A Statistical Method for Coupled Observational and NWP-based 1–4 hour Forecasts of Convective Storm Initiation (a.k.a. *Multi sensor approach to storm detection*)

John R. Mecikalski[1], Christopher P. Jewett[2], Ujjwal Narayan[3], Xuang Li[4] and Todd Berendes[4]

[1]Atmospheric Science Department, University of Alabama in Huntsville
[2]Earth Systems Science Center, University of Alabama in Huntsville
[3]NOAA Earth System Science Interdisciplinary Center
[4]Information Technology and Systems Center, University of Alabama in Huntsville

NOAA Award NA14OAR4830171

Convective Working Group Meeting
Florence, Italy, 4-9 April 2016

Motivation

- Despite extensive research on nowcasting CI, 1-4 hour forecasting of CI is a challenge
- Previous studies have indicated that CI process is a combined interaction of the mesoscale and synoptic scale settings, mesoscale processes as well as land-surface processes and orography that dictate boundary layer formation and local convergent circulations and moisture distributions
- **Goal** is to develop a probabilistic 1-4 hour CI nowcast product (30 min update, ~5 km resolution gridded product) using machine learning methods

Approach

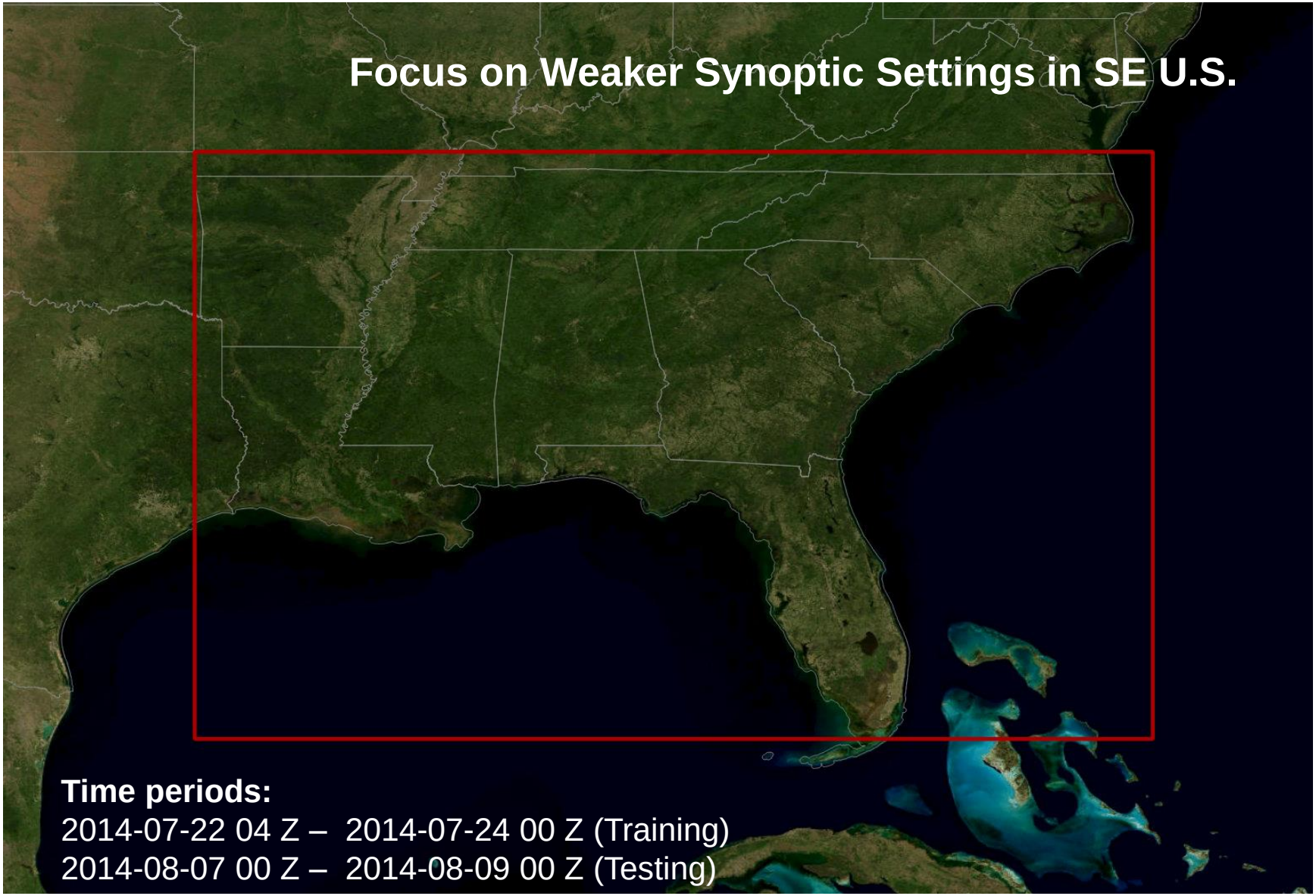
- **CI Event detection**: Define / detect CI as a ≥ 35 dBZ intensity radar echo at the surface or –10° C level at time ***t*** h. Delineate a 20 km radius Region Of Interest (ROI) centered on the CI event

- **Training database**: Characterize pre-thunderstorm atmospheric and land surface conditions for the ROI for times ***t, t-1, t-2, t-3 and t-4 h*** using NASA and NOAA satellite remote sensing fields as well as NOAA Rapid Update model fields

- **Machine learning**: Develop statistical models using machine learning techniques (e.g. Random Forest, Support Vector Machines, Logistic Regression) that may relate background conditions with occurrence / non-occurrence of CI within an ROI

