Melick,C.J, I.L. Jirak, A.R. Dean, J. Correia Jr, and S.J. Weiss, 2012: Real Time Objective Verification of Convective Forecasts: 2012 HWT Spring Forecast Experiment. Preprints, *37th Natl. Wea. Assoc. Annual Meeting*, Madison, WI, Natl. Wea. Assoc., P1.52.

P1.52 Real Time Objective Verification of Convective Forecasts: 2012 HWT Spring Forecast Experiment

CHRISTOPHER J. MELICK
NOAA/NWS/NCEP/Storm Prediction Center, Norman, Oklahoma
University of Oklahoma/CIMMS, Norman, Oklahoma

ISRAEL L. JIRAK
NOAA/NWS/NCEP/Storm Prediction Center, Norman, Oklahoma

ANDREW R. DEAN
NOAA/NWS/NCEP/Storm Prediction Center, Norman, Oklahoma

JAMES CORREIA JR.
NOAA/NWS/NCEP/Storm Prediction Center, Norman, Oklahoma
University of Oklahoma/CIMMS, Norman, Oklahoma

STEVEN J. WEISS
NOAA/NWS/NCEP/Storm Prediction Center, Norman, Oklahoma

ABSTRACT

Objective forecast verification was conducted for the first time in near real-time during the 2012 NOAA Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (2012 SFE). One of the daily activities was to test the value of verification metrics by comparing the scores to subjective impressions of the participants. For this purpose, 1-km simulated reflectivity from high-resolution model and ensemble guidance was selected for examination. The evaluation was conducted via web pages using spatial plots for distinct time frames as well as a table which summarized statistical results. Feedback from the five-week period of the 2012 SFE indicated that the "neighborhood" techniques applied were more useful than grid point verification methods in the evaluation process, with the fractions skill score often rated as the most preferred metric.

1. Introduction

The Storm Prediction Center (SPC) and National Severe Storms Laboratory (NSSL) have jointly conducted the Spring Forecast Experiment (SFE) every spring season to test new tools and techniques for improving the prediction of hazardous convective weather. Both Kain et al. (2003) and Clark et al. (2012b) provide a detailed history on the annual SFE from the first official program in 2000 through recent years after the move to the NOAA Hazardous Weather Testbed

Corresponding author address: Christopher J. Melick, NOAA/NWS/NCEP Storm Prediction Center, 120 David L. Boren Blvd, Norman, OK 73072; E-mail: chris.melick@noaa.gov

(HWT) at the National Weather Center. Over the course of this time frame, the ultimate goal has been to foster collaboration between researchers and operational forecasters with the hope of transferring promising ideas from research into operations. In order to facilitate this process, evaluations of experimental model and human forecasts have consistently been a part of the daily activities, with 2012 SFE being no different. In previous SFEs, however, these assessments were done solely in a subjective fashion with objective metrics computed post-experiment (e.g., Kain et al. 2008).

During the five-week period of the 2012 SFE (M-F; May 7 – June 8), objective verification of high-resolution model forecasts was conducted locally for the first time in near real-time. Because of its relevance to severe weather, simulated 1-km above ground level (AGL) reflectivity was evaluated using several convection-allowing models (CAMs) in both a deterministic and ensemble system framework. *The focus in the current investigation was placed on testing the utility of a few selected verification metrics on high-resolution data by comparing those statistics to subjective evaluations from the participants.* The goal is to refine the most useful approaches and measures for subsequent trials in the HWT and eventual application across the community to better assess high-resolution model performance.

2. Data and Methodology

a. Data

The performance of many operational and experimental CAMs was explored rigorously during the 2012 SFE. All forecasts considered in the evaluation had grid spacing of about 4-km, were initialized daily at 00 UTC, and covered the 25 weekdays from May 7th – June 8th. The verification of hourly output of 1-km simulated reflectivity was limited, however, to a 20-hour period (16-12 UTC) which corresponded to the time frame that experimental human forecasts were valid. These daily evaluations were also restricted to a mesoscale "area of interest" for possible severe convection. In order to define the spatial extent of the small domain, surface weather stations listed in Table 1 served as movable center points for locations across the United States.

Table 1. Description of the surface weather stations selected for each of the 25 days as center-points during 2012 SFE. All of the daily evaluations were restricted to a mesoscale "area of interest" for possible severe convection. This small domain was movable to locations in the eastern and central United States. Consult Fig. 1 for an example plot showing the spatial extent.

