A priori restrictions on the parameters of the generalized extreme value distribution for annual maxima of river discharge

Ronald van Nooijen*
Delft University of Technology, Delft, Netherlands - r.r.p.vannooyen@tudelft.nl

Alla Kolechkina
Delft University of Technology, Delft, Netherlands - r.r.p.vannooyen@tudelft.nl

Abstract

In water resource management the question of the probability of a future observation exceeding a certain value (flood level, discharge, precipitation intensity) occurs regularly. This question involves both hydrology and statistics. In hydrology many distributions are used ranging from log-normal and log-gamma to the traditional extreme value distributions. Practical arguments are given for disallowing certain parameter vectors when fitting the Generalized Extreme Value (GEV) distribution to river discharges. These restrictions may be used in three ways. They can be used after a distribution has been obtained to decide on its validity. If they can be associated with a probability then they may be used to restrict the fitting process with a known probability that the actual outcome conflicts with the restrictions. Finally in a Bayesian approach they can be used to obtain a prior parameter distribution density with bounded support for the Generalized Extreme Value (GEV) distribution.

Keywords: Statistics of extreme values; Hydrology; Bayesian inference.

1. Introduction

Hydrologists are under considerable pressure to provide predictions of extreme system behaviour based on extrapolation. For strong opinions on the results of this pressure please consult Klemeš (2000a,b). The problem lies in the relatively short data series of available data of 30 to 100 years and the interest of governments in floods with probability of occurrence of 0.01 or lower. For instance the Dutch government requires information on events with a probability of occurrence of 1×10^{-4} when considering the safety of the Western part of the country. Many approaches and distributions have been used, see for example Bobée and Rasmussen (1995).

This paper provides a series of assumptions that when taken together provide an internally consistent framework to function as prior knowledge for a Bayesian analysis of series of annual maximum discharges. A general discussion of Bayesian parameter estimation and more information on the Bayesian approach to extremes can be found in Coles and Powell (1996). After a short introduction to the basic mathematical model for extreme value analysis we state our assumptions and some of their consequences.

2. Choice of distribution family

According to Kirby and Moss (1987) at that time a report by the Hydrology Subcommittee (1982) of the U.S. Inter-agency Advisory Committee on Water Data prescribed: "The basic provision is that a Pearson Type III distribution be fitted to the logarithms of the annual peak flow magnitudes by use of the (logarithmic) sample mean and standard deviation and a weighted skew coefficient (method of moments)." See also Stedinger and Griffis (2008). An overview by Cunnane (1989) shows that the USA was not the only government that did not use an extreme value distribution to fit annual maxima. Canada is somewhat more open mind, see Watt (1989, page 50).

Ideally a distribution for the outcome of an experiment should be chosen based on a solid physical and statistical analysis. For cases where one tries to approximate an empirical distribution within the range of a sample, relaxing this requirement may not be harmful. However, for the case of high river discharges where

extrapolation to events outside, perhaps far outside the range of observations, this does not seem wise. This leaves two options:

- a derivation of a distribution based on physics of the distribution for a flood peak;
- a distribution based on the theory of extreme values, that is either the GEV distribution for series of maxima or the Generalized Pareto Distribution for peaks over a threshold.

As stated most recently in Naghettini (2017, page 154) "the extremal distributions are the only group of probability distributions that offers theoretical arguments that justify their use in the modelling of hydrological maxima ..." However, even recent publications in the field still allow for other distributions.
In this paper the choice is to consider only the GEV distribution. Its cumulative distribution function is given by

$$G_\beta(t) = \begin{cases} \exp(-\exp(-t)) & \beta = 0 \\ \begin{cases} 0 & : t \leq -\frac{1}{\beta} \\ \exp\left(-(1+\beta t)^{-1/\beta}\right) & : t > -\frac{1}{\beta} \end{cases} & \beta > 0 \\ \begin{cases} \exp\left(-(1+\beta t)^{-1/\beta}\right) & : t < \frac{1}{|\beta|} \\ 1 & : t \geq \frac{1}{|\beta|} \end{cases} & \beta < 0 \end{cases} \tag{1}$$

where $\beta = 0$ corresponds to a type I or Gumbel-type extreme value distribution, $\beta > 0$ corresponds to a type II or Fréchet-type extreme value and $\beta < 0$ corresponds to a type III or (reverse) Weibull-type extreme value distribution. The limit distribution H to be used for the annual maxima will have the form

$$H(x) = G_\beta\left(\frac{x-\xi}{\zeta}\right)$$

with quantile function

$$0 < p < 1 \,:\, H^{\leftarrow}(p) = \begin{cases} \xi - \zeta \log\left(\log\frac{1}{p}\right) & : \beta = 0 \\ \xi - \zeta\frac{1-\left(\log\frac{1}{p}\right)^{\beta}}{\beta} & : \beta \neq 0 \end{cases}$$

