Design flood estimation for global river networks based on machine learning models

Gang Zhao[1], Paul Bates[1,2], Jeffrey Neal[1,2], Bo Pang[3]

[1] School of Geographical Sciences, University of Bristol, Bristol, UK
[2] Fathom, Engine Shed, Station Approach, Bristol, UK
[3] College of Water Sciences, Beijing Normal University, Beijing, China

Correspondence to: Gang Zhao (gang.zhao@bristol.ac.uk)

Abstract. Design flood estimation is a fundamental task in hydrology. In this research, we propose a machine learning based approach to estimate design floods globally. This approach involves three stages: (*i*) estimating at-site flood frequency curve for global gauging stations by the Anderson-Darling test and a Bayesian MCMC method; (*ii*) clustering these stations into subgroups by a K-means model based on twelve globally available catchment descriptors, and (*iii*) developing a regression model in each subgroup for regional design flood estimation using the same descriptors. A total of 11793 stations globally were selected for model development and three widely used regression models were compared for design flood estimation. The results showed that: (1) the proposed approach achieved the highest accuracy for design flood estimation when using all twelve descriptors for clustering; and the performance of the regression was improved by considering more descriptors during training and validation; (2) a support vector machine regression provided the highest prediction performance amongst all regression models tested, with root mean square normalised error of 0.708 for 100-year return period flood estimation; (3) 100-year design floods in tropical, arid, temperate, cold and polar climate zones could be reliably estimated with relative mean relative biases (RBIAS) of -0.199, -0.233, -0.169, 0.179 and -0.091 respectively (i.e. <20% error); (4) the machine learning based approach developed in this paper showed considerable improvement over the index-flood based method introduced by Smith et al. (2015, https://doi.org/10.1002/2014WR015814) for design flood estimation at global scales; and (5) the average RBIAS in estimation is less than 18% for 10, 20, 50 and 100-year design floods. We conclude that the proposed approach is a valid method to estimate design floods anywhere on the global river network, improving our prediction of the flood hazard, especially in ungauged areas.

1 Introduction

Flood hazard is the primary weather-related disaster worldwide, affected 2.3 billion people and caused 662 billion US dollars in economic damage between 1995 and 2015 (CRED and UNISDR, 2015). The frequency and severity of flood events are expected to increase in the future because of climate change and socio-economic growth in flood-prone areas (Sharma et al., 2018;Wing et al., 2018;Winsemius et al., 2016). Flood hazard models are mature tools to identify flood-prone areas and have

been widely used in flood risk management at catchment or regional scales (Hammond et al., 2015;Teng et al., 2017). With the development of new remote sensing techniques and an increase in computing power, global flood hazard models (GFHMs) are now a practical reality and have been successfully applied for large-scale flood mapping and validated in several countries (Bates et al., 2020;Schumann et al., 2018). GFHMs can identify flood-prone areas in ungauged basins and provide a consistent and comprehensive understanding of the flood hazard at national, continental and global scales.

The 'cascade' model type and 'gauged flow data' model type are the two most frequently used approaches in global flood hazard modelling and both map the flood hazard based on return period flows (Trigg et al., 2016). The cascade model type, for example CaMa-UT (Yamazaki et al., 2011) and GLOFRIS (Winsemius et al., 2013), uses land surface models driven by climate reanalysis data to simulate streamflow. Return period flows along the river network in the cascade model type are derived by an at-site flood frequency analysis of the resulting land surface model streamflow. However, due to the coarse resolution (usually 0.5 degrees) of global climate and land surface models (Yang et al., 2019b;Liu et al., 2019), some downscaling and bias correction methods usually need to be adopted for high-resolution flood hazard mapping (Mueller Schmied et al., 2016;Frieler et al., 2017;Schumann et al., 2014a). Unlike the cascade model type, the gauged flow data model type uses observed gauged discharge and regional flood frequency analysis (RFFA) approaches to produce different return period flows along the global river network (Trigg et al., 2016). Compared with the cascade model type, the gauged flow data model type can be preferable for high-resolution flood hazard mapping as it avoids the uncertainties from rainfall-runoff modelling (Prihodko et al., 2008) and flood map downscaling (Schumann et al., 2014b). However, the performance of the RFFA approach adopted is highly dependent on the coverage and density of observed discharge stations (Hosking and Wallis, 1988;Lin and Chen, 2006).

RFFA has long been regarded as a reliable method to estimate design floods in engineering hydrology (Cunnane, 1988). Based on the basic assumption that there is a relationship between catchment descriptors and the design flood in a region, different types of RFFA approaches have been proposed and successfully applied in regional studies in the past five decades (Griffis and Stedinger, 2007;Merz and Bloschl, 2008b, a;Dalrymple, 1960). Amongst these, the index-flood method and the direct regression method are two of the most widely used procedures (Shu and Ouarda, 2008). The index-flood method assumes that the flood frequency curves at different sites in a region are identical except for one scale index (named as the index-flood) (Dalrymple, 1960;Bocchiola et al., 2003). Therefore, the index-flood method involves two steps; index-flood estimation and derivation of a single regional flood frequency relationship, often termed a growth curve. Unlike the index-flood method, the direct regression method predicts flood quantiles as a function of catchment descriptors directly (Shu and Ouarda, 2008). These two methods have been successfully applied at both regional and national scales (Salinas et al., 2013), but very few applications of such methods have been performed at the global scale. To date, Smith et al. (2015) have proposed an RFFA approach at the global scale based on the index-flood method (termed global RFFA). This approach has been applied successfully in high-resolution flood hazard mapping for ungauged areas in several countries (Sampson et al., 2015;Wing et al., 2017).

However, the global RFFA approach of Smith et al (2015) has shown some deficiencies during application and which we seek to address in this research as follows:

(1) the global RFFA approach of Smith et al (2015) was developed based on 945 stations in the Global Runoff Data Base (GRDB) and the United States Geological Survey (USGS) database. This limited number of stations means that it cannot provide a reliable estimation in some regions at global scale. To improve the coverage and density of discharge stations, nearly 12,000 stations from the newly published Global Streamflow Indices and Metadata (GSIM) archive were selected for model development in this research.

(2) the flood frequency curve in the Smith et al (2015) approach was assumed to obey a generalized extreme value (GEV) distribution for the global coverage. 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). In this research, the at-site flood frequency curve was selected from eight widely used distributions based on the Anderson-Darling goodness of fit tests.

(3) the global RFFA approach of Smith et al. (2015) adopted only three catchment descriptors (rainfall, slope and catchment area) for clustering and flood estimation. These three factors can only provide a basic description of global catchment characteristics. In this research, twelve catchment descriptors covering meteorological, physiographical, hydrological and anthropological aspects were considered for clustering and design flood estimation.

(4) the global RFFA approach of Smith et al. (2015) was proposed based on the index-flood method and the index-flood in ungauged areas was computed with a power-form function. As the coefficient of variation of flood flows generally varies from site to site, the index-flood method is not recommended if available samples contain more than 10 observations or are highly cross-correlated (Stedinger, 1983). Moreover, the simple power-form function may fail to capture complicated relationships within the data, since the relationship between flood and each descriptor is assumed to obey an explicit power-form function which may not always be appropriate. In this research, the design flood in ungauged areas is estimated based on direct regression method and machine learning models were adopted to describe the unknown relationship between catchment descriptors and at-site design floods. These machine learning based methods have shown advantages over ordinary regression approaches in RFFA at regional scales (Gizaw and Gan, 2016;Shu and Ouarda, 2008;Zhang et al., 2018), and are tested here for the first time in a global study.

The aim of this research is therefore to provide an improved method for reliable design flood estimation addressing the deficiencies identified with the previous global scale study of Smith et al (2015). A three-phase model framework is applied where Bayesian MCMC, K-means and SVM models were applied for at-site flood estimation, subgroup delineation and regional flood estimation, respectively. Specifically, 11,793 stations selected from the GSIM archive are used for model development and the stations were divided into 16 subgroups using a K-means model and twelve catchment descriptors. For each subgroup, 70% and 30% stations were randomly selected for model training and validation respectively. Three types of regressions including Power Form function (PF), Support Vector Machine (SVM) and Random Forest (RF) were compared

for regional flood estimation. Finally, the proposed direct regression approach is compared with the previous results of Smith et al. (2015) in each climate zone.

2 Study areas and data

2.1 Flood data

Observed annual maximum streamflow data in the GSIM archive were used to estimate at-site design floods (DFs) at different return periods. This archive is a collection of streamflow indices that includes more than 35,000 stations from seven national and five international streamflow databases (Gudmundsson et al., 2018;Do et al., 2018). Compared with a widely used database in large-scale hydrological studies, such as the Global River Discharge Database (GRDB) which contains data from 9500 stations, the GSIM archive can help to improve the understanding of large-scale hydrological processes by improving the coverage and density of streamflow observations. After quality control of the daily streamflow data, 15 types of time-series indices are provided at yearly, seasonal, and monthly resolutions in the GSIM archive (Do et al., 2018;Gudmundsson et al., 2018). In this research, time series of annual maximal floods in the GSIM archive were adopted for at-site design flood estimation. To date, this archive has been successfully used in flood classification (Stein et al., 2019), streamflow trend analysis (Do et al., 2019) and hydrological model evaluation (Yang et al., 2019a) at the global scale.

