nature communications

Article

https://doi.org/10.1038/s41467-025-58234-2

Machine learning shows a limit to rain-snow partitioning accuracy when using near-surface meteorology

Received: 21 June 2024

Accepted: 17 March 2025

Published online: 25 March 2025

Check for updates

Keith S. Jennings[1,2] ✉, Meghan Collins[3], Benjamin J. Hatchett[3,4], Anne Heggli[3], Nayoung Hur[2], Sonia Tonino[3], Anne W. Nolin[5], Guo Yu[3], Wei Zhang[6] & Monica M. Arienzo[3]

Partitioning precipitation into rain and snow with near-surface meteorology is a well-known challenge. However, whether a limit exists to its potential performance remains unknown. Here, we evaluate this possibility by applying a set of benchmark precipitation phase partitioning methods plus three machine learning (ML) models (an artificial neural network, random forest, and XGBoost) to two independent datasets: 38.5 thousand crowdsourced observations and 17.8 million synoptic meteorology reports. The ML methods provide negligible improvements over the best benchmarks, increasing accuracy only by up to 0.6% and reducing rain and snow biases by up to -4.7%. ML methods fail to identify mixed precipitation and sub-freezing rainfall events, while expressing their worst accuracy values from 1.0 °C–2.5 °C. A potential cause of these shortcomings is the air temperature overlap in rain and snow distributions (peaking between 1.0 °C–1.6 °C), which expresses a significant negative relationship ($p < 0.0005$) with partitioning accuracy. Thus, the meteorological characteristics of rain and snow are similar at air temperatures slightly above freezing with increasing overlap associated with decreasing performance. We suggest researchers switch their focus from marginally improving inherently limited precipitation phase partitioning methods using near-surface meteorology to creating new methods that assimilate novel data sources—e.g., crowdsourced precipitation phase observations.

Direct observations of precipitation phase are rare, necessitating the use of ancillary meteorological data—air temperature usually—to determine whether rain, snow, or mixed precipitation is falling at the land surface[1,2]. Precipitation phase partitioning should be relatively simple. The transition temperature between water's solid and liquid phases is 0 °C, suggesting a physically based threshold for partitioning precipitation into rain and snow. Warmer than 0 °C? Liquid. Cooler than 0 °C? Solid. Unfortunately, this simple assumption is often wrong. In fact, precipitation at the land surface is predominantly snowfall at 0 °C, with rainfall only increasing in prominence as the air temperature warms towards 1 °C–3 °C[3–6]. Moreover, the air temperature threshold at which rain and snow fall with equal probability varies regionally, from near 0 °C in maritime locations to over 4 °C in some upland, continental areas[6].

[1]University of Vermont Water Resources Institute, 210 Colchester Ave, Burlington, VT, USA. [2]Lynker, Boulder, CO, USA. [3]Desert Research Institute, Reno, NV, USA. [4]Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO, USA. [5]University of Nevada, Reno, Reno, NV, USA. [6]Utah State University, Department of Plants, Soils & Climate, Logan, UT, USA. ✉e-mail: keith.jennings@uvm.edu

As a result of this variability, there are many methods that partition precipitation into rain and snow with near-surface meteorological data such as air temperature, humidity, and pressure. These methods include thresholds, ranges, and statistical models[1,2], and comprise an important component of land surface and hydrologic models. Large-scale reanalysis products, such as the Global Precipitation Measurement (GPM) mission's Integrated Multi-satellitE Retrievals for GPM (IMERG) and Phase 2 of the North American Land Data Assimilation System (NLDAS-2), combine gridded meteorological data with an empirical phase partitioning method[7]. Most of these methods that use near-surface meteorology perform well at cold air temperatures below −5 °C and warm air temperatures above 10 °C when snow and rain, respectively, are consistently predictable. However, at air temperatures slightly above freezing (0 °C–4 °C), the best these methods can do is to properly identify precipitation phase ~65% of the time[6,8]. Model-based errors and uncertainties in rain-snow partitioning[9] then propagate into inaccurate and uncertain simulations of snow accumulation, snowmelt, and streamflow[10–12].

These challenges have motivated researchers to improve the techniques that use near-surface meteorology to partition the precipitation phase. We refer readers to the relevant review papers[1,2] for a full discussion of this work, but we cover some major themes here. The 1956 US Army Corps of Engineers[3] Snow Hydrology report was the first document we know of to show the prevalence of snowfall above 0 °C, a finding confirmed by Auer[4]. Subsequent studies over larger study domains demonstrated the spatial variability of the optimal air temperature threshold for partitioning rain and snow[13–15]. Other researchers have noted that the use of air temperature ranges[16,17] may improve partitioning performance relative to single-threshold methods, particularly over longer time scales[18]. The potential advantage here is that range-based methods use a lower threshold for snow and a higher threshold for rain, with mixed precipitation occurring between. Subsequent work at shorter time scales, however, has called this finding into question[9]. More recently, several papers have shown that including humidity in some manner—typically in the form of wet bulb or dew point temperature—aids in the correct identification of precipitation phase[18–20].

Comparative studies[6,9,12,21,22], meanwhile, have arrived at some general conclusions on partitioning rain, snow, and mixed precipitation using near-surface meteorology. For one, the inclusion of humidity improves predictive performance relative to methods that use only air temperature, with wet bulb temperature thresholds typically outperforming dew point temperature thresholds. Similarly, statistical models that use both air temperature and relative humidity, when optimized on observed data, can provide highly accurate estimates of precipitation phase. However, even when studies use the best available data and methods, shortcomings remain. While some methods exhibit low seasonal rain and snow biases, all methods present their worst accuracies between approximately 0 °C and 4 °C[8]. As of now, all studied methods relying on near-surface meteorology produce this dip in performance.

The lack of a skillful approach to estimating the precipitation phase worsens our representations of hydrologic processes in cold regions, which affects the populations and ecosystems dependent on snow-derived water resources[23–25]. These regions must balance managing snow as a resource (water supply) and as a source of hazards (flooding and mass movements) largely through investments in infrastructure[23,26]. The stress on natural resource managers to balance the risks to life, property, and ecosystem function between drought and flood will intensify as the climate continues to change[27] with projected increases in rain-on-snow magnitude and frequency[28] occurring prior to the ultimate warming-induced loss of snow[25]. This transition will be driven in part by increasingly warm temperatures on more frequent dry days[29] and a rise in snow levels during storms as precipitation increasingly falls as rain[26,30,31].

This produces an unsatisfying scenario where a seemingly simple problem has myriad societal implications yet no straightforward solution. Why is that? In previous work, we noted a concerning phenomenon: whether we were evaluating air, wet bulb, or dew point temperature, we found a 43.8–52.8% overlap in the distributions of rain, snow, and mixed precipitation[9]. We suggested that this overlap may act as a limiting factor in the performance of precipitation phase partitioning methods that use near-surface meteorology. Similar to other comparative studies, we only explored methods that used simple thresholds and ranges along with an optimized statistical model. Recent advances in machine learning in the hydrologic[32,33] and atmospheric sciences[34,35], however, have demonstrated the power of using data-driven approaches instead of prescribing a priori assumptions and techniques.

A few researchers have deployed machine learning methods, such as random forest[36] and artificial neural networks (ANN), to see if such techniques can improve precipitation phase partitioning performance. For example, Campozano et al. [37] showed that random forest models of the precipitation phase better predict rain and snow than a logistic regression model according to some metrics and data availability scenarios. They also found the tree-based random forest method often outperformed the more complex ANN. In addition, they noted that a relatively parsimonious set of predictors—air temperature, dew point temperature, and specific humidity—account for much of the improved performance[37]. Similarly, Filipiak et al. [35] found that air and dew point temperature were the two most important variables in predicting the precipitation phase with a random forest model. Both studies, however, focused on small, relatively climatologically homogenous regions, and neither study showed performance by air temperature to evaluate whether machine learning can avoid or mitigate the dip in performance that all other methods display when using near-surface meteorology. Thus, there is a need to explore more climatologically diverse study domains with large ranges in rain-snow air temperature thresholds and to evaluate machine learning performance by air temperature relative to benchmark partitioning methods.

There is a bigger, more worrying concern past those two considerations, though. What if machine learning techniques do not markedly improve precipitation phase prediction accuracy while minimizing rain and snow biases, particularly at air temperatures near freezing? Or, more simply put, is there a limit to rain-snow partitioning performance using near-surface meteorology? To answer these questions we employ two unique, independent datasets of precipitation phase and near-surface meteorology: crowdsourced observations of rain, snow, and mixed precipitation[9,38,39] and synoptic meteorology reports of rain and snow[5,6], which we refer to as the crowdsourced and synoptic datasets, respectively. We apply a selection of the highest performing rain-snow partitioning algorithms[6] as benchmark methods, to which we compare two tree-based machine learning techniques —random forest[36] and XGBoost[40]—as well as an ANN with one hidden layer. We also evaluate two more advanced machine learning approaches—an ANN with two hidden layers and a stacked ensemble model—on the crowdsourced dataset to determine if increasing complexity produces better outcomes. We close by evaluating the meteorological properties of rain and snow in the two datasets and relating these characteristics to the accuracy of the benchmark and machine learning phase partitioning methods.

Results

Benchmark precipitation phase partitioning performance

The benchmark methods (Table 1) exhibited variable performance in predicting rain and snow in the crowdsourced and synoptic datasets (Table 2). The methods incorporating humidity (e.g., the binary logistic regression model and the wet bulb and dew point temperature thresholds) generally performed better than the methods using only air temperature, with a median accuracy value across both datasets of

Table 1 | Information on the temperature thresholds and the binary logistic regression model that we use as benchmark methods in this study

Method Type	Met. Data	Value (°C)	Symbol	Note or Source
Threshold	T_a	1.0	$T_{a1.0}$	Performant values from large-scale study[6]
		1.5	$T_{a1.5}$	
	T_w	0.0	$T_{w0.0}$	
		0.5	$T_{w0.5}$	
		1.0	$T_{w1.0}$	
	T_d	0.0	$T_{d0.0}$	
		0.5	$T_{d0.5}$	
Binary Logistic Regression	T_a and RH	NA	Bin_{log}	Model form[21] with optimized coefficients[6]

Meteorological (met.) data include air (T_a), wet bulb (T_w), and dew point (T_d) temperature and relative humidity (RH).

