VERIFICATION OF AUTOMATED HAIL FORECASTS FROM THE 2016 HAZARDOUS WEATHER TESTBED SPRING EXPERIMENT

Joseph Nardi[1, 2], David John Gagne[3], Nathan Snook[4], Amy McGovern[3, 5]

[1]National Weather Center Research Experiences for Undergraduates Program, University of Oklahoma, Norman, OK
[2]Carleton College, Northfield, Minnesota
[3]School of Meteorology, University of Oklahoma, Norman, OK
[4]Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, OK
[5]School of Computer Science, University of Oklahoma, Norman, OK

ABSTRACT

Every spring, the Storm Prediction Center (SPC) and the National Severe Storms Laboratory (NSSL) run an experiment to improve the prediction of severe weather called the Hazardous Weather Testbed. One of the major goals of the experiment is to forecast individual hazards, such as hail. These hail forecasts are run on the Center for Analysis and Prediction of Storms (CAPS) mixed physics ensemble. This ensemble is run using the Advanced Research Weather Research and Forecasting (WRF-ARW) numerical weather prediction model with 9 ensemble members and horizontal grid-spacing of 3 km. Automated hail forecasts are run for a 24 hour period using three different methods: HAILCAST, the Thompson Hail Size Method, and the Gagne Machine Learning Method.

To verify the three hail forecasting methods, neighborhood ensemble probabilities are calculated for a 24 hour period for both 25 mm and 50 mm hail. These hail forecasting methods are verified against data from the NSSL Multi-Radar Multi-Sensor (MRMS) radar mosaic using the Maximum Expected Size of Hail (MESH) method. Relative Operating Characteristic (ROC) curves as well as Attribute Diagrams were created along with calculating the ROC Area Under the Curve (ROC AUC) and Brier Skill Score. A case study of May 26, 2016 was performed; on this day a large complex of storms moved over Nebraska, Kansas, Oklahoma, and Texas, producing 204 reports of severe hail, 183 reports of severe wind, and 21 tornado reports.

Overall, the Gagne Machine Learning Method has greater skill, in terms of the Brier Skill Score, than the other two hail forecasting methods. The Gagne Machine Learning Method also exhibits better discrimination for 25 mm hail in terms of the ROC AUC score. Lastly, the Gagne Machine Learning Method consistently performs well across all microphysics schemes because it is calibrated on each microphysics scheme. For the May 26, 2016 case study, the Gagne Machine Learning method exhibited greater capability to predict hail exceeding 25 mm in diameter while producing relatively few false alarms.

1. INTRODUCTION AND MOTIVATION

Hail is a severe weather phenomenon that can cause property damage, injury to humans, as well as damage to crops. It is estimated that each year, hail causes over 1 billion dollars in property loss and also over 1 billion dollars in crop damage (Jewell and Brimelow 2009). In terms of injuries, the hail storm in Fort Worth, Texas at Mayfest on May 5, 1995 resulted in over 100 people with bruises and broken limbs (Edwards and Thompson 1998). Some of these economic losses and injuries could be prevented or mitigated with precise hail forecasts allowing more time for preventative actions such as moving people, cars, or other property to a safe location.

Currently, it is challenging to forecast severe hail because of uncertainties in model forecasts and in the observations. Some of these uncertainties come from the microphysical parameterizations, which are necessary to predict the characteristics of hail producing storms and the surrounding environment (Snook et al. 2016). Microphysical parameterizations indirectly model the effects of cloud and precipitation formation as these processes cannot be explicitly modeled. Also, the prediction of hail is restricted by the quick development and evolution of storms that produce

[1] *Corresponding author address:* Joseph Nardi, Carleton College, 300 North College Street, Northfield, MN, 55057, Email: nardij@carleton.edu

hail (Gagne et. al 2015). However, with the use of ensemble convection-allowing numerical weather prediction models, a range of atmospheric conditions with their uncertainties can be better predicted. These ensemble models can partially resolve storms that are capable of producing hail up to a day in advance (Clark et al. 2012).

The 2016 Hazardous Weather Testbed Spring Experiment took place from May 2, 2016 to June 3, 2016 at the National Weather Center in Norman, Oklahoma. The Hazardous Weather Testbed is conducted by the Storm Prediction Center (SPC) and the National Severe Storms Laboratory (NSSL) to test new ideas and methods to improve forecasting of severe weather events. A major component of this experiment was the forecasting individual hazards, such as hail. Three different hail forecasting models were run: HAILCAST, the Gagne Machine Learning Method, and the Thompson Hail Size Method. The purpose of this paper is to verify the three hail forecasts produced by the 2016 Hazardous Weather Testbed Spring experiment for the development of more accurate operational hail forecasts.

2. DATA AND METHODS

2.1 Data

The three hail forecasting models were run on the Center for Analysis and Prediction of Storms (CAPS) mixed physics ensemble. This ensemble used the Advanced Research Weather Research and Forecasting (WRF-ARW) model. Ten ensemble members with various microphysics schemes were run, however only data from nine of the members are used in this study. Data from member 2 was not used as this member failed to run at all. All of the ensemble members were initialized on weekdays at 00 UTC and used a 3 km grid-spacing on a domain covering the contiguous United States (CONUS). The members had 1680 grid-points east-west and 1152 grid-points north-south. More information about the individual ensemble members can be found in Table 1.

To verify the hail forecasts that were produced, data from the NSSL Multi-Radar Multi-Sensor (MRMS) radar mosaic was used. Radar derived maximum expected size of hail (MESH) served as the observed hail for verification purposes (Witt et al. 1998). For the main part of the study, the SPC reports of hail were not used because reports are concentrated near populated areas, which limits their coverage (Cintineo et al. 2012). Also, since hail diameters are often reported using comparisons to common circular or spherical objects, there are unnatural peaks in the distribution of the hail size (Jewell and Brimelow 2009). SPC hail reports were used for the case study as reports for this day reflected the general coverage of severe hail well.

2.2 Hail Forecasting Methods

The three hail forecasting methods that were used are HAILCAST, the Gagne Machine Learning Method, and the Thompson Hail Size Method. HAILCAST is a one-dimensional physics based cloud and hail model to predict the maximum expected hail size at the surface (Brimelow 2002). In HAILCAST, hail embryos are grown based on the atmospheric conditions such as instability, shear, and moisture. HAILCAST was implemented and tested by the SPC and showed to have considerable skill in forecasting hail size (Jewell and Brimelow 2009).

