Evaluation of Hail Size Forecasting Models during the 2016 Hazardous Weather Testbed Spring Experiment

David John Gagne

National Center for Atmospheric Research

Coauthors: Rebecca Adams-Selin, Greg Thompson, Burkely Gallo, Amy McGovern, Glen Romine, Craig Schwartz, Nate Snook, and Ryan Sobash

Motivation

- Convection-allowing models can partially resolve individual storms but not their associated hazards
- Severe hazard diagnostics provide direct assessments of severe hazard potential and intensity, unlike storm surrogates
- Three hail diagnostic methods were evaluated during the 2016 NOAA Hazardous Weather Testbed Spring Experiment
- Both subjective and statistical evaluations of methods were performed

NWP Model Output and Hail Observations

Ensemble	CAPS Ensemble	NCAR Ensemble
Members	19 WRF-ARW	10 WRF-ARW
Grid Spacing	3 km	3 km
Microphysics	Thompson, Morrison, MY, P3	Thompson
PBL Schemes	MYJ, MYNN, YSU	MYJ
Data Assimilation	3DVAR + cloud analysis	DART Cycled Ens. Adjustment KF

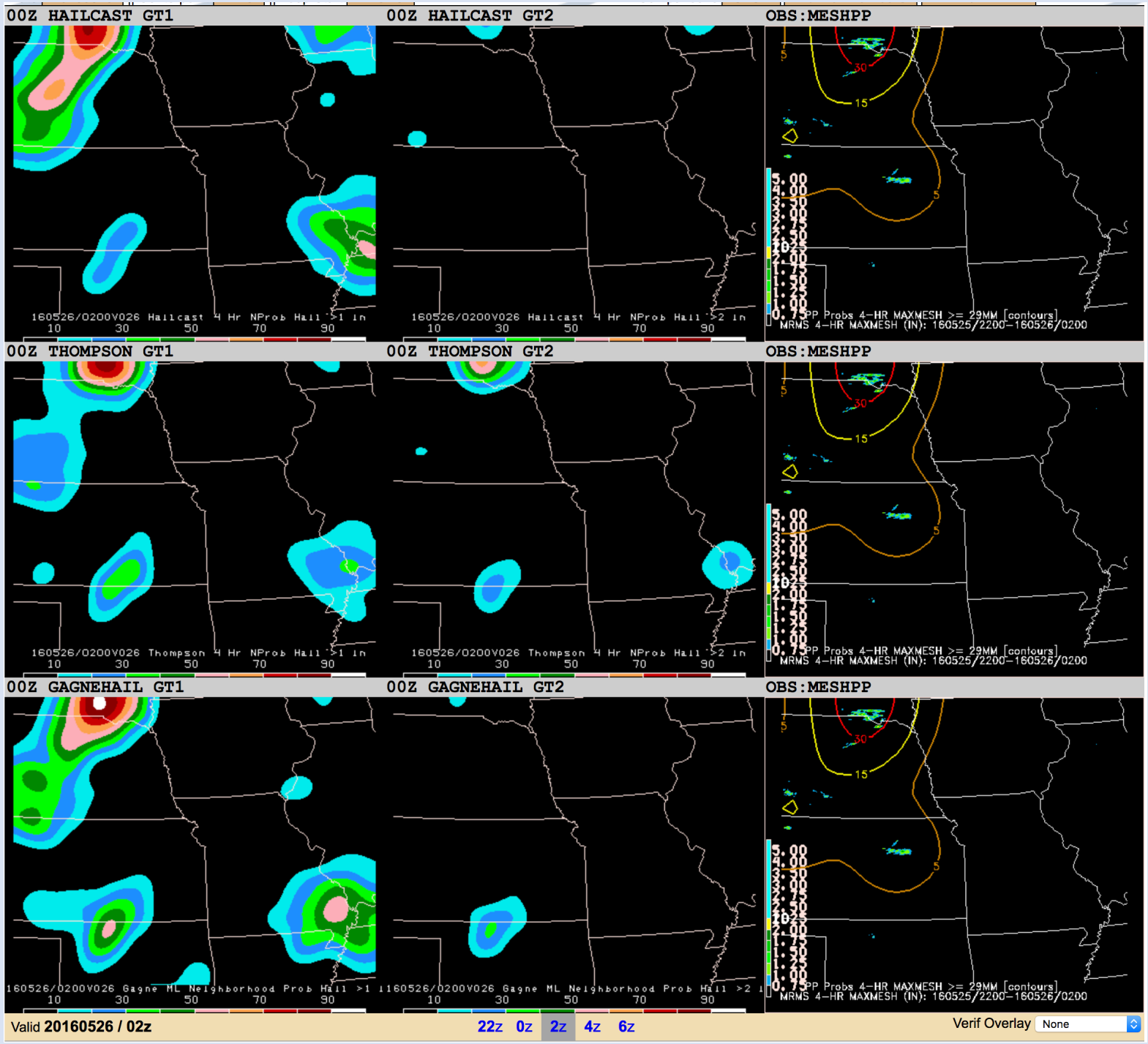
Hail Observations: NOAA NSSL Multi-Radar Multi-Sensor Maximum Expected Size of Hail: estimate of maximum hail size derived from radar reflectivity above freezing level. Estimate calculated from 3D radar mosaic.