- **Generate CI predictions** for 1, 2, 3 and 4 hours into the future

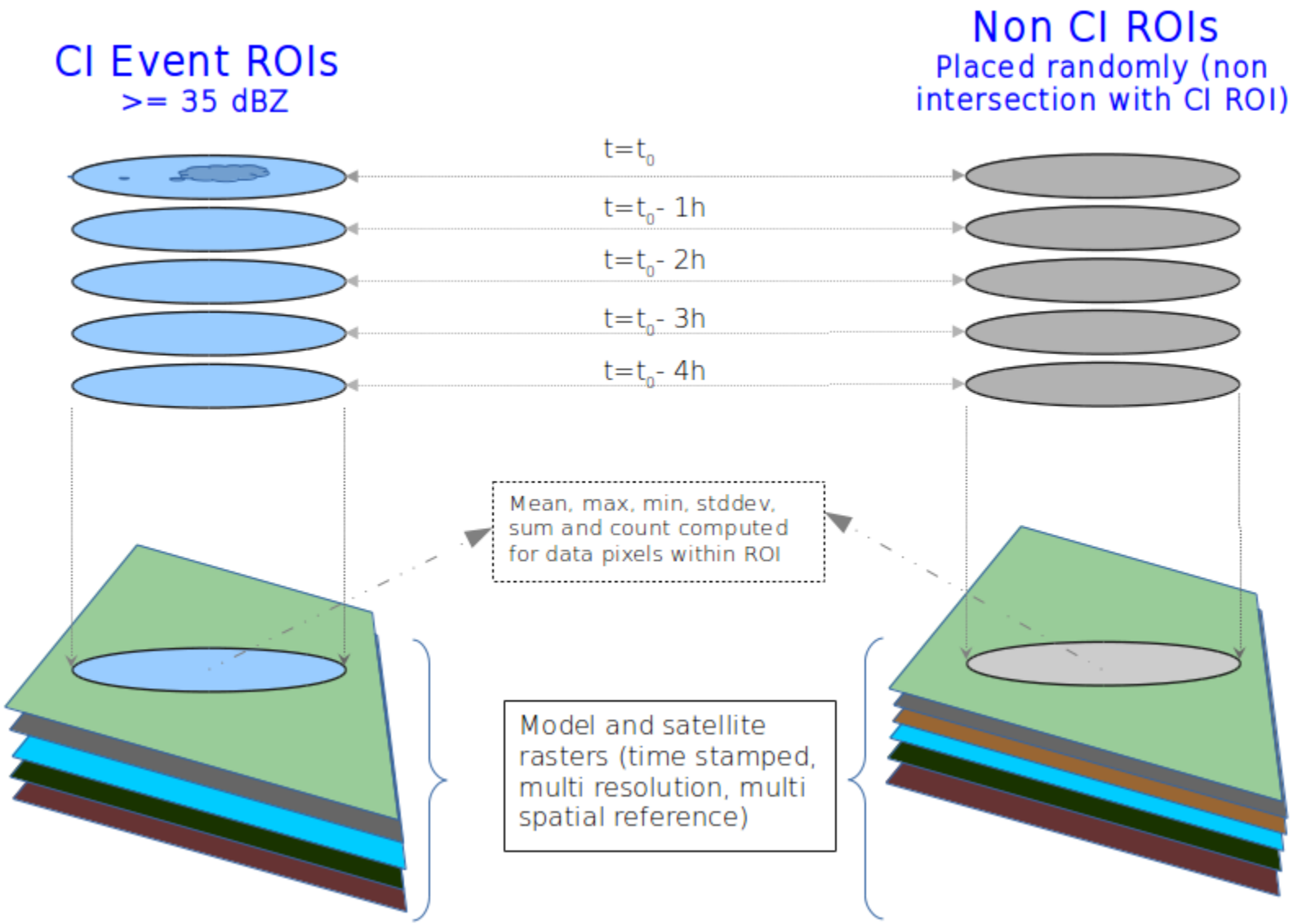
Study Domain



Data Sources

Data Source	Variables
MRMS Radar Reflectivity	– Rain/No-Rain regions – Detection of first 35 dBZ occurrences – For “truth” database of CI events
GOES Convective Cloud Mask	– CloudType (Cumulus & Towering Cumulus)
Topography (GTOPO30)	– Elevation – Slope and Aspect
NOAA **Antecedent Precipitation**	– Where did it rain yesterday? – How much rain has occurred regionally?
MODIS Vegetation (MOD13Q1)	– NDVI for active vegetation mapping – Provide qualification of evapotranspiration
MODIS Land Cover Type (MCD12Q1)	– Waterbody – Variability in land cover type
RAP Model Fields (Analysis)	– CAPE, HTFL (HGT, RH), CIN, UGRD, VGRD, WINDCONVERGENCE, DEWPOINT
LIS (NOAH) Soil Moisture / Temperature	– Soil moisture regulates evapotranspiration

Definition of Regions of Influence (ROI)

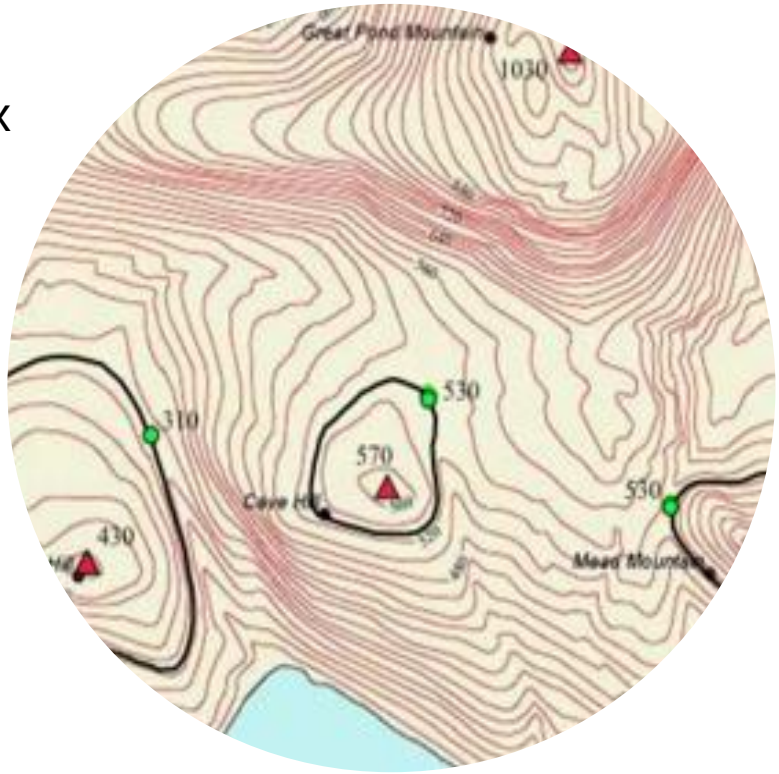


Mecikalski et al./CWG Meeting 2016

ROI & Feature Selection

Regional Variables

- Topography slope
- Topography aspect
- Wind–Topography angle
- Convergence in ROI – max
- Cumulus cloud type
- 1, 2 and 3 hour changes in cumulus cloud type
- Cumulus cloud coverage
- 1, 2 and 3 hour changes in cumulus cloud cover
- Vegetation
- Land cover type



