Journal: Hydrology and Earth System Sciences (HESS)
Title: Streamflow characteristics from modelled runoff time series - importance of calibration criteria selection
Manuscript: hess-2016-546
Authors: Sandra Pool, Marc J. P. Vis, Rodney R. Knight, and Jan Seibert

Dear editor,

We thank you for your efforts with our manuscript. The two reviewers provided valuable comments on our manuscript, which helped us to improve the quality of our manuscript. Below, we reply to each of the comments from the reviewers and indicate the changes that have been done accordingly (marked with blue color). We also did a few minor edits that were not suggested by the reviewers. References to pages (P), lines (L), chapters, figures and tables refer the track-changed revised version of our manuscript.

On the behalf of all Co-Authors,
Yours sincerely,
Sandra Pool

Response from the editor

A good discussion on an old subject - which indeed still needs attention. It can be also useful to remind young authors that "blind calibration" is worse than no calibration at all, and that "if you optimise something, first think well what is the objective function".
It can be seen that the authors appreciate the comments, and the (improved) version of the paper is expected.
Season's greetings to all!

Review 1: comments, response and modifications

Dear reviewer,

Thank you for your efforts with our manuscript. We greatly appreciated the comments, which helped us to improve the manuscript. 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 (Table 1 of this response), 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.

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 adapted the axis labels of Fig. 3, 5, and 7 and simplified the axis of Fig. 4. These graphs probably caused most of the misunderstanding. We also added a paragraph to the results part with the information that HBV-light model was capable of reproducing the observed runoff for the study catchments reasonably well when calibrated on the Nash-Sutcliffe efficiency (R_{eff}). Model calibration on R_{eff} resulted in R_{eff} values between 0.68 and 0.89 with a median of 0.79 in calibration. Model performance in calibration was above 0.7 for all except one catchment. The corresponding R_{eff} values in validation ranged from 0.62–0.86, whereby the median model efficiency was 0.77 and efficiencies were larger than 0.7 for 21 of 25 catchments." This means that it is in principle possible to get good performances in terms of R_{eff}, or, in other words, poorer performances are a result of the respective calibration criteria rather than an inappropriate model (structure).
Modification: P25-26 (Figure 3), P28-29 (Figure 4), P29-30 (Figure 5), P31 (Figure 7), P7 L11-13 (HBV calibration efficiency)

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 our reply to 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. We therefore decided to only report Nash-Sutcliffe efficiencies and not to show hydrographs. We also assume that the request for showing hydrographs came most out of the misunderstanding that the model would be very poor.

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 addressed 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 extended the introduction by a part about the drawback 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% and overestimation of low-flow related SFCs by 4-23%. In the current manuscript we analyze the estimation accuracy of the same SFCs when calibrated with the new approach that explicitly takes into account the SFCs of interest.
Modification: P3 L2-8

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 2.2 in the revised manuscript) 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 and we recommend reader who are interested in the exact mathematical formulations to check the R-package for detailed information on the calculation of the SFCs. For most readers, however, we argue that the idea of the indices as given in Table 1 is sufficient, especially since the 13 SFCs used in our study are commonly used SFCs and can be found in many ecohydrological studies. We further 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 used for calibration can be found in Table 2 (I_{Single}), 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: Thank you for these comments on the description of the SFCs. We adapted the text in the table according to your input.
Modification: P19 (Table 1)

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 (chapter 2.4.2 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 I_{Single_Reff}, I_{Multi}, I_{Multi_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 way 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 were not well selected and that they were misleading. We changed the axis labels to make this plot better readable. For example the axis label "R_{eff}" was changed to "Normalized TA1 error for model calibration with R_{eff}". Additionally, we changed the figure to a 3-D graph which on one hand allowed us to show all the relevant information in one single plot and on the other hand reduced the number of subplots from 6 to 2. For clarification we shortly describe the figure here: The figure shows the normalized TA1 error when calibrated with R_{eff}, I_{Single} and I_{Single_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 R_{eff} -axis. Thus, the value of approximately 0.1 at the R_{eff} -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 I_{Sinlge} or I_{Single_Reff}. These are the 13 values displayed at the I_{Sinlge}-axis or the I_{Single_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, the plot type and the figure caption it becomes clearer that we want to illustrate the variability of model performance and that we are not comparing calibration and validation.
Modification: P25-26 (Figure 3)

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 needed to be improved to clearly state the purpose and the method of this study. We therefore rewrote the abstract taking the comments into consideration. In our study we only work with gauged catchments. The aim was 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.
Modification: P1 L11-23

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 deleted this sentence and added the term "ecologically relevant" to the subsequent sentence.
Modification: P2 L3-4

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 tried to make sure in the abstract that our study is about gauged catchments (see comment P1 L10 in detailed comments RC 1 about text).

P3 L16: Other than?
Reply: We replaced "… or more other SFCs?" by "... or multiple SFCs?"
Modification: P4 L5

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 replaced "…and those where specific SFCs are included?" by "… and those where the SFCs of interest are included?"
Modification: P4 L6-7

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 ensure that the conditions of all the state variables are really in equilibrium. Besides requiring data, a longer warming-up period does not have a negative effect on the simulations and for certain parameterizations the typical one-year warming-up actually is too short if one looks in detail at the groundwater storage. We used 2 years and 9 months for warming up because it allowed an 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 added the dates for the warm-up periods and changed "A three-year calibration period…" to "A warm-up period…", because this might have been confusing.
Modification: P5 L28-29

P5 L24: consist of one single SFC that incoprorate 13 SFC.. I do not understand what you are saying here.
Reply: We adapted this part to make it more clear. It means that we defined an objective function that consists of one single SFC (I_{Sinlge}). Since we have 13 ecologically relevant SFCs in our study catchments, we also have 13 versions of that new objective function (I_{Sinlge}).
Modification: P6 L13-15

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 together:
We agree that the terms SFCs and objective function are not properly used. We adapted 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 I_{Multi} that will be relatively robust and informative.
Modification: P6 L23-25

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 important 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 changed the sentence from "…was evaluated by means of SFCs, R_{eff} and mean relative error (MARE)." to "…was evaluated by means of normalized SFC error, R_{eff} and mean absolute relative error (MARE)." to emphasize the different use of the SFCs in calibration and evaluation.
Modification: P6 L29

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: In this sentence we wanted to refer to the median efficiency and not the median parameter set. We adapted the term accordingly. We also added a short argument for using the median parameter set and extended the sentence to make it more clear what the median parameter set is. Here some more detailed information for clarification: 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.
Modification: P7 L1-2

P6 L10: Should be in the chp where SFC are described.
Reply: We moved the sentence to chapter 2.2.
Modification: Sentence moved from P7 L3 to P5 L1-2

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. We adapted the sentence to make that clear.
Modification: P7 L18

P6 L22: I assume the calibration results in a optimal parameterset given the criterion used. And not in a simulation...
Reply: A calibration results in both an optimal parameter set and its corresponding runoff simulation.

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 together:
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 R_{eff}, I_{Single} and I_{Single_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 (I_{Single}) can lead to good estimates. But it also means that not all SFCs used in I_{Single} result 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 representative example. But overall a similar pattern can be seen for all SFCs. We rearranged this paragraph (also because Fig. 4 in the unrevised manuscript was removed) and it should be more clear now.
Modification: Sentence was moved within paragraph from P7 L29 – L31 to P7 L19-21

P6 L28: illustrated by...
Reply: Was added.
Modification: P7 L27

P7 L1: Again...what is median simulation?
Reply: For each SFC we get each 13 normalized errors for the objective functions I_{Single} and I_{Single_Reff} (as shown in Fig. 3). With "median" we referred to the median of these 13 normalized SFC errors from calibrations with I_{Single} and $I_{Single\_Reff.}$ The sentence of P7 L1 was deleted due to restructuring of the paragraph.
Modification: P7 L28-29

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 was not displayed in Fig. 4 (here we refer the Fig.4 in the unrevised manuscript). 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 added some text in the results part of the manuscript and also adapted some figures to clarify the confusion regarding the Nash-Sutcliffe efficiency and to avoid any misunderstanding (see comment 2 of RC1). In the original version of Fig. 4 (corresponds to Fig. 5 in the unrevised

manuscript), the y-axis of R_{eff} (right axis) ranged 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 was above 0.5. Figure 4 was changed and now shows the difference in model performance between a model calibration with R_{eff} and model calibrations with I_{Single}, I_{Single_Reff}, I_{Multi} or I_{Multi_Reff} (see comment 2 in general comments RC1).

Review 2: comments, response and modifications

Dear Björn Guse,

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

Best regards,

Sandra Pool and Co-Authors

Major comments RC 3

Comment 1: In its current state, the article is in my opinion mainly related to hydrology (also by the selection of the journal). At several parts the authors emphasized its ecological importance (e.g. Page 1, Line 10, P.2, L.1-2, P.11, L.14-16). However, I do not really see a connection to ecology. Thus, either the article has to be fully focused on hydrology or it is required to emphasize its ecological relevance

Reply 1: We agree that our modelling approach for estimating SFCs is a typical hydrological approach. However, the motivation for our study is strongly related to ecohydrology because of its focus on the simulation of ecologically relevant SFCs of our study catchments (chapter 2.2). While many of our SFCs are also the focus of other ecohydrological studies, they are untypical for purely hydrological modelling studies where signatures such as segments of the FDC or runoff ratio are of interest. Given the importance of SFCs in ecology we find it important that the hydrological community addresses the issue how to compute these values for ecological studies and applications.

Comment 2: In this article, the Nash-Sutcliffe Efficiency is used as an example for a traditional calibration approach. However, in recent studies the use of several performance measures related to different parts of the hydrological system is recommended. In particular the use of typical performance metrics such as Reff or Kling-Gupta-Efficiency (KGE) in combination with signature measures is recommened (see eg. Van Werkhoven et al., 2009). Thus, I think that a comparison only with the Reff is not sufficient. I recommend to use two or three performance measures such as PBIAS or KGE to show that SFCs also outperforms a calibration approach based on NSE in combination with PBIAS (or other performance measures).

Reply 2: We agree that combined objective functions based on e.g. Nash-Sutcliffe efficiency (R_{eff}) and volume error are widely used for runoff model calibration. We actually used such combined objective functions (that are not based on SFCs) in our previous study (Vis et al., 2015) to estimate the same SFCs of the same catchments as in the current study. The average SFCs error (percent error) in calibration was lowest for calibrations with R_{eff} (error of -4.9 %), followed by calibrations with objective functions based on a) R_{eff}, LogReff and volume error (error of -5.1%), b) R_{eff} and volume error (error of -5.6%) and c) R_{eff}, MARE, spearman rank correlation and volume error (error of -6.1%). Some other combined objective functions were also tested, but resulted in clearly poorer SFC estimates than the ones listed above. We therefore decided to use R_{eff} as a benchmark in the current study. We added some information about the results of the preceding study in the introduction of our manuscript.

Modification: P3 L2-6 and P3 L25-28

Comment 3: Moreover, it needs at least to be discussed how a calibration approach based on these SFCs is related to recent studies using hydrological signatures such as segments of the FDC (see Yilmaz et al., 2008, Pfannerstill et al., 2014).
Reply 3: Thank you for making us aware of the study from Yilmaz et al. (2008) which we mentioned in the introduction in addition to the study of Pfannerstill et al. (2014). We also used the two studies in the first part of the discussion to briefly discuss our results.
Modification: P2 L25-26 and P3 L16-19 (introduction), P10 L33-P11 L1-3 (discussion)

Comment 4: P. 1, L. 24-27: I do not fully agree with this statement. There are different ways of how to calibrate a discharge time series. Certainly there are studies which are focused on certain parts such as on high or low flows. However, other studies aim to represent the whole hydrological system at best without neglecting or emphasizing certain parts. In the way towards a good representation of the hydrological system, the latter one should be the general goal and a strong focus on certain parts of the hydrological system should be a specific case.
Reply 4: We agree that some general agreement often is the calibration goal. Our statements aims at the fact that even with such a general agreement, specific aspects might be poorly simulated as we have shown in the preceding study (see Vis et al., 2015). We adapted the sentence to include your input.
Modification: P2 L1

Comment 5: P.2, L. 21: Here, SFCs are defined "as equivalent to hydrological signatures“. In this case, I do not understand the use of SFC. It is stated before that hydrological signatures are a more common term in hydrology. It is certainly required to justify why you used the term SFC since I do not really see the strong relationship to ecology.
Reply 5: We added some information about the reason for using the term SFC in our study in the introduction. As described in the reply of comment 1, the selection of the SFCs is motivated by their ecological relevance in the study catchments. While both terms, hydrological signature and SFC, describe characteristics of the hydrograph, only the term SFCs makes a distinct connection to ecology. The term SFC has also been used for many years, whereas the term hydrological signature has been introduced more recently.
Modification: P2 L 29-30

Comment 6: I think that the article would benefit from a more detailed interpretation of the results. For example on P. 7. L. 12-15; P.8, L.17-18: Can you explain these results or more specifically the behaviour of these SFCs?
Reply 6: The behavior of SFCs regarding robustness (comment on P8 L17-18) could be explained by grouping them into SFCs that represent catchment characteristics, average flow conditions or dependency on inter-annual weather changes. The striking behavior of the SFCs TL1 and MH10 might be related to the fact that they are calculated on one single value. These two SFCs as well as the pattern in the robustness are discussed in the second and third paragraph of chapter 4.1. We couldn't find a consistently different behavior for the remaining SFCs and therefore it was difficult to find reasonable explanation for their behavior (e.g. why is FL2 estimated similarly well with I_{Sinlge} and R_{eff}, but worse with $I_{\mathrm{Sinlge\_Reff}}$?) (comment on P7 L12-15). We adapted the text of P7 L12-14 to include the results about MH10 and removed the statement about FL2 and MA26. This should guide the reader's attention to the more exceptionally behaving SFCs.
Modification: P8 L14-17

Comment 7: It could be interesting to analyze the relationship/correlation between the SFCs.
Reply 7: We did a Spearman rank correlation test for the 13 SFCs using all 25 catchments (Table 1 of this response). We did the correlation test for both modelling time periods. The correlation values of the two time periods were similar and therefore averaged.