Hail Forecast Methods

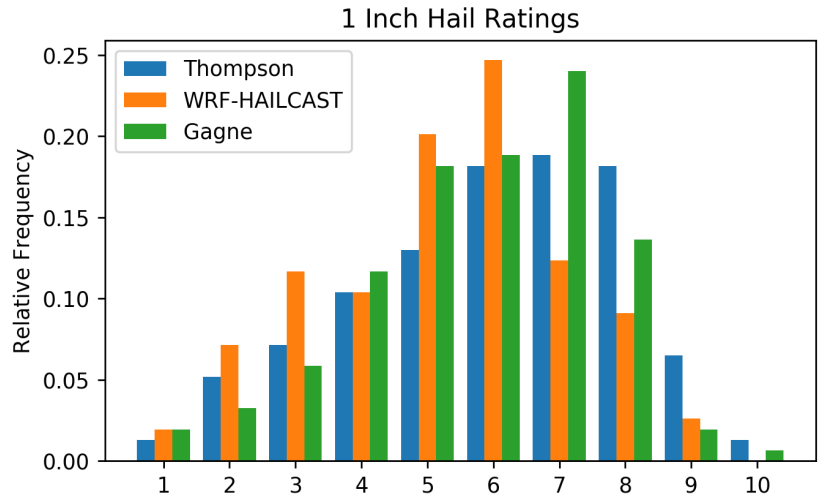
- Thompson Hail Size Method
  - Estimate hail size directly from microphysics size distribution
  - Finds largest diameter that exceeds a specified number concentration
- WRF-HAILCAST
  - 1D hail growth model embedded in each grid cell
  - Grows ensemble of hailstones based on vertical profile
  - Triggered when updraft speed exceeds a certain threshold
- Gagne Hail Model
  - Storm-based machine learning hail forecast method
  - Storm tracker identifies potential hailstorms, extracts storm and environment information
  - Machine learning models predict probability of hail and spatial hail size distribution

