SPRING FORECASTING EXPERIMENT 2013

Conducted by the

EXPERIMENTAL FORECAST PROGRAM

of the

NOAA HAZARDOUS WEATHER TESTBED

http://hwt.nssl.noaa.gov/Spring_2013/

HWT Facility – National Weather Center
6 May - 7 June 2013

Preliminary Findings and Results

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

1. Introduction

The 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. However, owing to National Weather Service (NWS) travel restrictions and forecaster staffing shortages at the SPC, the number of operational forecasters participating in SFE2013 was reduced considerably from previous years. 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 UKMET 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 and list of operational impacts can be found in Section 4.

2. Description

a) 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.

i. 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 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.

ii. 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.

iii. CAPS Storm Scale Ensemble Forecast (SSEF) 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.

iv. SPC Storm Scale Ensemble of Opportunity (SSEO)

The 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/budget 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.

v. Air Force Weather Agency (AFWA) 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 UKMET 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.

vi. UKMET Convection-Allowing Model Runs

The Unified Model (UM) is a generalized NWP system developed by the UKMET 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 used a 2D Smagorinsky boundary layer mixing scheme with single moment microphysics. The 4.4 km model used a CAPE limited closure shallow convective parameterization scheme, while the 2.2 km model did not utilize convective parameterization.

b) 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.

i. 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 ≥ 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.

ii. 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 UKMET 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

a) 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?

i) 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.

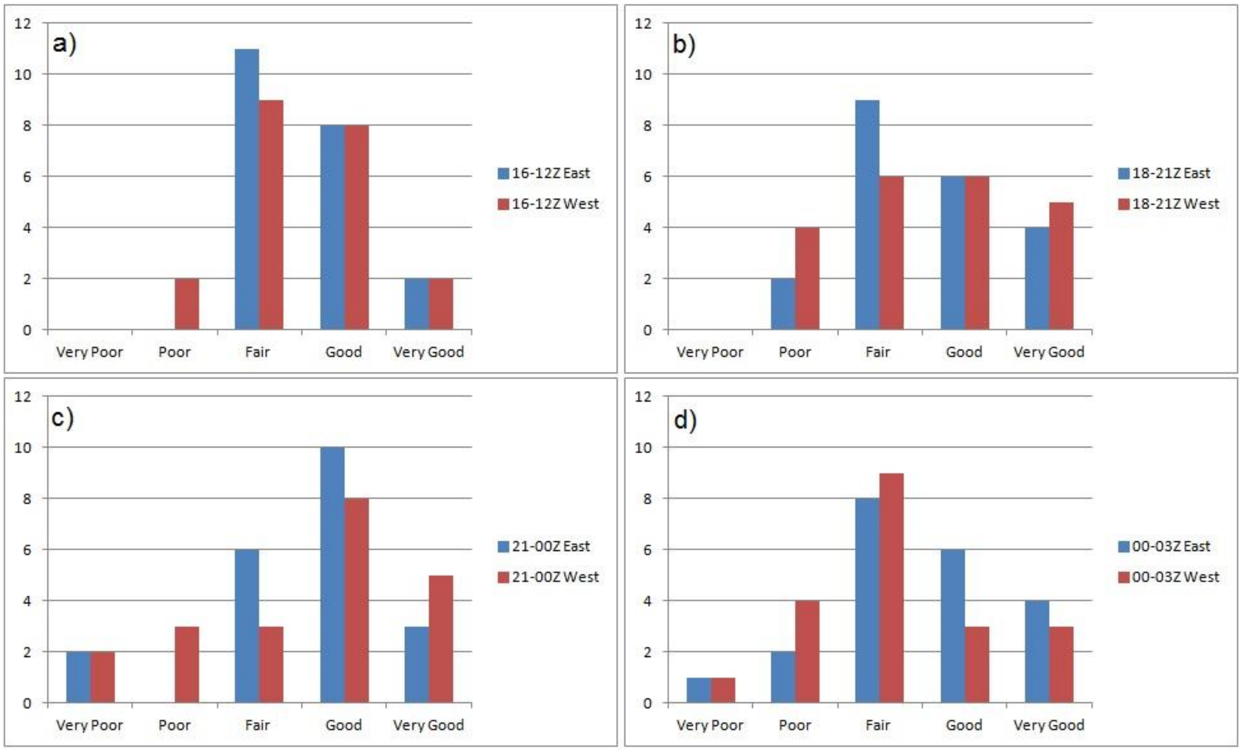


Figure 1. Subjective ratings assigned by participants to probabilistic total severe forecasts by the east and west teams valid from a) 1600-1200 UTC, b) 1800-2100 UTC, c) 2100-0000 UTC, and d) 0000-0300 UTC.

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-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.

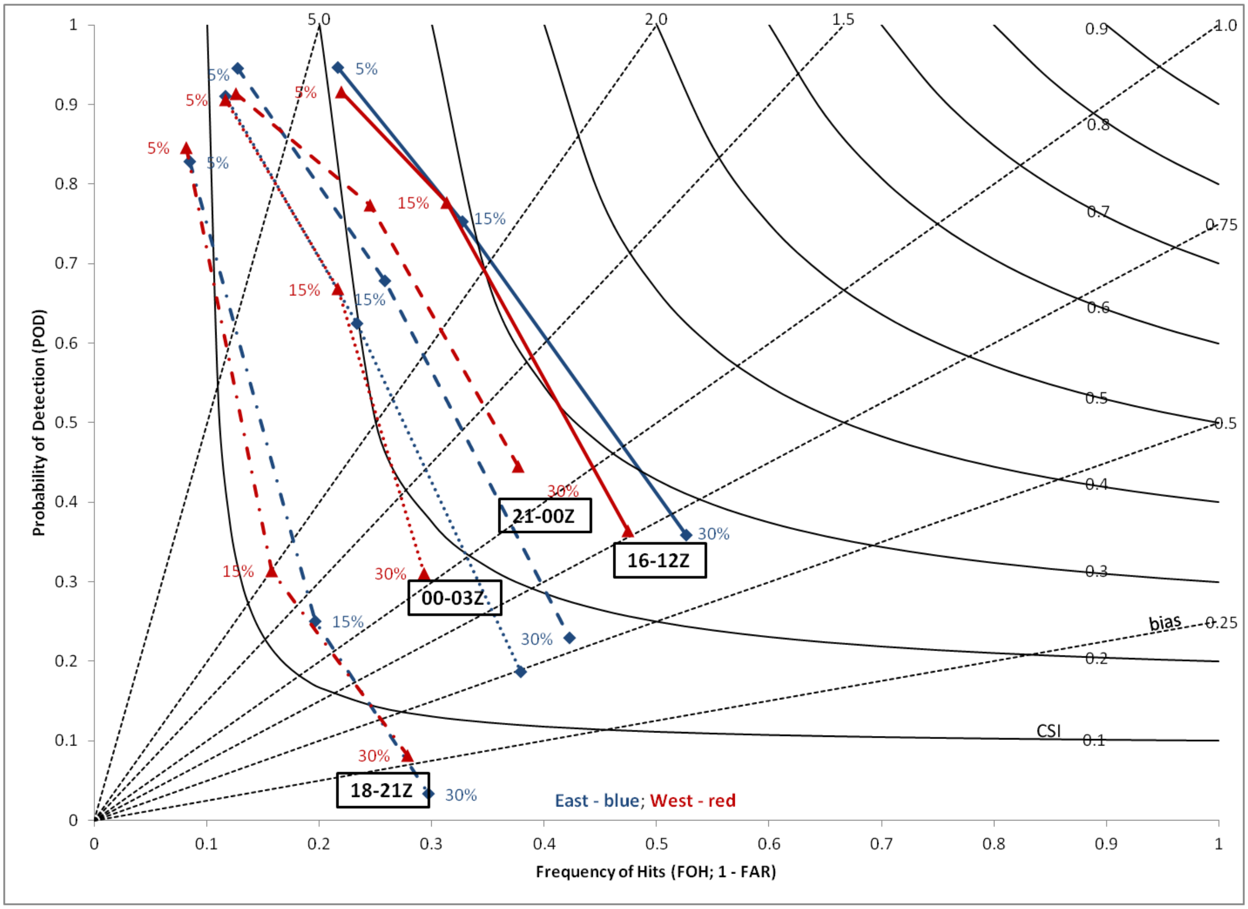


Figure 2. Performance diagram showing the accumulated statistics during SFE2013 for the final forecasts from the east and west teams. The full period (1600-1200 UTC – solid lines) and 3-h periods (1800-2100 UTC – dot/dash lines; 2100-0000 UTC – dash lines; 0000-0300 UTC – dotted lines) are shown. Data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

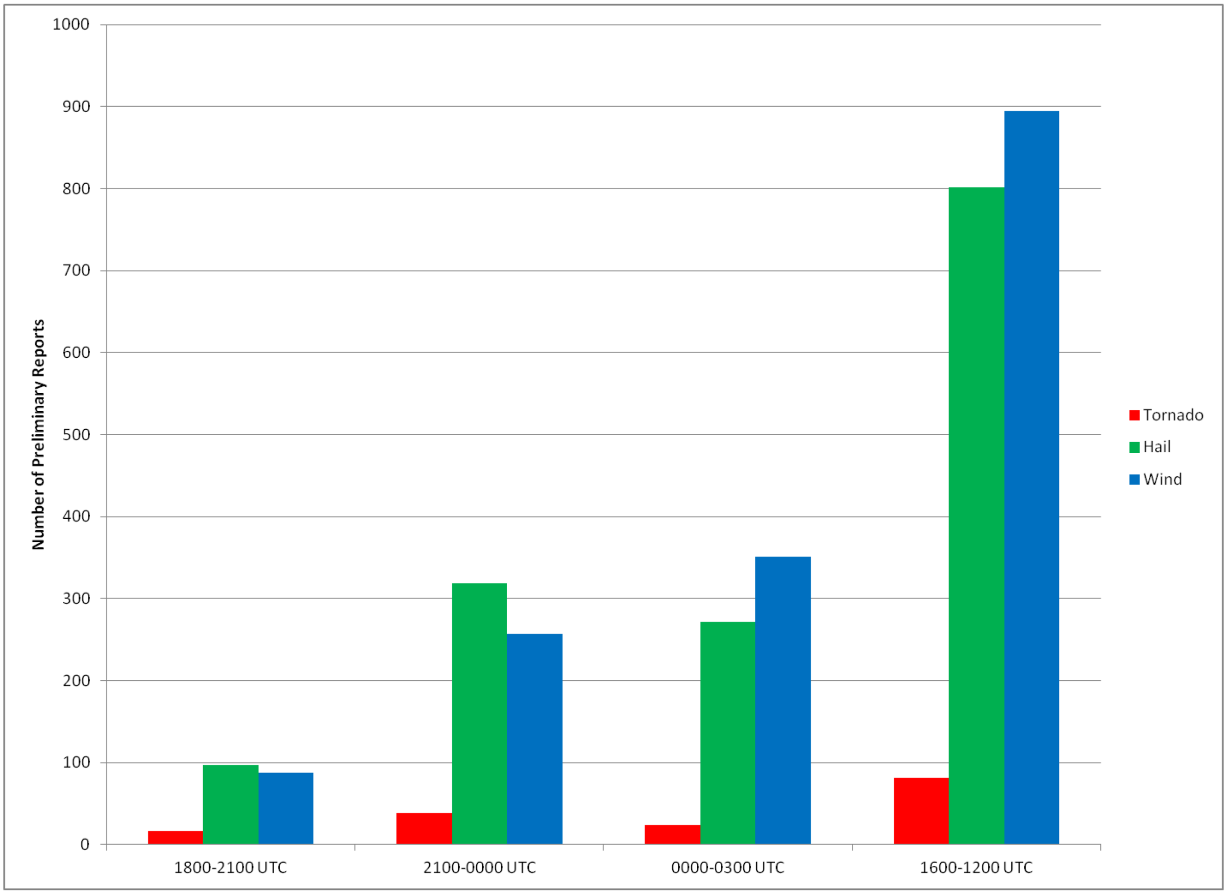


Figure 3. Total number of preliminary severe reports of tornadoes, hail, and wind within the daily mesoscale area of interest during SFE2013 for each of the forecast periods: 1800-2100, 2100-0000, 0000-0300, and 1600-1200 UTC.

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.

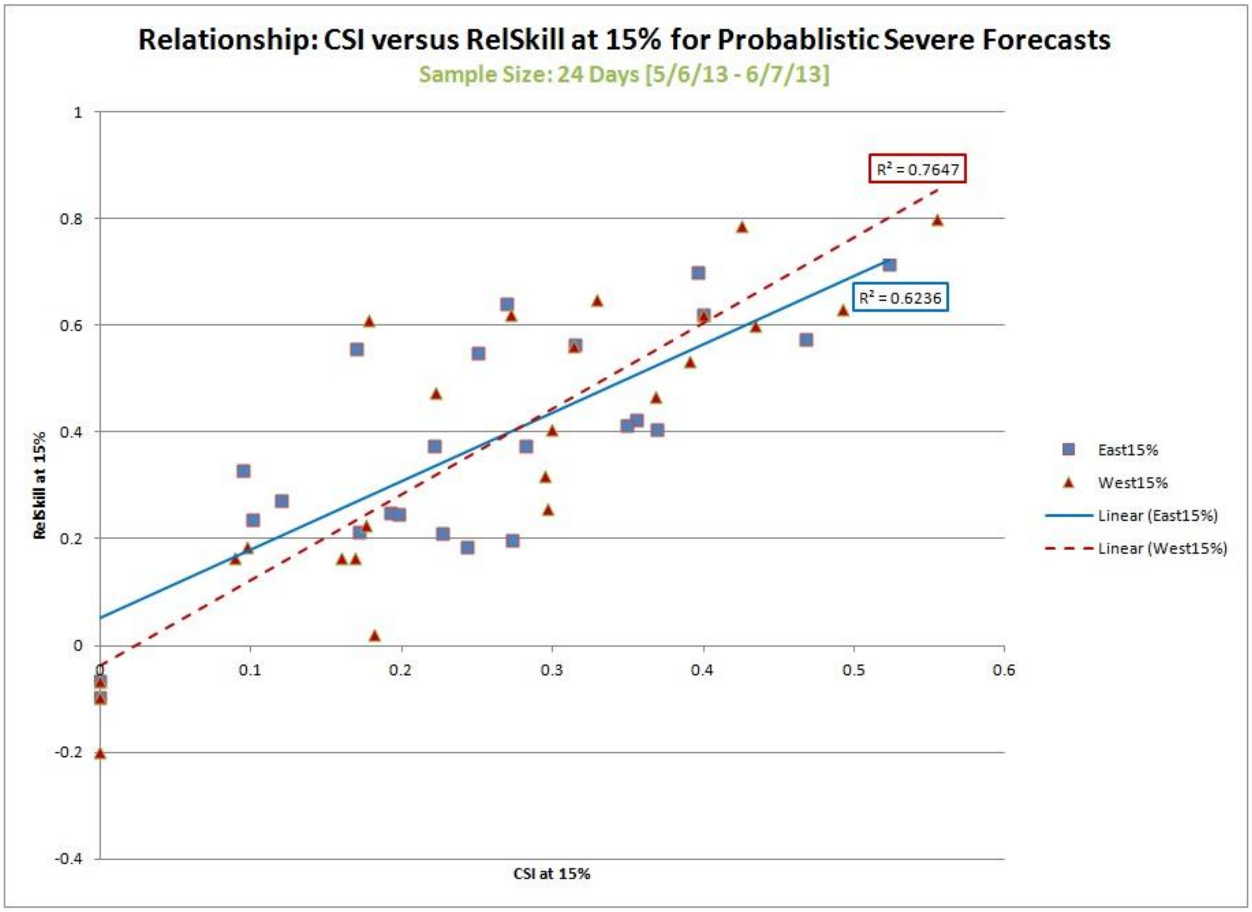


Figure 4. Scatter plot of CSI versus relative skill for the full-period (i.e., 1600-1200 UTC) probabilistic severe forecasts at 15% by the east and west teams during SFE2013.

ii) Temporal Disaggregation

Since the greatest forecaster skill typically occurs 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; comments in Table 1). The manual forecast was rated “worse” than the temporally disaggregated forecast only a small number of times over the five-week period.

