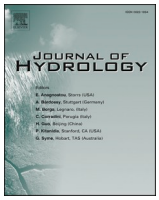
Journal of Hydrology 632 (2024) 130686

Contents lists available at ScienceDirect

Journal of Hydrology

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



Research papers

Are temporary stream observations useful for calibrating a lumped hydrological model?

Mirjam Scheller [a,*], Ilja van Meerveld [a], Eric Sauquet [b], Marc Vis [a], Jan Seibert [a]

[a] Department of Geography, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
[b] INRAE, UR RiverLy, Villeurbanne, France

ARTICLE INFO

Keywords:
Intermittent streams
Non-perennial streams
Value of data
Model calibration
Parameter uncertainty
7-day minimum flow

ABSTRACT

Multi-criteria model calibration can lead to a better representation of hydrological processes and reduce parameter uncertainty compared to calibration on streamflow data alone. However, the additional data may be difficult to collect or aggregate into a representative catchment average value that can be used to calibrate a lumped model. Temporary streams are highly dynamic, and their flow state can be observed visually. However, data on the state of temporary streams are still uncommon and rarely used in hydrological catchment modelling. In this study, we used a unique dataset with discrete flow state observations for temporary streams in France and evaluated how informative these data are for calibrating a lumped, bucket-type hydrological model. We calibrated the HBV model for 92 catchments using discharge or stream-level data at different temporal resolutions (daily, one daily value per month, or one daily value per season) and used the observed flow states of temporary streams as a proxy of groundwater storage. Temporary stream data generally did not result in a better overall discharge simulation for the validation period. For catchments for which the model performance based on the calibration on only discharge or stream-level data was poor, it was more likely to lead to an improvement in model performance. The use of temporary stream data in combination with discharge data reduced the uncertainties in the low-flow simulations for up to half of the catchments. This improvement was caused by a better-constrained storage coefficient for the slowest groundwater reservoir and the elimination of parameter sets that led to substantial variations in groundwater storage. However, the improvements in low-flow simulations or parameter uncertainty due to the inclusion of temporary stream data in model calibration were not related to catchment characteristics. Thus, it remains unclear for which catchments temporary stream data can help to improve low-flow simulations and reduce parameter uncertainty.

1. Introduction

Hydrological models are used for water management decisions, scenario analyses, and predictions. Lumped bucket-type models simulate the entire catchment as one unit. They require fewer data and parameters than physically based, fully distributed models, but the parameters cannot be measured directly and require calibration. The most common way to calibrate a lumped bucket-type hydrological model is to maximize the agreement between the simulated and observed discharge time series. However, for many streams, no or only limited discharge data are available for calibration (Oudin et al., 2008; Parajka et al., 2013; Sivapalan et al., 2003). Previous studies have found that a few discharge measurements can effectively constrain a hydrological model (Melsen et al., 2014; Seibert and Beven, 2009), and that water level or water level class data are informative for calibration as well (Jian et al., 2017; Seibert and Vis, 2016; van Meerveld et al., 2017), even if they are irregular in time and uncertain (Avellaneda et al., 2020; Etter et al., 2020a; Weeser et al., 2019).

Other data types can be used in model calibration as well, in particular to avoid overparameterization and reduce parameter uncertainty (Beldring, 2002; Seibert et al., 2019a) or to identify inappropriate model structures (Schaefli and Huss, 2011). Time series of groundwater levels (Pelletier and Andréassian, 2022), soil moisture (Dimitrova-Petrova et al., 2020; Parajka et al., 2006), remotely sensed water storage (Demirel et al., 2019), remotely sensed snow or glacier extent (Finger et al., 2015), hydrochemical data (Holmes et al., 2022), and isotope data (Nan et al., 2021) have all been used in combination with discharge data for multi-criteria (or multi-data) model calibration. 'Soft-data' based on

* Corresponding author.
E-mail address: Mirjam.scheller@geo.uzh.ch (M. Scheller).

https://doi.org/10.1016/j.jhydrol.2024.130686
Received 11 July 2023; Received in revised form 25 October 2023; Accepted 18 December 2023
Available online 3 February 2024
0022-1694/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

a general understanding of the functioning of a particular catchment can be used as well (Parajka et al., 2007; Seibert and McDonnell, 2002; Viaché et al., 2004). The inclusion of additional data sources in model calibration does not necessarily improve overall model performance and may lead to a slight deterioration of model performance in terms of the discharge simulations. However, the additional data are often still considered beneficial because the simulation of these variables improved (Finger et al., 2015; Mostafaie et al., 2018; Seibert, 2000) and/or parameter uncertainties were reduced (Beldring, 2002; Pelletier and Andréassian, 2022; Seibert, 2000). In other studies, the additional data led to a better simulation of specific parts of the hydrograph, e.g., flood events (Aubert et al., 2003) or low-flows (Seibert, 2000).

However, using additional data to calibrate a lumped bucket-type model is not straightforward because the field measurements are usually not directly comparable to the simulated variables. This means that the measured values need to be standardized, normalized, or transformed before they can be used in model calibration (Pelletier and Andréassian, 2022; Seibert, 2000; Széles et al., 2020). These adjustments can lead to a loss of information content or require additional model parameters (Pelletier and Andréassian, 2022; Seibert, 2000). For example, groundwater levels measured in wells across a catchment cannot be directly compared with the simulated groundwater storage in a lumped model and require the computation of an average water level time series. This averaging can be challenging because the groundwater level dynamics often differ for wells in the riparian zone and on the hillslopes (e.g., Detty and McGuire, 2010; van Meerveld et al., 2015). Furthermore, groundwater wells are more likely to be installed in riparian zones than hillslopes (Seibert et al., 2019a; Széles et al., 2020), which limits the representativeness of the individual measurements for the catchment average. Because groundwater level dynamics in the riparian zone are generally well correlated with the discharge (Seibert et al., 2003), these data may also not provide independent information beyond the observed discharge. Similarly, for soil moisture, the mean value of multiple soil moisture observations may not represent the actual catchment average soil moisture storage as soil moisture measurements are only representative of small areas (Martínez-Fernández and Ceballos, 2005). The average groundwater levels or soil moisture values can also not be used directly in model calibration because the measured values often do not correspond to the conceptual groundwater level or soil moisture storage in the model. Remotely sensed data (e.g., water storage, surface water extent, soil moisture, snow cover) was found to be more or less beneficial for the simulation of otherwise ungauged catchments (Kundu et al., 2017; Li et al., 2018; Mostafaie et al., 2018; Parajka et al., 2007; Revilla-Romero et al., 2015). However, they also come with high uncertainties, especially when aggregated to the catchment scale for use with a lumped model (Bennett et al., 2014; Samain and Pauwels, 2013). Remotely sensed data may, furthermore, not be available for small headwater catchments.

One less explored potential source of additional information for model calibration is the flow state of temporary (i.e., non-perennial) streams in headwater catchments. These streams are highly dynamic (Wohl, 2017) and common. More than half of the global river network is non-perennial (Datry et al., 2014; Larned et al., 2010; Messager et al., 2021). However, data about the flow state of these streams are limited. Only recently, approaches such as low-cost sensors (Assendelft and van Meerveld, 2019; Zanetti et al., 2022) and visual approaches (Beaufort et al., 2018; Datry et al., 2016; Durighetto et al., 2020; Godsey and Kirchner, 2014; Stubbington et al., 2017) have been developed and tested.

A few studies have looked at the value of temporary stream observations for calibrating spatially explicit models. Mahoney et al. (2023), for instance, investigated the dynamics of headwater stream drying with a process-based semi-distributed hydrological model and suggested that observations of the state of temporary streams can be useful for calibration. Similarly, Stoll and Weiler (2010) showed that information on the flowing stream network improved the calibration of the process-based Hill-Vi model. Several other modelling studies have focussed on temporary streams to assess the impacts of climate change on temporary stream dynamics (Beaufort et al., 2018; Beaufort et al., 2019; Botter and Durighetto, 2020; Gutiérrez-Jurado et al., 2019; Hammond et al., 2021; Jaeger et al., 2019; Sauquet et al., 2021). The effects of climate change on the regional probabilities of drying were, for example, simulated with the help of a lumped process-based model and different types of regressions using groundwater level, discharge, and data on the flow state of temporary streams (Beaufort et al., 2018; Sauquet et al., 2021). Virtual experiments with a fully integrated surface–subsurface hydrological model by Gutiérrez-Jurado et al. (2019) aimed to identify the factors that determine where and when a stream may dry up or start flowing again.

It has been suggested that data on the flow state of temporary streams can also be informative for the calibration of lumped hydrological models (van Meerveld et al., 2020), but this has so far not been tested in detail. Because temporary stream data can be obtained over large areas with citizen science approaches, evaluating the value of flow state observations for model calibration is interesting and useful. In this study, we take advantage of the ONDE (Observatoire National des Etiages Network, *https://onde.eaufrance.fr/*) dataset, which contains discrete observations of the state of temporary streams throughout France, to test if temporary stream data can improve the calibration of a lumped hydrological model. More specifically, we used the ONDE temporary stream data in combination with discharge or (synthetic) stream-level data to calibrate the HBV model, a typical lumped conceptual model. We assumed that the temporary stream observations are an indicator of catchment storage (i.e., that storage is low when the temporary streams in the headwater catchments are dry, and that storage is high when all temporary streams are flowing). We assessed the value of the temporary stream observations as indicators of catchment storage for model calibration in terms of overall discharge simulation performance, low-flow simulation performance, and parameter uncertainty. Because discharge data are lacking for many streams, we evaluated the value of temporary stream data in combination with different amounts of discharge or stream-level data (i.e., different temporal resolutions: daily data, one measurement per month, and one measurement per season). Finally, we tested for what kind of catchments, temporary stream data are useful for model calibration.

2. Methods

We calibrated the lumped process-based HBV model (Seibert, 2000) for 92 catchments throughout France with discharge or stream-level data of different temporal resolutions (daily, one daily value per month, or one daily value per season), either alone or in combination with temporary stream observations as indicators of catchment storage. Thus, in total, we compared 12 data scenarios for model calibration for each catchment.

2.1. Datasets and selection of catchments

We used daily discharge data from the French river discharge monitoring network (HYDRO, *https://www.hydro.eaufrance.fr/*). This dataset includes 632 catchments with limited human influences (Brigode et al., 2020; Caillouet et al., 2017). The catchment average daily precipitation and air temperature for the catchments were extracted from the gridded Safran dataset, which is based on a combination of measurements at meteorological stations and analyses from numerical weather prediction models and has an 8-km and up to an hourly resolution (Quintana-Seguí et al., 2008; Vidal et al., 2010). Daily reference evapotranspiration was determined using the Penman-Monteith formula (Allen et al., 1998).

In addition, we used the observations of the flow state of headwater streams from the Observatoire National des Etiages Network (ONDE, *https://www.onde.eaufrance.fr/*) of the French Office for Biodiversity.

This dataset contains monthly observations between April and September of the flow state of headwater streams since 2012. The state of the temporary streams is classified by trained staff as either dry, standing, or flowing (Beaufort et al., 2018; Sauquet et al., 2021). The fourth state (trickling) was not used consistently across the network, and these observations were, therefore, merged with the flowing state. The 3351 ONDE sites are evenly distributed throughout France, except that there are fewer sites in the French Alps (Beaufort et al., 2018).

Of the 632 catchments with limited human influence, 199 catchments had three or more ONDE sites, and were therefore considered for this study. We excluded catchments with an unreasonable water balance (P>(Q + PET)) or >10% missing data, and very large catchments (>5000 km^2). This reduced the dataset to 92 catchments (Fig. 1 and Table S1). Only a few of these catchments are influenced by snowmelt (median mean catchment elevation: 277 m.a.s.l.; 7, 3, and 2 of the 92 catchments have a mean elevation above 1000, 1500, and 2000 m.a.s.l;, respectively, Table 1).

The median number of ONDE sites per catchment for the 92 catchments was five (average: 6; maximum: 21). The median catchment size for these ONDE sites was 21 km^2. For 87 % of the 445 ONDE sites in the 92 catchments, the local drainage area was less than 100 km^2, and for 46 %, it was less than 20 km^2. The ratio of the median drainage area of the ONDE sites and the total catchment area varied between 0.003 and 0.68 (median: 0.03; average: 0.06; 90th percentile: 0.18).

For 32 % of the 445 ONDE sites, the streams were always flowing when the observations were made. For only 5 % of the sites, the stream was flowing for less than 20 % of the observations. For 51 % of the sites, there was at least one observation of a dry stream between 2012 and 2020, and for 60 % at least one observation of standing water. For 8 % and 17 % of the sites there were no observations of standing water or a dry streambed, respectively. Based on these flow statistics, we assume that the ONDE sites are (mainly) located on intermittent streams that are connected to the groundwater table (i.e., no or few ephemeral streams), and that a change in the flow state for these streams implies a change in groundwater storage.

