Probabilistic Forecasting Methods of Winter Mixed-Precipitation Events in New York State Utilizing a Random Forest

BRIAN C. FILIPIAK,[a] NICK P. BASSILL,[b] KRISTEN L. CORBOSIERO,[a] ANDREA L. LANG,[a] AND ROSS A. LAZEAR[a]

[a] *Department of Atmospheric and Environmental Science, University at Albany, State University of New York, Albany, New York*
[b] *Center of Excellence, University at Albany, State University of New York, Albany, New York*

(Manuscript received 3 November 2022, in final form 21 May 2023, accepted 20 June 2023)

ABSTRACT: Winter mixed-precipitation events are associated with multiple hazards and create forecast challenges that are due to the difficulty in determining the timing and amount of each precipitation type. In New York State, complex terrain enhances these forecast challenges. Machine learning is a relatively nascent tool that can help improve forecasting by synthesizing large amounts of data and finding underlying relationships. This study uses a random forest machine learning algorithm that generates probabilistic winter precipitation type forecasts. Random forest configuration, testing, and development methods are presented to show how this tool can be applied to operational forecasting. Dataset generation and variation are also explained because of their essential nature in the random forest. Last, the methodology of transitioning a machine learning algorithm from research to operations is discussed.

SIGNIFICANCE STATEMENT: Examining the role that machine learning can play in winter precipitation type forecasting is an area of research that has ample room for exploration, as much of the previous research has focused on applying machine learning to warm-season precipitation and severe weather events. Establishing a framework and methodology to successfully combine machine learning and weather research into effective operational tools is a valuable addition to the machine learning community. Because machine learning is increasingly being applied to meteorology, this work can act as a road map to help develop other meteorological tools based in machine learning.

KEYWORDS: Winter/cool season; Freezing precipitation; Mixed precipitation; Classification; Decision trees; Machine learning

1. Introduction

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 in this study as freezing rain or sleet, embedded within larger storms or on their own, can enhance difficulties in forecasting. Different precipitation types can cause a wide range of hazards while potentially occurring in adjacent meteorological environments or similar meteorological environments where the meteorological conditions differ on very fine scales. Differentiating between rain, freezing rain, sleet, and snow is essential to forecasting because of the unique hazards each one creates. In particular, freezing rain and heavy wet snow events can result in power failures and significant travel issues, which are not often associated with sleet or cold rain events.

In the United States between 1949 and 2000, catastrophic ice storm events (events with losses totaling over $1 million) generated $16.7 billion in losses; in particular, the Northeast 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–seven freezing rain days per year (Changnon 2003). Along with ice storms, the Northeast 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).

Significant damage and economic losses occur during mixed-precipitation type storms, and accurate forecasts of precipitation types are essential for decision-making and planning for organizations including city leaders, transportation departments, schools and universities, and many others. Accurate forecasts of precipitation type and timing can assist with decisions such as whether to pretreat roads and how to allocate snowplow and road salt operations. Because these weather hazards are destructive and costly and impact high-level decision-making (such as school closures), accurate mixed-precipitation forecasts are essential to protect lives and property.

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) 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. These changes in precipitation type have

Corresponding author: Brian Filipiak, bfilipiak@albany.edu

DOI: 10.1175/AIES-D-22-0080.1 e220080

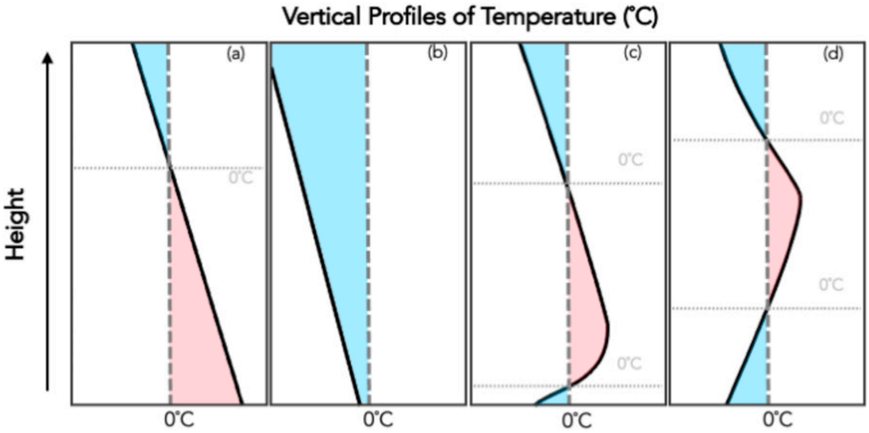

FIG. 1. Vertical temperature profiles for (a) rain, (b) snow, (c) freezing rain, and (d) sleet. Red and blue shaded areas respectively represent where the temperature is greater than and less than 0°C. The dotted horizontal gray lines represent intersection points between the profile and the (vertical dashed) 0°C line.

implications on the potential impacts of a storm, which highlights the importance of accurate vertical thermodynamic profiles.

Over the years, many methods have been developed to identify precipitation type within a forecast, both through implicit and postprocessing algorithms. Methods range from considering the properties of the temperature and humidity profiles, to the composition of the falling hydrometeors, to the use of model microphysical parameterizations to explicitly forecast precipitation type (Reeves 2016). Many of these methods are in operational use in numerical models, local forecast office guidance, and postprocessing of forecast information. For example, the North American Mesoscale Forecast System (NAM) uses a combination of methods (including some discussed in the next paragraph) based on the vertical environmental temperature profile to identify precipitation type (Manikin 2005). The High-Resolution Rapid Refresh (HRRR) uses a different approach where the determination of precipitation type comes from the cloud microphysics parameterization and three-dimensional hydrometeor mixing ratios (Benjamin et al. 2016).

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; Ellis et al. 2022). Numerous methods for identifying precipitation type are available in operations and models, including the commonly used Ramer (Ramer 1993), Baldwin (Baldwin et al. 1994), and Bourgouin (Bourgouin 2000) methods. Yet, even with all of the available forecasting 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). Even precipitation type forecasting methods for ensemble forecasting systems, like the Global Forecast System (GFS) and Integrated Forecast System (IFS), have struggled to forecast this differentiation (Scheuerer et al. 2017; Gascón et al. 2018). In addition, some of these methods have biases that impact their accuracy. For example, the Baldwin method has a known bias toward ice pellets (equivalently referenced as sleet in this study) and the Ramer method has a bias toward predicting freezing rain (Reeves et al. 2014). Probability of detection (POD) values for these methods can vary widely depending on the precipitation type and algorithm. The methods studied by Reeves et al. (2014) [e.g., Baldwin, Bourgouin, Ramer, and NSSL (Schuur et al. 2012)] showed a high level of accuracy when predicting rain or snow, with 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 (sleet in this study) and freezing rain ranged from 34.7% to 77% depending on the method. These values were even lower in other studies, such as Gascón et al. (2018) who focused on the IFS and found that the POD values were under 20% for both freezing rain and sleet. The range in success of these methods indicates the lack of an effective method for operational forecasters to rely on for predicting mixed precipitation.

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). The fact that there is no consensus on which method is the best presents an ongoing challenge as researchers and operational meteorologists look for effective methods for accurate precipitation type forecasting.

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 used 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 like 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), identification of large-scale weather features including atmospheric rivers and tropical cyclones (Chapman et al. 2019; Chen et al. 2020), prediction of visibility at airports (Herman and Schumacher 2016), and detection of atmospheric turbulence (McGovern et al. 2014; Williams 2014). Forecasting for these types of hazards can be operationally challenging and the use of ML algorithms by operational forecasters can aid in making an accurate forecast.

Many of the ML applications described above used a random forest (RF) technique (Breiman 2001) to successfully make strides in forecasting these operationally challenging events. These successful implementations of the RF indicate that it, along with other ML algorithms, can be used to help with the challenges in operational weather forecasting (McGovern et al. 2017, 2019). RF was selected as the ML method in the current 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. While ML is becoming a more common weather forecasting tool, one less studied application is highly impactful winter weather events. Considering the difficulties associated with forecasting winter precipitation types, the prospect of combining ML with the challenge of forecasting precipitation type represents an exciting and potentially informative forecasting tool that could be integrated into the operational forecasting process.

