Emerging Methods for Post-Processing

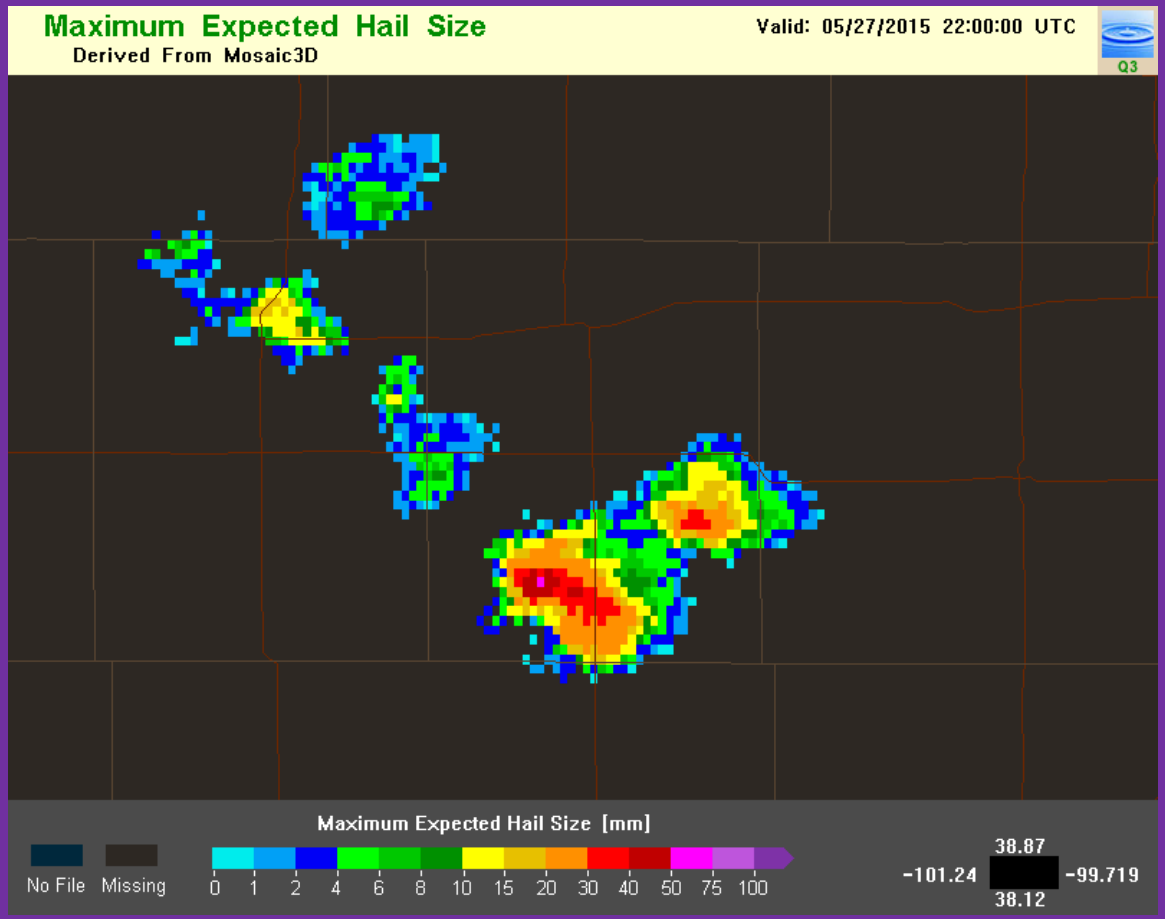
David John Gagne, University of Oklahoma/NCAR

The Future of Statistical Post-Processing in NOAA and the Weather Enterprise

20 January 2016

Trends Facing Statistical Post-Processing

- Finer model resolution
  - Convection-allowing models
  - Mesoscale global models
- Ensembles
  - Multi-model, multi-physics
- A wider range of observations
  - MRMS, GOES-R
  - Mesonets
  - Crowd-sourced observations
  - Connected vehicles

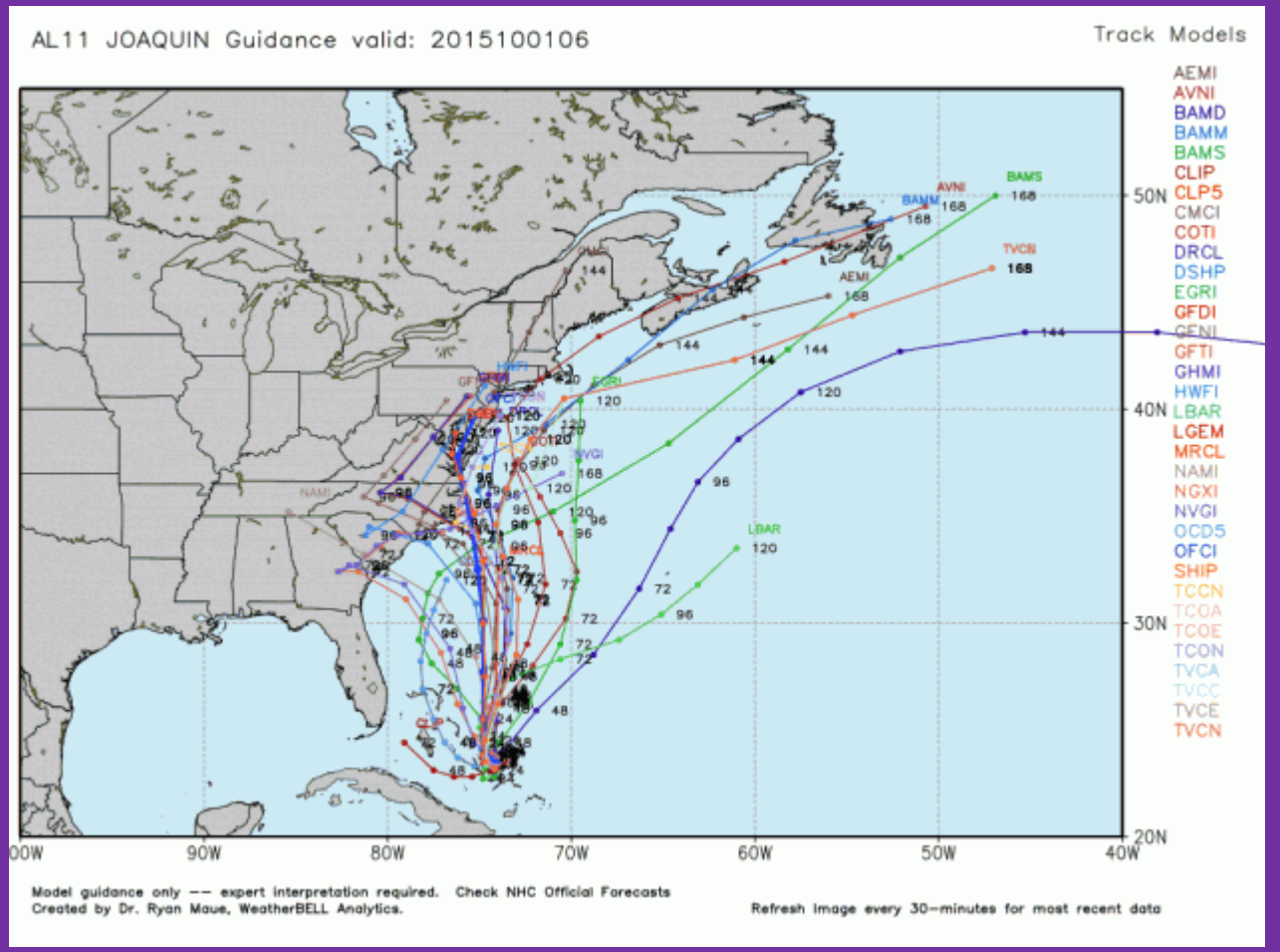


Future Focus Areas

- Object-based post-processing
- Machine Learning Models and Configurations
- Machine Learning products

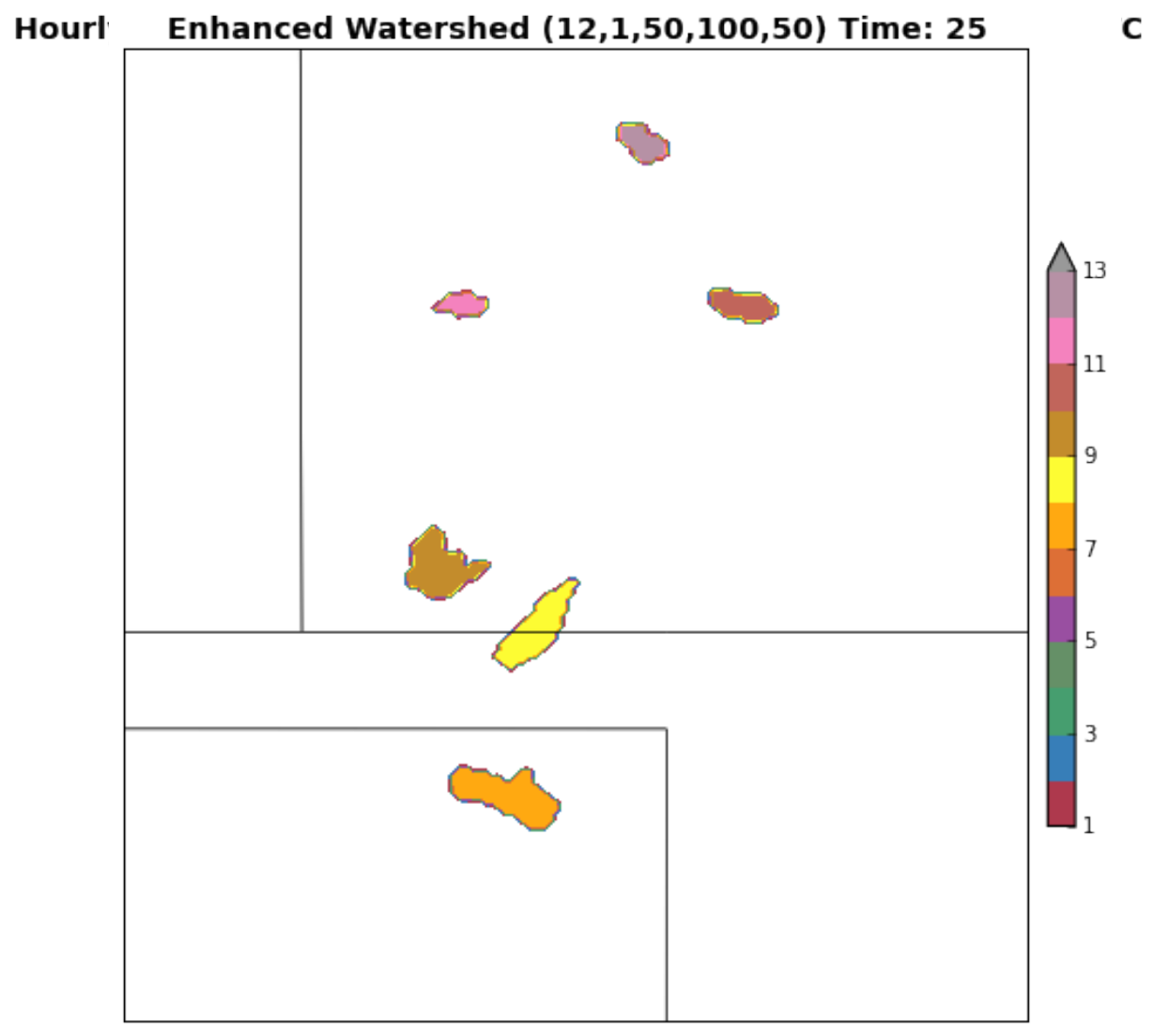
Object-Based Post-Processing

- Identify areas of interest in model output
- Track them through time
- Extract data about them
  - Intensities
  - Shape properties
  - Other collocated variables
- Analyze collective information
- Calibration and uncertainty estimation



http://nj1015.com/joaquin-now-a-powerful-category-3-hurricane-effects-start-in-n-j-soon/

