This MS attempts to derive a DF for global river basins using a large sample (n=11793) using state-of-the-art ML techniques. I welcome this effort and consider that that the HESS readership will do so too. Three reviewers have provided numerous comments that need to be addressed by the Authors in the revised manuscript. There are many comments regarding the notation, Methodology and the statistics that are not clear. In my opinion there are major shortcomings that need to be addressed before this MS is accepted in HESS:

Dear Luis,

We thank you for the constructive suggestions. We have carefully replied to all comments as follows and revised them in the manuscript in blue-coloured text. We hope that these responses and revisions meet your expectations.

Best wishes,
Gang Zhao, Paul Bates, Jeff Neal and Bo Pang

1. The authors should put in the appendix all statistical test that are standard, e.g., Mann-Kendall test and any test that has been already published. Writing once again text-book equations only adds bulk to the MS but not insight. Same for K-means et. If need, write all tests in an appendix for reference. Table 3 to appendix. The same with Anderson-Darling (AD) test.

Reply: Thanks. We put all statistical tests and Table 3 in the Appendices.

2. I am missing the coefficients of the potential models for the regionalization of the design floods (DF) in eq. 25 and their confidence intervals. See how we reported in Samaniego and Bárdossy, 2005 JoH (SB2005).

Reply: Thanks for your kind comments and references. In this research, the Power-form function (PF) is regarded as a benchmark model. We have not listed coefficients of PF

in the manuscript mainly because it showed much worst performance (RMSNE >1) than the machine learning models. This reveals the coefficients of PF cannot accurately describe the contribution of factors to the result. We have added a more detailed description of this point in Section 4.1.

In this regard too, how did you select the optimal number of predictors? Which methods was used? Are the number of predictors the same for all regions? How the parameters (\beta_i) vary from region-to-region? Is the change significative?

Reply: Sorry for the unclear description here. Before model development, we analysed the correlation of all factors. According to the criteria proposed by Evans (1996), no factor pairs show a strong correlation (Pearson's correlation coefficient > 0.6). This reveals no factor is replicated with others.

Table R1 correlation analysis of all factors

	CA	SL	AP	PS	AT	TR	CN	DC	LF	PD	LA	LO
CA	1.00	-0.15	-0.12	0.24	0.07	0.08	0.03	0.02	0.15	-0.13	-0.09	-0.02
SL	-0.15	1.00	0.17	0.01	-0.29	-0.27	-0.30	0.01	-0.07	-0.13	0.06	-0.02
AP	-0.12	0.17	1.00	0.00	0.39	-0.48	-0.01	0.02	-0.05	0.08	-0.27	0.13
PS	0.24	0.01	0.00	1.00	0.39	-0.23	0.22	0.02	-0.03	-0.01	-0.36	0.08
AT	0.07	-0.29	0.39	0.39	1.00	-0.57	0.50	0.00	-0.26	0.15	-0.70	0.27
TR	0.08	-0.27	-0.48	-0.23	-0.57	1.00	-0.12	0.02	0.20	-0.07	0.56	-0.42
CN	0.03	-0.30	-0.01	0.22	0.50	-0.12	1.00	-0.01	-0.32	0.16	-0.29	-0.03
DC	0.02	0.01	0.02	0.02	0.00	0.02	-0.01	1.00	0.13	0.00	0.00	-0.02
LF	0.15	-0.07	-0.05	-0.03	-0.26	0.20	-0.32	0.13	1.00	-0.05	0.18	-0.09
PD	-0.13	-0.13	0.08	-0.01	0.15	-0.07	0.16	0.00	-0.05	1.00	0.01	0.07
LA	-0.09	0.06	-0.27	-0.36	-0.70	0.56	-0.29	0.00	0.18	0.01	1.00	-0.40
LO	-0.02	-0.02	0.13	0.08	0.27	-0.42	-0.03	-0.02	-0.09	0.07	-0.40	1.00

The factor importance in each subgroup was evaluated using the RF model and Figure 7 (a) describes the range and average value of factor importance of all subgroups. We found that catchment area, annual precipitation, and latitude and longitude are the top four factors that contribute most to the results. To reduce model complexity, the type of descriptors for training and validation is the same for all subgroups. The optimal number of catchment descriptors on regression results was further selected based on the SVM regression and the factor importance order identified by the RF model. As shown

in Figure 7 (b)), we compared the model performances using the different number of predictors. We found that the range was stable after the top ten factors were considered for model development. The SVM showed the highest performance considering all factors for regression (mean RMSNE of 0.70 in the validation period). We more clearly described this point in the revised manuscript.

