2.5 Combining Probabilistic Ensemble Information from the Environment with Simulated Storm Attributes to Generate Calibrated Probabilities of Severe Weather Hazards

Israel L. Jirak[1*], Christopher J. Melick[1,2], and Steven J. Weiss[1]
[1]NOAA/NWS/NCEP/Storm Prediction Center, Norman, OK
[2]CIMMS, University of Oklahoma, Norman, OK

1. INTRODUCTION

The Storm Prediction Center (SPC) has been generating probabilistic forecasts of severe weather hazards (i.e., tornadoes, large hail, and damaging winds) as part of the Day 1 Convective Outlook for over a decade. Although the SPC forecasters have become proficient in generating reliable hazard probabilities over the years, there is a lack of calibrated probabilistic guidance to forecast these hazards. The recent proliferation of convection-allowing models (CAMs) in SPC operations allows for the unique opportunity to supplement traditional ingredients-based forecast assessments of the environment (e.g., CAPE and vertical wind shear: Johns and Doswell 1992) with explicit simulated storm-attribute characteristics (e.g., intensity and rotation; Kain et al. 2010).

The objective of this initial effort is to generate separate calibrated probabilities for tornadoes, large hail, and damaging winds by combining probabilistic environment information from the Short-Range Ensemble Forecast (SREF) system and probabilistic storm-attribute information from the SPC Storm-Scale Ensemble of Opportunity (SSEO). The following section will discuss the data and methodology used in creating calibrated hazard probabilities. Section 3 will provide forecast examples of the calibrated probabilities and a statistical analysis of the results for 2014. The final section will provide a summary of the findings.

2. DATA AND METHODOLOGY

2.1 Data

The SREF is a 21-member multi-model, multi-initial condition, and multi-physics ensemble run operationally at 16-km grid length at NCEP (Du et al. 2014). At this resolution, the SREF is primarily used by SPC forecasters to assess the forecast environment for severe weather potential. For example, an SPC forecaster may look at SREF forecasts of CAPE, vertical wind shear, and precipitation (among other fields) to determine the potential and likelihood of severe weather occurrence.

The SSEO, in comparison, is a 7-member multi-model, multi-initial condition, and multi-physics ensemble comprised of deterministic CAMs processed by SPC (Jirak et al. 2012). At this resolution, the models generate explicit convection, which allows for examination of simulated storm attributes, such as mode (e.g., Done et al. 2004; Weisman et al. 2008) and intensity [e.g., hourly maximum fields (HMFs); Kain et al. 2010]. An SPC forecaster may look at SSEO forecasts of reflectivity and updraft helicity (UH; Kain et al. 2008) to determine the likely storm mode and aspects of intensity in predicting the severe weather threat.

The concept of combining forecast information from these ensemble systems is similar to how a forecaster may analyze an ongoing convective event. For example in Fig. 1a, a forecaster might expect the discrete supercells in east-central Mississippi to have high potential to be tornadic given the favorable environment [i.e., significant tornado parameter (STP: Thompson et al. 2003) value over 4]. Analogously, a forecaster might predict a reasonably high probability of tornadoes given the forecast of a median STP value of 4 from the SREF valid while the NSSL-WRF is generating embedded supercells across central Mississippi (Fig. 1b).

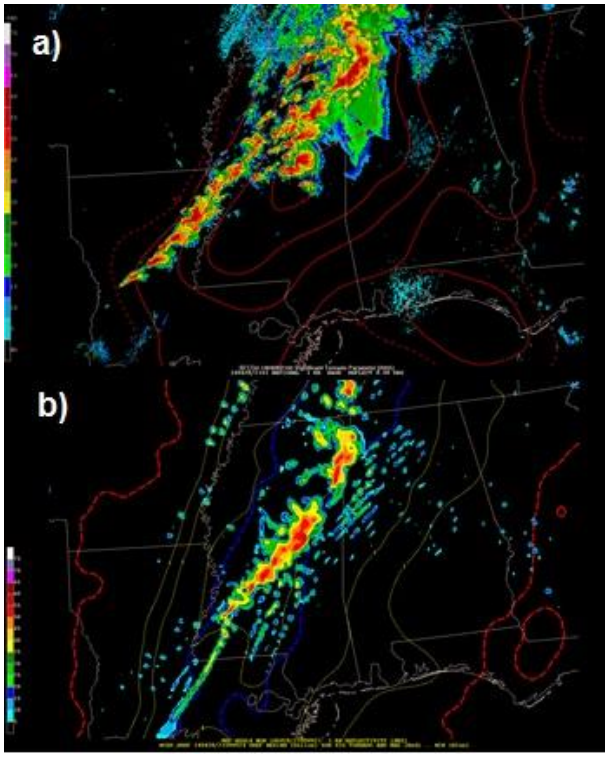

Figure 1. Valid at 2100 UTC on 28 April 2014 a)observed radar reflectivity (shaded) with mesoanalysis STP (contoured) and b) 21-h NSSL-WRF reflectivity forecast (shaded) and 24-h SREF STP forecast (contoured).

