Nat. Hazards Earth Syst. Sci., 20, 3521–3549, 2020
https://doi.org/10.5194/nhess-20-3521-2020

Natural Hazards
and Earth System
Sciences
Open Access

Downsizing parameter ensembles for simulations of rare floods

Anna E. Sikorska-Senoner[1], **Bettina Schaefli**[2,3,4], **and Jan Seibert**[1,5]

[1]University of Zurich, Department of Geography, Zurich, Switzerland
[2]University of Lausanne, Institute of Earth Surface Dynamics, Lausanne, Switzerland
[3]University of Bern, Institute of Geography, Bern, Switzerland
[4]Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland
[5]Swedish University of Agricultural Sciences, Department of Aquatic Sciences and Assessment, Uppsala, Sweden

Correspondence: Anna E. Sikorska-Senoner (anna.senoner@geo.uzh.ch)

Received: 12 March 2020 – Discussion started: 26 March 2020
Revised: 8 September 2020 – Accepted: 10 November 2020 – Published: 17 December 2020

Abstract. For extreme-flood estimation, simulation-based approaches represent an interesting alternative to purely statistical approaches, particularly if hydrograph shapes are required. Such simulation-based methods are adapted within continuous simulation frameworks that rely on statistical analyses of continuous streamflow time series derived from a hydrological model fed with long precipitation time series. These frameworks are, however, affected by high computational demands, particularly if floods with return periods > 1000 years are of interest or if modelling uncertainty due to different sources (meteorological input or hydrological model) is to be quantified. Here, we propose three methods for reducing the computational requirements for the hydrological simulations for extreme-flood estimation so that long streamflow time series can be analysed at a reduced computational cost. These methods rely on simulation of annual maxima and on analysing their simulated range to downsize the hydrological parameter ensemble to a small number suitable for continuous simulation frameworks. The methods are tested in a Swiss catchment with 10 000 years of synthetic streamflow data simulated thanks to a weather generator. Our results demonstrate the reliability of the proposed downsizing methods for robust simulations of rare floods with uncertainty. The methods are readily transferable to other situations where ensemble simulations are needed.

1 Introduction

The quantification of extreme floods and associated return periods remains a key issue for flood hazard management (Kochanek et al., 2014). Extreme-value analysis was largely developed in this field for the estimation of flood return periods (Katz et al., 2002); corresponding methods have been recently extended to bivariate approaches that assign probabilities jointly to flood peaks and flood volumes (Favre et al., 2004; De Michele et al., 2005; Brunner et al., 2016) and to trivariate approaches to assign probabilities jointly to flood peaks, volume and duration (Zhang and Singh Vijay, 2007); for a review of this field, see the work of Graler et al. (2013).

Most modern applications, however, require the estimation of not only extreme peak flow, associated flood volumes and duration but also of hydrograph shapes, in particular in the context of reservoir design or for safety checks of hydraulic infrastructure (Kochanek et al., 2014; Gaál et al., 2015; Zeimetz et al., 2018). The key is thus the construction of design hydrographs with different shapes, peak flows and volumes, with a corresponding probability of occurrence. Such approaches can be roughly classified into methods that identify the shape of these design hydrographs based on observed data (Mediero et al., 2010) or based on theoretical considerations (unit hydrographs) (Brunner et al., 2017) and regionalization (Tung et al., 1997; Brunner et al., 2018a) or methods that rely on streamflow simulations (Arnaud and Lavabre, 2002; Kuchment and Gelfan, 2011; Paquet et al., 2013).

Simulation-based methods for design or extreme-flood estimation have a long history in hydrology (for a review see Boughton and Droop, 2003) and started with the classical event-based simulation with selected design storms (Eagleson, 1972; Chow et al., 1988; American Society of Civil Engineers, 1996). Those event-based methods are based on the

Published by Copernicus Publications on behalf of the European Geosciences Union.

concept that the design storm and flood have the same return period. Moreover, as they usually do not simulate antecedent conditions prior to the event and do not account explicitly for storm patterns (duration, spatial and temporal variability), they may yield biased flood frequency distributions (Viglione and Blöschl, 2009; Grimaldi et al., 2012a). Although some modern extensions of this event-based concept account for variable initial conditions prior to the event through sensitivity tests (Filipova et al., 2019), most of the work using event-based simulations assume default initial conditions. Indeed, such event-based simulation is still in use, in particular in the context of probable maximum flood (PMF) estimation based on probable maximum precipitation (PMP) (Beauchamp et al., 2013; Gangrade et al., 2019).

Modern extensions of this approach, however, use continuous hydrological modelling for design flood estimation to generate either (i) a range of initial conditions for use in combination with design or randomly drawn storms (Paquet et al., 2013; Zeimetz et al., 2018) or (ii) long discharge time series from long observed-precipitation records or from synthetic precipitation time series obtained with a weather generator (Calver and Lamb, 1995; Cameron et al., 2000; Blazkova and Beven, 2004; Hoes and Nelen, 2005; Winter et al., 2019). The above approach (ii) is computationally intensive, especially if long time series are to be simulated using ensembles of hydrological parameter sets or if very high return periods (> 1000 years) have to be robustly estimated. But in exchange, return period analysis is straightforward for simulated peak flows or volumes. Full hydrographs for risk analysis are then obtained by either selecting a range of simulated extreme events or by scaling up an estimated synthetic design hydrograph by quantiles of extreme peak and volume estimated using frequency analysis (Pramanik et al., 2010; Serinaldi and Grimaldi, 2011).

These fully continuous simulation schemes are particularly useful for studies where recorded discharge time series are too short for extreme-flood analysis (Lamb et al., 2016; Evin et al., 2018). Although such an approach is based entirely on a continuous hydrological simulation, it is noteworthy that such a fully continuous approach might still be considered to be "semi-continuous" from a hydraulic perspective since corresponding studies often lack the final hydraulic routing step along the floodway (Grimaldi et al., 2013). For clarity, we therefore use the term "continuous hydrological simulation scheme" to distinguish it from the abovementioned hydraulic approach. These continuous hydrological simulation frameworks are still rare for time series ≥ 100 years due to heavy computational requirements (Grimaldi et al., 2013). An example is the work of Arnaud and Lavabre (2002), who use a continuous simulation framework to generate an ensemble of possible extreme hydrographs, which are then used as individual scenarios for hazard management. Another option is to summarize all simulated flood hydrographs into probability distributions for peak flow and flood volume (Gabriel-Martin et al., 2019).

High computational power is particularly needed in order to provide estimations for high to extreme return periods (up to 1000 years and higher) required for safety-related studies or for hydrological-hazard management. For such rare events, the large number of simulations in fully continuous frameworks can easily become prohibitive, in particular if the framework should also account for different sources of modelling uncertainty, such as input uncertainty (different weather generators) or the uncertainty in the hydrological model itself, which is often incorporated into the model parameter sets (using distribution of model parameters rather than a single best set) (Cameron et al., 1999). Using multiple parameter sets for a hydrological model is justified by the parameter equifinality (Beven and Freer, 2001; Sikorska and Seibert, 2018b). It has also been found that the model parameter uncertainty comprises important uncertainty sources in design floods that are based upon hydrological simulations (Brunner and Sikorska-Senoner, 2019). Other important uncertainty sources in hydrological modelling are linked to the calibration (discharge) data, input forcing (precipitation, temperature, evaporation) data and model structure (Sikorska and Renard, 2017; Westerberg et al., 2020).

Studies dealing with modelling or data uncertainties in such continuous simulation frameworks are rare as most previous studies have focused on the uncertainty related to the hydrological-model parameters only (e.g. Blazkova and Beven, 2002, 2004; Cameron et al., 1999). In addition to the uncertainties from seven hydrological-model parameters, Arnaud et al. (2017) investigated how the uncertainty related to six rainfall generator parameters propagates through the simulation framework using more than 1000 French basins with hydrological observation series of 40 years (median over all basins) and several hundreds of replicates. In their study they found that the uncertainty in the rainfall generator dominates the uncertainty in the simulated extreme-flood quantiles. With the exception of the work of Arnaud et al. (2017) using a simplified hydrological model, studies that deal with meteorological- and hydrological-modelling uncertainty in fully continuous simulation frameworks are currently missing. This is despite the fact that recent improvements in computational power with cluster and cloud computing theoretically open up the unlimited possibility of analysing different combinations of meteorological scenarios and parameter sets of a hydrological model within such ensemble-based simulation frameworks. Yet, computational constraints of hydrological models, especially at a high temporal resolution (subdaily or hourly), and data storage still remain bounding factors for simulation of long time series or for simulation of extreme floods with high return periods (up to 10 000 years).

Accordingly, for settings where full hydrological–hydraulic models are used for continuous simulation, some pre-selection of hydro-meteorological scenarios is often needed, particularly for computationally demanding complex hydrological or hydraulic models. How this selection should be completed, i.e. based on which quantitative criteria, remains unclear. The meteorological scenarios have the particularity that all scenarios generated with the same weather generator present different but equally likely realizations of the assumed climate condition; in other words, they represent the natural variability in the climate. Reducing the number of meteorological-input scenarios is not possible without simulating them with a hydrological model first as long as the continuous simulation scheme is of interest, i.e. if full time series are to be analysed without the possibility of extracting single events. This is due to the non-linear response of any hydrological model to meteorological input (scenario), which translates into hydrological scenarios with different statistical properties, albeit resulting from an ensemble of input scenarios having the same statistical properties.

We are therefore essentially left with finding ways to reduce at least the computational requirements associated with hydrological-model parameter uncertainty, apart from reducing the length of time series, which for analysis of extremes, is an unattractive option. Accordingly, in this work, we propose an assessment of different data-based methods to select a reduced-size ensemble of hydrological-model parameters for the use within a continuous simulation, ensemble-based hydro-meteorological framework. Our specific research questions are as follows. (1) How can we downsize (reduce) the hydrological-model parameter ensembles for simulation of rare floods so that the variability and the range of the full ensemble is preserved as closely as possible? (2) Can such a reduced hydrological-model parameter ensemble be assumed to be reliable for the simulation of rare floods during the reference period (used for parameter ensemble downsizing) as well as during an independent validation period? (3) Which metrics would be suitable to assess the performance of such a reduced hydrological-model parameter ensemble against the reference (full) ensemble? Specifically, three different methods of reducing a full hydrological-modelling parameter ensemble to a handful of parameter sets are proposed and tested for deriving the uncertainty ranges of simulated rare flood events (up to 10 000 years return period). All three methods rely on simulation of annual maxima and are tested on continuous synthetic data (simulated with a hydrological model) of 10 000 years. Using synthetic instead of observed data is important here as only recently Brunner et al. (2018b) have shown that the record length is one of the most important sources of uncertainty in design floods. Hence, using a simulation setting with synthetic data as a start for our analysis enables us (i) to provide long enough simulation periods for rare-flood analysis with return periods ≥ 100 years and (ii) to be able to focus entirely on the uncertainty in the hydrological response, while other uncertainty sources of a hydrological model (due to model calibration) are not explicitly considered. Note that way the hydrological model is calibrated lies outside of the scope of this paper. The principal idea underlying these selection methods is that the downsizing of the ensemble of hydrological-model parameters may be performed with a reduced length of input time series that is much shorter than the full simulation time frame and that then can be applied to the full time window for analysis of rare floods (up to return periods of 1000 years or more).

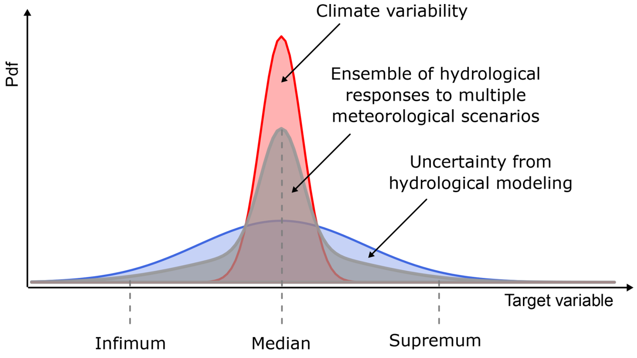
Climate variability
Pdf
Ensemble of hydrological responses to multiple meteorological scenarios
Uncertainty from hydrological modeling
Target variable
Infimum
Median
Supremum

Figure 1. Framework overview. The infimum and supremum refer to the largest interval bounding the ensemble simulation from below and the smallest interval bounding it from above.

2 Methods

2.1 Study framework and objectives

The focus of this study is a fully continuous hydro-meteorological ensemble-based simulation framework for estimation of rare floods. The underlying streamflow time series ensemble is built based on meteorological scenarios and multiple hydrological-model runs using a number of calibrated model parameter sets. A meteorological scenario represents a single realization from a stochastic weather generator with constant model parameters. These meteorological scenarios are equally likely model realizations that differ in the precipitation and temperature patterns, and together they represent the natural variability in the climate (and not the model uncertainty in a weather generator). These realizations are then used as inputs into a hydrological model to simulate the hydrological response. To account for hydrological-modelling uncertainty, a range of different hydrological-model parameter sets is used for each meteorological scenario. These two sources of hydrological variability then accumulate along the modelling chain and can be represented as an ensemble of possible hydrological responses (Fig. 1).

Within such a defined framework we first want to understand how variable the hydrological response simulation is and, second, develop methods to downsize the hydrological-model parameter ensemble to a smaller subset that could be dealt with within such a modelling chain for rare-flood simulations. This subset should represent the entire range of variability in the hydrological response but with little computational effort and should also be transferable to independent time periods. Hereafter, we call this subset the *representative parameter ensemble*.

Downsizing of the ensemble of hydrological-model parameters is particularly needed if (i) the probability distribution of the parameter sets is unknown because parameter sets result from independent calibrations or regionalization approaches, and only a limited number of sets can be run with the hydrological model, or (ii) the distribution is known (i.e. estimated from data), but due to time-consuming simulations it is not possible to run the hydrological model for a full ensemble of multiple meteorological scenarios.

The question of how many parameter sets are needed to cover most of the simulation range is important. However, here we set this value to a constant number and rather test different selection approaches. Hence, for the purpose of our work, we furthermore would like this *representative parameter ensemble* to be composed of only three sets, which should be representative of a lower (infimum), a middle (median) and an upper (supremum) interval of the full hydrological ensemble (Fig. 1). These intervals, together, should enable the construction of predictive intervals for rare-flood estimates that represent the full variability range of all ensemble members. The infimum (from the Latin – smallest) and supremum (from the Latin – largest) refer to the greatest lower bound and the least upper bound (Hazewinkel, 1994), i.e. the largest interval bounding the ensemble from below, here 5 %, and the smallest interval bounding it from above, here 95 %. Thus, the representative band should correspond to 90% predictive bands of a target variable. The choice of infimum and supremum is favourable over the maximum and minimum as the latter would imply a complete hydrological-model parameter ensemble range, whereas here we use the terms to describe the range of a certain ensemble.

The key challenge for such a downsizing is the fact that we would like to select hydrological-model parameter sets (i.e. select in the parameter space) but based on how representative the corresponding simulations are in the model response space. Moreover, the downsized ensemble should not only be representative of simulated time periods but also be transferable to independent time periods. The first question to answer is which model response space the selection should focus on. In the context of rare-flood estimation, focusing on the frequency distribution of annual maxima (AMs) is a natural choice; we thus propose to use the representation of AMs sorted by their magnitudes (i.e. frequency space) as the *reference model response space* for parameter selection.

The next step is the development of selection methods to select hydrological-model parameter sets that plot into certain locations in the model response space. Given the nonlinear relationship between model parameters and hydrological responses, this selection has to be obtained via a post-modelling approach; i.e. we have to first simulate all parameter sets and then decide which parameter sets fulfil certain selection criteria in the model response space.

For that purpose, we developed three methods, which are based on (a) ranking, (b) quantiling, and (c) clustering, described in detail in Sect. 2.2. The main idea behind all three methods is that the hydrological-parameter set selection is made based on the full ensemble with all hydrological-model simulations but using only a limited simulation period that is much shorter than the time window of full meteorological scenarios used within the simulation framework for which rare floods are to be estimated.

Next, for the purpose of this study, let us define the following variables:

- I is a number of hydrological-model parameter sets available, with $i = 1, 2, \ldots$ being a parameter set index.
- $\boldsymbol{\theta}_i$ is the ith parameter set of a hydrological model.
- J is a number of annual maxima (years) per hydrological simulation; $y = 1, 2, \ldots$ is a year of simulation (index of unsorted annual maxima); and $j = 1, 2, \ldots$ is an index of sorted annual maxima.
- X_j is the jth sorted annual maximum, and X_y is the unsorted annual maximum from the year y.
- M is a number of meteorological scenarios considered, with $m = 1, 2, \ldots$ being a meteorological scenario index.
- $\boldsymbol{S}_\mathrm{m}$ is the mth meteorological scenario.
- $H(\boldsymbol{\theta}_i|\boldsymbol{S}_\mathrm{m})$ is the hydrological simulation computed using the ith parameter set of a hydrological model and the mth meteorological scenario.
- $X_{y,i,\mathrm{m}}$ is the annual maximum for the year y extracted from $H(\boldsymbol{\theta}_i|\boldsymbol{S}_\mathrm{m})$.
- $\boldsymbol{\theta}_\mathrm{inf}$, $\boldsymbol{\theta}_\mathrm{med}$ and $\boldsymbol{\theta}_\mathrm{sup}$ are the representative parameter sets of the hydrological model, i.e. infimum, median and supremum that correspond to the intervals named in the same way.

