Clustering model responses in the frequency space for improved simulation-based flood risk studies: The role of a cluster number

Anna E. Sikorska-Senoner

Abstract
Hydrologic models are often employed in flood risk studies to simulate possible hydrologic responses. They are, however, linked with uncertainty that is commonly represented with uncertainty intervals constructed based on a simulation ensemble. This work adapts an alternative clustering-based approach to first, learn about hydrological responses in the frequency space, and second, select an optimal number of clusters and corresponding representative parameters sets for a hydrologic model. Each cluster is described with three parameter sets, which enable percentile and prediction intervals to be constructed. Based on a small Swiss catchment with 10,000 years of daily pseudo-discharge simulations, it was found that clustering the ensemble of 1000 members into 5–7 groups is optimal to derive reliable flood prediction intervals in the frequency space. This lowers the computational costs of using a hydrological model by 98%. The developed approach is suitable for probabilistic flood risk analysis with current or future climate conditions to assess hydrologic changes.

KEYWORDS
cluster analysis, ensemble, flood risk, frequency space, hydrologic model, k-means, representative parameter sets, uncertainty

1 | INTRODUCTION
Probabilistic flood risk assessment (PFRA) requires that several components in the complex chain between climate conditions, hydrologic discharge, inundation, and impact are considered (Ward et al., 2014). As each of these chain components is associated with uncertainty, represented with a large number of ensemble members (or scenarios), the dimension of possible options to be considered can increase very quickly. Within such a PFRA chain, continuous hydrologic models are often employed to simulate a possible hydrological response to a considered input and to estimate the extent of possible floods. In place of these hydrologic models, conceptual models are often employed as they are very fast in simulation and thus enable the use of long time series, many meteorological scenarios or different states of initial conditions (Blazkova & Beven, 2004; Cameron et al., 1999; Haberlandt & Radtke, 2014; Paquet et al., 2013; Winter et al., 2019; Zeimetz et al., 2018). They thus appear to be more suitable for PRFA than physically based models that, although do not need a direct calibration (Cunha et al., 2011), are slow in simulation. Such conceptual models require calibration with observed data and are...
thus linked with uncertainty (Renard et al., 2011; Sikorska & Renard, 2017; Sikorska & Seibert, 2018). Hence, it has been recommended against using only one parameter set for a hydrologic model (Beven & Freer, 2001; McIntyre et al., 2002; Seibert et al., 2018). This parametric uncertainty should be accounted for in flood risk management (Shah et al., 2018), as it can be an important uncertainty source in design floods (Brunner & Sikorska-Senoner, 2019).

A common way to consider uncertainty of hydrologic models is through resampling from the optimized parameter space (Sikorska & Renard, 2017) or via using multiple parameter sets derived from independent calibration runs (Westerberg et al., 2020). While the first approach requires defining a likelihood function to sufficiently sample from the posterior (Montanari & Koutsouyiannis, 2012; Sikorska et al., 2015), the latter approach requires running multiple calibration runs to minimize the risk of being trapped in a local minimum (Sikorska-Senoner et al., 2020). The set of resulting model simulations can be used as an ensemble of hydrologic responses.

Yet, in both cases, it is not clear how many parameter sets should be resampled to best represent the variability of the hydrologic ensemble. A common practice is to use at least 100 parameter sets. Use of more than 1000 parameter sets is rare due to computational limitations. For PFRA, these parameter sets must be further analyzed in the context of other variable uncertainty (or possible scenarios) and thus, the total number of model runs to be performed may quickly become very large. These simulation numbers further increase with additional settings to be considered (initial conditions, land use profile, etc.) for which hydrologic simulations must be performed. Furthermore, if long time series of meteorological inputs (≥100 years) must be simulated, running all model parameter sets may become prohibitive (Grimaldi et al., 2013).

While a reduction of meteorological scenarios or initial state condition set-ups is not straightforward due to a nonlinear input–output relationship in hydrologic models, a reduction of the parameter sample size seems to be a conceptually easier way to lower the computational needs of simulation-based PFRA. Recently, Sikorska-Senoner et al. (2020) proposed a very promising approach for such a parameter ensemble downsizing via direct analysis of hydrologic responses in the application space (flood frequency space [FFS]). Based on this, representative parameter sets (RPSs) that best represent the full hydrologic ensemble can be selected. Among three tested methods, the k-means clustering approach has been recommended due to its lowest bias and highest performance. Yet, the authors limited their analysis to the case with three arbitrary selected clusters and, based on these, selected only three RPSs that represent the upper, middle, and lower intervals of modeling uncertainty (one RPS per cluster). Although the number of three appears to be a good choice for representing the uncertainty ranges along the median simulation, it remains unclear how increasing the number of clusters (k) and the number of RPSs improves the identification of the simulation ensemble. It also remains unclear whether any optimal k number can be found and how k and RPSs relates to different prediction percentiles. In this work, the k-means clustering approach is adapted with the goal to test its sensitivity to different values of k and its relation to the return period. The two specific research questions are: (i) How many clusters (and RPSs) are needed to optimally represent the uncertainty intervals in the FFS while keeping model computational costs low? and (ii) Does this optimal number of clusters depend on the flood return period or the prediction percentile?

2 | METHODS

2.1 | Study motivation: Why clustering?

