Contents lists available at ScienceDirect

Journal of Hydrology

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

Research papers

Gauging ungauged catchments – Active learning for the timing of point discharge observations in combination with continuous water level measurements

Sandra Pool[a,b,*], Jan Seibert[c,d]

[a] Eawag, Swiss Federal Institute of Aquatic Science and Technology, Department Water Resources and Drinking Water, Überlandstrasse 133, 8600 Dübendorf, Switzerland
[b] Eawag, Swiss Federal Institute of Aquatic Science and Technology, Department Systems Analysis, Integrated Assessment and Modelling, Überlandstrasse 133, 8600 Dübendorf, Switzerland
[c] University of Zurich, Department of Geography, Winterthurerstrasse 190, 8057 Zurich, Switzerland
[d] Swedish University of Agricultural Sciences, Department of Aquatic Sciences and Assessment, 75007 Uppsala, Sweden

ARTICLE INFO

This manuscript was handled by Andras Bardossy, Editor-in-Chief

Keywords:
Prediction in ungauged basins
Value of data
Model calibration
Water-level time series
Point discharge observations

ABSTRACT

Hydrological models have traditionally been used for the prediction in ungauged basins despite the related challenge of model parameterization. Short measurement campaigns could be a way to obtain some basic information that is needed to support model calibration in these catchments. This study explores the potential of such field campaigns by i) testing the relative value of continuous water-level time series and point discharge observations for model calibration, and by ii) evaluating the value of point discharge observations collected using expert knowledge and active learning to guide when to measure streamflow. The study was based on 100 gauged catchments across the contiguous United States for which we pretended to have only limited hydrological observations, i.e., continuous daily water levels and ten daily point discharge observations from different hypothetical field trips conducted within one hydrological year. Water level data were used as a single source of information, as well as in addition to point discharge observations, for calibrating the HBV model. Calibration against point discharge observations was conducted iteratively by continually adding new observations from one of the ten field measurements. Our results suggested that the information contained in point discharge observations was especially valuable for constraining the annual water balance and streamflow response at the event scale, improving predictions based solely on water levels by up to 50% after ten field observations. In contrast, water levels were valuable to increase the accuracy of simulated daily streamflow dynamics. Informative discharge sampling dates were similar when selected with either active learning or expert knowledge and typically clustered during seasons with high streamflow.

1. Introduction

Many catchments that are of interest for research or practical purposes are ungauged or poorly gauged even in regions with a relatively dense hydrological observation network. Yet streamflow information is critical for the design and management of water infrastructures. Hydrological models are a commonly used tool to predict streamflow and its temporal variation under both current and future conditions. Parameter values of hydrological models are typically adapted to a specific catchment by calibration and validation against observed streamflow. The prediction of streamflow in ungauged catchments, that is, catchments without any observed discharge, is one of the major challenges in hydrology. This long-standing challenge has received renewed, community-wide attention through the PUB (Prediction in Ungauged Basins) initiative launched by the IAHS (International Association of Hydrological Sciences) (Hrachowitz et al., 2013).

Model calibration should ideally be based on long continuous discharge time series (Brath et al., 2004; Merz et al., 2009; Singh and Bárdossy, 2012; Tada and Beven, 2012; Vrugt et al., 2006). However, it has been demonstrated that much shorter time series between one and six months can lead to robust model parameter estimates (Brath et al., 2004; Melsen et al., 2014; Sun et al., 2017). Others have shown that

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

https://doi.org/10.1016/j.jhydrol.2021.126448
Received 6 January 2021; Received in revised form 7 May 2021; Accepted 7 May 2021
Available online 11 May 2021

some point discharge observations, taken at randomly chosen dates, can provide valuable information for model calibration (Kim and Kaluarachchi, 2009; Perrin et al., 2007). Collecting such individual discharge data points strategically by explicitly taking observations during peak flows or events and the subsequent recessions (Correa et al., 2016; McIntyre and Wheater, 2004; Pool et al., 2017; Seibert and McDonnell, 2015), could further lower the number of data points needed to reach acceptable model parameterizations. Results indicate that even a small sample of ten to sixteen observations can be highly informative (Pool et al., 2017; Seibert and Beven, 2009; Seibert and McDonnell, 2015), especially when the natural variability in streamflow is well represented and there are observations when dominant hydrological processes are active (Harlin, 1991; Singh and Bárdossy, 2012; Sun et al., 2017; Tan et al., 2008; Vrugt et al., 2006; Yapo et al., 1996). These findings are in line with results from influence diagnostic statistics, which demonstrated that the ten most influential observations cover a range of flow magnitudes (Wright et al., 2018), whereby the five most influential discharge observations have an order of magnitude more influence on model performance than any other observation in a ten-year time series (Wright et al., 2015). One potential solution to overcome the challenges related to predictions in data-scarce situations thus might be the collection of at least some hydrological data during field campaigns. However, such field campaigns are restricted by practicalities, such as the accessibility of the catchment, financial resources, or time, which make a careful choice of observation times essential. The expert knowledge gained from the previous studies could thereby provide guidance on the choice of sampling dates.

Active learning methods provide an alternative option to investigate the value of short and discontinuous discharge time series for model calibration from an explorative point of view, rather than by testing hypotheses as has been done so far. Active learning is a subfield of machine learning that has been widely applied in the domains of text processing, remote-sensing, or chemoinformatics. These domains typically face the challenge of having large unlabeled datasets (i.e., datasets with a large number of unknown samples) that need to be classified with a prediction model. The training of the model is based on labelled samples, whereby labelling (i.e., assigning a value to an unknown data point) is expensive (Cawley, 2011). Active learning provides a method to select and label the most informative samples from the pool of unlabeled data such that the most favourable model performance can be achieved with the smallest number of samples (Settles, 2012). Active learning is an iterative approach in which the model and the user regularly interact. Current model predictions of each sample are ranked by a performance criterion, and the user selects and labels the highest-ranked samples that are subsequently used to recalibrate the prediction model (Crawford et al., 2013). A commonly used performance criterion is prediction uncertainty (Lewis and Gale, 1994), which means that high ranks are assigned to samples that have been predicted with the least confidence. It is thereby assumed that samples are most informative for model parameter estimation for points at which model simulations disagree most (Crawford et al., 2013). In hydrology, we face a similar challenge when gauging an ungauged catchment: a hydrologist needs to measure (i.e., label) the most informative discharge observations (i.e., samples) for model calibration from a future time series (i.e., unlabeled dataset) with the least possible effort. We, therefore, hypothesize that active learning could be a powerful tool to decide on the timing of discharge observations for the calibration of hydrological models in previously ungauged catchments. Note that the term *sample* has different meanings in hydrology (Brunner et al., 2018), and is used here to refer to point discharge observations selected from an existing discharge time series.

Table 1
Statistics of catchment attributes of the 100 catchments used in this study. Climate indices and hydrological signatures were calculated for the hydrological years 1990–2009.

Catchment attribute	Minimum	5th quantile	Median	95th quantile	Maximum
Area (km^2)	14	39	339	2438	12,601
Mean elevation (m a.s.l.)	23	41	448	2729	3271
Annual precipitation (mm yr^{-1})	267	532	1275	2640	3160
Precipitation falling as snow (%)	0	0	9	63	71
Aridity index (–)[a]	0.24	0.27	0.75	1.77	3.68
Annual specific discharge (mm yr^{-1})	28	84	524	2029	2678
Baseflow (%)[b]	6	13	47	83	91

[a] The aridity index was calculated as the ratio of the sum of potential evapotranspiration and the sum of precipitation (ETo/P).

[b] Baseflow was calculated using the EflowStats R-Package from the U.S. Geological Survey (2014).

Seibert and Vis (2016) suggested that instead of performing discharge measurements at different points in time, it could be easier and less time consuming to install a water-level logger. Simulations for more than 600 catchments in the contiguous United States indicated a surprisingly high value of water-level time series for model calibration, especially in humid catchments. With increasing aridity, however, information on dynamics alone was not sufficient, and the lack of volume information steadily reduced model performance. Lebecherel (2015) proposed to combine water-level time series and point discharge observations to make the most out of field trips. Specifically, Lebecherel (2015) argues that a discharge observation at a given water level can be assumed to be representative for all occasions with a similar water level provided that the stage-discharge relationship is stationary and unique. Discharge time series created this way were successfully used to inform the regionalization of model parameters in 609 French catchments. While Lebecherel (2015) exclusively used discharge for calibration and disregarded water levels, Seibert and Vis (2016) proposed a simple method to use water levels for calibration without increasing the number of model parameters. Thus, combining continuous water-level time series and point discharge observations for calibration would allow the prediction of discharge in a previously ungauged basin with local information collected with a reasonable amount of effort.

The aim of this study was to provide further guidance on the optimal collection of streamflow data at a limited number of observation times to improve the prediction in ungauged basins. This study extends previous work on the value of data in ungauged basins by explicitly comparing the value of water levels and point discharge observations, and by testing a machine learning approach for guiding the timing of point discharge observations. To evaluate the value of point discharge observations and water-level time series across a wide spectrum of hydroclimatic conditions, we used a set of 100 gauged catchments distributed over the contiguous United States. Treating the catchments as poorly gauged catchments with only a limited amount of field observations, allowed the following two main objectives to be addressed:

1. Quantification of the relative value of individual point discharge observations, continuous water-level data, or a combination thereof for the calibration of hydrological models.
2. Evaluation of the potential of *active learning* for providing guidance on the timing of the most informative discharge observations as opposed to a prior decision based on *hydrological expert knowledge*.

2. Data and methods

2.1. Study catchments

In this study, data from 100 catchments across the contiguous United States were used. The catchments represent a wide range of topographic

