Replies to Editor

Editor Decision: Publish subject to minor revisions (further review by editor) (10 Sep 2018) by Nadav Peleg

Comments to the Author:
Dear Simon Etter and co-authors,

Thank you for posting your responses to the two referees' reports. The reviews are quite favorable, but they also raised some important comments and suggestions that I urge you to consider as they might improve the quality of the manuscript. Based on my own reading, I find this to be an interesting paper that fits the scope of HESS well and will be of interest to the community.

We thank the editor and reviewers for the positive feedback on our manuscript. Please see the answers to the individual comments of the editor and reviewers below.

In addition to the comments from the reviewers, I kindly ask you to also consider my two cents: (i) consider discussing how general are the presented results (e.g. considering different climates, terrain, hydrological model types and calibration techniques);

We agree that it is of interest to all readers to what extent these results can be generalized. We added a section at the beginning of the discussion in section 4.1 (page 10, lines 10-17): **"In this study, we evaluated the information content of streamflow estimates by citizen scientists for calibration of the bucket-type hydrological model for six Swiss catchments. Streamflow estimates by citizens are sometimes very different from the measured values, and the individual estimates can be dis-informative for model calibration (Beven, 2016; Beven and Westerberg, 2011). While the hydroclimatic conditions, the model or the calibration approaches might be different in other studies, these results should be applicable for a wide range of cases. However, for physically-based spatially distributed models that are usually not calibrated automatically, the use of limited streamflow data would probably benefit from a different calibration approach. Furthermore, our results might not be applicable in arid catchment cases where rivers fall dry for some period of the year because the linear reservoirs used in the HBV model are not appropriate for such systems."**

(ii) Fig. 3 – consider reducing the symbol size of the dots and zooming closer to the observed hydrograph, I found it difficult to see the match between obs. and sim. data;

We thank the editor for this helpful suggestion. We changed the figure in the manuscript. Because not all dots are within the figure limits anymore, we also added the number of outliers and a sentence to the figure caption to explain this.

(iii) Fig. 6 – consider focusing on one or two cases (e.g. calibration: dry and validation: dry and calibration: wet and validation: wet) to make the figure larger and clearer, and presenting the other cases as supplementary information.

We changed the figure 6 to contain only the two suggested cases and added the entire figure to the supplemental material. The plot in the supplementary material also shows the results for the two cases highlighted in Figure 6 to enable an easier comparison between all cases.

I invite you to upload a revised manuscript, incorporating the proposed changes and additions, and making any other modifications where you see fit (minor revision iteration). I look forward to receiving the revised manuscript.

Sincerely,
Nadav Peleg

Replies to Reviewer #1

Anonymous Referee #1
Received and published: 13 July 2018

The manuscript entitled “Value of uncertain streamflow observations for hydrological modelling” presents interesting and novel research on the worth of citizen science discharge observations for the calibration of lumped hydrological models. The manuscript is well structured and concise with a clear motivation. The presentation and the application of the methods are scientifically sound. My comments are mostly of minor character and therefore I hope to see this article soon published in HESS.

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

general comments:

-The presentation of the calibration experiments is clear and complete with regard to the model performance. However, I was wondering how the model robustness is affected by uncertain observations which was neglected by the authors. From a modelling point of view parameter uncertainty and its reduction through calibration is of high importance. Therefore I believe that an additional figure on that matter would improve the quality of the study. How do the different temporal resolutions of observations as well as the three applied error scenarios affect the parameter values and their uncertainty compared to the benchmark case? This issue should be discussed in light of model equifinality.

We thank the reviewer for this helpful comment. We included a short paragraph about the effects of the errors on the timing and amount of data used for model calibration on the range of parameter values in the revised version of the manuscript.

In Figures R1-R6, we show the boxplots with the parameter ranges for each of the six catchments. Each boxplot consists of 300 values (3 year characters x 100 calibration runs). We summarized these results in another plot (Figure R7) that shows the interquartile range of the parameter distribution for each catchment for the different scenarios. The effects of the errors in the data and the timing of the data used for model calibration on the interquartile range of parameter values are summarized in Table R1. The spread in the parameter values was smallest for the upper benchmark for almost all parameters and cases, although the differences were very small for some parameters (e.g. PERC, PCALT and CWH). The trend of increasing spread in the parameter range with increasing errors is clearest for the MAXBAS parameter, which is the routing parameter. The parameter range of some other parameters (e.g. TCALT, TT and BETA) also increased with increasing error in the data used for calibration, but for other parameters (e.g., CFMAX, FC, and SFCF) the temporal resolution and the number of data points used for calibration determined the range in the optimized parameter values. However, these changes in the range of model parameters differed significantly for the different catchments (see differences in Figures R1-R6 and spread of the dots in Figure R7). Because these trends are not very clear, we prefer not to include any of the Figures R1-R6 but added a section 3.5 in the results on the parameter distribution (section 3.4, page 9, lines 14-19): **“For most parameters the spread in the optimized parameter values was smallest for the upper benchmark. The spread in the parameter values increased**

with increasing errors in the data used for calibration, particularly for MAXBAS (the routing parameter) but also for some other parameters (e.g. TCALT, TT and BETA). However, for some parameters (e.g., CFMAX, FC, and SFCF) the range in the optimized parameter values was mainly affected by the temporal resolution of the data and the number of data points used for calibration. It should be noted though that the changes in the range of model parameters differed significantly for the different catchments and the trends weren't very clear."

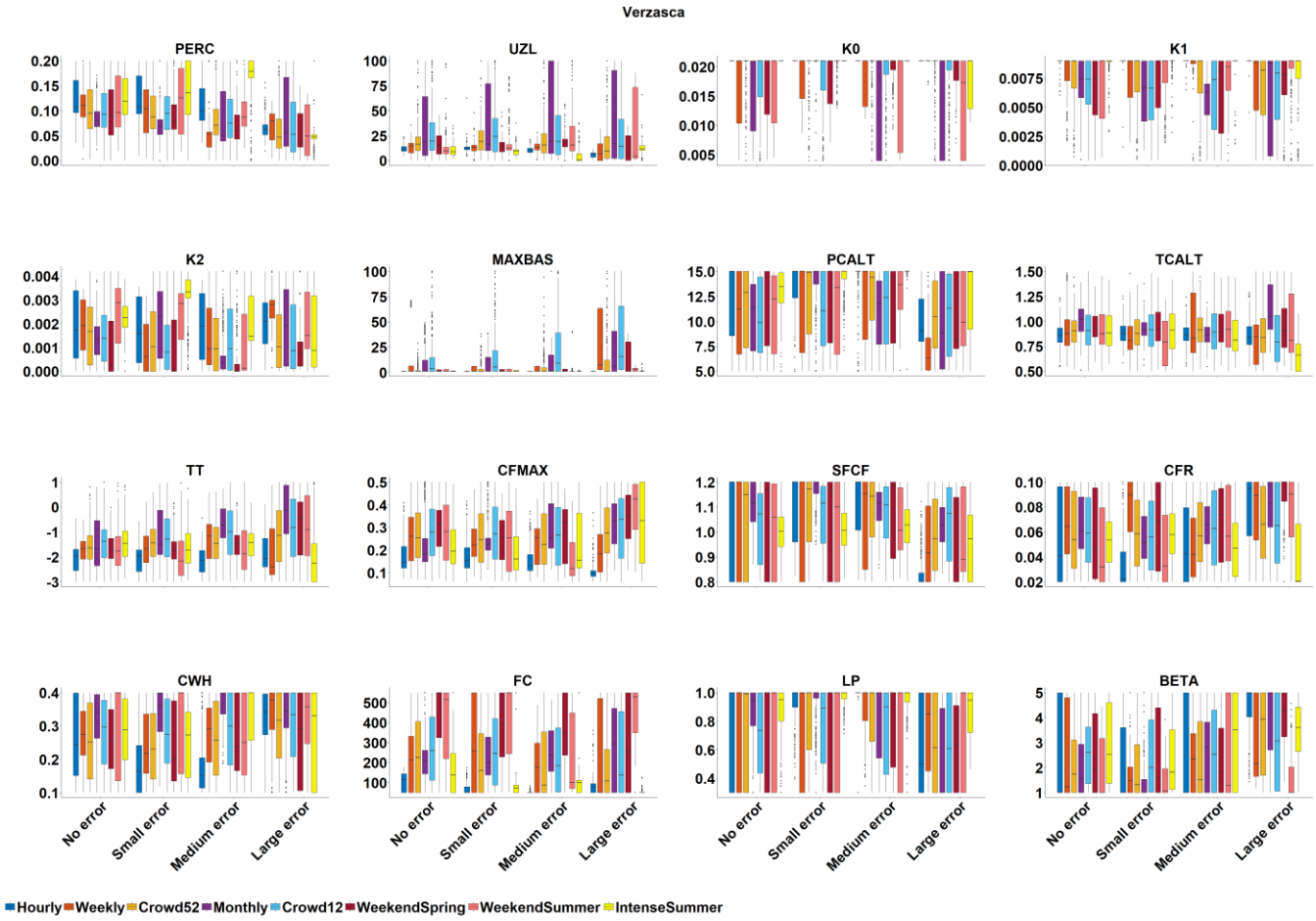


Figure R1 Boxplots of the model parameters for different combinations of errors and temporal resolutions of the data used for model calibration for the Verzasca catchment. Each subplot shows the range for one model parameter and consists of 300 values (3 year characters x 100 calibrations). The box represents the 25th and 75th percentile, the thick horizontal line the median, the whiskers extend to 1.5 times the interquartile range below the 25thpercentile and above the 75th percentile, and the dots represent the outliers. For a description of the model parameters and units see Table R1.

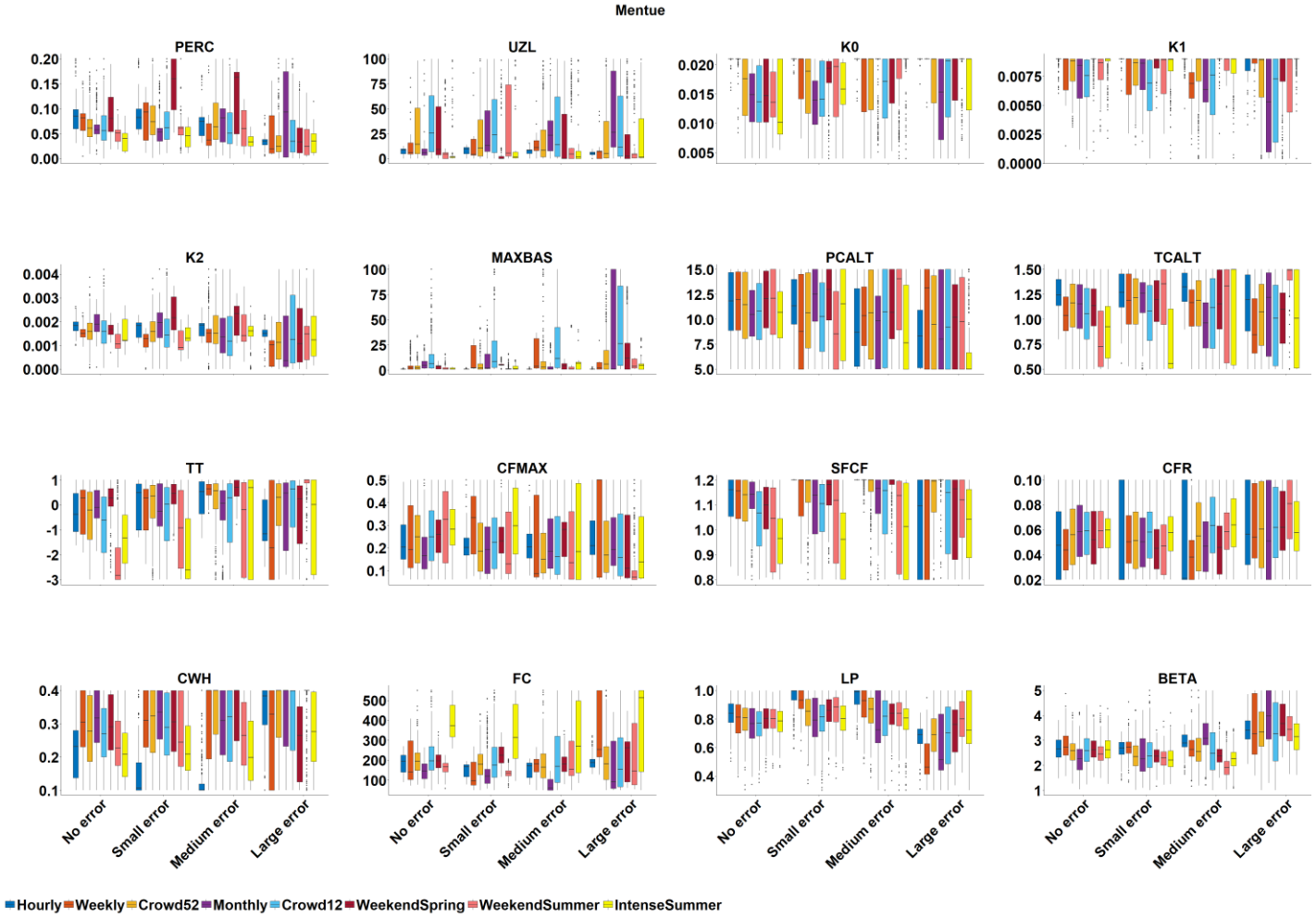


Figure R2 Boxplots of the model parameters for different combinations of errors and temporal resolutions of the data used for model calibration for the Mentue catchment. Each subplot shows the range for one model parameter and consists of 300 values (3 year characters x 100 calibrations). For a description of the box plots see Figure 1. For a description of the model parameters and units see Table R1.

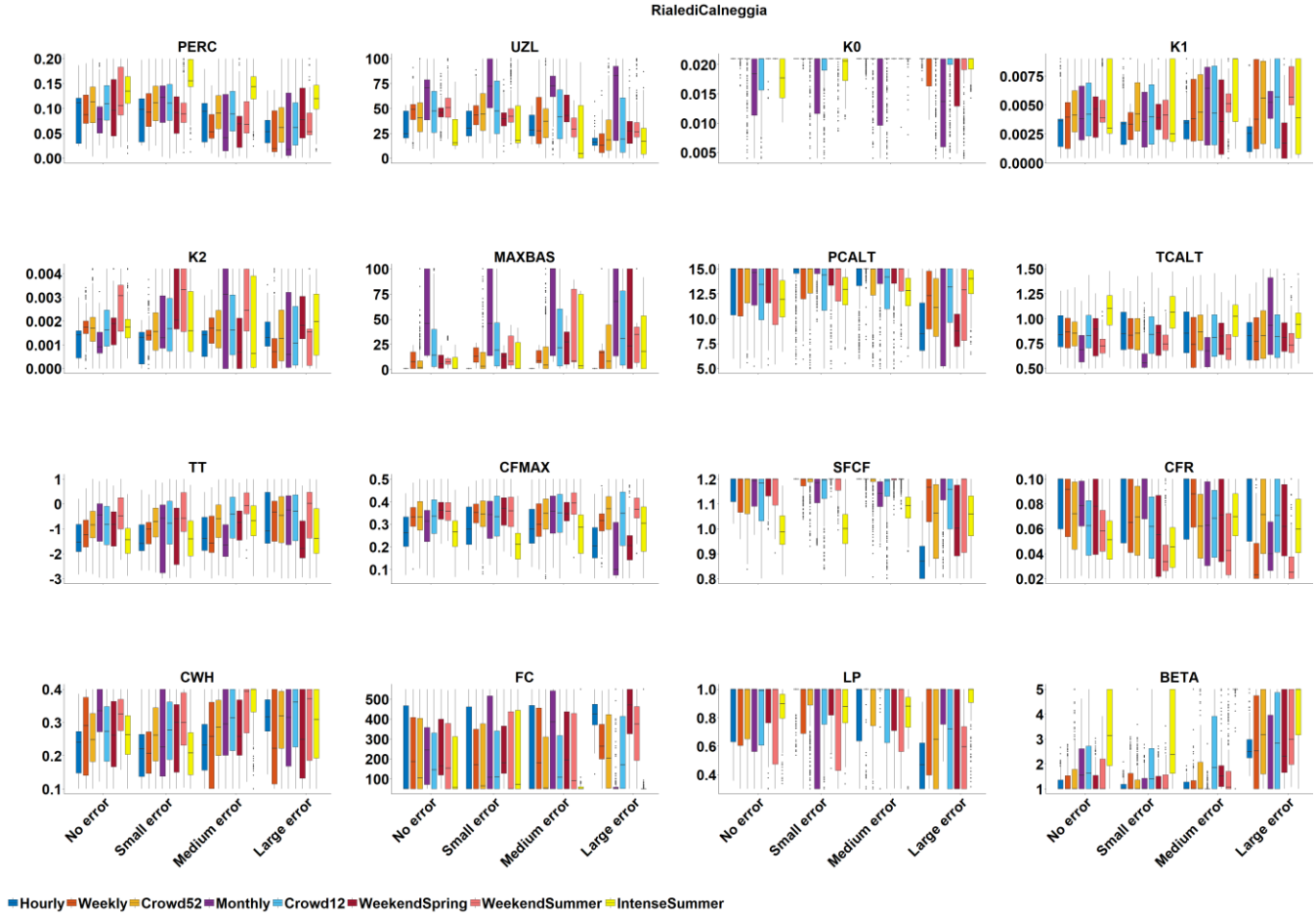


Figure R3 Boxplots of the model parameters for different combinations of errors and temporal resolutions of the data used for model calibration for the Riale di Calneggia catchment catchment. Each subplot shows the range for one model parameter and consists of 300 values (3 year characters x 100 calibrations). For a description of the box plots see Figure 1. For a description of the model parameters and units see Table R1.

