Data Fusion: A Machine Learning Tool for Forecasting Winter Mixed Precipitation Events

Brian Filipiak, Kristen Corbosiero, Andrea Lang, Ross Lazear, and Nick Bassill

University at Albany, SUNY Albany, NY

WPC HMT Winter Weather Experiment Seminar Series : 12/07/2021

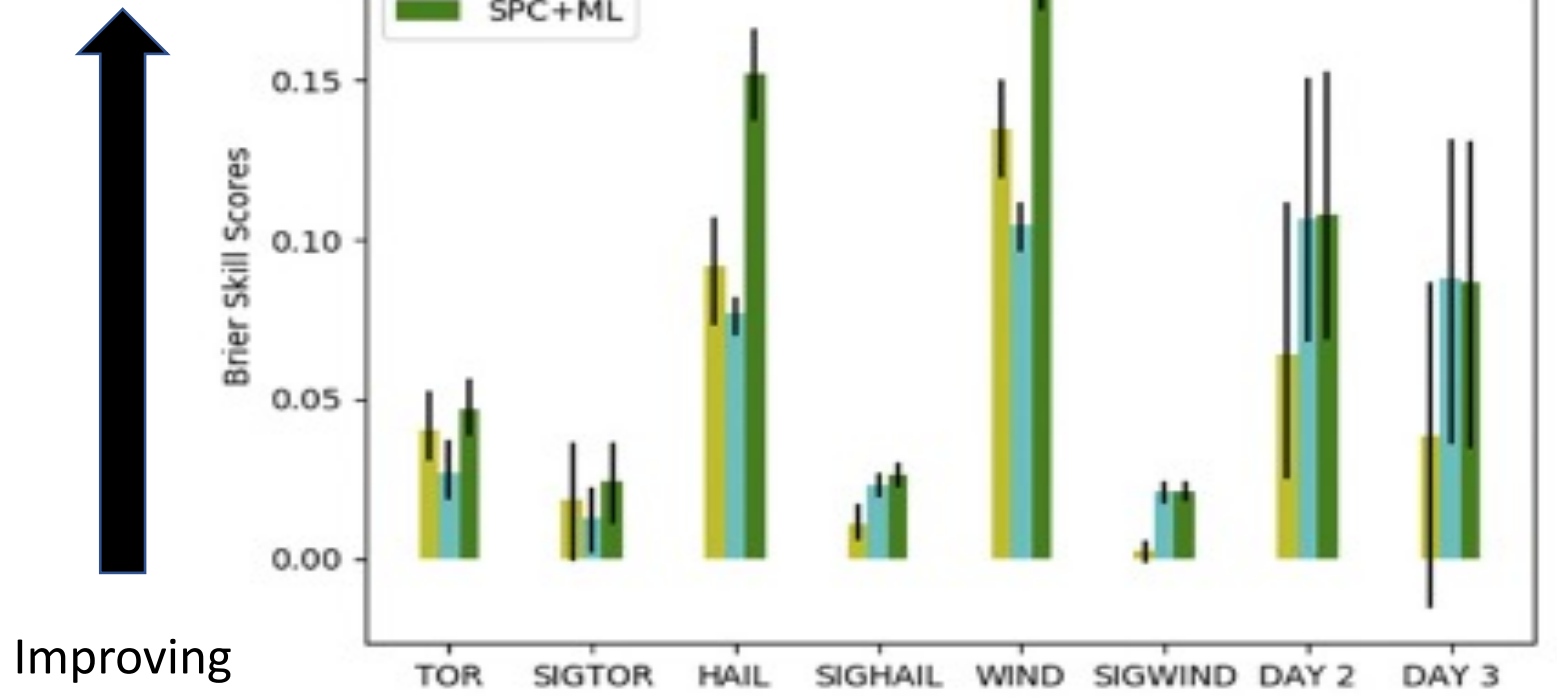
NWS Focal Points: Christina Speciale (ALB), Neil Stuart (ALB)

Background

- Numerous sources of data for forecasters
- Develop algorithm to breakdown conditions surrounding an event
- Identify important environmental variables for each type of precipitation
- Synthesize data and conditions to improve forecasting ability

Machine Learning in Operational Forecasting

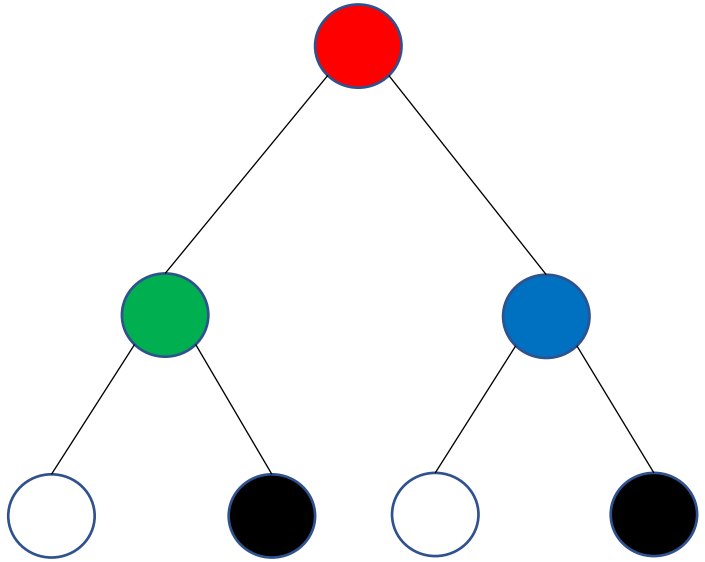
- Become more common in recent years
- Proven to be effective in several areas of forecasting
- Improve forecasts when used alone or in combination with forecasters



Hill et al. (2020)

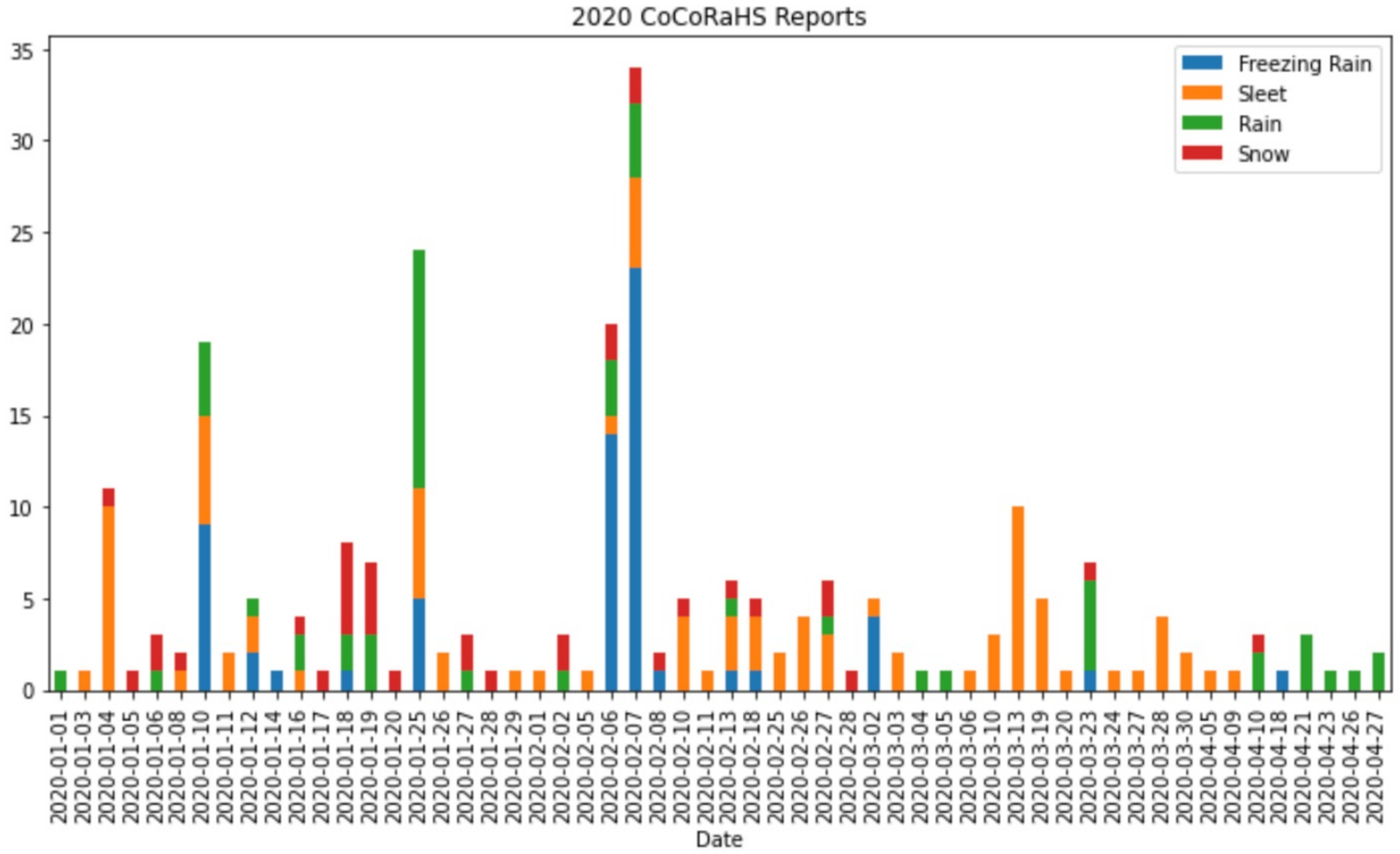
Random Forest

- ‘Forest’ of decision trees
- Identify patterns and nonlinear interactions in data
- Train the trees to make a prediction from its previous knowledge
- Generate a probabilistic outcome and relative feature importance



Case List

- Study period January 2017 – September 2020
- Using CoCoRaHS data to identify cases
- Precipitations types identified: Rain, Snow, Freezing Rain, and Sleet



Data Sources

- New York State Mesonet (NYSM)
  - Hourly Statistics
  - 5-minute observations
- Upper Air Soundings
  - BUF, ALB, OKX offices and WMW (Canada)
- North American Mesoscale (NAM) Forecast Model
  - 4km resolution from BUFKIT

10 m Winds (sonic)
10 m Winds (prop)
9 m Temperature
2 m Temperature
Relative Humidity
Solar Insolation
Precipitation
Snow Depth
Camera
Pressure
5 cm Soil
25 cm Soil
50 cm Soil

Features (Variables) in Datasets

NYSM Hourly	NYSM 5-Minute
Temperature (min, max, avg)	2m Temperature
Relative Humidity (min, max, avg)	Relative Humidity
Station Pressure (min, max, avg)	Station and Sea Level Pressure
Insolation (avg and total)	Insolation
Precipitation (hourly, daily, intensity)	Precipitation (hourly, daily, intensity)
Wind Speed (avg and max) and Direction (avg)	Wind Speed (avg and max) and Direction (avg)

Features (Variables) in Datasets

	Standard Pressure Levels (surface, 925, 850, 700, 500)	Differences between Standards Pressure Levels	Supplemental Variables	Added Variables
Sounding Profiles	• Temperature • Pressure • Dew point • Wind Speed and Direction • Geopotential Height • Wet Bulb Temperature • Relative Humidity	• Temperature • Precipitable Water Vapor • Wind Speed and Direction	• Critical thickness- (surface-850hPa and surface-500hPa)	• Max Wet Bulb Temperature 925-700hPa • Positive and Negative areas and ratio of Positive to Negative (Bourgouin 2000) • Critical Thickness- (850-700hPa and 700-500hPa) • Mean Relative Humidity surface-500hPa • Dew point depression • Mean Temperature- (surface-850hPa and surface-700hPa) • Min Temperature surface-850hPa • Max Temperature 850-700hPa

