Assessing the Trustworthiness of Crowdsourced Rainfall Networks: A Reputation System Approach

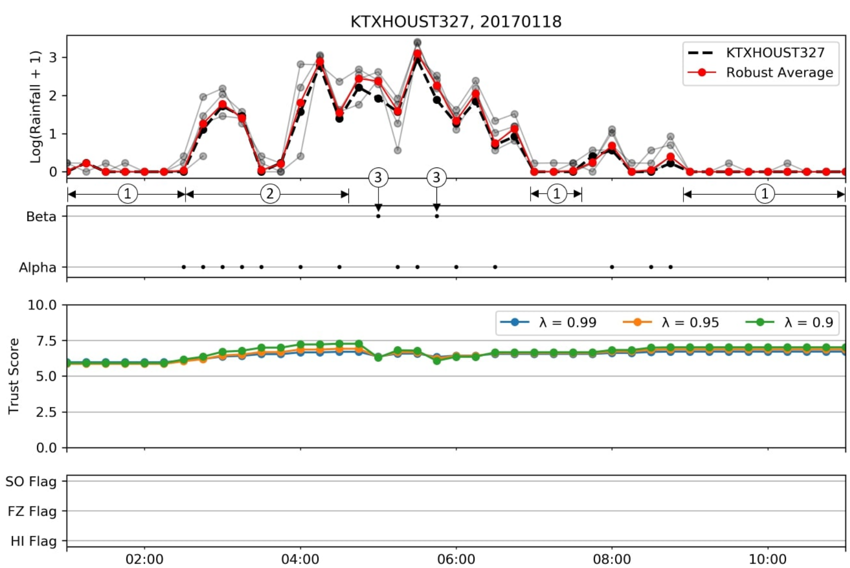
Alexander Byron Chen[1], Madhur Behl[1], and Jonathan L Goodall[1]

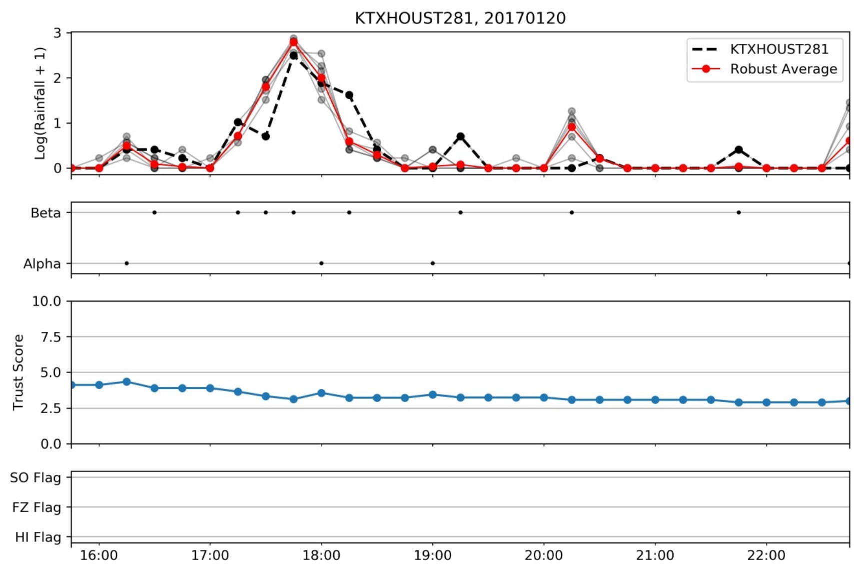
[1]University of Virginia

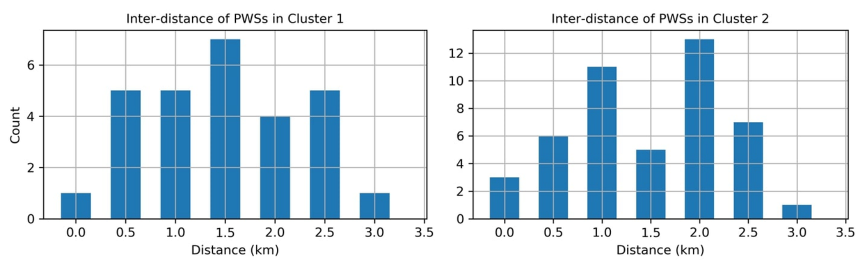
November 24, 2022

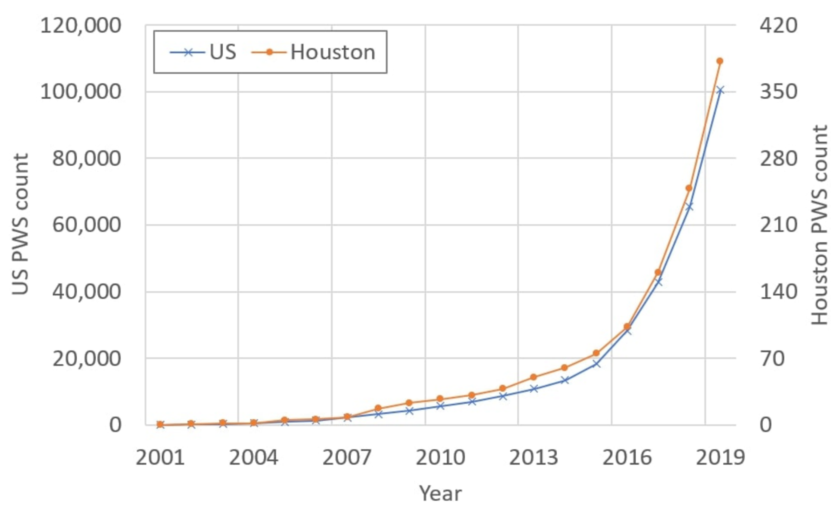
Abstract

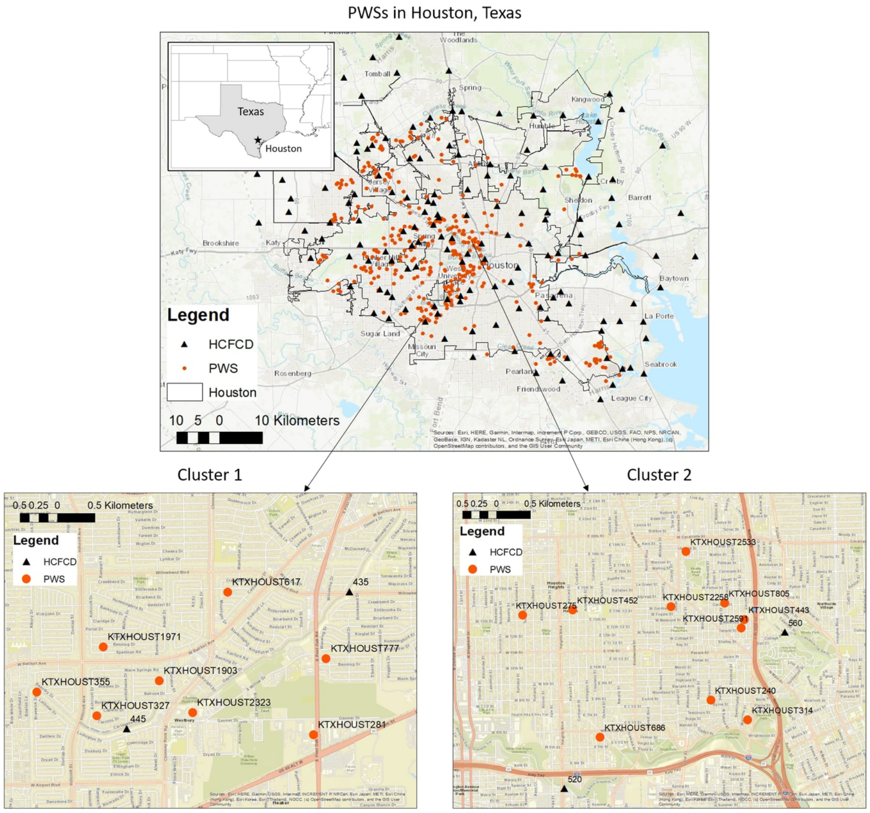
High resolution and accurate rainfall information is essential to modeling and predicting hydrological processes. Crowdsourced personal weather stations (PWSs) have become increasingly popular in recent years and can provide dense spatial and temporal resolution in rainfall estimates. However, their usefulness is limited due to a lack of trust in crowdsourced data compared to traditional data sources. Using crowdsourced PWSs data without an evaluation of its trustworthiness can result in inaccurate rainfall estimates as PWSs may be poorly maintained or incorrectly sited. In this study, we advance the Reputation System for Crowdsourced Rainfall Networks (RSCRN) to bridge this trust gap by assigning dynamic trust scores to the PWSs. Using rainfall data collected from 18 PWSs in two dense clusters in Houston, Texas USA as a case study, the results show that using RSCRN-derived trust scores can increase the accuracy of 15-min PWS rainfall estimates when compared to rainfall observations recorded at city's high-fidelity rainfall stations. Overall, RSCRN rainfall estimates improved for 77% (48 out of 62) of the analyzed storm events, with a median RMSE improvement of 27.3%. Compared to an existing PWS quality control method, results showed that while 13 (21%) storm events had the same performance, RSCRN improved rainfall estimates for 78% of the remaining storm events (38 out of 49), with a median RMSE improvement of 13.4%. Using RSCRN-derived trust scores can make the rapidly growing network of PWSs a more useful resource for urban flood management, greatly improving knowledge of rainfall patterns in areas with dense PWSs.




KTXHOUST281, 20170120
Log(Rainfall + 1)
KTXHOUST281
Robust Average
Beta
Alpha
Trust Score
SO Flag
FZ Flag
HI Flag
16:00
17:00
18:00
19:00
20:00
21:00
22:00


Inter-distance of PWSs in Cluster 1
Inter-distance of PWSs in Cluster 2
Count
Distance (km)


US
Houston
US PWS count
Houston PWS count
Year


PWSs in Houston, Texas
Texas
Houston
Legend
HCFCD
PWS
Houston
10 5 0 10 Kilometers
Cluster 1
Cluster 2
0.5 0.25 0 0.5 Kilometers
KTXHOUST617
435
KTXHOUST1971
KTXHOUST777
KTXHOUST1903
KTXHOUST355
KTXHOUST327
445
KTXHOUST2323
KTXHOUST281
KTXHOUST2533
KTXHOUST275
KTXHOUST452
KTXHOUST2258
KTXHOUST805
KTXHOUST443
KTXHOUST2591
560
KTXHOUST240
KTXHOUST314
KTXHOUST686
520

Assessing the Trustworthiness of Crowdsourced Rainfall Networks: A Reputation System Approach

Alexander B. Chen[1], **Madhur Behl**[1,2], **Jonathan L. Goodall**[1,2]

[1]Department of Engineering Systems and Environment, University of Virginia, Charlottesville, VA, USA
[2]Department of Computer Science, University of Virginia, Charlottesville, VA, USA

Key Points:

- We present a reputation system framework to measure the trustworthiness of crowdsourced personal weather stations (PWSs).
- PWSs are assigned a trust score based on their consensus with rainfall measured at neighboring stations.
- The accuracy of rainfall estimates based on crowdsourced PWSs can be improved by excluding PWSs with low trust scores.

Corresponding author: Jonathan L. Goodall, goodall@virginia.edu

Abstract

High resolution and accurate rainfall information is essential to modeling and predicting hydrological processes. Crowdsourced personal weather stations (PWSs) have become increasingly popular in recent years and can provide dense spatial and temporal resolution in rainfall estimates. However, their usefulness is limited due to a lack of trust in crowdsourced data compared to traditional data sources. Using crowdsourced PWSs data without an evaluation of its trustworthiness can result in inaccurate rainfall estimates as PWSs may be poorly maintained or incorrectly sited. In this study, we advance the Reputation System for Crowdsourced Rainfall Networks (RSCRN) to bridge this trust gap by assigning dynamic trust scores to the PWSs. Trust scores can be used when estimating rainfall for applications such as real-time flood management within urban areas with dense networks of PWSs. Using rainfall data collected from 18 PWSs in two dense clusters in Houston, Texas USA as case study, the results show that using RSCRN-derived trust scores can increase the accuracy of 15-min PWS rainfall estimates when compared to rainfall observations recorded at city's high-fidelity rainfall stations. Overall, RSCRN rainfall estimates improved for 77% (48 out of 62) of the analyzed storm events, with a median RMSE improvement of 27.3%. Compared to an existing PWS quality control method, results showed that while 13 (21%) storm events had the same performance, RSCRN improved rainfall estimates for 78% of the remaining storm events (38 out of 49), with a median RMSE improvement of 13.4%. Using RSCRN-derived trust scores can make the rapidly growing network of PWSs a more useful resource for urban flood management, greatly improving knowledge of rainfall patterns in areas with dense PWSs.

1 Introduction

Flooding is becoming commonplace in cities and communities worldwide, causing severe damage and loss of property (Wilby & Keenan, 2012; Salman & Li, 2018). As a result of climate change, rainfall extremes are expected to become more intense and highly heterogeneous (Ohba & Sugimoto, 2019; Sharma et al., 2018). Floods triggered by these increased storms often exhibit large variability both in space and time, especially in urban areas with a large portion of impervious surface (Quinn et al., 2019; Cristiano et al., 2017). Although recent advancements in computational power and modeling approaches have made it possible to accurately model flooding at increasingly high resolution (Saksena et al., 2019; Zahura et al., 2020; Shen et al., 2019; Mosavi et al., 2018; Savage et al., 2016), these models require measured rainfall observations as input at high spatial and temporal resolutions. However, the current resolution of observations through traditional rainfall networks is typically insufficient, or even unavailable, for certain flood-prone regions (Sadler et al., 2018; Cristiano et al., 2017; Zhu et al., 2018).

