P15.5 EVALUATION OF SHORT-RANGE ENSEMBLE FORECASTS DURING THE 2003 SPC/NSSL SPRING PROGRAM

David R. Bright*, Steven J. Weiss, and Jason J. Levit
NOAA/NWS Storm Prediction Center, Norman, OK

Matthew S. Wandishin
University of Arizona, Tucson, AZ

John S. Kain
CIMMS/University of Oklahoma, Norman, OK

David J. Stensrud
NOAA/National Severe Storms Laboratory, Norman, OK

1. INTRODUCTION

The Storm Prediction Center/National Severe Storms Laboratory (SPC/NSSL) Spring Program (hereafter referred to as simply the Spring Program, or SP) is a collaborative multi-week exercise held in Norman, OK coinciding with the climatological peak of severe convective weather (Kain et al. 2003). It is designed to bring together meteorologists from research and operational communities to investigate specific, applied research problems, and swiftly migrate positive results into SPC operations.

The 2003 SP ran from 14 April to 6 June with a primary objective of evaluating the ability of short-range ensemble forecasts (SREF) to aid in the prediction of severe convection. More specifically, the SP sought to: (1) determine if SREFs can benefit SPC real-time convective forecasting operations, and if so, (2) explore techniques that may assist the integration of SREFs into a historically deterministic forecast process. The full operations plan can be viewed online at http://www.spc.noaa.gov/exper/Spring_2003.

2. METHODOLOGY

a. Spring Program SREF Exercise

SP participants were asked to spend a full week (Monday through Friday) as part of a four member experimental forecast team at the SPC. Each four member team was anchored by a SPC forecaster with the other three participants from operational, research, or academic institutions. Participating institutions included the following: the Cooperative Institute for Mesoscale Meteorological Studies; NCEP's Environmental Modeling Center; NOAA's Forecast Systems Laboratory; the Norman, Oklahoma and White Lake, Michigan National Weather Service WFOs; the University of Arizona; the University of Oklahoma; the University of Washington; Iowa State University; Massachusetts Institute of Technology; the United Kingdom Meteorological Office; and Meteorological Services of Canada. Part-time observers from COMET and USWRP also participated.

The SP forecast exercise consisted of issuing experimental probabilistic outlooks of severe convection valid for Day 2 (i.e., beginning 12 UTC tomorrow through 12 UTC the following day). Focusing on Day 2 convection ensured proper emphasis was placed on numerical model and SREF guidance rather than observational data. The experimental outlooks were issued every Monday through Thursday and were similar in content to their operational counterpart.

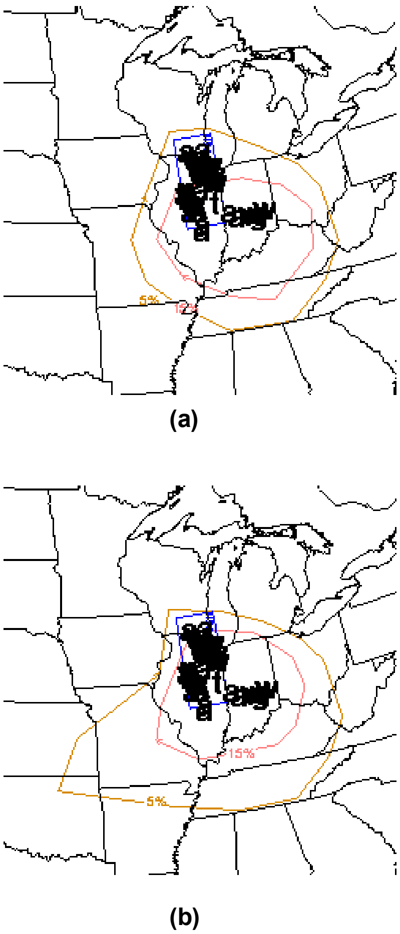
An initial Day 2 outlook was first issued during the late morning following a "traditional" forecast process incorporating several deterministic models with all SREF information withheld. Forecasters then examined SREF data and subsequently issued an updated (or final) outlook. The assumption is made that any change to the forecast is the result of incorporating SREF information.

b. Spring Program SREF Data

The NCEP Environmental Modeling Center (EMC) SREF (Du and Tracton 2001) was the primary system utilized during the SP. This system consisted of 15 members including 5 Eta members with Betts-Miller-Janjic convection (Eta-BMJ); 5 Eta members with Kain-Fritsch convection (Eta-KF); and 5 members from the NCEP Regional Spectral Model (RSM). The horizontal grid spacing was 48 km and each five member subset included one member with an unperturbed initial condition and four members with perturbed initial conditions using the breeding of growing modes technique.

