



Intl. J. River Basin Management Vol. 6, Special Issue 1 (2008), pp. 1–9
 © 2008 IAHR, INBO & IAHS

The importance of statistical uncertainties in selecting appropriate methods for estimation of extremes

P.H.A.J.M. VAN GELDER, *TU Delft, Faculty of Civil Engineering and Geosciences, Delft, The Netherlands.*

ABSTRACT

Statistical uncertainties in the estimation of extremes are normally not well considered in present practice or even present research. This paper shows the application of extreme distributions to river variables with non negligible levels of uncertainty and presents an approach to deal with them. Four distribution types (Gumbel, Pearson III (log shifted Gamma), Lognormal and Generalised Pareto) are fitted to peaks over threshold or annual maxima data sets of river discharges. Piecewise exponential distributions are used to summarise the four distribution types and bootstrapping methods for quantifying the uncertainty in the design discharges. The paper furthermore presents relationships between probability density functions, convergence theorems and statistical tests to judge the goodness of fits.

Keywords: Extremes; uncertainty; flood risk; bootstrap; statistics.

1 Introduction

The development of probabilistic methods is of real interest for flood risk analysis, and in particular design decisions for high reliability flood defence structures. One of the main stumbling blocks to the development of probabilistic methods is substantiation of probabilistic models. In fact, it is frequently necessary to estimate an extreme value based on a very small sample of existing data. The aims of this paper is to elaborate on the issues surrounding the development of realistic statistical methodology to substantiate a probability distribution, notably in the distribution tail.

The sections 2 to 3 give a reminder of the context of the data statistical treatment and the issues and problems surrounding structural reliability applications. Section 4 concentrates on the determination of adequacy of a distribution type. Section 5 recalls the relationships between extreme laws and the study of tail distribution. Section 6 describes a case study concerning extreme river discharges of the American River. The paper is completed with a conclusion (section 7) and list of references.

This paper adopts a probabilistic position to model the uncertainty associated with a value or sample dispersion: accordingly, other approaches (such as fuzzy sets, neural networks, etc.) will not be dealt with in the following. The aim of this article is not to compare the advantages and drawbacks of the various available methods, but to present representations of statistical uncertainties in the estimation of extremes.

2 Description of uncertainties associated with data

Tools to describe sample dispersion are taken from the statistics; however, their effectiveness is a function of the sample size.

References 1 and 2 describe methods that may be used to adjust a probability distribution to a sample, and then verify the adequacy of this adjusted distribution in the region with the highest probability density of the random variables. It is obvious that if data is lacking or scarce, these tools are difficult to use. Under such circumstances, it is entirely reasonable to refer to expert opinion in order to model uncertainty associated with a value, and then transcribe said information in the form of a probability distribution. This paper does not describe methods available in these circumstances, for example, the maximum entropy principle [2], and expert opinion methods based on Bayesian modelling [3].

The practical approach may be summarised by three scenarios:

Scenario 1. If a lot of data is available, the frequentist statistics is generally used. The objectivist or frequentist interpretation associates the probability with the observed frequency of an event. In this interpretation, the confidence interval of a parameter, p , has the property that the actual value of p is within the interval with a confidence level α ; this confidence interval is calculated based on measurements.

Scenario 2. If data is not as abundant, expert opinion may be used to obtain modelling hypotheses. The Bayesian analysis is used to correct *a priori* values established based on expert opinion as a function of observed events. The subjectivist (or Bayesian) interpretation treats probability as a degree of belief in a hypothesis. In this interpretation, the confidence interval is based on a probability distribution representing the analyst's degree of confidence in the possible values of the parameter and reflecting his/her knowledge of the parameter.

Received on April 2, 2007. Accepted on December 20, 2007.

Scenario 3. If no data are available, probabilistic methods may be used that are designed to reason based on a model that allows the value sought to be obtained from other values (referred to as the input parameters). The data to be gathered thus concerns the input parameters. The quality of the probabilistic analysis is a function of the credibility of statistics concerning these input parameters and that of the model.

The adjustment of a probability distribution and subsequent testing of the quality of the adjustment around the central section or the maximum failure probability region (the region where the joint density function of all random variables has a maximum and leading to failure or a negative limit state function) of the distribution constitute operations that are relatively simple to implement using the available statistical software packages (SAS, Matlab, SPSS, Stagraphics, Splus, Statistica [19]), for conventional laws in any case. However, the interpretation and verification of results still requires the expertise of a statistician. For example, the following points:

- the results of an adjustment based on a histogram is sensitive to the intervals width; as a rule of thumb the rounded value of \sqrt{n} is used as the number of classes in which the data is binned;
- the maximum likelihood or moment methods are not suitable for modelling a sample obtained by simultaneous occurring phenomena and for using the models outside a given limit of an observation variable;
- moment methods assume estimations of kurtosis and symmetry coefficients that are only usually specified for large bases (with at least one hundred values for kurtosis);
- most statistical tests, specifically, the most frequently used Kolmogorov-Smirnov, Anderson-Darling and Cramer-Von Mises tests, are asymptotic tests. These are goodness-of-fit tests and check if the data departs from a given distribution function. Stephens [21] found the Anderson-Darling test [20] to be one of the best statistics for detecting most departures from normality.
- in the Bayesian approach, the distribution selected *a priori* influences the result. Furthermore, the debate concerning whether or not the least informative law should be used has not been concluded.

In flood risk analysis studies, the characteristics of databases must be taken into consideration which may render implementation of an adjustment difficult. For example the following frequently encountered scenario:

- (1) the sample size is small and therefore asymptotic results need to be handled cautiously as well as approximations of more or less valid moments of an order greater than two;
- (2) sample data values are measured with an uncertainty;
- (3) sample homogeneity is not verified (mixture of samples taken from different populations, overlaying of phenomena (such as wave heights caused by local wind fields or by swell, or wind surges caused by storms from the west, or storms from the south), etc.);

- (4) if the area of interest is a distribution tail, it may be noted that the statistical theory and above all associated tools are less developed.

