Interactive comment on “Value of uncertain streamflow observations for hydrological modelling” by Simon Etter et al.

Anonymous Referee #2

Received and published: 3 August 2018

General Comments

The manuscript titled “Value of uncertain streamflow observations for hydrological modelling” is a helpful contribution to the growing body of literature on citizen science applications in hydrology. The article is scientifically significant, is of high quality, and is well presented. The objectives of the study are clearly stated, the methods are applicable, the results are clear, and the discussion and conclusions return to the original questions posed. The overall structure of the article is sound, and the prose is for the most part acceptable. However, efforts should be made to make the language more concise by separating long sentences and properly using commas and semi colons to join dependent and independent clauses, respectively.

We thank the reviewer for the positive comments about our manuscript and the helpful review comments, which we address in detail below.

The following are a few more general comments.

First, in the conclusion, only the first question regarding errors and not the frequency of observations is included; it is suggested that both questions be briefly addressed.

We thank the reviewer for this helpful comment. We agree that this should be addressed as well. We will include the following sentence in the revised manuscript: **“We, furthermore, demonstrated that realistic frequencies for citizen science projects (one observation on average per week or month) can be informative for model calibration. “**

Second, the “lower benchmark” is an important part of this study, and the one sentence dedicated to it (7-13/14) doesn’t provide enough information on how it was developed.

The use of upper and lower benchmarks to compare different model results follows the strategy of several recent studies (van Meerveld et al., 2017; Pool et al., 2017; Wang et al., 2017). Seibert et al. (2018) point out that it is important to assess what model performance is possible (upper benchmark) because the data used for model calibration and validation contain errors and a perfect model fit can't be expected, and to compare the model performance to what can be expected (lower benchmark) because the driving (precipitation and temperature) data often dictate that models can't be too far off for humid catchments, as long as the water balance is respected. The lower benchmark used in this study is therefore the median model performance for an uncalibrated model (based on 1000 random parameter sets).

We will extend the section in the revised manuscript:

“In humid climates, the input data (precipitation and temperature) often dictate that model simulations can't be too far off as long as the water balance is respected (Seibert et al., 2018). To assess the value of limited inaccurate streamflow data compared to a situation without any streamflow data, a lower benchmark (Seibert et al., 2018) was therefore used as well. Here the lower benchmark was defined as the median performance of the model ran with 1000 random parameters sets. By running the model with 1000 randomly chosen parameter sets, we represent a situation where no streamflow data for calibration are available and the model is driven only by the temperature and precipitation data. We used 1000 different parameter sets to cover most of the model variability due to the different parameter combinations.”

References:
van Meerveld, H J; Vis, Marc J P; Seibert, Jan (2017). Information content of stream level class data for hydrological model calibration. Hydrology and Earth System Sciences, 21(9):4895-4905.

Pool, Sandra; Vis, Marc J P; Knight, Rodney R; Seibert, Jan (2017). Streamflow characteristics from modeled runoff time series – importance of calibration criteria selection. Hydrology and Earth System Sciences, 21(11):5443-5457.

Seibert, Jan; Vis, Marc J P; Lewis, Elizabeth; van Meerveld, H J (2018). Upper and lower benchmarks in hydrological modelling. Hydrological Processes, 32(8):1120-1125.

Wang, Ling; van Meerveld, H J; Seibert, Jan (2018). Effect of observation errors on the timing of the most informative isotope samples for event-based model calibration. Hydrology, 5(1):4.

Finally, additional discussion of how training could possibly decrease errors in citizen science streamflow estimates should be included (perhaps this is also included in the other paper in review). For example, should the focus be on improving depth, width, or velocity measurements? Are there any simple tools that could be added to improve the estimates? For example, could photos of the site including a person for scale (for area) and short videos (for velocity) be used to identify (and possibly filter) high error estimates?

There are indeed multiple possibilities for training. These include tutorial videos, or providing a list with well-known streams and their ranges in width, depth, flow velocity and streamflow to indicate ball park numbers.
We will include a brief statement on potential training options. However we do not want to focus too much on potential training options because their advantages and effectiveness are not known yet: **“Options for training might be tutorial videos, as well as providing values for the width, average depth and flow velocity of some well-known streams (Strobl et al., in review).”**

Specific Comments (page # - line # - comment)
2-23/24 - The “stick-method” is unfamiliar and should have a reference or some description. Is this the same as the “float” method, or ?