Table 2 | Performance metrics for the benchmark precipitation partitioning methods (PPMs) using the crowdsourced and synoptic datasets

Dataset	PPM	Accuracy (%)	Snow bias (%)	Rain bias (%)
Crowdsourced	$T_{w1.0}$	88.7	4.5	−10.2
	$T_{w0.5}$	88.5	−3.3	7.4
	$T_{d0.5}$	87.7	7.9	−17.9
	$T_{d0.0}$	87.3	0.5	−1.1
	Bin_{log}	86.3	−8.6	19.5
	$T_{w0.0}$	86.1	−11.4	25.7
	$T_{a1.5}$	83	−13.2	29.7
	$T_{a1.0}$	80.5	−20.6	46.5
Synoptic	$T_{w0.5}$	93.1	6.2	−5.5
	Bin_{log}	93.1	5.6	−5
	$T_{w0.0}$	92.2	−2.7	2.3
	$T_{a1.0}$	92.1	4.5	−4
	$T_{a1.5}$	91.9	8.2	−7.2
	$T_{w1.0}$	91.8	14.1	−12.5
	$T_{d0.0}$	90.9	9.8	−8.7
	$T_{d0.5}$	90.7	13.5	−11.9

The rows are ordered by accuracy—from high to low—for each dataset. For more details on the PPMs, please refer to Table 1.

87.5% for the former compared to 81.8% for the latter. With a median rain bias value of 0.6%, the humidity-based methods performed markedly better in both datasets than the air-temperature-only methods (38.1%). The air-temperature-only methods also had a slightly worse median snow bias in both datasets than the humidity methods at -8.8% versus 5.9%.

When examining the two datasets separately, we found a few key differences (Table 2). Accuracy was generally lower in the crowdsourced dataset, ranging from 80.5% to 88.7%, a total spread of 8.2%. That is larger than the accuracy spread of the synoptic dataset (2.4%), which had a minimum value of 90.7% and a maximum of 93.1%. In the crowdsourced dataset, the top six methods, ranked by accuracy, all used humidity, while the bottom two relied on air temperature alone. The case was slightly different for the synoptic dataset where the two dew point temperature thresholds performed worse than the two static air temperature thresholds. Snow and rain biases had larger ranges than accuracy in both datasets. Snow bias varied by 28.5% and 16.8%, while rain bias varied by 64.4% and 14.8% in the crowdsourced and synoptic datasets, respectively. In general, the benchmarks performed better and with less variability when applied to the synoptic dataset, and methods with better accuracy values typically had lower rain and snow biases.

All of the benchmark methods performed worse at air temperatures near and slightly above the freezing point in terms of accuracy, with most of them reaching their minimum accuracy values between 1.5 °C and 2.0 °C in the crowdsourced dataset and between 0.5 °C and 1.5 °C in the synoptic dataset (Fig. 1). The best-performing benchmarks in terms of accuracy were $T_{w1.0}$ and $T_{w0.5}$ in the crowdsourced and synoptic datasets, respectively. For the former, $T_{w1.0}$ hit a minimum accuracy of 66.3% at 2 °C, while the latter reached a minimum accuracy of 68.7% at 1°C (Fig. 1a). In both datasets, most methods had slightly positive snow biases and markedly negative rain biases at air temperatures less than 1.0 °C (Fig. 1b,c). This is consistent with their limited, or complete lack of, ability to predict rain at sub-freezing temperatures. Conversely, most methods, with the exception of the dew point temperature thresholds, had largely negative snow biases and slightly positive rain biases at air temperatures above 2°C (Fig. 1b,c). Notably, the crowdsourced dataset expressed larger positive rain biases at higher temperatures (Fig. 1c), indicating the benchmark methods failed to capture snowfall occurring during warm near-surface conditions.

Machine learning precipitation phase partitioning performance

The three machine learning methods, an ANN with one hidden layer, random forest, and XGBoost, generally provided small performance gains relative to the best benchmark methods and larger, more consistent improvements relative to the average benchmark results (Table 3). In the crowdsourced dataset, ANN (89.2%) and XGBoost (88.8%) had slightly higher accuracy values than $T_{w1.0}$ (88.7%), while random forest had a slightly lower accuracy (88.3%) than the $T_{w1.0}$ threshold, which was the best-performing method from the benchmark exercise. Both XGBoost (93.3%) and random forest (93.7%) provided marginal improvements relative to $T_{w0.5}$ (93.1%), the best benchmark in the synoptic dataset as measured by the accuracy metric. ANN, however, expressed a slightly worse accuracy value (92.8%) than the best benchmark.

The three machine learning methods exhibited similar accuracy patterns by air temperature as the best benchmark and the average benchmark results, with performance degrading at air temperatures slightly above freezing (Supplementary Fig. 1). Both random forest and XGBoost achieved their minimum success rates of 66% and 68.6%, respectively, at 2 °C in the crowdsourced dataset. Meanwhile, ANN had its worst accuracy value of 70% at 2.5°C. Performance was higher in the synoptic dataset for random forest and XGBoost with minimum accuracy values of 74.2% and 70%, respectively. ANN, conversely, expressed slightly worse performance than in the crowdsourced dataset with a minimum accuracy of 67.1%. All minima occurred at 1 °C in the synoptic dataset. Relative improvements in accuracy compared to the best and average benchmarks varied by temperature for the machine learning methods (Fig. 2). In the crowdsourced dataset, relative improvements reached a maximum of 7.9% at 7.5°C compared to the best benchmark and a maximum of 19% at 1.5°C compared to the average benchmark values. In the synoptic dataset, relative improvements were similar for the best benchmark comparison, with a maximum increase of 8% at 1°C, and smaller for the average benchmark values, with a maximum increase of 14.7% at 1°C. Notably, ANN did not provide an accuracy improvement at any air temperature relative to the best benchmark in the synoptic dataset.

When compared to the snow and rain biases produced by the best benchmarks and average benchmark values, the machine learning methods generally provided larger absolute improvements than those recorded for accuracy (Table 3). While accuracy improvements were

https://doi.org/10.1038/s41467-025-58234-2

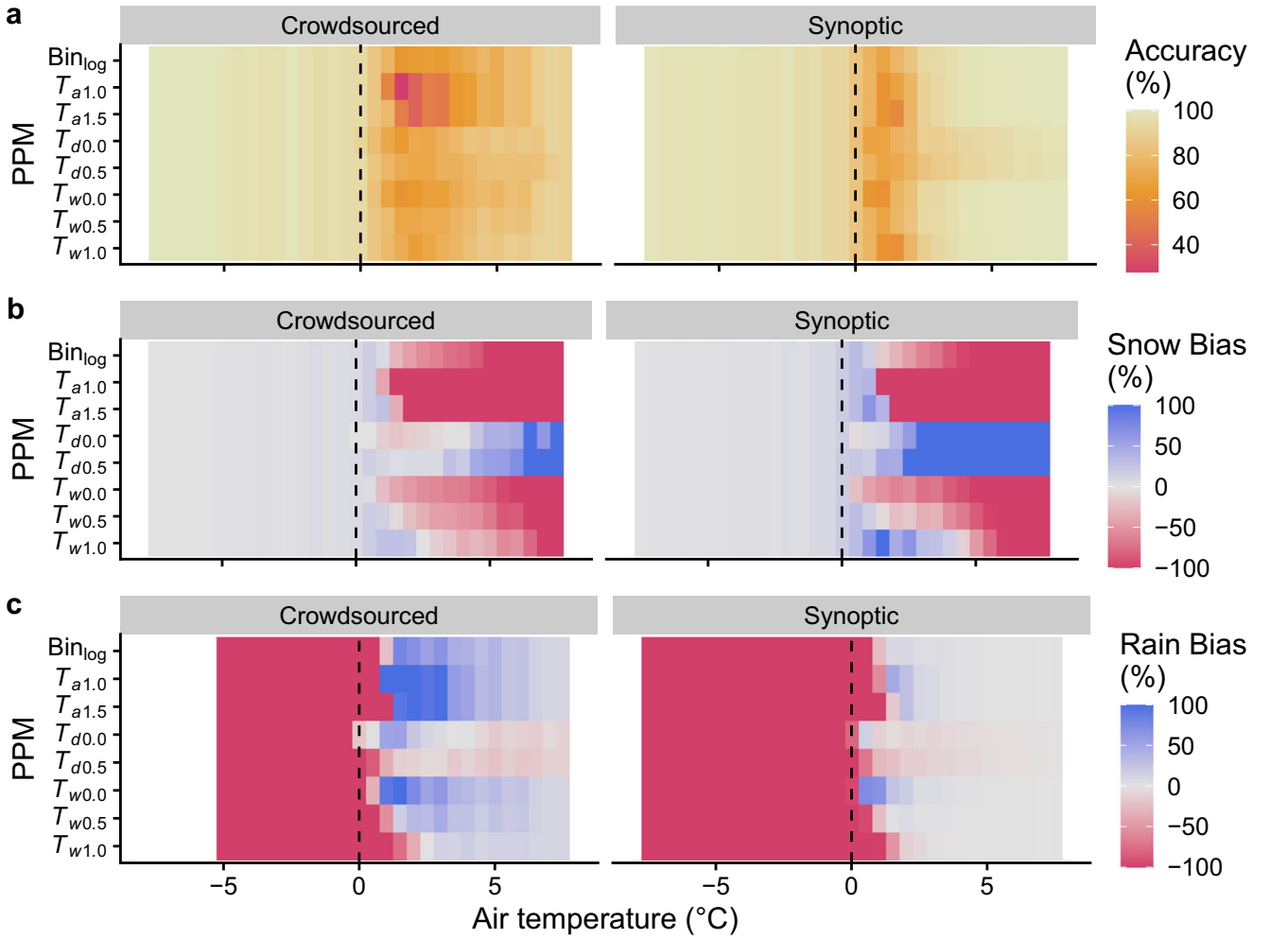

Fig. 1 | Performance metrics for each benchmark phase partitioning method (PPM). These metrics include accuracy (**a**), snow bias (**b**), and rain bias (**c**) plotted by air temperature for the crowdsourced and synoptic datasets. The vertical dashed line in each panel represents the 0 °C isotherm. Note: the −100% snow and rain bias values in (**b**) and (**c**), respectively, correspond to air temperatures where the method does not predict the given phase. For example, $T_{a1.0}$ predicts only rain above 1.0°C and thus presents a −100% bias for all observed snowfall above that threshold. Source data are provided as a Source Data file.

all less than or equal to 3.2%, the reduction in snow bias maxed out at −2% and −5.9% in the crowdsourced and synoptic datasets, respectively. Random forest, XGBoost, and ANN also provided consistent improvements in rain bias, with a maximum reduction of −4.5% in the crowdsourced dataset and -5.2% in the synoptic dataset. These results show that while accuracy improvements from the machine learning methods may be small, random forest, XGBoost, and ANN typically predict rain and snow with lower biases than the benchmark methods.