3 Selection of a distribution in practice

In practice, the criteria relevant to selecting a probabilistic model for a random variable would seem to be:

- Use a family compatible with the physical properties (bound value, symmetrical or not, exhibits exponential decay, etc.),
- the result of data adjustment,
- distribution not rejected by statistical tests,
- the selection of a distribution that is least “informative” with respect to data or available information (i.e., the introduction of too many hypotheses is avoided).

Statistical tests are used to decide whether or not the adequacy of a selected distribution should be rejected *a priori* with a confidence level. Whereas, for samples that are not very homogeneous or small in size, several distributions are frequently accepted to represent the sample or, on the contrary, no conventional distribution can be accepted.

To select a specific distribution, the following methods may be applied:

- either rely on selection based on expert opinion or current practice;
- or to rely on statistical convergence theorems which says that the average of a large number of i.i.d. (independent and identically distributed) random variables should converge to a normal distribution and the minimum and maximum of a large number to a generalised extreme value distribution, and the peaks over a threshold value to a generalised Pareto distribution;
- or compare confidence levels taken from various statistical tests for each accepted distribution and select the distribution associated with the highest confidence level;
- or further analyse adequacy in the zone of interest and select the distribution most graphically suited in the region of interest (for example a distribution tail for a reliability problem).

Another possibility is to use a resampling method: for example the Bootstrap method or cross-validation method, in order to select “the most relevant” distribution family (for example Refs 4 and 5). Consider that a random sample of observation, $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$, is used to obtain a sample estimate θ_s of a parameter of interest θ , which can be a quantile or some other statistic. The purpose of bootstrap simulations is to estimate the uncertainty (bias and variance) associated with the sample estimate θ_s . In the standard version of bootstrap, a random sample of size n is drawn with replacement from the ordered sample $\{X_{1:n}, X_{2:n}, \dots, X_{n:n}\}$ as

$$X_j^* = F_E^{-1}(p) = X_{[np]+1} \quad (\text{for } j = 1, \dots, n)$$

where $F_E^{-1}(p)$ denotes the empirical (sample) quantile function, p is a uniform random variable on (0-1), and $[np]$ denotes the integer floor function. This method is also known as non-parametric

bootstrap, as no parametric distribution function is used for simulating bootstrap samples.

Using a k th bootstrap sample, denoted by

$$\mathbf{X}^*(k) = \{X_1^*, X_2^*, \dots, X_n^*\}, \quad (k = 1, 2, \dots, b)$$

a new bootstrap estimate θ_k^* of θ_s can be obtained. Here, b denotes the number of bootstrap simulations. The set of estimates, $\theta^* = \{\theta_1^*, \theta_2^*, \dots, \theta_b^*\}$, constitutes the sampling distribution of θ_s .

As regards the non-parametric Bootstrap method, “the most relevant” family is defined as that which is almost always accepted by adjustment tests. This method assumes that sample size is not too small. Experience has also shown that the non-parametric Bootstrap method is rarely suited to the treatment of extreme values.

As regards the parametric Bootstrap method, “the most relevant” family is defined as the family whose confidence intervals on parameters are the smallest.

In the event that the graphic adjustment of a sample for various distribution features a “break”, the rejection of all conventional distributions may be noted. For example, in Fig. 1, graphical adjustment of the annual maxima of the River Osinovka (Russia), given by the red triangles, seems impossible by just one straight line, caused by “a break” halfway. These breaks may be explained by the fact that the sample is a mixture of several homogeneous sub-samples or is taken from a process combining several phenomena. Under such circumstances, if application is a reliability calculation or this distribution tail is, *a priori*, influencing, there is no choice, the model must be rejected. For an uncertainty propagation calculation, this model may be accepted if the number of simulations is not too high (overly frequent simulations in the distribution tail should be avoided).

Practical solutions to obtain a general model include the use of mix models and definition of a distribution through rejoining.

A mix model is assumed to be a weighted sum of random variables, for example, in the case of 2 Gaussian components:

$$G_{\theta,\gamma} = p \times \Phi_{\mu_1,\sigma_1} + (1 - p) \times \Phi_{\mu_2,\sigma_2}$$

where this distribution is a mixture of two normal distributions, means μ_1 and μ_2 respectively, standard deviations σ_1 and σ_2 respectively.

The principle of the rejoining method is to build a distribution of a random variable X by rejoining two distribution functions, marked F on the left and H on the right of a threshold u . Hence, for an observation x of X , it may be written:

$$\text{if } x \leq u \text{ then } X \rightarrow F \quad \text{and} \quad \text{if } x > u \text{ then } X \rightarrow H$$

The two laws F and H are therefore truncated. For every value a of $[0; 1]$, it is possible to construct a continuous rejoining in u based on the value a at this point.

$$G_a(x) = a \frac{F(x)}{F(u)} \quad \text{if } x \leq u$$

$$G_a(x) = 1 - (1 - a) \frac{1 - H(x)}{1 - H(u)} \quad \text{if } x > u$$

where G_a is the distribution resulting from rejoining.

4 Proposed method to determine adequacy of a distribution (central value)

This proposal is put forward in the context of a Kolmogorov-Smirnov test. However, it may be used for another *a priori* statistic, like Anderson-Darling, Cramer-von Mises, etc.

Take for example a random sample E , size N , order of magnitude y . The $f(x, c_0)$ type distribution (normal, Gumbel, etc.) was adjusted on this sample; c_0 , the vector of parameters obtained and KS_0 the value of the Kolmogorov-Smirnov test was noted. The following question was raised: what is the probability that a random sample of size N taken from the distribution $f(x, c_0)$ would allow a value of the Kolmogorov-Smirnov test equal to KS_0 to be obtained?