Random Forest Methods & Evaluation

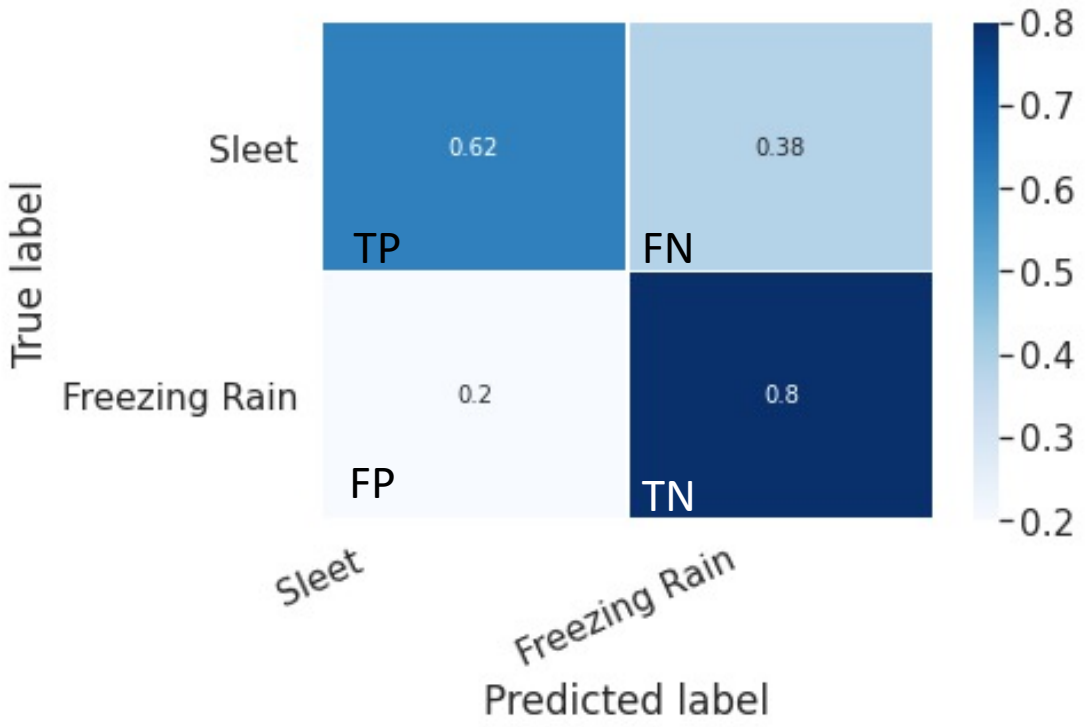
- Configuration: 650 trees, 75/25 training and testing split, stratified
- Accuracy, Precision, Recall, F1 Scores

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Cases}$$

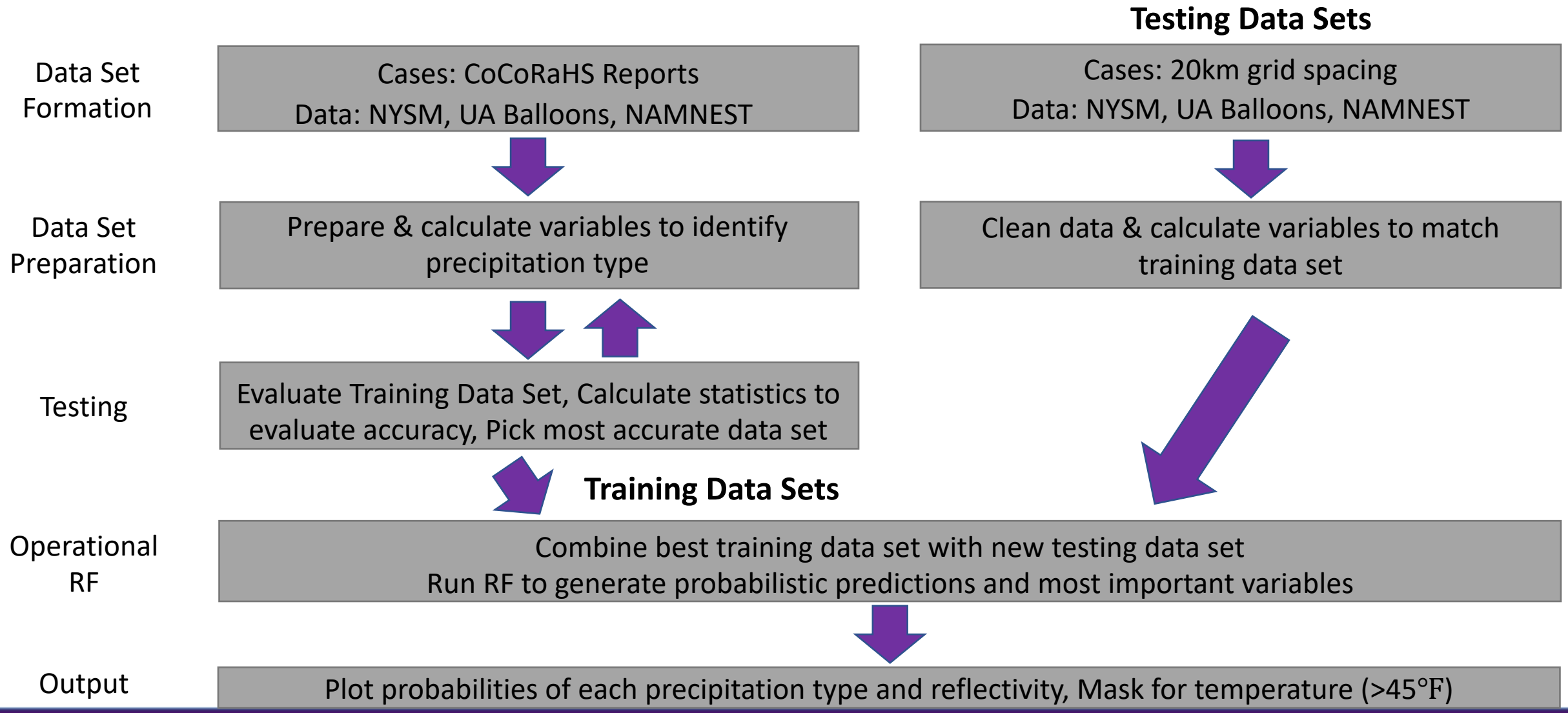
$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + True\ Negative}$$

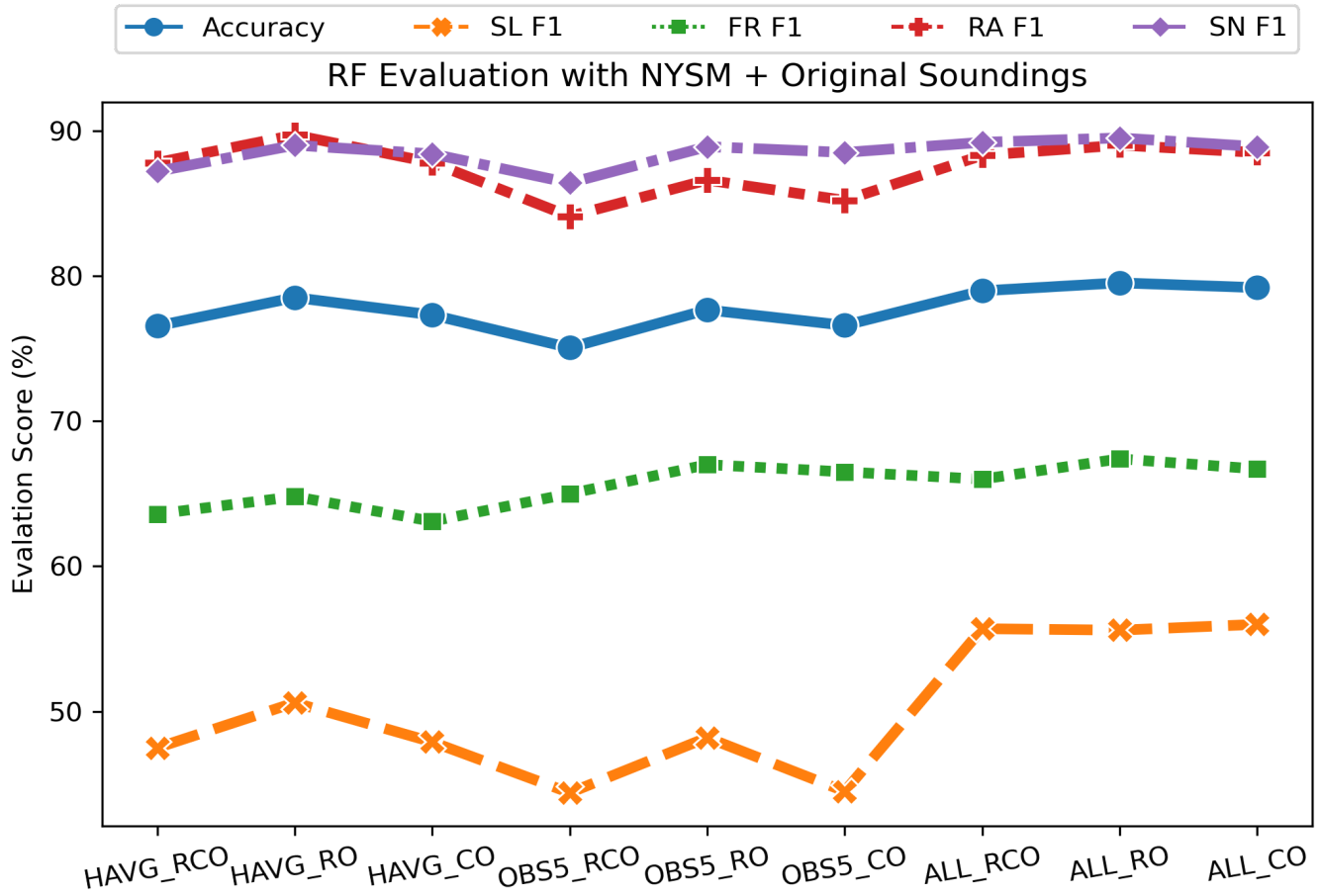
$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$



Random Forest Framework

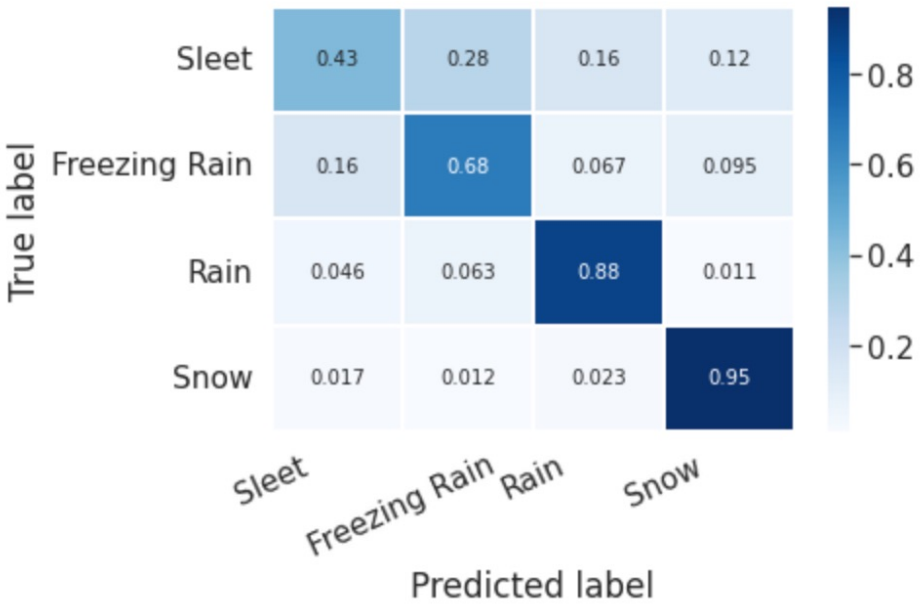


Random Forest Results



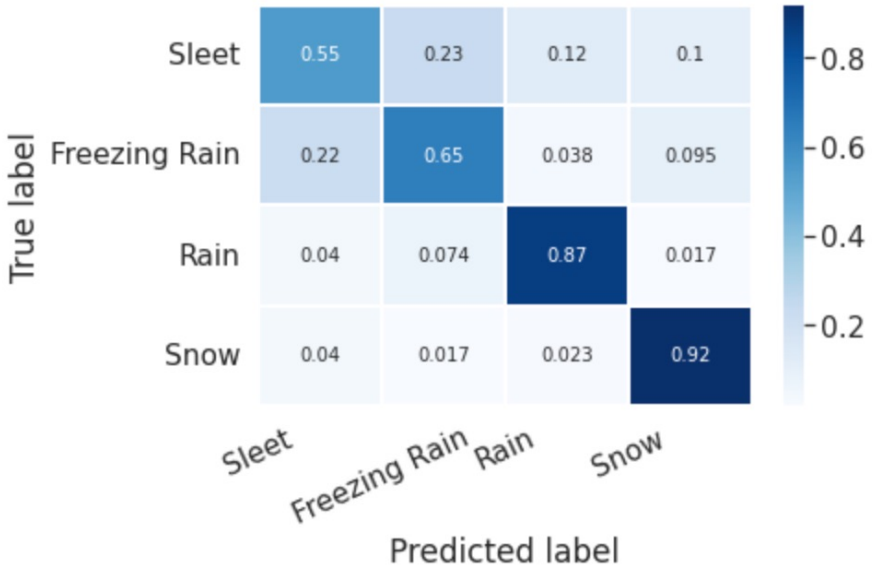
Original Soundings- NWS Buffalo, Albany, and Upton
HAVG_RCO: NYSM Hourly Averaged Raw and Calculated Original
HAVG_RO: NYSM Hourly Averaged Raw Original
HAVG_CO: NYSM Hourly Averaged Calculated Original
OBS5_RCO: NYSM 5-minute observations Raw and Calculated Original
OBS5_RO: NYSM 5-minute observations Raw Original
OBS5_CO: NYSM 5-minute observations Calculated Original
ALL_RCO: ALL NYSM Raw and Calculated Original
ALL_RO: ALL NYSM Raw Original
ALL_CO: ALL NYSM Calculated Original