2.2 Developed methods for selecting the representative parameter sets

For the sake of simplicity, let us choose a single meteorological scenario $\boldsymbol{S}_\mathrm{m}$ for now. Using $\boldsymbol{S}_\mathrm{m}$ as an input into a hydrological model combined with I parameter sets results in an ensemble of hydrological simulations, $H(\boldsymbol{\theta}_{1,2,\ldots}|\boldsymbol{S}_\mathrm{m})$. Now, the goal is to select a limited number (here three) of hydrological-model parameter sets, i.e. $\boldsymbol{\theta}_\mathrm{inf}$, $\boldsymbol{\theta}_\mathrm{med}$ and $\boldsymbol{\theta}_\mathrm{sup}$, from the available pool of I sets ($I \gg 3$) based on the simulation of annual maxima (AMs). These AMs are extracted from time series with continuous hydrological simulations, i.e. $H(\boldsymbol{\theta}_{1,2,\ldots}|\boldsymbol{S}_\mathrm{m})$, using a maximum approach that guarantees that the highest peak flow within each calendar year for each hydrological simulation is selected (Fig. 2). This assumption is made to cover the situation when different model realizations (i.e. for $i = 1, 2, \ldots$) lead to different flood events being classified as the largest event within the year. In

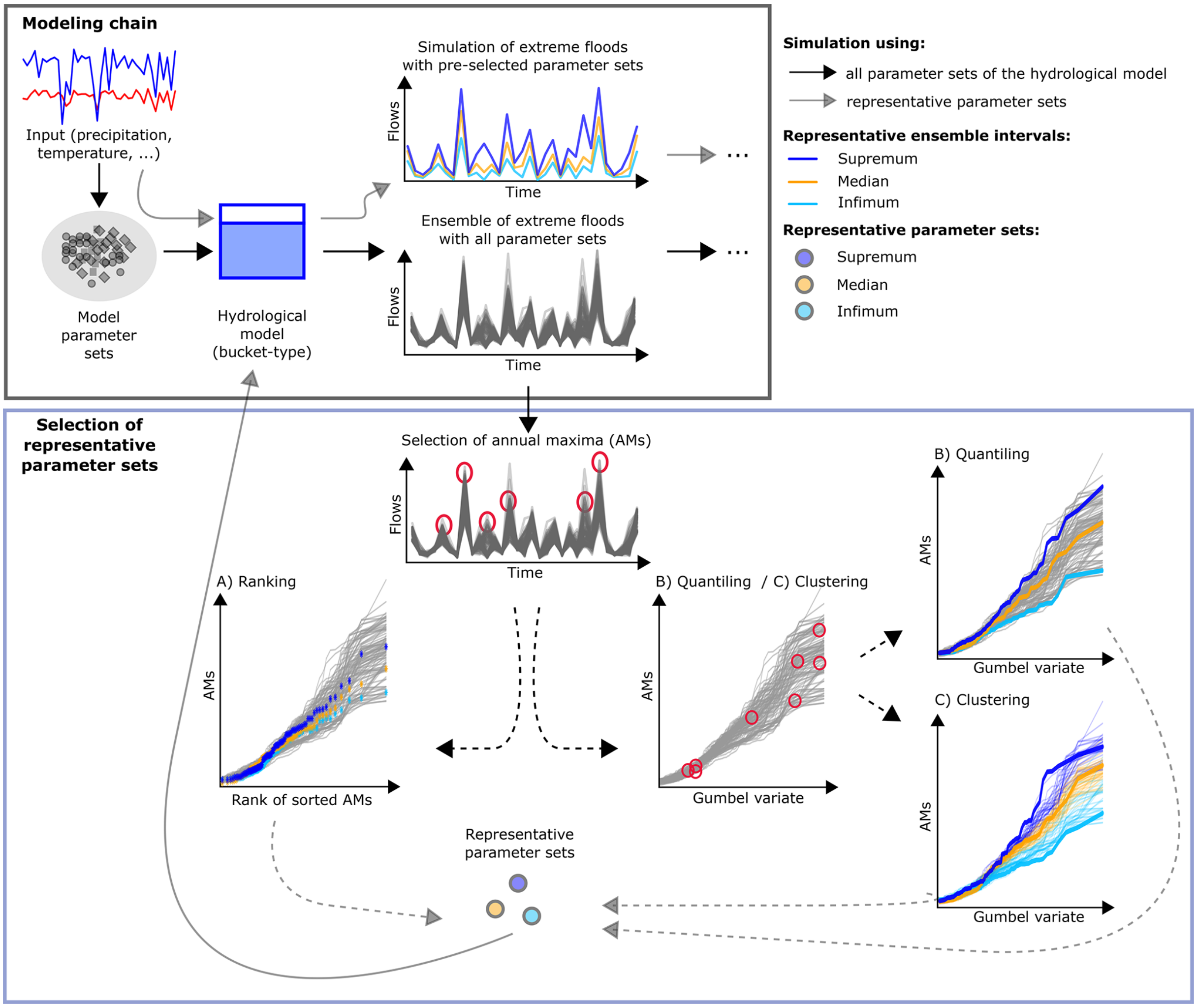
Modeling chain
Input (precipitation, temperature, ...)
Model parameter sets
Hydrological model (bucket-type)
Simulation of extreme floods with pre-selected parameter sets
Flows
Time
Ensemble of extreme floods with all parameter sets
Flows
Time
Simulation using:
all parameter sets of the hydrological model
representative parameter sets
Representative ensemble intervals:
Supremum
Median
Infimum
Representative parameter sets:
Supremum
Median
Infimum
Selection of representative parameter sets
Selection of annual maxima (AMs)
Flows
Time
A) Ranking
AMs
Rank of sorted AMs
B) Quantiling / C) Clustering
AMs
Gumbel variate
B) Quantiling
AMs
Gumbel variate
C) Clustering
AMs
Gumbel variate
Representative parameter sets

Figure 2. Overview of the modelling chain and the selection methods of the representative parameter sets; **(a)** delivery of hydrological-simulation ensembles and ensemble ranges; **(b)** three methods (A–C) proposed for selecting the representative parameter sets based on annual maxima (AMs) marked with red circles.

this case, we ensure that the largest flood event simulated within each yth year and each ith parameter set is selected. This means however that AMs selected for the same year y but with a different hydrological-model parameter set may originate from different flood events and even from a different dominant flood process, e.g. heavy rainfall or intensive snowmelt (Merz and Blöschl, 2003; Sikorska et al., 2015). This could be the case when one hydrological-model parameter set better represents processes driven by the rainfall excess, while others better represent processes driven by the snowmelt dynamics. For simplicity, we do not distinguish events by their different flood genesis and pool all AMs together.

Using the above notations, the selection of representative parameter sets can be summarized as follows.

1. Simulation of continuous streamflow times series: the hydrological model is run with all available I parameter sets of a hydrological model over the simulation period. This gives I different hydrological realizations (simulation ensemble members) covering the same time span.

2. Selection of annual maxima (AMs): for each ith hydrological realization, annual maxima are selected as the highest peak flow within each yth simulation year. This results in a J set of AMs per i hydrological simulation. The selection of AMs is repeated for all I hydrological simulations.

3. Selection of three representative parameter sets based on the simulation of AMs and following on from the three methods detailed below.

2.2.1 Ranking

a. AMs computed from I hydrological simulations (i.e. using I hydrological-model parameter sets) are sorted by their magnitude from the highest to the lowest within each yth simulation year independently (Fig. 2a).

b. For each yth simulation year, AMs which correspond to the 5th, 50th and 95th rank for that year are selected.

c. Parameter sets that correspond to the selected AM ranks are then attributed as 5th, 50th and 95th parameter sets per yth year independently.

d. The parameter sets selected in step (c) are compared over all J simulation years, and the sets which are chosen most often as the 5th, 50th and 95th ranks are retained as the parameter sets $\boldsymbol{\theta}_{\mathrm{R}5}$, $\boldsymbol{\theta}_{\mathrm{R}50}$ and $\boldsymbol{\theta}_{\mathrm{R}95}$ representative of the entire simulation period and for the entire hydrological-simulation ensemble.

2.2.2 Quantiling

a. For each ith hydrological-model parameter set, AMs computed with this parameter set are sorted by their magnitude over the entire simulation period (J years), thus creating the ensemble of sorted AMs simulated with different parameter sets.

b. The 5 %, 50 % and 95 % quantiles of these ensembles are computed at each jth point in the frequency space, resulting in quantiles $\boldsymbol{Q}_5$, $\boldsymbol{Q}_{50}$ and $\boldsymbol{Q}_{95}$ over the entire simulation period (Fig. 2b).

c. Next, for each ith ensemble member, a metric R_{MSE} is computed such that for each jth point of the ith ensemble member it measures distances from $\boldsymbol{Q}_5$, $\boldsymbol{Q}_{50}$ and $\boldsymbol{Q}_{95}$. This metric is somehow similar to the mean square error and is computed for $\boldsymbol{Q}_{50}$ as

$$R_{\mathrm{MSE},Q_{50},i} = \frac{1}{J}\sum_{j=1}^{J}\left(\boldsymbol{Q}_{50,j} - H_j\left(\boldsymbol{\theta}_i|\boldsymbol{S}_m\right)\right)^2 \qquad (1)$$

and in the same way for $\boldsymbol{Q}_5$ and $\boldsymbol{Q}_{95}$.

d. Finally, the ensemble members which lie closest to $\boldsymbol{Q}_5$, $\boldsymbol{Q}_{50}$ and $\boldsymbol{Q}_{95}$, i.e. that received the smallest values for R_{MSE,Q_5}, $R_{\mathrm{MSE},Q_{50}}$ and $R_{\mathrm{MSE},Q_{95}}$, respectively, are chosen as the ensemble members representative of the entire hydrological ensemble, and the parameter sets corresponding to these members, i.e. $\boldsymbol{\theta}_{Q_5}$, $\boldsymbol{\theta}_{Q_{50}}$ and $\boldsymbol{\theta}_{Q_{95}}$, are retained as representative.

2.2.3 Clustering

a. Similar to the quantiling method, for each ith parameter set, AMs computed with this parameter set are sorted by their magnitude over the entire simulation period, creating I ensemble members of sorted AMs simulated with different parameter sets.

b. These members are next clustered into three representative groups (clusters) based on all J simulation years using the k-means clustering with the Hartigan–Wong algorithm (Hartigan and Wong, 1979), as implemented in the function kmeans from the package "stats" (R Core Team, 2019); see Fig. 2c.

c. Next, these clusters are sorted based on cluster means by their magnitude by comparing percentiles in the upper tail of the distribution (here we used a 90th percentile). Use of a percentile from the upper tail is important as methods are focusing on rare floods. However, we found that the method was insensitive to the percentile choice as long as it lies in the upper tail (i.e. $\geq$ 80th percentile). Based on the percentiles computed for each cluster mean, the lower, middle and upper clusters are defined. Next, for the lower cluster a 5th percentile, for the upper a 95th percentile and for the middle a 50th percentile are computed, i.e. $\boldsymbol{P}_5$, $\boldsymbol{P}_{50}$ and $\boldsymbol{P}_{95}$. Note that we use here percentiles instead of cluster means to make this method comparable with the other two methods and to better cover the variability in the hydrological-model parameter sample. Use of the 5th and 95th percentiles appears to be a fair choice for asymmetrically spread clusters, which is most often the case as different parameter sets of a hydrological model may emphasize different hydrological processes in the catchment.

d. For each ith ensemble member, the metric R_{MSE} is computed in relation to three estimated cluster percentiles as, e.g. for $\boldsymbol{Q}_{50}$,

$$R_{\mathrm{MSE},P_{50},i} = \frac{1}{J}\sum_{j=1}^{J}\left(\boldsymbol{P}_{50,j} - H_j\left(\boldsymbol{\theta}_i|\boldsymbol{S}_m\right)\right)^2 \qquad (2)$$

and in the same way for $\boldsymbol{P}_5$ and $\boldsymbol{P}_{95}$.

e. For each of these three clusters, the ensemble member that lies closest to the cluster percentile, i.e. received the smallest value of R_{MSE}, is selected as the representative member for that cluster, and the parameter sets which correspond to these members, $\boldsymbol{\theta}_{P_5}$, $\boldsymbol{\theta}_{P_{50}}$ and $\boldsymbol{\theta}_{P_{95}}$, are retained as representative.

For visualizing the selection methods, we use the Gumbel space (generalized extreme-value distribution Type I) with the Gringorten's method (Gringorten, 1963) to compute the plotting positions of AMs in the Gumbel plots:

$$k_j = \frac{j - 0.44}{J + 0.12}, \qquad (3)$$

where k_j is a plotting position for the jth (sorted) AM in the Gumbel space.

Table 1. Comparison of three methods for selecting representative parameter sets based on annual maxima (AMs).

Criteria	Ranking	Quantiling	Clustering
Selection window	Year	All simulation years	All simulation years
Annual maxima (AMs)	Unsorted over years	Sorted over years	Sorted over years
Sorting space	Simulated AMs	AMs frequency, quantiling	AMs frequency, clustering
Sorting extent	AMs over simulations	AMs over years	AMs over years
Selection criteria	Ranks	R_{MSE}	R_{MSE}
Interpretation of pred. intervals	No	Yes	Yes
Parameter grouping	No	No	Yes

2.3 Estimation of the predictive intervals for rare-flood simulations

While the three methods described in Sect. 2.2 vary in the way the representative parameter sets are selected (see Sect. 2.4 for a summary), each of these selection methods results in three (different) representative hydrological-simulation ensemble members and can be thought of as representing the lower (infimum), upper (median) and middle (supremum) interval of the full simulation range. The hydrological-model parameter sets corresponding to these are then noted as $\boldsymbol{\theta}_{\mathrm{inf}}$, $\boldsymbol{\theta}_{\mathrm{med}}$ and $\boldsymbol{\theta}_{\mathrm{sup}}$. The simulations corresponding to these three parameter sets together create the so-called predictive interval, which can be used for rare-flood simulation studies. Here, these predictive intervals constructed based on representative parameter sets correspond to 90 % predictive intervals (PIs).

2.4 Comparison of three selection methods

The major difference between these three methods is that the ranking method is evaluated based on individual simulation years using simple ranking of flow maxima independently of their frequency; i.e. it works on unsorted annual maxima. Note that in this way, for each y simulation year, a different rank order of the I hydrological-model parameter sets may be achieved. In an extreme case, where for each year different parameter sets are chosen, a choice of the representative sets over all simulation years may become problematic due to difficulties in identifying the parameter sets most frequently selected over all simulation years. The derived predictive intervals thus are sensitive to individual years of simulations, and their interpretation may be difficult (as they do not result from any flow frequency analysis).

In contrast to the ranking method, both other methods, i.e. quantiling and clustering, are performed on sorted AMs over all simulation years, i.e. in the flow frequency space. This enables statistical statements to be made about the selected parameter sets and about the predictive intervals constructed with the help of these parameter sets (as they are constructed on the entire simulation ensemble). Furthermore, selected parameter sets can be assumed to be representative over the entire simulation period (see Table 1 for a detailed overview of three methods). Finally, the clustering method splits all ensemble members (hydrological simulations) into three clusters, and so each parameter set can be attributed to corresponding clusters. This could be useful if one would like to extract more information on each cluster behaviour.

2.5 Assessment of the behaviour of the approach

Testing the methods for a time period different than the one that was used for the parameter ensemble downsizing is crucial for assessing how well the reduced ensembles substitute the whole simulation ensemble for the selection of representative parameter sets. Thus, we propose to assess the behaviour of the developed approach by repeating the selection of the three representative parameter sets with the three proposed methods with multiple (M) meteorological scenarios. Using multiple meteorological scenarios first enables us to account for the natural variability in the hydrological response due to climate variability and, second, gives us the possibility to evaluate the bias of the approach. Particularly, with the help of multiple meteorological scenarios we explore how the choice of the representative parameter sets $\boldsymbol{\theta}_{\mathrm{inf}}$, $\boldsymbol{\theta}_{\mathrm{med}}$ and $\boldsymbol{\theta}_{\mathrm{sup}}$ depends on the meteorological scenario.