The goal of this paper is to detail the methodology of configuring a RF for producing effective, reliable, probabilistic products for operational winter precipitation type forecasting in New York State. Section 2 describes the data and RF methods. Section 3 explains the validation process of the RF through internal testing, which will be the basis for the operational RF forecast. Section 4 discusses challenges and hurdles faced during the transition from a research RF to an operational RF and suggests a flowchart to streamline the process. In addition, a framework for applying ML to operational forecasting that can be translated to other locations and weather types will be detailed.

2. Data and RF methods

a. *Data collection and processing*

To successfully create an RF, the basis for the training dataset first needs to be determined. 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 to use data from Automated Surface Observing Stations (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, however, which means they are not representative of complex terrain in a given region. This lack of representativeness 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 nonaugmented 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, so 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 accumulation (24-h period) across the country. All volunteers are trained in 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 is 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, for this specific effort, CoCoRaHS daily reports were chosen to identify cases for the training dataset because the observers were both trained and consistent, have a large spatial distribution, and collect reports of all precipitation types across a variety of terrain. The training dataset development began with daily CoCoRaHS reports dated between January 2017 and September 2020 for four precipitation types: rain, freezing rain, sleet, and snow. Once all these reports were obtained, the notes section of the reports was 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 daily reports do not have a specific precipitation type reporting section, the notes section of the reports was used to identify and categorize cases into the four target values (rain, freezing rain, sleet, or snow). Precipitation type, timing, and uncertainty were recorded for each individual report. While these are daily reports, the reported time for the precipitation observation was a singular time as described in the report's notes section. If a CoCoRaHS report contained multiple precipitation types, the less-common phenomena were selected first, assuming the atmospheric conditions corroborated the precipitation type. This meant that the order for classification of multiple precipitation types was sleet, freezing rain, snow, and rain. One CoCoRaHS report could generate multiple event labels if precipitation type changed during the reporting period; however, if it was unclear when the transition occurred or if there were multiple precipitation types occurring at the same time, those would not have been included in the event labels. To classify the

TABLE 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, whereas category-4 reports included no information about the time of precipitation.

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 2 pm and sleet began to mix in at 5 pm to all sleet by 6 pm to all rain by 7 pm."
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 in. 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 :("

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 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 category-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. Because of the significantly larger volume of CoCoRaHS reports for rain and snow relative to freezing rain and sleet, only those cases classified as category-1 reports were kept. For freezing rain and sleet, category-1–3 reports were used because of the limited volume of reports. After processing all of the CoCoRaHS reports, the final set of reports included 2617 viable cases: 750 for rain, 750 for snow, 619 for sleet, and 498 for freezing rain.

Once the cases were identified and verified, they were matched with meteorological data that are often used to forecast precipitation type. These meteorological variables would become the feature inputs into the RF and be used to make the predictions of the target values. This process is described more in section 2b. The main data sources used to match with the final CoCoRaHS reports were NYSM standard site data for surface observations, in situ radiosonde data, and Buffalo Toolkit (BUFKIT; Mahoney and Niziol 1997) model profiles for vertical profiles. Surface observations in this study came 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 subnetworks 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. Variables from each NYSM dataset used as features in the RF.

NYSM 5-min variables	NYSM hourly variables
2-m temperature	2-m temperature (min, max, and avg)
2-m relative humidity	Relative humidity (min, max, and avg)
Surface pressure	Station pressure (min, max, and avg)
Solar irradiance	Solar irradiance and total solar irradiance
Precipitation (5-min total, daily total, and intensity)	Precipitation (hourly total, daily total, and intensity)
10-m-avg wind speed and direction from sonic anemometer	10-m-avg wind speed and direction from sonic anemometer

TABLE 3. Variables used as features in the RF from the NAMNEST and radiosonde vertical profile datasets. Raw variables (C1) are at standard pressure levels including the surface and 925, 850, 700, and 500 hPa. All calculated variables (C2 and C3) were found between standard pressure levels unless otherwise noted.

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

Table 2 details the variables from the NYSM that are used as features in the training dataset. The features that were extracted from the NYSM data were matched to the nearest report time period (5-min or hourly for each CoCoRaHS report).

Complete profiles of the lower and middle troposphere are crucial for making precipitation type predictions, so in situ radiosonde data were used to complement the NYSM standard site data. Radiosonde data were collected from four sites in or near New York State at 0000 and 1200 UTC daily (Albany, Buffalo, and Upton in New York, and Maniwaki, Quebec). Each CoCoRaHS report was matched to the nearest and most recent radiosonde launch (i.e., if a CoCoRaHS observer reported snow at 0600 UTC, the snow report would be matched to the data from 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 6 h 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 interpolated as a radiosonde observation at specific locations. The BUFKIT generated NAMNEST profiles are available through the Iowa State Mesonet archive (Mtarchive) and in real time from the Pennsylvania State University (the data from these profiles will be referred to as NAMNEST profile data or just NAMNEST). 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.

All three datasets, NYSM, upper-air radiosondes, and the NAMNEST, used in the RF combined raw observed variables with calculated variables based on the raw data available from radiosonde data or model output. NYSM data were 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 [Table 3, columns 2 and 3 (C2 and C3)], including wet-bulb temperature, precipitable water vapor, and calculations of raw variables between standard pressure levels. Since these additional variables were not in the original profile datasets, they had to be calculated after the original profile was processed. Many of the advanced calculations (wet-bulb temperature, precipitable water vapor, etc.) were completed via the MetPy python package (version 1.4; May et al. 2022). The original calculated variables (Table 3, C2) were calculated before the new calculated variables (Table 3, C3), and they were both tested separately and together as features that were included in attempts to train the RF. Since the radiosondes and NAMNEST profiles have the same structure, one table can be used to summarize the features that were used from the profiles. Table 3 details all of the features used in the vertical profile datasets (radiosondes and NAMNEST profiles).

b. RF methods

RFs are a type of supervised ML consisting of an ensemble of individual decision trees trained on an example set of data. Here, the RF is being used as a classification task where it is given a separate, testing dataset and each tree votes for the most popular class based on the input features in the training dataset 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 is created by separating the training dataset by making decisions at nodes (points at which a value from a feature in the testing dataset is compared with the RF-determined threshold value of that feature in training dataset). Once split, the separated training dataset feeds into different branches that go to other nodes. This process at each node occurs continuously until the training dataset has been totally separated into the individual target values (the types of precipitation) or there are too few training dataset examples left to split (Hill et al. 2020). This process can be controlled via a

TABLE 4. RF parameter default configuration from the python scikit-learn package and the configuration determined from the hyperparameter tuning process.

Parameter	Default value	Value after tuning
No. of decision trees (N)	100	650
Min no. of samples to split at a node (min_samples_split)	2	10
Min no. of samples to be at a leaf node (min_samples_leaf)	1	1
No. of features to consider for best split (max_features)	Sqrt	Log2
Max depth of a decision tree (max_depth)	None	25
Bootstrap samples (bootstrap)	True	True

tuning parameter of the RF and creates end points (also called final nodes) where predictions of the target values can be made. Once the RF is trained and the decision trees have been created, real-time input features can be used to make predictions by following the decisions made at each node of the decision tree. At this point, a vote is made for the most popular class by the tree. While each tree may have many final nodes, there is only one prediction made from each tree for each set of features being processed. In the case of the different winter precipitation types, the target values for the RF are the four classes of precipitation in the testing data.

Once all the observed and simulated data were collected, processed, and matched with the CoCoRaHS reports, the reports and associated feature variables were combined to create different, unique data combinations with two distinct types of datasets: NYSM and upper-air data and NAMNEST profile data. This process was done by matching the timing of the reported precipitation type to the nearest available meteorological observations or model data at the specific time of the event (e.g., a report of freezing rain at 1145 UTC would be matched with 5-min NYSM data at 1145 UTC, hourly NYSM data covering 1100–1200 UTC, upper-air radiosonde data from 0000 UTC, and NAMNEST data from the 0600 UTC initialization at forecast hour 5). The input features from the different datasets are the unique combinations of features described in section 2a because the different data sources (NYSM, radiosondes, and NAMNEST) have different variables within them to select as features. These combinations were tested in the RF to determine which would be the final set of features used as the training dataset for the operational RF. The goal of testing the different data combinations was to address the important question: “What data sources and variables actually contribute to making a good precipitation type forecast?” By varying the data combinations and features in the different training datasets, the answer to this question will be apparent because the contributions of different data sources via their features will be evident in the RF’s ability to predict the target precipitation type values.