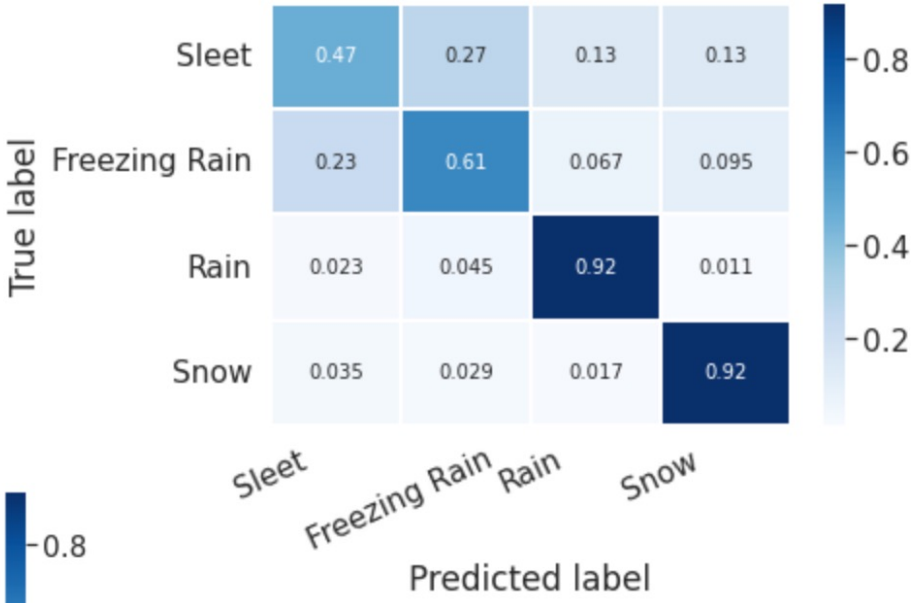
Random Forest Results



NYSM 5 Min and Sounding raw data

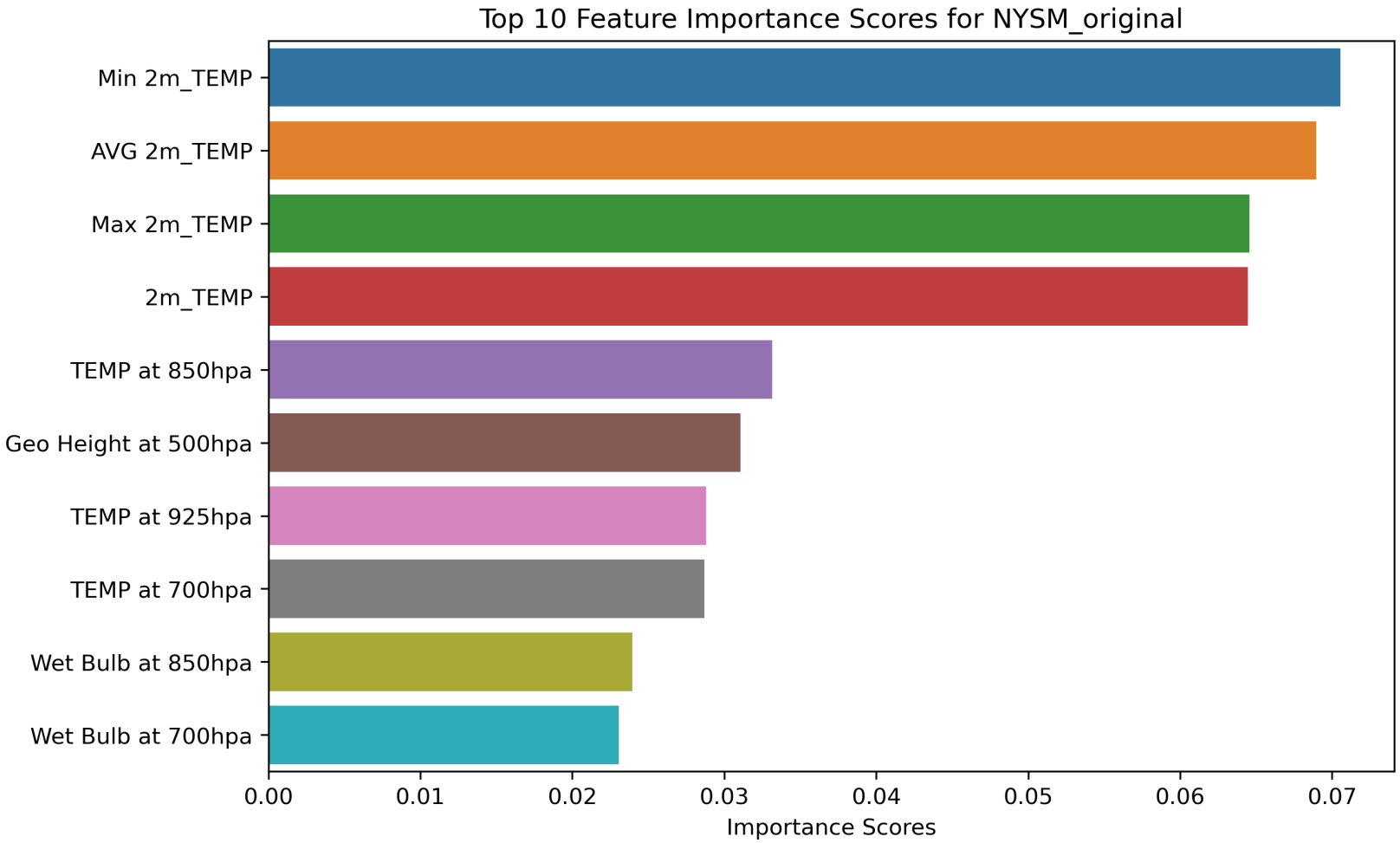
All NYSM and Sounding Raw data





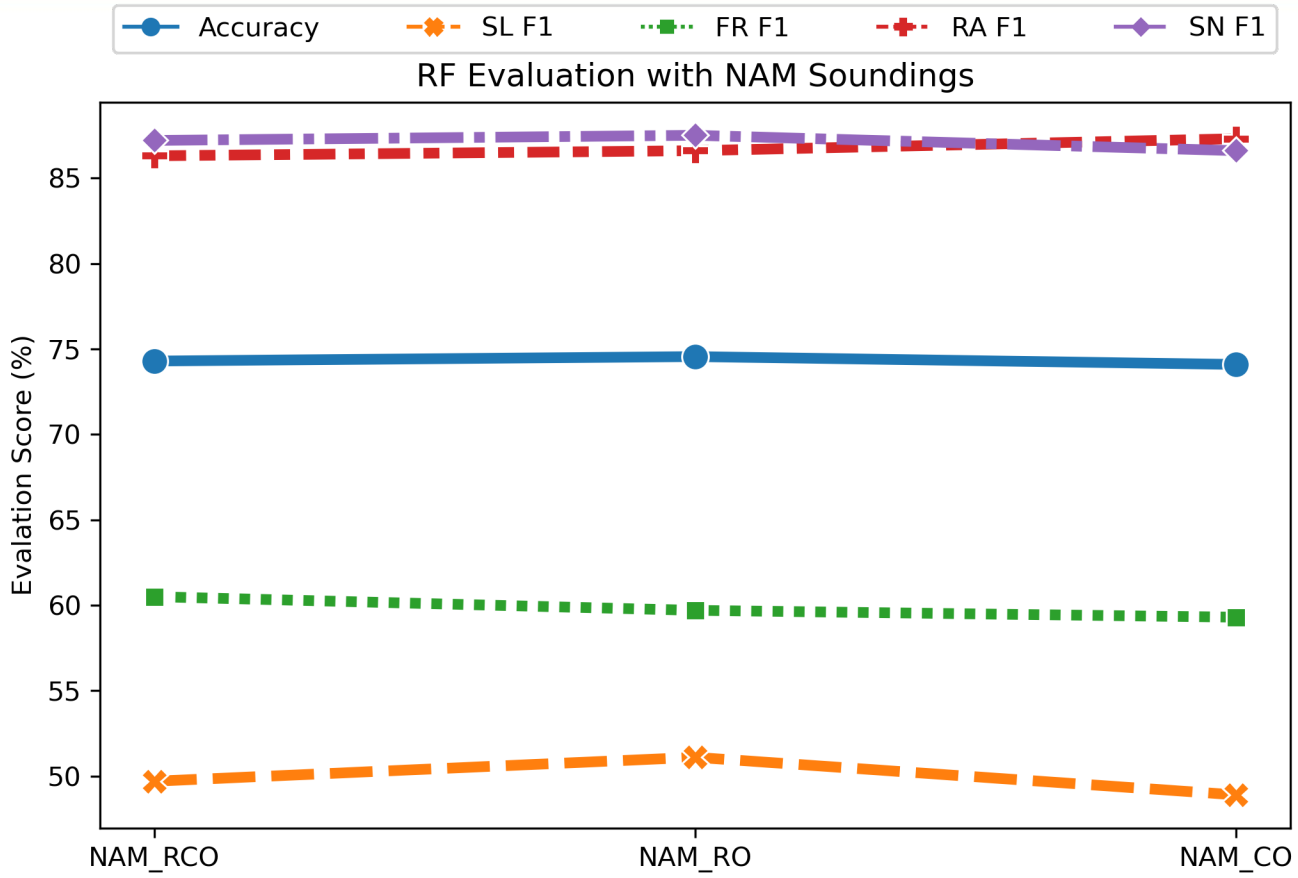
NYSM Hourly AVG and Sounding Raw data

Random Forest Results



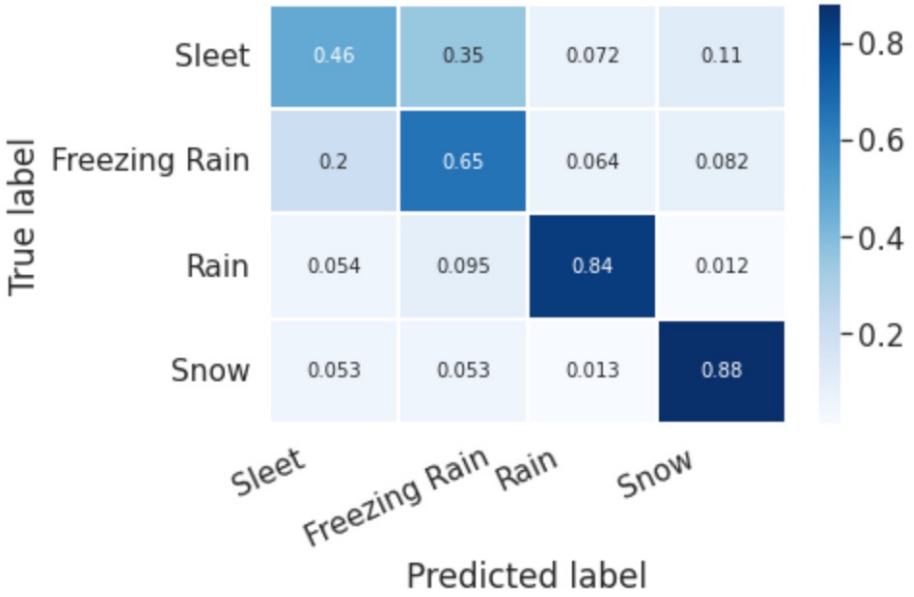
Random Forest Results

- NAM 4km sounding profile dataset
- Uses forecast hour timing to match with events



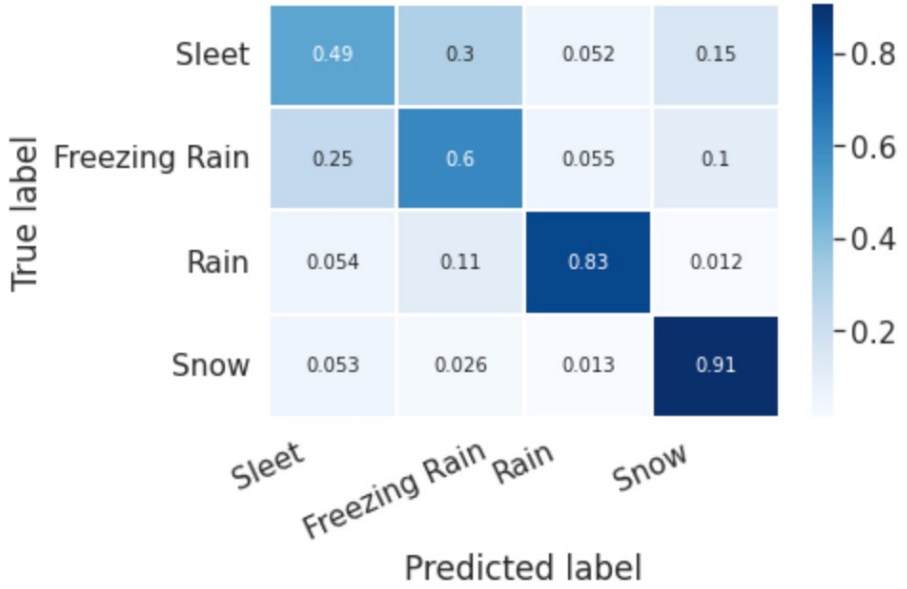
NAM_RCO: NAM Raw and Calculated Original
NAM_RO: NAM Raw Original
NAM_CO: NAM Calculated Original

