Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-72, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 14 February 2017

Hydrology and Earth System Sciences
Discussions

Information content of stream level class data for hydrological model calibration

Ilja van Meerveld[1], Marc Vis[1], Jan Seibert[1,2]

[1] Department of Geography, University of Zurich, Zurich, Switzerland
[2] Department of Earth Sciences, Uppsala University, Uppsala, Sweden

Correspondence to: Ilja van Meerveld (ilja.vanmeerveld@geo.uzh.ch)

Abstract. Citizen science can provide spatially distributed data over large areas, including hydrological data. Stream levels are easier to measure than streamflow and can be observed more easily by citizen scientists. However, the challenge with crowd-based stream level data is that observations are taken at irregular time intervals and with a limited vertical resolution. The latter is especially the case at sites where no staff gauge is available and relative stream levels are observed based on (in)visible features in the stream, such as rocks. In order to assess the potential value of crowd-based stream level observations for model calibration, we pretended that stream level observations were available at a limited vertical resolution by transferring streamflow data into stream level classes. A bucket-type hydrological model was calibrated with these hypothetical data sets and subsequently evaluated on the observed streamflow records. Our results indicate that stream level data can result in good streamflow simulations, even with a reduced vertical resolution of the observations. Time series of only two stream level classes, e.g. above or below a rock in the stream, were already informative, especially when the class boundary was chosen towards the highest stream levels. There was some added value in using up to five stream level classes but there was hardly any improvement in model performance when using more level classes. These results are encouraging for citizen science projects and provide a basis for designing observation systems that collect data that are as informative as possible for deriving model-based streamflow time series for previously ungauged basins.

Keywords. information content, stream levels, hydrological model calibration, citizen science, measurement resolution, ungauged catchments

1 Introduction

Streamflow data are crucial for water resources management decisions and the calibration of hydrological models. However, streamflow data are only available for a number of sites and gauging stations are not always installed at representative locations. There is, for instance, a lack of streamflow gauges in small headwater streams (Kirchner, 2006) and in developing countries (Mulligan, 2013). Although technological developments provide the possibility to expand the measurement network, the reality is that due to budget cuts, observation networks often shrink (Kundzewicz, 1997), rather than expand. Remote sensing images can be used to estimate stream levels or streamflow, particularly for wide lowland rivers (Milewski et al., 2009; Pavelsky, 2014; Smith, 1997; Van Dijk et al., 2016) but estimation of streamflow from satellite images is likely to remain problematic for small headwater streams.

Stream level data are easier to obtain than streamflow data because they do not require any information on the rating curve. Seibert and Vis (2016) tested if stream level data can be used to constrain a simple hydrological model. The results for ~600 catchments in the USA showed that level data can be surprisingly informative for hydrological model calibration. This applies especially for humid and wet catchments (defined as catchments where the annual precipitation is larger than the annual potential evapotranspiration), for which the median model efficiencies (Nash and Sutcliffe, 1970) of models calibrated with stream level data were generally only about 0.10-0.15 units below those of models calibrated with streamflow data and for all but one catchment the difference was less than 0.17. For dry catchments, additional information on the volume of streamflow, such as the annual mean flow or streamflow percentiles, were needed.

Even though the price for water level recorders has significantly gone down in recent years and their datalogging capacity has increased, it is not feasible to install a water level recorder in every ungauged catchment. It is, therefore, useful to also consider the use of other approaches to obtain water level data. Citizen science is now more frequently used to obtain environmental data over large areas (Bonney et al., 2009; Fohringer et al., 2015; Graham et al., 2011; Huddart et al., 2016; Savan et al., 2003; Wiseman and Bardsley, 2016). Little *et al.* (2016) gave citizen scientists water level sounders to measure groundwater levels in private wells and found that these measurements provided valuable data on groundwater levels across a large area in Alberta, Canada, and that the measurements were relatively accurate; the root mean square error between citizen scientist observed water levels and pressure transducer based water levels ranged between 3 and 11 cm. Lowry and Fienen (2013) installed staff gauges in rivers and asked citizen scientists to send stream level measurements via text message. They showed that the accuracy of the crowd-sourced measurements and pressure transducer data were similar to the staff gauge gradations (root mean square error of 0.5 cm). However, it is not feasible to install a staff gauges in every ungauged catchment or to equip all citizen scientists with water level recorders. Therefore, it is useful to also design citizen science approaches that do not require staff gauges or water level sensors. Citizen scientists have for example successfully mapped the occurrence of streamflow in intermittent streams (Turner and Richter, 2011) and water levels are a standard measurement in the Earthwatch FreshWater Watch program (https://freshwaterwatch.thewaterhub.org/). Estimates of relative stream levels or stream level classes based on features in the stream or on the streambank (i.e., whether the water is above or below a certain rock) are easier and can be done more quickly than actual water level measurements and are, therefore, likely suitable for citizen science projects where no staff gauges are available. However, the (vertical) resolution of these data is less than those of actual stream level measurements.

Information from time lapse cameras or webcams can also be used to obtain information on stream water level classes. Pixel classification or image recognition to determine whether the water level is above or below a certain point can be used to determine the relative stream water level, even if no other information about the stream or the cross section is available. Several studies have shown that cameras can be used for accurate streamflow estimation (Hilgersom and Luxemburg, 2012; Muste et al., 2011; Royem et al., 2012; Stumpf et al., 2016; Tsubaki et al., 2011) but these studies used dedicated cameras that focused directly on the stream and often required information about the stream channel cross section. While promising, it is unlikely that many of the ungauged streams will be equipped with these systems. However, streams are often included in the pictures of existing webcams and time lapse cameras that were installed for other reasons, e.g., to show the snow conditions on a ski-slope or to highlight the view from a hotel. The information from these webcams can be used to obtain information about the relative changes in the stream level or width but this information might not be very precise because of the sub-optimal angle of the camera. It is, thus, more likely that these images can be used to obtain information about the relative water level or stream width (class), rather than the actual water level. Remotely sensed satellite data can also be used to rank stream levels or stream width. These data, however, as promising as they are, have limitations regarding their accuracy and resolution (and will likely have so for the foreseeable future). Thus also for these measurements time series of level (or width) classes are more realistic than high-resolution time series of actual water levels.

