**Referee Report of**
**« Downsizing parameter ensembles for simulations of extreme floods»**
**(NHESS-2020-79)**

**General comments:**

The paper is well written and properly structured.

The scope of the study is clear, and the purpose of the selection of models among a given ensemble is relevant in the perspective of hydrological and hydraulic simulation for extreme flood estimation.

However, the presented methods, their scoring, and the interpretation of these scores deserve a better statistical assessment. For example, the presented scores, used in the comparison of the selection methods, rely only on the counts of simulated annual maxima being in or out of the predictive interval, without consideration of scale or frequency within their distribution. No assessment of the width of the predictive intervals is given relative to a global metric (standard deviation of annual maxima of the calibration data for example). Even if a relative ranking of the three methods can be provided here, it is difficult to have a proper statistical characterization of them in terms of robustness and reliability, which can be an issue for using them for extreme flood evaluation.

Furthermore, no interpretation of these selections in term of flood process or modelling is given. It is not supposed to be the core of this article, but it would help to connect the results to some hydro-climatological features and their impact in terms of variability. The problem of parameter equifinality is not evoked here, although it is the main factor of the parameters set variability, given that here, the 100 models have been calibrated by the same algorithm using the same data.

I would recommend then a major revision of this paper, in order to tackle with those main issues.

**Detailed comments/questions:**

Quoted sentences are written in italic.

Line 10: *10'000 years of synthetic streamflow data simulated with a weather generator.* Simulated "thanks to" a weather would be more appropriate, the weather generator doesn't generates streamflow directly, it feds the hydrological models.

Line 12: *The methods are readily transferable to other situations where ensemble simulations are needed*. This is only evoked as a perspective at the end of the paper, without providing an example of such "other" application. I am not sure it deserves to be in the abstract.

Line 39: *initial conditions for use in combination with design storms*. Regarding the SCHADEX method detailed in (Paquet el al., 2013), I would add "initial conditions for use in combination with design or randomly drawn storms".

Line 43: *especially if long time series are to be simulated using ensembles of hydrological parameter sets*. And also if very high return times (above 1000 years) have to be robustly estimated, thus implying several thousand years of simulation.

Line 45: *extrapolating a synthetic design hydrograph*. How? By scaling up a synthetic hydrograph thanks to estimated extreme quantiles of peak and volume values?

Line 51: *These continuous hydrologic simulation frameworks are still rare for time series >100 years*. 100 years of simulation merely allow to compute a robust 20 year return period estimation, which is pretty useless for dam safety for example. I would rather say that "high computational power are needed in order to provide estimations for high to extreme return times (up to 1000 years) required for safety-related studies" (although this is almost written in the same terms in line 57).

Lines 77-86: The problem is particularly well stated here.

Line 89: *to select a reduced-size parameter ensemble for the use with a hydrological model within a continuous simulation*. Here and later on I would always keep "parameter" linked to "hydrological model". Most simulation frameworks are heavily parametrized, and the uncertainty linked to the hydrological model is only one (important) of the numerous sources of uncertainty. I would then write "to select a reduced-size ensemble of hydrological model parameters for the use within a continuous simulation".

Line 92: *for simulation of extreme floods*. A recurrent formal remark about the word "extreme". Usually "extreme floods" refers to return times largely exceeding the observational range, currently more than 1000 years, thus being extrapolated (by FFA or simulation, or both). This is especially true in dam-safety related literature. In the presented case, the meteorological scenarios are 100 years long, meaning that only very few "extreme" floods are simulated in the whole experiment. At best a robust 1000 years estimation can be empirically inferred here given the fact than 100x100=10 000

meteorological years have been simulated. The whole set of AM being extracted can surely not be considered as a set of "extreme floods". The authors could consider using "intense floods", "rare floods" or more simply "floods" when they refer to the simulated floods.

Line 97: *simulated rare flood events*. Following the remarks above, the term "rare flood" is also appropriate alternative.

Line 99: *the aim is thus i) to provide long enough simulation periods for extreme flood analysis, ii) to avoid the propagation of errors due to data/model calibration etc. and iii) to be able to focus entirely on the uncertainty of the hydrological response*. I my opinion this goes farer than the actual results of the study. The uncertainty linked to model parameters is assessed, and properly summarized thanks to a reduced number of meteorological simulations. But I don't understand why this "provides long enough simulation", and why it "avoids the propagation of errors". If you have 100 "bad" models due to date, calibration, etc., three of them are selected in order to keep a good representation of their variability, but you still work with "bad" models (sorry for the term "bad", it only means "affected by uncertainty" !).