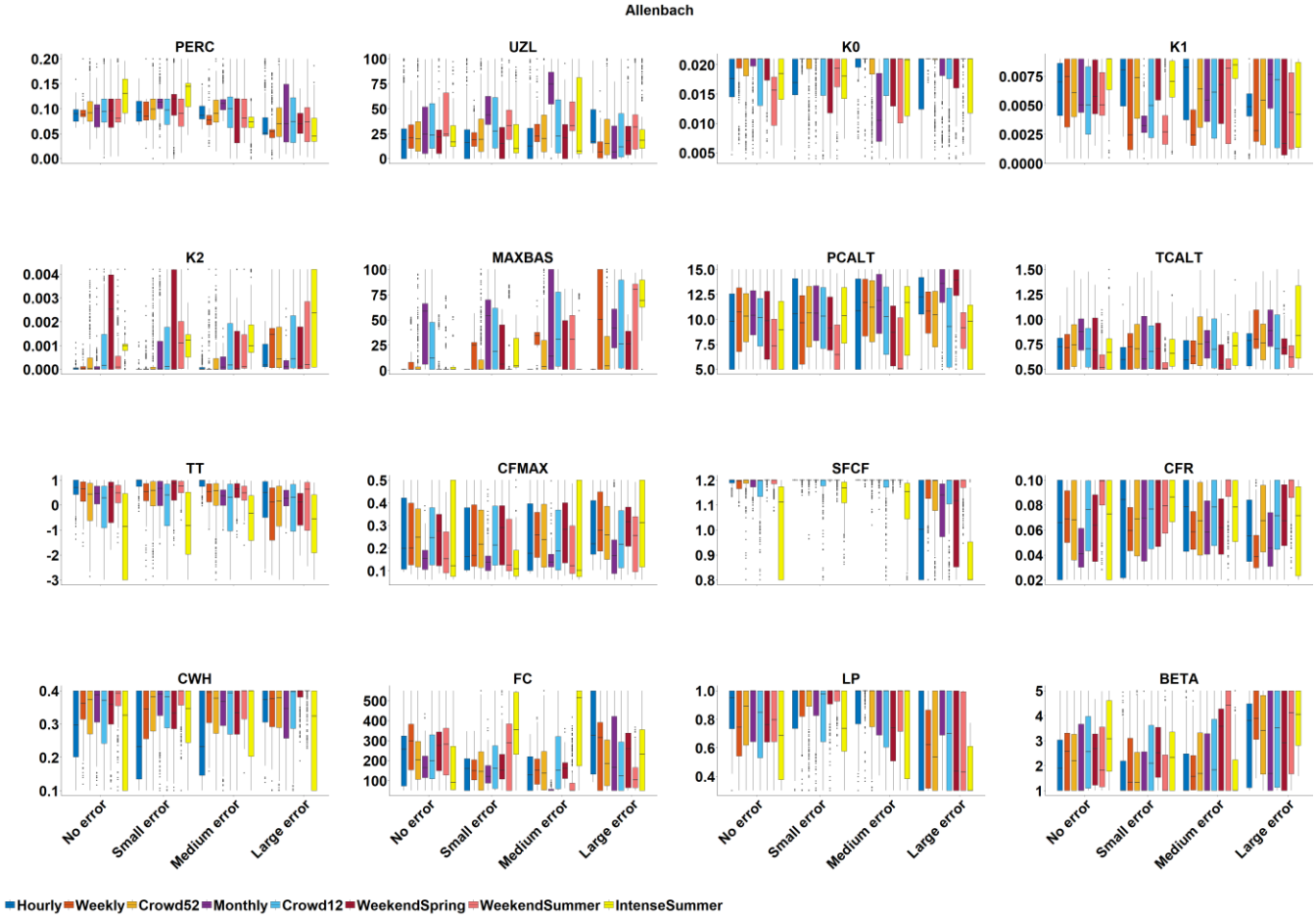


Figure R4 Boxplots of the model parameters for different combinations of errors and temporal resolutions of the data used for model calibration for the Allenbach catchment. Each subplot shows the range for one model parameter and consists of 300 values (3 year characters x 100 calibrations). For a description of the box plots see Figure 1. For a description of the model parameters and units see Table R1.

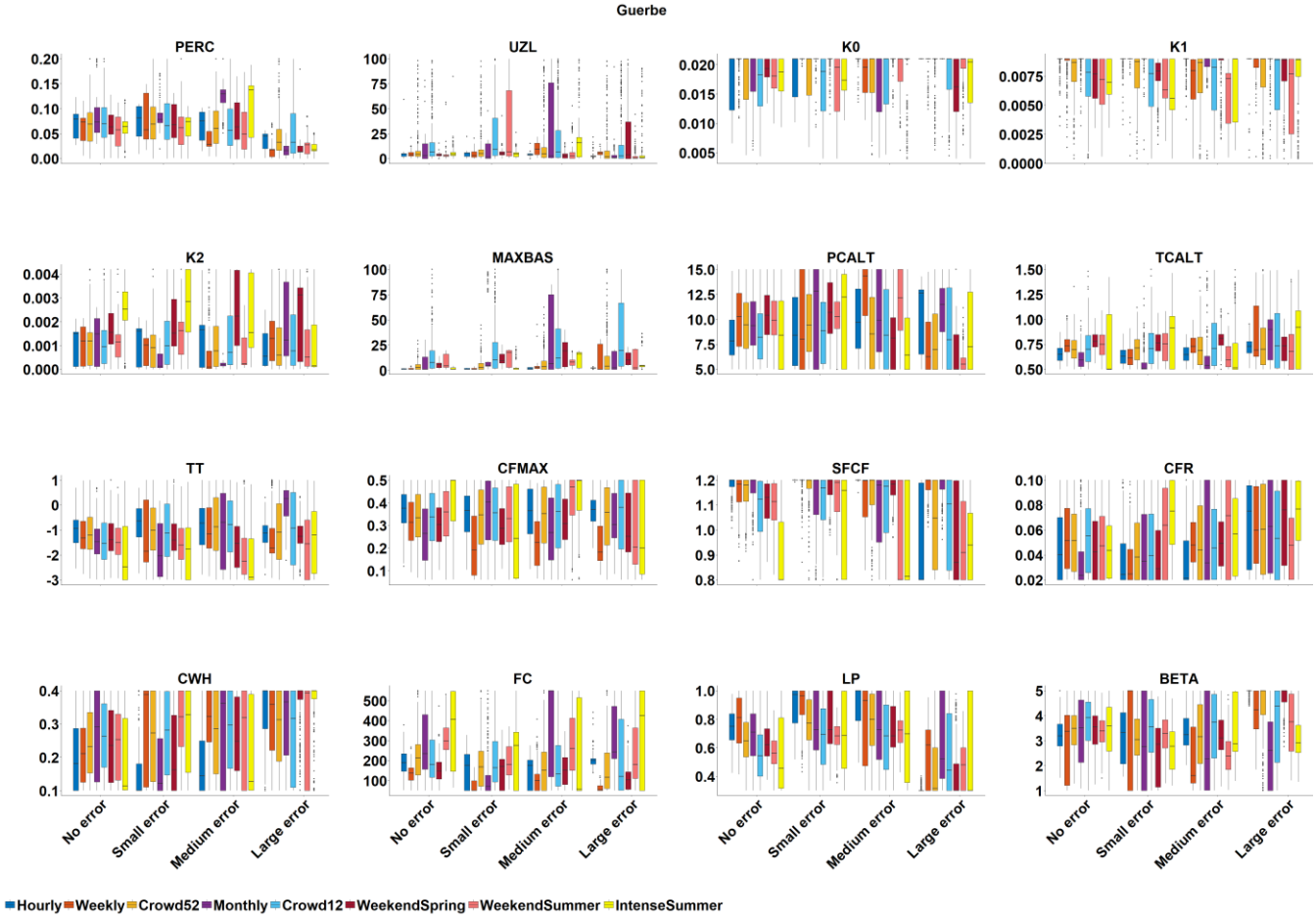


Figure R5 Boxplots of the model parameters for different combinations of errors and temporal resolutions of the data used for model calibration for the Guerbe catchment. Each subplot shows the range for one model parameter and consists of 300 values (3 year characters x 100 calibrations). For a description of the box plots see Figure 1. For a description of the model parameters and units see Table R1.

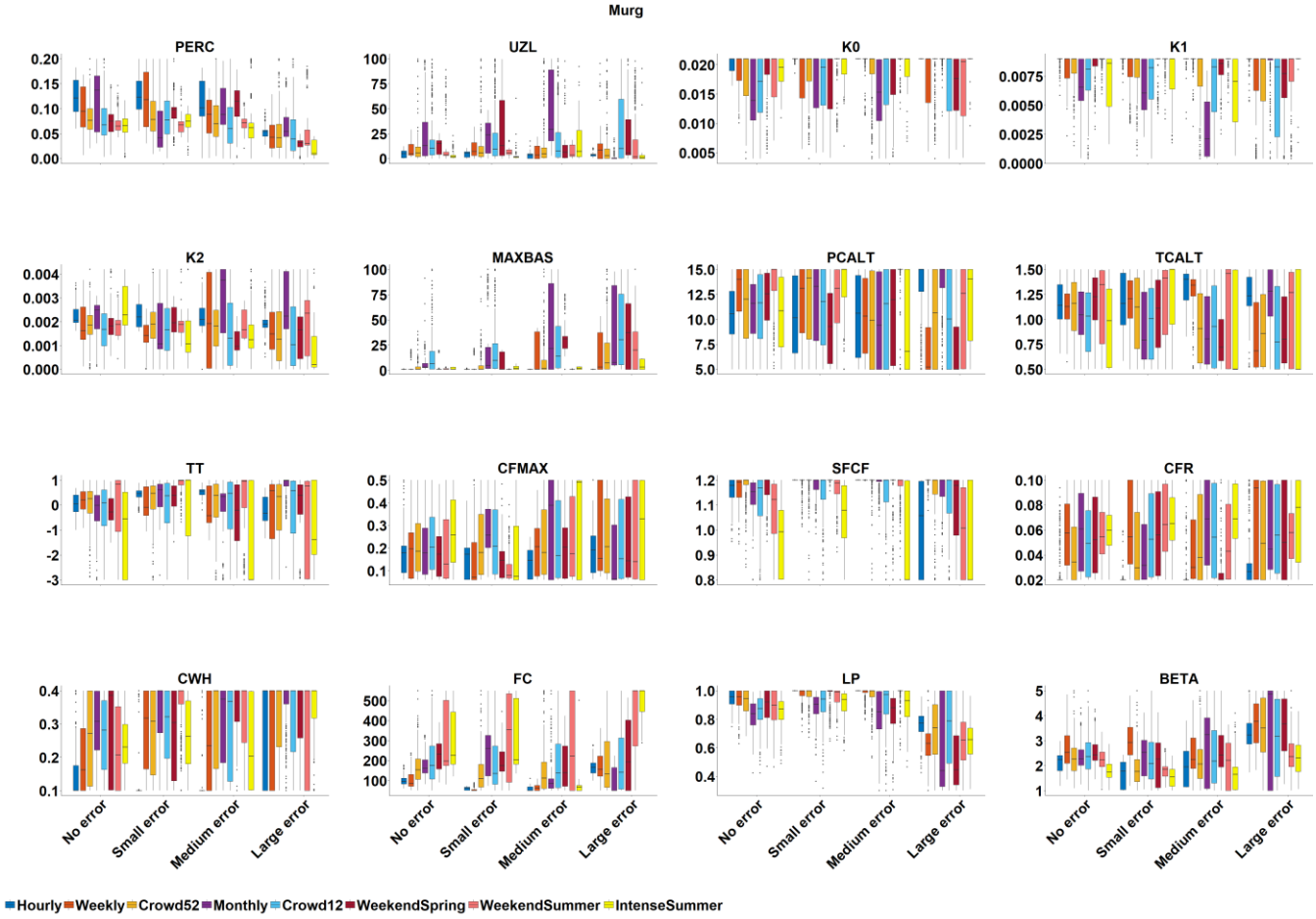


Figure R6 Boxplots of the model parameters for different combinations of errors and temporal resolutions of the data used for model calibration for the Murg catchment. Each subplot shows the range for one model parameter and consists of 300 values (3 year characters x 100 calibrations). For a description of the box plots see Figure 1. For a description of the model parameters and units see Table R1.

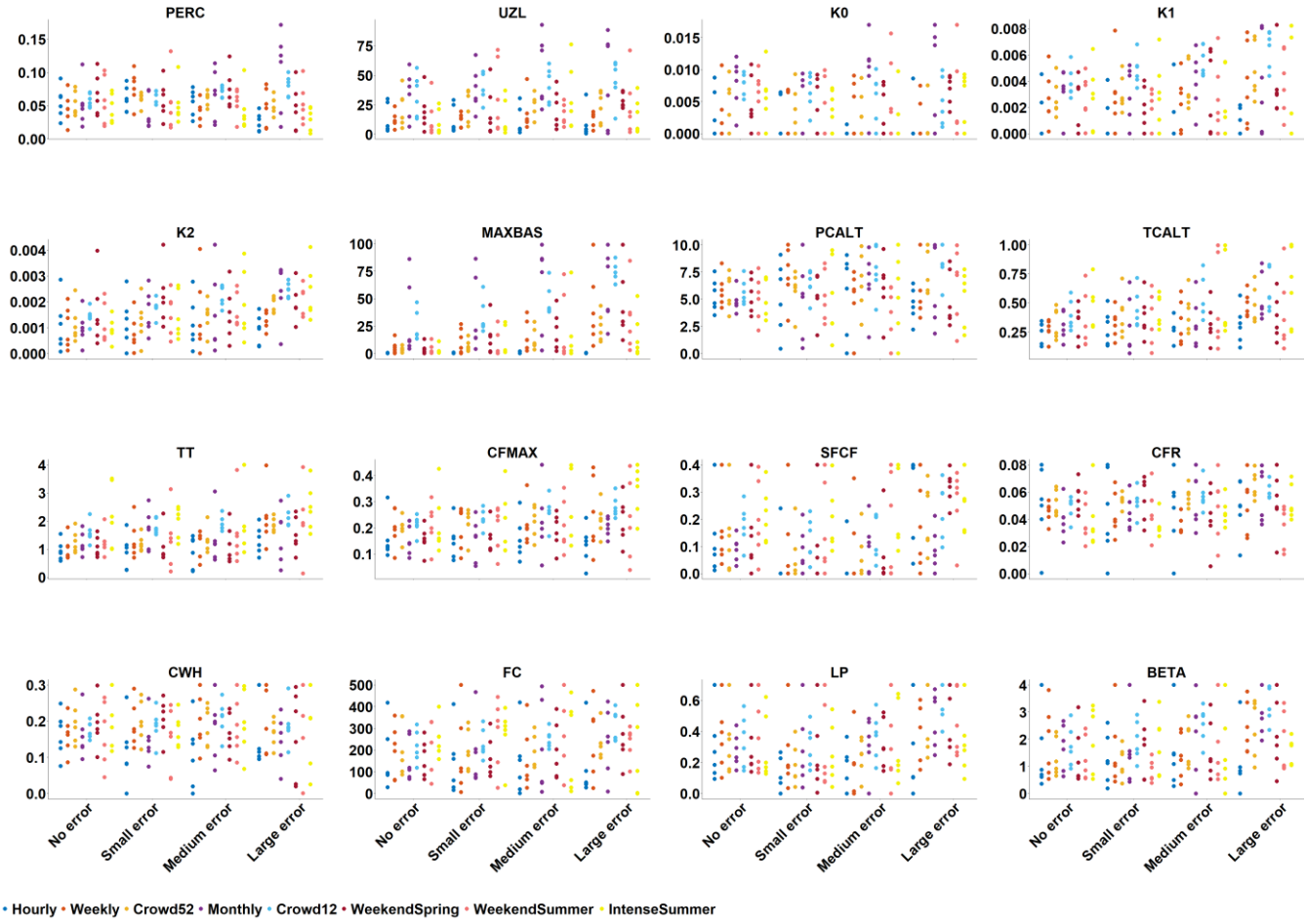


Figure R7 The interquartile range of the model parameters for the six catchments for the different combinations of errors and temporal resolutions of the data used for model calibration. Each dot represents the interquartile range for one catchment (i.e. is the size of the box in Figures R1-R6). For a description of the model parameters and units see Table R1.

5

10

Table R1 Effect of errors and timing of the data used for model calibration on the interquartile range of the calibrated parameters in the HBV-light model (See also Figure R7). For a description of the different data sets (names in italic) see the main text of the manuscript.

Parameter		Effect of errors	Effect of timing
PERC	Maximum percolation from upper to lower groundwater storage [mmd^{-1}]	No clear effect of errors, only *Monthly* dataset has larger range if large errors	Slightly larger range for *Monthly* and *Crowd12* data sets if large errors
UZL	Threshold parameter [mm]	No big effect, larger range for *Monthly* with increasing errors	Largely effect of timing
K0	Storage (or recession) coefficients [h^{-1}]	Slightly larger parameter range for medium and large errors	No clear effect of timing
K1	Storage (or recession) coefficients [h^{-1}]	Slightly larger range for *WeekendSpring*, *WeekendSummer IntenseSummer* datasets, smaller range for e.g. *hourly* dataset with increasing errors	*Hourly* dataset usually has the smallest range
K2	Storage (or recession) coefficients [h^{-1}]	No effect	No effect
MAXBAS	Length of triangular weighting function [H]	Increasing range with increasing errors	Large range for *WeekendSpring* dataset
PCALT	Change in precipitation with elevation [% (100 m)$^{-1}$]	Sometimes larger and sometimes smaller range with increasing errors	No clear effect of timing, *hourly* dataset has the smallest range
TCALT	Change in temperature with elevation [°C (10 m)$^{-1}$]	Increasing range with increasing errors	Some effect of timing, sometimes smaller range, sometimes larger range with less data (e.g. *Weekly* dataset)
TT	Threshold temperature for liquid and solid precipitation [°C]	Increasing range with increasing errors	Some effect of timing
CFMAX	Degree-day factor [mm d^{-1}°C^{-1}]	Only for largest errors increase in parameter range	Larger range for intense summer than for others
SFCF	Snowfall correction factor [-]	No effect	No effect
CFR	refreezing coefficient [-]	No effect	No effect, *Crowd52* dataset usually has the smallest range
CWH	Water holding capacity of the snow storage [-]	Larger range for *WeekendSpring* and *Intense Summer* datasets with increasing errors, for other datasets no clear trend	No effect
FC	Maximum soil moisture storage [Mm]	*IntenseSummer* and *Weekly* datasets have larger range with increasing errors, for other datasets no clear trend	No effect
LP	Soil moisture value above which actual evapotranspiration reaches potential evapotranspiration [-]	No effect	No effect
Beta	Shape factor for the function used to calculate the distribution of rain and snow melt going to runoff and soil box	No effect	No effect

-In my opinion the authors should be more specific that their study addresses lumped hydrological models. For integrated spatially distributed models such a study surely would have different implications. Therefore I suggest to clearly state this throughout the manuscript; especially in title, introduction and discussion.

We agree and adjusted the text in the introduction and the conclusion to more explicitly state that these results are for lumped, or as we prefer to call them, bucket-type hydrological models.

specific comments:

-In the introduction the authors provide a great overview on existing studies addressing the question how much data is needed to calibrate a hydrological model. I am wondering why the findings vary so drastically between days to years. Can the authors provide an explanation for this?

These studies all had a different focus, used different performance metrics and different definitions of what a good model performance is (see Table R2). Vrugt et al. (2006) and Yapo et al. (1996) defined stable parameters as a good calibration criterion. Others (Juston et al. (2009); Seibert and Beven (2009); Seibert and McDonnell (2015)) used benchmark calibrations and looked at the differences in the values of the objective functions. Pool et al. (2017) always used 12 streamflow data points and explored the best timing of these measurements. Juston et al. (2009) used a very long time series with possibly much more variation in streamflow than is observed within one year of data (as in this study) from which the subsets were drawn. Brath et al. (2004) used a spatially distributed model and concluded that three months was the absolute minimum.

We added more information to the introduction to describe why the different studies resulted in different minimum data sets and highlighted better that despite their differences they all find that limited datasets are useful (page 1, line 29-page 2, line 17): **"Therefore, several studies have addressed the question: how much data are needed to calibrate a model for a catchment? Yapo et al. (1996) and Vrugt et al. (2006a) using stable parameters as a criteria for satisfying model performance, concluded that most of the information to calibrate a model is contained in 2-3 years of continuous streamflow data and that no more value is added when using more than eight years of data. Perrin et al. (2007) using the Nash-Sutcliffe efficiency criterion (NSE), showed that streamflow data for 350 randomly sampled days out of a 39 year period were sufficient to obtain robust model parameter values for two bucket-type models, TOPMO, which is derived from TOPMODEL concepts (Michel et al. 2003), and GR4J (Perrin et al., 2003). Brath et al. (2004) using the volume-, relative peak- and time to peak error concluded that at least three months of continuous data were required to obtain a reliable calibration. Other studies have shown that discontinuous streamflow data can be informative for constraining model parameters (Juston et al., 2009; Pool et al., 2017; Seibert and Beven, 2009; Seibert and McDonnell, 2015). Juston et al. (2009) used a multi-objective calibration that included groundwater data and concluded that the information content of a subset of 53 days of streamflow data was the same as 1065 days of data from which the subset was drawn. Seibert and Beven (2009) using the NSE criterion, found that model performance reached a plateau for 8-16 streamflow measurements collected throughout a one-year period. They, furthermore, showed that the use of streamflow data for one event and the corresponding recession resulted in a similar calibration performance as data for the six highest measured streamflow values during a two-month period.**

These studies had different foci and used different model performance metrics but nevertheless their results are encouraging for the calibration of hydrological models for ungauged basins based on a limited number of high-quality measurements. However, the question remains: how informative are low(er)-quality data? ”

Table R2 Cited modelling studies focusing on the amount of streamflow data necessary to calibrate a hydrological model.