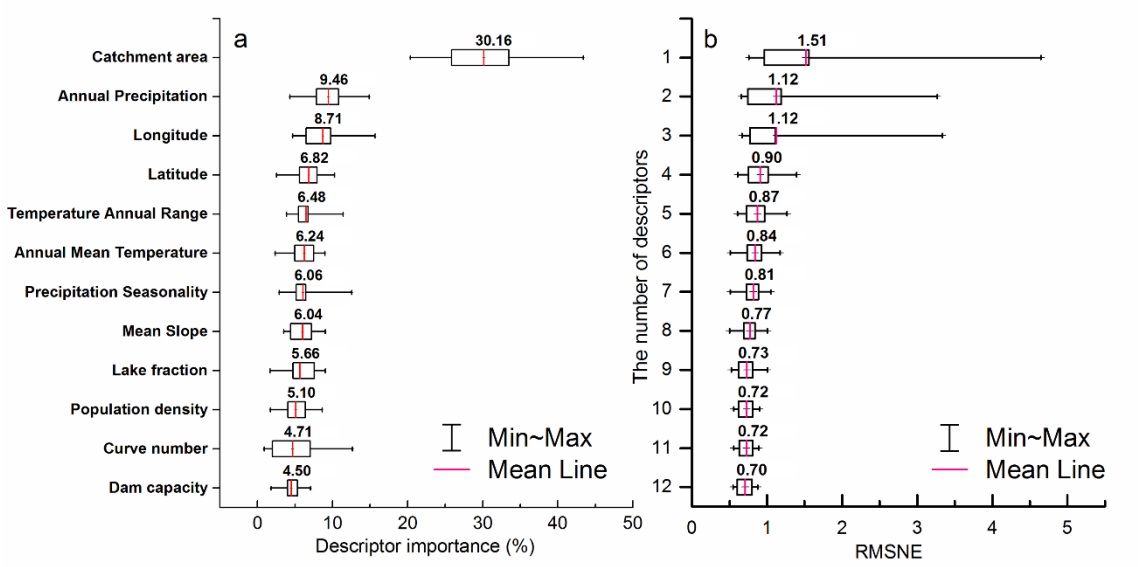

Figure 7 (a) Descriptor importance evaluated by RF model and (b) the optimal number of catchment descriptors for SVM regression

3. In the K-means phase: did you test another metric? For example Mahalanobis? There is not reason why the Euclidian metric describe the manifold of the predictors in the best way. Even if the Euclidean distance is the best, perhaps there is an embedding space (u) that betters describe the relationships between predictors (x). See other possibilities in Bárdossy, Pegram, & Samaniego WRR 2005 or in Samaniego, Bárdossy, and Schulz IEEE TGRS 2008.

Reply: This is very helpful. We compared four distance metrics in 100-year design flood estimation. As shown in Table R1, we found that changing the distance metric will not significantly affect the regression model results, and the squared Euclidean distance is the best metric amongst all those we compared. In this study, the subgroups are delineated based on a widely used K-means model using squared Euclidean distance. I agree with you: the K-means model is susceptible to outliers and noise and also cannot solve non-convex clusters. Other clustering models and distances metrics which can

better describe the highly nonlinear relationship between descriptors should be compared in the future study. We now consider this limitation in the discussion.

Table R1 100-year design flood estimation results using different distance metrics

Distance metrics	Training		Testing	
	RBIAS	RMSNE	RBIAS	RMSNE
sqeuclidean	-0.179	0.703	-0.174	0.708
mahalanobis	-0.192	0.728	-0.181	0.723
cosine	-0.181	0.767	-0.178	0.838
minkowski	-0.174	0.704	-0.188	0.742

4. Minor: use color as suggested in https://colorbrewer2.org for the map and other graphics. And 8. Use more pleasing colours in Fig. 2. (suggestion above)

Reply: This is very helpful. We used colours as suggested in the revised manuscript.

