13.2 A REPORT AND FEATURE-BASED VERIFICATION STUDY OF THE CAPS 2008 STORM-SCALE ENSEMBLE FORECASTS FOR SEVERE CONVECTIVE WEATHER

Amy R. Harless[1, 2], Israel Jirak[2], Russell Schneider[2], Steven Weiss[2], Ming Xue[1, 3], Fanyou Kong[3]

[1]School of Meteorology, University of Oklahoma, Norman, Oklahoma
[2]NOAA/NWS/Storm Prediction Center, Norman, Oklahoma
[3]Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma

1. INTRODUCTION

A persistent challenge to forecasters is predicting small-scale, high-impact convective weather events such as high wind, severe hail, and tornadoes. Afforded by the advancement of computing power, innovative numerical systems, and assimilation of observations at high spatial and temporal density, nonhydrostatic numerical models at fine resolutions (i.e., 4 km grid spacing) that are capable of explicitly resolving convection are being experimentally applied to forecasting convective precipitation and severe weather [e.g., NOAA Hazardous Weather Testbed (HWT) Spring Experiment (http://hwt.nssl.noaa.gov)]. These convection-allowing models can develop storms with unique attributes permitting forecasters to interrogate characteristics of simulated storms and relate them to the likelihood of severe weather occurrence.

Even with the advancement of storm-scale models and data assimilation techniques, forecast quality is hindered by limited sampling of the atmosphere coupled with imperfect understanding and inherent nonlinearity of the physical and dynamical atmospheric processes. These sources of unpredictability motivate the use of an ensemble forecast system as a means by which model error and uncertainty can be quantified in the forecast. Employing ensemble forecasting methods to the storm scale will provide information about a range of solutions, including the timing, and location of convective storms and their characteristics that are sensitive to small changes in the environmental conditions. High-resolution ensembles offer beneficial output relative to a single deterministic model by representing inherent uncertainties through probabilistic guidance.

This project investigated the capabilities of a storm-scale ensemble forecast system to predict simulated storm attributes by comparing them to observed severe weather events. The study focused on the 2008 Storm-Scale Ensemble Forecast (SSEF) system developed by the University of Oklahoma Center for Analysis and Prediction of Storms (CAPS) for experimental forecasting and assessment during the 2008 NOAA HWT Spring Experiment. A unique dataset of diagnostic convective storm attribute fields from the ensemble was examined, including updraft helicity (UH), maximum column vertical velocity (VVEL) associated with storm updraft, and 4 km AGL simulated reflectivity (REFL). The utility of probabilistic forecasts of these model-derived fields were analyzed as guidance for operational forecasts of high-impact, convective weather events. Probabilities were extracted using various methodologies (i.e., traditional and neighborhood) and were evaluated and compared based on skill, reliability, and bias scores. The primary objective of this study was to assess the predictability of storm-scale attributes and determine if this 10-member ensemble at 4-km grid spacing provides useful objective guidance for making informed decisions in severe weather forecasting.

* *Corresponding author's address:* Amy R Harless, NOAA/NWS/NCEP Storm Prediction Center, 120 David L. Boren Blvd., Norman, OK 73072; Email: amy.harless@noaa.gov

2. BACKGROUND

Ensemble forecasts of high-impact, rare weather events generated from models that explicitly resolve convection have shown potential for greater skill and operational value over a single deterministic model (e.g., Elmore et al. 2002a,b, 2003; Kong et al 2006, 2007). Elmore et al. (2002, 2003) employed a storm-scale ensemble initialized with horizontally homogeneous environments and used model-generated storms with lifetimes of at least 60 minutes as proxies for severe weather reports. Results based on this methodology indicate forecast skill on days that are more likely to experiencing severe weather. Kong et al. (2006, 2007) also extended ensemble forecasting to the convective scale by applying a full-physics numerical prediction system initialized with fine-scale observations from the WSR-88D radar network to a tornadic supercell event. Five-member ensemble systems were constructed on nested grids with 24-, 6-, and 3-km grid spacing. With explicit convection and assimilation of radar data, ensemble output from the 3-km configuration generally better predicted overall storm structure and translation compared to the output from the grids with decreased resolution.

Clark et al. (2009) compared the precipitation forecast skill of a 5-member convection-allowing ensemble on a 4-km grid to a 15-member convection-parameterized ensemble with 20-km grid spacing. The results revealed that the storm-scale ensemble had a tendency to improve upon the timing and location of rainfall systems when compared to the coarser-resolution ensemble. Additionally, it was found that the probability-matched mean precipitation forecasts from the convection-allowing ensemble

often outperformed the parameterized ensemble mean forecasts. ROC scores also demonstrated that probabilistic forecasts derived from the storm-scale ensemble outperformed those forecasts derived from the ensemble with parameterized convection.

Using model output from the CAPS 10-member storm-scale ensemble generated for the 2007 HWT Spring Experiment, Schwartz et al. (2010) evaluated the sensitivities of the WRF-ARW to varying microphysics and planetary boundary layer parameterization schemes. The findings showed that the ensemble members on average had a high precipitation bias. Additionally, a "neighborhood" approach to extract probabilities was introduced and compared to traditional point-based probabilities. Verification scores indicated that use of a neighborhood approach to calculate probabilities can enhance the forecast skill of high-resolution ensembles.

Playing a pivotal role in utilizing output from state-of-the-art numerical models for forecast guidance, the Storm Prediction Center (SPC) has worked closely with CAPS, the National Severe Storms Laboratory, and other partners within the HWT to develop and test ensemble applications for operational severe weather forecasting. Emphasis is currently being placed on evaluating the utility of ensemble output at convection-allowing resolutions in the annual real-time spring forecasting experiments. This study and the data used herein are a direct consequence of this unique collaborative program.

3. DATA

3.1. CAPS Storm Scale Ensemble Forecast System

During the 2008 HWT Spring Experiment, CAPS produced a Storm Scale Ensemble Forecast (SSEF) system that was employed as experimental guidance for the prediction of severe convective weather. The ensemble, comprised of 10 hybrid members, was generated from the Weather and Research Forecast (WRF) model which used the Advanced Research WRF (ARW) dynamic core version 2.2 (does not contain convective parameterization). At 4 km grid spacing, the ensemble's spatial domain covered approximately the eastern two-thirds of the CONUS (Fig. 1) and contained 903 x 675 x 53 grid points. Daily 30-hour forecasts were generated for the Spring Experiment beginning 14 April 2008 to 06 June 2008. The dataset for this study includes 37 days from this period for which there were no missing data.

The SSEF consisted of two control members (cn and c0) and eight perturbed members that are initialized at 00 UTC. Interpolations from the 00 UTC 12-km NAM analysis provided the background initial conditions for all members. Additionally, the 00 UTC 12-km NAM forecast provided the lateral boundary conditions for the two SSEF control members. Four pairs of negative/positive perturbations from NCEP Short Range Ensemble Forecast (SREF) members (two each from WRF-em, WRF-nmm, eta-KF, and eta-BMJ) are applied to the members to introduce mesoscale perturbations into the initial conditions. Lateral boundary condition perturbations were supplied by the corresponding three-hour forecasts of the eight 21UTC SREF members scaled to their initial perturbation amplitude. To introduce convective-scale perturbations, level-2 radial velocity and reflectivity data from operational WSR-88D radar network across the CONUS are assimilated into all but one of the control members (c0) using the CAPS Advanced Regional Prediction System (ARPS) Three-Dimensional Variational (3DVAR) system and cloud analysis package. Table 1 describes the configuration of each ensemble member including the IC and physics perturbations. In the table, NAM-a and NAM-f refer to 12-km NAM analysis and forecast, respectively.

