Evaluation of Real-Time Machine Learning Hail Forecasts from the NCAR Convection-Allowing Ensemble

David John Gagne II[1], Amy McGovern[2], Ryan A. Sobash[1], Sue Ellen Haupt,[1] John K. Williams[3], and Doug Nychka[1]

1. National Center for Atmospheric Research; 2. University of Oklahoma; 3. The Weather Company, An IBM Business

Email: dgagne@ucar.edu
Twitter: @DJGagneDos
Github: github.com/djgagne

Introduction

- **Motivation:** Evaluate different hail forecasting models on the same ensemble system for multiple months
- **Issue:** Previous evaluations focused on May and June and did not always have fixed NWP model configurations
- **Challenge:** Relative performance of each system may vary based on location and intensity threshold
- **Goal:** Evaluate hail forecast performance over a 4 month period and determine the skill of each method for different metrics, intensity thresholds, and regions

Data

NWP System: NCAR Ensemble

- 10 WRF-ARW members, Thompson microphysics
- 3 km grid spacing over Contiguous US
- Evaluation Period: 1 May to 31 August 2016

Observations: NOAA Multi-Radar Multi-Sensor (MRMS) Maximum Expected Size of Hail (MESH)

Methods

Storm Surrogate Probability Forecasts

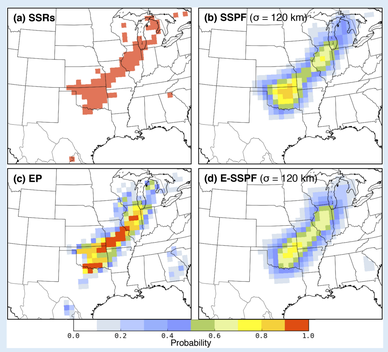


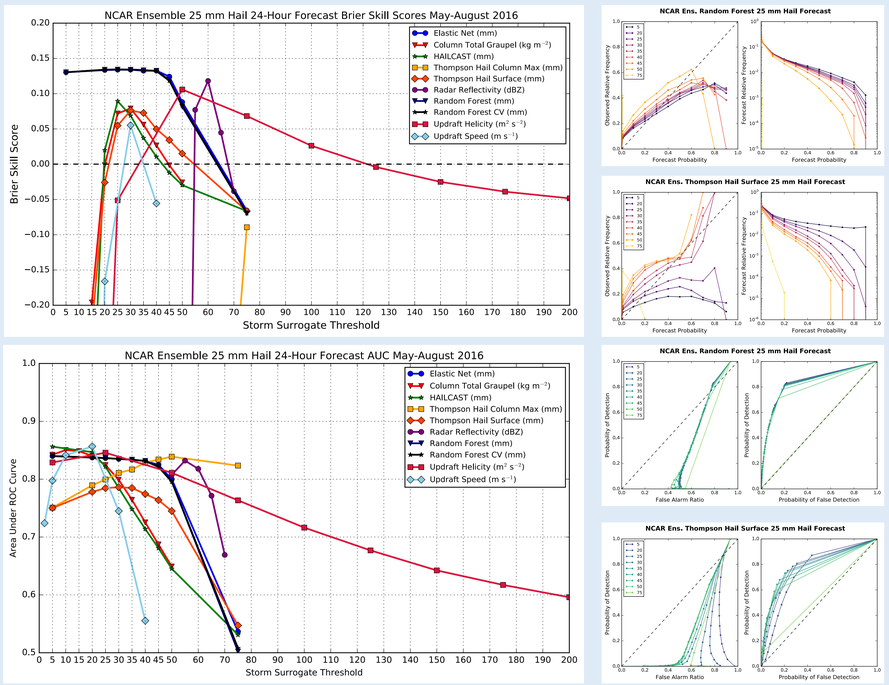
Figure from Sobash et al. (2016).

(a) Storm Surrogate Reports created when variable exceeds threshold within 42 km of each grid point.
(b) Apply (**σ**=84 km) Gaussian smoother to grid from (a) to spread probability across grid to show spatial uncertainty.
(c) Average storm surrogate reports across members to generate ensemble probabilities.
(d) Apply Gaussian smoother to ensemble probabilities.