For crowd-based (or citizen science) observations, but also for data from webcams or satellites, the resolution of the stream level data will be significantly poorer than for data obtained by a dedicated water level sensor. To determine the effect of this loss of information, we tested the usefulness of these new types of stream level class data for constraining a simple bucket-type hydrological model. The aim was to provide a basis for designing citizen science projects that collect data that are as informative as possible and that can be used to derive model-based streamflow time series. We pretended that stream level class observations were available continuously (daily) but only at a limited vertical resolution by transferring the streamflow data into stream level classes. We then tested how the number of stream level classes (i.e., the degree of resolution) influenced the information content of the data with regard to constraining the model. Furthermore, we studied the effect of different locations of the class boundaries on model performance.

2. Methods

2.1 Study catchments and dataset

This study largely followed the methodology of Seibert and Vis (2016), who used Spearman rank correlation to calibrate a hydrological model for almost 600 catchments in the contiguous US based on continuous, high-resolution stream level data. In this study the model was calibrated based on stream level class data for a subset of catchments. The 100 catchments used in this study were chosen randomly from the catchments used by Seibert and Vis (2016) and are spread across the contiguous US. The hydrometric data for these 1 to 12584 km^2 catchments were obtained from the dataset for 671 catchment of Newman *et al*. (2015). The mean annual precipitation (P) was derived from DAYMET (Thornton et al., 2012) and varied for the different catchments between 249 and 3113 mm y^{-1}. The potential evapotranspiration (E_{pot}) was calculated with the Priestley–Taylor equation. The annual average runoff ratios calculated based on the precipitation at the mean elevation varied between 0.05 and 1.18 (between 0.12 to 0.93 for 90% of the catchments). The aridity index (P/E_{pot}) varied between

Hydrology and Earth System Sciences Discussions
Open Access

0.25 and 4.33. Of the 100 catchments, 22 are considered dry ($P/E_{pot} \leq 1.0$), 62 are considered humid ($1.0 < P/E_{pot} < 2.0$) and 16 catchments are considered wet ($P/E_{pot} \geq 2.0$).

2.2 Transformation of streamflow data into stream level classes

In order to determine how many stream level classes are needed for model calibration, the streamflow data were converted into time series of n stream level classes, where n varied from 2 to 20. In real citizen science projects the class boundaries are likely chosen based on features in the stream or on the stream bank (e.g. above or below a certain rock or marker) but in this study we chose the boundaries so that each class contained the same number of data points (i.e. each class had observations for a fraction of n^{-1} of the entire time period). For the situations with two and three stream level classes, we also systematically varied the class boundaries by changing the fraction of the streamflow data in each class to determine the optimal location of the class boundaries.

2.3 Hydrological Model

The HBV (Hydroloiska Byråns Vattenavdelning) model (Bergström, 1992; Lindström et al., 1997) was used in the software implementation HBV light (Seibert and Vis, 2012). The HBV model is a frequently used bucket-type model and consists of different routines representing snow, soil, groundwater and stream routing processes. The HBV model, as it was applied here, has 14 free parameters, which are usually found by calibration or regionalisation. Elevation bands of 200 m were used to represent catchment topography, whereas only one lumped land-cover class was used for each catchment. The parameter ranges for the 14 model parameters in the HBV model were similar to those used by Seibert and Vis (2016) and represent the range of typical parameter values found in previous studies worldwide.

2.4 Model calibration and validation

For each catchment the HBV model was calibrated for the period 1.10.1982 - 30.9.1996 using a genetic optimization algorithm (Seibert, 2000). The data from the 1.1.1980 - 30.9.1982 period were used for warming up the model. For model calibration, we maximized the Spearman rank correlation coefficient (r_s) between the stream level class data and the simulated streamflow. The Spearman rank correlation evaluates the dynamics of the modeled streamflow but not the streamflow volume and is highest (r_s =1) when stream level and streamflow are monotonically related. The advantage of using the Spearman rank correlation for model calibration based on stream level class data is that no information about the rating curve is needed. The use of class data leads to a large number of ties (measurements with the same (mean) rank for the water level class) and r_s values of one can, thus, not be obtained. However, r_s can still be used for model calibration because its value is highest when the dynamics of the stream level classes and streamflow are most similar. For each catchment, the model was calibrated 100 times, with each calibration trial consisting of 3500 model runs.

The 100 calibration parameter sets for each catchment were validated by comparing the simulated streamflow to the observed streamflow data using the model efficiency measure (R_{eff}). The median value of R_{eff} for the 100 parameter sets for each catchment was used to represent the performance of the model for that catchment.

2.4 Benchmarks

Different benchmarks were used to assess the performance of the models calibrated with the stream level class data: an upper benchmark that represents how good the model simulation would be if continuous streamflow data were available, and two lower benchmarks that represent a model simulation in the absence of any streamflow or stream level data.

For the upper benchmark (R_{eff}), the model was calibrated for each catchment using the streamflow data and optimizing the model efficiency for the observed and modeled streamflow (100 calibration trials per catchment, each consisting of 3500 model runs). The median model efficiency of these 100 calibration trials was used as the upper benchmark value for each catchment. Because the goal of this study was to assess the value of stream level class data for model calibration, rather than to evaluate the ability of the model to simulate the streamflow, all model validation results for the stream level class data are given as the difference in R_{eff} relative to this upper benchmark.

In addition, the simulations based on the stream level class data were also compared to the simulations based on calibrations derived from high-resolution stream level data (r_{s_∞}). Here the model was calibrated by optimizing the Spearman rank correlation between the observed and modeled streamflow (c.f. Seibert and Vis, 2016). These simulations represent a situation where a water level recorder is installed in the catchment and this data is used for model calibration.

