

Diagnostic-based peaks-over-threshold flood frequency analysis of the Erzen River, Albania

Santino Spahiu^{1*} , Enkelejda Gjinali² , Andrin Kërpaçi³ 

^{1,3}Department of Hydraulics and Hydrotechnics, Polytechnic University of Tirana, Rruga Muhamet Gjollështa 54, 1023, Tirana, Albania

² Department of Environmental Engineering, Polytechnic University of Tirana, Rruga Muhamet Gjollështa 54, 1023, Tirana, Albania

* Corresponding author's e-mail: santino.spahiu@fn.edu.al

ABSTRACT

Reliable estimates of flood quantiles are a prerequisite for the safe design of hydraulic structures and for flood risk management. Traditionally, they are derived from annual maximum series that retain one event per year, discarding secondary peaks and small maxima from dry years. This study estimated the frequency of maximum discharges of the Erzen River at Ndroq station using a partial duration series built with the peaks-over-threshold approach, and compared its results with the conventional annual maximum series analysis. The daily discharge record (1979–2023) was screened over 100 candidate thresholds, and the threshold was selected by reading six diagnostic plots. Three are classical: the number of exceedances, the Generalised Pareto shape parameter and the mean excess plot. Three are rarely combined: the threshold-wise Anderson-Darling test of the Generalised Pareto fit, its parametric-bootstrap p -value and the dispersion index test of the Poisson occurrence model. Independence of the retained peaks was enforced by the Bulletin 17B inter-event time criterion (11 days) together with the Flood Estimation Handbook recession criterion (discharge receding below two-thirds of the preceding peak), and the sensitivity of the results to the time separation window was quantified. At the selected threshold of 445 m³/s the series comprises 34 independent events (0.76 events per year) whose magnitudes follow a Generalised Pareto distribution (shape $\kappa = -0.276$) fitted by L-moments, while the annual event counts are consistent with a Poisson process (dispersion index 0.91). Annual exceedance probabilities were obtained by the empirical Langbein transformation and by the analytical conversion of the fitted Pareto–Poisson model into an equivalent Generalised Extreme Value distribution. The 90% confidence intervals were computed by parametric bootstrap. The partial duration series yields a 1% annual exceedance probability discharge of 1114 m³/s, approximately 12% higher than the 996 m³/s obtained from the annual maximum series. The difference grows to about 51% at the 1000-year return period. Half of the annual maximum series members lie below the threshold, which produces a near-Gumbel upper tail that the threshold-censored sample contradicts. The peaks-over-threshold approach provides a more representative description of flood behaviour in the Erzen River and a more conservative basis for the design of hydraulic works and flood mitigation planning.

Keywords: flood frequency analysis, peaks over threshold, partial duration series, generalised Pareto distribution, threshold selection, Poisson process, annualisation.

INTRODUCTION

Accurate estimation of flood magnitudes of different return periods has major implications for the design of water infrastructure and for the estimation of flood risk (Salas et al., 2018). Both the

safety and the financial cost of these structures and water systems depend on it. In the frequentist approach, the conventional basis for estimating flood frequency is the annual maximum series (AMS). For each year of the record, the largest discharge is retained, and the resulting sample is fitted to an

extreme value distribution. In most of Europe this is usually the Generalised Extreme Value (GEV) distribution; under the United States federal guidance of Bulletin 17C it is the log-Pearson type III distribution (England et al., 2019).

The AMS approach has the advantage of being straightforward. The sample is easy to form: its size equals the number of years, and the annual spacing of the values makes their statistical independence almost automatic. It has, however, two weaknesses that have been recognised for decades. The first concerns what the AMS includes. In a dry year, the largest discharge is retained no matter how small it is, even when it is an ordinary in-bank flow that no engineer would call a flood. Such “opportunistic” maxima carry no information about flood behaviour, yet they enter the fitted distribution with the same weight as the major events. As they accumulate over the years, they drag the fitted curve towards them and distort its upper tail. United States guidance acknowledges the problem explicitly and prescribes a statistical test to detect and censor these potentially influential low floods before fitting (England et al., 2019). Klemeš (1986) formulated the underlying objection memorably. It is by no means hydrologically obvious, he observed, why the regime of the highest floods should be affected by the regime of flows in years when no floods occur. The second weakness concerns what the AMS excludes. Because only one value per year survives, the second- and third-largest floods of a wet year are discarded even when they exceed the annual maxima of most other years. Information that is directly relevant to the frequency of large floods is simply thrown away (Bezak et al., 2014). These values are not the largest observed in their year, but they can still be large enough to cause flooding. Disregarding them influences the reality of the frequency of floods in a given location.

The peaks-over-threshold (POT) approach addresses both weaknesses by changing the sampling rule: instead of taking the largest event of each year, it takes every independent peak that exceeds a suitably chosen high threshold, however many or few of them a year may contain. The resulting sample is called the partial duration series (PDS). Small annual maxima of dry years fall below the threshold and never enter the sample, which removes the first weakness; multiple genuine floods of a wet year all enter the sample, which removes the second. The statistical analysis for the PDS is well established. The magnitudes

of the threshold exceedances are modelled with the Generalised Pareto (GP) distribution, a choice justified by an asymptotic theorem of extreme value theory (Nerantzaki and Papalexiou, 2022; Pickands, 1975). The number of exceedances per year is modelled as a Poisson process. The combination of the two, GP magnitudes arriving in Poisson counts, can be converted analytically into an equivalent GEV distribution for the annual maximum. This restores the familiar annual probability scale required for design (Kolaković et al., 2023; Stedinger et al., 1993). Comparative studies have repeatedly found the POT approach at least as accurate as, and frequently preferable to, the AMS approach, particularly for short records and for quantiles beyond the 100-year return period (Mostofi Zadeh et al., 2019). Bezak et al. (2014) reached the same conclusion for a Slovenian catchment. Karim et al. (2017) and Pan and Rahman (2022) reported the same for Australian catchments. At the scale of 842 British stations, Kjeldsen and Prosdociami (2023) showed that POT data provide a suitable foundation for national flood frequency studies, whereas the GEV distribution fitted to annual maxima was found unsuitable for a large share of the stations. A recent scoping review nevertheless notes that the POT method remains underused in operational practice, largely because of the perceived subjectivity of its two key decisions: the threshold and the independence criteria (Pan et al., 2022).

These two decisions deserve a closer look, because they are the reason the POT method is not yet a routine tool despite its statistical advantages. The threshold embodies a genuine trade-off. It must be high enough for the GP approximation to hold and for every retained value to be a flood. But every increase of the threshold discards data, and a threshold that is too high leaves a sample too small for stable estimation, so the fitted parameters acquire large variances (Langousis et al., 2016; Nerantzaki and Papalexiou, 2022). Standard practice explores this trade-off by computing a set of diagnostics over a range of candidate thresholds: the number of retained events, the stability of the estimated GP shape parameter, and the linearity of the mean residual life plot (Pan et al., 2022; Kjeldsen and Prosdociami, 2023). More recent work has supplemented these graphical tools with formal goodness-of-fit measures, notably Anderson-Darling type statistics computed threshold by threshold (Choulakian and Stephens, 2001; Solari et al., 2017).

The independence criteria, in turn, must guarantee that two retained peaks represent two separate flood episodes rather than two crests of a single multi-peaked hydrograph. Operational rules link the required separation to the catchment size and to the recession between the peaks.

A search of the literature returned no peaks-over-threshold flood frequency study for Albanian rivers. Design practice there has traditionally relied on annual maxima. The present study addresses this gap using the Erzen River at Ndroq as a case study. The objectives of this study are: (1) to construct a partial duration series from a 44.5-year daily discharge record by selecting the computational threshold through six diagnostic plots, including the threshold-wise Anderson-Darling test and the Poisson dispersion index, which extend the classical three-plot toolbox; (2) to enforce and verify the independence and identical-distribution assumptions through the Bulletin 17B time-separation criterion, the Flood Estimation Handbook recession criterion, a sensitivity analysis of the separation window, and an examination of the seasonality of the retained events; (3) to model the series with the Pareto–Poisson framework fitted by L-moments and to transform the results to the annual probability scale by both the empirical Langbein and the analytical Madsen procedures; and (4) to compare the resulting frequency curve, with parametric-bootstrap confidence intervals (CIs), against the conventional AMS/GEV analysis up to the 1000-year return period. The remainder of the paper describes the study area and the data, develops the methodology together with the governing equations and the reasoning behind each step, and then reports the results of the threshold diagnostics and the rationale for selecting the computational threshold used in the frequency analysis. These are discussed in the light of the international literature and of current engineering practice. The paper can also serve as a blueprint for similar studies in which the analyst chooses the POT approach.

METHODS

Study area and data

The Erzen River drains a catchment of approximately 760 km² in central Albania. It rises on the slopes of the Mali me Gropa massif east of Tirana, flows for about 109 km through hilly terrain

and the Ndroq gorge, and discharges into the Adriatic Sea at Lalzi Bay, north of Durrës (Pano, 2015; AMBU, 2023). The climate is Mediterranean, characterised by mild and wet winters and hot and dry summers. Precipitation falls mostly between October and March, when Atlantic and Mediterranean frontal systems cross the region. Snow plays only a marginal role, because most of the catchment lies below the elevations at which a seasonal snowpack accumulates. The flow regime follows the precipitation regime. The river responds quickly to rainfall. Floods occur almost exclusively in the autumn and winter months, and summer discharges recede to a small fraction of the annual mean. The gauging site considered in this study is Ndroq, located in the lower-middle course of the river, with a contributing drainage area of 663 km² (Fig. 1). The catchment upstream of the station was delineated from the Copernicus DEM GLO-30 (30 m; European Space Agency, 2024) by D8 flow routing, giving its contributing area of 663 km²; the gauge (station Ndroq, identifier H74M07) lies at 19.660° E, 41.265° N. The lower valley between Ndroq and the coast hosts intensive agriculture, gravel extraction and expanding peri-urban settlement, which heightens the exposure of the reach to flood damage.

The analysis is based on a continuous series of daily mean discharges measured at the Ndroq gauging station, covering the period 1 January 1979 – 30 June 2023, i.e. 44.5 years or 45 hydrological years. The record is continuous and gap-free, comprising 16,252 daily values with no missing and no duplicated dates over the period. The Ndroq gauge measures stage, which the Albanian Institute of Geoscience converts to discharge through a stage-discharge rating. Because the independence of flood events matters in the statistical analysis, the data are grouped into hydrological years (1 October to 30 September) and labelled by the calendar year in which they end. This approach is essential because a wet season that begins in November and ends in March would otherwise straddle two calendar years, and a single flood-rich winter could contribute two “annual” maxima to the sample.

Over the record, the mean daily discharge is 34.5 m³/s, the mean annual maximum is 451 m³/s, the lowest annual maximum is 149 m³/s, and the largest daily mean discharge is 1129 m³/s, recorded on 10 November 2016 (Table 1). The ratio of the largest observed flood to the mean daily flow, roughly a factor of thirty, illustrates the torrential

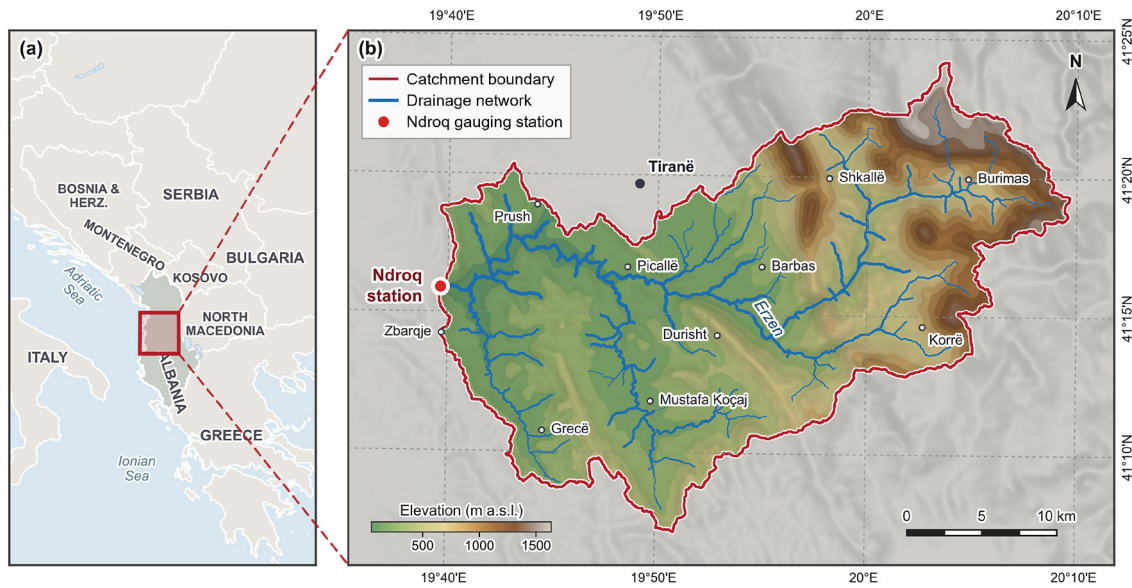


Figure 1. (a) Location of the Erzen River, Albania
(b) The catchment upstream of the Ndroq station and its drainage network

character of the river's flow regime. The second-largest event, 884 m³/s on 2 December 2017, corresponds to the widely reported floods of late 2017 (IFRC, 2018).

