EGU General Assembly 2019 Vienna | Austria | 7–12 April 2019

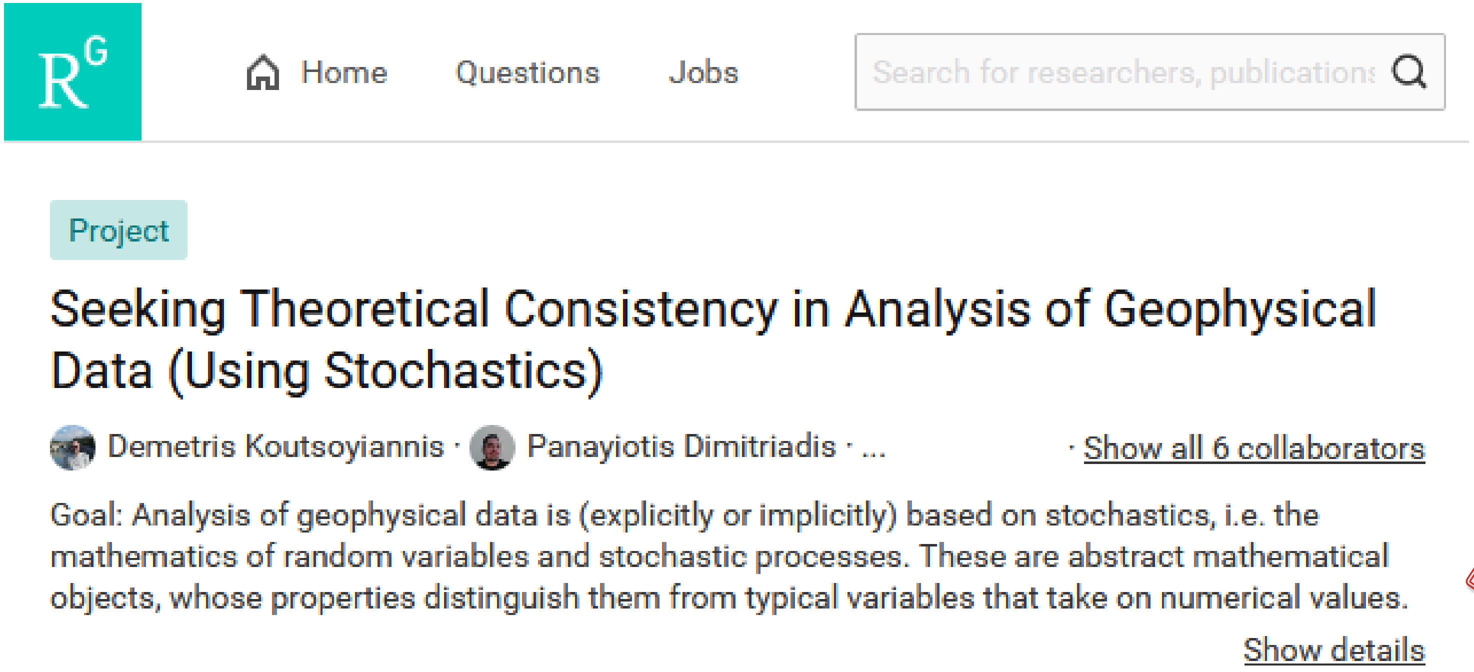
Session HS3.2: Spatio-temporal and/or (geo)statistical analysis of hydrological events, floods, extremes, and related hazards

Extreme-oriented selection and fitting of probability distributions

Demetris Koutsoyiannis
School of Civil Engineering
National Technical University of Athens, Greece
(dk@itia.ntua.gr, http://www.itia.ntua.gr/dk/)

Presentation available online: http://www.itia.ntua.gr/1942/

The general framework: Seeking theoretical consistency in analysis of geophysical data (Using *stochastics*)



Seeking Theoretical Consistency in Analysis of Geophysical Data (Using Stochastics)

Demetris Koutsoyiannis · Panayiotis Dimitriadis · ... · Show all 6 collaborators

Goal: Analysis of geophysical data is (explicitly or implicitly) based on stochastics, i.e. the mathematics of random variables and stochastic processes. These are abstract mathematical objects, whose properties distinguish them from typical variables that take on numerical values.

