3rd revision of manuscript

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 and the reviewer's efforts with our manuscript. The reviewer provided again valuable comments on our manuscript, which helped us to further improve the clarity of our manuscript. Below, we reply to each of the comments from the reviewer and indicate the changes that have been done accordingly (marked with blue color). References to pages (P), lines (L), chapters, figures and tables refer to the track-changed revised version of our manuscript. We also did a few additional minor edits that are marked in the track-changed manuscript.

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

Reviewer 3: comments to 2nd revision, response and modifications

Dear Sjur Kolberg,
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 your comments.
Best regards,
Sandra Pool and Co-Authors

Major comments

Comment 1: The paper contains no clear definition or distinction between a goodness-of fit measure based on an SFC, and a 'traditional calibration criterion' (TCC). In my view, a comparison of SFCs to TCCs should acknowledge that there may be a transition zone between the two, but still provide a clear distinction which makes the comparison meaningful. SFCs are (citing): 'often used to refer to <such> specific aspects of the flow regime', however, this is also true for R_{eff}, R_{eff}(ln), slope of FDC etc. In addition, the analysis use R_{eff} alone as TCC benchmark, although the necessity of combining different TCCs in calibration is well recognised. This constructs a comparison in favour of SFCs.
A definition of 'traditional criterion' could be used, for instance along the lines of 'A goodness of fit measure computed from all sim-obs residuals within the calibration period, possibly transformed'. Then distinct groups of SFCs could be recognised by being 'selected from specific seasons or situations', or 'based on duration of events or conditions' etc.

Reply 1: We agree that there is no clear boundary between traditional objective functions and SFC based objective functions. We account for this transition by not only calibrating on R_{eff} or SFCs, but also on combinations thereof (I_{Single_Reff} or I_{Multi_Reff}). The decision to use R_{eff} as benchmark in this study stems from the conclusions of the preceding study (Vis et al., 2015), where R_{eff} outperformed model calibrations with a multi-objective function with respect to SFCs (please see reply to comment 1 in the section 'detailed and specific comments').
In this study we use the term traditional objective function when referring to e.g. Nash-Sutcliffe efficiency or volume error. Such traditional objective functions mostly have a statistical background. In contrast, SFC based objective functions of this study focus on specific runoff aspects with ecological relevance in our study region. We therefore consider them as not being purely statistically

motivated. As suggested by the reviewer, we added a few sentences in the introduction to define traditional and SFC based objective functions.
Modification: P4 L10-13

Comment 2: With no rules for how SFCs are constructed, statements like 'High flow SFCs tend to be under-estimated' are meaningless. Flashiness index and concentration time both characterise high-flow behaviour, but underestimating one would mean overestimating the other. From table 1 the reader may verify that each SFC used in this paper is scaled so its value is positively related to flow magnitude for the relevant section of the FDC, but this should be asserted in the text when used in conclusions.
Reply 2: We agree that drawing conclusions over groups of flow conditions (e.g. high-flow) can lead to misleading conclusions if SFCs of that group represent contrasting hydrograph aspects (as shown in the example given by the reviewer). Fortunately, this is not the case for the flow conditions and SFCs used in this study (please see Table 1 for the definition of SFC). We adapted a sentence in the abstract (P1 L20) and in the results part (P9 L8-10) to make the reader aware of that the results for over- and underestimation only hold for the tested SFCs.
Five out of the 13 tested SFCs are scaled with mean, median or total runoff (Table 1). We agree, that this normalization leads to SFC values that are dependent on the flow magnitude. Although this has an influence on SFC errors, it does not affect the sign of the error. The normalization therefore does not change the conclusions about under- and overestimation of SFCs. We added a few sentences about the normalization of SFCs and its influence on the results in the methods section (P5 L4-7).
Modification: P1 L20, P9 L8-10, P5 L4-7

Comment 3: Most serious objection: The paper draws conclusions based on subjective evaluations of results. There is no reference to statistical significance or confidence intervals, and very few references to thresholds for 'good simulation', 'small error' etc. In section 2.4.1 on page 5, it is stated that an ensemble of 100 calibrated parameter sets were available for each objective function, allowing analysis of parameter uncertainty. I have not investigated the genetic algorithm used for calibration, but could the variability within this ensemble be used to assess which differences extend beyond mere noise?
Reply 3: As suggested by the reviewer, we conducted a parameter uncertainty analysis to estimate the significance of the results (especially the significance of the median value). As null hypothesis we assumed that the probability for over- and underestimation of a SFC is 50%. The two-sided binomial test was used to assess the significance of the results at a confidence level of 0.95. The results are presented in Fig. 9 and section 3.3.
Thresholds for good and poor SFC estimates which were originally introduced in chapter 3.3 were discarded in the adapted manuscript (for a discussion on thresholds please see comment 10 in the section ‘detailed and specific comments’). In the results chapters 3.1 and 3.2 we illustrate the difference in model performance related to the five objective functions relative to each other. Since the focus lies on which objective function gives the best SFC estimates, a clear threshold for good or poor simulations is not needed in these two chapters.
Modification: P7 L6-8 (methods), P9 L21–30 (results), P31-32 Fig. 9.

Detailed and specific comments

Comment 1: Page 3, lines 22-24. This sentence is unclear and possibly erroneous. What is meant?
Reply 1: We adapted the sentence to make it clearer. The content of the sentence was correct, meaning that in the cited study calibrations with R_{eff} outperformed calibrations with a multi-objective function (consisting of different combinations of R_{eff}, R_{eff} of log-transformed flow, volume error and Spearman rank correlation; please see P2 L33-35 and P3 L24-27) with respect to the investigated SFCs. For this reason we selected R_{eff} as benchmark for our study.
Modification: P3 L24-26

Comment 2: Page 3 line 29 and onwards. Please specify also what is **not** used. For instance in (1) make it clear that this SFC is used alone, in (2) whether or not the multi-SFC vector

includes the SFC being evaluated, and (3) if the TCCs are kept in or excluded when the SFCs are included.
Reply 2: Following the research questions we specified in brackets how these objective functions could be defined.
Modification: P4 L4-9

Comment 3: Page 4 line 13. The reference to NAVD 88 is unnecessary.
Reply 3: We removed the reference to NAVD 88.
Modification: P4 L23

Comment 4: Page 6 line 10: Not yet knowing that the SFCs are normalised to a common scale, the reader may wonder how R_{eff} and an arbitrary SFC can be equally weighted. Just put in a 'normalised (see below)'.
Reply 4: Thank you for this suggestion, we added the term ‘normalized’ to the mentioned sentence.
Modification: P6 L24