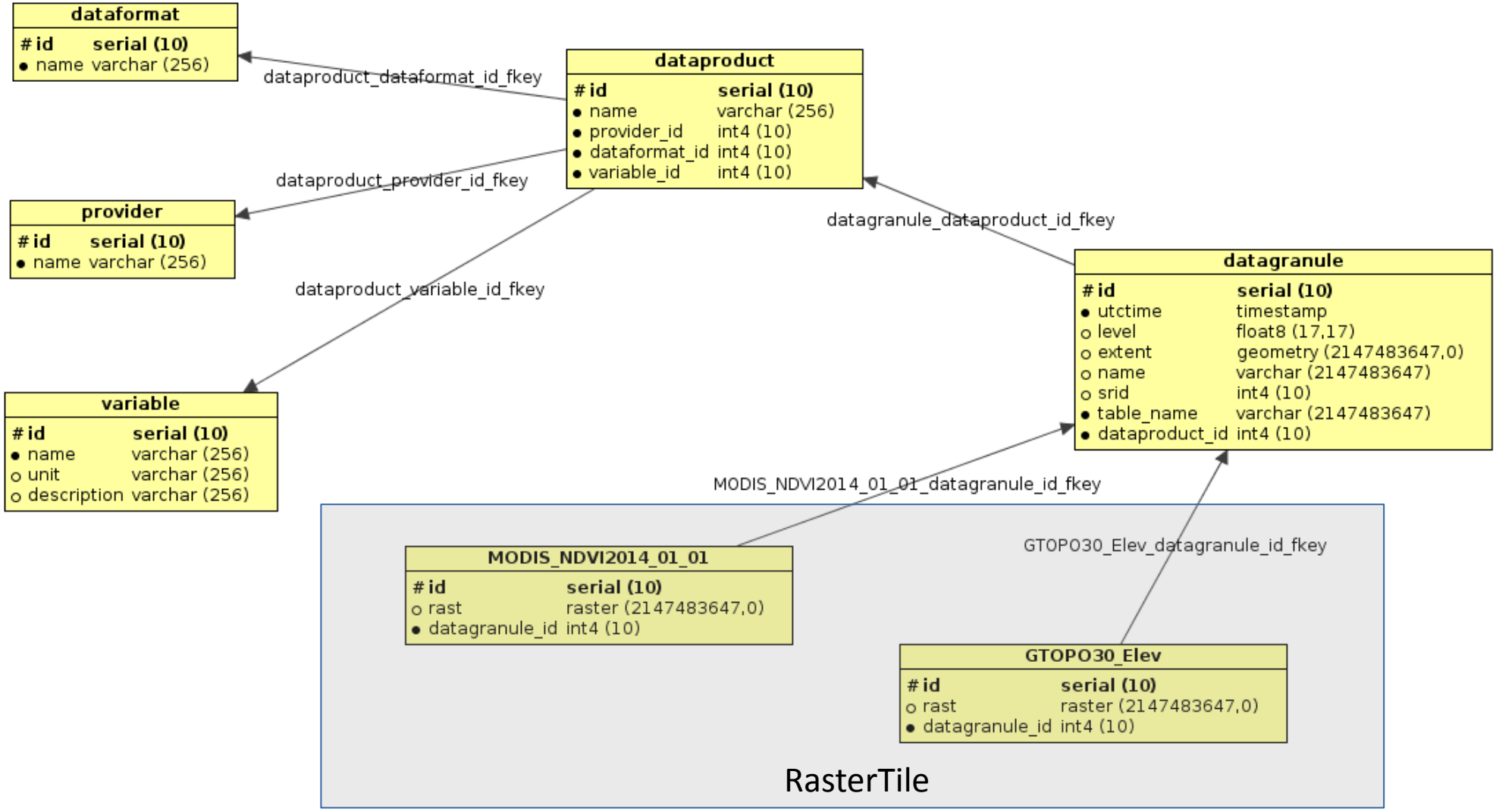
RAP Model Variables

- 0–3 hour minimum CAPE
- 0–3 hour mean CAPE
- 0–3 hour maximum CAPE
- 0–3 hour minimum CIN
- 0–3 hour mean CIN
- 0–3 hour maximum CIN
- Hourly changes in CAPE
- Hourly changes in CIN
- Model reflectivity/precip
- Hourly changes in precip
- Dew points & trends
- Temperature & trends
- Soil temperature
- Soil moisture
- Freezing level
- Convective temperature

Toward trying to capture all key variables that could be important to CI in the 0-3 hour timeframe, a total of 750+ variables were initially evaluated, which decreased to ~234, and then to 59.

PostGIS Geodatabase Schema

Data sources are in different formats, spatial resolution and coordinate systems (e.g., MODIS 500m, RAP 13km, GTOPO 1km), PostGIS allows homogenizing of datasets into a common schema to take advantage of relational database, geospatial and raster querying.



Mecikalski et al./CWG Meeting 2016

CI & non-CI Event ROI Generation

- MRMS data granules are searched for ≥35 dBZ intensity radar echo at the –10° C level and 20 km radius circle is drawn centered on the precipitating region. This is a CI event ROI (*type 1*) at time **t**.

- **For each CI event ROI**, additional co-located ROIs are inserted at time **t-1h, t-2h, t-3h and t-4h** (*for extracting prior time data*).