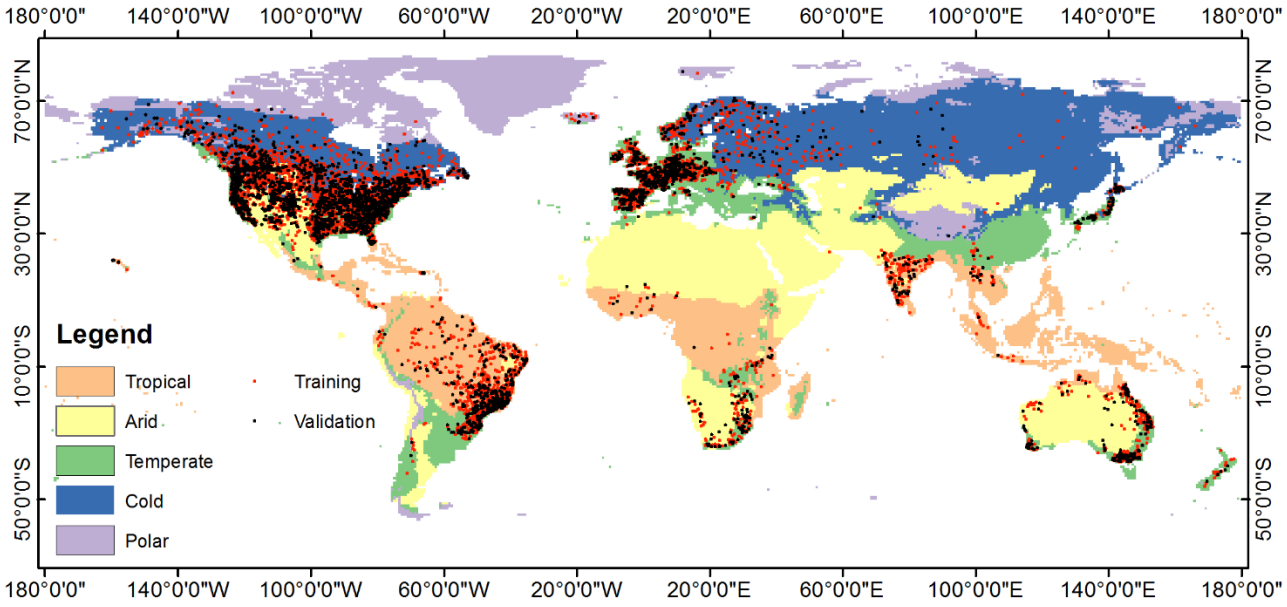


Figure 1 The distribution of discharge stations used for training and validation in this research.

To make reliable estimates for at-site DFs, a station selection is needed based on some quality control criteria. Firstly, the hydrological signatures of streamflow are reported to become stable during estimation with at least 20 years of record (Richter

et al., 1997). Therefore, only stations with a historical streamflow record of 20 years or longer were selected for the analysis. Second, the RFFA approach requires the assumption of flood stationarity. Stations experiencing obvious trends or sudden changes were detected by a Mann-Kendall test (MKT) (Hamed, 2008) and the standard normal homogeneity test (SNHT) (Alexandersson, 1986). Any stations exhibiting obvious non-stationarity, or which do not obey the given hypothetical distributions identified by Anderson-Darling goodness of fit tests were not considered further in model development. Lastly, to better estimate DFs at the global scale, the selected stations were first divided into several sub-groups based on a K-means clustering model, and separate regression models were developed for the different sub-groups. The streamflow stations in each sub-group were also required to pass discordancy and heterogeneity tests. The adopted stationarity, discordancy and heterogeneity measures are described in Section 3 and the distribution of training and validation stations after selection is shown in **Figure 1.**

Table 1 compares the number of stations used for model training and validation in each climate class between Smith et al. (2015) and this study. Compared with the research of Smith et al. (2015), the number of stations in this research is significantly improved in all climate classes.

Table 1 The number of discharge stations for model development in each climate class

Climate Class	Smith et al. (2015)		This research	
	Training	Validation	Training	Validation
Tropical	163	8	716	298
Arid	121	55	595	231
Temperate	296	99	4231	1816
Cold	109	80	2666	1099
Polar	14	N/A	100	41
Total	703	242	8308	3485

2.2 Catchment descriptors

The main idea of an RFFA is to characterise a relationship between at-site DFs and catchment descriptors and then apply this relationship to estimate flood magnitudes along the river network in similar ungauged catchments. In this research, the river network of the global coverage is extracted from the 1km resolution catchment area map for catchments exceeding a threshold area of 50 km^2. Twelve explanatory factors collected from open-access databases were selected as potential descriptors and applied for clustering and regression model development. A correlation analysis of all explanatory factors was done before model development to ensure none of them have strong correlation (Pearson's correlation coefficient > 0.6) with each other. The statistics and data sources of these factors are summarised in **Table 2**.

Table 2 Statistics and data source of the explanatory descriptors

No.	Factor name (Abbreviation)	Unit	Min	Max	Data source
1	Annual Precipitation (AP)	mm	0	7743	WorldClim
2	Precipitation Seasonality (PS)	-	0	219.35	WorldClim
3	Annual Mean Temperature (AT)	℃	-27.57	33.44	WorldClim
4	Temperature Annual Range (TR)	℃	0	74.79	WorldClim
5	Mean Slope (SL)	degree	0	32.78	MERIT DEM
6	Lake fraction (LF)	%	0	1	GLWD
7	Longitude (LO)	degree	-180	180	WGS 84
8	Latitude (LA)	degree	-60	85	WGS 84
9	Curve number (CN)	-	0	95	NRCS CN dataset
10	Catchment area (CA)	km^2	50	16198686	MERIT DEM
11	Dam capacity (DC)	million m^3	0	2648	GRanD
12	Population density (PD)	number/km^2	0	5317	GPW

These explanatory factors can be grouped into four categories as follows:

(1). *Meteorological factors* included annual precipitation (AP), precipitation seasonality (PS), annual mean temperature (AT), and annual temperature range (TR). AP and AT represent the average annual precipitation and mean temperature of the upstream catchment respectively. PS is the coefficient of variation of the monthly rainfall series and TR is the range between the maximal and minimal temperatures of the time series average over the upstream catchment. These four factors were collected from the WorldClim dataset (V2) at 30 arcsecond (~1km) resolution (Fick and Hijmans, 2017) and are used to represent the extreme value and seasonal distribution of precipitation and temperature.

(2). *Physiographical factors* comprised discharge station location, slope (SL) and lake fraction (LF) of the upstream catchment. Station location is defined by the longitude (LO) and latitude (LA) of the discharge station in degree units of the World Geodetic System 1984 (WGS84) spatial reference frame. SL reflects the average slope of the upstream catchment which is calculated based on the MERIT DEM (Yamazaki et al., 2017). LF represents the fraction of the catchment area upstream of the station covered with lakes. The location and area of global lakes are taken from the Global Lakes and Wetlands Database (GLWD) (Lehner and Doll, 2004).

(3). *Hydrological factors* Catchment area (CA) and Curve Number (CN) reflect the area and runoff capacity of the upstream catchment respectively. CA is defined as the upstream flow accumulation area based on the D8 algorithm and the Merit DEM. CN is an empirical parameter for runoff prediction in ungauged areas which is calculated from the tables in the National Engineering Handbook of the United States, Section-4 (NEH-4) (Mockus, 1964). The CN in this research is collected from a global CN dataset that utilizes the latest Moderate Resolution Imaging Spectroradiometer (MODIS) land cover information and the Harmonized World Soil Database (HWSD) soil data (Zeng et al., 2017).

(4). *Anthropological factors* include total dam capacity (DC) and population density (PD). DC is the total dam capacity of the upstream catchment. The location and capacity of each dam is provided by the Global Reservoir and Dam (GRanD) dataset which includes about 7000 dams globally (Lehner et al., 2011). PD is a widely used factor to reflect the impact of human activities on the environment. The PD map in this research is collected from the Gridded Population of the World (GPW) dataset as of the year 2015 (Doxsey-Whitfield et al., 2015).

3 Methodology

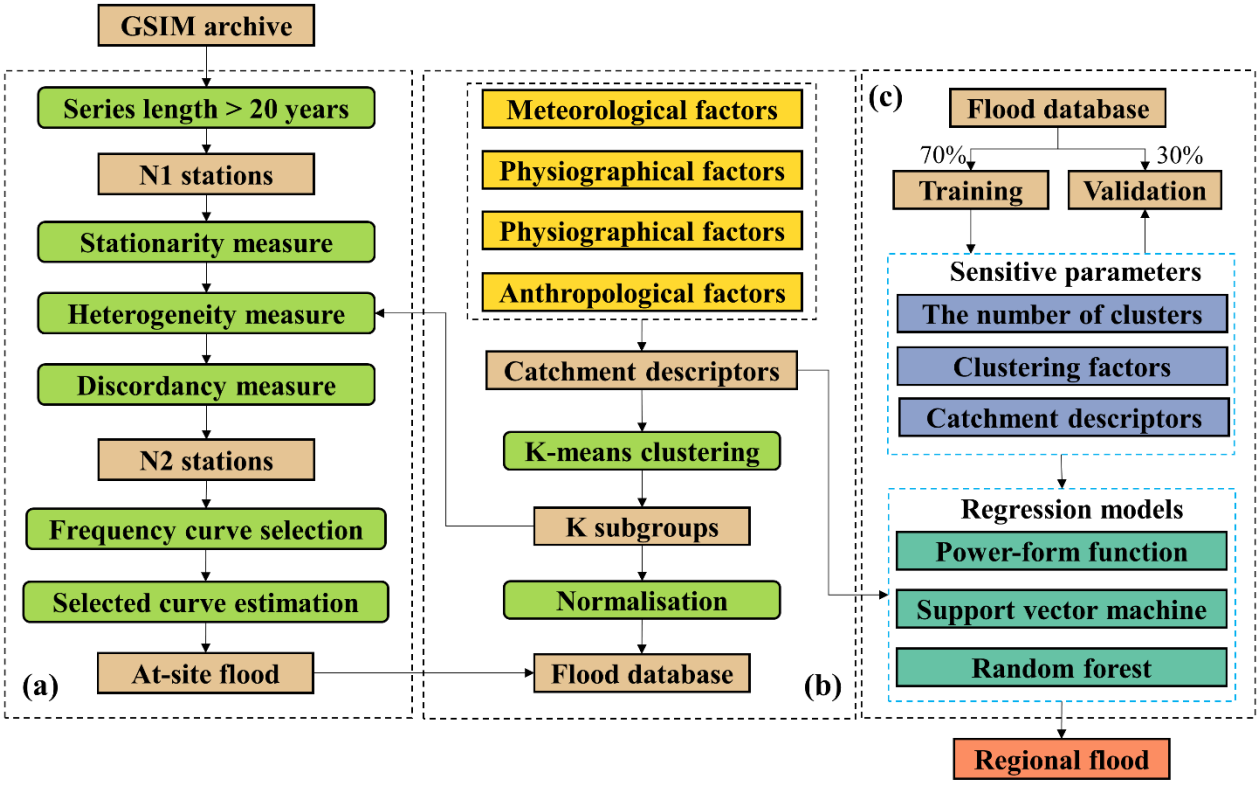


Figure 2 The framework of this research
(a) at-site flood estimation; (b) subgroups delineation; (c) regional flood estimation

The flowchart for this research has three parts which are shown in **Figure 2**. Part (a) includes the procedures of station selection and at-site design flood estimation that are described in Sections 3.1 and 3.2 respectively. Part (b) displays the procedure for subgroup delineation by using a K-means clustering model (in Section 3.3). Part (c) describes the procedure of regression model development and regional flood estimation in ungauged areas. Three regression models described in Section 3.4, were adopted for model comparison.

