Accuracy of crowdsourced streamflow and stream level class estimates

Strobl, Barbara ; Etter, Simon ; van Meerveld, H J ; Seibert, Jan

DOI: https://doi.org/10.1080/02626667.2019.1578966

Accuracy of crowdsourced streamflow and stream level class estimates

Barbara Strobl[a]*, Simon Etter[a], Ilja van Meerveld[a], and Jan Seibert[a,b]

[a] *Department of Geography, University of Zurich, Zurich, Switzerland;* [b] *Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden*

* barbara.strobl@geo.uzh.ch; ORCiD: https://orcid.org/0000-0001-5530-4632

Etter Simon – ORCiD: https://orcid.org/0000-0002-7553-9102

van Meerveld Ilja – ORCiD: https://orcid.org/0000-0002-7547-3270

Seibert Jan – ORCiD: https://orcid.org/0000-0002-6314-2124

Abstract Streamflow data are important for river management and the calibration of hydrological models. However, such data are only available for gauged catchments. Citizen science offers an alternative data source, and can be used to estimate streamflow at ungauged sites. We evaluated the accuracy of crowdsourced streamflow estimates for 10 streams in Switzerland by asking citizens to estimate streamflow directly, or based on the estimated width, depth and velocity of the stream. Additionally, we asked them to estimate the stream level class by comparing the current stream level with a picture that included a virtual staff gauge. To compare the different estimates, the stream level class estimates were converted into streamflow. The results indicate that stream level classes were estimated more accurately than streamflow, and more accurately represented high- and low-flow conditions. Based on this result, we suggest that citizen science projects focus on stream level class estimates.

Keywords citizen science; crowdsourcing; stream level; stream level class; streamflow; accuracy; CrowdWater

How to cite:
Barbara Strobl, Simon Etter, Ilja van Meerveld & Jan Seibert (2019): Accuracy of crowdsourced streamflow and stream level class estimates, Hydrological Sciences Journal, DOI: 10.1080/02626667.2019.1578966

1 Introduction

Streamflow data are important for many aspects of river management, including water allocation and the reduction of flood hazards. Streamflow data are also important for the calibration of hydrological models to predict floods and droughts or the impacts of climate change. Most hydrological models need at least a certain amount of data to be properly "tuned" to a particular catchment (Beven 2012).

Three important aspects define the usability of streamflow data: accuracy, spatial coverage and temporal resolution. Conventional streamflow gauging stations can provide detailed information with a high accuracy and temporal resolution, but the spatial coverage is limited. While data from gauging stations are considered accurate the data can still contain substantial errors due to sensor errors, interpolation and extrapolation of the rating curve and cross-section instability (McMillan et al. 2012). Typical relative errors for streamflow are ±50–100% for low flows and ±10–20% for medium or high flows (still within the streambank) (McMillan et al. 2012). Similar values were derived by Westerberg et al. (2011), who mention rating curve related errors of –60% to +90% for low flows and ±20% for medium to high flows.

The temporal resolution of gauging stations is often high. However due to financial and logistic constraints only a few sites have a gauging station, hence the spatial coverage is limited. Furthermore, these stations may not be installed at representative locations or might miss certain types of catchments, especially small headwater streams (Kirchner 2006, Bishop et al. 2008). Also relatively few measurement stations are located in developing countries. Thus, for many catchments there are no streamflow data available for water management decisions or model calibration.

Although new wireless sensor network technology provides the possibility to expand the measurement networks, the reality is that due to budget cuts observation networks often shrink rather than expand (Kundzewicz 1997, Ruhi et al. 2018). For example, Ruhi et al. (2018) showed that between 1947 and 2016 the number of stream gauges in river basins in the USA decreased by 21%.

Several studies have focused on the minimum number of measurements required to properly calibrate a hydrological model (Perrin et al. 2007, Juston et al. 2009, Seibert and Beven 2009, Seibert and McDonnell 2015, Vis et al. 2015) and have shown that even a few streamflow measurements can vastly improve the performance of a model

(Pool et al. 2017). While employees of agencies responsible for national or regional gauging station networks could perhaps take some additional measurements at a few ungauged streams, it is impossible for them to take a limited number of measurements at all ungauged streams. An interesting alternative to obtaining streamflow data for more streams is to ask citizen scientists or citizen observers to collect streamflow data.

Citizen science has been used in numerous environmental studies to obtain data with a much higher spatial resolution than is otherwise possible (Dickinson et al. 2010, Tulloch et al. 2013, Aceves-Bueno et al. 2017, Hadj-Hammou et al. 2017) and has been used to obtain hydrological data as well (Buytaert et al. 2014). For example citizen science data have been used to fill in spatial and temporal gaps in water quality and stream level data series (Lowry and Fienen 2013, Hadj-Hammou et al. 2017) and to obtain groundwater level data across large areas (Little et al. 2016). Citizen science could therefore be a complementary approach to collect the stream level and streamflow data that are needed for hydrological model calibration, particularly for the many streams that are currently ungauged. In order to involve as many citizens in data collection as possible and to obtain data for remote areas, approaches are needed to collect these data with very little time-effort and without special equipment.

Despite their potential to complement existing data sources, citizen science data are not without challenges, especially the accuracy of crowdsourced data is often discussed (Engel and Voshell 2002, Haklay 2010, See et al. 2013, Aceves-Bueno et al. 2017). Several studies have examined the accuracy of crowdsourced hydrological data (Turner and Richter 2011, Rinderer et al. 2012, 2015, Lowry and Fienen 2013, Peckenham and Peckenham 2014, Breuer et al. 2015, Le Coz et al. 2016, Little et al. 2016, Weeser et al. 2018). Lowry and Fienen (2013) found promising results in terms of the accuracy of stream level data from participants who read the level from a staff gauge in a stream close to a hiking path. The root mean square error (RMSE) of the crowdsourced stream level data was approximately 5 mm, which was almost as good as that of pressure transducer data. They concluded that the level of accuracy *"is encouraging since no training was given to the citizen scientists"* (Lowry and Fienen 2013, p.155). In a similar study by Weeser et al. (2018) in Kenya, data collected by citizens were comparable to that of conventional data loggers, although it had a low temporal resolution. Little et al. (2016) provided volunteers with equipment to measure the water level in their own wells. They found that the absolute difference of the well

reading errors ranged from 2 to 11 mm and concluded that *"community-based groundwater monitoring provides an effective and affordable tool for sustainable water resources management"* (Little et al. 2016, p.317). Peckenham and Peckenham (2014) analysed groundwater quality data collected by students and concluded that the accuracy varied, but "*it is possible to make precise and accurate measurements consistent with the methods specifications"* (Peckenham and Peckenham 2014, p.1477).

However, these previous hydrological citizen science studies are not easily scalable to many sites because they require the installation of staff gauges or other instrumentation. Therefore, it is useful to also develop and test citizen science approaches to collect streamflow or stream level data that do not require equipment or the installation of staff gauges but these new citizen science tasks should be designed "*with the skill of the citizens in mind*" (Aceves-Bueno et al. 2017, p.287). It is likely that many citizens who frequently pass by streams notice high and low flows throughout the seasons. These frequently visited locations could be turned into locations for streamflow or stream level class observations if citizens can accurately estimate streamflow or stream level classes.

Testing the accuracy of citizen science data before starting a citizen science project is crucial for every citizen science project. This ensures that the data collected are sufficiently accurate for the purpose of the project and avoids unnecessarily burdening citizens with tasks that result in data that are in hindsight of limited value due to data accuracy issues. The objective of this study was, therefore, to determine what types of parameters related to streamflow citizens can estimate accurately. We asked 517 citizens to estimate both the streamflow and stream level class and assessed whether one can be estimated more accurately than the other by calculating the corresponding streamflow for each stream level class estimate. Accuracy is defined here as the difference between the estimated value and the measured value, as well as the frequency of extreme outliers. The specific research questions of this study were:

(1) How well can stream level class, streamflow and the different factors of streamflow (width, depth, flow velocity) be estimated by citizens?
(2) To what extent do stream size and flow conditions affect the accuracy of the crowdsourced data?

2 Methodology

2.1 Basic approach and study sites

We conducted 16 field surveys where we asked people to estimate the streamflow, as well as the average width, depth and velocity of the stream, and the stream level class. For the surveys, we selected 10 locations (Table 1; see also Supplementary material, Fig. S1) where we expected enough people to pass by and have time for the survey. We divided the streams into four different size classes (XS, S, M, L) based on the mean annual streamflow, and when long-term time series were not available, based on the available measurements:

- XS (Chriesbach, Hornbach and Irchel): ≤ 1 m^3/s,
- S (Glatt, Magliasina, Schanzengraben, Sihl and Töss): >1–50 m^3/s,
- M (Limmat): >50–200 m^3/s, and
- L (Aare): >200 m^3/s.

To analyse whether the flow conditions affect the accuracy of the estimates, surveys were conducted under high -and low-flow conditions for three streams: Aare (L), Limmat (M) and Sihl (S).

The aim of the surveys was to get a sufficient number of streamflow estimates for a specific stream on a specific day (our aim was 30 participants per survey to assure statistical significance; Field et al. 2013). We therefore used a logistically simple sampling strategy, whereby we personally approached passers-by (similar to Breuer et al., 2015) and asked if they would complete the 5-minute survey (i.e. we did not use a targeted approach to capture responses of a representative group of citizens). No data were collected on the percentage of passers-by who participated, but we estimate that about every third person we approached agreed to participate in our survey. In addition, we asked high-school (Magliasina) and university students (Irchel, Glatt and Limmat) to fill out the survey during excursions. All surveys took place between October 2016 and September 2017. In total, we received 517 complete surveys: 372 passers-by, 61 participants from a university geography bachelor student excursion (Glatt and Chriesbach), 40 from a high school student excursion (Magliasina) and 44 from a summer school for PhD students from fields ranging from physics to social sciences (Limmat) (see Table 1). During the group excursions we emphasized the need for

individual estimates and limited discussions between the students for the duration of the survey.

The age distribution of all 517 participants corresponds to that of the inhabitants of Zurich (where most field surveys were conducted), although there were fewer participants over the age of 60 (13% of the participants *vs* 19% of the population in Zurich; see Supplementary material, Fig. S2(c–)) (Statistik Stadt Zürich 2017). There was an almost equal split between male and female participants (Fig. S2(a)). A large number of participants were university-educated, roughly 48% compared to 16% of the population in Zurich (Fig. S2(b)) (Statistik Stadt Zürich 2017).

[Table 1 near here]

2.2 Streamflow estimation

Participants were first asked to estimate the streamflow directly. For this direct estimate, we asked them to estimate the flow in m^3/s, or in l/s for the very small streams (XS). This directly estimated streamflow value is referred to as Q_{direct}. This task, understandably, proved to be difficult for some participants because streamflow quantification was difficult and they were unfamiliar with the units. A few participants refused to answer this question, even with a bit of prompting. Some decided to guess, even though they thought it was unlikely to be a realistic value and others deduced on their own that they could estimate the width, mean depth and flow velocity to get an approximate value.

