Melick,C.J, I.L. Jirak, J. Correia Jr, A.R. Dean, and S.J. Weiss, 2013: Utility of Objective Verification Metrics during the 2013 HWT Spring Forecasting Experiment. Preprints, *38th Natl. Wea. Assoc. Annual Meeting*, Charleston, SC, Natl. Wea. Assoc., P1.27.

# P1.27 Utility of Objective Verification Metrics during the 2013 HWT Spring Forecasting Experiment

CHRISTOPHER J. MELICK
Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma
NWS Storm Prediction Center, Norman, Oklahoma

ISRAEL L. JIRAK
NWS Storm Prediction Center, Norman, Oklahoma

JAMES CORREIA JR.
Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma
NWS Storm Prediction Center, Norman, Oklahoma

ANDREW R. DEAN
NWS Storm Prediction Center, Norman, Oklahoma

STEVEN J. WEISS
NWS Storm Prediction Center, Norman, Oklahoma

ABSTRACT

Objective forecast verification was conducted for the second year in near real-time during the 2013 Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (2013 SFE). As part of the daily activities, experimental probabilistic forecasts for severe thunderstorms were created. These forecasts were then evaluated the next day via webpages with preliminary local storm reports (LSR) serving as the verification dataset. The idea was to further explore the value of incorporating verification metrics by comparing various scores to subjective evaluations from the participants. In addition to the forecast verification metrics examined in the 2012 SFE, the relative skill score was introduced since it was designed with a baseline reference capable of measuring skill of rare-event forecasts (i.e. severe thunderstorms). Results suggested that the relative skill scores were generally better on days with more severe weather reports. Further, the participants generated skillful forecasts at the lower probability thresholds, as the relative skill scores were predominately positive in accordance with favorable subjective ratings.

## 1. Introduction

The Storm Prediction Center (SPC) and National Severe Storms Laboratory (NSSL) jointly conduct the Spring Forecasting Experiment (SFE) each spring in the Hazardous Weather Testbed (HWT) at the National Weather Center in Norman, OK. Historical descriptions of the annual SFE dating back to 2000 can be found in both Kain et al. (2003) and Clark et al. (2012). As in prior years, both model and forecast evaluations remained as an activity during the 2013 SFE. Nevertheless, for many of those years, a statistical assessment often waited until after the participants had left and the program had concluded (e.g., Kain et al. 2008). Starting with the 2012 SFE, a near real-time objective evaluation component was added to complement the traditional, subjective verification

*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

performed daily (Melick et al. 2012). Given the promising results from this initial exploration, use of forecast verification metrics resumed and was expanded in the 2013 SFE.

Next-day evaluations of severe weather forecasts produced by the participants occurred once again during the five-week period of the 2013 SFE (May 6 – June 7). The current work addresses the performance of these experimental probabilistic forecasts in predicting the occurrence of damaging winds, hail, and tornadoes associated with severe thunderstorms. An emphasis is placed on testing the utility of several forecast verification metrics by relating the objective results to the subjective impressions provided by the participants. This objective is similar to that of Melick et al. (2012), except their work dealt with verification of high-resolution model forecasts of simulated reflectivity during the 2012 SFE. The expectation is that these types of approaches will continue in some fashion for many years in the HWT, especially considering that SPC has already commenced internal testing of objective verification for probabilistic ensemble guidance and SPC operational convective outlooks.

## 2. Data and Methodology

### *a. Data*

The skill of experimental forecasts issued in the HWT was investigated during the course of the 2013 SFE. SFE participants from two separate teams (named: East and West) produced identical probabilistic products consisting of total severe (wind gusts ≥ 50 kt, hail ≥ 1" in diameter, and any tornado) forecasts that were valid within 25 miles [~40-km] of a point, as defined in SPC operational convective outlooks. More precisely, both teams used the same probability contours in the SPC Day 2 convective outlook product (5, 15, 30, 45, and 60%), but also had **the option of including extra contour lines (every 5%) for localized maxima.** The teams were also permitted to delineate an area for ≥10% probability of significant severe storms (i.e. hail ≥ 2" in diameter, wind gusts ≥ 65 kt). While it would be interesting in the future to examine significant severe events, computations of forecast verification metrics were specifically restricted here to just *any type of severe weather occurrence*.

All severe weather forecasts considered in the evaluation covered the 16Z-12Z forecast period for 24 weekdays from May 6th – June 7th (with no activities on Memorial Day). Additional forecasts of higher temporal resolution (3-hr; 18-21, 21-00, and 00-03Z) were also created but were not examined further since the sample size of verifying observations would be much smaller compared to the 20-hr, full-period forecast. Verification was obtained by utilizing preliminary local storm reports (LSR) received from the National Weather Service forecast offices through the valid forecast period (just after 12Z). These next-day evaluations were also restricted to a mesoscale "area of interest" for possible severe convection. Table 1 lists the surface weather stations that served as daily movable center-points along with the tally of verifying LSRs.

**Table 1.** Description of the surface weather stations selected for each of the 24 days as center-points during 2013 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. Also, the 16Z-12Z verifying tallies of LSRs over the restricted domain are displayed as well. Consult Fig. 1 for an example plot showing the spatial extent.

| | Center-point | Local Storm Reports |
|---|---|---|
| Date[YYMMDD] | Station Name, State (3-Char ID) | 16Z-12Z Verification |
| 130506 | Greensboro, NC (GSO) | 12 |
| 130507 | Gage, OK (GAG) | 27 |
| 130508 | Gage, OK (GAG) | 130 |
| 130509 | Corsicana, TX (CRS) | 100 |
| 130510 | College Station, TX (CLL) | 41 |
| 130513 | Lewistown, MT (LWT) | 12 |
| 130514 | Volk/Camp Douglas, WI (VOK) | 16 |
| 130515 | Austin, TX (AUS) | 40 |
| 130516 | North Platte, NE (LBF) | 12 |
| 130517 | Rapid City, SD (RAP) | 45 |
| 130520 | Muskogee, OK (MKO) | 189 |
| 130521 | Mount Pleasant, TX (OSA) | 121 |
| 130522 | Johnstown, PA (JST) | 127 |
| 130523 | Snyder/Winston, TX (SNK) | 90 |
| 130524 | Hill City, KS (HLC) | 40 |
| 130528 | Whiteman AFB, MO (SZL) | 89 |
| 130529 | Enid/Vance AFB, OK (END) | 160 |
| 130530 | Grove, OK (GMJ) | 141 |
| 130531 | Joplin, MO (JLN) | 150 |
| 130603 | Medicine Lodge, KS (P28) | 34 |
| 130604 | Enid/Vance AFB, OK (END) | 42 |
| 130605 | Graham Municipal, TX (RPH) | 111 |
| 130606 | Stephenville, TX (SEP) | 34 |
| 130607 | Cannon AFB/Clovis, NM (CVS) | 15 |