To test how well the RF would work, a training dataset and testing dataset needed to be identified. To do this, the full dataset (the full set of CoCoRaHS reports matched with features from different datasets) was randomly split 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 for each target value (precipitation type) were not equal, the training and testing datasets were split such that the proportion of each of the target values (rain, freezing rain, sleet, and snow) was equal in both datasets. The testing dataset was saved to evaluate the RF, which will be discussed in section 3.

The RF was configured using a thorough hyperparameter tuning process (the process of determining the optimal combination of parameters that control how the RF is structured through grid searches of all possible hyperparameter combinations), with cross validation to thoroughly test the setup. The process of cross validation is important in order to make sure the RF is not overfitting, especially with a smaller number of total (training and testing) cases. In addition, cross validation can help with determining RF skill with new data. Initial experiments used for evaluation of the RF (not shown) used 500 decisions trees and kept the rest of the default RF options (Table 4) from the python scikit-learn package (version 1.1.2; Pedregosa et al. 2011). This testing was done for multiple combinations of feature variables from different data sources. For example, one experiment only included features from the radiosondes, whereas the next one may include the same features from the radiosondes plus features from the 5-min NYSM data. This was done to better understand how the different combinations of data sources and features would work together, as well as which may perform the best. The hyperparameter tuning process began once the highest performing datasets were selected from the initial testing process (see Table 5 for list of datasets), which happened through examining which had the best internal statistics (accuracy and F1 scores). The training dataset, described above as 75% of the original CoCoRaHS reports matched with input features from different data sources, was used in multiple tests to determine the best set of hyperparameters. Since several different feature combinations would be tested later, the hyperparameter tuning was completed on two training datasets: one contained all NYSM features from both hourly and 5-min data sources with NWS upper-air radiosondes, while the other contained the NAMNEST features in the training dataset. This was done to cover both types of training datasets to be examined later, while keeping the hyperparameters the same. A random grid search with tenfold cross validation was conducted with 150 iterations, which was done across a wide range of values for all the parameters, with each iteration trying a different combination of parameters from the possible options. The process was done five times to get multiple grid outputs in order to narrow down the range of options for each parameter. Since there was a range of results, a full grid search with tenfold cross validation was completed over the narrower range of options from the random search. The result from

TABLE 5. Description of RF training datasets and name abbreviations. Descriptions indicate dataset (NYSM and NAMNEST); sounding location (original, updated, and NAMNEST); and type of sounding variables included as features [raw (C1) and calculated: original (C2) and new (C3)]. Sounding features are described in Table 3, and NYSM features are described in Table 2.

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 3: C1 and C2)	HAVG_RCO
NYSM hourly averaged surface variables with raw from original soundings (Table 3: C1)	HAVG_RO
NYSM hourly averaged surface variables with original calculated variables from original soundings (Table 3: C2)	HAVG_CO
NYSM 5-min surface obs with raw and original calculated variables from original soundings (Table 3: C1 and C2)	OBS5_RCO
NYSM 5-min surface obs with raw variables from original soundings (Table 3: C1)	OBS5_RO
NYSM 5-min surface obs with original calculated variables from original soundings (Table 3: C2)	OBS5_CO
All NYSM surface data with raw and original calculated variables from original soundings (Table 3: C1 and C2)	ALL_RCO
All NYSM surface data with raw variables from original soundings (Table 3: C1)	ALL_RO
All NYSM surface data with original calculated variables from original soundings (Table 3: C2)	ALL_CO
NAMNEST soundings with raw and original calculated variables (Table 3: C1 and C2)	NAM_RCO
NAMNEST soundings with raw variables (Table 3: C1)	NAM_RO
NAMNEST soundings with original calculated variables (Table 3: C2)	NAM_CO
NAMNEST soundings with raw and all calculated variables (Table 3: C1, C2, and C3)	NAM_RCN
NAMNEST soundings with raw variables (Table 3: C1)	NAM_RN
NAMNEST soundings with all calculated variables (Table 3: C2, C3)	NAM_CN
All NYSM surface data with raw and all calculated variables from updated soundings (Table 3: C1, C2, and C3)	ALL_RCN
All NYSM surface data with raw variables from updated soundings (Table 3: C1)	ALL_RN
All NYSM surface data with all calculated variables from updated soundings (Table 3: C2 and C3)	ALL_CN
All NYSM surface data with raw and all calculated variables from NAMNEST soundings (Table 3: C1, C2, and C3)	ALL_NAM_RCN
All NYSM surface data with raw variables from NAMNEST soundings (Table 3: C1)	ALL_NAM_RN
All NYSM surface data with all calculated variables from NAMNEST soundings (Table 3: C2 and C3)	ALL_NAM_CN

that search is the parameter configuration that was used in the full RF, and thus concluded the hyperparameter tuning process (Table 4).

After conducting the random grid and full grid searches, a full internal testing of the RF was performed using the testing dataset, the remaining 25% of original training dataset set aside for this purpose, to evaluate the RF performance with winter precipitation type classification of past events. This evaluation of the RF was done by using four key metrics: accuracy, precision, recall, and F1 score (Fig. 2). Accuracy indicates the overall number of correct predictions out of the total predictions of the RF; precision is the number of correct predictions divided by the number of total predictions for that precipitation type; recall is how often the correct prediction occurs in the RF; and F1 score is 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 the target values were predicted. These metrics were calculated for each run of the RF, defined as where the RF is trained with one training dataset with one set of features and makes predictions on a testing dataset that contains the same set of features. One independent RF run is defined as when the RF is trained and validated against the testing dataset; this will be referred to as a run throughout the rest of this study. Multiple independent runs are done by running the RF multiple times in a row with the same training and testing datasets; the results from each run, the probabilities and feature importance, can then be averaged. The numbers described later were averaged over 50 independent RF runs to remove any potential outlier runs.

Confusion matrices were used to evaluate the outcomes from the RF predictions (Figs. 2 and 4). 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 makes incorrect predictions, thereby allowing for corrections and necessary changes to the RF.

Section 3 will highlight which features were most important in the decision-making of the RF. To determine feature importance, the method of Gini 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 the feature is. Impurity importance was selected since it can

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

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

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

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

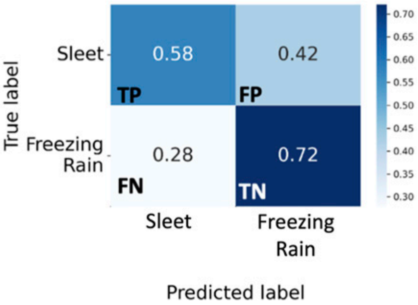

FIG. 2. (left) RF evaluation metrics and (right) a sample confusion matrix for sleet and freezing rain with the prediction of sleet set as the true positive. Evaluation metrics can be calculated from the case distribution in the confusion matrix, which shows the proportions of correct and incorrect predictions.

be done during the RF training process (McGovern et al. 2019); however, there are some limitations of this method including favoring features that appear early in the decision tree and not properly expressing the correct importance of two correlated features (either by ignoring one or splitting the importance between the two). While there are other methods of determining feature importance, like permutation importance, impurity importance was used in this study despite its limitations because there were little to no differences in the most important features between the impurity and permutation importance methods (not shown).

3. Results: Internal algorithm testing

Internal testing of the RF is an integral step to evaluate the impact of the different combinations of input features available to train the RF, which can be done by making predictions based on the testing dataset, 25% of original training dataset that was not included when training the RF. Starting with the NYSM and upper-air data, all combinations of features were considered to identify which would be most effective, including testing different combinations of features 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 3; Fig. 3). In evaluating training datasets, it was important to not only consider overall accuracy (denoted by the black circles), but also the individual F1 scores for different precipitation types because high accuracies in certain categories can mask other low accuracies. Figure 3 shows that the overall accuracy and F1 scores changed when considering different combinations of input features, which points to the importance of the input dataset in the RF algorithm. The most accurate combination of features was the NYSM 5-min and hourly features combined with observed sounding features (ALL_RO; Table 5 and Fig. 3). This combination of features 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 higher number of the training cases correctly), these values are, nonetheless, promising. It is important to note that it is unlikely the RF would achieve F1 scores close to 100% due to the inherent uncertainty in the observed precipitation type reports. The F1 scores are 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 represent different metrics for evaluating the datasets.