2.5.1 Leave-one-out cross-validation

To evaluate the three selection methods, we perform a leave-one-out cross-validation simulation study, in which a meteorological scenario S_{r} is removed from the analysis, and the selection of the representative parameter sets is executed based on all other remaining meteorological scenarios, i.e. using all $m = 1, 2, \dots M$ and $m \neq r$. The evaluation of selection methods is then executed against the one meteorological scenario initially removed from the set. In detail, the following steps are executed for each of the three methods independently:

a. Pick up and remove one meteorological scenario $\boldsymbol{S}_{\mathrm{r}}$ from $\boldsymbol{S}_{1,2,\dots\mathrm{M}}$ scenarios available.

b. Analyse all other meteorological scenarios $\{\boldsymbol{S}_{\mathrm{M-r}}\} = \{\boldsymbol{S}_{1,2,\dots\mathrm{M}}\} \setminus \{\boldsymbol{S}_{\mathrm{r}}\}$, each containing I ensemble members resulting from I hydrological-model parameter sets, $\{H(\boldsymbol{\theta}_i|\boldsymbol{S}_{\mathrm{m-r}})\}$, for $i = 1,2,\dots I$, $m = 1,2,\dots M$

and $m \neq r$ and based on the selected three representative parameter sets $\boldsymbol{\theta}_{\mathrm{inf,m-r}}$, $\boldsymbol{\theta}_{\mathrm{med,m-r}}$ and $\boldsymbol{\theta}_{\mathrm{sup,m-r}}$ as described in Sect. 2.2.

c. Estimate the predictive intervals of these $\boldsymbol{S}_{\mathrm{M-r}}$ meteorological scenarios as the band spread between $H(\boldsymbol{\theta}_{\mathrm{inf,m-r}}|\boldsymbol{S}_{\mathrm{m-r}})$ and $H(\boldsymbol{\theta}_{\mathrm{sup,m-r}}|\boldsymbol{S}_{\mathrm{m-r}})$, the interval defined in step (b).

d. Evaluate the meteorological scenario $\boldsymbol{S}_{\mathrm{r}}$ removed at step (a) against the predictive intervals of $\boldsymbol{S}_{\mathrm{M-r}}$ meteorological scenarios to assess how well the defined identified intervals represent the ensemble members of this $\boldsymbol{S}_{\mathrm{r}}$ meteorological scenario (see Sect. 2.6 for assessment criteria).

The simulation is repeated M times to use each meteorological scenario once. In other words, this test evaluates how well the selection methods applied to all but one scenario can predict the full simulation range of the left-out scenario.

2.5.2 Multi-scenario evaluation

To further evaluate the three methods, we perform a simulation study using multiple (M) meteorological scenarios. In this study, the three selection methods are executed on one meteorological scenario randomly (without replacement) selected from the M available scenarios and evaluated against all remaining scenarios. In detail, the following steps are executed for each of the three methods independently:

a. Pick up one meteorological scenario $\boldsymbol{S}_{\mathrm{p}}$ out of the $\boldsymbol{S}_{1,2,\ldots\mathrm{M}}$ scenarios available.

b. Analyse the I simulated hydrological ensemble members of this scenario $H(\boldsymbol{\theta}_i|\boldsymbol{S}_{\mathrm{p}})$, $i = 1, 2, \ldots I$, resulting from I hydrological-model parameter sets $\boldsymbol{\theta}_i$ for $\boldsymbol{S}_{\mathrm{p}}$, and select three representative parameter sets corresponding to $\boldsymbol{\theta}_{\mathrm{inf,p}}$, $\boldsymbol{\theta}_{\mathrm{med,p}}$ and $\boldsymbol{\theta}_{\mathrm{sup,p}}$, as described in Sect. 2.2.

c. For all other remaining meteorological scenarios $\{\boldsymbol{S}_{\mathrm{M-p}}\} = \{\boldsymbol{S}_{1,2,\ldots\mathrm{M}}\} \setminus \{\boldsymbol{S}_{\mathrm{p}}\}$, take all hydrological ensemble members $\{H(\boldsymbol{\theta}_i|\boldsymbol{S}_{\mathrm{m}})\}$ for $m = 1, 2, \ldots M$ and $m \neq p$ that correspond to $\boldsymbol{\theta}_{\mathrm{inf,p}}$, $\boldsymbol{\theta}_{\mathrm{med,p}}$ and $\boldsymbol{\theta}_{\mathrm{sup,p}}$. This results in $M-1$ model simulations for $\boldsymbol{\theta}_{\mathrm{inf,p}}$, $\boldsymbol{\theta}_{\mathrm{med,p}}$ and $\boldsymbol{\theta}_{\mathrm{sup,p}}$, one per meteorological scenario.

d. Compute the 5th percentile for $\{H(\boldsymbol{\theta}_{\mathrm{inf,p}}|\boldsymbol{S}_{\mathrm{m}})\}$, the 50th for $\{H(\boldsymbol{\theta}_{\mathrm{med,p}}|\boldsymbol{S}_{\mathrm{m}})\}$ and the 95th for $\{H(\boldsymbol{\theta}_{\mathrm{sup,p}}|\boldsymbol{S}_{\mathrm{m}})\}$ for $m = 1, 2, \ldots M$ and $m \neq p$. The computed 5th and 95th percentiles together are assumed to describe the predictive intervals.

e. Evaluate the predictive intervals against all $\boldsymbol{S}_{\mathrm{M-p}}$ meteorological scenarios for assessing how well the identified prediction intervals represent the ensemble members of these $\boldsymbol{S}_{\mathrm{M-p}}$ scenarios (see Sect. 2.6).

The steps (a)–(e) are repeated M times to use each meteorological scenario once. We call this evaluation a multi-scenario evaluation because the evaluation is performed using multiple meteorological scenarios at once ($\boldsymbol{S}_{\mathrm{M-p}}$) in contrast to the leave-one-out cross-validation (Sect. 2.5.1), where the evaluation is performed against only one meteorological scenario ($\boldsymbol{S}_{\mathrm{r}}$). This test quantifies how well the methods applied to a single scenario are transferable to all other scenarios.

2.6 Evaluation criteria

2.6.1 Visual assessment

The simplest way of assessing the behaviour of these three methods is a visual inspection of curves plotted in the frequency space (e.g. using Gumbel distribution for plotting), which can tell us how well the selected members reproduce the simulation ensemble and particularly whether the assignment of the representative parameter sets is correct or not. For this purpose, we propose to plot all simulated hydrological ensemble members together with the selected representative members in the frequency space for each considered meteorological scenario m individually and to visually assess the assignment of the three selected parameter sets, $\boldsymbol{\theta}_{\mathrm{inf,m}}$, $\boldsymbol{\theta}_{\mathrm{med,m}}$ and $\boldsymbol{\theta}_{\mathrm{sup,m}}$, and the corresponding intervals, i.e. $H(\boldsymbol{\theta}_{\mathrm{inf,m}}|\boldsymbol{S}_{\mathrm{m}})$, $H(\boldsymbol{\theta}_{\mathrm{med,m}}|\boldsymbol{S}_{\mathrm{m}})$ and $H(\boldsymbol{\theta}_{\mathrm{sup,m}}|\boldsymbol{S}_{\mathrm{m}})$. The order of the intervals' assignment is assumed to be correct if it holds in the frequency space that

$$H\left(\boldsymbol{\theta}_{\mathrm{inf,m}}|\boldsymbol{S}_{\mathrm{m}}\right) \leq H\left(\boldsymbol{\theta}_{\mathrm{med,m}}|\boldsymbol{S}_{\mathrm{m}}\right) \leq H\left(\boldsymbol{\theta}_{\mathrm{sup,m}}|\boldsymbol{S}_{\mathrm{m}}\right). \tag{4}$$

We further define a ratio of incorrectly attributed scenarios with mixed-up intervals, i.e. for which Eq. (4) does not hold, as a measure of the bias as

$$R_{\mathrm{bias}} = \sum_{m=1}^{\mathrm{M}} \frac{R_{\mathrm{m}}}{M}, \tag{5}$$

where R_{m} is computed for the mth scenario as

$$R_{\mathrm{m}} = \left\{ \begin{array}{ll} 0 & \text{if } H\left(\boldsymbol{\theta}_{\mathrm{inf,m}}|\boldsymbol{S}_{\mathrm{m}}\right) < H\left(\boldsymbol{\theta}_{\mathrm{med,m}}|\boldsymbol{S}_{\mathrm{m}}\right) < H\left(\boldsymbol{\theta}_{\mathrm{sup,m}}|\boldsymbol{S}_{\mathrm{m}}\right) \\ 1 & \text{else} \end{array} \right. . \tag{6}$$

2.6.2 Quantitative assessment

To quantitatively compare the three selection methods, we propose to compute the five following metrics:

I. The ratio of simulation points in the frequency space, i.e. sorted annual maxima, lying outside the predictive intervals is computed for each mth scenario as

$$R_{\mathrm{spo,m}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{R_{\mathrm{spo,m},i,j}}{I \cdot J}, \tag{7}$$

where $R_{\mathrm{spo,m},i}$ is the ratio for each ith hydrological-model parameter set of the meteorological scenario m and is computed for each simulation point j (sorted annual maximum) as

$$R_{\mathrm{spo,m},i,j} = \begin{cases} 0 & \text{if } H_j\left(\boldsymbol{\theta}_{\mathrm{inf,m}}|\boldsymbol{S}_{\mathrm{m}}\right) \leq H_j\left(\boldsymbol{\theta}_i|\boldsymbol{S}_{\mathrm{m}}\right) \leq H_j\left(\boldsymbol{\theta}_{\mathrm{sup,m}}|\boldsymbol{S}_{\mathrm{m}}\right) \\ 1 & \text{else} \end{cases}. \tag{8}$$

II. In the leave-one-out cross-validation, the ratio of hydrological-simulation ensemble members lying outside the predictive intervals is computed for each mth scenario as

$$R_{\mathrm{hso,m}} = \sum_{i=1}^{I} \frac{R_{\mathrm{hso,m},i}}{I}, \tag{9}$$

where $R_{\mathrm{hso,m},i}$ is the ratio computed for each ith ensemble member as

$$R_{\mathrm{hso,m},i} = \begin{cases} 0 & \text{if } H\left(\boldsymbol{\theta}_{\mathrm{inf,m}}|\boldsymbol{S}_{\mathrm{m}}\right) \leq H\left(\boldsymbol{\theta}_i|\boldsymbol{S}_{\mathrm{m}}\right) \leq H\left(\boldsymbol{\theta}_{\mathrm{sup,m}}|\boldsymbol{S}_{\mathrm{m}}\right) \\ 1 & \text{else} \end{cases}. \tag{10}$$

III. In the multi-scenario evaluation, the ratio of meteorological scenarios lying outside the predictive intervals is computed for each scenario p as

$$R_{\mathrm{mso,p}} = \sum_{m}^{M} \frac{R_{\mathrm{mso,m}}}{M-1} \quad m = 1, 2, \ldots M \text{ and } m \neq p, \tag{11}$$

where $R_{\mathrm{mso,m}}$ is computed as

$$R_{\mathrm{mso,m}} = \begin{cases} 0 & \text{if } H\left(\boldsymbol{\theta}_{\mathrm{inf,m}}|\boldsymbol{S}_{\mathrm{m}}\right) \leq H\left(\boldsymbol{\theta}_i|\boldsymbol{S}_{\mathrm{m}}\right) \leq H\left(\boldsymbol{\theta}_{\mathrm{sup,m}}|\boldsymbol{S}_{\mathrm{m}}\right) \forall i = 1, 2, \ldots I \\ 1 & \text{else} \end{cases}. \tag{12}$$

IV. Relative band spread of PIs ($R_{\Delta\mathrm{PIs}}$) is computed for both tests and compares the spread of PIs constructed with the representative parameter sets versus 90 % PIs of the full hydrological ensemble. In detail, $R_{\Delta\mathrm{PIs}}$ is computed for each mth scenario as

$$R_{\Delta\mathrm{PIs,m}} = \sum_{j}^{J} \frac{S_{\mathrm{PIs,repr.,m}}}{S_{\mathrm{PIs,full,m}}} \quad m = 1, 2, \ldots M \\ \text{and } j = 1, 2, \ldots J, \tag{13}$$

where $S_{\mathrm{PIs,repr.,m}}$ and $S_{\mathrm{PIs,full,m}}$ are band spreads of the 90 % PIs constructed with the representative parameter sets and with the full hydrological ensemble. The band spread is computed as a difference between the upper (or supremum) and the lower (or infimum) interval at each jth simulation point in the frequency space.

V. Overlapping pools of PIs (R_{OPPIs}) are computed for both tests in the frequency space by taking the Gumbel variate and discharge values of sorted AMs as coordinates of the PI pools. In detail, R_{OPPIs} of PIs constructed with the representative parameter sets is computed for each mth scenario as

$$R_{\mathrm{OPPIs,m}} = \sum_{j}^{J} \frac{\left(k_j - k_{j-1}\right)}{2} \left(H\left(\boldsymbol{\theta}_{\mathrm{sup,m},j}\right) + H\left(\boldsymbol{\theta}_{\mathrm{sup,m},j-1}\right) - H\left(\boldsymbol{\theta}_{\mathrm{inf,m},j}\right) - H\left(\boldsymbol{\theta}_{\mathrm{inf,m},j-1}\right)\right) \quad m = 1, 2, \ldots M \\ \text{and } j = 2, 3, \ldots J. \tag{14}$$

In a similar way, R_{OPPIs} is computed for the full hydrological ensemble using the pool restricted by the 90 % PIs, i.e. taking the 5 % and 95 % intervals as pool borders.

With respect to R_{spo}, the question arises of how to define the ratio of simulation points outside the predictive intervals if multiple hydrological simulations (leave-one-out cross-validation) or multiple meteorological scenarios (multi-scenario evaluation) are considered. Here we propose to use the 50th percentile to characterize the ratio of the majority of simulation points lying outside the computed predictive intervals for each of the methods.

In a similar way, for R_{hso} and R_{mso} an additional condition must be defined, i.e. how many out of J hydrological-simulation points for R_{hso} or how many out of I hydrological-simulation ensemble members for R_{mso} must lie outside the defined predictive intervals so that the hydrological simulation $H(\boldsymbol{\theta}_i|\boldsymbol{S}_{\mathrm{m}})$ or the meteorological scenario $\boldsymbol{S}_{\mathrm{m}}$ is considered to be outside these intervals. For this purpose we define the rejection threshold r_{thr} (dimensionless) that has to be reached so that the meteorological scenario or hydrological simulation is assumed to be outside the predictive intervals. In this work, we consider the two following values for r_{thr}: {0.50, 0.10}.

With regards to $R_{\Delta\mathrm{PIs}}$, we propose to compute the relative band spread as a mean over all sorted AMs at first. Also, to focus more on rare floods, we propose to compute means of rare floods limited by different Gumbel variates. Here we computed $R_{\Delta\mathrm{PIs}}$ for the upper half of AMs ($R_{\Delta\mathrm{PIs},j\geq 51}$), for the uppermost 10 AMs ($R_{\Delta\mathrm{PIs},j\geq 91}$) and the uppermost 5 AMs ($R_{\Delta\mathrm{PIs},j\geq 96}$).

These five metrics are computed for all three methods and for all M meteorological scenarios, and the median values over these M scenarios are taken as a measure for comparing the three methods.

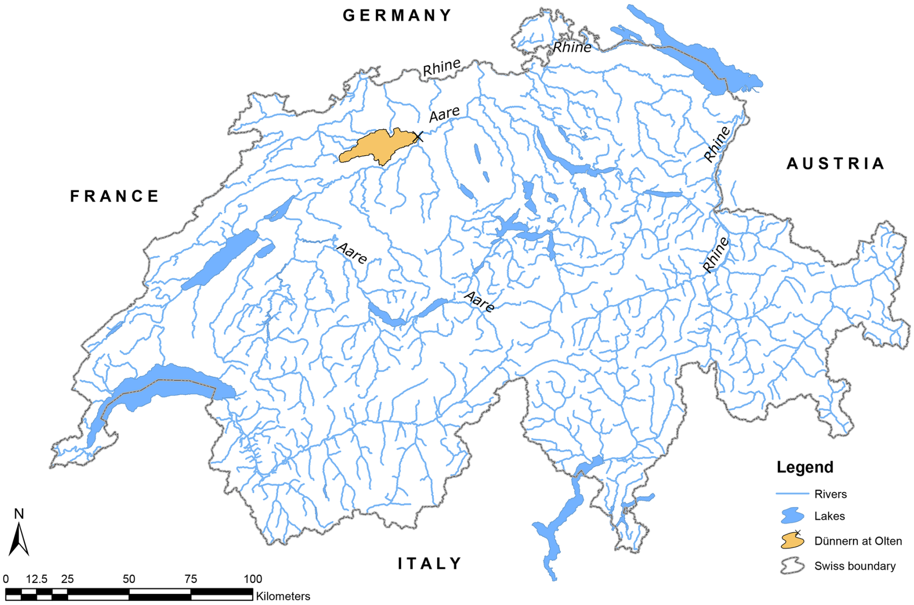

Figure 3. Location of the Dünnern at Olten catchment with a river network extracted from Swiss Map Vector 25 (SwissTopo, 2008).

