JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION

AMERICAN WATER RESOURCES ASSOCIATION

Exploring the Use of Decision Tree Methodology in Hydrology Using Crowdsourced Data

Di Wu, Elizabeth A. Del Rosario, and Christopher Lowry

Research Impact Statement: The decision tree methodology has good performance on figuring out uncertainty in hydrological citizen science data and can be applied as a quality control tool for crowdsourcing dataset.

ABSTRACT: To fill the observations gap on ungauged streams, crowdsourced distributed hydrologic measurements were considered as a potential supplement for observational data networks. However, citizen science data come with uncertainty as they are provided by the general public. In order to investigate this uncertainty, a decision tree methodology was applied to evaluate existing citizen science data of stream stage based on the CrowdHydrology (CH) network. Quality control (QC) flags were developed and applied to CH sites, dividing Level 1 dataset (raw dataset) into Level 2 (flagged dataset) and Level 3 (processed dataset). Error estimates were calculated to determine uncertainty in the citizen science data. The results indicate that the decision tree could provide reliable QC for citizen science data and demonstrate how uncertainty can be quantified in the QC datasets.

(KEYWORDS: rivers/streams; public participation; computational methods; crowd hydrology; citizen science; crowdsourcing.)

INTRODUCTION

As an enhancement to traditional research, citizen science projects based on crowdsourcing have the potential to complement existing observation networks, meeting the challenges of limited data availability (Davids et al. 2019; Njue et al. 2019; Seibert et al. 2019). Defined by National Oceanic and Atmospheric Administration (NOAA), citizen science is "a form of open collaboration where members of the public participate in the scientific process to address real-world problems in ways that include identifying research questions, collecting and analyzing data, interpreting results, making new discoveries, developing technologies and applications, and solving complex problems" (Dickinson et al. 2012; NOAA 2018). The revolution of mobile phones and the Internet in recent years provides citizen volunteers with easier and more efficient approaches to collect, store and communicate a large amount of data, which has contributed to the growth of citizen science in new fields with innovative methods (Sullivan et al. 2009; McCormick 2012; Hemmi and Graham 2014; Poelen et al. 2014).

These technological advances have encouraged a worldwide increase in measurements by citizen scientist in hydrological research with a wide range including streamflow estimation, floods prediction, hydrological database generation, and water quality monitoring (Lowry and Fienen 2013; Toivanen et al. 2013; Le Coz et al. 2016). Crowdsourcing hydrologic data, where data are provided by the crowd, could help fill the information gap on intermittence steams, vastly increase the number of monitored tributaries in a watershed, and expand understanding of when, where, and how streams flow (Lowry et al. 2019).

Increasing research and programs about crowdsourcing hydrology were launched in recent years.

Paper No. JAWR-19-0109-P of the *Journal of the American Water Resources Association* (JAWR). Received July 18, 2019; accepted July 21, 2020.

Envriomental Resources and Policy (Wu), Southern Illinois University-Carbondale Carbondale, Illinois,USA; Harte Research Institute (Del Rosario), Texas A&M University-Corpus Christi Corpus Christi, Texas,USA; and Department of Geology (Lowry), University at Buffalo Buffalo, New York,USA (Correspondence to Wu: di.wu@siu.edu).

Citation: Wu, D., E.A. Del Rosario, and C. Lowry. 2020. "Exploring the Use of Decision Tree Methodology in Hydrology Using Crowdsourced Data." *Journal of the American Water Resources Association* 1–11. https://doi.org/10.1111/1752-1688.12882.

Seibert et al. (2019) generated a smartphone app that allows the collection of stream level information occurring at places without physical staff gauges. It provided the public with easy access to set up a new measurement site and encouraged increased public participation in citizen science. Lowry et al. (2019) evaluated a citizen science hydrological program, CrowdHydrology (CH), in data accuracy, citizen participation, and station popularity to figure out the barriers that may inhibit public participation. Davids et al. (2019) evaluated three citizen science streamflow measurement methods, finding the preferred method and applying it to larger regions. Weeser et al. (2018) estimated the quality and quantity of data generated by the public in a remote Kenyan basin, demonstrating that water level data can be measured with enough quality and high temporal resolution by the public. Various citizen science programs related to hydrology have been launched by governments such as Volunteer Water Monitoring Programs (USA), the Risk-Scape Project (New Zealand), and Water-Watch Victoria monitoring network (Australia); which not only vastly increases the number and type of available hydrologic data with low-cost collection methods, but also promotes public understanding about hydrological processes and participation in science (Gauchat 2012; Kampf et al. 2018).

Uncertainty and error in citizen science measurements are a primary concern for the scientific community (Law et al. 2017). Fienen and Lowry (2012) found high-quality observations could be obtained without requiring trained observers and stated that with a simple filter, errors such as transcriptions could be removed from the dataset. The objective of this study was to develop a quality control (QC) method for finding errors in crowdsourced data so it can be used to expand the observational network into ungauged watersheds. A decision tree methodology was developed to apply a QC filter to citizen science data. Using this method, the citizen science data would pass through a QC process consisting of L1 (raw dataset) to L2 (flagged dataset) to L3 (processed dataset).

DATA COLLECTION

Citizen science data are considered to be a high risk for potential in error. This error can be reduced using reference datasets to flag atypical data points. Data quality can be influenced by factors such as sampling scale, frequency, location, collection method, etc.; therefore, multiple sources of reference data were utilized in this study. The data sources and locations used are listed in Table 1.

The locations of the dataset sites on the Boyne River, Michigan are shown in Figure 1.

Dataset Collection and Use

Crowd-Hydrology Data. To test the decision tree rule set and select an optimal data QC metric, citizen science datasets were used. Datasets were obtained for the CH stations Michigan 1022 to 1026 from May 2014 to June 2018 from www.crowdhydrology.com/data.