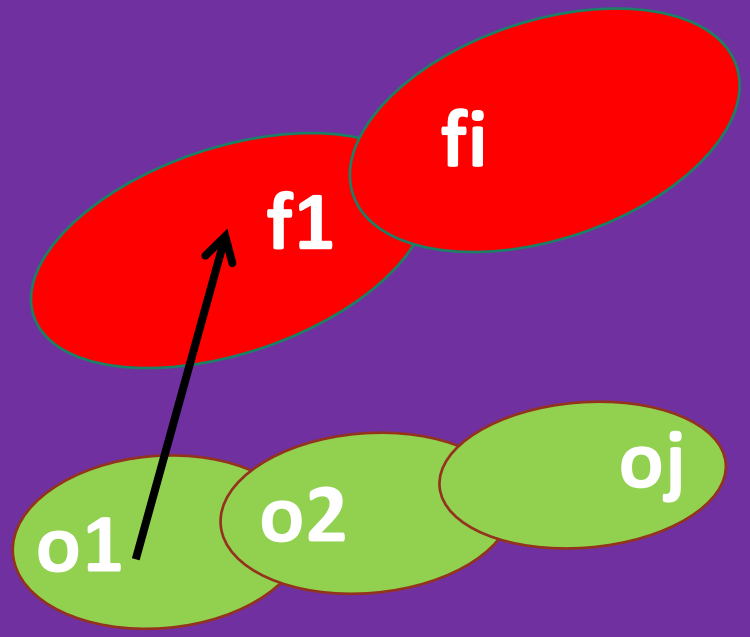
Potential Hailstorm Identification



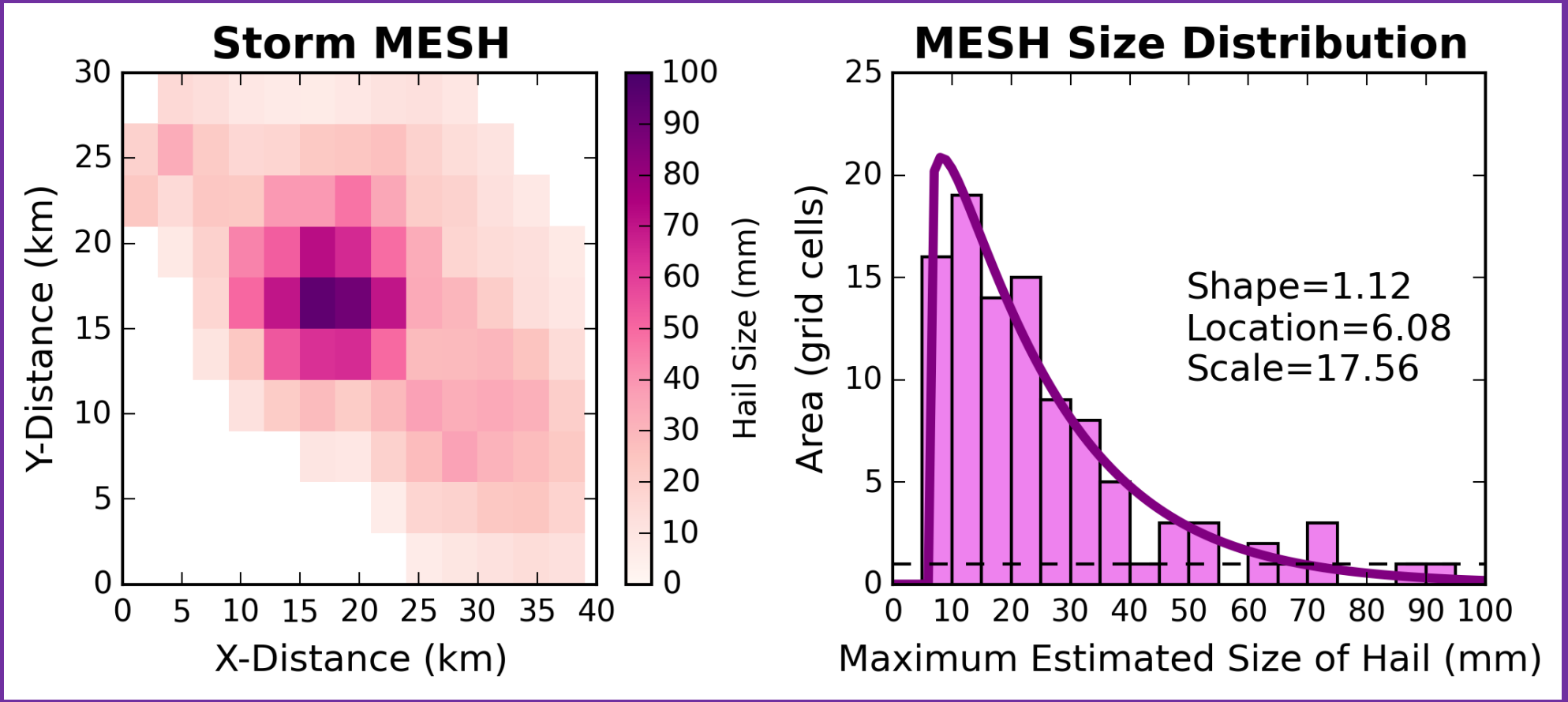
- Hailstorm Proxy: Hourly Max Column Integrated Graupel
- Enhanced watershed (Lakshmanan et al. 2009) used to identify storm “objects”
- Objects must have area within specified range

Object Matching

- Matching forecast and observed objects requires subjective choices about criteria
- Weighted average of space, time and track properties
  - 50%: Start centroid Euclidean distance
  - 30%: Start time absolute difference
  - 10%: Duration absolute difference
  - 10%: Mean area absolute difference
- Distance and time differences are tightly constrained
- Duration and area are loosely constrained



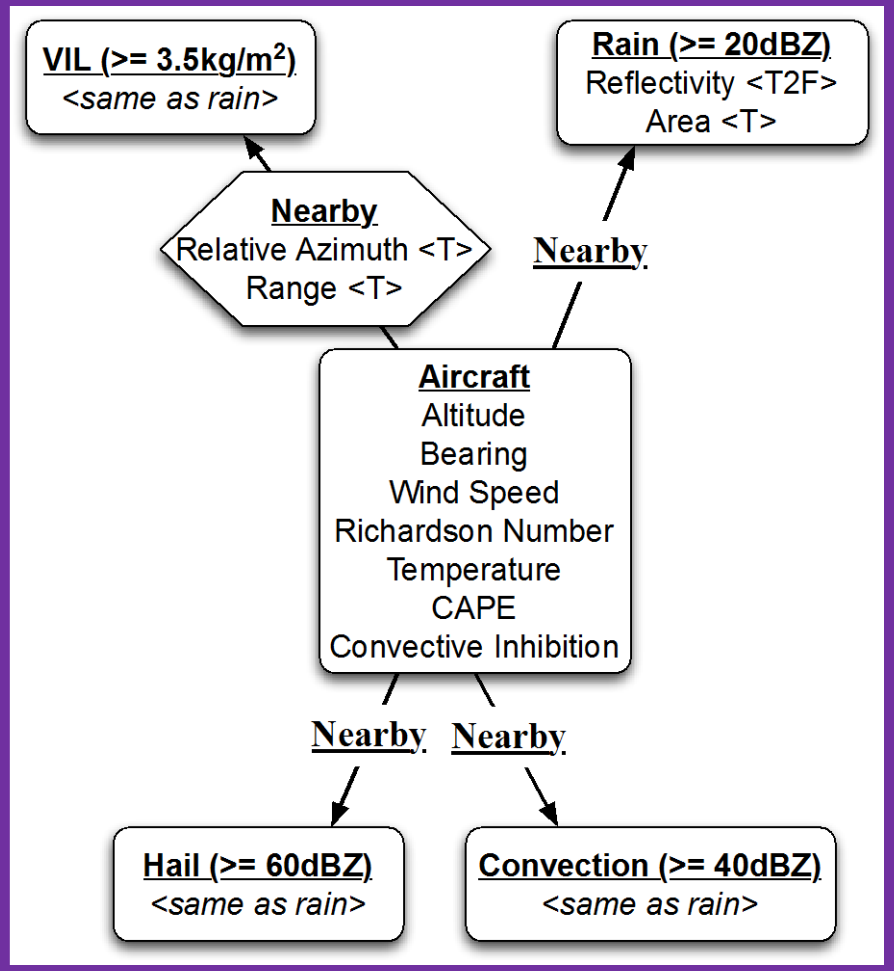
Gamma Hail Size Distribution

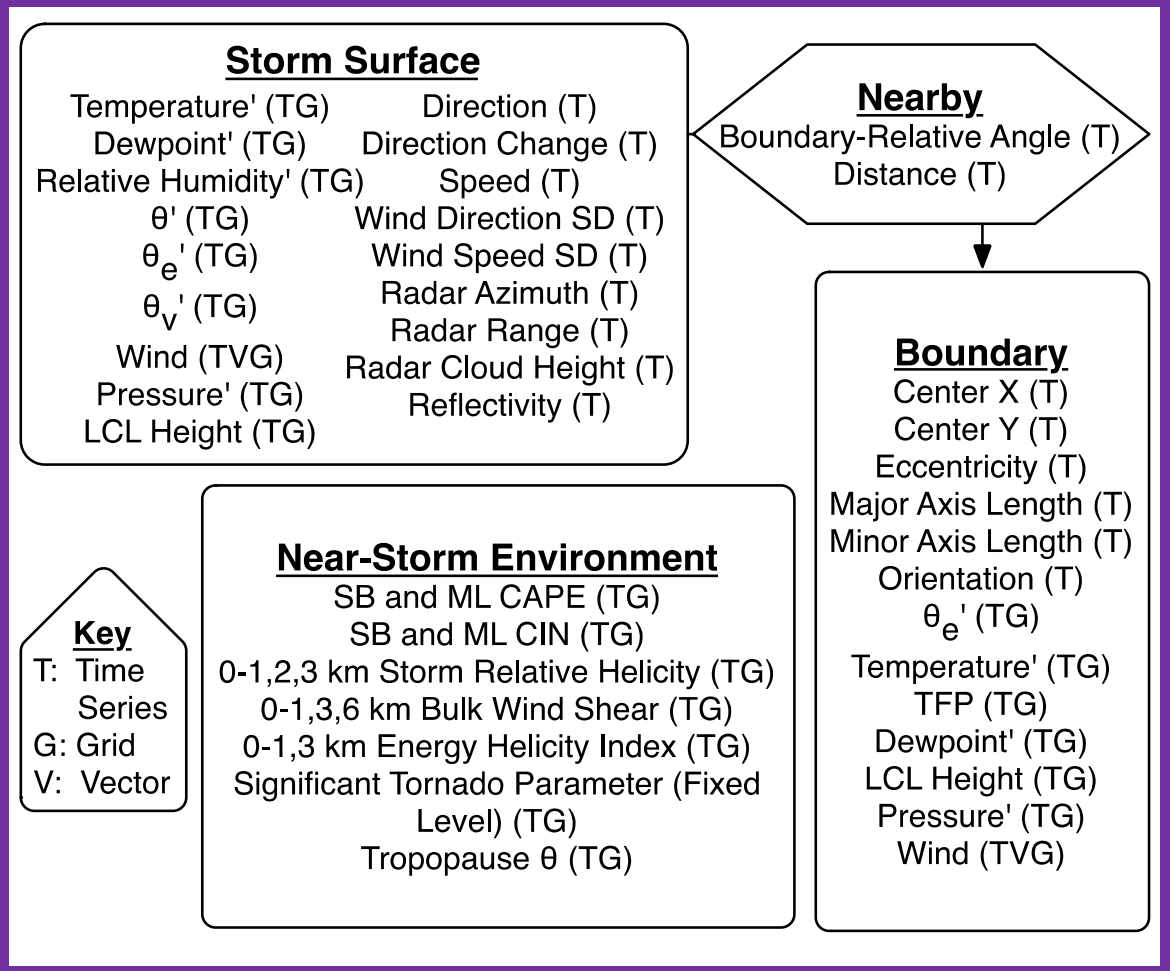


Wide range of hail sizes within MESH object
Large hail occurs within small area

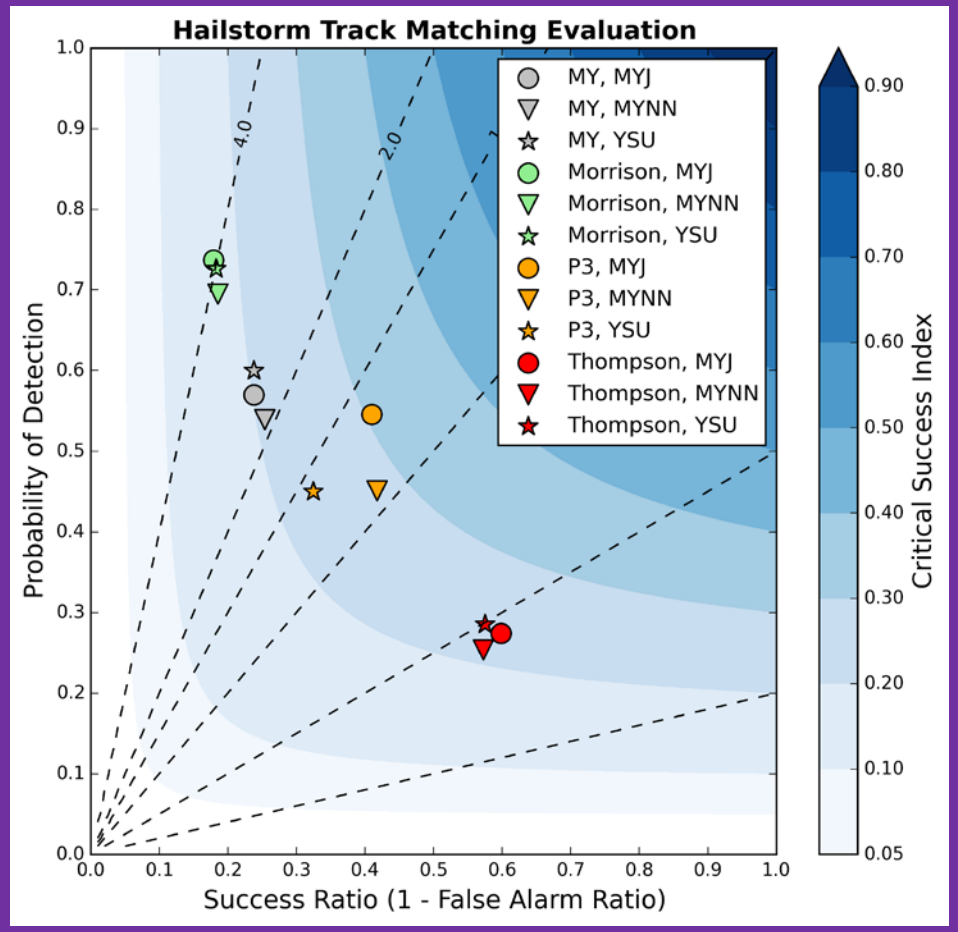
Distribution of MESH pixels
Fit gamma PDF to MESH values
Compress distribution information into 3 values