Study / Authors	Performance Metric	Temporal Resolution	Model	How much data was needed
Yapo et al., 1996	Daily root mean square estimation criterion and hetero-scedastic maximum likelihood error	Daily	NWSRFS_SMA model (Brazil, 1988)	No more added value after 8 years of data. If wettest years are chosen for calibration, model parameters were “properly identifiable”
Vrugt et al., 2006	RMSE	daily	Sacramento Soil Moisture Accounting model	Stable estimates for most of the parameters with 2-3 years of streamflow data
Perrin et al., 2007	NSE in calibration, NSE and LogNSE in validation	daily	TOPMO (derived from TOPMODEL concepts (Michel et al 2003) and GR4J (Perrin et al 2003)	350 random days out of a 39 year period including dry and wet conditions are sufficient to obtain robust model parameters
Brath et al., 2004	Relative volume error, relative peak error, Time to peak error	Hourly	Spatially distributed model	At least 3 months were required to obtain reliable calibration
Juston et al., 2009	Combination of NSE and groundwater performance index (multi-objective calibration)	daily	HBV-Forsmark	Information content of subset of 53 days was the same, as the entire 1065-day period from which the data was drawn
Pool et al., 2017	NSE and log NSE	daily	HBV	12 data points, different “sampling“ strategies (high flows, low flows, recession limbs, on the peak, etc)
Seibert and Beven, 2009	NSE	daily	HBV	Model performance plateaued after 8-16 streamflow measurements within a one year period
Seibert and McDonnell, 2015	the overall acceptability of a parameter set was defined by three components: (1) the model	10 min streamflow data	Variant of HBV	One event or 10 high flow measurements provided almost as much information as a

	efficiency (NSE) values (Nash and Sutcliffe 1970) for the hard runoff data (calculated based on subsets of the total runoff series), (2) the acceptability of the model simulations with regard to soft data, and (3) the acceptability of the parameter values based on the experimentalist's understanding.			3 months of data

-The applications of citizen science in hydrology are broad and go beyond the collection of data. For completion the authors could mention Koch et al. (2017) were the human perception was consulted to compare the similarity between simulated spatial patterns in order to evaluate spatial performance metrics.

Thank you for providing this reference. We included a statement that citizen science also includes data analysis in the introduction and refer to this paper there (page 2, lines 18-20): **"Citizen science has been proven to be a valuable tool to collect (Dickinson et al., 2010) or analyse (Koch and Stisen, 2017) various kinds of environmental data, including hydrological data (Buytaert et al., 2014)."**

-Extreme outliers are filtered with respect to maximum possible streamflow values. One could imagine a more thorough filtering based on the season. An extreme outlier during low flow season can be expected to be smaller than during high flow. Have the authors considered such an improved filtering?

We agree that when using real citizen science data more advanced filtering mechanisms are useful but these will have to be thoroughly tested first. The testing of different filtering methods is not within the scope of our study. Also, these filtering mechanisms need to be applicable for all places without measurements or local knowledge about flows. Low flows and high flows can occur in multiple seasons and differ drastically between stations (and years) because some catchments are snow influenced, others have glaciers, and others are only rain fed. Therefore, local knowledge or data about how low flows depend on the season are necessary. We considered the lowest and highest ever measured values for a particular catchment size for the filtering because these data may be available for different regions and provide a very simple filter to take out the most unrealistic values. Because no extreme low flow value was replaced with the lowest ever recorded flow and only a few high flow estimates were replaced, we assume that the results would not have been significantly different if a slightly more advanced filtering mechanism for low flows was used.

-I can imagine a better visualization of the data in Figure 3. Instead of nine subplots one could imagine three subplots, one for each temporal resolution. Then each error scenario could have a different color. In this way the graphs could be stretched over the entire page and the dynamics would be more visible.

Thank you for this suggestion for an improvement. We thought about the suggested graph but there are too many dots (or other symbols) that overlap, as can be seen in Figures R8-R9. This makes it hard to read the figures, even if the symbols are not filled. We however adapted the figure according to the editors' comment.

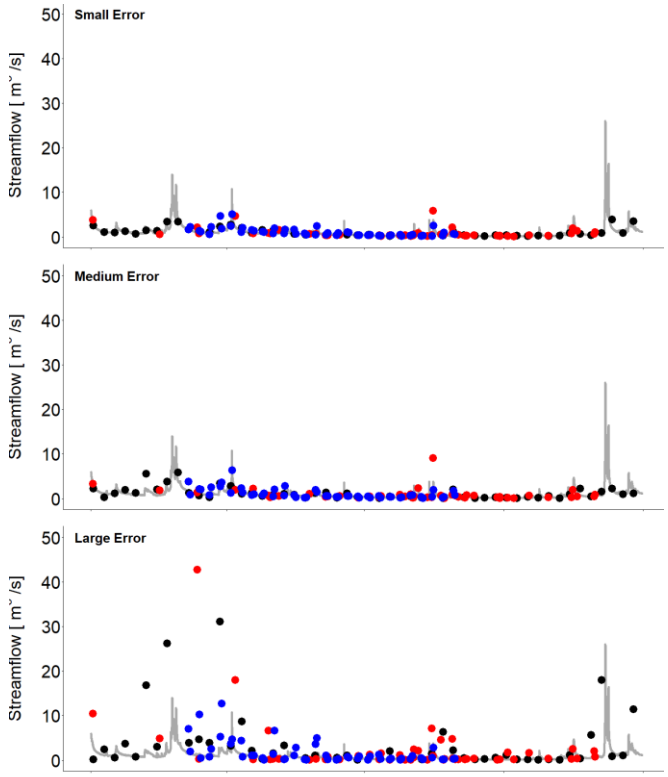


Figure R8 Alternative design of Figure 3 with filled symbols to represent the data used for model calibration for the different scenarios.

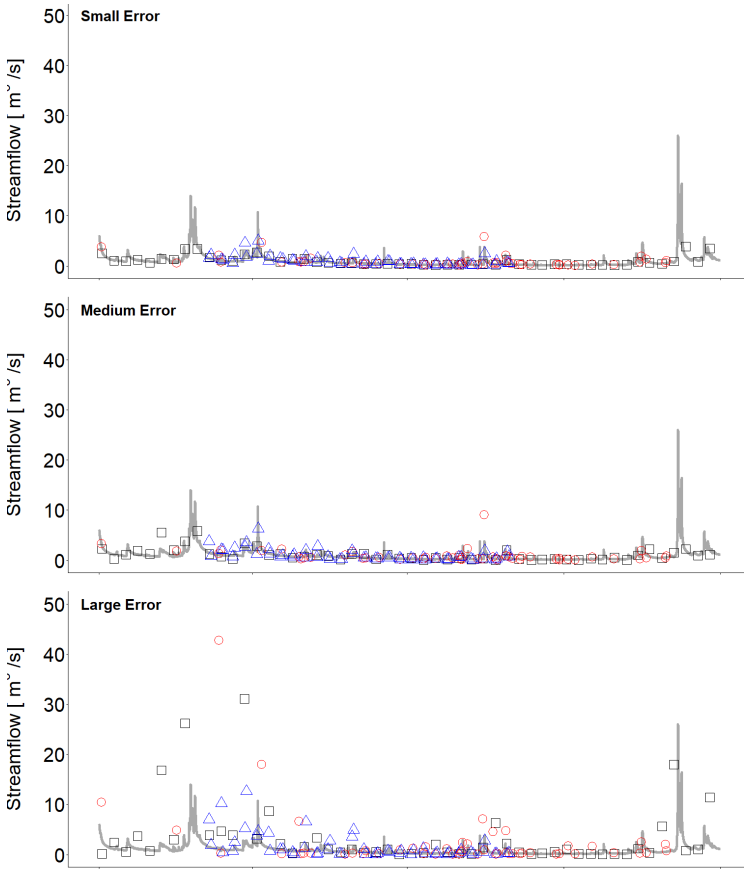


Figure R9 Alternative design of Figure 3 with different open symbols to represent the data used for model calibration for the different scenarios.

Replies to Reviewer #2

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. We have carefully read through the manuscript and shortened several of the long sentences.

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 noting this inconsistency. We included the following sentence in the revised manuscript (page 14, line 30 – page 15, line1): **“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 cannot 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 extended the section in the revised manuscript (page 7, lines 19-25): **“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 for model calibration compared to a situation without any streamflow data, a lower benchmark (Seibert et al., 2018) was used. 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 included 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 (page 13, lines 27-30): **"Options for training might be tutorial videos, as well as providing values for the width, average depth and flow velocity of 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 clarified this by rewriting these sentences in the following way (page 2, lines 27-30): "**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 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. (2018) and Morlot et al. (2018) showed that this correction factor is reasonable for most streams, except for concrete channels. Even if the exact value of the correction factor is uncertain (e.g. varies between 0.71 and 0.95 (Welber et al., 2016)), 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. (2018) Velocity profile and depth-averaged to surface velocity in natural streams: a review over a large sample of rivers. E3S Web of Conferences (doi: 10.1051/e3sconf/20184006015)

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.

Welber, M., Le Coz, J., Laronne, J. B., Zolezzi, G., Zamler, D., Dramais, G., Hauet, A. and Salvaro, M.: Field assessment of noncontact stream gauging using portable surface velocity radats (SVR), Water Resour. Res., 52, 1108–1126, doi:10.1002/2015WR017906.Received, 2016.

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, therefore, mention that the depth is particularly uncertain when we describe the options of training (page 13, lines 27-30).

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 to run the model are hourly averages 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 inserted the following sub-sentence in the manuscript (page 4, lines 5-6): **“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 (page 6, lines 21-24):

"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 test the effect of errors when the temporal resolution of the data is optimal (i.e., by comparing simulations for the models calibrated with the hourly FOEN data and those calibrated 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 (page 5, lines 18-25):

"To summarize, we tested the following four cases:

- ***No error*: The data measured by the FOEN, assumed to be (almost) 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 surveys 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 surveys."**

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 added a sentence in chapter 2.6 (page 7, line 3): **"The parameters were calibrated within their typical ranges (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 changed the sentence to (page 10, lines 22-23): **"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 changed 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 therefore rewrote this sentence (page 11, lines 22-25): **"…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 small. We changed the sentence into (page 11, lines 26-28): **"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 deleted it and rewrote the sentence (page 12, lines 7-9): **"This is good news for using citizen science data for model calibration as it suggests that the timing 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 deleted 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 (page 1, lines 17-19): "**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 changed it accordingly (page1, lines 29-30): **"Therefore, several studies have addressed the question: how much data are needed to calibrate a model for a catchment?"**

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 changed it to (page 2, lines 15-17): **"These studies had different foci and used different model performance metrics but nevertheless their results are encouraging for the calibration of hydrological models for ungauged basins based on a limited number of high-quality measurements. However, the question remains: how informative are low(er)-quality data?"**

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

Thanks for pointing this out, we changed it accordingly.

4-18/19 - sentence is incomplete

Thanks for pointing this out. We changed it to (page 4, lines 26-27): "**For the validation, we chose the year closest to the mean summer streamflow and the years with the lowest and the highest total summer streamflow (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 adapted the section in the revised manuscript.

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

Thank you for pointing this out, we corrected it.

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

Thanks for pointing at this shortcoming: We edited the sentence to clarify this (page 10, lines 3-4): **"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 changed the sentence in the revised version to (page 14, lines 28-29): **"… (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)."**

5

Reply to Interactive Comment by Anusha Pandey

Received and published: 8 August 2018

This is a very interesting paper which will add more value in citizen hydrology. My only comment is: When stating examples of citizen science projects that collect streamflow or stream level data (page 2 line 21), it would be helpful to include SmartPhones4Water Nepal (http://www.smartphones4water.org/category/news/) as an example project using references of either https://link.springer.com/article/10.1007/s00267-017-0872-x or https://link.springer.com/article/10.1007%2Fs10661-018-6687-2. This is especially important because literature about citizen science so far has been relatively focused on the west, so any example applications of citizen science in Asia should be properly included.

Thank you very much for making us aware of these interesting recent studies and the project SmartPhones4Water. We included a reference in the introduction in the revised manuscript where similar citizen science projects are mentioned (page 2, lines 24-27): "**Examples are the CrowdHydrology project (Lowry and Fienen, 2013), SmartPhones4Water in Nepal (Davids et al., 2018) and a project in Kenya (Weeser et al., 2018), which all ask citizens to read stream levels at staff gauges and to send these via an app or as text message to a central database.**"

Value of uncertain streamflow observations for hydrological modelling

Simon Etter[1], Barbara Strobl[1], Jan Seibert[1,2], Ilja van Meerveld[1]

[1]Department of Geography, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
[2]Department of Earth Sciences, Uppsala University, Villavägen 16, 752 36 Uppsala, Sweden

Correspondence to: Simon Etter (simon.etter@geo.uzh.ch)

Abstract. Previous studies have shown that ~~a~~ hydrological ~~model~~models can be parameterized using ~~on~~ a limited number of streamflow measurements~~ for otherwise ungauged basins.~~. Citizen science projects can collect such data for otherwise ungauged catchments but an important question is whether these observations are informative given that these streamflow estimates will be uncertain. We address the value of inaccurate streamflow estimates for calibration of a simple bucket-type runoff model for six Swiss catchments. We pretended that only a few observations were available and that these were affected by different levels of inaccuracy. The ~~initial~~level of inaccuracy~~ level~~ was based on a log-normal error distribution that was fitted to streamflow estimates of 136 citizens for medium-sized streams. Two additional levels of inaccuracy, for which the standard deviation of the error-distribution was divided by two and four, were used as well. Based on these error distributions, random errors were added to the measured hourly streamflow data. New time series with different temporal resolutions were created from these synthetic streamflow time series. These included scenarios with one observation each week or month~~ and~~, as well as scenarios that are more realistic for crowdsourced ~~datasets with~~data that generally have an irregular ~~distributions~~distribution of data points throughout the year, or ~~a~~ focus on ~~spring or summer~~a particular season. The model was then calibrated for the six catchments using the synthetic time series for a dry, an average and a wet year. The performance of the calibrated models was evaluated based on the measured hourly streamflow time series. The results indicate that streamflow estimates from untrained citizens are not informative for model calibration. However, if the errors can be reduced, the estimates are informative and useful for model ~~parameterization.~~calibration. As expected, the model performance increased when the number of observations used for calibration increased. The model performance was also better when the observations were more evenly distributed throughout the year. This study indicates that uncertain streamflow estimates can be useful for model calibration but that the estimates by citizen scientists need to be improved by training or more advanced data filtering before they are useful for model calibration.

1 Introduction

The application of hydrological models usually requires several years of precipitation, temperature and streamflow data for calibration, but these data are only available for a limited number of catchments. Therefore, several studies have addressed

the question: how much data are needed to calibrate a model for a catchment ~~where continuous streamflow data are not available.~~? Yapo et al. (1996) and Vrugt et al. (2006a) using stable parameters as a criteria for satisfying model performance, concluded that most of the information to calibrate a model is contained in 2-3 years of continuous streamflow data and that no more value is added when using more than eight years of data. Perrin et al. (2007) using the Nash-Sutcliffe efficiency criterion (NSE), showed that streamflow data for 350 randomly sampled days out of a 39 year period were sufficient to obtain robust model parameter values for two bucket-type models, TOPMO, which is derived from TOPMODEL concepts (Michel et al. 2003), and GR4J (Perrin et al., 2003). Brath et al. (2004) using the volume-, relative peak- and time to peak error concluded that at least three months of continuous data were required to obtain a reliable calibration. Other studies have shown that discontinuous streamflow data can be informative for constraining model parameters (Juston et al., 2009; Pool et al., 2017; Seibert and Beven, 2009; Seibert and McDonnell, 2015). Juston et al. (2009) used a multi-objective calibration that included groundwater data and concluded that the information content of a subset of 53 days of streamflow data was the same as 1065 days of data from which the subset was drawn. Seibert and Beven (2009) using the NSE criterion, found that model performance reached a plateau ~~level~~ for 8-16 streamflow measurements collected throughout a one-year period. They, furthermore, showed that the use of streamflow data for one event and the corresponding recession resulted in a similar calibration performance as data for the six highest measured streamflow values during a two-month period.

~~These~~These studies had different foci and used different model performance metrics but nevertheless their results are encouraging for the calibration of hydrological models for ungauged basins based on a limited number of high-quality measurements~~, but~~. However, the question remains: how informative are low(er)-quality data ~~are.~~? An alternative approach to high quality streamflow measurements in ungauged catchments is to use citizen science. Citizen science has been proven to be a valuable tool to collect (Dickinson et al., 2010) or analyse (Koch and Stisen, 2017) various kinds of environmental data, including hydrological data (Buytaert et al., 2014). Citizen science approaches use simple methods to enable a large number of citizens to collect data and allow local communities to contribute data to support science and ~~water~~environmental management. Citizen science approaches can be particularly useful in light of the declining stream gauging networks (Ruhi et al., 2018; Shiklomanov et al., 2002) and to complement the existing monitoring networks. However, citizen science projects that collect streamflow or stream level data in flowing ~~waterbodies~~water bodies are still rare. ~~Two~~ Examples are the CrowdHydrology project (Lowry and Fienen, 2013), SmartPhones4Water in Nepal (Davids et al., 2018) and a project in Kenya (Weeser et al., 2018), which ~~both~~all ask citizens to read stream levels at staff gauges and to send these via an app or as text ~~messages~~message to a central database. 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 ~~and thus~~, this allows them to obtain a rough estimate of the streamflow. However, these streamflow estimates may be so inaccurate that they are not useful for model calibration. It is therefore necessary to not only

evaluate the requirements of hydrological models~~, not only~~ in terms of the amount and temporal resolution of data, but also in terms of the achievable quality by the citizen scientists before starting a citizen science project.

The ~~effect~~effects of rating curve uncertainty on model calibration ~~has been quantified in recent studies~~ (e.g. McMillan et al. 2010; Horner et al. 2018) and the value of sparse datasets (Davids et al., 2017)~~but~~ have been quantified in recent studies. However, the potential value of sparse datasets in combination with large uncertainties (such as those from crowdsourced streamflow estimates) has not been evaluated so far. Therefore, the aim of this study was to determine the effects of observation inaccuracies on the calibration of bucket-type hydrological ~~model calibration~~models when only a limited number of observations are available. The specific objectives of this paper are to determine (*i*) whether the streamflow estimates from citizen scientists are informative for model calibration or if these errors need to be reduced (e.g. through training) to become useful and (*ii*) how the timing of the streamflow observations affects the calibration of a hydrological model. The latter is important for citizen science projects, as it provides guidance on whether it is useful to encourage citizens to contribute streamflow observations during a specific time of the year.