Extreme value theorems

There are two extreme value theorems (EVT) that are commonly used in modelling the frequency of flood events. In the first extreme value theorem (Fisher and Tippett, 1928; Gnedenko, 1943), the sample to be modelled is formed by what is called *block maxima*: the greatest peak flow in each year is selected, and subsequently an annual maximum series is formed. Based on the first extreme value theorem, this series converges to the Generalised Extreme Value family of distributions (Gumbel, 1958; Jenkinson, 1955). The cumulative distribution function of the GEV distribution is:

$$F(x) = \exp(-e^{-y}) \quad (1)$$

where: y is the reduced variate given by:

$$y = \begin{cases} -\kappa^{-1} \ln \left\{ 1 - \frac{\kappa (x - \xi)}{\alpha} \right\}, & \kappa \neq 0 \\ \frac{x - \xi}{\alpha}, & \kappa = 0 \end{cases} \quad (2)$$

and ξ is the location parameter, α ($\alpha > 0$) is the scale parameter and κ the shape parameter. Based

on the values the shape parameter takes, the GEV distribution corresponds to the Gumbel distribution for $\kappa = 0$, the Fréchet distribution for $\kappa < 0$ and the Weibull distribution for $\kappa > 0$ (Hosking and Wallis, 1997). In the Gumbel distribution, the tail decays exponentially. In the Fréchet distribution the tail decays as a power law, the distribution has no upper bound and rare quantiles grow rapidly with the return period, whereas the Weibull distribution has a finite upper bound. This study uses the L-moment sign convention of Hosking and Wallis (1997) throughout, in which the shape parameter κ carries the opposite sign to the ξ of the standard extreme-value parameterisation of Coles (2001); that is, $\kappa = -\xi$. Under this convention $\kappa < 0$ denotes the heavy-tailed Fréchet domain (unbounded above, power-law tail) and $\kappa > 0$ the bounded Weibull domain, and the convention is applied consistently in every equation and every numerical value below. Accordingly, the fitted flood tail $\kappa = -0.276$ is identical to $\xi = +0.276$ in the Coles notation: both describe an unbounded, heavy tail, and neither implies a finite upper bound.

The second extreme value theorem (Balkema and de Haan, 1974; Pickands, 1975) provides the theoretical foundation of the peaks-over-threshold approach. Rather than modelling a series formed from the highest peak of each year, only the peaks that surpass a threshold level u are selected. If the difference between each peak value and the threshold value u is recorded (i.e. $X-u$), the series

Table 1. The daily discharge record of the Erzen River at Ndoroq (1979–2023)

Characteristic	Value
Record period	1 Jan 1979 – 30 Jun 2023
Length	44.5 years (45 water years)
Daily observations	16,252
Mean daily discharge	34.5 m ³ /s
Median daily discharge	15.9 m ³ /s
Mean annual maximum	450.9 m ³ /s
Minimum annual maximum	149 m ³ /s
Maximum annual maximum	1,129 m ³ /s (10 Nov 2016)

formed by these differences is called the partial duration series. The second EVT states that, provided the selected threshold u is sufficiently high, this series converges to the Generalised Pareto distribution. The conditional distribution of the excess values of $X-u$, given $X>u$ is:

$$F_u(y) = Pr(X - u \leq y \mid X > u) = \frac{F(y + u) - F(u)}{1 - F(u)} \quad (3)$$

$F_u(y)$ expresses the frequency of occurrence of flow values given that a flood above the threshold u has already occurred. $F_u(y)$ converges to the Generalised Pareto distribution $G(x)$ regardless of the parent distribution:

$$\lim_{u \rightarrow \infty} F_u(y) \rightarrow G(x) \quad (4)$$

$$G(x) = 1 - \left\{ 1 - \frac{\kappa(x - u)}{\alpha} \right\}^{1/\kappa}, \quad \kappa \neq 0 \quad (5)$$

Equation (5) reduces to the exponential distribution $G(x)=1-\exp(-(x-u)/\alpha)$ for $\kappa=0$. The location parameter of the GP distribution is fixed at the threshold, $\zeta=u$, so only the scale α and the shape κ are estimated from the excesses; the shape parameter has the same interpretation and the same sign as in Equation (2). The admissible range of Equation (5) follows the same convention: the excesses are bounded above, $u \leq x \leq u + \alpha/\kappa$, only when $\kappa > 0$, whereas for $\kappa \leq 0$ they are unbounded, $x \geq u$. With the fitted $\kappa = -0.276$ the upper tail is therefore unbounded, which is the formal basis for the heavy-tailed reading of the shape parameter.

The two theorems are not independent descriptions of natural phenomena; they are two

views of the same process, and the link between them is what makes the POT method advantageous for design. If exceedances over the threshold occur in time as a Poisson process (with rate λ events per year), and their magnitudes follow a GP distribution, then the largest events in a year (i.e. the annual maxima) follow exactly a GEV distribution, whose three parameters are analytical functions of λ and of the GP parameters (Madsen et al., 1997; Pan et al., 2022). This equivalence, given in Equations (22-24) below, means that a PDS analysis loses nothing compared to an AMS analysis: it ends in the same GEV family but with its parameters estimated from a sample that contains only floods (i.e. hydrologically significant events). In addition, the equivalence transfers the shape parameter unchanged, so the tail behaviour diagnosed in the threshold excesses carries over directly to the annual frequency curve.

Modelling of the annual maximum series

The focus of this study is the application of the second extreme value theorem. However, since the first EVT is the more commonly used of the two, we also applied it here to compare the quantile estimates and the confidence intervals obtained by fitting the GEV distribution to the AMS with those obtained by fitting the GP distribution to the PDS. The AMS was extracted based on the hydrological year (1 October – 30 September), yielding a series of 45 annual maxima (sample size $n = 45$). The GEV distribution was fitted to this series using the method of L-moments. L-moments are summary statistics formed from linear combinations of order statistics (Hosking, 1990). The first L-moment λ_1 corresponds to the mean of the series, while the second L-moment λ_2 measures spread and is analogous to the standard deviation in the method of moments. The L-moment ratio $\tau_3=\lambda_3/\lambda_2$ (L-skewness) measures asymmetry, in analogy with the conventional coefficient of skewness but without squaring or cubing the data. This is the main advantage that the method of L-moments has compared to the traditional method of moments: no observation is ever raised to a power. Thus, the effect of outliers on the resulting moment estimates is attenuated, which makes the method of L-moments more robust and less biased than the method of moments, especially when dealing with maxima series, whose samples tend to be small and skewed.

The L-moments of the GEV distribution are related to its parameters by the following expressions:

$$\lambda_1 = \xi + \frac{\alpha\{1 - \Gamma(1 + \kappa)\}}{\kappa} \quad (6)$$

$$\lambda_2 = \frac{\alpha(1 - 2^{-\kappa})\Gamma(1 + \kappa)}{\kappa} \quad (7)$$

$$\tau_3 = \frac{2(1 - 3^{-\kappa})}{1 - 2^{-\kappa}} - 3 \quad (8)$$

where: $\Gamma(\cdot)$ is the gamma function.

The estimation proceeds first by computing the sample L-moments from the 45 ordered annual maxima, then Equation (8) is solved for κ , since the sample L-skewness alone determines the shape parameter. Next, using the standard rational approximation of Hosking et al. (1985), the values of α and ξ are computed from Equations (7) and (6), respectively. The resulting GEV curve serves as a baseline reference against which the partial duration series results are compared.

Modelling of the partial duration series

The partial duration series and its assumptions

The PDS is composed of the excesses $X-u$ of the peaks that are greater than the threshold value u . In order for this series to be valid as a statistical sample, three important assumptions should be fulfilled: the sample must be representative of the population, and its elements must be independent and identically distributed (Bezák et al., 2014).

Representativeness requires that the sample be composed only of floods. A sufficiently high threshold ensures that only genuine flood peaks enter the PDS, while a threshold that is too high limits the sample size. If this is not fulfilled, the model fitted to the sample does not actually represent the population it is meant to describe. Independence requires that each flood retained in the series be independent of the others. In practice, this means that enough time must pass between two consecutive peaks and that the hydrograph must sufficiently recede between them. If peaks belonging to the same flood event are retained, the frequency of large floods is overestimated, i.e. large floods appear to happen more often than they actually do. Identical

distribution requires that all the excesses in the series be generated by the same underlying process. This is violated when the retained peaks originate from different flood-generating mechanisms, or when the series is affected by a trend, so that the excesses no longer come from a single parent distribution. Thus, selecting a proper threshold value is crucial in fulfilling these three assumptions.

Flood separation: independence criteria

The time separation and the recession criteria are two independence criteria that were applied jointly to guarantee that the peaks retained in the PDS represent separate flood events and are neither part of the same hydrograph nor produced by the same meteorological event (Fig. 2).

Time separation represents the minimum duration of time between two independent peaks (Fig. 2). Following the criterion given in Bulletin 17B and retained in United States Army Corps of Engineers practice (IACWD, 1982; USACE, 1993), two peaks are considered independent only if a time interval of d days has passed between them:

$$d = [5 + \ln A] \quad (9)$$

where: A is the drainage area of the watershed in square miles and $[\cdot]$ denotes rounding up to the next integer, so that d is a whole number of days.

Larger watersheds produce larger floods with longer duration, and more time is needed for the flow rates to go back to baseflow levels; thus, the required separation d (Equation (9)) grows logarithmically with watershed size. In our case, $A = 256 \text{ mi}^2$ (i.e. 663 km^2), so the argument of Equation (9) evaluates to 10.55 days, hence $d = 11$ days. Because Equation (9) is empirical and its constants were not derived for Albanian watersheds, the sensitivity of the results to the separation interval was quantified by repeating the entire extraction and fitting with $d = 3$ and $d = 21$ days, a seven-fold range in the separation interval about the value given by Equation (9).

The recession criterion (i.e. magnitude differential) is needed because the first criterion on its own may not account for a long, slow-draining flood wave whose secondary crest arrives after the time separation window has elapsed (Fig. 2). Following the independence rule of the UK Flood

Estimation Handbook, two consecutive peaks are treated as independent, only if the discharge difference between them is lower than two-thirds of the first peak (Robson and Reed, 1999):

$$\min_{t_1 < t < t_2} Q(t) < \frac{2}{3} Q_{p1} \quad (10)$$

where: Q_{p1} is the first of the two peaks and t_1 and t_2 represent the time of occurrence of the first and second peak, respectively.

The criterion is evaluated on the earlier peak, Q_{p1} : the minimum discharge in the interval between the two peaks must fall below two-thirds of Q_{p1} for them to be treated as independent, and otherwise the smaller of the two peaks is removed.

Poisson occurrence model (modelling of counts)

Contrary to the AMS, where each year contributes a single element (the highest flood peak of that year), in the PDS the elements (events) arrive randomly: sometimes one in a year or several in a single year (wet years), as well as none in other years. To model the counts of these events, the Poisson distribution was used:

$$\Pr(N = n) = \frac{\lambda^n e^{-\lambda}}{n!}, \quad n = 0, 1, 2, \dots \quad (11)$$

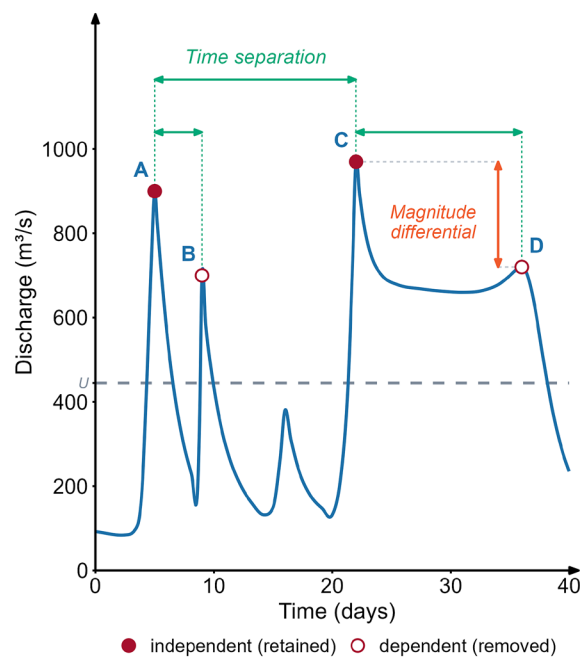


Figure 2. Separation of independent flood events, adapted after Van Campenhout et al. (2020)

where: λ is the average rate of events per year, estimated as $\lambda = m/M$, m is the total number of events and M is the record length in years (Kjeldsen and Prosdocimi, 2023).

The Poisson process assumes that while events arrive at a steady average rate λ , they are random and independent of one another. These assumptions are empirical, thus they should be tested. The dispersion index (DI) of Equation (16) below was used and tested again at the final computational threshold both by the χ^2 test on $(M-1) \cdot DI$ and by comparing the observed proportion of years with 0, 1, 2, ... events against the fitted probability mass function.