Line 111: *and not the model uncertainty of a weather generator*. This is perhaps one of the main limit of the study. At line 69, it is written that Arnaud et al. (2017) *found that the uncertainty of the rainfall generator dominates the uncertainty in the simulated extreme flood quantiles*. This uncertainty will not be considered here, and I wonder how far the results exposed here would still be useful to deal with the weather generator uncertainty (which of course is not to be confused with the variability of the scenarios generated thanks a given set of parameters). A comparison of both uncertainties (model and weather generator), even basic, would have been welcomed here.

Line 120: *(ii) the distribution is known*. I am not sure that knowing the probability of parameters is a reasonable perspective, in my opinion the problem of equifinality of parameters in models like HBV prevents an *a priori* expression of parameter probability, as different sets of parameters can "produce" the same model, i.e. models having the same behaviour for a given meteorological scenario. And this is one of the interesting outcome of this study, which focuses on the hydrological response of the models, and not on the actual values of the parameters. I think that this equifinality problem deserves more writing in this paper.

Line 129: *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 and the smallest interval bounding it from above*. This definition deserves to be connected to frequencies of the target variable, even if it's not straightforward. Does it (roughly) provides a 90, 95 or 99% confidence interval of the simulated variable? Given that follows in lines 184 to 234, with quantiles 5 and 95%, it "looks like" a 90% CI.

Line 138: *we thus propose to use the representation of AMs in the Gumbel space as the reference model response space for parameter selection*. The plotting in Gumbel is useful here to illustrate the rare to extreme quantiles, but doesn't explicitly play a role in the "parameter selection" (which doesn't imply any explicit reference to an implicit Gumbel distribution of AM in the statistical criterions/indicators used).

Line 142: *inverse modelling approach*. The term “inverse modelling” appears to me somehow excessive. An inverse hydrologic modelling would be for example to infer rainfall from discharges. Here it’s more a “post-modelling” approach.

Line 145: *the parameter set selection is made based on the full hydrological simulation ensemble but using only a limited simulation period*. To be more specific I suggest to write “based on the simulation with all the hydrological models but using […]”.

Line 189: *The parameter sets selected in step (d)*. Should be step (c).

Line 189: *the sets which are chosen most often as the 5th, 50th and 95th ranks are retained as the parameter sets […] representative for the entire simulation period*. The ranking methods yet shows its weakness: the $5^{th}$ and $95^{th}$ of a given year have very low chance to match to the overall corresponding quantiles, given the “climate variability” illustrated in Figure 1, thus preventing the parameters selected on a given year to have a global representativeness. I am not sure it’s worth keeping this method “in the game” for the rest of the paper…

Lines 194-206: I don’t understand why the “Gumbel space” is evoked here (three times!), and constantly throughout the paper. Apart from the plots of Figure 2 and others, what is “Gumbel specific“ in the metrics and statistics presented? For example, the $R_{MSE}$ scores are computed using each simulated annual maximum, regardless its empirical frequency.

Line 201-205, equations 1-3: This is the same equation for the three considered quantiles. Only one is necessary.

Line 211-234: same remark as above about the “Gumbel space”.

Lines 214: *These members are next clustered in the Gumbel space into three representative groups (clusters) based on all J simulation years using the k-means clustering*. If I understood properly it means that the clustering has been performed in the J-dimensional space of the full set of members values?

Line 217: *Next, these clusters are sorted by their magnitude*. What variable/quantile is used for this sorting?

Line 218: *Note that we use here percentiles instead of cluster means to make this method comparable with the other two methods*. I am not sure of that : say that each cluster regroups one third of the ensembles, and for a given quantile in the AM distribution (say the 50%), it is evenly distributed through all the members, the percentile 5% of the lower cluster would more or less correspond to a 0.05x0.33 = 0.17 global percentile. The 5% and 95% are more “rare” than their corresponding quantile in the quantiling method…

Line 223-227, equations 4-6: Same remark as for equations 1 to 3.

Line 233, equation 7: This mention of the plotting position mention could be moved at line 195.