Tables of critical values are available that were obtained for a sample size tending towards infinity. Using these values when N is small is therefore questionable.

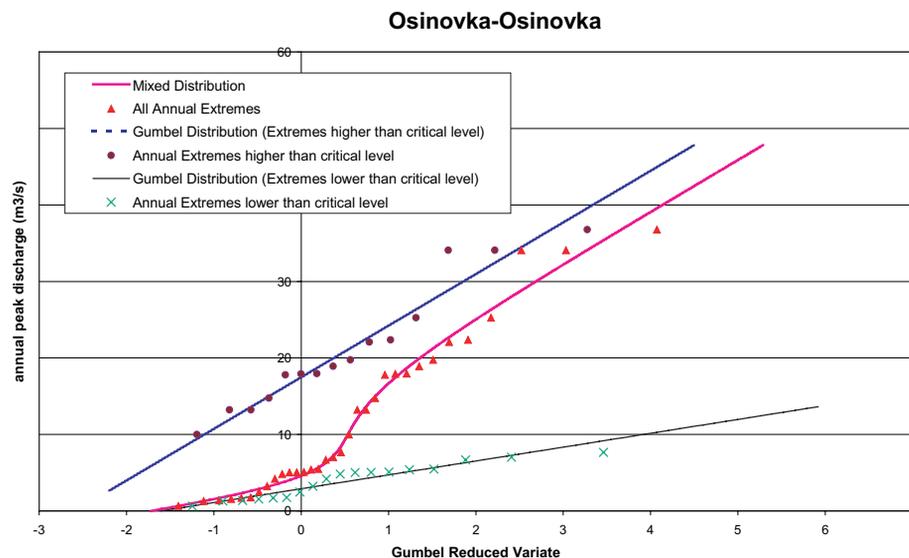


Figure 1 Example of a sample from a process combining several phenomena [18].

A confidence interval concerning the Kolmogorov–Smirnov test value is constructed as follows:

- (1) Based on the $f(x, c_0)$ distribution, that is to say that which has been adjusted, a large number p , of samples of size N are constructed (that is to say the same size as sample E).
- (2) For each of these samples, the distribution parameter vector is established and the value of the Kolmogorov–Smirnov test, noted KS_i is calculated.
- (3) This provides a sample with p values of KS_i . As p is very large, the fractiles of the statistic KS may be calculated with confidence as may be the probability of the KS_0 value.

Based on the above, a decision may be made whether or not to reject the hypothesis, that according to the Kolmogorov–Smirnov test the distribution $f(x, c_0)$ is an acceptable model of sample E .

Lindley [12] has shown that the elicitation and modulation of expert judgment concerning quantities of interest (such as model constants) often involves the use of the normal distribution.

The central limit theorem explains why we might see so many times a normal distribution in practice: a stochastic variable that is influenced by a large number of independent processes will be approximately normally distributed.

Roughly, the central limit theorem says that the sum of a number of (independent) samples taken from any distribution is approximately normally distributed. As we add more terms the approximation becomes better. This does not only apply to the sum but also to the average (which makes sense if one knows that if X has a normal distribution then it follows that also a X has a normal distribution). In mathematical terms, given a set X_1, X_2, \dots, X_n of i.i.d. random variables, suppose that each X_i has mean $E(X_i) = \mu$ and variance $\text{var}(X_i) = \sigma^2 < \text{infinity}$. Define $Y_n = \sum X_i$

Then:

$$\lim_{n \rightarrow \infty} \frac{Y_n - n\mu}{\sigma\sqrt{n}} \text{ has a standard normal } N(0,1) \text{ distribution.}$$

The central limit theorem for the sum of random variables can easily be applied to the product of random variables by noting that “ $\ln \prod = \sum \ln$ ”, and therefore the product of n i.i.d. random variables converges to a lognormal distribution.

More relationships between PDF’s can be derived. The following relations were found by Van Gelder [15]:

Table 1 Relationships between PDFs.

X	$Y = \exp(X)$
Normal	Shifted Log Normal
Gumbel	Frechet
Exponential	Pareto

The proof for the last two relationships is given by:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(e^X \leq y) = P(X \leq \ln(y)) \\ &= F_X(\ln(y)) = e^{-e^{-\frac{\ln(y)-\xi}{\alpha}}} = e^{-e^{-\frac{\ln(y)}{\alpha}} e^{\frac{\xi}{\alpha}}} = e^{-e^{\frac{\xi}{\alpha}} y^{-\frac{1}{\alpha}}} \\ &= e^{-(e^{-\xi} y)^{-\frac{1}{\alpha}}} = e^{-(y/e^{\xi})^{-\frac{1}{\alpha}}} \end{aligned}$$

This is a Frechet distribution with location 0, scale e^{ξ} and shape $1/\alpha$.

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(e^X \leq y) = P(X \leq \ln(y)) \\ &= F_X(\ln(y)) = 1 - e^{-e^{-\frac{\ln(y)-\xi}{\alpha}}} = 1 - e^{-e^{-\frac{\ln(y)}{\alpha}} e^{\frac{\xi}{\alpha}}} \\ &= 1 - e^{\frac{\xi}{\alpha}} y^{-\frac{1}{\alpha}} = 1 - \left(\frac{e^{\xi}}{y}\right)^{\frac{1}{\alpha}} \end{aligned}$$

This is a Pareto distribution with scale e^{ξ} and shape $1/\alpha$. Also the following relationship could be proven:

If X is Weibull then $Y = \ln(X - \zeta)$ is Gumbel for minima distributed:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(\ln(X - \zeta) \leq y) = P(X \leq e^y + \zeta) \\ &= 1 - \exp\left(-\frac{e^y}{\beta}\right)^{\delta} = 1 - \exp\left(-\frac{e^y}{e^{\ln(\beta)}}\right)^{\delta} \\ &= 1 - \exp\left(-e^{y-\ln(\beta)}\right)^{\delta} = 1 - e^{-e^{\frac{y-\ln(\beta)}{\delta-1}}} \end{aligned}$$

This is a Gumbel distribution for minima.

