Evaluation of design flood estimates – a case study for Norway

Florian Kobierska[1,2], Kolbjørn Engeland[1] and Thordis Thorarinsdottir[3]

[1]The Norwegian Water Resources and Energy Directorate, P.O.box 5091 Majorstua, NOR-0301 Oslo, Norway

[2]Western Norway University of Applied Sciences, Institute of Natural Sciences, Sogndal
[3]Norwegian Computing Center, P.O. Box 114 Blindern, NO-0314 Oslo, Norway

Email: fbaffie@gmail.com , koe@nve.no , thordis@nr.no

Abstract

The aim of this study was to evaluate the predictive fit of probability distributions to annual maximum flood data, and in particular to evaluate (i) which combination of distribution and estimation method gives the best fit and (ii) whether the answer to (i) depends on record length. These aims were achieved by assessing the sensitivity to record length of the predictive performance of several probability distributions. A bootstrapping approach was used by resampling (with replacement) record lengths of 30 to 90 years (50 resamples for each record length) from the original record and fitting distributions to these sub-samples. Subsequently, the fits were evaluated according to several goodness of fit measures and to the variability of the predicted flood quantiles. Our initial hypothesis that shorter records favor two-parameter distributions was not clearly supported. The ordinary moments method was the most stable while providing equivalent goodness of fit.

Keywords: *Design Floods, Flood Frequency Analysis, Probability Distributions, bootstrapping, reliability, stability*

Introduction

The motivation for this study is the need to revise guidelines for design flood estimation in Norway. The design flood estimates form the basis for hazard management related to flood risk and is a legal obligation when building infrastructure such as dams, bridges and roads close to water bodies. Flood inundation maps used for land use planning are also based on design flood estimates. Existing guidelines are given in Midttømme *et al.* (2011) and Castellarin *et al.* (2012), and summarized in Table 1. The approach is based on using annual maximum floods, and the recommendations depend on the length of the local data record. A minimum of 30 years of local observations is required for local flood frequency analysis and at least 50 years of data should be available to use three-parameter distributions. The Gumbel (two parameters) and GEV (three parameters) are the preferred distributions. More recently Glad *et al.* (2015) found that the Generalized Logistic is the preferred distribution for annual maximum floods in small catchments.

Table 1. Guidelines for flood frequency analysis according to data availability

Data availability	Procedure for calculation of the index flood	Procedure for calculation of growth curve for target return periods between Q200 and Q1000
>50 years	Not used	Calculated from 2- or 3-parameter distribution, based on observed series
30-50 years	Not used	Calculated from 2-parameter distribution, based on observed series
10-30 years	Calculated from observed series	Calculated by analysis of other long series in the area
< 10 years		Calculated by analysis of other long series in the area
None		Use of regional flood frequency curves

Other guidelines for flood frequency estimation include USA (Stedinger and Griffis, 2008 and 2011), Australia (Ball *et al.*, 2016), and Europe (Castellarin *et al.*, 2012). The four distributions that are most commonly used for annual maximum floods are the generalized extreme value (GEV) distribution (Australia, Austria, Cyprus, Germany, France, Italy, Lithuania, Slovakia, Spain) with the Gumbel distribution (Finland, Greece) as a special case, the generalized logistic (UK) and the log-Pearson III (United States, Australia, Lithuania, Poland, Slovenia). Two-component Gumbel distributions are recommended in Italy and Spain in order to account for different flood generating processes.

Four methods are commonly used to estimate distribution parameters: ordinary moments, linear moments, maximum likelihood and Bayesian. The method of linear moments has been recommended for its robustness with small sample sizes (Hosking *et al.*, 1990). In recent years Bayesian flood frequency estimation has got an increased attention in the research community (e.g. Coles and Tawn, 1996; Gaal *et al.*, 2010; Gaume *et al.*, 2010; Renard *et al.*, 2013), and is recommended in the operational guidelines in Australia (see chapter 2.6.3 in Ball *et al.*, 2016). The benefit of the Bayesian method is the flexibility in model formulation, the possibility to include prior and/or regional knowledge in the local estimation, and the possibility to account for errors in rating curves (Ball *et al.*, 2016).

The recommendations provided in the national guidelines are in most cases based on systematic evaluations. Renard *et al.* (2013) provide a short review of evaluation frameworks and distinguish between simulation based and data based frameworks. In the simulation based approach, the true distribution is known, and Monte-Carlo-generated samples from the true distribution are used to assess the performance of different distributions and/or parameter estimation methods (e.g. Hosking *et al.*, 1985). It is especially useful for assessing robustness (e.g. Stedinger and Cohn, 1986) and evaluating the estimates of standard errors (e.g. Stedinger *et al.*, 2008). For data based approaches, the true distribution is not known, and the aim of the evaluation is to assess if the observations might be realizations of the estimated distribution. Goodness of fit tests combined with split-sample or cross-validation are used in order to assess the predictive performance of the fitted distribution. The goodness of fit criterions measure the reliability, i.e. how well the model fits to (independent) data. Renard *et al.* (2013) introduced "stability" as an additional criterion. It measures the sensitivity of the design flood estimates to different subsets of data. Design flood estimates that depend strongly on the underlying data might lead to re-assessment of the design flood. This can for example result in large costs for dam owners as the design of dams has to be re-assessed every 20 years. Stability is therefore an important criterion in order to choose between the most reliable models.

The aim of this study is to perform a systematic evaluation of the predictive performance of local flood frequency distributions and estimation methods applied to annual maximum data. The results will later be used as a foundation for recommendations in new guidelines.

In this study we wanted to answer the following research questions:

(i) Which combination of distribution and estimation method best fits the data?

(ii) Does the answer to (i) depend on local data availability?

To answer these questions we set up a test bench for local flood frequency analysis using data based evaluation methods inspired by Renard *et al.* (2013) by using a bootstrapping-approach where we systematically evaluated how the predictive performance depends on record length. The final aim is to update the flood frequency analysis guidelines for Norway.

Data

We used annual maximum floods from 529 streamflow stations of the Norwegian hydrological database "Hydra II". We present here a brief summary of the dataset and associated quality control methods, which are described in detail in Engeland *et al.* (2016). All data influenced by river regulations were removed. In addition, quality controls of the data including quality assessment by the field hydrologist and of the rating curve for high flows, were used to select flood data with a sufficient quality. For all gauging stations, we extracted a set of catchment properties (for details see Engeland *et al.*, 2016). Figure 1 shows the histogram for record length, catchment areas, lake percentage, mean annual temperature and precipitation and the rain contribution to floods. Figure 2 presents a map of mean annual precipitation, temperature and floods and the rain contribution to floods. All climatological descriptors are based on the gridded temperature and precipitation data product in SeNorge (www.senorge.no). In this study we used 280 stations which have at least 30 years of record. Only 103 stations have more than 50 years of data. The catchment area spans between 0.5 and 20300 km^2 with 163 km^2 as the median. The presence of lakes influences flood sizes, and 494 of the catchments has more than 1 % of the catchment area covered by lakes. For these catchments the median lake percentage is 6.5 %. The mean annual precipitation ranges from 400 to 3140 mm with 986 mm as the median. We see a strong west-east gradient with the highest precipitation on the west coast. The mean annual temperature ranges from -3.75 to 7.62 ºC with 0.21 ºC as median. The temperatures are influenced by elevation as well as latitude (temperature decrease with elevation and longitude). The relative contribution of rain was estimated by calculating the ratio of accumulated rain and snowmelt in a time window prior to each flood and then averaging these ratios over all floods (for details see Engeland *et al.*, 2016). Rainfall processes dominate most coastal catchments and none of the catchments are completely dominated by snowmelt. A majority of stations, i.e. those where contribution from snow melt is important, show a prevalence of floods in spring and very few floods during winter. The catchments dominated by rainfloods do not show a clear seasonal pattern by frequently displaying floods in summer and winter. Both the flood records and the catchment properties datasets (catchment area, record length, mean annual runoff and several other catchment descriptors) are available as supplementary materials.