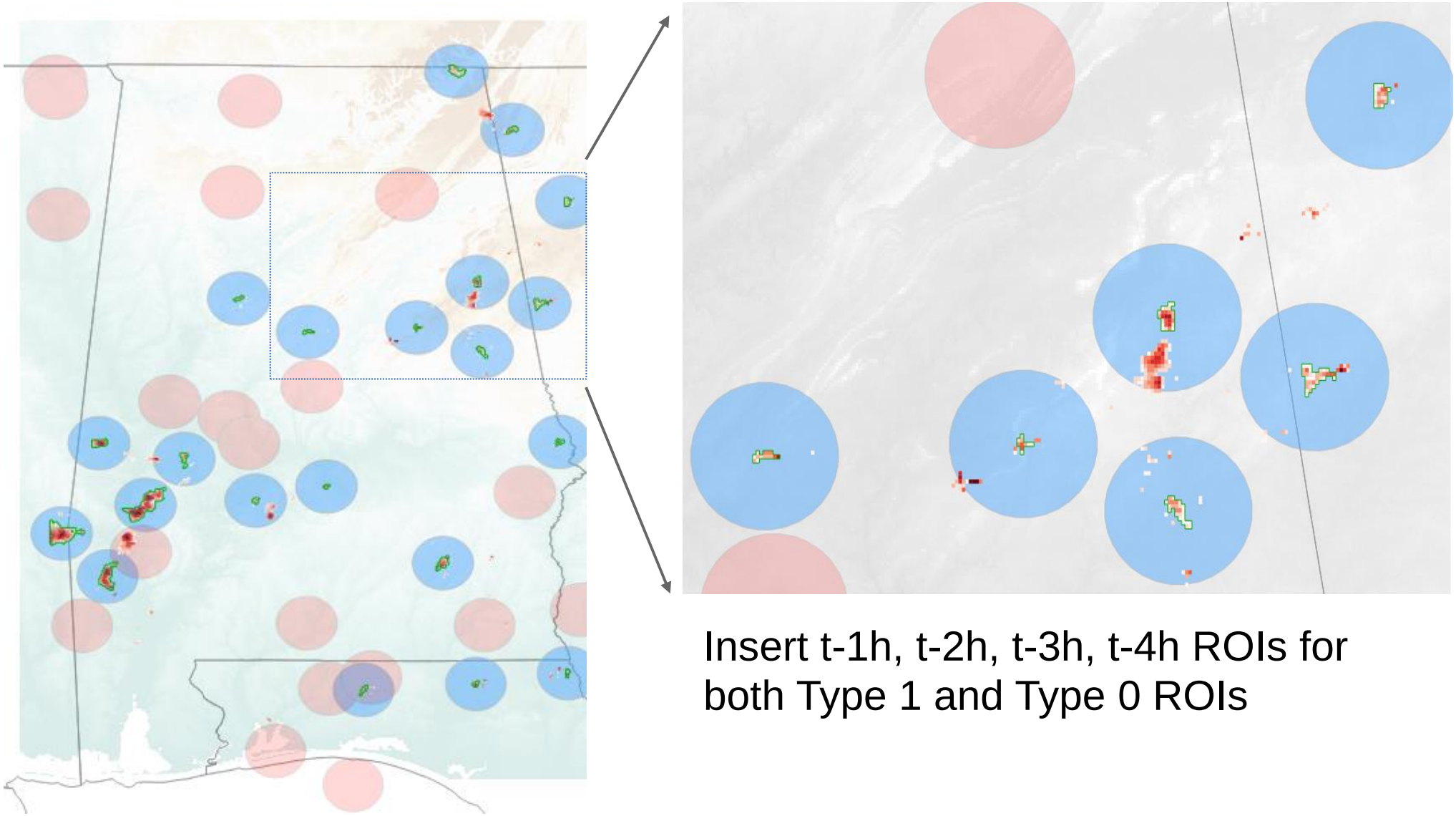
- CI event ROIs cannot intersect each other within 4 hours.

- Equal number of **randomly located non-CI event ROIs** (*type 0*) are also inserted.

- CI and non-CI event ROIs do not intersect each other.

CI (blue) and non-CI ROI Selection (red)

