HYDROLOGICAL SCIENCES JOURNAL
https://doi.org/10.1080/02626667.2020.1735638

SPECIAL ISSUE: HYDROLOGICAL DATA: OPPORTUNITIES AND BARRIERS

Hydrological model calibration with uncertain discharge data

Ida K. Westerberg[a], Anna E. Sikorska-Senoner[b], Daniel Viviroli[b], Marc Vis[b] and Jan Seibert[b,c]

[a]IVL Swedish Environmental Research Institute, Stockholm, Sweden; [b]Department of Geography, University of Zurich, Zurich, Switzerland; [c]Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden

ABSTRACT

Discharge data used to calibrate and evaluate hydrological models can be highly uncertain and this uncertainty affects the conclusions that we can draw from modelling results. We investigated the role of discharge data uncertainty and its representation in hydrological model calibration to give recommendations on methods to account for data uncertainty. We tested five different representations of discharge data uncertainty in calibrating the HBV-model for three Swiss catchments, ranging from using no information to using full empirical probability distributions for each time step. We developed a new objective function to include discharge data uncertainty, as quantified by these distributions directly in calibration to hydrological time series. This new objective function provided more reliable results than using no data uncertainty or multiple realizations of discharge time series. We recommend using the new objective function in combination with empirical or triangular distributions of the discharge data uncertainty.

ARTICLE HISTORY
Received 28 May 2019
Accepted 6 January 2020

EDITOR
A. Castellarin

GUEST EDITOR
S. Uhlenbrook

KEYWORDS
discharge data uncertainty; model calibration; rating curve uncertainty; objective function; HBV model; model evaluation

1. Introduction

Uncertainty in discharge data propagates to uncertainty about the conclusions that we can draw from hydrological analyses based on discharge data. For example, this uncertainty may obscure detection of temporal change (Juston *et al.* 2014, Wilby *et al.* 2017), identification of design floods (Di Baldassarre *et al.* 2012, Steinbakk *et al.* 2016), analyses of differences in catchment behaviour (Westerberg *et al.* 2016), and identification of reliable model-based estimates (Liu *et al.* 2009, Sikorska *et al.* 2013). Discharge data uncertainties can also directly propagate to increased costs and sub-optimal decisions in water management (McMillan *et al.* 2017a). It is, therefore, essential to take data uncertainties into account in the design of hydrological analyses so that their impacts can be quantified or excluded (Westerberg and McMillan 2015).

The main source of discharge data uncertainty is typically the indirect calculation of discharge from stage (water level) using a model of the stage–discharge relationship at the gauging site. This model (i.e., the rating curve) is fitted to stage–discharge gauging pairs, which have been measured simultaneously at different flow conditions. Ideally, such gauging pairs should be collected over the entire rating curve range, but this is usually impossible as extremely low and high flows occur, by definition, seldomly. The rating curve is, thus, often particularly uncertain at these extreme flows as a result of the rating curve extrapolation. Typical discharge uncertainties are in the order of ±20–80% for low flows, ±10–15% for average flows, and ±15–40% for high flows (McMillan *et al.* 2012, Westerberg *et al.* 2016), but there is a sizeable site-specific variability (Coxon *et al.* 2015), in particular where the stage–discharge relationship varies temporally (Jalbert *et al.* 2011) or is not unique because of hysteresis or backwater effects (Mansanarez *et al.* 2016).

Quantification of discharge uncertainty has long been researched (Herschy 1970, Pelletier 1988), and in recent years new methods for the estimation of rating-curve uncertainty have been developed, ranging from methods producing estimates of upper and lower uncertainty limits (Westerberg *et al.* 2011a, Coxon *et al.* 2015) to those using Bayesian Markov Chain Monte Carlo (MCMC) techniques to estimate a full posterior distribution of discharge (Petersen-Overleir *et al.* 2009, Sikorska *et al.* 2013, Le Coz *et al.* 2014, Juston *et al.* 2014, McMillan and Westerberg 2015). Apart from the uncertainty estimation technique, the methods differ in their treatment of temporal variability in the rating curve uncertainty (Tomkins 2012, Morlot *et al.* 2014) and in their assumptions about the separation of different uncertainty components (i.e., measurement, parameter, and structural uncertainty). Kiang *et al.* (2018) provide a comprehensive comparison of different rating curve uncertainty estimation methods and find the largest differences between methods at low and high flows, when the rating curve varies with time, and when it is extrapolated to ungauged flows. Uncertainty in rating curve estimates directly leads to uncertainty in discharge data based on these estimates.

In this study we investigated the impact of discharge data uncertainty on model calibration and evaluation. When calibrating a hydrological model, discharge data uncertainty can obscure and bias model parameter identification, affect simulation results, and lead to wrong conclusions about the model structure and its performance. Several methods to account for discharge data uncertainty in model calibration have been developed (e.g. Liu *et al.* 2009, Thyer *et al.* 2009, McMillan

CONTACT Ida K. Westerberg ✉ ida.westerberg@ivl.se IVL Swedish Environmental Research Institute, Stockholm, Sweden

Supplemental data for this article can be accessed here.

et al. 2010, Westerberg *et al.* 2011b, Sikorska and Renard 2017). Such calibration methods differ primarily in four methodological aspects linked to: (a) the type of discharge data uncertainty information that is used, e.g. upper and lower bounds (Coxon *et al.* 2013), or full empirical discharge realizations or distributions (McMillan *et al.* 2010); (b) the assumptions about the discharge data errors, e.g. assumptions about the temporal error autocorrelation; (c) the type of calibration, i.e., whether the objective function uses the discharge time series directly (Liu *et al.* 2009), or is based on hydrological signatures to which the discharge uncertainty is propagated (Blazkova and Beven 2009); and (d) the assumptions about how the uncertainty in discharge data interacts with other uncertainty components such as uncertainty in input data, model structure and model parameters (Renard *et al.* 2010, Krueger *et al.* 2010, Westerberg and Birkel 2015).

The aim of this study was to investigate the role of discharge data uncertainty in hydrological model calibration to give recommendations on methods to account for data uncertainty. We focused on the type of discharge data uncertainty information that is needed (the first of the four abovementioned methodological aspects) and we limited the study to model calibration against uncertain discharge time series directly, i.e., we did not investigate calibration to hydrological signatures (the third aspect above). We focused our analysis on three specific objectives:

(1) What is the impact of including or excluding discharge data uncertainty in model calibration on the resulting simulations?
(2) How much information about the discharge data uncertainty distribution is needed in the objective function to obtain reliable model simulations?
(3) What is the impact of allowing for model simulations outside the estimated discharge data uncertainty in the objective function?

To be able to address these questions, the discharge data uncertainty needs to be incorporated into the model calibration. We developed a new objective function capable of incorporating different types of information on discharge data uncertainty and tested it for three Swiss catchments with different properties and rating curves.

2. Study catchments, data and discharge data uncertainty estimation

2.1. Study catchments

For our study, we selected three Swiss meso-scale catchments (Fig. 1) that had comprehensive information available about rating curves, gaugings and site-specific conditions. One catchment is located in the Bernese Alps (Kander-Hondrich) and has a small areal glacier cover, whereas the other two catchments extend from the pre-alps to the Swiss Plateau (Broye-Payerne in the West and Wigger-Zofingen in northern central Switzerland) and are not glacierized (Table 1).

2.2. Data for discharge uncertainty estimation

We used the stage–discharge gauging data and the 10-minute water-level time series data (1980–2014) for each catchment outlet gauging station to estimate rating-curve and discharge uncertainty. For each gauging station, more than 150 stage–discharge gauging pairs were available for this period (193, 165, and 153 gaugings at Payerne, Hondrich, and Zofingen respectively) together with around 40 official rating curves at each station that had been used historically. These data were analysed for temporal variability, and we found neither major systematic variability nor major change to the stage–discharge relationship for the stations in the analysed period. A single rating-curve uncertainty estimate for the 35 years was therefore used for each of the stations.

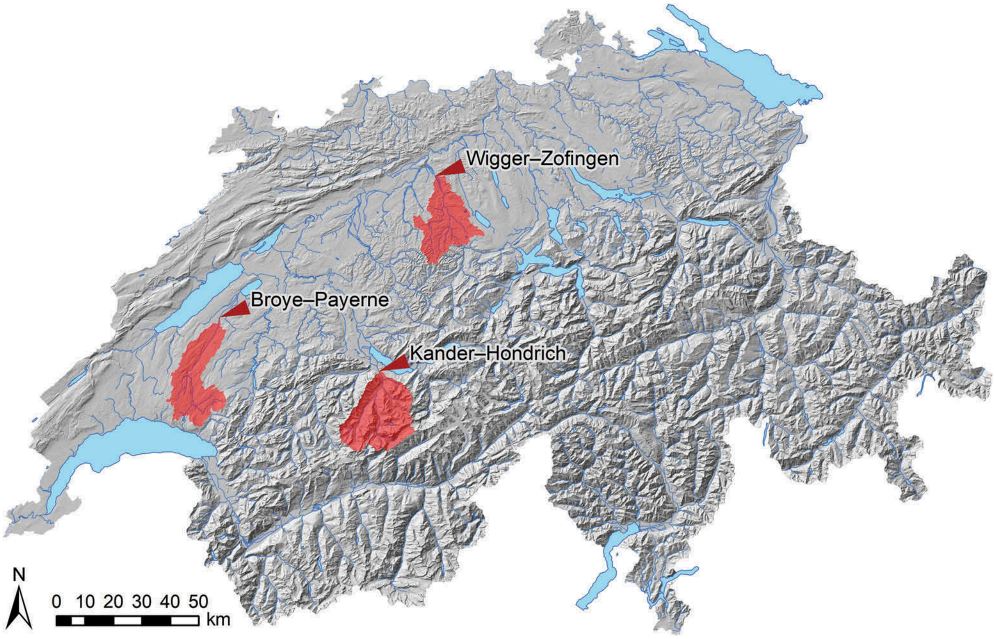

Figure 1. Map of Switzerland with the location of the three study catchments and their gauging stations.

Table 1. Main characteristics of the catchments used in this study (from Weingartner and Aschwanden 1992).

River and station	Catchment area (km^2)	Station elevation (m a.s.l.)	Mean catchment elevation (m a.s.l.)	Areal glacier cover (%)	Regime type
Kander-Hondrich	491	650	1900	7.9	b-glacio-nival (dominated by ice and snow melt runoff processes)
Broye-Payerne	392	441	710	0	pluvial inférieur (dominated by rainfall–runoff processes, related to lower elevation bands)
Wigger-Zofingen	368	426	660	0	pluvial inférieur

For Zofingen, the rating curve was extrapolated for the highest 1.5 m (i.e. the maximum water level in 1980–2014 was 1.5 m above the highest stage–discharge gauging). We extended the available measured stage–discharge data with two stage–discharge pairs calculated with the Gauckler-Manning-Strickler formula:

$$v = \frac{1}{n} . R^{2/3} . S^{1/2} \quad (1)$$

where v is the flow velocity, n the Manning roughness coefficient, R the hydraulic radius (wetted perimeter divided by cross-section area) and S the friction slope (see, for instance, Chow *et al.* 1988). Using a riverbed cross-section at the gauging site, discharge Q for a given stage h can be computed as v multiplied by the cross-section area A. As no accurate information was available for slope S at this site, we estimated S using the roughness coefficient n for which detailed information was available, together with the discharge from the two highest gaugings. Because these two gaugings deviated a lot from each other, we calculated one estimate of S for each gauging using Equation (1). To account for data uncertainties, we assumed normally distributed errors with 95% ranges of ±4% for gauged discharge Q_g, ±2 cm for gauged stage h_g, and ±20% for roughness coefficient n (Wyder 1998; Hanspeter Hodel, 10.03.2017, Federal Office for the Environment, personal communication). We then computed two sets of 1,000 realizations of Manning estimates of the discharge $Q_{\max}$ at the highest recorded stage $h_{\max}$, with random perturbations of Q_g, h_g, and n as specified above. These 1,000 estimates of $Q_{\max}$ were approximately normally distributed, and we used the average $Q_{\max}$ from each of the 1,000 estimates and its 95% uncertainty range to specify two additional stage–discharge gaugings with uncertainty. These two calculated gaugings were then added to the existing gauging dataset for Zofingen, to help constrain the high flows in the rating curve uncertainty estimation.