USGS Gauges Data. The United States Geological Survey (USGS) data are one of the most widely used reference data source, which provides long-term stage monitoring data across the United States (U.S.). Texas USGS Gauge No. 08211503 was used with added random noise, as a simulation of citizen science data. The simulated dataset was compared to the original dataset's corresponding values, and selected controls applied for flagging atypical data values. Michigan USGS Gauge No. 04127800 was used as a reference dataset for the Michigan citizen science datasets.

The stage data were not available for the USGS Michigan gauge for the dates corresponding with the CH data. A stage-discharge relation was created to calculate the stage from the recorded discharge. The stream stage was calculated, and values predicted for the Michigan USGS gauge by creating a rating curve from the 2018-gauge height and discharge data (Wahl et al. 1995; USGS 2016). The rating curve was made by inserting a linear regression line with the equation in the form $y = ax + b$, where x is discharge and y is gauge height. It is given by the following equation:

$$y = 0.0059x + 1.9865, \quad (1)$$

was calculated with an R^2 value of 0.9622 and a root mean square error of 0.0578.

TABLE 1. Data source and location for each station.

Data source	Station ID	Latitude	Longitude
CH	MI1022	45.214508	−85.011725
CH	MI1023	45.203904	−84.972731
CH	MI1024	45.196873	−84.958077
CH	MI1025	45.157571	−84.921393
CH	MI1026	45.1714449	−84.876804
CocoRaHS	MI-CX-7	45.18639	−85.1475
USGS TX	08211503	27.8969652	−97.625551
USGS MI	04127800	45.10250676	−85.098112

Notes: CH, CrowdHydrology; USGS TX, United States Geological Survey Texas; USGS MI, USGS Michigan.

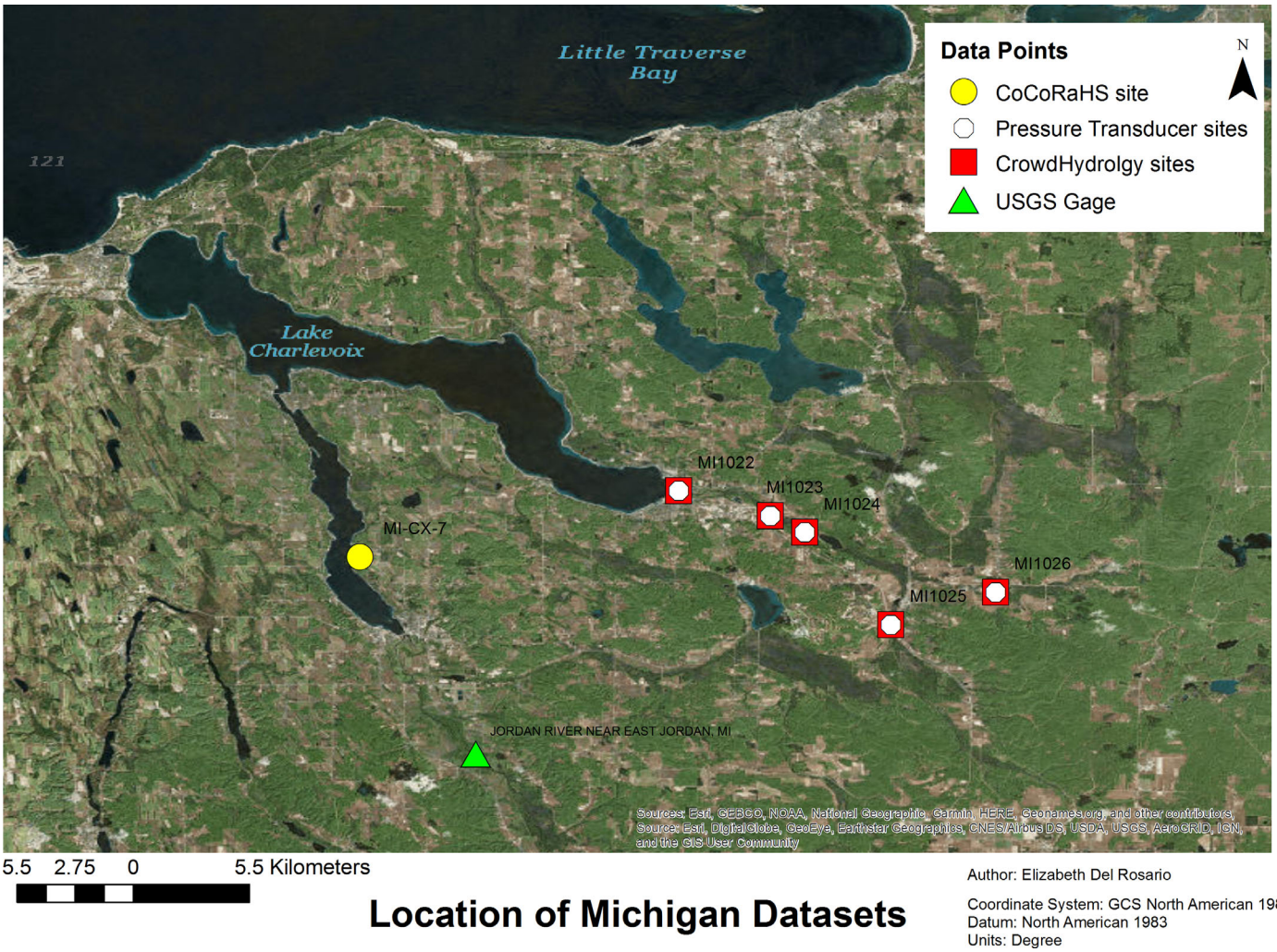

FIGURE 1. Location of Michigan sites on Boyne River.

Pressure Transducer Data. Pressure transducers (PT) were installed and corresponded with the CH Michigan data points on the Boyne River (PT1022–PT1026). These data obtained from the PT were used for qualifying uncertainty in the citizen science data and evaluating the classification capabilities of the decision tree by comparing with CH data.

