76 Exploration of the NSSL Maximum Expected Size of Hail (MESH) Product for Verifying Experimental Hail Forecasts in the 2014 Spring Forecasting Experiment

Christopher J. Melick*[,1,2], Israel L. Jirak[1], James Correia Jr.[1,2], Andrew R. Dean[1], and Steven J. Weiss[1]

[1]*NOAA/NWS/NCEP Storm Prediction Center*
[2]*Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma*

1. Introduction and Motivation

The 2014 Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (SFE) operated for a 5-week period (5 May – 6 June) at the National Weather Center in Norman, OK. The Storm Prediction Center (SPC) and National Severe Storms Laboratory (NSSL) jointly conduct the SFE every spring to test emerging concepts and technologies for improving the prediction of hazardous convective weather. More importantly, efforts to bridge the gap between research and operations continue as a key component of the HWT with each year designed to build successful collaborations. For more background details, historical summaries of the annual SFE since 2000 can be found in both Kain et al. (2003) and Clark et al. (2012).

Objective forecast verification of experimental severe weather forecasts was conducted for the third consecutive year during the 2014 SFE. The next-day subjective evaluations by SFE participants have been found to be more complete by incorporating forecast verification metrics in near real-time as opposed to performing statistical assessments only in a post-experiment fashion. As discussed in Melick et al. (2013), the subjective evaluations from the participants were generally consistent and agreed with the statistical results. The comparisons were facilitated by creating time-matched spatial plots of forecasts and observations for display on webpages linked from the SFE website. Skill scores could then be viewed for each forecast time period with the appropriate images and/or examined via table summaries. Preliminary local storm reports (LSR) have traditionally served as the primary verification dataset when computing objective performance metrics.

**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

Subjective assessments of the probabilistic forecast products created by the participants during the 2014 SFE were similar to what had been done in previous years. For the first time, however, individual hazard (tornado, wind, hail) probabilistic forecasts were produced by the Severe Desk led by the SPC instead of a single probabilistic forecast for total severe. This study addresses the performance of the experimental probabilistic severe hail forecasts in exploring additional verification datasets instead of solely using LSRs.

Images from multiple observation sources were made available for next-day subjective comparisons during the 2014 SFE. Of these, radar-derived maximum expected size of hail (MESH; Witt et al. 1998) from NSSL served as a valuable surrogate to document the occurrence of hail, especially in low-density population areas where there may be a scarcity of LSRs. A key goal is to further explore gridded MESH fields as an alternative dataset to verify the experimental hail forecasts. This also provides the opportunity to compare results in the objective forecast verification to those obtained from traditional LSRs.

2. Data and Methodology

2.1. Forecast and Verification Datasets

SFE team participants from the SPC Severe Desk produced probabilistic experimental forecasts for severe hail (hail ≥ 1" in diameter) that were valid within 25 miles (~40-km) of a point, as defined in SPC operational convective outlooks. These were made over a mesoscale area of interest that was moved each day to correspond to the greatest and/or most challenging severe threat area. Table 1 lists the surface weather stations that served as daily movable centerpoints. In order to generate the product, the team used the same probability threshold contours as the SPC Day 1 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). However, computations of forecast verification metrics were specifically restricted here to *just any severe hail occurrence.*

All experimental probabilistic forecasts considered in the current investigation covered the 16-12 UTC period for 23 weekdays from 5 May – 6 June (with no activities on Memorial Day). Additional forecasts of higher temporal resolution (i.e., 3-hr intervals; 18-21, 21-00, and 00-03 UTC) 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. Initially, verification was computed utilizing LSRs received from the National Weather Service forecast offices covering the valid forecast period. In addition, the retained MESH product files were on a 0.01^{o} Latitude x 0.01^{o} Longitude grid in a format where the grid points represent 60-minute maximum values (i.e. hourly MESH tracks). Unfortunately, a portion of these data were missing during the 2014 SFE, and only MESH files from 17 days were completely available. As a result, only matching time periods from both sets of verification data were utilized to conduct an appropriate comparison.

2.2. Contingency Table

Severe hail events were defined for both the forecasts and observations in order to perform the objective evaluation. These events were determined by placing all 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). Following the process described in Melick et al. (2013), grid-point probability values from the experimental forecast contours were obtained using a graph-to-grid routine in GEMPAK (GEneral Meteorological PAcKage; desJardins et al., 1991). This algorithm 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 completing format conversion, binary (yes/no) event grids were created from the probabilistic information by specifying various thresholds to define the forecast area.