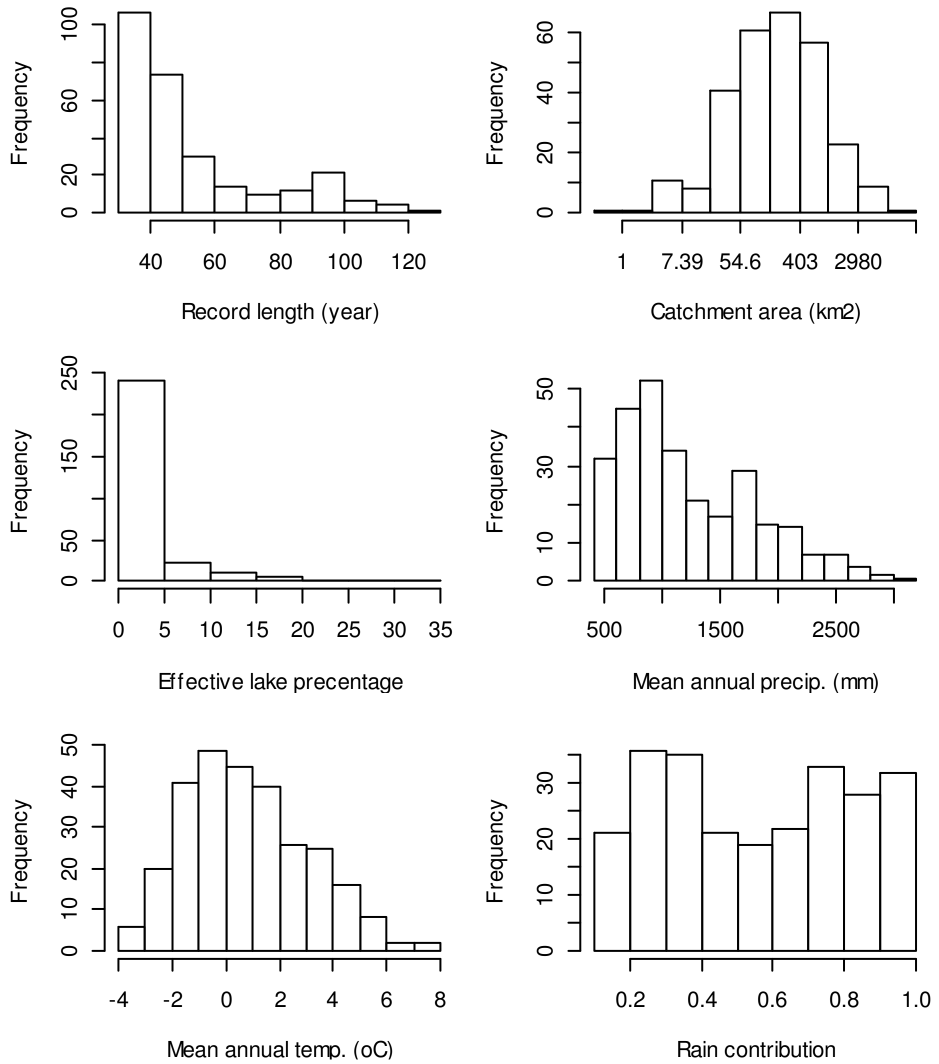

Figure 1. Histograms showing the distribution of record lengths, catchment area, lake percentage, mean annual precipitation and temperature; and the relative contribution from rain to floods.

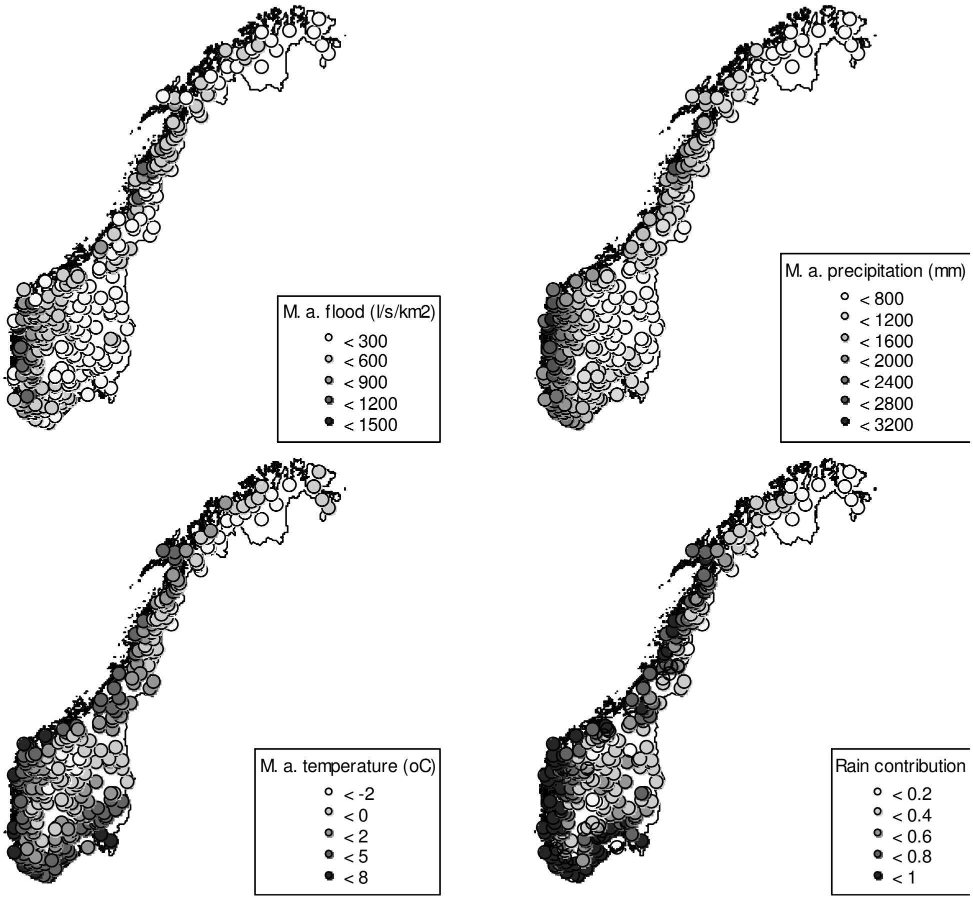

Figure 2. Maps showing the mean annual precipitation, temperature and flood (per unit area). The last map shows the contribution of rain precipitation to floods (index of flood generating processes).

Methods

Distributions

We evaluated five probability distributions: Generalized extreme value (GEV), Gumbel, Pearson III, Gamma and the Generalized Logistic (GL) distribution. The equations for the quantile functions and the probability density functions (pdf) are provided in the supplementary materials, below we provide the equations for the distribution functions. See also Bezak *et al.* (2014) for a recent overview.

Generalized extreme value distribution

The extreme value theorem is also known as the Fisher-Tippet theorem says that the maximum value from a sample of independent and identically distributed (iid) random variables follows the GEV distribution (e.g., (Embrechts *et al.*, 1997; Fisher and Tippett, 1928)

$$F(x) = \begin{cases} exp\left\{-\left[1-k\left(\frac{x-m}{\alpha}\right)\right]^{1/k}\right\} & k \neq 0 \\ exp\left\{-exp\left(-\frac{x-m}{\alpha}\right)\right\} & k = 0 \end{cases} \quad (1)$$

Where *m* is a location parameter, α scale parameter and *k* a shape parameter. Defined on the region $1 - k(x-m)/\alpha > 0$. The mean exists if $k > -1.0$, and the variance if $k > -0.5$. The shape parameter *k* is important in the GEV distribution as it shapes the tail of the distribution. A negative value indicates a heavy tail, whereas positive values describe a light tail and an upper limit for the variable *x*.

