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 will include a discussion about the effects of errors in the data and the effect of 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) which 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 for some catchments, 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 parameter values. However, these change 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).

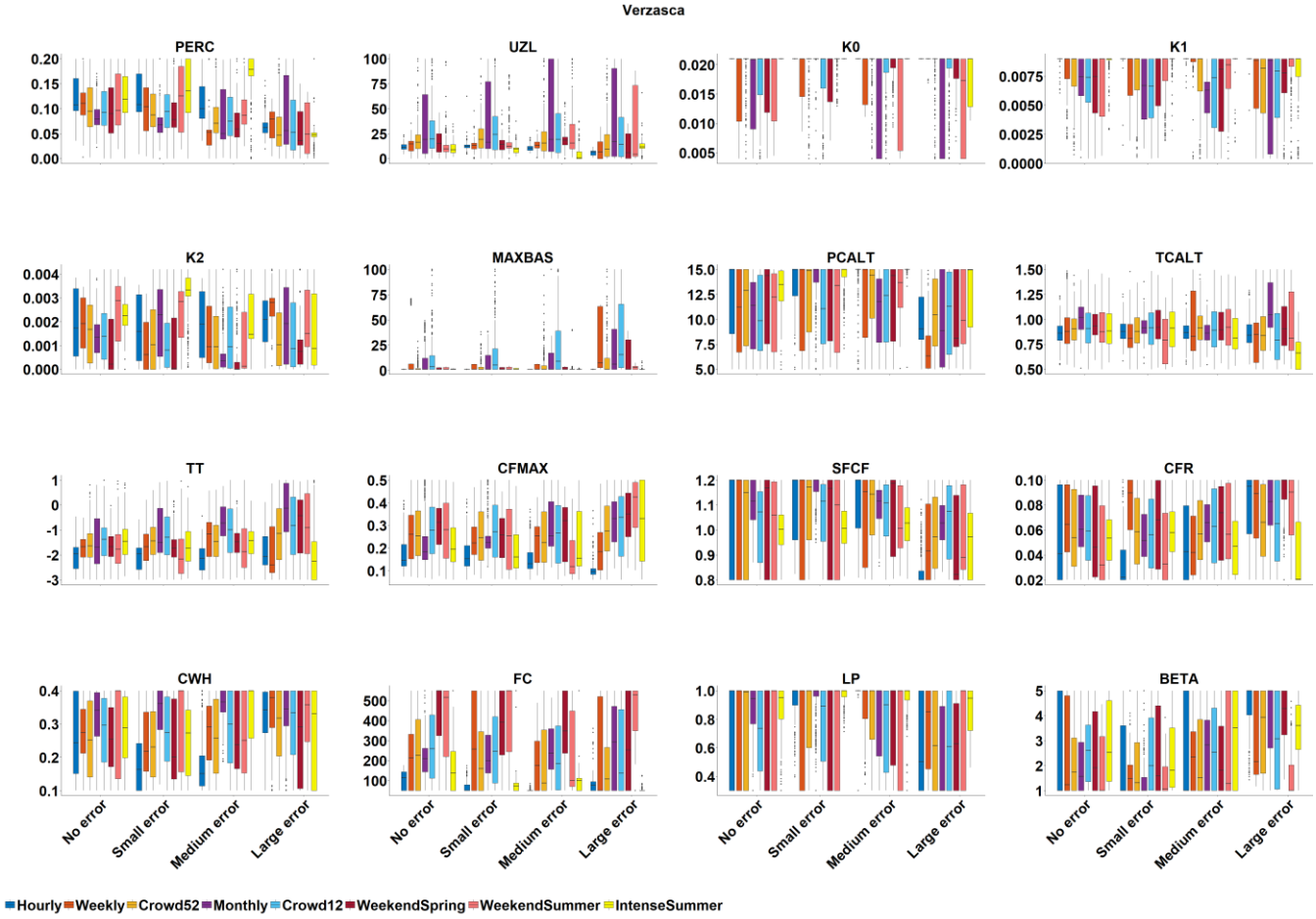

Figure R1 Boxplots of the model parameters for different combinations of errors and temporal resolutions of the data used for model calibration for the Verzasca catchment. Each subplot shows the range for one model parameter and consists of 300 values (3 year characters x 100 calibrations). The box represents 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 see Table R1 Error! Reference source not found.**.**