Comment 8: The combination of different metrics might outperform in general single-metric approaches. However, the more metrics are included, the more a trade-off might occur and the equifinality problem arises. In this context, can you give a recommendation for a good number of required SFCs? In the best case, a systematic way of how to select the best SFCs will be provided. Even though when I expect that it is difficult to find a precise number, it is worth discussing this point.
Reply 8: Adding more criteria into a combined objective function usually rather decreases than increases the equifinality issue according to our experience and other studies, although we agree with you that also a trade-off between the different criteria might occur.
We systematically selected SFCs for a multi-objective function based on the criteria of information value and robustness. Our results indicate the importance of jointly evaluating both criteria for the selection of SFCs for a multi-objective function. However, we cannot give a minimum number of SFCs required for such an objective function, because this will depend on the type and combination of SFCs one is interested in. We could show that the four SFCs used for the multi-objective function preserved similar hydrograph characteristics as the Nash-Sutcliffe efficiency (similar estimation accuracy of SFCs not included in the objective function; Fig. 4b and 7c). How much value additional SFCs have for the representation of the hydrograph characteristics would have to be evaluated using many more than 13 SFCs and eventually using synthetic data to also see the effect of redundant information (see discussion chapter 4.3). We made some small changes to chapter 4.3 to take some of the suggestions and questions of comment 8 into account.
Modification: P12 L23-25 and P12 L28-32

Comment 9: The figures 3-8 are very similar (at least visually). I think that the article would benefit from emphasizing the relevance of each figure. It is partly difficult to differentiate them. Maybe you can also thinking about reducing the number of figures to improve the overall message. For example, the figures 6 and 8 have almost the same figure caption. To summarize this point, it is easier to detect the whole message in the case of a more distinct presentation of the results. One example for this is the figure 9 which can be clearly distinguished from the other figures. These results are easier to understand.
Reply 9: We agree that some figures are similar and maybe hard to distinguish. We decided to change two figures and delete one to present the results in a more interesting or intuitive way in the revised manuscript: Figure 3 was changed to a 3-D graph which on one hand allowed us to show all the relevant information in one single plot and on the other hand reduced the number of subplots from 6 to 2. We removed Fig. 4 of the unrevised manuscript, because its main information/ conclusion is similar to the information of Fig. 3 and some of its information is contained in Fig.6b. Figure 4 was adapted so that it shows model performance as the difference between a model calibration with R_{eff} and model calibrations with I_{Single}, $I_{Single\_Reff,}$ I_{Multi} or I_{Multi_Reff}. This change helps to directly compare SFC-based calibration approaches with a traditional Nash-Sutcliffe calibration approach and therefore provides a more clear answer to question 3 in the introduction. We also adapted the axis labels of Fig.s 6, 8 and 10 (corresponds to Fig. 5, 7 and 9 in the revised manuscript).
Modification: P25-26 (Figure 3), P27 (Figure 4 of the unrevised manuscript), P28-29 (Figure 4), P29-30 (Figure 5), P31 (Figure 7), P33 (Figure 9)

Comment 10: P. 8, L.12-13: Could you specify how you can here differentiate between error dependence on time period or objective function?
Reply 10: We adapted the sentence. Generally it means that we looked for systematic patterns in the error magnitude (absolute value of normalized SFC error) of the SFCs. For some SFCs (e.g. TL1) the error magnitude could be considerably higher in one modelling time period than in the other one. We described the error of such SFCs as being time depended. Some other SFCs had clearly higher error magnitudes when calibrated on a certain objective function (e.g. MH10). The estimation accuracy of such SFCs was considered as being dependent on the objective function.
Modification: P9 L19-21

Comment 11: P.8, L.14-15: I do not understand this statement that the SFCs are neither related to flow components nor to flow conditions. Hydrological signatures (as an equivalent term) are known to be of special importance to explain the hydrological behaviour. Thus, what can we learn from using these SFCs in terms of the hydrological behavior in the catchment. And how is this related to the general idea of the hydrological signatures?
Reply 11: We are sorry for the confusion this statement evoked and made the sentence clearer. We meant that we could not relate the estimation accuracy (error magnitude, spread of error magnitude and dependency of error magnitude on modelling time period/ objective function (see comment 10)) of the analyzed SFCs to the flow components or flow conditions they belong to. E.g. we cannot say that all high-flow related SFCs had very low estimation accuracies.
Modification: P9 L21-22

Comment 12: P.8, L.26 to P. 9, L.21: I agree with this part which is clearly understandable, but certainly also not surprising. It is mostly existing knowledge of hydrological modelling. What can we learn here except of using several and different metrics. I recommend to shorten this part and emphasize the most important points from this study. In contrast, I really like to following passage (P. 9, L.21-31).
Please also discuss the impact of a SFC-based calibration for the process representation. Can you state that the hydrological system is overall better represented by using several SFCs?
Please discuss the benefit of optimizing one specific SFC. This leads to a modelled hydrograph which is able to represent a very specific condition but probably not the overall hydrograph. This implies that the part of the hydrological system which is not in the focus of this SFC is probably not adequately considered. This might be of particular relevance when using very specific SFCs such as MA26.
Reply 12: Thank you for this helpful comment. We adapted the discussion part you mentioned focusing on your proposed aspect of how the calibration with SFCs affects process representations: The benefit of optimizing one specific SFCs lies in the relatively accurate estimation of the respective SFC compared to a calibration with R_{eff} or a multi-SFC objective function. Model calibration on one single SFC clearly emphasizes the hydrograph aspects of the selected SFC which can negatively affect other hydrograph aspects. This implies that calibrations with I_{Single} can lead to poor model performance for SFCs not included in the objective function (Fig. 3). E.g. calibrating on the frequency of moderate floods (FH6) leads to poor model efficiencies for base flow (ML20) (Figure 6b) which indicates that the representation of the main runoff processes can suffer by SFC-specific model calibration. From the fact that a calibration with R_{eff} and a calibration with multiple SFCs lead to comparable SFC estimates we infer that the main hydrological processes of the catchments are similarly well represented with the two approaches. We assume that these two calibration criteria result in a better process representation than the calibration with a single SFC, because they outperform the calibration with I_{Single} for those SFCs not included in I_{Single}.
Modification: P10 L6-33

Minor comments R 3

P.1, L. 12: maybe "optimization“ instead of "minimization or maximization“.
Reply: We did the replacement as suggested.
Modification: P1 L13

P.1., L.16: Are these over- and underestimations a general aspect of these SFCs or are they case-specific?
Reply: These results represent the general tendency of the model and the results from a specific objective function and/or modelling time period can deviate from these general conclusions (see Fig. 8 in the revised manuscript).

P.2, L.16-19 and L. 30: I recommend to include the study from Yilmaz et al. (2008).
Reply: Thank you for this suggestion. We included the study of Yilmaz et al. (2008) in the introduction.
Modification: P2 L25-26 and P3 L16-20

P.2, L. 34: The meaning of "esoteric and subtle aspects of the flow regime“ is unclear.
Reply: Good point, we adapted the sentence.
Modification: P3 L15

P.3, L. 19: Please think about renaming the section to "Methods and materials“, since a catchment is not a method.
Reply: We changed the title from “Methods” to “Materials and methods”.
Modification: P4 L8

P.5, L.31: Why you have used 0.2 and 0.25 as weights?
Reply: I_{Multi} consists of four SFCs that we wanted to weight equally, which results in a weight of 0.25 for each of the SFCs. These four SFCs were combined with the Nash-Sutcliffe efficiency (I_{Multi_Reff}) and each of the components was again assigned the same weight (0.2). We weighted all components of I_{Multi} and I_{Multi_Reff} equally because there was no clear reason to weight one/some of the SFCs differently from the others.

P.6, L. 9: Could you specify "median parameter set“?
Reply: This sentence should say “…, the median model efficiency of each catchment was selected.” We changed this sentence.
Modification: P7 L1

P. 11, L.15: Could you specify "later application of simulated SFCs related to flow alteration – ecosystem change relationships“. This aspect was up to now neither in the focus of the article nor emphasized as an overall aim.
Reply: This statement refers to the first paragraph of the introduction where we motivate the need for accurate SFC estimates by giving the example of flow alteration – ecosystem change relationships used for sustainable flow management. We adapted the sentences related to the mentioned statement to make this connection more clear.
Modification: P13 L13-16

Table 1: TA1: runoff with two f
Reply: We added the missing f.
Modification: P19

Table 1: Why you have named the SFCs FH6 and FH7 and not FH3 and FH7.
Reply: We agree that the abbreviation can be confusing, but decided to follow the abbreviations used in previous publications and in the EflowStats R-package that was used for the calculation of the SFCs. The same abbreviations are commonly used in many studies (see e.g. Olden and Poff, 2003).

Fig. 5: Can you explain the outlier TL1?
Reply: We discussed the low estimation accuracy of TL1 in the discussion (last paragraph of chapter 4.1).

References used in the responses to the reviewer

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.

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.

Vis, M., Knight, R., Pool, S., Wolfe, W., and Seibert, J.: Model calibration criteria for estimating ecological flow characteristics, Water, 7, 2358–2381, doi:10.3390/w7052358, 2015.

Table 1: Spearman rank correlation coefficient between the 13 ecologically relevant streamflow characteristics used in this study.

	Mean-flow				Low-flow				High-flow				
	ma41	ma26	ra7	ta1	e85	fl2	tl1	ml20	mh10	dh13	dh16	fh6	fh7
ma41	-												
ma26	-0.529	-											
ra7	-0.449	0.843	-										
ta1	-0.844	0.653	0.632	-									
e85	0.803	-0.811	-0.801	-0.852	-								
fl2	0.213	-0.506	-0.697	-0.395	0.500	-							
tl1	-0.393	0.130	0.120	0.377	-0.165	-0.097	-						
ml20	0.546	-0.894	-0.962	-0.736	0.869	0.688	-0.148	-					
mh10	0.840	-0.425	-0.402	-0.725	0.689	0.198	-0.475	0.463	-				
dh13	-0.650	0.864	0.867	0.804	-0.918	-0.600	0.120	-0.945	-0.540	-			
dh16	0.513	-0.646	-0.722	-0.642	0.678	0.600	-0.319	0.750	0.545	-0.650	-		
fh6	-0.484	0.865	0.902	0.632	-0.796	-0.662	0.103	-0.934	-0.364	0.876	-0.677	-	
fh7	-0.614	0.884	0.895	0.762	-0.901	-0.609	0.112	-0.947	-0.503	0.966	-0.666	0.917	-

Streamflow characteristics from modelled runoff time series - importance of calibration criteria selection

Sandra Pool[1], Marc J. P. Vis[1], Rodney R. Knight[2], and Jan Seibert[1,3,4]

[1]Department of Geography, University of Zurich, Zurich, Switzerland
[2]U.S. Geological Survey Lower Mississippi—Gulf Water Science Center, 640 Grassmere Park, Suite 100, Nashville, TN 37211, USA
[3]Department of Earth Sciences, Uppsala University, Uppsala, Sweden
[4]Department of Physical Geography, Stockholm University, Stockholm, Sweden

Correspondence to: Sandra Pool (sandra.pool@geo.uzh.ch)

Abstract. Ecologically relevant streamflow characteristics (SFCs) of ungauged catchments are often estimated from simulated runoff of hydrologic models that were originally calibrated on gauged catchments. ~~Estimated SFCs~~ However, SFC estimates of the gauged donor catchments and subsequently the ungauged catchments can be substantially uncertain when models are calibrated using traditional approaches based on ~~minimization or maximization~~ optimization of statistical performance metrics (e.g. Nash~~–~~-Sutcliffe efficiency). An improved calibration strategy for gauged catchments is therefore crucial to help reducing the uncertainties of estimated SFCs for ungauged catchments. The aim of this study was to improve SFC estimates from modelled runoff time series in gauged catchments by explicitly including one or several SFCs in the calibration process. Different types of objective functions were defined consisting of the Nash-Sutcliffe efficiency, single SFCs or combinations thereof. ~~To evaluate model performance, we tested how well SFCs are simulated when the model objective function was calibrated using one or more SFCs.~~ We calibrated a bucket-type runoff model (HBV model) for 25 catchments in the Tennessee River basin and evaluated the proposed calibration approach on 13 ~~selected~~ ecologically relevant SFCs representing major flow regime components and different flow conditions. While the model generally tend~~s~~ed to underestimate SFCs related to mean and high-flow conditions, SFCs related to low flow ~~are~~ were generally overestimated. The highest estimation accuracies were achieved by a SFC-specific model calibration. Estimates of SFCs not included in the calibration process were of similar quality when comparing a multi-SFC calibration approach to a traditional Nash~~–~~-Sutcliffe efficiency calibration. For practical applications, this implies that SFCs should preferably be estimated from targeted runoff model calibration and modelled estimates need to be carefully interpreted.

1 Introduction

Reliable runoff information is fundamental for many water resources-related tasks such as flood prevention, drought mitigation, management of drinking water supply and hydropower, or river restoration. Runoff modelling is a tool that can be used to create runoff time series when observed time series are not available. Runoff model simulations usually focus on

either representing the general shape of the hydrograph or on accurately simulating specific ~~runoff~~ streamflow characteristics relevant to a respective application. The extraction of ~~runoff~~ streamflow characteristics (SFCs) from a simulated time series may produce poor estimates when these characteristics were not included in model calibration. ~~A typical example is the use of runoff simulations for the estimation of streamflow characteristics (SFCs).~~ Ecologically relevant SFCs are properties of the annual streamflow hydrograph defining the structure and functioning of aquatic and riparian biodiversity (Richter et al., 1996; Poff et al., 1997). The accurate prediction of streamflow characteristics is a core determinate to defining how streamflow and aquatic communities relate. A large number of SFCs have been suggested to characterize ecologically relevant aspects of the flow regime (Tharme, 2003) and have become the basis for decision-support systems integrating resource management with ecological response.

Multivariate regression or runoff models are used to estimate SFCs when observed streamflow time series data are not available (Hailegeorgis and Alfredsen, 2016). The estimation of SFCs with linear regression usually relates a single SFC to catchment characteristics such as climate, land cover, geographic, and geologic variables (e.g. Sanborn and Bledsoe, 2006; Carlisle et al., 2010; Knight et al., 2012). This approach is inflexible in a sense that the regression is SFC-specific and does not allow for analysis of potential water-use and land management (Murphy et al., 2013). These disadvantages can be partially overcome by applying runoff models. Simulated streamflow time series from runoff models can be used to calculate any SFC and by changing model input and parameters different scenarios such as climate change, groundwater withdrawals, land use and riverine change can be simulated (Poff et al., 2010; Murphy et al., 2013; Olsen et al., 2013; Shrestha et al., 2014). While runoff models provide flexibility in evaluating scenarios, statistical models such as multiple linear regressions often provide greater accuracy (Murphy et al., 2013).

Runoff models are used in both ecohydrology and hydrological modelling as tools to simulate specific aspects of the runoff regime. The terms, SFCs or ecological flow indices, are often used to refer to such specific aspects of the flow regime in ecohydrology studies, whereas the more recently introduced term, hydrological signatures, has been used in hydrological modelling (Jothityangkoon et al., 2001; Wagener et al, 2007). Hydrological signatures can often support a physical interpretation of the way a catchment functions and are seen as valuable metrics especially for modelling ungauged catchments (Jothityangkoon et al., 2001), for selecting appropriate model structures (Euser et al., 2013) or guide model parameter selection in a meaningful way (Yilmaz et al., 2008), and for classifying catchments (Wagener et al., 2007; Sawicz et al., 2011). Regardless of the terminology and the ultimate goal, the basic goal is the quantification of certain aspects of a streamflow time series ~~to answer various questions such as the response of aquatic health to changes in a flow regime~~. In this paper, we use the term SFC as equivalent to hydrological signature, but generally prefer the term SFC to emphasize their ecological relevance.