* *Corresponding author address*: Israel L. Jirak, NOAA/NWS/NCEP/Storm Prediction Center, 120 David L. Boren Blvd., Norman, OK 73072; e-mail: Israel.Jirak@noaa.gov

A key challenge of combining information from the SREF and SSEO in generating calibrated probabilities of severe weather hazards was selecting the appropriate fields and their threshold values from a limited number of archived fields at SPC. In this initial effort, physical reasoning and knowledge of favorable mesoscale environments and parameter magnitudes from ~4km CAMs was used to select first-guess predictor fields. The selection of fields for tornado forecasts was the most straightforward. The probability of STP ≥1 from the SREF was paired with the smoothed neighborhood probability (Harless et al. 2010) of UH ≥25 m^2s^{-2} from the SSEO to generate a calibrated tornado probability. Again, the idea is that given forecasts of explicitly rotating storms in an environment favorable for tornadoes should result in a higher probability of tornadoes.

Choosing fields for hail was more difficult given the larger variety of storm modes and environments that can produce severe hail. Ultimately, the probabilities of most-unstable CAPE (MUCAPE) ≥1000 Jkg^{-1} and effective shear ≥20 kts were selected for the SREF, based on environments in which the majority of severe hail reports occur (Fig. 2). From the SSEO, the smoothed neighborhood probability of updraft speed ≥10 ms^{-1} was chosen, as to not exclude non-rotating, multi-cell storms in the calibration of hail forecast probabilities.

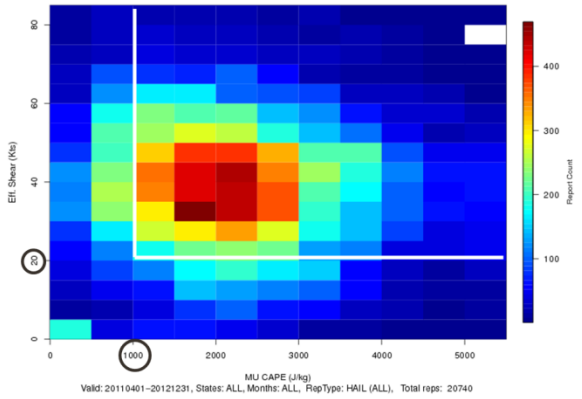

Figure 2. Number of severe hail reports (shaded) from 1 April 2011 to 31 December 2012 across the CONUS plotted against MUCAPE (x-axis) and effective shear (y-axis) values from the SPC mesoanalysis database (Dean et al. 2006).

Finally, the selection of fields for generating calibrated probabilities for wind was the most difficult. Damaging wind gusts are possible from a variety of storm modes and environments (Smith et al. 2013), ranging from a dry microburst in a deep, well-mixed boundary layer to a derecho in a very moist, unstable environment with sufficient shear. As a starting point, the probability of MUCAPE ≥250 Jkg^{-1} was chosen from the SREF to ensure convective processes were contributing to the surface winds. From the SSEO, the smoothed neighborhood probability of 10-m wind speeds ≥30 kts was selected. The wind speed probabilities are only considered if they coincide with non-zero probabilities of reflectivity ≥40 dBZ (i.e., to ensure convectively generated wind gusts).

2.2 Calibration Methodology

To create a calibrated probability [defined as within 25 miles (40 km) of a point, per SPC operational forecasts], the fields selected for each hazard needed to be combined. The method chosen to calibrate the data was a frequency-adjustment approach analogous to that used for the SREF calibrated products at SPC (Bright et al. 2005; Bright and Wandishin 2006). Using this basic approach, the input data at 3-h intervals were binned over the calibration period (defined below) and assigned the historical relative frequency of the event (i.e., report within 40 km) for those bins as the calibrated probability. A separate calibration table was created for each 3-h time period, but data across the entire CONUS were combined, owing to the rarity of severe weather events. For example, a 90% probability of STP ≥1 from the SREF and a 50% probability of UH ≥25 m^2s^{-2} resulted in a ~20% calibrated tornado probability for 1800-2100 UTC (Fig. 3; i.e., ~2 out of every 10 grid points that had STP and UH probabilities in those bins had a tornado report within 40 km during the calibration period).

