Journal of Hydrology 554 (2017) 613–622

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Prediction of hydrographs and flow-duration curves in almost ungauged catchments: Which runoff measurements are most informative for model calibration?

Sandra Pool [a,*], Daniel Viviroli [a], Jan Seibert [a,b]

[a] Department of Geography, University of Zurich, Zurich, Switzerland
[b] Department of Earth Sciences, Uppsala University, Uppsala, Sweden

ARTICLE INFO

Article history:
Received 3 April 2017
Received in revised form 18 July 2017
Accepted 19 September 2017
Available online 20 September 2017
This manuscript was handled by K. Georgakakos, Editor-in-Chief, with the assistance of Yasuto Tachikawa, Associate Editor

Keywords:
Runoff modelling
Hydrograph prediction
Flow-duration curve prediction
Ungauged catchment
Sampling strategy
Value of data

ABSTRACT

Applications of runoff models usually rely on long and continuous runoff time series for model calibration. However, many catchments around the world are ungauged and estimating runoff for these catchments is challenging. One approach is to perform a few runoff measurements in a previously fully ungauged catchment and to constrain a runoff model by these measurements. In this study we investigated the value of such individual runoff measurements when taken at strategic points in time for applying a bucket-type runoff model (HBV) in ungauged catchments. Based on the assumption that a limited number of runoff measurements can be taken, we sought the optimal sampling strategy (i.e. when to measure the streamflow) to obtain the most informative data for constraining the runoff model. We used twenty gauged catchments across the eastern US, made the assumption that these catchments were ungauged, and applied different runoff sampling strategies. All tested strategies consisted of twelve runoff measurements within one year and ranged from simply using monthly flow maxima to a more complex selection of observation times. In each case the twelve runoff measurements were used to select 100 best parameter sets using a Monte Carlo calibration approach. Runoff simulations using these 'informed' parameter sets were then evaluated for an independent validation period in terms of the Nash-Sutcliffe efficiency of the hydrograph and the mean absolute relative error of the flow-duration curve. Model performance measures were normalized by relating them to an upper and a lower benchmark representing a well-informed and an uninformed model calibration. The hydrographs were best simulated with strategies including high runoff magnitudes as opposed to the flow-duration curves that were generally better estimated with strategies that captured low and mean flows. The choice of a sampling strategy covering the full range of runoff magnitudes enabled hydrograph and flow-duration curve simulations close to a well-informed model calibration. The differences among such strategies covering the full range of runoff magnitudes were small indicating that the exact choice of a strategy might be less crucial. Our study corroborates the information value of a small number of strategically selected runoff measurements for simulating runoff with a bucket-type runoff model in almost ungauged catchments.

1. Introduction

Sustainable management of water resources and mitigation of natural hazards in ungauged catchments strongly rely on accurate and reliable runoff estimates often predicted by rainfall-runoff models (Sivapalan et al., 2003). Runoff models used in hydrology all consist of parameters representing different catchment characteristics. The effective values of these parameters cannot be measured directly, because of their conceptual meaning or incommensurability issues. As a consequence, parameter values need to be defined or adapted in a calibration process by comparing observed and simulated catchment runoff response (Beven, 2012). After a decade of research on prediction of runoff in ungauged basins (PUB), it still remains a considerable challenge to calibrate runoff models for data scarce catchments (Hrachowitz et al., 2013).

A variety of approaches have been developed to estimate model parameters for ungauged catchments. For example, regionalization methods were proposed that either estimate individual parameter values from regressions relating model parameters to catchment characteristics or that transfer entire parameter sets from gauged donor catchments to the ungauged target catchment based on

[*] Corresponding author.
E-mail address: sandra.pool@geo.uzh.ch (S. Pool).

https://doi.org/10.1016/j.jhydrol.2017.09.037
0022-1694/© 2017 Elsevier B.V. All rights reserved.

proximity or similarity measures (see e.g. Parajka et al. (2013) for an extended discussion). Hydrograph predictions from regionalization could be improved given that a few runoff measurements were available to further constrain model parameters (Rojas-Serna et al., 2006; Drogue and Plasse, 2014; Viviroli and Seibert, 2015; Rojas-Serna et al., 2016). Some authors assumed that a short and intensive field campaign could be carried out in the catchment of interest to collect data for model calibration. They tested the value of combining runoff data and additional data such as groundwater dynamics (Freer et al., 2004; Juston et al., 2009; Seibert and McDonnell, 2013), soil moisture (Hughes et al., 2014) or hydrochemical tracers (Uhlenbrook and Sieber, 2005) for model calibration.

The PUB initiative determined the evaluation of the value of runoff data for model calibration as one of their main objectives (Sivapalan et al., 2003). This induced a series of studies exploring the minimum length of a runoff time series necessary to obtain robust model calibrations. First studies typically tested model sensitivity related to continuously measured runoff. Between two and eight years of runoff data were reported as minimum requirement for robust model parameterizations independent of the selected calibration period (Harlin, 1991; Yapo et al., 1996; Xia et al., 2004; Vrugt et al., 2006; Merz et al., 2009). While there is a general agreement that model performance tends to improve with an increased length of calibration data, much smaller data sets have been shown to be of comparable value as long continuous time series (McIntyre and Wheater, 2004; Perrin et al., 2007; Seibert and Beven, 2009, Singh and Bárdossy, 2012; Seibert and McDonnell, 2013; Melsen et al., 2014). Perrin et al. (2007) successfully calibrated a runoff model with 350 runoff measurements selected randomly from an almost forty year continuous runoff series. Seibert and Beven (2009) reported that approximately sixteen runoff measurements randomly picked within one hydrological year could already provide information for an acceptable model calibration. An alternative to randomly extracting measurements from a time series is the selection of runoff samples in a strategic manner. Seibert and Beven (2009) demonstrated that maximum flows or a combination of maximum and recession data contained more information than minimum or mean flows. Results from Seibert and McDonnell (2013) indicated that one fully gauged event or ten observations during different high flow situations had a similar information value as three months of continuously measured data. Extracting unusual events from a time series, Singh and Bárdossy (2012) achieved reliable model simulations with less than 10% of the data from a continuous time series. Moreover, event based sampling strategies resulted in better model performances than strategies with measurements at fixed time intervals (McIntyre and Wheater, 2004; Juston et al., 2009; Seibert and McDonnell, 2013). Model calibration with a limited number of runoff measurements performed best in relatively wet catchments (Perrin et al., 2007; Sun et al., 2017), which is a common observation in rainfall runoff modelling even when long continuous time series are available, or when runoff samples are selected during a wet period (Yapo et al., 1996; Vrugt et al., 2006; Kim and Kaluarachchi, 2009; Melsen et al., 2014; Correa et al., 2016). In addition, the consideration of hydrological variability and of hydrologically important processes was found to be essential for the calibration process and the resulting simulation uncertainty (Harlin, 1991; Vrugt et al., 2006; Konz and Seibert, 2010; Singh and Bárdossy, 2012).

The present study aimed at finding the most informative runoff measurements for calibrating a hydrologic model with a limited number of strategically selected runoff samples in order to accurately simulate the hydrograph and the flow-duration curve (FDC) in almost ungauged catchments. Based on data from twenty gauged catchments in the eastern US, which were treated as hypothetically poorly gauged catchments, we evaluated the following assumptions:

1) There is an optimal strategy to decide on when to measure runoff in an ungauged catchment to obtain the most informative data for constraining a runoff model.
2) The optimal strategy is generally valid, i.e., does not depend on the catchment or simulation evaluation criteria.
3) Runoff measurements chosen with an optimal sampling strategy are of comparable value as a long continuous runoff time series.

In our study we assume that measurements actually can be taken at these strategic points in time such as on the day with maximum flow during a month. In practice, this is obviously not possible as the runoff during a month is not known beforehand. However, our study gives an indication on how useful a certain strategy could be at best.