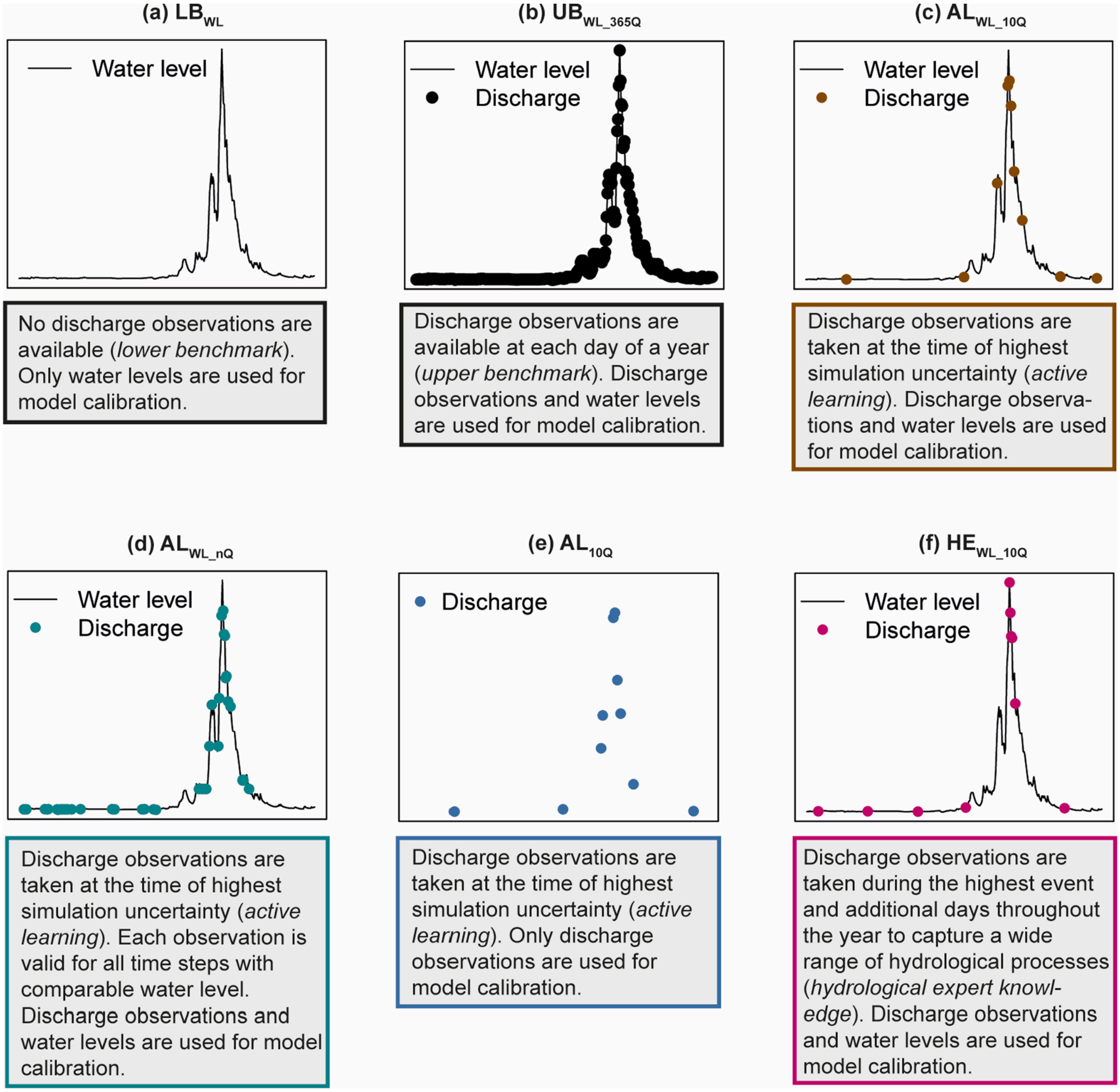

Fig. 1. Information used for model calibration with the lower and upper benchmark (LB$_{WL\ and}$ UB$_{WL\_365Q}$), the active learning-based data collection approaches (AL$_{WL\_10Q}$, AL$_{WL\_nQ}$, and AL$_{10Q}$), and the expert knowledge-based data collection approach (HE$_{WL\_10Q}$). A snow-dominated catchment in the Rocky Mountains is used as an example to indicate the temporal distribution of point discharge observations after ten sampling iterations. Note that the same colour scheme is used in Figs. 3–5 to differentiate the data collection approaches.

and hydroclimatic aspects (see Table 1 for statistics). They vary in size from 14 km^2 to 12,601 km^2, and their mean elevation ranges from 23 m a.s.l. up to 3271 m a.s.l. Annual precipitation is between 267 mm yr^{-1} and 3160 mm yr^{-1}, of which up to 71% falls as snow in some catchments. While the majority of catchments are either humid (47%) or temperate (35%), 18% can be classified as somewhat arid or arid (as defined by Coopersmith et al., 2014). Annual specific discharge ranges between 28 mm yr^{-1} and 2678 mm yr^{-1} with baseflow contributing 6% to 91% to annual streamflow.

The 100 catchments are a constrained-randomly selected subset of the Newman et al. (2015) dataset containing more than 600 U.S. catchments. The selection is consistent with a previous study (Pool et al., 2018) and was necessary to reduce the computational costs of the modelling experiments conducted in this study. The dataset was compiled by Newman et al. (2015) and provides time series of daily discharge, precipitation and temperature for each catchment. Furthermore, the dataset contains time series of different meteorological variables that were used to compute monthly potential evapotranspiration using the Priestley-Taylor equation (Priestley and Taylor, 1972). The dataset also includes basic information on catchment boundaries. However, detailed elevation data were downloaded from the SRTM digital elevation database (Jarvis et al., 2008). Further information on catchment attributes, such as topographic information, climatic indices, and hydrological signatures were extracted from the CAMELS dataset (version 1.0; Addor et al., 2017). Climate indices and hydrological signatures were recalculated for the hydrological years 1990–2009, which were used for model simulations in this study.

The model calibration experiments were based on both discharge amounts at individual points in time and water level time series. Since water levels are not part of the Newman et al. (2015) dataset, synthetic water-level time series were created for each catchment. This was done by replacing the discharge values for each day by their respective rank in the time series. In other words, we created time series that contained only the information about the temporal dynamics but not quantitative information. These series correspond to the information contained in water-level time series in the case of stationary stage-discharge relationships. In cases where there were shifts in the (real) rating curves used, these shifts were implicitly considered by our approach as we based the ranking on the estimated streamflow and not directly on the observed water levels.

2.2. Hydrological model

Continuous daily streamflow was simulated with the HBV runoff model (Hydrologiska Byråns Vattenbalansavdelning; Bergström 1976, Lindström et al. 1997) using the software implementation HBV-light (Seibert and Vis, 2012). The HBV model is a bucket-type model with a conceptual representation of hydrological processes typically dominating streamflow response at the catchment scale. Hydrological fluxes and state variables are represented by fourteen model parameters and four model routines, including a snow routine, soil routine, groundwater routine, and routing routine. Daily temperature and precipitation are used as input time series, together with long-term mean monthly potential evapotranspiration estimates. In the snow routine, a degree-day method is used to calculate snow accumulation and snowmelt. Snowmelt and rainfall supply water to the soil routine, in which simulated soil moisture content controls actual evapotranspiration and groundwater recharge. Recharge increases groundwater levels in the upper and lower reservoirs of the groundwater routine. The two reservoirs simulate the variable contribution of shallow and deep groundwater, or fast and slow runoff components, to total streamflow. Finally, in the routing routine, the sum of the three streamflow components is transformed by a triangular weighting function to simulate the hydrograph at the catchment outlet.

In this study, HBV was used in a semi-distributed way by dividing each catchment into elevation bands of 200 m. Computations in the snow and soil moisture routines were performed separately for each elevation zone, but using the same parameter values. The groundwater routine, on the other hand, was applied in a lumped way for the entire catchment. Daily temperature and precipitation input data were adjusted to each elevation band using lapse rates of 0.6 °C per 100 m (Wallace and Hobbs, 2006) and 10% per 100 m (Johansson, 2000), respectively. In contrast, monthly potential evapotranspiration values were assumed to be equal in all elevation bands.

2.3. Data collection approaches

We defined six data collection approaches representing different possible scenarios for the collection of streamflow information in a previously ungauged basin (see Fig. 1 for a visualization of the data collection approaches). The approaches mainly differ in the type of data measured, i.e., water-level time series or point discharge observations, and in the timing of the point discharge observations. The period considered for the collection of streamflow information was restricted to one hydrological year (October 1 to September 30) in each case, to reflect a situation as could be realistic in practice, where there is some limited time to collect data for a previously ungauged catchment. The data collection approaches were 'simulated' by selecting water level and discharge information from the observed time series of each catchment. A more detailed description of the approaches is provided in the following sections.

2.3.1. Benchmark approaches

A relatively simple data collection approach would be installing a water-level sensor for collecting continuous daily time series over an entire hydrological year (Fig. 1a). This approach calibrates the model against streamflow dynamics only and therefore served as a lower benchmark (LB_{WL}) for more advanced methods.

In contrast, the most data-rich approach would be the use of continuous water-level time series combined with discharge observations for each day of the hydrological year (Fig. 1b). Results from the calibration against the full dataset, including continuous water levels and continuous discharge, provide information about how good model simulations could be at best (see Section 2.4 for calibration details). These simulations, therefore, served as an upper benchmark (UB_{WL_365Q}).

2.3.2. Approaches based on active learning

Active learning could guide the decision about when to measure discharge to obtain the most informative data for model calibration. The basic idea is that discharge observations are most valuable for constraining a hydrological model on days of high model simulation uncertainty. In other words, we hypothesize that model parameterization needs most support from discharge observations when simulations disagree most.

Active learning is an iterative process, in which model parameterization is improved by adding (discharge) information at each iteration. Here, we conducted a total of ten sampling iterations that represent ten individual field trips. The active learning approach adopted here followed five main steps, whereby step two to five was repeated for each of the ten sampling iterations:

- First, an initial set of 100 parameter sets was obtained by calibration against water levels (LB_{WL}) or by a random selection of parameter values.
- Second, the model was run using these 100 parameter sets, which resulted in a range of possible hydrographs for the same forcing input.
- Third, simulation uncertainty was calculated at each time step using the difference between the 5th and 95th quantiles of the simulated discharge time series.
- Forth, a discharge measurement was selected at the time step with the highest simulation uncertainty. We selected discharge observations alternating from the highest absolute uncertainty and the highest relative uncertainty to give similar weight to different flow conditions. For example, the absolute uncertainty was used in the first iteration, the relative uncertainty was used in the second iteration, and so on.
- Fifth, the hydrological model was recalibrated taking into account the discharge observation(s) obtained in step four.

The information collected by active learning was used for model calibration in three different ways:

- AL_{WL_10Q}: Water-level time series and point discharge observations were used for model calibration (Fig. 1c). The date of the first discharge observation was defined from simulations with the lower benchmark only (LB_{WL}).
- AL_{WL_nQ}: The same procedure was applied as for AL_{WL_10Q}. However, the discharge time series was extended by assuming that an observed discharge value was representative for all time steps with a comparable water level (Fig. 1d). Comparable water levels were defined as levels for which the corresponding discharge was within +/- 5% of the discharge observation on that day (note that water levels were derived from discharge time series as described in Section 2.1).
- AL_{10Q}: Only point discharge observations were used for model calibration (Fig. 1e). The date of the first discharge observation was selected based on the uncertainty range of simulations with randomly selected parameter values.

2.3.3. Approach based on hydrological expert knowledge

Informative discharge days could alternatively be determined based on hydrological expert knowledge (Fig. 1f). Here we defined such an expert-based discharge collection strategy (HE_{WL_10Q}) using findings from a previous study (Pool et al., 2017). The strategy consisted of ten discharge observations collected at the annual peak, the first three subsequent recession days, and six observations at the 15th of every other month. In case that the 15th of a month coincided with the annual

Table 2
Performance metrics used for evaluating model performance in the validation period and metrics optimized during model calibration. The relative model performance metric R^* was calculated using the performance with limited data (R_{D}), the lower benchmark (R_{LB}), and the upper benchmark (R_{UB}). Abbreviations used in the equations refer to observed (*obs*) and simulated (*sim*) discharge (Q), time step i of a time series of length n, and the rank S of time step i within the time series.

Metric	Description	Formula
Evaluation metrics		
R_{NS}	Nash-Sutcliffe efficiency	$1 - \frac{\sum_{i=1}^{n}\left(Q_{obs(i)} - Q_{sim(i)}\right)^2}{\sum_{i=1}^{n}\left(Q_{obs(i)} - \overline{Q_{obs}}\right)^2}$; additionally calculated using square root-transformed ($R_{\mathrm{NS\_sqrtQ}}$) and log-transformed ($R_{\mathrm{NS\_logQ}}$) discharge.
R_{S}	Spearman rank correlation	$\frac{\sum_{i=1}^{n}\left(S_{obs(i)} - \overline{S_{obs}}\right)\left(S_{sim(i)} - \overline{S_{sim}}\right)}{\sqrt{\left(\sum_{i=1}^{n}\left(S_{obs(i)} - \overline{S_{obs}}\right)^2\right)\left(\sum_{i=1}^{n}\left(S_{sim(i)} - \overline{S_{sim}}\right)^2\right)}}$
R_{VE}	Volume error	$1 - \frac{\left\lvert\sum_{i=1}^{n}\left(Q_{obs(i)} - Q_{sim(i)}\right)\right\rvert}{\sum_{i=1}^{n}\left(Q_{obs(i)}\right)}$
R^*	Relative model performance	$R^* = \frac{R_D - R_{LB}}{R_{UB} - R_{LB}}100$
Calibration metrics		
$R_{\mathrm{NS\_sqrtQ\_adj}}$	Bounded Nash-Sutcliffe efficiency	$\frac{R_{\mathrm{NS\_sqrtQ}}}{2 - R_{\mathrm{NS\_sqrtQ}}}$
R_{S}	Spearman rank correlation	–

peak and its recession, we randomly selected an alternative day within that month. Model calibration was based on these point discharge observations and water-level time series, whereby discharge observations were iteratively added, starting with the peak flow information.

