

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

4 **J.A. Eisma¹, G. Schoups², J.C. Davids³, N. van de Giesen²**

5 ¹Department of Civil Engineering, University of Texas at Arlington, Arlington, Texas, USA

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

7 ³Department of Civil Engineering and College of Agriculture, California State University, Chico,

8 California, USA

9 **Key Points:**

- 10 • A Gaussian mixture of regressions explains the likelihood of citizen scientists sub-
11 mitting erroneous observations
- 12 • Citizen scientists are sorted into communities based on characteristics and the type
13 and frequency of errors in the data that they submit
- 14 • The distribution of errors in the data from citizen scientists evolves as they gain
15 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

48 scientific discoveries advanced by citizen scientists (Hunter, Alabri, & van Ingen, 2013;
49 Riesch & Potter, 2014; Sheppard & Terveen, 2011). Others, however, have embraced cit-
50 izen science as an effective means for increasing the spatial and temporal resolution of
51 scientific data. Successful citizen science programs investigate the type and frequency
52 of errors in the data collected by program participants and develop training initiatives
53 designed to reduce errors (Bird et al., 2014; Crall et al., 2011; Davids et al., 2019).

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

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

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

80 Most error analyses of citizen science data focus on identifying and removing outliers from
81 a dataset. Trained filters flag outliers by identifying observations that do not fit within
82 the expected range of values or classes, such as species range or allowable count (Bon-
83 ter & Cooper, 2012; Wiggins, Newman, Stevenson, & Crowston, 2011). Some citizen sci-
84 ence programs develop eligibility or trust rating procedures to identify users that are likely
85 to submit correct observations (Delaney, Sperling, Adams, & Leung, 2008; Hunter et al.,
86 2013). Ratings schemes that consider demographic and experience-related characteris-
87 tics have potential for describing the variability in citizen science data reliability (Kos-
88 mala, Wiggins, Swanson, & Simmons, 2016). However, some individual citizen scientists
89 do not submit enough observations to be accurately assigned a rating. To overcome such
90 limitations, Venanzi, Guiver, Kazai, Kohli, and Shokouhi (2014) based their error anal-
91 ysis on four communities of citizen scientists, each with a distinctive pattern of errors.

92 Machine learning algorithms and hierarchical, generalized linear, and mixed-effects
93 models have also been employed by a variety of citizen science programs to study errors
94 in citizen science data (Bird et al., 2014; Venanzi et al., 2014). Generalized linear mod-
95 els have largely been used to study whether and how characteristics of citizen scientists
96 affect the accuracy of their observations (Butt, Slade, Thompson, Malhi, & Riutta, 2013;
97 Crall et al., 2011; Delaney et al., 2008). Mixed-effects models add a random-effects fac-
98 tor to generalized linear models, permitting the study of errors in relation to an unin-
99 tended grouping effect, such as spatial clustering (Bird et al., 2014; Brunsdon & Comber,
100 2012). Alternatively, hierarchical models have been leveraged to study how citizen sci-
101 entist errors relate to effort and site-level effects (de Solla et al., 2005; Fink et al., 2010;
102 Miller et al., 2011). Lastly, machine learning has been used to study errors in qualita-
103 tive citizen science data, such as species identification and labeling tweets (Cox, Philip-
104 poff, Baumgartner, & Smith, 2012; Lukyanenko, Wiggins, & Rosser, 2019; Venanzi et
105 al., 2014). Machine learning has not yet been employed to identify erroneous citizen sci-
106 ence observations for quantitative data. In addition, most machine learning citizen sci-
107 ence research has focused on datasets that are relatively static or slow-moving in the fields
108 of biology and conservation (Lukyanenko et al., 2019). To our knowledge, the study pre-
109 sented here is the first attempt to leverage machine learning to assess errors in quanti-
110 tative citizen science data with high spatiotemporal variability. Despite the wide range
111 of existing research on citizen science errors, flexible methods for analyzing errors in quan-
112 titative citizen science data remains largely unexplored.

113 The objective of this study is to improve quality control of quantitative citizen sci-
114 ence data by developing a Bayesian inference model that discovers, explains, and pos-
115 sibly corrects the errors in observations submitted by citizen scientists. The following
116 research questions will be explored:

- 117 1. How can the type and magnitude of citizen science data errors be automatically
118 identified from citizen science data and corresponding ground truth?
- 119 2. Given a calibrated citizen scientist, to what extent can errors be detected and cor-
120 rected without ground truth?
- 121 3. To what extent do citizen scientist characteristics help in identifying and screen-
122 ing errors?

123 A probabilistic graphical model was developed to address these questions based on as-
124 sumptions about the probabilistic relationships between citizen scientists, their charac-
125 teristics, and the magnitude of their errors. The probabilistic graphical model includes
126 a regression clustering sub-model relating true and observed values and includes an un-
127 known number of linear regressions. The model also includes a probabilistic sub-model
128 relating citizen scientist characteristics to error types. Applied to the S4W-Nepal pro-
129 gram, the model identifies unique error types within the S4W-Nepal citizen scientist rain-
130 fall observations, and groups citizen scientists into communities based on their charac-
131 teristics and error profile. Each community is characterized by a distinct distribution of
132 error types which indicates the likelihood that a submitted observation should be reviewed
133 further. After testing and training, the model was applied to investigate three practi-
134 cal issues: the error evolution of citizen scientist data over time (research question 1),
135 multiple observations of a single rainfall event (research question 2), and observations
136 submitted by citizen scientists with unknown characteristics (research question 3).

137 2 Study Area

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

144 The majority of citizen scientists participating in S4W-Nepal’s rainfall data collection
 145 efforts reside in the Kathmandu Valley, home to about 10% of Nepal’s population (Vibhāga,
 146 2012). While the average annual precipitation is approximately 1,500 mm in the city of
 147 Kathmandu and 1,800 mm in the surrounding hills, it is highly variable and unpredictable
 148 (Thapa, Ishidaira, Pandey, & Shakya, 2017).

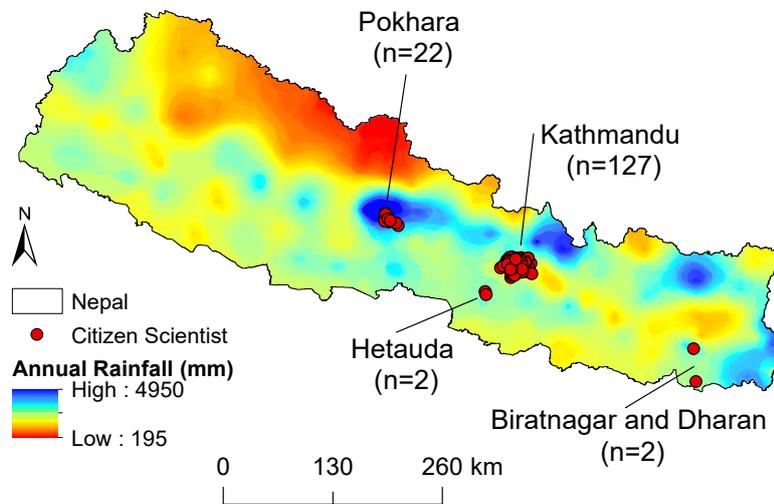


Figure 1. Locations of citizen scientists for which characteristics are known with the number of citizen scientists at specified locations shown in parentheses. Average annual rainfall grid created by USAID Nepal from observed data at 200 weather stations from 1980-2000.

149 3 Data

150 S4W-Nepal recruits citizen scientists to participate in a crowdsourced rainfall ob-
 151 servation program in Nepal. S4W-Nepal collects the submitted observations via the Open
 152 Data Kit application for smart phones. Submitted observations include geo-location data,
 153 time of measurement, citizen scientist-reported depth of rainfall in millimeters, and a pho-
 154 tograph of the rain gauge. The program is ongoing and has collected over 24,500 obser-
 155 vations from over 265 citizen scientists since 2016.