The procedure for constructing observed severe hail objects varied depending upon the verification dataset considered. In the case of LSRs, if ≥ 1 severe hail report occurred within a 40-km radius of influence (ROI) of the grid box during the 16-12 UTC period, the grid box was recorded as a severe hail event. On the other hand, the situation was not as straightforward for the MESH. First, 20-hour maximum MESH tracks covering the 16-12 UTC period were obtained by taking the maximum value at each grid point from the individual hourly fields. In addition, a separate filtered grid to eliminate isolated hail pixels was produced by using a two-dimensional Gaussian smoother (sigma=0.01^{o}) on the raw MESH tracks. A 40-km ROI neighborhood maximum was then applied to the high resolution analyses, thereby allowing a suitable point to interpolate both sets of MESH tracks (Raw and Filtered methods) to the NCEP 212 40 km common grid. As a final quality control check to remove potentially spurious or unrepresentative MESH data points, the existence of thunderstorms was confirmed via cloud-to-ground (CG) lightning flashes from the National Lightning Detection Network (NLDN). Specifically, MESH values were retained as long as ≥ 1 flash occurred within the 40-km grid box over the corresponding 20-hour period. Finally, separate Raw and Filtered datasets of MESH-derived severe hail events were created by determining if MESH ≥ 29 mm (see Cintineo et al. 2012 regarding threshold selection) at each grid point.

A direct grid-point-to-grid-point comparison between the forecasts and observational datasets was used to develop a 2x2 contingency table (Wilks 2006). From this verification approach, counts of hits, misses, false alarms, and correct nulls were obtained and standard verification metrics computed (e.g., Critical Success Index [CSI]) for all of the fixed SPC thresholds for severe hail (5, 15, 30, 45%), except for the 60% probability threshold, which was not forecast during the 2014 SFE. For the statistical analysis, a mask was also applied to include only grid points over the contiguous United States within the daily mesoscale "area of interest".

	Center-point
Forecast Date[YYMMDD]	Station Name, State (3-Char ID)
140506	Norfolk/Stefan Fld, NE (OFK)
140507	Fort Sill, OK (FSI)
140508	Carroll, IA (CIN)
140509	Poplar Bluff, MO (POF)
140514	Jackson/J. Carroll, KY (JKL)
140515	Lynchburg/P. Gleen, VA (LYH)
140516	Texarkana Rgnl/Webb, AR (TXK)
140519	Ainsworth Municipal, NE (ANW)
140520	Grand Rapids Intl, MI (GRR)
140521	Parkersburg/Wilson, WV (PKB)
140522	Baltimore/Wash Intl, MD (BWI)
140523	Charlotte/Douglas, NC (CLT)
140527	Waco-Madison Cooper, TX (ACT)
140528	Baker Municipal, MT (BHK)
140603	Omaha/Eppley Field, NE (OMA)
140604	Paducah/Barkley, KY (PAH)
140605	Springfield Muni, MO (SGF)

Table 1. Description of the surface weather stations selected for each of the 17 days as center-points during 2014 SFE. All of the daily evaluations were restricted to a mesoscale "area of interest" for possible severe convection. This domain was movable to locations in the eastern and central United States.

2.3. Practically Perfect Hindcasts

Melick et al. (2013) utilized a technique in their objective verification which relates a meaningful baseline to assess skill in severe weather forecasts using the collection of LSRs at SPC. As defined in Brooks et al. (1998), "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 grid box. This produces a probabilistic field which is considered to be consistent with what a forecaster would produce given prior (perfect) knowledge of the observations. Analogous to the LSRs, PP hindcasts were also constructed for MESH by applying the same smoother to the binary event grids (both Raw and Filtered). Thus, the utility of using MESH as a verification dataset will be partially judged by comparing PP areas against those created from LSRs.

2.4. Fractions Skill Score

The verification metrics discussed thus far have been constrained to evaluating whether or not a severe hail event was predicted and whether or not a severe hail event occurred. Instead of setting a threshold and converting the probabilistic forecast into a binary one, the PP hindcast can also serve as the verifying dataset. In this case, the probability fields from the experimental forecasts and PP hindcasts from LSRs and both versions of MESH are compared directly by calculating the fractions skill score (FSS; Schwartz et al. 2010), which is a variation on the Brier skill score. The range in FSS is from 0 to 1, with a value of 1 indicating a perfect forecast and a value of 0 revealing no skill and without overlap in non-zero probabilities. Similar to CSI, FSS was computed for the 17 day sample.

