https://doi.org/10.5194/hess-2021-325
Preprint. Discussion started: 6 August 2021

Hydrology and Earth System Sciences
Open Access
Discussions

Improving the Pareto Frontier in multi-dataset calibration of hydrological models using metaheuristics

Silja Stefnisdóttir[1,2], Anna E. Sikorska-Senoner[3], Eyjólfur I. Ásgeirsson[1,2], and David C. Finger[1,2,4]

[1]School of Technology, Reykjavik University, Reykjavik, Iceland
[2]Sustainability Institute and Forum (SIF), Reykjavik University, Reykjavik, Iceland
[3]Department of Geography, University of Zurich, Zürich, Switzerland
[4]Energy Institute, Johannes Kepler University, Linz, Austria

Correspondence: David C. Finger (davidf@ru.is)

Abstract. Hydrological models are crucial tools in water and environmental resource management but they require careful calibration based on observed data. Model calibration remains a challenging task, especially if a multi-objective or multi-dataset calibration is necessary to generate realistic simulations of multiple flow components under consideration. In this study, we explore the value of three metaheuristics, i.e. (i) Monte Carlo (MC), (ii) Simulated Annealing (SA), and (iii) Genetic Algorithm (GA), for a multi-dataset calibration to simultaneously simulate streamflow, snow cover and glacier mass balances using the conceptual HBV model. Based on the results from a small glaciated catchment of the Rhone River in Switzerland, we show that all three metaheuristics can generate parameter sets that result in realistic simulations of all three variables. Detailed comparison of model simulations with these three metaheuristics reveals however that GA provides the most accurate simulations (with narrowest confidence intervals) for all three variables when using both the 100 and the 10 best parameter sets for each method. However, when considering the 100 best parameter sets per method, GA yields also some worst solutions from the pool of all methods' solutions. The findings are supported by a reduction of the parameter equifinality and an improvement of the Pareto frontier for GA in comparison to both other metaheuristic methods. Based on our results, we conclude that GA-based multi-dataset calibration leads to the most reproducible and consistent hydrological simulations with multiple variables considered.

1 Introduction

One of the greatest challenges in hydrological modelling is the calibration of numerical models to obtain realistic simulation results. Research on this topic dates back to the 1950s when the first computer programmes allowed fast processing of data (Beard, 1962; Wallis, 1965). The model calibration can be performed both manually or automatically (Yilmaz et al., 2010). Manual calibrations using educated guesses of model parameters revealed to be tedious, time-consuming and inefficient and often required expert knowledge (Boyle et al., 2000). Hence, automatic methods have been developed since the 1960s (Gupta et al., 1999). With the development of more powerful computers, the Monte Carlo technique emerged in the 1970s (Freeze, 1975, 1980), enabling the testing and validating of millions of calculations within a short time (Beven, 2021). Nevertheless, further research revealed that MC techniques combined with manual calibration lead to a higher model efficiency (Boyle et al., 2000). With the development of more complex hydrological models that can process multiple output variables, the need for multi-objective calibration (MOC) emerged in the late 1990s (Gupta et al., 1998; Yapo et al., 1998; Efstratiadis and Koutsoyiannis, 2010). In contrast to a single-objective calibration, MOC involves several objective functions that are simultaneously optimised during the model calibration. As these functions can be contradicting to each other, a search for the best parameter set is shifted towards the search for multiple acceptable parameter sets or parameter distributions (McIntyre et al., 2002), according to the concept of the Pareto Frontier or optimality (Efstratiadis and Koutsoyiannis, 2010). Following this concept, one best solution for multi-objective criteria may not exist and thus several sub-optimal solutions, called Pareto sets, are searched for. The most frequently applied MOC are i) multiple criteria, using two or more objective functions, e.g. Nash-Sutcliffe (Nash and Sutcliffe, 1970) and peak efficiency (Seibert, 2003), for the same output variable (e.g. streamflow) (Madsen, 2003; Pool et al., 2017a; Sikorska et al., 2018) or using hydrologic signatures (Euser et al., 2013; Hanus et al., 2021); ii) multiple gauging sites with the same output variable (Bai et al., 2017), and iii) multiple output variables, e.g. streamflow and soil moisture (Brocca et al., 2012; Mostafaie et al., 2018). The latter MOC approach is particularly suitable if the model simulates several fluxes, such as streamflow or snowmelt, that all should be reliably represented. Such a multi-output calibration has been successfully adapted in water quality modelling (Rode et al., 2007; Reichert and Schuwirth, 2012; Sikorska et al., 2015a), limnological modelling (Reichert, 1995; Finger et al., 2007) and ecological modelling (Kuppel et al., 2018; Tang et al., 2018, 2019), as observations of water quality variables are often accompanied by basic measurement of streamflow. Yet, MOC application to hydrologic models is often limited by the data availability as streamflow records are often the only output variable directly measured in the catchment (Kuppel et al., 2018). Examples of multi-output calibration focus on the inclusion of groundwater levels (Seibert, 2000; Fenicia et al., 2005; Khu et al., 2008) or soil moisture observations (Downer and Ogden, 2003), in addition to streamflow.

A breakthrough for a multi-output calibration was achieved with increasing use of remote sensing data that has opened several new possibilities to utilize additional data (Silvestro et al., 2015). Radar-based or satellite-based estimates are of great value for providing information on temperature, evaporation, soil moisture, or snow cover, especially for poorly gauged, unpopulated and remote locations such as mountainous catchments. Remote sensed estimates of soil moisture are, in addition to streamflow, the most commonly utilized data in a multi-output calibration of hydrologic models (Campo et al., 2006; Brocca et al., 2012; Rajib et al., 2016; Budhathoki et al., 2020). In terms of high-altitude catchments with significant snow processes, a logical inclusion is to utilize observations on snow cover (Chen et al., 2017; Duethmann et al., 2014; Finger, 2018), and in glaciated catchments also on glacier mass balance (Konz and Seibert, 2010; Finger et al., 2011, 2015; Etter et al., 2017; He et al., 2019), in addition to streamflow. van Tiel et al. (2020) provide an extensive review on multi-objective calibration of hydrological models in glaciated catchments. Indeed multi-output calibration has proven to be very beneficial for realistically simulating hydrological processes in high-altitude catchments with significant snowmelt or glacier-melt contributions (Finger et al., 2011), while a single-output calibration tends to a large underestimation of these processes (Finger et al., 2015; He et al., 2018). An accurate simulation of different flow generation processes is also crucial for tracking long-term changes in hydrological processes in the catchment enhanced by climate change (Finger et al., 2012; Etter et al., 2017; De Niet et al., 2020).

Having more than one output variables included in the model calibration requires however a suitable calibration technique that explores information contained in all output variables to constrain multiple model parameters that represent a trade-off between different variables according to the Pareto Frontier. The existence of multiple sub-optimal parameter sets can be also justified by the modelling uncertainty and the parameter equifinality problem (Beven and Freer, 2001). Consequently, the issue of an efficient multi-output calibration can be shifted to a choice of a sufficient parameter search technique that, on the one hand, enables for accounting for modelling uncertainty, and on the other hand, lowers computational efforts through shortening the time needed for the search of optimal parameter solutions. In this respect, several promising techniques for model calibration have been proposed that can be classified into two groups: likelihood-based and likelihood-free techniques. The first group includes Bayesian methods and have been developed within a multi-output calibration framework that enables informing parameters of model errors along with the parameters of a hydrological model (e.g., Renard et al. (2011); Thyer et al. (2009); Sikorska et al. (2015a); Tang et al. (2018, 2019)). Yet, these Bayesian-based optimization techniques require a definition of the error model along with the likelihood function and thus making statements on model error properties (Montanari and Koutsoyiannis, 2012; Sikorska et al., 2015b). In addition, the inference may become computationally expensive if based on several model outputs or long time series, and the identifiability problem may occur in the case of an uninformative prior on the error model (Sikorska et al., 2015a). In contrast to the above, the second group is based on likelihood-free optimization techniques that do not require making any assumptions on model error properties. The identifiability issue of model parameters is dealt with via using multiple parameter sets as a proxy of model uncertainty (Sikorska-Senoner et al., 2020). The likelihood-free optimization techniques include metaheuristics methods that rely on learning strategies to find near-optimal solutions (Maier et al., 2014). Such metaheuristics involve an iterative generation process that consists of exploring and exploiting the defined search (parameter) space (Zufferey, 2012). Examples of metaheuristics include such optimization methods as Monte Carlo (MC), Simulated Annealing (SA) and Genetic Algorithm (GA) (see Sect. 3.2 for further details on these optimization techniques). Although these different metaheuristics have been independently tested in different studies focusing on multiple output calibration of hydrological models (Seibert, 2000; Finger et al., 2015; Etter et al., 2017; Mostafaie et al., 2018), a comprehensive comparative study of them in the context of hydrological modelling has been lacking so far. Specifically, it is not clear yet which of them has the greatest value for a multi-output calibration of a hydrologic model in a catchment of complex terrain with several processes being important such as a glaciated catchment.

In this study, we focus on such a multi-output calibration of a small glaciated catchment located in the Swiss Alps (headwater of the Rhone River). We test three different metaheuristic optimization techniques, i.e., Monte Carlo (MC), Simulated Annealing (SA) and Genetic Algorithm (GA), with a conceptual semi-distributed hydrological model (HBV) with streamflow, snow cover and glacier mass balance used as three output variables. All three output variables are weighted in the same way in order not to prioritize any of them. The efficiency of simulating these three output variables is assessed via a multi-criteria objective function developed by Finger et al. (2011) for the same study catchment and the same output data sets (see sect. 3.4). Specifically, we evaluate the value of three metaheuristics in terms of the uncertainty estimates of all three output variables represented with multiple model parameter sets and the behaviour of the Pareto frontier given the same computational effort. Thus, the novelty of our work lies in confronting for the first time these three metaheuristics most frequently applied in hy-

https://doi.org/10.5194/hess-2021-325
Preprint. Discussion started: 6 August 2021

drology within a multi-output calibration framework to derive practical recommendations for further applications. Our specific research questions are:

1. Which of the three metaheuristic methods tested, can constrain model parameter sets when calibrated against all three observational data sets given the same computational effort measured by the number of model evaluations needed?

2. Which metaheuristic does yield model parameter sets giving the most reliable simulations of the three modelled outputs, i.e. streamflow, snowmelt and glacier mass balance?

3. How does the ranking of the metaheuristics' performance depend on the number of best parameter sets selected?

4. Which metaheuristic does provide the most balanced Pareto frontier for three modelled output variables?

2 Material

2.1 Study Site

The headwaters of the Rhone River are delimited by the catchment of the gauging station Gletsch (Figure 1). The catchment covers an area of 38.9 km^2 and about 42% of it is covered by the Rhonegletscher. The highest point in the catchment reaches up to 3630 m asl. While mean precipitation in the watershed accounts for 1789.9 mm per annum, the average runoff in Gletsch is 2.8 m^3 s^{-1} (see Table 1). The difference between long-term mean discharge and precipitation can be attributed to intense glacier melt and shrinking of glacier mass. The Rhone glacier has been researched for over 120 years (Huss et al., 2008), resulting in an extensive set of historical data and is, therefore, a suitable location for this study.

2.2 Observed data

Hourly air temperature and precipitation observations are available from the weather station Grimsel Hochspitz (Figure 1) operated since 1903 by the Federal Office of Meteorology and Climatology (MeteoSwiss). These data serve as an input for a hydrological model (sect. 3.1).

For the multi-dataset calibration, three independent output datasets were available for the study site for the period 2001 to 2008: (i) discharge data, (ii) glacier mass balances from the Rhone glacier, and (iii) satellite-derived snow cover observations. Daily discharge data were obtained from the gauging station at Gletsch (Figure 1). The gauging station is operated by the Swiss Federal Office of Environment (FOEN). The daily satellite snow cover images were obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) (Hall et al., 2002). The snow cover data from all days with less than 10% cloud cover were used. The glacier mass balances were provided by Huss et al. (2008) and consist of glacier mass changes during the accumulation period (1. October to 30. April) and the ablation period (1st. May to 30th. September) in different altitudes bands of 100 m. All datasets were pre-processed as described in detail in Finger et al. (2011).