156 **3.1 Rain gauges**

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

165 **3.2 Citizen characteristics**

166 During the recruitment process, S4W-Nepal recorded characteristic data for 153
167 citizen scientists. Characteristics recorded were: motivation (paid/volunteer), recruit-
168 ment method (personal connection, random site visit, social media, outreach), age (≤ 18 ,
169 19-25, > 25), education ($<$ Bachelors, Bachelors, $>$ Bachelors), place of residence (urban,
170 semi-urban, rural), occupation (agriculture, student, other), and gender (male, female).
171 Citizen scientist characteristics will be used here to relate individual citizen scientists
172 with the likelihood of errors in the data they submit.

173 **4 Methods**

174 **4.1 Identification of erroneous observations**

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

184 **4.2 Model development**

185 ***4.2.1 Assumptions and model structure***

186 A Bayesian probabilistic graphical model was developed based on a number of as-
187 sumptions about the data being modeled. These assumptions were used to inform the
188 relationships between the variables and ensure the model accurately represents the mod-
189 eler’s understanding of the physical processes that underlie the data (Winn, Bishop, Di-
190 ethe, Guiver, & Zaykov, 2020). The following assumptions informed the development of
191 the citizen science errors inference model:

- 192 1. Each citizen scientist belongs to a single community.
- 193 2. A citizen scientist’s community is defined by their collective demographic and experience-
194 related characteristics and the type and frequency of errors they have made in prior
195 submissions.
- 196 3. Each citizen scientist in a particular community always submits an observation
197 with a community-specific error type distribution.
- 198 4. Each citizen scientist observation relates to an underlying true value with a sys-
199 tematic bias and random noise level that depends on the error type of the obser-
200 vation.

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

214 servation may be necessary at this stage of model development for application to other
 215 citizen science programs.

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

$$224 \quad \gamma_s \sim Dis(\mathbf{PCom}|s), \quad (1)$$

225 We use a lowercase subscript to denote a random variable index (e.g. γ_s indicates
 226 there is a community variable for each citizen scientist S). Greek letters represent la-
 227 tent (inferred) variables, and Latin letters to represent observable variables. The value
 228 $Z_{c,s}$ of citizen characteristic c for citizen scientist s is assumed to be from a discrete dis-
 229 tribution with probability vector **PChar** that depends on the characteristic c under con-
 230 sideration and the community γ_s the citizen scientist belongs to:

$$231 \quad Z_{c,s} \sim Dis(PChar_c|\gamma_s), \quad (2)$$

232 Equation 2 quantifies the probabilistic relationship between each citizen charac-
 233 teristic and each assigned community in the form of a conditional probability table. Sim-
 234 ilarly, Equation 3, below, describes the conditional probability table for each error type
 235 and community. The error type $\varepsilon_{s,e}$ of event e observed by citizen scientist s is assumed
 236 to be from a discrete distribution with probability vector **PErr** that depends on the com-
 237 munity γ_s that the citizen scientist belongs to:

$$238 \quad \varepsilon_{s,e} \sim Dis(PErr|\gamma_s), \quad (3)$$

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

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

$$244 \quad O_{s,e} \sim \mathcal{N}(\alpha_{\varepsilon_{s,e}} \vartheta_e + \beta_{\varepsilon_{s,e}}, \tau_{\varepsilon_{s,e}}), \quad (4)$$

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

256 *4.2.2 Model implementation*

257 We implemented the probabilistic model using Microsoft Research’s open source
 258 Infer.NET software framework (Minka et al., 2018). The Infer.NET framework provides
 259 adaptable tools to develop and run Bayesian inference for probabilistic graphical mod-
 260 els. The modeler must define the variables, the dependencies between variables, and pro-
 261 vide prior distributions for the variables that will be inferred. For implementation in In-
 262 fer.NET, Equations 1-4 are translated into a factor graph as shown in Figure 2. The fac-
 263 tor graph completely describes the joint posterior probability of the model (see Equa-
 264 tion A.5). The factor graph includes observable and latent (inferred) variables, factor
 265 nodes, edges (arrows), plates, and gates. Variables are depicted by shaded or unfilled el-
 266 lipses. A shaded variable is an observable value; an unfilled variable is a latent value. Fac-
 267 tor nodes are the small black boxes connected to variables, describing the relation be-
 268 tween variables connected to the factor. Edges (directional arrows) connect factor nodes
 269 to variables (Winn et al., 2020).

270 *Plates.* Plates are the large boxes outlined in gray surrounding portions of the fac-
 271 tor graph. Plates are a simplified way to express repeated structures. The number of times

272 a structure will be repeated is based on the index variable shown in the bottom right cor-
 273 ner of the plate (Winn et al., 2020). For example, in Figure 2, the structure within the
 274 characteristics plate is repeated nine times, because the model considers nine different
 275 CS characteristics: motivation, recruitment, age, education, place of residence, occupa-
 276 tion, gender, performance, and experience.

277 *Gates.* Gates are indicated by a dashed box, as seen around the Regression factor
 278 node in Figure 2. Gates essentially act as a switch, turning on and off depending on the
 279 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.

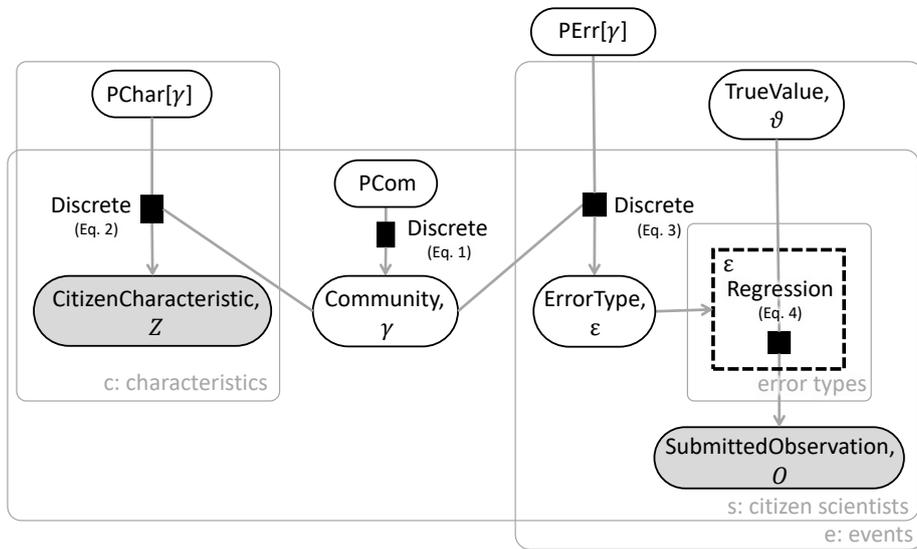


Figure 2. The citizen science error model depicted as a factor graph. A factor node represents a probabilistic relation between variables in the model and is shown by a black square. A variable is shown in an oval, with shading identifying observable variables. Arrows depict the output variable of each 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).

280

281 Infer.NET generates a computationally efficient code for the inference algorithm
 282 using one of three available inference algorithms: expectation propagation, variational
 283 message passing, and Gibbs sampling. The model developed here employs the expecta-
 284 tion propagation algorithm, because it is time efficient but reasonably accurate (Minka,
 285 2013). Expectation propagation is a deterministic approximate inference algorithm for

286 computing the marginal posterior distribution of each variable in the model (Minka, 2013).
287 Each posterior distribution is assumed to take a specific parametric form in an exponen-
288 tial family (e.g. Gaussian, Gamma, discrete). The algorithm then aims to find param-
289 eter values for each parametric posterior that result in a good approximation of the ex-
290 act posterior in terms of moment matching. For example, for a Gaussian approximation,
291 expectation propagation will find a Gaussian whose mean and variance approximate those
292 of the actual posterior. This is done using an iterative approach that starts from an ini-
293 tial guess for the approximate posteriors, and iteratively refines each posterior in turn
294 via moment matching. Since all individual posterior updates depend on each other, the
295 algorithm is iterated until all updates and posteriors stabilize (here, in <5 iterations).
296 The final posteriors are not necessarily unique and may depend on how the algorithm
297 was initialized. Here, we adopt a random initialization strategy for mixture models as
298 used in Nishihara, Minka, and Tarlow (2013) and Minka et al. (2018) and evaluate non-
299 uniqueness in the inferred posteriors using multiple runs with different random initial-
300 ization.