Spatiotemporal Relational Framework

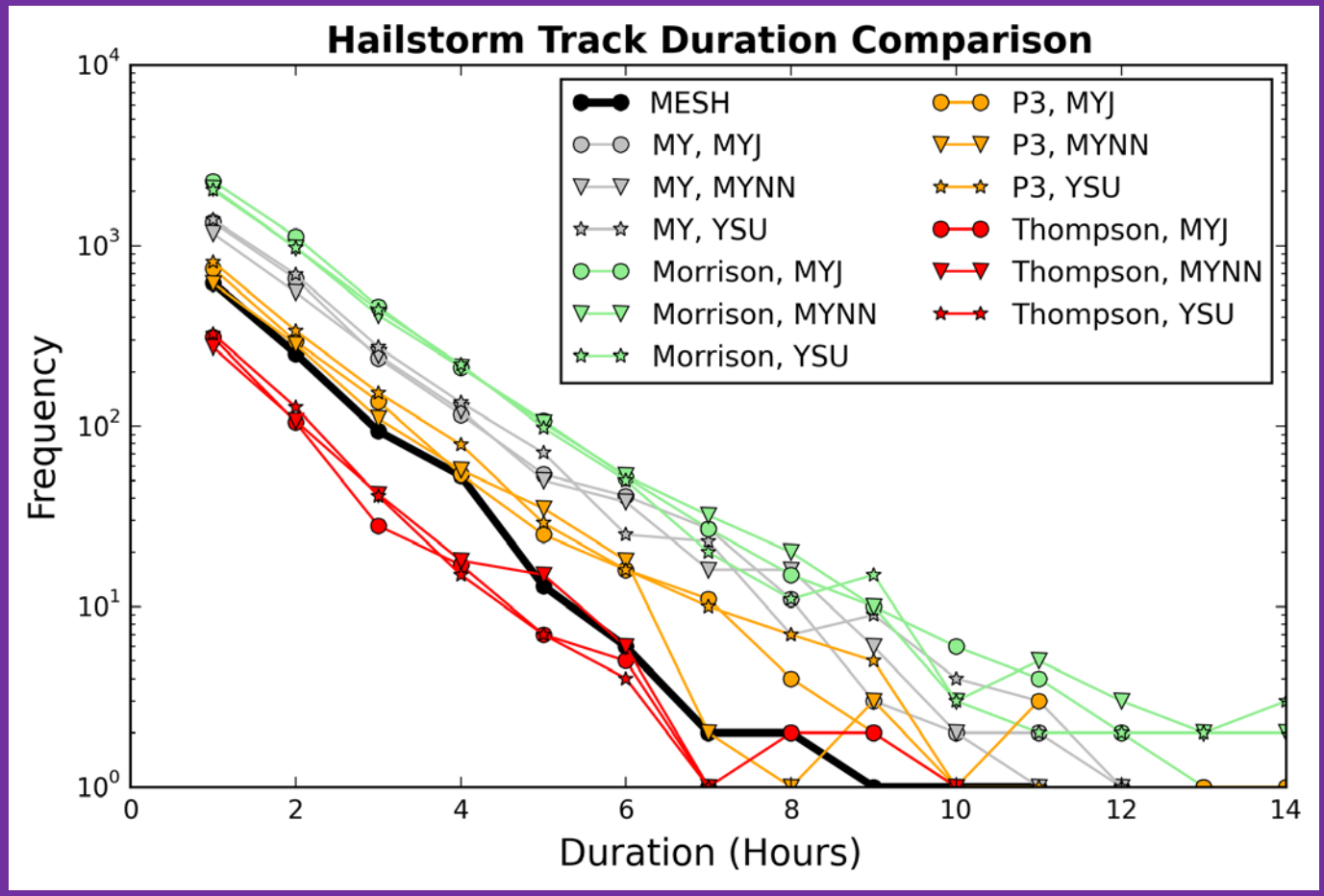




Modeled Storm Properties



Thompson underforecasts, P3 is nearly unbiased, Morrison overforecasts



Duration patterns compare well with MODE-TD analysis (Clark et al. 2014) and storm tracking by hand (Hocker and Basara 2008)

Future Resource: Model Object Databases

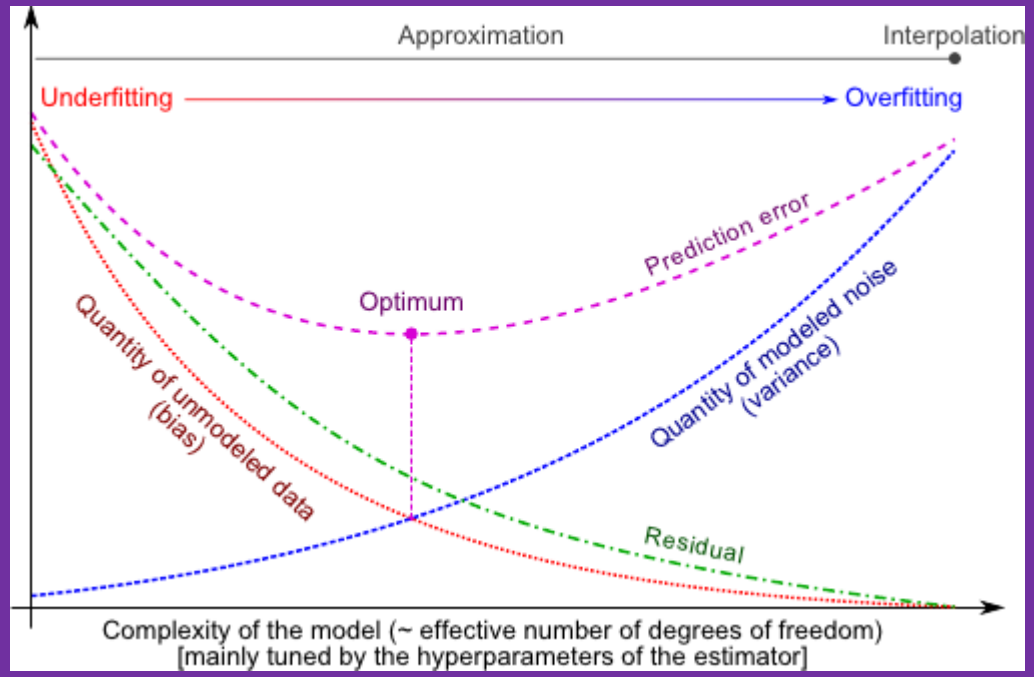
- Collections of objects extracted from operational model output for different phenomena
  - Thunderstorms
  - Fronts
  - Cyclones
  - Heavy rain
- Store information in database linked to API
- API handles requests for objects over time and in certain regions
- Delivers object information as JSON, XML, or CSV

Machine Learning

- Limitations of Multiple Least-Squared Linear Regression
  - Does not scale well with large numbers of input variables
  - Global fit to input data and fixed variance estimate
  - Sensitive to outliers
  - Requires data transformations, smart-subsetting to be effective
- Machine Learning Methods
  - Designed to make predictions from large, high-dimensional datasets
  - Can provide varying confidence and uncertainty estimates
  - Performs variable selection as part of model building process

Regularization and Model Tuning

- Regularization: add bias to model optimization to improve generalization
  - In regression: adding penalty terms
  - In decision trees: restricting the tree depth
- Model tuning: Find optimal model parameters
  - Grid search helps choose model parameters
  - Cross-validation helps mitigate overfitting



http://www.brnt.eu/phd/node14.html

Elastic Net

$$\min_{w} \frac{1}{2n_{samples}}||Xw - y||_2^2 + \alpha\rho||w||_1 + \frac{\alpha(1-\rho)}{2}||w||_2^2$$

(scikit-learn.org)

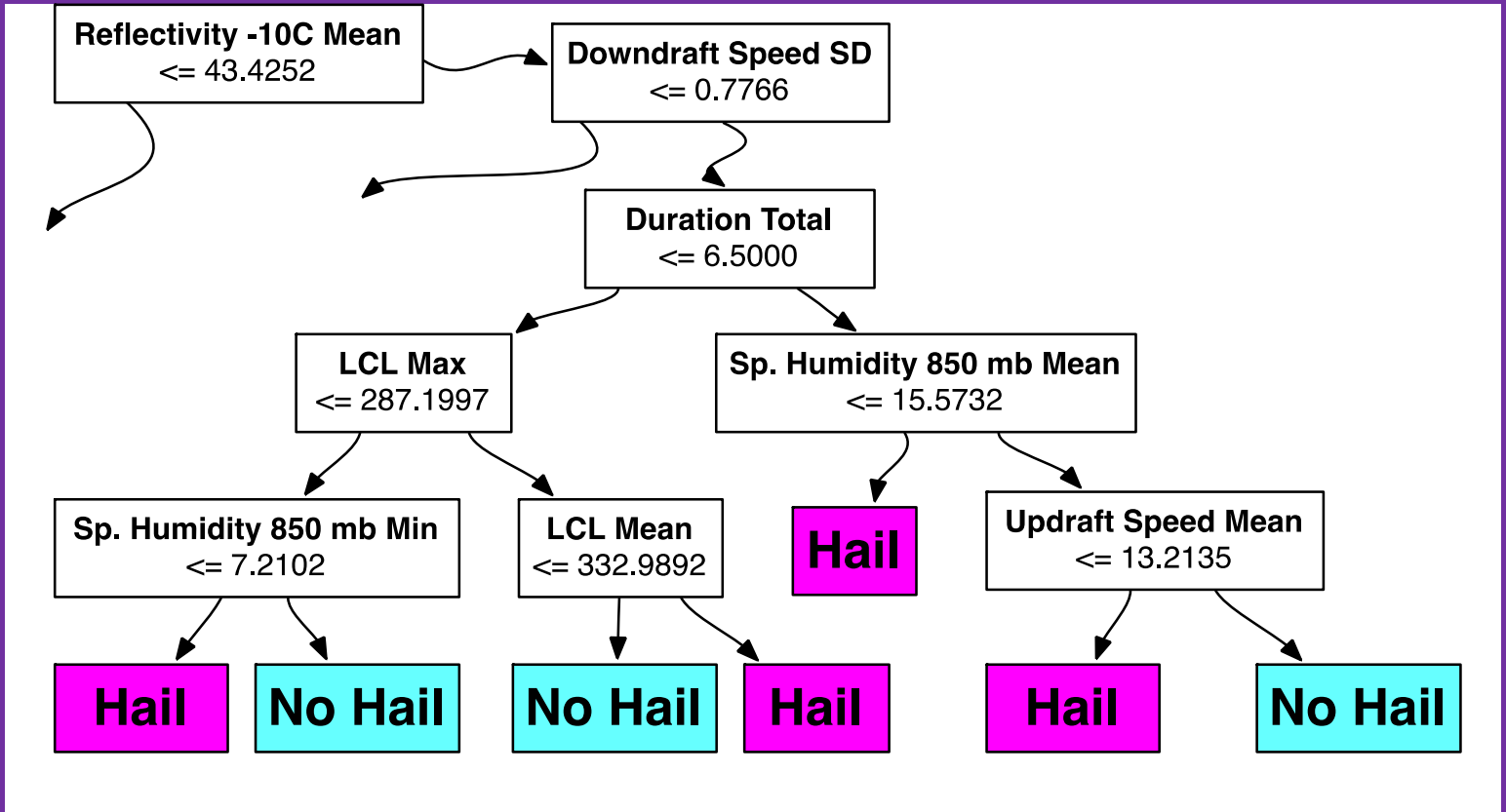
Mean squared training set error

Lasso (absolute weight sum)

Ridge term (squared weight sum)

- Generalized linear model that contains additional penalty terms in the optimization function to penalize large magnitude weights
- **Lasso term** favors sparse set of weights (mostly zeros)
- **Ridge term** favors smaller weights