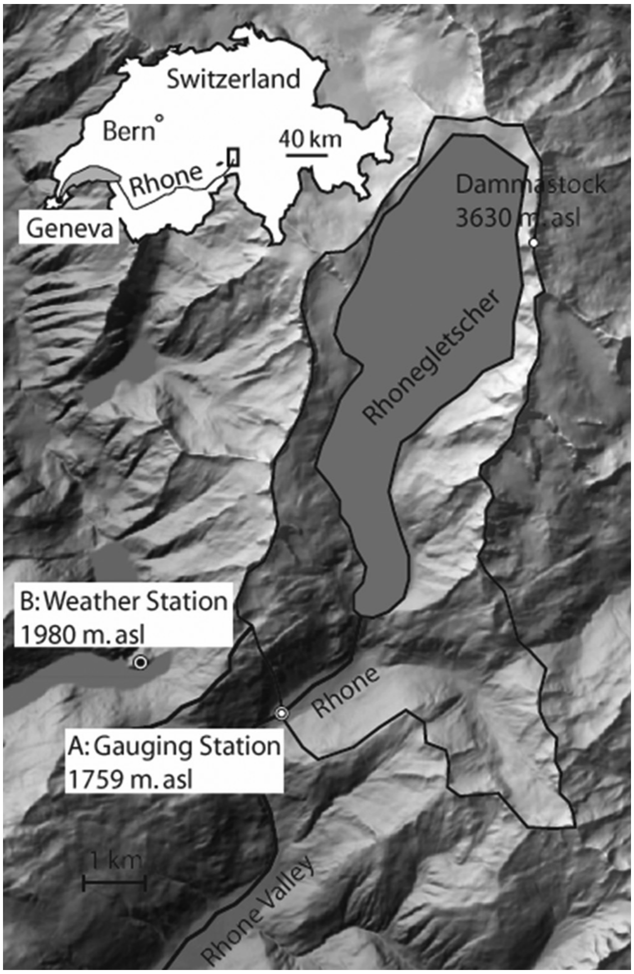

Figure 1. Rhone river catchment in Switzerland.

3 Methods

3.1 The HBV model

The *Hydrologiska Byråns Vattenbalansavdelning* (HBV) hydrological model is a conceptual precipitation-runoff model, developed by Sten Bergström (1976). The HBV model has been widely used in various versions and under different climate conditions (Pers, 2019). The version utilized in this study is identical to the one used by Finger et al. (2015). It relies on the HBV-light model (Seibert and Vis, 2012), being a simplified version of the HBV-6 model (Bergström, 1992) complemented with a glacier module (Finger et al., 2015) and a snow cover module to enable multi-dataset calibrations (Finger et al., 2011). The HBV-light is a semi-distributed model in which a catchment can be split into smaller sub-catchments with different elevation and vegetation zones (Seibert and Vis, 2012). It consists of five various routines which are: rainfall excess and snow

https://doi.org/10.5194/hess-2021-325
Preprint. Discussion started: 6 August 2021

Table 1. Catchment Characteristics of Rhone River.

River	Rhone
Gauging station	Gletsch
(Location CH 1903)[a]	670810 / 157200
Data period used	2001 - 2008
Catchment Area	38.9 km^2
Lowest Altitude	1761 m asl
Highest Altitude	3630 m asl
Mean Altitude	2719 m asl
Glacierization[b]	42.3%
Mean discharge[c]	2.8 $\mathrm{m}^3\ \mathrm{s}^{-1}$
	2770 mm a^{-1}
Weather Station	Grimsel Hospiz
Location	668583 / 158215
Distance to stream gauging	2.4 km (outside catchment)
Mean precipitation	1789.8 mm year^{-1}
Mean temperature	4.5 °C day^{-1}

[a] CH1905 system coordinates., [b] According to Huss et al. (2008), [c] According to the FOEN station data, available at http://www.hydrodaten.admin.ch.

routine, soil routine, groundwater routine, routing routine and glacier routine. The model input variables are daily precipitation (P) and air temperature (T). The phase of precipitation is determined by the air temperature. Using the inputs on a daily scale with the addition of potential evaporation (E), the model simulates daily discharge (Q) at the catchment outlet. Equation (1) shows the general water balance used by the HBV model:

$$P - E - Q = \frac{d}{dt}[SP + SM + UZ + LZ + lakes] \quad (1)$$

where SP is the snowpack, SM is moisture in the soil, LZ and UZ the lower and upper groundwater zones and $lakes$ represents the volume of the lake (Pers, 2019). The model's equations are described in detail by Seibert and Vis (2012).

The version of the HBV-light model used in this study has 21 parameters that must be optimized (see Table **??**). The optimization ranges for model parameters were taken from the previous study (Finger et al., 2015). This model is calibrated using the year 2008 within the multi-output framework with three output variables: discharge, snow cover and glacier mass balance. The model runs a warm-up period during the years 2005 to 2007 to set up initial catchment conditions. We use here the same calibration year as by Finger et al. (2015), to be able to compare our results with the previous work.

https://doi.org/10.5194/hess-2021-325
Preprint. Discussion started: 6 August 2021

3.2 Metaheuristic techniques used within the multi-output calibration

3.2.1 Monte Carlo

The Monte Carlo (MC) calibration procedure is performed by randomly selecting values for parameters from defined uniform distributions (Mooney and Sage Publications, 1997). The simulation is performed using multiple parameter sets and the outcome for each run is estimated and ranked based on the pre-defined objective function(s) (Seibert and Vis, 2012). Hence, all randomly selected parameter sets receive a rank which enables a selection of the most optimal solutions from all tested sets. The main drawback with the MC calibration procedure is that the number of runs required to achieve convergence can be quite significant (McIntyre et al., 2002) because there is no mechanism to improve selected solutions, i.e., the selection is completely random without any learning involved. As such, the MC procedure is focused only on exploration, i.e. the procedure is constantly trying out new parameter sets instead of trying to refine and improve good solutions. Despite this drawback, the MC calibration procedure has been shown to be able to identify good parameter sets if enough model trials are taken, and a selection of the 100 best parameter sets from a batch of 10'000 randomly generated has been shown to be sufficient for the parameter optimization (Konz and Seibert, 2010; Finger et al., 2011). Moreover, due to its simplicity, MC remains a frequently applied optimization technique in hydrology (Pool et al., 2017b; Finger, 2018; De Niet et al., 2020; Ferreira et al., 2021).

To further improve the MC search, a local search procedure could be added on top of the MC that would introduce the exploitation phase, where the method tries to improve the current solution by searching in its neighbourhood. If an improved solution is found, the current solution is updated.

Following Finger et al. (2011), in this paper, we randomly generated 10'000 parameter sets using the MC (i.e., without any local search), and for further analysis, we selected only 100 best based on the performance criteria described in sect. 3.3.

3.2.2 Simulated Annealing

Simulated Annealing (SA) is a flexible and generally applicable optimization technique first introduced by Kirkpatrick et al. (1983). In contrast to MC, SA involves both the exploration and exploitation by including the acceptance and rejection rate of randomly generated trails (Dougherty and Marryott, 1991). Moreover, SA tries to mitigate the problem of local optima by introducing a probability of accepting a worse solution, thereby allowing the hill-climbing search algorithm to escape from local optima. The algorithm uses a so-called temperature parameter t that decreases with each iteration of the algorithm and with it the probability of accepting a worse solution decreases also (Kirkpatrick et al., 1983). SA has been less frequently applied to hydrological modelling than MC but some recent works adapt SA for flood predictions (Zhu and Wu, 2013; Huang et al., 2018; Hosseini et al., 2020).

In this paper, the initial value for the temperature is 999 and the temperature decreases by 10% in each iteration, so $t_{i+1} = 0.9t_t$. The SA calibration procedure generates 200 initial random parameter sets, and then runs 50 iterations of SA for each parameter set. The total number of model evaluations is therefore 10'000, the same as for the MC calibration procedure.

3.2.3 Genetic Algorithm

The Genetic Algorithm (GA) is a metaheuristic that uses inspiration from natural selection to search for good solutions (Wan; Mitchell, 1998; Seibert, 2000; Sivanandam and Deepa, 2008). The algorithm maintains a set of solutions, called a population, and tries to improve the solution set using selection, crossover and mutation (Holland et al., 1992). New individuals are created using the crossover function, where two individuals (parents) are combined to create offspring. The selection method determines how the parents are selected and mutation allows for a possibility of the offspring having features that are not inherited directly from its parents. One of the keys to a good GA performance is to have a crossover function that fits the problem and the search space. GA based optimization techniques have been frequently applied in hydrology (Zhu and Wu, 2013; Brunner et al., 2018; Sikorska et al., 2018; Van Tiel et al., 2018; Sikorska-Senoner et al., 2020).

GA used in this study was introduced by Seibert (2000) and is designed for a multicriteria calibration of the HBV model. In this GA approach, the crossover function takes in two parameter sets (parents) and creates a new offspring using the following procedure for each parameter in the parameter set:

- with probability 0.41, the offspring receives the parameter value from Parent 1.
- with probability 0.41, the offspring receives the parameter value from Parent 2.
- with probability 0.16, the offspring receives a parameter value from a uniform distribution in the interval between the parameter values from Parent 1 and Parent 2.
- with probability 0.02, the offspring receives a parameter value from a uniform distribution in the interval between the minimum and maximum allowed values for the parameter, ignoring the parameter values of the parents. This is where mutation occurs.

The GA set-up used in this study uses a population size of 50 individuals and iterates for a total count of 40 generations, resulting in 2000 model runs. The process is performed 5 times resulting in a total of 10’000 model evaluations. In the end, a total of 250 parameter sets is delivered which are evaluated and the 100 best parameter sets are retained. The set-up of the GA was configured to have in total 10’000 model evaluations, i.e., the same number as both the MC and SA procedures.

3.3 Multi-criteria objective function for model calibration

The efficiency metric that is used to evaluate the model performance and the acceptance of model parameter sets during calibration consists of six criteria (see Table 2 for details). The accuracy of discharge (Q) simulations is determined by using the Nash-Sutcliffe efficiency (NSE) coefficient (Nash and Sutcliffe, 1970) which values may range from $[-\infty;1]$, where 1 describes a perfect fit between simulated and observed Q. NSE is computed over all discharge values and also for the logarithm values of discharges. Correctly predicted snow cover area (CPSC) is estimated for the whole year and also for the summer (1st. May to 30th. September), with the possible range between 0 and 1. The glacier mass balance (MB) simulation accuracy is estimated using root mean squared error (RMSE) for both the ablation (1st. May to 30th. September) and the accumulation

Table 2. The Efficiency criteria used within the multi-objective framework.

Objective[a]	Efficiency criteria	Equation
max	Nash-Sutcliffe of Q, (E_Q)[b]	$E_Q = 1 - \frac{\sum_{i=1}^{n} (q_{obs,i} - q_{sim,i})^2}{\sum_{i=1}^{n} (q_{obs,i} - \overline{q_{sim,i}})^2}$
max	Nash-Sutcliffe of $\log(\text{Q})$, (E_Q)[b]	$E_Q = 1 - \frac{\sum_{i=1}^{n} (\log q_{obs,i} - \log q_{sim,i})^2}{\sum_{i=1}^{n} (\log q_{obs,i} - \overline{\log q_{sim,i}})^2}$
max	Correctly predicted snow cover area, $(E_{SC,year})$[c]	$E_{SC} = \frac{1}{n} \sum_{i=1}^{n} (1 - \lvert a_{sim,i} - a_{obs,i} \rvert)$
max	Correctly predicted snow cover area, $(E_{SC,summer})$[c]	
min	Root mean square error of mass balance, $(E_{MB,acc})$[d]	$E_{MB} = \sqrt{\frac{1}{m} \sum_{j=1}^{m} (\Delta h_{ref,j} - \Delta h_{sim,j})^2}$
min	Root mean square error of mass balance, $(E_{MB,abl})$[d]	

[a]The objective signifies if the criterion is to be maximized or minimized. [b] q_{obs} represents the observed daily discharge, q_{sim} demonstrates the simulated daily discharge, both for time i. [c] a is the daily snow cover fraction for number of days, n. Also, sim stands for simulated, obs stands for observed estimations based on satellite images. [d] Δh is the combined change in snow and ice height [w. eq] for each altitude band (100 m), j during the indicated period.