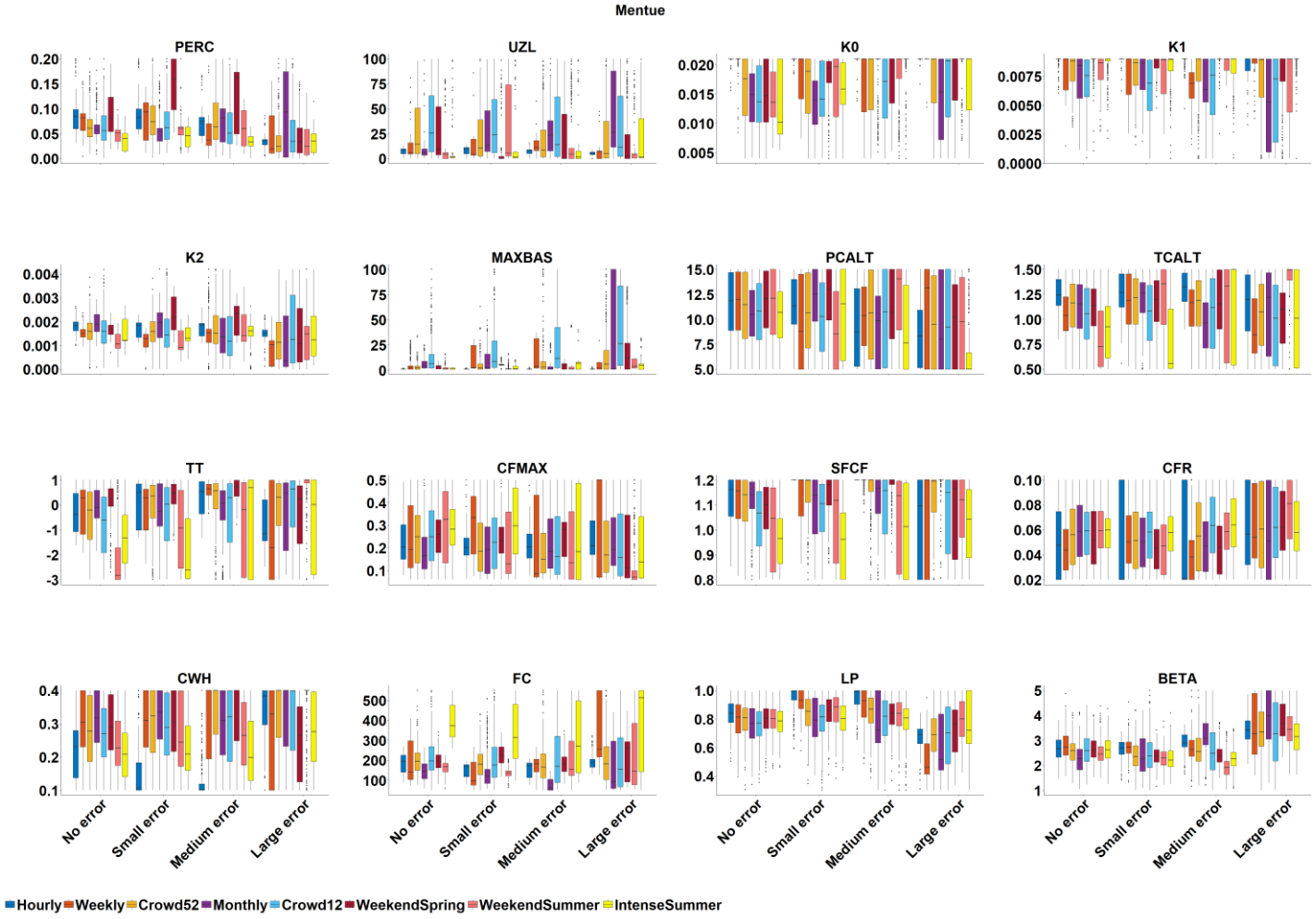

Figure R2 Boxplots of the model parameters for different combinations of errors and temporal resolutions of the data used for model calibration for the Mentue catchment. Each subplot shows the range for one model parameter and consists of 300 values (3 year characters x 100 calibrations). For a description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

