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

Gang Zhao1, Paul Bates1,2, Jeffrey Neal1,2, Bo Pang3

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 mainly involves three stages: (i) estimating at-site flood frequency curve for global gauging stations by the Anderson-Darling test and 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 regression was improved by considering more descriptors during the training and validation; (2) a support vector machine regression provide the highest prediction performance among all regression models tested, with root mean square normalised error of 0.708 for 100-year return period flood estimation; (3) 100-year design flood in tropical, arid, temperate, cold and polar climate zones could be reliably estimated with the relative mean relative biases (RBIAS) of -0.199, -0.233, -0.169, 0.179 and -0.091 respectively; (4) This machine learning based approach shows considerable improvement over the index-flood based method introduced by Smith et al. (2015, https://doi.org/10.1002/2014WR015814) for the design flood estimation at global scales; and 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, affecting 2.3 billion people and causing 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 (Kundzewicz et al., 2014; Wing et al., 2018a; Winsemius et al., 2016a; Wing et al., 2017; Winsemius et al., 2016b; Wing et al., 2018b). 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 (Sampson et al., 2015; Schumann et al., 2013). 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 land surface models (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 land surface model implementation (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, 2008a, b; Dalrymple, 1960). Amongst these, the index-flood method and the direct regression method are two of the most widely used procedures (Shu and Ouarda, 2006). 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 the best of our knowledge, only 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 GEV distribution for the global coverage. Studies have suggested alternate distributions, such as generalized pareto distribution, perform better in populated regions of Australia and a log pearson type 3 distribution was recommended in the southwestern United States (Vogel et al., 1993a; Vogel et al., 1993b). 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 regional floods 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 method 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 series 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, 5 types of time-series indices are provided at yearly, seasonal, and monthly resolutions in the GSIM archive. 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 ranges of streamflow parameters are reported to narrow substantially 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 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>
<thead>
<tr>
<th>Climate Class</th>
<th>Smith et al. (2015)</th>
<th>This research</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
<td>Validation</td>
</tr>
<tr>
<td>Tropical</td>
<td>163</td>
<td>8</td>
</tr>
<tr>
<td>Arid</td>
<td>121</td>
<td>55</td>
</tr>
<tr>
<td>Temperate</td>
<td>296</td>
<td>99</td>
</tr>
<tr>
<td>Cold</td>
<td>109</td>
<td>80</td>
</tr>
<tr>
<td>Polar</td>
<td>14</td>
<td>N/A</td>
</tr>
<tr>
<td>Total</td>
<td>703</td>
<td>242</td>
</tr>
</tbody>
</table>

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 km². Twelve explanatory factors collected from open-access databases were selected as potential descriptors and applied for clustering and regression model development. The statistics and data sources of these factors are summarised in Table 2.

<table>
<thead>
<tr>
<th>No.</th>
<th>Factor name (Abbreviation)</th>
<th>Unit</th>
<th>Min</th>
<th>Max</th>
<th>Data source</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Annual Precipitation (AP)</td>
<td>mm</td>
<td>0</td>
<td>7743</td>
<td>WorldClim</td>
</tr>
<tr>
<td>2</td>
<td>Precipitation Seasonality (PS)</td>
<td>-</td>
<td>0</td>
<td>219.35</td>
<td>WorldClim</td>
</tr>
</tbody>
</table>
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).
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

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.

![Figure 2](https://doi.org/10.5194/hess-2020-594)

Preprint. Discussion started: 23 December 2020
© Author(s) 2020. CC BY 4.0 License.
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. For a discharge series \( \{Y_i | i = 1, 2, \ldots, n\} \), the process of MKT is described in Eq. (1) to (4):

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

\[
sgn(y) = \begin{cases} 
1, & y > 0 \\
0, & y = 0 \\
-1, & y < 0 
\end{cases} \tag{2}
\]

The variance of \( S \) is computed as follows:

\[
Var(s) = \frac{1}{18} (n(n - 1)(2n + 5) - \sum_{i=1}^{n} t_i (i - 1)(2i + 5)), \tag{3}
\]

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:

\[
Z = \begin{cases} 
\frac{S - 1}{\sqrt{Var(s)}}, & S > 0 \\
0, & S = 0 \\
\frac{S + 1}{\sqrt{Var(s)}}, & S < 0 
\end{cases} \tag{4}
\]

The SNHT is described in Eq. (5) to (8).