Mixed-phase precipitation partitioning performance

Until this point, we have only considered precipitation in its solid and liquid forms. While the synoptic dataset does not include any mixed precipitation observations, the testing split of the crowdsourced dataset includes 1140 mixed phase observations, comprising 11.8% of the total. As noted in the methods section, we do not include any dual-threshold benchmark methods because of their poor historical performance. The machine learning methods analyzed here fared a little better. Including the mixed observations caused overall accuracy to markedly decline, going from 88.3% (rain and snow only) to 77.5% (rain, snow, and mixed) for random forest, from 88.8% (rain and snow only) to 79.2% (rain, snow, and mixed) for XGBoost, and from 89.2% (rain and snow only) to 79.2% (rain, snow, and mixed) for ANN.

Although XGBoost and ANN had higher accuracy values than random forest when we included mixed precipitation, they achieved this result at the expense of bias. The random forest method predicted that 4.3% of the observations were mixed, which was well short of the observed value of 11.8%, giving it a mixed bias of −63.9%. XGBoost and ANN fared even worse, predicting 0% of the observations to be mixed, giving them both a mixed bias of −100%. When including mixed precipitation, random forest had a lower snow bias (12.2%) than XGBoost (16.1%) and ANN (17.1%). It was the same story for rain bias with XGBoost (7.3%) and ANN (5.1%) having worse biases than random forest (0.4%). These findings appear even worse when viewing the confusion matrix (Fig. 3), which shows that random forest correctly predicted only 9.3% of the observed mixed precipitation and incorrectly predicted the mixed observations as rain 32.9% of the time and as snow the remaining 57.8%. XGBoost and ANN failed completely here, correctly predicting none of the mixed precipitation observations, identifying them as snow over 60% of the time.

More complex machine learning methods

We found that applying more complex machine learning methods to the crowdsourced data did not consistently improve performance (Supplementary Table 1). Specifically, we evaluated an ANN with two hidden layers (ANN-2) and a stacked ensemble that combines the outcomes of

https://doi.org/10.1038/s41467-025-58234-2

Table 3 | Accuracy, rain bias, and snow bias for the three machine learning phase partitioning methods (PPMs): the artificial neural network (ANN), random forest (RF), and XGBoost (XG) applied to the two datasets

			Performance Metric for Machine Learning Models			Absolute Difference of Performance Metric to Benchmark Values		
Dataset	PPM	Benchmark Comparison	Accuracy (%)	Snow Bias (%)	Rain Bias (%)	Accuracy (%)	Snow Bias (%)	Rain Bias (%)
Crowdsourced	ANN	Best	89.2	3.8	−8.6	0.5	−0.7	−1.6
	RF	Best	88.3	3.5	−7.9	−0.4	−1	−2.3
	XG	Best	88.8	4.7	−10.6	0.1	0.2	0.4
	ANN	Average	89.2	3.8	−8.6	3.2	−1.7	−3.9
	RF	Average	88.3	3.5	−7.9	2.3	−2	−4.5
	XG	Average	88.8	4.7	−10.6	2.8	−0.8	−1.8
Synoptic	ANN	Best	92.8	1.5	−1.4	-0.3	−4.7	−4.1
	RF	Best	93.7	3.6	−3.2	0.6	-2.6	−2.3
	XG	Best	93.3	5.4	−4.8	0.1	-0.8	−0.7
	ANN	Average	92.8	1.5	−1.4	0.8	−5.9	−5.2
	RF	Average	93.7	3.6	−3.2	1.7	−3.8	−3.3
	XG	Average	93.3	5.4	−4.8	1.3	−2	−1.7

This table also presents the absolute differences between the machine learning method accuracy, rain bias, and snow bias values and the best-performing benchmark and the average benchmark values. Note: a positive value for the accuracy absolute difference indicates the machine learning method provided an improvement relative to the benchmark, while a negative value for the snow and rain bias absolute differences corresponds to an improvement.

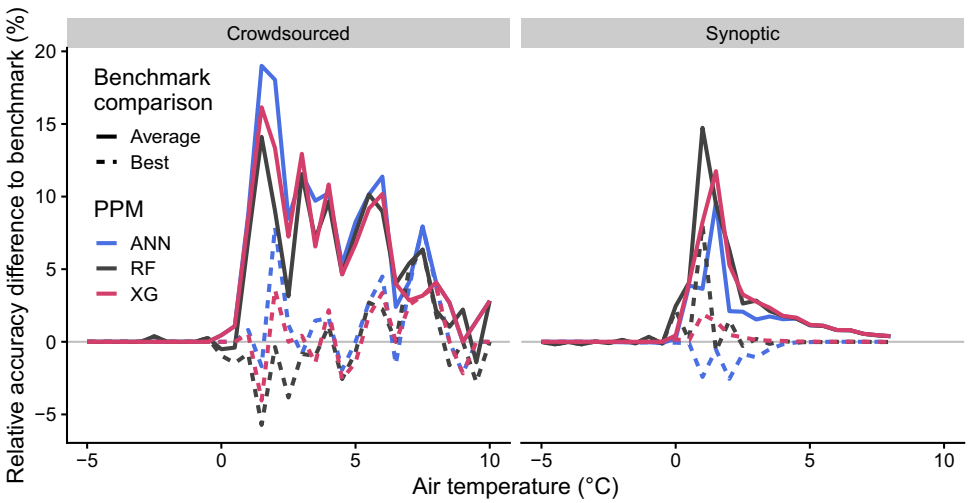

Fig. 2 | Relative accuracy differences by air temperature for the machine learning methods compared to the benchmarks. Dashed lines show the comparison relative to the best benchmark and the solid lines are relative to the average benchmark value. The three machine learning phase partitioning methods (PPM) are the artificial neural network (ANN), random forest (RF) and XGBoost (XG) applied to the crowdsourced (left panel) and synoptic datasets (right panel). Source data are provided as a Source Data file.

the three previously analyzed methods (ANN, random forest, and XGBoost) and predicts the precipitation phase with a regularized linear model. ANN-2 (88.8%) and the ensemble (88.9%) both had slightly lower accuracies than the ANN with only one hidden layer (89.2%), when evaluated on the crowdsourced data with rain and snow only. However, they were both 0.1% more accurate than the ANN with one hidden layer (79.2%) when evaluated against the crowdsourced rain, snow, and mixed data. Similar to the previously analyzed methods, the accuracy of ANN-2 and the stacked ensemble came at the expense of mixed bias in this dataset, with the former predicting no mixed precipitation and the latter predicting just 0.05% of the data to be mixed.

Relationship between meteorology and patterns of rain and snow

In this section, we return our focus to the predictions of only rain and snow given the poor mixed-phase results. The first consideration is whether the benchmark and machine learning methods match observed snowfall frequency curves (Supplementary Fig. 2). Two patterns emerge here. One, all phase partitioning methods fail to capture the sub-freezing rainfall events observed in both datasets. Although rainfall comprises a small proportion of total precipitation below 0 °C in the crowdsourced (2.5%) and synoptic (5.7%) datasets, these observations explain the rain biases approaching −100% shown in Fig. 1 to the left of the 0 °C isotherm. Two, the spreads between the minimum and maximum snowfall frequencies predicted by all of the benchmark and machine learning phase partitioning methods (pink shading in Supplementary Fig. 2) approach their largest values at air temperatures slightly above 0 °C. The largest difference (93.2%) occurs at 1.5°C in the crowdsourced dataset, which is the same temperature at which the synoptic dataset expresses its largest difference (66.9%).

Investigating the observed air temperature distributions of rain and snow, we find that the largest ranges in snowfall frequency

https://doi.org/10.1038/s41467-025-58234-2

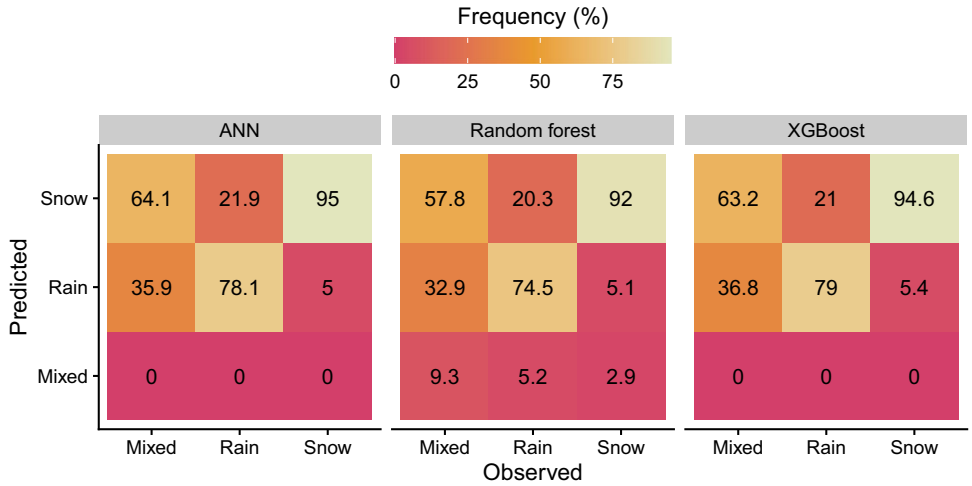

Fig. 3 | Confusion matrices showing the percentage of observations correctly and incorrectly predicted by the machine learning methods. These include the artificial neural network (ANN, left panel), random forest (center panel), and XGBoost (right panel) models. The x-axis shows the observed phase and the y-axis shows the predicted phase. To see the percentage of observations correctly predicted, find the cell where the observed phase matches the predicted phase. For example, if you line up mixed and mixed in the random forest panel, you see that 9.3% of observations were correctly predicted. To see the percentage of observations incorrectly predicted, line up the observed phase of interest with non-matching predicted phases. To illustrate, if you move up one cell from the previous example so the observed phase is mixed and the predicted phase is rain, then 32.9% of mixed precipitation was incorrectly predicted as rain by the random forest model. In the matrices, each column sums to 100%. Source data are provided as a Source Data file.

predicted by the different methods (Supplementary Fig. 2) correspond roughly to the air temperatures with the greatest degree of overlap in rain and snow occurrence. Specifically, there is a 33.7% overlap in the distributions of rain and snow as plotted by air temperature in the crowdsourced dataset and a 16% overlap in the synoptic dataset (Fig. 4). The maximum overlaps of the rain and snow distributions occur at 1.6 °C and 1 °C in the crowdsourced and synoptic datasets, respectively. What is more, if we only examine the 0 °C–4 °C range of poor performance, we find a 66.7% overlap in the crowdsourced dataset and 39.7% in the synoptic dataset. In other words, the near-surface air temperature characteristics of the two phases in both datasets are nearly indistinguishable between 0 °C and 4 °C.