Figure 4 shows the confusion matrix from a run of the RF with the ALL_RO training and testing data (Table 5), one of the highest performing RF runs using features from the NYSM and upper-air datasets. The F1 scores from Fig. 3 are very similar to the fraction of correct predictions for sleet (0.55 correct) and freezing rain (0.65 correct) in Fig. 4. An interesting result of these RF runs is that the algorithm's values of correct predictions for snow (0.92) and rain (0.87) events are 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 suggests the RF can more easily recognize three major types of precipitation: rain, mixed precipitation, and snow.

A similar method of evaluation can be followed for the NAMNEST-based training datasets. The model-derived precipitation type prediction from the NAMNEST is not used as a feature, so the RF creates a target value of precipitation type from meteorological features only. Two combinations of NAMNEST data were generated. The original datasets (ending in O in Table 5) were the first attempt to combine model features to predict the target values. The new datasets (ending in N in Table 5) were an attempt to improve upon the original datasets, which only contained the features in Table 3, column 1 (C1) and C2, by adding new calculated features, found in Table 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. 5). Additionally, the F1 scores for all four target values increased for RF runs utilizing the new datasets. Several of the new calculated features appeared in the top 10 features from runs of the NAM_RCN dataset (Fig. 6): positive area (defined as the integrated area where environmental temperature is greater than 0°C in a vertical temperature profile; Bourgouin 2000), maximum wet-bulb temperature between 925 and 700 hPa, minimum temperature between the

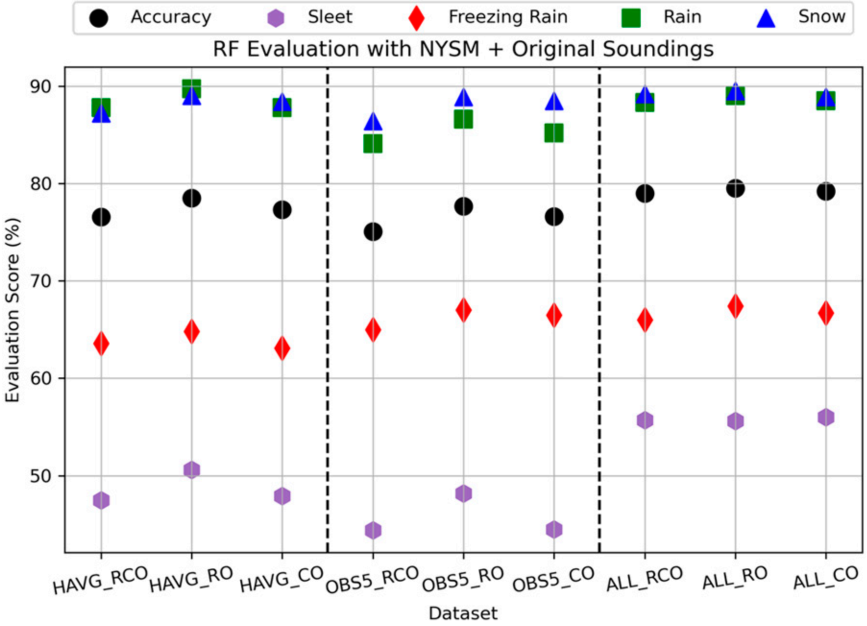

FIG. 3. Accuracy and F1 scores for different combinations of NYSM and upper-air input features. The training dataset abbreviations are in Table 5. Dashed lines represent divisions between different types of datasets. The left third represents the dataset built with features from hourly NYSM and upper-air data. The center third represents the dataset built with features from 5-min NYSM and upper-air data. The right third represents the dataset built with features from hourly and 5-min NYSM along with upper-air data. The black circles represent the overall accuracy of the RF. The colored shapes correspond to the F1 scores of the different target values (purple hexagons for sleet, red diamonds for freezing rain, green squares for rain, and blue triangles for snow).

surface and 850 hPa, average temperature between the surface and 850 hPa, and maximum temperature between 850 and 700 hPa. This result indicates that the new calculated features (Table 3, C3) likely made the difference in the higher scores.

From a meteorological perspective, these new calculated features 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 features were selected specifically because they have been identified in the literature (Ramer 1993; Baldwin et al. 1994; Bourgouin 2000; Manikin 2005; Benjamin et al. 2016) as important in determining mixed-precipitation type, and their usefulness is seen by an increase in F1 scores for sleet and freezing rain (Fig. 5). While improvement was made by incorporating features 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. This point is important because if they were more consistent, additional features could be extracted at important pressure levels in the lower levels of the atmosphere, which could provide more features to improve the RF.

Since the new calculated features (Table 3, C3) were successful in improving the NAMNEST RF, those same features were added into original NYSM datasets to see if the same improvement would occur. Figure 7 compares the best three original NYSM runs (Fig. 7, to the left of the first vertical dashed line) with NYSM runs with the new calculated features added into the datasets (Fig. 7, between the vertical

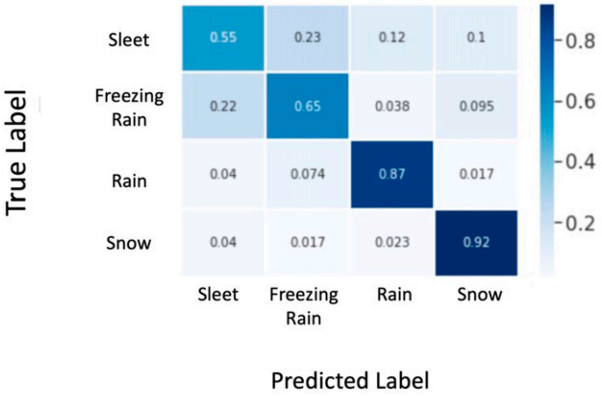

FIG. 4. Confusion matrix of values averaged over 50 independent RF runs for the ALL_RO NYSM and upper-air training dataset (Table 5).

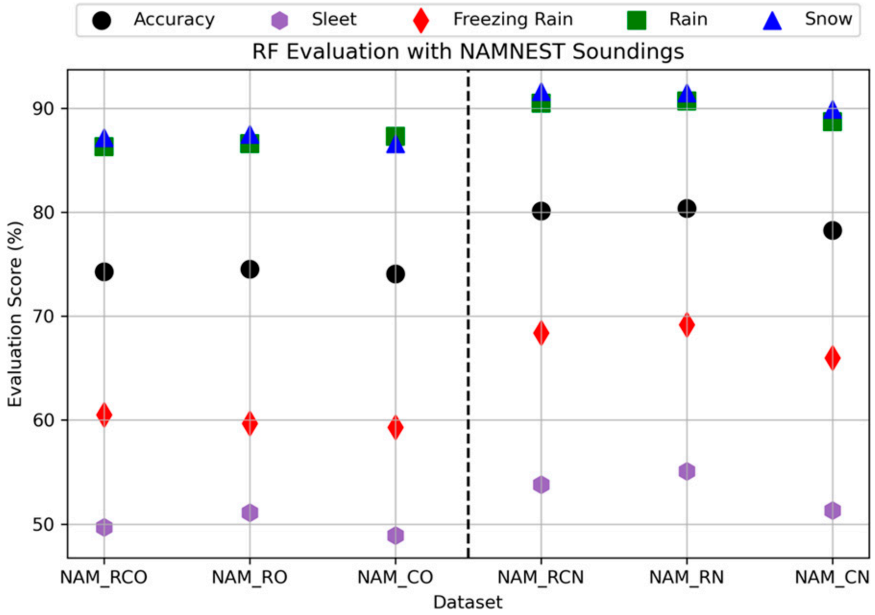

FIG. 5. Accuracy and F1 scores for different combinations of NAMNEST features. The dataset abbreviations are in Table 5. Dashed lines represent divisions between different types of datasets. The left half represents the dataset built from NAMNEST data with the raw and original calculated features (Table 3; C1 and C2). The right half represents the dataset built from NAMNEST data with the raw and original calculated features plus the new calculated features (Table 3; C1, C2, and C3).

dashed lines). In addition, combining the NYSM 5-min and hourly features with the NAMNEST profiles features (to the right of the second dashed vertical line) was tested to determine whether increasing the spatial density of vertical profiles would improve the RF. The new calculated features 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 is likely associated with the new calculated features adding noise to the decision-making process of the RF, which is examined in the next paragraph.