Traditionally, rainfall observations are obtained from gauges managed by federal or municipal agencies. These rain gauges, which we refer to as *high-fidelity rainfall stations* in this study, provide accurate measurements as they are installed and maintained by experts, but are limited in coverage (Villarini et al., 2008; Overeem et al., 2013). An alternative to rain gauges is the use of weather radars. However, radar rainfall is derived indirectly from radar reflectivity observed at certain heights in the atmosphere, which may not accurately represent rainfall at the ground level (Smith et al., 1996) and requires calibration with ground gauges (Krajewski & Smith, 2002). Recent advancement of dual-polarization technologies in weather radar addresses some of these limitations of using radar. However, further improvements of rainfall estimation using dual-polarization radars are needed. For example, range-dependent sampling errors and the uncertainties in identifying hydrometeor types with radar measurements may introduce larger bias in rainfall estimates, which cannot be easily corrected with ground gauges (Cunha et al., 2015).

Crowdsourcing could offer a potential solution to the need of high resolution and accurate rainfall estimates. Crowdsourcing is a broad term whereby data are obtained

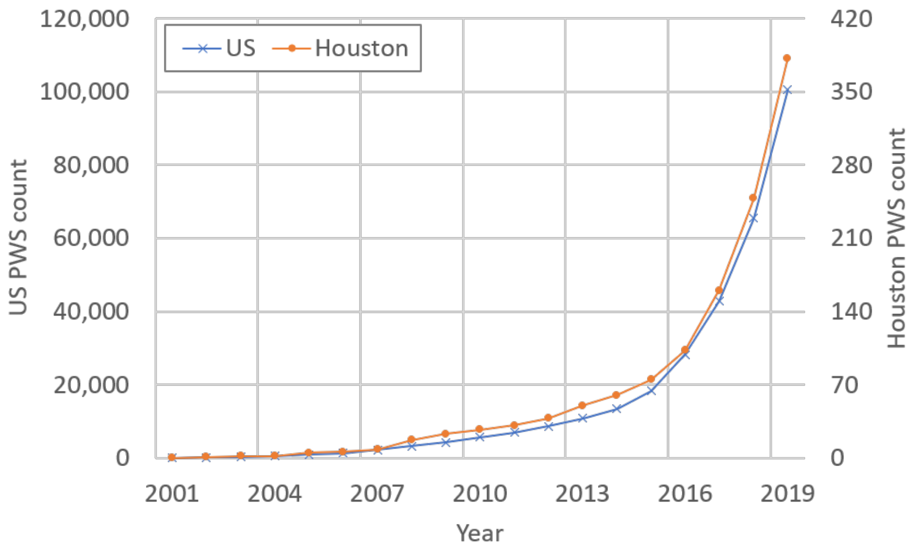


Figure 1. The number of Personal Weather Stations in Weather Underground network in the US and Houston, Texas has been growing exponentially in past 20 years.

through open calls to the general public for data collection, resulting in increased data coverage, but introducing the challenges associated with the data being collected by non-experts (Estellés-Arolas & González-Ladrón-De-Guevara, 2012). PWS are user-friendly and affordable off-the-shelf weather stations installed and maintained by individuals that offer a means for crowdsourcing weather data including rainfall observations (Gharesifard & Wehn, 2016). PWSs data can be easily shared through services such as Weather Underground, which enables real-time data gathering, integration, and visualization of weather data collected across a world wide network of PWSs via their online platforms and mobile applications. The growing adoption of PWSs in recent years has made crowdsourcing a powerful opportunity to supplement existing rainfall networks (Muller et al., 2015; de Vos et al., 2017; P. Yang & Ng, 2017; Weeser et al., 2019; Lowry & Fienen, 2013). Importantly, this crowdsourced data is growing rapidly, making it an increasingly valuable resource for hydrologists (Zheng et al., 2018). Based on our review of the Weather Underground data archive, the number of PWSs in the US has increased exponentially from 7,000 to 100,000 from 2010 to 2019 (Figure 1). In Houston, Texas, for example, the number of PWSs has grown from 99 to 382 over the three year period 2016 to 2019 (Figure 1), which equates to an increased density from 0.06 to 0.24 PWSs per square kilometer. If such exponential growth continues, the density of PWSs in populated areas in the US could reach one PWS per square kilometers in five years, which exceeds recommended spatial resolutions for rainfall observations required for urban hydrology (Berne et al., 2004; Fletcher et al., 2013).

The increased adoption of PWSs can be attributed to the openness of the crowdsourced networks that allows anyone to act as a data contributor. Such openness, however, also introduces challenges in assuring accurate data (Bell et al., 2013; Muller et al., 2015; Meier et al., 2017; Chapman et al., 2017). Crowdsourced networks are typically lightly controlled networks with few constraints and limited quality control processes. As a result, people have higher levels of confidence in data collected from high-fidelity rainfall stations and there are fewer sources of error in their observations compared to crowdsourced data (Hunter et al., 2013; Cox, 2011). PWSs can experience device errors, like high-fidelity rainfall stations, but can also suffer from compromised setups, lack of routine maintenance, and other sources of error that are less common in high-fidelity rainfall stations (de Vos et al., 2017; Meier et al., 2017). For example, improper installation of PWSs, such as siting the station under a tree canopy or next to a building, may lead to consistently incorrect readings. Likewise, the owner of the PWS might not routinely maintain and calibrate the device, which could lead to sensor drift and faulty observations. Beyond these cases, it is also possible in open crowdsourced networks that people might deliberately manipulate data to produce misleading evidence (Huang et al., 2014; Sanchez et al., 2018). Therefore, a method to evaluate the trustworthiness of crowdsourced PWSs is needed before this rich and growing dataset can be effectively used in decision making.

One approach for addressing this problem with crowdsourced PWS data would be to adopt quality control and quality assurance (QA/QC) methods to detect, flag or remove doubtful and erroneous data based on certain rules and thresholds (Estévez et al., 2011; Fiebrich et al., 2010; Blenkinsop et al., 2017; de Vos et al., 2019). If other data from more trusted sources is available, then another method would be to evaluate the quality of crowdsourced rainfall data by direct comparison with these more trusted sources (de Vos et al., 2017; Muller et al., 2015). Existing methods, however, may not adequately address the needs of crowdsourced weather and, specifically rainfall, observation. QA/QC methods designed for high-fidelity stations tend to focus on outlier detection that presumes a certain source of error (sensor malfunction) and may be less able to detect other sources of error (poor sensor siting or installation). For example, the Weather Underground designates a PWS as “Gold Star Weather Station” if it passes basic quality control criteria such as data validity and a sensor failure checks over the prior five days (The Weather Channel, 2018). Direct comparison with trusted data sources presumes that such data is available, but PWSs have reached a density of observation in space and time that cannot be matched with other, more trusted, measurement methods. There is a need and opportunity, therefore, to innovate on methods for assessing the data generated by PWSs at scale so that trustworthy stations can be used more confidently in decision making and, just as importantly, untrustworthy stations can be reported to owners with suggestions for improving data quality so that the overall observation network reaches its full potential.

In this study, we explore the use of reputations systems as an approach for measuring the trustworthiness of rainfall observations from PWSs. Reputation systems are commonly used to build trust between participants and foster good behavior in online crowdsourced systems (Jøsang et al., 2007). For example, online markets such as eBay and Amazon use reputation systems to enhance the buying and selling experiences. Such systems aggregate sellers’ past behavior and represent it as a trustworthiness rating for buyers to rely on (Resnick et al., 2000). Reputation systems have also been used for citizen science and crowdsdourced data. H. Yang et al. (2013) designed a reputation system framework for enhancing the data reliability of citizen science environmental acoustic data. Silvertown et al. (2015) utilized a reputation system to motivate and reward participants of a crowdsourced species identification website that improved the accuracy of species determinations. Huang et al. (2014) proposed a reputation system framework using the Gompertz function to compute device reputation score based on the trustworthiness of the contributed data in participatory sensing applications. However, limited work has investigated the use of reputation systems for crowdsourced PWS networks.

In our previous work, we presented an initial version of a system called the Reputation System for Crowsourced Rainfall Network (RSCRN) (Chen et al., 2018) to assign trust scores to PWSs. In this paper, we significantly enhance RSCRN and evaluate the method for storm events using Houston, Texas as a case study. The research questions guiding this work are: (i) How can we systematically evaluate the trustworthiness of crowdsourced PWSs? and (ii) To what extent could a reputation system approach improve rainfall estimates derived from PWSs?

The remainder of the paper is organized as follows. Section 2 describes the detail of the RSCRN algorithm and methods to evaluate the RSCRN for storm events. Section 3 provides a description of the study area, data used in the study, as well as storm events selection process. The results and discussion of this study are presented in Sections 4 and 5, followed by conclusions in Section 6.

2 Material and Methods

2.1 Data preparation

The RSCRN method begins with a crowdsourced rainfall network in a specific region having N PWSs. Given an analysis period of interest (say X time steps), the rainfall observations from these N PWSs can be collected into a matrix P

$$P_{i,j} = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,N} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ p_{X,1} & p_{X,2} & \cdots & p_{X,N} \end{pmatrix}$$

where $p_{i,j}$ is a rainfall observation measured at time step i and PWS j. This matrix P, with X rows for the rainfall observations and N columns of PWSs, will be used as the input for RSCRN.

2.2 RSCRN Algorithm

The RSCRN algorithm consists of three steps: *Cluster*, *Consensus* and *Score* (Algorithm 1). The objective of RSCRN is to evaluate the trustworthiness of PWSs based on their consensus with rainfall measured at neighboring stations. The *Cluster* step is to find clusters of neighboring PWSs. Next, the *Consensus* step is used to identify PWSs with rainfall observations that deviate from a cluster's consensus. Finally, the *Score* step uses the degree of deviation from consensus to assign a new trust score to each PWS on a given time step that represents the trustworthiness of that PWSs. Further detail for each step in the algorithm follows (Algorithm 1).

Algorithm 1 RSCRN algorithm

Cluster step:
Input PWS rainfall observation matrix $P(i,j) \mid i \in [1:X];\ j \in [1:N]$
X: number of time steps, N: number of PWSs
Output Clustered sub-dataset matrix $D^k_{M_k} \subset P \mid k \in [1:K]$
K: number of clusters, M_k: number of PWSs in the k-th cluster
Consensus/Score step:
Input Clustered sub-dataset matrix $D^k_{M_k}(i,j) \mid i \in [1:X];\ j \in [1:M_k]$
Output Trust score matrix $T_{i,j} \mid i \in [1:X];\ j \in [1:M_k]$

```
for k ∈ [1 : K] do
    for i ∈ [1 : X] do
        for j ∈ [1 : M_k] do
            Compute robust weight w_{i,j} using Equations 1 - 4
            Compute cooperative metric C_{i,j} using Equations 5
            Compute trust score T_{i,j} using Equations 8 - 10
```

Python codes for the RSCRN algorithm and datasets used in this study are available from Hydroshare (`https://www.hydroshare.org/resource/`

`cf7796cdeace42818dbbd7f95f8e1872/`). For the Weather Underground, the P matrix can be populated for a region using their Application Programming Interface (API). Example code for this process is also provided as a resource in Hydroshare with the same link as above. This code requires a Weather Underground key to use the API, which at the time of this writing can be obtained by connecting a PWS to the Weather Underground platform.

2.2.1 Cluster

Different methods can be used to define PWS clusters in RSCRN. In this work, we take a simple approach of defining clusters based on geographic proximity of stations. Thus, we used a buffering tool in a Geographic Information System (GIS) to identify clusters that consist of PWSs within a fixed distance to other neighboring PWSs and high-fidelity, government-operating rainfall stations that will be used for evaluation. We have explored other methods for clustering as well including k-means where clusters are identified not only based on geographic position, but also other factors including elevation (Chen et al., 2018). It is important to ensure there are sufficient PWSs (at least four and preferably five or more PWSs) in each cluster actively reporting rainfall observations during the analysis period of interest, because the RSCRN algorithm relies on the consensus of PWSs rainfall observations in a cluster. More active PWSs in a cluster will likely result in a more reliable consensus. Starting from the input matrix P, the resulting clustered matrices are denoted by $D^k_{M_k}$, where k is the k-th cluster, and M_k is the number of PWSs in the k-th cluster. These matrices will be the input for the consensus step.

2.2.2 Consensus

The input to the consensus step are clustered sub-datasets $D^k_{M_k}$, where each sub-dataset contains rainfall observations from PWSs that fall within the same cluster. Assuming the k-th clustered sub-dataset has m PWSs, this sub-dataset will be a matrix D^k_m

$$D^k_m(i,j) = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,m} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{X,1} & p_{X,2} & \cdots & p_{X,m} \end{pmatrix}$$

where $p_{i,j}$ represents the rainfall observation PWS j within the cluster k measured at time i. For the clustered sub-datasets $D^k_{M_k}$ $(k = 1, 2, \ldots, K)$, the consensus step computes a *cooperative metric* (denoted as $C_{i,j}$, which has the same dimension as D^k_m) based on the rainfall observations for each time-step $(i = 1, 2, \ldots, X)$ and each PWS $(j = 1, 2, \ldots, m)$.

We use the robust averaging algorithm (Chou et al., 2013) as the method for estimating a cluster's consensus. We selected this method for its effectiveness and efficiency in similar applications for wireless sensor networks and participatory sensing (Ganeriwal et al., 2008; Huang et al., 2014). Robust averaging is a type of weighted average method that is less affected by values that deviated from the average. For each time step i, this iterative algorithm works as follows

1. First, assign an initial (uniform) weight to every PWS j at iteration $l = 1$

$$w^{l=1}_{i,j} = \frac{1}{m} \tag{1}$$

   where m is the number of PWSs in the clustered sub-dataset D^K_m.