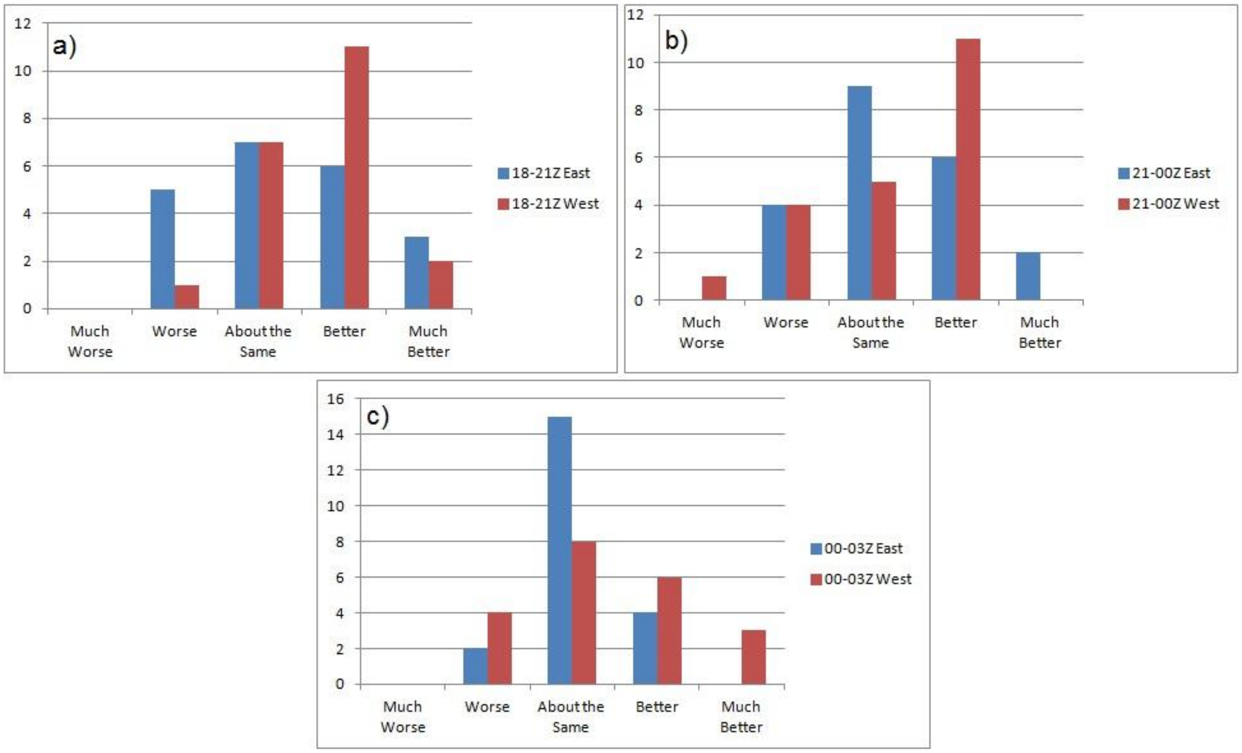


Figure 5. Subjective ratings assigned by participants to the initial forecasts relative to the temporally disaggregated first guess valid from a) 1800-2100 UTC, b) 2100-0000 UTC, and c) 0000-0300 UTC.

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.

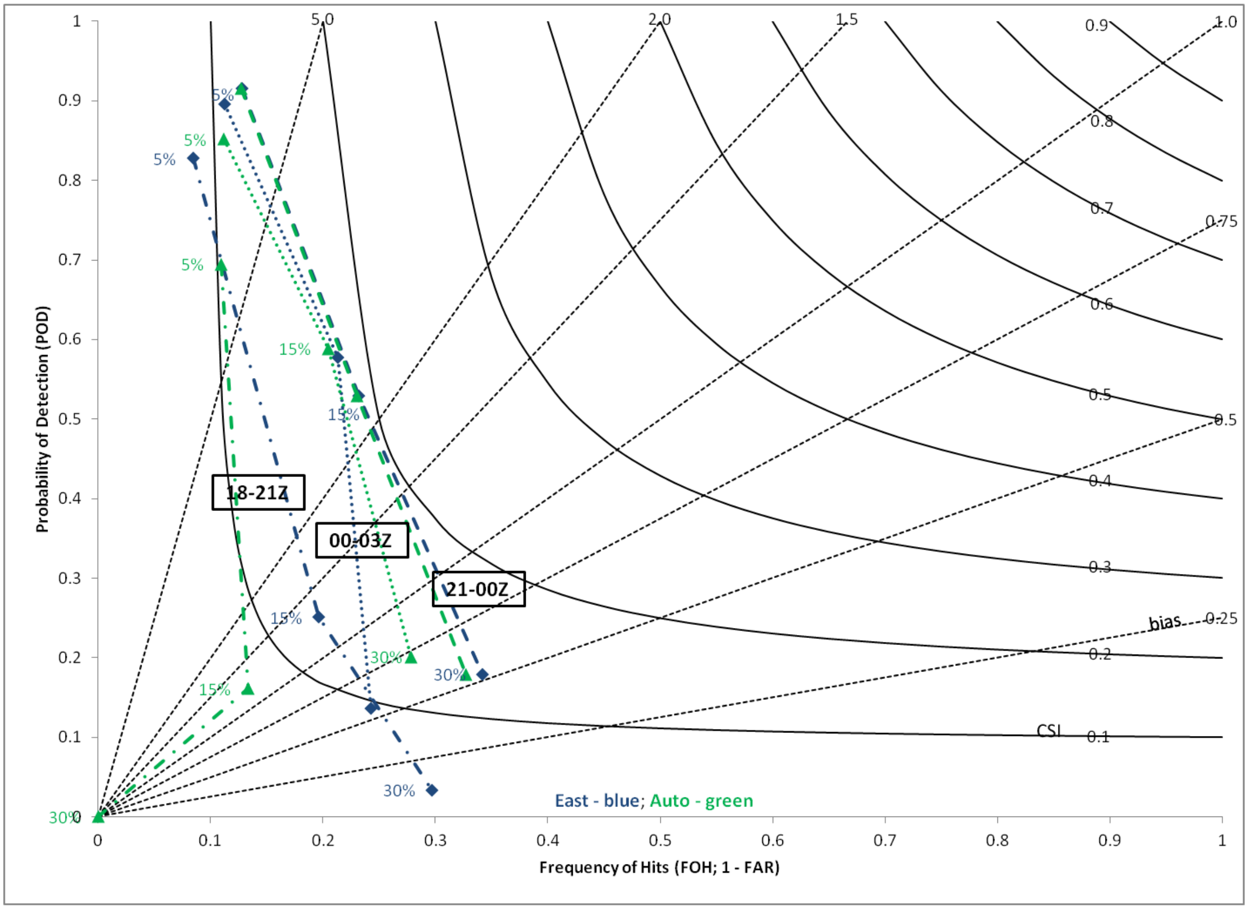


Figure 6. Performance diagram showing the accumulated statistics during SFE2013 for the initial forecasts from the east team and the temporally disaggregated forecasts from the east team full-period forecast. The individual 3-h periods (1800-2100 UTC – dot/dash lines; 2100-0000 UTC – dash lines; 0000-0300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

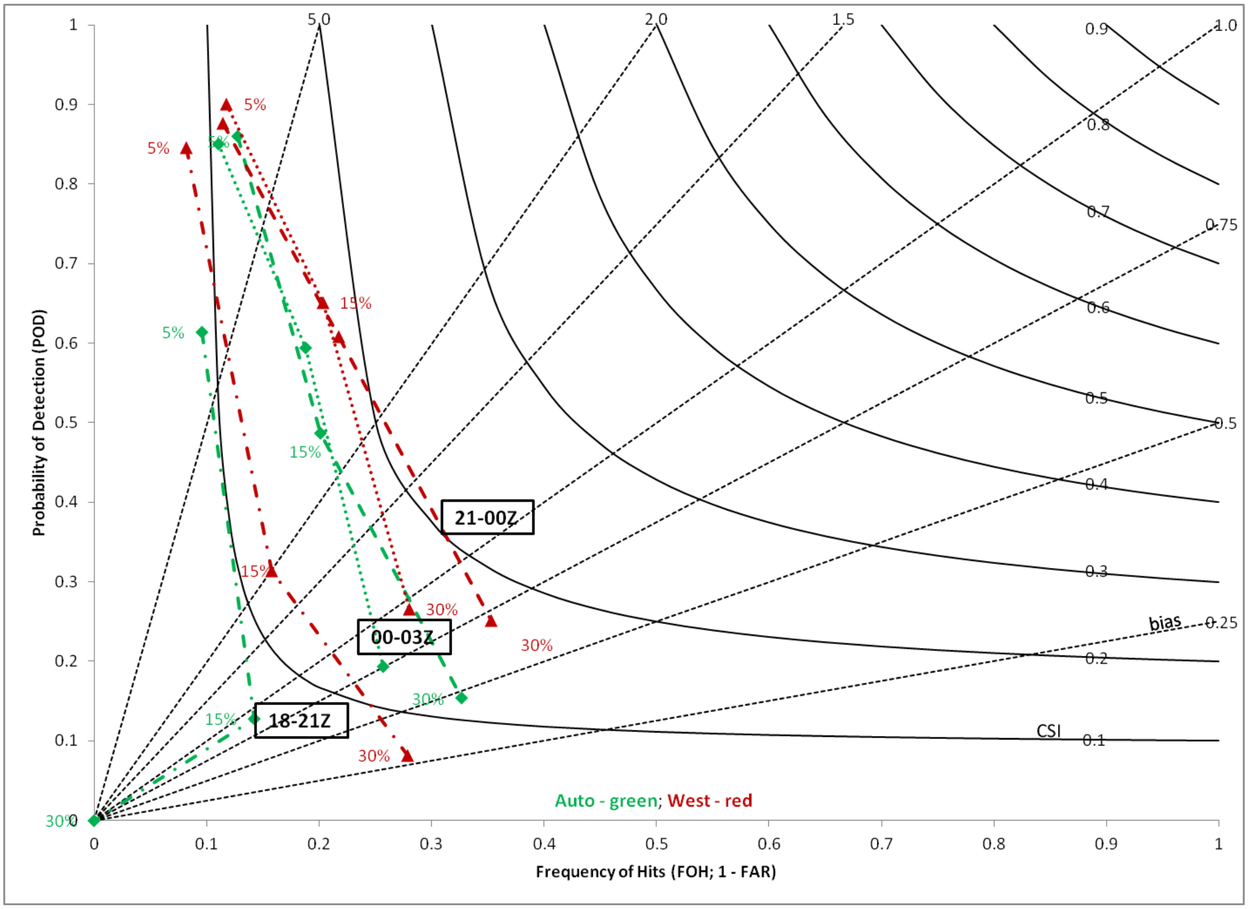


Figure 7. Performance diagram showing the accumulated statistics during SFE2013 for the initial forecasts from the west team and the temporally disaggregated forecasts from the west team full-period forecast. The individual 3-h periods (1800-2100 UTC – dot/dash lines; 2100-0000 UTC – dash lines; 0000-0300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

iii) 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; comments in Table 2). 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.

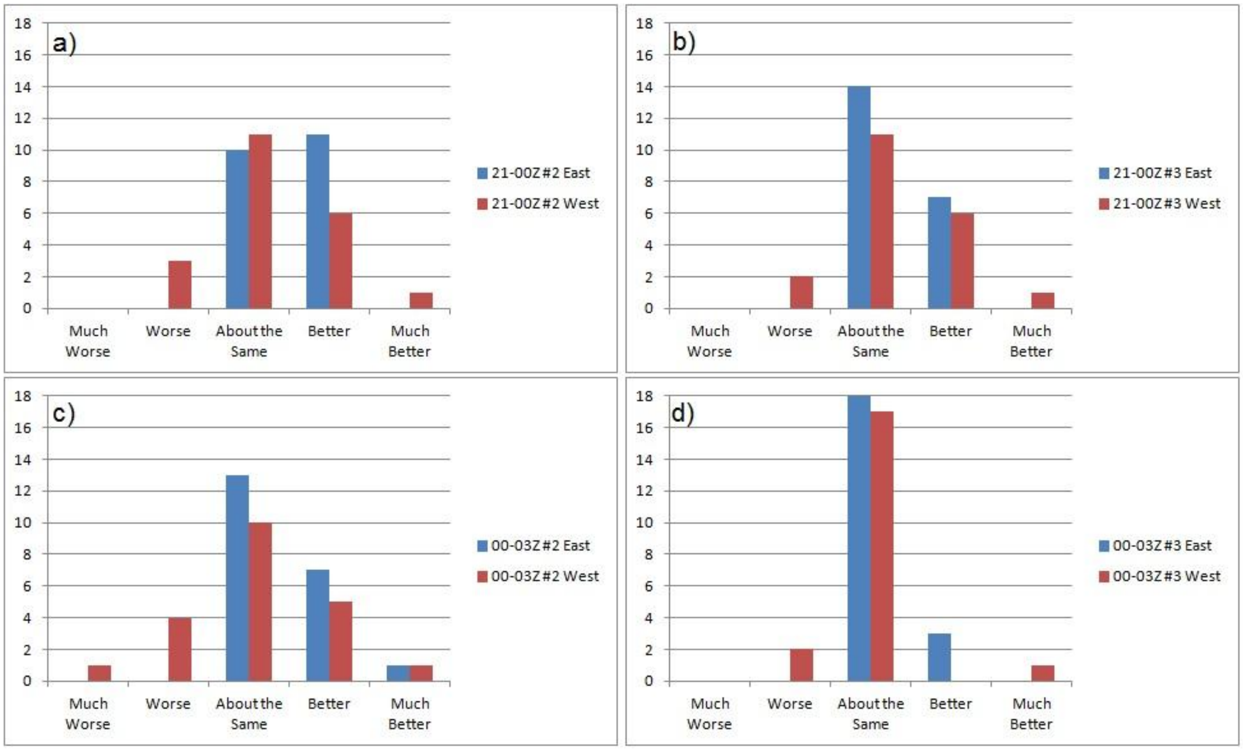


Figure 8. Subjective ratings assigned by participants to the forecast updates relative to the previous forecasts: a) 2100-0000 UTC Update, b) 2100-0000 UTC Update, c) 0000-0300 UTC Final, and d) 0000-0300 UTC Final.

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.