Gumbel distribution

The Gumbel distribution is a special case of the GEV distribution (shape parameter $k = 0$) and is written as:

$$F(x) = exp\left\{-exp\left(-\frac{x-m}{\alpha}\right)\right\} \quad (2)$$

Where *m* is a location parameter and α a scale parameter.

This distribution is often recommended for small datasets. Maximum values of random variables, with an exponential like upper tail (e.g. Normal, lognormal, Gamma), will theoretically follow a Gumbel distribution.

Generalized logistic

The Generalized logistic (GL) distribution (Hosking and Wallis, 1997) is recommended for flood frequency estimation in the United Kingdom (Robson and Reed, 1999) and was recently recommended for predicting floods in small ungauged catchments in Norway (Glad *et al.*, 2014). The distribution is a re-parameterization of the log-logistic distribution (Ahmad *et al.*, 1988), and has some similarities to the GEV distribution as shown in Equation 3:

$$F(x) = \begin{cases} \left\{1 + \left[1 - k\left(\frac{x-m}{\alpha}\right)\right]^{1/k}\right\}^{-1} & k \neq 0 \\ \left\{1 + exp\left(-\frac{x-m}{\alpha}\right)\right\}^{-1} & k = 0 \end{cases} \qquad (3)$$

Where m is a location parameter, α scale parameter and k a shape parameter. As for the GEV distribution, the GL distribution has an upper bound if $k > 0$. This is the case only when the skewness is negative whereas for the GEV distribution, there is also an upper bound for positive skewness, i.e. L-skewness < 0.17 (Robson and Reed, 1999). Thus for flood data we could expect the shape parameter to be between -0.5 and 0.2.

Gamma distribution

The gamma distribution is a flexible two-parameter distribution often used in environmental sciences.

$$F(x) = \frac{1}{\Gamma(k)}\gamma\left(k, \frac{x}{\alpha}\right) \qquad (4)$$

Here, Γ denotes the complete gamma function and γ the lower incomplete gamma function.

Pearson III

The Pearson type III distributions given as:

$$F(x) = \frac{1}{\Gamma(k)}\gamma\left(k, \frac{x-m}{\alpha}\right) \qquad (5)$$

Where m is a location parameter, α a scale parameter and k a shape parameter. For $m = 0$, the P3 distribution reduces to the gamma distribution. Applied to log-transformed floods, this distribution is recommended for flood frequency analysis in the USA (Stedinger and Griffis, 2008; Dawdy *et al.*, 2012) and Australia (Haddad and Rahman, 2008). Prior distributions are given in Reis and Stedinger (2005)

Fitting methods

Three methods for fitting the distributions to observed data were used: method of moments, method of linear moments and maximum likelihood.

Ordinary moments (O-moments)

The method of ordinary moments means that the moments (mean, variance and skewness) are estimated based on the data and subsequently the parameters of the selected distribution are calculated based on a theoretical relationship between the moments and the distribution parameters. Two parameter distributions need the estimates of mean and standard deviation whereas the three-parameter distributions would also require an estimate of the skewness. The specific equations for each distribution used in this study are given in Bezak *et al.* (2014) and are also provided as supplementary materials.

Linear moments (L-moments)

The method of linear moments is a popular method in hydrology since it is a direct analogue to the method of moments, easy to apply and the parameter estimates are less sensitive to outliers in the data (Hosking, 1990). As for the O-moments, the linear moments are estimated from the data and subsequently the parameters of the selected distribution are calculated based on a theoretical relationship between the L-moments and the distribution parameters. The specific equations for each distribution used in this study are given in Hosking (1990), and are also provided as supplementary materials.

Maximum likelihood (ML)

The maximum likelihood method chooses the values of the parameters estimates that maximize the probability of the data sample. This probability is the product of the probability density function evaluated at all observations (with a common parameter set) and is called the likelihood function $l(\theta \mid x)$ of the parameters θ *given* data $\boldsymbol{x}$. The objective is to maximize this function. The likelihood-functions are specified in Bezak *et al.* (2014). For numerical reasons, the log-likelihood (and not the likelihood) is maximized. For distributions used in flood frequency analysis, numerical optimization is needed for estimating the parameters. For small samples, the ML estimator is known to be more biased and to give larger estimation uncertainty compared to the two moment estimators for the GEV distribution (Hosking et al, 1985, Madsen *et al.*, 1997). It might also provide absurd estimates of the shape parameter (Martins and Stedinger, 2000). Those issues are most conveniently minimized by adding a prior likelihood for the shape parameter (Coles and Dixon, 1999; Martins and Stedinger 2000). An alternative estimation approach is suggested in Laio (2004). Finally, the shape parameter of the Pearson Type III distribution is challenging to estimate using the ML-approach (Arora and Singh, 1989). An estimation strategy is suggested in in Laio (2004).

Bayesian estimation

Bayes theorem combines the knowledge brought by the prior distribution and the data (through the likelihood) into the posterior distribution of parameters, whose pdf is noted $p(\theta|x)$.

$$p(\theta|x) = \frac{p(\theta)l(\theta|x)}{\int p(\theta)l(\theta|x)d\theta} \quad (6)$$

The Bayesian method might include prior knowledge that could be expert knowledge, regional information (e.g. Gaume *et al.*, 2010; Kuczera, 1982) or historical information (e.g. Reis and Stedinger, 2005; Viglione *et al.*, 2013). It is also possible to express the prior knowledge on the estimated quantiles, i.e. design floods (Coles and Tawn, 1996). It is also easy to extend it to non-stationary model accounting for trends or shifts in extremes (Benito *et al.*, 2004; Benjamin Renard *et al.*, 2013; Renard *et al.*, 2006). The Bayesian methods allows us to easily calculate predictive distributions, confidence intervals, and the median or mean of return levels based on the posterior sample from the distribution of parameters (Coles *et al.*, 2003; B. Renard *et al.*, 2013).

Evaluation methods

We followed the evaluation strategy specified by Renard *et al.* (2013) and evaluated goodness-of-fit according to both reliability and stability indices. Reliability evaluates how well the estimated model predicts return levels whereas stability measures to which degree the design flood estimates depend on the data used for estimation.

The approach used in Renard *et al.* (2013) is based on a split sample cross validation test where, at each station s, each sample is in turn used for estimation and evaluation. The aim of this study is to assess performance as a function of record length l. We therefore chose a bootstrapping strategy by drawing, with replacement, 50 random samples (noted m) for each record length l sampled every 5 years between 30 and 90 years (30, 35, 40…). Subsequently, for each sample, we fitted a distribution $F_{l,s,m}$, and derived the associated return levels $X_{T,l,s,m}$ and evaluation scores $H_{T,l,s,m}$ where T is the return period. The complete original flood data at each station was used for evaluation. Results were averaged over all subsamples to obtain average scores for each record length $H_{T,l,,s.}$ To yield general conclusions, station-specific results were then averaged over all sites and groups of similar sites in order to obtain evaluation score $H_{T,l,}$ as a function of record length, Both the fitted distribution parameters and the return levels were used for evaluation as described below.

Stability

The stability measure is a property of the statistical model only and we can thus evaluate it for any return period, including those greatly exceeding the length of record. Here we evaluated the stability by calculating the coefficient of variation (CV) of the return levels for each site s, each resampling record length l and each return period T over all subsamples $m = 1,…, 50$: $\mathrm{CV}_{T,l,s,.}$ Subsequently, we calculated the average coefficient of variation over all sites: $CV_{T,l}$. This allowed us to show CV as a function of record length for individual sites as well as averaged over several sites.