2.4. Model calibration using point discharge observations and water-level time series

The HBV model was calibrated for each study catchment using continuous daily meteorological input and streamflow information according to the six data collection approaches. Independent calibrations were run with data from the ten hydrological years between 1990 and 1999. The 33 months preceding the calibration periods were used for model warming-up to start model calibration from suitable initial state variables.

Parameters values were optimized within predefined feasible ranges using a genetic algorithm (Seibert, 2000) that selected and recombined an initial random set of fifty parameter values over 3500 model runs (note that no local Powel runs were conducted). Calibration was based on the two performance metrics $R_{\mathrm{NS\_sqrtQ\_adj}}$ and R_{S} (Table 2). $R_{\mathrm{NS\_sqrtQ\_adj}}$ is originally a bounded version of the Nash-Sutcliffe efficiency R_{NS} (Nash and Sutcliffe, 1970) that was proposed by Mathevet et al. (2006). $R_{\mathrm{NS\_sqrtQ\_adj}}$ was used to minimize the error between simulated and observed square root-transformed discharge observations. Model optimization against water levels was based on the Spearman rank correlation R_{S} (Spearman, 1904) as proposed by Seibert and Vis (2016). R_{S} transforms the values of a time series into a sequence of ranks and thereby reduces the information of a continuous discharge time series to its dynamical aspects. Both calibration metrics used in this study can vary between −1 and 1, with 1 representing a perfect fit. The two metrics were averaged arithmetically with equal weights for model calibrations against water levels and point discharge observation.

For each calibration step, 100 independent calibrations were conducted to account for parameter uncertainty, resulting in 100 possible hydrographs for the same forcing input. The described calibration procedure was repeated for each of the ten sampling iterations of the active learning and expert knowledge-based data collection approaches.

2.5. Evaluation of the value of point discharge observations and water-level time series

2.5.1. Characterization of point discharge observations

As a first step of the analyses, we characterized the sample of discharge observations resulting from the data collection approaches in terms of seasonal distribution and representation of streamflow classes. The seasonal distribution of discharge observations was analyzed using circular statistics (Pewsey et al., 2013). Circular statistics use the unit circle as the basis for the calculation of trigonometric moments, such as measures of location and concentration. Following the theory in Pewsey et al. (2013, Ch. 3.1–3.4, p.21–29) and the hydrological example provided in Hall and Blöschl (2018), we first converted the date of discharge observations to angular values as measured in radian. The mean sampling date (sample mean direction) and concentration index (sample mean resultant length) were then calculated to describe the sample distribution of the discharge observations from all ten sampling years. A concentration index of 1 indicates that discharge observations were tightly clustered around the mean sampling date. In contrast, smaller index values indicate a large spread of sampling dates (a uniform distribution around the year would result in a value of zero). For a more detailed description of circular statistics, we refer the reader to Pewsey et al. (2013).

To gain insights into the distribution of discharge observations at the event scale, discharge observations were classified by streamflow class. Four streamflow classes were considered including the event peak, falling limb of an event, rising limb of an event, and baseflow between two events. The classification was based on the event definition of Sikorska et al. (2015) that was used in Swiss catchments representing a range of runoff regimes. An event was defined as the period that includes a peak flow day, i.e., a day at which the flow reaches a maximum within any moving window of fifteen days. The start of an event was then defined as the day with the minimum flow over five days before an event peak. The first day after the event peak with streamflow of less than 20% of peak flow was considered the end of an event.

2.5.2. Model performance

In the second part of the analysis, we evaluated the model performance related to the six hypothetical data collection approaches. The approaches were thereby evaluated in an independent validation period covering the hydrological years 2000–2009. The continuous daily discharge simulations of the validation years were used to calculate five different performance metrics representing different aspects of the hydrograph (Table 2). R_{S} and R_{VE} were used to assess daily streamflow dynamics and annual volume separately. R_{NS} calculated from untransformed (R_{NS}), square root-transformed ($R_{\mathrm{NS\_sqrtQ}}$), and log-transformed ($R_{\mathrm{NS\_logQ}}$) time series served to evaluate the daily dynamics and magnitude of high, mean, and low flows.

In addition, each of the five performance metrics was input to the relative model performance metric R^*. As suggested by Girons Lopez and Seibert (2016), R^* was used as an indicator for the relative value of the active learning and expert knowledge-based data collection approaches compared to the lower and upper benchmarks.

Overall, model performance related to the six data collection approaches was calculated for 100 model parameterization in 10 calibration years and 100 catchments. Unless stated differently, model performances for each catchment were aggregated by calculating the median of the 100 simulations and the 10 sampling years.

Finally, the median model performance values were evaluated in terms of their spatial distribution. Maps of the value of water-level time series and point discharge observations, as quantified by the model performance improvement, were used to visually investigate which parts of the contiguous United States a particular type of data was most

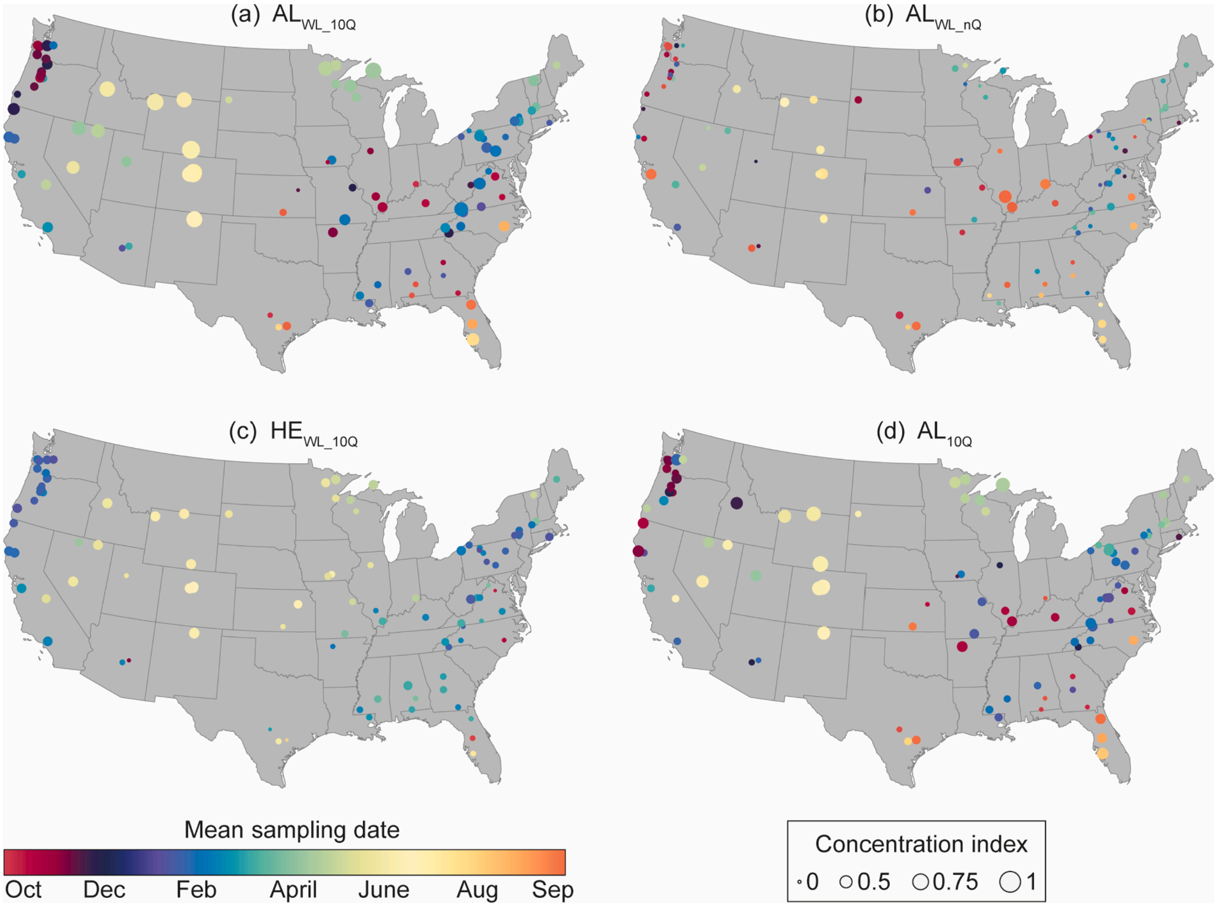

Fig. 2. Seasonal distribution of point discharge observations resulting from the data collection approaches (a) AL_{WL_10Q} (active learning with water levels and ten discharge observations), (b) AL_{WL_nQ} (active learning with water levels and *n* discharge observations), (c) AL_{10Q} (active learning with ten discharge observations), and (d) HE_{WL_10Q} (hydrological expert knowledge with water levels and ten discharge observations). The colours indicate the mean sampling date of all discharge observations after ten iterations. A large (small) marker size indicates that observations were strongly (weakly) concentrated around the mean sampling date. The mean sampling date and the concentration index were calculated from the discharge sampling dates of all ten sampling years.

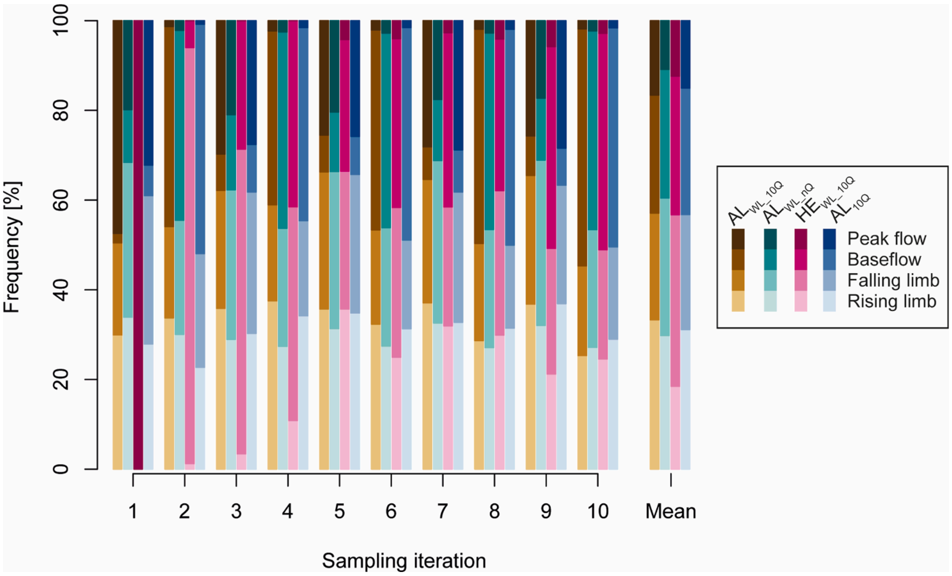

Fig. 3. Streamflow classes represented in the point discharge observations resulting from the data collection approaches AL_{WL_10Q} (active learning with water levels and ten discharge observations), AL_{WL_nQ} (active learning with water levels and *n* discharge observations), AL_{10Q} (active learning with ten discharge observations), and HE_{WL_10Q} (hydrological expert knowledge with water levels and ten discharge observations). The y-axis indicates the percentage of discharge observations in a given streamflow class in iteration 1–10, whereby values represent an average over all 100 catchments. The last column indicates the mean frequency of a streamflow class over all ten iterations. The total number of discharge observations in the final iteration was ten for AL_{WL_10Q}, AL_{10Q}, and HE_{WL_10Q}, and ranged from 32 to 303 for AL_{WL_nQ} (average of all catchments was 67 discharge observations).