Decision Tree



Set of hierarchical rules

Outputs

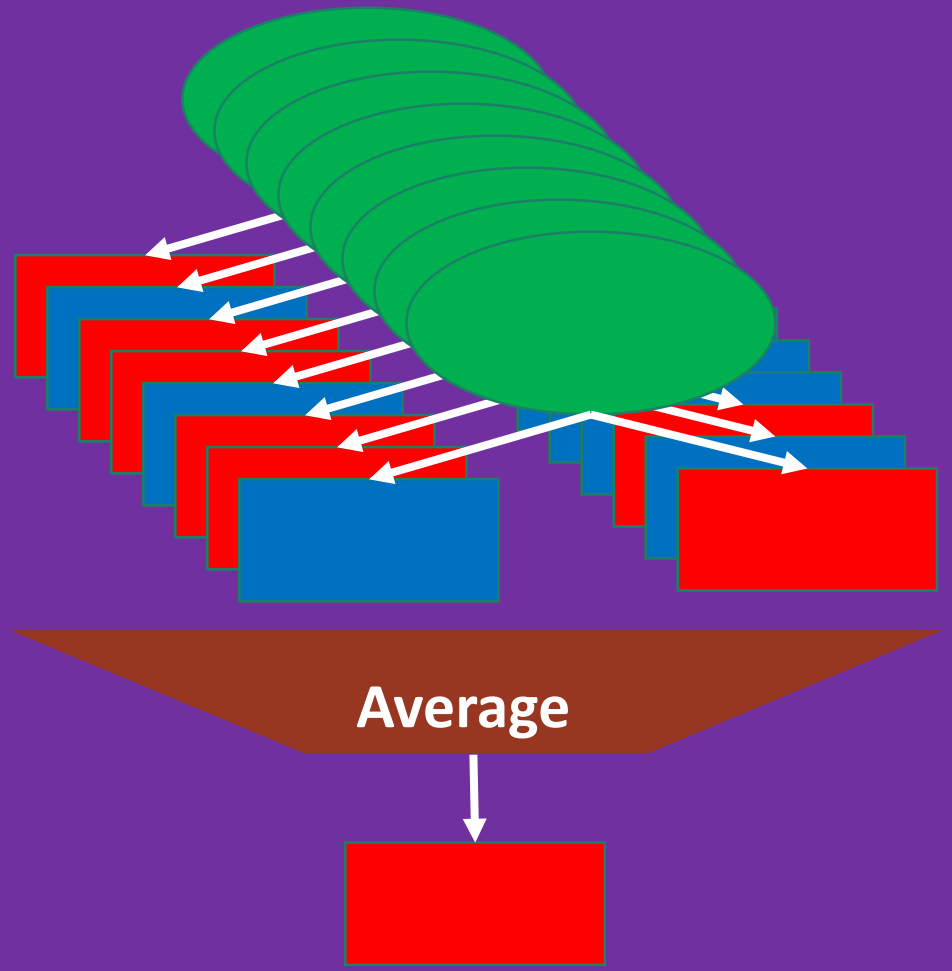
- Binary decision
- Probability
- Constant value

Easy to interpret

Poor predictive accuracy

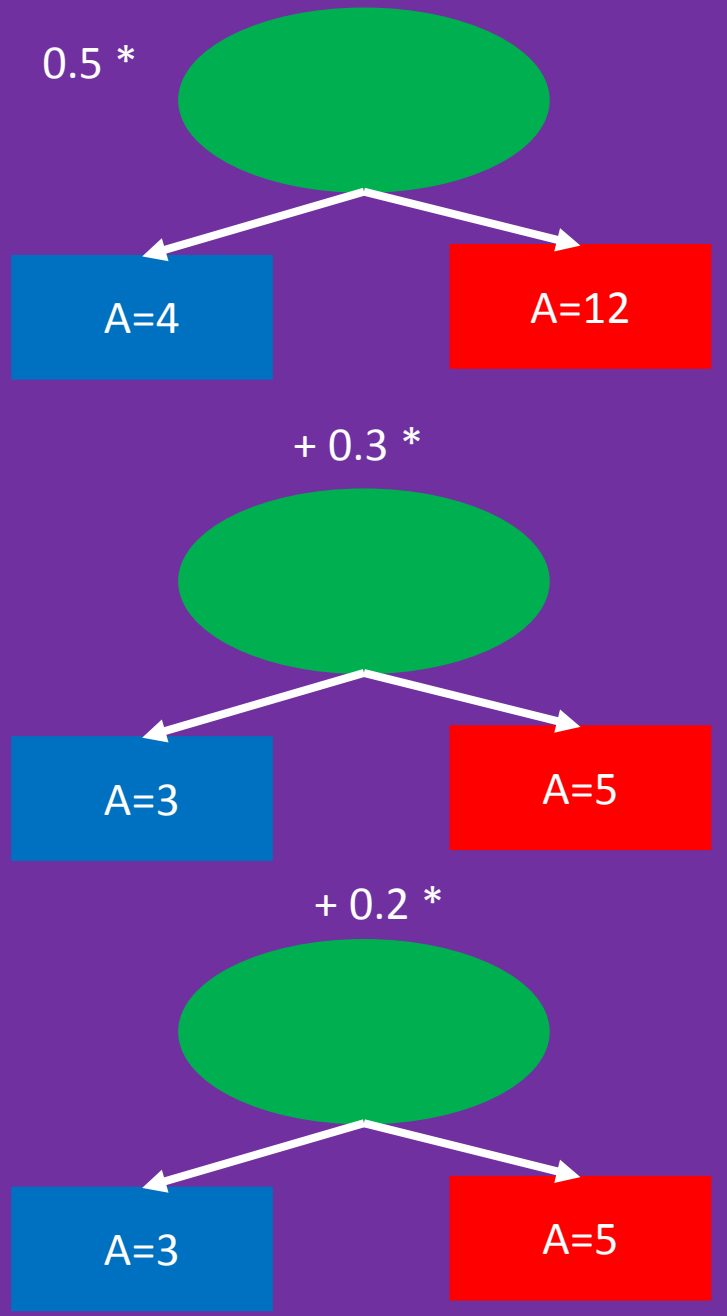
Random Forest

- Ensemble of randomized decision trees (Breiman 2001)
- Two forms of randomness
  - Bootstrap resample training data for each tree
  - Select random subset of input variables for evaluation at each node during training
- Predictions from trees averaged
- Special features
  - High prediction accuracy
  - Automatic feature selection
  - Fast and parallelizable
  - Requires little tuning

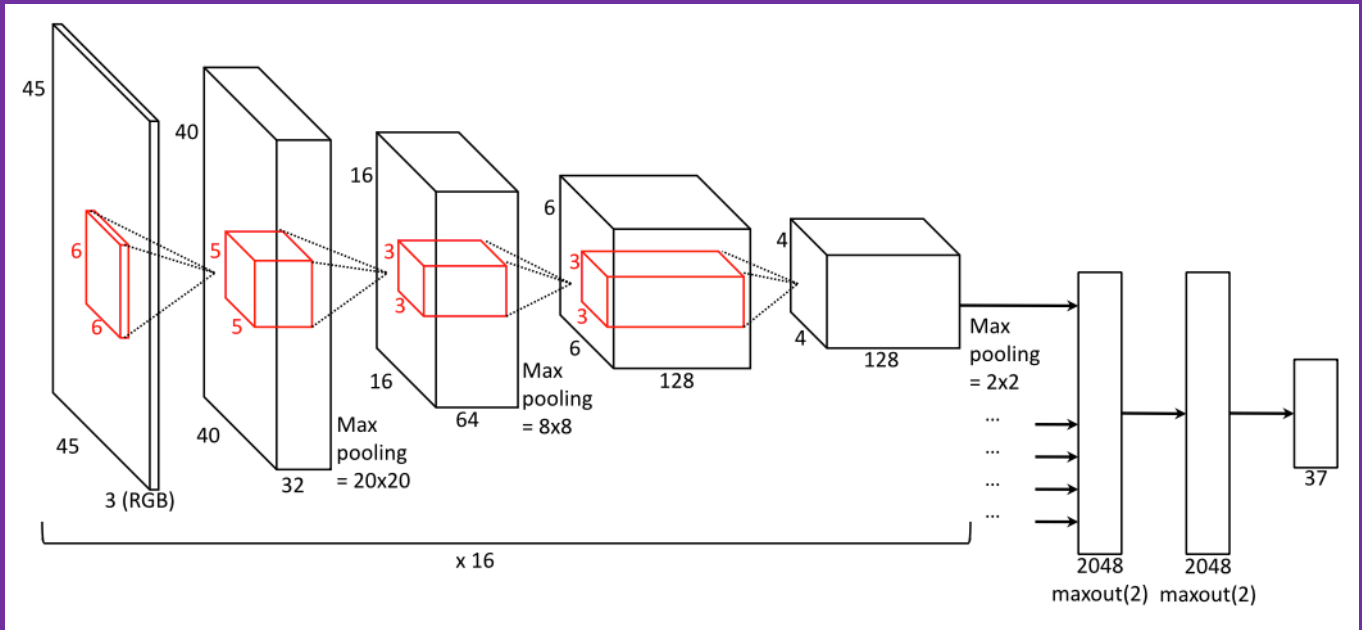


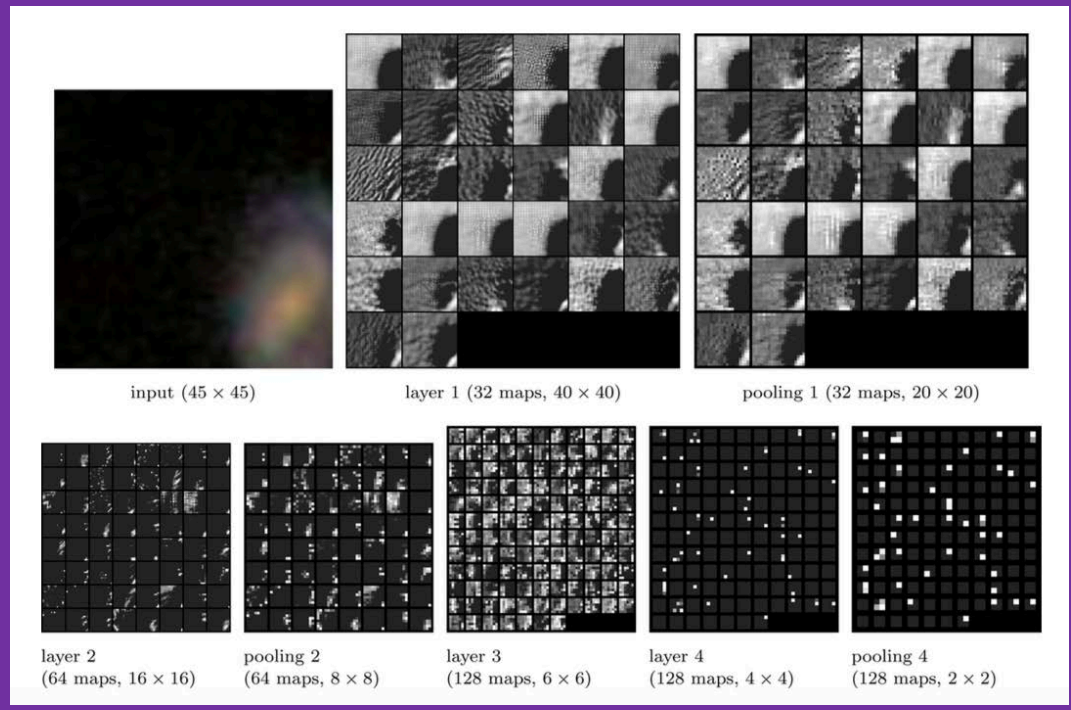
Gradient Boosting Trees

- Additive ensemble of weighted decision trees (Friedman 1999)
- Weights determined by performance on training data
- Focuses on improving predictions of most challenging examples
- Higher performance ceiling than random forest
- Requires more tuning