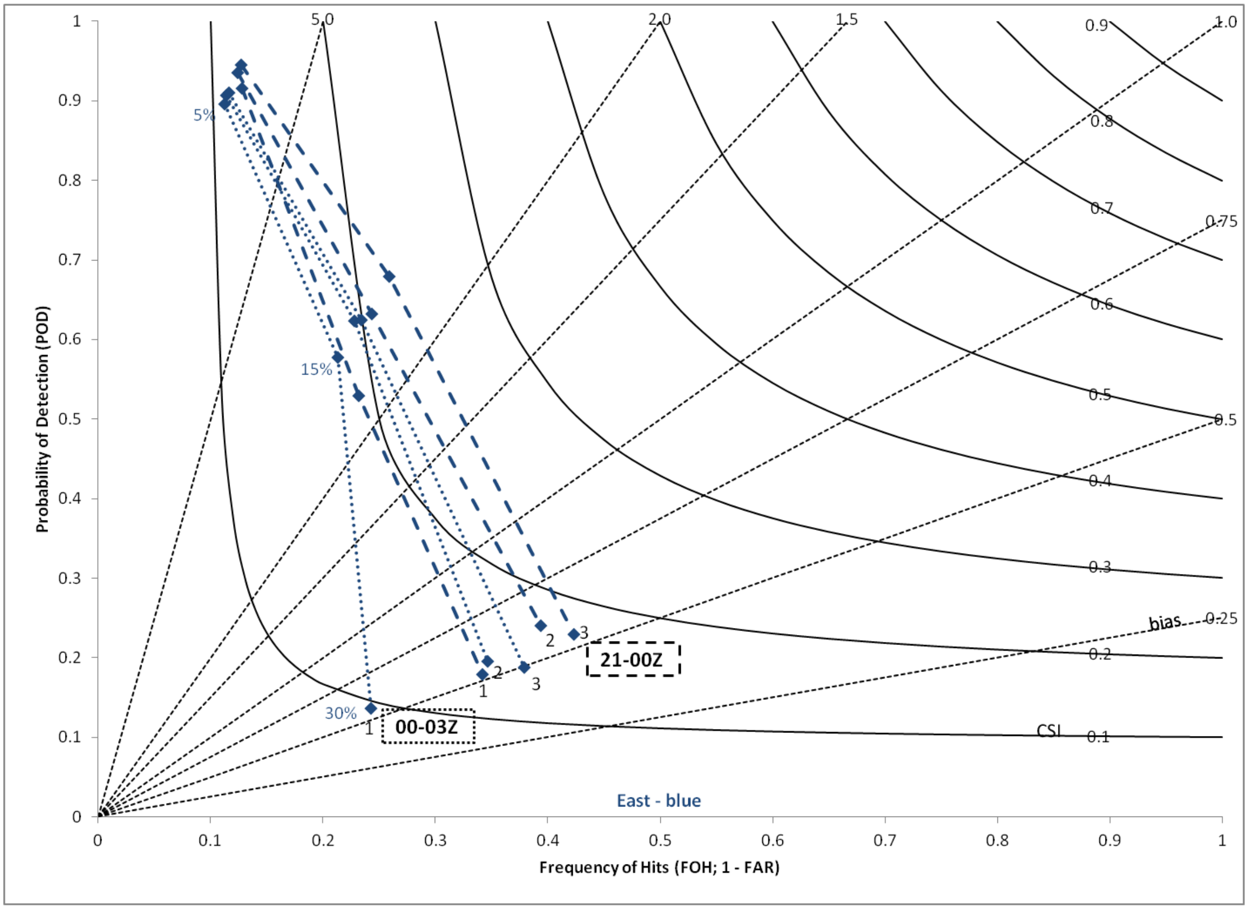


Figure 9. Performance diagram showing the accumulated statistics during SFE2013 for the initial (1), update (2), and final (3) forecasts from the east team for the 2100-0000 UTC (dashed lines) and 0000-0300 UTC (dotted lines) periods.

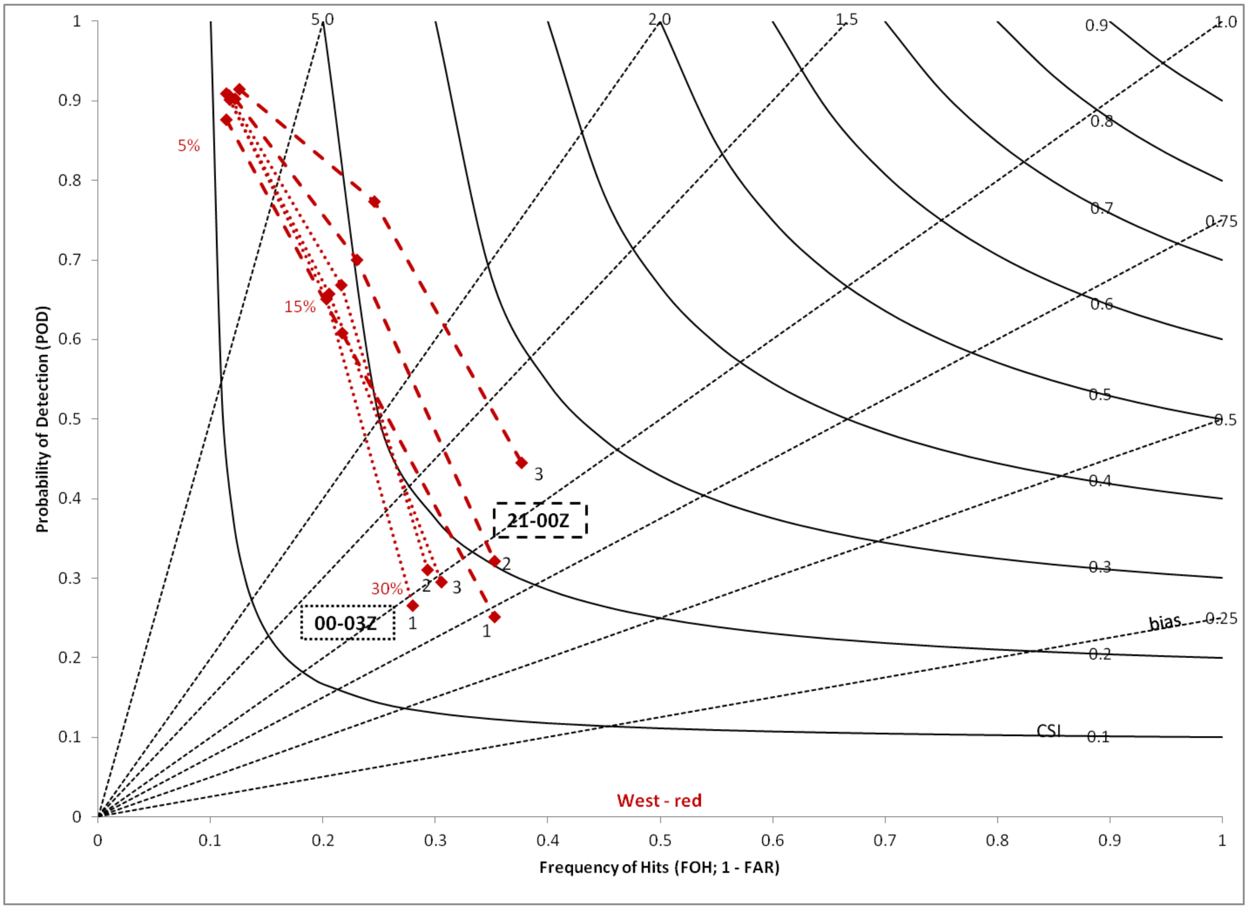


Figure 10. Performance diagram showing the accumulated statistics during SFE2013 for the initial (1), update (2), and final (3) forecasts from the west team for the 2100-0000 UTC (dashed lines) and 0000-0300 UTC (dotted lines) periods..

b) 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 (comments in Table 3). 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 ≥ 2°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.

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 (Table 3). 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.

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 (Table 3). 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 Forth 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.

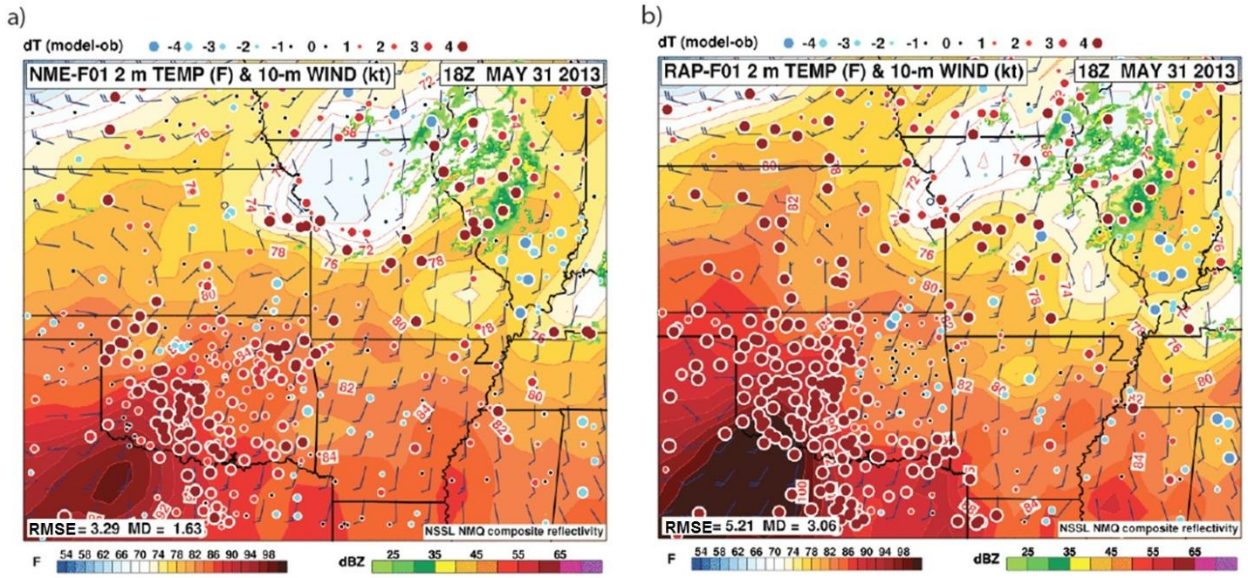

a)
dT (model-ob) -4 -3 -2 -1 0 1 2 3 4
NME-F01 2 m TEMP (F) & 10-m WIND (kt)
18Z MAY 31 2013
RMSE= 3.29 MD = 1.63
NSSL NMQ composite reflectivity
F 54 58 62 66 70 74 78 82 86 90 94 98
dBZ 25 35 45 55 65
b)
dT (model-ob) -4 -3 -2 -1 0 1 2 3 4
RAP-F01 2 m TEMP (F) & 10-m WIND (kt)
18Z MAY 31 2013
RMSE= 5.21 MD = 3.06
NSSL NMQ composite reflectivity
F 54 58 62 66 70 74 78 82 86 90 94 98
dBZ 25 35 45 55 65

Figure 11. 1-h forecast of 2-m temperature (°F) and 10-m winds (kts) valid at 1800 UTC 31 May 2013 from the a) NME and b) RAPv2. Red dots indicate points where the model temperature is higher than the observations, while blue dots indicate points where the model temperature is 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.

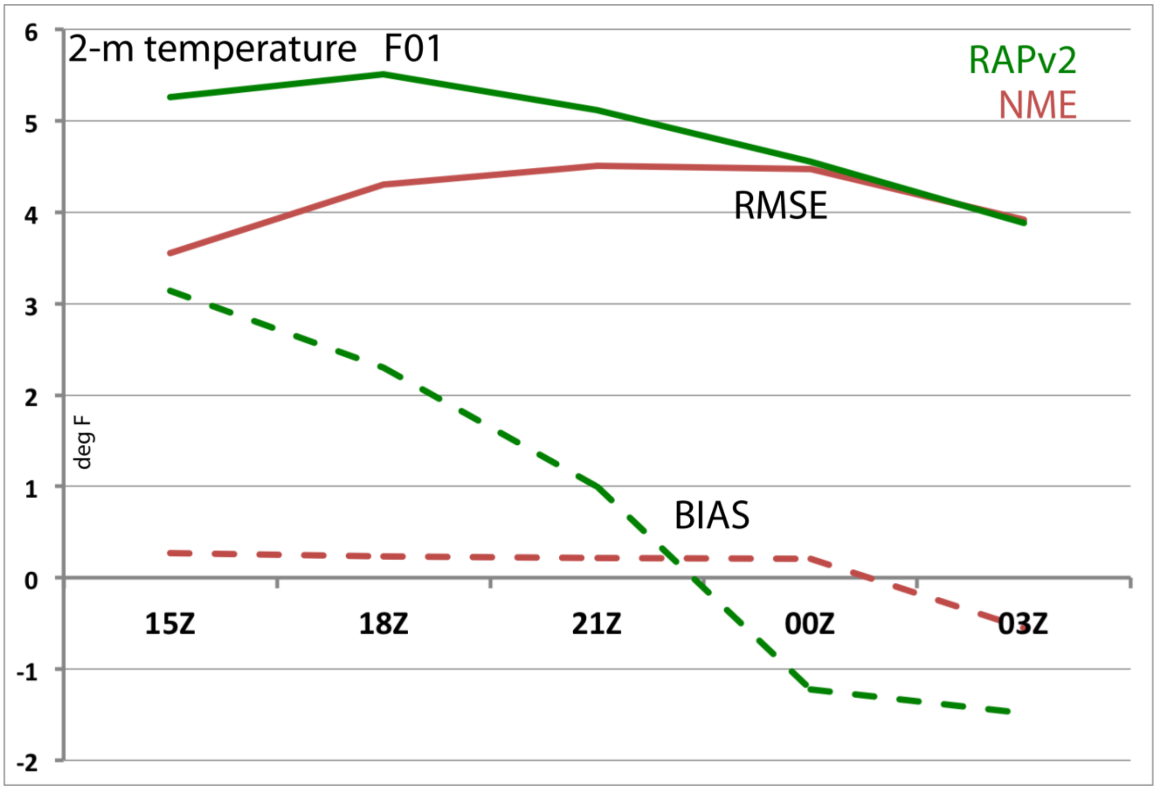


Figure 12. Cumulative RMSE (solid lines) and bias (MD; dashed lines) by valid time for one-hour forecasts of 2-m temperature for the RAPv2 (green) and the NME mean (red) over the mesoscale area of interest during SFE2013.