Time: 2014-07-22 22:58:00 CT

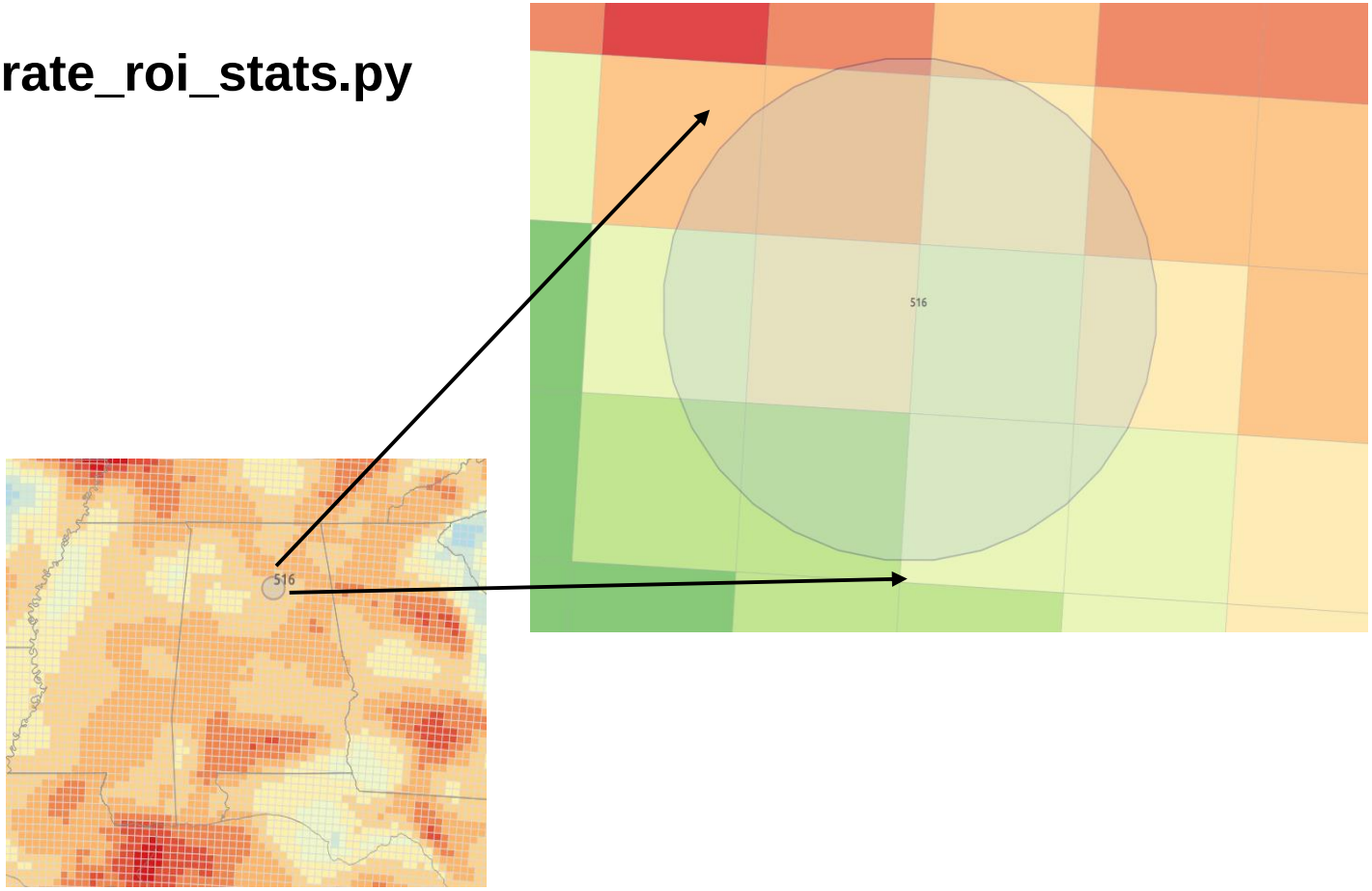


Insert t-1h, t-2h, t-3h, t-4h ROIs for both Type 1 and Type 0 ROIs

ROI Statistics

All CI and non-CI event ROIs are intersected with meteorological variable raster datasets and statistics (count, sum, mean, max, min) for raster pixel values computed.

$ python ci/generate_roi_stats.py



Machine Learning Setup

- Two months (July and August) of CI feature data are generated and used in the study.

- For the goal of predicting CI events, we use data at t-1, t-2, t-3 and t-4 for **1 hour prediction**; data at t-2, t-3 and t-4 for **2 hour prediction**. Time trends are also used, for example, for 2hr prediction we calculate change in model fields between t-2 and t-3 (1 hr diff) and t-2 and t-4 (2hr diff)

- Begin with 750+ variable database, and perform initial variable reduction

- Full training database consists of 162 features expanded to **234** features when differences are included. Data is separated for July and August with ~**3000** CI/non-CI samples each, we used July for training learning models and August data for testing (vice-versa led to similar results)

- Using **Random Forest** model we examined CI vs non CI feature classification accuracy when all available (234) features are used. We obtained an accuracy of ~69% which is indicative of the maximum separability we can obtain between CI and non CI events for our database.

Feature Selection

- Feature attribute selection using **information-gain algorithm** on combined (1 and 2 hour) feature set which ranks each of 234 features in decreasing order of importance

- Using information-gain algorithm, the identified top 20 features are:
  - aspect_count [number of topography–wind angles per ROI]
  - TCUM_CLOUD_sum_1hr_dif [1-hr change amount of GOES-observed cumulus clouds]
  - TCUM_CLOUD_count [amount of GOES-observed cumulus clouds]
  - cape2_min [minimum CAPE 2 hours ago]
  - cape0_min [minimum CAPE present time]
  - cape3_mean [mean CAPE 3 hours ago]
  - cape3_max [maximum CAPE 3 hours ago]
  - cape1_min [minimum CAPE 1 hour ago]
  - cape2_max [maximum CAPE 2 hours ago]
  - cape1_mean [mean CAPE 1 hour ago]
  - cape0_mean [mean CAPE present time]
  - cape0_max [maximum CAPE present time]
  - cape2_mean [mean CAPE 2 hours ago]
  - cape1_max [maximum CAPE 1 hour ago]
  - cape3_min [minimum CAPE 3 hours ago
  - TCUM_CLOUD_mean [mean GOES-observed cumulus cloud type]
  - aspect_sum [sumof topography–wind angles per ROI]
  - cin3_min [minimum CIN 3 hours ago]
  - cin2_min [minimum CIN 2 hours ago]
  - RAP_REFL_SUM_1hr_dif [RAP model reflectivity in ROI]

Random Forest Field Importance: Alabama Domain

CI 0-1 hour

Variable	Importance
CLOUDTYPE_count	0.07779
CLOUDTYPE_sum	0.04723
VGRD_min	0.01879
ASPECT_stddev	0.01783
HTFL_RH_sum	0.01622
HTFL_RH_stddev	0.01591
HTFL_HGT_min	0.01591
VGRD_sum	0.0155
LCT_count	0.0151
VGRD_max	0.0149
ASPECT_mean	0.01464
HTFL_HGT_mean	0.01452
HTFL_HGT_max	0.01434
VGRD_stddev	0.01419
VGRD_mean	0.01406
SLOPE_sum	0.01358
CLOUDTYPE_max	0.01312
ASPECT_max	0.01283
LCT_sum	0.01253

t hour: Variable importance when Random Forest algorithm used for Type 0 and Type 1 ROIs. 126 out of 173 events (72%) were predicted correctly while testing the classifier

t-2 hours: Variable importance when Random Forest algorithm used for Type 0 and Type -2 ROIs. 130 out of 173 events (75%) were predicted while testing the classifier

CI 2-3 hours

Variable	Importance
CAPE0_stddev	0.0341
HTFL_HGT_min	0.02489
HTFL_HGT_max	0.02469
HTFL_HGT_mean	0.0233
HTFL_RH_sum	0.01867
LCT_count	0.0183
HTFL_RH_max	0.01763
CIN3_sum	0.0166
HTFL_RH_min	0.01619
SLOPE_mean	0.01597
CAPE2_stddev	0.01583
CAPE2_mean	0.01539
LCT_mean	0.01441
CIN0_sum	0.01402
CIN0_max	0.01364
HTFL_HGT_stddev	0.01353
HTFL_RH_stddev	0.01352
CAPE2_sum	0.01332
HTFL_RH_mean	0.0132

Feature Selection

- Using **Random Forest** classifier with 10 fold cross validation, 12 experiments were run by selecting top 1, 3, 10, 20, 30, 50, 60, 90, 120, 150, 180 and 210 sub-features respectively which led to overall accuracies of 58.62%, 60.03%, 62.22%, 64.78%, 65.21%, 67.28%, 68.19%, 68.54%, 68.50%, **68.96%**, 68.95%, 68.61%

- Based on above results we determine optimal number of features is between 50 to 60 and **optimum accuracy is around 69%**.