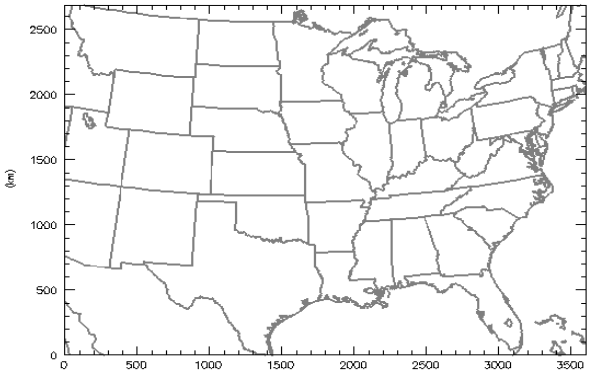

Fig. 1: SSEF spatial domain.

Member	Initial Conditions	Lateral Boundary Conditions	Assimilated Radar Data	Micro-Physics	Short-wave Physics	Planetary Boundary Layer Physics
cn	00Z NAM – a	00Z NAM – f	Yes	Thompson	Goddard	MYJ
c0	00Z NAM – a	00Z NAM – f	no	Thompson	Goddard	MYJ
n1	cn – em_pert	21Z SREF em_n1	yes	Ferrier	Goddard	YSU
p1	cn + em_pert	21Z SREF em_p1	yes	WSM 6-class	Dudhia	MYJ
n2	cn – nmm_pert	21Z SREF nmm_n2	yes	Thompson	Goddard	MYJ
p2	cn + nmm_pert	21Z SREF nmm_p2	yes	WSM 6-class	Dudhia	YSU
n3	cn – etaKF_pert	21Z SREF etaKF_n3	yes	Thompson	Dudhia	YSU
p3	cn + etaKF_pert	21Z SREF etaKF_p3	yes	Ferrier	Goddard	MYJ
n4	cn – etaBMJ_pert	21Z SREF etaBMJ_n4	yes	WSM 6-class	Goddard	MYJ
p4	cn + etaBMJ_pert	21Z SREF etaBMJ_p4	yes	Thompson	Dudhia	YSU

Table 1: SSEF member configuration. Long-wave radiation physics configuration is RRTM and surface physics configuration is Noah for all members. There is no cumulus parameterization.

3.2 Diagnostic Output Fields

Two-dimensional diagnostic fields generated from the ensemble were mined and verified against observed severe local storm reports. The forecast parameters investigated were top-of-the-hour updraft helicity (UH), 3-6 km column maximum vertical velocity (VVEL), and 4-km AGL simulated reflectivity (REFL). These fields were chosen due to their

predictive capabilities of hazardous convective weather (e.g., Sobash et al. 2008).

In convection-allowing numerical systems, rotating updrafts (i.e., supercells) can be identified by measuring the vertical component of helicity and integrating over a vertical layer to produce a measure of UH (Kain et al. 2008). This variable can be interpreted as a proxy for mesocyclone generation in the model. While updrafts can rotate cyclonically and anticyclonically, only cyclonically rotating updrafts, represented by positive UH values, were considered in this study. UH is computed by calculating the local product of vertical velocity and vertical vorticity and averaging over a vertical layer.

$$UH = \int w\zeta \, dz, \quad (1)$$

where ζ is the vertical component of relative vorticity (s^{-1}) and w is vertical velocity (ms^{-1}). In this study, UH was integrated from z_o = 2-km and z_t = 5-km AGL using a midpoint approximation. With data available every 1000 meters AGL, equation (1) is computed as

$$UH = \sum_{z=2km}^{z=5km} (\overline{w\zeta}\Delta z) = (\overline{w\zeta}_{2,3} + \overline{w\zeta}_{3,4} + \overline{w\zeta}_{4,5}) \times 1000 \quad (2)$$

The overbar indicates a layer average, delta z is the depth of the layer (1000 m), and the subscripts indicate the vertical layers over which the variable is averaged.

The 4-km AGL simulated REFL (units of dBZ) is a derived field based on the predicted concentration of hydrometeors. With simulated REFL not being mathematically equivalent to observed REFL, the REFL output is considered a proxy for observed REFL (Kain et al. 2008). The simulated REFL fields are based on hydrometeors at fixed altitudes while observed REFL is based on the detection by a radar beam at a given elevation angle. For the latter, hydrometeors closer to the radar are detected at lower altitudes than the hydrometeors farther from the radar.

The model climatology of each of the diagnostic fields used in this study was examined for the 2008 SSEF. For every member, the distribution and mean of each forecast field across the domain is tabulated. For UH and VVEL, the distributions of forecast values exceeding a specified threshold at each grid point coinciding with a minimum forecast REFL value were binned. This provided a distribution of values for the diagnostic variables where the model generated convection. Data were analyzed at grid points for UH (VVEL) meeting a minimum threshold of 25 m^2s^{-2} (4 ms^{-1}) co-located with REFL values of at least 20 dBZ. Results revealed that over one-third of these grid points averaged over all members have UH values exceeding 25 m^2s^{-2} while a small fraction (~1%) have UH values greater than 200 m^2s^{-2}. VVEL exceeds 10 ms^{-1} for just over 5% of the grid points.

Based on appropriate thresholds for each forecast field as determined by the climatological distribution, model data were contoured to create continuous 'features' for each date, forecast hour, and member. UH data were contoured at thresholds of 25, 50, 75, 100, 200 m^2s^{-2}. Vertical velocity data were contoured at thresholds of 10, 15, and 20 ms^{-1}. Lastly, 4-km AGL REFL data were contoured at thresholds of 30, 40, and 50 dBZ.

3.3. Verifying Storm Report Data

Hail, wind, and tornado report datasets from the NOAA National Climatic Data Center (NCDC) Storm Data publication were used as verification in this study. Report datasets include begin and end time, begin and end latitude and longitude, and magnitude (if available). Storm report locations are mapped to 4-km grids identical to the SSEF data grids and converted to XY space.

4. METHODOLOGY

As numerical models have increased in horizontal resolution, the scale of features resolved by models has decreased. As a result, verifying probabilistic forecasts of discrete diagnostic fields for rare events is a challenging task. Point-to-point verification methods of forecasts for rare events reveal considerable displacement errors and will often misrepresent the true skill of a forecast (Gallus 2002, Baldwin and Kain 2006). To account for this, multiple techniques were considered in this study for the development of probabilistic forecasts of diagnostic fields at various thresholds. The first method is a traditional at-the-grid point frequency-based probability. The second method utilizes a neighborhood approach by determining whether a threshold is exceeded within a given radius of the grid point to extract probabilities. For each method, a spatial smoother was applied to the probability fields which recognized the spatial uncertainty in probabilistic forecasts. Verification procedures were applied similarly to each probabilistic forecast extraction method in order to compare and discern the relative skill of each technique in forecasting the occurrence of severe weather.