Show details

Fully Unfunded

Book in preparation:

D. Koutsoyiannis, Stochastics of Hydroclimatic Extremes – A Cool Look at Risk (2020)

Data for illustration

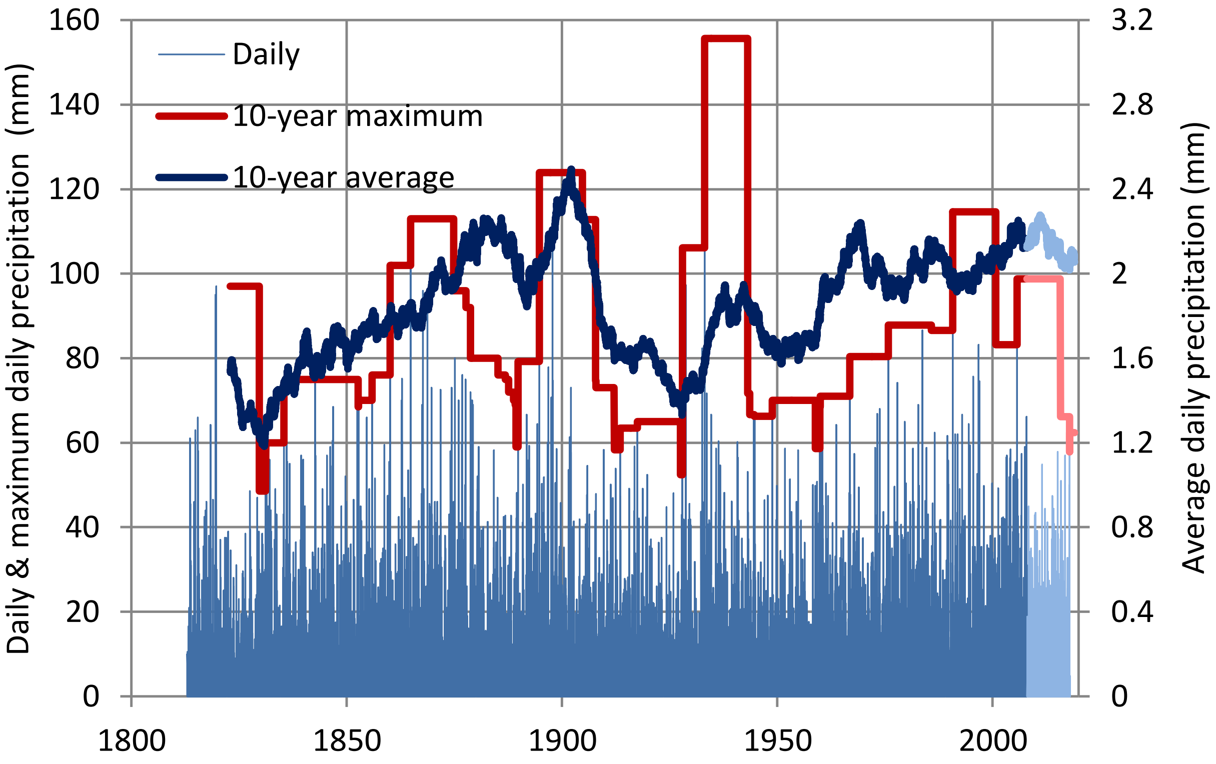
Bologna, Italy (44.50°N, 11.35°E, +53.0 m).

Available from the Global Historical Climatology Network (GHCN) - Daily.

Uninterrupted for the period 1813-2007: 195 years.

For the most recent period, 2008-2018 daily data are provided by the repository Dext3r of ARPA Emilia Romagna.

Total record length: 206 years.



Main observation:
The 10-year climatic variables have varied irregularly by a factor of 2 for the average daily precipitation and by a factor > 3 for the maximum daily precipitation.

Are “nonstationary” analyses and trend identification useful or necessary?

Author’s opinion: Such analyses are both fashionable and funny—but of little scientific value.