Additional Datasets. The external factors used in this study were precipitation and temperature. These datasets were used for setting controls to determine if they can also be used as potential references to flag erroneous citizen science data. These additional datasets include local precipitation and air temperature.

Precipitation Data. Precipitation may impact the stream stage. To generate a continuous dataset over the study period and provide transferability for future studies, two sources of precipitation data were selected: the National Climatic Data Center (NCDC)-NOAA, and CoCoRaHS dataset, a national citizen science-based precipitation observation network. For the NCDC dataset, the record was chosen from the gauge nearest to the specific citizen science data. For CoCoRaHS data, the precipitation is recorded as point measurements and extrapolated to a county scale. These datasets reported cumulative daily precipitation in inches from April 1, 2014 to June 29, 2018.

Temperature Data. Temperature may have potential influence not only on the stage itself from evaporation but also on volunteer activities, which are closely related with the sampling frequency of citizen science data. The daily average temperature data were collected from NCDC-NOAA for the observation station closest in distance from the citizen science site gauges. The temperature dataset was reported in degrees Fahrenheit.

METHODS

A decision tree methodology was applied in an EXCEL platform to evaluate uncertainty in citizen science-based stream stage time-series data and to flag erroneous observations in the dataset. Multiple sources of datasets were utilized to determine the ruleset of the decision tree, and its classification ability evaluated based on CH stage data of Boyne River, Michigan.

Decision Tree

Decision tree methods, also called recursive partitioning, were developed to segment the target dataset into subdivisions based on the predefined controls for each branch (Friedl and Brodley 1997; Lemon et al. 2003). As one of the most commonly used prediction models, the decision tree has incomparable advantages for binary classification by facilitating user's comprehension and simplifying the classification processes (Anyanwu and Shiva 2009). The typical framework of a simple decision tree includes one input dataset, several test rules, and a set of categories, which correspond to the root, branches, and leaf nodes. To ensure one-way data flow and avoid loops in the decision tree, a node is only allowed to have one parent node. Based on this structure, the input dataset can be subdivided sequentially according to the controls and fall into a certain class in the end (Friedl and Brodley 1997). Recent research indicates that the decision tree is a valuable tool for data uncertainty analysis. Tami et al. (2018) presented a reliable tree construction whose prediction and split rule take the uncertainty of each quantitative observation into account. Ma et al. (2016) extended classical decision trees to generate a tree approach, which can not only handle uncertain data but also reduce uncertainty by querying the most valuable uncertain cases within the learning procedure.

The controls in a decision tree define the classification rules. By breaking a complex decision into a set of sequentially independent controls, decision trees implement data categorization in a multistage approach (Safavian and Landgrebe 1991). Every obtained case should satisfy the ruleset, which is composed of controls along the path from the root to the corresponding leaf. For the same task, different control combinations may lead to different conclusions and accuracy (Quinlan 1987).

Control Design. A simulated citizen science dataset of Texas was used to set the classification rules for each control and to assess the decision tree's performance quantitatively. The Texas dataset was created by randomly deleting data points to simulate irregular sampling frequency, and by adding erroneous values as noise. Two types of data, incorrect data and atypical values, were flagged by the decision tree for having high potential to be sources of uncertainty in citizen science stream stage measurements. Data flagged as incorrect should be removed before data analysis, and atypical values checked for accuracy.

Seven controls were designed to be tested by the decision tree.

Positive. The data point will be flagged if negative or zero value. The CH staffing gauges are placed in the riverbed and start at zero, therefore the stage height cannot be negative. The zero stage values could potentially be true values as in extreme situations, such as drought, and should be verified for accuracy.

Local Stability. The data point will be flagged if the distance between the point value and the local average is greater than three times the standard deviation. Sharp changes in time-series data could indicate potential inaccuracies on a local scale. Values with large local variations were flagged using the standard deviation from the average stage value. Misclassifications were reduced by applying moving windows of four sizes (3-, 5-, 7-, and 14-day). The moving windows reflect variation within different time scales and, were used to calculate local average Aver_i and standard deviation $S_i (i = 3, 5, 7, 14)$.

Sampling Frequency. Data points will be flagged if the sampling interval is over a set threshold. Data gaps can be produced in time-series data from sampling intervals, which potentially reduce the reliability of the dataset. A threshold of 3 days between observations was used for this study.

Comparison with Reference Dataset. Data points will be flagged where the absolute difference from the reference data minus the average difference in citizen science data is greater than the standard deviation ($x0007C; \text{dif} - \overline{\text{dif}} x0007C; > S$). Citizen science data were compared with the reference USGS gauge data to test for consistent trends. This was done as follows:

1. Calculating absolute difference (dif) between citizen science data and reference data.
2. Pairing citizen science data with reference data and calculating slope for each dataset. Data points where the slope k_{cs} trends different from the corresponding k_{USGS} was grouped as "temporary flagged" in the citizen science dataset.
3. Calculating average difference ($\overline{\text{dif}}$) and standard deviation (S) for temporary flagged group. If $x0007C; \text{dif} - \overline{\text{dif}} x0007C; > S_{\text{flag}}$, data point remained flagged, otherwise data point was unflagged.

4. Calculating average difference ($\overline{\text{dif}}$) and standard deviation (S) for the ungrouped citizen science dataset. Data points where $x0007C;\text{dif} - \overline{\text{dif}}x0007C; > S$ were flagged.

Sampling Time. Data points will be flagged if during a specified time frame as the quality of the citizen science stage data correlates with visual perception factors, such as darkness. It was assumed that during the early morning (0000–0359) and late night (2100–2359) there was diminished accuracy in the data measurement observation.

Precipitation. Stage data points will be flagged if the rank of its slope does not trend with corresponding precipitation slope. The precipitation intensity, which is classified by the precipitation rate seven-level of American Meteorological Society (1959), was used to determine the rank of precipitation and stage change. The lowest precipitation rate level was ranked as 1 with the highest rate being ranked at 7. Precipitation slopes and stage slopes were calculated and paired. It was assumed that changes in stage correspond with the intensity of precipitation.

