Evaluation of sub-hourly MRMS quantitative precipitation estimates in mountainous terrain using machine learning

Phoebe White [1], **Peter A. Nelson** [1]

[1]Department of Civil and Environmental Engineering, Colorado State University, Fort Collins, CO, USA

Key Points:

- The error of radar QPEs is highly variable, even in remote areas with complex terrain where QPEs are typically assumed to be unreliable
- Longer lasting precipitation tends to have lower QPE error
- A spatiotemporally varying error model of sub-hourly radar QPEs is developed for the mountains of Colorado

---

Corresponding author: Phoebe White, `phoebe.white@colostate.edu`

Abstract

The Multi-Radar Multi-Sensor (MRMS) product incorporates radar, climate model, and gage data at a high spatiotemporal resolution for the contiguous United States. MRMS is subject to various sources of measurement error, especially in complex terrain. The goal of this study is to provide a framework for understanding the uncertainty of MRMS in mountainous areas with limited observations. We evaluate 8-hour time series samples of MRMS 15-minute intensity through a comparison to 204 gages located in the mountains of Colorado. This analysis shows that the MRMS surface precipitation rate product tends to overestimate rainfall with a median normalized root mean squared error (RMSE) of 42% of the maximum MRMS 15-minute intensity. For each time series sample, various features related to the physical characteristics influencing MRMS performance are calculated from the topography, surrounding storms, and rainfall observed at the gage location. A gradient-boosting regressor is trained on these features and is optimized with quantile loss, using the RMSE as a target, to model nonlinear patterns in the features that relate to a range of error. This model was used to predict a range of error throughout the mountains of Colorado during warm months, spanning 6 years, resulting in a spatiotemporally varying error model of MRMS for sub-hourly precipitation rates. Mapping of this dataset by aggregating normalized RMSE over time reveals that areas further from radar sites in higher elevation terrain show consistently greater error. However, the model predicts larger performance variability in these regions compared to alternative error assessments.

Plain Language Summary

Storms in mountainous regions can develop quickly and cause significant flooding. The lack of precipitation gages in mountainous remote areas inhibits detailed monitoring of these hazardous events. Radar estimates of precipitation can fill the gaps in areas where gages are sparse, but the signal can be blocked by mountains, depending on where the storm is relative to the radar site. Because the error of radar estimates of precipitation can change based on where the storm is located in relation to the surrounding terrain and location of the radar, the reliability of these precipitation estimates is variable, adding to the difficulty of monitoring storms in mountains. Here we develop a novel method of identifying where and when the radar estimates of precipitation are reliable, based on attributes of the region, rainfall, and storm events. The results can assist in deciding when to trust radar estimates of precipitation and in determining where more gages or radar sites are necessary. Unsurprisingly, areas that are far from radar sites and in more complex mountainous terrain have less reliable radar precipitation estimates. However, poor performance in these regions is not certain.

1 Introduction

In mountainous regions, in-situ precipitation observations from sparse gage networks are inadequate for accurate hydrologic modeling and monitoring of hazardous events (Lundquist et al., 2019). Post wildfire debris flows and flash floods are often initialized by brief high-intensity rainfall (Cannon et al., 2008; Doswell et al., 1996) and therefore require high-resolution precipitation observations for accurate modeling and validation of forecasts (Moody et al., 2013; Sokol et al., 2021). Increasing atmospheric moisture tied to global warming has been shown to amplify the intensity of short-duration rainfall (Fowler et al., 2021). With wildfires becoming more widespread in the western United States (Higuera et al., 2021), these events are more likely to be coupled with intense rainfall, increasing the likelihood of flooding and debris flows (Touma et al., 2022). Several studies have shown that the magnitudes of slow-rise floods are also sensitive to spatiotemporal variability of precipitation (Syed et al., 2003; Nicótina et al., 2008; Zhu et al., 2018). Accurate assessment of hydrologic risk requires precipitation observations at the scale of the event, which may be increasingly seen at the sub-hourly and sub-basin scale.

Gages provide precipitation data at high temporal resolutions for point locations, but mountains can cause significant spatial variation in precipitation (R. B. Smith, 2019). Many interpolation schemes exist to create spatially distributed precipitation estimates from gages; for example, the Parameter-Elevation Regressions on Independent Slopes Model (PRISM) is a continuous dataset for the United States developed from interpolated gage data (Daly et al., 2008). However, the accuracy of these interpolated gage datasets declines where gages are sparse (Lundquist et al., 2019). And, the gage itself is subject to various sources of measurement error, such as recording errors and under-catching due to advection of hydrometeors and interception by the tree canopy (Sevruk, 2005; Sieck et al., 2007). Henn et al. (2018) compared the annual accumulation from six gage-based gridded datasets across the Western United States and found that relative error varied from 5 to 60% with the largest error located in regions of high-elevation terrain.

Quantitative precipitation estimates (QPEs) are estimations of precipitation at locations or regions where ground or direct measurements do not exist. Satellite-based QPEs have been improving, but the resolution of QPEs developed from radar reflectivity is often superior (Derin et al., 2016). Radar-based precipitation estimates are high resolution and spatially continuous but are subject to several sources of uncertainty, made worse by mountainous terrain and sampling frequency (Seo & Krajewski, 2010). Radar-based QPE accuracy has improved with the adoption of dual-polarization schemes and improved algorithms to correct for beam blockage (Zhang et al., 2016). The Multi-Radar Multi-Sensor (MRMS) system provides a mosaic of radar for the contiguous United States, improves ground clutter with dual-polarization and better conversion algorithms, and assimilates additional sources of data (Zhang et al., 2016, 2014). The MRMS products are used in mountainous operational forecasting settings for flash flood (NOAA, n.d.; J. A. Smith et al., 2007) and debris flow monitoring (Force, 2005) and to validate satellite QPEs (Kirstetter et al., 2012; Sun et al., 2021). Despite the usefulness of this product, the imperfect conversion from reflectivity to rain rate, ground clutter, and beam blockage all contribute to significant and complex uncertainty (Berne & Krajewski, 2013; Bytheway et al., 2020). MRMS is especially uncertain in the mountainous West where the minimum height observed by radar is significantly higher in the atmosphere (Maddox et al., 2002).

Many studies have evaluated the error associated with QPEs; however, there is a lack of comprehensive error modeling, specifically at sub-hourly resolutions in mountainous terrain. Several case studies have evaluated MRMS at hourly and daily timescales over large regions of complex terrain by comparing precipitation estimates to gages. Moazami and Najafi (2021) compared hourly radar-only MRMS QPEs in Canada to gage data and found that MRMS tends to overestimate precipitation in the southern plains and underestimate precipitation in western and eastern parts of Canada. Bytheway et al. (2019) evaluated hourly gage-corrected and elevation-adjusted MRMS QPEs in the mountains of California and observed that MRMS failed to capture precipitation caused by small scale flow patterns and interactions with terrain. Additionally, rainfall originating closer to the surface was often too low to be picked up by radar. Case studies reveal weather processes and precipitation characteristics related to performance of the QPE, but the error ranges are not generalizable as they can vary by storm and location.

Stream gages have been used to evaluate the spatial variability of error associated with stream flow predictions from well-calibrated hydrologic models (Liao & Barros, 2023; Moreno et al., 2012). The QPE skill can be evaluated throughout the basin, rather than where precipitation gages exist, but this requires accurate flow records and many mountainous watersheds are ungaged. Other models incorporate the spatiotemporal variability in uncertainty by modeling particular components of QPE error. The Radar Quality Index (RQI) is a temporally varying estimate of the reliability of MRMS based on known sources of radar sampling error (Zhang et al., 2012). Probabilistic models estimate the distribution of uncertainty using dense gage networks as reference (Ciach et al., 2007; Villarini et al., 2014). Kirstetter et al. (2015) created a probabilistic QPE for the continental US using radar-only MRMS data by modeling the range of possible precipitation values given the reflectivity and assuming the distribution of possible values. There is still a need for an error model that makes limited assumptions on the structure of error, can be applied in regions where in-situ observations are lacking, and provides spatiotemporally varying estimates of error.

Machine learning excels in identifying trends in complex data with nonlinear and interconnected dependencies for which we have weak or incomplete theory and limited direct observations (Appling et al., 2022). Here we seek to develop a model for estimating error of sub-hourly MRMS rainfall estimates in a mountainous environment. The MRMS 15-minute intensity, an interval used in many applications such as flash flood and debris flow monitoring, is compared to gage observations throughout the mountains of Colorado. A gradient-boosting regressor is trained on sub-samples of time series at each gage location with associated storm and topographic features to predict a target root mean squared error (RMSE) calculated from the gage. The model learns what the expected range of RMSE should be, based on the feature values and target RMSE, and can then predict the error of MRMS where we have no gage records. A statewide dataset, including the features used in training, is developed to predict the theoretical RMSE (if a ground truth existed) for the MRMS-estimated 15-minute intensity. This dataset can be used to understand the range of expected error for sub-hourly intensity estimates for a specific time window and location. Feature importance is calculated to understand what features the model deems most useful in determining QPE error.