	Center-point	
Date[YYMMDD]	3-char ID	Station Name, State
120507	BMQ	Burnet, TX
120508	SJT	San Angelo, TX
120509	AGS	Augusta/Bush, GA
120510	SAT	San Antonio, TX
120511	SAT	San Antonio, TX
120514	DRT	Del Rio, TX
120515	RAC	Racine, WI
120516	RUT	Rutland State, VT
120517	GBD	Great Bend, KS
120518	MBG	Mobridge, SD
120521	CDS	Childress, TX
120522	MBG	Mobridge, SD
120523	BVN	Albion Municipal, NE
120524	RGK	Red Wing, MN
120525	ICT	Witchita, KS
120528	ORD	Chigao O'Hare, IL
120529	OKC	Oklahoma City, OK
120530	END	Enid/Vance AFB, OK
120531	MSL	Muscle Shoal, AL
120601	AVC	South Hill/Meckl, VA
120604	BZN	Bozeman, MT
120605	LGC	La Grange, GA
120606	LBF	North Platte, NE
120607	TIF	Thedford/Thomas, NE
120608	2WX	Buffalo, SD

1) DETERMINISTIC MODELS AND ENSEMBLES

Table 2 provides a description of the two deterministic models used in the examination. These models were chosen based on ease of accessibility and to compare a relatively new 4-km model to one that has been examined at the SPC for several years. One model was the WRF-ARW version produced in real time at NSSL (hereafter referred to as NSSL-WRF; Sobash et al. 2011). Since late 2006, forecasts out to 36-hours from this 4-km CAM have been directly transmitted to SPC. More recently in 2011, National Centers for Environmental Prediction (NCEP) began running the Nonhydrostatic Multiscale Model on a rotated, Arakawa B-grid (NMMB) in the North American Mesoscale (NAM) slot, replacing the WRF-NMM framework. One-way, higher resolution nests are possible in this new framework using lateral boundary conditions (LBCs) from the 12-km NMMB parent. A 4-km nest (hereafter referred to as NAM-Nest) was used for comparison during the 2012 SFE.

Table 3 describes configuration specifics for three different storm-scale ensembles considered in investigating ensemble probability thresholds of simulated reflectivity. As had been the case since 2007, the University of Oklahoma (OU) Center for Analysis and Prediction of Storms (CAPS) supplied the SFE with data from their 4-km grid length storm-scale ensemble forecast (SSEF) system. The SSEF employs three dynamical cores (ARW, NMM, ARPS) and is a multi-initial condition (IC), multi-lateral boundary condition (LBC), and multi-physics system with 12 core members included in the post-processed ensemble products (Clark et al. 2012a). As a practical alternative, SPC developed the storm-scale ensemble of opportunity (SSEO) in 2011 by processing five existing deterministic CAMs (Jirak et al. 2012). Although SPC has limited control over its configuration, the SSEO remains attractive because of the minimal computational costs involved. Finally, a direct feed to SPC from the Air Force Weather Agency (AFWA) allowed the testing of their operational, 10-member 4-km WRF-ARW ensemble. The multi-IC/LBCs of the AFWA storm-scale ensemble are comprised from various, downscaled global model forecasts (Clark et al. 2012a).

Table 2. Configuration of the 00Z initialized deterministic CAMs used in 2012 SFE study.

Model	Grid Spacing	Vert. Levels	Time Step	Fcst. Length	PBL	Micro
NSSL-WRF	4-km	35	24 s	36 h	MYJ	WSM6
NAM-Nest	4-km	60	8.89 s	60 h	MYJ	Ferrier

Table 3. Configuration of the 00Z initialized storm-scale ensembles used in 2012 SFE study.

Ensemble	Grid Spacing	Core # Members	Fcst. Length	Reference
SSEO	4-5 km	7	36 h	Jirak et al. (2012)
AFWA	4-km	10	72 h	Clark et al. (2012a)
SSEF	4-km	12	36 h	Clark et al. (2012a); Kong et al. (2012)

2) VERIFICATION DATA

Verification of high-resolution reflectivity forecasts was performed using gridded radar observations of mosaic hybrid-scan reflectivity from the National Mosaic and Multi-Sensor QPE (NMQ) System (Zhang et al. 2011). Given the very fine resolution (0.01x0.01 degree) of the NMQ, a direct one-to-one comparison was achieved through interpolation of all the data onto common GEMPAK (GEneral Meteorological PAcKage; desJardins et al., 1991) grids (one for deterministic model evaluation and one for the ensemble component. For the statistical analysis, a mask was also applied to ignore grid points outside of the contiguous United States for those days when the mesoscale "area of interest" included such regions.

b. Methodology: Verification Metrics and Techniques

The verification process required creation of forecast- and observed-storm grid points from the reflectivity by specifying a threshold, in this case ≥40 dBZ. A traditional (i.e., "at-the-grid point") method was used as one way for evaluating the deterministic CAMs. In this case, a 2x2 contingency table (Wilks 2006) was constructed from the binary (yes/no) event grids. After counts of hit, misses, false alarms, and correct nulls were obtained each forecast hour, some standard verification metrics were computed (Critical Success Index [CSI] and Gilbert Skill Score [GSS] – also known as Equitable Threat Score).