### *b. Methodology: Verification Metrics*

#### 1) RELIABILITY DIAGRAM

The reliability diagram (Wilks 2006) was utilized to illustrate the performance of probability forecasts for severe weather events by determining the

observed relative frequency as a function of forecast probability. This allowed for a quick visual means to understand properties of the probabilistic forecasts relative to "perfect reliability", a 1:1 diagonal line shown on the reliability diagram. For this application, the probability values from the experimental team forecasts and "practically perfect" hindcasts (Brooks et al. 1998; see description below) were grouped into six bins (0%, 5% ,15% , 30%, 45%, 60%) by rounding down to the nearest bin. Then, counts of grid points with one or more severe reports were evaluated for each probability bin and summed over all of the days. Similarly, sample sizes were also computed for the total number of forecast grid points that corresponded to each of the forecast probability bins. As a result, the ratio of these two results for each probability bin produced the relative frequency for the observations (i.e., the ordinate in the reliability diagram).

### 2) Contingency Table Metrics

Defining severe storm events for both the forecasts and observations was necessary in order to accomplish the objective evaluation. These events were determined by first placing the datasets on a 40-km grid (NCEP 212; http://www.nco.ncep.noaa.gov/pmb/docs/on388/tableb.html), similar to verification procedures used at SPC (e.g., Bright and Wandishin 2006). In the case of the experimental forecasts produced by the participants, grid point values of the probabilities were obtained from the drawn contours by a graph-to-grid routine in GEMPAK (GEneral Meteorological PAcKage; desJardins et al., 1991). More specifically, the technique produced non-continuous forecast probabilities, meaning that grid point values were constant between the contour lines and set to the lower probability contour (e.g., entire area between 5 and 15% contour lines is set to 5%). In the case of bounds for the minimum (maximum), anything less (greater) than 5% (60%) was set to 0% (60%). After the conversion in formats, binary (yes/no) event grids could be specified from the probabilistic information by specifying various thresholds to define the forecast area. As for the verification, if ≥ 1 severe weather report occurred within a 40-km radius of influence (ROI) of the grid box, it was recorded as a severe event.

A direct grid-point-to-grid-point comparison between the forecasts and observations can result in only four possible outcomes from the discrete predictands (i.e., yes/no). Thus, a 2x2 contingency table (Wilks 2006) was developed for each probability threshold to tally all possible combinations. After counts of hits, misses, false alarms, and correct nulls were obtained, standard verification metrics were computed (e.g., Critical Success Index [CSI]) for all of the fixed SPC thresholds (5, 15, 30, 45, 60%), as well as the probability for which the maximum CSI value occurred. For the statistical analysis, a mask was also applied to include only grid points over the contiguous United States within the small "area of interest".

### 3) Practically Perfect Hindcasts

Brooks et al. (1998) presented a technique to produce a meaningful baseline to relate to severe weather forecasts using the collection of LSRs recorded at SPC. Following their approach, "practically perfect" [PP] hindcasts were created by applying a two-dimensional Gaussian smoother (sigma=120-km) to the occurrence of one or more severe reports within 25 miles of a 40-km x 40-km grid box. In addition, another grid of analyzed, *significant* severe probabilities was created using one or more *significant* severe reports at a grid point. The PP method produced a probabilistic field which was considered to be consistent with what a forecaster would produce given prior (perfect) knowledge of the observations (Brooks et al. 1998). As with the experimental products issued by both teams, comparable scores from the 2x2 contingency table were determined by treating PP like a forecast and specifying identical probability thresholds (i.e. 5, 15, 30, 45, and 60%). Consequently, this allowed for reference in measuring the performance of severe weather forecasts from day to day, which was particularly beneficial since attaining high scores from traditional verification metrics can often be challenging.

### 4) Relative Skill Score

The notion of a relative skill score in verifying rare event forecasts was described by Hitchens et al. (2013). Their work utilized PP hindcasts as a reference to evaluate SPC convective outlook slight risk areas from 1973 to 2011. In its formulation, the relative skill score (RelSkill) is given by:

$$\text{RelSkill} = \left[ \frac{\text{CSI}_{\text{Forecast}} - \text{CSI}_{\text{MinPP}}}{\text{CSI}_{\text{MaxPP}} - \text{CSI}_{\text{MinPP}}} \right] \quad \textbf{(1)},$$

where $\text{CSI}_{\text{Forecast}}$ represents the value of CSI from the forecast being verified, $\text{CSI}_{\text{MaxPP}}$ is the maximum (upper bound) value of CSI from PP, and $\text{CSI}_{\text{MinPP}}$ is the minimum (lower bound) value of CSI from PP. Although the choice of performance measure is arbitrary in computing relative skill, usage of CSI as the metric was retained in the current investigation. In order to determine the upper and lower bounds in equation **(1)**, binary severe weather events were created for every PP probability threshold at an interval of one percent, similar to that in Hitchens et al. (2013). For the case of $\text{CSI}_{\text{MaxPP}}$, the maximum probability threshold was reached once the increase in CSI had terminated (going up from 1%; see Fig. 4 in Hitchens et al. (2013) for an example). On the other hand, the $\text{CSI}_{\text{MinPP}}$ was the theoretical CSI value when approaching 0% based on a downward extrapolation of CSI from one percent using the slope between the one and two percent threshold values (i.e., $\text{CSI}_{\text{MinPP}} \equiv \text{CSI}_{1\%} - (\text{CSI}_{2\%} - \text{CSI}_{1\%})$).