Convolutional Deep Neural Networks



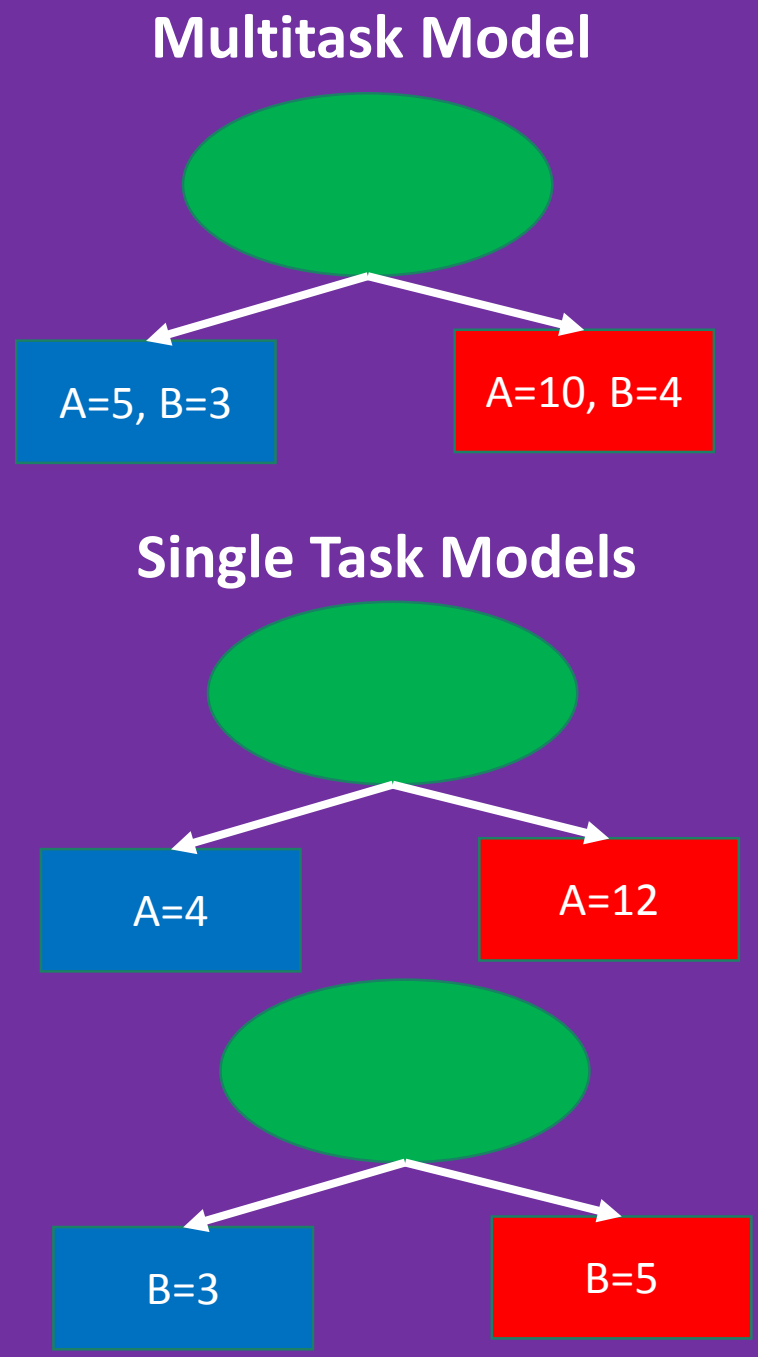


- Extract features from raw image output using convolutional filters
- Images decomposed into feature maps with different scales
- Feature maps fed into neural network and used to generate predictions
- Has caused dramatic improvements in image and speech recognition in recent years

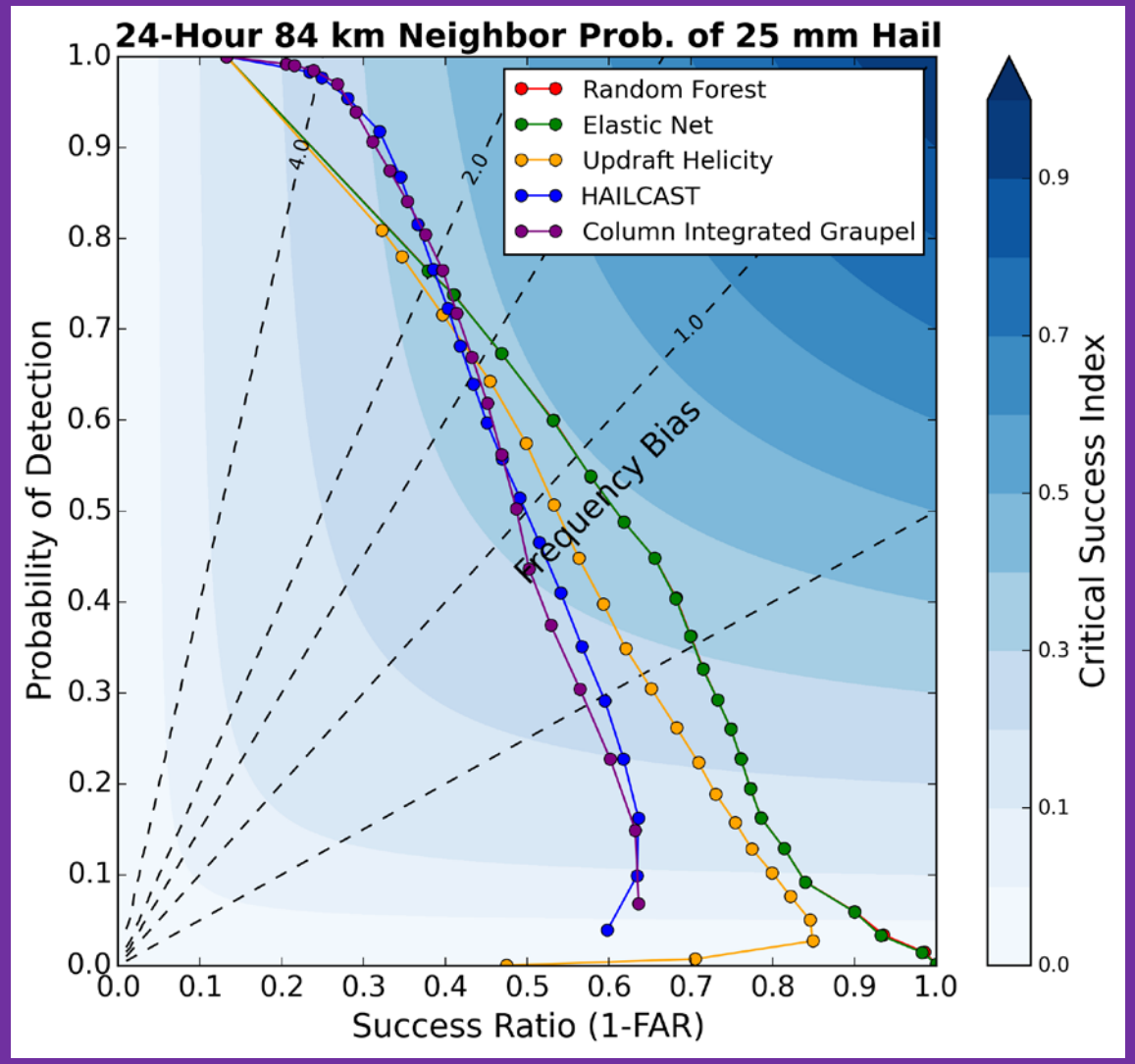
Images from Dieleman et al. (2015), which classified properties of galaxies

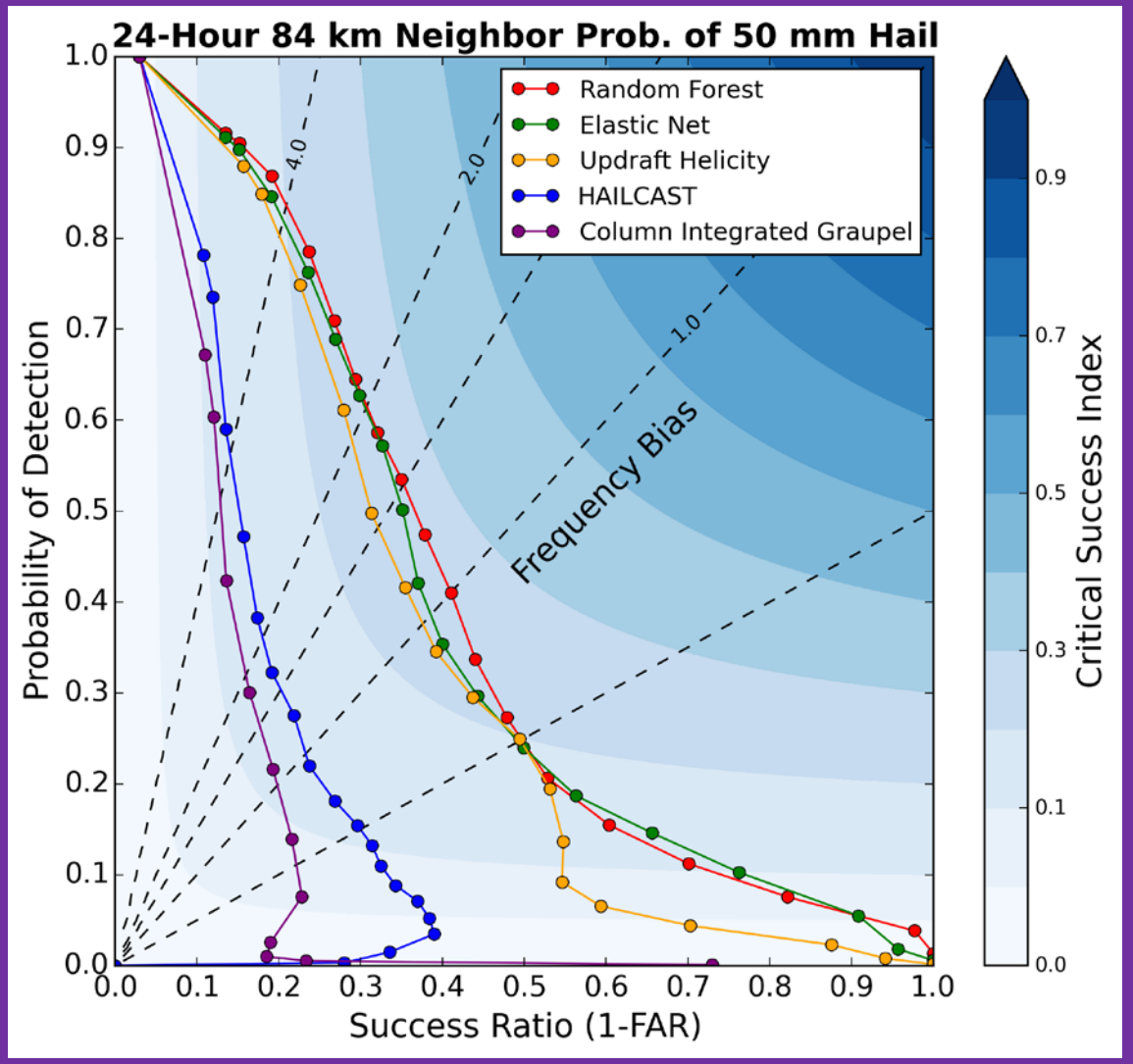
Multitask Learning

- Caruana (1997): multiple labels simultaneously
- Benefits: if labels are correlated, multitask models maintain correlations in predictions
- For Random Forest: splits selected using total mean squared error of normalized labels
- For Elastic Net: regression weights optimized for total normalized mean squared error