205 season (1st. October to 30th. April). RMSE is scale-dependent, meaning the scale of RMSE depends on the scale of the data being estimated (Hyndman and Koehler, 2006). An RMSE value of 0 would describe a perfect fit to observed data. In addition, a normalized mass balance efficacy is computed for the ablation season.

All objective criteria described above are equally weighted during the model calibration to not prioritize any of the modelled variables. For the validation period, a normalized mass balance efficacy $(E_{MB,norm})$ is computed as:

210 $$E_{MB,norm} = 1 - \frac{E_{MB,abl}}{h_{ref,j,mean}} \tag{2}$$

where $E_{MB,abl}$ is the RMSE of mass balance and $h_{ref,j,mean}$ is the mean Δh for the entire validation period, which is the combined change in snow and ice height for each altitude band (100 m) during the indicated period.

3.4 Overall model performance in three methods

The overall consistency performance, P_r^{OAnorm} introduced by Finger et al. (2011), is used for comparison between values of 215 different model runs and optimization methods. P_r^{OAnorm} consists of normalized average ranking values obtained from all six efficiency criteria. The ranking value P_r^i is defined as

$$P_r^i = \frac{(N+1) - R_r^i}{N}, \tag{3}$$

https://doi.org/10.5194/hess-2021-325
Preprint. Discussion started: 6 August 2021

where i represents one of the six efficiency criteria, r each model run and N the total number of model runs per method. R_r^i represents the rank for criterion i in run r. As we intend all three variables to be equally well modelled, the overall consistency performance, P_r^{OA} is given by

$$P_r^{OA} = \frac{1}{3}\left(P_r^{Q} + P_r^{SC} + P_r^{MB}\right), \tag{4}$$

where P_r^{Q}, P_r^{SC} and P_r^{MB} are aggregated ranking values for discharge, snow cover and glacier mass balance for run r computed as an average of considered efficiency criteria for a respective variable and run. For instance, P_r^{Q} is an average of values achieved for NSE and log-NSE for the run r. This ranking approach is also used to select the best 100 and best 10 parameter sets per method.

The P_r^{OA} is then normalized by its maximum value achieved among all methods resulting in:

$$P_r^{OAnorm} = \frac{P_r^{OA}}{P_{max}^{OA}}. \tag{5}$$

Thus, P_r^{OAnorm} takes values from 0 (minimum) to 1 (maximum), whereas the minimum and maximum are defined based on obtained simulation results. In our case, as we compare 3 different methods of model calibration, we define the minimum and maximum based on all three methods. Hence, the minimum is the lowest rank and the maximum is the highest rank achieved overall three calibration methods, i.e. MC, SA and GA.

3.5 Statistical analysis of differences between calibration methods

Statistical analysis is performed to estimate if the difference in results between the three studied methods is statistically significant (Ross, 2014). We use a two-sample t-test for the three variables, i.e. discharge, snow cover and glacier mass balance, for the calibration year and the validation period. The level of significance is set as $\alpha = 0.05$ to estimate if the differences in the mean values for the three outputs are statistically significant.

4 Results

The number of iterations in each of three tested methods, MC, SA, and GA, was chosen to have the same number of computational runs for each method. From all optimised parameter sets, for each method 100 best parameter sets were chosen based on the ranking approach as described in Sect. 3.4. In addition, the efficiency criteria for glacier mass balance were normalized using the maximum value obtained for mass balance in the MC procedure. This proved to be necessary for the comparison of the SA and GA algorithm.

4.1 Optimized model parameters

Optimised HBV model parameters (mean and standard deviation values) of the best performing parameter sets for each method are presented in Table 3 and box plots of all 100 best parameter sets are illustrated in Figure 2. It can be seen that the MC method led to the largest span of optimised parameters, followed by the SA method, while the GA provided the narrowest box plots for

https://doi.org/10.5194/hess-2021-325
Preprint. Discussion started: 6 August 2021

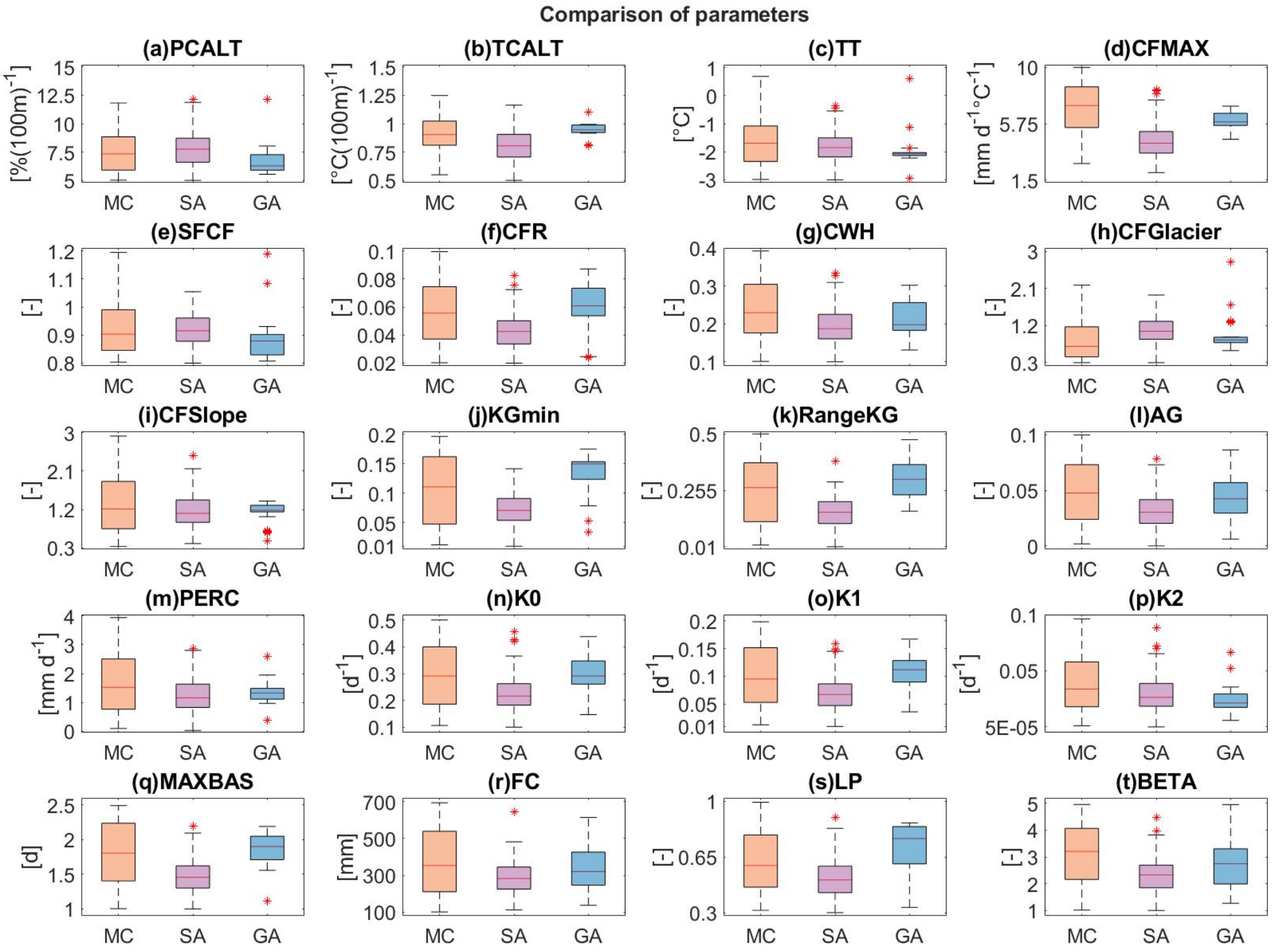

Figure 2. Box plots of 100 best-optimised parameter values of the HBV model for three metaheuristic methods: MC - Monte Carlo, SA - Simulated Annealing, GA - Genetic Algorithm. The red line within each box shows the median, the whiskers contain the most extreme data points, except for the outliers which are visualized by a red asterisk. The y-axis of each plot spans the parameter's optimisation range and unit. See Table 3 for the parameters' description.

all parameters. Both SA and GA parameter values lie within the ranges of MC optimised parameters for most of the parameters. This finding highlights the difference in the parameter search between these three methods. Namely, the MC-based search relies only on the exploration phase without any exploitation phase, which resulted in wide box plots. In contrast, both SA and GA involve an exploitation phase in addition to the exploration phase and this led to narrower parameter box plots.

https://doi.org/10.5194/hess-2021-325
Preprint. Discussion started: 6 August 2021

4.2 Model performance during the calibration period

The comparison of the model performance for all three methods is based on the simulations using the 100 best parameter sets and additionally also for the 10 best parameter sets selected based on the ranking approach. Table 4 demonstrates the mean efficiency criteria and standard deviations for all three methods using 100 and 10 best parameter sets. For the calibration year (2008) GA yielded the best results overall, both for 100 best runs and 10 best runs. The mean efficiency criterion from the 100 best runs resulted in the criteria values for discharge $E_{Q,GA} = 0.918 \pm 0.063$, snow cover $E_{SC,GA} = 0.917 \pm 0.013$ and glacier mass balance $E_{MB,GA} = 951.2 \pm 813.4$. The 10 best runs resulted in $E_{Q,GA} = 0.936 \pm 0.004$, $E_{SC,GA} = 0.917 \pm 0.003$ and $E_{MB,GA} = 639.8 \pm 157.6$. Both other methods led to slightly poorer model performance for all three runoff components than the GA method, and they delivered criteria values similar to each other for both 100 best and 10 best runs. However, a significant difference was measured neither between GA and SA for snow cover and glacier mass balance nor for GA and MC for snow cover from the 10 best runs (see Table 6 for results from the t-test). A significant difference was only measured between MC and SA for snow cover values from the 100 best runs, where MC delivered better results than SA (Table 6).

A visual comparison of the three metaheuristic methods versus the observed values is presented in Figure 3 for the 100 best simulations with a 90% simulation range and the 10 best simulations in Figure 4. It can be seen that GA, for both the 100 best and 10 best simulations, provides the narrowest simulation ranges for all three runoff components. Simulation ranges of both other methods (SA and MA) are wider in comparison to the GA. This is in agreement with our observations on model parameter ranges (sect. 4.1), where the GA method led to much narrower parameter ranges in comparison to the other two methods. It can be also seen from the comparison of Figures 3 and 4 that moving from 100 best to 10 best simulations leads to a substantial narrowing of the simulation ranges for all three methods and all three runoff components.

Further differences between the three methods can be seen from the analysis of monthly discharge and snow cover that are presented in Figure 5. Note that the monthly calculations for the glacier mass balance are not possible as the mass balance is calculated only on an annual scale. Similarly to time series analysis, also monthly analysis shows that both MC and SA provided wider simulation ranges than the GA. It can be also seen that respecting the discharge, all three methods reproduce quite well the monthly values with an exception of the period of September-November. For this period all methods slightly overestimate the monthly values and the simulation ranges do not cover the observed values. Regarding the snow cover, the coverage of the observation data by the simulation ranges is worse than for the discharge. Here two periods are poorly reproduced by the model, i.e. the period of May-July and September-October. The poorer performance of the model in these two periods for the snow cover and one period for the discharge overlap with the period of the most intensified snow and glacier melt (May-July) and the beginning of the snow accumulation period (September-November). Nevertheless, for both variables, i.e., discharge and snow cover, the peak values could be captured by the simulation ranges.