Block maxima, values over threshold, or all data?

Traditionally, hydrometeorological records are analysed in two ways:

- *Block maxima*—typically *annual maxima* (most frequent). We extract the highest of all recorded values for a given time period (typically year) and form a statistical sample with size equal to the number of blocks (typically years) in the record.
- *Values over threshold* (VOT, aka peaks-over-threshold—POT). We form a sample of values exceeding a certain threshold irrespective of the time they occurred. Usually the threshold is chosen so that the sample size is again equal to the number of years of the record.

At first glance, the block maxima option corresponds to the extreme value theory and the resulting limiting distributions. In fact, however, the latter are only approximations as the convergence to the limit is very slow. What is useful to keep from the extreme value theory is this:

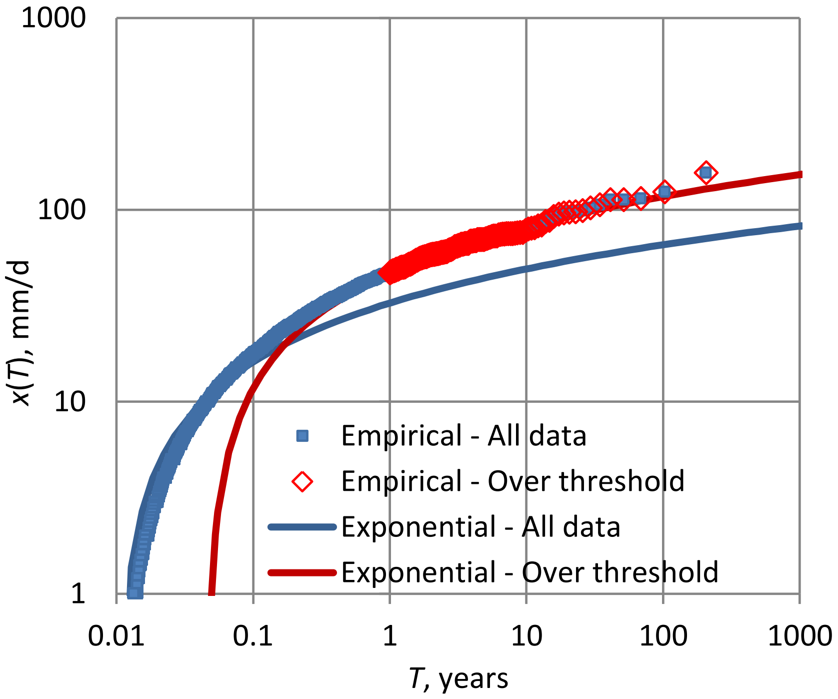
- If the limiting distribution of maxima is Gumbel, then the tail of the parent distribution is exponential.
- If the limiting distribution of maxima is type II, then the tail of the parent distribution is Pareto.

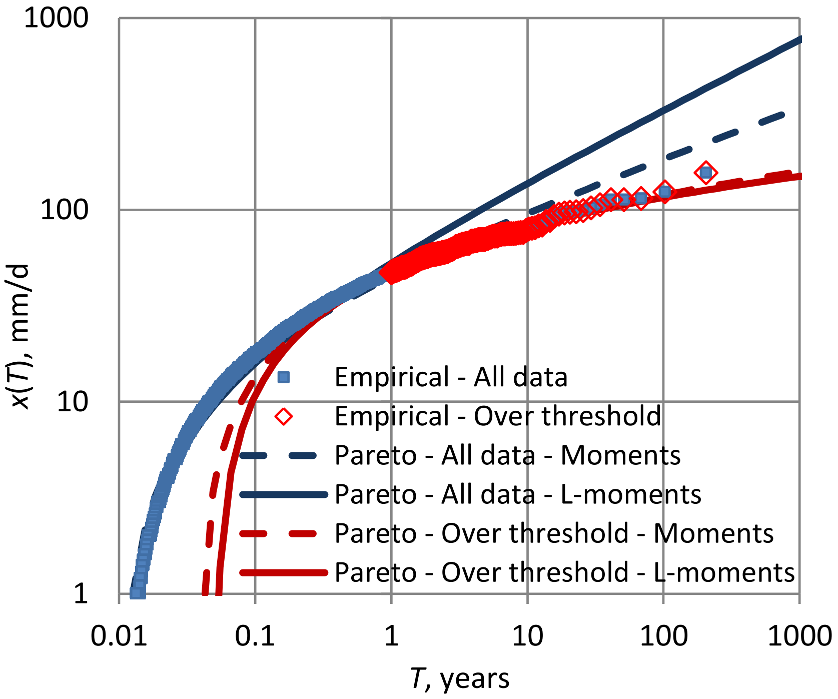
Today there is abundance of hydrometeorological data on daily and sub-daily scales and there is no need to extract annual or seasonal maxima. Thus, it is preferable to use the entire observational record; second best option is the VOT. Only if the available observations are originally given for time-blocks (e.g., annual maxima), it is justified to refer to extreme value distribution.