A secondary SREF system available to SP forecasters was configured by the NSSL and executed on the University of Oklahoma supercomputer (Levit et al. 2004). This 32 member SREF utilized a single version of NCAR's MM5 with initial perturbations constructed via forecaster determined regions of uncertainty and the MM5 adjoint. Using a WEB-based interface, 16 parameters of concern were identified during the 12-48 hour forecast period, allowing the MM5 adjoint to then produce reasonably scaled initial perturbations owing their existence to forecaster diagnosed uncertainty (Xu et al. 2001). These forecasts often arrived too late to be of consequence in the final outlook. Thus, the NCEP SREF was the primary system used in developing the final outlook. Initial verification results of the MM5 adjoint SREF are in Homar et al. (2004).

* *Corresponding author address*: David R. Bright, 1313 Halley Circle, Norman, OK 73069; e-mail: david.bright@noaa.gov



(a)

(b)

Fig. 1. The (a) initial and (b) final experimental Day 2 outlooks issued during the Spring Program on 27 May 2003. SREF information was withheld from the forecasters in creating the initial outlook, while the final outlook includes SREF guidance. The valid period of the outlooks is 12 UTC 28 May to 12 UTC 29 May 2003.

3. SPRING PROGRAM VERIFICATION

The difference between initial and final outlooks was often subtle and generally consisted of minor shifts in areal coverage or probabilistic values. For example, the experimental outlooks issued on 27 May 2003 show the area encompassed by a 15% chance of severe thunderstorms (SPC policy equates a 15% probability to a "slight risk" of severe weather) is shifted northward approximately 100 miles, removing northern Tennessee and southern Kentucky from a slight risk while adding the Chicago metropolitan area (Fig. 1). The 5% area was also extended southwestward along the cold front. This particular day demonstrates a relatively large adjustment to the outlook; most days involved lesser modification.

In order to address the utility and skill of integrating SREF guidance in the forecast process, one subjective and two objective measures were used to verify the 31 initial and final outlooks. The subjective measure was based on an after-the-fact team evaluation of the outlooks, with each forecast receiving a subjective rating of 0 to 10. This rating reflects the team's consensus subjective opinion as to the usefulness of the forecast. Because teams varied from week to week, the raw magnitude of the distinct ratings are not uniformly calibrated. However, the difference between an initial and final rating does provide a subjective measure of usefulness that is comparable from week to week. Over the entire 31 days of the experiment, the final outlook was considered an improvement 14 times and degradation 6 times (Fig. 2). (For the example shown in Fig 1, the initial outlook received a rating of 6 and the final outlook scored an 8.)

As an objective measure of performance, the Brier score and area under the Relative Operating Characteristic (ROC) curve were calculated. The Brier score is commonly used to verify probabilistic forecasts, ranging from a perfect score of 0 to a worst-possible value of 1. Similarly, the ROC is useful for verifying probabilistic forecasts and their ability to discriminate occurrences from non-occurrences. If the area under the ROC curve is integrated values range from 0 to a perfect score of 1, with an area greater than 0.7 considered to represent reasonable discriminating ability.

Fig. 3 shows the Brier score results for each day of the experiment graphed as the percentage improvement of the final outlook compared to the initial outlook. For larger-change days (arbitrarily chosen to be ± 1%) SREF-adjusted outlooks improved the forecast 14 times and degraded the forecast 4 times. When all 31 days are considered collectively, the Brier score indicates a 0.1% improvement in SREF-adjusted outlooks. Similarly, the daily results from the ROC area show general improvement (Fig. 4), with larger-change days (arbitrarily chosen at ± 2%) indicating SREF-adjusted improvement 10 times and degradation 5 times. Over the entire 31 days collectively, the ROC area is 5% better in SREF-adjusted outlooks. (For the case in Fig. 1, the Brier score and ROC area both indicated the final outlook provided about 4% improvement.)