Temperature. Data points with corresponding temperatures out of a specified range will be flagged. It was assumed that extreme temperatures correlate with sampling frequency due to impacts on human activities. For this study, the range 15°C–25°C was selected as adequate for citizen scientist data observations' validity.

The flag assignment rules are shown in Table 2. For each control, the results involve two classes, "Flag" and "Unflag," which correspond to Flag 1 and Flag 0.

TABLE 2. The flag assignment rules for each control.

	Control	Flag Assignment Rules				
1	Positive	$Stage > 0$				0
		$Stage \leq 0$				1
2	Local Stability	$\lvert Citizen\ Science\ data - Aver_i \rvert \leq 3 * S_i \quad i = 3,5,7,14$				0
		At least one of $\lvert Citizen\ Science\ data - Aver_i \rvert > 3 * S_i \quad i = 3,5,7,14$				1
3	Sampling Frequency	$\leq 3\ days$				0
		$> 3\ days$				1
4	Reference Comparison	$k_{CS} \times k_{USGS}$		$\lvert dif - \overline{dif} \rvert$		
		< 0	Temporary Flag	$> S_{flag}$		1
		$= 0$ Only one slope equals to 0	Temporary Flag	$\leq S_{flag}$	$> S$	1
		$= 0$ Both slopes are equal to 0	Temporary Unflag		$\leq S$	0
		> 0	Temporary Unflag		$\leq S$	0
5	Sampling Time	$21{:}00 - 3{:}59 (+1\ day)$				0
		$4{:}00 - 20{:}59$				1
6	Precipitation	$Rank_{precipitation} = Rank_{Citizen\ science\ data}$				0
		$Rank_{precipitation} \neq Rank_{Citizen\ science\ data}$				1
7	Temperature	$15°C - 25°C$				0
		$< 15°C\ or > 25°C$				1

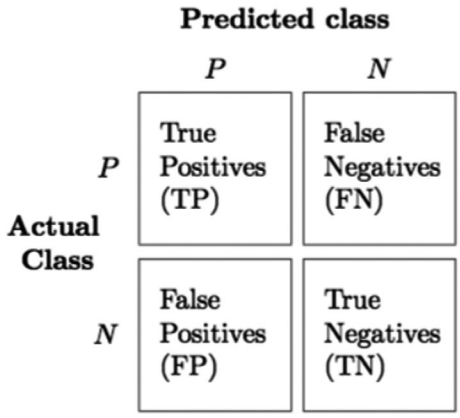

FIGURE 2. Binary confusion matrix.

Ruleset Generation. The ruleset is composed of ranked controls with the capability to categorize data points into appropriate classes. Every control was tested individually by simulation of citizen science data. To reflect the classification's results, a binary confusion matrix was utilized (Figure 2). Four performance indicators were considered to quantitatively estimate the classification accuracy of each control.

Precision, Accuracy, and Error Estimation. In this study, precision describes the proportion of the data points flagged by controls that should be flagged, according to the ruleset. Recall expresses the ability of controls to find flag-data in the dataset. Precision P, Recall r, F1-score F_1, and "Accuracy" were computed (Goutte and Gaussier 2005; Powers 2011). The equations of precision and recall are as follows:

$$P = \frac{TP}{TP + FP}, \tag{2}$$

$$r = \frac{TP}{TP + FN}. \tag{3}$$

The harmonic average of precision and recall yield the F1-score, which gives equal weight to both measures and avoids bias on extreme values. The F1-score was calculated using:

$$F_1 = 2 \times \frac{P \times r}{P + r}. \tag{4}$$

The "Accuracy" reflects the overall accuracy of classification and is computed by the following equation:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}. \tag{5}$$

In the above equation, FN represents data points that are misclassified into the unflagged class. To reduce the rate of FN, controls were selected to form the rules in which combinations tended to have complementary effects. FN was separated into nonpositive data and positive data with bias to establish the complementary effects of the controls and to determine their operation orders. The performance of selected control combinations was assessed with the confusion matrix and performance indicators on the individual controls.

Accuracy evaluation was conducted with the CH data from the five stations on the Boyne River, Michigan, as input datasets for the decision tree. To test the flexibility of the decision tree, USGS data from the nearest gauge and pressure transducer data from the same five stations on the Boyne River were used as references (Figure 1). Since two different collection methods where used, PT and staff gauge observations, to obtain the compared datasets, percent difference was used to calculate error (NCSU 2010). It was computed by the following equation:

$$\text{Percent difference} = \frac{\frac{|E_1 - E_2|}{E_1 + E_2}}{2} \times 100\%. \tag{6}$$

RESULTS AND DISCUSSION

The results section presents the formulated decision tree and its reliability. Improvements in the citizen science dataset are evaluated quantitatively based on this decision tree.

Control Selection

The confusion matrix allows visualization of the performance for each control. The classification

TABLE 3. Classification confusion matrix for simulated citizen science data.

Control		Actual class	Prediction class: Flag	Prediction class: Unflag
1	Positive	Flag	403	169
		Unflag	0	103,458
2	Local stability	Flag	96	476
		Unflag	725	103,209
3	Sampling frequency	Flag	0	572
		Unflag	1	103,457
4	Reference comparison	Flag	94	478
		Unflag	0	103,458
5	Sampling time	Flag	160	412
		Unflag	30,192	73,266
6	Precipitation	Flag	209	363
		Unflag	50,543	52,915
7	Temperature	Flag	334	238
		Unflag	60,603	42,855

TABLE 4. Classification performance estimation.