The limits of using near-surface meteorology to partition rain and snow

This lack of differentiability at temperatures from 0° to 4°C reveals an issue with using near-surface meteorology to partition rain and snow. The left panels of Fig. 5 below present the average accuracy of the benchmark and machine learning phase partitioning methods plus the overlap in rain and snow distributions from the two datasets. We plot both variables against air temperature to demonstrate how the dip in performance corresponds to the increase in overlap. The right panel then shows that there is a statistically significant ($p < 0.0005$) negative relationship between the distribution overlap and average accuracy, where accuracy decreases as overlap increases. Using ordinary least squares regression, we compute an r^2 of 0.79 for this relationship in the crowdsourced dataset and 0.73 in the synoptic.

This relationship (Fig. 5) provides information on the limits of using near-surface meteorology to partition rain and snow. Recall that the benchmark and machine learning methods, when applied to the synoptic dataset, expressed higher accuracy values than in the crowdsourced dataset (Tables 2 and 3). Consistent with the findings in this section, the higher accuracy in the synoptic dataset is complemented by a lower percentage of rain and snow distribution overlap (16%) than in the crowdsourced dataset (33.7%). This, plus the fact that the air temperature distribution overlap between rain and snow explains at least 73% of the variance in precipitation phase partitioning accuracy, indicates that the key limiting factor to performance is the overlap of the air temperature distributions of the different phases. The higher the overlap, the worse the performance. Even the machine learning methods are beholden to this phenomenon (Fig. 2, Table 3). Thus, as of now, regions with a high degree of rain-snow air temperature distribution overlap (i.e., those with characteristics of the crowdsourced dataset) are likely to see minimum benchmark accuracy values approaching 66.3%, while regions with less overlap (i.e., those with characteristics of the synoptic dataset) may see improved minimum accuracies, approaching 68.7%. Although the machine learning methods may moderately improve this performance dip in the crowdsourced (up to 70%) and synoptic datasets (up to 74.2%), they do not eliminate it, demonstrating a limit to precipitation phase partitioning accuracy using near-surface meteorology.

Discussion

What is the take-home message of this work? In brief, we suggest that researchers stop investing time and money into making new or analyzing old precipitation phase partitioning methods that rely solely on near-surface meteorology. Previous research has already shown that threshold-based and statistical methods using near-surface meteorology to partition precipitation into rain and snow exhibit worse performance at air temperatures slightly above freezing[6,8,41]. We originally assumed that machine learning techniques could erase the performance dip, yet the ANN, random forest, and XGBoost methods only marginally improved accuracy and bias relative to the benchmarks. These modest performance increases fall well short of the revolutions brought about by machine learning in hydrologic modeling[32].

We consider two aspects of this issue before moving on to additional discussion points. The first is the complexity of the techniques we used. Many recent advances in hydrologic model performance have come through the application of Long Short Term Memory (LSTM), a type of recurrent neural network. Unlike the ANN tested here, an LSTM has a memory of previous states, produced by time series forcing, that facilitate improved predictions of streamflow[42] and snow water equivalent[43]. We did not apply an LSTM in this work primarily because it would require serially complete, continuous predictor data that are not available in either the crowdsourced or synoptic dataset. We did, however, evaluate benchmark and machine learning methods that

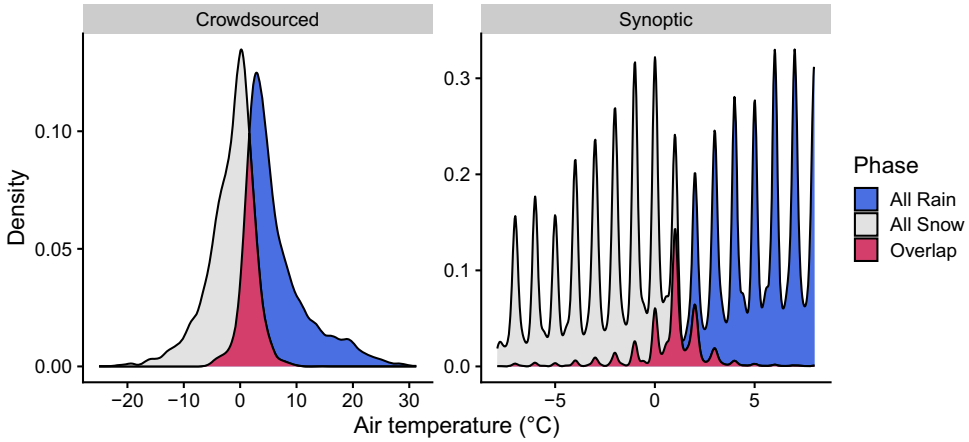

Fig. 4 | Kernel density plots showing the overlap in rain and snow distributions by air temperature in the crowdsourced and synoptic datasets. The waves in the synoptic dataset are a result of reporting practices that favor the submission of integer air temperature values. Source data are provided as a Source Data file.

span a range of complexity. The threshold benchmarks, which use only one near-surface meteorological variable apiece, performed nearly as well as the machine learning methods that used four or five predictor variables. Similarly, we did not find consistent improvement when applying the more complex ANN-2 and stacked ensemble approaches. This is in line with prior research that has shown a binary logistic regression model with three predictor variables does not outperform one with two predictor variables[6] and that an ANN does not always outperform a random forest[37] when predicting precipitation phase. While we found that increasing method complexity does not produce better performance, we do not suggest that there will be no future advancements in machine learning predictions of the precipitation phase. Rather, these advances are likely to result from the incorporation of different types and spatiotemporal dimensions of data.

This gets to the second—and the likely more important—consideration: the similar near-surface meteorological properties of rain and snow at air temperatures near 0 °C. The overlapping air temperature distributions of rain and snow in this work and others[9] suggest it will be practically impossible to effectively distinguish between solid and liquid precipitation using methods relying solely on near-surface meteorology between 0 °C and 4 °C. At best, such methods will return seasonally non-biased responses while still miscategorizing 25% or more of precipitation between 0.5 °C and 2.0 °C. Given the performance improvements produced by machine learning in other hydrologic science subdisciplines, we note here that this shortcoming is likely not an inherent issue of the techniques themselves, but rather of the near-surface meteorological predictor data at a single point in time and space. Previous work using LSTMs has shown that in addition to continuous time series forcing and hidden memory states, the addition of ancillary spatial information improves performance[44]. We therefore suggest that researchers stop trying to marginally improve the rain-snow partitioning accuracy of methods that use near-surface meteorology alone and instead focus on developing methods that integrate additional data sources.

These alternatives include networks and initiatives that provide precipitation phase observations, such as the Mountain Rain or Snow project presented in this manuscript or the Meteorological Phenomena Identification Near the Ground (mPING) program[45]—another citizen science effort—from the National Oceanic and Atmospheric Administration's (NOAA) National Severe Storms Laboratory. There is also NOAA's Local Climatological Dataset (LCD), which consists of visual observations, automated sensor reports, and combinations thereof from larger airports[46]. Some states also have road weather information systems (RWIS) that can infer precipitation types using automated present weather sensors[47].

Such point-based observational data can be complemented by weather radars that produce operational, real-time information over large spatial extents. These instruments provide publicly accessible data on rain, snow, and other precipitation types in both the horizontal and vertical directions[48–50]. Upwards-looking freezing level radars provide information on the precipitation phase only in the vertical direction by estimating the brightband elevation and using that level as a proxy for the melting of solid-phase precipitation into liquid[51,52].

Satellite-based radars provide an additional way of monitoring the precipitation phase over large spatial extents. One such instrument is the Dual-frequency Precipitation Radar (DPR) from the joint NASA-JAXA GPM mission. According to its algorithm theoretical basis document (ATBD), the DPR estimates the precipitation phase by calculating the brightband elevation, similar to the previously discussed freezing-level radars[53]. These data are available in a series of range bins representing different elevation bands of the atmosphere, while the Level 2A phaseNearSurface product classifies precipitation as rain, snow, or mixed in the lowest range bin without surface clutter. In a previous study, we found little spatiotemporal overlap of DPR overpasses with crowdsourced observations of the precipitation phase, thus limiting the product's utility in that study domain[9].

Large-scale climate, weather, and hydrology products also include estimates of the precipitation phase; however, many of these rely on modeled near-surface meteorological conditions to partition precipitation into rain and snow[54]. The IMERG Level 3 product, for example, uses wet bulb temperature from the fifth generation of the European Centre for Medium-Range Weather Forecasts reanalysis product (ERA5)[55] to produce spatially and temporally continuous estimates of the probability of liquid precipitation[56,57]. As such, this product suffers from the same shortcomings as other methods using near-surface meteorology, such as poor performance at air temperatures between approximately 0 °C and 4 °C and an inability to capture mixed precipitation[7,9].