4.3 Model performance during the validation period

Similarly to the calibration period, the model performance in the validation period (2001-2007) was assessed for the 100 best runs and the 10 best runs, and the mean and standard deviation values are summarized in Table 5.

https://doi.org/10.5194/hess-2021-325
Preprint. Discussion started: 6 August 2021

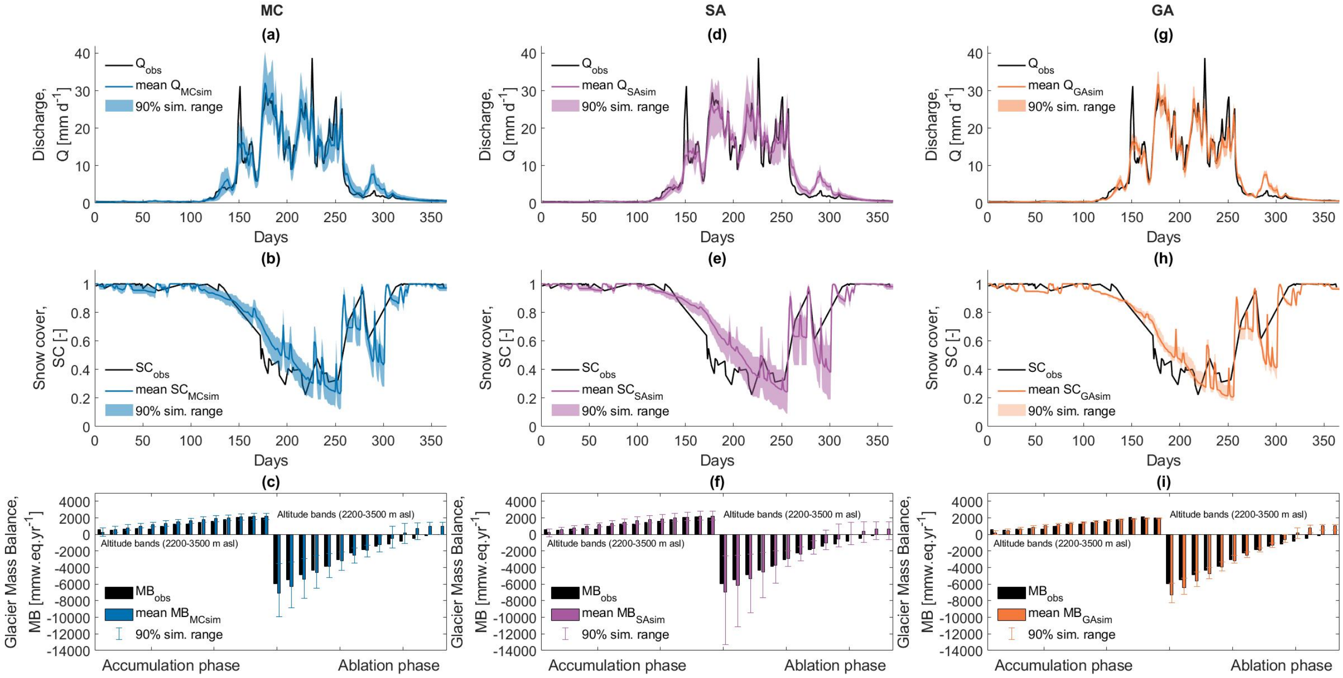

Figure 3. Model performance of the 100 best runs for each method for the Rhone catchment during the calibration year (2008). Plots a, d, g present the mean simulated discharge in a different colour line for each method together with the 90% simulation range in a lighter colour. The observed discharge is marked with a black line. Plots b, e, h demonstrate the mean simulated snow cover and the 90% simulation range versus the observed snow cover (black line). Plots c, f, i present the mean simulated glacier mass balance and the whiskers demonstrate the 90% simulation range. The observed glacier mass balance illustrated by black bars. Each bar group stands for a single altitude band.

Unlike the calibration year where the GA provided the best estimates, SA delivered the best results for discharge during the validation period, both from the 100 best runs ($E_{Q,SA} = 0.855 \pm 0.079$) and 10 best runs ($E_{Q,SA} = 0.833 \pm 0.037$). Note however that the simulation results for GA were only slightly worse than these of SA. Hence, a significant difference was not measured between SA and GA for discharge from the 100 best runs (Table 7). Regarding the snow cover, MC yielded the best results from the 100 best runs ($E_{SC,MC} = 0.901 \pm 0.022$) but GA delivered the best results from the 10 best runs ($E_{SC,GA} = 0.902 \pm 0.017$. Again, a significant difference between MC and GA for snow cover simulations was found neither from the 100 best runs nor the 10 best runs. Concerning the glacier mass balance, similarly to the calibration year (2008), GA delivered the best results for both the 100 best runs ($E_{MB,GA} = 0.673 \pm 0.300$) and 10 best runs ($E_{MB,GA} = 0.772 \pm 0.089$). Also here, a significant difference was not detected between the GA and SA runs from the 100 and 10 best runs.

Visual comparison of the model performance in the validation period including the calibration year is presented in Figure 6 for the 100 best runs and in Figure 7 for the 10 best runs with box plots. Looking at the performance of the 100 best sets for runoff components, it can be noticed that MC led to the widest spread of the model performance for the discharge. For snow cover, SA resulted in the widest spread and for the glacier mass balance, both SA and MC led to a wide spread of the

https://doi.org/10.5194/hess-2021-325
Preprint. Discussion started: 6 August 2021

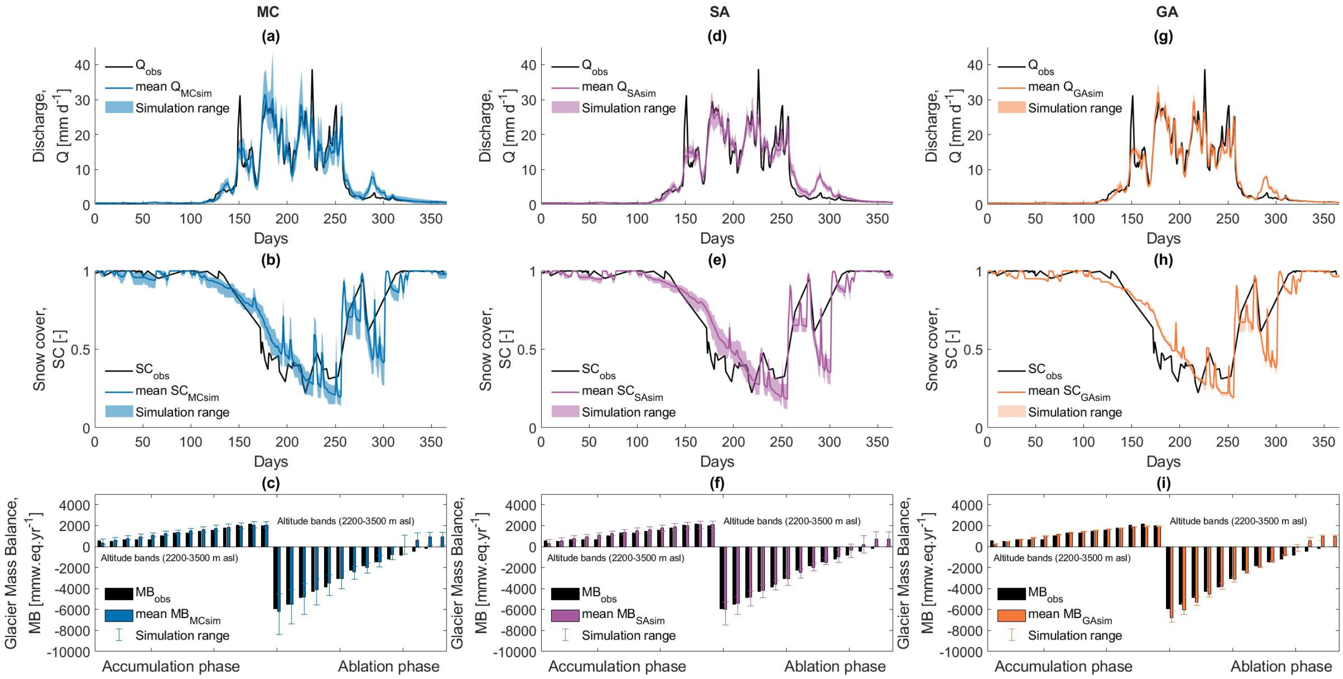

Figure 4. Model performance of the 10 best runs for each method for the Rhone catchment during the calibration year (2008). Description similar as in Figure 3 but this time simulation ranges represent all 10 best simulations instead of 90% ranges.

model performance. For all three runoff components, GA led to the narrowest box plots. However, GA yielded also some 'bad' solutions particularly for the discharge, as seen by the outliers lying below the 0-value.

When comparing the model performance evolution for three runoff components from the 100 best to the 10 best simulations, it can be noticed that all box plots are becoming narrower and grouped around higher values for efficiency criteria for all three methods. Still, the patterns visible for the 100 best parameter sets are present for the 10 best sets with the GA providing the narrowest box plots and MC and GA wider box plots. As it could be expected, 'bad' solutions are sieved out from the pool of the 10 best sets as seen by the disappearance of all outliers in Figure 7.

The validation of the modelling results furthermore reveals that the multi-dataset calibration is sufficiently robust to simulate exceptional hydro-meteorological events that occurred in the studied catchment over the observation period. Two examples of such events were the heatwave in 2003 (Schär et al., 2004) that caused an exceptional drought summer with very low flows, and the unprecedented precipitation event with subsequent devastating flooding in summer 2005 (Barredo, 2007). Model performance regarding EQ, EMB, and ESC show efficiency values for these two years 2003 and 2005 similar to these observed during the rest of the validation period (Figure 6 and 7). Comparing different calibration techniques, while efficiencies for the two exceptional years 2003 and 2005 for all three calibration methods are similarly high, results obtained with the GA method provide a slightly better model performance, underpinning the robustness of GA for multiple dataset calibrations.

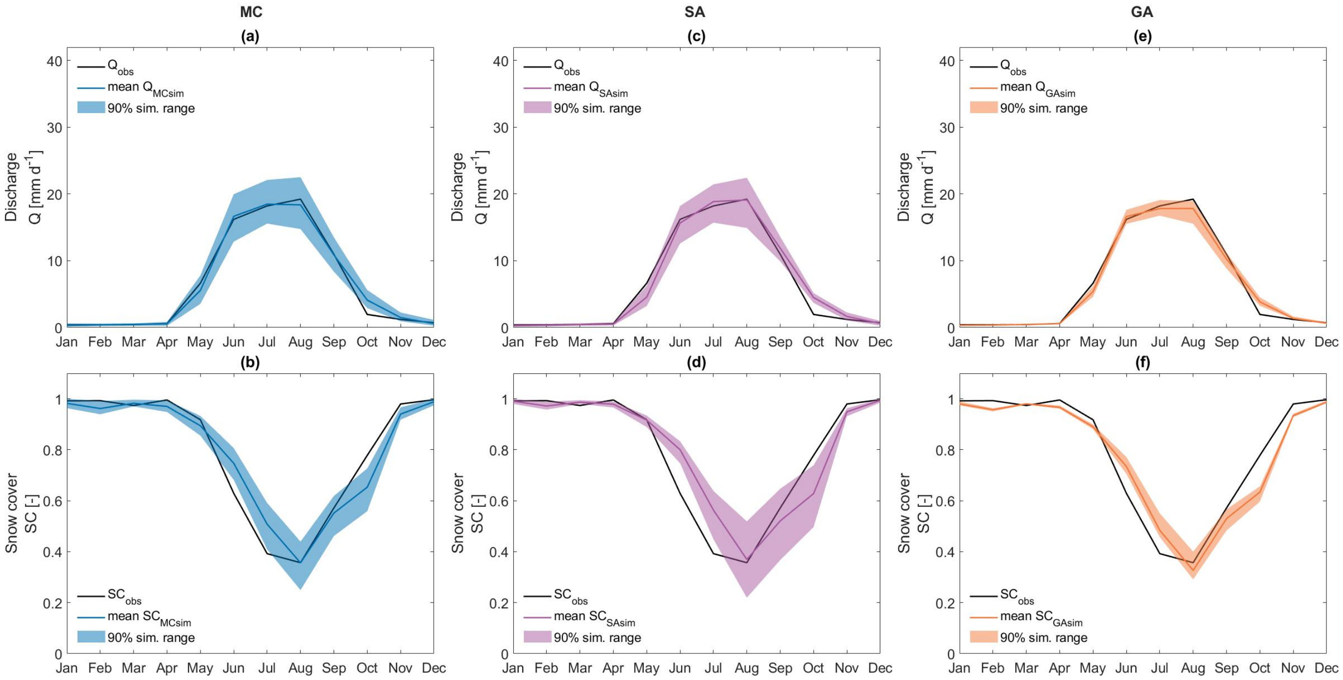

Figure 5. Monthly discharge and snow cover values of the 100 best simulation runs for each method for the Rhone catchment during the calibration year (2008). Plots a, d, e demonstrate the mean simulated discharge and the total simulation range versus the observed discharge. Plots b, d, f demonstrate the mean simulated snow cover and the total simulation range versus the observed snow cover.

