Probabilistic Forecasting of Winter Mixed Precipitation Types in New York State Utilizing a Random Forest

By

Brian C. Filipiak

A Thesis

Submitted to the University at Albany, State University of New York

In Partial Fulfillment of

the Requirements for the Degree of

Master of Science

College of Arts & Sciences

Department of Atmospheric and Environmental Sciences

Fall 2022

ABSTRACT

Operational forecasters face a plethora of challenges when making a forecast; they must consider multiple data sources ranging from radar and satellites to surface and upper air observations, to numerical weather prediction output. Forecasts must be done in a limited window of time, which adds an additional layer of difficulty to the task. These challenges are exacerbated by winter mixed precipitation events where slight differences in thermodynamic profiles or changes in terrain create different precipitation types across small areas. In addition to being difficult to forecast, mixed precipitation events can have large-scale impacts on our society.

To aid forecasts for these events, the goal of this thesis is to take the multiple data sources used by forecasters and combine them together using machine learning to improve forecasting ability for mixed precipitation events. The expectation is that by employing a machine learning framework, forecasters will have more time to spend analyzing the most difficult portions of the forecast.

In order to achieve this goal, Community Collaborative Rain, Hail and Snow Network (CoCoRaHS) daily observations from trained reporters between January 2017 to September 2020 were used to identify precipitation events that included rain, snow, freezing rain, and sleet. The data associated with the timing of these mixed precipitation events were collected from the New York State Mesonet, National Weather Service upper air soundings, High-Resolution Rapid Refresh model (HRRR), and North American Mesoscale forecast model with a nested domain (NAMNEST). A random forest (RF) machine learning algorithm was trained and tested on cases identified from the CoCoRaHS reports that were matched with the meteorological datasets.

Internal testing identified the best combinations of meteorological variables and data sources to make operational forecasts with.

An operational website was developed to display the products made with the output of the RF. The website was operational in real-time during the 2021-2022 winter season, and the RF was also run to create the same products for the 2020-2021 winter season. This was done to increase the sample size of the forecast guidance to do verification on. Verification was completed for the winters of 2020-2021 and 2021-2022 by using ASOS and mPING observations as ground truth. Results from the verification process gave a positive indication that certain products can provide accurate precipitation type forecasts to be employed in combination with analysis by an operational forecaster during winter mixed precipitation events. The operational products were expanded for the winter of 2022-2023, and the operational website currently has five operational forecasting tools, four of which provide probabilistic winter precipitation type forecasts for rain, freezing rain, sleet, and snow.

ACKNOWLEDGEMENTS

As my Master’s degree concludes, I would be remiss to not say thank you to the many people who have helped me on my journey. First, I would like to thank my entire advising team, Kristen Corbosiero, Andrea Lang, Nick Bassill, and Ross Lazear. Under their advisement for the last two and a half years, I have truly grown as student and researcher throughout all our conversations and meetings. They came up with idea of a “Data Fusion” project, and I know we all would agree this project exceeded our expectations. I am indebted to them for selecting me for this project, which I have enjoyed so much.

I also must thank all the professors, staff and individuals I interacted with in the Department of Atmospheric and Environmental Sciences, the Atmospheric Science Research Center, and the New York State Mesonet. I have learned so much from all the excellent classes I have taken, classes I have been a teaching assistant for, projects I have participated in, and conversations about all things related to the atmosphere and beyond. I would also like thank Kevin Tyle for his continued technological assistance, as well as all the conversations we had during my time in Albany.

My National Weather Service focal points, Christina Speciale and Neil Stuart, have provided me with an abundance of advice, suggestions, and recommendations throughout the thesis. One of my project goals was to make sure they were involved as much as possible; without them, this project would not have been as successful. I am grateful for their assistance across all aspects of this project.

A very special thank you to all my fellow graduate students, particularly to the cohort of 2020. We began our journey at Albany during one of the most challenging times in the world,

and through our shared classes and experiences have come out better for it. I am grateful to have colleagues who I can call my friends, and I look forward to seeing where we go from here.

Lastly, I would thank the people who provided support throughout my entire graduate school journey at Albany. My sincerest gratitude to my parents, Sara and Tim, my brother, Eric, my partner, Avery, and our dog, Lola, for all their love and support. They have all been by my side throughout my ups and downs in graduate school. Without them listening to my various presentations, reading my abstracts and papers, and being my voice of reason, I would not have been as successful as I have been during my time in Albany.

This research was funded by the National Oceanic and Atmospheric Administration (NOAA) grant NA19NWS4680006, awarded to the University at Albany, SUNY, as part of the Collaborative Science, Technology, and Applied Research (CSTAR) program.

TABLE OF CONTENTS

1. Introduction

1.1. Motivation

Winter weather hazards can hinder travel, utility operations, and day-to-day activities for individuals and businesses. Forecasting and communicating the impacts of winter storms, particularly on the East Coast of the United States, can be challenging due to complex terrain, continental–marine boundaries, and high-density population centers, which make accurate forecasts for these events essential (e.g., Ralph et al. 2005). Areas of mixed precipitation, defined here as freezing rain or sleet, embedded within larger storms or on their own, can enhance difficulties in forecasting, as different precipitation types can cause a wide range of hazards while potentially occurring in similar or adjacent meteorological environments. Differentiating between rain, freezing rain, sleet, and snow is essential to forecasting because of the unique hazards each one generates. In particular, freezing rain and heavy wet snow events can create extreme hazards that result in power failures (Theriault et al. 2022), damage to infrastructure (Changnon 2003), and significant travel issues, which are not often associated with sleet or cold rain events.

In the United States between 1949–2000, catastrophic ice storm events (events with losses totaling over $1 million) generated $16.7 billion in losses; in particular, the Northeast United States had the greatest number of these events with 39, causing over $4 billion in damage (Changnon 2003). New York State alone experienced 31 of the 39 (79%) events, with five to seven freezing rain days per year (Changnon 2003). Along with ice storms, the Northeast United States is susceptible to significant snowstorms. Between 1980 and 2021, 19 billion-dollar winter storm disaster events affected the Northeast Climate Region (Consumer Price Index-adjusted); these events totaled $79.8 billion in estimated costs (NOAA NCEI 2022).

Because significant damage and economic losses occur during mixed-precipitation type storms, accurate forecasts of precipitation types are essential for decision making and planning for organizations including city councils, transportation departments, public utilities, schools and universities, and many others. Accurate forecasts of precipitation type and timing can assist with decisions such as whether to pre-treat roads and how to allocate snowplow and road salt operations to the assignment of repair crews to areas where significant power outages may occur. Because these weather hazards are destructive, costly, and impact high-level decision making (such as school closures), accurate mixed-precipitation forecasts are vital to protect lives and property.

1.2. Precipitation Type Forecasting: Challenges and Methods

Precipitation type forecasts are challenging because slight variations in thermodynamic profiles and surface conditions can result in significant changes to weather conditions and impacts. The typical vertical temperature profiles for rain, snow, freezing rain, and sleet (Fig. 1.1) illustrate how slight differences in the vertical temperature profile can change the precipitation type. For example, minor changes in the depth of a near-surface freezing layer or an above freezing layer aloft can cause a change in precipitation type, such as rain to freezing rain or freezing rain to sleet. While each of the different environments for the precipitation types have a distinct profile, the environments can occur in close proximity and can be modified by background features in the environment like areas of complex terrain. When the different environments are close to each other spatially or the vertical profile resembles that of multiple precipitation types, the precipitation occurring at the surface can change quickly making forecasts challenging. These changes in precipitation type have implications on the potential impacts of a storm, which highlights the importance of accurate vertical thermodynamic profiles.

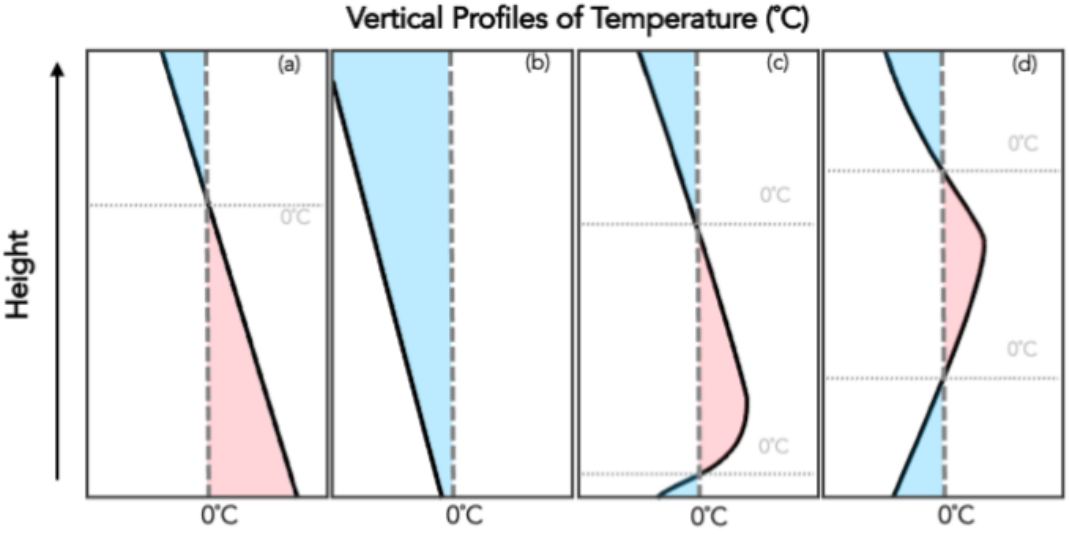


Fig. 1.1: Vertical temperature profiles for (a) rain, (b) snow, (c) freezing rain, and (d) sleet. Red (blue) shaded areas represent where the temperature is greater (less) than 0°C. The dotted horizontal gray lines represent intersection points between the profile and 0°C line.

Over the years, many methods have been developed to identify precipitation types both through implicit and post-processing algorithms. Methods range from considering the properties of the temperature and humidity profiles to the composition of falling hydrometeors to the use of model microphysical parameterizations to explicitly forecast precipitation type (Reeves 2016). These different methods present various solutions to solving the challenges around winter precipitation type forecasting, especially as many of these methods are currently being used in operational forecasting settings.

The Baldwin algorithms (Baldwin et al. 1994) are based on vertical profiles from both observed and numerical model soundings for select storms. There are two separate algorithms, Baldwin1 and Baldwin2, which only differ in the step for distinguishing snow and ice pellets (sleet in this thesis). Both methods first identify a precipitation generation layer and then calculate the area between the wet-bulb temperature profile and the 273K or 269K isotherms that are below the precipitation layer. To decipher the precipitation type, the size of the areas is used to decide if melting aloft or refreezing is possible. Both these methods can diagnose four categories of precipitation: rain, freezing rain, ice pellets, and snow.

The Ramer algorithm (Ramer 1993) is similar to the Baldwin algorithms in that it also uses observed and numerical model soundings as its basis dataset. Ramer is a statistically derived algorithm that uses a top-down method which follows a hydrometeor through the atmosphere from where it would be generated to the surface. To decide between precipitation types, the ice fraction of the hydrometeors in each layer is determined from the wet-bulb temperature and relative humidity. If the starting wet-bulb temperature is greater than 266.55K, it is assumed that the starting hydrometeor is liquid. As the hydrometeor is followed down through the atmosphere, the value of the ice fraction is used to determine rain, freezing rain, ice pellets, and snow.

The Bourgoin algorithm (Bourgouin 2000) was created by using 173 observed soundings, which differentiates it from Ramer and Baldwin as those use model soundings. Bourgouin determines precipitation type by examining the observed temperature profile in comparison to the 0°C isotherm. If a profile does not cross the 0°C isotherm, then snow is identified because the hydrometeor is assumed to start as frozen; otherwise, the number of times the 0°C isotherm is crossed, as well as how long a hydrometeor spends on each side of 0°C isotherm, decides precipitation type. To determine precipitation, the area of the temperature on each side of the 0°C isotherm can be estimated by multiplying the mean temperature of a layer by the height of that layer. The area to the left (right) of the 0°C isotherm is labeled a negative (positive) area. Rain will have only one positive area identified since it only crosses the 0°C isotherm once. To determine freezing rain and ice pellets, the ratio of negative to positive area and the relative size of the two areas are used.

Schuur et al. (2012) (also known as National Severe Storms Laboratory, NSSL, algorithm) predicts precipitation type through examining the wet-bulb temperature below the precipitation generation layer. It uses four distinct sounding types in determining precipitation type through the wet-bulb temperature profile: 1) If the temperature profile is completely below the 0°C isotherm, snow will be diagnosed; 2) if the profile crosses the 0°C isotherm once, then either snow or rain will be predicted based on the height of the crossing point; 3) if the profile crosses the 0°C isotherm three times, ice pellets or freezing rain will be diagnosed; and, 4) if the profile crosses the 0°C isotherm two times, ice pellets or freezing rain will be predicted. To determine between ice pellets and freezing rain in sounding types 3 and 4, the maximum wet-bulb temperature aloft and the minimum wet-bulb temperature in the lowest layer that has temperatures below 0°C are examined.

While these four precipitation type forecasting methods are post-processing numerical model profiles or for observed soundings, numerical models have their own separate precipitation type determination methods. Manikin (2005) describes the North American Mesoscale (NAM) Forecast System method for determining precipitation type. Before 2003, the NAM used Baldwin1 as its method to determine precipitation type. Forecasters indicated that there was a potential bias in this algorithm, so in 2003 the NAM moved to a mini-ensemble for predicting precipitation type. This mini-ensemble encompasses multiple precipitation type algorithms (Baldwin1, Baldwin2, Bourgouin, and Ramer) available to generate a consensus result. This method combines individual precipitation type methods to potentially eliminate biases.

The High-Resolution Rapid Refresh (HRRR) model has also switched between different explicit precipitation forecasting schemes. Ikeda et al. (2013) presented potential biases including a warm (cold) bias in surface temperature when mixed precipitation was observed and caused rain (snow) to be forecasted. This impacted the HRRR's precipitation type algorithm (operational between January 2011 and late 2015), which was based on a logic table that made decisions using the snow to rain ratios and snow, graupel, and rain mixing ratios from the microphysics scheme as well as model surface temperature. An updated precipitation type scheme (Benjamin et al. 2016a) was developed in response to Ikeda et al. (2013). The new approach uses the cloud microphysics parameterizations and three-dimensional hydrometeor mixing ratios at the ground combined with a flowchart to determine precipitation type. This algorithm allows for the prediction of more than one precipitation type at a time.

Even with all the available methods, it is still difficult to consistently produce accurate precipitation type predictions, especially when events include mixed-phase precipitation like freezing rain and sleet (Manikin 2005; Wandishin et al. 2005; Reeves et al. 2014; Ikeda et al. 2017). Reeves et al. (2014) compared the forecasts of the Baldwin1, Baldwin2, Bourgouin, NSSL, and Ramer algorithms. Using Heidke skill scores (HSSs), Reeves et al. (2014) showed that snow and rain were well forecasted with all methods, but many methods struggle with mixed precipitation types (freezing rain and ice pellets). None of the algorithms had HSSs over 0.332 for ice pellets, while only Ramer was above 0.5 in the freezing rain category. Reeves et al. (2014) also showed that the algorithms had a high level of detecting rain or snow, with probability of detection (POD) values ranging from 96.1% to 99.6% for rain and 86.7% to 94.9% for snow. For the mixed precipitation categories, the POD values were significantly lower. Values for predicting the combination of ice pellets and freezing rain ranged from 34.7% to 77% depending on the method. Biases in these algorithms leave forecasters with no clear preferred method. The Baldwin algorithms has biases toward predicting ice pellets, while Bourgouin is limited by its use of the temperature profile, as opposed to the wet-bulb temperature profile, and its limited base dataset that derived the threshold for freezing rain and ice pellets. NSSL tends to favor a mix of ice pellets and freezing rain as opposed to either of them individually. Ramer is sensitive to the temperature threshold of 266.55K and experiments show the POD values change significantly if the threshold temperature is shifted (Reeves et al. 2014).

Whereas Reeves et al. (2014) focused on evaluating precipitation type algorithms that are generally used to post-process observational or model-derived soundings, other studies have focused on how models evaluate precipitation types with different precipitation algorithms. Wandishin et al. (2005) evaluated five different precipitation type algorithms, including Ramer, Baldwin2, and Bourgouin, through different experiments with the Eta Model (precursor to NAM). They noted that none of the algorithms used were a consensus "best" algorithm as individual algorithms performed better in certain situations. An additional challenge when evaluating precipitation type algorithms in numerical models is the fact that the algorithm must be evaluated as well as the model accuracy of the local meteorological conditions verified. Temperature biases in the model can impact the forecast precipitation type (Ikeda et al. 2017).

Precipitation type algorithms continue to be researched because there is still much room for improvement. Birk et al. (2021) revised the Bourgouin algorithm based on results from previous research evaluation (Reeves et al. 2014) and operational forecasting experiences. To improve upon Bourgouin, Birk et al. (2021) used a larger developmental dataset and wet-bulb temperature profiles in conjunction with allowing the algorithm to diagnose freezing precipitation from situations without ice nucleation. The output of this updated method allowed for probabilistic precipitation type forecasts and combinations of winter precipitation types. Their revised method improved upon the original Bourgouin method by 0.17 in critical success index (CSI) and 0.17 in HSS. In addition, each of the individual precipitation type values for CSI and HSS increased compared to the original algorithm. They also stated that the probabilistic output helps to offset the non-inclusion of variables like precipitation rate or other hydrometeor characteristics.

In summary, previous analyses of precipitation type forecasting methods resulted in important information about their accuracy and biases (e.g., Bourgouin 2000; Reeves et al. 2014; Reeves 2016; McCray et al. 2019; Birk et al. 2021; Ellis et al. 2022), but there is no consensus on which precipitation type identification method is the most accurate. This uncertainty presents an ongoing challenge as researchers and operational meteorologists attempt to accurately forecast precipitation type.

1.3. Random Forest Applications in Operational Forecasts

Machine learning (ML) tools are increasingly used in the earth sciences to examine complex problems. Atmospheric science is no exception, and these tools are being utilized to solve problems and process significant volumes of data that previously were too large for analysis systems to process. ML has recently been used in several critical atmospheric applications such as quantitative precipitation forecasts and forecasting of flooding events (Gagne et al. 2014; Herman and Schumacher 2018a,b; Erickson et al. 2019), hail and severe weather prediction (Gagne et al. 2017; Hill et al. 2020), and predicting visibility at airports (Herman and Schumacher 2016).

One popular technique for ML applications are random forests (RFs) (Breiman 2001). RFs can assist with the challenges in operational weather forecasting as they have been shown to be successfully implemented in many of the applications listed above (McGovern et al. 2017, 2019). RF was selected as the ML method in this study because of its ability to handle large datasets (McGovern et al. 2017), the popularity of subjective (human-derived) decision trees in weather forecasting (McGovern et al. 2017), and the ease of explanation to end users, which is important when transitioning the algorithm to operations.

Herman and Schumacher (2018a,b) described how a RF could successfully be used to predict extreme rainfall events across the continental United States (CONUS). Herman and Schumacher (2018a) focused on the development of a RF to predict extreme rainfall events over a 1- and 10-year average recurrence intervals (ARIs) two and three days in advance. They utilized 11 years of NOAA's Second Generation Global Ensemble Forecast System Reforecast (GEFS/R) to train the RF. They also compared the RF directly to model output from GEFS/R and a logistic regression method that used the same data as the RF. When comparing the final model forecasts, they found that the RF algorithm performed significantly better than the raw GEFS/R model across all ARIs and forecast periods. In some cases, the forecast skill nearly doubled for the RF compared to the GEFS/R (Herman and Schumacher 2018a). Herman and Schumacher (2018b) examined the impact of the type of model that data was passed into in Herman and Schumacher (2018a). They used a RF, logistic regression, and raw GEFS/R as the three base models. Herman and Schumacher (2018b) also looked at the impact of reduced dimensionality by techniques such as principal component analysis (PCA). The results were important to understanding the impact of data and the base model when predicting extreme rainfall. Both the logistic regression and RF identified key variables that a forecaster would look for. This result can give forecasters confidence that these algorithms make physical sense. They also discussed how work to remove the "black box" around RF or other ML algorithms is important because it allows a forecaster to understand the potential limitations and pitfalls that come with these algorithms. The work in Herman and Schumacher (2018a,b) was applied in the Weather Prediction Center's Flash Flood and Intense Rainfall experiment (Erickson et al. 2019), which shows there is interest in putting ML algorithms into operational forecasters' hands.

Another successful implementation of a RF into operational weather forecasting used a RF to generate improvements of severe weather forecasts made by the Storm Prediction Center (SPC) (Hill et al. 2020). A similar setup to Herman and Schumacher (2018a,b) was implemented with 11 years of GEFS/R data used to train a RF-based ML algorithm to predict the same categorical outlooks that the SPC produces, including Day 1 probabilistic outlooks of tornadoes, significant tornadoes, hail, significant hail, wind, and significant wind, as well as Day 2 and 3 total severe probabilities. This setup allowed for direct comparison with SPC forecasts in addition to examining how the forecaster and RF algorithm would work together. The RF was able to identify key features in the dataset that would be similar to what operational forecasters would look for while also pointing out relationships that were not as clear. The produced outlook forecasts had considerable skill by outperforming the SPC outlooks on Days 2 and 3; they underperformed slightly when compared to SPC outlooks on Day 1. The weighted blend of forecasts made using both SPC forecasts and the RF algorithm outperformed SPC-only outlooks on all products and at all lead times (Hill et al. 2020). This work proves that combining effective ML algorithms with operational forecasters can provide a boost to operational forecasts.

While ML is becoming a more common weather forecasting tool and has been proven to be successful, one less studied application is highly-impactful winter weather events. Considering the difficulties associated with forecasting winter precipitation types as described in Section 1.2, the prospect of combining ML with the challenge of forecasting precipitation type represents an exciting and potentially effective forecasting tool that could be integrated into the operational forecasting process.

1.4. Research Objectives

This thesis will examine the development of a RF-based winter precipitation type forecasting algorithm across New York state. Since winter precipitation type forecasting can create a number of challenges for forecasters, there is the opportunity to provide significant improvement to precipitation type forecasts based on previous implementation of RF algorithms. RFs also provide important information on feature importance which can give confidence to forecasters, as well as additional understanding on how the algorithm is functioning and making predictions. The key research goals for this thesis are:

- Develop a RF that can accurately predict winter precipitation type across New York
- Use the base RF structure developed to expand the RF's input data sources to encompass both observational and model datasets
- Create and maintain a suite of operational probabilistic winter precipitation type products available for forecasters across New York
- Evaluate different RFs' success at identifying winter precipitation types to convey potential strengths and limitations of the available product suite

Chapter 2 describes the data and RF methodology for training and validation. Chapter 3 explains the validation process of the RF through internal testing, which will be the basis for the operational RF forecast. Chapter 4 discusses challenges and hurdles faced during the transition from a research RF to an operational RF, suggests a flowchart to streamline the process, and outlines the necessary collaboration to develop a successful ML operation product. In addition, a framework for applying ML to operational forecasting that can be translated to other locations and weather types will be detailed. Chapter 5 analyzes the effectiveness of the operational probabilistic forecasts over the winters of 2020–2021, 2021–2022 and for select individual events. Chapter 6 summarizes the overall findings from the thesis as well as present suggestions for future work and expansion of the current operation products being made.

2. Data and Random Forest Methodology

2.1. Data Collection and Processing

2.1.1. Random Forest Datasets

2.1.1.1. Training Dataset Case List

To successfully create a RF, the basis for the training dataset needs to be determined first. While there is not a complete archive for winter mixed precipitation events, there are several options to use as ground truth observations in a training dataset. One option is data from Automated Surface Observing Systems (ASOS), which is logical because ASOS stations have present weather sensors to detect precipitation types and some stations are augmented by trained observers who can change precipitation type reports as necessary (NOAA 1998). ASOS locations are primarily at airports, which means they are not representative of the complex terrain in a given region. This lack of representation is an important consideration as complex terrain, including mountains and valleys, can modify conditions locally and alter precipitation type. In addition to terrain challenges, prior research has indicated that non-augmented stations can have biases in identifying precipitation types like sleet (Reeves 2016).

Another option for ground truth observations is the Meteorological Phenomena Identification Near the Ground (mPING; Elmore et al. 2014) dataset. mPING reports are precipitation type observations submitted by anyone with the mobile app on their phone or tablet; this means the reports can cover a much wider area than ASOS stations. One downside to mPING is that reports are reliant on people having and correctly using the application, so reports may be more sporadic than desired. In addition, the observer making the report may not have a background in meteorological observations and, while there are online resources to help observers make decisions on precipitation type, there is no guarantee it will be reported accurately.

A third option for ground truth observations is the Community Collaborative Rain, Hail, and Snow Network (CoCoRaHS; Cifelli et al. 2005), which is a volunteer network of trained observers who report daily precipitation reports across the country. All volunteers are trained how to report and measure different precipitation types. While these daily reports generally do not record the exact timing of precipitation like mPING, the notes section of these reports are filled with information about the time and precipitation type. However, since not all observer notes are the same, only certain reports are useful to identify precipitation timing.

While there are many potentially useful options to develop a training dataset, CoCoRaHS reports were chosen to identify cases for the training dataset because they use trained and consistent observers, have a large spatial distribution, and collect reports of all precipitation types across a variety of terrain. The training dataset development began with CoCoRaHS reports between January 2017 and September 2020 for four precipitation types: rain, freezing rain, sleet, and snow. Once all the reports were obtained, the notes section of the reports were individually reviewed and subjectively verified using New York State Mesonet (NYSM; Brotzge et al. 2020) standard station weather data, NEXRAD radars, and Weather Prediction Center surface analyses. The verification process ensured the meteorological conditions around the report location were commensurate with the reported observation.

Since CoCoRaHS reports do not have a specific precipitation type reporting section, the notes section of the reports was used to identify and categorize cases. Precipitation type, timing, and uncertainty were recorded for each individual report. To classify the reports, a qualitative coding classification was performed using a scale from 1 to 4, with 1 being the most informative reports and 4 being the least informative reports (see Table 2.1 for examples). Specific times were the most helpful for determining event timing and most reports that had specific times were classified as Category 1 reports. Other reports used terms and phrases that indicated timing with less certainty, including phrases like “around 9 PM” or “in the early morning”. These notes provided a moderate amount of confidence and were classified as Category 2 or 3 reports, depending on the event type and the meteorological information available. Reports that gave no information regarding timing or contained irrelevant information were classified as Category 4 reports.