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 of the atmosphere should create more representative input feature values. However, this was not the case with the NYSM and NAMNEST datasets. The decrease connected to these results is likely because there is an overlap in the features in the combined datasets creating noise in the RF. This was seen in Wang et al. (2022) where, if highly correlated features were run through an ML algorithm, those features had little impact on the final predictions. Additionally, the surface feature values can conflict or not be meteorologically consistent with the NAMNEST feature values due to model initialization and data assimilation methods and/or the geographical locations of the datapoints from which the values for the feature were extracted. For example, pairing surface observing sites with a location at significantly higher elevation than a NAMNEST profile site can cause a discrepancy. A similar issue was pointed out in Mital et al. (2020) and it is important to note that it may be better to match stations at similar elevations. While these results show a decrease in performance, it prompts a response to conduct future experimentation with different combinations of features from the NYSM and NAMNEST profile datasets to find the best possible combination. This point is a key takeaway from the process of testing the RF and determining the best possible dataset because each type and combination of features needs to be treated differently. Throughout the extensive testing process, the best training datasets for each combination of features were ALL_RO for the NYSM and Upper-air and NAM_RN for the NAMNEST because they had the best overall performance for both accuracy as well as individual F1 scores. These combinations of features form the final training datasets that were used in the operational forecasts.

4. Research-to-operations methodology

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 combinations of features that already exist. The challenge occurs when trying to apply the algorithm as a real-time tool to make forecasts and nowcasts. Transitioning this algorithm from research to operations, defined here as a forecast product that is made with real-time data and produced at the University at Albany, was a multistep process with potential issues associated with real-time processing of incoming data from various sources, formatting the real-time dataset, and runtime issues. Some previous work (Taillardat and Mestre 2020;

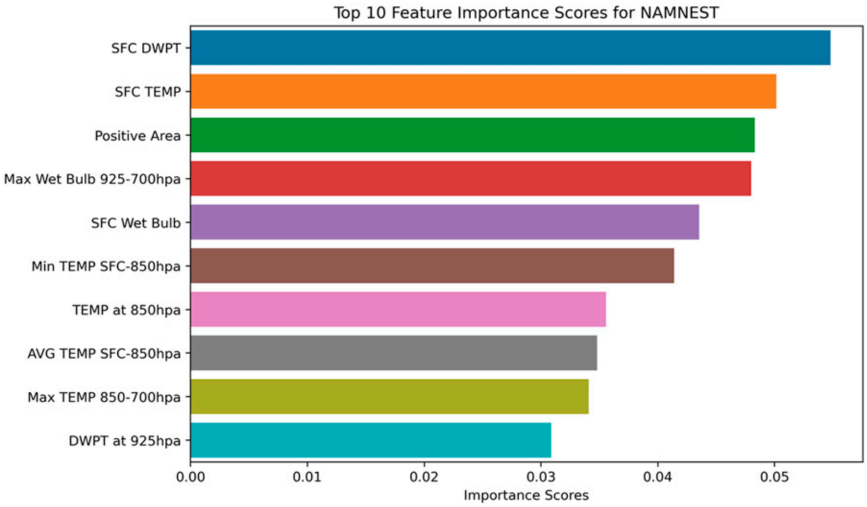

FIG. 6. The top 10 most important features from RF runs of the new NAM_RCN dataset. The values of importance have been averaged over 50 runs. The higher the value is, the more important are the features. The full names of the features in order of importance are as follows: surface dewpoint, surface temperature, positive area, maximum wet-bulb temperature from 925 to 700 hPa, surface wet-bulb temperature, minimum temperature from the surface to 850 hPa, temperature at 850 hPa, average temperature from the surface to 850 hPa, maximum temperature from 850 to 700 hPa, and dewpoint at 925 hPa.

Haupt et al. 2021; Vannitsem et al. 2021) has detailed parts of the research to operation transition process for postprocessing weather forecasts, but there is more work to be done to make sure this process is clear, and the associated challenges known. This section describes the transition of this research-oriented ML algorithm to an operational forecasting setting.

The path for transitioning this RF from research to operations is represented in the flowchart in Fig. 8. Sections 2 and 3

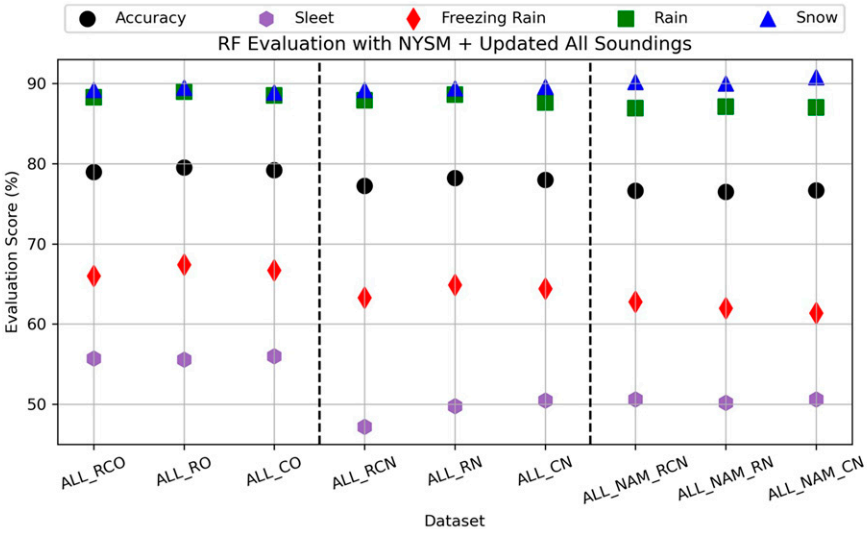

FIG. 7. Accuracy and F1 scores for different combinations of NYSM, NAMNEST, and upper-air features. The dataset abbreviations are in Table 5. Dashed lines represent divisions between different types of datasets. The left third represents the dataset built with features from hourly and 5-min NYSM and upper-air data with raw and original calculated variables (Table 3; C1 and C2). The center third represents the dataset built with features from hourly and 5-min NYSM and upper-air data with raw and original calculated variables plus the new calculated variables (Table 3; C1, C2, and C3). The right third represents the dataset built with features from hourly and 5-min NYSM and NAMNEST data raw and original calculated variables plus the new calculated variables (Table 3; C1, C2, and C3).

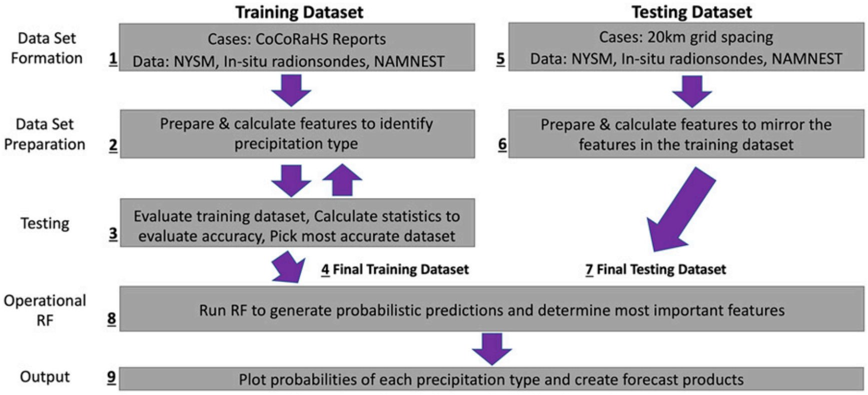
Training Dataset
Testing Dataset
Data Set Formation
1 Cases: CoCoRaHS Reports Data: NYSM, In-situ radionsondes, NAMNEST
5 Cases: 20km grid spacing Data: NYSM, In-situ radionsondes, NAMNEST
Data Set Preparation
2 Prepare & calculate features to identify precipitation type
6 Prepare & calculate features to mirror the features in the training dataset
Testing
3 Evaluate training dataset, Calculate statistics to evaluate accuracy, Pick most accurate dataset
4 Final Training Dataset
7 Final Testing Dataset
Operational RF
8 Run RF to generate probabilistic predictions and determine most important features
Output
9 Plot probabilities of each precipitation type and create forecast products