Threshold selection

The threshold u is the most consequential choice in the peak over threshold method, since it determines which peaks enter the PDS. A threshold that is too low admits non-extreme values, so the asymptotic argument behind the second extreme value theorem does not hold and the fitted GP is biased. A threshold that is too high leaves too few exceedances, so the parameter estimates become unstable and increasingly sensitive to individual events. No single criterion identifies the optimal, or computational, threshold on its own; therefore, six diagnostic plots were computed across the candidate range and read jointly: the number of exceedances, the Generalised Pareto shape parameter, the mean excess, the right-tail-weighted Anderson-Darling statistic, its parametric-bootstrap p -value and the dispersion index of the annual counts. The first five inform the choice of the threshold, while the sixth validates the Poisson occurrence model at the value selected.

To locate the computational threshold, the six diagnostics were computed over 100 equally spaced candidate thresholds from 150 m³/s to 700 m³/s, augmented by the selected threshold so that 101 points were evaluated (Figs. 4–5). The value $u = 150$ m³/s corresponds approximately to the smallest annual maximum, essentially capturing every maximum in the series, together with numerous secondary peaks. The two contrasting demonstration thresholds of Fig. 3, $u = 150$ m³/s and $u = 550$ m³/s, bracket the region within which the computational threshold is sought. The threshold cannot be lower than the lowest observed annual maximum, and above about 550 m³/s the partial duration series retains only about a dozen events, too few for a stable Generalised Pareto fit;

the diagnostics were nonetheless computed and plotted up to 700 m³/s to display the behaviour at high thresholds.

Number of exceedances plot

When the number of events lying above a threshold value u is plotted, the number of events decreases as the threshold level increases from 150 m³/s upwards. The number-of-exceedances graph practically shows the size of each PDS for a certain level of threshold. The threshold level at which the number of events corresponds to the number of total years, a mean event rate of one per year, is a conventional anchoring point, because the PDS and AMS then have equal sample size. Although the point itself carries no special statistical meaning, the plot is nevertheless useful for threshold selection, since it shows how quickly the sample size shrinks as the threshold rises, and therefore represents graphically the size of the resulting PDS. Below the anchoring point of one event per year the sample is larger than the AMS and admits smaller peaks, whereas above it the sample size is smaller than the AMS and the estimates might become increasingly sensitive to individual events.

Generalised Pareto shape parameter plot

For each candidate threshold, the shape parameter κ of the Generalised Pareto distribution is estimated from the corresponding PDS above that threshold by the method of L-moments. These κ values are then plotted against the corresponding threshold value. The Generalised Pareto

distribution has a threshold invariance property: if excesses of some threshold u_0 follow a GP distribution with shape κ , then excesses over any higher threshold follow a GP distribution with the *same* κ (Coles, 2001). Once the threshold is high enough for the second EVT to hold, the estimated shape parameter should therefore stay approximately constant as the threshold rises. Erratic behaviour at low thresholds indicates contamination by non-extreme values, while instability at very high thresholds reflects sampling noise due to the small sample size. The optimal threshold is therefore located where the plot flattens into a plateau of approximately constant κ values, or at a clear break in the curve.

Mean excess plot

For each candidate threshold, the average of the deviations of each peak over the threshold plotted against the corresponding threshold value represent the mean excess plot. The sample mean excess for a threshold u is given by:

$$e(u) = \frac{1}{m(u)} \sum_{i=1}^{m(u)} (x_i - u) \quad (12)$$

where: $m(u)$ is the number of peaks that exceed the threshold u and $x_i - u$ are the corresponding deviations.

If the excesses over a threshold u follow a Generalised Pareto distribution with scale α and shape κ ($\kappa > -1$), then, by the threshold invariance

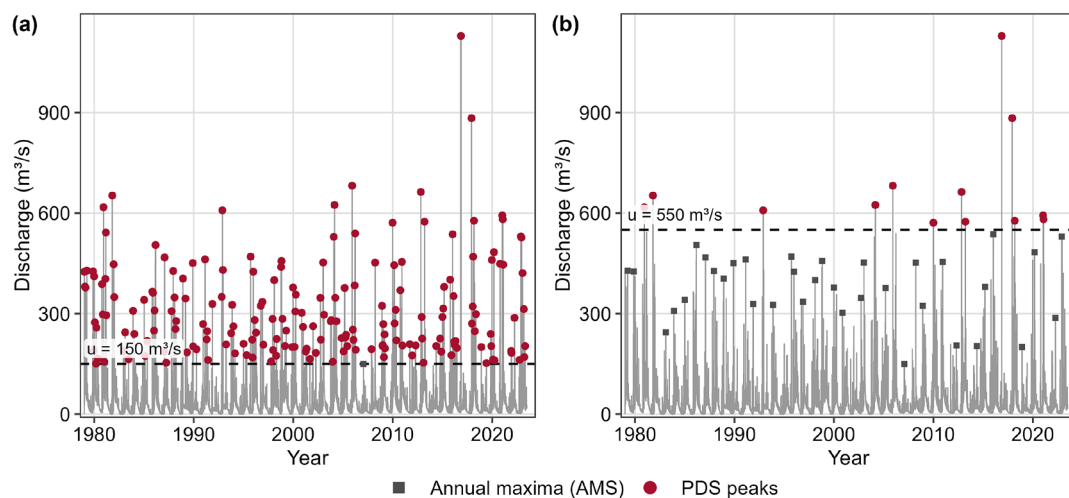


Figure 3. Partial duration series of the Erzen River at Ndreq for the two extreme demonstration thresholds: (a) $u = 150$ m³/s, approximately the smallest annual maximum; (b) $u = 550$ m³/s, near the highest observed peak values

property, the mean excess over any higher threshold u' is a linear function of the threshold:

$$e(u') = \frac{\alpha_u - \kappa(u' - u)}{1 + \kappa}, \quad u' > u \quad (13)$$

where: u' is any threshold higher than u and α_u is the scale parameter of the excesses over u .

The line therefore has slope $-(\kappa)/(1+\kappa)$. Thus, the mean excess plot of $e(u)$ against u should be approximately a straight line above the lowest threshold at which the GP model is adequate (Langousis et al., 2016). The optimal threshold is therefore located at the value above which the mean excess plot becomes approximately linear, since linearity indicates that the second extreme value theorem holds from that threshold upwards. As with the shape parameter plot, the departures from linearity at low thresholds reflect the presence of non-extreme values in the sample, whereas the widening scatter at high thresholds reflects the small number of exceedances that remain (Coles, 2001).

Anderson-Darling goodness of fit across thresholds

The choice of the optimal threshold made using the Generalised Pareto shape parameter and the mean excess graphs can be further confirmed with a Right-Tail-Weighted Anderson-Darling statistic (RTWAD) of the GP fit at each candidate threshold. This test, in contrast with the classical Anderson-Darling statistic, emphasises the upper tail, where large and rare discharges are located. After the selection of the optimal threshold and the extraction of the PDS, for the ordered sample of the m exceedances above u , $x_{(1)} \leq \dots \leq x_{(m)}$, with fitted GP probabilities $z_i = G(x_{(i)})$, the RTWAD statistic is calculated by:

$$A^2 = \frac{m}{2} - 2 \sum_{i=1}^m z_i - \sum_{i=1}^m \left(2 - \frac{2i-1}{m} \right) \ln(1 - z_i) \quad (14)$$

This measures the distance between the empirical distribution of the sample and the fitted GP (Sinclair et al., 1990; Solari et al., 2017). The argument $\ln(1-z_i)$ gives more weight to the upper-tail values, rather than weighting both tails

equally. The values of this test are then plotted against the threshold. Lower values practically mean the fitted GP is close to the data in the upper tail. However, this does not mean that the threshold values which have the lowest RTWAD value are the best. There can be very high threshold values which can well have low RTWAD values, but the PDS in these cases is of small size. Too few events in the test mean lower power of the test. The RTWAD test thus favours very high thresholds. Therefore, the graph must be read together with the other graphs, and a low statistic alone is not a reason to choose a threshold. This graph is a relative diagnostic, good for spotting where the fit is stable or deteriorating, but a poor absolute verdict on its own.

An alternative variant of the RTWAD is its representation through a parametric bootstrap. Given that the GP parameters are estimated from the same data that are being tested, the usual critical values of A^2 do not apply. Thus, the p -values were obtained by parametric bootstrap. First the GP was fit to the real peaks and the A^2 was recorded. Then, synthetic dataset of the same size was generated using the fitted GP. The refit is applied to the new GP and A^2 is recomputed. This procedure is repeated many times ($B=999$ per threshold). This pile of simulated A^2 value represent the null hypothesis: peaks follow the GP distribution. The p -value is calculated as the ratio of the simulated samples that fit as badly as, or worse than, the real one (Choulakian and Stephens, 2001):

$$p = \frac{1 + \#\{A_b^2 \geq A_{obs}^2\}}{B + 1} \quad (15)$$

$B = 999$ replications were used across the threshold scan and $B = 4999$ at the final threshold. The GP hypothesis is not rejected wherever $p > 0.05$, meaning the misfit is ordinary and there is no evidence against GP, and rejected for $p \leq 0.05$, where the misfit is too large to be attributed to chance. Using the Anderson-Darling p -value in this way can be used to screen candidate thresholds automatically (Solari et al., 2017). However, for the reasons mentioned above, this test is used here as supporting evidence and does not replace the graphical evidence. A high p -value might look like “good news”, in that the model appears to be adequate. But not rejected is not proven perfect and a very high p -value at a high threshold partly reflects few events, rather than a genuine, better fit.

Dispersion index of the annual counts

In the peak over threshold method, a combination of the Pareto and Poisson models is used. The first, the GP distribution, is used to model the magnitude of flood events, while the Poisson process models the counts (numbers) of these events happening each year. The Poisson process is used to model rare, independent events happening at a constant average rate. It has the defining property that the variance of the annual number of events equals its mean. The ratio of the two is called the dispersion index, given by:

$$DI(u) = \frac{s_N^2(u)}{\bar{N}(u)} \quad (16)$$

where: $\bar{N}(u)$ and $s^2N(u)$ are the mean and the variance of the annual event counts, including the years with zero events.

The dispersion index should be close to 1, if the counts are to behave like a Poisson process. Values well above 1 indicate that the counts vary more than the Poisson process would allow: floods cluster, bunching into some years, thus breaking the independence assumption (Fischer et al., 2023). Values under 1 indicate under-dispersion, with floods arriving more regularly than pure chance would produce. Under the Poisson hypothesis, if M annual counts come from a Poisson Process, the scaled dispersion index $(M-1) \cdot DI$ follows a chi-square distribution with $M-1$ degrees of freedom. DI should remain inside the 95% acceptance band across the threshold range if the Poisson occurrence model is to be trusted (Kjeldsen and Prodocimi, 2023; Lang et al., 1999). The dispersion index plot depicts graphically the value of the DI at each candidate threshold. Rather than helping to select the optimal computational threshold, this plot serves to validate and confirm that the Poisson count model is trustworthy at the chosen threshold.

Based on the evidence output from the five diagnostic plots, the computational threshold was chosen. It is important to emphasise that no single diagnostic plot is decisive on its own, and the selection is based on their convergence, as far as possible. Based on the selected threshold, the final PDS was extracted. The PDS was formed by elements located about the computational threshold value, and which at the same time fulfil the independence criteria (i.e. time separation and recession criteria).

Elements that did not fulfil these criteria were discarded and do not form part of the PDS.

Generalised Pareto fit and goodness of fit

The peaks above the computational threshold $u = 445 \text{ m}^3/\text{s}$ were selected and recorded in the series, while at the same time the independence criteria for time separation and recession (Equations 9 and 10) were applied. The resulting series represents the partial duration series, on which the calculations that follow are based. The GP distribution was fitted to this series using the method of L-moments with the location parameter fixed at the threshold ($\xi = u$). The L-moments of the GP are related to its parameters by the following expressions:

$$\lambda_1 = u + \frac{\alpha}{1 + \kappa} \quad (17)$$

$$\lambda_2 = \frac{\alpha}{(1 + \kappa)(2 + \kappa)} \quad (18)$$

where the parameters follow in the closed form from the sample L-moments (Hosking and Wallis, 1987, 1997):

$$\hat{\kappa} = \frac{\hat{\lambda}_1 - u}{\hat{\lambda}_2} - 2, \quad (19)$$

$$\hat{\alpha} = (1 + \hat{\kappa})(\hat{\lambda}_1 - u)$$

The empirical non-exceedance probabilities were plotted using the Weibull plotting position $p_i = i/(m+1)$, which is unbiased in probability, where i is the rank of the ordered excess and m is the number of events in the PDS. The quality of the fit of the GP to the PDS was assessed in three ways: with the probability plot of the ordered events against the fitted quantile function, with the direct comparison of the empirical and fitted cumulative distribution functions, and with the right-tail-weighted Anderson-Darling test of Equations (14)-(15), using $B = 4999$ bootstrap replicates.