After this initial guess of the streamflow, we explained to the participants that it is possible to estimate the individual factors (width, mean depth and flow velocity) and to derive the streamflow by multiplying these values (eq. 1). The participants were then asked to estimate the average width, mean depth and velocity of the stream. We also asked them to classify the streambed material. Equation 1 was used to calculate the streamflow using these factors:

$$Q_{factor} = w \cdot d \cdot v \cdot k \quad (1)$$

where Q_{factor} is the estimated streamflow (m^3/s), w is the estimated width (m), d is the estimated mean depth (m), v is the estimated surface flow velocity (m/s) and k is the correction factor to obtain the average velocity from the surface velocity. While some participants still found the quantification difficult, they were more familiar with these

units, compared to m^3/s or L/s. Often a value of 0.85 is used for the correction factor k (Welber et al. 2016); it can also be estimated using the logarithmic velocity distribution (Prandtl-von Kármán equation) for turbulent flow based on the surface flow velocity, grain size and stream depth (Dingman 2015). This calculated factor for the mean flow velocity varied for the different estimates of the participants (even for the same stream). For two thirds of all estimates, the calculated velocity factor was not within the typical range of 0.71–0.95 (Welber et al. 2016) due to an unrealistic ratio between the estimated average water depth and estimated streambed roughness. Values lower than 0.71 were adjusted to 0.71 (52% of estimates) and values over 0.95 were adjusted to 0.95 (1% of estimates). When no estimate for streambed roughness was available (this happened only occasionally, except for the entire field survey at Magliasina), the typical velocity correction factor of 0.85 was used (including the participants at Magliasina this corresponds to 13% of all surveys). During the university excursion at the Glatt and Chriesbach, we did not ask for direct stream estimates because most geography bachelor students would likely have applied the indirect estimation method (Q_{factor}) because of lectures on streamflow during their education.

To assess the accuracy of crowdsourced streamflow data, the streamflow estimates were compared to measured streamflow data. Streamflow was measured before or after the surveys (Chriesbach, Hornbach, Irchel and Schanzengraben) or obtained from official gauging station data when these were located near the survey location (Aare, Limmat, Magliasina and Sihl, stations of the Swiss Federal Office for the Environment (FOEN); Glatt and Töss, stations of the Office of Waste, Water, Energy and Air of Canton Zurich (WWEA)) (see Table 1). The methods for the reference measurements for width, mean depth and flow velocity depended on the size and accessibility of the river. These measurements included direct measurements for width and depth with measurement tapes, data on the stream cross-section from FOEN for width and depth (when available), an estimate of the width of the river from Google Maps for wide rivers (Aare and Limmat) and the stick method for flow velocity. Even though these measurements are likely also affected by errors, they were assumed to be the "true" data to which the citizen science estimates could be compared. We assumed that the uncertainty for the measured values is 10% for streamflow (Pelletier 1988), 0.5% for width and 1-3% for depth (Herschy 1971) and roughly 10% for flow velocity (based on our own measurements).

2.3 Stream level class estimation

We also asked participants to estimate the stream level class. Stream level refers to the height of the water in a stream. A stream level class means that this height is expressed on a discrete scale of classes, rather than on a continuous scale. Stream level class data only provide information about whether the stream level is higher or lower than previously, but earlier studies have shown that stream level class data are useful for hydrological model calibration (van Meerveld et al. 2017). Thus, the participants were not asked to estimate the stream level in centimetres but to estimate the stream level class. The participants compared the current stream level with a photo of the same stream (taken at an earlier time) with a digitally inserted staff gauge with 10 level classes (Fig. 1, also Supplementary material, Section S2). The staff gauge was scaled so that the highest class represented the highest in bank flood level and the lowest class represented the likely lowest stream level. The height of the classes is arbitrary and varied for each location, depending on the size of the river and on how the virtual staff gauge was placed in the picture. A small staff gauge would have a higher resolution, but the stream level for very high and low flows may be above or below the staff gauge, whereas a large staff gauge would imply a lower resolution of the observations as the stream level would fluctuate across fewer classes. In this study we tried to place the staff gauges so that the staff gauge covered both high and low in bank flows. The number of classes was a compromise between resolution and usability. A larger number of classes provides higher resolution data but also makes it more difficult (or even impossible) for participants to determine the stream level class. Based on a previous model study model calibration results do not improve much when more than five stream level classes are used (van Meerveld et al. 2017). The number of ten classes was chosen to ensure observable stream level fluctuations even in cases where the virtual staff gauge is placed so that some classes are never or very rarely reached. The staff gauge was scaled so that the highest class represented the highest flood level and the lowest class represented the likely lowest stream level. The correct stream level class value was determined by us by carefully choosing appropriate references and individually (but unanimously) deciding on the correct stream level class.

For the Limmat, results are given for all five field surveys for streamflow, but stream level class estimates are given for only four surveys because a slightly different virtual staff gauge was used for the first survey.

[Figure 1 near here]

2.4 Data analyses

To be able to compare the accuracy of the streamflow estimates for different streams, relative estimates (in percent) were calculated by dividing the streamflow estimate by the measured value (i.e. considered true value). A value of 100% corresponds to a perfect estimate, smaller values represent an underestimation and larger values represent an overestimation. The quality of the data was then assessed by statistical measures, such as the interquartile range and median. In addition, we determined the number of outliers as they are likely disinformative for model calibration (Beven & Westerberg 2011) and can be worse than having no data. Even though filters can be used to remove outliers in citizen science data, in practise, it may be difficult to filter out all outliers. All relative estimates below 50% and above 150% were considered to be outliers.

For comparison between streamflow and stream level class estimates, stream level classes and the errors in this classification were converted to an equivalent streamflow (m^3/s), named Q_{level} in the remainder of the manuscript. For the stream locations with a nearby FOEN gauging station (Sihl, Limmat, Aare), the classes of the virtual staff gauge were converted to a metric value by determining the stream depth that corresponded to each stream level class (i.e. mid-point and upper and lower stream level for each class) and using the FOEN rating curve to convert these stream levels to a streamflow estimate. For the sites where no rating curve was available (Hornbach, Irchel, Schanzengraben and Töss), additional measurements of the stream profile and water surface slope (estimated based on the slope of the streambed) were used to estimate the streamflow for each stream level class using the Manning-Strickler formula (Manning 1891).This curve was fitted through the streamflow measured on the day of the surveys by adjusting the roughness coefficient within predefined boundaries based on the streambed material. The roughness coefficient used for the Manning-Strickler formula introduces some subjectivity and thereby likely increases the uncertainty of the conversion of the stream level class to streamflow compared to FOEN rating curve measurements. Because the stream level classes represent a range of values, rather than just one value, the streamflow was not only calculated for the centre value of the level class, but also the class boundaries to obtain the possible range of streamflow values. The estimates from Chriesbach, Glatt and Magliasina were excluded from this analysis

(101 of the 517 estimates) because the relevant data were not collected at the time of the surveys.

The differences in the median relative estimates for the different stream size classes were tested for significance using the Kruskal-Wallis test with the post hoc procedure based on Dunn (1964). Differences in the median relative streamflow estimates between high and low flow conditions were tested for significance using the Mann-Whitney test. A p-value of 0.05 was used for all statistical tests, unless otherwise indicated.

3 Results

3.1 Streamflow estimates

Although there was a large spread in the streamflow estimates, the median values were surprisingly close to the measured streamflow (Figs 2 and 3). Across all surveys the median of the direct streamflow estimates (Q_{direct}) was closer to the measured value than the estimate based on the factors (Q_{factor}) (median relative estimates of 93% and 80% respectively, when all surveys were analysed together). However, the interquartile range was smaller for the streamflow calculated from the estimated factors (the first and third quartiles were, respectively, 26% and 309% for Q_{direct} and 39% and 172% for Q_{factor};

Fig. 3

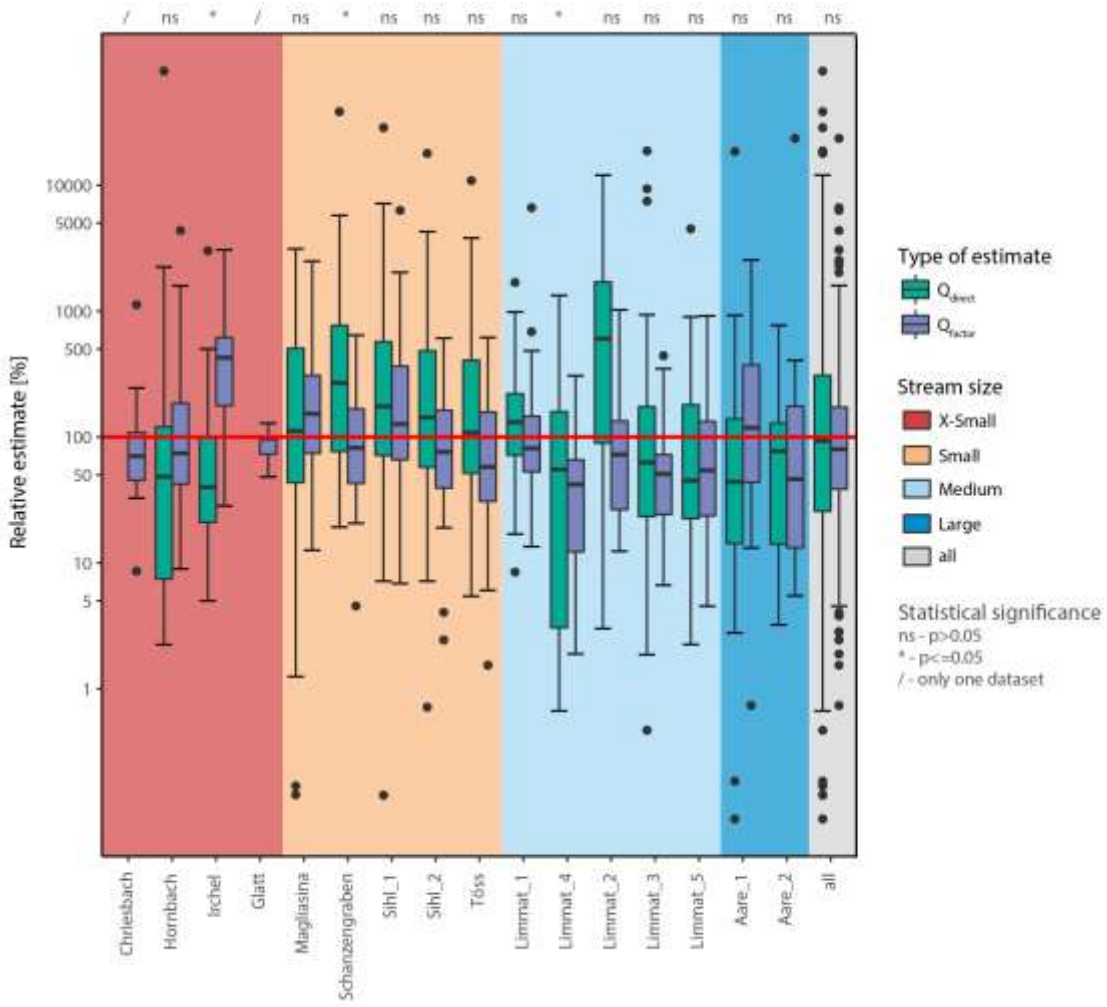

Figure), meaning that the streamflow estimates were closer to the measured value for the estimates based on the factors.

The differences between the median estimates of Q_{direct} and Q_{factor} were statistically significant ($p < 0.05$) for three out of the 14 surveys with both Q_{direct} and Q_{factor} estimates, but not for all surveys combined (Fig. 3). Of these three surveys, two had a median estimate for Q_{direct} that was closer to the measured value. The interquartile range was smaller for Q_{factor} for two of those three surveys.

[Figure 2 near here]

[Figure 3 near here]

3.2 Streamflow factor estimates

There were also numerous outliers for the relative estimates of width, mean depth and flow velocity (Fig. 4). The median relative estimates for the width, depth and flow

velocity were all significantly different from each other (Fig. 4). The width was generally underestimated (median relative estimate of 75%, and third quartile of 95% when all stream surveys were analysed together), the mean depth was generally overestimated (median relative estimate of 126% when all stream surveys were analysed together), while the median flow velocity was surprisingly accurate (median relative estimate of 100% when all stream surveys were analysed together). However, the interquartile range suggests that width can be estimated most accurately (interquartile range of relative estimates from 57 to 95% when looking at all surveys together) and mean depth (interquartile range of relative estimates from 86 to 180%) and flow velocity (interquartile range of relative estimates from 57 to 143%) can be estimated less accurately. The percentage of relative estimates below 50% or above 150% shows the same pattern, with width having fewer outliers (26%) than flow velocity (39%) and mean depth (41%) (Fig. 4).