The reliability of all final outlooks is shown in Fig. 5. At first glance, it appears severe events may be

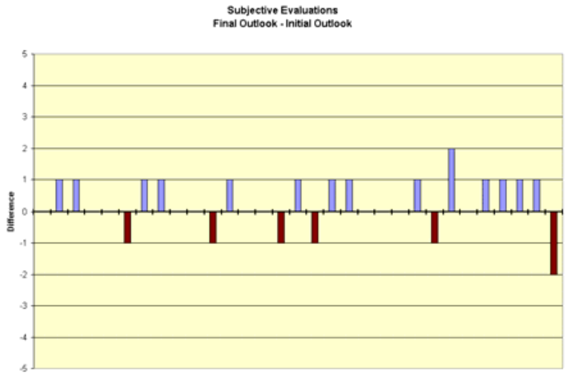


Fig. 2. The difference (final outlook – initial outlook) in subjective ratings for all 31 Spring Program days from 14 April to 6 June 2003. The usefulness of each outlook was verified subjectively and received a score from 0 to 10.

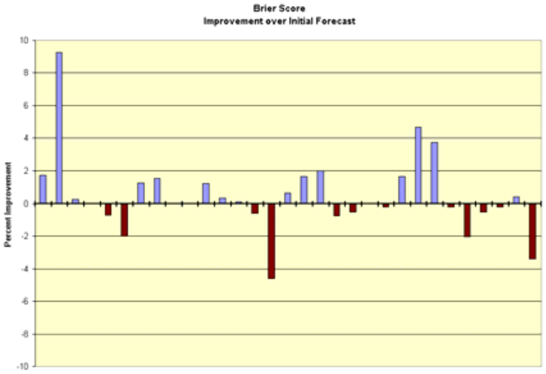


Fig. 3. As in Fig. 2, except the percentage improvement of the final outlook Brier score compared to the initial outlook Brier score. Severe reports on the 80 km AWIPS grid 211 were used to objectively verify the outlooks.

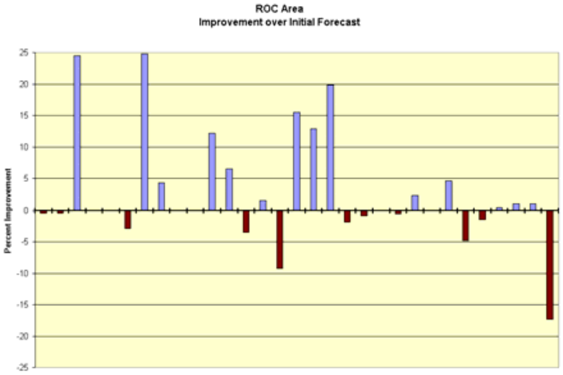


Fig. 4. As in Fig. 3, except the percentage improvement of the area under the ROC curve.

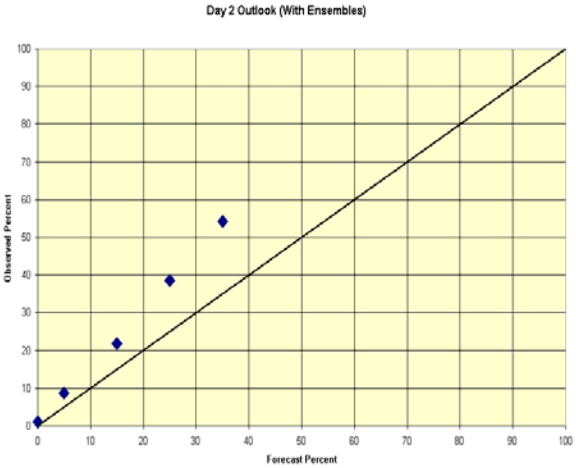


Fig. 5. Reliability diagram of Spring Program outlooks for the 31 days of the experiment. The results plotted are for the final, SREF-adjusted forecast. The reliability of the initial outlook (SREF information withheld) is nearly identical and not shown.

slightly underforecast, particularly at higher probabilities. However, SPC Day 2 probabilities are issued at 10% increments from 5% to 35% such that a 15% value represents all probabilities from ≥ 15% to < 25%. Thus, underforecasting is really only occurring at 25% and 35%. Some of the underforecasting at 35% is almost certainly due to the self-imposed SPC policy that Day 2 probabilities shall not exceed 35%.