Comment 5: Page 6 line 20: Can you specify how small error are required for a SFC to be robust, and how 'relatively good simulations for other SFC' are required for being informative? These limits have the impact of restricting which SFCs enter the Multi alternative.
Reply 5: We agree that a clear threshold value for robustness and information value helps the reader to understand why a particular SFC was selected for I_{Multi}. Instead of using a fix threshold value (e.g. normalized model error smaller than 25%), we decided to evaluate the robustness and information value of a SFCs relative to other SFCs (e.g. a potential SFC for I_{Multi} should rank among the SFCs with highest robustness). This enables a more flexible selection of potential SFCs for I_{Multi} with focus on acceptable trade off solutions for all SFCs. We adapted the text accordingly to make our selection procedure clearer.
Modification: P6 L30-32, P7 L1-2

Comment 6: Sections 3.1.1 and 3.1.2: Would these come more naturally in opposite order? I perceive the I_{Single} experiment as the simplest and most obvious, the one-to-any SFC as more involved.
Reply 6: We changed to order of these two catchments as suggested. We therefore also had to change the order of Fig. 3 and Fig. 4 and adapt the figure references in the text accordingly.
Modification: P7 section 3.1.1, P8 section 3.1.2, P25-26 (Figures)

Comment 7: Would it be informative to summarise in a table for all SFCs what is illustrated in fig 3 for TA1?
Reply 7: Fig. 3 shows TA1 estimates for calibrations with each of the 13 I_{Single} objective functions and illustrates that calibrating on a single SFCs often leads to poor estimates of another SFC. This result is not very surprising. Since TA1 can also be considered as a representative example for the other 12 SFCs, we argue that it is not necessary or of major interest to show the results for each of the 13 SFCs.

Comment 8: Page 8 line 6. What is required to deserve a 'well simulated' mark? Is there an a priori defined threshold?
Reply 8: The term ‘well simulated’ was not meant in absolute, but rather in relative terms. We adapted the sentence accordingly. Together with the suggested text modification in the methods part (see comment 5) it should become clear that we refer to relative performance of SFCs when evaluating robustness and information value. For the discussion of using well defined values for classifying good and poor SFC estimates please see answer to comment 10.
Modification: P8 L20-21

Comment 9: Page 8 line 8. Se discussion above about SFC estimates being high or low.
Reply 9: Please see answer for comment 10

Comment 10: Section 3.3, page 8: This is one of the weaker parts. See the above point on subjective conclusions. The categorisation in lines 19 and 20 are well defined, but then not used for anything. The lowest error class collects everything from perfect match to 10% error, capturing the result for all the mean-flow SFCs, 3 out of 4 low-flow SFCs and 4 out of 5 high-flow SFCs. Still, this paragraph states that all high-flow SFCs and three of four mid-flow SFCs are under-estimated, whereas all except one low-flow SFC were over-estimated' (lines 21-22). Such conclusions need to refer to at least a clearly stated threshold, but preferably to statistical significance.

The SFCs listed as having small vs medium absolute errors in line 23 does not correspond to a sorted grouping of the errors in fig 8. The FH6 error is characterised as 'small', but is larger than the DH16, MA41 and TA1 errors listed as medium, as well as the MA26 and FL2 errors not listed in these two lowest groups. Likewise it is difficult to see from figure 8 why MA26, DH13 and FH7 are identified in line 25 as having large error ranges, while MA41, TA1 or DH16 are not. One gets to suspect that the text is referring to another version of figure 8.

Reply 10: We reorganized section 3.3. First, we discarded the categorization of the error magnitudes because they were, as pointed out by the reviewer, to some degree subjective. This made the paragraph about error magnitudes shorter and more concise (P8 L2-10). Secondly, we extended the part of section 3.3 which is about the uncertainty of the over- and underestimation of SFCs (please see comment 3 of major comments).

Modification: P8 L2-10

Comment 11: The statement in lines 15-17 suggests an identification between goodness of fit and process representation, which this paper neither investigates nor justifies.

Reply 11: We agree that we don't explicitly analyze the link between a particular objective function and process representation by e.g. analyzing hydrographs. However, we argue that this link is implicitly present in all model calibration studies. Objective functions are usually chosen to steer calibration in the direction of certain hydrograph aspects. In our case calibration with a particular SFC aims at simulating the hydrological process represented by that SFC. Similarly, calibration with R_{eff} or I_{Multi} aims at simulating many different hydrograph aspects (and thus hydrological processes) simultaneously.

Comment 12: Figure 9: The term 'absolute normalised SFC errors' are used. The figure seems to display signed errors.

Reply 12: The figure indeed displays the signed errors. We therefore deleted the word "absolute" from the figure capture.

Modification: P32 (Fig. 9)

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. 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 optimization of statistical performance metrics (e.g. Nash-Sutcliffe model 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 model efficiency, single SFCs or combinations thereof. 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 ecologically relevant SFCs representing major flow regime components and different flow conditions. While the model generally tended to underestimate the tested SFCs related to mean and high-flow conditions, SFCs related to low flow 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 model 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 simulations usually focus on either representing the general shape of the hydrograph or on accurately simulating specific streamflow characteristics relevant to a respective application. However, the extraction of streamflow characteristics (SFCs) from a simulated time series may produce poor estimates when these characteristics were not included in model calibration. 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 (Cartwright et al., 2017).

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 statistical models such as multiple linear regressions often provide greater accuracy (Murphy et al., 2013), runoff models provide opportunities for also evaluating climate or land-use change scenarios.

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. 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 model 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 based on only the model efficiency with calibrations based on the combination of multiple objectives such as model efficiency, model efficiency of log-transformed 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 model efficiency (Nash and Sutcliffe, 1970). However, runoff models calibrated using FDC have to be constrained by additional SFCs if one is interested in 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. In a recent study, Kiesel et al. (2017) compared estimates of ecologically relevant SFCs simulated from model calibrations using different objective functions including SFCs and the Kling-Gupta efficiency (Gupta et al., 2009). They found that including all SFCs of interest into the model calibration resulted in better SFC estimates than a calibration using the Kling-Gupta efficiency. 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 the model efficiency (R_{eff}) outperformed multi-objective model calibrations (different combinations of R_{eff}, of log-transformed flow, volume error and Spearman rank correlation) for the investigated SFCs. ~~and i~~It was furthermore 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 a model efficiency based calibration approach. 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~~ as the model objective function? (Objective function is the SFC of interest)

(2) How well is a single SFC simulated when the model objective function contains one or multiple other SFCs? (Objective function can include the SFC of interest, but generally contains one or multiple other SFCs)

(3) How does the accuracy of estimated SFCs vary between traditional calibration approaches and those where the SFCs of interest are included? (Objective functions are different combinations of SFC(s) and the model efficiency)

Throughout this study, we refer to traditional and ‘SFC based’ objective functions. Traditional objective functions were defined as efficiency criteria based on statistical performance metrics computed from (transformed) model residuals (e.g R_{eff} or volume error). In contrast, ‘SFC based’ objective functions evaluate specific hydrograph aspects, such as event frequencies, timing or variability of runoff, that are of ecological relevance in our study region.