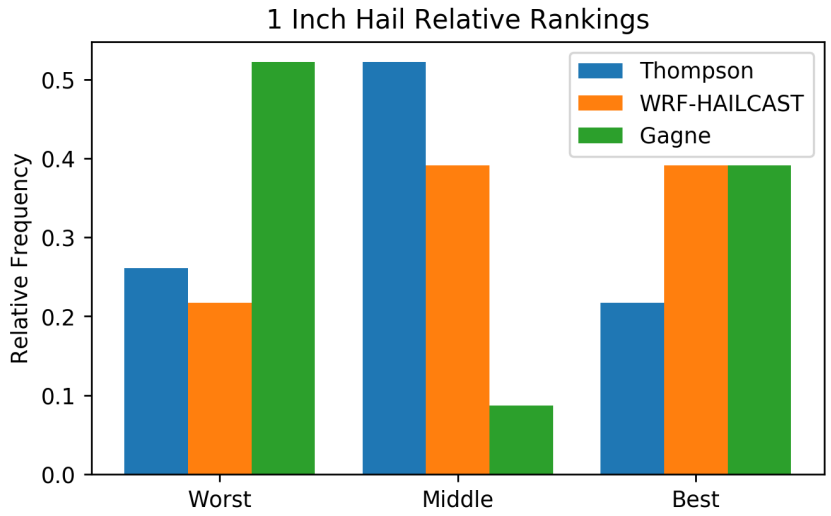
Subjective Verification

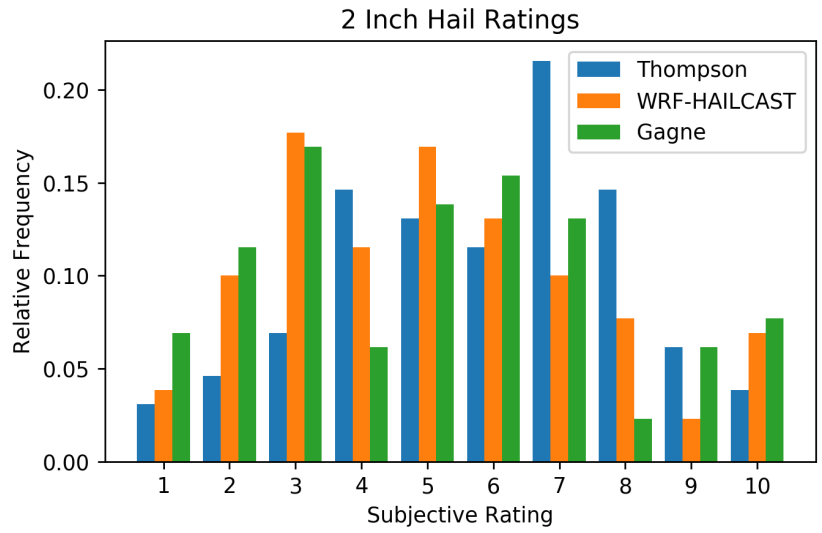
- A group of forecasters viewed a time series of 4-hour probabilities of 1 inch and 2 inch hail from each method
- Compared with MESH and preliminary storm reports
- Rated the overall forecast quality from 1 (poor) to 10 (great)
- Performance over time was aggregated
- Separate ratings for 1 and 2 inch hail

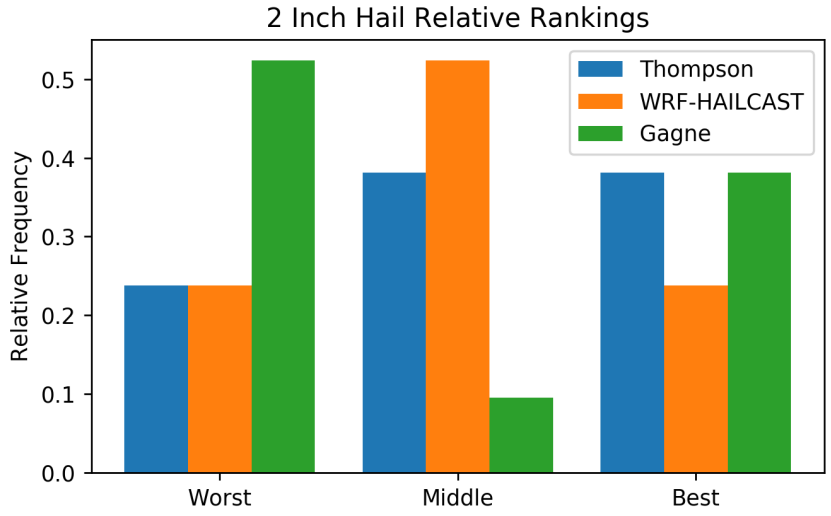

00Z HAILCAST GT1
00Z HAILCAST GT2
OBS:MESHPP
160526/0200V026 Hailcast 4 Hr NProb Hail >1 in
160526/0200V026 Hailcast 4 Hr NProb Hail >2 in
10 30 50 70 90
MRMS 4-HR MAXMESH (IN): 160525/2200-160526/0200
00Z THOMPSON GT1
00Z THOMPSON GT2
OBS:MESHPP
160526/0200V026 Thompson 4 Hr NProb Hail >1 in
160526/0200V026 Thompson 4 Hr NProb Hail >2 in
10 30 50 70 90
MRMS 4-HR MAXMESH (IN): 160525/2200-160526/0200
00Z GAGNEHAIL GT1
00Z GAGNEHAIL GT2
OBS:MESHPP
160526/0200V026 Gagne ML Neighborhood Prob Hail >1
160526/0200V026 Gagne ML Neighborhood Prob Hail >2
10 30 50 70 90
MRMS 4-HR MAXMESH (IN): 160525/2200-160526/0200
Valid 20160526 / 02z
22z 0z 2z 4z 6z
Verif Overlay None

