Authors' response to interactive comment by Anonymous Reviewer #2

Overall Comments: The paper proposes a machine learning based approach to estimate design floods globally. It includes three stages. First is using Anderson-Darling test and Bayesian MCMC method to choose suitable distribution and estimate parameters. Then at-site frequency curve will be sure. Second is clustering these stations into subgroups by a L-means model based on 12 globally available catchment descriptors. Third is developing a regression model in each subgroup for regional design flood estimation using the same descriptors. 11793 stations' data is used to predict regional flood and a support vector machine regression provide the highest prediction performance with root mean square normalized error of 0.708 for 100-year return period flood estimation and relative mean relative biases of all climate types being less than 20%. This paper proposes a large-scale regional flood estimation method by machine learning which covers 11793 stations globally. The method performance is also satisfactory compared with previous work. However, there are still some shortcomings. Some explanations should be complemented and the negative value of RBIAS should be analysed more.

We thank the reviewer for his/her valuable comments and suggestions that have undoubtedly helped us improve our manuscript. We revised the manuscript (corrections noted in blue-colored text) and reply to each comment as follows. We hope that these responses and revisions meet your expectations.

Specific Comments:

(1): Page 6: "... These explanatory factors can be grouped into four categories as follows: ...". A correlation analysis of all factors can be done to make sure they have weak correlation with each other.

Reply: This is very helpful. We analysed the correlation of all factors and the result is shown in Table R1. According to the criteria proposed by Evans (1996), no factor pairs show a strong correlation (Pearson's correlation coefficient > 0.6). We clarified this point in the revised manuscript.

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

(2): Page 9: "... The adopted Anderson-Darling test and Bayesian MCMC method are briefly described as follows. ...". A brief introduction of distributions is necessary. For example, Pearson type three distribution is used widely in China.

Reply: Thanks for your kind suggestion. We added more description about the adopted distributions in the revised manuscript.

Line 70: Studies have suggested alternate distributions, such as , GEV, generalized logistic (GENLOGIS), Pearson type III (P3), whilst lognormal (LN3) distributions are mandated for use in the US (Committee, 1981).

Line 210: Table 3 describes the tested hypothetical distributions which have been widely recommended in flood frequency analysis in different countries. For example, the generalized pareto distribution performs better in populated regions of Australia and

P3 is recommended as the national standard distribution for flood frequency analysis in China (Gao et al., 2019;Vogel et al., 1993;Wang et al., 2015).

(3): Page 11: "... The adopted Bayesian MCMC method was proposed by Reis and Stedinger (2005) and is reported to provide better parameter estimates than the MOM and MLE approaches in some studies ...". L-moment method is a valid method on estimating the parameters. Bayesian MCMC method should compare with it.

Reply: I agree with you. The L-moment method is also a valid technique for estimating the parameters.

We compared the 100-year return period flow calculate by L-moments and Bayesian MCMC methods. As shown in Figure R3, the 100-year return period flows calculated by the two methods are similar (Pearson's correlation coefficient = 0.99).

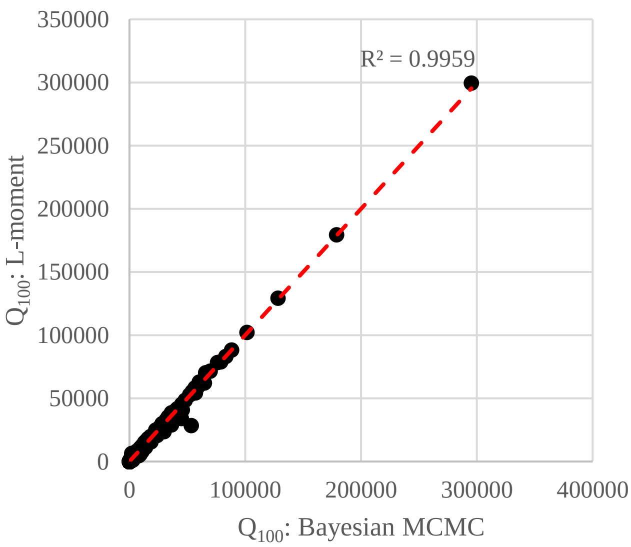

Figure R3 result of 100-year flow calculated by the Bayesian MCMC versus L-moment