2. Data and methods

2.1. Study catchments and runoff model

This study was based on twenty catchments across the eastern US (Fig. 1). Catchment data was extracted from the freely available large scale dataset of Newman et al. (2015). The dataset with over 600 basins spread over the contiguous US includes catchments with only minimal human disturbances and complete thirty-year forcing and runoff data series. We selected twenty catchments that are similar in terms of wetness and precipitation seasonality, but different regarding the importance of snow related runoff processes. This small catchment sample can be considered as a relatively controlled subset of the large dataset with small hydroclimatic variation, but representing some of the most common runoff regime types in the US. The selected catchments (Table 1) vary in area from 148 to 2925 km^2 with steepest elevation gradients in or close to the Appalachian Mountains. Some catchments are to a large degree composed of wetlands and lakes account for up to 6% of the area of three of these catchments (C1, C2 and C20 in Table 1; Lehner and Döll, 2004). All catchments are humid and receive precipitation throughout the entire year. Snow processes dominate the runoff regime in northern latitudes where 10–28% of the annual precipitation falls as snow. The contribution of baseflow to runoff varies between the catchments from 23 to 69% indicating a large variation in runoff response characteristics.

Continuous daily runoff time series at the catchment outlets were simulated with a bucket-type runoff model, namely the HBV model (Hydrologiska Byråns Vattenbalansavdelning; Bergström, 1976; Lindström et al., 1997) in the version HBV-light (Seibert and Vis, 2012). The HBV model is forced with daily temperature and precipitation and monthly potential evaporation data. Hydrological processes are modelled with four model routines representing snow, soil water, groundwater and routing related processes. Snow accumulation and snowmelt are calculated in the snow routine using a degree-day method. Together with rainfall and potential evaporation, snowmelt is used to determine the actual evaporation and groundwater recharge in the soil routine. The groundwater routine consists of a shallow and a deep groundwater storage where the contribution of groundwater to peak runoff, intermediate runoff and baseflow is calculated. The routing routine transforms these three runoff components into the hydrograph at the catchment outlet by a triangular weighting function.

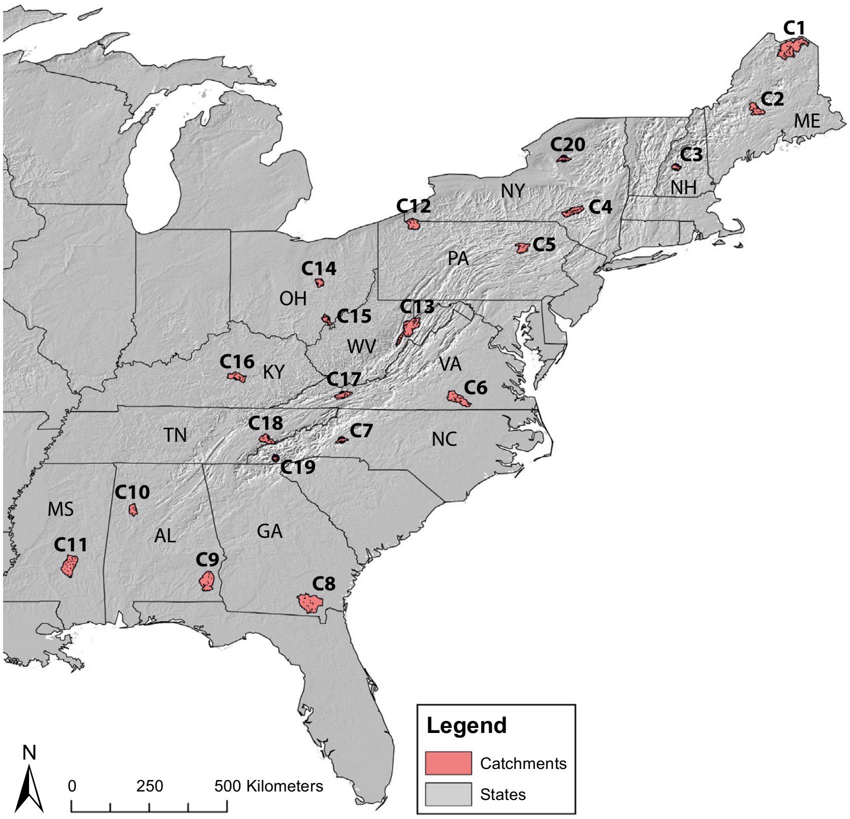


Fig. 1. Location of the twenty study catchments across the eastern US (catchment shapefiles from Newman et al. (2015); state boundaries and shaded relief from ESRI and U. S. Geological Survey (2011)).

Table 1
Information on the twenty study catchments. Snow [%]: percentage of annual precipitation falling as snow; precipitation seasonality: calculated according to Coopersmith et al. (2014), low seasonality for values ~<0.25; aridity index: ratio of sum of potential evaporation and sum of precipitation; runoff coefficient: ratio of runoff and sum of precipitation; baseflow [%]: percentage of runoff classified as baseflow, calculated based on the minimum runoff in fixed 5 day time intervals using the U.S. Geological Survey (2014) EflowStats R-package; wetland area [%]: percentage of catchment area covered by partial wetlands according to Lehner and Döll (2004).

ID	USGS station number and name	Area [km^2]	Mean elevation [m a.s.l.]	Snow [%]	Precipitation seasonality	Aridity index	Runoff coefficient	Baseflow [%]	Wetland area [%]
C1	01013500 Fish River near Fort Kent, ME	2260	379	27.6	0.17	0.63	0.54	68.9	92.2
C2	01031500 Piscataquis River near Dover-Foxcroft, ME	771	452	24.5	0.12	0.60	0.58	43.2	95.9
C3	01078000 Smith River near Bristol, NH	222	486	19.7	0.11	0.62	0.49	44.3	97.8
C4	01423000 West Branch Delaware River at Walton, NY	860	690	18.3	0.11	0.62	0.49	46.0	5.1
C5	01539000 Fishing Creek near Bloomsburg, PA	709	478	12.5	0.11	0.69	0.51	46.1	9.1
C6	02051500 Meherrin River near Lawrenceville, VA	1429	124	3.5	0.07	0.85	0.27	40.7	0.0
C7	02143000 Henry Fork near Henry River, NC	215	593	2.2	0.06	0.76	0.39	61.5	0.0
C8	02314500 Suwannee River at US 441 at Fargo, GA	2925	69	0.0	0.26	0.88	0.19	69.5	99.1
C9	02361000 Choctawhatchee River near Newton, AL	1776	127	0.0	0.16	0.82	0.31	52.5	0.0
C10	02464000 North River near Samantha, AL	577	157	0.9	0.12	0.70	0.37	29.6	0.0
C11	02472000 Leaf River near Collins, MS	1924	131	0.3	0.14	0.75	0.32	31.5	28.4
C12	03015500 Brokenstraw Creek at Youngsville, PA	831	486	16.3	0.14	0.63	0.54	40.2	21.4
C13	03069500 Cheat River near Parsons, WV	1869	984	16.4	0.11	0.61	0.60	36.2	21.6
C14	03144000 Wakatomika Creek near Frazeysburg, OH	362	308	7.4	0.13	0.84	0.36	36.6	0.0
C15	03159540 Shade River near Chester, OH	404	246	5.9	0.10	0.82	0.34	25.6	0.0
C16	03285000 Dix River near Danville, KY	823	349	3.5	0.10	0.77	0.40	23.0	0.0
C17	03488000 N F Holston River near Gate City, VA	572	976	6.8	0.11	0.81	0.38	46.1	0.0
C18	03498500 Little River near Maryville, TN	696	1141	2.9	0.11	0.64	0.41	51.8	0.0
C19	03500240 Cartoogechaye Creek near Franklin, NC	148	1121	2.4	0.09	0.55	0.45	68.1	0.0
C20	04256000 Independence River at Donnattsburg, NY	230	478	24.7	0.11	0.60	0.62	47.8	97.6

The HBV model allows runoff to be simulated in a semi-distributed way by disaggregating a catchment into elevation bands. We therefore split the catchments into elevation bands of 200 m using SRTM elevation data (Shuttle Radar Topography Mission; Jarvis et al., 2008). Temperature and precipitation data for each elevation band were interpolated with lapse rates of 0.6 °C per 100 m and 10% per 100 m, respectively. Potential evaporation was assumed to be uniform over all elevation bands and was calculated with the Priestley-Taylor equation.