4. DIAGNOSTIC PRODUCTS

Some products found particularly useful for viewing SREF output are now described. These products are demonstrated using the NCEP SREF forecast from 09 UTC 27 May 2003 and correspond to the example shown in Fig. 1. All forecasts are valid 03 UTC 29 May 2003 (forecast hour 42) unless otherwise noted. (Many of these products are available on the SPC real-time SREF webpage at http://www.spc.noaa.gov/exper/sref/.)

The mean and standard deviation are a classic way of viewing ensemble forecasts. At 500 hPa, the SREF mean shows a jet maximum over Iowa while the standard deviation indicates the largest uncertainty in its magnitude and/or location in the left-exit region (Fig. 6). Another useful plot for displaying central tendency is the median member with "spatial range" overlaid (Fig. 7). For example, the median of surface-based CAPE shows a maximum over northern Illinois and southern Wisconsin (Fig. 7, solid contours), with at least one member (i.e., the maximum or union of the ensemble) exceeding 500 J/kg of CAPE as far west as central Iowa (Fig. 7, red dashed) and all members (i.e., the minimum or intersection of the ensemble) exceeding 500 J/kg over a small area of northern Illinois and southern Wisconsin (Fig. 7, blue dashed). While the standard

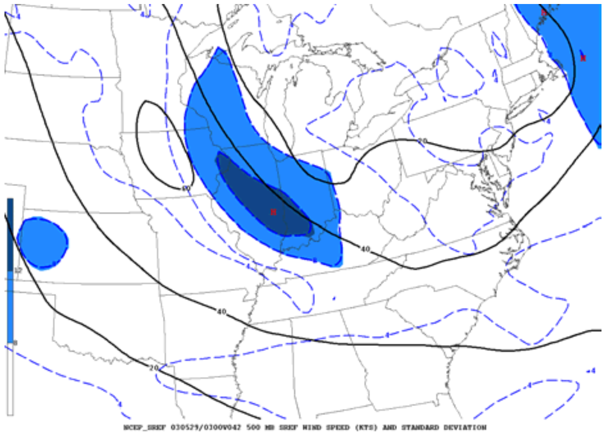

60
40
40
20
NCEP_SREF 030529/0300V042 500 MB SREF WIND SPEED (KTS) AND STANDARD DEVIATION

Fig. 6. SREF mean (solid black) and standard deviation (shaded) of isotachs (kts) at 500 hPa valid 03 UTC 29 May 2003 (forecast hour 42). The largest uncertainty in the position of the mid level jet maximum is in the the left-exit region over IL.

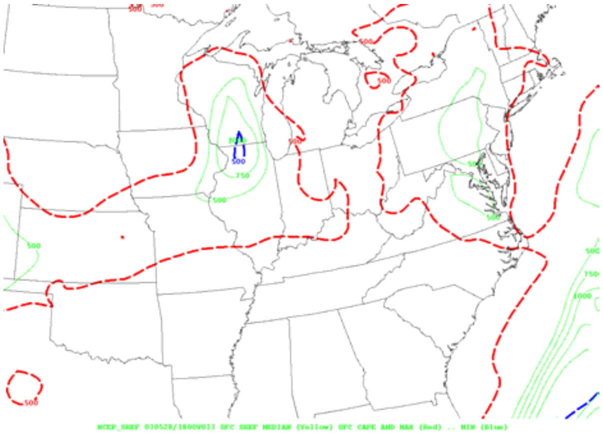

NCEP_SREF 030528/1800V033 SFC SREF MEDIAN (Yellow) SFC CAPE AND MAX (Red) .. MIN (Blue)

Fig. 7. SREF median (solid green) and spatial range of surface-based CAPE valid 03 UTC 29 May 2003 (forecast hour 42). The union (or maximum) of any member with at least 500 J/kg is shown by the dashed red line, while the intersection (or minimum) of all members with ≥ 500 J/kg is indicated by the dashed blue line. All members are predicting at least 500 J/kg CAPE over northern IL/southern WI (blue dashed), while at least one member predicts 500 J/kg as far south as southern IL but only as far west as central IA (red dashed).

deviation provides a measure of variability at every grid point, the median with spatial range provides information on the areal coverage, or spatial range, of possible solutions.