5. Math notations should be consistent. Regression models must have an error term. (e.g., in eq. 25). By the way, are the errors of the DF models homoscedastic? Please comments and show tests.

Reply: Thank you very much for your constructive comments. We revised all notations to avoid inconsistency. The error term is added in eq. 25. We added a boxplot to describe the range of errors of training, validation and eliminated stations (in Figure 9 (b)). We found the errors of the DF in training and testing stations are homoscedastic.

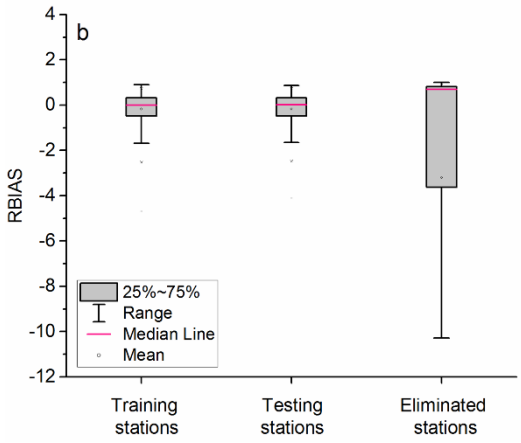

Figure 9 (b) RBIAS of 100-year return period flood estimation in training, testing and eliminated stations.

For the regression models, which estimators were used? Please provide Jackknifed (leave one out) bias and RMSE (see paper SB2005). this is important to understand if the coefficients are robust of just the result of one outlier, which is very likely in this global data set. L1 estimators are recommended in this case.

Reply: In this research, we adopted a simple hold-out strategy that 70% and 30% of stations were randomly selected for training and testing, respectively. This strategy is widely used in machine learning studies when the samples in training and testing datasets are sufficiently representative. To adopt Jack-knifed validation in this research, 11793 models will be developed. This will significantly increase the computational demand and increase the difficulty during model selection. Meanwhile, even though each model's test error is unbiased in Jack-knifed validation, it has a high variability as only one sample is validated for prediction. A sufficient number of hold-out samples are needed to demonstrate prediction ability of the model in terms of new data.

I agree with you, we should develop a robust model to avoid one particular dataset having too great an influence on the results. In the revised manuscript, a 10-fold cross-validation strategy was adopted to test the influence of hold-out samples on the model results in the discussion section. All stations in one subgroup are randomly divided into 10 folds. For every fold i (i=1, 2, …, 10), the Model_i is trained by the remaining 9 folds (except the ith fold) and is validated by the ith fold. Using this strategy, 10 models are developed for each subgroup, and each sample in the original dataset was used for validation once. Table 5 describes the best, worse and median results validated by samples from 10 folds. We found that the selection of hold-out samples had a moderate impact on the PF model and the RMSNE performance ranged from 0.97 to 1.49. Both SVM and RF showed stable performances and SVM provides the narrowest range. We suggest using ensemble results from SVM models using different split training samples

to reduce the errors generated by sample selection. We demonstrated this in the Section 4.5.

Table 5 Results of 100-year design flood estimation using 10-fold cross validation.

Regressions	RBIAS			RMSNE		
	Worse	Best	Median	Worse	Best	Median
PF	-0.21	-0.37	-0.29	1.49	0.97	1.12
SVM	-0.16	-0.20	-0.19	0.74	0.69	0.71
RF	-0.14	-0.21	-0.18	0.78	0.69	0.74

6. Descriptions of RF, SVM, MCMC, etc. are out of the scope of this manuscript. Here the authors should indicate how they make assumptions, set thresholds, etc. Then show results, and compare the various methods. Show which is the best in cross-validation experiments.

Reply: Thanks. We revised the description of RF, SVM, etc. based on editor and reviewers' comments. We added the 10-fold cross-validation results in the revised manuscript (4.5).

7. Provide uncertainty bounds of the statistics in Table 4. The same for figure 6.

Reply: Thanks for this kind suggestion. In this research, we proposed a hybrid model framework for regional flood frequency analysis at a global scale. The uncertainty sources of the proposed framework may be induced by the errors in model inputs (discharge data and catchment descriptors), at-site design flood estimation, and regional design flood estimation (clustering and regression models). To best of our knowledge, there still lack mature method to quantitatively assess the RFFA uncertainty bounds for hybrid model framework at the global scale. A robust uncertainty analysis approach considering these complicated model structure and uncertainty sources should be developed in the future works.