We further tested the results of 100-year flow estimation in the training and validation periods using different design floods calculated by L-moments and Bayesian MCMC methods. We found that the methods of at-site flood estimation will not affect the results, and both are valid methods for flood regional flood frequency analysis. Some studies have shown that at-site quantiles are estimated using at-site L-moments which are known to be inferior to regional flood quantiles (Martins and Stedinger, 2000). Compared with L-moments, the Bayesian MCMC method can provide credible intervals for flood estimation. Therefore, we adopted the Bayesian MCMC method in this study.

Table R2 Results of 100-year flood using different methods for local flood estimation

Methods	**Training**		**Testing**	
	Mean BIAS	Mean RMSE	Mean BIAS	Mean RMSE
L-moment	-0.175	0.692	-0.171	0.698
Bayesian MCMC	-0.179	0.703	-0.174	0.708

(4): Page 11: "... The detail of the Bayesian MCMC method is comprehensively described in the research of Reis and Stedinger (2005) ...". Although its numerical method will be complicated, a brief explanation is still essential.

Reply: Thanks. We added a detailed description of the Bayesian MCMC in the section 3.2.2 and as follows.

The adopted Bayesian inference consisted of three steps as follows:

(a): Prior distribution and likelihood function calculation. The first step is to determine the prior distribution for Bayesian analysis. As non-prior knowledge (i.e. population distribution function) was considered, a non-informative Normal distribution was selected as a prior distribution. The likelihood function $f(q|\theta)$ can be computed as Eq. (15).

$$f(q|\theta) = \prod_{i=1}^{n} f(q_i;\theta) \tag{15}$$

Where: $q = (q_1, q_2, q_3, \dots, q_n)$ are the given samples and θ is a parameter vector;

(b): Posterior distribution calculation. The posterior distribution $f(\theta|q)$ is computed using Bayesian inference as in Eq. (16). As the integral in Eq. (16) cannot be solved analytically, the Metropolis-Hastings MCMC method was used to generate samples from the posterior distribution.

$$f(\theta|q) = \frac{f(q|\theta)\pi(\theta)}{\int f(q|\theta)\pi(\theta)\, d\theta} \qquad (16)$$

Where $\pi(\theta)$ is the density function of the prior distribution;

(c): parameter and design flood estimations. The final parameters $\hat{\theta}$ can be calculated by the expected value of the posterior distribution; The probability density function of design flood Q can be described as Eq. (17).

$$f(Q|q) = \int_{\Theta} f(Q|\theta) f(\theta|q) d\theta \qquad (17)$$

The three common methods for estimating the parameters of at-site flood frequency curves are based on moments (MOM), maximum likelihood (MLE) and Bayesian inference. Compared with MOM and MLE, the Bayesian approach can provide credibility intervals for the estimated design flood. The details of the adopted approach is comprehensively described in the research of Reis and Stedinger (2005) and the calculation is implemented based on a *Bayesian MCMC* function (nsRFA package) in the R software.

(5): Page13: "... SVM regression has shown advantages in solving complicated non-linear problems in the field of hydrology ...". As a major method of this article, the introduction of SVM may be too simple. More detailed description can be added.

Reply: This is very helpful. We added more descriptions about the implementation of SVM in Section 3.4.2 and as follows.

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).

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

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

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).

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

$$\text{s.t.} \begin{cases} f(x_i) - y_i \leq \varepsilon + \xi_i \\ y_i - f(x_i) \leq \varepsilon + \hat{\xi}_i \\ \xi_i \geq 0, \hat{\xi}_i \geq 0, i = 1,2, \ldots, N \end{cases}$$

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).

(6): Page 13: "... RF regression is a representative type of ensemble machine learning model ...". Math is the best language of science. Several mathematical formulas of RF will help readers to understand it abstractly.