Classification Category	CoCoRaHS Report Notes
1	• 10 min. snow flurry 8:20 a.m. yesterday. Freezing rain began around 7 p.m. Dusting snow around 9:30 p.m. Raining at obs. time - 32 degrees. Ice on tree branches but no wind or storm damage. Hard to estimate snow fall depth. • 27F and sleet at obs time. Little hard pellets, accumulation recorded under new snowfall. It's not clear when this started - during the night.
2	• Some sleet just before observation. Intermittent showers only. • 20F at obs time. Precip started as freezing rain and sleet about 3 pm then changed to snow. • Light snow began at 2pm and sleet began to mix in at 5pm to all sleet by 6pm to all rain by 7pm.
3	• Sleet on and off overnight but only a trace on the ground • Mixed precipitation event, with snow mixing with sleet during the day. Temperatures rose in the evening, with snow changing to rain for several hours.

4	• A combination of sleet, freezing rain and rain. 47 degrees this morning! • Measured 1.48 inches of rain before the change over to snow. There was some sleet and freezing rain in my collector, but could not get a proper measurement of that. Included in snow measurement. • Yesterday was a mostly grey day no precipitation. Temps were around the freezing point. Right now is cloudy with very little wind and warmer temps … Have not seen the crows yet - they usually are flying overhead by this time. I did see a rabbit this morning right before dawn and I heard a chickadee and saw a squirrel in a tree at 8 am. Wind advisory for tonight-trash day is tomorrow :(

Table 2.1. Examples of CoCoRaHS reports and the qualitative scoring used to categorize their usefulness for this study. Category 1 reports included specific information about time of precipitation while Category 4 reports included no information about the time of precipitation.

To ensure the CoCoRaHS reports covered all of New York, the rain and snow reports were sorted by nearest NYSM site, and only those cases classified as Category 1 reports were kept due to the significantly larger volume of CoCoRaHS reports for rain and snow compared to freezing rain and sleet. The final training dataset includes 2617 viable training dataset cases: 750 rain, 750 snow, 619 sleet, and 498 freezing rain.

Once the final training dataset cases were selected, those cases need to be matched with meteorological data that would be used in the RF. This process is described more in section 2.2. The main data sources used to match with the training dataset cases were NYSM standard site data for surface observations, in-situ radiosonde data, NAM BUFKIT (BUFfalo toolKIT; Mahoney and Niziol 1997) profiles, and HRRR (High-Resolution Rapid Refresh; Benjamin et al. 2016b) model data.

2.1.1.2. New York State Mesonet (NYSM)

Surface observations in this study come from the NYSM (Brotzge et al. 2020), a high-quality network of weather stations installed in New York State between 2015 and 2018. The network consists of 126 "standard" sites (used for this analysis) as well as a variety of specialized sub-networks including profiler, flux, and snow networks. Standard sites are evenly distributed throughout the state and measure temperature at two heights, relative humidity, redundant wind speed and direction measurements, snow depth, irradiance, precipitation, soil temperature and moisture at three depths, and surface pressure. Each site is also equipped with a camera. Data are collected, archived, and disseminated every five minutes and undergo a series of automatic and manual quality control procedures. A dedicated team of field technicians perform regular maintenance on all sites to ensure data quality. Table 2.2 details the variables from the NYSM that are used in the surface variable dataset.

NYSM 5-min data	**NYSM Hourly Variables**
• 2-m temperature • 2-m relative humidity • Surface pressure • Solar irradiance • Precipitation (5-min total, daily total, intensity) • 10-m average wind speed and direction from sonic anemometer	• 2-m temperature (minimum, maximum, average) • Relative humidity (minimum, maximum, average) • Station pressure (minimum, maximum, average) • Solar irradiance and total solar irradiance • Precipitation (hourly total, daily total, intensity) • 10-m average wind speed and direction from sonic anemometer

Table 2.2. Variables from each NYSM dataset used in the RF.

2.1.1.3. Vertical Profile Datasets

Complete profiles of the lower and middle troposphere are crucial for making precipitation type predictions. On the observational side, in-situ radiosonde data was used to complement the NYSM standard site data. Radiosonde data was collected from four sites in or near New York State at 0000 and 1200 UTC daily (Albany, Buffalo, and Upton in NY, and Maniwaki, QC). CoCoRaHS reports were matched to the nearest and most recent radiosonde launch (i.e., if a CoCoRaHS report was at 0600 UTC, the report would be matched to the 0000 UTC launch).

As described in Mahoney and Niziol (1997), BUFKIT profiles were used to create a dataset of forecast vertical profiles. The NAM nested domain (NAMNEST) was selected for these profiles as it offered the highest resolution of the available models that have BUFKIT archives throughout the period of study. The NAMNEST is a 3-km grid spacing (4-km prior to March 2017, representing up to two months of CoCoRaHS reports) nested domain of the larger 12-km NAM. The model is initialized every six hours with hourly model output and 60 vertical levels, with 27 levels in the lowest 3 km starting at 20 m. The BUFKIT program generates vertical profiles from model forecast data that have the same data structure as radiosondes. The CoCoRaHS reports were matched to the most recent NAMNEST profile, so a report at 1000 UTC would be matched to forecast hour 4 from the 0600 UTC NAMNEST simulation.

Raw Variables (C1)	Original Calculated Variables (C2)	New Calculated Variables (C3)
• Temperature • Pressure • Dew point • Wind speed and direction • Geopotential height • Wet bulb temperature • Relative humidity	• Temperature difference between standard pressure levels • Precipitable water vapor difference between standard pressure levels • Wind speed and direction difference between standard pressure levels • Critical thickness (sea level –850 hPa and sea level–500 hPa)	• Max wet bulb Temperature 925–700 hPa • Positive and negative areas and ratio of positive to negative (Bourgouin 2000) • Critical thickness- (850–700 hPa and 700–500 hPa) • Mean relative humidity sea level–500 hPa • Dew point depression • Mean temperature- (sea level–850 hPa and sea level–700 hPa) • Minimum temperature sea level–850 hPa • Maximum temperature 850–700 hPa

Table 2.3. Variables used in the RF from the NAMNEST and in-situ radiosonde vertical profile datasets. Raw Variables (C1) are at standard pressure levels including the surface, 925, 850, 700, and 500 hPa). All calculated variables (C2, C3) were found between standard pressure levels unless otherwise noted.

The NYSM, in-situ radiosondes, and NAMNEST datasets used in the RF combine raw variables with calculated variables based on the raw data available from in-situ radiosonde data or model output. NYSM data was supplemented with derived sea-level pressure using existing NYSM data and metadata. For radiosonde and NAMNEST profiles, numerous variables were calculated to give additional information, including wet-bulb temperature, precipitable water vapor, and calculations of raw variables between standard pressure levels. Table 2.3 details all the variables used in the vertical profile datasets (in-situ radiosondes and NAMNEST profiles).

As described in Benjamin et al. (2016b), the HRRR is a 3-km grid spacing forecast model that is initialized every hour with hourly model output and 51 vertical levels. While BUFKIT profiles are available for the HRRR, they do not exist for the entirety of the training dataset cases. Instead, information about the lower and middle troposphere from HRRR was collected on a 40-km grid across New York, which generates greater than three times more vertical profiles for New York (107 HRRR vertical soundings vs 32 NAMNEST soundings). Since the HRRR is initialized every hour, CoCoRaHS reports were matched to the most recent HRRR run: a report at 1000 UTC would be matched to forecast hour 1 from the 0900 UTC HRRR simulation.

HRRR pressure files were used to match with the training dataset cases as well. These pressure files contain variables at every pressure level from 1000 hPa to 50 hPa in 25-hPa increments. In addition to the variables at pressure levels, other variables, including ones located at the surface, sub surface, and upper atmosphere, were available. Table 2.4 displays the different vertical data combinations for the HRRR.

All HRRR Variables (H1)	Reduced HRRR Variables (H2)	Simplified HRRR Variables (H3)
• Temperature • Dew point • Wind speed and direction (u, v components, and magnitude and direction) • Vertical velocity • Geopotential height • Wet bulb temperature • Relative humidity • 2-m temperature • 2-m potential temperature • 2-m wet bulb temperature • 2-m dewpoint • 10-m wind speed and direction (u, v components, and magnitude and direction) • Max wet bulb Temperature 925–700 hPa • Critical thickness (975 hPa–850 hPa, 975 hPa–700 hPa and 850 hPa–700 hPa) • Temperature difference between standard pressure levels • Precipitable water vapor difference between standard pressure levels • Wind speed and direction difference between standard pressure levels • Mean relative humidity sea level–700 hPa • Dew point depression between standard pressure levels	• Temperature • Dew point • Wind speed and direction • Vertical velocity • Geopotential height • Wet bulb temperature • Relative humidity • 2-m temperature • 2-m potential temperature • 2-m wet bulb temperature • 2-m dewpoint • 10-m wind speed and direction • Max wet bulb Temperature 925–700 hPa • Critical thickness (975 hPa–850 hPa, 975 hPa–700 hPa and 850 hPa–700 hPa) • Temperature difference between standard pressure levels • Precipitable water vapor difference between standard pressure levels • Wind speed and direction difference between standard pressure levels • Mean relative humidity sea level–700 hPa • Dew point depression between	• *Temperature* • *Dew point* • *Wind speed and direction* • *Vertical velocity* • *Geopotential height* • *Wet bulb temperature* • *Relative humidity* • 2-m temperature • 2-m potential temperature • 2-m wet bulb temperature • 2-m dewpoint • 10-m wind speed and direction • Max wet bulb temperature 925–700 hPa • Critical thickness (975 hPa–850 hPa, 975 hPa–700 hPa and 850 hPa–700 hPa) • Mean relative humidity sea level–700 hPa • Dew point depression between standard pressure levels • Mean temperature- (sea level–850 hPa and sea level–700 hPa) • Minimum temperature sea level–850 hPa

• Mean temperature (sea level–850 hPa and sea level–700 hPa) • Minimum temperature sea level–850 hPa • Maximum temperature 850–700 hPa • Pressure at highest freezing level • Height at highest freezing level • RH at highest freezing level • Pressure at 0°C isotherm • Height at 0°C isotherm • RH at 0°C isotherm • Precipitable water vapor in entire atmospheric column	standard pressure levels • Mean temperature (sea level–850 hPa and sea level–700 hPa) • Minimum temperature sea level–850 hPa • Maximum temperature 850–700 hPa • Height at highest freezing level • RH at highest freezing level • Height at 0°C Isotherm • RH at 0°C isotherm • Precipitable water vapor in entire atmospheric column	• Maximum temperature 850–700 hPa • Pressure at highest freezing level • Height at highest freezing level • RH at highest freezing level • Pressure at 0°C isotherm • Height at 0°C isotherm • RH at 0°C isotherm • Precipitable water vapor in entire atmospheric column

Table 2.4. Variables used in the RF from the HRRR vertical profile datasets. Variables are at all pressure levels between 975 and 700 hPa unless otherwise noted. The number of variables in Simplified HRRR Variables (H3) is significantly reduced compared to the other two columns (H1, H2) because the italicized variables are only kept at 700, 850, 925, 950, and 975 hPa.

2.1.2. Verification Datasets

To verify the results of the RF output, a similar argument to Section 2.1.1.1. can be made. ASOS, mPING, and CoCoRaHS reports are potential options to use as ground truth for verification because they all give precipitation type reports with the associated timing. One difference between verification and training dataset development is that more reports across a variety of locations will give more confidence to the verification results. Another consideration is how other studies have verified winter precipitation forecasts because, if the same verification report sources are used, comparison can be made between studies and forecasting methods. For these reasons, ASOS and mPING reports were used to verify the RF. Combined, they provide good spatial coverage with a high enough volume to produce multiple reports from across areas of interest during precipitation events.

2.2. Random Forest Development and Configuration

RFs are a type of supervised ML consisting of an ensemble of individual decision trees trained on an example set of data. The RF is then given a separate, testing dataset and each tree votes for the most popular class based on the predictors it was given (Breiman 2001; McGovern et al. 2017). The relative frequencies of the votes in the ensemble of decision trees create the probabilistic forecasts for each class being predicted by the RF (Herman and Schumacher 2018a). A higher number of trees in the RF increases the diversity of the decision trees because of the different combinations of data used to train and make predictions (McGovern et al. 2017; Hill et al. 2020).

Looking at the internal process for one decision tree in the RF, the trained decision tree processes the testing data by making decisions at nodes (points where the testing data is compared to the training data for a specific variable) and, once split, the separated testing data feeds into different branches that go to another node. This process occurs continuously until the testing data has been totally separated into individual classes or there are too few cases left to split (Hill et al. 2020). At this point, a vote is made for the most popular class by the tree. In the case of the different winter precipitation types, the RF is attempting to isolate the four classes of precipitation (rain, freezing rain, sleet, and snow) in the testing data.

Once all the observed and simulated data was collected, processed, and matched with the CoCoRaHS reports, the reports and associated data were combined to create different, unique data combinations with four distinct datasets: NYSM and upper-air data, NAMNEST vertical profiles, HRRR vertical profiles, and a NYSM surface observation with HRRR vertical profiles.

These combinations were tested in the RF to determine what would be the final training data for the operational RF.

The RF was then configured using a significant hyperparameter tuning process (the process of determining the optimal combination of parameters that control how the RF is structured), with cross validation to thoroughly test the setup. Initial testing used 500 decisions trees and kept the default RF options from the python scikit-learn package (Version 1.1.2, Pedregosa et al. 2011). Once the best datasets were selected, through an examination of which had the best internal statistics (accuracy and F1 scores), the hyperparameter tuning process was complete. A random grid search with 10-fold cross validation was conducted with 150 iterations, which was done across a wide range of values for all the parameters. The process was run five times to get multiple grid outputs; since there were a range of results, a full grid search with 10-fold cross validation was completed over the narrower range of options from the random search. The result from that search is the parameter configuration that was used in the full RF (Table 2.5).

Parameter	**Value**
Number of decision trees (N)	650
Minimum number of samples to split at a node (min_samples_split)	10
Minimum number of samples to be at a leaf node (min_samples_leaf)	1
Number of features to consider for best split (max_features)	‘log2’
Maximum depth of a decision tree (max_depth)	25
Bootstrap samples (bootstrap)	‘True’

Table 2.5. Random forest parameter configuration determined from the hyperparameter tuning process.

2.3. Random Forest Evaluation Methodology

2.3.1. Evaluation of Internal Training Datasets

After conducting the random grid and full grid searches, a full internal testing of the RF was performed on the training dataset to evaluate the RF performance with winter precipitation type classification of past events. This testing included a random split of the original training dataset into subsets creating a training dataset (75% of original training dataset) and a testing dataset (25% of original training dataset). Since the number of CoCoRaHS reports were not equal, these datasets were split such that the proportion of each type of precipitation report was equal in both datasets. This internal testing considered four key metrics: accuracy, precision, recall, and F1 Score (Fig. 2.1). Accuracy indicated the overall number of correct predictions out of the total predictions of the RF; precision was the number of correct predictions divided by the number of total predictions for that precipitation type; recall was how often the correct prediction occurs in the RF; and F1 Score was the combination of precision and recall and represents how well the RF is predicting that precipitation type. Section 3 focuses on accuracy and F1 Scores because they represent the overall RF success and how well individual precipitation types were predicted. These metrics were calculated for each run of the RF and the numbers described later were averaged over 50 independent RF runs.

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

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

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

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

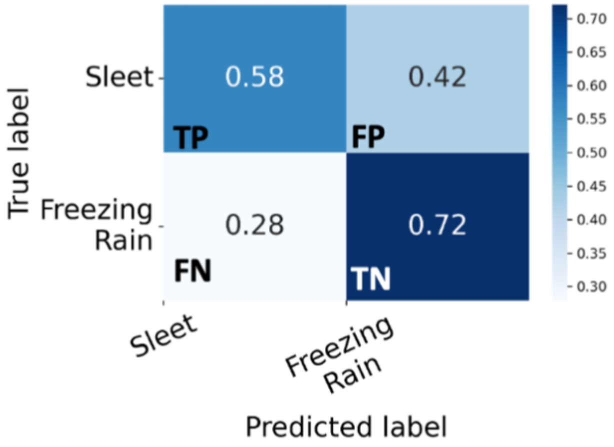


Fig. 2.1. RF evaluation metrics and a sample confusion matrix for sleet and freezing rain with the prediction of sleet set as the true positive. Evaluation metrics can be calculated from the case distribution in the confusion matrix, which shows the proportions of correct and incorrect predictions.

Confusion matrices were used to evaluate the outcomes from the RF predictions (Figs. 2.1). The diagonal from the upper-left to the bottom-right corner of a confusion matrix indicates the correct predictions for each precipitation type (between 0 and 1, with 1 equal to 100%). The off-diagonal values are important when evaluating the RF as they can elucidate the scenarios when the RF made incorrect predictions, thereby allowing for corrections and necessary changes to the RF.

Chapter 3 will highlight which features (data variables) were most important in the running and decision-making of the RF. To determine feature importance, the method of impurity importance as described in Breiman (2001) and McGovern et al. (2019) was used. Importance is determined using this method by how well a decision at a node isolates the known training cases in the RF. In the example of winter precipitation types, the more a decision at a node splits one precipitation type out from the rest, the more important it is as a feature. Examining the feature importance can play a large role in understanding the decision-making process of the RF.

2.3.2. Evaluation of Operational Forecasts

To evaluate the operational forecasts made by the RF, forecasts from 1 November to 31 March were compared to ASOS and mPING reports at each of the valid forecast times. The ASOS and mPING reports were compared to the nearest RF forecast location, so long as there was a RF forecast point within 40-km of the observation. Only the NYSM and upper-air, NYSM and upper-air reduced (described in Section 4.2.1), and NAMNEST products were available to be evaluated for the winters of 2020 k–2021 and 2021–2022, as the HRRR, and HRRR and NYSM, products were only developed in time for the winter of 2022–2023.

To match the ASOS and mPING reports with the RF forecast, the valid time of each individual forecast issued by the RF was used to determine the reports for comparison. Because ASOS sites report precipitation from present weather sensors at a variety of time scales, generally at either 1- or 5-minintervals, only ASOS reports at the valid forecast were used to verify RF forecasts. ASOS present weather sensors have a variety of codes for the weather that is occurring. These codes were simplified and matched to fit the four categories for which the RF makes predictions. If there were multiple types of precipitation occurring, the report was duplicated; therefore, only one precipitation type would be associated with each type of report.

mPING reports went through a similar selection process as the ASOS reports, but there are some important differences. mPING reports were selected by locating all reports that were 30 min before or after the valid forecast time. For a forecast valid at 1000 UTC, mPING reports from 0930 to 1030 UTC were used to verify the RF output. This window is due to there being no mandatory reporting window for mPING reports, unlike ASOS and helped to collect more reports at the valid forecast time while still giving an accurate depiction of what precipitation was occurring. mPING reports also can contain multiple types of precipitation occurring at the same time; these reports were also duplicated, so only one precipitation type would be associated with each type of report.

After the reports were matched for each valid forecast time, maps and confusion matrices were made to display the comparison between the ASOS and mPING observations, and the random forest predictions (Figs. 2.2 and 2.3). The maps allow forecasters to see in which areas the RF did the best and worst. The confusion matrices illustrate numerically where the correct and incorrect predictions end up. Since the confusion matrices are also broken down by verification report, any potential bias in the observations can be examined. The confusion matrices for individual forecasts can be combined to make event or whole winter season confusion matrices.

It is important to note that the forecast guidance must be evaluated both deterministically and probabilistically. Deterministic evaluation, e.g., did the RF make the correct precipitation type forecast, can be done by skill scores and confusion matrices. Probabilistic evaluation, e.g., what do the RF output probabilities mean for forecasters' confidence in the forecast guidance, can be done by the comparing observations to the nearest RF output probability. Chapter 5 will discuss this in more detail as well as the successes and issues with the NYSM and upper-air, NYSM and upper-air reduced, and NAMNEST forecast products both over the past two winters and for specific case studies.

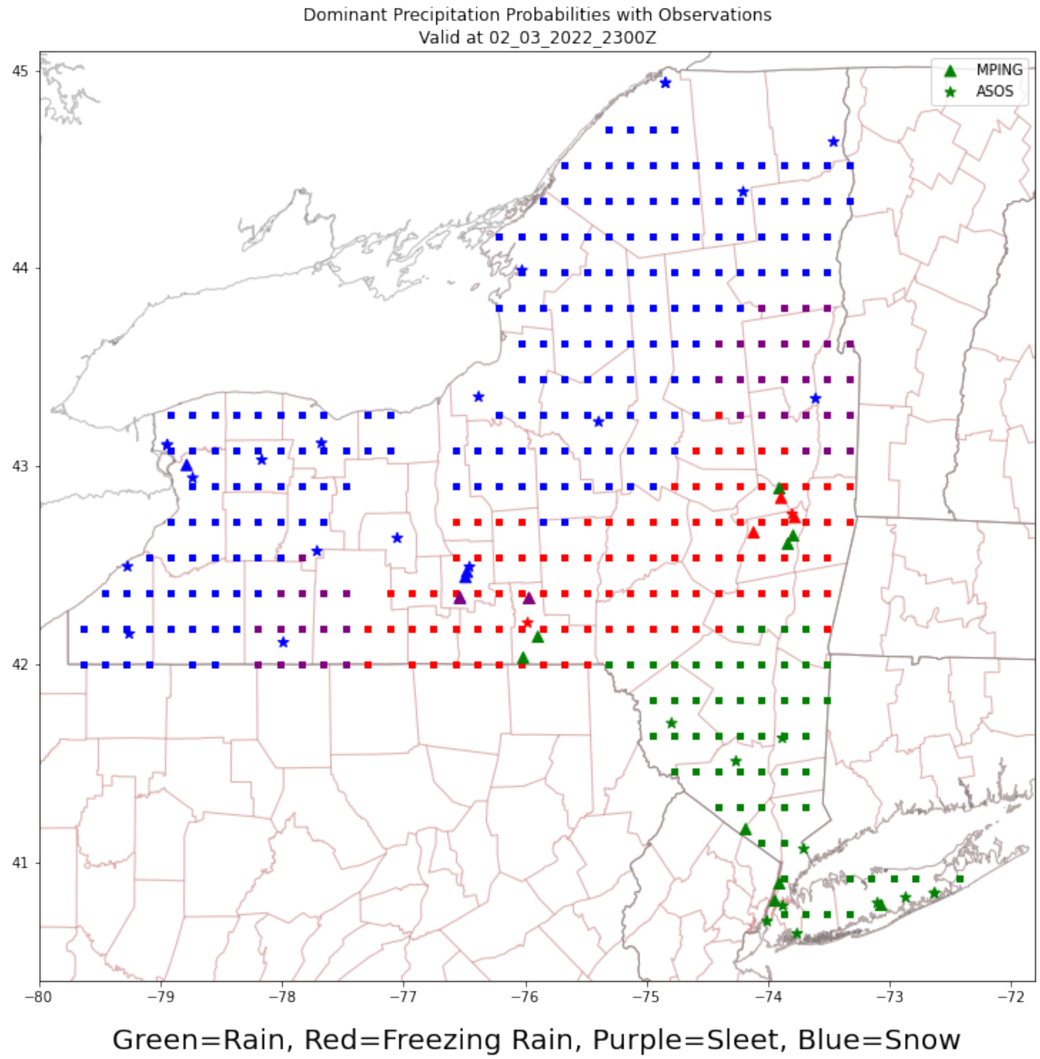

Dominant Precipitation Probabilities with Observations
Valid at 02_03_2022_2300Z
MPING
ASOS
45
44
43
42
41
−80
−79
−78
−77
−76
−75
−74
−73
−72
Green=Rain, Red=Freezing Rain, Purple=Sleet, Blue=Snow

Fig. 2.2. Map of RF forecast from the 1800 UTC NAMNEST model run at forecast hour five on 3 February 2022 with mPING (triangles) and ASOS (stars) observations overlayed. Each forecast point and observation are color coded (color key underneath plot).

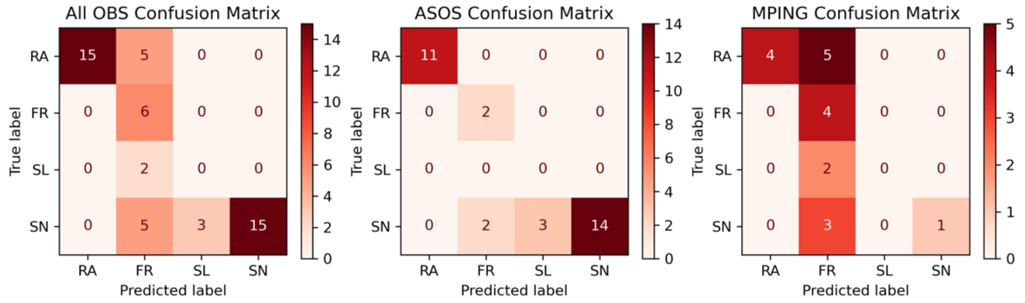


Fig. 2.3. Confusion matrix evaluating the RF forecast from the 1800 UTC NAMNEST model run at forecast hour five on 3 February 2022. The matrices are split up by observation data source (ASOS or mPING) as well as having a total matrix (ASOS+mPING). Darker red colors indicate more reports in that box.

3. Random Forest Internal Evaluation

3.1. NYSM and Upper-air

Internal testing of the RF is an integral step to evaluate the impact of the different combinations of input data and variables (features) available to train the RF, which can be done by making predictions based on the subset of testing data that was not included when training the RF. Starting with the NYSM and upper-air data, all combinations of data were considered to identify which variables would be most effective, including testing different combinations of variables such as only raw data variables, taken directly from the observations or radiosondes, or only calculated variables, calculated from the raw observations or radiosonde data (Table 2.3 and Fig. 3.1). In evaluating training datasets, it was important to not only consider overall accuracy (denoted by the black circles), but the individual F1 scores for different precipitation types because high accuracies in certain categories can mask other low accuracies. Fig. 3.1 shows that the overall accuracy and F1 scores changed when considering different combinations of data, which points to the importance of the dataset in the RF algorithm. The most accurate data combination overall was the NYSM 5-min and hourly data combined with observed soundings (Table 3.1 and Figure 3.1). This combination aligned with the highest F1 scores for the mixed precipitation categories of sleet and freezing rain. While F1 scores around 55% and 65% are not ideal (a higher F1 score indicates that the RF is identifying a majority of the training cases correctly), these values were nevertheless promising. The F1 scores were especially encouraging considering the challenge of forecasting mixed precipitation as noted by the range of POD values from Reeves et al. (2014), with the caveat that these values represented different metrics for evaluating the datasets.