[Figure 4 near here]

3.3 Stream level class estimates

About half of the participants (48%) selected the correct stream level class and most of the remaining participants (40%) were out by only one class. There were only a few outliers (13% of participants had an error of two classes or more; the total does not add to 100% due to rounding) (Fig. 5(a)). The largest overestimation was six classes and the largest underestimation was three classes. These errors likely occurred due to a misunderstanding of the method.

[Figure 5 near here]

3.4 Comparison of stream level class and streamflow estimates

To allow comparison of the streamflow and stream level class estimates, the latter were translated into corresponding streamflow values. These calculated streamflow values had a narrower interquartile range than the streamflow estimates based on the factors (67–157% compared to 30–163% for Q_{level} and Q_{factor} respectively, when all estimates are compared together) and also had fewer outliers (see Fig. 6). Only 39% of the streamflow estimates derived from the stream level class estimates (compared to 66% for Q_{factor}) were significantly overestimated (relative estimate >150%) or underestimated (relative estimate <50%). Furthermore, only 3% of the estimates were more than a

factor of 10 'off target' (compared to 11% for Q_{factor}). Even when taking the uncertainty in streamflow for the upper and lower stream level class boundaries into account (Fig. 7), the stream level class estimates resulted in streamflow values that were more accurate and had fewer outliers than those determined from the estimated width, mean depth and flow velocity.

Only for the small-sized streams was the interquartile range for streamflow calculated from stream level classes larger than the streamflow determined from the estimated width, depth and flow velocity (Fig. 6). When taking a closer look at the surveys for the different streams, it is clear that mainly the first survey at the Sihl and partly the survey at the Töss caused the large variation in the estimated streamflow from the stream level class data (see Supplementary material, Fig. S3).

[Figure 6 near here]

[Figure 7 near here]

3.5 Effect of stream size on streamflow and stream level class estimates

3.5.1 Streamflow

When estimating streamflow directly (Q_{direct}), participants made larger errors for the small streams (S; first to third quartile of relative estimates: 55–542%), than for the XS (19–112%), M (23–233%) and L (14–134%) streams. However, general statements on the effect of stream size on the accuracy of streamflow estimates are difficult to make because there were significant differences within each size class as well (Fig. 3).

The interquartile range of the Q_{factor} estimates was significantly smaller for the small (first to third quartile of relative estimates: 49–175%) and medium (27–117%) streams compared to Q_{direct} (Fig. 6). The Q_{factor} estimates were less accurate for XS (interquartile range: 47–293%) and L (17–226%) streams than for S and M streams. For the XS streams this difference is largely based on the estimates from Irchel, where direct streamflow estimates were more accurate than those derived from the estimated factors. For the Hornbach (another XS stream), there was no significant difference between the median relative estimates of Q_{direct} and Q_{factor} (for the Chriesbach there was no directly estimated streamflow data). The reasons for this different pattern in the Irchel are unknown, but could be due to the lower streamflow in the Irchel (0.01 m^3/s) compared to the Hornbach stream (0.13 m^3/s).

3.5.2 Stream level classes

Stream level class estimates were also analysed according to the distance between the participants and the virtual staff gauge, because the distance was not always related to the stream size. For the Limmat the virtual staff gauge was positioned on a bridge pillar rather than the opposite streambank (Fig. 1).

The stream level class estimates were generally more accurate if the staff gauge was closer to the observer (Fig. 5). For a distance of 0–10 m, 53% of participants selected the correct stream level class, while 35% selected a stream level that was only one class away. For a distance of 10–20 m, no-one selected a stream level class more than one class from the true value, and 73% of the participants selected the correct class, while for a distance of 20–30 m, 32% of participants were correct and 45% were one class away. For a distance of 50–60 m, 30% of participants chose the correct stream level class and 60% a neighbouring stream level class (Fig. 5(b)). This is not surprising, as, in cases where the virtual staff gauge is far away, it is more difficult to discern the stream level class and the reference, such as stones or other helpful objects, on the streambank.

3.6 High vs *low flow estimates*

One issue with hydrological data based on citizen science is the accuracy of the estimated streamflow, but another issue is whether changes in these estimates reflect differences in streamflow over time. Comparison of the estimated streamflow values for the Limmat, Sihl and Aare shows that the median estimated streamflow (Q_{factor}) was higher when the flow was higher, but the differences were not sufficiently higher to fully correspond to the increased streamflow (Fig. 8) and were not significant for the Aare (Fig. 8(b) and (c)). For the Limmat there were significant differences between the surveys, but these differences did not correspond fully to the measured values, as participants underestimated both high and low flow and the differences of estimates between the surveys are seemingly random regardless of high or low flow (Fig. 8(a)).

The variations in streamflow were better represented by the streamflow derived from the stream level class estimates (Q_{level}; Fig. 8(d)–(f)), for which the median estimated streamflow was indeed significantly higher when the flow was higher for seven out of eight surveys. The exception is the median streamflow for the survey on June 2017 at the Limmat for which the median estimated streamflow (Q_{level}) was not

significantly different from the median estimated streamflow during the July and April 2017 surveys, although the first and third quartiles were higher than for the July and April 2017 surveys (see Table 2 and Fig. 8(d)). The accuracy of the estimated variation in streamflow is therefore better represented by streamflow derived from stream level class estimates than by streamflow derived by the factors.

[Figure 8 near here]

[Table 2 near here]

4 Discussion

4.1 Can citizens estimate streamflow accurately?

The results of the streamflow estimation surveys demonstrated the "wisdom of the crowd" effect (Surowiecki 2004, Nielsen 2011) as the median estimates were close to the measured values. However, in practice there will be, at a certain location, only one or at most a few estimates for a certain point in time, so for hydrological citizen science projects focusing on streamflow the accuracy of the individual estimates is more important than the accuracy of the median estimate.

As expected, estimation of the individual streamflow factors (width, mean depth and flow velocity) led to more accurate streamflow estimates than the direct estimation of streamflow. The reduction in the number of extreme outliers for estimates based on the streamflow factors is likely due to the more intuitive units in which the estimates have to be given. For non-scientists the unit cubic metres per second (m^3/s) is difficult to visualize and not easy to relate to everyday experiences. Width and depth in metres (m) and flow velocity in metres per second (m/s) are easier to visualize and estimate for most people. The unit litres per second (L/s) is likely more tangible (as one knows the volume of a litre from drink containers and can estimate how long it takes to fill a bottle or a bucket). This might explain why, for the very small Irchel stream, direct streamflow estimates were more accurate than the streamflow derived from the estimated width, depth and velocity, which included the multiplication of three different types of error. For the Hornbach, another very small stream, there was no significant difference between Q_{direct} and Q_{factor}, possibly because it had more streamflow than can fit in a bucket in a second.

The direct streamflow estimates for the Aare (L) were also surprisingly accurate. After the survey, we learned that there used to be a digital display of the current streamflow at the FOEN gauging station, close to the location of our surveys. That display was dismantled before our survey, but it is possible that some participants walked by this site regularly and had a 'ballpark' value for the streamflow of the Aare in the back of their minds. Nevertheless, based on our dataset, estimating the streamflow factors rather than the streamflow directly is especially suitable for small and medium streams. It is, however, also important to note that, within the same stream size class, the accuracy of estimates varied for each stream, and even the accuracy of the estimates for the same stream location, can vary for different flow conditions (Figs 3 and 8). There was no clear pattern in the relative streamflow estimates (Q_{factor} or Q_{level}) to suggest that either low or high flows are more accurately estimated (see Fig. 8 and Table 2; also supplementary Fig. S4).

Many participants estimated the flow velocity fairly accurately if they threw a twig or leaf into the stream, as we suggested, or even just watched something like a bubble in the stream pass by. The differences between these approaches could not be quantified, as it was unfortunately not sufficiently documented who chose which approach.

Even though width and mean depth are measured in the same units, width could be estimated more accurately than mean depth. This is consistent with a study by Wahl (1977), in which trained participants measured both the width and depth of a stream, but measured width with more consistency than depth. In our case this is likely due to the refraction of light in water, as well as the inability to see the bottom of the stream because the water is murky or deep, which was the case for the Sihl at high flow (S), Limmat at high flow (M) and both surveys for the Aare (L). Also in some cases – Hornbach (XS), Irchel (XS), Glatt (S), Sihl (S), Töss (S) and Limmat (M) – it was feasible to pace the width along a bridge, in order to gain a better estimate, which made the width estimates more accurate; of course this could not be done for depth.

According to Gibson and Bergman (1954), distance estimation can be trained and constant over- and under-estimation of distances can be improved.

Training is implemented in many citizen science projects to ensure high quality data (Bonney et al. 2009, Haklay et al. 2010, See et al. 2013, Stepenuck and Genskow 2017). Participants in our survey received no training, had no prior experience and

(presumably) only estimated streamflow and its factors once. The effect of a one-time training was tested for some citizen science projects (Crall et al. 2013, Rinderer et al. 2015) and has been shown to improve the data-collection ability of the participants. Training options for our study could be in the form of online tutorial videos, or a list of well-known streams and their range in streamflow to indicate approximate numbers for streamflow, as well as width, depth and flow velocity. If participants can improve the accuracy of their estimates and the number of outliers can be reduced sufficiently, streamflow estimates might be usable for hydrological model calibration (Etter et al. 2018). Further research will test the applicability of quality control methods, such as outlier detection and the effect of training on the accuracy of streamflow estimates.

The inaccuracies of the streamflow estimates should be seen in light of the rating curve errors that are included in conventional measurements, which have a range of ±20% for medium to high flows and substantially higher errors ranging from –60% to +90% for low flows (McMillan et al. 2012). Only 29% and 63% of the Q_{direct} estimates were within ±20% and ±90% of the measured streamflow value. For the Q_{factor} estimates, the respective values were 15% and 73%.

Ensuring, and possibly improving, the accuracy of the crowdsourced data is an important aspect in any citizen science project. The inaccurate estimates of streamflow might be excluded from analyses by quality control methods. A comprehensive overview of data validation methods in the field of citizen science, such as expert review, photo submission or automatic filtering is provided by Wiggins et al. (2011) and many of these methods are likely also applicable to crowdsourced hydrological estimates.

Video imagery is an alternative way to estimate streamflow. These methods have great potential, especially for more accurately determining flow velocities (Bradley et al. 2002, Tsubaki et al. 2011, Lüthi et al. 2014, Le Coz et al. 2016, Tauro et al. 2018) and have benefits, such as being more objective and possibly allowing a higher accuracy than visual streamflow estimates. By using advanced and sophisticated technology, they also create a curiosity factor that can motivate people. However, there are also some limitations of these approaches in citizen science projects. Issues include light requirements, camera restrictions and the need for initial *in situ* channel measurements as a reference (Lüthi et al. 2014). To encourage more participants to join a citizen science project, we were interested to keep the ‘installation’ of new sites and

the observation approach as easy as possible. The visual estimates used in this study are easier to apply for many citizens and, thus, can potentially be used to provide more observations. The different methodologies complement each other and different methods might be most suitable for different locations, participant groups or observation goals. Tauro et al. (2018) express a similar opinion: "*Reconciling and complementing observations from such an abundant pool of methodologies, devices and platforms is the ultimate goal of the research community towards an improved understanding of hydrological processes.*" (Tauro et al. 2018, p.187). Many of the current limitations in video imagery will likely be resolved in the future, making this approach a more usable alternative for streamflow or stream level estimates. A possibility in the future might also be to develop a virtual staff gauge in an augmented reality setting, thereby facilitating participants' stream level class estimates.