Annualisation of the PDS model

The quantiles given by the GP distribution do not represent annual probabilities, given that it was fitted to the PDS, which is formed by elements that, unlike the AMS, do not happen annually. The GP distribution of Equation (5) is a conditional distribution describing the probability $G(x)$ that an event does not exceed x , given that this event has exceeded a threshold u . It does

not describe the annual occurrence of the flood quantiles. This means that the quantiles given by the conditional probability function $G(x)$ must be converted into annual exceedance probabilities. For this, two methods were used, one empirical by Langbein (1949) and the other analytical by Madsen et al. (1997).

Empirical transformation

Based on Langbein (1949) the relations between the PDS cumulative distribution function $G(x)$ and its annualised counterpart $F(x)$ for a given mean rate λ are:

$$G(x) = \frac{[\ln(F(x)) + \lambda]}{\lambda} \quad (20)$$

$$F(x) = \exp(-\lambda[1 - G(x)]) \quad (21)$$

This conversion follows an elementary counting argument. The annual maximum is smaller than x exactly when no event during the year exceeds x . Under the Poisson model, events exceeding the threshold arrive at rate λ , and of these a fraction $1 - G(x)$ exceeds x , so exceedances of x arrive at a rate of $\lambda \cdot [1 - G(x)]$. The Poisson probability of observing none of them in a year gives directly the annual non-exceedance probability given in Equation (21). This equation reproduces the classical correspondence between partial-duration and annual return periods, $T_A = 1/\{1 - \exp(-1/T_p)\}$, tabulated by Langbein (1949) and given a rigorous basis by Takeuchi (1984); for return periods beyond about ten years the two scales practically coincide, but at frequent quantiles the distinction is appreciable. Applied to the Weibull plotting positions of the observed peaks, Equation (21) places the partial duration series on the annual probability axis without any distributional assumption, which allows a direct graphical comparison with the annual maxima; its limitation is that, being empirical, it cannot extrapolate beyond the largest observed event.

Analytical transformation

By combining Equations (5) and (11), the fitted Pareto–Poisson model can be transformed into an equivalent GEV distribution of annual maxima (Stedinger et al., 1993). The relationships between the Generalised Pareto distribution parameters (ξ , α , κ) and the Generalised Extreme Value parameters (ξ^* , α^* , κ^*) are (Madsen et al., 1997):

$$\xi^* = u + \frac{\alpha}{\kappa}(1 - \lambda^{-\kappa}), \quad \kappa \neq 0 \quad (22)$$

$$\alpha^* = \alpha \lambda^{-\kappa} \quad (23)$$

$$\kappa^* = \kappa \quad (24)$$

with $\xi^* = u + \alpha \ln \lambda$ and $\alpha^* = \alpha$ for the limiting case $\kappa = 0$. The conversion changes location and scale, both of which depend on how often events occur, while the shape parameter is transferred unchanged by Equation (24), so the tail heaviness diagnosed in the threshold excesses is inherited intact by the annual frequency curve. Design quantiles for a return period T then follow from the GEV quantile function evaluated at $F = 1 - 1/T$:

$$x_T = \xi^* + \frac{\alpha^*}{\kappa^*} \left\{ 1 - \left[-\ln \left(1 - \frac{1}{T} \right) \right]^{\kappa^*} \right\} \quad (25)$$

It is worth noting that the two annualisation methods used here are not rivals: under the Pareto–Poisson model, Equations (21) and (22)–(24) are mathematically equivalent for quantiles above the threshold and yield identical values. The Langbein transformation was used for the empirical plotting positions, where its distribution-free character is the advantage, and the Madsen conversion for the parametric curve and its extrapolation.

Uncertainty analysis

All analyses reported in this section were performed in R programming language (R Core Team, 2022). Sampling uncertainty of both curves was quantified by parametric bootstrap (Efron and Tibshirani, 1993; Schendel and Thongwichian, 2015). Treating the fitted model as a working truth, many records of the same length as the real one were simulated from it, and the model was refitted to each simulated record. Then it was observed how widely the re-estimated quantiles scattered, which locates the 90% confidence interval.

For the PDS model, each of $B = 10,000$ replicates was generated by drawing the number of events N from a Poisson distribution with mean $\lambda \cdot n_y$, drawing N magnitudes from the fitted GP distribution, re-estimating the GP parameters by L-moments (with the location fixed at the threshold), and re-applying Equations (22)–(24). Replicates with fewer than 10 events, for which the L-moment fit is unreliable, were redrawn, a restriction that affects a negligible fraction of draws at the

observed event rate. All bootstrap computations used a fixed seed (`set.seed(42)` in R), re-seeded before the final goodness-of-fit test and before the quantile bootstrap so that the reported intervals are independent of the earlier threshold scan. The threshold was held fixed at $u = 445 \text{ m}^3/\text{s}$ in every replicate, the excesses were resampled but not re-thresholded, so the intervals quantify the sampling uncertainty of the magnitudes and rate at the selected threshold. The fewer-than-ten-events rule is immaterial in practice: with a mean of $\lambda \cdot n_y \approx 34$ events the Poisson probability of drawing fewer than ten is below 10^{-6} , so essentially no replicate is affected. For the AMS model, each replicate consisted of 45 annual maxima drawn from the fitted GEV distribution and refitted by L-moments. The 5th and 95th percentiles of the B quantile estimates define the 90% confidence interval at each return period.

The quantiles of the two models, together with their 90% confidence intervals, were compared with each other, and the differences are shown both graphically and numerically. Both frequency curves (i.e. AMS/GEV and PDS/GP–Madsen) were computed for annual exceedance probabilities from 0.5 down to 0.001 (return periods of 2 to 1000 years). R version 4.2.2 was used, with the packages POT for declustering (Ribatet and Dutang, 2022), `lmom` and `lmomco` for the L-moment fits and the Langbein transformation (Asquith, 2022; Hosking, 2022). The exact package versions were POT 1.1-10, `lmom` 3.0, `lmomco` 2.5.1, `tidyverse` 2.0.0 and `scales` 1.3.0, with `trend` 1.1.6 for the stationarity tests.

RESULTS AND DISCUSSION

Threshold selection

The number of exceedances plot (Fig. 4a) shows a continual decrease as the threshold value increases, meaning fewer and fewer events are included above the threshold value. At $n=45$ the number of events corresponds to the number of years of record (i.e. the AMS sample size). At this horizontal line the rate of exceedances (events/year) is equal to 1, at a threshold of approximately $400 \text{ m}^3/\text{s}$. This crossing point anchors the search interval: to its left the event rate exceeds one per year and the sample increasingly admits non-extreme values, to its right the sample shrinks below the AMS size, which is acceptable only if the loss

of sample size is compensated by a closer agreement with the asymptotic model, i.e. a shape parameter that has stabilised and a mean excess plot that is approximately linear from that threshold upwards. Other than informing us about the size of the resulting PDS, this plot cannot further aid in the selection of the computational threshold.

The GP shape parameter (Fig. 4b) plot shows how the value of κ parameter changes across the PDS formed at each of the thresholds shown on the abscissa. The graph is “messy”, as unfortunately real measured data are. Ideally, we look for approximately flat regions, where κ takes stable values. In our case, there is an apparent plateau between roughly 200 and $350 \text{ m}^3/\text{s}$. However, the flows in this region are low and the corresponding samples are dominated by peaks below the mean annual maximum, which are de facto non-floods, so selecting a threshold here would violate the representativeness assumption and, with it, the conditions under which the second EVT holds. From about $350 \text{ m}^3/\text{s}$ the estimate declines systematically, crosses zero near $390 \text{ m}^3/\text{s}$ and reaches $\kappa = -0.27$ at $445 \text{ m}^3/\text{s}$. Between 420 and $455 \text{ m}^3/\text{s}$ the curve forms a valley, within which κ ranges from -0.22 to -0.30 . At $445 \text{ m}^3/\text{s}$, the curve changes shape, turning from declining to increasing. Above $445 \text{ m}^3/\text{s}$ the curve rises steadily to a local maximum near $470 \text{ m}^3/\text{s}$, then falls again and varies erratically as the threshold increases further. At these high thresholds, fewer and fewer events are retained, so the PDS becomes small in size and the optimal threshold is unlikely to lie in this range: the number of events is 21 at $470 \text{ m}^3/\text{s}$, 19 at $500 \text{ m}^3/\text{s}$ and 13 at $540 \text{ m}^3/\text{s}$ (Fig. 4a). Since the valley between 420 and $455 \text{ m}^3/\text{s}$ shows no erratic variation in κ , varying by only 0.08 (from -0.22 to -0.30), and since the graph offers no definitively stable region elsewhere, we assume that the optimal threshold lies within this range.

In the mean excess plot (Fig. 4c) what we look for is the threshold from which there is a clear linearity in the graph. From $230 \text{ m}^3/\text{s}$ up to $445 \text{ m}^3/\text{s}$ the graph declines quasi-linearly, with the presence of some small noise throughout. From $445 \text{ m}^3/\text{s}$ there is a clear linear growth up to $472 \text{ m}^3/\text{s}$, and a quasi-linear growth from $445 \text{ m}^3/\text{s}$ to $483 \text{ m}^3/\text{s}$. After that the graph varies erratically, with some regions of linear growth or decrease at the very high thresholds. In these regions, the optimal threshold is unlikely to be found, since the potential PDS would have very few events, hampering the second EVT and making it difficult to fit a GP

distribution. Based on the GP shape parameter plot discussed in the previous paragraph, we concluded that the optimal threshold might lie within the range 420–455 m³/s. In the same region of the mean excess plot, a linear growth begins exactly at 445 m³/s. Therefore, since there is enough agreement between the GP shape parameter and the mean excess plot, we adopt 445 m³/s as the computational threshold for the analysis, provided that the Anderson-Darling goodness-of-fit test does not contradict it.

The Anderson-Darling (right-tail weighted) plot (Fig. 5a) depicts the value the A^2 quantity takes across thresholds. The smaller this value, the better the fit between the empirical distribution and the GP curve. In two regions of the plot, A^2 takes its lowest values: from 150–178 m³/s with values fluctuating from 0.15 to 0.17, while from 500 m³/s and upwards A^2 values range from 0.20 down to 0.12, making this the interval with the lowest RTWAD values. In another part of the graph, the interval 395–430 m³/s presents relatively calm values oscillating around 0.20. If we were to look only at the RTWAD plot as a means to select the optimal threshold, while disregarding the other plots, neither of the two extreme

regions (i.e. 150–178 m³/s and 500 m³/s and upwards) could be considered, given the very small and very high thresholds that would result in these cases. Thus, a low statistic value alone is not a reason to choose the optimal threshold. Generally, the RTWAD test tends to favour very high thresholds as the optimal threshold. This is why the literature reports mixed performance when this test is used as an automated method to select the optimal threshold (Scarrott and MacDonald, 2012; Solari et al., 2017). In our case, at 445 m³/s, which was the best candidate threshold based on the GP shape parameter and mean excess plots, the RTWAD value is 0.33. While a local maximum, this test value, decreasing on both sides of 445 m³/s, is hard to read on its own. The alternative parametric bootstrap version of the Anderson-Darling test can further clarify whether 445 m³/s is an adequate optimal threshold or should be disregarded.