4.1 Probabilistic Forecast Extraction

4.1.1 Traditional Approach

A traditional method for computing forecast probabilities exploits model output from each ensemble member at individual grid points. The traditional ensemble probabilities (TEPs) in this study were determined by taking the average of the binary probabilities (BPs) from each of the ensemble members. BPs were generated based on the occurrence or nonoccurrence of a grid point exceeding a specified threshold, x, of the forecast variable. At each grid point, the BP is given by

$$BP_{ij} = \begin{cases} 1 & if \quad F_{ij} \geq x \\ 0 & if \quad F_{ij} < x \end{cases}, \qquad (3)$$

where F_{ij} is a diagnostic field forecast and x is a threshold value. Subscript *i* refers ith ensemble member and *j* denotes the jth grid point. The TEPs were then computed as the following

$$TEP_j = \frac{1}{N} \sum_{i=1}^{N} BP_{ij}, \qquad (4)$$

where N is the total number of ensemble members.

4.1.2 Neighborhood Approach

An alternative technique to extracting probabilities is based on a 'neighborhood' approach and is employed to account for displacement errors that are inherent in convective scale prediction but are not accounted for by the traditional method. To create the TEP grids with relaxed spatial criteria, the sum of members with forecasts exceeding a specified threshold within the radius of influence (ROI) of a grid point was divided by the total number of members. This quantity is referred to as a Binary Neighborhood Ensemble Probability (BNEP) and defined mathematically as

$$BNEP_{jROI} = \frac{1}{N} \sum_{i=1}^{N} BP_{iROI}, \qquad (5)$$

where *j* is the grid point, *i* is the ensemble member, N is the total number of ensemble members, and *ROI* is the specified radius of influence.

4.1.3 Gaussian Smoother

A two-dimensional Gaussian kernel smoothing operator was applied to the ensemble probabilities generated from the various methods to arrive at smoothed probabilities. This allows for spatial uncertainty to be represented in the probabilistic forecasts. This method is based on that of Brooks et al. (1998). Brooks et al. (1998) sought to produce "Practically Perfect" forecasts which are forecasts as accurate as can be expected given prior knowledge of locations of events. Practically Perfect forecasts are generated by smoothing events using a nonparametric density estimation with a two-dimensional Gaussian kernel (Silverman 1986). This is given by

$$f = \sum_{n=1}^{N} \frac{1}{2\pi\sigma^2} \exp\left(-0.5 \left[\frac{d_n}{\sigma} \right]^2 \right), \qquad (6)$$

where N is the total number of grid points, σ is a weighting function, and d_n is the distance from the forecast grid point to the nth location where an event occurred (Brooks et al. 1998). The values of σ tested in this study were 5, 10, 20, and 30 grid points which correspond to 20-, 40-, 80-, and 120-km, respectively, where a larger magnitude of σ reflects an increase in uncertainty.

4.2 Verification Methods

Spatial and temporal post processing procedures were applied to the raw ensemble data before performing verification. To allow for temporal uncertainty in the model forecasts, output grids were created that contain the maximum value of each parameter within 60 minutes of the valid forecast hour at each grid point. For example, the 20-hour forecast grid contains the maximum value at each grid point from the 19-, 20-, and 21-hour forecast grids. Second, a land/sea mask was applied to the model output data to account for only the region where storm report data were available. Only model data bounded by the borders of the CONUS, excluding the Great Lakes, were verified.

Model data were extracted from the 20-28 forecast hours (valid 20 – 04 UTC) for verification. This time frame was chosen to focus on the period with the highest climatological frequency of severe weather occurrence.

The quality of the probabilistic forecasts was evaluated quantitatively through several verification measures. Contingency tables were created to calculate various skill scores. The reliability and resolution of the forecasts were assessed subjectively through reliability diagrams and relative operating characteristic (ROC) curves, and objectively through the Brier score and area under the ROC curve (AUC) metrics.

When computing the verification metrics, storm reports located within 25 miles (~40 km) of a forecast location were deemed correct 'yes' forecasts. This procedure is consistent with the Storm Prediction Center outlook forecasts.

4.2.1 Contingency Table Statistics

Conventionally, verification data of deterministic forecasts are expressed using an I x J contingency table which represents the joint distribution of forecasts and observations. Dichotomous forecasts are represented by a 2 x 2 table which shows the frequencies of paired 'yes' forecasts and occurrence of the event, 'no' forecasts and the nonoccurrence of an event, 'yes' forecasts and the nonoccurrence of an event, and 'no' forecasts and the occurrence of an event. The construction of a contingency table was applied to the probability forecasts of output variables defined by a threshold. By setting probability thresholds, a probabilistic forecast was converted to a dichotomous yes/no forecast. As labeled in Fig. 2, quadrant **a** corresponds to *hits*, **b** is referred to as

false alarms, **c** is called *misses,* and **d** is called *correct negative*. A variety of scalar performance metrics and skill scores can be derived from the contingency table.

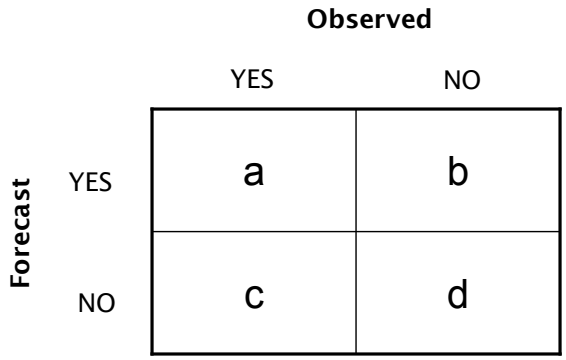

Fig. 2: 2 x 2 contingency table (Wilks 1995)

4.2.1.1 HIT RATE

One metric, known as either the hit rate (H) or probability of detection (POD), is the ratio of correct forecasts to the total number of occurrences of the event. The score ranges from 0 to 1, where 1 is a perfect forecast. Using the categories from the contingency table, H is defined as

$$H = \frac{a}{a + c}. \quad (7)$$

4.2.1.2 FALSE ALARM RATE

The ratio of false alarms to the total number of nonoccurrences of the an event is called the false alarm rate (F), or probability of false detection (POFD), and is given by

$$F = \frac{b}{b + d}. \quad (8)$$

Values range from 0 to 1 with a perfect score being 0.

4.2.1.3 BIAS

The bias score measures the ratio of frequency of forecasted events to the occurrence of observed events. Scores range from 0 to ∞, where 1 is a perfect score, less than one indicates the model tends to underforecast, and greater than one reflects a tendency to overforecast. This metric is given by

$$\text{Bias} = \frac{a + b}{a + c}. \quad (9)$$

4.2.1.4 CRITICAL SUCCESS INDEX

The Critical Success Index (CSI) measures the correspondence of forecasted events to observed events. CSI scores take on values ranging from 0 to 1. A score of 0 indicates no skill and 1 is a perfect score. CSI is defined as

$$\text{CSI} = \frac{a}{a + b + c}. \quad (10)$$

4.2.1.5 HEIDKE SKILL SCORE

To measure the fraction of correct forecasts after which forecasts that would be correct by pure coincidence are removed, the Heidke Skill Score (HSS) is computed. The range of values is from -∞ to 1, where 0 indicates no skill and 1 is a perfect score. HSS is given by

$$\text{HSS} = \frac{2(ad + bc)}{[(a + c)(c + d) + (a + b)(b + d)]}. \quad (11)$$

4.2.2 Brier Score

The Brier Score (BS) is the mean squared error of the probability forecasts. Scores range from 0 to 1 with 0 being a perfect forecast. The BS is calculated by

$$\text{BS} = \frac{1}{N} \sum_{j=1}^{N} (p_j - o_j)^2, \quad (12)$$

where *N* is the total number of grid points, *j* is the grid point, *p* is the probability forecast, and *o* is the observation in binary format.

