A Bayesian model for quantifying errors in citizen science data: application to rainfall observations from Nepal

Jessica A Eisma[1,1], Gerrit Schoups[2,2], Jeffrey Davids[3,3], and Nick Van de Giesen[2,2]

[1]Purdue University
[2]Delft University of Technology
[3]California State University Chico

November 30, 2022

Abstract

High quality citizen science data can be instrumental in advancing science toward new discoveries and a deeper understanding of under-observed phenomena. However, the error structure of citizen scientist (CS) data must be well-defined. Within a citizen science program, the errors in submitted observations vary, and their occurrence may depend on CS-specific characteristics. This study develops a graphical Bayesian inference model of error types in CS data. The model assumes that: (1) each CS observation is subject to a specific error type, each with its own bias and noise; and (2) an observation's error type depends on the error community of the CS, which in turn relates to characteristics of the CS submitting the observation. Given a set of CS observations and corresponding ground-truth values, the model can be calibrated for a specific application, yielding (i) number of error types and error communities, (ii) bias and noise for each error type, (iii) error distribution of each error community, and (iv) the error community to which each CS belongs. The model, applied to Nepal CS rainfall observations, identifies five error types and sorts CSs into four model-inferred communities. In the case study, 73% of CSs submitted data with errors in fewer than 5% of their observations. The remaining CSs submitted data with unit, meniscus, and unknown errors. A CS's assigned community, coupled with model-inferred error probabilities, can identify observations that require verification. With such a system, the onus of validating CS data is partially transferred from human effort to machine-learned algorithms.

A Bayesian model for quantifying errors in citizen science data: application to rainfall observations from Nepal

J.A. Eisma[1], G. Schoups[2], J.C. Davids[3], N. van de Giesen[2]

[1]Department of Civil Engineering, University of Texas at Arlington, Arlington, Texas, USA

[2]Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, the Netherlands

[3]Department of Civil Engineering and College of Agriculture, California State University, Chico, California, USA

Key Points:

- A Gaussian mixture of regressions explains the likelihood of citizen scientists submitting erroneous observations
- Citizen scientists are sorted into communities based on characteristics and the type and frequency of errors in the data that they submit
- The distribution of errors in the data from citizen scientists evolves as they gain experience

Corresponding author: J.A. Eisma, `jessica.eisma@uta.edu`

Abstract

High quality citizen science data can be instrumental in advancing science toward new discoveries and a deeper understanding of under-observed phenomena. However, the error structure of citizen scientist (CS) data must be well-defined. Within a citizen science program, the errors in submitted observations vary, and their occurrence may depend on CS-specific characteristics. This study develops a graphical Bayesian inference model of error types in CS data. The model assumes that: (1) each CS observation is subject to a specific error type, each with its own bias and noise; and (2) an observation's error type depends on the error community of the CS, which in turn relates to characteristics of the CS submitting the observation. Given a set of CS observations and corresponding ground-truth values, the model can be calibrated for a specific application, yielding (i) number of error types and error communities, (ii) bias and noise for each error type, (iii) error distribution of each error community, and (iv) the error community to which each CS belongs. The model, applied to Nepal CS rainfall observations, identifies five error types and sorts CSs into four model-inferred communities. In the case study, 73% of CSs submitted data with errors in fewer than 5% of their observations. The remaining CSs submitted data with unit, meniscus, and unknown errors. A CS's assigned community, coupled with model-inferred error probabilities, can identify observations that require verification. With such a system, the onus of validating CS data is partially transferred from human effort to machine-learned algorithms.

1 Introduction

Communities worldwide face increasing uncertainty regarding extreme weather events due to climate change. Reliable weather forecasts allow a community to initiate proactive measures when anticipating an extreme event—measures that sometimes save hundreds, if not thousands of lives. Unfortunately, sparse weather data in many regions of the world inhibit coordinated response efforts of local and regional governments (Teague & Gallicchio, 2017, p. 218). Citizen science can help bridge such data gaps.

Citizen science programs, organized efforts to collect scientific data from members of the public, have become increasingly popular as advances in technology have made the data collection and submission process more accessible (Bonney et al., 2009; Newman et al., 2012). However, some traditional scientists continue to question the quality of data submitted by members of the public, and have yet to accept the legitimacy of scientific discoveries advanced by citizen scientists (Hunter, Alabri, & van Ingen, 2013; Riesch & Potter, 2014; Sheppard & Terveen, 2011). Others, however, have embraced citizen science as an effective means for increasing the spatial and temporal resolution of scientific data. Successful citizen science programs investigate the type and frequency of errors in the data collected by program participants and develop training initiatives designed to reduce errors (Bird et al., 2014; Crall et al., 2011; Davids et al., 2019).

Most citizen scientist programs conduct quality control of the data submitted by their participants. For example, citizen scientists report when they feel an earthquake and rank its strength for the United States Geological Survey's (USGS) Did You Feel It? program. The USGS removes outliers and aggregates reported intensities at zip code or city-level after processing the data through the Community Decimal Intensity algorithm (USGS, n.d.). Other citizen scientist programs invest significant time and energy into assuring the quality of their data. For example, citizen scientists submit rainfall depth observations to the SmartPhones4Water-Nepal (S4W-Nepal) program. S4W-Nepal checks the value of each submitted rainfall observation against an accompanying photograph of the rain gauge and manually corrects erroneous observations (Davids et al., 2019). The range of time and effort dedicated to conduct quality control for citizen science data varies greatly across programs.

Rainfall observations submitted by citizen scientists have immense potential to increase the scientific community's understanding of rain events which are, by nature, highly heterogeneous in space and time. Currently, only about 1.6% of the land surface on Earth lies within 10 km of a rain gauge, and rain gauges are notoriously inconsistent (Kidd et al., 2017). So much so that the correlation coefficient for rain gauges 4 km apart in the midwestern United States was less than 0.5 for instantaneous rainfall (Habib, Krajewski, & Ciach, 2001). Citizen science rainfall observation programs must contend with the systematic errors inherent in measuring rainfall, as well as the errors induced by the citizen scientists. Detailed investigations into the errors made by citizen scientists, such as the efforts of S4W-Nepal, can help increase the utility of citizen science data and inform future program development, and is the subject of this study.

Motivated by the need to reduce the time-cost for quality control of citizen science data without sacrificing effectiveness, this study seeks to develop a reliable, semi-automated method for identifying citizen science observations that require additional verification.

Most error analyses of citizen science data focus on identifying and removing outliers from a dataset. Trained filters flag outliers by identifying observations that do not fit within the expected range of values or classes, such as species range or allowable count (Bonter & Cooper, 2012; Wiggins, Newman, Stevenson, & Crowston, 2011). Some citizen science programs develop eligibility or trust rating procedures to identify users that are likely to submit correct observations (Delaney, Sperling, Adams, & Leung, 2008; Hunter et al., 2013). Ratings schemes that consider demographic and experience-related characteristics have potential for describing the variability in citizen science data reliability (Kosmala, Wiggins, Swanson, & Simmons, 2016). However, some individual citizen scientists do not submit enough observations to be accurately assigned a rating. To overcome such limitations, Venanzi, Guiver, Kazai, Kohli, and Shokouhi (2014) based their error analysis on four communities of citizen scientists, each with a distinctive pattern of errors.

Machine learning algorithms and hierarchical, generalized linear, and mixed-effects models have also been employed by a variety of citizen science programs to study errors in citizen science data (Bird et al., 2014; Venanzi et al., 2014). Generalized linear models have largely been used to study whether and how characteristics of citizen scientists affect the accuracy of their observations (Butt, Slade, Thompson, Malhi, & Riutta, 2013; Crall et al., 2011; Delaney et al., 2008). Mixed-effects models add a random-effects factor to generalized linear models, permitting the study of errors in relation to an unintended grouping effect, such as spatial clustering (Bird et al., 2014; Brunsdon & Comber, 2012). Alternatively, hierarchical models have been leveraged to study how citizen scientist errors relate to effort and site-level effects (de Solla et al., 2005; Fink et al., 2010; Miller et al., 2011). Lastly, machine learning has been used to study errors in qualitative citizen science data, such as species identification and labeling tweets (Cox, Philippoff, Baumgartner, & Smith, 2012; Lukyanenko, Wiggins, & Rosser, 2019; Venanzi et al., 2014). Machine learning has not yet been employed to identify erroneous citizen science observations for quantitative data. In addition, most machine learning citizen science research has focused on datasets that are relatively static or slow-moving in the fields of biology and conservation (Lukyanenko et al., 2019). To our knowledge, the study presented here is the first attempt to leverage machine learning to assess errors in quantitative citizen science data with high spatiotemporal variability. Despite the wide range of existing research on citizen science errors, flexible methods for analyzing errors in quantitative citizen science data remains largely unexplored.

The objective of this study is to improve quality control of quantitative citizen science data by developing a Bayesian inference model that discovers, explains, and possibly corrects the errors in observations submitted by citizen scientists. The following research questions will be explored:

1. How can the type and magnitude of citizen science data errors be automatically identified from citizen science data and corresponding ground truth?
2. Given a calibrated citizen scientist, to what extent can errors be detected and corrected without ground truth?
3. To what extent do citizen scientist characteristics help in identifying and screening errors?