The Gagne Machine Learning Method uses machine learning, specifically random forest to fit a model between atmospheric variables and observed hail size (Gagne 2016). Random forests are ensembles of decision trees in which each

(1) CAPS: mixed phys + radar						
Members	**IC**	**BC**	**Microphysics**	**LSM**	**PBL**	**Model**
core01	NAMa+3DVAR	NAMf	Thompson	NOAH	MYJ	arw
core02	RAPa+3DVAR	GFSf	Thompson	RUC	MYNN	arw
core03	core01+arw-p1_pert	arw-p1	P3	NOAH	YSU	arw
core04	core01+arw-n1_pert	arw-n1	MY	NOAH	MYNN	arw
core05	core01+arw-p2_pert	arw-p2	Morrison	NOAH	MYJ	arw
core06	core01+arw-n2_pert	arw-n2	P3	NOAH	YSU	arw
core07	core01+nmmb-p1_pert	nmmb-p1	MY	NOAH	MYNN	arw
core08	core01+nmmb-n1_pert	nmmb_n1	Morrison	NOAH	YSU	arw
core09	core01+nmmb-p2_pert	nmmb-p2	P3	NOAH	MYJ	arw
core10	core01+nmmb-n2_pert	nmmb-n2	Thompson	NOAH	MYNN	arw

Table 1: Specifications for the CAPS Mixed Physics Ensemble (NSSL)

decision tree is randomized by bootstrap resampling of the training data and random feature subset selection to increase its independence from the other trees. (Breiman 2001). The machine learning method first decides if a storm will produce hail, and if it does, a separate model is trained to predict the size distribution of hail it will produce (Gagne 2016). The Gagne Machine learning method is calibrated using members from the 2015 CAPS ensemble that share the same microphysics scheme.

The Thompson Hail Size Method calculates the maximum hail size directly from graupel and the hail size distribution produced by the microphysics scheme at the lowest model level. The Thompson Hail Size Method is different than the Thompson microphysics scheme as the method is run on all of the ensemble members, of which there are four different microphysics schemes. The Thompson Hail Size Method works by identifying the largest hail or graupel diameter that exceeds a specified number concentration threshold (Thompson 2003).

2.3 Methods

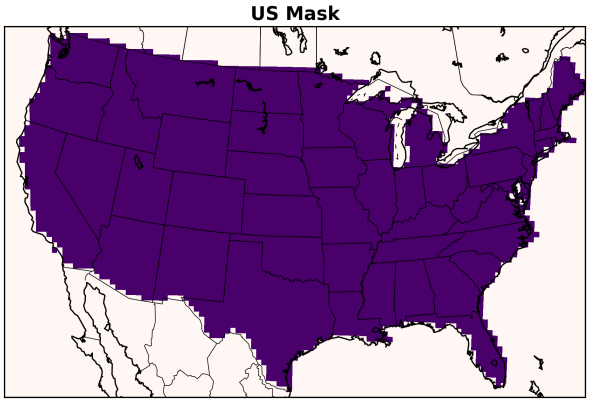


Figure 1. Areas within the US Mask used for verification

To verify the hail forecasts, a method of neighborhood ensemble probabilities was used. This is to account for the spatial errors in formation and evolution of storms (Schwartz et al. 2010). For this study, a US mask was applied to the data to only include points within the contiguous United States, excluding areas over Mexico, Canada, and bodies of water where observation coverage is poor or non-existent. The mask region is plotted in Figure 1.

A coarse grid was used in the verification of the hail forecasts. This was done by plotting a grid point for an ensemble member if hail of a certain threshold occurred within 42 km of that point. This number was chosen as it is close to the spatial verification requirements outlined by the NWS, which is 40 km (Sobash 2016). Threshold values of 25 mm and 50 mm hail, or 1 inch and 2 inch hail respectively were used as these are the thresholds of severe hail and significant severe hail by the NWS (Melick et. al 2014). Data plotted on the coarse grid can be found in Figure 2a. This same procedure was also performed for the NSSL MRMS data with the MESH algorithm.

This procedure was applied for each ensemble member of the hail forecast. Then the probability from all the ensemble members were averaged to create an ensemble mean. The ensemble mean data is plotted in Figure 2b. A Gaussian filter was applied to smooth out the probabilities so that they more closely match human forecasts (Gagne et al. 2015). A standard deviation of 42 km, or 1 coarse grid point, was used for the Gaussian kernel. The Gaussian kernel performs smoothing up to 4 standard deviations away from the center point. A picture of the neighborhood ensemble probabilities plotted on a map can be found in Figure 2c.

Contingency Table

		Observed		
		yes	no	Total
Forecast	yes	*hits*	*false alarms*	*forecast yes*
	no	*misses*	*correct negatives*	*forecast no*
Total		*observed yes*	*observed no*	*total*

Figure 3. A contingency table, Source: CAWCR

2.4 Verification Methods

Two methods were used in the verification of the hail forecasts, Relative Operating Characteristic (ROC) Curves and Attributes Diagrams. A ROC Curve (Mason 1982) is a measure of the ability of the forecast to discriminate between an event happening or not. At a specific threshold of a probability value, a contingency table can be constructed. A contingency table shows the frequencies of “yes” and “no” for the forecast and observations (CAWCR). An example of a contingency table can be found in Figure 3.

With a contingency table, different verification statistics can be calculated. The ROC curve plots probability of detection (POD) verses

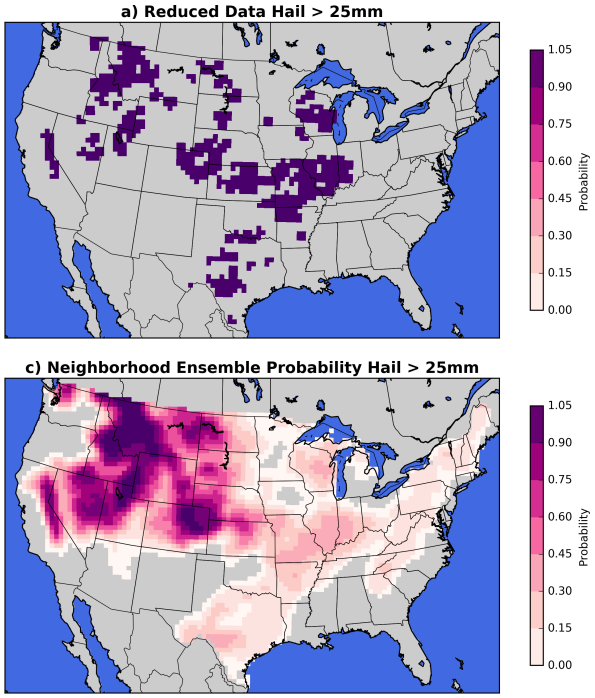


Figure 2. a) This shows the data reduced to a 42 km grid b) Shows the ensemble mean c) Shows the data after the Gaussian filter was applied

false alarm rate (POFD). POD is the fraction of observed "yes" events that were correctly forecasted (Wilks 2011). POFD is the fraction of observed "no" events that were wrongly forecasted as "yes" (Wilks 2011). The equations for POD and POFD are below in equations 1 and 2 respectively.