2 Methods

To assess the potential value of crowdsourced streamflow estimates for hydrological model calibration, the HBV ~~model (Bergström, 1972)~~ (Hydrologiska Byråns Vattenbalansavdelning) model (Bergström, 1976) was calibrated against streamflow time series for six Swiss catchments, as well as different subsets of the data that represent citizen science data in terms of errors and temporal resolution. Similar to the approach used in several recent studies (Ewen et al., 2008; Finger et al., 2015; Fitzner et al., 2013; Haberlandt and Sester, 2010; Seibert and Beven, 2009), we pretended that only a small subset of the data were available~~.~~ for model calibration. In addition, various degrees of inaccuracy were assumed. The value of these data for model calibration was then evaluated by comparing the model performance for ~~the subset~~these subsets of data to the performance of the model calibrated with the complete measured streamflow time series.

2.1 HBV model

The HBV ~~(Hydrologiska Byråns Vattenbalansavdelning)~~ model was originally developed at the Hydrologiska Byrans Vattenavdelning unit at the Swedish Meteorological and Hydrological Institute (SMHI) by Bergström (1976). The HBV model is a bucket-type model that represents snow, soil, groundwater and stream routing processes in separate routines. In this study, we used the version HBV-light (Seibert and Vis, 2012).

2.2 Catchments

The HBV-light model was set up for six 24-186 km^2 catchments in Switzerland (Table 1 and Figure 1). The catchments were selected based on the following criteria: *i*) there is little anthropogenic influence, *ii*) they are gauged at a single location, *iii*) they have reliable streamflow data during high flow and low flow conditions (i.e. no complete freezing during winter and a cross section that allows accurate streamflow measurement at low flows), and *iv*) there are no glaciers. The six selected

catchments (Table 1) represent different streamflow regime types (Aschwanden and Weingartner, 1985). The snow dominated highest elevation catchments (Allenbach and Riale di Calneggia) have the largest seasonality in streamflow, i.e. the biggest differences between the long-term maximum and minimum Pardé coefficients, followed by the rain and snow dominated Verzasca catchment. The rain dominated catchments (Murg, Guerbe and Mentue ~~consequently~~) have the lowest seasonal variability in streamflow (Table 1). The mean elevation of the catchments varies from 652 to 2003 m asl (Table 1). The elevation range of each individual catchment was divided in 100 m elevation bands for the simulations.

2.3 Measured data

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). The average hourly areal precipitation amounts were extracted for each study catchment from the gridded CombiPrecip dataset from MeteoSwiss (Sideris et al., 2014). This dataset combines gauge and radar precipitation measurements at an hourly timescale and 1 km^2 spatial resolution and is available since 2005.

We used hourly temperature data from the automatic monitoring network of MeteoSwiss (see Table 1 for the stations) and applied a gradient of -6 °C per 1000 m to adjust the temperature of each weather station to the mean elevation of the catchment. Within the HBV model, the temperature was then ~~adapted~~adjusted for the different elevation ~~zones~~bands using a calibrated lapse rate.

As recommended by Oudin et al. (2005), potential evapotranspiration was calculated using the temperature-based potential evapotranspiration model of McGuinness and Bordne, (1972) using the day of the year, the latitude and the temperature. This rather simplistic approach was considered sufficient because this study focused on differences in model performance relative to a benchmark calibration.

2.4 Selection of years for model calibration and validation

The model was calibrated for an average, a dry and a wet year to investigate the influence of wetness conditions and the amount of streamflow on the calibration results. The years were selected based on the total streamflow during summer (July-September). The driest and the wettest years of the period 2006-2014 were selected based on the smallest and largest sum of streamflow during the summer. The average streamflow years were selected based on the proximity to the mean summer streamflow for all the years 1974-2014 (1990-2014 for Verzasca). For each catchment the years that were the 2^{nd}-closest to the mean summer streamflow for all years, as well as the years with the 2^{nd}-lowest and 2^{nd}-highest streamflow sum were chosen for model calibration (see Table 2). We did this separately for each catchment because for each catchment a different year was dry, average or wet. For the validation, we chose the year closest to the mean summer streamflow and the years with the lowest and the highest total summer streamflow ~~sums~~ (see Table 2). We used each of the parameter sets obtained from calibration for the dry, average or wet years to validate the model for each of the three validation years, resulting in nine validation combinations for each catchment (and each dataset, as described below).

2.5 Transformation of data sets to resemble citizen science data quality

2.5.1 Errors in crowdsourced streamflow observations

Strobl et al. (in review) asked 517 participants to estimate streamflow based on the stick method at ten streams in Switzerland. Here we use the estimates for the medium sized streams Töss, Sihl and Schanzengraben in the Canton of Zurich and the Magliasina in Ticino (n=136), which had a similar streamflow range at the time of the estimations (2.6 – 28 m^3/s) as the mean annual streamflow of the six streams used for this study (1.2 – 10.8 m^3/s). We calculated the streamflow from the estimated width, depth and flow velocities using a factor of 0.8 to adjust the surface flow velocity to the average velocity (Harrelson et al., 1994). The resulting streamflow estimates were normalized by dividing them by the measured ~~value~~streamflow. We then combined the normalized estimates of all four rivers and log transformed the relative estimates. A normal distribution with a mean of 0.12 and a standard deviation of 1.30 fits the distribution of the log-transformed relative estimates well, with a standard error of the mean of 0.11 and a standard error of the standard deviation of 0.08 (Figure 2).

To create synthetic datasets with data quality characteristics that represent the observed crowdsourced streamflow estimates, we assumed that the errors in the streamflow estimates are uncorrelated (as they ~~would~~are likely ~~be~~ provided by different people). For each time step, we randomly selected a relative error value from the lognormal distribution of the relative estimates (Figure 2) and multiplied the measured streamflow with this relative error. To simulate the effect of training and to obtain time series with different data quality, two ~~more~~additional streamflow time series were created using a standard deviation divided by two (standard deviation of 0.65) ~~for the medium error~~ and by four (standard deviation of 0.33~~) for the small error.~~). This reduces the spread in the data (but did not change the small systematic overestimation of the streamflow), so that large outliers are still possible, but are less likely. ~~The benchmark in terms of quality were the no error datasets for which we used the FOEN data directly.~~To summarize, we tested the following four cases:

- *No error*: The data measured by the FOEN, assumed to be (almost) 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 surveys 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 surveys.

2.5.2 Filtering of extreme outliers

Usually ~~citizen science data undergo~~ some form of quality control is used before ~~they~~ citizen science data are analyzed. Here, we used a very simple check to remove unrealistic outliers from the synthetic datasets. This check was based on the ~~upper limit of~~ likely minimum and maximum streamflow ~~values~~ for a given catchment area. We defined an upper limit of possible streamflow values as a function of catchment area using ~~a~~the dataset of maximum streamflow from 1500 Swiss catchments provided by Scherrer AG, Hydrologie und Hochwasserschutz (2017). To account for the different precipitation intensities

north and south of the Alps, different ~~limits~~curves were ~~used~~created for the catchments on each side of the Alps. All streamflow observations, i.e., modified streamflow measurements, above the maximum observed streamflow for a particular catchment size including a 20 % buffer (), were replaced by the value of the maximum streamflow for a catchment of that size. This affected less than 0.5 % of all data points. A similar procedure was used for low flows based on a dataset of the FOEN with the lowest recorded mean streamflows over seven days but this resulted in no replacements.

2.5.3 Temporal resolution of the observations

Data entries from citizen scientists are not as regular as data from sensors with a fixed temporal resolution. Therefore, we decided to test eight scenarios with a different temporal resolution and a different distribution of the data throughout the year to simulate different patterns in citizen contributions~~. These scenarios were based on our own experiences within the CrowdWater project () and information from the CrowdHydrology project .~~:

~~We used the same selection of days, including the same times of the day for each of the four different error groups, years and catchments to allow comparison of the different model results.~~

- *Hourly*: One data point per hour ($8760 \leq n \leq 8784$, depending on the year)
- *Weekly*: One data point per week, every Saturday, randomly between 6 am and 8 pm ($52 \leq n \leq 53$)
- *Monthly*: One data point per month on the 15th of the month, randomly between 6 am and 8 pm (n=12)
- *IntenseSummer*: One data point every other day from July until September, randomly between 6 am and 8 pm (~15 observations per month, n=46).
- *WeekendSummer*: one data point each Saturday and each Sunday between May and October, randomly between 6 am and 8 pm ($52 \leq n \leq 54$)
- *WeekendSpring*: one data point on each Saturday and each Sunday between March and August, randomly between 6 am and 8 pm ($52 \leq n \leq 54$)
- ~~In addition, we also tested two scenarios (*Crowd52* and *Crowd12*) with a random interval between data points. *Crowd52* had 52~~*Crowd52* had 52 random data points (in order to be comparable to the *Weekly*, *IntenseSummer*, *WeekendSummer* and *WeekendSpring* time series~~), whereas~~)
- *Crowd12* had ~~only~~ 12 random data points (comparable to the *Monthly* data).

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 test the effect of errors when the temporal resolution of the data is optimal (i.e., by comparing simulations for the models calibrated with the hourly FOEN data and those calibrated with hourly data with errors). In the two scenarios Crowd 52 and Crowd12 with random intervals between data points we assigned higher probabilities for periods when people are more likely to be outdoors (i.e., higher probabilities for summers than winters, higher probabilities for weekends than weekdays, higher probabilities outside office hours; Table 4). Times without daylight (dependent on the season) were always

excluded. We used the same selection of days, including the same times of the day for each of the four different error groups, years and catchments to allow comparison of the different model results.

2.6 Model calibration

For each of the 1728 cases (6 catchments, 3 calibration years, 4 error groups, 8 temporal resolutions) the HBV model was calibrated by optimizing the overall consistency performance P_{OA} (Finger et al., 2011) using a genetic optimization algorithm (Seibert, 2000). The overall consistency performance P_{OA} is the mean of four objective functions with an optimum value of one: *i*) ~~the Nash-Sutcliffe efficiency (~~NSE~~),~~ *ii*) the NSE for the logarithm of streamflow, *iii*) the volume error, and *iv*) the mean absolute relative error (MARE). The parameters were calibrated within their typical ranges (see Supplemental Material - Table 1). To consider parameter uncertainty, the calibration was performed ~~independently~~ 100 times, which resulted in 100 parameter sets for each case. For each case, the preceding year was used for the warm-up period. For the *Crowd52* and *Crowd12* time series, we used 100 different random selections of times for which data were available, whereas for the regularly spaced time series the same times were used for each case.

2.7 Model validation and analysis of the model results

The 100 parameters from the calibration for each ~~data set~~case were used to run the model for the validation years (Table 2). For each case (i.e. each catchment, year, error magnitude and temporal resolution), we determined the median validation P_{OA} for the 100 parameter sets for each validation year. We analysed the validation results of all years combined and for all nine combinations of dry, mean and wet years separately.

Because the focus of this study was on the value of limited inaccurate streamflow observations for model calibration, i.e. the difference in the performance of the models calibrated with the synthetic data series compared to the performance of the models calibrated with hourly FOEN data, all model validation performances are expressed relative to the average P_{OA} of the model calibrated with the hourly FOEN data (our upper benchmark, representing the fully informed case when continuous high quality streamflow data are available). A relative P_{OA} of 1 indicates that the model performance is as good as the performance of the model calibrated with the hourly FOEN data, whereas lower P_{OA} values indicate a poorer performance.

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 also~~. To assess the value of limited inaccurate streamflow data for model calibration compared to a situation without any streamflow data, a lower benchmark (Seibert et al., 2018) was used. 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 ~~random parameters~~ randomly chosen parameter sets, we represent a situation where no streamflow data for ~~every catchment~~calibration are available and ~~year.~~ 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. The Mann Whitney U-Test was used to evaluate whether the median P_{OA} for a specific error group and temporal resolution of the data was significantly different from the median P_{OA}

for the lower benchmark (i.e. the model ~~run~~runs with random parameters~~.~~). We furthermore checked for differences in model performance for models calibrated with the same data errors but ~~a~~ different temporal ~~resolution~~resolutions using a Kruskal-Wallis test. By applying a Dunn-Bonferroni post-hoc test (Bonferroni, 1936; Dunn, 1959, 1961) we analysed which of the validation results were significantly different from each other.

~~showed that the use of monthly maximum streamflow data in model calibration leads to a better model performance than streamflow data for more average conditions.~~ The random generation of the 100 crowdsourced-like datasets (i.e. for the *Crowd52* and *Crowd12* scenario) for each of the catchments and year characteristics resulted in time series with a different number of high flow estimates. In order to find out whether the inclusion of more high flow values resulted in a better validation performance, we defined the threshold for high flows as the streamflow value that was exceeded 10 % of the time in the hourly FOEN streamflow dataset. The *Crowd52* and *Crowd12* datasets were then divided into a group that had more than the expected 10 % high flow observations and a group that had fewer high flow observations. To determine if more high flow data improves model performance, the Mann-Whitney-U-test was used to compare the relative median P_{OA} of the two groups.

3 Results

3.1 Upper benchmark results

The model was able to reproduce the measured streamflow reasonably well when the complete and unchanged hourly-FOEN datasets were used for calibration, although there were also a few exceptions. The average validation P_{OA} was 0.61 (range: 0.19 – 0.83; Table 3). The validation performance was poorest for the Guerbe ~~had the lowest (~~validation P_{OA} ~~(~~= 0.19) because several high flow peaks were missed or underestimated by the model for the wet validation year. Similarly, the validation for the Mentue for the dry validation year ~~2009 for the Mentue~~ resulted in a low P_{OA} (0.23) because a very distinct peak at the end of the year was missed and summer low flows were overestimated. The third lowest P_{OA} value was ~~again from~~also for the Guerbe (~~2013,~~ dry validation year) but already had a P_{OA} of 0.35. Six out of the nine lowest P_{OA} values were for dry validation years. Validation for wet years for the models calibrated with data from wet years resulted in the best validation results (i.e., highest P_{OA} values; Table 3).

3.2 Effect of errors on the model validation results

Not surprisingly, increasing the errors in the streamflow data ~~decreased~~used for model calibration led to a decrease in the model performance (Figure 4). For the small error category, the median validation performance was better than the lower benchmark for all temporal resolutions (Figure 4 and Supplemental Material - Table 2). For the medium error category, the median validation performance ~~of all scenarios~~ was also better than the lower benchmark for all scenarios, except for the *Crowd12* dataset. For the model calibrated with the dataset with large errors only the *Hourly* data set was significantly better than the lower benchmark (Table 5).

3.3 Effect of the ~~measurement~~data resolution on the model validation results

The *Hourly* measurement scenario resulted in the best validation performance for each error group, followed by the *Weekly* data, and then usually the *Crowd52* data (Figure 4). Although the median ~~model~~validation performance of the models calibrated with the *Weekly* datasets was ~~always~~ better than for the *Crowd52* dataset for all error cases, the difference was only statistically significant for the no error category (Figure 5).

The validation performance of the models calibrated with the *Weekly* and *Crowd52* datasets was better than for the ~~measurement~~ scenarios focused on spring and summer observations (*WeekendSpring*, *WeekendSummer* and *IntenseSummer*). The median model performance for the *Weekly* dataset was significantly better than the datasets focusing on spring and summer for the no, small and medium error groups. The median performance of the *Crowd52* dataset was only significantly better than all three measurement scenarios focusing on spring or summer for the small error case (Figure 5). The model validation performance for the *WeekendSummer* and *IntenseSummer* scenarios decreased faster with increasing ~~error~~errors compared to the *Weekly*, *Crowd52* or *WeekendSpring* datasets (Figure 5).

~~The median model performance for the *Weekly* dataset was significantly better than the other datasets for the no, small and medium error groups; the median performance of the *Crowd52* dataset was only significantly different from the measurement scenarios focusing on spring or summer for the medium error case ().~~ The median validation P_{OA} ~~was better~~ for the models calibrated with the *WeekendSpring* observations was better than for the ~~model~~models calibrated with the *WeekendSummer* and *IntenseSummer* datasets but the differences were only significant for the small, medium and large error groups. The differences in the model performance results ~~of~~for the observation strategies ~~focussing~~focussed on summer (*IntenseSummer* and *WeekendSummer*) were not ~~significantly different in~~significant for any of the error groups (Figure 5).

The median model performance for the regularly spaced *Monthly* datasets with 12 observations was similar to the median performance for the three datasets focusing on summer with 46-54 measurements (*WeekendSpring*, *WeekendSummer* and *IntenseSummer*), except for the case of large errors for which the monthly dataset performed worse. The irregularly spaced *Crowd12* time series resulted in the worst model performance for each error group but the difference from the performance for the regularly spaced *Monthly* data was only significant for the dataset with large errors.

3.4 Effect of errors and data resolution on the parameter ranges

For most parameters the spread in the optimized parameter values was smallest for the upper benchmark. The spread in the parameter values increased with increasing errors in the data used for calibration, particularly for MAXBAS (the routing parameter) but also for some other parameters (e.g. TCALT, TT and BETA). However, for some parameters (e.g., CFMAX, FC, and SFCF) the range in the optimized parameter values was mainly affected by the temporal resolution of the data and the number of data points used for calibration. It should be noted though that the changes in the range of model parameters differed significantly for the different catchments and the trends weren't very clear.

~~3.4~~3.5 Influence of the calibration and validation year and number of high flow data points on the model performance

The influence of the validation year on the model performance was larger ~~for~~ than the effect of the calibration year (Figure 6 and Supplemental Material – Figure 2). In general model performance was poorest for the dry validation years. The model performances of all datasets with fewer observations or bigger errors than the *Hourly* datasets without errors were not significantly better than the lower benchmark for the dry validation years, except for ~~the~~ *Crowd52* in the no error group when calibrated with data from a wet year. However, even for the wet validation years some observation scenarios of the no error and small error group did not lead to significantly better model validation results compared to the median validation ~~results~~performance for the random parameters. Interestingly, the *IntenseSummer* data set in the no error group resulted in ~~some~~a very good ~~performances~~performance when the model was calibrated for a dry and also validated in a dry year compared to its performance in the other calibration and validation year combinations. The median model performance was however not significantly better than the lower benchmark due to the ~~two very~~ low performances ~~of~~for the Guerbe and Allenbach (outliers beyond figure margins in Figure 6). The validation results of one of these two catchments always resulted in the worst performance for all the no error - *IntenseSummer* datasets for all calibration and validation year combinations.

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 ~~model~~validation performance than the Crowd52 ~~dataset~~datasets with more high -streamflow data points. For six of them the difference in model performance was significant. For none of the five cases where more high flow data points led to a better model performance was the difference significant. Also when the results were analysed by year character or catchment there was no improvement when more high flow values were ~~present~~included in the calibration dataset.

4 Discussion

4.1 Usefulness of inaccurate streamflow data for hydrological model calibration

~~If~~In this study, we evaluated the information content of streamflow estimates by citizen scientists for calibration of the bucket-type hydrological model for six Swiss catchments. Streamflow estimates by citizens are sometimes very differ-ent from the measured values, and the individual estimates can be dis-informative for model calibration (Beven, 2016; Beven and Westerberg, 2011). While the hydroclimatic conditions, the model or the calibration approaches might be different in other studies, these results should be applicable for a wide range of cases. However, for physically-based spatially distributed models that are usually not calibrated automatically, the use of limited streamflow data would probably benefit from a different calibration approach. Furthermore, our results might not be applicable in arid catchment cases where rivers fall dry for some period of the year because the linear reservoirs used in the HBV model are not appropriate for such systems. The

results show that if the streamflow estimates by citizen scientists would be available at a high temporal resolution (hourly~~)~~), these data are informative for the calibration of a bucket-type hydrological model ~~calibration~~ despite their high uncertainties. However, ~~such detailed~~ observations with such a high resolution are very unlikely to be obtained in practice. All ~~the~~ scenarios with error distributions that represent the estimates from citizen scientists with fewer observations were no better than the lower benchmark (using random parameters). ~~Streamflow estimates are sometimes very different from the measured values, and individual estimates can be dis-informative .~~ With ~~reduced~~medium errors, however, and one data point per week on average or regularly spaced monthly data, the ~~time series with the medium errors~~data were informative for model parameterization. Reducing the standard deviation of the error-distribution by a factor of four, led to significantly improved model performance for all the observation scenarios compared to the ~~random parameter datasets. In reality this~~lower benchmark.