2 Materials and 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 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. Please note that some of the tested SFCs (DH13, ML20, MA26, DH16 and FL2) are defined as scaled with the median, mean or total runoff. The scaling leads to SFC values that are dependent on flow magnitudes. The magnitude of the simulation error for DH13, ML29, MA26, DH16 and FL2 is therefore dependent on runoff magnitudes, whereas the sign of the simulation error is not affected by the normalization.

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 0.6° 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 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 model efficiency (R_{eff}, Nash and Sutcliffe, 1970). The model 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 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 ($I_{\text{Single\_Reff}}$). 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 equally weighted normalized SFCs was defined (I_{Multi}, Table 2). This 'SFC based' efficiency measure was again combined with R_{eff} ($I_{\text{Multi\_Reff}}$). For the resulting combined objective function, the same weights 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 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 relatively small errors over the full range of catchments in both validation time periods compared to other SFCs. A SFC was regarded as being informative, when it also yielded relatively good simulations for other SFCs. Robustness and information value of a SFC was therefore assessed relative to other SFCs enabling acceptable trade off solutions for all SFCs with a minimum number of SFCs being potentially representative for (most of) the thirteen SFCs.

2.4.3 Evaluation of model performance

Model performance in calibration and validation was evaluated by means of normalized SFC 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 model efficiency of each objective function, validation period and catchment was selected as representative value for the model efficiency distribution. Simulation uncertainty stemming from the 100 parameter sets was assessed by a two-sided binomial test with the null hypothesis that the probability for over- and underestimation of a SFC is equal to 50%.

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 10th and 90th 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 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 was used as an objective function (I_{Single}) for model calibration (the 13 absolute nSFCs varied between 0.000 and 0.005 for calibrations with I_{Single}). Both I_{Single} and the combined objective function $I_{\mathrm{Single\_Reff}}$ clearly outperformed model calibrations based on R_{eff} with regard to the estimation of SFCs (Fig. ~~4a~~3a). However, calibration with I_{Single} yielded poor model performances when evaluated in terms of R_{eff} whereas R_{eff} efficiencies of calibrations with either $I_{\mathrm{Single\_Reff}}$ or R_{eff} were comparable (Fig. ~~4a~~3a).

Validation results (Fig. ~~4b~~3b) exhibited a similar pattern in model performance as the calibration results. The median absolute normalized error of the 13 SFCs was relatively low for model runs based on the objective functions I_{Single} and $I_{\text{Single\_Reff}}$ compared to model calibration with R_{eff}. The comparable SFC estimation accuracy of I_{Single} and $I_{\text{Single\_Reff}}$ 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).

3.1.2 How informative is a SFC for estimating any SFC?

The calibrations for all 13 versions of I_{Single} and $I_{\text{Single\_Reff}}$ resulted in total in 26 different runoff simulations that were evaluated by calculating the normalized SFC error for the calibration and validation periods. The SFC TA1 (stability of runoff; Fig. ~~3a~~ 4a and b) was selected as a representative example to illustrate that the use of SFCs as a single objective function (I_{Single}) generally resulted in poor SFC estimates for those SFCs not included in I_{Single} in both model calibration and validation when compared to model calibrations with $I_{\text{Single\_Reff}}$ or R_{eff}. Estimation accuracies from calibrations with $I_{\text{Single\_Reff}}$ and R_{eff} were often of comparable magnitude. Error magnitudes from the three described objective function types (I_{Single}, $I_{\text{Single\_Reff}}$ and R_{eff}) could vary considerably between time periods (illustrated by triangles and circles respectively in Fig. ~~3a~~ 4a and b).

3.2 The use of multiple SFCs for model calibration

Figure 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. 6a) were RA7, ML20, FH6, E85 and MA41. All these five SFCs could be used for the objective function I_{Multi}, however E85 (lowest 15% of daily runoff) was discarded as potential SFC for I_{Multi} because of its redundant information with ML20 (base flow). The information value of the remaining four SFCs for each of the 13 SFCs is presented in Fig. 6b. All 13 SFCs were relatively well simulated by model calibrations with I_{Single} of either RA7, ML20, FH6, or MA41 (coloured circles in Fig. 6b) compared to calibrations with other SFCs.

Median estimates of the 13 SFCs in the calibration period were slightly lower when the model was calibrated with I_{Multi} rather than $I_{\text{Multi\_Reff}}$. Both of these objective functions led to better model performance for SFCs than calibrating with R_{eff} alone (Fig. ~~4a~~3a). Model performance for the validation period with $I_{\text{Multi\_Reff}}$ had lower median error for SFCs than the error associated with using I_{Multi} as objective function (Fig. ~~4b~~3b). The comparison of I_{Multi} and $I_{\text{Multi\_Reff}}$ for all SFCs separately (Fig. 7a) revealed that for most SFCs both objective functions resulted in similar estimates. $I_{\text{Single\_Reff}}$ was better for estimating SFCs than $I_{\text{Multi\_Reff}}$, especially for SFCs not included in the $I_{\text{Multi\_Reff}}$ objective function (Fig. 7b). Comparing simulations from $I_{\text{Multi\_Reff}}$ and R_{eff} revealed a smaller median error of the SFCs when calibrating with $I_{\text{Multi\_Reff}}$ (Fig. ~~4b~~ 3b and 7c). Yet, for most SFCs not explicitly incorporated into the objective function $I_{\text{Multi\_Reff}}$, the objective function R_{eff} performed equally well or slightly better than $I_{\text{Multi\_Reff}}$ (Fig. 7c).

3.3 Estimation accuracy for SFCs

Figure 8 provides an overview (median of all 25 catchments) 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.~~ Error magnitudes ranged between -25% and 25% for the majority of SFCs. Considerably higher estimation accuracy was achieved for ML20 (-5% to 2%), whereas estimation accuracies were lowest for MH10 and TL1 with error magnitudes up to 40% and 77% percent, respectively. For some SFCs (e.g. MA26 and TL1) the error tended to be higher in one of the two modelling time periods whereas for other SFCs (e.g. RA7 and MH10) 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).

The median error (illustrated by stars in Fig. 8) was used for the evaluation of the under- or overestimation of SFCs. Among the tested SFCs, ~~A~~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. 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).~~ This overall pattern was less evident when evaluating each objective function and time period separately (Fig. 8 and Fig. 9).

~~Normalized errors for the high-flow conditions,~~ The SFCs DH16 and MH10 ~~, for all 25 catchments and for both modelling time periods~~ indicate two typically observed deviations in the overall pattern. ~~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 (median values in Fig. 9a). However, for objective functions or modelling time periods with a low magnitude in the median bias, the underestimation of the SFC was not statistically significant. Even in case of a median pointing to statistically significant underestimation, there might be a substantial number of catchments ~~that show overestimation of~~ for which DH16 was overestimated. A second commonly observed phenomenon is shown by the SFC MH10 (Fig. 9b). While MH10 had mostly small but statistically significant 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. 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. 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. Considering that SFCs not incorporated in the objective function I_{Multi} 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 when aiming at estimating a suite of SFCs. This is surprising because the SFCs selected for I_{Multi} or $I_{\mathrm{Multi\_Reff}}$ 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 calibrating a runoff model for estimating many different SFCs from one single hydrograph 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. 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).