A probabilistic graphical model was developed to address these questions based on assumptions about the probabilistic relationships between citizen scientists, their characteristics, and the magnitude of their errors. The probabilistic graphical model includes a regression clustering sub-model relating true and observed values and includes an unknown number of linear regressions. The model also includes a probabilistic sub-model relating citizen scientist characteristics to error types. Applied to the S4W-Nepal program, the model identifies unique error types within the S4W-Nepal citizen scientist rainfall observations, and groups citizen scientists into communities based on their characteristics and error profile. Each community is characterized by a distinct distribution of error types which indicates the likelihood that a submitted observation should be reviewed further. After testing and training, the model was applied to investigate three practical issues: the error evolution of citizen scientist data over time (research question 1), multiple observations of a single rainfall event (research question 2), and observations submitted by citizen scientists with unknown characteristics(research question 3).

2 Study Area

SmartPhones4Water Nepal (S4W-Nepal) partners with citizen scientists across Nepal to collect rainfall observations (see Figure 1). Across Nepal, rainfall is highly heterogeneous in space and time. Average annual rainfall in Nepal varies from 250 mm on the leeward side of the Himalayas to over 3,000 mm in the center of the country near Pokhara (Figure 1) (Nayava, 1974). The South Asian summer monsoon brings approximately 80% of Nepal's annual precipitation during the months of June to September (Nayava, 1974).

The majority of citizen scientists participating in S4W-Nepal's rainfall data collection efforts reside in the Kathmandu Valley, home to about 10% of Nepal's population (Vibhāga, 2012). While the average annual precipitation is approximately 1,500 mm in the city of Kathmandu and 1,800 mm in the surrounding hills, it is highly variable and unpredictable (Thapa, Ishidaira, Pandey, & Shakya, 2017).

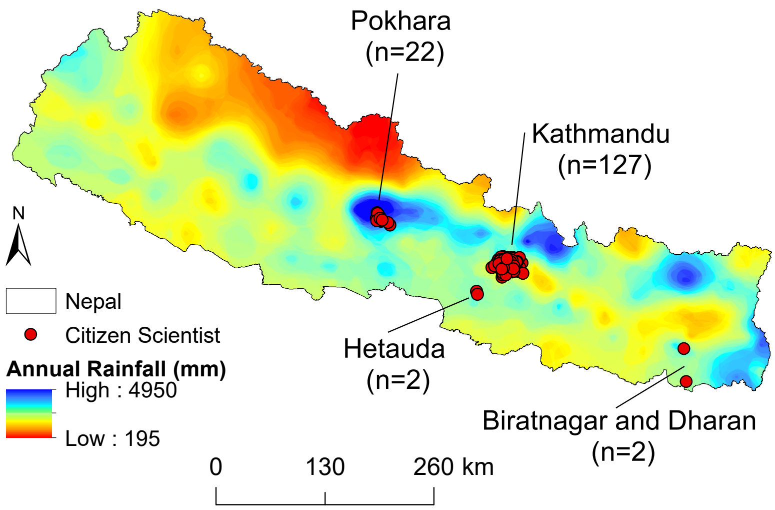

Figure 1. Locations of citizen scientists for which characteristics are known with the number of citizen scientists at specified locations shown in parentheses. Average annual rainfall grid created by USAID Nepal from observed data at 200 weather stations from 1980-2000.

3 Data

S4W-Nepal recruits citizen scientists to participate in a crowdsourced rainfall observation program in Nepal. S4W-Nepal collects the submitted observations via the Open Data Kit application for smart phones. Submitted observations include geo-location data, time of measurement, citizen scientist-reported depth of rainfall in millimeters, and a photograph of the rain gauge. The program is ongoing and has collected over 24,500 observations from over 265 citizen scientists since 2016.

3.1 Rain gauges

The participants were given a rain gauge constructed by S4W-Nepal and provided instructions on the proper installation and recording of rainfall data. The rain gauges were constructed from a re-purposed clear plastic bottle with a 100 mm diameter. The bottle was filled with a few centimeters of concrete to provide stability and a level measuring surface. The lid of the bottle was cut off where the taper ends, inverted, and placed flush with the top of the bottle to reduce evaporation losses. Finally, a ruler with millimeter precision was attached to the bottle to assist the reading of the rainfall depth (Davids et al., 2019).

3.2 Citizen characteristics

During the recruitment process, S4W-Nepal recorded characteristic data for 153 citizen scientists. Characteristics recorded were: motivation (paid/volunteer), recruitment method (personal connection, random site visit, social media, outreach), age (≤18, 19-25, >25), education (<Bachelors, Bachelors, >Bachelors), place of residence (urban, semi-urban, rural), occupation (agriculture, student, other), and gender (male, female). Citizen scientist characteristics will be used here to relate individual citizen scientists with the likelihood of errors in the data they submit.

4 Methods

4.1 Identification of erroneous observations

To detect erroneous rainfall observations submitted by citizen scientists, S4W-Nepal checks the value of each submitted rainfall observation against the accompanying rain gauge photograph. If they detect an error, the correct rain depth is recorded while preserving the record of the original value submitted by the citizen scientist. This allows S4W-Nepal to track the types and frequencies of errors made by the citizen scientists. Overall, approximately 9% of submitted rainfall observations are erroneous. Meniscus errors are the most common (58% of errors; records capillary rise), followed by unknown errors (33%), and unit errors (8%; records data in centimeters rather than millimeters) (Davids et al., 2019).

4.2 Model development

4.2.1 Assumptions and model structure

A Bayesian probabilistic graphical model was developed based on a number of assumptions about the data being modeled. These assumptions were used to inform the relationships between the variables and ensure the model accurately represents the modeler's understanding of the physical processes that underlie the data (Winn, Bishop, Diethe, Guiver, & Zaykov, 2020). The following assumptions informed the development of the citizen science errors inference model:

1. Each citizen scientist belongs to a single community.
2. A citizen scientist's community is defined by their collective demographic and experience-related characteristics and the type and frequency of errors they have made in prior submissions.
3. Each citizen scientist in a particular community always submits an observation with a community-specific error type distribution.
4. Each citizen scientist observation relates to an underlying true value with a systematic bias and random noise level that depends on the error type of the observation.

While the tendency of citizen scientists to make errors may change as they gain experience, the model developed here assumes that a citizen scientist will not change communities over time. This simplifies the model while also including the potential impact of experience as a citizen characteristic. Citizen scientist demographic information was assumed to be a factor in determining community, because demographics, such as age, experience, and education, are a useful predictor in citizen scientist performance (Crall et al., 2011; Delaney et al., 2008; Sunde & Jessen, 2013). Furthermore, motivation and recruitment method were predictive factors in citizen scientist participation rate (Davids et al., 2019). The predictive power of demographics in determining community will be assessed. An additional assumption was incorporated, due to the nature of rainfall data: the inferred true value of rainfall was assumed to be between 0 and 540 mm. Rainfall events cannot result in negative rainfall, and 540 mm is the maximum one-day rainfall recorded for Nepal. Similar assumptions unique to a specific type of citizen science observation may be necessary at this stage of model development for application to other citizen science programs.

These assumptions are translated into the following set of equations describing the probabilistic relationship between model variables. The terminology and symbology used here is based on probabilistic graphical models (Winn et al., 2020). We first state the main statistical relations used in the model and have provided clarifications for the wider hydrological community. The community γ to which citizen scientist S belongs is a discrete random variable drawn from a discrete distribution denoted by Dis with probability vector **PCom** that specifies the prior probability of each community occurring within the citizen scientist population:

$$\gamma_s \sim Dis(\mathbf{PCom}|s), \tag{1}$$

We use a lowercase subscript to denote a random variable index (e.g. γ_s indicates there is a community variable for each citizen scientist S). Greek letters represent latent (inferred) variables, and Latin letters to represent observable variables. The value $Z_{c,s}$ of citizen characteristic c for citizen scientist s is assumed to be from a discrete distribution with probability vector **PChar** that depends on the characteristic c under consideration and the community γ_s the citizen scientist belongs to:

$$Z_{c,s} \sim Dis(PChar_c|\gamma_s), \tag{2}$$

Equation 2 quantifies the probabilistic relationship between each citizen characteristic and each assigned community in the form of a conditional probability table. Similarly, Equation 3, below, describes the conditional probability table for each error type and community. The error type $\varepsilon_{s,e}$ of event e observed by citizen scientist s is assumed to be from a discrete distribution with probability vector **PErr** that depends on the community γ_s that the citizen scientist belongs to:

$$\varepsilon_{s,e} \sim Dis(PErr|\gamma_s), \tag{3}$$

As seen in Equations 1-3, the model assigns each citizen scientist to a single community based on their characteristics and the type and frequency of errors they make.

Next, we quantify systematic (bias) and random (noise) differences between observations and underlying true values by means of a linear regression model parameterized by an error-type specific slope α, offset β and precision (inverse variance) τ:

$$O_{s,e} \sim \mathcal{N}(\alpha_{\varepsilon_{s,e}} \vartheta_e + \beta_{\varepsilon_{s,e}}, \tau_{\varepsilon_{s,e}}), \tag{4}$$

where $O_{s,e}$ represents the observed amount of rainfall in event e submitted by citizen scientist s, and ϑ_e is the corresponding true rainfall amount for event e. Given the error type of an observation, the observed value is thus drawn from a Gaussian distribution with mean equal to an error-type specific linear function of the true value and an error-type specific variance. α, β, and τ depend on error type $\varepsilon_{s,e}$. It follows that unconditionally, i.e. without knowing the error type, the relation between observed and true value is a mixture of error-type specific Gaussian distributions, with the weight of each Gaussian distribution in the mixture given by the probability of the corresponding error type. Finally, the model is completed by specifying priors for the regression parameters (α, β, τ) and the probability vectors ($\mathbf{PCom}$, $\mathbf{PChar_c}$, $\mathbf{PErr}$). The priors were different for the training and testing phases and are detailed below.