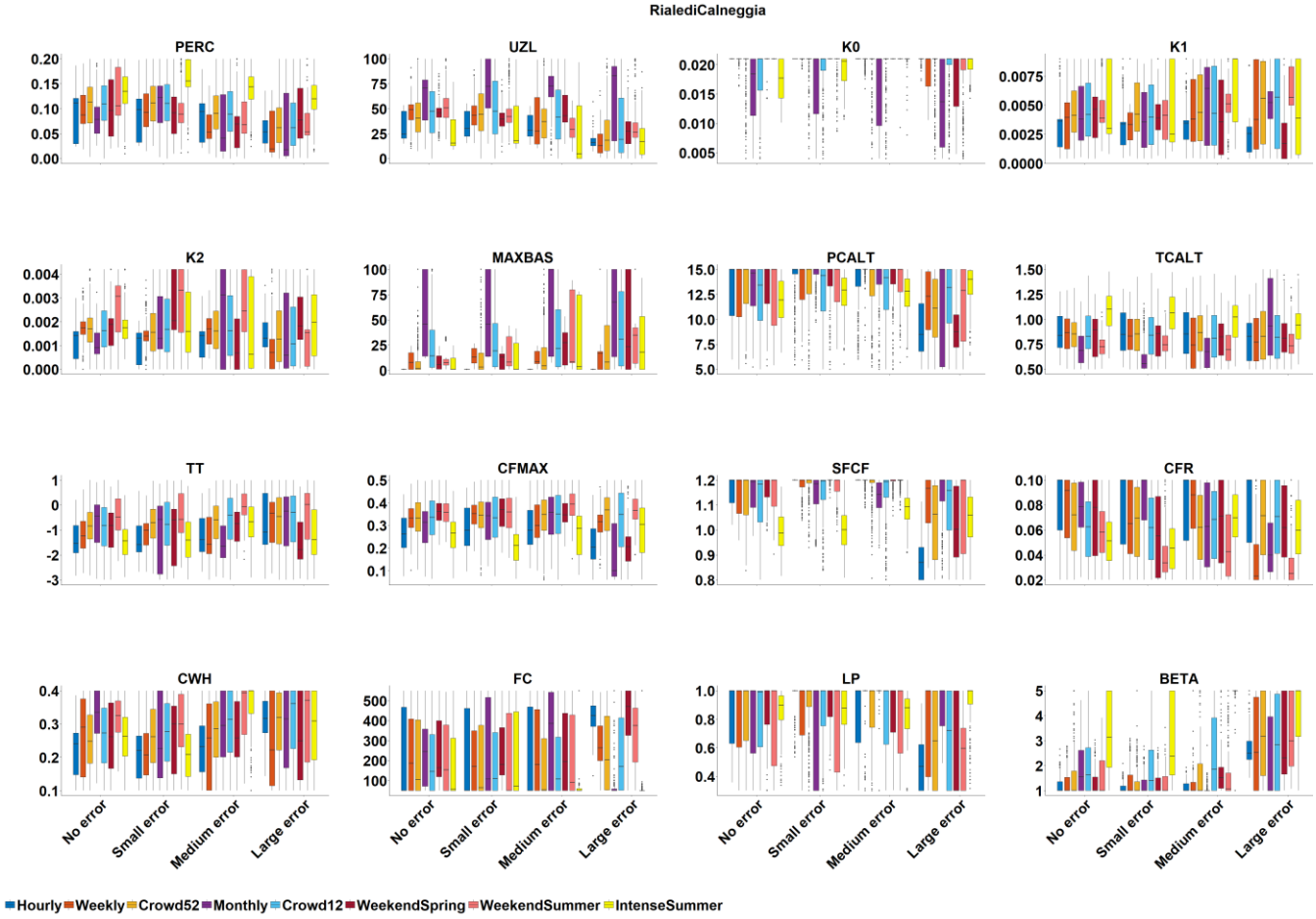

Figure R3 Boxplots of the model parameters for different combinations of errors and temporal resolutions of the data used for model calibration for the Riale di Calneggia catchment catchment. Each subplot shows the range for one model parameter and consists of 300 values (3 year characters x 100 calibrations).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