2. Next, compute the robust average RA_i^l, such that

$$RA_i^l = \sum_{j=1}^{m} w_{i,j}^l \cdot p_{i,j} \tag{2}$$

where $p_{i,j}$ is the rainfall observation of PWS j for time step i.

3. Next, compute the squared difference of PWS j's rainfall observation $p_{i,j}$ from the robust average RA_i^l

$$v_{i,j}^l = (p_{i,j} - RA_j^l)^2. \tag{3}$$

4. Finally, compute the new robust weight at iteration $l+1$

$$w_{i,j}^{l+1} = (\frac{1}{\frac{v_{i,j}^l}{\sum_{j=1}^{m} v_{i,j}^l} + \epsilon})/(\sum_{i=1}^{m} \frac{1}{\frac{v_{i,j}^l}{\sum_{j=1}^{m} v_{i,j}^l + \epsilon}}). \tag{4}$$

The algorithm continues iterating until the convergence $|w_{i,j}^l - w_{i,j}^{l+1}| < \nu$ is achieved, i.e., the robust weights converge to a value with difference less than ν. Note that the ϵ is a small positive constant that is set to 0.1, determined by trial and error based on the convergence of the algorithm (Chou et al., 2013).

The *cooperative metric* is then defined as

$$C_{i,j} = \frac{w_{i,j} - \overline{W_i}}{\sigma(W_i)} \tag{5}$$

where $\overline{W_i}$ and $\sigma(W_i)$ are the average and standard deviation, respectively, of the i-th row of the robust weight matrix. This metric represents the level of deviation of the final robust weight from the initial weight. A positive cooperative metric indicates agreement with the consensus (robust average) within the cluster, while a negative cooperative metric represents disagreement with the consensus. The resulting *cooperative metric* is used as the input for score step.

We further extended the algorithm to accommodate two data exception cases: (i) all zero observations and (ii) missing observations in certain time steps. In a clustered sub-dataset, the first case occurs when all rainfall observations are zero on a given time step. In this case, cooperative metrics will be invalid because the standard deviation of the robust weight matrix is zero. Therefore, the cooperative metric of every PWS in this case is set to zero. The second case occurs when PWSs have intermittent missing observations. In this case, for those time steps that PWS has missing observations, this particular PWS is excluded from the robust average calculation, and the cooperative metric of this PWS will be set to zero. Additionally, if there are too many PWSs with missing observations resulting in a low number of active PWSs reporting data on a time step, the cooperative metrics of all PWSs on that time step will also be set to zero, because the consensus computed from the robust average algorithm may be unreliable. The definition of this low number can be determined based on the data availability during the analysis period.

2.2.3 Score

As described in Section 2.2.2, the *cooperative metric* can be interpreted as a measure of the PWS deviation from the robust average for each time step. To evaluate the trustworthiness of the PWSs, this step assumes neutral initial *trust scores* for every PWS without the knowledge of any past behaviors (rainfall observations in this case), and integrates this cooperative metric to update the *trust score* for every PWS for each time step.

We used the beta reputation system (Josang & Ismail, 2002) for its advantages of simplicity, flexibility, and ability to counter most arbitrary device faults in wireless sensor networks (Ganeriwal et al., 2008). The beta reputation uses a statistical approach to provide a mathematical basis for trust management. The idea is that the trust score, which is computed based on the beta probability density function (PDF), is gradually updated as new observations are made available. The beta PDF is a continuous family of distribution functions indexed by two parameters: α and β. It is denoted by $beta(p|\alpha, \beta)$ and can be expressed using the gamma function Γ as

$$beta(p|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \tag{6}$$

where $0 <= p <= 1$, α and $\beta > 0$. The expectation value of the beta distribution is given by

$$E(p) = \alpha/(\alpha+\beta) \tag{7}$$

where $0 < E(p) < 1$.

In each clustered sub-dataset, the prior distribution is assumed to be a uniform beta PDF with $\alpha_1 = 1$, $\beta_1 = 1$, and $E(p)_1 = 0.5$ for every PWS at time step $i = 1$ before any data is collected. This can be interpreted as the neutral trust for these PWSs which indicates that the relative frequency of reporting trustworthy or untrustworthy observations is equal. After observing new data, the posterior distribution will be the beta PDF with updated α and β parameters. In RSCRN, these parameters are updated using the cooperative metrics $C_{i,j}$ computed from the consensus step as

$$\begin{aligned} \alpha_{i+1,j} = \alpha_{i,j} \times \lambda + C_{i,j}, \quad \beta_{i+1,j} = \beta_{i,j} \times \lambda \quad \text{if } C_{i,j} > 0 \\ \alpha_{i+1,j} = \alpha_{i,j} \times \lambda, \quad \beta_{i+1,j} = \beta_{i,j} \times \lambda + |C_{i,j}| \quad \text{if } C_{i,j} < -1 \\ \alpha_{i+1,j} = \alpha_{i,j}, \quad \beta_{i+1,j} = \beta_{i,j} \quad \text{if } C_{i,j} = 0 \text{ or } -1 < C_{i,j} < 0. \end{aligned} \tag{8}$$

There are four possible outcomes for updating the α and β parameters: (i) A *positive outcome* is defined if the cooperative metric is greater than zero. In a positive outcome, the alpha parameter increases by the value of the cooperative metric. (ii) A *negative outcome* is defined if the cooperative metric is less than -1, which implies significant deviation from the consensus because the final robust weight is more than one standard deviation lesser from the average weight. In a negative outcome, the beta parameter increases by the absolute value of the cooperative metric. (iii) A *zero cooperative metric outcome* indicates either all observations at the time step were zero or a missing observation from a single PWS. In this case, both alpha and beta parameters are held constant with the previous time step values. (iv) In a *minor negative outcome* which we define as when the cooperative metrics is less than zero but greater than -1, both alpha and beta parameters will also be held constant with the previous time step values because the deviation from the consensus is insignificant. In addition, to focus the evaluation on time steps when rain is reported, the algorithm is set to only update trust scores on time steps when at least one PWS in the cluster is reporting more than one tick (0.25mm) of rainfall.

A forgetting factor λ is introduced in Equation 8 to avoid the trust score being overly weighted on the past information. The λ parameter, which ranges from 0 to 1, is used to give old information less weight than more recent information. A forgetting factor of 1.0 indicates no forgetting at all, whereas a forgetting factor of 0 indicates forgetting all past information except for the previous time step.

Given the updated alpha and beta parameters by the cooperative metrics, the expected value of the posterior beta PDF becomes

$$E(p)_{i+1,j} = \frac{\alpha_{i+1,j}}{\alpha_{i+1,j} + \beta_{i+1,j}}. \tag{9}$$

Finally, the *trust score* $T_{i,j}$ is computed by re-scaling the expectation value to be between 0 and 10 for each PWS j at time step i

$$T_{ij} = 10 \cdot E(p)_{i,j}. \tag{10}$$

2.3 Comparison with a PWS Quality Control Method

The performance of RSCRN approach is evaluated against a quality control method recently proposed for PWSs (de Vos et al., 2019). This quality control approach (hereinafter referred as PWS QC method) consists of three major filters to flag PWS rainfall observations. These filters are (i) a high influx (HI) filter to capture PWS observations with observations much higher than neighboring stations, (ii) a faulty zero (FZ) filter to identify erroneous zeros, and (iii) a station outlier (SO) filter to flag PWSs with low correlation of rainfall time series with neighboring stations. Using the same clustered sub-datasets as input for the PWS QC method, individual PWS observations were flagged with SO flags, FZ flags and SO flags, and these flags were used to compare to RSCRN trust scores.

2.4 Validation using High-Fidelity Rainfall Stations on Storm Events

Using a binary trust score threshold, PWSs were classified as trustworthy or untrustworthy PWSs for storm events with durations of X time steps that begin on time step T_1 and end on time step T_2. The trust score thresholds of the each PWS are defined as follows

$$\text{Trustworthy PWS:} \quad \sum_{i=T_1}^{T_2} \frac{T_{i,j}}{X} > \gamma \tag{11}$$

$$\text{Untrustworthy PWS:} \quad \sum_{i=T_1}^{T_2} \frac{T_{i,j}}{X} < \gamma \tag{12}$$

where γ is the threshold value that ranges from 0 to 10.

Using $\gamma = 5.0$ as an example, trustworthy PWSs are stations that received average trust scores higher than 5.0 during this storm event. Generally, these PWSs are more likely to be reporting trustworthy data during this storm event because they have been consistently contributing observations that agreed with the consensus from neighboring PWSs before the storm event. On the other hand, untrustworthy PWSs are PWSs that received average trust scores lower than 5.0, which indicates these PWSs have been reporting observations that disagreed with the consensus from neighboring PWSs. These PWSs are, therefore, less likely to report trustworthy data during the storm event.

To validate if using a trust score threshold method can improve rainfall estimates from the crowdsourced rainfall network, we compute the root-mean squared error (RMSE) of the PWS rainfall observation with the nearest high-fidelity rainfall station for the storm event as

$$RMSE = \sqrt{\frac{1}{X}\sum_{i=T_1}^{T_2}(c_i - h_i)^2} \tag{13}$$

where c_i is the rainfall time series of the PWS, h_i is the rainfall time series of the high-fidelity rainfall station, and X is the duration of the storm event. Consider a cluster with M PWSs, the average RMSE of all PWSs in the cluster (denoted as R_{all}) becomes

$$R_{all} = \sum_{j=1}^{M}\frac{RMSE_j}{M} \tag{14}$$

where $RMSE_j$ is the RMSE of the j-th PWS in the cluster. This R_{all} is used to benchmark the improvement made from the RSCRN and PWS QC methods.

Assuming that the RSCRN trust score threshold revealed that in these M PWSs, there are U trustworthy PWSs that received trust scores above the threshold, the average RMSE of trustworthy PWSs in the cluster (denoted as R_{RSCRN}) can be computed as

$$R_{RSCRN} = \sum_{j=1}^{U}\frac{RMSE_j}{U} \tag{15}$$

The RMSE of trustworthy PWSs is further comapred with the RMSE of QC PWS. Assuming there are V unflagged PWSs (stations without any flags filtered by the PWS QC method) during a storm event, the average RMSE of QC rainfall estimates (denoted as R_{QC}) can be be computed as

$$R_{QC} = \sum_{j=1}^{V}\frac{RMSE_j}{V}. \tag{16}$$

Lower RMSE values indicate agreement with the high-fidelity rainfall station observations. Therefore, the comparison of R_{all}, R_{RSCRN}, and R_{QC} can then be used to determine the improvements of rainfall estimates made by each method in providing accurate rainfall estimation from a network of PWSs.

3 Case study

3.1 Study Area

To demonstrate and evaluate RSCRN, we focus on PWSs in Houston, Texas as a case study. The City of Houston is in a sub-tropical climate with average annual rainfall of 1,250 mm. Flooding has been a recurring issue in Houston because of urbanization and the increase in frequency and intensity of severe storms (W. Zhang et al., 2018). The growing adoption of PWSs in Houston in recent years significantly increases ground gauge rainfall networks coverage. Extracting trustworthy rainfall information from the PWSs could potentially supply denser point rainfall time series and, thus, improve the knowledge of rainfall patterns to better model and control flooding.

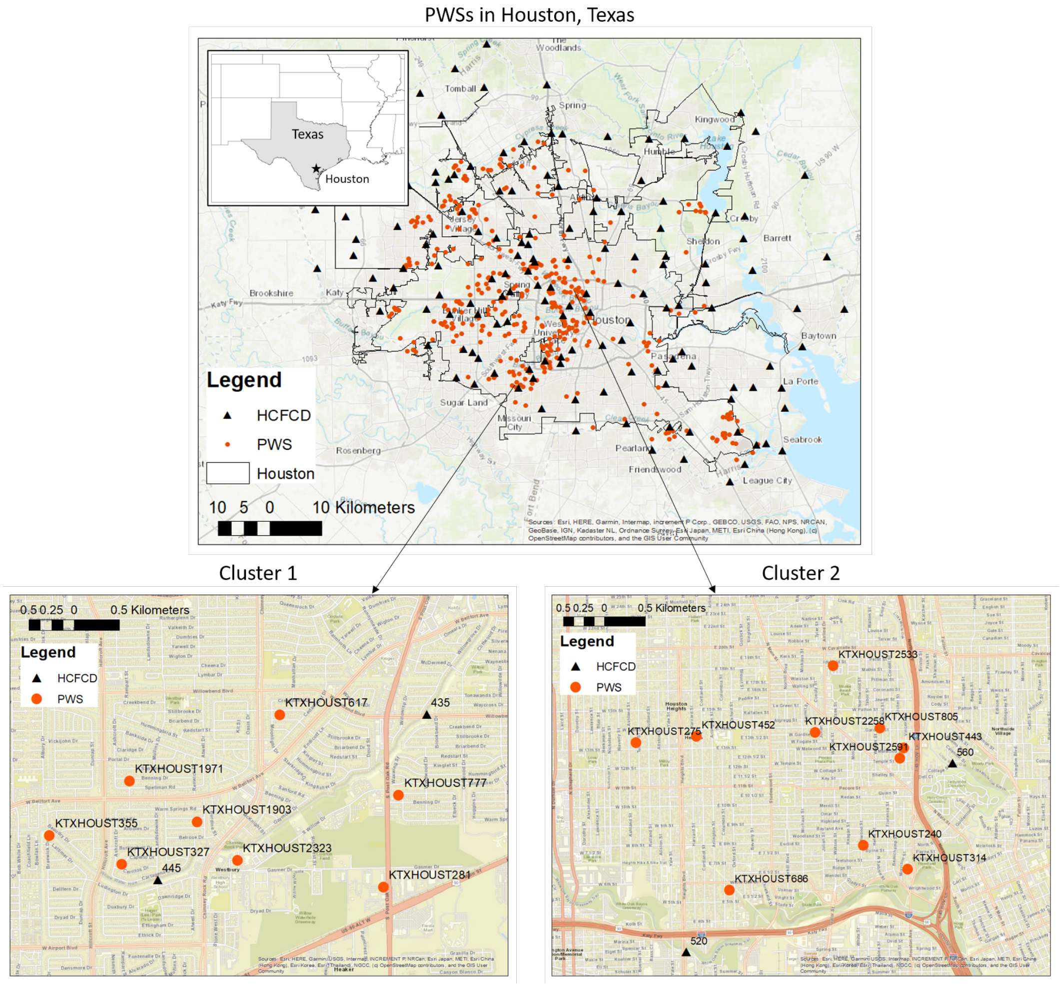


Figure 2. Two clusters of crowdsourced PWSs in Houston, Texas were selected as case study for evaluating the RSCRN.