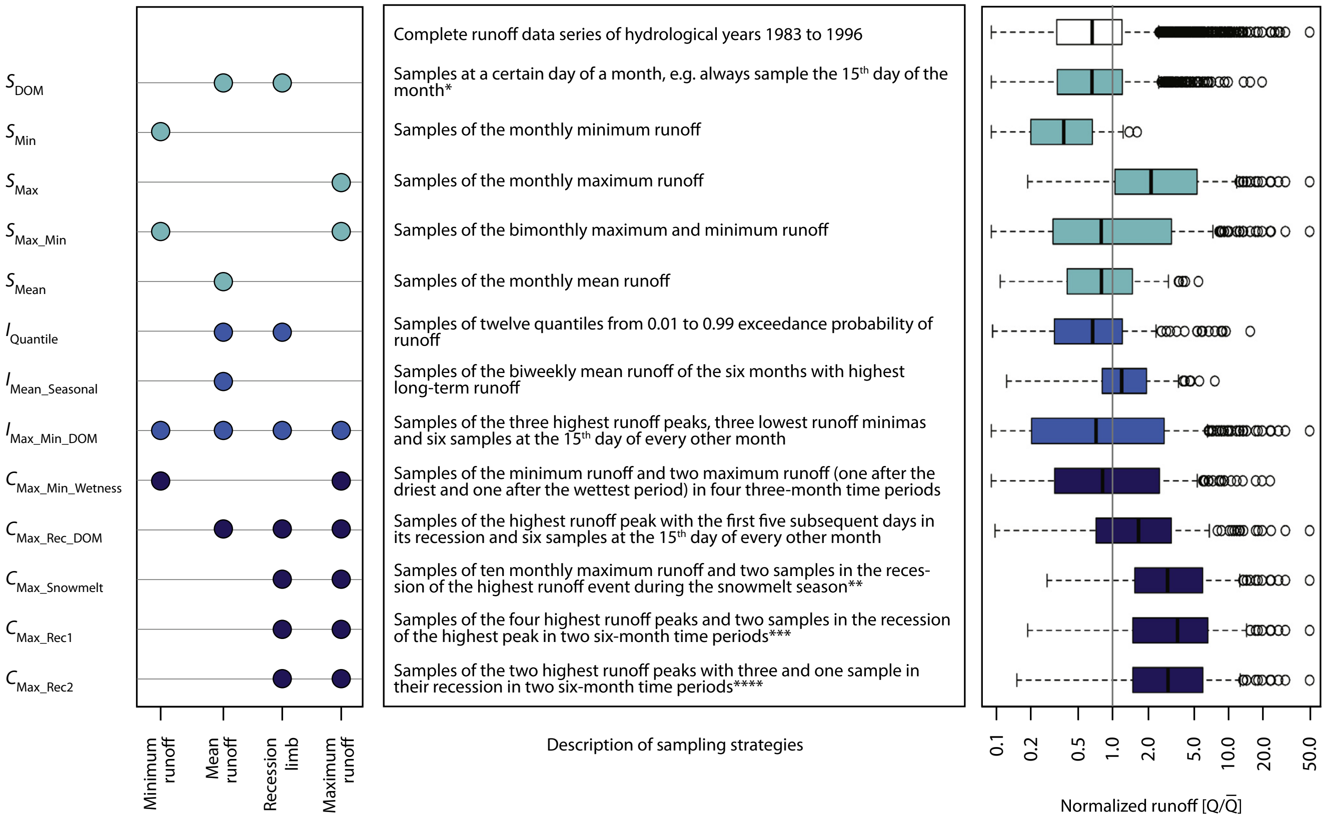
2.2. Definition of sampling strategies

Sampling strategies were defined considering both existing hydrological knowledge from previous studies (see Section 1) and practical aspects for the implementation of a runoff monitoring in the ungauged catchment of interest (Fig. 2). We defined a total of thirteen sampling strategies that were categorized as simple (*S*), intermediate (*I*) or complex (*C*) according to their hydrological background. For practical reasons it was interesting to examine sampling strategies with runoff samples at a fixed time interval (e.g. S_{DOM}). Runoff samples of event peaks or during low flow (e.g. S_{Max} or S_{Min}) could also be collected with relatively little effort as long as the exact timing was not crucial. From a hydrological point of view, strategies capturing runoff variability or dominant runoff processes could be promising. For example, the strategy I_{Quantile} contains samples over the full range of runoff magnitudes, $C_{\text{Max\_Min\_Wetness}}$ takes into account the different runoff response of catchments after dry and wet periods or additional samples are taken during the snowmelt season with $C_{\text{Max\_Snowmelt}}$. All tested sampling strategies were restricted to twelve runoff samples within a single hydrological year (1st of October until 30th of September) that were extracted from the continuous runoff time series of each catchment. The decision to test the temporal distribution of runoff at twelve times within a year was chosen to represent a balance between a minimum number of measurements assumed to be necessary for model calibration and the practical limitations of measuring runoff at several times.

2.3. Modelling approach

The runoff model was calibrated for the twenty study catchments with a limited number of runoff samples. To run the model, twelve runoff samples selected from different hydrological years and the continuous precipitation and temperature data series were used in all cases. The data of fourteen hydrological years from 1983 to 1996 were used for independent model calibrations. A warm-up period of 2.75 years preceded each calibration period to ensure suitable initial values for the state variables. Model parameters of each calibration period were evaluated in an independent continuous validation time period from 1997 to 2010 in terms of how well the simulated runoff represented the observed hydrograph and the flow-duration curve. The two modelling time periods (1983–1996 and 1997–2010) were generally similar with respect to the yearly sum of precipitation, the yearly sum of runoff, the mean annual temperature and the percentage of precipitation falling as snow in each of the twenty study catchments (statistically evaluated using a non-parametric Mann-Whitney-*U* test). The detailed modelling steps were as follows:

1. 100,000 parameter sets were randomly generated within predefined parameter ranges (Table 2) and assuming a uniform parameter distribution.
2. The model was run for each parameter set. The simulated runoff was compared to the twelve observed runoff samples of each sampling strategy and calibration period. The objective functions used for comparison were the model efficiency (Nash and Sutcliffe, 1970) calculated directly on the runoff data (R_{eff}) and the model efficiency calculated on the log-transformed runoff data ($R_{\text{eff\_logQ}}$). The 100 best parameter sets of each calibration period were retained for each strategy and objective function.



Complete runoff data series of hydrological years 1983 to 1996

Samples at a certain day of a month, e.g. always sample the 15th day of the month*

Samples of the monthly minimum runoff

Samples of the monthly maximum runoff

Samples of the bimonthly maximum and minimum runoff

Samples of the monthly mean runoff

Samples of twelve quantiles from 0.01 to 0.99 exceedance probability of runoff

Samples of the biweekly mean runoff of the six months with highest long-term runoff

Samples of the three highest runoff peaks, three lowest runoff minimas and six samples at the 15th day of every other month

Samples of the minimum runoff and two maximum runoff (one after the driest and one after the wettest period) in four three-month time periods

Samples of the highest runoff peak with the first five subsequent days in its recession and six samples at the 15th day of every other month

Samples of ten monthly maximum runoff and two samples in the recession of the highest runoff event during the snowmelt season**

Samples of the four highest runoff peaks and two samples in the recession of the highest peak in two six-month time periods***

Samples of the two highest runoff peaks with three and one sample in their recession in two six-month time periods****

Fig. 2. Definition of the thirteen sampling strategies used for model calibration. Each sampling strategy consisted of twelve runoff samples. From left to right: abbreviation of sampling strategies, conceptual idea of runoff represented by strategies, description of strategies and normalized runoff magnitudes sampled with the strategies (normalized runoff corresponds to the sampled runoff Q divided by the mean catchment runoff Q̄; data of catchment 17 (see Table 1) is shown). *S_{DOM}: we tested the strategy with samples at the 1st, 5th, 10th, 15th, 20th and 25th day of the month and finally calculated the mean performance of all these six versions; **$C_{\text{Max\_Snowmelt}}$: maximum runoff of the ten months with highest long-term runoff and recession samples taken at 80% and 60% of highest runoff peak in the snowmelt season (February to May); ***$C_{\text{Max\_Rec1}}$: recession samples taken at 80% and 40% of highest runoff peak; ****$C_{\text{Max\_Rec2}}$: recession samples taken at 80%, 60% and 40% of highest runoff peak and 80% of second highest runoff peak.