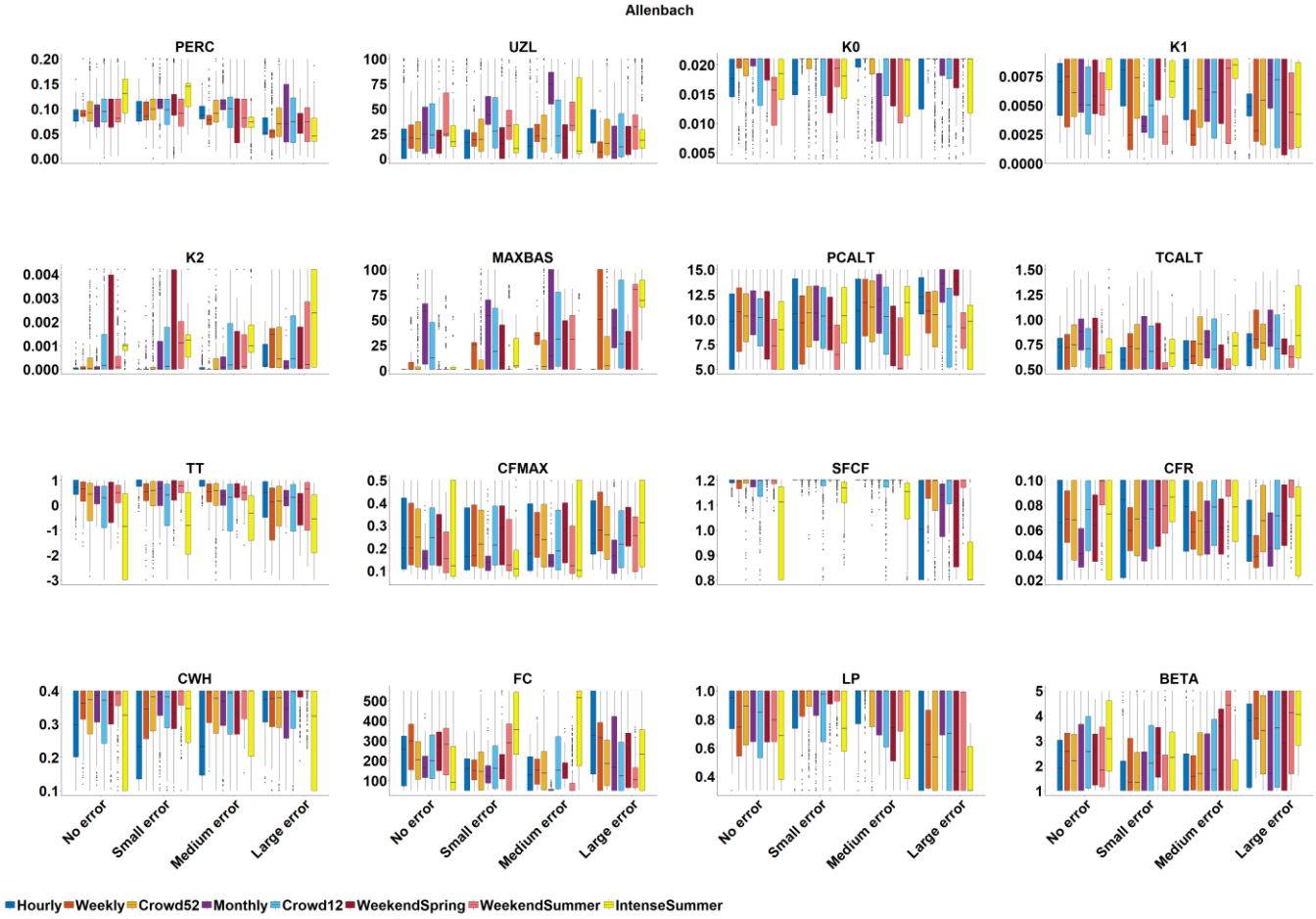

Figure R4 Boxplots of the model parameters for different combinations of errors and temporal resolutions of the data used for model calibration for the Allenbach catchment. Each subplot shows the range for one model parameter and consists of 300 values (3 year characters x 100 calibrations). For a description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