Dataset Description	**Abbreviations**
NWS Buffalo, Albany, and Upton radiosondes	Original Soundings
NWS Buffalo, Albany, and Upton and Maniwaki, Quebec radiosondes	Updated Soundings
NYSM hourly-averaged surface variables with raw and original calculated variables from original soundings (Table 2.3 C1,C2)	HAVG_RCO
NYSM hourly-averaged surface variables with raw from original soundings (Table 2.3 C1)	HAVG_RO
NYSM hourly-averaged surface variables with original calculated variables from original soundings (Table 2.3 C2)	HAVG_CO
NYSM 5-min surface observations with raw and original calculated variables from original soundings (Table 2.3 C1,C2)	OBS5_RCO
NYSM 5-min surface observations with raw variables from original soundings (Table 2.3 C1)	OBS5_RO
NYSM 5-min surface observations with original calculated variables from original soundings (Table 2.3 C2)	OBS5_CO
All NYSM surface data with raw and original calculated variables from original soundings (Table 2.3 C1,C2)	ALL_RCO
All NYSM surface data with raw variables from original soundings (Table 2.3 C1)	ALL_RO
All NYSM surface data with original calculated variables from original soundings (Table 2.3 C2)	ALL_CO
NAMNEST soundings with raw and original calculated variables (Table 2.3 C1,C2)	NAM_RCO
NAMNEST soundings with raw variables (Table 2.3 C1)	NAM_RO
NAMNEST soundings with original calculated variables (Table 2.3 C2)	NAM_CO
NAMNEST soundings with raw and all calculated variables (Table 2.3 C1,C2,C3)	NAM_RCN
NAMNEST soundings with raw variables (Table 2.3 C1)	NAM_RN

NAMNEST soundings with all calculated variables (Table 2.3 C2,C3)	NAM_CN
All NYSM surface data with raw and all calculated variables from updated soundings (Table 2.3 C1,C2,C3)	ALL_RCN
All NYSM surface data with raw variables from updated soundings (Table 2.3 C1)	ALL_RN
All NYSM surface data with all calculated variables from updated soundings (Table 2.3 C2,C3)	ALL_CN
All NYSM surface data with raw and all calculated variables from NAMNEST soundings (Table 2.3 C1,C2,C3)	ALL_NAM_RCN
All NYSM surface data with raw variables from NAMNEST soundings (Table 2.3 C1)	ALL_NAM_RN
All NYSM surface data with all calculated variables from NAMNEST soundings (Table 2.3 C2,C3)	ALL_NAM_CN

Table 3.1. Description of random forest dataset and abbreviations. Descriptions indicate dataset (NYSM and NAMNEST), sounding location (original, updated, NAMNEST) and type of sounding variables (raw (C1) and calculated, original (C2), and new (C3)). Sounding variables are described in Table 2.3 and NYSM variables are described in Table 2.2.

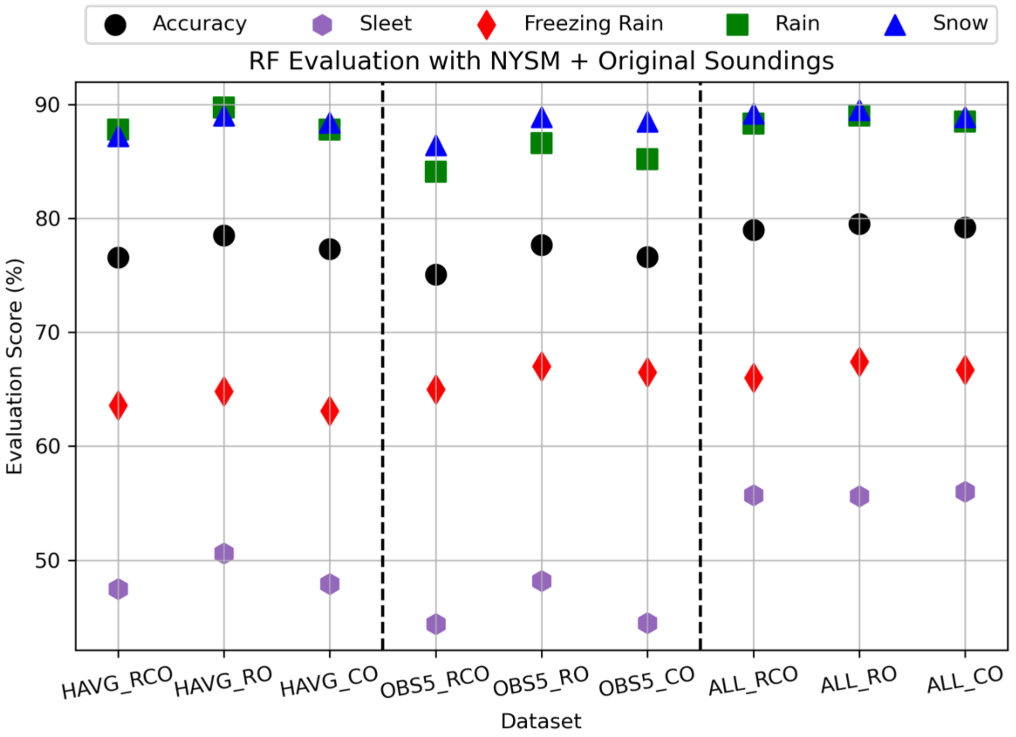


Fig. 3.1. Accuracy and F1 scores for different combinations of NYSM and upper-air data. The dataset abbreviations are in Table 3.1. Dashed lines represent divisions between different datasets. The left third represents the dataset built from hourly NYSM and upper-air data. The middle third represents the dataset built from 5-min NYSM and upper-air data. The right third represents the dataset built from hourly and 5-min NYSM, and upper-air data. The black circles represent the overall accuracy of the RF. The colored shapes correspond to the F1 scores of the different precipitation types (purple hexagons for sleet, red diamonds for freezing rain, green squares for rain, and blue triangles for snow).

Fig. 3.2 shows the confusion matrix from a test run of the RF with the ALL_RO training data (Table 3.1), one of the highest performing RF runs of the NYSM and upper-air data. The F1 scores from Fig. 3.1 were very similar to the fraction of correct predictions for sleet (0.55 correct) and freezing rain (0.65 correct) in Fig. 3.2. An interesting feature of these RF runs was that the algorithm’s values of correct predictions for snow (0.92) and rain (0.87) events were similar to the combined confusion matrix value when predicting a mixed precipitation type (sleet or freezing rain) for a true freezing rain or sleet event (0.87 for freezing rain and 0.78 for sleet, respectively). This result suggested the RF can more easily recognize three major types of precipitation: rain, mixed precipitation, and snow.

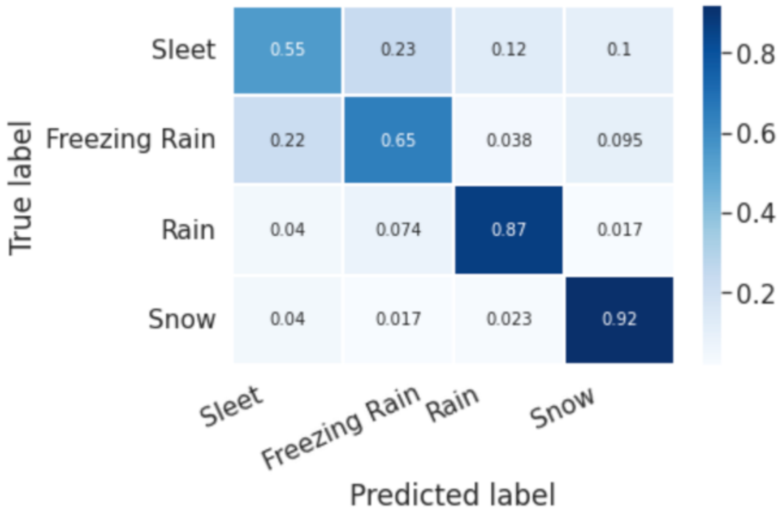


Fig. 3.2. Confusion matrix of values averaged over 50 independent RF runs for the ALL_RO NYSM and upper-air dataset (Table 3.1).

3.2. NAMNEST

A similar method of evaluation was followed for the NAMNEST dataset. There was no model-derived precipitation type in the training dataset, so the RF created a precipitation type from meteorological variables only. Two combinations of NAMNEST data were generated: the original datasets were the first attempt to pull model data to predict the winter mixed precipitation events and the new datasets were an attempt to improve the original dataset, which only contained the variables in Table 2.3 C1 and C2, by adding new calculated variables, found in Table 2.3 C3 to the original dataset. The new datasets were a clear improvement over the original datasets with a roughly 5% jump in overall accuracy (Fig. 3.3), which reinforced the premise that different variable combinations in the dataset can impact the RF. Additionally, the F1 scores for all four types of precipitation increased for runs utilizing the new datasets. Several of the new calculated variables appeared in the top 10 features from runs of the NAM_RCN dataset (Fig. 3.4): positive area (defined as the area between the 0°C isotherm and environmental temperature in a vertical temperature profile; Bourgouin 2000), maximum wet-bulb temperature between 925–700 hPa, minimum temperature between the surface and 850 hPa, average temperature between the surface and 850 hPa, and maximum temperature between 850 hPa and 700 hPa. This result verified that the new calculated variables (Table 2.3 C3) that were added led to the difference in the higher scores that were seen.

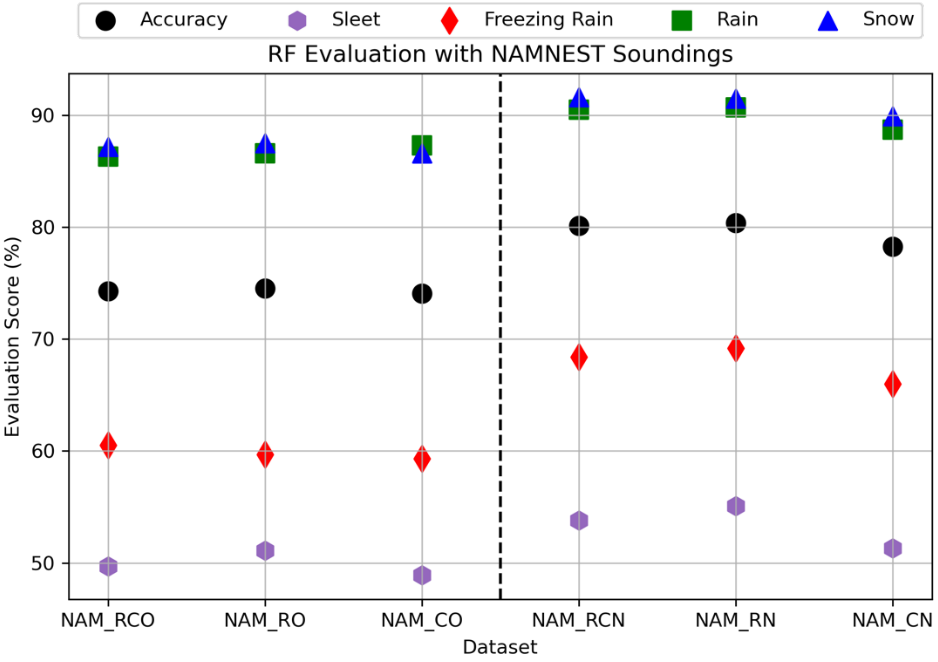


Fig. 3.3. Accuracy and F1 scores for different combinations of NAMNEST data. The dataset abbreviations are in Table 3.1. Dashed lines represent divisions between different datasets. The left half represents the dataset built from NAMNEST data with the raw and original calculated variables (Table 2.3 C1 and C2). The right half represents the dataset built from NAMNEST data with the raw and original calculated variables plus the new calculated variables (Table 2.3 C1, C2, and C3).

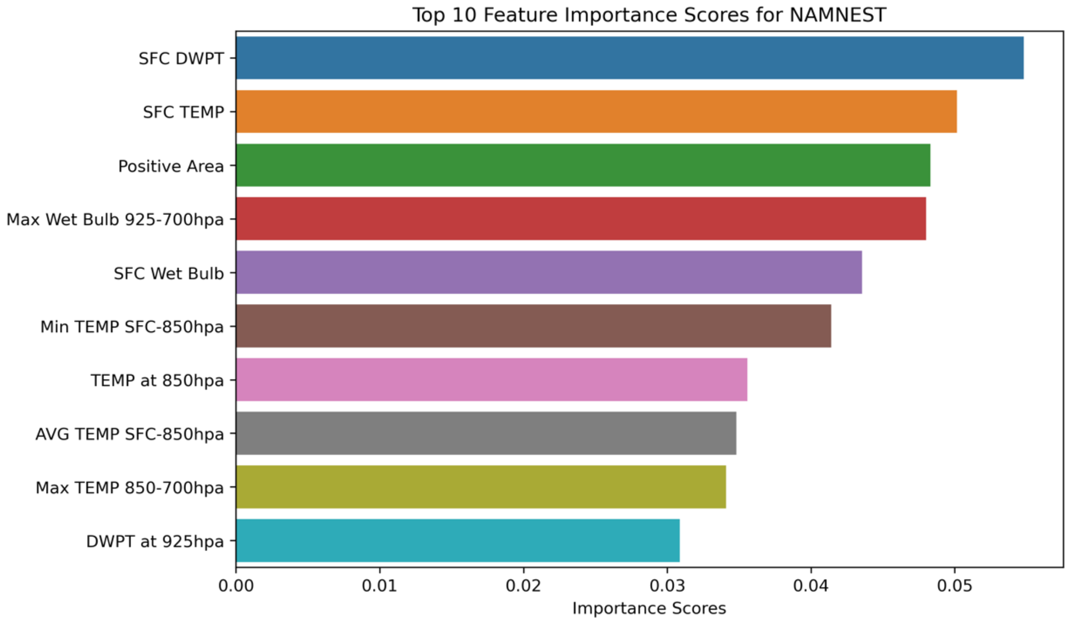


Fig. 3.4. Top 10 most important variables from RF runs of new NAM_RCN dataset. The values of importance have been averaged over 50 runs. The higher the value the more important the variable.

From a meteorological perspective, these new calculated variables tended to better quantify how temperature varied between mandatory pressure levels and more clearly captured the vertical temperature profile in the lowest part of the atmosphere. These variables were selected specifically because they have been identified in the literature (Ramer 1993; Baldwin et al. 1994; Bourgouin 2000; Manikin 2005; Benjamin et al. 2016a) as important in determining mixed precipitation type, and their usefulness was seen by an increase in F1 scores for sleet and freezing rain (Fig. 3.3). While improvement was made by incorporating variables that gave a better sense of the entire vertical temperature profile, there were challenges with using different datasets for vertical temperature profiles. One challenge with radiosonde and NAMNEST profiles was that the pressure levels were not consistent throughout due to no two soundings ever being the same, which made it difficult to get values at consistent locations aside from mandatory pressure levels.

Since the new calculated variables (Table 2.3 C3) were successful in improving the NAMNEST RF, those same variables were added into the original NYSM and radiosonde datasets to see if the same improvement would occur. Fig. 3.5 compares the best three original NYSM runs (left third) to NYSM runs with the new calculated variables added into the datasets (middle third). Additionally, combining the NYSM 5-min and hourly variables with the NAMNEST profiles (right third) was tested to determine whether increasing the spatial density of vertical profiles would improve the RF. The new calculated variables added to the NYSM and upper-air profiles caused a slight decrease in overall accuracy. In particular, the sleet and freezing rain F1 scores decreased by 6–8% and 3–4%, respectively. This decrease was likely associated with the new calculated variables adding conflict in the decision-making process of the RF leading to incorrect predictions. Conflicting decisions occur because the new calculated variables in the dataset may end up creating decisions that are at odds with each other about how to split the dataset. Without the new calculated variables, these decisions were previously in agreement with each other.

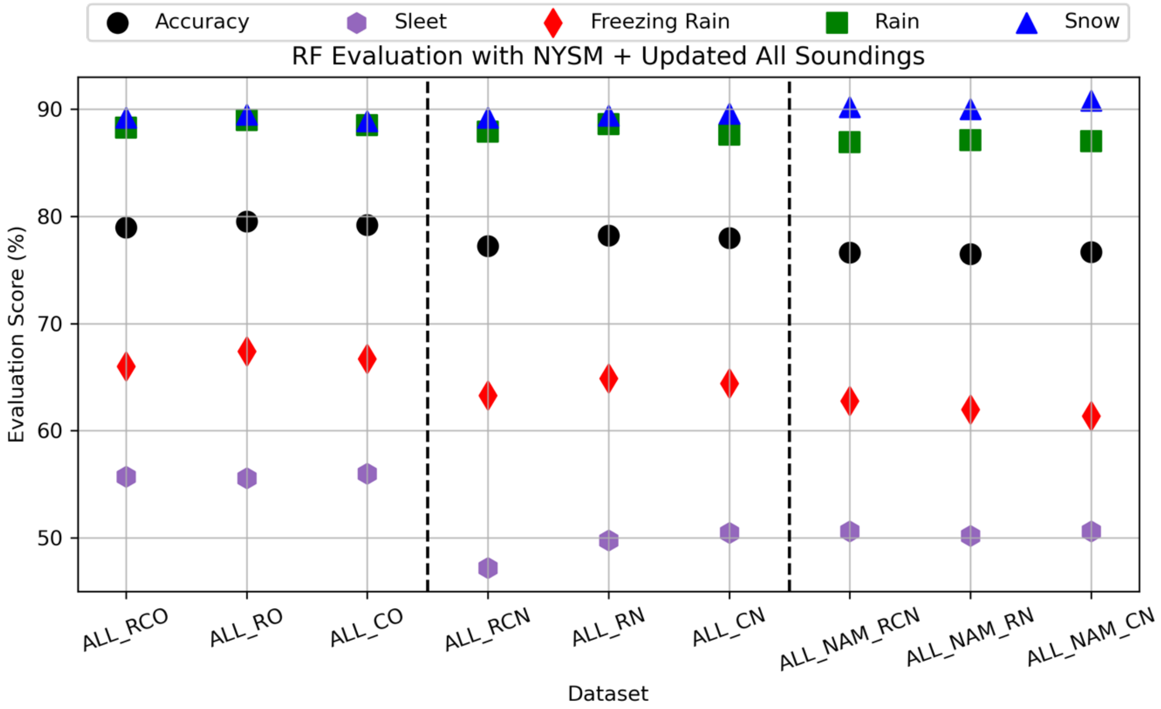


Fig. 3.5. Accuracy and F1 scores for different combinations of NYSM, NAMNEST, and upper-air data. The dataset abbreviations are in Table 3.1. Dashed lines represent divisions between different datasets. The left third represents the dataset built from hourly and 5-min NYSM and upper-air data with raw and original calculated variables (Table 2.3 C1 and C2). The middle third represents the dataset built from hourly and 5-min NYSM and upper-air data with raw and original calculated variables plus the new calculated variables (Table 2.3 C1, C2, and C3). The right third represents the dataset built from hourly and 5-min NYSM and NAMNEST data raw and original calculated variables plus the new calculated variables (Table 2.3 C1, C2, and C3).

When testing whether increased spatial resolution of the vertical profiles would increase accuracy, one would expect the RF to do better because more vertical sampling points of the atmosphere should create more representative data. This was not the case, however, with the NYSM and NAMNEST datasets. The decrease connected to these results was most likely because there was an overlap in the variables in the combined datasets. Also, the surface data could conflict or not be meteorologically consistent with the NAMNEST data due to the geographical locations of the datapoints. Pairing surface data from a location at significantly higher elevation than a NAMNEST profile site can cause a discrepancy. While these results showed a decrease in performance, it prompted a response to conduct future experimentation with different combinations of the NYSM and NAMNEST profile data to find the best possible combination. This point was a key takeaway from the process of testing the RF and determining the best possible dataset because each type and combination of data needs to be treated differently.

3.3. HRRR

Similar to the NYSM and upper-air (Section 3.1) and NAMNEST (Section 3.2), the matched training dataset cases with the HRRR vertical profile data was put through internal testing to select the best training dataset. Along the same lines as the NAMNEST, there was no model-derived precipitation type in the training dataset, which limited the variables to be meteorological only. Since the HRRR products were develop after the NAMNEST and NYSM and upper-air products had been operational and had verification done on them, the HRRR dataset evaluation did not focus on changing the composition of raw and calculated variables. Instead, the focus was on selecting the best combination of variables possible now that more information was available because the HRRR data had consistent pressure levels, meaning that there were significantly more variables available in the HRRR than in the NAMNEST.

The left half of Fig. 3.6 shows the three different HRRR datasets that were built and their accuracy and F1 scores. The three datasets decreased in total number of variables from left to right; this decrease in total variables was accompanied by a general increase across accuracy and F1 scores. While this may seem counter intuitive because more data should mean more information about what is occurring, one thing that can impact the success of a RF is the relatedness of its variables. If there are similar variables that capture the same information, overlaps in the dataset can occur that can add conflict to the RF's decision-making process. Essentially, the extra variables create extra noise in the random forest; hence, the simplified dataset, HRRR_simp, was most effective due to no overlapping variables and having a total number of variables more similar to the NAMNEST datasets. One way to reduce the number of variables in a dataset is by doing PCA, as mentioned in Section 1.3. This process reduces the number of variables in the dataset by identifying relationships between variables; one downside to PCA is that the set of reduced variables was not as intuitive to understand as the non-reduced set of variables. Thus, this technique was not applied to the data used in the random forest.

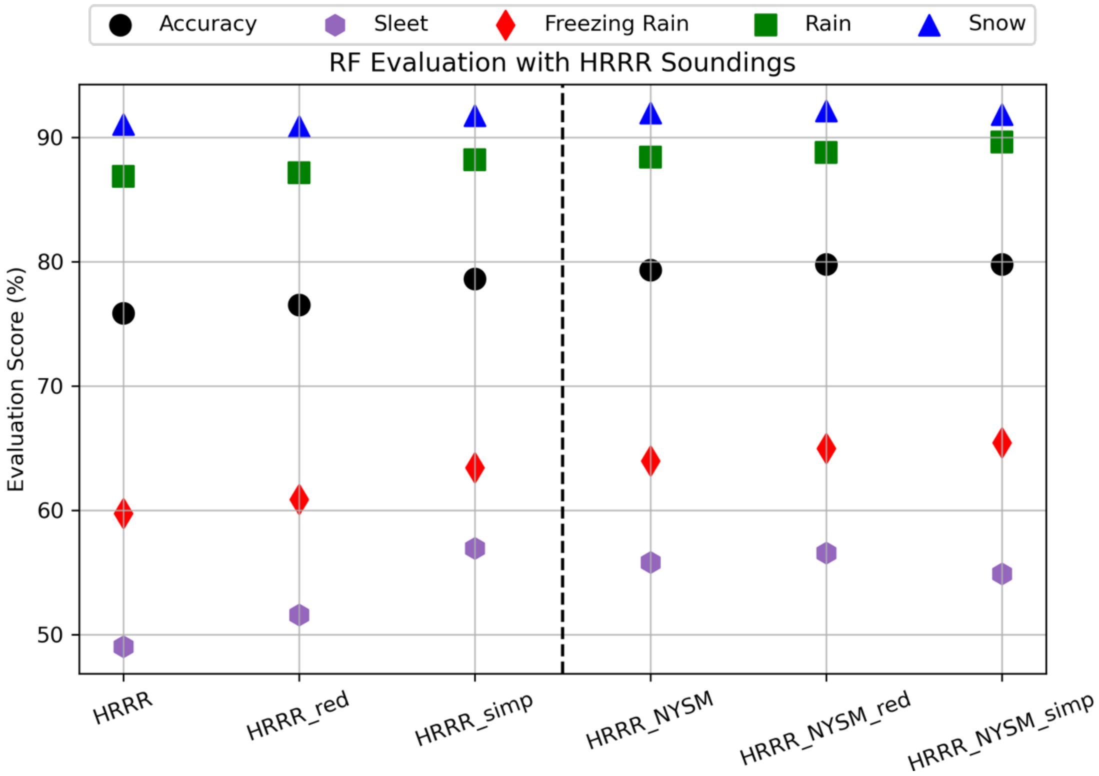


Fig. 3.6. Accuracy and F1 scores for different combinations of HRRR and 5-min NYSM data. The HRRR dataset abbreviations are in Table 2.4. Dashed lines represent divisions between different datasets. The left half represents the dataset built from only HRRR data (Table 2.4 H1, H2, and H3). The right half represents the dataset built from HRRR data (Table 2.4 H1, H2, and H3) and 5-minute NYSM data.

The right half of Fig. 3.6 shows the same three combinations of HRRR datasets to the left, except they have been combined with surface observations in the form of 5-min NYSM data. The NYSM data has replaced all surface information (2-m and 10-m variables) from the HRRR. While there was no definitive pattern as compared to the left half of Fig. 3.6, there was still interesting differences between the datasets on the right half. The dataset that was selected as the final training dataset was HRRR_NYSM_red; this was because the accuracy was the highest, just higher than HRRR_NYSM_simp, and its F1 scores for freezing rain and sleet were the best pair. It would be remiss to not note that the F1 score for freezing rain in the HRRR_NYSM_simp dataset was higher, but the sleet F1 score was lower. It was important to not sacrifice the F1 score of sleet for slight improvements in freezing rain considering how difficult they both are to forecast. In working with the HRRR data, there was an increase in the amount of information that the RF was exposed to. This increase in information caused for increased understanding of what information was helpful and harmful to the RF besides that data sources need to be treated independently from each other.

4. Research to Operations Framework and Transition

Developing a functioning ML algorithm is an intensive process and can be time consuming to properly configure and test. Once configured and tested, it is straightforward to test on datasets that already exist. The challenge occurs when trying to apply the algorithm as a real-time tool to make forecasts. Transitioning this algorithm from research to operations was a multistep process with potential issues associated with real-time processing of incoming data from various sources, formatting the datasets, and runtime issues. This section describes the transition of this research-oriented ML algorithm to an operational setting as well as the operational products available to forecasters.