Table 1: *Sorting space = Gumbel space*. Once again, I don't undertand how "Gumbel-specific" the sorting process is for Quantiling and Clustering.

Table 1: *Interpretation of pred. intervals / Parameter grouping*. I don't see to what these lines refer in the text before.

Line 254: *assessing how well the reduced ensembles cover the reference simulation ensemble*. More specifically I would rather say "how well the reduced ensembles substitute the whole simulation ensemble for the selection of representative parameter sets".

Line 273: *assess how well the defined identified intervals represent the ensemble members of this Sr meteorological scenario*. What metric is used to do this assessment?

Line 288: *Compute the 5$^{th}$ percentile […] and the 95$^{th}$ for {H($\theta_{sup,p}$/$S_m$)}*. The mention "for m=1,2…,M and m≠p" could be added for more clarity.

Line 290: same question as for line 273.

Line 295 and below: as evoked above, the recurrent mention to the Gumbel space is, in my opinion, useless, and over time tedious to read. In the paper, it is quickly implicit that the plots and the metrics used are defined in the frequency space (or frequency domain), being Gumbel or not, without need to repeat it.

Line 330: *Here we propose to use different percentiles, i.e., the 5th, 50th, and 95th percentiles, to characterize the ratio of the simulation points lying outside the computed predictive intervals for each of the methods*. I don't understand this? Why not using only the 50$^{th}$ percentiles of this ratio? Refer to comments on Table 3 for a more detailed version of this question.

Line 333: *how many out of J hydrological simulation points […] must lie outside the defined predictive intervals*. I think that the problem of such a simple "count" of points (simulated annual maximum) outside the predictive interval doesn't take into account their position in the simulated distribution. As written in the title, the methods exposed here are supposed to be used in the estimation of *extreme floods*, which in any post-treatment of the hydrological simulation will strongly rely on the high simulated quantiles. The scores should somehow reflect this focus on high quantiles, which is not the case here. Instead of this count of "outside points", the area outside the predictive interval could be computed, using the Gumbel variable (as x) and the discharge value (as y), thus giving a contrasted score in which lying outside the predictive interval for high quantiles is more important than for low values.

Line 337: *In this work, we consider the following values for $r_{thr}$ = {0.50,0.25,0.10,0.05}*. Following the preceding comment, a metric accounting for the scale or the frequency of the points being outside the predictive interval would avoid to distinguish such thresholds, which apart from the $r_{tht}$=0.50 or 0.10 have little statistical meaning in this context.

Line 343: *For testing the methods developed here, a small natural catchment is preferable*. Why small?

Line 358: *In this study, the version HBV light […] with 15 calibrated parameters is used*. The considered model can be then considered as heavily parametrized, and thus fully affected by the equifinilaty problem of its parameter evoked in the remark made for line 120.

Line 374: *for details on $R_{PEAK}$ and $R_{MARE}$, see the work of Vis et al. (2015)*. A brief description of $R_{PEAK}$ and $R_{MARE}$ would be welcomed here, especially as within the calibration process, the parameters conditioning the modelling of floods are surely strongly conditioned by the $R_{PEAK}$ score.

Line 377: *The available observational datasets are split into a calibration (1990-2005 years) and a validation (2006-2014 years) period*. What is the point of having a validation period here? This validation period is never used in that follows.

Line 381: *The calibration is repeated 100 times resulting in 100 independent optimal parameter sets*. I am surprised by the variability of the parameters obtained by these 100 calibration runs, performed on the same calibration data with the same objective function. I would like to read a comment from the authors on that. Mine is that the optimization is not complete, seeming to depend on the aleatory exploration performed by the genetic algorithm, somehow "trapped" in local optimums, and/or affected by a strong equifinality problem (yes, once again, sorry). An alternative strategy for the generation of model parameter sets, in my opinion providing more "independent" models, could be to bootstrap 12 years among the 24 years available in order to generate 100 truly different calibration & validation samples.

Line 381: *The median model efficiency measured with Fobj over all 100 runs is 0.7*. To better assess the quality and the variability of the models generated at this step, it would be useful to show the distribution of NSE (Nash & Sutcliffe Efficiency) for both calibration & validation, and the ensemble plots of daily regime and classified discharge distribution for all the generated models. The ensemble simulation of the biggest observed would also be very pedagogic.