Estimated streamflow characteristics are prone to significant errors when calculated from simulated time series (Murphy et al., 2013; Shrestha et al., 2014; Vis et al., 2015). This is due in part to the objective functions used for evaluating the model error such as the commonly used Nash–Sutcliffe efficiency (Nash and Sutcliffe, 1970) or volume error, which do not ensure that a model is reproducing particular streamflow characteristics. These objective functions subsequently guide model

parameter calibration, which strongly influences the simulated hydrograph (for an overview see Pfannerstill et al., 2014) in terms of annual, seasonal, and monthly volumes and magnitudes. For example, Vis et al. (2015) compared model simulation from calibrations with Nash-Sutcliffe efficiency only with calibrations based on the combination of multiple objectives such as Nash-Sutcliffe efficiency, Nash-Sutcliffe efficiency of logarithmic flow, volume error and Spearman rank correlation. All these calibration approaches tended to overestimate low-flows and underestimate medium and high-flow related SFCs. Estimation accuracy varied greatly between SFCs with absolute biases between 3% and 33%. Large differences in estimation accuracy are also reported by Shrestha et al. (2014) and Ryo et al. (2015). Their multi-objective calibration approach resulted in runoff simulations favouring high-flows at the expense of the estimation accuracy of low-flows. The large variability in estimated SFC accuracy as well as the bias in the estimates can generally be observed independent of the model used to simulate the runoff time series (Caldwell et al., 2015). A remedy to this large variability and bias is to incorporate SFCs into model calibration schemes. For example, Westerberg et al. (2011) and Pfannerstill et al. (2014) focused on specific evaluation points or segments of the flow-duration curve (FDC) during model calibration. Both studies report better overall performance for the simulated hydrograph with a FDC-based calibration compared to a more traditional calibration approach using, for example, the Nash–Sutcliffe efficiency (Nash and Sutcliffe, 1970). However, runoff models calibrated using FDC have to be constrained by additional SFCs if one is interested in ~~more esoteric and subtle aspects of the flow regime such as the~~the exact timing of events or when snow-related runoff processes are of importance (Westerberg et al., 2011). Yilmaz et al. (2008) combined information on different segments of the FDC with the runoff ratio and the rainfall-runoff lag time to guide model parameter selection in terms of primary catchment functions. These hydrologically meaningful signatures generally improved hydrograph simulation, but their value was limited for the process of vertical redistribution of excess rainfall in the catchment. Instead of aiming at a well-simulated, general hydrograph, Hingray et al. (2010) and Olsen et al. (2013) focused on certain aspects of the streamflow regime that were considered most important. Their results, which are echoed by Murphy et al. (2013), suggest that the runoff model performs reasonably well for the aspects on which it is calibrated, whereas it only modestly represents other runoff characteristics. Hence, developing an approach to increase the accuracy of estimated SFCs from runoff model time series continues to be an open challenge in hydrological modelling.

This study extends on the study of Vis et al. (2015) where various combinations of traditionally used objective functions were evaluated with respect to a suite of ecologically relevant SFCs. Their model calibrations with Nash-Sutcliffe efficiency outperformed multi-objective model calibrations and it was hypothesized that the explicit consideration of SFCs in runoff model calibration could reduce bias in estimated SFCs. The main objective of this study was therefore to assess the potential for a runoff model calibrated using specific aspects of the flow regime to more accurately estimate a suite of SFCs as compared to using ~~more traditional~~ a Nash-Sutcliffe efficiency based calibration approach~~es~~. The general approach was based on the idea that most information essential for estimating SFCs is preserved in the simulated hydrograph by including selected SFCs in model calibration. Our modelling approach relies on catchments with observed runoff time series and therefore does not answer the question of how to simulate SFCs in ungauged or altered catchments. However, the prediction of runoff for ungauged catchments benefits from an improved and informed calibration strategy for gauged catchments,

which is used in the subsequent regionalisation. For regionalization approaches we refer to studies such as Yadav et al. (2007), Viglione et al. (2013) or Westerberg et al. (2016).

The following questions are addressed in this paper:

(1) How well is a single SFC simulated when that SFC is used in the model objective function?

(2) How well is a single SFC simulated when the model objective function contains one or ~~more other~~multiple SFCs?

(3) How does the accuracy of estimated SFCs vary between traditional calibration approaches and those where ~~specific~~ the SFCs of interest are included?

2 Materials and ~~M~~methods

2.1 Catchment locations and characteristics

The study catchments are all located in the 106000 km^2 Tennessee River basin in the southeastern United States (Fig. 1), which is one of the most diverse temperate freshwater ecosystems in the world (Abell et al., 2000). A large number of endemic fish species and a unique assemblage of mussels, crayfish and salamanders make the Tennessee River basin an excellent area for ecohydrological studies (Abell et al., 2000). From a study published by Knight et al. (2008), 25 catchments in the Tennessee River basin having observed streamflow time series (U.S. Geological Survey, 2016b), precipitation (U.S. Department of Commerce, 2007a), temperature (U.S. Department of Commerce, 2007b) and potential evaporation data (Rotstayn et al., 2006) were selected. The sizes of catchment areas range between 100 and 4800 km^2 with elevations ranging from 174 to 937 m (U.S. Geological Survey, 2016a) above the North American Vertical Datum of 1988 (NAVD 88). Land cover for the study catchments is predominantly hardwood forest and pasture. Air temperature and precipitation varies between catchments according to both catchment elevation and longitude. Mean annual air temperature in the 25 catchments varies between 9.3 and 14.7° C, and annual precipitation varies from 1500 to 2020 mm with autumn being slightly drier and less than 8% of annual precipitation falling as snow. Runoff is highest in winter and lowest in summer, ranging from 400 to 1300 mm a^{-1} (millimeters per year). Variability in soil thickness (Omernik, 1987), regolith thickness, karst development and topographic slope (Hoos, 1990; Wolfe et al., 1997; Law et al., 2009) are documented as asserting the most influence on runoff.

2.2 Selection of SFCs

Thirteen SFCs assessed in this study were chosen for use in model scenarios based on discernible functional connections with fish community diversity (Knight et al., 2008; Knight et al., 2014). This set of 13 SFCs represents each of the major flow regime components commonly used in ecological studies (e.g. Olden and Poff, 2003; Arthington et al., 2006; Caldwell et al., 2015): magnitude, ratio, frequency, variability and date (Table 1). For this study the SFCs were additionally grouped according to flow conditions (mean, low and high flow), because different aspects of the hydrograph have been shown to be

sensitive to the objective function used for model calibration (for an overview see Pfannerstill et al., 2014). The SFCs were calculated using the U.S. Geological Survey (2014) EflowStats R-package.

2.3 The runoff model

The HBV (Hydrologiska Byråns Vattenavdelning) model (Bergström, 1976; Lindström et al., 1997) is a bucket-type hydrologic model for simulating continuous runoff series. Model inputs are daily rainfall and air temperature, as well as daily potential evaporation values. Hydrologic processes are represented by four different routines corresponding to snow, soil water, groundwater, and runoff routing, with a combined total of 16 parameters. In the snow routine, snow accumulation and snowmelt are calculated by a degree-day method. Snowmelt together with rainfall and potential evaporation are input to the soil-water routine, where the actual evaporation and the groundwater recharge are computed based on the soil-moisture storage. The groundwater (or response) routine consists of a connected shallow and deep groundwater reservoir and simulates peak flow, intermediate runoff and baseflow. These three runoff components are taken together and transformed by a triangular weighting function during the routing process to calculate the runoff at the catchment outlet. Runoff can be modelled in a semi-distributed way by separating a catchment into elevation bands. Thereby, the snow and soil-water routines are calculated for each elevation band, whereas the groundwater storage and the runoff routing routines are treated as a lumped representation of the entire catchment. HBV exists in different versions, whereby the general structure of the model remains the same. The version applied in this study is HBV-light (Seibert and Vis 2012). Like for all bucket-type models, parameters in the HBV model cannot be determined *a priori*, they are identified by model calibration instead. More detailed information on the HBV model can be found in Bergström (1976), Lindström et al. (1997) and Seibert and Vis (2012).

2.4 Modelling approach

2.4.1 Model setup

For each of the 25 catchments the number of elevation bands was defined by splitting the catchment into elevation zones of 200 m. Elevation zones covering less than 5% of the catchment area were merged with the adjacent elevation zone. For the resulting elevation bands, air temperature and rainfall were computed with a lapse rate of 6° C per 100 m and 10% per 100 m, respectively. Potential evaporation was assumed to be uniform over the whole catchment.

Model simulations were run for two time periods, one lasting from the hydrological years (1st of October until 30th of September) 1984 to 1996 and the other lasting from 1997 to 2009. The approximately three years preceding each simulation period (January 1982 to September 1984 and January 1995 to September 1997 respectively) served to establish state variables of the model. A ~~calibration~~ warm-up period was needed to ensure that the different state variables at the beginning of the simulation period were consistent with the preceding meteorological conditions and parameter values. The two simulation periods were used for model calibration and validation. For calibration, a genetic algorithm (Seibert, 2000) was

used and the range of possible parameter values was specified based on previous studies (Lindström et al., 1997; Seibert, 1999; Table 2 in Vis et al., 2015). The 100 independent calibration trials allowed to account for parameter uncertainty or equifinality (Beven and Freer, 2001) and resulted in a set of 100 calibrated parameter sets for each objective function (Fig. 2).

2.4.2 Choice of objective functions for model calibration

The complete model calibration process was conducted for 25 catchments and using data from all five different types of objective functions (see Table 2 for the exact equations) that focused on different aspects of the hydrograph. In the first step, model parameters were constrained maximizing the Nash–Sutcliffe efficiency criterion (R_{eff}, Nash and Sutcliffe, 1970). The Nash–Sutcliffe efficiency is the most widely used objective function in hydrological modelling, and it served as a benchmark for the objective functions that included SFCs. Model calibration with R_{eff} tends to reduce simulation errors in magnitude and timing of high-flow conditions at the expense of errors in low-flow conditions (Legates and McCabe, 1999; Krause et al., 2005).

Next, a new efficiency measure that consisted of one single SFC (I_{Single}) was defined to explicitly incorporate ~~each of the 13~~ individual SFCs in model calibration (Table 2). Each of the 13 selected SFCs was used separately for model calibration resulting in 13 versions of I_{Single}. Additionally~~,~~ each SFC efficiency measure was combined with R_{eff}, whereby both metrics were equally weighted (I_{Single_Reff}). The use of a single SFC as the objective function allowed calibration to focus on a specific aspect of the hydrograph, while adding R_{eff} helped to improve the overall shape of the hydrograph including the magnitude and timing of events.

Based on the results from the individual SFCs, an objective function consisting of four different and equally weighted SFCs was defined (I_{Multi}, Table 2). This SFC based efficiency measure was again combined with R_{eff} (I_{Multi_Reff}). For the resulting combined objective function, weights of 0.2 were assigned to each metric to make sure the individual SFCs had sufficient influence on the model calibration and were not dominated by R_{eff}. The number of SFCs constituting I_{Multi} was not previously fixed. Instead, a minimum number of SFCs was ~~defined~~ selected so that the resulting objective function was both robust and informative. These two requirements for the objective function could be achieved by only including SFCs that are robust and informative. A SFC was considered as robust when the SFC calculated from a model simulation with I_{Single} had small errors over the full range of catchments in both validation time periods. A SFC was regarded as being informative, when it also yielded relatively good simulations for other SFCs.

2.4.3 Evaluation of model performance

Model performance in calibration and validation was evaluated by means of normalized SFC~~s~~ error, R_{eff} and mean absolute relative error (MARE) (see Table 3 for the exact equations). These evaluation criteria were calculated for all 100 runoff simulations based on the five different types of objective functions in both validation time periods and for all 25 catchments.

For the interpretation of the results, the median ~~parameter set~~model efficiency of each objective function, validation period ~~of each~~and catchment was selected as representative value for the model efficiency distribution.

~~The SFCs were calculated using the U.S. Geological Survey (2014) EflowStats R package.~~ As there are significant differences in the SFC ranges, a normalization was needed that allowed comparison of the different SFCs. Instead of normalizing in terms of relative error, an approach was applied that normalizes the SFC estimation error. The normalization of a SFC was computed as the absolute simulation error divided by the range of possible values for that SFC in the respective catchment (Table 3). To calculate these SFC ranges, 10000 Monte Carlo simulations were run for each respective catchment using randomly chosen parameter values from the previously identified parameter space (Lindström et al., 1997; Seibert, 1999; Table 2 in Vis et al., 2015). The Monte Carlo simulations represented the potential variation in a certain SFC if no information was available to constrain the runoff model. The range was then calculated as the difference between the 10^{th} and 90^{th} percentiles of the simulated SFC values.

3 Results

The HBV model was capable of reproducing the observed runoff for the study catchments reasonably well. Model calibration on R_{eff} resulted in R_{eff} values between 0.68 and 0.89 with a median of 0.79. The corresponding R_{eff} values in validation ranged from 0.62 to 0.86 with a median of 0.77.

3.1 The use of single SFCs as objective functions in model calibration

3.1.1 How informative is a SFC for estimating any SFC?

The calibrations for all 13 versions of I_{Single} and I_{Single_Reff} resulted in total in 26 ~~in 13~~ different runoff simulations that were evaluated by calculating the normalized SFC~~s~~ error for the calibration and validation periods. The use of SFCs as a single objective function (I_{Single}) generally resulted in poor SFC estimations for those SFCs not included in I_{Single} in both model calibration and validation. The SFC estimates became substantially better when I_{Single} was combined with R_{eff}. The SFC TA1 (stability of runoff) was selected as a representative example to illustrate that model calibration with I_{Single} resulted in greater variability in model performance than the calibrations with either I_{Single_Reff} or R_{eff}, independent of the considered time period (Fig. 3, where the spread along the I_{Single}-axis is larger than the spread along the I_{Single_Reff} or R_{eff}-axis). While estimation accuracies with I_{Single_Reff} and R_{eff} are often of comparable magnitude, they both outperform most simulations with I_{Single}. Error magnitudes from the three described objective function types (I_{Single}, I_{Single_Reff} and R_{eff}) can vary considerably between time periods (illustrated by triangles and circles respectively in Fig. 3).

~~The median simulation error of all 13 versions for each objective function (I_{Single} and I_{Single_Reff}) with each SFC is presented in Fig. 4. The use of SFC as a single objective function (I_{Single}) generally resulted in poor SFC estimations for those SFCs not included in I_{Single} in both the model calibration and validation. The SFC estimates became substantially better, with narrow spread and lower median of the absolute normalized SFC error, when I_{Single} was combined with R_{eff}.~~

3.1.2 Estimation accuracy using SFC-specific model calibrations

Model calibration results for the 13 SFCs confirmed that HBV-light is capable of estimating different SFCs with a high level of precision if the respective SFC is used as an objective function (I_{Single}) for model calibration (~~Fig. 5a~~almost 100% estimation accuracy for all SFCs with I_{Single}). ~~Using the combined objective function I_{Single_Reff} gave similar, although slightly less precise results, whereas calibrations using R_{eff} as the objective function resulted in the least accurate estimates.~~ Both I_{Single} and the combined objective function I_{Single_Reff} clearly outperformed model calibrations based on R_{eff} with regard to the estimation of SFCs. However, calibration with I_{Single} yielded poor model performances ~~in relation to~~ when evaluated in terms of R_{eff} ~~if~~ whereas R_{eff} efficiencies of calibrations with I_{Single_Reff} and R_{eff} were comparable. ~~was not combined with the objective function I_{Single}.~~

Validation results exhibited a similar pattern in model performance (Fig. ~~5b~~4b). The median absolute normalized error of the 13 SFCs was relatively low for model runs based on the objective functions I_{Single} and I_{Single_Reff} ~~and was higher for simulations based on the~~compared to model calibration with R_{eff}. ~~The inclusion of R_{eff} into the objective function had a negative effect on the model performance, especially for FL2 and MA26 (Fig. 5a-c). Except for MH10, which was best estimated with the objective function R_{eff}, SFCs can be regarded as valuable for model calibration.~~ The comparable SFC estimation accuracy of I_{Single} and I_{Single_Reff} that often outperformed model simulations with R_{eff} confirms the value of SFCs for model calibration aiming at a respective SFC. An exceptional behaviour can be observed for MH10, where the estimation accuracy was negatively affected by a calibration based on the SFC itself (Fig. 5a-c).

3.2 The use of multiple SFCs for model calibration

Figure ~~7a~~ 6a shows simulation results for the objective function I_{Single} for all 25 catchments and both modelling time periods. The five SFCs with the highest robustness (less variability in error; Fig. ~~7a~~6a) were RA7, ML20, FH6, E85 and MA41. The information value of these five SFCs varied, but all together each of the 13 SFCs were well simulated by at least one of these five (Fig. ~~7b~~6b). However, since the information value of ML20 (base flow) and E85 (lowest 15% of daily runoff) was redundant, E85 was discarded as a potential SFC for the objective function I_{Multi}.