2.3. Discharge data uncertainty estimation

Using the available and extended stage–discharge pairs, we then estimated rating curve uncertainty for each site in a Monte Carlo analysis using the Voting Point likelihood method (McMillan and Westerberg 2015). This method accounts for uncertainties in the measured stage–discharge gauging data and in the rating-curve model approximation of the (unknown) true stage–discharge relationship at the cross-section of the gauging station. We assumed a power-law rating-curve function, commonly applied in rating-curve estimation and hydraulic studies (Le Coz *et al.* 2014). Based on information from the local monitoring agency (Hanspeter Hodel, 10.03.2017, Federal Office for the Environment, personal communication), we assumed normally distributed errors for discharge gauging uncertainty with 95% bounds at ±4% (current meters), ±15% (float gaugings), and ±6% (ADCP, salt dilution, and other techniques) of the measured values. For stage we used a uniform error of ±5 mm for low to medium stages and ±20 mm for high stages (i.e. >95th percentile of the stage time series). The rating curve parameter priors were set to standardized ranges, similarly to Westerberg *et al.* (2016), and were adjusted for each station if necessary.

The rating curve uncertainty assessment resulted in a posterior distribution of 40,000 feasible power-law parameter sets for each station. This rating curve uncertainty was then propagated to the discharge time series by calculating a corresponding discharge value from each rating curve realization for each stage value in the stage time series. This resulted in a set of 40,000 discharge time series realizations that can be used directly in the model calibration. Alternatively, a distribution such as the empirical probability distribution function (pdf) or a triangular distribution can be estimated from the 40,000 discharge values for each time step, and then used for the model calibration. Note that these distributions are identical for the same stage values but that they vary in shape with stage and therefore vary along the estimated discharge time series.

We used five different ways to represent the discharge data uncertainty in the time series used for the model calibration. Apart from using the time series realizations and the empirical pdf, we investigated the use of uniform and triangular distributions as well as the case when no discharge data uncertainty is considered (i.e. the typical model calibration approach here used as a benchmark). The lower and upper bounds of the uniform and triangular distributions were derived from the 0.05^{th} and 99.95^{th} percentiles of the empirical distribution to represent ranges similar to typical fuzzy estimates (e.g. Blazkova and Beven 2009). The optimal realization from the MCMC rating-curve estimation (i.e. the maximum *a posteriori* probability estimate) was used to derive the best-estimate discharge for the triangular distribution and the no-data-uncertainty case.

2.4. Data for hydrological modelling

The following meteorological data were available for each catchment: time series of hourly precipitation sums, time series of hourly mean temperature, long term means of daily temperature, and seasonally varying daily estimates of potential evaporation. All of these meteorological variables were computed as areal mean values for each catchment using the Thiessen-polygon method. The precipitation time series were checked for water balance consistency against the discharge data, and the precipitation data were corrected accordingly to close the water balance in the Hondrich catchment. For the

Zofingen station, we observed hydropeaking (from short-term hydropower regulation), evident from numerous artificial peaks in the low flow range. Periods affected by such hydropeaking were removed from the calibration and validation data.

3. Methods

We organized the study method according to the three specific objectives (Section 1). We addressed these research questions as shown in the flow chart in Fig. 2. We used the five different representations of discharge data uncertainty (Section 2.3) and two objective functions (Section 3.2) to evaluate different ways to incorporate discharge data uncertainty in the calibration of a hydrological model (the HBV model as described in Section 3.1). For this purpose, we developed a new objective function that incorporates information about different distributions of discharge data uncertainty (i.e., three of the five uncertainty representations). As the new objective function cannot be used without discharge data uncertainty, we used a standard multi-objective calibration for the other two uncertainty representations where data uncertainty is not incorporated in the objective function. Finally, we evaluated the simulated results using a set of hydrological signatures and scaled model residuals analysis, which also accounted for the uncertainty in the observed discharge data (Section 3.3).

3.1. *Model description*

The HBV model (Bergström 1976, Lindström *et al.* 1997) is a typical bucket-type, semi-distributed hydrological model. Several routines are used to represent the catchment functioning, i.e. the transfer from precipitation to catchment discharge. In the snow routine, a degree-day approach is used to simulate snow accumulation and melt. Snow melt and rainfall enter the soil routine where the groundwater recharge is computed based on the antecedent soil water storage, actual evaporation is estimated based on the relative soil storage filling, and a simple water balance accounting routine is used to update the soil water storage. Groundwater recharge enters the groundwater routine, where two reservoirs are used to represent groundwater storage and its control on runoff, which is computed by three linear outflows and parameters defining thresholds. The simulated runoff is then modified using a simple routing routine, which attenuates and delays discharge peaks. To allow for variable snow dynamics within a catchment, the catchment is subdivided into elevation zones (typically bands of 100–200 m) for the computations

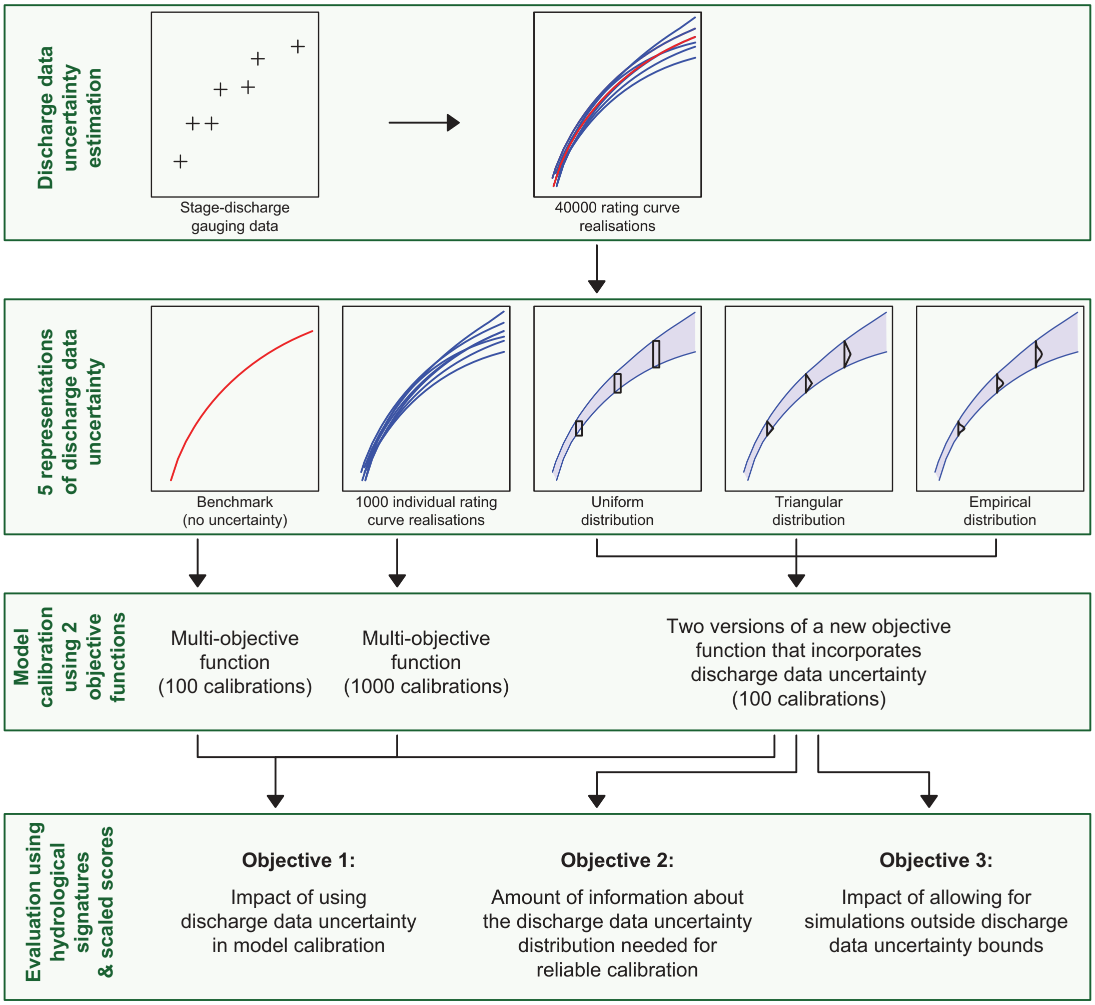
Discharge data uncertainty estimation
Stage-discharge gauging data
40000 rating curve realisations
5 representations of discharge data uncertainty
Benchmark (no uncertainty)
1000 individual rating curve realisations
Uniform distribution
Triangular distribution
Empirical distribution
Model calibration using 2 objective functions
Multi-objective function (100 calibrations)
Multi-objective function (1000 calibrations)
Two versions of a new objective function that incorporates discharge data uncertainty (100 calibrations)
Evaluation using hydrological signatures & scaled scores
Objective 1:
Impact of using discharge data uncertainty in model calibration
Objective 2:
Amount of information about the discharge data uncertainty distribution needed for reliable calibration
Objective 3:
Impact of allowing for simulations outside discharge data uncertainty bounds

Figure 2. Flow chart of the method used in the study for each objective.

in the snow and soil routines. In addition, a glacier routine can be activated for catchments where glaciers contribute significantly to catchment hydrology (Seibert *et al.* 2018).

The HBV model (without glacier routine) has typically 10–15 model parameters, which are usually set through calibration or regionalization. As in most other bucket-type models, the parameters take effective values at the catchment scale and lump together different processes. Any direct measurements of the parameter values are, thus, not possible.

The HBV model exists in many different versions, and we used the version HBV-light (Seibert and Vis 2012) with 15 parameters for the catchments Wigger-Zofingen and Broye-Payerne and with 6 additional parameters for the Kander-Hondrich catchment, for which the model included the glacier routine (Seibert *et al.* 2018). Each catchment was subdivided into 100 m elevation zones.

While the HBV model is often applied using a daily time step, we here used an hourly time step, as this preserves more information about the hydrological behaviour around flow peaks in fast-responding catchments such as those used in this study (e.g. multiple flow peaks that would be averaged out in daily data). For some analyses, we compared the hourly model calibration results to daily-scale simulations to assess impacts of a more detailed precipitation input, but also a potentially greater impact of precipitation data errors on an hourly compared to daily scale. The model was calibrated using data for the period 2001–2008 and evaluated using data for 2009–2010. The calibration of the model was performed using a Genetic Algorithm and Powell optimization (GAP) approach (Seibert 2000) using the two objective functions and the five uncertainty representations (Section 3.2 and Fig. 2).