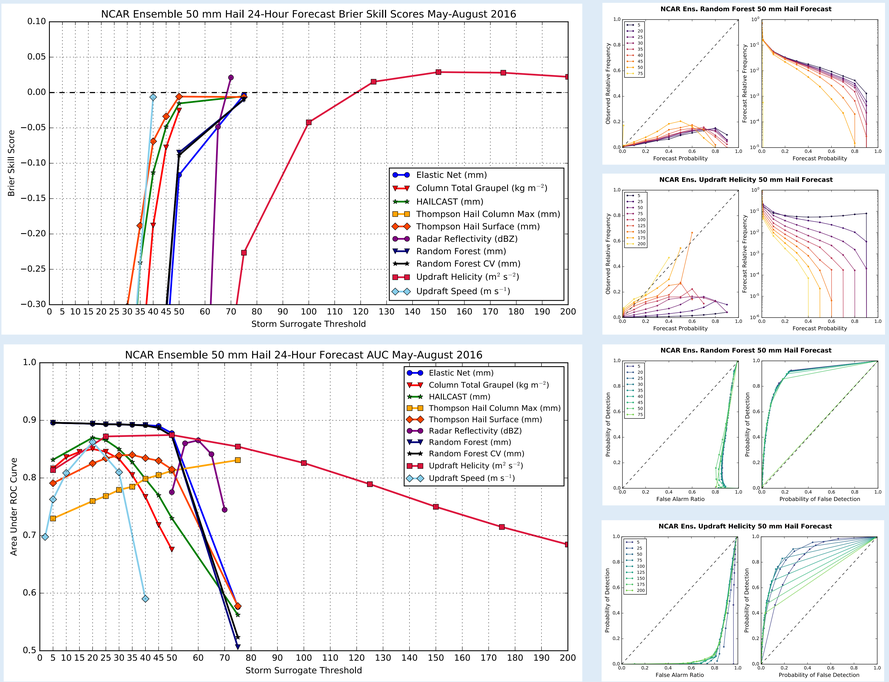
Storm Surrogate Variables

- Random Forest Hail Size (mm)
- Elastic Net Linear Regression Hail Size (mm)
- HAILCAST Hail Size (mm)
- Thompson Microphysics Column Max Hail (mm)
- Thompson Microphysics Surface Max Hail (mm)
- Column Total Graupel (kg m^{-2})
- Updraft Helicity (m^2 s^{-2})
- Updraft Speed (m s^{-1})
- Radar Reflectivity (dBZ)

Intensity Threshold Sensitivity



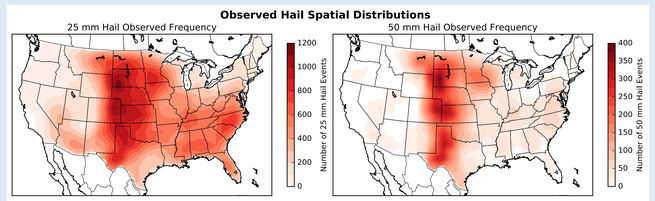
Comparison of Brier Skill Scores and Area Under the ROC Curve for storm surrogate probability forecasts predicting **25 mm hail**. Side panels show how the reliability diagram, ROC curve, and performance curves vary for Random Forest and Thompson Hail forecasts. Random Forest is less sensitive to changes in threshold than others.



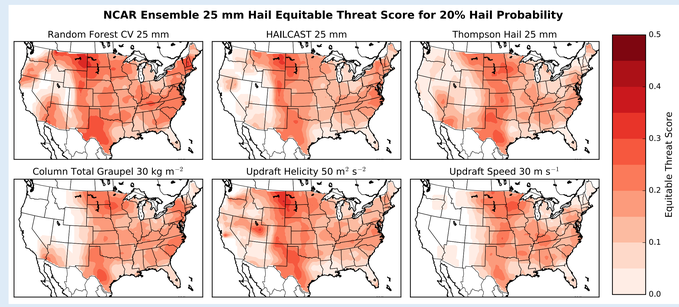
Comparison of Brier Skill Scores and Area Under the ROC Curve for storm surrogate probability forecasts predicting **50 mm hail**. All methods tend to be overconfident at most intensity thresholds but do detect most 50 mm hail events. Those detections come at the expense of a high false alarm ratio.

Spatial Trends

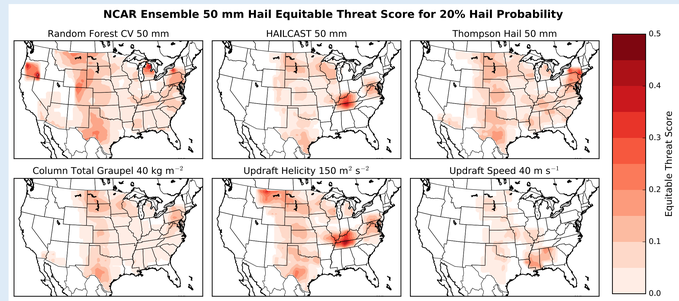
We evaluate the spatial variability of performance by aggregating binary contingency tables over a 336 km wide box and calculating the equitable threat score within that area for every grid point.



Most hail occurs in the High Plains region, especially 50 mm or larger hail. There is a secondary maximum of smaller hail in the Piedmont region of the Southeast.



Random Forest performs better than other methods in the High Plains. In the Southeast, most methods do well in the Carolinas, but Updraft Speed performs best in Alabama.



Similar regional differences in performance are seen at 50 mm. Some of the more extreme variance in performance occurs in areas with few events.

Conclusions

- Most storm surrogate variables are sensitive to the choice of intensity threshold.
- Random forests are less sensitive because most of skill comes from predicting which modeled hail storms will not produce hail. Size model adds little skill.
- Relative differences in forecast performance vary by region based on event frequency and storm type.

Acknowledgements. This National Center for Atmospheric Research is sponsored by the National Science Foundation. This work is also part of the Severe Hail Analysis, Representation and Prediction project funded by NSF Grant AGS-1261776. The NCAR Ensemble and the analysis for this work were performed on the Yellowstone Supercomputer.