- And... After Random Forest experiments, **further reduction and selection in features to be used could be done**.

- Using information gain algorithm results and some domain knowledge we manually selected **59 out of 234 features** for training algorithm

Training (Evaluation of Algorithms)

- **7 classification algorithms are used** – **BayesNet** (BN), **Naive Bayes** (NB), **Logical Model Trees** (LMT), **Logistic Regression** (LR), **Multilayer Perceptron** (MP), **Random Forest** (RF), **Support Vector Machines** (SVM). *Waikato Environment for Knowledge Analysis* (WEKA) package was used for machine learning analysis

- **August data (59 features) was used to train each model** and **tested on July data**. For classifiers that allow parameter optimization (Random Forest, SVM, etc.) a range of parameter options are tested and best models were selected

	1hr Training	1hr Testing	2hr Training	2hr Testing
BN	58.6%	62.9%	54.6%	58.0%
NB	57.6%	64.7%	54.9%	58.7%
LMT	61.4%	65.0%	55.2%	56.8%
LR	59.7%	66.8%	55.8%	58.5%
MP	59.8%	64.7%	55.9%	59.5%
RF	**63.4%**	**66.8%**	**58.0%**	**59.1%**
SMO	64.0%	66.4%	59.3%	59.1%

Prediction using Ensemble of Classifiers

For this component of the study, several **ensembles of classifies** were evaluated as a means to predict CI 1-3 hours into the future.

CI event probability is a weighted average of 7 individual classifiers with weight value being the overall accuracy for classifier.

In this way, probability of 100% for CI occurrence is assigned to a sample if all seven classifiers predict the occurrence of a CI event and 0 if none predicts CI

1 hour Prediction

	Predict non-CI	Predict CI
Truth non-CI	1074	527
Truth CI	487	1016

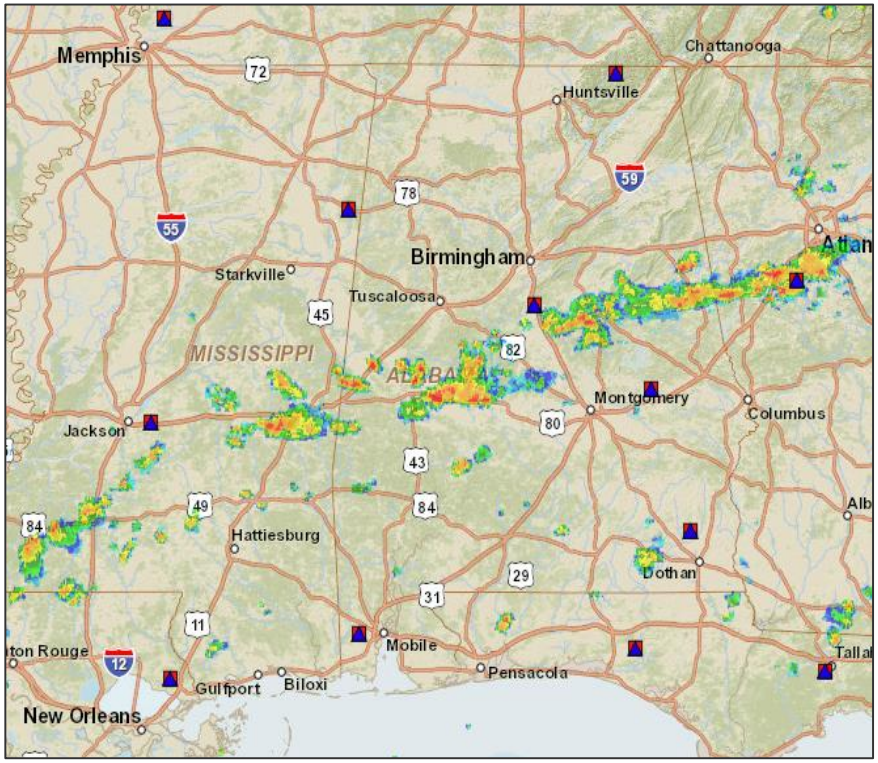
Overall Accuracy: 67.33%

2 hour Prediction

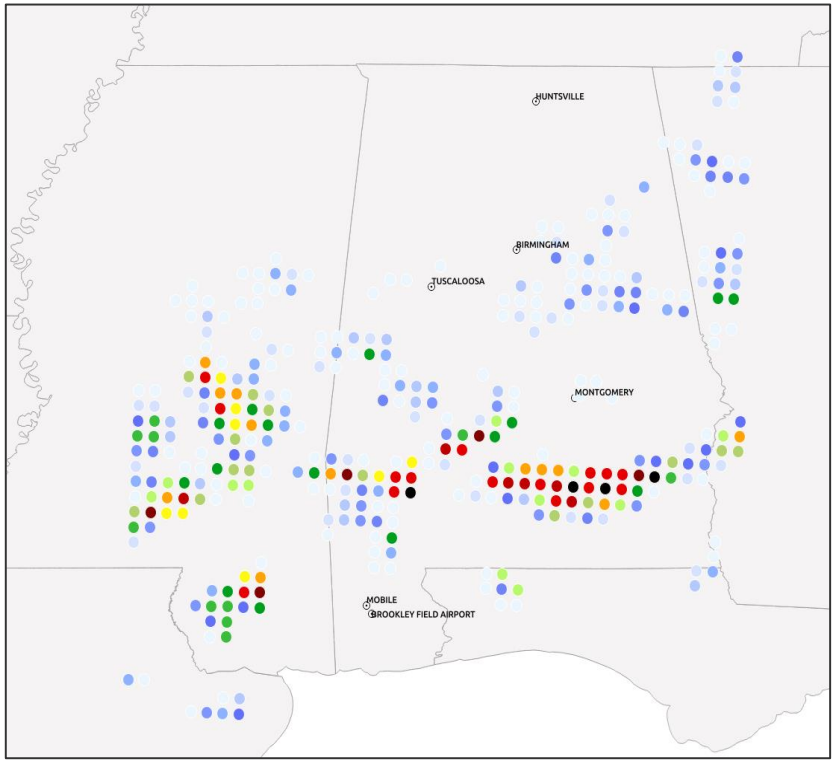
	Predict non-CI	Predict CI
Truth non-CI	1134	466
Truth CI	726	777

