Hydrol. Earth Syst. Sci. Discuss.,
doi:10.5194/hess-2016-546-AC2, 2016

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Streamflow characteristics from modelled runoff time series – importance of calibration criteria selection" *by* Sandra Pool et al.

Sandra Pool et al.

sandra.pool@geo.uzh.ch

Received and published: 5 December 2016

Dear reviewer,

Thank you for your efforts with our manuscript. We greatly appreciated the comments. Please see below our detailed response to each of the comments.

Best regards,

Sandra Pool and Co-Authors

General comments RC 1

Comment 1: It is a very good idea and a well know idea, that the models should be calibrated using criterions relevant to the purpose of the model and simulations. The criterions here are many but seems relevant to ecological studies. The relevance of the choosen SFC should be argued for in relevance to the purpose. Here are maybe to many SFC and the it becomes difficult to keep track of them. Are they correlated, are they in contradiction (to fulfill one means that another suffer) etc. It might be an idea to pick the most important for the kind of studies the model should be used in and discuss this closer.

Reply 1: We agree that we used many SFCs. There are more than 150 different SFCs used in ecohydrological studies and many of them are correlated or redundant (Olden and Poff, 2003). In our study we use 13 SFCs that have been shown to be the most relevant ones for the ecological integrity of the study catchments in previous studies (see chapter 2.2 for the motivation of the selected SFCs). We are aware that some of the selected 13 SFCs are correlated, e.g. FH6 and FH7 (we will add a correlation analysis in the revised manuscript), but still decided to use these ecologically most relevant SFCs in our study. As you suggest, selecting the most important SFC would allow a deeper analysis but also reduce the potential for general conclusions. Our 13 ecologically relevant SFCs have the advantage of representing different flow components and flow conditions.

Calibrating the model for one SFC can indeed have a negative effect on the estimation accuracy of another SFC (e.g. calibrating for FH6 (high flow) could negatively affects ML20 (low flow)) (see also our discussion in chapter 4.1). This effect seems to be inevitable in runoff modelling because model calibration is a trade-off where usually no perfect parameterization can be found due to different uncertainty sources. In case of a perfect situation with no model, parameter and data uncertainty, all SFCs could be perfectly estimated with the same runoff simulation.

Olden, J. D., Poff, N. L.: Redundancy and the choice of hydrologic indices for characterizing streamflow regimes, River Research and Applications, 19(2), 101-121, doi:

10.1002/rra.700, 2003.

Comment 2: But before it is relevant to discuss other criterions and SFC used for calibration and how well these can be recreated you need to demontrate that teh model is able to reproduce observed flow to a certain degree. I can not see that this is the case here. As long as this is not the case and is demontrated the remaining work becomes irrelevant. If the model is not able to get an Reff higher than 0.7 the interesting discussion is then if it is able to or relevant to use for simulation of other SFC.

Reply 2: Thank you for this useful comment. We agree that it is important to know that HBV can reproduce observed runoff to an acceptable degree. Based on your feedback we realized that we didn't clearly formulate this, and especially the presentation and labels of some of the figures were misleading. We will adapt the text and figures where needed to avoid any possible misunderstanding by the reader.

The HBV model is capable of reproducing the observed runoff for the study catchments reasonably well when calibrated on the Nash-Sutcliffe efficiency (Reff). Figure 5a and 5b show that the median Nash-Sutcliffe efficiency of the 25 catchments is 0.79 in calibration and 0.77 in validation when calibrated with Nash-Sutcliffe (note that values corresponding to Reff are on the y-axis at the right side of the plot). Over all 25 catchments the Nash-Sutcliffe efficiency ranges from 0.68-0.89 in calibration (24 of 25 catchments with efficiencies larger than 0.7) and 0.62-0.86 in validation (21 of 25 catchments with efficiencies larger than 0.7).