\[
T_d = d \bar{z}_d^2 + (n - d) \bar{z}_2^2, \quad d = 1, 2, \ldots, n, \tag{5}
\]

\[
\bar{Z}_1 = \frac{1}{d} \sum_{i=1}^{d} (Y_i - \bar{Y}) / s, \tag{6}
\]

\[
\bar{Z}_2 = \frac{1}{n-d} \sum_{i=d+1}^{n} (Y_i - \bar{Y}) / s, \tag{7}
\]

The SNHT statistic \( T \) is given by:

\[
T = \max(T_d). \tag{8}
\]

One advantage of these two tests is that they provide an index reflecting the degree of trend or change. 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 \( T \) 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. (9) and z values greater than 3 are regarded as outliers (Garcia, 2012).

\[
z_i = \frac{R_i - \bar{R}}{s_d(R)},
\]  

(9)

Where \(R_i\) is the mean annual runoff of station \(i\) in one subgroup, \(\bar{R}\) is the mean value of \(R_i\) and \(s_d(R)\) is the standard deviation of \(R_i\).

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. (10) to (12). As suggested by Hosking and Wallis, a station is considered as discordant if \(D_i \geq 3\) when \(N_k\) is larger than 15.

\[
D_i = \frac{1}{2}N_k(U_i - \bar{U})^TS^{-1}(U_i - \bar{U}),
\]  

(10)

\[
\bar{U} = \frac{1}{N_k}\sum_{i=1}^{N_k} U_i,
\]  

(11)

\[
S = \sum_{i=1}^{N_k}(U_i - \bar{U})(U_i - \bar{U})^T,
\]  

(12)

Where \(N_k\) is the number of stations in cluster \(k\), and \(U_i = [t_i^c, t_i^s, t_i^k]\) is the vector of L-CV, L-skewness and L-kurtosis of the station \(i\) as defined by Hosking and Wallis (2005).

3.2 At-site flood estimation

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. Table 3 describes the adopted hypothetical distributions. 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.

<table>
<thead>
<tr>
<th>No.</th>
<th>Hypothetical distributions</th>
<th>Probability density function (f(x)) or cumulative distribution function (F(X))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Normal distribution (NORM)</td>
<td>[f(x) = \frac{1}{\sqrt{2\pi}b^2} \exp\left(-\frac{(x-a)^2}{2b^2}\right)]</td>
</tr>
</tbody>
</table>
Exponential distribution
(EXP)
\[ F(x) = 1 - \exp\left(-\frac{x - a}{b}\right) \]

Gumbel distribution
(GUMBEL)
\[ F(x) = \exp\left[-\exp\left(-\frac{x - a}{b}\right)\right] \]

Generalized extreme value distribution
(GEV)
\[ F(x) = \exp \left[-(1 - b \frac{x - a}{c})^\frac{1}{c}\right] \]

Generalized logistic distribution
(GENLOGIS)
\[ F(x) = \frac{1}{1 + \exp\left(-\frac{x - a}{b}\right)^c} \]

Generalized Pareto distribution
(GENPAR)
\[ F(x) = 1 - \left(1 + c \frac{x - a}{b}\right)^{-\frac{1}{c}} \]

Lognormal distribution
(LN3)
\[ f(x) = \frac{1}{\sqrt{2\pi}b(x-c)} \exp\left(-\frac{1}{2} \left(\frac{x-c-b}{a}\right)^2\right) \]

Pearson type III distribution
(P3)
\[ f(x) = \frac{1}{aF(b)} \left(\frac{x-c}{a}\right)^{b-1} \exp\left(-\frac{x-c}{a}\right) \]

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

### 3.2.1 Distribution selection

The Anderson–Darling (AD) test can detect whether a given sample of data follows a hypothetical distribution. Given a sample of data \( x_i (i = 1, 2, \ldots, m) \), the AD test compares the cumulative frequency function of \( F_m(x) \) and hypothetical distribution \( F(x, \theta) \) as Eq. (13) to (15), where \( \theta \) is the parameter estimated by the sample \( x_i \).

\[
\begin{aligned}
F_m(x) &= 0, \quad x < x_{(1)} \\
F_m(x) &= \frac{i}{m}, \quad x_{(i)} \leq x < x_{(i+1)}, \\
F_m(x) &= 1, \quad x_{(m)} \leq x 
\end{aligned}
\]  

\[ Q^2 = m \int_x [F_m(x) - F(x, \theta)]^2 \Psi(x) \, dF \]  

\[ \psi(x) = [F(x, \theta) \left(1 - F(x, \theta)\right)]^{-1} \]  

Where: \( x_{(i)} \) 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).