POD = hits / (hits + misses). (1)

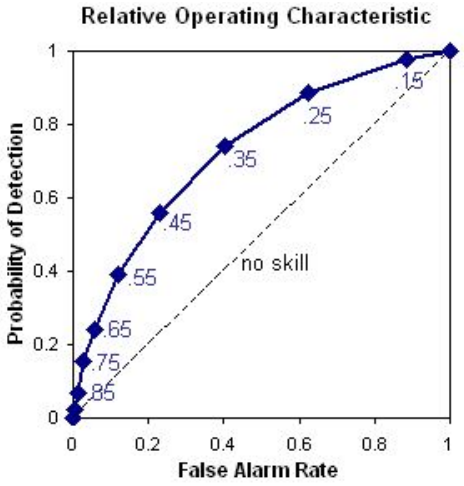
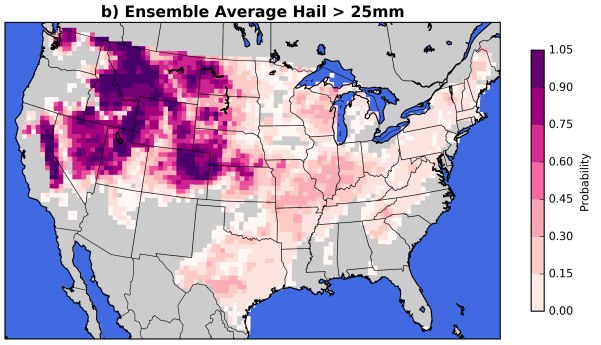


Figure 4. Example of a ROC Curve, Source: CAWCR



POFD = false alarms / (correct negatives + false alarms). (2)

A ROC Curve graphs POD and POFD for a series of threshold values. In this case, the threshold values used were .01 and also 0 to 1 with a step of .05. An example of a ROC Curve can be found in Figure 4. A "perfect" ROC Curve would be a vertical line as the forecast would a POD of 1 and a POFD of 0, meaning that the forecast can perfectly discriminate events from non-events. The diagonal line on the ROC curve represents no skill as the POD equals the POFD. The area under the ROC Curve (ROC AUC) can be calculated to quantify resolution. A perfect ROC AUC score would be 1 and a ROC AUC score with no skill would be 0.5.

An attributes diagram shows the reliability of a forecast by graphing the forecast probability and observed relative frequency for a set of probability bins (Hsu and Murphy 1986). An attributes diagram examines how well the predicted probabilities of an event correspond to the frequencies that were observed (CAWCR). An example of an attributes diagram can be found in Figure 5. With an attributes diagram, a "perfect" forecast would be on the 1 to 1 line as the forecast probabilities equaled the observed relative frequency. If the curve is above the perfect reliability line, this means that the forecast underpredicts hail. On the other hand, if the curve is below the perfect reliability line, the forecast overpredicts hail. The horizontal line represents the climatology probability, so a curve along this line would have no resolution. Half way in between the no resolution line and the perfect reliability line is the no skill line. Points above the no skill line contribute positively to the Brier Skill Score. The spread of the forecast distribution across bins shows the sharpness of the forecasts (CAWCR).

3. RESULTS

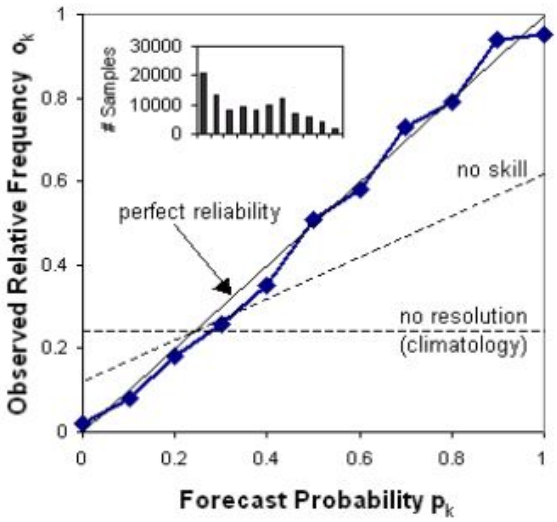


Figure 5. Example of an Attributes Diagram, Source: CAWCR

3.1 Overall Verification

ROC curves were created for all of the forecasts for both 25 mm and 50 mm hail. The ROC curve for 25 mm hail can be found in Figure 6a and the ROC curve for 50 mm hail can be found in Figure 6b. In terms of the 25 mm or greater hail forecasts, the Gagne Machine Learning method had the highest ROC AUC score with 0.867. This indicates that the Gagne Machine Learning Method performs the best in discriminating storms that produce hail greater than 25 mm and storms that do not. HAILCAST had a ROC AUC score close to that of the Gagne Machine Learning Method with that of 0.856. The Thompson Hail Size Method performed the worst with a ROC AUC score of 0.734. It is important to note that all three hail forecasting models had a ROC AUC score above 0.7, which is considered to be operationally useful (Strensrud and Yussouf 2007).

In terms of the 50 mm or greater hail forecast, the Thompson Hail Size Method scored the highest ROC AUC score with 0.828. This indicates that the Thompson Hail Size Method was the best in discriminating storms that would produce hail greater than 50 mm due to the model having higher PODs. However, the Thompson Hail Size Method had higher POFDs when compared to other models when forecasting 50 mm hail. The Gagne Machine Learning Method had a ROC AUC score reasonably close to that of the Thompson Hail Size Method with a ROC AUC score of 0.779. The worst performing method was HAILCAST with a ROC AUC score of 0.704. Again, all of the hail forecasting methods had ROC AUC scores above 0.7, suggesting that they have operationally useful skill for forecasting 50 mm hail or greater. However, the hail forecasting methods had worse ROC AUC scores with 50 mm hail or larger than 25 mm hail or larger.