3 Experimental set-up

3.1 Study catchment

For testing the methods developed here, a small close-to-natural catchment is preferable, i.e. with only little anthropogenic influence, in which hydrological responses are transparent, and the generation of rare floods (peaks) is not affected by human constructions (dams, bridges). For this purpose, the Dünnern stream at Olten catchment with an area of 196 km^2 is selected, located in the Jura region in Switzerland (Fig. 3). The Dünnern stream is a tributary of the Aare river and belongs to the basin of the Rhine river. The mean elevation of the Dünnern at Olten catchment is 711 m a.s.l., with an elevation span from 400 to 760 m a.s.l. The flow regime is defined as nival pluvial jurassien (Weingartner and Aschwanden, 1992; Schürch et al., 2010), with high flows in winter and spring and low flows in autumn. With no direct human influence within the entire catchment known, it can be assumed to be close to natural (BAFU, 2017). This catchment is part of a large-scale extreme-flood modelling effort in Switzerland for the entire Aare catchment (Viviroli et al., 2020).

3.2 Hydrological model and calibration data

To simulate the hydrological catchment responses to meteorological scenarios, the HBV model (Hydrologiska Byråns Vattenbalansavdelning) is used. The HBV model is a semi-distributed bucket-type model, and it consists of four main routines: (1) precipitation excess, snow accumulation and snowmelt; (2) soil moisture; (3) groundwater and streamflow responses; and (4) run-off routing using a triangular weighting function. Due to the presence of the snow component, the HBV model is applicable to mountainous catchments (e.g. Jost et al., 2012; Addor et al., 2014; Breinl, 2016; Griessinger et al., 2016; Sikorska and Seibert, 2018b).

In this study, the version HBV light (Seibert, 1997; Seibert and Vis, 2012) with 15 calibrated parameters is used; see Table A1 for details on model parameters and their calibration ranges. Such a set-up of the HBV light was previously successfully applied in Swiss catchments (e.g. Sikorska and Seibert, 2018a; Brunner et al., 2018c; Brunner and Sikorska-Senoner, 2019; Müller-Thomy and Sikorska-Senoner, 2019; Westerberg et al., 2020). Model inputs are time series of precipitation and air temperature and long-term averages of seasonally varying estimates of potential evaporation, all being area-average values for the entire catchment. These inputs are next redistributed along predefined elevation bands using two different constant altitude-dependent correction factors for precipitation and temperature. The model output is streamflow at the catchment outlet at time steps identical to input data (hourly in this study).

For the study catchment, meteorological inputs (hourly precipitation totals, hourly air temperature means, average hourly evaporation sums) for the HBV model are derived from observed records from meteorological stations and are averaged to the mean catchment values using the Thiessen polygon method. The recorded continuous hourly streamflow data at the catchment outlet (Olten station) cover the period 1990–2014.

3.3 Identification of multiple HBV parameter sets

To derive multiple parameter sets of the HBV model, we propose a heuristic approach that relies on multiple independent model calibration trials using a genetic-algorithm (GA) approach (Appendix A). By using independent model runs, the possibility of being trapped in the same local optimum should be reduced. The genetic algorithm is used together with a multi-objective function F_{obj} with three scores: the Kling–Gupta efficiency (R_{KGE}), the efficiency for peak flows (R_{PEAK}) and a measure based on the mean absolute relative error (R_{MARE}). R_{PEAK} is defined in a similar way to the Nash–Sutcliffe efficiency but using peak flows instead of the entire time series. While both R_{KGE} and R_{PEAK} focus on high (peak) flows, R_{MARE} is sensitive to low flows. See Appendix B for equations.

F_{obj} is obtained through weighing these metrics as follows:

$$F_{\mathrm{obj}} = 0.3R_{\mathrm{KGE}} + 0.5R_{\mathrm{PEAK}} + 0.2R_{\mathrm{MARE}}. \tag{15}$$

The weights in F_{obj} are chosen following our previous experience in modelling Swiss catchments (Sikorska et al., 2018; Westerberg et al., 2020). The available observational datasets are split into a calibration (1990–2005) and a validation (2006–2014) period. Evaluation of the model in the independent period is important as the model is applied to simulate time series outside the calibration period. To set up the initial conditions, 1 year of model simulations is discarded from the calibration simulation, and the remaining are used for model performance computation. For the validation period, the initial conditions are taken from the calibration simulation.

Here, the genetic algorithm is run 100 times, resulting in 100 independent optimal parameter sets (see Fig. C1). These 100 parameter sets represent similarly likely solutions to model hydrological responses in this catchment and can be explained by the equifinality of hydrological-model parameters (Beven and Freer, 2001). The median model efficiency measured with F_{obj} over all 100 runs is 0.7 in the calibration and in the validation periods, which can be assumed to be a good model performance on an hourly scale. Also, diagnostics of the Nash–Sutcliffe efficiency and the peak efficiency demonstrate that the model performs well in the range of high flows, which are mostly important for simulation of rare floods studied in this paper (see Fig. C2).

Note that the way to derive 100 parameter sets described above is one possible approach, and other calibration methods could be used (e.g. Monte Carlo or bootstrapping).

3.4 Generation of synthetic meteorological scenarios using a weather generator

Meteorological scenarios of synthetic precipitation and temperature data for the Dünnern at Olten catchment are generated with the weather generator model GWEX developed by Evin et al. (2018) and referred to in their paper as GWEX_Disag. This stochastic model is a multi-site precipitation and temperature model that reproduces the statistical behaviour of weather events on different temporal and spatial scales. The major property of GWEX is that it uses marginal heavy-tailed distributions for generating extreme-precipitation and extreme-temperature conditions. Moreover, it has been developed to generate long-term (≈ 1000 years) meteorological scenarios. GWEX_Disag generates precipitation amounts at a 3 d scale and then disaggregates them to a daily scale using a method of fragments (for details on the precipitation model, see the work of Evin et al., 2018, and for details on the temperature model, see the work of Evin et al., 2019).

The meteorological scenarios used in this study are a subset from the long-term meteorological scenarios developed for the entire Aare river basin using recorded data from 105 precipitation stations and from 26 temperature stations in Switzerland (Evin et al., 2018, 2019). For this larger-scale research project, GWEX_Disag was set up using daily precipitation and temperature data from the period 1930–2015 and hourly records of precipitation and temperature from 1990–2015 for the Aare river basin. The daily values generated with GWEX_Disag were then disaggregated to hourly values using the meteorological analogues method, which for each day in the simulated dataset finds an analogue day in observed data, i.e. with a known hourly time structure. Next, catchment means were computed using the Thiessen polygon method (using three stations located close by).

For the present study, 100 different meteorological scenarios (precipitation and temperature) covering the same time frame of 100 years at an hourly time step are available for the Dünnern at Olten catchment (see Fig. D1 for an overview of meteorological scenarios). The simulated data are assumed to be representative of current climate conditions, i.e. neither variation due to climate or land use change nor river modification is considered. Thus, differences between scenarios are exclusively due to the natural variability in the meteorological time series and modelled by the GWEX_Disag weather generator.

3.5 Generation of synthetic hydrological-simulation ensembles

Finally, for our analysis, 100 meteorological scenarios with continuous data of 100 years of precipitation and temperature and 100 calibrated parameter sets of the HBV model are available. This number of 100 was chosen as a compromise between minimizing the intensive model calibrations and the simulations at an hourly time step and maximizing the information content of the hydrological-parameter sample and the climate variability. We have chosen the same number of 100 for meteorological scenarios, parameter sets and simulation years to not favour any of these components in the methods' comparison. These 100 meteorological scenarios are used as input into the HBV model to generate streamflow time series with 100 different HBV parameter sets. To set up the initial conditions of the model, a 1-year warming-up period is always used prior to the simulation period. To get an overview of the variability in such hydrological ensembles, see Fig. D2.

From each of these continuous hydrological simulations, 100 annual maxima (AMs; one per calendar year) are selected (see Fig. 4). This results in the following analysis set-up:

- $I = 100$ and $i = 1, 2, \ldots 100$;
- $J = 100$, $y = 1, 2, \ldots 100$ and $j = 1, 2, \ldots 100$;
- $M = 100$ and $m = 1, 2, \ldots 100$,

with $100 \times 100 \times 100$ combinations of the annual maximum × hydrological-model parameter set × meteorological scenario.

These series of AMs are next used to test the developed methods of selecting the representative parameter sets from the ensemble of 100 available sets.

4 Results

4.1 Representative parameter sets

The representative parameter sets selected with each of the three methods are summarized over all 100 meteorological scenarios in Table 2, which presents the three most frequently chosen hydrological parameter sets for each method.

Although different parameter sets are usually selected by different methods, in a few cases the same set is chosen with more than one selection method. Among the first three most frequently chosen sets, the same parameter set is selected as the median set once for all three methods and several times for at least two methods.

The variability in the selected hydrological parameter sets is presented in Fig. 5. As can be seen from the figure, some parameters presented smaller and others larger variability in selected sets. It also appears that different values are selected for the infimum, median and supremum set but not always. Among the three selection methods, the ranking method (marked in green) has the largest spread of parameter values for most of the parameters. The clustering (blue) and quantiling (yellow) selection methods seem to choose more extreme parameter values for both, i.e. infimum and supremum sets. Looking at different model routines no clear patterns could be seen regarding the choice of parameter sets. It appears however that the representative parameters from the response (blue) and soil moisture (yellow) routines have a smaller spread than those from the snow routine (grey) as they are more often outside and further away from the interquartile ranges (grey boxplots).

4.2 Infimum, median and supremum intervals

Using the selected representative sets, representative intervals for rare-flood estimations are constructed for each of the 100 meteorological scenarios and each of the three selection methods. Examples of these intervals for two meteorological scenarios are presented in Figs. 6 and 7. Note that apart from selecting representative intervals, the clustering method leads to grouping all ensemble members into three selected clusters.

According to a first visual assessment, these three methods lead to slightly different constructed frequency intervals particularly in the upper tail of the distribution, i.e. for the most rare (highest) flows, which are of highest interest. Moreover, the ranking method leads to less symmetrically spread intervals, with the median and infimum intervals lying close to each other. The other two methods lead to more symmetrically spread intervals.

For the quantitative assessment, the ratio of scenarios incorrectly attributed, i.e. with intervals being mixed up (R_{bias}), varies between the three methods and is the highest for the ranking method ($R_{\text{bias}} = 0.54$). For the clustering method, the three intervals are always correctly ordered for all 100 meteorological scenarios tested ($R_{\text{bias}} = 0.0$). For the quantiling method, this ratio is equal to $R_{\text{bias}} = 0.02$ and thus is also very low. Hence, we can conclude that both clustering and quantiling methods provide correctly attributed intervals with a bias $\leq 2\,\%$. For the ranking method, the correctness of the interval attribution is poor, and in more than 50 % of the meteorological scenarios, the simulations corresponding to the selected parameter sets lead to mixed-up frequency intervals.

4.3 Evaluation of the three selection methods

The behaviour of the three selection methods is further evaluated with the 100 meteorological scenarios using the leave-one-out cross-validation test (Sect. 2.5.1) and the multi-scenario evaluation method (Sect. 2.5.2) and corresponding metrics (Sect. 2.6.2). Examples for two meteorological scenarios are presented in Fig. 8 for the leave-one-out cross-validation test and in Fig. 9 for the multi-scenario evaluation. From the visual assessment, it is difficult to judge the methods as they seem to perform similarly well. However, the range of the predictive intervals obtained with 99 meteorological scenarios (one left out) is considerably narrower for ranking and quantiling on one hand and much wider for clustering on the other hand (Fig. 8). Accordingly, the correspondence between the prediction interval and the full simulation range of the left-out scenario differs between the methods (Fig. 9).

This is reflected in the quantitative assessment of the methods' behaviour, summarized in Table 3. Namely, the leave-one-out cross-validation reveals that the quantiling method receives the highest values for both evaluation criteria, i.e.

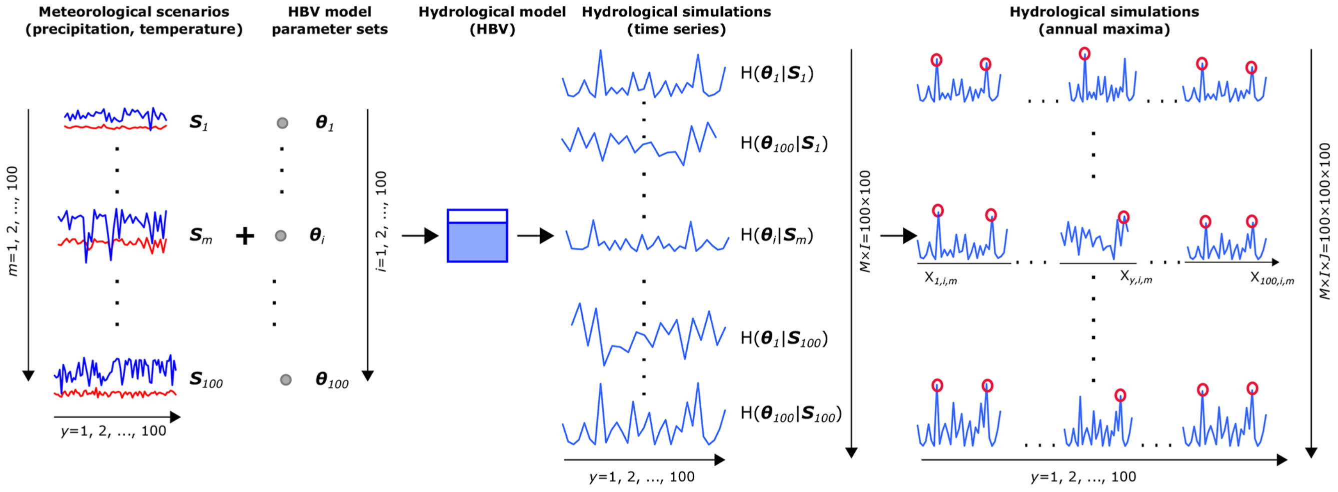

Figure 4. Set-up of the experimental study.

Table 2. The three representative parameter sets $\boldsymbol{\theta}_{\text{inf}}$, $\boldsymbol{\theta}_{\text{med}}$ and $\boldsymbol{\theta}_{\text{sup}}$ most frequently selected with three methods. The i stands for the set index and ct. for the number of counts. The expression $\sum$ ct. stands for the sum of counts for the first three most frequently selected sets. Bold font indicates parameter set indices which are selected as representative with at least two methods among the three sets most frequently chosen.

Method	Ranking						Quantiling						Clustering					
Repr. set	$\boldsymbol{\theta}_{\text{inf}}$		$\boldsymbol{\theta}_{\text{med}}$		$\boldsymbol{\theta}_{\text{sup}}$		$\boldsymbol{\theta}_{\text{inf}}$		$\boldsymbol{\theta}_{\text{med}}$		$\boldsymbol{\theta}_{\text{sup}}$		$\boldsymbol{\theta}_{\text{inf}}$		$\boldsymbol{\theta}_{\text{med}}$		$\boldsymbol{\theta}_{\text{sup}}$	
Par. set	i	ct.	i	ct.	i	ct.	i	ct.	i	ct.	i	ct.	i	ct.	i	ct.	i	ct.
1st	**97**	21	1	11	20	25	**47**	78	**2**	22	**19**	32	**47**	62	**2**	13	34	48
2nd	16	15	**2**	7	**19**	13	**66**	10	93	11	86	15	**97**	35	**46**	11	22	33
3rd	6	12	14	7	57	9	67	4	**46**	9	69	11	**66**	2	62	11	98	7
$\sum$ ct.		48		25		47		92		42		58		99		35		88

the ratio of simulation points lying outside the predictive intervals (R_{spo}) and the ratio of hydrological-simulation ensemble members lying outside the predictive intervals (R_{hso}), both presented as median values over all scenarios. Thus, this method performed the poorest among all three methods tested here. Yet, with $R_{\text{spo}} \leq 0.14$ for the 50th percentile and $R_{\text{hso}} \leq 0.05$ for the threshold $r_{\text{thr}} \geq 0.50$, even this method can be qualified as behaving well based on the leave-one-out cross-validation. For the ranking and the clustering methods, similar values for these two metrics are achieved, with slightly lower values for the ranking method.

In summary, it can be said that all criteria values are relatively low for all three methods, and thus the computed criteria values can only be used to order the methods by their behaviour, while none of the methods are rejected.