We will clarify this by rewriting these sentences in the following way: “**Estimating streamflow is obviously more challenging than reading levels from a staff gauge but citizens can apply the stick or float method, where they measure the time it takes for a floating object (e.g., a small stick) to travel a given distance to estimate the flow velocity. Combined with estimates for the width and the average depth of the stream, this allows them to obtain a rough estimate of the streamflow**.”

4-7 - USBR Water Measurement Manual 2001 Ch 13.10 recommends variable surface velocity with depth

We are unfortunately not exactly sure what this comment refers to. We used a factor of 0.8 to correct for the decline in flow velocity with depth and to obtain an average velocity from the surface velocity. Text books (e.g. Harrelson, Rawlins, & Potyondy, 1994) recommend this correction factor. Hauet et al. (in review) and Morlot et al. (2018) showed that this correction factor is reasonable for most streams, except for concrete channels (see page 5, line 7 in the first submitted version). Even if the exact value of the correction factor is uncertain (e.g. varies between 0.6 and 0.95), the impact on the estimated streamflow is small compared to the errors in the estimates of the velocity, width and depth.

References:
Harrelson, C.C., Rawlins, C.L. & Potyondy, J.P., 1994. Stream channel reference sites: an illustrated guide to field technique. (http://www.treesearch.fs.fed.us/pubs/20753)
Hauet, A., Morlot, T. & Daubagnan L. (in review) Velocity profile and depth-averaged to surface velocity in natural streams: a review over a large sample of rivers.
Morlot T., Hauet, A., & L. Daubagnan, L., 2018. Computation of the coefficient relating depth-averaged velocities to surface velocity over a large sample of French cross-sections gauged with a current meter, Geophysical Research Abstracts, Vol. 20, EGU2018-1874.

5-2 - do you have raw velocity and area data to further evaluate if the errors come more frequently from velocity or area estimates? Perhaps if you have the width and depth estimates this can also help to unpack uncertainty in areas estimates further.

Strobl et al. (in review) show that the width can generally be estimated better than the depth and velocity. Here we would like to focus on the value of the resulting streamflow estimates for hydrological modelling. We will mention that the depth is particularly uncertain when we describe the options of training.

6-8 - Is the one point per hour randomly selected or ??? Is hourly data a plausible citizen science output? You later say (9-21/22) that this frequency is “very unlikely.” What was the frequency of the original data?

The measurements from the Swiss Federal Office for the Environment (FOEN) have a 10 minute interval. The values we used were hourly averages. Hourly data were used to run the model because this is the resolution of the precipitation data, and represents the highest resolution that is regularly used for hydrological models in Switzerland and the HBV-model. We used the hourly data also for the simulations with error, even though it is very unlikely to get such a high contribution rate for citizen science projects, because this allowed us to draw conclusions about the effects of errors (i.e. for cases where the temporal resolution is “optimal” and only the quality is bad).
We will insert the following sub-sentence in the manuscript: **“Hourly runoff time series (based on 10 minute measurements) for the six study catchments were obtained from the Federal Office for the Environment (FOEN; see Table 1 for the gauging station numbers).”**
and the following text after we describe the different scenarios:
"Except for the hourly data, these scenarios were based on our own experiences within the CrowdWater project (www.crowdwater.ch) and information from the CrowdHydrology project (Lowry and Fienen, 2013). The hourly dataset was included to be able to test the effect of the errors when the temporal resolution of the data is optimal (i.e., by comparing simulations with the hourly FOEN data and those with hourly data with errors)."

5-12 - it might be nice to more explicitly include a summary (e.g. bullet points) here of the four levels of error that you refer to later: none, low, medium, and high

In the revised paper we will include the following list:
“To summarize, we tested the following four cases:

- **No error: The data measured by the FOEN, assumed to be error-free, the benchmark in terms of quality.**
- **Small error: random errors according to the log-normal distribution of the snapshot campaigns with the standard deviation divided by 4.**
- **Medium error: random errors according to the log-normal of the snapshot campaigns with the standard deviation divided by 2.**
- **Large error: typical errors of citizen scientists, i.e. random errors according to the log normal distribution of errors from the snapshot campaigns.”**

7-13/14 - perhaps the range bounds on the parameters for the random selections need to be discussed further

We agree that Table 1 in the supplemental material with the range of the parameters should be mentioned in the text. We will add a sentence in chapter 2.6: **“The parameters were calibrated within the typical ranges of the parameters (see Supplemental Material – Table 1).”**

9-25 - rather than “reduced errors” it would be better to specific either low or medium like you do later in the sentence

We agree. We will change the sentence into: **“With medium errors, however, and one data point per week on average or regularly spaced Monthly data, the data were informative for model parameterization.”**

9-27/28 - it would be good to consistently use either “lower benchmark” or “random parameter datasets”

We agree. We will change it to **“lower benchmark”**.