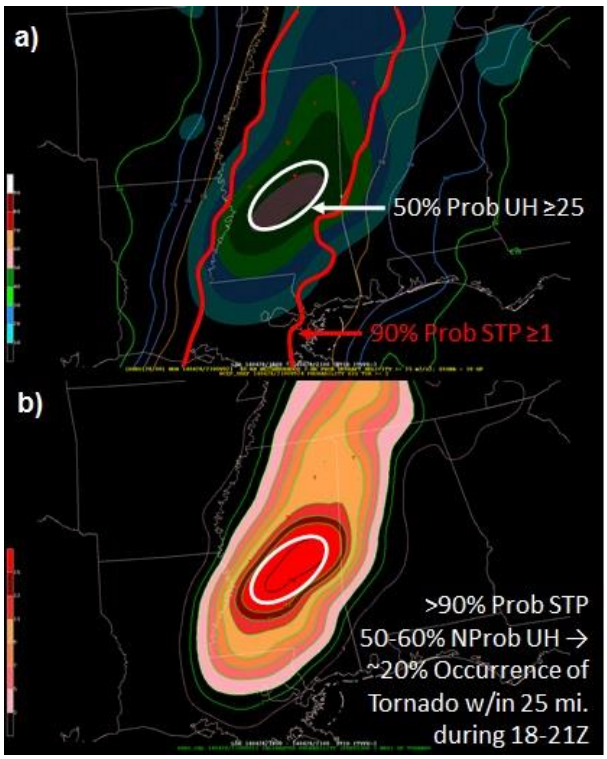

Figure 3. Valid at 2100 UTC on 28 April 2014 a) 24-h SREF forecast of STP ≥1 (contoured) and 21-h SSEO smoothed neighborhood probability of UH ≥25 m^2s^{-2} (hatched fill) and b) 3-h calibrated probability of tornado.

Forecasts from the 2100 UTC SREF and 0000 UTC SSEO cycles were used to create the calibration dataset valid in 3-hour periods from 1200 UTC on Day 1 (i.e., f15 for the SREF and f12 for the SSEO) to 1200 UTC on

the following day (i.e., f39 for the SREF and f36 for the SSEO). In addition, a second frequency-adjustment calibration process was applied to the eight 3-hour periods to create a 24-hour calibrated probability (valid for the convective day, 1200-1200 UTC). Unfortunately, the calibration periods were limited by the availability of SREF and SSEO data archived at SPC. The tornado calibration period included 14 April 2011-23 June 2011 and 14 February 2013-5 December 2013. The hail and wind calibration periods were identical running from 1 April 2011-31 December 2013. In addition to the limited data sample, the SREF configuration changed during the calibration period, so these issues likely had an impact on the calibration results. Nevertheless, this study highlights the practical approach of combining environmental and storm-attribute information to create calibrated probabilities. Presumably, a more robust forecast dataset (e.g., reforecast dataset with stationary ensembles) would yield even better results than those shown in the following section.

3. RESULTS

After generating the calibration tables for each hazard, forecasts were created for an independent dataset from 1 April 2014-18 October 2014. By the end of this period, the SREF and several members of the SSEO had changed once again. Regardless, the utility of this concept can still be demonstrated through forecast examples and verification statistics, including comparison to operational SPC convective outlooks.

3.1 Example Calibrated Forecasts

To provide a subjective perspective on the performance of the 24-h calibrated hazard probabilities, examples of the best and worst forecasts [i.e., in terms of critical success index (CSI)] during the independent data period of 2014 are shown for each hazard. The best forecast was defined as having the highest CSI across the CONUS at 10%, 15%, and 15% thresholds for tornado, hail, and wind, respectively, while the worst forecast was defined as having the most missed events (CSI=0) at those thresholds.

The best tornado forecast occurred for the tornado outbreak across Mississippi and Alabama on 28 April 2014 (Fig. 4a). While the probability values are too high across southern Alabama, the primary corridor of tornado activity was well captured by relatively high tornado probabilities. The worst tornado forecast occurred on 8 July 2014 where the calibrated tornado probabilities were too low from northwestern Ohio into northern Pennsylvania and New York.

The best hail forecast occurred on 25 April 2014 where a focused hail event across eastern North Carolina was well captured by the 15% probability contour (Fig. 5a). The 3 June 2014 hail event across Nebraska with numerous significant hail reports was also well forecast (second highest CSI; not shown). The worst hail forecast occurred on 27 May 2014 in which the overlap of even low probabilities (5%) with hail reports was poor (Fig. 5b).