3. Arguments for restricting the type of extreme value distribution under consideration

In Koutsoyiannis (2004, 2005) it is argued that, for extreme precipitation at least, the type II or Fréchet-type distributions should be used. In Koutsoyiannis (2007, section 6) arguments are given that the type II or Fréchet-type should be used for extreme discharges as well.
An argument for type I or type III could be made by considering that distributions with a finite upper bound on their support cannot be of type II, because of the domain of attraction of type II does not contain such distributions, see Falk et al. (2011, Theorem 2.1.1). However, Koutsoyiannis (2004, 2005, 2007) provide counter arguments.
Finally one could perhaps argue that, in a subjective Bayesian context, the willingness to extrapolate from historical data indicates a high level of certainty that no floods will occur that permanently alter the upstream catchment in a way that renders the extrapolation meaningless. This would mean that we are sure that the sample ω that we take from the probability space Ω is such that the corresponding values $X_j(\omega)$ lie in a subset corresponding to floods below a certain threshold. This would argue against type II. In this paper we put no restrictions on the type of extreme value distribution.

4. A priori assumptions

From a hydrological point of view not all parameter vectors (β, ξ, ζ) lead to quantiles that make sense. It would seem prudent to formulate assumptions that, when violated, constitute a reason to assume the results are invalid. We treat the cases $\beta < 0$, $\beta = 0$ and $\beta > 0$ separately. We start by stating our basic assumptions. We will use the following notation:

- F is the distribution of the separate values over which we take the maximum;

- n is the length in years of our sample, that is the series of measurements of annual maxima;
- X_j is the random variable representing the j-th yearly maximum;
- $s_{\min}$ is the sample mininum and $s_{\max}$ is the sample maximum.

Assumption 1. *Standard Extreme Value theory (EV-theory) applies.*

This assumption is very difficult to test. For instance, a goodness of fit test with a 0.05 significance level should reject 1 out every 20 data sets for the distribution we are testing for. If we have a large number of data sets, this makes rejection of some of them almost inevitable. On the other hand, acceptance of a data set by the test says nothing about whether or not the sample is actually from the distribution, as significance levels are intended to protect us from rejecting a sensible null hypothesis on the basis of insufficient evidence, not to prevent us from accepting a nonsensical alternative hypothesis. In the case of floods we are almost always extrapolating to predict values not yet observed. This means that good prior knowledge is needed to support the choice of distribution, as the choice of distribution will have a large effect on the outcome. However, barring new developments that provide theoretical or physical evidence in support of distributions outside the extreme value family (or the Generalized Pareto for peaks over a threshold), assumption 1 seems the only one supported by theory. While large scale studies claim good fits for several other distributions, it seems to us that a good fit on realtively short observation series cannot be considered evidence of the applicability of these distributions to floods with recurrence periods exceeding the length of the period of observations.

Assumption 2. *We assume that* $H^{\leftarrow}\left(\frac{1}{4}\right)$ *and* $H^{\leftarrow}\left(\frac{3}{4}\right)$ *lie within the range of observed maxima.*

If Assumption 1 holds and we have a reasonable number of observations, it is highly likely that this holds. Even if we disregard that a hydrologist would spot signs of the violation of this assumption in the landscape, the probability of taking a sample of size n that contains no points in the lower or upper quartile for an arbitrary distribution is given by

$$\Pr\left(s_{\min} \geq H^{\leftarrow}\left(\frac{1}{4}\right) \text{ or } s_{\max} \leq H^{\leftarrow}\left(\frac{3}{4}\right)\right) = 2 \times \left(\frac{3}{4}\right)^n - \left(\frac{1}{2}\right)^n \tag{2}$$

which decreases rapidly with sample size n, for $n = 50$ it is 1.1×10^{-6}. This assumption implies that $H^{\leftarrow}\left(\exp\left(-1\right)\right)$ lies within the range of observed maxima so

$$s_{\min} \leq \xi \leq s_{\max} \tag{3}$$

Assumption 3. *The yearly discharges vary significantly on time scales relevant to the application.*

If this assumption does not hold then one would expect to a hydrologist to find signs of this in the landscape. Assigning a precise probability to mistakenly assuming this is difficult.