Overall Accuracy: 61.59%

CI Probability for 2014-07-24 2300 UTC



NEXRAD Radar composite

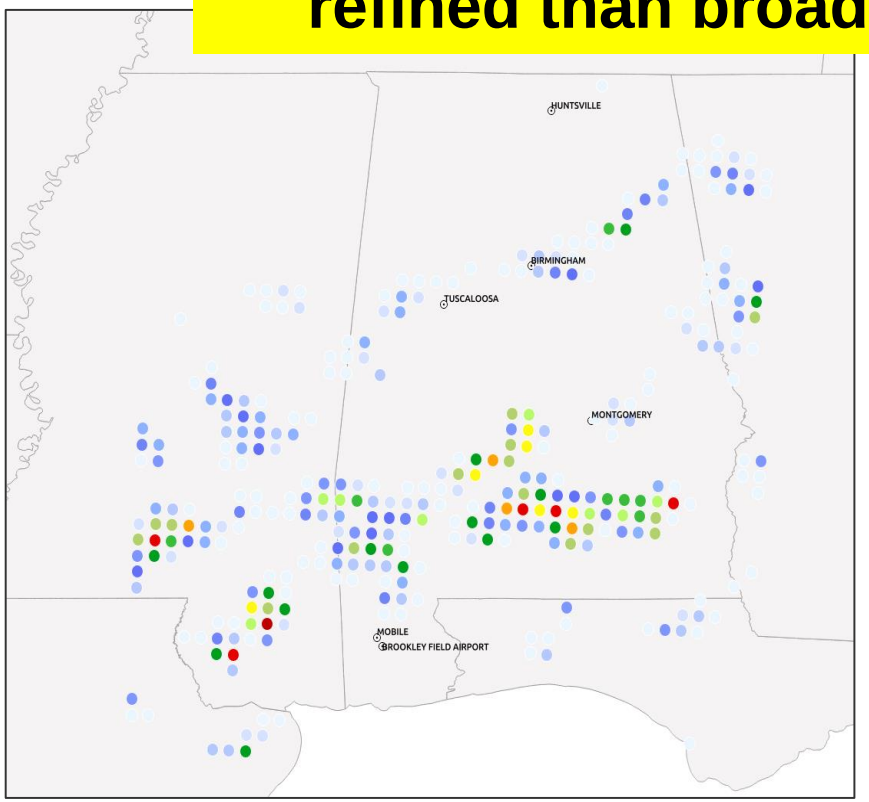


0 hour (CI nowcast)

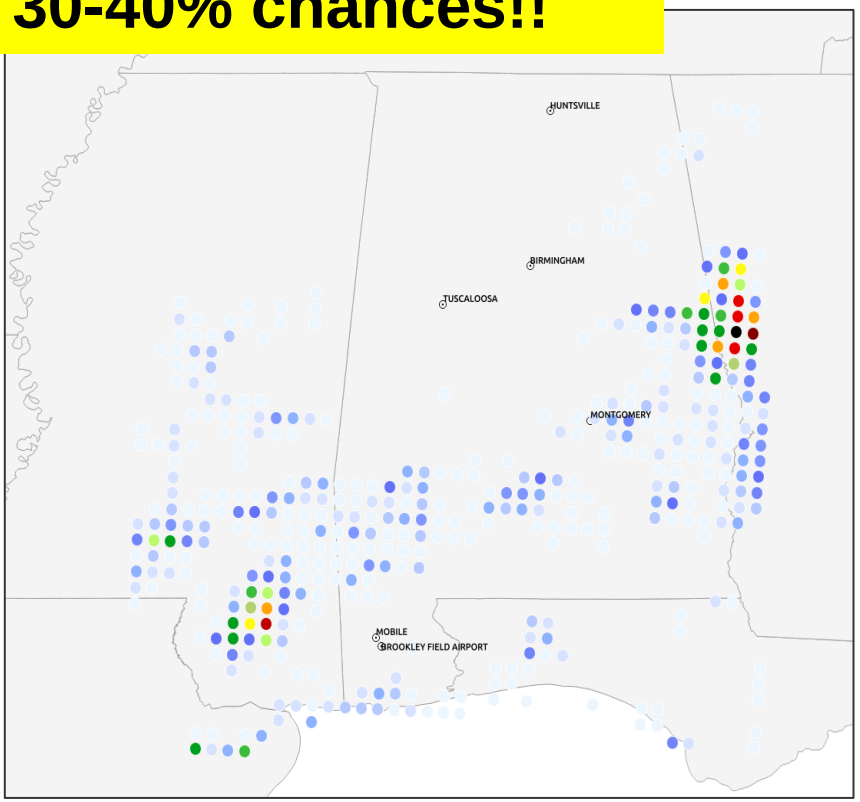
- 50.00 - 51.50
- 51.50 - 53.00
- 53.00 - 54.50
- 54.50 - 56.00
- 56.00 - 57.50
- 57.50 - 59.00
- 59.00 - 60.50
- 60.50 - 62.00
- 62.00 - 63.50
- 63.50 - 65.00
- 65.00 - 66.50
- 66.50 - 68.00
- 68.00 - 69.50
- 69.50 - 71.00
- 71.00 - 72.50
- 72.50 - 74.00
- 74.00 - 75.50
- 75.50 - 77.00
- 77.00 - 78.50
- 78.50 - 80.00

CI Probability for 2014-07-24 2300 UTC

These CI nowcasts are significantly more refined than broad 30-40% chances!!