The usefulness of single contour charts (or "spaghetti charts" as they're commonly known) was found to be situation dependent as their interpretation could sometimes be difficult, particularly if an ineffective contour value were selected. Nonetheless, spaghetti charts could be valuable in assessing spread, clustering, and predicted extremes or outliers. Furthermore, due to the offset start time of the SREF (i.e., 09 and 21 UTC), linkages between the latest high-resolution operational Eta and the SREF were deemed critically important. Spaghetti charts provide an easy method of building such a linkage. For example, see the spaghetti chart of surface based CAPE (39 hour forecast valid 00 UTC 29 May 2003; single contour value of 1000 J/kg) shown in Fig. 8. The red contours are the various Eta-BMJ members, the yellow contours the Eta-KF members, the blue contours the RSM members, and the solid black line the 12 UTC Eta. In this case the 12 UTC Eta predicted greater CAPE farther south through much of Kentucky and is clearly an outlier as compared to all 15 SREF members. The natural inclination of the SP forecaster was to reject the older SREF in favor of the more recent Eta solution. Nevertheless, the purpose of the SP experiment was to test SREF utility, such that in this case the Eta was deemed an outlier and therefore a less likely solution. As a result, the slight risk area was shifted northward in accordance with SREF guidance (see Fig. 1). This adjustment proved worthwhile (as discussed in the previous section).

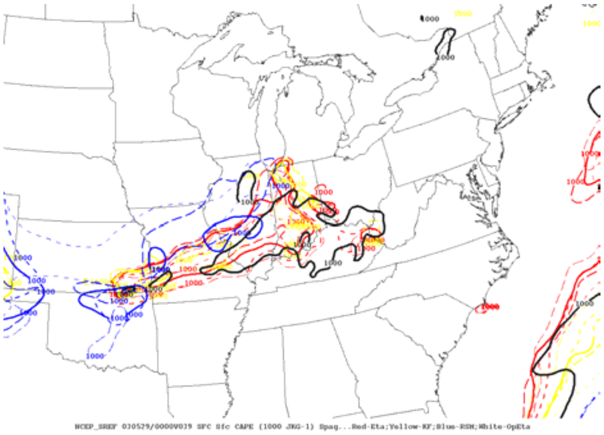

NCEP_SREF 030529/0000V039 SFC Sfc CAPE (1000 JKG-1) Spag...Red-Eta;Yellow-KF;Blue-RSM;White-OpEta

Fig. 8 SREF spaghetti chart of surface-based CAPE valid 00 UTC 29 May 2003 (forecast hour 39). The single contour is 1000 J/kg using a red, yellow, blue (Eta-BMJ, Eta-KF, RSM, respectively) color scheme. Solid contours are the control members. The solid black line is the 36 hour forecast from the 12 UTC operational Eta showing it to be an outlier over KY as compared to all SREF members.

Another noteworthy application of spaghetti charts was their use in conjunction with probabilistic forecasts (figure not shown). Spaghetti charts helped to discern where probabilities were low due to the clustering or phase shifting of the members versus where members simply did not meet the given threshold.

Probabilistic forecasts (uncalibrated) were considered the most useful SREF product during the SP. An example of their utility is now presented. In order to demonstrate a point, let's begin with some basic ingredients of supercell formation: instability, shear, and convective initiation. Translating these ingredients to probabilities provided by the SREF leads to the probability of CAPE exceeding 1000 J/kg, surface-to-6 km shear exceeding 30 kts, and convective precipitation exceeding 0.01" (Figs. 9-11, respectively). These ingredients intersect somewhere over the upper Mississippi river valley, but joint probabilities (i.e., the probability of multiple events occurring) are required to assess the situation more precisely. Joint probabilities were in fact found to be extremely useful, but required pre-calculation or customized software to generate them on-the-fly. A practical alternative is to simply treat the probabilities as if they're independent and take the product of the probabilities. The resulting "combined probability" allows the forecaster to quickly invoke an ingredients-based approach for ensemble interrogation which serves as a proxy for true joint probabilities. Multiplying the three probabilistic forecasts shown in Figs. 9-11 yields a combined probability of severe thunderstorms shown in Fig. 12. This combined probability delineates spatially where the ingredients for supercell thunderstorms intersect and provides some indication as to the likelihood of severe weather (provided the treatment as independent variables has not oversimplified the problem). In this case, the greatest SREF-based threat of severe thunderstorms is over greater Illinois. Indeed, severe weather did occur over much of this area (Fig. 13). The combined probability approach allows for on-the-fly creation of