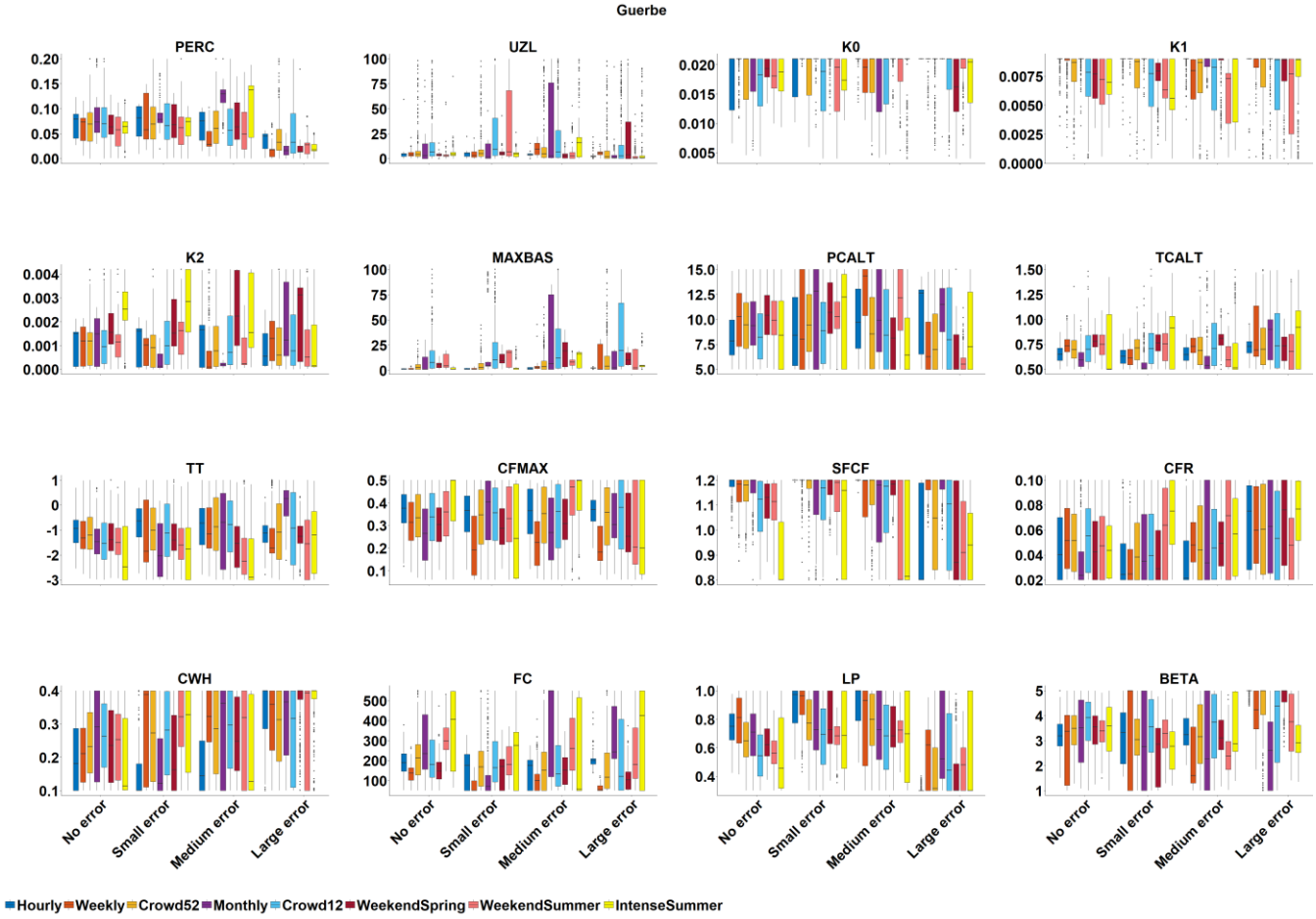

Figure R5 Boxplots of the model parameters for different combinations of errors and temporal resolutions of the data used for model calibration for the Guerbe catchment. Each subplot shows the range for one model parameter and consists of 300 values (3 year characters x 100 calibrations). For a description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