3.2. Model calibration with different discharge data uncertainty representations

The five different representations of discharge data uncertainty that we used in model calibration differed in the amount of information about the uncertainty characteristics that they include: from no information (benchmark) to little information (uniform and triangular distributions) to much information (empirical frequency distribution and individual rating curve realizations).

The three uncertainty representations, i.e. uniform, triangular and empirical distributions, required that we developed a new objective function to include distributional information about discharge uncertainty for each time step directly into the model calibration (Section 3.2.3). Note, however, that including the discharge data uncertainty directly into the calibration procedure in the new objective function makes this function dependent on the availability of the discharge data uncertainty estimates. Thus, this function cannot be used without any discharge data uncertainty estimates and other (standard) objective functions had to be used for the other two uncertainty representations, i.e. the individual rating curve realizations and the no uncertainty benchmark (Sections 3.2.1–3.2.2).

3.2.1. Independent realizations of discharge time series from rating curve realizations

With this uncertainty representation, 1,000 independent rating curve realizations were randomly sampled from the estimated 40,000 posterior parameter distributions (Section 2.3). For each sampled rating curve, the corresponding discharge time series was used to run a multi-objective model calibration, resulting in 1,000 model simulation realizations in total (i.e. one simulation for each model calibration). Note that information about the autocorrelation of the discharge uncertainty and the bias in low or high flows is implicitly accounted for when using multiple individual realizations in this approach. Similarly, parameter uncertainty was also implicitly considered by calibration to the 1,000 individual discharge realizations. For the model calibration we used a multi-objective function that combined the Kling-Gupta efficiency (R_{KGE}, Gupta *et al.* 2009), the efficiency for peak flows (R_{peak}, Seibert 2003), and the mean absolute relative error (i.e. the MARE efficiency, R_{MARE}, Dawson 2007). Both R_{KGE} and R_{peak} are sensitive to peak flows and R_{MARE} is sensitive to low flows. These metrics were weighted following Sikorska *et al.* (2018) but using R_{MARE} instead of the Nash-Sutcliffe efficiency calculated using the logarithm-transformed discharge values to avoid problems with zero discharge values. The resulting objective function was:

$$F_{\mathrm{obj}} = 0.3 \cdot R_{\mathrm{KGE}} + 0.2 \cdot R_{\mathrm{MARE}} + 0.5 \cdot R_{\mathrm{peak}} \tag{2}$$

The R_{KGE} value was computed as

$$R_{\mathrm{KGE}} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{3}$$

where r is the correlation, α is a measure of the relative variability in the simulated and observed values, and β is a bias. R_{MARE} and R_{peak} were computed as:

$$R_{\mathrm{MARE}} = 1 - \frac{1}{n}\sum \frac{|Q_{\mathrm{obs}} - Q_{\mathrm{sim}}|}{Q_{\mathrm{obs}}} \tag{4}$$

$$R_{\mathrm{peak}} = 1 - \frac{\sum \left(Q_{\mathrm{obs,peak}} - Q_{\mathrm{sim,peak}}\right)^2}{\sum \left(Q_{\mathrm{obs,peak}} - \overline{Q_{\mathrm{obs,peak}}}\right)^2} \tag{5}$$

where $\overline{Q_{\mathrm{obs,peak}}}$ is the average observed peak discharge, and $Q_{\mathrm{obs,peak}}$ and $Q_{\mathrm{sim,peak}}$ are the peaks of the observed and simulated discharges ($Q_{\mathrm{sim,peak}}$ corresponds to the highest value within a ±3 day window of each observed peak).

3.2.2. "No uncertainty" benchmark: optimization with the best-estimate discharge

As a benchmark for comparing to the calibration using the 1000 rating curve realizations, we used the same multi-objective calibration as in Section 3.2.1, but only with the best-estimate discharge from the rating curve estimation, i.e. without considering the uncertainty in the discharge time series. This corresponds to the standard calibration approach used in hydrology when the model is calibrated only against the best-estimate discharge time series and its uncertainty is neglected. We used 100 individual calibration trials to account for some parameter uncertainty without being too demanding computationally considering the total number of calibrations in the study.

3.2.3. New objective function for discharge time series with uncertainty distributions

Discharge uncertainty information on time series data have been used before in model calibration, e.g. when using triangular distributions in GLUE limit of acceptability approaches and accepting simulations as behavioural if they are always inside the limits, or inside for a certain fraction of time (Liu *et al.* 2009). An alternative approach is to take an average of the deviations relative to the observed distribution at individual time steps. This time-step based method was previously used by Krueger *et al.* (2010) who use a uniform observed distribution and average the performance over different hydrograph aspects (non-driven quick and slow flow, and driven quick flow), and by McMillan *et al.* (2010) who accounted for autocorrelation by including the effective sample size in the calculation of the aggregated likelihood. Some weighting of the performance at each time step is necessary as taking the equally-weighted average of the deviations at individual time steps would bias the simulations to low flows, which occur more frequently in discharge time series. To reduce this bias, we investigated using a flow-weighted average that gives relatively higher weights to deviations at high flows, which occur more seldom. The objective function was defined for all three discharge data uncertainty distributions in the following way:

$$F_{\text{obj}} = \frac{\sum_{t=1}^{T} w(t) \cdot \left(Q_{\text{obs}}(t)\right)^2}{\sum_{t=1}^{T} \left(Q_{\text{obs}}(t)\right)^2} \tag{6}$$

where t is the time step, T is the total number of time steps, $w(t)$ is a weight at time step t assigned depending on the position of the simulated discharge value in relation to the observed data uncertainty distribution, and $Q_{\text{obs}}(t)$ is an observed discharge at time step t. The measure has a maximum value of one (representing a perfect match to the mode of the uncertainty distribution at each time step) and a minimum value of zero (representing a model simulation that is consistently outside the uncertainty distribution). The assigning of the weight $w(t)$ at each time step varies depending on the type of discharge data uncertainty representation that is used and is explained below for the three distributions we used (uniform, triangular, empirical). Obviously, this objective function cannot be used without discharge data uncertainty because the weights $w(t)$ cannot be computed if the uncertainty bounds for the discharge time series are not given.

To address the third study objective (Section 1), we investigated a modification of the objective function in Equation (6) that allows for simulations outside the defined discharge uncertainty bounds by penalising the model more, the further away from the bounds the model simulations are (see below). This modification is similar to the approach of Krueger *et al.* (2010) and was done as an attempt to allow for, e.g. model input errors, which can cause even an acceptable model simulation to be outside the uncertainty bounds. As for the benchmark simulations, we used 100 individual calibration trials to account for parameter uncertainty.

3.2.3.1. Upper and lower bounds (uniform distribution).

This uncertainty representation uses the simplest type of uncertainty information, where only lower and upper discharge uncertainty bounds are available and are used with a uniform distribution in-between the bounds. For each time step, each simulation that is outside the bounds was assigned a weight $w(t)$ of zero, whereas a value of one was assigned when the simulation was within the bounds:

$$w(t) = \begin{cases} 1 & \text{if } Q_L(t) \le Q_{\text{sim}}(t) \le Q_U(t) \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $Q_{\text{sim}}(t)$ is the discharge simulated at the time step t, and $Q_L(t)$ and $Q_U(t)$ are, respectively, the lower and the upper bounds at time step t.

This approach could be used with most types of discharge uncertainty estimation techniques, including fuzzy rating-curve approaches (Section 1). We used the 0.05th and 99.95th percentiles of the discharge data uncertainty distribution as the lower and upper bounds of discharge uncertainty (see Section 2.3).

With Equation (7), a weight of zero is assigned to all discharge simulations lying outside the bounds regardless of the distance from the upper/lower bound. The weights were assigned independently at each time step. To allow for simulations outside the bounds (that may result from, for instance, precipitation data errors), we investigated a modification of the above approach. For all simulations lying outside the bounds, a linear extrapolation was performed based on the interval between the best-estimate discharge and the upper/lower bound. In this way, simulations outside the bounds receive a negative weight, and the weights become more negative the further the simulations are from the bounds.

$$w(t) = \begin{cases} \frac{Q_{\text{sim}}(t) - Q_L(t)}{Q_B(t) - Q_L(t)} & \text{if } Q_{\text{sim}}(t) < Q_L(t) \\ 1 & \text{if } Q_L(t) \le Q_{\text{sim}}(t) \le Q_U(t) \\ \frac{Q_{\text{sim}}(t) - Q_U(t)}{Q_B(t) - Q_U(t)} & \text{if } Q_{\text{sim}}(t) > Q_U(t) \end{cases} \tag{8}$$

where $Q_B(t)$ is the best-estimate (i.e. optimal from the MCMC) observed discharge value at time step t as estimated from the rating curve (Section 2.3).

3.2.3.2. Triangular distribution.

For this uncertainty representation, the same upper and lower bounds as for the uniform distribution were used together with the best-estimate discharge estimate to define a triangular distribution. Triangular distributions have been commonly used in fuzzy discharge data uncertainty estimation (e.g. Westerberg *et al.* 2011a). Here the best-estimate discharge gets a weight of 1, whereas a weight of 0 is assigned to all simulations lying outside the bounds. Linear interpolation was applied to all values between the lower bound, the optimum, and the upper bound:

$$w(t) = \begin{cases} \frac{Q_{\text{sim}}(t) - Q_L(t)}{Q_B(t) - Q_L(t)} & \text{if } Q_L(t) \le Q_{\text{sim}}(t) \le Q_B(t) \\ \frac{Q_{\text{sim}}(t) - Q_U(t)}{Q_B(t) - Q_U(t)} & \text{if } Q_B(t) < Q_{\text{sim}}(t) \le Q_U(t) \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

As for the uniform distribution, an extrapolation outside the bounds could also be performed to allow evaluation of simulations that are outside the bounds with non-zero (i.e. negative) values, and the weights are then assigned as:

$$w(t) = \begin{cases} \frac{Q_{\text{sim}}(t) - Q_L(t)}{Q_B(t) - Q_L(t)} & \text{if } Q_{\text{sim}}(t) \leq Q_B(t) \\ \frac{Q_{\text{sim}}(t) - Q_U(t)}{Q_B(t) - Q_U(t)} & \text{if } Q_{\text{sim}}(t) > Q_B(t) \end{cases} \tag{10}$$

3.2.3.3. *Empirical frequency distribution.*

The empirical pdf of the discharge uncertainty distribution for each time step was represented by 100 equally distributed points (Q_i) in-between the minimum and maximum discharge values and their corresponding frequency values (W_i). Values lying outside the minimum and the maximum values were assigned a value of zero. This uncertainty representation allows incorporating information about uncertainty distributions that change shape (skew, uni- or multimodality, heavy-tailed, etc.) across the flow range (e.g. Le Coz *et al.* 2014). The weights are assigned as:

$$w(t) = \begin{cases} W_1(t) & \text{if } Q_1(t) \leq Q_{\text{sim}}(t) < Q_2(t) \\ W_2(t) & \text{if } Q_2(t) \leq Q_{\text{sim}}(t) < Q_3(t) \\ \cdots & \cdots \\ W_{99}(t) & \text{if } Q_{99}(t) \leq Q_{\text{sim}}(t) < Q_{100}(t) \\ W_{100}(t) & \text{if } Q_{\text{sim}}(t) = Q_{100}(t) \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

As for both previous approaches, an extrapolation outside the bounds is possible, and the weights are assigned in the following way:

$$w(t) = \begin{cases} \frac{Q_{\text{sim}}(t) - Q_1(t)}{Q_B(t) - Q_1(t)} & \text{if } Q_{\text{sim}}(t) < Q_1(t) \\ W_1(t) & \text{if } Q_1(t) \leq Q_{\text{sim}}(t) < Q_2(t) \\ W_2(t) & \text{if } Q_2(t) \leq Q_{\text{sim}}(t) < Q_3(t) \\ \cdots & \cdots \\ W_{99}(t) & \text{if } Q_{99}(t) \leq Q_{\text{sim}}(t) < Q_{100}(t) \\ W_{100}(t) & \text{if } Q_{\text{sim}}(t) = Q_{100}(t) \\ \frac{Q_{sim}(t) - Q_{100}(t)}{Q_B(t) - Q_{100}(t)} & \text{if } Q_{\text{sim}}(t) > Q_{100}(t) \end{cases} \tag{12}$$

3.3. Evaluation of the different model calibration methods

The reliability of the simulations from the different model calibration strategies was evaluated for the calibration and evaluation period using two main types of analysis that took account of the observed discharge data uncertainty. First, we investigated how well the model simulations reproduced observed (uncertain) signature values, and secondly, how the model residuals, when scaled to the observed uncertainty intervals, behaved for different aspects of the hydrograph.

