VERIFICATION OF PROXY STORM REPORTS DERIVED FROM ENSEMBLE UPDRAFT HELICITY

MALLORY ROW[12],JAMES CORRIEA JR. [3], AND PATRICK MARSH[3]

[1]*National Weather Center Research Experiences for Undergraduates Program*
Norman, Oklahoma 73072

[2]*Valparaiso University Department of Geography and Meteorology*
Valparaiso, Indiana 46383

[3]*National Oceanic and Atmospheric Administration*
Storm Prediction Center
Norman, Oklahoma 73072

ABSTRACT

Convection-allowing models (CAMs) are one of the newest improvements the area of numerical weather prediction (NWP) has seen in the last 10 years. One of the new diagnostic fields these models output is updraft helicity (UH), a measure of rotation in modeled storms. Data collected from Storm Scale Ensemble of Opportunity (SSEO) and its individual members in 2012 is used to create proxy storm reports derived from UH track-like objects. Daily probabilistic forecasts are created from the reports allowing for a direct comparison to the observed for that day. 2x2 contingency tables are constructed daily to gain insight to if UH provides a skillful and reliable probabilistic serve weathers forecast and understand the characteristics of the SSEO and members. Various verification metrics are calculated along with looking at correlation data and probabilistic outlooks to provide a fuller understanding. The SSEO is found to have good skill and reliability throughout the year with especially good skill in the spring time (March to June).

1. INTRODUCTION

While severe weather forecasting remains difficult for meteorologists, the process of verifying these forecasts is just as troublesome. Without having a meaningful method of forecast verification, forecasters cannot learn and improve. Since the first attempt of tornado forecast verification in 1884 by J.P. Finley, however, both aspects of forecasting have seen tremendous improvements.

Finley's method of verification involved a dichotomous forecast, answering yes or no for forecasted and observed events (2x2 verification). After Finley, others were quick to point out his flaws and derive their own methods of forecast verification. Some have taken into account the difficulty of rare event forecasting, while others have set out to apply Finley's methods to other weather phenomenon besides tornadoes (Murphy 1996). It still remains, though, when assessing the quality of a forecast, numerous forecast verification metrics must be analyzed.

Severe weather forecasting methods have been aided by the development of convection-allowing models, numerical weather prediction models with the ability to develop convection and severe storms. Newly developed model parameters help minimize the amount of data forecasters look at when producing a forecast. One of these new convection-allowing model parameters is updraft helicity (UH).

Updraft helicity is a characteristic of rotating storms. It is mathematically defined as the integral of vertical vorticity multiplied by the updraft velocity between the 2 and 5km above ground layer, Kain et al. (2008). It allows forecasters to see simply on one map where rotating storms are occurring.

[1] *Corresponding author address:* Mallory Row, National Weather Center Research Experiences for Undergraduates Program, Center for Analysis and Prediction of Storms, The University of Oklahoma, National Weather Center, 120 David L. Boren Blvd, Suite 2500, Norman, OK 73072
Email: mallory.row@valpo.edu

Recent work by Clark et al. (2013) has shown a strong correlation between updraft helicity and tornado tracks. However, there are other types of severe weather out there besides tornadoes. This prompts the question if UH is a useful forecasting method for severe weather. Sobash et al. (2011) used UH in probabilistic forecasting from a single CAM. A grid point above a certain value of UH was flagged as a surrogate storm report. This forecast was verified against an observed probabilistic forecast, and was found to have reliability and skill. This paper extends this work using ensemble data and maxima from UH objects as surrogate reports. The data and methods will be discussed at more length in Section 2. Section 3 will provide the results of the research discussing the skill and reliability of this method, and provide insight into the characteristics of the SSEO and the members. Section 4 will provide a discussion on the main conclusions of the paper and possible future work.

2. DATA AND METHODS

a. Forecast Method

The dataset used is from the approximate 4 km grid space Storm Scale Ensemble of Opportunity (SSEO) with a domain over the central and eastern United States. The SSEO is comprised of 7 individual members including the 5.1 km Weather Research and Forecasting (WRF) –Advanced Research WRF (ARW), a time-lagged ARW (ARWL), the 4km National Severe Storms Laboratory –WRF (NSSL), Nonhydrostatic Mesoscale Model (NMM), a time-lagged NMM (NMML), a NMM with a domain over the continental United States (NMMC), and NMMB Nest (NMMB). The data was collected from the SSEO and the individual members from January 8, 2012 – December 26, 2012, where only days with observed storms reports are in the dataset. The SSEO was initialized at 0 UTC and run out to 36 hours. From this, forecast hours 13 to 36 (12-12 UTC) were used. Updraft helicity values for each grid point were obtained from the hourly maximum value.