Median estimates of the 13 SFCs in the calibration period were slightly lower when the model was calibrated with I_{Multi} rather than I_{Multi_Reff}. Both of these objective functions led to much better model performance for SFCs than calibrating with R_{eff} alone. While MARE were generally smaller (i.e. better) in calibrations with I_{Multi} and I_{Multi_Reff} than in calibrations with R_{eff}, ~~T~~the inverse pattern was observed when evaluating model performance in terms of R_{eff} ~~and MARE~~ (Fig. ~~5a~~4a).

Model performance for the validation period with I_{Multi_Reff} had lower median error for SFCs than the error associated with using I_{Multi} as objective function (Fig. ~~5b~~4b). The comparison of I_{Multi} and I_{Multi_Reff} for all SFCs separately revealed that for most SFCs both objective functions resulted in similar estimates (Fig. ~~8a~~7a). While the two objective functions had a comparable performance in terms of SFC and MARE, ~~the result diverged when evaluating their efficiency for R_{eff} and MARE. The two criteria, R_{eff} and MARE, were~~was better simulated with R_{eff} being part of the objective function (Fig. ~~5b~~4b).

As could be expected, there was a ~~pronounced~~ difference in median estimates of SFCs between model simulations with the objective functions I_{Multi_Reff} and I_{Single_Reff}. I_{Single_Reff} was ~~clearly~~ better for estimating SFCs, especially for SFCs not included in the I_{Multi_Reff} objective function (Fig. ~~8b~~7b). Comparing simulations from I_{Multi_Reff} and R_{eff} revealed a smaller median error of the SFCs and MARE ~~(Fig. 8c)~~ but poorer efficiencies for R_{eff} ~~and MARE~~ when calibrating with I_{Multi_Reff} (Fig. ~~5b~~4b and 7c). Yet, for most SFCs not explicitly incorporated into the objective function I_{Multi_Reff}, R_{eff} performed equally well or slightly better than I_{Multi_Reff} (Fig. ~~5b~~4b).

3.3 Estimation accuracy for SFCs

Figure ~~9~~ 8 provides an overview of how well SFCs were simulated by presenting the results for both modelling time periods and all five objective function types. Performance values were categorized as small (< 10%), medium (11–20%), large (21–30%) and very large (>30%) errors. The median error (illustrated by stars in Fig. 8) was used for the evaluation of the under- or overestimation. An underestimation of SFC values was observed for all five SFCs representing high-flow conditions as well as for three of four mean-flow related SFCs. With one exception, low-flow SFCs were overestimated. The magnitude of the absolute error varied from generally small for RA7, ML20, MH10 and FH6, to medium for MA41, TA1 and DH16, and up to very large magnitude for TL1. A considerable range, from small to large errors, was observed in the individual objective functions for FL2, MA26, E85, MH10, DH13, FH7, and TL1. ~~Except for four SFCs, the magnitude of the simulation error depended either on the time period (MA26, E85, TL1, DH13, DH16) or the objective function (RA7, MH10, FH6, FH7) considered. These groups of SFCs regarding magnitude, spread and dependence of the error did not seem to be related to the flow components (magnitude, ratio, frequency, variability and date) or flow conditions (low, medium and high flow).~~For some SFCs (MA26, E85, TL1, DH13 and DH16) the error tended to be higher in one of the two modelling time periods whereas for other SFCs (RA7, MH10, FH6 and FH7) the objective function had a distinct influence on the error magnitude. There was no evidence that the estimation accuracy depends on flow components (magnitude, ratio, frequency, variability and date) or flow conditions (low, medium and high flow).

Normalized errors for the high-flow conditions, DH16 and MH10, for all 25 catchments and for both modelling time periods indicate two typically observed phenomena regarding uncertainty due to differences in catchments. DH16 is an example of a SFC that could be regarded as being clearly underestimated by the model, because of its negative bias in nine out of ten cases (Fig. ~~10a~~9a). However, for objective functions or modelling time periods with a low magnitude in the median bias, there might be a substantial number of catchments that show overestimation of DH16. A second commonly observed phenomenon is shown by the SFC MH10 (Fig. ~~10b~~9b). While MH10 had mostly small median errors, there were many catchments with considerably higher errors. Although MH10 was the most extreme example, it illustrates that small median errors do not guarantee good results for all catchments.

4 Discussion

4.1 On the importance of the choice of the objective function

The results demonstrated that the objective function used for model calibration strongly influences the estimation accuracy of SFCs. This finding confirms the findings of previous studies (e.g. Hingray et al., 2010; Westerberg et al., 2011; Murphy et al., 2013; Olsen et al., 2013; Pfannerstill et al., 2014; Shrestha et al., 2014; Caldwell et al., 2015; Vis et al., 2015) and points out the importance of making a careful choice of the objective function for model calibration. ~~As can be expected, a particular SFC is best estimated when the model calibration is based on that SFC (I_{Single}). However, a SFC-specific model calibration generally results in rather poorly simulated hydrographs, which negatively affects the estimation accuracy of SFCs that were not included in the model calibration. This poor estimation of SFCs can be improved by constraining model parameters not only to one SFC but also to R_{eff} (I_{Single_Reff}). Based on the study results it could be expected that the application of an objective function that addresses multiple aspects of a hydrograph improves runoff simulations for calculating a suite of SFCs. Calibration approaches based on simulating the general shape of the hydrograph (I_{Multi}, I_{Multi_Reff} and R_{eff}) reveal distinct results regarding individual SFCs, R_{eff} and MARE. R_{eff}, and to a lesser extent MARE, are improved with the more weight R_{eff} has in model calibration, whereas SFC estimates tend to be more accurate when SFCs are part of the objective function (in combination with R_{eff}). The results confirm that the objective functions I_{Multi} and I_{Multi_Reff} constrain the model better for simulating the general shape of the hydrograph and thus are more suited for model simulations aiming at many different SFCs than SFC-specific model calibrations. However, considering that SFCs not incorporated in the objective function showed little change in estimation error brings into question the benefit of including SFCs into model calibration instead of applying a traditional calibration approach based on R_{eff}.~~ The benefit of optimizing one specific SFC lies in the relatively accurate estimation of the respective SFC compared to a calibration with R_{eff} or a multi-SFCs objective function. Model calibration on one single SFC clearly emphasizes the hydrograph aspects of the selected SFC possibly neglecting an adequate representation of other hydrograph characteristics. This implies that calibrations with I_{Single} can lead to poor model performance for SFCs not included in the objective function. However, it is important to be aware of that an excellent calibration with I_{Single} does not guarantee that the respective SFC is well simulated in validation (see next two paragraphs for a discussion about the robustness of SFCs). The fact that a calibration with R_{eff} and a calibration with multiple SFCs lead to comparable estimates for most SFCs indicates that the main hydrological processes of the catchments are similarly well represented with the two approaches. We assume that these two calibration criteria result in a better process representation than the calibration with a single SFC, because they outperform the calibration with I_{Single} for those SFCs not included in I_{Single}. Considering that SFCs not incorporated in the objective function showed little change compared to calibrations with R_{eff} brings into question the benefit of including SFCs into model calibration instead of applying a traditional calibration approach. This is surprizing because the SFCs selected for I_{Multi} or I_{Multi_Reff} provide information on high-flows, recession rate, percentage of base flow and annual runoff volume and therefore should help constraining the model with respect to different important runoff processes. These results are different from those of Yilmaz et al. (2008) and

Pfannerstill et al. (2014) whose multi-metric runoff model calibration resulted in an improved general shape of the hydrograph. Although their calibration approach was mainly based on various segments of the flow duration curve, it is unclear why the conclusions differ that much. From the above discussion it becomes evident that ~~C~~calibrating a runoff model for estimating many different SFCs from one single hydrograph ~~becomes~~ is a trade-off between finding a parameterization that is general enough to represent different aspects of the hydrograph and that simultaneously emphasizes specific SFCs. ~~As stated by Caldwell et al. (2015), there is little chance to find an objective function suitable to estimate all SFCs because fitting model parameters to some hydrograph aspects inevitably disregards other aspects. Similar conclusions were drawn by Zhang et al. (2016) who calibrated a runoff model with a multi objective function consisting of 16 SFCs of interest to capture an overall flow regime. While applying the multi objective function resulted in an increased performance for low-flow and high-flow magnitudes, they reported a decreased model performance for mean flow magnitude related SFCs.~~ These trade-off situations are common as perfect model parameterizations are usually not possible due to a variety of uncertainty sources, such as model structural uncertainty and input and runoff data uncertainty (Beven, 2016). ~~In addition, various parameterizations can also have their strengths and weaknesses for different parts of the hydrograph.~~

A noticeable result from the current study is the distinct difference in model performance in calibration and validation when using the objective function I_{Single}. While almost perfect fits are achieved in calibration for all catchments and SFCs, model errors tend to be much higher in validation with a considerable spread between catchments as well as a clear difference depending on the SFC. This observation confirms that the model is able to simulate the SFCs well, but also outlines that a good model calibration does not imply robust simulations in validation. In general, it seems that SFCs that are strongly related to physical catchment properties (e.g. rate of streamflow recession) are the most robust, followed by SFCs representing average flow condition with a moderate robustness. SFCs that are a measure of more extreme high-flow conditions are the least robust, possibly because these conditions are subject to inter-annual weather changes and are more difficult to model due to their dynamic behaviour. A low robustness could also indicate that the model structure might be suboptimal for some catchments.

The two least robust SFCs are MH10 and TL1. MH10 simulations with I_{Single} yield by far the poorest results of all objective function types with very large normalized error in both positive and negative directions. In comparison, the high estimation errors for TL1 depend on the modelling time period. The high estimation errors for TL1 in period 2 stem from years where the minimum runoff was simulated in late winter while the observed minimum was in late fall. By visually analyzing the temperature and runoff time series, it can be hypothesized that such model simulations mainly happened in years with successive weeks of continuously little precipitation during late winter. Such prolonged drier periods occurred more often in one of the two modelling time periods and thus evoked the distinct bias in model accuracy depending on the simulation period. Both TL1 and MH10 are calculated from a single value per year, as opposed to e.g. RA7, which is based on all recessions. In model calibration, many parameter sets are derived that perfectly simulate this single value. However, a good simulation of either TL1 or MH10 is not so much dependent on an accurate representation of dominant runoff processes.

Thus, model results for the validation period using input data of identical quality can fail to accurately simulate either SFC because of parameter sets 'tuned' to the data as opposed to being based on modelling the process.

4.2 Model performance regarding SFCs

The runoff model tends to underestimate SFCs related to mean and high-flow conditions, while SFCs representing low-flow conditions are generally overestimated. These results are consistent with those of Olsen et al. (2013), Caldwell et al. (2015), and Vis et al. (2015) and can partly be explained by the model behaviour characterized by a less pronounced runoff response to precipitation events but increased groundwater discharge to the stream during drier periods compared to the observed data (Vis et al., 2015). The observations that average flow conditions are better simulated than extremes (Caldwell et al., 2015; Vis et al., 2015) or that high-flow related SFCs are more accurately estimated than those related to low flow (Shrestha et al., 2014; Ryo et al., 2015) cannot be confirmed with our results. None of these earlier studies explicitly included SFCs into model calibration and the deviating results could be attributed to the differing approaches to defining the objective function(s). This presumption is supported by the previously described differences in results of Vis et al. (2015) although they applied the same runoff model, catchments and SFCs.

4.3 How to select SFCs for a multi-index calibration approach

The current study supports the assumption that including SFCs into model calibration helps to preserve most hydrograph aspects relevant to those SFCs. Thus, an objective function based on several SFCs is expected to result in a hydrograph from which a suite of SFCs can be calculated. Not knowing which SFCs will be relevant for a given study, a guideline as to which SFCs the model calibration could be based on would be helpful. The first step towards a guideline consists of selecting SFCs that are potentially valuable for model calibration. This selection was based on the concept of robustness and information value of SFCs, which is comparable to the approach used by Euser et al. (2013) who assessed the realism of model structures. Like Euser et al. (2013), results from the current study indicated that high robustness was not necessarily related to high information value, emphasizing the importance of selecting SFCs by jointly evaluating robustness and information value. The concept of information value and robustness favours simulations that preserve important hydrograph characteristics as can be seen from the slightly improved median estimation accuracy of SFCs with the objective functions I_{Multi} or I_{Multi_Reff} compared to estimations with R_{eff} only.

A model calibrated on certain flow conditions (low, medium and high flow) is beneficial for SFCs representing these flow conditions (see e.g. Murphy et al., 2013), so it was hypothesized that the information value of the selected SFCs is highest for SFCs belonging to the same group of flow conditions. The confirmation of this hypothesis would allow to draw general conclusions about a minimum number of SFCs required for model calibration. Surprisingly the results did not reveal any pattern related to flow conditions and thus no recommendation for the final selection of SFCs can be made. It seems that the selection of SFCs for an informative and robust objective function depends on the type and the combination of SFCs one is interested in. Since this study was based on a limited number of SFCs it could be interesting to test the hypothesis by

analyzing a greater number of SFCs. Testing a larger number of SFCs might reveal relations that are difficult to see with a small sample. Furthermore, more knowledge about the effect of single SFCs or the combination of SFCs used as objective functions on runoff simulations could be gained by using synthetic data and a modelling approach where an excellent hydrograph fit is possible (e.g. HBV-land in Seibert and Vis, 2012).

4.4 Objective functions, their estimation accuracy and consequences for practical applications

The emphasis of SFC-related modelling studies changed in recent years from estimating single SFCs to simulating a suite of SFCs (Olden and Poff, 2003). The modelling design of this study combined both approaches for the same SFCs and catchments and thus enabled a direct comparison of the results. Ideally, the runoff model could be calibrated to simulate a hydrograph for each catchment from which any SFC can be calculated. Such an approach ensures a relatively small calibration effort, which is especially valuable if one is interested in modelling many catchments and/or various scenarios. However, results indicate that SFCs related to a more generally calibrated model (e.g. R_{eff}, I_{Multi} or I_{Multi_Reff}) are less accurate than when they are estimated from hydrographs based on targeted model calibrations (e.g. I_{Single} or I_{Single_Reff}). This fact has substantial implications for the later application of simulated SFCs in decision-support systems for integrated resource management. ~~related to flow alteration – ecosystem change relationships.~~ As stated by Carlisle et al. (2010), with high errors in SFC estimates, only considerable flow departures from natural conditions can be detected. Also, inaccurate SFC values can impede the generation of more robust flow alteration – ecosystem change relationships that are ultimately needed for sustainable flow management guidelines (Arthington et al., 2006; Poff and Zimmermann, 2010; Gillespie et al., 2015).

As with regional statistical approaches, incorporating SFCs into model objective functions implies that a modeller knows which SFCs are relevant and that the model must be recalibrated if one is interested in additional SFCs. The advantage of runoff models over multivariate regressions and observed streamflow series includes their use for climate scenario analysis or for simulating runoff in ungauged catchments with the latter being one of the ultimate aims in the ELOHA framework (Poff et al., 2010). Modelling SFCs gets even more challenging when moving from a gauged to an ungauged catchment. An appropriate calibration strategy targeted to the main simulation goal is crucial for any subsequent regionalization.

4.5 Choice of the runoff model for estimating SFCs

When comparing SFCs estimated from simulations of different runoff models (e.g. HBV, Precipitation runoff modelling system (PRMS), etc.), the question can be raised whether the results depend on the selected model. This question is especially important for resource managers who need to make decisions based on model results from different studies (Caldwell et al., 2015). A comparison of runoff models with different spatial scales that rely on different data inputs was conducted by Caldwell et al. (2015). Their results do not indicate that a certain runoff model is more suited for predicting SFCs than others, but rather that the calibration process probably has as much influence as the model structure. Thus, it can be assumed that the conclusions of this study would be similar if a different calibrated runoff model was applied.