4.4 Ranking of the model performance in three methods

Previous comparisons of the model performance in the three methods are based on the visual analysis of the model simulation performance and the model criteria. Yet, it would be of interest to see how the 100 best and the 10 best solutions compare to each other. For this purpose, the 100 best solutions of each method were put through a normalized ranking procedure (sect. 3.4) and compared to each other. Thus, all 100 best solutions for each method were put together creating a pool of 300 best solutions and ranked accordingly to their normalized rank received. Similarly, 10 best solutions for each method were pooled together creating a group of 30 solutions and ranked according to their normalized ranks. Figure 8 demonstrates how the method's runs were ranked when compared directly to each other when selecting the 100 best and 10 best solutions.

As it appears from the Figure and the normalized ranks, most of the 300 best solutions are arriving from the GA method. This method yielded however also some of the worst solutions. In contrast to that, both SA and MC yielded middle-good solutions with the SA yielding worse solutions than MC. When comparing only the 10 best solutions for each method (i.e., 30 in total), GA yielded all 10 best solutions. MC was placed second and SA third with all worst solution from all 30 considered.

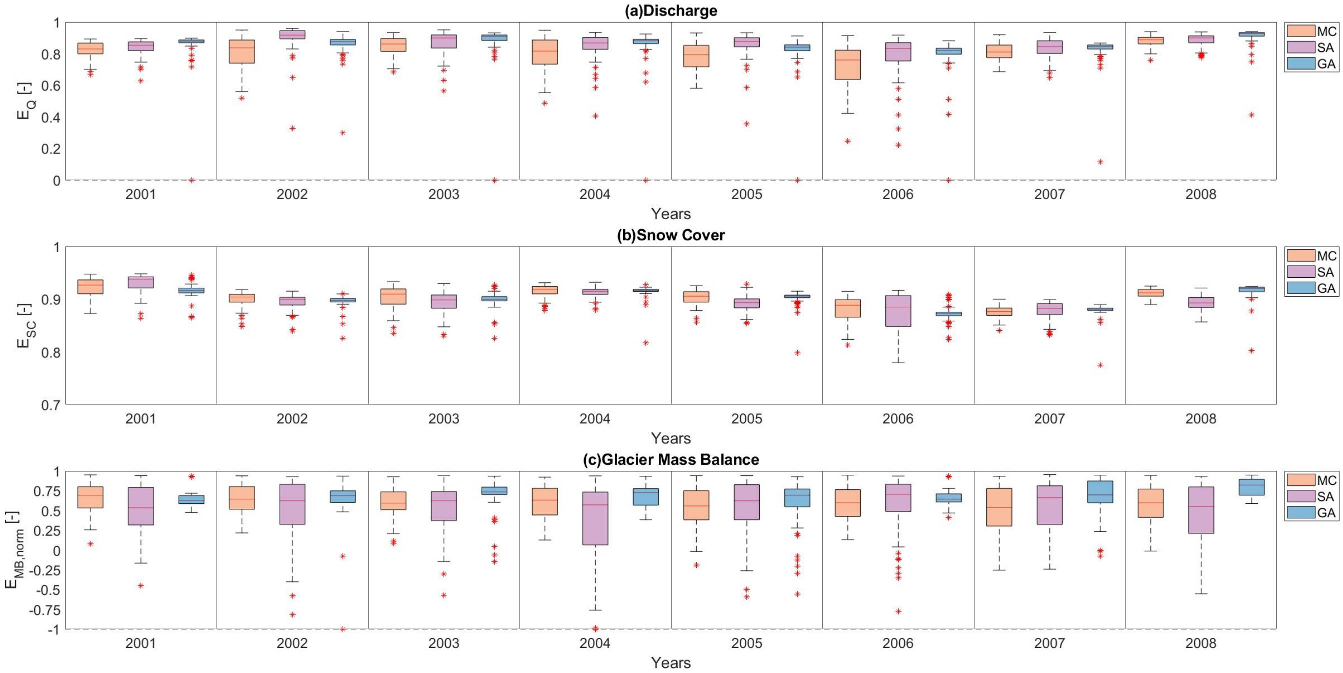

Figure 6. A comparison of the 100 best runs for each method for the Rhone catchment during the entire simulation period (2001-2008). The year 2008 is the calibration year and the years 2001-2007 are validation years. The red line within each box shows the median, the whiskers contain the most extreme data points, except for the outliers which are visualized by a red asterisk. Some outliers landed below the y-axis selected here and are therefore identified on the axis with the axis being dashed. (a) The discharge efficacy, (b) the snow cover efficacy, and (c) the normalized glacier mass balance efficacy.

4.5 Pareto frontier in three methods

As in any multi-criteria calibration, a competition between competing criteria occurs. This is illustrated by sub-optimal solutions for each of the considered criteria, i.e. by the occurrence of the Pareto frontier. This was also the case in our multi-output calibration with discharge, snow cover and glacier mass balance data. Therefore, we analysed the behaviour of the Pareto frontier, to identify the non-dominated runs which are runs where none of the three objective functions could be improved without the others worsening. Figure 9 demonstrates the 100 best runs of each of the three methods (panel a) and shows the 63 non-dominated runs (panel b), i.e. the Pareto frontier. Among three compared methods, GA yielded the highest number of non-dominated runs, i.e. 46 runs. The MC method yielded 13 runs and SA only 4 non-dominated runs. Thus, the GA method led to the largest improvement of the Pareto frontier by finding the most sub-optimal solutions for all three variables considered during the calibration.

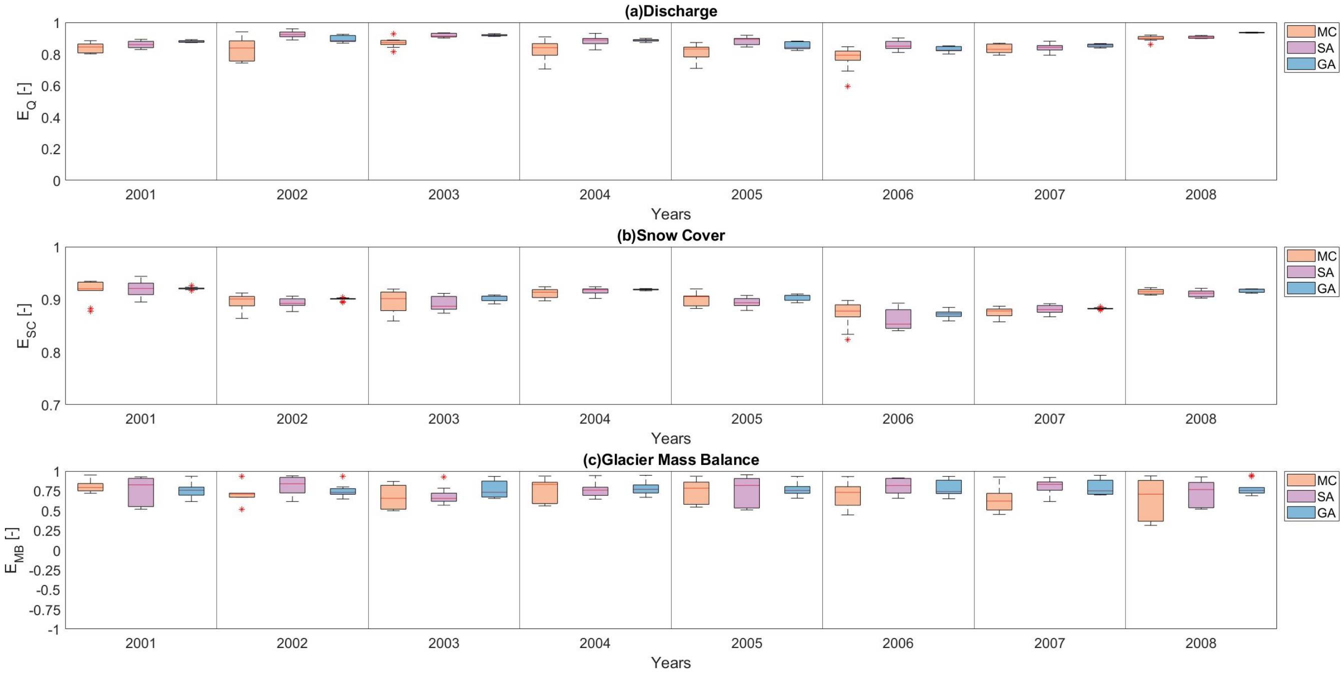

Figure 7. A comparison of the 10 best runs for each method for the Rhone catchment during the simulation period (2001-2008).The axis is the same as in Figure 6 for comparison reasons. Description as in Figure 6.

5 Discussion

5.1 Comparison of three metaheuristic methods for multi-output calibration

Given the same number of model runs, the comparison of three calibration methods, i.e. Monte Carlo (MC), Simulated Annealing (SA), and Genetic Algorithm (GA) reveals that all three methods can identify parameter sets that generate robust solutions, which can generate simulations that adequately reproduce all three observational datasets (streamflow, snow cover and glacier mass balance). These solutions lead to an ensemble of adequate simulations since an improvement for one observational dataset implies a decline in the performance of another dataset. However, all three calibration methods identify ensemble solutions that reveal a deviation from observed data points that are smaller than the standard deviation of the ensemble solution (Fig. 4). This is also valid for an 8-year validation period (2001 to 2008), including a record-breaking heatwave in 2003 and an unprecedented precipitation event in 2005 (Fig. 7). During the entire validation period, the results generated by GA reveal the highest mean performance and the lowest scattering of the model parameters and the efficiency values. This indicates that GA is most efficient in identifying high-performance and robust solutions that pass the validation for all three datasets. Hence, it appears that GA approaches faster good solutions than the other two methods tested here, which shows a great potential of this method for model calibration within a multi-data approach. We, however, do not neglect the other two methods which may reach similar model efficiency in an extended number of model runs. Nevertheless, the GA appeals to be the most promising

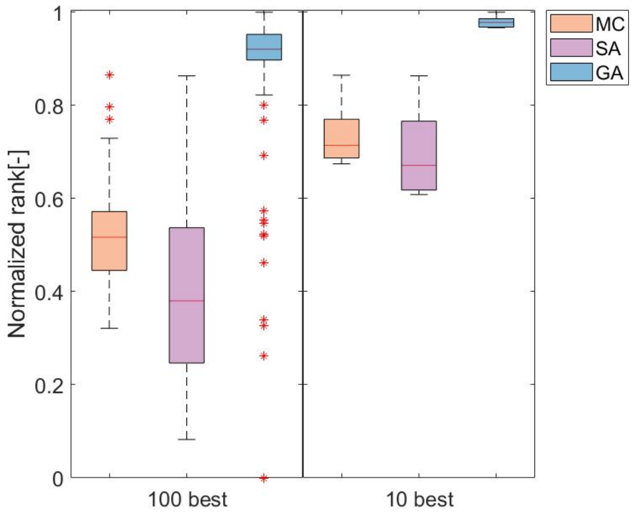

Figure 8. A comparison of the normalized ranking of the 100 best and the 10 best runs for each method for the Rhone catchment during the calibration year (2008). The best run among all three methods is indicated with the value of 1, whereas the worst run received the value of 0. The red line within each box shows the median, the whiskers contain the most extreme data points, except for the outliers which are visualized by a red asterisk.

method for multi-data calibrations when a computational cost should be kept low. In addition, moving from 100 best to 10 best solutions sieved out poorer solutions of the GA method. This feature could be further adapted to improve the model simulation skills, e.g. by increasing the number of initially performed model simulation runs to increase the number of good solutions retained. This is however linked with increased computational cost. Alternatively, further improvement in model performance skills could be achieved by coupling an optimized hydrological model with a data-driven approach to mimic model residuals (Sikorska-Senoner and Quilty, 2021).

