Response to interactive comment on "Information content of stream level class data for hydrological model calibration" by W. Buytaert (Referee)

This is a nice paper - very clearly written and overall well presented. The topic is novel and relevant - indeed I think that the insights are useful beyond citizen science, and help understanding the usefulness of other unconventional data sources such as cameras, and low-resolution sensors.

We thank the reviewer for the positive comments.

I have only a couple of concerns/queries:

1. The impact of measurement frequency on the performance of the models The simulation of "citizen science" data pretends that stream level data are available at a daily level (p.3/16). This is a lot, and probably unrealistic for real citizen science applications. This matters, because the Nash Sutcliffe efficiency and many other performance measures are quite sensitive to timing errors, and daily measurements, even only of water level, will make it possible to calibrate the timing related parameters of a hydrological model (e.g. overland and channel flow velocities) pretty well for all but the smallest catchments. I expect that the constraining power of the data will decrease strongly if the frequency of measurement reduces. So it is a pity that this was not studied. Alternatively, it may be useful to evaluate the model performance using a measure that puts more weight on the water balance (e.g., bias), because this is of course the specific weakness of using water level data for calibration instead of streamflow data.

We agree that daily data is unlikely for citizen science projects (except perhaps in rare cases where there is a dedicated volunteer who takes daily measurements near his/her house). However, daily data is certainly likely for webcam or time-lapse camera images, which are usually renewed multiple times per hour.

We already discussed these limitations in section 4.3, where we also mentioned that daily data contains a lot of redundant information and that previous studies have shown that a handful of measurements can be sufficient for model calibration (Rojas-Serna et al., 2016; Seibert and Beven, 2009). In response to this reviewer comment, we have now calibrated the model also with weekly (instead of daily) data. We then validated these parameterisations with the daily streamflow data. We did this for the case that weekly data are available for two, three and five stream level classes, stream levels and streamflow (Figure 1). The results show that the deterioration in model performance when weekly data are used instead of daily data is small, particularly for the stream level class data. We will include a description of these results in the revised version of the manuscript and if the editor thinks it is useful, can include the figure as well.

To more realistically represent citizen science data we are working on a follow up study for Swiss catchments, where we will test the effects of different measurement intervals and the effect of the temporal distribution of the citizen science data on model calibration. Here, we will also include the effects of data errors. Because there are many possible scenarios to represent citizen science data, this leads to a very large number of simulations. We feel that it is too much to include all this information in this manuscript and that it would take the focus away from the central message that stream level class data are useful.