To produce the proxy storm reports, UH track objects were created. The object is based on four pairings of thresholds 20 and 30, 40 and 50, 70 and 90, and 100 and 125 m^2s^{-2}. The pairing thresholds were based on the thresholds used in the National Oceanic and Atmospheric Administration (NOAA) Hazardous Weather Testbed (HWT). The object's lower threshold value is triggered when there are 4 contiguous pixels present, while the higher threshold is triggered by 2 pixels. The UH object's maximum value is then found, and identified a surrogate storm report. This allows for a direct comparison to observed storm reports in forecast verification. Figure 1 provides a visual of this process.

A 40km neighborhood method and 120km Gaussian smoother are applied to the surrogate storm reports to create a daily probabilistic outlook for severe weather. The highest percentage from the probabilistic outlook is recorded as the maximum value for the ensemble and the members to serve as a measure of the severity the model predicts for the day. For this research, the 15% and greater threshold is verified.

In addition, a dataset composed of the ensemble and individual member's correlation with the observations is used. The correlation data provides further explanation into the relationship between the forecast and observations.

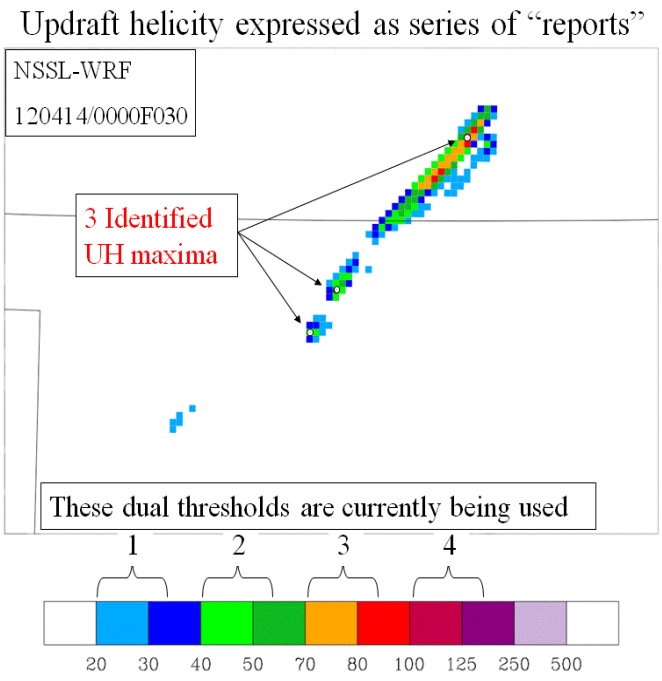


Figure 1. Example of how UH track objects are created

b. Forecast Verification

A 2x2 contingency table is used to verify the dataset. These tables are created daily based on the probabilistic forecasts for the ensemble and individual members on a grid point to grid point basis. Again, for this research, the threshold for verification is greater than or equal to 15%. A in the contingency table is defined as a hit, where for both the forecasted and observed grid point meet this threshold. B is defined as a false alarm, where the forecasted grid point meets the threshold but observed is below. C is defined as a miss, where the observed grid point meets the threshold but the forecasted is below. D is defined as a correct null forecast, where both the observed and forecasted are below the threshold. Figure 2 provides a visual explanation of the contingency table used.

From the 2x2 contingency tables, various forecast verification metrics are calculated to provide insightful information on the skillfulness and reliability of the forecast method. Each metric provides a different explanation about the forecasting method, thus numerous metrics must be calculated to tell the entire story. In this project, the following metrics were calculated: Probability of Detection (POD), False Alarm Ratio (FAR), Frequency of Hits (FOH), Bias, Critical Success Index (CSI), and Hiedke Skill Score (HSS). The monthly mean was found by adding each A, B, C, and D for the month, and calculating the specified metric from those values. These metrics will be discussed in further detail in Section 3 with the formulas for each seen below.

$$POD = A / (A+C)$$
$$FAR = B / (A+B)$$
$$FOH = 1 – FAR$$
$$Bias = (A+B) / (A+C)$$
$$CSI = A / (A+B+C)$$
$$HSS = 2(AD-BC)/ [(A+C)(C+D) + (A+B)(B+D)]$$