The next question is: how do we integrate alternative data sources? Previous research has shown that crowdsourced programs[9,38,45,58,59] and observational networks[5,6,8,21] are indispensable for validating rain-snow partitioning methods. However, we need to move beyond validation. Recent studies have already shown the potential benefits of integrating crowdsourced data into other facets of numerical weather prediction[60,61]. One option is to deploy modeled or observed upper atmosphere information, which the weather forecasting field has been doing for decades[62]. However, weather forecasters struggle with the same challenges as hydrologic modelers: variability in precipitation phase partitioning performance near 0 °C[63,64] and issues with predicting mixed phase precipitation[65].

https://doi.org/10.1038/s41467-025-58234-2

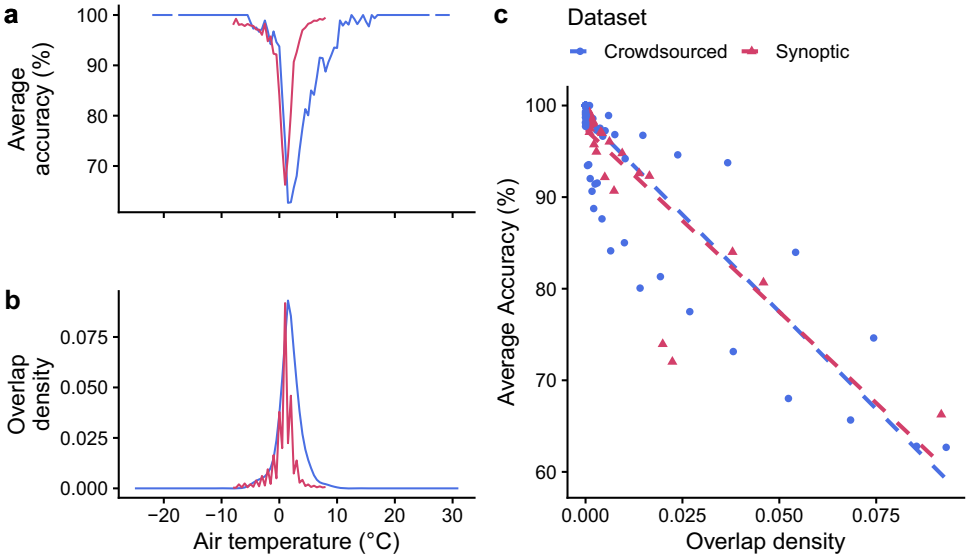

Fig. 5 | Relationship between rain-snow distribution overlap density and method accuracy. The average accuracy (**a**) of all precipitation phase partitioning methods (benchmarks and machine learning methods) and the rain-snow distribution overlap (**b**) for the synoptic and crowdsourced datasets are both plotted against air temperature. Average phase partitioning method accuracy is plotted against the rain-snow distribution overlap, where the dashed lines are the lines of best fit computed from ordinary least squares regression (**c**). Source data are provided as a Source Data file.

Despite these shared challenges, one study suggested the skill of weather models may exceed the ability of ground-based networks to capture rain-snow dynamics[66] while another noted the promise of using upper atmosphere information in a random forest model to predict precipitation phase[35]. A recent investigation showed that modeled atmospheric snow line data improved snowfall and snow water equivalent performance compared to a benchmark partitioning method using near-surface wet bulb temperature[67]. However, the same study also showed that forcing the model with the precipitation phase from a weather model's microphysics scheme degraded performance relative to the wet bulb temperature benchmark. These contradictory findings indicate there is much work left to do in this regard.

Another integration option is the extension and use of data assimilation (DA) techniques, although there are several significant challenges here. One is that while there are myriad DA methods in the earth sciences[68], the majority of them can only process quantitative data[69]. This contrasts to the commonly categorical observations of the precipitation phase. In this case, one may choose a DA method that can handle categorical data[70] or convert precipitation phase observations into continuous values, allowing for direct insertion of these observations. The Crocus snow model, for example, accepts continuous data for the precipitation phase, enabling the use of upper atmosphere information[67] and radar and disdrometer data[12]. An additional DA challenge is that most climate data products are spatially and temporally continuous while crowdsourced precipitation phase observations are reported opportunistically from variable locations. Furthermore, stations that report precipitation phase at regular intervals from the same location—e.g., stations in the LCD—only offer point observations and there is, as of now, no validated method we know of to accurately spatially distribute categorical phase observations. However, there tends to be high spatial coherence in rainfall and snowfall patterns[71], and many large-scale data products, such as the NLDAS-2[72,73], already assimilate continuous data (e.g., gage observations of precipitation) from point locations that do not fit the scale or spacing of their spatial grid. This suggests that additional research in this field could yield promising results.

As of now, however, most land surface and hydrologic models cannot directly deploy such methods nor assimilate their data without modifications. That means it may be necessary to rely on near-surface meteorology to partition rain and snow in the coming years. In that case, there is one guideline: use wet bulb temperature. In study after study, wet bulb temperature has proven to be the optimal choice, producing accurate phase predictions, improving snowpack simulations, and increasing the performance of hydrologic models[10,11,18,20,67]. In this study, we showed that the machine learning methods only slightly improve phase partitioning performance relative to the wet bulb temperature thresholds despite their increased complexity and computational expense. A look at the variable importance plots (Supplementary Fig. 3) indicates why. In every panel of Supplementary Fig. 3, we see the primacy of wet bulb temperature as the most important variable, meaning it provides the greatest positive contribution to accuracy in the tree-based machine learning methods, random forest, and XGBoost, in both datasets.

In this study, we assume that the crowdsourced observations and synoptic reports represent the true precipitation phase. Our group has previously shown that most crowdsourced observations pass quality control checks[9,38,39] while also noting that the crowdsourced data do not represent the true rain and snow frequencies of a given location because they are reported opportunistically. We have found, consistent with other research[71], that volunteers tend to report rain less frequently than snow. This could affect the outcomes of our machine learning predictions as these methods can be sensitive to class imbalances[74].

We had more snow (69.3%) than rain (30.7%) in the testing split of the crowdsourced data without mixed precipitation. When we included mixed precipitation, the breakdown was 61.1% snow, 27.1% rain, and 11.8% mixed. Our data preparation workflow did not use any over- or undersampling method to adjust class imbalances, so we reran the ANN, random forest, and XGBoost hyperparameter tuning exercise (Supplementary Table 2) for the no mixed and all-phase scenarios. We found that when using the synthetic minority oversampling technique (SMOTE)[75], accuracy declined in every case relative to using the imbalanced data directly (Supplementary Table 2). These accuracy declines ranged from −0.7% for XGBoost predicting rain and snow to −13.1% for ANN predicting rain, snow, and mixed precipitation. Rebalancing the data with SMOTE consistently produced more predictions

of mixed precipitation relative to the outcomes in Fig. 3; however, these new results were highly biased for random forest (74.7%), XGBoost (90.3%), and ANN (179.8%), suggesting there is more work to do on accurately classifying mixed precipitation.

Methods

Data

Citizen science observations. Our first dataset—hereafter the crowdsourced dataset—consists of 38.5 thousand crowdsourced observations of the precipitation phase collected between 2020-01-08 and 2023-07–24 from two related citizen science projects: Tahoe Rain or Snow and Mountain Rain or Snow (Fig. 6a). The former began in 2020 as a way to gather rain, snow, and mixed-phase observations in the Lake Tahoe region of California and Nevada[38]. This work then evolved into Mountain Rain or Snow (in 2022), with the team expanding the project to cover multiple mountain regions of the United States[39]. For both projects, volunteers submit reports via a mobile app that automatically geotags and timestamps each observation of the precipitation phase. We then access these reports from a private, password-protected database and process them to share publicly.

The processed dataset includes ancillary data such as elevation, modeled meteorological data, and quality control (QC) flags[9,38]. For each crowdsourced data point, we use the latitude, longitude, and timestamp to identify meteorological stations within a predefined search radius. We then download the data reported at those stations within a time window and remove anomalously high and low temperature and relative humidity data. Next, we select the air temperature values from each station that are closest in time to the crowdsourced report and use inverse distance weighting to distribute the data across the landscape[76]. We then compute a time-varying lapse rate from the stations and their elevations to estimate the air temperature corresponding to the time and location of the citizen science observation.

Most meteorological stations provide air temperature measurements, but wet bulb and dew point temperature and relative humidity are available less frequently. As such, we use empirical equations from Alduchov and Eskridge[77] and Stull[78] to estimate the values using the meteorological quantities provided by the stations. When station-observed dew point temperature data are available, or when we can estimate station dew point temperature using relative humidity and air temperature, we distribute and lapse dew point temperature using the same approach as for air temperature. These approaches produced low biases (0.05 °C for air temperature and 0.14 °C for dew point temperature) and high r^2 values (0.84 for air temperature and 0.72 for dew point temperature) when validating the estimates against observations on days in the record with precipitation and air temperatures between −5 °C and 10 °C[9].

In this study, we only used observations that passed our six QC checks. The first removes duplicate observations, which we define as those submitted by the same user at the same timestamp. The second QC step removes empty reports—i.e. those that do not include precipitation phase—from the database. The third and fourth QC checks ensure that observations correspond to reasonable air temperature (snow <10°C; rain > −5 °C) and relative humidity (> 30%) values. The last two QC checks require that each report has at least five nearby meteorological stations reporting air temperature and dew point temperature, respectively.

Northern hemisphere observations. The second dataset—hereafter the synoptic dataset—is a processed collection of 17.8 million rain and snow observations from land-based stations in the northern hemisphere (Fig. 6b), excluding the tropics, between 1978-01-01 and 2007-02–25[6]. These data come from the National Center for Environmental Prediction (NCEP) Automated Data Processing (ADP) Operational Global Surface Observations program, which includes 3 h and 6 h synoptic reports of weather conditions and precipitation phase along with concomitant measurements of air temperature and dew point temperature from a set of global meteorological stations[5].

Some of the reports (80.5%) included surface pressure. For the remaining reports, Jennings et al. [6] used the 1980–2007 average December, January, and February surface pressure from the Modern-Era Retrospective analysis for Research and Applications version 2 (MERRA−2) reanalysis dataset for the grid cell containing the observation[79,80]. Additionally, Jennings et al. [6] calculated relative humidity for each report as a function of air and dew point temperature and surface pressure as in Dai[81]. Here, we also calculated the wet bulb temperature for each report using the Stull[78] equation.

Each observation in the dataset uses the rain and snow classifications from Jennings et al. [6]. These are based on the World Meteorological Organization precipitation phase categories described in detail in Dai[5,82]. Unlike the crowdsourced dataset, the synoptic observations do not include mixed-phase precipitation because some of the mixed categories are ambiguous as to what type of precipitation was occurring at the time of observation. For example, code 26 corresponds to: Shower(s) of snow, or of rain and snow.