Through the gage comparison, modeling of error, and interpolation of error throughout Colorado, we seek to address a series of questions. Firstly, when and where are sub-hourly MRMS precipitation estimates trustworthy in the mountains of Colorado? This information can help forecasters and modelers understand when to trust MRMS. Also, trends in error can help identify where additional observations through gap-filling radar or more gages might be a priority. Secondly, what are the controls on performance or the circumstances that lead to low error? Lastly, how well can error be predicted by the physical aspects of the region, rainfall patterns, and storm characteristics? By examining these factors, we aim to gain a deeper understanding of the predictability of error in MRMS precipitation estimates.

2 Methods

2.1 Study area and data extents

The study area is confined to the state of Colorado, west of 104.5° longitude. This region includes several distinct mountain ranges with varying elevations and aspects. The topographic complexity of this region contributes to a spatially diverse precipitation climatology (Mahoney et al., 2015). The relief of the mountains in Colorado is significant enough to cause disturbances in lower atmospheric flow, which can cause significant spatiotemporal variability in precipitation (R. B. Smith, 2019). The area was selected to test the model's ability to learn the influence of terrain blockage on QPE performance with interactions between various characteristics of storms unique to mountain precipitation. The terrain of the study area is shown in Figure 1 with the locations of gages and radar for reference.

Both gages and radar-based QPEs are subject to several sources of error associated with freezing precipitation (Zhang et al., 2016), so the study period is constrained to warmer months (May through September) to limit the evaluation of MRMS with unreliable gage data. Based on availability of archived MRMS data, a total of 30 months were included in the modeling from years 2018 through 2023.

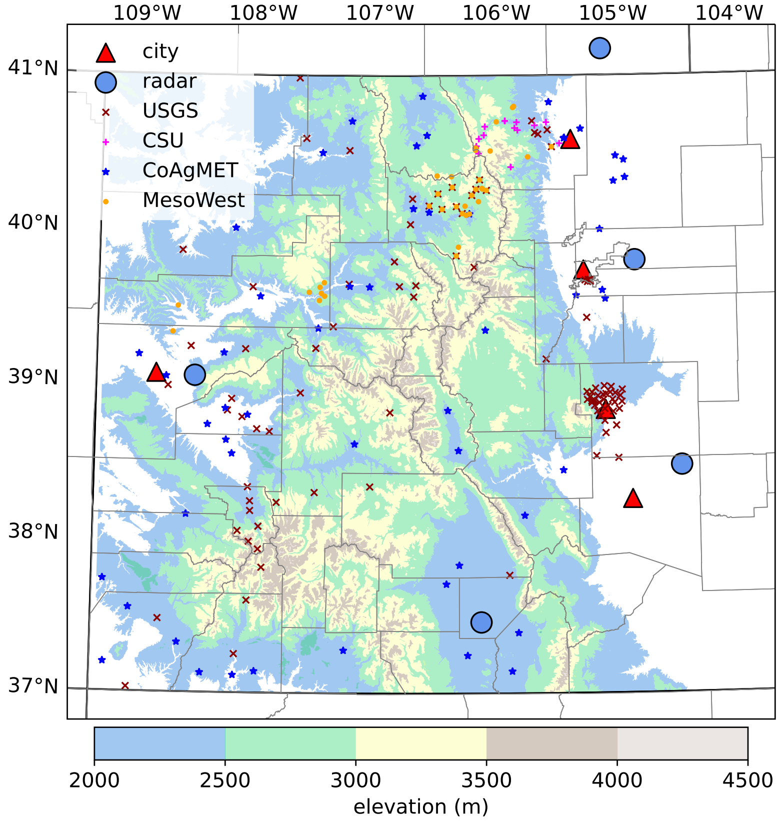

Figure 1. Terrain and instruments within the study area. Gages used in the study are labeled according to their source.

2.2 Data sources and processing

2.2.1 MRMS and gage data

MRMS includes a collection of products with a coverage of the contiguous United States and southern Canada. This study evaluates the surface precipitation rate product, a 2-minute precipitation rate estimate from a mosaic of the Weather Surveillance Radar-1988 Doppler (WSR-88D) radar network and out-of-network gap-filling radar with a spatial resolution of 1 km (Zhang et al., 2016). This product includes several quality controls and an evaporative correction from the Rapid Refresh (RAP) model data (NOAA, n.d.). Missing MRMS data was filled with zero for calculations. All data were accessed through the Iowa Environmental Mesonet MRMS archived daily data (Mesonet, n.d.). Each file's spatial coverage was resampled to the extents of Colorado using wgrib2 (NOAA/CPC, 2023).

Initially, we explored how the 1-hour multi-sensor QPE Pass 2 might be used to improve the 2-minute precipitation rate accuracy. The multi-sensor QPE Pass 2 includes additional information from several numerical weather prediction (NWP) models, various gage networks, and climatological data such as elevation corrections from the PRISM dataset. The inputs vary depending on radar coverage, the type of precipitation, and local topography (NOAA, n.d.). A simple multiplicative correction factor, using a ratio of the 1-hr multi-sensor QPE to the 1-hr radar-only QPE, was applied to the 2-minute precipitation rate product to decrease error. A factor of 1 was used where the 1-hr radar-only QPE was equal to zero.

The corrected and uncorrected radar-only 2-minute precipitation rate products were compared to the gage-observed 15-minute intensity. The multi-sensor correction did not significantly improve the RMSE error, with a median normalized RMSE of 41% of maximum MRMS 15-minute intensity for years 2021 through 2022 (compared to 42% for the radar only rate), see Supporting Information Figure S1 for a comparison of the overall distributions of errors. Because the multi-sensor product did not significantly improve the sub-hourly precipitation estimates, and the multi-sensor product is not archived before 2021, we did not include the correction in this analysis.

The Radar Quality Index (RQI) was used to train the model and as a baseline to compare the performance of model estimated error. This product has a temporal resolution of 2 minutes and provides an estimate of uncertainty, represented by a floating-point number ranging from 0 to 1, that is related to terrain blockage, higher beam heights, and the beam position with respect to the freezing level (Zhang et al., 2012). An RQI of 1.0 suggests limited uncertainty associated with these problems. RQI is set to 0.0 when terrain blockage is greater than 50%.

The gage data was sourced from instruments monitored by Colorado State University (White & Nelson, 2024a), the U.S. Geological Survey (USGS) (*USGS National Water Information System*, 2023; Rengers et al., 2023), CoAgMET (Colorado State University, 2023), and other networks aggregated by MesoWest (*Mesowest - Weather and Climate Data*, Accessed 2023). The locations of these stations are shown in Figure 1. The set of gages includes various models of tipping buckets and one disdrometer. Several gages record at regular time intervals (ranging from 1 minute to 5 minutes) and some gages record tips with varying record lengths. A total of 204 gages were used to develop the dataset. Gages were quality controlled by comparing the magnitude of RMSE across the study area. One gage, located in central Fort Collins, was removed from the dataset because it had significantly higher error, with a mean RMSE of 42.8 mm/hr compared to the overall dataset mean of 3.43 mm/hr.

2.2.2 Feature extraction and labeling

The labeled dataset is comprised of 37,215 samples, consisting of an 8-hour non-overlapping window of 15-minute intensity. Each sample has 28 features evaluated within the 8 hours and is labeled with a target RMSE calculated from the gage and MRMS intensity with the gage values assumed to be the ground truth. The sampling start times include 0800, 1600, and 0000 UTC. The purpose of the 8-hour window is to focus on chunks of time impacted differently by the diurnal cycle. In the labeled dataset, features and error are evaluated in relation to the MRMS grid location closest to the gage. Only 8-hour samples with an MRMS total accumulation above 1 mm were included to prevent sampling below the minimum accumulation that can be observed by a gage. The purpose of the model is to predict the error threshold of a non-zero time series of MRMS 15-minute intensity; the model cannot determine if the time series is a false negative because we did not include zero-value accumulations.

Feature values were calculated for each 8-hour time series window, and include both point and storm attributes. Storm attributes were averaged throughout the storm footprint. When several storms occur during an 8-hour window, storm values were averaged. Table 1 provides a description of all features with a comparison of the mean and standard deviation of feature values for the labeled and statewide datasets. Features were chosen to capture the variability and magnitude of the rainfall, the orographic influences, the radar QPE uncertainty, and space-time characteristics of the storms.