Table 2
Specification of HBV-light model parameters calibrated in this study according to Seibert and Vis (2012).

Parameter	Meaning	Unit	Minimum	Maximum
Snow routine				
TT	Threshold temperature	°C	−2	2.5
CFMAX	Degree-day factor	mm°C^{-1} d^{-1}	0.5	10
SFCF	Snowfall correction factor	–	0.5	1.2
SCR	Refreezing coefficient	–	0	0.1
CWH	Water holding capacity	–	0	0.2
Soil routine				
FC	Maximum soil moisture storage (SM)	mm	100	550
LP	Threshold for reduction of evaporation (SM/FC)	–	0.3	1
BETA	Shape coefficient	–	1	5
Groundwater routine				
PERC	Maximal flow from upper to lower box	mm d^{-1}	0	4
UZL	Maximal storage in the soil upper zone	mm	0	70
K0	Recession coefficient of fast response	d^{-1}	0.1	0.5
K1	Recession coefficient of intermediate response	d^{-1}	0.01	0.2
K2	Recession coefficient of baseflow	d^{-1}	0.00005	0.1
Routing routine				
MAXBAS	Routing, length of weighting function	d	1	5

3. The 100 best parameter sets were used to simulate runoff in the validation period. An ensemble mean hydrograph and ensemble mean FDC were calculated from the 100 runoff simulations. The ensemble mean hydrograph was evaluated in terms of R_{eff}. The ensemble mean FDC was evaluated by calculating the mean absolute relative error at 99 evaluation points of the FDC (R_{FDC}). The evaluation points were selected at equally spaced intervals of runoff volume between 0.1 and 0.99 exceedance probability, which is a similar approach to that suggested by Westerberg et al. (2011).

Model performance values in validation were normalized by relating them to an upper and a lower benchmark (Eq. (1)) as suggested by Girons Lopez and Seibert (2016). The upper benchmark represented the best possible model performance that could be achieved for a particular catchment. It was calculated with the simulation approach described above with the exception that the model was calibrated against the full continuous runoff time series of all fourteen years. While the upper benchmark parameter sets for the hydrograph were selected by applying R_{eff} or R_{eff_logQ}, R_{FDC} was used in the case of the FDC. The lower benchmark was calculated from 1000 randomly selected parameter sets and was a measure of how well the model would simulate runoff without any runoff information for a calibration. The identical normalization was applied for R_{eff} and R_{FDC} using the equation

$$R^{*} = \frac{R_{ss} - R_{lb}}{R_{ub} - R_{lb}} \quad (1)$$

with R^{*} as the normalized model performance (specifically R^{*}_{eff} and R^{*}_{FDC}), R_{ss} as the model performance based on the sampling strategy, R_{ub} as the model performance of the upper benchmark and R_{lb} as the model performance of the lower benchmark. Normalized performance values ranged from –inf to 1. A normalized performance of one indicates that model calibration with a particular sampling strategy was as good as a well-informed model calibration, whereas values below zero reveal that model calibration with a small number of strategically selected runoff measurements performs worse than simulations with random parameter sets.

Additionally, we evaluated the influence of the thirteen different sampling strategies for constraining model parameters. Since parameter values vary between catchments, we evaluated the range of parameter values, which had been calibrated based on a particular sampling strategy. Parameter ranges after calibration (0.05–0.95 quantile of all 100 parameter values) were normalized by their allowed range before calibration to make the different parameters comparable.

3. Results

When calibrated against the complete runoff time series, model performances were generally good for both the hydrograph (R_{eff}) and the FDC (R_{FDC}) (median R_{eff} 0.76 and median R_{FDC} 0.15; Fig. 3a and b, where the best possible model performance is 1.0 for the hydrograph and 0.0 for the FDC). As expected, model performances were poorer for simulations with a random parameterization (median R_{eff} 0.45 and median R_{FDC} 0.43). Model calibrations based on twelve runoff values selected by the different sampling strategies mostly resulted in performances between the two benchmarks. The hydrograph efficiency R_{eff} for all catchments and all strategies (Fig. 3a) ranged from 0.45 to 0.74 (median of 0.64) when parameter sets were selected based on R_{eff}. Calibrating the model with R_{eff_logQ} resulted in similar model performance for the hydrograph (R_{eff} from 0.48 to 0.74 with a median of 0.66) as calibrations with R_{eff}. Simulations of the FDC with a limited number of measurements (Fig. 3b) were considerably better when using the objective function R_{eff_logQ} instead of R_{eff}. Median R_{FDC} was 0.26 (range from 0.16 to 0.97) for calibrations with R_{eff_logQ} and 0.34 (range from 0.19 to 5.45) for calibrations with R_{eff}.

Model calibration with runoff data of a sampling strategy resulted in fourteen ensemble mean efficiencies for each catchment. The median of these fourteen values is an indicator of the information value of a particular strategy for model calibration. Ranking sampling strategies according to their median R^{*}_{eff} and R^{*}_{FDC} values revealed an interesting pattern with marked differences for the two evaluation criteria (Fig. 4a and b). The best ranked strategies for simulating the hydrograph (Fig. 4a) consisted of maximum runoff values mostly in combination with data in the recession of an event (e.g. $C_{Max\_Snowmelt}$). Strategies that combine maximum runoff with minimum runoff or runoff taken at a fixed time interval ranked in the middle (e.g. S_{Max_Min}). The poorest model performance was achieved by sampling minimum and mean runoff or by taking samples at a fixed time interval (e.g. S_{Min}). The described ranking pattern for the hydrograph was almost reversed when strategies were evaluated in terms of their information value for the FDC (Fig. 4b). The rank of each strategy was more consistent between the study catchments for the FDC than for the hydrograph. The differences in the ranking of strategies between catchments for the hydrograph simulation could partly be explained by catchment area and snowfall

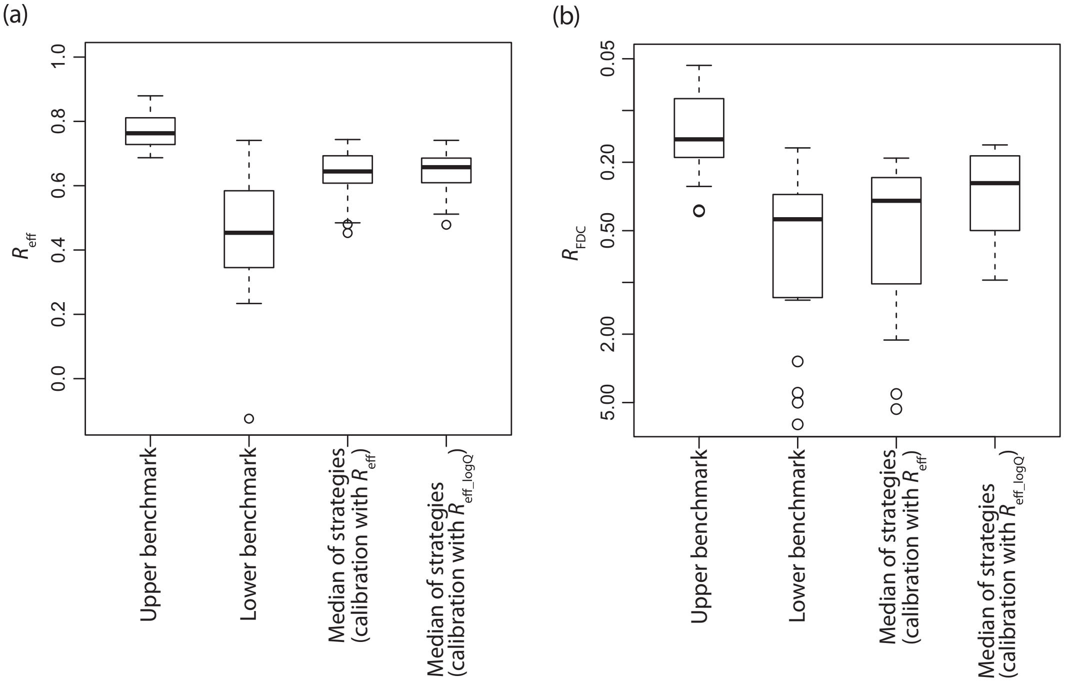


Fig. 3. Model performance for the twenty catchments as validated in terms of a) hydrograph efficiency R_{eff} and b) FDC efficiency R_{FDC} for model calibrations with the upper benchmark (continuous fourteen year calibration period), the lower benchmark (random generation of parameter sets) and the sampling strategies (twelve runoff samples) using either R_{eff} or R_{eff_logQ} as objective function. Best possible model performance is 1.0 for R_{eff} and 0.0 for R_{FDC}. Model performance related to the benchmarks was calculated as the median ensemble mean model performance of all calibration years for each catchment. Model performance of the sampling strategies is summarized by the median model performance of all strategies for each catchment. Strategy performance was calculated on the basis of the median ensemble mean performance of all calibration years.