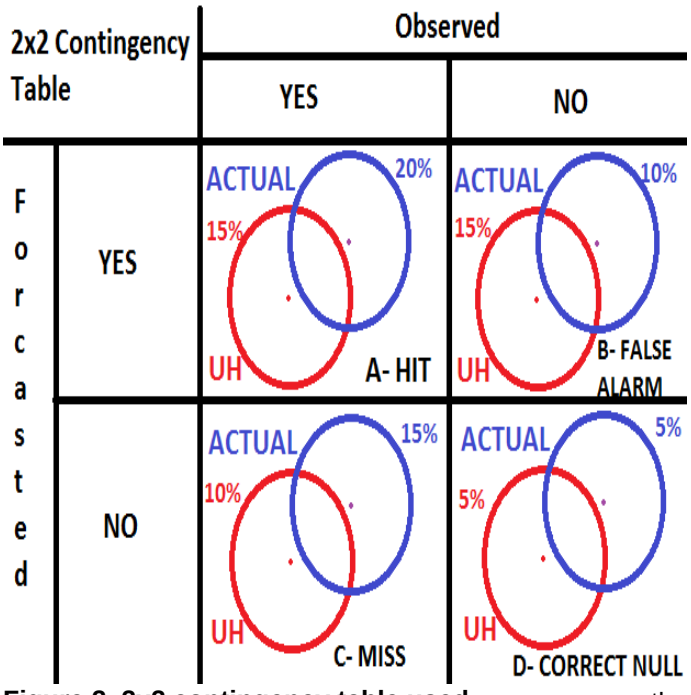


Figure 2. 2x2 contingency table used

3. RESULTS

a. 2012 Analysis

The number of days with a maximum probabilistic forecast value at or above 15% for each month is seen in Figure 3 (the ensemble in on the left hand side, and the NSSL on the right). Both models have their peaks in May, and see approximately half of the days being at or above the verifying threshold between March through June.

Probability of Detection (POD) gives insight into what fraction of observed “yes” events were forecasted correctly. This metric can be increased (improved) through issuing a larger area or number of “yes” forecasts. Thus, this leads to an increase in the number of hits and ultimately POD. Figure4a shows the POD of each month’s mean plotted through the year. The SSEO (black) falls right in the middle of the members, which is to be expected. Some members, such as the NSSL and NMM, show higher POD values than the ensemble, while other members, such as the ARW and ARWL, show lower POD values. Interesting to see is the POD time series shows the seasonality of severe weather; the higher values in March through June as the primary season with the spike in September and October the “second” season.

False Alarm Ratio (FAR) gives the fraction of forecasted “yes” events that did not occur. Figure 4b shows the time series the FAR throughout the year. The ensemble has some of the lowest FAR values with only the ARW and ARWL members being lower.

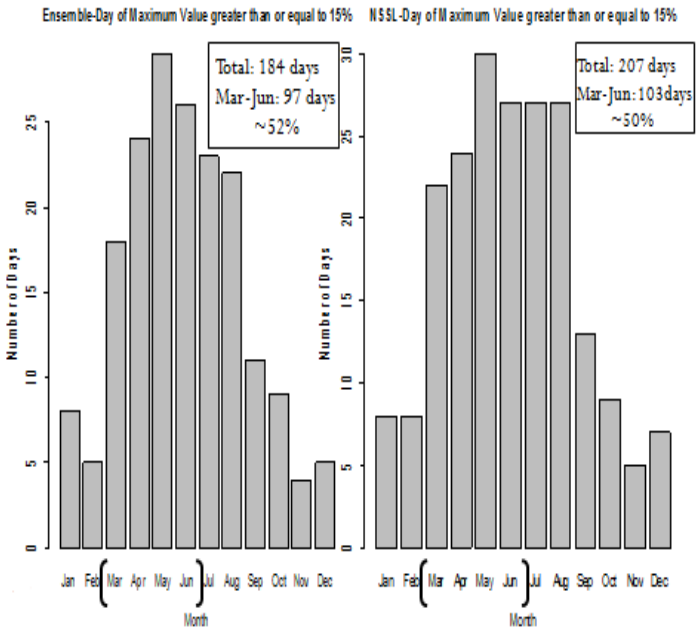


Figure 3. Days at or above 15% for SSEO (left) and NSSL (right)

Members such as the NMMC and NMM have some of the highest FAR throughout the year. Consistent for the ensemble and all members, though, is the lowest FAR are seen in the primary severe weather season.