Rainfall statistics were calculated from nonzero values in the 8-hour sample window. Elevation, aspect, and slope were extracted from the NASA Shuttle Radar Topography 1-arc second dataset (NASA Jet Propulsion Laboratory, 2022). Aspect and slope were calculated using ArcMap (Esri, 2020). Non-precipitation values with missing data were filled with NaN. All feature values were standardized for training of models by subtracting the mean and scaling it to the standard deviation of the feature values.

Storm regions were identified using the scikit-image label module after converting the MRMS precipitation rate product to a binary array, with a value of 1 assigned to values greater than zero (van der Walt et al., 2014). The label module assigns a unique storm-id to neighbors in the array with equal values. Diagonal neighbors were not considered connected to avoid stringing together objects. Objects are assigned storm ids throughout time, latitude, and longitude. Storm values were processed in monthly chunks to manage memory resources and values were not computed across chunks.

A statewide (west of 104° longitude) dataset was developed to visualize the predictions of the model and create a spatially continuous error estimate of MRMS where no gage exists. This dataset includes 598,310 8-hour time series of MRMS intensity with the 28 features sampled throughout the state where the total accumulation is greater than 1 mm. To decrease the size of the statewide dataset, the original MRMS dataset was sampled every 10 grid coordinates, resulting in a resolution of approximately 10 km. Data were sampled from May through September from 2018 through 2023.

MRMS is subject to several temporal sampling errors mainly caused by advection of hydrometeors (Fabry et al., 1994). Rather than directly compare the gage and MRMS time series, the time series was first upsampled to 10 minute intervals. The maximum 15-minute intensity was selected from each 10 minute interval. A more complex approach that includes correcting for sampling error using the movement of the storm such as the method presented by Seo and Krajewski (2015), would likely decrease the error further.

Table 1. Feature Descriptions and Associated Mean and Standard Deviations for the Labeled and the Statewide Datasets

Feature	Mean labeled	Mean statewide	Std dev Labeled	Std dev statewide
Point				
maximum 15-min intensity, mm/hr	8.68	7.41	13.02	11.91
median of nonzero intensity	1.41	1.25	2.16	2.05
std dev of nonzero intensity	2.40	2.03	3.73	3.44
variance of nonzero intensity, $(\mathrm{mm/hr})^2$	19.69	16.00	71.92	68.71
mean of nonzero intensity	2.30	1.98	3.12	2.91
maximum accumulation, mm	0.47	0.39	0.70	0.64
total accumulation	5.09	4.52	6.42	5.62
median of nonzero accumulation	0.06	0.05	0.10	0.10
std dev of nonzero accumulation	0.12	0.10	0.19	0.17
variance of nonzero accumulation, mm^2	0.05	0.04	0.16	0.16
mean of nonzero accumulation	0.11	0.09	0.14	0.14
duration of nonzero accumulation values, min	234.33	242.20	134.50	133.23
month of 8-hour window	6.85	6.85	1.28	1.30
starting hour of 8-hour window	9.39	9.28	7.46	7.44
latitude of MRMS closest to gage	39.32	39.01	0.93	1.16
longitude of MRMS closest to gage	254.18	253.91	1.17	1.36
min Radar Quality Index	0.78	0.68	0.29	0.30
max Radar Quality Index	0.82	0.74	0.25	0.28
mean std dev Radar Quality Index	0.81	0.72	0.26	0.28
median Radar Quality Index	0.82	0.73	0.26	0.28
std dev Radar Quality Index	0.01	0.01	0.03	0.04
elevation at the MRMS coordinate, m	2161.07	2422.43	449.31	616.19
slope aspect at the MRMS coordinate, deg	178.53	169.67	107.17	103.50
topographic slope at the MRMS coordinate, deg	7.15	9.16	8.01	8.46
Storm				
mean of variance of total accumulation between time steps of storm object	698.29	647.15	522.76	510.81
mean of variance of accumulation between coordinates of storm object	5.48	5.08	4.67	4.45
mean of storm footprint area, km^2	91515.87	87971.39	62453.04	62948.86
mean velocity of storm objects, m/s	13.94	13.56	4.63	4.76

2.3 Modeling

2.3.1 Model description

The dataset labeled with true RMSE values from gages is used to train a tree-based ensemble model, which is then used to predict the error of MRMS with the statewide dataset and understand how the extracted features are associated with performance. The output of the model is a range of RMSE values for the MRMS time series sample. To predict this range, a gradient-boosting regressor is trained with quantile loss, using the implementation from scikit-learn (Pedregosa et al., 2011). Boosting methods ensemble weak learners to produce strong learners, improving the balance of bias and variance in individual decision trees, with the aim of decreasing over-fitting and increasing accuracy (Géron, 2021). The gradient-boosting algorithm learns sequentially from the residual error of an ensemble of decision trees (Friedman, 2001). Each ensemble of decision trees is essentially attempting to improve on the residual error from its predecessor. An additional benefit of gradient-boosting models is that there are relatively few parameters to tune.

The residual error or training loss is calculated as mean quantile loss, otherwise known as mean pinball loss, shown in Equation 1:

$$L(y,\hat{y}) = \frac{1}{N}\sum_{i=0}^{N-1} \alpha \cdot \max(y_i - \hat{y}_i, 0) + (1-\alpha)\cdot \max(\hat{y}_i - y_i, 0) \tag{1}$$

Optimizing the loss involves minimizing the difference in the predicted quantile and a target group of size N. We essentially have three models with different α values, representing the quantile, to predict the median and 90% confidence interval. In Equation 1, when $\alpha = 0.50$ the loss is equal to half of the mean absolute error. The loss (L) is a function of the observed value (y) and the predicted value ($\hat{y}$).

Model performance was initially evaluated on a diverse set of single-output regression models to compare performance. The models included k-nearest neighbors, decision trees, various decision tree ensemble models, linear regression, multilayer perceptron, and support vector machines. Models were chosen based on common use and varying complexity. A brief description of these models can be found in the Supporting Information Table S1. After hyperparameter tuning, the coefficient of determination and mean absolute error (MAE) was 0.80 and 0.96 mm/hr for the highest-performing model, random forest regression. The model is not robust to outliers and makes many predictions significantly higher and lower than the target value (Table S2, Figure S2 in Supporting Information S1). The quantile regression model encapsulates these erroneous predictions by providing a range of error. The confidence interval can also account for heteroscedasticity in the data. Linear quantile regression was evaluated for comparison, the performance is shown in section 3.3.

2.3.2 Evaluation

Models were evaluated using 5-fold cross validation and a separate test set. The validation set is used to understand how each model generalizes to new data, compare model performance during training, select hyperparameter values without leaking knowledge from the test set, and to compare feature importance and model calibration to the test set. During k-fold cross validation, the model is trained on k-1 folds then validated with the remaining fold. 20% of the dataset was kept entirely separate to evaluate the final model performance.

Performance of spatiotemporally correlated data can easily be inflated if dependencies in the sample feature values are not accounted for when splitting the training and testing data (Roberts et al., 2017). To limit correlation between data folds and testing data, samples were grouped by location. Cross folds were manually set to keep samples with the same location in the same fold.

The models tested in this study assume independence between features. Correlated features must be removed to interpret how features are used by the model. Several features listed in Table 1 are unsurprisingly highly correlated. The Spearman's rank correlation test was used to determine the correlation between features and Ward's linkage was used to sort clusters of correlated features. A threshold of 0.55 was used for the distance linkage, and only one feature was selected from each cluster below this threshold. Each permutation of features below the threshold was evaluated by the $\alpha = 0.50$ model to determine the best performing group, with a remaining correlation high of 0.61.

After feature selection, randomized search with cross fold validation was used to tune hyperparameters for each quantile separately. This method randomly samples parameter values from a specified range through multiple iterations then outputs the parameter values for the highest performing estimators, based on negative mean quantile loss. Hyperparameters were altered to increase performance while decreasing the tendency to over-fit. See Supporting Information Table S3 for the hyperparameters tuned and the final values chosen.

The performance of the tuned models was evaluated with mean quantile loss in validation and testing. Additionally, the fraction of target values that fall outside of the predicted 90% confidence interval were calculated to get a sense of how well calibrated the predicted confidence interval is; this value should be close to 10%. Feature importance was evaluated by randomly permuting each feature for both the validation and test dataset and determines importance based on the mean increase in mean quantile loss when permuting the feature. The comparison of validation and test importance shows how the model might be over-fitting with the features.

3 Results

3.1 MRMS performance trends

Figure 2a compares the distributions of normalized RMSE (nRMSE) values for all target values, test target values, the test $\alpha = 0.50$ predictions, and the statewide $\alpha = 0.50$ predictions. The values are normalized with the maximum MRMS 15-minute intensity. The distributions peak at approximately 0.4 or 40% of the maximum intensity. The target nRMSE values appear to have a higher density for lower nRMSE values than both the test and statewide predictions, with the statewide predictions having a higher density towards larger nRMSE values. This suggests the model is slightly conservative, or tends to overestimate the error.

