Authors' response to interactive comment by Dr Eric Gaume

The manuscript presents and extension of a work published in 2015 by Smith et al. and aiming attesting methods suited for design flood estimation at global scale. The article is based on the analysis of the very rich international streamflow database GSIM. Numerous other international datasets are processed to derive climatological, physiographical and hydrological descriptors for each of the considered 12000 watersheds worldwide. Three different regression methods are tested to relate the locally estimated annual maximum discharge quantiles and the watershed descriptors, namely a power function, a support vector and a random forest model. A split sample test is used to assess the performances (bias, root mean square error) of the various implemented and tested approach. All methods are described in the manuscript. The manuscript is overall of high quality, comprehensive and based on the best available datasets and methods. It deserves without doubt a publication in HESS. Its content could be slightly improved in several ways.

We thank Dr Eric Gaume for his valuable comments and suggestions that will undoubtedly help us improve our manuscript. Below we reply to each of the comments and explain how we have incorporated them into the manuscript.

(1): Some figures appear complex and difficult to understand without the explanations provided in the text. The legends and captions could be improved and enriched (see comments in the attached pdf)

Reply: This is very helpful. We revised the captions according to your kind comments in pdf as follows.

1) The caption of Figure 4 has been enriched.

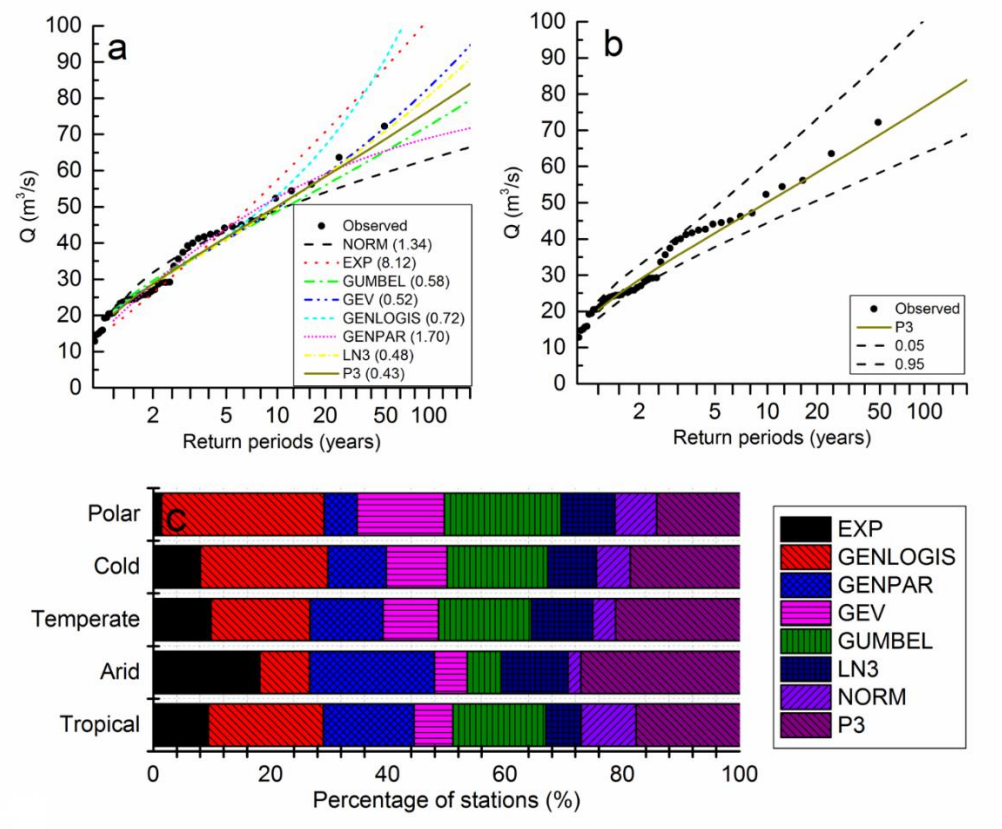

Figure 4 Flood frequency curve selection and estimation
(a: Flood frequency curve selection of the station No. AT0000032; b: Design flood estimation of the station No. AT0000032 using the Bayesian MCMC interface. c: Selection results of flood frequency curve for all stations)

2) We clearly describe the labels rather than acronyms in Figure 7.