For the first lower benchmark (L_{random}), the model was run for each catchment 1000 times using randomly chosen parameters within the parameter ranges that were used for model calibration. For the second lower benchmark ($L_{regional}$), the model was run 9900 times using the 100 calibrated parameter sets of each of the 99 other catchments.

3. Results

3.1 Model performance as a function of the number of water level classes

Not surprisingly, the model efficiency (R_{eff}) of the models calibrated with the stream level class data was lower than for the models calibrated with the streamflow data. However, the differences between the models calibrated with the high-resolution stream level data and the models calibrated with water level class data was relatively small, as long as at least five stream level classes were used for model calibration (compare results for r_{s_5} and r_{s_∞} in Figure 1). The median difference in model performance for the models calibrated on high-resolution water level data and the models calibrated on five stream level classes was only 0.01. The median difference was 0.06 when two stream level classes were used. These differences are relatively small compared to the 0.17 difference in median model efficiency for the models calibrated on continuous streamflow (R_{eff}) and the high-resolution stream level data (r_{s_∞}).

A more detailed analysis of the increase in model performance with an increasing number of water level classes suggests that for the wet catchments model performance increased only slightly when increasing the number of water level classes from two to five but that for some of the dry catchments model performance increased significantly when using more than two water level classes (Figure 2). In general, the increase in model performance with an increasing number of stream level classes was largest for the catchments with the largest difference in model performance between the upper and lower benchmarks (Figure 2).

3.2 Comparison with the benchmarks

Comparison of the performance of the models calibrated with stream level class data to the upper benchmark suggests that especially for the wet catchments the differences between traditional model calibration based on continuous streamflow data and the calibration based on the stream level class data were small (Figure 3a-b). For the dry catchments, model calibration based on stream level class data led to larger errors in the simulated streamflow (Figure 3a-b).

Comparison of model performance for the models calibrated with the stream level class data to the lower benchmarks suggests that the inclusion of stream level class data led to a huge improvement in model performance for some of the dry catchments (Figure 3c-d). However, the median improvement in model efficiency when using the two stream level class data compared to the lower benchmark (L_{random}) between the wet, humid and dry catchments was small (0.23, 0.23 and 0.15, respectively) and not statistically significant (Kruskal Wallace test p=0.09). The differences in the median improvement in the efficiency when using the five stream level class data compared to the lower benchmark (L_{random}) between the wet, humid and dry catchments were also small (0.23, 0.32 and 0.22, respectively) but statistically significant (Kruskal Wallis test p=0.018).

3.3 Optimal location of class boundaries

In order to determine the optimal location of the class boundaries, we systematically varied them for the cases with two and three water level classes. The results show that model performance generally improved when at least one class boundary was located at high stream levels. For example, for the case with two classes, the median model performance for the 100 catchments was highest when the class boundary was chosen so that the stream water level was in the lower class for 94% of the time and in the upper class for 6% of the time. The smallest median difference between the model performance for two classes and the upper benchmark occurred at the class boundaries of 93-7% (Figure 4a). The variability in model performance also decreased when the boundary was chosen at a higher stream water level, so that for fewer catchments the difference between the median model performance (i.e., median performance of the 100 calibration parameter sets) and the upper benchmark was larger than 0.20 (the difference in R_{eff} for models calibrated on streamflow data and models calibrated with data for two water level classes was larger than 0.20 for 86, 61, 22, and 22% of the catchments when the boundary was set at 10-90, 50-50, 90-10, and 94-6% of the time, respectively). There was no clear spatial pattern in the optimal location of the class boundaries and for a few catchments the optimal boundary was located at a much lower stream level (Figure 4b).

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-72, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 14 February 2017

For the case with the three water level classes, on average for the 100 catchments, better model results were obtained when the boundary for the upper class was at the high water level, but the other boundary could either be at a high level or at a low level (Figure 5). Intermediate values for the other boundary resulted in poorer model performance. The median performance of the models calibrated with three water level classes for the 100 catchments was highest when the class boundaries were set so that the water level was in the lowest, medium and highest class 94, 5 and 1% of the time, respectively.

4. Discussion

4.1 Usefulness of stream level class data

The results of this study show that five stream level classes are as informative for model calibration as stream level data with a very high vertical resolution. This is good news for citizen science projects or webcam based analyses, as it is much easier to determine the stream level class when there are only a few classes than when there are many classes. The small difference between the performance of the models calibrated on data for a few stream level classes and the upper benchmark (Figure 3a-b) suggests that the stream level class data from citizen science approaches or webcam images is most useful for model calibration for wet catchments and that stream level class data for these catchments can be used in combination with a model to obtain time series of streamflow. This is useful, as it is much harder for citizen scientists to estimate streamflow than stream level classes and this way the streamflow data that are needed for water management or flood- or drought forecasting can be obtained from the stream level class data.

On the other hand, the large improvement of the models calibrated with stream level class data compared to the lower benchmark for some of the dry catchments (Figure 3c-d) suggests that stream level class data may be especially useful in improving model performance in some dry catchments when no other streamflow or stream level data are available. For these catchments, the model performance of the lower benchmark (i.e. based on the random parameter sets) was very poor, while for the wet catchments the model performance of the lower benchmarks was already reasonably good (see color coding in Figure 2 and Figure 3). Thus the biggest gain in adding stream level class data was seen for the dry catchments, even though the absolute model performance was much poorer than for models calibrated on streamflow data. Seibert and Vis (2016) showed that model calibration based on high-resolution stream level data worked best for wet catchments and that for dry catchments, additional data on the water balance were needed. Using such additional information may also improve model performance based on stream level class data for the dry catchments. What kind of additional information might be most useful in combination with stream level class data remains to be explored.

4.2 Location of the class boundaries

In practice, the boundaries between the different water level classes will be chosen based on features in the river or the stream bank that are easy to observe. The results from this study suggest that for most streams the optimal class boundaries should be located at the high flow levels, but not at the very highest flows. This high optimal class boundary is good news for model calibration based on opportunistic webcam images because high flows are usually easier to observe in these images, while it may be difficult to see the water level at low flows when the camera does not focus directly on the stream. Citizen scientists, on the other hand, are perhaps more likely to go out and estimate stream levels during nice weather conditions and low flow periods. However, people also tend to look at rivers when the water level is particularly high. The still relatively long time that the water level is in the highest class (e.g. 6% of the time or 22 days per year for the case with only 2 water level classes) suggests that there is ample time for citizen scientist to observe the water levels during the high water level period. These results thus suggest that citizen science projects should communicate to the participants that measurements during high water levels are important and worth transmitting.