Clustering involves partitioning of data into groups (clusters) based on some similarity measures. As a result of clustering, elements belonging to the same cluster are more similar to each other than to the elements from other clusters (Notaro et al., 2018). Clustering has been continuously gaining in popularity in hydrologic applications and has been successfully applied to describe catchment functionality (Jehn et al., 2020; Kuentz et al., 2017), to regionalization and predictions in ungauged catchments (Rao & Srinivas, 2006; Singh et al., 2016), or for reducing model computational requirements (Ehret et al., 2020; Notaro et al., 2018; Sikorska-Senoner et al., 2020). The latter work has been the first application of the cluster analysis in the FFS.

In this work, the clustering approach developed by Sikorska-Senoner et al. (2020) is further adapted to analyze the role of the cluster number k in deriving reliable uncertainty estimates in the FFS for hydrologic-model based simulations within PFRA. A clustering is involved to search for similarities in hydrologic responses in the FFS and thus to learn about the hydrologic model behavior, and next to select the RPSs. By learning from a training data set, one can better describe the variability of simulation ensemble in the context of new applications or settings.

2.2 | Clustering of hydrological responses in the FFS

One of the most widely applied clustering methods is the k-means (MacQueen, 1967) due to its computational
efficiency and ease in interpreting the results (Sharghi et al., 2018), and is also employed in this study for clustering hydrologic responses following Sikorska-Senoner et al. (2020). A hydrologic model response is understood as a single deterministic model-based discharge simulation, with one ith parameter set of a hydrologic model that consists of J simulation years. A model simulation ensemble thus represents a set of possible model responses resulting from single model runs with different parameter sets. For each of these simulations, floods with the largest flood peak within each ith simulation year are selected, creating the vector with J annual maxima (AMs). These AMs are next passed to the cluster analysis, adapting the approach of Sikorska-Senoner et al. (2020).

The clustering steps of the current approach are summarized as follows:

a. For each ith parameter set (i = 1, 2, ..., I), AMs computed with this parameter set are sorted by their magnitude over the entire simulation period, creating I ensemble members of J sorted AMs.

b. These ensemble members are next clustered into k groups (clusters) based on all J simulation years using the k-means clustering.

c. Next, these k clusters are sorted by their magnitude based on the clusters' means by comparing the percentiles in the upper tail of the distribution (e.g., a 90th percentile).

d. For each kth cluster, the lower, middle, and the upper interval are computed that correspond to the 5th, 50th, and 95th percentiles. These percentiles are assumed to represent the cluster.

e. For each cluster, a search is made for the RPSs that correspond to simulations lying closest to each of these percentiles using the Euclidean distance (ED) as an agreement measure. Parameter sets that receive the smallest value of ED are selected as RPSs for the corresponding percentile under consideration. Thus, three RPSs per cluster are retained.

f. Next, the hyper-percentiles over all simulation ensembles are attributed by computing the 5th, 50th, and 95th percentiles based on selected simulations corresponding to the subset of RPSs.

g. Finally, the hyper RPSs are selected (from the available RPS subset) that correspond to the hyper-intervals and are assumed to describe the entire ensemble simulation (i.e., the 90% predictive intervals [90% PIs] along with the median value).

h. The entire analysis is redone with different k-values.

Thus, the major differences between the current study and Sikorska-Senoner et al. (2020) are, first, the k-value here is a variable that is defined based on the sensitivity analysis (see Section 2.6) whereas it was set arbitrarily to the value of 3 in the previous work, and second, each cluster is defined with three RPSs instead of only one RPS.

In this study, clustering is done using the Hartigan-Wong algorithm (Hartigan & Wong, 1979), as implemented in the function kmeans from the package “stats” (R Core Team, 2019).

### 2.3 Characterizing the interval variability with different percentiles

For some practical applications, more information on the interval variability may be desired than only the 90% PIs. For instance, 68% PIs and 95% PIs are a common choice in discharge uncertainty analysis (Kiang et al., 2018). To extract such additional information on the simulation ensemble, the clustering approach is extended to compute the following percentiles, that is, 0.01, 0.1, 1, 2.5, 5, 10, 16, 20, 50, 80, 84, 90, 95, 97.5, 99, 99.1, and 99.91.

These percentiles were chosen as they allow the following PIs to be derived: 99.9%, 99%, 98%, 95%, 90%, 80%, 68%, and 60% prediction intervals, which are a common choice in uncertainty analysis (Westerberg et al., 2020). The computation of these percentiles and the selection of corresponding RPSs is done by extending the steps (f)–(g) of the clustering approach in Section 2.2 to a computation of these additional hyper-percentiles. As a result, one RPS for each percentile is chosen from the subset of RPSs. Note that the same RPS can be chosen for more percentiles.

### 2.4 Benchmark

To evaluate the results from the clustering approach, a benchmark is needed that is defined as the full simulation ensemble that contains AMs generated with 1000 parameter sets. This benchmark can be seen as a 1000-cluster case, when each ensemble member is assigned to its own cluster and thus selection of the percentiles and attribution of the RPSs is based upon the entire ensemble.