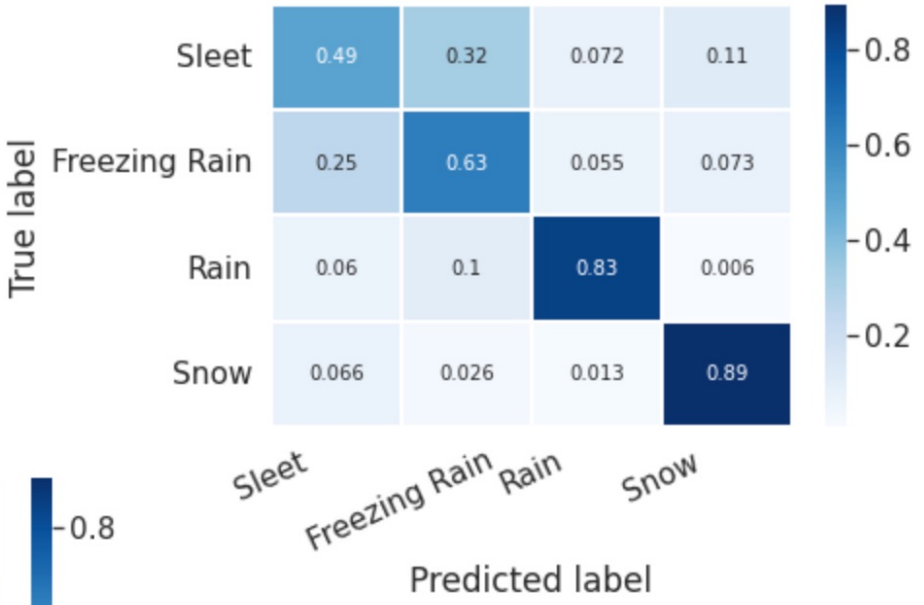
Random Forest Results



NAM 4km Soundings Raw data

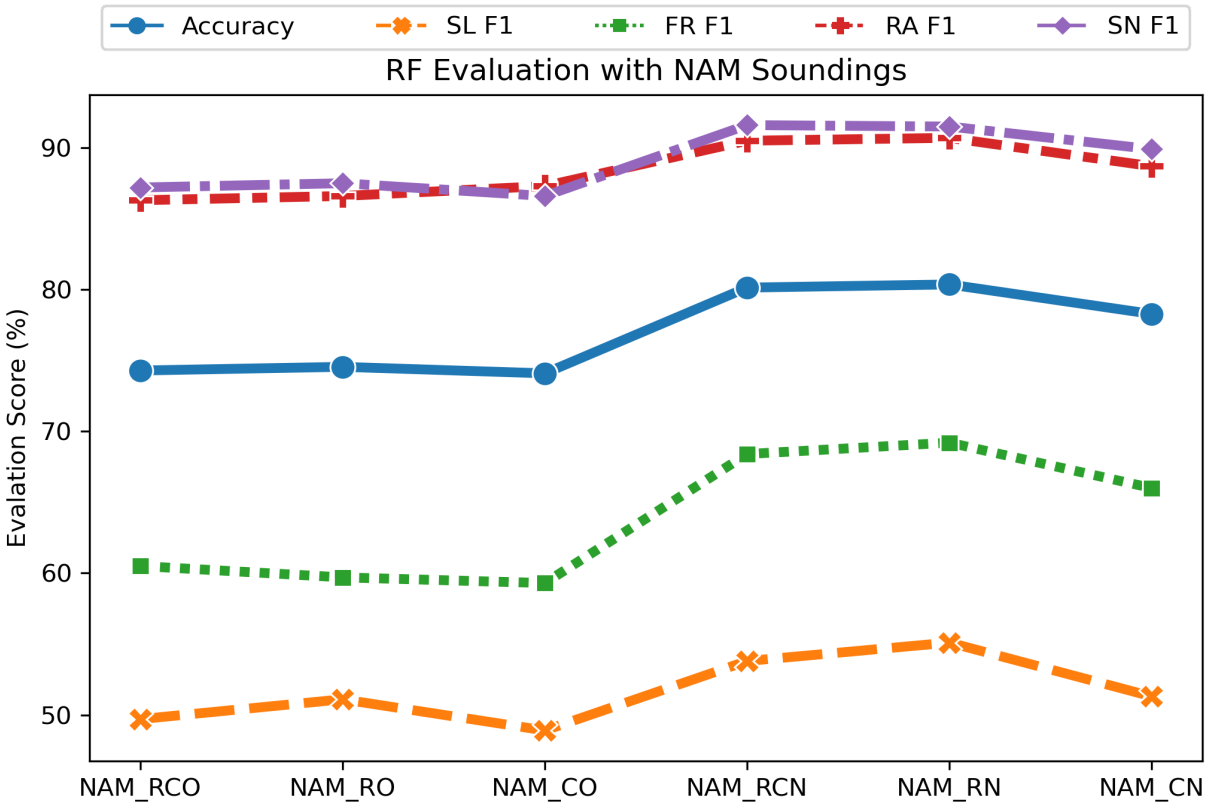
NAM 4km Soundings Calculated data





NAM 4km Soundings Raw and Calculated data

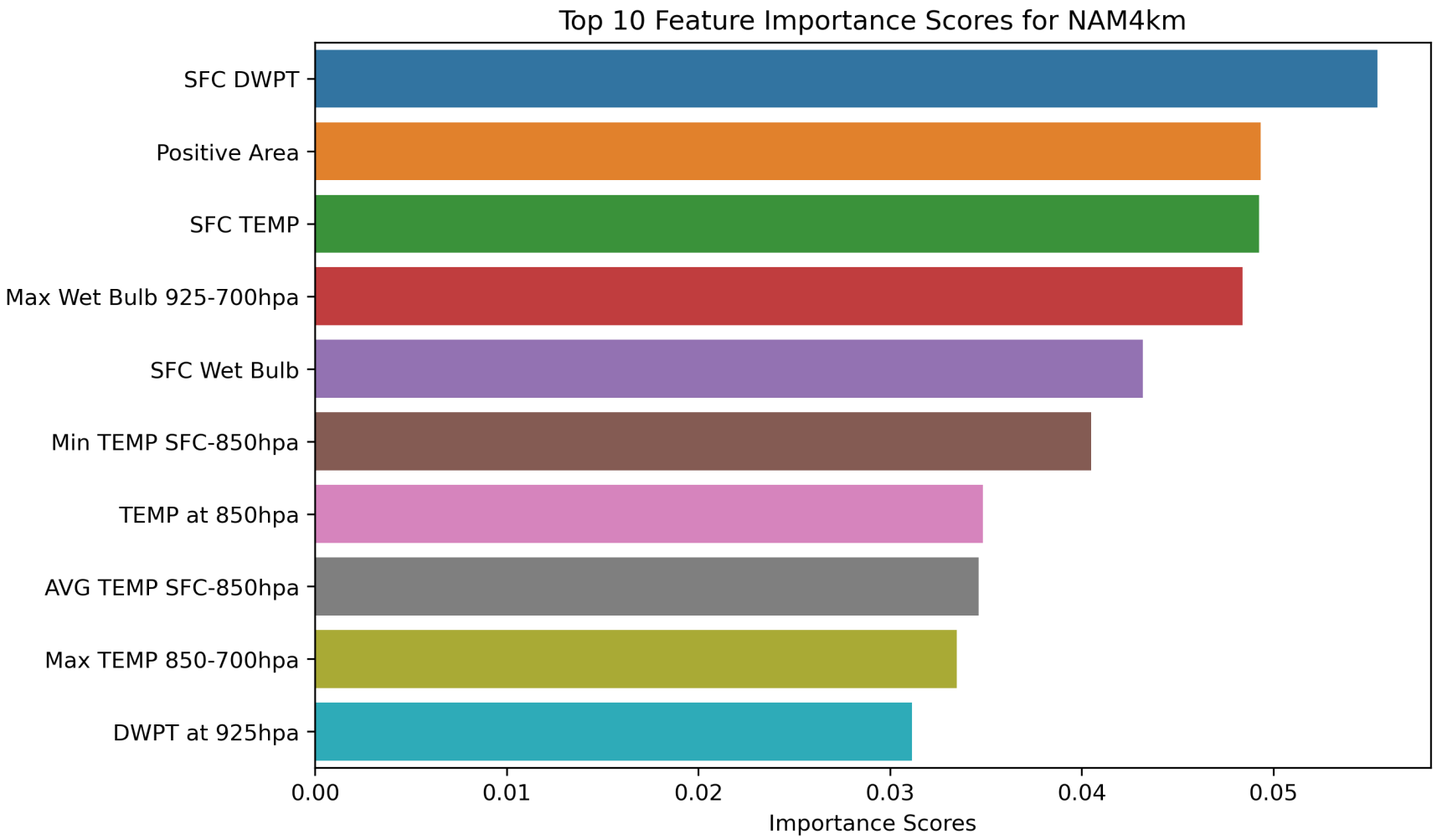
Random Forest Results



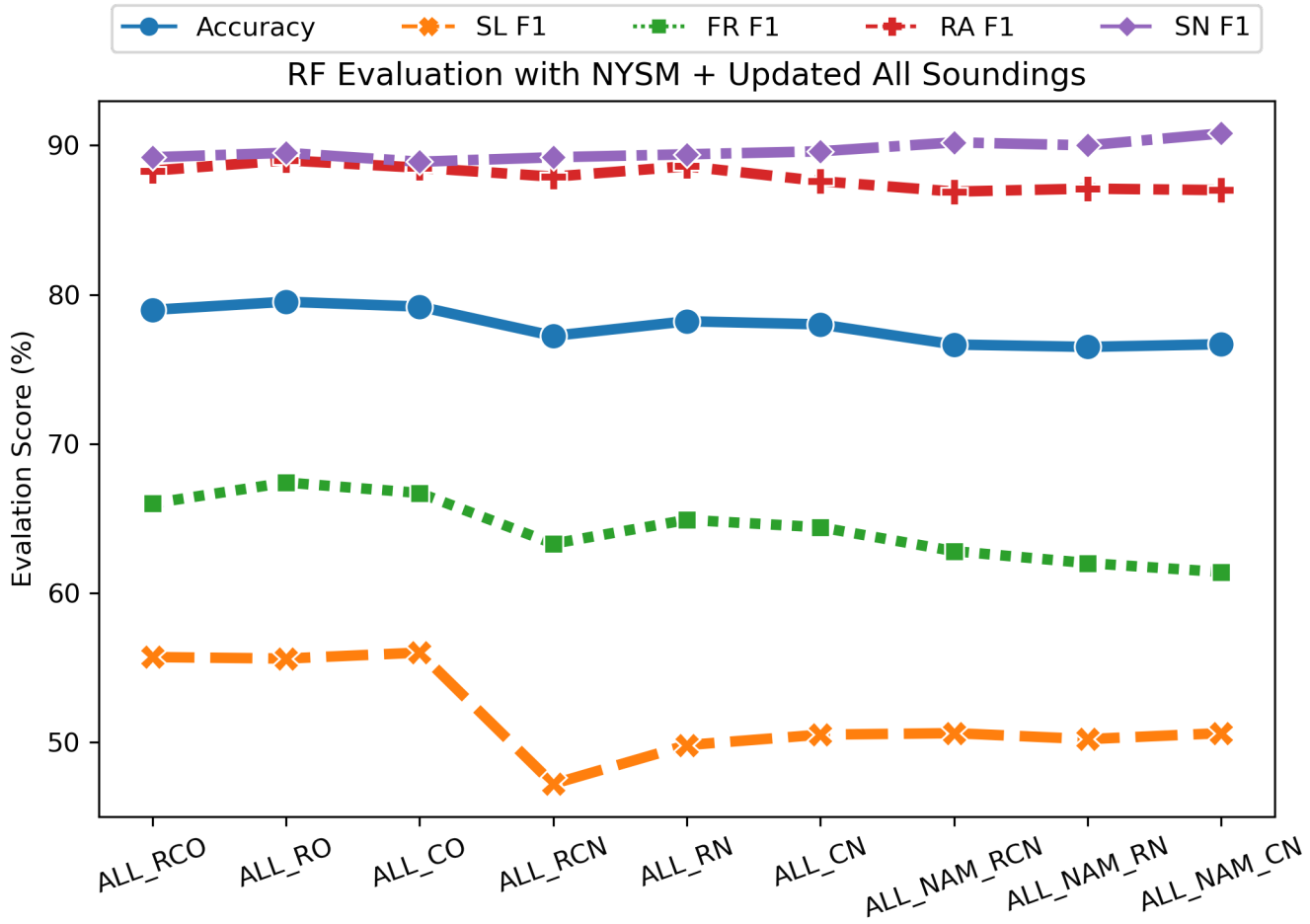
NAM_RCO: NAM Raw and Calculated Original
NAM_RO: NAM Raw Original
NAM_CO: NAM Calculated Original