The reasons that for the majority of the catchments the optimal boundary between the different water level classes is located at high stream levels are related to the data, the model and the choice of the model evaluation criterion. The choice of a high water level class boundary helps to avoid the selection of a model parameter set that leads to a too flashy streamflow response because the water level is in the upper water level class for only a limited fraction of time. The information content of the water level class data, and thus its value for hydrological model calibration, is higher when we know that for some events the water level doesn't cross this boundary and for another set of events it does. If for every event, the water level crosses the boundary because it is set at a low level, then it is not possible to distinguish between the responses of the different events. Similarly, if the level is set too high, then the water level may cross the water level class boundary only a very few times so that no distinction can be made for the response of the majority of the events. For the optimal boundary for two classes at 94-6% of the time, the streams crossed the class boundaries on average between 1.1 and 13.6 times per year (median: 7.2; 25th and 75th percentile 4.0 and 8.5 respectively). One could also argue that the water level class data is most informative when the class boundaries are crossed as often as possible in the actual time series. For two stream level classes this would mean a class boundary between 60 and 80% for three quarters of the catchments.

The optimal location of the water level class boundaries is also dependent on the model validation criterion that is used. We used the Nash-Sutcliff model efficiency to evaluate model performance, which is known to give more weight to the evaluation of high flows (Krause et al., 2005; Schaefli and Gupta, 2007). A high water level class boundary provides more information for these high flows. Using a different model evaluation criterion that focuses less on the high flows would result in lower optimal class boundaries. For example, using the efficiency of the log-transformed streamflow to evaluate the model performance the model efficiency values (again median for the 100 catchments) would be highest when the class boundary is chosen so that the stream water level is in the lower class for about 60% of the time when there are only two stream level classes, and the water level was in the lower, middle and upper class for about 10, 60, and 30% of the time when there are three water level classes. In other words, the exact location of the optimal water level class boundaries depends on the model evaluation criteria and should be chosen based on the objective of the study (simulation of the peaks, low flow periods or the water balance). Because in real citizen science projects the boundaries will not be chosen based on optimality as discussed above, but will be chosen by citizens based on local conditions, such as identifiable features in the stream, this means that the usefulness of citizen science based water level class data for the simulation of different aspects of the

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-72, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 14 February 2017

hydrograph will differ. However, the investigation of theoretically optimal class boundaries is still valuable for at least two reasons. Firstly, these results can be used to provide guidance to citizen scientists on how to choose boundary levels if at all possible. Secondly, such results can help to decide which citizen science based water level class data might be especially useful for the simulation of certain aspect of the hydrographs.

4.3 Limitations of this study when faced with the reality of citizen science based data collection

A challenge with citizen science-based stream level data is that observations are taken at irregular time intervals, with a limited vertical resolution and may contain errors. In this study, we addressed the issue of the limited vertical resolution by assessing the value of stream level class data. More work is needed on the issue of irregular data to determine the number of observations that are needed and the best times of these observations. Model calibration based on streamflow measurements suggest that continuous streamflow data are not needed and only a few streamflow measurements, particularly during rainfall events, are already useful to constrain hydrological models because many of the streamflow measurements contain redundant information (Rojas-Serna et al., 2016; Seibert and Beven, 2009).

In this study, we pretended to have stream level class data by transforming the streamflow data to stream level classes. This data, therefore, does not include any errors in the stream level classes. In reality, citizen science data may contain errors and misclassification of the water levels. The effects of data errors on model results needs to be tested as well. However, in this respect, it has to be mentioned that several studies have shown that citizen science data can be quite accurate (Cohn, 2008; Lowry and Fienen, 2013; Tye et al., 2016) (but not always, e.g. (Savan et al., 2003)) and that traditional streamflow data also can have significant uncertainties and may even contain dis-informative information that affects model calibration as well (Beven and Westerberg, 2011; McMillan et al., 2010).

5. Conclusion

This study demonstrates that stream level class data can be useful for constraining hydrological models. The results confirm the conclusions from a previous study (Seibert and Vis, 2016) but more importantly extend the findings towards the use of stream level data for model calibration to cases where data is available at only a limited vertical resolution, such as in citizen science-based observation approaches or webcam image analysis. The results show that a small number of stream level classes contain almost as much information as high-resolution water level data for hydrological model calibration. This is good news for citizen science approaches. We also found that class boundaries at high water levels result in the most informative water level class time series. While the class boundaries in practice are likely determined by the local situation (such as a rock that is covered by water at a certain level), the importance of high levels shows the importance of motivating the public to collect data during high flow situations.

More generally, this study demonstrates how hydrological modeling can be used to evaluate the potential value of certain types of data. Similar approaches can be used to evaluate how much the information content of water level class data might decrease if observations are made at irregular times or with a certain amount of error. This information is crucial for the optimal design and implementation of citizen science-based observation approaches.

Acknowledgements

We thank Andy Newman and Martyn Clark for making the data used in this study available. The ScienceCloud provided by S3IT at the University of Zurich enabled us to run the computational-intensive simulations on virtual machines.

References

Bergström, S.: The HBV model: Its structure and applications, Swedish Meteorological and Hydrological Institute, 1992.

Beven, K., and Westerberg, I.: On red herrings and real herrings: disinformation and information in hydrological inference, Hydrological Processes, 25, 1676-1680, 2011.

Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., and Shirk, J.: Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy, Bioscience, 59, 977-984, 10.1525/bio.2009.59.11.9, 2009.

Cohn, J. P.: Citizen Science: Can Volunteers Do Real Research?, Bioscience, 58, 192-197, 10.1641/b580303, 2008.