A table with the factors used for the regressions should be included. I can't find them. in Table 4, which are the Meteorological, Physiographical, Hydrological and

Anthropological factors? See in SB2005 how we described them. How the optimal nr. was found? The data set should be made available too. I see a description in L370, but a table with basic descriptors (max, min, mean) is useful in the appendix

Reply: Thanks. We described the max and min value and data source of the basic descriptors in Table 2. All factors are used for clustering and regression. During clustering, we compared the model results using all factors, the meteorological, physiographical, hydrological and anthropological factors, respectively. As shown in Table 3, the best combination is to use all factors for clustering. We added the abbreviation of factors in Table 3 to make it easier for readers to track.

Table 3 The impact of clustering factors on regional 100-year flood estimation

Clustering factors	**Optimal K**	**Training**		**Validation**	
		RBIAS	RMSNE	RBIAS	RMSNE
All factors	16	-0.179	0.703	-0.174	0.708
Meteorological factors (AP, PS, AT, TR)	11	-0.202	0.768	-0.185	0.746
Physiographical factors (SL, LF, LO, LA)	5	-0.214	0.829	-0.244	0.824
Hydrological factors (CA, CN)	30	-0.207	0.781	-0.243	0.876
Anthropological factors (DC, PD)	7	-0.379	1.235	-0.909	1.788

During regression, the RF model is adopted to identify the factor contribution by using the out-of-bag (OOB) samples approach. To reduce model complexity, the type of factors for training and validation is the same for all subgroups. The optimal number of catchment descriptors on regression results was further selected based on the SVM regression and the factor importance order identified by the RF model. We added the boxplot (Figure 7 (a)) to describe the max, min, and mean value of factor importance of all subgroups.

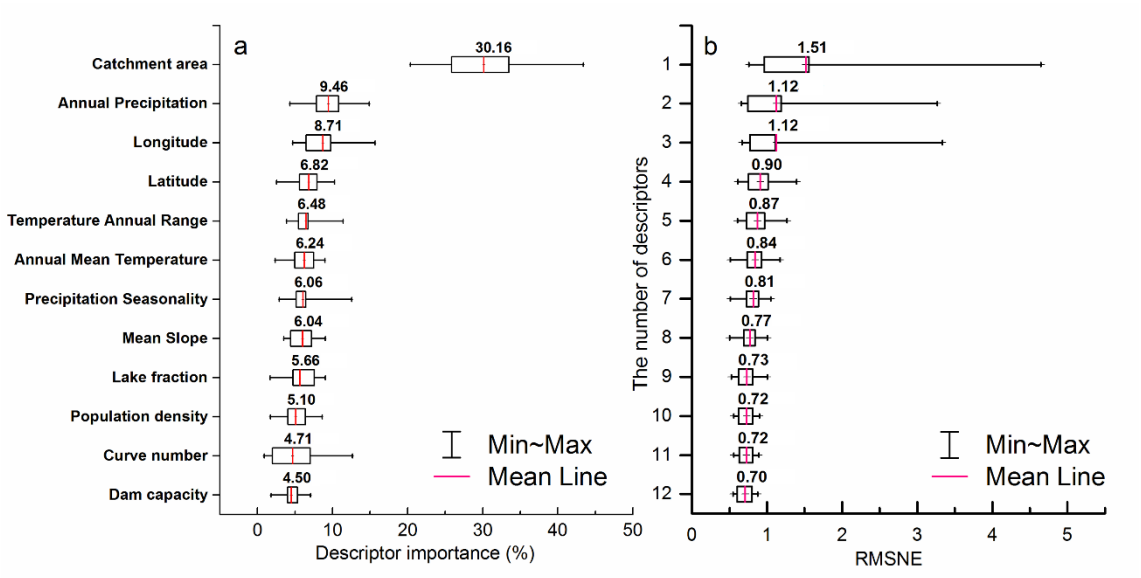

Figure 7 (a) Descriptor importance evaluated by RF model and (b) the optimal number of catchment descriptors for SVM regression