Finally it is possible to examine the tail behaviour of distributions. The following relations were found (properties of the Halphen distributions are described in Refs 13 and 14):

Table 2 Tail behaviour as a function of the return interval T .

X	Tail
Log-normal	$\exp(\sqrt{\ln(T)})$
Gamma	$\ln(T)$
Normal	$\sqrt{\ln(T)}$
Halphen	T^{ν}

5 Extreme laws and tail distributions

Assume that X_1, X_2, \dots, X_n are independent and identically distributed random variables coming from a parent distribution with a cumulative distribution function $F(x)$ and probability distribution function $f(x)$. Define $H_n = \max(X_1, X_2, \dots, X_n)$ then H_n has a cumulative density function (CDF) given by:

$$H_n(x) = P(\max(X_1, X_2, \dots, X_n) < x) = F^n(x)$$

Notice that the percentiles of H_n move to the right with increasing n , approaching the upper and lower end points if they are bounded, or going to ∞ if they are unbounded. When n goes to infinity, we have

$$\begin{aligned} \lim_{n \rightarrow \infty} H_n(x) &= 1 \quad \text{if } F(x) = 1 \\ \lim_{n \rightarrow \infty} H_n(x) &= 0 \quad \text{if } F(x) < 1 \end{aligned}$$

that is, the limit distribution degenerates to a Dirac function. To avoid this degeneracy, we transform the random variable X by means of constants a_n and b_n such that

$$\lim_{n \rightarrow \infty} H_n(a_n + b_n x) = \lim_{n \rightarrow \infty} F^n(a_n + b_n x) = H(x)$$

where $H(x)$ is a no degenerated CDF.

For instance:

$$F_{X_{\max}}(x) = (1 - e^{-\lambda x})^n \rightarrow \infty \quad (n \rightarrow \infty, \forall x).$$

However, if x also goes to infinity with $a_n + b_n x$ with $a_n = \ln(n)/\lambda$ and $b_n = 1/\lambda$ then:

$$\begin{aligned} F_{X_{\max}}(x) &= F_{X_{\max}}(a_n + b_n x) = F_{X_{\max}}(\ln(n)/\lambda + x/\lambda) \\ &= (1 - e^{-\lambda(\ln(n)/\lambda + x/\lambda)})^n = (1 - e^{-x}/n)^n \\ &\rightarrow \exp(-e^{-x}) \quad (n \rightarrow \infty). \end{aligned}$$

So X_{\max} converges to Gumbel. $F_{X_{\max}}(a_n + b_n x)$ is Gumbel distributed with parameters (0,1) for $n \rightarrow \infty$. So X_{\max} is Gumbel distributed with parameters (a_n, b_n) for $n \rightarrow \infty$.

$$\begin{aligned} \text{So } : E(X_{\max}) &\rightarrow \ln(n)/\lambda + 0.57772/\lambda \quad (n \rightarrow \infty) \quad \text{and} \\ \sigma(X_{\max}) &\rightarrow 1.2825/\lambda \quad (n \rightarrow \infty). \end{aligned}$$

The above can be illustrated to the data of the American River. The American River watershed drains the west side of the Sierra Nevada mountain range and is located just east of Sacramento, California. It contains 1,888 square miles of drainage area. The American River flows through the city of Sacramento. The Yuba River Watershed is located approximately 50 miles north of Sacramento, California, and is tributary to the Feather River. The Yuba River joins the Feather River at the location of the cities of Marysville and Yuba City.

Assume the river discharges of the American River are exponentially distributed with parameter λ . In Fig. 2 the ten highest river discharges (> 80000 CFS (cube feet per second)) are depicted. In the same figure the distribution of X_{\max} (being $(1 - e^{-\lambda x})^n$) and the limit distribution (being Gumbel) are given.

When $F(x)$ satisfies the limit $\lim_{n \rightarrow \infty} F^n(a_n + b_n x) = H(x)$, we say that $F(x)$ belongs to the domain of attraction of $H(x)$. The surprising result is that there are only 3 possible CDF's for $H(x)$ [8]. They are given by:

$$\begin{aligned} \text{Frechet:} \quad & H(x) = \exp(-x^{-c}), \quad x \geq 0, c > 0 \\ \text{Weibull} \\ \text{for max.:} \quad & H(x) = \exp(-(-x)^{-c}) \quad x \leq 0, c < 0 \\ \text{Gumbel:} \quad & H(x) = \exp(-\exp(-x)), \quad -\infty < x < +\infty \end{aligned}$$

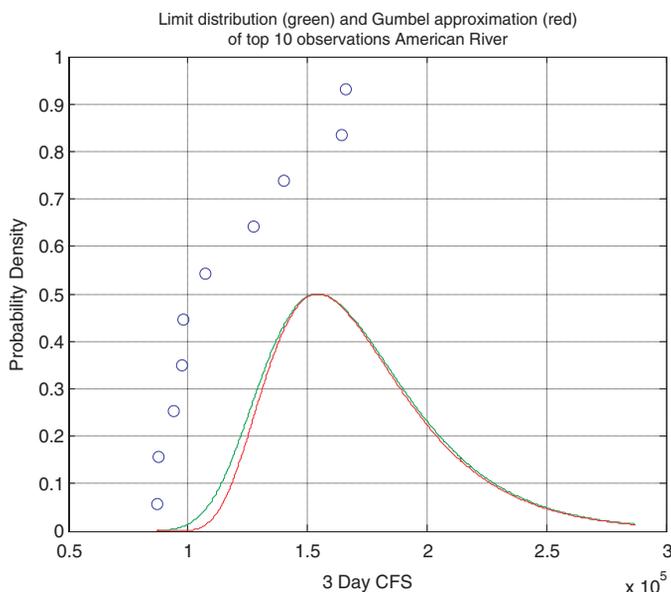


Figure 2 Extreme values with $n = 10$ (parent is Exponential).