Figure 5b shows graphically the right-tail-weighted bootstrap p-values across all the thresholds. Across the whole range these p-values lie above the 5% significance level, meaning that our null hypothesis stands: the GP distribution is not rejected at any threshold, and the

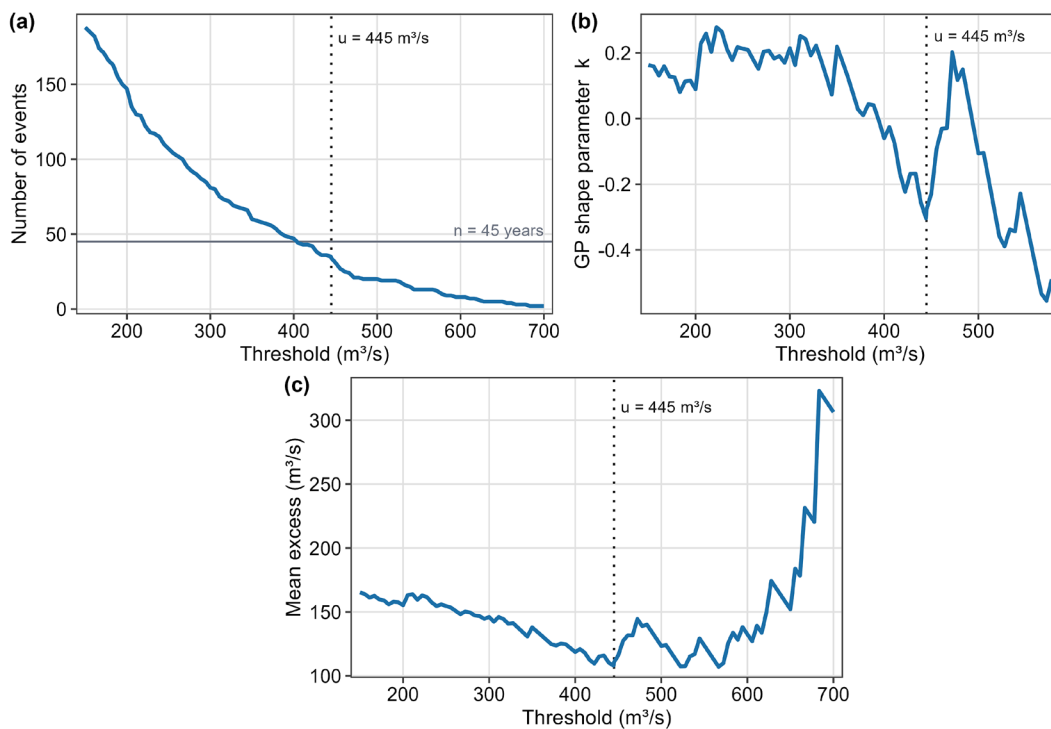


Figure 4. Threshold diagnostic plots for the Erzen River at Ndreq station: (a) number of exceedances versus threshold, with the horizontal line at the number of years of record; (b) generalised Pareto shape parameter versus threshold; (c) mean excess plot. The dotted vertical line marks the selected threshold $u = 445$ m³/s

fit is therefore acceptable wherever it is applied. The curve does approach the 5% level at a few isolated thresholds between about 310 and 350 m^3/s , and it passes through a local minimum in the immediate neighbourhood of the candidate threshold: 0.082 at 445 m^3/s (marginal), recovering to 0.12 by 450 m^3/s and to above 0.20 by 460 m^3/s . The scan uses $B = 999$ replicates per threshold; the definitive goodness-of-fit test at the selected threshold uses $B = 4999$ and returns $p = 0.076$ for the same statistic ($A^2 = 0.333$). The two are the same quantity estimated at different bootstrap resolutions, not a discrepancy, and $p = 0.076$ is the value reported throughout. This local minimum is not accidental. Three moderate peaks lie between 445 and 450 m^3/s . These are the events that arrest the decline of the shape trace and that terminate the declining limb of the mean excess. They are simultaneously the part of the sample that the fitted generalised Pareto distribution reproduces least well. The same three events that make 445 m^3/s the natural change point therefore also make the fit at that threshold its weakest, which is why the p -value dips exactly there. The dip nevertheless remains above 0.05, and, being a noisy Monte-Carlo estimate, it should not be over-interpreted. While this fact might not be useful in selecting the optimal threshold, it can serve as

a further confirmation that our best candidate threshold of 445 m^3/s , offered by the other diagnostic plots, cannot be rejected. Thus, Figure 5b's role is to confirm the other diagnostics of the adequacy of the GP model rather than to single out one threshold by itself.

The dispersion index plot of the annual event counts (Fig. 5c) shows that the flood occurrences remain inside the 95% χ^2 acceptance band over the entire candidate range. The acceptable band extends with DI values from 0.63 to 1.46. As a result, the annual flood counts are consistent with a Poisson process at every candidate threshold. At the selected threshold $u=445 \text{ m}^3/\text{s}$, $\text{DI} = 0.91$ ($p=0.64$), sitting just below the ideal DI value of 1, and confidently inside the band. The underlying counts are 20 years with no floods, 18 with one, 5 with two and 2 with three – 34 events in 45 years. $\text{DI} = 0.91$ is statistically indistinguishable from 1 ($p = 0.64$): the counts are Poisson to within sampling noise, and the dip below 1 is not significant. With $\lambda = 0.76$, most years have no flood or just one (38 of the 45 years), and 20 have none at all. The dip below 1 means only that the counts vary slightly less from year to year than a pure Poisson process would. At $p = 0.64$ the record is too short for that difference to mean anything. The problematic case would have been the opposite: a DI drifting above the

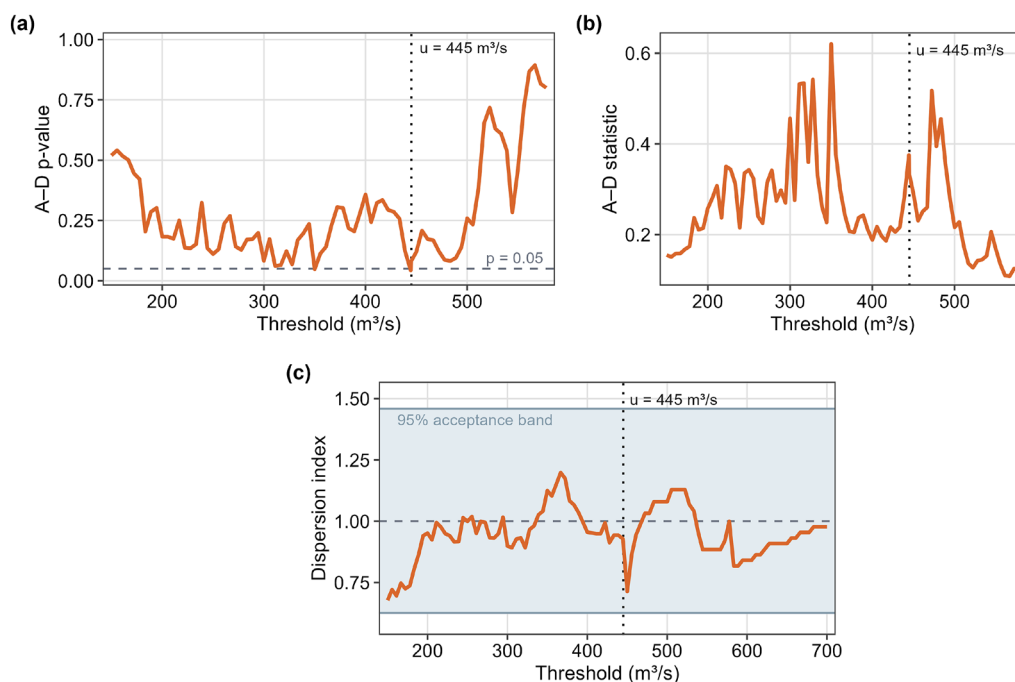


Figure 5. (a) Anderson-Darling (right-tail-weighted) statistics across thresholds, (b) bootstrap p -values of the right-tail-weighted Anderson-Darling test of the GP fit across thresholds, (c) dispersion index of the annual event counts versus threshold with the 95% acceptance band of the Poisson hypothesis

upper bound (over-dispersion), which signals flood clustering and would have required a different count model, such as the negative binomial (Bhunya et al., 2013). The dispersion index plot validates the selected computation threshold rather than selecting it, further confirming the Poisson count model is trustworthy at the chosen threshold $u = 445 \text{ m}^3/\text{s}$.

Reading the six diagnostic plots together, we conclude that the optimal threshold for computation is $u = 445 \text{ m}^3/\text{s}$. At this threshold the sample retains 34 exceedances, an event rate of 0.76 per year, slightly below the one-per-year anchor, and the GP shape parameter is $\kappa = 0.276$, inside the 420 - 455 m^3/s valley in which κ varies by only 0.08 and shows no erratic behaviour. The mean excess at $u = 445 \text{ m}^3/\text{s}$ is 110.6 m^3/s and marks the start of the linear limb of the mean excess plot, the behaviour expected when the excesses over the threshold follow a Generalised Pareto distribution. The right-tail weighted Anderson-Darling statistic is 0.33 and the bootstrap p -value at the same threshold is 0.082, on the 101-point scan (0.076 in the definitive $B = 4999$ test), above the 0.05 cut-off, so the GP fit at 445 m^3/s cannot be rejected. The dispersion index is $DI = 0.91$, comfortably inside the 95% χ^2 acceptance band, confirming that the annual event counts at this threshold are consistent with a Poisson process. Taken together, the diagnostics support $u = 445 \text{ m}^3/\text{s}$ as the computational threshold in the case of our study.

Final PDS and application of the independence criteria

At $u = 445 \text{ m}^3/\text{s}$, the extraction with the 11-day separation of Equation (9) yields 34 peaks. This

separation is an empirical rule considered reliable for watersheds ranging between 100 and 10,000 mi^2 , a range within which the Ndroq watershed falls comfortably. The subsequent application of the recession criterion of Equation (10) did not remove any additional peaks, meaning that between every pair of consecutive retained peaks the discharge had already receded below two-thirds of the first peak. This is physically expected: in a 663 km^2 watershed fed by Mediterranean climate inputs, flood waves rise and drain within days, so any two peaks more than a week apart are almost automatically separate hydrological events. Our final sample therefore consists of independent and identically distributed values, with the final PDS comprising 34 independent events in 45 years ($\lambda = 0.76$ events per year), see Figure 6.

The sensitivity of the results to the time separation criterion is summarised in Table 2. Shortening the time interval from 11 to 3 days adds a single event (35 peaks), while lengthening it to 21 days removes four (30 peaks). Across this sevenfold variation, the 1% annual quantile changes by less than 1% (1122, 1114 and 1121 m^3/s for 3, 11 and 21 days respectively). In our case, then, all three time-intervals give practically the same frequency curve. This is worth stating, because de-clustering rules are often criticised as a subjective step in POT analysis (Pan et al., 2022). In the Erzen, a flashy, single-season regime, flood waves rise and drain quickly, so an 11-day window is comfortably longer than the events it separates. In large, slow-responding basins the same formula can return a separation interval too short to keep successive peaks independent.

Two further checks illuminate the composition of the PDS and its relation to the annual maximum series. First, the retained events are

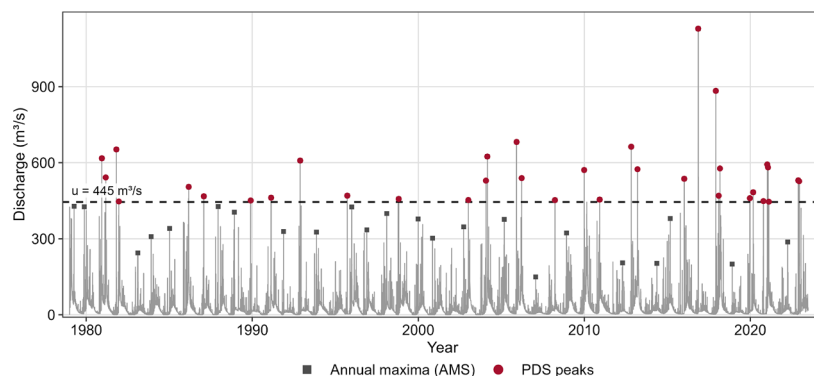


Figure 6. Final partial duration series of the Erzen River at Ndroq ($u = 445 \text{ m}^3/\text{s}$, 11-day separation, recession criterion)

strongly seasonal: 33 of the 34 peaks (97%) occurred between October and March, the only exception occurring in late September. The PDS is thus generated almost exclusively by the autumn–winter frontal rainfall regime, which supports the identical distribution assumption directly. Seasonal concentration does not by itself prove identical distribution; it is corroborating evidence that a single flood-generating population is being sampled, whereas the independence of the retained peaks is enforced by the separation and recession criteria and the stationarity of the magnitudes is tested below. There is no second flood-producing mechanism whose events would need to be modelled separately. This matters because if this assumption fails, there are consequences for the upper tail of the fitted distribution. In watersheds where distinct runoff-generating mechanisms coexist (e.g. snowmelt against convective storms, or tropical against frontal rainfall), the correct treatment is to split the partial duration series by mechanism provenance and then model each sub-population separately and recombine the curves (Breverman et al., 2023). One of the advantages of the POT method is that such a split remains feasible, because each mechanism retains its own events rather than competing for a single annual event as in the AMS. In the case of the Erzen at Ndroq, the seasonal evidence indicates that no split is needed, and the single-population assumption implicit in most flood frequency analyses is well supported for this catchment rather than merely assumed. Second, the overlap between the two samples is revealing: 22 of the 34 events coincide with the annual maximum of their hydrological year, while 12 are secondary peaks that the annual maximum series discards. On the other hand, 23 of the 45 annual maxima (one half of the AMS sample) fall below the threshold and

are thus classified as non-floods by the partial duration series. The two samples, often loosely imagined as “almost the same data”, in fact share only 22 members: half of the AMS and roughly two-thirds of the PDS. The 2021 hydrological year illustrates the mechanism concretely. It contributed four independent floods to the partial duration series, two of them (593 m³/s and 582 m³/s) seventeen days apart in January 2021; the annual maximum series retains only the larger, although the second exceeds the annual maxima of most other years in the record.