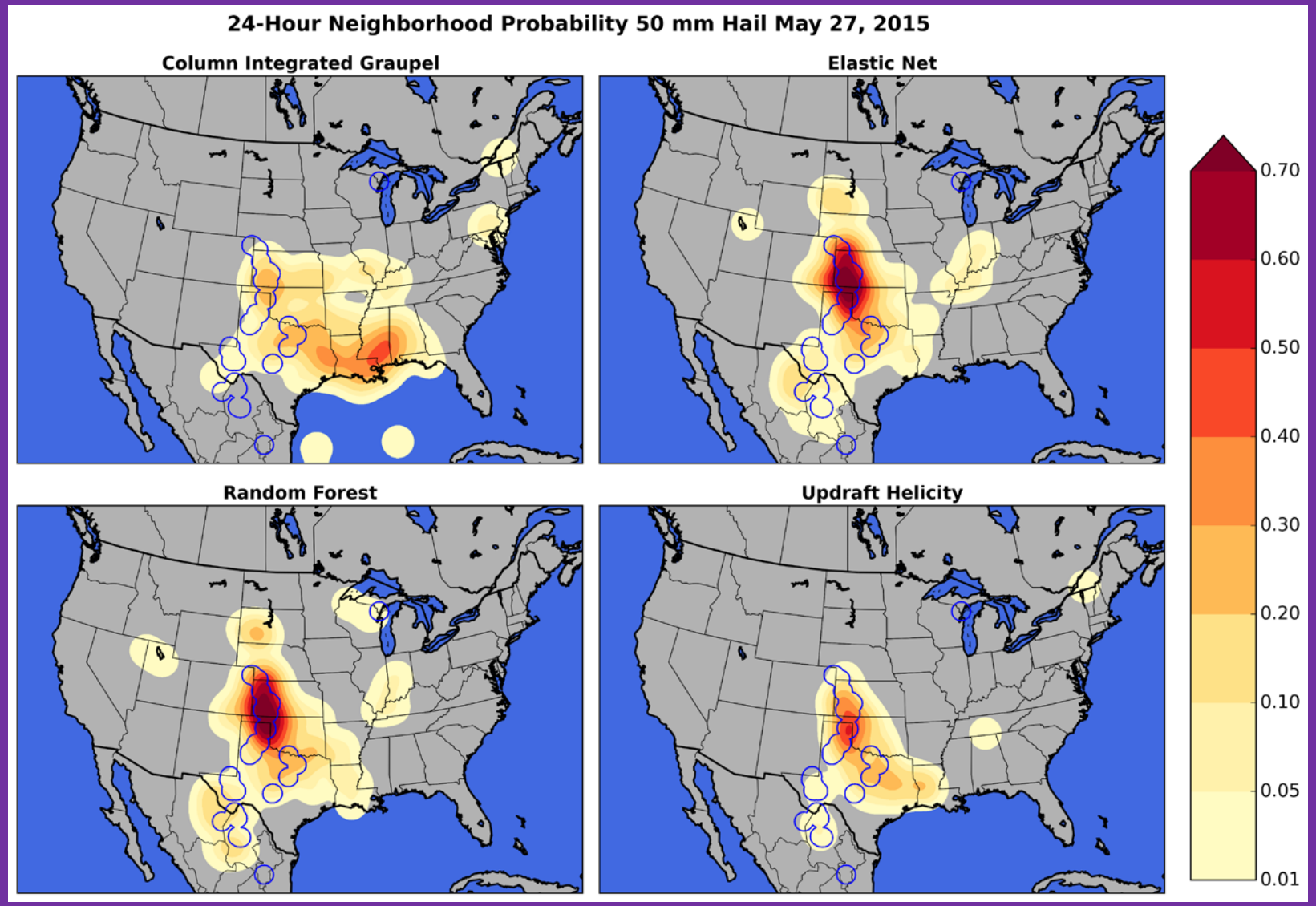


Hail Neighborhood Probability Performance



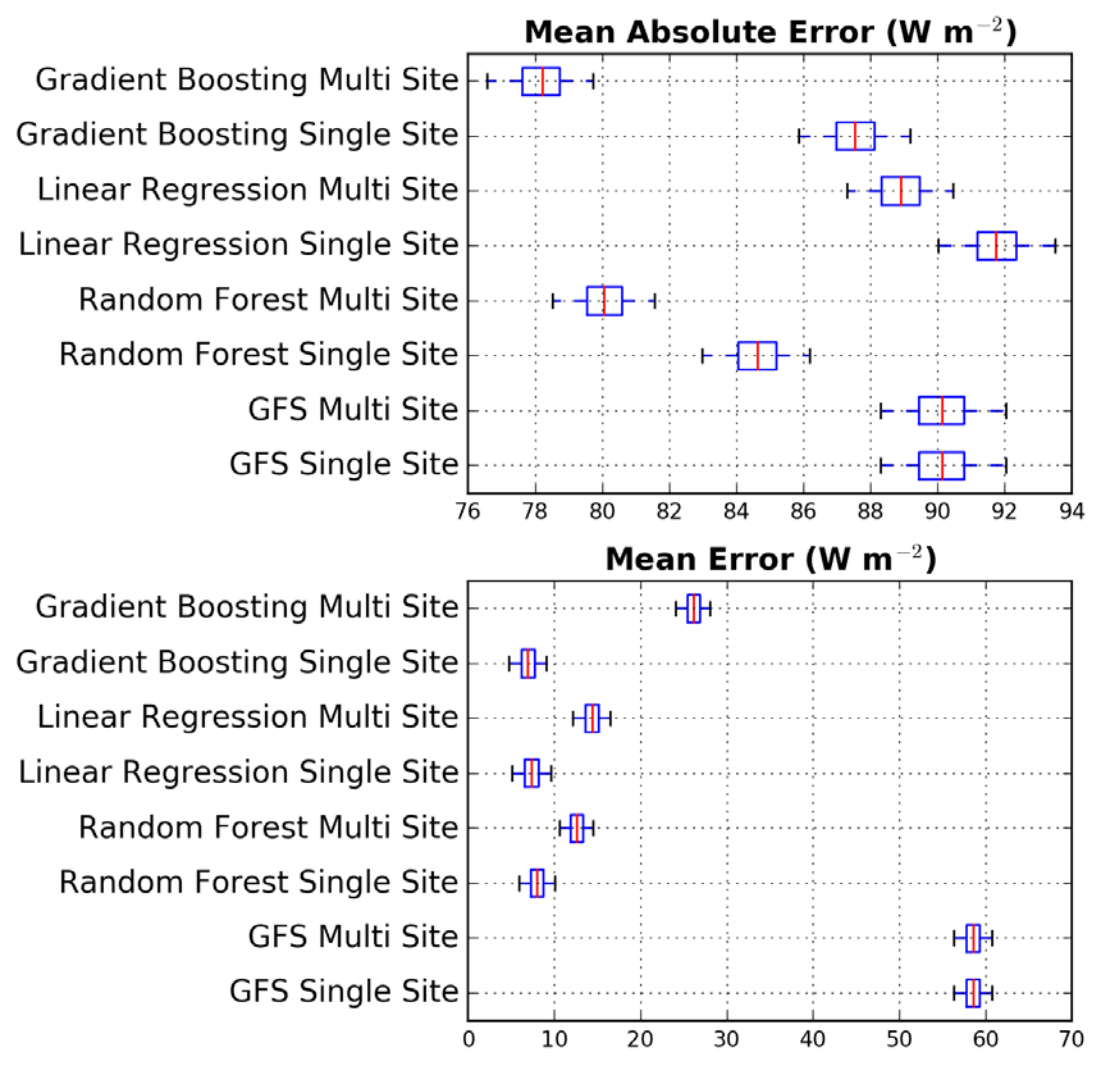


Hail Case Study: May 27, 2015



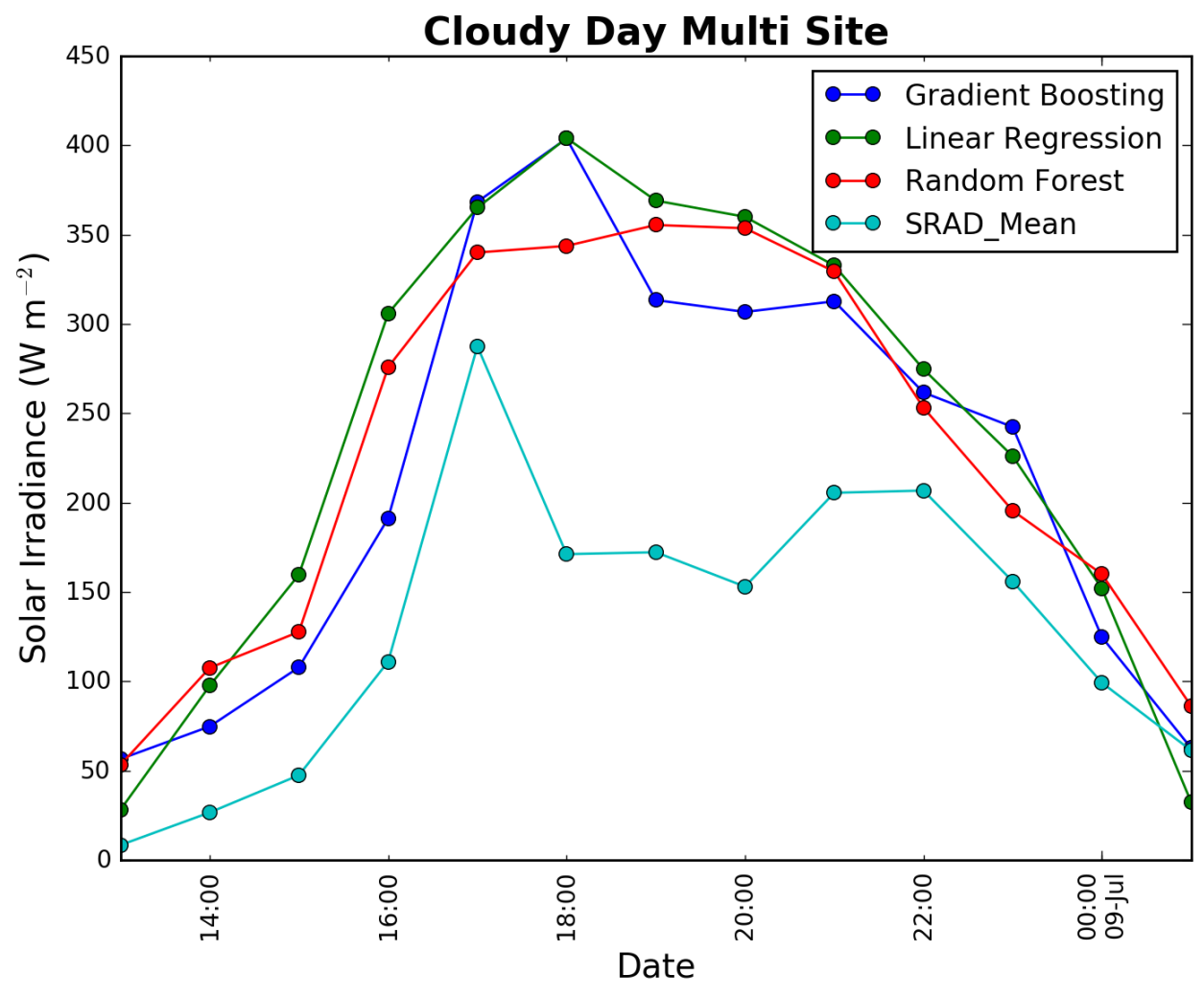
- Observed 50 mm or larger hail within 84 km is shown with the blue contours
- Column-integrated graupel probabilities highest in areas with non-hail-producing storms
- ML models match highest probabilities with largest hail
- Updraft helicity has smallest false alarm area

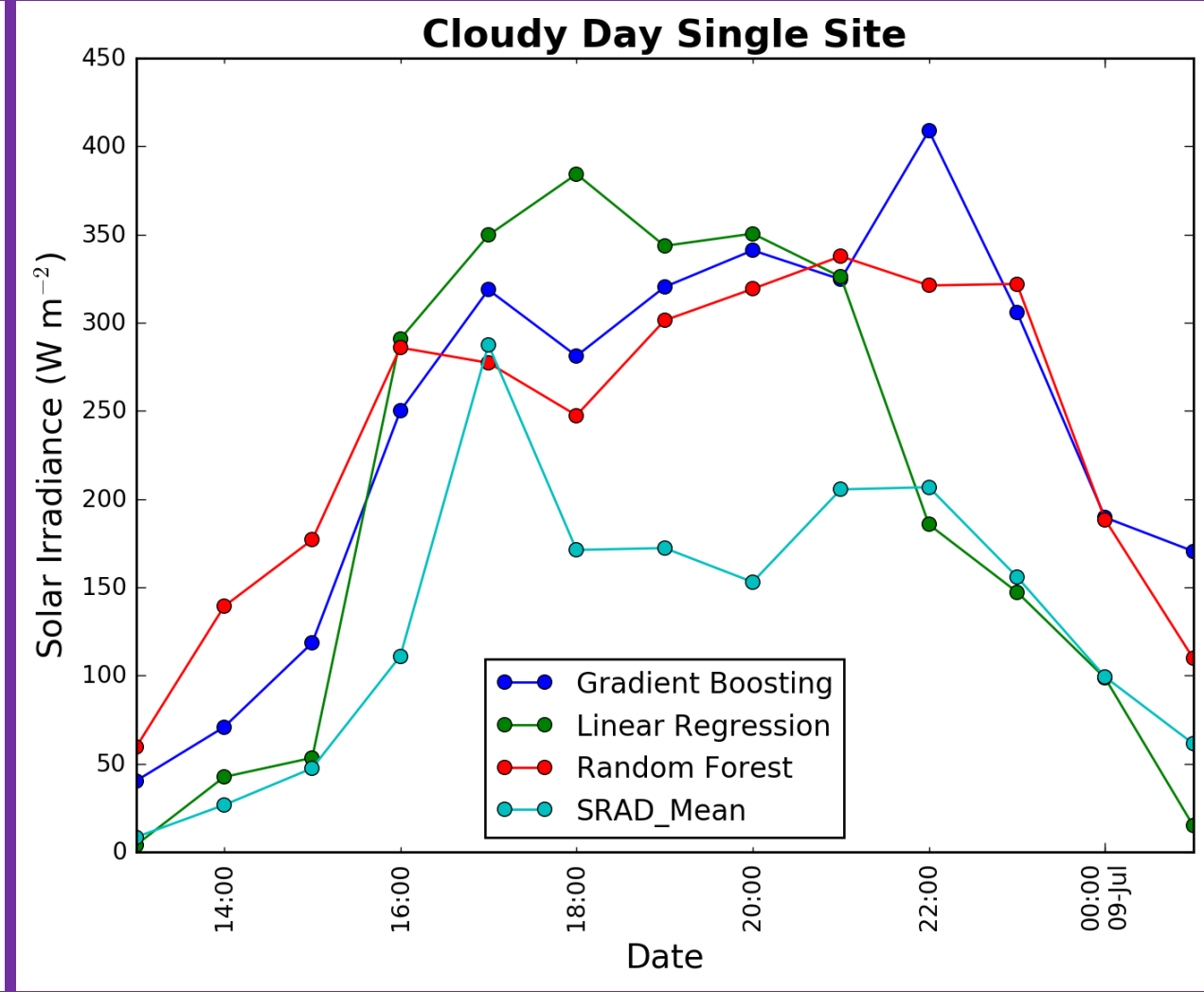
Solar Irradiance: Overall Errors by Model