3.1 Station selection methods

3.1.1 Stationarity measures

Two widely used stationarity measures in hydrology, the Mann-Kendall test (MKT) (Hamed, 2008) and the standard normal homogeneity test (SNHT) (Alexandersson, 1986), were adopted for trend and change-point detection respectively in the research. One advantage of these two tests is that they provide an index reflecting the degree of trend or change. The details of the MKT and SNHY methods are described in Appendix A and B respectively.

The Z index in the MKT test was used to evaluate the trend of descending (negative values) and ascending (positive values) in the time series. The change-points occurred at the largest value of T index among the time series T_d, and a larger T represents a higher change. After stationarity measures were evaluated, each station has corresponding Z_i and T_i indices. The stations experiencing obvious trends and/or sudden changes were removed based on the 95% quantile value of the $|Z_i|$ and T_i indices.

3.1.2 Discordancy measures

The aim of using the discordancy measures was to detect potential outliers in each subgroup. Two discordancy measures are applied in this research to the time series of observed discharge. The first is the Z-score method (Shiffler, 1988) which is based on the value of mean annual runoff (R). The global gauging stations are divided into several subgroups and the R in each subgroup is normalised using Eq. (C1) in Appendix C and Z-score greater than 3 are regarded as outliers (Garcia, 2012).

Another discordancy measure was proposed by Hosking and Wallis based on L-moments ratios (Hosking and Wallis, 2005). The discordancy of each station can be evaluated by D_i as defined in Eq. (C2) to (C4) in Appendix C. As suggested by Hosking and Wallis (2005), a station is considered as discordant if $D_i \geq 3$ when N_k is larger than 15.

3.2 At-site flood estimation

Table D1 (in Appendix D) describes the tested hypothetical distributions which have been widely recommended in flood frequency analysis in different countries. For example, a generalized pareto distribution performs better in populated regions of Australia whilst 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). The at-site flood estimation involves two steps. Firstly, the flood frequency curve for each station was selected from eight widely used hypothetical distributions based on the Anderson-Darling goodness of fit

test. After the preferred distribution was selected, the parameters and design floods for this station were derived by a Bayesian MCMC method. The adopted Anderson-Darling test and Bayesian MCMC method are briefly described as follows.

3.2.1 Distribution selection

The Anderson-Darling (AD) test can detect whether a given sample of data follows a hypothetical distribution. It has been widely used in flood frequency analysis as it shows good skill for small sample sizes and heavy-tailed distributions (Haddad and Rahman, 2011). For each station, the A^2 statistic was calculated for all hypothetical distributions and the distribution with the minimum A^2 was selected as the flood frequency curve for that station. The calculation process for the A^2 statistic is described in Appendix D. After selection of flood frequency curve for all stations, a threshold $A^2_{95\%}$ was determined by the 95% quantile value of A^2_i. The stations at which A^2_i exceeded the threshold were regarded as not obeying all hypothetical distributions and these stations were not considered further for model development.

3.2.2 Parameter estimation

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. (1).

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

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. (2) cannot be solved analytically, the Metropolis-Hastings MCMC method was used to generate samples from the posterior distribution.

$$f(\theta|q) = \frac{f(q|\theta)\pi(\theta)}{\int f(q|\theta)\pi(\theta)\,d\theta}, \tag{2}$$

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. (3).

$$f(Q|q) = \int_{\Theta} f(Q|\theta) f(\theta|q) d\theta, \tag{3}$$

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.

3.3 Clustering method

3.3.1 K-means model

The standard K-means model was adopted for clustering in this research. This model has been widely used in the area of hydrology including in RFFA studies (Smith et al., 2015;Lin and Chen, 2006). The K-means model consists of two steps:

(a) An input dataset $D = \{x_j | j = 1,2,3 \dots, N\}$ where N is the number of samples. The K centroids $C = \{C_k | k = 1,2,3 \dots, \mathrm{K}\}$ are first randomly selected from the input dataset D. Each sample x_j is assigned to its nearest C_k based on the squared Euclidean distance as in Eq. (4).

$$\arg\min_{c_k \in C} [dis(C_k, x_j)]^2, \quad (4)$$

Where: $dis(-)$ is the function for Euclidean distance calculation.

(b) The centroid C_k in step (a) is updated as in Eq. (5). Then, the algorithm iterates between centre assignment (a) and centre update (b) until the maximum number of iterations is reached or cluster assignments do not change.

$$\mathrm{C}_k = \frac{1}{N_i} \sum_{j=1}^{N_i} x_j, \; x_k \in A_k, \quad (5)$$

Where A_k is the subset of D and N_i is the number of samples in A_k.

The K-means model was implemented based on the MATLAB *kmeans* function and the maximum number of iterations was set to 1000 (https://www.mathworks.com/help/stats/kmeans.html). Two sensitive parameters, the number of clusters and the clustering factors, were selected to maximize model performance during the validation procedure.

3.3.2 Heterogeneity measure

The heterogeneity of subgroup delineation is measured by a Davies-Bouldin index (Davies and Bouldin, 1979). This index considers both intra subgroup diversity and inter subgroup distances. The intra subgroup diversity of a subgroup *i* is computed as Eq. in (6)

$$S_i = \left\{ \frac{1}{T_i} \sum_{j=1}^{T_i} |x_j - C_i|^2 \right\}^{1/2}, \quad (6)$$

Where C_j is the centroid of subgroup i; x_j is the clustering factors of subgroup j; T_i is the number of samples in subgroup *i*.

The inter subgroup distance between subgroup *j* and subgroup *k* is measured as Eq. (7):

$$R_{i,j} = \frac{s_i + s_j}{dis(C_{i,}, C_j)}, \quad (7)$$

Where $dis(C_{i,}, C_j)$ is the Euclidean distance between C_i and C_j;

The Davies-Bouldin index is computed as Eq. (8) and (9)

$$R_i = \max_{i \neq j}\{R_{i,j}\}, \quad (8)$$

$$DBI = \frac{1}{K}\sum_{i=1}^{K} R_i, \quad (9)$$

Where K is the number of subgroups; DBI is called the Davies–Bouldin index and a lower value means that the subgroups are better clustered.

3.4 Regression method

The direct regression method estimates design floods directly with catchment descriptors using a regression model as described in Eq. (10). Three types of regression models were compared in this research and are described below. The optimal regression model for the direct regression method was selected according to the highest prediction accuracy during the validation period. To best of our knowledge, this research is the first three different regression models have been tested in an RFFA at the global scale.

$$DF = f_{FQ}(d_1, d_2, \dots, d_N), \quad (10)$$

Where DF is the design flood for specific return periods; f_{FQ} is the regression model for design flood estimation in each subgroup;

3.4.1 Power-form function (PF)

Power-form (PF) function is a simple non-linear regression model that has been successfully used in RFFA studies at a global scale (Smith et al., 2015). The basic formula of PF regression is described in Eq. (11).

$$DF = \beta_0 d_1^{\beta_1} d_2^{\beta_2} d_3^{\beta_3} \dots d_N^{\beta_N} \pm \varepsilon, \quad (11)$$

Where β_m are the model parameters and *d* is the explanatory descriptor used for model development; *N* is the total number of descriptors and ε is the error term.

3.4.2 Support vector machine (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 $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, 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. (12).

$$\min_{w,b} \frac{1}{2}\|w\|^2 , \quad (12)$$

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

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. (13).

$$\min_{w,b,\xi_i,\hat{\xi}_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i+\hat{\xi}_i), \quad (13)$$

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

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. (13) is described in Garmdareh et al. (2018);Gizaw and Gan (2016).

3.4.3 Random forest (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 = $\{(x_1, y_1), (x_2, y_2),\dots, (x_n, y_n)\}$, an RTM is developed by minimizing the loss as Eq. (14).

$$min\frac{1}{n}\sum_{m=1}^{M}\sum_{x_i\in R_m}(p_m - y_i), \quad (14)$$

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. (15).

$$E_{OOB} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i}), \quad (15)$$

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) .

3.5 Validation indices

Two widely used indices, the root mean square normalised error (RMSNE) and the relative mean bias (RBIAS) were applied for model evaluation (Salinas et al., 2013;Shu and Ouarda, 2008;Smith et al., 2015). The RMSNE and RBIAS are computed as in Eq. (16) and (17) respectively. RMSNE is a negatively oriented index where lower value means better model performance. Since the errors in RMSNE are squared before they are averaged, it should be more useful than RBIAS when large errors are particularly undesirable. However, as well as its use in error evaluation, RBIAS can describe if the model has positive bias (underestimation) or negative bias (overestimation).

$$\mathrm{RMSNE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{q_i-\widehat{q_i}}{q_i}\right)^2}, \tag{16}$$

$$\mathrm{RBIAS} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{q_i-\widehat{q_i}}{q_i}\right), \tag{17}$$

Where q_i is the discharge for specific return periods derived by observed data, $\widehat{q_i}$ is the discharge for specific return periods estimated by the RFFA method and N is the total number of stations.