2.5. Reliability Diagram

The performance of probability forecasts can also be assessed by comparing the observed relative frequency as a function of forecast probability to determine statistical reliability. A reliability diagram (Wilks 2006) provides a visual means to understand properties of the probabilistic forecasts relative to "perfect reliability", a 1:1 diagonal line on the reliability diagram. For this application, the probability values from the experimental severe hail forecasts were grouped into five bins (0%, 5%, 15%, 30%, 45%). Then, counts of grid points with one or more severe hail reports from LSRs were determined for each probability bin and summed over all of the days. For MESH data, an equivalent methodology resulted in tallies of grid points with hail sizes $\geq$ 29 mm. 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 the forecast counts to the observation counts for each probability bin produced the relative frequency for the observations (i.e., displayed on the ordinate in the reliability diagram).

3. Results

3.1. Accumulated Results: Contingency Table

Roebber (2009) demonstrated that multiple verification metrics derived from the contingency table could be summarized on one graph called a performance diagram. This includes information on probability of detection (POD), false alarm ratio (FAR), frequency of hits (FOH), bias, and CSI for an integrated diagnosis of forecast accuracy without needing to examine separate tables or graphs. Figure 1 was constructed using the accumulated, multiple-day results based on the verification datasets discussed in Section 2. The overall statistics are shown at all probability thresholds for the 16-12 UTC experimental hail forecasts and verified using LSRs, Raw MESH, and Filtered MESH.

One evident feature in Fig. 1 is the CSI values were maximized generally at the 15% severe hail probability threshold. At this threshold, CSI values either ranged from slightly over 0.2 (LSRs, Filtered MESH) to exceeding 0.3 (Raw MESH). This behavior matches results identified in Melick et al. (2013) for experimental forecasts created in the HWT for total severe weather. Similarly, a very large POD (near or exceeding 0.9) was apparent at the 5% level and high FOH (low FAR) values above (below) 0.5 for forecast probabilities at or above 30%. More importantly, though, using MESH resulted in a similar POD but higher FOH compared to using LSRs as the verifying database. For instance, a more significant discrepancy occurs at the 15% probability threshold with a doubling of FOH for the Raw MESH approach. Statistical results using Filtered MESH were more comparable to those based on LSRs but still showed a slight improvement. Finally, caution should be taken in interpreting metrics at the 45% probability threshold (Fig. 1), as the sample size was very small with input only from one forecast date (3 June).

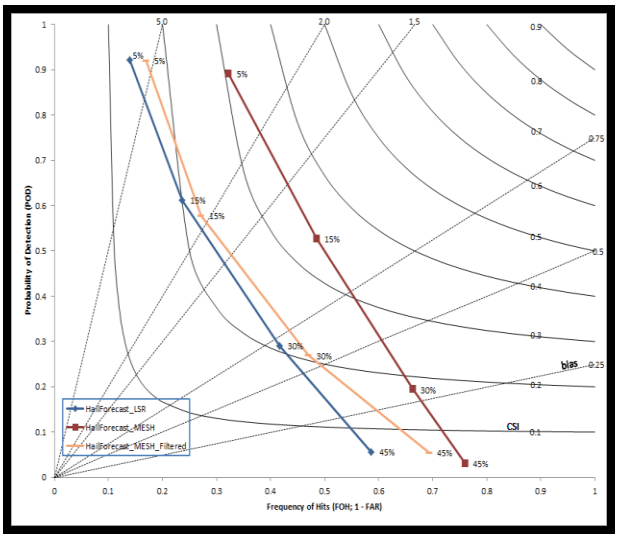

Fig. 1. Performance diagram (Roebber 2009) showing accumulated multiple day results for contingency table forecast verification metrics of the 16-12 UTC probabilistic severe hail forecasts from 17 days (5/5/2014 – 6/6/2014) of the 2014 SFE. The color code legend reveals the matching type of verification (LSR, Raw MESH, Filtered MESH) with the probability thresholds labeled next to the corresponding scores.

3.2. Accumulated Results: Reliability Diagram

Figure 2 reveals the observed frequency of severe hail events for the experimental forecasts based on the three types of verification data. Across the five forecast probability bins, the forecasts evaluated using either LSRs or Filtered MESH were nearly reliable for probabilities up to and including the 15% bin where sample sizes were on the order of 1,000 grid points (Fig. 2). The small under-prediction bias increased at the 30% threshold where the observed frequency of occurrence was closer to 40%. Further, there was a larger under-prediction (i.e., more observed events than forecast events) at all thresholds for the Raw MESH dataset. For instance, even at the lowest threshold (5%), the relative observed frequency was much higher around the 20% value. For the highest probability bin (45%), the findings were inconclusive once again as the total number of forecast grid points had decreased to less than a hundred (Fig. 2).