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 Kiesel et al. (2017) 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 $I_{\mathrm{Multi\_Reff}}$ 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 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 $I_{\mathrm{Multi\_Reff}}$) are less accurate than when they are estimated from hydrographs based on targeted model calibrations (e.g. I_{Single} or $I_{\mathrm{Single\_Reff}}$). This fact has substantial implications for the later application of simulated SFCs in decision-support systems for integrated resource management. 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; Cartwright et al., 2017).

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, 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 model 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 co-authors. Writing of the paper was led by Sandra Pool with contribution of all co-authors.

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. We ~~would like to~~ thank the reviewers Björn Guse, ~~and~~ Oddbjørn Bruland and Sjur Kolberg for their constructive and detailed comments that helped to improve the quality of our manuscript.

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.

Cartwright, J., Caldwell, C., Nebiker, S., and Knight, R.: Putting flow–ecology relationships into practice: A decision-support system to assess fish community response to water-management scenarios, Water, 9, 196, doi:10.3390/w9030196, 2017.

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.

Gupta, H. V., Kling, H., Yilmaz, K. K. and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377, 80-91, doi:10.1016/j.jhydrol.2009.08.003, 2009.

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., 33, 233-248, 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.

Kiesel, J., Guse, B., Pfannerstill, M., Kakouei, K., Jähnig, S. C. and Fohrer N.: Improving hydrological model optimization for riverine species, Ecol. Indic., 80, 376-385, doi:10.1016/j.ecolind.2017.04.032, 2017.

Knight, R. R., ~~Brian~~ Gregory, M. B., 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.

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	Further explanation	Flow condition	Unit
Magnitude				
Mean annual runoff	MA41	Mean annual daily runoff	mean-flow	[mm d^{-1}]
Maximum October runoff	MH10	Mean of October runoff maxima for each year	high-flow	[mm d^{-1}]
Lowest 15% of daily runoff	E85	Daily mean runoff that is exceeded 85% of the time for the period of record	low-flow	[mm d^{-1}]
Rate of runoff recession	RA7	Median change in log of runoff 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 runoff 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 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 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 runoff over the period of record divided by the mean runoff 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 25th percentile) divided by the mean low-flow pulse counts multiplied by 100	low-flow	[%]

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
Model efficiency	R_{eff}	$1-\frac{\sum(Q_{\mathrm{obs}}-Q_{\mathrm{sim}})^2}{\sum(Q_{\mathrm{obs}}-\overline{Q_{\mathrm{obs}}})^2}$	1
Efficiency for each individual SFC[1]	I_{Single}	$1-\frac{\lvert I_{\mathrm{obs}}-I_{\mathrm{sim}}\rvert}{I_{\mathrm{obs}}}$	1
SFC and model efficiency	$I_{\mathrm{Single\_Reff}}$	$0.5\ (I_{\mathrm{Single}}+R_{\mathrm{eff}})$	1
Efficiency for the selected SFCs[2]	I_{Multi}	$\frac{1}{n}\ (I_{\mathrm{Single}_1}+\ldots+I_{\mathrm{Single\_n}})$	1
SFCs and model efficiency	$I_{\mathrm{Multi\_Reff}}$	$\frac{n-1}{n}\ I_{\mathrm{Multi}}+\frac{1}{n}\ R_{\mathrm{eff}}$	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
Model efficiency	R_{eff}	$1-\frac{\sum(Q_{\mathrm{obs}}-Q_{\mathrm{sim}})^2}{\sum(Q_{\mathrm{obs}}-\overline{Q_{\mathrm{obs}}})^2}$	1
Mean absolute relative error[1]	MARE	$1-\frac{1}{n}\sum\frac{\lvert Q_{\mathrm{obs}}-Q_{\mathrm{sim}}\rvert}{Q_{\mathrm{obs}}}$	1
Normalized SFC error[2]	nSFC	$\frac{I_{\mathrm{obs}}-I_{\mathrm{sim}}}{R_{\mathrm{obs}}}$	0

[1] n is the number of days.