The range in values from RelSkill can vary between negative (i.e., when $\text{CSI}_{\text{Forecast}} < \text{CSI}_{\text{MinPP}}$) to greater than one (i.e., when $\text{CSI}_{\text{Forecast}} > \text{CSI}_{\text{MaxPP}}$). From the analysis work performed by Hitchens et al. (2013), little to no RelSkill was noted in SPC convective outlook slight risk areas until the mid-1990s, after which a steady increase occurred. One of the main differences from their study was the testing of several thresholds with the probabilistic experimental forecasts.

5) FRACTIONS SKILL SCORE

The verification metrics discussed thus far have been constrained to evaluating whether or not a severe weather event was predicted and whether or not a severe event occurred. Instead of setting a threshold and converting the probabilistic forecast into a binary one, the PP hindcast could serve as the verifying dataset. In this case, the probabilities from both the experimental forecasts and PP could be directly compared by calculating the fractions skill score (FSS; Schwartz et al. 2010), which is a variation on the brier skill score. The range on FSS is from 0 to 1, with the highest score indicating a perfect forecast and the lowest score revealing no skill without any overlap in non-zero probabilities. Similar to CSI and the relative skill score, computations of FSS were performed for the 24 days of the five-week period of the 2013 SFE.

## 3. Results

### *a. 2013 SFE Website*

One of the objectives in conducting the objective verification during the 2013 SFE was to provide a means for the HWT participants to quickly evaluate the experimental severe weather forecasts. This was accomplished by incorporating the ability to showcase various forecast verification metrics the next day from the 2013 SFE website (http://hwt.nssl.noaa.gov/Spring_2013/). Time-matched images of forecasts and observations were created and displayed on web pages along with the computed statistics. An example snapshot highlighting probabilistic forecast comparisons from both the East and West teams are presented in Fig. 1. A related survey question sought the subjective impressions of participants regarding the experimental severe weather forecasts.

In addition, the SFE participants were able to retrieve a summary of the objective results for all five weeks in a tabular format. The table creation (Fig. 2) was driven on a separate web page by choosing the forecast time period and then selecting the verification metric and probability threshold from drop-down menus. Another available feature was the accumulated statistic, which was offered through dynamic calculation in PHP.

### *b. Reliability Diagram*

The reliability diagram provided one method to get insight into aspects of the probabilistic forecast system. Figure 3 revealed the observed frequency of severe weather reports in each of the six forecast probability bins for the experimental forecasts and PP hindcasts. The forecasts were nearly reliable for probabilities up to and including the 15% bin where substantial sample sizes (on the order of 10,000 grid points) were present. Both the East team and PP tended to slightly under-predict at 30% with a more substantial under-prediction for all forecasts at 45% (Fig. 3). With respect to the highest probability bin (60%), results for the West team indicated reliability with a slight over-prediction for the East team. Still, limited confidence should be placed in the

findings at the 45%, and particularly, the 60% probability bins where the total number of forecast grid points had substantially diminished to an order of a few hundred or less (Fig. 3). Nonetheless, the overall impression was that the experimental probabilistic forecasts were fairly reliable over the relatively short five-week period, even more reliable than the PP hindcasts, especially at higher probabilities.

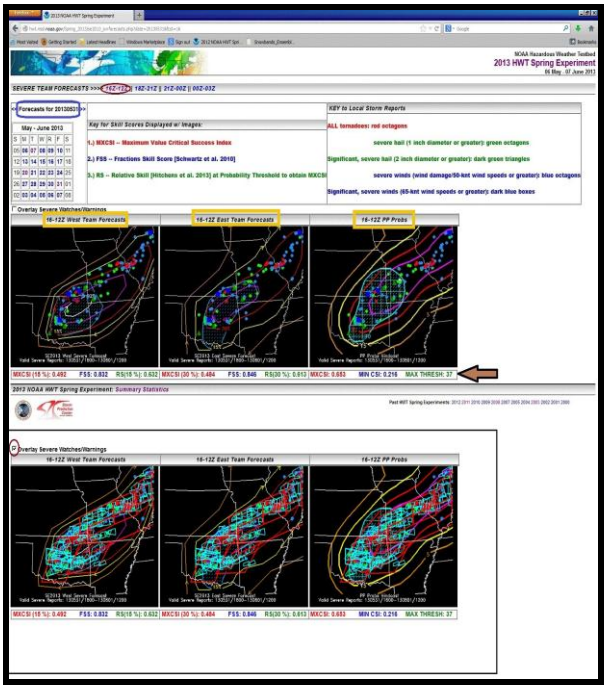

**Figure 1.** Sample spatial plots from 2013 SFE website illustrating the ability to display verification metric scores for experimental severe weather forecasts. The probabilistic forecasts for 16Z-12Z are valid starting on May 31st, 2013 for a mesoscale "area of interest" centered on Joplin, MO. For the top row, the far-left panel shows the probability contours from the West team, the middle represents those from the East team, and the far-right matches to the PP hindcast. In addition, a 10% or greater hatched area for significant severe storms is also predicted/analyzed, with the verifying observations from the LSRs overlaid on top of each of the plots. Beneath each of the experimental forecasts, the corresponding maximum threshold Critical Success Index (CSI), Fractions Skill Score (FSS), and the relative skill score (RS) obtained from the maximum threshold CSI are displayed as well. The upper and lower bounds of CSI from PP to calculate relative skill are given below the PP hindcast. Finally, the bottom row shows an overlay of severe thunderstorm/tornado watches and warnings issued by the NWS. The figure is annotated to highlight some of the details relevant to the date, type of forecast, forecast time period, forecast verification metrics, and other functionality. See text for more details.