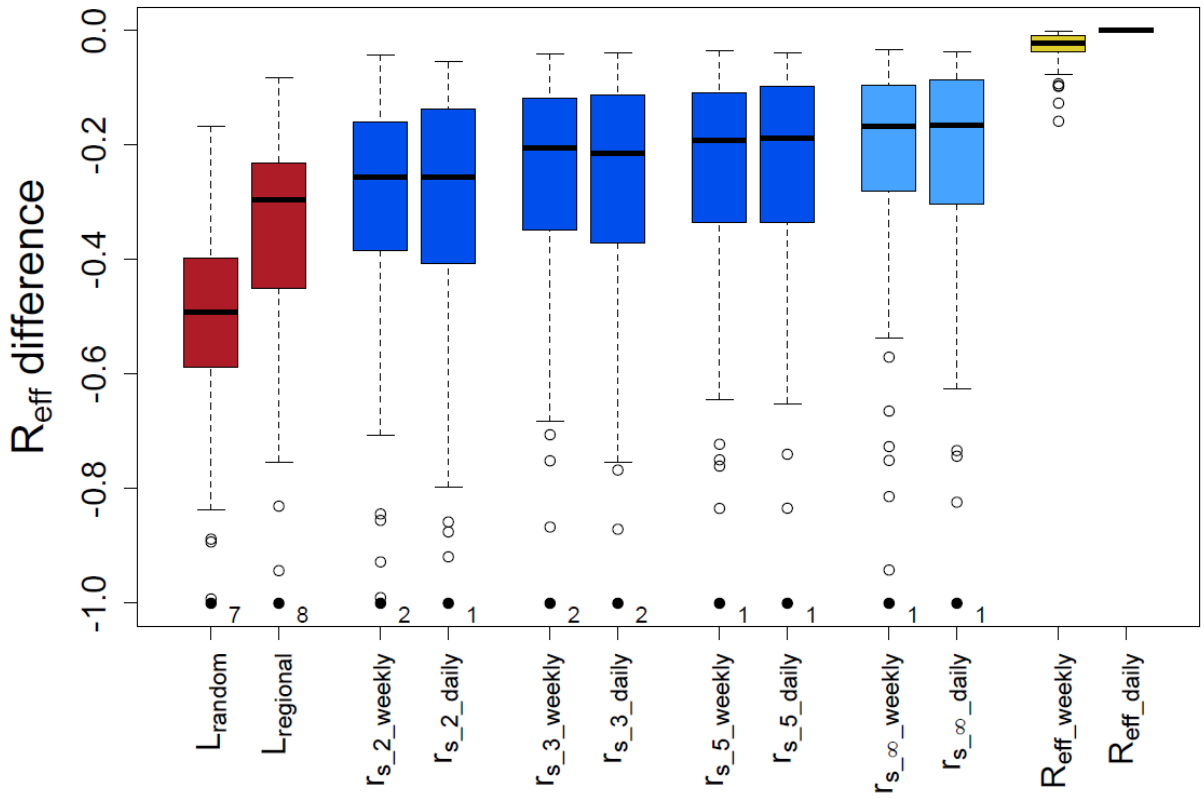

Figure 1. Box plots of model performance relative to the upper benchmark (R_{eff_daily}) for the the 100 catchments for model calibration with daily and weekly data for 2, 3, and 5 water level classes (r_{s_n}), stream level data (r_{s_∞}) and weekly streamflow data (R_{eff_weekly}). The results for the lower benchmarks are shown for comparison as well (note that for the lower benchmarks the model is not calibrated and there is thus no difference in model performance for daily and weekly data). The number of catchments for which the difference in model efficiency with the upper benchmark was >1 is given above the x-axis (indicated with the solid circles). The results for the lower benchmarks and the daily data are the same as those shown in Figure 1 in the manuscript.

2. The reporting of the model efficiency. The model efficiency measure R_eff is not defined (p.5/2). Only much further in the text, it is suggested that the Nash Sutcliffe efficiency is used (p.8/22-23). Is that correct? Irrespective of the definition of R_eff, I think that it would be useful to report the actual performance of the "upper benchmark", i.e. the models calibrated with stream- flow data. This is useful to get an idea of the order of magnitude of model performance that can be obtained with the citizen science data (irrespective of the difference with a fully calibrated model).

We thank the reviewer for pointing out that we did not define model efficiency at the start of the manuscript. This is indeed the Nash Sutcliffe efficiency and we will make this clearer in the revised version of the manuscript.

We will also include a table with the minimum, maximum and median Nash Sutcliffe efficiencies for the upper benchmark (R_{eff}), the models calibration with high resolution water level data (r_{s_∞}), the models calibrated with two, three and five water level classes (r_{s_n}) and the two lower benchmarks.

Data used for model calibration		All catchments (n=100)	Dry catchments (n=22)	Humid catchments (n=62)	Wet catchments (n=16)
Streamflow data (upper benchmark, R_{eff})	Median	0.77*	0.77	0.75	0.86
	Max	0.92	0.92	0.90	0.92
	Min	0.53	0.56	0.53	0.64
Water level data (r_{s_∞})	Median	0.58	0.32	0.58	0.80
	Max	0.89	0.61	0.79	0.89
	Min	-1.48	-1.48	0.13	0.53
5 stream level classes (r_{s_5})	Median	0.56	0.29	0.57	0.79
	Max	0.88	0.62	0.79	0.88
	Min	-1.68	-1.68	0.10	0.53
3 stream level classes (r_{s_3})	Median	0.54	0.27	0.55	0.76
	Max	0.88	0.57	0.79	0.88
	Min	-1.71	-1.71	-0.14	0.52
2 stream level classes (r_{s_2})	Median	0.49	0.28	0.49	0.72
	Max	0.87	0.65	0.77	0.87
	Min	-0.57	-0.57	-0.12	0.47
Parameters from other catchments ($L_{regional}$)	Median	0.43	0.21	0.43	0.70
	Max	0.79	0.50	0.65	0.79
	Min	-5.56	-5.56	-2.54	0.43
Random parameters (L_{random})	Median	0.25	0.11	0.26	0.56
	Max	0.76	0.38	0.66	0.76
	Min	-6.04	-6.04	-1.60	0.13

** For the 600+ catchments studied by Seibert and Vis (2016) the median efficiency was 0.74*

Table 1. Median, maximum and minimum Nash Sutcliff efficiency for the 100 catchments for model calibrations using different types of data and the two lower benchmarks. Note that the difference in the median Nash Sutcliff efficiency for the model calibrations with all streamflow data (R_{eff}) and the median Nash Sutcliff efficiency for the model calibrations with data for n water level classes (r_{s_n}) is not the same as the median of the differences in efficiency between the model calibrated with all streamflow data and the model calibrated with the stream level class data that is reported in the text and shown in the figures of the manuscript.

3. Model calibration The procedure used to calibrate the models is not clear to me. The manuscript states that "the model was calibrated 100 times, with each calibration trial consisting of 3500 model runs.", but I do not understand how exactly this is done. I suppose that the 3500 runs refer to different (sampled?) parameter sets, but what do the 100 times refer to? It suggests a kind of equifinality approach, but then I don't understand how this results in a single performance measures. Similarly, I don't understand how the 1000 randomly chosen parameters of the first lower benchmark (L_random), result in a single performance measure. I think that this needs to be clarified to make sure that it is reproducible, if only for confused minds like mine.

We thank the reviewer for pointing out this unclarity and will improve the description in the next version of the manuscript. In short, we used 100 independent model calibration trials resulting in 100 parameter sets for each catchment (one for each model calibration). For each of these (100) calibration trials, a total of 3500 model runs were done to find the

optimum parameter set with the genetic algorithm. Thus, indeed 100 parameter sets were found for each dataset for each catchment. The median of the model performance for these 100 parameter sets is described in the text (and compared to the median performance of the model calibrated with the streamflow data).

For the lower benchmark, the 1000 random parameter sets result in 1000 model simulation results. We used the median model performance from these 1000 simulations to represent the performance of a model with random parameters.