2.2. Flowiness definition

We used a lumped model and thus considered each catchment as one unit (i.e., we did not consider the spatial variability in catchment storage). Therefore, we derived a measure that represents the average flow state for all ONDE sites in a catchment. First, we converted the observed flow states to numeric values: dry: 0, standing: 0.5, and flowing: 1) and then calculated the average flow state for all ONDE sites in a catchment. We refer to this average value of the flow state as the *flowiness*. Since this flowiness was based on ordinal data, it provides only an indicator of catchment wetness but allows for a ranking of catchment storage states. Observations at the different ONDE sites in a catchment were not made on the same day and, therefore, we used a buffer of five days around the observations to calculate the flowiness. If at least 60 % of the ONDE sites were observed within this time window, the mean value of the flow state was calculated (see example in Fig. 2c and Fig. S1).

The range of flowiness values (i.e., the difference between the minimum and maximum flowiness) for each catchment varied between 0.1 and 1.0 (10th percentile: 0.20, median: 0.58, 90th percentile: 0.89, average: 0.55). Given the nature of the data, there were a lot of similar values of flowiness. On average, slightly more than half of the data points for a catchment were tied.

We also tested different approaches to determine the flowiness time series. We used a different value for the standing water class to indicate its closer proximity to a dry streambed (0.15), and merged the standing water and dry streambed class into a no-flow class. We also weighted the ONDE sites by their corresponding drainage area. For all of these methods, the resulting flowiness time series were highly correlated to the original flowiness data (Spearman rank correlation (r_s) > 0.9, except for two catchments) and the effects on the model results were minimal (Table S2). We, therefore, do not discuss these results in the main text.

2.3. Catchment characteristics

Several catchment characteristics were calculated for the 92 selected catchments (Table 1; Fig. S2). The mean catchment elevation was

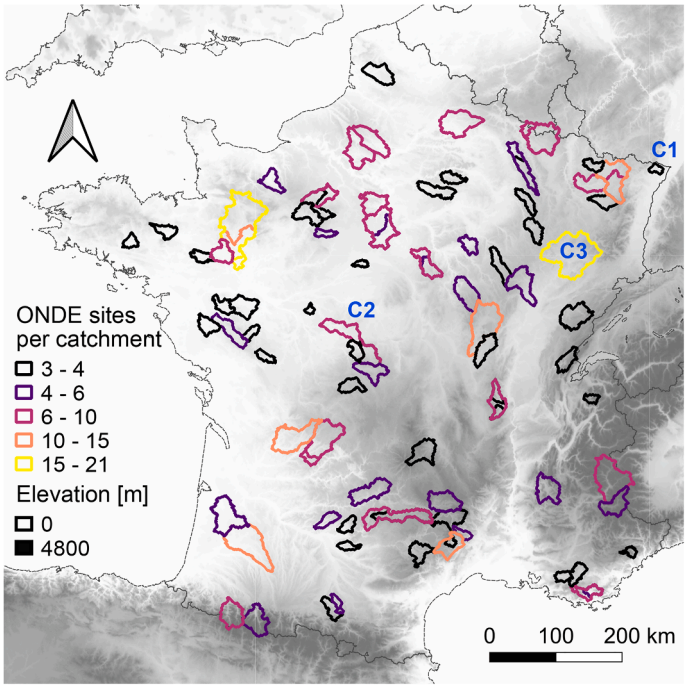


Fig. 1. Map showing the boundaries of the 92 selected catchments colour coded by the number of ONDE sites within each catchment (see Table S1 for a description of the catchments). The three selected catchments (C1, C2 and C3) for which the results are described in more detail in the text and shown in Figs. 2, 4, 5, and 7, are indicated with blue labels. The grey background shading represents the elevation (in m.a.s.l.).

Table 1
Main characteristics of the 92 selected catchments. For additional information, see Table S1 and Fig. S2. The short name is the label used in Fig. 9. The last column indicates if this catchment characteristic was used in the Principal Component Analysis (PCA) of the catchment characteristics.

Category	Catchment characteristic	Short name	Min	Median	Mean	Max	Used in PCA
Catchment	Catchment area [km^2]	Area	128	815	979	4166	Yes
	Mean catchment elevation [m.a.s.l.]	Elevation	77	277	393	2104	Yes
	Mean slope [°]	Slope	0.42	2.39	4.76	18.44	No
	Stream density [km^{-1}]	–	0.14	0.29	0.33	0.68	Yes
	Aridity Index (P/PET) [-]	Aridity	0.92	1.44	1.49	2.79	Yes
Response	Low flow [mm/d]	Q_{95}	0.0004	0.10	0.16	1.14	No
	High flow [mm/d]	Q_5	0.59	2.8	3.3	15.7	No
	Flashiness Index [-]	Flashiness	0.03	0.21	0.22	0.45	Yes
	Base Flow Index [-]	BFI	0.25	0.55	0.58	0.95	Yes
ONDE data set	Number of ONDE sites per catchment [-]	#ONDE	3	5	6	21	Yes
	Mean ONDE drainage area per catchment [km^2]	A_{mean}	3	21	43	818	Yes
	Mean ONDE drainage area as fraction of catchment area [-]	–	0.003	0.031	0.057	0.679	No
	Maximum fraction of ties [-]	–	0.20	0.61	0.62	0.98	No
	Coefficient of variation of flowiness (standard deviation/mean) [-]	COV Fl	0.01	0.18	0.19	0.28	Yes
	Spearman rank correlation between flowiness and discharge [-]	r_s	0.00	0.73	0.65	0.91	Yes

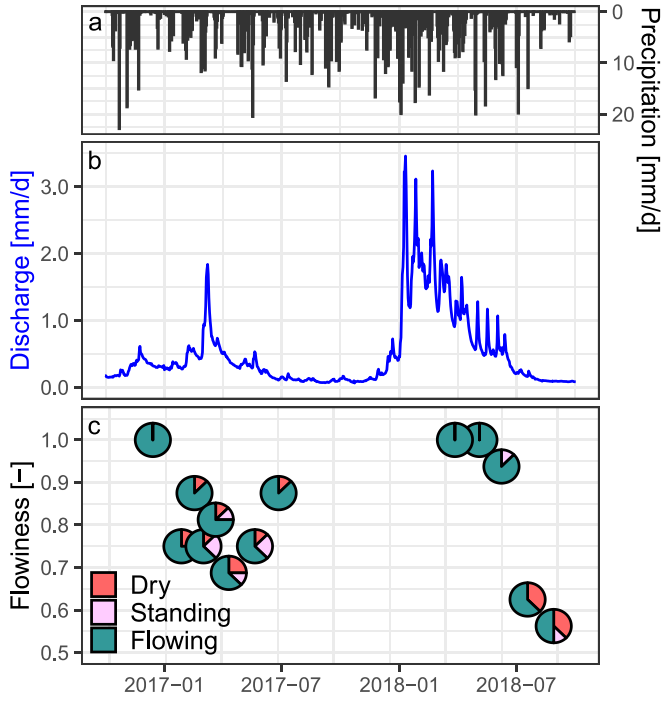


Fig. 2. Time series of a) precipitation, b) discharge, and c) flowiness based on observations at eight ONDE sites in catchment C2 (ID: K7312610). The pie charts show the distribution of flow states for each flowiness value. The Spearman rank correlation (r_s) between flowiness and discharge for this catchment is 0.87.

derived from a 25 m resolution Digital Elevation Model (Copernicus Land Monitoring Service, EU-DEM v1.1. https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1). The stream density was based on the ratio of the stream length (European Environment Agency, 2012) and the catchment area. For each catchment, we also determined several hydrological characteristics for the entire study period (2001–2020). The aridity index was calculated as the ratio of the mean annual precipitation and mean annual potential evaporation (P/E_{pot}). The base flow index (BFI) was used as a measure of the importance of groundwater flow and was calculated according to the method of the Institute of Hydrology (1980). The Richards-Baker Flashiness Index (Baker et al., 2004) was used as a measure of the responsiveness of the catchment to precipitation. A catchment with a high flashiness index responds rapidly to rainfall, while a catchment with a low flashiness index has more stable streamflow. A low flashiness index corresponds to a high baseflow index (r_s = -0.90, Fig. S3). Finally, the discharge that was exceeded for 5 and 95 percent of the time during the 2001–2020 period, were used to characterize the high and low flows, respectively.

2.4. Model

2.4.1. The HBV Model

The HBV model (Hydrologiska Byråns Vattenbalansavdelning; Bergström (1992), Lindström et al. (1997)) is a lumped bucket-type model. Discharge is simulated based on the observed precipitation, air temperature, and estimates of the long-term monthly mean potential evaporation. The model consists of different routines representing snow-, soil-, and groundwater storage and stream routing. The groundwater storage is subdivided into an upper (SUZ) and lower zone (SLZ), representing a faster- and a slower-draining reservoir. In this study, we used the model implementation HBV-light (Seibert and Vis, 2012), version 4.0.0.25, with the standard model structure with 13 parameters (Table 2) and a daily time step.

2.4.2. Model setup

The catchments were divided into 100 m elevation zones. For model calibration, we used a genetic optimization algorithm GAP (Seibert, 2000) consisting of 5,000 model runs that were repeated in 100 independent calibration trials, each resulting in one optimized parameter set. The calibration period lasted from 01.10.2010 to 30.09.2019, with a one-year (hydrological year 2010) warm-up period. The validation period lasted from 01.10.1999 to 30.09.2010, with also a one-year warm-up period. The calibration and validation periods were chosen based on the availability of the ONDE data to calculate the flowiness data for the calibration. The ONDE data were not available before the summer of 2012, and therefore the model was calibrated on the later period and validated on the earlier period.

2.4.3. Data sets used for model calibration

From the daily discharge time series, we created a dataset with monthly (i.e., one daily value per month) and seasonal (i.e., one daily value per season) discharge data to determine if the value of the flowiness data for model calibration depends on the amount of streamflow data that are available for calibration, i.e., if flowiness data are more useful in situations where there are only limited streamflow data. To obtain a realistic monthly time series of discharge, the discharge on the first day of each month was selected. For the seasonal time series, only the discharge on the first day of March, June, September, and December were selected. The model was calibrated for each of these three datasets, with and without the inclusion of the flowiness data (see section 2.4.4). We performed all model calibrations also with streamflow data for the 15th day of the month but did not find an apparent influence of the

Table 2
HBV model parameters and parameter ranges used for the calibration (modified from Seibert and Vis (2012)).

Routine	Parameter	Description	Unit	Lower value	Upper value
Snow Routine	TT	Threshold temperature	°C	−2.0	2.5
	CFMAX	Degree-day factor	mm/(°C*d)	0.5	10
	SFCF	Snowfall correction factor	–	0.5	1.2
	CFR	Refreezing coefficient	–	0	0.1
	CWH	Water holding capacity	–	0	0.2
Soil Moisture Routine	FC	Maximum soil moisture storage	mm	100	550
	LP	Relative threshold for reduction of evaporation	–	0.3	1.0
	BETA	Shape coefficient for the computation of recharge	–	1.0	5.0
Response Routine	PERC	Maximum flow from upper to lower groundwater box	mm/d	0	4.0
	Alpha	Non-linearity coefficient	–	0	5.0
	K1	Recession coefficient (upper groudwater box)	1/d	0.01	0.2
	K2	Recession coefficient (lower groundwater box)	1/d	0.001	0.1
Routing Routine	MAXBAS	Length of triangular weighting function	d	1.0	5.0

selection of the day on the overall model results. Therefore, only the results for the calibration with the discharge on the first day of the month are shown in the manuscript.

To test the value of temporary stream data for cases where only stream-level data but no discharge data are available, we followed the approach of Etter et al., (2020a) and Seibert and Vis (2016). We did not convert the streamflow time series into a stream-level time series (nor did we have any stream-level data). Instead, we used a different objective function for model calibration (i.e., the Spearman rank correlation (r_s); see section 2.4.4). This approach assumes that discharge and stream-level are strictly monotonic related to each other (i.e., when discharge is high, the stream-level is also high). The advantage of this approach is that it does not require any information on the rating curve to convert discharge to stream-level (or vice versa). For the monthly and seasonal stream-level data, we used the same measurement dates as for the discharge.

2.4.4. Model calibration

For each catchment, the model was calibrated for each of the six discharge or stream-level data sets, with and without the flowiness data (Table 3). Additionally, a lower benchmark was created with 10,000 Monte Carlo runs of randomly selected parameters within the parameter ranges (Table 2).