### 2.5 Reference space: FFS

Clustering is performed independently from the selected FFS. For visualizing the cluster analysis in the FFS, however, the Gumbel space is applied (generalized extreme value distribution type-I) with the method of Gringorten (1963) to compute the plotting positions:
where $x_j$ is a plotting position for the $j$th (sorted) AM in the Gumbel space. The relationship between the $x_j$ and the probability $p_j$ of the event occurrence is simply:

$$p_j = 1 - x_j$$

Knowing $p_j$, the return period of the event $T_j$ can be computed:

$$T_j = \frac{1}{1 - p_j}$$

### 2.6 Sensitivity analysis to the number of clusters

$k$-means clustering has only one parameter, the $k$ value, which is the number of clusters. In the work of Sikorska-Senoner et al. (2020), this number was set arbitrarily to 3. Having a higher number of clusters should theoretically provide more insights into the behavior of the simulation ensemble but is linked with a higher computational cost (more RPSs are retained). Hence, choosing the proper $k_0$ value can be seen as a search for the number of clusters that maximizes the information contained in the simulation ensemble (with multiple RPSs) and minimizes the model computational costs. Note that, in that sense, $k_0$ does not stand for the optimal partitioning of the data, as it does in the classical approach (Rao & Srinivas, 2006), because it is also conditioned on the computational costs of a hydrological model.

To identify $k_0$, a sensitivity analysis is performed in which the $k$-value varies between 1 and 1000 (steps b–h in Section 2.2), where 1 corresponds to the one-cluster case and 1000 corresponds to the benchmark. Only $k$-values that could be justified for practical purposes are considered and thus the following $k$-values are explored: $k = c[1, 2, 3, 5, 7, 10, 20, 50, 100, 200, 500]$.

### 2.7 Evaluation criteria

The ED is computed as a measure to assess how well the percentiles from the clustering approach are able to replicate the benchmark percentiles.

The overlapping pools of prediction intervals (OPPIs) (Sikorska-Senoner et al., 2020), are computed in the FFS by taking the variate of the FFS, here the Gumbel variate, and discharge values of prediction intervals (PIs), that is, sorted AMs, as coordinates of the pools. Hence, the OPPIs index measures the overlapping of the pool computed based on PIs resulting from the cluster analysis and the reference pool (benchmark PIs). The area of each PIs pool ($A_{OPPIs}$) is computed as (Figure 1a):

$$A_{OPPIs} = \sum_j \frac{1}{2} (\text{PI}_{Uj} + \text{PI}_{Uj-1} - \text{PI}_{Lj} - \text{PI}_{Lj-1}),$$

$$j = 2, 3, \ldots, J$$

where $x$ is the Gumbel variate, $j$ is the index of sorted AMs, and PI$_{U}$ and PI$_{L}$ are the upper and lower intervals of the PIs. By dividing the area of the cluster-based pool ($A_{OPPIs,C}$) by the area of the benchmark pool ($A_{OPPIs,B}$), the relative value can be derived:

$$\Delta A_{OPPIs} = \frac{A_{OPPIs,C}}{A_{OPPIs,B}}$$

The best coverage is achieved for $\Delta A_{OPPIs}$ equal to 1, meaning that both pools perfectly overlap. Values $<1$ and $>1$ for $\Delta A_{OPPIs}$ mean that the pools are underrepresented (smaller than the benchmark) or overestimated (larger than the benchmark). This metric is computed for all PIs considered here.

#### 2.7.1 Sensitivity measure: Slope stability

As the tested $k$-values of clusters are not linearly increasing, a selection of the $k_0$ is not straightforward. To support such analysis, the slope stability of the curves, constructed as the relationship between the ED and the $k$-values, is computed. This slope stability (ratio of ratios of ED) corresponds to the second derivative of ED and is a function $f$ of the derivative of the derivative of ED, that is, $f(f'(ED))$. Hence, it measures how the rate of change of ED is itself changing with an increase in the $k$-value. The number $k_0$ can be identified as the point at the $f''$ (ED) curve where the second derivative reaches its first local maximum (Figure 1b).

### 3 EXPERIMENTAL SETUP

#### 3.1 Study catchment, data set, meteorological scenarios, and hydrologic model

The clustering approach is tested on the catchment of Dünnern stream at Olten (Switzerland) that has an area of 196 km$^2$ and an elevation span from 400 to 760 m
above sea level (see Figure 2). For this catchment, 25 years of continuous observations on discharge, precipitation, temperature, and evaporation were available for years 1989–2014 at an hourly time step and were aggregated to a daily time step for the purpose of this study.

For Dünnerg, 100 meteorological scenarios of 100 years of continuous precipitation and temperature time series generated using a weather generator were available from the previous study (Sikorska-Senoner et al., 2020). Similar to the previous study, a continuous HBV-light model (Seibert & Vis, 2012) with 15 adjustable parameters is used to simulate pseudo-observed discharge time series from meteorological scenarios. However, in the current study the model is run at a daily time step, in contrast to the previous study which was run at an hourly time step. Thus, the model parameters require recalibration at a new temporal resolution (Table 1). The reason to run the model at a daily time step is the possibility to analyze more parameter sets (1000 vs. 100) while preserving the realism of flood simulations for that catchment.

**Figure 1** Explanation of the overlapping pool computation (a) and defining an optimum number ($k_0$) of clusters (b) based on the second derivative.

**Figure 2** Study catchment, Dünnerg at Olten in Switzerland. The figure is reproduced from Sikorska-Senoner et al. (2020) under a CC BY 4.0 license.
3.2 | Delivery of a heterogenic parameter sample and hydrologic ensemble

To generate an ensemble of parameter sets, a heuristic approach proposed by Sikorska-Senoner et al. (2020) is also applied here that uses multiple independent calibration trials. In addition, a multiperiod calibration is performed that consists of four different subperiods, as indicated in Table 2.