NAM_RCN: NAM Raw and Calculated New
NAM_RN: NAM Raw New
NAM_CN: NAM Calculated New

Random Forest Results



Random Forest Results

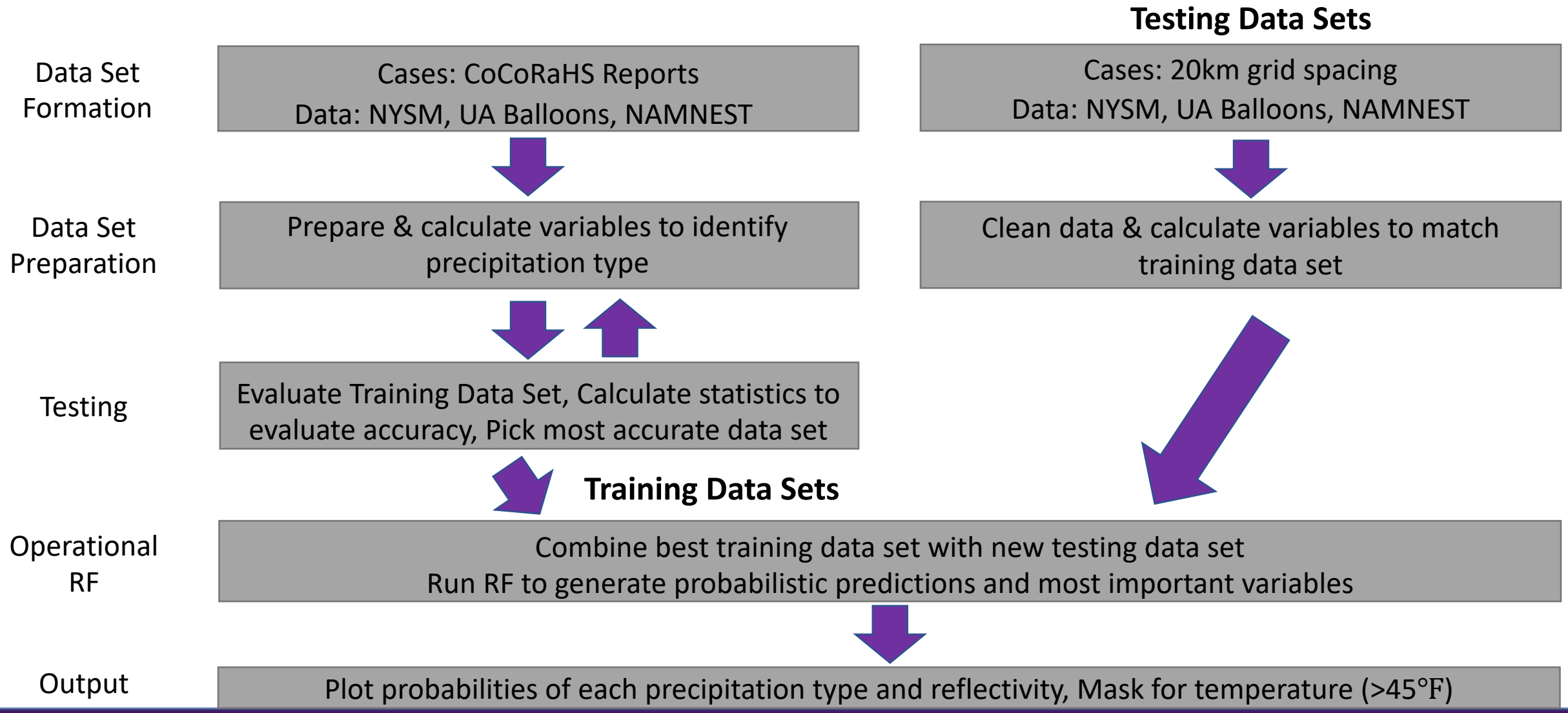


Original Soundings- NWS Buffalo, Albany, and Upton
Updated Soundings- NWS Buffalo, Albany, Upton and Maniwaki, Quebec or NAM
ALL_RCO: ALL NYSM Raw and Calculated Original
ALL_RO: ALL NYSM Raw Original
ALL_CO: ALL NYSM Calculated Original

ALL_RCN: ALL NYSM Raw and Calculated New
ALL_RN: ALL NYSM Raw New
ALL_CN: ALL NYSM Calculated New

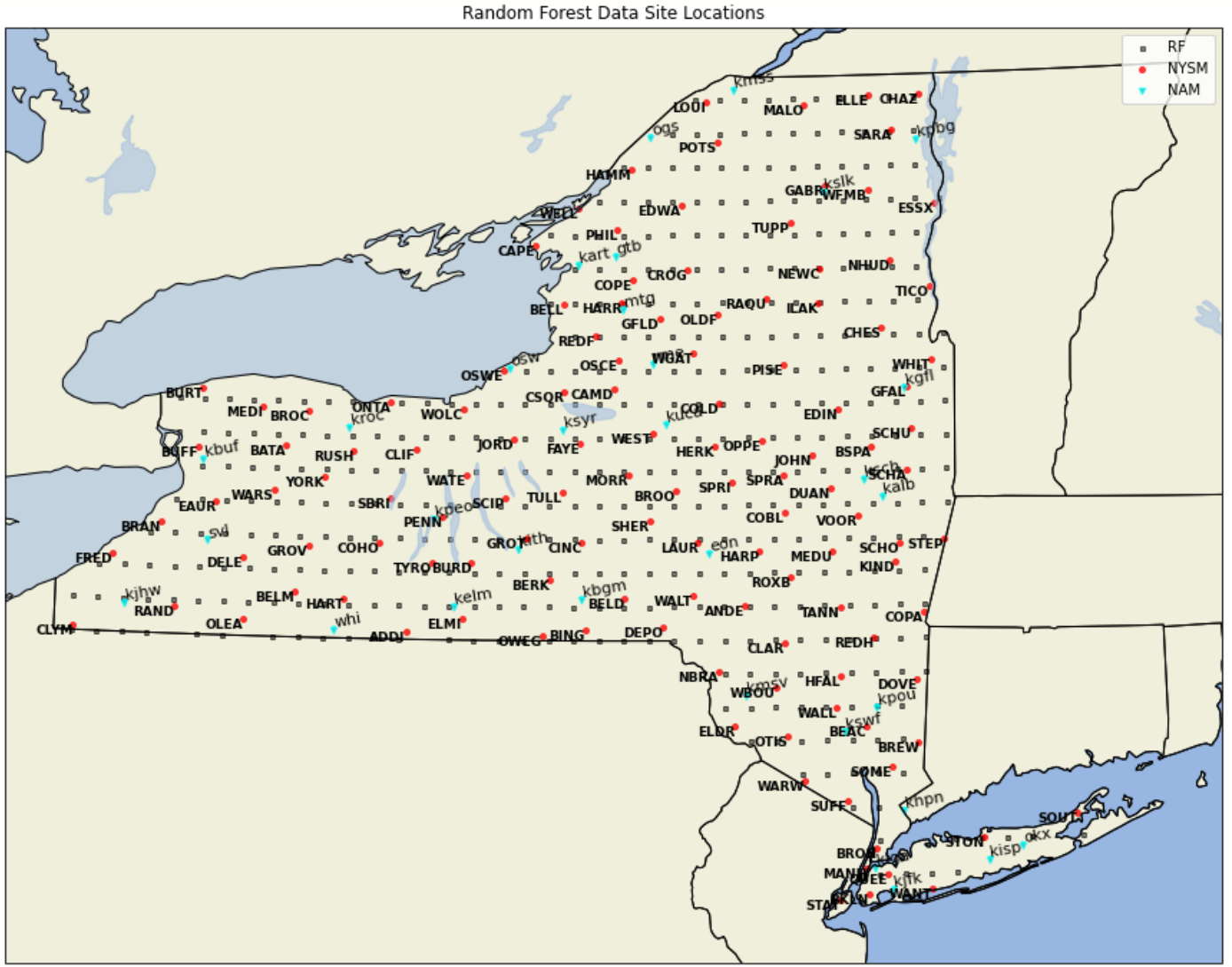
ALL_NAM_RCN: ALL NYSM and NAM Raw and Calculated New
ALL_NAM_RN: ALL NYSM and NAM Raw New
ALL_NAM_CN: ALL NYSM and NAM Calculated New