- Whiskers indicate 95% bootstrap confidence interval
- Random Forest and Multi Site Gradient Boosting significantly outperform raw GFS
- Single site linear regression actually worse
- All statistical learning models reduce bias
- Every model still has slight positive bias

Solar Irradiance Forecast Example: Cloudy





Machine Learning Data Needs

- Complex ML models perform better with more data
- Large archives of model output/reforecasts
- Mix of ingredient variables
- Metadata
- Spatial and temporal neighborhood data
- More observations and analyses
- More complex ML models require more storage and computation for the model itself

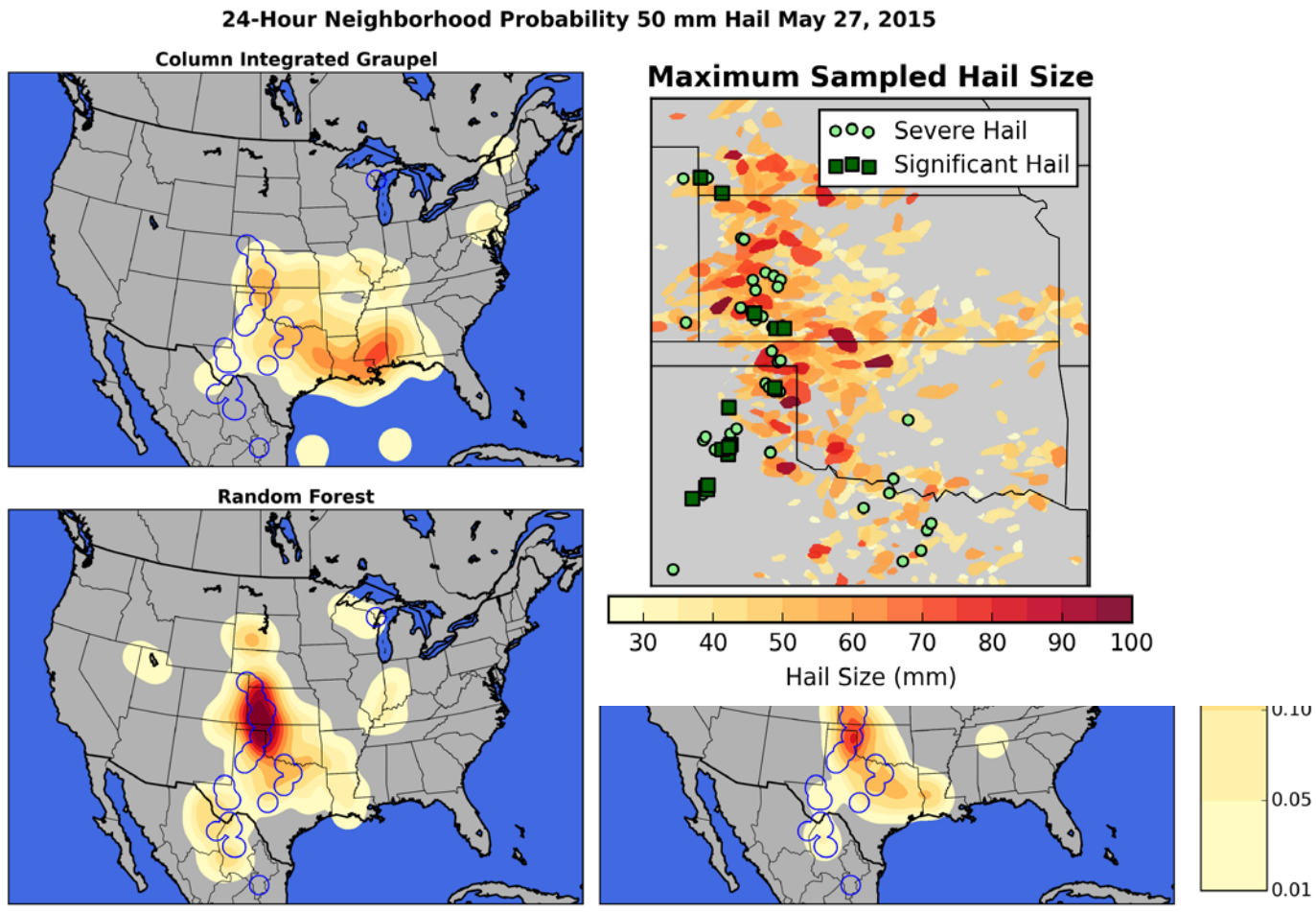
Machine Learning Software

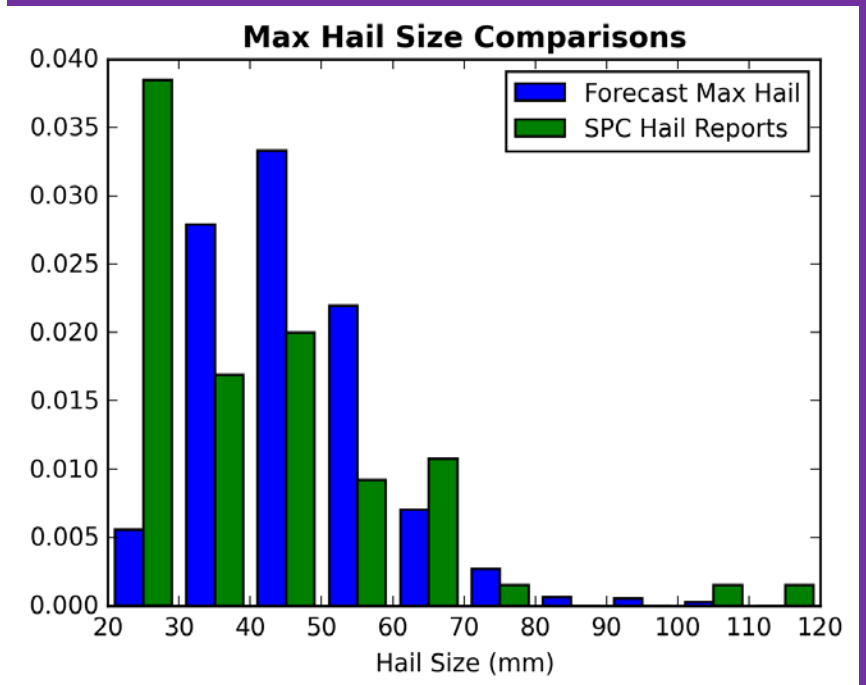
- Scikit-learn (Python + Cython)
  - Supports linear regression, decision tree ensembles, clustering, SVMs
- Xgboost (C++ with Python and R interfaces)
  - Random forest and gradient boosting
- Theano (Python with GPU interfaces)
  - Deep learning
- Various R libraries
- MATLAB

Interactive Post-Processing Products

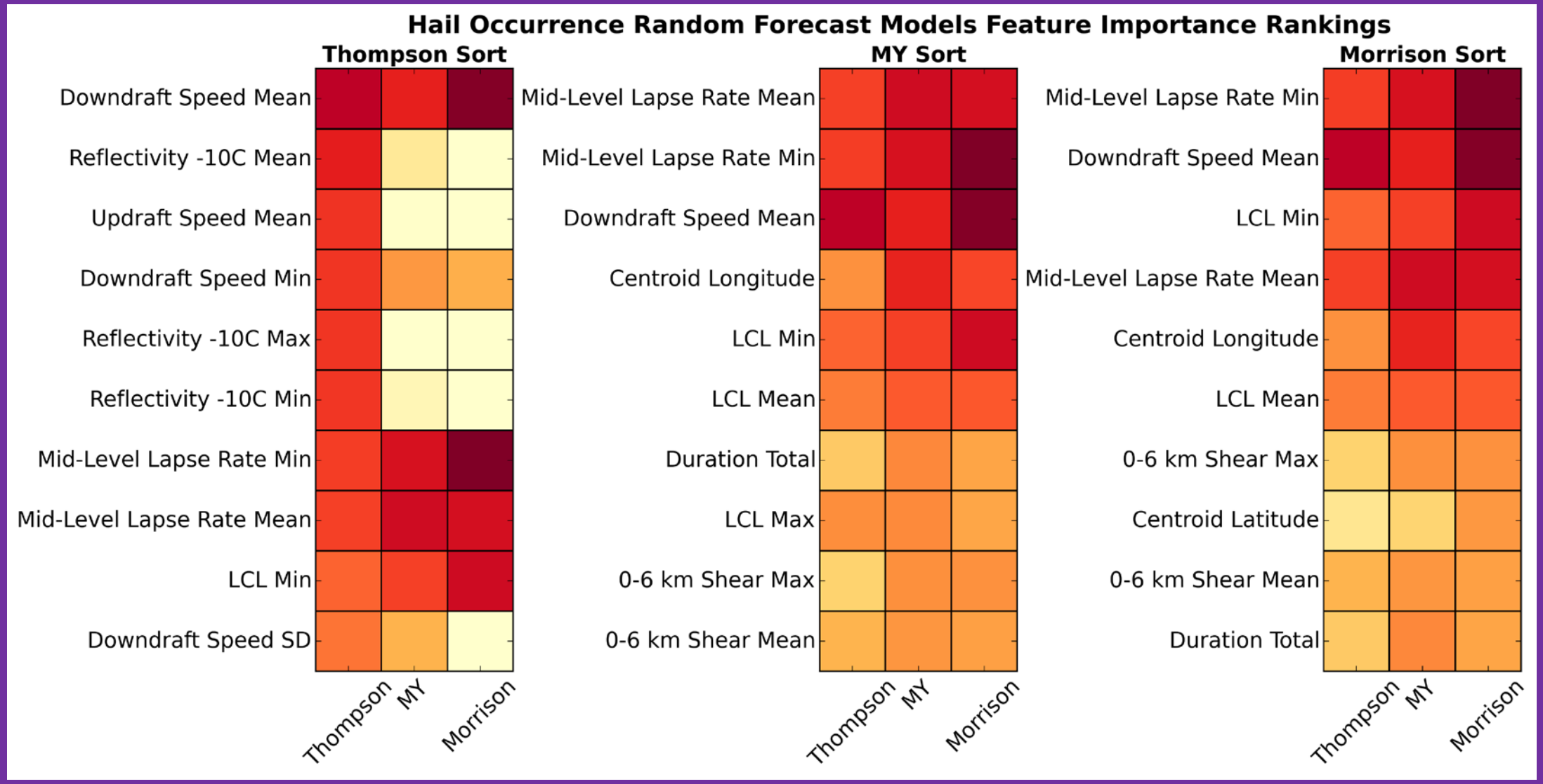
- Statistical models convey forecasts and uncertainty in many ways
  - Probabilities
  - Quantiles
  - CDF
  - Consensus statistic
- How do we best convey information from post-processing methods to forecasters and users?
  - Need to move beyond static maps and time series
  - Embrace more interactive visualization techniques