In contrast to the above findings, the multi-scenario evaluation reveals different results, with R_{spo} being the lowest for clustering and the largest for the ranking method. Similarly, the ratio of meteorological scenarios lying outside the predictive intervals (R_{mso}) is the lowest for clustering and the highest for the ranking method (r_{thr} in Table 3).

Also, here all computed criteria values are relatively low, with $R_{\text{spo}} \leq 0.05$ for the 50th percentile and $R_{\text{mso}} = 0$ for the threshold $r_{\text{thr}} \geq 0.50$ for the poorest-behaving method (ranking). Hence, again here all three methods can be qualified as behaving well based on the multi-scenario evaluation.

Analysis of overlapping PI pools (R_{OPPIs}) and relative band spreads ($R_{\Delta\text{PIs}}$) shows that in the cross-validation test all methods provide bands that are wider than the 90% PIs computed using the full simulation ensemble. This should not be surprising as the selection of relative parameter sets is based on a larger sample of hydrological-model realizations (i.e. 99 scenarios) than the full ensemble for model assessment (i.e. single scenario). However, these metrics show large differences in the multi-scenario test, in which the clustering method outperforms the other two selection methods, particularly when the focus lies on rare floods (compare $R_{\Delta\text{PIs},\,j\geq51}$, $R_{\Delta\text{PIs},\,j\geq91}$ and $R_{\Delta\text{PIs},\,j\geq96}$ in Table 3). The quantiling was the second-best method, while the ranking performed the worst. These observations are also confirmed when looking at the variability in these two metrics for different return periods (Fig. 10). A better performance of the

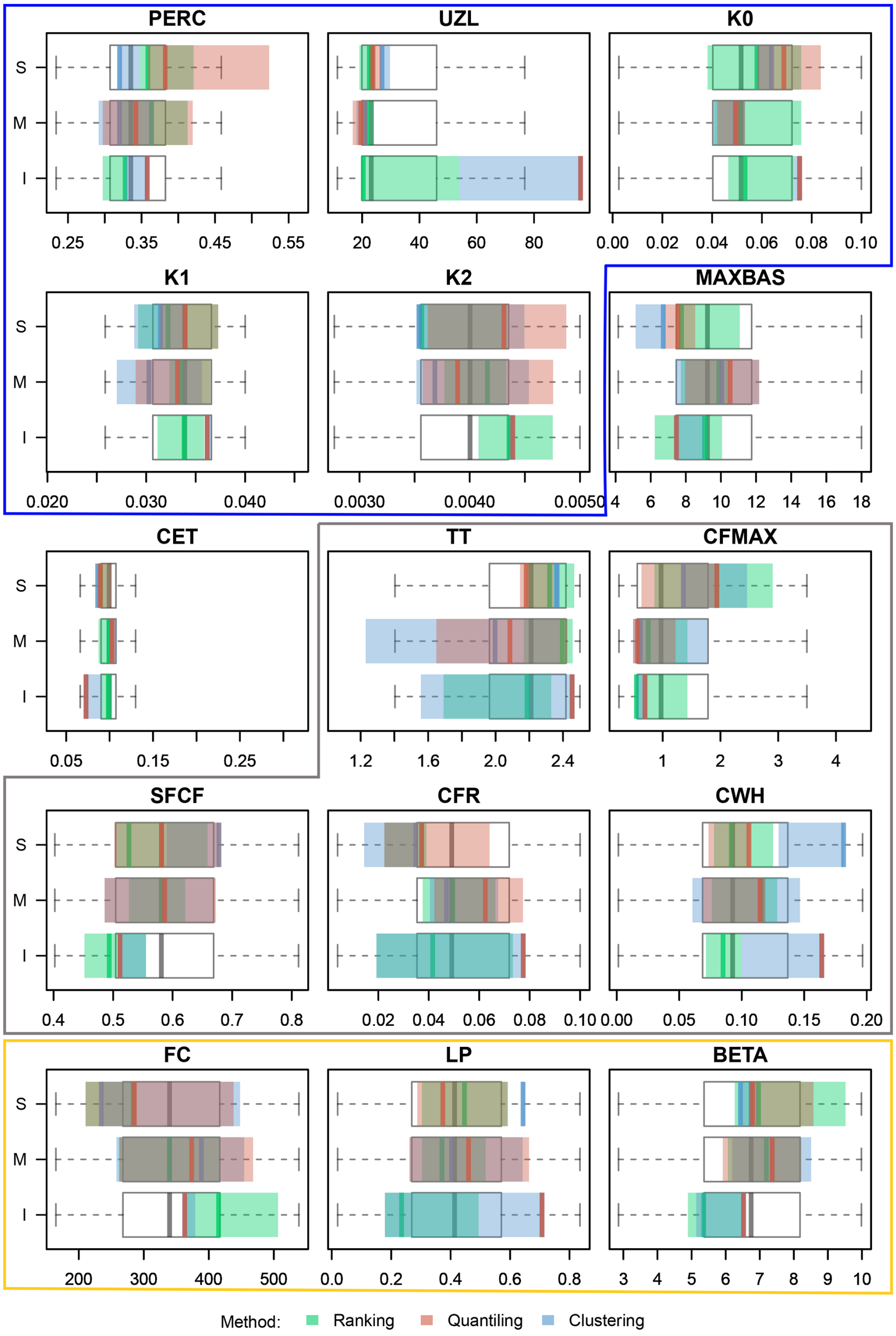

Figure 5. Box plots showing the variability in the hydrological parameter sets selected as the representative parameter sets over 100 meteorological scenarios chosen with three methods. The white box plots illustrate the entire parameter ensemble (i.e. 100 sets); outliers are not presented. I: infimum; M: median; and S: supremum set. Units as in Table A1. The blue box surrounds parameters from the response routine, the grey box from the snow routine and the yellow from the soil moisture routine. MAXBAS is the only parameter from the routing routine, and CET is a potential-evaporation correction factor.

clustering method can be again noticed in the range of rare floods. While quantiling performed worse than clustering, it was still better than the ranking method.

As it appears from the above, the rejection or acceptance of one of the three methods tested here is not straightforward. Apart from the ranking method, which was linked to a huge bias, both other methods, i.e. quantiling and clustering, performed similarly well. Yet, these methods provide quite different intervals (of a different spread). The validity and usefulness of these methods for selecting the representative

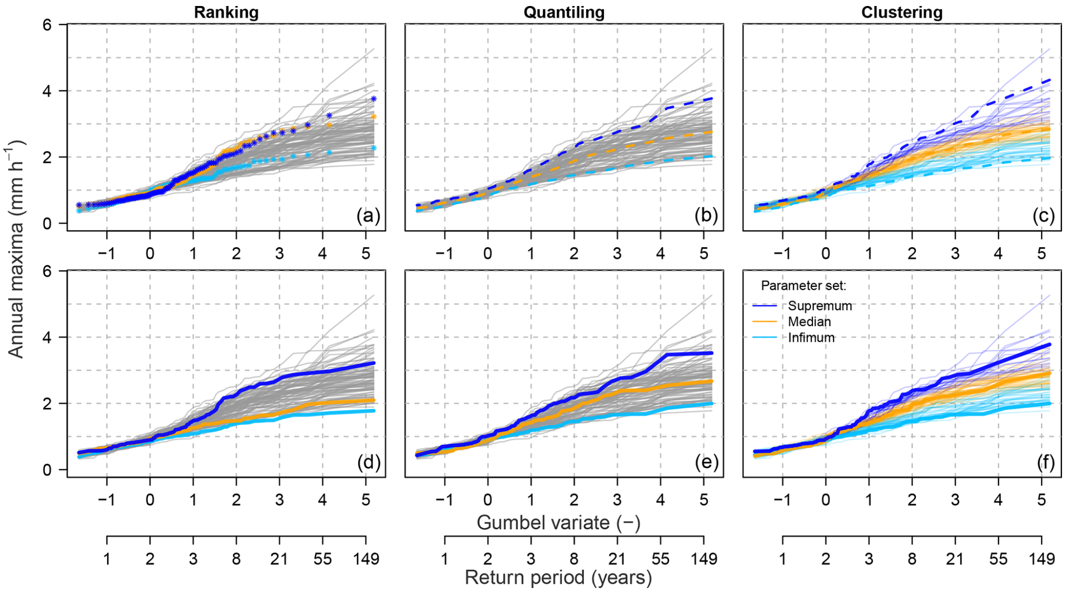

Figure 6. Example of the representative parameter sets' selection with three methods in the Dünnern at Olten catchment (meteorological scenario $m = 14$). The top panel presents intermediate steps of selecting the representative sets and the bottom panel the finally constructed intervals, i.e. infimum, median and supremum. The dashed lines (top panel) indicate the computed representative intervals (i.e. steps a–c in ranking and clustering and a–b in quantiling), and the solid lines (bottom panel) indicate the hydrological-simulation members corresponding to the parameter sets selected as representative (step d in ranking and quantiling and e in clustering).

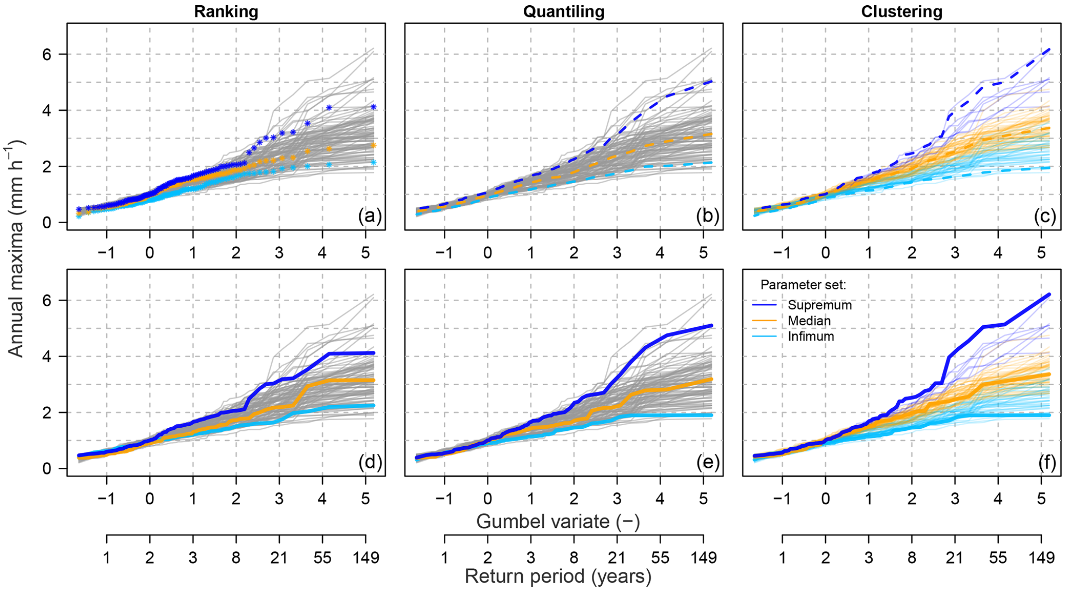

Figure 7. Example of the representative parameter sets' selection with three methods in the Dünnern at Olten catchment (meteorological scenario m =93); description as in Fig. 6.

parameter sets are thus further discussed below in Sect. 5.1. The detailed analysis of the relative band spread and the overlapping pools indicated however that the clustering method performed the best, particularly in the range of rare floods. The quantiling method was scored as the second best, while the ranking method performed poorest.

5 Discussion

5.1 Behaviour of three selection methods

The results from our experimental study demonstrate that generally all three methods are capable of selecting representative parameter sets that yield reliable predictive intervals in

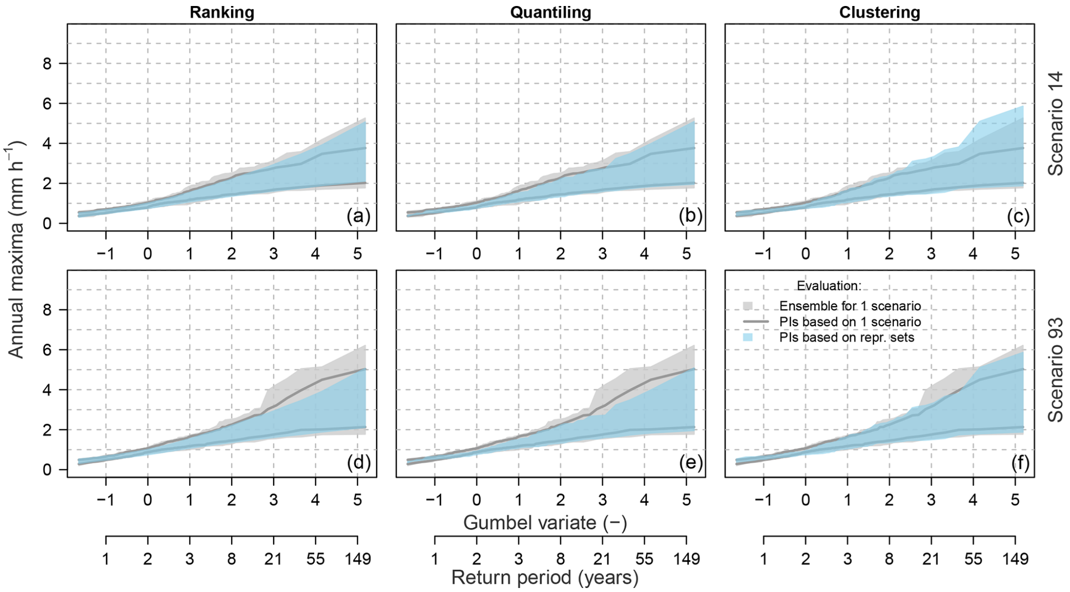

Figure 8. Example of leave-one-out cross-validation for the three selection methods and two meteorological scenarios. PIs represent the 90 % predictive intervals.

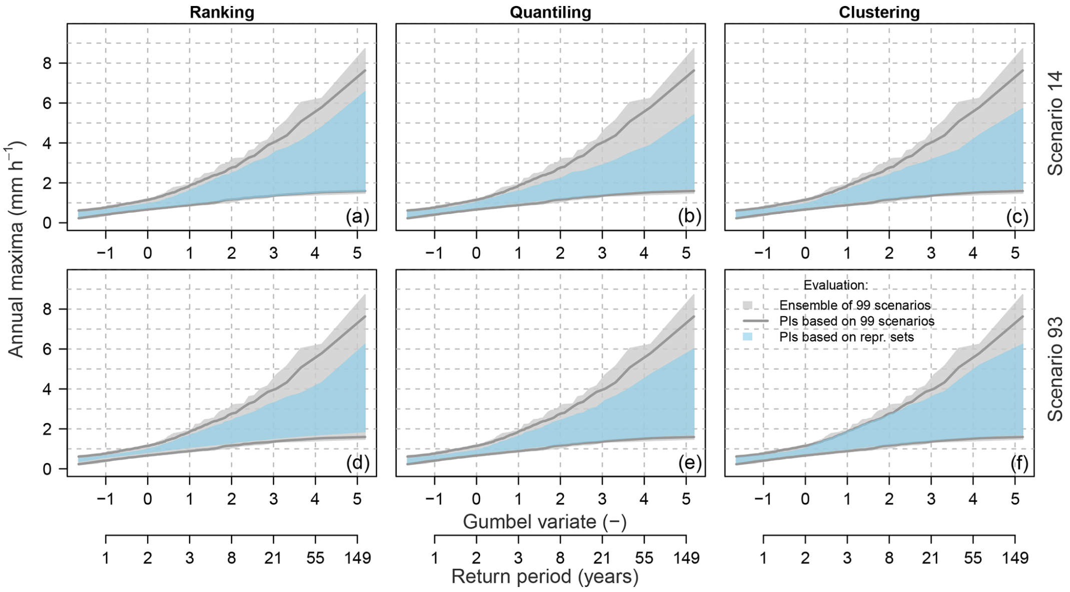

Figure 9. Example of multi-scenario evaluation for the three selection methods and two meteorological scenarios. PIs represent the 90 % predictive intervals.

the frequency domain, i.e. all three methods are fit for purpose for extreme-flood simulation, with the ranking method performing, however, clearly less well than the others (larger bias, as visible in Sect. 4.2). As the developed methods rely on selecting three representative sets as infimum, median and supremum, they respect the maximum variability between individual ensemble members for a given meteorological scenario.

In the validation tests, the behaviour scores of the three methods, however, were attributed differently depending on the evaluation criteria. To further compare the methods, we provide a detailed discussion of the major differences below and present a synthesis of how the methods rank on average (averaged across all scenarios) for the quantitative evaluation criteria, which we support with further qualitative evaluation criteria (Table 4).

Table 3. Metrics of the behaviour of the approach for three methods of selecting representative parameter sets and the predictive intervals in the leave-one-out cross-validation and in the multiple-scenario evaluation. The values represent the median values over all 100 scenario runs.