The practical importance of this result is that, when we are dealing with extremes, and n is large enough, the infinite many degrees of freedom we have with the distribution function for these extremes, are reduced to 3 parametric families. In other words, no matter what $F(x)$ is the exact parent distribution, one of the above given 3 distributions can be used for approximating the extremes of $F(x)$. The problem is now, which distribution (or domain of attraction) is associated with our CDF $F(x)$? With the following result from [9], this question can be solved:

A necessary and sufficient condition for $F(x)$ to belong to the domain of attraction $H(x)$ is that:

$$\begin{aligned} \lim_{\varepsilon \rightarrow 0} (F^{-1}(1 - \varepsilon) - F^{-1}(1 - 2\varepsilon)) \\ \times (F^{-1}(1 - 2\varepsilon) - F^{-1}(1 - 4\varepsilon))^{-1} = 2^\varepsilon \end{aligned}$$

with the following correspondence:

$c > 0$	H is Frechet
$c = 0$	H is Gumbel
$c < 0$	H is Weibull (for max.)

Sometimes physical considerations can be used to reduce the three limiting distributions to two. Parent distributions with a finite endpoint (like wave heights in shallow water) cannot lie in a Frechet-type domain of attraction (because they have their domain for x to infinity). Moreover, if we consider that the Gumbel distribution can be approximated as closely as desired by Weibull for maxima or Frechet, we conclude that the limit distribution can be selected solely from physical considerations. *If we are dealing with random variables limited in the tail to the right, then a Weibull for maxima distribution is the limiting distribution.*

From the assumption that we only have a set of extreme observations where the parent distribution $F(x)$ is unknown, we would like to determine the domain of attraction. An estimator for the c -parameter can be found with the Pickands' method. Reference 6 shows that this c -parameter is the same as the one in a Generalized Pareto distribution given by:

$$GPA(x; a, c) = 1 - (1 + cx/a)^{-1/c}$$

Fitting its 2 parameters on the data gives us automatically the domain of attraction. A curvature observation of the data plotted on Gumbel probability paper can also be used:

Convex curve	H is Weibull
Linear curve	H is Gumbel
Concave curve	H is Frechet

If we have concluded for a Weibull for maxima domain of attraction:

$$H(x) = \exp(-(-x)^{-c}) \quad x \leq 0, c < 0$$

Then, with $\lim_{n \rightarrow \infty} F^n(a_n + b_n x) = H(x)$, we can derive:

$$\begin{aligned} F(x) &= H^{1/n}((x - a_n)/b_n) \\ &= \exp(-((\lambda - x)/\delta)^\beta) \quad x \leq \lambda, \beta > 0 \end{aligned}$$

the general form of a Weibull for maxima distribution in which λ , δ and β are the unknown parameters, to be fitted to the data. λ is the location parameter, δ is the scale parameter and β is the shape parameter. Note that the maximum value that the random variable can adopt is given by λ .

Caution: there is also another type of Weibull distribution known in practice:

$$F(x) = \exp(-((x - \lambda)/\delta)^\beta), \quad x \leq \lambda, \beta > 0$$

This is the type which was suggested by Weibull [10]. This is in fact the limiting distribution for minima, but is nowadays used a lot for fitting maximum data as well. It is also commonly used as a lifetime distribution and in corrosion engineering [11].

The problem concerning the statistical treatment of extreme values is intrinsically a non-parametric statistical problem, however, based on tools currently available; it may only be dealt with parametrically. Thus, a problem specific to distribution tails is transformed into an extrapolation problem based on the hypothesis that a model selected based on analysis in a maximum probability region is valid in “extreme” regions.

References 7 and 16 propose statistical tests to establish whether or not an extreme quantile may be estimated based on an accepted distribution in the maximum probability domain, through an adequacy test, and especially the ET (Exponential Tail) test, that is adapted for probability distributions exhibiting exponential decay in the Gumbel domain, for example normal, log-normal, exponential, Weibull distributions. The reader is referred to Refs 7 and 16 concerning this theme for a description of this test. Another approach, referred to as the Bayesian regularisation procedure, was proposed in Ref. 7 to try and improve a probabilistic model that was accepted in the maximum probability region but was rejected in the extreme section by the ET test. The gain in the distribution tail assumes that a degradation in the region of maximum probability is acceptable. A compromise must be reached between minimum degradation in this region and maximum gain in the distribution tail. It can be noted that mixing or rejoining methods may also be of interest in treating distribution tails. The methods are recommended for further exploration.

6 Application

This section describes the results of a so-called Dutch flood frequency analysis on a data set of the American River (1 day peak flow (in cfs) from 1905 to 1997 and 3-day flows which go from 1905 to 2006). In a Dutch flood frequency analysis (according to the Boertien Committee, 1995 [17]) an analysis is made where four distribution functions (Gumbel, Pearson III (log shifted Gamma), Lognormal and Generalised Pareto) are fitted to peaks over threshold or annual maxima data sets. Due to the large uncertainties involved in the extrapolation, as well as all other kind of uncertainties (in climate change, river basin development, land use, etc.), Dutch authorities in water management tend to prefer a simple distribution to describe the low probabilities of extreme flood events. The simplest form and yet good correspondence to

an extreme value distribution is the exponential curve, which is a straight line on a log-scale. Therefore, the four fitted distributions are averaged and replaced by an as-good-as-possible approximation of a piecewise exponential distribution, which is then used for a flood frequency analysis (FFA) of the discharges. Boertien’s FFA is adopted by the politics and nowadays used to determine the design discharges on Dutch rivers.