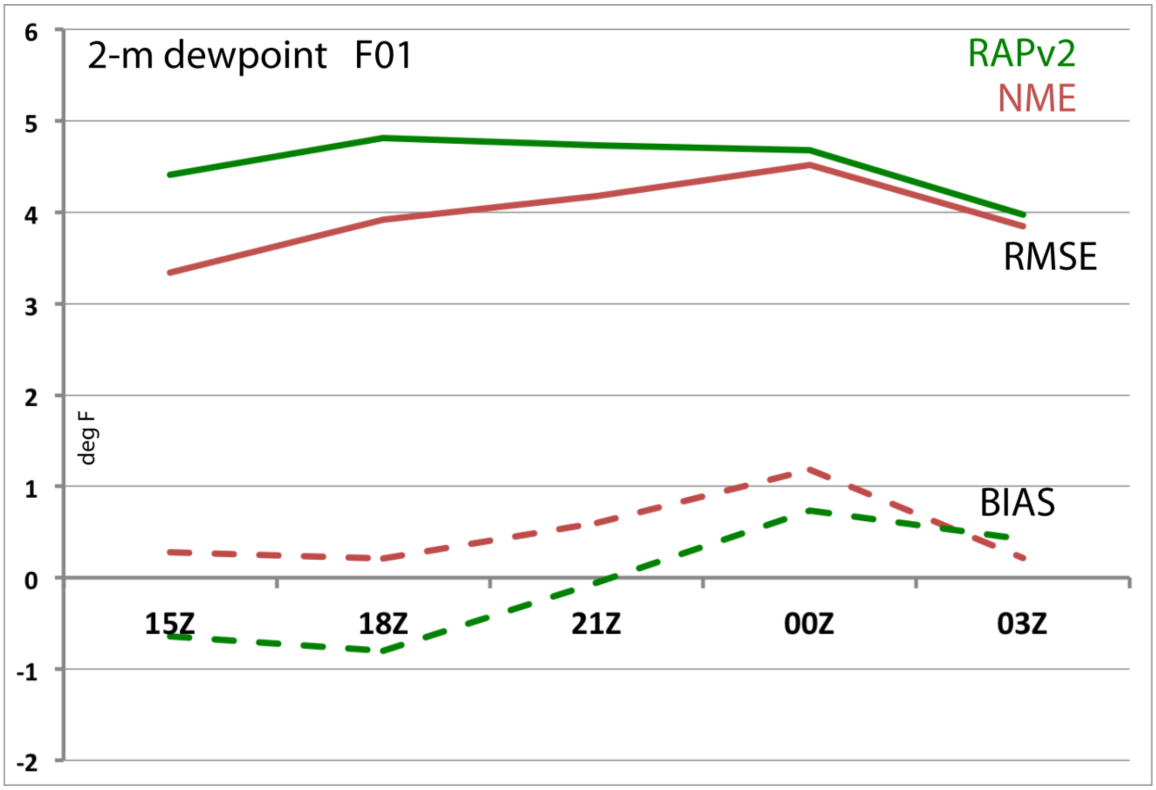


Figure 13. Same as Fig. 12, except for one-hour forecasts of 2-m dewpoint temperature.

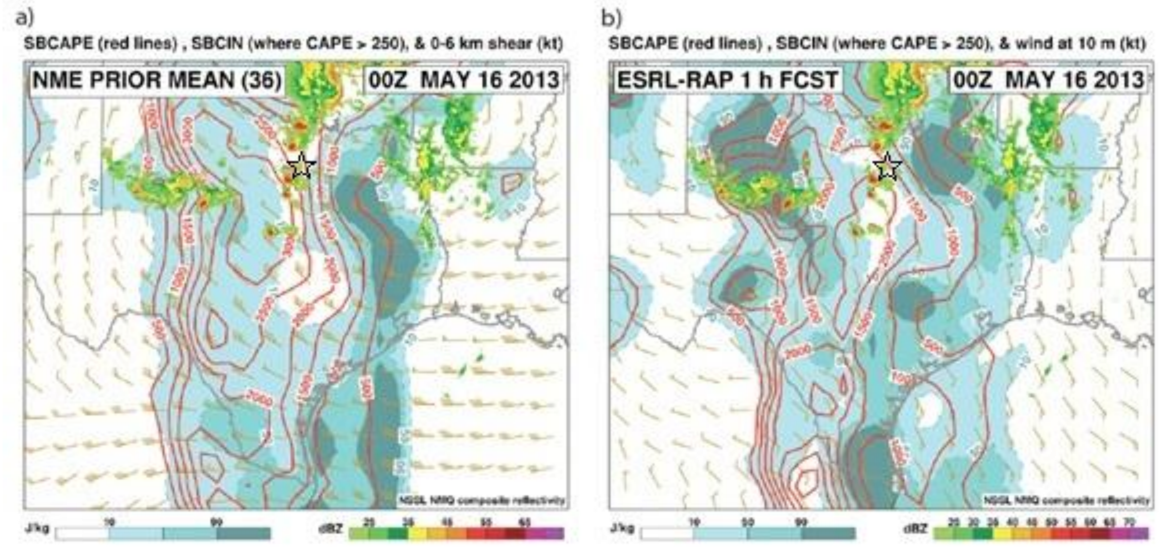

a)
SBCAPE (red lines) , SBCIN (where CAPE > 250), & 0-6 km shear (kt)
NME PRIOR MEAN (36)
00Z MAY 16 2013
NSSL NMQ composite reflectivity
J/kg
10
99
dBZ
25
35
45
55
65
b)
SBCAPE (red lines) , SBCIN (where CAPE > 250), & wind at 10 m (kt)
ESRL-RAP 1 h FCST
00Z MAY 16 2013
NSSL NMQ composite reflectivity
J/kg
10
50
99
dBZ
20 25 30 35 40 45 50 55 60 65 70

Figure 14. 1-h forecast of surface-based convective available 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.

c) 1200 UTC 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; comments in Table 4). 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.

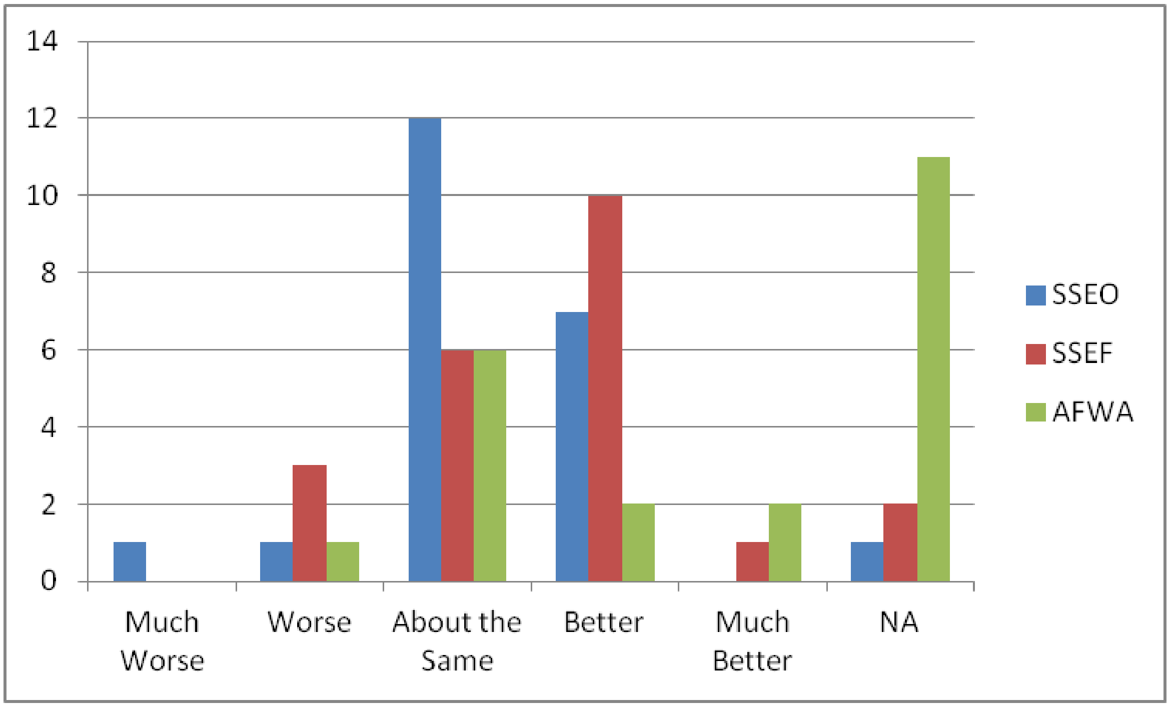


Figure 15. Subjective ratings assigned by participants to the 1200 UTC ensemble neighborhood probability forecasts of 1-km AGL reflectivity ≥40 dBZ relative to the same forecasts from the 0000 UTC ensembles (SSEO - blue, 8-member SSEF - red, and AFWA - green) valid during the 1300-0600 UTC forecast period.

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).

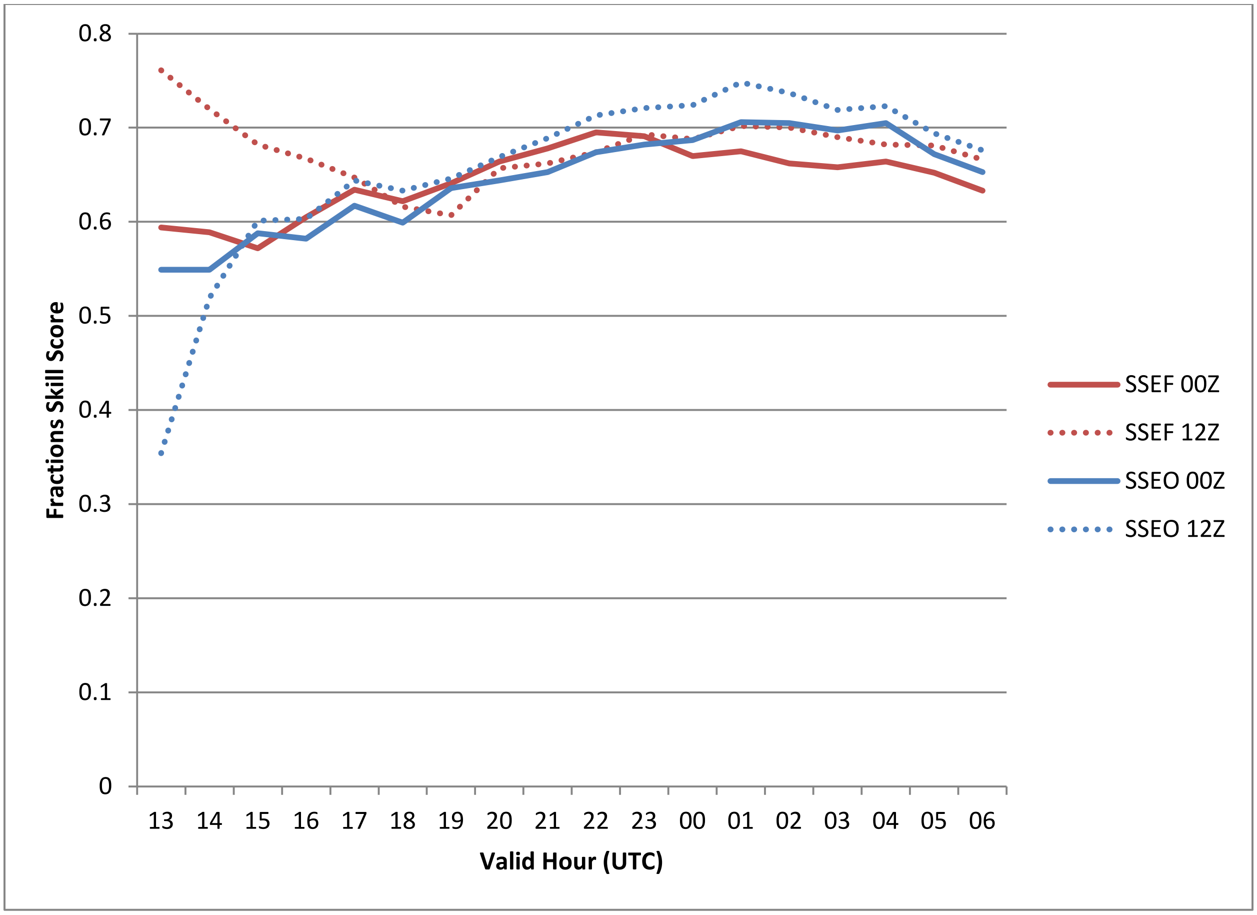


Figure 16. Accumulated fractions skill score by forecast hour for neighborhood probabilities of reflectivity ≥40 dBZ for the SSEF and SSEO.

Interestingly, the subjective ratings of the ensemble forecasts of hourly maximum storm-attribute fields (HMFs; Fig. 17; comments in Table 5) 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.

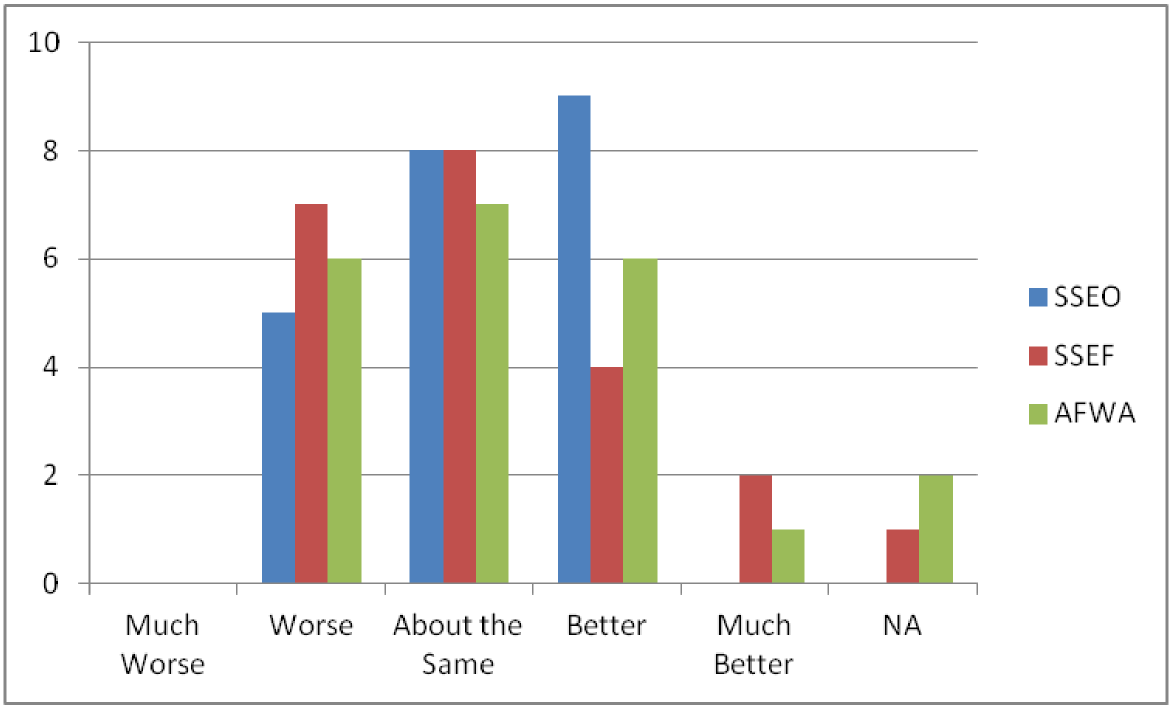


Figure 17. Subjective ratings assigned by participants to the 1200 UTC ensemble forecasts of hourly maximum storm-attribute fields relative to the same forecasts from the 0000 UTC ensembles (SSEO - blue, 8-member SSEF - red, and AFWA - green).

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.