As with tornadoes, the best wind forecast occurred on 28 April 2014 (Fig. 6a). Despite the large number of wind reports, however, the calibrated wind probabilities did not reach the 30% threshold. The worst wind forecast occurred on 2 September 2014 and reveals the difficulty in producing quality forecasts of severe wind, as there is little correspondence of the wind probabilities to the wind reports on this day.

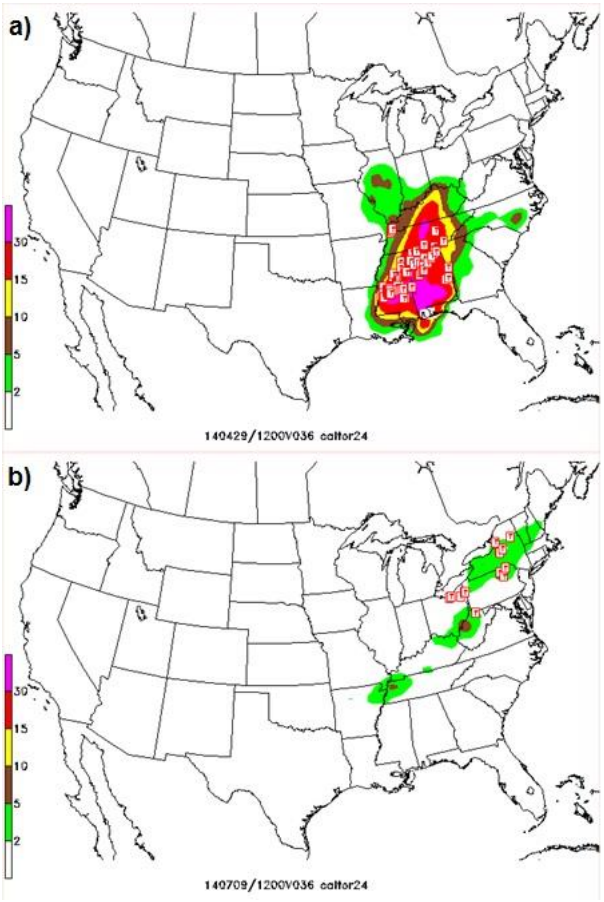

Figure 4. Example 24-hour calibrated tornado probability forecasts (shaded; %) for the a) best forecast: 28 April 2014 and b) worst forecast: 8 July 2014. The preliminary tornado reports are indicated with a red "T".

3.2 Statistical Verification

Verification statistics were calculated for the entire independent data period from 1 April 2014- 18 October 2014. A 2x2 contingency table was tallied to calculate several standard metrics including probability of detection (POD), frequency of hits (FOH), bias, and CSI for each forecast period at all probability thresholds (i.e., 2, 5, 10, 15, 30, 45%) for the 24-hr calibrated probabilities, using local storm reports to verify the forecasts. These metrics are concisely displayed on a performance diagram (Roebber 2009) to compare the forecasts of the individual hazards.

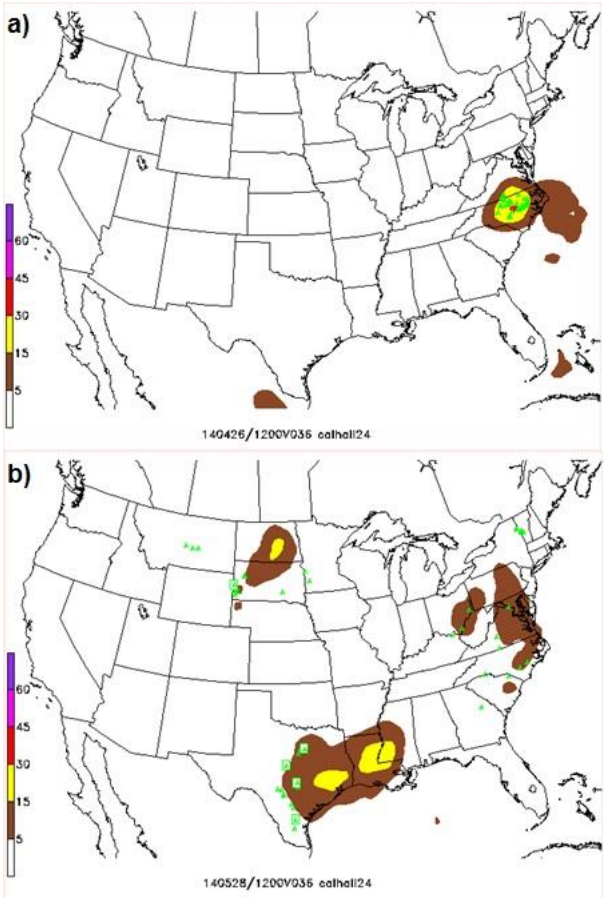

Figure 5. Example 24-hour calibrated hail probability forecasts (shaded; %) for the a) best forecast: 25 April 2014 and b) worst forecast: 27 May 2014. The preliminary hail reports (≥1") are indicated with a green "A" while significant hail reports (≥2") have a box around the letter.