301 *4.2.3 Community and error selection*

302 To select the appropriate number of communities to capture the differences among
303 the citizen scientists, model evidence was used. Model evidence indicates which model
304 best explains the data relative to the model's complexity (MacKay, 2003, p. 343-386).
305 While the model evidence is notoriously hard to compute, expectation propagation pro-
306 vides a convenient estimate as a by-product of its posterior approximations. Model ev-
307 idence calculation in Infer.NET is achieved by inferring posterior component weights of
308 a mixture consisting of two components, i.e. the entire model and the empty model (Minka,
309 2000).

310 Too many communities may lead to overfitting, whereas too few communities may
311 lead to underfitting. The model evidence automatically makes this trade-off and iden-
312 tifies the optimal number of communities. Model evidence was computed for models with
313 one to ten communities. The number of communities that resulted in the largest model
314 evidence was selected as the correct number of communities for the model and data. Sim-
315 ilarly, model evidence was used to determine how many error types were present in the
316 data. Model evidence was computed for one to twelve error types while using the op-
317 timal number of communities. The number of error types that resulted in the largest model

318 evidence was selected as the number of error types for the model and data. After select-
319 ing the number of error types, model evidence was again checked to verify that the op-
320 timal number of communities remained constant. Selecting the error types via model ev-
321 idence may identify more error types than expected, but the Bayesian model accounts
322 for all possibilities and selects the one that most accurately represents the data.

323 *4.2.4 Training and testing the model*

324 The inference model was trained and tested to ensure model performance was con-
325 sistent across different groups of data. During training and testing, the following char-
326 acteristics were known for each citizen scientist: motivation, recruitment, age, education,
327 place of residence, occupation, gender, performance, and experience. The first seven char-
328 acteristics were recorded by S4W-Nepal (as explained in Section 3). The last two char-
329 acteristics, performance and experience, were defined based on the observations submit-
330 ted by each citizen scientist. Performance is simply the percentage of observations sub-
331 mitted by a citizen scientist that did not require correction. A performance of 90% in-
332 dicates that 90% of that citizen scientist’s submitted observations matched the true value
333 shown in the associated photograph. Experience is a count of how many observations
334 a citizen scientist submitted through the 2018 monsoon season. Performance and expe-
335 rience rates were split into three levels based on natural breakpoints in their respective
336 histograms.

337 *Splitting the data.* Rainfall observations submitted by citizen scientists with known
338 characteristics from 2016 to 2018 were randomly split into a training data set and a test-
339 ing data set. The training set consisted of 92% of available observations, representing
340 6,091 observations submitted by 152 citizen scientists. The citizen scientists in the train-
341 ing set submitted anywhere from 1 to 159 observations, with the average number of sub-
342 missions being 43.5. The testing set consisted of the remaining 8% of available obser-
343 vations, representing 527 observations from 109 citizen scientists. The citizen scientists
344 in the testing set submitted anywhere from 1 to 159 observations, with the average num-
345 ber of submissions being 57.4. All citizen scientists in the testing set were also in the train-
346 ing set. Note that individual observations in each group were unique.

347 *Training the model.* Before training the model, prior distributions were set for the
348 variables that were inferred. Uniform prior distributions were set for the citizen char-

349 characteristics (see Equation A.1), community (see Equation A.2), and error (see Equation A.3).
350 The prior distribution for the true value parameter was a Gaussian distribution with a
351 mean equal to the average value of all submitted observations (15) and the four times
352 the variance of the entire dataset (2400; see Equation A.4). A true value prior variance
353 of 2400 was chosen to reduce small event bias and accommodate inference of large rain-
354 fall observations. The prior distributions for the Gaussian mixture parameters (α , β , and
355 τ) were assigned based on the magnitude of unit, meniscus, and unknown errors clas-
356 sified by Davids et al. (2019).

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

363 *Testing the model.* To test the model, prior distributions for latent variables were
364 set to the associated posterior distribution calculated during training. The character-
365 istics for each citizen scientist and the values of the submitted observations were set. The
366 model inferred the community for each citizen scientist, the probable error type for each
367 observation, and provided a posterior distribution for the true value of the submitted ob-
368 servation. The performance of the model was assessed based on the whether the inferred
369 posterior distribution for true value (ϑ) covered the true value identified in the accom-
370 panying photograph submitted by the citizen scientist and whether the mode of the true
371 value posterior matched the actual true value.

372 A synthetic rainfall event was created to explore how many observations of a sin-
373 gular event are needed to produce a reliable estimate of the event's true value. A synthetic
374 observation of the event was created by first assigning an error type to each citizen sci-
375 entist based on the distribution of errors for their respective error communities (see Ta-
376 ble 2). Then, the value of the synthetic observation was calculated using Equation 4, the
377 α , β , and τ values from Table 1 with a true value of 15 mm. Multiple synthetic events
378 were created with two to three observations of the same event with one to two erroneous
379 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.

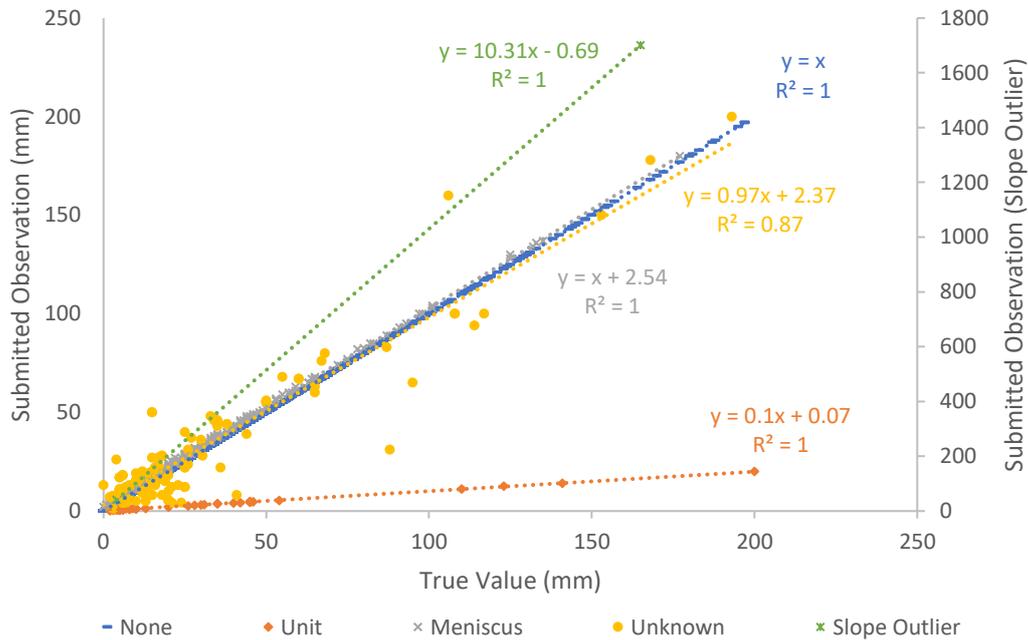
5.2 Error analysis

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.