Rain-snow partitioning methods

Benchmarks. We use standard rain-snow partitioning methods that incorporate near-surface meteorology from the literature as benchmarks in this work. These include air, wet bulb, and dew point temperature thresholds along with a binary logistic regression model using optimized coefficients from Jennings et al. [6]. Each of the threshold-based methods uses only one near-surface meteorological variable, while the binary logistic regression model uses two predictor variables. For ease of interpretation, we only include the most performant methods from previous studies (Table 1). As such, we do not consider dual-threshold methods that assign precipitation to be snowfall when air temperature is below one threshold and rainfall above another, with a mix of the two phases in between. We forgo these methods because they performed poorly relative to observations in our previous study[9]. For example, the dual-threshold range from −0.5 °C to 0.5 °C correctly identified the precipitation phase in only 47.1% of the observations, which was 20.0% worse than the next best method.

Machine learning methods. We explore the applicability of machine learning methods to see if we can leverage data-driven techniques to better predict precipitation phase than the benchmarks using near-surface meteorology (Fig. 7). Here, we focus on the outcomes from three methods of varying complexity: random forest[36], XGBoost[40], and an ANN with one hidden layer (specifically a multilayer perceptron, which is a feed-forward neural network[83,84]). The first two, random forest and XGBoost, are conventional tree-based machine learning methods. The former uses a simple majority vote classification derived from the output of the random forest's many trees, whereas the latter uses a more complex gradient boosting algorithm that iteratively builds the weighted ensemble of trees[85]. The ANN, in turn, is the most complex of the three as a neural network with one input layer, one hidden layer with a tunable number of hidden units, and one output layer.

We apply each method using the Tidymodels[86] package in R to classify the precipitation phase in the two datasets. This provides two scenarios for the crowdsourced dataset (1: rain and snow only; 2: rain, snow, and mixed precipitation) and one for the synoptic dataset (rain and snow only). We use the near-surface meteorological data associated with each dataset as the predictor variables. For the crowdsourced dataset, these four variables are air, wet bulb, and dew point temperature, and relative humidity. We use these same four variables plus surface pressure, for a total of five predictor variables, in the synoptic dataset. This means the machine learning methods

https://doi.org/10.1038/s41467-025-58234-2

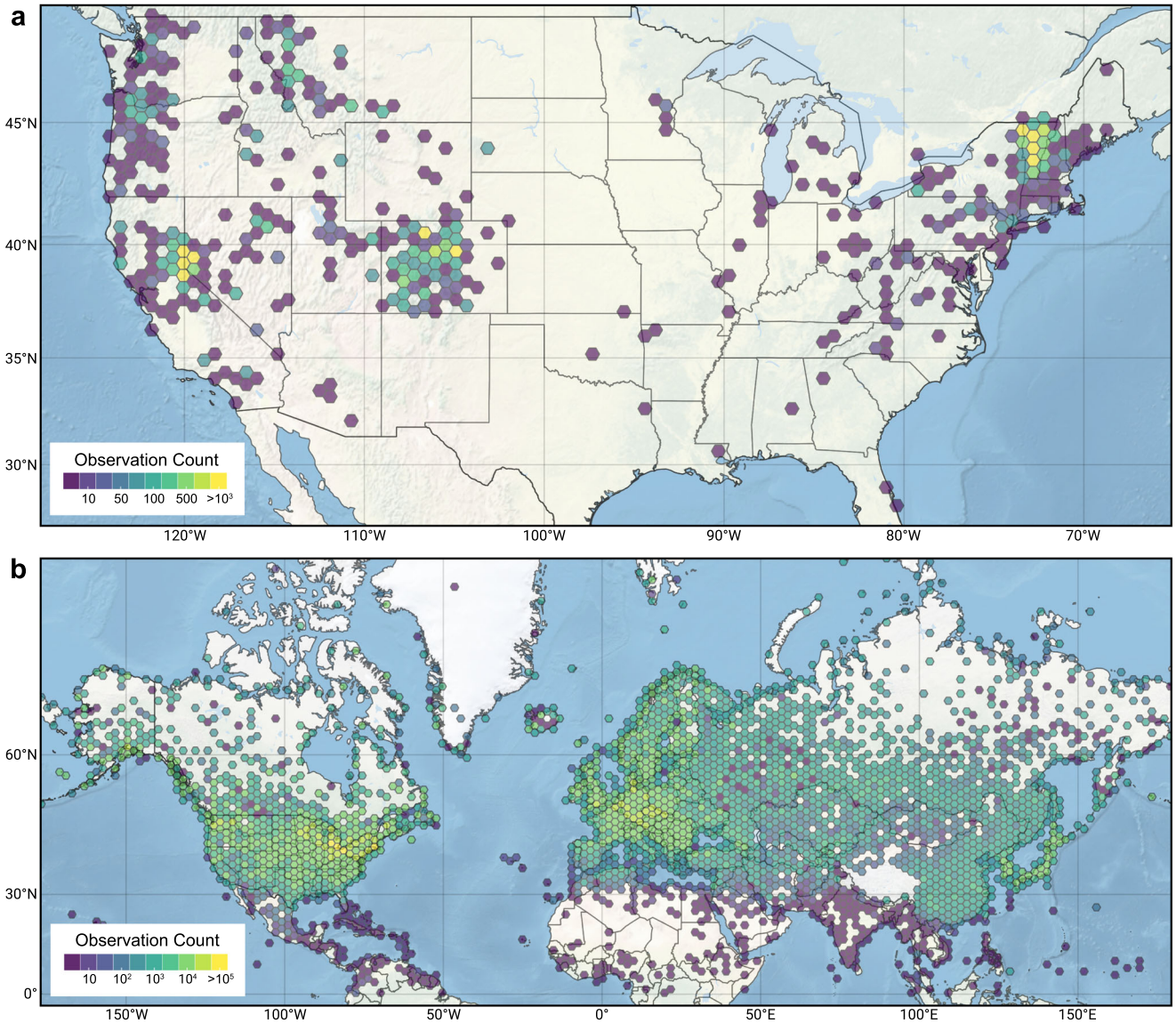

Fig. 6 | Heat maps of crowdsourced precipitation phase observations. Data shown are from the Mountain Rain or Snow dataset[38,39] (**a**) and the synoptic dataset[5,6] (**b**). The color of each hexagon corresponds to the number of observations within that cell over the study period. Note: the color scale differs in the two maps.

incorporate more near-surface meteorological data (four or five variables, depending on the dataset) than the benchmark methods (one variable per threshold and two variables for the binary logistic regression model).

To tune and deploy the machine learning models, we randomly split each dataset into training data and testing data using a 75% training, 25% testing allocation. We stratify the split using the precipitation phase to have approximately equal proportions of each phase in the training and testing data. We then run a hyperparameter tuning exercise on 10 random, independent folds of the training data to identify the best model configurations for each dataset, scenario, and algorithm. The Hyperparameter Tuning section of the Supplementary Information provides further details for each method. Specifically, we provide the hyperparameter tuning ranges (Supplementary Tables 3, 8, and 13); mean performance values for the crowdsourced rain and snow (Supplementary Tables 4, 9, and 14), the crowdsourced rain, snow, and mixed (Supplementary Tables 5, 10, and 15), and the synoptic rain and snow (Supplementary Tables 6, 11, and 16) tuning exercises; and the optimized hyperparameters for each method (Supplementary Tables 7, 12, and 17). We then fit the models to all the training data using the tuned hyperparameters. We next apply the fitted models to the testing data and compare the resulting predictions to the observed data to calculate the performance metrics detailed in the following subsection.

In addition to the three previously detailed methods, we evaluate whether increasing model complexity leads to better performance in the crowdsourced dataset. Here, we include an ANN with two hidden layers (ANN-2) and a stacked ensemble of the random forest, XGBoost, and ANN output. We tune the hyperparameters and fit the ANN-2 model using the same protocol described above for the ANN with one hidden layer (Supplementary Tables 18–20). For the stacked ensemble, we fit the three constituent models using their previously tuned hyperparameters. We then use the R stacks package to combine the three individual precipitation phase predictions into a single outcome with a regularized linear model[87].

Method comparison

To predict rain versus snow, we apply both the benchmark methods and the tuned machine-learning techniques to the crowdsourced and synoptic datasets and compare the outcomes to the reported phase. To prevent data leakage and enforce direct comparisons between the benchmarks and machine learning methods, we perform our