For the calibrations without flowiness data, the model was calibrated by optimizing the fit between the observed and simulated streamflow. For the three discharge scenarios (daily, one daily value per month, or one daily value per season), we used the Kling-Gupta efficiency (KGE; Gupta et al., 2009). For the three stream-level scenarios, we maximized the Spearman rank correlation coefficient (r_s) between the observed and simulated streamflow (Table 3). Note that the observed streamflow can be used for the observed stream-level due to using the Spearman rank correlation as objective function (see section 2.4.3).

For the calibrations with flowiness data, we used a multi-criteria calibration. Again, the simulated discharge was compared to the observed discharge using the KGE or r_s (for the discharge or stream-level scenario, respectively). In addition, the time series of the sum of the simulated storage in the upper and lower groundwater reservoir (named *total groundwater storage* from hereon) was compared to the time series of the flowiness using r_s (Table 3). The objective functions for the discharge and the storage were weighted equally.

Initial tests indicated that the calibration results were more plausible when the flowiness was compared with the total groundwater storage than when it was compared to only one of the two individual groundwater reservoirs. When the fit between the flowiness and storage in only one of the two groundwater reservoirs was optimized, we obtained unrealistic results because the storage in the other groundwater reservoir would become very large and no longer change seasonally. Initial tests where we calibrated the model only with flowiness data (i.e., no discharge or stream-level data) led to a poor model performance (Fig. S4), regardless of which measure was used to describe the model fit. Tests with model calibration based on minimizing the mean absolute relative error (MARE) did not provide any additional insights beyond those obtained for model calibration based on the KGE (and the results are, therefore, not shown).

2.4.5. Evaluation of the value of flowiness data for model calibration

The performance of the different calibrations was assessed based on the simulated discharge for the validation period. We simulated the discharge for the 100 calibrated parameter sets for the validation period and determined the mean discharge for each day to obtain the time series of the ensemble mean discharge for the validation period. This was done for each data scenario and catchment. We compared the time series of this ensemble mean discharge to the observed discharge and

Table 3
Overview of the model scenarios and calibration criteria (KGE = Kling-Gupta Efficiency, r_s = Spearman rank correlation coefficient). The model was calibrated by maximizing the sum of the two objective functions. The lower benchmark, based on 10,000 Monte Carlo simulations, is not included in the table. Note that monthly refers to one value per month and seasonal to one value per season (and thus not the monthly or seasonal average discharge).

	Data scenario		Objective function 1	Objective function 2
1	Daily discharge	-	KGE	-
2	Daily discharge	Irregular flowiness	KGE	r_s
3	Monthly discharge	–	KGE	–
4	Monthly discharge	Irregular flowiness	KGE	r_s
5	Seasonal discharge	–	KGE	–
6	Seasonal discharge	Irregular flowiness	KGE	r_s
7	Daily stream-level	–	r_s	–
8	Daily stream-level	Irregular flowiness	r_s	r_s
9	Monthly stream-level	–	r_s	–
10	Monthly stream-level	Irregular flowiness	r_s	r_s
11	Seasonal stream-level	–	r_s	–
12	Seasonal stream-level	Irregular flowiness	r_s	r_s

used the KGE to determine the overall model performance. To compare this overall model performance for the different catchments, we scaled the KGE values for the validation period by an upper and lower benchmark (cf. Seibert et al., 2018). For the upper benchmark, we used the KGE value for the calibration period when the model was calibrated with daily discharge data (i.e., we used the KGE value for the best possible fit). For the lower benchmark, we used the KGE value for the ensemble mean discharge calculated from 10,000 Monte Carlo runs with randomly selected parameters within the parameter ranges (Table 2). We refer to this scaled model performance as the *relative model performance* E_{rel}. A simulation with a relative model performance (E_{rel}) of one is as good as the simulation of the streamflow for the calibration period when it is calibrated with daily discharge data. A value of zero indicates a model performance similar to the uncalibrated (i.e., uninformed) model for the validation period. A value less than zero indicates a model fit that is worse than the uncalibrated model for the validation period.

To assess the model performance for the low-flow simulations, we used the root mean squared error of the simulated minimum seven-day mean discharge for each hydrological year (R_{Q7_min}). We specifically looked at the model performance for the low-flow period because the ONDE data were collected during the summer low-flow period. We expected the flowiness data to affect the simulated groundwater storage mainly during this period and, thus, the low-flow simulations during this period (Beaufort et al., 2018). We did not scale R_{Q7_min}, because it was not used as an objective function in the model calibration.

We determined the difference in E_{rel} and R_{Q7_min} for the calibrations with and without the flowiness data (ΔE_{rel} and ΔR_{Q7_min}) for each of the discharge and stream-level scenarios. Positive values of ΔE_{rel} and negative values of ΔR_{Q7_min} indicate an improvement in model performance when flowiness data are included in model calibration (i.e., a higher value for the KGE or a lower value for the R_{Q7_min}). In the following, we multiplied ΔR_{Q7_min} by minus one for easier comparisons. We used the Wilcoxon matched-pairs signed-rank test (*stats* package; R Studio Version 4.2.2) to assess if the effect of the flowiness data on model performance (E_{rel} and R_{Q7_min}) was statistically significant.

We also analysed the effect of including flowiness data in model calibration on the range of the low-flow simulations. We calculated the range in R_{Q7_min} for the 100 simulations per catchment. Again, we used the Wilcoxon matched-pairs signed-rank test to assess if the effect was statistically significant.

Finally, we analysed the effect of including the flowiness data in the calibration on parameter uncertainty. First, the calibrated parameter values were scaled by the upper and lower boundary values (Table 2), ranging between zero and one. Then, we determined the difference between the 5th and the 95th-percentile of the 100 calibrated parameter values for each catchment and data scenario. We used the Wilcoxon matched-pairs signed-rank test to determine if the inclusion of the flowiness data in model calibration affected the spread of the parameter values. The Wilcoxon test was used instead of a *t*-test for all significance tests because neither the values of the model performance nor the parameter ranges were normally distributed.

2.5. Evaluation of the characteristics of the catchments for which flowiness data improved model performance

We calculated the spatial autocorrelation of the changes in model performance (ΔE_{rel} and ΔR_{Q7_min}) and range in parameter values with the Moran's I test (Bivand and Wong, 2018). The catchment centroid was used to select three, four, five, six, and seven neighbouring catchments weighted by distance. Afterwards, the correlation of the changes in model performance for the neighbouring catchments was determined.

To further assess the influence of catchment characteristics on the value of flowiness data for calibration, we related the change in model performance to catchment characteristics using Spearman rank correlation. The change in model performance (ΔE_{rel}, ΔR_{Q7_min}, parameter range) was also related to the first two principal components for ten characteristics that describe the catchment itself, its hydrological response, and the ONDE dataset (Table 1; Fig. S5) (*stats* package, R Studio Version 4.2.2). We related the change in model performance to the principal components because we did not expect that the value of temporary stream data for model calculation depends on only one catchment characteristic. Furthermore, many of the catchment characteristics are correlated (see Table 1 and Fig. S3).

3. Results

3.1. Model calibration with discharge data only

The calibration with daily discharge data resulted in good model fits. The median KGE for the ensemble mean for the calibration period was 0.92 (range: 0.67 to 0.98, Fig. S6). For 15 of the 92 catchments (16 %) the KGE was less than 0.85 and for 12 catchments (13 %) it was less than 0.70. The results for the validation period were slightly lower (median KGE for the ensemble mean discharge: 0.85, range: 0.35 to 0.96). The E_{rel} values varied between −2.44 and 1.20 (median: 0.76) (Fig. 3a). The median KGE of the lower benchmark was 0.60 (range: 0.02 to 0.88). The difference between the lower and upper benchmark varied between 0.05 and 0.71 (median: 0.28).

Calibration with only one discharge measurement per month (monthly data) led to lower but still reasonable model fits. The median KGE for the ensemble means for the validation period was 0.76 (range: 0.01 to 0.96; Fig. S6). For 24 of the 92 catchments (26 %) the KGE was less than 0.70. The median E_{rel} for the validation period was 0.61 (range: −2.85 to 1.09). Only for 10 of the 92 catchments (11 %) was E_{rel} less than zero, i.e., the validation results were worse than the lower benchmark (Fig. 3a).

The decrease in model performance was larger when using only one discharge measurement per season (seasonal data). The median KGE of the ensemble mean for the validation period did not change much compared to the results for the monthly data (0.74; range: −0.22 to 0.92; Fig. S6). For 44 catchments (48 %) was the KGE less than 0.70. The median E_{rel} was 0.31 (range: −5.16 to 0.94). For 28 of the 92 catchments (30 %) was E_{rel} less than zero (Fig. 3a).

3.2. Model calibration if only stream-level data were available

As expected, calibration for the stream-level scenario resulted in poorer model fits than calibration with discharge data (Fig. 3a). The median KGE for the validation period was 0.69, 0.67, and 0.66 (range: −0.08 to 0.94, 0.05 to 0.93, −0.27 to 0.93) for the calibration for the daily, monthly, and seasonal stream-level data scenarios, respectively (Fig. S6). For 46, 46, and 56 catchments, was the median KGE for the validation period less than 0.70 for the calibration with the daily, monthly, and seasonal stream-level data scenarios, respectively. The median E_{rel} values were 0.33, 0.26, and 0.12 (range: −2.02 to 1.41, −2.05 to 0.92, −5.56 to 0.77), respectively. For 18 (20 %), 20 (22 %), and 43 (46 %) of the 92 catchments, the calibration for the daily, monthly, and seasonal stream-level scenarios resulted in a model performance for the validation period that was worse than the lower benchmark ($E_{rel} < 0$; Fig. 3a).

3.3. Model calibration results when also using flowiness data

3.3.1. Ensemble mean streamflow

Using the flowiness data in model calibration had a minimal effect on the overall model performance (Fig. 3). The change in the model performance for the ensemble mean discharge for the validation period was statistically significant only for the scenarios with daily discharge and seasonal stream-level data (p-values: 0.003 and 0.039, respectively) but the differences were so small (median ΔE_{rel}: −0.01; range: −0.50 to 0.10) that they are in practice not important. For all other scenarios, the change in E_{rel} due to the use of the flowiness data in model calibration

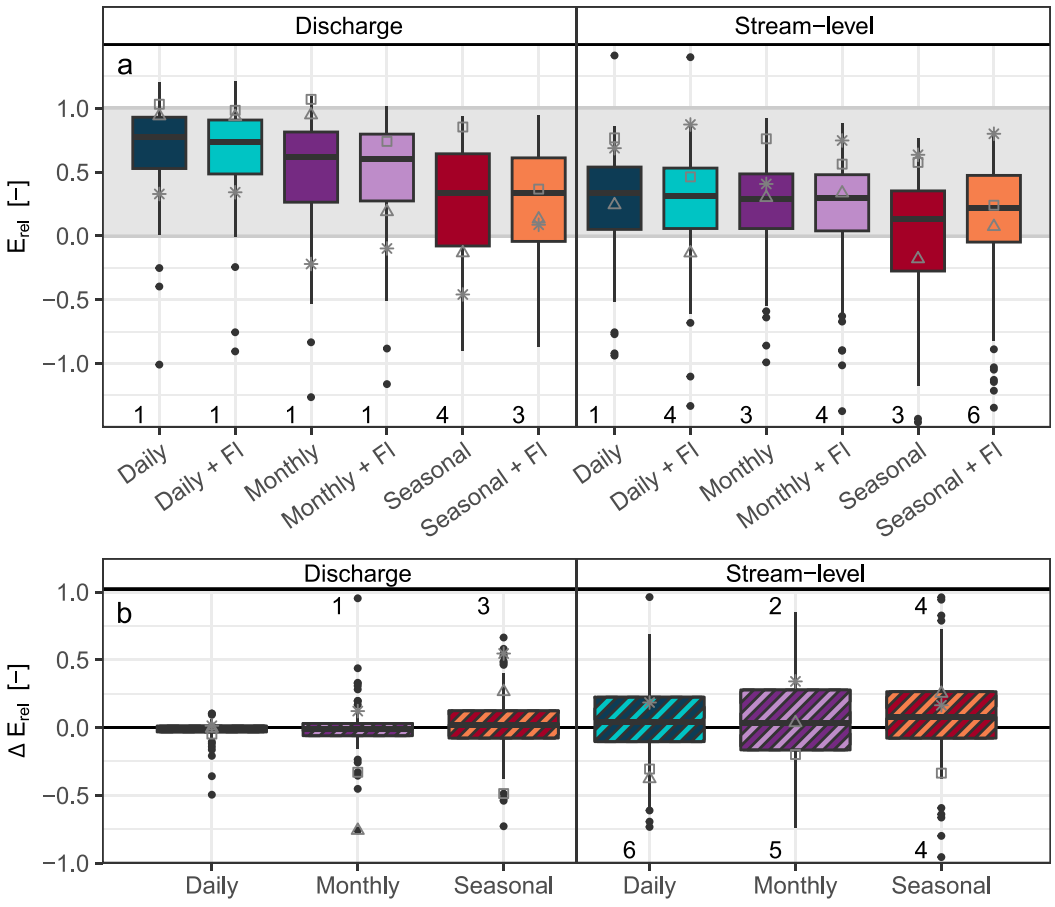


Fig.3. Boxplots of a) the relative performance (E_{rel}) of the ensemble mean discharge for the validation period for the discharge and stream-level scenarios with different temporal resolutions, with and without flowiness (Fl) data, and b) the difference in the relative model performance (ΔE_{rel}) for the models calibrated with and without flowiness data, where positive values indicate an improvement in model performance when flowiness data are used for model calibration. The box represents the interquartile range, and the thick line the median. The whiskers extend to 1.5 times the interquartile range. The numbers above and below the boxplots show the number of outliers outside the displayed y-axis range. The grey square, triangle, and star represent the values for catchments C1, C2, and C3, respectively. The grey shading in (a) indicates a model performance between the upper and lower benchmark. The different temporal resolutions are: daily, monthly (a daily value on the first day of the month) and seasonal (a daily value on the first of March, June, September, and December). For the results of the calibration with data taken on the 15th day of the month, see Fig. S7.