informative for. The interpretation of the maps was supported by calculating Spearman rank correlations between the model performance improvements and six hydroclimatic catchment attributes. Based on the studies of Berghuijs et al. (2014) and Addor et al. (2018), we selected aridity, the fraction of precipitation falling as snow, and precipitation seasonality as three important climatic indices. To quantify the

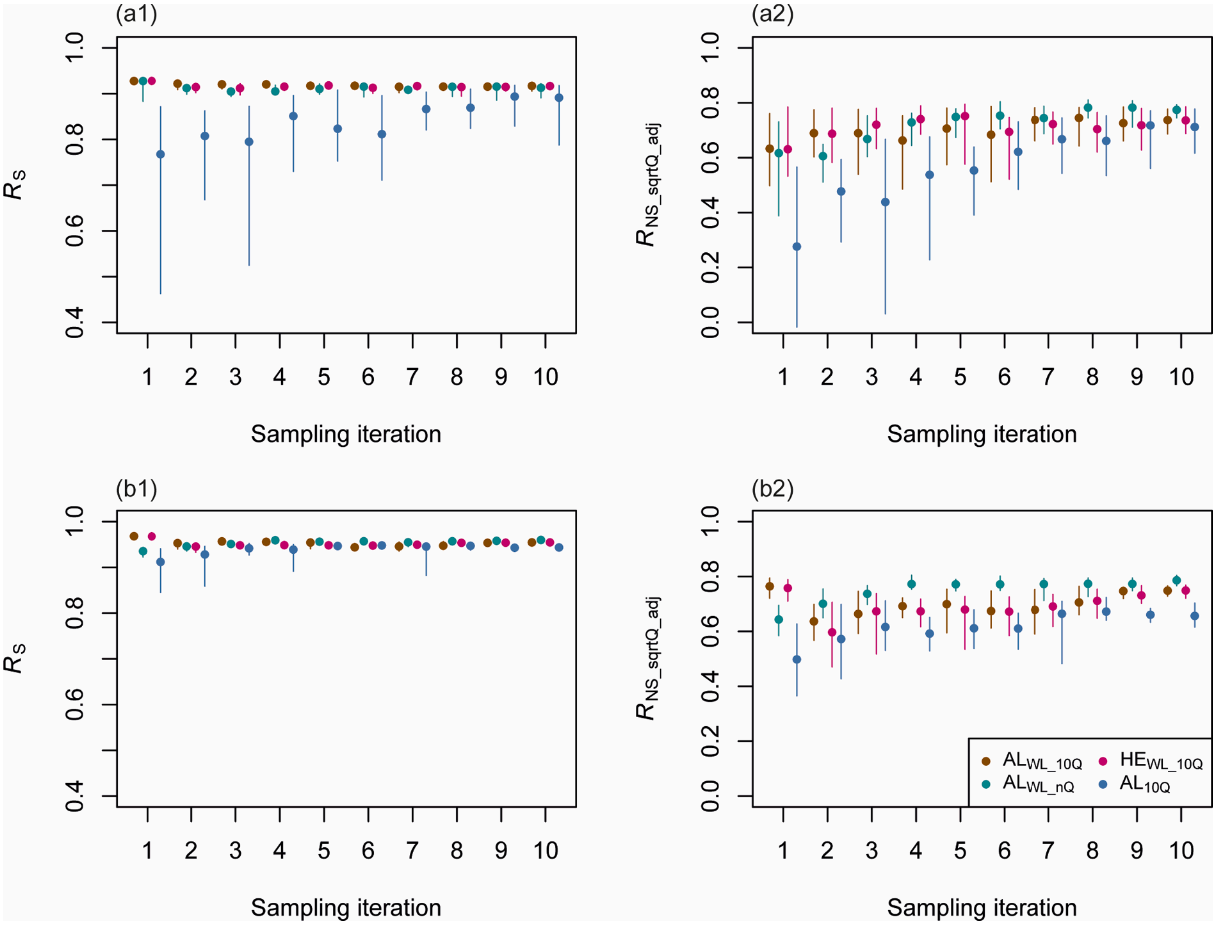

Fig. 4. Learning curves in the calibration period for the model performance metrics R_S and $R_{NS\_sqrtQ\_adj}$ (calibration metrics) as a function of the point discharge observations collected at ten sampling iterations using the data collection approaches AL_{WL_10Q} (active learning with water levels and ten discharge observations), AL_{WL_nQ} (active learning with water levels and *n* discharge observations), AL_{10Q} (active learning with ten discharge observations), and HE_{WL_10Q} (hydrological expert knowledge with water levels and ten discharge observations). The learning curves are shown for a year with an average precipitation of a) a snow-dominated catchment in the Northeast (top row; a1 for $R_S$and a2 for $R_{NS\_sqrtQ\_adj}$) and b) a rain-dominated catchment in the Northwest (bottom row; b1 for $R_S$and b2 for $R_{NS\_sqrtQ\_adj}$). The points indicate the median performance of all 100 calibration runs and the lines range from the 5th to 95th performance quantile. Note that the y-axis limits are different for the two performance metrics.

hydrological regime, we additionally selected the hydrological signatures of mean daily discharge, runoff ratio, and baseflow index. Maps with information on the geographical regions, hydrological regimes, climatic indices, and hydrological signatures can be found in the appendix (Figs. A.1 and A.2)

3. Results

3.1. *When are point discharge observations most informative?*

The characterization of discharge observations in terms of mean sampling date, seasonal concentration, and streamflow class allowed us to explore the timing of the most informative observations (Figs. 2 and 3). Using active learning to decide on the timing of discharge observations resulted in a strong spatial variability in the mean sampling date that tended to follow the annual peak discharge season. Mean sampling dates were thus observed in fall and winter for the Pacific Northwest and the mountainous regions of the Atlantic Coast states, in spring and early summer in the Rocky Mountains, the adjoining Great Basins and the Great Lakes Region, and in fall along the Gulf Coast. The seasonality in sampling dates was indirectly reflected in the distribution of streamflow classes that indicated a tendency towards observations during the peak and falling limb of events.

For AL_{WL_10Q} and AL_{10Q}, the concentration of informative discharge observation dates was most pronounced in snow-dominated catchments located in the Rocky Mountains, the Great Basins and the Great Lakes Region. As could be expected, discharge observations were spread across the year if an observation was assumed to exist at all time steps with a similar water level (AL_{WL_nQ}). The number of observations collected after ten iterations with AL_{WL_nQ} ranged from 32 to 303 (with an average of 67), whereby more observations were collected with increasingly arid conditions or with increasing importance of baseflow.

The use of active learning (AL_{WL_10Q} and AL_{10Q}) or hydrological expert knowledge (HE_{WL_10Q}) led to surprisingly similar selections of sampling dates as characterized by their mean. However, discharge observations were generally more distributed over the year with HE_{WL_10Q}, which was a direct result of collecting a range of flow classes at different days of the year.

3.2. *Learning curves: change of model performance with increasing availability of point discharge observations*

Learning curves illustrate the learning effect of a model (here, change in model performance) as a function of the additional information. These curves answer the practical question of how many sampling iterations are needed to reach a certain model performance. The learning curves in calibration and validation indicated that the iterative addition of point discharge observations for the calibration of HBV

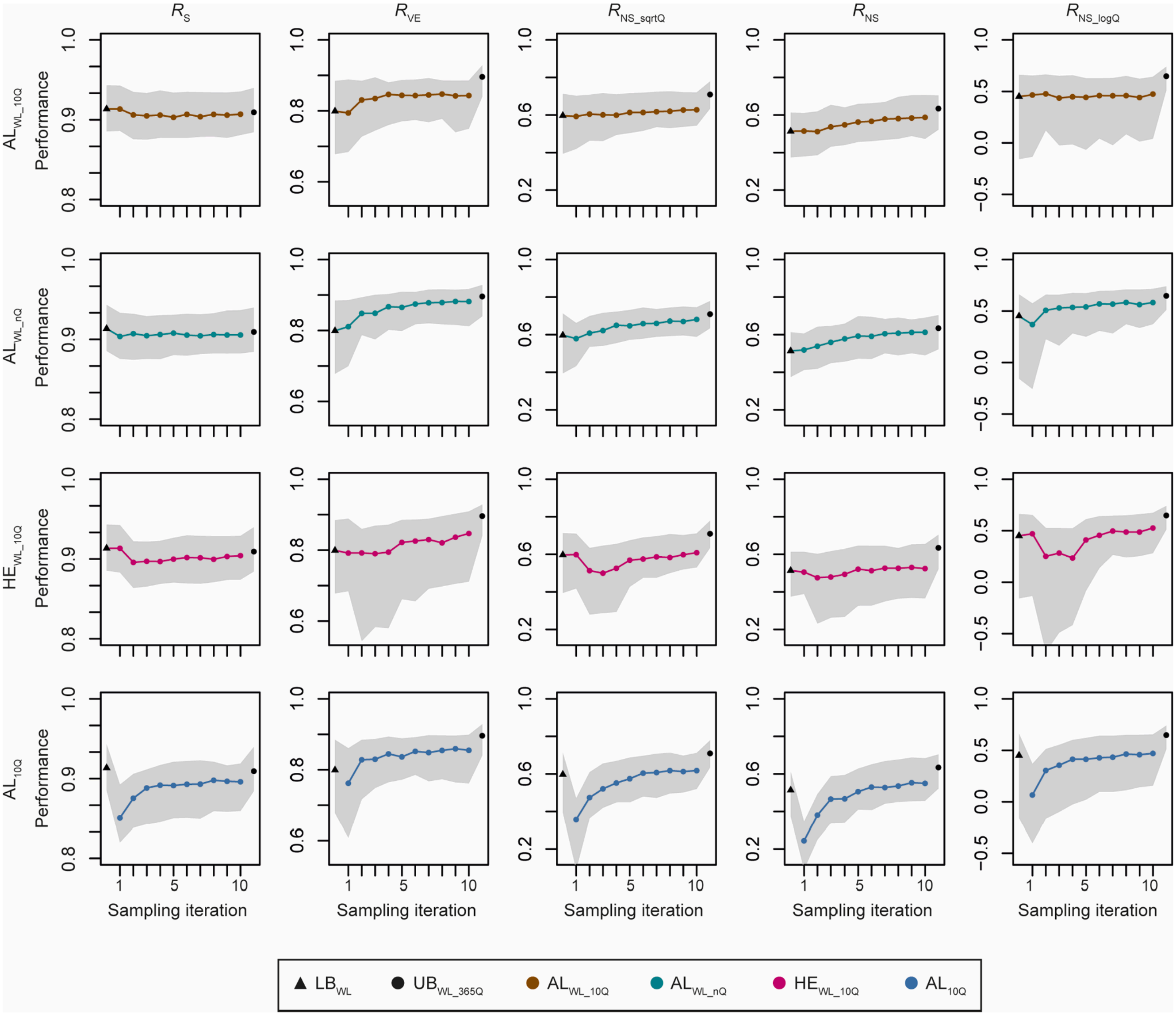

Fig. 5. Learning curves in the validation period for the model performance metrics R_S, R_{VE}, R_{NS_sqrtQ}, R_{NS}, and R_{NS_logQ} as a function of the point discharge observations collected at ten sampling iterations using the data collection approaches AL_{WL_10Q} (active learning with water levels and ten discharge observations), AL_{WL_nQ} (active learning with water levels and n discharge observations), AL_{10Q} (active learning with ten discharge observations), and HE_{WL_10Q} (hydrological expert knowledge with water levels and ten discharge observations). The coloured line indicates the median performance of all 100 catchments and the grey shaded area represents the 25th and 75th performance quantile. Note that the y-axis limits are different for different performance metrics. Some values extend below the lower limit of the y-axis and are plotted onto the x-axis.

generally increased model performance continuously. The added value of an additional observation decreased as the number of sampling iterations increased (Figs. 4 and 5).