4.2.2 Model implementation

We implemented the probabilistic model using Microsoft Research's open source Infer.NET software framework (Minka et al., 2018). The Infer.NET framework provides adaptable tools to develop and run Bayesian inference for probabilistic graphical models. The modeler must define the variables, the dependencies between variables, and provide prior distributions for the variables that will be inferred. For implementation in Infer.NET, Equations 1-4 are translated into a factor graph as shown in Figure 2. The factor graph completely describes the joint posterior probability of the model (see Equation A.5). The factor graph includes observable and latent (inferred) variables, factor nodes, edges (arrows), plates, and gates. Variables are depicted by shaded or unfilled ellipses. A shaded variable is an observable value; an unfilled variable is a latent value. Factor nodes are the small black boxes connected to variables, describing the relation between variables connected to the factor. Edges (directional arrows) connect factor nodes to variables (Winn et al., 2020).

Plates. Plates are the large boxes outlined in gray surrounding portions of the factor graph. Plates are a simplified way to express repeated structures. The number of times a structure will be repeated is based on the index variable shown in the bottom right corner of the plate (Winn et al., 2020). For example, in Figure 2, the structure within the characteristics plate is repeated nine times, because the model considers nine different CS characteristics: motivation, recruitment, age, education, place of residence, occupation, gender, performance, and experience.

Gates. Gates are indicated by a dashed box, as seen around the Regression factor node in Figure 2. Gates essentially act as a switch, turning on and off depending on the value of the selector variable, which is the error type here (Minka & Winn, 2008). When gates are used to define a distribution, that distribution is a mixture.

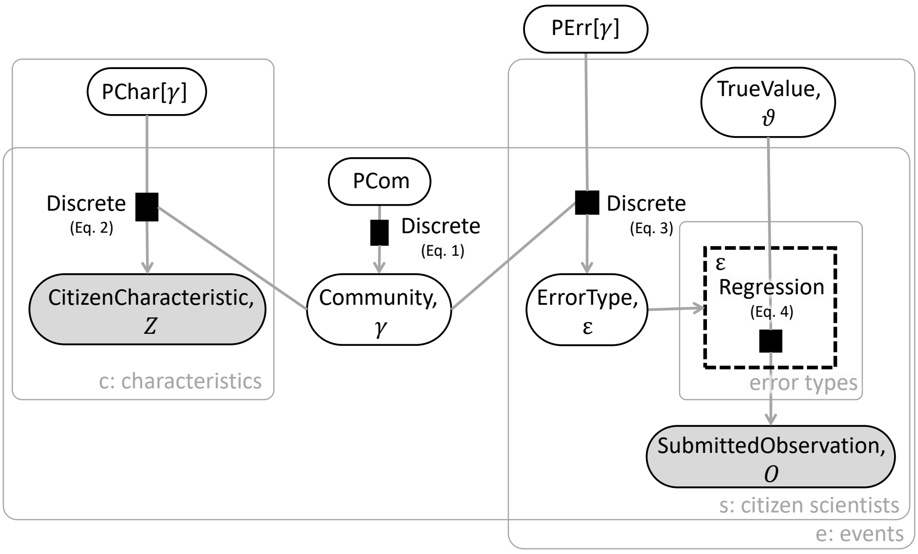

Figure 2. The citizen science error model depicted as a factor graph. A factor node represents a probabilistic relation between variables in the model and is shown by a black square. A variable is shown in an oval, with shading identifying observable variables. Arrows depict the output variable of each factor. A gate is represented by a dashed box. Plates are represented by gray rectangles with rounded corners. Symbols adopted from Winn et al. (2020).

Infer.NET generates a computationally efficient code for the inference algorithm using one of three available inference algorithms: expectation propagation, variational message passing, and Gibbs sampling. The model developed here employs the expectation propagation algorithm, because it is time efficient but reasonably accurate (Minka, 2013). Expectation propagation is a deterministic approximate inference algorithm for computing the marginal posterior distribution of each variable in the model (Minka, 2013). Each posterior distribution is assumed to take a specific parametric form in an exponential family (e.g. Gaussian, Gamma, discrete). The algorithm then aims to find parameter values for each parametric posterior that result in a good approximation of the exact posterior in terms of moment matching. For example, for a Gaussian approximation, expectation propagation will find a Gaussian whose mean and variance approximate those of the actual posterior. This is done using an iterative approach that starts from an initial guess for the approximate posteriors, and iteratively refines each posterior in turn via moment matching. Since all individual posterior updates depend on each other, the algorithm is iterated until all updates and posteriors stabilize (here, in <5 iterations). The final posteriors are not necessarily unique and may depend on how the algorithm was initialized. Here, we adopt a random initialization strategy for mixture models as used in Nishihara, Minka, and Tarlow (2013) and Minka et al. (2018) and evaluate non-uniqueness in the inferred posteriors using multiple runs with different random initialization.

4.2.3 Community and error selection

To select the appropriate number of communities to capture the differences among the citizen scientists, model evidence was used. Model evidence indicates which model best explains the data relative to the model's complexity (MacKay, 2003, p. 343-386). While the model evidence is notoriously hard to compute, expectation propagation provides a convenient estimate as a by-product of its posterior approximations. Model evidence calculation in Infer.NET is achieved by inferring posterior component weights of a mixture consisting of two components, i.e. the entire model and the empty model (Minka, 2000).

Too many communities may lead to overfitting, whereas too few communities may lead to underfitting. The model evidence automatically makes this trade-off and identifies the optimal number of communities. Model evidence was computed for models with one to ten communities. The number of communities that resulted in the largest model evidence was selected as the correct number of communities for the model and data. Similarly, model evidence was used to determine how many error types were present in the data. Model evidence was computed for one to twelve error types while using the optimal number of communities. The number of error types that resulted in the largest model evidence was selected as the number of error types for the model and data. After selecting the number of error types, model evidence was again checked to verify that the optimal number of communities remained constant. Selecting the error types via model evidence may identify more error types than expected, but the Bayesian model accounts for all possibilities and selects the one that most accurately represents the data.

4.2.4 Training and testing the model

The inference model was trained and tested to ensure model performance was consistent across different groups of data. During training and testing, the following characteristics were known for each citizen scientist: motivation, recruitment, age, education, place of residence, occupation, gender, performance, and experience. The first seven characteristics were recorded by S4W-Nepal (as explained in Section 3). The last two characteristics, performance and experience, were defined based on the observations submitted by each citizen scientist. Performance is simply the percentage of observations submitted by a citizen scientist that did not require correction. A performance of 90% indicates that 90% of that citizen scientist's submitted observations matched the true value shown in the associated photograph. Experience is a count of how many observations a citizen scientist submitted through the 2018 monsoon season. Performance and experience rates were split into three levels based on natural breakpoints in their respective histograms.

Splitting the data. Rainfall observations submitted by citizen scientists with known characteristics from 2016 to 2018 were randomly split into a training data set and a testing data set. The training set consisted of 92% of available observations, representing 6,091 observations submitted by 152 citizen scientists. The citizen scientists in the training set submitted anywhere from 1 to 159 observations, with the average number of submissions being 43.5. The testing set consisted of the remaining 8% of available observations, representing 527 observations from 109 citizen scientists. The citizen scientists in the testing set submitted anywhere from 1 to 159 observations, with the average number of submissions being 57.4. All citizen scientists in the testing set were also in the training set. Note that individual observations in each group were unique.

Training the model. Before training the model, prior distributions were set for the variables that were inferred. Uniform prior distributions were set for the citizen characteristics (see Equation A.1), community (see Equation A.2), and error (see Equation A.3). The prior distribution for the true value parameter was a Gaussian distribution with a mean equal to the average value of all submitted observations (15) and the four times the variance of the entire dataset (2400; see Equation A.4). A true value prior variance of 2400 was chosen to reduce small event bias and accommodate inference of large rainfall observations. The prior distributions for the Gaussian mixture parameters (α, β, and τ) were assigned based on the magnitude of unit, meniscus, and unknown errors classified by Davids et al. (2019).

While running the model in the training phase, the characteristics for each citizen scientist, the submitted observations, and the true values were known. The community for each citizen scientist, the error type for each submitted observation, the conditional probability tables for each characteristic and error type, and parameters for the Gaussian mixture were inferred (see Equations 2-4 and Figure 2). The training phase provided posterior distributions that were then used while testing the model.

Testing the model. To test the model, prior distributions for latent variables were set to the associated posterior distribution calculated during training. The characteristics for each citizen scientist and the values of the submitted observations were set. The model inferred the community for each citizen scientist, the probable error type for each observation, and provided a posterior distribution for the true value of the submitted observation. The performance of the model was assessed based on the whether the inferred posterior distribution for true value (ϑ) covered the true value identified in the accompanying photograph submitted by the citizen scientist and whether the mode of the true value posterior matched the actual true value.

A synthetic rainfall event was created to explore how many observations of a single event are needed to produce a reliable estimate of the event's true value. A synthetic observation of the event was created by first assigning an error type to each citizen scientist based on the distribution of errors for their respective error communities (see Table 2). Then, the value of the synthetic observation was calculated using Equation 4, the α, β, and τ values from Table 1 with a true value of 15 mm. Multiple synthetic events were created with two to three observations of the same event with one to two erroneous observations per event. The true value of each synthetic event was predicted by the model.