Looking at the POD and FAR, the members with higher POD also have higher FAR. While on the other hand, the members with lower POD have lower FAR. Looking at the bias time series (Figure 4c) provides a further understanding into these findings. The NMMC, NMM, NMML, and NSSL show with a bias consistently higher than 1 (above the bold black line). A bias greater than one states the model has a tendency to overforecast. When comparing Figures 4a and b with this, these "overforecasters" have high POD values and high FAR. It is concluded that these members tend to cover a larger forecast area, thus they are rewarded with high POD but are punished with a high FAR. A bias less than one (below the bold black line) states the model has a tendency to underforecast. This is the case with ARW, ARWL, and NMMB members. When comparing to Figure 4a-c, these underforecasting models have low POD and low FAR. These "underforecasters" forecast a smaller area, and are rewarded with a low FAR but punished with a low POD. A bias of 1 states a properly sized forecast, what forecasters strive to attain. The ensemble has a relatively high POD, lower FAR, and a bias value hovering around 1. From this, it is concluded that the ensemble shows relative skill in this forecasting method.

Figure4d shows the Performance Diagram, also known as the Roebber Diagram (Roebber 2009). The best forecast in this diagram lies in the upper right corner, where the forecast has a high POD, high FOH(low FAR) along with high CSI. The overforecasters are above the bias of 1 line with their high POD and lower FOH. They are covering a larger forecast area so they are detecting area, but also have many areas that are not getting hits. The underforecasters, though the area they cover is smaller, they are still getting hits in the areas they are forecasting. The ensemble, though, has POD and FOH values that are nearly equal.