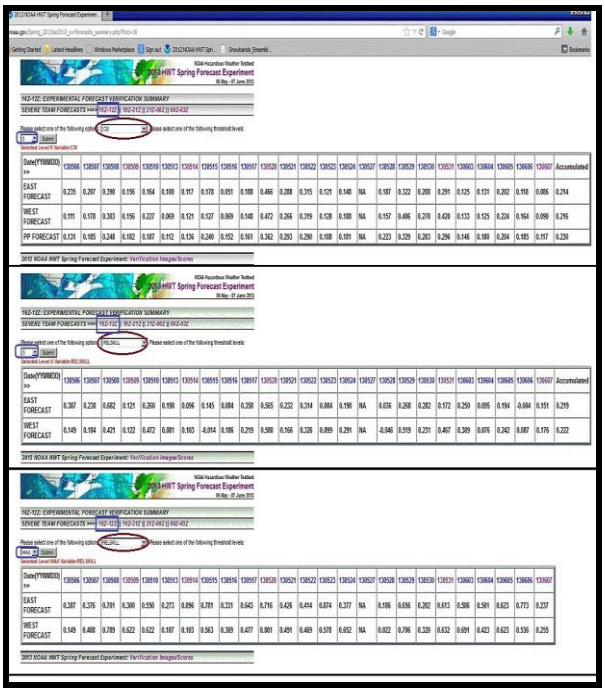

**Figure 2.** Sample composite of several tables created from 2013 SFE website which summarizes the multiple day (5/6/2013 – 6/7/2013) verification metrics for the experimental severe weather forecasts (and PP hindcast in some cases). The table is created from a variety of user options: forecast time period, forecast verification metric, and probability threshold (CSI at 5% being the default). Skill score results are binned according to the day of the SFE forecast with the rows separating the East team, West team, and PP results. Finally, accumulated statistics for a few appropriate forecast verification metrics are offered through dynamic calculation in PHP. The three tables displayed in this example are valid for the 16-12Z time period and show the CSI values at the 5% threshold (top), RS values at the 5% threshold (middle), and RS values from the maximum threshold CSI (bottom). Again, annotation is used to emphasize some options and functionality.

### *c. Accumulated Results: Contingency Table Verification*

Figure 4 presents the accumulated, multiple day results for contingency table forecast verification metrics using the performance diagram (Roebber 2009). The performance diagram (Roebber 2009) is appealing as it is able to summarize information on probability of detection [POD], false alarm ratio [FAR], frequency of hits [FOH], bias, and CSI from all forecasts in one illustration. Specifically, the overall statistics are shown at all fixed probability thresholds and broken up by each of the 16Z-12Z experimental team forecasts as well as the PP hindcast.

The first noticeable feature in Fig. 4 was that the 2013 SFE contingency table verification metric scores were higher at all probability thresholds for the PP hindcast compared to the experimental team forecasts

for severe weather. This was concurrent with the fact that most of the probability thresholds tended to occur more often with PP, as revealed in Fig. 4. More importantly, though, the synopsis for the entire SFE was that most scores were favorable or maximized (i.e. CSI) at the 15% probability level for both of the experimental team forecasts and at the 30% probability level for the PP hindcast. At these optimal thresholds, CSI values reached slightly under 0.3 for the forecasts produced by the participants and PP slightly exceeded 0.45 for CSI (Fig. 4).

For the other fixed probability thresholds, a very large POD was evident in Fig. 4 at the 5% level (greater than 0.9) since the 5% forecast areas often captured a significant majority of the observed severe weather reports. On the other hand, FOH (FAR) values started to go above (below) 0.5 at or above the 30% probability threshold for the experimental team forecasts, the result of false alarms falling substantially relative to hits as the spatial coverage diminished. Finally, the score trend inconsistency from the 45% to 60% probability threshold (Fig. 4), especially for the East team, was presumably the result of a very small sample size of days (Fig. 5) and grid points (Fig. 3).

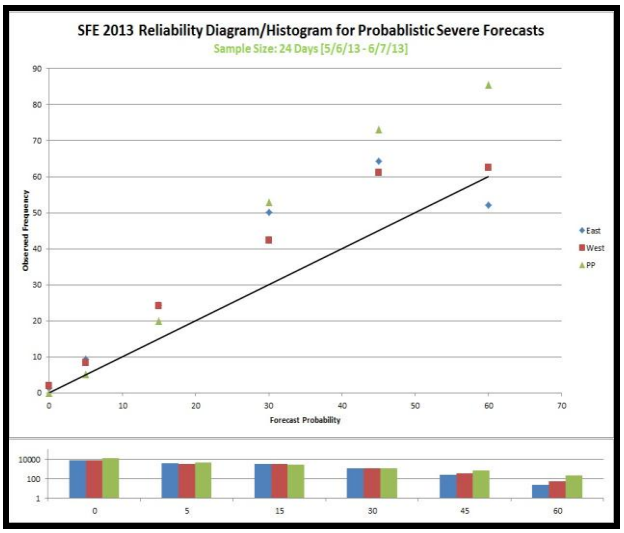


**Figure 3.** Reliability diagram for 16-12Z probabilistic severe forecasts using accumulated grid point tallies over the 24 days (5/7/2012 – 6/8/2012) of the 2013 SFE. The inset histogram displayed below gives the forecast subsample sizes computed each of the forecast probability bins (0%, 5%, 10%, 15%, 30%, 45%, and 60%). The y-axis on the histogram is a logarithmic scale so as to represent the large disparity in occurrences between the lower and higher probability thresholds. The color code legend for the markers reveals the matching type of forecast (East team, West team, PP hindcast).