Validating "grid-point" to "grid-point" is inherently problematic, though, when considering the occurrence of a relatively rare weather phenomenon (e.g., thunderstorm; Ebert 2009). Instead, a better technique to account for spatial uncertainty relies on setting a radius of influence (ROI) to incorporate a "neighborhood" around each grid point. A 40-km ROI was used for the current investigation to be consistent with SPC Convective Outlooks (i.e., within 25 miles of a point). Thus, the goal was to evaluate measures that account for spatial uncertainty in the forecasts.

Roberts and Lean (2008) and Schwartz et al. (2010) described a process for calculating "neighborhood" fractional probabilities from a single model. By applying their formula to the NSSL-WRF and NAM-Nest at each grid point, the number of grid boxes with 1-km AGL simulated reflectivity ≥ 40 dBZ within a 40-km ROI was divided by the total number of boxes within that "neighborhood". As a result, a smoother field is obtained which shows predicted coverage of storms as opposed to precise placement or intensity of localized features. The same technique was then utilized to form corresponding probabilistic fields for the observations. For the fractions skill score (FSS; Schwartz et al. 2010), fractional coverage values of

the observations and models were directly compared . Alternatively, CSI was calculated at a threshold of 10% for comparison among the models and ensembles.

Ensemble probabilities of an event occurrence were examined using the SSEO, AFWA, and SSEF systems. In this case, the probabilities were computed as the fraction of members with one or more grid points meeting or exceeding the threshold (40 dBZ) within the ROI (40-km). Harless (2010) introduced this concept as a binary neighborhood ensemble probability (BNEP) and found it to be skillful in quantifying forecast uncertainty associated with severe weather outbreaks. Similar to her work, a 2-D Gaussian kernel operator was also utilized with the weighting function set to 10 grid points, thereby effectively spreading the response to a 40-km distance. This acted to smooth the ensemble probabilities and create spatial probability distributions for the observed events given that there was only one source of radar data (i.e., 100% probability of occurrence). Again, an objective evaluation was performed by using CSI and FSS each forecast hour for the storm-scale ensembles.

3. Results

a. 2012 SFE Website

The afternoon evaluation component of the daily activities in the HWT sought participant feedback in comparing the verification metrics with their subjective impressions of the high-resolution model forecast performance. To facilitate a simple and quick diagnosis, the various forecast metrics were available on the 2012 SFE website (http://hwt.nssl.noaa.gov/Spring_2012/) for next-day evaluations. Time-matched images of forecasts and observations were created and displayed on web pages with the option to overlay the computed statistics. In order to illustrate this functionality, example snapshots highlighting ensemble and deterministic model plot comparisons are presented in Figs. 1 and 2, respectively.

In addition, the SFE participants were able to get a summary of the hourly objective results for a particular day in a tabular format. The table creation (Fig. 3) was driven on a separate web page by the choice of a date, verification method summary, and verification metric selected from a drop down menu. To cover an overall analysis for all five weeks, another summary table was also made available for examining trends across multiple days (Fig. 4), which was offered through dynamic calculation in PHP.

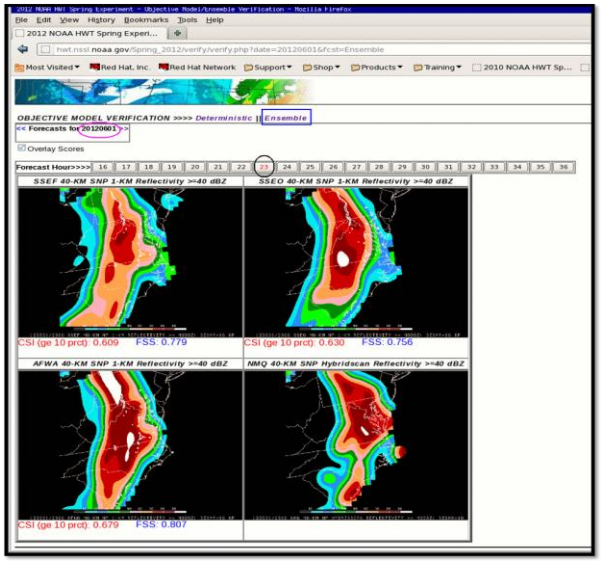

Figure 1. Sample spatial plots from 2012 SFE website illustrating the ability to overlay verification metric scores for storm-scale ensemble guidance. The ensemble probabilistic forecasts for 1-km simulated reflectivity are valid at 23Z on June 1st, 2012 for a mesoscale "area of interest" centered over south-central VA. The upper-left panel shows the CAPS SSEF, the upper-right shows the SPC SSEO, the lower-left displays the AFWA ensemble, with the observations of hybrid-scan reflectivity from the NMQ system located in the lower right. Beneath each forecast, the corresponding CSI and FSS are displayed as well. The figure is annotated to highlight the date, forecast time, and type of display.