10-25 - it is unclear whether “fewer data points” here is referring specifically to calibrations with only 12 observations or to calibrations with even fewer than 12 observations (which wasn’t evaluated)

We agree that this statement is unclear. With “Fewer data points” we meant that the performance of models generally decreased faster with increasing errors if 12 instead of 48-52 data points were available. We will rewrite this sentence: **“…the results of this study also suggest that the performance of models decreases faster with increasing errors when fewer data points are available (i.e. there was a faster decline in model performance with increasing errors for models calibrated with 12 data points than for the models calibrated with 48-52 data points).”**

10-27/28 - only if the errors don’t contain systematic bias; please clarify

Indeed, errors only average out when more data points are included if the errors don’t contain a systematic bias. Our errors include a small overestimation but apparently the effect of this small bias is not strong. We will change the sentence: **“These findings can be explained by the compensating effect of the number of observations and their accuracy because the random errors for the inaccurate data average out when a large number of observations are used, as long as the data do not have a large bias.”**

11-7/8/9 - this sentence doesn’t seem to match the main point discussed earlier in the paragraph. Earlier you state that monthly performed better than IntenseSummer and WeekendSummer which had roughly 5 times more measurements. The you say it is “easier to get a certain number of observations…” Is it rather easier to get measurements spread out through the entire year than a certain number of measurements with citizen science?

Thanks for pointing at these confusing statements
The statement at 11-7/8/9 points to the fact that it is likely easier to obtain a certain number of observations distributed over the year than at very specific times or flow conditions because people can contribute whenever they want . The goal of the statement that the Monthly dataset performs better than the IntenseSummer and WeekendSummer datasets is to make it clear that fewer data can be more useful if they are distributed over the entire year (likely because they contain more information on the streamflow-variability). The term “certain number” is therefore confusing and not necessary. We will delete it and rewrite the sentence: **“This is good news for using citizen science data for model calibration as it suggests that the tim-**

ing is not as important as the number of observations because it is likely much easier to get observations throughout the year than during specific periods or flow conditions.”

Technical Corrections (page # - line # - comment)

1-7 - “….model can be parameterized using on a limited…” need to either remove “on” or modify sentence otherwise

Thanks for this suggestion for improvement. We will delete the word “on”.

1-16/17 - suggest using more commas to properly phrase the content (also the last sentence of the abstract could benefit from the same)

Thanks for making us aware of this, we changed the indicated sentence accordingly: “**These included scenarios with one observation each week or month, as well as scenarios that are more realistic for crowdsourced data that generally have an irregular distribution of data points throughout the year, or focus on a particular season**.”

1-29 - punctuation for the question within the sentence should be used: …question: how much data… are not available?”

Thanks for this suggestion. We will change it accordingly.

2-14 - same issue here where you end the sentence without a question mark. Either edit similar to above or rephrase: “but the question of how informative low quality data are remains.”

Thanks for pointing out also the second case. We will change it to: **“These results are encouraging for the calibration of hydrological models for ungauged basins based on a limited number of high quality measurements, but the question remains: how informative are low quality data?”**

3-5 - should define HBV here (first use) instead of below

Thanks for pointing this out, we will change it accordingly.

4-18/19 - sentence is incomplete

Thanks for pointing this out. We changed it to: “**For the validation, we chose the year closest to the mean summer streamflow and the years with the lowest and the highest summer streamflow sums (see Table 2).”**

6-17 - it seems more logical to include Crowd52 and Crowd12 in the bullet list of the six other temporal resolutions presented

We agree: We will adapt the section in the revised manuscript.

9-2 - correct grammar error “…was larger for than the…”

Thank you for pointing this out, we will correct it.

9-13 - which year are you referring to here: calibration or validation?

Thanks for pointing at this shortcoming: We will edit the sentence to clarify this: **“For 13 out of the 18 catchment and year combinations, the Crowd52 datasets with fewer than 10 % high streamflow data points led to a better validation performance than the Crowd52 datasets with more high streamflow data points.”**

13-19 - “…this data was not statistically significant better…” needs to be revised to possible “…these data did not show statistically significant improvements in model performance…”

Thanks for this suggestion for improvement. We will change the sentence in the revised version into: **“… (i.e. the median performance of the models calibrated with these data was not significantly better than the median performance of the models with random parameter values).”**