Studying the complete series of observations has the advantage of not wasting information, in accord to the motto "*Save hydrological observations!*" (Volpi et al., 2019).

A final advantage is that the design quantities do correspond to the parent distribution, rather than the artificially induced maxima over an arbitrarily defined time period.

Entire data set or values above threshold?





The graph shows the fitting of 2 theoretical distributions:

- exponential (with 1 parameter for all data and 2 for VOT),
- Pareto (with 2 parameters for all data and 3 for VOT).

Two samples of daily rainfall in Bologna were used, namely:

- the sample of all nonzero values (size: 19426 for 206 years – 94.3 rain days per year on average),
- the sample above a threshold of 47 mm (VOT, size: 206).

The plots of empirical distributions were based on order statistics (see p. 8) with $T\left(x_{(i)}\right) = \frac{n+1}{n+1-i} d$, where $x_{(i)}$ is the ith smallest value of the sample of size n and d is a time unit.

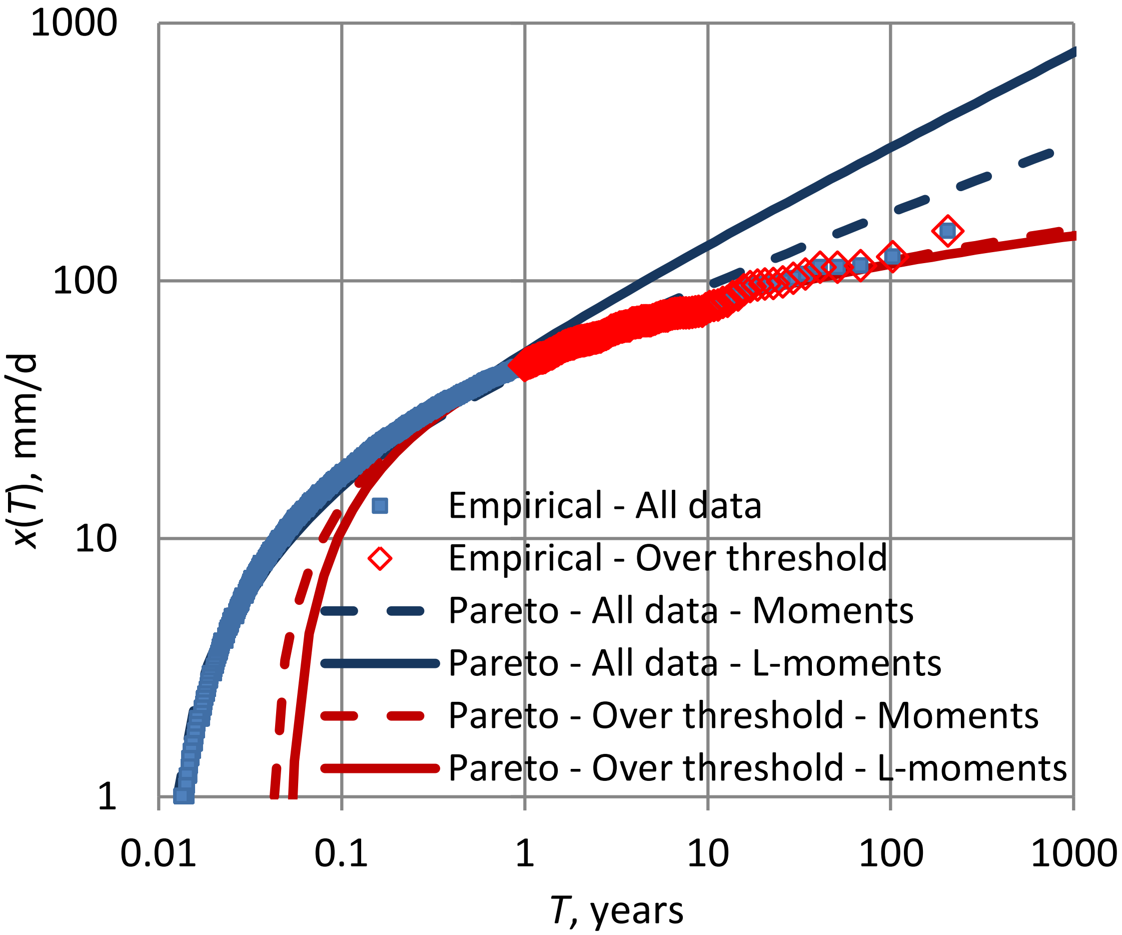
Clearly, the second option (VOT) gives a better fitting on the maxima—but at the expense of an additional parameter and an unrealistic nonzero minimum. Can we improve the first option?