4.2 Can citizens estimate stream level classes accurately?

Stream level classes were introduced to simplify the stream level estimation task for the participants. In theory we could have also asked participants to estimate a metric value above or below some fixed point. However, the depth estimates (Fig. 4) for Q_{factor} suggest that this approach would lead to estimates with a low accuracy. The high accuracy of stream level class estimates and small number of outliers (i.e. estimates that are more than one class off target) indicates that this is a suitable parameter for citizen science projects. The major benefits of the virtual staff gauge approach is that estimates can be done quickly and that relative variations can be estimated with small uncertainties, but, on the down side, they would also have a lower resolution. A participant can be no more than 10 classes off target (which never happened; 0.7% of participants were four classes off and <0.5% of participants were five or six classes off).

Participants only needed to compare the current stream level to a previous stream level using structures, streambanks or stones as a reference. If the virtual staff gauge is well placed (i.e. there is a suitable structure on the stream bank or in the stream), the participant only needs to look for the reference and then determine the corresponding stream level class. In general, the vast majority of participants had no problem understanding the concept and estimated the stream level class correctly; outliers in the estimated stream level classes were very rare. However, there were also a few clearly wrong stream level class estimates, which might suggest a misunderstanding

of the concept by some participants. The two most extreme overestimations were both at the Limmat, the most extreme underestimations at the Aare. Most participants (49%) underestimated the stream level class at the Aare. The reasons are unknown, but potentially this could be attributed to a staff gauge placement during an exceptionally low stream level (less than a 2-year low according to official measurements; BAFU 2017), meaning that the zero value was already very low. This might have confused participants as they may have thought that the staff gauge represents the average streamflow condition.

The stream levels class estimates were especially accurate for smaller streams where the opposite stream banks, at which the virtual staff gauges were located in the photo, were close to the participant. The Limmat is a wider stream, but was an exception as the virtual staff gauge was placed on a bridge pillar, which was relatively close to the observer. This is most likely the reason why the stream level class estimates for the Limmat were more accurate than for the Aare (the only stream where the references for the virtual staff gauge were 50–60 m away from the participant), even though the widths of the actual streams were similar (50 and 52 m, respectively). This shows that, for stream level class estimates, the placement of the virtual staff gauge is important. One of the very small streams (Irchel) had a poorly placed staff gauge (the image was taken looking down onto the stream rather than horizontally from the height of the stream level, which distorted the virtual staff gauge relative to the wall behind the stream) and made it more difficult to read. The median relative estimate for Q_{level} for the Irchel was 12%, whereas the median relative estimate for Q_{level} for all surveys was 101%.

Several studies have examined the accuracy of crowdsourced data (Haklay et al. 2010, Crall et al. 2011, See et al. 2013, Isaac and Pocock 2015, Tye et al. 2016, Aceves-Bueno et al. 2017, Mengersen et al. 2017), mentioning case studies such as OpenStreetMaps, where Volunteered Geographic Information (VGI) data are collected online and verified by other participants (Haklay et al. 2010), and discussing issues such as presence-only data for crowdsourced species classification (Isaac and Pocock 2015, Tye et al. 2016, Mengersen et al. 2017). While hydrological studies have also discussed crowdsourced data accuracy (Turner and Richter 2011, Rinderer et al. 2012, 2015, Lowry and Fienen 2013, Peckenham and Peckenham 2014, Breuer et al. 2015, Le Coz et al. 2016; Little et al. 2016, Weeser et al. 2018), most of these studies looked at

crowdsourced measurements rather than estimates (Lowry and Fienen 2013, Peckenham and Peckenham 2014, Little et al. 2016, Weeser et al. 2018). While others, such as Turner and Richter (2011), looked at class estimates, they mainly looked at two class options (wet or dry stream), but unfortunately do not mention data accuracy apart from the fact that participants were trained for consistency. Rinderer et al. (2012, 2015), who also looked at classed data, analysed participants' ability to estimate relative soil moisture classes and found that, in one case study, 95% of participants were no more than one class off (Rinderer et al. 2012), and in another study with various groups, 81–93% of the participants were no more than one class off (Rinderer et al. 2015). However, as far as we are aware, our study is the first to address the accuracy of participants' estimates of stream level classes.

In addition to being more accurate, the stream level class estimation process is also very quick, which is a big advantage for a citizen science project. It is hoped that offering a fast procedure to document stream levels will encourage citizen observers to contribute data to a project regularly (Eveleigh et al. 2014). It is very common for citizen science projects that the majority of the contributions come from a small group of high contributors (Lowry and Fienen 2013, Eveleigh et al. 2014, Sauermann and Franzoni 2015). For example, in the CrowdHydrology project, one participant walked past a particular station three to four times a week, which led to this station having almost 10 times as many measurements as the station with the next highest number of data submissions (Lowry and Fienen 2013). This highlights the extreme value of these high contributors and shows that it is important to be able to take measurements quickly.

4.3 Are citizens likely to observe variations in streamflow?

Having data for high and low flows, or relative variations in streamflow is crucial in order to determine how a stream reacts to precipitation, snowmelt events or long periods without rainfall, and for hydrological model calibration. Hence, it is important to know if crowdsourced data can properly reflect such variations in streamflow and whether the accuracy of the data depends on the flow conditions. The results from the surveys suggest that the temporal dynamics in streamflow will be relatively poorly represented by citizen-based streamflow estimates. For two of the three streams (Sihl and Aare), the median streamflow was overestimated at low flows and underestimated at high flows,

which indicates insufficient adjustment of the streamflow estimates to the variation in flow conditions. For the Limmat, the significant difference in the streamflow estimates does not seem to correspond to the differences in the measured streamflow (Fig. 8(a–c)). This is partly due to the problem that width (and to a lesser degree velocity) estimates were more accurate compared to depth estimates (Fig. 4). As long as a high flow stays within the streambank, the width of the streams in our survey does not vary significantly between low and high flows. Thus the majority of the variation in flow conditions is due to the variation in depth, which was most difficult to estimate.

During the surveys we did not ask the same persons to estimate the flow during high- and low-flow conditions. The results for an individual who reports the streamflow at different times may be different, because the participant might consistently over- or underestimate the flow and therefore the relative variations might be more accurate than indicated by our results (Rinderer et al. 2015). Thus further research is needed to determine if the streamflow dynamics are better described by the streamflow estimates when the majority of the contributions for a particular stream are made by one (or a few) active citizen(s) (Lowry and Fienen 2013).

The high- and low-flow patterns are better reflected in the stream level class estimates, with the median flow derived from these estimates (Q_{level}) being significantly different between high and low flows for all streams. For the Limmat, the *post hoc* tests showed a significant difference between the high flow and all other survey campaign estimates. This underlines the benefits of collecting stream level class estimates, particularly for model calibration (see additional discussion below).

4.4 Should citizen science projects focus on streamflow or stream level class estimates?

The reduction of the number of outliers in the streamflow estimates calculated from the stream level class data (Q_{level}) compared to the direct streamflow estimates (Q_{direct}) and streamflow estimates based on the streamflow factors (Q_{factor}) can partly be explained by the limited number of potential entries for the virtual staff gauge (i.e. participants can only choose one out of 10 available classes for the stream level estimate). For Q_{direct} and Q_{factor}, participants were able to state any value for their estimates, even values that are physically impossible for a particular stream. Hence, with regard to the reduction of outliers, estimating stream level classes seems advantageous for citizen science projects.

Additionally, our results suggest that stream level class estimates appear to be better suited to represent variations in flow conditions. Thus, the results of this study suggest that citizen science projects should focus on stream level class estimates instead of streamflow estimates, although this needs to be tested for different climatic, geographical and socio-economic settings.

However, it should be noted that part of the difference in accuracy for the stream level class estimates and streamflow estimates is due to the difference between relative and absolute values. For our approach, it would be impractical to use classes for streamflow estimates, as we would need many classes, or the resolution of the data would be very low (i.e. the flow for a given stream is likely to always be within the same class). However, as mentioned above, lists of well-known streams, giving their streamflow range to indicate orders of magnitude for the expected streamflow, as well as width, depth and flow velocity, could be provided to make it easier for citizens to make the estimates and to improve the accuracy of the estimates.

One of the disadvantages of the stream level classes is that each class represents a range of potential streamflow values, rather than one specific value. If a participant estimates that the stream level is in class two, it is unclear whether that means the upper, middle or lower part of the class. The other disadvantage is that these estimates do not provide information on streamflow volumes. However, the usability of stream level class data for hydrological model calibration was tested by van Meerveld et al. (2017), who showed that stream level class data can be used to calibrate a simple bucket type hydrological model, and suggested that simple hydrological models can be used to convert stream level class data to time series of streamflow. The value of stream level data for hydrological model calibration, especially for humid catchments, was demonstrated recently by Seibert and Vis (2016). The value of crowdsourced stream level data (photographs of a fixed staff gauge) together with rainfall and flood observations was also shown by Starkey et al. (2017). They used community-based observations of rainfall (manual raingauges), river levels (manual staff gauge) and flood-related evidence (anecdotes, photographs or videos) alongside traditional information (tipping bucket raingauge, official raingauge measurements, six pressure transducers for water level measurements and flow gauging for the discharge-rating curve), in order to fill spatial and temporal gaps in hydrometric data for a 42 km^2-catchment in the UK to improve a physically-based, spatially-distributed catchment

model (SHETRAN). Etter et al. (2018) calibrated a bucket type model with synthetic crowdsourced streamflow data with different degrees of error (including errors that are comparable to those observed in this study) and varying temporal resolutions, and indeed found that such streamflow estimates do not contain sufficient information to improve the model compared to random parameter sets. However, they also showed that, if the standard deviation of the log-normal distribution that was used to describe the errors of crowdsourced streamflow estimates could be reduced by a factor of two, one estimate per week would lead to a significant improvement in the model simulations.

5 Conclusion

We asked 517 citizens to estimate streamflow directly and indirectly by estimating the stream width, depth and flow velocity. We also asked them to estimate the stream level class. The survey results allowed us to quantify the accuracy of the estimates and is, thus, a basis for evaluating the potential value of citizen science based estimates of streamflow and stream level classes. The median estimated streamflow values were close to the measured streamflow, but there were also many outliers, and the variations in the flow conditions were not fully discernible in the streamflow estimates. The stream level class estimates, which were converted into streamflow values for comparison, had far fewer outliers and were significantly different for the different flow conditions. Stream level class estimates also seemed to be quicker and easier to estimate and are thus considered preferable for citizen science approaches. Hydrological models can then be parameterized based on these stream level class observations to obtain streamflow time series. The study was conducted in Switzerland and, while we do not expect significant differences, we recommend testing the accuracy of citizen science based estimates of streamflow and stream level classes in different climatic, geographical or socio-economic settings and for rivers with different sizes.

Acknowledgements

We thank all study participants for their time and interest in this research project and for sharing their hydrological estimates with us, as well as the FOEN (Federal Office for the Environment) and WWEA (Office of Waste, Water, Energy and Air of Canton Zurich) for providing the streamflow data used for comparison with the estimates.

Funding

This study was funded by the Swiss National Science Foundation (project 163008, CrowdWater).

References

Aceves-Bueno, E. et al., 2017. The Accuracy of Citizen Science Data : A Quantitative Review. *The Bulletin of the Ecological Society of America*, 98(4), pp.278–290.

BAFU, 2017. *Niedrigwasserwahrscheinlichkeit (Jahresniedrigwasser NM7Q) Aare-Brugg (EDV: 2016)*, Available at: https://www.hydrodaten.admin.ch/lhg/sdi/nq_studien/nq_statistics/2016nq.pdf [Accessed 2 May 2018].

