A Machine-Learning System for Hydrometeor Identification with Polarimetric Radar

C. A. M. Gerlach[1]

[1]Department of Atmospheric Science, University of Alabama Huntsville, Huntsville, Alabama, USA

Corresponding author: C. A. M. Gerlach, `CAM.Gerlach@Gerlach.CAM`

1 Introduction

Since the advent of polarimetric radar, insight into the phase and morphology of the sampled targets, and in turn the determination of likely hydrometeor classification, has been cited as perhaps their foremost benefit for both research and operational applications, such as microphysical research, rain/hail size classification, and the disambiguation of mixed precipitation types in a winter weather event (Dusan S. Zrnic and Ryzhkov 1999). With the upgrade of the National Weather Service WSR-88D S-band radar network to support dual-polarization capabilities, particular interest developed in applying these data toward enabling better forecaster diagnosis and prediction of rapidly developing weather events, such as severe convection and winter storms (Elmore 2011). Therefore, to formalize and automate the inference of target type from the radar base moments, numerous hydrometeor identification (HID) algorithms have been developed to objectivity classify the bulk composition of each radar volume.

Early such algorithms, such as those by J. M. Straka and D. S. Zrnic (1993) and Höller et al. (1994) applied a boolean-logic decision tree to the polarimetric parameter fields to classify the likely precipitation type at each radar gate. Vivekanandan et al. (1999) pioneered an approach making use of fuzzy-logic principles, using probabilistic membership functions for each variable rather than a strictly ordered, binary-rule-based method, increasing accuracy near parameter-space boundaries, and most following work through to the present day (e.g. Dolan et al. 2013 and Park et al. (2009)) have made use of similar techniques.

Both general methodologies share the common feature of relying on threasholds and membership functions determined *a priori* through independent laboratory and empirical research (see Jerry M. Straka, Zrnić, and Ryzhkov 2000 for a canonical example). These are typically rigorous, extensive and compiled from multiple independent sources, have clear theoretical basis, and enable algorithms developed from them to be applied immediately to any well-calibrated radar of similar wavelength. However, these pre-built tables require substantial adjustment or complete replacement when attempting to apply such algorithms to a radar of different wavelength, sub-optimal calibration, or under different assumptions regarding climate regime, storm morphology, or other characteristics deviating from those under which they were developed, and have no intrinsic means to autogenously adapt themselves to themselves to such.

Considering the hydrometeor classification algorithm (Park et al. 2009) developed for the National Weather Service WSR-88D, the radar of greatest interest for operational diagnosis of precipitation type near the ground, these strengths and limitations are evident. The product was developed principally for use in warm-season deep convection, utilizes simple fuzzy-logic-based techniques applied to theoretically and empirically derived relationships, relies only on native dual-polarization radar data, and was verified in limited field campaigns and case studies. Accordingly, its utility is naturally quite limited when applied to situations outside these design goals, such in winter weather cases, determining actual precipitation type at the ground, and generating accurate probabilistic output, and there are clear opportunities for improvement using larger datasets, exogenous variables, and more sophisticated machine-learning techniques (Elmore 2011; and others cited therein).

Later work by Schuur et al. (2012) have demonstrated improvements by adding NWP-model-derived temperature profile data to the algorithm; however, the other mentioned gaps remain largely unaddressed in the referred literature. The recent mPING project (Elmore, Flamig, et al. 2014), crowdsourcing precipitation-type observations from the public via a smartphone "app" to help fill these voids, offers a uniquely broad-coverage, temporally continuous, very large sample size dataset to serve as approximate "ground truth" for precipitation type at the surface. Despite the relatively informal collection method and untrained observers, these data can be employed to verify a variety of diagnostic and forecast tools (e.g. Elmore, Grams, et al. 2015).

Therefore, we leverage these data to develop a random-forest-based machine learning modeling system to classify winter precipitation type at the surface over the domain of a WSR-88D radar, and characterize its performance. Furthermore, we evaluate the relative predictive value of a variety of the parameters employed, both radar moments and metadata. In section 2 discusses our data and approach, section 3 summarizes the results, and section 4 discusses this work's limitations and outlines possibilities for future improvement.

2 Methodology

In summary, for this work we develop and utilize a system to automatically ingest all mPING precipitation reports meeting user-specified criteria, acquire the appropriate radar volumes for each, extract the relevant radar fields and metadata for each observation point, and train and test a random forest machine-learning model to predict the surface precipitation type at any point domain given the radar moments. Here, we review its basic process flow, input data, and methods used.