The findings presented here indicate that the reliability of the experimental probabilistic severe hail forecasts was strongly dependent upon the observed dataset utilized in the evaluation. The highest reliability in probabilities was noted for LSRs, followed by the Filtered MESH verification (Fig. 2). The abundance of severe hail observations from Raw MESH resulted in higher skill using standard forecast verification metrics (Fig. 1), but also

resulted in much higher spatial coverage of observed severe hail events. In addition, the very isolated nature of some of the radar-derived Raw MESH objects highlighted challenges in using very high space/time resolution data compared to traditional coarser resolution LSRs in verifying the experimental hail forecasts.

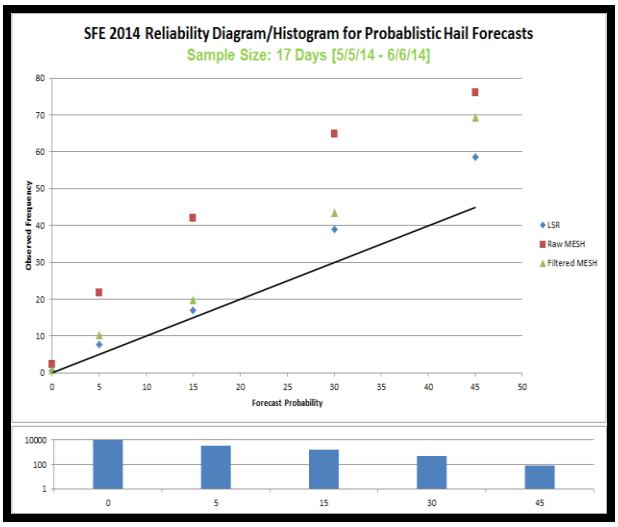

Fig. 2. Reliability diagram for 16-12 UTC probabilistic severe hail forecasts using accumulated grid point tallies over 17 days (5/5/2014 – 6/6/2014) of the 2014 SFE. The inset histogram displayed below gives the forecast subsample sizes computed each of the forecast probability bins (0%, 5%, 10%, 15%, 30%, and 45%,). 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 verification (LSR, Raw MESH, Filtered MESH).

3.3. *Daily Results: Severe Hail Events*

Characteristics from the individual daily 20-hr time periods is not apparent from the summary type analysis just presented. In order to consider these details, daily scatter plots relating the various methods for constructing severe hail events are shown in Fig. 3. Specifically, counts of LSRs were matched against tallies from either Raw MESH or Filtered MESH datasets. Across each daily mesoscale area of interest, both MESH approaches tended to produce more grid point objects compared to LSRs (Fig. 3). This finding is consistent with the accumulated results presented in Fig. 2, and is more pronounced for the Raw MESH as demonstrated by several dates where the coverage of Raw MESH was 1.5 to 2 times more than that from LSRs. This assessment for Raw MESH is also supported in Fig. 3 by the comparative slopes from the fitted linear relationship as well as the elevated FOH in the accumulated results (Fig. 1).

The degree of spatial overlap in defining the verification at each grid point is examined in Fig. 4, which highlights how often LSRs corresponded with areas of MESH objects using running trends during the 2014 SFE. Interestingly, the coincident sum of Raw MESH values at or above 29 mm was slightly less than the overall number of LSRs, with more of a difference noted for the Filtered MESH dataset (top of Fig. 4). The high spatial agreement, especially in the case of the former, was persistent over the 17 days as calculated percentages (bottom of Fig. 4) remained largely between 85-90% and 65-70% for Raw and Filtered MESH, respectively.

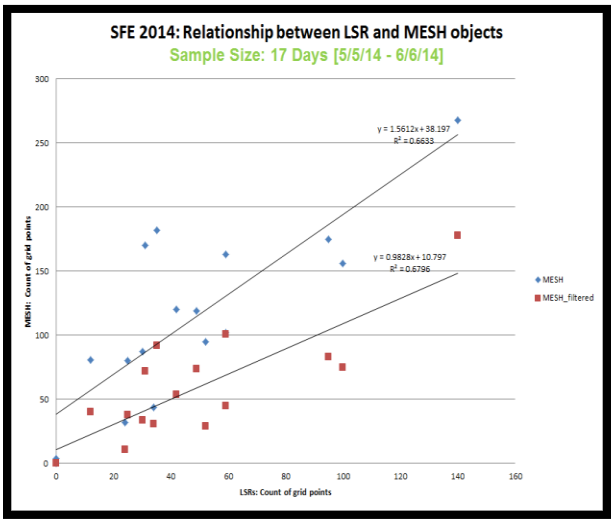

Fig. 3. Scatter plots showing relationship between observed hail objects identified by LSRs versus those from Raw and Filtered MESH. Daily matched counts for the 16-12 UTC forecast period on 17 days were determined across the limited areas of interest. Linear trend lines and the coefficient of determination are included.