410 **5.2.1 Error distribution within communities**

411 The distribution of errors committed by citizen scientists varied depending on the
 412 assigned community, as seen in Table 2. Each community was named based on its re-
 413 spective error distribution: Few, Few-MUn, Mensicus, and Random Unknown (RandU).
 414 The Few community makes very few errors—only 2% of submitted observations are er-
 415 roneous. Of the erroneous submissions, members in the Few community are most likely
 416 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.

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

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

438 The distribution of errors within each community is a useful tool not only for se-
439 lecting which submitted observations might require verification, but also for identifying
440 opportunities to improve or maintain the overall accuracy of submitted observations. Cit-
441 izen science project organizers can use targeted training to help specific communities im-
442 prove their performance (Budde et al., 2017; Sheppard & Terveen, 2011). For example,
443 S4W-Nepal could occasionally send feedback messages to the meniscus community mem-
444 bers reminding them to read the rainfall depth from the bottom of the meniscus. As an-
445 other example, members in the Few community might positively respond to general feed-
446 back messages acknowledging their strong record of accurate observations and choose
447 to remain engaged with the program. Knowing the error structure of observations sub-
448 mitted by different communities may help improve the overall effectiveness of citizen sci-
449 ence programs.

450 5.3 Community composition

451 The model grouped citizen scientists into four distinct communities with a unique
452 combination of characteristics and probability of making errors. The Few community is
453 the largest with 47% of citizen scientists in the training group assigned to this commu-
454 nity (see Table 2). The RandU community is the smallest with only 7% of citizen sci-
455 entists classified into this group. The remaining citizen scientists are grouped into the
456 Few-Un (19%) and Meniscus (16%) communities. Overall, only 24% of participating cit-
457 izen scientists are likely to make errors in more than 8.3% of their submitted observa-
458 tions.

459 The probability that a citizen scientist will belong to a specific community depends,
460 in part, on the unique characteristics of that citizen scientist. Figure 4 provides the pos-
461 terior probability that a citizen scientist with a particular characteristic would belong
462 to each community, offering insight into the characteristic composition of each commu-
463 nity. Singular characteristics may have a large impact on the tendency of a citizen sci-
464 entist to make errors, and therefore to be assigned to a specific community. However,
465 it is also true that any combination of characteristics could contribute to the probabil-
466 ity of a citizen scientist being assigned to a community. In some cases, citizen scientists
467 are likely to possess a similar combination of characteristics, which surfaces in the com-
468 munity distributions. For example, Figure 4 indicates that citizen scientists recruited dur-
469 ing a random visit, older than 25 years of age, holding less than a bachelor’s degree, and

470 with an “other” occupation make up 20% of all citizen scientists in the project and have
 471 a similar community distribution. While community assignment trends for singular char-
 472 acteristics can be enlightening, the impact of multiple citizen scientists with a similar
 473 combination of characteristics must be acknowledged.

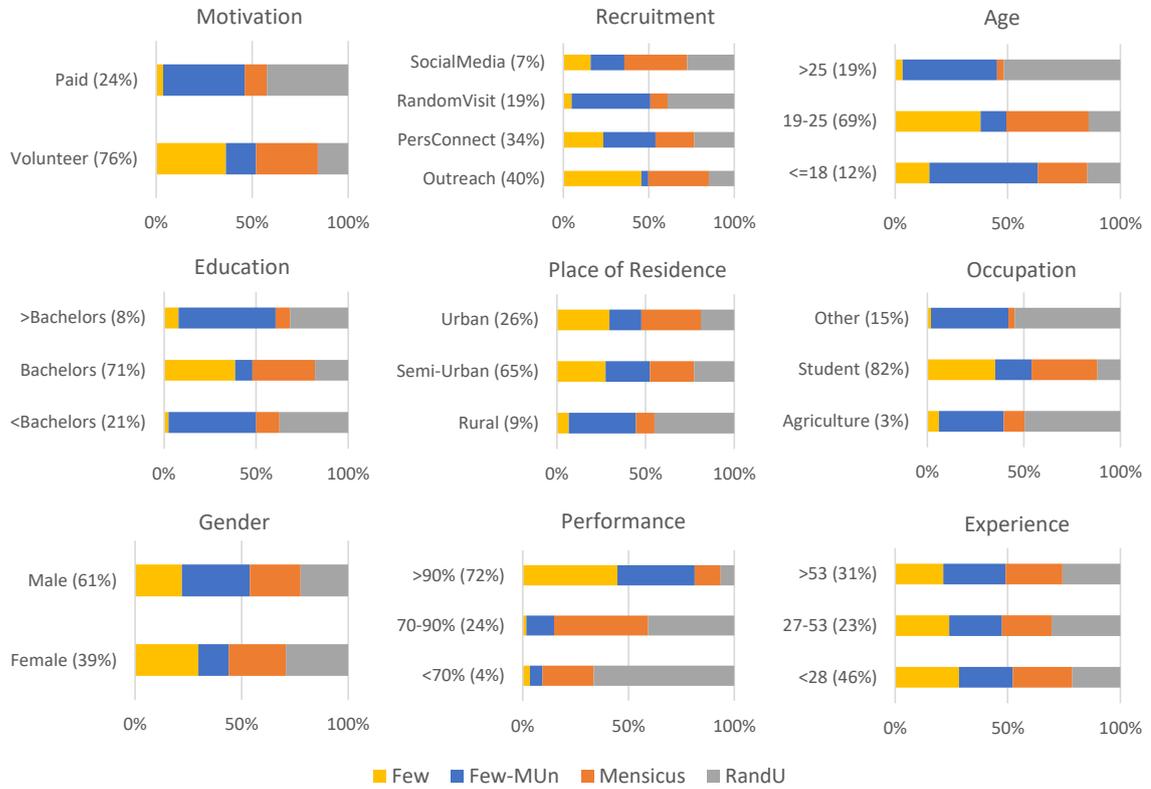


Figure 4. Community composition for each characteristic. The percentage of participating citizen scientists with the associated characteristic is shown in parentheses.

474 **5.4 Sensitivity of α , β , and τ Priors and algorithm initialization**

475 In the model application examined here, Davids et al. (2019) provided prior infor-
 476 mation on the types of errors in the data, but such information will not always be avail-
 477 able. Prior information on the types of errors in the data is useful but not necessary to
 478 identify some of the errors made by participating citizen scientists. When prior error in-
 479 formation is known, the model reliably infers the same five errors, even when the uncer-
 480 tainty of this information is high (i.e. high variance assigned to the Gaussian prior dis-
 481 tributions). When no prior information is known about the potential types of errors present

482 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 er-
 483 ror type and splits the meniscus error into two error types—a 2-mm meniscus error and
 484 a 3.8-mm meniscus error. The two remaining error types identified are variations on the
 485 unknown error with relatively low R^2 values, 0.79 and 0.09 compared with 1.0 for the
 486 none and meniscus errors. The model may fail to identify the unit error type, because
 487 it occurs in only 0.7% of submitted observations. Multiple local optima exist for the er-
 488 ror types, and the model may fail to identify all unique errors if no prior information on
 489 the errors is known. Regardless of whether error information is known previously, model
 490 evidence indicated that four communities and five error types best capture the variance
 491 in the data. The model may require many more iterations (possibly up to 100) to con-
 492 verge when the priors are vague.

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

500 **5.5 Inferring the true value of a submitted observation**

501 In addition to providing insight into the error structure of the submitted observa-
 502 tions and the relationship between citizen scientist characteristics and error tendencies,
 503 the model provides information about the true value of submitted observations. Test-
 504 ing the model reveals that the model can infer a previously unknown true value based
 505 on the value of the submitted observation and the characteristics of the citizen scientist.
 506 The inferred true value differs from the actual true value by a median percent error of
 507 0.9%. The standard deviation of percent error is, however, 98.8%. With a wide true value
 508 prior distribution (here, 24,000; see Eq. A.4), the model has a tendency to over-predict
 509 unit errors for a small number of observations submitted with a value of 6 mm or lower
 510 which causes the large standard deviation (see Figure 5a). In most cases, the actual true
 511 value of the submitted observation falls within the range of the posterior distribution in-
 512 ferred for the true value variable as seen in Figures 5b,c. However, as Figures 5b,c show,

513 the mode of the posterior distribution is not always a good estimate of the actual true
 514 value.

