 UTC NME analysis using a Google-maps-based interactive comparison interface. Comparisons will also be made to the Met Office’s convection-allowing model.

1100 – 1200: Update forecast #1 –Both groups will use 1400 UTC initialized NME forecasts and all other available observations and guidance to issue forecasts for the 1800-2100, 2100-0000, and 0000-0300 UTC time periods. A first guess for each time period will be generated using temporal disaggregation applied to the full-period forecast

issued earlier in the morning. The same products as from the initial forecast will be issued (i.e., probabilities of total and significant severe).

1200 – 1300: Lunch and possible collaboration with the EWP.

1300 – 1330: Weather Briefing – Highlights from yesterday, general overview, discussion of forecast challenges and products. In addition, each group will discuss reasoning for their forecasts.

1330 – 1430: Update forecast #2 – Same as #1, except for just the 2100-0000 and 0000-0300 UTC periods. The 1600 UTC initialized NME and 1200 UTC initialized convection-allowing ensembles will be available.

1430 – 1445: Break and possible collaboration with the EWP.

1445 – 1500: Open time period for discussion and questions of the day.

1500 – 1600: Update forecast #3 – Same as #2. The 1800 UTC initialized NME will be available and possible collaboration with the EWP.