Reliability

Evaluation of distributions

The Anderson-Darling (AD) test measures the integral of the distance between empirical and fitted cumulative distribution functions. Here $F_{l,s,m}$ is the fitted distribution to subsample *m* for record length *l* at site *s* and $F_{n,s}$is the empirical distribution at site *s* with *n* data. It places more importance on the tail of the distribution than the Kolmogorov-Smirnoff test.

$$A_{l,s,m} = n \int \frac{\left(F_{n,s}(x) - F_{l,s,m}(x)\right)^{\wedge}2}{F_{l,s,m}(x) * \left(1 - F_{l,s,m}(x)\right)} dF_{l,s,m}(x) \qquad (7)$$

The Kolmogorov-Smirnov (KS) test evaluates how well an empirical distribution fits to a parametric one. The statistics is based on the maximum distance between the two cumulative distributions and should therefore be as small as possible:

$$D_{l,s,m} = \sup_{q} \left|F_{n,s}(x) - F_{l,s,m}(x)\right| \qquad (8)$$

Evaluation of thresholds

Since the aim of flood frequency analysis is to assess critical design flood, it is relevant to evaluate the fitted distributions according to how well they predict thresholds.

The Brier score (Brier, 1950) is commonly used for evaluating, and was used in this paper for evaluating the predicted T-years event for flood frequency distributions. The Brier score (BS) compares the predicted probability of the exceedance of a threshold $u_{T,s}$ (given by $1 - F_{l,s,m}(u_{T,s})$) to actual exceedance of the threshold by independent data (given by $\mathbb{I}\{x_{s,i} > u_{T,s}\}$):

$$B_{l,s,m}(F_{l,s,m}|u_{T,s}) = \frac{1}{n_s}\sum_{i=1}^{n_s}\left(1 - F_{l,s,m}(u_{T,s}) - \mathbb{I}\{x_{s,i} > u_{T,s}\}\right)^2 \qquad (9)$$

Where $u_{T,s}$ is the threshold defined by a return period T and $\mathbb{I}$ is an indicator function that is one if $x_{s,i} > u_{T,s}$ and otherwise zero.

The Quantile score (QS) compares observed floods $x_{s,i}$ to the estimated flood quantile $F_{l,s,m}^{-1}(1-1/T)$ for a given return period *T* and gives the difference a low weight if the observed flood is smaller than the estimated quantile.

$$Q_{l,s,m}\left(F_{l,s,m}\middle|T\right) = \left(x_{s,i} - F_{l,s,m}^{-1}\left(1-\frac{1}{T}\right)\right)\left(\left(1-\frac{1}{T}\right) - \mathbb{I}\left\{x_{s,i} \leq F_{l,s,m}^{-1}\left(1-\frac{1}{T}\right)\right\}\right) \quad (10)$$

Since the shortest records have 30 years of data, BS and QS were evaluated for return periods up to 30 years (2, 5, 15, 20 and 30). In particular, we selected the threshold $u_{T,s}$ in the BS equation from the empirical distribution of the complete dataset. For each station we applied the Hazen plotting position in Equation 11 (Makkonen, 2008), where *i* is the rank of the observation $Q_{(i)}$, *n* is the number of observations and $\hat{P}'_{(i)}$ is the estimated cumulative probability:

$$\hat{P}'_{(i)} = \frac{i-0.5}{n} \quad (11)$$

Evaluation of empirical L-moments

The L-moment ratio diagram compares sample estimates of τ_2, τ_3 and τ_4 (standard deviation, skewness and kurtosis) to the theoretical population for parametric distributions by plotting the relationship between τ_4 and τ_3 for three parameter distributions and between τ_3 and τ_2 for two parameter distributions. It was introduced by Hosking (1990), and approximations for several distributions are given in Hosking and Wallis (1997). The advantage of this evaluation is that we visually compare how several theoretical distributions fit to our data sample, and it has become a standard tool in regional flood frequency analysis (Peel *et al.*, 2001).

Results

Computational methodology

For the application of the Bayesian approach, we specified the priors for the shape parameters in the GEV and GL distributions to be Normally distributed with mean and standard deviations specified as *N(0, 0.2)* and *N(-0.15, 0.175)* respectively. The prior for the GEV parameters is suggested in Martins and Stedinger (2000), whereas the prior for the GL parameters were obtained from scatter plots of the L-moment skewness for flood data in UK (Robson and Reed, 1999).

Based on the methods presented above, our research approach was highly multi-dimensional and involved saving a high amount of data. For this reason, we chose to save the input and model data into a NetCDF database. The full computational chain was carried out with the R software (R Core Team, 2016)). The following libraries were used. *RNetCDF* (Michna and Woods, 2016) for managing the NetCDF files, *doSNOW* (Revolution Analytics and Weston, 2015a) and doMC for parallel backend on Windows and Linux respectively, *foreach* (Revolution Analytics and Weston, 2015b) for parallel computation. In addition the following libraries were used for fitting the distributions: *evd* (Stephenson, 2002), *nsRFA* (Viglione, 2014), *fitdistrplus* (Delignette-Muller, and Dutang 2015), *ismev* (Heffernan and Stephenson. 2016) and *pracma* (Borchers, 2017).

Two packages were created to facilitate the re-usability of this work. Code and data are available at https://github.com/NVE/FlomKart and https://github.com/NVE/fitdistrib.

Given the size and multidimensionality of both NetCDF files (estimated parameters and goodness-of-fit indices), an easy-to-use visualization tool was required to analyse the data. The R package *Shiny* (Cheng *et al.*, 2016) was used to create a browser-based graphical user interface. In addition the following libraries were used to create the graphical interface: *shinyBS* (Bailey, 2015), *leaflet* (Cheng and Xie, 2016), *DT* (Xie, 2015) and *formattable* (Ren and Russell, 2015).

The code of this visualization tool was organized as in R package available there: https://github.com/NVE/FlomKart_ShinyApp. For every station, key plots can be drawn to compare the modelled probability distribution to the empirical distribution of data, and the evaluation criterions are shown for each station. Since we in this study were interested in extracting general conclusions, we chose to present results aggregated over all stations.

Station averaged results

We starts by presenting the evaluation of reliability as average values over all stations and subsamples. The reliability measures, i.e. Kolmogorov-Smirnov test statistics, Anderson-Darling test statistics, Brier score, and quantile scores (QS), are shown in Figures 3-6 respectively. All 280 stations with more than 30 years of data were used, and the reliability measures are plotted as a function of the length of the sub-sample used for estimating distribution parameters. This allowed us to evaluate how the performance depends on the length of the available data. We made one subplot for each distribution and one line for each estimation procedure. In these plots, the lowest value indicates the best performance.

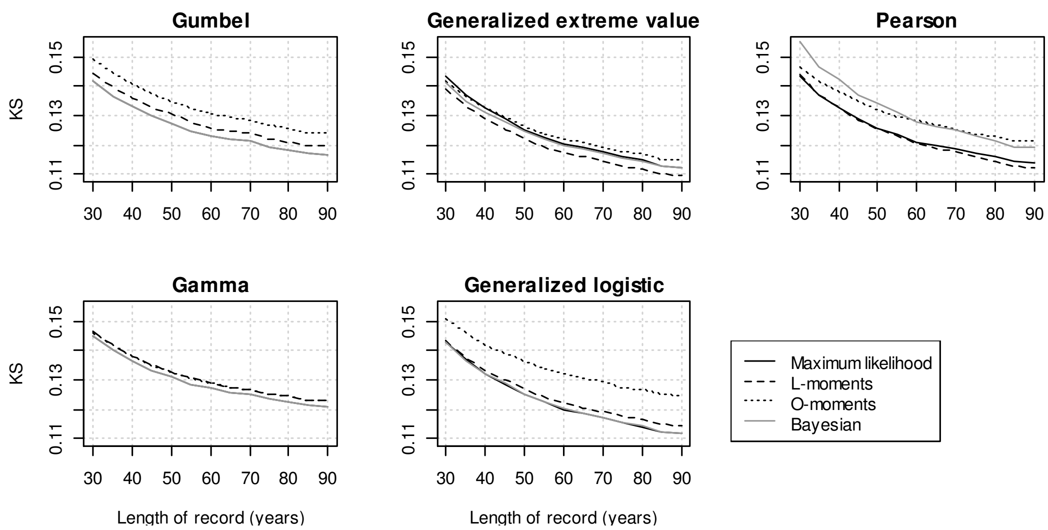

Figure 3. Evolution of KS, as a function of length of record, averaged over all stations with more than 30 years of record.