Fohringer, J., Dransch, D., Kreibich, H., and Schröter, K.: Social media as an information source for rapid flood inundation mapping, Nat. Hazards Earth Syst. Sci., 15, 2725-2738, 10.5194/nhess-15-2725-2015, 2015.

Graham, E. A., Henderson, S., and Schloss, A.: Using mobile phones to engage citizen scientists in research, Eos, Transactions American Geophysical Union, 92, 313-315, 10.1029/2011eo380002, 2011.

Hilgersom, K., and Luxemburg, W.: Technical Note: How image processing facilitates the rising bubble technique for discharge measurement, Hydrology and Earth System Sciences, 16, 2012, 2012.

Huddart, J. E. A., Thompson, M. S. A., Woodward, G., and Brooks, S. J.: Citizen science: from detecting pollution to evaluating ecological restoration, Wiley Interdisciplinary Reviews: Water, 3, 287-300, 10.1002/wat2.1138, 2016.

Krause, P., Boyle, D. P., and Bäse, F.: Comparison of different efficiency criteria for hydrological model assessment, Adv. Geosci., 5, 89-97, 10.5194/adgeo-5-89-2005, 2005.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S.: Development and test of the distributed HBV-96 hydrological model, Journal of Hydrology, 201, 272-288, http://dx.doi.org/10.1016/S0022-1694(97)00041-3, 1997.

Little, K. E., Hayashi, M., and Liang, S.: Community-Based Groundwater Monitoring Network Using a Citizen-Science Approach, Groundwater, 54, 317-324, 10.1111/gwat.12336, 2016.

Lowry, C. S., and Fienen, M. N.: CrowdHydrology: Crowdsourcing Hydrologic Data and Engaging Citizen Scientists, Ground Water, 51, 151-156, 2013.

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, Hydrological Processes, 24, 1270-1284, 2010.

Milewski, A., Sultan, M., Yan, E., Becker, R., Abdeldayem, A., Soliman, F., and Gelil, K. A.: A remote sensing solution for estimating runoff and recharge in arid environments, Journal of Hydrology, 373, 1-14, 2009.

Mulligan, M.: WaterWorld: a self-parameterising, physically based model for application in data-poor but problem-rich environments globally, Hydrol. Res., 44, 748–769, 2013.

Muste, M., Ho, H. C., and Kim, D.: Considerations on direct stream flow measurements using video imagery: Outlook and research needs, Journal of Hydro-environment Research, 5, 289-300, 2011.

Nash, J. E., and Sutcliffe, J. V.: River flow forecasting through conceptual models part I — A discussion of principles, Journal of Hydrology, 10, 282-290, http://dx.doi.org/10.1016/0022-1694(70)90255-6, 1970.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., and Duan, Q.: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance, Hydrol. Earth Syst. Sci., 19, 209-223, 10.5194/hess-19-209-2015, 2015.

Pavelsky, T. M.: Using width-based rating curves from spatially discontinuous satellite imagery to monitor river discharge, Hydrological Processes, 28, 3035-3040, 10.1002/hyp.10157, 2014.

Rojas-Serna, C., Lebecherel, L., Perrin, C., Andréassian, V., and Oudin, L.: How should a rainfall-runoff model be parameterized in an almost ungauged catchment? A methodology tested on 609 catchments, Water Resources Research, 52, 4765-4784, 10.1002/2015wr018549, 2016.

Royem, A. A., Mui, C. K., Fuka, D. R., and Walter, M. T.: Technical note: Proposing a low-tech, affordable, accurate stream stage monitoring system, Transactions of the ASABE, 55, 2237-2242, 2012.

Savan, B., Morgan, J. A., and Gore, C.: Volunteer Environmental Monitoring and the Role of the Universities: The Case of Citizens' Environment Watch, Environmental Management, 31, 0561-0568, 10.1007/s00267-002-2897-y, 2003.

Schaefli, B., and Gupta, H. V.: Do Nash values have value?, Hydrological Processes, 21, 2075-2080, 2007.

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

Seibert, J., and Beven, K. J.: Gauging the ungauged basin: how many discharge measurements are needed?, Hydrology and Earth System Sciences, 13, 883-892, 2009.

Seibert, J., and Vis, M. J. P.: Teaching hydrological modeling with a user-friendly catchment-runoff-model software package, Hydrol. Earth Syst. Sci., 16, 3315-3325, 2012.

Seibert, J., and Vis, M. J. P.: How informative are stream level observations in different geographic regions?, Hydrological Processes, 30, 2498-2508, 10.1002/hyp.10887, 2016.

Smith, L. C.: Satellite remote sensing of river inundation area, stage, and discharge: A review, Hydrological Processes, 11, 1427-1439, 1997.

Stumpf, A., Augereau, E., Delacourt, C., and Bonnier, J.: Photogrammetric discharge monitoring of small tropical mountain rivers: A case study at Rivière des Pluies, Réunion Island, Water Resources Research, 52, 4550-4570, 10.1002/2015wr018292, 2016.

Tsubaki, R., Fujita, I., and Tsutsumi, S.: Measurement of the flood discharge of a small-sized river using an existing digital video recording system, Journal of Hydro-environment Research, 5, 313-321, 2011.

Turner, D., and Richter, H.: Wet/Dry Mapping: Using Citizen Scientists to Monitor the Extent of Perennial Surface Flow in Dryland Regions, Environmental Management, 47, 497-505, 10.1007/s00267-010-9607-y, 2011.

Tye, C. A., McCleery, R. A., Fletcher, R. J., Greene, D. U., and Butryn, R. S.: Evaluating citizen vs. professional data for modelling distributions of a rare squirrel, Journal of Applied Ecology, n/a-n/a, 10.1111/1365-2664.12682, 2016.

Van Dijk, A. I. J. M., Brakenridge, G. R., Kettner, A. J., Beck, H. E., De Groeve, T., and Schellekens, J.: River gauging at global scale using optical and passive microwave remote sensing, Water Resources Research, n/a-n/a, 10.1002/2015wr018545, 2016.