ratio, whereby large catchments or small snow-dominated catchments tended to form clusters with a slightly different ranking of the sampling strategies. Other catchment characteristics such as mean elevation, precipitation seasonality, aridity, importance of baseflow or percentage of wetland area did not help to explain the mentioned variations. Not all strategies were more informative for model calibration than the lower benchmark with random parameter sets (Fig. 4). Especially catchments with a high model performance of the lower benchmark (R_{eff} >0.7), such as catchment C3, C9 and C11, had many sampling strategies with a negative normalized model performance for the hydrograph. Negative R^*_{FDC} values were most prominent in the low ranked sampling strategies ($I_{Mean\_Seasonal}$, $C_{Max\_Rec2,}$ $C_{Max\_Snowmelt,}$ and C_{Max_Rec1}), suggesting that these strategies cannot be considered as an acceptable option for deciding on when to make runoff measurements in many catchments.

To evaluate the impact of using either R_{eff} or R_{eff_logQ} as objective function on the evaluation of the different sampling strategies, we focused on the median R^*_{eff} and median R^*_{FDC} values of a strategy over all catchments (Fig. 5a and b). Samples of maximum runoff were always crucial for a good hydrograph simulation, whereby the magnitude or timing of additional samples seemed to be of minor importance (e.g. S_{Max} or $C_{Max\_Rec\_Dom}$). R^*_{eff} values were between 0.52 and 0.72 for strategies containing high runoff values, independent of which of the two objective functions was applied in model calibration. In contrast, R^*_{FDC} clearly differed for some strategies as a function of the objective function. All sampling strategies with high runoff values poorly constrained model parameters for FDC simulations when calibrated based on R_{eff}. Using the objective function R_{eff_logQ} for model calibration strongly improved R^*_{FDC} for strategies combining maximum runoff with minimum runoff or with runoff samples at a fixed time interval ($I_{Max\_Min\_Dom}$, $C_{Max\_Rec\_Dom}$, S_{Max_Min} and $C_{Max\_Min\_Wetness}$). Sampling strategies covering low and mean flows (S_{Min}, S_{Mean}, S_{DOM} and $I_{Quantile}$) mostly led to good R^*_{FDC} values with slightly higher model performance for calibrations based on R_{eff_logQ} (R^*_{FDC} from 0.78 to 0.92). Model calibration on R_{eff_logQ} guided parameter selection in a way that some sampling strategies provided informative runoff samples for both hydrograph and FDC, whereas the value of sampling strategies was restricted to either of these simulation aims for calibrations with R_{eff} (Fig. 5a and b).

Model performance generally varied greatly between calibration periods for all strategies and catchments (Fig. 6a and b; standard deviation shown on y-axis). However, it was not possible to establish any relation between hydroclimatic conditions (e.g. yearly or seasonal precipitation, runoff or snowfall) or variations in runoff measurement magnitudes and model performance of the calibrated model. The differences in yearly model performance were smaller for model calibrations with informative sampling strategies, which was indicated by the negative correlation between the median model performance and the standard deviation of the model performance for calibrations based on R_{eff_logQ} (Fig. 6a and b). Also, the relative value of sampling strategies for the simulation of the hydrograph or the FDC was consistent over the fourteen calibration periods (Fig. 7).

We were further interested in how sampling strategies constrained the different model parameters during calibration (Fig. 8). Parameters of the snow routine had mostly large normalized parameter ranges for all sampling strategies indicating that model simulations were often not sensitive to the parameter value. This was different for the five catchments with the highest percentage of precipitation falling as snow, where TT, CFMAX and SFCF were clearly better constrained with normalized ranges as low as 0.42, 0.25, and 0.65. Parameters influencing the water balance (soil routine and PERC of groundwater routine) were better constrained by strategies that sample low and mean flow. However, hydrograph related parameters (UZL, K0 and MAXBAS in

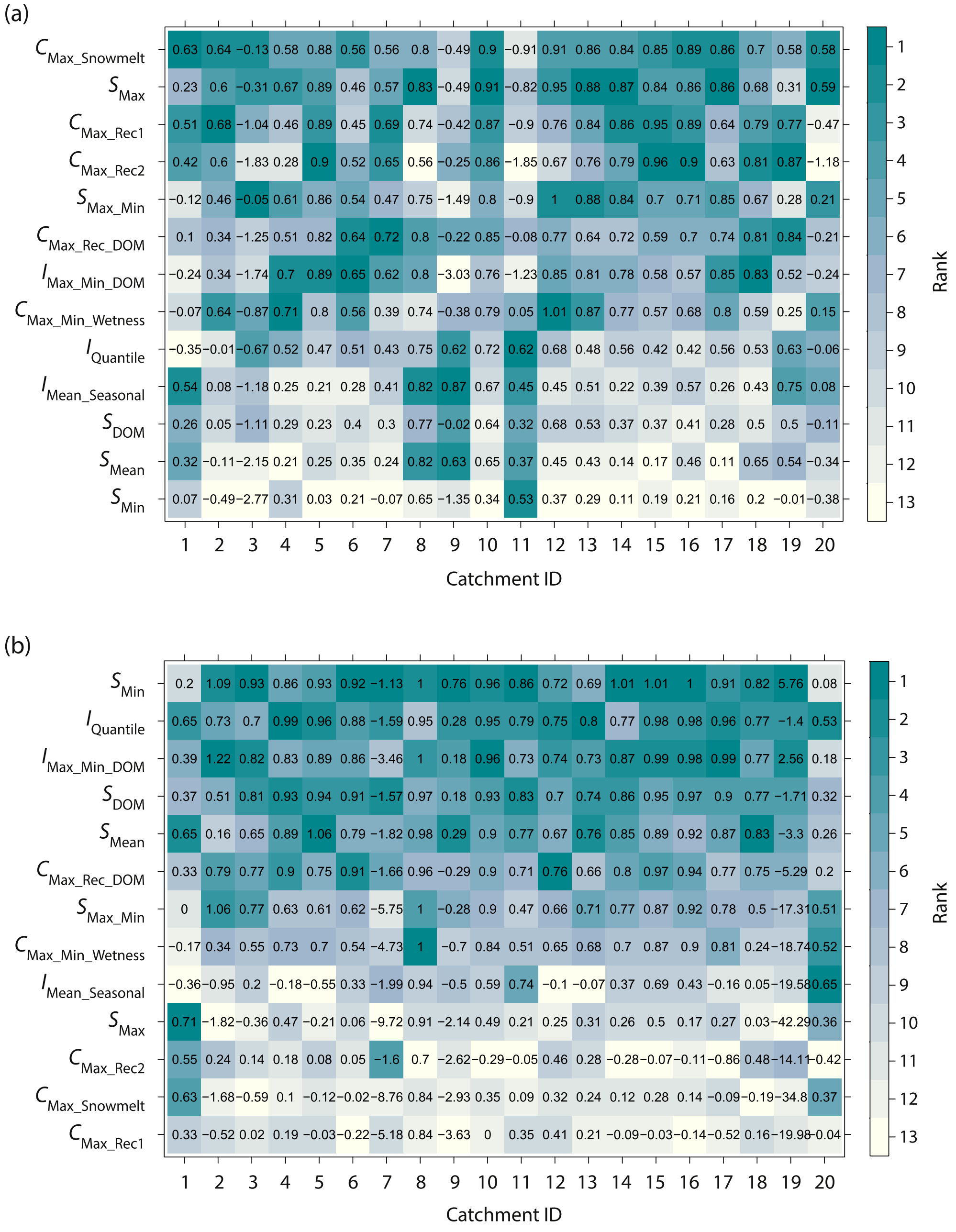


Fig. 4. Normalized model performance as validated for a) the hydrograph (R^{*}_{eff}) and b) the FDC (R^{*}_{FDC}) for model calibrations with the sampling strategies using R_{eff_logQ} as objective function. The normalized performance values correspond to the median ensemble mean of all calibration years. Sampling strategies were ranked according to their model performance. Sampling strategies on the y-axis are ordered by their mean rank over all catchments. Colours indicate the rank of a sampling strategy for a particular catchment. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the groundwater and routing routine) were generally more similar if the model was calibrated with sampling strategies containing maximum runoff.