Stationarity and independence

Because the generalised Pareto and Poisson models assume stationary flood magnitudes and serially independent peaks, both were tested rather than assumed. The flood magnitudes show no trend: the Mann-Kendall test gives $\tau = 0.036$ ($p = 0.73$) for the 45 annual maxima and $\tau = 0.016$ ($p = 0.91$) for the 34 partial-duration peaks, and Sen’s slope for the annual maxima is +0.8 m³/s per year, about 0.2% of the mean annual maximum and not significant. The Pettitt test detects no change point in the annual maxima ($p = 1.0$). Successive partial-duration peaks are not significantly correlated: the lag-one autocorrelation of the ordered peak magnitudes is 0.29, within the $\pm 2/\sqrt{34} = \pm 0.34$ band expected of white noise (Ljung-Box $p = 0.08$), and the shortest interval between two retained peaks is 17 days, comfortably beyond the 11-day separation. The annual exceedance counts show a weak positive tendency (Mann-Kendall $\tau = 0.28$, $p = 0.02$); this concerns how often the threshold is crossed rather than the magnitude of the floods, the counts remain consistent with a Poisson dispersion (Fig. 5c), and it therefore leaves the magnitude-frequency curve unchanged, though it is noted as

Table 2. Sensitivity of the partial duration series and of the fitted Generalised Pareto to the minimum time separation between peaks (threshold $u = 445$ m³/s), extended to the 100-, 200-, 500- and 1000-year quantiles

Time separation	Events	λ (yr ⁻¹)	Mean excess (m ³ /s)	Scale α (m ³ /s)	Shape κ	Q_{100} (m ³ /s)	Q_{200} (m ³ /s)	Q_{500} (m ³ /s)	Q_{1000} (m ³ /s)
3 days	35	0.787	107.8	75.7	-0.298	1122	1337	1697	2043
11 days (Bulletin 17B)	34	0.764	110.6	80.0	-0.276	1114	1317	1653	1969
21 days	30	0.674	117.2	84.5	-0.279	1121	1330	1677	2004

Note: Location $\xi = u = 445$ m³/s in all cases. Q_{100} is the 1% annual exceedance probability (AEP) quantile (Langbein); Q_{200} , Q_{500} and Q_{1000} are the 0.5%, 0.2% and 0.1% AEP quantiles from the Madsen-equivalent GEV. Across the sevenfold change in the window the 100-year quantile changes by less than 1% and the 1000-year quantile by less than 4%.

a possible non-stationarity of occurrence in the Limitations. Thus, the design-relevant quantity (the flood magnitude) is stationary and the peaks are independent, so the stationary Pareto-Poisson model is appropriate for this record.

Poisson occurrence and generalised Pareto fit

The annual event counts at the final threshold behave like a Poisson process would predict. No event occurred in 44% of the years, exactly one in 40%, two in 11% and three in 4%. These proportions agree closely with the fitted probability mass function with an average rate of $\lambda = 0.76$ (Fig. 7a). In absolute counts, 20, 18, 5, 2 and 0 years recorded 0, 1, 2, 3 and ≥ 4 events, against Poisson expectations of 21.0, 16.0, 6.1, 1.6 and 0.4 years respectively. The dispersion index is $DI = 0.91$, giving $(M - 1) \cdot DI = 40.1$ which, referred to a χ^2 distribution with 44 degrees of freedom (one for each of the 45 years, less one), gives $p = 0.64$. The slight under-dispersion present in the counts (marginally more regular than purely random arrivals) is far from statistical significance, so the data provide no grounds for rejecting the Poisson count model, which is consistent with the threshold screening of Figure 5c; a non-significant test is read as failure to reject the model, not as proof of it. Annual counts slightly more regular than a Poisson process would produce are a common finding where a single wet season delivers a small number of independent floods per year, which is precisely the case in the Erzen basin. In the case of an over-dispersion the Poisson

hypothesis is rejected (Kjeldsen and Prosdociimi, 2023) and negative-binomial occurrence models are the standard remedy (Bhunya et al., 2013), but no such extension is necessary here. The histogram of the peak magnitudes (Fig. 7b) shows the concentration of events near the threshold and the rapidly thinning upper tail, which is the characteristic shape that the GP distribution is designed to describe.

Fitting the PDS composed of 34 elements, to the GP distribution using the method of L-moments gives the following distribution parameters: $\alpha = 80.0 \text{ m}^3/\text{s}$, $\kappa = -0.276$, with the location fixed at $\zeta = u = 445 \text{ m}^3/\text{s}$ (Table 3). In the L-moment convention adopted here (see Methods), this negative shape places the excesses in the heavy-tailed Fréchet domain, equivalently $\xi = +0.276$ in the standard EVT sign convention, both denoting an unbounded power-law tail: the record's behaviour above the threshold is consistent with a power-law tail where occasional events fall far above the typical overshoot, as in the November 2016 flood of $1129 \text{ m}^3/\text{s}$, set against a mean excess of only $111 \text{ m}^3/\text{s}$. The fitted GP quantile function follows the empirical distribution of the events closely over the full probability range (Fig. 8a), and the comparison of the empirical and fitted cumulative distribution functions (Fig. 8b) shows no systematic departure in the upper tail.

The right-tail-weighted Anderson-Darling test at the final threshold yields $A^2 = 0.333$ with a bootstrap p -value of 0.076 ($B = 4999$). The generalised Pareto hypothesis is therefore not rejected at the 5% level. This is the definitive

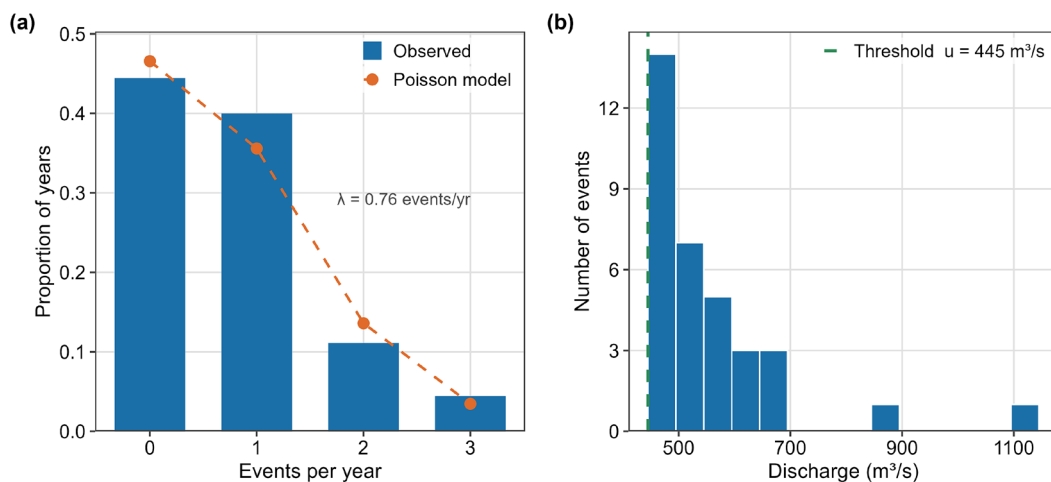


Figure 7. The final partial duration series at $u = 445 \text{ m}^3/\text{s}$: (a) observed proportion of years with 0, 1, 2 and 3 events and the fitted Poisson probability mass function ($\lambda = 0.764$); (b) histogram of the peak magnitudes

test at the selected threshold; the 0.082 shown for 445 m³/s in the scan of Fig. 5b is the same statistic estimated with the coarser $B = 999$ used across the 101 points, and 0.076 is used throughout. Because the statistic weights the upper tail, the test targets exactly the region that governs the flood quantiles; however, this is a moderate rather than a strong acceptance. As mentioned before, the test tends to favour high thresholds, and gives good statistic values only for very high thresholds. In our case, there are specific reasons for this result. First, the value of the statistics is not a failure of the upper tail: Figure 8b shows the empirical and fitted curves separating slightly in the middle of the distribution, among the cluster of moderate excesses just above the threshold that were identified earlier as the events that influence the test. Second, a goodness-of-fit test on 34 observations has little power, so a p -value in this range is weak evidence in either direction; and we can say only that the model is retained. Third, the value is specific to this threshold rather than to the model: at 450 m³/s the same test on the same record returns $p = 0.13$, and the generalised Pareto distribution is not rejected anywhere in the admissible range. Finally, the quantity driving the AMS–PDS divergence is the negative shape parameter, which is not a consequence of the threshold choice: it is negative throughout the admissible range above 390 m³/s. Thus, the generalised Pareto model is retained at the final threshold on evidence that is moderate rather than strong, but stable: the fit is

not rejected anywhere in the admissible range, and the negative shape parameter that drives the AMS–PDS divergence persists across it.

Annualisation, frequency curves, quantile comparison and uncertainty

Using the Langbein transformation, the 34 peaks of the PDS are plotted on an annual probability scale (non-exceedance) against the annual maxima series (Fig. 9a). Comparing the two series, the largest events plot identically in both series, since the top of the two samples is the same set of floods. In the intermediate probability range, however, the partial duration series points lie systematically above the annual maxima. In the lower half, by contrast, the AMS is populated by the sub-threshold maxima of dry years, which pull the empirical AMS curve downward. The PDS stops at the threshold value of 445 m³/s, and contains no elements below it.

The GEV distribution is fitted to the 45 annual maxima, giving the following parameters: $\zeta = 372.4$ m³/s, $\alpha = 137.5$ m³/s and $\kappa = 0.007$. The shape parameter is statistically indistinguishable from zero, so the AMS fit reduces to a Gumbel distribution, characterised by its exponentially decaying tail. Using the Madsen conversion of the Pareto–Poisson model, Equations (22)–(24), yield the equivalent GEV parameters $\zeta^* = 424.3$ m³/s, $\alpha^* = 74.3$ m³/s and $\kappa^* = -0.276$. Here the two analyses of the same record part ways in the most consequential respect: the AMS sample suggests an exponential upper tail, while the

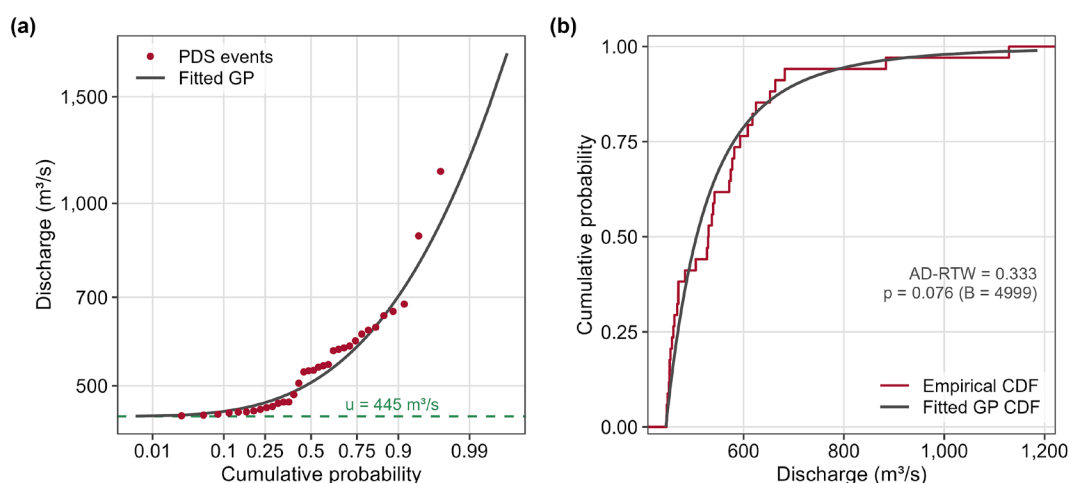


Figure 8. (a) Probability plot of the partial duration series (Weibull plotting positions) with the fitted Generalised Pareto quantile function. (b) Empirical cumulative distribution of the 34 peaks and fitted generalised Pareto distribution, with the right-tail-weighted Anderson-Darling statistic and its parametric-bootstrap p -value

Table 3. Summary of the final partial duration series model and of the annual maximum series model: threshold, separation, number of events, event rate, mean excess, generalised Pareto parameters (ξ , α , κ), equivalent GEV parameters (ξ^* , α^* , κ^*) from Equations (20)–(24), and GEV parameters fitted to the annual maxima

Model	ξ (m ³ /s)	α (m ³ /s)	κ
Generalised Pareto, PDS excesses ($u = 445$)	445.0	80.0	-0.276
Equivalent GEV, Madsen (Eqs. 22–24)	424.3	74.3	-0.276
GEV, annual maxima	372.4	137.5	0.007

Note: PDS series: $n = 34$ events, $\lambda = 0.764$ yr⁻¹, mean excess = 110.6 m³/s, separation $d = 11$ days. AMS: $N = 45$ water years. Madsen parameters are the equivalent ξ^* , α^* , κ^* ($\kappa^* = \kappa$ by construction).

PDS approach identifies a heavy, power-law-type tail distribution. The shape parameter controls how quickly quantiles grow with return period. Everything that follows, including the divergence of the two curves and the difference in design values, comes from this one disagreement. Its origin is examined below. Figure 9b compares the two

fitted frequency curves on the annual probability scale, and Figure 9c extends the comparison to the 1000-year return period with the 90% parametric-bootstrap confidence bands.