Beven, K. & Westerberg, I., 2011. On red herrings and real herrings: disinformation and information in hydrological inference. *Hydrological Processes*, 25(10), pp.1676–1680.

Beven, K.J., 2012. *Rainfall-Runoff Modelling: The Primer* 2nd ed., Wiley-Blackwell.

Bonney, R. et al., 2009. Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. *BioScience*, 59(11), pp.977–984.

Bradley, A.A. et al., 2002. Flow measurement in streams using video imagery. *Water Resources Research*, 38(12), pp.51-1-51–8.

Breuer, L. et al., 2015. HydroCrowd: a citizen science snapshot to assess the spatial control of nitrogen solutes in surface waters. *Scientific Reports*, 5(16503).

Buytaert, W. et al., 2014. Citizen science in hydrology and water resources: opportunities for knowledge generation, ecosystem service management, and sustainable development. *Frontiers in Earth Science*, 2(26), p.21. Available at: http://journal.frontiersin.org/article/10.3389/feart.2014.00026/abstract

Le Coz, J. et al., 2016. Crowdsourced data for flood hydrology : feedback from recent citizen science projects in Argentina , France and New Zealand. *Journal of Hydrology*, 541, pp.766–777.

Crall, A.W. et al., 2011. Assessing citizen science data quality: An invasive species case study. *Conservation Letters*, 4(6), pp.433–442.

Crall, A.W. et al., 2013. The impacts of an invasive species citizen science training program on participant attitudes, behavior, and science literacy. *Public Understanding of Science*, 22(6), pp.745–764.

Dickinson, J.L., Zuckerberg, B. & Bonter, D.N., 2010. Citizen Science as an Ecological Research Tool: Challenges and Benefits. *Annual Review of Ecology, Evolution, and Systematics*, 41(1), pp.149–172.

Dingman, S.L., 2015. *Physical Hydrology* 3rd ed., Long Grove: Waveland Press, Inc.

Dunn, O.J., 1964. Multiple Comparisons Using Rank Sums. *Technometrics*, 6(3), pp.241–252.

Engel, S.R. & Voshell, J.R., 2002. Volunteer biological monitoring: can it accurately assess the ecological condition of streams? *American Entomologist*, 48, pp.164–177.

Etter, S. et al., 2018. Value of uncertain streamflow observations for hydrological modelling. *Hydrology and Earth System Sciences*, 22, pp.5243–5257.

Eveleigh, A. et al., 2014. Designing for dabblers and deterring drop-outs in citizen science. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*. pp. 2985–2994.

Field, A., Miles, J. & Field, Z., 2013. *Discovering statistics using R*, Los Angeles: Sage.

Gibson, E.J. & Bergman, R., 1954. The effect of training on absolute estimation of distance over the ground. *Journal of Experimental Psychology*, 48(6), pp.473–482.

Hadj-Hammou, J. et al., 2017. Getting the full picture: Assessing the complementarity of citizen science and agency monitoring data. *PLoS ONE*, 12(12), pp.1–18.

Haklay, M., 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environment and Planning B: Planning and Design*, 37(4), pp.682–703.

Haklay, M. (Muki) et al., 2010. How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information. *The Cartographic Journal*, 47(4), pp.315–322.

Herschy, R.W., 1971. *The magnitude of errors at flow measurement stations*, Technical report, Water Resources Board, Reading, UK.

Isaac, N.J.B. & Pocock, M.J.O., 2015. Bias and information in biological records. *Biological Journal of the Linnean Society*, 115(3), pp.522–531.

Juston, J., Seibert, J. & Johansson, P., 2009. Temporal sampling strategies and uncertainty in calibrating a conceptual hydrological model for a small boreal catchment. *Hydrological Processes*, 23(21), pp.3093–3109. Available at: http://doi.wiley.com/10.1002/hyp.7421.

Kundzewicz, Z.W., 1997. Water resources for sustainable development. *Hydrological Sciences Journal*, 42(4), pp.467–480.

Little, K.E., Hayashi, M. & Liang, S., 2016. Community-Based Groundwater Monitoring Network Using a Citizen-Science Approach. *Groundwater*, 54(3), pp.317–324.

Lowry, C.S. & Fienen, M.N., 2013. CrowdHydrology: Crowdsourcing Hydrologic Data and Engaging Citizen Scientists. *Ground Water*, 51(1), pp.151–156. Available at: http://doi.wiley.com/10.1111/j.1745-6584.2012.00956.x.

Lüthi, B., Philippe, T. & Peña-Haro, S., 2014. Mobile device app for small open-channel flow measurement. In D. P. Ames, N. W. T. Quinn, & A. E. Rizzoli, eds. *7th International Congress on Environmental Modelling and Software*. San Diego, CA, USA. Available at: http://www.iemss.org/sites/iemss2014/papers/iemss2014_submission_112.pdf.

Manning, R., 1891. On the flow of water in open channels and pipes. *Transactions of the Institution of Civil Engineers of Ireland*, 20, pp.161–207.

McMillan, H., Krueger, T. & Freer, J., 2012. Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality. *Hydrological Processes*,

26(26), pp.4078–4111.

van Meerveld, H.J., Vis, M.J.P. & Seibert, J., 2017. Information content of stream level class data for hydrological model calibration. *Hydrology and Earth System Sciences*, 21(9), pp.4895–4905.

Mengersen, K. et al., 2017. Modelling imperfect presence data obtained by citizen science. *Environmetrics*, 28(5), pp.1–29.

Nielsen, M., 2011. *Reinventing Discovery: The New Era of Networked Science*, Princeton University Press.

Peckenham, J.M. & Peckenham, S.K., 2014. Assessment of quality for middle level and high school student-generated water quality data. *Journal of the American Water Resources Association*, 50(6), pp.1477–1487.

Pelletier, P.M., 1988. Uncertainties in the single determination of river discharge: a literature review. *Canadian Journal of Civil Engineering*, 15(5), pp.834–850. Available at: http://www.nrcresearchpress.com/doi/10.1139/l88-109.

Perrin, C. et al., 2007. Impact of limited streamflow data on the efficiency and the parameters of rainfall—runoff models. *Hydrological Sciences Journal*, 52(1), pp.131–151. Available at: http://www.tandfonline.com/doi/abs/10.1623/hysj.52.1.131.

Pool, S., Viviroli, D. & Seibert, J., 2017. Prediction of hydrographs and flow-duration curves in almost ungauged catchments: Which runoff measurements are most informative for model calibration? *Journal of Hydrology*, 554, pp.613–622.

Rinderer, M. et al., 2015. Qualitative soil moisture assessment in semi-arid Africa - the role of experience and training on inter-rater reliability. *Hydrology and Earth System Sciences*, 19, pp.3505–3516.

Rinderer, M. et al., 2012. Sensing with boots and trousers - qualitative field observations of shallow soil moisture patterns. *Hydrological Processes*, 26(26), pp.4112–4120. Available at: http://doi.wiley.com/10.1002/hyp.9531 [Accessed March 27, 2014].

Ruhi, A., Messager, M.L. & Olden, J.D., 2018. Tracking the pulse of the Earth's fresh waters. *Nature Sustainability*, 1(4), pp.198–203.

Sauermann, H. & Franzoni, C., 2015. Crowd science user contribution patterns and their implications. *Proceedings of the National Academy of Sciences*, 112(3), pp.679–684.

See, L. et al., 2013. Comparing the Quality of Crowdsourced Data Contributed by Expert and Non-Experts. , 8(7), pp.1–11.

Seibert, J. & Beven, K.J., 2009. Gauging the ungauged basin: how many discharge measurements are needed? *Hydrology and Earth System Sciences*, 13(6), pp.883–892. Available at: http://www.hydrol-earth-syst-sci.net/13/883/2009/.

Seibert, J. & McDonnell, J.J., 2015. Gauging the Ungauged Basin : Relative Value of Soft and Hard Data. *Journal of Hydrologic Engineering*, 20(1), pp.A4014004-1–6. Available at: https://ascelibrary.org/doi/10.1061/%28ASCE%29HE.1943-5584.0000861.

Seibert, J. & Vis, M.J.P., 2016. How informative are stream level observations in different geographic regions? *Hydrological Processes*, 30(14), pp.2498–2508.

Starkey, E. et al., 2017. Demonstrating the value of community-based ('citizen science') observations for catchment modelling and characterisation. *Journal of Hydrology*, 548, pp.801–817.

Statistik Stadt Zürich, 2017. Statistisches Jahrbuch der Stadt Zürich 2017. , pp.188–201. Available at: https://www.stadt-zuerich.ch/prd/de/index/statistik/publikationen-angebote/publikationen/Jahrbuch/statistisches-jahrbuch-der-stadt-zuerich_2017.html [Accessed October 12, 2017].

Stepenuck, K.F. & Genskow, K.D., 2017. Characterizing the Breadth and Depth of Volunteer Water Monitoring Programs in the United States. *Environmental Management*, 61(1), pp.46–57.

Surowiecki, J., 2004. *The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations.*,

London, UK: Little Brown.

Tauro, F. et al., 2018. Measurements and observations in the XXI century (MOXXI): Innovation and multi-disciplinarity to sense the hydrological cycle. *Hydrological Sciences Journal*, 63(2), pp.169–196. Available at: https://doi.org/10.1080/02626667.2017.1420191.

Tsubaki, R., Fujita, I. & Tsutsumi, S., 2011. Measurement of the flood discharge of a small-sized river using an existing digital video recording system. *Journal of Hydro-Environment Research*, 5(4), pp.313–321.

Tulloch, A.I.T. et al., 2013. Realising the full potential of citizen science monitoring programs. *Biological Conservation*, 165, pp.128–138.

Turner, D.S. & Richter, H.E., 2011. Wet/dry mapping: using citizen scientists to monitor the extent of perennial surface flow in dryland regions. *Environmental Management*, 47(3), pp.497–505.

Tye, C.A. et al., 2016. Evaluating citizen versus professional data for modelling distributions of a rare squirrel. *Journal of Applied Ecology*, pp.1–10.

Vis, M. et al., 2015. Model calibration criteria for estimating ecological flow characteristics. *Water (Switzerland)*, 7(5), pp.2358–2381.

Wahl, K.L., 1977. Accuarcy of channel measurements and the implications in estimating streamflow characteristics. *Journal Research of the U.S. Geological Survey*, 5(6), pp.811–814.

Weeser, B. et al., 2018. Citizen science pioneers in Kenya – A crowdsourced approach for hydrological monitoring. *Science of The Total Environment*, 632, pp.1590–1599.

Welber, M. et al., 2016. Field assessment of noncontact stream gauging using portable surface velocity radats (SVR). *Water Resources Research*, 52, pp.1108–1126.

Westerberg, I. et al., 2011. Stage-discharge uncertainty derived with a non-stationary rating curve in the Choluteca River, Honduras. *Hydrological Processes*, 25(4), pp.603–613.

Wiggins, A. et al., 2011. Mechanisms for data quality and validation in citizen science. In *Seventh IEEE International Conference on e-Science Workshops*. pp. 14–19.

Table 1. Information on the streams where the field surveys took place. Size classes XS: ≤ 1 m^3/s; S: >1–50 m^3/s, M: >50–200 m^3/s and L: >200 m^3/s. A map with the survey locations is given in the Supplementary material (Fig. S1). Survey dates given as dd.mm.yyyy.