Control		Precision	Recall	F1-score	Accuracy
1	Positive	1	0.7045	0.8267	0.9984
2	Local stability	0.1169	0.1678	0.1378	0.9885
3	Sampling frequency	0	0	NA	0.9945
4	Reference comparison	1	0.1643	0.2823	0.9954
5	Sampling time	0.0053	0.2797	0.0103	0.7058
6	Precipitation	0.0041	0.3654	0.0081	0.5107
7	Temperature	0.0055	0.5839	0.0109	0.4152

TABLE 5. Flag-class data misclassified into unflag class.

		Misclassified flag-class data				
			Type			
Control		Total	Negative or 0		Bias	
1	Positive	169	0	0%	169	100%
2	Local stability	476	353	74.16%	123	25.84%
3	Sampling frequency	572	403	70.45%	169	29.55%
4	Reference comparison	478	373	78.03%	105	21.97%
5	Sampling time	412	308	74.76%	104	25.24%
6	Precipitation	363	228	62.81%	135	37.19%
7	Temperature	238	144	60.50%	94	39.50%

confusion matrix is shown in Table 3 for the simulated citizen science data from the USGS Texas gauge. Classification performance was estimated by four indicators, providing a quantitative comparison between different controls (Table 4).

Control orders relate to the efficiency and result of the decision tree. Misclassified types in FN control order are described in Table 5.

To establish an optimal ruleset, controls with complementary classifications effects were chosen. According to Tables 3 and 4, Control 3 (Sampling frequency) should be removed from the ruleset, since it shows the worst performance on identifying flag type data. Based on the F1-score and Accuracy, Control 1 (Positive), Control 4 (Reference Comparison) and Control 2 (Local stability) have high potential for flagging atypical values and should therefore be used in the ruleset. Based on the result of Table 5, among these three controls, Control 1 has preeminent capability to locate nonpositive values but fails with data containing biases. On the contrary, Control 4 and 2 have a low misclassification rate for biased data and a high error rate on nonpositive type data. Thus, the combination including Controls 1, 2, 4 potentially provides substantial information without large duplication and lead to satisfactory results of categorizing data into flag and unflag classes. The finalized structure decision tree from the optimized ruleset is shown in Figure 3.

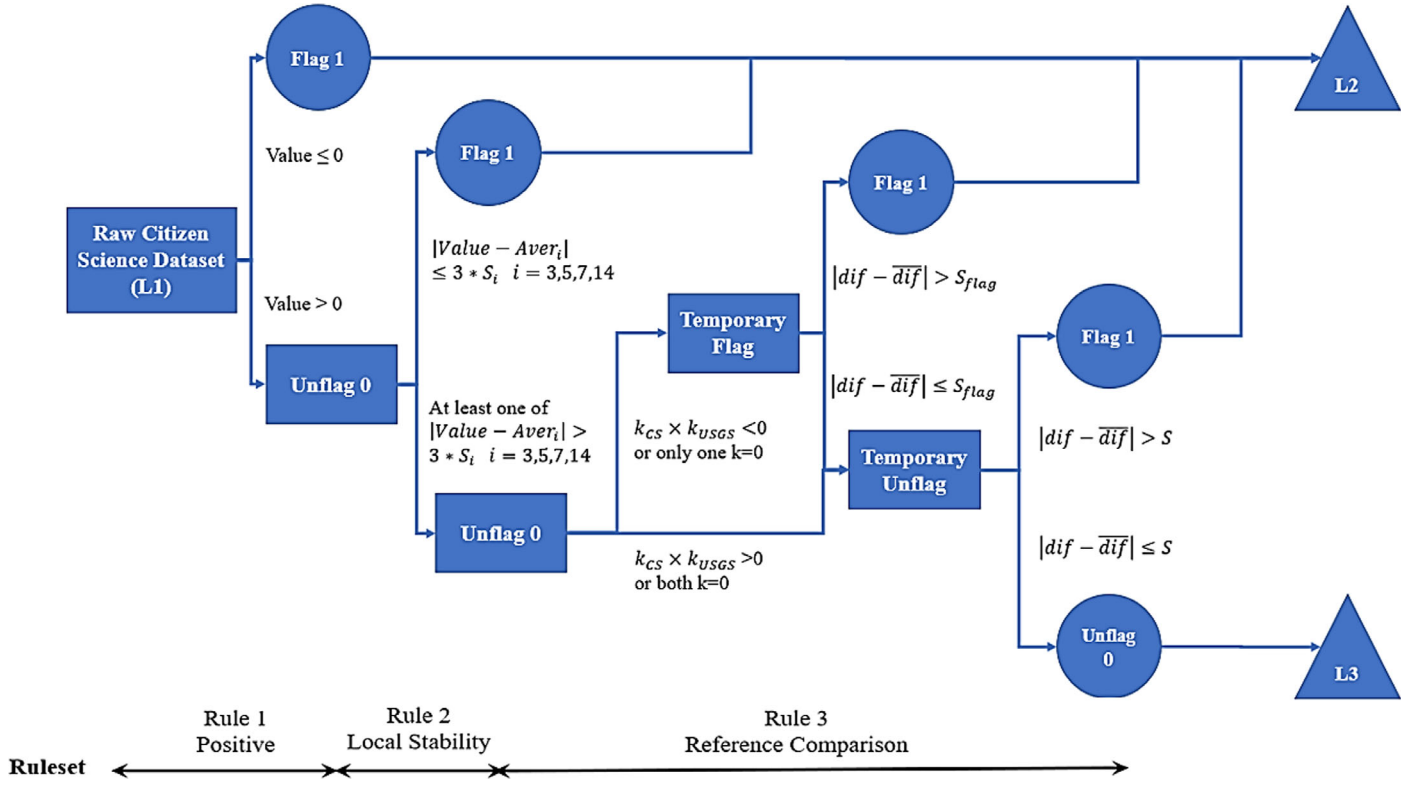

FIGURE 3. Decision tree structure.