Annual quantiles of both models for the standard return periods with their 90% confidence intervals are given in Table 4. The two frequency curves are practically coincident between the 5- and 25-year return periods, with differences lower than 4%. As a practical result, we can say that for frequent design events the choice of the method, whether an annual maxima approach or a peak over threshold approach is immaterial, at least for this watershed. At $T=25$ years the two curves cross and then diverge monotonically: the partial duration series quantiles exceed the AMS quantiles by 5% at $T = 50$, by 12% at $T = 100$ (1114 m³/s against 996 m³/s), by 21% at $T = 200$ and by 51% at $T = 1000$ (1969 m³/s against 1301 m³/s). Two reference points put these numbers in perspective. At the 100-year level, the partial duration series estimate of 1114 m³/s practically reproduces the largest discharge actually observed in the 44.5-year record (1129 m³/s, November 2016), to which it assigns a return period

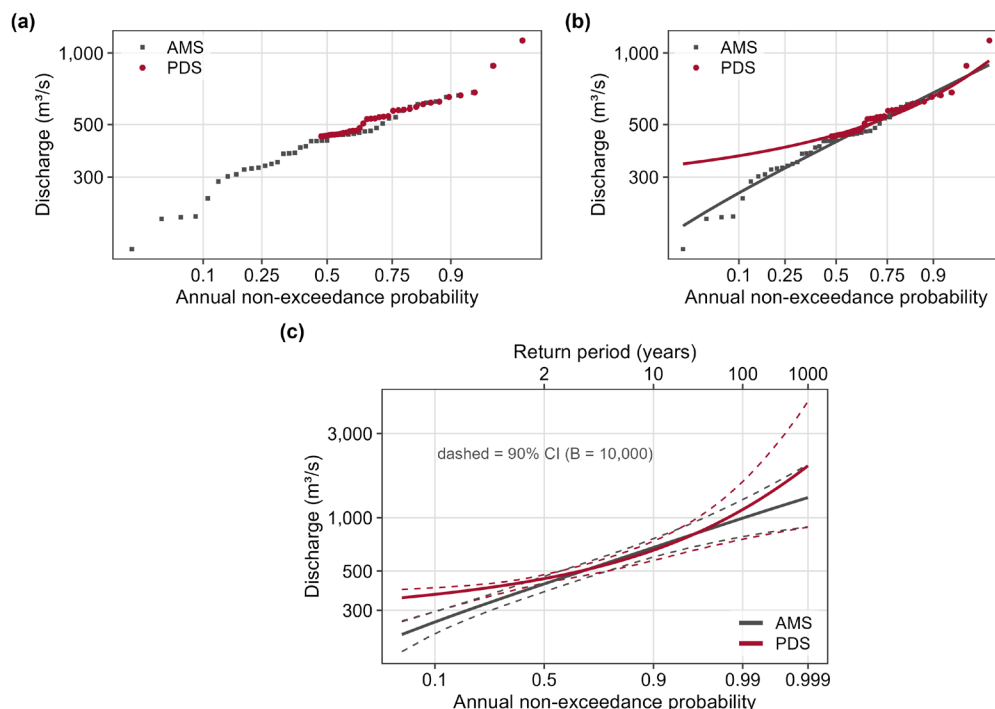


Figure 9. (a) Annual maxima and Langbein-transformed partial duration series peaks on the annual non-exceedance probability scale. (b) Frequency curves of the Erzen River at Ndroq: GEV fitted to the annual maximum series and equivalent GEV derived from the Pareto–Poisson model of the partial duration series. (c) Frequency curves extrapolated to the 1000-year return period with 90% parametric-bootstrap confidence bands ($B = 10,000$)

of 106 years, whereas the AMS curve places the same event at 272 years. Directly, the 1129 m³/s maximum has a non-exceedance probability of 0.9906 under the PDS/Madsen GEV and 0.9963 under the AMS GEV, i.e. return periods of 106 and 272 years. This looks like an uncomfortable result for a discharge that was recorded within nearly 45 years. At the 1000-year level, relevant to safety-check floods, the AMS estimate of 1301 m³/s exceeds the once-observed 2016 flood by barely 15%. The confidence intervals of the two models overlap at all return periods, but they behave differently. Both are asymmetric (longer upward than downward, a generic property of heavy-tailed distributions). The PDS intervals widen much faster (756–1603 m³/s at $T = 100$ and 885–4557 m³/s at $T = 1000$, against 782–1269 and 888–2007 m³/s for the AMS), reflecting the propagation of uncertainty in a heavy shape parameter estimated from only 34 events.

It is important to understand the causes of divergence between the two frequency curves, since this helps us decide which curve deserves belief. Under the Pareto–Poisson model, the shape parameter of the excess distribution is inherited exactly by the annual maxima ($\kappa^* = \kappa$, Equation (24)). If the flood process truly has a heavy tail with $\kappa \approx -0.28$, the annual maxima drawn from that process should display the same shape. However, the GEV fitted to the actual annual maxima instead returns $\kappa \approx 0.01$, i.e. a Gumbel-type tail. Thus, the two estimates cannot both reflect the flood process, and the composition of the AMS sample explains which one does not. One half of the AMS (23 of 45 values) consists of maxima below 445 m³/s. These are dry-year maxima that the

threshold analysis classifies as non-floods. While these values can be informative about ordinary wet-season hydrology, they are not very informative about floods, yet they enter the GEV fit with the same weight as the major events. Their effect is systematic: they stretch the range of the sample downward, inflate its L-scale relative to its upper portion, and compress the sample L-skewness. The L-skewness alone determines the fitted GEV shape through Equation (8). The lighter tail of the AMS fit is thus a direct arithmetic consequence of averaging floods with non-floods. This is precisely the distortion noted in the Introduction, where flows of years in which no flood occurred influence the estimate of the largest floods, and it is the same phenomenon that United States guidance treats by censoring potentially influential low floods before fitting (England et al., 2019). The POT approach removes the distortion by construction, without any censoring test: everything below the threshold is simply outside the sample, and the tail is estimated from floods alone.

The pattern and magnitude of the divergence are consistent with the published literature. In the systematic comparisons of Bezak et al. (2014) and Karim et al. (2017), the PDS/GP model reproduced heavy-tailed design events better than the AMS/GEV alternative. Pan and Rahman (2022) reached the same conclusion. Likewise, Kjeldsen and Prosdocimi (2023) found the GEV distribution unsuitable for annual maxima across much of Britain, precisely because observed flood tails were heavier than the fitted GEV shapes implied. The Erzen result adds a quantitative Albanian data point to this pattern, and its practical implication is serious: the divergence emerges at exactly the

Table 4. Discharge quantiles (m³/s) of the annual maximum series model and of the partial duration series model with 90% parametric-bootstrap confidence intervals

T (yr)	AEP	AMS GEV (90% CI)	PDS Madsen (90% CI)	Diff
2	0.500	423 (382–467)	453 (427–477)	+7.1%
5	0.200	578 (518–638)	562 (512–614)	–2.7%
10	0.100	680 (600–763)	656 (573–738)	–3.5%
25	0.040	808 (685–941)	806 (654–968)	–0.2%
50	0.020	902 (737–1096)	946 (707–1228)	+4.8%
100	0.010	996 (782–1269)	1114 (756–1603)	+11.9%
200	0.005	1088 (821–1463)	1317 (799–2145)	+21.0%
500	0.002	1210 (862–1754)	1653 (850–3256)	+36.6%
1000	0.001	1301 (888–2007)	1969 (885–4557)	+51.3%

Note: Quantiles in m³/s; CIs are 90% parametric-bootstrap ($B = 10,000$). The two curves cross near $T = 25$ yr; the PDS curve is higher for rarer events, reflecting its heavier tail ($\kappa = -0.276$ vs 0.007).

return periods used to size major infrastructure. At the 100-year level, the difference of 118 m³/s is of the same order as the freeboard allowances typically provided for levees, embankments, and bank revetments, and the largest observed event (1129 m³/s in November 2016) has an annual exceedance probability of 0.95% under the PDS curve but of 0.37% under the AMS curve. Encountering such an event within a 44.5-year window is an unremarkable outcome under the PDS curve and considerably less likely under the AMS curve. Over a 44.5-year record the expected number of such exceedances is about 0.42 under the PDS curve and about 0.16 under the AMS curve, so observing one is an ordinary outcome under the former and comparatively unlikely under the latter. The partial duration description is thus more consistent with the single direct piece of evidence the record provides about its own upper tail. This is a statement of statistical consistency with the observed series, not a claim that a particular return period is preferable.

The two curves cross below $T \approx 25$ years, with the PDS curve marginally lower in the 5–25-year range. This is a secondary but genuine feature. The AMS curve is lifted in this range for the same reason that its tail is depressed. A single near-Gumbel shape parameter must accommodate the sub-threshold maxima at one end and the major floods at the other. The fit therefore overshoots slightly in the middle. The PDS curve, by contrast, is anchored by the threshold near the mean annual flood and shaped by the floods alone. It therefore sits marginally lower in this range. For frequent design events ($T \leq 25$ years) the two approaches are operationally interchangeable in this watershed. The choice of method matters exactly where extrapolation begins.

The bootstrap intervals of Table 4 are indicative of caution in overinterpreting either curve at long return periods. The POT method is usually regarded in literature as more representative. In our case, however, its wider confidence intervals at long return periods appear to contradict this. Three reasons explain these wider intervals. First, the width of the PDS interval reflects a real uncertainty. If the flood tail is genuinely heavy, the AMS analysis does not escape that uncertainty so much as understate it. The AMS analysis does not escape that uncertainty; it merely understates it. A heavy shape parameter estimated from 34 events is genuinely uncertain, and the Madsen conversion amplifies it as it propagates into the

tail. The narrower AMS interval may be a consequence of the near-zero shape on which it is built, which the inclusion of sub-threshold maxima can bias towards zero, a form of model misspecification. Should that be the case, a narrower interval around a misspecified curve would not be more informative for being narrow, and the stated 90% confidence would not be attained. Second, design decisions are driven by the central estimates, not by the interval widths. The two central curves differ materially from $T = 50$ upwards. Because the intervals overlap, that difference cannot be declared statistically significant at the 90% level. Design practice, however, requires a single design value. Where two curves compete, the better-founded one must be adopted rather than a compromise between them. Third, both intervals derive from a 44.5-year record extrapolated to return periods more than twenty times longer. At that range, no statistical device substitutes for more information: longer records, historical flood marks, or regional pooling. The 1000-year estimates in particular (1969 m³/s for the PDS model and 1301 m³/s for the AMS model) rest on extrapolating a 44.5-year record by more than a factor of twenty and on a shape parameter estimated from 34 events; their sensitivity to the threshold (Table 5) is large, and they should be read as indications of tail behaviour rather than as reliable design values. The regional index-flood extension of the PDS framework is therefore the natural next step for the Albanian network context (Mostofi Zadeh et al., 2019; Kjeldsen and Prosdoci, 2023).

Sensitivity of the design quantiles to the threshold

Because the threshold is the most consequential and most subjective choice in the method, the sensitivity of the fitted model and of the design quantiles to it was quantified directly by refitting the entire model at ten thresholds spanning the credible range, 400–480 m³/s (Table 5). The 100-year quantile – the value most relevant to design – is robust: across the defensible neighbourhood 430–460 m³/s it varies between 1030 and 1114 m³/s (about 8%), and the partial duration series estimate exceeds the annual-maximum estimate of 996 m³/s across essentially the whole range. Above 460 m³/s the sample falls below 26 events, where the shape parameter turns positive and the estimates become unreliable, as the shape and mean-excess plots already indicated. The

Table 5. Sensitivity of the fitted Generalised Pareto model and of the design quantiles to the threshold (11-day separation; Q from the Madsen-equivalent GEV)

u (m ³ /s)	Events	λ (yr ⁻¹)	α (m ³ /s)	κ	Q100	Q200	Q500	Q1000
400	47	1.056	111.6	-0.060	999	1103	1248	1363
410	44	0.989	105.1	-0.098	1019	1138	1308	1447
420	43	0.966	85.1	-0.221	1091	1266	1543	1792
430	37	0.832	97.5	-0.160	1056	1201	1419	1606
440	35	0.787	86.1	-0.234	1093	1273	1561	1824
445	34	0.764	80.0	-0.276	1114	1317	1653	1969
450	31	0.697	89.5	-0.228	1090	1268	1551	1807
460	26	0.584	116.3	-0.089	1030	1150	1320	1458
470	22	0.495	155.7	+0.112	962	1029	1111	1167
480	21	0.472	148.1	+0.083	968	1040	1130	1193

1000-year quantile, by contrast, is markedly more sensitive to the threshold (from about 1170 to 1970 m³/s), reflecting the inevitable amplification of shape-parameter uncertainty under extrapolation; this is the reason the millennial estimate is treated as highly uncertain above. The selected threshold of 445 m³/s was fixed by the diagnostic plots, not by the resulting quantile; that it also carries the most negative shape reflects that this is where the flood-only sample is least contaminated by sub-threshold discharges. The recession criterion removed no events at any of these thresholds.

CONCLUSIONS

A peaks-over-threshold flood frequency analysis was applied to a 44.5-year daily discharge series for the Erzen River at Ndroq, Albania. Six diagnostic plots identified a computational threshold of 445 m³/s. The resulting partial duration series contains 34 independent peaks, at a rate of 0.76 events per year, and these fall almost entirely in autumn and winter. Their annual counts follow a Poisson process, and their magnitudes follow a generalised Pareto distribution with a heavy upper tail ($\kappa = -0.276$). The design quantiles are robust to both subjective steps of the method. The 1% quantile changes by less than 2% when either the declustering window or the threshold is varied, although the rarer 0.1% quantile is more sensitive. The 100-year design quantile is robust to the two subjective steps of the method: it changes by less than 1% across declustering windows of 3 to 21 days and by about 8% across the credible threshold range 430–460 m³/s, remaining above the annual-maximum estimate throughout. The

1000-year quantile is considerably more sensitive to both choices and is correspondingly uncertain.