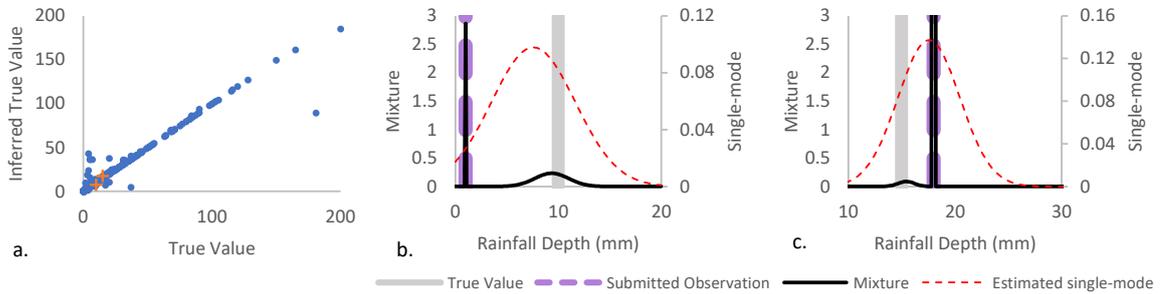


Figure 5. a. The inferred true value is usually a good estimate of the true value of the submitted observation. In some erroneous submissions, the mode of the estimated single-mode posterior is not equal to the true value, however an exact Gaussian mixture of the true value posterior distributions has 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.).

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

524 While the predicted single-mode true value posterior distribution does not always
 525 estimate the actual true value of an erroneous submission well, the exact Gaussian mix-
 526 ture posterior often exhibits a local peak at the actual true value (see Figures 5b,c). The
 527 mode of the Gaussian mixture posterior usually presents at the value of the submitted
 528 observation because of the high precision associated with the none error type (see Ta-
 529 ble 1). Only 8.7% of submitted observations have greater than a 20% probability of be-
 530 ing 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.	Error Types	Inferred		No.	Error Types	Inferred	
		True Value	Variance			True Value	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

531 distribution may be examined in conjunction with the Gaussian mixture posterior to pro-
532 vide additional information on the probability of each error type. For example, despite
533 the mode of the Gaussian mixture posterior being located at the value of the submit-
534 ted observation in Figure 5b, the probability of a none error type is only 0.23, and the
535 unit error probability is 0.73. The Gaussian mixture posterior and the error type pos-
536 terior distributions may provide a more accurate representation of the true value of a
537 submitted observation than the approximated single-mode Gaussian posterior distribu-
538 tion.

539 *5.5.1 Multiple observations of a single event*

540 If only a single observation of a rainfall event is available, the predicted error type
541 is based on the error types observed during model training. However, analyzing multi-
542 ple observations of a single rainfall event should improve the accuracy of the inferred er-
543 ror type and true value of rainfall.

544 For each of the simulations described below, the model was not given any infor-
545 mation about the error types associated with the submitted observations. The model
546 inferred the true value solely based on what it learned during model training. When only
547 one error was made out of two observations submitted, the model predicted the true value
548 every time except for instances of a slope outlier error (see Table 3 column 1). In such
549 cases, the ability of the model to correctly infer the event true value was related to the
550 error communities of the citizen scientists. Through 12 trials (not shown) with differ-
551 ent algorithm initialization and combinations of citizen scientists from the Few-MUn and
552 Meniscus communities, the model correctly inferred the true value only twice. However,
553 the model was able to infer the true value if one submitted observation had a slope out-
554 lier for other combinations of citizen scientist communities (see Table 3 column 1). If one
555 slope outlier observation was paired with two or more correct observations, the model
556 consistently failed to infer the correct true value. The low probability of a slope outlier
557 combined with the relatively high probability of unit and meniscus errors cause the model
558 to infer the slope outlier as a meniscus error and the correct observations as unit errors.
559 When one slope outlier error was paired with another error, the model required an ad-
560 ditional correct observation to accurately predict the true value (see Table 3 column 2).
561 For the best performance, the slope outlier error needs to be paired with at least one other
562 erroneous observation and a correct observation. When two errors were made out of two
563 observations submitted, the model often failed to correctly predict the true value. How-
564 ever, when a third observation without an error was included, the model predicted the
565 true value every time (see Table 3). Overall, the model inferred the correct error types
566 when the inferred true value was also correct.

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

574 **5.6 Further model applications**

575 The trained model was tested for two unique applications that provide insight into
576 the utility of the model in practical applications and the distribution of errors in citi-
577 zen science data over time.

578 **5.6.1 Citizen scientists with unknown characteristics**

579 As citizen scientist programs expand, recording complete characteristics data for
580 each participating citizen scientist may become challenging. The model's ability to in-
581 fer the correct community for citizen scientists with unknown characteristics and the cor-
582 rect true value for the observations they submit was investigated. The characteristics
583 for each unknown citizen scientist were selected from a discrete distribution estimated
584 from the characteristics data of citizen scientists observed during training. The prior dis-
585 tribution of the community, *PCom*, was set to a discrete distribution equal to the over-
586 all community posterior distribution of the training set. The community for each citi-
587 zen scientist and the true values of their submitted observations were inferred and com-
588 pared to the communities and true values inferred when the characteristics were known
589 precisely, but the community was also unknown.

590 The model performed quite well while inferring the community of unknown citi-
591 zen scientists and the true values of observations submitted by unknown citizen scien-
592 tists. Communities of citizen scientists with known characteristics were correctly pre-
593 dicted 0.9% more than citizen scientists with unknown characteristics. The coefficient
594 of determination between the actual true values and predicted true values was 0.015 higher
595 for known citizen scientists than for unknown citizen scientists. While the predicted true
596 values for known and unknown citizen scientists were similar, the uncertainty of the true
597 values predicted from observations submitted by unknown citizen scientists was higher.
598 The average variance of the inferred true value posteriors was 140.2 mm² for unknown
599 citizen scientists and 125.6 mm² for known citizen scientists. Overall, the value of sub-
600 mitted observations has greater influence on the inferred true values of rainfall than the
601 characteristics of the associated citizen scientist. While knowing the characteristics of
602 all citizen scientists increases the accuracy of predicting the true value of submitted ob-
603 servations, it is not essential.

604 **5.6.2 Evolution of error structure within communities**

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

612 The probability that a citizen scientist in each community would commit a type
613 of error changed from the 2017 to 2018 to 2019 S4W-Nepal program years (see Figure 6).
614 In 2017, only 16 citizen scientists for whom characteristics are known submitted obser-
615 vations (see Table 4). The 2017 community error distributions, particularly the Few-MUn,
616 Meniscus, and Unit-MUn communities, are highly uncertain due to the small sample size.
617 Overall, citizen scientists became increasingly active as S4W-Nepal's program progressed
618 through the years. Citizen scientists submitted an average of just over 8 observations in
619 2017, growing to 80 by 2019. In the first full year of rainfall submissions (2017), most
620 citizen scientists were assigned to the Few-MUn community. In the following two years,
621 active citizen scientists were most often in the Few community, followed by the Few-MUn
622 community. In all three years of S4W-Nepal's program, the RandU community repre-
623 sented the smallest fraction of active citizen scientists.