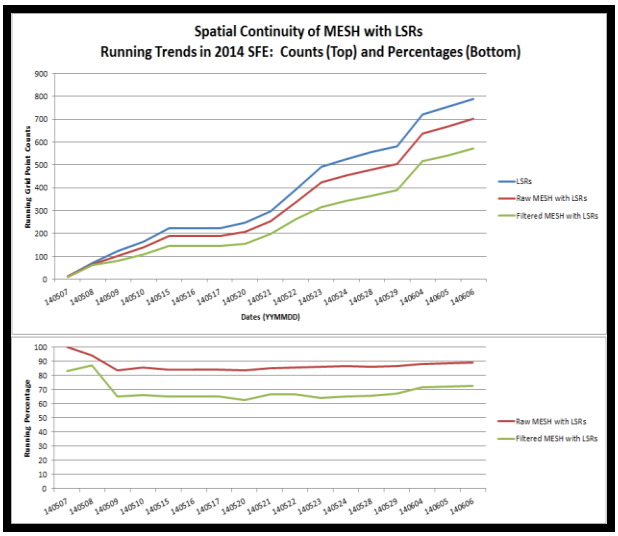

Fig. 4. Plots showing spatial agreement of MESH with severe hail LSRs during the course of the 2014 SFE. The top panel shows the running tally of grid point counts from severe hail LSRs, Raw MESH with severe hail LSRs, and Filtered MESH with severe hail LSRs. The bottom panel shows the cumulative percentage of severe hail LSR grid points with either Raw or Filtered MESH objects. The corresponding 17 dates (in YYMMDD format) are indicated in progression along the x-axis.

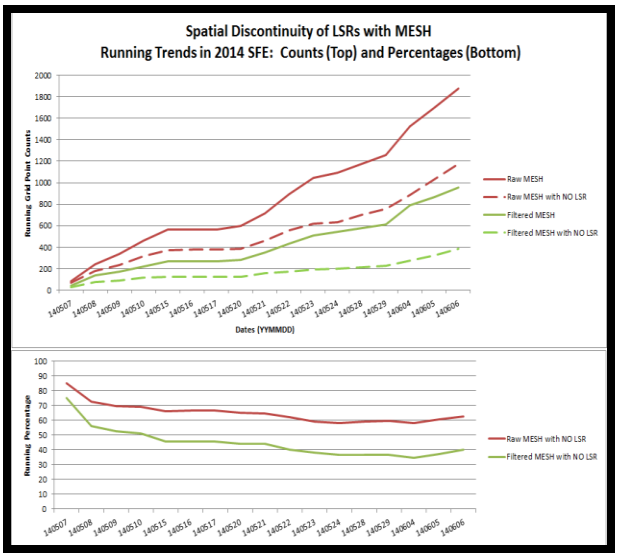

Fig. 5. Plots showing spatial disagreement of severe hail LSRs with MESH objects during the course of the 2014 SFE. The top panel shows the running tally of grid point counts from Raw MESH and Raw MESH without a severe hail LSR as solid lines and the Filtered MESH and Filtered MESH without a severe hail LSR as dashed lines. The bottom panel shows the cumulative percentage of Raw and Filtered MESH grid points without a severe hail LSR. The corresponding 17 dates (in YYMMDD format) are indicated in progression along the x-axis.

Alternatively, Fig. 5 reveals how often LSRs did not correspond with areas of MESH objects. As seen earlier (e.g., Fig. 3), the MESH verification had higher coverage of observations than LSRs (compare sums in top of Fig. 5 to top of Fig. 4). Thus, it was expected to find that at many of the grid points where MESH objects were located, no severe hail reports were identified. In fact, the relative percentages in the bottom of Fig. 5 reveal that around 60% and 40% for the Raw and Filtered approaches respectively do not have a corresponding LSR. The sharp contrast in absolute numbers should also be emphasized as well here. The aggregate count of Raw MESH grid points with no LSR was over 1000 by the end of the 17 days, whereas the Filtered MESH method only totaled a few hundred (top of Fig. 5).