4.2.3 Reliability Diagrams

A reliability diagram is a graphical method for assessing the agreement between probabilistic forecasts of weather events and the occurrence of the events, with the observation relative frequency plotted against the forecast probabilities. A perfect forecast would result in a line that is oriented from the lower left corner to the upper right corner. The forecasts are said to be conditionally biased when the resultant curve deviates from the perfect reliability line. Over-forecasting is represented when the curve is below the line and under-forecasting is signified when the curve is oriented above the line. Reliability diagrams are produced for TEPs and BNEPs.

4.2.4 ROC Curves

A ROC curve is a graphical tool developed by Mason (1972) for evaluating the resolution of a forecast system by plotting the hit rate and false alarm rates that are derived from the contingency table statistics for the probability thresholds. For a single quantitative measure from the ROC curve, the area under the curve (AUC) is computed. A perfect forecast would be represented by an AUC = 1 while an AUC = 0.5 signifies random forecasts (Marzban 2004). AUC values greater than ~0.7 are generally considered to represent useful probabilistic forecast that discriminate between events and non-events (Stensrud and Yussouf 2007). ROC curves were generated and AUCs were calculated for the probabilities generated using the various methods described above.

5. Results

5.1 Traditional Ensemble Probabilities

Computation of the various skill scores for the TEP forecasts yielded results that were inferior to the results from the BNEP forecasts, which revealed the difficulty in forecasting rare events at the grid point. Skill scores for UH (Fig. 3a) and VVEL (not shown) tended to peak at the lowest probabilities (~10%) for the lowest thresholds of each parameter. This is a reflection of the low frequency of members forecasting discrete variables at the same grid point. The scores decreased for these two parameters with an increase in threshold. Applying a smoother slightly increased the score but only for the lowest thresholds.

Trends in the skill scores for REFL deviated from the tendencies associated with the other two parameters (not shown). The highest skill scores were associated with slightly higher probabilities and increased when the forecast threshold was raised to 40 dBZ. The values also increased when a 20-grid point smoother was applied to the forecasts. Overall, TEP forecasts of REFL remained inferior to the BNEP forecasts of REFL despite the improvements associated with the application of a smoothing function.

Low ROC AUC values (< 0.7) for UH (Fig. 3b) and VVEL (not shown) further illustrated the shortcomings of the forecasts associated with TEPs.

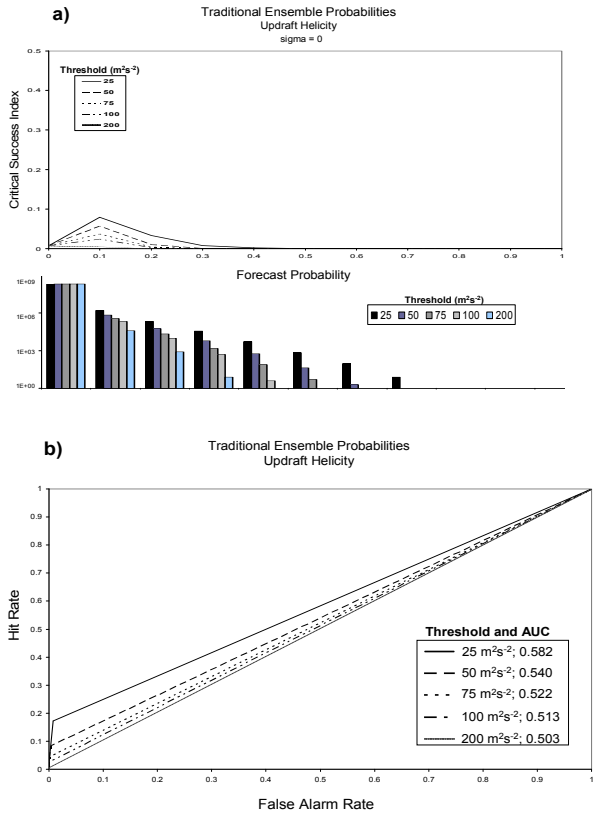

Fig. 3: a) CSI values for UH with distribution frequency for each probability threshold binned on the bottom and b) ROC curves for UH with AUC totals noted in the inset.

5.2 Binary Neighborhood Ensemble Probabilities

5.2.1 CSI and HSS

CSI and HSS values for UH were greater than the values corresponding to the VVEL and REFL parameters. Forecasts of UH ≥ 25 m^2s^{-2} (Fig. 4) provided the most skillful results as increasing the threshold dampened the peak in CSI and HSS. VVEL forecasts resulted in more favorable scores when the threshold was increased from 10 ms^{-1} to 15 ms^{-1}; however, increasing the threshold further worsened the skill scores. For the REFL forecasts (not shown), skill scores improved with each increase in threshold. All three parameters exhibited a shift in the maximum skill towards higher probabilities as the ROI increased. Beyond a 50-mile ROI, the values of the maximum HSS and CSI began to decrease. Adding a smoother to each of the forecasts proved to be beneficial for the lowest thresholds. A 30-grid point smoother applied to forecasts of UH ≥ 25 m^2s^{-2} within a 50- mile ROI yielded the maximum CSI and HSS values among all forecasts. This is indicated in CSI plots for UH (Fig. 4) in which the shift and increase in the peak CSI value correspond to expanding ROI to 50 miles and increasing the smoother to 30 grid points. The effect of adding a smoother to the forecasts is more evident in an upcoming example.

HSS scores mirrored the CSI magnitude and trends and are not shown.

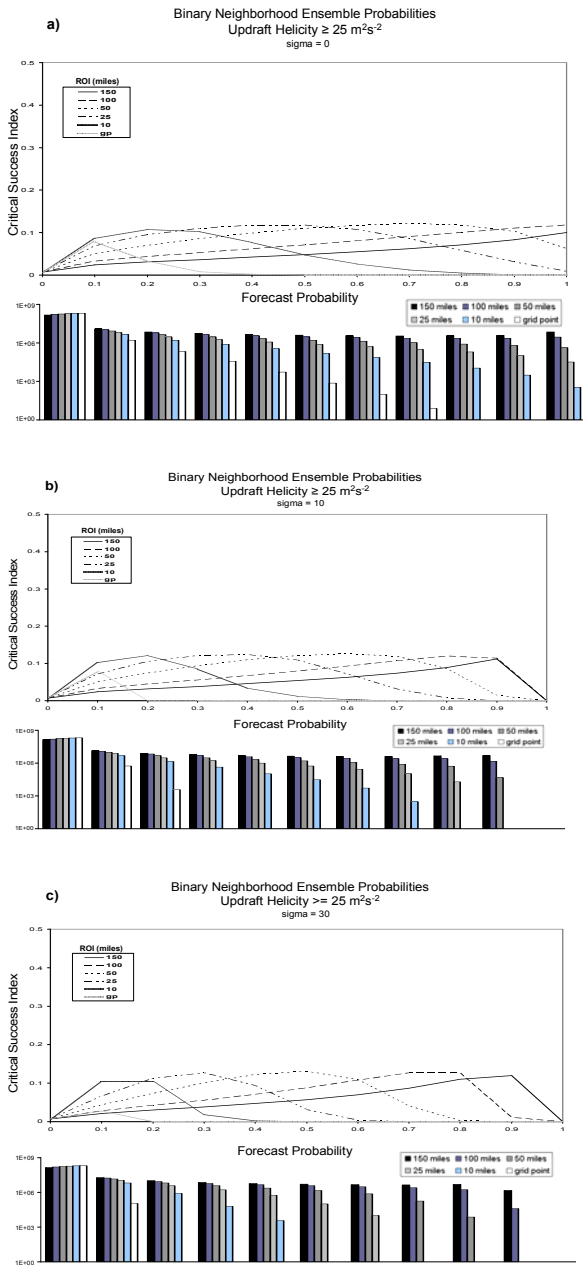

Fig. 4: CSI for forecasts of UH ≥ 25 m^2s^{-2} for the five ROI distances with a) no smoothing, b) sigma = 10 grid points, and c) sigma = 30 grid points with forecast frequency distribution binned below the plot.