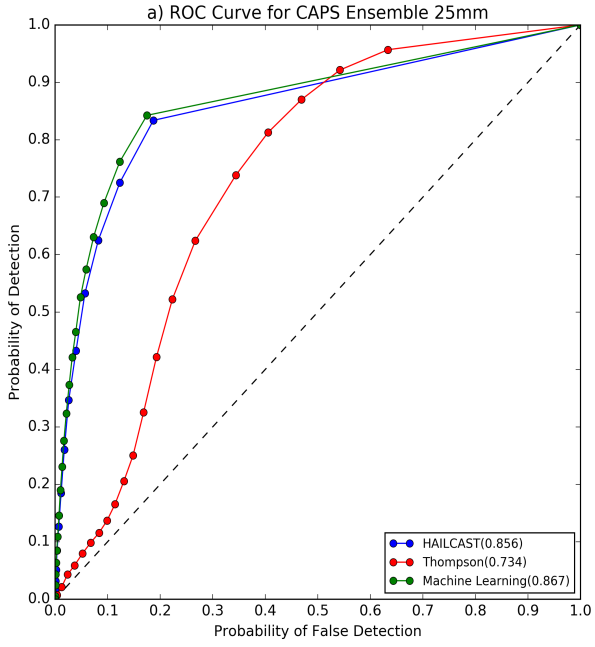
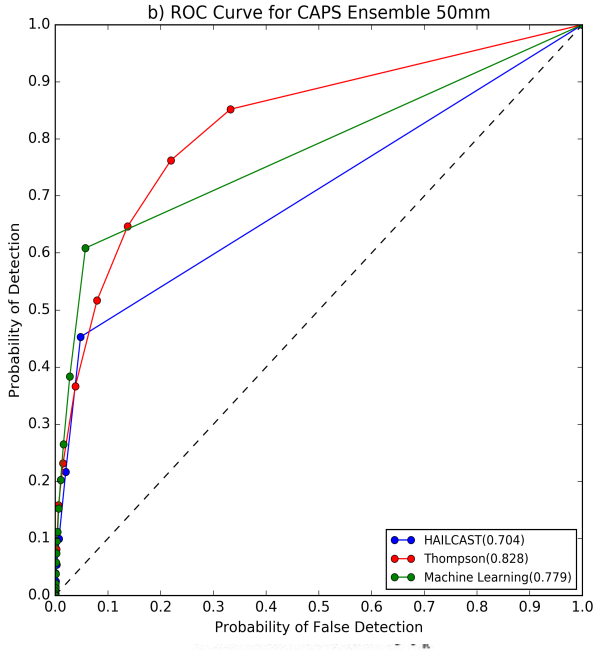


Figure 6. a) ROC Curve for CAPS Ensemble 25 mm



b) ROC Curve for CAPS Ensemble 50 mm

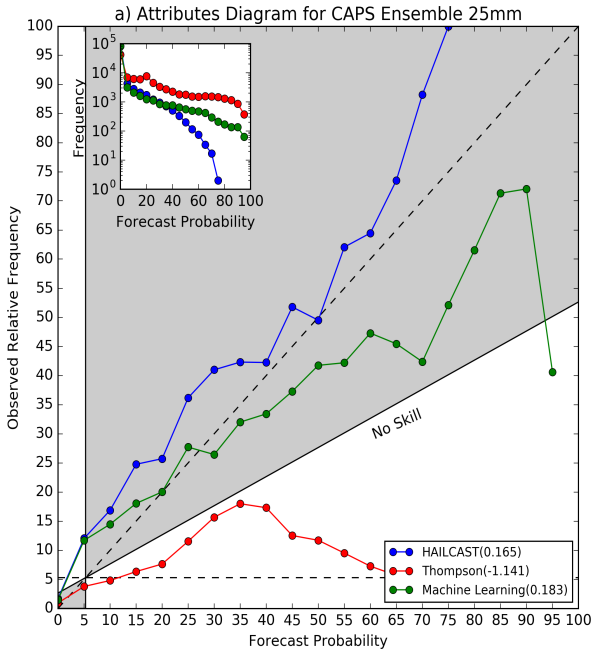
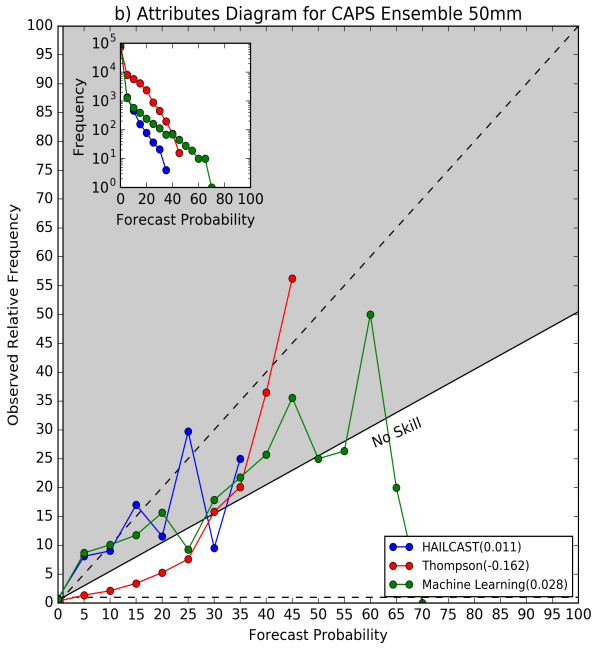