3.2 Data

3.2.1 High-fidelity rainfall network

In Houston, the Harris County Flood Control District (HCFCD) manages a rainfall monitoring network of 174 rainfall stations that can be used as the ground truth of the rainfall observation to evaluate RSCRN (Figure 2).

3.2.2 Crowdsourced rainfall network

The crowdsourced rainfall network used in this study consists of PWSs that are available through the Weather Underground. We accessed the data through the API provided by the Weather Underground. The PWS observation sampling interval varies from station to station. Most of the sampling intervals are about 5-10 minutes per observation. Based on the available PWSs in Houston area queried from the Weather Underground API, there were 99 PWSs in January 2016 and 382 PWSs in just a few years later in April 2019.

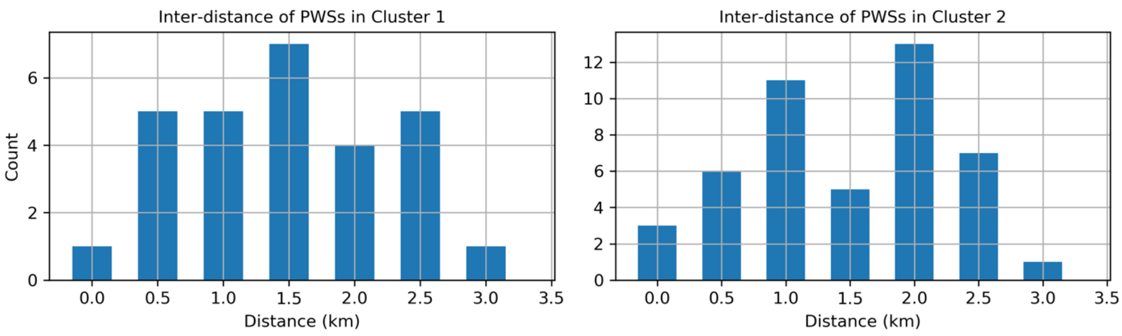


Figure 3. The distribution of PWSs inter-distance in both clusters. Both clusters contain two HCFCD rainfall stations in close proximity.

Table 1. The metadata of PWSs used in this study

	ID	Elevation	Latitude	Longitude	Start Time	Station Type
Cluster 1	KTXHOUST281	24	29.65	-95.46	9/14/2012	N/A
	KTXHOUST617	21	29.66	-95.47	7/6/2015	Ambient Weather WS-1400-IP (Wireless)
	KTXHOUST1971	16	29.66	-95.49	5/7/2017	AcuRite Pro Weather Center
	KTXHOUST2323	21	29.65	-95.48	3/18/2018	AcuRite Pro Weather Center
	KTXHOUST327	24	29.65	-95.49	10/27/2013	Davis Vantage Pro2 (Cabled)
	KTXHOUST1903	18	29.66	-95.48	12/29/2016	Ambient Weather WS-900-IP (Wireless)
	KTXHOUST355	21	29.65	-95.50	6/29/2014	N/A
	KTXHOUST777	21	29.66	-95.46	3/24/2016	Ambient Weather WS-1001-WiFi (Wireless)
Cluster 2	KTXHOUST240	15	29.79	-95.38	10/15/2010	Davis Vantage Pro2 Plus (Wireless)
	KTXHOUST805	20	29.80	-95.38	5/6/2016	AcuRite Pro Weather Center
	KTXHOUST443	21	29.79	-95.37	12/26/2014	AcuRite Pro Weather Center
	KTXHOUST2591	21	29.79	-95.37	1/1/2019	Ambient Weather WS-2902
	KTXHOUST314	28	29.78	-95.37	5/2/2013	Davis Vantage Pro2 Plus (Wireless)
	KTXHOUST686	25	29.78	-95.39	11/20/2015	Ambient Weather WS-1001-WiFi (Wireless)
	KTXHOUST452	24	29.80	-95.40	1/19/2015	Ambient Weather WS-1200-IP (Wireless)
	KTXHOUST2533	26	29.80	-95.38	10/29/2018	AcuRite 5-in-1 Weather Station with AcuRite Access
	KTXHOUST275	20	29.79	-95.40	6/23/2012	Davis Vantage Pro 2
	KTXHOUST2258	21	29.80	-95.38	1/27/2018	Ambient Weather WS-1001-WiFi (Wireless)

3.3 PWS cluster

Following the cluster method mentioned in Section 2.2.1, by setting the buffer distance to 2 kilometer for computing the number of neighboring PWSs, two clusters with the most active PWSs available at the beginning of the analysis period were used to evaluating RSCRN (see Figure 2). The first cluster consists of 8 PWSs and is located in southwestern Houston, Texas. The second cluster consists of 10 PWSs and is located in northwest of downtown Houston. The inter-distance of PWSs in both clusters is less than 3 kilometers (Figure 3). Table 1 shows the available metadata from the Weather Underground API. Each PWS has different start times, which is when the station joined Weather Underground and started reporting data to Weather Underground database. Among these PWSs, there are three major PWS brands: Ambient Weather, AcuRite and Davis Instrument.

The 15-min rainfall time series from the 18 PWSs in the two clustered sub-datasets for the analysis period from 1/1/2017 to 3/28/2019 were used as the input to the RSCRN. In this study, the minimum number of valid rainfall observations on a time step for computing the cooperative metrics was set to 5, given that there are 5-7 PWSs that are actively reporting rainfall within a cluster for the majority of the time steps during the analysis period. The forgetting factor λ was set to 0.95, which approximately retains 20% of the prior knowledge that is more than 25 time steps (6 hours) old. This is to ensure that the trust score computed by RSCRN will not

Table 2. Summary information of the selected 33 storms events for cluster 1.

No	Storm Event Date	Season	HCFCD Rain Gauge 445			PWS	
			Duration (hr)	Max. Rainfall Intensity (mm/hr)	Total Rainfall (mm)	Active PWSs	Median PWS Total Rainfall (mm)
1	20170102	Winter	0.5	81.3	30.5	6	29.3
2	20170118	Winter	4.8	89.4	119.9	6	100.6
3	20170120	Winter	2.0	77.2	35.6	6	33.8
4	20170305	Winter	8.8	24.4	61.0	6	48.5
5	20170329	Winter	2.5	93.5	49.8	6	40.4
6	20170418	Summer	4.0	16.3	25.4	6	14.7
7	20170522	Summer	3.3	52.8	32.5	7	30.5
8	20170529	Summer	1.8	105.7	62.0	7	56.1
9	20170604	Summer	2.5	56.9	53.8	7	44.5
10	20170624	Summer	1.8	56.9	31.5	6	20.8
11	20170715	Summer	3.3	113.8	47.8	6	28.7
12	20170802	Summer	3.8	36.6	42.7	7	37.3
13	20170808	Summer	2.3	48.8	26.4	6	24.8
14	20170825	Summer	10.8	61.0	101.6	7	67.1
15	20170918	Summer	2.5	52.8	53.8	7	44.7
16	20171203	Winter	1.3	69.1	39.6	7	42.9
17	20171216	Winter	1.8	40.6	25.4	7	23.4
18	20180210	Winter	6.8	32.5	87.4	7	86.9
19	20180329	Winter	4.3	52.8	66.0	8	55.0
20	20180421	Summer	1.5	56.9	41.7	8	44.2
21	20180521	Summer	2.0	69.1	35.6	7	20.8
22	20180704	Summer	6.5	89.4	164.6	8	144.3
23	20180731	Summer	1.0	81.3	37.6	7	35.6
24	20180909	Summer	1.5	61.0	29.5	6	35.8
25	20181015	Summer	0.8	89.4	27.4	6	20.1
26	20181031	Summer	3.3	101.6	82.3	5	94.2
27	20181207	Winter	8.5	52.8	124.0	7	118.1
28	20181213	Winter	2.0	28.4	26.4	7	30.0
29	20181227	Winter	3.3	20.3	36.6	6	39.2
30	20190102	Winter	2.8	21.3	28.4	7	35.6
31	20190119	Winter	0.8	65.0	25.4	7	22.9
32	20190123	Winter	4.8	20.3	26.4	4	26.2
33	20190226	Winter	3.3	28.4	27.4	6	28.6

be overweighted by past observations so that it is able to accommodate temporary behavioral changes, especially during storm events. The sensitivity to this forgetting factor is further explored later in the paper. For the PWS QC method, the neighboring stations for each PWSs were set to all other PWSs in the cluster identified by the RSCRN. Several parameter choices were evaluated and the best one were chosen based on the data availability and rainfall characteristics of the collected PWS data.

3.4 Storm Events Selection

In this study, a storm event is defined as the accumulated rainfall greater than 25.4 mm within a 12-hour rolling window. Rainfall time series from the high fidelity rainfall network (HCFCD rainfall stations 445 and 560) were used to identify storm events for cluster 1 and cluster 2, respectively. As shown in Tables 2 and 3, 33 and 29 storm events with various rainfall statistics that occurred in what we referred to as winter (November to March) and summer (April to October) seasons during the analysis period (1/1/2017 to 3/28/2019) were identified. In these storm events, durations ranged from 0.5 to 10.8 hours, maximum rainfall intensity from 20.3 to 113.8 mm/hr, and total rainfall from 25.4 to 164.6 mm.

Table 3. Summary information of the selected 29 storms events for cluster 2.

No.	Storm Event Date	Season	HCFCD Rain Gauge 560			PWS	
			Duration (hr)	Max. Rainfall Intensity (mm/hr)	Total Rainfall (mm)	Active PWSs	Median PWS Total Rainfall (mm)
1	20170102	Winter	1.0	56.9	25.4	5	21.3
2	20170118	Winter	6.0	81.3	157.5	6	165.4
3	20170120	Winter	2.8	73.2	42.7	6	39.2
4	20170220	Winter	6.3	12.2	37.6	5	40.4
5	20170305	Winter	7.5	16.3	38.6	5	38.6
6	20170329	Winter	2.0	101.6	47.8	5	46.5
7	20170411	Summer	4.3	48.8	36.6	6	35.6
8	20170522	Summer	2.5	69.1	37.6	7	39.1
9	20170604	Summer	3.0	105.7	61.0	7	65.3
10	20170625	Summer	2.3	81.3	57.9	7	25.4
11	20170713	Summer	1.0	81.3	30.5	7	27.2
12	20170807	Summer	1.8	48.8	27.4	7	22.4
13	20170826	Summer	18.8	113.8	389.1	6	417.7
14	20170918	Summer	1.8	48.8	25.4	6	30.5
15	20180108	Winter	2.5	40.6	27.4	7	16.5
16	20180210	Winter	5.5	20.3	57.9	8	59.6
17	20180329	Winter	5.5	52.8	67.1	8	62.4
18	20180421	Summer	1.8	65.0	45.7	8	46.4
19	20180520	Summer	2.0	44.7	26.4	8	26.7
20	20180704	Summer	5.8	40.6	110.7	6	148.6
21	20180909	Summer	1.3	81.3	45.7	7	37.9
22	20180922	Summer	2.0	56.9	25.4	7	16.5
23	20181015	Summer	1.3	61.0	27.4	7	31.2
24	20181031	Summer	3.0	44.7	34.5	8	27.8
25	20181207	Winter	9.0	44.7	116.8	8	109.0
26	20181213	Winter	2.0	36.6	27.4	7	27.9
27	20181227	Winter	3.5	36.6	43.7	7	39.1
28	20190102	Winter	2.3	28.4	30.5	8	26.4
29	20190123	Winter	5.0	16.3	26.4	8	25.7

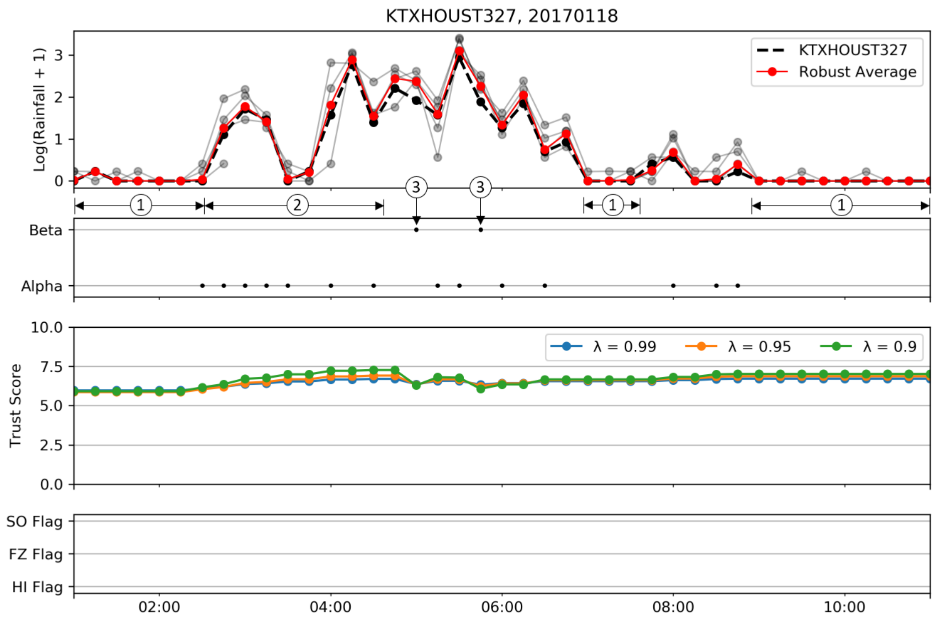


Figure 4. Example of a trustworthy PWS for a storm event. Trust score steadily increases during a storm event when the observed rainfall of a PWS (black dashed line) matches well with the consensus (robust average, shown in red line) of neighboring stations' reported rainfall (gray lines) consensus. No flags were identified by the PWS QC method in this storm event.