(ΔE_{rel}) was not statistically significant ($p = 0.06$ for the scenarios with seasonal discharge data, $p > 0.25$ for all other scenarios).

Using flowiness data in the calibration could improve or worsen overall model performance for the validation period, with the differences being larger for the stream-level scenarios (Fig. 3b; Fig. S8a-b). Of the 552 comparisons (92 catchments times six data scenarios), model performance improved ($\Delta E_{rel} > 0.01$) 261 times, worsened ($\Delta E_{rel} <$ -0.01) 241 times, and did not change (-0.01 $< \Delta E_{rel} >$ 0.01) 50 times due to the inclusion of flowiness data in the calibration.

When the model performance was worse than the lower benchmark ($E_{rel} < 0$), adding flowiness data in the calibration was more likely to improve model performance than when the initial model performance was already good. The Spearman rank correlation between ΔE_{rel} and E_{rel} for the calibrations without flowiness data was −0.39. Adding flowiness data in the calibration improved the model performance ($\Delta E_{rel} > 0.01$) for 91 of the 123 comparisons (74 %), for which the calibration without the flowiness data was worse than the lower benchmark ($E_{rel} < 0$). On the contrary, for 213 out of 429 comparisons (50 %) for which the calibration without the flowiness data was better than the lower benchmark ($E_{rel} > 0$), model performance decreased when using flowiness data in the calibration ($\Delta E_{rel} <$ -0.01) (median ΔE_{rel}: −0.10; range: −4.66 to −0.01).

3.3.2. *Example time series for three catchments*

To better show the effects of the inclusion of flowiness data in model calibration on the simulated discharge, we present the observed and simulated discharge for the driest hydrological year of the validation period for three catchments: Catchments C1 (station ID: A3832010; catchment area: 204 km^2), C2 (K7312610; 1706 km^2), and C3 (M1034020; 267 km^2). The addition of flowiness data did not noticeably affect the simulations with daily discharge for C1 and C2 (Fig. 4a-d). For catchment C3, it reduced the uncertainty in the low-flow simulations (Fig. 4e-f).

The effect of including the flowiness data in the calibration was larger for the seasonal stream-level scenario. For catchment C1, the use of flowiness data in the calibration led to a poorer overall simulation (ΔE_{rel} = -0.36) (Fig. 3b) and poorer low-flow simulations during the validation period (i.e., the R_{Q7_min} for the ensemble mean increased from 0.21 to 0.34 mm/d) (Fig. 5a-b). For catchment C2, the overall simulation performance improved ($\Delta E_{rel} = 0.16$): the simulation of the highest peak flow deteriorated, but the other high flow events were simulated better (Fig. 5c). Using flowiness data in the calibration also eliminated the underestimation of the low flows, but caused an overestimation of the minimum discharge (Fig. 5c-d), resulting in an increase in the R_{Q7_min} from 0.02 to 0.16 mm/d. For catchment C3, the overall model performance improved (ΔE_{rel} = 0.25) and the lowest flows were underestimated less when flowiness data were used in the calibration. The

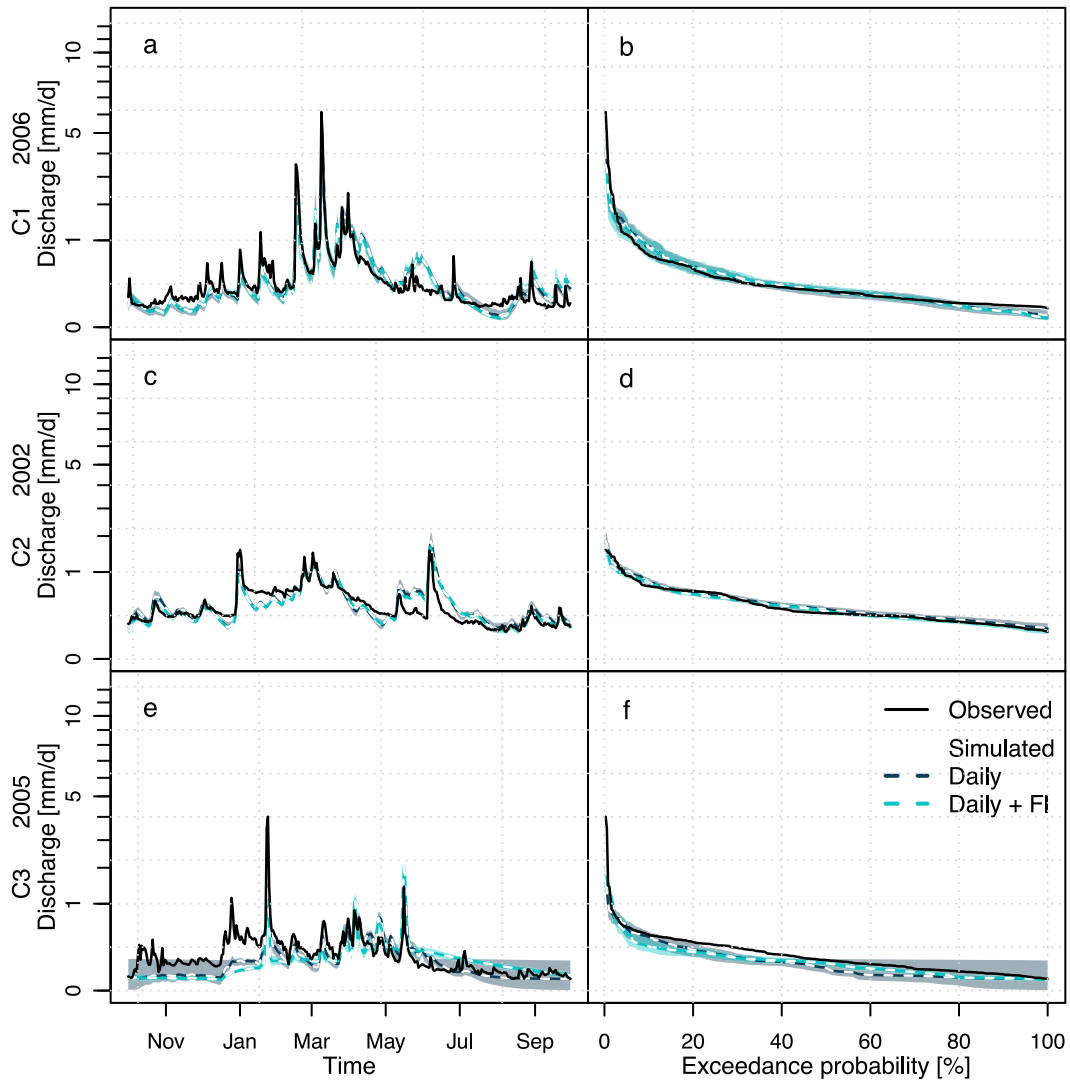


Fig.4. Time series of the observed and simulated discharge (left) and flow duration curves (right) when the model was calibrated with daily discharge data or daily discharge data and flowiness (Fl) data for the driest year of the validation period for three selected catchments (C1: top; C2: middle; C3: bottom). The shaded area shows the uncertainty bands (5th to 95th percentile) for the 100 model calibrations; the thick dashed lines indicate the ensemble means. Note that the y-axis shows the discharge on a square-root scale to better visualize the low flows. For catchments C1 and C2, the uncertainty bands are very narrow and partly overlap. For the time series of the flowiness data and groundwater simulations for the driest year of the calibration and validation period, see Fig. S9.

discharge in fall was also less overestimated, but the small responses between May and September were overestimated (Fig. 5e-f). As a result, the R_{Q7_min} of the ensemble mean discharge increased from 0.02 to 0.08 mm/d when flowiness data were used in the calibration.

3.3.3. *Low-flow simulations and uncertainty*

As suggested by the time series for the selected catchments (Fig. 4 and Fig. 5), the low-flow simulations did not necessarily improve by including flowiness data in the calibration (Fig. S10). The difference in the R_{Q7_min} of the ensemble means for the calibrations with and without flowiness data for the 92 catchments was not significant for any of the data scenarios (Fig. 6a: $p > 0.10$). However, the addition of the flowiness data in the calibration with discharge data reduced the uncertainty of the low-flow simulations (Fig. 6b). For 21 of the 92 catchments (23 %) the range in the R_{Q7_min} for the 100 parameter sets decreased by more than 0.01 mm/d when flowiness data were used in addition to the daily discharge data. For the monthly and seasonal discharge data scenarios, this was the case for 27 (29 %) and 48 (52 %) catchments, respectively. However, the difference in the median values of the ranges of R_{Q7_min} was only significant for the seasonal discharge data scenario ($p < 0.01$). For the stream-level data scenarios, the use of flowiness data did not have an apparent effect on the range of the R_{Q7_min} ($p > 0.08$) (Fig. 6b; Fig. S8d).

3.3.4. *Parameter uncertainty*

The median values of the scaled parameter ranges (5th – 95th percentile) for all 92 catchments are given for each of the twelve calibration scenarios in Fig. 7. The parameters of the snow routine were most uncertain, but this is not surprising as few catchments are influenced by snowmelt. The FC and Alpha parameters were the least uncertain. We expected that the inclusion of the flowiness data would mainly affect the parameters of the response routine (PERC, Alpha, K1 and K2) and the routing routine (MAXBAS) because the flowiness was compared to the groundwater storage (cf. Demirel et al., 2019). However, the parameter ranges of FC, LP, K2 and MAXBAS changed significantly for all discharge scenarios when flowiness data were included. For FC, LP and K2 the median parameter range decreased, and these parameters thus became more certain when flowiness data were used in the calibration. According to the Wilcoxon test, these differences were significant ($p < 0.04$ for FC, and $p < 0.01$ for LP, K2 and MAXBAS).