This means that it is in principle possible to get good performances in terms of Reff, or, in other words, poorer performances are a result of the respective calibration criteria rather than an inappropriate model (structure).

Comment 3: It might be that you are able to acheive good simulations in this catchment. I that case show this also by showing hydrographs. If you are not able to achieve a Reff up to 0.7 you need to disuss why first and then move to other STC that you can argue for the model is able to recreate in spite of a poor Reff. If the Reff is OK, but

as you initially says is not enough for some studies, you can compare calibration with other criterions againt Reff and each other. I beileve that it is the latter that you try to do, but as long as I as a reader am not able to see that your model and data actually are representative or good enough to reporduce the observed the discussion becomes not relevant or interesting.

Reply 3: Yes, we indeed calibrated the model with different objective functions and then compared the resulting model efficiencies.

As discussed in comment 2 of RC1, HBV was able to reasonably well reproduce the hydrographs of the 25 catchments. We decided to only report the Nash-Sutcliffe efficiencies and not to show the hydrographs. The hydrographs of the complete simulation period from 1984-2009 do not provide much information when shown in one graph (low visibility of information) and the details of one year might not be representative. However, the Nash-Sutcliffe efficiency is an integrated measure of the hydrograph fit over all years.

Comment 4: So before revising this deeper several clification have to be made initially. But the topic is very intersting and relevant so I hope you are able to structure this in a way that make many readers interested and enlightened.

Reply 4: We address your more detailed comments in the parts below (see detailed comments RC 1)

General comments RC 2

Comment 1: You mention in generalm terms that high peaks and low flow is not well enough simulated. Could this and other features the "good Reff model" does not capture be illustrated in the introduction as a background for the study and choice of additional criterions?

Reply 1: This is a good idea. We will extend the introduction by a part about the draw-

back of some commonly used calibration criteria for low and high flow related SFCs. In our previous study model calibrations with widely used objective functions such as Nash-Sutcliffe efficiency, volume error, MARE, etc. resulted in the underestimation of high-flow related SFCs by 13-33

Comment 2: The paper would also benefit of a short argumentation for the choice of SFC as calibration criterions. And also how these are calculated. Ieff does not tell med how you find the value for the criterions and the table does not say so either.

Reply 2: Table 1 in the manuscript indeed only gives a description of the SFCs used in this study. For the calculation of the SFCs the EflowStats R-Package from the USGS (chapter 2.4.3) was used. More details about the exact calculation can be found in the R scripts and the corresponding documentation. The R-package is freely available for everybody and we recommend the interested reader to check the R-package for detailed information on the calculation of the SFCs.

The 13 SFCs used in our study are commonly used SFCs of many ecohydrological studies. We discuss the choice of the SFCs in comment 1 in RC 1.

Detailed comments RC 1: tables and figures

P17 L1: Table 1- But how are these used. For instance how do you evaluate simulated against observed for TA1? This should be explained for all SFC's.

Reply: We use SFCs for model calibration and validation, whereby simulated SFC values are always evaluated against observed SFC value. The exact definition of the objective function as used for calibration can be found in table 2 (ISingle), whereas the performance measure used for validation is given in table 3 (nSFC). The same equations are used for comparison of simulated and observed values for all SFCs.

P17 L3-30: Table 1 - smaller comments on the description of SFCs

Reply: Thanks for these smaller comments on the description of the SFCs. We will

adapt the text in the table according to your input.

P19 L4: Table 3 - It is not clear for me why you need two tables for evaluation/calibration criterions. I assume you use them all in different calibrations....and then also the same in evaluating the performance? From fig 2 I get the idea. But this should be better described in the text. Reff is used both in calibration and evaluation. Describe why.