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

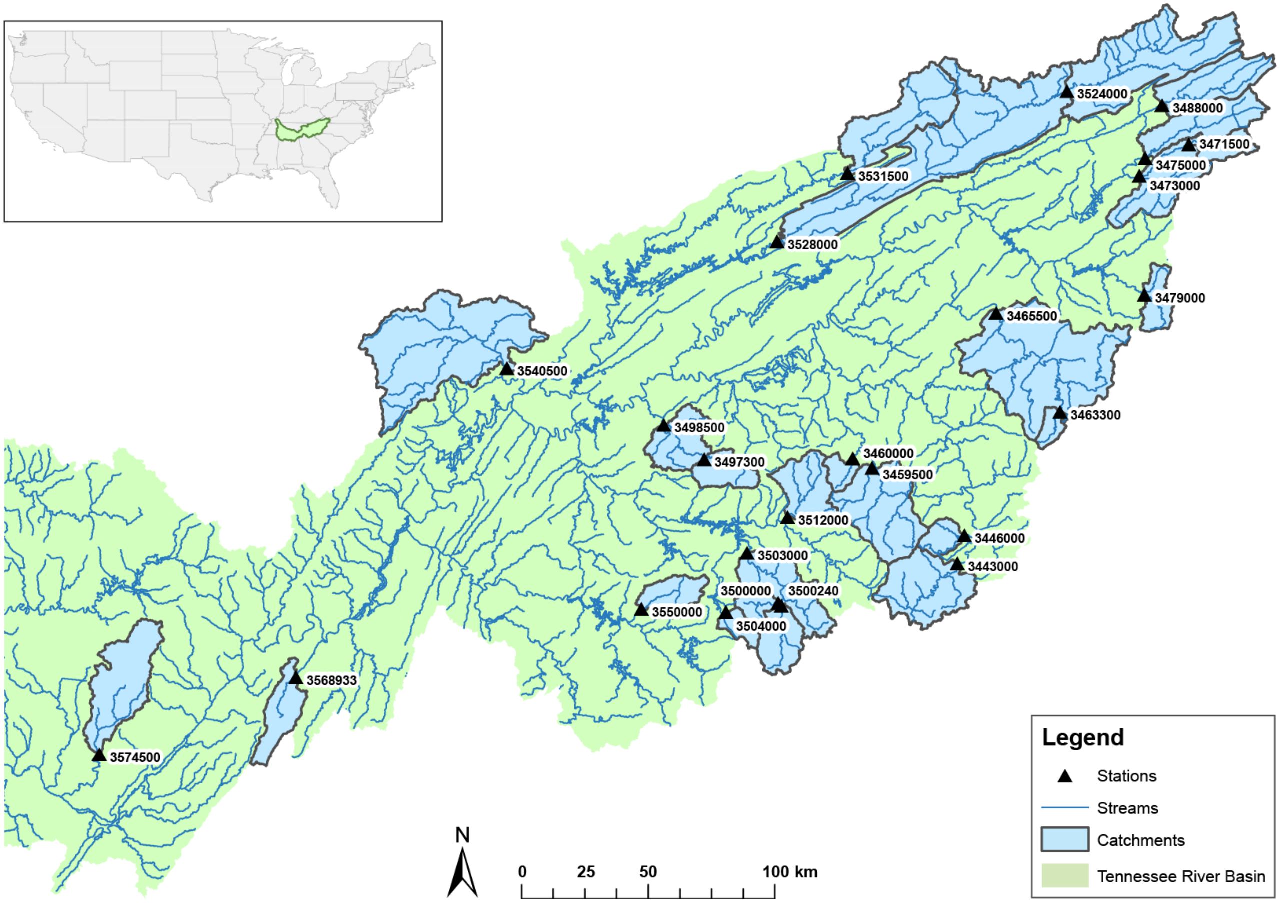


Figure 1. Location of the 25 study catchments in the Tennessee River basin (Table 1 in Vis et al. (2015) for more information).

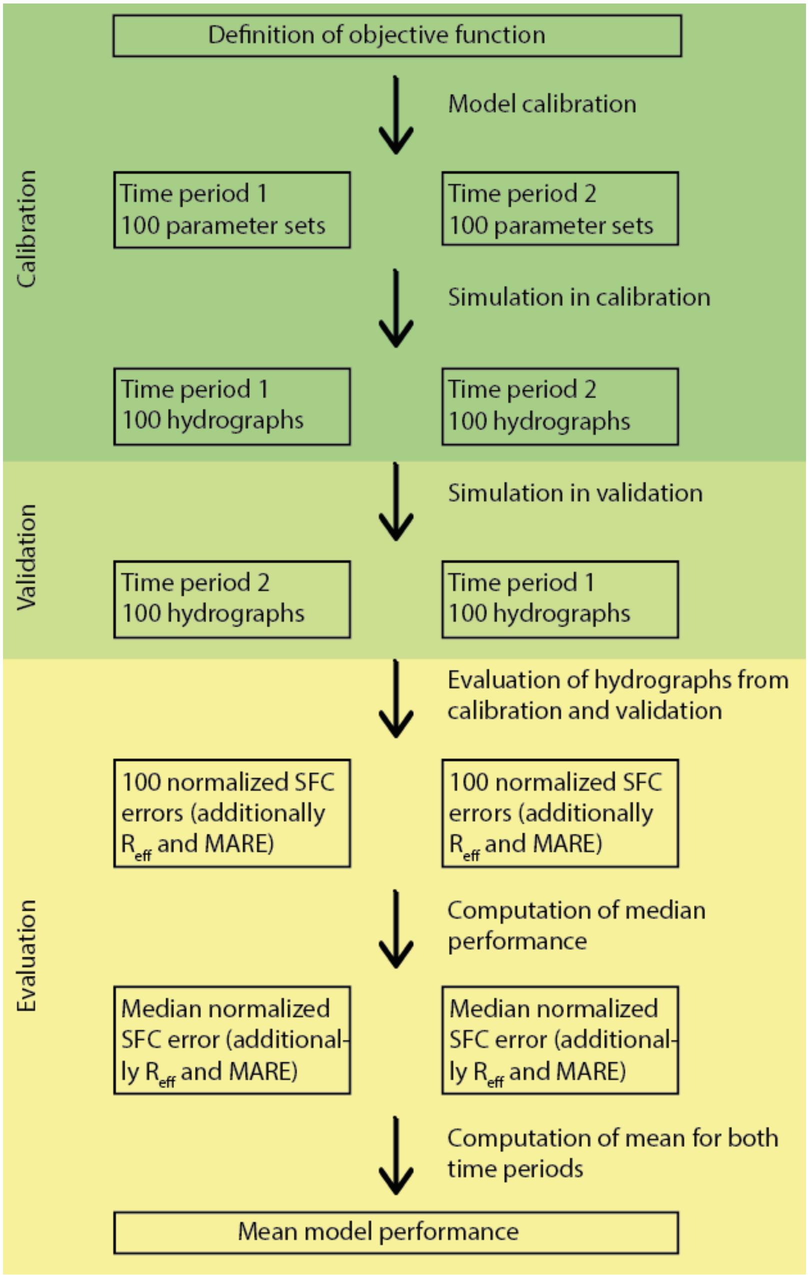

Definition of objective function
Model calibration
Time period 1
100 parameter sets
Time period 2
100 parameter sets
Simulation in calibration
Time period 1
100 hydrographs
Time period 2
100 hydrographs
Simulation in validation
Time period 2
100 hydrographs
Time period 1
100 hydrographs
Evaluation of hydrographs from calibration and validation
100 normalized SFC errors (additionally R_{eff} and MARE)
100 normalized SFC errors (additionally R_{eff} and MARE)
Computation of median performance
Median normalized SFC error (additional-ly R_{eff} and MARE)
Median normalized SFC error (additional-ly R_{eff} and MARE)
Computation of mean for both time periods
Mean model performance
Calibration
Validation
Evaluation

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}.