4.1. Operational Framework and Challenges

The path for transitioning this RF from research to operations is represented in the flow chart in Fig. 4.1. Chapters 2 and 3 detailed the process of preparing and testing the training dataset until the best dataset was determined (Fig. 4.1, steps 1–4). The next step in the transition to an operational RF was creating the testing dataset. Since the process of collecting CoCoRaHS reports as training cases was qualitative, they did not make sense to use in real-time due to the quality control needed and their limited spatial distribution. As an alternative, a 20-km grid of New York State was generated to create synthetic locations for which predictions of precipitation type can be generated. These points were matched with NYSM, upper-air, NAMNEST, and HRRR profile locations using the same process as the CoCoRaHS reports (Fig. 4.1, step 5; Fig. 4.2). Each time the RF was run to make a prediction, the incoming data was compiled, cleaned to make sure there is no missing data, and prepared to conform to the training dataset used in the RF. The processing of the incoming data was an essential step because the RF will not run if the incoming dataset did not match the training dataset or there were missing values. (Fig. 4.1, step

6). If reliable data sources were not available, it was difficult to produce consistent forecast guidance for end users. For example, if radiosondes were not launched or if other data sources were not uploaded with consistency, it can be difficult to process and make a complete prediction in a specific window of time. This lack of data may occur from computer/power outage issues or more significant issues like helium shortages for radiosondes (e.g., NOAA NWS 2022a). Once the incoming data in the testing dataset matched the training dataset (Fig. 4.1, step 7), the RF can be run. The outputs of the RF were probabilistic predictions of each precipitation type at each location in the testing dataset and can be processed to create maps, graphics, or tables to be displayed in an operational setting (Fig. 4.1, steps 8–9).

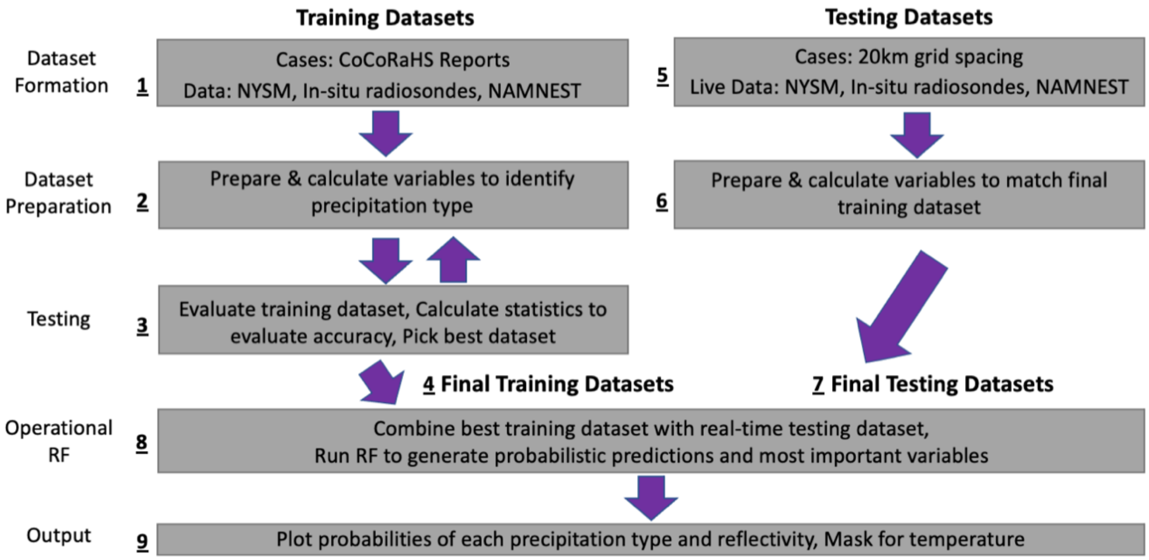

Training Datasets
Testing Datasets
Dataset Formation
1
Cases: CoCoRaHS Reports
Data: NYSM, In-situ radiosondes, NAMNEST
5
Cases: 20km grid spacing
Live Data: NYSM, In-situ radiosondes, NAMNEST
Dataset Preparation
2
Prepare & calculate variables to identify precipitation type
6
Prepare & calculate variables to match final training dataset
Testing
3
Evaluate training dataset, Calculate statistics to evaluate accuracy, Pick best dataset
4 Final Training Datasets
7 Final Testing Datasets
Operational RF
8
Combine best training dataset with real-time testing dataset,
Run RF to generate probabilistic predictions and most important variables
Output
9
Plot probabilities of each precipitation type and reflectivity, Mask for temperature

Fig. 4.1. Flowchart of the methodology to create an operational RF for predicting winter mixed precipitation types. This flowchart can be generalized for other ML algorithms and meteorological events.

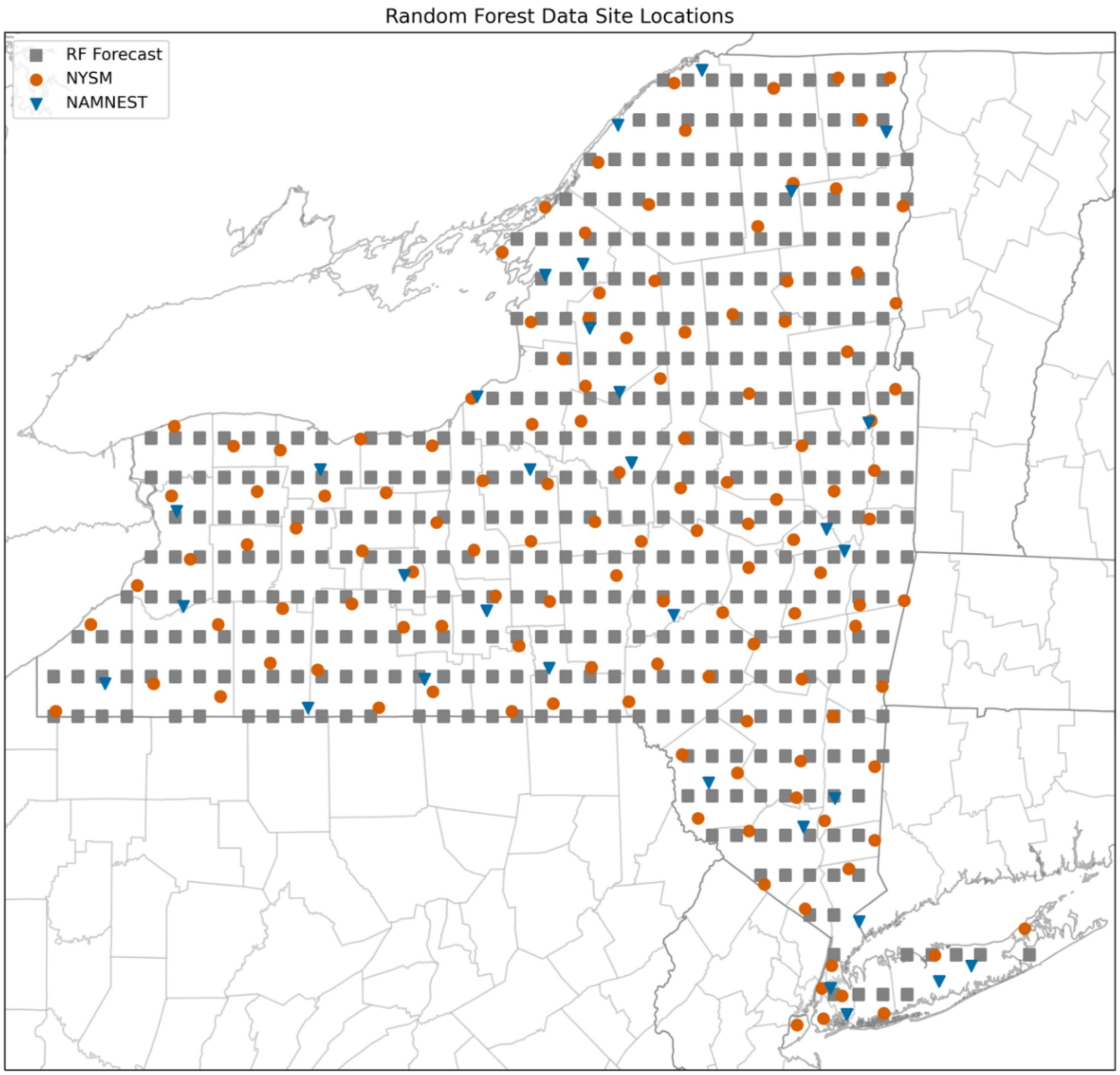


Fig. 4.2. Map of New York displaying locations of NYSM (orange circles) and NAMNEST profile sites (blue triangles). The gray squares denote 20-km grid spacing of the RF prediction locations.

4.2. Operational Forecasting Tool

4.2.1. Product Development

One important issue that occurred throughout the development of the RF was the continual reminder that the RF output had to be displayed in an effective manner. This process was not just to limit how numbers were displayed on a map but was expanded to what products were being made and when can be made based on the data sources available. These considerations were impactful because if a product takes 1 hour or more to be made, but is meant to be a nowcasting tool, it limits the usefulness of the tool. This subsection will focus on the product development timeline for all products available on the operational website.

The initial two groups of datasets that were the focus of the RF development were the NYSM and upper air, and the NAMNEST. This allowed the RF to be implemented on observational datasets that could be used for nowcasting and a model-based dataset that could allow precipitation forecasts to be made in the future. When the products for an operational website were being developed, it was key that the products being made for both the NYSM and upper air, and the NAMNEST, could be used for other datasets as well. Because of this, six base products were developed that include probabilities for precipitation types as well as radar or model reflectivity overlayed. The base six products are described in more detail in section 4.2.2. Going into the of winter 2021–2022, the available products included nowcasts by the NYSM and upper air data, and forecasts by the NAMNEST data, which made predictions for 10 forecast hours, including forecast hour 0, thereby giving operational forecasters up to 5 hours of lead time since operational forecasts became available in forecast hour 4.

During the winter of 2021–2022, the focus was on how well the NYSM and upper air, and NAMNEST, RF models forecast precipitation type and what were the next plausible combinations or data sources that should be implemented. No active verification occurred during this period, but both RF models' operational products were put through the eye test to determine reasonableness of the probabilities and precipitation types that were being displayed. One issue discovered was that the NYSM and upper-air RF model had an unrealistically large number of sleet predictions (discussed more in Chapter 5). This issue was so prevalent a mid-winter assessment occurred to determine how this RF model could be improved. After examining the key variables in the dataset, it was clear that the temperature range of the radiosondes was not as representative as it should be due to issues in the spatial and temporal resolution of the data. The spatial resolution at which the radiosondes were being used to make predictions was identified as being too large. For example, the NWS Albany radiosonde was being used from the Canadian border, throughout the entirety of the Adirondacks, and as far south as Poughkeepsie and Newburgh. Temporally, the radiosondes were being used for about 12 hours, so by the end of this period the radiosonde profile may not accurately depict the current vertical profile. This prompted the development of the NYSM and upper-air reduced RF model. The NYSM and upper-air reduced RF model used the same data sources as the original NYSM and upper-air, except it only combines 5-min NYSM surface data with the upper-air radiosondes. In addition, the area in which predictions were being made was limited spatially to within a 100-km radius of the launch site and temporally to within 4 h of launch time (0000 UTC to 0400 UTC or 1200 UTC to 1600 UTC). The results of the NYSM and upper-air reduced RF model will be discussed in Chapter 5.

In preparation for the winter of 2022–2023, focus was on what new data sources and upgrades could be incorporated into the operational website. For new data sources, the HRRR was added into the operational website because the spatial resolution of the HRRR was significantly better compared to the NAMNEST (Fig 4.3). This attempt to increase the spatial resolution came from the results of the NAMNEST (see Chapter 5) over the previous two winters and feedback from NWS Albany. The HRRR contains 107 vertical sounding locations, evenly spaced throughout New York on a 40-km grid, compared to the NAMNEST which has 32 vertical sounding locations unevenly distributed throughout New York. Based on results in Chapter 3, a combination of HRRR and NYSM data produced enough confidence in internal testing to develop a HRRR and NYSM product. While other combinations of model and observational datasets were tested internally, this product was the first operational tool that combined NYSM surface observations with model vertical profiles. In addition to the new RF models available for the winter of 2022–2023, the NAMNEST RF model showed skill at predicting precipitation type with 5 h of lead time (Chapter 5 has more discussion on this topic). To upgrade the HRRR and NAMNEST RF models, they will now forecast out to 12 hours of lead time. This expansion will give forecasters more lead time on challenging precipitation type forecasts.

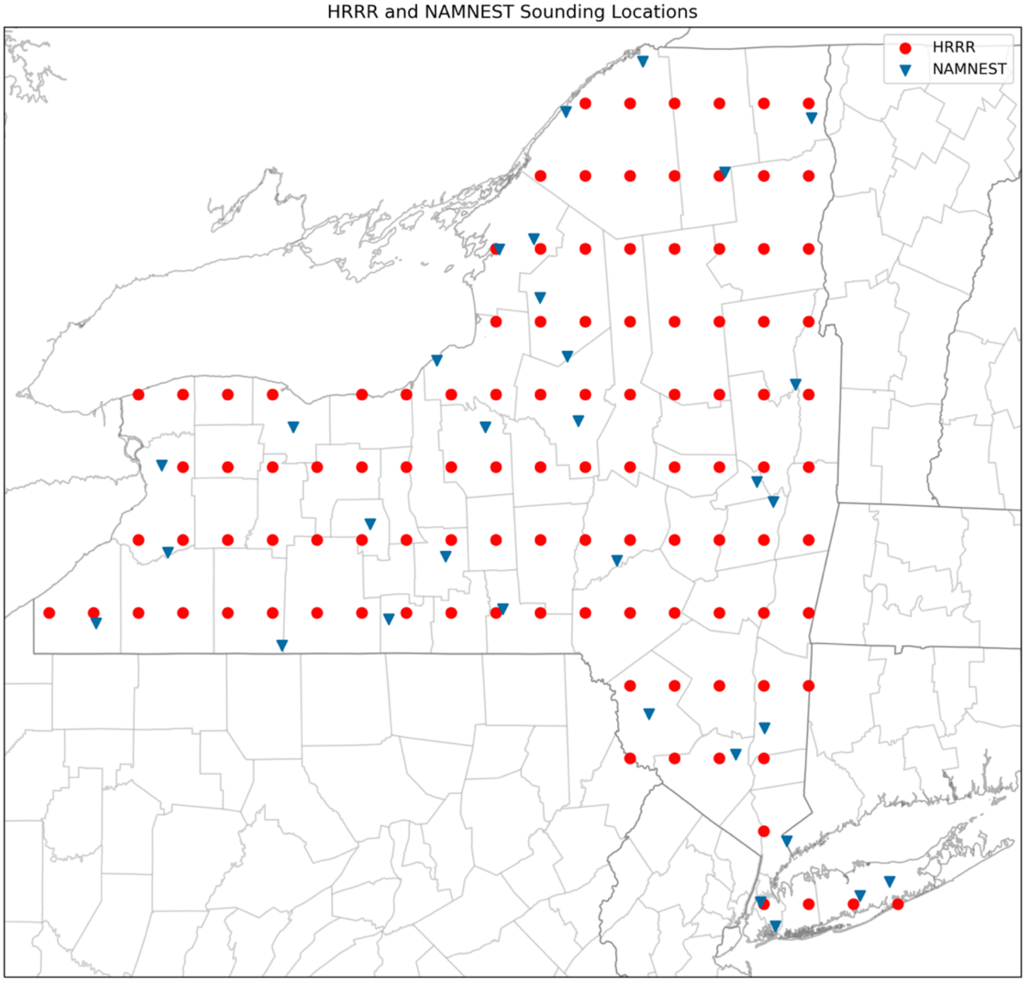


Fig. 4.3. Map of New York displaying locations of NAMNEST (blue triangles) and HRRR profile sites (red circles).

In addition to all the RF models displayed on the operational website, a new model–observation comparison product will be available for the winter of 2022–2023. This product (Fig. 4.4) displays the 2-m temperature difference for either, or both, the NAMNEST and HRRR compared to the NYSM 5-min 2-m temperature. Since 2-m temperature is very important to precipitation type forecasts and the RF models, this product was developed to show which model is most representative of current conditions. This comparison can help forecasters increase their confidence in RF model forecasts. Section 4.2.2 provides a summary of all available products, latency times, and the timing when each product was made for the current website configuration.

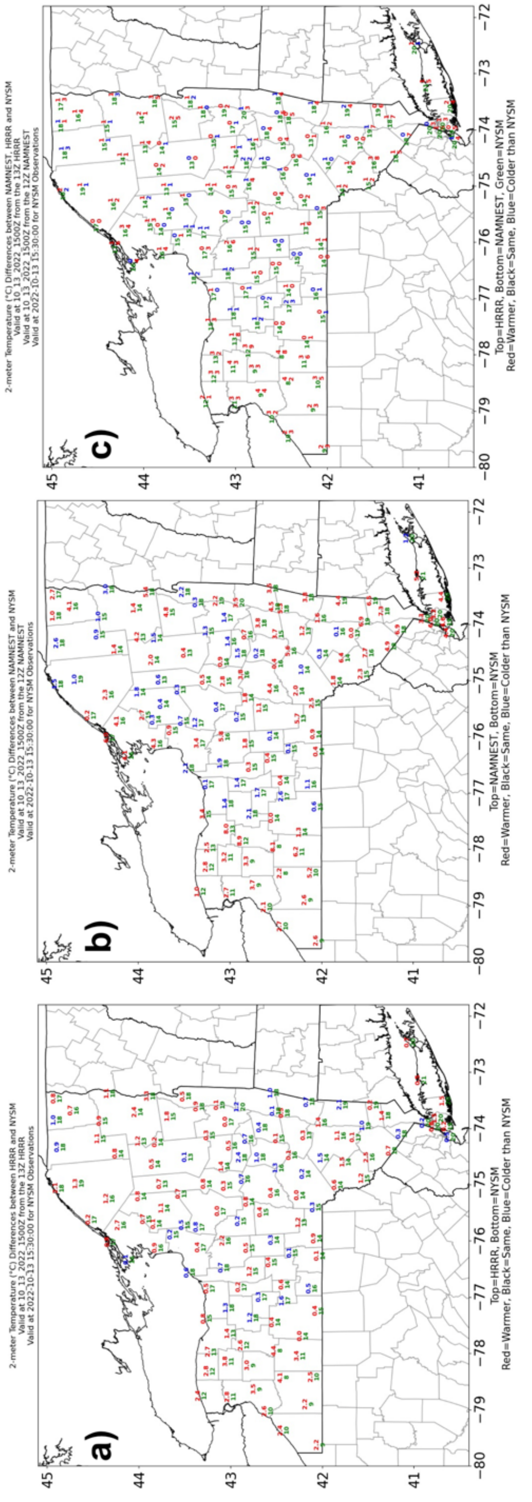

a)
2-meter Temperature (°C) Differences between HRRR and NYSM
Valid at 10_13_2022_1500Z from the 13Z HRRR
Valid at 2022-10-13 15:30:00 for NYSM Observations
Top=HRRR, Bottom=NYSM
Red=Warmer, Black=Same, Blue=Colder than NYSM
b)
2-meter Temperature (°C) Differences between NAMNEST and NYSM
Valid at 10_13_2022_1500Z from the 12Z NAMNEST
Valid at 2022-10-13 15:30:00 for NYSM Observations
Top=NAMNEST, Bottom=NYSM
Red=Warmer, Black=Same, Blue=Colder than NYSM
c)
2-meter Temperature (°C) Differences between NAMNEST, HRRR and NYSM
Valid at 10_13_2022_1500Z from the 13Z HRRR
Valid at 10_13_2022_1500Z from the 12Z NAMNEST
Valid at 2022-10-13 15:30:00 for NYSM Observations
Top=HRRR, Bottom=NAMNEST, Green=NYSM
Red=Warmer, Black=Same, Blue=Colder than NYSM
45
44
43
42
41
−80
−79
−78
−77
−76
−75
−74
−73
−72

Fig. 4.4. Map of New York displaying observations of 2-m temperature from the NYSM (green), HRRR, and NAMNEST. a) Compares HRRR to NYSM, b) compares NANMNEST to NYSM, and c) compares both NAMNEST and HRRR to NYSM.

4.2.2. Operational Website and Archive

The probabilities from the RF runs are located on an operational website: http://www.atmos.albany.edu/student/filipiak/op/. The website displays multiple RF products being made in real-time throughout the 2022–2023 winter season. The full lineup of products includes RF models for the NYSM and upper-air reduced, NAMNEST, HRRR, HRRR and NYSM datasets, as well as the model–observation comparison product. The latency on the NYSM and upper-air reduced product is under 10 min and is made with the forecast valid at 30 min past each hour for each of the hours available (0000 UTC, 0100 UTC, 0200 UTC, 0300 UTC, 1200 UTC, 1300 UTC, 1400 UTC, and 1500 UTC). The latency for the NAMNEST product updated with each new model run is about one hour for a 17-h forecast period including forecast hour 0. The NAMNEST is available 4 h, including time to run RF, after model initialization time (0000 UTC, 0600 UTC, 1200 UTC, 1800 UTC). The latency for the HRRR product updated with each new model run is about one hour for a 16-h forecast period including forecast hour 0. The HRRR is available 3 h, including time to run RF, after model initialization time (0000 UTC, 0300 UTC, 0600 UTC, 0900 UTC, 1200 UTC, 1500 UTC, 1800 UTC, 2100 UTC). The latency for the HRRR and NYSM product is under 10 min and is available every hour 7 min past each hour. The latency for the model observation comparison product is under 1 min and is available every hour 38 minutes past each hour.

The products display the probabilities of the different types of precipitation, if precipitation is occurring. Even when there is no precipitation occurring, the RF products are being made, and they can be used to understand the current atmospheric conditions. The products available that display the RF output include the probabilities of the four main precipitation types (rain, Fig. 4.5a; freezing rain, Fig. 4.5b; sleet, Fig. 4.5c; snow, Fig. 4.5d), an all-mixed precipitation (with the addition of sleet and freezing rain probabilities; Fig. 4.5e), and a dominant precipitation type (shows color coded probabilities to display highest value at each location; Fig. 4.6). This last plot type (Fig. 4.6) is made with the assumption that the product will mostly be used during periods of active weather. The variety of products available allows for end users to have multiple views of which winter weather hazards are present or being forecast. In addition, the probabilistic nature of the RF allows for end users to have a sense of confidence in forecast precipitation type.

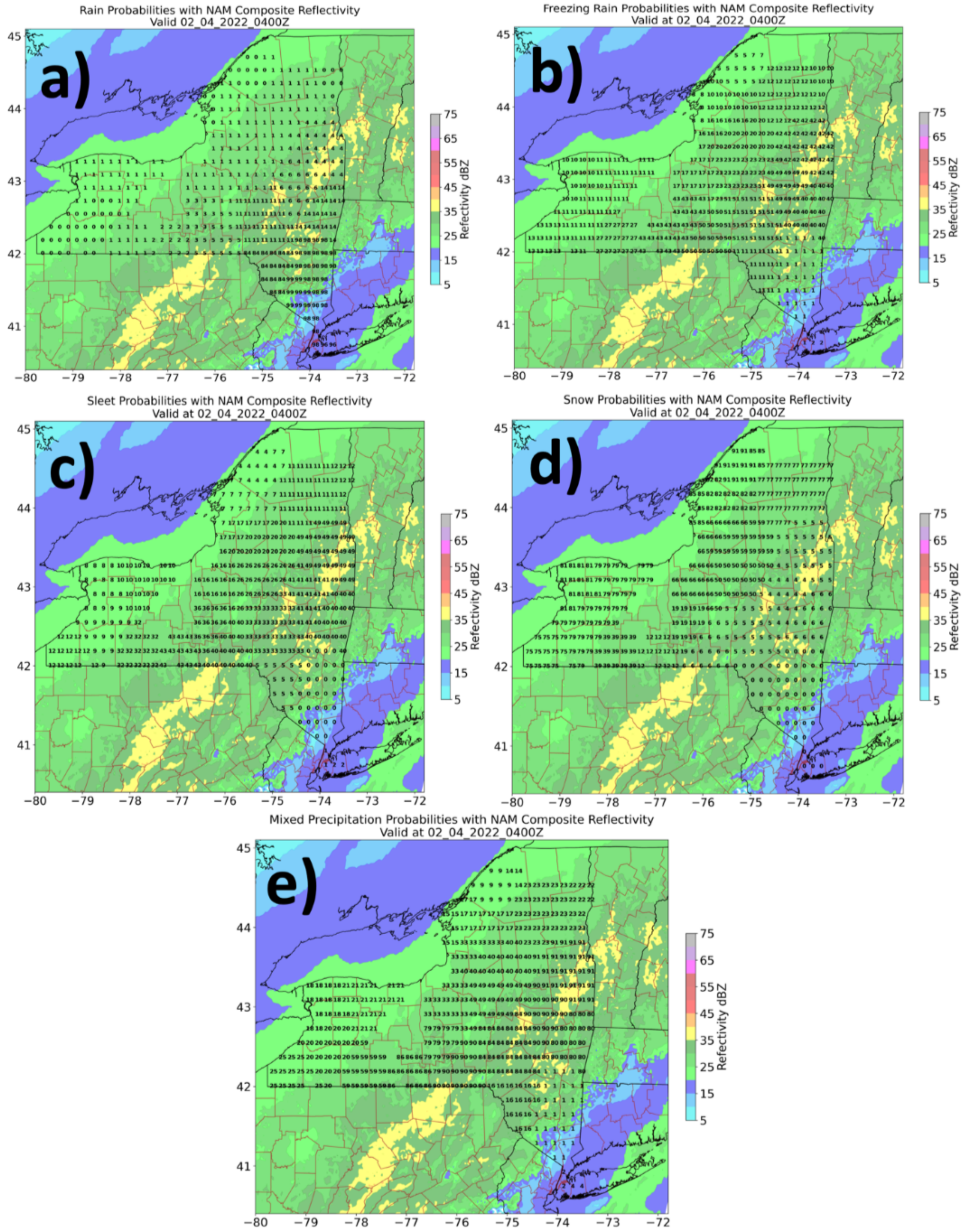

Rain Probabilities with NAM Composite Reflectivity
Valid 02_04_2022_0400Z
a)
Freezing Rain Probabilities with NAM Composite Reflectivity
Valid at 02_04_2022_0400Z
b)
Sleet Probabilities with NAM Composite Reflectivity
Valid at 02_04_2022_0400Z
c)
Snow Probabilities with NAM Composite Reflectivity
Valid at 02_04_2022_0400Z
d)
Mixed Precipitation Probabilities with NAM Composite Reflectivity
Valid at 02_04_2022_0400Z
e)
45
44
43
42
41
−80
−79
−78
−77
−76
−75
−74
−73
−72
Refectivity dBZ
75
65
55
45
35
25
15
5

Fig. 4.5: Forecast probabilities (black text) of a) rain, b) freezing rain, c) sleet, d) snow, and e) mixed precipitation (freezing rain + sleet) from the 0000 UTC NAMNEST model run at forecast hour four on 4 February 2022 with NAMNEST composite reflectivity (dBZ, shaded).