Calibration results for two example catchments, one snow-dominated (Fig. 4a) and one rain dominated (Fig. 4b), suggest a more consistent performance over all 100 calibration runs with an increasing number of point discharge observations. This effect is stronger for calibration against point discharge observations only (AL_{10Q}) than for calibration against discharge observations and water levels (AL_{WL_10Q}, AL_{WL_nQ}, and HE_{WL_10Q}). The effect is also more pronounced for $R_{NS\_sqrtQ\_adj}$ than for R_S as the latter is by definition relatively well simulated by using continuous water level time series.

Validation results for all 100 catchments indicate that the value of point discharge observations for model calibration varied considerably between catchments (grey area in Fig. 5). The variability was lowest when data were collected using active learning (as opposed to using hydrological expert knowledge) or when simulations were evaluated focusing on mean or high flows (R_{NS}, R_{NS_sqrtQ}, and R_{VE}). Furthermore, the value of point discharge observations tended to become more similar across catchments for an increasing number of sampling iterations.

The median performance for all catchments was used to analyze the learning curves for the relative model performance R^* in the validation period (Fig. 6). Results indicated that the value of point discharge observations for improving water-level based model calibration was on average highest for annual volume estimates followed by high flows, mean flows and low flows. More specifically, model performance after ten sampling iterations improved by 58%-84% for R^*_{VE}, by 38%–93% for R^*_{NS}, by 22%–79% for $R^*_{NS\_sqrtQ}$, and by 10%–83% for $R^*_{NS\_logQ}$.

The majority of the simulation results are encouraging for the approach of collecting a few point discharge observations, whereby as few as two to six observations are typically already (highly) beneficial for model calibration. However, it is important to note that a small number of observations could, in some cases, also be disinformative for model calibration. This was especially the case for the collection of

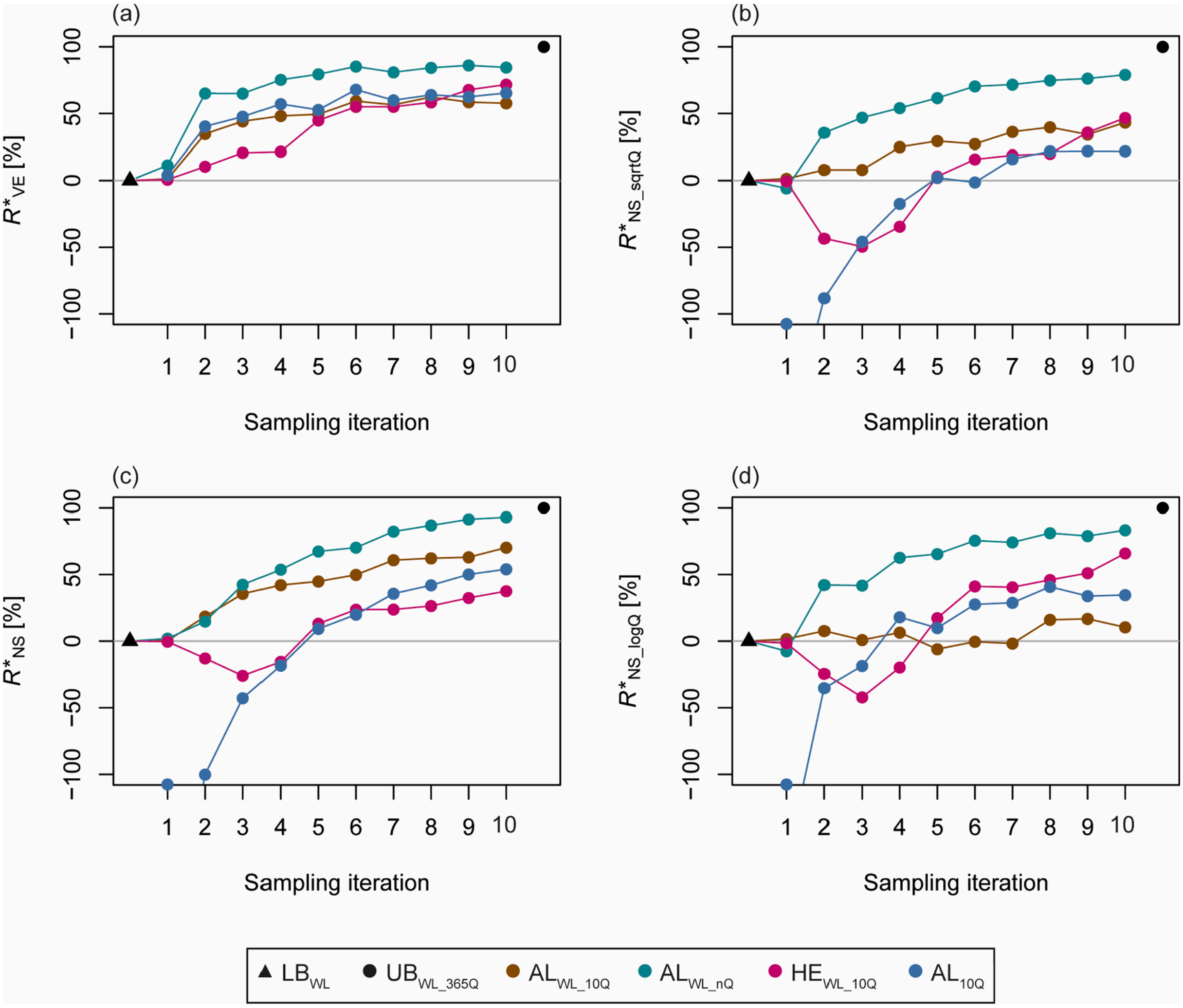

Fig. 6. Learning curves in the validation period for the relative model performance metrics R^*_{VE}, $R^*_{NS\_sqrtQ}$, R^*_{NS}, and $R^*_{NS\_logQ}$ as a function of the point discharge observations collected at ten sampling iterations using the data collection approaches AL_{WL_10Q} (active learning with water levels and ten discharge observations), AL_{WL_nQ} (active learning with water levels and *n* discharge observations), AL_{10Q} (active learning with ten discharge observations), and HE_{WL_10Q} (hydrological expert knowledge with water levels and ten discharge observations). The curves show the median performance for all 100 catchments. Note that values below the lower limit of the y-axis are plotted onto the x-axis.

discharge based on hydrological expert knowledge, where a calibration with less than five point discharge observations negatively affected the simulation of mean and low flows (Fig. 6). A further exception were simulations evaluated with R_S, for which model performance decreased when adding any discharge observations to a previous calibration against water levels (Fig. 5; note that this was expected since a calibration against water levels was based on R_S).

3.3. *Relative value of point discharge observations and water-level time series*

By looking at the relative value of discharge and water levels for model calibration, we analyzed for which hydrograph aspects (represented by the evaluation performance metrics) and for which catchments the two different types of data were more informative. The analysis was based on the validation model performance for each catchment after ten sampling iterations (Fig. 7). Spatial differences in the value of discharge and water levels are presented with a focus on the performance metric that showed the highest benefit of a certain data type (Figs. 8 and 9).

3.3.1. *Value of point discharge observations*

The comparison of performance metrics between AL_{WL_10Q} and LB_{WL} suggested that point discharge observations inform model calibration with information on streamflow volumes that is missing when only water levels were available (Fig. 7a). Point discharge observations were beneficial for all performance metrics (except for R_S), whereby the effect was most pronounced for R_{NS} (high flows). While simulated high flows were improved in the majority of catchments all over the contiguous United States, calibration against water levels and discharge was most valuable in (semi-) arid catchments (Figs. 8a and 9).

Point discharge observations collected with both AL_{WL_10Q} and HE_{WL_10Q} generally improved model performance. Yet the choice of the data collection approach could make a difference in the effectiveness of point discharge observations when evaluating simulations in terms of R_{NS} and R_{NS_logQ} (Fig. 7b). In arid catchments as well as in baseflow- or snowfall-dominated catchments low flows were better simulated when point discharge observations were selected based on HE_{WL_10Q}. In contrast, observations from AL_{WL_10Q} were more informative in these catchments for the simulation of high flows (Figs. 8b and 9). In relatively humid or rain dominated catchments R_{NS} and R_{NS_logQ} were not distinctly different if point discharge observations were selected based on HE_{WL_10Q} or AL_{WL_10Q}.

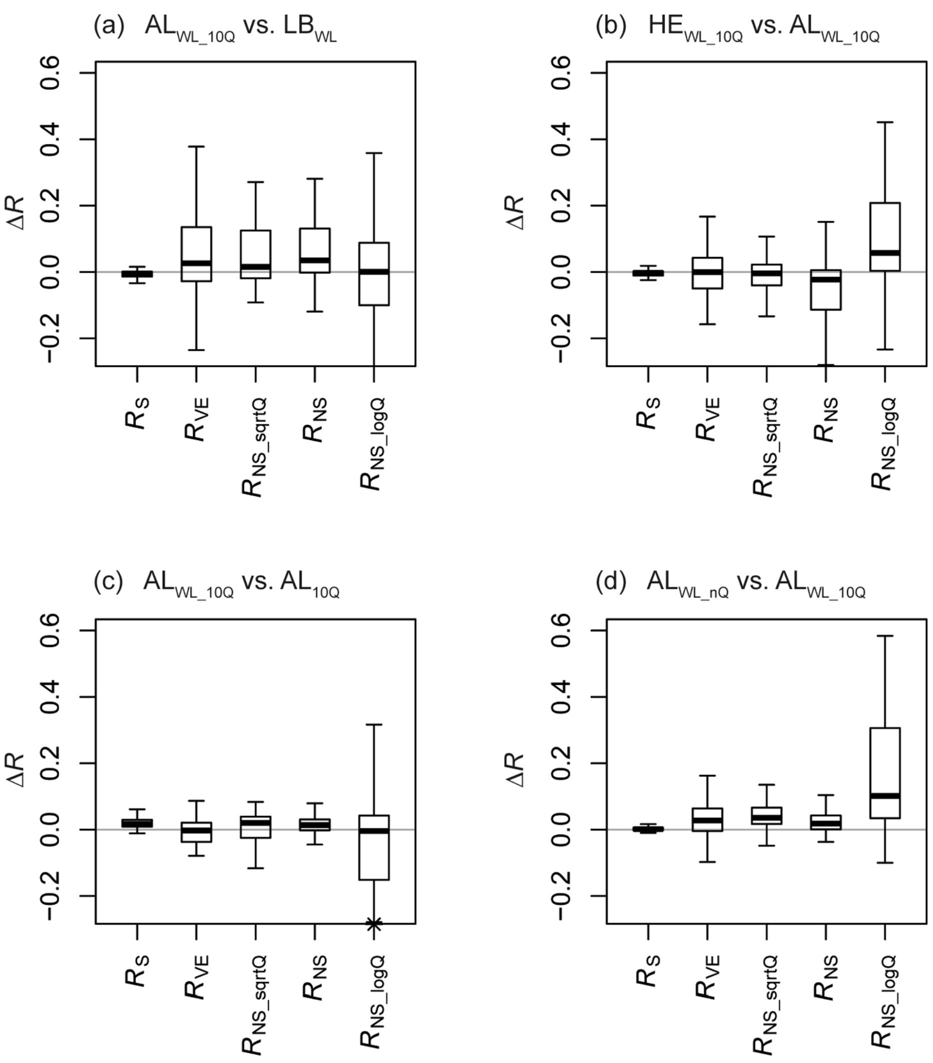

Fig. 7. Relative value of point discharge observations and water-level time series for model calibration. The relative value corresponds to the performance difference (ΔR) in the validation period between two data collection approaches after ten sampling iterations for each catchment. Positive values indicate an increased performance if (a) point discharge observations were used in addition to water-level time series, (b) hydrological expert knowledge was used to select point discharge observations as opposed to the use of active learning, (c) water-level time series are used in addition to point discharge observations only, and (d) point discharge observations were assumed to be representative for all dates with a similar water level. AL_{WL_10Q} is active learning with water levels and ten discharge observations, AL_{WL_nQ} is active learning with water levels and n discharge observations, AL_{10Q} is active learning with ten discharge observations, HE_{WL_10Q} is hydrological expert knowledge with water levels and ten discharge observations, and LB_{WL} is the lower benchmark with water levels. Note that the lower boxplot whisker extends to −0.47 in the case of R_{NS_logQ} in subplot (c) (marked by *).