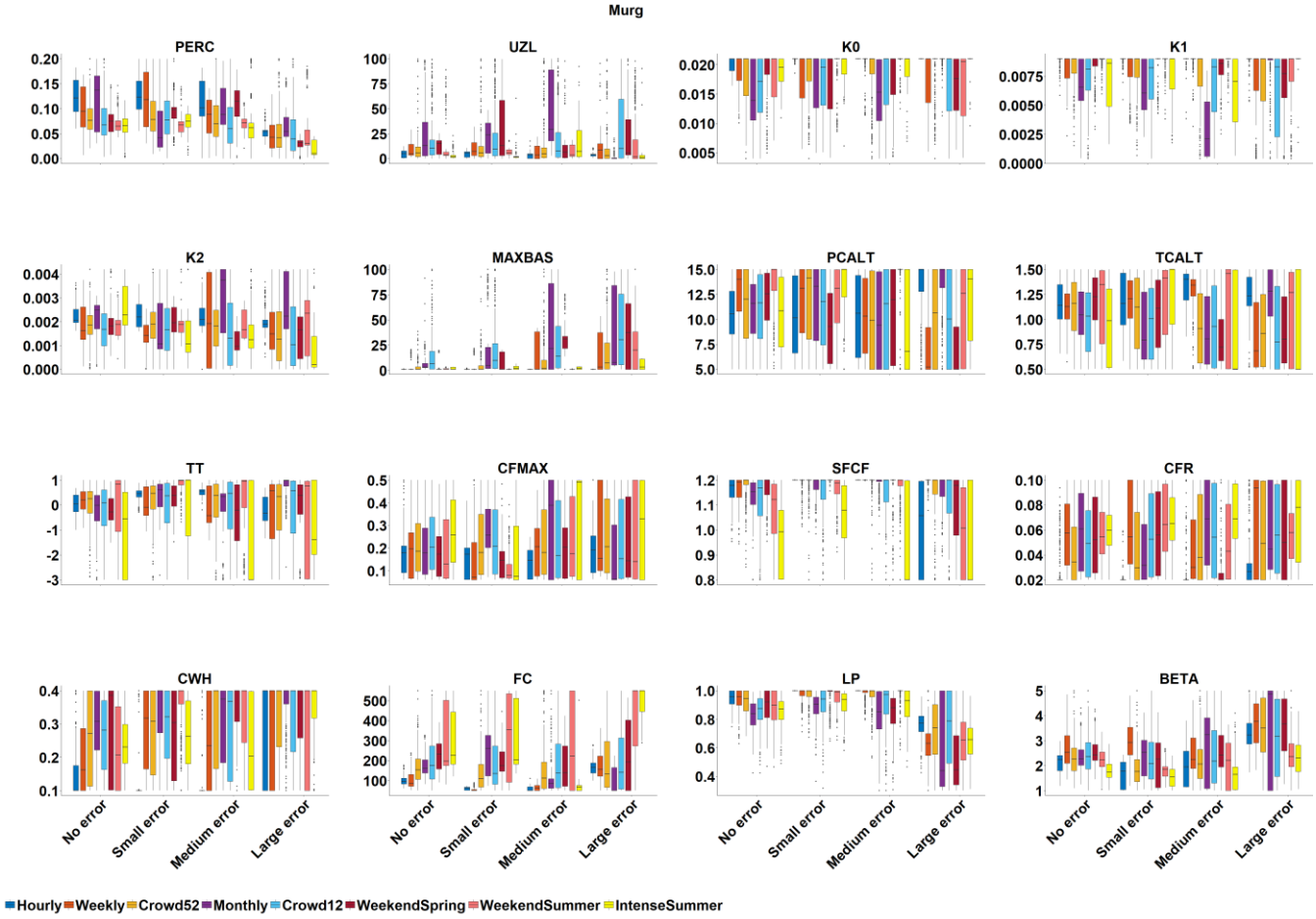

Figure R6 Boxplots of the model parameters for different combinations of errors and temporal resolutions of the data used for model calibration for the Murg catchment. Each subplot shows the range for one model parameter and consists of 300 values (3 year characters x 100 calibrations). For a description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