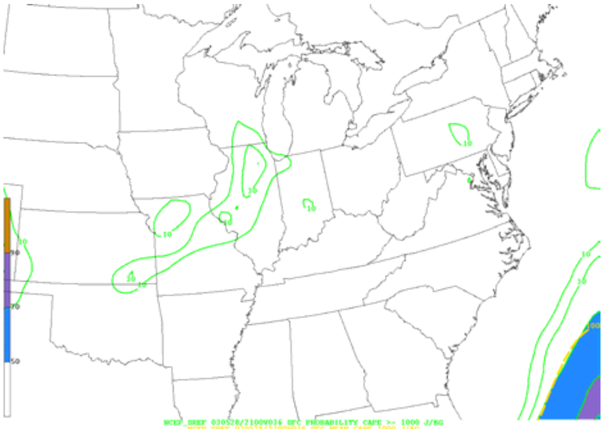


Fig. 9. Percentage of SREF members with surface-based CAPE ≥ 1000 J/kg valid at 03 UTC 29 May 2003 (forecast hour 42). The dashed gold line is the SREF mean at 1000 J/kg.

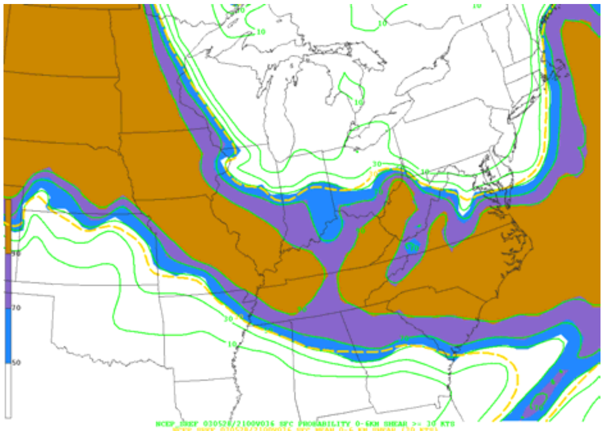


Fig. 10. As in Fig. 9, except percentage of SREF members with surface-to-6 km shear ≥ 30 kts. The dashed gold line is the SREF mean at 30 kts.

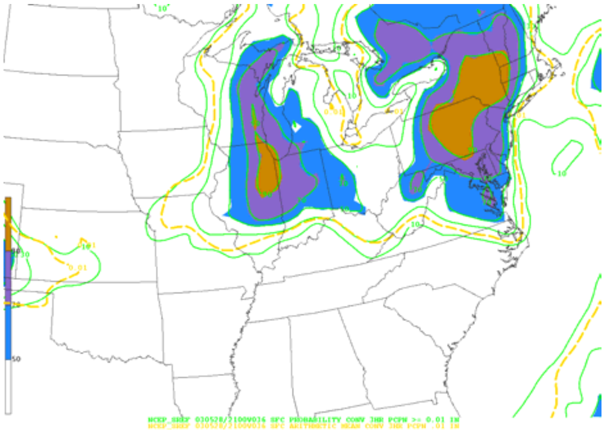


Fig. 11. As in Fig. 9, except percentage of SREF members with convective precipitation ≥ 0.01". The dashed gold line is the SREF mean at 0.01".

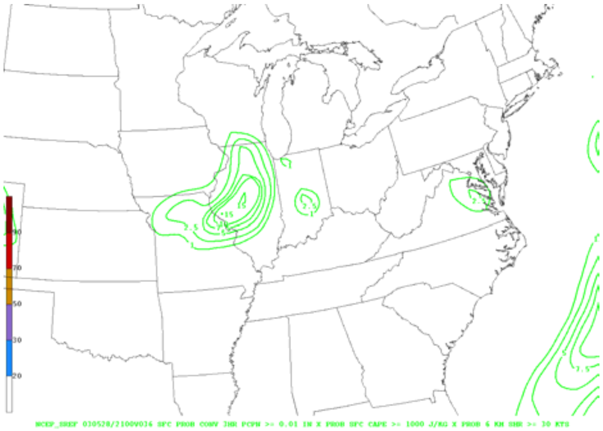


Fig. 12. The product of the probabilities shown in Figs. 9-11, quantifying the juxtaposition of ingredients for supercell thunderstorms (based on the thresholds given in Figs. 9-11).