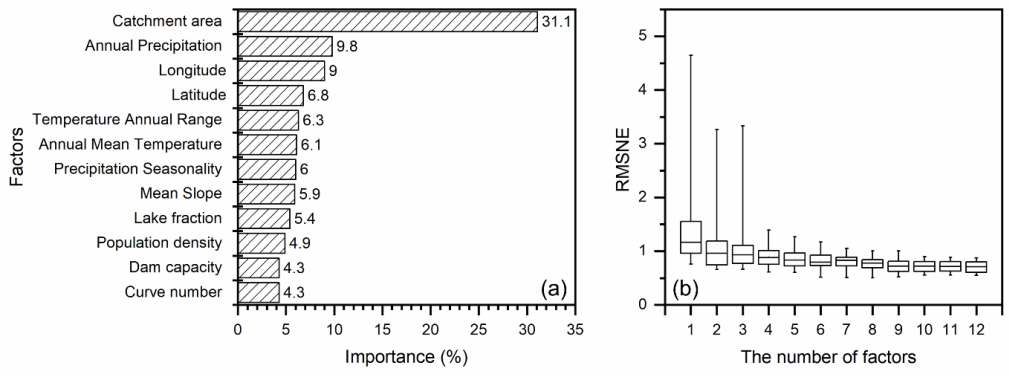

Figure 7 (a) Factor importance evaluated by RF model and (b) the impact of catchment descriptors for SVM regression

3) Subtitles and caption in Figure 6 have been corrected.

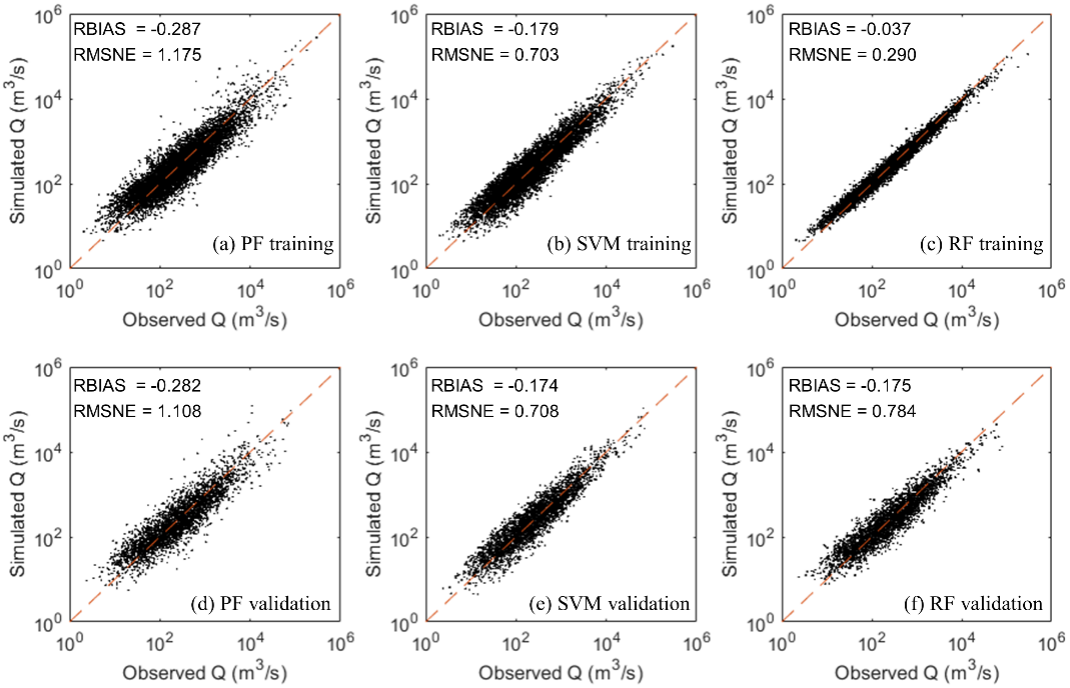

Figure 6 Comparison of three regression models (PF, SVM, and RF) during training and validation periods