3.3.2. Value of water-level time series

The value of water-level time series for model calibration was first evaluated by comparing simulations based on water levels and discharge (AL_{WL_10Q}) with simulations based on discharge only (AL_{10Q}). Results demonstrated that water levels improved the simulation of streamflow dynamics (R_S) in all catchments (Fig. 7c). Also, model performance for metrics sensitive to the dynamics of flow magnitudes (in particular R_{NS} and R_{NS_sqrtQ}) could often be slightly improved from the combined use of water levels and discharge. Spatially, water levels were most informative for model calibration in catchments with rather dry conditions and summer rainfall (Figs. 8c and 9).

Finally, the use of water-level time series to extend the observed discharge time series (AL_{WL_nQ}) led to an increased model performance for all metrics that evaluate volume-related hydrograph aspects (Fig. 7d). Thereby, low-flow simulations improved the most, especially in catchments with prolonged periods of relatively constant (low) flow conditions, such as arid or snow-influenced catchments (Figs. 8d and 9).

4. Discussion

4.1. Value of point discharge observations and water-level time series

Our results indicated that the collection of water-level and discharge data during a limited number of field visits could be highly valuable for predicting streamflow in previously ungauged catchments. Results thereby confirm earlier findings suggesting that a few months of continuous discharge observations (Brath et al., 2004; Melsen et al., 2014; Sun et al., 2017), or a small number of strategically timed discharge observations (Correa et al., 2016; McIntyre and Wheater, 2004; Pool et al., 2017; Seibert and McDonnell, 2015), can be very informative for model calibration.

Assuming that there is the opportunity to perform a number of streamflow observations, one needs to decide on when to measure which variable, i.e., discharge or water levels (Seibert and McDonnell, 2015). Our findings suggested that the combination of both types of data is advantageous over the use of either water levels or discharge. While continuous water-level time series provided information about streamflow dynamics, selected point discharge observations helped to link these dynamics to streamflow volumes.

As demonstrated by Seibert and Vis (2016), volume information is essential for the prediction of discharge in (semi-) arid catchments. In these catchments, the annual water balance is sensitive to actual evapotranspiration, and the corresponding model parameters could only be constrained if some information on volumes was also available. Independent of the hydroclimatic context, volume information was also found to be important at the event scale. While including point discharge observations in calibration improved the simulation of all flow conditions, model performance improved the most for annual volume estimates (R_{VE}). The improvement was furthermore larger for high flows than for low flows. This was likely because both active learning and hydrological expert knowledge resulted in the collection of a considerable number of observations at the peak and the falling limb of events,

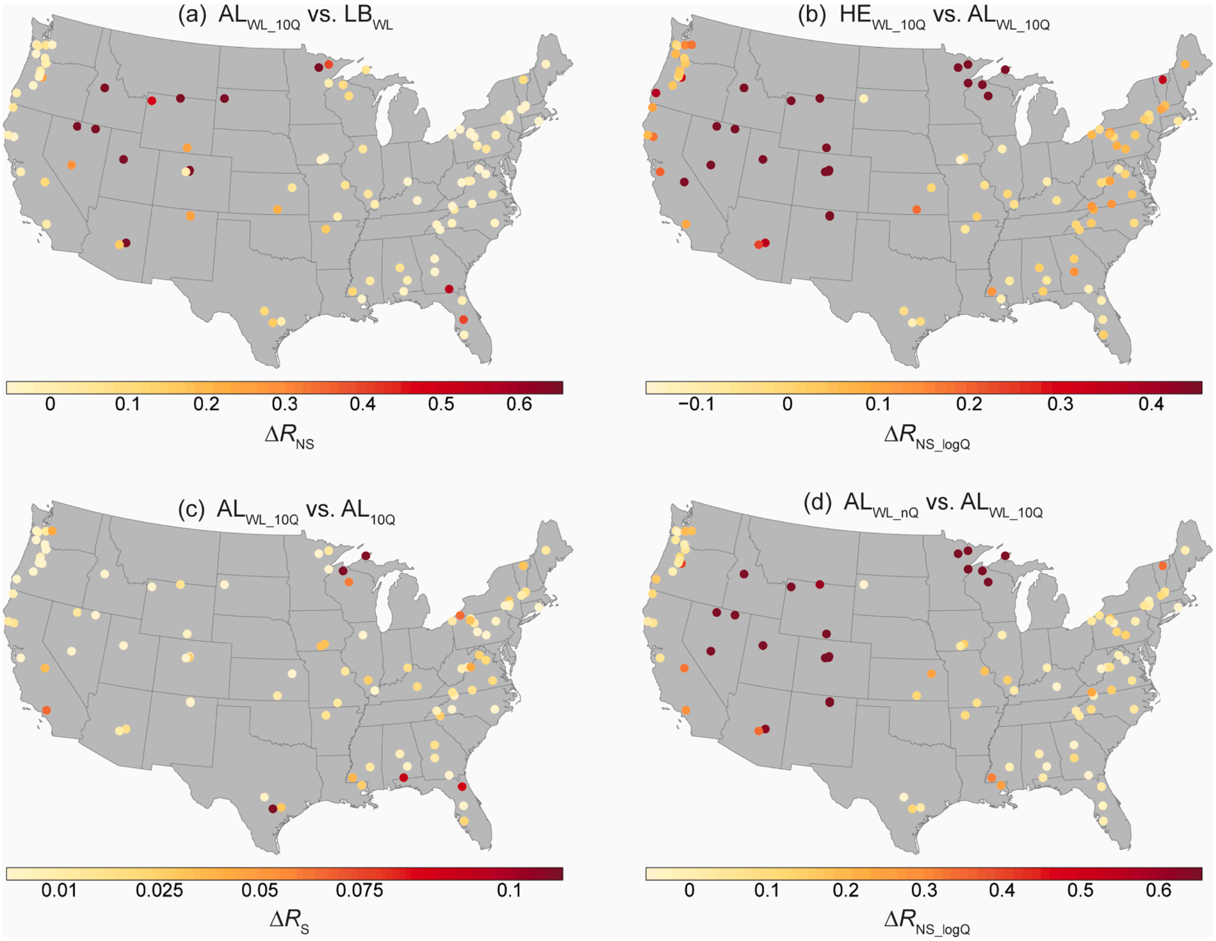
(a) $\mathrm{AL_{WL\_10Q}}$ vs. $\mathrm{LB_{WL}}$
0
0.1
0.2
0.3
0.4
0.5
0.6
ΔR_{NS}
(b) $\mathrm{HE_{WL\_10Q}}$ vs. $\mathrm{AL_{WL\_10Q}}$
−0.1
0
0.1
0.2
0.3
0.4
$\Delta R_{\mathrm{NS\_logQ}}$
(c) $\mathrm{AL_{WL\_10Q}}$ vs. $\mathrm{AL_{10Q}}$
0.01
0.025
0.05
0.075
0.1
ΔR_{S}
(d) $\mathrm{AL_{WL\_nQ}}$ vs. $\mathrm{AL_{WL\_10Q}}$
0
0.1
0.2
0.3
0.4
0.5
0.6
$\Delta R_{\mathrm{NS\_logQ}}$

Fig. 8. Spatial distribution of the relative value of point discharge observations and water-level time series for model calibration. The relative value corresponds to the performance difference (ΔR) in the validation period between two data collection approaches after ten sampling iterations for each catchment. Positive values indicate an increased performance if (a) point discharge observations were used in addition to water-level time series, (b) hydrological expert knowledge was used to select point discharge observations as opposed to the use of active learning, (c) water-level time series are used in addition to point discharge observations only, and (d) point discharge observations were assumed to be representative for all dates with a similar water level. AL$_{\mathrm{WL\_10Q}}$ is active learning with water levels and ten discharge observations, AL$_{\mathrm{WL\_nQ}}$ is active learning with water levels and *n* discharge observations, AL$_{\mathrm{10Q}}$ is active learning with ten discharge observations, HE$_{\mathrm{WL\_10Q}}$ is hydrological expert knowledge with water levels and ten discharge observations, and LB$_{\mathrm{WL}}$ is the lower benchmark with water levels. Note that the colour scales are different in (a) to (d).

which better constrained model parameters, influencing the intensity of the streamflow response to a given precipitation event.

Our results suggested that the installation of a water-level logger at the beginning of a field campaign is beneficial for two reasons. First, as opposed to a calibration exclusively based on point discharge observations, considering water-level information improved the simulation of any hydrograph characteristic related to streamflow dynamics in most of the catchments. This was likely due to the high temporal resolution of the water-level time series used for model calibration. The benefit of water levels was therefore especially pronounced in catchments where active learning and hydrological expert knowledge led to a temporally concentrated collection of point discharge observations. Second, as demonstrated by Lebecherel (2015), using water-level time series to extend the observed discharge time series could be an effective way to reduce the number of field trips, in particular, if a catchment is characterized by prolonged periods of similar flow conditions. However, results reported here have to be considered as optimistic, because model calibration for AL$_{\mathrm{WL\_nQ}}$ was based on the actual discharge values and not on values approximated by the originally ‘observed’ value.

4.2. *Value of active learning for the collection of point discharge observations*

A main objective of this study was to explore the value of active learning for selecting the most informative points in time for discharge observations as opposed to a decision based on hydrological expert knowledge. The use of active learning and hydrological expert knowledge resulted in surprisingly similar mean sampling dates. These sampling dates were typically aligned with hydrologically ‘active’ season(s). As a consequence, sampling dates were most concentrated in catchments with a pronounced annual peak flow, such as snow-dominated or winter-precipitation dominated catchments. Our results thereby indicate that active learning (i.e., model uncertainty) could guide the timing of point discharge observations towards hydrologically meaningful periods, which are generally in agreement with an expert’s decision on the timing of informative sampling dates. Furthermore, the results confirmed the importance of observations during subperiods of high parameter sensitivity (Harlin, 1991), especially when constraining model parameters with limited data.