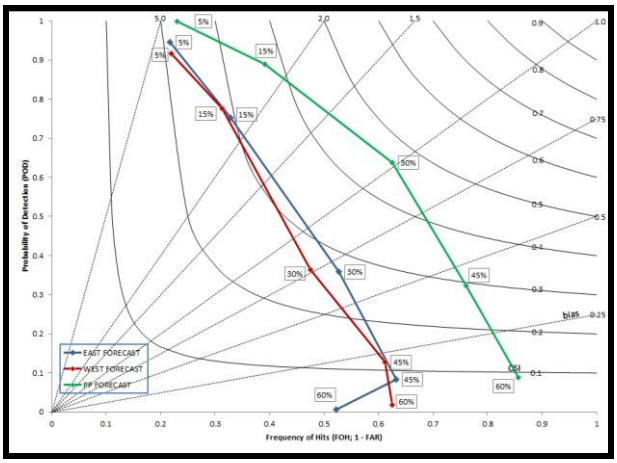


**Figure 4.** Performance diagram (Roebber 2009) showing accumulated multiple day results for contingency table forecast verification metrics of the 16-12Z forecasts from the 24 days (5/7/2012 – 6/8/2012) of the 2013 SFE. The color code legend reveals the matching type of forecast (East team, West team, PP hindcast) with the probability thresholds labeled next to the corresponding scores.

### *d. Daily Distribution of Maximum Threshold CSI*

During the 2012 SFE, a similar evaluation of the experimental forecasts using calculations of CSI at the 5% probability threshold was performed. While this provided a quick means to document whether the event was captured by a low probability, the approach was incapable of diagnosing the best (or optimal) threshold for maximizing CSI. Such an investigation was possible for the 2013 SFE, with daily distributions of probability threshold for maximizing CSI and the corresponding CSI score displayed in Fig. 6.

In order to create the histogram in Fig. 6, the best probability values from the East team, West team, and PP hindcast were rounded down to the nearest fixed probability threshold (e.g., 20% classified under 15% bin). The frequency counts in Fig. 6 indicated that the highest scores were often at or just above the 15% probability threshold. It is also interesting to note that the best results sometimes extended into the 30% probability threshold bin, especially for PP hindcasts. Further, the maximum threshold CSI values were generally in the 0.2 to 0.4 range for the experimental team forecasts with a shift upward to around 0.6 for PP hindcasts (Fig. 6). Thus, a more complete representation of CSI (or any contingency table metric) was available by examining multiple probability thresholds than looking at just one.

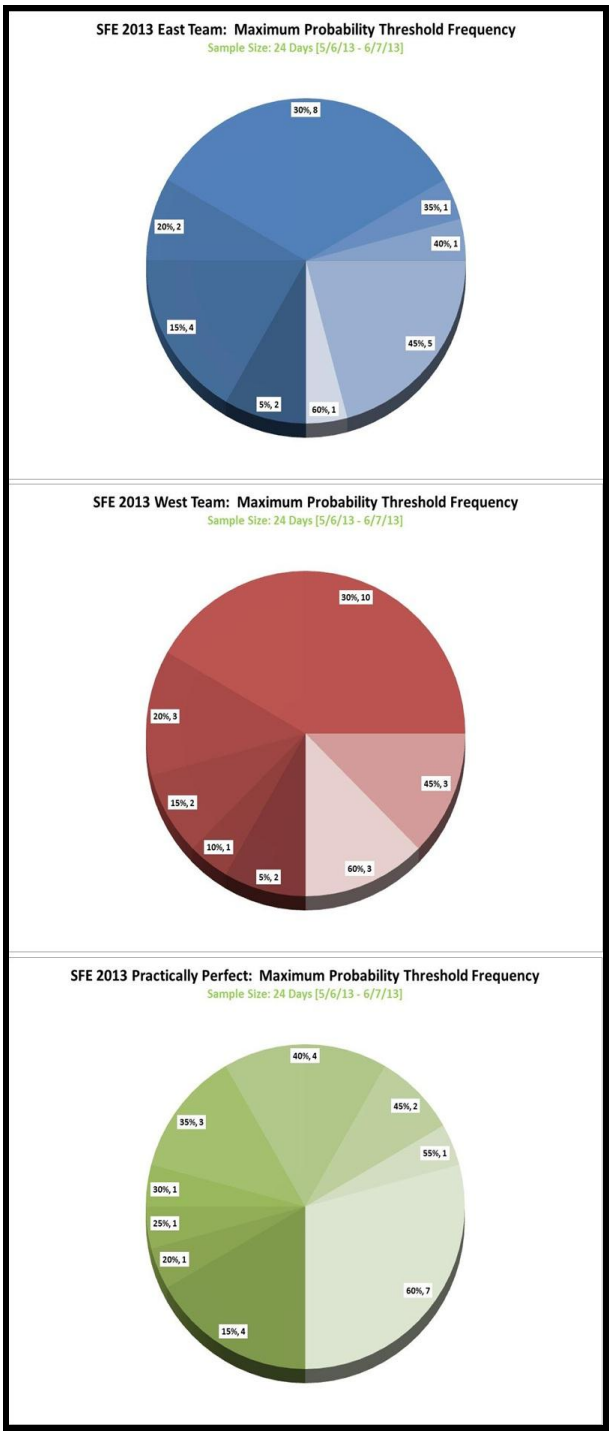


**Figure 5.** Pie chart showing the daily frequency for the maximum probability threshold for the 16-12Z forecasts from the 24 days (5/7/2012 – 6/8/2012) of the 2013 SFE. The top, middle, and bottom panels present the outcomes from the East team, West team, and PP hindcast, respectively. The continuous probabilities from PP were sorted into 5% increment bins which match those possible from the two teams. For the data labels, the count of occurrence is given next to each probability bin value in the range from 5% to 60%.

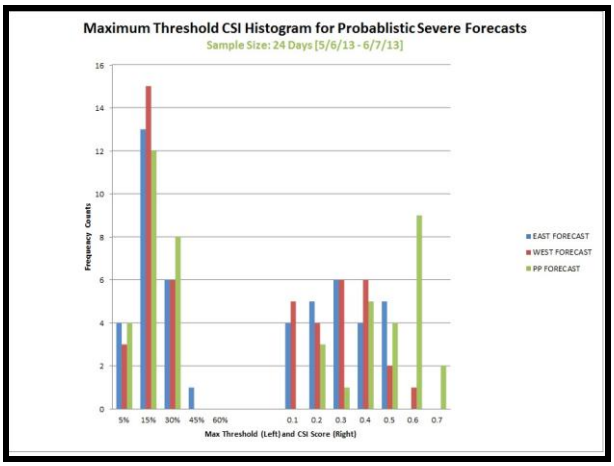


**Figure 6.** Histogram plot displaying maximum threshold CSI for the 16-12Z probabilistic severe forecasts from the 24 days during the 2013 SFE. East team, West team, and PP hindcast frequency counts for the best threshold for maximizing CSI and the resulting maximum CSI are given to the left and right side of the figure, respectively. The probability and CSI score bins are constructed as described in the text.