3.3.1. Reproduction of time series and signatures

We chose a set of 16 signatures that defined key aspects of the catchment behaviour: flow distribution, event frequency and duration, and flow dynamics (Table 2). These types of signatures have been used in many previous studies of, e.g. flow variability, model calibration and regionalization (Jowett and Duncan 1990, Yadav *et al.* 2007, Euser *et al.* 2013, Vigiak *et al.* 2018). To reduce impacts of data uncertainty related to the signature design (McMillan *et al.* 2017b), the event frequency and duration signatures were defined using a flow percentile threshold instead of a multiplier of median flow (Westerberg and McMillan 2015). The signature values were calculated for each of the 1,000 observed discharge data time series estimated from the 1,000 rating curve realizations and for each model simulation. The distributions of the 1,000 observed signature values were then compared with those from the model simulations for each calibration (Fig. 2).

3.3.2. Analysis of scaled model residuals

We assessed the model residuals quantitatively by scaling the residual values to the observed discharge uncertainty intervals for each time step and by analysing their characteristics for different parts of the hydrograph. Analysis of scaled residuals, also called scaled scores (Liu *et al.* 2009, Westerberg *et al.* 2011b), allows the performance of individual model realizations to be analysed in relation to the observed discharge data uncertainty. This enables an analysis not only of the position of the simulated values within the observed uncertainty distributions (such as when using a rank histogram (McMillan *et al.* 2010) or a predictive quantile-quantile plot (Thyer *et al.* 2009), but also their positive respective negative distance when the simulations are outside the observed upper respective lower bounds. Deviations outside the bounds are important to consider in model evaluation since even a perfect model can deviate from the observed data uncertainty distribution because of other errors than model structure, such as errors in the input data.

We followed the method of Westerberg and Birkel (2015) for assigning and analysing the scores. This method uses the best-estimate discharge and the upper and lower bounds

Table 2. Signatures used for evaluation of model performance (based on Westerberg *et al.* 2016). For evaluating the daily simulations, the signatures were calculated in a corresponding way using daily data.

Signature type	Name	Description	Unit
Flow distribution	Flow percentiles ($Q_{0.01}$, $Q_{0.1}$, Q_1, Q_5, Q_{10}, Q_{20}, Q_{50}, Q_{80}, Q_{90}, Q_{99})	Low and high flow exceedance percentiles from the flow duration curve (FDC).	mm h^{-1}
	Mean flow (Q_{MEAN})	Average flow in the analysis period	mm h^{-1}
Event characteristics	High flow event frequency (Q_{HF})	Average number of hourly high-flow events per year, with a threshold equal to the 5th exceedance percentile	$year^{-1}$
	High flow event duration (Q_{HD})	Average duration of hourly flow events higher than a threshold equal to the 5th exceedance percentile	h
Flow dynamics	Base-flow index (Q_{BFI})	Contribution of base flow to total streamflow, calculated from daily flows using the Flood Estimation Handbook method (Gustard *et al.* 1992)	-
	Overall flow variability (Q_{CV})	Coefficient of variation in streamflow, i.e. standard deviation divided by mean flow (e.g. Jowett and Duncan 1990)	-
	Flow autocorrelation (Q_{AC})	Autocorrelation for 1 day (24 h). E.g. used by Winsemius *et al.* (2009)	-

(Section 2.3) to assign the scores: i.e. a value of 0 for the best estimate, and −1 and 1 for the lower and upper bound respectively (Fig. 3). Values lying inside these borders were linearly interpolated, whereas those outside were extrapolated linearly. For example, a simulated value that is two times the magnitude of the upper bound will, therefore, get a score of +2, and a value that is two times the lower bound will receive a score of −2. These scores were then analysed for six different aspects of the hydrograph: large and small peaks, base flows, rising and falling limbs, and troughs. The hydrograph aspects were defined for both hourly and daily time series following Westerberg and Birkel (2015), with large peaks defined as flows >0.8 mm/h (>12 mm/d), base flows for flows <0.27 mm/h (<1.5 mm/d), and the time-window parameter for the definition of the other aspects set to 3 hours (1 day).

4. Results

4.1. Rating curve and discharge uncertainty results

The estimated rating curve and discharge uncertainty distributions differed in-between the three stations (Fig. 4). Zofingen had a wide uncertainty distribution at high flows because of the scatter in the high-flow gauging data and the extrapolation of the highest 1.5 m of the rating curve. Hondrich had a more centred distribution but with a wide 5–95% interval, likely as a result of scatter in the gaugings in the mid flow range. At Payerne, the uppermost 2 m of the rating curve were constrained only by one gauging, which led to a constrained but heavy-tailed distribution: the upper (99.95%) and lower (0.05%) bounds at high flows had the widest interval of all stations. The half-widths of the 5–95% uncertainty bounds for hourly low, medium and high flow (i.e. Q_{90}, Q_{20} and $Q_{0.1}$) were ±30%, ±12% and ±8.6% respectively at Payerne, ±23%, ±10%, and ±20% at Hondrich, and ±22%, ±10%, and ±15% at Zofingen. Note that at Zofingen the best-estimate realization from the MCMC at extrapolated high flows was not in the centre of the discharge distribution. This was likely because of conflicting information in the middle to high flow gaugings (i.e. scatter in the gauging data) in combination with the lack of gaugings at extremely high flows.

4.2. What is the impact of including/excluding discharge data uncertainty in model calibration on the resulting simulations?

To address this research question, we first investigated the impact of considering uncertainties in model calibration by comparing the 100 multi-objective calibrations using the best-estimate data (i.e. the no uncertainty benchmark) to the 1,000 rating-curve realizations calibrated with the same method. We then compared these results to the simulations with the new objective function (see objective 1 in Fig. 2).

4.2.1. Use of the multi-objective function

The simulated hydrological signatures for these two calibrations using the multi-objective function had similar uncertainty distributions, but there was generally a wider uncertainty for the rating-curve realizations (results not shown). This is not surprising given that the rating-curve realization simulations allow for the observed discharge data uncertainty. The difference in the number of simulated realizations between these two methods did not influence the results: we checked that the resulting signature distributions and the other results for the rating-curve realizations were the same, but less smooth, when using 100 instead of all 1,000 rating-curve realizations.

We then analysed the model residuals scaled to the observed discharge data uncertainty interval (i.e. the scaled scores, Section 3.3.2) for different hydrograph aspects for simulations

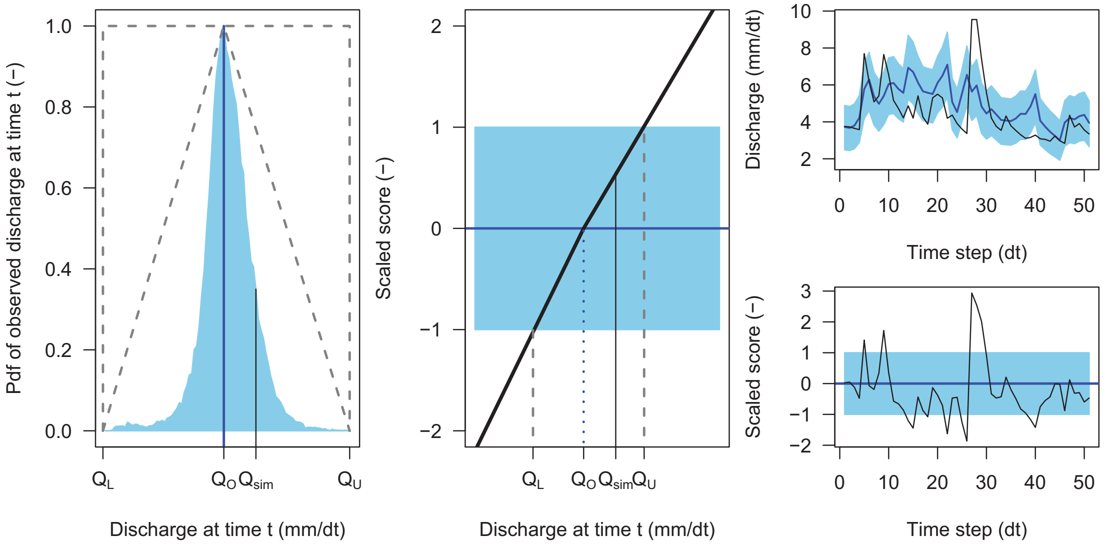

Figure 3. Left panel, discharge uncertainty representations: empirical, triangular and uniform distributions for an observed discharge at time step *t*. Middle panel, calculation of scaled scores for a simulated discharge value Q_{sim} at time step *t* using a linear rescaling in the range between the lower bound (score of −1), best-estimate discharge (score of 0) and upper bound (score of 1) and linear extrapolation outside the bounds. Top right panel, time series of observed and simulated discharge, and bottom right panel, time series of corresponding scaled scores. Q_{sim} (thin black line) is the simulated discharge at time *t*, Q_O (dark blue line) is the best-estimate discharge, Q_L and Q_U are the lower and upper bounds, and *dt* is the time step of the simulations (hourly or daily).