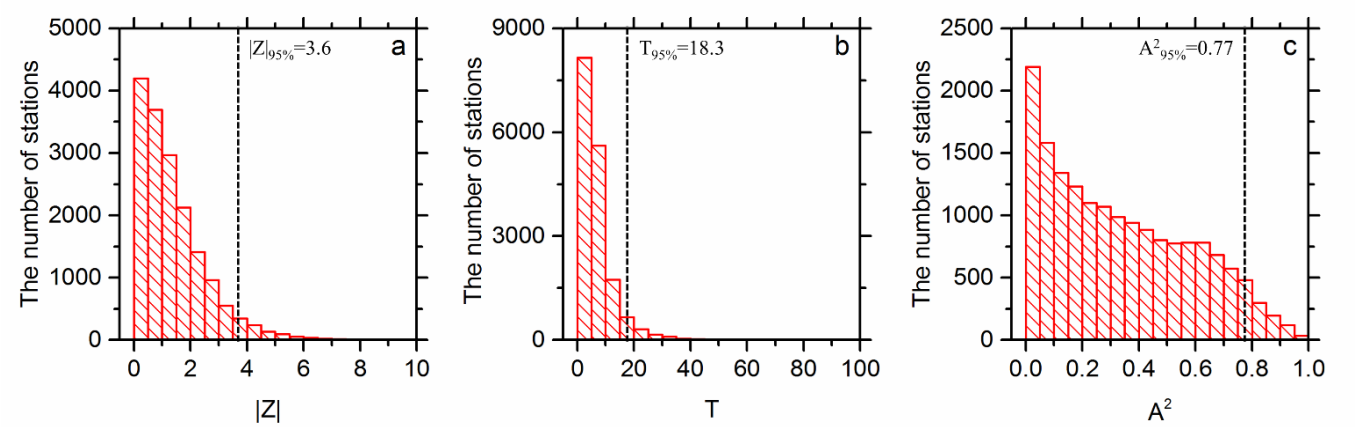
4 Results and discussion

4.1 At-site flood estimation

Firstly, 16854 stations with a historical streamflow record of 20 years or longer were selected from all stations in the GSIM archive. Then, the MK, SNHT, and AD tests were applied to these stations. As shown in **Figure 3**, the thresholds of the MK, SNHT and AD tests were derived based on the 95% quantile value of the |Z|, T and A^2 indices, respectively. A total number of 1951 stations which experience obvious trends (or sudden changes), or at which the discharge data did not obey the given hypothetical distributions were not considered further. The selected stations were then clustered into several subgroups based on a K-means clustering model and the discordancy measures described in 3.1.2 were further applied to these stations in each subgroup. Finally, 11793 stations were selected for model development in this research.

Taking one selected station (No. AT0000032) as an example, as shown in **Figure 4** (a), the P3 distribution was selected as the optimal flood frequency curve for this station as this gives the minimal A^2 of 0.43 among all hypothetical distributions. After the flood frequency curve was defined, the parameters and uncertainties of the P3 curve were estimated based on the Bayesian MCMC method described in the section 3.2.2. As shown in **Figure 4** (b), the uncertainties (band between 0.05-0.95 confidence interval) improved with the increase of return periods and the 100-year return period could be estimated within 25% RBIAS for this station. The results of flood frequency curve selection and 100-year return period flood estimation for all stations are

shown in **Figure** 4 (c) and (d). We found that P3 and GENLOGIS are two most favoured distributions and only a small percentage of stations obey a NORM distribution.



335 **Figure 3 Thresholds for MK (a), SNHT (b) and AD (c) tests during station selection**

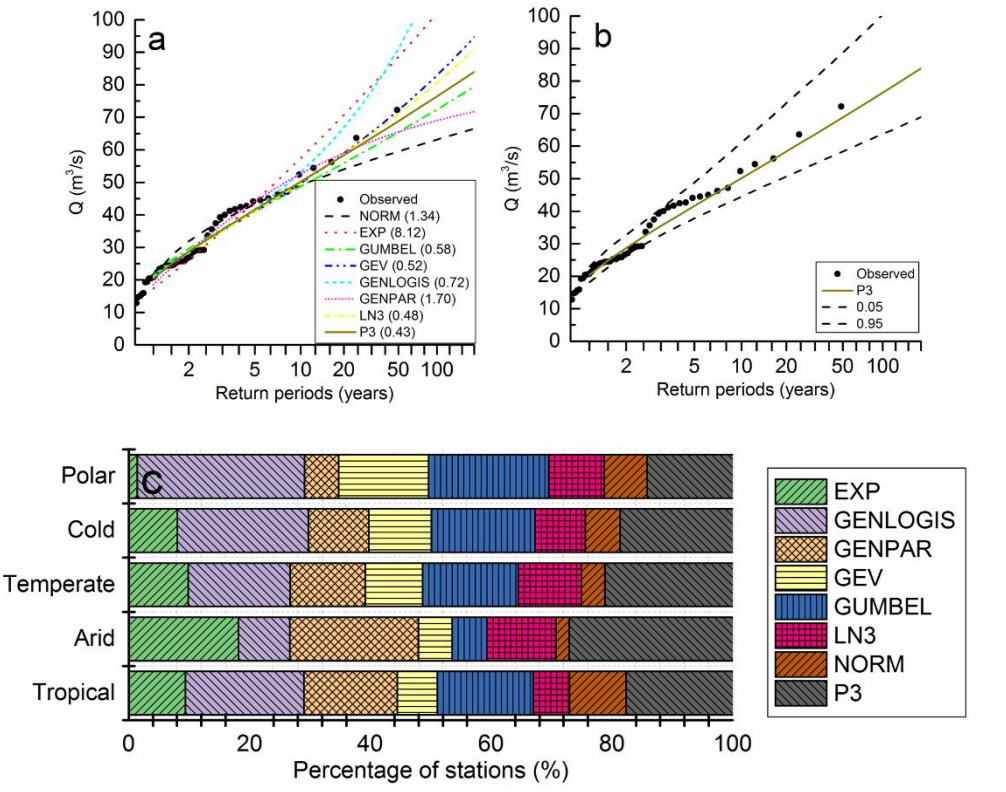


Figure 4 Flood frequency curve selection and estimation
(a: Flood frequency curve selection of the station No. AT0000032; b: Design flood estimation of the station No. AT0000032 using the Bayesian MCMC interface. c: Selection results of flood frequency curve for all stations)

4.2 Subgroups delineation

For subgroup delineation, the number of clusters (K) and the clustering factors are two sensitive parameters. The value of K is selected by considering both the heterogeneity (reflected by the DB index) and the number of stations in the subgroup. **Figure 5** describes the selection process of K when using all factors for clustering. As shown in **Figure 5** (a), the DB index reached an optimal value at K=16 and then fluctuated with the increase of K. From **Figure 5** (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 is not considered in this research.

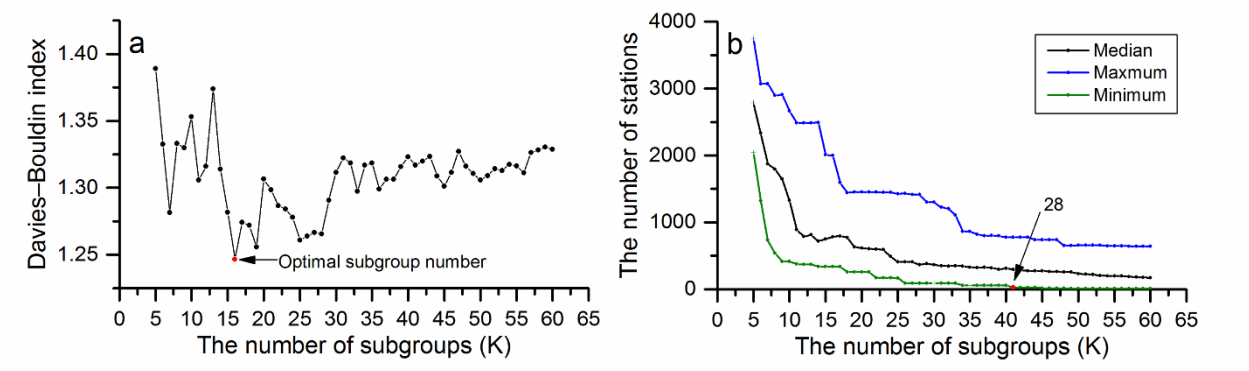


Figure 5 Optimal K selection by (a) DB index and (b) the number of stations in subgroups during delineation

Taking the 100-year flood estimation as an example, different combinations of clustering factors were tested using the optimal K. As shown in **Table 3**, the best combination is to use all factors for clustering, and this result reaches the best RMSNE of 0.708 and RBIAS of -0.174 among all combinations for the validation periods. The design flood still could be satisfactorily estimated (RBIAS<0.25) when using meteorological, physiographical, or hydrological factors for clustering, respectively. Anthropological factors alone are not recommended for clustering as this results in the worst RMSNE of 1.788 and RBIAS of -0.909 among all combinations.

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

4.3 Regression model comparison

Three regression models, PF, SVM, and RF were compared for 100-year flood estimation. **Figure 6** describes the accuracy of the three models during training and validation periods. We found that PF showed the worse accuracy with RMSNE of 1.175 and 1.1089 in training and validation periods respectively. This reveals the coefficients of PF cannot accurately describe the contribution of factors to the results. RF gave the highest fitting ability during the training period with the RMSNE of 0.290 and RBIAS of -0.037. However, the SVM model outperformed the RF model during the validation period with the RMSNE of 0.708 and RBIAS of -0.174. Therefore, SVM was selected for flood estimation for ungauged areas in this research.