Reply: Thanks. We added mathematical formulas of RF in Section 3.4.3 as follows.

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).

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

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).

$$E_{OOB} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y_i}) \tag{29}$$

Where n is the total number of OOB samples; $\widehat{y_i}$ 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) .

(6): Page 18: "... Figure 7 (a) Factor importance evaluated by RF model and (b) the impact of catchment descriptors for regression ...".

Reply: Thanks. We revised this figure as follows to make it easier for readers to understand.

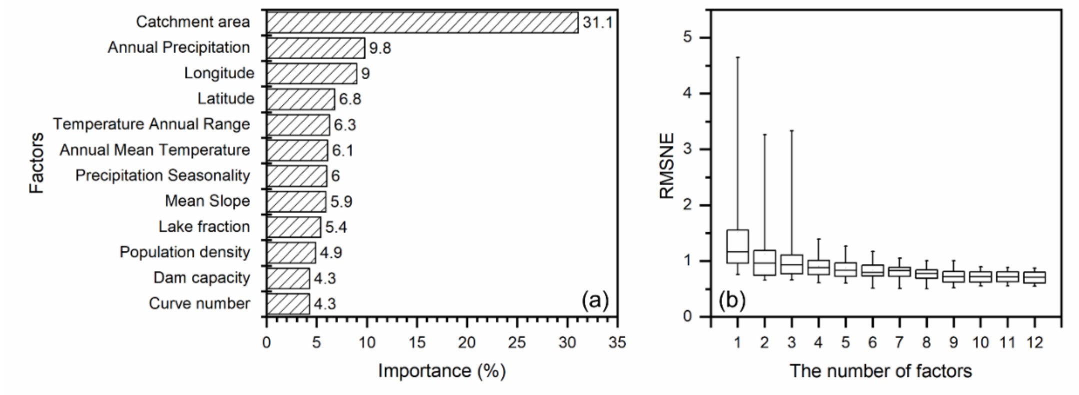

Figure 7 (a) Factor importance evaluated by RF model and (b) the impact of catchment descriptors for SVM regression

(7): Page 19: "... The negative value of RBIAS reflected some overestimation which mainly occurred due to low discharge in small catchments ...". It is normal to underestimate 100-year return period floods, but why all the RBIAS indexes are negative? RBIAS index is just a relative index so the absolute value of discharge should not take much effect on the index. Please analyze more about it.

Reply: This is very helpful. We analysed the relationship between R-BAIS and catchment size in the revised manuscript.

As shown in Figure R4, both over and underestimations were found from small to large catchments. The range of RBIAS in small catchments 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. The mean RBIAS represents the overall model performance of a subgroup and it is highly dependent on the nature of the model and the estimated at-site design floods. Some studies have suggested that the decreasing (increasing) trends of observed discharge can lead to overestimation (underestimation) of flood quantiles if such nonstationarity is not taken into consideration in an RFFA (Kalai et al., 2020;O'Brien and Burn, 2014). The negative value of mean RFFA may be caused by the overall decreasing trend for global discharge. This negative value is also consistent with some similar studies using machine learning models in RFFA (Desai and Ouarda, 2021;Shu and Ouarda, 2008). We have analysed this point in the discussion section.