Figure 7. a) Attributes Diagram for CAPS Ensemble 25 mm



b) Attributes Diagram for CAPS Ensemble 50 mm

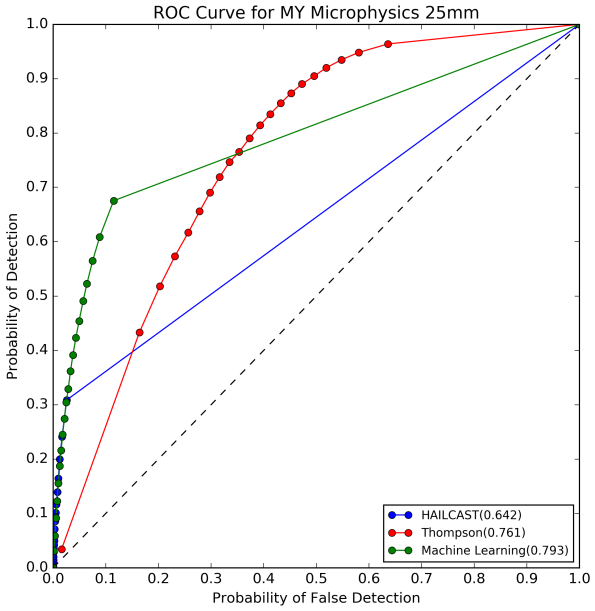
Attributes diagrams were also created for all of the forecast for both 25 mm and 50 mm hail. The attributes diagram for 25 mm hail can be found in Figure 7a and the attributes diagram for 50 mm hail can be found in Figure 7b. For 25 mm hail, the Gagne Machine Learning Method had the highest Brier Skill Score with 0.183. HAILCAST was not far behind with a Brier Skill score of 0.165. However, HAILCAST tended to slightly underpredict hail as the curve is above the dashed line and the Gagne Machine Learning Method tended to slightly overpredict hail as the curve is below the dashed line. The Thompson Hail Size Method consistently overpredicted hail and thus had a negative Brier Skill Score of -1.141.

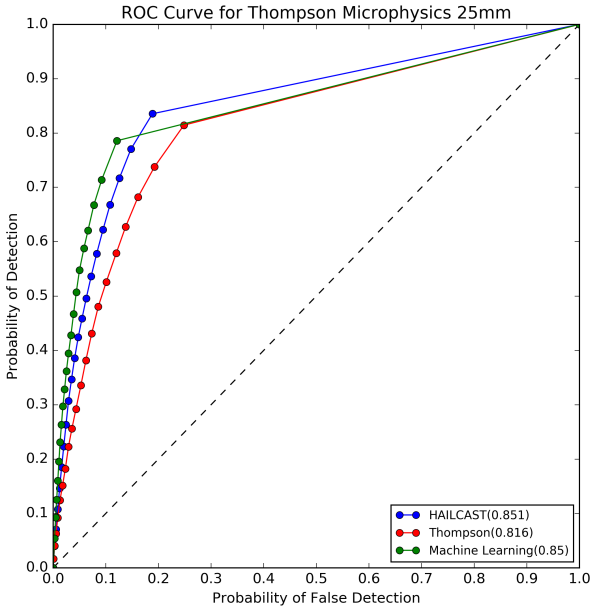
When looking at the 50 mm or greater forecast, the Gagne Machine Learning Method has the highest Brier Skill Score of 0.028. Again, HAILCAST was close to the Gagne Machine Learning Method with a Brier Skill Score of 0.011. Just as with the 25 mm forecasts, the Thompson Hail Size Method has a negative Brier Skill Score, that of -0.162. Similar to the ROC AUC scores, the 25 mm hail forecasts had a higher Brier Skill Score than the 50 mm hail forecasts for the Gagne Machine Learning Method and HAILCAST. For 25 mm hail forecasts the Gagne Machine Learning Method and HAILCAST showed considerable skill with the Brier Skill Score. However, with the 50 mm hail forecasts, all three models overpredicted hail; this can be seen in Brier Skill Scores, which are around zero or negative, implying little to no skill.

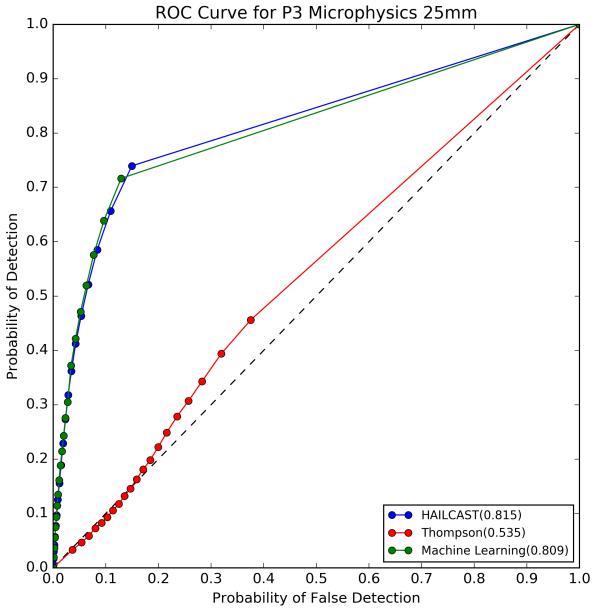
3.2 Verification with Microphysics Schemes

To evaluate the different hail forecast models with the four microphysics schemes, ROC Curves and attributes diagrams were created for the different microphysics schemes for 25 mm hail for greater. ROC curves for each microphysics scheme for 25 mm hail can be found in Figure 8. Attribute diagrams for 25 mm hail can be found in Figure 9.

At the 25 mm threshold, the Gagne Machine Learning Method has a similar ROC AUC score over all four of the microphysics schemes, all around 0.8. The Gagne Machine Learning Method is more consistent over all the microphysics schemes as it is calibrated to each scheme. HAILCAST scores the highest ROC AUC scores on the Thompson and P3 microphysics schemes with ROC AUC scores of above 0.8. HAILCAST performed almost the same in terms of ROC AUC scores on both the Thompson and P3 microphysics. However, HAILCAST also scored low ROC AUC scores on the MY and Morrison microphysics with ROC AUC scores around 0.6, meaning that these forecasts have little to no skill.