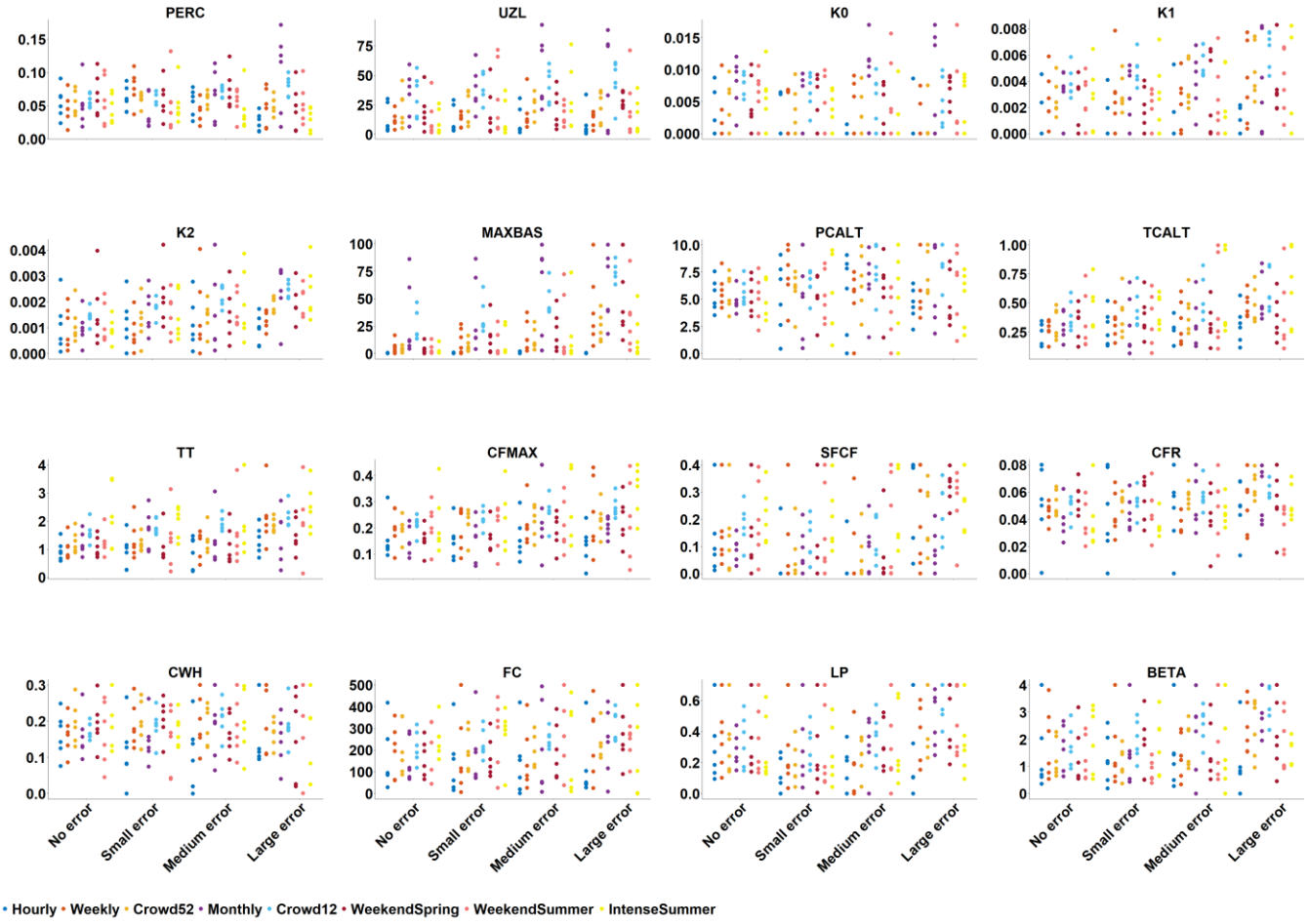

Figure R7 The interquartile range of the model parameters for the six catchments for the different combinations of errors and temporal resolutions of the data used for model calibration. Each dot represents the interquartile range for one catchment (i.e. is the size of the box in Figures R1-R6). For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

5

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

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.

Thank you very much for pointing this out. We agree and will adjust the text to more explicitly state that these results are for lumped 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 values 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 3 months were the absolute minimum.

We will add more information to the introduction to describe why the different studies resulted in different minimum data sets and highlight even better that despite their differences they all find that limited datasets are useful.

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 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.	10 min streamflow data	Variant of HBV	One event or 10 high flow measurements provided almost as much information as a 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 will include a sentence that states that citizen science includes more than data collection in the introduction and reference this paper there.