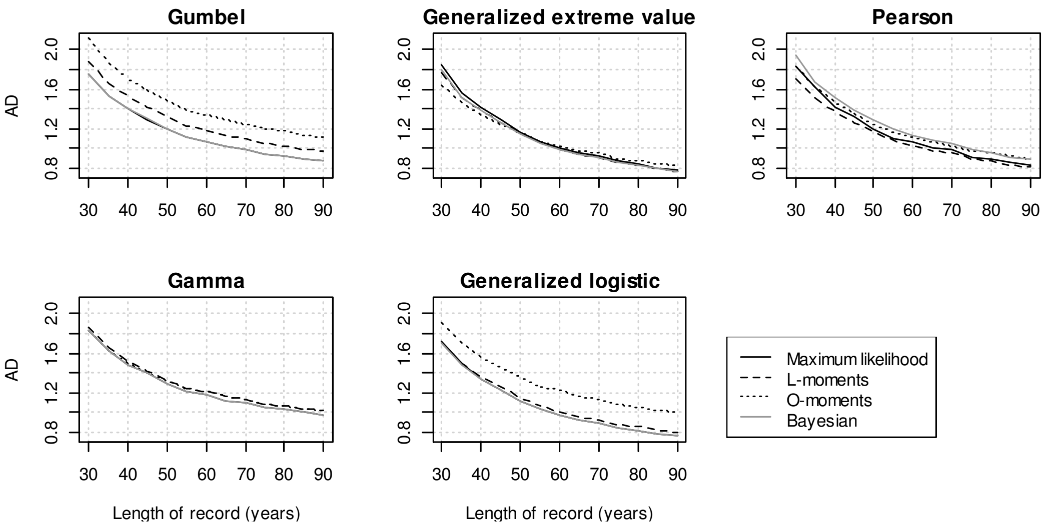

Figure 4. Evolution of AD, as a function of length of record, averaged over all stations with more than 30 years of record.

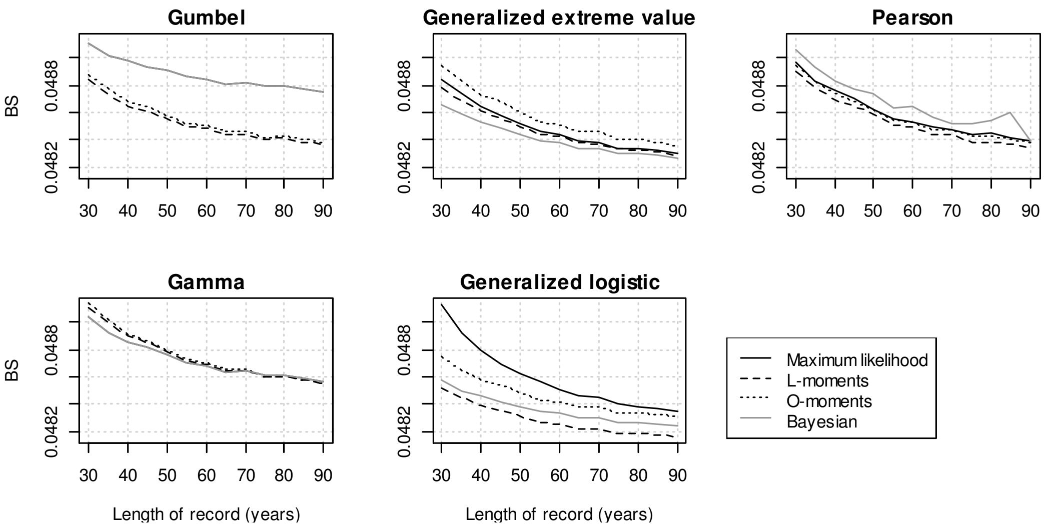

Figure 5. Evolution of BS, as a function of length of record, averaged over all stations with more than 30 years of record.

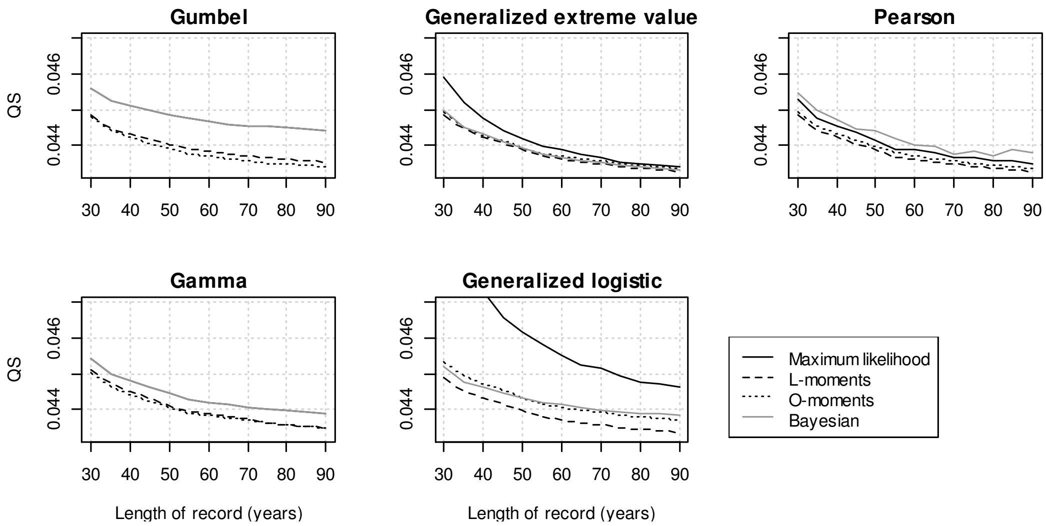

Figure 6. Evolution of QS, as a function of length of record, averaged over all stations with more than 30 years of record.

The evaluation according to stability is shown in Figure 7 where the average coefficient of variation in return levels is plotted as a function of record length. The calculation of the CV was based on the 100 sub-samples for each record length. All distributions and methods become more stable as record length increases.

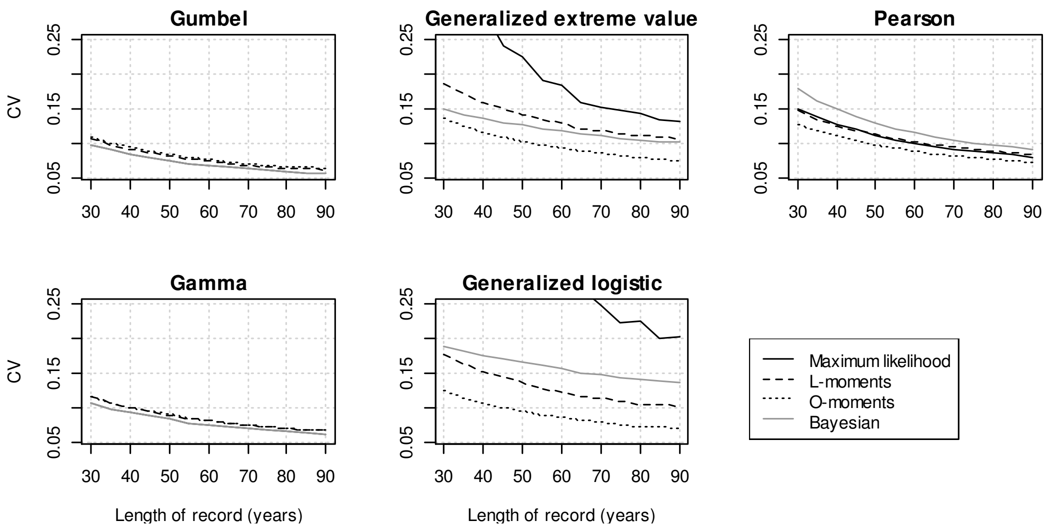

Figure 7. Evolution of the coefficient of variation (CV) of return levels averaged over all stations with more than 30 years of data.