5 Results and Discussion

5.1 Number of communities and error types

Model evidence indicated that there are four communities and five error types present in the data, given the model structure. In comparison, S4W-Nepal identified four error types in the data based on visual inspection of the submitted observations. The inference model, however, is a much more powerful tool for uncovering nuances in the data than graphical techniques. Therefore, the number of communities and error types inferred from the model were used for the remaining analysis.

5.2 Error analysis

Parameters for the error-specific linear regressions were inferred for the five error types in the submitted rainfall observations (see Table 1 and Figure 3). The inferred parameters included the mean and precision, τ, of the Gaussian distribution, where the mean is based on a linear regression α, β, and ϑ as shown in Equation 4. Four of the five error types align well with the error types identified by Davids et al. (2019): none, unit, meniscus, and unknown. Meniscus errors occur when a citizen scientist reports the top of a concave meniscus rather than the bottom of the meniscus. Unit errors indicate instances where a citizen scientist submitted an observation in units of centimeters rather than millimeters, resulting in a unit error slope, α, of 0.10. Unknown errors do not present a discernible pattern that would explain their origin, as indicated by the low inferred precision (0.01) for this error type. Figure 3 shows that the model-inferred error types are accurate, with only the unknown error type encompassing highly variable submitted observation/true value pairs.

The inference model identified one error type that was overlooked during the Davids et al. (2019) analysis of errors in the Nepal citizen science data: slope outliers. Slope outliers signify a case where the citizen scientist's reported observation was approximately ten times greater than the true value evident in the accompanying photograph of the rainfall gauge. The underlying cause of outlier errors is unclear, but these outliers can likely be attributed to typos (e.g. adding an additional zero) or a mistake made by reading the gauge from the wrong direction (e.g. top down). Of the 6,091 observations included in the training data, only two were labelled as slope outliers.

Table 1. Inferred regression parameters for the different error types

Error Type	Slope, α	Intercept, β	Precision, τ
None	1.00	0.00	55750.04
Unit	0.10	0.07	36.89
Meniscus	1.00	2.54	1.74
Unknown	0.97	2.37	0.01
Slope Outlier	10.31	-0.69	1.50

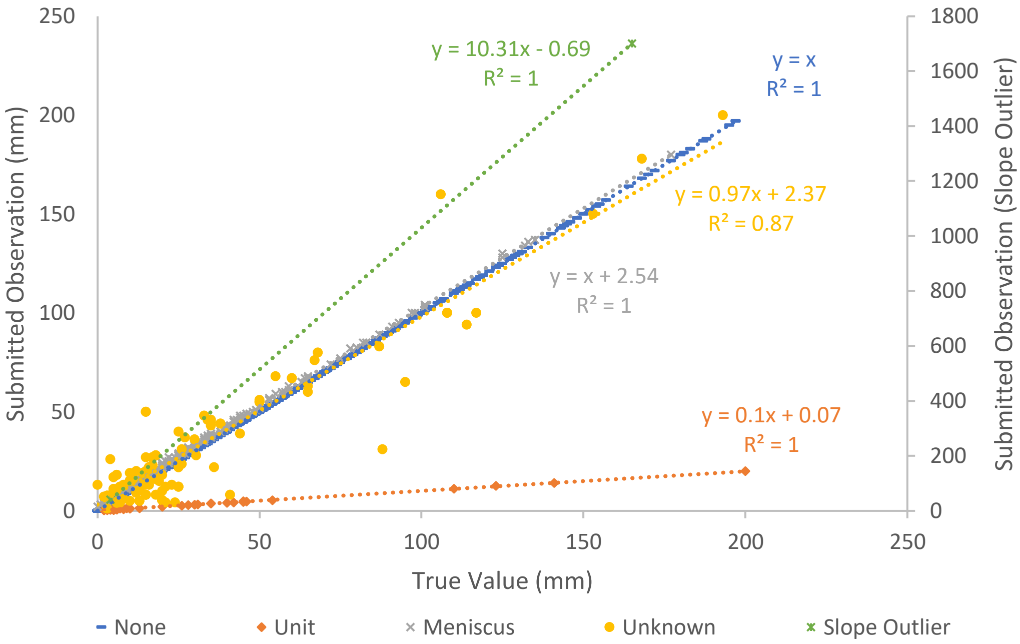

Figure 3. Inferred error types for each pair of submitted observation and true value of rainfall in the training dataset.

5.2.1 Error distribution within communities

The distribution of errors committed by citizen scientists varied depending on the assigned community, as seen in Table 2. Each community was named based on its respective error distribution: Few, Few-MUn, Mensicus, and Random Unknown (RandU). The Few community makes very few errors—only 2% of submitted observations are erroneous. Of the erroneous submissions, members in the Few community are most likely to make meniscus or unknown errors (1% each). The Few-MUn community also makes

Table 2. Distribution of errors made by citizen scientists in each community

Community	None	Unit	Meniscus	Unknown	Slope Outlier
Few (0.47)	0.98	0.00	0.01	0.01	0.00
Few-MUn (0.26)	0.95	0.00	**0.03**	**0.02**	0.00
Meniscus (0.20)	0.80	0.01	**0.17**	0.02	0.00
RandU (0.07)	0.78	0.06	0.06	**0.11**	0.00

Note : The probability of each community is shown in parentheses after the community name. Bold values indicate the most common error type(s) for each community. The probabilities may not add to 1 due to rounding.

relatively few mistakes but does so at a rate of 5%. Members of the Few-MUn community are almost equally likely to make meniscus errors (3%) and unknown errors (2%). The two other communities, Meniscus and RandU, are much more likely to submit erroneous rainfall observations. The Meniscus community submits erroneous observations at a rate of 20%. These observations are largely erroneous due to citizen scientists reading the meniscus of the water incorrectly (17%). Lastly, the RandU community makes the most errors, with 22% of its observations requiring correction. While the RandU community makes primarily unknown errors (11%), meniscus (6%) and unit (6%) errors still represent a large portion of the erroneous submissions. Members of the RandU community are prone to making a wide variety of errors.

The Few community members may have a high degree of scientific literacy; more than 97% of Few community members have at least a Bachelor's degree. The Few-MUn community members may also have high scientific literacy but occasionally make mistakes. Citizen scientists that were initially error prone but were able to correct their misunderstandings based on the feedback provided by S4W-Nepal may also be assigned to the Few-MUn community. For example, one citizen scientist in the Few-MUn community made 3 mistakes in the first 16 submissions, but then submitted 44 observations over the next 1.5 years without making a mistake. The Meniscus community largely misunderstands how to correctly read the depth of water in the rain gauge. The RandU community has several misunderstandings that cross multiple error types, therefore citizen scientists in this community make a mix of errors.

The distribution of errors within each community is a useful tool not only for selecting which submitted observations might require verification, but also for identifying opportunities to improve or maintain the overall accuracy of submitted observations. Citizen science project organizers can use targeted training to help specific communities improve their performance (Budde et al., 2017; Sheppard & Terveen, 2011). For example, S4W-Nepal could occasionally send feedback messages to the meniscus community members reminding them to read the rainfall depth from the bottom of the meniscus. As another example, members in the Few community might positively respond to general feedback messages acknowledging their strong record of accurate observations and choose to remain engaged with the program. Knowing the error structure of observations submitted by different communities may help improve the overall effectiveness of citizen science programs.

5.3 Community composition

The model grouped citizen scientists into four distinct communities with a unique combination of characteristics and probability of making errors. The Few community is the largest with 47% of citizen scientists in the training group assigned to this community (see Table 2). The RandU community is the smallest with only 7% of citizen scientists classified into this group. The remaining citizen scientists are grouped into the Few-Un (19%) and Meniscus (16%) communities. Overall, only 24% of participating citizen scientists are likely to make errors in more than 8.3% of their submitted observations.

The probability that a citizen scientist will belong to a specific community depends, in part, on the unique characteristics of that citizen scientist. Figure 4 provides the posterior probability that a citizen scientist with a particular characteristic would belong to each community, offering insight into the characteristic composition of each community. Singular characteristics may have a large impact on the tendency of a citizen scientist to make errors, and therefore to be assigned to a specific community. However, it is also true that any combination of characteristics could contribute to the probability of a citizen scientist being assigned to a community. In some cases, citizen scientists are likely to possess a similar combination of characteristics, which surfaces in the community distributions. For example, Figure 4 indicates that citizen scientists recruited during a random visit, older than 25 years of age, holding less than a bachelor's degree, and with an "other" occupation make up 20% of all citizen scientists in the project and have a similar community distribution. While community assignment trends for singular characteristics can be enlightening, the impact of multiple citizen scientists with a similar combination of characteristics must be acknowledged.