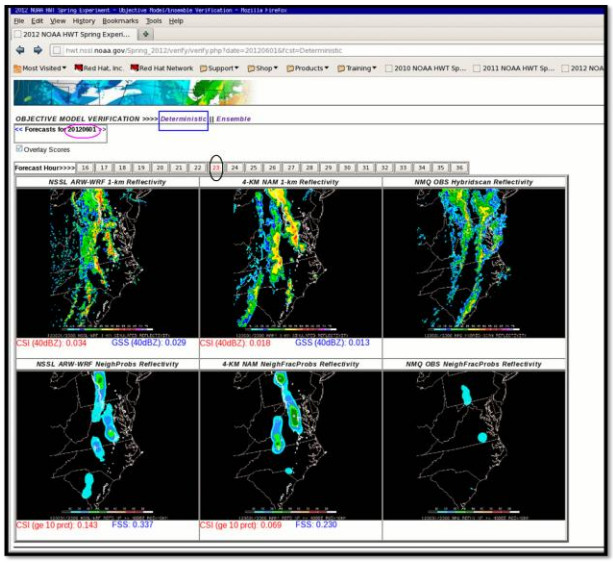

Figure 2. Similar to Fig. 1 except for deterministic guidance from the NSSL-WRF (left panels) and NAM-Nest (middle panels) models. The raw 1-km AGL simulated reflectivity is displayed in the top row of plots while probabilities from a fractional "neighborhood" approach are given in the bottom row. The verifying observations on the far right again come from the NMQ system. Besides computing CSI for verification at the grid point, the traditional verification metrics displayed in the top row also include Gilbert Skill Score (GSS).

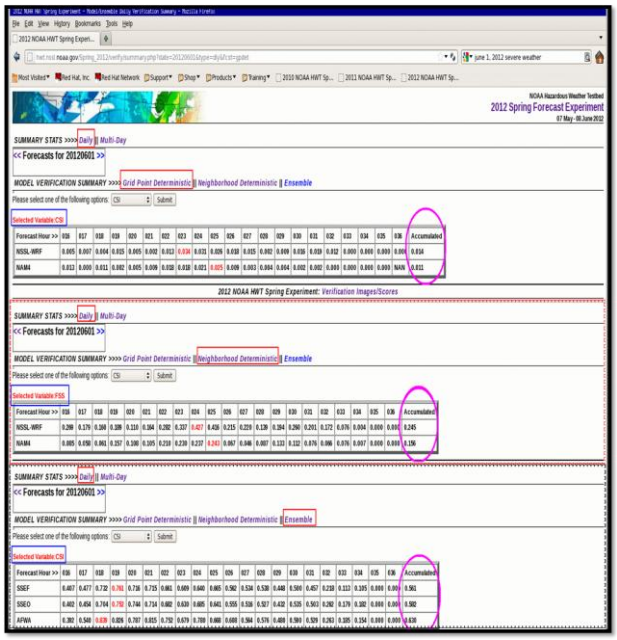

Figure 3. Sample composite of several tables created from 2012 SFE website which summarizes daily verification metrics for 1-km AGL simulated reflectivity. The table is created from a variety of user options: choice of a date, the type of verification method summary, and selection from a drop down list of skill scores (CSI being the default). Skill score results are binned by forecast hours (columns) from 16Z-12Z and models/ensemble systems (rows). The three tables displayed in this example from June 1st, 2012 are CSI values for grid point comparison of deterministic models (top), FSS values using a fractional "neighborhood" method of deterministic models (middle), and FSS values from an evaluation of storm-scale ensemble probabilities. Again, annotation is used to emphasize some options and functionality.

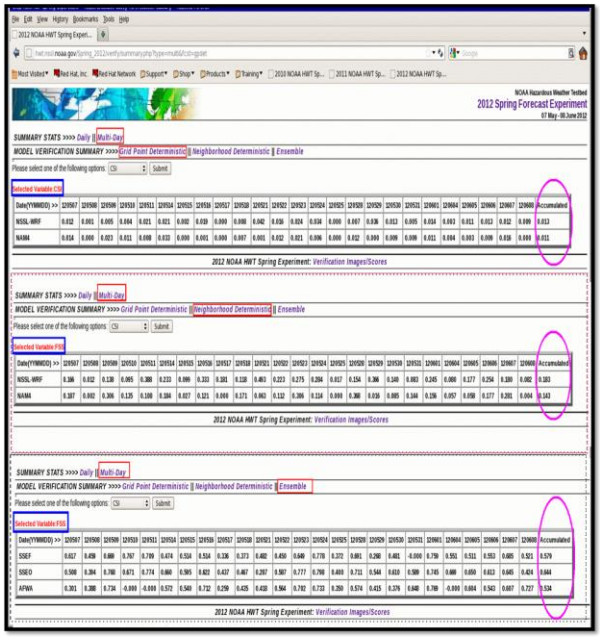

Figure 4. Similar to Fig. 3 except for tables created from the 2012 SFE website showcasing multiple day statistical overview. In this case, the daily accumulated scores from each day are presented, with the right most column representing the final outcome accumulated over the entire twenty five days of the 2012 SFE (5/7/2012 – 6/8/2012).