The parameter Sample A overlaps with the calibration period used in Sikorska-Senoner et al. (2020), while Sample B uses the validation period of the Sample A for calibrating the model. In Sample C, the model is calibrated on the entire observation period (1990–2014), while Sample D uses only 2 years with the highest flood events (2006–2007) to infer model parameters. Calibrating the model with different subperiods covers all ranges of different conditions from the observation period.

For all calibration periods, a 1-year warm-up period is used to set the catchment initial conditions. For each calibration strategy, 250 model trials are run resulting in 250 optimized parameter sets. The optimization algorithm used is a genetic algorithm (GA) as developed by Seibert (2000) and a multiobjective function (MOF) as in Sikorska-Senoner et al. (2020), that is, using the Kling-Gupta efficiency (Gupta et al., 2009), with efficiency for flow peaks and Mare measure weighted as 0.3, 0.5, and 0.2. This MOF has been tested in other studies (Sikorska et al., 2018; Sikorska-Senoner et al., 2020; Westerberg et al., 2020). A GA search is initiated by randomly selecting parameter sets within the given ranges (Table 1) and is terminated at the maximal number of interactions allowed (3000). Finally, the optimized parameter sets from all four calibrations emerge into one parameter ensemble with 1000 sets, which is next used for generating pseudo-observations of discharge time series with 1000 members for each meteorological scenario.

4 | RESULTS

4.1 | Characteristics of the parameter sample and model performance

The optimized 1000 parameters are presented in Figure 3. For some parameters (i.e., KO, CFR, CWH, PERC, BETA), the resulting violin plots indicate the existence of different optimal solutions selected during four calibration periods (seen by stretched violin plots), whereas for all other parameters, the calibration strategies led to similar optimal solutions (seen by smoothed violin plots).

The hydrologic ensemble resulting from the parameter ensemble is illustrated in Figure 3 together with observed values. The KGE criteria was from 0.82 to 0.85 in the calibration and from 0.70 to 0.83 in the validation for all four calibration periods (median over all 250 runs) (right panels in Figure 4).

Overall, the model performs well in both calibration and validation periods during all four strategies. However, the performance for peaks drops slightly in the validation for individual strategies as compared to the calibration period (while KGE, MARE, and NSE stay at similar levels). Thus using only one calibration period (e.g., Sample A)

### Table 1: HBV parameter description and ranges for model calibration at a daily time step

| Parameter | Unit | Min | Max | Description |
|-----------|------|-----|-----|-------------|
| PERC      | mm d⁻¹ | 0   | 4   | Percolation parameter |
| UZL       | mm   | 0   | 70  | Groundwater runoff threshold parameter |
| K0        | d⁻¹  | 0.2 | 0.5 | Recession coefficient; K0 > K1 |
| K1        | d⁻¹  | 0.01| 0.3 | Recession coefficient; K1 > K2 |
| K2        | d⁻¹  | 0.00001| 0.1 | Recession coefficient |
| MAXBAS    | d    | 1   | 10  | Length of triangular weighting function |
| CET       | °C⁻¹ | 0   | 0.5 | Correction factor for potential evaporation |
| TT        | °C   | −2.5| 2.5 | Threshold temperature |
| CFMAX     | mm d⁻¹ °C⁻¹ | 0.5 | 6   | Degree-day factor |
| SFCF      | —    | 0   | 2   | Snowfall correction factor |
| CFR       | —    | 0   | 0.2 | Refreezing correction factor |
| CWH       | —    | 0   | 0.2 | Water holding capacity |
| FC        | mm   | 100 | 600 | Maximum moisture storage in soil box |
| LP        | mm   | 0   | 1   | Threshold for reduction of evaporation |
| BETA      | —    | 1   | 10  | Shape coefficient |
**TABLE 2** Calibration strategies for the HBV model

| Parameter sample | Calibration period | Validation period | Notes               |
|------------------|--------------------|-------------------|---------------------|
| A                | 1990–2005          | 2006–2014         |                     |
| B                | 2006–2014          | 1990–2006         |                     |
| C                | 1990–2014          | —                 | No validation possible |
| D                | 2006–2007          | 1990–2014         |                     |

**FIGURE 3** Box plots of the calibrated parameters (violin plots) versus calibration ranges (gray polygons). Units as in Table 1
would have led to an underestimation of large flood peaks in the validation phase, which was richer in large floods. This greatly justifies the need for a multiperiod calibration and is further discussed in Section 5.1.

### 4.2 Representation of the parameter uncertainty with RPSs versus the number of clusters

The hydrologic ensemble of all AMs resulting from 1000 optimized parameter sets for each meteorological scenario were the basis for applying the clustering approach to the hydrological responses in the FFS and selecting RPSs. The number of RPSs selected varied depending on the \( k \) number. The effect of the \( k \) value on the performance of such selected RPSs versus the benchmark (full parameter ensemble from Section 4.1) is presented in Figure 5 for the demonstration scenario (median).