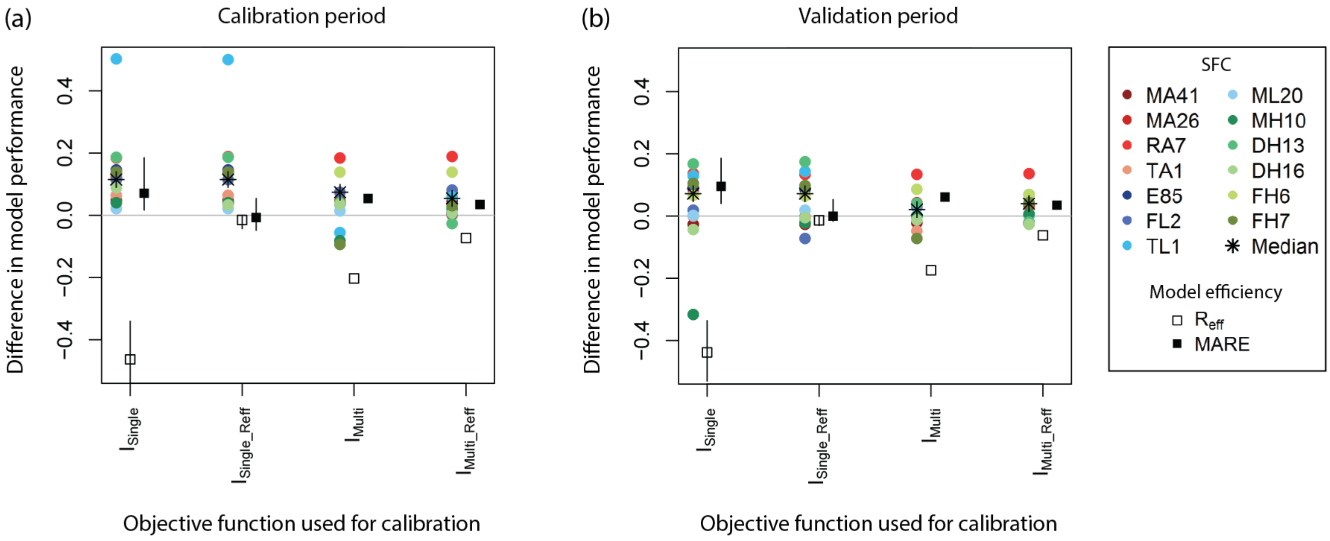


Figure ~~4~~3. Model performance in a) calibration and b) validation in terms of absolute normalized SFC errors (nSFC), R_{eff} and MARE depending on the objective function used in calibration. Model performance is shown as the difference between a model calibration with R_{eff} and model calibrations with I_{Single}, $I_{\mathrm{Single\_Reff}}$, I_{Multi} or $I_{\mathrm{Multi\_Reff}}$ (positive values indicate that model calibration with I_{Single}, $I_{\mathrm{Single\_Reff}}$, I_{Multi} or $I_{\mathrm{Multi\_Reff}}$ resulted in better model performance than model calibration with R_{eff}; negative values indicate that model calibration with I_{Single}, $I_{\mathrm{Single\_Reff}}$, I_{Multi} or $I_{\mathrm{Multi\_Reff}}$ 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.

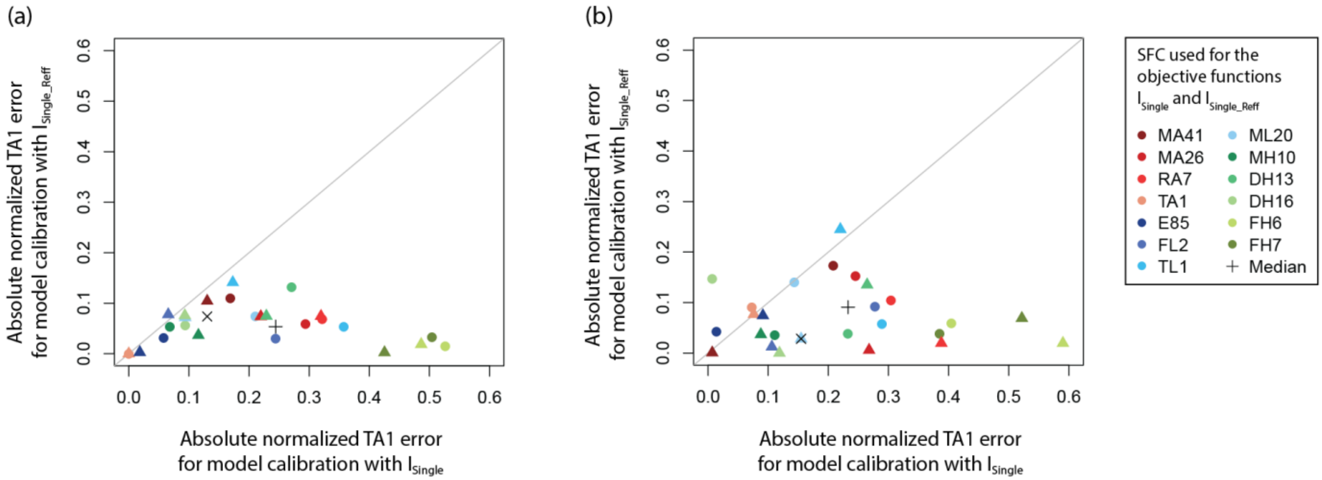


Figure ~~3~~4. Absolute normalized TA1 error (nSFC) in a) calibration and b) validation calculated from model calibrations with the objective functions I_{Single} and $I_{\text{Single\_Reff}}$. Absolute normalized SFC errors correspond to the median of the 25 catchments and are shown separately for both 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. (Absolute normalized TA1 error for model calibrations with the objective function R_{eff} was 0.08 (period 1) and 0.05 (period 2) in calibration and 0.002 (period 1) and 0.15 (period 2) in validation.)

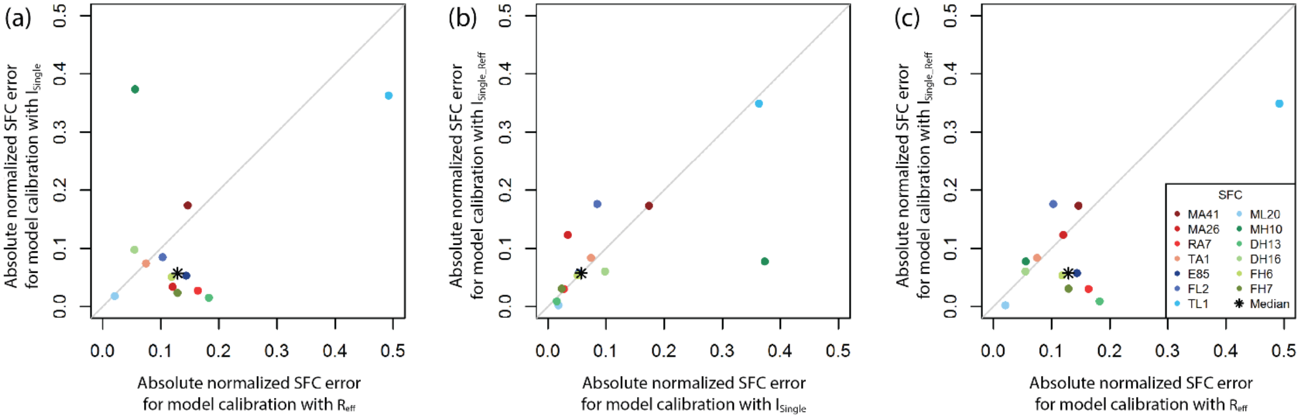


Figure 5. Comparison of absolute normalized SFC errors (nSFC) in validation calculated from model calibrations with the objective functions R_{eff}, I_{Single} and $I_{\mathrm{Single\_Reff}}$. Absolute normalized SFC errors correspond to the median of the 25 catchments and the mean of both modelling time periods.

