Authors' response to interactive comment by Anonymous Reviewer #3

This study aims to provide a reliable design flood estimation at global scale using improved methods and an expansive discharge dataset. The authors developed a three-phase model framework consists of standard parameter estimation methods and novel machine learning regressions. The framework mainly includes three parts: (1) estimating the gage wise flood frequency curve using data from a global discharge station network. (2) clustering these stations into subgroups based on basin characteristics. (3) developing a machine learning-based regression model in each subgroup for design flood estimation based on the subgroup's shared basin characteristics. The authors also compared the accuracy of the results in different regions globally and compared the performance of the three machine learning regressions in flood estimation. This study employs innovative methods to study a topic that is very relevant to HESS. The manuscript is generally well-written, the methods are sound, and the result presentation is clear. I suggest considering the following comments in the revision:

Thank you so much for your helpful comments. These comments significantly improved this manuscript. We revised the manuscript (traced in blue-coloured text) and reply to each comment as follows. We hope that these responses and revisions meet your expectations.

(1): The minimum drainage area of this study is 50km^2 (Line 131). How is this cutoff selected? I wonder if there are gages with smaller drainage basins, and how would the method work with relatively low flows? You can also compare the flood estimation accuracy as a function of drainage area size based on the gages used in the study.

Reply: This is very helpful. The cut off is an empirical value and is selected based on the definition of a small catchment in some studies (Djodjic et al., 2021;Niadas, 2005;Tsegaw et al., 2019). These catchments were not considered for model development as the floods in the very small catchments (which usually are regarded as

flash floods) have different characteristics from larger river floods. Flash floods usually last less than 6h and are difficult to describe using the traditional design flood estimation techniques. We cited these references and clarified this point in the revised manuscript.

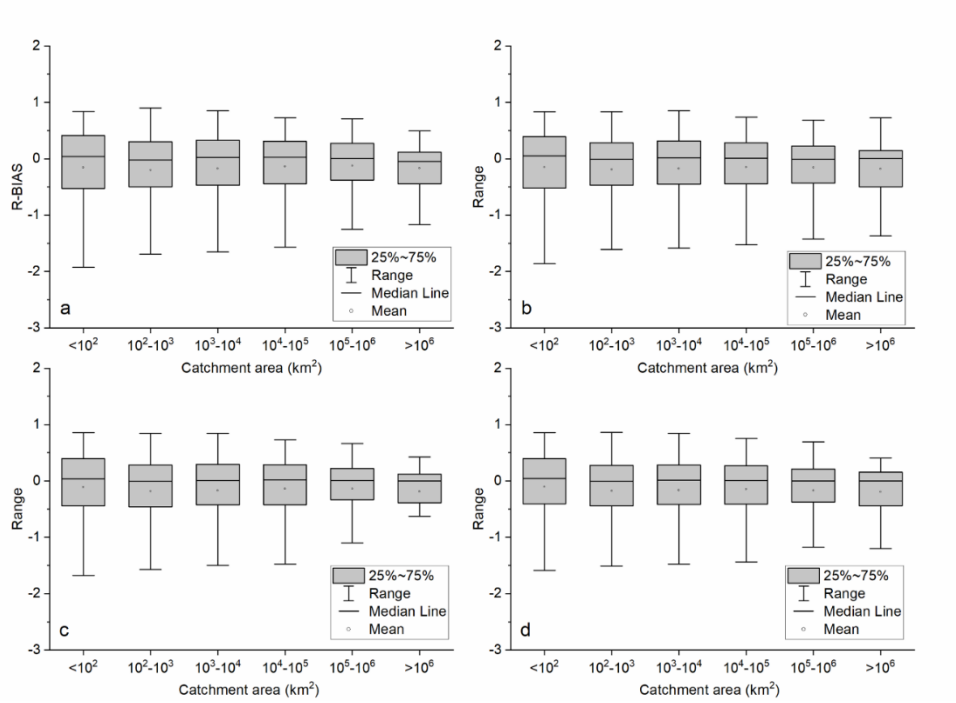

Figure R5 R-BIAS of (a) 100-year, (b) 50-year; (c) 20-year; and (d) 10-year flows in different catchment size.

As per your kind suggestion, we compare the flood estimation accuracy under different catchment sizes. As shown in Figure R5, both over and underestimations were found from small to large catchments. The range of RBIAS in the small catchment is typically wider than that in the large catchments. This reveals that design floods in small catchments are more difficult to estimate than in large catchments. We added these points to the revised manuscript.

(2): Can you provide insights into which of the four factors contributes the most to improving the system's estimation skills? The contribution is relatively easy to quantify with the traditional methods like multivariate regression, but it is not immediately clear with machine-learning methods that work like black boxes. I understand a quantitative

analysis of this is non-trivial and is out of the scope of this study; a brief discussion could serve as a future work/direction.