2.1 Data and Predictors

For this phase of the project, to limit the computing time and resources required, we limit ourselves to mPING precipitation type reports and their associated radar data spanning two winters: 2017-2018 for training the model, and 2016-2017 for testing its results. Our domain comprised all reports between 10-100 km from the KLWX (Sterling/Baltimore-Washington) WSR-88D dual-polarization radar, to reduce the relative variability in beam heights and sizes at varying ranges, as well as avoid ground clutter at closer ranges and below-beam effects at further ones, as we are working within native radar polar coordinates. In all, these amounted to $n = 2421$ reports for the training data, and $n = 2257$ reports for the test set.

We also greatly reduce the large number of categories reported by mPING to a more reasonable number, to reduce the effects of reporter inexperience, small numbers of reports per category and precipitation-type ambiguity. Initially, we only consider mPING reports in the "Rain/Snow" category, and prune reports of mixed precipitation (due to small relative numbers and likely reporter inexperience) freezing rain or drizzle, due to the latter's lack of a discernible difference in radar properties their non-freezing equivalents, and partition the remaining into liquid ("Rain", "Drizzle") and ice ("Snow/Graupel", "Sleet/Ice Pellets") phase categories for classification.

In all, 759 and 1035 reports of the training and test dataset, respectively, were classified into the former category, while 1662 and 1222 fell into the latter. The discrepancy between the relative numbers per category for the training and test dataset is not insignificant, and another likely artifact of the relatively small number of winter weather events and the public's tendency to concentrate their reports during them, thus leading to large variability between seasons (and likely decreased algorithm performance).

The standard polarimetric base moments (Z_h, Z_{DR}, ρ_{hv} and ϕ_{DP} were read from each file, along with Dopplar spectrum width and velocity. Other intrinsic metrics read from the file or the mPING report and used as candidate predictors include the ground range to the radar, current VCP in use, range and azimuth information to the report location, longitude and latitude of the report location, month and hour of observation time, and the elevation of each scan used. However, not all of these were eventually incorporated into the final algorithm, due to overfitting issues arising from being derived from a relatively small number of spatially-correlated event days.

2.2 Methods

This work's analysis procedure and machine-learning model were implemented as a modular, open-source Python package, to be available on PyPI and Github, to enable re-use and reproducibility of these results by other researchers, as well as practical employment of the machine learning modeling system devised. Aside from the below-mentioned libraries, the standard PyData stack was used throughout for data handling, analysis and visualization.

After querying the OU mPING API to obtain the relevant mPING reports within the study domain, filtering them for minimum range and reducing their categories, we match each with the WSR-88D radar file, as obtained from the Nexrad AWS API, with a volume start time closest to that of the mPING observation. We then retrieved the resulting Nexrad Archive Level II files, and calculated the radar-relative ground distance for each report for later use. PyART (Helmus and Collis 2016) was used to aid the radar portion of data processing, analysis and visualization. We match each report with the range/azimuth bin closest to the report location at each elevation angle closest, which we use as the nominal center point for our radar-based retrievals.

We then average the base moments within a variable-resolution grid from the point of interest, comprising an array of 5 gates (1.25 km) in range, and either 5 gates (2.5 °) in azimuth for Super-Res tilts, or 3 gates (3 °) for non-Super-Res elevations, producing a roughly balanced area at varying ranges. The lowest five elevations are used for each scan, as every volume regardless of VCP has at least five vertical elevations even with AVSET enabled. Only the “non-Doppler” (low-PRF) sweeps are used for the split cut elevations, as it are these variables we are primarily concerned with for determining precipitation type, whereas maximizing data quality and avoiding range-aliased data are of greater interest; likewise, SAILS scans are also skipped.

We then post-process the gathered data to prepare it for the machine-learning classifier, trimming the dataset to the selected predictor and predictand columns (by using feature selection to only employ those that enhanced algorithm performance), filling missing values with a suitable replacement, and concatenating the remaining classifications down to the desired two (liquid/ice). Scikit-Learn (Pedregosa et al. 2011) was used to train a random forest classifier on the training dataset, the former of which essentially comprises a large, bootstrapped ensemble of decision trees. Each of these is not dissimilar to those employed by simpler Boolean HID methods, but each accounting for multiple dimensions as appropriate to classify the data and the large number and randomization reducing overfitting bias and allowing for probabilistic output and a more robust characterization of model uncertainty.

Subsequently, the trained model was then used to predict the precipitation classification for the radar gates corresponding to each of the mPING observations in the test dataset, and the results compared for accuracy. Furthermore, the algorithm hyperparameters and selected features were iteratively tuned for optimal classification accuracy. Finally, various statistics were gathered on these data, and final tables and plots produced. In order to calculate more reliable mean values and confidence intervals for such statistics, we trained 100 models with difference random seeds on bootstrap samples of the dataset and computed means and standard deviations for each.