5.2.2 ROC AUC

Application of a ROI > 10 miles resulted in AUC values that were greater than 0.7 for each parameter, which was not realized with the TEP forecasts. For UH, this criterion was met with an ROI of 10 miles (Fig. 5). The ROI must be increased further, however, as the threshold values increase in order to reach an AUC ≥ 0.7. All VVEL forecasts (not shown) reached the 0.7 criterion with an ROI ≥ 25 miles. REFL forecasts (not shown) at all thresholds resulted in AUC totals ≥ 0.7 at the lowest ROI distance (i.e., 10 miles). Increasing the UH and VVEL forecast thresholds decreased the AUC values; however, increasing the REFL threshold led to an increase in the AUC. Applying a smoother slightly improved the results for each parameter at a given threshold and ROI. ROC curves for the largest AUC found in this study (i.e., UH ≥ 25 m^2s^{-2}) are plotted in Fig. 5, which indicate AUC values are improved when the threshold is increased out to 150 miles and sigma is increased to 30 grid points. This reiterates that grid point predictability is low on the convective scale, and neighborhood approaches are needed to supply more operationally valuable probability forecasts.

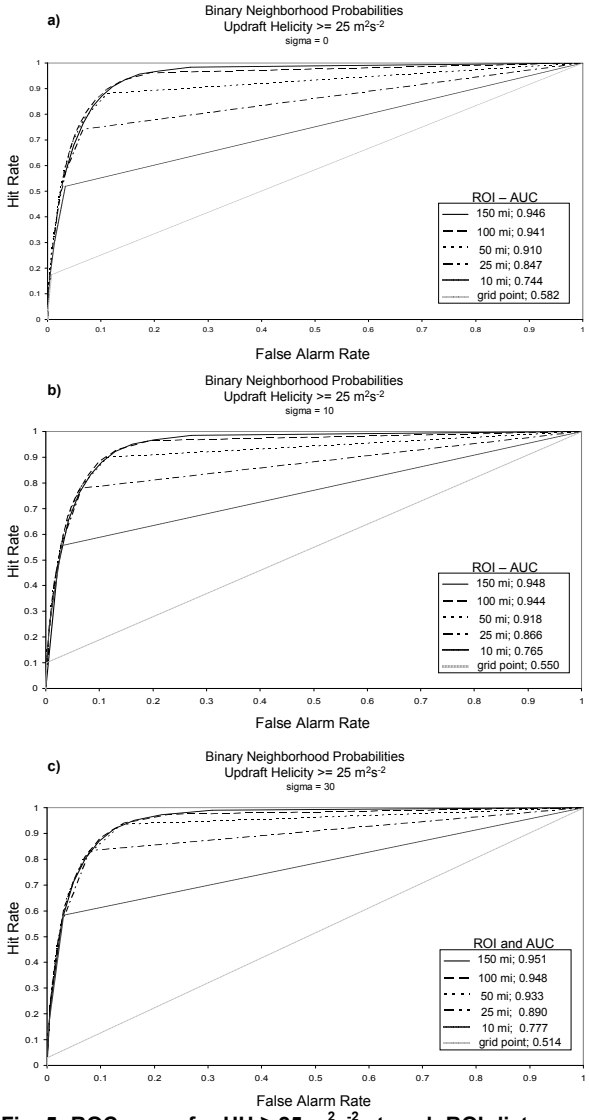

Fig. 5: ROC curve for UH ≥ 25 m^2s^{-2} at each ROI distance with forecasts with a) no smoothing, b) sigma = 10 grid points, and c) sigma = 30 grid points. AUC totals are indicated in the inset for the different forecasts.

5.2.3 Bias Score

High bias scores were associated with BNEP forecasts for the three parameters at the varying thresholds and degrees of smoothing (not shown). Similar to the TEP forecasts, the values were > 100 at the lowest probabilities then decreased towards zero with increasing probabilities. The bias never stabilized at a particular value.

Increasing the threshold tended to decrease the high bias values and shift the point at which the scores indicated a transition from overforecasting to underforecasting to higher probabilities. Decreasing the ROI and increasing the smoother had the same effect as increasing the threshold. UH forecasts had lower bias scores compared to the other two variables.

5.2.4 BS

The magnitude of the BS was similar for all probabilistic forecasts of UH and VVEL (Table 2) at the varying thresholds, ROI, and sigma values. The low scores associated with the forecasts give the impression of very skillful forecasts. It should be recognized, however, that the low scores are in part due to the very large number of correct-negative forecasts. The lowest BS scores (highest skill) are realized at the lower ROI, largely due to smaller forecast area. Forecast of REFL ≥ 30 dBZ had the largest values (i.e., lowest skill). This can be related to the more frequent and widespread forecasts of this field with respect to severe storm occurrence overforecasting.

Brier Score						
Updraft Helicity (m^2s^{-2})						
sigma	**grid point**	**10 miles**	**25 miles**	**50 miles**	**100 miles**	**150 miles**
0	0.0066	0.0067	0.0094	0.0184	0.0471	0.0847
10	0.0065	0.0062	0.0076	0.0136	0.0342	0.0618
30	0.0066	0.0061	0.0067	0.0109	0.0288	0.0546
Vertical Velocity (ms^{-1})						
sigma	**grid point**	**10 miles**	**25 miles**	**50 miles**	**100 miles**	**150 miles**
0	0.0066	0.0075	0.0128	0.0262	0.0623	0.1045
10	0.0064	0.0065	0.0097	0.0190	0.0452	0.0763
30	0.0065	0.0062	0.0082	0.0153	0.0387	0.0682
Reflectivity (dBZ)						
sigma	**grid point**	**10 miles**	**25 miles**	**50 miles**	**100 miles**	**150 miles**
0	0.0158	0.0392	0.0731	0.1262	0.2285	0.3259
10	0.0113	0.0268	0.0510	0.0900	0.1658	0.2382
30	0.0096	0.0214	0.0413	0.0760	0.1479	0.2183

Table 2: Brier Score for a) updraft helicity (UH), b) vertical velocity (VVEL), and c) reflectivity (REFL) with varying values of sigma (grid points).

Reliability diagrams for UH exemplify the propensity for the discrete variables to overforecast majority of the time (Fig. 8). Applying a smoother adjusts for this; even leading to underforecasting dominating at the lower probabilities. The curves associated with the largest values of sigma are closer to the diagonal line which indicates more reliability and this is reflected in the BS for the varying sigma values. The most reliable forecasts in terms of BS and reliability diagrams correspond to UH ≥ 25 m^2s^{-2} at a 10-mile ROI with a 30-grid point smoother.