Reply: Yes. The RF model can identify the factor importance by using the out-of-bag (OOB) samples approach (samples not selected by the bootstrap method). Once an RF is developed, the error of OOB samples (E_{OOB}) can be computed as Eq. (R1).

$$E_{OOB} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \quad \text{(R1)}$$

Where n is the total number of OOB samples; $\hat{y}_i$ is the predicted value of RF.

Each factor in the OOB samples is permuted one at a time, and the permuted E_{OOB} can be computed with the permuted OOB samples and the trained RF model. The RF estimates the factor importance by comparing the difference between original and permuted E_{OOB} while all others are unchanged. The results of factor importance are shown in Figure R6 (a), and we found that catchment area, annual precipitation, and latitude and longitude are the top four factors that contribute most to the results. This rank was further validated using the SVM model (in Figure R6 (b)) where we found that the SVM model performances were stable after the four top factors were considered for model development. We discuss more about this point in the revised manuscript.

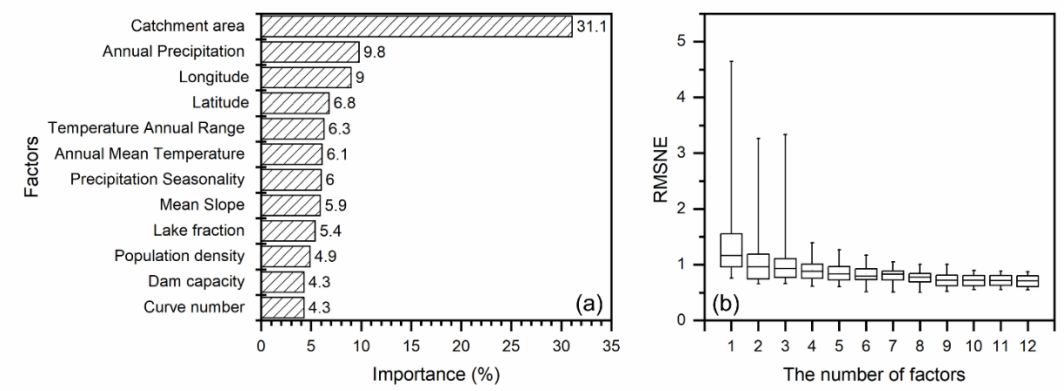

Figure R6 (a) Factor importance evaluated by RF model and (b) the impact of catchment descriptors for SVM regression

(3): The introductions of SVM and RF in section 3.4.2 and 3.4.3 are out of this study's context. I suggest including details on how these methods are implemented with the flood and other ancillary data in this study.

Reply: Thanks for your kind suggestions. We have added more description of the SVM and RF models especially on how these methods are implemented in section 3.4.2 and 3.4.3 as follows.

(1) SVM

SVM has shown advantages in solving complicated non-linear problems in the field of hydrology. The adopted SVM regression model was proposed by Drucker et al. (1997) and successfully used in forecasting of flood, drought, groundwater etc. For a given training dataset $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$, where N is the number of training samples, the overall goal of SVM regression is to find a function $f(x)$ that has at most ε deviation from the observed y_i. Thus, the SVM regression model can be described as a convex optimization problem as Eq. (26).

$$\min_{w,b} \frac{1}{2}\|w\|^2 \tag{26}$$

$$\text{s. t. } \begin{cases} y_i - w^T x_i - b \le \varepsilon \\ w^T x_i + b - y_i \le \varepsilon \end{cases},$$

where w and b are hyperplane parameters and ε is the insensitive loss.

The SVM regression is formulated as follows by adding two slack variables in Eq. (27).

$$\min_{w,b,\xi_i,\hat{\xi}_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\left(\xi_i + \hat{\xi}_i\right) \tag{27}$$

$$\text{s. t.} \begin{cases} f(x_i) - y_{\mathrm{i}} \le \varepsilon + \xi_i \\ y_{\mathrm{i}} - f(x_i) \le \varepsilon + \hat{\xi}_i \\ \xi_{\mathrm{i}} \ge 0, \hat{\xi}_{\mathrm{i}} \ge 0, i = 1,2, \ldots, \mathrm{N} \end{cases}$$

where ξ_i and $\hat{\xi}_i$ are the two slack variables; and C is a parameter that controls the trade-off between the support line and training samples. The solution of Eq. (27) is described in Garmdareh et al. (2018);Gizaw and Gan (2016).

(2) RF

RF regression is a representative type of ensemble machine learning model. Unlike SVM, which makes decisions based on a single trained model, RF is based on the

average result of numerous independent regression tree models (RTM). In RF, N subsets were selected using a Bootstrap aggregating method from the whole training samples, where n is the number of subsets. For each subset T $= \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, an RTM is developed by minimizing the loss as Eq. (28).

$$min \frac{1}{n} \sum_{m=1}^{M} \sum_{x_i \in R_m} (p_m - y_i) \quad (28)$$

Where *x* is the input; and *y* is the observed training target; M is the amount of leaf of an RTM; R is the subset of whole model inputs; p_m is the predicted value of leaf *m*.