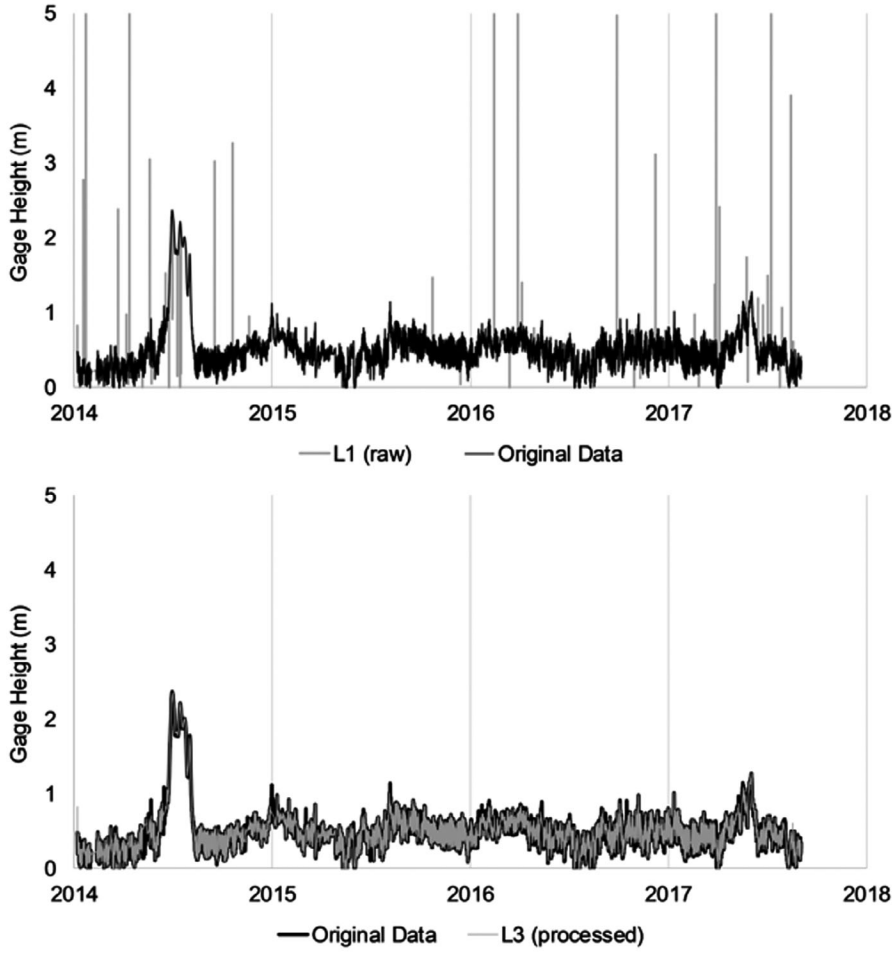

FIGURE 4. L1-raw (top), L3-processed (bottom) simulated citizen science dataset with original USGS Texas gauge dataset.

TABLE 6. Classification result of decision tree.

	Prediction class	
Actual class	**Flag**	**Unflag**
Confusion matrix		
Flag	572	0
Unflag	726	102,732

Precision	**Recall**	**F1-score**	**Accuracy**
Classification performance estimation			
0.4407	1	0.6118	0.9930

TABLE 7. Error calculations for Michigan CH and pressure transducers (PT) datasets.

Site	**Dataset**	**% Difference**
MI1022	L1	0.0706
	L2	0.0686
	L3	0.0717
MI1023	L1	0.1197
	L2	0.1173
	L3	0.1177
MI1024	L1	0.0266
	L2	0.0174
	L3	0.0335
MI1025	L1	0.1642
	L2	0.2473
	L3	0.1781
MI1026	L1	0.1661
	L2	0.7898
	L3	0.0527

Application of Decision Tree Methodology

The simulated citizen science dataset consisted of 104,029 (n) observations in the L1 (raw) dataset; 88,099 points flagged by the decision tree methodology in the L2 dataset; and 15,930 (n) observations in the L3 (processed) dataset. The original USGS Texas stage record observations (Figure 4) were modified to add noise in order to test the decision tree methodology. The decision tree methodology was then implemented to flag potentially erroneous data points to be

TABLE 8. Error estimates based on range ± meters for the CH and PT datasets.

Site	Dataset	Error (m)	Uncertainty (m)	Difference (m)
MI1022	L1	0.00282	0.00011	0.00071
	L3	0.00287	0.00037	0.00072
MI1023	L1	0.00478	0.00019	0.00120
	L3	0.00470	0.00060	0.00118
MI1024	L1	0.00106	0.00004	0.00027
	L3	0.00134	0.00018	0.00033
MI1025	L1	0.01613	0.00065	0.00164
	L3	0.00710	0.00091	0.00178
MI1026	L1	0.00657	0.00021	0.00166
	L3	0.00210	0.00022	0.00053

TABLE 9. Number of data points in each dataset under two-standard deviations.

Station ID	$L1_{2\ std}$	$L2_{2\ std}$	$L3_{2\ std}$
MI1022	15	0	15
MI1023	20	3	17
MI1024	7	0	7
MI1025	23	1	22
MI1026	13	2	11

TABLE 10. Accuracy comparison for Michigan CH based on different thresholds.

Site	Dataset	% Difference
MI1022	L1	0.0706
	$L3_{1\ std}$	0.0717
	$L3_{2\ std}$	0.0706
MI1023	L1	0.1197
	$L3_{1\ std}$	0.1177
	$L3_{2\ std}$	0.1162
MI1024	L1	0.0266
	$L3_{1\ std}$	0.0335
	$L3_{2\ std}$	0.0266
MI1025	L1	0.1642
	$L3_{1\ std}$	0.1781
	$L3_{2\ std}$	0.1173
MI1026	L1	0.1661
	$L3_{1\ std}$	0.0527
	$L3_{2\ std}$	0.0527

removed, resulting in a L3 dataset. The decision tree results for the simulated Texas data indicated high performance (Table 6).

