Responses to Reviewer #1's Comments:

Please see general responses to your helpful comments below in blue (original comments in black). Once all the edits (per the details below) are finalized, a marked-up version of the manuscript showing all changes, along with specific responses to reviewers' comments will be provided.

This manuscript aims at studying the potential for citizen science streamflow measurement methods. Citizen science is so far underused in hydrology and studies on this topic are, thus, much welcome. The manuscript starts with a well-written introduction, where several relevant studies are cited. After this promising start, however, I was rather disappointed by the study. I really like the aim of this study, and I appreciate the attempt to evaluate the suitability of different streamflow gauging methods, but in the end, I have three major concerns. These are related to 1) the study design and data collection, 2) the data analyses and 3) statements that are not supported by the data analyses presented in the manuscript. A more detailed discussion of these issues and some minor comments are provided below.

RESPONSE: Your three main concerns are well received. It was timely to receive your comments when we did because we were just starting our post-monsoon Citizen Science (CS) Flow campaign in Kathmandu. Based on your comments, and Reviewer #2's feedback, we were able to design and implement additional data collection which was performed from 18 to 20 of September 2018. We feel that once these data have been incorporated along with the other suggested edits the revised manuscript will be strengthened.

From 18 to 20 of September 2018, we facilitated measurements at 15 sites in two different watersheds in the Kathmandu Valley. Ten CS Flow groups, each comprised of three students, performed all three methods (i.e. float, salt dilution, and Bernoulli) at each site. At the same time, an "expert" group (authors) performed the same three methods at the same sites, along with a FlowTracker ADV reference flow measurement. After the field measurements, all the CS Flow participants completed a survey about their experiences with (and perceptions of) each simple measurement method.

In the original version of the manuscript, there was some confusion about the two different data sets being evaluated, including who actually had generated the data. An additional table has been added to clarify the three phases (and datasets) of the study: (1) initial evaluation (authors), (2) citizen scientist evaluation (authors and CS Flow groups), and (3) citizen scientist application (CS Flow groups).

#	Phase	Description	Performed By	Period	Season
1	Initial Evaluation	Initial evaluation of three simple flow measurement methods (i.e. float, salt dilution, and Bernoulli) along with FlowTracker ADV reference flow measurements at 20 sites within the Kathmandu Valley. Reference flows ranging from 6.4 to 240 liters per second (L s^{-1}).	Authors	March/ April 2017	Pre-monsoon

2	Citizen Scientist Evaluation	Citizen Scientist evaluation of three simple flow measurement methods (i.e. float, salt dilution, and Bernoulli) along "expert" and FlowTracker ADV reference flow measurements at 15 sites within the Kathmandu Valley. Reference flows ranging from 4.2 to 896 L s^{-1}.	Authors for "expert" and reference flows PLUS 10 Citizen Science Flow groups for simple methods	September 2018	Post-monsoon
3	Citizen Scientist Application	Salt dilution measurements at roughly 150 sites in the 10 perennial watersheds of the Kathmandu Valley. Float measurements with a small number of sub-sections (e.g. 3 to 5) performed at each site to determine salt quantities.	17 Citizen Science Flow groups (7 from April and 10 from September)	April and September 2018	Pre and Post Monsoon

I am afraid that the concerns related to the available data require additional data to be collected. Frankly, I would say the presented work is an interesting pre-study, but a better study design and data collection are needed to obtain useful results. Publishing the preliminary results as presented here could do more harm than good as people might use the conclusions without being aware that there actually was little data evidence. Given the importance of the topic, I hope the authors will be able to do this and will resubmit a study, which addresses the issues they raise in this manuscript.

1. There are several severe flaws in the study design and in the end I am afraid the authors did not collect the data that would be needed to address the questions they wanted to study.
    a. It is highly unfortunate that there are no concurrent flow measurements for the 'true' flow available. Flow measurements taken a few weeks apart are just not the basis for a serious evaluation. It is also surprising the different 'citizen scientists' were asked to measure streamflow at different sites. It would have much more informative to let them measure the same stream and about the same conditions.

RESPONSE: At all 15 sites, the new data we collected has CS measurements and "expert" measurements of each simple method along with a reference flow with the FlowTracker ADV. We believe this now provides the data we need to make a comparison.

    b. The authors mention that three salt dilution measurements were excluded as outliers. While they present some explanation (which I do not fully agree), they do not present anything that would help to detect such cases in an application where there is no comparison with any other gauging. In other words, in a real application, these values would pass undetected, and the potential error, thus, would be much larger than reported here. Note that almost half of the cases with comparison streamflow data were excluded! Again, it is unfortunate that the authors need to very speculative about what might have happened because of the study design.

RESPONSE: We will re-evaluate this concern in light of the new data. Your point about not being able to detect such errors without a reference flow is well received and will be incorporated into the revised version of the manuscript.

c. Related to the above comment, one potential issue are mistakes that could be done by 'citizen scientists'. With a better study design (e.g., more groups at the same place, 'secret' observer, . . .), this could have been addressed.