b. Multiple Day Forecast Verification Metrics

Figure 5 shows the accumulated multiple day forecast verification metrics for simulated reflectivity. The FSS from all three storm-scale ensemble systems (SSEO, AFWA, SSEF) surpassed the skill value of 0.5 with the SSEO even exceeding 0.6. The accumulated CSI results ranged from a little above 0.3 for AFWA to slightly over 0.4 with SSEO (Fig. 5). The skill scores were much lower for the deterministic CAMs. Still, a noticeable distinction existed as the fractional "neighborhood" CSI and FSS revealed higher values than the traditional, grid-point CSI and GSS. Between the latter two metrics, grid-point GSS was slightly worse but both showed values near zero. When comparing the verification metrics amongst the deterministic CAMs, the NSSL-WRF had higher values than the NAM-Nest.

Similar multiple-day statistics by forecast hour are displayed in Fig. 6. From this perspective, the separate panels present trend lines from 16-12 UTC for all of the methods and verification metrics investigated. Results indicated that the maximum skill often occurred in mid-afternoon when convective activity was generally at its peak across the mesoscale "area of interest". Generally, the storm-scale ensemble performance peak was higher and broader than the deterministic CAMs. For the latter, all four metrics showed low skill scores for all forecast hours with the fractional "neighborhood" approach showing the highest value near 0.25 at 20 UTC (middle right panel in Fig. 6). When comparing the two deterministic CAMs, the NSSL-WRF usually performed better for most forecast hours. With respect to the ensembles, the SSEO FSS revealed the highest skill values ranging from 0.6 to 0.7 (bottom right panel in Fig. 6).

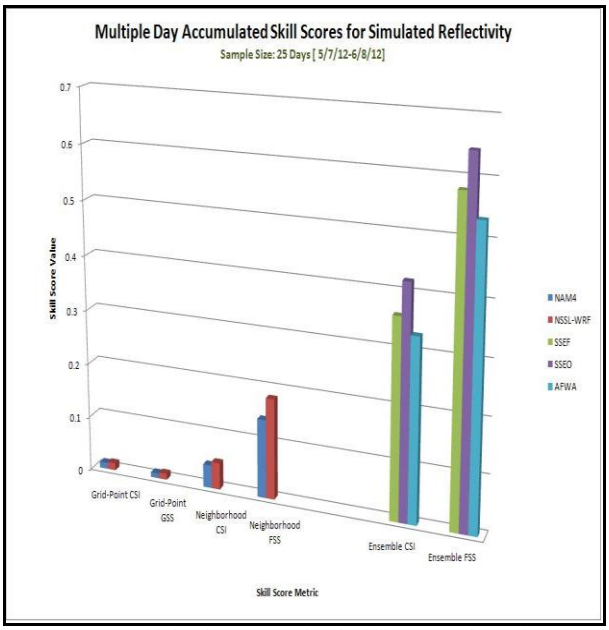

Figure 5. Multiple day accumulated skill scores for 1-km AGL simulated reflectivity from the 25 days (5/7/2012 – 6/8/2012) of the 2012 SFE. The following verification metrics are presented in this graph: CSI/GSS for traditional grid-point comparisons of deterministic models, CSI/FSS for fractional "neighborhood" comparisons of deterministic models, and CSI/FSS for evaluation of storm-scale ensemble systems. It should be noted that the AFWA ensemble data was unavailable for three days with the CAPS SSEF ensemble data missing for one.

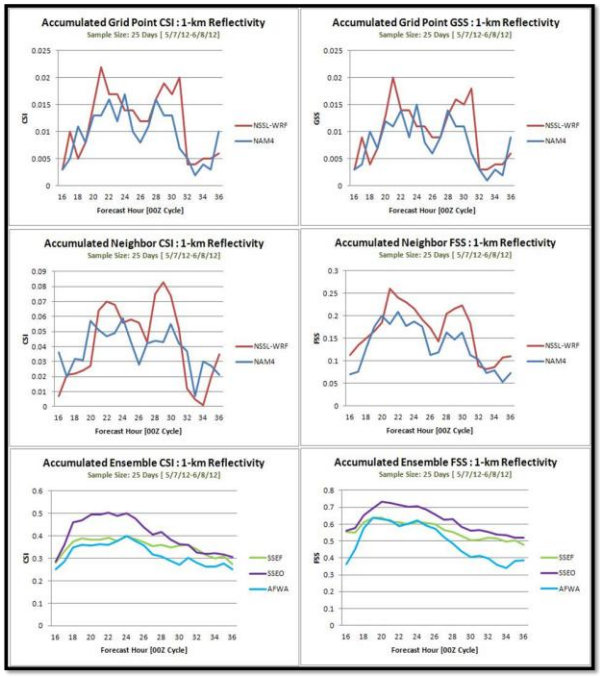

Figure 6. 2012 SFE multiple day accumulated skill scores by forecast hour for 1-km AGL simulated reflectivity. Multiple panels display diurnal trend lines for deterministic model evaluations at the grid point [CSI (top left) and GSS (top right)] and from a fractional "neighborhood" approach [CSI (middle left) and FSS (middle right). The bottom two panels show forecast hour trend lines for storm-scale ensemble metrics [CSI (left) and FSS (right)].