Overall, it appears that for a small value of \( k \) (\( k < 5 \)), RPSs do not represent well the full parameter ensemble as the coverage of the resulting RPS parameter range and benchmark is very limited. With an increasing value of \( k \), the coverage of the RPS-based range versus the benchmark improves. For most model parameters, the value of \( k = 7 \) seems to provide a good representation of the full

---

**FIGURE 4** Flow duration curves and model performance in the calibration and validation periods
FIGURE 5 The variability of the parameter ensemble for a varying number of clusters (k) illustrated for one demonstration (median) scenario. B corresponds to the benchmark, that is, the full parameter ensemble containing 1000 parameter sets originating from the model calibration. The red dots indicate the median value of the RPSs, and the black dot, the median value of the benchmark.
ensemble variability, whereas for other parameters (i.e., MAXBAS, CFMAX) the parameter ensemble should optimally be represented with 20–200 clusters. Note that for these two parameters, the distribution of the optimized parameter ensemble is strongly asymmetric whereas the RPSs appear to be selected close to the median value of the benchmark. Similar patterns were observed for all other scenarios.

4.3 Uncertainty prediction percentiles for different numbers of clusters

Percentiles computed for different \( k \) numbers are pictured in Figure 6 for the demonstration scenario together with the benchmark. As can be seen, depending on the \( k \)-value chosen for clustering the simulation ensemble, the resulting percentiles are attributed in a different way. While for the benchmark, the percentiles take smoother curves and the PIs cover a larger span of the full ensemble particularly for wider PIs (>95% PIs), for a small number of clusters (1–3) the same simulations and RPSs are chosen to represent different percentiles. For the one-cluster case, three simulations and three RPSs represent all percentiles considered. It can be also seen that the prediction percentiles are spread asymmetrically within the full ensemble for all \( k \)-values, as judged by the fact that the 50th percentile does not lie in the middle of the simulation ensemble.

Moreover, it can be seen that different realizations and RPSs are chosen as best representing the percentiles depending on the \( k \)-value. This attribution is based on minimizing the ED between the model realization and the percentile computed based on the benchmark. As also seen from Figure 7, ED values vary for different percentiles and for different \( k \)-values (\( k = \{1, 2, ..., 500\} \)) and return periods.

Note that the benchmark (\( k = 1000 \)) is not plotted, as ED is computed with reference to this benchmark. Generally, for each return period considered, the same patterns can be observed. Taking the return period of 2 years (\( T \geq 2 \)) as an example (top right in Figure 7), it can be seen that ED is higher for the percentiles from the lower and upper tails of the distribution. The distances are smallest for the median and for the percentiles located at the lower site of the median. Comparing different \( k \)-values, ED decreases when moving from smaller \( k \) values (1) toward larger \( k \) values (500). Inspecting different return periods, ED is larger for smaller return periods and smaller for higher return periods. It also seems that percentiles from the upper tail are more difficult to be correctly attributed than the percentiles from the lower tail of the distribution, particularly for return periods smaller than 10 years. Generally, it appears that for all return periods increasing the \( k \)-value leads to decreasing ED. Yet, based on this assessment \( k_O \) cannot be defined.

4.4 Sensitivity of the prediction intervals to the \( k \)-value

The computed overlapping pools are presented in Figure 8 for eight PIs considered. The values are presented in reference to the pools computed for the benchmark, that is, as \( \Delta A_{OPPIs} \). It can be seen that \( \Delta A_{OPPIs} \) tend to reach the value of 1, the more we move from smaller \( k \)-values toward the larger \( k \)-values. For narrower PIs (60%–80%), the values of \( \Delta A_{OPPIs} \) are close to 1 for \( k \geq 7 \). For the 90% PIs, 2–3 clusters provide a good overlap with the benchmark, whereas for wider intervals (95%–99.9%) the relative values are a bit lower than 1, even for \( k = 500 \). However, after about 7–10 clusters, the increase in the relative values is no longer visible and thus having more clusters (and more RPSs) than this does not improve the overlap. It appears though that for intervals 80%–95%, a good overlap can be achieved when using cluster analysis, whereas for wider uncertainty intervals, a slight underestimation of the PIs is observed for all \( k \)-values but it becomes most pronounced for \( k < 7 \). For narrow PIs (60% or 80%), the one-cluster case leads to much wider uncertainty bands than the benchmark.

These findings show, first, that increasing the cluster number beyond 7 does not necessarily increase the overlap of the PIs with the benchmark, and second, that PIs wider than 95% are underestimated irrespective of the \( k \)-value (see further discussion).

4.5 Optimal number of clusters

To define the \( k_O \), a second derivative of ED was computed with reference to \( k \). The resulting \( k_O \) are presented in Figure 9 for different return periods and percentiles.

This \( k_O \) value was found to be between 3 and 20 for all percentiles and all return periods. Overall, smaller \( k_O \) values were defined for percentiles from the upper tail of the distribution, and larger \( k_O \) values for percentiles from the lower tail of the distribution. However, \( k_O \) of 7–10 was also found for the percentile lying close to the median values (16th). It can be generally said
**FIGURE 6**  Percentiles corresponding to the selected parameter sets based on clustering of hydrologic responses versus the number of clusters ($k$) used to describe the full simulation ensemble (marked in gray). The results are presented for one demonstration (median) scenario.
that for rare floods (with a return period \( \geq 10 \) years), the simulation ensemble should be clustered into 3–7 clusters, with 3–5 for common PIs, that is, 90% or 95%.