5.2 Value of the multi-data calibration for glaciated catchments

The presented results fortify the conclusion that multi-data calibration is essential to constrain the parameter uncertainty, reduce the equifinality problem and improve the overall consistency performance of hydrological models for different runoff components being modelled. This is in agreement with previous studies (Finger et al., 2015; Sikorska et al., 2015a; Tarasova et al., 2016; van Tiel et al., 2020). While it remains important to maximise the model performance of a primary target variable, which is in most cases the total water streamflow, only the simultaneous calibration with relevant components of the water cycle can improve the overall robustness of the model calibration and can ensure the correct calibration for the right reason (Kirchner, 2006). Robust model calibration is extremely important for model applications to other periods or conditions. As demonstrated by our validation results, a multi-data calibration with three variables and one observation year was sufficient to constrain robust model parameters that are transferable to other even extreme conditions, such as extremely dry or wet periods observed in 2003 and 2005 year. In the particular case of glacierized high alpine catchments, as an example studied here, the most

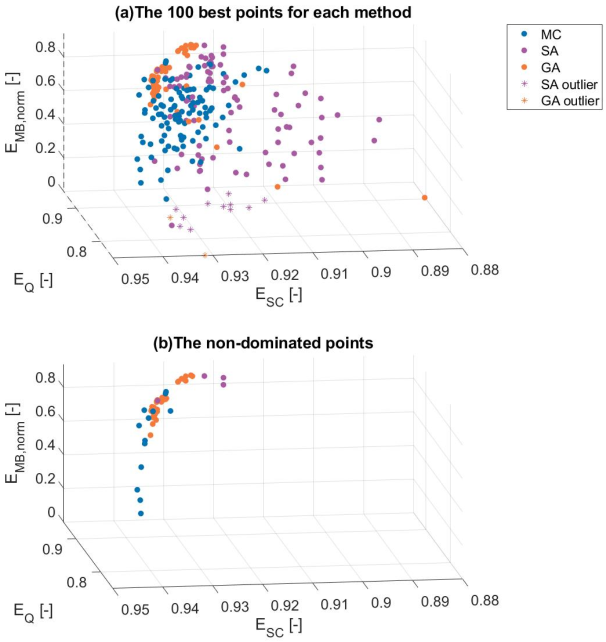

Figure 9. Visual representation of the results by the discharge efficacy, the snow cover efficacy, and the glacier mass balance efficacy for the Rhone catchment during the calibration year (2008). Some runs landed outside the axis and are therefore identified on the axis by an asterisk in the colour of its method, for clarifications of this purposes the axis is dashed. (a) A comparison of the 100 best points for each method and (b) isolated non-dominated points.

relevant processes generating runoff are snowmelt (observed in the daily satellite images), glacier volume change (monitored by biannual glacier mass balances) and surface runoff (observed in the short term dynamics of the total streamflow). Our results reveal that giving equal weights to all three datasets enabled us to obtain robust and consistent hydrological simulations for all these three runoff components (streamflow, snowmelt and glacier melt). Although we did not test other weight configurations, giving a higher weight to one of these datasets would presumably slightly improve the performance for that component but at the price of a drop in model performance for the other two runoff components. Similarly, calibrating the model only against the discharge data would not provide reliable information on changes in snow and glacier dynamics. Therefore, it appears to be essential to give equal weights to all modelled runoff components if all of them are of interest or long-term changes in water resources should be studied. The model ability to preserve the water balance and to realistically simulate water resources over a longer-term period is of particular importance for glaciated catchments, i.e., with slowly occurring changes in water resources (Hanzer et al., 2016; van Tiel et al., 2020).

5.3 Value of the multi-data calibration and Genetic Algorithm for improving the Pareto frontier

All three methods (MC, SA and GA) identify the limits of possible best solutions expressed by the Pareto Frontier (Fig. 9). The Pareto frontier represents the mathematically optimal solutions of the model to reproduce the observational datasets and reveals the essence of the equifinality problem (Beven and Freer, 2001) that occurs between two or more competitive calibration criteria (Efstratiadis and Koutsoyiannis, 2010). While the scattering of the final results is highest using the MC method, the application of GA identifies results that have a higher mean performance and shows the lowest scattering of the simulation results. This is valid for all three observational datasets, i.e., glacier mass balances, satellite-derived snow cover, and total runoff in the downstream valley. In addition, the comparison of the parameter estimation by the tree calibration methods reveals that some parameters are constrained better by GA than MC or SA (e.g., TT, PCALT, and TCALT). As argued by Finger et al. (2011, 2015), the value of datasets to constrain model parameters depends on the relevance of the parameter for the calculation of a specific runoff component according to this dataset. Accordingly, parameters describing snow melt are better constrained by satellite-derived snow cover images, parameters describing precipitation distribution and ice melt are best constrained by glacier mass balances, and parameters governing the short-term dynamic in the runoff are best constrained with streamflow data. The smaller scattering of parameters identified by GA indicates that this method is more efficient in identifying optimal parameters for all three runoff components than MS or SA. Since GA increases simultaneously the performance regarding all three datasets, it can be concluded that GA improves the Pareto Frontier, decreases the equifinality and enhances the parameter estimation, given the same number of model runs for all three methods.

5.4 Recommendations and outlook

Based on our results, we recommend a multi-data calibration in combination with the GA method for optimal calibration of a hydrological model and constraining model parameters. GA has been shown to be the most efficient in finding optimal Pareto solutions at a limited computational cost (expressed by the number of runs needed to be performed) in comparison to two other methods. As for glaciated catchments, it is essential to use simultaneously glacier mass balances, snow cover images, and discharge to obtain adequate and robust simulations of different runoff components. These three datasets coupled into a multi-data calibration are also used most often for calibration a hydrological model in glaciated catchments if three variables are considered (van Tiel et al., 2020). Optionally, further datasets could be included in multi-data approaches such as evapotranspiration data, soil moisture, stable isotopes and other datasets in catchments of multiple processes contributing to the runoff generation. For instance, in arid and semi-arid catchments, an additional dataset on evaporation could be of great value for the model (Dembélé et al., 2020). Such multi-dataset calibrations are in particular valuable in remote catchments with limited in-situ data and where multiple datasets can provide additional information for constraining the model parameters. The high performance of results obtained with GA demonstrates the potential of this approach for the multi-data calibration with an application to other environmental models, such as sediment transport models, water quality models or ecohydrological models (e.g., Finger et al. (2007); Sikorska et al. (2015a); Kuppel et al. (2018)), which should be investigated in further studies.

6 Conclusions

Three calibration methods were implemented in this study within a multi-data calibration approach to assess their applicability for providing robust model simulations of three different runoff components in a glaciated catchment; i.e., streamflow, snowmelt and glacier-melt. All three calibration techniques, i.e. Monte Carlo (MC), Simulated Annealing (SA) and Genetic Algorithm (GA) were constrained in the way that the same number of model runs is performed during the model calibration. Our results indicated that, first, among the three tested methods, the GA provided the most robust simulations of all three runoff components and the most optimal Pareto solutions. Hence, this method overperformed the other two methods as it approaches faster to good solutions and thus has a lower computational requirement. Second, our results demonstrated once again the value of the multi-dataset calibration for realistically simulating different runoff components, which is of particular importance for glaciated catchments or other catchments with multiple processes contributing to the runoff generation. Based on that, we recommend using a multi-data calibration in combination with the GA method for optimal calibration of hydrological models at a limited computational cost.

Author contributions. SS, AESS, EIA and DCF developed the concept and design the study details. The data and model code was provided by DCF. SS performed analysis, model calibrations and generated figures. All authors contributed to interpreting the results. SS and AESS wrote the manuscript draft which was revised and edited by DCF and EIA.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The first author was partially financed by the Icelandic Ministry of Social Affairs through the "Summer Jobs for Students" project. The authors would like to thank the Icelandic Ministry of Social Affairs, the School of Technology and the Sustainability Institute and Forum (SIF) at Reykjavik University, the Department of Geography at University of Zurich and the Energy Institute at Johannes Kepler University.

References

Bai, J., Shen, Z., and Yan, T.: A comparison of single- and multi-site calibration and validation: a case study of SWAT in the Miyun Reservoir watershed, China, Front. Earth Sci., 11, 592–600, https://doi.org/10.1007/s11707-017-0656-x, 2017.

Barredo, J.: Major flood disasters in Europe: 1950–2005, Nat Hazards, 42, 125–148, https://doi.org/10.1007/s11069-006-9065-2, 2007.

Beard, L.: Statistical Methods in Hydrology, Civil Works Investigations Project CW-151, U.S.Army Engineer District, Corps of Engineers, 1962.

Bergström, S.: Development and Application of a Conceptual Runoff Model for Scandinavian Catchments, Department of Water Resources Engineering, Lund Institute of Technology, University of Lund, 1976.

Bergström, S.: The HBV Model: Its Structure and Applications, SMHI, 1992.

Beven, K.: An epistemically uncertain walk through the rather fuzzy subject of observation and model uncertainties1, Hydrological Processes, 35, e14 012, https://doi.org/10.1002/hyp.14012, 2021.

Beven, K. and Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, Journal of Hydrology, 249, 11–29, https://doi.org/10.1016/S0022-1694(01)00421-8, 2001.

Boyle, D. P., Gupta, H. V., and Sorooshian, S.: Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, Water Resources Research, 36, 3663–3674, 2000.

Brocca, L., Moramarco, T., Melone, F., Wagner, W., Hasenauer, S., and Hahn, S.: Assimilation of Surface- and Root-Zone ASCAT Soil Moisture Products Into Rainfall–Runoff Modeling, IEEE Transactions on Geoscience and Remote Sensing, 50, 2542–2555, https://doi.org/10.1109/TGRS.2011.2177468, 2012.

Brunner, M., Sikorska, A., and Seibert, J.: Bivariate analysis of floods in climate impact assessments, Science of The Total Environment, 616-617, 1392–1403, https://doi.org/10.1016/j.scitotenv.2017.10.176, 2018.

Budhathoki, S., Rokaya, P., Lindenschmidt, K.-E., and Davison, B.: A multi-objective calibration approach using in-situ soil moisture data for improved hydrological simulation of the Prairies, Hydrological Sciences Journal, 65, 638–649, https://doi.org/10.1080/02626667.2020.1715982, 2020.

Campo, L., Caparrini, F., and Castelli, F.: Use of multi-platform, multi-temporal remote-sensing data for calibration of a distributed hydrological model: an application in the Arno basin, Italy, Hydrological Processes, 20, 2693–2712, https://doi.org/10.1002/hyp.6061, 2006.

Chen, X., Long, D., Hong, Y., Zeng, C., and Yan, D.: Improved modeling of snow and glacier melting by a progressive two-stage calibration strategy with GRACE and multisource data: How snow and glacier meltwater contributes to the runoff of the Upper Brahmaputra River basin?, Water Resources Research, 53, 2431–2466, https://doi.org/10.1002/2016WR019656, 2017.