4 Results

4.1 Reputation System For Crowdsourced Rainfall Networks (RSCRN)

4.1.1 PWS Trust Score Assignment

Figure 4 shows the resulting trust scores based on RSCRN for an example storm event. In this example, the majority of rainfall observations of the target PWS (KTXHOUST327) matched well with the robust average computed from neighboring stations in its cluster. At the beginning of the storm (marked with circle 1 in Figure 4), the trust scores remained unchanged because the RSCRN only updates trust scores when at least one PWS in the cluster reports rainfall greater than 0.25 mm. Similarly, in the middle and end of the storm (also marked with circle 1), the trust scores remain constant because all reporting rainfall is lower than 0.25 mm. The PWSs began to observe heavier rainfall starting at 2:45am. As can be seen in Figure 4, the rainfall time series during the following time interval (marked with circle 2) agreed well with the robust average. Therefore, the PWS received several positive outcomes (marked with the black dot on the Alpha axis) and the trust score steadily increased. There were two time steps when the PWS received negative outcomes (marked with circle 3 and black dots on the Beta axis) because its observations disagreed with the consensus and the trust score decreased accordingly. Note that the change of trust score with regard to the same positive or negative outcomes was higher when using a smaller forgetting factor, because less prior knowledge was retained, and thus the trust score change was more sensitive. Expectedly, The PWS QC method did not identify any flags for the PWS during this storm event. Thus, both approaches agree this is a trustworthy PWS.

Figure 5 shows examples when trust scores decrease for the majority of the time steps during a storm event for two untrustwortyhy PWSs. In the first example, the rainfall time series of the target PWS (KTXHOUST452) frequently deviated from the robust average. Although this PWS captured some of the peak values of the storm, there were several time steps between those peaks where rainfall observations significantly deviated from the consensus of the neighboring PWSs. For example, the consensus of rainfall observations among the neighboring stations were showing that it had been raining heavily between the time interval 3:00 to 6:00. However, this PWS was either reporting zero rainfall or underreporting rainfall, which resulted in receiving many negative outcomes (black dots on the Beta axis). Therefore, the trust score decreased and remained low for the entire storm. Using the PWS QC method, several time steps were identified with the FZ flag, which agreed with the RSCRN that this station is likely to be untrustworthy. In the second example, the rainfall observation from the target PWS (KTXHOUST1903) was underreporting (0:00 to 0:30) and reporting zero value rainfall while the neighboring stations showed strong consensus of a certain rainfall magnitude. This station also overreported rainfall at 3:30 while the consensus of the neighboring stations showed that the storm had stopped. Using the PWS QC method, this PWS was first flagged with FZ flags for several intervals, followed by an HI flag where this PWS was reporting 49.27 mm while other neighboring stations all reported zero. Thus, both approaches agree these are untrustworthy PWSs.

4.1.2 PWS Trustworthiness Assessment

The RSCRN trust score evolution over all analyzed storm events is shown in Figure 6. Note that the trust scores of each PWS were computed for every time step during the analysis period (1/1/2017 0:00 to 3/28/2019 23:45) and the trust score for the analyzed storm events were extracted to assess the trustworthiness of a PWS during a particular storm event. In this figure, each dot represents the average trust score for a storm event. The trust score evolution shows that some PWSs were assigned high trust scores throughout the analyzed storm events (e.g., KTXHOUST327 in cluster 1 and KTXHOUST805 in cluster 2), while other PWSs consistently received low trust scores (e.g., KTXHOUST617 in cluster 1 and KTXHOUST452 in cluster 2). However, there are PWSs with trust scores that fluctuated over time, which indicates that perhaps these stations had state changes over the analysis period.

Table 4 shows the overall assessment of PWS trustworthiness for the analyzed storm events. Based on the results for cluster 1, KTXHOUST617 was the least trustworthy PWS. If we assume 4.0 as the trust score threshold γ, of the 23 active storm events for which this PWS reported valid rainfall observations, 18 (78%) were classified as untrustworthy. If we use a more restrictive trust score threshold $\gamma = 5.0$, 20 (87%) storm events were classified as untrustworthy. KTXHOUST281 was the second least trustworthy PWS in this cluster, as its trust score fluctuated between 4.0 and 6.0, and eventually dropped below 2.5. Of the 31 active storm events this PWS actively reported, 10 (32%) were classified as untrustworthy with trust score threshold $\gamma = 4.0$. Notably, as shown in Figure 6, KTXHOUST1903 initially received high trust scores, but dropped below 5.0 during several storm events. However, its trust score was restored to above 5.0 after storm event 20170715, and remained mostly trustworthy for the rest of the time. Other PWSs, such as KTXHOUST1971, KTXHOUST327, and KTXHOUST355, received relatively higher trust scores and were classified as trustworthy for most of the storm events (Figure 6). In cluster 2, KTXHOUST452 and KTXHOUST240 were the least trustworthy PWSs with an average trust score less than the threshold $\gamma = 4.0$ for 83% and 61% of the storm events, respectively. KTXHOUST443, with 29% of the storm events evaluated as untrustworthy, received a high trust score at the beginning of the analysis period

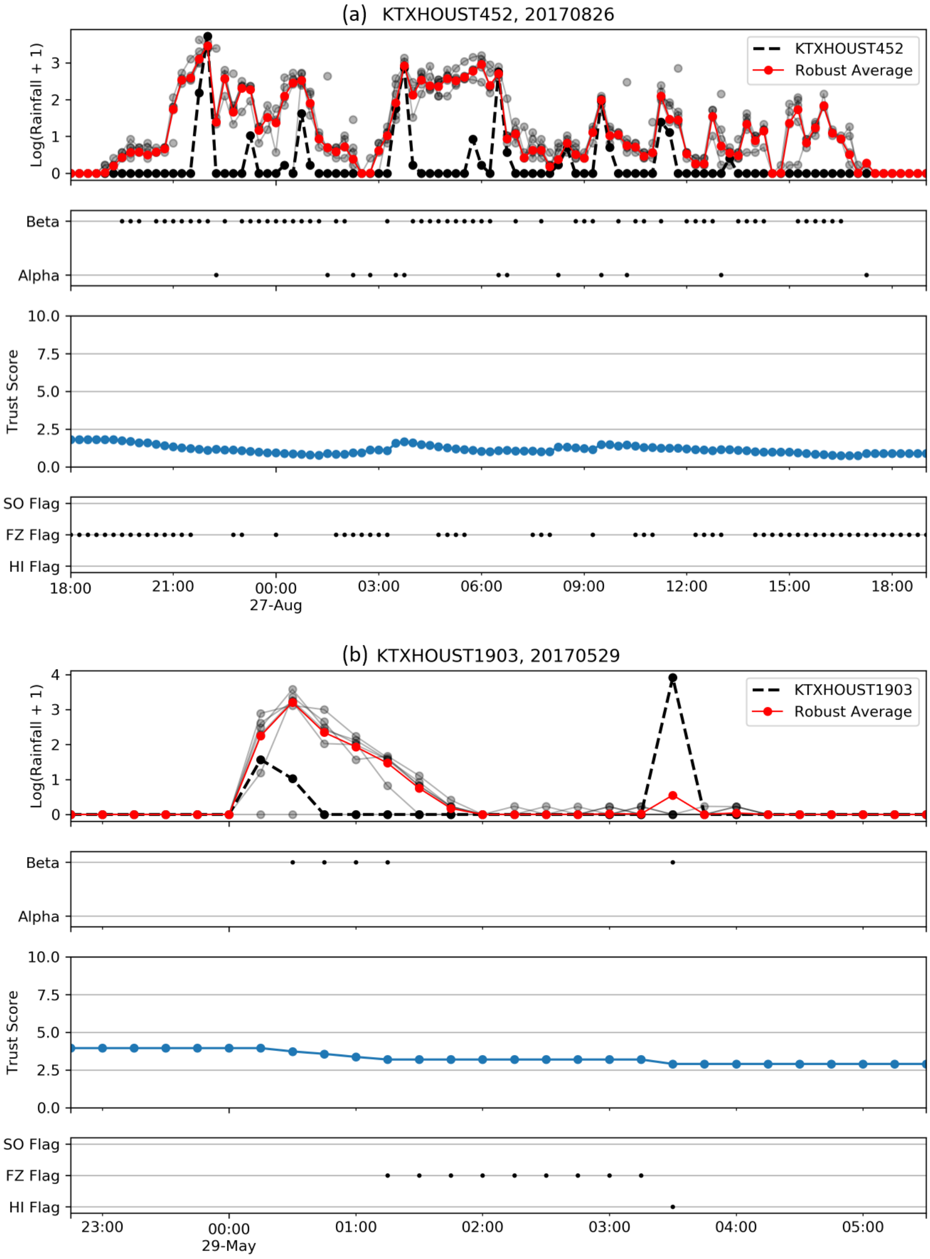


Figure 5. Examples of untrustworthy PWSs. Trust scores decrease when the reported rainfall of a PWS disagreed with the neighboring consensus. Faulty zero and high influx flags were also detected by the PWS QC method in these storm events.

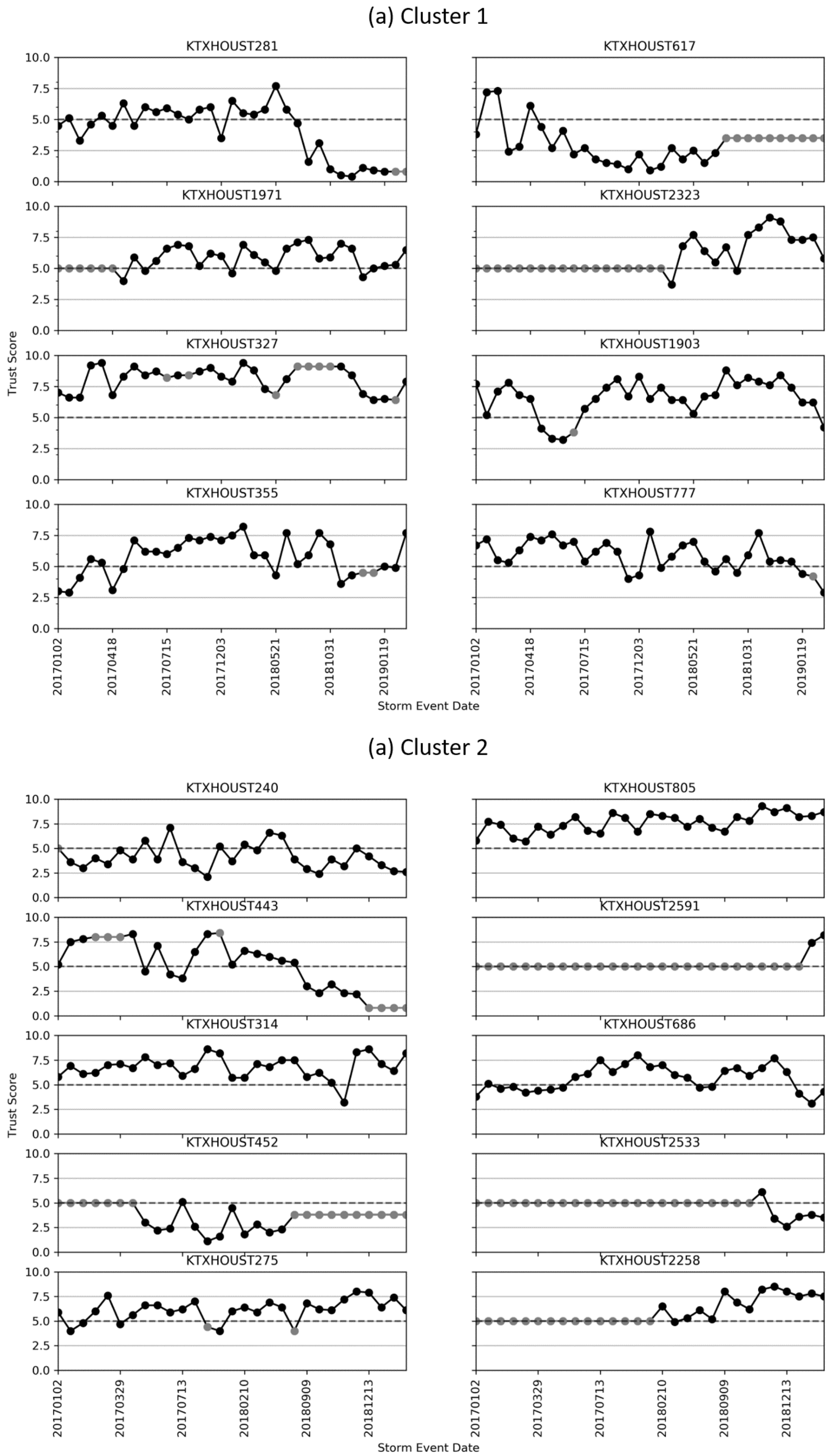


Figure 6. RSCRN trust score evolution during analyzed storm events. The gray markers indicate the PWS was not reporting any data during the storm events, thus the trust score remained constant.

Table 4. Overall assessment of the RSCRN trust score and PWS QC methods for the analyzed storm events.