624 As S4W-Nepal gained experience in operating a citizen science program, the partic-
625 ipating citizen scientists also gained skills in collecting and submitting accurate rain-
626 fall observations. The Meniscus community had an increasing probability of submitting
627 correct observations in each year after 2017, while the Few-MUn community maintained
628 a low probability of submitting an erroneous observation (see Figure 6). The Few and
629 RandU communities also increased their probability of submitting a correct observation
630 in 2018 but saw a decrease in 2019. As the years progressed, all communities submit-
631 ted the same or successively fewer meniscus errors. Similarly, unit errors tended to de-
632 crease or remain the same as citizen scientists gained experience. Interestingly, while menis-
633 cus type errors and unit errors decreased over time, 2019 saw relatively high rates of un-
634 known errors. The reason for an increase in unknown errors is difficult to diagnose but
635 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.

636 regression parameters for this analysis are inferred rather than held constant, the un-
637 known error β decreases from 2.4 in 2017 to 1.7 in 2019. The error structure of obser-
638 vations submitted by citizen scientists is evolving as both S4W-Nepal and the partici-
639 pating citizen scientists gain experience, a common trend in citizen science programs (Kos-
640 mala et al., 2016).

641 S4W-Nepal uses various training techniques and feedback methods to increase the
642 scientific literacy of citizen scientists (Davids et al., 2019). Their methods have been ef-
643 fective in reducing the magnitude and frequency of errors committed by the citizen sci-
644 entists. Perhaps the best evidence for this change is the reduction in meniscus errors com-
645 mitted by citizen scientists in the Meniscus community. From 2018 to 2019, the prob-
646 ability of meniscus errors in the Meniscus community decreased from 19.0 to 8.1%. Sim-
647 ilarly, unit errors committed by those in the RandU community decreased from 6.4% in
648 2018 to 4.2% in 2019. While a trend in reduced meniscus and unit errors over two years
649 is promising, additional analysis after multiple years of collecting citizen scientist obser-

650 vations would provide more conclusive evidence for increased scientific literacy of the par-
 651 ticipants.

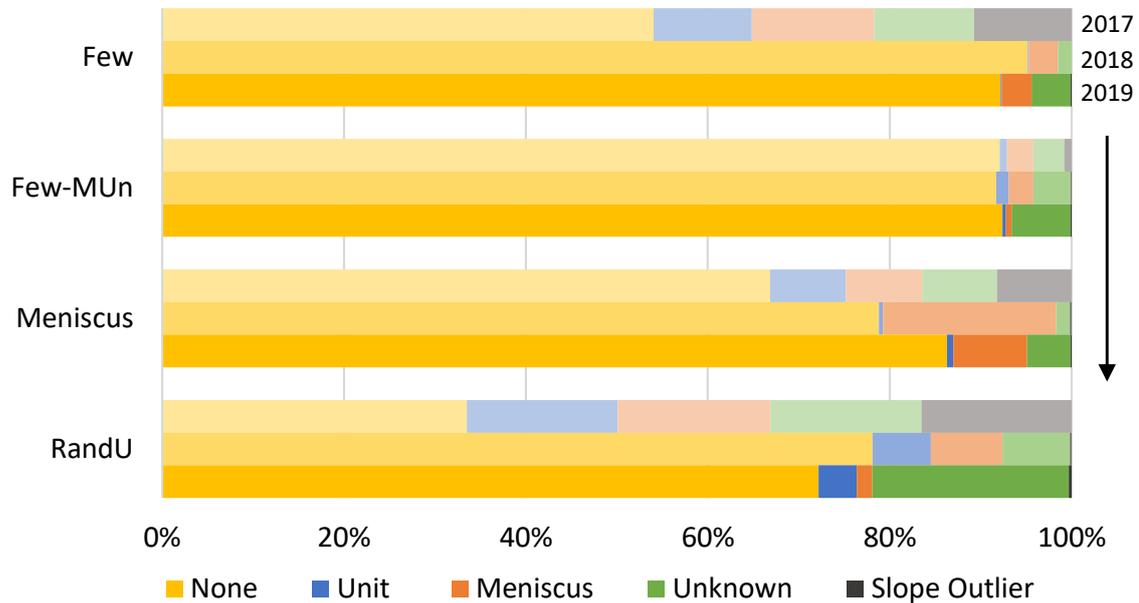


Figure 6. Change in the distribution of errors for each community over time. Note that the 2017 error distributions for the Few, Meniscus, and RandU communities are poorly informed due to the low number of active citizen scientists assigned to those communities.

652 5.7 Utility and limitations in application

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

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

681 **6 Summary and conclusions**

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

689 Analysis of the error structure in citizen scientist rainfall observations revealed that
690 individuals can be characterized by one of four error patterns: not error prone, mostly
691 not error prone, meniscus error prone, and random or various error prone. While the Bayesian
692 inference model developed here used communities to relate citizen scientist character-
693 istics to error tendencies, the magnitude and type of errors committed is the crux of ev-
694 ery community assignment. The distribution of characteristics within each community
695 is useful for investigating potential reasons for making errors rather than for identify-
696 ing individuals who might be particularly error prone.

697 The Bayesian inference model developed using Infer.NET’s software framework un-
698 covered five error types and their probability distribution within each of the four error-
699 based communities. The community assignments are a useful tool for discerning which
700 citizen scientists are more likely to submit erroneous observations that require further
701 review. In addition, community-specific training and feedback messages may be a pow-
702 erful tool for increasing the quality and frequency of submissions. The Bayesian prob-
703 abilistic model was often able to predict the true value of a submitted observation, and
704 the model extrapolated useful error probabilities for each observation. These error prob-
705 abilities, in conjunction with the model’s inferred error-specific regression and precision
706 parameters, can be used to calculate a Gaussian mixture distribution that provides more
707 information about the probable true value of submitted observations than Infer.NET’s
708 single-mode true value prediction. As citizen science programs expand to include mul-
709 tiple participants submitting observations of a single event, the model’s ability to pre-
710 dict the true value for that event will likely increase. However, the model’s potential may
711 be limited in regions where the target parameter is highly heterogeneous in space and
712 time.

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

722 **A Prior and Posterior Distributions**

723 The prior distribution for each inferred model variable was a uniform Dirichlet dis-
724 tribution, with the exception of the true value prior. The prior distribution for true value
725 was a Gaussian distribution with a mean of 15 and variance of 2400. The variance for
726 the true value prior was selected is four times the variance of the entire true value dataset
727 (i.e., twice the standard deviation). Note that Equation A.5 is the posterior distribution
728 for the model. The posterior is obtained by writing the joint distribution over latent vari-

ables $X = (PChar, PCom, PErr, \vartheta, \varepsilon, \gamma, \alpha_\varepsilon, \beta_\varepsilon, \tau_\varepsilon)$ and observed variables $D = (Z, O)$,
 followed by conditioning on the observations.

$$PChar_c | \gamma \sim Dirichlet(Uniform), \quad (A.1)$$

$$PCom_s \sim Dirichlet(Uniform), \quad (A.2)$$

$$PErr | \gamma \sim Dirichlet(Uniform), \quad (A.3)$$

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

$$p(X|D) \propto Dir(PCom|s) Dir(PErr|\gamma) \prod_{\varepsilon} \mathcal{N}(\mu_\alpha, \sigma_\alpha^2 | \varepsilon) \mathcal{N}(\mu_\beta, \sigma_\beta^2 | \varepsilon) Gamma(A, B | \varepsilon) \quad (A.5)$$

$$\prod_{c=1}^C Dir(PChar_c) \prod_{e=1}^E \mathcal{N}(\vartheta_e | \mu_e, \sigma_e^2)$$

$$\prod_{s=1}^S \left\{ Dis(\gamma_s | PCom) \prod_{c=1}^C \left\{ Dis(Z_{s,c} | \gamma_s, PChar_c) \right\} \right.$$

$$\left. \prod_{e=1}^E \left\{ Dis(\varepsilon_{s,e} | \gamma_s, PErr) \prod_{\varepsilon} \mathcal{N}(O_{s,e} | \alpha_\varepsilon \vartheta_e + \beta_\varepsilon, \tau_\varepsilon)^{\delta(\varepsilon_{s,e} - \varepsilon)} \right\} \right\},$$

where the Dirac delta function $\delta(\cdot)$ 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.