De Niet, J., Finger, D. C., Bring, A., Egilson, D., Gustafsson, D., and Kalantari, Z.: Benefits of Combining Satellite-Derived Snow Cover Data and Discharge Data to Calibrate a Glaciated Catchment in Sub-Arctic Iceland, Water, 12, https://doi.org/10.3390/w12040975, 2020.

Dembélé, M., Ceperley, N., Zwart, S., Salvadore, E., Mariethoz, G., and Schaefli, B.: Potential of satellite and reanalysis evaporation datasets for hydrological modelling under various model calibration strategies, Advances in Water Resources, 143, 103 667, https://doi.org/10.1016/j.advwatres.2020.103667, 2020.

Dougherty, D. E. and Marryott, R. A.: Optimal Groundwater Management: 1. Simulated Annealing, Water Resources Research, 27, 2493–2508, https://doi.org/10.1029/91WR01468, 1991.

Downer, C. W. and Ogden, F. L.: Prediction of runoff and soil moistures at the watershed scale: Effects of model complexity and parameter assignment, Water Resources Research, 39, https://doi.org/10.1029/2002WR001439, 2003.

Duethmann, D., Peters, J., Blume, T., Vorogushyn, S., and Güntner, A.: The value of satellite-derived snow cover images for calibrating a hydrological model in snow-dominated catchments in Central Asia, Water Resources Research, 50, 2002–2021, https://doi.org/10.1002/2013WR014382, 2014.

Efstratiadis, A. and Koutsoyiannis, D.: One decade of multi-objective calibration approaches in hydrological modelling: a review, Hydrological Sciences Journal, 55, 58–78, https://doi.org/10.1080/02626660903526292, 2010.

Etter, S., Addor, N., Huss, M., and Finger, D.: Climate change impacts on future snow, ice and rain runoff in a Swiss mountain catchment using multi-dataset calibration, Journal of Hydrology: Regional Studies, 13, 222–239, 2017.

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., and Savenije, H. H. G.: A framework to assess the realism of model structures using hydrological signatures, Hydrology and Earth System Sciences, 17, 1893–1912, https://doi.org/10.5194/hess-17-1893-2013, 2013.

Fenicia, F., Zhang, G., Rientjes, T., Hoffmann, L., Pfister, L., and Savenije, H.: Numerical simulations of runoff generation with surface water–groundwater interactions in the Alzette river alluvial plain (Luxembourg), Physics and Chemistry of the Earth, Parts A/B/C, 30, 277 – 284, https://doi.org/10.1016/j.pce.2004.11.001, dealing with Floods within Constraints, 2005.

Ferreira, D., Scapulatempo Fernandes, C. V., E., K., and Bleninger, T.: Calibration of river hydrodynamic models: Analysis from the dynamic component in roughness coefficients, Journal of Hydrology, 598, 126 136, https://doi.org/10.1016/j.jhydrol.2021.126136, 2021.

Finger, D.: The value of satellite retrieved snow cover images to assess water resources and the theoretical hydropower potential in ungauged mountain catchments, Jokull, 68, 47–66, 2018.

Finger, D., Schmid, M., and Wuest, A.: Comparing effects of oligotrophication and upstream hydropower dams on plankton and productivity in perialpine lakes, Water Resources Research, W12404, https://doi.org/10.1029/2007WR005868, 2007.

Finger, D., Pellicciotti, F., Konz, M., Rimkus, S., and Burlando, P.: The value of glacier mass balance, satellite snow cover images, and hourly discharge for improving the performance of a physically based distributed hydrological model, Water Resources Research, 47, 2011.

Finger, D., Heinrich, G., Gobiet, A., and Bauder, A.: Projections of future water resources and their uncertainty in a glacierized catchment in the Swiss Alps and the subsequent effects on hydropower production during the 21st century, Water Resources Research, 48, https://doi.org/10.1029/2011WR010733, 2012.

Finger, D., Vis, M., Huss, M., and Seibert, J.: The value of multiple data set calibration versus model complexity for improving the performance of hydrological models in mountain catchments, Water Resources Research, 51, 1939–1958, 2015.

Freeze, R. A.: A stochastic-conceptual analysis of one-dimensional groundwater flow in nonuniform homogeneous media, Water Resources Research, 11, 725–741, https://doi.org/10.1029/WR011i005p00725, 1975.

Freeze, R. A.: A stochastic-conceptual analysis of rainfall-runoff processes on a hillslope, Water Resources Research, 16, 391–408, https://doi.org/10.1029/WR016i002p00391, 1980.

Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, Water Resources Research, 34, 751–763, https://doi.org/10.1029/97WR03495, 1998.

Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Status of Automatic Calibration for Hydrologic Models: Comparison with Multilevel Expert Calibration, Journal of Hydrologic Engineering, 4, 135–143, 1999.

Hall, D. K., Riggs, G. A., Salomonson, V. V., DiGirolamo, N. E., and Bayr, K. J.: MODIS snow-cover products, Remote Sensing of Environment, 83, 181–194, 2002.

Hanus, S., Hrachowitz, M., Zekollari, H., Schoups, G., Vizcaino, M., and Kaitna, R.: Timing and magnitude of future annual runoff extremes in contrasting Alpine catchments in Austria, Hydrology and Earth System Sciences Discussions, 2021, 1–35, https://doi.org/10.5194/hess-2021-92, 2021.

Hanzer, F., Helfricht, K., Marke, T., and Strasser, U.: Multilevel spatiotemporal validation of snow/ice mass balance and runoff modeling in glacierized catchments, The Cryosphere, 10, 1859–1881, https://doi.org/10.5194/tc-10-1859-2016, 2016.

He, Z., Vorogushyn, S., Unger-Shayesteh, K., Gafurov, A., Kalashnikova, O., Omorova, E., and Merz, B.: The Value of Hydrograph Partitioning Curves for Calibrating Hydrological Models in Glacierized Basins, Water Resources Research, 54, 2336–2361, https://doi.org/10.1002/2017WR021966, 2018.

He, Z., Unger-Shayesteh, K., Vorogushyn, S., Weise, S. M., Kalashnikova, O., Gafurov, A., Duethmann, D., Barandun, M., and Merz, B.: Constraining hydrological model parameters using water isotopic compositions in a glacierized basin, Central Asia, Journal of Hydrology, 571, 332 – 348, https://doi.org/10.1016/j.jhydrol.2019.01.048, 2019.

Holland, J. H. et al.: Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence, MIT press, 1992.

Hosseini, F., Choubin, B., Mosavi, A., Nabipour, N., Shamshirband, S., Darabi, H., and Haghighi, A.: Flash-flood hazard assessment using ensembles and Bayesian-based machine learning models: Application of the simulated annealing feature selection method, Science of The Total Environment, 711, 135 161, https://doi.org/10.1016/j.scitotenv.2019.135161, 2020.

Huang, C.-L., Hsu, N.-S., Liu, H.-J., and Huang, Y.-H.: Optimization of low impact development layout designs for megacity flood mitigation, Journal of Hydrology, 564, 542–558, https://doi.org/10.1016/j.jhydrol.2018.07.044, 2018.

Huss, M., Bauder, A., Funk, M., and Hock, R.: Determination of the seasonal mass balance of four Alpine glaciers since 1865, Journal of Geophysical Research: Earth Surface, 11, 2008.

Hyndman, R. J. and Koehler, A. B.: Another look at measures of forecast accuracy, International Journal of Forecasting, 22, 679–688, 2006.

Khu, S.-T., Madsen, H., and di Pierro, F.: Incorporating multiple observations for distributed hydrologic model calibration: An approach using a multi-objective evolutionary algorithm and clustering, Advances in Water Resources, 31, 1387 – 1398, https://doi.org/10.1016/j.advwatres.2008.07.011, 2008.

Kirchner, J. W.: Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, Water Resources Research, 42, https://doi.org/10.1029/2005WR004362, 2006.

Kirkpatrick, S., Gelatt, C., and Vecchi, M.: Optimization by Simulated Annealing, Science (New York, N.Y.), 220, 671–80, 1983.

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

Kuppel, S., Tetzlaff, D., and Maneta, M.: What can we learn from multi-data calibration of a process-based ecohydrological model?, Environmental Modelling and Software, 101, 301–316, https://doi.org/10.1016/j.envsoft.2018.01.001, 2018.

Madsen, H.: Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives, Advances in Water Resources, 26, 205–216, https://doi.org/10.1016/S0309-1708(02)00092-1, 2003.

Maier, H., Kapelan, Z., Kasprzyk, J., Kollat, J., Matott, L., Cunha, M., Dandy, G., Gibbs, M., Keedwell, E., Marchi, A., Ostfeld, A., Savic, D., Solomatine, D., Vrugt, J., Zecchin, A., Minsker, B., Barbour, E., Kuczera, G., Pasha, F., Castelletti, A., Giuliani, M., and Reed, P.: Evolutionary algorithms and other metaheuristics in water resources: Current status, research challenges and future directions, Environmental Modelling Software, 62, 271–299, https://doi.org/10.1016/j.envsoft.2014.09.013, 2014.

McIntyre, N., Wheater, H., and Lees, M.: Estimation and propagation of parametric uncertainty in environmental models, Journal of Hydroinformatics, 4, 177–198, 2002.

Mitchell, M.: L.D. Davis, Handbook of Genetic Algorithms*, Artificial Intelligence, 100, 325–330, 1998.

Montanari, A. and Koutsoyiannis, D.: A blueprint for process-based modeling of uncertain hydrological systems, Water Resources Research, 48, https://doi.org/10.1029/2011WR011412, 2012.

Mooney, C. and Sage Publications, i.: Monte Carlo Simulation, no. Nr. 116 in Monte Carlo Simulation, SAGE Publications, https://books.google.de/books?id=xQRgh4z_5acC, 1997.

Mostafaie, A., Forootan, E., Safari, A., and Schumacher, M.: Comparing multi-objective optimization techniques to calibrate a conceptual hydrological model using in situ runoff and daily GRACE data, Comput Geosci, 22, 789–814, https://doi.org/10.1007/s10596-018-9726-8, 2018.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I — A discussion of principles, Journal of Hydrology, 10, 282–290, 1970.

Pers, C.: HBV, https://www.smhi.se/en/research/research-departments/hydrology/hbv-1.90007, 2019.

Pool, S., Vis, M., Knight, R., and Seibert, J.: Streamflow characteristics from modeled runoff time series – importance of calibration criteria selection, Hydrology and Earth System Sciences, 21, 5443–5457, https://doi.org/10.5194/hess-21-5443-2017, 2017a.

Pool, S., Viviroli, D., and Seibert, J.: Prediction of hydrographs and flow-duration curves in almost ungauged catchments: Which runoff measurements are most informative for model calibration?, Journal of Hydrology, 554, 613–622, https://doi.org/10.1016/j.jhydrol.2017.09.037, 2017b.

Rajib, M. A., Merwade, V., and Yu, Z.: Multi-objective calibration of a hydrologic model using spatially distributed remotely sensed/in-situ soil moisture, Journal of Hydrology, 536, 192–207, https://doi.org/10.1016/j.jhydrol.2016.02.037, 2016.

Reichert, P.: Design techniques of a computer program for the identification of processes and the simulation of water quality in aquatic systems, Environmental Software, 10, 199–210, https://doi.org/10.1016/0266-9838(95)00010-I, 1995.

Reichert, P. and Schuwirth, N.: Linking statistical bias description to multiobjective model calibration, Water Resources Research, 48, https://doi.org/10.1029/2011WR011391, 2012.

Renard, B., Kavetski, D., Leblois, E., Thyer, M., Kuczera, G., and Franks, S. W.: Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation, Water Resour. Res., 47, W11 516, https://doi.org/10.1029/2011WR010643, 2011.

Rode, M., Suhr, U., and Wriedt, G.: Multi-objective calibration of a river water quality model—Information content of calibration data, Ecological Modelling, 204, 129–142, https://doi.org/10.1016/j.ecolmodel.2006.12.037, 2007.

Ross, S. M.: Introduction to Probability and Statistics for Engineers and Scientists, Academic Press, 2014.