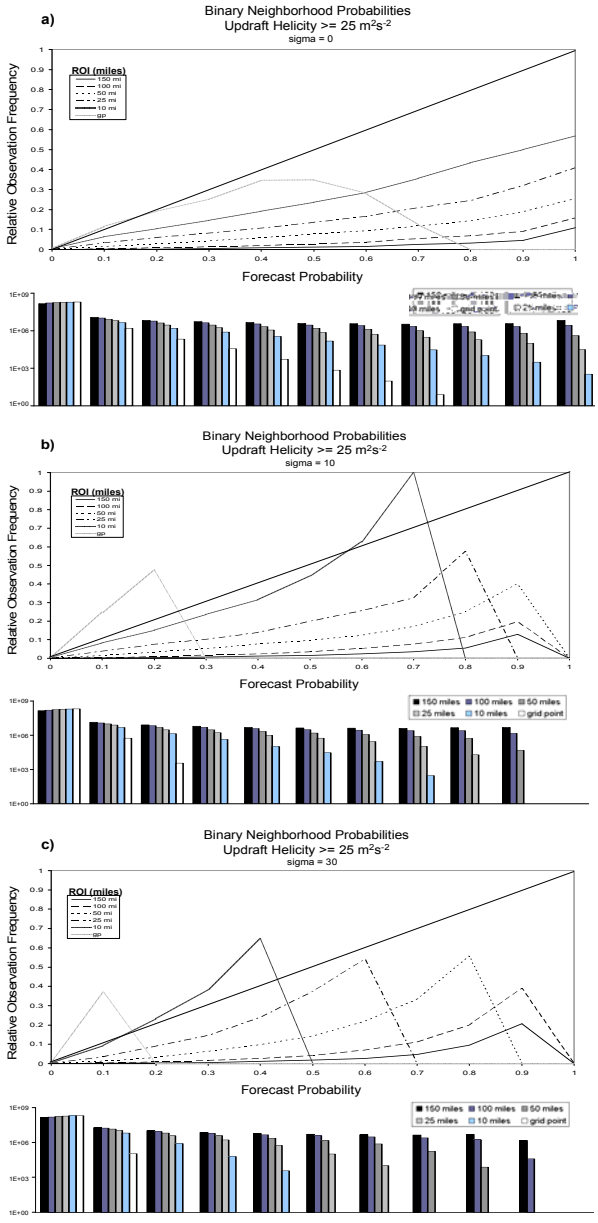

Fig. 8: Reliability diagram for forecasts of UH ≥ 25 m2s-2 at each ROI with a) no smoothing, b) sigma = 10 grid points, and c) sigma = 30 grid points. Below the reliability diagram is a histogram of forecast frequency for the varying ROI.

6. Discussion

Various skill scores were presented that quantitatively described the quality of the forecasts derived from the TEPs and the BNEPs. The lower skill scores associated with the TEP forecasts relative to BNEP forecasts demonstrated the benefit of using a temporal and spatial neighborhood approach. This is especially true for UH which can be used as a powerful tool for isolating locations of rotating updrafts. UH is an infrequent and discrete variable that proved to be difficult to verify at the grid point.

When using a neighborhood approach, values of CSI, HSS, BS and AUC indicated UH at the lowest threshold (25 m^2s^{-2}) was the most effective in identifying potential regions of severe convective weather. These scores further indicated that applying a smoother yielded more skillful forecasts of UH. For this variable, AUC values were the greatest for an ROI of 150 miles while the BS indicated that forecasts within a 10 mile ROI were the most reliable. The reliability diagrams showed that increasing the ROI out to 150 miles led to overforecasting. Given these results, it appears appropriate to opt for an ROI in between 10 and 150 miles (i.e., 50 miles). This is substantiated by the CSI and HSS results for this parameter which showed a peak in skill scores at the 50 mile ROI. In summary, the results suggest forecasts of UH ≥ 25 m^2s^{-2} subject to a 30-grid point smoother at an ROI of 50 miles have more skill in predicting severe weather events.

7. Case Study

On 2 May 2008, the Storm Prediction Center issued a convective outlook that included a "Moderate Risk" area valid during the afternoon and evening hours for the likelihood of severe convection developing ahead of a surface cold front. The region with the greatest potential for significant severe weather included northeastern Louisiana, southeastern Arkansas and northern Mississippi (Fig. 9.), where a "hatched region" was included in the tornado outlook indicating the >10% potential for EF2-EF5 tornadoes within 25 miles of a point (Fig. 9b).

TEP and BNEP forecasts of UH ≥ 25 m^2s^{-2} with a ROI of 50 miles (i.e., the most skillful combination found in this study) are compared for this event to highlight the potential benefits of the neighborhood approach. Figure 10 displays TEP without smoothing for 23 UTC on 02 May 2008 through 02 UTC on 03 May 2008. Storm reports that occur during each corresponding hour are overlaid with the model data. As expected, the TEPs are characterized by lower probabilities of generally <30%. This reflects the inherent lower predictability of convective features, as evidenced by the relative lack of agreement among all members of the ensemble at the grid point when forecasting discrete variables (e.g., UH). The highest of these probabilities match up with the location of the outlined moderate risk region and associated threats for significant tornadoes and higher probabilities of wind and hail. Additionally, probability maxima show some agreement with the location of severe events. However, the relatively low probabilities associated with TEPs of UH do not provide convincing visual representation of a high-impact event.

Applying a neighborhood approach results in higher probabilities for this event (Fig. 11, left panel) compared to the TEPs, which provides more confidence to the forecaster that severe weather is likely to occur. Comparing the forecasts for the different smoothing values (Fig. 11), BNEPs with the lowest value of sigma have the highest probabilities (even 100% in some areas). These probability maxima correspond reasonably well with observed events. As the smoothing is increased to sigma of 30 grid points (Fig. 11, right panel), there is some evidence of higher skill and less displacement error owing to the removal of the spurious regions of lower probabilities. This is represented quantitatively in the previous sections.

Details in the forecasts afforded by a storm-scale model can be lost, however, when smoothing is applied to the probabilities. Additionally, smoothing the probabilities can remove regions where only one or two members are indicating the potential for a severe event. This can degrade a forecast of isolated events and suggest that perhaps displays of smoothed and non-smoothed output fields may both have benefits to forecasters.