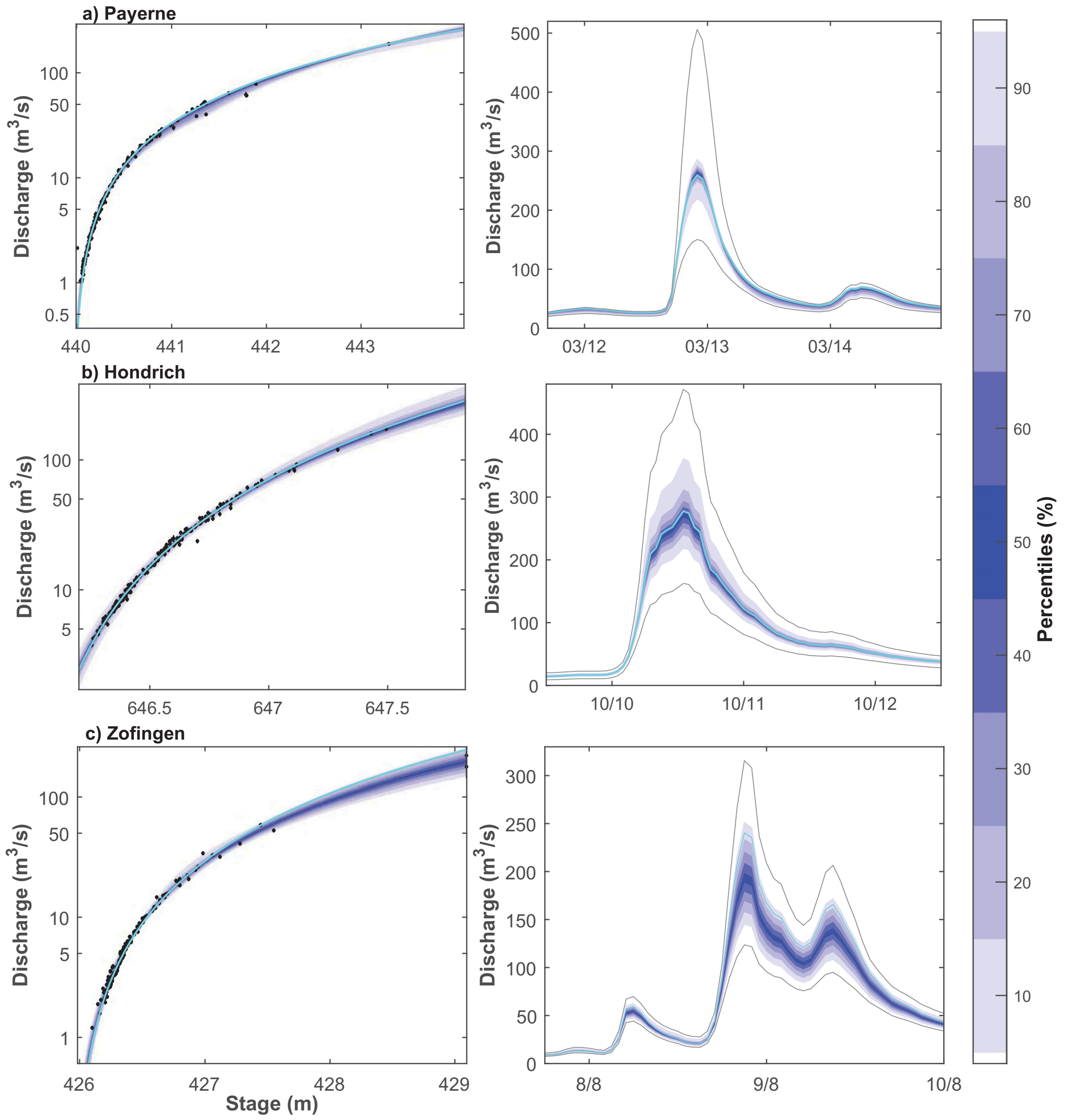

Figure 4. Left column: rating curve uncertainty distribution in log scale for the three study catchments (blue bands), best-estimate rating curve in light blue and gaugings (black dots with black lines representing discharge measurement uncertainty, note that the lines are barely visible as uncertainty is low). Right column: corresponding uncertainty distribution on the hourly discharge time series for the highest flow event in the study period (blue bands), time series for best-estimate rating curve in light blue, and for upper and lower uncertainty bounds in grey. In both plots, the blue bands show the uncertainty distribution of discharge from all the 40,000 estimated rating curves at each stage value or time step, from the 5th to 95th percentiles. Where the blue bands are narrower, there is a higher probability density for the discharge compared to where they are wider.

performed at both hourly and daily time steps. This analysis showed that at hourly time steps the scaled score distributions for the benchmark simulations and the rating-curve realizations were mostly similar for Hondrich, where the model performed best (Fig. 5, top panel). The main difference occurred for large peak flows, where calibration to the rating curve realizations resulted in more underestimation (i.e., scaled scores <-1) compared to the benchmark simulations. At Zofingen both simulations had larger deviations for all time steps compared to at Hondrich, and the scaled score distributions were less similar between the two calibrations for rising limbs and troughs (Fig. 5, bottom). At Payerne, the model had the worst performance for all hydrograph aspects apart for from base flows, for which the model performed best among all three catchments (Figure S1 in supplementary material). For Payerne, the distributions were also similar when including or excluding discharge uncertainty in the multi-objective calibration, but using the rating-curve realizations led to slightly more underestimation of peak flows, troughs, and rising limbs. For the evaluation period, the results were generally similar to the calibration period for all three catchments, but with more underestimation at Payerne when using the rating curve realizations. For the daily scale simulations in both calibration and evaluation periods, there was more underestimation at Hondrich for all flow aspects and for base flows at Zofingen, while at Payerne no apparent differences could be identified. In summary, the results showed that using multiple realizations of the time series to represent discharge data uncertainty did not give better

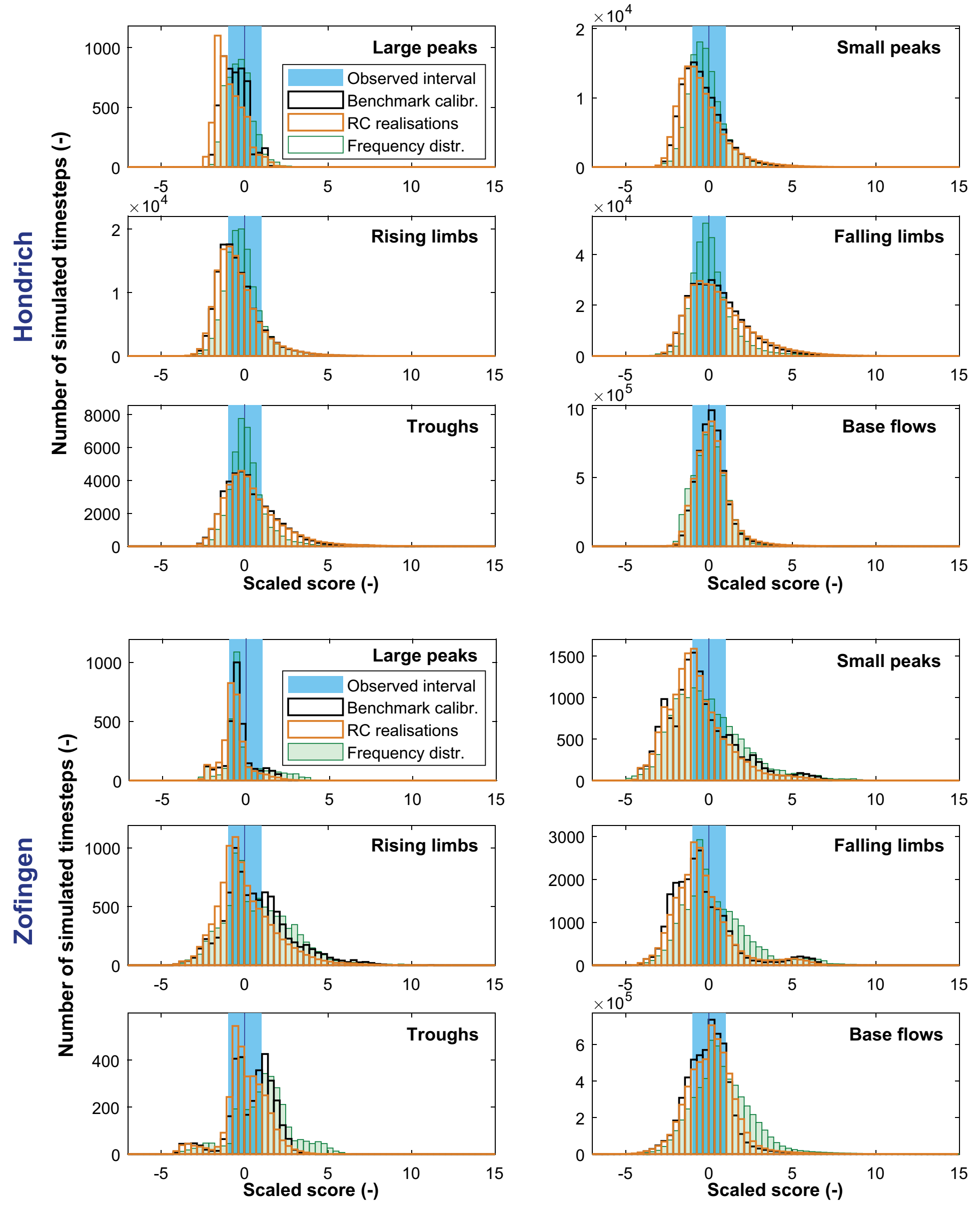

Figure 5. Histograms of scaled scores for six different aspects of the hydrograph for the Zofingen and Hondrich catchments for the benchmark calibration, the rating curve realizations and the calibration using the empirical frequency distribution. A scaled score within ±1 means that the simulation is inside the observed discharge uncertainty, and a scaled score of 0 represents an exact match to the best-estimate discharge.

results than the no-uncertainty benchmark when using the multi-objective function. In the next section we therefore investigate whether including information about the discharge data uncertainty directly in the objective function is a better approach.

4.2.2. *Use of the new objective function*

We compared the simulated results from the multi-objective calibrations (with and without observed uncertainty) to the simulations using the new objective function (i.e. those based on Equation (6)) that integrates information about the observed discharge uncertainty distribution. This new objective function requires a quantification of discharge data uncertainty, wherefore we compare these results to the multi-objective calibrations. The two multi-objective calibrations had worse results than the representations using the new objective function for all flow aspects apart from base flows at Hondrich (Fig. 5, top panel). While only results for the empirical frequency distribution are shown for clarity, the differences to the multi-objective calibrations were similar for the other methods. At Zofingen the results were more mixed, with more underestimation of small peaks and falling limbs and less overestimation of troughs for the multi-objective calibrations, but better results for base flows and no clear differences for rising limbs and large peaks (Fig. 5, bottom panel). At Payerne, there was more underestimation of large peaks for the multi-objective calibrations, similar base flow performance and equally poor results for the other flow aspects for all calibrations. The results were in general similar for the evaluation period in all catchments with more underestimation for the multi-objective calibrations for most flow aspects, but with slightly better results for base flows. Overall, the objective functions that integrate information about the observed discharge data uncertainty distribution led to better simulation results than both approaches using the multi-objective function, i.e., the approach with the rating curve realizations and the no-uncertainty benchmark approach. Including discharge data uncertainty in the model calibration therefore gave better results than not including discharge data uncertainty, but only when the information about the discharge data uncertainty was included directly into the (new) objective function.

4.3. How much information about the discharge data uncertainty distribution is needed in the objective function to obtain reliable model simulations?

We compared model calibration using the three uncertainty representations with different observed discharge distributions in the objective function, i.e. using uniform, triangular, or empirical frequency distributions (see Fig. 2, objective 2). We found that there was a small difference for low to medium flows and for average flows between the different simulations in all catchments (Q_{90}–Q_{20} and Q_{MEAN}, Fig. 6). The simulated results showed the same differences for the calibration and evaluation periods. For intermediate and high flows, the differences between the simulated uncertainty distributions were larger, but they were still overlapping (Q_5–$Q_{0.01}$). For the Q_{BFI}, Q_{CV}, Q_{AC}, Q_{HD}, and Q_{HF} signatures (Table 2), the uncertainty distributions overlapped to a large extent for all the three calibration methods. An analysis of the scaled residuals showed that the simulated distributions were less centred on zero within the observed uncertainty bounds when using the uniform distribution compared to the triangular or empirical frequency distributions. This occurs because the uniform distribution gives equal weights within the uncertainty bounds. This finding suggests that the uniform distribution may be less useful when there is some confidence in the observed discharge uncertainty distributions or best-estimate values. In summary, the differences between the three types of calibrations that incorporate the observed discharge distributions in the objective function were small. However, using the empirical or triangular pdf was preferable to the uniform distribution, and therefore having information about at least the discharge data uncertainty bounds and the best-estimate discharge is preferable for model calibration.