### *e. Daily Distribution of RelSkill*

A majority of the investigation thus far has been concentrated on just using contingency table verification metrics without understanding any context of the difficulty of the experimental forecasts. For this purpose, box-and-whisker plots for the daily RelSkill at each of the fixed probability thresholds and the threshold for which CSI maximized were produced (Fig. 7). In looking at the RelSkill of the forecasts issued during the SFE, both teams exhibited positive RelSkill for almost all of the days at both the 5% and 15% probability thresholds and more than half of the days at the 30% probability threshold. The RelSkill was rarely positive at the 45% and 60% thresholds, which is not unexpected given that these higher probability areas are not drawn to capture all of the reports (i.e., discriminate between occurrence and non-occurrence). Correspondingly, Fig. 7 showed the magnitudes for the median RelSkill with increasing thresholds increased from approximately 0.2 to 0.4, decreased to about 0.2, before trending below zero at the two highest probability thresholds. While most of the distributions were small and concentrated, the 30% threshold had the broadest distribution of RelSkill. As for the RelSkill calculated at the threshold for which CSI maximized, all of the results were similarly good as at the 15% probability threshold, but shifted slightly higher at all percentiles (Fig. 7).

It is worth noting that the results shown were conditional on how Equation **(1)** was formulated and would subsequently be sensitive to the strengths and weaknesses associated with the selection of CSI. For instance, the strong association between the two variables is illustrated in Fig. 8, in which the scatter plot reveals a high, positive correlation (R~0.8-0.9). Thus, this measure of skill will suffer when there are few hits in the forecast relative to the number of false alarms and misses. It would be interesting in the future to calculate RelSkill by applying a different verification metric (e.g., Bias, POD, etc.) as a comparison to the results obtained using CSI.

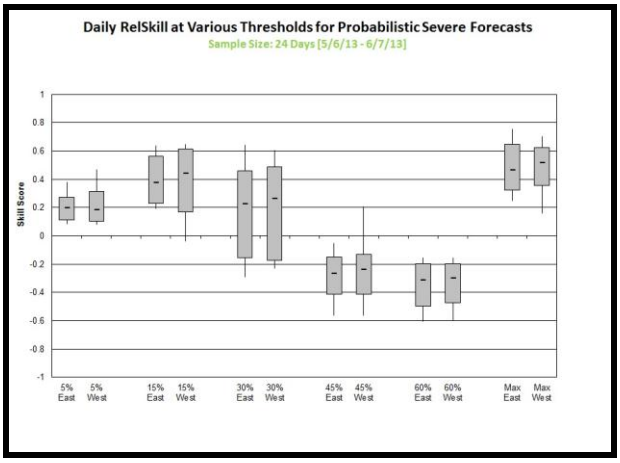


**Figure 7.** Box-and-whisker plots of daily RelSkill at various probabilistic thresholds for the 16-12Z severe weather forecasts issued by the East and West teams. Starting from left to right, results are calculated and displayed at 5%, 15%, 30%, 45%, 60%, as well as for the threshold at which CSI was maximized. The whiskers correspond to the 10th and 90th percentile rankings from the 24 days during the 2013 SFE.

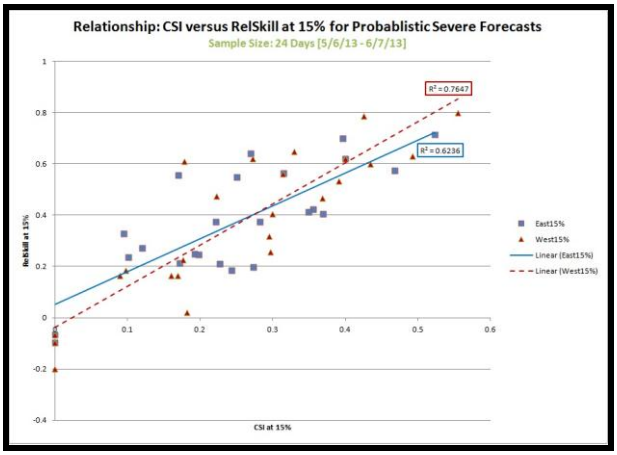


**Figure 8**. Scatter plot showing relationship between CSI and RelSkill values at the 15% probability threshold for the 16-12Z severe forecasts issued by the East and West teams during the 2013 SFE. Linear trend lines and the coefficient of determination are included.

### *f. Distribution of FSS*

The FSS was examined during the 2013 SFE because of its advantage in evaluating probabilistic type information in a straightforward manner. During the 2012 SFE, FSS had some of the highest scores calculated in the objective evaluation of high-resolution guidance. Further, these high scores were supported by it often being rated the most preferred metric by the participants (Melick et al. 2012). Similarly, excellent FSS values for the 2013 SFE experimental severe forecasts are indicated in Fig. 9. As such, nearly all of the daily results resided above 0.5 for both teams with tight distributions centered between 0.7-0.8. This assessment coincided with the observation that a substantial portion of the forecast probabilistic threat areas over the 24 days aligned themselves well with that from PP. Still, subjective appraisals of FSS suggest it may not distinguish subtle, but important differences in forecast performance.