Line 383: *which can be assumed to be a good model performance on an hourly scale*. As mentioned above, this really need to be illustrated more richly.

Line 397: *The daily values generated with GWEX_Disag were then disaggregated to hourly values using the meteorological analogues method*. More details would be welcomed on that disaggregation: what fields/variable are used for analogy, what analogy criterion, what about seasonality (i.e. are the analogues identified within period of the year similar to the one of the simulation to be disaggregated, etc.).

Line 399: *Next, catchment means were computed using the Thiessen polygon method*. On how many simulated precipitation stations do the Thiessen average rely on for the considered catchment? How many simulated stations lie within the catchment?

Line 403: *Thus, differences between scenarios are exclusively due to the natural variability of the meteorological time series*. "[…] and modelled by the GWEX weather generator" could be added. Similarly to the models, the variability of these scenarios deserve to be illustrated, and compared to the observations, e.g. thanks to their average and standard deviation of the annual maximum daily precipitations.

Line 406: *These 100 meteorological scenarios are used as input into the HBV model to generate streamflow time series with 100 different HBV parameter sets*. I am not sure that this sentence is useful. The simulation scheme is clear from the beginning.

Figure 4: The title of the second plot should be "Quantiling" instead of *Quantailing*.

Lines 419-434: I find the results of this paragraph difficult to interpret. Some violin plots show odd parameter selection patterns (like in Clustering/Infimum), other show weak parameter discrimination (Quantiling/Median). The Table 2 is quite difficult to read/interpret with so many counts exposed. In this paragraph and in the following ones, some "illustrations" of the most selected parameter sets should be provided, e.g. by presenting the range of hydrological responses to observed meteorological data of the selected models compared to the full ensemble. In other words, some interpretation in term of modelling and hydrological processes would be welcomed.

Line 434: *Interestingly, for the supremum set in the clustering method, only four parameter sets among all 100 available are chosen over all 100 scenarios*. Given that, I don't understand why in Table 2, column Clustering/$\theta$sup, 5 parameters sets (# 34, 22, 98, 86, 50) are identified.

Table 2: a graphical alternative or a complement to that table deserves to be presented, to better assess the "density" of parameter sets selected by the different methods.

Line 437: *intervals for extreme flood predictions*. The term "extreme flood estimations" could be more appropriate.

Line 441: *According to a first visual assessment, these three methods lead to slightly different constructed frequency intervals particularly in the upper tail of the distribution*. To ease this visual assessment, horizontal lines marking these intervals for the upper values could be added to the plots of Figure 5.

Line 446: *the three intervals are always correctly attributed*. I would temper this in writing that "the three intervals are always correctly ordered" as this exactly what it is measured in $R_{bias}$.

Line 456: From the visual assessment, it is difficult to judge the methods. See remark on Figures 7-8.
Figure 5 to 8: Instead of having an x-axis graduated with the Gumbel variable U, some ticks at remarkable return times (2, 5, 10, 20, 50 & 100 years) could be added in order to ease the reading of these plots, and avoid the long caption *The Gumbel variates etc.* in Figure 4.

Line 463: *the highest values for both evaluation criteria, i.e., the median ratio of simulation points lying outside the predictive intervals (Rspo) and the median ratio of hydrological simulation ensemble […].* Given the definitions of §2.6.2, this is more a mean ratio than a median ratio.

Line 475: *Hence, again here all three method can be qualified as behaving well based on the multi-scenario evaluation, and only the order of their behavior can be established*. Honestly, at the end of this paragraph, I have no clear idea of the absolute performance of each method. One important point is that the methods provide rather different intervals (like illustrated in the Figure 6), thus a method providing wide intervals will have good "in/out" scores (like the ones in Table 3), better than for narrow intervals, but the question of the statistical relevance of such intervals is not solved.

Figure 7-8: I found the plots of the top panels of both figures rather counterintuitive: *the prediction interval resulting from selecting representative parameter sets for 99 meteorological scenarios and compared to the full simulated range with all 100 parameter sets* seems narrower than "statistically expected" (more and less a 5-95% confidence interval given the quantiles or percentiles involved in the process). For the highest simulated quantiles, the prediction interval seems only to cover about 50 to 66% of their variability. In the bottom plot of the Figure 7 for clustering, a second blue interval is plotted without being identified in the legend nor in the caption.