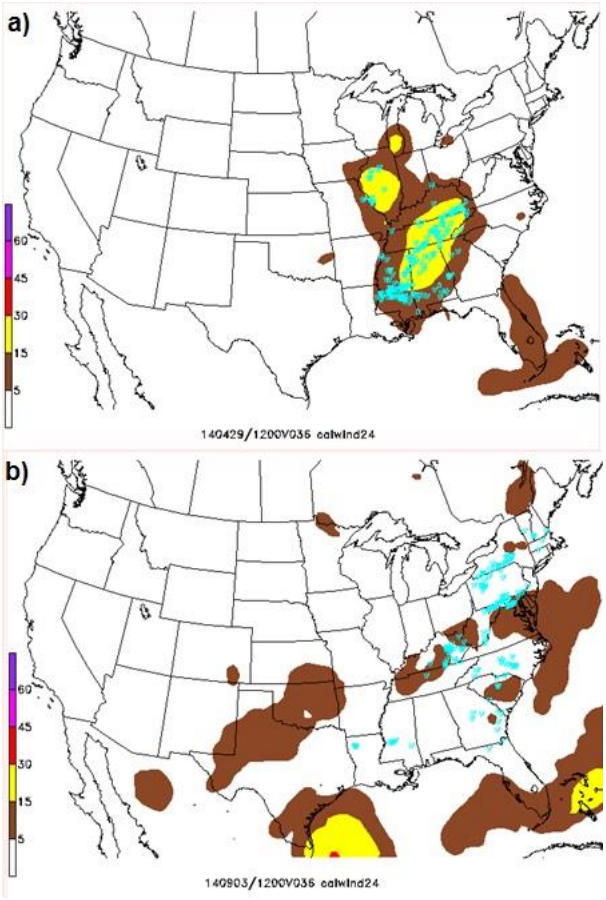

Figure 6. Example 24-hour calibrated wind probability forecasts (shaded; %) for the a) best forecast: 28 April 2014 and b) worst forecast: 2 September 2014. The preliminary wind reports are indicated with a blue "W".

Although contingency-table metrics are more appropriate for dichotomous forecasts (i.e., occurrence vs. non-occurrence) than for probabilistic forecasts, owing to artificially lower POD values for higher forecast probabilities, the performance diagram provides a convenient way to summarize verification information. An inspection of the performance diagram for 24-hr calibrated probability forecasts for tornado, hail, and wind reveals overall low verification scores for all hazards at all thresholds (Fig. 7). The 15% hail forecast was the only forecast with a CSI value above 0.1 during the period. As expected, the POD was highest for each hazard at the lowest probability threshold and decreased for higher probability thresholds. The calibrated hail forecasts generally verified better than the calibrated tornado and wind forecasts, as indicated by being farther toward the upper right of the performance diagram. The wind forecasts, especially at 15% and 30%, were notably the worst forecasts. For example, the 15% calibrated wind forecast had a much lower POD than the 5% forecast (i.e., 0.2 vs. 0.6), yet the FAR was similar at both thresholds.

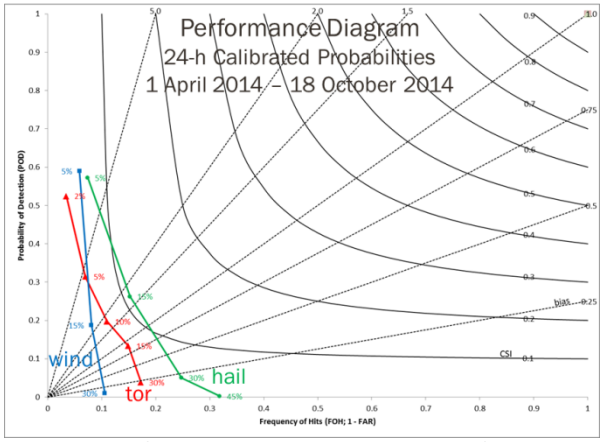

Figure 7. Performance diagram (Roebber 2009) for the 24-hr calibrated probabilities of tornado (red), hail (green), and wind (blue) for the period of 1 April 2014-18 October 2014.

Reliability diagrams (Wilks 2006) can provide additional insight into the characteristics and behavior of probabilistic forecasts. A reliability diagram plots the observed frequency of events for each forecast probability threshold to identify forecast bias and resolution. A reliable, or well-calibrated, forecast will have a 1:1 correspondence between forecast probability and observed frequency (e.g., a reliable forecast of 30% should have observed events 30% of the time over a sufficient sample).