SPC Day 1 Convective Outlook/Valid 20 UTC 02 May 2008 – 12 UTC 03 May 2008

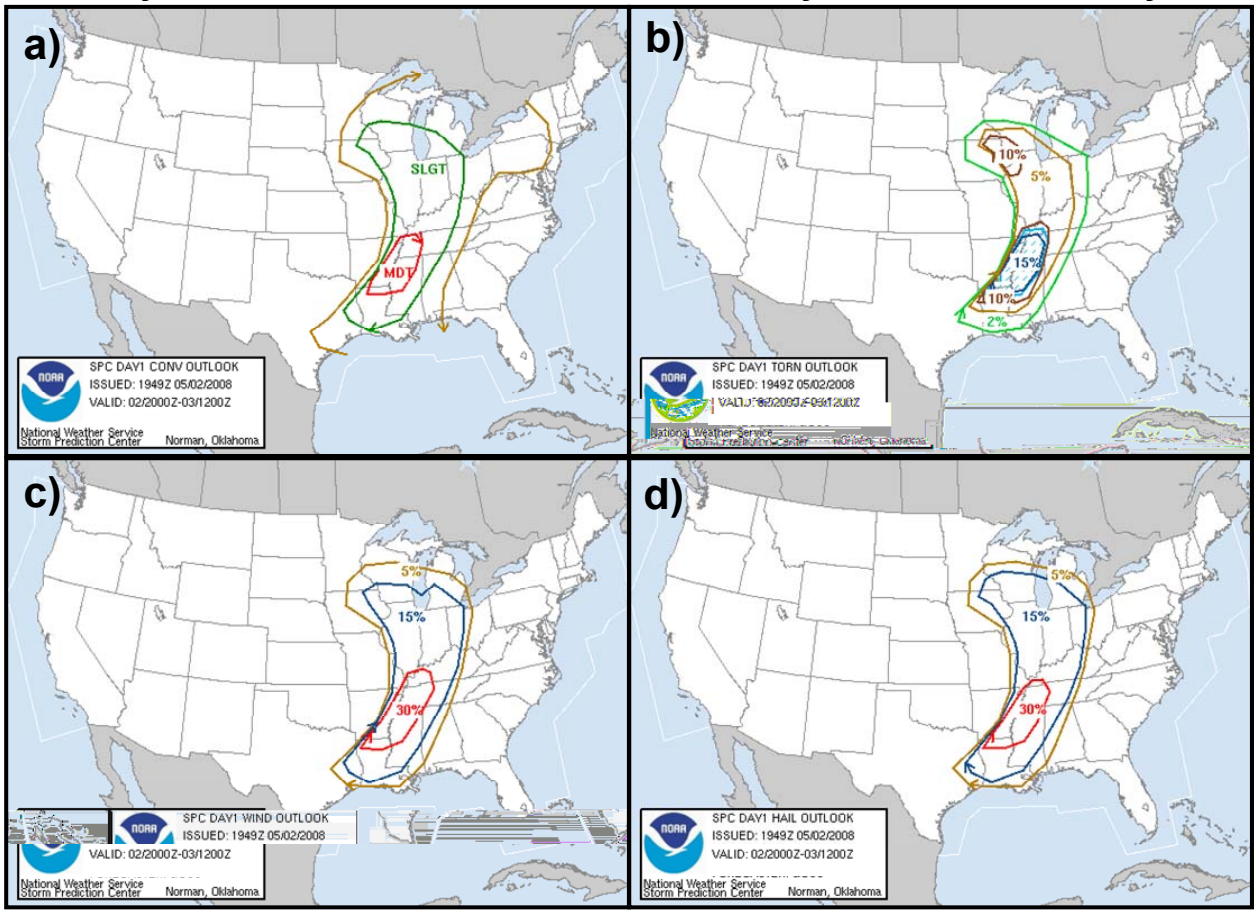

Fig. 9: (a) SPC Day 1 Convective Outlook, (b) probabilistic tornado outlook, (c), probabilistic wind outlook, and (d) probabilistic hail outlook.

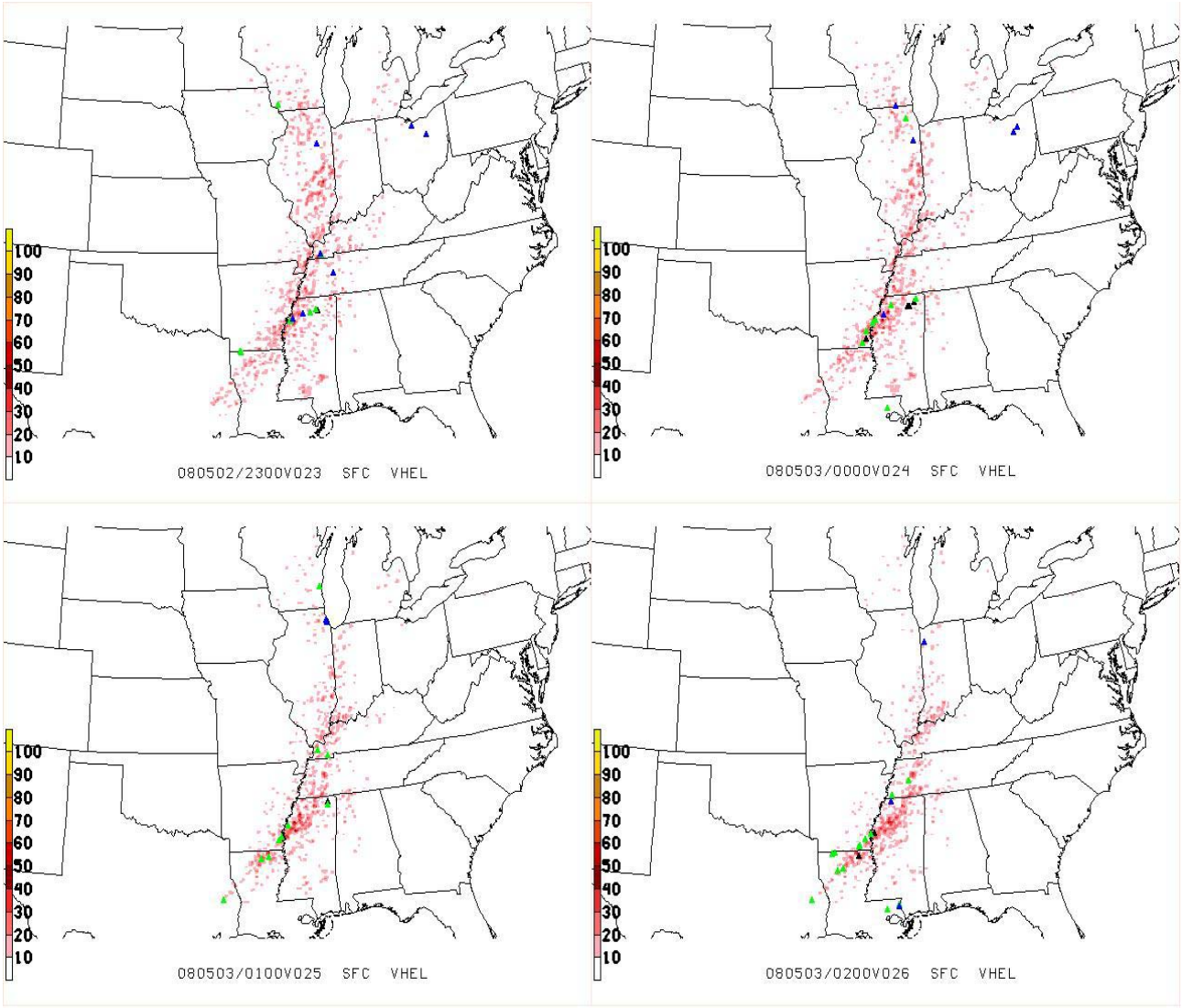

Fig. 10: Hourly SSEF TEP forecasts of UH ≥ 25 m^2s^{-2} valid 23 UTC 02 May 2008 through 02 UTC 03 May 2008 with no smoother. Triangles represent tornado (black), hail (green), and wind (blue) reports.