4. Discussion

The modelling results indicate that a limited number of strategically selected runoff samples is informative for hydrograph and FDC simulations in almost ungauged catchments. Different combinations of runoff samples had a different information value for simulating the hydrograph and the FDC. Possible factors contributing to this difference could be the runoff distribution resulting from a particular sampling strategy (boxplots in Fig. 2) and the model parameters most sensitive at the point in time a runoff sample was provided for calibration. Model parameters of the groundwater and the routing routine that define the timing and the shape of the hydrograph had the least uncertainty when the model was calibrated with runoff samples of high flows and recessions. The benefit of maximum runoff and event data for model calibration was also reported by Seibert and Beven (2009) and Seibert and McDonnell (2013). Our results also confirm the conclusion of several studies (Yapo et al., 1996; Vrugt et al., 2006; Kim and Kaluarachchi, 2009; Melsen et al., 2014; Correa et al., 2016) that rather average and dry runoff periods, represented by samples of mean and minima flows, are less informative for hydrograph prediction than wet periods. For FDC simulations it is crucial to accurately model runoff magnitudes, whereas the exact shape of the hydrograph is less important. Therefore, sampling strategies

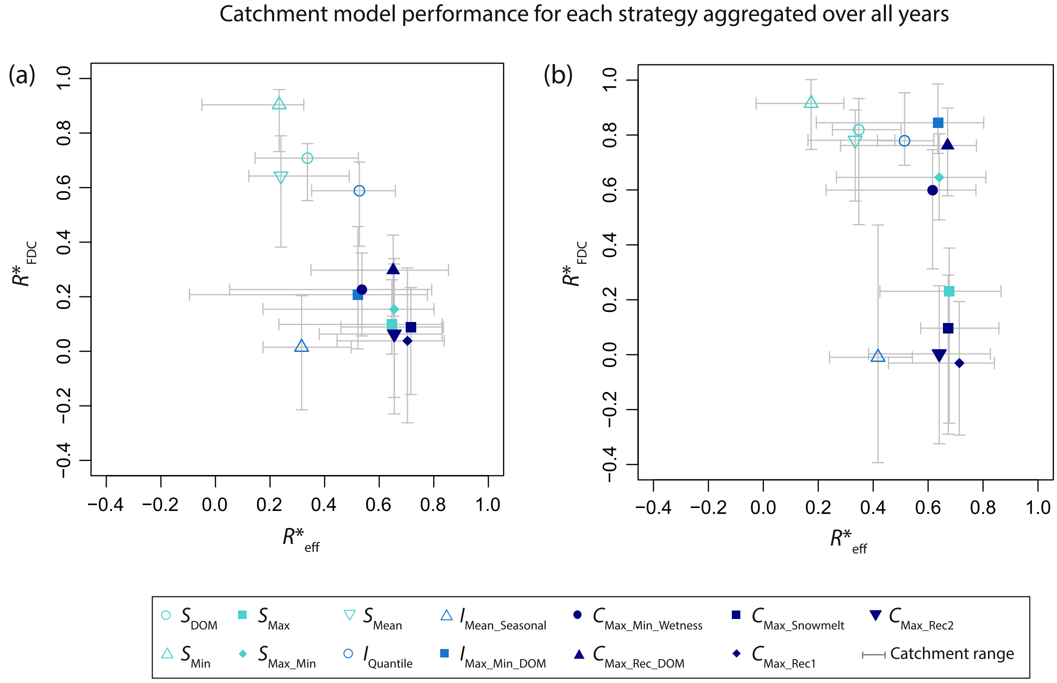


Fig. 5. Normalized model performance as validated for the hydrograph (R^*_{eff}) and the FDC (R^*_{FDC}) for model calibrations with the sampling strategies using (a) R_{eff} and (b) $R_{\mathrm{eff\_logQ}}$ as objective functions. Each symbol represents the median model performance for a particular strategy over all catchments. It was calculated on the basis of the median ensemble mean of all calibration years. Error bars indicate the 0.25–0.75 quantile model performance of all catchments for the respective strategy. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

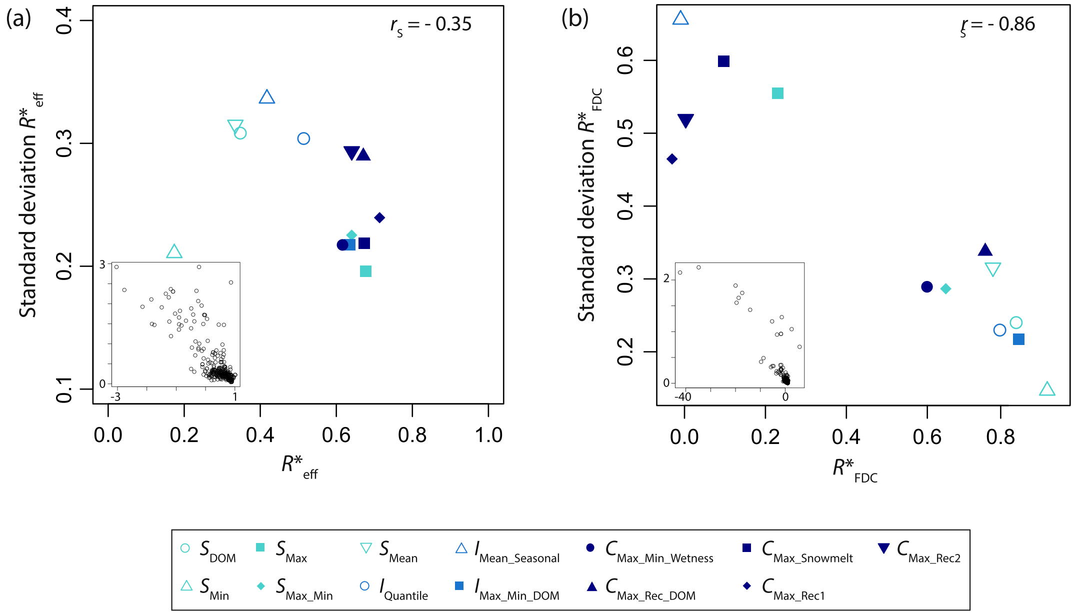


Fig. 6. Comparison of the normalized model performance and the standard deviation of the normalized model performance as validated for a) the hydrograph (R^*_{eff}) and b) the FDC (R^*_{FDC}) for model calibrations with the sampling strategies using $R_{\mathrm{eff\_logQ}}$ as objective function. Each coloured symbol represents the median model performance and the median standard deviation of the model performance for a particular strategy over all catchments. The median and the standard deviation were calculated on the basis of the ensemble mean of all calibration years. r_S corresponds to the Spearman's rank correlation coefficient between the median R^*_{eff} and the standard deviation of R^*_{eff}. The inset plot makes the same comparison, but indicating the values for each catchment separately. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

resulting in a comparable runoff distribution as a continuous long-term runoff time series were most valuable for simulating the FDC. These strategies, e.g. S_{DOM}, S_{Mean} or I_{Quantile}, were most effective in constraining parameters with strong impact on the water balance (soil routine and percolation parameters). None of the sampling strategies noticeably reduced the high uncertainty of snow related model parameters, probably because many study catchments had no or little snowfall.