Figure 2b shows the distribution of $\alpha = 0.50$ predictions for the state, compared to the distributions of $\alpha = 0.50$ state predictions grouped by coordinates with lower (median prediction below 0.1 quantile) and higher (median prediction above 0.9 quantile) errors. The distributions of error not only shift based on the separation of high versus low error, but for the higher error, the density increases in the right tail. And, the opposite occurs for the low error distribution.

To understand the occurrence of false negatives the labeled dataset was constrained to a total accumulation of 1 mm for either the gage or MRMS (rather than just MRMS). 4.4% of the 8-hour samples had 0 mm MRMS total accumulation when the gage was above 1 mm (false negative). 24.2% showed greater than 1 mm MRMS total accumulation when the gage was 0 mm (false positive). 71% of the 8-hour samples had a higher total accumulation value for MRMS.

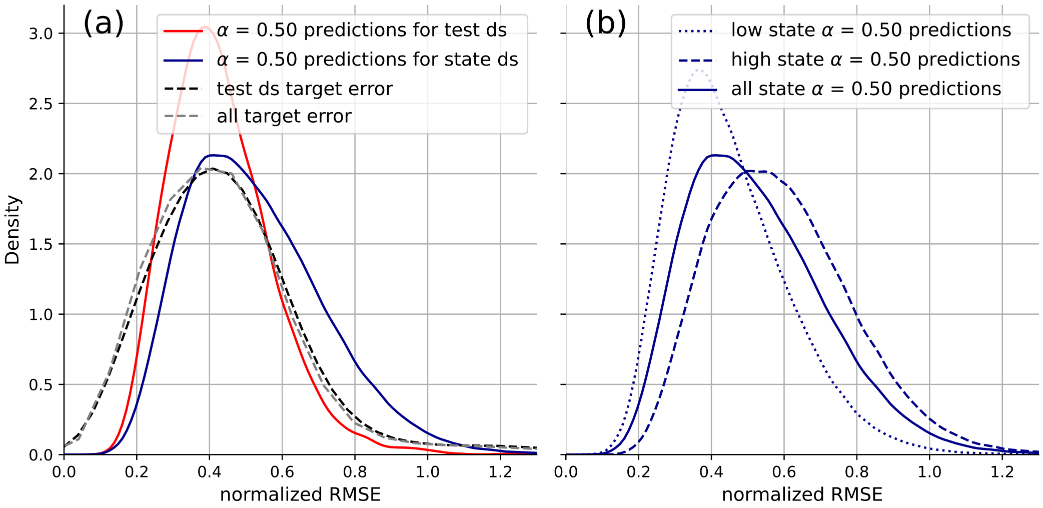

Figure 2. (a) Kernel density estimations of the distributions of normalized RMSE (nRMSE) for all target error, the test target error, the test predicted error, and the statewide predicted error. (b) The distribution of error for the coordinates in the state dataset with a median nRMSE below the 0.1 quantile, above the 0.9 quantile, and all state predicted error values. The values only include the $\alpha = 0.50$ model predictions.

The spatial distribution of predicted median error is shown in Figure 3a. The MRMS grid locations with co-located gages are colored by the proportion of samples in the validation/testing labeled dataset. The lack of spatial uniformity in sampling for the labeled dataset does not appear to influence the spatial trends. There is a band of poor performance that arcs from the northwest corner of the state to the southwest corner that may be influenced by the terrain west of the continental divide and the distance from radar sites. However, areas on the map that show higher median nRMSE still have a distribution of error that includes lower nRMSE values, shown by the map of the first quartile of nRMSE in Figure 3b.

3.2 Features related to MRMS performance

The permutation feature importance for all quantile models for the validation and test dataset predictions is shown in Figure 4. The error bars represent the standard deviation from 10 iterations of randomly permuting each feature. Unsurprisingly, the standard deviation of MRMS 15-minute intensity was most important. Linear regression, using only standard deviation of intensity, results in a coefficient of determination of 0.76 and MAE of 1.03 mm/hr. The standard deviation of intensity was removed from Figure 4 to understand if other features influence the model predictions or are used to overfit in training.

There is some agreement between the feature importance in the testing and training data for the 0.50 and 0.95 quantile models. The 0.05 quantile model seems to rely entirely on the standard deviation and median of intensity. The discrepancy in the ranking of features between models and between the training and testing feature importance values within models is evidence that the models did over-fit. Although, these differences are marginal considering the magnitude of permutation importance. The RQI minimum and duration of positive values were both important for the 0.50 and 0.95 quantile models. This consistency suggests that these features are slightly useful in determining error.

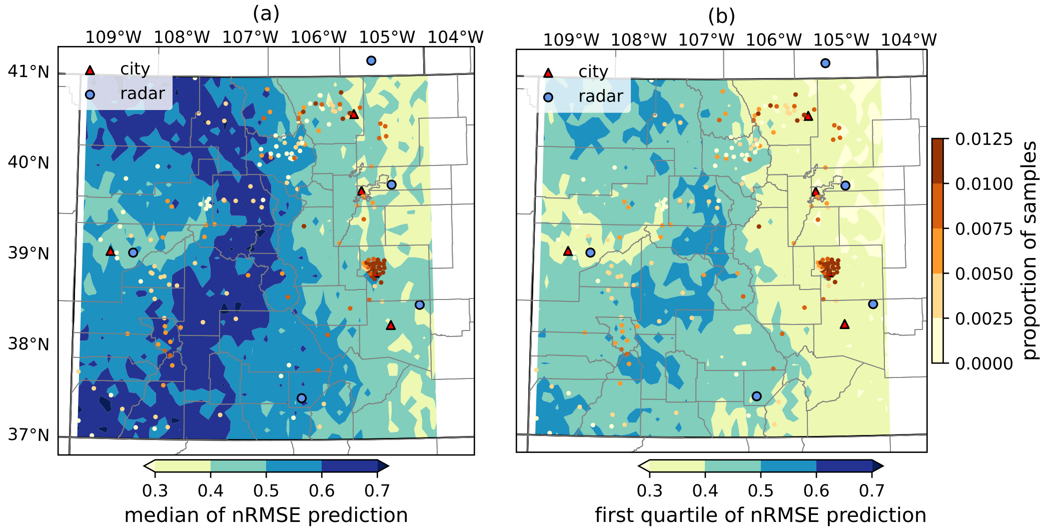

Figure 3. (a) The median normalized RMSE for each coordinate, predicted by the $\alpha = 0.50$ model. (b) The first quartile of normalized RMSE for each coordinate, predicted by the $\alpha = 0.50$ model. The gage locations are shown with colors representing the proportion of samples in the labeled dataset sourced from each location.

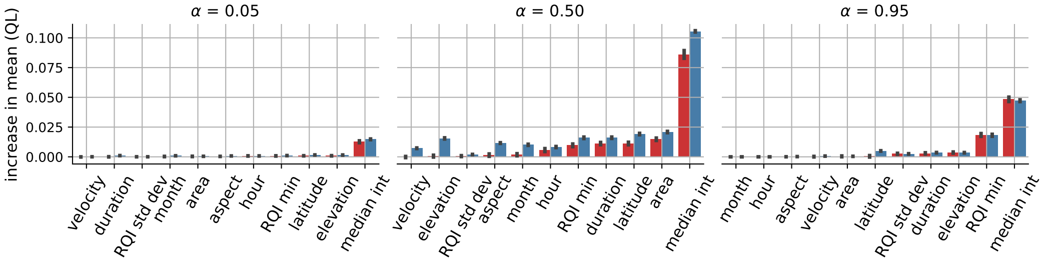

Figure 4. Permutation feature importance, defined by an increase in mean quantile loss (QL), across the three quantile models. The intensity standard deviation is excluded to focus on features not highly correlated with RMSE. The error bar represents the standard deviation between the 10 permutations. Importance is sorted from left to right by test importance (red). The training importance is shown in blue.

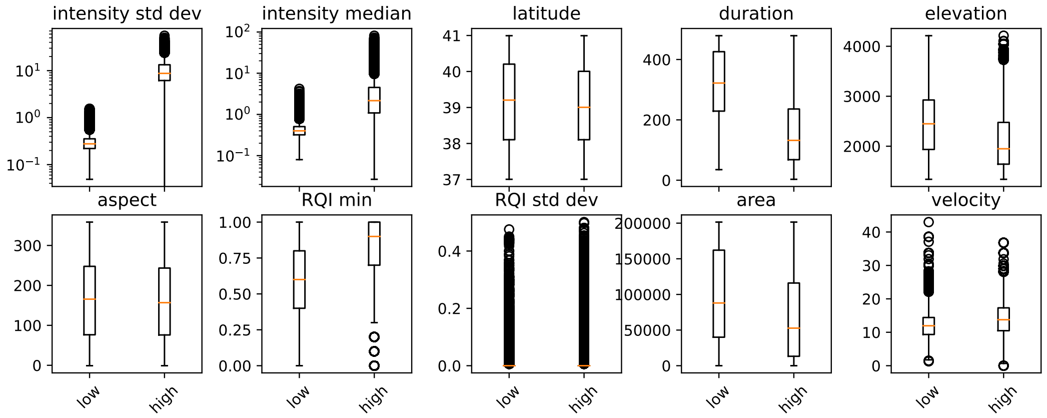

Figure 5. Comparison of feature value distributions for samples with low (0.95 mm/hr) and high (5.60 mm/hr) RMSE predictions for the $\alpha = 0.50$ statewide predictions. See Table 1 for feature units.