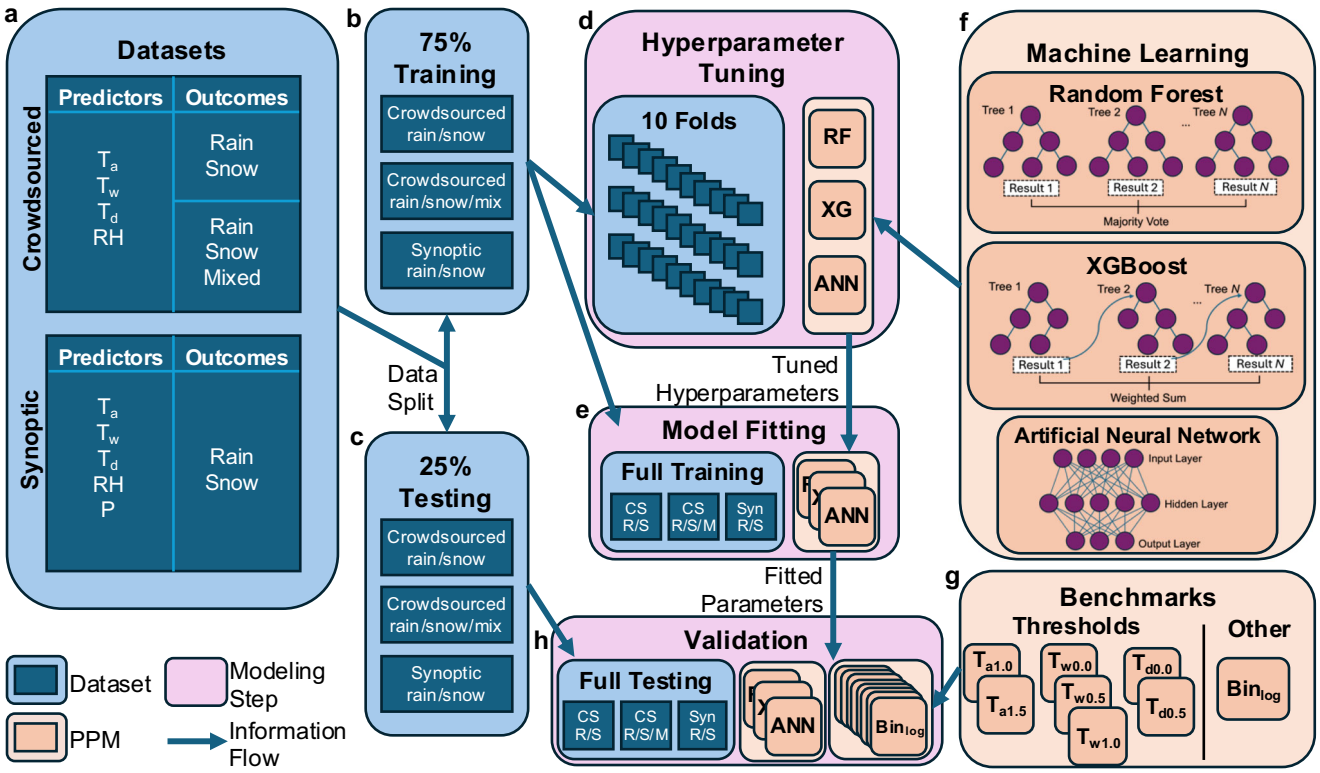

Fig. 7 | Overview schematic of our project workflow. Here, we show the datasets (blue boxes), the tuning, fitting, and validation steps (pink boxes), and the phase partitioning methods (PPMs, peach boxes). We start with complete datasets (**a**), which include observations of rain (R), snow (S), and mixed precipitation (M) along with air (T_a), wet bulb (T_w), and dew point (T_d) temperature, relative humidity (RH), and pressure (P). We split these datasets into training (**b**) and testing (**c**). We tune the hyperparameters (**d**) and fit (**e**) the random forest (RF), XGBoost (XG), and artificial neural network (ANN) machine learning models (**f**) using the training data. We then apply the machine learning models and the benchmarks (**g**, Table 1) to the testing data to validate their precipitation phase predictions (**h**).

evaluations using only the testing data, even for the benchmarks. These data, representing 25% of all observations in each dataset, are not used in training the machine learning models, making them an independent validation source of method performance.

To quantify performance we focus primarily on accuracy, which we define as the total percent of precipitation phase reports correctly predicted by a given method. Accuracy ranges from 0% to 100%, where 0% corresponds to no phase reports correctly predicted and 100% corresponds to all reports correctly predicted. We also present the relative rain and snow biases for each method. A value of 0% equates to non-biased, while negative and positive biases correspond, respectively, to under and overestimates of the given phase. We present accuracy and bias statistics as aggregate values for each phase partitioning method and we also present these figures by 0.5 °C air temperature bins for each method to explore the effects of air temperature on method performance.

Evaluation of rain-snow frequency and near-surface meteorology

After comparing the performance of the different methods, we also evaluate the meteorological properties of the precipitation phase categories—rain, snow, and mixed—to determine their unique and shared characteristics. We consider the distributions of air, wet bulb, and dew point temperature from the two datasets along with the relative proportions of each precipitation phase by 0.5 °C temperature bins.

Data availability

The crowdsourced data used in this study are available in the Mendeley Data database under accession code https://doi.org/10.17632/x84hy7yky4.1. The synoptic data used in this study are available in the DataDryad database under accession code https://doi.org/10.5061/dryad.c9h35. Source data are provided with this paper.

Code availability

All of the data preparation, modeling, analysis, and plotting code can be found on the lead author's GitHub page (https://github.com/SnowHydrology/MountainRainOrSnow/tree/main/manuscripts/jennings_et_al_limits). A DOI version of the full repository is also available (https://doi.org/10.5281/zenodo.14968963).

References

1. Feiccabrino, J., Graff, W., Lundberg, A., Sandström, N. & Gustafsson, D. Meteorological knowledge useful for the improvement of snow rain separation in surface based models. *Hydrology* **2**, 266–288 (2015).
2. Harpold, A. A. et al. Rain or snow: hydrologic processes, observations, prediction, and research needs. *Hydrol. Earth Syst. Sci.* **21**, 1–22 (2017).
3. United States Army Corps of Engineers. Snow hydrology. *US Army North Pac. Div. Portland Or.* (1956).
4. Auer, A. H. Jr. The rain versus snow threshold temperatures. *Weatherwise* **27**, 67–67 (1974).
5. Dai, A. Temperature and pressure dependence of the rain-snow phase transition over land and ocean. *Geophys. Res. Lett.* **35**, L12802 (2008).
6. Jennings, K. S., Winchell, T. S., Livneh, B. & Molotch, N. P. Spatial variation of the rain-snow temperature threshold across the Northern Hemisphere. *Nat. Commun.* **9**, 1148 (2018).

7. Yu, G. et al. Crowdsourced data reveal shortcomings in precipitation phase products for rain and snow partitioning. *Geophys. Res. Lett.* **51**, e2024GL112853 (2024).
8. Ding, B. et al. The dependence of precipitation types on surface elevation and meteorological conditions and its parameterization. *J. Hydrol.* **513**, 154–163 (2014).
9. Jennings, K. S. et al. Crowdsourced data highlight precipitation phase partitioning variability in rain-snow transition zone. *Earth Space Sci.* **10**, e2022EA002714 (2023).
10. Harder, P. & Pomeroy, J. W. Hydrological model uncertainty due to precipitation-phase partitioning methods. *Hydrol. Process.* **28**, 4311–4327 (2014).
11. Jennings, K. S. & Molotch, N. P. The sensitivity of modeled snow accumulation and melt to precipitation phase methods across a climatic gradient. *Hydrol. Earth Syst. Sci.* **23**, 3765–3786 (2019).
12. Leroux, N. R., Vionnet, V. & Thériault, J. M. Performance of precipitation phase partitioning methods and their impact on snowpack evolution in a humid continental climate. *Hydrol. Process.* **37**, e15028 (2023).
13. Motoyama, H. Simulation of seasonal snowcover based on air temperature and precipitation. *J. Appl. Meteorol. Climatol.* **29**, 1104–1110 (1990).
14. L'hôte, Y., Chevallier, P., Coudrain, A., Lejeune, Y. & Etchevers, P. Relationship between precipitation phase and air temperature: comparison between the Bolivian Andes and the Swiss Alps/Relation entre phase de précipitation et température de l'air: comparaison entre les Andes Boliviennes et les Alpes Suisses. *Hydrol. Sci. J.* **50**, 997 (2005).
15. Kienzle, S. W. A new temperature based method to separate rain and snow. *Hydrol. Process.* **22**, 5067–5085 (2008).
16. Quick, M. C. & Pipes, A. UBC WATERSHED MODEL/Le modèle du bassin versant UCB. *Hydrol. Sci. J.* **22**, 153–161 (1977).
17. Tarboton, D. G. & Luce, C. H. *Utah Energy Balance Snow Accumulation and Melt Model (UEB)*. (1996).
18. Harder, P. & Pomeroy, J. Estimating precipitation phase using a psychrometric energy balance method. *Hydrol. Process.* **27**, 1901–1914 (2013).
19. Marks, D., Winstral, A., Reba, M., Pomeroy, J. & Kumar, M. An evaluation of methods for determining during-storm precipitation phase and the rain/snow transition elevation at the surface in a mountain basin. *Adv. Water Resour.* **55**, 98–110 (2013).
20. Behrangi, A., Yin, X., Rajagopal, S., Stampoulis, D. & Ye, H. On distinguishing snowfall from rainfall using near-surface atmospheric information: Comparative analysis, uncertainties and hydrologic importance. *Q. J. R. Meteorol. Soc.* **144**, 89–102 (2018).
21. Froidurot, S., Zin, I., Hingray, B. & Gautheron, A. Sensitivity of precipitation phase over the swiss alps to different meteorological variables. *J. Hydrometeorol.* **15**, 685–696 (2014).
22. Hu, J., Che, T., He, Y., Huang, W. & Yang, X. An intercomparison of empirical schemes for partitioning precipitation phase. *J. Hydrol. Reg. Stud.* **53**, 101757 (2024).
23. Huss, M. et al. Toward mountains without permanent snow and ice. *Earths Future* **5**, 418–435 (2017).
24. Immerzeel, W. W. et al. Importance and vulnerability of the world's water towers. *Nature* **577**, 364–369 (2020).
25. Siirila-Woodburn, E. R. et al. A low-to-no snow future and its impacts on water resources in the western United States. *Nat. Rev. Earth Environ.* **2**, 800–819 (2021).
26. Lynn, E. et al. Technical note: Precipitation-phase partitioning at landscape scales to regional scales. *Hydrol. Earth Syst. Sci.* **24**, 5317–5328 (2020).
27. Heggli, A., Hatchett, B., Schwartz, A., Bardsley, T. & Hand, E. Toward snowpack runoff decision support. *iScience* **25**, 104240 (2022).
28. Musselman, K. N. et al. Projected increases and shifts in rain-on-snow flood risk over western North America. *Nat. Clim. Change* 1 https://doi.org/10.1038/s41558-018-0236-4 (2018).
29. Polade, S. D., Gershunov, A., Cayan, D. R., Dettinger, M. D. & Pierce, D. W. Precipitation in a warming world: Assessing projected hydro-climate changes in California and other Mediterranean climate regions. *Sci. Rep.* **7**, 10783 (2017).
30. Rhoades, A. M. et al. Asymmetric emergence of low-to-no snow in the midlatitudes of the American Cordillera. *Nat. Clim. Change* **12**, 1151–1159 (2022).
31. Shulgina, T. et al. Observed and projected changes in snow accumulation and snowline in California's snowy mountains. *Clim. Dyn.* **61**, 4809–4824 (2023).
32. Kratzert, F. et al. Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resour. Res.* **55**, 11344–11354 (2019).
33. Frame, J. M. et al. Deep learning rainfall–runoff predictions of extreme events. *Hydrol. Earth Syst. Sci.* **26**, 3377–3392 (2022).
34. McGovern, A. et al. Using Artificial Intelligence to Improve Real-Time Decision-Making for High-Impact Weather. *Bull. Am. Meteorol. Soc.* **98**, 2073–2090 (2017).
35. Filipiak, B. C., Bassill, N. P., Corbosiero, K. L., Lang, A. L. & Lazear, R. A. Probabilistic Forecasting Methods of Winter Mixed-Precipitation Events in New York State Utilizing a Random Forest. *Artif. Intell. Earth Syst.* **2**, e220080 (2023).
36. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
37. Campozano, L. et al. Parsimonious models of precipitation phase derived from random forest knowledge: intercomparing logistic models, neural networks, and random forest models. *Water* **13**, 3022 (2021).
38. Arienzo, M. M., Collins, M. & Jennings, K. S. Enhancing engagement of citizen scientists to monitor precipitation phase. *Front. Earth Sci.* https://doi.org/10.3389/feart.2021.617594 (2021).
39. Collins, M. et al. Effective engagement while scaling up: lessons from a citizen science program transitioning from single- to multi-region scale. *Citiz. Sci.: Theory Pract.* **8**, 65 (2023).
40. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, San Francisco California USA, 2016). https://doi.org/10.1145/2939672.2939785.
41. Ye, H., Cohen, J. & Rawlins, M. Discrimination of solid from liquid precipitation over northern eurasia using surface atmospheric conditions*. *J. Hydrometeorol.* **14**, 1345–1355 (2013).
42. Kratzert, F., Klotz, D., Brenner, C., Schulz, K. & Herrnegger, M. Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrol. Earth Syst. Sci.* **22**, 6005–6022 (2018).
43. Meyal, A. Y. et al. Automated Cloud based long short-term memory neural network based SWE prediction. *Front. Water* **2**, 574917 (2020).
44. Kratzert, F. et al. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrol. Earth Syst. Sci.* **23**, 5089–5110 (2019).
45. Elmore, K. L. et al. MPING: Crowd-sourcing weather reports for research. *Bull. Am. Meteorol. Soc.* **95**, 1335–1342 (2014).
46. Landolt, S. D. et al. The impacts of automation on present weather–type observing capabilities across the conterminous united. *S. J. Appl. Meteorol. Climatol.* **58**, 2699–2715 (2019).
47. White, S. P., Thornes, J. E. & Chapman, L. A guide to road weather information systems. *Version* **2**, 1–83 (2006).
48. Serafin, R. J. & Wilson, J. W. Operational weather radar in the united states: progress and opportunity. *Bull. Am. Meteorol. Soc.* **81**, 501–518 (2000).