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

where $\mu_\alpha = (1, 0.1, 1.002, 0.9, 7)$, and $\sigma_\alpha^2 = (0.5, 0.5, 2, 50, 70)$. And,

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

where $\mu_\beta = (0, 0.02, 2.3, 4.2, 3)$, and $\sigma_\beta^2 = (0.5, 0.5, 0.2, 50, 30)$. The α and β mean and variance for the first four ε error types were based on the mean and variance of a series

746 of slopes and intercepts from linear regressions fit to subsets of (ϑ, O) pairs correspond-
 747 ing to the four error types identified by Davids et al. (2019). Note that the σ^2 values used
 748 are larger than calculated to provide a wider prior distribution. The mean and variance
 749 for the remaining error type was selected randomly, since there was no information avail-
 750 able regarding this error prior to training the model.

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

$$753 \quad \tau_{\varepsilon} \sim \text{Gamma}(A, B|\varepsilon) \quad (\text{A.8})$$

754 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
 755 and rate for the first four ε error types were calculated based a Gamma distribution fit
 756 to observations that corresponded to the four error types identified by Davids et al. (2019).
 757 The shape and rate for the remaining error type was selected randomly, since there was
 758 no information available regarding this error prior to training the model.

759 **Notation**

760 ***Dir*** Dirichlet distribution

761 ***Dis*** Discrete distribution

762 **\mathcal{N}** Gaussian distribution

763 **C** characteristic

764 **S** citizen scientist

765 **ε** error type

766 **e** event

767 **γ** Community

768 **O** SubmittedObservation

769 **ϑ** TrueValue

770 **Acknowledgments**

771 The dataset analyzed for this study can be accessed in the Supplementary Material pub-
 772 lished by Davids et al. (2019). This research has been supported by the National Sci-
 773 ence Foundation, Division of Graduate Education (grant no. DGE-1333468) and the Dutch
 774 Research Council. Data collection and quality control was supported by the Swedish In-

775 ternational Development Agency (grant no. 2016-05801) and by SmartPhones4Water
 776 (S4W). The authors declare that they have no conflict of interest. The authors would
 777 like to thank S4W's Saujan Maka for instrumental guidance.

778 References

- 779 Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J.,
 780 ... Frusher, S. (2014, May). Statistical solutions for error and bias in
 781 global citizen science datasets. *Biological Conservation*, *173*, 144–154. Re-
 782 trieved 2020-05-02, from [https://linkinghub.elsevier.com/retrieve/pii/](https://linkinghub.elsevier.com/retrieve/pii/S0006320713002693)
 783 [S0006320713002693](https://linkinghub.elsevier.com/retrieve/pii/S0006320713002693) doi: 10.1016/j.biocon.2013.07.037
- 784 Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg,
 785 K. V., & Shirk, J. (2009, December). Citizen Science: A Developing Tool
 786 for Expanding Science Knowledge and Scientific Literacy. *BioScience*,
 787 *59*(11), 977–984. Retrieved 2020-05-02, from [https://academic.oup.com/](https://academic.oup.com/bioscience/article-lookup/doi/10.1525/bio.2009.59.11.9)
 788 [bioscience/article-lookup/doi/10.1525/bio.2009.59.11.9](https://academic.oup.com/bioscience/article-lookup/doi/10.1525/bio.2009.59.11.9) doi:
 789 [10.1525/bio.2009.59.11.9](https://academic.oup.com/bioscience/article-lookup/doi/10.1525/bio.2009.59.11.9)
- 790 Bonter, D. N., & Cooper, C. B. (2012, August). Data validation in citizen science:
 791 a case study from Project FeederWatch. *Frontiers in Ecology and the Environ-*
 792 *ment*, *10*(6), 305–307. Retrieved 2020-05-03, from [http://doi.wiley.com/10](http://doi.wiley.com/10.1890/110273.1890/110273)
 793 [.1890/110273](http://doi.wiley.com/10.1890/110273.1890/110273) doi: 10.1890/110273
- 794 Brunsdon, C., & Comber, L. (2012). Assessing the changing flowering date of the
 795 common lilac in north america: a random coefficient model approach. *Geoin-*
 796 *formatica*, *16*(4), 675–690.
- 797 Budde, M., Schankin, A., Hoffmann, J., Danz, M., Riedel, T., & Beigl, M. (2017,
 798 September). Participatory Sensing or Participatory Nonsense?: Mitigating
 799 the Effect of Human Error on Data Quality in Citizen Science. *Proceedings of*
 800 *the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(3),
 801 1–23. Retrieved 2020-05-03, from <https://dl.acm.org/doi/10.1145/3131900>
 802 doi: 10.1145/3131900
- 803 Butt, N., Slade, E., Thompson, J., Malhi, Y., & Riutta, T. (2013). Quantifying
 804 the sampling error in tree census measurements by volunteers and its effect on
 805 carbon stock estimates. *Ecological Applications*, *23*(4), 936–943.
- 806 Cox, T., Philippoff, J., Baumgartner, E., & Smith, C. (2012). Expert variability pro-

- 807 vides perspective on the strengths and weaknesses of citizen-driven intertidal
808 monitoring program. *Ecological Applications*, *22*(4), 1201–1212.
- 809 Crall, A. W., Newman, G. J., Stohlgren, T. J., Holfelder, K. A., Graham, J., &
810 Waller, D. M. (2011, December). Assessing citizen science data qual-
811 ity: an invasive species case study: Assessing citizen science data qual-
812 ity. *Conservation Letters*, *4*(6), 433–442. Retrieved 2020-05-02, from
813 <http://doi.wiley.com/10.1111/j.1755-263X.2011.00196.x> doi:
814 10.1111/j.1755-263X.2011.00196.x
- 815 Davids, J. C., Devkota, N., Pandey, A., Prajapati, R., Ertis, B. A., Rutten, M. M.,
816 ... van de Giesen, N. (2019, March). Soda Bottle Science—Citizen Science
817 Monsoon Precipitation Monitoring in Nepal. *Frontiers in Earth Science*, *7*,
818 46. Retrieved 2020-04-23, from [https://www.frontiersin.org/article/](https://www.frontiersin.org/article/10.3389/feart.2019.00046/full)
819 10.3389/feart.2019.00046/full doi: 10.3389/feart.2019.00046
- 820 Delaney, D. G., Sperling, C. D., Adams, C. S., & Leung, B. (2008, January). Ma-
821 rine invasive species: validation of citizen science and implications for national
822 monitoring networks. *Biological Invasions*, *10*(1), 117–128. Retrieved 2020-
823 05-03, from <http://link.springer.com/10.1007/s10530-007-9114-0> doi:
824 10.1007/s10530-007-9114-0
- 825 de Solla, S. R., Shirose, L. J., Fernie, K. J., Barrett, G. C., Brousseau, C. S., &
826 Bishop, C. A. (2005). Effect of sampling effort and species detectability on
827 volunteer based anuran monitoring programs. *Biological Conservation*, *121*(4),
828 585–594.
- 829 Fink, D., Hochachka, W. M., Zuckerberg, B., Winkler, D. W., Shaby, B., Munson,
830 M. A., ... Kelling, S. (2010). Spatiotemporal exploratory models for broad-
831 scale survey data. *Ecological Applications*, *20*(8), 2131–2147.
- 832 Habib, E., Krajewski, W. F., & Ciach, G. J. (2001). Estimation of rainfall intersta-
833 tion correlation. *Journal of Hydrometeorology*, *2*, 621–629. doi: 10.1175/1525-
834 7541(2001)002<0621:EORIC>2.0.CO;2
- 835 Hunter, J., Alabri, A., & van Ingen, C. (2013, February). Assessing the quality and
836 trustworthiness of citizen science data. *Concurrency and Computation: Prac-*
837 *tice and Experience*, *25*(4), 454–466. Retrieved 2020-05-03, from [http://doi](http://doi.wiley.com/10.1002/cpe.2923)
838 [.wiley.com/10.1002/cpe.2923](http://doi.wiley.com/10.1002/cpe.2923) doi: 10.1002/cpe.2923
- 839 Kidd, C., Becker, A., Huffman, G. J., Muller, C. L., Joe, P., Skofronick-Jackson, G.,