It is interesting that strategies combining samples of maximum, minimum and recession flow could become informative for the

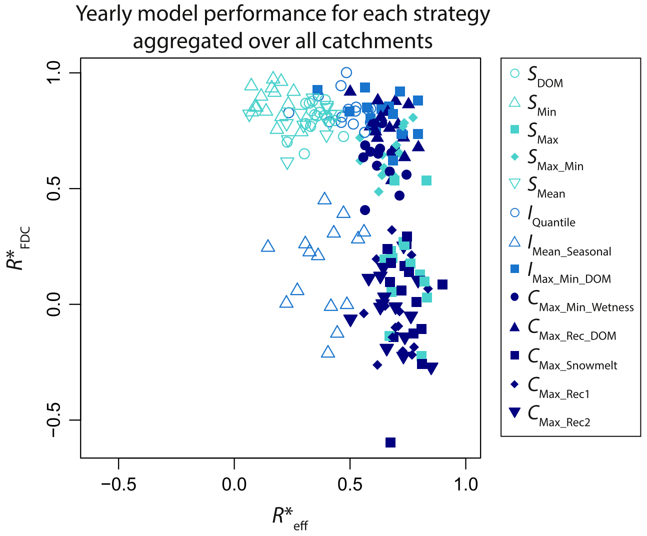


Fig. 7. Normalized model performance as validated for the hydrograph (R^*_{eff}) and the FDC (R^*_{FDC}) for model calibrations with the sampling strategies using R_{eff_logQ} as objective function. Each symbol represents the median model performance for a particular strategy over all catchments for one calibration year. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

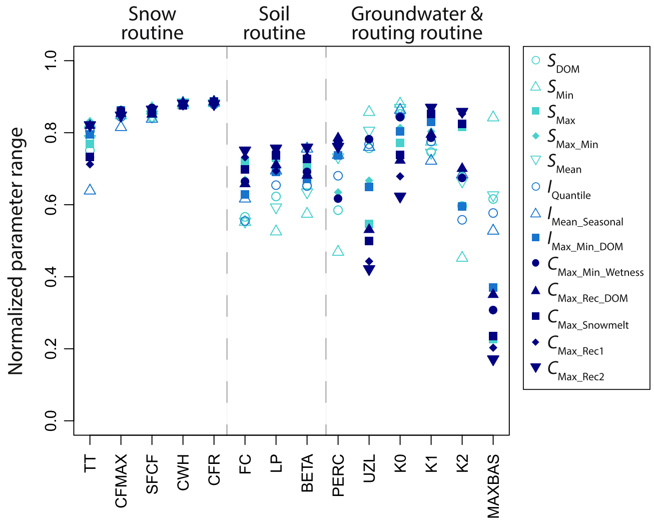


Fig. 8. Normalized model parameter ranges resulting from model calibrations with the sampling strategies using R_{eff_logQ} as objective function. Parameter ranges (0.05–0.95 quantile) after calibration were normalized by their allowed range before calibration. The symbols represent the median normalized parameter range of all catchments related to a particular strategy. This range was calculated on the basis of the median normalized parameter range of all calibration years. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

prediction of the FDC when HBV was calibrated with R_{eff_logQ} instead of R_{eff}. This considerable change could be explained by the distinct focus of the two objective functions during calibration. R_{eff_logQ} emphasises low and mean flow giving more weight to the accurate simulation of a range of magnitudes, while the timing of peak flows is of minor importance. This result demonstrates the importance of carefully choosing the objective function used to optimize model simulations.

The ranking of sampling strategies according to their related model performance (Fig. 4a and b) was clearly less consistent between the twenty catchments for the hydrograph than for the FDC. We tested various catchment characteristics to explain these ranking differences, but no variable was found that could clearly explain the results. Similarly, it was not possible to establish consistently strong relationships between catchment characteristics and the yearly model performance. The sample of twenty catchments might have been too small to find strong relationships between catchment characteristics and model performance as observed by Perrin et al. (2007) in a comparable modelling study framework.

In this study we decided to analyse the modelling results in relation to benchmarks instead of focusing on absolute model performance values. As suggested by Girons Lopez and Seibert (2016), we related model performance based on a limited number of runoff measurements to model calibrations of a well and a non-informed situation. The concept of benchmarks is especially beneficial when predicting runoff for almost ungauged catchments, where the value of taking a few runoff measurements compared to investing efforts in long-term gauging stations is of interest. Absolute model performance becomes more important for practical applications as efficiencies are too low for a reasonable runoff simulation. At this point it is also important to note that low normalized performance does not imply a poor model calibration. For example, the catchments C3, C9 and C11 had many negative normalized performance values due to high Monte Carlo efficiencies. However hydrographs of these catchments were all well simulated in absolute terms. We would also like to stress that the interpretation of the results was not affected by the use of benchmarked performances, because the normalization of model performance did not change the hierarchy of the thirteen sampling strategies within a catchment.

The proposed sampling strategy approach was implemented assuming that one can take a runoff measurement exactly at a certain point in time, such as at the monthly maximum runoff. This is not possible in practice as the runoff is not known at the beginning of a month or a year. The results in our study give an indication of what could be achieved at best and the question is how much the results might have been affected when the runoff was observed at slightly different points in time. Our modelling results suggested that there is some flexibility in taking runoff samples, because none of the tested sampling strategies proved to be superior for model calibration. In the case of hydrograph prediction it was most important to sample high flows preferably in combination with recession data. The most informative sampling strategies for simulating the FDC are not very time sensitive and it was more essential to sample a representative runoff distribution of the particular catchment.

5. Conclusion

This study evaluated the information value of a small number of runoff measurements for calibrating a runoff model for almost ungauged catchments. Our calibration approach has some interesting implications for the prediction of runoff in almost ungauged catchments. It shows the potential of calibrating a runoff model with as few as twelve strategically sampled runoff measurements. Since the exact timing of taking runoff samples was not a major constraint for model calibration, taking samples could be a realistic and efficient alternative to installing a long-term gauging station. Additionally, we applied a runoff model that only requires daily temperature, precipitation and monthly potential evaporation as input, which are variables often available in many regions around the world. The proposed calibration approach could therefore be especially valuable for water management decisions and the mitigation of natural hazards in data scarce regions. However, in case of remote catchments, it might not be time and cost effective to take twelve runoff samples distributed over a hydrological year.

Different strategies for sampling runoff at higher time resolutions within the duration of a short field campaign could be tested to evaluate the value of data for these catchments. Furthermore, our results are limited to humid catchments with little precipitation seasonality and dominated by rain or snow processes. Further investigations are required to evaluate the value of individual runoff measurements, for e.g., arid and glaciated catchments or catchments with a marked precipitation seasonality.

Acknowledgements

The authors thank all the people taking the time to discuss this study and supporting us with new ideas and comments for sampling strategies. We also thank the University of Zurich for funding this work. The two anonymous reviewers are acknowledged for their valuable comments on an earlier draft of this manuscript. Hydrometerological data and catchment shapefiles were made available from Newman et al. (2015). SRTM elevation data was used from Jarvis et al. (2008).

References

Beven, K.J., 2012. Rainfall-Runoff Modelling – The Primer. Wiley and Sons, Chichester.

Bergström, S., 1976. Development and Application of a Conceptual Runoff Model for Scandinavian Catchments. SMHI, Norrköping, Sweden, No. RHO 7, pp 134.

Correa, A., Windhorst, D., Crespo, P., Célleri, R., Feyen, J., Breuer, L., 2016. Continuous versus event-based sampling: how many samples are required for deriving general hydrological understanding on Ecuador's páramo region? Hydrol. Process. 30, 4059–4073. https://doi.org/10.1002/hyp.10975.

Coopersmith, E.J., Minsker, B.S., Sivapalan, M., 2014. Patterns of regional hydroclimatic shifts: an analysis of changing hydrologic regimes. Water Resour. Res. 50, 1960–1983. https://doi.org/10.1002/2012WR013320.