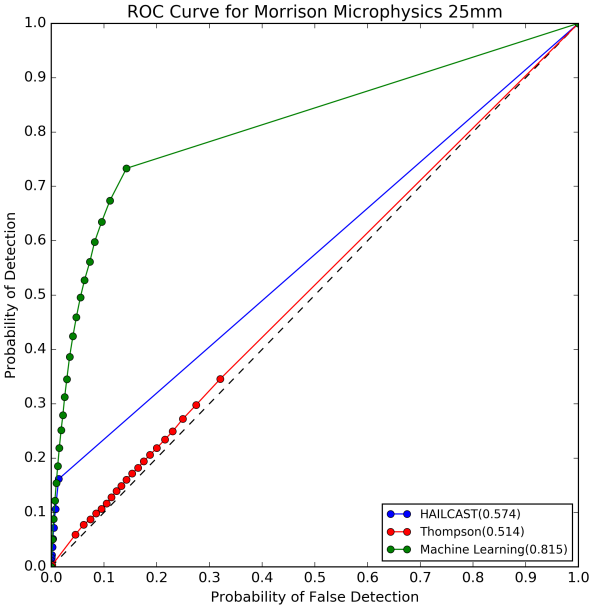
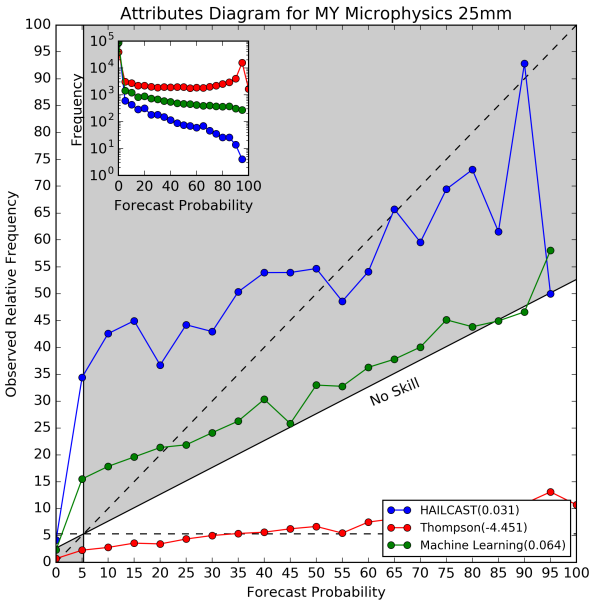
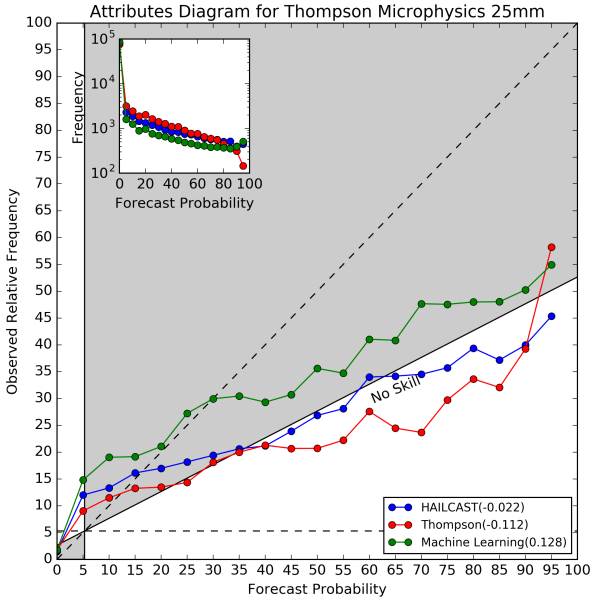


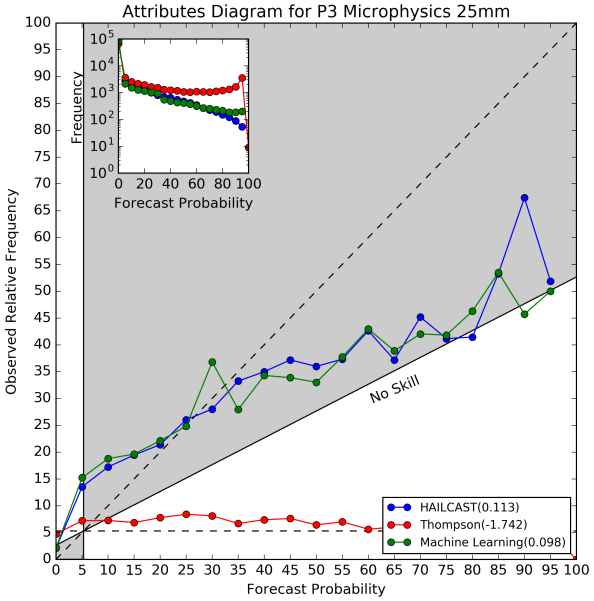
Figure 8. ROC Curves at 25 mm Hail for a) MY Microphysics b) Thompson Microphysics c) P3 Microphysics d) Morrison Microphysics

The Thompson Hail Size Method had ROC AUC scores of around 0.5 with both the P3 and Morrison microphysics, which means that these forecasts had no skill. All of the hail forecast models had ROC AUC scores of above 0.85 on the Thompson microphysics scheme.

As mentioned above, attributes diagrams were also created for 25 mm hail or larger for the four different microphysics schemes. Once again, the Gagne Machine Learning Method is consistent over all of the different microphysics schemes, as explained above. The Gagne Machine Learning Method has the highest skill on the Thompson and P3 microphysics, with Brier Skill Scores of around 0.1. HAILCAST has the highest skill on the P3 microphysics scheme, outperforming the Gagne Machine Learning Method in terms of the Brier Skill Score. The Thompson Hail Size Method has a negative Brier Skill Score for all of the different microphysics schemes showing no skill. This is due







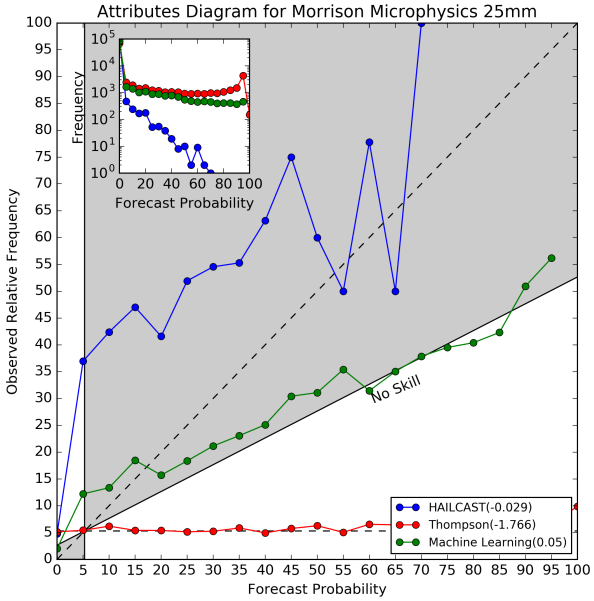


Figure 9. Attributes Diagram at 25 mm Hail for a) MY Microphysics b) Thompson Microphysics c) P3 Microphysics d) Morrison Microphysics

to the fact that the Thompson Hail Size Method overpredicts the extent of hail. Almost all of the hail forecast methods over-predict hail, except for HAILCAST on the MY and Morrison microphysics schemes. All of the hail forecast models had similar skill for forecasts using the Thompson microphysics scheme, as it was the case with the ROC AUC scores. When splitting up the hail forecast models by the different microphysics schemes, many of the hail forecast models had little to no skill in forecasting hail 25 mm or larger.