Reply: We decided to have two tables to make clear which criteria are used for model calibration and which criteria are used for model evaluation. We also used a separate chapter for the description of the calibration and evaluation criteria (chapter2.4.2 and 2.4.3) to make the difference more clear. Nash-Sutcliffe efficiency is the only criteria used in both calibration and evaluation. SFCs are used for model calibration and evaluation, but the exact definition is slightly different. For model evaluation the SFCs had to be normalized to make the results comparable (see chapter 2.4.3). The combined objective functions ISingle_Reff, IMulti, IMulti_Reff are only used in model calibration, whereas MARE is only used in model evaluation. We think it is helpful for the reader to keep those two tables separated, especially because the use of the SFCs is different.

As you mentioned, we use the Nash-Sutcliffe efficiency in model calibration and evaluation. The reason for that is that the Nash-Sutcliffe efficiency is an established measure used in many modelling studies in the same we as we do. It shows how well the model is reproducing the criteria it was calibrated on in an independent time period. It is also a measure of how well the general shape of the hydrograph is simulated, although with a focus on peaks. The use of Nash-Sutcliffe in our model evaluation is useful for the comparison to other studies.

P22 L4: Figure 3 - You refer too this figure as an illustration of variability of model perfomance...and her say it is comaprison between calibration and validation period. This is no consistent. You also say it is about A1 but shows all SFC. I do not get it. Neither that you have stablke and very poor Reff...

Reply: We apologize for the confusion this figure evoked. We admit that the axis labels

are not well selected and that they are misleading. We will change the axis labels to make this plot better readable. For example the axis label “Reff” will be changed to “Normalized TA1 error for models calibrated based on Reff”.

For clarification we shortly describe the figure here: The figure shows the normalized TA1 error when calibrated with Reff, ISingle and ISingle_Reff.

When calibrating the model based on the Nash-Sutcliffe efficiency, we get one normalized TA1 error for each modelling time period. This results in two values of the normalized TA1 error displayed at the Reff -axis. Thus, the value of approximately 0.1 at the Reff -axis is the normalized TA1 error and not a Nash-Sutcliffe efficiency.

Since we have 13 different SFCs that we use for objective functions, we get 13 normalized TA1 error values from calibrations with ISinlge or ISingle_Reff. These are the 13 values displayed at the ISinlge-axis or the ISingle_Reff-axis. Each of the 13 different objective functions are colored based on the SFCs it is based on.

We hope that with the change of the axis labels it also becomes clear that we do not compare calibration and validation. The figure caption says that figure 5a-c are calibration results and figure 5d-f are validation results for both modelling time periods.

Detailed comments RC 1: text

P1 L10: What is the purpose? It does not come clearly forward here. Is it to simulate streamflow characteristic in ungauged basin. In that case is the simulation results relevant for ungauged basins? Is the test sites ungauged, is it split samle testing? Or is it ordinary calibrated models for guaged basin. In cas of latter, how representativ is this test then for ungauged basins?

P1 L14: I assume it’s here it is satde what the actual purpose is. But could the strategy and the purpose be stated clearer. To a person only reading the abstract to consider if this is interesting I do not think this tells enough.

P1 L19: For ungauged basisn estimates or in general?

Reply: The three comments above all address the abstract, so we answer them in one paragraph:

We agree that the abstract should be improved to clearly state the purpose and the method of this study. We will rewrite it. For some clarification right now: In our study we only work with gauged catchments. The aim is to test whether the consideration of SFCs in model calibration improves the estimation accuracy of SFCs compared to more traditional calibration approaches using e.g. the Nash-Sutcliffe efficiency. Our results help to improve model calibration for estimating SFCs which is of great importance for a subsequent regionalization of SFCs for ungauged catchments. The ultimate aim is therefore to have improved model calibration approaches for the regionalization of runoff to ungauged catchments. For gauged catchments we don’t need to model runoff for the estimation of SFCs because they can be calculated form the observed data.

P1 L27: I can not see many other applications for runoff simulations than recreating streamflow characteristics. so in that sense this sentence does not make sense. But if it is ecological SFC then I can see this is referred to as spesific SFC. So it migth be that ecological shoudl be added. Here and other places in the document.