c. Daily Distribution of Forecast Verification Metrics

The findings thus far have been accumulated over the entire five week period of the 2012 SFE. To get an indication of the variability in daily skill scores, separate box-and-whisker plots for the deterministic (Fig. 7) and ensemble (Fig. 8) forecasts were produced. All of the percentile rankings for grid-point CSI and GSS were close to zero (top panel in Fig. 7). The distribution in FSS displayed a wider range (bottom panel in Fig. 7) and a predominant upward shift in the distribution of the NSSL-WRF results compared to the NAM-Nest. An examination of CSI and FSS from the ensemble in Fig. 8 suggests a substantial overlap amongst all three storm-scale ensembles. Nevertheless, AFWA subjectively showed a tendency to be an outlier and have more forecasts with lower scores, this being implied by the lower 25th percentile value for FSS.

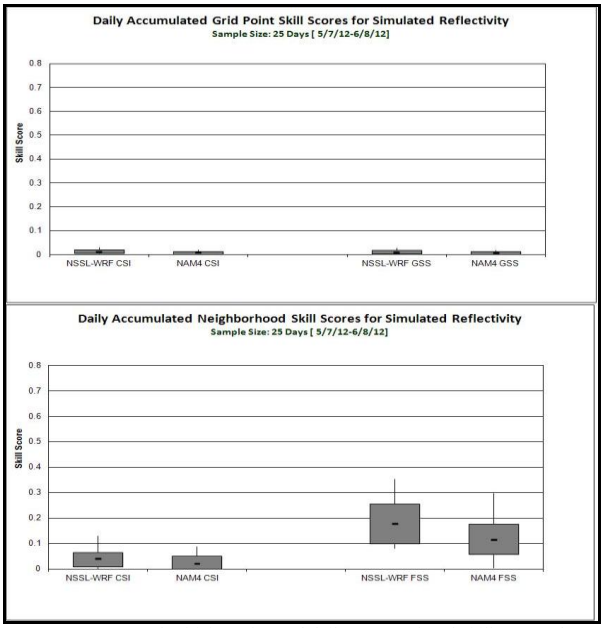

Figure 7. Box and whisker plots of daily accumulated skill scores for 1-kmAGL simulated reflectivity from deterministic model solutions. The top (bottom) row presents results from the traditional grid-point ("neighborhood") method. The whiskers correspond to the 10th and 90th percentile rankings from the 25 days during 2012 SFE.

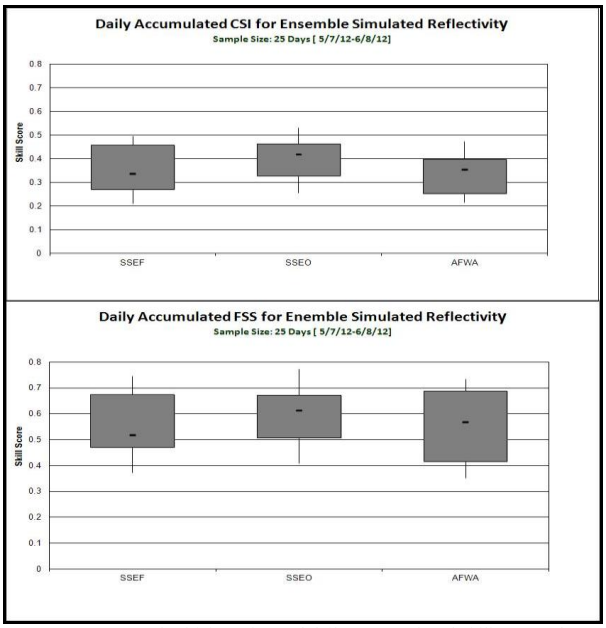

Figure 8. Box and whisker plots of daily accumulated skill scores for 1-km AGL simulated reflectivity from storm-scale ensemble systems. The top (bottom) row presents results from the FSS (CSI). The whiskers correspond to the 10th and 90th percentile rankings for 2012 SFE but the sample size is smaller for the AFWA (22 days) and CAPS SSEF (24 days) systems.

d. Participant Feedback

Another major purpose of the research was to compare the participant feedback to the objective results for 1-km AGL simulated reflectivity. Figures 9 and 10 present tallies gathered from the responses on 23 days of forecast verification technique/metric, respectively. While there were occasional differences, subjective impressions typically matched objective results during the 2012 SFE daily evaluation activity. In particular, the NAM-Nest was frequently rated "Worse" to "Much Worse" than the NSSL-WRF, and the SSEO forecasts received the most ratings of "Good" (Fig. 9). One of the discrepancies was the subjective ranking of the SSEF with respect to the AFWA ensemble. Here, AFWA was rated "Good" more often than SSEF and rated "Poor" or "Very Poor" less often than SSEF (Fig. 9) while the objective metrics preferred the SSEF over the AFWA.