For frequent floods (with a return period <10 years), 10–20 clusters should be defined. Thus, it appears that \( k_O \) is sensitive to the return period.
FIGURE 8 Relative overlapping pools ($\Delta A_{\text{OPPIs}}$) for different predictive intervals (PIs) versus the number of clusters ($k$) for all scenarios. The median value is marked in blue.
5 | DISCUSSION

5.1 | Value of the multiperiod calibration for flood risk studies

In this study, a multiobjective calibration was coupled with a multiperiod approach in order to extract the most information from the observed data and to derive an optimized parameter ensemble. The results show that such a multiperiod calibration is very beneficial for flood studies because it increases the reliability of large flood simulations. In contrast, using a single-period calibration, even if combined with multicriteria, may lead to underrepresentation of large floods, particularly if the validation period is richer in floods or contains unusual floods (i.e., events with extraordinary patterns). The split of the observation period into calibration and validation phases always raises the question of how such a split should be performed (Liu et al., 2018), and has a risk of falling into split-sample pitfalls, that is, where good parameter sets may be discarded (Arsenault et al., 2018). Thus, using a multiperiod calibration overcomes this issue by minimizing the effect such a selection may play in the derivation of the optimized parameter sets. Along the same line, Singh and Bardossy (2012) have recommended calibration using the entire observation period or calibrating the model with unusual events. Yet, none of these calibration strategies guarantee that all possible flood patterns can be captured with optimized parameter sets, particularly if unusual events occur in the application period only. This is a general calibration issue for any hydrological or environmental model which is then applied to a new period of very different patterns or for extreme flood simulations. Nevertheless, combining a multiperiod calibration with a long enough observation period may greatly minimize this risk, but cannot exclude it.

5.2 | Search for RPSs and an optimal $k$-value

The $k$-value in the k-means clustering is the only variable that must be determined by the user prior to data clustering. The choice of this number often depends on subjective preferences or practical issues. In this study, an optimal number $k_o$ is searched for via optimizing the information contained in the resulting RPS-based simulations and the computational effort of running these simulations. Thus, this paper provides an objective way of defining $k_o$ and RPSs for all simulation-based flood risk studies. The importance of defining $k_o$ has been shown through the analysis of the RPSs versus the parameter ensemble and the RPS-based simulations versus the simulation ensemble in the FFS.

Findings from the case study reported that RPSs represent well the entire parameter ensemble but their representability depends largely on the $k$-value. It was found that seven clusters (with 21 RPSs) already provides a good representation of the full parameter ensemble for most of the model parameters. Less than five clusters (<10 RPSs) demonstrated a very poor representation of the parameter ensemble. Similar findings were found from the analysis of the hydrological simulation ensemble based on ED which reported that using 3–7 clusters (with corresponding RPSs) already provides reliable information on the simulation ensemble. If only commonly applied prediction intervals in hydrology (90% or 95%) are of interest, even 3–5 clusters are enough. Note that the number of final RPSs selected depends on the number of clusters (three RPSs per cluster) but also on the detail required. If only PIs are of interest, three hyper RPSs can be selected from the subset of all RPSs. The number of three is also a common choice in practice for
communicating uncertainty in flood analysis and corresponds to the median value along with prediction limits (Lamb & Kay, 2004) and has also been adapted by Sikorska-Senoner et al. (2020). One has to keep in mind however that using three sets only does not contain any information on the PIs intervariability and whether they are symmetric or not. If this information is required, all RPSs retained for all clusters should be analyzed.

It was also found that increasing the number of clusters above seven does not necessarily improve the computation of PIs, nor percentiles, because their accuracy may even slightly drop. This finding can be explained by the fact that having too many clusters may split the ensemble sample into artificial groups that potentially have some similarity with other groups. Hence, a learning process from the training data set is suboptimal and a selection of the RPSs may result in suboptimal solutions, while the computational efforts will increase with increasing number of clusters. Thus, the optimal $k$-value of 5–7 clusters can be recommended for similar frameworks or settings as studied here. Because this optimal $k_o$ value was defined using only a single catchment, it cannot be recommended for a general use. Thus, for all other settings, an optimal $k_o$ value should be defined via the clustering method developed here. To generalize the value of $k_o$, a large sample study with many catchments of different properties should be performed. These findings highlight the need to define an optimal cluster number and the value of an objective criteria-based selection of the cluster number.

### 5.3 Methodological issues

One methodological issue is linked with the way in which RPSs are selected, that is, as representing the 90% PIs per cluster following Sikorska-Senoner et al. (2020). As seen, percentiles lying in the FFS very far from these intervals were more poorly simulated than those lying closer to the 90% PIs. Hence, for a more accurate simulation of these intervals, a wider coverage of clusters may be required (e.g., use of 95% or 98% instead of 90%). The major limitation arises however from the need of having long-enough (pseudo-)discharge observation records that cover very different flood conditions for a sufficient learning process on model responses and deriving the parameter ensemble. Only then, can reliable RPSs be chosen. Optimally, this data should contain floods of different types (Sikorska-Senoner & Selbert, 2020), from different seasons (Fischer, 2018), and for unusual events (Singh & Bárđossy, 2012) for best learning. Covering a large spectrum of possible rare floods ensures the reliability of PIs for different flood conditions not studied in the training dataset. A multiperiod calibration can be recommended to extract the most information on flood behaviors from the entire observation data, as discussed in Section 5.1.