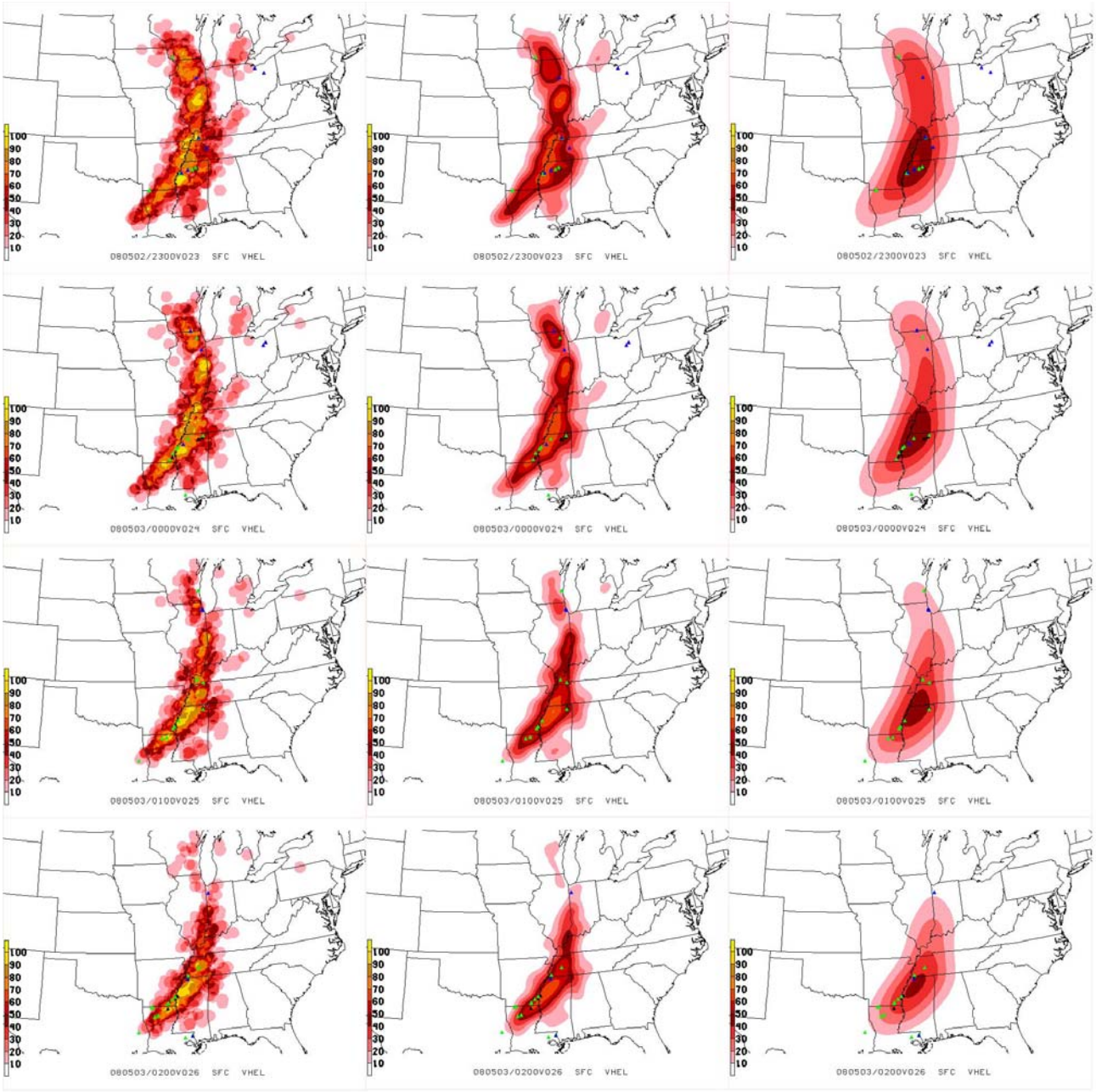

Fig. 11: Hourly SSEF BNEP forecasts of UH ≥ 25 m^2s^{-2} BNEP valid 23 UTC 02 May 2008 (first row) through 02 UTC 03 May 2008 (bottom row) with sigma = 0 (left panel), sigma = 10 grid points (center panel), and sigma = 30 grid points (right panel). Triangles represent tornado (black), hail (green), and wind (blue) reports.

8. Conclusions and Future Work

This study analyzed probabilistic forecasts of diagnostic variables derived from a storm-scale ensemble forecast system. Two methods of extracting probabilities were presented and the skill of each at predicting convective weather events was discussed. The verification scores demonstrated the quality of the forecast was improved when a neighborhood approach was applied to the probability extraction methodology. Scores were maximized at the lower thresholds for lower probabilities and the peak shifted toward higher probabilities with increase in ROI. Quantitatively, the scores indicated that applying a smoother improved the forecasts. However, the smoothed UH plots for the case study illustrate that probabilities are reduced and some details are lost in the forecast, which may inhibit a forecaster's confidence in isolating regions of severe convective weather.

Future work will explore the statistical significance in the differences among the scores associated with the varying variables, thresholds, ROI, and sigma values. Additionally, another method for extracting

probabilities based on a fractional neighborhood approach will be studied and the findings will be compared to the results presented here.

9. References

Baldwin, M.E., and J. S. Kain, 2006: Sensitivity of several performance measures to displacement error, bias, and event frequency, Wea. Forecasting, **21**, 636-648.

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

Clark, A. J., W. A. Gallus, M. Xue, F. Kong, 2009: A Comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Wea. Forecasting*, **24**, 1121–1140.

Elmore, K. L., D. J. Stensrud, K. C. Crawford, 2002a: Ensemble cloud model applications to forecasting thunderstorms. *J. Appl. Meteor.*, **41**, 363–383.

Elmore, K. L., D. J. Stensrud, K. C. Crawford, 2002b: Explicit cloud-scale models for operational forecasts: a note of caution. *Wea. Forecasting*, **17**, 873–884.

Elmore, K. L., S. J. Weiss, P. C. Banacos, 2003: forecasts: some preliminary results. *Wea. Forecasting*, **18**, 953–964.

Gallus, William A., 2002: Impact of verification grid-box size on warm-season QPF skill measures. *Wea. Forecasting*, **17**, 1296–1302.

Kain, J.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.

Kong, F., K. K. Droegemeier, N. L. Hickmon, 2006: Multiresolution ensemble forecasts of an observed tornadic thunderstorm system. Part I: comparison of coarse- and fine-grid experiments. *Mon. Wea. Rev.*, **134**, 807–833.

Kong, F., K. K. Droegemeier, N. L. Hickmon, 2007: Multiresolution ensemble forecasts of an observed tornadic thunderstorm system. Part II: storm-scale experiments. *Mon. Wea. Rev.*, **135**, 759–782.

Marzban, C., 2004: The ROC curve and the area under it as performance measures. *Wea. Forecasting*, **19**, 1106–1114.

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

Sobash, R. A., D. R. Bright, A. R. Dean, J. S. Kain, M. Coniglio, S. J. Weiss, and J. J. Levit, 2008: Severe storm forecast guidance based on explicit identification of convective phenomena in WRF-model forecasts, Preprints, 24th Conference on Severe Local Storms, Savannah, GA, Amer. Meteor. Soc., 11.3.

Stensrud, D. J., and N. Yussouf, 2007: Reliable probabilistic quantitative precipitation forecasts from Short Range Ensemble Forecasting System. Wea. Forecasting, **22**, 3-17.

Wilks, D. S., 1995: Statistical Methods in the Atmospheric Sciences: An Introduction. Academic Press, 261 pp.