Stream	Size		Date of survey	No. of partici-pants, *n*	Stream-flow (m^3/s)	Source for measured streamflow*	Approx. distance to virtual staff gauge (m)	Comments
Chriesbach (Zurich)	XS		29.09.2017	30	0.38	Salt dilution	5	BSc students: no direct streamflow estimates
Hornbach (Zurich)	XS		19.02.2017	33	0.134	Salt dilution	8	
Irchel (Zurich)	XS		11.03.2017	25	0.01	Salt dilution	1	
Glatt (Zurich)	S		29.09.2017	31	2.8	WWEA, station: 533	11	BSc students: no direct streamflow estimates
Magliasina (Magliaso)	S		28.04.2017	40	16	FOEN, station: 2461	14	High school students: no stream level class estimates
Schanzen-graben (Zurich)	S		01.04.2017	31	2.6	Salt dilution	16	
Sihl (Zurich)	S	1	18.02.2017	33	7	FOEN, station: 2176	32	Low flow
		2	26.07.2017	31	28			High flow
Töss (Winterthur)	S		12.03.2017	35	9	WWEA, stations: 518, 520 and 581	29	Interpolation between three nearby stations for reference value
Limmat (Zurich)	M	1	29.10.2016	38	59	FOEN, station: 2099	7	No stream level class estimates
		2	08.04.2017	27	83			
		3	02.06.2017	31	107			
		4	09.07.2017	44	75			PhD students Low flow
		5	13.11.2017	31	222			High flow
Aare (Brugg)	L	1	07.01.2017	27	108	FOEN, station: 2016	53	Low flow
		2	10.05.2017	30	389			High flow

* The measured streamflow data were obtained from: the Federal Office of the Environment (FOEN; http://hydrodaten.admin.ch/), the Office of Waste, Water, Energy and Air of Canton Zurich (WWEA; www.hydrometrie.zh.ch/), or by salt dilution gauging (Salt dilution).

Table 2. Descriptive statistics of the streamflow derived from the estimated width, mean depth and flow velocity (Q_{factor}; m^3/s) (and relative estimate, %) and the stream level classes for the Aare, Limmat and Sihl under different flow conditions.

Stream	Date	Streamflow, Q_{factor} (m^3/s) (relative Q_{factor}, %)				Stream level class			
		Measured	Quartile			Measured	Quartile		
			25%	50%	75%		25%	50%	75%
Sihl	18.02.2018	7 (100)	5 (66)	9 (127)	26 (365)	0	0	1	1
	26.07.2018	28 (100)	11 (39)	21 (76)	46 (163)	1	2	2	3
Limmat	29.10.2016	59 (100)	31 (53)	48 (81)	86 (146)				
	08.04.2017	83 (100)	22 (27)	60 (73)	111 (134)	–2	–2	–1	–1
	02.06.2017	107 (100)	26 (24)	54 (51)	78 72)	–1	–1	–1	0
	09.07.2017	75 (100)	9 (12)	32 (42)	49 (66)	–2	–2	–1	–1
	13.11.2017	222 (100)	53 (24)	120 (54)	296 (133)	1	1	1	2
Aare	07.01.2017	108 (100)	47 (44)	128 (118)	404 (374)	0	–1	0	1
	10.05.2017	389 (100)	51 (13)	182 (47)	684 (176)	4	3	3	4

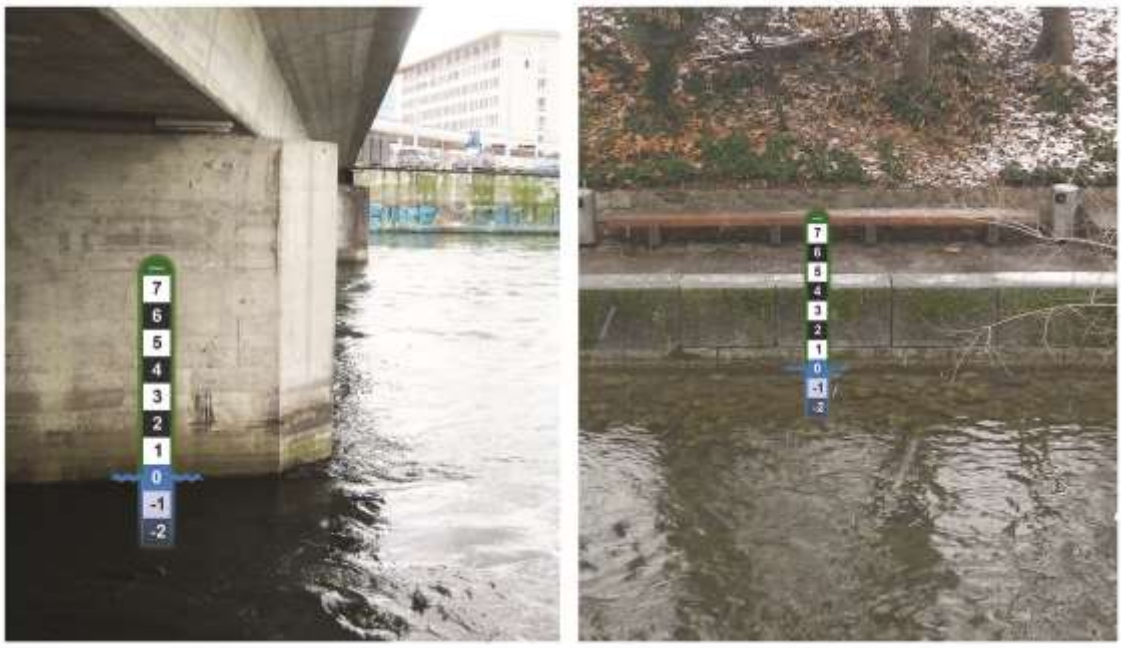

Figure 1. Example of a virtual staff gauge in the pictures used for the field survey at Limmat (left) and Schanzengraben (right). Photographs taken on 29.06.2016 when the streamflow was 165m^3/s (Limmat) and on 05.01.2017 (unknown streamflow; Schanzengraben). For the dates and the flow conditions during the surveys see Table 1.

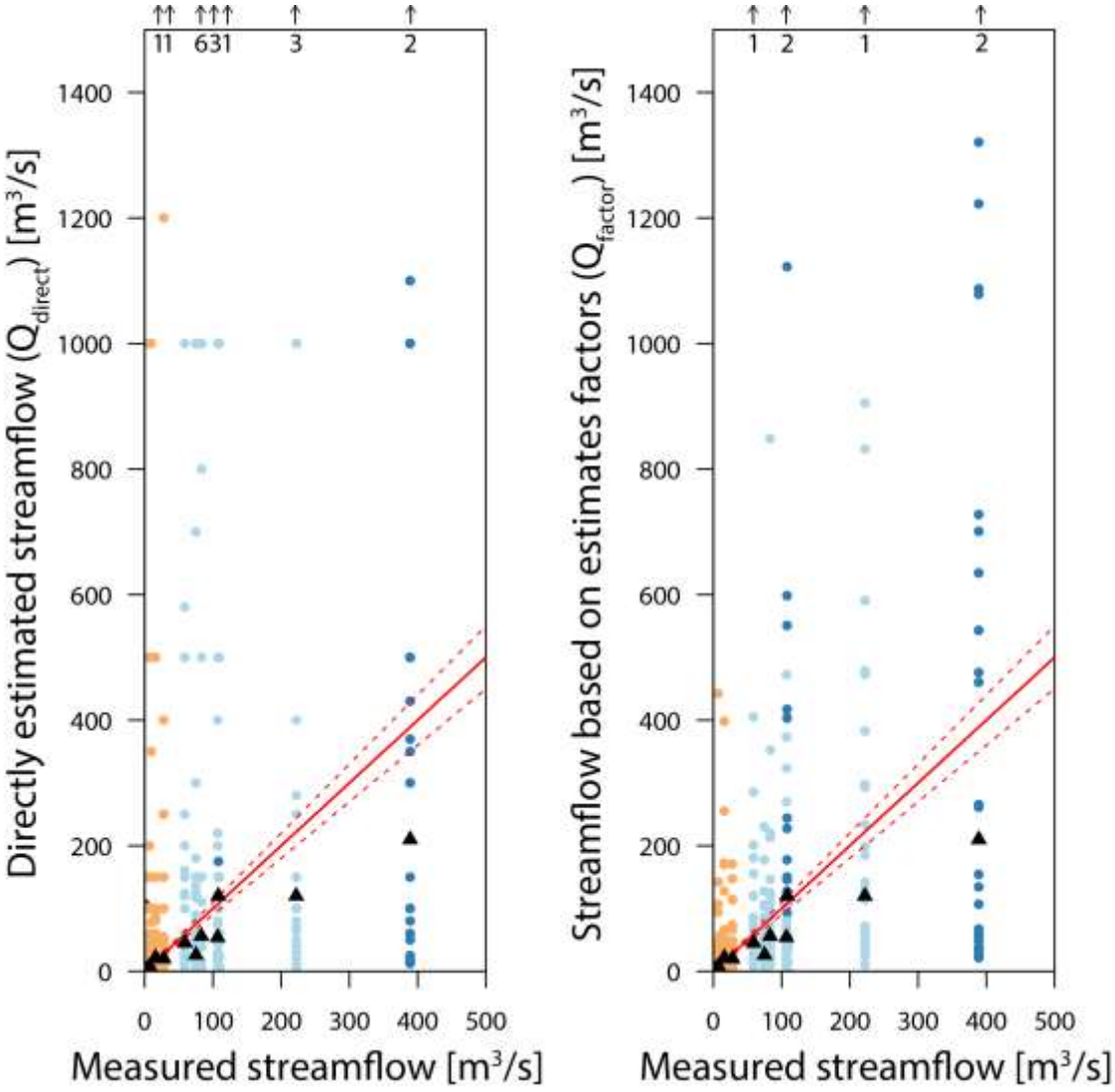

Figure 2. Scatterplots showing the spread of Q_{direct} (left) and Q_{factor} (right) for each field survey. The data points are colour-coded according to the stream size: from left to right, XS to L are red, orange, light blue and dark blue, respectively. ▲: median estimated streamflow per survey; solid (red) line: the 1:1 line; and dashed (red) lines: the 10% uncertainty band for measured streamflow. The number at the top of the graph indicates the number of extreme outliers (1–6, not shown).