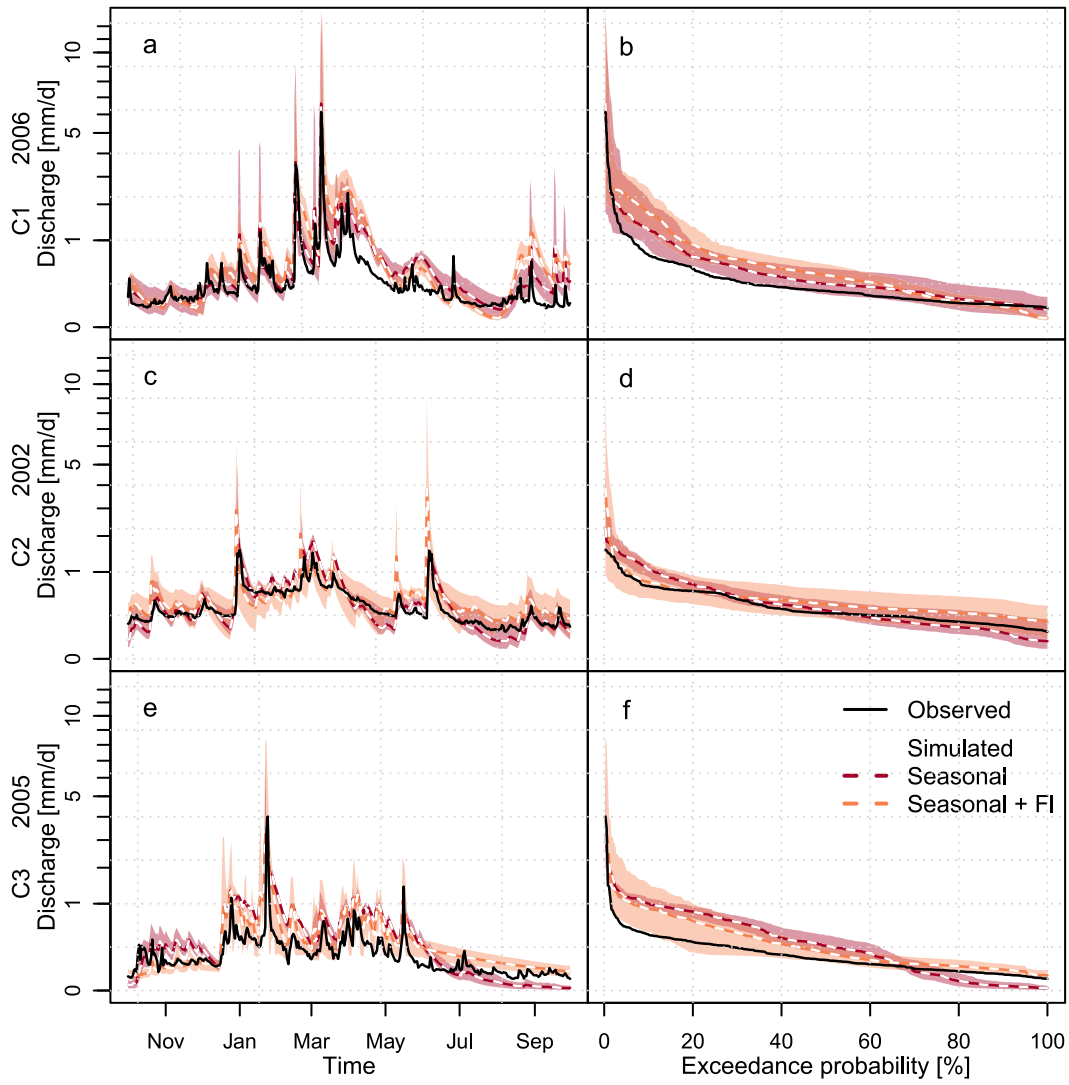


Fig. 5. Time series of the observed and simulated discharge (left) and flow duration curves (right) for the seasonal stream-level scenario (one measurement per season, on the 1st of March, June, September, and December), with and without the flowiness (Fl) data for the driest year of the validation period for the three selected catchments (C1: top; C2: middle; C3: bottom). The shading shows the uncertainty bands (5th to 95th percentile) for the 100 model calibrations; the dashed lines indicate the ensemble means. Note that the y-axis shows the discharge on a square-root scale to visualize the low flows better.

For the 21 times that the parameter ranges changed significantly by including flowiness data for the stream-level scenarios, they became more uncertain 17 times (Fig. 7). In particular, the parameter ranges of BETA and PERC were affected by including flowiness data. For the daily and monthly stream-level scenarios, the change was significant for nine out of the thirteen parameters. For the seasonal stream-level scenario, including the flowiness data in model calibration affected the parameter range for only three parameters (LP, BETA, PERC).

3.3.5. *Uncertainty of the recession coefficient K2 parameter*

Because the use of flowiness data in model calibration decreased the uncertainty for parameter K2 the most for all discharge data scenarios (Fig. 7), we looked more closely at the effect of including flowiness data in the calibration on the calibrated K2 parameter values. Parameter K2 determines the outflow of the (s)lower groundwater reservoir (SLZ) as a function of the storage in this reservoir. This parameter mainly affects the simulation of low-flow conditions. For 72 of the 92 catchments, the parameter range (5th – 95th percentile) was reduced when flowiness data were used in addition to daily discharge data (Fig. S8e). It increased for 12 catchments. For eight catchments, the range of the K2 parameter values did not change.

For the three exemplarily catchments discussed earlier, the calibrated parameter range for K2 decreased differently (Fig. 8). For catchment C1, the range became smaller, but the calibrated values were at the upper edge of the predefined parameter range. However, the median simulated water level in the (s)lower groundwater reservoir and the overall- and low-flow model performance (KGE and R_{Q7_min}) were not considerably influenced by this change in the calibrated value of the K2 parameter (Fig. 8a,d,g). The calibrated parameter value was at the upper edge of the predefined parameter range for 16 of the 92 catchments when flowiness data were used together with daily discharge data for the calibration. This was also the case for two catchments when only daily discharge data were used.

For catchment C2, the parameter range of K2 was smaller when flowiness data were included in model calibration. Including flowiness data in the calibration avoided parameter sets with small values for K2 that lead to very large (median) water levels in the (s)lower groundwater reservoir. The very small K2 values and very high median water levels (i.e., SLZ values) were even more pronounced for catchment C3. The changes in the calibrated values for the K2 parameters for catchments C2 and C3 affected the total simulated storage, as well as the dynamic storage (Fig. S9). The uncertainty (5th – 95th percentile) of the

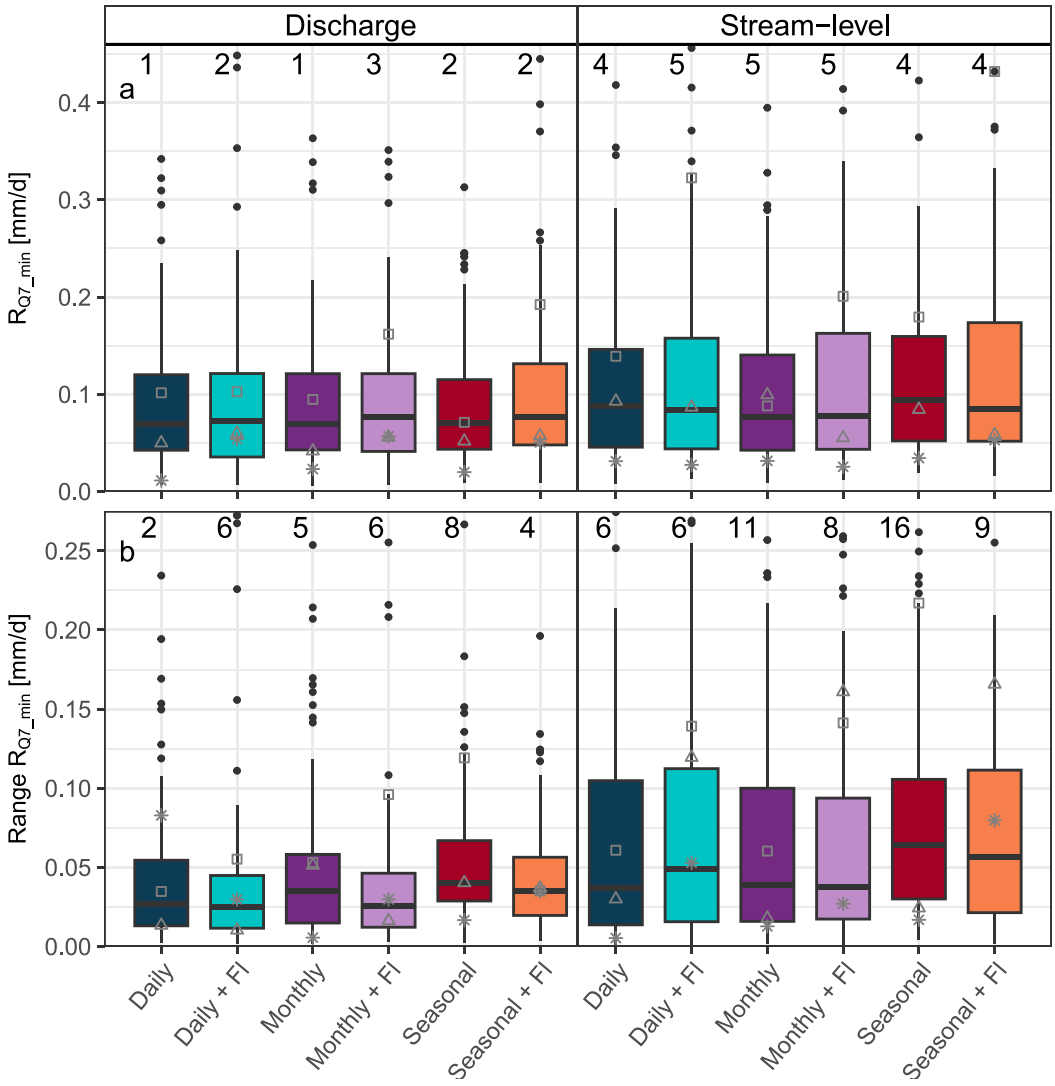


Fig. 6. Boxplots of the a) root mean squared error of the seven-day minimum average flow (R_{Q7_min}) for the ensemble mean for the validation period when the model was calibrated with the different datasets and b) the 5th to 95th percentile range of R_{Q7_min} (for the 100 calibration runs) for each of the 92 catchments. For comparison, the median R_{Q7_min} for the lower benchmark was 0.14 mm/d (range: 0.03 to 0.97 mm/d) and the median R_{Q7_min} of the upper benchmark was 0.07 mm/d (range: 0.00 to 0.38 mm/d). The grey square, triangle and star represent the values for catchments C1, C2, and C3, respectively. The numbers above the boxplots represent the number of outliers outside the displayed y-axis range. The differences between the R_{Q7_min} for the ensemble mean for the model calibrated with and without flowiness data are shown in Fig. S10.

average dynamic storage for the validation period was reduced from 17 mm to 7 mm for catchment C3 when flowiness data was used in the calibration with daily discharge data. For catchment C1, it remained unchanged at 7 mm, and for catchment C2 it reduced only slightly (from 6 mm to 5 mm). For catchment C3 the value of the objective function (KGE) increased slightly when very small values for parameter K2 were chosen (Fig. 8c), but this led to a very poor model performance for low flows (Fig. 8i). Not surprisingly, the Spearman rank correlation coefficient between flowiness data and groundwater level (SUZ + SLZ) in the calibration period increased for all data scenarios when flowiness data were used in the calibration (Fig. S11-Fig. S13).

When using only discharge data for calibration, there was at least one parameter set with K2 values at the lower edge of the parameter range for 40 of the 92 catchments. This was solved when flowiness data were used in the calibration for 34 catchments. For only five catchments were there more calibrated parameter sets with small K2 values when flowiness data were used in model calibration than for the calibration without flowiness data. In other words, for eleven catchments the calibrated parameter sets still included small, optimized values for the K2 parameter when flowiness data were included in model calibration.

Although the inclusion of flowiness data in calibration also changed the uncertainty of the FC and LP parameters, there was no clear pattern, and the parameter values became smaller or bigger when including flowiness data in the calibration. The calibrated parameter values mainly were far away from the parameter bounds, except for 13 catchments for FC and 20 catchments for LP. Adding flowiness data to the calibration did not lead to any changes regarding parameter values near these boundaries. The parameters FC and LP are highly dependent on each other. It is, thus, difficult to unravel the effects of the flowiness data on these parameters.