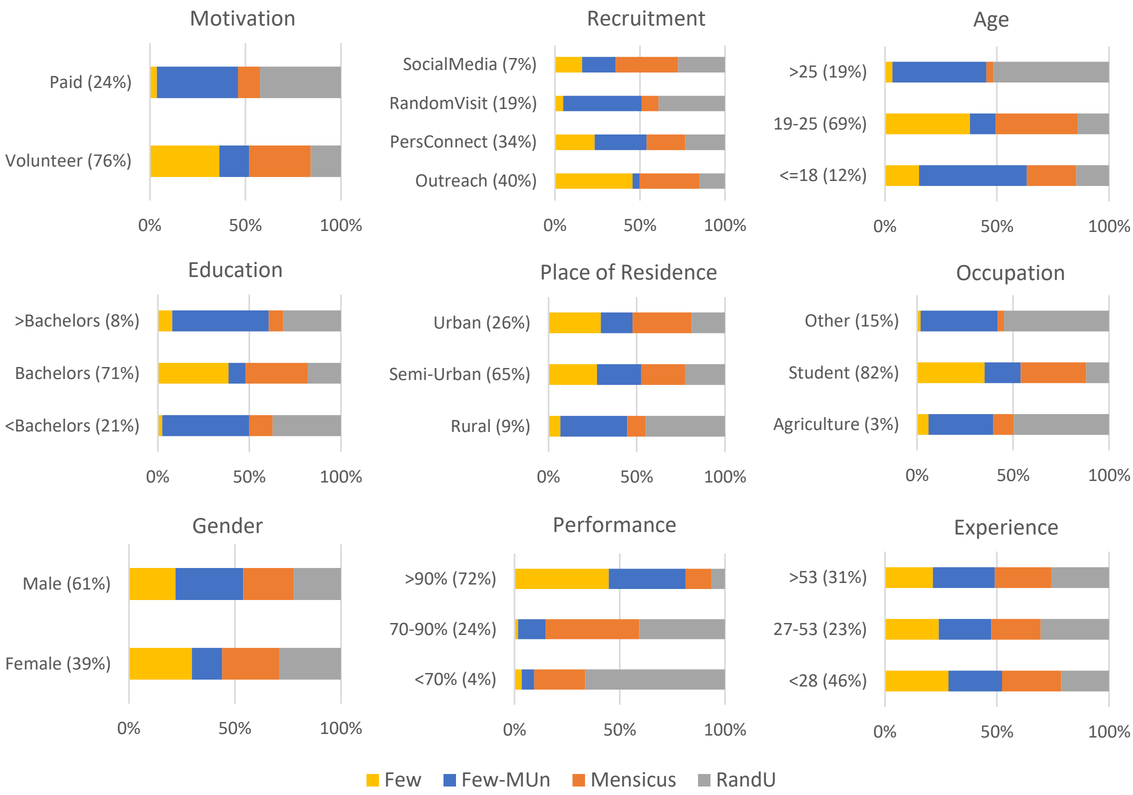

Figure 4. Community composition for each characteristic. The percentage of participating citizen scientists with the associated characteristic is shown in parentheses.

5.4 Sensitivity of α, β, and τ Priors and algorithm initialization

In the model application examined here, Davids et al. (2019) provided prior information on the types of errors in the data, but such information will not always be available. Prior information on the types of errors in the data is useful but not necessary to identify some of the errors made by participating citizen scientists. When prior error information is known, the model reliably infers the same five errors, even when the uncertainty of this information is high (i.e. high variance assigned to the Gaussian prior distributions). When no prior information is known about the potential types of errors present in the data (i.e. $\alpha_\varepsilon \sim \mathcal{N}(1, 100)$, $\beta_\varepsilon \sim \mathcal{N}(0, 100)$), the model reliably infers the no error type and splits the meniscus error into two error types—a 2-mm meniscus error and a 3.8-mm meniscus error. The two remaining error types identified are variations on the unknown error with relatively low R^2 values, 0.79 and 0.09 compared with 1.0 for the none and meniscus errors. The model may fail to identify the unit error type, because it occurs in only 0.7% of submitted observations. Multiple local optima exist for the error types, and the model may fail to identify all unique errors if no prior information on the errors is known. Regardless of whether error information is known previously, model evidence indicated that four communities and five error types best capture the variance in the data. The model may require many more iterations (possibly up to 100) to converge when the priors are vague.

There is also some variation in the inferred posterior distributions that is based on how the algorithm is initialized, but the variation is insignificant ($p>0.05$). Changing the algorithm initialization during inference minimally affects the posterior distributions of the error types. For example, with a different initialization, the α, β, and τ of the slope outlier change from (10.31, -0.69, 1.5) to (10.31, -0.24, 1.5). The α, β, and τ values of the remaining error types are more consistent than the slope outlier type, regardless of how the algorithm is initialized.

5.5 Inferring the true value of a submitted observation

In addition to providing insight into the error structure of the submitted observations and the relationship between citizen scientist characteristics and error tendencies, the model provides information about the true value of submitted observations. Testing the model reveals that the model can infer a previously unknown true value based on the value of the submitted observation and the characteristics of the citizen scientist. The inferred true value differs from the actual true value by a median percent error of 0.9%. The standard deviation of percent error is, however, 98.8%. With a wide true value prior distribution (here, 24,000; see Eq. A.4), the model has a tendency to over-predict unit errors for a small number of observations submitted with a value of 6 mm or lower which causes the large standard deviation (see Figure 5a). In most cases, the actual true value of the submitted observation falls within the range of the posterior distribution inferred for the true value variable as seen in Figures 5b,c. However, as Figures 5b,c show, the mode of the posterior distribution is not always a good estimate of the actual true value.

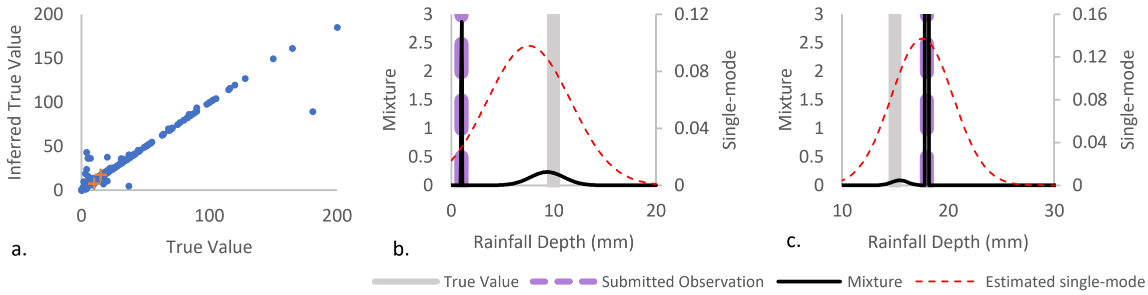

Figure 5. a. The inferred true value is usually a good estimate of the true value of the submitted observation. In some erroneous submissions, the mode of the estimated single-mode posterior is not equal to the true value, however an exact Gaussian mixture of the true value posterior distributions has a local peak at the true value of an observation submitted with a (b.) unit error and a (c.) meniscus error. The points shown in (b.) and (c.) are indicated by a plus (+) in (a.).

To increase the computational efficiency of an inference algorithm that sometimes needs to consider thousands of variables, expectation propagation approximates a multi-mode posterior distribution with a single-mode distribution (Minka et al., 2018) by minimizing the Kullback-Leibler divergence between the two (Minka, 2005). In many applications, this method works very well. However, here, the mixture distribution covers values ranging from 10% (unit error) of the true value up through 1,000% (slope outlier error) of the true value. Such a wide range of possible true values results in a predicted true value posterior with high variance and a mode that is occasionally shifted left or right of the true value (see Figures 5b,c).

While the predicted single-mode true value posterior distribution does not always estimate the actual true value of an erroneous submission well, the exact Gaussian mixture posterior often exhibits a local peak at the actual true value (see Figures 5b,c). The mode of the Gaussian mixture posterior usually presents at the value of the submitted observation because of the high precision associated with the none error type (see Table 1). Only 8.7% of submitted observations have greater than a 20% probability of being erroneous in this example application. Therefore, the inferred error type posterior

Table 3. Synthetic tests inferring true value from multiple observations submitted for a single event with a true value of 15 mm

No. Obs.	Error Types	Inferred True Value	Inferred Variance	No. Obs.	Error Types	Inferred True Value	Inferred Variance
2	0, 1	14.98	6.26E-2	2	2, 4	168.10	2.89
2	0, 2	15.00	1.39E-3	3	0, 2, 4	15.00	6.54E-5
2	0, 3	14.99	1.39E-2	3	0, 3, 4	15.00	5.43E-5
2	0, 4	153.85	3.45	2	1, 3	16.69	4.23E-1
	Different	15.00	9.04E-5	3	0, 1, 3	14.99	1.66E-2
	CS community	15.00	5.94E-2	2	1, 2	17.35	5.96E-1
	combinations	15.00	5.94E-2	3	0, 1, 2	15.00	3.27E-3
	...	15.00	5.94E-2	2	2, 3	17.59	1.91E-1
3	0, 0, 4	150.70	1.22	3	0, 2, 3	15.00	3.29E-3
4	0, 0, 0, 4	150.50	0.64				

Note : Error Types: 0=None, 1=Unit, 2=Meniscus, 3=Unknown, 4=Slope Outlier

distribution may be examined in conjunction with the Gaussian mixture posterior to provide additional information on the probability of each error type. For example, despite the mode of the Gaussian mixture posterior being located at the value of the submitted observation in Figure 5b, the probability of a none error type is only 0.23, and the unit error probability is 0.73. The Gaussian mixture posterior and the error type posterior distributions may provide a more accurate representation of the true value of a submitted observation than the approximated single-mode Gaussian posterior distribution.

5.5.1 Multiple observations of a single event

If only a single observation of a rainfall event is available, the predicted error type is based on the error types observed during model training. However, analyzing multiple observations of a single rainfall event should improve the accuracy of the inferred error type and true value of rainfall.