Reply: We will change this sentence.

P3 L9: But initially you stated that this is what is the challange...."Ecologically relevant streamflow characteristics (SFCs) of ungauged catchments are often estimated from simulated runoff of hydrologic models." ... and the rest of the paper is about gauged basin where this challange is substantially lower... this confuses the reader as this is two very different challanges. You can discuss this but make clear already in abstract that this paper is about gauged catchments.

Reply: True, we will make sure in the abstract that our study is about gauged catch-

ments (see comment P1 L10 in detailed comments RC 1 about text).

P3 L16: Other than?

Reply: We will replace "... or more other SFCs?" by "... or multiple SFCs?"

P3 L17: Spesific SFC is included also in traditional calibration (max and mean is spesific SFC). But maybe not those intersting for a spesific purpose. Understand the point, but is not clearly formulated.

Reply: We will replace "...and those where specific SFCs are included?" by "... and those where the SFCs of interest are included?"

P4 L1: Is usually a challenge if water drain out uncontrolled. Is this a problem in this catchment?

Reply: We agree that karst can have a strong influence on runoff modelling. In our study the influence of karst on the catchment scale is relatively small which is reflected on the reasonable well simulated hydrographs.

P5 L8: Was it really necessary with 3 years spin up time to establish state variables in HBV.... it is commonly done over much shorter time. This need some explanation. Usually one prefer to have as long calibration period as possible to catch as many different met variations and combinations as possible.

Reply: For many cases a warm-up period of one year will be sufficient for HBV-light. However, longer warm-up periods improve the conditions of the state variables and don't have a negative effect on the simulations. We used 2 years and 9 months for warming up because it allowed us the most optimal use of the time series with two equally long modelling time periods covering full hydrological years. The runoff time series lasted from January 1982 to December 2009.

P5 L9: This is not clear. Did you use the first period for calibration 84-96 where 84 to 87 was spin up time, and 97-09 for validation and 97-00 as spin up... or

Reply: For the simulation period of October 84 - October 96 the warming up was from January 82 – September 84. For the simulation period of October 97 - October 09 the warming up was from January 95 – September 97.

We will change "A three-year calibration period..." to "A three-year warm-up period...", because this might have been confusing.

P5 L24: consist of one single SFC that incoprorate 13 SFC.. I do not understand what you are saying here.

Reply: We will adapt the sentence to make it more clear. It means that we defined an objective function that consists of one single SFC (ISinlge). Since we have 13 ecologically relevant SFCs in our study catchments, we also have 13 versions of that new objective function (ISinlge).

P6 L1: ...are you not mixing the term objective funtion and SFC here... if not this is very unclearly formulated...

P6 L2: I think I understand what you try to say....but is this formulation good? Rewrite to make it clearer and separate between the function and the SFC's

Reply: The two comments above relate to each other and thus we answer them in one paragraph:

We agree that the terms SFCs and objective function are not properly used. We will adapt the sentence so that it becomes clear that we selected SFCs that resulted in robust and informative estimates when used as objective function. The two selection criteria of robustness and information value should help to define an IMulti that will be relatively robust and informative.

P6 L5: Combine evaluation and calibration chp these are so cloes that they should be in same chp. And the sfc discussion should be with sfc description.

Reply: Model calibration and evaluation seem to be close, but there are some impor-

tant differences in the criteria we use or how we use criteria (see comment P19 L4 in detailed comments RC 1 about tables and figures). Thus, we prefer to keep the two parts separated to emphasize the difference and reduce the potential for confusion.

P6 L6: ? do you try to say that you use the five criterion's in table 2 and Mare. You have already stated how SFC's are included in thes and do not need to say so again.

Reply: We only use the criteria in table 3 for model evaluation (see comment P19 L4 in detailed comments RC 1 about tables and figures). We will change the sentence from ". . .was evaluated by means of SFCs, Reff and mean relative error (MARE)." to ". . .was evaluated by means of normalized SFC error, Reff and mean relative error (MARE)." to emphasize the different use of the SFCs in calibration and evaluation.