	Leave-one-out cross-validation			Multi-scenario evaluation		
Metric	Ranking	Quantiling	Clustering	Ranking	Quantiling	Clustering
R_{spo} [–], 50th percentile	0.02	0.13	0.065	0.048	0.02	0
R_{hso} [–], $r_{\mathrm{thr}} \geq 0.50$	0.02	0.05	0.02	–	–	–
R_{hso} [–], $r_{\mathrm{thr}} \geq 0.10$	0.26	0.57	0.41	–	–	–
R_{mso} [–], $r_{\mathrm{thr}} \geq 0.50$	–	–	–	0	0	0
R_{mso} [–], $r_{\mathrm{thr}} \geq 0.10$	–	–	–	0.20	0.091	0.03
$R_{\Delta\mathrm{PIs}}$ [–], mean	1.26	1.01	1.30	0.58	0.57	0.67
$R_{\Delta\mathrm{PIs},j\geq 51}$ [–]	1.23	1.16	1.56	0.59	0.70	0.81
$R_{\Delta\mathrm{PIs},j\geq 91}$ [–]	1.18	1.15	1.62	0.51	0.68	0.82
$R_{\Delta\mathrm{PIs},j\geq 96}$ [–]	1.21	1.21	1.67	0.49	0.67	0.80
R_{OPPIs} [–]	1.21	1.19	1.64	0.52	0.65	0.79

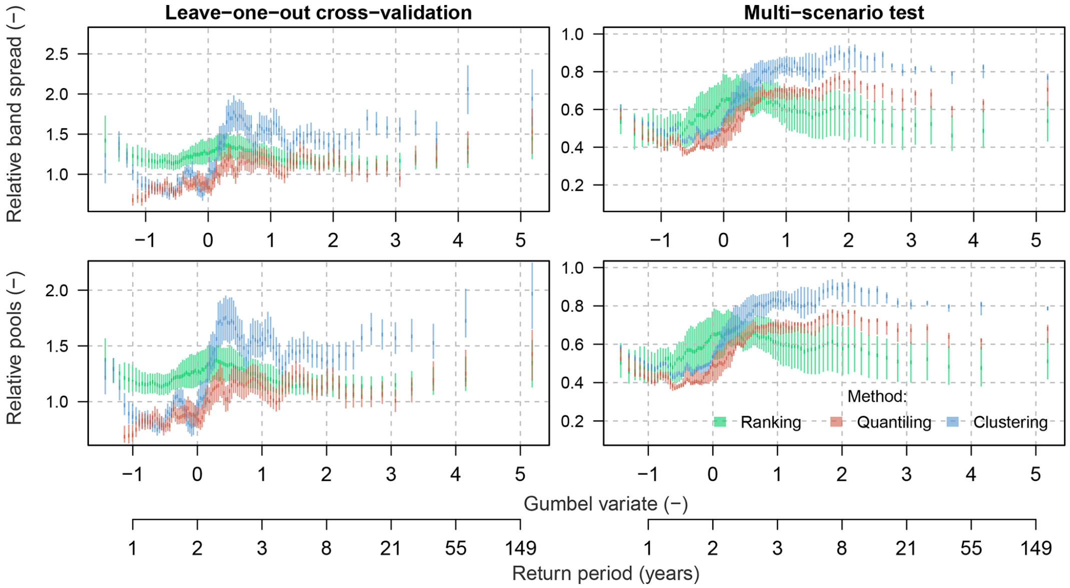

Figure 10. Evaluation of the leave-one-out cross-validation and the multi-scenario test for the three selection methods using the relative band spread ($R_{\Delta\mathrm{PIs}}$) and the relative overlapping pools (R_{OPPIs}), both computed with reference to the 90 % PIs of the full hydrological-simulation ensemble.

From the visual assessment, i.e. based on the method bias (R_{bias}), it clearly appears that the ranking method is the most biased method (with more than half of all meteorological scenarios having mixed-up intervals), while the other two methods can be considered to be unbiased with correctly attributed intervals for 98 % (quantiling) or more (clustering) of all meteorological scenarios considered here (Sect. 4.2, unbiasedness in Table 4). As expected, these findings are further confirmed by the results from the multi-scenario evaluation that yield the best behaviour for the clustering method and the worst for the ranking method (Sect. 4.3), particularly if the focus lies on rare floods as assessed by the relative band spread and the overlapping pools.

Interestingly, the leave-one-out cross-validation study, in contrast to the multi-scenario evaluation, attributes the lowest criteria value to the ranking method, i.e. ranks it as the best method (Table 4). This requires a careful interpretation and understanding of how the predictive intervals are constructed in both evaluation studies. In the leave-one-out cross-validation study, the representative parameter sets are selected, and the predictive intervals are constructed based on 99 meteorological scenarios and then evaluated against the full simulation range corresponding to the left-out scenario. In the multi-scenario evaluation, the representative parameter sets are selected based on a single scenario, and the predictive intervals are then assessed by applying these three selected sets (selected based upon a single scenario) to

Table 4. Synthesis of scoring ranks attributed to the three methods for selecting representative parameter sets (based on quantitative metrics). The ranks are attributed descending from the best (1) to the worst (3) behaviour. The median scoring rank (last line) corresponds to the median over all criteria.

Score criteria	Ranking	Quantiling	Clustering
Unbiasedness (not mixed-up intervals)	3	2	1
Leave-one-out cross-validation	1	3	2
Multi-scenario evaluation	3	2	1
Independence from meteorological scenario	3	1	1
Independence from simulation years	3	1	1
Ease in application	1	3	3
Interpretability of prediction intervals	3	1	1
Median scoring rank	3	2	1

the other 99 meteorological scenarios. Hence, by comparing findings from these two evaluation studies, it appears that the ranking method performs poorly if using a single scenario for selecting the representative sets (multi-scenario evaluation). In exchange, the ranking method outperforms the two other methods when a high number of meteorological scenarios are used for selecting the representative parameter sets (leave-one-out cross-validation). This means that the ranking method strongly depends on the meteorological scenario choice, while the other two methods result in representative parameter sets that are transferable to other meteorological scenarios. We hence introduce here a criterion *independence from meteorological scenario*, which defines how strongly the selected sets depend on the meteorological scenario used for selection of representative parameter sets.

In a similar way, *independence from simulation years* will define how strongly the selected sets depend on the simulation years used for selection of the representative parameter sets. To make statements on that, one needs to recall how the selection methods are constructed: the ranking method, in fact, depends strongly on the selected simulation period (and hence on the meteorological scenario) because the selection of the representative parameter sets is performed on unsorted annual maxima for each simulated year independently. The other two methods are performed over the entire simulation period, which makes them less strongly dependent on individual simulation years. Nevertheless, the ranking method can be considered to be the (computationally and methodologically) easiest in application due to its selection criteria relying purely on ranking within individual simulation years. We call this criterion *an ease in application*. The other two methods need to be performed in the frequency space on sorted annual maxima over the entire simulation period and, in the case of the clustering method, require some additional computational effort (which remains low, however, compared to the hydrological simulation). The use of the frequency space in selecting the representative parameter sets helps, however, to interpret the constructed prediction intervals and to directly assign return periods to them. This speaks for their higher *interpretability of prediction intervals* as compared to the ranking method, in which interpretation of intervals is very limited (as they are selected without any flow frequency analysis).

Overall based on scoring results from Table 4, it appears that the clustering method behaves the best (with a median scoring rank of 1) due to its unbiasedness and due to a good performance achieved for all evaluation criteria, for both the leave-one-out cross-validation and the multi-scenario evaluation. Finally, this method was proven to perform the best if the focus lies on rare-flood simulations.

5.2 Limitations and perspectives

We should emphasize that the presented methods are independent of the selected hydrological-model calibration approach or from the selected hydrological-response model and are thus readily transferable to any similar simulation setting. Despite the fact that the calibration of a hydrological model lies beyond the scope of this paper, it is assumed that (at least) 100 parameter sets of a hydrological model can be made available for selecting the representative parameter sets. For that purpose, a hydrological model should be calibrated with observed data of a long enough record that covers rare floods so that rare floods could be realistically simulated. In this work, to derive 100 parameter sets, we proposed a heuristic approach that relies on multiple independent model calibration trials using a genetic-algorithm approach and a multi-objective function. This method represents an interesting solution to systematic sampling of the posterior parameter distributions (e.g. via Markov chain Monte Carlo sampling) or to any Monte Carlo method relying on a very high number of model runs. Its strength is that it can be applied for selecting parameter sets from independent model calibration settings (with different scores, calibration periods, etc.).

Note however that for the purpose of deriving 100 parameter sets, a continuous hydrological model does not necessarily require continuous calibration data, and it could be also calibrated to discrete data (e.g. using hydrological signatures; Kavetski et al., 2018). If no observed data or only too short records are available, model parameters can be acquired through regionalization approaches (see the work of Brunner et al., 2018a, for an overview of regionalization methods). The developed methods are of use for applications when a hydrological model should be employed for simulations of rare floods. If the use of a hydrological model is not possible, i.e. neither information for calibration nor sufficient information for parameter regionalization is available, these methods cannot be applied. Moreover, although the methods are tested with a bucket-type hydrological model, the most valuable application of the proposed methods would be to computationally more demanding hydrological models that can profit even more from a reduced computational demand.

Furthermore, the proposed approach is tested here using synthetic hydrological data, i.e. using streamflow simulations of the hydrological model in response to meteorological scenarios. We chose to use synthetic instead of real observed data to work with long enough continuous simulations that cover rare events and to minimize the focus of the model error arising from the calibration data and procedure. By using synthetic data as a reference (instead of observed data), the latter error can be neglected here. The proposed methods should be tested with more catchments and other models to verify the scoring of methods that was achieved in this study.

Selection methods proposed in this study enable one to choose representative parameter sets of a hydrological model and based on those to construct uncertainty predictive intervals (PIs) for extreme-flood analysis in the frequency space. Here, we tested the methodology using 100 meteorological scenarios that should represent the natural climate variability and in this way should provide independent conditions for methods' evaluation. Such a method for constructing PIs from a hydrological model ensemble is a powerful tool that opens several avenues for further detailed uncertainty analysis. For instance, one may be interested in contributions of different uncertainty sources into the total PIs constructed, e.g. coming from the hydrological model or the natural climate variability. As these two components are not linearly additive, their separation is not straightforward.

In addition, any ensemble simulation also encompasses other uncertainty sources of the modelling chain, such as those resulting from the weather generator, from the structure of the selected hydrological model, from the prediction of very rare flood events, etc. (Lamb and Kay, 2004; Schumann et al., 2010; Kundzewicz et al., 2017). To assess individual contributions of interest, a simple sensitivity analysis based on the variance variability could be recommended here, in which one uncertainty source is propagated through the method at once while other sources are kept at their mode or median values and in which the resulting PI spread is compared.

Downsizing the hydrological-model parameter sample can only aim to understand and characterize the hydrological part of the full hydrological ensemble resulting from a combination of multiple parameter sets and multiple meteorological scenarios. The variability in hydrological-model parameters arises from the parameter equifinality (Beven and Freer, 2001), and it can be overcome by using several hydrological-model parameter sets that should encompass the parametric and (implicitly) also other uncertainty sources. Our selection methods thus enable one to choose representative parameter sets from the hydrological-response point of view and in this way to cover the variability in hydrological responses with a reduced number of hydrological-model runs needed. These methods are however not applicable for characterizing the climate variability (nor for downsizing the number of meteorological scenarios needed).

Moreover, in developing the selection methods, we did not distinguish between different flood types such as heavy rainfall excess or intensive snowmelt events (Merz and Blöschl, 2003; Sikorska et al., 2015). Also, as we focused only on large annual floods (annual maxima), we did not represent the flood seasonality in our analysis. Yet, some recent works emphasize the need to include such information on the flood type (Brunner et al., 2017) or on flood seasonality (Brunner et al., 2018c) into bivariate analysis of floods or to represent a mixture of both flood type and flood seasonality in flood frequency analysis (Fischer et al., 2016; Fischer, 2018). Thus, the proposed selection methods could potentially be extended to account for different flood types during representative parameter selection, e.g. using automatic methods of flood type attribution from long discharge series (Sikorska-Senoner and Seibert, 2020). For that purpose, peak-over-threshold (POT) selection criteria of flood peaks could be more appropriate instead of a block selection (annual maximum) used here in constructing the simulated distributions of hydrological responses in order to cover a range of different flood processes.

Finally, we downsize the hydrological-model parameter sample to three sets which represent the predictive intervals of the full ensemble of hydrological responses fairly well given different meteorological scenarios. This number of three sets is motivated by the fact that it can be readily processed within a fully continuous ensemble-based framework using numerous climate settings. This is common practice in flood frequency analysis, and the three sets emulate the common practice of communicating median values along with prediction limits (Cameron et al., 2000; Blazkova and Beven, 2002; Lamb and Kay, 2004; Grimaldi et al., 2012b). For safety studies, these representative intervals should be additionally statistically proved.

Optionally, one could further downsize the hydrological-model parameter sample to two sets (i.e. infimum and supremum), which would represent the intervals only. Downsizing to more than three parameter sets (e.g. five or more) could have the advantage of containing more information on uncertainty intervals, e.g. in the case that they are asymmetric, and should be explored in further studies.

Possible applications of these selection methods include all studies where computational requirements are an issue, e.g. rare-flood analysis in safety studies concerning dams or bridge breaks; climate scenarios of these; and evaluation of rare floods due to changes in climatic variables using several emission scenarios and different uncertainty source propagation. Finally, these methods could be used for quantifying different uncertainty source contributions in rare-flood estimates but with less effort from the hydrological model as due to parametric uncertainty propagation.

6 Conclusions

In this study, we propose and test three methods for selecting the representative parameter sets of a hydrological model to be used within fully continuous ensemble-based simulation frameworks. The three selection methods are based on ranking, quantiling and clustering of simulation of annual maxima within a limited time window (100 years) that is much shorter than the full simulation period of thousands of years underlying the simulation framework. Based on a synthetic case study, we demonstrate that these methods are reliable for downsizing a hydrological-model parameter sample composed of 100 parameter sets to three representative sets that represent most of the full simulation range in the frequency space. Among the tested methods, the clustering method that selects parameter sets based on cluster analysis in the frequency space appears to outperform the others due to its unbiasedness, its transferability between meteorological scenarios and a better performance for rare floods. The ranking method, which is the only tested method that completes the parameter selection on non-sorted annual maxima, can clearly not be recommended for typical settings since it (i) tends to result in mixed-up prediction intervals in the frequency space and (ii) depends too strongly on the simulation period used for parameter selection and thus lacks transferability to other periods or other meteorological scenarios. Possible applications of these methods include all fully continuous simulation schemes for rare-flood analysis and particularly those for which computational constraints arise, such as safety studies or scenario analysis.

Appendix A: Details on the HBV model parameters and model calibration

For searching the best hydrological-model parameter sets within the defined parameter ranges (Table A1), a genetic-algorithm and Powell optimization (GAP) approach (Seibert, 2000) is used. This approach is executed in two major steps. Firstly, the GA optimization is performed and relies on an evolutionary mechanism of selection and recombination of a user-defined number of parameter sets (i.e. parameter population) randomly selected within the defined parameter ranges. The principle idea of this searching relies on regenerating the parameter sets from the subgroup of parameter sets selected using the defined objective function F_{obj} as a criterion to choose the parameters that give the highest value of F_{obj} at the previous step of the model calibration. The search for the best parameter set is terminated at a user-defined maximum number of model interactions and results in a selected optimal parameter set. Secondly, the optimal parameter set obtained at the previous step is used as a starting point for a local optimization search using Powell's quadratically convergent method (Press et al., 2002). The parameter set finally achieved from the local optimization is retained as the best set. In this study, the total number of model interactions is set to 2500 for the GA and 500 for the local Powell's optimization.

Table A1. Parameter ranges for the calibration of the HBV model.

Parameter	Unit	Min.	Max.	Description
PERC	$\mathrm{mm\,h^{-1}}$	0	1	Percolation parameter
UZL	mm	0	100	Groundwater run-off threshold parameter
$K0$	$\mathrm{h^{-1}}$	1×10^{-4}	0.2	Recession coefficient
$K1$	$\mathrm{h^{-1}}$	1×10^{-5}	0.1	Recession coefficient
$K2$	$\mathrm{h^{-1}}$	1×10^{-8}	0.05	Recession coefficient*
MAXBAS	h	1	100	Length of triangular weighting function
CET	$\mathrm{^{\circ}C^{-1}}$	0	0.5	Correction factor for potential evaporation
TT	°C	-2.5	2.5	Threshold temperature
CFMAX	$\mathrm{mm\,h^{-1}\,^{\circ}C^{-1}}$	1×10^{-3}	5	Degree–hour factor
SFCF	–	0.4	1.6	Snowfall correction factor
CFR	–	0	0.1	Refreezing correction factor
CWH	–	0	0.2	Water-holding capacity
FC	mm	50	550	Maximum moisture storage in soil box
LP	–	0	1	Threshold for reduction in evaporation
BETA	–	1	10	Shape coefficient

* For recession coefficients the following condition must be fulfilled: $K0 > K1 > K2$.