The 24-hour calibrated hazard forecasts were generally an overforecast (i.e., below the diagonal) during the independent data period though the calibrated hail forecasts were generally the most reliable (Fig. 8). The calibrated wind forecasts stood out as the worst once again, showing poor resolution with a strong overforecast bias. The calibrated tornado forecasts generally verified between the hail and wind forecasts at the higher probability thresholds.

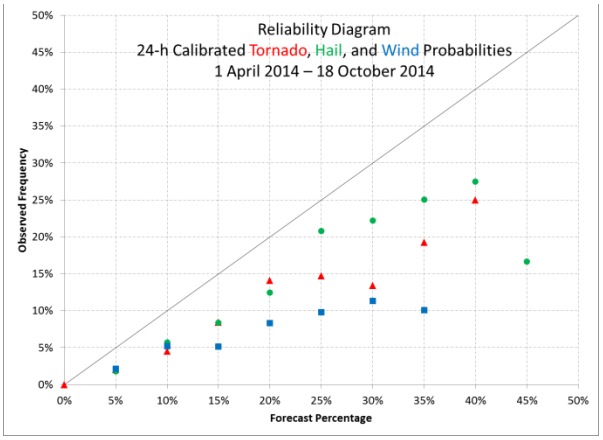

Figure 8. Reliability diagram for the 24-hr calibrated probabilities of tornado (red triangles), hail (green circles), and wind (blue squares) for the period of 1 April 2014-18 October 2014.

Even though the verification metrics and reliability of the calibrated forecasts revealed plenty of room for improvement, it is difficult to assess their skill without comparing to a reference forecast. Therefore, the 24-hour calibrated hazard probability forecasts were compared to the 0600 UTC SPC Day 1 probability outlooks. Given that the SPC outlooks are valid over the same 24-hour period and use the same probabilistic definition as the calibrated guidance, the two could be directly compared during the independent forecast period from 1 April 2014- 18 October 2014.

In comparing the calibrated tornado probabilities to the SPC tornado outlook, the two verify similarly, especially at the 10% threshold (Fig. 9a). At 2% and 5%, the SPC outlooks have a noticeably higher POD than the calibrated forecasts with a similar FAR. By 15%, the calibrated forecasts actually have a higher POD (and CSI) than the SPC outlooks. The SPC tornado outlook probabilities are more reliable than the calibrated forecast probabilities (Fig. 9b). However, the sample size for the SPC outlooks is noticeably lower at all probability bins, indicating fewer/smaller probabilistic forecast areas compared to the calibrated guidance. In fact, the 0600 UTC SPC tornado outlook did not have any forecast probabilities greater than 15% during this period while there were several hundred grid points above that threshold for the calibrated probabilities (e.g., Fig. 4a). This trait is a promising characteristic of the calibrated tornado guidance.

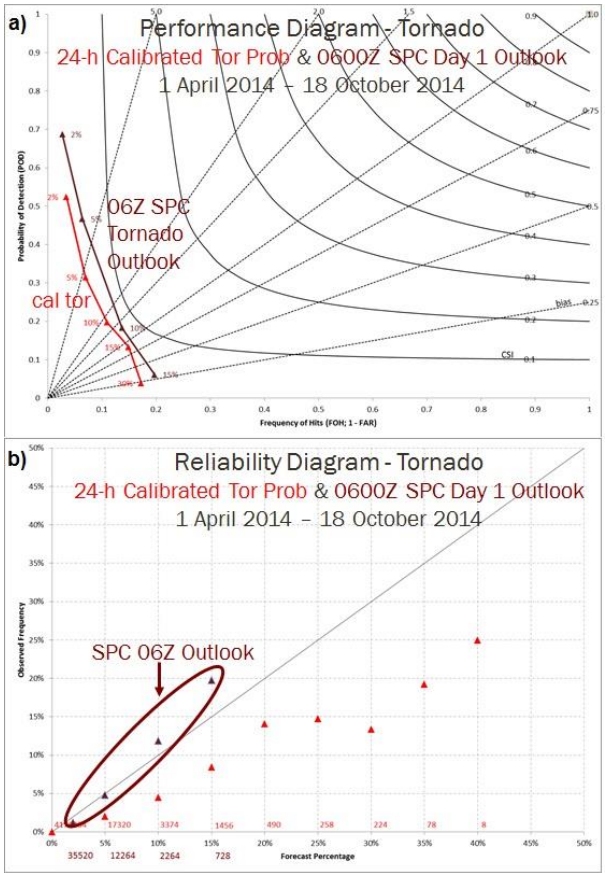

Figure 9. Comparison of 24-hour calibrated tornado probabilities (red) with 0600 UTC SPC Tornado Outlooks (dark red) in a) performance diagram and b) reliability diagram from 1 April 2014-18 October 2014. The forecast sample size is noted for each bin in the respective colors.