A reduction in the errors of the streamflow estimates could be ~~done~~achieved by training of citizen scientists~~,~~ (e.g. videos), improved information about feasible ~~value~~ ranges for stream depth, with and velocity, or examples of streamflow values for ~~a given stream. Furthermore,~~well-known streams. Filtering of extreme outliers ~~has the potential to~~can also reduce the spread of the estimates. This could be done with existing knowledge of feasible streamflow values for a catchment of a given area or the amount of rainfall right before the estimate is made to determine if streamflow is likely to be higher or lower than ~~earlier estimates.~~for the previous estimate. More detailed research is necessary to test the effectiveness of such methods.

Le Coz et al. (2014) reported an uncertainty in stage-discharge streamflow measurements of around 5-20 ~~%, whereas~~%. McMillan et al. (2012) ~~in a more detailed review~~ summarized streamflow uncertainties from stage-discharge relationships in a more detailed review and gave a range of ±50-100 % for low flows, ±10-20 % for medium or high (in-bank) flows and ±40 % for out-of-bank flows. The errors for the most extreme outliers in the citizen estimates are considerably higher, as they can differ by a factor of up to 10’000 from the measured value in the most extreme but rare cases (Figure 2). Even with reduced standard deviations of the error distribution by a factor of two or four, the observations in the most extreme cases can still differ by a factor of 100 and 10. The percentage of values beyond 200 % of the measured value in the synthetic datasets with streamflow observations was 33 % for the large error group, 19 % in the medium error group and 4 % in the small error group. Only 3 % were more than 90 % below the measured value in the large error group and 0 % ~~in~~for both in the medium and small error classes. If such observations are used for model calibration without filtering, they are seen as extreme droughts or floods, even if the actual conditions may be close to average flow. Beven and Westerberg (2011) suggest to isolate periods of dis-informative data. It is therefore beneficial to identify such extreme outliers, independent of a model, e.g. with ~~previous~~ knowledge of feasible maximum and minimum streamflow quantities, as ~~performed~~used in this study, with the help of the maximum regionalized specific streamflow values for a given catchment area.

4.2 Number of streamflow estimates required for model calibration

In general, one would assume that the calibration of a model becomes better when there is more data (Perrin et al., 2007), although others have shown that the increase in model performance plateaus after a certain number of measurements (Juston

et al., 2009; Pool et al., 2017; Seibert and Beven, 2009; Seibert and McDonnell, 2015). In this study, we limited the length of the calibration period to one year because in practice it may be possible to obtain a limited number of measurements during a one year period for ungauged catchments before the model results are needed for a ~~practical~~certain application, as has been assumed in previous studies (Pool et al., 2017; Seibert and McDonnell, 2015). While a limited number of observations (12) was informative for model calibration when the data uncertainties were limited, the results of this study also suggest that the performance of ~~the models calibrated with fewer data points decreased faster as the errors increased.~~bucket-type 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). This finding was most pronounced when comparing the model performance for the small and the medium error groups (Figure 4). These findings can be explained by the compensating effect of the number of observations and their accuracy because the random errors for ~~less accurate~~the inaccurate data average out when a ~~larger~~large number of observations are used, as long as the data do not have a large bias.

4.3 Best timing of streamflow estimates for model calibration

The performance of the parameter sets depended on the ~~observation~~ timing and the error distribution of the data used for model calibration. The model performance was generally better if the observations were more evenly spread throughout the year. For example for the cases of no and small errors, the ~~model~~ performance ~~for~~of the model calibrated with the *Monthly* dataset with 12 observations ~~performed~~was better than for the *IntenseSummer* and *WeekendSummer* scenarios with 46-54 observations. Similarly, the less clustered observation scenarios performed better than the more clustered scenarios (i.e. *Weekly* vs. *Crowd52*, *Monthly* vs. *Crowd12*, *Crowd52* vs. *IntenseSummer*, etc.). This suggests that more regularly distributed data over the year leads to a better model calibration. Juston et al. (2009) compared different subsamples of hydrological data for a 5.6 km^2 Swedish catchment and found that including inter-annual variability in the data used for the calibration of the HBV model reduced the model uncertainties. More evenly distributed observations throughout the year might represent more of the within-year streamflow variability and therefore result in improved model performance. This is good news for using citizen science data for model calibration as it suggests that the timing is not as important as the number of observations because it is likely much easier to get ~~a certain number of~~ observations throughout the year than ~~observations~~ during specific periods or flow conditions.

When comparing the *WeekendSpring*, *WeekendSummer* and *IntenseSummer* datasets, it seems that it was in most cases more beneficial to include data from spring rather than summer. This tendency was more pronounced with increasing data errors. The reason for this might be that the *WeekendSpring* scenario includes more snow melt or rain-on-snow event peaks, in addition to usually higher baseflow ~~values~~ and therefore contains more information on the inter-annual variability in streamflow.

By comparing different variations of 12 data points to calibrate the HBV model, Pool et al. (2017) found that ~~more high flow data points resulted in an improved model performance. In our study, this could not be observed, which might be due to the~~

~~fact that~~a dataset that contains a combination of different maximum (monthly, yearly etc.) and other flows in model calibration led to the best model performance but also that the differences between the different datasets covering the range of flows were small. In our study we did not specifically focus on the high or low flow data points, and therefore did not have datasets that contained only high flow estimates, which would be very difficult to obtain with citizen science data. ~~Here we tested only scenarios~~However, our findings similarly show that ~~are realistic~~ for ~~citizen science projects. In contrast to, we conclude that~~model calibration for catchments with seasonal variability in streamflow it is beneficial to obtain ~~streamflow~~ data ~~of all~~for different magnitudes ~~for model calibration.~~of flow. Furthermore, we found that data points during relatively dry periods are beneficial for validation or prediction in another year and might even be beneficial for years with the same characteristics, as was shown ~~with~~for the improved validation performance of the *IntenseSummer* dataset compared to the other datasets when data from dry years were used for calibration (Figure 6).

4.4 Effects of different types of years on model calibration and validation

The calibration year, i.e. the year in which the observations were made, was not decisive for the model performance. Therefore, a model calibrated with data from a dry year can still be useful for simulations ~~in~~for an average or wet years. This also means that data in citizen science projects can be collected during any year and that this data is useful for simulating ~~the~~ streamflow for most years, except the driest years. However, model performance did vary significantly for the different validation years. The results during dry validation years were almost never significantly better than the lower benchmark (Supplemental Material – Figure 2). This might be due to the objective function that was used in this study. Especially the NSE was lower for dry years, because the flow variance (i.e., the denominator in the equation) is smaller when there is a larger variation in streamflow. Also, these results are based on six median model performances and therefore, outliers have a big influence on the significance of results (Supplemental Material – Figure 2).

Lidén and Harlin (2000) used the HBV-96 model by Lindström et al. (1997) with changes suggested by Bergström et al. (1997) for four catchments in Europe, Africa and South America. They achieved better model results for wetter catchments and argued that during dry years evapotranspiration plays a bigger role and therefore the model performance is more sensitive to inaccuracies in ~~processes concerning~~the simulation of the evapotranspiration processes. The fact that we used a very simple method to calculate the potential evapotranspiration (McGuinness and Bordne, 1972), might also explain why the model performed less well during dry years.

The model parametrisation, obtained from calibration using the *IntenseSummer* data set resulted in a surprisingly good performance for the validation for a more extreme dry year for four out of the six catchments. For the two ~~poorly performing~~ catchments for which the performance for the *IntenseSummer* data set was poor (Guerbe and Allenbach~~,~~), the weather stations are located outside the catchment boundaries. Especially during dry periods missed streamflow peaks due to misrepresentation of precipitation can affect model performance a lot. The fact that always one of these two catchments had the worst model performance for all the no error – *IntenseSummer* runs, furthermore indicates that the July-September period might not be suitable to represent characteristic runoff events for these catchments. The bad performance for these two

catchments ~~resulted in the insignificant improvements compared to the lower benchmark of~~for the *IntenseSummer* – no error run with calibration and validation in the dry year resulted in the insignificant improvement in model performance compared to the lower benchmark. Because the wetness of a year was based on the summer streamflow, these findings suggest that data obtained during times of low flow, result in improved validation performance during dry years compared to data collected during other times (Supplemental Material – Figure 2). This suggests that if the interest is in understanding the streamflow response during very dry years, it is important to obtain data during the dry period. To ~~verify~~test this ~~assumption~~hypothesis more detailed analyses are needed.

4.5 Recommendations for citizen science projects

Our results show that streamflow estimates from citizens are not informative for hydrological model calibration, unless the errors in the estimates can be reduced through training or ~~advance~~advanced filtering of the data to reduce the errors (i.e. to reduce the number of extreme outliers). In order to make streamflow estimates useful, the standard deviation of the estimation-error-distribution needs to be reduced by a factor of two. ~~Research of~~ Gibson and Bergman (1954) ~~suggests~~suggest that ~~the error~~errors in distance ~~estimations~~estimates can be reduced from 33 % to 14 % with very little training. ~~Those~~These findings are encouraging, although their tests covered distances larger than 365 meters (400 yards) and the widths of the medium sized rivers ~~of~~ for which the streamflow was estimated were less than 40 meters (Strobl et al., in review). ~~In order to determine the effect of~~. Options for training ~~on streamflow estimates further research has to~~might be ~~done because especially~~tutorial videos, as well lists with values for the width, average depth ~~estimates are very inaccurate~~and flow velocity of well-known streams (Strobl et al., in review). In order to determine the effect of training on streamflow estimates further research has to be done because especially the depth estimates were inaccurate (Strobl et al., in review).

The findings of this study suggest the following recommendations for citizen science projects that want to use streamflow estimates:

- Collect as much data as possible: In this study hourly data always led to the best model performance. It is therefore beneficial to collect as much data as possible. Because it is unlikely to obtain hourly data, we suggest to aim for (on average) one observation per week. Provided that the standard deviation of the streamflow estimates can be reduced by a factor of two, 52 observations (as in the *Crowd52* data series) are informative for model calibration. Therefore, it is essential to invest in advertisement of a project and to find suitable locations where many people can potentially contribute, as well as to communicate to the citizen scientists that it is beneficial to submit observations regularly.
- Encourage ~~evenly distributed~~ observations throughout the year: To further improve the model performance, or to allow for greater errors, it is beneficial to have observations at all types of flow conditions during the year~~.~~, rather than during a certain season.

Observations during high streamflow conditions were in most cases not more informative than flows during other times of the year. Efforts to ask citizens to submit observations during specific flow conditions (e.g. by sending reminders to the

citizen observers) do not seem very effective in light of the above findings. It is rather more beneficial to remind them to submit observations regularly.

Instead of focussing on training to reduce the errors in the streamflow estimates, an alternative approach for citizen science projects is to switch to a parameter that is easier to estimate, such as stream levels (Lowry and Fienen, 2013). Recent studies successfully used daily stream level data (Seibert and Vis, 2016) and stream level class data (van Meerveld et al. 2017) to calibrate hydrological models, and other studies demonstrated the potential value of crowdsourced stream level data for providing information on e.g. baseflow (Lowry and Fienen, 2013) or to improve flood forecasts (Mazzoleni et al., 2017). However, further research is needed to determine if real crowdsourced stream level (class-) data is informative for the calibration of hydrological models.

5 Conclusions

The results of this study extend previous studies on the value of limited hydrological data for hydrological model calibration or the best timing of streamflow measurements for model calibration (Juston et al., 2009; Pool et al., 2017; Seibert and McDonnell, 2015) that did not consider observation errors. This is an important aspect, especially when considering citizen science approaches to obtain streamflow data. Our results show that inaccurate streamflow data can be useful for model calibration, as long as the errors are not too large. When the distribution of errors in the streamflow data represented the distribution of the errors in the ~~estimates of~~ streamflow estimates from citizen scientists, this information was not informative for model calibration (i.e. the median performance of the models calibrated with ~~this~~these data was not ~~statistically significant~~significantly better than the median performance of ~~a model~~the models with random parameter values). However, if the standard deviation of the estimates is reduced by a factor two, then the (less) inaccurate data would be informative for model calibration. We, furthermore, demonstrated that realistic frequencies for citizen science projects (one observation on average per week or month) can be informative for model calibration. The findings of studies such as the one presented here provide important guidance on the design of citizen science projects, and also other, observation approaches.

6 Author contribution

While Jan Seibert and Ilja van Meerveld had the initial idea, the concrete study design was ~~designed~~ based on input from all authors. Simon Etter and Barbara Strobl conducted the field surveys to ~~collect~~determine the typical errors in the streamflow estimates. The simulations and analyses were performed by Simon Etter. The writing of the manuscript was ~~written~~led by Simon Etter ~~with valuable comments of~~ ; all co-authors contributed to the writing.

7 Data availability

The data are available from FOEN (streamflow) and MeteoSwiss (precipitation and temperature). The HBV software is available from https://www.geo.uzh.ch/en/units/h2k/Services/HBV-Model.html.

8 Acknowledgements

We thank all citizen scientists who participated in the field surveys, as well as the Swiss Federal Office for the Environment for providing the streamflow data, MeteoSwiss for providing the weather data, ~~and~~ Maria Staudinger, Jan Schwanbeck and Scherrer AG for the permission to use their datasets~~.~~, and the reviewers for the useful comments. This project was funded by the Swiss National Science ~~Fund~~Foundation (project CrowdWater).

References

Aschwanden, H. and Weingartner, R.: Die Abflussregimes der Schweiz, Publ. Gewässerkd., 65, 237, 1985.

Bergström, S.: Development and application of a conceptual runoff model for Scandinavian catchments, Bull. Ser. A, 52, 134, 1976.

Bergström, S., Carlsson, B., Grahn, G. and Johansson, B.: A More Consistent Approach to Watershed Response in the HBV Model., Vannet i Nord., 4, 1997.

Beven, K. and Westerberg, I.: On red herrings and real herrings: disinformation and information in hydrological inference, Hydrol. Process., 25(10), 1676–1680, doi:10.1002/hyp.7963, 2011.

Beven, K. J. and Kirkby, M. J.: A physically based, variable contributing area model of basin hydrology / Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant, Hydrol. Sci. Bull., 24(1), 43–69, doi:10.1080/02626667909491834, 1979.

Bonferroni, C. E.: Teoria statistica delle classi e calcolo delle probabilità., st. Super. di Sci. Econom. e Commerciali di Firenze., 1936.

Brath, A., Montanari, A. and Toth, E.: Analysis of the effects of different scenarios of historical data availability on the calibration of a spatially-distributed hydrological model, J. Hydrol., 291(3–4), 232–253, doi:10.1016/j.jhydrol.2003.12.044, 2004.

Buytaert, W., Zulkafli, Z., Grainger, S., Acosta, L., Alemie, T. C., Bastiaensen, J., De BiÃ¨vre, B., Bhusal, J., Clark, J., Dewulf, A., Foggin, M., Hannah, D. M., Hergarten, C., Isaeva, A., Karpouzoglou, T., Pandeya, B., Paudel, D., Sharma, K., Steenhuis, T., Tilahun, S., Van Hecken, G. and Zhumanova, M.: Citizen science in hydrology and water resources: opportunities for knowledge generation, ecosystem service management, and sustainable development, Front. Earth Sci., 2, doi:10.3389/feart.2014.00026, 2014.

Le Coz, J., Renard, B., Bonnifait, L., Branger, F. and Le Boursicaud, R.: Combining hydraulic knowledge and uncertain

gaugings in the estimation of hydrometric rating curves: A Bayesian approach, J. Hydrol., 509, 573–587, doi:10.1016/j.jhydrol.2013.11.016, 2014.

Davids, J. C., van de Giesen, N. and Rutten, M.: Continuity vs. the Crowd—Tradeoffs Between Continuous and Intermittent Citizen Hydrology Streamflow Observations, Environ. Manage., 60(1), 12–29, doi:10.1007/s00267-017-0872-x, 2017.

Davids, J. C., Rutten, M. M., Shah, R. D. T., Shah, D. N., Devkota, N., Izeboud, P., Pandey, A. and Giesen, N. Van De: Quantifying the connections — linkages between land-use and water in the Kathmandu Valley , Nepal, Environ. Monit. Assess., 190(5), doi:10.1007/s10661-018-6687-2, 2018.

Dickinson, J. L., Zuckerberg, B. and Bonter, D. N.: Citizen Science as an Ecological Research Tool: Challenges and Benefits, Annu. Rev. Ecol. Evol. Syst., 41(1), 149–172, doi:10.1146/annurev-ecolsys-102209-144636, 2010.

Dunn, O. J.: Estimation of the Medians for Dependent Variables, Ann. Math. Stat., 30(1), 192–197, doi:10.1214/aoms/1177706374, 1959.

Dunn, O. J.: Multiple Comparisons among Means, J. Am. Stat. Assoc., 56(293), 52–64, doi:10.1080/01621459.1961.10482090, 1961.

Ewen, T., Brönnimann, S. and Annis, J.: An extended Pacific-North American index from upper-air historical data back to 1922, J. Clim., 21(6), 1295–1308, doi:10.1175/2007JCLI1951.1, 2008.

Finger, D., Pellicciotti, F., Konz, M., Rimkus, S. and Burlando, P.: The value of glacier mass balance, satellite snow cover images, and hourly discharge for improving the performance of a physically based distributed hydrological model, Water Resour. Res., 47(7), n/a-n/a, doi:10.1029/2010WR009824, 2011.

Finger, D., Vis, M., Huss, M. and Seibert, J.: The value of multiple data set calibration versus model complexity for improving the performance of hydrological models in mountain catchments, Water Resour. Res., 51(4), 1939–1958, doi:10.1002/2014WR015712, 2015.

Fitzner, D., Sester, M., Haberlandt, U. and Rabiei, E.: Rainfall Estimation with a Geosensor Network of Cars – Theoretical Considerations and First Results, Photogramm. - Fernerkundung - Geoinf., 2013(2), 93–103, doi:10.1127/1432-8364/2013/0161, 2013.

Gibson, E. J. and Bergman, R.: The effect of training on absolute estimation of distance over the ground., J. Exp. Psychol., 48(6), 473–482, doi:10.1037/h0055007, 1954.

Haberlandt, U. and Sester, M.: Areal rainfall estimation using moving cars as rain gauges – a modelling study, Hydrol. Earth Syst. Sci., 14(7), 1139–1151, doi:10.5194/hess-14-1139-2010, 2010.

Harrelson, C. C., Rawlins, C. L. and Potyondy, J. P.: Stream channel reference sites: an illustrated guide to field technique., 1994.

Horner, I., Renard, B., Le Coz, J., Branger, F., McMillan, H. K. and Pierrefeu, G.: Impact of Stage Measurement Errors on Streamflow Uncertainty, Water Resour. Res., 54(3), 1952–1976, doi:10.1002/2017WR022039, 2018.

Juston, J., Seibert, J. and Johansson, P.: Temporal sampling strategies and uncertainty in calibrating a conceptual

hydrological model for a small boreal catchment, Hydrol. Process., 23(21), 3093–3109, doi:10.1002/hyp.7421, 2009.