5. Consequences of the assumptions for Type I parameters

This corresponds to $\beta = 0$. Assumption 2 implies that $H^{\leftarrow}\left(p\right) \geq 0$ because flows are non-negative so

$$\xi - \zeta \log\left(\log 4\right) \geq 0 \tag{4}$$

in other words $\xi \geq \log\left(\log 4\right)\zeta$ or approximately $\xi \geq 0.32\zeta$. We translate Assumption 3 into

$$\frac{H^{\leftarrow}\left(\frac{3}{4}\right) - H^{\leftarrow}\left(\frac{1}{4}\right)}{H^{\leftarrow}\left(\exp\left(-1\right)\right)} \geq \lambda$$

where determination of the value λ to be used on the right hand side would depend on the situation, a reasonably conservative choice would be 1/10, we get

$$\zeta\left(\log\left(\log 4\right) - \log\left(\log 4 - \log 3\right)\right) \geq \lambda\xi \tag{5}$$

or, for $\lambda \approx 1/10$, $\xi \leq 15.7\zeta$. We now have bounds on ζ in terms of ξ

$$\frac{\lambda}{\log(\log 4) - \log(\log 4 - \log 3)}\xi \leq \zeta \leq \frac{1}{\log(\log 4)}\xi \tag{6}$$

From assumption 2 we get (3), so we have finite bounds on the support of the a-priori parameter distribution.

6. Consequences of the assumptions for Type II parameters

This corresponds to $\beta > 0$. Assumption 2 implies that $H^{\leftarrow}\left(\frac{1}{4}\right) \geq 0$ because flows are non-negative, so for $\beta > 0$ we find

$$\xi \geq \zeta \frac{1 - (\log 4)^{-\beta}}{\beta} \tag{7}$$

We translate Assumption 3 into

$$\frac{H^{\leftarrow}\left(\frac{3}{4}\right) - H^{\leftarrow}\left(\frac{1}{4}\right)}{H^{\leftarrow}(\exp(-1))} \geq \lambda$$

where determination of the exact value to be used for λ would depend on the situation. This results in

$$\zeta \geq \lambda \frac{\beta(\log 4 - \log 3)^{\beta}}{1 - \left(\frac{\log 4 - \log 3}{\log 4}\right)^{\beta}}\xi \tag{8}$$

From assumption 2 we get

$$s_{\min} \leq \xi \leq s_{\max}$$

and from (7) and (8)

$$\lambda \frac{\beta(\log 4 - \log 3)^{\beta}}{1 - \left(\frac{\log 4 - \log 3}{\log 4}\right)^{\beta}}\xi \leq \zeta \leq \frac{\beta}{1 - \left(\frac{1}{\log 4}\right)^{\beta}}\xi$$

As the tail increases in thickness for increasing β the techniques we use to find a bound on β with type I and III will not work here. There is however such a thing as a tail that is too thick. Especially when we take into account the probability of large (10% and up) errors in measurements for large flows. If we have n measurements and we wish to determine a $H^{\leftarrow}(1/m)$ with $m > n$ and the distance from $H^{\leftarrow}(\exp(-1))$ to $H^{\leftarrow}(1/n)$ is less than ϵ (with for example $\epsilon = 0.01$) of the distance $H^{\leftarrow}(\exp(-1))$ to $H^{\leftarrow}(1/m)$ we might want to re-examine the result.

In other words we would like to impose a condition of the form

$$\frac{H^{\leftarrow}(1 - 1/n) - H^{\leftarrow}(\exp(-1))}{H^{\leftarrow}(1 - 1/m) - H^{\leftarrow}(\exp(-1))} \geq \epsilon$$

which reduces to

$$\frac{\left(\log\left(1 + \frac{1}{n}\right)\right)^{-\beta} - 1}{\left(\log\left(1 + \frac{1}{m}\right)\right)^{-\beta} - 1} \geq \epsilon$$

which for fixed n and m is a decreasing function of β and will provide an upper bound for β.