4.4. What is the impact of allowing for simulations outside the discharge data uncertainty bounds?

With the first type of objective function (Equations (7), (9) and (11)), the simulated discharge had a zero weight at all time steps during which the simulated discharge was outside the observed uncertainty bounds. However, model simulations could be outside the observed uncertainty bounds because of an input data error such as precipitation data that is wrongly measured or not representative of the whole catchment. To account for such errors, we tested the modified objective functions (Equations (8), (10) and (12)) that, instead of assigning a zero weight, assign increasingly negative weights the further the simulated discharge is from the observed uncertainty bounds (i.e. extrapolating outside the uncertainty bounds).

We found that for all three catchments, allowing for simulations outside the bounds by assigning non-zero weights resulted in less variability among the 100 simulated realizations (Fig. 7). The largest impact on the simulated results was observed for Payerne. Here the rising limbs, falling limbs, and troughs were much better represented with this modified version of the objective function, but at the same time the model consistently underestimated the peak flows. At Payerne, there was only one high flow gauging available (Fig. 4) to constrain the upper two meters of the rating curve, and the high flows are therefore much more uncertain than intermediate or low flows. However, the observed discharge uncertainty bounds were very wide at high flows, which means that it is improbable that the true discharge would not be within these bounds. The poor model performance at high flows, with considerable underestimation of peak flows even when allowing for simulations outside the bounds (Figs. 7 and 8), suggests that either the model structure is not well adapted to reproduce the fast precipitation–runoff response in this catchment, or there may be some inconsistencies in the observed data (Beven and Westerberg 2011). The ambiguous delineation of the Payerne catchment from the Digital Elevation Model points at possible groundwater flows towards Lake Geneva (that are, thus, not recorded at the gauge in Payerne), and a drinking water diversion may also contribute to data inconsistencies and consequently poor model reproduction of flow peak behaviour (Parriaux 1981, Bultot *et al.* 1994).

At Zofingen, the model simulations were in general lower when accounting for simulations outside the bounds. In particular, the high flows were more underestimated but still within or close to the uncertainty bounds. This also occurred at Hondrich, where the average deviations were more centred within the bounds for the simulations without extrapolation outside the bounds. In summary, allowing for deviations outside the uncertainty bounds led to smaller variability between the 100 simulated realizations but also to consistent underestimation of peak flows. This modified objective function therefore appears to be a less useful strategy than the original approach with assigning zero weights to all simulated values lying outside the uncertainty bounds. However, note that the way of assigning the weights, particularly to high flows, may also play a role here, and this is further discussed in Section 5.1.

5. Discussion

5.1. Model calibration with uncertain discharge data

Incorporating information about the discharge uncertainty distribution at each time step directly into the objective function gave overall the best simulation results in our study. In contrast, using multiple rating curve realizations and calibrating the model once to each corresponding time series gave poorer results with underestimated high flows. In addition, this approach did not lead to better results than the benchmark (no data uncertainty) calibration and was much more computationally demanding (a factor of 10 compared to the calibrations with the new objective function and the uniform,

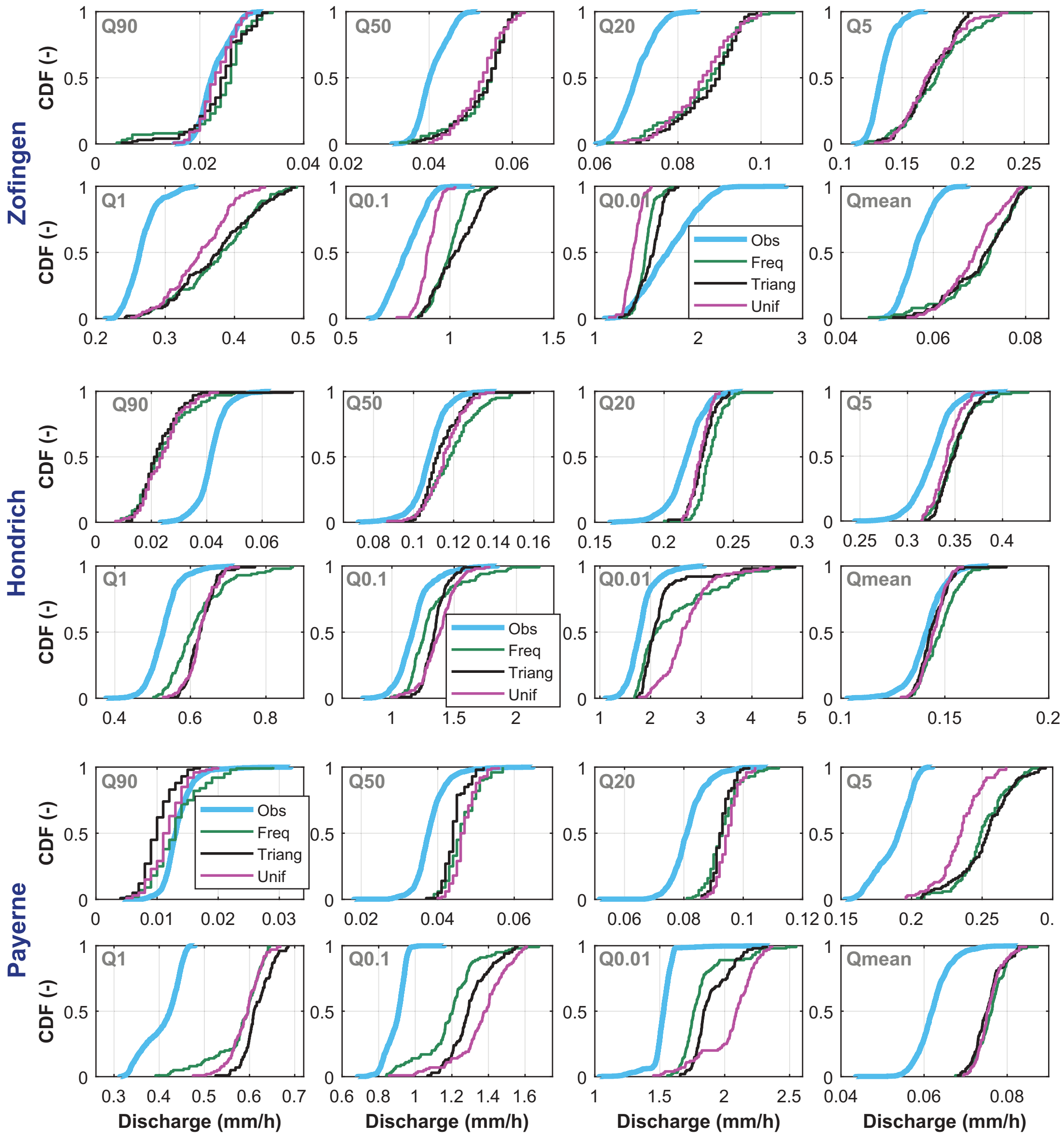

Figure 6. Observed and simulated signature uncertainty distributions for the three catchments (top two rows Zofingen, middle two rows Hondrich and bottom two rows Payerne) for three calibrations including observed discharge uncertainty as an empirical frequency distribution (Freq), a triangular distribution (Triang), and a uniform distribution (Unif).

triangular and empirical distributions, which were roughly the same). The number of discharge time series realizations could be reduced from 1,000 to lower the computational demand, but it would mean that some information about the discharge time series uncertainty would be lost. It is worth mentioning that the no data uncertainty approach (benchmark) in fact accounts for some parametric uncertainty of the hydrological model and by this it may partly compensate for the discharge data uncertainty that is not explicitly considered in this representation. Similarly, the representation using the multiple (1,000) realizations of the rating curve and doing 1,000 model calibrations implicitly accounts for parameter uncertainty of the hydrological model. Such effects can be more comprehensively studied using more advanced Bayesian likelihood approaches (see Section 5.2).

When using the new objective function that incorporates the observed uncertainty distribution, the information about the discharge uncertainty is considered directly in the evaluation between the simulated and observed values at each time step. In contrast, when using multiple realizations, the discharge uncertainty is considered indirectly in the calibration, but on the other hand, the information about the autocorrelation of the discharge errors is preserved (as the entire realization is used). The latter could be an advantage in some applications targeting, e.g., flow recession behaviour, but would not work with the VPM rating curve estimation method we used here if there are major shifts in the stage–discharge relationship at the discharge gauging site. This is because the rating curve uncertainty needs to be estimated separately before and after the shift and there is no link between the individual rating curve realizations for the two periods. In such cases, rating-curve uncertainty estimation methods that model the temporal evolution of the stage–discharge relationship and the uncertainty explicitly can be used instead (see methods overview in Kiang *et al.* 2018). Alternatively, the rating curve uncertainty can be estimated separately before and after the rating shift. In both these cases

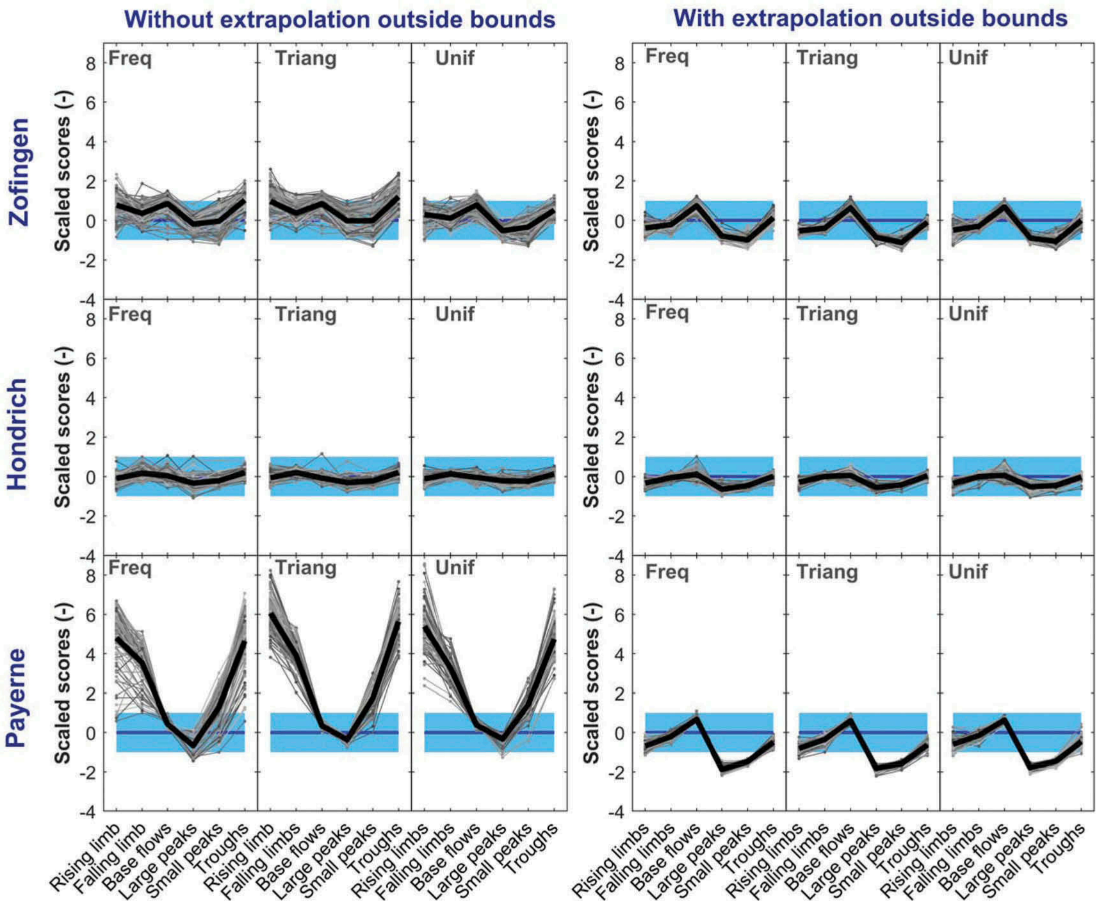

Figure 7. Comparison of model calibration without (left) and with (right) accounting for deviations outside the uncertainty distribution bounds for the three catchments using the objective function that incorporates information about the observed discharge uncertainty distribution (Freq: empirical frequency distribution, Triang: triangular distribution, and Unif: uniform distribution). The graphs show the average scaled scores (residuals scaled to the observed uncertainty interval) for the six flow types for each of the 100 realizations (grey lines) and the average for all realizations (thick black line).

the new objective function that incorporates the discharge uncertainty distribution at each time step can be used.