3.4. *Influence of catchment characteristics on the value of flowiness data*

The change in overall model performance due to the inclusion of flowiness data in the calibration (ΔE_{rel}) was spatially autocorrelated only for the daily discharge scenario ($p < 0.002$, for all tests with three to seven neighbours), but these changes were very small (see Fig. 3). The change in model performance for the low-flow simulations (ΔR_{Q7_min})

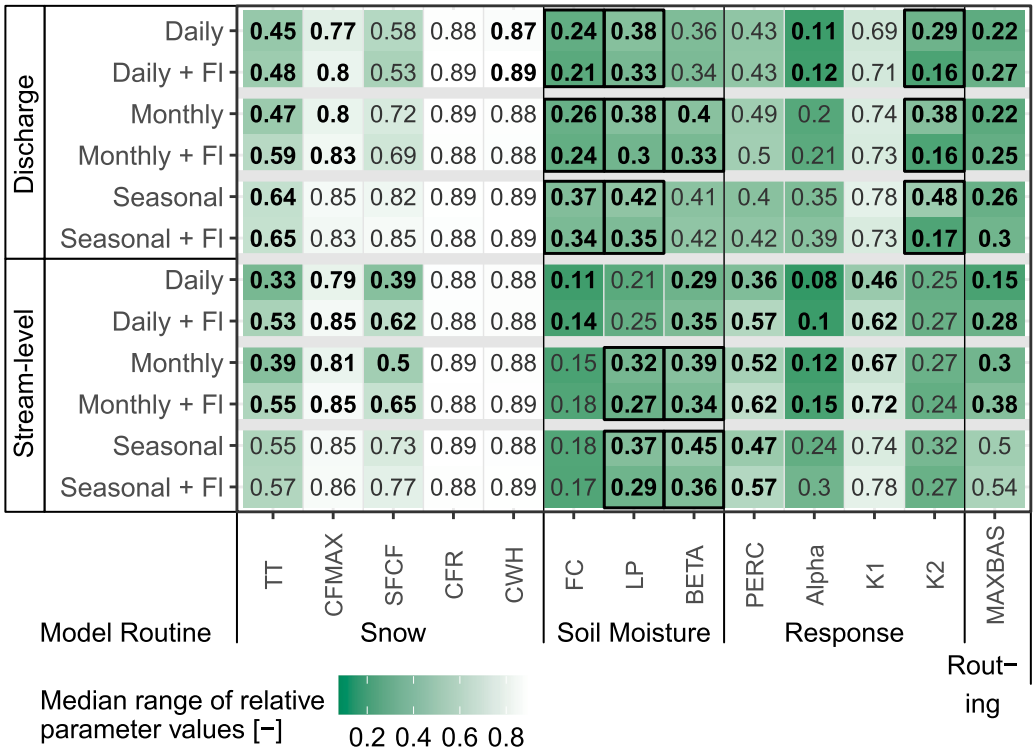


Fig. 7. The median value of the 5th to 95th percentile range of the calibrated model parameters for all data scenarios. A darker green shading indicates a smaller range (i.e., less parameter uncertainty). The parameters for which the use of flowiness (Fl) data in the calibration led to a statistically significant change in the parameter range are printed in bold. The parameters for which the median range became smaller when including flowiness data are highlighted with a black rectangle. For a description of the parameter values, see Table 2. The different temporal resolutions are: daily, monthly (one value per month) and seasonal (one value per season).

and the change in the range for parameter K2 were not spatially autocorrelated ($p > 0.18$) (Fig. S8c-d).

For a couple of catchments with high stream densities and small mean ONDE drainage areas, the low-flow predictions worsened by using flowiness data for the calibration ($\Delta R_{Q7_min} < 0$), but the overall correlations with catchment characteristics were poor. The change in overall model performance (ΔE_{rel}), the performance for the low-flow simulations (ΔR_{Q7_min}), and the range for parameter K2 were not well correlated with any of the catchment characteristics ($|r_s| < 0.36$; Table S3). The changes were also not well correlated to the principal components of the catchment characteristics ($|r_s| < 0.29$; Table S3). There was no clear pattern in the improvement or decline in model performance across the principal component space either (Fig. 9), except that the low-flow simulations became worse by adding flowiness data for the seasonal stream-level scenario for catchments with negative values for the two dimensions of the principal component analysis (PCA) (Fig. 9c-d). The first dimension describes the dynamics of the catchments and their size, and the second dimension describes the topography and climate (Fig. S5).

4. Discussion

4.1. Effect of flowiness data on the discharge simulations

Multi-criteria (or multi-data) model calibration leads to a (slightly) worse overall model fit for the calibration period than calibration on the discharge data alone because the model is no longer optimized to only fit the discharge data (cf. Beldring, 2002; Finger et al., 2015; Mostafaie et al., 2018; Parajka et al., 2007; Pelletier and Andréassian, 2022; Schaefli and Huss, 2011; Seibert, 2000). However, multi-criteria calibration is expected to lead to a more robust model that better represents the hydrological processes in the catchment. Therefore, the model fit for the validation period can be lower, higher, or nearly the same as for the calibration based on discharge data alone. It was, thus, not surprising that there was no clear improvement in the overall model performance for the ensemble mean when flowiness data were used in the calibration of the HBV model together with the daily discharge data. This was, for example, also reported by Pelletier and Andréassian (2022), who found that the combined use of monthly groundwater level data and daily discharge data in the calibration of the GR6J model for a set of catchments in France (partly overlapping with the ones in this study) did not improve the overall discharge simulation.

However, if the discharge simulation for the validation period was poor when only discharge or stream-level data were used in model calibration, the model performance often improved when flowiness data were used in the calibration as well. Thus, it seems that flowiness data can help to avoid poor model fits but not improve an already good model. This suggests that for most catchments discharge data or water level data are already so informative for model calibration that even if there is only one value per month or season (cf. Etter et al., 2020a; Jian et al., 2017; Seibert and Vis, 2016; van Meerveld et al., 2017; Weeser et al., 2019) the use of additional data, such as temporary stream observations, does not improve the overall model performance any further. Possible other reasons for the lack of a clear improvement in model performance for the majority of the catchments could be the choice of the model or a bias in the observed data.

Adding flowiness data to the calibration did affect the low-flow simulations. For 57 % of the catchments, the use of flowiness data in combination with daily discharge data improved the simulation of the minimum seven-day average flow. For 51 % of the catchments, the uncertainty of low-flow predictions decreased. The changes in low-flow uncertainties were mainly caused by a more well-defined value of

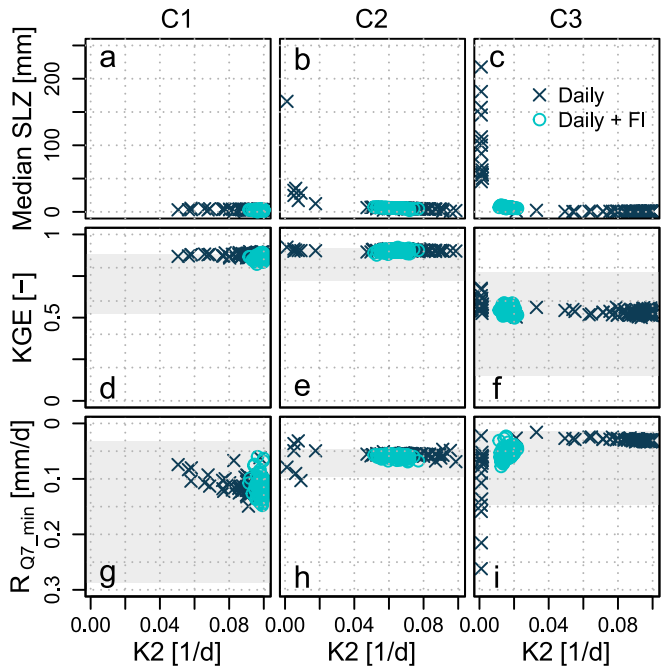


Fig. 8. Relation between the calibrated value of the K2 parameter and a-c) the median storage in the (s)lower groundwater reservoir (SLZ), d-f) the Kling-Gupta efficiency for the validation period (KGE), and g-i) the root mean squared error for the minimum seven-day mean discharge per hydrological year (R_{Q7_min}) for the 100 calibrated model parameter sets for the three selected catchments (C1, C2, C3). The model was calibrated with either daily discharge data (crosses) or daily discharge data and flowiness (FI) data (circles). Note that the Y-axis for R_{Q7_min} is reversed as smaller errors indicate a better simulation of low flows. The grey shaded areas for KGE and R_{Q7_min} indicate the range between the upper and lower benchmarks.

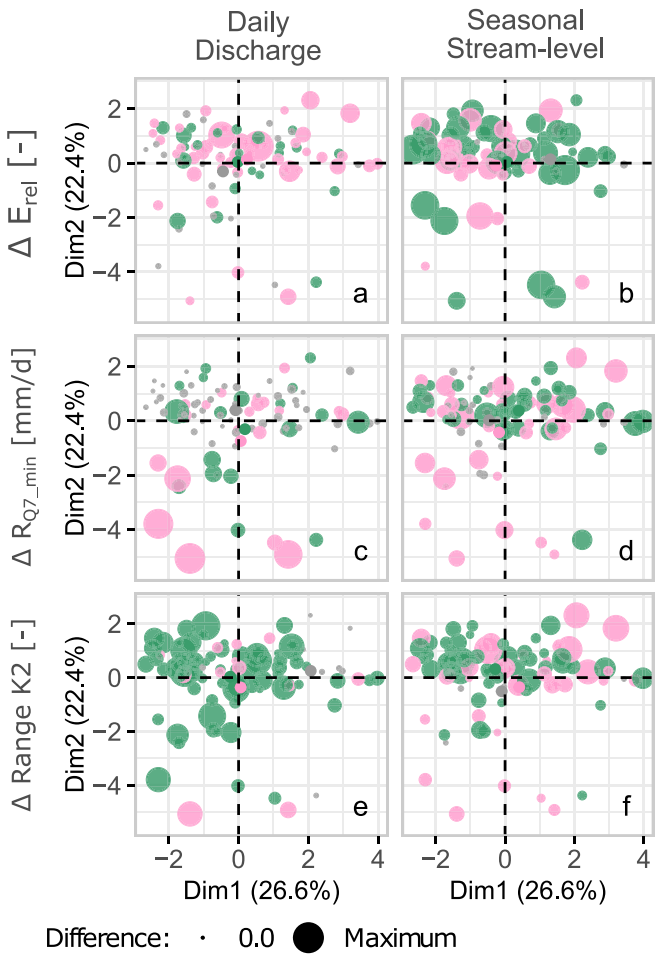


Fig. 9. The differences in model performance when flowiness data are used in model calibration for the daily discharge (left) and seasonal stream-level (one value per season) (right) scenarios for the 92 catchments plotted in the principal component space: the difference in the relative Kling-Gupta efficiency (ΔE_{rel}; a-b), the difference in the root mean squared error for the minimum seven-day mean discharge per hydrological year (ΔR_{Q7_min}; c-d), and the difference in the scaled range for parameter K2 (e-f). Green circles imply an improvement in model performance when flowiness data were included in model calibration, and pink circles a deterioration in model performance. Small differences (±0.01) are shown in grey. The size of the circles represents the absolute change in model performance, with the biggest circles representing a value of 1.5 for ΔE_{rel}, 0.33 mm/d for ΔR_{Q7_min} and 1.0 for the difference in the scaled range for K2. The first component of the PCA was mainly influenced by the flashiness index (24 %, negative correlation), catchment area (22 %, positive correlation), base flow index (22 %, positive correlation), and number of ONDE sites per catchment (17 %, positive correlation). The second component was mainly affected by the mean elevation (34 %, negative correlation), stream density (33 %, negative correlation), and aridity index (19 %, negative correlation) (see Fig. S5).

parameter K2 (see also discussion section 4.2). Seibert (2000) similarly showed that multi-data calibration with discharge and groundwater level data can better constrain the parameters that influence low-flow simulations. The fact that the additional data improved low-flow simulations is helpful because most objective functions for model calibration give relatively little weight to the low-flow simulations (Oudin et al., 2006).

There was no clear pattern for which type of catchments, the addition of flowiness data improved the overall discharge or low-flow simulations. Thus, we cannot predict for which catchments observations of the flow state of temporary streams will be useful for model calibration. A similar result was found by Pelletier and Andréassian (2022), who also did not find a spatial pattern in the improvement of model performance when groundwater level data were used in model calibration. Rakovec et al. (2016) calibrated a model for 83 catchments throughout Europe and did not find a pattern in the change in model performance when including satellite-based estimates of total water storage in model calibration either.

4.2. *Effects of flowiness data on parameter uncertainty*

Using flowiness data in the model calibration helped to constrain parameter K2. This is useful as the parameters that influence the low-flow simulations are often not well-constrained (Seibert, 2000). Thus, adding flowiness data to the calibration of the model can be useful, even if it does not visibly improve the overall mean model performance. A similar result was found for studies that included groundwater level data in model calibration (Pelletier and Andréassian, 2022; Seibert, 2000).

Abebe et al. (2010) and Karimi et al. (2022) reported that small values for the K2 parameter (approximately < 0.01 d^{-1}) can lead to a good model fit and a small improvement in the model's objective function compared to more typical values for parameter K2, but unrealistically high groundwater storage, with small or no seasonal fluctuations. They also showed that this leads to a worse, almost constant low-flow or baseflow simulation. For catchments in the southeastern USA, this occurred when the K2 parameter values ranged between 0.001 and 0.015 d^{-1} (Abebe et al., 2010). Karimi et al. (2022), therefore, suggested to change the minimum value of the K2 parameter to 0.02 d^{-1}. We set the minimum value of K2 to 0.001 d^{-1} (Table 2) to be consistent with the parameter ranges used by Seibert (2000), Seibert and Vis (2012), and Etter et al., (2020a). Including temporary stream data in the calibration with daily discharge solved the issue of unrealistically high water levels in the (s)lower groundwater storage for most catchments. However, the

optimized K2 values were sometimes still smaller than 0.02 d^{-1} (Fig. 8e, f, i). Therefore, we do not recommend changing the minimum value for the K2 parameter to an arbitrary value but rather to use additional information in the model calibration. Otherwise, the "real" optimum model parameter value may be missed.