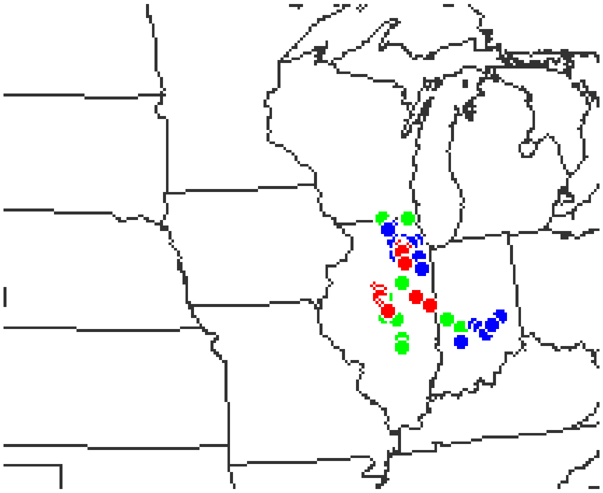


Fig. 13. Severe weather reports around the valid time of the prediction shown in Fig. 12. The red, green, and blue circles are the location of tornadoes, hail (>= 0.75"), and wind (>= 50 kts), respectively.

joint probabilities for most conceivable diagnostics, and may compensate for model biases by treating ingredients as if they're independent. Additional work needs to be undertaken to determine the impact of neglecting any dependence between ingredients.

5. SUMMARY

The 2003 SPC/NSSL Spring Program focused on the utility of short-range ensemble forecasts to determine if SREFs could benefit SPC operations. Results found that utilizing SREF output does provide a small but positive contribution to the Day 2 outlook process. Three metrics, including two objective measures (Brier score and area under the ROC curve) and a subjective evaluation, all showed a small but positive contribution to the Day 2 outlook when SREF diagnostics were considered.

The fact that improvement was small is testimonial to the skill of SPC forecasters and their ability to assess uncertainty in the forecast in the absence of SREF guidance. Indeed, most of the SPC forecasters that participated in the 2003 SP have several years of experience issuing probabilistic severe forecasts and already construct a "poor-person's ensemble" by examining the output of several operational numerical prediction models. Furthermore, it takes time to learn how new datasets should be integrated with existing information. Thus, the SP results are encouraging and suggest that SREFs can play a positive role in SPC operations.

One of the best applications of spaghetti charts is their ability to link the latest, higher-resolution deterministic forecast to an earlier ensemble forecast. If the deterministic forecast is found to be an outlier relative to the ensemble and that position in phase space cannot be explained through meteorological reasoning (e.g., new raob information provided a better initialization to the new model), then it may be prudent to apply less than customary weighting to the deterministic result.

Probability charts were the most popular SREF product used during the SP, and combining (or multiplying) probabilistic forecasts together in an ingredients-based approach proved helpful. These combined probabilities can serve as a convenient substitute for true joint probabilities that may require pre-calculation or more sophisticated software than currently available to the SPC.

Finally, real-time SREF output is now available on the SPC website at the following URL: http://www.spc.noaa.gov/exper/sref/.

6. REFERENCES

Du, J., and M. S. Tracton, 2001: Implementation of a real-time short range ensemble forecasting system at NCEP: An update. Preprints, *Ninth Conf. On Mesoscale Processes*, Fort Lauderdale, FL, Amer. Meteor. Soc., 355-356.

Homar, V., D. J. Stensrud, and J. J. Levit, 2004: Severe weather forecasts from an ensemble of human-perturbed simulations using an adjoint model. Preprint, *22nd Conf. on Severe Local Storms*, Hyannis, MA, Amer. Meteor. Soc.

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

Levit, J., D. Stensrud, D. Bright, and S. Weiss, 2004: Evaluation of short-range ensemble forecasts during the SPC/NSSL 2003 Spring Program. Preprint*, 16th Conf. On Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc.

Xu, M., D. J. Stensrud, J-W Bao, and T. T. Warner, 2001: Applications of the adjoint technique to short-range ensemble forecasting of mesoscale convective systems. *Mon. Wea. Rev.*, **129**, 1395-1418.