Classical moments or L-moments?

- We fit to all data a 2-parameter Pareto distribution with zero lower bound:

$$F(x) = 1 - (1 + \kappa\, x/\lambda)^{-1/\kappa}$$

- To estimate the two parameters we need two moments (or more?).
- The fitting (blue lines) is quite unsatisfactory for the distribution tail (extremes), with the classical moments showing better performance than the L-moments.
- If we used the sample over threshold (red lines) and a 3-parameter Pareto, classical moments and L-moments give fittings very close to each other (with slight advantage of the latter on both small and high values).



Questions

1. How can we use the entire data set and fit on the distribution tail?
2. How can we compare (validate) the distribution fitting? Any means better that order statistics?
3. How many moments do we need for fitting and comparison? Of what order and of which type?

Classical, L-, or Probability Weighted Moments?

- If we need to go to moment order higher than 2 to 3, we should exclude classical moments (cf. Lombardo et al., 2014: "*Just two moments*") because they are *unknowable* (Koutsoyiannis, 2019).
- L-moments are popular and have unbiased estimators for high orders, and thus are preferable.
- However, L-moments are connected to Probability Weighted Moments via one-to-one relationships; the latter are more directly defined and estimated and therefore are preferable. (In fact the estimation of L-moments is made from that of Probability Weighted Moments).
- Definition of Probability Weighted Moments:

$$\beta_p \coloneqq \mathrm{E}\left[\underline{x}\left(F(\underline{x})\right)^p\right]$$

- Relationships of Probability Weighted Moments and L-moments (for the first four orders):

$$\lambda_1 = \beta_0 = \mu, \qquad \lambda_2 = 2\beta_1 - \beta_0, \qquad \lambda_3 = 6\beta_2 - 6\beta_1 + \beta_0,$$

$$\lambda_4 = 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0$$

Probability Weighted Moments or K-moments?

The newly introduced (Koutsoyiannis, 2019, 2020) *knowable* (opposite to unknowable) *moments* or *K-moments* contain as special cases (or are one-to-one connected to) classical moments, Probability Weighted Moments and L-moments, as well as order statistics. The related definitions follow:

Noncentral knowable moment of order $(p, 1)$ [**analogous to Probability Weighted Moments**]

$$K'_{p1} := p\mathrm{E}\left[\left(F(\underline{x})\right)^{p-1}\underline{x}\right], \qquad p \geq 1$$

Noncentral knowable moment (or *noncentral K-moment*) *of order* (p, q) [**recovering classical noncentral moments for *p* = *q***]:

$$K'_{pq} := (p-q+1)\mathrm{E}\left[\left(F(\underline{x})\right)^{p-q}\underline{x}^{q}\right], \qquad p \geq q$$

Central knowable moment of order (p, q) [**recovering classical central moments for *p* = *q***]

$$K_{pq} := (p-q+1)\mathrm{E}\left[\left(F(\underline{x})\right)^{p-q}(\underline{x}-\mu)^{q}\right], \qquad p \geq q$$

where μ is the mean of $\underline{x}$, i.e., $\mu := \mathrm{E}\left[\underline{x}_{(p)}\right] \equiv K'_{11}$.