The difference in feature values for the statewide dataset for high and low predicted error extremes from the $\alpha = 0.50$ model are shown in Figure 5. Distributions of feature values are separated by low (RMSE less than 0.95 mm/hr or the 0.10 quantile of predictions) and high (RMSE greater than 5.60 mm/hr or the 0.90 quantile of predictions) predictions. The hour and month features are excluded, because they are essentially categorical features.

There is a clear difference in the inter-quartile range for standard deviation of intensity and median intensity. The inter-quartile range is lower for the low RMSE values for standard deviation and median intensity features, as expected. The inter-quartile range for duration is non-overlapping between groups, suggesting that MRMS is more accurate when longer duration rainfall occurs. The distributions for other features differ across performance groups, but are highly overlapping, meaning that there is no obvious decision boundary observed in these individual features.

3.3 Model performance

A summary of validation and test results is shown in Table 2 for both the gradient-boosting (GB) and linear (L) regressors. A perfect value for quantile loss (QL) is zero. A perfect value for the fraction of values outside of the 90% confidence interval (FO) is 0.10. The validation standard deviations for the quantile loss and outside of the 90% confidence interval show how the model performance varies across different data folds. Model performance on the test data is slightly lower, which is expected given that features were chosen and hyperparameters were adjusted based on the validation data. The linear regressor is slightly better calibrated, but the gradient-boosting regressor has lower variation in performance across folds and lower mean quantile loss for all models in both validation and testing.

Figure 6 shows the target RMSE as a function of the predicted median RMSE with the predicted 90% confidence interval the background for the test dataset. A 1:1 line is plotted to show how the median prediction either over or underestimates with growing RMSE magnitude. The confidence interval widens as the target RMSE values increase. The kernel density estimation for the target values as a function of the median predicted RMSE is shown in cyan contours. The kernel density contours show that the median predictions slightly underestimate the RMSE. The scatter points depict the target RMSE

Table 2. Cross Fold Validation and Test Results [a]

	Validation				Test	
	Mean QL	Std QL	Mean FO	Std FO	QL	FO
GB α =0.05	0.100	0.013			0.104	
GB α =0.50	0.456	0.040	0.111	0.015	0.454	0.120
GB α =0.95	0.189	0.020			0.199	
L α =0.05	0.110	0.018			0.115	
L α =0.50	0.474	0.044	0.102	0.020	0.483	0.105
L α =0.95	0.208	0.018			0.215	

[a] Results for quantile loss (QL) and fraction of outliers (FO) for both gradient-boosting (GB) and linear (L) regressor models. Validation results include the mean and standard deviation across folds.

values, the red indicates values outside and blue indicates values inside the 90% confidence interval.

4 Discussion

4.1 When and where are sub-hourly MRMS precipitation estimates trustworthy in the mountains of Colorado?

To better understand the quality of MRMS in the mountains of Colorado, sub-hourly MRMS precipitation estimates were compared to gage records from 2018 through 2023 during warm months. The gage record comparison shows a wide distribution of nRMSE with a median of 42% and a standard deviation of 40% of the maximum MRMS 15-minute intensity. A gradient-boosting regressor was used to interpolate error estimates where no gages exist and provide a spatiotemporally varying record of error in Colorado. The model predictions of nRMSE and RQI aggregated through time at each coordinate show some spatial agreement for high error and uncertainty, shown in Figures 3a and 7a. According to our model results, these areas still see lower error values, perhaps when longer-duration storms occur higher in the atmosphere rather than during orographically enhanced or smaller convective disturbances. Figure 7b shows that areas with the lowest median RQI minimum (high uncertainty) remain low even in the upper quartile. This contrasts with the model results, shown in Figure 3b. The lower variability in RQI is caused by the essentially fixed beam blockage factor. The model shows that MRMS error in remote areas with poor radar coverage and complex terrain is not as consistently unreliable as suggested by RQI or the study by Maddox et al. (2002).

4.2 What are the circumstances that lead to low error?

The standard deviation and median of intensity were unsurprisingly the most important predictors of RMSE. Although, the RQI minimum and duration of rainfall did add some predictive skill, see Figure 4. Separation of very low and very high RMSE predictions for the state dataset, shown in Figure 5, indicates that longer duration precipitation tends to have lower predicted error. This suggests that radar is more skilled at estimating precipitation for storms producing continuous rainfall. Permutation importance is lower for features associated with the terrain and spatiotemporal attributes of storms, meaning that the model did not find these features useful in determining the RMSE.

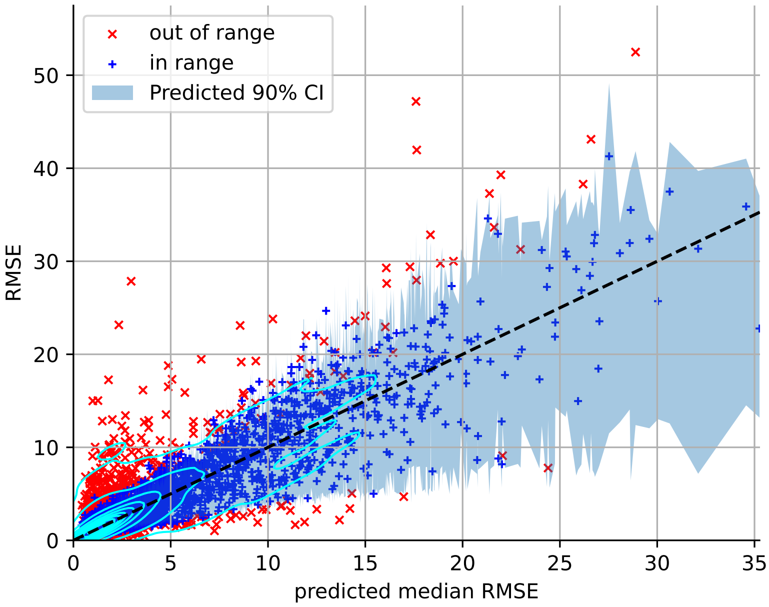

Figure 6. The target values as a function of the median predicted RMSE for the test dataset. Each data point represents an 8-hour sample RMSE between gage and MRMS plotted on the predicted. The predicted 90% confidence interval (CI) is plotted. Truth values in range are green, those that fall outside the cutoff are red. The kernel density estimation for the target values as a function of the median predicted RMSE is shown in cyan contours.

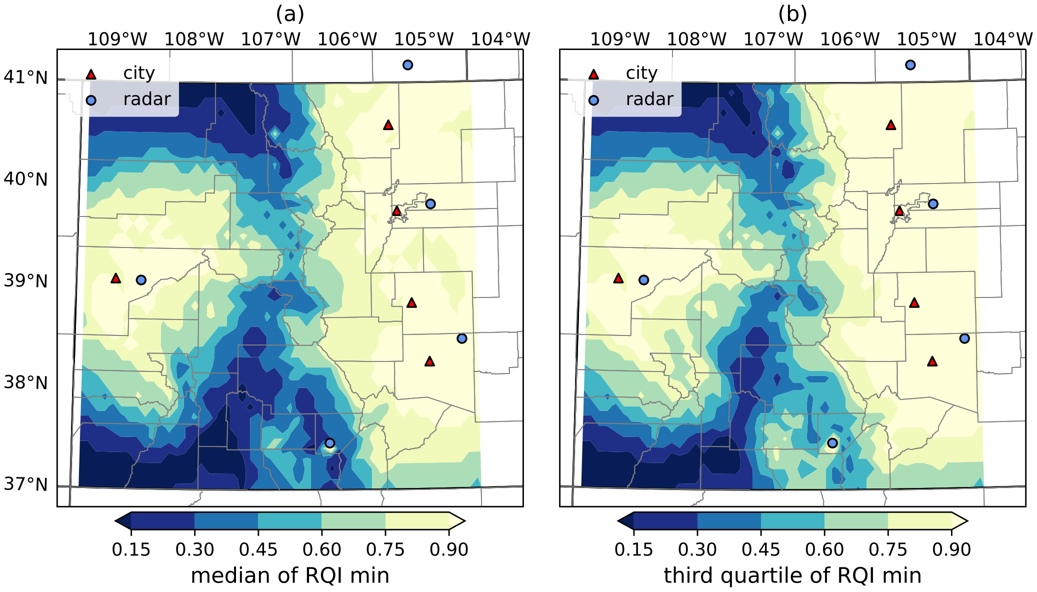

Figure 7. (a) Median of all sample RQI minimums for each coordinate in the statewide dataset across the 6 years. (b) The third quartile of RQI minimums.