49. Maddox, R. A., Zhang, J., Gourley, J. J. & Howard, K. W. Weather Radar coverage over the contiguous United States. *Weather Forecast.* **17**, 927–934 (2002).
50. Kitzmiller, D., Miller, D., Fulton, R. & Ding, F. Radar and multisensor precipitation estimation techniques in national weather service hydrologic operations. *J. Hydrol. Eng.* **18**, 133–142 (2013).
51. White, A. B. et al. Developing a performance measure for snow-level forecasts. *J. Hydrometeorol.* **11**, 739–753 (2010).
52. Hatchett, B. J. et al. Winter snow level rise in the Northern Sierra Nevada from 2008 to 2017. *Water* **9**, 899 (2017).
53. Iguchi, T. et al. GPM/DPR Level-2 Algorithm Theoretical Basis Document. (2018).
54. Xiong, W., Tang, G., Wang, T., Ma, Z. & Wan, W. Evaluation of IMERG and ERA5 precipitation-phase partitioning on the global scale. *Water* **14**, 1122 (2022).
55. Hersbach, H. et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).
56. Skofronick-Jackson, G. et al. The Global Precipitation Measurement (GPM) mission's scientific achievements and societal contributions: reviewing four years of advanced rain and snow observations. *Q. J. R. Meteorol. Soc.* **144**, 27–48 (2018).
57. Huffman, G. et al. NASA Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG) Prepared for: Global Precipitation Measurement (GPM) National Aeronautics and Space Administration (NASA). *Algorithm Theoretical Basis Document (ATBD) Version 06* 26–26 (2019).
58. Elmore, K. L., Grams, H. M., Apps, D. & Reeves, H. D. Verifying forecast precipitation type with mPING. *Weather Forecast.* **30**, 656–667 (2015).
59. Casellas, E. et al. Surface precipitation phase discrimination in complex terrain. *J. Hydrol.* **592**, 125780 (2021).
60. Nipen, T. N., Seierstad, I. A., Lussana, C., Kristiansen, J. & Hov, Ø. Adopting citizen observations in operational weather prediction. *Bull. Am. Meteorol. Soc.* **101**, E43–E57 (2020).
61. Garcia-Marti, I. et al. From proof-of-concept to proof-of-value: Approaching third-party data to operational workflows of national meteorological services. *Int. J. Climatol.* **43**, 275–292 (2023).
62. Heppner, P. O. G. Snow versus Rain: Looking beyond the "Magic" Numbers. *Weather Forecast.* **7**, 683–691 (1992).
63. Reeves, H. D., Elmore, K. L., Ryzhkov, A., Schuur, T. & Krause, J. Sources of uncertainty in precipitation-type forecasting. *Weather Forecast.* **29**, 936–953 (2014).
64. Minder, J. R. et al. P-Type processes and predictability: The Winter Precipitation Type Research Multiscale Experiment (WINTRE-MIX). *Bull. Am. Meteorol. Soc.* **104**, E1469–E1492 (2023).
65. Ikeda, K., Steiner, M. & Thompson, G. Examination of mixed-phase precipitation forecasts from the high-resolution rapid refresh model using surface observations and sounding data. *Weather Forecast.* **32**, 949–967 (2017).
66. Lundquist, J., Hughes, M., Gutmann, E. & Kapnick, S. Our skill in modeling mountain rain and snow is bypassing the skill of our observational networks. *Bull. Am. Meteorol. Soc.* **100**, 2473–2490 (2019).
67. Vionnet, V. et al. Snow level from post-processing of atmospheric model improves snowfall estimate and snowpack prediction in mountains. *Water Resour. Res.* **58**, e2021WR031778 (2022).
68. Reichle, R. H. Data assimilation methods in the Earth sciences. *Adv. Water Resour.* **31**, 1411–1418 (2008).
69. Li, X. et al. Land data assimilation: harmonizing theory and data in land surface process studies. *Rev. Geophys.* **62**, e2022RG000801 (2024).
70. Hu, X., Liu, F., Qi, Y., Zhang, J. & Li, X. A novel strategy to assimilate category variables in land-use models based on Dirichlet distribution. *Environ. Model. Softw.* **149**, 105324 (2022).
71. Reeves, H. D. The uncertainty of precipitation-type observations and its effect on the validation of forecast precipitation type. *Weather Forecast.* **31**, 1961–1971 (2016).
72. Cosgrove, B. A. et al. Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project. *J. Geophys. Res. Atmos.* **108**, 8842 (2003).
73. Xia, Y. et al. Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *J. Geophys. Res. Atmos.* **117**, D03109 (2012).
74. Tyralis, H., Papacharalampous, G. & Langousis, A. A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water* **11**, 910 (2019).
75. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
76. Bavay, M. & Egger, T. MeteoIO 2.4.2: a preprocessing library for meteorological data. *Geosci. Model Dev.* **7**, 3135–3151 (2014).
77. Alduchov, O. A. & Eskridge, R. E. Improved Magnus form approximation of saturation vapor pressure. *J. Appl. Meteorol.* **35**, 601–609 (1996).
78. Stull, R. Wet-bulb temperature from relative humidity and air temperature. *J. Appl. Meteorol. Climatol.* **50**, 2267–2269 (2011).
79. Global Modeling and Assimilation Office (GMAO). MERRA-2 tavg1_2d_slv_Nx: 2d,1-Hourly,Time-Averaged,Single-Level,Assimilation,Single-Level Diagnostics V5.12.4. *Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC)* (2015).
80. Global Modeling and Assimilation Office (GMAO). MERRA-2 tavg1_2d_flx_Nx: 2d,1-Hourly,Time-Averaged,Single-Level,Assimilation,Single-Level Diagnostics V5.12.4. *Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC)* (2015).
81. Dai, A. Recent climatology, variability, and trends in global surface humidity. *J. Clim.* **19**, 3589–3606 (2006).
82. Dai, A. Global precipitation and thunderstorm frequencies. Part I: Seasonal and interannual variations. *J. Clim.* **14**, 1092–1111 (2001).
83. Gardner, M. W. & Dorling, S. R. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmos. Environ.* **32**, 2627–2636 (1998).
84. Popescu, M.-C., Balas, V. E., Perescu-Popescu, L. & Mastorakis, N. Multilayer perceptron and neural networks. *WSEAS Trans. Cir. Sys* **8**, 579–588 (2009).
85. Bentéjac, C., Csörgő, A. & Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **54**, 1937–1967 (2021).
86. Kuhn, M. & Wickham, H. Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles. R package version 1.2.0. (2020).
87. Couch, S. & Kuhn, M. stacks: Tidy Model Stacking. R package version 1.0.5. (2024).

Acknowledgements

Community engagement, data collection, and analysis were funded by the National Aeronautics and Space Administration (NASA) under grants 80NSSC21K0855 (K.S.J., M.C., M.M.A., B.J.H., and A.W.N.) and 80NSSC22K1916 (K.S.J., M.C., M.M.A., B.J.H., and A.W.N.) issued through the Citizen Science for Earth Systems Program (CSESP). Tahoe Rain or Snow was funded under grant 80NSSC19M0056 (M.M.A., M.C., and K.S.J.) issued through the Nevada NASA Established Program to Stimulate Competitive Research (EPSCoR). In addition to our funding sources, we are grateful to the community observers who graciously volunteered their time to provide rain, snow, and mixed precipitation reports. We also acknowledge the Citizen Science Tahoe Collaborative which provided the platform that hosted the original *Tahoe Rain or Snow* app.

Author contributions

K.S.J. performed the analyses and wrote the manuscript draft. M.C. led community engagement and the collection of crowdsourced data with substantive support from A.H., N.H., S.T., and M.M.A. N.H. processed the crowdsourced data. B.J.H., A.W.N., G.Y., and W.Z. contributed extensively to iterations on study design and manuscript content. K.S.J., M.C., B.J.H., A.H., N.H., S.T., A.W.N., G.Y., W.Z., and M.M.A. edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-58234-2.

Correspondence and requests for materials should be addressed to Keith S. Jennings.

Peer review information *Nature Communications* thanks Sarmad Latif and Yuan-Heng Wang for their contribution to the peer review of this work. A peer review file is available.