Hypercentral knowable moment (or *central K-moment*) *of order* (p, q) [**analogous to L-moments**]

$$K^{+}_{pq} := (p-q+1)\mathrm{E}\left[\left(2F(\underline{x})-1\right)^{p-q}(\underline{x}-\mu)^{q}\right], \qquad p \geq q$$

Noncentral K-moments of order $(p, 1)$ are closely related to order statistics and, through them, to the extremes.

K-moments or order statistics?

Let $\underline{x}$ be a random variable and $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_p$ be copies of it, independent and identically distributed (iid), forming a sample. The *highest* (pth) *order statistic* of $\underline{x}$ is by definition:

$$\underline{x}_{(p)} \coloneqq \max\left(\underline{x}_1, \underline{x}_2, \dots, \underline{x}_p\right)$$

It is readily obtained that if $F(x)$ is the distribution function of $\underline{x}$ and $f(x)$ its probability density function, then those of $\underline{x}_{(p)}$ are

$$F^{(p)}(x) = \left(F(x)\right)^p, \qquad f^{(p)}(x) = pf(x)\left(F(x)\right)^{p-1}$$

Based on the definition of K-moments it is readily seen that

$$K'_{p1} = \mathrm{E}\left[\underline{x}_{(p)}\right] = \mathrm{E}\left[\max\left(\underline{x}_1, \underline{x}_2, \dots, \underline{x}_p\right)\right]$$

More generally, K-moments of all categories represent expected values of maxima. For example, for odd q or for nonnegative $\underline{x}$, $K'_{pq} = \mathrm{E}\left[\underline{x}^q_{(p-q+1)}\right]$.

In a sample of size n arranged in ascending order ($\underline{x}_{(1)} \le \underline{x}_{(2)} \le \cdots \le \underline{x}_{(n)}$), the random variable $\underline{x}_{(r)}$, $r = 1, \dots, n$, is the rth *order statistic*. Its distribution function is (David and Nagaraja, 2004, p. 10):

$$F^{(r,n)}(x) = P\left\{\underline{x}_{(r)} \le x\right\} = P\left\{F\left(\underline{x}_{(r)}\right) \le F(x)\right\} = \frac{\mathrm{B}_{F(x)}(r, n-r+1)}{\mathrm{B}(r, n-r+1)}$$

which means that the random variable $\underline{u}_r \coloneqq F\left(\underline{x}_{(r)}\right)$ has beta distribution and its mean is $\mathrm{E}\left[\underline{u}_r\right] = \mathrm{E}\left[F\left(\underline{x}_{(r)}\right)\right] = \frac{r}{n+1}$ (hence, the so-called Weibull *plotting positions*). For $r = n$, $\mathrm{E}\left[\underline{u}_n\right] = \frac{n}{n+1}$.

Similar to order statistics, where we use orders up to the sample size n, we should use high orders, up to n, for K-moments, with higher importance given to the highest moments.

Are those high-order K-moments knowable?

Yes, because we can construct an unbiased estimator with good properties such as small variance.