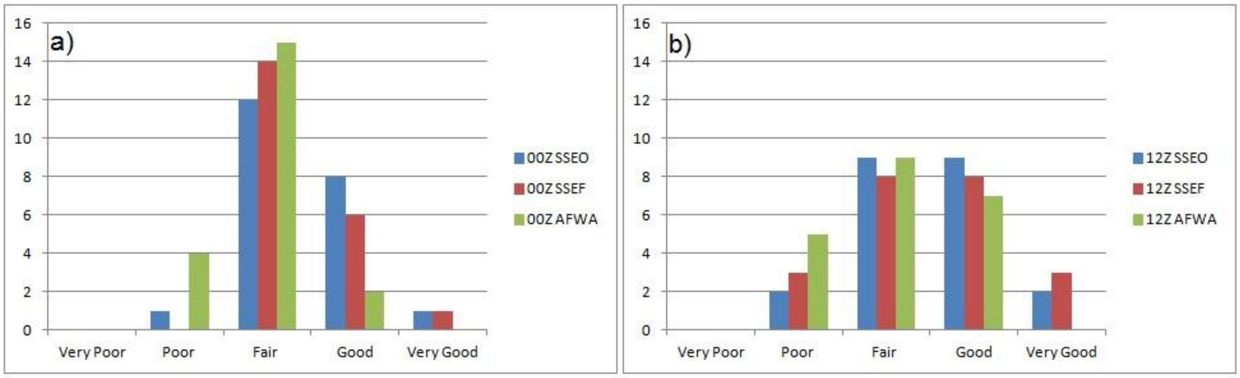


Figure 18. Subjective ratings assigned by participants to the a) 0000 UTC and b) 1200 UTC ensemble forecasts of hourly maximum storm-attribute fields on the usefulness to a severe weather forecaster (SSEO - blue, 8-member SSEF - red, and AFWA - green).

d) NSSL-WRF Comparisons

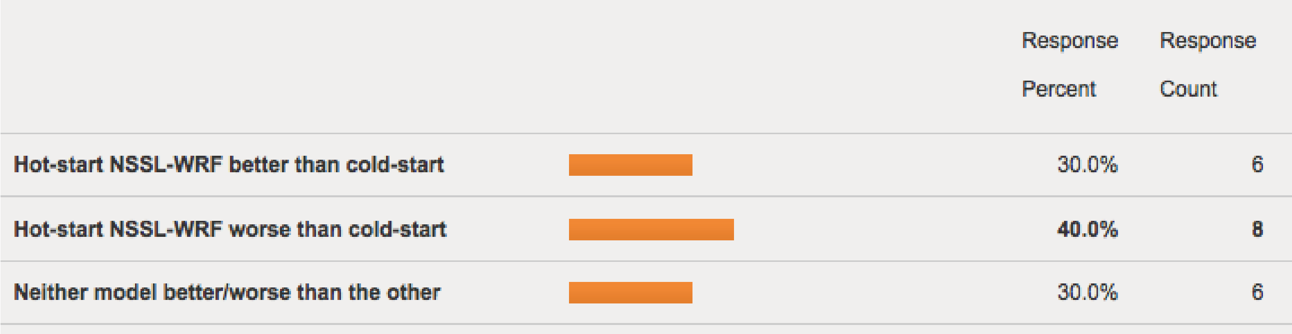
The cold-start NSSL-WRF was compared subjectively to three other convection-allowing runs during SFE2013: the hot-start NSSL-WRF, the 4.4 km UKMET UM, and the 2.2 km UKMET UM. The hot-start NSSL-WRF was configured identically to the cold-start NSSL-WRF run so that the impact of the NME analyses in initializing the forecasts could be evaluated. In addition, the NSSL-WRF was also compared to the two deterministic convection-allowing models run by the UKMET 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 UKMET 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?"

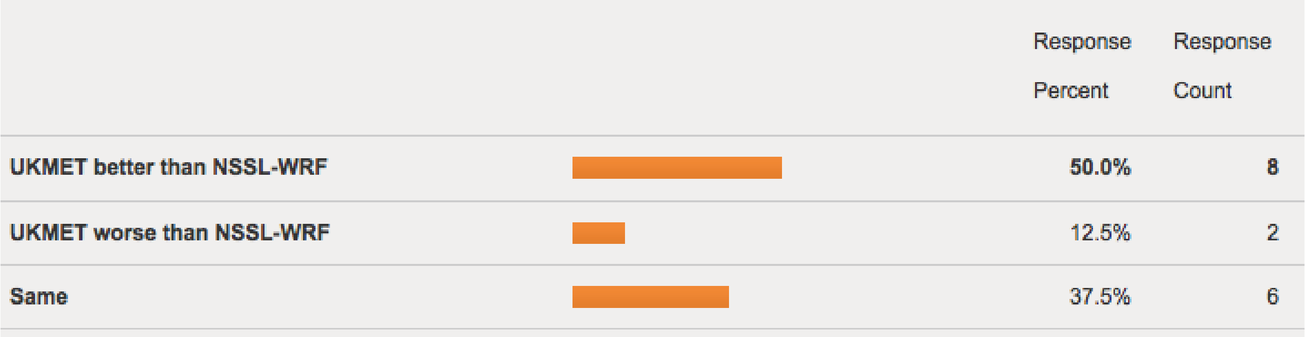
For item 1, there were a total of 20 responses, which are summarized in Fig. 19. Slightly more often (40% of the time), it was determined that the hot-start run was worse than the cold start. However, 30% of responses rated the hot start runs as better and another 30% rated the runs as not being better or worse. Given the small sample size, the results should be used with caution. The daily responses along with corresponding comments are provided in Table 6. 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.

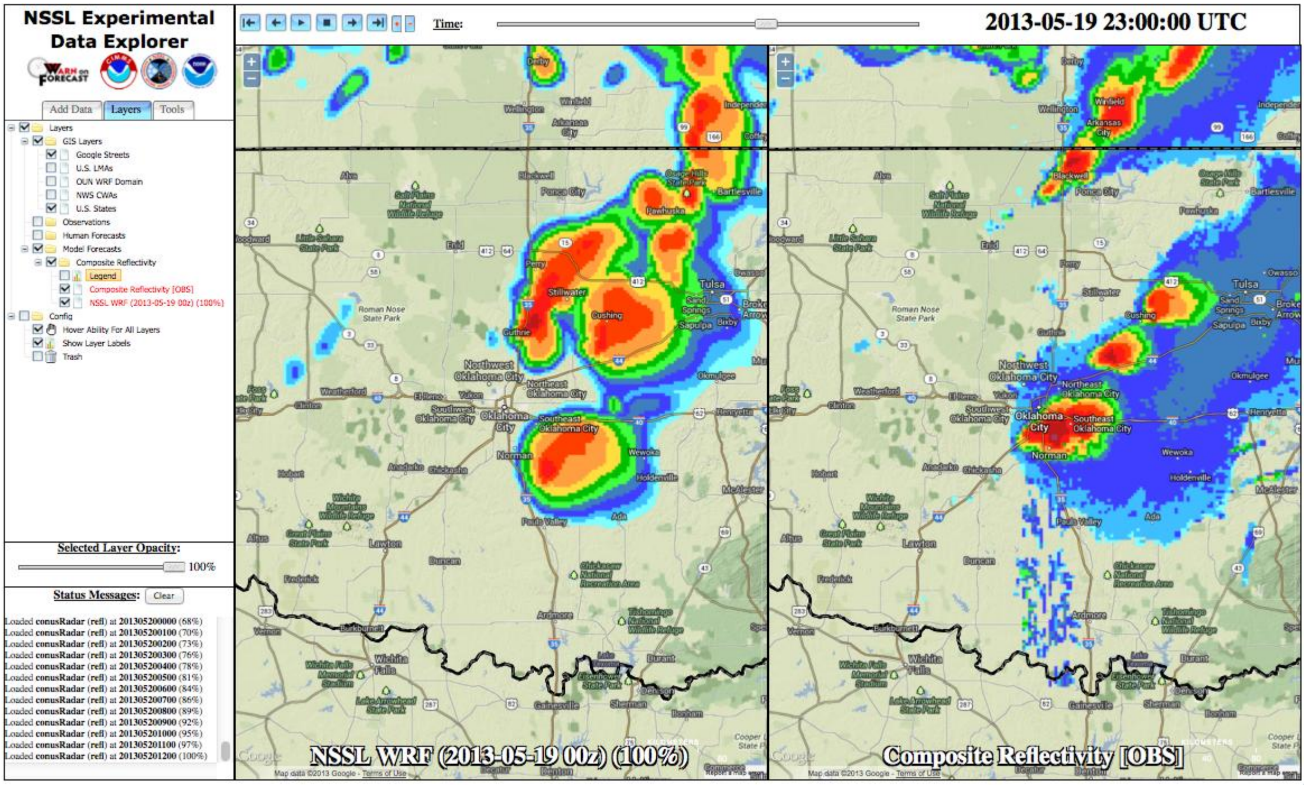
For item 2 there were a total of 16 responses. The majority of responses indicated that the 4.4-km UKMET forecast was better (50%) or the same (37.5%) relative to the cold-start NSSL-WRF, with only two cases (12.5%) in which the NSSL-WRF was rated as better than the UKMET (Fig. 20). For the cases in which UKMET performed better than NSSL-WRF there was a wide variety of reasons, which can be seen in Table 7. These reasons include the following: 1) the UKMET better depicting an MCV and related convection, 2) the UKMET suppressing convection in the correct locations, and 3) the UKMET better depicting timing and placement of convection. Perhaps one flaw that was noticed in the UKMET 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 UKMET model would depict large clusters of intense storms that never organized into coherent lines. It was speculated that the UKMET 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 UKMET. 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 UKMET versus NSSL-WRF comparisons.

Item 3 asked participants to comment on the utility of the NSSL Interactive Data Explorer (e.g., Fig. 21), and the responses are shown in Table 8. 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.



NSSL Experimental Data Explorer

Add Data | Layers | Tools

- ☑ Layers
  - ☑ GIS Layers
    - ☑ Google Streets
    - □ U.S. LMAs
    - □ OUN WRF Domain
    - □ NWS CWAs
    - ☑ U.S. States
  - □ Observations
  - □ Human Forecasts
  - ☑ Model Forecasts
    - ☑ Composite Reflectivity
      - □ Legend
      - ☑ Composite Reflectivity [OBS]
      - ☑ NSSL WRF (2013-05-19 00z) (100%)
- □ Config
  - ☑ Hover Ability For All Layers
  - ☑ Show Layer Labels
  - □ Trash

Selected Layer Opacity: 100%

Status Messages: Clear

```
Loaded conusRadar (refl) at 201305200000 (68%)
Loaded conusRadar (refl) at 201305200100 (70%)
Loaded conusRadar (refl) at 201305200200 (73%)
Loaded conusRadar (refl) at 201305200300 (76%)
Loaded conusRadar (refl) at 201305200400 (78%)
Loaded conusRadar (refl) at 201305200500 (81%)
Loaded conusRadar (refl) at 201305200600 (84%)
Loaded conusRadar (refl) at 201305200700 (86%)
Loaded conusRadar (refl) at 201305200800 (89%)
Loaded conusRadar (refl) at 201305200900 (92%)
Loaded conusRadar (refl) at 201305201000 (95%)
Loaded conusRadar (refl) at 201305201100 (97%)
Loaded conusRadar (refl) at 201305201200 (100%)
```

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).

e) Microphsyics 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 SSEF ensemble members with identical configurations except for their microphysical parameterization. During SFE2013, the following microphysics parameterizations were systemically 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)."*

Table 9 lists all the daily responses that were collected. 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.

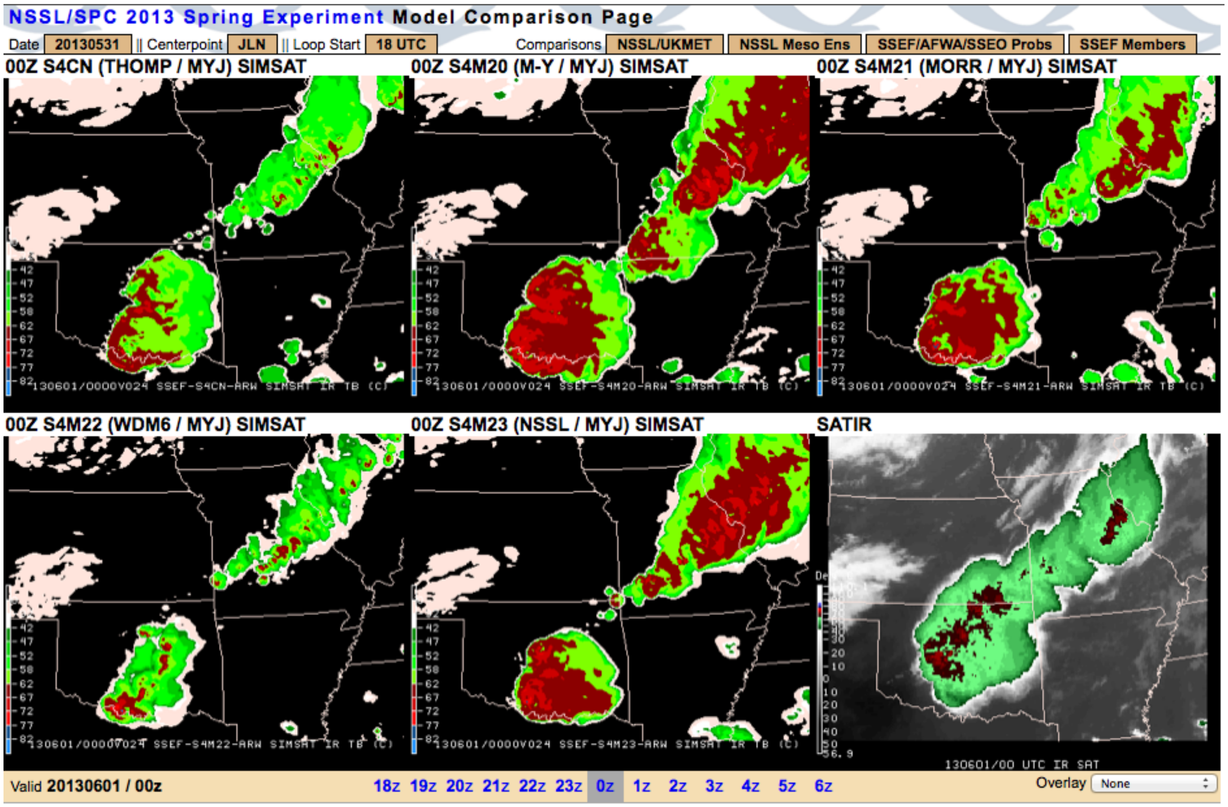

NSSL/SPC 2013 Spring Experiment Model Comparison Page
Date 20130531 || Centerpoint JLN || Loop Start 18 UTC
Comparisons NSSL/UKMET NSSL Meso Ens SSEF/AFWA/SSEO Probs SSEF Members
00Z S4CN (THOMP / MYJ) SIMSAT
00Z S4M20 (M-Y / MYJ) SIMSAT
00Z S4M21 (MORR / MYJ) SIMSAT
00Z S4M22 (WDM6 / MYJ) SIMSAT
00Z S4M23 (NSSL / MYJ) SIMSAT
SATIR
130601/0000V024 SSEF-S4CN-ARW SIMSAT IR TB (C)
130601/0000V024 SSEF-S4M20-ARW SIMSAT IR TB (C)
130601/0000V024 SSEF-S4M21-ARW SIMSAT IR TB (C)
130601/0000V024 SSEF-S4M22-ARW SIMSAT IR TB (C)
130601/0000V024 SSEF-S4M23-ARW SIMSAT IR TB (C)
130601/00 UTC IR SAT
-42 -47 -52 -58 -62 -67 -72 -77 -82
Valid 20130601 / 00z
18z 19z 20z 21z 22z 23z 0z 1z 2z 3z 4z 5z 6z
Overlay None

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.