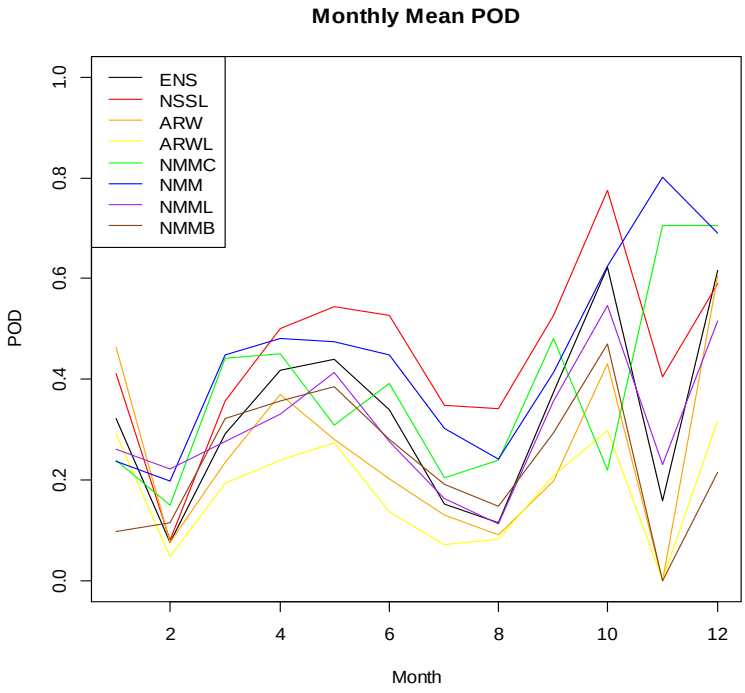


Figure 4a. Monthly mean POD time series

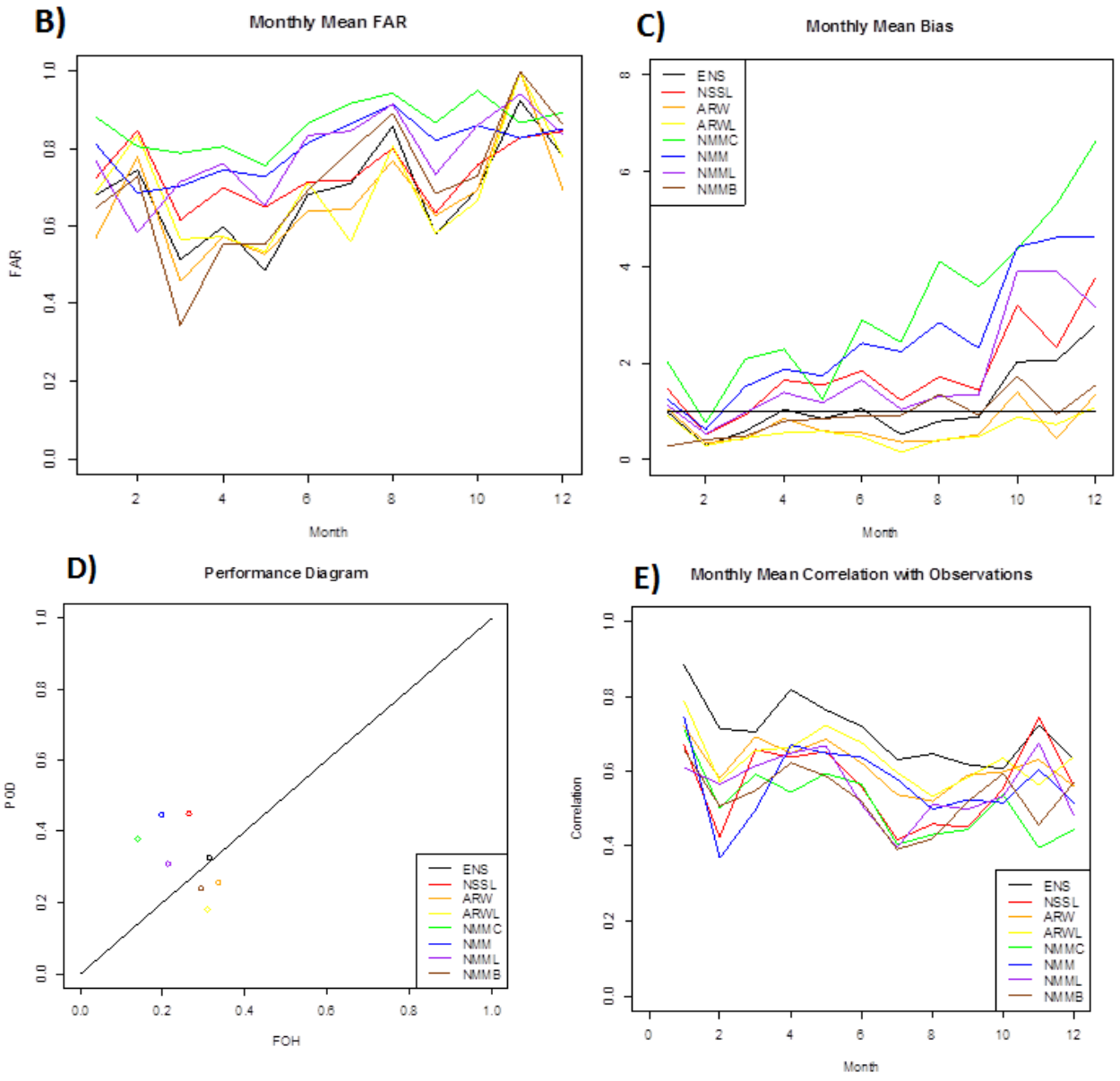


Figure 4 (b) Monthly mean FAR time series, upper left (c) Monthly mean Bias time series, upper right (d) Performance Diagram (e) Monthly mean Correlation with observations time series

Comparing Figures 4a-d, a compensation affect between the members of the SSEO is seen. Some members overforecast and some members underforecast, but together they create an overall decently skillful ensemble mean forecast. The overforecasters are toned down by the underforecasters, but by the same token the overforecasters work to enhance the underforecasters. Some members are picking up for other member's weaknesses. By this, the ensemble forecast has decent POD and FOH. The skill of the ensemble forecast is seen in Figure 4e, which shows the model's correlation with the observations. The ensemble has the highest

correlation with the observations showing the forecast has relative skill and reliability throughout the year.

b. March to June Analysis

Spring is the primary season for severe weather in the United States. Along with March to June having some of the highest POD values for the year, this span of months is also when the ensemble and members have relatively low FAR values, and contains over half of days in the dataset with a maximum probalistic forecast of 15% or greater. This prompted a further investigation of the ensemble and member performance during these months, and the affect of compensation.