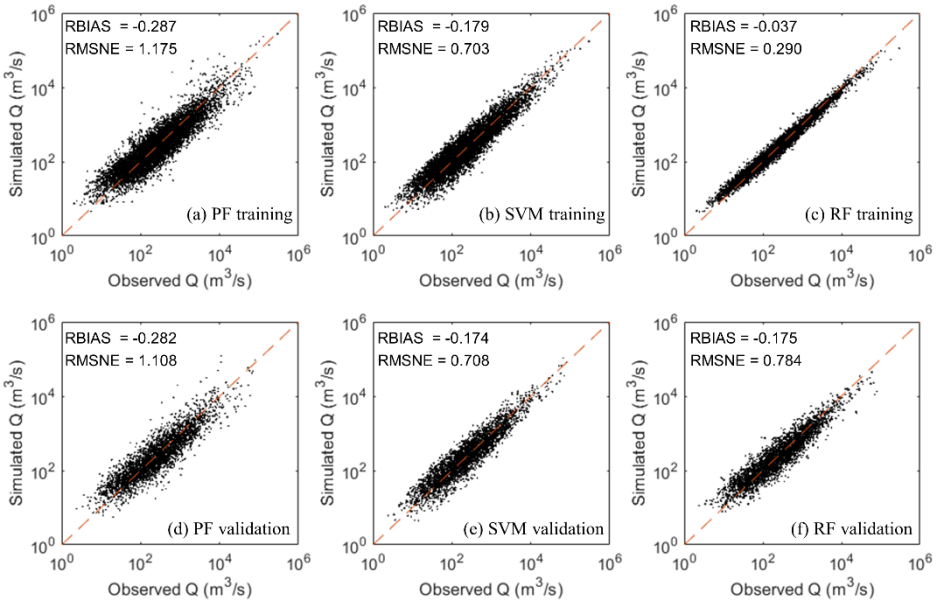
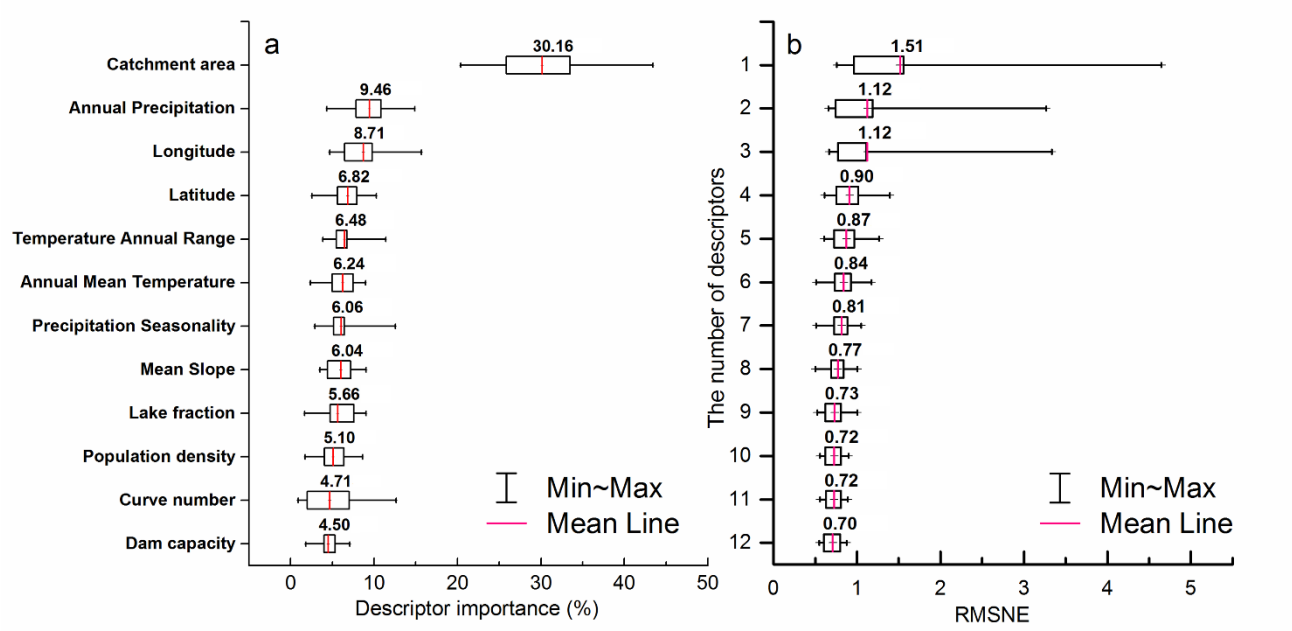


Figure 6 Comparison of three regression models (PF, SVM, and RF) during training and validation periods

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. The top four important factors are the catchment area (CA) and annual precipitation (AP), Longitude (LO) and latitude (LA). Meanwhile, dam capacity (DC) and curve number (CN) are the two least important factors and make up a relatively low proportion (average importance < 5%) of the total importance among all factors.



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

To reduce model complexity, the type of descriptors for training and validation was the same in all subgroups. The optimal number of catchment descriptors used in the regression results was further tested based on the SVM regression and the factor importance order identified by the RF model. **Figure 7** (b) described the RMSNE for all subgroups during the validation period 375 considering different combinations of catchment descriptors. AP and CA are the two most important factors and also were recommended as the catchment descriptors in the global RFFA approach of Smith et al. (2015) with a small number of samples (242 stations) for validation. When only using the CA and AP (top 2 factors) for regression and a large number of stations (3485 stations) for validation, the design flood cannot be accurately estimated for some subgroups (RMSNE>1.0). As shown in **Figure 7 (b),** the prediction accuracy can be improved by considering more importance factors for model training and 380 validation. After considering the top 10 factors for regression, the RMSNE of subgroups becomes stable and does not significantly improved by adding the further factors DC and CN. The SVM showed the highest performance considering all factors for regression (mean RMSNE of 0.70 in the validation period).

4.4 Regional flood estimation

Figure 8 (a-e) demonstrates the simulated results from the optimal SVM model in each climate zone and **Figure 8** (f) displays 385 the cumulative distribution function (CDF) between the simulated and observed discharge for all stations. Specially, the RMSNEs found were 0.746, 0.785, 0.722, 0.704 and 0.480 for tropical, arid, temperate, cold, polar climates respectively. Although there were some large individual errors, the SVM model showed good performance in tropical, temperature, cold and polar climate zones with an RBIAS of -0.199, -0.169, -0.179 and -0.091 respectively. The SVM model showed a higher

error of flood estimation in arid zones with an RBIAS of -0.233. As shown in **Figure 8** (f), the simulated result can successfully fit the observed discharge curve given its likely error.

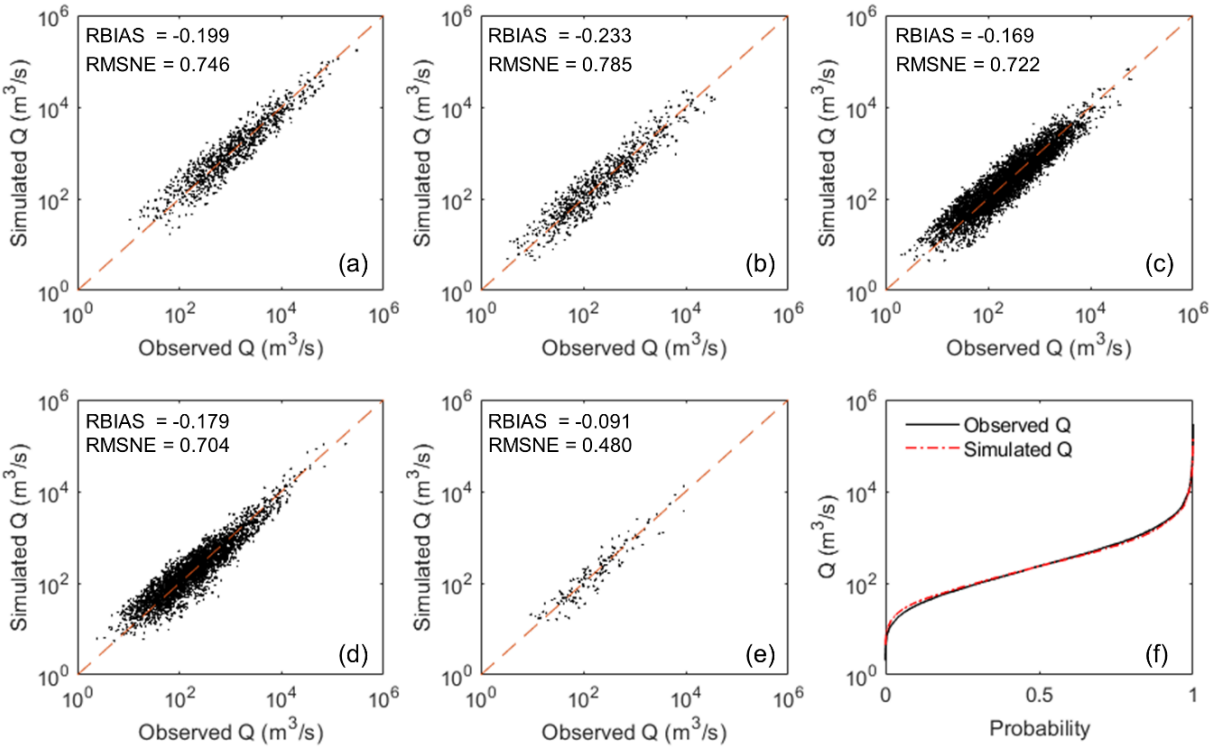


Figure 8 Results of simulated Q100 in different climate zones (a) Tropical; (b) Arid; (c) Temperate; (d) Cold; (e) Polar; (f) Cumulative distribution function of simulated and observed discharge.

Table 4 Result for different return period flood estimation

Return periods	Training		Testing	
	RBIAS	RMSNE	Mean RBIAS	RMSNE
10	-0.160	0.655	-0.165	0.664
20	-0.165	0.663	-0.168	0.672
50	-0.175	0.695	-0.166	0.684
100	-0.179	0.703	-0.174	0.708

Table 4 compares the training and validation results of 10, 20, 50, and 100-year return period floods derived by the proposed direct regression method. By using the proposed direct regression method, all design floods could be estimated within the RBIAS of 0.18 (18%). We found, unsurprisingly, that the prediction accuracy decreases with a larger flood magnitude. This mainly because the at-site design flood derived by observed data is still not truth and large return period floods are more difficult to be reliably estimated with limited observed data (Gaume, 2018). Regarding the natural and epistemic uncertainties

in flood frequency analysis, the RBIAS of at-site design flood estimation was reported as large as 30% in some studies (Di Baldassarre et al., 2012;Di Baldassarre and Montanari, 2009;Halbert et al., 2016;Merz and Thieken, 2005). As the at-site design flood is regarded as a training target, this error will be introduced in the regression and will directly affect the approach accuracy. Therefore, RBIAS of 20% or less is very impressive considering the uncertainties of at-site design flood estimation.