Appendix B: Equations used in the multi-objective function

$$R_{\mathrm{KGE}} = 1 - \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2}, \tag{B1}$$

where r, α and β are the correlation, a measure of the relative variability in the simulated and observed values, and a bias.

$$R_{\mathrm{PEAK}} = 1 - \frac{\sum \left(Q_{\mathrm{o,peak}} - Q_{\mathrm{s,peak}}\right)^2}{\sum \left(Q_{\mathrm{o,peak}} - \overline{Q}_{\mathrm{o,peak}}\right)^2}, \tag{B2}$$

where $Q_{\mathrm{o,peak}}$ and $Q_{\mathrm{s,peak}}$ are observed and simulated values for flood peaks, and $\overline{Q}_{\mathrm{o,peak}}$ is the average value of $Q_{\mathrm{o,peak}}$.

$$R_{\mathrm{MARE}} = 1 - \frac{1}{n}\sum \frac{|Q_{\mathrm{o}} - Q_{\mathrm{s}}|}{Q_{\mathrm{o}}}, \tag{B3}$$

where n is the number of observation points, and Q_{o} and Q_{s} are observed and simulated discharge.

For further details on R_{KGE} see the work of Gupta et al. (2009), for details on R_{PEAK} see the work of Seibert (2003), and for details on R_{MARE} see the work of Dawson et al. (2007).

Appendix C: Model calibration results

The optimized hydrological-model parameter sets are presented in Fig. C1, whereas diagnostics of the model performance during the calibration and validation periods are presented in Fig. C2.

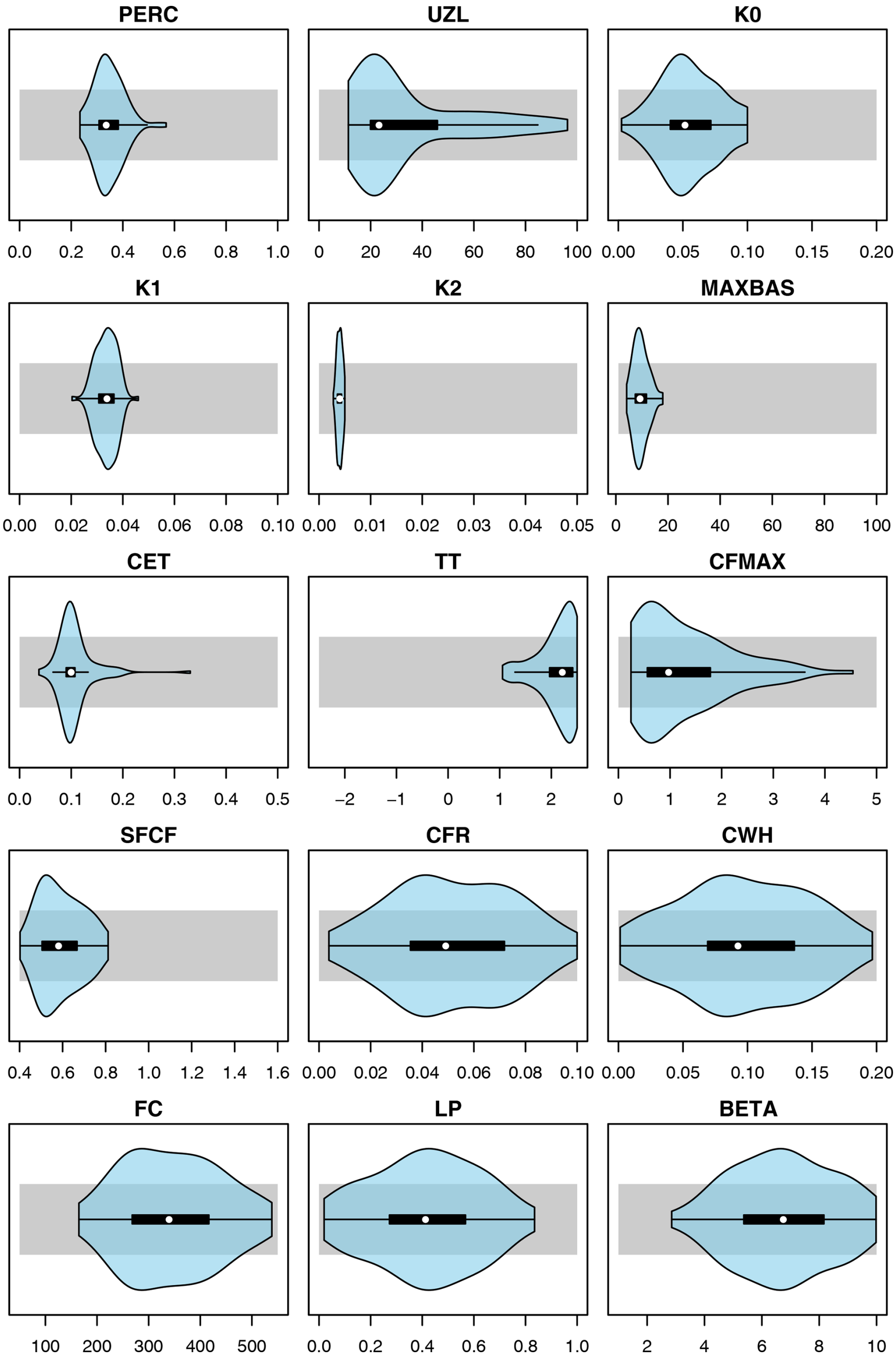

Figure C1. Violin plots (blue) summarizing 100 optimized parameter sets of the HBV model for the Dünnern at Olten catchment vs. initial calibration ranges (grey). Units as in Table A1.

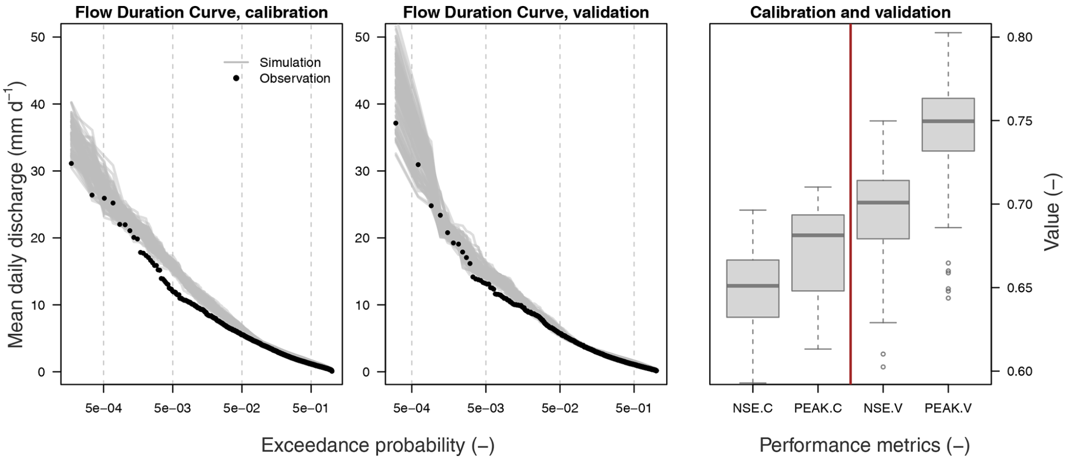

Figure C2. Flow duration curves and model performance metrics for calibration and validation periods over all 100 optimized parameter sets.

Appendix D: Scenario variability

The variability in the precipitation depth (annual daily maxima) and temperature (annual daily minima and maxima) of 100 meteorological scenarios used in this study is presented in Fig. D1. It can be seen that, in comparison to the observations, the meteorological scenarios are generally slightly colder ($\mu = 22.9\,^\circ\mathrm{C}$ and $\sigma = 1.5\,^\circ\mathrm{C}$ vs. $\mu = 25.8\,^\circ\mathrm{C}$ and $\sigma = 1.4\,^\circ\mathrm{C}$ for the annual daily maxima and $\mu = -11.9\,^\circ\mathrm{C}$ and $\sigma = 3.1\,^\circ\mathrm{C}$ vs. $\mu = -7.3\,^\circ\mathrm{C}$ and $\sigma = 3.5\,^\circ\mathrm{C}$ for the annual daily minima) and wetter ($\mu = 46.1\,\mathrm{mm}$ and $\sigma = 12.4\,\mathrm{mm}$ vs. $\mu = 45.9\,\mathrm{mm}$ and $\delta = 12.0\,\mathrm{mm}$ for the annual maximal daily precipitation depths). The variability in resulting hydrological scenarios is presented in Fig. D2 together with observations.

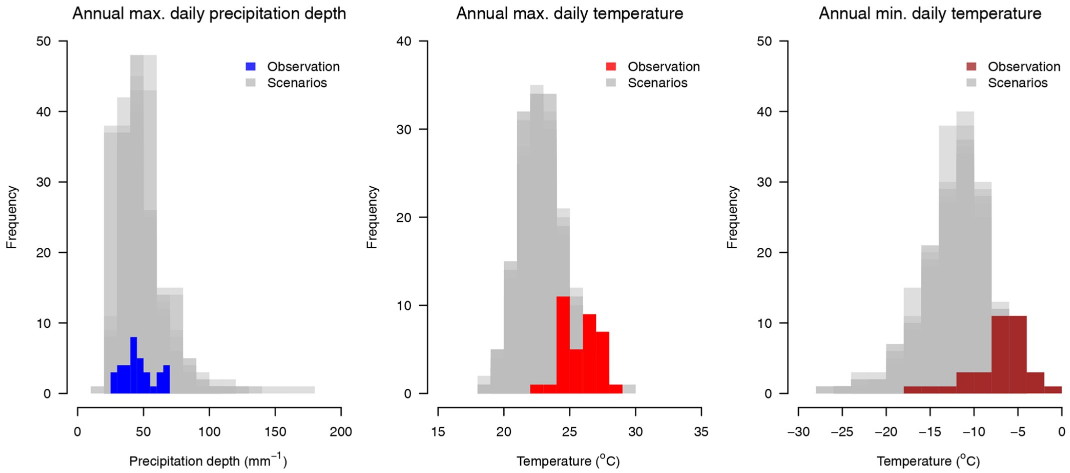

Figure D1. Variability in 100 meteorological scenarios used in this study vs. observations.

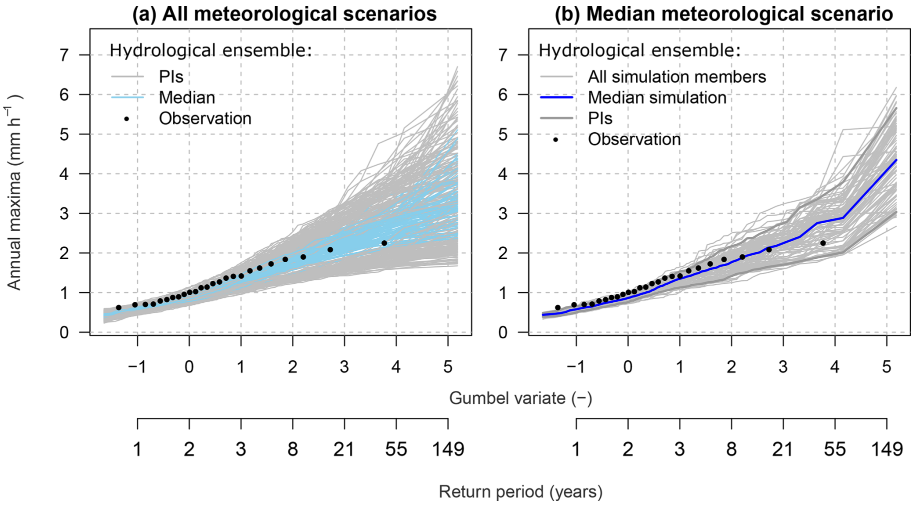

Figure D2. Variability in 100 hydrological scenarios used in this study; **(a)** hydrological ensemble with all meteorological scenarios and all hydrological-model parameters; **(b)** hydrological ensemble with all hydrological-model parameters but for the median meteorological scenario only. PIs represent the 90 % predictive intervals.

Data availability. The observed discharge data for calibrating the hydrological model can be ordered from the FOEN (https://www.bafu.admin.ch, last access: 12 February 2020) (FOEN, 2020), the observed meteorological data from MeteoSwiss (http://www.meteoswiss.ch, last access: 12 February 2020) (MeteoSwiss, 2020) and the topographic data from Swisstopo (http://www.swisstopo.ch last access: 12 February 2020) (SwissTopo, 2008).

Author contributions. AESS and JS developed the idea. AESS and BS jointly designed the details of this study. AESS performed the analyses and produced the figures. AESS, with a contribution from BS, wrote the first draft of the manuscript. The manuscript was edited by BS and revised by JS.

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

Acknowledgements. The research was funded through the Forschungskredit of the University of Zurich, grant no. FK-18-118. The ScienceCloud provided by S3IT at the University of Zurich enabled the computation-intensive simulations to be run on virtual machines. Meteorological scenarios used in this study were made available through the project EXAR (Hazard information for extreme flood events on the rivers Aare and Rhine), project no. 15.0054.PJ/O503-1381, funded by the Swiss Federal Office for Environment (FOEN). The authors wish to give special thanks to Guillaume Evin, Anne-Catherine Favre and Benoit Hingray for preparing the meteorological scenarios; to Marc Vis for setting-up virtual machines; and to Tracy Ewen for proofreading the manuscript. We thank the editor Paolo Tarolli, Emmanuel Paquet and one anonymous referees for their constructive comments, which improved the quality of the manuscript.

Financial support. This research has been supported by the Forschungskredit of the University of Zurich (grant no. FK-18-118).

Review statement. This paper was edited by Paolo Tarolli and reviewed by Emmanuel Paquet and one anonymous referee.

References

Addor, N., Rössler, O., Köplin, N., Huss, M., Weingartner, R., and Seibert, J.: Robust changes and sources of uncertainty in the projected hydrological regimes of Swiss catchments, Water Resour. Res., 50, 7541–7562, https://doi.org/10.1002/2014WR015549, 2014.

American Society of Civil Engineers: Hydrology Handbook, 2nd Edn., American Society of Civil Engineers, New York, https://doi.org/10.1061/9780784401385, 1996.

Arnaud, P. and Lavabre, J.: Coupled rainfall model and discharge model for flood frequency estimation, Water Resour. Res., 38, 1075, https://doi.org/10.1029/2001WR000474, 2002.

Arnaud, P., Cantet, P., and Odry, J.: Uncertainties of flood frequency estimation approaches based on continuous simulation using data resampling, J. Hydrol., 554, 360–369, https://doi.org/10.1016/j.jhydrol.2017.09.011, 2017.

BAFU: Hochwasserstatistik Stationsbericht Dünnern – Olten, Hammermüuhle, BAFU Bericht, available at: https://www.hydrodaten.admin.ch/lhg/sdi/hq_studien/hq_statistics/2434_hq_Bericht.pdf (last access: 16 December 2020), 2017.

Beauchamp, J., Leconte, R., Trudel, M., and Brissette, F.: Estimation of the summer–fall PMP and PMF of a northern watershed under a changed climate, Water Resour. Res., 49, 3852–3862, https://doi.org/10.1002/wrcr.20336, 2013.

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

Blazkova, S. and Beven, K.: Flood frequency estimation by continuous simulation for a catchment treated as ungauged (with uncertainty), Water Resour. Res., 38, 14–1–14–14, https://doi.org/10.1029/2001WR000500, 2002.

Blazkova, S. and Beven, K.: Flood frequency estimation by continuous simulation of subcatchment rainfalls and discharges with the aim of improving dam safety assessment in a large basin in the Czech Republic, J. Hydrol., 292, 153–172, https://doi.org/10.1016/j.jhydrol.2003.12.025, 2004.

Boughton, W. and Droop, O.: Continuous simulation for design flood estimation – a review, Environ. Model. Softw., 18, 309–318, https://doi.org/10.1016/S1364-8152(03)00004-5, 2003.

Breinl, K.: Driving a lumped hydrological model with precipitation output from weather generators of different complexity, Hydrolog. Sci. J., 61, 1395–1414, https://doi.org/10.1080/02626667.2015.1036755, 2016.

Brunner, M. I. and Sikorska-Senoner, A. E.: Dependence of flood peaks and volumes in modeled discharge time series: Effect of different uncertainty sources, J. Hydrol., 572, 620–629, https://doi.org/10.1016/j.jhydrol.2019.03.024, 2019.