Subjective Verification Statistics





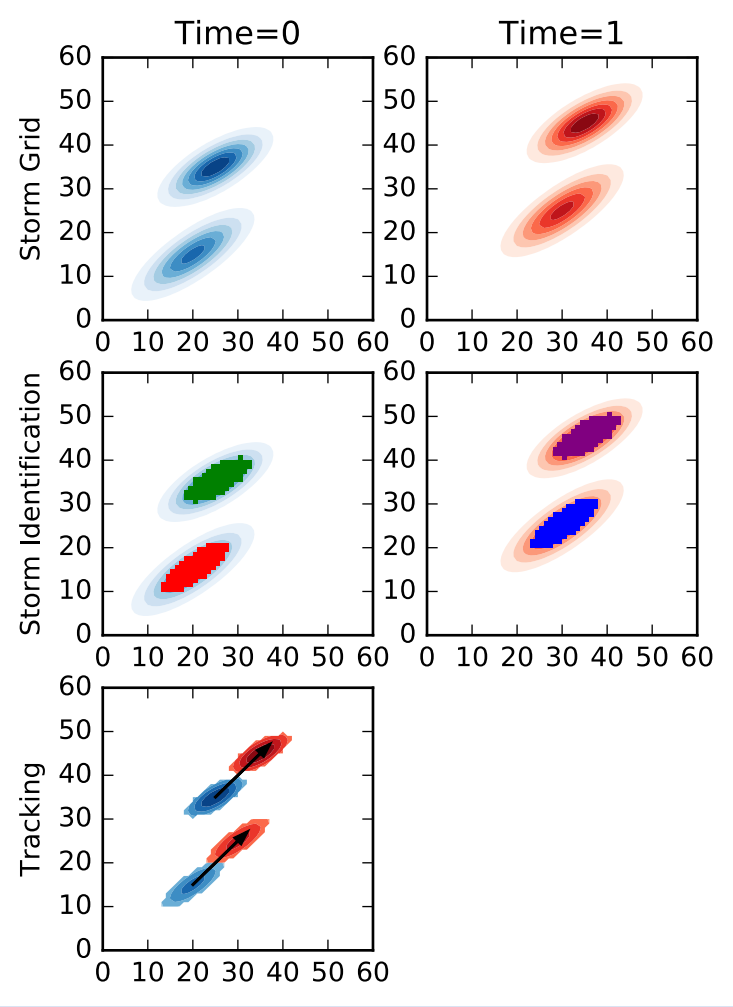




Observations

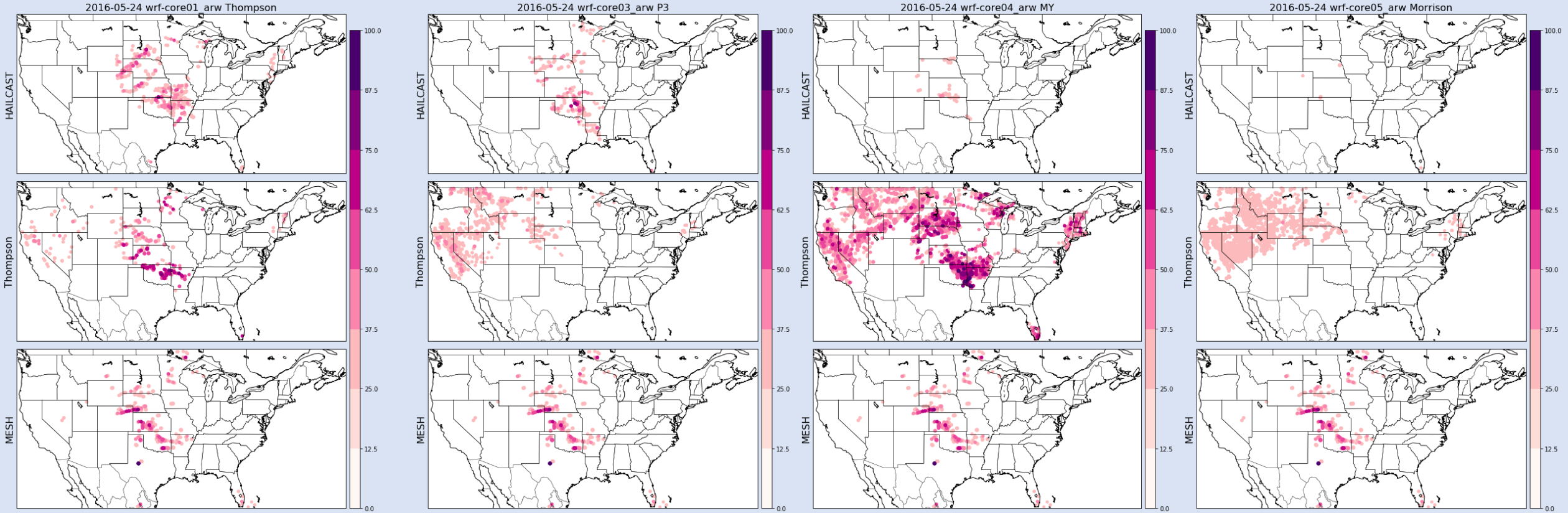
- None of the hail diagnostics are clearly superior
- Gagne hail tends to be either best or worst method on a given day
- Gagne hail is more conservative with large hail, which benefits it on days with no large hail
- The average rating for Thompson is higher for 2 inch hail than 1 inch hail