- 840 & Kirschbaum, D. B. (2017, January). So, How Much of the Earth's Surface
 841 Is Covered by Rain Gauges? *Bulletin of the American Meteorological Society*,
 842 98(1), 69–78. Retrieved 2020-05-02, from [http://journals.ametsoc.org/
 843 doi/10.1175/BAMS-D-14-00283.1](http://journals.ametsoc.org/doi/10.1175/BAMS-D-14-00283.1) doi: 10.1175/BAMS-D-14-00283.1
- 844 Kosmala, M., Wiggins, A., Swanson, A., & Simmons, B. (2016, December). As-
 845 sessing data quality in citizen science. *Frontiers in Ecology and the Environ-
 846 ment*, 14(10), 551–560. Retrieved 2020-05-03, from [http://doi.wiley.com/10
 847 .1002/fee.1436](http://doi.wiley.com/10.1002/fee.1436) doi: 10.1002/fee.1436
- 848 Lukyanenko, R., Wiggins, A., & Rosser, H. K. (2019). Citizen science: An informa-
 849 tion quality research frontier. *Information Systems Frontiers*, 1–23.
- 850 MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*.
 851 Cambridge: Cambridge University Press.
- 852 Miller, D. A., Nichols, J. D., McClintock, B. T., Grant, E. H. C., Bailey, L. L., &
 853 Weir, L. A. (2011). Improving occupancy estimation when two types of ob-
 854 servational error occur: Non-detection and species misidentification. *Ecology*,
 855 92(7), 1422–1428.
- 856 Minka, T. (2000). *Bayesian linear regression* (Tech. Rep.). Citeseer.
- 857 Minka, T. (2005). *Divergence measures and message passing* (Technical Report No.
 858 TR-2005-173). Microsoft Research.
- 859 Minka, T. (2013). Expectation propagation for approximate bayesian inference.
 860 *arXiv preprint arXiv:1301.2294*.
- 861 Minka, T., & Winn, J. (2008). Gates. *Advances in Neural Information Processing
 862 Systems 21*, 1073–1080.
- 863 Minka, T., Winn, J., Guiver, J., Zaykov, Y., Fabian, D., & Bronskill, J. (2018). *In-
 864 fer.NET 0.3*. Microsoft Research Cambridge. Retrieved from [http://dotnet
 865 .github.io/infer](http://dotnet.github.io/infer)
- 866 Nayava, J. L. (1974, December). Heavy monsoon rainfall in Nepal. *Weather*, 29(12),
 867 443–450. Retrieved 2020-04-23, from [http://doi.wiley.com/10.1002/j.1477
 868 -8696.1974.tb03299.x](http://doi.wiley.com/10.1002/j.1477-8696.1974.tb03299.x) doi: 10.1002/j.1477-8696.1974.tb03299.x
- 869 Newman, G., Wiggins, A., Crall, A., Graham, E., Newman, S., & Crowston, K.
 870 (2012, August). The future of citizen science: emerging technologies and shift-
 871 ing paradigms. *Frontiers in Ecology and the Environment*, 10(6), 298–304.
 872 Retrieved 2020-05-02, from <http://doi.wiley.com/10.1890/110294> doi:

- 873 10.1890/110294
- 874 Nishihara, R., Minka, T., & Tarlow, D. (2013). Detecting parameter symmetries in
875 probabilistic models. *arXiv preprint, arXiv:1312.5386*.
- 876 Riesch, H., & Potter, C. (2014, January). Citizen science as seen by scien-
877 tists: Methodological, epistemological and ethical dimensions. *Public*
878 *Understanding of Science*, 23(1), 107–120. Retrieved 2020-05-02, from
879 <http://journals.sagepub.com/doi/10.1177/0963662513497324> doi:
880 10.1177/0963662513497324
- 881 Schoups, G., & Nasser, M. (2020). Gracefully closing the water balance: a data-
882 driven probabilistic approach applied to river basins in iran.
- 883 Sheppard, S. A., & Terveen, L. (2011). Quality is a verb: the operationaliza-
884 tion of data quality in a citizen science community. In *Proceedings of the*
885 *7th International Symposium on Wikis and Open Collaboration - WikiSym*
886 *'11* (p. 29). Mountain View, California: ACM Press. Retrieved 2020-05-
887 03, from <http://dl.acm.org/citation.cfm?doid=2038558.2038565> doi:
888 10.1145/2038558.2038565
- 889 Sunde, P., & Jessen, L. (2013). It counts who counts: an experimental evaluation
890 of the importance of observer effects on spotlight count estimates. *European*
891 *Journal of Wildlife Research*, 59(5), 645–653.
- 892 Teague, K. A., & Gallicchio, N. (2017). *The evolution of meteorology: a look into*
893 *the past, present, and future of weather forecasting*. Hoboken, NJ: John Wiley
894 & Sons, Inc.
- 895 Thapa, B. R., Ishidaira, H., Pandey, V. P., & Shakya, N. M. (2017, February). A
896 multi-model approach for analyzing water balance dynamics in Kathmandu
897 Valley, Nepal. *Journal of Hydrology: Regional Studies*, 9, 149–162. Re-
898 trieved 2020-04-23, from [https://linkinghub.elsevier.com/retrieve/pii/](https://linkinghub.elsevier.com/retrieve/pii/S2214581816303342)
899 [S2214581816303342](https://linkinghub.elsevier.com/retrieve/pii/S2214581816303342) doi: 10.1016/j.ejrh.2016.12.080
- 900 USGS. (n.d.). *DYFI Scientific Background*. Retrieved 2020-05-05, from [https://](https://earthquake.usgs.gov/data/dyfi/background.php)
901 earthquake.usgs.gov/data/dyfi/background.php
- 902 Venanzi, M., Guiver, J., Kazai, G., Kohli, P., & Shokouhi, M. (2014). Community-
903 based bayesian aggregation models for crowdsourcing. In *Proceedings of the*
904 *23rd international conference on World wide web - WWW '14* (pp. 155–164).
905 Seoul, Korea: ACM Press. Retrieved 2020-05-03, from <http://dl.acm.org/>

- 906 `citation.cfm?doid=2566486.2567989` doi: 10.1145/2566486.2567989
- 907 Vibhāga, N. K. T. (2012). *National Population and Housing Census 2011: National*
908 *report* (Vol. 1). Government of Nepal, National Planning Commission Secre-
909 tariat, Central
- 910 Wiggins, A., Newman, G., Stevenson, R. D., & Crowston, K. (2011, Decem-
911 ber). Mechanisms for Data Quality and Validation in Citizen Science. In
912 *2011 IEEE Seventh International Conference on e-Science Workshops* (pp.
913 14–19). Stockholm, Sweden: IEEE. Retrieved 2020-05-03, from [http://](http://ieeexplore.ieee.org/document/6130725/)
914 ieeexplore.ieee.org/document/6130725/ doi: 10.1109/eScienceW.2011.27
- 915 Winn, J., Bishop, C., Diethe, T., Guiver, J., & Zaykov, Y. (2020). *Model-based ma-*
916 *chine learning* (early access ed.). online: Microsoft Research. Retrieved from
917 www.mbmlbook.com