3 Results

Overall, our algorithm performed reasonably well, with a mean accuracy of 0.692 (95% confidence interval width 0.0107) on the independent test dataset. Given an expected accuracy near 0.5 given the relative equal distribution of rain and snow events (we do not accurately estimate it, given our limited season data availible), this is a statistically significant improvement relative to chance, whereas in Elmore (2011) the NSSL HCA did not demonstrate such in seperating the liquid/ice cases. However, we note that the out of bound accuracy estimate is significantly higher, at 0.850, which indicates a substantial degree of overfitting is present.

3.1 Predictor Analysis

Examining the mean relative importance values calculated from the models, we find that no one predictor dominates, with all the predictors (except for elevation 0) having importance weights between 0.01 and 0.05 (B.1). The only exception, elevation 0, likely was rejected by the random forest model due to exhibiting little to no variance between the events, as every WSR-88D volume scan begins at the 0.5 ° elevation angle regardless of VCP. However, elevation 3 had a statistically significantly ($p < 0.05$) higher importance weight than the other elevations, possibly due to the greater variation for this scan altitude between the VCP x12 and VCP 2x/3x families.

Overall, all polarimetric moment predictors had a similar pattern, with a maximum importance value at the first or second elevation, with decreasing importance as one moves higher in the volume. This is quite physically reasonable given the circumstances, as one would expect a stronger association between the lowest elevation angles and precipitation falling at the ground as opposed those at higher altitudes. Interestingly, all of the moments had relatively similar magnitudes of values, particularly near their maximum, with differences only barely statistically significant. Therefore, this hints at the importance of considering each in any hydrometeor ID algorithm.

Surprisingly, differential reflectivity had a higher importance than any other parameter at most elevations (though not to a statistically significant degree), considering it is not very meaningful by itself without knowledge of its rate of change. Similarly, velocity and spectrum width would not be expected from the literature to have such high predictive value, particularly the former which depends entirely on orientation from the radar. However, a likely reason for the inflated importance of all of these is overfitting, which despite a random forest being resilient to is unavoidable in this case due to the strong spatial and temporal correlation of the small number of events for a season.

4 Discussion

4.1 Limitations

While this technique has been demonstrated to show considerable promise for its specific application, it possesses significant limitations—some intrinsic to the approach itself and the operation of current operational dual-polarization radars, and some that can be addressed through further extensions of this preliminary investigation.

Chief among the latter is the limited size of the test and particularly training datasets, comprising only a relatively small number of spatially and temporally correlated winter storm events over the course of each season. As a result, the random forest algorithm lacked the suitably large and independent dataset needed to fully refine its trees and develop a more accurate picture of the high-dimensional parameter space incorporating dual-pol moments and metadata at multiple radar levels, relative to the relatively low number of dimensions (typically no more than a few pairs of 2D parameter spaces, e.g. in a 2D Boolean or fuzzy logic approach) found in traditional HID algorithms.

Furthermore, many predictors that would likely otherwise show some positive predictive ability (e.g. month, range, azimuth, lat, lon) due to geographic and seasonal variations associated with precipitation type and variations in radar presentation due to increase beam height, reduction in minimum detectable reflectivity, beam broadening and azimuth-specific beam blockage actually reduced accuracy on the test dataset, as these were overfit due to specific values being favored in such events, without being representative of the underlying pattern.

This combination of low sample number and small number of major events drawing the public's attention, with their attended inter-seasonal variability in character, frequency and magnitude, makes it difficult to assess the impact of tuning, parameter selection and other design choices on model accuracy, as sample variability is sufficiently large to land most such changes well within the margin of error (95% confidence interval) of true model performance, as determined from a bootstrapped sample of random forests with varying seeds (which in turn can be viewed as a result of a boostrap sample of the data itself). Therefore, it cannot be determined to a suitably reliable extent whether adding or removing many parameters (i.e. data denial) has a statistically significant effect on ultimate performance (and thus implying an association between the threadsholds of the parameter and precipitation type) unless the magnitude of the resulting change is sufficiently large.

Additional sources of error stemmed from the basic limitations of modern radar. Most prominently, as is ultimately true of any real-world radar, phase or other changes below the beam height were not accounted for, which increased with increasing range. Therefore, the radar data could be missing evidence of a refreezing or melting layer below the beam height, as well as other changes in hydrometeor characteristics. Although somewhat limited by the domain chosen, there remained a substantial discrepancy in the coverage of the 3/5 x 5 averaging areas over the one order of magnitude increase in range due to operating in native radar coordinates, which due to the limited dataset size could only be partially accounted for by the model itself. Similarly, some affects of beam broadening, reduced sensitivity, and the like as mentioned previously may remain adjusted for.