3.4. Daily FSS Distributions

The FSS is used to evaluate spatial correlations between forecast and observed probabilistic areas of severe hail. SFE participants had previously regarded this metric using a spatial neighborhood approach to be most useful in objective evaluations of simulated reflectivity from convection-allowing model guidance (Melick et al. 2012). This was due, in part, to the larger FSS values associated with forecasts that were subjectively considered to correspond well with observations. In addition, Melick et al. (2013) noted these favorable FSS results extended to the experimental probabilistic forecasts created in the HWT (e.g., in the 0.7-0.8 range). Figure 6 reveals that this trend continued for the 2014 SFE experimental severe hail forecasts, although values were not as high as the prior year. A majority of the FSS daily scores occurred above 0.5, with a clustering around 0.5-0.7 for the PP hindcast using LSRs or Filtered MESH. Still, a relative 25% decrease in the inter-quartile range is noted in the Raw MESH. Interestingly, the contingency table statistics (Fig. 1) using the Raw MESH were higher than the other verification datasets while the probabilistic verification using the Raw MESH (Figs. 2 and 6) was lower than the other

datasets. In order to examine the correspondence between the verification datasets, Fig. 6 also shows FSS as determined by comparing different PP probabilistic areas. From this perspective, the resemblance in forecast performance for LSRs versus Filtered MESH is confirmed by the fact that the median FSS value between the two was near 0.9 (Fig. 6). In contrast, much more discrepancy is suggested when relating PP probabilities from LSRs to Raw MESH as the daily score distribution is broader and lower at every percentile ranking, especially the lower quartile.

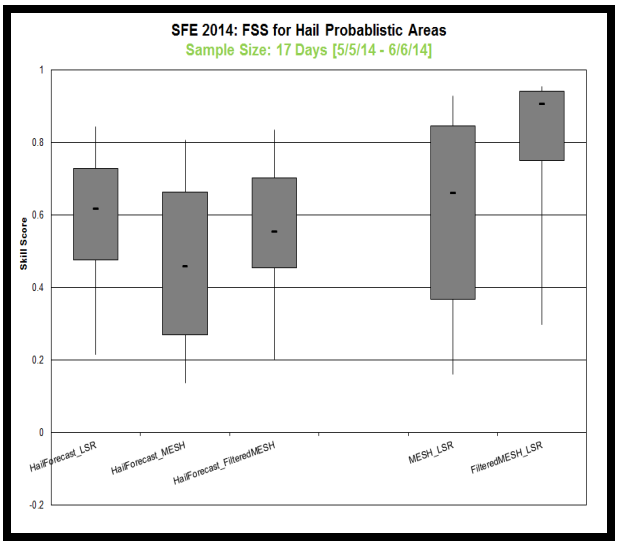

Fig. 6. Box-and-whisker plots of daily FSS for the 16-12 UTC probabilistic severe hail forecasts based on verification (LSR, Raw MESH, and Filtered MESH) as well as the direct comparison of 16-12 UTC PP probabilistic areas derived from LSRs and MESH (Raw and Filtered). The whiskers correspond to the 10th and 90th percentile rankings from the 17 days during the 2014 SFE.

3.5. Case Studies: Verification Comparisons

The bulk statistics examined have indicated that MESH provided a higher occurrence of severe hail within the forecast domain region compared to the traditional LSRs. To investigate further, data from 6 May (Fig. 7) and 4 June (Fig. 8) during the 2014 SFE are shown. Spatial plots were created to highlight the contrast in PP analyses obtained from LSRs (left panel), Raw MESH (middle panel), and Filtered MESH (right panel) along with the underlying verification and experimental hail forecast products.

Figure 7 revealed limited PP hindcast verification at 5% based on isolated hail LSRs, but MESH tracks over Minnesota, Nebraska, and Wyoming showed greater spatial coverage. The 30% contours based on Raw MESH was reduced to 15% for the Filtered MESH (compare middle and right panels in Fig. 7). On the 4 June case, all LSRs occurred within the 5% forecast contour but the PP hindcast was disjointed (left panel of Fig. 8). Instead, the connected probability area from MESH PP hindcast shown in the middle and right panels of Fig. 8 appeared more representative given the distribution of the MESH tracks. One of the main challenges in using Raw MESH for verification was the tendency to produce much larger PP spatial areas at higher probabilities compared to LSRs. However, after filtering the MESH, the 15% PP corresponded more closely with the sparser coverage of LSRs in Kentucky, as opposed to the original 45% and embedded 60% Raw MESH contours (compare middle and right panels of Fig. 8).