Koch, J. and Stisen, S.: Citizen science: A new perspective to advance spatial pattern evaluation in hydrology, PLoS One, 12(5), 1–20, doi:10.1371/journal.pone.0178165, 2017.

Lidén, R. and Harlin, J.: Analysis of conceptual rainfall–runoff modelling performance in different climates, J. Hydrol., 238(3–4), 231–247, doi:10.1016/S0022-1694(00)00330-9, 2000.

Lindström, G., Johansson, B., Persson, M., Gardelin, M. and Bergström, S.: Development and test of the distributed HBV-96 hydrological model, J. Hydrol., 201(1–4), 272–288, doi:10.1016/S0022-1694(97)00041-3, 1997.

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

Mazzoleni, M., Verlaan, M., Alfonso, L., Monego, M., Norbiato, D., Ferri, M. and Solomatine, D. P.: Can assimilation of crowdsourced data in hydrological modelling improve flood prediction?, Hydrol. Earth Syst. Sci., 21(2), 839–861, doi:10.5194/hess-21-839-2017, 2017.

McGuinness, J. and Bordne, E.: A comparison of lysimeter-derived potential evapotranspiration with computed values., 1972.

McMillan, H., Freer, J., Pappenberger, F., Krueger, T. and Clark, M.: Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions, Hydrol. Process., 24(10), 1270–1284, doi:10.1002/hyp.7587, 2010.

McMillan, H., Krueger, T. and Freer, J.: Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality, Hydrol. Process., 26(26), 4078–4111, doi:10.1002/hyp.9384, 2012.

van Meerveld, H. J., Vis, M. J. P. and Seibert, J.: Information content of stream level class data for hydrological model calibration, Hydrol. Earth Syst. Sci., 21(9), 4895–4905, doi:10.5194/hess-21-4895-2017, 2017.

Michel, C., Perrin, C. and Andreassian, V.: The exponential store: a correct formulation for rainfall—runoff modelling, Hydrol. Sci. J., 48(1), 109–124, doi:10.1623/hysj.48.1.109.43484, 2003.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F. and Loumagne, C.: Which potential evapotranspiration input for a lumped rainfall–runoff model?, J. Hydrol., 303(1–4), 290–306, doi:10.1016/j.jhydrol.2004.08.026, 2005.

Perrin, C., Michel, C. and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, J. Hydrol., 279(1–4), 275–289, doi:10.1016/S0022-1694(03)00225-7, 2003.

Perrin, C., Ouding, L., Andreassian, V., Rojas-Serna, C., Michel, C. and Mathevet, T.: Impact of limited streamflow data on the efficiency and the parameters of rainfall—runoff models, Hydrol. Sci. J., 52(1), 131–151, doi:10.1623/hysj.52.1.131, 2007.

Pool, S., Viviroli, D. and Seibert, J.: Prediction of hydrographs and flow-duration curves in almost ungauged catchments: Which runoff measurements are most informative for model calibration?, J. Hydrol., 554, 613–622, doi:10.1016/j.jhydrol.2017.09.037, 2017.

Ruhi, A., Messager, M. L. and Olden, J. D.: Tracking the pulse of the Earth's fresh waters, Nat. Sustain., 1(4), 198–203, doi:10.1038/s41893-018-0047-7, 2018.

Scherrer AG: Verzeichnis grosser Hochwasserabflüsse in schweizerischen Einzugsgebieten. Auftraggeber: Bundesamt für Umwelt (BAFU), Abteilung Hydrologie., Reinach., 2017.

Seibert, J.: Multi-criteria calibration of a conceptual runoff model using a genetic algorithm, Hydrol. Earth Syst. Sci., 4(2), 215–224, doi:10.5194/hess-4-215-2000, 2000.

Seibert, J. and Beven, K. J.: Gauging the ungauged basin: how many discharge measurements are needed?, Hydrol. Earth Syst. Sci., 13(6), 883–892, doi:10.5194/hess-13-883-2009, 2009.

Seibert, J. and McDonnell, J. J.: Gauging the Ungauged Basin : Relative Value of Soft and Hard Data, J. Hydrol. Eng., 20(1), A4014004-1–6, doi:10.1061/(ASCE)HE.1943-5584.000086, 2015.

Seibert, J. and Vis, M.: Teaching hydrological modeling with a user-friendly catchment-runoff-model software package, Hydrol. Earth Syst. Sci., 16(9), 3315–3325, doi:10.5194/hess-16-3315-2012, 2012.

Seibert, J. and Vis, M. J. P.: How informative are stream level observations in different geographic regions?, Hydrol. Process., 30(14), 2498–2508, doi:10.1002/hyp.10887, 2016.

Seibert, J., Vis, M. J. P., Lewis, E. and van Meerveld, H. J.: Upper and lower benchmarks in hydrological modelling, Hydrol. Process., 32(8), 1120–1125, doi:10.1002/hyp.11476, 2018.

Shiklomanov, A. I., Lammers, R. B. and Vörösmarty, C. J.: Widespread decline in hydrological monitoring threatens Pan-Arctic Research, Eos, Trans. Am. Geophys. Union, 83(2), 13, doi:10.1029/2002EO000007, 2002.

Sideris, I. V., Gabella, M., Erdin, R. and Germann, U.: Real-time radar-rain-gauge merging using spatio-temporal co-kriging with external drift in the alpine terrain of Switzerland, Q. J. R. Meteorol. Soc., 140(680), 1097–1111, doi:10.1002/qj.2188, 2014.

Strobl, B., Etter, S., van Meerveld, I. and Seibert, J.: Accuracy of Crowdsourced Streamflow and Stream Level Class Estimates, Hydrol. Sci. J., (special issue on hydrological data: opportunities and barriers), n.d.

Strobl, B., Etter, S., van Meerveld, I. and Seibert, J.: Accuracy of Crowdsourced Streamflow and Stream Level Class Estimates, Hydrol. Sci. J., (special issue on hydrological data), n.d.

Vrugt, J. A., Gupta, H. V., Dekker, S. C., Sorooshian, S., Wagener, T. and Bouten, W.: Application of stochastic parameter optimization to the Sacramento Soil Moisture Accounting model, J. Hydrol., 325(1–4), 288–307, doi:10.1016/j.jhydrol.2005.10.041, 2006a.

Vrugt, J. A., Gupta, H. V., Nualláin, B. and Bouten, W.: Real-Time Data Assimilation for Operational Ensemble Streamflow Forecasting, J. Hydrometeorol., 7(3), 548–565, doi:10.1175/JHM504.1, 2006b.

Weeser, B., Stenfert Kroese, J., Jacobs, S. R., Njue, N., Kemboi, Z., Ran, A., Rufino, M. C. and Breuer, L.: Citizen science pioneers in Kenya – A crowdsourced approach for hydrological monitoring, Sci. Total Environ., 631–632, 1590–1599, doi:10.1016/j.scitotenv.2018.03.130, 2018.

Yapo, P. O., Gupta, H. V. and Sorooshian, S.: Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data, J. Hydrol., 181(1–4), 23–48, doi:10.1016/0022-1694(95)02918-4, 1996.

Tables

Table 1 Characteristics of the six Swiss catchments used in this study. For the location of the study catchments see Figure 1. Long-term averages are for the period 1974-2014, except for Verzasca for which the long term average is for the 1990-2014 period. Regime types are classified according to (Aschwanden and Weingartner, 1985).

Catchment		**Murg**	**Guerbe**	**Allenbach**	**Riale di Calneggia**	**Mentue**	**Verzasca**
Gauging station (FOEN station number)		Waengi (2126)	Belp Mülimatt (2159)	Adelboden (2232)	Cavergno, Pontit (2356)	Yvonand La Mauguettaz (2369)	Lavertezzo, Campiòi (2605)
Area [km^2]		79	117	29	24	105	186
Elevation [m asl]	Min	465	522	1297	885	445	490
	Max	1035	2176	2762	2921	927	2864
Regime Type		Pluvial-inférieur	Pluvial-superieur	Nival-alpin	Nival-méridional	Pluvial-jurassien	Nivo-pluvial-méridional
Min - Max Pardé coefficients	Dry year	0.29 - 1.61	0.44 - 1.93	0.40 - 2.48	0.13 - 3.22	0.22 - 2.37	0.16 - 2.92
	Average year	0.58 - 2.16	0.61 - 1.65	0.39 - 2.44	0.09 - 2.84	0.23 - 2.66	0.23 - 3.17
	Wet year	0.34 - 1.69	0.42 - 2.14	0.32 - 2.12	0.10 - 3.48	0.35 - 2.39	0.26 -2.64
	Long-term	0.68 - 1.34	0.77 - 1.39	0.35 - 2.70	0.14 - 2.70	0.46 - 1.57	0.23 - 2.22
Annual runoff-rainfall ratio	Dry year	0.72	0.37	0.86	1.301	0.41	0.98
	Average year	0.55	0.48	1.731	1.381	0.52	0.66
	Wet year	0.56	0.54	0.78	0.98	0.50	1.321
	Long-term	0.56	0.57	0.94	1.061	0.38	0.9
Long-term mean annual streamflow [m^3/s]		1.84	2.75	1.23	1.43	1.64	10.76
Weather stations		Aadorf-Taenikon, Hörnli	Plaffeien, Bern-Zollikofen	Adelboden	Robiei	Mathod, Pully	Acquarossa, Cimetta, Magadino, Piotta

Table 2 Calibration years (2nd-most extreme and 2nd-closest to average years) and validation years (most extreme and closest to average years) for each catchment. Numbers in parenthesis are the ranks over the period 1974-2014 (or 1990-2014 for Verzasca).

[1] In Verzasca, Allenbach,and Riale die Calneggia there are some streamflow-rainfall ratios >1 because the weather stations are located outside the catchment and precipitation is highly variable in this alpine terrain.

Year character	Murg	Guerbe	Allenbach	Riale di Calneggia	Mentue	Verzasca
			Calibration			
Wet	2007 (3)	2007 (2)	2007 (4)	2009 (11)	2014 (7)	2011 (4)
Dry	2013 (8)	2011 (8)	2009 (11)	2012 (8)	2010 (4)	2013 (5)
Average	2008 (6)	2008 (17)	2013 (7)	2013 (2)	2006 (6)	2007 [7]
			Validation			
Wet	2014 [1]	2014 [1]	2014 [1]	2008 [9]	2007 [1]	2008 [1]
Dry	2009 (7)	2013 (5)	2012 (9)	2006 (5)	2009 (3)	2010 (4)
Average	2011 (4)	2006 (13)	2011 (6)	2011 (1)	2013 (2)	2006 (4)

Table 3 Median and the full range of P_{OA} scores for the upper benchmark (hourly-FOEN data). The upper benchmark values for the dry, average and wet calibration years were used as the upper benchmarks for the evaluation based on the year character (Figure 6 and Supplemental Material – Figure 2); the values in the "overall median"-column were used as the benchmarks in the overall median performance evaluation shown in Figure 4.

Calibration year	*Dry*	*Average*	*Wet*	*Overall median*
Validation wet year				
Upper benchmark	0.63 (0.19 - 0.79)	0.65 (0.36 - 0.8)	0.66 (0.45 - 0.8)	
Lower benchmark	0.34 (-0.02 - 0.47)			**Upper benchmark** 0.61 (0.19 - 0.83)
Validation average year				
Upper benchmark	0.59 (0.49 - 0.64)	0.61 (0.45 - 0.78)	0.53 (0.36 - 0.77)	
Lower benchmark	0.36 (0.03 - 0.59)			**Lower benchmark** 0.34 (-0.02 - 0.59)
Validation dry year				
Upper benchmark	0.51 (0.35 - 0.71)	0.59 (0.41 - 0.83)	0.53 (0.23 - 0.74)	
Lower benchmark	0.35 (0.09 - 0.52)			

Table 4 Weights assigned to specific seasons, days and times of the day for the random selection of data points for *Crowd52* and *Crowd12*. The weights for each hour were multiplied and normalized. We then used them as probabilities for the individual hours. For times without daylight the probability was set to zero.

Variable		Weight
Season		
December – February		2
March – May / September – November)		6
June – August)		10
Day		
Saturdays – Sundays		3
Monday – Friday		1
Time		
Times when people usually have breaks	6 am – 8:00 am, 12 am-1 pm, 5 pm-9 pm	3
Times with daylight in winter (Dec-Feb)	8 am – 4 pm	1
Times with daylight in spring/fall (Mar-May/Sept-Nov):	7 am – 7 pm	1
Times with daylight in summer (Jun-Aug)	6 am – 9 pm	1
Other times (depending on season)		0

5

Figures

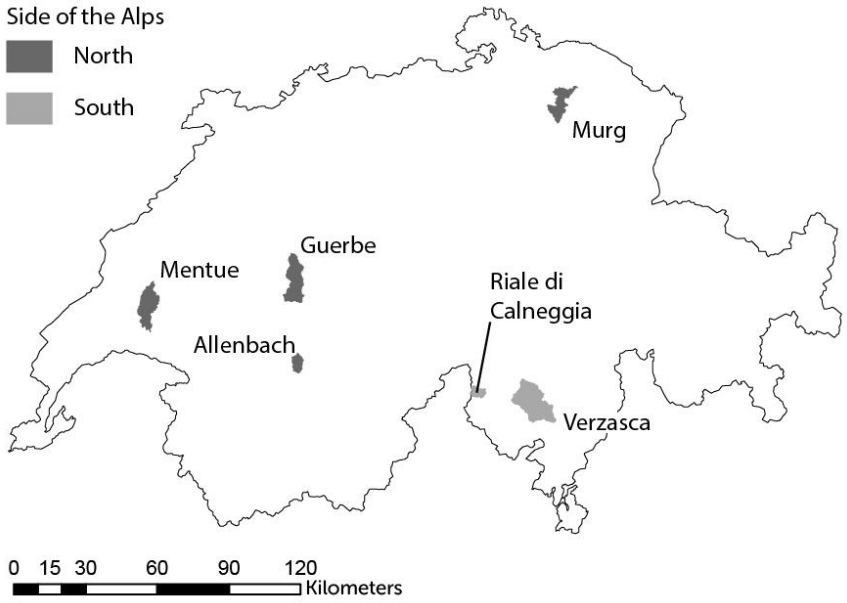


Figure 1 Location of the six study catchments in Switzerland. Shading indicates whether the catchment is located on the north or south side of the Alps. See Table 1 for the characteristics of the study catchments.

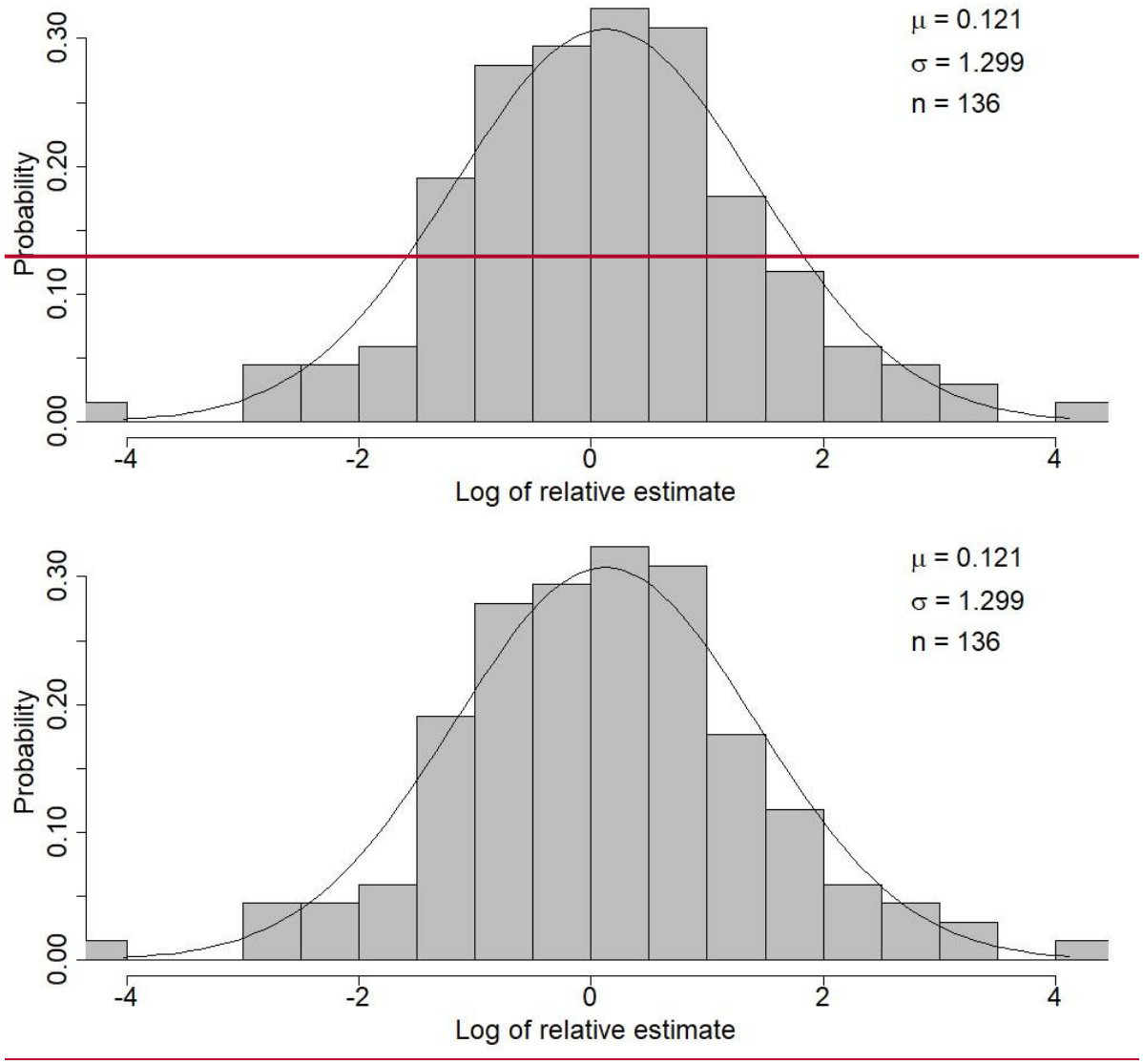
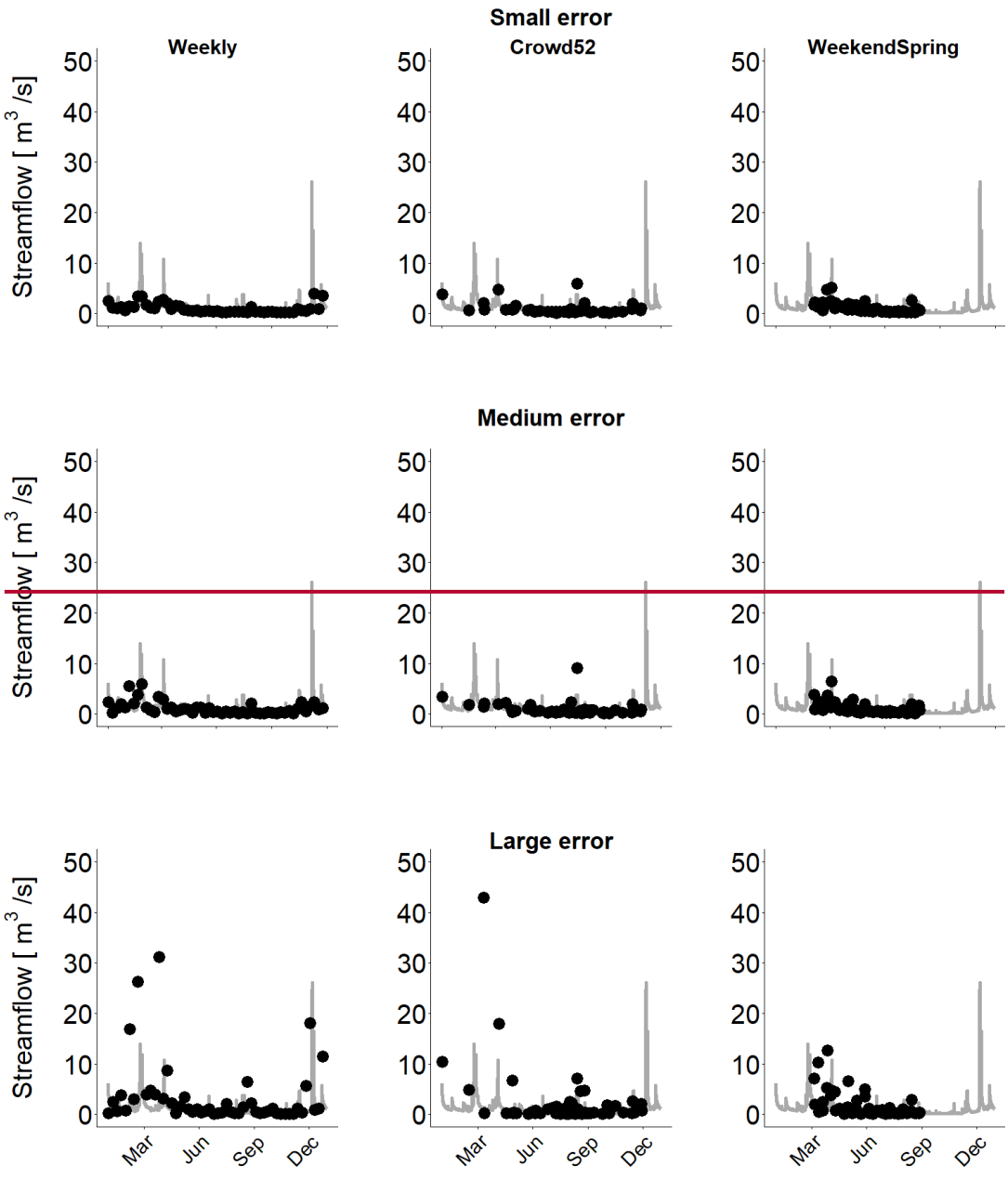


Figure 2 Fit of the normal distribution to the frequency distribution of the log transformed relative streamflow estimates (ratio of the estimated streamflow and the measured streamflow).