Random Forest Framework

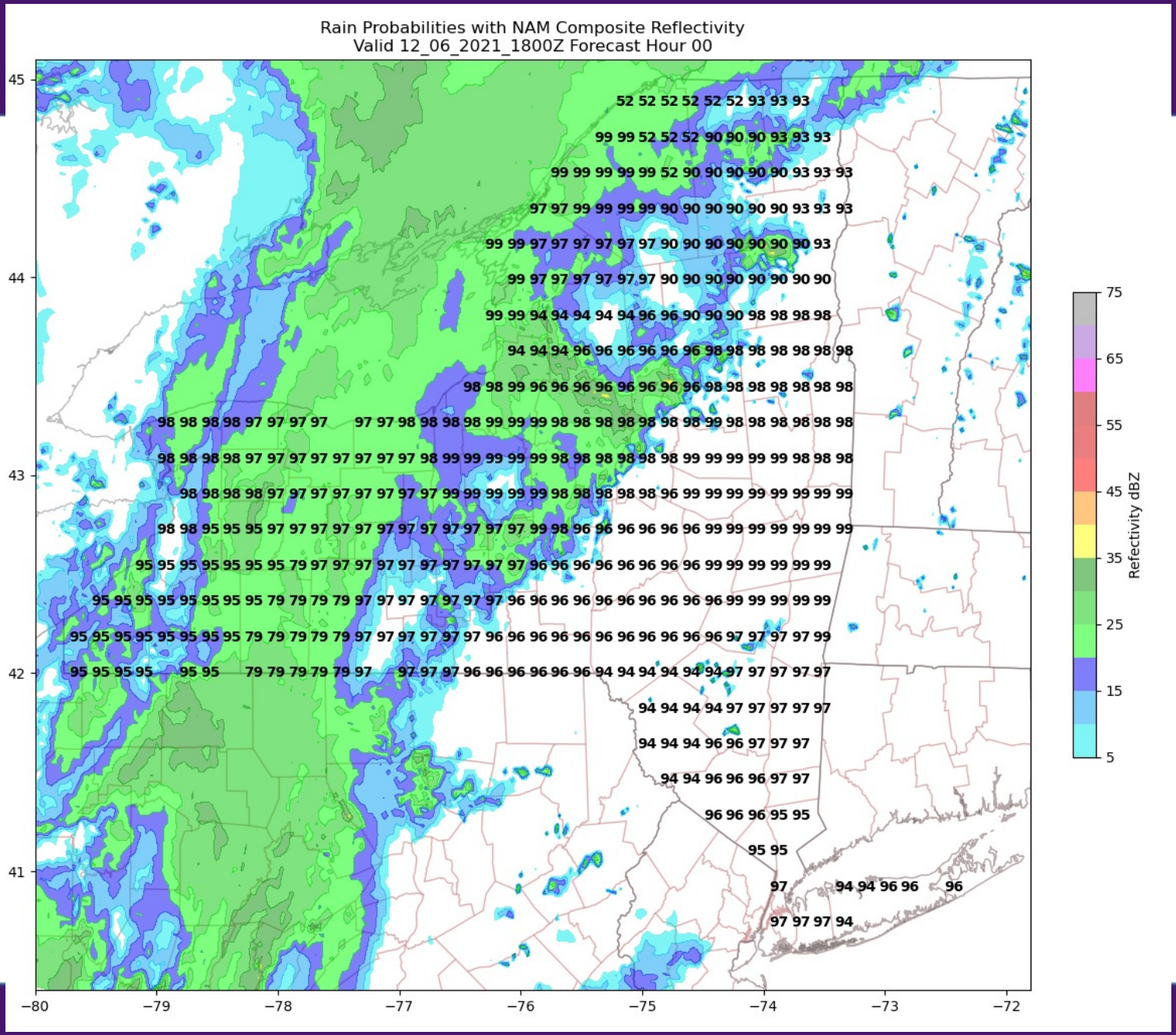


Operational Product

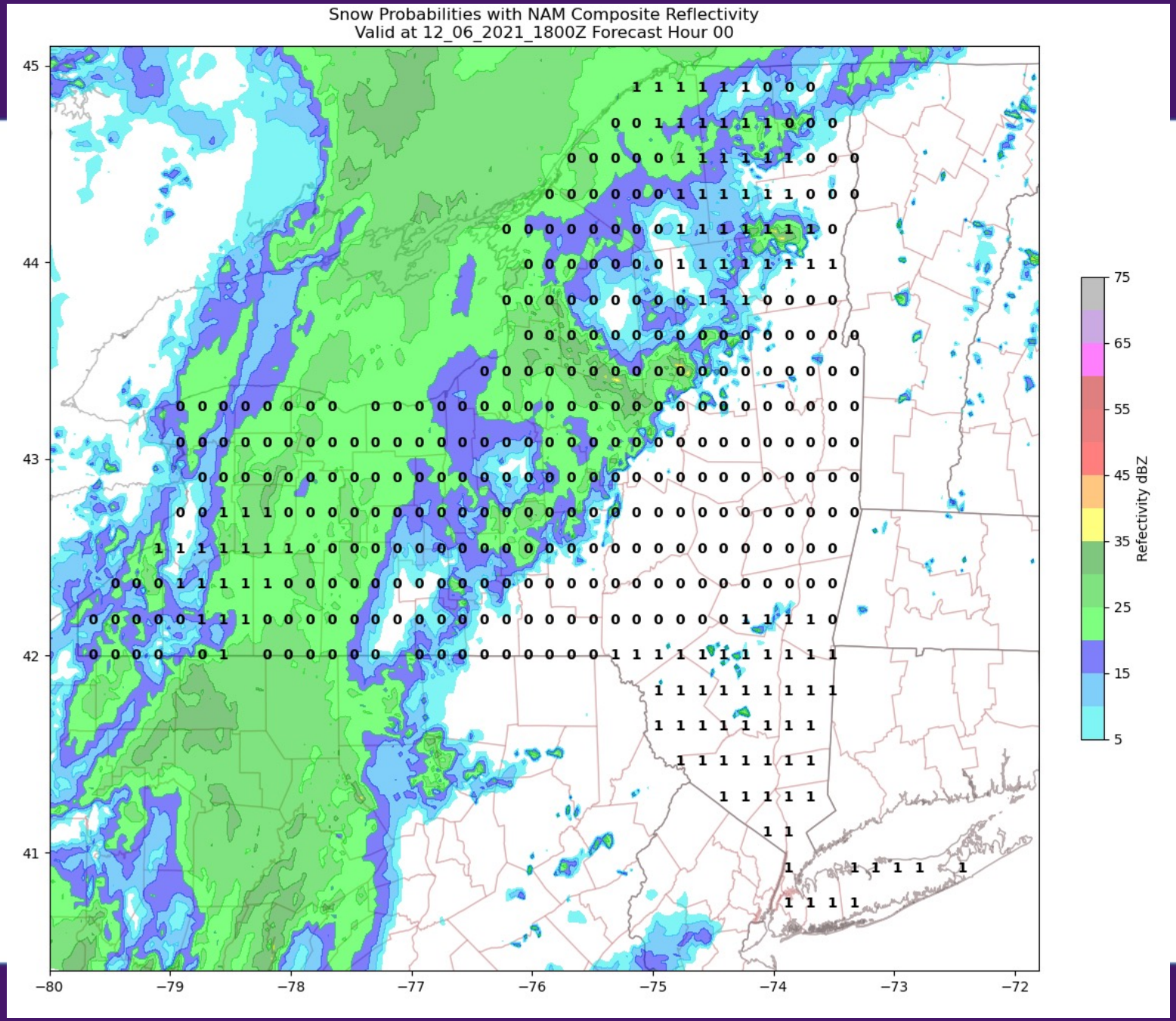
- Develop web-based product for operational use
- Live updating map of probabilities with radar/reflectivity
- Incorporate information about most important variables



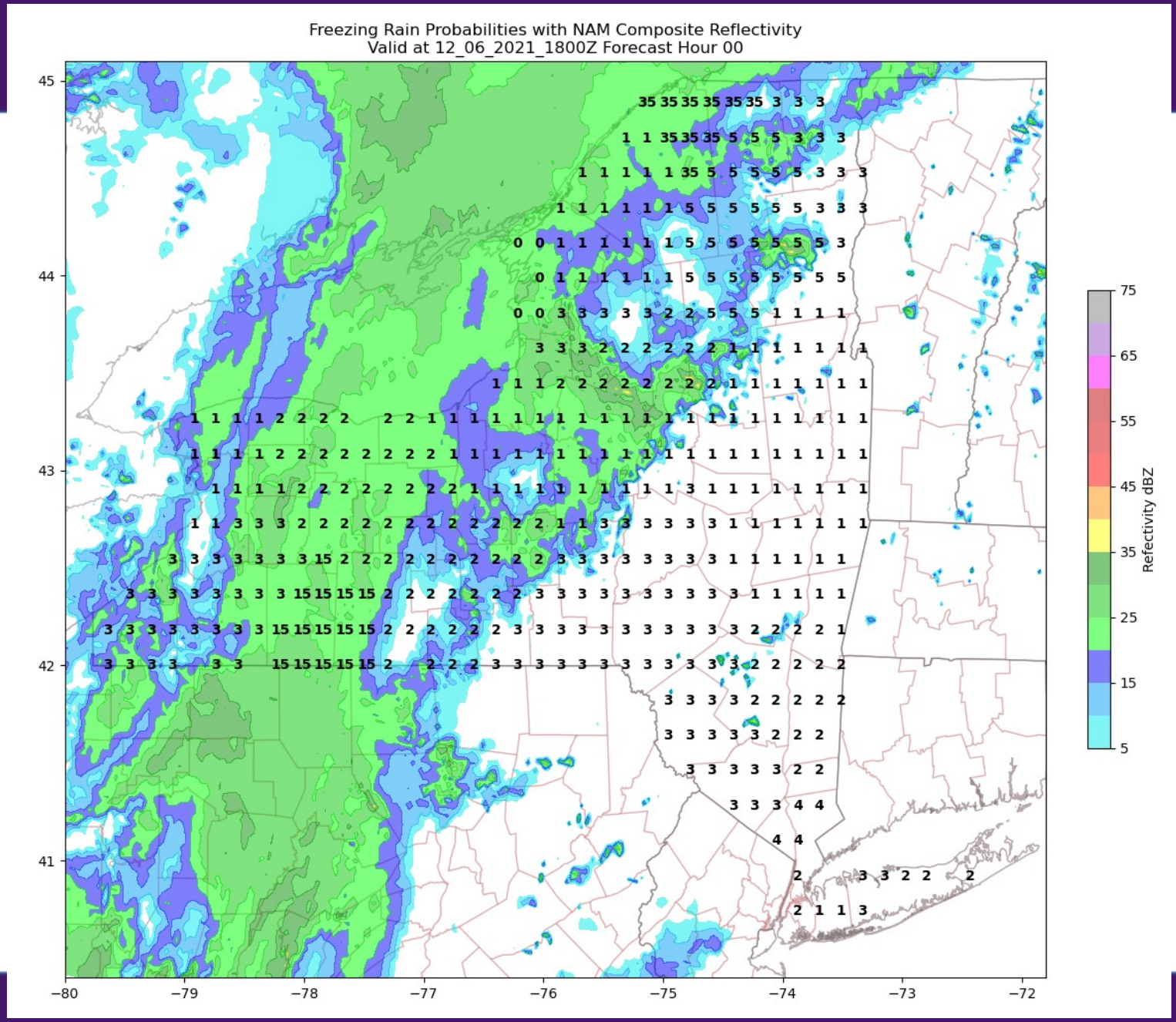
Products


Rain Probabilities with NAM Composite Reflectivity
Valid 12_06_2021_1800Z Forecast Hour 00
Refectivity dBZ
75
65
55
45
35
25
15
5
45
44
43
42
41
−80
−79
−78
−77
−76
−75
−74
−73
−72

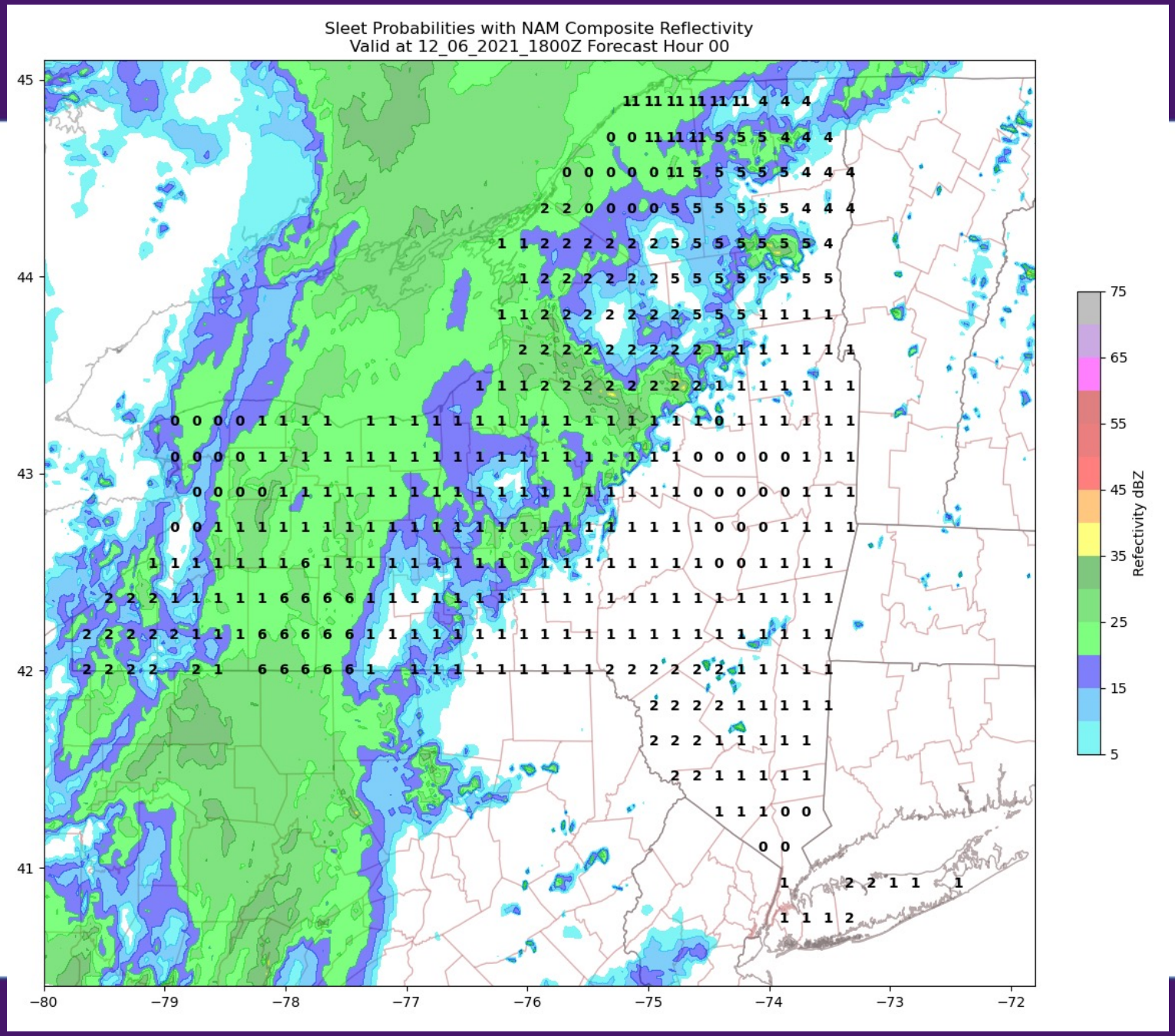
Products


Snow Probabilities with NAM Composite Reflectivity
Valid at 12_06_2021_1800Z Forecast Hour 00
45
44
43
42
41
−80
−79
−78
−77
−76
−75
−74
−73
−72
75
65
55
45
35
25
15
5
Refectivity dBZ