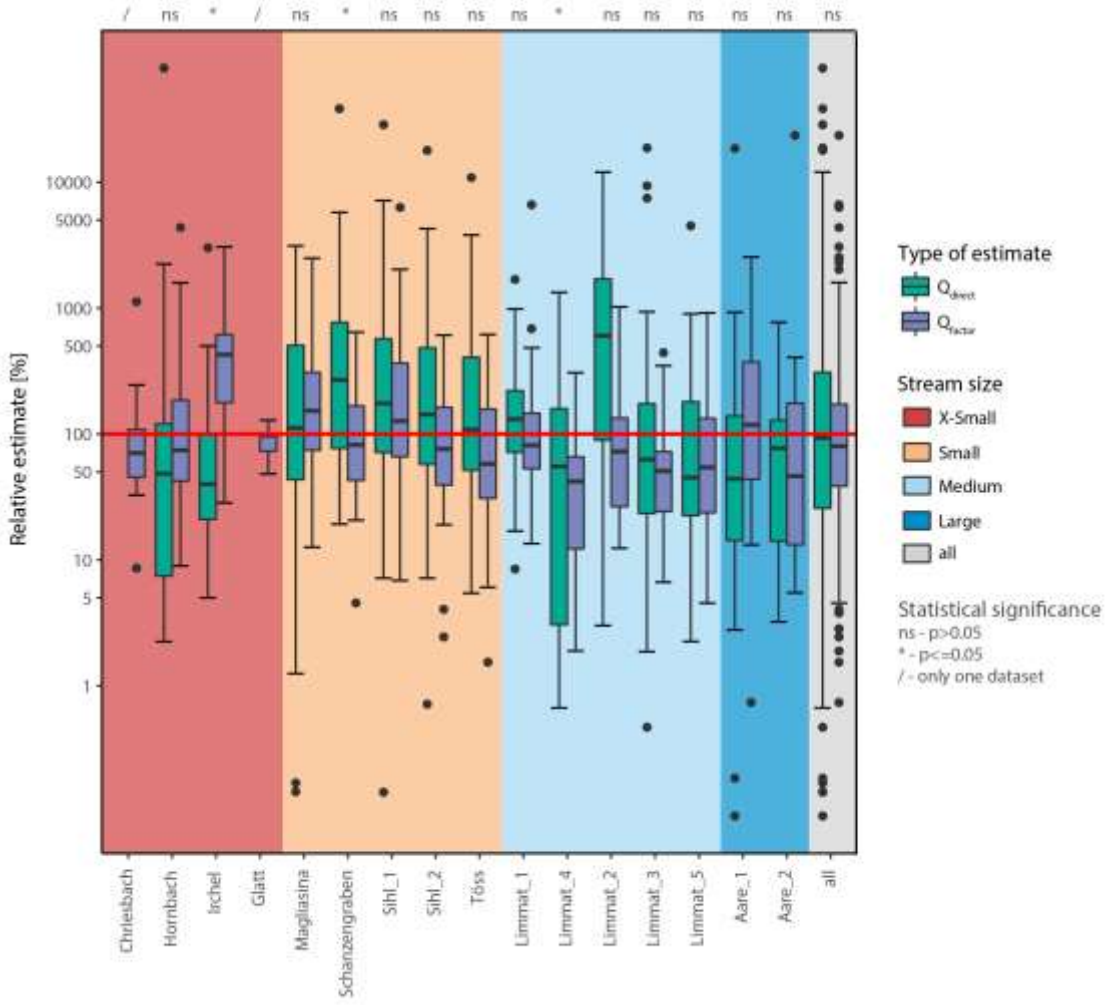

Figure 3. Boxplots of the relative estimates of streamflow (ratio of estimated *vs* measured streamflow) for Q_{direct} and Q_{factor} for each surveyed stream, and for all streams combined (all). Statistical significance, i.e. difference in median relative streamflow estimate for the two methods, is shown across the top. The data for the Sihl, Limmat and Aare are ordered from low- to high-flow conditions (see Table 1). The box represents the interquartile range, the black line the median, the whiskers extend to 1.5-times the interquartile range below/above the first/ third quartile, and the dots represent values beyond 1.5-times the interquartile range. Note the log scale.

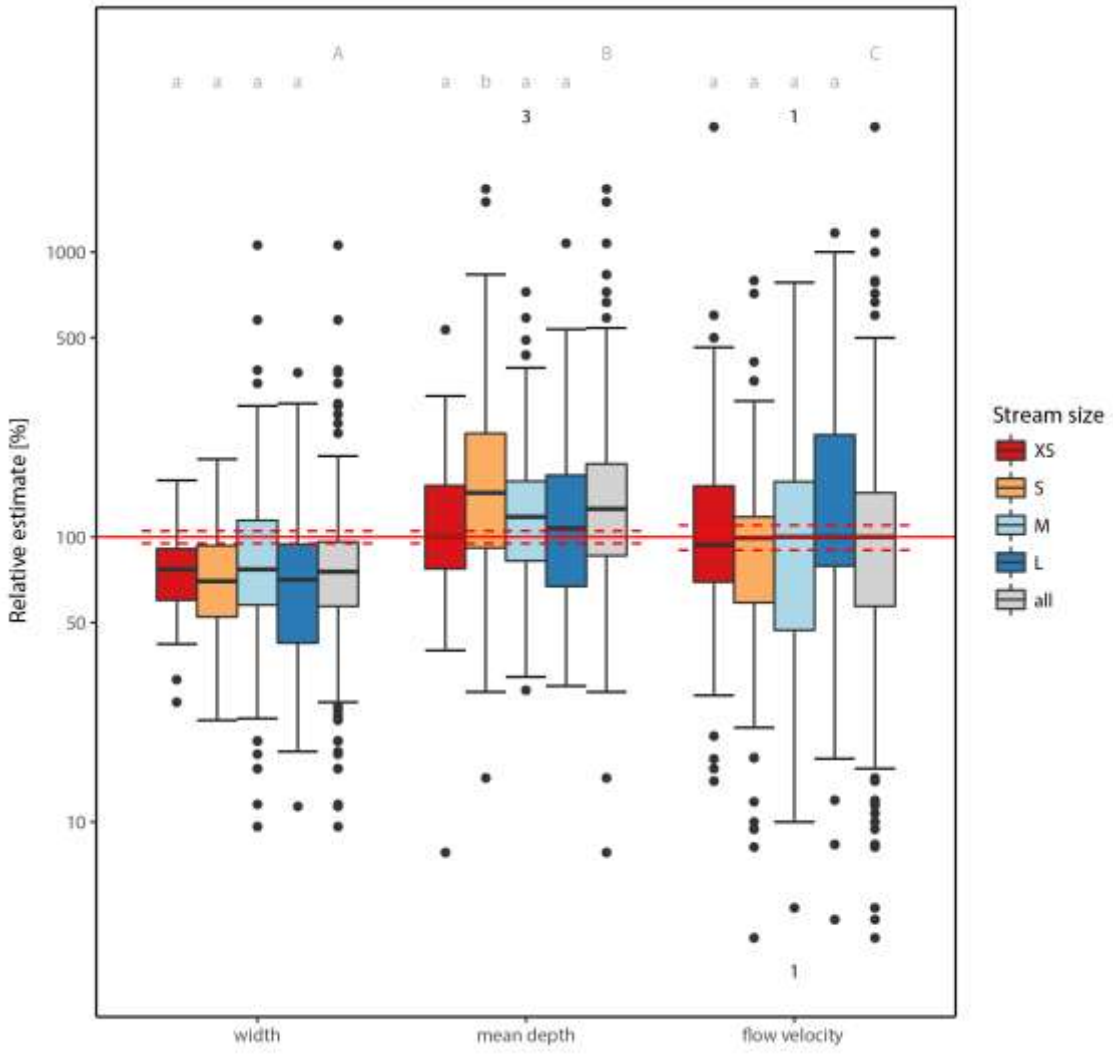

Figure 4. Boxplots of the relative estimates of width, mean depth and flow velocity for each stream size class and all streams together. Median relative estimates of width, mean depth and flow velocity, individually (lower case letters) and for data from all surveys combined (upper case letters) indicate that they are all significantly different. The solid red line (100%) indicates that the estimate is the same as the measured value; dashed red lines indicate the 5% (width and mean depth) and 10% (flow velocity) uncertainty bands. The numbers above and below the boxplots indicate the number of outliers not shown. Note the log scale.

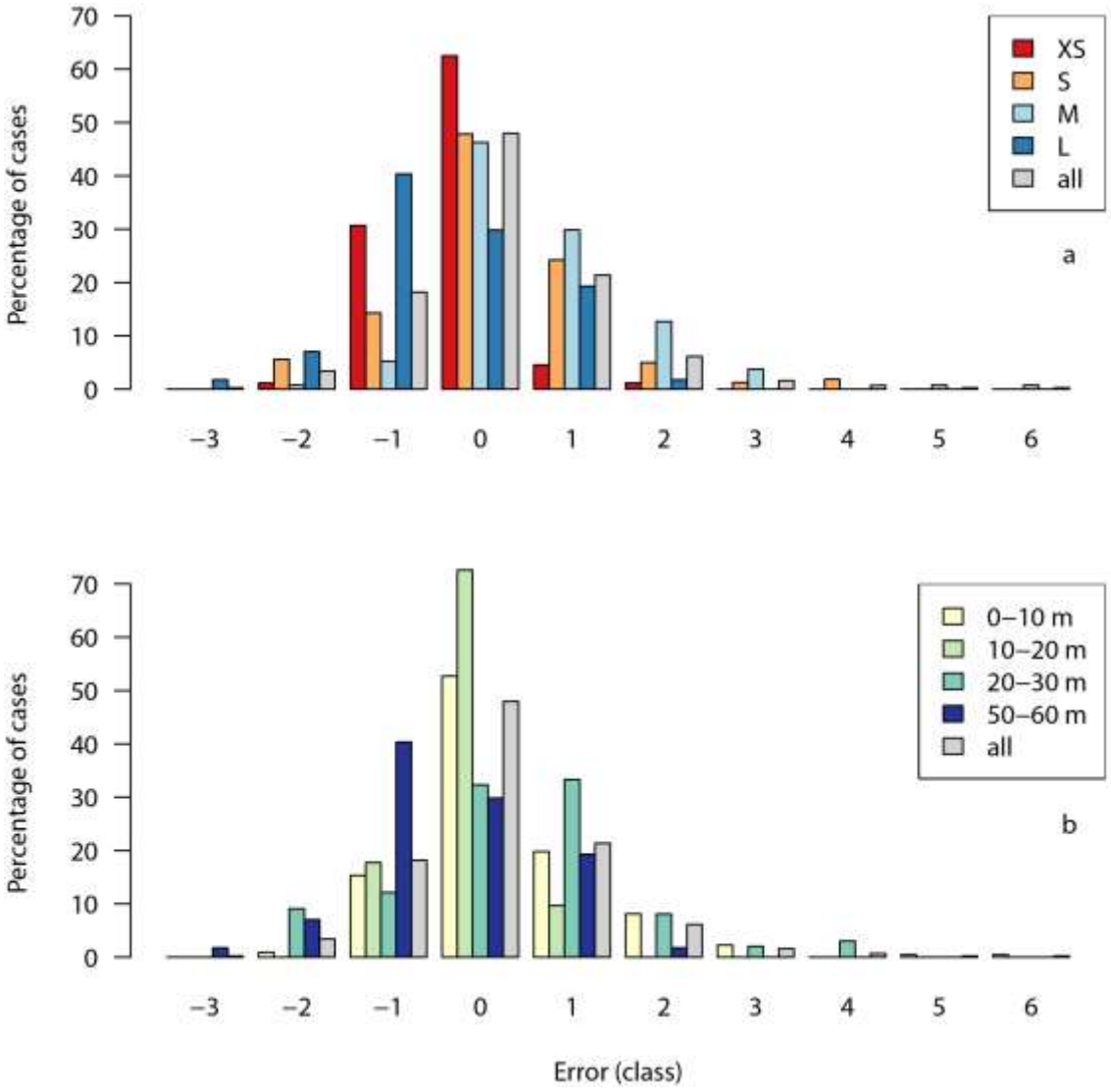

Figure 5. (a) Distribution of error in stream level class estimates (0: no error, –1: one class lower than the actual stream level class, and 1: one class higher than the actual class) for streams of different sizes; and (b) the distance between participant and the virtual staff gauge, as well as all estimates together. There were no streams where the virtual staff gauge was 30–50 m away from the participants.