For each of the simulations described below, the model was not given any information about the error types associated with the submitted observations. The model inferred the true value solely based on what it learned during model training. When only one error was made out of two observations submitted, the model predicted the true value every time except for instances of a slope outlier error (see Table 3 column 1). In such cases, the ability of the model to correctly infer the event true value was related to the error communities of the citizen scientists. Through 12 trials (not shown) with different algorithm initialization and combinations of citizen scientists from the Few-MUn and Meniscus communities, the model correctly inferred the true value only twice. However, the model was able to infer the true value if one submitted observation had a slope outlier for other combinations of citizen scientist communities (see Table 3 column 1). If one slope outlier observation was paired with two or more correct observations, the model consistently failed to infer the correct true value. The low probability of a slope outlier combined with the relatively high probability of unit and meniscus errors cause the model to infer the slope outlier as a meniscus error and the correct observations as unit errors. When one slope outlier error was paired with another error, the model required an additional correct observation to accurately predict the true value (see Table 3 column 2). For the best performance, the slope outlier error needs to be paired with at least one other erroneous observation and a correct observation. When two errors were made out of two observations submitted, the model often failed to correctly predict the true value. However, when a third observation without an error was included, the model predicted the true value every time (see Table 3). Overall, the model inferred the correct error types when the inferred true value was also correct.

For instances when multiple observations of a single event are submitted, at least one error-free observation is likely necessary to ensure that the model predicts the true value with minimal uncertainty. When multiple erroneous observations are submitted, the model performs best when at least one correct observation is submitted of that same event. Given that over 90% of submitted observations do not have an error, it is unlikely that an erroneous observation would be submitted without a complementary error-free observation, assuming that additional citizen scientists are active.

5.6 Further model applications

The trained model was tested for two unique applications that provide insight into the utility of the model in practical applications and the distribution of errors in citizen science data over time.

5.6.1 Citizen scientists with unknown characteristics

As citizen scientist programs expand, recording complete characteristics data for each participating citizen scientist may become challenging. The model's ability to infer the correct community for citizen scientists with unknown characteristics and the correct true value for the observations they submit was investigated. The characteristics for each unknown citizen scientist were selected from a discrete distribution estimated from the characteristics data of citizen scientists observed during training. The prior distribution of the community, $PCom$, was set to a discrete distribution equal to the overall community posterior distribution of the training set. The community for each citizen scientist and the true values of their submitted observations were inferred and compared to the communities and true values inferred when the characteristics were known precisely, but the community was also unknown.

The model performed quite well while inferring the community of unknown citizen scientists and the true values of observations submitted by unknown citizen scientists. Communities of citizen scientists with known characteristics were correctly predicted 0.9% more than citizen scientists with unknown characteristics. The coefficient of determination between the actual true values and predicted true values was 0.015 higher for known citizen scientists than for unknown citizen scientists. While the predicted true values for known and unknown citizen scientists were similar, the uncertainty of the true values predicted from observations submitted by unknown citizen scientists was higher. The average variance of the inferred true value posteriors was 140.2 mm^2 for unknown citizen scientists and 125.6 mm^2 for known citizen scientists. Overall, the value of submitted observations has greater influence on the inferred true values of rainfall than the characteristics of the associated citizen scientist. While knowing the characteristics of all citizen scientists increases the accuracy of predicting the true value of submitted observations, it is not essential.

5.6.2 Evolution of error structure within communities

The change in error distribution over time within each community was studied. The observations submitted by citizen scientists with known characteristics were divided into years 2017, 2018, and 2019. The same communities assigned to each citizen scientist during training were assigned, and the α, β, and τ for each error type inferred during training were made static. In addition, a uniform prior was set for the community error distributions to reduce skew in the posterior distribution. Then, the inference model was run to infer the error distribution for each community during each year.

The probability that a citizen scientist in each community would commit a type of error changed from the 2017 to 2018 to 2019 S4W-Nepal program years (see Figure 6). In 2017, only 16 citizen scientists for whom characteristics are known submitted observations (see Table 4). The 2017 community error distributions, particularly the Few-MUn, Meniscus, and Unit-MUn communities, are highly uncertain due to the small sample size. Overall, citizen scientists became increasingly active as S4W-Nepal's program progressed through the years. Citizen scientists submitted an average of just over 8 observations in 2017, growing to 80 by 2019. In the first full year of rainfall submissions (2017), most citizen scientists were assigned to the Few-MUn community. In the following two years, active citizen scientists were most often in the Few community, followed by the Few-MUn community. In all three years of S4W-Nepal's program, the RandU community represented the smallest fraction of active citizen scientists.

As S4W-Nepal gained experience in operating a citizen science program, the participating citizen scientists also gained skills in collecting and submitting accurate rainfall observations. The Meniscus community had an increasing probability of submitting correct observations in each year after 2017, while the Few-MUn community maintained a low probability of submitting an erroneous observation (see Figure 6). The Few and RandU communities also increased their probability of submitting a correct observation in 2018 but saw a decrease in 2019. As the years progressed, all communities submitted the same or successively fewer meniscus errors. Similarly, unit errors tended to decrease or remain the same as citizen scientists gained experience. Interestingly, while meniscus type errors and unit errors decreased over time, 2019 saw relatively high rates of unknown errors. The reason for an increase in unknown errors is difficult to diagnose but may be due to an evolution in the magnitude of errors committed. For example, if the

Table 4. Yearly Observations and Community Sizes

	2017	2018	2019
	Number of Observations		
Min.	1	1	1
Max.	30	216	409
Average	8.1	46.7	80.0
Std. Dev.	9.6	47.6	93.0
Total	130	6915	4878
Community	Probability (Count)		
Few	0.25 (4)	0.46 (68)	0.30 (18)
Few-MUn	0.56 (9)	0.26 (38)	0.33 (20)
Meniscus	0.13 (2)	0.20 (29)	0.28 (17)
RandU	0.06 (1)	0.08 (12)	0.10 (6)

Note : The number of citizen scientists in each community is shown in parentheses.

regression parameters for this analysis are inferred rather than held constant, the unknown error β decreases from 2.4 in 2017 to 1.7 in 2019. The error structure of observations submitted by citizen scientists is evolving as both S4W-Nepal and the participating citizen scientists gain experience, a common trend in citizen science programs (Kosmala et al., 2016).

S4W-Nepal uses various training techniques and feedback methods to increase the scientific literacy of citizen scientists (Davids et al., 2019). Their methods have been effective in reducing the magnitude and frequency of errors committed by the citizen scientists. Perhaps the best evidence for this change is the reduction in meniscus errors committed by citizen scientists in the Meniscus community. From 2018 to 2019, the probability of meniscus errors in the Meniscus community decreased from 19.0 to 8.1%. Similarly, unit errors committed by those in the RandU community decreased from 6.4% in 2018 to 4.2% in 2019. While a trend in reduced meniscus and unit errors over two years is promising, additional analysis after multiple years of collecting citizen scientist observations would provide more conclusive evidence for increased scientific literacy of the participants.

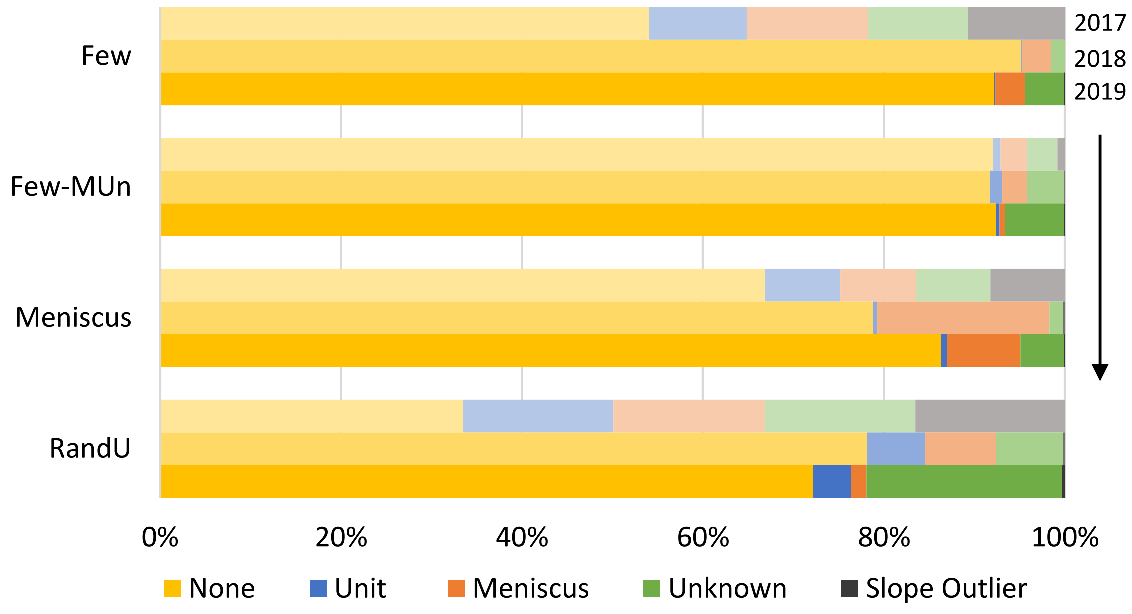

Figure 6. Change in the distribution of errors for each community over time. Note that the 2017 error distributions for the Few, Meniscus, and RandU communities are poorly informed due to the low number of active citizen scientists assigned to those communities.