4.5 Discussion

In this research, we adopted a simple hold-out validation strategy such that 70% and 30% of stations were randomly selected for training and validation, respectively. To test the influence of hold-out samples on the model results, a 10-fold cross-validation strategy (Sharifi Garmdareh et al., 2018;Lee et al., 2020) was applied. All stations in one subgroup were randomly divided into 10 folds. For every fold i (i=1, 2, …, 10), the Model_i was trained by the remaining 9 folds (except the ith fold) and validated by the ith fold. Using this strategy, 10 models were developed for each subgroup, and each sample in the original dataset was used for validation once. Table 5 describes the best, worse and median model performances validated by the different 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 more stable performances than PF and SVM provided the narrowest range. We suggest using ensemble results from SVM models trained by different split samples to reduce the errors in sample selection.

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

Models	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

We compared the flood estimation accuracy in different catchment sizes (Figure 9 (a)). Both over and underestimations were found from small to large catchments. The range of R-BIAS in small catchments is wider than that in large catchments. This reveals that design floods in small catchments are more difficult to estimate than large catchments. This is mainly because the catchment descriptors derived from global satellite images can have large uncertainties (McCabe et al., 2017) when describing small catchments and can lead to more substantial errors at low discharge. The negative value of RBIAS reflected some overestimation. Some studies suggested that the trends in observed discharge can lead to overestimation (or underestimation) of flood quantiles if such nonstationarity is not taken into consideration in RFFA (Kalai et al., 2020;O'Brien and Burn, 2014). Before model development, some nonstationary or discordant stations were eliminated from the whole dataset using a threshold of 5%. This threshold is an empirical value and may remove some reliable stations if the sample is homogeneous. Figure 9 (b) presents the model performances of training, testing and eliminated stations. We found that the proposed approach showed

stable performance in the training and testing stations. However, the wide range of RBIAS in eliminated stations reveals that there may be some downsides in applying the proposed approach to a station where flows are changing rapidly. To address this limitation, the nonstationary RFFA approach should be explored in further research.

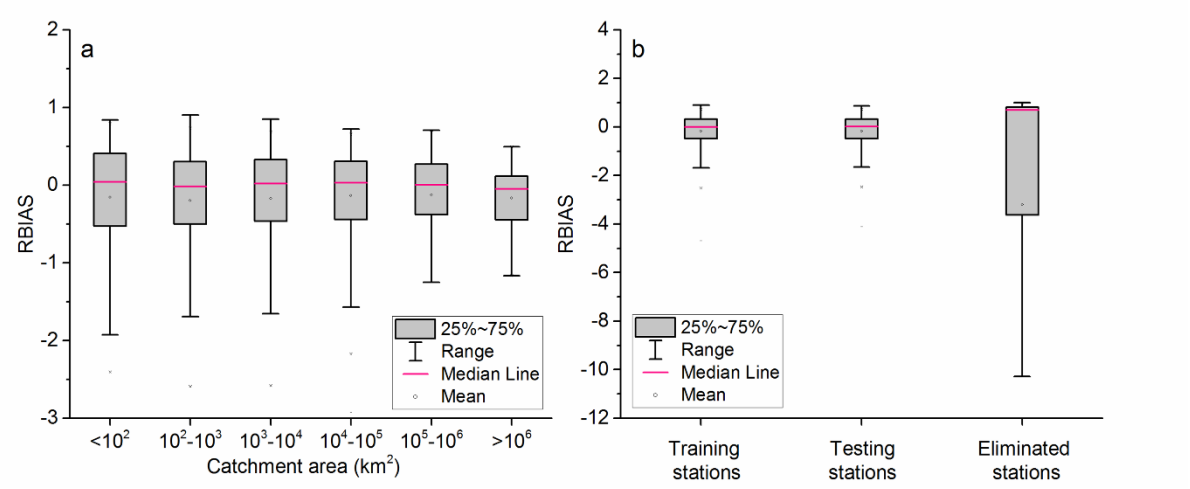


Figure 9 R-BIAS of 100-year return period flood estimation (a) in different catchment sizes; (b) in training, testing and eliminated stations.

Compared with the global RFFA of Smith (2015), this study significantly improved the station density and model accuracy in all climate zones. The proposed approach achieved similar performances to some regional studies by adopting a global coverage of stations (Shu and Ouarda, 2008;Salinas et al., 2013). However, there can still be large errors in design flood estimation, especially in arid zone. This is consistent with previous research and is likely to be a result of dryer regions being more heterogeneous (Salinas et al., 2013;Smith et al., 2015). In this study, the subgroups are delineated based on a widely used K-means model using squared Euclidean distance. Some studies showed that this model is not recommended for solving highly nonlinear problems (Bárdossy et al., 2005;Samaniego et al., 2008;Raykov et al., 2016) and this may make it difficult to properly reflect the heterogeneous characteristics of the study area. Other clustering models and distance metrics that can better describe the nonlinear relationship between descriptors should be compared in future studies.

5 Conclusions

A three-phase machine learning model-based approach was proposed to estimate design floods along river networks at the global scale. The at-site floods for global gauging station data were estimated based on Bayesian MCMC method. The global stations were then clustered into several subgroups using a K-means clustering model and SVM regression were developed in each subarea for design flood estimation. Three widely used regression techniques, Power-form function (PF), Support Vector Machine (SVM) and Random Forests (RF), were compared for regression in each subgroup. The regional floods were final estimated with the using the same descriptors and the optimal SVM regression model.

The main conclusions of this study are summarised below:

(1) To make reliable estimates, 11793 stations from the GSIM archive were selected for model implementation using stationarity, discordancy and homogeneity measures. For clustering, the impact of clustering number and clustering factors on estimation accuracy were tested. The best clustering number (K) was 16 which gave the lowest RMSNE of 0.708 for 100-year flood estimation when using all factors for clustering in a K-means model.

(2) Three regression models were compared for 100-year flood estimation and the SVM method showed the highest accuracy during the validation period with the RMSNE of 0.708 and RBIAS of -0.174. Factor importance was identified using the RF model, and different combinations of factors were tested for the optimal SVM regression implementation. When only using the top 2 factors (annual precipitation and catchment area) for regression and a large number of stations (3485 stations) for validation, the design flood cannot be accurately estimated (RMSNE>1.0) for some subgroups. After considering the top 10 factors for regression, the RMSNE for subgroups become stable and did not significantly improve by further adding factors.

(3) The proposed approach showed good performance in tropical, temperate, cold, and polar climate zones with an RBIAS of -0.199, -0.169, -0.179 and -0.091 respectively (i.e. the bias was less than 20%). A satisfying performance was found in arid areas with an RBIAS of -0.233. The negative value of RBIAS reflected some overestimation in RFFA.

This approach shows considerable promise for the estimation of extreme flood magnitudes at global scales and average RBIAS in estimation is less than 18% for all design floods. This is likely to yield a significant improvement in the skill of global flood inundation models.

6 Appendices

Appendix A Mann-Kendall test (MKT)

For a discharge series $\{Y_i | i = 1,2, \dots, n\}$, the process of MKT is described in Eq. (A1) to (A4):

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sgn(Y_j - Y_i), \tag{A1}$$

$$sgn(y) = \begin{cases} 1, & y > 0 \\ 0, & y = 0, \\ -1, & y < 0 \end{cases} \tag{A2}$$

The variance of S is computed as follows:

$$Var(s) = \frac{1}{18}(n(n-1)(2n+5) - \sum_{i=1}^{n} t_i i(i-1)(2i+5)), \tag{A3}$$

Where n is the total number of time series Y_i; $Var(s)$ is the variance of S; t_i is the number of data points contained in the i-th tie group;

The MKT statistic Z is given by:

$$Z = \begin{cases} \frac{S-1}{\sqrt{Var(s)}}, & S > 0 \\ 0, & S = 0 \\ \frac{S+1}{\sqrt{Var(s)}}, & S < 0 \end{cases}, \quad \text{(A4)}$$

Appendix B Standard normal homogeneity test (SNHT)

The SNHT is described in Eq. (B1) to (B4).

$$T_d = d\bar{z}_1^2 + (n-d)\bar{z}_2^2, d = 1,2,\dots n, , \quad \text{(B1)}$$

$$\bar{Z}_1 = \frac{1}{d}\sum_{i=1}^{d}(Y_i - \bar{Y})/s, \quad \text{(B2)}$$

$$\bar{Z}_2 = \frac{1}{n-d}\sum_{i=d+1}^{n}(Y_i - \bar{Y})/s, \quad \text{(B3)}$$

The SNHT statistic T is given by:

$$T = \max(T_d), \quad \text{(B4)}$$

Appendix C Discordancy measure methods

The Z-score (ZS) method is described in Eq. (C1).

$$zs_i = \frac{R_i - \overline{\mathrm{R}}}{sd(R)}, \quad \text{(C1)}$$

Where R_i is the mean annual runoff of station *i* in one subgroup, $\overline{\mathrm{R}}$ is the mean value of R_i in one subgroup; and $sd(R)$ is the standard deviation of R_i in one subgroup.

The L-moments-based method is presented in Eq. (C2) to (C4).

$$D_i = \frac{1}{3}N_k(U_i - \bar{U})^T S^{-1}(U_i - \bar{U}), \quad \text{(C2)}$$

$$\bar{U} = \frac{1}{N_k}\sum_{1}^{N_k} U_i, \quad \text{(C3)}$$

$$S = \sum_{1}^{N_k}(U_i - \bar{U})(U_i - \bar{U})^T, \quad \text{(C4)}$$

Where N_k is the number of stations in cluster k, and $U_i = [t^i, t_3^i, t_4^i]$ is the vector of L-CV, L-skewness and L-kurtosis of the station *i* as defined by Hosking and Wallis (2005).