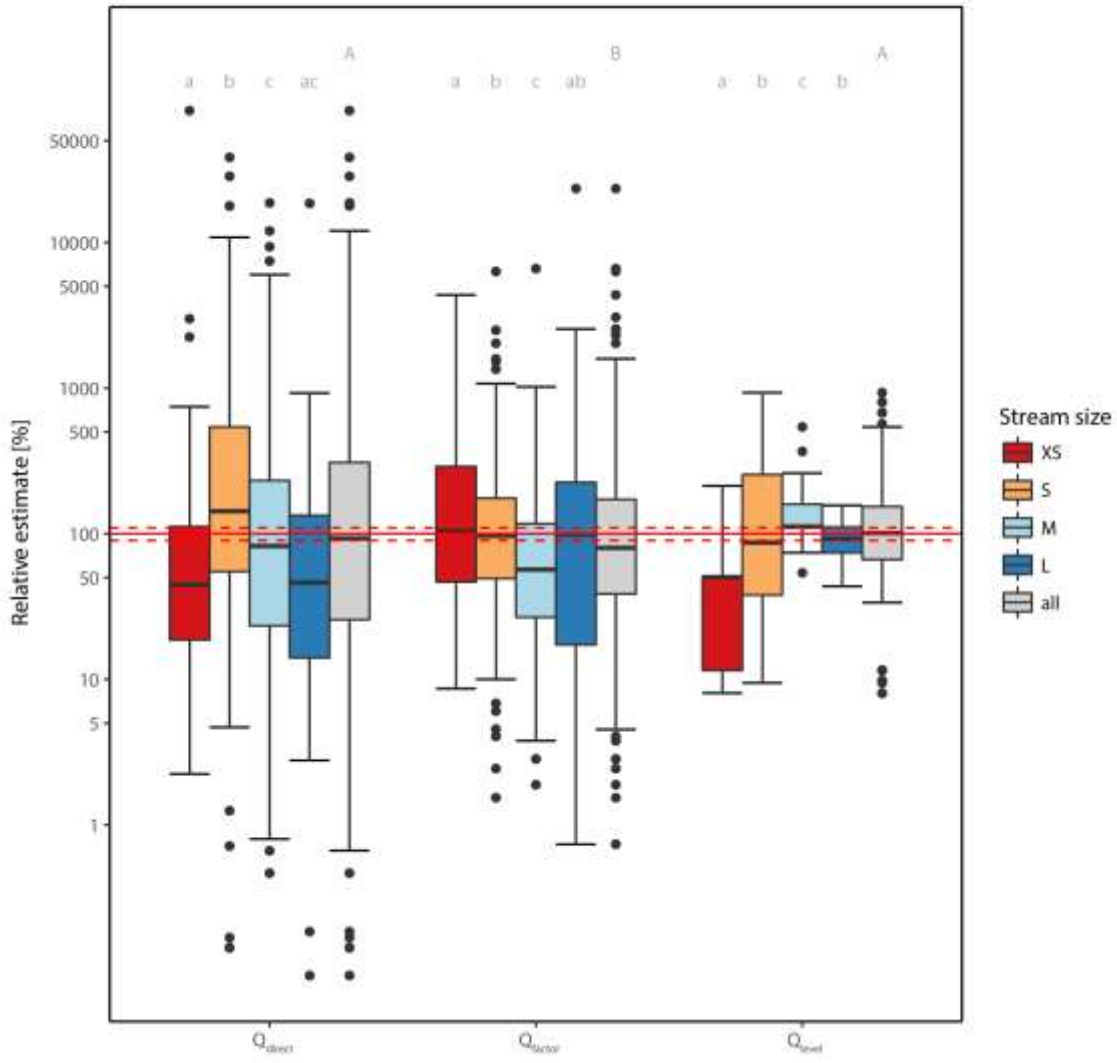

Figure 6. Boxplot of the relative estimates of Q_{direct}, Q_{factor} and Q_{level} for each stream size class and all surveys combined. Median relative estimates of Q_{direct}, Q_{factor} and Q_{level}, individually (lower case letters) and for combined data from all surveys (upper case letters) indicate that they are all significantly different. The solid (red) line at 100% indicates that the estimate is the same as the measured value and the dashed (red) lines indicate the 10% uncertainty band for the measured streamflow.

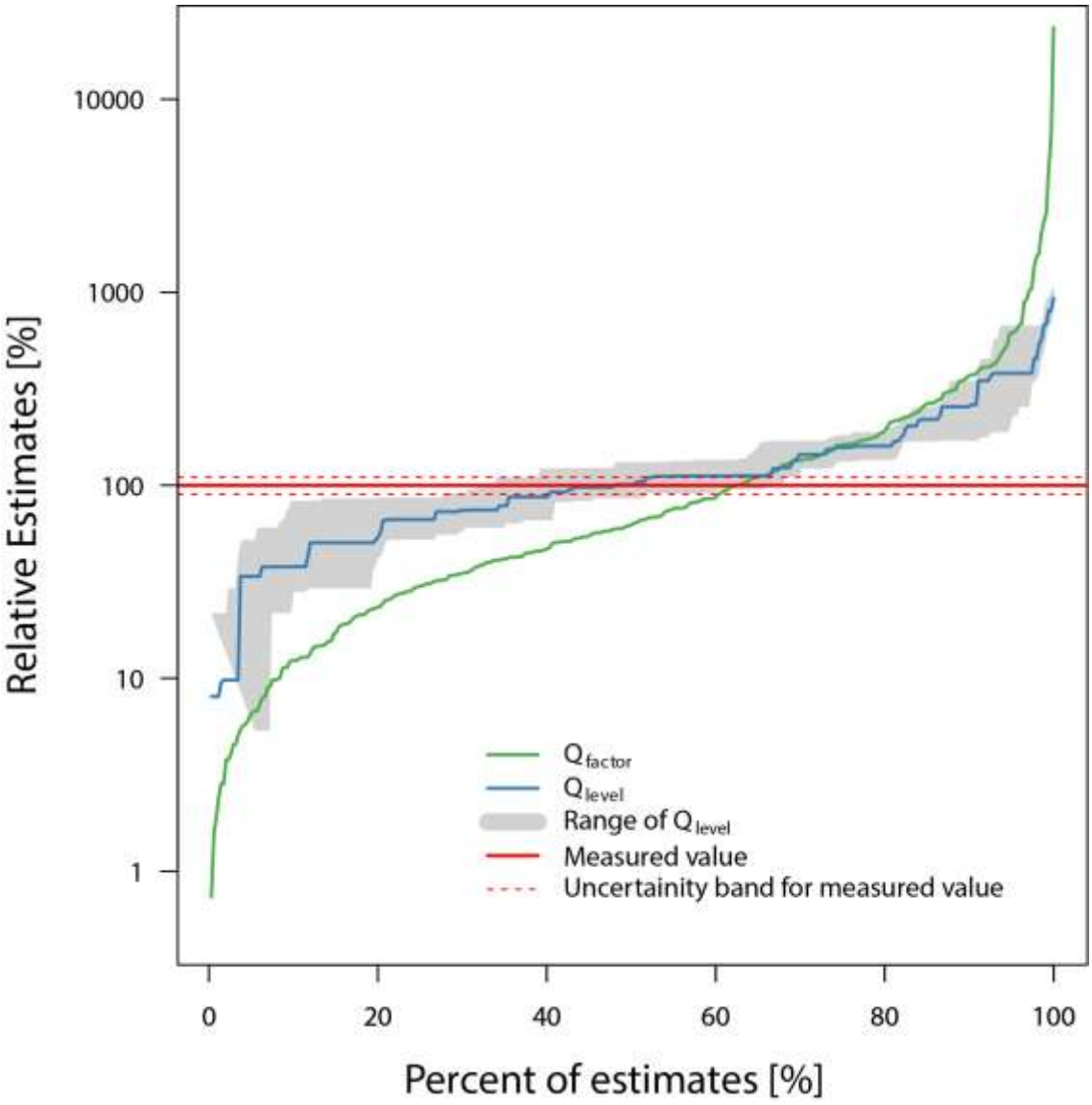

Figure 7. Frequency distribution of the relative streamflow estimate for Q_{factor} and Q_{level}. The shaded (grey) band indicates the upper and lower streamflow for each stream level class. The lower streamflow for each stream level class does not reach the 0% mark, as there were 18 zero-values, which cannot be displayed on a log scale.

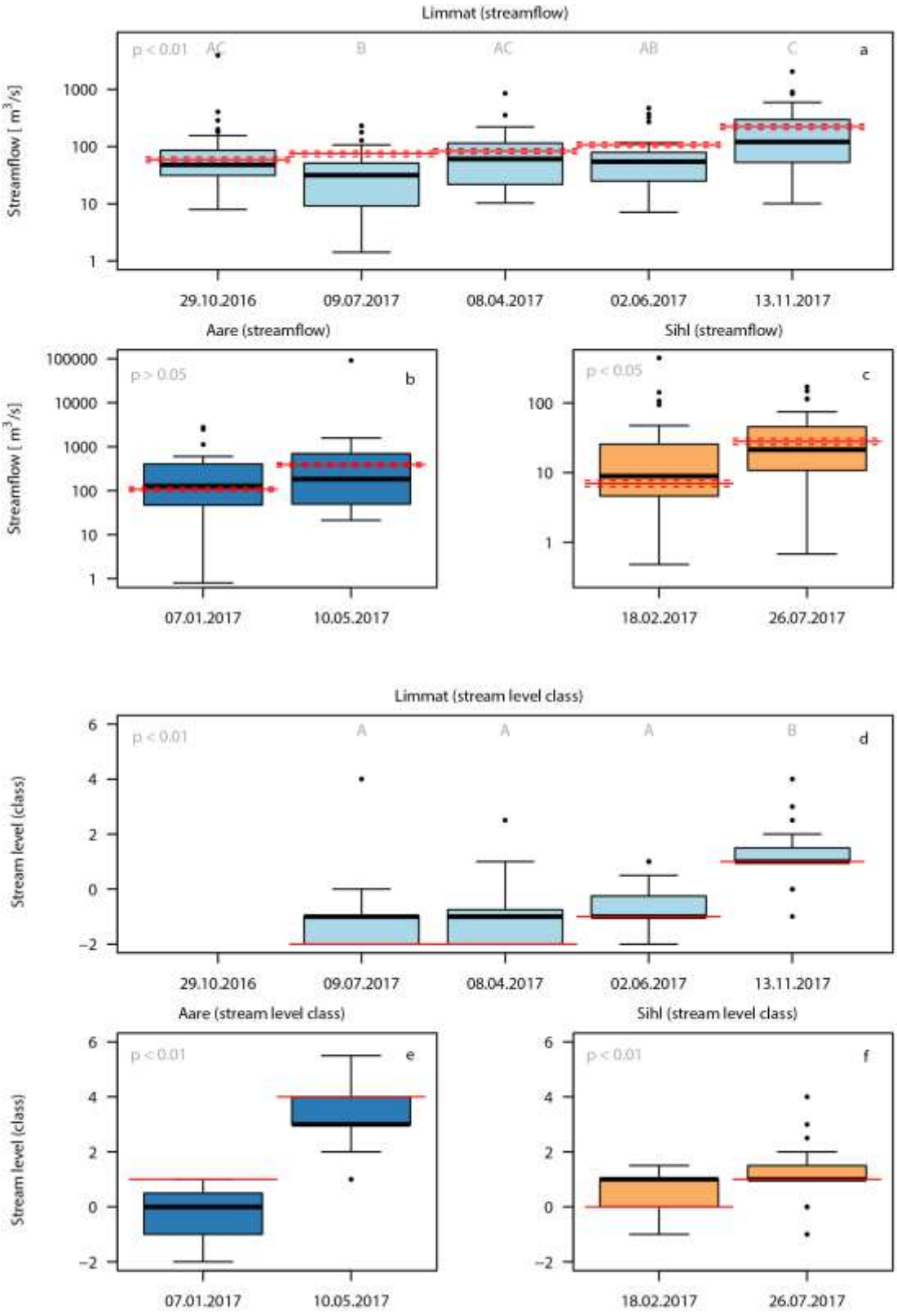

Figure 8. Boxplots of the streamflow based on (a–c) Q_{factor} and (d–f) the estimated stream level classes for different flow conditions for three streams (low flow to high flow in each subplot; see Table 1 for details). Solid and dashed (red) lines as described in Fig. 6 caption. Note: the axis ranges are different for each stream. The *p* values

indicate the results of the Mann-Whitney (Sihl and Aare) and Kruskal-Wallis (Limmat) tests to determine whether the median estimated streamflow/stream level class estimates of the different surveys are significantly different or not. For the Limmat surveys with the same upper case letter (e.g. A) the Dunn *post hoc* test showed that median streamflow/stream level class estimates are not significantly different from each other.