Drogue, G.P., Plasse, J., 2014. How can a few streamflow measurements help to predict daily hydrographs at almost ungauged sites? Hydrol. Sci. J. 59, 2126–2142. https://doi.org/10.1080/02626667.2013.865031.

ESRI and U.S. Geological Survey, 2011. Shaded relief, medium resolution, USA. Available at: ArcGIS online maps and data, last access: March 2017.

Girons Lopez, M., Seibert, J., 2016. Influence of hydro-meteorological data spatial aggregation on streamflow modelling. J. Hydrol. 541, 1212–1220. https://doi.org/10.1016/j.jhydrol.2016.08.026.

Freer, J.E., McMillan, H., McDonnell, J.J., Beven, K.J., 2004. Constraining dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures. J. Hydrol. 291, 254–277. https://doi.org/10.1016/j.jhydrol.2003.12.037.

Harlin, J., 1991. Development of a process oriented calibration scheme for the HBV hydrological model. Nordic Hydrol. 22, 15–36.

Hrachowitz, M., Savenije, H.H.G., Blöschl, G., McDonnell, J.J., Sivapalan, M., Pomeroy, J.W., Arheimer, B., Blume, T., Clark, M.P., Ehret, U., Fenicia, F., Freer, J.E., Gelfan, A., Gupta, H.V., Hughes, D.A., Hut, R.W., Montanari, A., Pande, S., Tetzlaff, D., Troch, P.A., Uhlenbrook, S., Wagener, T., Winsemius, H.C., Woods, R.A., Zehe, E., Cudennec, C., 2013. A decade of predictions in ungauged basins (PUB): a review. Hydrol. Sci. J. 58, 1198–1255. https://doi.org/10.1080/02626667.2013.803183.

Hughes, D.A., Gush, M., Tanner, J., Dye, P., 2014. Using targeted short-term field investigations to calibrate and evaluate the structure of a hydrological model. Hydrol. Process. 28, 2794–2809. https://doi.org/10.1002/hyp.9807.

Jarvis, A., Reuter, H.I., Nelson, A., Guevara, E., 2008. Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m. Available at: http://srtm.csi.cgiar.org, last access: January 2016.

Juston, J., Seibert, J., Johansson, P.O., 2009. Temporal sampling strategies and uncertainty in calibrating a conceptual hydrological model for a small boreal catchment. Hydrol. Process. 23, 3093–3109. https://doi.org/10.1002/hyp.7421.

Kim, U., Kaluarachchi, J.J., 2009. Hydrologic model calibration using discontinuous data: an example from the upper Blue Nile River Basin of Ethiopia. Hydrol. Process. 23, 3705–3717. https://doi.org/10.1002/hyp.7465.

Konz, M., Seibert, J., 2010. On the value of glacier mass balances for hydrological model calibration. J. Hydrol. 385, 238–246. https://doi.org/10.1016/j.jhydrol.2010.02.025.

Lehner, B., Döll, P., 2004. Development and validation of a global database of lakes, reservoirs and wetlands. J. Hydrol. 296, 1–22. https://doi.org/10.1016/j.jhydrol.2004.03.028.

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

McIntyre, N.R., Wheater, H.S., 2004. Calibration of an in-river phosphorus model: prior evaluation of data needs and model uncertainty. J. Hydrol. 290, 100–116. https://doi.org/10.1016/j.jhydrol.2003.12.003.

Melsen, L.A., Teuling, A.J., Berkum, S.W., Torfs, P.J.J.F., Uijlenhoet, R., 2014. Catchments as simple dynamical systems: a case study on methods and data requirements for parameter identification. Water Resour. Res. 50, 5577–5596. https://doi.org/10.1002/2013WR014720.

Merz, R., Parajka, J., Blöschl, G., 2009. Scale effects in conceptual hydrological modeling. Water Resour. Res. 45, W09405. https://doi.org/10.1029/2009WR007872.

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

Parajka, J., Viglione, A., Rogger, M., Salinas, J.L., Sivapalan, M., Blöschl, G., 2013. Comparative assessment of predictions in ungauged basins-Part 1: Runoff-hydrograph studies. Hydrol. Earth Syst. Sci. 17, 783–1795. https://doi.org/10.5194/hess-17-1783-2013.

Perrin, C., Oudin, L., Andreassian, V., Rojas-Serna, C., Michel, C., Mathevet, T., 2007. Impact of limited streamflow data on the efficiency and the parameters of rainfall-runoff models. Hydrol. Sci. J. 52, 131–151. https://doi.org/10.1623/hysj.52.1.131.

Rojas-Serna, C., Lebecherel, L., Perrin, C., Andreassian, V., Oudin, L., 2016. How should a rainfall-runoff model be parameterized in an almost ungauged catchment? A methodology tested on 609 catchments. Water Resour. Res. 52, 4765–4784. https://doi.org/10.1002/2015WR018549.

Rojas-Serna, C., Michel, C., Perrin, C., Andreassian, V., Hall, A., Chahinian, N., Schaake, J., 2006. Ungauged catchments: How to make the most of a few streamflow measurements? Large sample basin experiments for hydrological model parameterization: results of the model parameter experiment – MOPEX. IAHS Publ. 307, 230–236.

Seibert, J., Beven, K.J., 2009. Gauging the ungauged basin: How many discharge measurements are needed? Hydrol. Earth Syst. Sci. 13, 883–892. https://doi.org/10.5194/hess-13-883-2009.

Seibert, J., McDonnell, J.J., 2013. Gauging the ungauged basin: relative value of soft and hard data. J. Hydrol. Eng. 20, A4014004. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000861.

Seibert, J., Vis, M.J.P., 2012. Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. Hydrol. Earth Syst. Sci. 16, 3315–3325. https://doi.org/10.5194/hess-16-3315-2012.

Singh, S.K., Bárdossy, A., 2012. Calibration of hydrological models on hydrologically unusual events. Adv. Water Resour. 38, 81–91. https://doi.org/10.1016/j.advwatres.2011.12.006.

Sivapalan, M., Takeuchi, K., Franks, S.W., Gupta, V.K., Karambiri, H., Lakshim, V., Liang, X., McDonnell, J.J., Mendiondo, E.M., Connell, O., Oki, T., Pomeroy, J.W., Schertzer, D., Uhlenbrook, S., Zehe, E., 2003. IAHS decade on Predictions in Ungauged Basins (PUB), 2003–2012: shaping an exciting future for the hydrological sciences. Hydrol. Sci. J. 48, 857–880. https://doi.org/10.1623/hysj.48.6.857.51421.

Sun, W., Wang, Y., Wang, G., Cui, X., Yu, J., Zuo, D., Xu, Z., 2017. Physically based distributed hydrological model calibration based on a short period of streamflow data: case studies in four Chinese basins. Hydrol. Earth Syst. Sci. 21, 251–265. https://doi.org/10.5194/hess-21-251-2017.

Uhlenbrook, S., Sieber, A., 2005. On the value of experimental data to reduce the prediction uncertainty of a process-oriented catchment model. Environ. Modell. Softw. 20, 19–32. https://doi.org/10.1016/j.envsoft.2003.12.006.

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

Viviroli, D., Seibert, J., 2015. Can a regionalized model parameterisation be improved with a limited number of runoff measurements? J. Hydrol. 529, 49–61. https://doi.org/10.1016/j.jhydrol.2015.07.009.

Vrugt, J.A., Gupta, H.V., Dekker, S.C., Sorooshian, S., Wagener, T., Bouten, W., 2006. Application of stochastic parameter optimization to the Sacramento soil moisture accounting model. J. Hydrol. 325, 288–307. https://doi.org/10.1016/j.jhydrol.2005.10.041.

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

Xia, Y., Yang, Z.L., Jackson, C., Stoffa, P.L., Sen, M.K., 2004. Impacts of data length on optimal parameter and uncertainty estimation of a land surface model. J. Geophys. Res. Atmos. 109, D07101. https://doi.org/10.1029/2003JD004419.

Yapo, P.O., Gupta, H.V., Sorooshian, S., 1996. Automatic calibration of conceptual rainfall-runoff models: Sensitivity to calibration data. J. Hydrol. 181, 23–48. https://doi.org/10.1016/0022-1694(95)02918-4.