The partial duration and annual maximum series agree up to the 25-year return period, but diverge beyond it. At the 1% annual exceedance level the partial duration series gives 1114 m³/s against 996 m³/s for the annual maximum series, a 12% difference that grows to about 51% at the 1000-year level. The cause is sample composition. Half of the annual maxima fall below the threshold and represent dry-year discharges rather than floods. These low values compress the sample L-skewness and force a near-Gumbel tail, which contradicts the heavy tail measured on the flood peaks themselves. The record confirms this: the largest observed flood (1129 m³/s, November 2016) is an ordinary once-per-century event under the partial duration series, but a far rarer one under the annual maximum series. This discrepancy emerges precisely at the return periods used to size hydraulic structures, so the choice of method is decisive for design rather than academic.

For the Erzen River at Ndroq, and within the limits of a single 44.5-year record, the peaks-over-threshold description of the flood tail is more internally consistent with the observed flood peaks than the annual-maximum fit, because it estimates the tail from floods alone rather than from a sample half of which lies below the flood threshold. Its advantage at the most extreme return periods nevertheless remains uncertain given the shortness of the record. The central conclusion is therefore best framed not as which model yields the larger millennial quantile, but as which model has the stronger statistical and hydrological support for this series, here, the

Pareto-Poisson model for the upper tail, stated together with the uncertainty that a single short record entails. The natural next step is regional index-flood pooling across the Albanian gauging network, which would reduce the uncertainty that no single-site record can overcome. These results can be relevant to researchers and engineers in the region studying the Erzen River and other Albanian rivers of similar characteristics.

REFERENCES

1. AMBU (Agjencia e Menaxhimit të Burimeve Ujore) (2023). *Plani i Menaxhimit të Basenit Ujor për Basenin e Lumit Erzen, Raporti Kryesor* [River basin management plan for the Erzen River basin, Main report] (in Albanian). Drafti përfundimtar, Shkurt 2023. Agjencia e Menaxhimit të Burimeve Ujore, Tiranë, 211 pp. https://konsultimipublik.gov.al/documents/RENJK_607_PMBU_Erzen.pdf
2. Asquith, W.H. (2022). *lmomco — L-moments, censored L-moments, trimmed L-moments, L-comoments, and many distributions* [Computer software, R package]. <https://CRAN.R-project.org/package=lmomco>
3. Balkema, A.A., & de Haan, L. (1974). Residual life time at great age. *The Annals of Probability*, 2(5), 792–804. <https://doi.org/10.1214/aop/1176996548>
4. 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(5), 959–977. <https://doi.org/10.1080/02626667.2013.831174>
5. Bhunya, P.K., Berndtsson, R., Jain, S.K., & Kumar, R. (2013). Flood analysis using negative binomial and Generalized Pareto models in partial duration series (PDS). *Journal of Hydrology*, 497, 121–132. <https://doi.org/10.1016/j.jhydrol.2013.05.047>
6. Breverman, A., Bartles, M., Karlovits, G., & Arabi, M. (2023). Addressing mixed populations in flood frequency analyses: A case study in eastern Pennsylvania. *SEDHYD 2023 — Sedimentation and Hydrologic Modeling Conference*, St. Louis, MO. https://www.hec.usace.army.mil/confluence/sspdocs/ssptr/files/105580369/213684392/1/1728072320244/SEDHYD_Breverman_Bartles_Karlovits_Mixed_Populations_Lehigh_River.pdf
7. Choulakian, V., & Stephens, M.A. (2001). Goodness-of-fit tests for the generalized Pareto distribution. *Technometrics*, 43(4), 478–484. <https://doi.org/10.1198/00401700152672573>
8. Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer. <https://doi.org/10.1007/978-1-4471-3675-0>
9. Efron, B., & Tibshirani, R.J. (1993). *An introduction to the bootstrap*. Chapman & Hall. <https://doi.org/10.1007/978-1-4899-4541-9>
10. England, J.F., Cohn, T.A., Faber, B.A., Stedinger, J.R., Thomas, W.O., Veilleux, A.G., Kiang, J.E., & Mason, R.R. (2019). *Guidelines for determining flood flow frequency — Bulletin 17C (Techniques and Methods 4-B5)*. U.S. Geological Survey. <https://doi.org/10.3133/tm4B5>
11. European Space Agency (2024). *Copernicus DEM — Global and European Digital Elevation Model, GLO-30 (COP-DEM_GLO-30-DGED, 2023_1)* [Data set]. European Space Agency. <https://doi.org/10.5270/ESA-c5d3d65>
12. Fischer, S., Lun, D., Schumann, A.H., & Blöschl, G. (2023). Detecting flood-type-specific flood-rich and flood-poor periods in peaks-over-threshold series with application to Bavaria (Germany). *Stochastic Environmental Research and Risk Assessment*, 37(4), 1395–1413. <https://doi.org/10.1007/s00477-022-02350-8>
13. 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(2), 180–190. <https://doi.org/10.1017/S0305004100015681>
14. Gnedenko, B. (1943). Sur la distribution limite du terme maximum d’une série aléatoire [On the limiting distribution of the maximum term of a random series]. *Annals of Mathematics*, 44(3), 423–453 (in French). <https://doi.org/10.2307/1968974>
15. Gumbel, E.J. (1958). *Statistics of extremes*. Columbia University Press. <https://doi.org/10.7312/gumb92958>
16. 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 B*, 52(1), 105–124. <https://doi.org/10.1111/j.2517-6161.1990.tb01775.x>
17. Hosking, J.R.M. (2022). *lmom — L-moments* [Computer software, R package]. <https://CRAN.R-project.org/package=lmom>
18. Hosking, J.R.M., & Wallis, J.R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, 29(3), 339–349. <https://doi.org/10.1080/00401706.1987.10488243>
19. Hosking, J.R.M., & Wallis, J.R. (1997). *Regional frequency analysis: An approach based on L-moments*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511529443>
20. 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(3), 251–261. <https://doi.org/10.1080/00401706.1985.10488049>
21. IACWD (1982). *Guidelines for determining flood flow frequency — Bulletin 17B*. Interagency Advisory Committee on Water Data, Hydrology Subcommittee, U.S. Geological Survey. <https://doi.org/10.3133/70275162>
 22. IFRC (2018). *Albania: Floods — Emergency plan of action final report* (MDRAL007). International Federation of Red Cross and Red Crescent Societies. <https://reliefweb.int/report/albania/albania-floods-emergency-plan-action-final-report-operation-n-mdral007>
 23. Jenkinson, A.F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, 81(348), 158–171. <https://doi.org/10.1002/qj.49708134804>
 24. Karim, F., Hasan, M., & Marvanek, S. (2017). Evaluating annual maximum and partial duration series for estimating frequency of small magnitude floods. *Water*, 9(7), 481. <https://doi.org/10.3390/w9070481>
 25. Kjeldsen, T.R., & Prosdocimi, I. (2023). Use of peak over threshold data for flood frequency estimation: An application at the UK national scale. *Journal of Hydrology*, 626, 130235. <https://doi.org/10.1016/j.jhydrol.2023.130235>
 26. Klemeš, V. (1986). Dilettantism in hydrology: Transition or destiny? *Water Resources Research*, 22(9S), 177S–188S. <https://doi.org/10.1029/WR022i09Sp0177S>
 27. Kolaković, S., Mandić, V., Stojković, M., Jeftenić, G., Stipić, D., & Kolaković, S. (2023). Estimation of large river design floods using the peaks-over-threshold (POT) method. *Sustainability*, 15(6), 5573. <https://doi.org/10.3390/su15065573>
 28. Lang, M., Ouarda, T.B.M.J., & Bobée, B. (1999). Towards operational guidelines for over-threshold modeling. *Journal of Hydrology*, 225(3–4), 103–117. [https://doi.org/10.1016/S0022-1694\(99\)00167-5](https://doi.org/10.1016/S0022-1694(99)00167-5)
 29. Langbein, W.B. (1949). Annual floods and the partial-duration flood series. *Eos, Transactions American Geophysical Union*, 30(6), 879–881. <https://doi.org/10.1029/TR030i006p00879>
 30. Langousis, A., Mamalakis, A., Puliga, M., & Deidda, R. (2016). Threshold detection for the generalized Pareto distribution: Review of representative methods and application to the NOAA NCDC daily rainfall database. *Water Resources Research*, 52(4), 2659–2681. <https://doi.org/10.1002/2015WR018502>
 31. 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(4), 747–757. <https://doi.org/10.1029/96WR03848>
 32. Mostofi Zadeh, S., Durocher, M., Burn, D.H., & Ashkar, F. (2019). Pooled flood frequency analysis: A comparison based on peaks-over-threshold and annual maximum series. *Hydrological Sciences Journal*, 64(2), 121–136. <https://doi.org/10.1080/02626667.2019.1577556>
 33. Nerantzaki, S.D., & Papalexiou, S.M. (2022). Assessing extremes in hydroclimatology: A review on probabilistic methods. *Journal of Hydrology*, 605, 127302. <https://doi.org/10.1016/j.jhydrol.2021.127302>
 34. Pan, X., & Rahman, A. (2022). Comparison of annual maximum and peaks-over-threshold methods with automated threshold selection in flood frequency analysis: A case study for Australia. *Natural Hazards*, 111(2), 1219–1244. <https://doi.org/10.1007/s11069-021-05092-y>
 35. Pan, X., Rahman, A., Haddad, K., & Ouarda, T.B.M.J. (2022). Peaks-over-threshold model in flood frequency analysis: A scoping review. *Stochastic Environmental Research and Risk Assessment*, 36(9), 2419–2435. <https://doi.org/10.1007/s00477-022-02174-6>
 36. Pano, N. (2015). *Pasuritë ujore të Shqipërisë* [Water resources of Albania] (in Albanian). Akademia e Shkencave e Shqipërisë (printed by Kristalina-KH), Tiranë, 634 pp. ISBN 978-99956-10-87-6.
 37. Pickands, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1), 119–131. <https://doi.org/10.1214/aos/1176343003>
 38. R Core Team (2022). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
 39. Ribatet, M., & Dutang, C. (2022). *POT: Generalized Pareto distribution and peaks over threshold* [Computer software, R package]. <https://CRAN.R-project.org/package=POT>
 40. Robson, A., & Reed, D. (1999). *Flood estimation handbook, Volume 3: Statistical procedures for flood frequency estimation*. Institute of Hydrology.
 41. Salas, J.D., Obeysekera, J., & Vogel, R.M. (2018). Techniques for assessing water infrastructure for nonstationary extreme events: A review. *Hydrological Sciences Journal*, 63(3), 325–352. <https://doi.org/10.1080/02626667.2018.1426858>
 42. Scarrott, C., & MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT — Statistical Journal*, 10(1), 33–60. <https://doi.org/10.57805/revstat.v10i1.110>
 43. Schendel, T., & Thongwichian, R. (2015). Flood frequency analysis: Confidence interval estimation by test inversion bootstrapping. *Advances in*

- Water Resources*, 83, 1–9. <https://doi.org/10.1016/j.advwatres.2015.05.004>
44. Sinclair, C.D., Spurr, B.D., & Ahmad, M.I. (1990). Modified Anderson Darling test. *Communications in Statistics – Theory and Methods*, 19(10), 3677–3686. <https://doi.org/10.1080/03610929008830405>
 45. Solari, S., Egüen, M., Polo, M.J., & Losada, M.A. (2017). Peaks over threshold (POT): A methodology for automatic threshold estimation using goodness of fit p-value. *Water Resources Research*, 53(4), 2833–2849. <https://doi.org/10.1002/2016WR019426>
 46. Stedinger, J.R., Vogel, R.M., & Foufoula-Georgiou, E. (1993). Frequency analysis of extreme events. In D.R. Maidment (Ed.), *Handbook of hydrology* (Chapter 18). McGraw-Hill.
 47. Takeuchi, K. (1984). Annual maximum series and partial-duration series — Evaluation of Langbein’s formula and Chow’s discussion. *Journal of Hydrology*, 68(1–4), 275–284. [https://doi.org/10.1016/0022-1694\(84\)90148-6](https://doi.org/10.1016/0022-1694(84)90148-6)
 48. USACE (1993). *Hydrologic frequency analysis* (Engineer Manual EM 1110-2-1415). U.S. Army Corps of Engineers. https://www.publications.usace.army.mil/portals/76/publications/engineermanuals/em_1110-2-1415.pdf
 49. Van Campenhout, J., Houbrechts, G., Peeters, A., & Petit, F. (2020). Return period of characteristic discharges from the comparison between partial duration and annual series, application to the Walloon rivers (Belgium). *Water*, 12(3), 792. <https://doi.org/10.3390/w12030792>