The lack of environmental data leaves the model without much of the key context a human, or a sophisticated fuzzy-logic HCA, would be able to rely on to easily decide many cases, and improve predictability on many others. Furthermore, given only minimal QC filtering was applied to the mPING data, anomolous reports may remain confusing the training and testing process. Finally, reliance on only ϕ_{DP} rather than a proper K_{DP} estimation or at least an analog further limits the algorithm relative to existing approaches, which all incorporate it.

4.2 Extensions and Future Work

While limited in scope, this work is designed to be easily extended in the future to alleviate many of these shortcomings and further explore the possibilities inherent in applying a machine-learning method to these data. Perhaps the most obvious improvement is to simply consider more years in the analysis for a given radar, particularly for the training dataset, given that more are available.

Another simple addition is calculating K_{DP} alongside ϕ_{DP}; while a computationally expensive method could be employed for the entire radar volume, a conveniently-calculated proxy for our purposes would simply be to compute the finite difference along each radial in the 3 x 5/5 x 5 sample around the point of interest, and then azimuthally average them; this would correct for system offset, backscatter phase would be minimal at S-band, and much noise would be smoothed out (the length along the radial could be increased, as necessary).

Better quality control of the initial mPING reports could eliminate spurious observations, also improving both the algorithm's training performance and its consistency on the test dataset. Furthermore, these improvements could enable more sophisticated feature selection and hyperparameter tuning on the model due to reduced sample variability, thus optimizing performance. Similarly, additional classes (principally ice pellets, perhaps along with a common "mixed" class) could be included and predicted for, if sufficient data was available to characterize them accurately.

More fundamental changes could include adding exogenous temperature and other environmental data as predictiors; an initial implementation could merely retrieve the relevant values from the nearest surface station in the NCEI archive, while a more complete solution would ingest RAP vertical profiles and other such NWP variables, which have been shown to significantly increase accuracy in other HCAs (Schuur et al. 2012). This algorithm could be trained and tested on other radar sites as well, either individually or in aggregation.

Additionally, in along with the random forest, an ensemble of machine learning classifiers (SVM, ANN, etc) could be applied to reduce the weaknesses in any one approach. Finally, an ultimate extension could be to the MRMS dataset, which could allow easy incoperation of additional data, extension to a true national scale, increased data quality with reduced non-meteorological variability, and easy Cartesian gridding of results. This could even enable replacement of these relatively simple machine-learning techniques with a more sophisticated, "deep learning" convolution neural network in the future that could detect complex patterns and incorporate conceptual models, such as the bright band location and refreezing level, truly rivaling human radar interpretation.

A: Figures

B: Tables

Table B.1. Relative Importance of each Predictor in Random Forest Model

Notes: Predictor names follow the field names as used in the original data, with the incrementing integer denoting successively higher volume scans. Importance is the mean relative weight over a bootstrap sample of 100 trained models, and ConfidenceInterval is the width of the 95% confidence interval for each importance value calculated from the bootstrap sample.

	Predictor	Importance	ConfidenceInterval
0	ground_range	0.0403	0.0026
1	VCP	0.0110	0.0022
2	cross_correlation_ratio_0	0.0372	0.0028
3	cross_correlation_ratio_1	0.0414	0.0028
4	cross_correlation_ratio_2	0.0389	0.0029
5	cross_correlation_ratio_3	0.0278	0.0024
6	cross_correlation_ratio_4	0.0219	0.0025
7	differential_phase_0	0.0432	0.0040
8	differential_phase_1	0.0422	0.0049
9	differential_phase_2	0.0372	0.0040
10	differential_phase_3	0.0375	0.0048
11	differential_phase_4	0.0375	0.0052
12	differential_reflectivity_0	0.0346	0.0030
13	differential_reflectivity_1	0.0341	0.0027
14	differential_reflectivity_2	0.0279	0.0026
15	differential_reflectivity_3	0.0202	0.0018
16	differential_reflectivity_4	0.0186	0.0021
17	elevation_0	0.0000	0.0000
18	elevation_1	0.0183	0.0057
19	elevation_2	0.0183	0.0055
20	elevation_3	0.0410	0.0085
21	elevation_4	0.0183	0.0068
22	reflectivity_0	0.0372	0.0029
23	reflectivity_1	0.0403	0.0040
24	reflectivity_2	0.0307	0.0035
25	reflectivity_3	0.0250	0.0025
26	reflectivity_4	0.0242	0.0021
27	spectrum_width_2	0.0376	0.0034
28	spectrum_width_3	0.0197	0.0018
29	spectrum_width_4	0.0180	0.0018
30	velocity_2	0.0322	0.0025
31	velocity_3	0.0249	0.0024
32	velocity_4	0.0252	0.0024
33	hour	0.0377	0.0026