Wiseman, N. D., and Bardsley, D. K.: Monitoring to Learn, Learning to Monitor: A Critical Analysis of Opportunities for Indigenous Community-Based Monitoring of Environmental Change in Australian Rangelands, Geographical Research, 54, 52-71, 10.1111/1745-5871.12150, 2016.

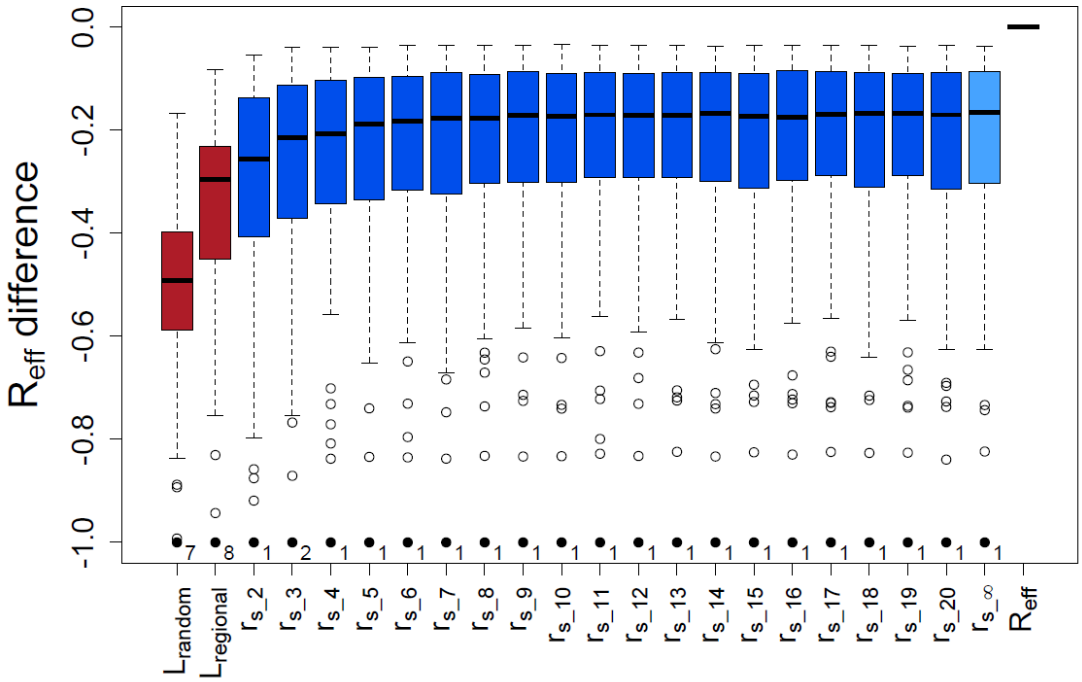

Figure 1. Box plots of the median model validation results (R_{eff}) relative to the upper benchmark for the models calibrated on stream level class data (2 to 20 classes; r_{s_n}), models calibrated on high-resolution stream level data (r_{s_∞}) and the two lower benchmarks (L_{random} and $L_{regional}$) for all 100 catchments.

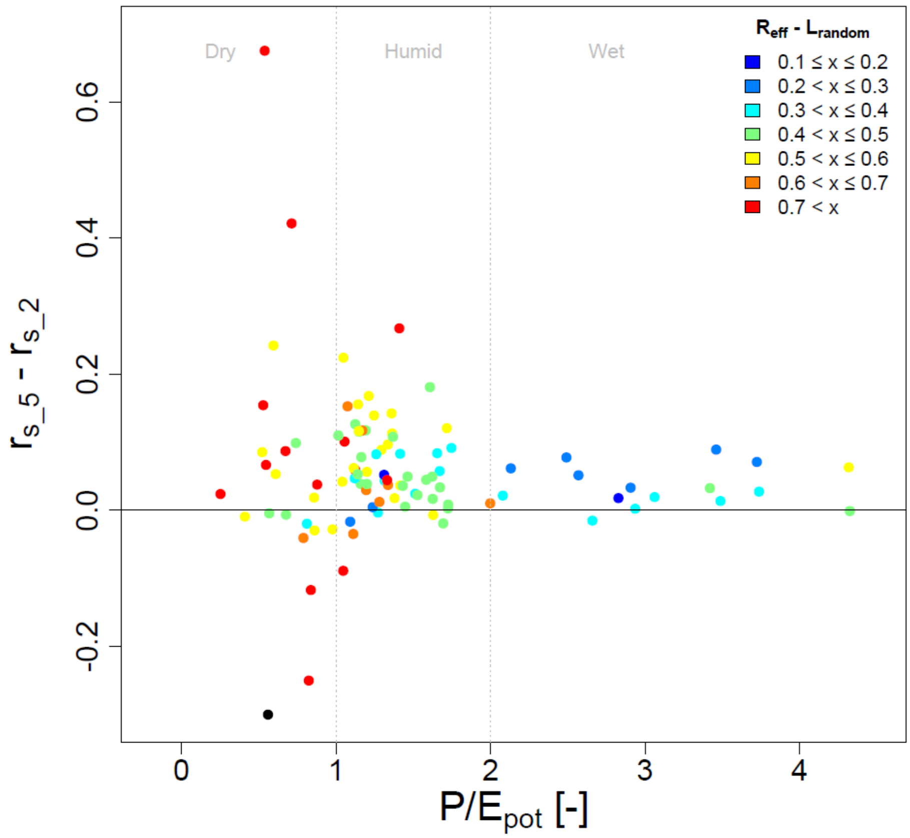

Figure 2. Difference in the median model validation results (R_{eff}) for the models calibrated using two water level classes (r_{s_2}) and five water level classes (r_{s_5}) for all 100 catchments as a function of the aridity index. The color of the symbols represents the difference between the upper and the lower benchmark (i.e. the difference in the median model performance when the model is calibrated with all available streamflow data (R_{eff}) and when the model is run with randomly selected parameters (i.e. without any calibration; L_{random})).