Hail Analytics Example

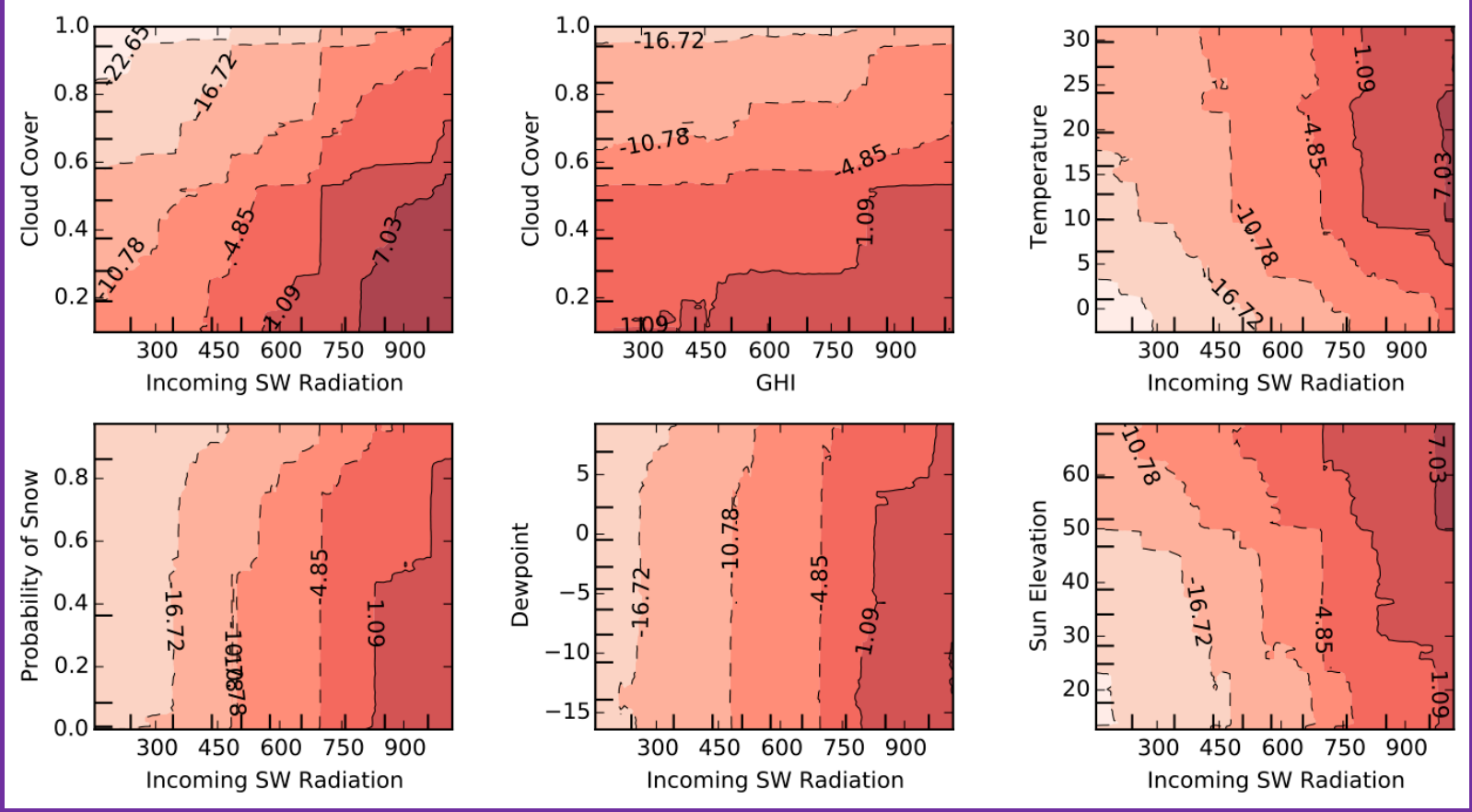




CAPS Ensemble Feature Importance Rankings



Partial Dependence Plots

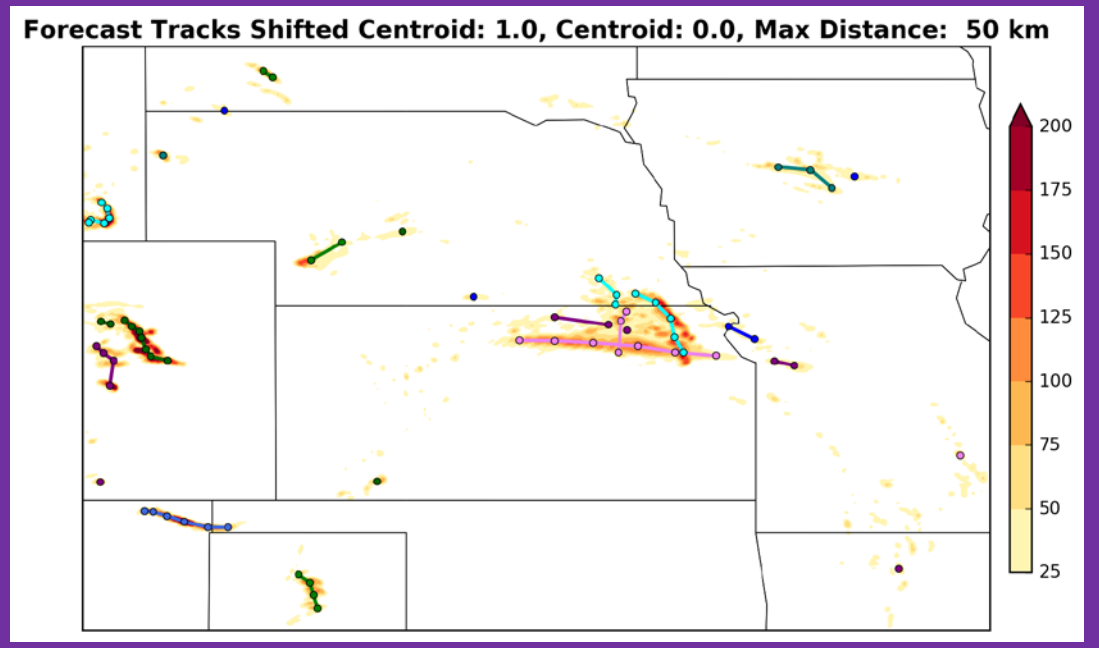


- Explore joint distributions of input variables and ML forecast values
- Highlights ranges of input values that have a strong impact on predictions

Personal Example: Hagelslag

- Python package for object-based post-processing and evaluation
- Supports parallel processing
- Storm identification and tracking
- Machine learning diagnostic integration
- Distributed forecast evaluation
- Github.com/djgagne/hagelslag

Seriouseats.com



Discussion Questions

- What kinds of weather objects/fields should be archived?
- What machine learning methods best fit community needs?
- How do we translate machine learning probabilities and diagnostics into useful tools for forecasters and the public?

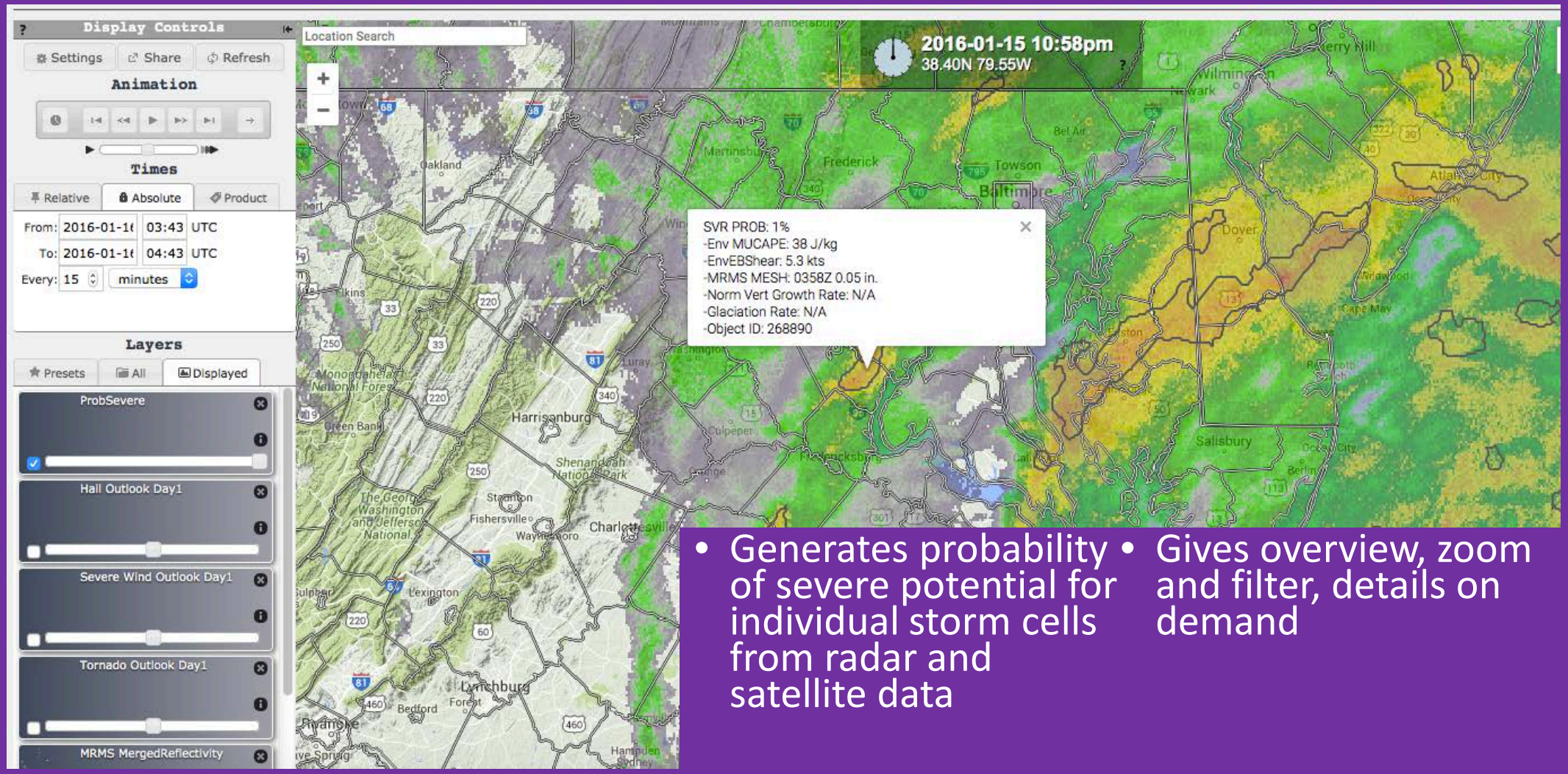
Email: djgagne@ou.edu
Twitter: @DJGagne

Visual Analytics Principles

- Analyze first
- Show the important
- Zoom, filter, and analyze further
- Details on Demand

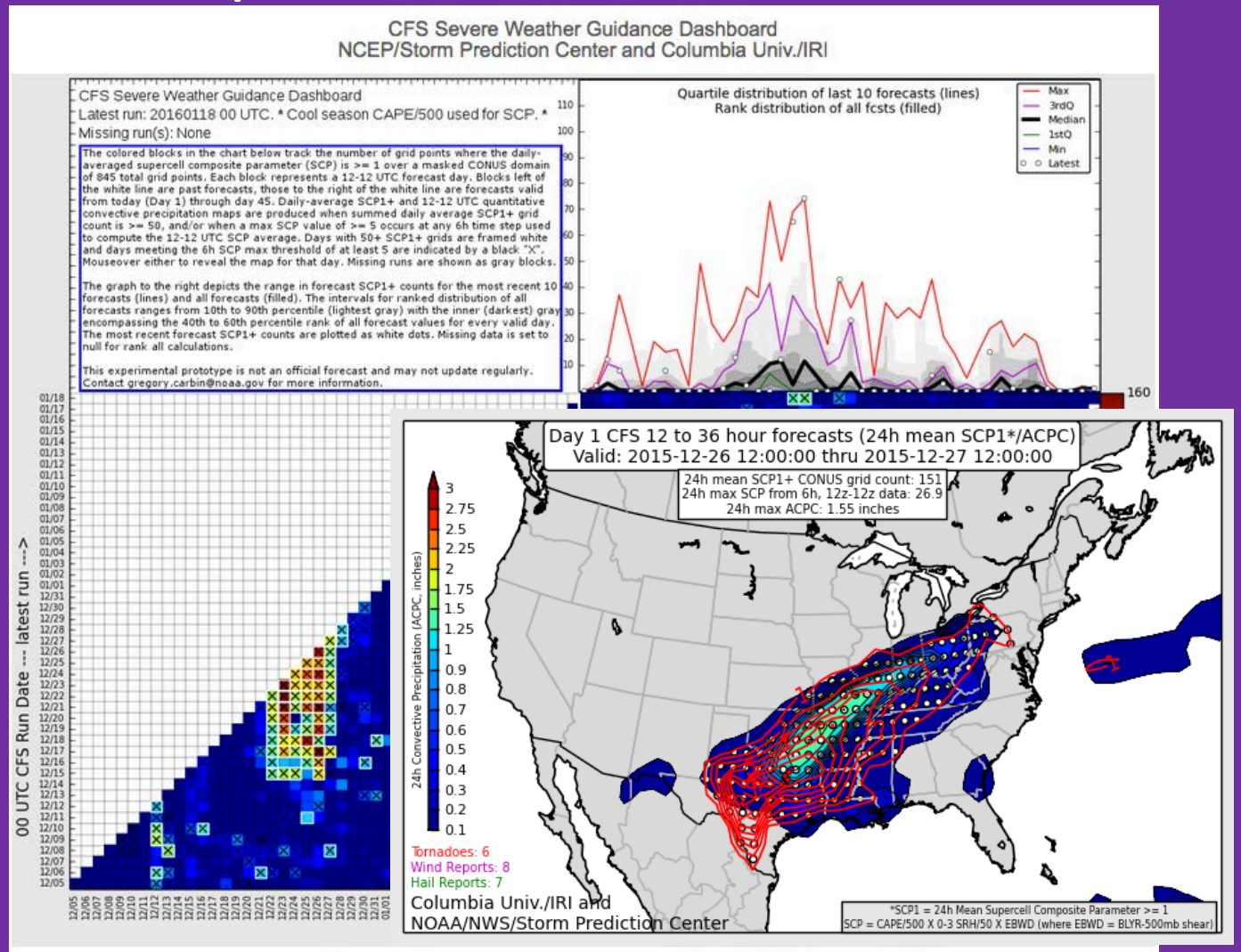
D. Keim, et al., 2008: Visual Analytics: Scope and Challenges,

Example: CIMSS Prob Severe



- Generates probability of severe potential for individual storm cells from radar and satellite data
- Gives overview, zoom and filter, details on demand

Example: SPC CFS Chiclet Charts



- Visualization for severe weather likelihood from CFS output
- Displays summary information from many model runs
- Users can mouse over individual runs to see details of forecast

Interactive Visualization Infrastructure

- Who should handle the burden of visualization computation?
  - Users (native/mobile/Javascript apps)
  - NOAA (computing clusters)
  - Cloud (Spin up Amazon/Microsoft/IBM/Google/etc. servers as needed)
- Data storage and transmission
  - Balance of local and remote data
- Latency
  - Effective visual analytics requires fast response to user queries