5.7 Utility and limitations in application

The model proposed here can be implemented by a wide array of citizen science programs. The model is flexible, and thus can be adapted to both qualitative and quantitative citizen science observations. For example, the model could be directly used to assess errors in citizen science water quality measurements or river stage observations. The model could also be adapted to assess the quality of count data submitted by citizen scientists, for example, in the Audubon Society's Christmas Bird Count. Here, the error variable would likely need to be further informed by physiographic features that influence bird habitat and migration. As a qualitative example, the model could be adapted to assess errors in galaxy identification conducted by citizen scientists. Here, the Gaussian mixture of regressions factor would be replaced by a simple discrete distribution wherein the correct galaxy label is assumed to be from a community-specific probability distribution of possible galaxy labels.

While the model has potential for adaptation to a wide variety of citizen science programs, it has limitations. For example, the model is data intensive, because a large dataset is required for training and testing the model. This limits its utility for small-scale or newly developed citizen science programs. In addition, a record of erroneous data is required for training the model, which must be identified and corrected by the citizen science program. This may require a large effort and, depending on the type of data collected, may be difficult to achieve. It could be interesting to investigate to what extent the model can be trained without the availability of error-free ground truth data. For example, Schoups and Nasseri (2020) showed that fusion of multi-source data with unknown noise and bias (in their case, water balance data from remote sensing) is possible in the absence of ground truth data. Lastly, the model design requires that citizen scientists are registered with the program, and that submissions can be linked to registered individuals. This is not the case for all citizen science programs- some do not require registration and some do not track the submission record of their participants. The model can be implemented for quality assessment in many citizen science programs, but the model is not universally useful or without limitations.

6 Summary and conclusions

This study developed a probabilistic model to investigate the type and frequency of errors in citizen science data. The model assigns citizen scientists to a community based on the characteristics of the citizen scientist and their tendency to submit erroneous observations. This helps to target manual corrections of CS data. The model then infers a posterior distribution of the true value of a submitted observation from the value of the observation and the community of the participating citizen scientist. Designed thus, the model can be adapted to a wide array of citizen science datasets.

Analysis of the error structure in citizen scientist rainfall observations revealed that individuals can be characterized by one of four error patterns: not error prone, mostly not error prone, meniscus error prone, and random or various error prone. While the Bayesian inference model developed here used communities to relate citizen scientist characteristics to error tendencies, the magnitude and type of errors committed is the crux of every community assignment. The distribution of characteristics within each community is useful for investigating potential reasons for making errors rather than for identifying individuals who might be particularly error prone.

The Bayesian inference model developed using Infer.NET's software framework uncovered five error types and their probability distribution within each of the four error-based communities. The community assignments are a useful tool for discerning which citizen scientists are more likely to submit erroneous observations that require further review. In addition, community-specific training and feedback messages may be a powerful tool for increasing the quality and frequency of submissions. The Bayesian probabilistic model was often able to predict the true value of a submitted observation, and the model extrapolated useful error probabilities for each observation. These error probabilities, in conjunction with the model's inferred error-specific regression and precision parameters, can be used to calculate a Gaussian mixture distribution that provides more information about the probable true value of submitted observations than Infer.NET's single-mode true value prediction. As citizen science programs expand to include multiple participants submitting observations of a single event, the model's ability to predict the true value for that event will likely increase. However, the model's potential may be limited in regions where the target parameter is highly heterogeneous in space and time.

As a graphical, assumption-based Bayesian inference model, the citizen science error model presented here has immense potential for adaptation to other citizen science programs with diverse data types. The implementation of error-based communities provides a simple, yet effective method for tracking changes in the types and frequency of errors committed by citizen scientists. The communities also provide opportunities for targeted training and feedback to improve citizen science data at the point of collection, rather than at the point of correction. Improving the quality of citizen science data at every step enables increasingly more citizen scientist-supported decision-making and scientific discoveries.

A Prior and Posterior Distributions

The prior distribution for each inferred model variable was a uniform Dirichlet distribution, with the exception of the true value prior. The prior distribution for true value was a Gaussian distribution with a mean of 15 and variance of 2400. The variance for the true value prior was selected is four times the variance of the entire true value dataset (i.e., twice the standard deviation). Note that Equation A.5 is the posterior distribution for the model. The posterior is obtained by writing the joint distribution over latent variables $X = (PChar, PCom, PErr, \vartheta, \varepsilon, \gamma, \alpha_\varepsilon, \beta_\varepsilon, \tau_\varepsilon)$ and observed variables $D = (Z, O)$, followed by conditioning on the observations.

$$PChar_c|\gamma \sim Dirichlet(Uniform), \tag{A.1}$$

$$PCom_s \sim Dirichlet(Uniform), \tag{A.2}$$

$$PErr|\gamma \sim Dirichlet(Uniform), \tag{A.3}$$

$$\vartheta_e \sim \mathcal{N}(15, 2400), \tag{A.4}$$

$$\begin{aligned} p(X|D) \propto Dir(PCom|s)Dir(PErr|\gamma)\prod_{\varepsilon}\mathcal{N}(\mu_\alpha, \sigma_\alpha^2|\varepsilon)\mathcal{N}(\mu_\beta, \sigma_\beta^2|\varepsilon)Gamma(A, B|\varepsilon) \\ \prod_{c=1}^{C} Dir(PChar_c)\prod_{e=1}^{E}\mathcal{N}(\vartheta_e|\mu_e, \sigma_e^2) \\ \prod_{s=1}^{S}\left\{Dis(\gamma_s|PCom)\prod_{c=1}^{C}\left\{Dis(Z_{s,c}|\gamma_s, PChar_c)\right\}\right. \\ \left.\prod_{e=1}^{E}\left\{Dis(\varepsilon_{s,e}|\gamma_s, PErr)\prod_{\varepsilon}\mathcal{N}(O_{s,e}|\alpha_\varepsilon\vartheta_e + \beta_\varepsilon, \tau_\varepsilon)^{\delta(\varepsilon_{s,e}-\varepsilon)}\right\}\right\}, \end{aligned} \tag{A.5}$$

where the Dirac delta function $\delta()$ in the exponent on the last line is used to mathematically represent the mixture of linear regressions (i.e. the gate in Fig. 2), as documented in Minka and Winn (2008).

The prior distributions for the α and β parameters in Eq. 4 were set to a Gaussian distribution parameterized by mean and variance.

$$\alpha_\varepsilon \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2|\varepsilon) \tag{A.6}$$

where μ_α = (1, 0.1, 1.002, 0.9, 7), and σ_α^2 = (0.5, 0.5, 2, 50, 70). And,

$$\beta_\varepsilon \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2|\varepsilon) \tag{A.7}$$

where μ_β = (0, 0.02, 2.3, 4.2, 3), and σ_β^2 = (0.5, 0.5, 0.2, 50, 30). The α and β mean and variance for the first four ε error types were based on the mean and variance of a series of slopes and intercepts from linear regressions fit to subsets of (ϑ, O) pairs corresponding to the four error types identified by Davids et al. (2019). Note that the σ^2 values used are larger than calculated to provide a wider prior distribution. The mean and variance for the remaining error type was selected randomly, since there was no information available regarding this error prior to training the model.

The prior distributions for the τ parameter in Eq. 4 were set to a Gamma distribution parameterized by shape (A) and rate (B).

$$\tau_\varepsilon \sim Gamma(A, B|\varepsilon) \tag{A.8}$$

where $A = (0.25, 0.75, 1.5, 0.5, 15)$, and $B = (0.05, 0.25, 0.05, 0.01, 10)$. The τ shape and rate for the first four ε error types were calculated based a Gamma distribution fit to observations that corresponded to the four error types identified by Davids et al. (2019). The shape and rate for the remaining error type was selected randomly, since there was no information available regarding this error prior to training the model.

Notation

$\boldsymbol{Dir}$ Dirichlet distribution

$\boldsymbol{Dis}$ Discrete distribution

$\boldsymbol{\mathcal{N}}$ Gaussian distribution

$\boldsymbol{C}$ characteristic

$\boldsymbol{S}$ citizen scientist

$\boldsymbol{\varepsilon}$ error type

$\boldsymbol{e}$ event

$\boldsymbol{\gamma}$ Community

$\boldsymbol{O}$ SubmittedObservation

$\boldsymbol{\vartheta}$ TrueValue

Acknowledgments

The dataset analyzed for this study can be accessed in the Supplementary Material published by Davids et al. (2019). This research has been supported by the National Science Foundation, Division of Graduate Education (grant no. DGE-1333468) and the Dutch Research Council. Data collection and quality control was supported by the Swedish International Development Agency (grant no. 2016-05801) and by SmartPhones4Water (S4W). The authors declare that they have no conflict of interest. The authors would like to thank S4W's Saujan Maka for instrumental guidance.

References

Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., ... Frusher, S. (2014, May). Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, *173*, 144–154. Retrieved 2020-05-02, from https://linkinghub.elsevier.com/retrieve/pii/S0006320713002693 doi: 10.1016/j.biocon.2013.07.037

Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., & Shirk, J. (2009, December). Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. *BioScience*, *59*(11), 977–984. Retrieved 2020-05-02, from https://academic.oup.com/bioscience/article-lookup/doi/10.1525/bio.2009.59.11.9 doi: 10.1525/bio.2009.59.11.9

Bonter, D. N., & Cooper, C. B. (2012, August). Data validation in citizen science: a case study from Project FeederWatch. *Frontiers in Ecology and the Environment*, *10*(6), 305–307. Retrieved 2020-05-03, from http://doi.wiley.com/10.1890/110273 doi: 10.1890/110273

Brunsdon, C., & Comber, L. (2012). Assessing the changing flowering date of the common lilac in north america: a random coefficient model approach. *Geoinformatica*, *16*(4), 675–690.