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-72, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 14 February 2017

Hydrology and Earth System Sciences
Discussions

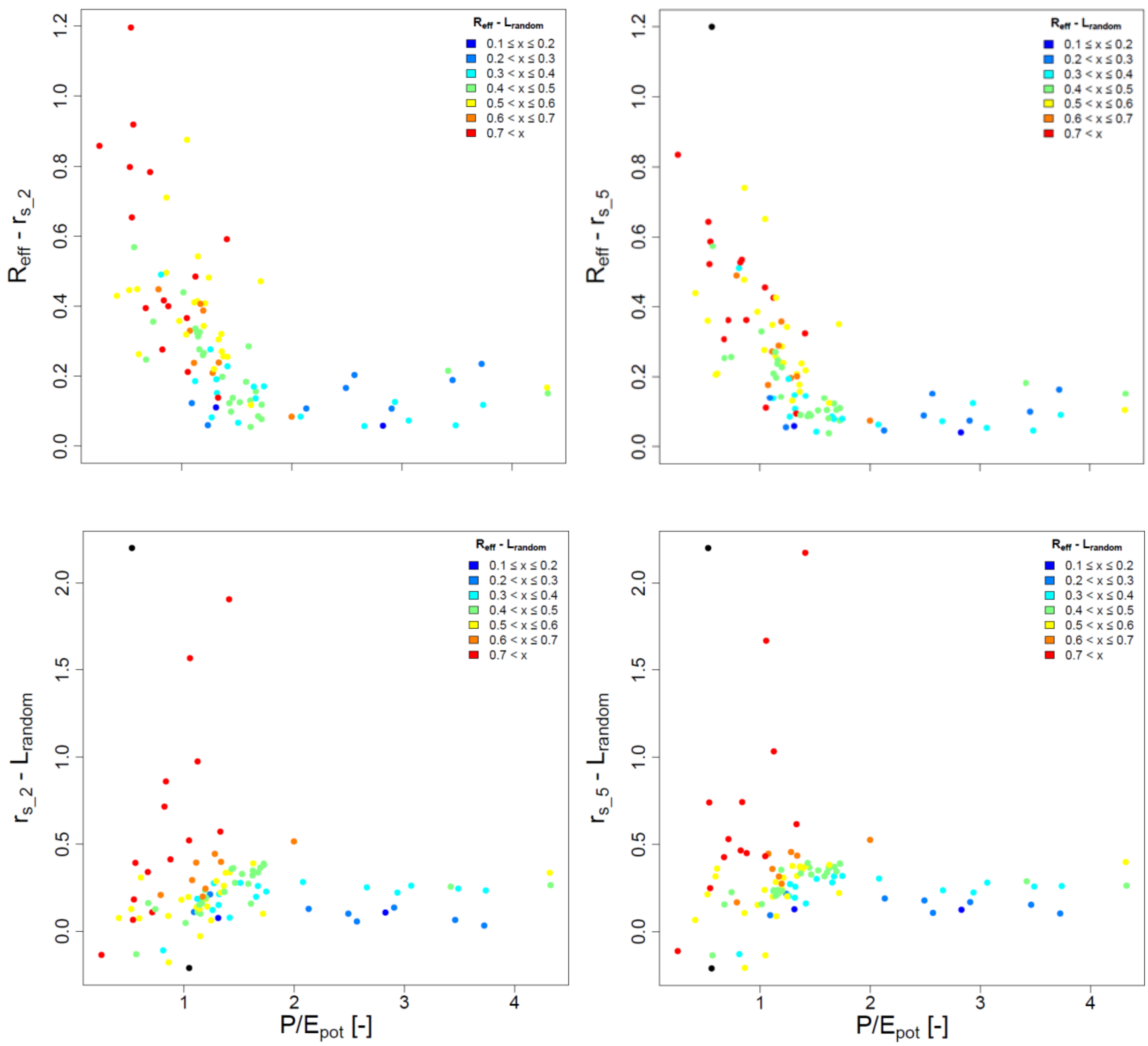

Figure 3. Difference in model validation results (R_{eff}) for the models calibrated with data from two (r_{s_2}; left) and five (r_{s_5}; right) stream level classes with the upper benchmark (upper row) and the lower benchmark (L_{random}; bottom row) as a function of the aridity index. Each dot represents one catchment; the color of the symbol represents the difference between the upper and lower benchmark for that catchment. Note the difference in the scale of the y-axis for the comparison to the upper benchmark (upper row; a and b) and the lower benchmark (lower row; c and d).