\[ A^2 = -m - \frac{1}{m} \sum_{l=1}^{m} \left\{ (2l - 1) \ln[F(x_{(l)}, \theta)] + (2m + 1 - 2l) \ln[1 - F(x_{(l)}, \theta)] \right\} \]  

The AD statistic is 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. After selection of flood
frequency curve for all stations, a threshold $A_{295\%}$ 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 three common methods for estimating the parameters of at-site flood frequency curve are based on moments (MOM), maximum likelihood (MLE) and Bayesian inference. 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 (Nazmi et al., 2020; Haddad and Rahman, 2011; Gaume, 2018). The Bayesian MCMC approach derives an empirical approximation of the posterior distribution of parameters and flood quantiles for the selected distribution. The posterior distribution is computed using Bayesian inference as in Eq. (17).

$$P(\theta|D) = \frac{l(D|\xi(\theta))}{\int l(D|\xi(\theta)) d\theta},$$  \hspace{1cm} (17)

Where $D$ is the given sample; $\theta$ is a parameter vector; $P(\theta|D)$ is the posterior distribution of the parameters $\theta$; $\xi(\theta)$ is the prior distribution of parameters $\theta$; $l(D|\theta)$ is the likelihood function.

As the integral in Eq. (17) is insolvable analytically in practise, the Metropolis-Hastings MCMC method was used to generate samples from the posterior distribution. The detail of the Bayesian MCMC method is comprehensively described in the research of Reis and Stedinger (2005).

### 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 $D = \{x_j | j = 1, 2, 3, \ldots, N\}$ where $N$ is the number of samples. The $K$ centroids $C = \{C_k | k = 1, 2, 3, \ldots, K\}$ 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. (18).

$$\text{arg min}_{k \in C} [\text{dis}(C_k, x_j)]^2,$$  \hspace{1cm} (18)

Where: $\text{dis}(-)$ is the function for Euclidean distance calculation.

(b) The centroid $C_k$ in step (a) is updated as in Eq. (19). 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.

$$C_k = \frac{1}{N_k} \sum_{j=1}^{N_k} x_j, \quad x_k \in A_k.$$  \hspace{1cm} (19)
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 cluster diversity and inter cluster distances. The intra cluster diversity of a cluster $j$ is computed as Eq. in (20):

$$S_j = \frac{1}{N_j} \sum_{p_i=j} \text{dis}(C_j, x_i)^2,$$

(20)

The inter cluster distance between cluster $j$ and cluster $k$ is measured as Eq. (21):

$$R_{j,k} = \frac{s_j + s_k}{|C_j - C_k|}.$$  

(21)

The Davies-Bouldin index is computed as Eq. (22) and (23)

$$R_j = \max_{l \neq j} R_{j,l}.$$  

(22)

$$\text{DBI} = \frac{1}{K} \sum_{j=1}^{K} R_j.$$  

(23)

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 Direct regression method

The direct regression method estimates design floods directly with catchment descriptors using a regression model as described in Eq. (24). 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 time the direct regression method has been tested in an RFFA at the global scale.

$$DF = f_{FQ}(d_1, d_2, ..., d_N).$$  

(24)

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. (25).
\[ DF = \beta_0 x_1^{\beta_1} x_2^{\beta_2} x_3^{\beta_3} \cdots x_M^{\beta_M}, \]  

(25)

Where \( \beta_m \) are the model parameters and \( x_m \) is the explanatory descriptors used for model development; \( M \) is the total number of the descriptors.

### 3.4.2 Support vector machine (SVM)

SVM regression has shown advantages in solving complicated non-linear problems in the field of hydrology. 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 \( \varepsilon \) deviation from the observed \( y_i \). Thus, the SVM regression model can be described as Eq. (26).

\[
\begin{align*}
\min_{w,b,\xi_i,\xi^*_i} & \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} (\xi_i + \xi^*_i), \\
\text{s.t.} & \quad f(x_i) - y_i \leq \varepsilon + \xi_i, \\
& \quad y_i - f(x_i) \leq \varepsilon + \xi^*_i, \\
& \quad \xi_i \geq 0, \xi^*_i \geq 0, i = 1,2,\ldots,N
\end{align*}
\]  

(26)

where \( \varepsilon \) is the insensitive loss, \( w \) and \( b \) are hyper-plane parameters; \( \xi_i \) and \( \xi^*_i \) are two slack variables; and \( C \) is a parameter that controls the trade-off between the support line and training samples.