Appendix D Hypothetical distributions

Table D1 Hypothetical distributions adopted in this research

No.	Hypothetical distributions	Probability density function f(x) or cumulative distribution function F(X)
1	Normal distribution (NORM)	$\mathrm{f}(x) = \frac{1}{\sqrt{2\pi b^2}} \exp\left(-\frac{(x-a)^2}{2b^2}\right)$
2	Exponential distribution (EXP)	$\mathrm{F}(x) = 1 - \exp\left(-\frac{x-a}{b}\right)$
3	Gumbel distribution (GUMBEL)	$\mathrm{F}(x) = \exp[-\exp(-\frac{x-a}{b})]$
4	Generalized extreme value distribution (GEV)	$\mathrm{F}(x) = \exp\left[-(1-\frac{b}{a}(x-c))^{\frac{1}{b}}\right]$
5	Generalized logistic distribution (GENLOGIS)	$\mathrm{F}(x) = \frac{1}{(1+\exp(\frac{x-a}{b}))^c}$
6	Generalized Pareto distribution (GENPAR)	$\mathrm{F}(x) = 1-(1+c(\frac{(x-a)}{b}))^{-\frac{1}{c}}$
7	Lognormal distribution (LN3)	$\mathrm{f}(x) = \frac{1}{\sqrt{2\pi}b(x-c)} \exp\left(-\frac{1}{2}(\frac{\log(x-c)-b}{a})^2\right)$
8	Pearson type III distribution (P3)	$\mathrm{f}(x) = \frac{1}{a\Gamma(b)}\left(\frac{x-c}{a}\right)^{b-1} \exp\left(-\left(\frac{x-c}{a}\right)\right)$

Note: x is the random variable; a, b, c are parameters of distributions

Appendix E Anderson-Darling (AD) test

Given a sample of data x_i (i = 1,2, …, m), the AD test will determine if a given data set comes from a candidate cumulative distribution function $F(x, \theta)$ as Eq. (13) and (14), where θ is the parameter estimated by the sample x_i.

$$\begin{cases} F_m(x) = 0, x < x_{(1)} \\ F_m(x) = \frac{i}{m}, x_{(i)} \le x < x_{(i+1)}, \\ F_m(x) = 1, x_{(m)} \le x \end{cases} \quad \text{(D1)}$$

Where: $F_m(x)$ is the empirical cumulative distribution function; $x_{(i)}$ is the i-th element of sample in increasing order; the AD test statistic A^2 can be calculated as in Eq. (16) (Ahmad et al., 1988;Laio, 2004).

$$A^2 = -m - \frac{1}{m}\sum_{i=1}^{m}\{(2i-1)\ln[F(x_{(i)},\theta)] + (2m+1-2i)\ln[1-F(x_{(i)},\theta)]\}, \quad \text{(D2)}$$

7 Code availability

The MKT and SNHT were implemented using the *trend* package (https://CRAN.R-project.org/package=trend). The discordancy and at-site design floods were estimated based on the nsRFA package (https://CRAN.R-project.org/package=nsRFA). The heterogeneity measure, K-means, SVM and RF models are available from https://www.mathworks.com/help/stats/kmeans.html, https://uk.mathworks.com/help/stats/support-vector-machine-regression.html, and https://www.stat.berkeley.edu/~breiman/RandomForests/ respectively.

8 Data availability

The Global Streamflow Indices and Metadata Archive (GSIM) is freely available at https://doi.pangaea.de/10.1594/PANGAEA.887477. To access WorldClim - Global Climate Data, visit http://www.worldclim.org/. The adopted terrain data is the MERIT-DEM at http://hydro.iis.u-tokyo.ac.jp/~yamadai/MERIT_DEM/. The Gridded Population of the World (GPW) dataset is available from https://sedac.ciesin.columbia.edu/data/collection/gpw-v4. The Global Lakes and Wetlands Database (GLWD) is available from https://www.worldwildlife.org/pages/global-lakes-and-wetlands-database. The data sources of NRCS Curve Number dataset are described at https://doi.org/10.1080/2150704X.2017.1297544. The Global Reservoir and Dam Database (GRanD) can be downloaded by visiting http://globaldamwatch.org/grand/.

9 Author contribution

Conceptualization: Gang Zhao, Paul Bates and Jeff Neal; Methodology: Gang Zhao; Formal analysis: Gang Zhao; Writing-original draft preparation: Gang Zhao; writing—review and editing: Paul Bates, Jeff Neal and Bo Pang; Visualization, Gang Zhao; Supervision, Paul Bates and Jeff Neal;

10 Competing interests

The authors declare that they have no conflict of interest.

11 Acknowledgments

Gang Zhao would like to thank Andrew Smith at Fathom for helpful discussions. Paul Bates is supported by a Royal Society Wolfson Research Merit award. Jeff Neal is supported by NERC grants (NE/S003061/1 and NE/S006079/1). Gang Zhao is supported by the China Scholarship Council-University of Bristol Joint PhD Scholarships Programme (No.201700260088). Bo Pang is supported by National Natural Science Foundation of China (No. 51879008).

References

Ahmad, M. I., Sinclair, C., and Spurr, B.: Assessment of flood frequency models using empirical distribution function statistics, Water Resour Res, 24, 1323-1328, 1988.

Alexandersson, H.: A Homogeneity Test Applied to Precipitation Data, J Climatol, 6, 661-675, DOI 10.1002/joc.3370060607, 1986.

Bárdossy, A., Pegram, G. G., and Samaniego, L.: Modeling data relationships with a local variance reducing technique: Applications in hydrology, Water Resour Res, 41, 2005.

Bates, P. D., Quinn, N., Sampson, C., Smith, A., Wing, O., Sosa, J., Savage, J., Olcese, G., Neal, J., and Schumann, G.: Combined modelling of US fluvial, pluvial and coastal flood hazard under current and future climates, Water Resour Res, e2020WR028673, 2020.

Bocchiola, D., De Michele, C., Rosso, R. J. H., and Discussions, E. S. S.: Review of recent advances in index flood estimation, 7, 283-296, 2003.

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

CRED, and UNISDR: The Human Cost of Weather Related Disasters - 1995 - 2015, United Nations Office for Disaster Risk Reduction (UNISDR) and Centre for Research on the Epidemiology of Disasters (CRED), 2015.

Cunnane, C.: Methods and Merits of Regional Flood Frequency-Analysis, J Hydrol, 100, 269-290, Doi 10.1016/0022-1694(88)90188-6, 1988.

Dalrymple, T.: Flood-frequency analyses, manual of hydrology: Part 3, USGPO, 1960.

Davies, D. L., and Bouldin, D. W.: A cluster separation measure, IEEE transactions on pattern analysis and machine intelligence, 224-227, 1979.

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

Di Baldassarre, G., and Montanari, A.: Uncertainty in river discharge observations: a quantitative analysis, Hydrology & Earth System Sciences, 13, 2009.

Di Baldassarre, G., Laio, F., and Montanari, A.: Effect of observation errors on the uncertainty of design floods, Physics and Chemistry of the Earth, Parts A/B/C, 42, 85-90, 2012.

Do, H., Zhao, F., Westra, S., Leonard, M., and Gudmundsson, L.: A global-scale comparison of modeled and observed trends in magnitude and frequency of flooding, Geophysical Research Abstracts, 2019,

Do, H. X., Gudmundsson, L., Leonard, M., and Westra, S.: The Global Streamflow Indices and Metadata Archive (GSIM) - Part 1: The production of a daily streamflow archive and metadata, Earth Syst Sci Data, 10, 10.5194/essd-10-765-2018, 2018.

Doxsey-Whitfield, E., MacManus, K., Adamo, S. B., Pistolesi, L., Squires, J., Borkovska, O., and Baptista, S. R. J. P. i. A. G.: Taking advantage of the improved availability of census data: a first look at the gridded population of the world, version 4, 1, 226-234, 2015.

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.

Fick, S. E., and Hijmans, R. J.: WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas, Int J Climatol, 37, 4302-4315, 10.1002/joc.5086, 2017.

Frieler, K., Lange, S., Piontek, F., Reyer, C. P., Schewe, J., Warszawski, L., Zhao, F., Chini, L., Denvil, S., and Emanuel, K. J. G. M. D.: Assessing the impacts of 1.5 C global warming–simulation protocol of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2b), 2017.

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.

Garcia, F. A. A.: Tests to identify outliers in data series, Pontifical Catholic University of Rio de Janeiro, Industrial Engineering Department, Rio de Janeiro, Brazil, 2012.

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.

Gaume, E.: Flood frequency analysis: The Bayesian choice, Wiley Interdisciplinary Reviews: Water, 5, e1290, 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, 2016.

Griffis, V. W., and Stedinger, J. R.: Log-Pearson Type 3 distribution and its application in flood frequency analysis. I: Distribution characteristics, J Hydrol Eng, 12, 482-491, 10.1061/(Asce)1084-0699(2007)12:5(482), 2007.

Gudmundsson, L., Do, H. X., Leonard, M., and Westra, S.: The Global Streamflow Indices and Metadata Archive (GSIM) - Part 2: Quality control, time-series indices and homogeneity assessment, Earth Syst Sci Data, 10, 10.5194/essd-10-787-2018, 2018.

Haddad, K., and Rahman, A.: Selection of the best fit flood frequency distribution and parameter estimation procedure: a case study for Tasmania in Australia, Stochastic Environmental Research and Risk Assessment, 25, 415-428, 2011.

Halbert, K., Nguyen, C. C., Payrastre, O., and Gaume, E.: Reducing uncertainty in flood frequency analyses: A comparison of local and regional approaches involving information on extreme historical floods, J Hydrol, 541, 90-98, 2016.

Hamed, K. H. J. J. o. h.: Trend detection in hydrologic data: the Mann–Kendall trend test under the scaling hypothesis, 349, 350-363, 2008.