A box and whisker plot of the ensemble and members maximum values is found in Figure 5a. The median values (shown by the solid black line in each box) are varied from member to member. The NSSL and NMMC have medians above 0.4 with the NMM and NMML just below, and it important to note that these are the "overforecasters". The ARW, ARWL, and NMMB are above the 0.2, and these are

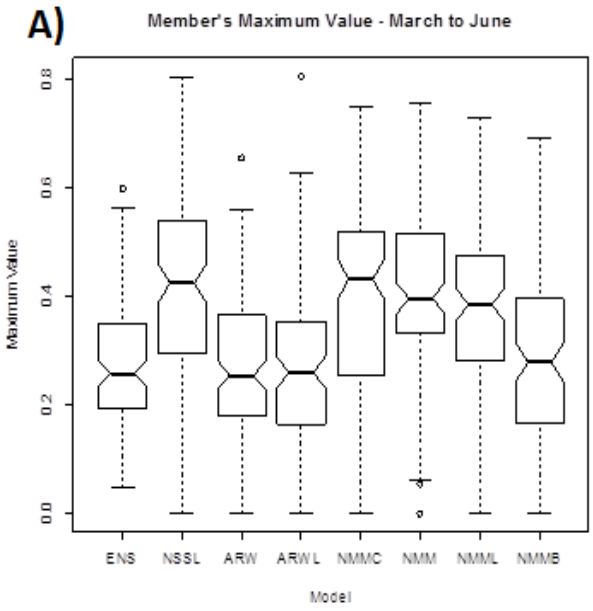
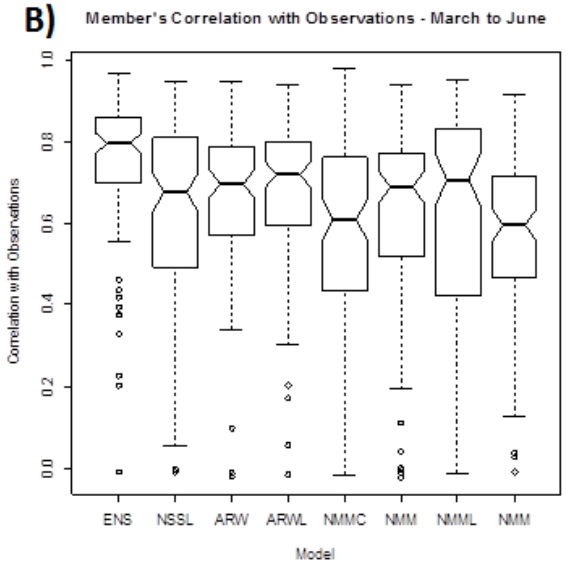


Figure 5 (a) Box Plot of member's maximum values from March to June (b) Box Plot of member's correlation with observations

"underforecasters". These models are exhibiting their influence against the "overforecasters" seen in the ensemble that has a maximum value median around 0.25. The underforecasters are working to dampen down the severity of the day in which the overforecasters predict, and bring the maximum probabilistic value to a lower value.

Figure 5b is the box and whiskers plot of the correlation with the observations. The ensemble's median is at an impressive 0.8. While all the member's correlation median is below the ensemble, it shows that each member holds an important influence in creating the ensemble forecast. The individual member's forecasts alone could not stand as a very useful, but taking these 7 models together in an ensemble mean shows predictive skill in the March through June months, where one would like to see the highest skill.



c. August Analysis

While the ensemble mean show particularly good skill in March through June, it must be noted that this skill is not consistent throughout the year. August sees a little over of 10% of the days at the threshold. The POD values are significantly lower than March to June at, and the FAR values are much high as well. Biases among the individual members become even more varied with the NMMC with a having the highest bias at about 4. The correlation with observations is at some of their lowest values for the year in August. Looking at Figure 4a-c and e, forecasts for August are being made, but they are covering the wrong

areas. The poor performance from the members leads to an ensemble mean lacking skill in this month. Thus, using the forecast would not provide much benefit to forecasters.

d. Case Studies

A further look into the aerial and maximum value compensation affect by members in March to June was done by using case studies. An outbreak day (April 14, 2012) and a lower end severe weather day (April 28, 2012) are discussed in the following sections.

1. April 14, 2012

On April 14, 2012, over 461 storm reports were collected from Nebraska to the Texas Panhandle. For these case studies, this day is considered an outbreak day. Figure 6a shows the probabilistic forecasts for the ensemble (upper left), NSSL (lower left), ARW (upper right), and NMM (lower right) plotted with the observations plotted as solid grey lines overlaid on each plot. The ARW forecasts 3 areas of maximum value around 30% in central Oklahoma, the northern Missouri and Kansas border, and Kansas. These 3 areas are displaced from the single observed area. The NMM's maximum value forecast is higher with a bullseye at about 50% in southeastern Kansas, which is only slightly displaced from what was observed. The NSSL exhibits a maximum value bullseye in the similar area as in NMM, but is elongated into eastern Oklahoma. The NSSL also has a maximum value area that is similar to the