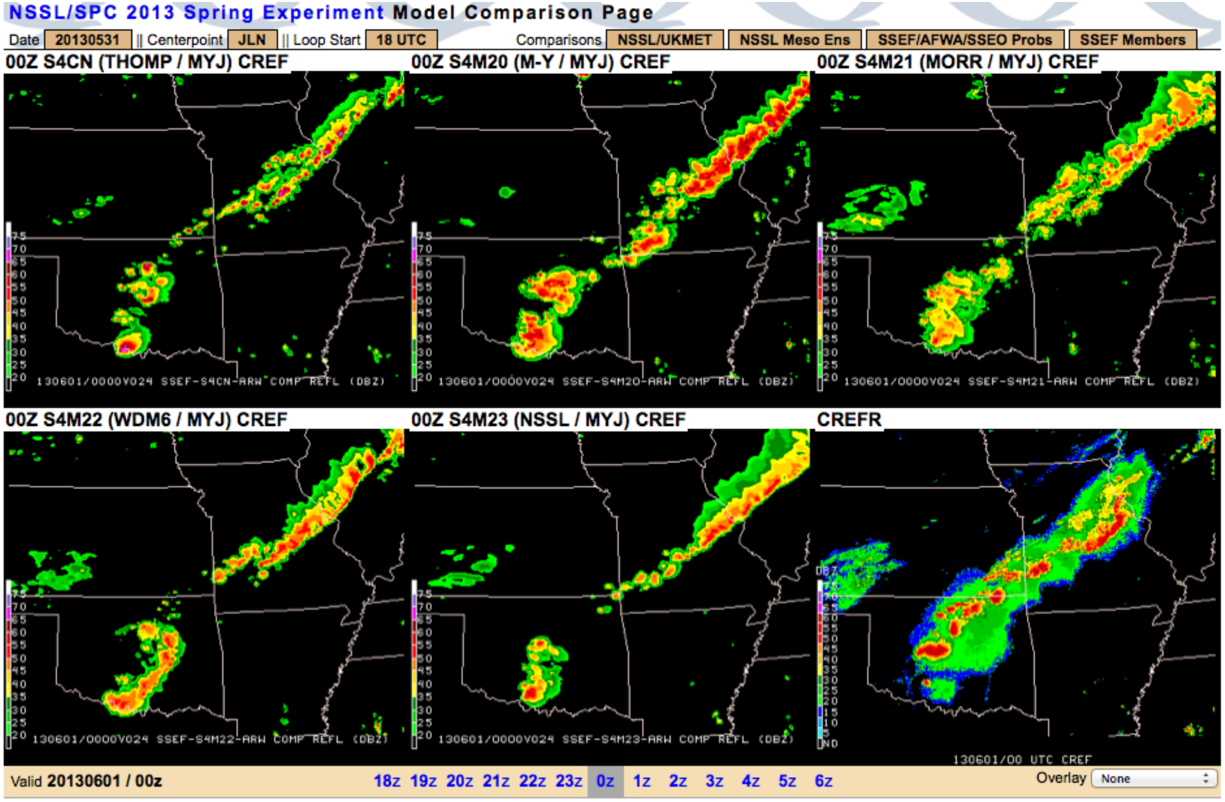

NSSL/SPC 2013 Spring Experiment Model Comparison Page
Date 20130531 || Centerpoint JLN || Loop Start 18 UTC
Comparisons NSSL/UKMET NSSL Meso Ens SSEF/AFWA/SSEO Probs SSEF Members
00Z S4CN (THOMP / MYJ) CREF
00Z S4M20 (M-Y / MYJ) CREF
00Z S4M21 (MORR / MYJ) CREF
00Z S4M22 (WDM6 / MYJ) CREF
00Z S4M23 (NSSL / MYJ) CREF
CREFR
130601/0000V024 SSEF-S4CN-ARW COMP REFL (DBZ)
130601/0000V024 SSEF-S4M20-ARW COMP REFL (DBZ)
130601/0000V024 SSEF-S4M21-ARW COMP REFL (DBZ)
130601/0000V024 SSEF-S4M22-ARW COMP REFL (DBZ)
130601/0000V024 SSEF-S4M23-ARW COMP REFL (DBZ)
130601/00 UTC CREF
75 70 65 60 55 50 45 40 35 30 25 20
Valid 20130601 / 00z
18z 19z 20z 21z 22z 23z 0z 1z 2z 3z 4z 5z 6z
Overlay None

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

4. Summary and Operational Impacts

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 findings from SFE2013 are relevant to near-term operations or are likely to have a direct or indirect impact on operations in the future:

- Travel restrictions within NOAA and forecaster vacancies at SPC prevented participation from NWS forecasters, which hindered the effectiveness of the O2R/R2O process during SFE2013.
- 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 UKMET 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 UKMET Office were vital to the success of SFE2013. In particular, Ming Xue (OU CAPS), Fanyou Kong (OU CAPS), Kevin Thomas (OU CAPS), Keith Brewster (OU CAPS), Yunheng Wang (OU CAPS), Evan Kuchera (AFWA), Scott Rentschler (AFWA), Steve Willington (UKMET), and Dan Suri (UKMET) were essential in generating and providing access to model forecasts examined on a daily basis. Finally, Steve Weiss (SPC) provided a helpful and thorough review that resulted in significant improvements throughout this report.

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 UKMET'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.

Table 1. Daily comments (when available) for the comparison of manual 3-h forecasts to the temporally disaggregated 3-h forecasts. The date refers to the day the forecasts were made.

7 May - West: West team has better placement of 15% in KS.
10 May - West: only slight modification, some good some decent.
14 May - West: 00-03 reduced false alarm but still miss the event. **East:** 18-21 - Same; 21-00 - Same; 00-03 - Same - false alarm area was increased in manual guidance, but effectively would not have mattered.
15 May - East: 21-00Z - increased probs where reports occurred, but added too much area into south TX where nothing happened. 00-03Z - adjusted 15% area to better encompass reports
16 May - East: 21-00Z - magnitudes were brought down and there were no reports so that was improvement. 00-

03Z - got rid of false alarm areas, but trim out area where there was a report so these cancelled each other out.
21 May - West: Slightly better in the 18-21Z period. The 15% was somewhat worse. **East:** For the 18-21Z period, the manual forecast added a 30% probability which verified. The automated forecast had no 30% probs. For the final two periods, improvements and detriments cancelled each other out.
22 May - East: 18-21: The 15% area was cut down in area, which made for less false alarm. 21-00: There were shifts but nothing that really made too much difference 00-03: Northward extension of the 5% area covered some additional reports.
23 May - West: For all periods, the manual forecasts do a better job of adapting to the evolution of the convection after initiation.
28 May - East: 18-21Z: Enlarged 5% area way too much.
31 May - West: For 18-21Z, a large false alarm area in the manual forecast, but the automated missed a significant severe report. For 21-00Z, manual forecast was overdone across northeastern OK For 00-03Z, both are fairly similar.
3 June - East: 18-21: added false alarm 21-00: added false alarm again 00-03: communicated there would be higher probs than automated, but higher probs weren't placed correctly
4 June - East: 18-21: exact same 21-00: added too much false alarm
5 June - East: 18-21: Added some false alarm area 21-00: Minor adjustments, but in the end not much difference

Table 2. Daily comments (when available) for the comparison of update forecasts. The date refers to the day the forecasts were made.

7 May - East 2100-0000 UTC first update forecast better than the previous forecast: They separated the 15% area.
8 May - West 2100-0000 UTC first update forecast about the same as the previous forecast: Extended 15% north into KS, but 30% extended too far south. **West 2100-0000 UTC final update forecast about the same as the previous forecast:** Some aspects better, some worse **West 0000-0300 UTC first update forecast worse than the previous forecast:** Expanded 15% erroneously to the west into SERN CO. **West 0000-0300 UTC final update forecast about the same as the previous forecast:** Some changes, but balanced out.
9 May - East 2100-0000 UTC first update forecast better than the previous forecast: Trimmed down some area where there were no reports
10 May - West 2100-0000 UTC first update forecast about the same as the previous forecast: some changes but either way the 30 is in between the clustered reports **West 2100-0000 UTC final update forecast about the same as the previous forecast:** few changes **West 0000-0300 UTC first update forecast about the same as the previous forecast:** few changes
13 May - East 2100-0000 UTC first update forecast about the same as the previous forecast: not much change
15 May - West 2100-0000 UTC first update forecast much better than the previous forecast: capturing the reports much better **West 2100-0000 UTC final update forecast better than the previous forecast:** good change **West 0000-0300 UTC first update forecast better than the previous forecast:** some cancellation of errors with trimming the 5 and moving the 30 to cover reports **West 0000-0300 UTC final update forecast about the same as the previous forecast:** axis changed slightly, but really it was pretty close with the exception of the extension of the 10 down to south texas. **East 0000-0300 UTC final update forecast better than the previous forecast:** Added a sig and 20% that verified quite well.
16 May - West 2100-0000 UTC first update forecast worse than the previous forecast: upped it to 15 and still missed warnings in 15 area **West 2100-0000 UTC final update forecast worse than the previous forecast:** expanded 10 and 15 areas

capturing warnings but no reports

East 2100-0000 UTC first update forecast about the same as the previous forecast: outlooks very similar

East 2100-0000 UTC final update forecast about the same as the previous forecast: outlooks very similar

21 May - West 2100-0000 UTC first update forecast worse than the previous forecast: The northward expansion of the 30% was worse.

West 2100-0000 UTC final update forecast worse than the previous forecast: Adding a 45% was a bad move into WRN TN.

West 0000-0300 UTC first update forecast worse than the previous forecast: Expansion of 45% was worse.

West 0000-0300 UTC final update forecast about the same as the previous forecast: Good trimming of 15% in AR, but poor addition of 45% into WRN TN.

East 0000-0300 UTC first update forecast much better than the previous forecast: Good refinement from the early forecast, did an excellent job capturing the higher probs.

22 May - East 2100-0000 UTC first update forecast about the same as the previous forecast: Both improvement and degradation cancelled each other out...

East 0000-0300 UTC final update forecast about the same as the previous forecast: Not really much change in the outlook

23 May - West 2100-0000 UTC first update forecast about the same as the previous forecast: 30% prob across SW Texas is a detriment.

West 2100-0000 UTC final update forecast much better than the previous forecast: Area of higher probs was correctly trimmed off on the northern part across the central TX panhandle.

West 0000-0300 UTC final update forecast much better than the previous forecast: Probs were refined well across the central TX panhandle mostly attributable to observational trends.

28 May - East 2100-0000 UTC first update forecast better than the previous forecast: The update shifted higher probs slightly more north and east capturing more of the observed reports

East 2100-0000 UTC final update forecast better than the previous forecast: slight changes resulted in capturing slightly more reports

East 0000-0300 UTC first update forecast about the same as the previous forecast: only slight changes were made

East 0000-0300 UTC final update forecast about the same as the previous forecast: only slight changes were made

31 May - West 2100-0000 UTC first update forecast better than the previous forecast: It became apparent that convection would develop further north, and the forecast was adjusted accordingly.

West 2100-0000 UTC final update forecast better than the previous forecast: False alarm area was decreased and higher probabilities were shifted slightly north

West 0000-0300 UTC first update forecast better than the previous forecast: 15% probs were expanded to account for the convective development across the St. Louis metro area.

West 0000-0300 UTC final update forecast worse than the previous forecast: Narrowing the higher probs across northeastern OK, southeastern KS, and southwestern MO was a mistake.

3 June - East 2100-0000 UTC first update forecast about the same as the previous forecast: Exactly the same

East 2100-0000 UTC final update forecast about the same as the previous forecast: Exactly the same

4 June - West 2100-0000 UTC first update forecast about the same as the previous forecast: Some improvements, but not enough to call it better.

5 June - West 2100-0000 UTC first update forecast worse than the previous forecast: They trimmed the 5% into small areas that didn't verify and there were reports in between 5% areas.

Table 3. Daily comments (when available) for the comparison of 1-hr forecasts of 2-m temperature, 2-m dewpoint, and instability from the NME to the RAPv2. The date refers to the day the forecasts were made.

6 May - NME mean about the same as the RAPv2 for 1-hr forecast of 2-m T: They were very similar.

NME mean better than the RAPv2 for 1-hr forecast of 2-m Td: The NME-mean was slightly more moist than SFCOA and in some regions of North and South Carolina there was less of a dry bias as indicated by the dot plots....

NME mean about the same as the RAPv2 for 1-hr forecast of instability: Biggest differences were offshore the coast of the outerbanks, where the NME-mean was more unstable. Also, big noticeable differences in SRH - probably based on co-location of convection.

7 May - NME mean better than the RAPv2 for 1-hr forecast of 2-m T: RAP was very warm relative (+4F) to observations over the domain. NME had a cool, moist bias in western OK (possible soil moisture issue) ; rap/nme data missing after 23Z

NME mean about the same as the RAPv2 for 1-hr forecast of 2-m Td: RAP too dry early then become very similar by 23Z; data void regions have bigger differences than NME

NME mean much better than the RAPv2 for 1-hr forecast of instability: SBCAPE: big differences in areas; look at LMN at 17 UTC with RAP has a bullseye of 1000 while NME agrees with obs cape of ~600. RAP OUN sounding at 00 UTC was better with 500 as opposed to 1000-1200 in NME.

8 May - NME mean much better than the RAPv2 for 1-hr forecast of 2-m T: NME much better over the domain with consistently much lower mean difference values. RAP cools down too much in presence of deep convection over wrn/cntrl OK in the 21-00z time frame. Mean difference values become closer through 00z but NME still better throughout.

NME mean about the same as the RAPv2 for 1-hr forecast of 2-m Td: NME mean missed development and extend of dry pocket across cntrl OK. RAPv2 showed moist bias in axis across cntrl OK. Both show dry bias either side of moist axis.

NME mean about the same as the RAPv2 for 1-hr forecast of instability: Both models forecast MLCAPE magnitudes too high compared to observed soundings. NME somewhat better with lower MLCAPE in areas with confirming soundings (LMN, OUN).

9 May - NME mean better than the RAPv2 for 1-hr forecast of 2-m T: Again, the RAP was much too warm in the early hours, then converged closer to the NME mean by evening. The ESRL RAP didn't stabilize the region behind the MCS in Texas enough, whereas the NME cooled it a bit too much. The only areas where the NME seems to struggle is with smaller areas of convection- there's a lag in getting the cold pools in.

NME mean better than the RAPv2 for 1-hr forecast of 2-m Td: The NME was better in the first few hours as the ESRL RAP was too moist in OK and too dry in the post frontal airmass in the TX panhandle. Otherwise they were similar beyond ~20Z.

NME mean about the same as the RAPv2 for 1-hr forecast of instability: No good raob sites to evaluate.

10 May - NME mean about the same as the RAPv2 for 1-hr forecast of 2-m T: starts off better then becomes a wash; multiple factors associated with convection. RAP heated up too fast with departing MCS and clearing through LA, convective response behind the supercell/line was too strong in the models in west texas.

NME mean better than the RAPv2 for 1-hr forecast of 2-m Td: change in dew point errors along north TX and OK around 20 UTC; mexico rapid increase in moisture with very large 20F differences compared to NME.

13 May - NME mean about the same as the RAPv2 for 1-hr forecast of 2-m T: early NME is good then transitions to about the same and then back to the RAP

NME mean about the same as the RAPv2 for 1-hr forecast of 2-m Td: pretty similar with this Montana moisture pool developed from parameterized convection: either tendencies of moisture or directly from precip. Which verified just not in the areal coverage sense.

NME mean about the same as the RAPv2 for 1-hr forecast of instability: nme was overdone for the Glasgow sounding ; just did comparison with the observed sounding and the spatial plot of sbcape.