Storm-Based Verification

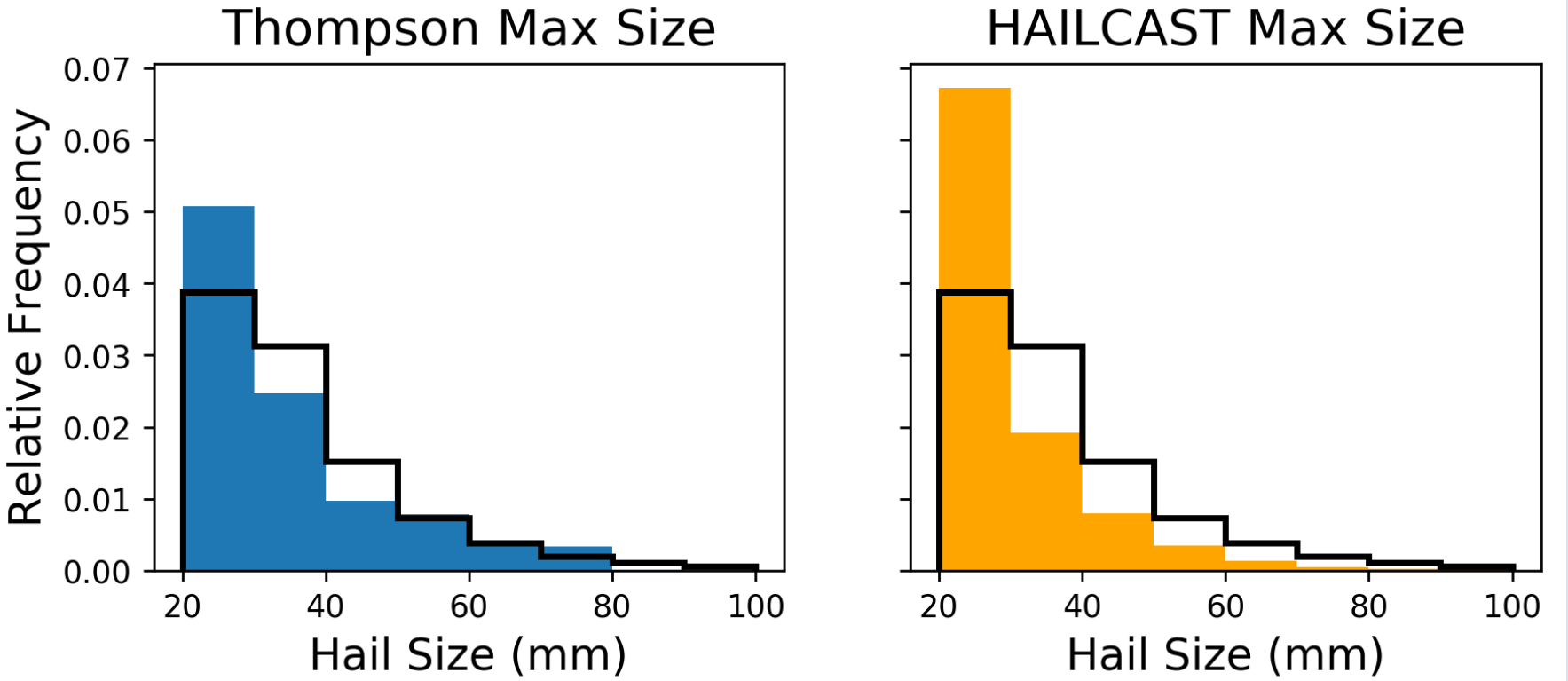


1. Use enhanced watershed to identify storms in the output of each hail forecast method
2. Apply enhanced watershed to MRMS MESH data
3. Extract statistics on storm and environment variables within each hail object
4. Compare place of objects from each hail model