In order to summarize the relative performance of the different distributions and estimation methods, Figure 8 contains a subplot of each of the performance measures. For each distribution, the estimation method providing the best performance was selected. For the three-parameter distributions, we excluded the maximum likelihood methods from the reliability criterions since it was only marginally performing better and provide unstable results. When selecting the estimation methods for the coefficient of variation, we excluded the method of moments from the three-parameter distributions, since this method never obtained the most reliable predictions. Figure 8 thus allowed us to compare the performance of the different distributions for the estimation method that performs the best for each of them.

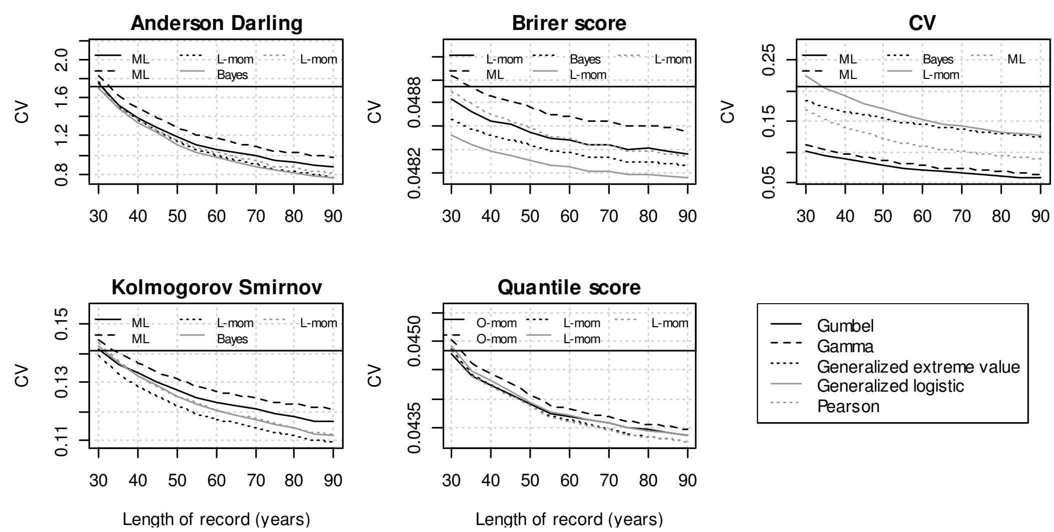

Figure 8. Plot of the best estimation method for each of the distribution as a function of record length.

The L-moments ratios plotted in Figure 9 give a good visual impression of the spread in L-kurtosis and L-skewness across all stations. A moving average of L-skewness along L-kurtosis removes much of the scatter and thus helps analysing the data.

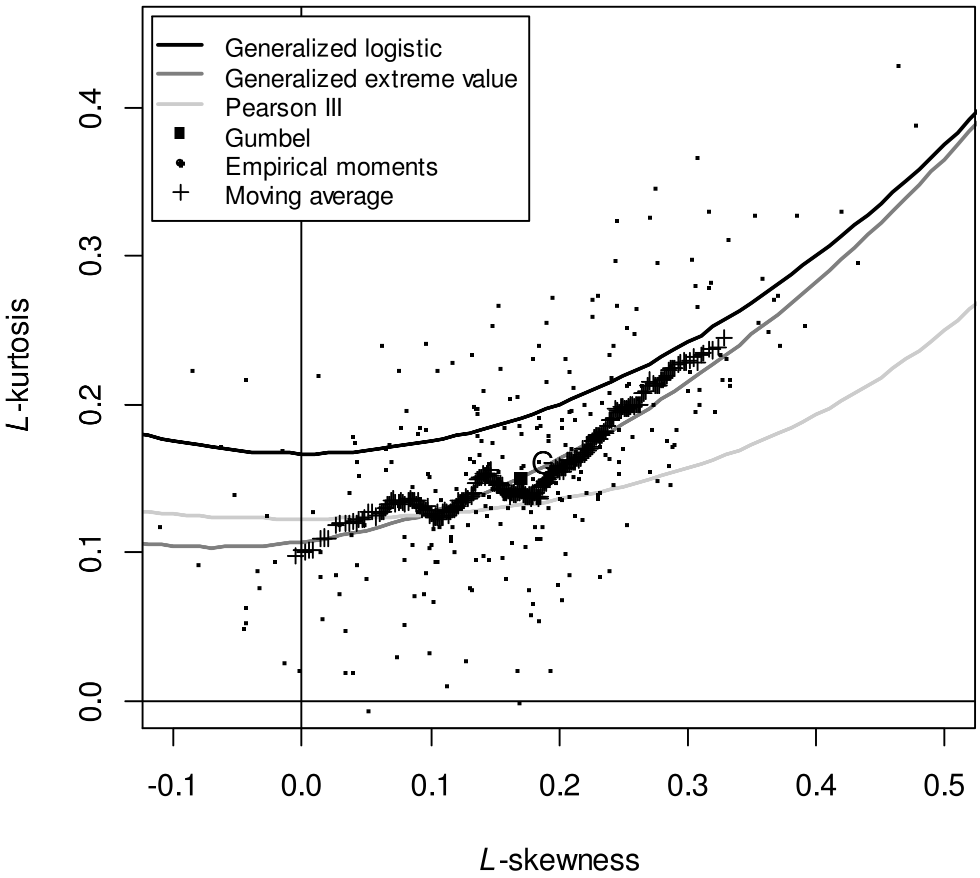

Figure 9. L-moment ratios for the 280 stations, the moving average of L-skewness over L-kurtosis, together with the theoretical distributions used in this study. Gamma and Pearson overlap. The black square is for Gumbel.

Discussion

The first research question raised in the introduction sought to determine which combination of distribution and estimation method best fits the data. From the results presented herein, we see that it is difficult to disentangle the performance of the estimation methods from the performance of the distributions, and that the combinations of estimation method and distribution that give the best performance vary between the performance measures. The interpretation of the results in order to answer the research questions, is therefore challenging.

From the performance of the reliability criterions, we see that the best estimation methods for the three-parameter distributions perform, in general, equally well or better than the best estimation methods for two-parameter distributions for all record lengths (Figure8). The gain in using a three-parameter distribution increases with record length. The only exception is the quantile score, where the Gumbel distribution is equally good as the three parameter distributions (Figure 8). Among the three-parameter distribution, the GEV and the GL distributions give the best performance. The GL distribution is better than the GEV distribution for the Brier score, whereas for the two other scores, the GEV distribution slightly outperforms the GL distribution. The GL distribution seems to be more challenging to estimate than the GEV distribution, since it is rather sensitive to the estimation methods used. Taking into account the stability criterion, the method of moments is most stable with the GL distribution. However, choosing to look only at the L-moments and Bayesian estimators that are the most reliable, we see that the difference in stability between the GEV and GL distribution according to stability is small (Figure 7). This indicates a slight preference for the GEV distribution.

Concerning the choice of estimation methods, the ML method should not be used in combination with three-parameter distributions since this combination provides very unstable results (Figure 7) and is, in some cases, only marginally better than the Bayesian and L-moment approaches (Figures 4, 5 and 6). The method of moments is the most stable method for all distributions (Figure 7), but it also provides the most unreliable results in for several scores (Figures 4, 5 and 6). For all three-parameter distributions, either the L-moments or the Bayesian methods is preferred (Figure 8).