Most importantly, the forecast verification metrics with the deterministic CAMs "Agreed" to "Strongly Agreed" with the subjective impressions more so for the fractional "neighborhood" technique than the grid point approach (Fig. 10). With regard to the best metric from the deterministic forecasts, the 13 day tally for fractional "neighborhood" FSS indicated that it was preferred over the other metrics. Finally, feedback suggested that the ensemble CSI and FSS from the BNEP method was consistent with the subjective impressions of the participants for well over half of the five week period.

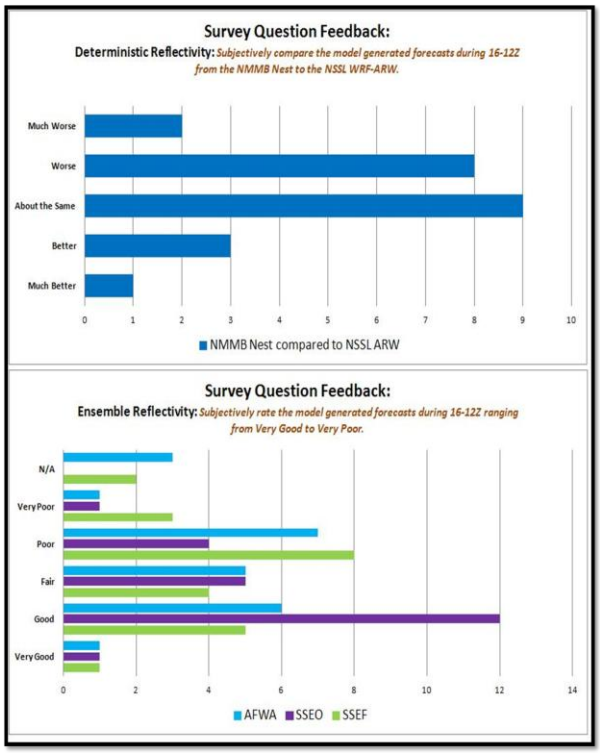

Figure 9. Participant feedback tallies gathered during 2012 SFE daily activity evaluation. The results obtained from the two survey questions covered subjective comparisons of 1-km reflectivity from deterministic models (top graph) and storm-scale ensemble systems (bottom graph). The wording of the questions is given as: "Deterministic Reflectivity: Subjectively compare the model generated forecasts during 16-12UTC from the NMMB Nest to the NSSL-WRF" and "Ensemble Reflectivity: Subjectively rate the model generated forecasts during 16-12UTC ranging from Very Good to Very Poor". The sample size was 23 since no evaluations were performed on Memorial Day (5/28/2012) nor the very last day (6/8/2012) of the 2012 SFE.

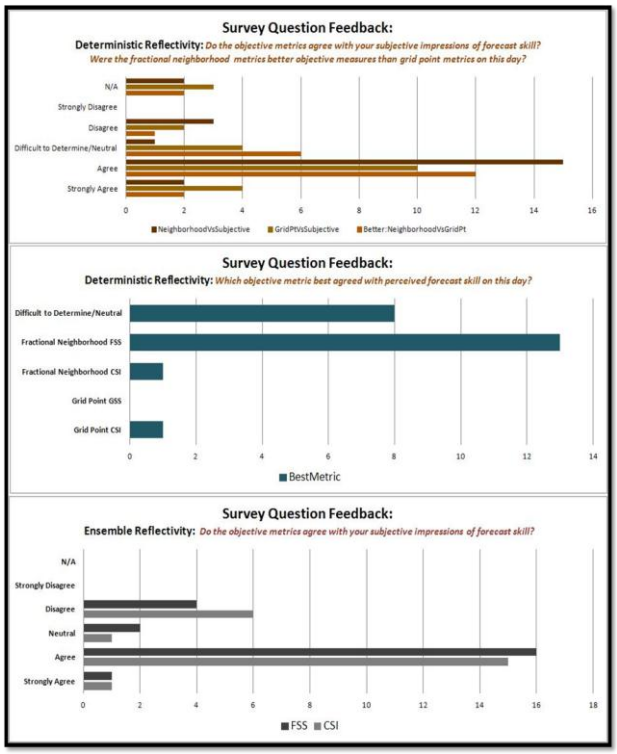

Figure 10. Same as in Fig. 9 except for survey questions relevant to assessing the forecast verification metrics examined from the deterministic models (top two graphs) and storm-scale ensemble systems (bottom graph). The wording of the questions is given as: "Deterministic Reflectivity: Do the objective metrics agree with your subjective impressions of forecast skill? Were the fractional neighborhood metrics better objective measures than grid point metrics on this day?", "Deterministic Reflectivity: Which objective metric best agreed with perceived forecast skill on this day?", and "Ensemble Reflectivity: Do the objective metrics agree with your subjective impressions of forecast skill?".