Small error
Weekly
Crowd52
WeekendSpring
Medium error
Large error
Streamflow [m3 /s]
0
10
20
30
40
50
Mar
Jun
Sep
Dec

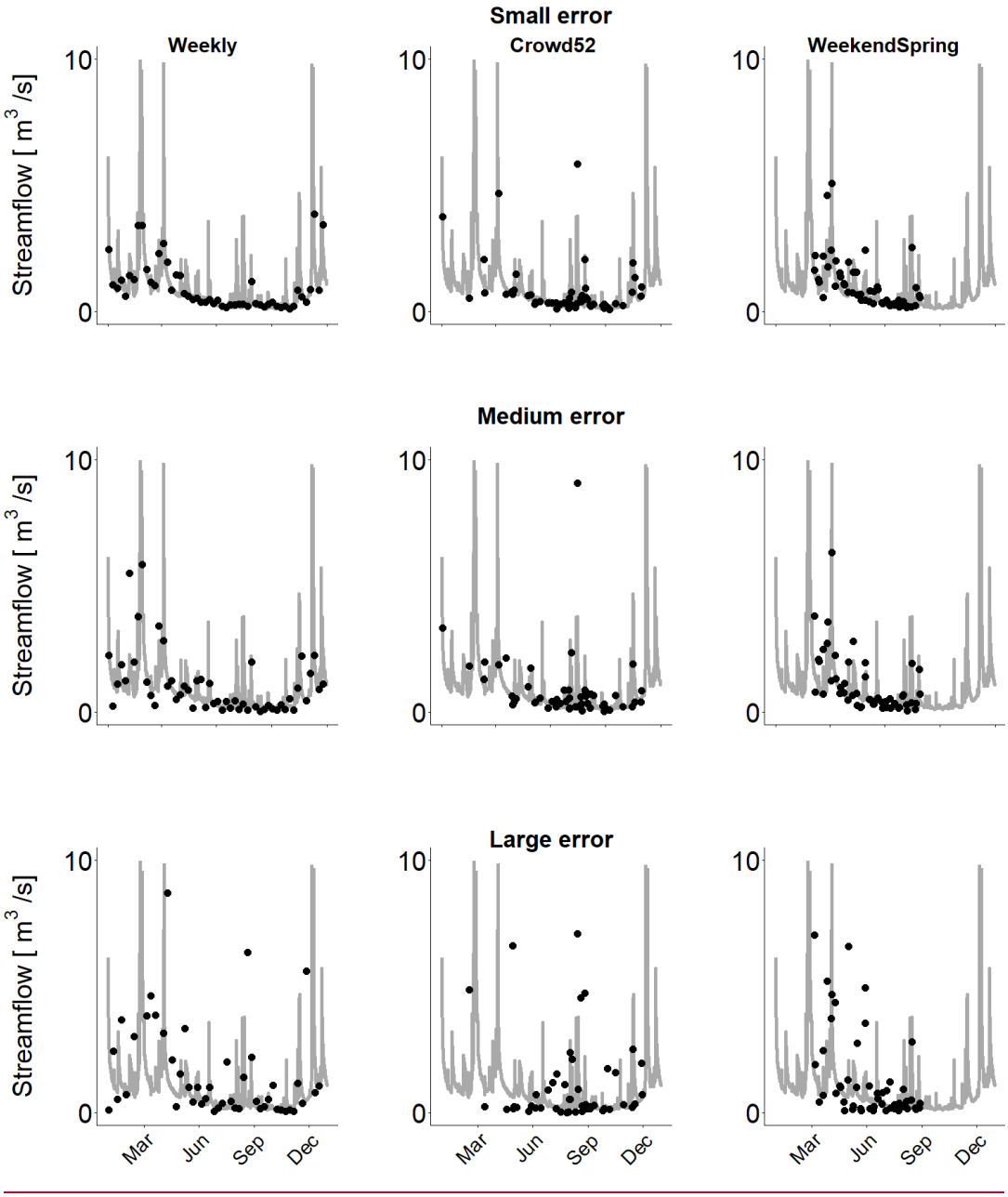
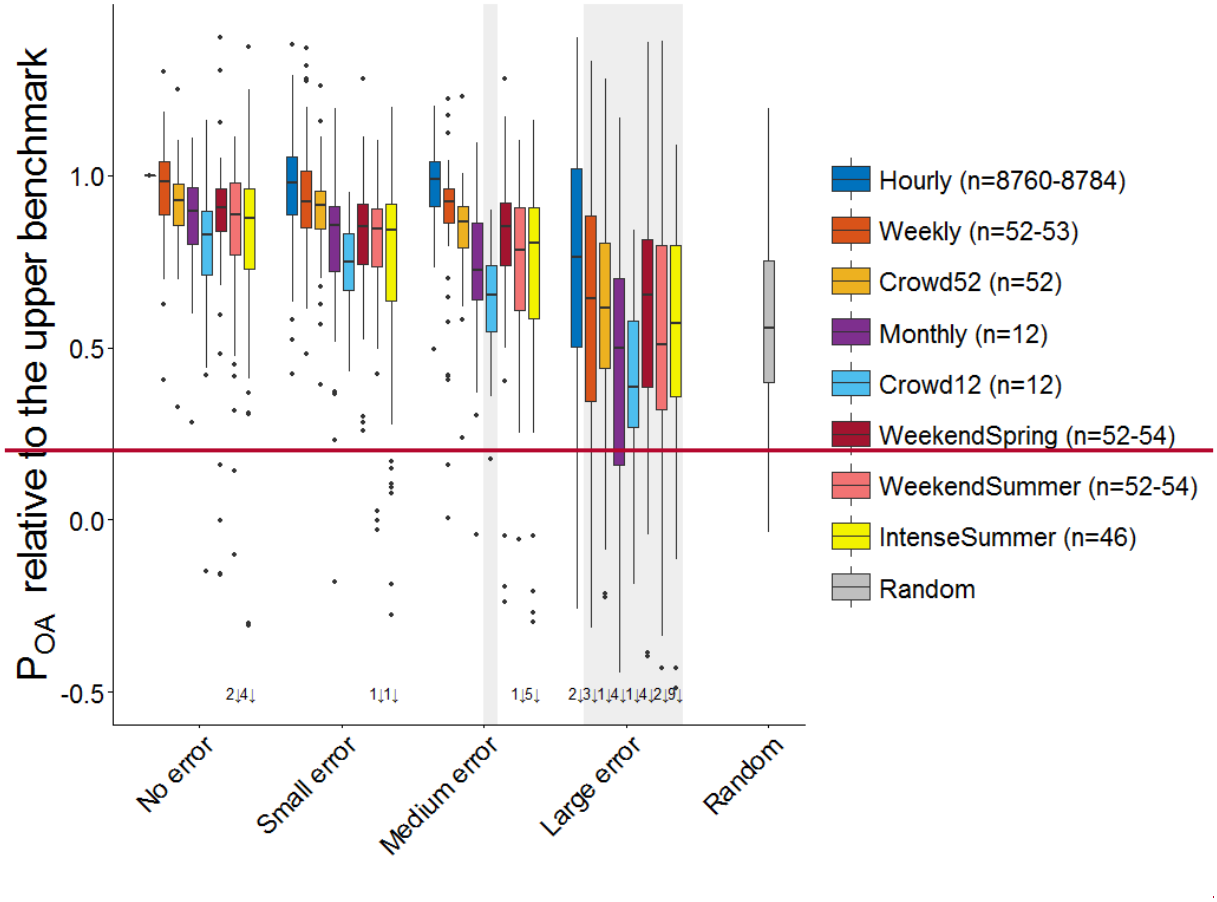


Figure 3 Example of different streamflow time series used for calibration with small, medium and large errors and the temporal resolutions (*Weekly*, *Crowd52* and *WeekendSpring*) for the Mentue in 2010. Large error: adjusted FOEN data with errors resulting from the log-normal distribution fitted to the streamflow estimates from citizen scientists (see Figure 2). Medium error: same as large error, but the standard deviation of the log normal distribution was divided by 2. Small error: same as the large error, but the standard deviation of the log normal distribution was divided by 4. The grey line represents the measured streamflow, the dots

the derived time series of streamflow observations. **Note that especially in the large error category some dots lie outside the figure margins.**


P_{OA} relative to the upper benchmark
1.0
0.5
0.0
-0.5
2↓4↓
1↓1↓
1↓5↓
2↓3↓1↓4↓1↓4↓2↓9↓
No error
Small error
Medium error
Large error
Random
Hourly (n=8760-8784)
Weekly (n=52-53)
Crowd52 (n=52)
Monthly (n=12)
Crowd12 (n=12)
WeekendSpring (n=52-54)
WeekendSummer (n=52-54)
IntenseSummer (n=46)
Random

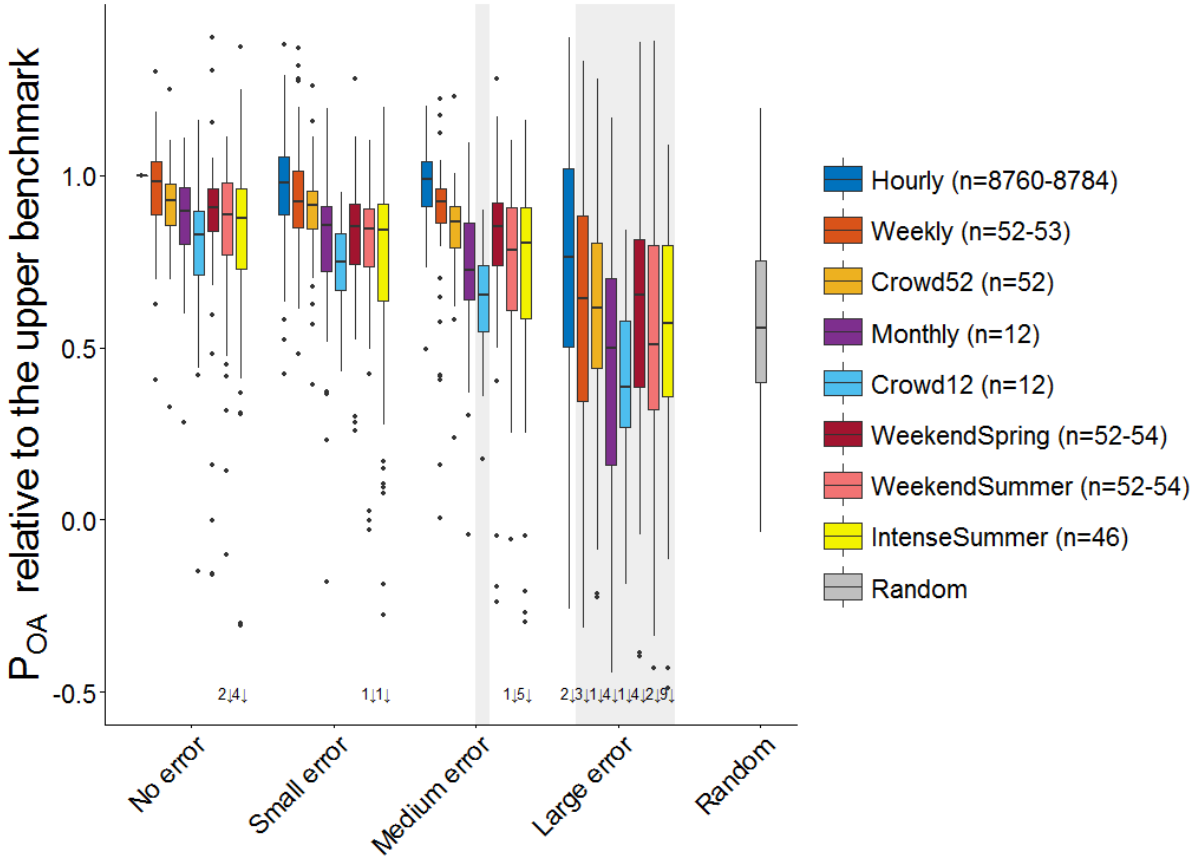


Figure 4 Boxplots of the median model performance relative to the upper benchmark for all datasets. The grey rectangles around the boxes indicate non-significant differences in median model performance compared to the lower benchmark with random parameter sets. The box represents the 25th and 75th percentile, the thick horizontal line the median, the whiskers extend to 1.5 times the interquartile range below the 25thpercentile and above the 75th percentile, and the dots represent the outliers. The numbers at the bottom indicates the number of outliers beyond the figure margins. *n* is the number of streamflow observations used for model calibration. The result of the hourly-benchmark FOEN dataset has some spread because the results of the 100 parameters sets were divided by their median performance. A relative P_{OA} of 1 indicates that the model performance is as good as the performance of the model calibrated with the hourly FOEN data (upper benchmark).

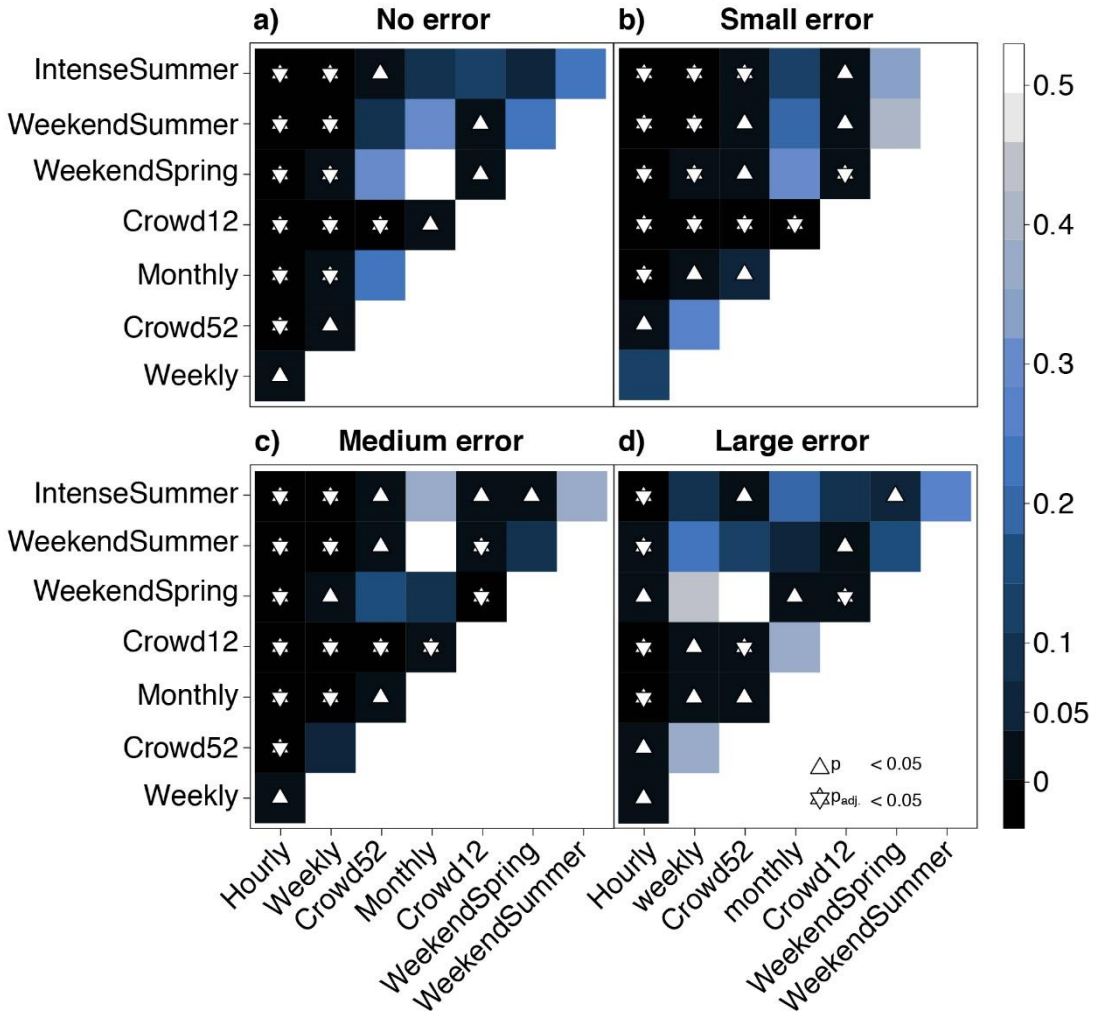
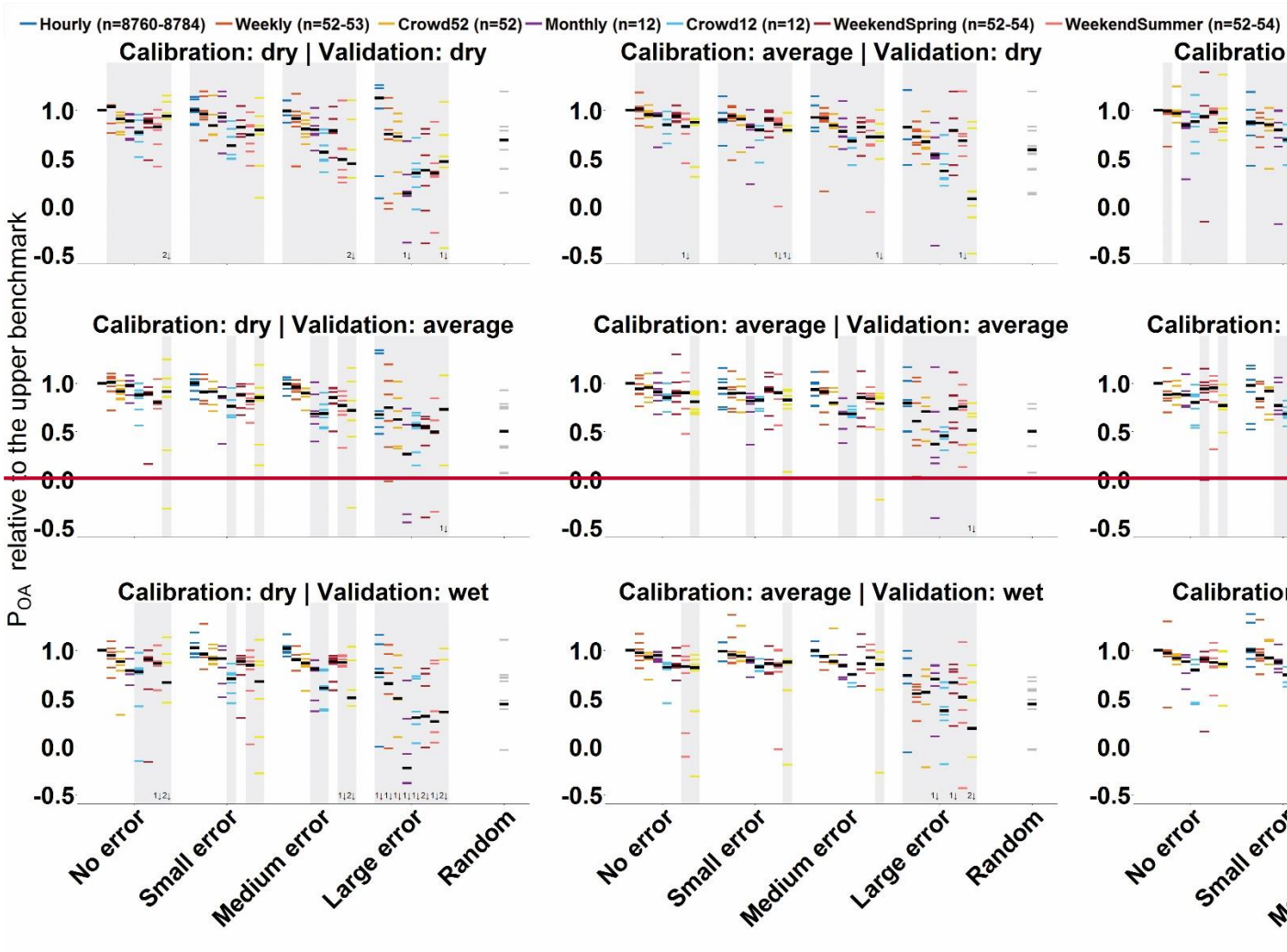