4.3 How well can error be predicted by the physical aspects of the region, rainfall patterns, and storm characteristics?

Given the features we developed, the model is well calibrated to the 90% confidence interval and the mean quantile error is low for the test results, shown in Table 2. Model predictions can be used to evaluate the reliability of historic sub-hourly MRMS time series and identify where more observations are needed. This method also provides an approach to estimate error without making assumptions on the sources of uncertainty that are most important or the distribution of error. The type of algorithm used to model the error does not appear to be critical, as evidenced by the marginal performance gains of the gradient-boosting regressor over the linear quantile regressor.

Model results are limited by the assumption that the gage network is a reliable ground truth for MRMS. The difference in spatial sampling scale (point versus 1 km^2 grid) as well as the estimation of surface precipitation from a high radar beam height is known to cause significant uncertainty in error estimates (Tang et al., 2018; Villarini et al., 2008; Ciach & Krajewski, 1999). Several studies have avoided the ground truth uncertainty by comparing multiple QPEs. For example, Bytheway (2022) created a probabilistic QPE by combining multiple high resolution precipitation estimates, including MRMS and forecasts. As satellite data and forecasts increase in temporal resolution this type of probabilistic precipitation estimate might be possible at sub-hourly resolutions.

5 Conclusion

In mountainous regions, high-resolution data is often lacking but necessary to quantify the unique hydrometeorology and its related hazards. In this study, we evaluate MRMS sub-hourly precipitation estimates in the mountains of Colorado to understand general trends in performance. We also evaluate the physical attributes that are most important in determining error and the characteristics that are associated with various error values. The performance and feature evaluation are enhanced by a novel method for predicting radar-based QPE error that can be used to predict error where gages are sparse. The spatial patterns of performance generally align with RQI; however, gage comparisons and model predictions show that the error is highly variable throughout the state. The feature importance and predictions show that duration of rainfall is a useful indicator of the performance of MRMS.

Given a sub-hourly and spatially continuous reference dataset, a logical next step would be to improve the QPE accuracy and reliability, rather than estimate error. Recent studies have successfully improved satellite and radar based QPEs using more complex machine learning methods such as a convolutional neural network (CNN). CNNs excel in identifying spatial dependencies by scanning parts of an image, with matrix weights, rather than ingesting all data as does a fully connected network (Géron, 2021). Osborne et al. (2023) used a CNN with the radar-only MRMS QPE and other reflectivity parameters as input and a gage network as the target precipitation to obtain more accurate hourly precipitation. The CNN QPE showed some improvement over the radar-only QPE, a valuable result in operational forecasting where the latency of the multi-sensor QPE hinders its usefulness. Hilburn et al. (2021) used a CNN to create synthetic MRMS radar reflectivity images from GOES-R that preserves the spatial resolution of MRMS using a U-NET architecture. The approach outperformed traditional data assimilation techniques and has the potential to generate more precise reflectivity estimates from satellites in regions where ground radar blockage is notable, such as the western United States. Model architectures such as visual transformers may be used to leverage spatial as well as temporal context during learning. Our analysis shows that the error of MRMS is often high in several regions of Colorado, and likely unacceptable for many applications. It is worth exploring the potential of more advanced algorithms to enhance high resolution QPEs in areas of complex terrain.

Open Research Section

The statewide RMSE prediction results and features used to make predictions, along with the training dataset features and intensity time series, can be accessed at http://www.hydroshare.org /resource /95aa5dbcb9ab4345ae589b28d95582c2 (White & Nelson, 2024a). Code used for the development of the datasets and results in this study can be accessed at https://doi.org/ 10.5281/zenodo.10734150 (White & Nelson, 2024b). Three additional USGS gages used in the study cited in the text as Rengers et al. (2023) can be accessed directly at https://www.sciencebase.gov /catalog/item/63617bebd34ebe4425065664.

Acknowledgments

This research was funded by the Joint Fire Science Program project 22-1-01-21 and USDA National Institute of Food and Agriculture Hatch project COL00422. Authors acknowledge Matt Gorbett (Department of Computer Science, Colorado State University) for consultations on best practices in machine learning.

References

Appling, A. P., Oliver, S. K., Read, J. S., Sadler, J. M., & Zwart, J. A. (2022). Machine Learning for Understanding Inland Water Quantity, Quality, and Ecology. In *Encyclopedia of Inland Waters* (pp. 585–606). Elsevier. Retrieved 2022-10-25, from `https://linkinghub.elsevier.com/retrieve/pii/B9780128191668001213` doi: 10.1016/B978-0-12-819166-8.00121-3

Berne, A., & Krajewski, W. (2013, January). Radar for hydrology: Unfulfilled promise or unrecognized potential? *Advances in Water Resources*, *51*, 357–366. Retrieved 2022-09-04, from `https://linkinghub.elsevier.com/retrieve/pii/S0309170812001157` doi: 10.1016/j.advwatres.2012.05.005

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. *Wadsworth, Belmont, CA*, *37*(15), 237–251.

Bytheway, J. L. (2022, January). Demonstrating a Probabilistic Quantitative Precipitation Estimate for Evaluating Precipitation Forecasts in Complex Terrain. *Weather and Forecasting*, *37*(1), 45–64. Retrieved 2022-09-21, from `https://journals.ametsoc.org/view/journals/wefo/37/1/WAF-D-21-0074.1.xml` doi: 10.1175/WAF-D-21-0074.1

Bytheway, J. L., Hughes, M., Mahoney, K., & Cifelli, R. (2019, March). A Multiscale Evaluation of Multisensor Quantitative Precipitation Estimates in the Russian River Basin. *Journal of Hydrometeorology*, *20*(3), 447–466. Retrieved 2022-07-13, from `http://journals.ametsoc.org/doi/10.1175/JHM-D-18-0142.1` doi: 10.1175/JHM-D-18-0142.1

Bytheway, J. L., Hughes, M., Mahoney, K., & Cifelli, R. (2020, May). On the Uncertainty of High-Resolution Hourly Quantitative Precipitation Estimates in California. *Journal of Hydrometeorology*, *21*(5), 865–879. Retrieved 2022-07-18, from `https://journals.ametsoc.org/view/journals/hydr/21/5/jhm-d-19-0160.1.xml` doi: 10.1175/JHM-D-19-0160.1

Cannon, S. H., Gartner, J. E., Wilson, R. C., Bowers, J. C., & Laber, J. L. (2008, April). Storm rainfall conditions for floods and debris flows from recently burned areas in southwestern Colorado and southern California. *Geomorphology*, *96*(3-4), 250–269. Retrieved 2021-12-15, from `https://linkinghub.elsevier.com/retrieve/pii/S0169555X07001778` doi: 10.1016/j.geomorph.2007.03.019

Ciach, G. J., & Krajewski, W. F. (1999, February). On the estimation of radar rainfall error variance. *Advances in Water Resources*, *22*(6), 585–595. Retrieved 2022-06-16, from `https://linkinghub.elsevier.com/retrieve/pii/`

`S0309170898000438` doi: 10.1016/S0309-1708(98)00043-8

Ciach, G. J., Krajewski, W. F., & Villarini, G. (2007, December). Product-Error-Driven Uncertainty Model for Probabilistic Quantitative Precipitation Estimation with NEXRAD Data. *Journal of Hydrometeorology*, *8*(6), 1325–1347. Retrieved 2023-01-11, from `http://journals.ametsoc.org/doi/10.1175/2007JHM814.1` doi: 10.1175/2007JHM814.1

Colorado State University, C. (2023). *Colorado agricultural meteorological network.* Retrieved from `https://coagmet.colostate.edu/`

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27.