The performance diagram for hail forecasts showed a clear separation between the SPC hail outlooks and the calibrated hail probabilities (Fig. 10a). The SPC hail outlooks verified better at every probability threshold with a much higher POD than the calibrated hail forecasts while maintaining a similar FAR. While both forecasts display a similar overforecast bias (Fig. 10b), the SPC hail outlooks (unlike the SPC tornado outlooks) have more/larger forecast areas than the calibrated hail probabilities at ≥15%. Although other verification aspects of the calibrated hail forecasts were favorable, the inability to produce high probabilities was a negative characteristic of this guidance.

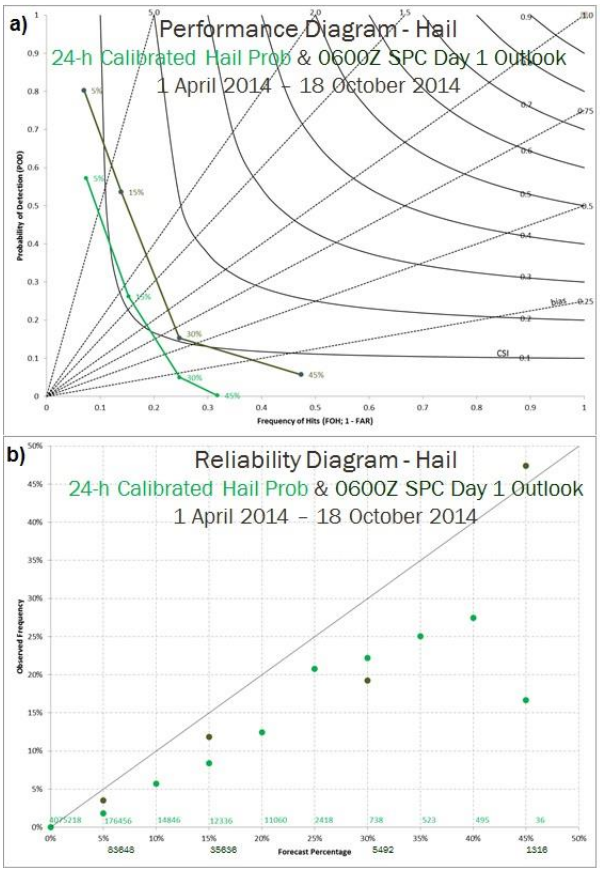

Figure 10. Same as Fig. 9, except for 24-hour calibrated hail probabilities (green) and 0600 UTC SPC Tornado Outlooks (dark green).

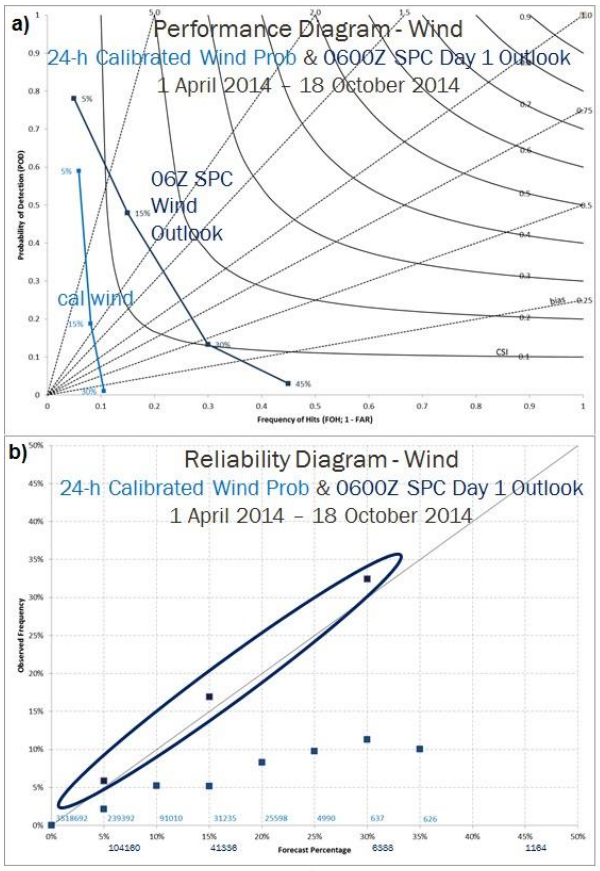

Figure 11. Same as Fig. 9, except for 24-hour calibrated wind probabilities (blue) and 0600 UTC SPC Tornado Outlooks (dark blue).