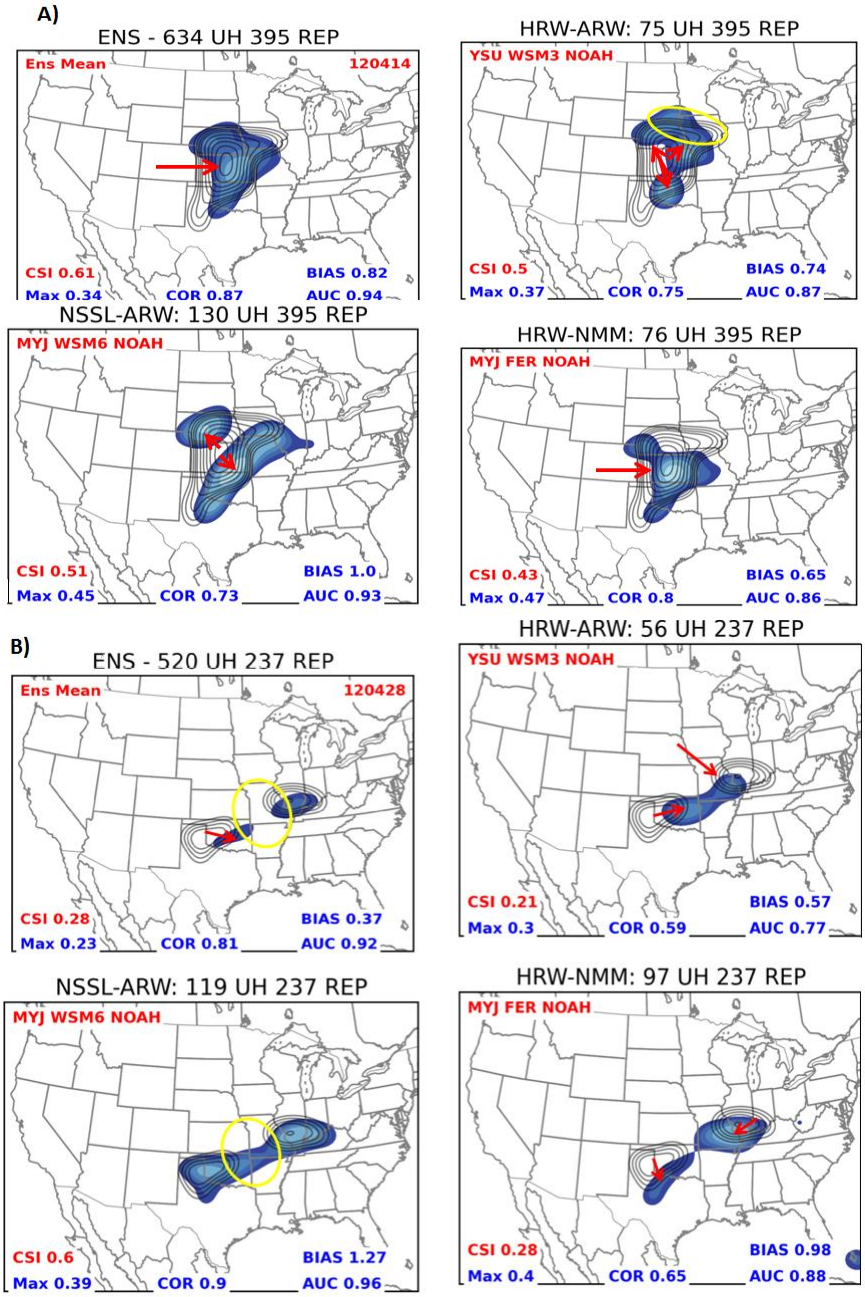


Figure 6. Probabilistic forecast outlooks with maximum value shading (blue) and observed probabilistic forecast (gray lines) for the SSEO (upper left), ARW (upper right), NSSL (lower left), NMM (lower right) for April 14, 2012 (a) and April 28, 2012 (b)

ARW in western Nebraska. The ensemble picks up on the members' bullseyes, producing a maximum value bullseye that is only slightly displaced from the observations in Kansas. The compensation affect is also seen in the forecast coverage area. The ARW and NSSL cover area into Iowa, in which the NMM lacks. This area is then covered in the ensemble and is verified in the observations. The individual members lack characteristics to stand alone as a good forecast for the day, but coming together to form a good ensemble forecast. Members compensate where other members are lacking in maximum value and spatial coverage to output the ensemble forecast with a correlation value with the observation of 0.87.