5 Conclusions

In this study, we evaluated the value of using SFCs for the calibration of a runoff model used to estimate SFCs. The results suggest that the choice of the objective function used for model calibration strongly influences the estimation accuracy of SFCs. While the model was capable of correctly simulating any of the tested SFCs, a good reproduction of a particular SFC was generally achieved when this SFC was included in the objective function. SFC estimates from model simulations with an objective function consisting of a representative selection of SFCs resulted in comparable accuracies to the estimates from model runs based on the commonly used Nash–Sutcliffe efficiency when evaluated against SFCs not included in the objective function. Estimates of SFCs that are less dependent on the short-term weather input or SFCs representing average flow conditions were more robust than other SFCs. Since the results imply that one has to consider significant uncertainties when simulated time series are used to derive SFCs that were not included in the calibration, we strongly recommend calibrating the runoff model explicitly for the SFCs of interest.

Data availability

Data used in this study is available at the U.S. Department of Commerce (2007a, 2007b,) and the U.S. Geological Survey (2016a, b).

Author contributions

Sandra Pool, Marc Vis, Rodney Knight and Jan Seibert designed this study based on a previous collaboration; Marc Vis performed the runoff simulations; Sandra Pool analyzed the results that were discussed with all coauthors. Writing of the paper was led by Sandra Pool with contribution of all coauthors.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

This paper is a product of discussions and activities that took place at the U.S. Geological Survey John Wesley Powell Center for Analysis and Synthesis as part of the workgroup focusing on Water Availability for Ungauged Rivers (https://powellcenter.usgs.gov/). Funding for this research was provided by the U.S. Geological Survey Cooperative Water Program and the University of Zurich. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Abell, R. A., Olson, D. M., Dinerstein, E., Hurley, P. T., Diggs, J. T., Eichbaum, W., Walters, S., Wettengel, W., Allnutt, T., Loucks, C. J., and Hedao, P. (Eds.): Freshwater ecoregions of North America: A conservation assessment, Island Press, Washington, DC, USA, 2000.

Arthington, A. H., Bunn, S. E., Poff, N. L., and Naiman, R. J.: The challenge of providing environmental flow rules to sustain river ecosystems, Ecol. Appl., 16, 1311–1318, doi:10.1890/1051-0761(2006)016[1311:TCOPEF]2.0.CO;2, 2006.

Bergström, S.: Development and application of a conceptual runoff model for Scandinavian catchments, SMHI, Norrköping, Sweden, No. RHO 7, 134 pp., 1976.

Beven, K.: Facets of uncertainty: Epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication, Hydrol. Sci. J., 61, 1652–1665, doi:10.1080/02626667.2015.1031761 , 2016.

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.

Caldwell, P. V., Kennen, J. G., Sun, G., Kiang, J. E., Butcher, J. B., Eddy, M. C., Hay, L. E., LaFontaine, J. H., Hain, E. F., Nelson, S. A. C., and McNulty, S. G.: A comparison of hydrologic models for ecological flows and water availability, Ecohydrology, 8, 1525–1546, doi:10.1002/eco.1602, 2015.

Carlisle, D. M., Falcone, J., Wolock, D. M., Meador, M. R., and Norris, R. H.: Predicting the natural flow regime: models for assessing hydrological alteration in streams, River Res. Appl., 26, 118–136, doi:10.1002/rra.1247, 2010.

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., and Savenije, H. H. G.: A framework to assess the realism of model structures using hydrological signatures, Hydrol. Earth Syst. Sci, 17, 1893–1912, doi:10.5194/hess-17-1893-2013, 2013.

Gillespie, B. R., Desmet, S., Kay, P., Tillotson, M. R., and Brown, L. E.: A critical analysis of regulated river ecosystem responses to managed environmental flows from reservoirs, Freshwater Biol., 60, 410–425, doi:10.1111/fwb.12506, 2015.

Hailegeorgis, T. T., and Alfredsen, K.: Regional statistical and precipitation-runoff modelling for ecological applications: Prediction of hourly streamflow in regulated rivers and ungauged basins, River Res. Appl., in press, doi:10.1002/rra.3006, 2016.

Hingray, B., Schaefli, B., Mezghani, A., and Hamdi, Y.: Signature-based model calibration for hydrological prediction in mesoscale Alpine catchments, Hydrol. Sci. J., 55, 1002–1016, doi:10.1080/02626667.2010.505572, 2010.

Hoos, A. B.: Recharge rates and aquifer hydraulic characteristics for selected drainage basins in middle and east Tennessee, U.S. Geological Survey, Nashville, Tennessee, USA, Water Resources Investigations Report 90–4015, 39 pp., 1990.

Jothityangkoon, C., Sivapalan, M., and Farmer, D. L.: Process controls of water balance variability in a large semi-arid catchment: Downward approach to hydrological model development, J. Hydrol., 254, 174–198, doi:10.1016/S0022-1694(01)00496-6, 2001.

Knight, R. R., Brian Gregory, M., Wales, A. K.: Relating streamflow characteristics to specialized insectivores in the Tennessee River Valley: A regional approach, Ecohydrology, 1, 394–407, doi:10.1002/eco.32, 2008.

Knight, R. R., Gain, W. S., and Wolfe, W. J.: Modelling ecological flow regime: an example from the Tennessee and Cumberland River basins, Ecohydrology, 5, 613–627, doi:10.1002/eco.246, 2012.

Knight, R. R., Murphy, J. C., Wolfe, W. J., Saylor, C. F., and Wales, A.K. Ecological limit functions relating fish community response to hydrologic departures of the ecological flow regime in the Tennessee River basin, United States, Ecohydrology, 7, 1262–1280, doi:10.1002/eco.1460, 2014.

Krause, P., Boyle, D. P., and Bäse, F.: Comparison of different efficiency criteria for hydrological model assessment, Adv. Geosci., 5, 89–97, doi:1680-7359/adgeo/2005-5-89, 2005.

Law, G. S., Tasker, G. D., and Ladd, D. E: Streamflow-characteristic estimation methods for unregulated streams of Tennessee, U.S. Geological Survey, Reston, Virginia, USA, Scientific Investigations Report 2009–5159, 212 pp., 2009.

Legates, D. R., and McCabe, G. J.: Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation, Water Resour. Res., 35, 233–241, doi:10.1029/1998WR900018, 1999.

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

Murphy, J. C., Knight, R. R., Wolfe, W. J., & S Gain, W.: Predicting ecological flow regime at ungauged sites: A comparison of methods, River Res. Appl., 29, 660–669, doi:10.1002/rra.2570, 2013.

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

Olden, J. D., and Poff, N. L.: Redundancy and the choice of hydrologic indices for characterizing streamflow regimes, River Res. Appl., 19, 101–121, doi:10.1002/rra.700, 2003.

Olsen, M., Troldborg, L., Henriksen, H. J., Conallin, J., Refsgaard, J. C., and Boegh, E.: Evaluation of a typical hydrological model in relation to environmental flows, J. Hydrol., 507, 52–62, doi:10.1016/j.jhydrol.2013.10.022, 2013.

Omernik, J. M.: Ecoregions of the Conterminous United States, Ann. Assoc. Am. Geogr., 77, 118–125, doi:10.1111/j.1467-8306.1987.tb00149.x, 1987.

Pfannerstill, M., Guse, B., and Fohrer, N.: Smart low flow signature metrics for an improved overall performance evaluation of hydrological models, J. Hydrol., 510, 447–458, doi:10.1016/j.jhydrol.2013.12.044, 2014.

Poff, N. L., and Zimmerman, J. K.: Ecological responses to altered flow regimes: A literature review to inform the science and management of environmental flows, Freshwater Biol., 55, 194–205, doi:10.1111/j.1365-2427.2009.02272.x, 2010.

Poff, N. L., Allan, J. D., Bain, M. B., Karr, J. R., Prestegaard, K. L., Richter, B. D., Sparks, R. E., and Stromberg, J. C.: The natural flow regime, BioScience, 47,769–784, doi:10.2307/1313099, 1997.

Poff, N. L., Richter, B. D., Arthington, A. H., Bunn, S. E., Naiman, R. J., Kendy, E., Acreman, M., Apse, C., Bledsoe, B. P., Freeman, M. C., Henriksen, J., Jacobson, R. B., Kennen, J. G., Merritt, D. M., O`Keeffe, Y. H., Olden, J. D., Rogers, K., Tharme, R. E., and Warner, A.: The ecological limits of hydrologic alteration (ELOHA): A new framework for developing regional environmental flow standards, Freshwater Biol., 55, 147–170, doi:10.1111/j.1365-2427.2009.02204.x, 2010.

Richter, B. D., Baumgartner, J. V., Powell, J., and Braun, D. P.: A method for assessing hydrologic alteration within ecosystems, Conserv. Biol., 10, 1163–1174, doi:10.1046/j.1523-1739.1996.10041163.x, 1996.

Rotstayn, L. D., Roderick, M. L., and Farquhar, G. D: A simple pan-evaporation model for analysis of climate simulations: Evaluation over Australia. Geophys. Res. Lett., 33, L7715, doi:10.1029/2006GL027114, 2006.

Ryo, M., Iwasaki, Y., and Yoshimura, C.: Evaluation of spatial pattern of altered flow regimes on a river network using a distributed hydrological model, PloS ONE, 10, e0133833, doi:10.1371/journal.pone.0133833, 2015.

Sanborn, S. C., and Bledsoe, B. P.: Predicting streamflow regime metrics for ungauged streams in Colorado, Washington, and Oregon, J. Hydrol., 325, 241–261, doi:10.1016/j.jhydrol.2005.10.018, 2006.

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A., and Carrillo G.: Catchment classification: Empirical analysis of hydrologic similarity based on catchment function in the eastern USA, Hydrol. Earth Syst. Sci., 15, 2895–2911, doi:10.5194/hess-15-2895-2011, 2011.

Seibert, J.: Regionalization of parameters for a conceptual rainfall-runoff model, Agric. For. Meteorol., 98–99, 279–293, doi:10.1016/S0168-1923(99)00105-7, 1999.

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

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, doi:10.5194/hess-16-3315-2012, 2012.

Shrestha, R. R., Peters, D. L., and Schnorbus, M. A.: Evaluating the ability of a hydrologic model to replicate hydro-ecologically relevant indicators, Hydrol. Process., 28, 4294–4310, doi:10.1002/hyp.9997, 2014.

Tharme, R. E.: A global perspective on environmental flow assessment: Emerging trends in the development and application of environmental flow methodologies for rivers, River Res. Appl., 19, 397–441, doi:10.1002/rra.736, 2003.

U.S. Department of Commerce: Divisional normals and standard deviations of temperature, precipitation, and heating and cooling degree days 1971–2000 (and previous normals periods), section 2 precipitation, United States Department of Commerce, Washington, DC, USA, Climatography of the United States No. 85, 2007a.

U.S. Department of Commerce: Divisional normals and standard deviations of temperature, precipitation, and heating and cooling degree days 1971–2000 (and previous normals periods), section 1 temperature, United States Department of Commerce: Washington, DC, USA, Climatography of the United States No. 85, 2007b.

U.S. Geological Survey: EflowStats R-package, https://github.com/USGS-R/EflowStats, last access: July 2016, 2014.

U.S. Geological Survey: The National Map, 3D Elevation Program Products and Services Web page, http://nationalmap.gov/3DEP/3dep_prodserv.html, last access: November 2015, 2016a.

U.S. Geological Survey: National Water Information System - Web interface, http://dx.doi.org/10.5066/F7P55KJN, last access: October 2016, 2016b.

Viglione, A., Parajka, J., Rogger, M., Salinas, J. L., Laaha, G., Sivapalan, M., and Blöschl, G.: Comparative assessment of predictions in ungauged basins-Part 3: Runoff signatures in Austria, Hydrol. Earth Syst. Sci, 17, 2263–2279, doi:10.5194/hess-17-2263-2013, 2013.

Vis, M., Knight, R., Pool, S., Wolfe, W., and Seibert, J.: Model calibration criteria for estimating ecological flow characteristics, Water, 7, 2358–2381, doi:10.3390/w7052358, 2015.

Wagener, T., Sivapalan, M., Troch, P., and Woods, R.: Catchment classification and hydrologic similarity, Geogr. Compass, 1, 901–931, doi:10.1111/j.1749-8198.2007.00039.x, 2007.

Westerberg, I. K., Guerrero, J. L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., Freer, J. E., and Xu, C. Y.: Calibration of hydrological models using flow-duration curves, Hydrol. Earth Syst. Sci., 15, 2205–2227, doi:10.5194/hess-15-2205-2011, 2011.

Westerberg, I. K., Wagener, T., Coxon, G., McMillan, H. K., Castellarin, A., Montanari, A., and Freer, J.: Uncertainty in hydrological signatures for gauged and ungauged catchments, Water Resour. Res., 52, 1847–1865, doi:10.1002/2015WR017635, 2016.

Wolfe, W., Haugh, C., Webbers, A., Diehl, T.: Preliminary conceptual models of the occurrence, fate, and transport of chlorinated solvents in karst regions of Tennessee, U.S. Geological Survey, Nashville, Tennessee, USA, Water Resources Investigations Report 97–4097, 88 pp., 1997.

Yadav, M., Wagener, T., and Gupta, H.: Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, Adv. Water Resour., 30, 1756–1774, doi:10.1016/j.advwatres.2007.01.005, 2007.

Yilmaz, K. K., H. V. Gupta, and Wagener T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resour. Res., 44, W09417, doi:10.1029/2007WR006716, 2008.

~~Zhang, Y., Shao, Q., Zhang, S., Zhai, X., She, D.: Multi-metric calibration of hydrological model to capture overall flow regimes, J. Hydrol., 539, 525–538, doi:10.1016/j.jhydrol.2016.05.053, 2016.~~

Table 1. Description of streamflow characteristics used to calibrate the runoff model (adapted from Knight et al., 2014; U.S. Geological Survey, 2014) [mm d-1, millimeters per day; -, no units; a-1, per annum; %, percent]

Streamflow characteristic	Abbre-viation	Definition	Flow condition	Unit
Magnitude				
Mean annual runoff	MA41	Annual mean daily streamflow	mean-flow	[mm d^{-1}]
Maximum October runoff	MH10	Mean maximum October streamflow across the period of record	high-flow	[mm d^{-1}]
Lowest 15% of daily runoff	E85	Daily mean streamflow that is exceeded 85% ~~exceedance of daily mean streamflow~~ of the time for the period of record	low-flow	[mm d^{-1}]
Rate of streamflow recession	RA7	Median change in log of streamflow for days in which the change is negative across the period of record	mean-flow	[mm d^{-1}]
Ratio				
Average 30-day maximum runoff	DH13	Mean annual maximum of a 30-day moving average streamflow divided by the median for the entire record	high-flow	[-]
Base flow	ML20	Ratio of total base flow to total flow. Base flow is the minimum flow magnitude in a 5-day window if 90% of that minimum flow magnitude is less than the minimum flow magnitude of the 5 day-window before and after the considered ~~block~~ window	low-flow	[-]
Stability of runoff	TA1	Measure of the constancy of a flow regime by dividing daily flows into predetermined flow classes. The 11 flow classes capture flow ranging from flow less than 0.1 times the logarithmic mean flow to flow more than 2.25 times the logarithmic mean flow	mean-flow	[-]
Frequency				
Frequency of moderate floods	FH6	Average number of high-flow events per year that are equal to or greater than three times the median annual flow for the period of record	high-flow	[a^{-1}]
Frequency of ~~moderate~~ larger floods	FH7	Average number of high-flow events per year that are equal to or greater than seven times the median annual flow for the period of record	high-flow	[a^{-1}]
Variability				
Variability of March runoff	MA26	Standard deviation for March streamflow over the period of record divided by the mean streamflow for March over the period of record	mean-flow	[%]
Variability in high-flow pulse duration	DH16	Standard deviation for the yearly average high-flow pulse duration (daily flow greater than the 75th percentile) divided by the mean of the yearly average high-flow pulse duration multiplied by 100	high-flow	[%]
Variability of low-flow pulse count	FL2	Standard deviation for the average number of yearly low-flow pulses (daily flow less than the	low-flow	[%]

		25th percentile) divided by the mean low-flow pulse counts multiplied by 100			
Date					
Timing of annual minimum runoff	TL1	Julian date of annual minimum flow occurrence	low-flow	[Julian day]	

Table 2. Objective functions used in model calibration. Objective functions were calculated with observed (obs) and simulated (sim) runoff (Q) or SFCs (I).