14 May - NME mean about the same as the RAPv2 for 1-hr forecast of 2-m T: models heated up quickly then not enough after 17 UTC, RAP cools down too rapidly at 00z. Sea breeze issue throughout the day. domain wide cooler in nme then shifts at 21 z to differences maximized at the front in IA; 00z NME is much warmer domain wide

NME mean about the same as the RAPv2 for 1-hr forecast of 2-m Td: nme and rap are equally moist in central and eastern IA; almost domain wide excessive moisture in both

NME mean about the same as the RAPv2 for 1-hr forecast of instability: looking at DVN sounding in the core of the sbcape: observed 1800, NME has greater than 1500, RAP has greater than 1000.

15 May - NME mean better than the RAPv2 for 1-hr forecast of 2-m T: convection in the panhandle was absent in the NME but it's in the RAP. dallas area around 00 uTC was better in nme. where its convecting its worse in nme.

NME mean about the same as the RAPv2 for 1-hr forecast of 2-m Td: substantial dryline difference in south TEXAS , rap is too far east. and in the rap a cold pool is too moist. some very interesting differences in various places.

NME mean much better than the RAPv2 for 1-hr forecast of instability: sbcape: 2000+ in NME, RAP is slightly lower 1500-2000. looking at FWD sounding: NME was better with sbcape. profile was too cool at the sfc , none have the 500 inversion but do have a stability change, RAP is bit worse with the profile stp is much better in nme than rap and also scp. sig differences in SRH0-3km!

16 May - NME mean about the same as the RAPv2 for 1-hr forecast of 2-m T: nme struggles to produce cold pools in the convective areas due to the mean. discussing whether this approach of comparing ens mean to one member is appropriate. could choose closest member/representative.

NME mean about the same as the RAPv2 for 1-hr forecast of 2-m Td: pretty distinct moisture differences domain wide. but the diffs are relatively small scale. scale of differences is on par with the differences in observations...down in the noise.

20 May - NME mean about the same as the RAPv2 for 1-hr forecast of 2-m T: where convection is the NME is much warmer than the rap. this is using the nam to start.

NME mean about the same as the RAPv2 for 1-hr forecast of 2-m Td: similarly poor wrt obs. otherwise there were dryline (dewpoint gradients) differences.

21 May - NME mean about the same as the RAPv2 for 1-hr forecast of 2-m T: With the significant areal coverage of ongoing convection, the NME struggled getting the cold pools correct, which was a big factor in convective development later in the day. The warm sector across NERN TX was somewhat better during the afternoon in the NME.

NME mean about the same as the RAPv2 for 1-hr forecast of 2-m Td: In the warm sector, the NME was slightly better as noted by Jeremy. Overall, the RAP might have been better including where precipitation had occurred.

22 May - NME mean about the same as the RAPv2 for 1-hr forecast of 2-m Td: need big-to-small dot ratio

23 May - NME mean better than the RAPv2 for 1-hr forecast of 2-m T: RAP 2-m temps are way too warm over most of the domain for a large portion of the period.

NME mean better than the RAPv2 for 1-hr forecast of 2-m Td: RAP moves the dryline too far east too quickly. NME is better across OK and in the southern TX panhandle region around 21Z.

24 May - NME mean better than the RAPv2 for 1-hr forecast of 2-m T: Smaller differences across NE CO and SW NE where convection would eventually develop.

NME mean better than the RAPv2 for 1-hr forecast of 2-m Td: NME has much better representation of the dryline across eastern CO/western KS. RAP is too moist across southern NE, which could influence the convective evolution there. NME better across southeastern CO late in the period.

NME mean about the same as the RAPv2 for 1-hr forecast of instability: NME was better at DDC. ESRL-RAP was better at LBF.

31 May - NME mean comparison to RAPv2 for 1-hr forecast of 2-m T: Several forecast hours were missing from RAP. The RAP was much too warm across WRN OK. The NME was as well, but to a lesser extent. The RAP also had precipitation along the dryline too soon, which caused the model to be too cool for many stations.

3 June - NME mean about the same as the RAPv2 for 1-hr forecast of 2-m T: The NME is much better early on where the RAP is too hot, but the RAP is better in the region of the convection and ahead of the convective line after 00Z.

NME mean better than the RAPv2 for 1-hr forecast of 2-m Td: The NME was slightly better in areas where convection was expected in NWRN OK, where the RAP was too dry and the differences to obs were less in NME.

NME mean about the same as the RAPv2 for 1-hr forecast of instability: Both NME and RAP were very good for OUN. There were more differences for DDC.

4 June - NME mean about the same as the RAPv2 for 1-hr forecast of 2-m T: he NME was better ahead of the convective line in ERN NM, as the RAP had already convected ahead of the actual storms.

NME mean about the same as the RAPv2 for 1-hr forecast of 2-m Td: Similar to above, but the RAP was too

moist.
6 June - NME mean about the same as the RAPv2 for 1-hr forecast of 2-m T: Differences across southern TX at 18Z, NME is cooler than obs, whereas RAP is warmer. Both handle the cold pool behind the developing MCS across southeastern TX comparably. **NME mean better than the RAPv2 for 1-hr forecast of 2-m Td:** NME better across southeastern TX ahead of the developing MCS. RAP pools way too much moisture across the Big Bend area of TX. **NME mean about the same as the RAPv2 for 1-hr forecast of instability:** In regards to the DRT sounding, both the NME and RAP do not capture an inversion between 700-850 mb that most likely prevented convection in this area.

Table 4. Daily comments (when available) for the comparison of the reflectivity forecasts from the 1200 UTC ensembles to the 0000 UTC ensembles. The date refers to the model initialization time.

6 May - 1200 UTC SSEF about the same as the 0000 UTC SSEF: The placement was better in 00Z ensemble, but magnitudes were better in the 12Z. **1200 UTC AFWA about the same as the 0000 UTC AFWA:** The probabilities were overdone at the end of the period.
7 May - 1200 UTC SSEF better than the 0000 UTC SSEF: Forecast the gap correctly in the vicinity of the TX panhandle. **1200 UTC AFWA worse than the 0000 UTC AFWA:** The 12Z had too much confidence farther south in isolated convection, and better placement in the north.
8 May - 1200 UTC SSEF about the same as the 0000 UTC SSEF: Ensembles more similar to themselves than the observations. **1200 UTC SSEO better than the 0000 UTC SSEO:** Slightly better for the 12Z **1200 UTC AFWA much better than the 0000 UTC AFWA:** 12Z better extended farther south into OK.
9 May - 1200 UTC SSEF about the same as the 0000 UTC SSEF: lots of error cancellation across the domain. time series was the deciding factor **1200 UTC SSEO about the same as the 0000 UTC SSEO:** lots of error cancellation 12z better with ongoing, but 00z better with other random areas **1200 UTC AFWA about the same as the 0000 UTC AFWA:** looked really identical
10 May - 1200 UTC SSEF better than the 0000 UTC SSEF: 12z better early but still overdone west, and by late in the forecast the 00 UTC is better. FSS was not greatly better but captured enough of the ongoing events to be pretty good. **1200 UTC SSEO much worse than the 0000 UTC SSEO:** seems to completely miss the relevant convective episodes in location, timing and evolution. FSS says 12z much better but the overlap is dominant at 12z without coherent signals reflecting the event. **1200 UTC AFWA about the same as the 0000 UTC AFWA:** forecasts look much more like each other than the atmosphere; equally BAD
13 May - 1200 UTC SSEF better than the 0000 UTC SSEF: domain cut off; better in the area of the eastern MT region especially in the transition of the cluster to the line to the west which has a better depiction. **1200 UTC SSEO about the same as the 0000 UTC SSEO:** more like each other than obs. **1200 UTC AFWA better than the 0000 UTC AFWA:** overdone at 12 UTC; fss was saying higher FSS at 12 UTC with higher probs coverage: overfcst
15 May - 1200 UTC SSEF better than the 0000 UTC SSEF: central and southern tx way overdone in 12UTC but the areas were decent in North texas. 00z really latches on late and is really good as opposed to the 12 UTC **1200 UTC SSEO better than the 0000 UTC SSEO:** some aspects of 12z early are better, but 00z gets much better around 3z but this was not reflected in the FSS; both timing have differences that result in spatial locations east (12) and west (00). Just a little better
16 May - 1200 UTC SSEF worse than the 0000 UTC SSEF: 00z better, just matching the observed areas slightly better (timing)

1200 UTC SSEO about the same as the 0000 UTC SSEO: only eval until 23z. 00z run had a better chance to capture the NE CO convection
20 May - 1200 UTC SSEF about the same as the 0000 UTC SSEF: prob magnitude better in the 12z overall, though phase error was different. 12z was aggressive early and off on initiation but recovered. **1200 UTC SSEO better than the 0000 UTC SSEO:** 12z prob magnitude better; have nearly the same phase. biggest difference is early.
21 May - 1200 UTC SSEF much better than the 0000 UTC SSEF: The 12Z initialized storms well and carried on that benefit throughout the forecast. **1200 UTC SSEO better than the 0000 UTC SSEO:** The 12Z had better timing of convective line across NRN TX. **1200 UTC AFWA better than the 0000 UTC AFWA:** The 12Z had slightly better timing of the convective line and less false alarm across LA.
23 May - 1200 UTC SSEF better than the 0000 UTC SSEF: 12Z is better due to the depiction of initiation along the outflow boundary across NW Texas. **1200 UTC SSEO about the same as the 0000 UTC SSEO:** Early on 12z takes awhile to spin up, but the two are fairly similar for the remainder of the forecast. 12Z is earlier with initiation across the TX Panhandle, which was an improvement.
24 May - 1200 UTC SSEF worse than the 0000 UTC SSEF: FSS is higher through much of the forecast period in the 12Z when compared to the 00Z. However, 12Z produces higher probabilities across western KS, but it does not verify. There were no observed reflectivity across western KS. **1200 UTC SSEO worse than the 0000 UTC SSEO:** 00Z depicts the two higher probability areas which verified better when compared to the 12Z.
30 May - 1200 UTC SSEO worse than the 0000 UTC SSEO: FSS didn't agree with our subjective impression of 00z better early, 12z better much later.
31 May - 1200 UTC SSEF about the same as the 0000 UTC SSEF: 12Z is generally better, at later periods it has higher probabilities along the axis of convection. 00Z under-does the probabilities. **1200 UTC SSEO better than the 0000 UTC SSEO:** 12Z has higher probabilities across OK, which matches the observed probabilities better. **1200 UTC AFWA much better than the 0000 UTC AFWA:** 12Z is a much better match to the observed probabilities.
3 June - 1200 UTC SSEF better than the 0000 UTC SSEF: Evolution between 00Z and 12Z more like each other than reality, but 12Z a bit better. **1200 UTC SSEO about the same as the 0000 UTC SSEO:** The skill scores were better for 00Z, but too slow in initiation.
4 June - 1200 UTC SSEF better than the 0000 UTC SSEF: The 12Z did better during the critical 00-03Z period. **1200 UTC SSEO about the same as the 0000 UTC SSEO:** Both perform poorly

Table 5. Daily comments (when available) for the comparison of the HMF forecasts from the 1200 UTC ensembles to the 0000 UTC ensembles. The date refers to the model initialization time.

6 May - 1200 UTC SSEF better than the 0000 UTC SSEF: The 12Z was slightly better with higher probabilities closer to observed reports. **1200 UTC SSEO worse than the 0000 UTC SSEO:** The 1200 UTC SSEO UH probs were lower and orientation/concentration of updraft speeds were worse.
7 May - 1200 UTC SSEF better than the 0000 UTC SSEF: Reduced false alarm and better gapping in the south. **1200 UTC AFWA worse than the 0000 UTC AFWA:** High probs in TX overdone.
8 May - 1200 UTC AFWA much better than the 0000 UTC AFWA: 0000 UTC AFWA did not extend probs far enough south and held on too long in KS.

9 May - 1200 UTC SSEF worse than the 0000 UTC SSEF: ongoing cluster neighborhood probs of uh was way overdone **1200 UTC SSEO worse than the 0000 UTC SSEO:** ongoing cluster neighborhood probs of uh was overdone **1200 UTC AFWA worse than the 0000 UTC AFWA:** ditto; lesson learned that agreement in guidance doesn't mean much in this case
10 May - 1200 UTC SSEF worse than the 0000 UTC SSEF: 00z better overlay with reports while 12z was too far south. **1200 UTC SSEO worse than the 0000 UTC SSEO:** 12z had one too many areas of UH probs **1200 UTC AFWA about the same as the 0000 UTC AFWA:** similarly bad. Looked like the signals for separate events were conflated with each other.
13 May - 1200 UTC SSEF about the same as the 0000 UTC SSEF: looked at wind and uh, relying on wind. **1200 UTC AFWA better than the 0000 UTC AFWA:** latter part 12z is better, more reports are in attention-getting maxima
14 May - 1200 UTC SSEO better than the 0000 UTC SSEO: phase error but covers more report area in WI
15 May - 1200 UTC SSEF about the same as the 0000 UTC SSEF: 12z favored early, 00z was favored late. 12z ensemble was really awesome at 00 UTC **1200 UTC SSEO better than the 0000 UTC SSEO:** better correspondence of reports at 00z, both do well at 03z **1200 UTC AFWA about the same as the 0000 UTC AFWA:** similar areas
16 May - 1200 UTC SSEF about the same as the 0000 UTC SSEF: updraft/wind speed much better than UH to eval this, fair ratings above are generous **1200 UTC SSEO about the same as the 0000 UTC SSEO:** higher probs in 12z but overall the same **1200 UTC AFWA about the same as the 0000 UTC AFWA:** 12z is much higher probs but areas covered are slightly larger
21 May - 1200 UTC SSEO better than the 0000 UTC SSEO: The 12Z had better timing of wind threat. **1200 UTC AFWA better than the 0000 UTC AFWA:** Slightly better for the 12Z with timing.
22 May - 1200 UTC SSEF about the same as the 0000 UTC SSEF: it's a wash, some areas are good in each model time periods, and they are not synced. **1200 UTC SSEO about the same as the 0000 UTC SSEO:** it's a wash, some areas are good in each model time periods, and they are not synced. **1200 UTC AFWA about the same as the 0000 UTC AFWA:** it's a wash, some areas are good in each model time periods, and they are not synced.
23 May - 1200 UTC SSEO worse than the 0000 UTC SSEO: 1200 UTC SSEO has too high of a signal across southwest OK.
31 May - 1200 UTC SSEF about the same as the 0000 UTC SSEF: Both forecasts have same general pattern, differences are small. **1200 UTC SSEO about the same as the 0000 UTC SSEO:** Both forecasts have same general pattern, differences are small. **1200 UTC AFWA better than the 0000 UTC AFWA:** 12Z forecast captures higher probabilities across central/northeastern OK during the 21-00Z time frame and across central OK in the 00-03Z time frame.
3 June - 1200 UTC SSEO better than the 0000 UTC SSEO: Jeff liked the more northern solution for the 00Z for watch purposes (i.e., not extending into TX).
4 June - 1200 UTC AFWA worse than the 0000 UTC AFWA: The WRN OK 100% probability was a bad distraction.

Table 6. Daily responses for the NSSL-WRF hot versus cold start comparisons. The date refers to the model initialization time.

6 May - Neither model better/worse than the other: Neither forecast was very good. Both had the general areas of precipitation correct on larger scales associated with the upper low, but It was hard to ascertain any significant differences in the smaller scale details between the two.

