Reviewer 2

General Comments

The manuscript "A Bayesian model for quantifying errors in citizen science data: application to rainfall observations from Nepal" presented by J. Eisma et al. introduces a graphical Bayesian inference model to (1) analyze and categorize various error types present in citizen science data and (2) classify the citizens into groups (communities) according to the error distribution within each group. By considering specific error types, the model allows a comprehensive understanding of the error structure of crowdsourced data.

The model was applied to real crowdsourced rainfall observations collected within the project SmartPhones4Water in Nepal. The model identified five distinct error types and classified the citizens into four inferred communities based on their error patterns. Leveraging this information, the model enables the identification of observations that require further verification, reducing the burden of data validation on human efforts by employing machine-learned algorithms.

I enjoyed reading the manuscript and acknowledge the potential of using a Bayesian model as a novel approach to improve the efficiency and accuracy of data validation in citizen science. The findings underscore the importance of well-defined error structures in citizen science data and demonstrate the value of a graphical Bayesian inference models in understanding and harnessing such data effectively, which becomes more and more relevant with the increasing amount of crowdsourced data. Overall, the manuscript is well structured and contributes relevant insights to the emerging field of citizen science data collection and, subsequently, the use of such data for further research. I recommend considering this manuscript for publication in HESS with minor revision.

We appreciate your thoughtful review of our manuscript.

Specific Comments

L24: Consider removing the word "traditional" in front of scientists. What are "non-traditional" scientists – and – in general, all scientists should be concerned about data quality from whatever source.

Good point. The word "traditional" has been removed.

L229: Are the data also checked/calibrated by automatic rain gauges installed according to certain quality assurance standards? If so, this section may need to be briefly expanded to include a comparison of the overall data quality between CS data and automatically collected data. However, as the overall quality of CS data is not the focus of this manuscript, this comparison is not critical.

We did not do this comparison, but the original data paper (Davids et al., 2019) compared the rain gauges used with a few different standard rain gauges. The line to which you are referring has been removed, and the Davids et al. (2019) paper is referenced for a detailed description of the dataset. See section 3.1.

L252: As an additional filter for the data, the authors set a maximum limit of 540 mm of rainfall per day. My concern with this limit is that citizens may not be able to record such an event because the rain collector would overflow. In this case, the maximum amount of precipitation that can be measured by a CS station per day/measurement might be a more realistic limit. The authors should also report the number and percentage of data points that exceeded the upper (and lower) limit.

This is a great point. The S4W-Nepal rain gauges had an upper limit of 200 mm. We changed the value in the model and re-ran the simulations. It did not change the results for this particular dataset. The start of section 3.3 has been edited to read:

Before training and testing, an additional assumption was incorporated due to the nature of rainfall data: the inferred value of rainfall was assumed to be between 0 and 200 mm. Rainfall events cannot result in negative rainfall, and 200 mm is the maximum rainfall depth that can be recorded in a single measurement using the S4W-Nepal rain gauges per Davids et al., (2019). The CS rainfall observations analyzed here include no observations below 0 mm and two observations (0.03%) greater than 200 mm. One could include overflow as another type of error but given the rare occurrence, that was not included here.

L320: Maybe name the error type that was introduced with the model here (slope outliers). It is mentioned in section 4.3, but I was missing this information in this section. It might be also valuable to expand this section slightly to explain why "slope outliers" have been identified as an error type. When looking at the distribution of errors made within the communities (Table 2), slope outliers never occurred. The relevance of this error type remains unclear to me.

The name and a brief description of the significance of the error type have been added to section 4.2.

The model inferred one previously unidentified error type in addition to the four error types that were identified by S4W-Nepal's visual inspection of the submitted observations (Davids et al., 2019). The additional, model-inferred error type, named slope outlier, is significantly different from the other identified error types (see Table 2) and only occurs twice in the training and testing data. Each identified error type will be explored more fully in the next section.

L344: I would recommend using a different term for the Few-MUn group. The "few group" also makes only Meniscus and Unknown errors – similar to the Few-MUn group. The only difference is the overall amount of errors (2 % vs 5%). Hence, the groups could be named according to the amount of errors (such as p2 and p5 group, or minor and few, etc.). This would also improve the readability of the manuscript.

The Few-MUn community has been renamed to the Few+ community throughout the manuscript. For example, see Table 2 and Figure 5.

L457: The authors mention that a set of erroneous data is required to train the model and that these data need to be identified and corrected by the CS program, which can be a significant effort. Other studies have shown that this task could also be done in collaboration with the community (e.g., Strobl et al. https://doi.org/10.1371/journal.pone.022257). It may be of interest to the readers of this

study to include some information on this approach here. This is currently listed in Section 6 (Future work) but may fit better within the discussion in Section 4.6.

Thanks for making this connection for us. CrowdWater has a great solution to the identified limitation. Section 4.6 has been updated and now includes:

This may require a large effort and may be difficult to achieve, but at least one citizen science program, CrowdWater, has an innovative solution. The CrowdWater Application collects CS observations of stream stage, and the CrowdWater Game crowdsources the true value of the submitted stage observations (Seibert et al., 2019; Strobl et al., 2019).

L486: A limitation of this study is that it was only tested with one CS project in one region. The authors should mention this limitation more clearly in the conclusion, as it remains unclear whether the method and model developed will work equally well in different settings.

L476 has been expanded to read:

While the results are promising, the model was only tested with one citizen science program deployed in one country. Further testing with datasets from different citizen science programs is required to assess whether the method and model perform equally well. Applying the model to different citizen science datasets may require some of the model assumptions to be tailored to the specific application (e.g., range of acceptable values, censored data, etc.). However, the flexibility of the modelling tool used, Infer.NET, makes it simple to vary the model to suit the specific needs of different CS datasets.