1 hour CI prediction



2 hour CI prediction

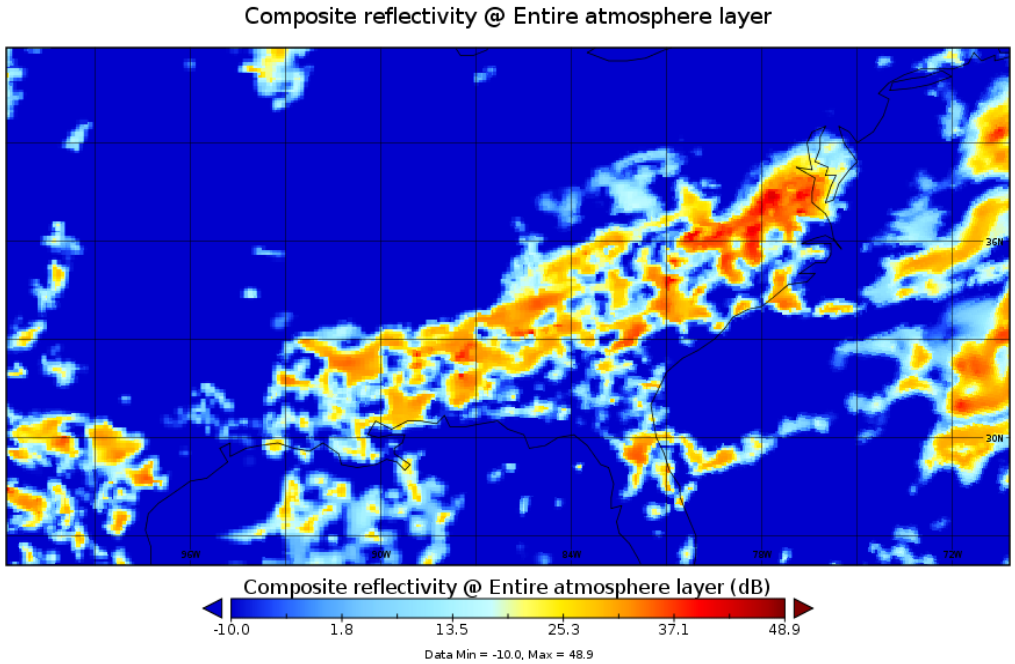
- 50.00 - 51.50
- 51.50 - 53.00
- 53.00 - 54.50
- 54.50 - 56.00
- 56.00 - 57.50
- 57.50 - 59.00
- 59.00 - 60.50
- 60.50 - 62.00
- 62.00 - 63.50
- 63.50 - 65.00
- 65.00 - 66.50
- 66.50 - 68.00
- 68.00 - 69.50
- 69.50 - 71.00
- 71.00 - 72.50
- 72.50 - 74.00
- 74.00 - 75.50
- 75.50 - 77.00
- 77.00 - 78.50
- 78.50 - 80.00

CI probabilities <50% not shown

Mecikalski et al./CWG Meeting 2016

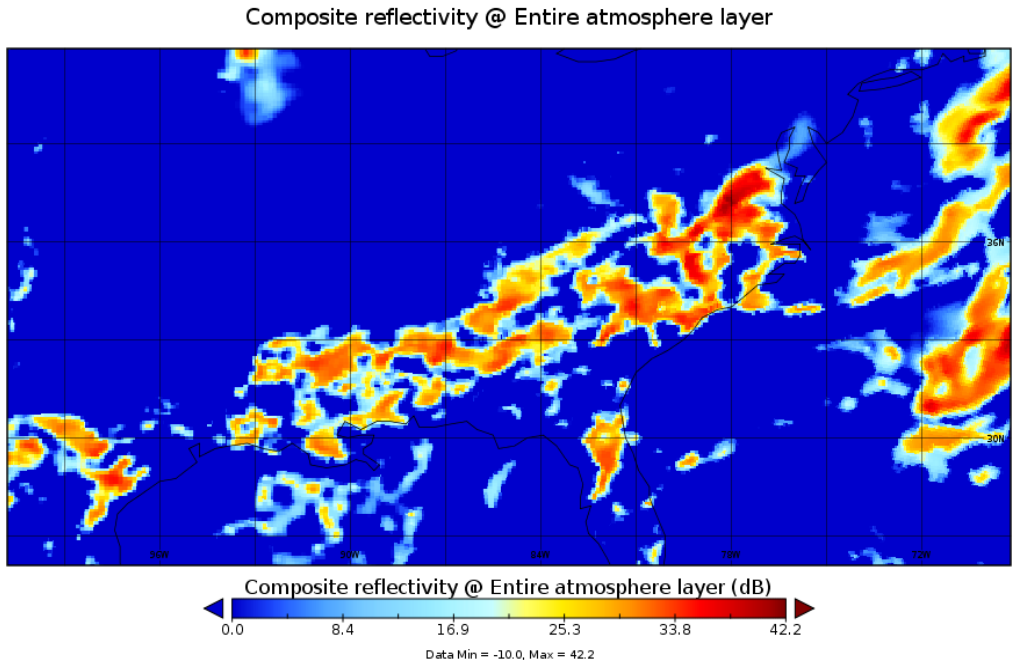
RAP Model Composite Reflectivity

Rainfall at 2014-07-24 2300 UTC



For Comparison: > 34 dBZ masked to show CI only

2 hour forecast at 2014-07-24 2100 UTC



Summary & Ongoing Work

1. Further evaluate 3-4 hour predictability for currently considered variables.

2. Evaluate improvement in CI prediction relative to RAP model by comparison with forecast simulated radar reflectivity.

1. Further expand training database.

2. Implement parameter tuning for machine learning models.

3. Eventually form a gridded product (30 min, 5 km resolution) that operates on ROIs in real-time.