The set of point discharge observations collected with active learning is the result of minimizing prediction uncertainty with the least number of observations. For this reason, the timing of such observations is typically model-specific (Crawford et al., 2013). For hydrological

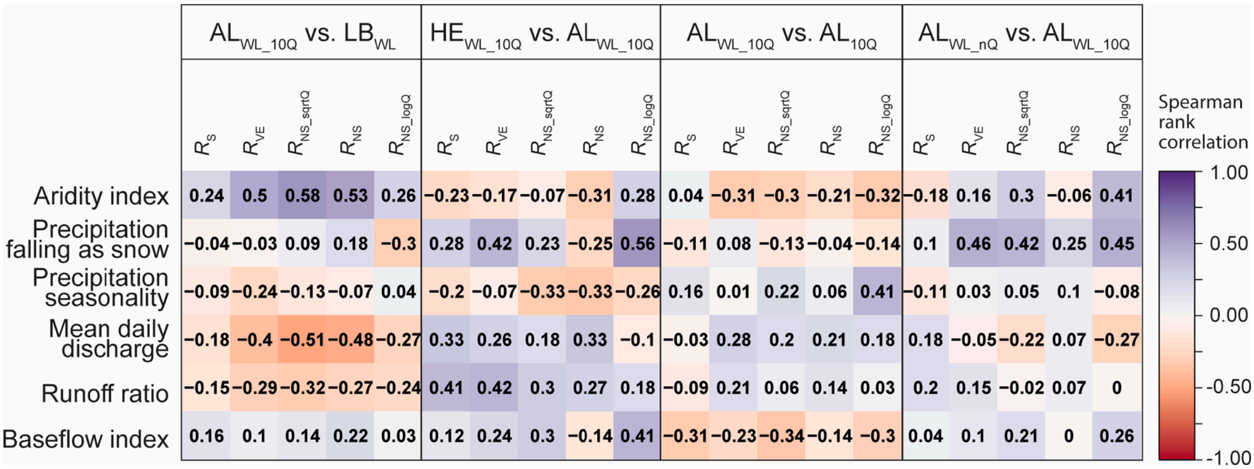

Fig. 9. Spearman rank correlation between catchment attributes and the relative value of point discharge observations and water-level time series for model calibration. The relative value corresponds to the performance difference (ΔR) in the validation period between two data collection approaches after ten sampling iterations for each catchment. Positive values indicate an increased performance if (a) point discharge observations were used in addition to water-level time series, (b) hydrological expert knowledge was used to select point discharge observations as opposed to the use of active learning, (c) water-level time series are used in addition to point discharge observations only, and (d) point discharge observations were assumed to be representative for all dates with a similar water level. AL_{WL_10Q} is active learning with water levels and ten discharge observations, AL_{WL_nQ} is active learning with water levels and n discharge observations, AL_{10Q} is active learning with ten discharge observations, HE_{WL_10Q} is hydrological expert knowledge with water levels and ten discharge observations, and LB_{WL} is the lower benchmark with water levels.

applications, this not only means that results could be different for different models, but also that the timing of the final set of point discharge observations collected by active learning is subject to model uncertainty and input uncertainty (in particular disinformative events; Beven and Westerberg, 2011). In contrast, the timing of point discharge observations based on hydrological expert knowledge is defined a priori, and their selection is therefore not directly affected by model uncertainty and input uncertainty. However, in the context of this study, active learning and hydrological expert knowledge were applied to the same hydrological model under identical forcing input. Results presented here for active learning and hydrological expert knowledge should therefore be directly comparable.

In this study, we applied active learning for the collection of point discharge observations during a hydrological year without respecting the temporal sequence of the observations. Results, and in particular model performance from calibrations with active learning, therefore provide an indication of how valuable active learning could be at best. Given the conceptual advantages and the practical limitations of active learning, we argue that active learning is especially valuable for complementing the collection of data based on expert knowledge. More specifically, expert knowledge could be used to decide on the timing of the first few field observations. Subsequently, active learning could guide the timing of additional measurements by providing information on flow situations that would be most informative.

4.3. Limitations of the study set-up

Our findings provided evidence that the prediction of streamflow in a previously ungauged basin can be greatly improved by the collection of a relatively small amount of local hydrological information. These encouraging results are based on a number of idealized assumptions that might be challenged when moving from a modelling study into practice.

The first major assumption made in this study was the perfectly known forcing time series. In practice, uncertain weather forecasts can lead to a too early or a delayed collection of point discharge observations. Results from previous studies with a limited number of streamflow or water level information suggested that a good coverage of a range of flow conditions is likely more important than the exact timing of observations (Etter et al., 2020; Pool et al., 2017). However, the importance of timing might depend on the flow regime of a catchment. Wright et al. (2015) thereby showed that the influence of single discharge observations on model performance could be considerably larger in an arid catchment than in a humid catchment. The effect of a mismatch in the timing of observations might also differ among flow classes. We expect that the importance of an accurate timing in observations is strongest for peak flows as they indicate the reactivity of a catchment to precipitation. In contrast, the timing might be less relevant during event recessions or baseflow conditions when similar hydrological processes dominate over a longer period. The sensitivity of model performance to the timing of point discharge observations is probably similar for active learning and expert knowledge because both data collection approaches led to a similar frequency of streamflow classes.

A further simplification of this study is the use of mean daily streamflow values as opposed to the use of instantaneous measurements taken during field visits. The difference between instantaneous discharge (discharge reported at 15-minutes interval) and mean daily discharge of the CAMELS dataset was small during recession and low-flow periods, but could be considerable during peak flows. This difference is expected to be most pronounced for either catchments or days with high streamflow variability and probably requires some attention when such field observations are used for model calibration.

Another basic assumption of this study were time-invariante rating curves. In practice, rating curves can change considerably due to changes in the cross-section of a river, backwater, or hysteresis effects (McMillan and Westerberg, 2015). Such changes can affect our modelling results in two ways. First, water-level time series were derived from discharge time series (see Section 2.1) and substantial intra-annual changes in the rating curve could mislead model parameterization. Second, the success of the active learning approach in which point discharge observations were assumed to be valid for all time steps with comparable water level (AL_{WL_nQ}) relies on a time-invariant rating curve. The value of AL_{WL_nQ} might therefore be overestimated in catchments with considerable rating curve changes within a hydrological year.

5. Conclusions

Long continuous discharge time series representing a variety of hydrological conditions are usually seen as a requirement for model calibration. In practice, many catchments have no, or only limited,

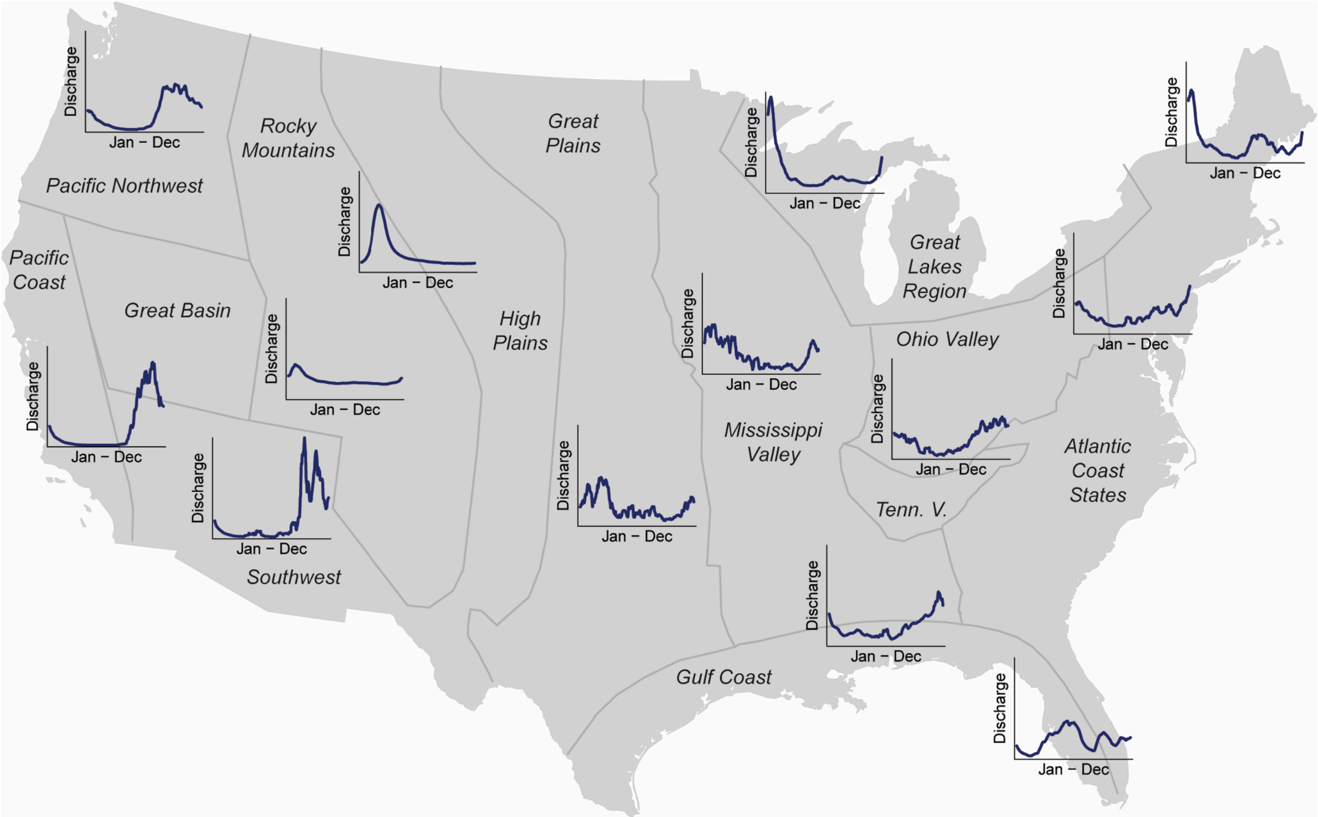

Fig. A1. Hydrological regimes across different regions of the contiguous United States. The regimes are shown for a selection of catchments representative for the 100 study catchments. The regimes were calculated as the mean daily discharge of each day of the hydrological years 1990–2009, whereby daily values were smoothed by calculating a moving average over 15 days. The regimes were normalized by their mean streamflow value. The regions are delineated according to NOAA (2020).

discharge data. Understanding which, and how much data is most valuable for model calibration, is essential to improve the prediction in ungauged basins. In this study, we contributed to an improved understanding by explicitly comparing the relative value of water-level time series and point discharge observations for model calibration, and by testing a machine learning approach to determine when to collect such a limited number of discharge observations. Based on results from simulation experiments for 100 hydroclimatically diverse catchments, the following conclusions can be drawn:

- A small number of point discharge observations contained, surprisingly, a lot of information, and can considerably improve model calibrations based on water-level time series with respect to annual and event-scale streamflow volumes. While model performance continuously improved as the number of observations increased, the incremental improvements were most considerable for the first two to six observations. The value of point discharge observations was highest for (semi-) arid catchments and for the simulation of annual volumes.
- Continuous water-level time series provided valuable information for the simulation of daily streamflow dynamics. Furthermore, water-level time series could reduce the number of field trips if a point discharge observation was assumed to exist at all time steps with a similar water level. Such an extension of the number of discharge observations was most effective in catchments with prolonged periods of relatively constant flow conditions.
- Choosing the date of point discharge observations based on active learning led to similar sampling dates as the selection of dates according to hydrological expert knowledge. In both cases, most observations were selected for the seasons with the highest flows.

Our findings encourage the approach to gauge an ungauged catchment with discharge observations on strategically selected dates. Independent of the geographic region, the most informative sampling dates are typically expected to take place during hydrologically active periods, such as the annual peak discharge, other discharge peaks, and recession situations. Two to six observations during these periods can be already very informative. However, increasing the number of observations to ten allows the collection of additional discharge observations in periods of more constant flow conditions, which is beneficial for a more balanced evaluation of different flow conditions during model calibration. The exact timing of the first few discharge observations could be defined with hydrological expert knowledge, active learning could then be valuable for guiding the timing of the additional observations. Combining such a small number of point discharge observations with (short) continuous water level time series is a promising way towards improved predictions in ungauged basins.