P6 L9: Why choosing the median parameter set? And what is median parameter set? Is it not normal to use the optimal parameter set? I do not understand the purpose of the interpretation using a median parameter set? Unless clearly described later this must be explained here.

Reply: The 100 calibrations done with each objective function result in 100 optimized parameter sets. These hundred different parameter sets lead to very similar model performance in calibration, which is a common observation usually referred to as equifinality (Beven and Freer, 2001). The equifinality concept rejects the idea of one single best parameter set. We calculated the median model performance of all 100 parameter sets in calibration and validation to not select the best, but rather a representative value from the efficiency distribution.

Beven, K., and Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, J. Hydrol. 249, 11–29, doi:10.1016/S0022-1694(01)00421-8, 2001.

P6 L10: Should be in the chp where SFC are described.

Reply: Ok, we will move that sentence to chapter 2.2.

P6 L22: Did you not optimize on both these criterions...if not that must be mead clear earlier. If you did...then it should result in 26 parametersets...

Reply: Correct, it's two times 13 resulting in totally 26.

P6 L22: I assume the calibration results in a optimal parameterset given the criterion used. And not in a simulation...

Reply: We will adapt the sentence to state it more precise.

P6 L24: But what about variability? Is it not better to have high variability and some good model simulations than low variability around poor simulations (as I assume a low score indicates)

P6 L24: here you say you uses both Isingle and the combined I signle and I reff ...my previous comment asks about this...

Reply: The two comments above relate to each other and thus we answer them in one paragraph:

The optimal value for the normalized SFC error is 0 (see Table 3). In the text we describe the variability of the error value and its magnitude for the objective functions Reff, ISingle and ISingle_Reff. The best situation would be a low variability at a low error magnitude. Low variability at a high error magnitude would indicate that the related objective function is not suitable for model calibration aiming at the respective SFC. High error variability combined with some low error magnitudes would indicate that certain SFCs used as objective function (ISingle) can lead to good estimates. But it also means that not any SFCs used in ISingle results in good SFC estimates. This fact supports our main conclusion that SFCs should preferably be estimated from targeted runoff model calibration.

P6 L27: In general or for TA1?

Reply: The description refers to the results of TA1 which was selected as a represen-

tative example. But overall a similar pattern can be seen for all SFCs.

P6 L28: illustrated by...

Reply: Will be added.

P7 L1: Again...what is median simulation?

Reply: For each SFC we get 13 normalized errors for the objective functions ISingle and ISingle_Reff (as shown in Figure 3). The median of these 13 normalized SFC errors is displayed in Figure 4 for each SFC.

P7 L1: Reff of <0.1 is very poor and telle me that there are something substatially wrong. Do I misunderstand? Reff should be higher than at least 0.5 before you can say you have representative data and higher than at least 0.7 before you can say that you are able to model a catchment satisfactory.

Reply: The Nash-Sutcliffe efficiency is not displayed in Figure 4. For more discussion on the Nash-Sutcliffe efficiency please see comment 1 in RC 1.

P7 L6: The Reff is very low for all catchments it seems like. only one above 0.5. This tells me the opposite of you comment here! Why is Reff so low. Poor precip data og runoff data or??? Is a acheived criterion below 0.2 good agreement with obeserved? In that case you have to explain how.

Reply: We will add some text in the results part of the manuscript to clarify the confusion regarding the Nash-Sutcliffe efficiency and to avoid any misunderstanding. In figure 5, the y-axis of Reff (right axis) ranges from 0 on top to 1 at the bottom. The closer to 1 (thus to the bottom), the better is the Nash-Sutcliffe efficiency. For all 5 types of objective functions the value of Nash-Sutcliffe is above 0.5.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-546, 2016.