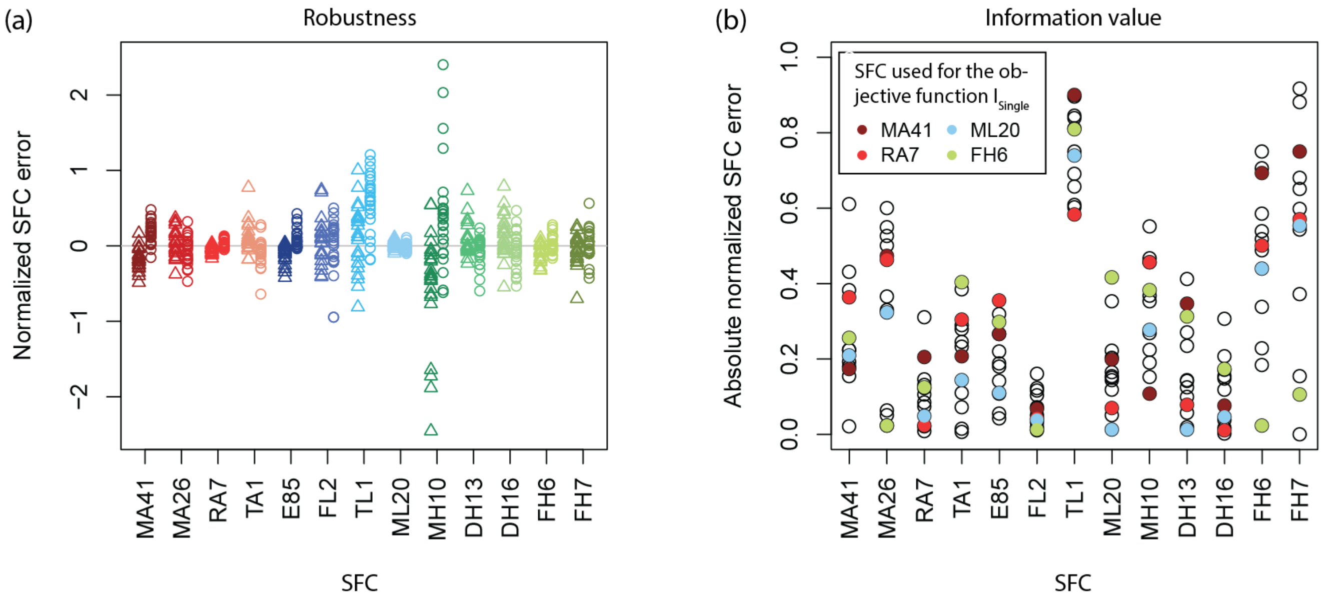


Figure 6. a) Robustness: normalized SFC errors (nSFC) in validation calculated from model calibrations with the objective function I_{Single} for the respective SFC. Values are shown for all 25 catchments and both modelling time periods (triangles for period 1 (1984 - 1996) and circles for period 2 (1997 - 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 open circle represents one of the 13 SFC used for I_{Single}. The coloured circles refer to the final selection of SFCs for the objective function I_{Multi}.

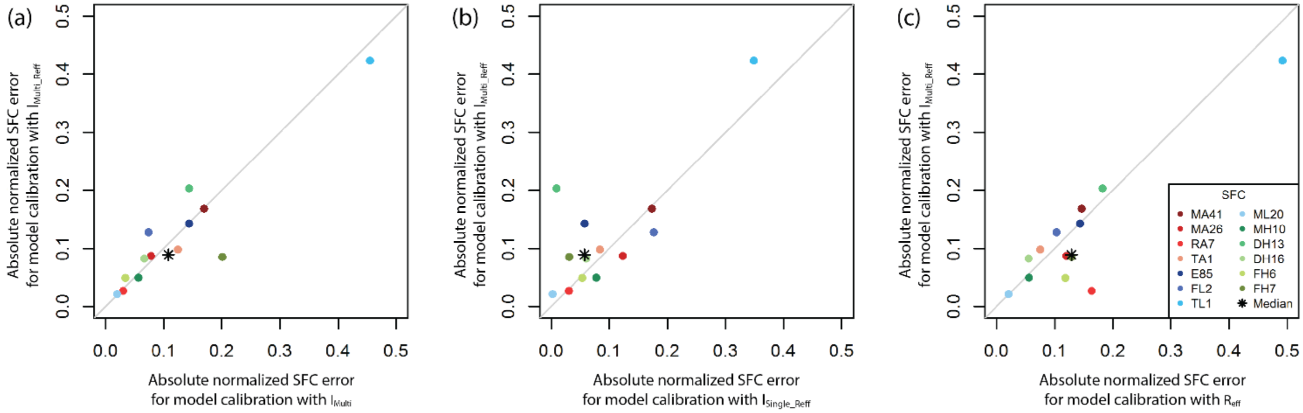


Figure 7. Comparison of absolute normalized SFC errors (nSFC) in validation calculated from model calibrations with the objective functions R_{eff}, $I_{\mathrm{Single\_Reff}}$, I_{Multi} and $I_{\mathrm{Multi\_Reff}}$. Absolute normalized SFC errors correspond to the median of the 25 catchments and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