### 5.4 Practical value of the clustering approach for flood studies

The proposed clustering approach with a search for an optimal $k$-value is independent from the hydrologic model applied and the case study. The approach aims at selecting RPSs through clustering of model responses in the FFS. Selected RPSs are obviously linked to the model used, and the reliability of obtained PIs will thus depend on the model performance obtained on the training dataset. It has been found that 5–7 clusters with 15–21 RPSs selected out of 1000 sets provide an optimal representation at the parameter ensemble and at the simulation ensemble levels in the FFS. As such, the clustering proposed here appears to be capable of selecting RPSs for flood risk studies to provide highly reliable uncertainty estimates at low computational costs. Its clear advantage is that it requires the clustering and the selection of an optimal $k$-value and corresponding RPSs to be performed only once, whereas further applications can profit directly from these preselected RPSs. Hence it helps to greatly reduce the computational costs linked to the use of a hydrologic model within complex simulation-based flood risk analysis with many modeling components included. How much this reduction can be achieved depends however on the detail on the ensemble required. Thus, the number of RPSs may vary between 3 (only PIs and median) and 21 (if 7 clusters with all RPSs are considered). In this study, one simulation run of one meteorological scenario with 100 simulation years at a daily time step and one parameter set of a hydrologic model equated to 1.61 s on 1 CPU. Thus, with a reduction of 98% of the total computational time, the computational time of simulating one scenario could be lowered by approximately 0.44 h. The reduction in the total computation time will be much higher for complex distributed hydrologic models. There is thus great potential for all continuous-based modeling approaches for which long-period simulations of more than 100 years are still rare because of computational limitations (Grimaldi et al., 2013).

The proposed clustering-based approach can be applied everywhere where computational power is an issue, that is, in PFRAs, hazard scenario analysis, climate scenario analysis, or multiple-site analysis. Further possible applications include: uncertainty frameworks where hydrologic model uncertainty is propagated through a modeling chain, as it decreases the number of hydrologic simulations (i.e., using 3–21 RPSs instead of 1000);
uncertainty frameworks using (flood) hydrologic signatures (Sikorska-Senoner 2021); or hydrological forecasting where parameter sets preselected off-line could be directly used for speeding-up the real-time computations. Reducing the number of parameter sets of a hydrological model to a small subset of 21 RPSs also enables climate impact studies to focus on other uncertainty sources, that is, those related to the use of emission scenarios, global climate models, downscaling methods, and the natural variability of the climate (Troin et al., 2018).

6 | CONCLUSIONS

In this work, a clustering-based selection of RPSs of a hydrologic model has been developed. The clustering is used as a learning tool to first select an optimal number of clusters in an objective criteria-based way, and then to select corresponding RPSs, and is performed in the FFS on sorted AMs. Having RPSs retained enables the direct application for constructing uncertainty intervals and percentiles for new modeling setups at much lower computational costs. Thus they support uncertainty analysis in all complex flood modeling chains. Based on a small Swiss catchment with 10,000 years of daily pseudo-discharge simulations used as a training data set, it was found that as few as 5–7 clusters are optimal to derive uncertainty information in the frequency space. The number of RPSs retained depends on the detail of the simulation ensemble needed (only ranges or also inter-variability) and may vary from 3 to 21 sets. The proposed approach is suitable for probabilistic flood risk analysis with current or future climate conditions to assess hydrologic changes or possible hydrologic responses with hydrologic models at low computational costs.

ACKNOWLEDGMENTS

The research was funded through the Forschungskredit of the University of Zurich, grants no. FK-18-118 and FK-20-124. The ScienceCloud provided by S3IT (University of Zurich) enabled simulations to be run on virtual machines. The author wishes to give special thanks to Jan Seibert for fruitful discussions, Marc Vis for setting-up virtual machines, and Tracy Ewen for proofreading the manuscript. The associate editor and two anonymous reviewers are acknowledged for providing very useful comments to the earlier version of this manuscript.

DATA AVAILABILITY STATEMENT

The implementation details on the clustering approach are provided in the Supporting Information. The observed discharge data can be ordered from the FOEN (https://www.bafu.admin.ch), the meteorological data from MeteoSwiss (http://www.meteoswiss.ch). The HBV-light model can be downloaded from https://www.geo.uzh.ch/en/units/h2k/Services/HBV-Model.html.

ORCID

Anna E. Sikorska-Senoner https://orcid.org/0000-0002-5273-1038

REFERENCES

Arsenault, R., Brissette, F., & Martel, J.-L. (2018). The hazards of split-sample validation in hydrological model calibration. Journal of Hydrology, 566, 346–362. https://doi.org/10.1016/j.jhydrol.2018.09.027

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

Blazkova, S., & Beven, K. (2004). 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. Journal of Hydrology, 292, 153–172. https://doi.org/10.1016/j.jhydrol.2003.12.025

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

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

Cunha, L., Krajewski, W., & Mantilla, R. (2011). A framework for flood risk assessment under nonstationary conditions or in the absence of historical data. Journal of Flood Risk Management, 4, 3–22. https://doi.org/10.1111/j.1753-318X.2010.01085.x