In summary, both case studies illustrated that numerous MESH tracks were present, sometimes displaced from LSR locations. From the verification datasets explored, using separate verification approaches for severe hail revealed some spatial overlap, but still distinct differences in placement and magnitude of the PP hindcast probabilities. Thus, including an alternative source of observations served to both substantiate LSR events and to fill report gaps in low-density population areas, which has been a known weakness of the LSRs. Still, because of the very high-resolution automated nature of the MESH, additional convective-scale details are identified that can be difficult to confirm, especially if they are very isolated in time/space. Thus, it is recommended that quality control to filter the MESH data be conducted before using it to verify probabilistic severe hail forecasts.

4. Summary and Conclusions

Objective verification of experimental forecasts continued in near real-time for the third consecutive year during the 2014 SFE. While probabilistic products for total severe (tornado, wind, hail) weather had been created in the HWT in the past, forecasts for individual severe hazards were evaluated for the first time. This study specifically focused on the experimental severe hail forecasts created by the HWT participants and considered alternative sources for verification. Similar to prior years, LSRs provided the primary observations for next-day subjective and objective evaluations, but MESH plots were also created to test alternative verifying data.

A more formal comparison of skill using MESH was conducted post-SFE where a similar procedure as used with LSRs was followed. In an attempt to eliminate isolated pixels in the high resolution MESH tracks, separate filtered grids were also created. After the observed events were defined using either LSRs or MESH (Raw and Filtered), separate sets of forecast verification metrics were computed for each of the 16-12 UTC daily forecast periods. In addition, PP hindcasts were created to provide valuable baselines to measure the skill of the probabilistic severe hail forecasts during the 2014 SFE.

The results show the high-resolution MESH to have a much greater number of observed objects in contrast to LSRs resulting in lower false alarms and higher CSI at all probability thresholds. The best reliability was noted for LSRs, followed by the Filtered MESH verification. However, this finding may reflect the fact that SPC severe weather forecasters have been "calibrated" by LSRs over the years in the issuance of probabilistic forecasts. Thus, it is not surprising that the experimental forecasts issued during the SFE are more reliable when LSR data are used in the verification process. Similarly, the Filtered MESH has been created, in part, to more closely correspond to LSR-based PP hindcasts compared to those created using the Raw MESH, so forecast reliability may be similar when Filtered MESH data are used.

The reliability data also revealed an under-prediction bias for all verification approaches, but this was much more apparent for the Raw MESH datasets at all probability thresholds, consistent with the higher coverage of Raw MESH hail objects. Accordingly, a higher FSS was computed for the Filtered MESH than the Raw MESH, as it more closely corresponded with the forecast and PP analyses from LSRs. This similarity was illustrated in two example cases where the Filtered MESH tended to be more compatible with the geographical extent and magnitude of severe hail episodes identified by LSRs.

In conclusion, the evaluations presented here demonstrated the importance of including an alternative source for forecast verification. In the case of the experimental hail forecasts, the implied skillfulness varied substantially depending upon the use of LSR or MESH data for verifying observations. Overall, the MESH tracks appeared to be potentially useful in identifying events in low-density population areas, and as an independent dataset to supplement hail LSRs. Future efforts will incorporate forecast verification metrics from both LSRs and MESH in a side-by-side diagnosis in subsequent SFE years. In addition, methods to combine both verification datasets will be investigated.

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.

Cintineo, J.L., T.M. Smith, V. Lakshmanan, H.E. Brooks, and K.L. Ortega, 2012: An objective high-resolution hail climatology of the contiguous United States. *Wea. Forecasting*, **27**, 1235–1248.

Clark, A.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.

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.

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.

_____, 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.

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.

Witt, A., M. D. Eilts, G. S. Stumpf, J. T. Johnson, E. D. Mitchell, and K. W. Thomas, 1998: An enhanced hail detection algorithm for the WSR-88D. *Wea. Forecasting*, **13**, 286–303.