4. Summary and Conclusions

SPC conducted objective verification of high-resolution model forecasts of 1-km AGL simulated reflectivity during the 2012 SFE in near real-time. This provided the opportunity to assess the utility of selected verification metrics in relation to subjective evaluations of model performance. Time-matched spatial plots of forecasts and observations were displayed on the SFE 2012 webpage for visual comparison. Unlike prior years, however, objective skill scores were also calculated for each forecast time period and overlaid with the appropriate images as well as included in table summaries.

A major finding was that the "neighborhood" objective measures best agreed with the subjective evaluations for the majority of the five week period. In particular, feedback from participants indicated that utilizing a "neighborhood" technique was usually better than evaluating high-resolution models using grid point methods. The FSS was often rated the most preferred metric, with the accumulated daily and multiple daily results continuously showing the highest values. The skillful scores (values over 0.5) from the SSEO, SSEF, and AFWA indicate the usefulness of a probabilistic approach through the use of ensembles. Future plans for the 2013 SFE include possibly incorporating more probability thresholds to CSI, allow for timing uncertainty between the forecasts and observations, and exploring the use of different ROI.

Acknowledgements. This research was supported by an allocation of advanced computing resources provided by the National Science Foundation. The computations for the CAPS SSEF were performed on Kraken (a Cray XT5) at the National Institute for Computational Science (NICS; http://www.nics.tennessee.edu/). We would like to thank Ming Xue, Fanyou Kong, and Kevin Thomas from OU CAPS for generating and providing SSEF data and Jack Kain and Patrick Marsh of NSSL for making the data available to SPC. Evan Kuchera and Scott Rentschler of AFWA generously provided AFWA data. We would also like to thank the participants from the 2012 SFE for providing useful feedback for this evaluation.

REFERENCES

Clark, A.J, and Coauthors, 2012a: Spring Forecasting Experiment 2012 Program Overview and Operations Plan. [Available online at http://hwt.nssl.noaa.gov/Spring_2012/OPS_plan_draft.pdf

_____, and Coauthors, 2012b: An Overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74.

desJardins, M.L., K.F. Brill, and S.S. Schotz, 1991: Use of GEMPAK on Unix workstations, *Proc. 7th International Conf. on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology*, New Orleans, LA, Amer. Meteor. Soc., 449-453.

Ebert, E.E., 2009: Neighborhood verification: A strategy for rewarding close forecasts. *Wea. Forecasting*, **24**, 1498–1510.

Harless, A.R., 2010: A report-based verification study of the CAPS 2008 storm-scale ensemble forecasts for severe convective weather. M.S. Thesis, School of Meteorology, University of Oklahoma, 143 pp.

Jirak, I.L, S.J. Weiss, and C.J. Melick, 2012: The SPC Storm-Scale Ensemble of Opportunity: Overview and results from the 2012 Hazardous Weather Testbed Spring Forecasting Experiment. *26th Conf. Severe Local Storms*, Nashville, TN, Amer. Meteor. Soc., Paper P9.137.

Kain, J.S., P.R. Janish, S.J. Weiss, R.S. Schneider, M.E. Baldwin, and H.E. Brooks, 2003: Collaboration between forecasters and research scientists at the NSSL and SPC: The Spring Program. *Bull. Amer. Meteor. Soc.*, **84**, 1797–1806.

____, S.J. Weiss, D.R. Bright, M.E. Baldwin, J.J. Levit, G.W. Carbin, C.S. Schwartz, M.L. Weisman, K.K. Droegemeier, D.B. Weber, and K.W. Thomas, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection allowing NWP. *Wea. Forecasting*, **23**, 931-952.

Roberts, N.M., and H.W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97.

Schwartz, C. S., and Coauthors, 2010: Toward improved convection-allowing ensembles: model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263–280.

Sobash, R.A., J.S. Kain, D.R. Bright, A.R. Dean, M.C. Coniglio, S.J. Weiss, 2011: Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme phenomena in convection-allowing model forecasts. *Wea. Forecasting*, **26**, 714–728.

Wilks, D.S., 2006: Forecast Verification. *Statistical methods in the atmospheric sciences*, 2nd Edition. Academic Press, 260-268.

Zhang, Jian, and Coauthors, 2011: National Mosaic and Multi-Sensor QPE (NMQ) System: Description, results, and future plans. *Bull. Amer. Meteor. Soc.*, **92**, 1321–1338.