RESPONSE: Because the new measurements are from the same 15 sites, we are now able to identify the variability in CS Flow group measurements, compared to the actual value, and "expert" values with the same methods (i.e. float, salt dilution, or Bernoulli).

d. Basically, there are two separate questions: 1) which of the 'simple' gauging methods provides best results (with 'perfect' persons) and 2) how re the methods used by 'citizen scientists'. By deciding the best methods already after the first step, the authors, unfortunately, do not fully explore which method is most suitable for citizen science approaches.

RESPONSE: To help answer both of these questions, the CS Flow groups now performed all three methods. Participants also completed a quantitative evaluation of their experience with (and perception of) each measurement methods to improve our understanding of citizen science suitability.

e. The accuracy of the salt dilution measurement depends largely on the selected site (mixing, flow volume, and velocity. . .), and depending on the site, thus, different methods might be most suiatble. Again, this is an important aspect that could have been adddressed with a better study design.

RESPONSE: Each CS Flow group was allowed to select their specific salt dilution measurement reach independently. They also performed Float and Bernoulli measurements, so now we will be able to make a full comparison of which method performed best, and how this varied with the type of site.

f. A minor point related to the study design: when the aim is to obtain relations of the calibrated k-factors with elevation or other variables, k should have been determined at as may places as possible and not just half of them. I am also not sure whether it is reasonable to use the mean k value was for the 10 locations without individual measurements, but the individual values for the others. I would rather have expected to use the mean or some regionalized values for ALL locations to ensure comparability.

RESPONSE: For the first 20 measurement sites performed by the authors, we will use the average of the 10 K values obtained for all 20 sites.

2. The data analyses contain some questionable use of statistics:
   a. Averaging of errors (tables 12): averaging positive and negative errors just does not make any sense, this makes the results look much better than they are. Instead, one should base the analyses on the absolute values so that positive and negative errors do not cancel out each other

RESPONSE: The averages for errors now average the absolute values so that positive and negative errors do not cancel each other out.

b. The correlations shown in figures 3 or 4 (and reported in the abstract) are misleading. These are spurious correlations! Comparing streams of different size, of course, one gets high r2 values. Imagine two persons would measure the height of a group of people, even if the individual measurements would be off by 5 cm, the correlation of the heights would still be large simply because some people are much taller than others. Please be more careful when using statistics.

RESPONSE: Instead of scatter plots, we will instead present box plots showing the distribution of error for each site.

3. The statements in section 4.4 are not really supported by the data in this study. The number of persons needed in each group, for instance, has not been tested. Also, the inexpensive EC meter has not been tested (or has it? comparison?).

RESPONSE: Participants in the post-monsoon (i.e. phase 2) measurements completed quantitative evaluations of the methods, so now we have additional evidence to answer these questions. Additionally, testing of the inexpensive EC meters has been performed and incorporated into the manuscript.

a. As another example, the statement of enjoyability seems not to be supported, and actually, the other methods have not been tested with 'citizen scientists. The authors also need to explain much better which type of citizen scientist they refer to. Form the title (where the term citizen science is used twice!), the abstract, the introduction and section 4.4 one gets the impression that this is about citizen science in a broad meaning. However, looking more closely at what has been done, it seems that the work does not address the participation of the general public in science but is based on selected individuals, which received a significant education. This is fine, but is a rather special case of citizen science.

RESPONSE: Participants in the post-monsoon (i.e. new) measurements completed quantitative evaluations of the methods, so now we have additional evidence to answer these questions. Additional clarifying language has been added to make it clear that we are initially targeting students as citizen scientists, but that this represents a narrow swatch of possible citizen scientists. Our experience has shown that in countries like Nepal, students provide an important "first wave" of citizen scientists, who can later promote citizen science to local community members. Over time, it is our goal to continue to expand the type of citizen scientists that we target and engage.

4. Minor comments:
    a. P4: which factor for c was used in the end? Variable or constant. This needs to be included in the steps.

RESPONSE: A constant C factor of 0.8 was used. This has been added to the steps.

b. P8: too little information is given about the 'citizen scientists': how old? Gender? Students, but which topic (how much hydrology or environmental engineering?), how large groups, ...

RESPONSE: A summary of the age, gender, major, and size of groups has now been included.

c. What is the purpose of showing figure 2?

RESPONSE: The original purpose was to show the breakthrough curves in order to illustrate that they are similar to standard curves. However, due to the large number of figures from the newly collected data, these graphs have now been removed.

d. P5L15: where does the value of 1667 g per m3/2 come from. Moore (2005) recommend a different value

RESPONSE: We have clarified the text to now state that an approximate average of values shown in Table 1 of Moore (2005) are the basis for our salt dose recommendations. The average mass of the studies in Moore (2005) is 1600 g per m^3 s^{-1}.

e. Tables 1 and 2: providing runoff with four digits seems a bit too accurate, especially given that the observations actually were weeks apart.

RESPONSE: Table 2 has now been removed. Because all flows observed are less than 1 m^3 s^{-1}, we have decided to present flows in liters per second (L s^{-1}). Flow greater than or equal to 100 L s^{-1} are now shown to the nearest integer. Flows less than 100 L s^{-1} are shown with two significant digits.

f. Please check the author guidelines, especially with regard to the date format and equations

RESPONSE: The author guidelines for date format and equations has been checked and the necessary revisions have been made.