The adopted SVM regression was proposed by Mariette et al. (2015) and the solution of Eq. (26) 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, each RTM is developed with the samples selected by a bootstrap aggregating method from the whole training sample. The out-of-bag (OOB) samples (samples not selected by the bootstrap method) are applied to test its accuracy. This strategy means RF usually has good performance in terms of overfitting, outliers and noise. The model structure of RF can be found in the previous studies (Zhao et al., 2020; Zhao et al., 2018). Once an RTM is developed, the error of OOB samples can be computed (termed \( E_{OOB1} \)). Each factor in the OOB samples is permuted one at a time, and the \( E_{OOB2} \) can be computed with the permuted OOB samples and the trained RF model. The RF estimates the factor importance by comparing the difference between \( E_{OOB1} \) and \( E_{OOB1} \) while all others are unchanged. The 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), but this is the first time the method has been applied in an RFFA to the best of the author’s knowledge.
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. (27) and (28) respectively. Both RMSNE and RBIAS are negatively oriented indices where lower values mean 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).

\[
\text{RMSNE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left( \frac{q_i - \hat{q}_i}{q_i} \right)^2},
\]
\[
\text{RBIAS} = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{q_i - \hat{q}_i}{q_i} \right).
\]

Where \(q_i\) is the flood runoff for specific return periods derived by observed data, \(\hat{q}_i\) is the flood runoff 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. The NORM and GENLOGIS are more suitable for high flow areas and EXP is recommended for low flow areas.
Figure 3 Thresholds for MK (a), SNHT (b) and AD (c) tests during station selection

Figure 4 Flood frequency curve selection and estimation
4.2 Subgroups delineation

The selected stations were clustered into several subgroups using the K-means model and an SVM regression model was developed in each subgroup for regional flood estimation. 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 during subgroups delineation](image)

Taking the 100-year flood estimation as an example, different combinations of clustering factors were tested using the optimal K. As shown in Table 4, the best combination is to use all factors for clustering, and this result reaches the best RRMSE 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>
<thead>
<tr>
<th>Clustering factors</th>
<th>Optimal K</th>
<th>Training</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RBIAS</td>
<td>RMSNE</td>
</tr>
<tr>
<td>All factors</td>
<td>16</td>
<td>-0.179</td>
<td>0.703</td>
</tr>
<tr>
<td>Meteorological factors</td>
<td>11</td>
<td>-0.202</td>
<td>0.768</td>
</tr>
<tr>
<td>Physiographical factors</td>
<td>5</td>
<td>-0.214</td>
<td>0.829</td>
</tr>
<tr>
<td>Hydrological factors</td>
<td>30</td>
<td>-0.207</td>
<td>0.781</td>
</tr>
<tr>
<td>Anthropological factors</td>
<td>7</td>
<td>-0.379</td>
<td>1.235</td>
</tr>
</tbody>
</table>
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. 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.

The factor importance in each subgroup was evaluated using the RF model and Figure 7 (a) describes the average value of factor importance of all subgroups. The top five important factors are the catchment area (CA) and annual precipitation (AP), Longitude (LO) and latitude (LA) and temperature range (TR). Meanwhile, population density (PD) and dam capacity (DC) and curve number (CN) are the three least important factors and make up a relatively low proportion (< 5%) of the total importance among all factors.
The impact of catchment descriptors on regression results was tested based on the optimal 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 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 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 0.708 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 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 error points, 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. However, some overestimation was found at low discharges, which is also reflected in the negative value of RBIAS during evaluation. It is logical that modelled values are usually less dispersed than observed values as the training target focuses on the mean value. Moreover, 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.

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 5 Result for different return period flood estimation

<table>
<thead>
<tr>
<th>Return periods</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean BIAS</td>
<td>Mean RMSE</td>
</tr>
<tr>
<td>10</td>
<td>-0.160</td>
<td>0.655</td>
</tr>
<tr>
<td>20</td>
<td>-0.165</td>
<td>0.663</td>
</tr>
<tr>
<td>50</td>
<td>-0.175</td>
<td>0.695</td>
</tr>
<tr>
<td>100</td>
<td>-0.179</td>
<td>0.703</td>
</tr>
</tbody>
</table>

Table 5 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 true, 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 0.30 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.

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 K-means model considering at-site explanatory factors. Therefore, the subgroups are not necessarily spatially consistent with climate zone and it may be difficult to properly reflect the heterogeneous characteristics of climate zones. This limitation is also noted in Dallaire et al. (2019) and could be improved by combining the unsupervised and supervised clustering models for subgroup delineation.

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.
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 which mainly occurred due to low discharge in small catchments.

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 Code availability


7 Data availability


8 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;

9 Competing interests

The authors declare that they have no conflict of interest.
10 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