The poor performance of the calibrated wind probabilities was especially apparent when compared to the 0600 UTC SPC wind outlooks. The SPC wind outlooks verified with much higher POD and lower FAR than the calibrated wind forecasts, especially above 5% (Fig. 11a). In addition, the SPC wind outlooks were much more reliable and produced more/larger forecast areas at high probabilities (i.e., ≥15%) than the calibrated wind probabilities. These results highlight the very challenging nature of creating useful and reliable calibrated severe wind guidance, likely owing to the variety of modes/environments that are capable of producing storms with damaging wind gusts and the suspect quality of the wind verification database (e.g., Trapp et al. 2006).

4. SUMMARY AND CONCLUSIONS

The concept of combining probabilistic environmental and storm-attribute forecast information was presented for generating calibrated guidance for severe weather hazards: tornadoes, large hail, and damaging winds. There are a large number of challenges in developing this type of probabilistic guidance: underdispersive and changing ensemble systems, limited calibration sample for rare events, representativeness issues with observations (i.e., reports), etc. Despite these challenges, the results presented in using this approach were promising, especially for tornadoes and hail. Additionally, the technique appeared to produce the best results on the biggest severe weather days (e.g., 28 April 2014). While improvements can likely be made by modifying the calibration fields, adjusting the statistical approach, and expanding the calibration sample, the encouraging preliminary results support continued exploration of this overall concept.

Acknowledgements. We would like to thank SPC forecasters and other participants who looked at this guidance and provided feedback during SFE2014.

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.

Bright, D.R., M.S. Wandishin, R.E. Jewell, and S.J. Weiss, 2005: A physically based parameter for lightning prediction and its calibration in ensemble forecasts. *Preprints*, Conf. on Meteor. Applications of Lightning Data, San Diego, CA, 4.3.

Dean, A.R., R.S. Schneider, and J.T. Schaefer*, 2006: Development of a comprehensive severe weather forecast verification system at the Storm Prediction Center. *Preprints*, 23nd Conf. Severe Local Storms, St. Louis, MO, P2.3.

Done, J., C. Davis, and M. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the Weather Research and Forecasting (WRF) model. *Atmos. Sci. Lett.*, **5** (6), 110–117.

Du, J. and Co-authors, 2014: NCEP regional ensemble update: current systems and planned storm-scale ensembles. *Preprints,* 26th Conf. on Wea. Forecasting, Atlanta, GA, Amer. Meteor. Soc., J1.4.

Harless, A.R., S.J. Weiss, R.S. Schneider, M. Xue, and F. Kong, 2010: A report and feature-based verification study of the CAPS 2008 storm-scale ensemble forecasts for severe convective weather. *Preprints*, 25th Conf. Severe Local Storms, Denver CO. Amer. Meteor. Soc., 13.2.

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. *Preprints*, 26th Conf. Severe Local Storms, Nashville, TN. Amer. Meteor. Soc., P9.137.

Johns, R. H. and C. A. Doswell III, 1992: Severe local storms forecasting. *Wea. Forecasting*, **7**, 588–612.

Kain, J. S., 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, K. W. Thomas, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931-952.

Kain, J. S., S. R. Dembek, S. J. Weiss, J. L. Case, J. J. Levit, and R. A. Sobash, 2010: Extracting unique information from high resolution forecast models: Monitoring selected fields and phenomena every time step. *Wea. Forecasting*, **25**, 1536–1542.

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

Smith, B. T., T. E. Castellanos, A. C. Winters, C. M. Mead, A. R. Dean, and R. L. Thompson, 2013: Measured severe convective wind climatology and associated convective modes of thunderstorms in the contiguous United States, 2003–09. *Wea. Forecasting*, **28**, 229–236.

Thompson, R. L., R. Edwards, J.A. Hart, K.L. Elmore and P.M. Markowski, 2003: Close proximity soundings within supercell environments obtained from the Rapid Update Cycle. *Wea. Forecasting*, **18**, 1243–1261.

Trapp, R. J., D. M. Wheatley, N. T. Atkins, R. W. Przybylinski, and R. Wolf, 2006: Buyer beware: Some words of caution on the use of severe wind reports in postevent assessment and research. *Wea. Forecasting*, **21**, 408–415.

Weisman, M. L., C. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with 0–36-h explicit convective forecasts with the WRF-ARW model. *Wea. Forecasting*, **23**, 407–437.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 627 pp.