An unexpected results, is the relatively low performance, as measured by the Brier- and Quantile scores, when the Bayesian and ML methods are used to fit the data to the Gumbel distribution. In contrast, these two estimation methods perform relatively well for the AD and KS test statistics (Figures 3 and 4). Further investigations revealed that this low performance is, to a large degree, controlled by the skewness for the original data. The relatively low performance for the Maximum Likelihood and Bayesian methods happens when the L-skewness is lower than 0.15, which is slightly lower than the L-skewness of the Gumbel distribution (0.17). This indicates that, for the Gumbel distribution, the ML and Bayesian estimators are more sensitive to low outliers in the dataset than the other estimation methods, and that they should be avoided when the L-skewness of the data is close to zero or negative.

The second research question was whether the answer to (i) depends on local data availability. To answer this question, we plotted all evaluation scores as a function of record length. As expected, for all evaluation scores, the performance improves with increasing record length. The difference in reliability between the distributions increases with record length, indicating that for the shortest record lengths, there is little gain in choosing a three-parameter distribution (Figure 8). The Brier score is an exception where the three parameter distributions are better than the two parameter distributions for all record lengths (Figure 5). With the exception of the method of moments, three-parameter distributions show lower stability than two-parameter distributions, even for the longest record length. There is no clear threshold in record length above which one should rather use a three-parameter distribution rather than a two-parameter distribution. A threshold at 50 years of record for switching from two- to three- parameter distributions could be justified if we only looked at the AD and QS test statistics. The difference between the GEV and Gumbel distributions is indeed small with those criterions. The Gumbel distribution is however considerably more stable for any length of record (Figure 8, upper right panel).

The results presented herein might be influenced by several factors that are not directly related to the choice of distribution. For the Bayesian method in particular, the choice of prior distribution might influence our conclusions. For the GEV distribution, values were chosen from the literature. Less information is available for the GL distribution, and the prior for the shape-parameter was set subjectively based on previous studies. For the Pearson-III distribution, we used a non-informative prior. We might therefore expect the performance of the Pearson-III distribution to be lower than for the other two. The results are prior-sensitive, in particular for the shortest record lengths. Providing different priors might change our conclusions. In addition, many of the algorithms used herein, require numerical solutions, and the convergence of these algorithms might in some cases be misleading. For the MCMC in particular, we could not monitor the convergence of the more than 390 000 chains that were estimated using our resampling approach.

The re-sampling with replacement approach allowed us to compare all stations with sample sizes longer than 30 years, i.e. resampled records of lengths up to 90 years were created from the original record of 30 years. The benefit of using this approach was that more stations could be included in the evaluation. We used 280 stations of which only 35 of them had record lengths of 90 years or more. The drawback of this approach wass that stations with short record lengths will got resampled several times. By grouping stations according to their length of record and plotting the group-averaged coefficient of variation of return levels for each group, we saw that (i) the average CV is the lowest for the shortest record lengths, and (ii) the spread in CV is the largest for the shortest record lengths. An explanation for the second issue is that the resampling approach used here might be sensitive to outliers in the underlying data, as those might be sampled several times for short records. We identified three stations that may exhibit this behaviour, but excluding them from the evaluation showed little influence on the average performance.

Conclusions and outlook

The aim of this study was to evaluate the predictive fit of probability distributions to annual maximum flood data, and in particular to evaluate (i) which combination of distribution and estimation method gives the best fit and (ii) whether the answer to (i) depends on record length. These aims were achieved by assessing the sensitivity to record length of the predictive performance of several probability distributions. A bootstrapping approach was used by resampling (with replacement) record lengths of 30 to 90 years (50 resamples for each record length) from the original records and fitting distributions to these sub-samples. Subsequently, the fits were evaluated according to several goodness of fit measures and to the variability of the predicted flood quantiles.

Based on the results presented herein we conclude that:

- The GEV and GL distribution provided the most reliable results.
- The method of linear moments or the Bayesian method are the recommended estimation methods.
- The maximum likelihood method was particularly unstable with three-parameter distributions, even for short return periods. This method should therefore be avoided.
- For the Gumbel distribution, the L-moment approach is recommended. The Bayesian approach was sensitive to the skewness of the data.
- The method of ordinary moments was consistently the most stable estimation method. This stability results in a light but consistent trade-off on goodness of fit against the method of linear moments.

- There is no clear threshold in record length above which one should rather use a three-parameter distribution rather than a two-parameter distribution.
- We focused on developing a reproducible workflow so that the methodology can be reused and improved as more data becomes available.

The results herein shows that the use of the GEV or the GL distribution is challenging since, in particular, the shape parameter is sensitive to the underlying data resulting in more unstable results. Alternative approaches, i.e. using a mixture of two parameter distributions, should therefore be investigated.

Acknowledgements

This work was jointly funded by The Research Council of Norway and Energy Norway (grant ES519956 FlomQ) and internal research funding at the Norwegian Water Resources and Energy Directorate. The data were extracted from the national hydrological database at the Norwegian Water Resources and Energy Directorate and are available upon request to the main author. The source code for estimating flood frequency distribution is done in the statistical programming language R (http://www.Rproject.org/) and is available on GitHub.

References

Ahmad M.I., Sinclair C.D. & Werritty A. 1988 Log-logistic flood frequency analysis. *Journal of Hydrology* 98, 205–224. doi:10.1016/0022-1694(88)90015-7

Arora K. & Singh, V.P. 1989 A comparative evaluation of the estimators of the log Pearson type (LP) 3 distribution. *Journal of Hydrology* 105, 19–37. doi:10.1016/0022-1694(89)90094-2

Benito G., Lang M., Barriendos M., Llasat C., Francés F., Ouarda T., Varyl R., Enzel Y. & Bardossy A. 2004 Use of systematic, paleoflood and historical data for the improvement of flood risk estimation. Review of scientific methods. *Natural Hazards* 31, 623–643. doi:10.1023/B:NHAZ.0000024895.48463.eb

Bezak N., Brilly M. & Šraj M. 2014 Comparison between the peaks-over-threshold method and the annual maximum method for flood frequency analysis. *Hydrological Sciences Journal* 59, 959–977. doi:10.1080/02626667.2013.831174

Brier, G.W. 1950 Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78, 1–3. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Castellarin A., Kohnová S., Gaál L., Fleig A., Salinas J.L., Toumazis A., Kjeldsen T.R. & Macdonald, N. 2012 European procedures for flood frequency estimation: *Review of Applied-Statistical Methods For Flood-Frequency Analysis in Europe*. ISBN: 978-1-906698-32-4. 2012)

Coles, S.G., Dixon, M.J., 2000. Likelihood-Based Inference for Extreme Value Models. *Extremes* 2, 5–23. doi:10.1023/A:1009905222644

Coles S., Pericchi L.R. & Sisson, S. 2003 A fully probabilistic approach to extreme rainfall modeling. *Journal of Hydrology* 273, 35–50. doi:10.1016/S0022-1694(02)00353-0

Coles S. & Tawn J. A. 1996 A Bayesian Analysis of Extreme Rainfall Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 45(4) 463-78. doi:10.2307/2986068.

Dawdy D.R., Griffis V.W. & Gupta V.K. 2012 Regional Flood-Frequency Analysis: How We Got Here and Where We Are Going. *Journal of Hydrologic Engineering* 17, 953–959. doi:10.1061/(ASCE)HE.1943-5584.0000584

Embrechts P., Klüppelberg C. & Mikosch, T. 1997 Modelling Extremal Events. Springer Berlin Heidelberg, Berlin, Heidelberg. doi:10.1007/978-3-642-33483-2

Engeland K., Schlichting L., Randen F., Nordtun K.S., Reitan T., Wang T., Holmqvist E., Voksø A. & Eide V. 2016 *Flood data: selection and quality control of flood data for flood frequency analyses.* (original title: Flomdata : utvalg og kvalitetssikring av flomdata for flomfrekvensanalyser), NVE Rapport, 2016:85, Oslo, Norway, 71 pp.