Objective function	Abbreviation	Definition	Optimal value
Nash-Sutcliffe efficiency	R_{eff}	$1 - \frac{\sum(Q_{\text{obs}} - Q_{\text{sim}})^2}{\sum(Q_{\text{obs}} - \overline{Q_{\text{obs}}})^2}$	1
Efficiency for each individual SFC[1]	I_{Single}	$1 - \frac{\lvert I_{\text{obs}} - I_{\text{sim}} \rvert}{I_{\text{obs}}}$	1
SFC and Nash-Sutcliffe efficiency	$I_{\text{Single\_Reff}}$	$0.5\ (I_{\text{Single}} + R_{\text{eff}})$	1
Efficiency for the selected SFCs[2]	I_{Multi}	$0.25\ (I_{\text{Single}_1} + \ldots + I_{\text{Single\_n}})$	1
SFCs and Nash-Sutcliffe efficiency	$I_{\text{Multi\_Reff}}$	$0.8\ I_{\text{Multi}} + 0.2\ R_{\text{eff}}$	1

[1]For each of the 13 SFCs a specific I_{Single} exists.

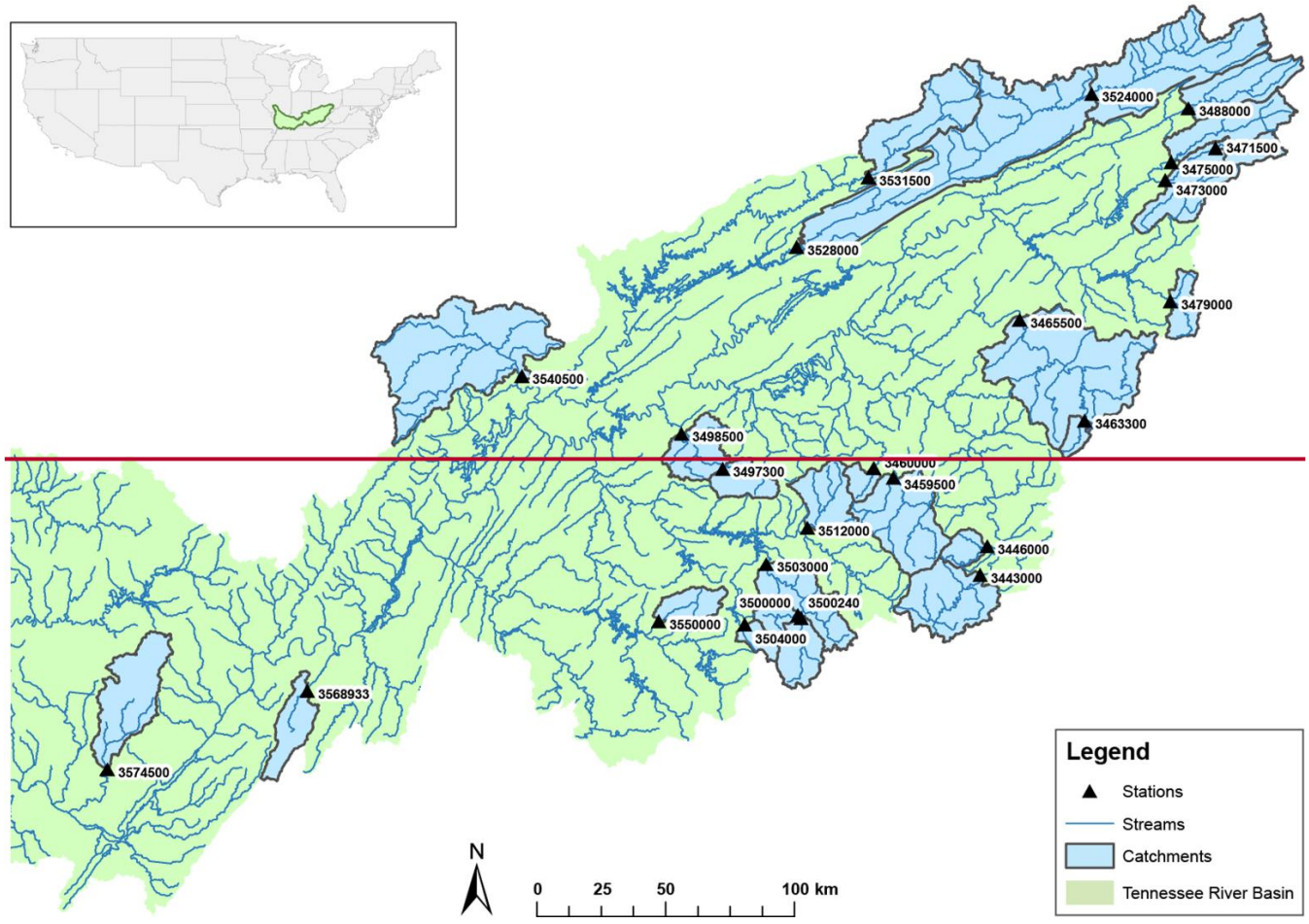
[2]I_{Multi} consists of the n most robust and informative SFCs.

Table 3. Performance measures used in model evaluation. Performance measures were calculated with observed (obs) and simulated (sim) runoff (Q) or SFCs (I).

Performance measure	Abbreviation	Definition	Optimal value
Nash-Sutcliffe efficiency	R_{eff}	$1 - \frac{\sum(Q_{\text{obs}} - Q_{\text{sim}})^2}{\sum(Q_{\text{obs}} - \overline{Q_{\text{obs}}})^2}$	1
Mean absolute relative error[1]	MARE	$1 - \frac{1}{n}\sum \frac{\lvert Q_{\text{obs}} - Q_{\text{sim}} \rvert}{Q_{\text{obs}}}$	1
Normalized SFC error[2]	nSFC	$\frac{I_{\text{obs}} - I_{\text{sim}}}{R_{\text{obs}}}$	0

[1] n is the number of days.

[2] R is the range of possible values of a SFC for the respective catchment.


3524000
3488000
3471500
3475000
3473000
3531500
3528000
3479000
3465500
3540500
3463300
3498500
3497300
3460000
3459500
3512000
3446000
3503000
3443000
3500000
3500240
3550000
3504000
3568933
3574500
N
0
25
50
100 km
Legend
Stations
Streams
Catchments
Tennessee River Basin

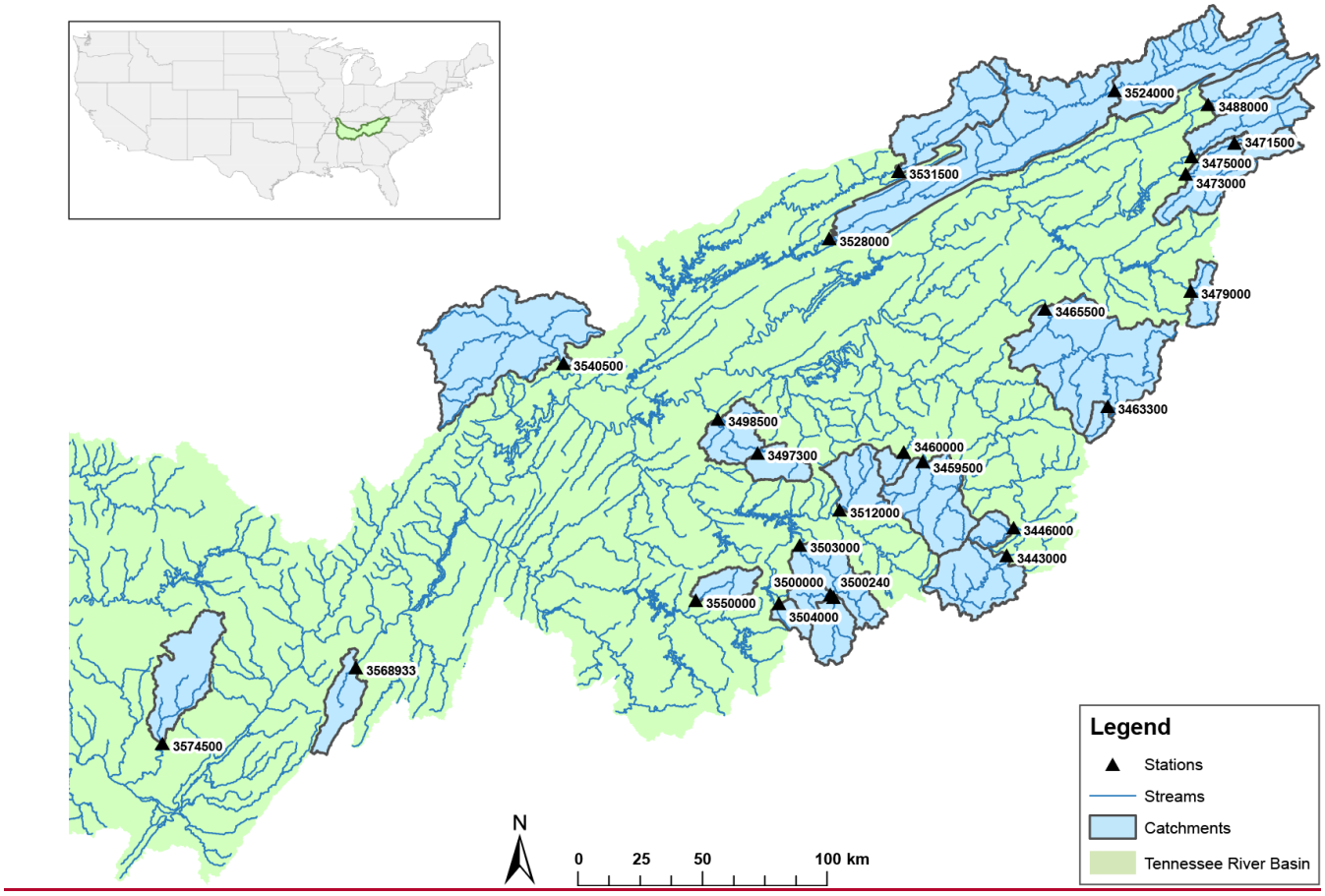


Figure 1. Location of the 25 study catchments in the Tennessee River basin (Table 1 in Vis et al. (2015) for more information).

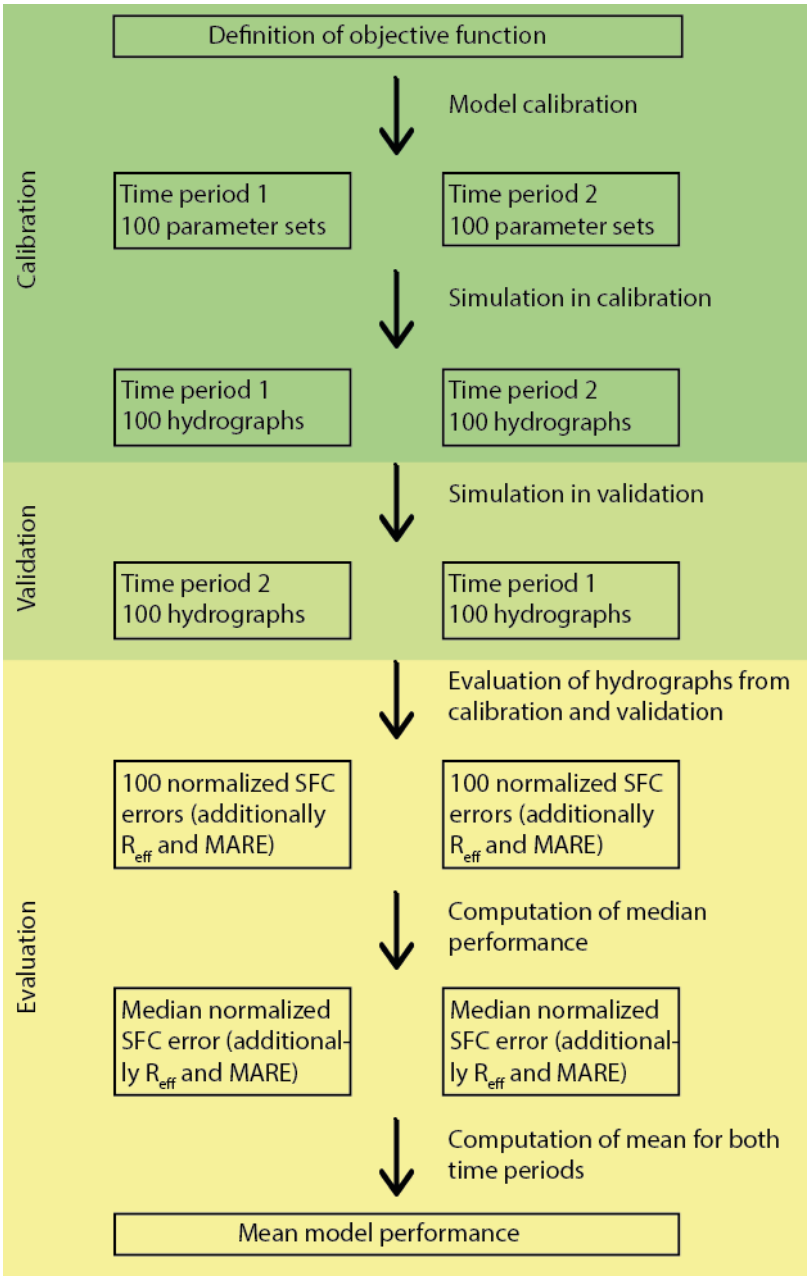
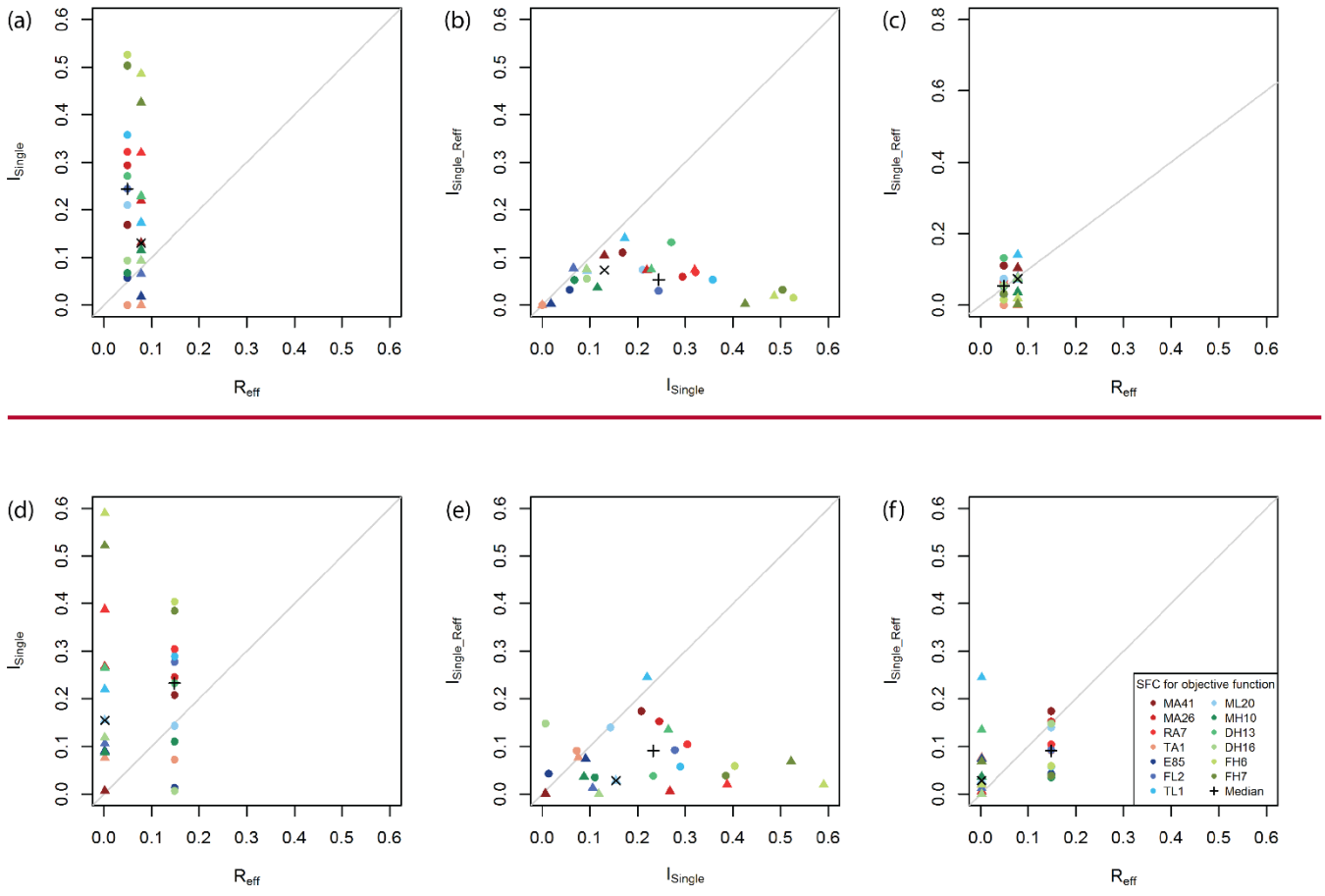