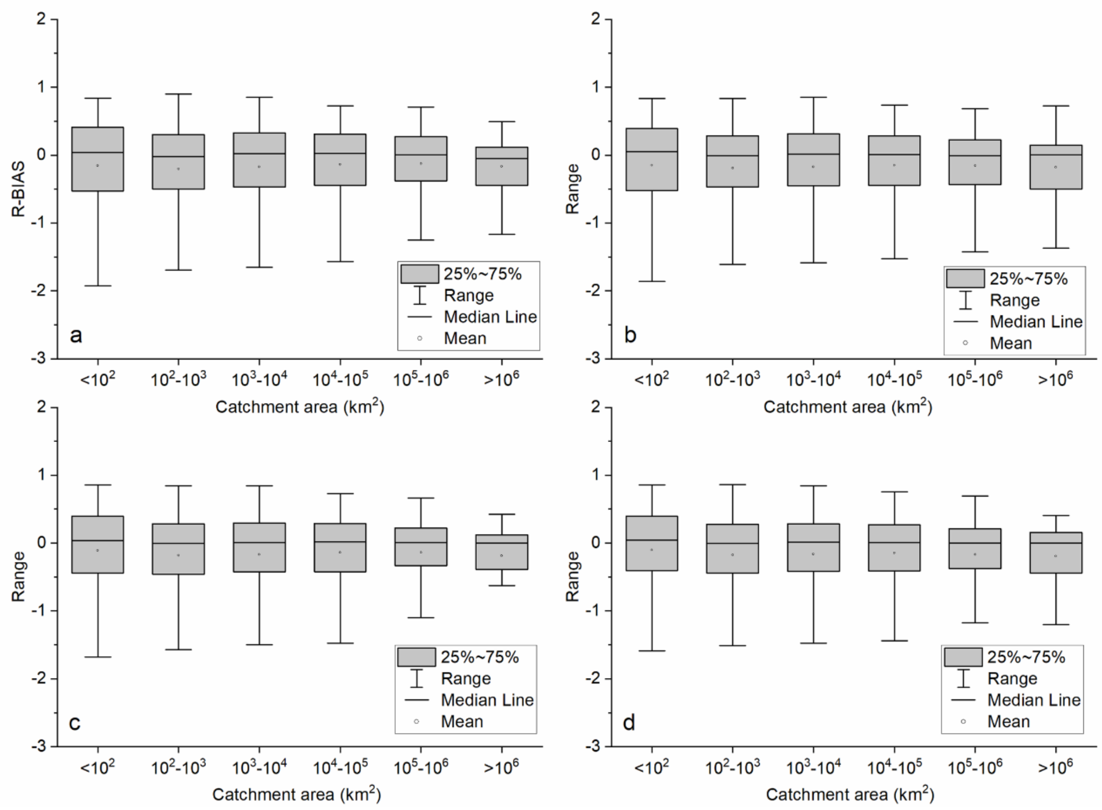

Figure R4 R-BIAS of (a) 100, (b) 50; (c) 20; and (d) 10-year return period flood in different catchment sizes.

References

Martins, E.S. and J.R. Stedinger, Generalized maximum-likelihood extreme-value quantile estimators for hydrologic data, Water Resources Research, 36(3), 2000.

Committee, W. R. C. H.: Guidelines for determining flood flow frequency, US Water Resources Council, 1981.

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

Gao, S., Liu, P., Pan, Z., Ming, B., Guo, S., Cheng, L., and Wang, J.: Incorporating reservoir impacts into flood frequency distribution functions, J Hydrol, 568, 234-246, 2019.

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.

Kalai, C., Mondal, A., Griffin, A., and Stewart, E.: Comparison of nonstationary regional flood frequency analysis techniques based on the index-flood approach, J Hydrol Eng, 25, 06020003, 2020.

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.

O'Brien, N. L., and Burn, D. H.: A nonstationary index-flood technique for estimating extreme quantiles for annual maximum streamflow, J Hydrol, 519, 2040-2048, 2014.

Reis, D. S., and Stedinger, J. R.: Bayesian MCMC flood frequency analysis with historical information, J Hydrol, 313, 97-116, 2005.

Shu, C., and Ouarda, T. B. M. J.: Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system, J Hydrol, 349, 31-43, 10.1016/j.jhydrol.2007.10.050, 2008.

Vogel, R. M., McMahon, T. A., and Chiew, F. H.: Floodflow frequency model selection in Australia, J Hydrol, 146, 421-449, 1993.

Wang, J., Liang, Z., Hu, Y., and Wang, D.: Modified weighted function method with the incorporation of historical floods into systematic sample for parameter estimation of Pearson type three distribution, J Hydrol, 527, 958-966, 2015.

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.