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., . . . Pasteris, P. P. (2008, December). Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology*, *28*(15), 2031–2064. Retrieved 2024-01-19, from `https://rmets.onlinelibrary.wiley.com/doi/10.1002/joc.1688` doi: 10.1002/joc.1688

Derin, Y., Anagnostou, E., Berne, A., Borga, M., Boudevillain, B., Buytaert, W., . . . Yilmaz, K. K. (2016, June). Multiregional Satellite Precipitation Products Evaluation over Complex Terrain. *Journal of Hydrometeorology*, *17*(6), 1817–1836. Retrieved 2024-03-01, from `http://journals.ametsoc.org/doi/10.1175/JHM-D-15-0197.1` doi: 10.1175/JHM-D-15-0197.1

Doswell, C. A., Brooks, H. E., & Maddox, R. A. (1996, December). Flash Flood Forecasting: An Ingredients-Based Methodology. *Weather and Forecasting*, *11*(4), 560–581. Retrieved 2022-09-09, from `http://journals.ametsoc.org/doi/10.1175/1520-0434(1996)011<0560:FFFAIB>2.0.CO;2` doi: 10.1175/1520-0434(1996)011⟨0560:FFFAIB⟩2.0.CO;2

Esri. (2020). *ArcGIS Desktop (version 10.8.1).* Retrieved from `https://desktop.arcgis.com/en/`

Fabry, F., Bellon, A., Duncan, M. R., & Austin, G. L. (1994, September). High resolution rainfall measurements by radar for very small basins: the sampling problem reexamined. *Journal of Hydrology*, *161*(1-4), 415–428. Retrieved 2024-01-18, from `https://linkinghub.elsevier.com/retrieve/pii/0022169494901384` doi: 10.1016/0022-1694(94)90138-4

Force, N.-U. D. F. T. (2005). *Noaa-usgs debris-flow warning system - final report* (Report No. 1283). Reston, VA: USGS Publications Warehouse. Retrieved from `https://pubs.usgs.gov/publication/cir1283` (Accessed: January 19, 2024)

Fowler, H. J., Lenderink, G., Prein, A. F., Westra, S., Allan, R. P., Ban, N., . . . Zhang, X. (2021, January). Anthropogenic intensification of short-duration rainfall extremes. *Nature Reviews Earth & Environment*, *2*(2), 107–122. Retrieved 2023-05-02, from `https://www.nature.com/articles/s43017-020-00128-6` doi: 10.1038/s43017-020-00128-6

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference*, *96*, 148-156.

Friedman, J. H. (2001, October). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). Retrieved 2024-01-19, from `https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full` doi: 10.1214/aos/1013203451

Géron, A. (2021). *Hands-on machine learning with scikit-learn, keras, and tensorflow: Concepts, tools, and techniques to build intelligent systems* (3rd ed.). O'Reilly Media.

Henn, B., Newman, A. J., Livneh, B., Daly, C., & Lundquist, J. D. (2018, January). An assessment of differences in gridded precipitation datasets in complex terrain. *Journal of Hydrology*, *556*, 1205–1219. Retrieved 2022-07-20, from `https://linkinghub.elsevier.com/retrieve/pii/S0022169417301452` doi: 10.1016/j.jhydrol.2017.03.008

Higuera, P. E., Shuman, B. N., & Wolf, K. D. (2021, June). Rocky Mountain subalpine forests now burning more than any time in recent millennia. *Proceedings of the National Academy of Sciences*, *118*(25), e2103135118. Retrieved 2022-02-12, from `http://www.pnas.org/lookup/doi/10.1073/pnas.2103135118` doi: 10.1073/pnas.2103135118

Hilburn, K. A., Ebert-Uphoff, I., & Miller, S. D. (2021, January). Development and Interpretation of a Neural-Network-Based Synthetic Radar Reflectivity Estimator Using GOES-R Satellite Observations. *Journal of Applied Meteorology and Climatology*, *60*(1), 3–21. Retrieved 2023-11-09, from `https://journals.ametsoc.org/view/journals/apme/60/1/jamc-d-20-0084.1.xml` doi: 10.1175/JAMC-D-20-0084.1

Kirstetter, P.-E., Gourley, J. J., Hong, Y., Zhang, J., Moazamigoodarzi, S., Langston, C., & Arthur, A. (2015). Probabilistic precipitation rate estimates with ground-based radar networks. *Water Resources Research*, *51*(3), 1422-1442. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014WR015672` doi: https://doi.org/10.1002/2014WR015672

Kirstetter, P.-E., Hong, Y., Gourley, J. J., Chen, S., Flamig, Z., Zhang, J., . . . Amitai, E. (2012, August). Toward a Framework for Systematic Error Modeling of Spaceborne Precipitation Radar with NOAA/NSSL Ground Radar–Based National Mosaic QPE. *Journal of Hydrometeorology*, *13*(4), 1285–1300. Retrieved 2023-04-22, from `http://journals.ametsoc.org/doi/10.1175/JHM-D-11-0139.1` doi: 10.1175/JHM-D-11-0139.1

Liao, M., & Barros, A. P. (2023, January). *Toward Optimal Rainfall for Flood Prediction in Headwater Basins - Orographic QPE error modeling using machine learning* (preprint). Preprints. Retrieved 2023-04-22, from `https://essopenarchive.org/users/565534/articles/618619-toward-optimal-rainfall-for-flood-prediction-in-headwater-basins-orographic-qpe-error-modeling-using-machine-learning?commit=9df7d4760888277f1bbecb80a295903ae56f27fe` doi: 10.22541/essoar.167390516.60866247/v1

Lundquist, J., Hughes, M., Gutmann, E., & Kapnick, S. (2019, December). Our Skill in Modeling Mountain Rain and Snow is Bypassing the Skill of Our Observational Networks. *Bulletin of the American Meteorological Society*, *100*(12), 2473–2490. Retrieved 2023-06-08, from `https://journals.ametsoc.org/view/journals/bams/100/12/bams-d-19-0001.1.xml` doi: 10.1175/BAMS-D-19-0001.1

Maddox, R. A., Zhang, J., Gourley, J. J., & Howard, K. W. (2002, August). Weather Radar Coverage over the Contiguous United States. *Weather and Forecasting*, *17*(4), 927–934. Retrieved 2023-11-15, from `http://journals.ametsoc.org/doi/10.1175/1520-0434(2002)017<0927:WRCOTC>2.0.CO;2` doi: 10.1175/1520-0434(2002)017⟨0927:WRCOTC⟩2.0.CO;2

Mahoney, K. M., Ralph, F. M., Wolter, K., Doesken, N. J., Dettinger, M. D., Gottas, D. J., . . . White, A. B. (2015, April). Climatology of Extreme Daily Precipitation in Colorado and Its Diverse Spatial and Seasonal Variability. *Journal of Hydrometeorology*, *16*(2), 781–792. (MAG ID: 2096033735 S2ID: 627aaa8a082f843e6cc7199ff5759366fa02e08c) doi: 10.1175/jhm-d-14-0112.1

Mesonet, I. E. (n.d.). *Mtarchive daily selected files.* Retrieved from `https://mtarchive.geol.iastate.edu/`

Mesowest - Weather and Climate Data. (Accessed 2023). Mesowest. Retrieved from `{https://mesowest.utah.edu/cgi-bin/droman/mesomap.cgi?state=CO}`

Moazami, S., & Najafi, M. (2021, March). A comprehensive evaluation of GPM-IMERG V06 and MRMS with hourly ground-based precipitation observations across Canada. *Journal of Hydrology*, *594*, 125929. Retrieved 2022-07-13, from `https://linkinghub.elsevier.com/retrieve/pii/S0022169420313901` doi: 10.1016/j.jhydrol.2020.125929

Moody, J. A., Shakesby, R. A., Robichaud, P. R., Cannon, S. H., & Martin, D. A. (2013, July). Current research issues related to post-wildfire runoff and erosion processes. *Earth-Science Reviews*, *122*, 10–37. Retrieved 2021-10-10, from `https://linkinghub.elsevier.com/retrieve/pii/S0012825213000536` doi: 10.1016/j.earscirev.2013.03.004

Moreno, H. A., Vivoni, E. R., & Gochis, D. J. (2012, May). Utility of Quantitative Precipitation Estimates for high resolution hydrologic forecasts in mountain watersheds of the Colorado Front Range. *Journal of Hydrology*, *438-439*, 66–83. Retrieved 2022-09-21, from `https://linkinghub.elsevier.com/retrieve/pii/S002216941200217X` doi: 10.1016/j.jhydrol.2012.03.019

NASA Jet Propulsion Laboratory. (2022). *NASA Shuttle Radar Topography Mission 1-arc second Global Land Data.* NASA LP DAAC. (`https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-shuttle-radar-topography-mission-srtm-1`)

Nicótina, L., Alessi Celegon, E., Rinaldo, A., & Marani, M. (2008, December). On the impact of rainfall patterns on the hydrologic response: RAINFALL PATTERNS AND HYDROLOGIC RESPONSE. *Water Resources Research*, *44*(12). Retrieved 2022-09-06, from `http://doi.wiley.com/10.1029/2007WR006654` doi: 10.1029/2007WR006654

NOAA, W. D. T. D. W. (n.d.). *Virtual lab.* `https://vlab.noaa.gov/web/wdtd/`. (Accessed: January 19, 2024)

NOAA/CPC. (2023). *Wgrib2.* `https://www.cpc.ncep.noaa.gov/products/wesley/wgrib2/`.