In each RTM, the factors were randomly selected for model development and the final prediction of the RF model is calculated as the average of the results of different RTMs. This strategy means RF usually has good performance in terms of reducing overfitting, outliers and noise (Zhao et al., 2020;Zhao et al., 2018).

The out-of-bag (OOB) samples (samples not selected by the bootstrap method) are applied to test its accuracy. Once an RF is developed, the error of OOB samples can be computed as Eq. (29).

$$E_{OOB} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i) \quad (29)$$

Where n is the total number of OOB samples; $\hat{y}_i$ is the predicted value of RF.

Each factor in the OOB samples is permuted one at a time, and the permuted E_{OOB} can be computed with the permuted OOB samples and the trained RF model. The RF estimates the factor importance by comparing the difference between the original and permuted E_{OOB}while all others are unchanged. RF has been successfully applied for tasks such as flood assessment, discharge prediction and ranking of hydrological signatures (Zhao et al., 2018;Hutengs and Vohland, 2016;Li et al., 2016), including RFFA at regional scales (Desai and Ouarda, 2021) .

(4): Section 3.5, both metrics used for validation, i.e. the RMSNE and RBIAS, focus on evaluating the deviation of the results to the truth. I suggest adding other metrics (such as the KGE and NSE) that account for the correlation, bias, and scattering at the same time.

Reply: Thanks for your kind comments. The most important for RFFA is the deviation of the simulated at-site discharge to the truth. Therefore, RMSNE and RBIAS are widely used in RFFA evaluation and easy to compare with other regional studies. KGE and NSE are widely used for evaluation in time-series problems. We calculated the NSE and KGE value under different return periods as shown in Table R3.

Table R3 Result for different return period flood estimation

Return periods	Metrics in the testing period			
	Mean BIAS	Mean RMSNE	Mean KGE	Mean NSE
10	-0.165	0.664	0.586	-0.177
20	-0.168	0.672	0.580	-0.174
50	-0.166	0.684	0.573	-0.189
100	-0.174	0.708	0.552	-0.394

We found that NSE shows conflicting results with the other three metrics. This is mainly because the mean flow of a subgroup is used as a benchmark in the NSE calculation. The stations in this research were delineated based on the K-means method and twelve catchment descriptors. The deviation of the at-site discharge to the mean observed flow of the subgroup cannot be regarded as a good benchmark. We found that very few RFFA studies adopted KGE and the KGE metric shows similar results with the RBIAS and RMSNE in this research. Therefore, we still used RMSNE and RBIAS for evaluation.

(5): There are a few places where acronyms are used without definition, e.g. line 71. They all need to be defined the first time they appear in the text.

Reply: Thanks. We revised it according to your kind suggestion in the revised manuscript.

References

Desai, S., and Ouarda, T. B.: Regional hydrological frequency analysis at ungauged sites with random forest regression, J Hydrol, 594, 125861, 2021.

Djodjic, F., Bieroza, M., and Bergström, L.: Land use, geology and soil properties control nutrient concentrations in headwater catchments, Sci Total Environ, 145108, 2021.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., and Vapnik, V.: Support vector regression machines, Advances in neural information processing systems, 9, 155-161, 1997.

Garmdareh, E. S., Vafakhah, M., and Eslamian, S. S.: Regional flood frequency analysis using support vector regression in arid and semi-arid regions of Iran, Hydrolog Sci J, 63, 426-440, 10.1080/02626667.2018.1432056, 2018.

Gizaw, M. S., and Gan, T. Y.: Regional Flood Frequency Analysis using Support Vector Regression under historical and future climate, J Hydrol, 538, 387-398, 10.1016/j.jhydrol.2016.04.041, 2016.

Hutengs, C., and Vohland, M.: Downscaling land surface temperatures at regional scales with random forest regression, Remote Sens Environ, 178, 127-141, 10.1016/j.rse.2016.03.006, 2016.

Li, B., Yang, G. S., Wan, R. R., Dai, X., and Zhang, Y. H.: Comparison of random forests and other statistical methods for the prediction of lake water level: a case study of the Poyang Lake in China, Hydrol Res, 47, 69-83, 10.2166/nh.2016.264, 2016.

Niadas, I. A.: Regional flow duration curve estimation in small ungauged catchments using instantaneous flow measurements and a censored data approach, J Hydrol, 314, 48-66, 2005.

Tsegaw, A. T., Alfredsen, K., Skaugen, T., and Muthanna, T. M.: Predicting hourly flows at ungauged small rural catchments using a parsimonious hydrological model, J Hydrol, 573, 855-871, 2019.

Zhao, G., Pang, B., Xu, Z. X., Yue, J. J., and Tu, T. B.: Mapping flood susceptibility in mountainous areas on a national scale in China, Sci Total Environ, 615, 1133-1142, 10.1016/j.scitotenv.2017.10.037, 2018.

Zhao, G., Bates, P., and Neal, J.: The impact of dams on design floods in the Conterminous US, Water Resour Res, 56, e2019WR025380, 2020.