Figure 2. Flow chart of the modelling approach consisting of calibration, validation and evaluation in time period 1 (1984 - 1996) and time period 2 (1997 - 2009) and completed for each of the five objective function types R_{eff}, I_{Sinlge}, I_{Single_Reff}, I_{Multi}, I_{Multi_Reff}.


(a)
(b)
(c)
(d)
(e)
(f)
I_{Single}
I_{Single_Reff}
R_{eff}
SFC for objective function
MA41
MA26
RA7
TA1
E85
FL2
TL1
ML20
MH10
DH13
DH16
FH6
FH7
Median

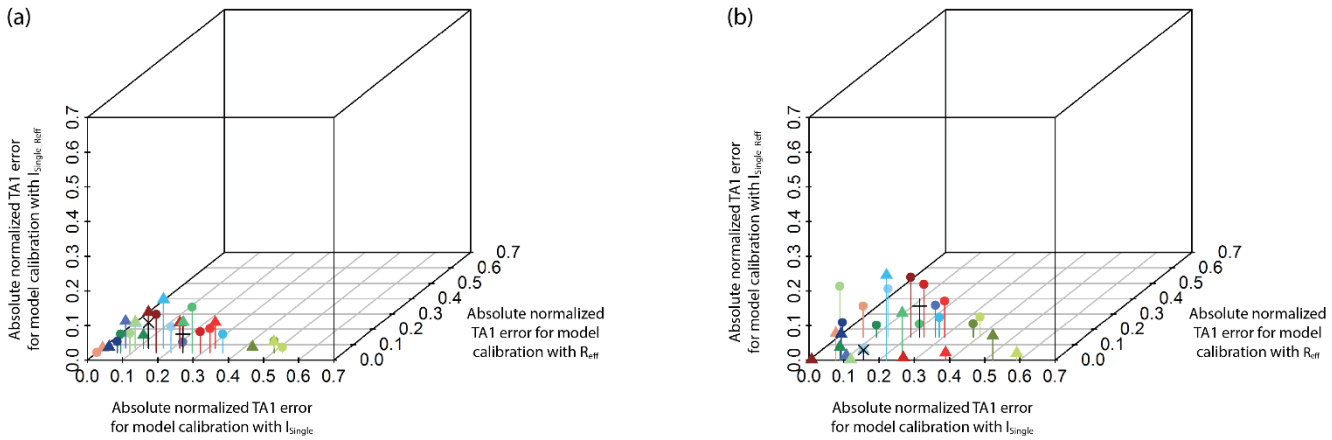


Figure 3. ~~Comparison of a~~Absolute normalized TA1 error (nSFC) in a) calibration ~~(a-c)~~ and b) validation ~~(d-f)~~ calculated from model calibrations with the objective functions R_{eff}, I_{Single} and I_{Single_Reff}. Absolute normalized SFC errors correspond to the median of the 25 catchments and are shown separately for both modelling time periods (triangles for period 1 (1984 - 1996) and circles for period 2 (1997 - 2009)). The x and plus symbols represent the median of period 1 and period 2 respectively.

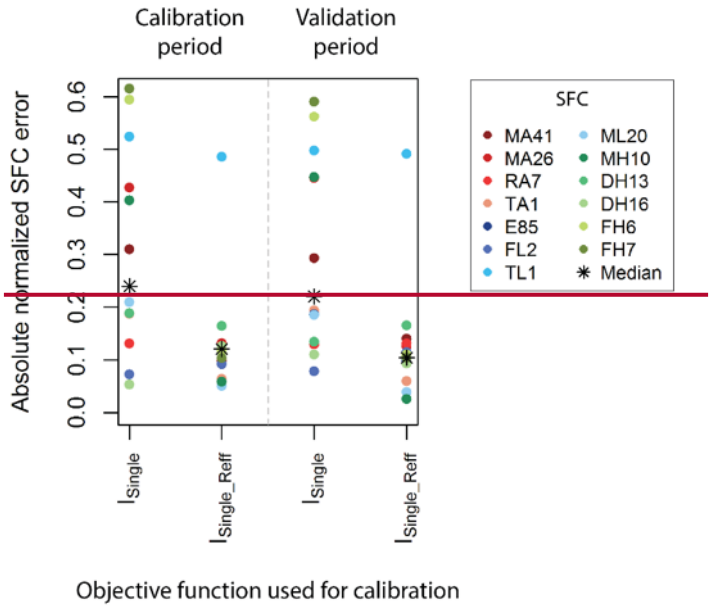
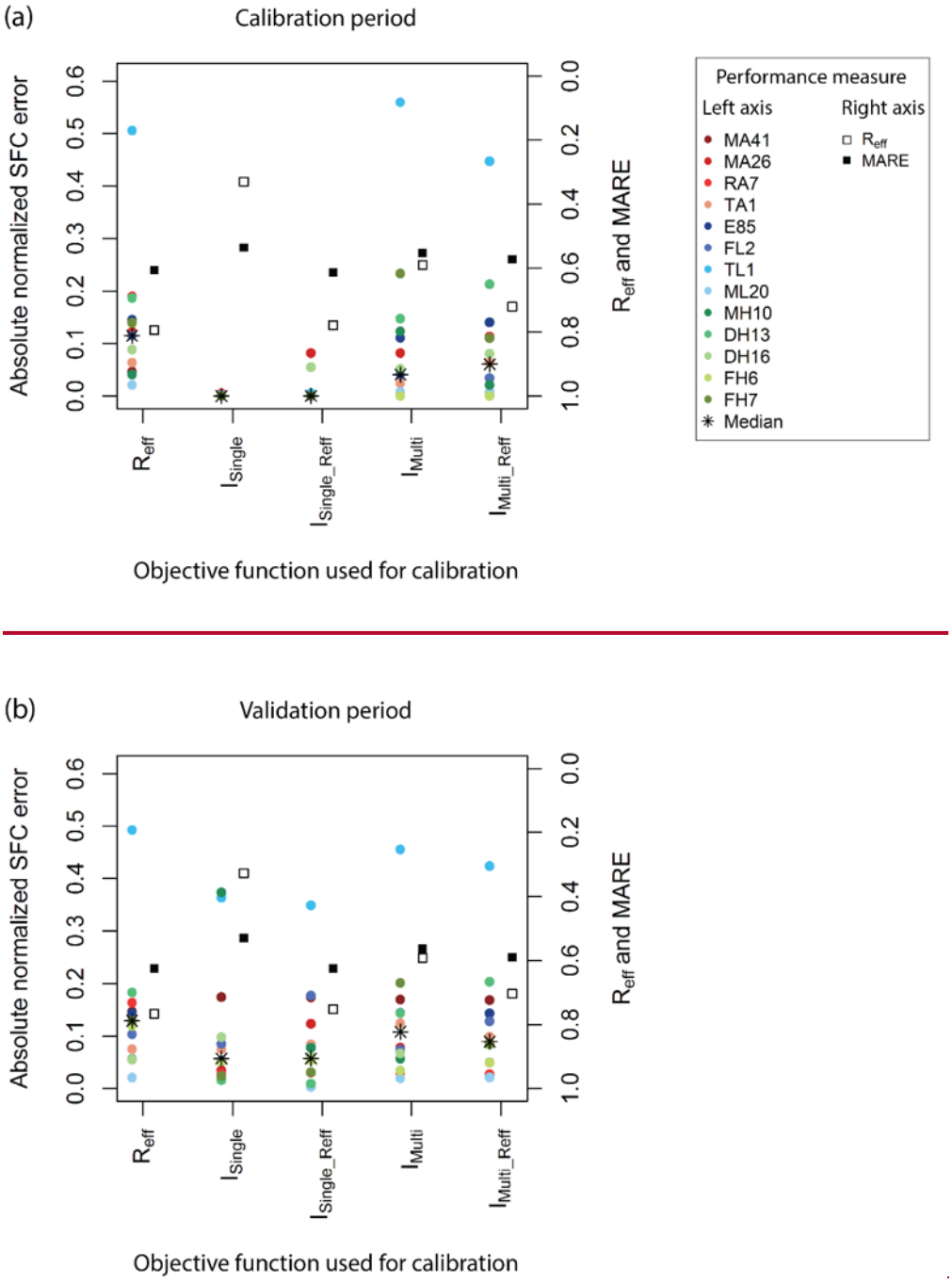


Figure 4. Absolute normalized SFC error (nSFC) for the model calibration (left side) and model validation periods (right side) calculated from model calibrations with the objective functions I_{Single} and I_{Single_Reff}. Values correspond to the median error of all 13 objective function versions and were calculated from the median of the 25 catchments and the mean of both modelling time periods.



~~**Figure 5.** Model performance in a) calibration and b) validation for absolute normalized SFC errors (nSFC) as well as R_{eff} and MARE depending on the the objective function used in calibration (optimal value is one for R_{eff} and MARE and zero for all SFC related performance measures). Model performance values correspond to the median of the 25 catchments and the mean of both modelling time~~

~~periods. R_{eff} and MARE values for the objective functions I_{Single} and I_{Single_Reff} were calculated as the median over all 13 versions. Note that in calibration with I_{Single} and I_{Single_Reff} the values of all or most absolute normalized SFCs plot at the same value.~~

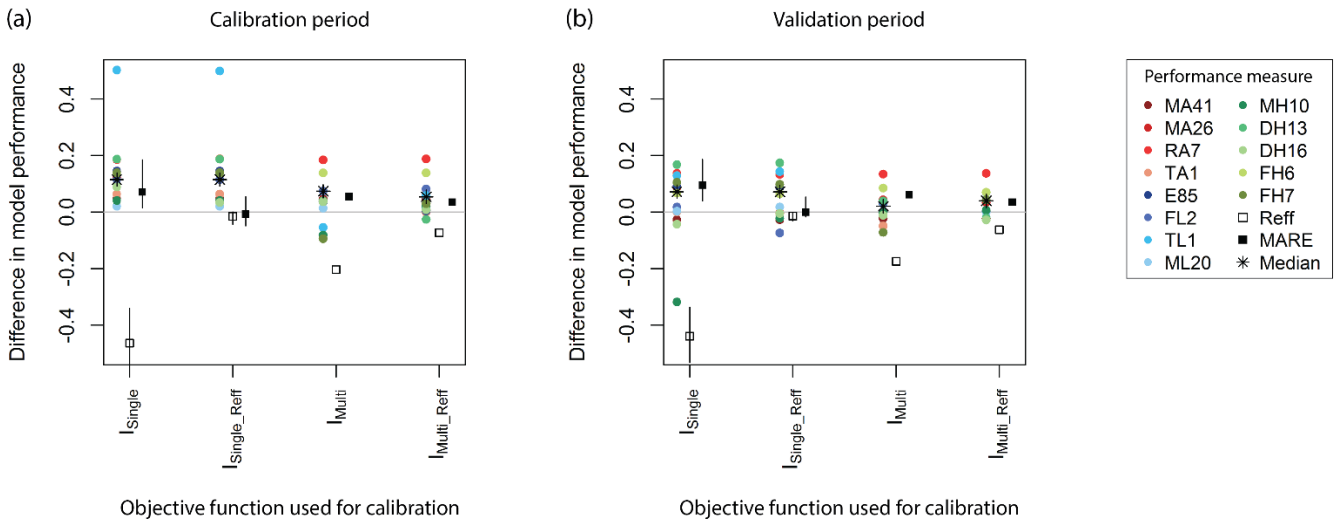
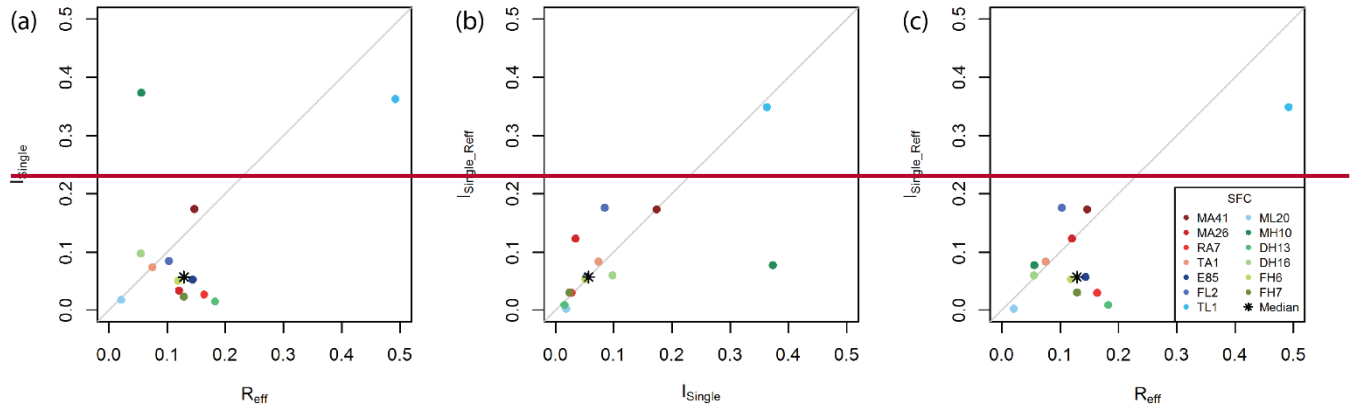


Figure 4. Model performance in a) calibration and b) validation for absolute normalized SFC errors (nSFC) as well as R_{eff} and MARE depending on the objective function used in calibration. Model performance is shown as the difference between a model calibration with R_{eff} and model calibrations with I_{Single}, I_{Single_Reff}, I_{Multi} or I_{Multi_Reff} (positive values indicate that model calibration with I_{Single}, I_{Single_Reff}, I_{Multi} or I_{Multi_Reff} resulted in better model performance than model calibration with R_{eff}; negative values indicate that model calibration with I_{Single}, I_{Single_Reff}, I_{Multi} or I_{Multi_Reff} resulted in poorer model performance than model calibration with R_{eff}). Model performance values correspond to the median of the 25 catchments and the mean of both modelling time periods.