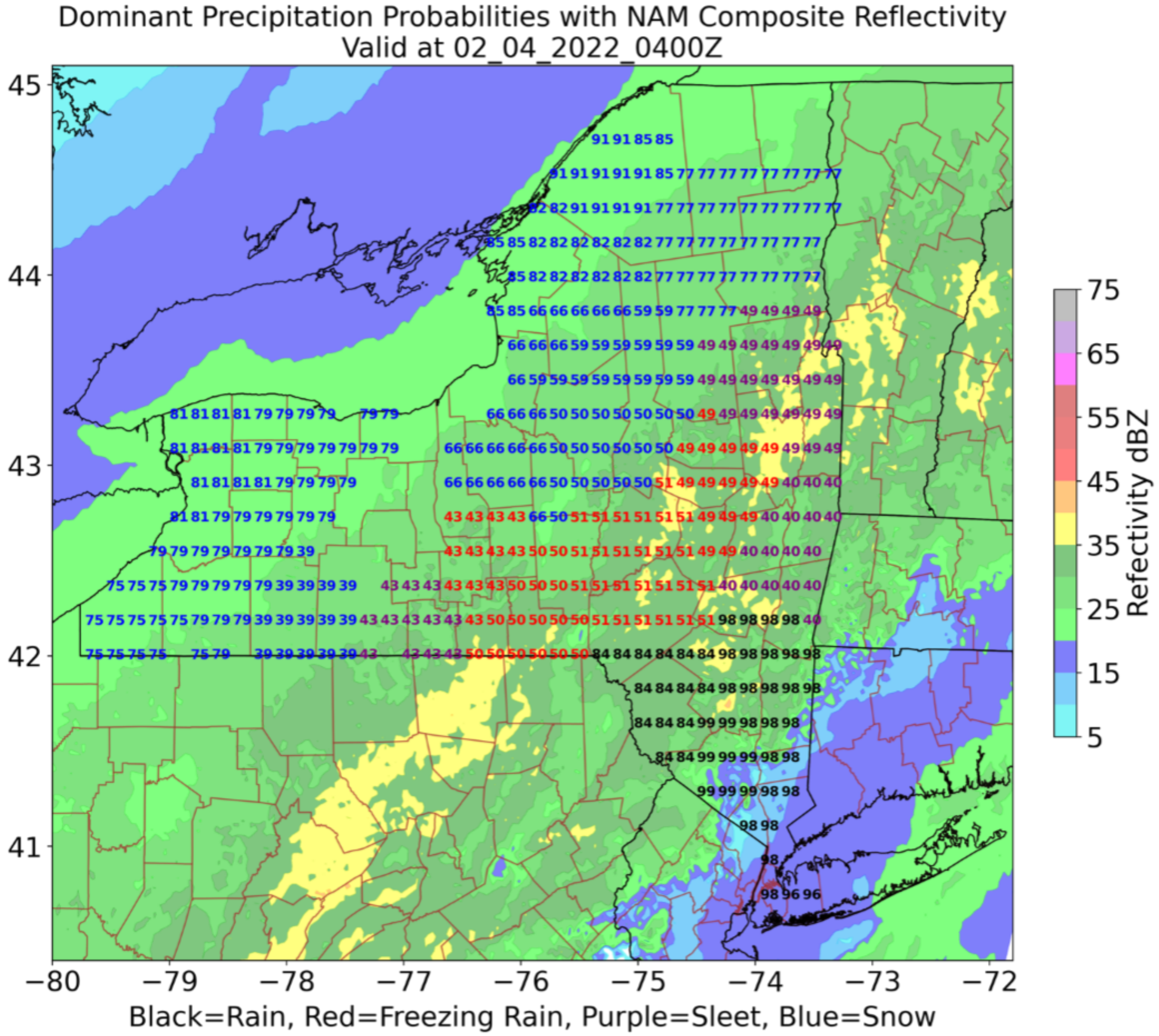

Dominant Precipitation Probabilities with NAM Composite Reflectivity
Valid at 02_04_2022_0400Z
45
44
43
42
41
−80
−79
−78
−77
−76
−75
−74
−73
−72
Black=Rain, Red=Freezing Rain, Purple=Sleet, Blue=Snow
Refectivity dBZ
75
65
55
45
35
25
15
5

Fig. 4.6: Forecast probabilities of the dominant precipitation type at each location (colored text, color key underneath plot) from the 0000 UTC NAMNEST model run at forecast hour four on 4 February 2022 with NAMNEST composite reflectivity (dBZ, shaded).

In addition to the operational website, a public archive was developed to allow end users to view products for previous dates at their own convenience. This archive updates every 4 h and contains archived forecast and verification products for NYSM and upper-air, NYSM and upper-air reduced, and NAMNEST from the winters of 2020–2021 and 2021–2022. This archive will be updated in real time for the winter of 2022–2023 where all forecast guidance products (NYSM and upper-air, NYSM and upper-air reduced, NAMNEST, HRRR, HRRR and NYSM, and model observation comparison) will be available. This will allow forecasters to see previous runs of any of the RF models they want for any winter dates.

4.3. Partnership with NWS Albany

Key to the operational component of this thesis were the interactions and feedback from NWS Albany. As a part of the development of this operational website and RF models, NWS Albany focal points, operational forecasters for NWS Albany, were collaborators across all areas the project and provided feedback and recommendations, assisted with collecting feedback, and helped with testing the operational products. Discussions on what type of ground truth observations would be best to base the RF models on led to the selection of the CoCoRaHS reports. NWS Albany focal points also helped in making recommendations of data combinations and variables they wanted included. This collaboration came to fruition via products like the HRRR and NYSM product, and trying to increase the spatial resolution of the RF model data. At the same time, NWS Albany was influential in the operational component of the project. They were extremely helpful in making suggestions for how the products could be displayed in a tool that fit into their forecasting framework. Not only did they provided suggestions on the how operational products could be designed to make the most sense to a forecaster viewing a product, but they helped collect feedback from forecasters on what could be improved with the website and its products.

Lastly, NWS Albany focal points helped by evaluating the product in real time and asking questions as to why certain results were seen. This occurred mostly during the winter of

2021–2022. This collaboration helped to make the product better and easier for forecasters to understand. The operational website was mentioned in an area forecast discussion (AFD) that was issued at 9:34 PM EST on 3 February2022, during a highly-impactful winter storm that had a large area of mixed precipitation. The AFD read:

> Based on reports from spotters, social media and data from the NY State Mesonet, and experimental precipitation type CSTAR output, sleet and freezing rain occurring to the Johnstown/Amsterdam area and even near the Herkimer sawtooth. Some slight reductions in the snow forecasts out there and a light increase in the ice forecasts from the Capital Region east and south. (e.g., NOAA NWS 2022b)

This AFD illustrates how the operational tool was valuable for forecasters to utilize and consider the output from the RF model.

5. Verification of Random Forest Forecast Guidance

5.1. Winters of 2020–2021 and 2021–2022

Section 2.3.2 described in detail how verification of operational RF forecasts would occur. This section will focus on how the RF performed over a period of two winters (2020–2021 and 2021–2022), both deterministically and probabilistically. To perform the analysis of the two winters, RF forecasts made from 1 November to 31 March of each year were evaluated with the results from each forecast being combined to create evaluation metrics. Only verification for the NYSM and upper air, NYSM and upper-air reduced, and NAMNEST will appear here as the HRRR, and HRRR and NYSM, products were developed to start in the winter of 2022–2023.

First, an evaluation from a deterministic perspective will be provided. Fig. 5.1 displays the confusion matrix for the entire two-winter period for each of the three products. As mentioned in Section 4.2.1, the NYSM and upper-air forecasts produced a large number of sleet predictions in the winter of 2021–2022. This over forecast of sleet led to the development of the NYSM and upper-air reduced product. The confusion matrix for the NYSM and upper-air product (Fig. 5.1a) indicates there is a majority fraction of sleet prediction in each observed precipitation type. While 93.4% of sleet cases were accurately predicted for the NYSM and upper-air product, this was most likely a by-product of all predictions being sleet predictions. There were 20,256 sleet predictions over the two-winter period compared to 970 rain, 328 freezing rain, and 8,864 snow predictions. This result clearly indicates that the training dataset, while sufficient in internal testing, was not effective at making precipitation type determinations in an operational setting.

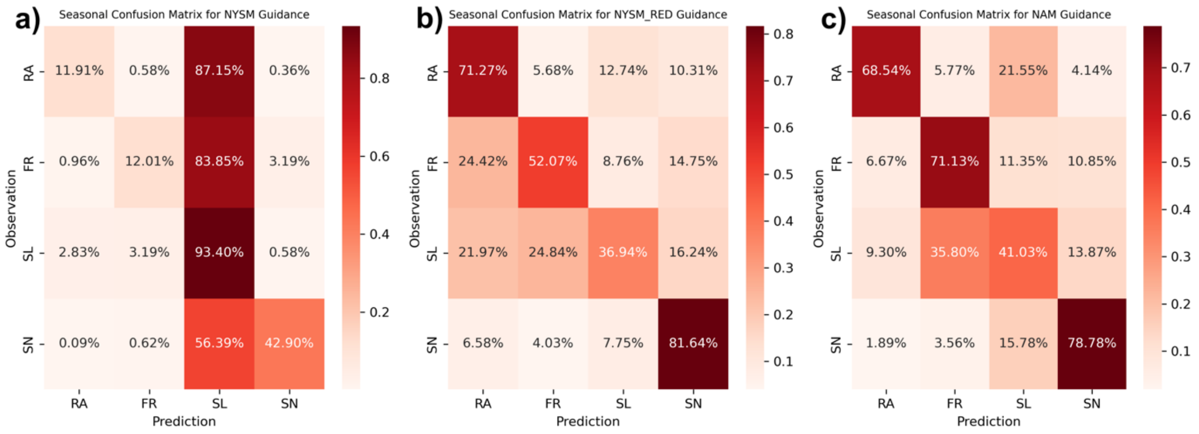


Fig. 5.1. Confusion matrices for NYSM and upper air (5.1a), NYSM and upper-air reduced (5.1b), and NAMNEST (5.1c). The numbers show the percentage of observations that occur in each of the boxes for the two-winter evaluation period. Darker red boxes indicate a higher percentage of observations in that box.

The NYSM and upper-air reduced product, developed to replace the original NYSM and upper-air product, did show significant improvement in its ability to forecast different precipitation types (Fig. 5.1b). The diagonal from top left to bottom right (diagonal of correct forecasts) had the highest percentages in each row. Snow and rain were well forecast with correct prediction percentages above 80% and 70%, respectively. Freezing rain was predicted corrected over 50% of the time, mostly being confused with rain predictions when incorrectly forecasted (24% of the time). Sleet observations were not well forecasted (about 37% correct); incorrect sleet forecasts were generally split evenly across the other precipitation types. These results indicate that while the NYSM and upper-air reduced product is an improvement, data used in the training dataset did not have much distinction between sleep and the other precipitation types.

The NAMNEST forecast guidance was the most successful of all three products. Overall correct forecasts for rain and snow were lower than the NYSM and upper-air reduced product by a couple percentage points to 68.5% and 78.8%, but the mixed precipitation forecasts (freezing rain and sleet) increased in their accuracy. Freezing rain observations were correctly predicted

71.1% over the two winters and sleet observations were correctly predicted 41% of the time. While correct sleet forecasts were still significantly less likely than any of the other precipitation types, by nearly 30% on average, the NAMNEST was still the most successful prediction of sleet of the three products. Another interesting result is that of the sleet observations, the highest incorrect forecast prediction was freezing rain, occurring in 76.8% of the observed sleet cases. This result is not surprising because of the difficultly in deciphering between freezing rain and sleet, but it also gives confidence in that the RF is making realistic predictions.

One of the other important things to consider with the deterministic verification is how the NAMNEST forecasts made predictions because they can help forecasters be more confident, particularly at times further away from model initialization time. Fig. 5.2 shows the confusion matrices for both the NAMNEST analysis (Fig. 5.2a) and forecast periods (Fig. 5.2b). These confusion matrices looked very similar to each other and the overall confusion matrix for the two-year period. A slight drop was seen in overall accuracy (diagonal from top left to bottom right) from the analysis to the forecast periods, as expected, but only by a handful of percentage points.

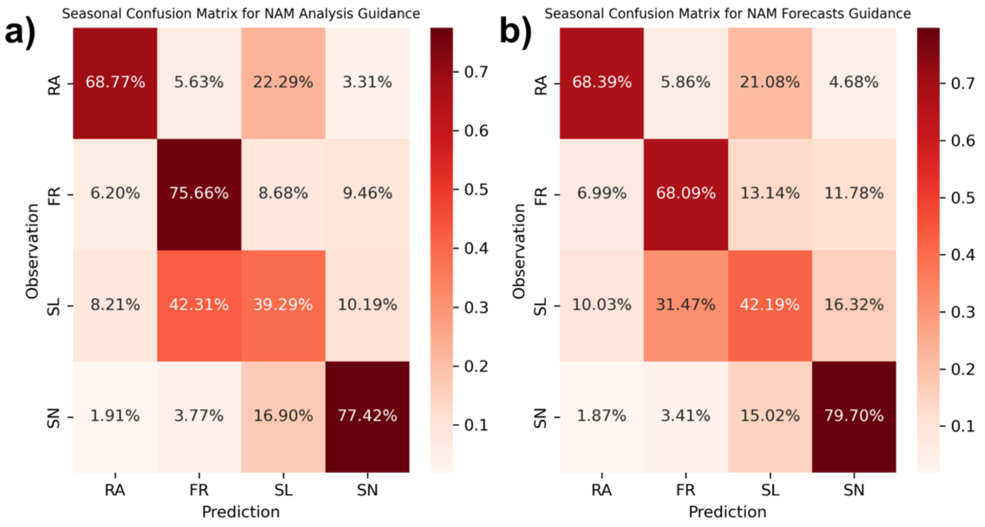
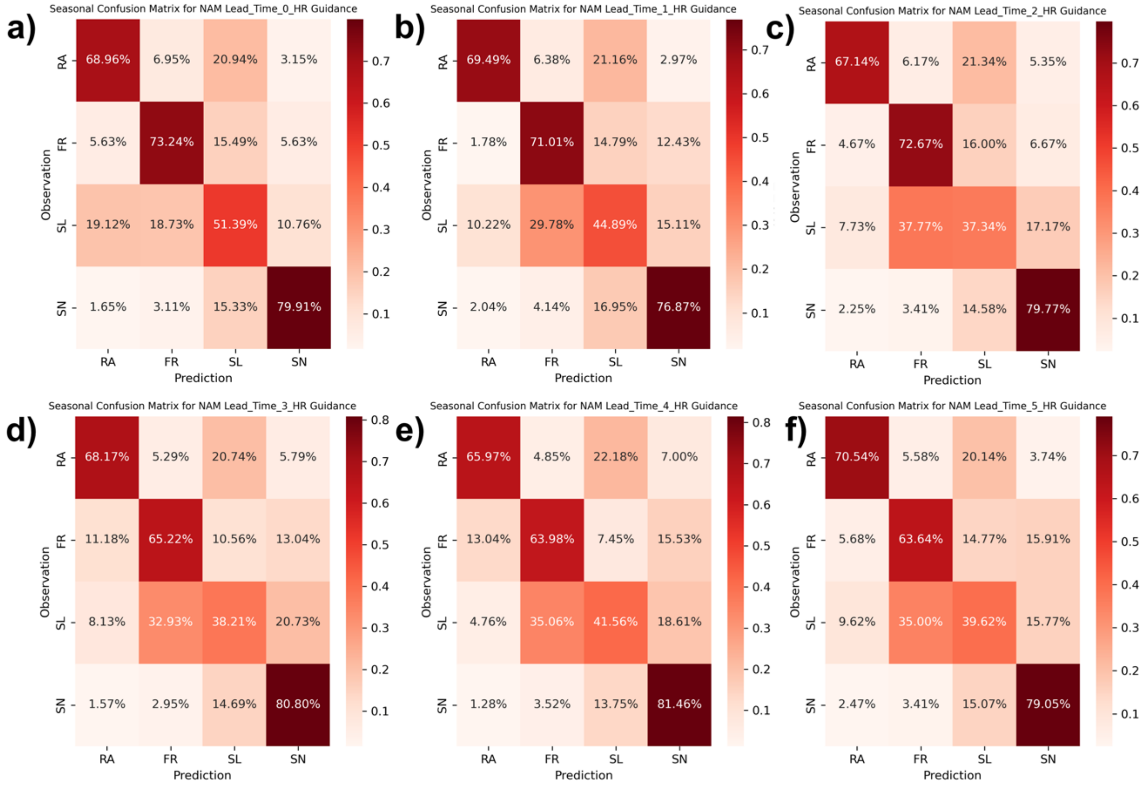


Fig. 5.2. Confusion matrices for NAMNEST separated by forecasting period:(a) is the matrix for forecast hours 0–3 (the analysis) and (b) is forecast hours 4–9 (the forecasts). The numbers show the percentage of observations that occur in each of the boxes for the two-winter evaluation period. Darker red boxes indicated a higher percentage of observations in that box.

Along with comparing the analysis and forecast periods for the NAMNEST, it was important to examine the forecast period specifically to see if forecast skill dropped off significantly with increased lead time. The confusions matrices for NAMNEST by lead time is shown in Fig. 5.3. Generally, all the confusion matrices had a similar appearance. Rain and snow correct forecast percentages were fairly consistent through all six of the lead time periods and only vary by 2–3 percent in either direction. Freezing rain and sleet both experienced a steadier decrease across the time periods, dropping by about 10% on average from the nowcast (0 h lead time).



a) Seasonal Confusion Matrix for NAM Lead_Time_0_HR Guidance

Observation \ Prediction	RA	FR	SL	SN
RA	68.96%	6.95%	20.94%	3.15%
FR	5.63%	73.24%	15.49%	5.63%
SL	19.12%	18.73%	51.39%	10.76%
SN	1.65%	3.11%	15.33%	79.91%

b) Seasonal Confusion Matrix for NAM Lead_Time_1_HR Guidance

Observation \ Prediction	RA	FR	SL	SN
RA	69.49%	6.38%	21.16%	2.97%
FR	1.78%	71.01%	14.79%	12.43%
SL	10.22%	29.78%	44.89%	15.11%
SN	2.04%	4.14%	16.95%	76.87%

c) Seasonal Confusion Matrix for NAM Lead_Time_2_HR Guidance

Observation \ Prediction	RA	FR	SL	SN
RA	67.14%	6.17%	21.34%	5.35%
FR	4.67%	72.67%	16.00%	6.67%
SL	7.73%	37.77%	37.34%	17.17%
SN	2.25%	3.41%	14.58%	79.77%

d) Seasonal Confusion Matrix for NAM Lead_Time_3_HR Guidance

Observation \ Prediction	RA	FR	SL	SN
RA	68.17%	5.29%	20.74%	5.79%
FR	11.18%	65.22%	10.56%	13.04%
SL	8.13%	32.93%	38.21%	20.73%
SN	1.57%	2.95%	14.69%	80.80%

e) Seasonal Confusion Matrix for NAM Lead_Time_4_HR Guidance

Observation \ Prediction	RA	FR	SL	SN
RA	65.97%	4.85%	22.18%	7.00%
FR	13.04%	63.98%	7.45%	15.53%
SL	4.76%	35.06%	41.56%	18.61%
SN	1.28%	3.52%	13.75%	81.46%

f) Seasonal Confusion Matrix for NAM Lead_Time_5_HR Guidance

Observation \ Prediction	RA	FR	SL	SN
RA	70.54%	5.58%	20.14%	3.74%
FR	5.68%	63.64%	14.77%	15.91%
SL	9.62%	35.00%	39.62%	15.77%
SN	2.47%	3.41%	15.07%	79.05%

Fig. 5.3. Confusion matrices for NAMNEST separated by hours of lead time. Forecast hour four is defined as zero hours of lead time since that is when the NAMNEST product is available on the website. The numbers show the percentage of observations that occur in each of the boxes for the two-winter evaluation period. Darker red boxes indicated a higher percentage of observations in that box.

Along with verifying RF forecast guidance in a deterministic sense, it was imperative to also verify probabilistically because understanding what the displayed probabilities mean is important to forecasters and for understanding how to better frame the RF guidance in the context of operational forecasts. Figs. 5.4–5.6 display the observed frequencies for the probability distribution for each of the winter precipitation types correct and incorrect predictions across the two-winter period. In addition, these figures show the percentage of correct predictions in each probability bin, located at top of each bar. This information is important because, ideally, the percent of correct predictions should match the probability bin. The pattern that should appear is that higher predicted probabilities should equal more correct predictions. Examining this part of the RF guidance will allow for better understanding of what the probabilities mean and for ways to correct the guidance to be more representative.

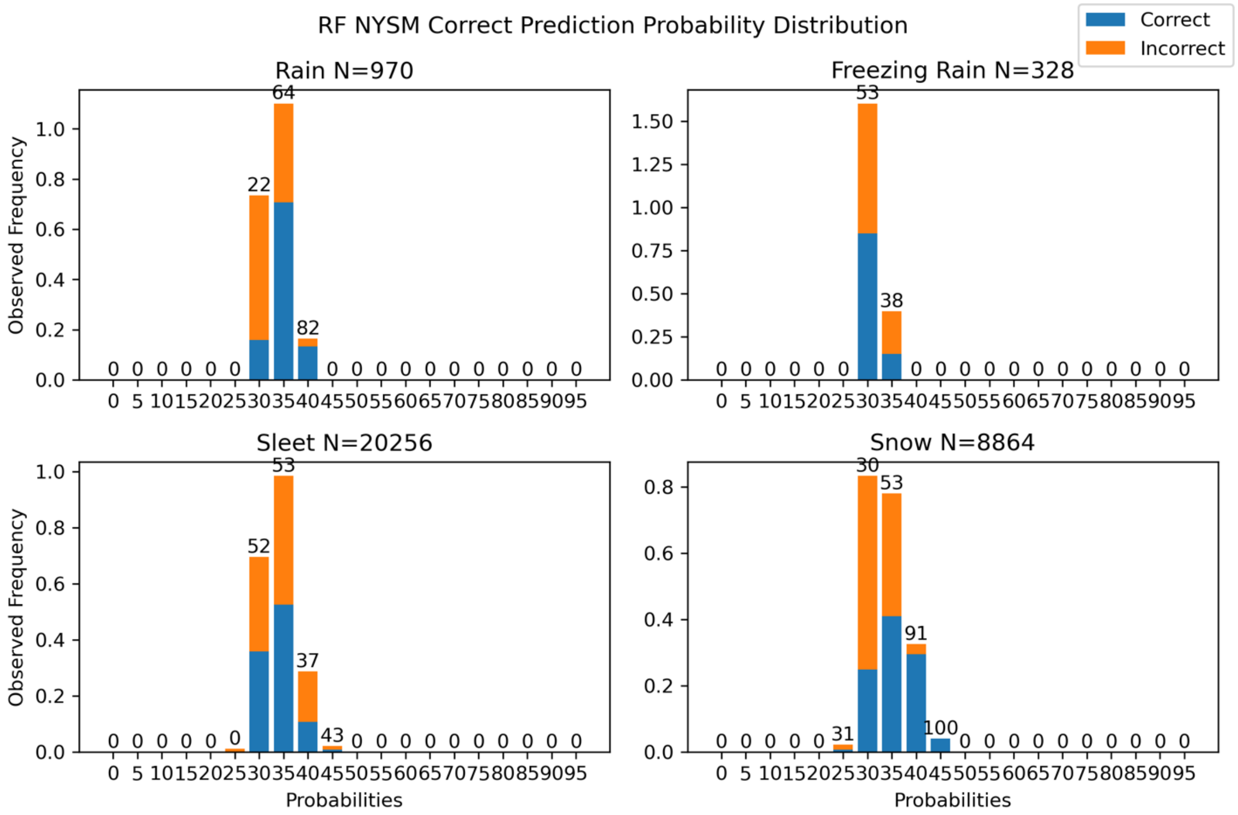


Fig. 5.4. Observed frequencies for the probability distribution of both correct and incorrect predictions of NYSM and upper-air RF product for all precipitation types. N value is number of predictions for each type of precipitation in the two-winter observing period. The number on the top of the frequency bars represents the percentage of correct predictions in each bin.

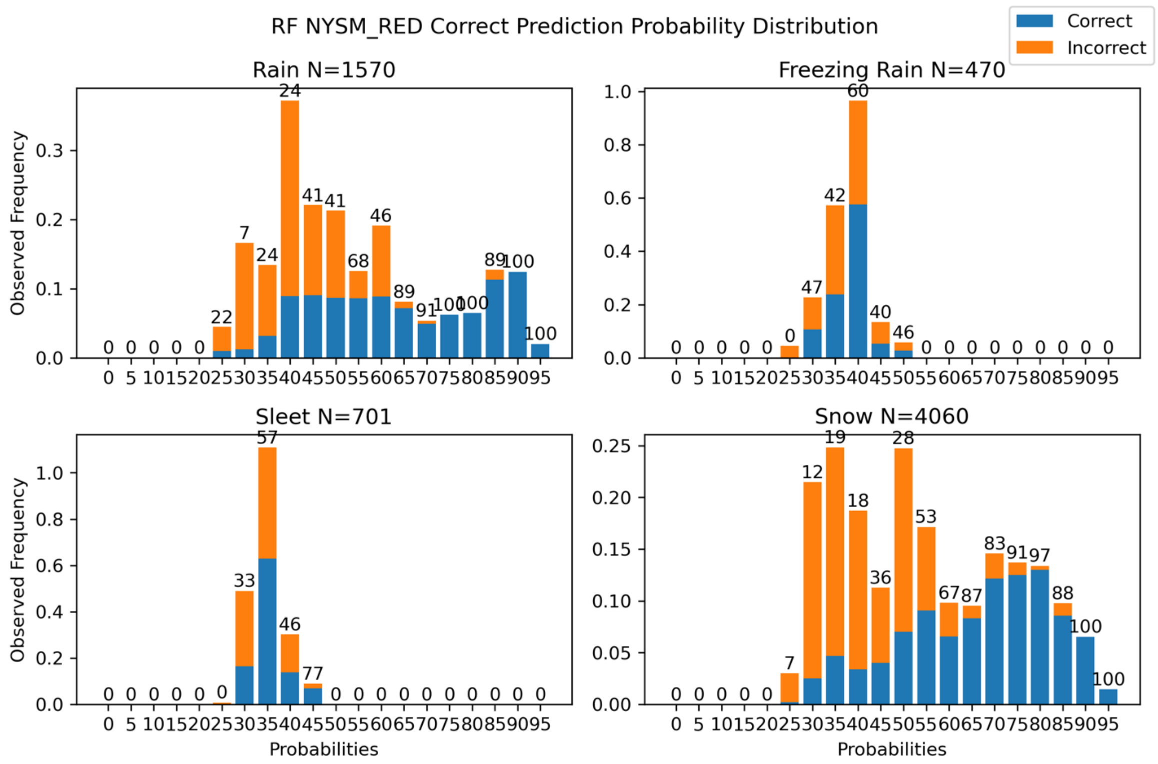


Fig. 5.5. Same as in Fig. 5.4, but for NYSM and upper-air reduced RF product.