Microphysics Effects, or the Problem with Multi-Physics Ensembles



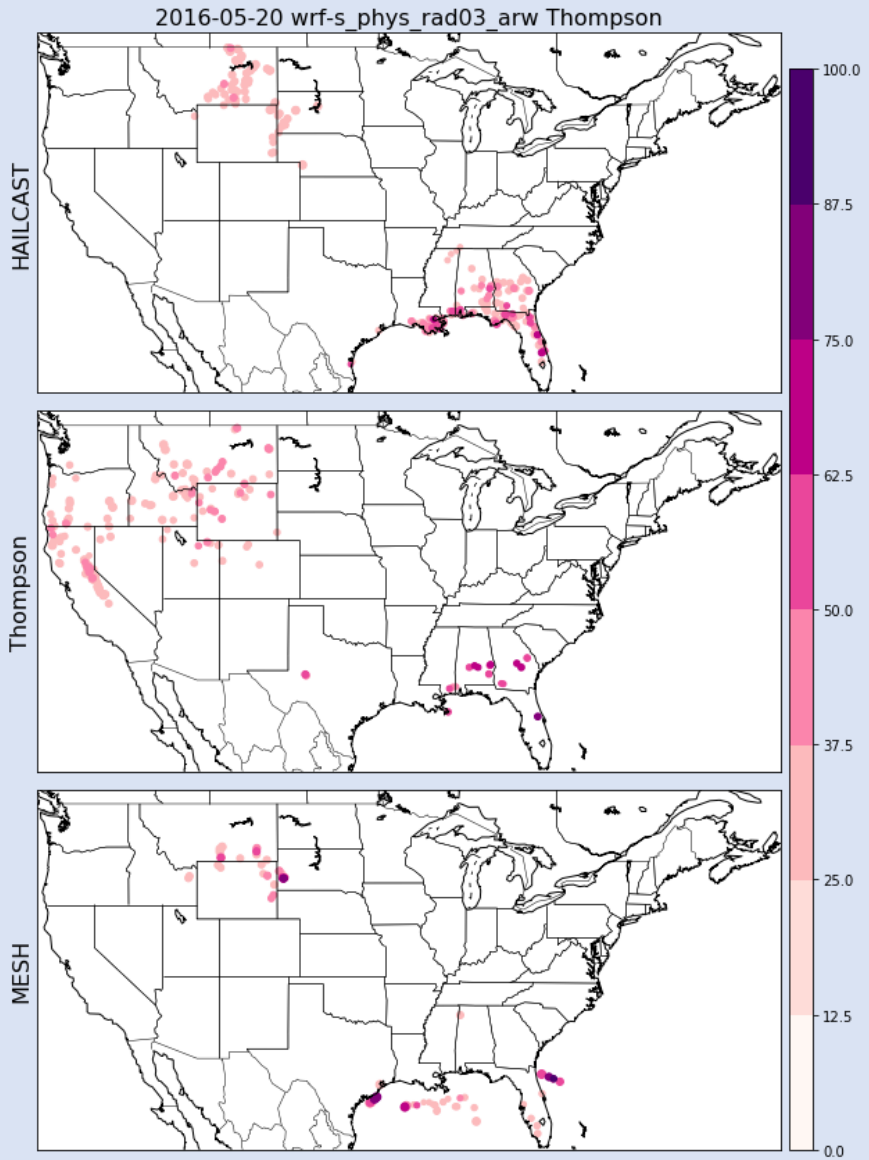
Hail Size Distribution Comparisons



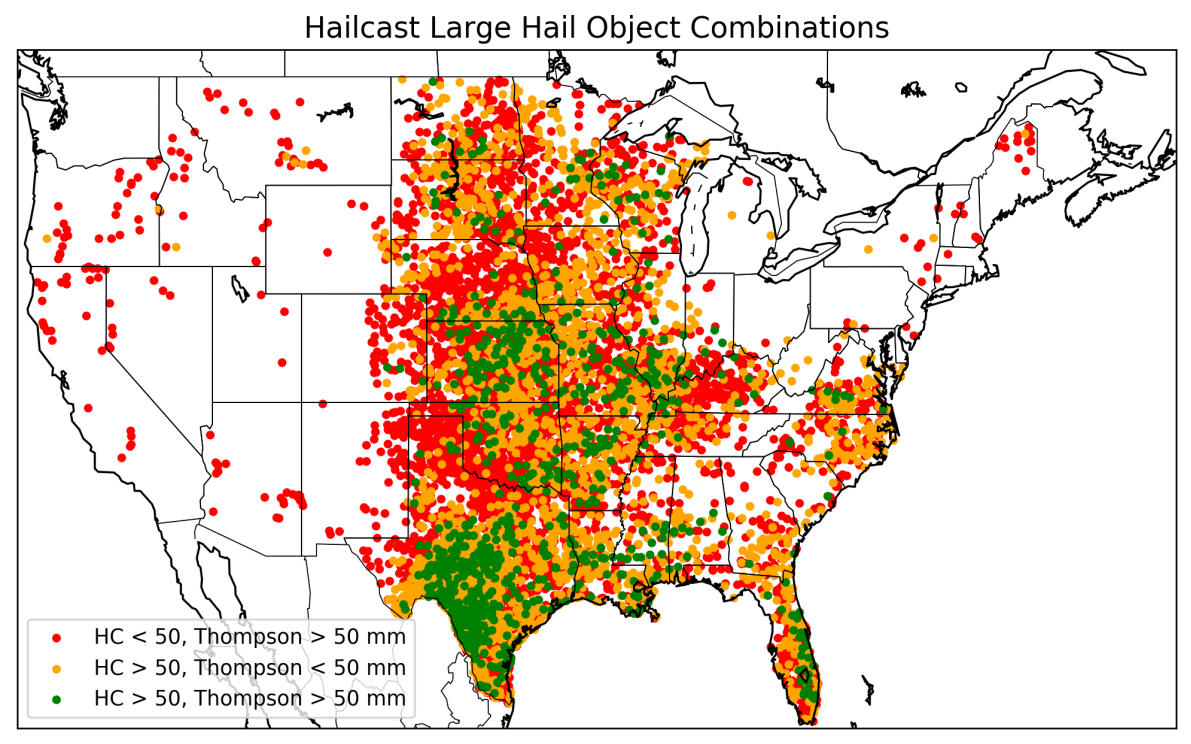
- WRF-HAILCAST overestimates the frequency of small hail and underestimates the frequency of large hail.
- Thompson matches the relative frequency of MESH for 50 to 80 mm hail but underestimates the frequency otherwise except between 20 and 30 mm