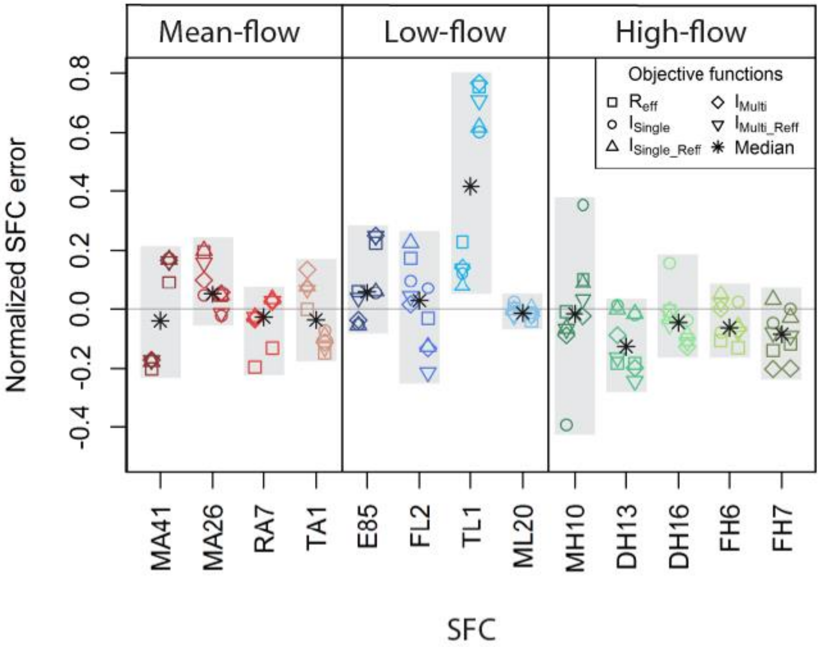
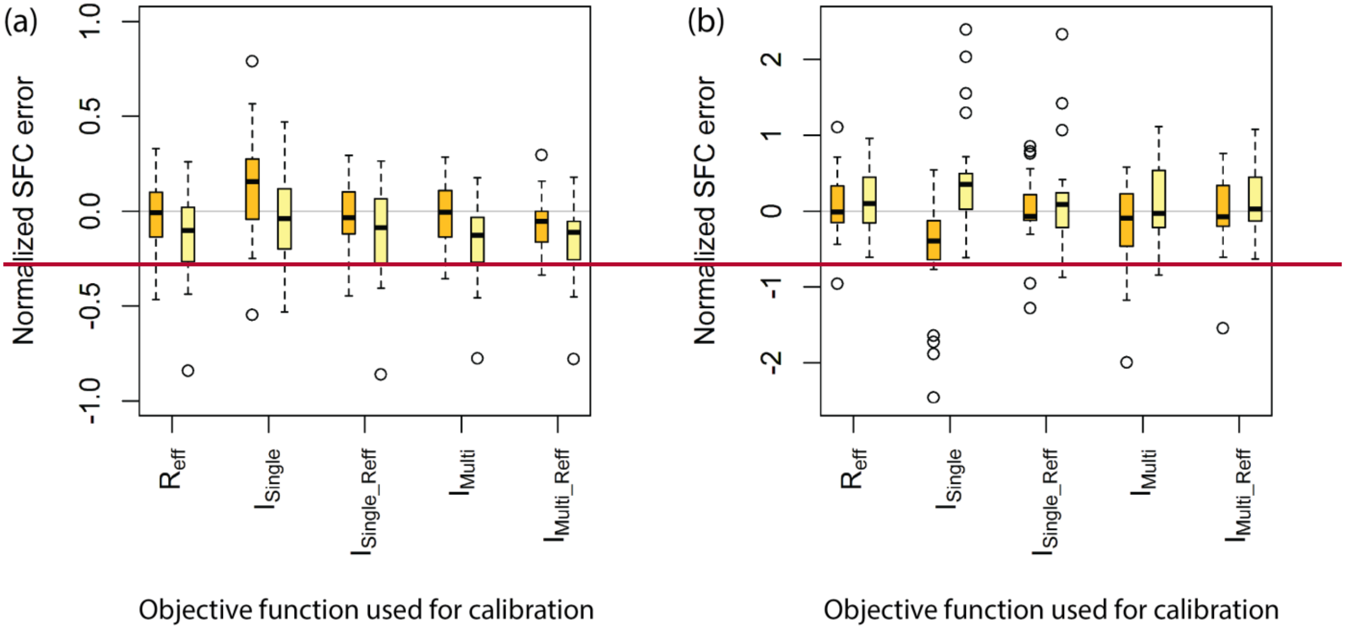


Figure 8. Normalized SFC errors (nSFC) in validation depending on the objective function used in calibration. Model performance values correspond to the median of the 25 catchments and are shown for both modelling time periods (period 1 (1984 - 1996) on the left side and period 2 (1997 - 2009) on the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


(a)
Normalized SFC error
1.0
0.5
0.0
-0.5
-1.0
R_{eff}
I_{Single}
I_{Single_Reff}
I_{Multi}
I_{Multi_Reff}
Objective function used for calibration
(b)
Normalized SFC error
2
1
0
-1
-2
R_{eff}
I_{Single}
I_{Single_Reff}
I_{Multi}
I_{Multi_Reff}
Objective function used for calibration

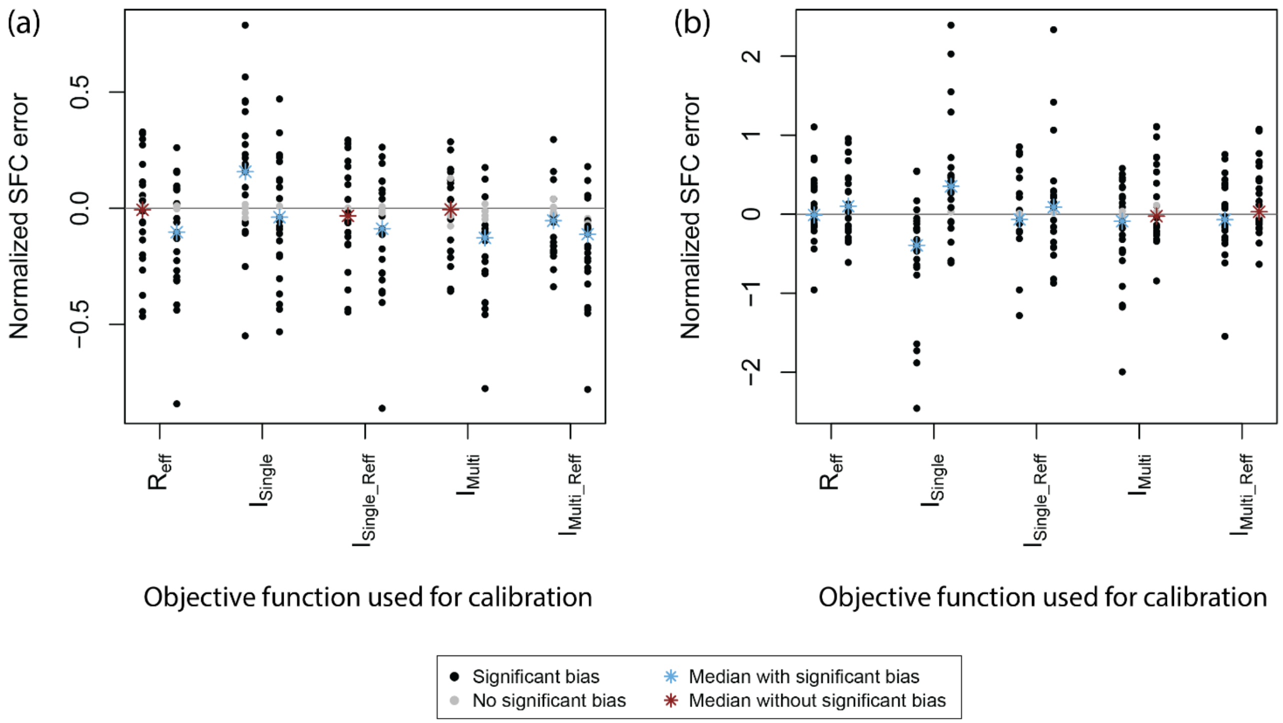


Figure 9. a) Normalized DH16 errors (nSFC) and b) normalized MH10 errors (nSFC) in validation depending on the objective function used in calibration. ~~Absolute n~~Normalized SFC errors are shown for all 25 catchments and for both modelling time periods (period 1 (1984 - 1996) ~~in orange~~ on the left side and period 2 (1997 - 2009) ~~in yellow~~ on the right side). Colours indicate the significance of the results assessed by a two-sided binomial test at a confidence level of 0.95. Note the difference in the y-axis.