Other studies have also reported a reduction in parameter uncertainty when using multi-data calibration (e.g., Beldring, 2002; Oudin et al., 2003; Riboust et al., 2019). The uncertainty decreased especially for the parameters that describe the processes to which the additional data were compared (Nan et al., 2021; Parajka et al., 2007; Pelletier and Andréassian, 2022; Rajib et al., 2016). For example, Seibert (2000) showed that the use of groundwater level data in the calibration of the HBV model reduced the uncertainties of the parameters in the response routine (e.g., PERC, Alpha, K1, and K2). However, it increased the uncertainty of the parameters outside this routine (e.g., MAXBAS). In our case, we did not find a consistent and significant change in the uncertainty for the other parameters of the response routine (other than K2) when flowiness data were used for calibration. This was somewhat surprising as parameters K2 and PERC are known to be very dependent (Abebe et al., 2010). However, the uncertainties of the parameters FC and LP, both used in the soil moisture routine, tended to decrease when flowiness data was used together with discharge data in model calibration.

4.3. Advantages and disadvantages of flowiness as a metric for groundwater storage

In previous studies (e.g., Finger et al., 2015; Mostafaie et al., 2018; Seibert, 2000), the additional variable used in model calibration (e.g., groundwater levels or soil moisture) could be compared to a simulated variable. For example, Pelletier and Andréassian (2022) showed that the use of groundwater level data in model calibration improved the simulation of groundwater levels for a variety of catchments across France. Remotely sensed soil moisture and groundwater data improved the simulation of soil moisture and groundwater storage of the HBV model for the Moselle River Basin (Demirel et al., 2019). We could not do this because the additional data (flowiness) are only a proxy of the simulated variable (groundwater storage). The advantage of visual observations of the state of temporary streams over actual groundwater data is that they are much easier to collect. A disadvantage is that the relation between flowiness and storage reaches a plateau when all streams flow. This is also the case for the relation between flowiness and discharge (see example in Fig. S1). Thus, flowiness data may only provide information on the storage during relatively low flow periods, whereas groundwater level data will provide information at times of higher catchment wetness (and discharge) as well.

Furthermore, flowiness was based on only three different, discrete flow states. Thus, the number of potential values for flowiness is limited if the flow state is only observed for a few temporary streams. This resulted in many ties in the flowiness time series. Etter et al., (2020a) showed that discrete observations of the water level in a stream (i.e., water level class data) were still informative for the calibration of a lumped model, even when there were only three different classes and thus many ties, as long as the observations had a weekly resolution. An increase in the observation frequency for temporary streams could be achieved by citizen science. Citizen scientists have, for example, observed the water level in streams at a high frequency (e.g., 271 observations in one year in Kenya (Weeser et al., 2019); 505 observations in one year and nine months for a stream in Austria (Etter et al., 2020b)). However, data of 'higher' frequency and quality does not necessarily lead to an improvement in model simulations for multi-data calibration. The value of higher resolution flowiness data for model calibration still needs to be tested. These data are rare, but several projects are currently collecting this type of data (e.g., Datry et al., 2016; Kampf et al., 2018; Seibert et al., 2019b). Therefore, tests with higher-resolution temporary stream data may be possible in a few years from now.

Finally, aggregating the spatially distributed observations of stream states into a time series of flowiness leads to a loss of information. This aggregation would not be necessary for a semi- or fully-distributed model. Previous studies that used semi- or fully-distributed models for multi-data calibration had a similar number of observation points relative to the catchment size (e.g., Holmes et al. (2022) (semi-distributed), Madsen (2003) (fully-distributed)). As a next step, the value of temporary stream observations (such as those in the ONDE data set) could be assessed for distributed models. The ONDE data could, for instance, be useful for stream network modelling in ungauged headwater catchments (Stoll and Weiler, 2010).

5. Conclusions

We studied the value of observations of the flow state of temporary (i.e., non-perennial) headwater streams as a proxy for groundwater storage for calibrating a lumped hydrological model. The temporary stream observations for the 92 catchments in France did not affect the overall performance of the discharge simulation. For 25 % of the catchments, the overall model performance for the validation period improved ($\Delta E_{ref} > 0.01$) and for 47 % it worsened ($\Delta E_{ref} < -0.01$) when flowiness data were used in the calibration with daily discharge data. If only seasonal stream-level data had been available, the simulations improved for 25 % of the catchments and became worse for another 25 % of the catchments. For the other 50 % of the catchments, the change in model performance was negligible (i.e., ΔE_{ref} changed by less than $\pm$ 0.01). However, including temporary stream observations in model calibration was more likely to improve the overall model performance if calibration based on only discharge or stream-level data led to a poor model fit for the validation period. In other words, calibration with flowiness data could avoid poor model fits but not improve already good model fits. Using the temporary stream observations in model calibration mainly affected the low-flow simulations and the uncertainty of the parameter that influences low flows. However, there was no spatial pattern in the improvement in the simulation of the low flows or parameter uncertainty, nor any correlation to catchment characteristics. Thus, it remains difficult to predict for which types of catchments observations of the flow state of temporary streams are useful for model calibration.

Funding

This study was funded by the Swiss National Science Foundation (SNSF, project 200020_192125, CrowdWater II).

CRediT authorship contribution statement

Mirjam Scheller: Conceptualization, Methodology, Software, Writing – original draft, Visualization. **Ilja van Meerveld:** Conceptualization, Methodology, Writing – review & editing, Supervision. **Eric Sauquet:** Resources, Writing – review & editing. **Marc Vis:** Conceptualization, Methodology, Software, Writing – review & editing. **Jan Seibert:** Conceptualization, Methodology, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare no financial interests/personal relationships which may be considered as potential competing interests.

Data availability

Data partly open access, partly available from agencies upon request.

Acknowledgements

We thank the French Office for Biodiversity for maintaining the Observatoire National des Etiages (ONDE) Network and Météo-France for providing access to the Safran database. We thank Maria Staudinger for insightful discussions on low-flow simulations, and the editors and reviewers for their valuable comments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2024.130686.

References

Abebe, N.A., Ogden, F.L., Pradhan, N.R., 2010. Sensitivity and uncertainty analysis of the conceptual HBV rainfall–runoff model: Implications for parameter estimation. J. Hydrol. 389 (3), 301–310. https://www.sciencedirect.com/science/article/pii/S0022169410003422.

Allen, R.G., Pereira, L.S., Raes, D., Smith, M., et al., 1998. Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56. FAO, Rome 300 (9), D05109.

Assendelft, R.S., van Meerveld, H.J.I., 2019. A Low-Cost, Multi-Sensor System to Monitor Temporary Stream Dynamics in Mountainous Headwater Catchments. Sensors 19 (21). https://www.mdpi.com/1424-8220/19/21/4645.

Aubert, D., Loumagne, C., Oudin, L., 2003. Sequential assimilation of soil moisture and streamflow data in a conceptual rainfall–runoff model. J. Hydrol. 280 (1), 145–161. https://www.sciencedirect.com/science/article/pii/S0022169403002294.

Avellaneda, P.M., Ficklin, D.L., Lowry, C.S., Knouft, J.H., Hall, D.M., 2020. Improving Hydrological Models with the Assimilation of Crowdsourced Data. Water Resour. Res. 56 (5). https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2019WR026325.

Baker, D.B., Richards, R.P., Loftus, T.T., Kramer, J.W., 2004. A new flashiness index: Characteristics and applications to midwestern rivers and streams. JAWRA J. Am. Water Resour. Assoc. 40 (2), 503–522.

Beaufort, A., Lamouroux, N., Pella, H., Datry, T., Sauquet, E., 2018. Extrapolating regional probability of drying of headwater streams using discrete observations and gauging networks. Hydrol. Earth Syst. Sci. 22 (5), 3033–3051. https://hess.copernicus.org/articles/22/3033/2018/.

Beaufort, A., Carreau, J., Sauquet, E., 2019. A classification approach to reconstruct local daily drying dynamics at headwater streams. Hydrol. Process. 33 (13), 1896–1912.

Beldring, S., 2002. Multi-criteria validation of a precipitation–runoff model. J. Hydrol. 257 (1), 189–211. https://www.sciencedirect.com/science/article/pii/S0022169401005418.

Bennett, J.C., Robertson, D.E., Shrestha, D.L., Wang, Q.J., Enever, D., Hapuarachchi, P., Tuteja, N.K., 2014. A System for Continuous Hydrological Ensemble Forecasting (SCHEF) to lead times of 9 days. J. Hydrol. 519, 2832–2846.

Bergström, S., 1992. The HBV model – its structure and applications. 0283-1104 (4).

Bivand, R.S., Wong, D.W.S., 2018. Comparing implementations of global and local indicators of spatial association. TEST 27 (3), 716–748.

Botter, G., Durighetto, N., 2020. The Stream Length Duration Curve: A Tool for Characterizing the Time Variability of the Flowing Stream Length. Water Resour. Res. 56 (8) e2020WR027282.

Brigode, P., Génot, B., Loblígeois, F., Delaigue, O., 2020. Summary sheets of watershed-scale hydroclimatic observed data for France. Recherche Data Gouv. Accessed.

Caillouet, L., Vidal, J.-P., Sauquet, E., Devers, A., Graff, B., 2017. Ensemble reconstruction of spatio-temporal extreme low-flow events in France since 1871. Hydrol. Earth Syst. Sci. 21 (6), 2923–2951. https://hess.copernicus.org/articles/21/2923/2017/.

Datry, T., Larned, S.T., Tockner, K., 2014. Intermittent Rivers: A Challenge for Freshwater Ecology. Bioscience 64 (3), 229–235.

Datry, T., Pella, H., Leigh, C., Bonada, N., Hugueny, B., 2016. A landscape approach to advance intermittent river ecology. Freshw. Biol. 61 (8), 1200–1213.

Demirel, M.C., Özen, A., Orta, S., Toker, E., Demir, H.K., Ekmekcioğlu, Ö., Tayşi, H., Eruçar, S., Sağ, A.B., Sarı, Ö., Tuncer, E., Hancı, H., Özcan, T.İ., Erdem, H., Koşucu, M.M., Başakın, E.E., Ahmed, K., Anwar, A., Avcuoğlu, M.B., Vanlı, Ö., Stisen, S., Booij, M.J., 2019. Additional Value of Using Satellite-Based Soil Moisture and Two Sources of Groundwater Data for Hydrological Model Calibration. Water 11 (10). https://www.mdpi.com/2073-4441/11/10/2083.

Detty, J.M., McGuire, K.J., 2010. Topographic controls on shallow groundwater dynamics: implications of hydrologic connectivity between hillslopes and riparian zones in a till mantled catchment. Hydrol. Process. 24 (16), 2222–2236.

Dimitrova-Petrova, K., Geris, J., Wilkinson, M.E., Rosolem, R., Verrot, L., Lilly, A., Soulsby, C., 2020. Opportunities and challenges in using catchment-scale storage estimates from cosmic ray neutron sensors for rainfall-runoff modelling. J. Hydrol. 586, 124878.

Durighetto, N., Vingiani, F., Bertassello, L.E., Camporese, M., Botter, G., 2020. Intraseasonal Drainage Network Dynamics in a Headwater Catchment of the Italian Alps. Water Resour. Res. 56 (4).

Etter, S., Strobl, B., Seibert, J., van Meerveld, H.J.I., 2020a. Value of Crowd-Based Water Level Class Observations for Hydrological Model Calibration. Water Resour. Res. 56 (2).

Etter, S., Strobl, B., van Meerveld, I., Seibert, J., 2020b. Quality and timing of crowd-based water level class observations. Hydrol. Process. 34 (22), 4365–4378.

European Environment Agency, 2012. European catchments and Rivers network system (Ecrins). https://www.eea.europa.eu/data-and-maps/data/european-catchments-and-rivers-network. Accessed 12.9.22.

Finger, D., Vis, M., Huss, M., Seibert, J., 2015. The value of multiple data set calibration versus model complexity for improving the performance of hydrological models in mountain catchments. Water Resour. Res. 51 (4), 1939–1958.

Godsey, S.E., Kirchner, J.W., 2014. Dynamic, discontinuous stream networks: hydrologically driven variations in active drainage density, flowing channels and stream order. Hydrol. Process. 28 (23), 5791–5803.

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377 (1), 80–91. https://www.sciencedirect.com/science/article/pii/S0022169409004843.

Gutiérrez-Jurado, K.Y., Partington, D., Batelaan, O., Cook, P., Shanafield, M., 2019. What Triggers Streamflow for Intermittent Rivers and Ephemeral Streams in Low-Gradient Catchments in Mediterranean Climates. Water Resour. Res. 55 (11), 9926–9946.