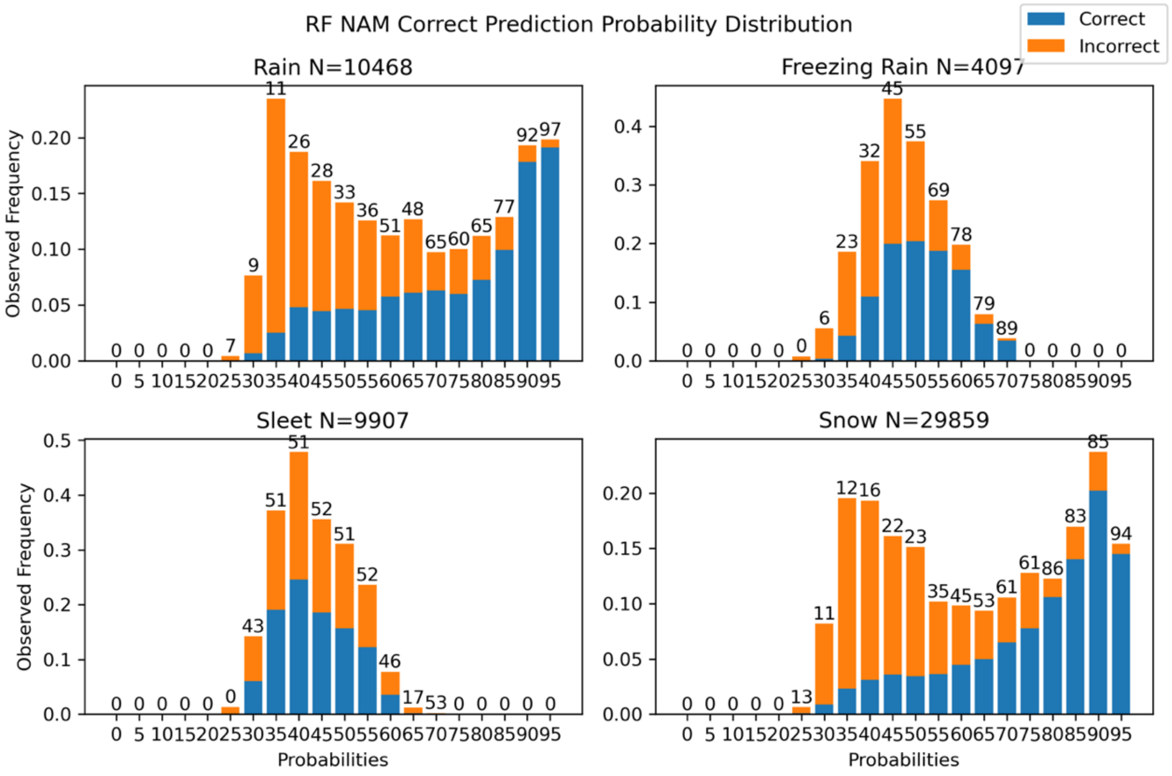


Fig. 5.6. Same as in Fig. 5.4, but for NAMNEST RF product.

Examining the probability distribution for the NYSM and upper-air product (Fig. 5.4), trends that were visible in the confusion matrix appear here as well. There is a narrow distribution of probabilities which is not ideal in an operational forecast tool. In addition, the ideal trend of increased probability indicating more likelihood of occurrence did not appear for all precipitation types: it is true for rain and snow, but not the mixed precipitation types. This distribution of predictions and the limited range of probabilities reinforced that developing the NYSM and upper-air reduced product was a good decision.

There was a significant difference between the NYSM and upper-air reduced product (Fig. 5.5) and the original NYSM and upper-air product (Fig. 5.4). The probability distribution for the NYSM and upper-air reduced product was much broader and what a probabilistic forecast distribution should, ideally, look like. The distribution had a larger range than in Fig. 5.4 and it matched with a more realistic distribution of RF forecasts for each precipitation types. For the most part, the ideal trend of increased forecast probability leading to increased likelihood of occurrence held true: only freezing rain was unreliable in this sense. Interestingly, the distribution of probabilities for the non-mixed precipitation types was much larger than the mixed precipitation types. While this may be due to limited sample size, it could, more importantly, point to the fact that the mixed precipitation types are harder to predict, so there was less of a consensus with these types of precipitation.

The NAMNEST product (Fig. 5.6) was more similar to the NYSM and upper-air reduced distribution (Fig. 5.5) than the original NYSM and upper-air distribution (Fig. 5.4). The distribution across the four precipitation types was realistic and had a wide range of values. The ideal distribution theory was generally true here as in Fig. 5.5, where three of four precipitation types followed the theory, with sleet being the outlier. One interesting finding with the rain and snow distributions was that once the probabilities get up to 70 or 80%, the accuracy of correct predictions was generally at or higher than the RF output probability. This result was important to convey to forecasters, so that they are aware of what the RF probabilities actually mean. A similar point can be made for the freezing rain distribution as well freezing rain probabilities were maximized in the 70–75% probability bin and the percentage of correct predictions indicated that 89% of cases verified. This fact is important to convey because the percentages displayed must give an accurate representation of the likelihood of occurrence to the forecasters using the product. By communicating this upgrade in likelihood to forecasters, the RF product can be used more effectively.

5.2. Case Studies

5.2.1. 15–16 February 2021

The winter storm of 15–16 February 2021 produced significant sleet and freezing rain across much of NY state. The storm formed on 15 February associated with a deep trough across the western and central US (Fig. 5.7a,b). During that time, there was a narrow area with a strong 850-hPa temperature gradient in western and central NY (Fig. 5.8a,b). This area of enhanced temperature gradient was just on the colder side of 0°C. At the surface, high pressure was present across NY state (Fig. 5.9a–c).

At 0000 UTC 16 February, a strong jet had developed in Ontario and Quebec (Fig. 5.7c) bringing strong upper-level southwesterly winds across southern Ontario and Quebec as well as the Northeast United States, including NY. Warm air advection was occurring at lower levels (850 hPa) across central and eastern NY (Fig. 5.8c). Temperatures at this level were slightly above freezing to just below freezing indicating the potential for a profile conducive to mixed precipitation. After 0000 UTC, the main precipitation shield entered NY. During the period of the heaviest precipitation (0300 to 1200 UTC), an interesting surface set up existed. At 0600 UTC, a double surface low was pushing closer to NY with one low located inland at the intersection of Ohio, Pennsylvania, and West Virginia, and the other low off the Delmarva coast (Fig. 5.9d). The surface low location and previous path, from the Gulf of Mexico, had a classic Miller type A nor'easter storm setup. By 0900 UTC, the double surface low was slightly more inland with the interior low at the intersection of Ohio, Pennsylvania, and New York, and the coastal low located over Virginia (Fig. 5.9e). This setup coincided with most of the precipitation being located in the eastern half of NY. By 1200 UTC, most of the precipitation had exited NY. At upper levels, most of the country was engulfed in a deep trough and winds were out the south or southwest in NY (Fig. 5.7d). At lower levels, there was significant warm air aloft with the upper air analysis at 850 hPa indicating 7°C at Albany as well as winds out the south or southwest (Fig. 5.8d). At the surface, the interior low was located over the Southern Tier, while the other low was over central New Jersey (Fig. 5.9f). This pattern created a double warm front setup in NY with temperatures generally above freezing across the state.

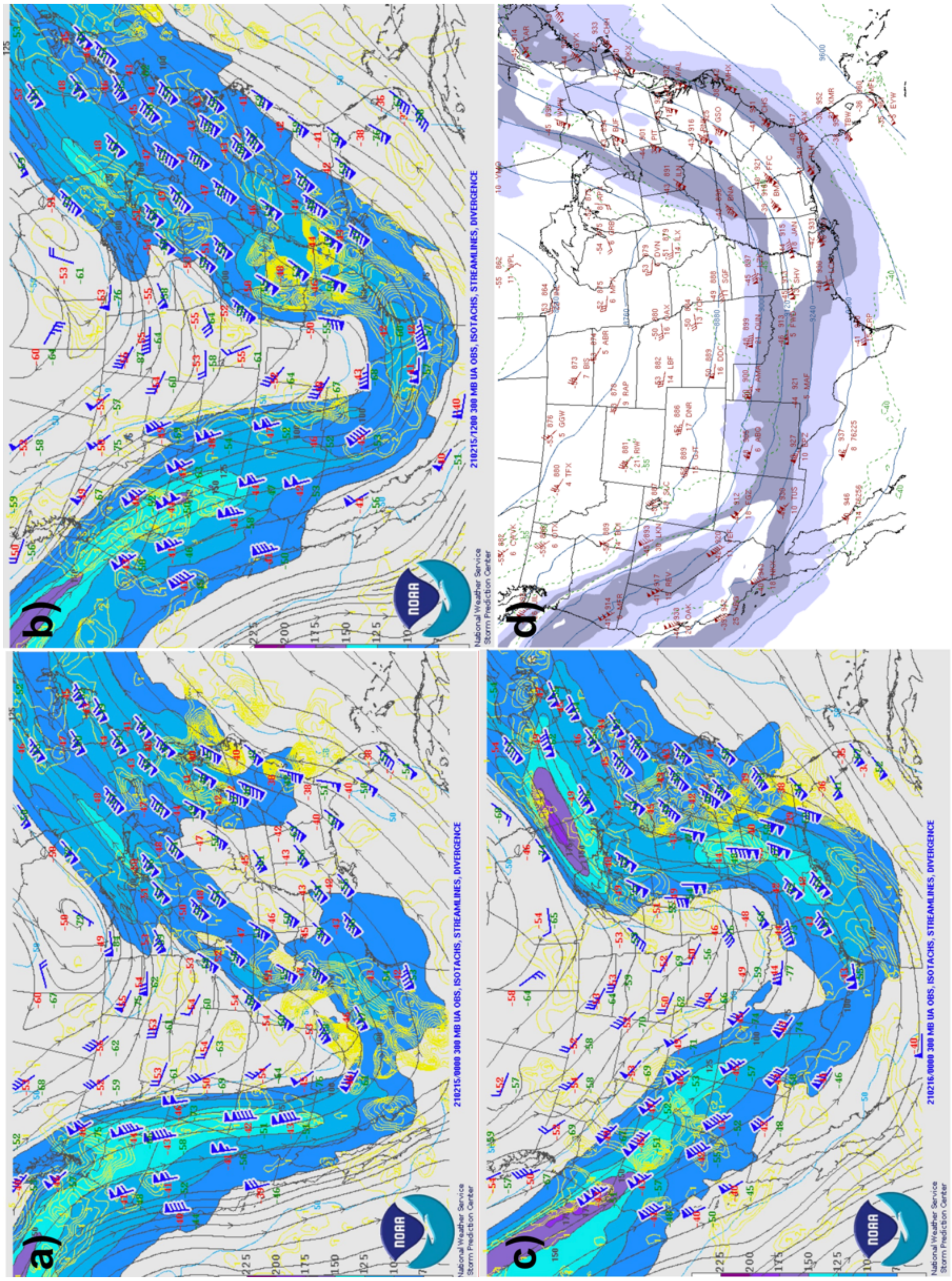

a)
b)
c)
d)
210215/0000 300 MB UA OBS, ISOTACHS, STREAMLINES, DIVERGENCE
210215/1200 300 MB UA OBS, ISOTACHS, STREAMLINES, DIVERGENCE
210216/0000 300 MB UA OBS, ISOTACHS, STREAMLINES, DIVERGENCE
National Weather Service
Storm Prediction Center
NOAA
225
200
175
150
125
100
75

Fig. 5.7. Storm Prediction Center 300-hPa analysis at a) 0000 UTC 15 February, b) 1200 UTC 15February, c) 0000 UTC 16 February, and d) 1200 UTC16 February 16 2021.

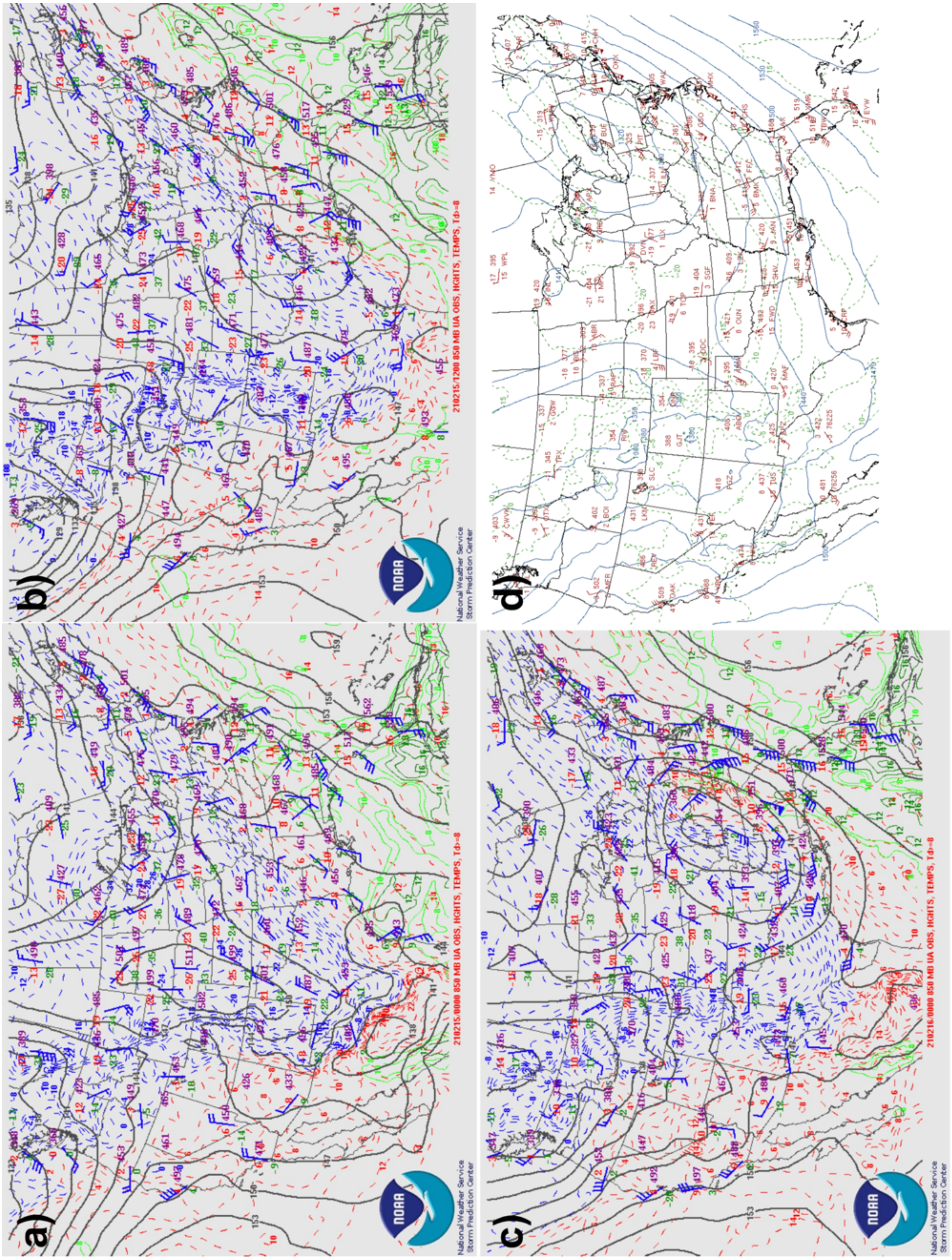


Fig. 5.8. Same as Fig. 5.7, except for 850 hPa.

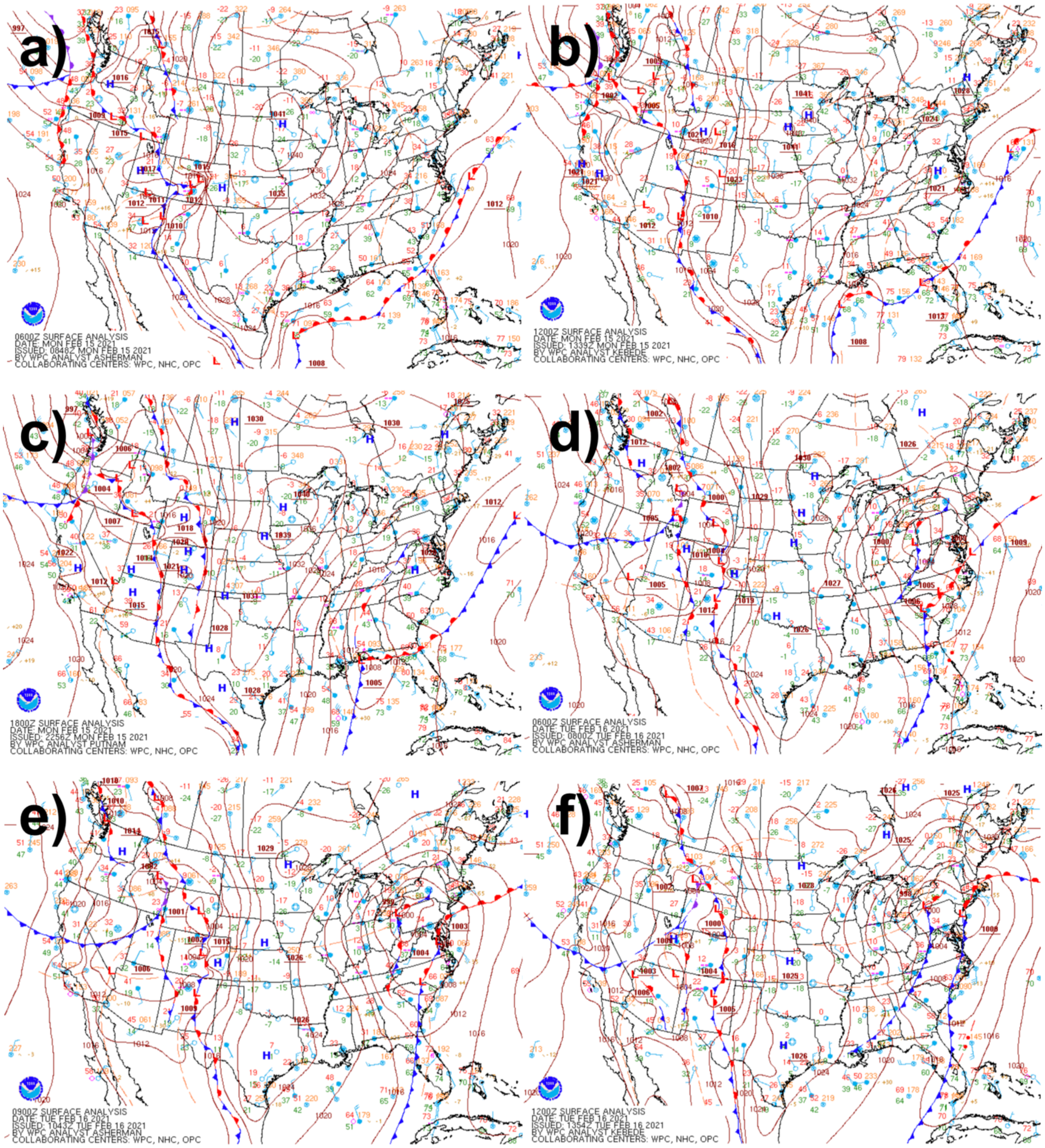


Fig. 5.9. Weather Prediction Center surface analysis at a) 0600 UTC 15 February, b) 1200 UTC 15 February, c) 1800 UTC15 February, d) 0600 UTC 16 February, e) 0900 UTC16 February, and f) 1200 UTC16 February 2021.

Looking at the NWS local storm reports (LSRs) from the storm (Fig. 5.10), the heaviest snowfall tended to be in western NY around Buffalo. There was an area of transitioning precipitation type across the Southern Tier and south-central NY where most of the sleet reports occurred. Freezing rain was mostly in eastern NY and the Hudson Valley area, likely due to the warm air advection with the south/southwesterly winds; however, there were other freezing rain reports scattered across the state, including a significant freezing rain report in Oswego, near Syracuse, of over 4.1 in of flat ice. In the Hudson Valley and Capital District Areas, flat ice totals were between 0.5 and 0.75 in.

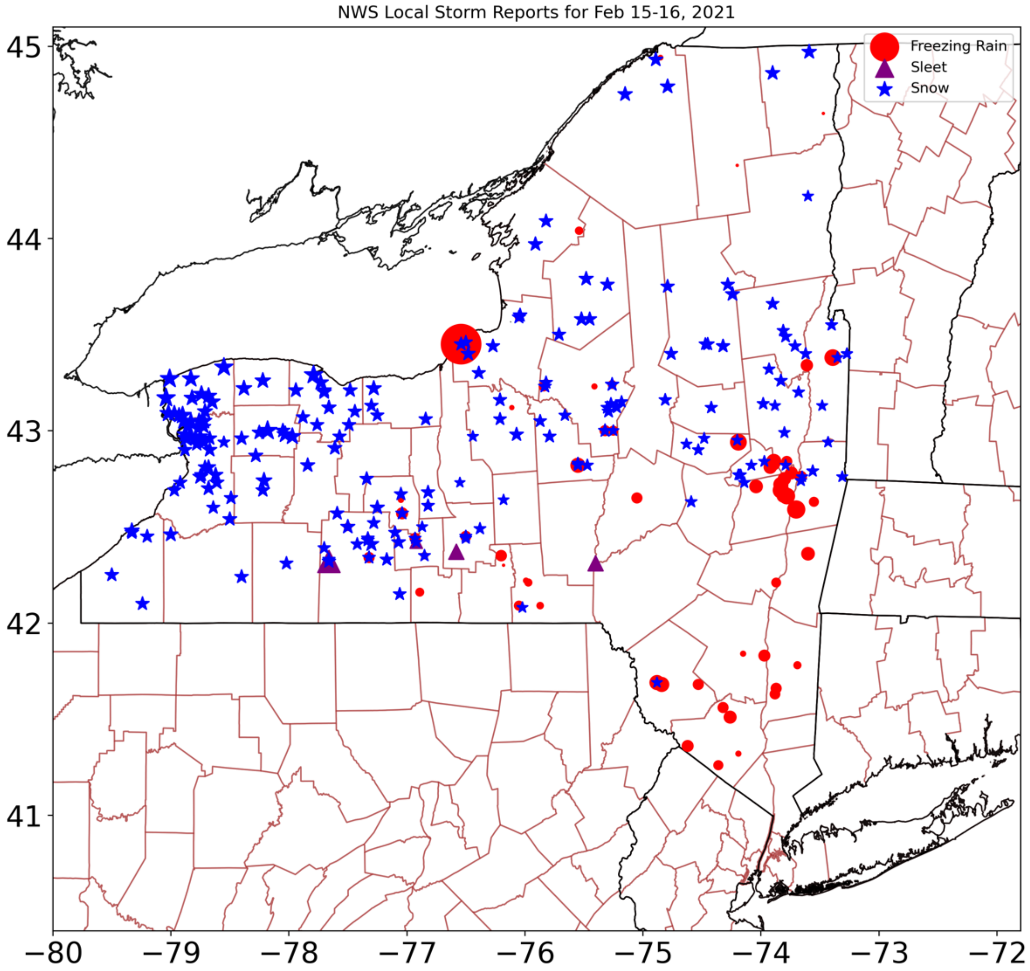


Fig. 5.10. NWS LSR map of wintry precipitation types in NY for 15–16 February2021. The marker size varies according to the magnitude of the precipitation reported.

In examining how the RF forecast guidance did deterministically, there are some similar trends to the results of the verification for the two full winters (Section 5.1). The original NYSM and upper-air forecasts continued to issue mostly sleet predictions with a handful of other precipitation types mixed in (Fig. 5.11a). The NYSM and upper-air reduced forecasts were much more accurate. While most of the freezing rain and snow reports (83.33% and 69.70%) were predicted correctly (Fig. 5.11b), the RF struggled with both sleet and rain during this event. For both of these precipitation types, about half of the predictions were for freezing rain (50% and 44.44%). It was worth noting that the second highest predicted precipitation type for snow reports was also freezing rain. This result indicates that the NYSM and upper-air reduced guidance may have overpredicted freezing rain during this event. The NAMNEST RF forecasts were very similar to the NYSM and upper-air reduced guidance: it did well in correctly predicting 73.39% of freezing rain and 72.70% of snow reports (Fig. 5.11c), but it also struggled with accurately predicting rain and sleet, being biased towards freezing rain even more so than the NYSM and upper-air reduced guidance.

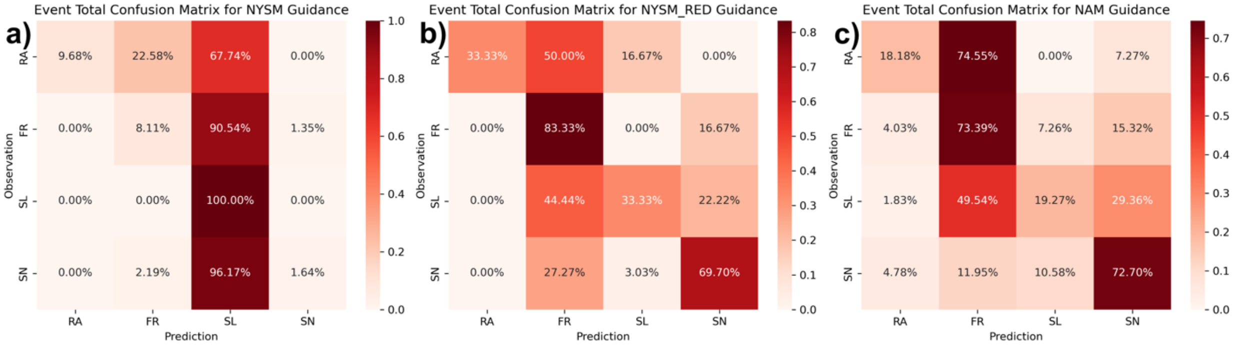


Fig. 5.11. Same as Fig. 5.1 except for 15–16 February 2021.