Brunner, M. I., Seibert, J., and Favre, A.-C.: Bivariate return periods and their importance for flood peak and volume estimation, WIREs Water, 3, 819–833, https://doi.org/10.1002/wat2.1173, 2016.

Brunner, M. I., Viviroli, D., Sikorska, A. E., Vannier, O., Favre, A.-C., and Seibert, J.: Flood type specific construction of synthetic design hydrographs, Water Resour. Res., 53, 1390–1406, https://doi.org/10.1002/2016WR019535, 2017.

Brunner, M. I., Furrer, R., Sikorska, A. E., Viviroli, D., Seibert, J., and Favre, A. C.: Synthetic design hydrographs for ungauged catchments: a comparison of regionalization methods, Stoch. Environ. Res. Risk A., 32, 1993–2023, https://doi.org/10.1007/s00477-018-1523-3, 2018a.

Brunner, M. I., Sikorska, A. E., Furrer, R., and Favre, A.-C.: Uncertainty Assessment of Synthetic Design Hydrographs for Gauged and Ungauged Catchments, Water Resour. Res., 54, 1493–1512, https://doi.org/10.1002/2017WR021129, 2018b.

Brunner, M. I., Sikorska, A. E., and Seibert, J.: Bivariate analysis of floods in climate impact assessments, Sci. Total Environ., 616–617, 1392–1403, https://doi.org/10.1016/j.scitotenv.2017.10.176, 2018c.

Calver, A. and Lamb, R.: Flood frequency estimation using continuous rainfall-runoff modelling, Phys. Chem. Earth, 20, 479–483, https://doi.org/10.1016/S0079-1946(96)00010-9, 1995.

Cameron, D., Beven, K., Tawn, J., Blazkova, S., and Naden, P.: Flood frequency estimation by continuous simulation for a gauged upland catchment (with uncertainty), J. Hydrol., 219, 169–187, https://doi.org/10.1016/S0022-1694(99)00057-8, 1999.

Cameron, D., Beven, K., Tawn, J., and Naden, P.: Flood frequency estimation by continuous simulation (with likelihood based uncertainty estimation), Hydrol. Earth Syst. Sci., 4, 23–34, https://doi.org/10.5194/hess-4-23-2000, 2000.

Chow, V. T., Maidment, D. R., and Mays, L. W.: Applied Hydrology, Mc Graw-Hill, New York, 1988.

Dawson, C., Abrahart, R., and See, L.: Hydrotest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts, Environ. Model. Softw., 22, 1034–1052, https://doi.org/10.1016/j.envsoft.2006.06.008, 2007.

De Michele, C., Salvadori, G., Canossi, M., Petaccia, A., and Rosso, R.: Bivariate Statistical Approach to Check Adequacy of Dam Spillway, J. Hydrol. Eng., 10, 50–57, https://doi.org/10.1061/(ASCE)1084-0699(2005)10:1(50), 2005.

Eagleson, P.: Dynamics of flood frequency, Water Resour. Res., 8, 878–898, https://doi.org/10.1029/WR008i004p00878, 1972.

Evin, G., Favre, A.-C., and Hingray, B.: Stochastic generation of multi-site daily precipitation focusing on extreme events, Hydrol. Earth Syst. Sci., 22, 655–672, https://doi.org/10.5194/hess-22-655-2018, 2018.

Evin, G., Favre, A.-C., and Hingray, B.: Stochastic generators of multi-site daily temperature: comparison of performances in various applications, Theor. App. Climatol., 135, 811–824, https://doi.org/10.1007/s00704-018-2404-x, 2019.

Favre, A. C., El Adlouni, S., Luc Perreault, L., Thiémonge, N., and Bobé, B.: Multivariate hydrological frequency analysis using copulas, Water Resour. Res., 40, W01101, https://doi.org/10.1029/2003WR002456, 2004.

Filipova, V., Lawrence, D., and Skaugen, T.: A stochastic event-based approach for flood estimation in catchments with mixed rainfall and snowmelt flood regimes, Nat. Hazards Earth Syst. Sci., 19, 1–18, https://doi.org/10.5194/nhess-19-1-2019, 2019.

Fischer, S.: A seasonal mixed-POT model to estimate high flood quantiles from different event types and seasons, J. Appl. Stat., 45, 2831–2847, https://doi.org/10.1080/02664763.2018.1441385, 2018.

Fischer, S., Schumann, A., and Schulte, M.: Characterisation of seasonal flood types according to timescales in mixed probability distributions, J. Hydrol., 539, 38–56, https://doi.org/10.1016/j.jhydrol.2016.05.005, 2016.

FOEN: Historical data from the hydrometric stations on Swiss watercourses and lakes, Switzerland, available at: https://www.bafu.admin.ch, last access: 12 February 2020.

Gaál, L., Szolgay, J., Kohnová, S., Hlavčová, K., Parajka, J., Viglione, A., Merz, R., and Blöschl, G.: Dependence between flood peaks and volumes: a case study on climate and hydrological controls, Hydrolog. Sci. J., 60, 968–984, https://doi.org/10.1080/02626667.2014.951361, 2015.

Gabriel-Martin, I., Sordo-Ward, A., Garrote, L., and García, J. T.: Dependence Between Extreme Rainfall Events and the Seasonality and Bivariate Properties of Floods. A Continuous Distributed Physically-Based Approach, Water, 11, 1896, https://doi.org/10.3390/w11091896, 2019.

Gangrade, S., Kao, S. C., Dullo, T. T., Kalyanapu, A. J., and Preston, B. L.: Ensemble-based flood vulnerability assessment for probable maximum flood in a changing environment, J. Hydrol., 576, 342–355, https://doi.org/10.1016/j.jhydrol.2019.06.027, 2019.

Graler, B., van den Berg, M. J., Vandenberghe, S., Petroselli, A., Grimaldi, S., De Baets, B., and Verhoest, N. E. C.: Multivariate return periods in hydrology: a critical and practical review focusing on synthetic design hydrograph estimation, Hydrol. Earth Syst. Sci., 17, 1281–1296, https://doi.org/10.5194/hess-17-1281-2013, 2013.

Griessinger, N., Seibert, J., Magnusson, J., and Jonas, T.: Assessing the benefit of snow data assimilation for runoff modeling in Alpine catchments, Hydrol. Earth Syst. Sci., 20, 3895–3905, https://doi.org/10.5194/hess-20-3895-2016, 2016.

Grimaldi, S., Petroselli, A., and Serinaldi, F.: Design hydrograph estimation in small and ungauged watersheds: continuous simulation method versus event-based approach, Hydrol. Process., 26, 3124–3134, https://doi.org/10.1002/hyp.8384, 2012a.

Grimaldi, S., Petroselli, A., and Serinaldi, F.: A continuous simulation model for design-hydrograph estimation in small and ungauged watersheds, Hydrolog. Sci. J., 57, 1035–1051, https://doi.org/10.1080/02626667.2012.702214, 2012b.

Grimaldi, S., Petroselli, A., Arcangeletti, E., and Nardi, F.: Flood mapping in ungauged basins using fully continuous hydrologic–hydraulic modeling, J. Hydrol., 487, 39–47, https://doi.org/10.1016/j.jhydrol.2013.02.023, 2013.

Gringorten, I. I.: A plotting rule for extreme probability paper, J. Geophys. Res., 68, 813–814, https://doi.org/10.1029/JZ068i003p00813, 1963.

Gupta, H., Kling, H., Yilmaz, K., and Martinez, G.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Hartigan, J. A. and Wong, M. A.: Algorithm AS 136: A K-Means Clustering Algorithm, J. Roy. Stat. Soc. Ser. C, 28, 100–108, 1979.

Hazewinkel, M.: Upper and lower bounds, In: Encyclopedia of Mathematics, edited by: Hazewinkel, M., Springer Science + Business Media B.V./Kluwer Academic Publishers, the Netherlands, 1994.

Hoes, O. and Nelen, F.: Continuous simulation or event-based modelling to estimate flood probabilities?, WIT Trans. Ecol. Environ., 80, 3–10, https://doi.org/10.2495/WRM050011, 2005.

Jost, G., Moore, R. D., Menounos, B., and Wheate, R.: Quantifying the contribution of glacier runoff to streamflow in the upper Columbia River Basin, Canada, Hydrol. Earth Syst. Sci., 16, 849–860, https://doi.org/10.5194/hess-16-849-2012, 2012.

Katz, R. W., Parlange, M. B., and Naveau, P.: Statistics of extremes in hydrology, Adv. Water Resour., 25, 1287–1304, 2002.

Kavetski, D., Fenicia, F., Reichert, P., and Albert, C.: Signature-Domain Calibration of Hydrological Models Using Approximate Bayesian Computation: Theory and Comparison to Existing Applications, Water Resour. Res., 54, 4059–4083, https://doi.org/10.1002/2017WR020528, 2018.

Kochanek, K., Renard, B., Arnaud, P., Aubert, Y., Lang, M., Cipriani, T., and Sauquet, E.: A data-based comparison of flood frequency analysis methods used in France, Nat. Hazards Earth Syst. Sci., 14, 295–308, https://doi.org/10.5194/nhess-14-295-2014, 2014.

Kuchment, L. and Gelfan, A.: Assessment of extreme flood characteristics based on a dynamic-stochastic model of runoff generation and the probable maximum discharge, J. Flood Risk Manage., 4, 115–127, https://doi.org/10.1111/j.1753-318X.2011.01096.x, 2011.

Kundzewicz, Z. W., Krysanova, V., Dankers, R., Hirabayashi, Y., Kanae, S., Hattermann, F. F., Huang, S., Milly, P. C. D., Stoffel, M., Driessen, P. P. J., Matczak, P., Quevauviller, P., and Schellnhuber, H.-J.: Differences in flood hazard projections in Europe – their causes and consequences for decision making, Hydrolog. Sci. J., 62, 1–14, https://doi.org/10.1080/02626667.2016.1241398, 2017.

Lamb, R. and Kay, A. L.: Confidence intervals for a spatially generalized, continuous simulation flood frequency model for Great Britain, Water Resour. Res., 40, W07501, https://doi.org/10.1029/2003WR002428, 2004.

Lamb, R., Faulkner, D., Wass, P., and Cameron, D.: Have applications of continuous rainfall–runoff simulation realized the vision for process-based flood frequency analysis?, Hydrol. Process., 30, 2463–2481, https://doi.org/10.1002/hyp.10882, 2016.

Mediero, L., Jimenez-Alvarez, A., and Garrote, L.: Design flood hydrographs from the relationship between flood peak and volume, Hydrol. Earth Syst. Sci., 14, 2495–2505, https://doi.org/10.5194/hess-14-2495-2010, 2010.

Merz, R. and Blöschl, G.: A process typology of regional floods, Water Resour. Res., 39, W1340, https://doi.org/10.1029/2002WR001952, 2003.

MeteoSwiss: MeteoSwiss ground level monitoring networks, Switzerland, available at: http://www.meteoswiss.ch, last access: 12 February 2020.

Müller-Thomy, H. and Sikorska-Senoner, A. E.: Does the complexity in temporal precipitation disaggregation matter for a lumped hydrological model?, Hydrolog. Sci. J., 64, 1453–1471, https://doi.org/10.1080/02626667.2019.1638926, 2019.

Paquet, E., Garavaglia, F., Garçon, R., and Gailhard, J.: The SCHADEX method: A semi-continuous rainfall–runoff simulation for extreme flood estimation, J. Hydrol., 495, 23–37, https://doi.org/10.1016/j.jhydrol.2013.04.045, 2013.

Pramanik, N., Panda, R., and Sen, D.: Development of design flood hydrographs using probability density functions, Hydrol. Process., 24, 415–428, https://doi.org/10.1002/hyp.7494, 2010.

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B.: Numerical recipes in C++: the art of scientific computing, in: xxvii, 2nd Edn., Cambridge University Press, Cambridge, UK, New York, 2002.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, available at: https://www.R-project.org/ (last access: 16 December 2020), 2019.

Schumann, A. H., Nijssen, D., and Pahlow, M.: Handling uncertainties of hydrological loads in flood retention planning, Int. J. River Basin Manage., 8, 281–294, https://doi.org/10.1080/15715124.2010.512561, 2010.

Schürch, M., Kozel, R., Biaggi, D., and Weingartner, R.: Typisierung von Grundwasserregimen in der Schweiz – Konzept und Fallbeispiele, Gas Wasser Abwasser, 11/2010, 955–965, 2010.

Seibert, J.: Estimation of Parameter Uncertainty in the HBV Model, Hydrol. Res., 28, 247–262, https://doi.org/10.2166/nh.1998.15, 1997.

Seibert, J.: Multi-criteria calibration of a conceptual runoff model using a genetic algorithm, Hydrol. Earth Syst. Sci., 4, 215–224, https://doi.org/10.5194/hess-4-215-2000, 2000.

Seibert, J.: Reliability of model predictions outside calibration conditions, Nord. Hydrol., 34, 477–492, https://doi.org/10.2166/nh.2003.0019, 2003.

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

Serinaldi, F. and Grimaldi, S.: Synthetic Design Hydrographs Based on Distribution Functions with Finite Support, J. Hydrol. Eng., 16, 434–446, https://doi.org/10.1061/(ASCE)HE.1943-5584.0000339, 2011.

Sikorska, A. and Seibert, J.: Appropriate temporal resolution of precipitation data for discharge modelling in pre-alpine catchments, Hydrolog. Sci. J., 61, 1–16, https://doi.org/10.1080/02626667.2017.1410279, 2018a.

Sikorska, A. E. and Seibert, J.: Value of different precipitation data for flood prediction in an alpine catchment: A Bayesian approach, J. Hydrol., 556, 961–971, https://doi.org/10.1016/j.jhydrol.2016.06.031, 2018b.

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

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

Sikorska-Senoner, A. E. and Seibert, J.: Flood-type trend analysis for alpine catchments, Hydrolog. Sci. J., 65, 1281–1299, https://doi.org/10.1080/02626667.2020.1749761, 2020.

Sikorska, A. E., Viviroli, D., and Seibert, J.: Flood-type classification in mountainous catchments using crisp and fuzzy decision trees, Water Resour. Res., 51, 7959–7976, https://doi.org/10.1002/2015WR017326, 2015.

SwissTopo: Vector25 – The digital landscape model of Switzerland, Wabern, Switzerland, available at: http://www.swisstopo.ch, (last access: 12 February 2020), 2008.

Tung, Y. K., Yeh, K. C., and Yang, J. C.: Regionalization of unit hydrograph parameters: 1. Comparison of regression analysis techniques, Stoch. Hydrol. Hydraul., 11, 145–171, https://doi.org/10.1007/BF02427913, 1997.

Viglione, A. and Blöschl, G.: On the role of storm duration in the mapping of rainfall to flood return periods, Hydrol. Earth Syst. Sci., 13, 205–216, https://doi.org/10.5194/hess-13-205-2009, 2009.

Viviroli, D., Kauzlaric, M., Sikorska-Senoner, A. E., Staudinger, M., Keller, L., Whealton, C., Nicolet, G., Evin, G., Raynaud, D., Chardon, J., Favre, A.-C., Hingray, B., Weingartner, R., and Seibert, J.: Estimation of extremely rare floods in a large river basin from continuous hydrometeorological simulations, in: Proceedings of INTERPRAEVENT, 11–14 May 2020, Bergen, Norway, 2020.

Weingartner, R. and Aschwanden, H.: Abflussregimes als Grundlage zur Abschätzung von Mittelwerten des Abflusses in Hydrologischer Atlas der Schweiz, available at: https://hydrologischeratlas.ch/produkte/druckausgabe# (last access: 16 December 2020), 1992.

Westerberg, I. K., Sikorska-Senoner, A. E., Viviroli, D., Vis, M., and Seibert, J.: Hydrologic Model Calibration with Uncertain Discharge Data, Hydrolog. Sci. J., https://doi.org/10.1080/02626667.2020.1735638, in press, 2020.

Winter, B., Schneeberger, K., Dung, N., Huttenlau, M., Achleitner, S., Stötter, J., Merz, B., and Vorogushyn, S.: A continuous modelling approach for design flood estimation on sub-daily time scale, Hydrolog. Sci. J., 64, 539–554, https://doi.org/10.1080/02626667.2019.1593419, 2019.

Zeimetz, F., Schaefli, B., Artigue, G., García Hernández, J., and Schleiss, A.: A new approach to identify critical initial conditions for extreme flood simulations based on deterministic and stochastic simulation, J. Hydrol. Eng., 23, 04018031, https://doi.org/10.1061/(ASCE)HE.1943-5584.0001652, 2018.

Zhang, L. and Singh Vijay, P.: Trivariate Flood Frequency Analysis Using the Gumbel–Hougaard Copula, J. Hydrol. Eng., 12, 431–439, https://doi.org/10.1061/(ASCE)1084-0699(2007)12:4(431), 2007.