Ehret, U., van Pruissen, R., Bortoli, M., Loritz, R., Azmi, E., & Zehe, E. (2020). Adaptive clustering: Reducing the computational costs of distributed (hydrological) modelling by exploiting time-variable similarity among model elements. Hydrology and Earth System Sciences, 24, 4389–4411. https://doi.org/10.5194/hess-24-4389-2020

Fischer, S. (2018). A seasonal mixed-POT model to estimate high flood quantiles from different event types and seasons. Journal of Applied Statistics, 45, 2831–2847. https://doi.org/10.1080/02664763.2018.1441385

Gringorten, I. I. (1963). A plotting rule for extreme probability paper. Journal of Geophysical Research, 68(3), 813–814. https://doi.org/10.1029/JZ068i003p00813

Grimaldi, S., Petroselli, A., Arcangeletti, E., & Nardi, F. (2013). Flood mapping in ungauged basins using fully continuous hydrologic-hydraulic modeling. Journal of Hydrology, 487, 39–47.

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

Haberlandt, U., & Radtke, I. (2014). Hydrological model calibration for derived flood frequency analysis using stochastic rainfall and probability distributions of peak flows. Hydrology and Earth System Sciences, 18, 353–365. https://doi.org/10.5194/hess-18-353-2014
Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: a k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), 100–108.

Jehn, F. U., Bestian, K., Breuer, L., Kraft, P., & Houska, T. (2020). Using hydrological and climatic catchment clusters to explore drivers of catchment behavior. *Hydrology and Earth System Sciences*, 24, 1081–1100. https://doi.org/10.5194/hess-24-1081-2020

Kiang, J., Gazsoorian, C., McMillan, H., Coxon, G., Le Coz, J., Westerberg, I., Belleville, A., Sevrez, D., Sikorska, A. E., Petersen-Overleir, A., Reitan, T., Freer, J., Renard, B., Mansanarez, V., & Mason, R. (2018). A comparison of methods for streamflow uncertainty estimation. *Water Resources Research*, 54, 7149–7176. https://doi.org/10.1029/2018WR022708

Kuentz, A., Arheimer, B., Hundecha, Y., & Wagener, T. (2017). Application of different clustering approaches to hydroclimatological catchment regionalization in mountainous regions, a case study in Utah State. *Journal of Mountain Science*, 15, 461–484. https://doi.org/10.1007/s11629-017-4454-4

Seibert, J., & Vis, M. J. P. (2012). Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. *Hydrology and Earth System Sciences*, 16, 3315–3325. https://doi.org/10.5194/hess-16-3315-2012

Seibert, J., Vis, M. J. P., Lewis, E., & van Meerveld, H. J. (2018). Upper and lower benchmarks in hydrological modelling. *Hydrological Processes*, 32, 1120–1125. https://doi.org/10.1002/hyp.11476

Shah, M., Rahman, A., & Chowdhury, S. (2018). Challenges for achieving sustainable flood risk management. *Journal of Flood Risk Management*, 11, S352–S358. https://doi.org/10.1111/jfr3.12211

Sharghi, E., Nourani, V., Soleimani, S., & Sadikoglu, F. (2018). Statistics for sample splitting for the calibration and validation of hydrological models. *Stochastic Environmental Research and Risk Assessment*, 32, 3099–3116. https://doi.org/10.1007/s00477-018-1539-8

Singh, S. K., McMillan, H., Bardsossy, A., & Chebana, F. (2016). Nonparametric catchment clustering using the data depth function. *Hydrological Sciences Journal*, 61(15), 2649–2667. https://doi.org/10.1080/02626667.2016.1168927

Singh, S. K., Bardsossy, A. (2012). Calibration of hydrological models on hydrologically unusual events. *Advances in Water Resources*, 35, 321–325. https://doi.org/10.1016/j.advwatres.2011.12.006

Singh, S. K., McMillan, H., Bardsossy, A., & Chebana, F. (2016). Nonparametric catchment clustering using the data depth function. *Hydrological Sciences Journal*, 61(15), 2649–2667. https://doi.org/10.1080/02626667.2016.1168927

Ward, P., van Pelt, S., de Keizer, O., Aerts, J., Beersma, J., van den Hurk, B., & te Linde, A. (2014). Probabilistic flood risk assessment. *Journal of Flood Risk Management*, 7, 141–151. https://doi.org/10.1111/jfr3.12029

Westerberg, I. K., Sikorska-SENONER, A. E., Viviroli, D., Vis, M., & Seibert, J. (2020). Hydrologic model calibration with uncertain
Winter, B., Schneeberger, K., Dung, N., Huttenlau, M., Achleitner, S., Stötter, J., Merz, B., & Vorogushyn, S. (2019). A continuous modelling approach for design flood estimation on sub-daily time scale. *Hydrological Sciences Journal*, 64, 539–554. https://doi.org/10.1080/02626667.2019.1593419

Zeimetz, F., Schaefli, B., Artigue, G., García Hernández, J., & Schleiss, A. (2018). A new approach to identify critical initial conditions for extreme flood simulations based on deterministic and stochastic simulation. *Journal of Hydrologic Engineering*, 23(8), 04018031. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001652

**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of the article at the publisher’s website.

**How to cite this article:** Sikorska-Senoner, A. E. (2022). Clustering model responses in the frequency space for improved simulation-based flood risk studies: The role of a cluster number. *Journal of Flood Risk Management*, 15(1), e12772. https://doi.org/10.1111/jfr3.12772