Acknowledgments

This work was generously supported and assisted by Dr. Lawrence Carey of the University of Alabama in Huntsville, and Dr. Kim Elmore of the University of Oklahoma provided much helpful input. The mPING report data were obtained from the University of Oklahoma mPING API, access to which is available for free (gratis) for research use, and the WSR-88D NEXRAD Archive Level II radar data employed can be freely obtained (*libre* and *gratis*) from the National Centers for Climate Information, Amazon AWS or Google Cloud.

References

Dolan, Brenda et al. (Sept. 2013). "A Robust C-Band Hydrometeor Identification Algorithm and Application to a Long-Term Polarimetric Radar Dataset". In: *Journal of Applied Meteorology and Climatology* 52.9, pp. 2162–2186. DOI: 10.1175/jamc-d-12-0275.1.

Elmore, Kimberly L. (Oct. 2011). "The NSSL Hydrometeor Classification Algorithm in Winter Surface Precipitation: Evaluation and Future Development". In: *Weather and Forecasting* 26.5, pp. 756–765. DOI: 10.1175/waf-d-10-05011.1.

Elmore, Kimberly L., Z. L. Flamig, et al. (Sept. 2014). "MPING: Crowd-Sourcing Weather Reports for Research". In: *Bulletin of the American Meteorological Society* 95.9, pp. 1335–1342. DOI: 10.1175/bams-d-13-00014.1.

Elmore, Kimberly L., Heather M. Grams, et al. (June 2015). "Verifying Forecast Precipitation Type with mPING". In: *Weather and Forecasting* 30.3, pp. 656–667. DOI: 10.1175/waf-d-14-00068.1.

Helmus, Jonathan J. and Scott M. Collis (July 2016). "The Python ARM Radar Toolkit (Py-ART), a Library for Working with Weather Radar Data in the Python Programming Language". In: *Journal of Open Research Software* 4. DOI: 10.5334/jors.119.

Höller, H. et al. (Sept. 1994). "Life Cycle and Precipitation Formation in a Hybrid-Type Hailstorm Revealed by Polarimetric and Doppler Radar Measurements". In: *Journal of the Atmospheric Sciences* 51.17, pp. 2500–2522. DOI: 10.1175/1520-0469(1994)051<2500:lcapfi>2.0.co;2.

Park, Hyang Suk et al. (June 2009). "The Hydrometeor Classification Algorithm for the Polarimetric WSR-88D: Description and Application to an MCS". In: *Weather and Forecasting* 24.3, pp. 730–748. DOI: 10.1175/2008waf2222205.1.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Schuur, Terry J. et al. (Apr. 2012). "Classification of Precipitation Types during Transitional Winter Weather Using the RUC Model and Polarimetric Radar Retrievals". In: *Journal of Applied Meteorology and Climatology* 51.4, pp. 763–779. DOI: 10.1175/jamc-d-11-091.1.

Straka, J. M. and D. S. Zrnic (1993). "An algorithm to deduce hydrometeor types and contents from multiparameter radar data". In: *26th International Confrence on Radar Meteorology*. Norman, OK: American Meteorological Society, pp. 513–515.

Straka, Jerry M., Dusan S. Zrnić, and Alexander V. Ryzhkov (Aug. 2000). "Bulk Hydrometeor Classification and Quantification Using Polarimetric Radar Data: Synthesis of Relations". In: *Journal of Applied Meteorology* 39.8, pp. 1341–1372. DOI: 10.1175/1520-0450(2000)039<1341:bhcaqu>2.0.co;2.

Vivekanandan, J. et al. (Mar. 1999). "Cloud Microphysics Retrieval Using S-band Dual-Polarization Radar Measurements". In: *Bulletin of the American Meteorological Society* 80.3, pp. 381–388. DOI: 10.1175/1520-0477(1999)080<0381:cmrusb>2.0.co;2.

Zrnic, Dusan S. and Alexander V. Ryzhkov (Mar. 1999). "Polarimetry for Weather Surveillance Radars". In: *Bulletin of the American Meteorological Society* 80.3, pp. 389–406. DOI: 10.1175/1520-0477(1999)080<0389:pfwsr>2.0.co;2.