2. April 28, 2012

April 28, 2012 was a lower end severe weather day with half the reports seen on April 14, 2012. Figure 6b is the same as 6a, but is for April 28, 2012. The ARW, NSSL and NMM all forecast two maximum values areas of severe weather but of different magnitudes and placement. The NMM's is in southern Illinois highest area of maximum value is in southern Illinois with a lesser area extending down into central Texas. The southern maximum area is not covered while the Illinois maximum is only slightly displaced. The ARW forecasts a very small area in southern Illinois and a more potent area in the eastern half of Oklahoma and covers portions of the observed. The NSSL forecasts proves to be extremely accurate hitting the two maximum value areas, but does overforecast by expanding the area covered having these two areas connected. The ensemble forecasts two separate maximum value areas, but these areas are displaced. The NSSL's near perfect is brought down significantly when taking an ensemble mean with the other poor performing members. Even on a low end severe weather day, though, the ensemble proves to have some skill with a correlation value of 0.81.

4. DISCUSSION

While a compensating affect between the members is expected in an ensemble mean forecast, it is important to understand how and in what ways this compensating affect is occurring. The NMMC, NSSL, NMM, and NMML members of the SSEO show characteristics of overforecasting (High POD, High FAR, Bias > 1). The ARW, ARWL, and NMMB member of the SSEO show characteristics of underforecasting (Low POD, Low FAR, and Bias < 1). These models compensate for each other to produce a skillful ensemble forecast with high POD, low FAR, the highest CSI (visually derived from Figure 4d), and the high correlations with observations. A specific look into spring time months (March – June) showed that the SSEO showed good skill over the span on many events.

However, Figures 4a-c and e, show there are other time frames that the ensemble mean does not show such strong of a skill. The month of August has around 20 days that are at the verifying threshold. Yet, this month has low POD, high FAR and poorer correlations with the observations. These facts are important to note when using the SSEO as a forecasting tool. More trust can be put into the forecast during the spring, but not necessarily in the summer.

This research only touches the surface on the work to be done. There are many possibilities for future work such as looking more in depth at other months and days in the dataset, looking at influence of the removal of members from the ensemble, looking into the relationship between the members, and the inclusion of data from other years to create a larger sample size.

Such further work will provide insight that will aid forecasters into when to use the SSEO as a key forecasting tool. As stated above, while the SSEO shows relativity good skill throughout the year, but there are times during the year it performs better and sometimes were it performs worse. Understanding the events (discrete supercells versus squall line) and conditions taking place in the different months would provide would be invaluable to forecasters in trusting forecasts from the SSEO, which will in turn can help better severe weather forecasts overall.

5. ACKNOWLEDGMENTS

The author would like to thank Daphne LaDue, Madison Miller, the National Weather Center Real-World Experiences for Undergraduates 2013 Interns, and especially the co authors for their continuous support. This material is based upon work supported by the National Science Foundation under Grant No. AGS-1062932.

6. REFERENCES

Brooks, H.E., cited 2013: The 2x2 Problem. [Available Online at http://www.nssl.noaa.gov/users/brooks/public_html/feda/note/2x2.html]

Clark, A.J., J. Gao, P.T. Marsh, T. Smith, J.S. Kain, J. Correia Jr., M. Xue, and F. Kong: Tornado Pathlength Forecasts from 2010 to 2011 Using Updraft Helicity

Kain, J.S., and Coauthors, 2008: Some Practical Considerations Regarding Horizontal Resolution in the First Generation of Operational Convection-Allowing NWP. *Wea. Forecasting*, **23**, 931-952.

Murphy, A.H., 1996: The Finley Affair: A Signal Event in the History of Forecast Verification. *Wea. Forecasting*. **11**, 3-20.

NOAA HWT, cited 2013: 2013 Spring Experiment: Forecast Verification Metrics. [Available Online at http://hwt.nssl.noaa.gov/Spring_2013/SkillScoresDescriptionSE2013.pdf]

Roebber, P.J., 2009: Visualizing Multiple Measures of Forecast Quality. *Wea. Forecasting*. **24**. 601-608.

Sobash, R.A., J.S. Kain, D.R. Bright, A.R. Dean, M.C. Coniglio, S.J. Weiss: Probabilistic Forecast Guidance for Severe Thunderstorms Based on the Identification of Extreme Phenomena in Convection-Allowing Model Forecasts. *Wea. Forecasting*. **26**. 714-728.