3.3 Case Study

A case study of May 26, 2016 was performed to evaluate the hail forecast models. This day was chosen as it was the worst hail even during the experiment. On this day a large complex of storms moved over Nebraska, Kansas, Oklahoma, and Texas, producing 204 reports of

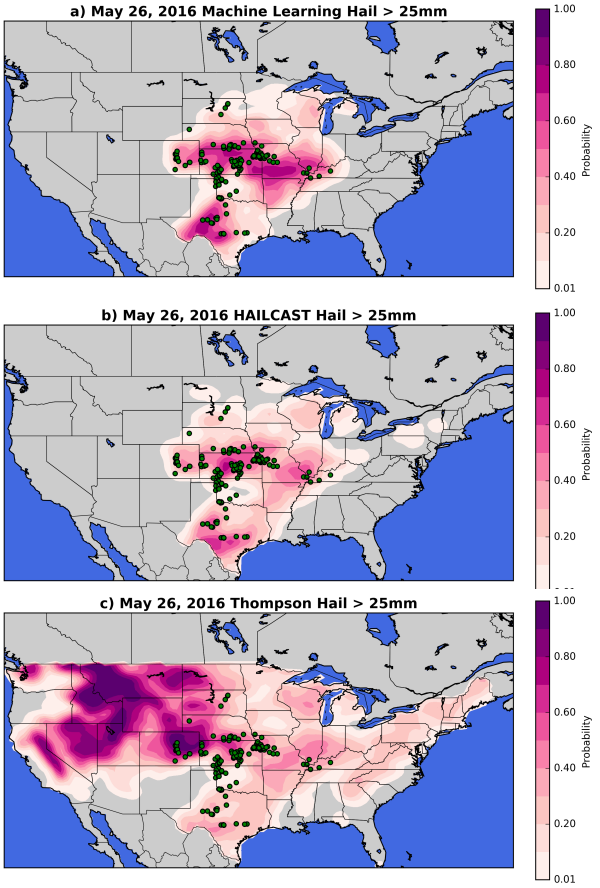


Figure 10. Case Study of 25 mm Hail and Reports a) Gagne Machine Learning Method b) HAILCAST c) Thompson Hail Size Method

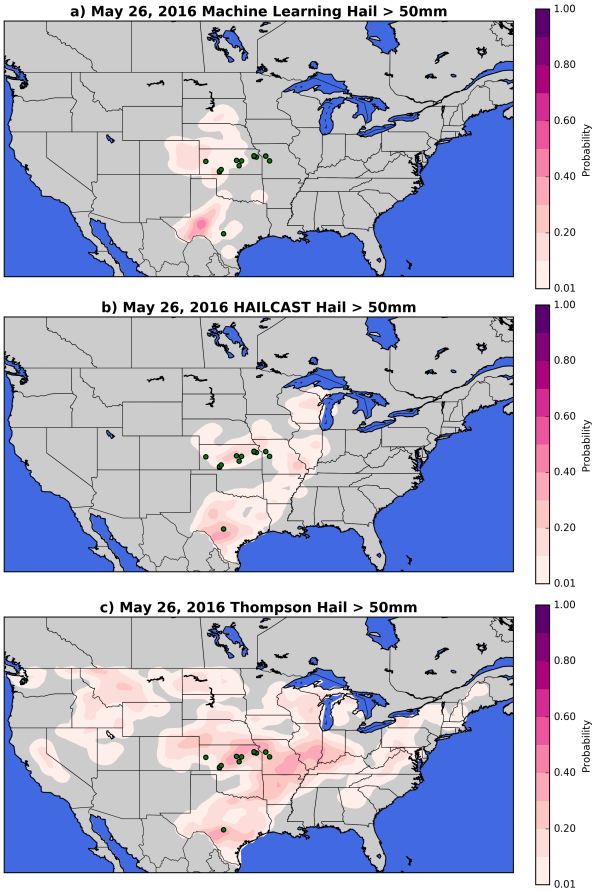


Figure 11. Case Study of 50 mm Hail and Reports a) Gagne Machine Learning b) HAILCAST c) Thompson Hail Size Method

severe hail, 183 reports of severe wind, and 21 tornado reports. In terms of the synoptic environment, there was strong instability with CAPE values over 4000 to 5000 J/kg. Shear ranged from 30 knots in Oklahoma to over 50 knots in Kansas. Maps with the probabilities hail exceeding 25 mm overlaid with SPC storm reports of hail greater than 25 mm for each hail forecast model can be found in Figure 10. Maps for the probabilities of hail greater than 50 mm with the SPC storm reports for each of the hail forecasts can also be found in Figure 11.

For the 25 mm hail forecasts, the Thompson Hail Size Method has high probabilities of hail all over the western part of the United States. However, the Thompson Hail Size Method misses or has very low probabilities of the hail that occurred in Oklahoma. This means that while the Thompson Hail Size Method correctly predicted most of the hail reports, there is a large false alarm rate with all of the hail forecasted in the western United States. The Gagne Machine Learning Method and HAILCAST have similar areas where hail is forecasted, however, the probabilities are a bit different. Both models predict hail where almost all the SPC storm reports occurred, however the Gagne Machine Learning Method has much higher probabilities of hail than HAILCAST. Also, the Gagne Machine Learning Method appears to have a slight bias to the south as many of the SPC storm reports occur north of the area of highest probabilities.

In terms of the 50 mm size hail or larger threshold, the Gagne Machine Learning Method has very low probabilities where reports of 50 mm hail occurred showing that this method is not as sharp as other methods. Also, the Gagne Machine Learning Method has higher probabilities in Texas where there are no SPC storm reports. HAILCAST has the correct area of hail greater than 50 mm, however, forecasts low probabilities. The Thompson Hail Size Method correctly predicts all of

the SPC storm reports that are greater than 50 mm. However, the Thompson Hail Size Method predicts a large area of which hail larger than 50 mm could occur, so there would be many false alarms associated with this forecast.

Overall, HAILCAST had the correct areas associated with the SPC storm reports but predicted low probabilities in some cases. The Gagne Machine Learning Method performed well at the 25 mm threshold but missed the main placement of the 50 mm hail.

4. SUMMARY AND CONCLUSION

Hail causes a lot of economic losses in property damage, crop loss, and injury each year. Automated hail forecasts from the 2016 Hazardous Weather Testbed Spring Experiment were verified in order to develop more accurate operational hail forecasts in the future.