The differences in results between using the uniform, triangular and empirical distributions were generally small, suggesting that any of these three approaches could be used. However, when evaluating the scaled model residuals, we could see that the simulations were more widely spread and less centred on the best-estimate discharge value when using the uniform distributions. This finding makes the uniform distribution less suitable than triangular and empirical distributions when, as in many situations, there is confidence about the best-estimate discharge or discharge distribution. Then empirical or triangular discharge distributions are recommended instead of a uniform distribution. If an empirical discharge distribution is available, it is advisable to use it instead of a triangular distribution as the former incorporates more information about the discharge uncertainty and therefore leads to a more appropriate weighting (particularly when the distribution is heavy-tailed or changes shape across the flow range).

The modification of the new objective function, which gave increasingly lower weights the further the simulations are from the distribution bounds (instead of assigning a zero weight), was a simple approach to allow for the effect of, e.g. input errors that can cause even a perfect model to be outside the uncertainty bounds. However, this modified objective function resulted in a consistent underestimation of peak flows and is therefore not recommended in its current form. This underestimation may have resulted from the squared flow weighting in the objective function, which may still be giving too little weight to the highest flows. This approach may therefore work better with a revised flow-weighting method, which could be tested in future studies, together with comparing to the approach of Krueger *et al.* (2010) of averaging the deviations for different flow aspects.

5.2. *Method considerations and limitations*

A major limitation of our study is that it was not possible to use the same objective function for all the five uncertainty representations, as the new objective function cannot be used without any information on discharge data uncertainty (i.e. the weights cannot be assigned). This means that when comparing the simulation results for the new objective function with those of the multi-objective calibrations, part of the difference in the results is due to the use of different objective functions and not to the inclusion or not of the discharge data uncertainty.

The new objective function, which we propose in this paper, can incorporate different distributions of discharge data uncertainty directly into the calibration of a hydrological model with discharge time series. Compared to full-scale Monte Carlo analyses using informal or formal Bayesian likelihood approaches to account for discharge data uncertainty (e.g.

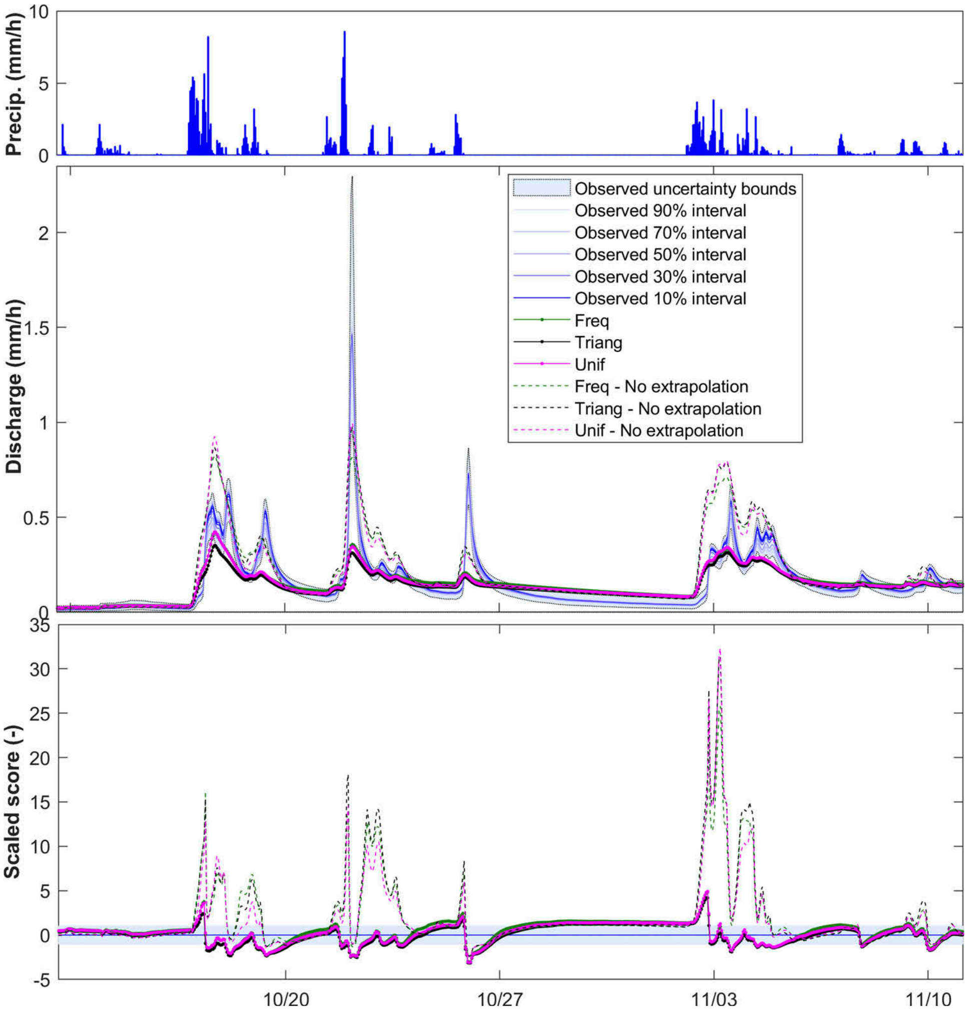

Figure 8. Precipitation time series (top), observed discharge (uncertainty bounds and distribution and averaged simulated discharge (middle), and average scaled scores (residuals scaled to the observed discharge uncertainty interval (bottom) for October–November 2002 in the calibration period in the Payerne catchment. The simulated discharge and corresponding scaled scores are shown for calibrations with and without accounting for extrapolation outside the uncertainty bounds, using the Freq (empirical frequency distribution), Triang (triangular distribution) and Unif (uniform distribution) to represent observed uncertainty in the objective function.

Krueger *et al.* 2010, Sikorska and Renard 2017), our approach in this study provides a less in-depth analysis of total uncertainty as it does not consider input or model structural errors and only partly parametric uncertainty. On the other hand, it provides a simpler and faster (much less computationally demanding) way to account for discharge data uncertainty in model calibration, as the objective function can be directly used with traditional optimization techniques without the need to sample the full parameter space, as in Bayesian approaches. This approach is primarily suitable for applications where it is not feasible to run a full-scale uncertainty analysis (e.g., due to computational issues).

When using limits of acceptability in GLUE applied directly to the time series as in Liu *et al.* (2009), this typically requires defining a threshold of acceptable time steps, for which the model simulation can be allowed outside the observed data distribution (i.e., the limits of acceptability), as simulations are generally not inside the limits at all time steps. However, this can lead to some systematic errors, as the time steps for which the simulations are outside the limits may be the hydrologically most interesting time steps (e.g., droughts or floods). In comparison, averaging the deviations for all time steps in a hydrologically meaningful way (Krueger *et al.* 2010, McMillan *et al.* 2010) can give a balanced simulation. We used a flow-weighted average of the deviations relative to the observed discharge data uncertainty interval to give deviations of different flow magnitudes a similar weight regardless of their occurrence. We found that this flow-weighting component played a large role for the calibration results. When developing the method, we tested a different flow-weighting that gave less weight to high flows, and this resulted in simulations that only had good performance for low flows. Using an inappropriate flow weighting could thus directly impact on what we learn about the hydrology of a modelled catchment. It is therefore important to check that the weighting is appropriate for a particular application, using approaches such as scaled model residuals to evaluate simulation performance for different flow conditions. The weighting function will be partly dependent on the flow regime (relative contribution of baseflows, length of peak flows, etc). We therefore recommend that different flow-weighting options are evaluated in future studies to investigate if there are better approaches than the squared weighting we used.

An alternative approach to the one we took here would be to translate the discharge data uncertainty to uncertainty in hydrological signatures (Westerberg and McMillan 2015), and then use these signatures for model calibration instead of discharge time series (Blazkova and Beven 2009, Schaefli 2016). Such an approach may be more robust to input errors at individual time steps (Westerberg *et al.* 2011b). Further comparisons between such signature-based approaches to account for discharge data uncertainty and time-series based approaches like those used here should be made. Such comparisons should investigate which approaches are most robust when considering input errors and other errors that affect hydrological model calibration apart from the discharge data uncertainty explored in this study.

6. Conclusions and recommendations

We investigated the role of discharge data uncertainty in hydrological model calibration and evaluation. Based on our findings from the three Swiss catchments, we conclude with the following recommendations for including discharge data uncertainty in model calibration (noting that the choice of the method will also be dependent on the resources, purpose and data availability in any study):

(1) Including information about the discharge data uncertainty distribution directly in the objective function resulted in better discharge simulations than using multiple realizations of discharge time series and optimizing the model to each realization.
(2) When sufficient discharge data uncertainty information is available, using triangular or empirical distributions is better than using uniform distributions, as it leads to a more appropriate weighting in the objective function, and using empirical distributions is in turn better than triangular distributions.
(3) Evaluating model simulations while taking the observed data uncertainty into account is important to interpret model results correctly when discharge data uncertainty is high. We recommend an approach using both hydrological signatures and scaled model deviations, which complement each other by assessing model performance for both overall flow statistics and at individual time steps.

Acknowledgements

The ScienceCloud computational and storage infrastructure provided by S3IT at the University of Zurich enabled the computation-intensive simulations. We thank the Swiss Federal Office for the Environment (FOEN) and MeteoSwiss for providing hydrological and meteorological data making this study possible. Special thanks to Hanspeter Hodel at FOEN for comprehensive personal communications on rating curves and site characteristics. We thank Tanja de Boer-Euser and an anonymous reviewer for constructive comments that helped to clarify and improve the paper. The new objective function that incorporates different discharge data distributions is implemented in the HBV-light model software, available at https://www.geo.uzh.ch/en/units/h2k/Services/HBV-Model.html

Disclosure statement

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding

This research has been conducted as a part of the EXAR project (Hazard information for extreme flood events on the rivers Aare and Rhine) funded by the Swiss Federal Office for the Environment (FOEN), the Swiss Federal Nuclear Safety Inspectorate (ENSI), the Swiss Federal Office of Energy (SFOE), the Swiss Federal Office for Civil Protection (FOCP) and the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss) under Project No. 15.0054.PJ/O503-1381. IW's time was partly funded by Svenska Forskningsrådet Formas (2018-02900, ERA-NET).