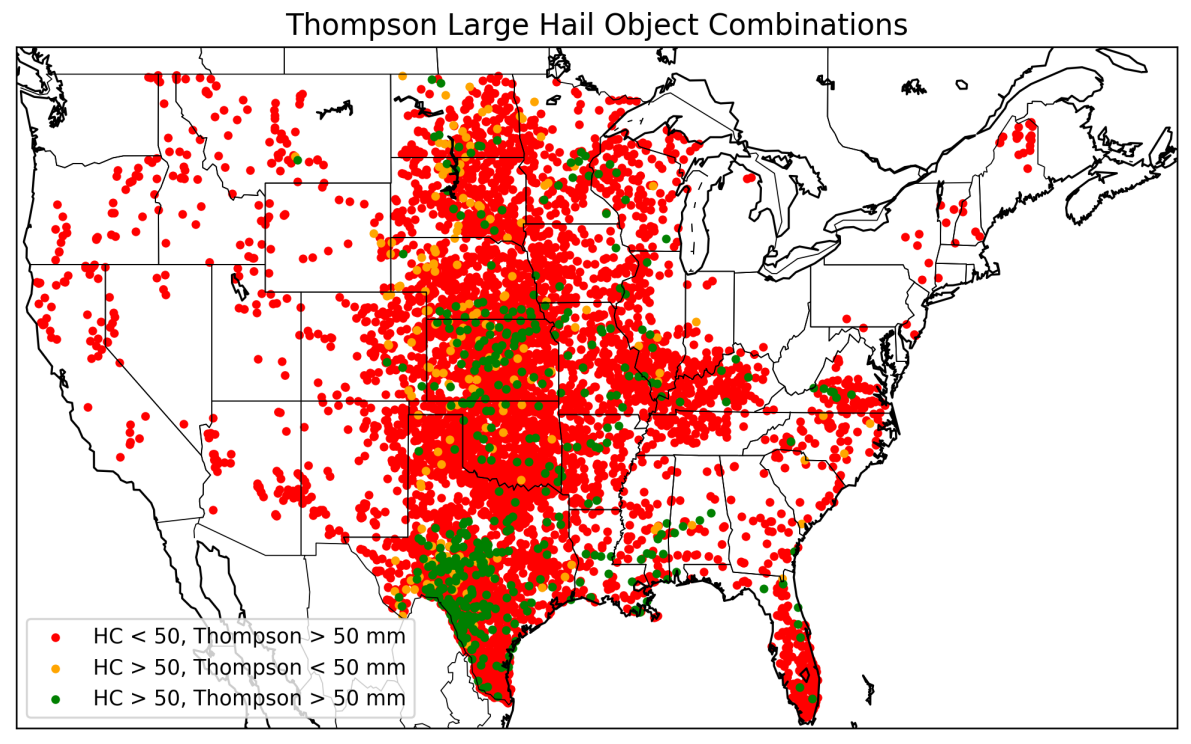
Hail Storm Biases

- HAILCAST produces hailstorms in the Southeast along the coast associated with wind reports but not hail
  - May not be melting hail enough in those environments
- Thompson under-produces hail in the Southeast but produces many spurious hailstorms in the Mountain West
  - Tied to graupel reaching high altitude surfaces