FIG. 8. A flowchart of the methodological steps to create an operational RF for predicting winter mixed-precipitation types. This flowchart can be generalized for other ML algorithms and meteorological events.

detailed the process of preparing and testing the training dataset until the highest performing training datasets were determined (Fig. 8, steps 1–4). The next step in the transition to an operational RF is creating the input dataset with features from real-time data. To create the locations where predictions will be made, a 20-km grid of New York State was generated to create synthetic locations for which predictions of target values can be generated. These points were matched with features from NYSM, upper-air, and NAMNEST profile locations using the same process as the CoCoRaHS reports (Fig. 8, step 5; Fig. 9). The real-time datasets were accessed via multiple tools. The NYSM data are accessed via a local shared disk with the Department of Atmospheric and Environmental Sciences at the University at Albany. The radiosonde data are accessed via Siphon data request package (https://doi.org/10.5065/D6CN72NW; May et al. 2017). The NAMNEST profiles are accessed via a BUFKIT processing package made by Carter Humphreys and is available on Github (bufkit-api; https://github.com/HumphreysCarter/bufkit-api) with the actual NAMNEST profile data coming from The Pennsylvania State University.

Each time the RF is given a set of real-time data features to make a prediction, the incoming features are compiled and cleaned to make sure there are no missing features. The processing of the incoming features from real-time datasets is an essential step because the RF will not be able to process the features if the real-time data features do not match the training dataset features or there are missing feature values (Fig. 8, step 6). Currently, if an RF prediction location has a missing feature or a whole real-time dataset, no prediction is made at that location. This is a simple solution to this issue; however, there are ways to fill in missing values via imputation using data from previous times at the same location, interpolating from neighboring locations, or a combination of these methods with or without using ML techniques (Mital et al. 2020; Dwivedi et al. 2022). If reliable data sources are not available, it is difficult to produce consistent forecast guidance for end users. For example, if radiosondes are not launched or if other data sources are 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 2022). Once the real-time features in the testing dataset match the training dataset (Fig. 8, step 7), the RF can be run. For a forecast 5 h in the future, this would be possible for only the NAMNEST products since the other products rely on observation data. For NAMNEST, the latest model run would be used and the data from forecast hour 5 would be selected as feature inputs for the RF. For the observational data sources used in the nowcast products, the latest data available would be used. For example, a nowcast product using NYSM and upper-air features at 0500 UTC would use features from NYSM data from 0500 UTC and the NWS sounding at 0000 UTC to be used in making the nowcast from the RF. The outputs of the RF are probabilistic predictions of each target value (precipitation type) at each location in the testing dataset. This information can be processed to create maps, graphics, or tables to be displayed in an operational setting (Fig. 8, steps 8–9).

The probabilities from the RF runs are displayed on an operational website (http://www.atmos.albany.edu/student/filipiak/op/). The website displayed multiple RF products made in real time throughout the 2021/22 winter season. The latency on the NYSM and upper-air data products is under 10 min, such that the NYSM and upper-air product can be made at 30 min past each hour. The latency for the NAMNEST product updated with each new model run is about one hour for a 10-h forecast period including forecast hour 0. The products are hourly forecasts and display the probabilities of the different types of precipitation if precipitation were to occur. Even when there is no precipitation occurring, the RF products are being made and can be used to understand the current atmospheric conditions. The products include the hourly probabilities of the four main precipitation types (rain in Fig. 10a; freezing rain in Fig. 10b; sleet in Fig. 10c; snow in Fig. 10d), an all mixed precipitation (addition of sleet and freezing rain probabilities; Fig. 10e), and a dominant precipitation type (shows

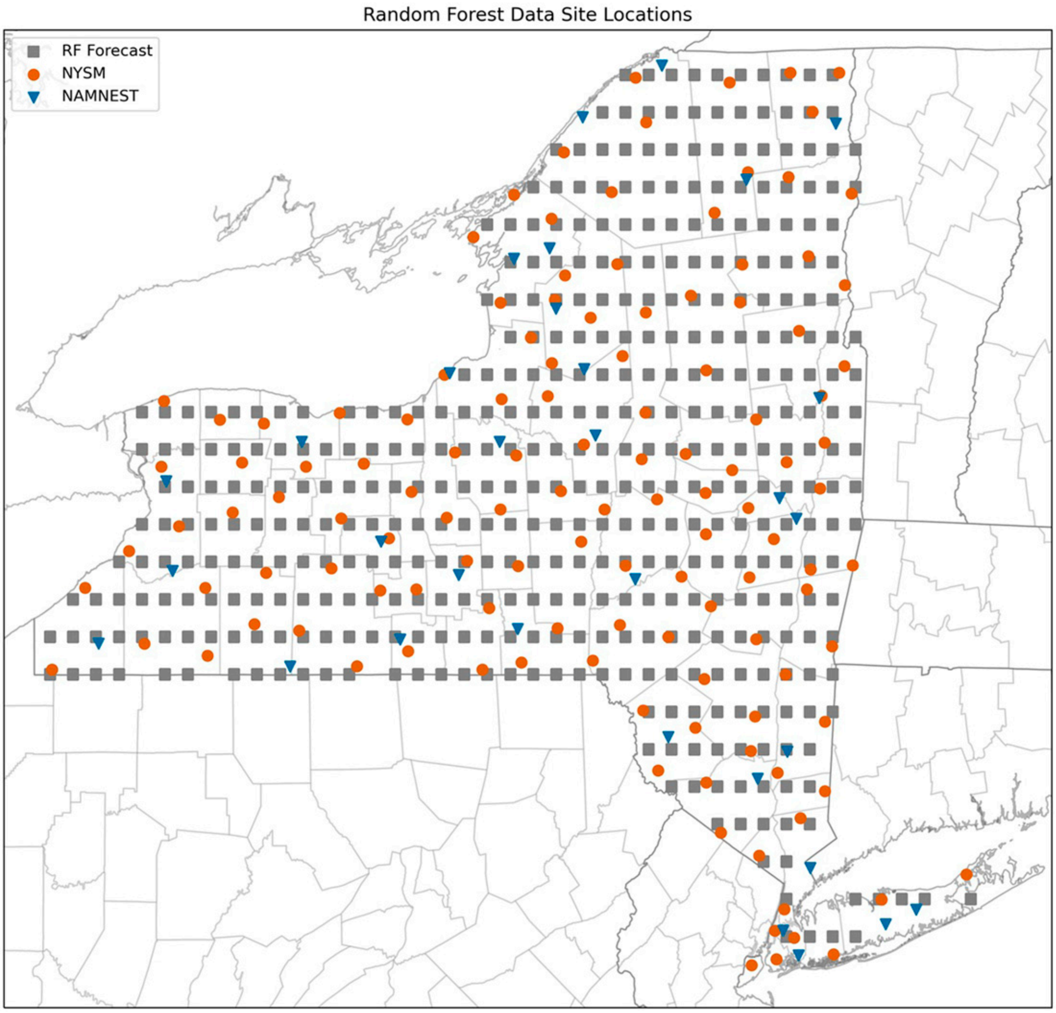

FIG. 9. Map of New York displaying locations of NYSM (orange circles) and NAMNEST profile sites (blue triangles). The gray squares denote 20-km grid spacing of the RF prediction locations.

color-coded probabilities to display highest value at each location; Fig. 11). This last plot type (Fig. 11) was 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.