	ID	Active Events	RSCRN: Untrustworthy Events (Avg. Trust Score <Threshold) Threshold = 5.0	Threshold = 4.0	Threshold = 3.0	PWS QC method (de Vos et al., 2019): Flagged Events
	KTXHOUST281	31	15 (48%)	10 (32%)	7 (23%)	8 (26%)
Cluster 1	KTXHOUST617	23	20 (87%)	18 (78%)	17 (74%)	20 (87%)
	KTXHOUST1971	27	5 (19%)	0 (0%)	0 (0%)	0 (0%)
	KTXHOUST2323	15	2 (13%)	1 (7%)	0 (0%)	0 (0%)
Analyzed	KTXHOUST327	25	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Storm	KTXHOUST1903	32	4 (13%)	2 (6%)	0 (0%)	6 (19%)
Events: 33	KTXHOUST355	31	9 (29%)	4 (13%)	1 (3%)	3 (10%)
	KTXHOUST777	32	7 (22%)	1 (3%)	0 (0%)	7 (22%)
	KTXHOUST240	28	21 (75%)	17 (61%)	5 (18%)	5 (18%)
	KTXHOUST805	29	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Cluster 2	KTXHOUST443	21	8 (38%)	6 (29%)	3 (14%)	6 (29%)
	KTXHOUST2591	2	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Analyzed	KTXHOUST314	29	1 (3%)	1 (3%)	0 (0%)	1 (3%)
Storm	KTXHOUST686	29	12 (41%)	2 (7%)	0 (3%)	6 (21%)
Events: 29	KTXHOUST452	12	11 (92%)	10 (83%)	9 (75%)	12 (100%)
	KTXHOUST2533	6	5 (83%)	5 (83%)	1 (16%)	1 (17%)
	KTXHOUST275	27	4 (15%)	1 (4%)	0 (0%)	2 (7%)
	KTXHOUST2258	14	1 (7%)	0 (0%)	0 (0%)	0 (0%)

but decreased overtime and eventually dropped below 2.5. Other PWSs were mostly trustworthy during the storm events based on the average trust scores they received.

4.1.3 Comparison with PWS Quality Control Method

A comparison with PWS QC method (Table 4) shows that for each PWS, the number of untrustworthy events (defined as storm events that average trust scores below a threshold) identified by RSCRN generally agreed with the number of flagged events (defined as storm events that had at least one observation flagged by the PWS QC method) for the analyzed storm events. In cluster 1, PWSs with a large number of untrustworthy events were also frequently flagged by the PWS QC method, whereas PWSs that were assigned higher trust scores usually received fewer or no flagged events. In cluster 2, PWSs with a higher percentage of untrustworthy events also received several flags from the PWS QC method. Using a different trust score threshold for classifying PWSs, the comparison showed that the number of untrustworthy events agreed the most with the flagged events from PWS QC method when using $\gamma = 4.0$.

As most of the agreements between the RSCRN and PWS QC method were for high influx and faulty zero flags (Figure 5), there were cases where RSCRN identified additional untrustworthy behavior while there were no flags determined by the PWS QC method. Using the storm event 20170120 as an example (Figure 7), the rainfall observed from this PWS (KTXHOUST281) received several negative outcomes from the RSCRN. At 17:30 and 18:15, the reported rainfall were 1.02 and 4.06 mm, while the robust average computed from the neighboring PWSs were 5.06 and 0.82 mm, respectively. This caused the trust score of the PWS to drop lower than the threshold value for this storm event. However, no observations were flagged by the PWS QC method in this event because none of the observations in this storm event met the predefined filter threshold of FZ, HI, and SO flags.

In a second example using PWS KTXHOUST443 (cluster 2) and storm event 20170625 (Figure 8), the rainfall reported from this station was much higher than the neighboring consensus, which resulted in a decrease in the trust score because of

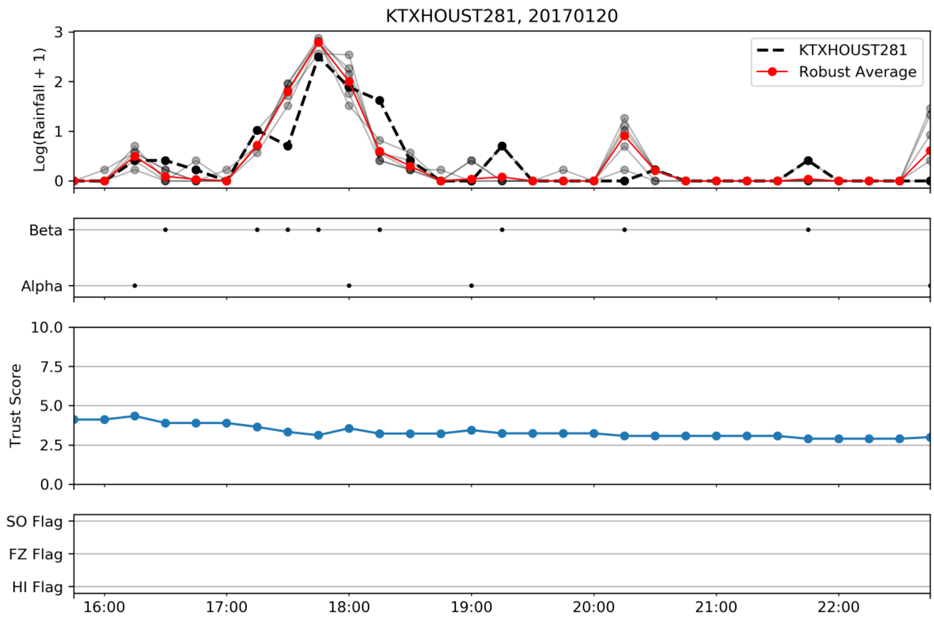


Figure 7. RSCRN algorithm assigned low trust scores to a PWS during a storm event while no flags were detected by the PWS QC method.

a couple time steps of negative outcomes identified by the RSCRN. However, in this particular storm event, the actual consensus of rainfall observations might be uncertain because greater spatial variability existed across PWSs in the cluster. Based on the RSCRN algorithm, this was interpreted as the PWS was untrustworthy because of overreporting, which deviated from the consensus. However, as larger rainfall variability exists on this time step, the actual trustworthiness of this station might be uncertain. Further work should explore the role of rainfall variability within a cluster, and not just the robust average, in assigning negative outcomes in RSCRN.

4.2 Validation using High-Fidelity Rainfall Stations

To validate if RSCRN can result in higher accuracy of PWS-derived rainfall estimates, the RMSE between rainfall observations from PWSs to high-fidelity rainfall stations (HCFCD) at storm events were computed. As shown in Figure 2, the HCFCD rain gauges (445 and 435 in cluster 1, 520 and 560 in cluster 2) were in close proximity (mostly less than 1 kilometer) with PWSs in the clusters and thus were used as the ground truth of actual rainfall observations for validation.

Table 5 and 6 show the RMSE comparison of PWS rainfall estimates for the analyzed storm events. In these comparisons, the RMSEs were computed using all PWSs (Equation 14, denoted as R_{all}), trustworthy PWSs (Equation 15, denoted as R_{RSCRN}), and QC PWSs (Equation 16, denoted as R_{QC}). The resulting R_{all} ranged from 0.43 to 3.41mm across the analyzed storm events, except for the storm event 20170305 in cluster 1 for which a single PWS (KTXHOUST355) reported an extreme value of 1462.53 mm, which resulted in much higher R_{RSCRN} for this particular storm event (column 1 in Table 5). It is worth noting that, because the RSCRN method did not rely on only a single observation to determine the trustworthiness of a PWS, it did not identify this station as untrustworthy for this storm event, which resulted worse performance at this particular storm event. This

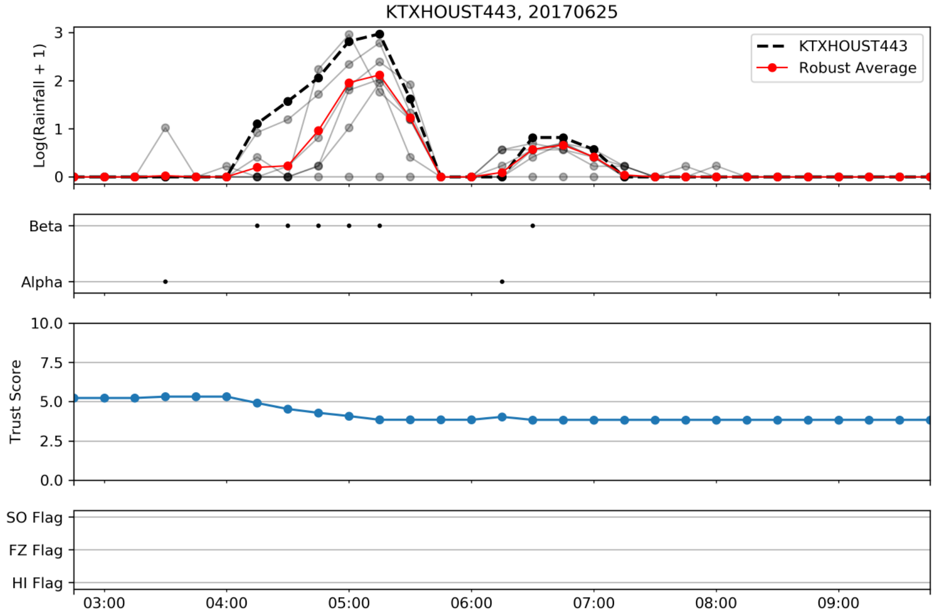


Figure 8. RSCRN algorithm identified several negative outcomes (mostly overreporting rainfall) in a storm event with large spatial and temporal rainfall variability.

suggests, however, that RSCRN could be used along with basic outlier detection method to improve its results.

The overall performances for both methods are shown in Table 7. Using the RSCRN method, the results showed that of the 33 analyzed storm events in cluster 1, R_{RSCRN} outperformed R_{all} for 25 (76%) of the events, with a median RMSE improvement (ΔRMSE) of 0.35 (24.5%) (bold values in column 4 of Table 5, which is computed by subtracting R_{RSCRN} with R_{all}). Of the 29 analyzed storm events in cluster 2, R_{RSCRN} outperformed R_{all} for 23 (79%) of the events, with a median RMSE improvement of 0.41 (29.8%) (bold values in column 4 of Table 6). Using the PWS QC method, results showed that R_{QC} improved for 22 (67%) of the events in cluster 1, and 16 (55%) of the events in cluster 2 (bold values in column 5 of Tables 5 and 6). This demonstrates that both approaches made significant improvements in PWS rainfall estimates for the majority of the storm events.

The comparison of R_{RSCRN} and R_{QC} showed that RSCRN generally outperformed the PWS QC method (Table 7). In cluster 1, the results showed that 7 (21%) storm events have the same performance. However, in the remaining 26 storm events, RSCRN outperformed PWS QC method in 18 (81%) of the storm events, with a median RMSE improvement of 0.11 (8.6%) (shown in bold values in column 6 of Table 5), while the PWS QC method outperformed RSCRN in 5 (15%) of the storm events (shown in italic values in the column 6 of Table 5). In cluster 2, the results showed that 6 (21%) storm events had the same performance. However, in the remaining 23 storm events, RSCRN outperformed the PWS QC method in 17 (74%) storm events, with a median RMSE improvement of 0.29 (25.0%), while PWS QC method outperformed RSCRN in 6 (23%) storm events. This suggests that the RSCRN apporach identified additional untrustworthy PWSs that were not flagged by the PWS QC method, and thus improved the rainfall estimates from PWS network.

Table 5. Comparison of RMSE improvements for PWS rainfall estimates across storm events using RSCRN and PWS QC methods for cluster 1. Column 1 to 3 are the average RMSE of rainfall estimates computed from all PWSs, trustworthy PWSs only and QC PWSs only, respectively. Column 4 shows the improvements made from trustworthy PWSs (R_{RSCRN} - R_{all}); Column 5 shows the improvements made from QC PWSs (R_{QC} - R_{all}); Column 6 shows the improvement between RSCRN and PWS QC method (R_{QC} - R_{RSCRN}).

Storm Event Date	(1) R_{all}	(2) R_{RSCRN}	(3) R_{QC}	(4) R_{all} - R_{RSCRN}	(5) R_{all} - R_{QC}	(6) R_{QC} - R_{RSCRN}
20170102	1.78	0.87	1.78	**0.92**	0.00	**0.92**
20170118	1.47	1.34	1.47	**0.13**	0.00	**0.13**
20170120	1.22	1.13	1.22	**0.09**	0.00	**0.09**
20170305	27.82	41.13	0.67	*-13.31*	**27.15**	*-40.46*
20170329	2.77	2.52	2.70	**0.25**	**0.07**	**0.18**
20170418	1.27	1.72	0.40	*-0.46*	**0.87**	*-1.32*
20170522	1.56	1.10	1.34	**0.46**	**0.22**	**0.24**
20170529	3.41	1.15	1.26	**2.26**	**2.15**	**0.11**
20170604	2.86	1.20	1.43	**1.66**	**1.43**	**0.23**
20170624	0.87	0.80	0.88	**0.07**	*-0.01*	**0.08**
20170715	2.10	2.14	2.30	*-0.04*	*-0.20*	**0.16**
20170802	1.26	1.33	1.33	*-0.08*	*-0.08*	0.00
20170808	1.23	1.05	1.14	**0.18**	**0.09**	**0.09**
20170825	2.80	2.22	2.22	**0.58**	**0.58**	0.00
20170918	1.99	1.92	1.87	**0.07**	**0.12**	*-0.05*
20171203	1.36	0.88	1.05	**0.48**	**0.31**	**0.18**
20171216	0.83	0.54	0.52	**0.29**	**0.31**	*-0.02*
20180210	1.19	0.88	0.88	**0.31**	**0.30**	**0.00**
20180329	2.20	2.00	1.97	**0.20**	**0.23**	*-0.03*
20180421	1.73	1.37	1.37	**0.35**	**0.35**	0.00
20180521	1.23	0.75	1.20	**0.48**	**0.03**	**0.45**
20180704	1.93	1.44	1.44	**0.48**	**0.48**	0.00
20180731	1.00	0.80	0.83	**0.20**	**0.17**	**0.03**
20180909	1.18	0.78	0.78	**0.40**	**0.40**	0.00
20181015	1.77	1.33	1.38	**0.43**	**0.39**	**0.05**
20181031	1.25	1.08	1.20	**0.17**	**0.05**	**0.12**
20181207	1.51	0.82	0.82	**0.69**	**0.69**	0.00
20181213	0.60	0.60	0.70	0.00	*-0.10*	**0.10**
20181227	0.47	0.48	0.56	*-0.01*	*-0.09*	**0.08**
20190102	2.37	0.45	0.46	**1.92**	**1.91**	**0.01**
20190119	0.80	0.80	0.93	0.00	*-0.13*	**0.13**
20190123	0.43	0.40	0.43	**0.03**	0.00	**0.03**
20190226	0.55	0.55	0.55	0.00	0.00	0.00

Table 6. The comparison of RMSE improvements for PWS rainfall estimates across storm events using RSCRN and PWS QC method for cluster 2.