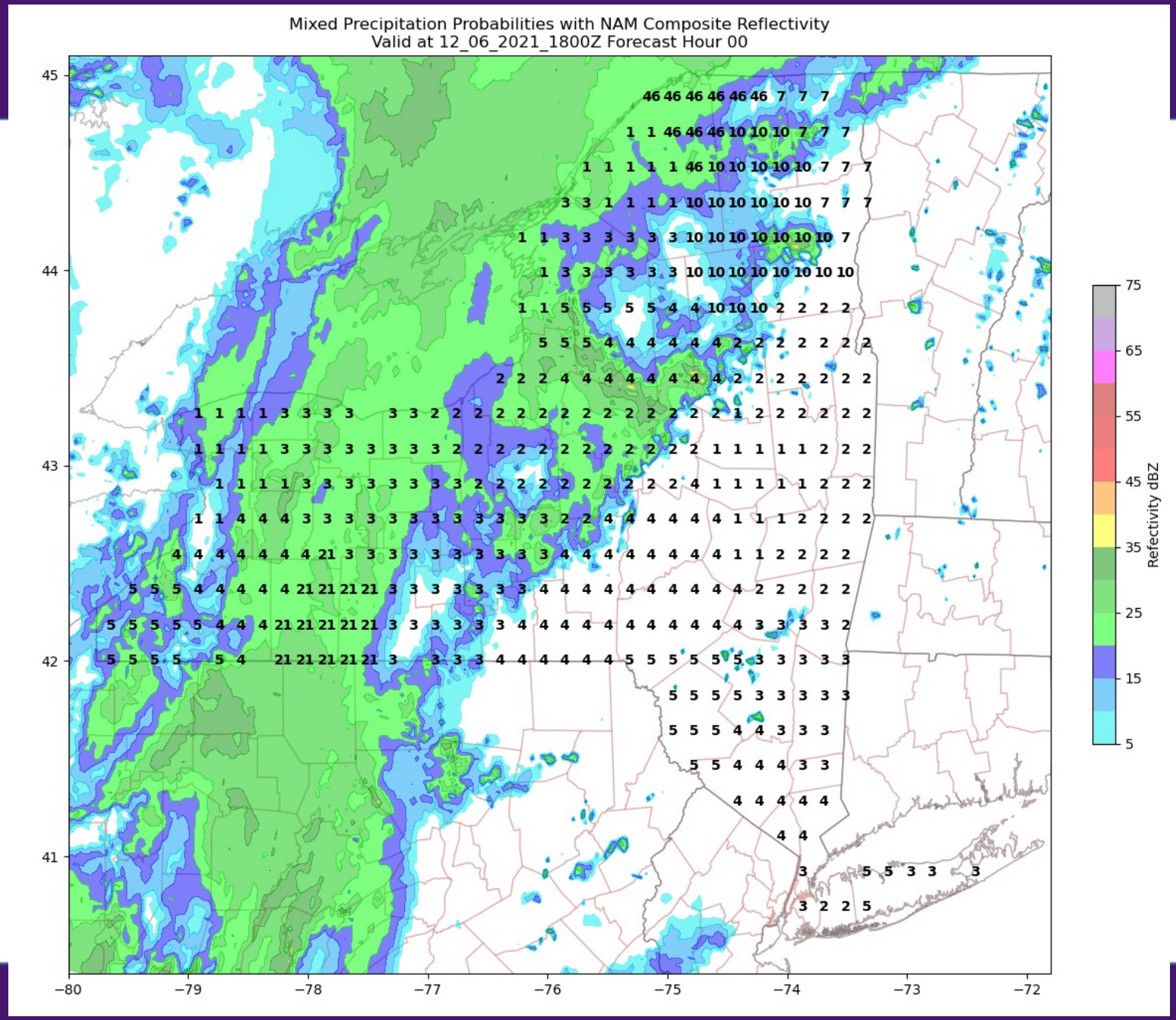
Products


Freezing Rain Probabilities with NAM Composite Reflectivity
Valid at 12_06_2021_1800Z Forecast Hour 00
45
44
43
42
41
−80
−79
−78
−77
−76
−75
−74
−73
−72
Refectivity dBZ
75
65
55
45
35
25
15
5

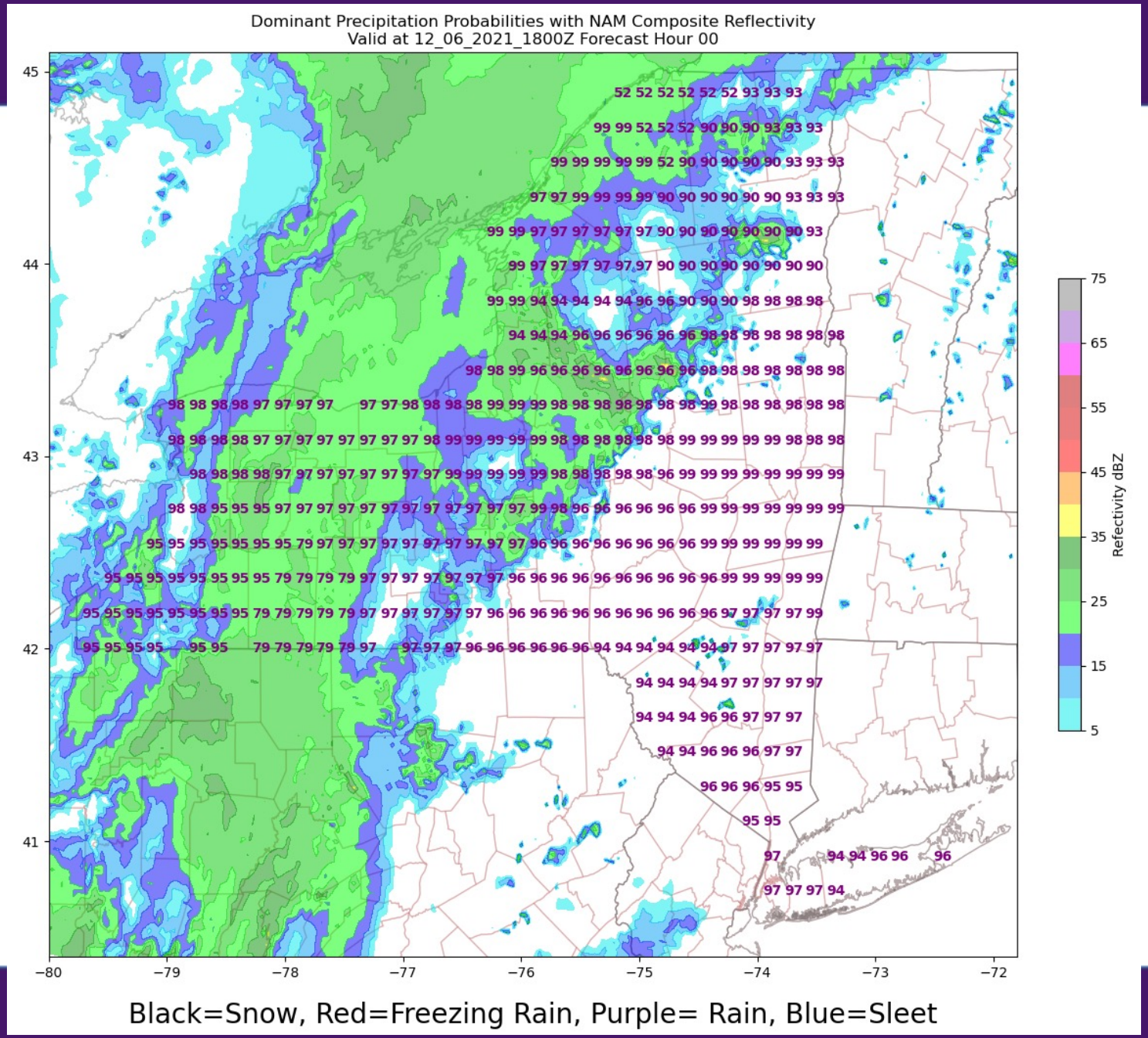
Products


Sleet Probabilities with NAM Composite Reflectivity
Valid at 12_06_2021_1800Z Forecast Hour 00
45
44
43
42
41
−80
−79
−78
−77
−76
−75
−74
−73
−72
75
65
55
45
35
25
15
5
Refectivity dBZ

Products


Mixed Precipitation Probabilities with NAM Composite Reflectivity
Valid at 12_06_2021_1800Z Forecast Hour 00
45
44
43
42
41
−80
−79
−78
−77
−76
−75
−74
−73
−72
Refectivity dBZ
75
65
55
45
35
25
15
5

Products


Dominant Precipitation Probabilities with NAM Composite Reflectivity
Valid at 12_06_2021_1800Z Forecast Hour 00
45
44
43
42
41
−80
−79
−78
−77
−76
−75
−74
−73
−72
Refectivity dBZ
75
65
55
45
35
25
15
5
Black=Snow, Red=Freezing Rain, Purple= Rain, Blue=Sleet

Operational Product

Data Fusion Operational Website:

http://www.atmos.albany.edu/student/filipiak/op/

Welcome to the Data Fusion Winter Weather AI Project!

This page has been developed through research conducted as a part of a NOAA CSTAR grant

All plots and data displayed is experimental in nature and should not be used other than in a research framework

To learn more about the algorithm, please visit the Data and Methods section
For information about how to use this product, please visit the Training tab

About this project:

This project is a part of the University at Albany's long-standing collaboration with the National Weather Service (NWS) through the Collaborative Science, Technology, and Applied Research (CSTAR) Program; this specific project is supported by NOAA Award Number NA19NWS4680006. To learn more about this and other CSTAR projects from the University at Albany, please visit the VLAB page.

The goal of the Data Fusion project is to merge a multitude of data sources forecasters can use to make a forecast and succinctly and effectively combine them to improve our forecasting ability of hazardous weather events like winter mixed precipitation events. To translate this work to operations, this website was created to share the live updating plots with winter precipitation probabilities. In addition to the website, training modules or quick references for the forecasters will be created as another tool to share what information and data is most important to determining precipitation type.

This tool uses machine learning as its base because of its ability to synthesize and learn how to differentiate between precipitation types. A random forest machine learning algorithm will be used in this case because of its ability to detail, clearly and explicitly, the decision-making process behind the algorithm, while handling large amounts of data. This tool is meant to be used in conjunction with a forecaster because the computer does not take things like local terrain into account, so local NWS forecasters can use this tool and their local knowledge to make the best possible prediction They can also use the output of the most important variables to help when making their own assessment of what the precipitation type will be based on information in other locations.

About the Data Fusion Team:

Brian Filipiak is currently a graduate student at the University at Albany. His work on this project is supported by his University at Albany advisors: Kristen Corbosiero, Andrea Lang, Nick Bassill, and Ross Lazear.

In addition to the Data Fusion team at the University at Albany, Brian's NWS focal points for the project are Christina Speciale, Neil Stuart, and Mike Evans at the Albany WFO. They have and continue to provide valuable guidance, suggestions, and feedback throughout the process.

Contact: Brian Filipiak bfilipiak@albany.edu

Data and Methods

Data Sources

New York State Mesonet (NYSM)
University of Wyoming Upper Air Data
Penn State BUFKIT NAMNEST (4-km)
Community Collaboration Rain, Hail and Snow (CoCoRaHS) Network Reports

Methods

To start the process of analyzing winter mixed precipitation events, CoCoRaHS reports were used to identify mixed precipitation (freezing rain and sleet) events in New York between January 2017 and September 2020. In addition to the mixed precipitation events, rain and snow events during the cold season (October-April) over the same time period were also identified using CoCoRaHS reports. The cases identified with one of the 4 precipitation types (freezing rain, sleet, rain and snow) were examined and verified using New York State Mesonet, radar, and satellite observations. This process resulted in about 3000 verified cases forming the training dataset to make all future predictions from.

Data to help train the machine learning algorithm was incorporated into the CoCoRaHS reports through locating the nearest Mesonet site, balloon launch site, and BUFKIT profile site and using the data they provide at the nearest time to the event.

Machine learning was applied to the verified CoCoRaHS reports to generate probabilistic values for each precipitation type in nowcasts or forecasts. This algorithm uses a random forest machine learning algorithm; the random forest is an ensemble decision tree machine learning method where numerous decision trees are used to make a prediction based on previous learned data.

In our algorithm, the random forest classifier method is used to generate predictions for each class or type of precipitation. This means that the predictions made by the random forest represents what type of precipitation is currently or forecasted to be occurring as opposed to how much of that precipitation will fall. The probabilities for each location are based on each tree in the forest voting for a precipitation type and coming to a consensus. The random forest is run 50 times, so the final calculated probability is the average across the 50 runs.

Live data is brought in from the New York State Mesonet, University of Wyoming and Penn State to form the different data sets that are fed into the random forest algorithm. The live data sets are collected based on points that are spaced in a 20-km grid across New York. This spacing encompasses all the BUFKIT sites as well as 125 of the 126 Mesonet stations, with Manhattan as the only station not used. The final data sets given to the random forest are filtered by temperature and/or precipitation. The Mesonet and Upper Air data is filtered only on 2-meter temperature based on the 95th percentile 2-meter temperature (about 45°F) of the training data. The NAMNEST data is filtered on surface temperature based on the 95th percentile surface temperature (about 45°F) of the training data and if there is predicted precipitation during that forecast hour at the BUFKIT location.

The output of the random forest probabilities is plotted on the maps of New York and surrounding states with county outlines and radar or reflectivity overlay. A blank map or blank areas on a map indicate that there is missing data or that those points have been filtered out because of temperature or precipitation.