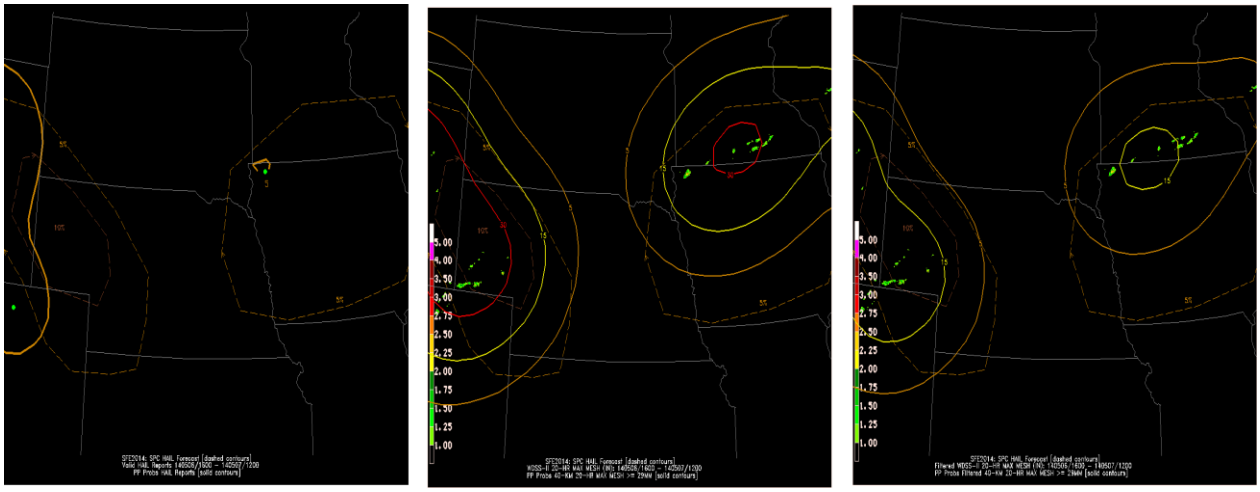

Fig. 7. Case study spatial plots highlighting differences in verification for the forecast date of May 6th, 2014. The probabilistic severe hail forecasts for 16-12 UTC are valid for a mesoscale "area of interest" and are displayed in all panels as dashed contours. The left panel shows PP probability contours derived from LSRs with the verifying severe hail reports overlaid on top of the plot. The middle and right panels shows PP probability contours derived from Raw and Filtered MESH, respectively. Also, the corresponding, original 20-hour maximum MESH tracks from both approaches are overlaid on both plots.

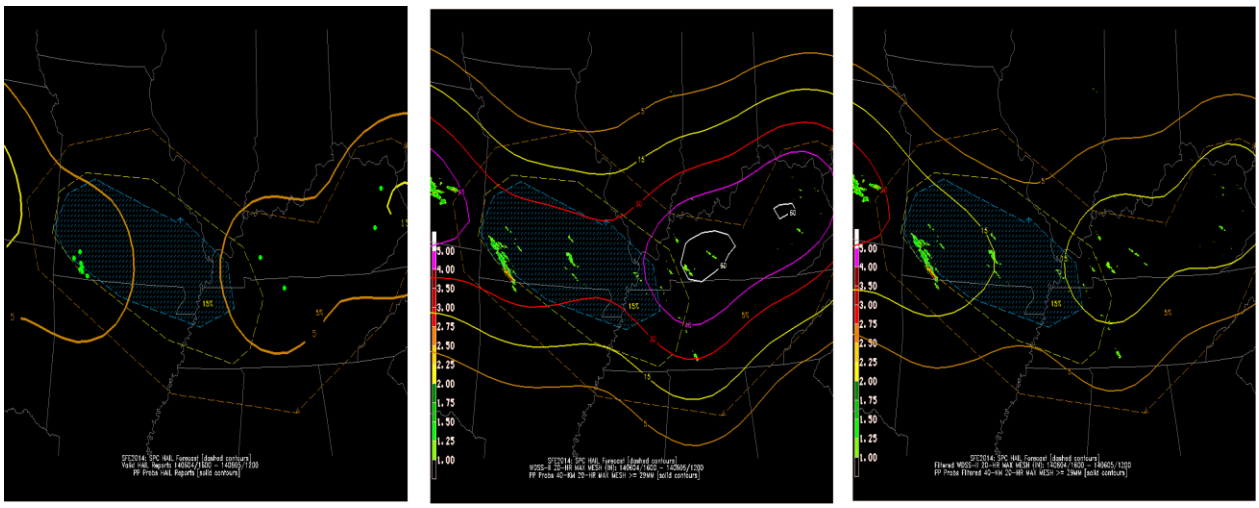

Fig. 8. Same as in Fig. 7 except for the forecast date of June 4th, 2014. In addition, a 10% or greater hatched area for significant severe hail is also predicted for this date.