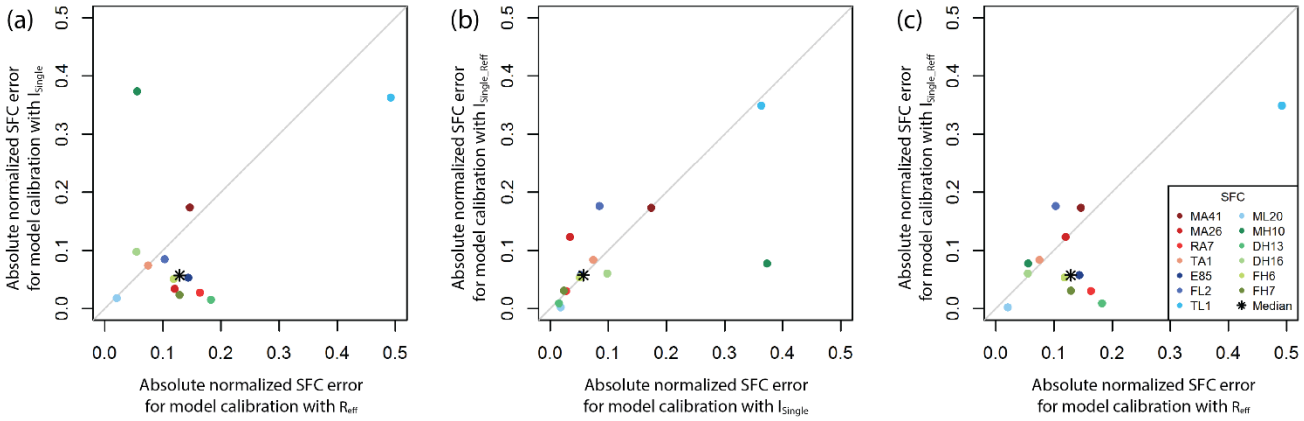


Figure 65. Comparison of absolute normalized SFC errors (nSFC) in validation calculated from model calibrations with the objective functions R_{eff}, I_{Single} and I_{Single_Reff}. Absolute normalized SFC errors correspond to the median of the 25 catchments and the mean of both modelling time periods.

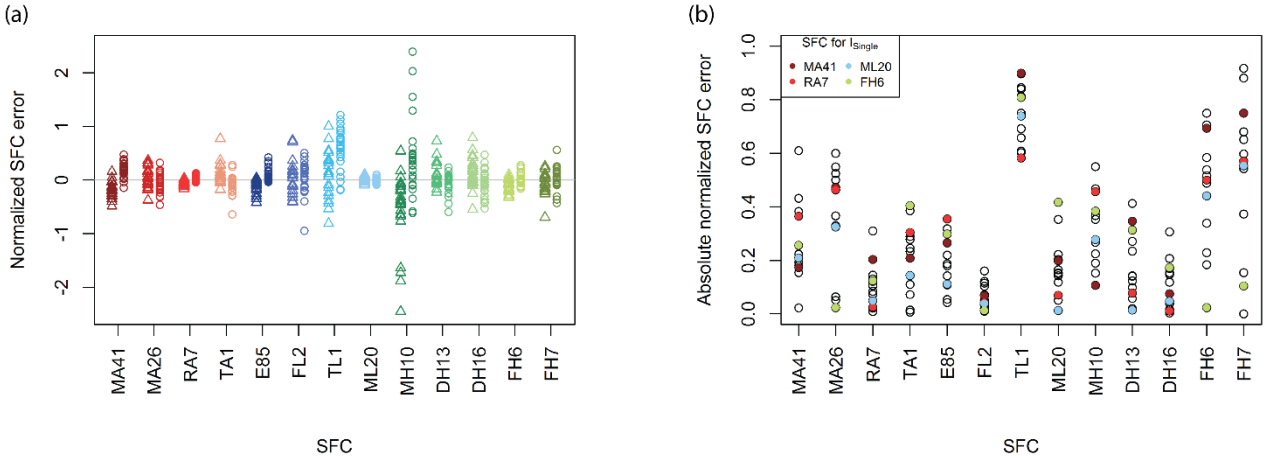


Figure 67. a) Robustness: normalized SFC errors (nSFC) in validation calculated from model calibrations with the objective function I_{Single} for the respective SFC. Values are shown for all 25 catchments and both modelling time periods (triangles for period 1 (1984 - 1996) and circles for period 2 (1997 - 2009)). b) Information value: absolute normalized SFC errors (nSFC) in validation calculated from model calibrations with all 13 objective functions I_{Single}. Model performance values correspond to the median of the 25 catchments and the mean of both modelling time periods. Each circle represents a SFC used for I_{Single}. The coloured circles show the information value of the final selection of SFCs for the objective function I_{Multi}.

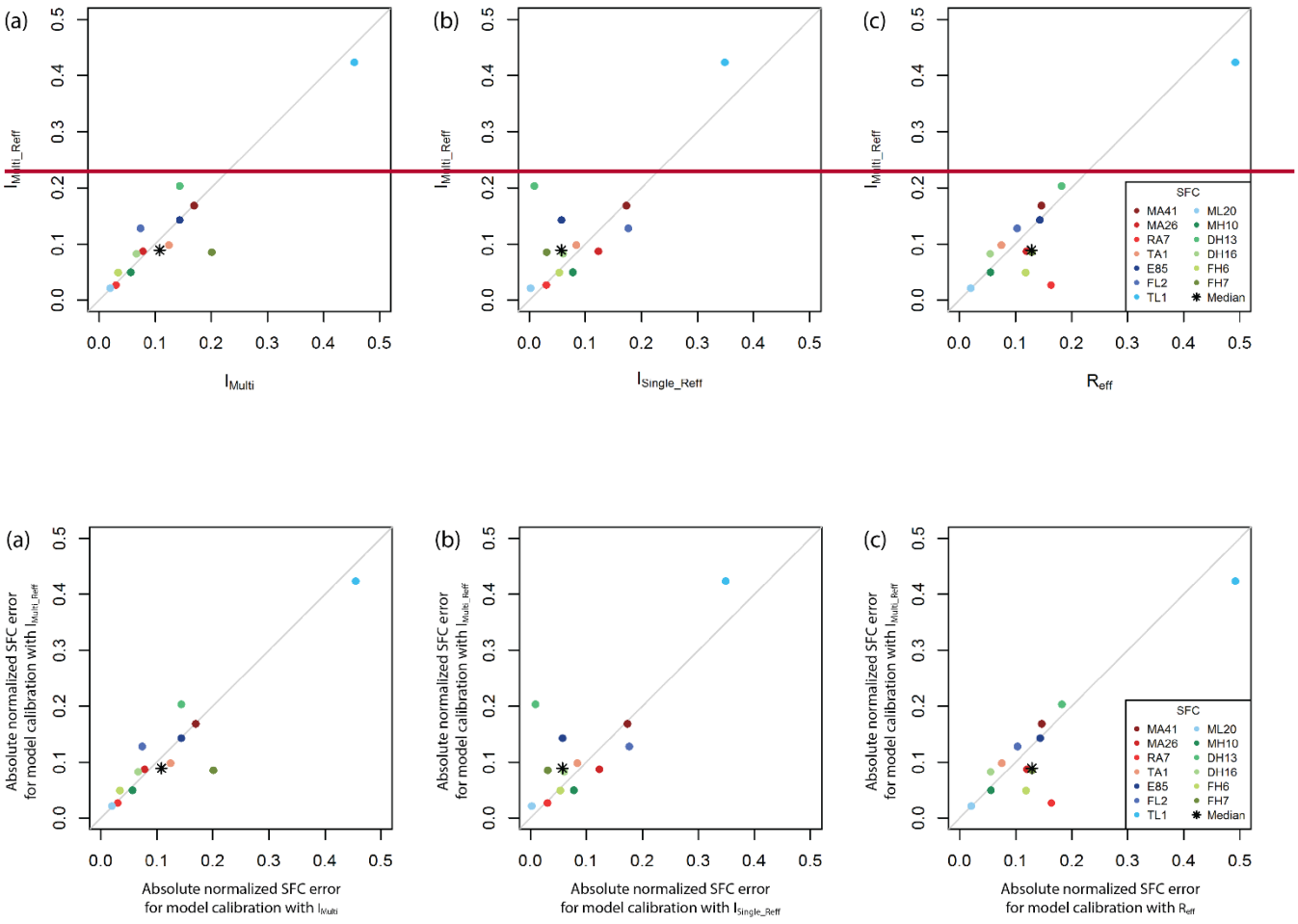


Figure ~~8~~7. Comparison of absolute normalized SFC errors (nSFC) in validation calculated from model calibrations with the objective functions R_{eff}, I_{Single_Reff}, I_{Multi} and I_{Multi_Reff}. Absolute normalized SFC errors correspond to the median of the 25 catchments and the mean of both modelling time periods.

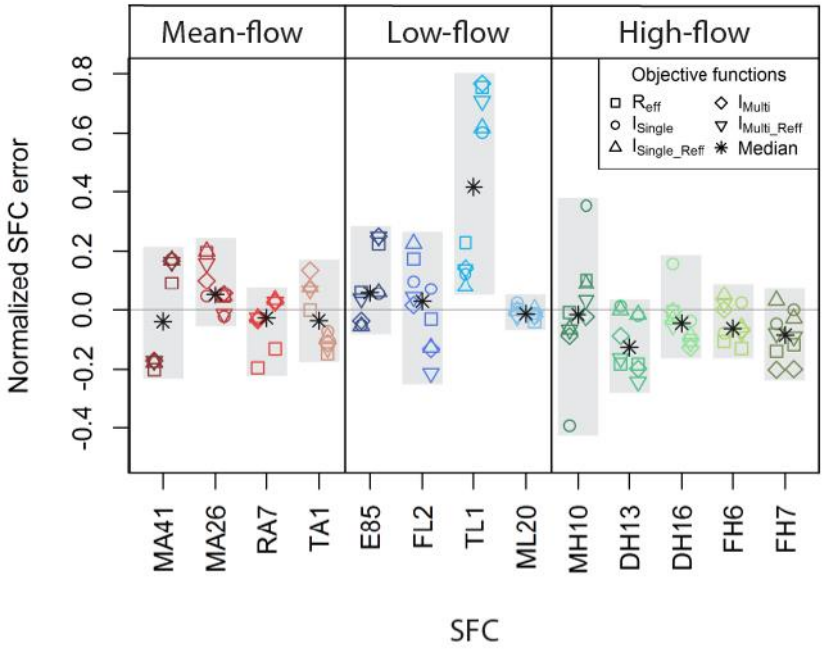


Figure ~~9~~8. Normalized SFC errors (nSFC) in validation depending on the objective function used in calibration. Model performance values correspond to the median of the 25 catchments and are shown for both modelling time periods (period 1 (1984 - 1996) on the left side and period 2 (1997 - 2009) on the right side).

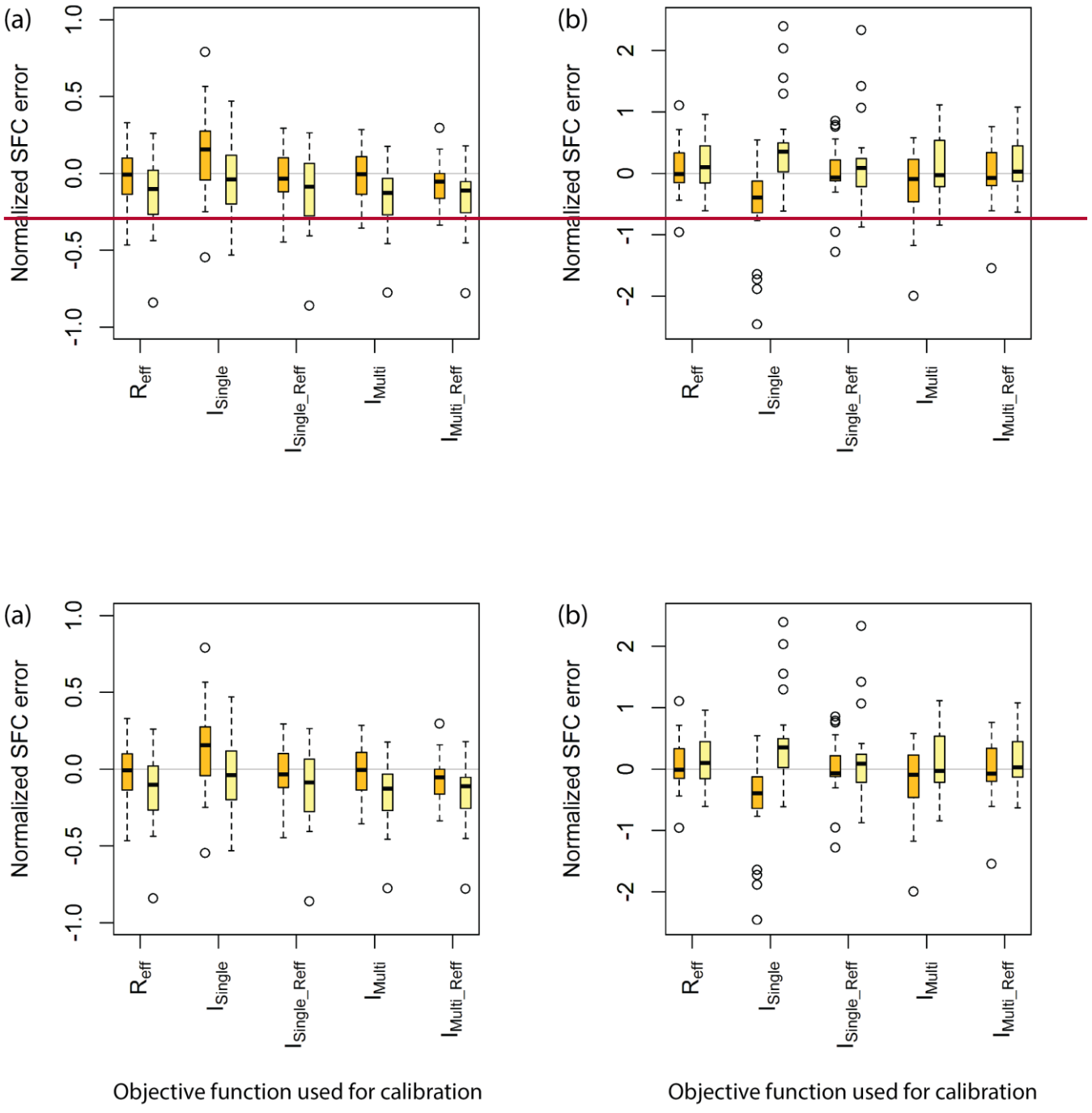


Figure ~~10~~9. a) Normalized DH16 errors (nSFC) and b) normalized MH10 errors (nSFC) in validation depending on the objective function used in calibration. Absolute normalized SFC errors are shown for all 25 catchments and for both modelling time periods (period 1 (1984 - 1996) in orange on the left side and period 2 (1997 - 2009) in yellow on the right side). Note the difference in y-axis.