The Mesonet and Upper Air plots update 45 minutes past each hour, and the NAMNEST plots update about 4 hours after model initialization.

Contact: Brian Filipiak bfilipiak@albany.edu

Training

A training guide will be posted soon. Please check back at a later time!

Contact: Brian Filipiak bfilipiak@albany.edu

NYSM and Upper Air Raw and Calculated Data

Home | NYSM & Upper Air | NAMNEST | Data and Methods | Training

Dec 7, 2021 01:31:11 UTC

******EXPERIMENTAL******

Please click any button to view the map of your choice

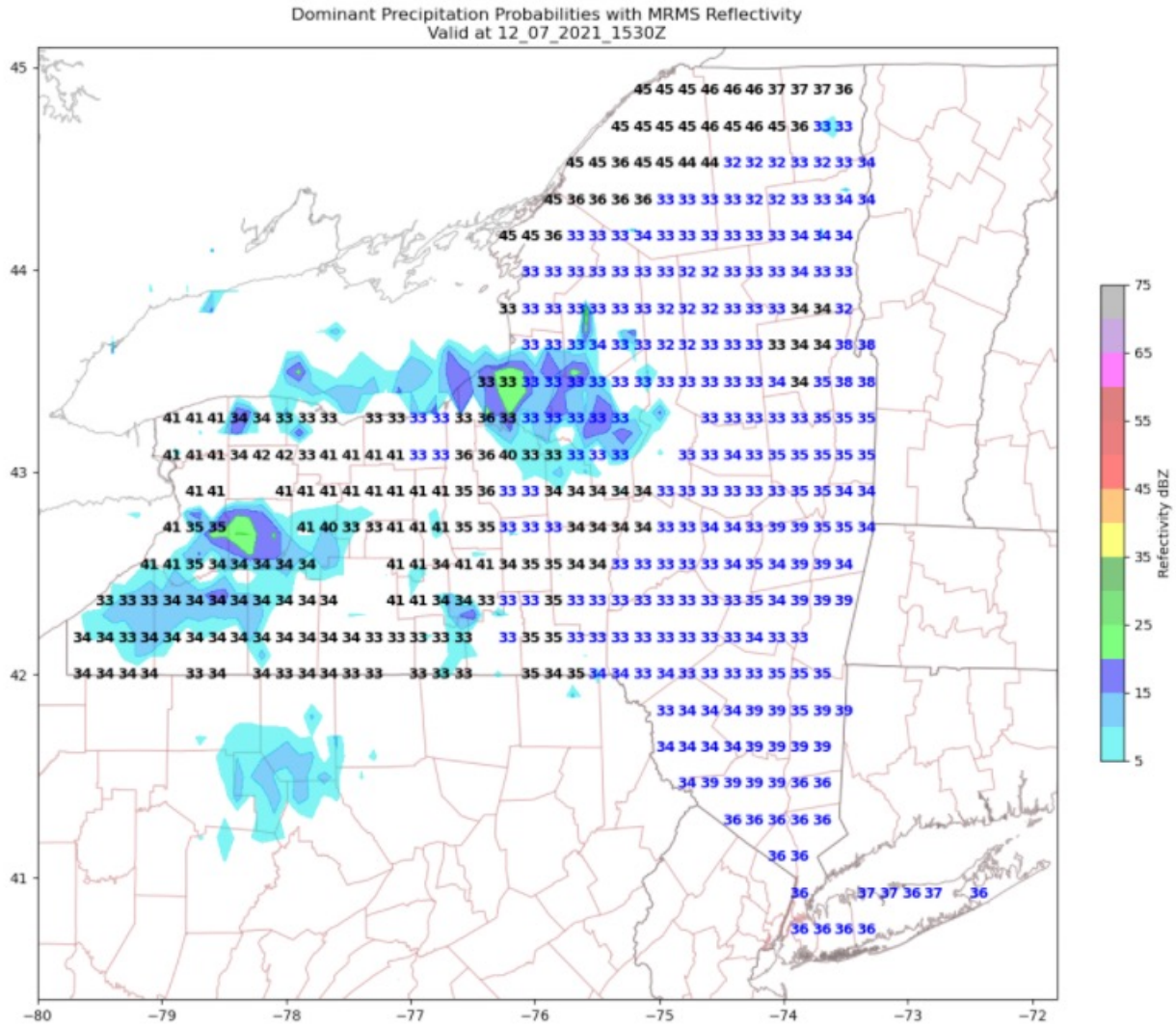
Dominant | Snow | Rain | Freezing Rain | Sleet | All Mixed

Top 10 Most Important Variables

temp_2m_min [degC]	temp_2m_avg [degC]	temp_2m_max [degC]	temp_2m [degC]	Geopotential Height at 500hpa	Temperature at 850hPa	Temperature at 700hPa	Temperature at 925hPa	700hpa Wet Bulb Temperature	850hpa Wet Bulb Temperature
0.07293	0.068204	0.06617	0.065254	0.032656	0.031072	0.02952	0.02824	0.024154	0.023014

Contact: Brian Filipiak bfilipiak@albany.edu

Dominant | Snow | Rain | Freezing Rain | Sleet | All Mixed



Black=Snow, Red=Freezing Rain, Purple= Rain, Blue=Sleet

NAMNEST Raw and Calculated Data

Home | NYSM & Upper Air | NAMNEST | Data and Methods | Training

Dec 7, 2021 01:31:24 UTC

******EXPERIMENTAL******

Please click any button to view the map of your choice

Dominant | Snow | Rain | Freezing Rain | Sleet | All Mixed

Feature Table for Each Forecast Hour

Contact: Brian Filipiak bfilipiak@albany.edu

NAMNEST Raw and Calculated Data

****EXPERIMENTAL****

Top 10 Most Important Variables

FH 00

Dewpoint at surface	Temperature at surface	Positive Area	Max Wet Bulb 925-700hPa	Surface Wet Bulb Temperature	Min Temperature surface-850hPa	Temperature at 850hPa	Max Temperature 850-700hPa	Mean Temperature surface to 850hpa	Dewpoint at 925hPa
0.05826	0.049744	0.0489	0.04838	0.04403	0.04156	0.034428	0.033562	0.033138	0.032682

FH 01

Dewpoint at surface	Temperature at surface	Positive Area	Max Wet Bulb 925-700hPa	Surface Wet Bulb Temperature	Min Temperature surface-850hPa	Temperature at 850hPa	Max Temperature 850-700hPa	Mean Temperature surface to 850hpa	Dewpoint at 925hPa
0.057708	0.04988	0.049002	0.047884	0.043892	0.041896	0.035088	0.03395	0.03296	0.032724

FH 02

Dewpoint at surface	Temperature at surface	Positive Area	Max Wet Bulb 925-700hPa	Surface Wet Bulb Temperature	Min Temperature surface-850hPa	Temperature at 850hPa	Max Temperature 850-700hPa	Mean Temperature surface to 850hpa	Dewpoint at 925hPa
0.05869	0.049522	0.04938	0.048358	0.04392	0.0419	0.03456	0.034166	0.033068	0.032248

FH 03

Dewpoint at surface	Temperature at surface	Positive Area	Max Wet Bulb 925-700hPa	Surface Wet Bulb Temperature	Min Temperature surface-850hPa	Temperature at 850hPa	Max Temperature 850-700hPa	Mean Temperature surface to 850hpa	Dewpoint at 925hPa
0.057878	0.050328	0.04967	0.047902	0.043498	0.041026	0.03461	0.034424	0.033648	0.03309

FH 04

Dewpoint at surface	Temperature at surface	Max Wet Bulb 925-700hPa	Positive Area	Surface Wet Bulb Temperature	Min Temperature surface-850hPa	Temperature at 850hPa	Max Temperature 850-700hPa	Mean Temperature surface to 850hpa	Dewpoint at 925hPa
0.058066	0.049712	0.04833	0.048046	0.043532	0.04244	0.03542	0.033908	0.033506	0.032206

FH 05

Dewpoint at surface	Temperature at surface	Positive Area	Max Wet Bulb 925-700hPa	Surface Wet Bulb Temperature	Min Temperature surface-850hPa	Max Temperature 850-700hPa	Temperature at 850hPa	Mean Temperature surface to 850hpa	Dewpoint at 925hPa
0.05804	0.049316	0.048892	0.048846	0.04378	0.041572	0.034822	0.034428	0.033132	0.033042

NAMNEST Raw and Calculated Data

Home | NYSM & Upper Air | NAMNEST | Data and Methods | Training

Dec 7, 2021 01:31:24 UTC

******EXPERIMENTAL******

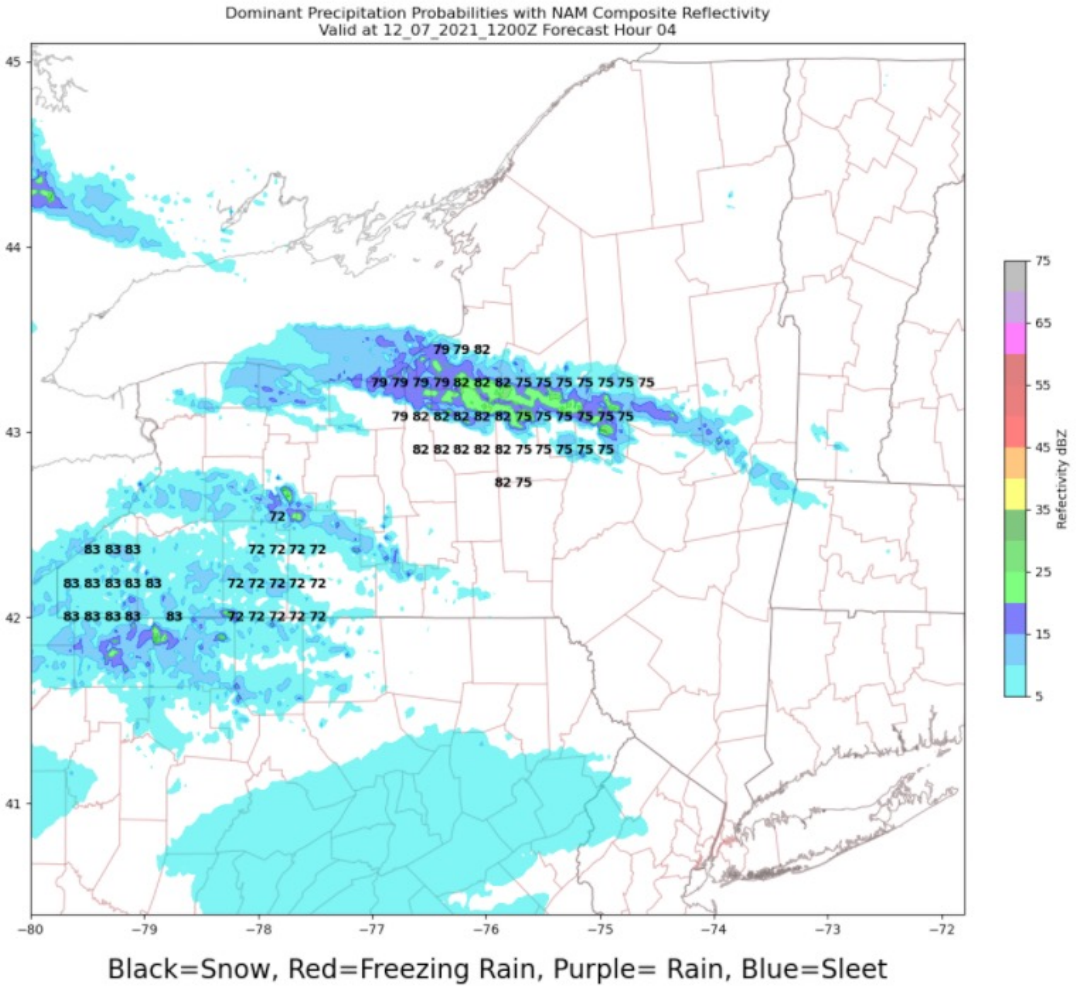
Please click any button to view the map of your choice

Dominant | Snow | Rain | Freezing Rain | Sleet | All Mixed

Feature Table for Each Forecast Hour

Contact: Brian Filipiak bfilipiak@albany.edu

Dominant | Snow | Rain | Freezing Rain | Sleet | All Mixed



Black=Snow, Red=Freezing Rain, Purple= Rain, Blue=Sleet

‹ ›

Feature Table for Each Forecast Hour

Future Work

- Add in additional data sources
- Understand importance of individual features on specific precipitation types
- Trial running reduced random forest with only most important variables
- Test over winter months

Conclusions

- Many sources of information when creating a forecast
- Random forests can accurately identify different precipitation types
- Specific data types and combinations of may need to be treated differently
- There will always be room for human interpretation of computer-generated guidance

Questions?

Contact: Brian Filipiak, bfilipiak@albany.edu

http://www.atmos.albany.edu/student/filipiak/op/