Results of the fitting procedures on the American River are presented in Fig. 3. Four distribution functions are fitted to the data. The goodness of fit of three distribution functions is high (Pearson III, Pareto and Lognormal), but the goodness of fit of the Gumbel is rather poor. In the next step the average of the four distribution functions is calculated (each distribution with a weight factor of 0.25). It can be observed for that a strong curvature in the distribution occurs at 140,000 cfs. In the last step of the flood frequency analysis, the average distribution is described by a piecewise distribution consisting of exponential distributions.

The exact formulations of the above piecewise exponential distributions are given by:

For the 1-day floods:

$$F_Q(q) = 1 - \exp\left(-\frac{q}{48399}\right) \quad \text{for } q < 190000 \text{ cfs}$$

$$F_Q(q) = 0.1496 \left(\exp\left(-\frac{q}{93631}\right)\right) \quad \text{for } q > 190000 \text{ cfs}$$

For the 3-day floods:

$$F_Q(q) = 1 - \exp\left(-\frac{q}{34379}\right) \quad \text{for } q < 140000 \text{ cfs}$$

$$F_Q(q) = 0.1675 \left(\exp\left(-\frac{q}{61155}\right)\right) \quad \text{for } q > 140000 \text{ cfs}$$

These expressions can be inverted analytically to calculate the quantile levels corresponding to 10, 50, 100, 200, 500, 1000, 2000 and 10000 years.

The quantile levels corresponding to 10 and 50 can be calculated from the first part of the exponential distribution:

$$1 \text{ day: } Q_{10} = -48399 \ln(1/10) = 111440 \text{ cfs}$$

$$Q_{50} = -48399 \ln(1/50) = 189340 \text{ cfs}$$

$$3 \text{ day: } Q_{10} = -34379 \ln(1/10) = 79161 \text{ cfs}$$

$$Q_{50} = -34379 \ln(1/50) = 134490 \text{ cfs}$$

The quantile levels corresponding to the other return periods can be calculated from the second part of the exponential distribution:

$$1 \text{ day: } Q_{100} = -93631 \ln((1/100)/0.1496) = 253310 \text{ cfs}$$

$$Q_{200} = -93631 \ln((1/200)/0.1496) = 318210 \text{ cfs}$$

$$Q_{500} = -93631 \ln((1/500)/0.1496) = 404000 \text{ cfs}$$

$$Q_{1000} = -93631 \ln((1/1000)/0.1496) = 468900 \text{ cfs}$$

$$Q_{2000} = -93631 \ln((1/2000)/0.1496) = 533800 \text{ cfs}$$

$$Q_{10000} = -93631 \ln((1/10000)/0.1496) = 684490 \text{ cfs}$$

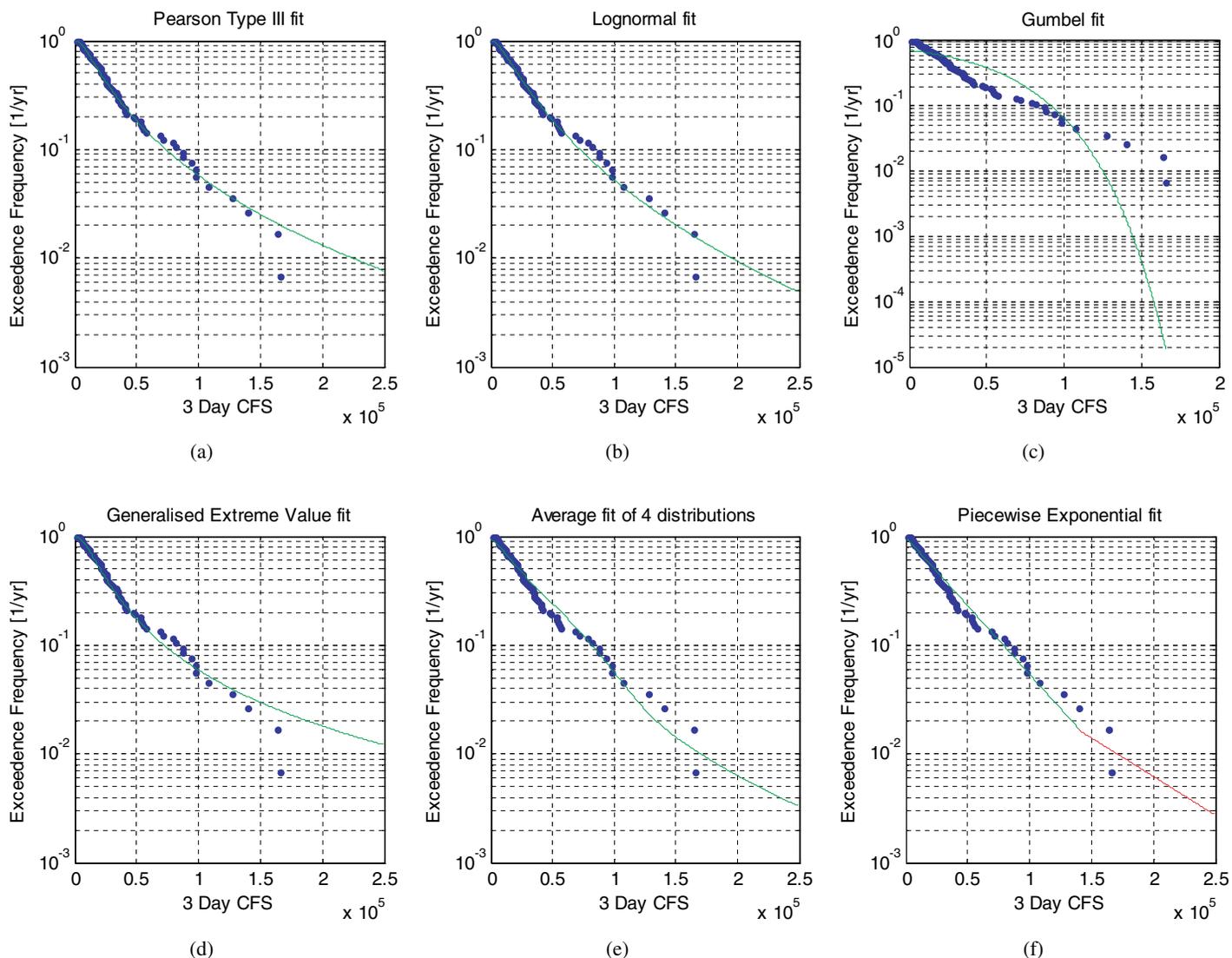


Figure 3 Distribution fitting to extreme river discharges of the American River.

$$\begin{aligned}
 3 \text{ day: } Q_{100} &= -61155 \ln((1/100)/0.1675) = 172360 \text{ cfs} \\
 Q_{200} &= -61155 \ln((1/200)/0.1496) = 214750 \text{ cfs} \\
 Q_{500} &= -61155 \ln((1/500)/0.1496) = 270780 \text{ cfs} \\
 Q_{1000} &= -61155 \ln((1/1000)/0.1496) = 313170 \text{ cfs} \\
 Q_{2000} &= -61155 \ln((1/2000)/0.1496) = 355560 \text{ cfs} \\
 Q_{10000} &= -61155 \ln((1/10000)/0.1496) = 453990 \text{ cfs}
 \end{aligned}$$

There is a large uncertainty in the extrapolation of the distribution functions. With a bootstrap analysis such uncertainty can be easily quantified, as shown in Fig. 4

Using the average distribution of the above bootstrapped lines the following extrapolated N-year flood’s one day – and three day discharges in cubic feet per second (cfs) have been calculated in Table 5.

7 Conclusions

One of the most significant criticisms concerning the application of probabilistic methods in the domain of flood risk analysis, as well as in other domains, concerns validation of the probability distribution used to model the various sources of uncertainty.

This paper presents various aspects of this issue (criteria for selecting a distribution, extreme laws, distribution tails, sample size, etc.) and presents promising practical leads to provide solutions. The optimal selection of a probability distribution involves “objective” and “subjective” parts, in which the objective parts are clarified in the paper with convergence theorems and relationships between distributions and the subjective parts are illustrated with the so called “Dutch approach”. The Dutch approach, according to the Boertien Committee, is presented in which four distribution types are fitted to Peaks-Over-Threshold data. A piecewise exponential fit is used as an average distribution. The approach is applied to data of the American River. The uncertainty of the 10^{-3} and 10^{-4} river discharge quantiles is extremely large. The knowledge of hydrological – and runoff processes might help to decrease these large uncertainties, although new uncertainties would be introduced (model uncertainties). Further experiments are necessary to validate some results, notably to allow a comparison with mixed physical-based statistical methods, in order to develop a methodology suitable for engineering- and environmental applications and industrial requirements.

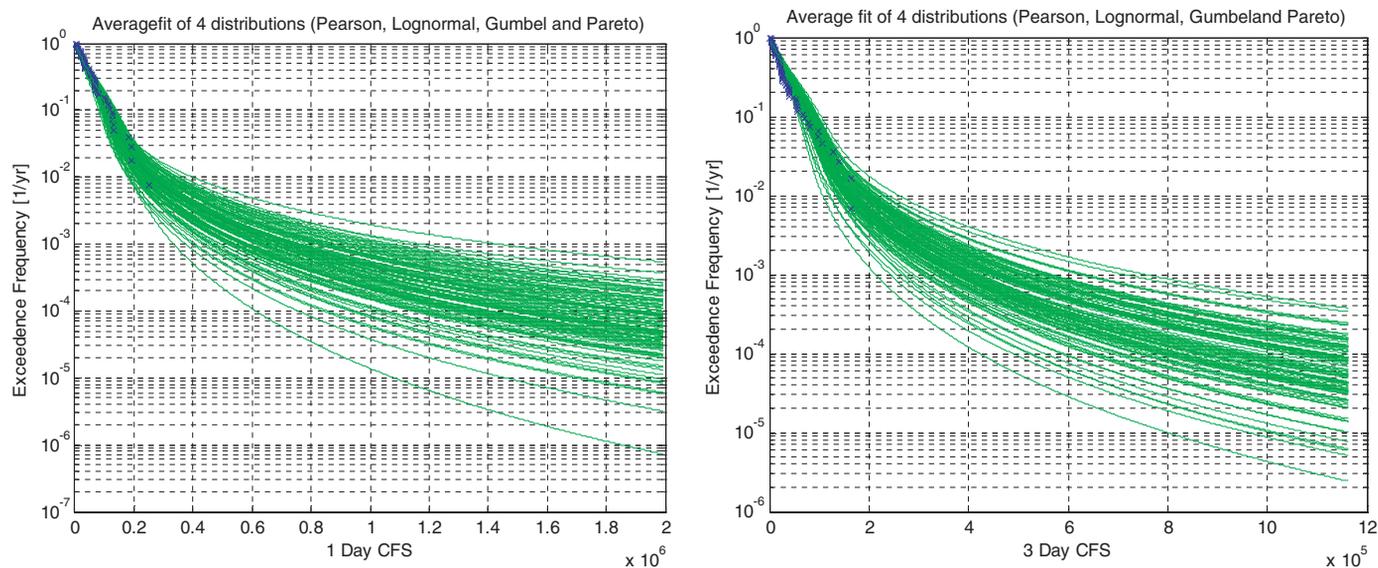


Figure 4 Bootstrap results for extreme river discharges.