Examining the NAMNEST RF forecast guidance further, locations of interest were identified based on the NWS LSRs. These locations of interest were determined by locating the NWS LSRs with the largest magnitude of precipitation for each of the freezing rain, sleet, and snow reports. The first site examined was the unique freezing rain report of 4.1 in of flat ice in Oswego, NY. The probability timeseries for that location (Fig. 5.12a) did not indicate much freezing rain or mixed precipitation. Prior to 0600 UTC 16 February, snow was the likeliest precipitation type to occur. From 0600 UTC to 0900 UTC, freezing rain and sleet had similar probabilities. By the end of this period, when most of the precipitation was in eastern NY away from Oswego, the freezing rain probabilities started to increase and became dominant. This increase occurred until 1600 UTC when rain briefly became dominant and, eventually, snow returned as the dominant precipitation once the precipitation was well clear of the region. The RF output did not corroborate the magnitude of freezing rain seen in Oswego due to it only appearing as dominant for a short period of time.

The second site examined was Hornell, NY in the Southern Tier where 1.3 in of sleet fell during the event. The probability timeseries for Hornell did indicate the potential for mixed precipitation. Prior to 0000 UTC 16 February, snow was the dominant probability for precipitation type. Starting at 0000 UTC, mixed precipitation types became dominant. From 0000 to 0400 UTC, freezing rain and sleet were both close to each other in terms of probabilities, and they switched back and forth between being dominant. After 0400 UTC, freezing rain became the dominant precipitation type until precipitation had moved well out of Hornell. It is not possible to clearly determine if the dominant precipitation output from the RF makes sense with the observed sleet totals. However, the RF was able to identify that the environment was conducive for mixed precipitation. In addition, Hornell was a significant distance, over 20 mi, from the nearest NAMNEST profile site used to generate probabilities; the NAMNEST site is also about 1,000 ft higher in elevation than Hornell, which could impact the vertical profile and final results.

The final site examined was in Batavia, NY, where 13 in of snow fell during the event. The probability timeseries for Batavia (Fig. 5.12c) looked very representative of the precipitation that occurred. Snow was the dominant precipitation predicted by the NAMNEST RF for the entirety of the event. Only around 0900 UTC the probability of sleet did get close to the probability of snow. Since snow was dominant for so long, it is possible to conclude that the RF output matched well with the precipitation observed in Batavia. Overall, the RF output from the NYSM and upper-air reduced, and NAMNEST, models did a good job at representing freezing rain and snow forecasts, while under forecasting rain and sleet. Generally, freezing rain was over predicted leading to misses in the forecast for the rain and sleet observations.

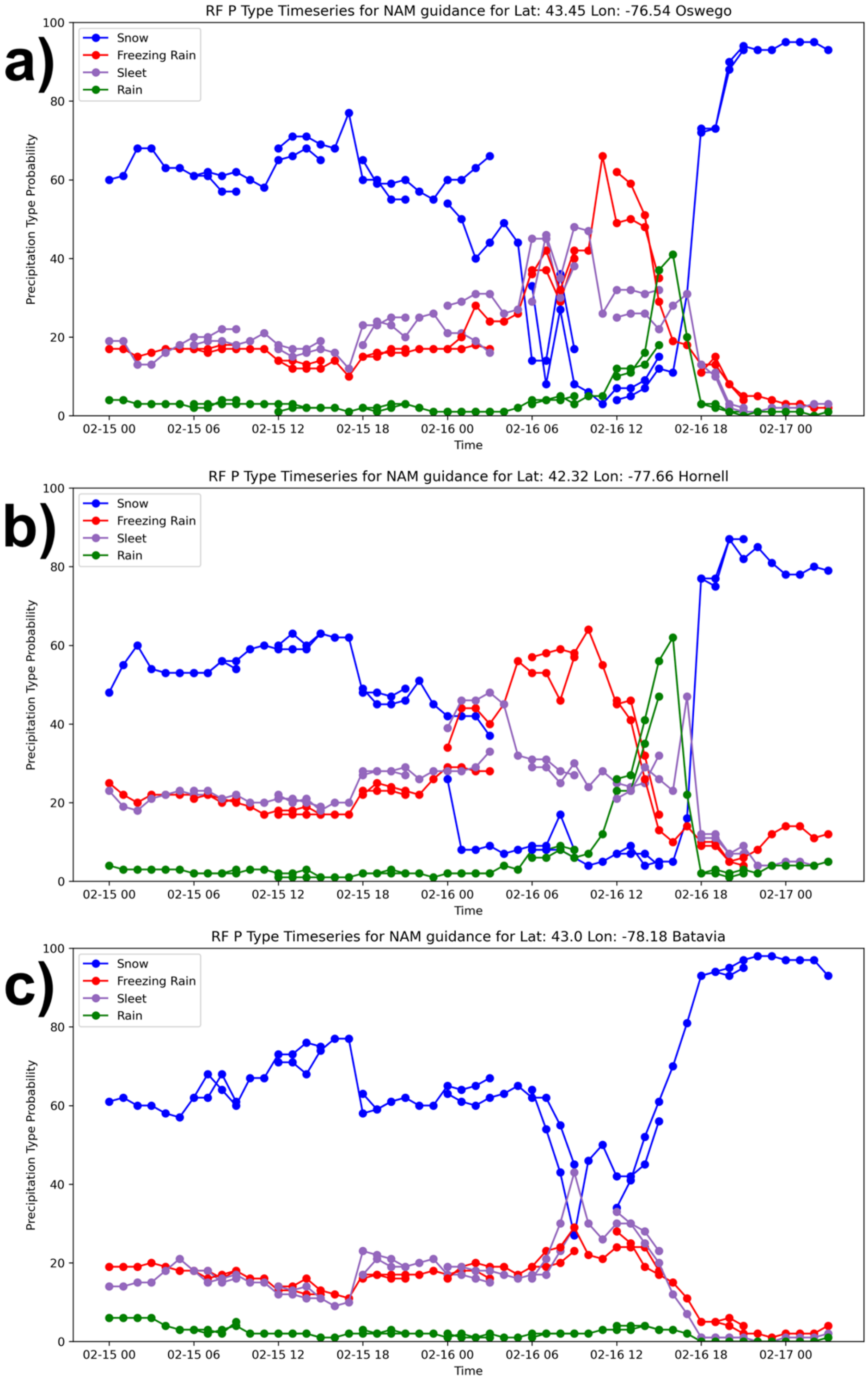


Fig. 5.12. Timeseries of RF probabilities at locations of interest during 15–16 February 15-16. NWS LSRs with the highest reported magnitude of each precipitation type were selected as the locations of interest:(a) Oswego, NY, (b) Hornell, NY, and (c) Batavia, NY

5.2.2. 9–10 January 2022

The 9–10 January winter storm had two distinct phases: on 9 January, a frontal system associated with precipitation passed through NY, and on 10 January, a prolonged lake-effect snow event developed across NY. At 0000 UTC 9 January, a trough–ridge pattern existed across the US. The jet was located over the upper Midwest and was associated with an upper-level low over central Canada (Fig. 5.13a). At lower levels, winds were from the southwest over NY with temperatures at 850 hPa in ranging from right above 0°C to near -10°C (Fig. 5.14a). There was light precipitation across western NY.

Approaching 1200 UTC, the first round of light precipitation moved through most of NY with a second round starting in western NY. At upper levels, the winds had strengthened across the Northeast and became a bit more southwesterly (Fig. 5.13b). At lower levels, the southwesterly winds had strengthened and the temperatures across NY were split, below freezing in western NY and above freezing in eastern NY (Fig. 5.14b). This setup of warm air advection, as well as the temperature gradient in NY, was favorable for mixed precipitation, particularly in eastern NY. At the surface, there was a warm front across northern NY, and southern Quebec and Ontario (Fig. 5.15a). The cold front associated with a low-pressure system in Hudson Bay stretched down through Michigan all the way to Texas. By 1800 UTC, most of the main precipitation had moved through NY leaving only lake effect precipitation across Lake Erie and Lake Ontario. At the surface, winds off the lake were from the southwest, which was very conducive for lake-effect precipitation (Fig. 5.15b).

Between 1800 UTC 9 February and 0000 UTC 10 January, non lake-effect precipitation was sparse across upstate NY. Some storms developed in the lower Hudson Valley and over

Long Island after 2000 UTC. At 0000 UTC, a strong upper-level jet developed over southern Ontario and Quebec with winds over NY increasing in strength (Fig. 5.13c). The winds at lower levels became more westerly and the temperatures outside New York City were below freezing at 850 hPa (Fig. 5.14c). The surface cold front split NY from Plattsburgh to Binghamton and a secondary cold front was located across the border in southern Canada (Fig. 5.15c). Starting at 0300 UTC, light lake-effect precipitation developed off of Lake Ontario and, by 0900 UTC, a very narrow lake-effect band stretched from Lake Ontario to the Capital District. At that time, both cold fronts had cleared through NY and the winds in western and central NY were westerly (Fig. 5.15d). Upper-air analysis at 1200 UTC 10 January showed the strong jet directly overly eastern NY (Fig. 5.13d) and cold westerly winds at lower levels (Fig. 5.14d).

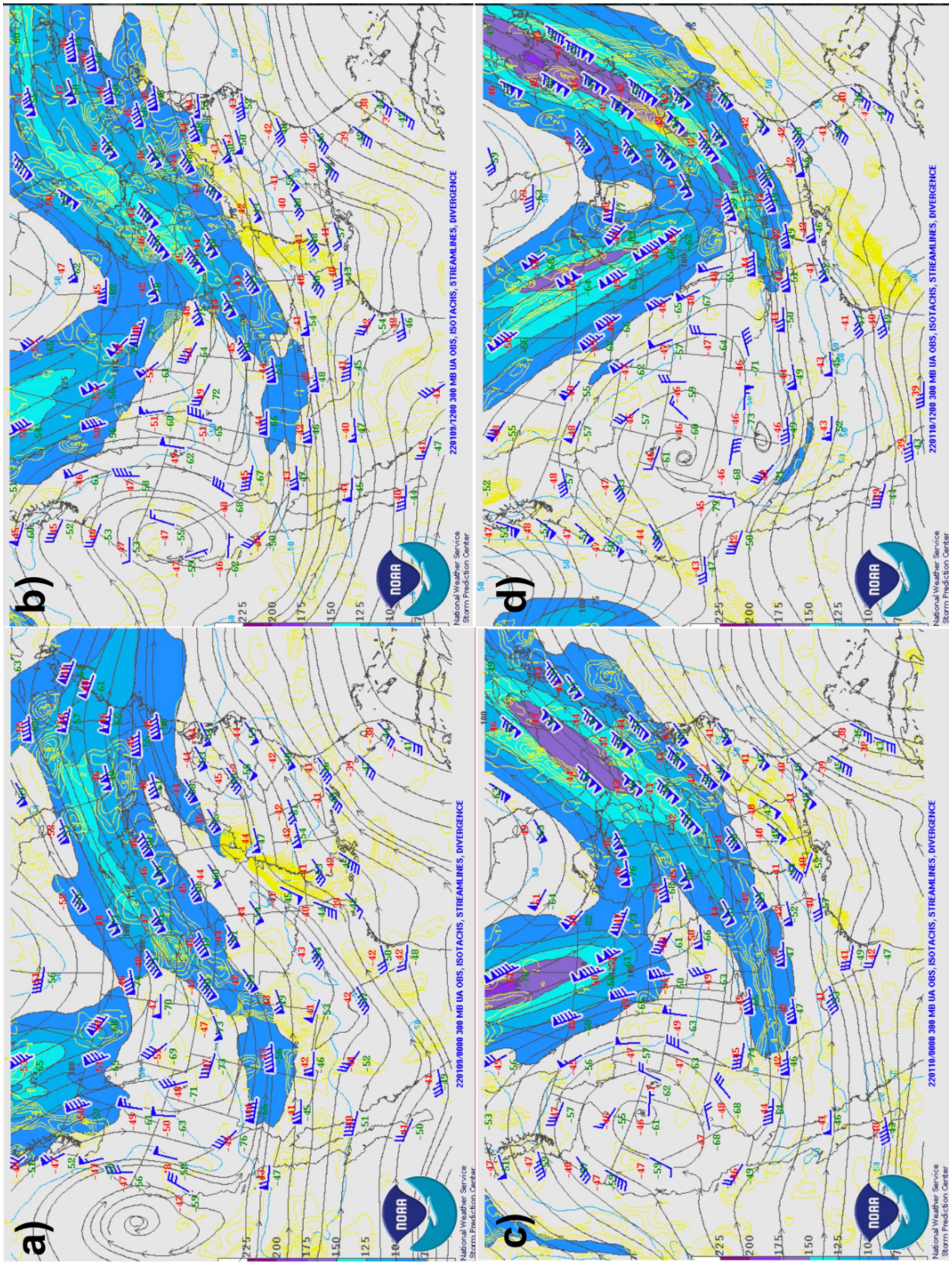

a)
b)
c)
d)
220109/0000 300 MB UA OBS, ISOTACHS, STREAMLINES, DIVERGENCE
220109/1200 300 MB UA OBS, ISOTACHS, STREAMLINES, DIVERGENCE
220110/0000 300 MB UA OBS, ISOTACHS, STREAMLINES, DIVERGENCE
220110/1200 300 MB UA OBS, ISOTACHS, STREAMLINES, DIVERGENCE
National Weather Service
Storm Prediction Center
NOAA
225
200
175
150
125
100

Fig. 5.13. Storm Prediction Center 300-hPa upper-air analysis at a) 0000 UTC 9 January, b) 1200 UTC 9 January, c) 0000 UTC 10 January, and d) 1200 UTC 10 January 2022.

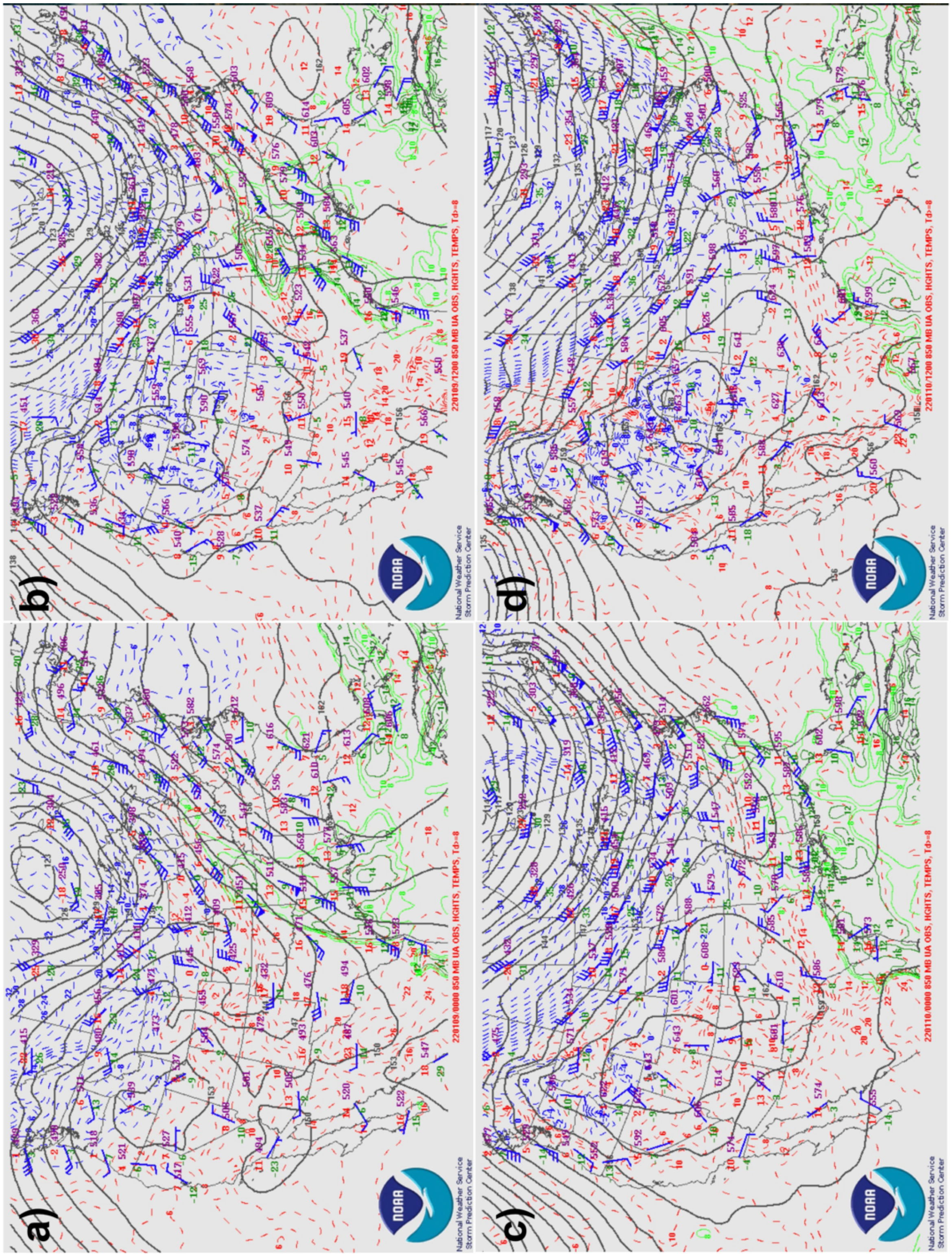

a)
b)
c)
d)
220109/0000 850 MB UA OBS, HGHTS, TEMPS, Td>-8
220109/1200 850 MB UA OBS, HGHTS, TEMPS, Td>-8
220110/0000 850 MB UA OBS, HGHTS, TEMPS, Td>-8
220110/1200 850 MB UA OBS, HGHTS, TEMPS, Td>-8
NOAA
National Weather Service
Storm Prediction Center

Fig. 5.14. Same as Fig. 5.13 except at 850 hPa.

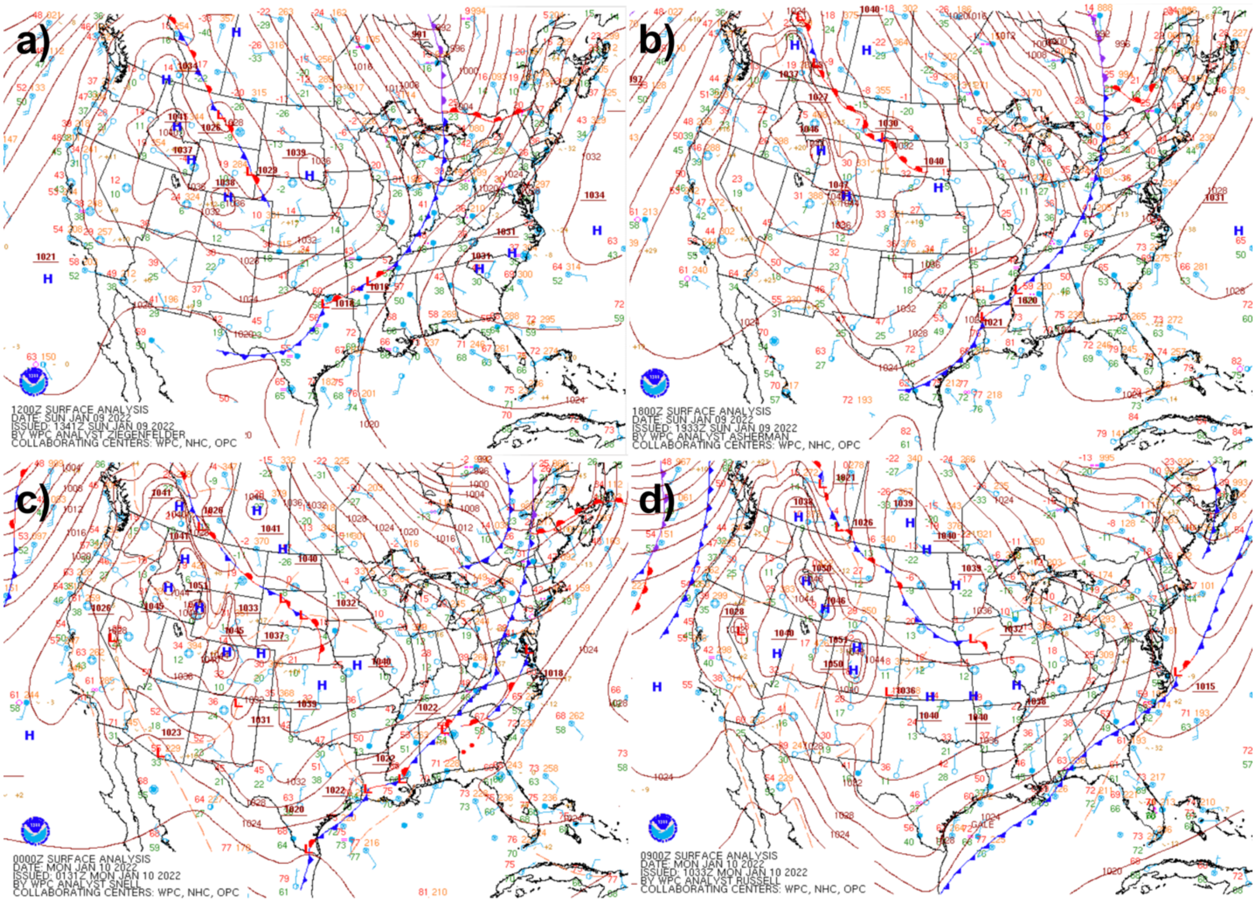

a)
1200Z SURFACE ANALYSIS
DATE: SUN JAN 09 2022
ISSUED: 1341Z SUN JAN 09 2022
BY WPC ANALYST ZIEGENFELDER
COLLABORATING CENTERS: WPC, NHC, OPC
b)
1800Z SURFACE ANALYSIS
DATE: SUN JAN 09 2022
ISSUED: 1933Z SUN JAN 09 2022
BY WPC ANALYST ASHERMAN
COLLABORATING CENTERS: WPC, NHC, OPC
c)
0000Z SURFACE ANALYSIS
DATE: MON JAN 10 2022
ISSUED: 0131Z MON JAN 10 2022
BY WPC ANALYST SNELL
COLLABORATING CENTERS: WPC, NHC, OPC
d)
0900Z SURFACE ANALYSIS
DATE: MON JAN 10 2022
ISSUED: 1033Z MON JAN 10 2022
BY WPC ANALYST RUSSELL
COLLABORATING CENTERS: WPC, NHC, OPC

Fig. 5.15. Weather Prediction Center surface analysis at a) 1200 UTC 9 January, b) 1800 UTC 9 January, c) 0000 UTC 10 January, and d) 0900 UTC 10 January 2022.

The 9–10 January event had two distinct phases across NY, clearly shown in the NWS LSRs. The map of LSRs (Fig. 5.16) shows a larger area of small to moderate snow and freezing rain totals across south central NY and the Hudson Valley regions, mostly from the first phase of the event on 9 January. There was also a narrow band of larger snow reports across upstate NY and the Mohawk River Valley, which came from the lake-effect event on 10 January. The environmental setup was conducive to freezing rain across eastern NY with warm air aloft and sub-freezing surface temperatures (Fig. 5.17) and the LSRs corroborated this, as many of the highest totals of freezing rain were located in central NY and Hudson Valley. Freezing rain totals across the Capital District area were generally above 0.25 in of flat ice, and Central NY flat ice totals were mostly between 0.1 and 0.25 in. The region of large snow totals from the lake-effect precipitation accumulated up to 30 in. Snow totals in general were between 5 and 20 in across much of the impacted area.

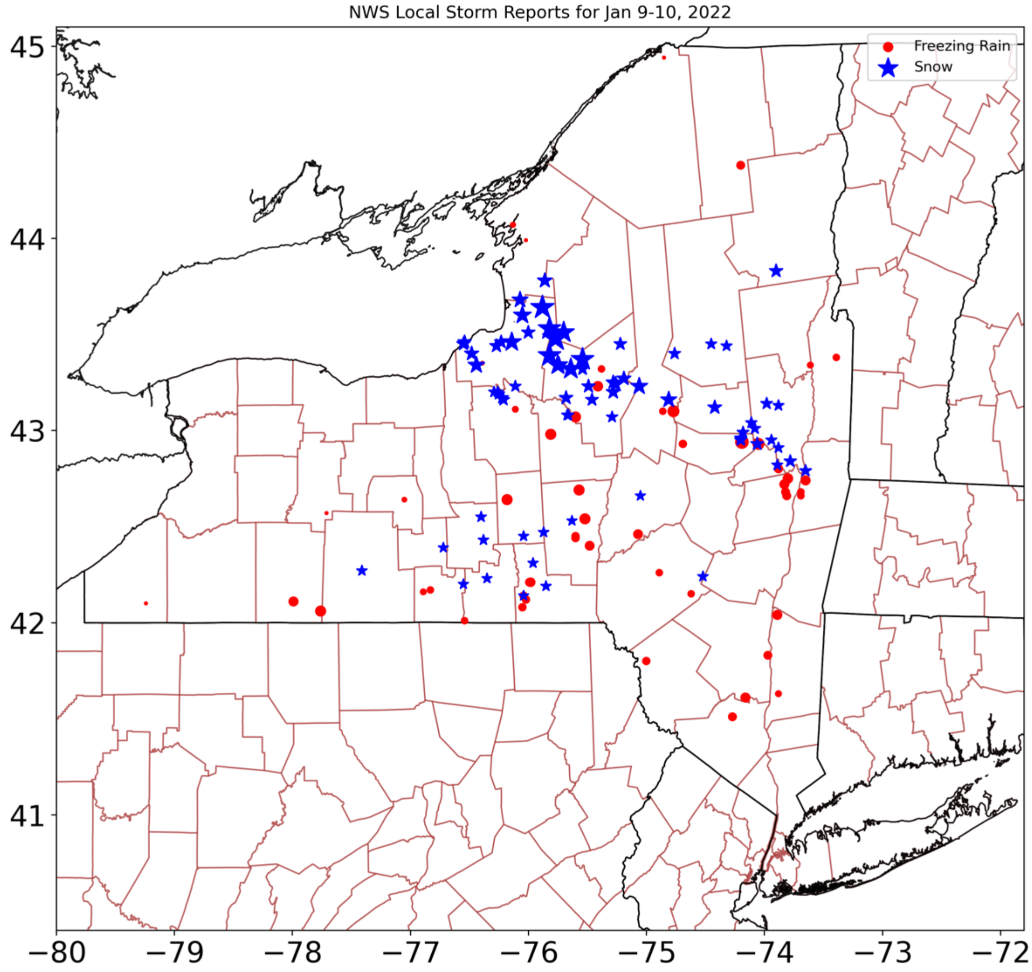


Fig. 5.16. Same as Fig. 5.10, but for 9–10 January 2022.