Osborne, A. P., Zhang, J., Simpson, M. J., Howard, K. W., & Cocks, S. B. (2023, April). Application of Machine Learning Techniques to Improve Multi-Radar Multi-Sensor (MRMS) Precipitation Estimates in the Western United States. *Artificial Intelligence for the Earth Systems*, *2*(2), 220053. Retrieved 2024-01-11, from `https://journals.ametsoc.org/view/journals/aies/2/2/AIES-D-22-0053.1.xml` doi: 10.1175/AIES-D-22-0053.1

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rengers, F., Bower, S., Knapp, A., Kean, J., & Staley, D. (2023, October). *Debris flow, precipitation, and volume measurements in the grizzly creek burn perimeter june 2021-september 2022, glenwood canyon, colorado.* U.S. Geological Survey data release. Retrieved from `https://doi.org/10.5066/P9Z7RROL` doi: 10.5066/P9Z7RROL

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., ... Dormann, C. F. (2017, August). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. Retrieved 2023-09-15, from `https://onlinelibrary.wiley.com/doi/10.1111/ecog.02881` doi: 10.1111/ecog.02881

Seo, B.-C., & Krajewski, W. F. (2010, October). Scale Dependence of Radar Rainfall Uncertainty: Initial Evaluation of NEXRAD's New Super-Resolution Data for Hydrologic Applications. *Journal of Hydrometeorology*, *11*(5), 1191–1198. Retrieved 2022-09-21, from `http://journals.ametsoc.org/doi/10.1175/2010JHM1265.1` doi: 10.1175/2010JHM1265.1

Seo, B.-C., & Krajewski, W. F. (2015, December). Correcting temporal sampling error in radar-rainfall: Effect of advection parameters and rain storm characteristics on the correction accuracy. *Journal of Hydrology*, *531*, 272–283. Retrieved 2022-08-09, from `https://linkinghub.elsevier.com/retrieve/pii/S002216941500267X` doi: 10.1016/j.jhydrol.2015.04.018

Sevruk, B. (2005, October). Rainfall Measurement: Gauges. In M. G. Anderson & J. J. McDonnell (Eds.), *Encyclopedia of Hydrological Sciences* (p. hsa038). Chichester, UK: John Wiley & Sons, Ltd. Retrieved 2023-05-17, from `https://onlinelibrary.wiley.com/doi/10.1002/0470848944.hsa038` doi: 10.1002/0470848944.hsa038

Sieck, L. C., Burges, S. J., & Steiner, M. (2007, January). Challenges in obtaining reliable measurements of point rainfall: RELIABLE MEASUREMENTS OF POINT RAINFALL. *Water Resources Research*, *43*(1). Retrieved 2023-05-17, from `http://doi.wiley.com/10.1029/2005WR004519` doi: 10.1029/2005WR004519

Smith, J. A., Baeck, M. L., Meierdiercks, K. L., Miller, A. J., & Krajewski, W. F. (2007, October). Radar rainfall estimation for flash flood forecasting in small urban watersheds. *Advances in Water Resources*, *30*(10), 2087–2097. Retrieved 2021-12-16, from `https://linkinghub.elsevier.com/retrieve/pii/S0309170807000553` doi: 10.1016/j.advwatres.2006.09.007

Smith, R. B. (2019, January). 100 Years of Progress on Mountain Meteorology Research. *Meteorological Monographs*, *59*, 20.1–20.73. Retrieved 2022-09-03, from `http://journals.ametsoc.org/doi/10.1175/AMSMONOGRAPHS-D-18-0022.1` doi: 10.1175/AMSMONOGRAPHS-D-18-0022.1

Sokol, Z., Szturc, J., Orellana-Alvear, J., Popová, J., Jurczyk, A., & Célleri, R. (2021, January). The Role of Weather Radar in Rainfall Estimation and Its Application in Meteorological and Hydrological Modelling—A Review. *Remote Sensing*, *13*(3), 351. Retrieved 2022-09-06, from `https://www.mdpi.com/2072-4292/13/3/351` doi: 10.3390/rs13030351

Sun, L., Chen, H., Li, Z., & Han, L. (2021, October). Cross Validation of GOES-16 and NOAA Multi-Radar Multi-Sensor (MRMS) QPE over the Continental United States. *Remote Sensing*, *13*(20), 4030. Retrieved 2022-12-15, from `https://www.mdpi.com/2072-4292/13/20/4030` doi: 10.3390/rs13204030

Syed, K. H., Goodrich, D. C., Myers, D. E., & Sorooshian, S. (2003, February). Spatial characteristics of thunderstorm rainfall fields and their relation to runoff. *Journal of Hydrology*, *271*(1-4), 1–21. Retrieved 2023-05-16, from `https://linkinghub.elsevier.com/retrieve/pii/S0022169402003116` doi: 10.1016/S0022-1694(02)00311-6

Tang, G., Behrangi, A., Long, D., Li, C., & Hong, Y. (2018, April). Accounting for spatiotemporal errors of gauges: A critical step to evaluate gridded precipitation products. *Journal of Hydrology*, *559*, 294–306. (MAG ID: 2793745286 S2ID: 2ccc8ef53cbcede6c79d54ae70ce548c2e959e7b) doi: 10.1016/j.jhydrol.2018.02.057

Touma, D., Stevenson, S., Swain, D. L., Singh, D., Kalashnikov, D. A., & Huang, X. (2022, April). Climate change increases risk of extreme rainfall following wildfire in the western United States. *Science Advances*, *8*(13), eabm0320. Retrieved 2022-05-12, from `https://www.science.org/doi/10.1126/sciadv.abm0320` doi: 10.1126/sciadv.abm0320

Usgs national water information system. (2023). Retrieved from `https://nwis.waterservices.usgs.gov/`

van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., … the scikit-image contributors (2014, 6). scikit-image: image processing in Python. *PeerJ*, *2*, e453. Retrieved from `https://doi.org/10.7717/peerj.453` doi: 10.7717/peerj.453

Villarini, G., Mandapaka, P. V., Krajewski, W. F., & Moore, R. J. (2008, June). Rainfall and sampling uncertainties: A rain gauge perspective. *Journal of Geophysical Research*, *113*(D11), D11102. Retrieved 2022-07-13, from `http://doi.wiley.com/10.1029/2007JD009214` doi: 10.1029/2007JD009214

Villarini, G., Seo, B.-C., Serinaldi, F., & Krajewski, W. F. (2014, January). Spatial and temporal modeling of radar rainfall uncertainties. *Atmo-*

spheric Research, *135-136*, 91–101. Retrieved 2022-08-09, from `https://linkinghub.elsevier.com/retrieve/pii/S0169809513002482` doi: 10.1016/j.atmosres.2013.09.007

White, P., & Nelson, P. (2024a). *Evaluation of sub-hourly mrms quantitative precipitation estimates in colorado's mountains using machine learning.* HydroShare. Retrieved from `http://www.hydroshare.org/resource/95aa5dbcb9ab4345ae589b28d95582c2`

White, P., & Nelson, P. (2024b, February). *psylw/mrms-eval-with-gages-in-co: submit.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.10734150` doi: 10.5281/zenodo.10734150

Zhang, J., Howard, K., Langston, C., Kaney, B., Qi, Y., Tang, L., … Kitzmiller, D. (2016, April). Multi-Radar Multi-Sensor (MRMS) Quantitative Precipitation Estimation: Initial Operating Capabilities. *Bulletin of the American Meteorological Society*, *97*(4), 621–638. Retrieved 2022-07-13, from `https://journals.ametsoc.org/doi/10.1175/BAMS-D-14-00174.1` doi: 10.1175/BAMS-D-14-00174.1

Zhang, J., Qi, Y., Langston, C., & Kaney, B. (2012). Radar quality index (rqi) – a combined measure for beam blockage and vpr effects in a national network. *Weather Radar and Hydrology*, *351*, 388–393.

Zhang, J., Qi, Y., Langston, C., Kaney, B., & Howard, K. (2014, October). A Real-Time Algorithm for Merging Radar QPEs with Rain Gauge Observations and Orographic Precipitation Climatology. *Journal of Hydrometeorology*, *15*(5), 1794–1809. Retrieved 2021-11-28, from `http://journals.ametsoc.org/doi/10.1175/JHM-D-13-0163.1` doi: 10.1175/JHM-D-13-0163.1

Zhu, Z., Wright, D. B., & Yu, G. (2018). The Impact of Rainfall Space-Time Structure in Flood Frequency Analysis. *Water Resources Research*, *54*(11), 8983–8998. Retrieved 2024-01-19, from `https://onlinelibrary.wiley.com/doi/abs/10.1029/2018WR023550` (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2018WR023550) doi: 10.1029/2018WR023550