Storm Event Date	(1) R_{all}	(2) R_{RSCRN}	(3) R_{QC}	(4) R_{all} - R_{RSCRN}	(5) R_{all} - R_{QC}	(6) R_{QC} - R_{RSCRN}
20170102	0.74	0.88	0.74	*-0.14*	0.00	*-0.14*
20170118	3.15	2.78	3.02	**0.38**	**0.13**	**0.25**
20170120	2.07	1.93	2.07	**0.13**	0.00	**0.13**
20170220	0.80	0.50	0.80	**0.30**	0.00	**0.30**
20170305	1.10	0.63	1.10	**0.47**	0.00	**0.47**
20170329	3.50	0.75	2.33	**2.75**	**1.18**	**1.58**
20170411	0.97	0.68	0.90	**0.29**	**0.07**	**0.23**
20170522	1.19	1.53	1.40	*-0.34*	*-0.21*	*-0.13*
20170604	1.69	0.98	1.30	**0.71**	**0.39**	**0.32**
20170625	1.99	2.06	2.02	*-0.07*	*-0.03*	*-0.04*
20170713	2.51	1.64	1.06	**0.87**	**1.45**	*-0.58*
20170807	1.07	0.90	0.97	**0.17**	**0.10**	**0.07**
20170826	7.40	4.10	4.10	**3.30**	**3.30**	0.00
20170918	1.70	1.85	1.72	*-0.15*	*-0.02*	*-0.13*
20180108	2.17	2.76	2.47	*-0.59*	*-0.30*	*-0.29*
20180210	1.21	1.10	1.10	**0.11**	**0.11**	0.00
20180329	1.81	1.50	1.59	**0.31**	**0.23**	**0.09**
20180421	2.00	1.61	1.61	**0.39**	**0.39**	0.00
20180520	1.20	1.20	1.20	0.00	0.00	0.00
20180704	3.10	2.18	3.10	**0.93**	0.00	**0.93**
20180909	2.46	0.98	0.98	**1.48**	**1.48**	0.00
20180922	1.16	0.62	0.62	**0.54**	**0.54**	0.00
20181015	2.33	1.92	2.08	**0.41**	**0.25**	**0.16**
20181031	1.40	0.46	1.37	**0.94**	**0.03**	**0.91**
20181207	1.55	1.06	1.60	**0.49**	*-0.05*	**0.54**
20181213	1.44	1.02	1.44	**0.42**	0.00	**0.42**
20181227	1.11	0.88	1.17	**0.24**	*-0.05*	**0.29**
20190102	0.67	0.48	0.57	**0.19**	**0.10**	**0.09**
20190123	0.60	0.50	0.57	**0.10**	**0.03**	**0.07**

Table 7. Overall comparison of RMSE improvements using RSCRN and PWS QC method.

		Number of Storm Events			Median ΔRMSE (%)		
		R_{all} - R_{RSCRN}	R_{all} - R_{QC}	R_{QC} - R_{RSCRN}	R_{all} - R_{RSCRN}	R_{all} - R_{QC}	R_{QC} - R_{RSCRN}
Cluster 1	ΔRMSE >0	25	22	21	0.35 (24.5%)	0.33 (22.3%)	0.11 (8.6%)
	ΔRMSE <0	5	6	5	-0.08 (-5.2%)	-0.10 (-13.1%)	-0.05 (-3.8%)
	ΔRMSE = 0	3	5	7	-	-	-
Cluster 2	ΔRMSE >0	23	16	17	0.41 (29.8%)	0.24 (13.9%)	0.29 (25.0%)
	ΔRMSE <0	5	7	6	-0.15 (-18.2%)	-0.05 (-4.0%)	-0.28 (-10.4%)
	ΔRMSE = 0	0	6	6	-	-	-

5 Discussion

5.1 The Potential Crowdsourced Rainfall Data

While this study focuses on crowdsourced rainfall data collected from PWSs, the proposed RSCRN can be beneficial to ensure the trustworthiness of other emerging crowdsourcing rainfall data collection methods as well. Beyond the case of PWSs, recent advancement of crowdsourcing methods has further enabled rainfall observations to be collected from connected vehicles (Bartos et al., 2019), surveillance cameras (Jiang et al., 2019), and mobile phones (Guo et al., 2018). The availability of these crowdsourcing methods greatly facilitates more crowd-participation, but also raises concerns of increased uncertainty associated with data contributors, highlighting the need for evaluating the trustworthiness of crowdsourced data (Gharesifard & Wehn, 2016; Hunter et al., 2013). RSCRN should be viewed as a starting point for creating algorithms capable of systematically assigning the trustworthiness of these data based on physical principle able to be applied at scale for quickly growing networks.

5.2 The Availability and Reliability of Crowdsourced Data

PWS adoption has been growing rapidly thanks to the advancement of technologies that made PWSs easy to install and affordable, as well as software able to connect and share the data through online platforms. This increase in PWS data openly shared on the Internet has transformed the value of PWSs from serving the owners' interests to anyone in the broader community that might benefit from the information (Gharesifard & Wehn, 2016). However, PWS data accessibility depends heavily on the platform that the PWSs are connected to. For example, Weather Underground recently ended a freely available service of the Weather Underground API and replaced it with a new API service that only allows PWS contributors to utilize the service (WXForum, 2018a). Weather Underground has also stopped the the automatic connection with PWSs of certain brands (e.g. Netatmo) to their platform (WXForum, 2018b), resulting in abrupt changes to the number of sensors available in the system. These kind of sudden changes might happen in any crowdsourced platform without warnings, which could further compromise the accessibility and reduce the utility of crowdsourced data. The community would benefit from more standardization of open networks and data sharing agreements to make the most from this emerging data resource.

5.3 The Assumption of Consensus in Rainfall Observations

One of the premises behind RSCRN is that consensus in crowdsourced rainfall observations exists at some scale in space and time and can, therefore, be used to judge trustworthiness of stations within a cluster. Such consensus-based ideas are widely used across disciplines to identify errors in data (J. Zhang et al., 2017; Foody et al., 2018; Strobl et al., 2019). Strong consensus in rainfall observations occurs when rain gauges are located in close proximity, but the exact distance and other factors that should be used for defining a cluster are uncertain. This idea is not new, however. For example, the United States Climate Reference Network (USCRN), an extremely high-fidelity rainfall network, uses three distinct tipping bucket sensors installed next to each other on the same site for immediate detection of single sensor failure (Diamond et al., 2013). Better accounting for factors that influence consensus in rainfall observations (e.g., geography, climate, observation frequency) are possible extentions to the approach used in this study. In this work, a cluster was identified based on a group of PWSs that were in close distance with each other. However, large rainfall variability may exist even over short distances, especially for high frequency rainfall observations or if stations have large elevation differences.

For example, this particular case study focused on Houston, Texas, which has only slight topographical variation across the region, thus elevation was not a factor in station clustering. However, for regions with greater variation in elevation such as mountainous areas, the clustering results should include elevation to reflect where the consensus actually exists (Buytaert et al., 2006). Rainfall types can also be one of the factors that affect the consensus. For example, a convective storm may produce rainfall over a small area that is only captured by a single PWS in a cluster if clusters are not carefully created. In these cases, incorporating additional variables of the PWS location into the clustering method may better capture the consensus and thus result in more meaningful trust scores.

5.4 Feedback to Data Collectors for Improved Crowdsourced Data Quality

People are motivated by various kinds of incentives to adopt a PWS. Examples of these incentives include obtaining useful weather data for personal purposes or having a sense of belonging to a community of friends with shared interests (Gharesifard & Wehn, 2016). As the need of higher spatial and temporal resolution of rainfall data increases, the role of PWS data may be shifted from serving personal interests to benefiting the society at large (Gharesifard & Wehn, 2016). In this case, because people who need to utilize the data are interested in knowing the quality of the data they contributed, PWS owners might become what Jøsang et al. (2007) described as *service providers*. To manage their *provision trust*, they may be willing to demonstrate their competence in collecting data and arguably welcome any feedback to improve their data quality. As a result, RSCRN could assist by making the trust score information available to the PWS owners. PWS owners could be notified of a drop in the trust score, actions could be taken to correct the erroneous observations (e.g., cleaning the clogged rain gauge). Such efforts not only help restore the trust scores, which maintain their *provision trust*, but also greatly improve the overall data quality of the crowdsourced rainfall network in the long term. Future improvements to RSCRN could focus on identifying particular types of errors to more effectively advise users on steps to improve their trust score.

5.5 Limitation of Binary Trust Score Threshold

Trust scores derived from the RSCRN represent the relative frequency of a PWS reporting trustworthy rainfall observations in the future. This continuous form is computationally efficient for reputation systems to calculate and update overtime (Ruan & Durresi, 2016). However, to better enable reputation-based decision making, a discrete format of trust scores is often used (Mousa et al., 2015), as humans are often better able to understand discrete verbal statements than continuous measures (Jøsang et al., 2007). In this study, we used a RSCRN-derived trust score threshold approach to classify PWSs as either trustworthy or untrustworthy. While using a binary trust score threshold is simple and straightforward for enabling decision making (e.g., include or ignore rainfall from a specific PWS), it does not represent the varying trustworthiness of PWSs (Ruan & Durresi, 2016). For example, a PWS with an extremely low trust score and a PWS with a trust score just below the threshold will both be categorized as an untrustworthy PWS, despite their difference in the extent of untrustworthiness. Alternatives can be dynamically adjusting the binary trust score threshold to optimize decision-making or use multinomial discrete values such as *very trustworthy, trustworthy, untrustworthy, very untrustworthy* to account for a broader extent of trustworthiness across PWSs (Ruan & Durresi, 2016). Future work could explore extensions like this so that RSCRN is able to weight information from PWSs based on their trust score rather simply including or excluding measurements using a threshold method.

6 Conclusion

In this study, we presented a Reputation System for Crowdsourced Rainfall Network (RSCRN) for ensuring the trustworthiness of PWSs in a crowdsourced rainfall network. The RSCRN assigned trust scores to PWSs are calculated by (i) clustering the PWSs into groups with similar rainfall characteristics, (ii) computing the rainfall observation consensus within each cluster using a robust average method and (iii) deriving trust scores using a beta reputation system.

Using PWS rainfall data collected from Houston, Texas as a case study, we demonstrated how RSCRN is able to identify PWSs with untrustworthy rainfall data. By ignoring rainfall from untrustworthy PWSs using a RSCRN-derived trust score threshold, the accuracy of the resulting 15-min rainfall estimates better matched rainfall observations observed from high-fidelity rainfall stations for 77% (48 out of 62) of the analyzed storm events, with a median RMSE improvement of 27.3%. Compared to a PWS quality control method, results showed that while 13 (21%) storm events had the same performance, RSCRN improved rainfall estimates for 78% of the remaining storm events (38 out of 48), with a median RMSE improvement of 13.4%.

We returned to the research questions mentioned in section 1 that guided this work to provide answers based on the research's outcomes.

(i) How can we evaluate the trustworthiness of crowdsourced PWSs?

This study demonstrated that a reputation system approach could be useful in evaluating the trustworthiness of crowdsourced PWSs. Unlike a traditional QA/QC method, the reputation system approach collectively evaluates the trustworthiness of a PWS itself over time rather than single observations collected at a gauge. The RSCRN presented in this study assigns trust scores to PWSs based on their agreement or disagreement with current and historical rainfall observations from neighboring PWS, and is able to converge to a confident trust score in 20-30 time steps, as well as accommodate sudden changes of PWSs trust levels during storm events in the case of system changes (e.g., a malfunctioning station).

(ii) To what extent could a reputation system approach improve rainfall estimates from PWSs?

The reputation system can be used to improve rainfall estimates in direct and indirect ways. First, the reputation system approach ensures the rainfall estimates were produced from trustworthy PWSs. Using RSCRN-derived trust scores threshold, PWSs were classified as trustworthy or untrustworthy. By judging whether a PWS should be included in the rainfall estimation process, the resulting trustworthy rainfall estimates were greatly improved in accuracy for matching rainfall observed from high-fidelity rainfall stations. Second, the reputation system approach has the potential to encourage PWS owners to maintain and contribute high quality data, which indirectly improves rainfall estimates from PWSs in the long term.