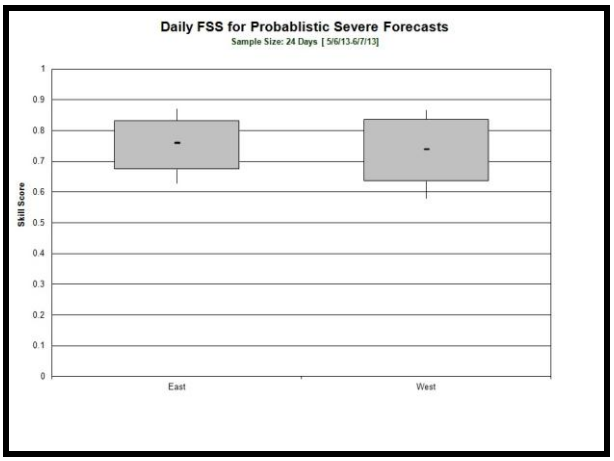


**Figure 9.** Box-and-whisker plots of daily FSS for the 16-12Z probabilistic severe weather forecasts issued by the East and West teams. The whiskers correspond to the 10th and 90th percentile rankings from the 24 days during the 2013 SFE.

### *g. Comparison of CSI, RelSkill, and FSS*

Some verification metrics (e.g., RelSkill at 30% in Fig. 7) analyzed in the 2013 SFE showed substantial variations in forecast performance while other measures showed less variability (e.g., FSS in Fig. 9) across the five weeks. In order to explore these differences, two case studies are offered from the 16-12Z forecasts on June 3rd (Fig. 10) and June 4th (Fig. 11). The statistical analysis revealed very good FSS with a slight increase of about 0.1-0.2 on the later

date, the consequence of the severe probability regions issued having slightly better resemblance to that of PP hindcasts. In terms of the 15% probability, scores from CSI (not shown) were about 0.17-0.18 for both team forecasts on both days. Alternatively, the 15% RelSkill (not shown) was much higher on June 3rd (0.56-0.61) compared to June 4th (about 0.22), when there were a few more LSRs (Table 1). The key distinction resides in the fact that the spatial coverage for the verifying reports was spread out more on June 3rd (compare Fig. 10 versus Fig. 11) which caused the upper bound from the baseline (i.e. PP hindcasts) to have a lower CSI (0.24 in contrast to 0.44). Consequently, the forecast was comparatively more challenging on June 3rd, and the teams were rewarded favorably from RelSkill even though the CSI values were similar on both days.

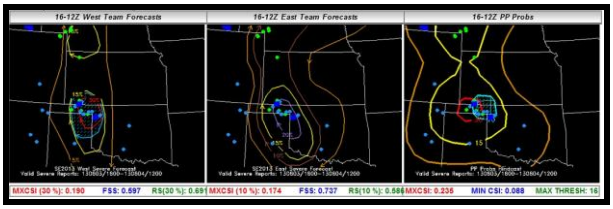


**Figure 10.** Same as in Fig. 1, except for the spatial plots and associated skill scores for June 3rd, 2013. The surrounding webpage information has been eliminated in order to focus in on the details for the case study.

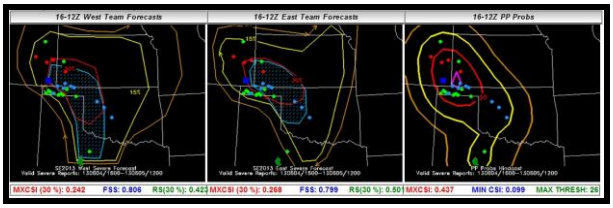


**Figure 11.** Same as in Fig. 1, except for the spatial plots and associated skill scores for June 4th, 2013. The surrounding webpage information has been eliminated in order to focus in on the details for the case study.

The ability of the objective metrics to discriminate based on the magnitude of the severe weather event was also explored. For this purpose, the values from each of the forecast verification metrics were sorted based on ranking each day by the number of LSRs recorded. Subsequently, box-and-whisker diagrams were created for 15% CSI, 15% RelSkill, and FSS for the bottom 12 LSR tally days as well as for the top 12 LSR tally days (Fig. 12). In the comparison of less active severe convective days to more active ones, an upward shift in the statistical distributions was noted for all three metrics at most percentile thresholds. The values for FSS exhibited the smallest increase, yet were substantially higher compared to the other two verification metrics regardless of the number of LSRs (compare panels in Fig. 12). In terms of 15% CSI and 15% RelSkill, very small positive to negative scores for the 10th to 25th percentile were reserved only for those days with a minimal collection of reports. Not surprisingly, an evaluation in this manner revealed the participants performed better in their probabilistic forecasts for more active severe weather days. The large distribution in the RelSkill stayed nearly the same regardless of the amount of severe weather, which presumably indicates that using PP hindcasts as a baseline reference is attempting to make days with different levels of severe activity more comparable.

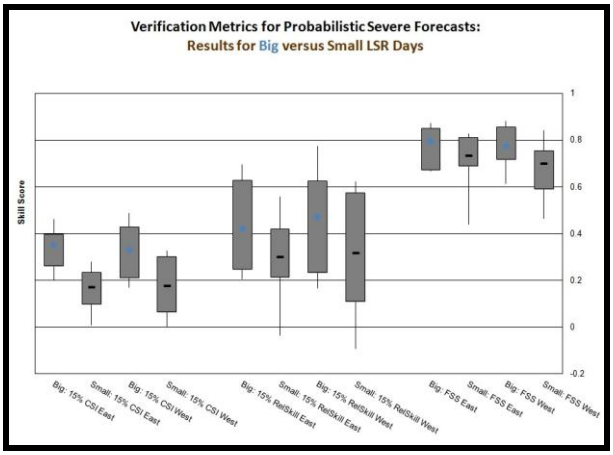


**Figure 12.** Box-and-whisker plots of daily 15% CSI, 15% RelSkill, and FSS for the 16-12Z probabilistic severe weather forecasts issued by the East and West teams. Results for the bottom 12 (small) LSR tally days are presented alongside the top 12 (big) LSR tally days for comparison. The whiskers correspond to the 10th and 90th percentile rankings during the 2013 SFE.