Schär, C., Vidale, P., Lüthi, D., Frei, C., Häberli, C., Liniger, M. A., and Appenzeller, C.: The role of increasing temperature variability in European summer heatwaves, Nature, 427, 332–336, https://doi.org/10.1038/nature02300, 2004.

Seibert, J.: Multi-criteria calibration of a conceptual runoff model using a genetic algorithm, Hydrology and Earth System Sciences, 4, 215–224, https://doi.org/10.5194/hess-4-215-2000, publisher: Copernicus GmbH, 2000.

Seibert, J.: Reliability of Model Predictions Outside Calibration Conditions, Hydrology Research, 34, 477–492, https://doi.org/10.2166/nh.2003.0019, 2003.

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

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

Sikorska, A. E., Del Giudice, D., Banasik, K., and Rieckermann, J.: The value of streamflow data in improving TSS predictions – Bayesian multi-objective calibration, J. Hydrol., 530, 241–254, https://doi.org/10.1016/j.jhydrol.2015.09.051, 2015a.

Sikorska, A. E., Montanari, A., and Koutsoyiannis, D.: Estimating the Uncertainty of Hydrological Predictions through Data-Driven Resampling Techniques, J. Hydrol. Eng., 20, A4014 009, https://doi.org/10.1061/(ASCE)HE.1943-5584.0000926, 2015b.

Sikorska-Senoner, A. and Quilty, J.: A novel ensemble-based conceptual-data-driven approach for improved streamflow simulations, Environmental Modelling & Software, p. 105094, https://doi.org/10.1016/j.envsoft.2021.105094, 2021.

Sikorska-Senoner, A. E., Schaefli, B., and Seibert, J.: Downsizing parameter ensembles for simulations of rare floods, Natural Hazards and Earth System Sciences, 20, 3521–3549, https://doi.org/10.5194/nhess-20-3521-2020, 2020.

Silvestro, F., Gabellani, S., Rudari, R., Delogu, F., Laiolo, P., and Boni, G.: Uncertainty reduction and parameter estimation of a distributed hydrological model with ground and remote-sensing data, Hydrology and Earth System Sciences, 19, 1727–1751, https://doi.org/10.5194/hess-19-1727-2015, 2015.

Sivanandam, S. and Deepa, S.: Genetic Algorithms, pp. 15–37, Springer Berlin Heidelberg, Berlin, Heidelberg, https://doi.org/10.1007/978-3-540-73190-0_2, 2008.

Tang, Y., Marshall, L., Sharma, A., and Ajami, H.: A Bayesian alternative for multi-objective ecohydrological model specification, Journal of Hydrology, 556, 25–38, https://doi.org/10.1016/j.jhydrol.2017.07.040, 2018.

Tang, Y., Marshall, L., Sharma, A., Ajami, H., and Nott, D. J.: Ecohydrologic Error Models for Improved Bayesian Inference in Remotely Sensed Catchments, Water Resources Research, 55, 4533–4549, https://doi.org/10.1029/2019WR025055, 2019.

Tarasova, L., Knoche, M., Dietrich, J., and Merz, R.: Effects of input discretization, model complexity, and calibration strategy on model performance in a data-scarce glacierized catchment in Central Asia, Water Resources Research, 52, 4674–4699, https://doi.org/10.1002/2015WR018551, 2016.

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W., and Srikanthan, S.: Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis, Water Resour. Res., 45, W00B14, https://doi.org/10.1029/2008WR006825, 2009.

Van Tiel, M., Teuling, A. J., Wanders, N., Vis, M. J. P., Stahl, K., and Van Loon, A. F.: The role of glacier changes and threshold definition in the characterisation of future streamflow droughts in glacierised catchments, Hydrology and Earth System Sciences, 22, 463–485, https://doi.org/10.5194/hess-22-463-2018, 2018.

van Tiel, M., Stahl, K., Freudiger, D., and Seibert, J.: Glacio-hydrological model calibration and evaluation, WIREs Water, 7, e1483, https://doi.org/10.1002/wat2.1483, 2020.

Wallis, J. R.: Multivariate statistical methods in hydrology—A comparison using data of known functional relationship, Water Resources Research, 1, 447–461, https://doi.org/https://doi.org/10.1029/WR001i004p00447, 1965.

Yapo, P. O., Gupta, H. V., and Sorooshian, S.: Multi-objective global optimization for hydrologic models, Journal of Hydrology, 204, 83–97, https://doi.org/10.1016/S0022-1694(97)00107-8, 1998.

Yilmaz, K., Vrugt, J., Gupta, H., and Sorooshian, S.: Model Calibration in Watershed Hydrology, in: Advances in Data-Based Approaches for Hydrologic Modeling and Forecasting, pp. 53–105, World Scientific Publishing Co., 2010.

Zhu, C. and Wu, J.: Hybrid of genetic algorithm and simulated annealing for support vector regression optimization in rainfall forecasting, International Journal of Computational Intelligence and Applications, 12, 1350 012, https://doi.org/10.1142/S1469026813500120, 2013.

https://doi.org/10.5194/hess-2021-325
Preprint. Discussion started: 6 August 2021

Zufferey, N.: Metaheuristics: some principles for an efficient design, Comput. Technol. Appl., 3 (6), 446e462, 2012.

Table 3. The model parameters, range and mean values for each method during the calibration year (2008).

Parameter	Description	Range	SA		MC		GA	
		[min, max]	mean	std	mean	std	mean	std
Rescaling parameters of input data								
PCALT	Change of precipitation with elevation [%(100m)$^{-1}$]	[5, 15]	7.722	1.473	7.560	1.761	6.609	0.933
TCALT	Change of temperature with elevation [°C(100m)$^{-1}$]	[0.5, 1.5]	0.810	0.144	0.911	0.153	0.932	0.067
Snowmelt parameters								
TT	Threshold temperature for liquid and solid precipitation [°C]	[-3, 1]	-1.820	0.559	-1.620	0.871	-2.048	0.306
CFMAX	Degree-day factor [mm d^{-1}°C^{-1}]	[1.5, 10]	4.500	1.305	6.950	1.886	5.966	0.669
SFCF	Snowfall correction factor [-]	[0.8, 1.2]	0.914	0.057	0.926	0.101	0.874	0.055
CFR	Refreezing coefficient [-]	[0.02, 0.1]	0.043	0.013	0.057	0.024	0.058	0.018
CWH	Water holding capacity of the snow storage [-]	[0.1, 0.4]	0.195	0.050	0.240	0.081	0.220	0.049
Ice melt parameters								
CFGlacier	Glacier melt correction factor [-]	[0.3, 3]	1.066	0.317	0.860	0.493	0.919	0.306
CFSlope	Slope snowmelt correction factor [-]	[0.3, 3]	1.196	0.421	1.332	0.722	1.130	0.252
KGmin	Minimum value for the outflow coefficient representing conditions with poorly developed glacial drainage systems in late winter [-]	[0.01, 0.2]	0.072	0.028	0.106	0.059	0.137	0.023
RangeKG	Range of the annual outflow coefficient variation [-]	[0.01, 0.5]	0.157	0.073	0.259	0.143	0.306	0.087
AG	Calibration parameter defining the sensitivity of the outflow coefficient to changes in the snow storage [-]	[0, 0.1]	0.031	0.017	0.048	0.029	0.043	0.019
Soil parameters								
PERC	Maximum percolation from upper to lower groundwater storage [mm d^{-1}]	[0,4]	1.244	0.607	1.663	1.076	1.346	0.323
K0	Storage (or recession) coefficient [d^{-1}]	[0.1, 0.5]	0.228	0.068	0.295	0.122	0.299	0.077
K1	Storage (or recession) coefficient 1 [d^{-1}]	[0.01, 0.2]	0.068	0.033	0.100	0.055	0.110	0.029
K2	Storage (or recession) coefficient 2 [d^{-1}]	[5E-05, 0.1]	0.029	0.017	0.038	0.026	0.024	0.008
MAXBAS	Length of triangular weighting function [d]	[1, 2.5]	1.455	0.257	1.807	0.455	1.878	0.210
FC	Maximum soil moisture storage [mm]	[100, 700]	286.1	92.	373.1	190.1	337.	101.3
LP	Relative soil water storage below which AET is reduced linearly [-]	[0.3, 1]	0.510	0.124	0.629	0.194	0.719	0.130
BETA	Shape factor for the function used to calculate the distribution of rain and snowmelt going to runoff and soil box, respectively [-]	[1, 5]	2.333	0.693	3.072	1.154	2.667	0.829

Table 4. A comparison of the efficiency criteria during the calibration year (2008).

			100 best runs			10 best runs		
Objective	Efficiency criteria		MC	SA	GA	MC	SA	GA
max	Discharge, E_Q $[-]$	mean	0.885	0.883	0.918	0.901	0.910	0.936
		std	0.027	0.037	0.063	0.017	0.008	0.004
max	Snow cover, $E_{SC,summer}$ $[-]$	mean	0.911	0.894	0.917	0.915	0.912	0.917
		std	0.008	0.014	0.013	0.005	0.006	0.003
min	Mass balances, $E_{MB,abl}$ [mmw.eq.]	mean	1216.6	1413.6	951.2	813.7	545.1	639.8
		std	450.1	905.1	813.4	218.5	269.5	157.6

The gray shaded areas represent the best results for each objective.

Table 5. A comparison of the efficiency criteria during the validation period.

			100 best runs			10 best runs		
Objective	Efficiency criteria		MC	SA	GA	MC	SA	GA
max	Discharge, E_Q $[-]$	mean	0.811	0.855	0.852	0.835	0.883	0.881
		std	0.093	0.079	0.137	0.060	0.037	0.036
max	Snow cover, $E_{SC,summer}$ $[-]$	mean	0.901	0.897	0.900	0.898	0.896	0.902
		std	0.022	0.025	0.020	0.022	0.022	0.017
max	Mass balances, $E_{MB,norm}$ $[-]$	mean	0.596	0.515	0.673	0.711	0.768	0.772
		std	0.223	0.365	0.300	0.156	0.130	0.089

The gray shaded areas represent the best results for each objective.

https://doi.org/10.5194/hess-2021-325
Preprint. Discussion started: 6 August 2021

Table 6. Results of t-test for the calibration year ($\alpha = 0.025$).

		100 best runs			10 best runs		
Efficiency criteria		MC and SA	GA and SA	GA and MC	MC and SA	GA and SA	GA and MC
Discharge, E_Q $[-]$	t-Stat	0.54	-4.79	-4.75	-1.26	-9.97	-6.26
	P-value	0.59	3.8E-06	5.2E-06	0.23	1.9E-07	9.4E-05
Snow cover, $E_{SC,summer}$ $[-]$	t-Stat	11.45	-12.50	-3.69	1.19	-8.26	-1.41
	P-value	1.3E-22	8.5E-27	0.0003	0.25	2.7E-06	0.18
Mass balances, $E_{MB,norm}$ $[-]$	t-Stat	-1.95	3.80	2.86	2.45	-0.96	2.04
	P-value	0.137	0.0002	0.005	0.03	0.35	0.06

Table 7. Results of t-test for the validation period ($\alpha = 0.025$).

		100 best runs			10 best runs		
Efficiency criteria		MC and SA	GA and SA	GA and MC	MC and SA	GA and SA	GA and MC
Discharge, E_Q $[-]$	t-Stat	-10.10	0.43	-7.05	-6.12	-4.55	-5.92
	P-value	2.8E-23	0.67	2.8E-12	9.9E-09	1.8E-05	2.7E-08
Snow cover, $E_{SC,summer}$ $[-]$	t-Stat	4.09	-2.86	1.53	0.42	-1.93	-1.43
	P-value	2.3E-05	0.004	0.13	0.68	0.06	0.16
Mass balances, $E_{MB,norm}$ $[-]$	t-Stat	-5.35	9.44	5.80	2.49	0.24	3.03
	P-value	1.0E-07	1.3E-20	8.2E-09	0.01	0.81	0.003