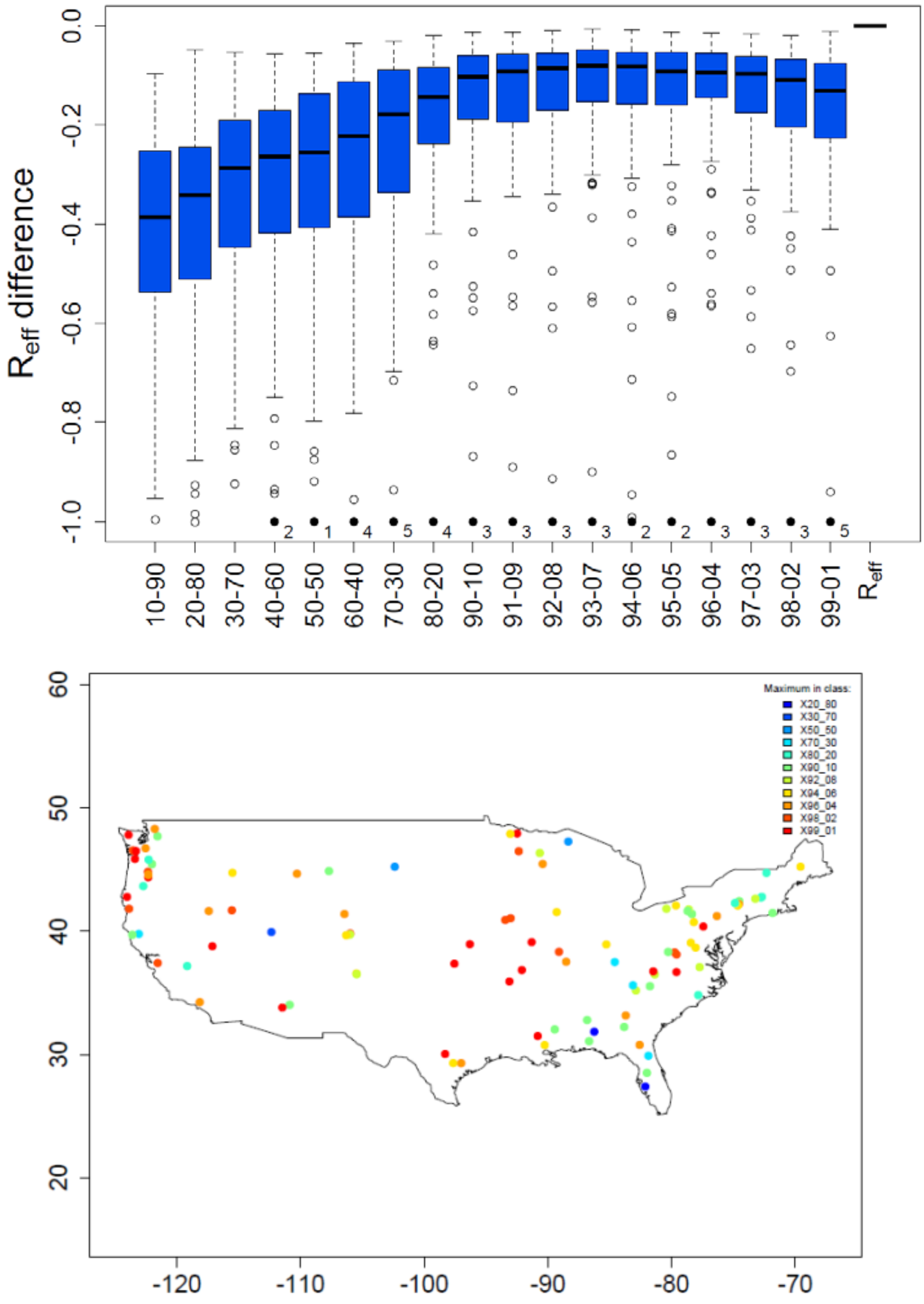

Figure 4. Difference in median model validation results (R_{eff}) relative to the upper benchmark for models calibrated with two water level classes for different class boundaries (a) and map of the optimal class boundary for each catchment (b). 10-90 indicates that streamflow was in the lower water level class for 10% of the time and in the upper class for 90% of the time. The median difference in model efficiency is smallest when the class boundaries are set at 93 and 7%.

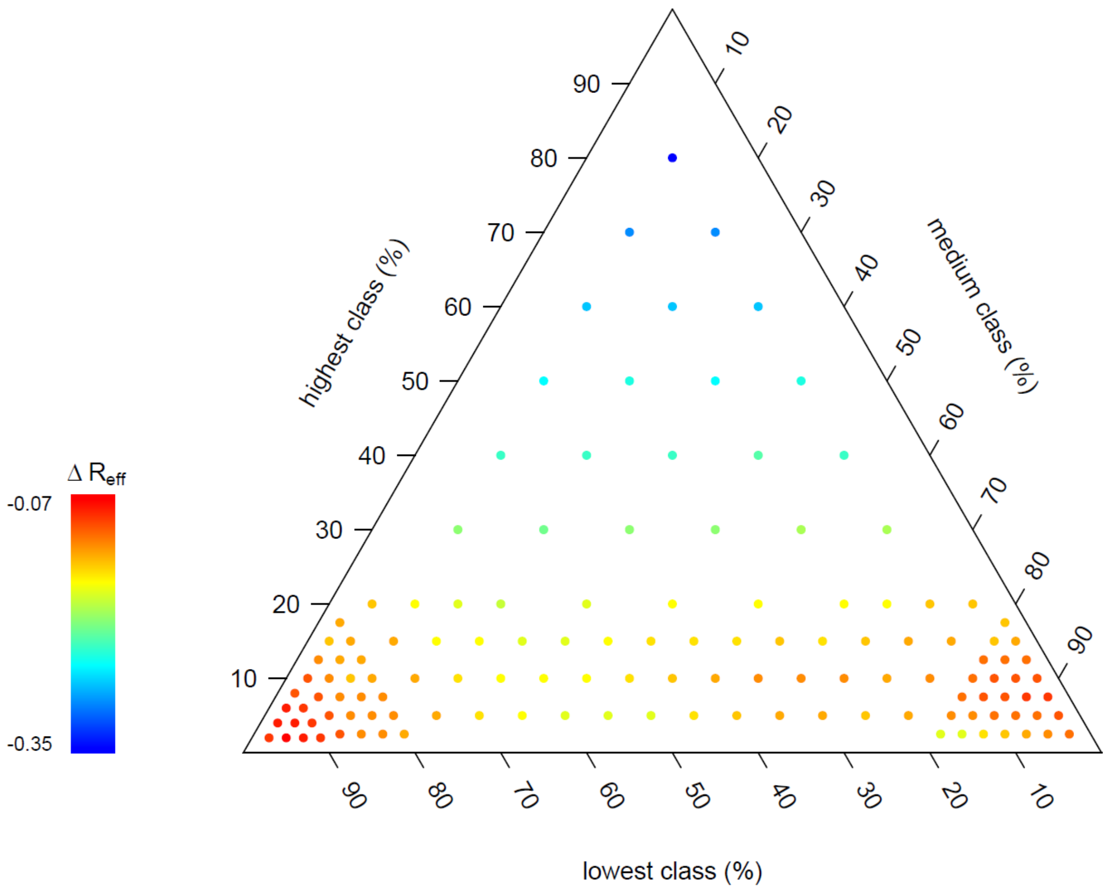

Figure 5. Median difference in model efficiency (R_{eff}) for models calibrated with data for three water level classes relative and the upper benchmark for different class boundaries.