Hammond, J.C., Zimmer, M., Shanafield, M., Kaiser, K., Godsey, S.E., Mims, M.C., Zipper, S.C., Burrows, R.M., Kampf, S.K., Dodds, W., Jones, C.N., Krabbenhoft, C.A., Boersma, K.S., Datry, T., Olden, J.D., Allen, G.H., Price, A.N., Costigan, K., Hale, R., Ward, A.S., Allen, D.C., 2021. Spatial Patterns and Drivers of Nonperennial Flow Regimes in the Contiguous United States. Geophys. Res. Lett. 48 (2) e2020GL090794.

Holmes, T.L., Stadnyk, T.A., Asadzadeh, M., Gibson, J.J., 2022. Variability in flow and tracer-based performance metric sensitivities reveal regional differences in dominant hydrological processes across the Athabasca River basin. J. Hydrol.: Reg. Stud. 41, 101088. https://www.sciencedirect.com/science/article/pii/S221458182200101X.

Institute of Hydrology, 1980. Low flow studies. Research Report 1, Institute of Hydrology, Wallingford, UK.

Jaeger, K.L., Sando, R., McShane, R.R., Dunham, J.B., Hockman-Wert, D.P., Kaiser, K.E., Hafen, K., Risley, J.C., Blasch, K.W., 2019. Probability of Streamflow Permanence Model (PROSPER): A spatially continuous model of annual streamflow permanence throughout the Pacific Northwest. Journal of Hydrology X 2, 100005. https://www.sciencedirect.com/science/article/pii/S2589915518300051.

Jian, J., Ryu, D., Costelloe, J.F., Su, C.-H., 2017. Towards hydrological model calibration using river level measurements. J. Hydrol.: Reg. Stud. 10, 95–109. https://www.sciencedirect.com/science/article/pii/S2214581816303500.

Kampf, S., Strobl, B., Hammond, J., Anenberg, A., Etter, S., Martin, C., Puntenney-Desmond, K., Seibert, J., van Meerveld, I., 2018. Testing the Waters: Mobile Apps for Crowdsourced Streamflow Data. Eos 99.

Karimi, S., Seibert, J., Laudon, H., 2022. Evaluating the effects of alternative model structures on dynamic storage simulation in heterogeneous boreal catchments. Hydrol. Res. 53 (4), 562–583.

Kundu, D., Vervoort, R.W., van Ogtrop, F.F., 2017. The value of remotely sensed surface soil moisture for model calibration using SWAT. Hydrol. Process. 31 (15), 2764–2780.

Larned, S.T., Datry, T., ARSCOTT, D.B., Tockner, K., 2010. Emerging concepts in temporary-river ecology. Freshwater Biology 55 (4), 717–738.

Li, Y., Grimaldi, S., Pauwels, V.R., Walker, J.P., 2018. Hydrologic model calibration using remotely sensed soil moisture and discharge measurements: The impact on predictions at gauged and ungauged locations. J. Hydrol. 557, 897–909.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., Bergström, S., 1997. Development and test of the distributed HBV-96 hydrological model. J. Hydrol. 201 (1), 272–288. https://www.sciencedirect.com/science/article/pii/S0022169497000413.

Madsen, H., 2003. Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. Adv. Water Resour. 26 (2), 205–216. https://www.sciencedirect.com/science/article/pii/S0309170802000921.

Mahoney, D.T., Christensen, J.R., Golden, H.E., Lane, C.R., Evenson, G.R., White, E., Fritz, K.M., D'Amico, E., Barton, C.D., Williamson, T.N., Sena, K.L., Agouridis, C.T., 2023. Dynamics of streamflow permanence in a headwater network: Insights from catchment-scale model simulations. J. Hydrol. 620, 129422.

Martínez-Fernández, J., Ceballos, A., 2005. Mean soil moisture estimation using temporal stability analysis. J. Hydrol. 312 (1), 28–38. https://www.sciencedirect.com/science/article/pii/S0022169405000764.

Melsen, L.A., Teuling, A.J., van 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 (7), 5577–5596.

Messager, M.L., Lehner, B., Cockburn, C., Lamouroux, N., Pella, H., Snelder, T., Tockner, K., Trautmann, T., Watt, C., Datry, T., 2021. Global prevalence of non-perennial rivers and streams. Nature 594 (7863), 391–397.

Mostafaie, A., Forootan, E., Safari, A., Schumacher, M., 2018. Comparing multi-objective optimization techniques to calibrate a conceptual hydrological model using in situ runoff and daily GRACE data. Comput. Geosci. 22 (3), 789–814.

Nan, Y., Tian, L., He, Z., Tian, F., Shao, L., 2021. The value of water isotope data on improving process understanding in a glacierized catchment on the Tibetan Plateau. Hydrol. Earth Syst. Sci. 25 (6), 3653–3673. https://hess.copernicus.org/articles/25/3653/2021/.

Oudin, L., Weisse, A., Loumagne, C., Le Hégarat-Mascle, S., 2003. Assimilation of soil moisture into hydrological models for flood forecasting: a variational approach. Can. J. Remote. Sens. 29 (6), 679–686.

Oudin, L., Andréassian, V., Mathevet, T., Perrin, C., Michel, C., 2006. Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations. Water Resour. Res. 42 (7).

Oudin, L., Andréassian, V., Perrin, C., Michel, C., Le Moine, N., 2008. Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments. Water Resour. Res. 44 (3).

Parajka, J., Naeimi, V., Blöschl, G., Wagner, W., Merz, R., Scipal, K., 2006. Assimilating scatterometer soil moisture data into conceptual hydrologic models at the regional scale. Hydrol. Earth Syst. Sci. 10 (3), 353–368.

Parajka, J., Merz, R., Blöschl, G., 2007. Uncertainty and multiple objective calibration in regional water balance modelling: case study in 320 Austrian catchments. Hydrol. Process. 21 (4), 435–446.

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 (5), 1783–1795.

Pelletier, A., Andréassian, V., 2022. On constraining a lumped hydrological model with both piezometry and streamflow: results of a large sample evaluation. Hydrol. Earth Syst. Sci. 26 (10), 2733–2758. https://hess.copernicus.org/articles/26/2733/2022/.

Quintana-Seguí, P., Le Moigne, P., Durand, Y., Martin, E., Habets, F., Baillon, M., Canellas, C., Franchisteguy, L., Morel, S., 2008. Analysis of near-surface atmospheric variables: Validation of the SAFRAN analysis over france. J. Appl. Meteorol. Climatol. 47 (1), 92–107. https://journals.ametsoc.org/view/journals/apme/47/1/2007jamc1636.1.xml.

Rajib, M.A., Merwade, V., Yu, Z., 2016. Multi-objective calibration of a hydrologic model using spatially distributed remotely sensed/in-situ soil moisture. J. Hydrol. 536, 192–207. https://www.sciencedirect.com/science/article/pii/S0022169416300713.

Rakovec, O., Kumar, R., Attinger, S., Samaniego, L., 2016. Improving the realism of hydrologic model functioning through multivariate parameter estimation. Water Resour. Res. 52 (10), 7779–7792.

Revilla-Romero, B., Beck, H.E., Burek, P., Salamon, P., de Roo, A., Thielen, J., 2015. Filling the gaps: Calibrating a rainfall-runoff model using satellite-derived surface water extent. Remote Sens. Environ. 171, 118–131.

Riboust, P., Thirel, G., Le Moine, N., Ribstein, P., 2019. Revisiting a Simple Degree-Day Model for Integrating Satellite Data: Implementation of Swe-Sca Hystereses. Journal of Hydrology and Hydromechanics 67 (1), 70–81.

Samain, B., Pauwels, V.R.N., 2013. Impact of potential and (scintillometer-based) actual evapotranspiration estimates on the performance of a lumped rainfall–runoff model. Hydrol. Earth Syst. Sci. 17 (11), 4525–4540.

Sauquet, E., Beaufort, A., Sarremejane, R., Thirel, G., 2021. Predicting flow intermittence in France under climate change. Hydrol. Sci. J. 66 (14), 2046–2059.

Schaefli, B., Huss, M., 2011. Integrating point glacier mass balance observations into hydrologic model identification. Hydrol. Earth Syst. Sci. 15 (4), 1227–1241. https://hess.copernicus.org/articles/15/1227/2011/.

Seibert, J., 2000. Multi-criteria calibration of a conceptual runoff model using a genetic algorithm. Hydrol. Earth Syst. Sci. 4 (2), 215–224. https://hess.copernicus.org/articles/4/215/2000/.

Seibert, J., Beven, K.J., 2009. Gauging the ungauged basin: how many discharge measurements are needed? Hydrol. Earth Syst. Sci. 13 (6), 883–892.

Seibert, J., Bishop, K., Rodhe, A., McDonnell, J.J., 2003. Groundwater dynamics along a hillslope: A test of the steady state hypothesis. Water Resour. Res. 39 (1).

Seibert, J., McDonnell, J.J., 2002. On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration. Water Resour. Res. 38 (11), 23-1-23-14.

Seibert, J., Staudinger, M., van Meerveld, H.J.I., 2019a. Validation and Over-Parameterization–Experiences from Hydrological Modeling. In: Beisbart, C., Saam, N.J. (Eds.), Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives. Springer International Publishing, Cham, pp. 811–834.

Seibert, J., van Meerveld, H.J., Etter, S., Strobl, B., Assendelft, R., Hummer, P., 2019b. Wasserdaten sammeln mit dem Smartphone – Wie können Menschen messen, was hydrologische Modelle brauchen? Bundesanstalt fuer Gewaesserkunde, Accessed.

Seibert, J., Vis, M.J.P., 2012. Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. Hydrol. Earth Syst. Sci. 16 (9), 3315–3325. https://hess.copernicus.org/articles/16/3315/2012/.

Seibert, J., Vis, M.J.P., 2016. How informative are stream level observations in different geographic regions? Hydrol. Process. 30 (14), 2498–2508.

Seibert, J., Vis, M.J.P., Lewis, E., van Meerveld, H.J., 2018. Upper and lower benchmarks in hydrological modelling. Hydrol. Process. 32 (8), 1120–1125.

Sivapalan, M., Takeuchi, K., Franks, S.W., Gupta, V.K., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J.J., Mendiondo, E.M., O'Connell, P.E., 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 (6), 857–880.

Stoll, S., Weiler, M., 2010. Explicit simulations of stream networks to guide hydrological modelling in ungauged basins. Hydrol. Earth Syst. Sci. 14 (8), 1435–1448. https://hess.copernicus.org/articles/14/1435/2010/.

Stubbington, R., England, J., Wood, P.J., Sefton, C.E., 2017. Temporary streams in temperate zones: recognizing, monitoring and restoring transitional aquatic-terrestrial ecosystems. WIREs Water 4 (4), e1223.

Széles, B., Parajka, J., Hogan, P., Silasari, R., Pavlin, L., Strauss, P., Blöschl, G., 2020. The Added Value of Different Data Types for Calibrating and Testing a Hydrologic Model in a Small Catchment. Water Resour. Res. 56 (10).

Vaché, K.B., McDonnell, J.J., Bolte, J., 2004. On the use of multiple criteria for a posteriori model rejection: Soft data to characterize model performance. Geophys. Res. Lett. 31 (21), n/a-n/a.

van Meerveld, H.J.I., Seibert, J., Peters, N.E., 2015. Hillslope–riparian-stream connectivity and flow directions at the Panola Mountain Research Watershed. Hydrol. Process. 29 (16), 3556–3574.

van Meerveld, H.J.I., Vis, M.J.P., Seibert, J., 2017. Information content of stream level class data for hydrological model calibration. Hydrol. Earth Syst. Sci. 21 (9), 4895–4905.

van Meerveld, H.J.I., Sauquet, E., Gallart, F., Sefton, C., Seibert, J., Bishop, K., 2020. Aqua temporaria incognita. Hydrol. Process. 34 (26), 5704–5711.

Vidal, J.-P., Martin, E., Franchistéguy, L., Baillon, M., Soubeyroux, J.-M., 2010. A 50-year high-resolution atmospheric reanalysis over France with the Safran system. Int. J. Climatol. 30 (11), 1627–1644.

Weeser, B., Jacobs, S., Kraft, P., Rufino, M.C., Breuer, L., 2019. Rainfall-Runoff Modeling Using Crowdsourced Water Level Data. Water Resour. Res. 55 (12), 10856–10871.

Wohl, E., 2017. The significance of small streams. Frontiers of Earth Science 11 (3), 447–456.

Zanetti, F., Durighetto, N., Vingiani, F., Botter, G., 2022. Technical note: Analyzing river network dynamics and the active length–discharge relationship using water presence sensors. Hydrol. Earth Syst. Sci. 26 (13), 3497–3516.