Prediction Accuracy of the Decision Tree

Error estimates between the CH datasets and the pressure transducer (PT) datasets are shown in Table 7. The smallest error estimates for the L1 and L3 datasets are in blue. Percent difference was decreased at sites MI1023 and MI1026.

Table 8 shows that error is estimated in a range for the datasets, ±instead of %, we divide the table % values by 100 and result in ± meters. Using this method of error calculation, the error associated with the CH data is in the same range as the reported USGS staff gauge accuracy of ±0.01 ft (Office of Surface Water 1992; Lowry and Fienen 2013).

Removing data points to create the L3 dataset increased all error estimates for MI1024; most likely due to the small sample size that corresponded with the pressure transducer. This decrease in accuracy when the flagged points are removed may also be explained by mis-flagging caused by an over-narrow threshold. To improve the flagging results, other empirical threshold was tested. The threshold of temporary unflagged dataset in Control 4 (USGS reference comparison), was modified amplifying it to two times the standard deviations. Considering the extreme values were already removed in a previous step, it is reasonable for us to enlarge the accepted verge to reduce the mis-flagging in the remainder of the dataset.

Table 9 shows that how many points remain unflagged with two standard deviations. Table 10 shows that when more points are kept as unflagged data using two standard deviations vs. one, it improves the data quality estimates. The smaller values in comparison of $L3_{1\ std}$ and $L3_{2\ std}$ are marked in blue.

For MI1022 and MI1024, whose error and uncertainty were increased under one-standard-deviation threshold (there were no points mis-flagged, which avoids the loss of data quality caused by flagged data being removed). The results of MI1026 remained unchanged. The number of flagged points in MI1023 and MI1025 were reduced and their error decreased compared with those in the L3 under one standard deviation.

CONCLUSION

There are many crowdsourced databases involving various data types such as hydrology, precipitation, and water quality, covering the regions where no systematic monitoring existed previously. To improve the quality of citizen science data, a binary decision tree model has been generated to flag potentially erroneous data points. Optimal categorization rules were selected based on their performance for finding "incorrect record" (negative stage value) and "bias record" (extreme value). Considering the performance evaluated by precision, recall, F1-score and accuracy, "positive," "local stability," and "reference comparison": were shown to be the most appropriate rule set for data quality. The overall classification accuracy of the decision tree shows

potential. This research demonstrated the application of decision tree methodology to hydrological data and was shown to be an effective tool for finding errors or outliers in datasets. The methodology demonstrated in this study could be applied to a broader aspect and incorporated as a QC tool for data processing.

ACKNOWLEDGMENTS

We thank at the National Water Center: NOAA-Trey Flowers, Fernando Salas, Fred Ogden, and Ed Clark; USGS-Dave Blodgett and Martin Briggs; NCAR: Aubrey Dugger and Katelyn FitzGerald; CUAHSI: Jared Bales, Lauren Grimley, and Fernando Aristizabal. We also thank Ben Ruddell, Sagy Cohen, John Brackins, and Azbina Rahman, Nishani Moragoda. Di Wu thank her funding source CUAHSI, and her academic advisor Ruopu Li for his supports. Elizabeth Del Rosario acknowledge her funding source NOAA EPP CCME Program (National Oceanic and Atmospheric Administration, Office of Education Educational Partnership Program Award [NA16SEC4810009]).

AUTHORS' CONTRIBUTIONS

Di Wu: Data curation; methodology; validation; writing-original draft; writing-review & editing. **Elizabeth A. Del Rosario:** Data curation; methodology; validation; writing-original draft; writing-review & editing. **Christopher Lowry:** Data curation; methodology; writing-review & editing.

LITERATURE CITED

American Meteorological Society. 1959. *Glossary of Meteorology*. Boston, MA: American Meteorological Society.

Anyanwu, M.N., and S.G. Shiva. 2009. "Comparative Analysis of Serial Decision Tree Classification Algorithms." *International Journal of Computer Science and Security* 3 (3): 230–40.

Davids, J.C., M.M. Rutten, A. Pandey, N. Devkota, W.D. van Oyen, R. Prajapati, and N. van de Giesen. 2019. "Citizen Science Flow — An Assessment of Simple Streamflow Measurement Methods." *Hydrology and Earth System Sciences* 23 (2): 1045–65.

Dickinson, J.L., J. Shirk, D. Bonter, R. Bonney, R.L. Crain, J. Martin, T. Phillips, and K. Purcell. 2012. "The Current State of Citizen Science as a Tool for Ecological Research and Public Engaugement." *Frontiers Ecology & Environment* 10 (6): 291–97. https://doi.org/10.1890/110236.

Fienen, M.N., and C.S. Lowry. 2012. "Social.Water — A Crowdsourcing Toll for Environmental Data Acquisition." *Computers & Geosciences* 49: 164–69. https://doi.org/10.1016/j.cageo.2012.06.015.

Friedl, M.A., and C.E. Brodley. 1997. "Decision Tree Classification of Land Cover from Remotely Sensed Data." *Remote Sensing of Environment* 61 (3): 399–409. https://doi.org/10.1016/S0034-4257(97)00049-7.

Gauchat, G. 2012. "Politicization of Science in the Public Sphere: A Study of Public Trust in the United States, 1974 to 2010." *American Sociological Review* 77 (2): 167–87. https://doi.org/10.1177/0003122412438225.

Goutte, C., and E. Gaussier. 2005. "A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation." *European Conference on Information Retrieval*, edited byC. Goutte and E. Gaussier, 345–59. Berlin, Heidelberg: Springer.

Hemmi, A., and I. Graham. 2014. "Hacker Science versus Closed Science: Building Environmental Monitoring Infrastructure." *Information, Communication & Society* 17 (7): 830–42. https://doi.org/10.1080/1369118X.2013.848918.