5. Summary and future directions

Winter mixed-precipitation events can be difficult to forecast because they present numerous challenges to forecasters, ranging from areas of complex terrain to challenging winter storms where precipitation types can transition across very short distances. ML algorithms can help with these challenges by combining multiple big datasets to account for local terrain variation or the widespread area covered by a large winter storm. An RF was trained to make easy to interpret probabilistic predictions of precipitation type to help ease the burden on forecasters who try to synthesize datasets with a large number of variables 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 and upper air) and model (NAMNEST) datasets to create combinations of features in multiple training datasets to be tested extensively for their accuracy in identifying winter mixed-precipitation types, as well as for the best configuration of the RF and features in the training datasets. Slight changes in the composition of these datasets created differences between the RF runs. This work reinforces the idea that the features and data combinations used to train an RF can impact the ability of the RF to skillfully predict precipitation types. Additionally, each data source and combination of features must be treated differently because combinations of features may not be transferrable. This effect was found after seeing the same additional features improve the NAMNEST

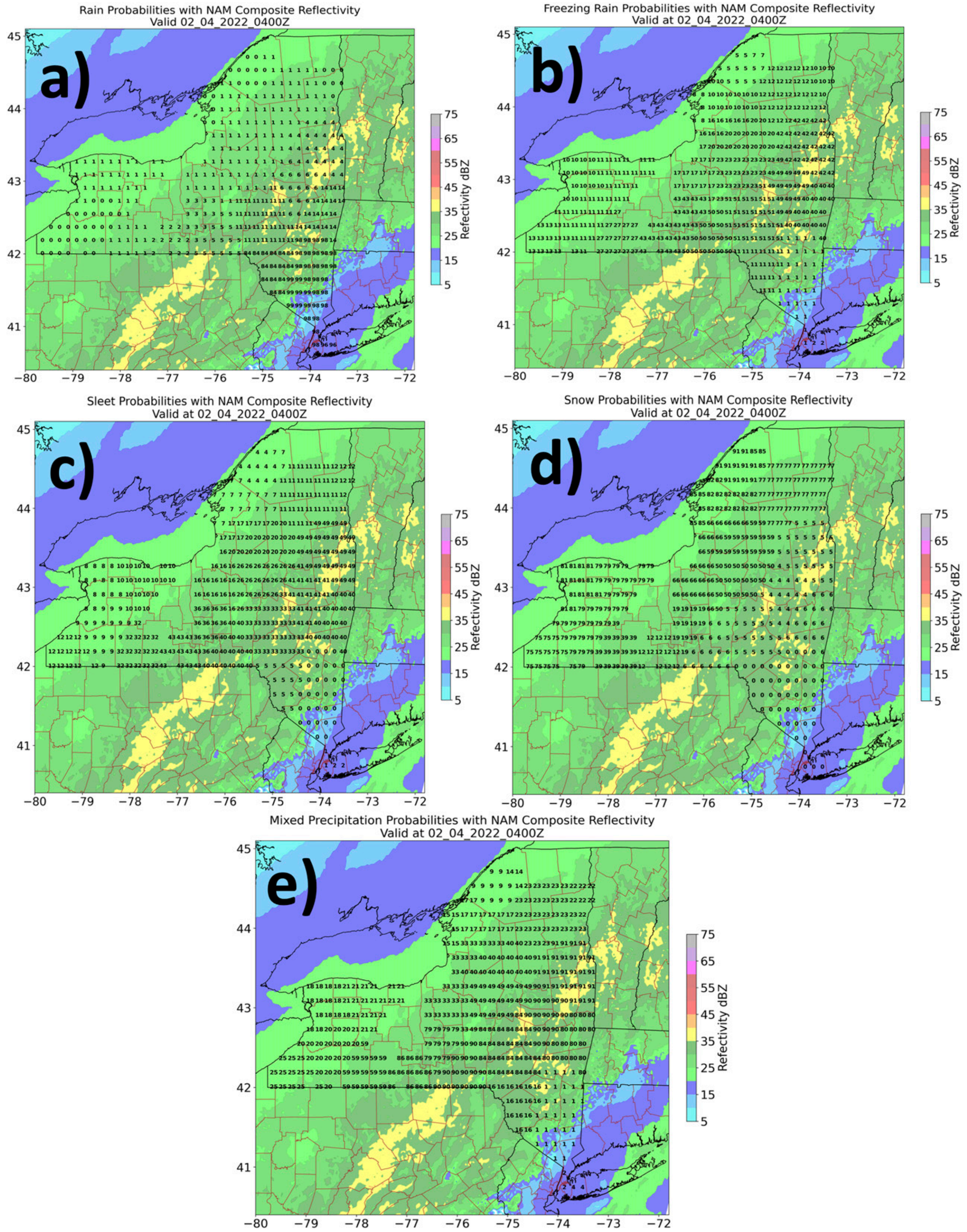
Rain Probabilities with NAM Composite Reflectivity
Valid 02_04_2022_0400Z
a)
Freezing Rain Probabilities with NAM Composite Reflectivity
Valid at 02_04_2022_0400Z
b)
Sleet Probabilities with NAM Composite Reflectivity
Valid at 02_04_2022_0400Z
c)
Snow Probabilities with NAM Composite Reflectivity
Valid 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. 10. 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 4 on 4 Feb 2022 with NAMNEST composite reflectivity (dB*Z*; shaded).

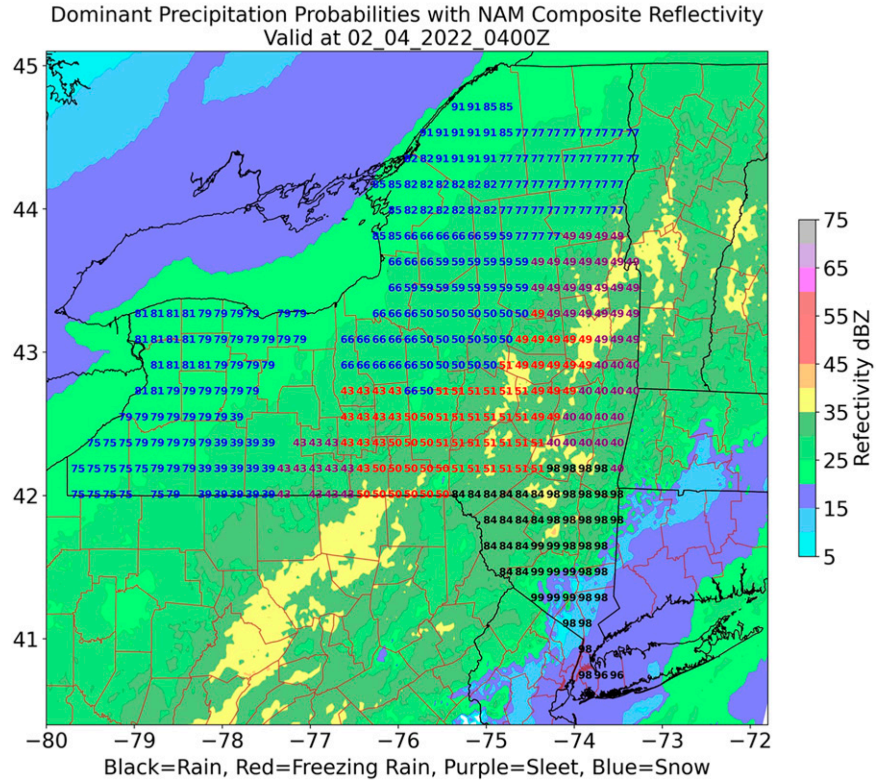
Dominant Precipitation Probabilities with NAM Composite Reflectivity
Valid at 02_04_2022_0400Z
45
44
43
42
41
−80
−79
−78
−77
−76
−75
−74
−73
−72
Black=Rain, Red=Freezing Rain, Purple=Sleet, Blue=Snow
Refectivity dBZ
75
65
55
45
35
25
15
5

FIG. 11. Forecast probabilities of the dominant precipitation type at each location (colored text; the color key is underneath the plot) from the 0000 UTC NAMNEST model run at forecast hour 4 on 4 Feb 2022 with NAMNEST composite reflectivity (dBZ; shaded).

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. This method described how real-time data were processed and matched with grid locations that produced real-time datasets to be compatible with the RF. Data challenges occurred during this transition process such as dealing with missing datasets and real-time datasets not being complete due to hardware or data transmission issues (both for observational devices and data disseminated online). Finally, examples of products available to end users were shown as they allow for end users to have information on all possible precipitation types.