### *b. Participant Feedback*

Another goal of the research work was to compare the objective results to the participant feedback from the survey questions. Figure 13 present tallies gathered from the responses on 21 days for subjective evaluations of the experimental severe forecasts. Both the East and West team forecasts for the 16-12Z time period were rated "Fair" to "Good" for more than seventy five percent of the forecasts. In order to relate these assessments to that of the forecast verification metrics, a cursory examination of Fig. 12 shows that the daily scores were at the very least reasonably favorable a greater

part of the 2013 SFE. Additionally, a follow-up inquiry in Fig. 14 revealed that the participants usually "Agreed" that the RelSkill matched their subjective impressions of forecast performance. It should be noted, though, the sample sizes for this survey question were slightly smaller (e.g., 15 and 19 days) compared to the previous one in Fig. 13.

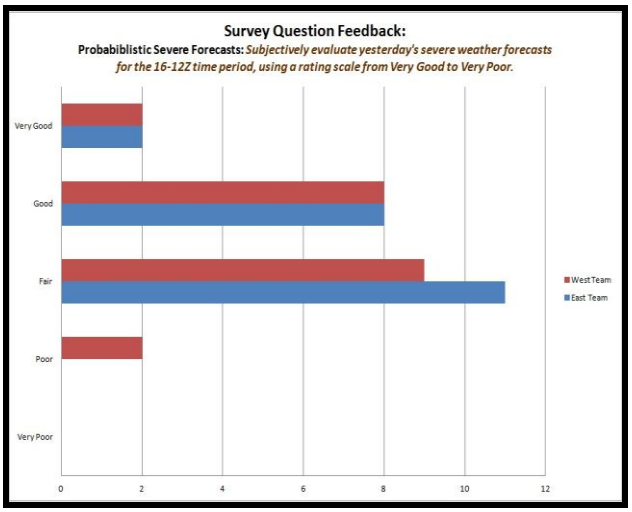


**Figure 13.** Participant feedback tallies gathered during 2013 SFE daily activity evaluation. The results obtained from the survey question covered subjective evaluations of the severe weather forecasts for the 16-12Z time period. The wording of the question is given in italics in the titles and the sample size was 21 days.

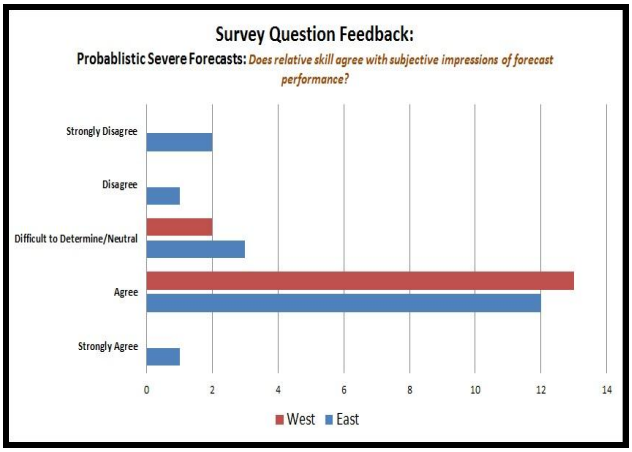


**Figure 14.** Same as in Fig. 13 except for a survey question relevant to comparing the results from the relative skill score against the subjective evaluations of forecast performance. The sample sizes were smaller compared to Fig. 13 at 19 and 15 days for the East and West teams, respectively.

## 4. Summary and Conclusions

SPC conducted objective verification of experimental severe forecasts during the 2013 SFE in near real-time. This second attempt built upon the success from the 2012 SFE in testing the value of selected verification metrics in relation to subjective evaluations. Besides examining CSI and FSS from the prior year, the RelSkill was added to the suite of forecast verification metrics examined daily in the HWT. In this approach, PP hindcasts were constructed to provide a valuable baseline to measure the skill of the probabilistic severe forecasts during the 2013 SFE.

The process for conducting an effective evaluation of the 16-12Z probabilistic forecasts mimicked that from the 2012 SFE. Specifically, time matched spatial plots of forecasts and observations were displayed on webpages linked from the 2013 SFE website for visual comparison. Skill scores were also calculated for each forecast time period to be viewed with the appropriate images or to be examined via table summaries. By incorporating more than one forecast verification metric and then comparing these statistical results to the participant feedback, a more complete picture was obtained in the evaluation process.

One notable finding was that RelSkill provided unique information regarding forecast performance over that of traditional forecast verification metrics. Since RelSkill includes the PP hindcast as a baseline reference, some measure of the difficulty of the forecast is included in the metric. The forecasts at the lower probability thresholds (i.e. 5% and 15% probability contours) nearly always had positive RelSkill. As demonstrated by the case study comparisons, this measure of skill was sensitive to multiple factors, including the spatial distribution of LSRs. Finally, it was noted that generally better statistical results occurred on days with more severe weather reports, something which was observed with both CSI and FSS as well.

Another conclusion was that the subjective evaluations during the five-week period were generally consistent and agreed with the statistical results. Participants usually rated the severe forecasts as "Fair" to "Good" with verification metrics being generally favorable, especially at the lower probability thresholds. Consequently, the encouraging results of performing objective verification have persisted for the last two years in the HWT (and locally at SPC) and support continued efforts in future SFEs.

*Acknowledgements*. This extended abstract was prepared by Chris Melick with funding provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement #NA11OAR4320072, U.S. Department of Commerce. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of NOAA or the U.S. Department of Commerce.

## REFERENCES

Bright, D.R. and M.S. Wandishin, 2006: Post processed short range ensemble forecasts of severe convective storms. Preprints, 18th Conf. Probability and Statistics in the Atmos. Sciences, Atlanta GA, Amer. Meteor. Soc., 5.5.

Brooks, H. E., M. Kay, and J. A. Hart, 1998: Objective limits on forecasting skill of rare events. Preprints, 19th Conf. on Severe Local Storms, Minneapolis, MN, Amer. Meteor. Soc., 552–555.

Clark, Adam J., and Coauthors, 2012: 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.

Hitchens, N.M., H.E. Brooks, and M.P. Kay, 2013: Objective limits on forecasting skill of rare events. Wea. Forecasting, 28, 525–534.

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.

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.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608.

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.

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