Line 480: […] *selecting representative parameter sets that yield reliable predictive intervals in the frequency domain*. Following the comment on line 475, I see no statistical demonstration of the reliability of the predictive interval (like the one that could be done by controlled random generation of a given variable to which a statistical test is applied, then a proper statistical scoring). I agree with the authors on that a ranking between the three methods (according to the presented scores) is however established.

Table 3: Three quantiles of the $R_{spo}$ score are given, although the caption mentions that *the values represent the median values over all 100 scenario runs*. What for providing the $5^{th}$ and $95^{th}$ quantile of a score measuring *the ratio of simulation points […] lying outside the predictive intervals* (line 312), which should be, on average, close to 10% (once again given the quantiles involved in the selection process)? In the low part of the Table 3, the *Metric method* is written as $R_{hso}$ *(*$R_{mso}$*).* Which scores are the ones provided?

Line 481: *all three methods are fit-for-purpose for extreme flood simulation*. Following the preceding comment, if the presented method cannot be statistically demonstrated, it can be considered as an *ad-hoc* heuristic, build for a given purpose, here extreme flood simulation/estimation. This last step is not evoked in the paper, then depriving the reader from assessing the relevancy/robustness of this heuristic.

Line 487: *for additional ease of use criteria*. I don't understand this sentence.

Line 488: *From the visual assessment*. Based on which figure?

Table 4: The different ranking features should be linked to the scores presented in Table 3. Some of them deserve to be better explained in the text: *Independence from meteorological scenario,*

*Independence from simulation years, Ease in application, Interpretation of prediction intervals*. They don't refer explicitly to scores, statistics or plots presented before.

Line 497-502: This lines could be put in section 2.5 in order to better describe the assessment of the approach.

Line 514: *The other two methods need to be performed in the Gumbel space over the entire simulation period and, in the case of the clustering method, require some additional computational effort*. Once again this reference to "Gumbel space" is unappropriated given the scores computed, and the additional computational effort doesn't seem significant, completely justified by the added robustness.

Line 516: *The use of the Gumbel space in selecting the representative parameter sets helps, however, to interpret the constructed prediction intervals and to directly assign return periods to them*. Same remark as for line 295.

Lines 522-534: These lines are, in my opinion, a short summary of the study, and do not fit in this section (Limitations and perspectives).

Line 541: *This use of synthetic data makes the approach results independent from the catchment properties and limits the effect of the hydrological model error and errors in calibration data on the methods comparison results*. I may be more cautious on that, given that the scores and the ranking of the methods are somehow linked to 1) the variability of the ensemble models, which depends on the equifinality of the model's parameters, the calibration data and FOs, etc. and to 2) the meteorological variability of intense events generated by the weather generator which depends on the climatology, the scale etc. Only some tests on different catchments (in scale and climatology) could ground this assertion.

Line 544: *We can, however, not directly assess here how much variability in the full hydrological ensemble is due to the climate variability and how much is due to the uncertainty resulting from the hydrological model parameters, because these two components are not linearly additive*. A simple exercise could help by 1) choosing a "median" model (in term of median response on the meteorological ensemble) and plotting/scoring the variability of simulations for the whole meteorological ensemble, and 2) choosing similarly a median meteorological scenario and simulating with the whole set of models and then 3) comparing the spread/variance of definite quantiles in the simulations. In my opinion, this is an indispensable complement to the presented results.

Line 560: *Thus, the proposed selection methods could potentially be extended to account for different flood types*. Another option could be to consider Peak-Over-Threshold selection instead of a block selection (annual maximum) in building the simulated distributions. If different flood processes are present above a certain intensity threshold, flood type and seasonality sampling will be relevant.

Line 566: *the three sets emulate the common practice of communicating median values along with prediction limits*. But in that case, these predictive intervals have to be statistically calibrated (or checked) in order to be used in safety studies, especially if these studies lead to engineering or compliance check.

Line 572-584: Here again, the term *Gumbel space* could be replaced by "frequency domain" or equivalent.

Appendix A & Figure A1: Interesting but rather off-topic here, as only continuous hydrological simulation has been used in this study.

Figure A3: For a better assessment of the distribution of calibrated parameters, I suggest that the scale of the horizontal axis of the violin plots (parameter values) exactly matches the corresponding calibration range written in the Table A1.