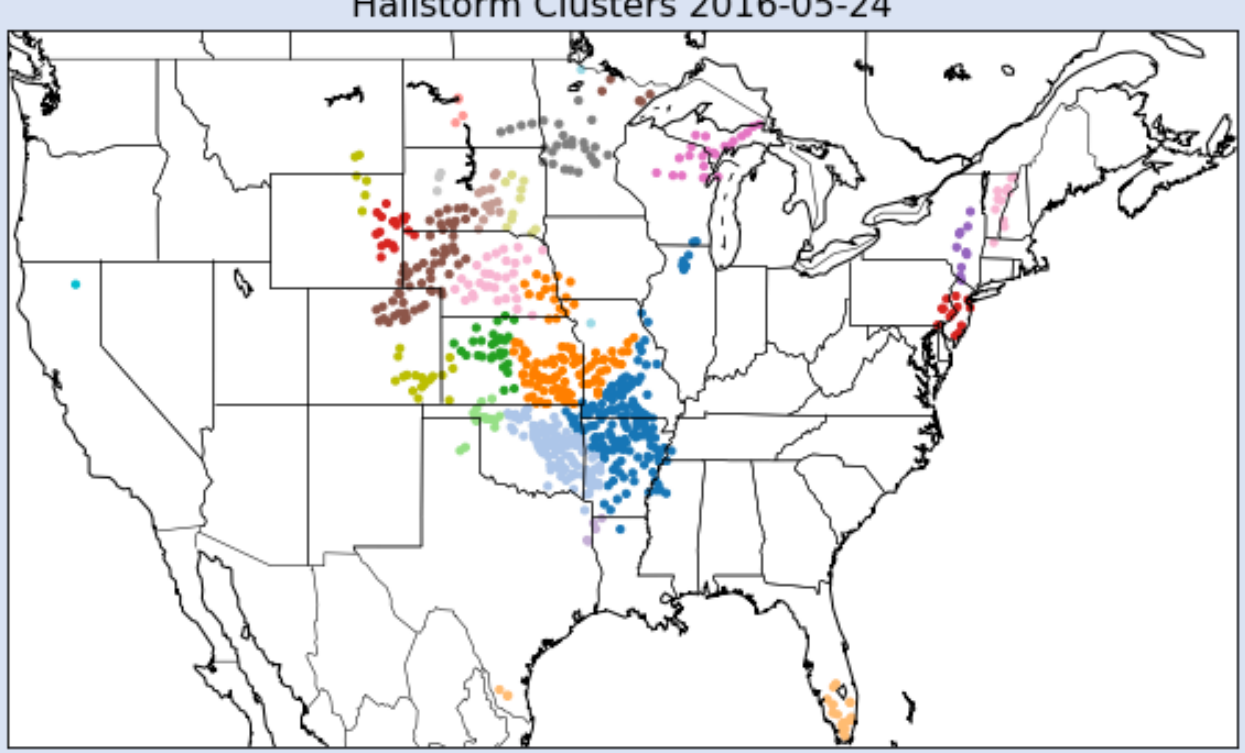
Large Hail Object Pairings





Ongoing Work

- Hailstorm clustering
  - Grouping storms spatially and examining overlap with observed storms
  - Off-the-shelf clustering methods do not produce desired results
  - Best option so far: agglomerative clustering with maximum distance constraint
- Verification conditioned on environment
  - Want to evaluate each algorithm on how it performs in different parts of storm environment feature space
  - Interested in different combinations of CAPE and shear



Summary

- The 2016 Hazardous Weather Testbed Spring Experiment provided an opportunity to compare 3 hail size diagnostics for convection-allowing models
- Subjective verification shows no model consistently performs the best
- More work needs to be done to optimize other hail algorithms for microphysics other than Thompson
- Both HAILCAST and Thompson produce large hail but often not in the same storms

Contact Information
Email: dgagne@ucar.edu
Twitter: @DJGagneDos
Github: djgagne