Kampf, S., B. Strobl, J. Hammond, A. Anenberg, S. Etter, C. Martin, K. Puntenney-Desmond, J. Seibert, and I. van Meerveld. 2018. "Testing the Waters: Mobile Apps for Crowdsourced Streamflow Data." *Eos* 99: 30–34. https://doi.org/10.1029/2018EO096355.

Law, E., K.Z. Gajos, A. Wiggins, M. Gray, and A. William. 2017. "Crowdsourcing as a Tool for Research: Implications of Uncertainty." *Proceedings of the ACM*,1544–61. https://doi.org/10.1145/2998181.2998197.

Le Coz, J., A. Patalano, D. Collins, N.F. Guillén, C.M. García, G.M. Smart, J. Bind et al. 2016. "Crowdsourced Data for Flood Hydrology: Feedback from Recent Citizen Science Projects in Argentina, France and New Zealand." *Journal of Hydrology* 541: 766–77. https://doi.org/10.1016/j.jhydrol.2016.07.036.

Lemon, S.C., J. Roy, M.A. Clark, P.D. Friedmann, and W. Rakowski. 2003. "Classification and Regression Tree Analysis in Public Health: Methodological Review and Comparison with Logistic Regression." *Annals of Behavioral Medicine* 26 (3): 172–81. https://doi.org/10.1207/S15324796ABM2603_02.

Lowry, C.S., and M.N. Fienen. 2013. "CrowdHydrology: Crowdsourcing Hydrologic Data and Engaging Citizen Scientists." *Ground Water* 51 (1): 151–56. https://doi.org/10.1111/j.1745-6584.2012.00956.x.

Lowry, C.S., M.N. Fienen, D.M. Hall, and K.F. Stepenuck. 2019. "Growing Pains of Crowdsourced Stream Stage Monitoring Using Mobile Phones: The Development of CrowdHydrology." *Frontiers in Earth Science* 7: 128. https://doi.org/10.3389/feart.2019.00128.

Ma, L., S. Destercke, and Y. Wang. 2016. "Online Active Learning of Decision Trees with Evidential Data." *Pattern Recognition* 52: 33–45.

McCormick, S. 2012. "After the Cap: Risk Assessment, Citizen Science and Disaster Recovery." *Ecology and Society* 17 (4). https://doi.org/10.5751/ES-05263-170431.

NCSU (North Carolina State University). 2010. "Appendix B: Percent Error and Percent Difference." In *Labs for Colleges Physics Mechanics* (Second Edition). www.webassign.net/labsgraceperiod/ncsulcpmech2/appendices/appendixB/appendixB.html.

Njue, N., J.S. Kroese, J. Gräf, S.R. Jacobs, B. Weeser, L. Breuer, and M.C. Rufino. 2019. "Citizen Science in Hydrological Monitoring and Ecosystem Services Management: State of the Art and Future Prospects." *Science of the Total Environment* 693 (13): 133531.

NOAA (National Oceanic and Atmospheric Association). 2018. "Citizen Science and Crowdsourcing." http://www.noaa.gov/office-education/citizen-science-crowdsourcing.

Office of Surface Water. 1992. "Technical Memorandum No. 93.07, Policy Statement on Stage Accuracy." United States Geological Survey. https://water.usgs.gov/admin/memo/SW/sw93.07.html.

Poelen, J.H., J.D. Simons, and C.J. Mungall. 2014. "Global Biotic Interactions: An Open Infrastructure to Share and Analyze Species-Interaction Datasets." *Ecological Informatics* 24: 148–59. https://doi.org/10.1016/j.ecoinf.2014.08.005.

Powers, D.M. 2011. "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation." *Journal of Machine Learning Technologies* 2 (1): 37–63. https://bioinfopublication.org/files/articles/2_1_1_JMLT.pdf.

Quinlan, J.R. 1987. "Generating Production Rules from Decision Trees." *International Joint Conferences on Artificial Intelligence* 87: 304–07.

Safavian, S.R., and D. Landgrebe. 1991. "A Survey of Decision Tree Classifier Methodology." *IEEE Transactions on Systems, Man, and Cybernetics* 21 (3): 660–74. https://doi.org/10.1109/21.97458.

Seibert, J., B. Strobl, S. Etter, P. Hummer, and H.I. van Meerveld. 2019. "Virtual Staff Gauges for Crowd-Based Stream Level Observations." *Frontiers in Earth Science* 7: 70.

Sullivan, B.L., C.L. Wood, M.J. Iliff, R.E. Bonney, D. Fink, and S. Kelling. 2009. "eBird: A Citizen-Based Bird Observation Network in the Biological Sciences." *Biological Conservation* 142 (10): 2282–92. https://doi.org/10.1016/j.biocon.2009.05.006.

Tami, M., M. Clausel, E. Devijver, E. Gaussier, and J.M. Aubert. 2018. "Decision Tree for Uncertainty Measures." JDS 2018: 50èmes Journées de statistique, May 2018, Paris-Saclay, France. ffhal-01815637.

Toivanen, T., S. Koponen, V. Kotovirta, M. Molinier, and P. Chengyuan. 2013. "Water Quality Analysis Using an Inexpensive Device and a Mobile Phone." *Environmental Systems Research* 2 (1): 9. https://doi.org/10.1186/2193-2697-2-9.

USGS (United States Geological Survey). 2016. "How Streamflow is Measured Part 3: The Stage-Discharge Relation." https://water.usgs.gov/edu/streamflow3.html.

Wahl, K.L., W.O. Thomas, Jr., and R.M. Hirsch. 1995. "Stream-Gaging Program of the U.S. Geological Survey." https://pubs.usgs.gov/circ/circ1123/collection.html.

Weeser, B., J.S. Kroese, S.R. Jacobs, N. Njue, Z. Kemboi, A. Ran, M.C. Rufino, and L. Breuer. 2018. "Citizen Science Pioneers in Kenya — A Crowdsourced Approach for Hydrological Monitoring." *Science of the Total Environment* 631: 1590–99.