Hammond, M. J., Chen, A. S., Djordjevic, S., Butler, D., and Mark, O.: Urban flood impact assessment: A state-of-the-art review, Urban Water J, 12, 14-29, 10.1080/1573062x.2013.857421, 2015.

Hosking, J. R. M., and Wallis, J. R.: The Effect of Intersite Dependence on Regional Flood Frequency-Analysis, Water Resour Res, 24, 588-600, DOI 10.1029/WR024i004p00588, 1988.

Hosking, J. R. M., and Wallis, J. R.: Regional frequency analysis: an approach based on L-moments, Cambridge University Press, 2005.

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.

Laio, F.: Cramer–von Mises and Anderson-Darling goodness of fit tests for extreme value distributions with unknown parameters, Water Resour Res, 40, 2004.

Lee, J.-Y., Choi, C., Kang, D., Kim, B. S., and Kim, T.-W.: Estimating Design Floods at Ungauged Watersheds in South Korea Using Machine Learning Models, Water, 12, 3022, 2020.

Lehner, B., and Doll, P.: Development and validation of a global database of lakes, reservoirs and wetlands, J Hydrol, 296, 1-22, 10.1016/j.jhydrol.2004.03.028, 2004.

Lehner, B., Liermann, C. R., Revenga, C., Vorosmarty, C., Fekete, B., Crouzet, P., Doll, P., Endejan, M., Frenken, K., Magome, J., Nilsson, C., Robertson, J. C., Rodel, R., Sindorf, N., and Wisser, D.: High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management, Front Ecol Environ, 9, 494-502, 10.1890/100125, 2011.

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.

Lin, G. F., and Chen, L. H.: Identification of homogeneous regions for regional frequency analysis using the self-organizing map, J Hydrol, 324, 1-9, 10.1016/j.jhydrol.2005.09.009, 2006.

Liu, X., Liu, W., Yang, H., Tang, Q., Flörke, M., Masaki, Y., Müller Schmied, H., Ostberg, S., Pokhrel, Y., and Satoh, Y.: Multimodel assessments of human and climate impacts on mean annual streamflow in China, Hydrol Earth Syst Sc, 23, 1245-1261, 2019.

McCabe, M. F., Rodell, M., Alsdorf, D. E., Miralles, D. G., Uijlenhoet, R., Wagner, W., Lucieer, A., Houborg, R., Verhoest, N. E., and Franz, T. E.: The future of Earth observation in hydrology, Hydrol Earth Syst Sc, 21, 3879, 2017.

Merz, B., and Thieken, A. H.: Separating natural and epistemic uncertainty in flood frequency analysis, J Hydrol, 309, 114-132, 2005.

Merz, R., and Bloschl, G.: Flood frequency hydrology: 1. Temporal, spatial, and causal expansion of information, Water Resour Res, 44, Artn W08432
10.1029/2007wr006744, 2008a.

Merz, R., and Bloschl, G.: Flood frequency hydrology: 2. Combining data evidence, Water Resour Res, 44, Artn W08433
10.1029/2007wr006745, 2008b.

Mockus, V.: National engineering handbook, Section, 1964.

Mueller Schmied, H., Adam, L., Eisner, S., Fink, G., Flörke, M., Kim, H., Oki, T., Portmann, F. T., Reinecke, R., Riedel, C. J. H., and Sciences, E. S.: Variations of global and continental water balance components as impacted by climate forcing uncertainty and human water use, 20, 2877-2898, 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.
Prihodko, L., Denning, A. S., Hanan, N. P., Baker, I., and Davis, K.: Sensitivity, uncertainty and time dependence of parameters in a complex land surface model, Agr Forest Meteorol, 148, 268-287, 10.1016/j.agrformet.2007.08.006, 2008.
Raykov, Y. P., Boukouvalas, A., Baig, F., and Little, M. A.: What to do when K-means clustering fails: a simple yet principled alternative algorithm, PloS one, 11, e0162259, 2016.
Reis, D. S., and Stedinger, J. R.: Bayesian MCMC flood frequency analysis with historical information, J Hydrol, 313, 97-116, 2005.
Richter, B. D., Baumgartner, J. V., Wigington, R., and Braun, D. P.: How much water does a river need?, Freshwater Biol, 37, 231-249, DOI 10.1046/j.1365-2427.1997.00153.x, 1997.
Salinas, J. L., Laaha, G., Rogger, M., Parajka, J., Viglione, A., Sivapalan, M., and Blöschl, G.: Comparative assessment of predictions in ungauged basins – Part 2: Flood and low flow studies, Hydrol. Earth Syst. Sci., 17, 2637-2652, 10.5194/hess-17-2637-2013, 2013.
Samaniego, L., Bárdossy, A., and Schulz, K.: Supervised classification of remotely sensed imagery using a modified $ k $-NN technique, IEEE Transactions on Geoscience and Remote Sensing, 46, 2112-2125, 2008.
Sampson, C. C., Smith, A. M., Bates, P. B., Neal, J. C., Alfieri, L., and Freer, J. E.: A high-resolution global flood hazard model, Water Resour Res, 51, 7358-7381, 10.1002/2015wr016954, 2015.
Schumann, G., Bates, P. D., Apel, H., and Aronica, G. T.: Global flood hazard mapping, modeling, and forecasting: challenges and perspectives, Global Flood Hazard: Applications in Modeling, Mapping, and Forecasting, 239-244, 2018.
Schumann, G. J.-P., Andreadis, K. M., and Bates, P. D. J. J. o. h.: Downscaling coarse grid hydrodynamic model simulations over large domains, 508, 289-298, 2014a.
Schumann, G. J. P., Andreadis, K. M., and Bates, P. D.: Downscaling coarse grid hydrodynamic model simulations over large domains, J Hydrol, 508, 289-298, 10.1016/j.jhydrol.2013.08.051, 2014b.
Sharifi Garmdareh, E., Vafakhah, M., and Eslamian, S. S.: Regional flood frequency analysis using support vector regression in arid and semi-arid regions of Iran, Hydrological sciences journal, 63, 426-440, 2018.
Sharma, A., Wasko, C., and Lettenmaier, D. P.: If precipitation extremes are increasing, why aren't floods?, Water Resour Res, 54, 8545-8551, 2018.
Shiffler, R. E. J. T. A. S.: Maximum Z scores and outliers, 42, 79-80, 1988.
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.
Smith, A., Sampson, C., and Bates, P.: Regional flood frequency analysis at the global scale, Water Resour Res, 51, 539-553, 10.1002/2014wr015814, 2015.
Stedinger, J. R.: Estimating a regional flood frequency distribution, Water Resour Res, 19, 503-510, 1983.
Stein, L., Pianosi, F., and Woods, R.: Event-based classification for global study of river flood generating processes, Hydrol Process, 2019.
Teng, J., Jakeman, A. J., Vaze, J., Croke, B. F. W., Dutta, D., and Kim, S.: Flood inundation modelling: A review of methods, recent advances and uncertainty analysis, Environ Modell Softw, 90, 201-216, 10.1016/j.envsoft.2017.01.006, 2017.
Trigg, M. A., Birch, C. E., Neal, J. C., Bates, P. D., Smith, A., Sampson, C. C., Yamazaki, D., Hirabayashi, Y., Pappenberger, F., Dutra, E., Ward, P. J., Winsemius, H. C., Salamon, P., Dottori, F., Rudari, R., Kappes, M. S., Simpson, A. L., Hadzilacos, G., and Fewtrell, T. J.: The credibility challenge for global fluvial flood risk analysis, Environ Res Lett, 11, Artn 094014
10.1088/1748-9326/11/9/094014, 2016.
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.
Wing, O. E., Bates, P. D., Sampson, C. C., Smith, A. M., Johnson, K. A., and Erickson, T. A. J. W. R. R.: Validation of a 30 m resolution flood hazard model of the conterminous U nited S tates, 53, 7968-7986, 2017.
Wing, O. E., Bates, P. D., Smith, A. M., Sampson, C. C., Johnson, K. A., Fargione, J., and Morefield, P.: Estimates of present and future flood risk in the conterminous United States, Environ Res Lett, 13, 034023, 2018.

Winsemius, H., Van Beek, L., Jongman, B., Ward, P., and Bouwman, A.: A framework for global river flood risk assessments, 2013.

Winsemius, H. C., Aerts, J. C., Van Beek, L. P., Bierkens, M. F., Bouwman, A., Jongman, B., Kwadijk, J. C., Ligtvoet, W., Lucas, P. L., and Van Vuuren, D. P.: Global drivers of future river flood risk, Nat Clim Change, 6, 381-385, 2016.

Yamazaki, D., Kanae, S., Kim, H., and Oki, T.: A physically based description of floodplain inundation dynamics in a global river routing model, Water Resour Res, 47, Artn W04501 10.1029/2010wr009726, 2011.

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., Sampson, C. C., Kanae, S., and Bates, P. D.: A high-accuracy map of global terrain elevations, Geophys Res Lett, 44, 5844-5853, 10.1002/2017gl072874, 2017.

Yang, T., Sun, F., Gentine, P., Liu, W., Wang, H., Yin, J., Du, M., and Changming, L.: Evaluation and machine learning improvement of global flood simulations, AGUFM, 2019, H33L-2122, 2019a.

Yang, T., Sun, F., Gentine, P., Liu, W., Wang, H., Yin, J., Du, M., and Liu, C.: Evaluation and machine learning improvement of global hydrological model-based flood simulations, Environ Res Lett, 14, 114027, 2019b.

Zeng, Z. Y., Tang, G. Q., Hong, Y., Zeng, C., and Yang, Y.: Development of an NRCS curve number global dataset using the latest geospatial remote sensing data for worldwide hydrologic applications, Remote Sens Lett, 8, 528-536, 10.1080/2150704x.2017.1297544, 2017.

Zhang, Y., Chiew, F. H., Li, M., and Post, D.: Predicting runoff signatures using regression and hydrological modeling approaches, Water Resour Res, 54, 7859-7878, 2018.

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.