(2): The authors made an important effort to provide to the readers all the necessary mathematical background and equations. But notations in the equations and the indices are not consistent throughout the manuscript introducing sometimes some confusion. This could be corrected (see attached document).

Reply: Thanks. We checked the notations and indices in the manuscript to avoid confusion.

(3): The authors use a Bayesian MCMC inference framework to derive local discharge quantile estimates but they only make use, in fact, of the most-probable estimated value (i.e. the maximum likelihood estimate if a non-informative prior is used which is what I suspect). They do not valuate the credibility intervals for the estimated quantiles in the analysis or discussion of the manuscript. In fact, a maximum likelihood local estimation is used, even if it is through a Bayesian MCMC framework. This should be

clearly stated in the manuscript and the first paragraph of section 3.2.2 has to be reformulated accordingly (see comments in the pdf).

Reply: I agree with you. The Bayesian MCMC inference is only applied to derive the most-probable design floods. This method is also based on likelihood but can provide much richer results than traditional maximum likelihood method. We reformulated section 3.2.2 to more clearly make this point.

(4): The proposed method is based on a clustering approach and regressions conducted on each cluster separately. But the clusters are optimized based on the SVM regression method. It is then not really surprising that the SVM method provides the best performances if compared to the two other tested method and especially to the RF. The comparison between the approaches is not totally fair. This should be mentioned in the discussion.

Reply: This is very helpful. The previous description of subgroup delineation was not clear. The number of subgroups is selected by considering both the heterogeneity (reflected by the Davies-Bouldin index) and the number of stations in the subgroup. Figure R1 describes the selection process of K when using all factors for clustering. As shown in Figure R1 (a), the DB index reached an optimal value at K=16 and then fluctuated with increasing K. From Figure R1 (b) we found that the number of stations in subgroups reduced with a larger K. To ensure a sufficient number of stations for model development for each subgroup, K greater than 40 was not considered in this research.

This procedure provides a fair comparison between different regression models. We revised section 4.2 to make it easier for readers to follow. Meanwhile, we understand that it is a high risk to mention that RF overperformed SVM as the model performance is highly dependent on the specific split datasets. Therefore, we clarified this point in the revised manuscript.

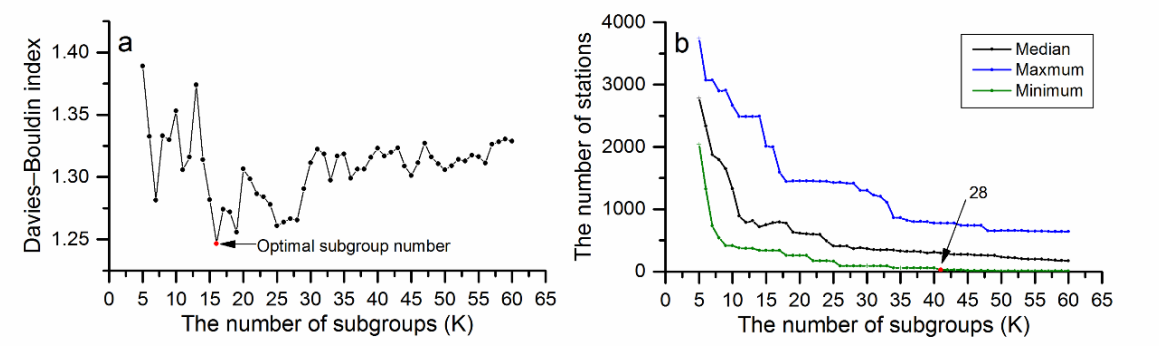

Figure R1 Optimal K selection by (a) DB index and (b) the number of stations in subgroups during delineation