Author contributions

SP and JS designed this study; SP performed the hydrological simulations; SP and JS analyzed and discussed the results; writing of the paper was led by SP with contributions of JS.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

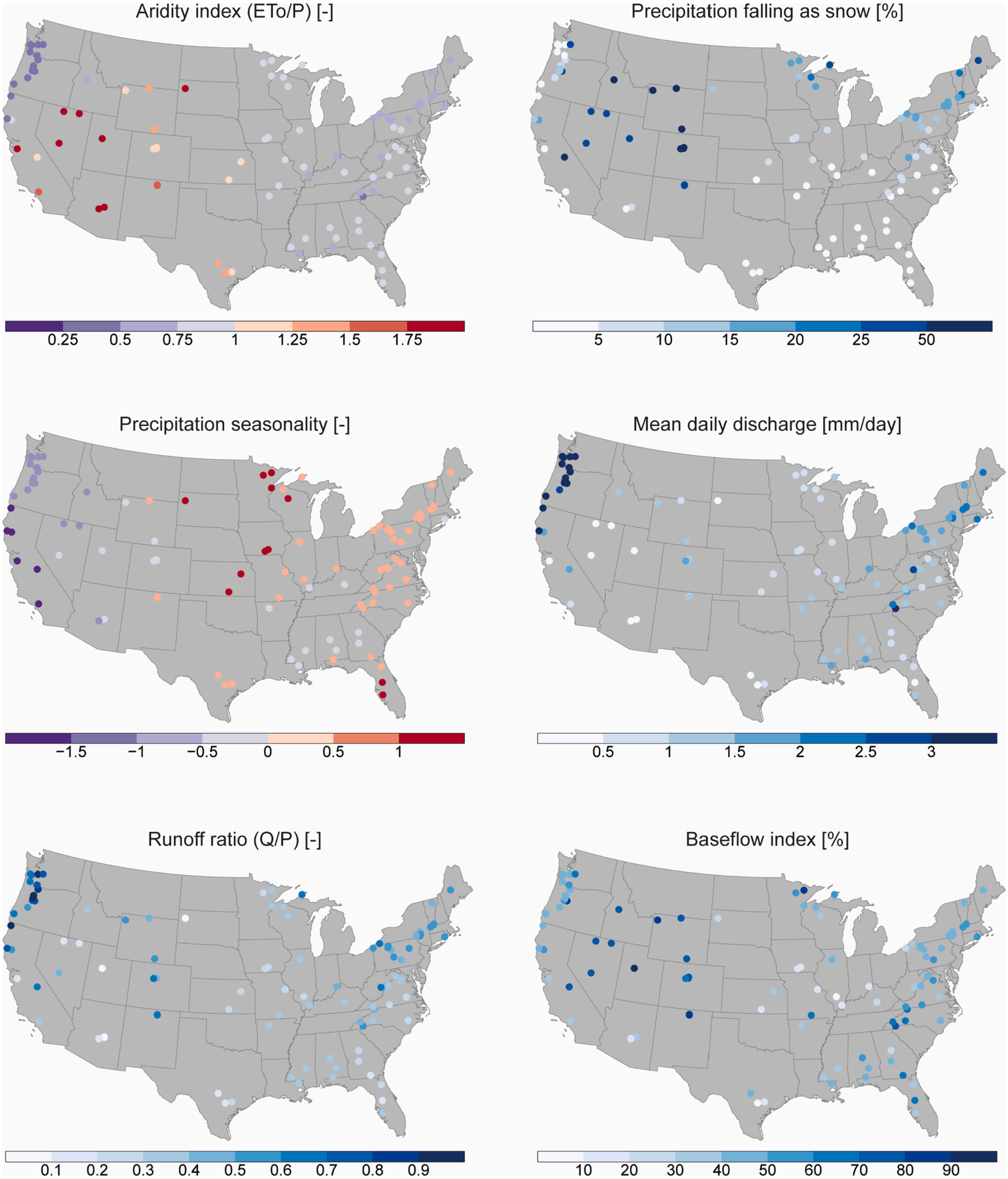

Fig. A2. Hydroclimatic attributes of the 100 study catchments used to explain the spatial distribution in model performance. The aridity index was calculated as the ratio of the sum of potential evapotranspiration and the sum of precipitation (ETo/P). The runoff ratio was calculated as the ratio of the sum of discharge and sum of precipitation (Q/P). The baseflow index was calculated using the EflowStats R-Package from the U.S. Geological Survey (2014).

Acknowledgements

The computations required for this study were only possible thanks to the cloud computing opportunities provided by S3IT at the Faculty of Science of the University of Zurich. Hydrometeorological time series, catchment attributes, and catchment boundaries were extracted from the data sets of Newman et al. (2015) and Addor et al. (2017). SRTM elevation data were used from Jarvis et al. (2008).

Appendix A:. Runoff regimes and hydroclimatic characteristics of the 100 study catchments

References

Addor, N., Newman, A.J., Mizukami, N., Clark, M.P., 2017. The CAMELS data set: catchment attributes and meteorology for large-sample studies. Hydrol. Earth Syst. Sci. 21 (10), 5293–5313.

Addor, N., Nearing, G., Prieto, C., Newman, A.J., Le Vine, N., Clark, M.P., 2018. Selection of hydrological signatures for large-sample hydrology. Water Resour. Res. 54 (11), 8792–8812.

Berghuijs, W.R., Sivapalan, M., Woods, R.A., Savenije, H.H.G., 2014. Patterns of similarity of seasonal water balances: a window into streamflow variability over a range of time scales. Water Resour. Res. 50 (7), 5638–5661.

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

Beven, K., Westerberg, I., 2011. On red herrings and real herrings: disinformation and information in hydrological inference. Hydrol. Process. 25 (10), 1676–1680.

Brath, A., Montanari, A., Toth, E., 2004. Analysis of the effects of different scenarios of historical data availability on the calibration of a spatially-distributed hydrological model. J. Hydrol. 291 (3-4), 232–253.

Brunner, M., Pool, S., Kiewiet, L., Acheson, E., 2018. The other's perception of a streamflow sample: from a bottle of water to a data point. Hydrol. Process. 32 (18), 2922–2927.

Cawley, G.C., 2011. Baseline methods for active learning. In: Proceedings of the Workshop on Active Learning and Experimental Design, pp. 47–57.

Coopersmith, E.J., Minsker, B.S., Sivapalan, M., 2014. Patterns of regional hydroclimatic shifts: an analysis of changing hydrologic regimes. Water Resour. Res. 50 (3), 1960–1983.

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 (22), 4059–4073.

Crawford, M.M., Tuia, D., Yang, H.L., 2013. Active learning: any value for classification of remotely sensed data? Proc. IEEE 101 (3), 593–608.

Etter, S., Strobl, B., Seibert, J., van Meerveld, H.J.I., 2020. Value of crowd-based water level class observations for hydrological model calibration. Water Resour. Res. 56 e2019WR026108.

Girons Lopez, M., Seibert, J., 2016. Influence of hydro-meteorological data spatial aggregation on streamflow modelling. J. Hydrol. 541, 1212–1220.

Hall, J., Blöschl, G., 2018. Spatial patterns and characteristics of flood seasonality in Europe. Hydrol. Earth Syst. Sci. 22 (7), 3883–3901.

Harlin, J., 1991. Development of a process oriented calibration scheme for the HBV hydrological model. Hydrol. Res. 22 (1), 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 (6), 1198–1255.

Jarvis, A., Reuter, H., Nelson, A., and Guevara, E., 2008. Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m. Retrieved from http://srtm.csi.cgiar.org.

Johansson, B., 2000. Areal precipitation and temperature in the Swedish mountains: an evaluation from a hydrological perspective. Hydrol. Res. 31 (3), 207–228.

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.: Int. J. 23 (26), 3705–3717.

Lebecherel, L., 2015. Sensibilité des calculs hydrologiques à la densité des réseaux de mesure hydrométrique et pluviométrique. Doctoral dissertation. Irstea, AgroParisTech, Antony, FR.

Lewis, D.D., Gale, W.A., 1994. In: SIGIR `94. Springer London, London, pp. 3–12. https://doi.org/10.1007/978-1-4471-2099-5_1.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., Bergström, S., 1997. Development and test of the distributed HBV-96 hydrological model. J. Hydrol. 201 (1-4), 272–288.

Mathevet, T., Michel, C., Andreassian, V., Perrin, C., 2006. A bounded version of the Nash-Sutcliffe criterion for better model assessment on large sets of basins. IAHS Publ. 307, 211–219.

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 (1-2), 100–116.

McMillan, H.K., Westerberg, I.K., 2015. Rating curve estimation under epistemic uncertainty. Hydrol. Process. 29 (7), 1873–1882.

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.

Merz, R., Parajka, J., Blöschl, G., 2009. Scale effects in conceptual hydrological modeling. Water Resour. Res. 45 (9). W09405.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models. Part I - A discussion of principles. J. Hydrol. 10 (3), 282–290.

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 (1), 209–223.

NOAA, 2020. Geographical reference maps, available at: <https://www.ncdc.noaa.gov/temp-and-precip/drought/nadm/geography>.

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 (1), 131–151.

Pewsey, A., Neuhäuser, M., Ruxton, G.D., 2013. Circular statistics in R. Oxford University Press, Oxford, UK.

Pool, S., Viviroli, D., Seibert, J., 2017. Prediction of hydrographs and flow-duration curves in almost ungauged catchments: Which runoff measurements are most informative for model calibration? J. Hydrol. 554, 613–622.

Pool, S., Vis, M., Seibert, J., 2018. Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency. Hydrol. Sci. J. 63 (13-14), 1941–1953.

Priestley, C.H.B., Taylor, R.J., 1972. On the assessment of surface heat flux and evaporation using large-scale parameters. Mon. Weather Rev. 100 (2), 81–92.

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

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., Vis, M.J., 2012. Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. Hydrol. Earth Syst. Sci. 16 (9), 3315–3325.

Seibert, J., McDonnell, J.J., 2015. Gauging the ungauged basin: relative value of soft and hard data. J. Hydrol. Eng. 20 (1). A4014004.

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

Settles, B., 2012. Active learning. In: Brachman, R.J., Cohen, W.W., Dietterich, T. (Eds.), Synthesis Lectures on Artificial Intelligence and Machine Learning (Lecture 18). Morgan and Claypool, San Rafael, CA.

Sikorska, A.E., Viviroli, D., Seibert, J., 2015. Flood-type classification in mountainous catchments using crisp and fuzzy decision trees. Water Resour. Res. 51 (10), 7959–7976.

Singh, S.K., Bárdossy, A., 2012. Calibration of hydrological models on hydrologically unusual events. Adv. Water Resour. 38, 81–91.

Spearman, C., 1904. The proof and measurement of association between two things. Am. J. Psychol. 15 (1), 72–101.

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 (1), 251–265.

Tada, T., Beven, K.J., 2012. Hydrological model calibration using a short period of observations. Hydrol. Process. 26 (6), 883–892.

Tan, S.B., Chua, L.H., Shuy, E.B., Lo, E.-M., Lim, L.W., 2008. Performances of rainfall-runoff models calibrated over single and continuous storm flow events. J. Hydrol. Eng. 13 (7), 597–607.

U.S. Geological Survey, 2014. EflowStats R-package, available at: https://www.github.com/USGS-R/EflowStats.

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 (1-4), 288–307.

Wallace, J.M., Hobbs, P.V., 2006. Atmospheric science: an introductory survey. In: Dmowksa, R., Hartmann, D., Rossby, H.T. (Eds.), International Geophysics Series, second ed. Academic Press, Canada.

Wright, D.P., Thyer, M., Westra, S., 2015. Influential point detection diagnostics in the context of hydrological model calibration. J. Hydrol. 527, 1161–1172.

Wright, D.P., Thyer, M., Westra, S., McInerney, D., 2018. A hybrid framework for quantifying the influence of data in hydrological model calibration. J. Hydrol. 561, 211–222.

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.