7. Consequences of the assumptions for Type III parameters

This corresponds to $\beta < 0$. Assumption 2 implies that $H^{\leftarrow}\left(\frac{1}{4}\right) \geq 0$ because flows are non-negative so for $\beta < 0$ we find

$$\xi \geq \zeta \frac{(\log 4)^{-\beta} - 1}{(-\beta)} \geq 0.32\zeta \tag{9}$$

We translate Assumption 3 into

$$\frac{H^{\leftarrow}(\exp(-\exp(-1))) - H^{\leftarrow}(\exp(-1))}{H^{\leftarrow}(1) - H^{\leftarrow}(\exp(-1))} \leq \frac{99}{100}$$

so

$$\frac{H^{\leftarrow}(1) - H^{\leftarrow}(\exp(-\exp(-1)))}{H^{\leftarrow}(1) - H^{\leftarrow}(\exp(-1))} \geq \frac{1}{100}$$

so

$$\exp(\beta) \geq \frac{1}{100} \tag{10}$$

which suggests that $\beta > -4.7$ which provides a lower bound on β. We would also expect that for $\beta < 0$

$$\frac{H^{\leftarrow}(1) - H^{\leftarrow}\left(\exp\left(-(1+\beta)^{-1/\beta}\right)\right)}{H^{\leftarrow}(\exp(-1))} \geq \frac{H^{\leftarrow}(1) - H^{\leftarrow}(\exp(-1))}{H^{\leftarrow}(\exp(-1))} \geq \frac{1}{100}$$

so

$$\zeta \geq \frac{1}{100}\xi \tag{11}$$

From assumption 2 we get

$$s_{\min} \leq \xi \leq s_{\max}$$

and from (11) and (9) we get

$$\frac{1}{100}\xi \leq \zeta \leq \frac{(-\beta)}{(\log 4)^{-\beta} - 1}\xi \leq \frac{1}{\log(\log 4)}\xi$$

and from (10) and the fact that we are dealing with $\beta < 0$

$$-\log(100) \leq \beta < 0$$

We now have finite bounds on the support of the a-priori distribution.

8. Conclusions

Based on practical considerations the range of parameters to be accepted as realistic when fitting the GEV to maximum annual discharges for rivers can be reduced substantially. This can be applied in two ways. Either we can use the assumptions to restrict the support of the a priori parameter distribution for a Bayesian approach or we can use the derived bounds on the parameters as a sanity check on parameters obtained from a fit with an uninformative prior or a frequentist calculation that did not use these constraints.

References

Bobée, B. and Rasmussen, P. F. (1995). Recent advances in flood frequency analysis. *Rev. Geophys.*, 33(S2): 1111–1116.

Coles, S. G. and Powell, E. A. (1996). Bayesian methods in extreme value modelling: A review and new developments. *Int. Stat. Rev.*, 64(1):119–136.

Cunnane, C. (1989). *Statistical Distributions for Flood Frequency Analysis*, volume 33 of *WMO Operational Hydrology Report*. Secretariat of the World Meteorological Organisation.

Falk, M. , Hüsler, J. , and Reiss, R.-D. (2011). *Laws of Small Numbers: Extremes and Rare Events*. Springer Science + Business Media, third edition.

Hydrology Subcommittee (1982). Guidelines for determining flood frequency. Technical report, U.S. Interagency Advisory Committee on Water Data. Bulletin 17-B (revised and corrected).

Kirby, W. and Moss, M. (1987). Summary of flood-frequency analysis in the United States. *J. Hydrol.*, 96 (1-4):5–14.

Klemeš, V. (2000a). Tall tales about tails of hydrological distributions. I. *J. Hydrol. Eng.*, 5(3):227–231.

Klemeš, V. (2000b). Tall tales about tails of hydrological distributions. II. *J. Hydrol. Eng.*, 5(3):232–239.

Koutsoyiannis, D. (2004). Statistics of extremes and estimation of extreme rainfall: I. theoretical investigation / statistiques de valeurs extrêmes et estimation de précipitations extrêmes: I. recherche théorique. *Hydrolog. Sci. J.*, 49(4):575-590.

Koutsoyiannis, D. (2005). Uncertainty, entropy, scaling and hydrological stochastics. 1. marginal distributional properties of hydrological processes and state scaling. *Hydrolog. Sci. J.*, 50(3):381–404.

Koutsoyiannis, D. (2007). A critical review of probability of extreme rainfall: principles and models. *Advances in Urban Flood Management*, 139–166.

Naghettini, M. , editor (2017). *Fundamentals of Statistical Hydrology*. Springer Nature.

Stedinger, J. R. and Griffis, V. W. (2008). Flood frequency analysis in the United States: Time to update. *J. Hydrol. Eng.*, 13(4):199–204.

Watt, W. E. , editor (1989). *Hydrology of floods in Canada - a guide to planning and design*. National Research Council Canada, Associate Committee on Hydrology.