

Interactive comment on “Information content of stream level class data for hydrological model calibration” by Ilja van Meerveld et al.

W. Buytaert (Referee)

w.buytaert@imperial.ac.uk

Received and published: 28 April 2017

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.

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 per-

C1

formance 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.

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 streamflow 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).

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.