Future work could be aimed at (i) a larger analysis of crowdsourced rainfall networks to identify and quantify the extent of untrustworthy PWSs across cities and regions in the world, (ii) enhancing the reputation system algorithm to account for rainfall variability in complex topography and finer-temporal scales and (iii) leveraging crowdsourced rainfall estimates to improve hydrological modeling such as rainfall-runoff and flood prediction. With a reputation system able to ensure the trustworthiness of PWSs and improve the data quality collected through crowdsourced rainfall networks, this growing data resource can be more confidently adopted and trusted for not only urban flood applications but other water resources management and decision-making, challenges as well.

Acknowledgments

This research is supported by the National Science Foundation under Grant No. CBET-1735587. We gratefully acknowledge Weather Underground and Harris County Flood Control District for access to their data. Python codes of the RSCRN and datasets used in this study can be accessed from Hydroshare (`https://www.hydroshare.org/resource/cf7796cdeace42818dbbd7f95f8e1872/`).

References

Bartos, M., Park, H., Zhou, T., Kerkez, B., & Vasudevan, R. (2019). Windshield wipers on connected vehicles produce high-accuracy rainfall maps. *Scientific Reports*, *9*(1), 170.

Bell, S., Cornford, D., & Bastin, L. (2013). The state of automated amateur weather observations. *Weather*, *68*(2), 36–41.

Berne, A., Delrieu, G., Creutin, J.-D., & Obled, C. (2004). Temporal and spatial resolution of rainfall measurements required for urban hydrology. *Journal of Hydrology*, *299*(3-4), 166–179.

Blenkinsop, S., Lewis, E., Chan, S. C., & Fowler, H. J. (2017). Quality-control of an hourly rainfall dataset and climatology of extremes for the uk. *International Journal of Climatology*, *37*(2), 722–740.

Buytaert, W., Celleri, R., Willems, P., De Bievre, B., & Wyseure, G. (2006). Spatial and temporal rainfall variability in mountainous areas: A case study from the south ecuadorian andes. *Journal of Hydrology*, *329*(3-4), 413–421.

Chapman, L., Bell, C., & Bell, S. (2017). Can the crowdsourcing data paradigm take atmospheric science to a new level? a case study of the urban heat island of london quantified using netatmo weather stations. *International Journal of Climatology*, *37*(9), 3597–3605.

Chen, A. B., Behl, M., & Goodall, J. L. (2018). Trust me, my neighbors say it's raining outside: ensuring data trustworthiness for crowdsourced weather stations. In *Proceedings of the 5th Conference on Systems for Built Environments* (pp. 25–28).

Chou, C. T., Ignjatovic, A., & Hu, W. (2013). Efficient computation of robust average of compressive sensing data in wireless sensor networks in the presence of sensor faults. *IEEE Transactions on Parallel and Distributed Systems*, *24*(8), 1525–1534.

Cox, L. P. (2011). Truth in crowdsourcing. *IEEE Security & Privacy*, *9*(5), 74–76.

Cristiano, E., Veldhuis, M.-c. t., & Giesen, N. v. d. (2017). Spatial and temporal variability of rainfall and their effects on hydrological response in urban areas–a review. *Hydrology and Earth System Sciences*, *21*(7), 3859–3878.

Cunha, L. K., Smith, J. A., Krajewski, W. F., Baeck, M. L., & Seo, B.-C. (2015). Nexrad nws polarimetric precipitation product evaluation for ifloods. *Journal of Hydrometeorology*, *16*(4), 1676–1699.

de Vos, L. W., Leijnse, H., Overeem, A., & Uijlenhoet, R. (2017). The potential of urban rainfall monitoring with crowdsourced automatic weather stations in amsterdam. *Hydrology and Earth System Sciences*, *21*(2), 765–777.

de Vos, L. W., Leijnse, H., Overeem, A., & Uijlenhoet, R. (2019). Quality control for crowdsourced personal weather stations to enable operational rainfall monitoring. *Geophysical Research Letters*, *46*(15), 8820–8829.

Diamond, H. J., Karl, T. R., Palecki, M. A., Baker, C. B., Bell, J. E., Leeper, R. D., ... others (2013). Us climate reference network after one decade of operations: Status and assessment. *Bulletin of the American Meteorological Society*, *94*(4), 485–498.

Estellés-Arolas, E., & González-Ladrón-De-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science*, *38*(2), 189–200.

Estévez, J., Gavilán, P., & Giráldez, J. V. (2011). Guidelines on validation procedures for meteorological data from automatic weather stations. *Journal of Hydrology*, *402*(1-2), 144–154.

Fiebrich, C. A., Morgan, C. R., McCombs, A. G., Hall Jr, P. K., & McPherson, R. A. (2010). Quality assurance procedures for mesoscale meteorological data. *Journal of Atmospheric and Oceanic Technology*, *27*(10), 1565–1582.

Fletcher, T. D., Andrieu, H., & Hamel, P. (2013). Understanding, management and modelling of urban hydrology and its consequences for receiving waters: A state of the art. *Advances in Water Resources*, *51*, 261–279.

Foody, G., See, L., Fritz, S., Moorthy, I., Perger, C., Schill, C., & Boyd, D. (2018). Increasing the accuracy of crowdsourced information on land cover via a voting procedure weighted by information inferred from the contributed data. *ISPRS International Journal of Geo-Information*, *7*(3), 80.

Ganeriwal, S., Balzano, L. K., & Srivastava, M. B. (2008). Reputation-based framework for high integrity sensor networks. *ACM Transactions on Sensor Networks (TOSN)*, *4*(3), 15.

Gharesifard, M., & Wehn, U. (2016). To share or not to share: Drivers and barriers for sharing data via online amateur weather networks. *Journal of Hydrology*, *535*, 181–190.

Guo, H., Huang, H., Sun, Y.-E., Zhang, Y., Chen, S., & Huang, L. (2018). Chaac: Real-time and fine-grained rain detection and measurement using smartphones. *IEEE Internet of Things Journal*, *6*(1), 997–1009.

Huang, K. L., Kanhere, S. S., & Hu, W. (2014). On the need for a reputation system in mobile phone based sensing. *Ad Hoc Networks*, *12*, 130–149.

Hunter, J., Alabri, A., & van Ingen, C. (2013). Assessing the quality and trustworthiness of citizen science data. *Concurrency and Computation: Practice and Experience*, *25*(4), 454–466.

Jiang, S., Babovic, V., Zheng, Y., & Xiong, J. (2019). Advancing opportunistic sensing in hydrology: A novel approach to measuring rainfall with ordinary surveillance cameras. *Water Resources Research*, *55*(4), 3004–3027.

Josang, A., & Ismail, R. (2002). The beta reputation system. In *Proceedings of the 15th bled electronic commerce conference* (Vol. 5, pp. 2502–2511).

Jøsang, A., Ismail, R., & Boyd, C. (2007). A survey of trust and reputation systems for online service provision. *Decision Support Systems*, *43*(2), 618–644.

Krajewski, W., & Smith, J. A. (2002). Radar hydrology: rainfall estimation. *Advances in Water Resources*, *25*(8-12), 1387–1394.

Lowry, C. S., & Fienen, M. N. (2013). Crowdhydrology: crowdsourcing hydrologic data and engaging citizen scientists. *GroundWater*, *51*(1), 151–156.

Meier, F., Fenner, D., Grassmann, T., Otto, M., & Scherer, D. (2017). Crowdsourcing air temperature from citizen weather stations for urban climate research. *Urban Climate*, *19*, 170–191.

Mosavi, A., Ozturk, P., & Chau, K.-w. (2018). Flood prediction using machine learning models: Literature review. *Water*, *10*(11), 1536.

Mousa, H., Mokhtar, S. B., Hasan, O., Younes, O., Hadhoud, M., & Brunie, L. (2015). Trust management and reputation systems in mobile participatory sensing applications: A survey. *Computer Networks*, *90*, 49–73.

Muller, C., Chapman, L., Johnston, S., Kidd, C., Illingworth, S., Foody, G., . . . Leigh, R. (2015). Crowdsourcing for climate and atmospheric sciences: current status and future potential. *International Journal of Climatology*, *35*(11), 3185–3203.

Ohba, M., & Sugimoto, S. (2019). Differences in climate change impacts between weather patterns: possible effects on spatial heterogeneous changes in future extreme rainfall. *Climate Dynamics*, *52*(7-8), 4177–4191.

Overeem, A., Leijnse, H., & Uijlenhoet, R. (2013). Country-wide rainfall maps from cellular communication networks. *Proceedings of the National Academy of Sci-*

ences, *110*(8), 2741–2745.
Quinn, N., Bates, P. D., Neal, J., Smith, A., Wing, O., Sampson, C., . . . Heffernan, J. (2019). The spatial dependence of flood hazard and risk in the united states. *Water Resources Research*, *55*(3), 1890–1911.
Resnick, P., Kuwabara, K., Zeckhauser, R., & Friedman, E. (2000). Reputation systems. *Communications of the ACM*, *43*(12), 45–48.
Ruan, Y., & Durresi, A. (2016). A survey of trust management systems for online social communities–trust modeling, trust inference and attacks. *Knowledge-Based Systems*, *106*, 150–163.
Sadler, J. M., Goodall, J., Morsy, M., & Spencer, K. (2018). Modeling urban coastal flood severity from crowd-sourced flood reports using poisson regression and random forest. *Journal of Hydrology*, *559*, 43–55.
Saksena, S., Dey, S., Merwade, V., & Singhofen, P. J. (2019). A computationally efficient and physically based approach for urban flood modeling using a flexible spatiotemporal structure. *Water Resources Research*.
Salman, A. M., & Li, Y. (2018). Flood risk assessment, future trend modeling, and risk communication: a review of ongoing research. *Natural Hazards Review*, *19*(3), 04018011.
Sanchez, L., Rosas, E., & Hidalgo, N. (2018). Crowdsourcing under attack: Detecting malicious behaviors in waze. In *IFIP International Conference on Trust Management* (pp. 91–106).
Savage, J. T. S., Pianosi, F., Bates, P., Freer, J., & Wagener, T. (2016). Quantifying the importance of spatial resolution and other factors through global sensitivity analysis of a flood inundation model. *Water Resources Research*, *52*(11), 9146–9163.
Sharma, A., Wasko, C., & Lettenmaier, D. P. (2018). If precipitation extremes are increasing, why aren't floods? *Water Resources Research*, *54*(11), 8545–8551.
Shen, Y., Morsy, M. M., Huxley, C., Tahvildari, N., & Goodall, J. L. (2019). Flood risk assessment and increased resilience for coastal urban watersheds under the combined impact of storm tide and heavy rainfall. *Journal of Hydrology*, 124159.
Silvertown, J., Harvey, M., Greenwood, R., Dodd, M., Rosewell, J., Rebelo, T., . . . McConway, K. (2015). Crowdsourcing the identification of organisms: A case-study of ispot. *ZooKeys*(480), 125.
Smith, J. A., Seo, D. J., Baeck, M. L., & Hudlow, M. D. (1996). An intercomparison study of nexrad precipitation estimates. *Water Resources Research*, *32*(7), 2035–2045.
Strobl, B., Etter, S., van Meerveld, I., & Seibert, J. (2019). The crowdwater game: A playful way to improve the accuracy of crowdsourced water level class data. *PloS one*, *14*(9).
The Weather Channel. (2018). *What is a gold star weather station?* Retrieved 2021-01-18, from `https://support.weather.com/s/article/What-is-a-Gold-Star-weather-station?language=en_US`
Villarini, G., Mandapaka, P. V., Krajewski, W. F., & Moore, R. J. (2008). Rainfall and sampling uncertainties: A rain gauge perspective. *Journal of Geophysical Research: Atmospheres*, *113*(D11).
Weeser, B., Jacobs, S., Kraft, P., Rufino, M., & Breuer, L. (2019). Rainfall-runoff modelling using crowdsourced water level data. *Water Resources Research*.
Wilby, R. L., & Keenan, R. (2012). Adapting to flood risk under climate change. *Progress in physical geography*, *36*(3), 348–378.
WXForum. (2018a). *Netatmo weather stations do not appear on weatherunderground anymore.* Retrieved 2021-01-18, from `https://www.wxforum.net/index.php?topic=35059.0`
WXForum. (2018b). *Netatmo weather stations do not appear on weatherunderground anymore.* Retrieved 2021-01-18, from `https://www.wxforum.net/index.php`

?topic=34815.0
Yang, H., Zhang, J., & Roe, P. (2013). Reputation modelling in citizen science for environmental acoustic data analysis. *Social Network Analysis and Mining*, *3*(3), 419–435.
Yang, P., & Ng, T. L. (2017). Gauging through the crowd: A crowd-sourcing approach to urban rainfall measurement and storm water modeling implications. *Water Resources Research*, *53*(11), 9462–9478.
Zahura, F. T., Goodall, J. L., Sadler, J. M., Shen, Y., Morsy, M. M., & Behl, M. (2020). Training machine learning surrogate models from a high-fidelity physics-based model: Application for real-time street-scale flood prediction in an urban coastal community. *Water Resources Research*, *56*(10), e2019WR027038.
Zhang, J., Sheng, V. S., Li, Q., Wu, J., & Wu, X. (2017). Consensus algorithms for biased labeling in crowdsourcing. *Information Sciences*, *382*, 254–273.
Zhang, W., Villarini, G., Vecchi, G. A., & Smith, J. A. (2018). Urbanization exacerbated the rainfall and flooding caused by hurricane harvey in houston. *Nature*, *563*(7731), 384–388.
Zheng, F., Tao, R., Maier, H. R., See, L., Savic, D., Zhang, T., … others (2018). Crowdsourcing methods for data collection in geophysics: state of the art, issues, and future directions. *Reviews of Geophysics*, *56*(4), 698–740.
Zhu, Z., Wright, D. B., & Yu, G. (2018). The impact of rainfall space-time structure in flood frequency analysis. *Water Resources Research*, *54*(11), 8983–8998.