Overall, the Gagne Machine Learning Method has greater skill, shown by the Brier Skill Score, than the other two hail forecasting methods. The Gagne Machine Learning Method also had better discrimination for 25 mm hail in terms of the ROC AUC score. HAILCAST performed nearly as well as the Gagne Machine Learning Method, however the Gagne Machine Learning had slightly higher ROC AUC and Brier Skill Scores. Lastly, Gagne Machine Learning Method performs better across all of the different microphysics schemes because it is calibrated on each microphysics scheme. In terms of the case study, the Gagne Machine Learning method captured more storm reports and had higher probabilities where hail exceeding 25 mm in diameter occurred without having a lot of false alarms. HAILCAST performed better at forecasting hail 50 mm or greater in the case study.

In conclusion, the Gagne Machine Learning Method shows advantages in predicting hail over other currently used hail forecasting methods. The Gagne Machine Method can be used in an operational setting to predict hail up to a day in advance. Further development of machine learning models and numerical weather prediction should lead to more accurate hail forecasts.

5. ACKNOWLEDGMENTS

The CAPS ensemble was generated on the Texas Advanced Computing Center (TACC) Stampede Supercomputer.

This work was prepared by the authors with funding provided by National Science Foundation Grant No. AGS-1560419 and AGS- 1261776, and NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement #NA11OAR4320072, U.S. Department of Commerce. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, NOAA, or the U.S. Department of Commerce.

6. REFERENCES

Breiman L. 2001: Random forests. *Mach. Learn.* **45**, 5-32.

Brimelow J.C., Reuter G.W., and Poolman E.R., 2002: Modeling Maximum Hail Size in Alberta Thunderstorms. *Wea. Forecasting* **17**, 1048–1062, doi: 10.1175/1520-0434(2002)017<1048:MMHSIA>2.0.CO;2.

CAWCR, 2009: Forecast Verifiction – Issues, Methods and FAQ. Accessed 2 August 2016. [Available online at http://www.cawcr.gov.au/projects/verification/verif_web_page.html]

Cintineo J.L., Smith T.M., Lakshmanan V., Brooks H.E., and Ortega K.L., 2012: An Objective High-Resolution Hail Climatology of the Contiguous United States. *Wea. Forecasting* **27**, 1235–1248, doi: 10.1175/WAF-D-11-00151.1.

Clark J., and Coauthors: An Overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Amer. Meteor. Soc.* **93**, 55–74, doi: 10.1175/BAMS-D-11-00040.1.

Edwards R., and Thompson R., 1998: Nationwide Comparisons of Hail Size with WSR-88D Vertically Integrated Liquid Water and Derived Thermodynamic Sounding Data. *Wea. Forecasting* **13**, 277-285, doi:10.1175/15200434(1998)013<0277:NCOHSW>2.0.CO;2.

Gagne D.J., McGovern A., Brotzge J., Coniglio M., Correria J., and Xue M., 2015: Day-Ahead Prediction Integrating Machine Learning with

Storm-Scale Numerical Weather Models. *Twenty-Seventh IAAI Conference*, Austin, TX, Association for the Advancement of Artificial Intelligence [Available online at http://www.aaai.org/ocs/index.php/IAAI/IAAI15/paper/view/9724/9898]

Gagne, D. J., 2016: Coupling Data Science Techniques and Numerical Weather Prediction Models for High-Impact Weather Prediction. Ph.D. dissertation, University of Oklahoma, 156 pp

Hsu, W.-R. and A. H. Murphy, 1986: The Attributes Diagram: A Geometrical Framework for Assessing the Quality of Probability Forecasts. International Journal of Forecasting, 2, 285–293. doi:10.1016/0169-2070(86)90048-8.

Jewell R., and Brimelow J., 2009: Evaluation of Alberta Hail Growth Model Using Severe Hail Proximity Soundings from the United States. *Wea. Forecasting* **24**,1592–1609, doi: 10.1175/2009WAF2222230.1.

Mason, I., 1982: A model for assessment of weather forecasts. Aust. Meteor. Mag., **30**, 291–303.

Melick, C.J., Jirak, I.L., Correia Jr, J., Dean, A.R. and Weiss, S.J., 2014: Exploration of the NSSL Maximum Expected Size of Hail (MESH) Product for Verifying Experimental Hail Forecasts in the 2014 Spring Forecasting Experiment. *Proceedings, 27th Conference on Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., 76. [Available online at https://ams.confex.com/ams/27SLS/webprogram/Paper254292.html].

Snook N., Jung Y., Brotzge J., Putnam B., and Xue M., 2016: Prediction and Ensemble Forecast Verification of Hail in the Supercell Storms of 20 May 2013. *Wea. Forecasting* **31**, 811–825, doi: 10.1175/WAF-D-15-0152.1.

Sobash R. A., Schwartz C. S., Romine G.S., Fossell K.R., and Weisman M.L., 2016: Severe Weather Prediction Using Storm Surrogates from an Ensemble Forecasting System. Wea. Forecasting **31**, 255–271, doi: 10.1175/WAF-D-15-0138.1.

Stensrud D.J., and Yussouf N., 2007: Reliable Probabilistic Quantitative Precipitation Forecasts from a Short-Range Ensemble Forecasting System. *Wea. Forecasting* **22**, 3–17, doi: 10.1175/WAF968.1.

Swartz W. H., Zhu X., Yee J.-H., Talaat E. R., and Coy E., 2010: A Spectral Parameterization of Drag, Eddy Diffusion, and Wave Heating for a Three-Dimensional Flow Induced by Breaking Gravity Waves. J. Atmos. Sci. **67**, 2520–2536, doi: 10.1175/2010JAS3302.1.

Thompson G., and Coauthors, 2003: Improvement of Microphysical Parameterization through Observational Verification Experiment. Bull. Amer. Meteor. Soc. **84**, 1807–1826, doi: 10.1175/BAMS-84-12-1807.

Wilks, D. S., 2011: Statistical Methods in the Atmospheric Sciences, 3rd ed. Academic Press, 676 pp.

Witt A., Eilts M.D., Stumpf G.J., Johnson J.T., Mitchell E. D.W., and Thomas K. W., 1998: An Enhanced Hail Detection Algorithm for the WSR-88D. *Wea. Forecasting* **13**, 286–303, doi: 10.1175/1520-0434(1998)013<0286:AEHDAF>2.0.CO;2.