Table 5 Discharge Quantiles (cfs) of N-year floods (one day and three days).

		1-Day flood						
10	50	100	200	500	1000	2000	10000	
113700	198600	263800	355600	529000	712400	956500	1882400	
		3-Day flood						
10	50	100	200	500	1000	2000	10000	
81900	137900	177600	232700	333700	437700	572500	1057800	

Note that these quantile estimates are much larger than the earlier exponential fits.

Acknowledgments

The authors would like to extend their sincere thanks to the following persons for discussions concerning this theme: Nicolas Devictor and Michel Marques (CEA, Cadarache), Joseph Countryman (MBK Engineers), Gary Estes, for providing the data of the American River, E. Meister and E. Ardillon (Electricité de France), J. Héliot (Framatome), J. Diebolt (Marne-la-Vallée University), M. Garrido (INRIA Rhône-Alpes), M. Baudot and Kadiri (Ligeron SA), M-P. Valeta, M. Eid and P. Lamagnère, S. Boulègue (CEA). Furthermore, we are extremely grateful to the EU-FLOODsite project (task 2 and 7) under contract no. GOCE-CT-2004-505420 and both EDF and Framatome for the financial support of part of this work.

References

1. SHANNON, C.E. (1948). “A Mathematical Theory of Communication,” *Bell Systems Technical Journal*, 27, 379–423 and 623–656.
2. D’AGOSTINO, R.B. and STEPHENS, M.A. (1986). *Goodness of Fit Techniques*, Dekker.
3. ROBERT, C. (2005). *The Bayesian Choice*, Springer Texts in Statistics.
4. CHEVALIER, M., PEREZ, D., BIASSE, J.-M. and MITTERAND, J.-M. (1997). ”Resampling and Simulation on Small-Size

- Samples,” in GUEDES SOARES, C. (ed.), *Advances in Safety & Reliability*, Pergamon.
5. DEVICTOR, N., CHAPULIOT, S. and EID, M. (1999). “Monte-Carlo Simulation to Assess Structures Reliability,” in *Proceedings of Esrel’99*, Edinburgh.
6. PICKANDS, J. (1975). “Statistical Inference using Extreme Order Statistics,” *The Annals of Statistics*, 3, 119–131.
7. DIEBOLT, J., GARRIDO, M. and TROTTIER, C. (2001). *A Bayesian Regularization Procedure for a Better Extremal Fit*, INRIA Research Report.
8. GNEDENKO, B. (1943). “Sur la Distribution limite du terme maximum d’une série aléatoire,” *Annals of Mathematics, Second Series*, 44, 423–453.
9. CASTILLO, E. (1988). *Extreme Value Theory in Engineering*, Academic Press Inc.
10. WEIBULL, W. (1961). *Fatigue Testing and Analysis of Results*, Pergamon, London.
11. SCARF, P.A. and LAYCOCK, P.J. (1996). “Estimation of Extremes in Corrosion Engineering,” *Journal of Applied Statistics*, 23(6), 621–643.
12. LINDLEY, D.V. (1983). “Reconciliation of Probability Distributions,” *Operations Research*, 31, 866–880
13. PERREAULT, L., BOBÉE, B. and RASMUSSEN, P.F. (1999). “Halphen Distribution System. I: Mathematical and Statistical Properties,” *Journal of Hydrologic Engineering*, 4(3), 189–196.

14. PERREAULT, L., BOBÉE, B. and RASMUSSEN, P.F. (1999). “Halphen Distribution System. II: Parameter and Quantile Estimation,” *Journal of Hydrologic Engineering*, 4(3), 200–208.
15. VAN GELDER, P.H.A.J.M. (2000). “Statistical Methods for the Risk-Based Design of Civil Structures,” in *Communications on Hydraulic and Geotechnical Engineering*, TU Delft.
16. GARRIDO, M. (2002). “TU-0753 - Modélisation des événements rares et estimation des quantiles extrêmes, méthodes de sélection de modèles pour les queues de distribution,” *Thèse de : Grenoble 1*, 231, 12 pp., juin 2002 - Document en français; Titre anglais : “Modelling of Rare Events and Estimation of Extreme Quantiles, Model Selecting Methods for Distribution Tails”.
17. Design Discharges on Dutch Rivers (in Dutch) (2002) Report from the Ministry of Water Management.
The Boertien Committee Approach to Extrapolation of River Discharges, Downloadable from: www.rijkswaterstaat.nl/rws/riza/home/publicaties/rapporten/2002/rr_2002_014.pdf
18. ALEXANDER, M.R. BAKKER, R. and WIM, M.J. (2005). Luxemburg, Heterogeneous Distributions within Flood Frequency Analysis, ISSH – Stochastic Hydraulics 2005, 23 and 24 May 2005, Nijmegen, The Netherlands.
19. Statistical Software Packages: SAS, Matlab, SPSS, Stargraphics, Splus, Statistica, Backgrounds of these Packages can be Conveniently Found via Internet Search Engines.
20. ANDERSON, T.W. and DARLING, D. A. (1952). “Asymptotic Theory of Certain ‘Goodness-of-Fit’ Criteria Based on Stochastic Processes,” *Annals of Mathematical Statistics*, 23, 193–212.
21. STEPHENS, M.A. (1974). “EDF Statistics for Goodness of Fit and Some Comparisons,” *Journal of the American Statistical Association*, 69, 730–737.