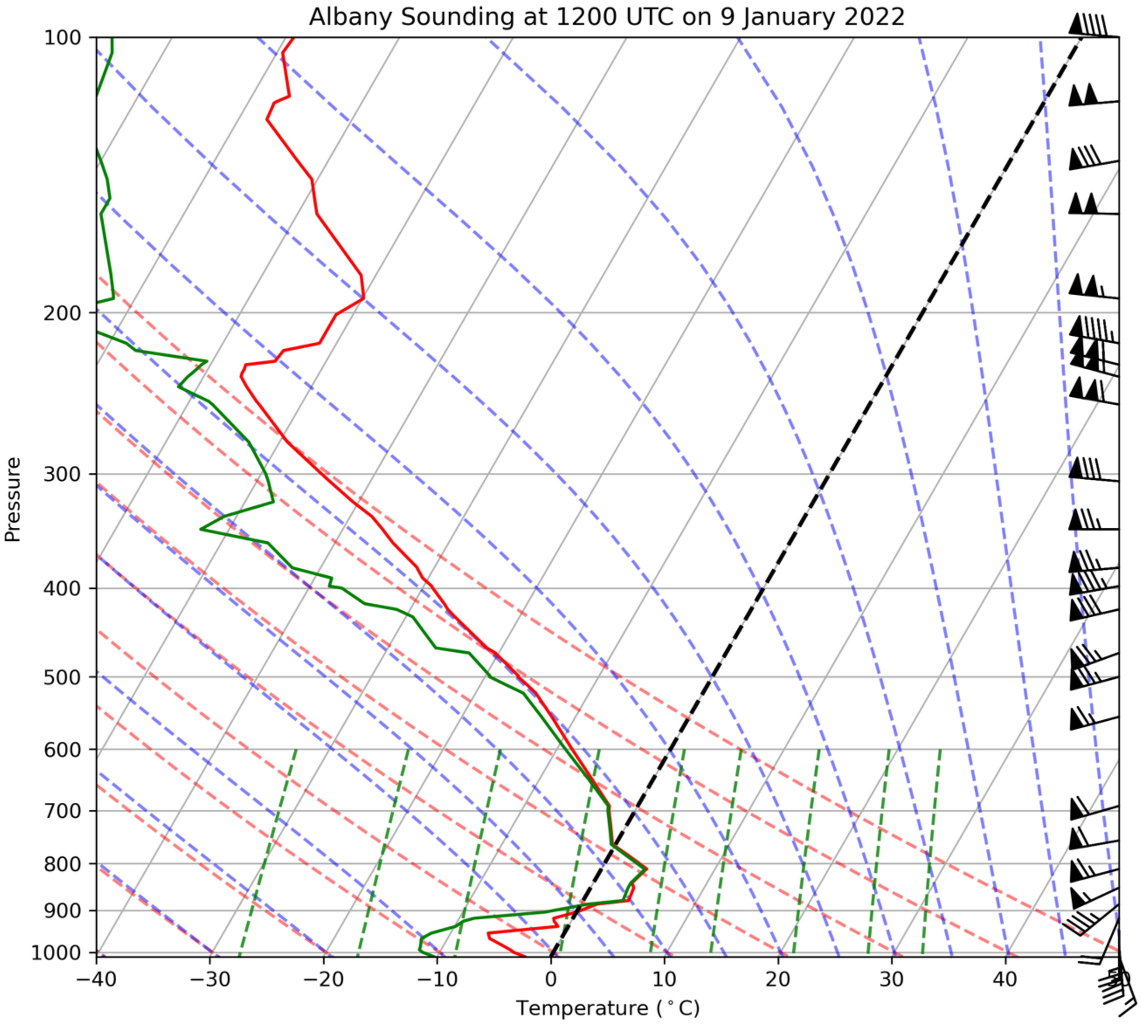


Fig. 5.17. Albany Sounding at 1200 UTC 9 January 2022. Solid red line is temperature. Solid green line is dewpoint. Dashed black line is 0°C Isotherm.

Looking at the how the RF did deterministically, the original NYSM and upper-air forecasts did better than expected with there being a relatively diverse set of predictions (Fig. 5.18a). It predicted the freezing rain and snow observations with over 50% accuracy, and rain observations were correctly predicted 32.46% of the time. These numbers were not great, but it much better than only sleet predictions, which was seen in the 2021 case study (Section 5.2.1). As for the NYSM and upper-air reduced forecast, it was near perfect for this event. Apart from the sleet observations, all other precipitation types were predicted with greater than 87%

accuracy, and rain and freezing rain were over 93% accurate (Fig. 5.18b). The NAMNEST was also very successful in accurately predicting precipitation types for the event. Again except for sleet, all other precipitation types were predicted with greater than 78% accuracy, with a higher number of observations to compare to (Fig. 5.18c). Rain was the lowest at 78.78%, freezing rain predictions were right 91.79% of the time, and snow predictions were correct 83.33% of the time. None of the RF models were able to capture any of the sleet observations; instead, they forecasted mostly, or exclusively, freezing rain. This forecast bust highlighted the difficulty of forecasting sleet as well as the RF models' potential bias for predicting freezing rain over sleet.

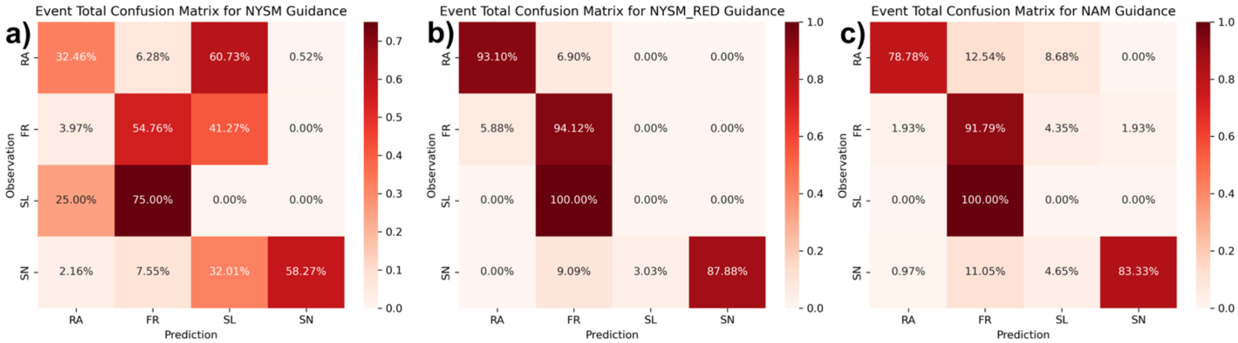


Fig. 5.18. Same as Fig. 5.1 except for only 9–10 January 2022.

The location of interest examined for this storm was the freezing rain in Amsterdam, NY, which had the largest freezing rain report of 0.4 in of flat ice. The probability timeseries for Amsterdam (Fig. 5.19a) showed a significant period of mixed precipitation being the highest probability. This occurred from 0500 UTC 9 January to 0400 UTC 10 January. Freezing rain was the dominant precipitation type most of this time (0900–2100 UTC 9 January). This period matched with the period of the main precipitation on 9 January and had the conditions to produce freezing rain including warm air advection, above freezing air aloft and below freezing air at the surface.

The other location examined for this event was in Osceola, NY, where 30 in of snow fell. The probability timeseries for Osceola (Fig. 5.19b) appeared to be representative of the precipitation occurring there. Most of the snow that fell in Osceola was from the lake-effect event that occurred on 10 January. The NAMNEST RF guidance had snow exclusively as the dominant precipitation type. This guidance corroborated 30 in of snow falling in Osceola, especially given its location and the long-duration lake-effect snow event that occurred. Overall, the available RF forecast guidance was successful at capturing all the precipitation types, except sleet, for the 9–10 January event. There was a bias towards freezing rain in all RF models as no sleet observations were accurately captured. In addition, the largest magnitude events for freezing rain and snow were well captured by the NAMNEST RF guidance.

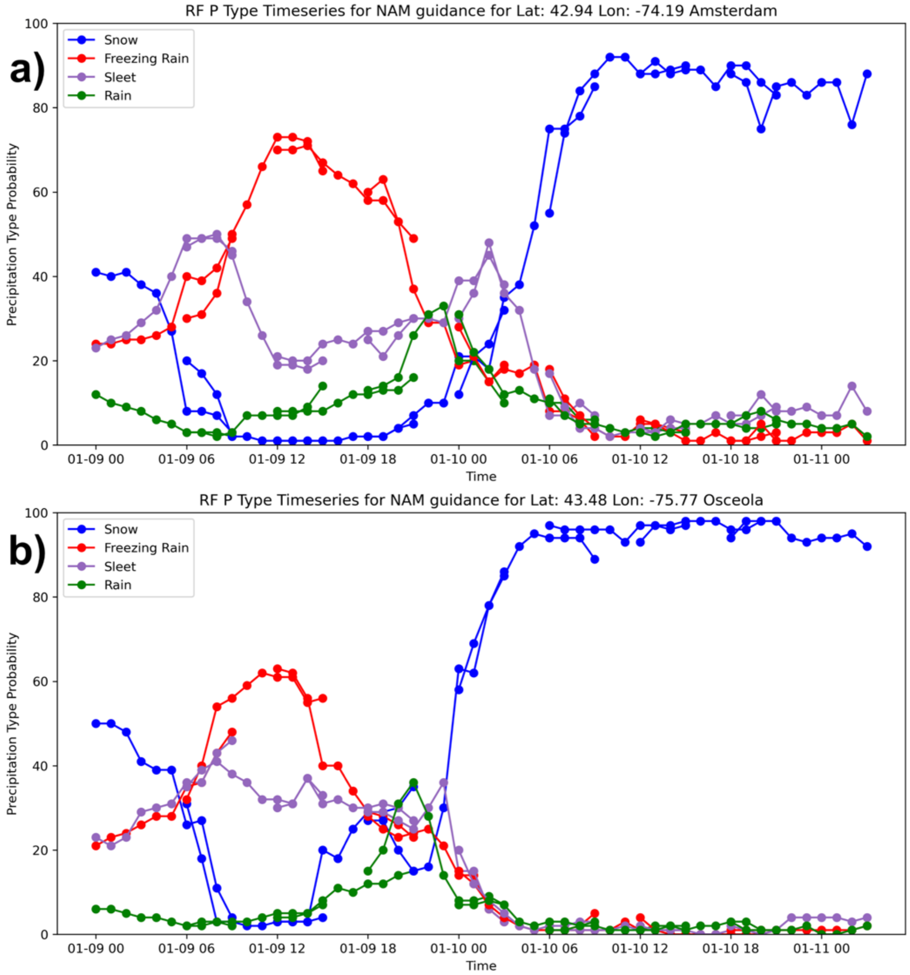


Fig. 5.19. Timeseries of RF probabilities at locations of interest during the 9–10 January storm. NWS LSRs with the highest reported magnitude of each precipitation type were selected as the locations of interest:(a) Amsterdam, NY and (b) Osceola, NY.

6. Future Work and Conclusions

6.1. Future Work

While the RF winter precipitation models and operational forecasts have great potential to aid operational precipitation type forecasting, future work can be done to both improve and complement what is currently available. Focusing on the RF algorithm, refinements and changes in the training datasets may provide increased accuracy. While CoCoRaHS reports were selected as the basis for the training datasets, it would be beneficial to test the algorithm with training datasets based on ASOS and mPING reports. The inclusion of these other datasets could expand the number of cases available for training data, including the ability to predict mixes of precipitation types and providing useful information in comparing the three types of winter precipitation reports. In addition to updating the basis of the training datasets, the winters from 2020–2021 onward could be incorporated into the training datasets thereby increasing the number of cases.

The RF could also be improved through additional internal testing, testing different configurations, and creating additional data combinations. By updating the training datasets, more internal testing would be needed to redetermine what datasets are the best options. Also, there is the ability to try techniques like principal component analysis (PCA) on the different datasets and data combinations to see if that improves the RF results by reducing the dimensions of the dataset. The configuration of the RF could also be re-examined to see if it can be improved. Updated training datasets would require different RF configurations. In addition, alternative ML techniques could be applied to forecasting winter precipitation types. Another area that may create improvement is in the data combinations. While multiple data combinations have been demonstrated, there are still a multitude of possibilities to increase the number of data sources. On the observational side, potential collaboration with the NYSM to use the profiler network would allow for the development of a state-wide observational product. On the model side, discussions with NWS Albany indicate that the Rapid Refresh (RAP) and National Blend of Models (NBM) may be other models to incorporate. For the data fusion component, combining model and observation data, like the HRRR and NYSM product, could be done for the other NWP models. In addition, there is the possibility to create an ensemble or weighted product of all the RF output. This product could create the potential for more accurate forecasts by giving ranges of probabilities based on the ensemble of model data. Along with adding more models, the vertical resolution and forecast lead time could continue to be increased through evaluation of the HRRR RF model. One more important area of future work is continuing to evaluate the RF models in different ways including comparing to numerical models' output against other methods used in operational forecasting.

On the operational forecast guidance side, improvements could be made to enhance the user experience and ease of use, which would likely increase the number of end users. By making the forecast maps interactive, users could zoom in or out, add or remove layers, and display multiple types of information on the same plot. These enhancements would eliminate the need to produce so many plots and make it easier for users to customize what they see. An additional way to improve the plots is to utilize the information from the probabilistic verification. Knowing what the numbers actually represent is important, and this could be implemented on each product by replacing the current probabilities with what that value has correlated to historically, or could include different levels of confidence contours, similar to the SPC Severe Weather Outlooks. Additional upgrades to the website and archive would make them more user friendly. The end goal is to make the website as easy to use as possible for end users. Some ways to achieve this could be to include the interactive graphics previously mentioned, improving the slideshow format to be more like a slider bar, and making it easy to get summaries of events through the archive.

6.2. Conclusions

Winter mixed precipitation events present numerous challenges to forecasters, ranging from areas of complex terrain to winter storms where precipitation types can transition across very short distances. ML algorithms can aid with these challenges by combining multiple large datasets to help account for local terrain variation or the widespread area covered by a large winter storm. A RF was trained to produce probabilistic predictions of precipitation type to help ease the data burden on forecasters who try to synthesize large datasets in real-time. To create this RF, CoCoRaHS reports were collected for four categories of precipitation: rain, freezing rain, sleet, and snow. These reports were then matched with observational (NYSM surface observations and upper-air radiosondes) and model (NAMNEST and HRRR) datasets to create combinations of training datasets. These datasets were then tested extensively for their accuracy in identifying winter mixed precipitation types, as well as for the best configuration of the RF and the variables in the datasets. Slight changes in the composition of these datasets created differences between the RF runs, reinforcing the idea that the variables and data combinations used to train a RF can impact the ability of the RF. Additionally, each data source and combination must be treated differently because combinations of variables may not be transferrable. This effect was found after seeing the same additional variables improve the NAMNEST RF runs while causing a decrease in the skill of the NYSM RF runs.

After finalizing the training datasets and understanding their strengths and weaknesses, a method was described that focused on transferring the research RF into an operational RF. Real-time data was processed and matched with grid locations to produce datasets compatible with the RF. Data challenges occurred during this transition process, such as dealing with missing datasets and incomplete real-time datasets due to hardware or data transmission issues (both for observational devices and data disseminated online). Product development methodology was also presented for context as why certain products were made or updated. Finally, examples of products available to end users were shown as they allow for end users to access information on all precipitation types.

Verification of the original NYSM and upper-air, NYSM and upper-air reduced, and NAMNEST forecast guidance was completed for the winters of 2020–2021 and 2021–2022. ASOS and mPING reports were used to verify both deterministically and probabilistically how well the RF forecast guidance did predicting precipitation types and what the probabilities displayed truly represent. It was shown that the original NYSM and upper-air product was limited by its ability to generally produce only sleet prediction. The NYSM and upper-air reduced product was an improvement on the original NYSM and upper-air product because it generated more realistic probability distributions and was more accurate overall. The NAMNEST RF model was successful at identifying different precipitation types and generating realistic probability distributions. To help forecasters and users to learn how long guidance can be trusted, the NAMNEST model was examined to see how well it did at predicting at different lead time. The NAMNEST guidance was fairly consistent in the accuracy of its predictions across five hours of lead time. The mixed precipitation types had the largest average drop off in accuracy when comparing the six different length forecasts. In addition to verifying how these products did over the two-winter period, two case studies showed the effectiveness of the different RF models over different types of winter precipitation events. Over both the two-winter period and individual case studies, there was success at predicting different precipitation types, particularly for rain, freezing rain, and snow. However, in both the NYSM and upper-air product reduced, and NAMNEST models, it appeared that there was a bias in mixed precipitation prediction where the model favored freezing rain over sleet. More extensive testing and comparison to other precipitation type algorithms will help to improve the RF models.

RFs, and other ML algorithms, can provide improvements in forecasting difficult weather events. While developing these algorithms is one part of the process, effectively discussing and learning what needs to be done to transition these algorithms into operations is an equally important part of the process. Evaluating these algorithms on an ongoing basis is important because understand their successes and limitations is necessary in order to effectively understand the RF guidance and communicate about the algorithms. For winter mixed precipitation events, working with end users to create meaningful products and training tools will help increase understanding and improve RF forecast guidance.

References

Baldwin, M., R. Treadon, and S. Contorno, 1994: Precipitation type prediction using a decision tree approach with NMC's meso-scale eta model. Preprints, 10th Conf. on Numerical Weather Prediction, Portland, OR, Amer. Meteor. Soc., 30–31.

Benjamin, S. G., J. M. Brown, and T. G. Smirnova, 2016a: Explicit precipitation-type diagnosis from a model using a mixed-phase bulk cloud-precipitation microphysics parameterizations. *Wea._Forecasting*, **31** ,609–619, doi:10.1175/WAF-D-15-0136.1.

Benjamin, S. G., and Coauthors, 2016b: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, https://doi.org/10.1175/MWR-D-15-0242.1.

Birk, K., E. Lenning, K. Donofrio, and M. T. Friedlein, 2021: A revised Bourgouin precipitation-type algorithm. *Wea. Forecasting*, **36**, 425–438, https://doi.org/10.1175/WAF-D-20-0118.1.

Breiman, L., 2001: Random Forests. *Mach. Learn.*, **45**, 5–32, https://doi.org/10.1023/A:1010933404324.

Brotzge, J. A, and Coauthors, 2020: A technical overview of the New York State Mesonet standard network. *J. Atmos. Oceanic Technol.,* **37**, 1827–1845, https://doi.org/10.1175/JTECH-D-19-0220.1.

Bourgouin, P., 2000: A method to determine precipitation types. *Wea._Forecasting*, **15**, 583–592, https://doi.org/10.1175/1520-0434(2000)015<0583:AMTDPT>2.0.CO;2.

Changnon, S. A., 2003: Characteristics of ice storms in the United States. *J. Appl. Meteor.*, **42**, 630–639, https://doi.org/10.1175/1520-0450(2003)042<0630:COISIT>2.0.CO;2.

Cifelli, R., N. Doesken, P. Kennedy, L. D. Carey, S. A. Rutledge, C. Gimmestad, and T. Depue, 2005: The Community Collaborative Rain, Hail, and Snow Network: Informal education for scientists and citizens. *Bull. Amer. Meteor. Soc.,* **86***,* 1069– 1078*,* https://doi.org/10.1175/BAMS-86-8-1069.

Ellis, A. W., S. J. Keighton, S. E. Zick, A. S. Shearer, C. E. Hockenbury, and A. Silverman. 2022: Analysis of model thermal profile forecasts associated with winter mixed precipitation within the United States mid-Atlantic region. *J. Operational Meteor.*, **10**, 1–17, https://doi.org/10.15191/nwajom.2022.1001.

Elmore, K., Z. Flamig, V. Lakshmanan, B. Kaney, V. Farmer, H. Reeves, and L. Rothfusz, 2014: mPING: Crowd-sourcing weather reports for research. *Bull. Amer. Meteor. Soc.*, **95**, 1335–1342, https://doi.org/10.1175/BAMS-D-13-00014.1.

Erickson, M. J., J. S. Kastman, B. Albright, S. Perfater, J. A. Nelson, R. S. Schumacher, and G. R. Herman, 2019: Verification Results from the 2017 HMT-WPC Flash Flood and Intense Rainfall Experiment. *J. Appl. Meteor. Climatol.,* **58**, 2591–2604, https://doi.org/10.1175/JAMC-D.19.0097.1.

Gagne, D. J., A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Wea._Forecasting*, **32**, 1819–1840. https://doi.org/10.1175/WAF-D-17-0010.1.

Herman, G. R., and R. S. Schumacher, 2016: Using reforecasts to improve forecasting of fog and visibility for aviation. *Wea._Forecasting*. **31**, 467–482, https://doi.org/10.1175/WAF-D-15-0108.1.

Herman, G. R., and R. S. Schumacher, 2018a: “Dendrology” in numerical weather prediction: What random forests and logistic regression tell us about forecasting extreme precipitation. *Mon. Wea. Rev.*, **146**, 1785–1812, https://doi.org/10.1175/MWR-D-17-0307.1.

Herman, G. R., and R. S. Schumacher, 2018b: Money doesn’t grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Mon. Wea. Rev.*, **146**, 1571–1600, https://doi.org/10.1175/MWR-D-17-0250.1.

Hill, A. J, G. R. Herman, and R. S. Schumacher, 2020: Forecasting severe weather with random forests. *Mon. Wea. Re.*, **148**, 2135–2161, https://doi.org/10.1175/MWR-D-19-0344.1.

Ikeda, K., M. Steiner, J. Pinto, and C. Alexander, 2013: Evaluation of cold-season precipitation forecasts generated by the hourly updating High-Resolution Rapid Refresh model. *Wea. Forecasting*, **28**, 921–939, https://doi.org/10.1175/WAF-D-12-00085.1.

Ikeda, K., M. Steiner, and G. Thompson, 2017: Examination of mixed-phase precipitation forecasts from the High-Resolution Rapid Refresh model using surface observations and sounding data. *Wea. Forecasting*, **32**, 949–967, https://doi.org/10.1175/WAF-D-16-0171.1.

Mahoney, E. A., and T. A. Niziol, 1997: BUFKIT: A software application tool kit for predicting lake-effect snow. Preprints, *13th Int. Conf. on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology*, Long Beach, CA, Amer. Meteor. Soc., 388–391.

Manikin, G. S., 2005: An overview of precipitation type forecasting using NAM and SREF data. *24th Conf. on Broadcast Meteorology/21st Conf. on Weather Analysis and Forecasting/17th Conf. on Numerical Weather Prediction*, Washington, DC, Amer. Meteor. Soc., 8A.6. [Available online at https://ams.confex.com/ams/pdfpapers/94838.pdf]

May, R**.** M**.**, Arms, S**.** C**.**, Leeman, J**.** R**.**, and Chastang, J**.**, 2017:_Siphon: A collection of Python Utilities for Accessing Remote Atmospheric and Oceanic Datasets**.** Unidata, Accessed 131 March 2021. [Available online at https://github.com/Unidata/siphon**.**] doi:10.5065/D6CN72NW**.**

McCray, C. D., E. H. Atallah, and J. R. Gyakum, 2019: Long-duration freezing rain events Over North America: Regional climatology and thermodynamic evolution. *Wea. Forecasting.* **34**, 665–681, https://doi.org/10.1175/WAF-D-18.0154.1.

McGovern, A., K. L. Elmore, D. J. Gagne, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith, and J. K. Williams, 2017: Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bull. Amer. Meteor. Soc.*, **98**, 2073–2090, https://doi.org/10.1175/BAMS-D-16-0123.1.

McGovern, A., R. Lagerquist, D. J. Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Amer. Meteor. Soc.*, **100**, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

NOAA National Centers for Environmental Information (NCEI), 2022: U.S. Billion-Dollar Weather and Climate Disasters. https://www.ncei.noaa.gov/access/billions/, DOI: 10.25921/stkw-7w73

NOAA National Weather Service (NWS) Public Information Statement 22-17, 2022a: Changes in Weather Balloon Launch Frequency Effective March 29, 2022 https://mesonet.agron.iastate.edu/wx/afos/p.php?pil=PNSWSH&e=202203291920.

NOAA National Weather Service (NWS) Albany Area Forecast Discussion, 2022b: https://mesonet.agron.iastate.edu/wx/afos/p.php?pil=AFDALY&e=202202040234.

NOAA, 1998: Automated Surface Observing System user's guide. NOAA/NWS, 58 pp. [Available from ASOS Program Office, NWS, 1325 East–West Hwy., Silver Spring, MD 20910.]

Ralph, F. M., and Coauthors, 2005: Improving short-term (0–48 h) cool-season quantitative precipitation forecasting: Recommendations from a USWRP workshop. *Bull. Amer. Meteor. Soc.*, **86**,1619–1632, https://doi.org/10.1175/BAMS-86-11-1619.

Ramer, J., 1993: An empirical technique for diagnosing precipitation type from model output. Preprints, Fifth Int. Conf. on Aviation Weather Systems, Vienna, VA, Amer. Meteor. Soc., 227–230.

Reeves, H. D., K. L. Elmore, A. Ryzhkov, T. Schuur, and J. Krause, 2014: Sources of uncertainty in precipitation-type forecasting. *Wea. Forecasting*, **29**, 936–953, https://doi.org/10.1175/WAF-D-14-00007.1.

Reeves, H. D., 2016: The Uncertainty of precipitation-type observations and its effect on the validation on forecast precipitation type. *Wea. Forecasting*, **31**, 1961–\1971, https://doi.org/10.1175/WAF-D-16-0068.1.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Schuur, T. J., H.-S. Park, A. V. Ryzhkov, and H. D. Reeves, 2012: Classification of precipitation types during transitional winter weather using the RUC model and polarimetric radar retrievals. *J. Appl. Meteor. Climatol.*, **51**, 763–779, https://doi.org/10.1175/JAMC-D-11-091.1.

Thériault, J. M., McFadden, V., Thompson, H. D., & Cholette, M, 2022: Meteorological factors responsible for major power outages during a severe freezing rain storm over Eastern Canada. *J. Appl. Meteor. Climatol.*, **61**, 1239-1255, https://doi.org/10.1175/JAMC-D-21-0217.1.

Wandishin, M. S., M. E. Baldwin, S. L. Mullen, and J. V. Cortinas Jr., 2005: Short-range ensemble forecasts of precipitation type. *Wea. Forecasting,* **20**, 609–626, https://doi.org/10.1175/WAF871.1.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346 USA