There is much work to be done now that the operational RF is running consistently. The operational application of the RF is being evaluated for the winter season of 2021/22; this verification will be instrumental in making improvements to the RF for end users. Case studies and verification metrics such as skill scores will be conducted and calculated to understand how the RF algorithm performed in real-time and operational situations, which will be reported upon in a follow-up paper. This understanding will improve the RF and will help explain to potential users the limitations and successes of these products. In addition, an increased number of data sources will be added into the RF to continue to expand the number of possible features that can be synthesized. Along with increasing the data sources, the model data resolution will be examined as a method to generate improvements in the RF, because increasing the number of atmospheric sounding points should increase the accuracy of the RF. With the increase in data sources, new products can be developed. This development will be done in consultation with end users to ensure that the products will be valuable to them. This step is important because making sure the information conveys what a forecaster needs is key to incorporating ML algorithms into operations.

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. Working with end users to create meaningful products and training tools will help increase understanding and improve product utility particularly for winter mixed-precipitation events.

Acknowledgments. This research was funded by NOAA Collaborative Science, Technology, and Applied Research (CSTAR) Program Grant NA19NWS4680006 and was also made possible by the New York State (NYS) Mesonet. Original funding for the

NYS Mesonet was provided by Federal Emergency Management Agency Grant FEMA-4085-DR-NY, with the continued support of the NYS Division of Homeland Security and Emergency Services; the state of New York; the Research Foundation for the State University of New York (SUNY); the University at Albany, SUNY (UAlbany); the Atmospheric Sciences Research Center at UAlbany; and the Department of Atmospheric and Environmental Sciences at UAlbany. The authors thank Noah Newman from CoCoRaHS for his help in accessing the archived reports. In addition, the authors thank the NWS Albany Forecast Office staff and CSTAR "focal points" Neil Stuart, Christina Speciale, and Mike Evans for their feedback, guidance, and collaboration in the development of this project.

Data availability statement. CoCoRaHS reports and data are publicly available and accessible on their website (https://www.cocorahs.org). New York State Mesonet data can be requested through their website (http://www.nysmesonet.org/data/requestdata). Radiosonde data were used from the University of Wyoming's archive (http://weather.uwyo.edu/upperair/sounding.html) and accessed via Siphon data request package (https://doi.org/10.5065/D6CN72NW; May et al. 2017). BUFKIT data were used from both the Iowa State Mtarchive (https://mtarchive.geol.iastate.edu/) and the Pennsylvania State University BUFKIT Data Distribution System (http://www.meteo.psu.edu/bufkit/CONUS_NAMNEST_12.html). These data were accessed via a BUFKIT processing package made by Carter Humphreys that is available on Github (bufkit-api; https://github.com/HumphreysCarter/bufkit-api).

REFERENCES

Baldwin, M., R. Treadon, and S. Contorno, 1994: Precipitation type prediction using a decision tree approach with NMC's mesoscale 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, 2016: Explicit precipitation-type diagnosis from a model using a mixed-phase bulk cloud-precipitation microphysics parameterizations. *Wea. Forecasting*, **31**, 609–619, https://doi.org/10.1175/WAF-D-15-0136.1.

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.

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.

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.

Chapman, W. E., A. Subramanian, L. Delle Monache, S.-P. Xie, and F. M. Ralph, 2019: Improving atmospheric river forecasts with machine learning. *Geophys. Res. Lett.*, **46**, 10 627–10 635, https://doi.org/10.1029/2019GL083662.

Chen, R., W. Zhang, and X. Wang, 2020: Machine learning in tropical cyclone forecast modeling: A review. *Atmosphere*, **11**, 676, https://doi.org/10.3390/atmos11070676.

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.

Dwivedi, D., and Coauthors, 2022: Imputation of contiguous gaps and extremes of subhourly groundwater time series using random forests. *J. Mach. Learn. Model. Comput.*, **3** (2), 1–22, https://doi.org/10.1615/JMachLearnModelComput.2021038774.

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. Oper. Meteor.*, **10** (1), 1–17, https://doi.org/10.15191/nwajom.2022.1001.

Elmore, K. L., Z. L. Flamig, V. Lakshmanan, B. T. Kaney, V. Farmer, H. D. Reeves, and L. P. 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., II, A. McGovern, and M. Xue, 2014: Machine learning enhancement of storm-scale ensemble probabilistic quantitative precipitation forecasts. *Wea. Forecasting*, **29**, 1024–1043, https://doi.org/10.1175/WAF-D-13-00108.1.

——, ——, 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.

Gascón, E., T. Hewson, and T. Haiden, 2018: Improving predictions of precipitation type at the surface: Description and verification of two new products from the ECMWF ensemble. *Wea. Forecasting*, **33**, 89–108, https://doi.org/10.1175/WAF-D-17-0114.1.

Haupt, S. E., W. Chapman, S. V. Adams, C. Kirkwood, J. S. Hosking, N. H. Robinson, S. Lerch, and A. C. Subramanian, 2021: Towards implementing artificial intelligence post-processing in weather and climate: Proposed actions from the Oxford 2019 workshop. *Philos. Trans. Roy. Soc.*, **A379**, 20200091, https://doi.org/10.1098/rsta.2020.0091.

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.

——, and ——, 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.

——, and ——, 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. Rev.*, **148**, 2135–2161, https://doi.org/10.1175/MWR-D-19-0344.1.

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, https://ams.confex.com/ams/pdfpapers/94838.pdf.

May, R. M., S. C. Arms, J. R. Leeman, and J. Chastang, 2017: Siphon: A collection of Python utilities for accessing remote atmospheric and oceanic datasets. Unidata, accessed 17 March 2021, https://github.com/Unidata/siphon.

——, and Coauthors, 2022: MetPy: A meteorological Python library for data analysis and visualization. *Bull. Amer. Meteor. Soc.*, **103**, E2273–E2284, https://doi.org/10.1175/BAMS-D-21-0125.1.

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., D. J. Gagne II, J. K. Williams, R. A. Brown, and J. B. Basara, 2014: Enhancing understanding and improving prediction of severe weather through spatiotemporal relational learning. *Mach. Learn.*, **95**, 27–50, https://doi.org/10.1007/s10994-013-5343-x.

——, 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.

——, R. Lagerquist, D. J. Gagne II, 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.

Mital, U., D. Dwivedi, J. B. Brown, B. Faybishenko, S. L. Painter, and C. I. Steefel, 2020: Sequential imputation of missing spatio-temporal precipitation data using random forests. *Front. Water*, **2**, 20, https://doi.org/10.3389/frwa.2020.00020.

NOAA, 1998: Automated Surface Observing System (ASOS) user's guide. ASOS Program Office NOAA/NWS, 74 pp., https://www.weather.gov/media/asos/aum-toc.pdf.

NOAA NCEI, 2022: U.S. billion-dollar weather and climate disasters. Accessed 17 March 2021, https://doi.org/10.25921/stkw-7w73.

NOAA NWS, 2022: Public information statement 22-17. NOAA, 2 pp., https://www.weather.gov/media/notification/pdf2/pns22-17_weather_balloon_launch_frequency_aaa.pdf.

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

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., 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.

——, 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.

Scheuerer, M., S. Gregory, T. M. Hamill, and P. E. Shafer, 2017: Probabilistic precipitation-type forecasting based on GEFS ensemble forecasts of vertical temperature profiles. *Mon. Wea. Rev.*, **145**, 1401–1412, https://doi.org/10.1175/MWR-D-16-0321.1.

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.

Taillardat, M., and O. Mestre, 2020: From research to applications–examples of operational ensemble post-processing in France using machine learning. *Nonlinear Processes Geophys.*, **27**, 329–347, https://doi.org/10.5194/npg-27-329-2020.

Vannitsem, S., and Coauthors, 2021: Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bull. Amer. Meteor. Soc.*, **102**, E681–E699, https://doi.org/10.1175/BAMS-D-19-0308.1.

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.

Wang, X., B. Gao, and X.-S. Wang, 2022: Investigating the ability of deep learning on actual evapotranspiration estimation in the scarcely observed region. *J. Hydrol.*, **607**, 127506, https://doi.org/10.1016/j.jhydrol.2022.127506.

Williams, J. K., 2014: Using random forests to diagnose aviation turbulence. *Mach. Learn.*, **95**, 51–70, https://doi.org/10.1007/s10994-013-5346-7.