(5): Some stations are discarded from the analysis after clustering based on a discordancy test (first paragraph of section 4.1). This may be problematic from a methodological and a statistical point of view. First, eliminating data based on a threshold from a statistical test is questionable from a statistical point of view and may be seen as an over-interpretation of the results of the statistical tests. At a significance level of 5%, one would expect the p-value to be exceeded for 5% of the available sample on average: i.e. if the sample is homogeneous, the p-value will be exceeded 5% of the time. Does it then make sense to eliminate these 5% of stations from the sample ? Second, the method is developed to be applied on ungauged watershed. Each ungauged site will be affected to one of the defined clusters, but it will not be possible to verify that the specific site is not discordant with the rest of the cluster. It is to be foreseen that the proportion of "discordant" sites will be equivalent or even higher for the ungauged sites than for the gauged sites (the clusters were adjusted on the gauged sites). The "discordant" sites may be discarded in the calibration process of the regression method but should be included in the validation dataset to provide a proper estimate of the performances of the proposed approach: i.e. an estimate of their performance if implemented in real-life applications.

Reply: Thanks for this constructive comment. Before model development, it is important to eliminate a small number of 'discordant' stations from the whole dataset. For example, a station which only records the discharge after dam construction shows

a significant ‘discordant’ behaviour in discharge and will be identified as a ‘discordant’ station of the subgroup. The reasons for ‘discordance’ may be induced by the uncertainty of the discharge data, catchment descriptors, the clustering process etc. Therefore, it is a high risk to use these ‘discordant’ stations for model development and validation. We therefore agree with you. The threshold (5%) is an empirical value. If the samples are homogeneous, some ‘useful’ sites may also be removed.

As your kind suggestion, we compared the model results of training, testing and eliminated sites (in Figure R2). The threshold is selected to ensure the training and testing performances are stable using different randomly split datasets. We found that some eliminated sites still can be reliably estimated which reveals that some useful stations are removed with this threshold. This is just a small number of stations and will not significantly change the results. However, the wide range of R-BIAS for eliminated stations reveals that it is a high risk to apply the proposed approach to a discordant or nonstationary site. To address this limitation, nonstationary RFFA approaches should be explored in further research. We now discuss this point in the revised manuscript.

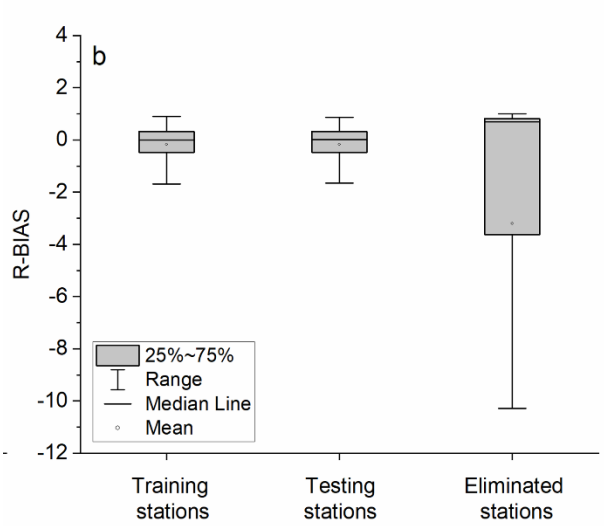

Figure R2 R-BIAS of 100-year return period flood estimation in training, testing and eliminated stations.

(7): Some sentences appeared a little simplistic to me (see attached commented pdf). Some nuance should probably be introduced at several places. With the hope that this

review will help the authors to improve their manuscript and looking forward to seeing this interesting paper being published in the near future in HESS.

Reply: Yes, we revised these according to your kind comments in the attached commented pdf. We are very grateful for your sentence-by-sentence suggestions.