Fisher R.A. & Tippett L.H.C. 1928 Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society* 24, 180–190. doi:10.1017/S0305004100015681

Gaál L., Szolgay J., Kohnová S., Hlavčová K. & Viglione, A. 2010 Inclusion of historical information in flood frequency analysis using a Bayesian MCMC technique: a case study for the power dam Orlík, Czech Republic. *Contributions to Geophysics and Geodesy* 40, 121-147, doi:10.2478/v10126-010-0005-5

Gaume E., Gaál L., Viglione A., Szolgay J., Kohnová S. & Blöschl G. 2010 Bayesian MCMC approach to regional flood frequency analyses involving extraordinary flood events at ungauged sites. *Journal of Hydrology* 394, 101–117. doi:10.1016/j.jhydrol.2010.01.008

Glad P.A., Reitan T. & Stenius, S. 2014 *Regional formulas for flood frequency estimation in small catchments.* (original title: Regionalt formelverk for flomberegning i små nedbørfelt : naturfareprosjektet) NVE Rapport 2014:62, NVE, Oslo, Norway, 45 pp

Haddad K. and Rahman A. 2008 Investigation of at-site flood frequency analysis in south-east Australia. *IEM Journal, The Journal of The Institution of Engineers, Malaysia*, 69(3), 59-64.

Hosking J.R.M., Wallis J.R., Wood E.F. 1985 Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics* 27, 251–261.

Hosking J. R. M. 1990 L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*. doi:10.2307/2345653

Hosking J. R. M. & Wallis J. R. 1997 *Regional frequency analysis - an approach based on L-moments*, 224 pp., Cambridge Univ. Press, New York.

Kuczera G., 1982 Combining site-specific and regional information: An empirical Bayes Approach. *Water Resources Research* 18, 306–314. doi:10.1029/WR018i002p00306

Laio F. 2004 Cramer-von Mises and Anderson-Darling goodness of fit tests for extreme value distributions with unknown parameters, *Water Resources Research* 40, W09308, doi:10.1029/2004WR003204.

Madsen, H. Rasmussen, P.F. & Rosbjerg, D. 1997 Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events: 1. At-site modeling. *Water Resources Research* 33, 747–757. doi:10.1029/96WR03848

Makkonen, L. 2008 Problems in the extreme value analysis. *Structural Safety* 30(5), 405–419. doi:10.1016/j.strusafe.2006.12.001

Martins E.S. & Stedinger, J.R. 2000 Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research* 36, 737–744. doi:10.1029/1999WR900330

Midttømme G., Pettersson L.E., Holmqvist E., Nøtsund Ø., Hisdal H. & Sivertsgård R. (2011) *Guidelines for flood estimation.* (original title: Retningslinjer for flomberegninger), NVE retningslinjer, 4/2011, Oslo, Norway, 59 pp.

Peel M. C., Wang Q.J., Vogel R. M. & McMahon T. A. 2001 The utility of L-moment ratio diagrams for selecting a regional probability distribution. *Hydrological Sciences Journal* 46(1). doi: 10.1080/02626660109492806

Reis D.S. & Stedinger, J.R. 2005 Bayesian MCMC flood frequency analysis with historical information. *Journal of Hydrology* 313, 97–116. doi:10.1016/j.jhydrol.2005.02.028

Renard B., Kochanek K., Lang M., Garavaglia F., Paquet E., Neppel L., Najib K., Carreau J., Arnaud P., Aubert Y., Borchi F., Soubeyroux J.-M., Jourdain S., Veysseire J.-M., Sauquet E., Cipriani T. & Auffray, A. 2013 Data-based comparison of frequency analysis methods: A general framework. *Water Resources Research* 49, 825–843. doi:10.1002/wrcr.20087

Renard B., Lang M. & Bois P. 2006 Statistical analysis of extreme events in a non-stationary context via a Bayesian framework: case study with peak-over-threshold data. *Stochastic Environmental Research and Risk Assessment* 21, 97–112. doi:10.1007/s00477-006-0047-4

Renard B., Sun X. & Lang M. 2013 Bayesian methods for non-stationary extreme value analysis. In: AghaKouchak A, Easterling D, Hsu K, Schubert S, Sorooshian S (eds) *Extremes in a changing climate*, water science and technology library, 65. Springer, Netherlands, 39–95, doi:10.1007/978-94-007-4479-0_3

Robson A., & Reed D. 1999 *Statistical procedures for flood frequency estimation., Flood Estimation Handbook Vol 3,* Institute of Hydrology, Wallingford, UK

Stedinger, J. R & Griffis V. W. 2008 Flood frequency analysis in the United States: time to update. *Journal of Hydrological Engineering* 13(4), 199-204. Doi: 10.1061/(ASCE)1084-0699(2008)13:4(199)

Stedinger J. R & Griffis V. W. 2011 Getting from here to where? Flood frequency analysis and climate. *Journal of the American Water Resources Association* 47(3) 2011. doi: 10.1111/j.1752-1688.2011.00545.x

Viglione A., Merz R., Salinas J.L. & Blöschl, G., 2013. Flood frequency hydrology: 3. A Bayesian analysis. *Water Resources Research* 49, 675–692. doi:10.1029/2011WR010782

R-packages:

Bailey, E. 2015 shinyBS: Twitter Bootstrap Components for Shiny. R package version 0.61. https://CRAN.R-project.org/package=shinyBS

Borchers, H.W. 2016 pracma: Practical Numerical Math Functions. R package version 1.9.3. https://CRAN.R-project.org/package=pracma

Cheng J. & Xie Y. 2016 leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library. R package version 1.0.1. https://CRAN.R-project.org/package=leaflet

Cheng J, & Xie, Y. & McPherson, J. 2016 shiny: Web Application Framework for R. R package version 0.13.2. https://CRAN.R-project.org/package=shiny

Delignette-Muller, M.L. & Dutang, C. 2015 fitdistrplus: An R Package for Fitting Distributions. Journal of Statistical Software, 64(4), 1-34. URL http://www.jstatsoft.org/v64/i04/.

Heffernan, J.E. & Stephenson, A.G. 2016 ismev: An Introduction to Statistical Modeling of Extreme Values. R package version 1.41. https://CRAN.R-project.org/package=ismev, Original S functions written by Janet E. Heffernan with R port and R documentation provided by Alec G. Stephenson.

Michna, P & Woods, M. 2016 RNetCDF: Interface to NetCDF Datasets. R package version 1.8-2. https://CRAN.R-project.org/package=RNetCDF

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ren, K. & Russell, K. 2015 formattable: Formattable Data Structures. R package version 0.1.5. https://CRAN.R-project.org/package=formattable

Revolution Analytics & Weston, S. 2015a doSNOW: Foreach Parallel Adaptor for the 'snow' Package. R package version 1.0.14. https://CRAN.R-project.org/package=doSNOW

Revolution Analytics & Weston, S. 2015b foreach: Provides Foreach Looping Construct for R. R package version 1.4.3. https://CRAN.R-project.org/package=foreach

Stephenson, A.G. 2002 evd: Extreme Value Distributions. R News, 2(2):31-32, June 2002. URL: http://CRAN.R-project.org/doc/Rnews/

Viglione, A. 2014 nsRFA: Non-supervised Regional Frequency Analysis. R package version 0.7-12. https://CRAN.R-project.org/package=nsRFA

Xie, Y. 2015 DT: A Wrapper of the JavaScript Library 'DataTables'. R package version 0.1. https://CRAN.R-project.org/package=DT