7 May - Neither model better/worse than the other: Timing of initiation around 22Z handled well in both simulations. HS concentrated a few storms in western KS similar to what was observed while the regular NSSl-WRF had too many storms extending south into OK and TX panhandles. Examining 2m TDs, the regular NSSL-WRF better handled a dryline moving from E. CO to W. KS. Both models didn't depict the storms moving into east-central KS later in the period very well.
13 May - Hot-start NSSL-WRF worse than cold-start: Hot start is earlier with initiation over wrn/central MT. Cold start did a great job locating two separate areas of convection in MT. Convection better placed in cold start.
14 May - Hot-start NSSL-WRF better than cold-start: cold start created too strong of storms in IA and the hot start only showed meager cells (which is what happened). Hot start also showed some cells in WI, while the cold start did not show anything (and there ended up being high based storms in WI with some wind damage). Hot start has slightly warmer sfc temperatures than cold start, but both under did warming. Hot start was better on dewpoint forecasts than cold start, though both were too high! as a result both greater overestimated SBCAPE on both models with slightly lower values on hot start. 10M WINDS...
15 May - Hot-start NSSL-WRF better than cold-start: hot start better with circulation/MCV over nrn Tx (though located in SWRN OK) and did better with much less storms in srn TX (which no storms occurred). Also storms on cold start in ERN TX,during the afternoon, were greatly exaggerated in intensity than storms on the hot start.
16 May - Hot-start NSSL-WRF better than cold-start: It handled storm development in NEB early on and has area of convection with SWRN NEB MCV and NERN NEB MCV well depicted.
19 May - Hot-start NSSL-WRF worse than cold-start: Cold start made one of the best convective forecasts ever with 2 supercells in metro (N OKC) and near Norman and pretty much predicted when, where and how many storms would develop.
20 May - Hot-start NSSL-WRF worse than cold-start: Hot start too slow initiating storms in TX. Hot start too far east with convective development in central/srn OK.
21 May - Hot-start NSSL-WRF better than cold-start: Hot start had a better representation of convective evolution during the daylight hours of May 21st. The squall line was too slow and too intense compared to reality, partly due to the fact that both models did not handle the cold and more stable air mass behind morning MCS.
22 May - Hot-start NSSL-WRF worse than cold-start: over the nern US eahc model similarly bad with each having having better/worse spots than the other. Hot start much better with overnight convection in OK than cold start.
23 May - Hot-start NSSL-WRF worse than cold-start: cold start had a great forecast on elevated convection of OK, while hot start totally missed elevated storm and consequent significant/strong outflow. Hot start also way over forecast convection across OK (huge storms too) due to no outflow boundary from morning storms. cold start handled storms developing near boundary in west TX the best.
24 May - Neither model better/worse than the other: Both runs missed linear MCS in NEB and each was better/worse than the other at different times.
28 May - Hot-start NSSL-WRF worse than cold-start: significantly better
29 May - Neither model better/worse than the other: both had right/wrong solutions to reality, depending on time and location.
30 May - Hot-start NSSL-WRF worse than cold-start: Hot start never developed convection over OK like cold start. More sharply defined dryline in dewpoint field in cold start. hot start dryline in OK gradient was quite broad and did not fit reality.
31 May - Hot-start NSSL-WRF worse than cold-start: Neither models handled the evolution of convection very well, but the cold-start had the timing better and southward extend of convection associated with a southward moving cold front very late in the forecast period.
3 June - Neither model better/worse than the other: minor differences
4 June - Neither model better/worse than the other: Neither was any good and provided bad guidance.
5 June - Hot-start NSSL-WRF better than cold-start: Hot start handled storms/linear system ewd across OK/KS Tue night/Wed morning and even forecast elevated convection that developed ahead of approaching upper wave.

However, both starts did poorly with convective evolution Wed afternoon/evening.
6 June - Hot-start NSSL-WRF better than cold-start: hot start got the evolution of mcv/ convection early part of period across nrn Tx. cold start was better later on...but hot start better for a longer period of time during this forecast period.

Table 7: Daily responses for the UKMET versus NSSL-WRF comparisons. The date refers to the model initialization time.

13 May – Same: n/a
14 May – Same: No convection UKMET, while cold start did, but cold start blew up a couple of big storms, which did not occur, so felt they were about the same.
15 May – UKMET better than NSSL-WRF: Much better at picking on MVC moving enewd through srn OK and develops discrete very intense storms in NWRN TX than NSSL. Actually a very good forecast. The 2km UKMET initiates storms further west and earlier than 4 km UKMET near dryline. 4 KM was slightly better than 2 KM for this particular case in location and evolution of severe convection in NRN TX.
16 May – Same: Better in different locations, but overall the differences were small. 2 km vs 4 km...very similar
20 May - UKMET better than NSSL-WRF: UKMET handled frontal convection in KS better than WRF. UKMET a little slow to develop convection in TX late afternoon,. UKMET has too much convection in MO overnight. UKMET does an excellent job developing e-w band of convection across central-ern OK overnight.
21 May - UKMET better than NSSL-WRF: UKMET handled southward and eastward motion of initial MCS across AR and SRN OK better than and location of second linear MCS better than NSSL (though both were slower than forecast).
22 May - UKMET better than NSSL-WRF: Overall UKMET captured overall convective evolution better than NSSL WRF.
23 May - UKMET worse than NSSL-WRF: UKMET did not seem to pick up on morning convective outflow boundary. Developed storms in the right location over the TX panhandle, but with no outflow boundary, moved storms newd into OK instead of propagating them swd through wrn TX.
27 May – Same: UKMET did better with overnight MCS moving ewd through NEb, but over forecast convection in central plains prior to that time.
28 May - UKMET better than NSSL-WRF: UKMET did much better in generating less convection across central wrn/central OK and also developed severe convection in correct orientation/location. 2km UKMET slightly better than 4 km UKMET late in period as it has slightly less FAR in NEB.
29 May – Same: UKMET did reall well with development/ evolution of bow in OK, but did not handle convection as well in KS. Each model had good/bad points that in the end made them the equally good (or bad). 2km vs 4 km - 2 km more realistic in echo shapes and better than 4 km
30 May - UKMET better than NSSL-WRF: UKMET handled convection in OK better than NSSL (though convective development was maybe an hour slow to develop in OK). Also UKMET handled overnight convection moving ewd through KS/MO better than NSSL. 2km better than 4 km in location and structure of storms. Excellent forecast!
3 June - UKMET better than NSSL-WRF: UKMET did a much better job of moving evening convection ewd across nrn ok (while the cold starts drove it ssewd through wrn OK/nrn TX). UKMET also caught redevelopment of convection in wrn OK around 06z and moved it into central OK by 12z (which is what happened). 2km vs 4 km UKMET...2 KM a little better after midnight...but about the same prior to that time.
4 June - UKMET worse than NSSL-WRF: After 06z, UKMET developed e-w band of convection across KS instead of linear MCS moving e-w oriented band of convection. However, UKMET Handles afternoon convection/development in CO sewd into nwrn TX much better than WRF. 2KM VS 4 KM....no big difference between the two.
5 June – Same: UKMET better than cold start Tue night/Wed morning and about the same thereafter. UKMET and

NSSL hot start very similar. 2km vs 4km... very similar overall looking pattern.
6 June - UKMET better than NSSL-WRF: Handled MCV/convection evolution better than cold start...though 2KM UKMET was too slow with ewd propagation of linear MCS and cold pool. 2km vs 4 km were about the same

Table 8. Comments regarding the utility of the NSSL Interactive Data Explorer.

"Side by side is always good"!
Need multi-panel displays...need better labels
More clear time labels on plots
Looking at any point and look at sounding. Add 500-700 mb lapse rates to data fields.
Like 4 panel display
great
Great
Incorporate more ensemble information into explorer.
Seems useful, would like to use it for other parameters.
Awesome.
Excellent tool for evaluating model differences

Table 9. Responses collected from the microphysics evaluations conducting during the 2013 Spring Forecasting Experiment. The date refers to the model initialization time.

6 May: M-Y has more higher reflectivities. NSSL has more 35+ dBZ areas than the others. Morrison was pretty hot in the higher reflectivity areas. M-Y clearly had more coverage in the cold cloud tops, followed by Morrison. WDM6 clearly had too few cloud top coverage. It's hard to compare to observations given the differences in color scales so we just did relative differences. For the Thompson/Thompson Mod comparison, the reflectivities were essentially the same outside of small differences. Same thing with the simulated IR. Maybe this is because of the weak instability/cold core set up with small, transient convective cores?
7 May: morrison overdoes pcpn in nwrn ok...all over did convection in tx panhandle...none had enough storms in wrn ks.... ref...ted's scheme has more intense cores...wdm convection goes away faster than other schemes... hourly max CR...TEDS..MJY...AND MORRISON TOO HOT with initial elevated convection...really high values in MJY SIMULATED SATELLITE...myj has much expansive CI shield with elevated convection...wdm6 has least extensive convection anvils compared to other schemes...actual colder storm tops move sewd, while the schemes move colder tops mostly ewd. wdm6 is terrible with convection clouds.
8 May: sb cape...all models too high on instability....wdm6 produces giant cold pools and wipes out instability across much of wrn/central ok.. Thompson not nearly as aggressive with storm outlfow's. MYJ warmest cold pools...Morrison coolest cloud tops. Morrison best depiction of storm evolution (over central OK and near dryline far ern TX Panhandle). simulated satellite...wdm6 warmest and least amount of anvil (despite very active convection). Hot/cold start comparison...hot start did not develop convection in central OK like cold start did cold start showed boundary across central ok...while hot start had it much further nwd.
9 May: simulated satellite little difference between Thompson and Thompson modified. Storms a little slower to go in mod Thompson in north TX. 2M TEMPERATURES....small differences in swrn ok where the regular Thompson warmed up more than the modified, though quite slight. hot/cold comparisons...
10 May: CREF...MORRISON/MYJ - overemphasized storms oriented e-w instead of n-s....wdm has weakest convection. especially after 00z. Temp....all forecasts are a little too warm in central TX area compared to observed. SB CAPE...all forecasts too unstable central TX ewd into LA compared to observed....wdm6 has more expansive stable layer (cold pool) than other schemes.
13 May: dewpoints 15-20 degrees lower in s central mt than most schemes...with WDM 6 having the highest dewpoints (50-55 compared to 30s in reality). BREF...Storm coverage less (and smaller in size) in WDM6 than other schemes. MYJ and NSSL had too high of reflectivities. hourly max updraft speed....NSSL scheme has fewer updrafts

(not obvious looking at reflectivity). simulated satellite...MYJ has coldest tops of all schemes....Spurious storm in E central MT had coldest and most expansive anvil tops (storm occurred in all schemes, though storm did not in reality).
15 May: MY overdid storms in south TX compared to others (which had less storms in the area that none occurred)...MY has bigger cores and very little in the way of stratiform/weaker echoes. all schemes performed poorly...though the WDM6 did a good job with evolving storms into a similar linear looking system from after 02z. simulated satellite...WDM6 does not have enough cold tops, while M-Y is too expansive. SBCAPE....Much more higher instability on NSSL scheme than other models, though temperatures and dewpoints similar. modified Thompson showing stronger instability and stronger storms than non-modified Thompson in NW TX near dryline.
16 May: CREF......all missed developing convection in nern CO. All missed convection with MCV moving ewd through srn NEB. Simulated satellite....Cloud tops in NERN NEB were much warmer on all schemes than reality due to much stronger storms associated with MVC that was not depicted in the schemes. SBCAPE....Appears that the schemes sbcape was at least 1000 j/kg more than reality in most locations. Modified Thompson and Thompson appeared to be the same...both were bad!
20 May: WDM6 and NSSL microphysics too slow to develop convection. Morrison and M-Y had large hook echoes with storms in srn OK and central TX that verified extremely well(with tornado events at 00z (though M-Y reflectivities were overdone). Simulated satellite...Morrison is the best overall. Thompson/WDM6 have considerably less high level cirrus than other schemes and reality.
21 May: Thompson has best depiction of actual pcpn evolution...though it was misplaced too far west with initial MCS. All schemes were too slow were ewd movement of severe squall line that moved ewd from TX to MS. Simulated satellite....Thompson/Morrison look a little better with colder cloud tops.
22 May: messy days...storms everywhere. Not good correlation with storm locations/mode from various models and what actually happened. MY too hot...WDM6 too cold with depicted reflectivities. SBCAPE...NSSL WRF has more SBCAPE and less convection. different representation of instability depending on storm location and coverage from scheme to scheme.
23 May: MY echoes are too hot.weak. Weakly forced situation with outflow boundaries primary source for convection...resulting in different development and evolution. NSSL did better job of developing storms in SRN TX panhandle and evolving it swd. NSSL scheme was the hottest with SBCAPE. Subtle differences in outflow locations from morning convection. Hourly max updraft much better in NSSL scheme with both location and very high localized values. Once again.NSSL cold tops very similar to what really happened....especially first 3-4 hours after convective initiation. Excellent cirrus top forecast by NSSL scheme! All schemes were cooler behind the outflow boundary than happened in reality. Thompson vs modified Thompson...SBCAPE greater on modified Thompson than Thompson...especially after 00z.
27 May: wdm6 is worse and handles convective development poorly...Morrison/Thompson and NSSL best depict evolution. simulated satellite...Thompson does best with extent of cold cloud temperatures and extent. NSSL's cold tops are too extensive compared to reality and other models.
28 May: all of them overdid convection in central and WRN OK. modified thompson satellite...no differences between the 2 schemes...CAPE fields....very little difference between the 2 schemes as well as the temp/dewpoint.
30 May: Thompson killed storms off prematurely. Lots of varied solutions due to cold pool issues from morning storms. simulated satellite...MY and NSSL had very cold cloud tops that moved south into N TX and did not occur. The other schemes did not do this.
31 May: All schemes except for NSSL were 1 h early initiating convection in OK. All schemes missed backbuilding convection that persisted until 10Z into central OK. Large MCS in IL with clear transition from heavy to light to stratiform precip - none of the schemes can predict the transition region.
3 June: NSSL has stronger looking cores than other schemes. All schemes too fast moving convection sewd out of OK panhandle, though Thompson scheme was the closest and best placed throughout the forecast period. satellite...NSSL way too cold and expansive cloud tops (and in the wrong place). Thompson handled the redevelopment in NWRN OK around 05/06 the best. temperatures...WDM6 cold pool is much colder than other schemes and surges swd too fast and results in large area of storms in the wrong place. SFC DEWPOINTS...very notable differences in schemes, especially MY.
4 June: all schemes developed discrete cells over central/wrn OK during the later afternoon hours and no storms developed.. WDM6/Thompson/Morrison by 06z have indication of linear MCS which is what happened. satellite...MY has massive cold tops from missed massive supercell in wrn OK and that prevents linear MCS from developing (probably due to resultant stable cold outflow from supercell that did not occur).

5 June: Morrison scheme completely off..including not developing significant convection in NMt. None of the schemes did a good job of convective evolution/upscale growth as the storms moved off the higher terrain of nm ewd through wrn/nrn TX Wed night. satellite..NSSL looks a bit better on cold cloud top extension/location, though it was off on convective pcpn evolution. SBCAPE...WDM6 has less areal coverage of SBCAPE than other schemes. Modified Thompson scheme had not as expansive or cold cloud tops as the regular scheme. not much difference in temperature fields between the 2 schemes.
6 June: Thompson was the best...although all of the schemes were too slow in ewd progression of convection located ahead of MCV moving through N TX. Satellite..my way cold cold and expansive cold cloud tops. WDM6 too little cold cloud tops. Thompson vs modified Thompson...regular Thompson created storms ahead of mcv in ern TX, while non modified did not and. Regular Thompson was very close to actual convective evolution. more notable differences in runs with microphysics than the PBL schemes.