ORCID

Ida K. Westerberg http://orcid.org/0000-0002-9382-0782
Anna E. Sikorska-Senoner http://orcid.org/0000-0002-5273-1038
Daniel Viviroli http://orcid.org/0000-0002-1214-8657
Marc Vis http://orcid.org/0000-0002-5589-2611
Jan Seibert http://orcid.org/0000-0002-6314-2124

References

Bergström, S., 1976. *Development and application of a conceptual runoff model for Scandinavian catchments*. Norrköping, Sweden: SMHI.

Beven, K.J. and Westerberg, I.K., 2011. On red herrings and real herrings: disinformation and information in hydrological inference. *Hydrological Processes*, 25 (10), 1676–1680. doi:10.1002/hyp.v25.10

Blazkova, S. and Beven, K.J., 2009. A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic. *Water Resources Research*, 45, W00B16. doi:10.1029/2007WR006726

Bultot, F., *et al.*, 1994. Effects of climate-change on snow accumulation and melting in the Broye catchment (Switzerland). *Climatic Change*, 28 (4), 339–363. doi:10.1007/BF01104078

Chow, V.T., Maidment, D.R., and Mays, L.W., 1988. *Applied hydrology*. New York: McGraw-Hill Book Company.

Coxon, G., *et al.*, 2013. Diagnostic evaluation of multiple hypotheses of hydrological behaviour in a limits-of-acceptability framework for 24 UK catchments. *Hydrological Processes*, 28 (25), 6135–6150. doi:10.1002/hyp.10096

Coxon, G., *et al.*, 2015. A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations. *Water Resources Research*, 51 (7), 1944–7973. doi:10.1002/wrcr.v51.7

Dawson, C.W., Abrahart, R.J., and See, L.M, 2007. Hydrotest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. *Environmental Modelling & Software*, 22 (7), 1034–1052.

Di Baldassarre, G., Laio, F., and Montanari, A., 2012. Effect of observation errors on the uncertainty of design floods. *Physics and Chemistry of the Earth*, 42–44, 85–90. doi:10.1016/j.pce.2011.05.001

Euser, T., *et al.*, 2013. A framework to assess the realism of model structures using hydrological signatures. *Hydrology and Earth System Sciences*, 17, 1893–1912. doi:10.5194/hess-17-1893-2013

Gupta, H.V., *et al.*, 2009. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. *Journal of Hydrology*, 377 (1–2), 80–91. doi:10.1016/j.jhydrol.2009.08.003

Gustard, A., Bullock, A., and Dixon, J.M., 1992. *Low flow estimation in the United Kingdom*. Wallingford, UK: Institute of Hydrology, 108.

Herschy, R.W., 1970. The evaluation of errors at flow-measurement stations. *In*: ed. *Paper presented at International Symposium on Hydrometry*, Koblenz, 109–131.

Jalbert, J., Mathevet, T., and Favre, A.C., 2011. Temporal uncertainty estimation of discharges from rating curves using a variographic analysis. *Journal of Hydrology*, 397 (1–2), 83–92. doi:10.1016/j.jhydrol.2010.11.031

Jowett, I.G. and Duncan, M.J., 1990. Flow variability in New Zealand rivers and its relationship to in-stream habitat and biota. *New Zealand Journal of Marine and Freshwater Research*, 24 (3), 305–317. doi:10.1080/00288330.1990.9516427

Juston, J., Jansson, P.-E., and Gustafsson, D., 2014. Rating curve uncertainty and change detection in discharge time series: case study with 44-year historic data from the Nyangores River, Kenya. *Hydrological Processes*, 28, 2509–2523. doi:10.1002/hyp.9786

Kiang, J.E., *et al.*, 2018. A comparison of methods for streamflow uncertainty estimation. *Water Resources Research*, 54 (10), 7149–7176. doi:10.1029/2018WR022708

Krueger, T., *et al.*, 2010. Ensemble evaluation of hydrological model hypotheses. *Water Resources Research*, 46, W07516. doi:10.1029/2009WR007845

Le Coz, J., *et al.*, 2014. Combining hydraulic knowledge and uncertain gaugings in the estimation of hydrometric rating curves: a Bayesian approach. *Journal of Hydrology*, 509, 573–587. doi:10.1016/j.jhydrol.2013.11.016

Lindström, G., *et al.*, 1997. Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology*, 201 (1–4), 272–288. doi:10.1016/S0022-1694(97)00041-3

Liu, Y., *et al.*, 2009. Towards a limits of acceptability approach to the calibration of hydrological models: extending observation error. *Journal of Hydrology*, 367, 93–103. doi:10.1016/j.jhydrol.2009.01.016

Mansanarez, V., *et al.*, 2016. Bayesian analysis of stage-fall-discharge rating curves and their uncertainties. *Water Resources Research*, 52 (9), 7424–7443. doi:10.1002/2016WR018916

McMillan, H.K., *et al.*, 2010. Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions. *Hydrological Processes*, 24 (10), 1270–1284.

McMillan, H.K., *et al.*, 2017a. How uncertainty analysis of streamflow data can reduce costs and promote robust decisions in water management applications. *Water Resources Research*, 53, 5220–5228. doi:10.1002/2016WR020328

McMillan, H.K., Krueger, T., and Freer, J., 2012. Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality. *Hydrological Processes*, 26 (26), 4078–4111. doi:10.1002/hyp.v26.26

McMillan, H.K., Westerberg, I., and Branger, F., 2017b. Five guidelines for selecting hydrological signatures. *Hydrological Processes*, 31, 4757–4761. doi:10.1002/hyp.v31.26

McMillan, H.K. and Westerberg, I.K., 2015. Rating curve estimation under epistemic uncertainty. *Hydrological Processes*, 29 (7), 1873–1882. doi:10.1002/hyp.v29.7

Morlot, T., *et al.*, 2014. Dynamic rating curve assessment for hydrometric stations and computation of the associated uncertainties: quality and station management indicators. *Journal of Hydrology*, 517, 173–186. doi:10.1016/j.jhydrol.2014.05.007

Parriaux, A., 1981. *Contribution à l'étude des ressources en eau du bassin de la Broye*. Lausanne: EPFL.

Pelletier, P., 1988. Uncertainties in the single determination of river discharge: a literature review. *Canadian Journal of Civil Engineering*, 15 (5), 834–850. doi:10.1139/l88-109

Petersen-Overleir, A., Soot, A., and Reitan, T., 2009. Bayesian rating curve inference as a streamflow data quality assessment tool. *Water Resources Management*, 23 (9), 1835–1842. doi:10.1007/s11269-008-9354-5

Renard, B., *et al.*, 2010. Understanding predictive uncertainty in hydrologic modeling: the challenge of identifying input and structural errors. *Water Resources Research*, 46, W05521. doi:10.1029/2009WR008328

Schaefli, B., 2016. Snow hydrology signatures for model identification within a limits-of-acceptability approach. *Hydrological Processes*, 30 (22), 4019–4035. doi:10.1002/hyp.v30.22

Seibert, J., 2000. Multi-criteria calibration of a conceptual runoff model using a genetic algorithm. *Hydrology and Earth System Sciences*, 4 (2), 215–224. doi:10.5194/hess-4-215-2000

Seibert, J., 2003. Reliability of model predictions outside calibration conditions. *Nordic Hydrology*, 34 (5), 477–492.

Seibert, J., *et al.*, 2018. Technical note: representing glacier geometry changes in a semi-distributed hydrological model. *Hydrology and Earth System Sciences*, 22 (4), 2211–2224. doi:10.5194/hess-22-2211-2018

Seibert, J., and Vis, M., 2012. Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. *Hydrology and Earth System Sciences*, 16, 3315–3325. doi:910.5194/hess-16-3315-2012

Sikorska, A.E., *et al.*, 2013. Considering rating curve uncertainty in water level predictions. *Hydrology and Earth System Sciences*, 17 (11), 4415–4427. doi:10.5194/hess-17-4415-2013

Sikorska, A.E. and Renard, B., 2017. Calibrating a hydrological model in stage space to account for rating curve uncertainties: general framework and key challenges. *Advances in Water Resources*, 105, 51–66. doi:10.1016/j.advwatres.2017.04.011

Sikorska, A.E., Viviroli, D., and Seibert, J., 2018. Effective precipitation duration for runoff peaks based on catchment modelling. *Journal of Hydrology*, 556, 510–522. doi:10.1016/j.jhydrol.2017.11.028

Steinbakk, G.H., *et al.*, 2016. Propagation of rating curve uncertainty in design flood estimation. *Water Resources Research*, 52 (9), 6897–6915. doi:10.1002/2015WR018516

Thyer, M., *et al.*, 2009. Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis. *Water Resources Research*, 45, W00b14. doi:10.1029/2008WR006825

Tomkins, K., 2012. Uncertainty in streamflow rating curves: methods, controls and consequences. *Hydrological Processes*. doi:10.1002/hyp.9567

Vigiak, O., *et al.*, 2018. Uncertainty of modelled flow regime for flow-ecological assessment in Southern Europe. *Science of the Total Environment*, 615, 1028–1047. doi:10.1016/j.scitotenv.2017.09.295

Weingartner, R. and Aschwanden, H., 1992. Discharge Regime – the basis for the estimation of average flows. *Hydrological atlas of Switzerland*. Bern: Swiss Federal Office for the Environment.

Westerberg, I.K., *et al.*, 2011a. Stage-discharge uncertainty derived with a non-stationary rating curve in the Choluteca River, Honduras. *Hydrological Processes*, 25 (4), 603–613. doi:10.1002/hyp.7848

Westerberg, I.K., *et al.*, 2011b. Calibration of hydrological models using flow-duration curves. *Hydrology and Earth System Sciences*, 15 (7), 2205–2227. doi:10.5194/hess-15-2205-2011

Westerberg, I.K., *et al.*, 2016. Uncertainty in hydrological signatures for gauged and ungauged catchments. *Water Resources Research*, 52 (3), 1847–1865. doi:10.1002/2015WR017635

Westerberg, I.K. and Birkel, C., 2015. Observational uncertainties in hypothesis testing: investigating the hydrological functioning of a tropical catchment. *Hydrological Processes*, 29 (23), 4863–4879. doi:10.1002/hyp.v29.23

Westerberg, I.K. and McMillan, H.K., 2015. Uncertainty in hydrological signatures. *Hydrology and Earth System Sciences Discussions*, 12, 4233–4270. doi:10.5194/hessd-12-4233-2015

Wilby, R.L., *et al.*, 2017. The 'dirty dozen' of freshwater science: detecting then reconciling hydrological data biases and errors. *WIREs Water*, 4, e1209. doi:10.1002/wat2.1209

Winsemius, H.C., *et al.*, 2009. On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information. *Water Resources Research*, 45, W12422. doi:10.1029/2009WR007706

Wyder, D., 1998. *Handbuch der Pegelmessung. Hydrologische Mitteilungen*, 26. Bern: Swiss Federal Office for the Environment.

Yadav, M., Wagener, T., and Gupta, H., 2007. Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. *Advances in Water Resources*, 30 (8), 1756–1774. doi:10.1016/j.advwatres.2007.01.005