-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), since 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 designing the graph as suggested 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 therefore prefer to use the nine subplots that we used in the submitted version of the manuscript. We are, however, open for other suggestions on how to improve the figure.

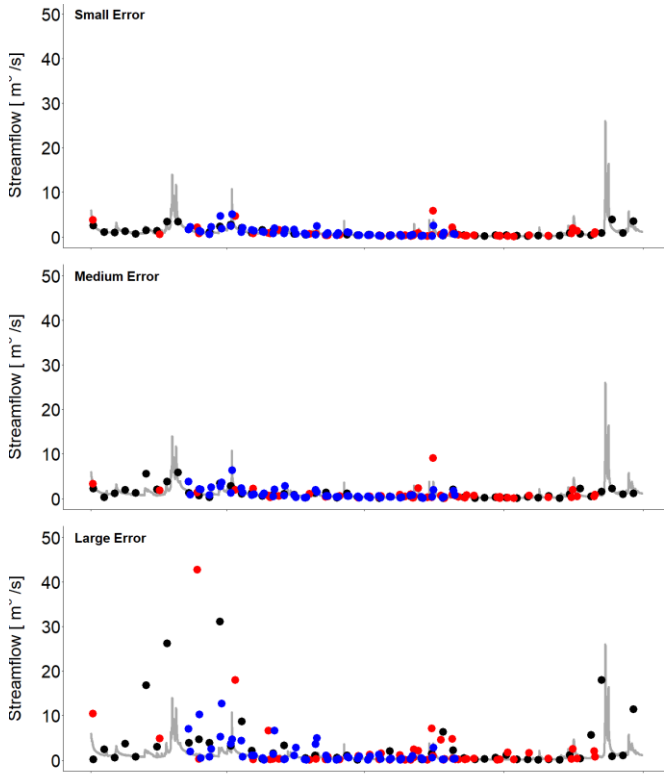

Figure R8 Alternative design of Figure 3 with filled symbols to represent the data used for model calibration for the different scenarios.

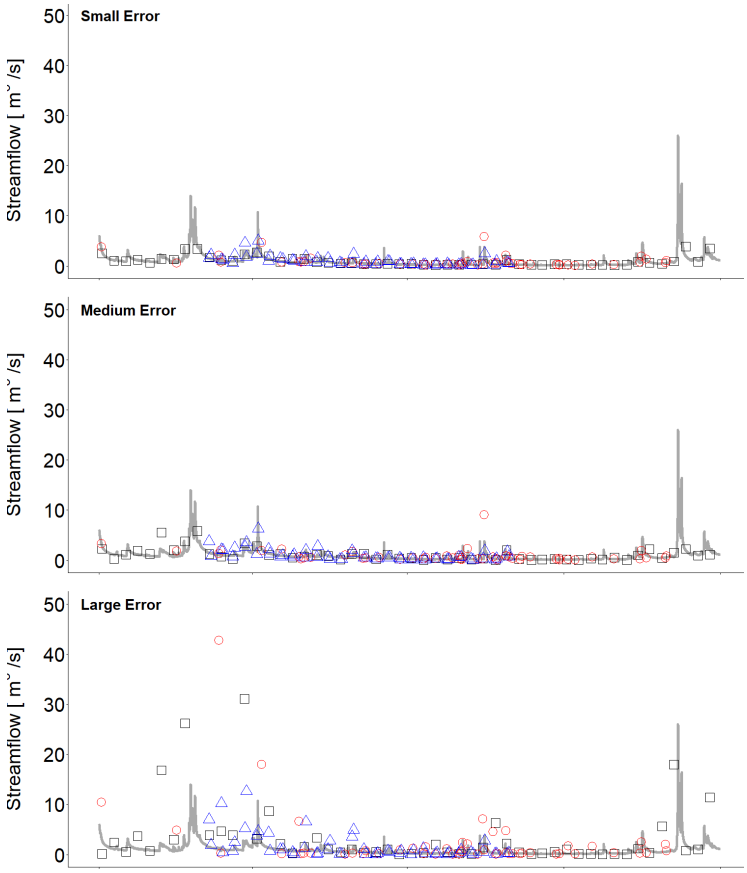

Figure R9 Alternative design of Figure 3 with different open symbols to represent the data used for model calibration for the different scenarios.