Figure 5 Results (p-values) of the Bonferroni Post-Hoc test to determine the significance of the difference in the median model performance for the data with different temporal resolutions within each data quality group (no error (a), small error (b), medium error (c), and large error (d)). Blue shades represent the p-values. White triangles indicate p-values < 0.05 and white stars indicate p-values that, when adjusted for multiple comparisons, are still < 0.05.

1


Hourly (n=8760-8784)
Weekly (n=52-53)
Crowd52 (n=52)
Monthly (n=12)
Crowd12 (n=12)
WeekendSpring (n=52-54)
WeekendSummer (n=52-54)
Calibration: dry | Validation: dry
Calibration: average | Validation: dry
Calibration: dry | Validation: average
Calibration: average | Validation: average
Calibration: dry | Validation: wet
Calibration: average | Validation: wet
P_OA relative to the upper benchmark
1.0
0.5
0.0
-0.5
No error
Small error
Medium error
Large error
Random

2

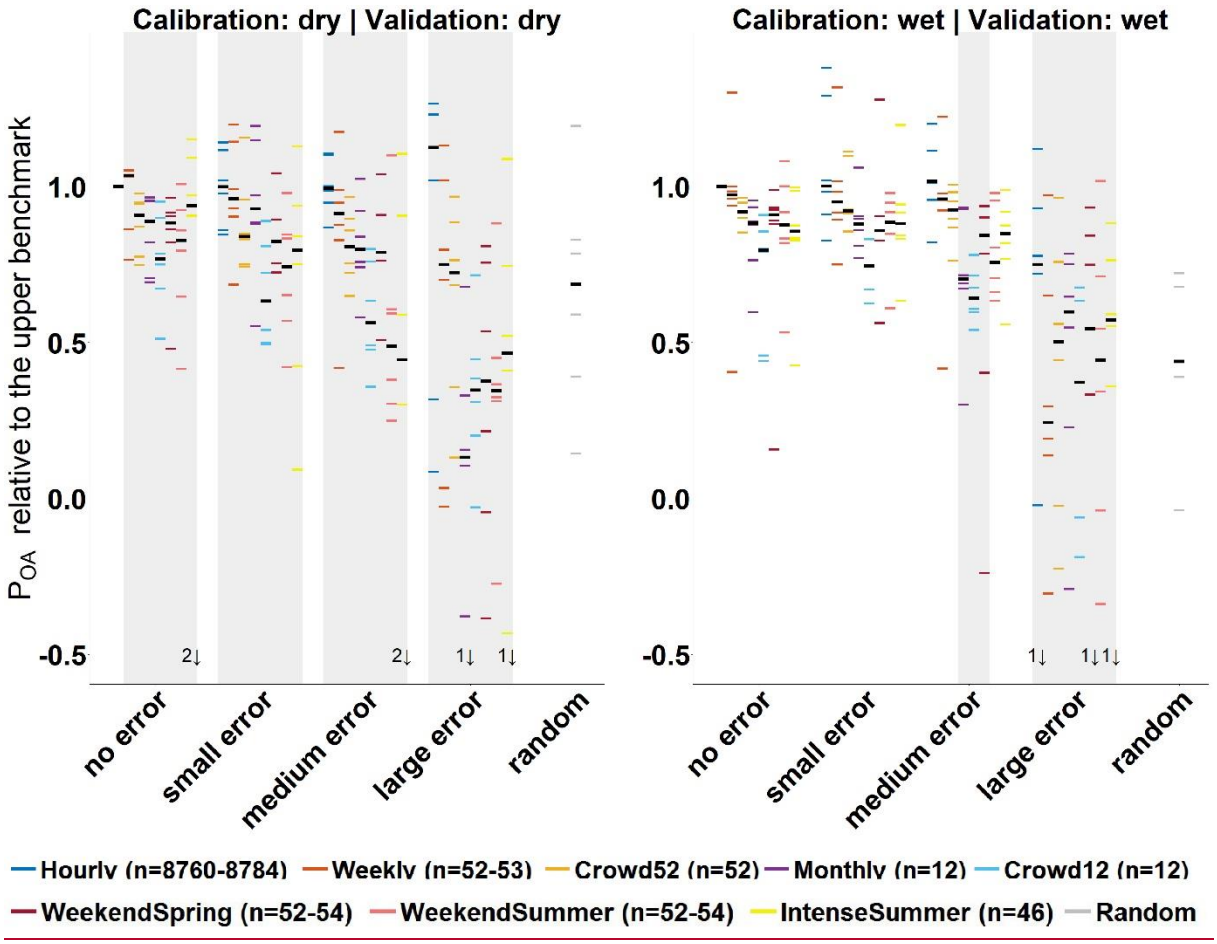


Figure 6 Median model validation performance for ~~all~~the datasets ~~used for calibration during the different validation periods~~calibrated and validated both in a dry year and both in a wet year. Each horizontal line represents the median model performance for one catchment. The black bold line represents the median for the six catchments. The grey rectangles around the boxes indicate non-significant differences in median model performance for the six catchments compared to the lower benchmark with random parameters. The numbers at the bottom indicate the number of outliers beyond the figure margins. For the individual P_{OA} values of the upper benchmark (no error – *Hourly* dataset) in the different calibration and validation years see Table 3~~Table 4.~~.

Supplemental Material

Model parameters

Supplemental Material - Table 1 Parameter ranges used for calibration of the HBV-model

Parameter	Description[a]	Unit	Min
Rescaling Parameters of Input Data			
PCALT	change in precipitation with elevation	% $(100m)^{-1}$	5
TCALT	change in temperature with elevation	°C $(10m)^{-1}$	0.5
Snow and ice melt parameters			
TT	threshold temperature for liquid and solid precipitation	°C	-3
CFMAX	degree-day factor	$mmd^{-1}°C^{-1}$	0.06
SFCF	snowfall correction factor	-	0.4
CFR	refreezing coefficient	-	0.001
CWH	water holding capacity of the snow storage	-	0.001
Soil Parameters			
PERC	maximum percolation from upper to lower groundwater storage	mm d^{-1}	0
UZL	threshold parameter	mm	0
K0	storage (or recession) coefficient 0	d^{-1}	0.001
K1	storage (or recession) coefficient 1	d^{-1}	0.0001
K2	storage (or recession) coefficient 2	d^{-1}	2E-06
MAXBAS	length of triangular weighting function	H	1
FC	maximum soil moisture storage	Mm	50
LP	soil moisture value above which actual evapotranspiration reaches potential evapotranspiration	-	0.3
Beta	shape factor for the function used to calculate the distribution of rain and snow melt going to runoff and soil box, respectively	-	1

[a]a detailed description of the model parameters is given in (Seibert and Vis, 2012).

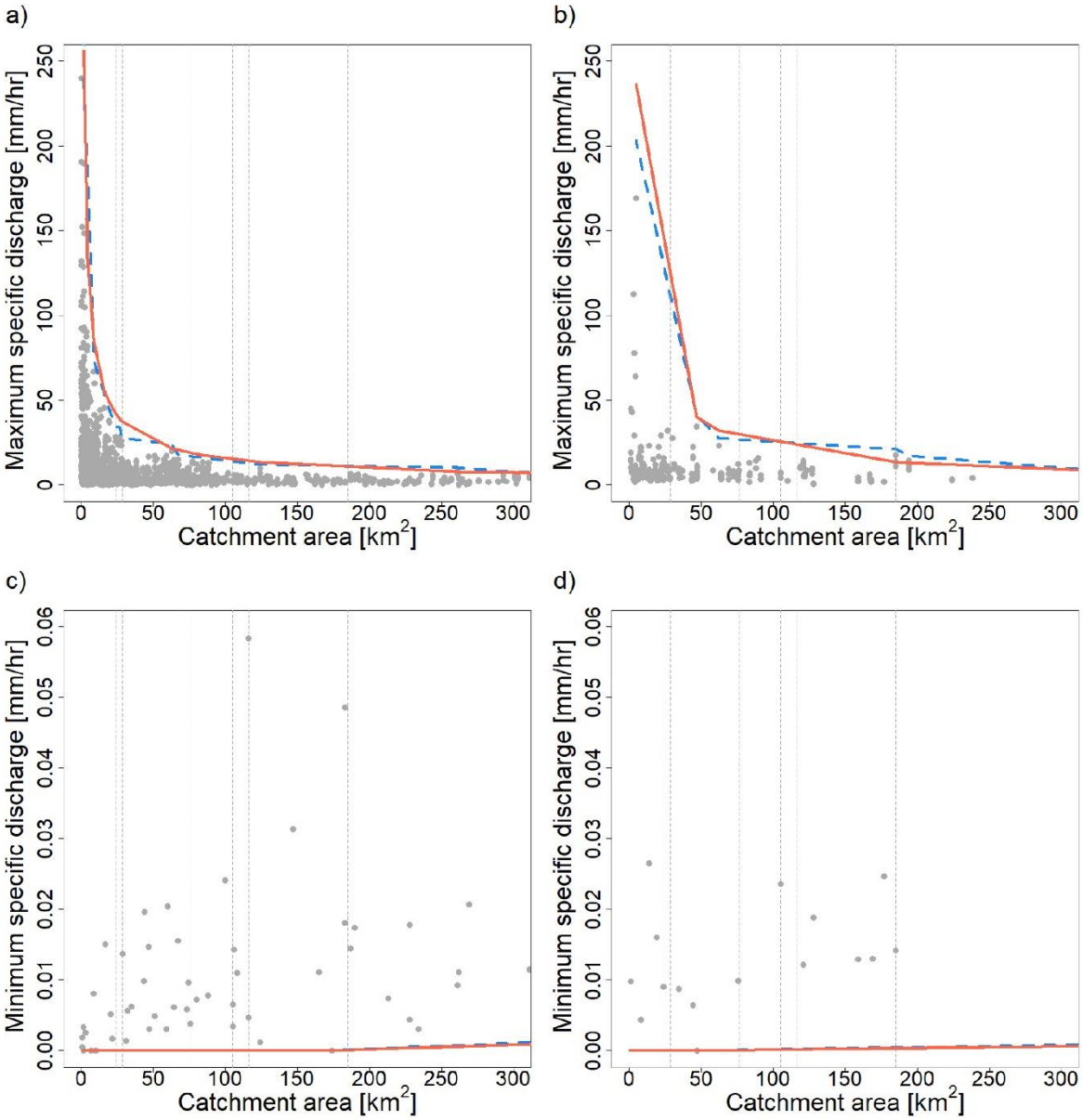
Significance of median model performance compared to the lower benchmark

Supplemental Material - Table 2 Significance of the differences in median model performance for each temporal resolution and an error group compared to the lower benchmark (Mann-Whitney U-test). The p-values of the Kruskal-Wallis test for the within group variability in the lowermost row shows that the median model performance of the different error groups was significantly different.

	No Error	Small Error	Medium Error	Large Error
Hourly	<0.01	<0.01	<0.01	<0.01
Weekly	<0.01	<0.01	<0.01	0.75
Crowd52	<0.01	<0.01	<0.01	0.40
Monthly	<0.01	<0.01	<0.01	0.03*
Crowd12	<0.01	<0.01	0.11	<0.01*
WeekendSpring	<0.01	<0.01	<0.01	0.40
WeekendSummer	<0.01	<0.01	<0.01	0.46
IntenseSummer	<0.01	0.01	0.04	0.21
Within error group	<0.01	<0.01	<0.01	<0.01

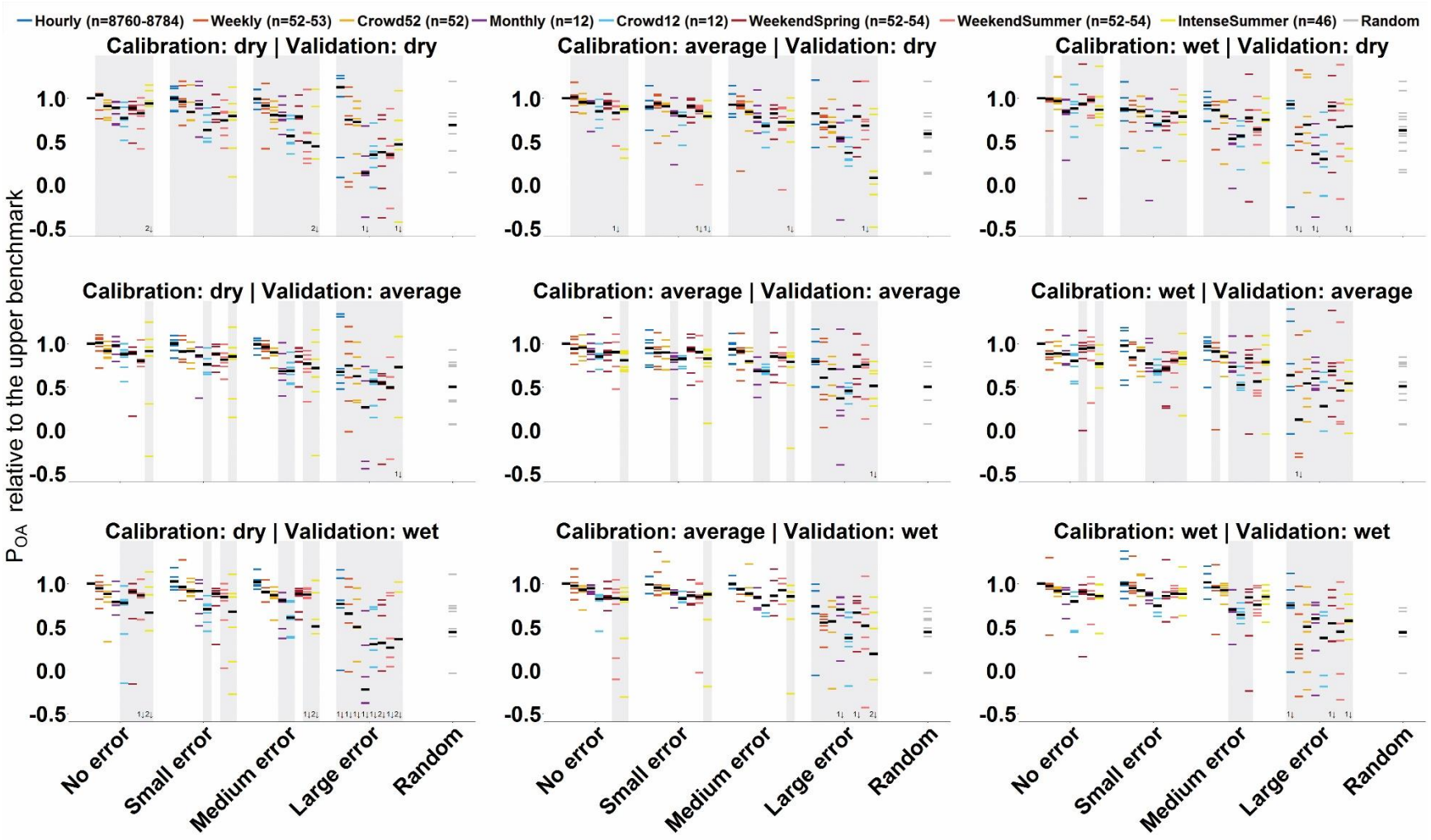
* These datasets result in significantly worse results than random parameters.

Extreme outlier removal for the northern and southern side of the Alps



Supplemental Material – Figure 1 Relation between catchment area and maximum (a, b) and minimum (c, d) specific streamflow for catchments on the north (a, c) and south (b, d) of the Alps. The dashed light blue line is the Pareto front including the 20 % buffer. The red lines are the fitted logarithmic models used to find the maximum and minimum possible flow for each catchment.

~~**Supplemental Material – Figure 1 Relation between catchment area and maximum (a, b) and minimum (c, d) specific streamflow for catchments on the north (a, c) and south (b, d) of the Alps. The dashed light blue line is the Pareto front including the 20 % buffer. The red lines are the fitted logarithmic models used to find the maximum and minimum possible flow for each catchment.**~~



Supplemental Material – Figure 2 Median model validation performance for all datasets used for calibration during the different validation periods. Each horizontal line represents the median model performance for one catchment. The black bold line represents the median for the six catchments. The grey rectangles around the boxes indicate non-significant differences in median model performance for the six catchments compared to the lower benchmark with random parameters. The numbers at the bottom indicate the number of outliers beyond the figure margins. For the individual P_{OA} values of the upper benchmark (no error – *Hourly* dataset) in the different calibration and validation years see Table 3.