Budde, M., Schankin, A., Hoffmann, J., Danz, M., Riedel, T., & Beigl, M. (2017, September). Participatory Sensing or Participatory Nonsense?: Mitigating the Effect of Human Error on Data Quality in Citizen Science. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(3), 1–23. Retrieved 2020-05-03, from https://dl.acm.org/doi/10.1145/3131900 doi: 10.1145/3131900

Butt, N., Slade, E., Thompson, J., Malhi, Y., & Riutta, T. (2013). Quantifying the sampling error in tree census measurements by volunteers and its effect on carbon stock estimates. *Ecological Applications*, *23*(4), 936–943.

Cox, T., Philippoff, J., Baumgartner, E., & Smith, C. (2012). Expert variability provides perspective on the strengths and weaknesses of citizen-driven intertidal monitoring program. *Ecological Applications*, *22*(4), 1201–1212.

Crall, A. W., Newman, G. J., Stohlgren, T. J., Holfelder, K. A., Graham, J., & Waller, D. M. (2011, December). Assessing citizen science data quality: an invasive species case study: Assessing citizen science data quality. *Conservation Letters*, *4*(6), 433–442. Retrieved 2020-05-02, from http://doi.wiley.com/10.1111/j.1755-263X.2011.00196.x doi: 10.1111/j.1755-263X.2011.00196.x

Davids, J. C., Devkota, N., Pandey, A., Prajapati, R., Ertis, B. A., Rutten, M. M., … van de Giesen, N. (2019, March). Soda Bottle Science—Citizen Science Monsoon Precipitation Monitoring in Nepal. *Frontiers in Earth Science*, *7*, 46. Retrieved 2020-04-23, from https://www.frontiersin.org/article/10.3389/feart.2019.00046/full doi: 10.3389/feart.2019.00046

Delaney, D. G., Sperling, C. D., Adams, C. S., & Leung, B. (2008, January). Marine invasive species: validation of citizen science and implications for national monitoring networks. *Biological Invasions*, *10*(1), 117–128. Retrieved 2020-05-03, from http://link.springer.com/10.1007/s10530-007-9114-0 doi: 10.1007/s10530-007-9114-0

de Solla, S. R., Shirose, L. J., Fernie, K. J., Barrett, G. C., Brousseau, C. S., & Bishop, C. A. (2005). Effect of sampling effort and species detectability on volunteer based anuran monitoring programs. *Biological Conservation*, *121*(4), 585–594.

Fink, D., Hochachka, W. M., Zuckerberg, B., Winkler, D. W., Shaby, B., Munson, M. A., … Kelling, S. (2010). Spatiotemporal exploratory models for broad-scale survey data. *Ecological Applications*, *20*(8), 2131–2147.

Habib, E., Krajewski, W. F., & Ciach, G. J. (2001). Estimation of rainfall interstation correlation. *Journal of Hydrometeorology*, *2*, 621–629. doi: 10.1175/1525-7541(2001)002⟨0621:EORIC⟩2.0.CO;2

Hunter, J., Alabri, A., & van Ingen, C. (2013, February). Assessing the quality and trustworthiness of citizen science data. *Concurrency and Computation: Practice and Experience*, *25*(4), 454–466. Retrieved 2020-05-03, from http://doi.wiley.com/10.1002/cpe.2923 doi: 10.1002/cpe.2923

Kidd, C., Becker, A., Huffman, G. J., Muller, C. L., Joe, P., Skofronick-Jackson, G.,

& Kirschbaum, D. B. (2017, January). So, How Much of the Earth's Surface *Is* Covered by Rain Gauges? *Bulletin of the American Meteorological Society*, *98*(1), 69–78. Retrieved 2020-05-02, from `http://journals.ametsoc.org/doi/10.1175/BAMS-D-14-00283.1` doi: 10.1175/BAMS-D-14-00283.1

Kosmala, M., Wiggins, A., Swanson, A., & Simmons, B. (2016, December). Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, *14*(10), 551–560. Retrieved 2020-05-03, from `http://doi.wiley.com/10.1002/fee.1436` doi: 10.1002/fee.1436

Lukyanenko, R., Wiggins, A., & Rosser, H. K. (2019). Citizen science: An information quality research frontier. *Information Systems Frontiers*, 1–23.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms.* Cambridge: Cambridge University Press.

Miller, D. A., Nichols, J. D., McClintock, B. T., Grant, E. H. C., Bailey, L. L., & Weir, L. A. (2011). Improving occupancy estimation when two types of observational error occur: Non-detection and species misidentification. *Ecology*, *92*(7), 1422–1428.

Minka, T. (2000). *Bayesian linear regression* (Tech. Rep.). Citeseer.

Minka, T. (2005). *Divergence measures and message passing* (Technical Report No. TR-2005-173). Microsoft Research.

Minka, T. (2013). Expectation propagation for approximate bayesian inference. *arXiv preprint arXiv:1301.2294*.

Minka, T., & Winn, J. (2008). Gates. *Advances in Neural Information Processing Systems 21*, 1073–1080.

Minka, T., Winn, J., Guiver, J., Zaykov, Y., Fabian, D., & Bronskill, J. (2018). *Infer.NET 0.3.* Microsoft Research Cambridge. Retrieved from `http://dotnet.github.io/infer`

Nayava, J. L. (1974, December). Heavy monsoon rainfall in Nepal. *Weather*, *29*(12), 443–450. Retrieved 2020-04-23, from `http://doi.wiley.com/10.1002/j.1477-8696.1974.tb03299.x` doi: 10.1002/j.1477-8696.1974.tb03299.x

Newman, G., Wiggins, A., Crall, A., Graham, E., Newman, S., & Crowston, K. (2012, August). The future of citizen science: emerging technologies and shifting paradigms. *Frontiers in Ecology and the Environment*, *10*(6), 298–304. Retrieved 2020-05-02, from `http://doi.wiley.com/10.1890/110294` doi:

10.1890/110294

Nishihara, R., Minka, T., & Tarlow, D. (2013). Detecting parameter symmetries in probabilistic models. *arXiv preprint, arXiv:1312.5386*.

Riesch, H., & Potter, C. (2014, January). Citizen science as seen by scientists: Methodological, epistemological and ethical dimensions. *Public Understanding of Science*, *23*(1), 107–120. Retrieved 2020-05-02, from `http://journals.sagepub.com/doi/10.1177/0963662513497324` doi: 10.1177/0963662513497324

Schoups, G., & Nasseri, M. (2020). Gracefully closing the water balance: a data-driven probabilistic approach applied to river basins in iran.

Sheppard, S. A., & Terveen, L. (2011). Quality is a verb: the operationalization of data quality in a citizen science community. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration - WikiSym '11* (p. 29). Mountain View, California: ACM Press. Retrieved 2020-05-03, from `http://dl.acm.org/citation.cfm?doid=2038558.2038565` doi: 10.1145/2038558.2038565

Sunde, P., & Jessen, L. (2013). It counts who counts: an experimental evaluation of the importance of observer effects on spotlight count estimates. *European Journal of Wildlife Research*, *59*(5), 645–653.

Teague, K. A., & Gallicchio, N. (2017). *The evolution of meteorology: a look into the past, present, and future of weather forecasting.* Hoboken, NJ: John Wiley & Sons, Inc.

Thapa, B. R., Ishidaira, H., Pandey, V. P., & Shakya, N. M. (2017, February). A multi-model approach for analyzing water balance dynamics in Kathmandu Valley, Nepal. *Journal of Hydrology: Regional Studies*, *9*, 149–162. Retrieved 2020-04-23, from `https://linkinghub.elsevier.com/retrieve/pii/S2214581816303342` doi: 10.1016/j.ejrh.2016.12.080

USGS. (n.d.). *DYFI Scientific Background.* Retrieved 2020-05-05, from `https://earthquake.usgs.gov/data/dyfi/background.php`

Venanzi, M., Guiver, J., Kazai, G., Kohli, P., & Shokouhi, M. (2014). Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of the 23rd international conference on World wide web - WWW '14* (pp. 155–164). Seoul, Korea: ACM Press. Retrieved 2020-05-03, from `http://dl.acm.org/`

citation.cfm?doid=2566486.2567989 doi: 10.1145/2566486.2567989

Vibhāga, N. K. T. (2012). *National Population and Housing Census 2011: National report* (Vol. 1). Government of Nepal, National Planning Commission Secretariat, Central . . . .

Wiggins, A., Newman, G., Stevenson, R. D., & Crowston, K. (2011, December). Mechanisms for Data Quality and Validation in Citizen Science. In *2011 IEEE Seventh International Conference on e-Science Workshops* (pp. 14–19). Stockholm, Sweden: IEEE. Retrieved 2020-05-03, from http://ieeexplore.ieee.org/document/6130725/ doi: 10.1109/eScienceW.2011.27

Winn, J., Bishop, C., Diethe, T., Guiver, J., & Zaykov, Y. (2020). *Model-based machine learning* (early access ed.). online: Microsoft Research. Retrieved from www.mbmlbook.com