The unbiased estimator of the noncentral moment K'_{p1} and its extension for q > 1 are

$$\underline{\hat{K}}'_{p1} = \sum_{i=1}^{n} b_{inp}\, \underline{x}_{(i)}\,, \qquad \underline{\hat{K}}'_{pq} = \sum_{i=1}^{n} b_{i,n,p-q+1}\, \underline{x}_{(i)}^{q}$$

with (Koutsoyiannis, 2020):

$$b_{inp} = \begin{cases} 0, & i < p \\ \dfrac{p}{n}\,\dfrac{\Gamma(n-p+1)}{\Gamma(n)}\,\dfrac{\Gamma(i)}{\Gamma(i-p+1)}, & i \geq p \geq 0 \end{cases}$$

where p is any positive number (usually, but not necessarily, integer). It can be verified that

$$\textstyle\sum_{i=1}^{n} b_{inp} = 1$$

which is a necessary condition for unbiasedness. Furthermore, for p = 1, $b_{in1} = 1/n$, while for p = 2, the quantity $(n/2)b_{in2}$ is the estimator $\hat{F}\left(x_{(i)}\right)$, i.e.,

$$\hat{F}\left(x_{(i)}\right) = \frac{i-1}{n-1}$$

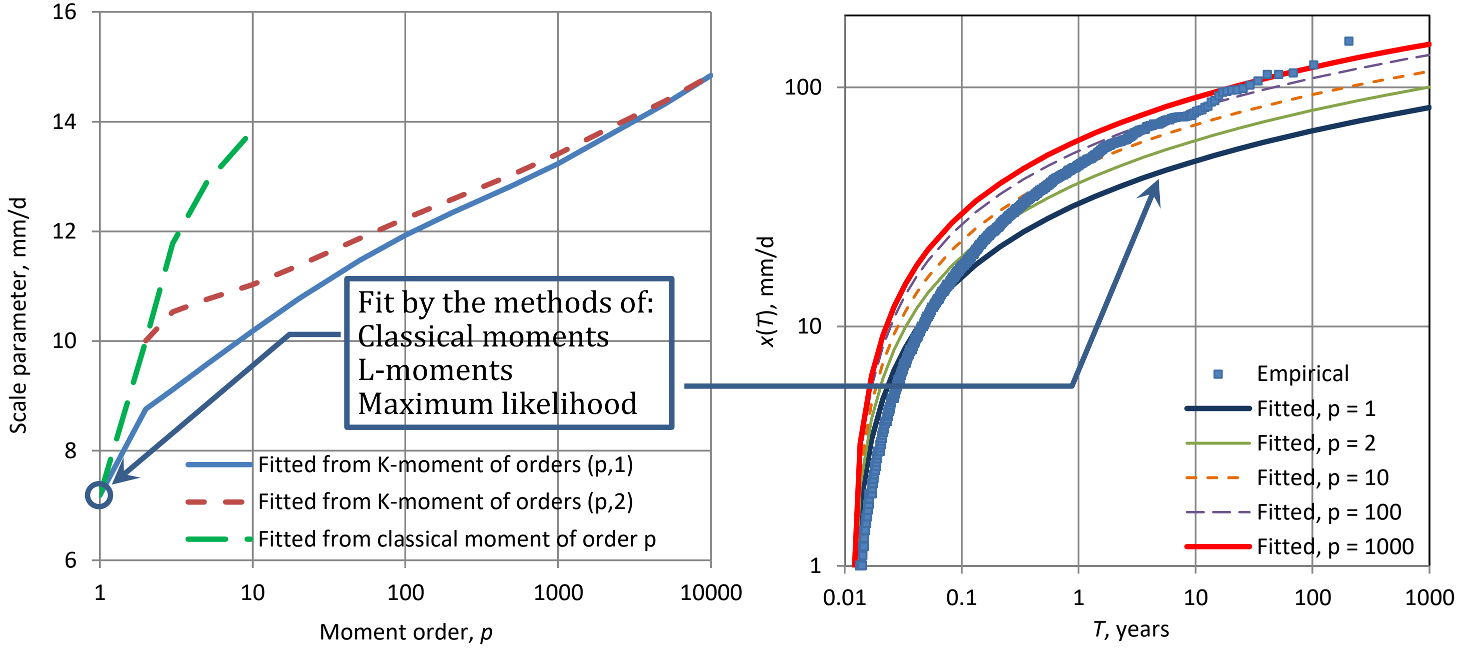
The fact that $b_{inp} = 0$ for i < p suggests that, as the moment order increases, progressively, fewer data values determine the moment estimate, until it remains only one, the maximum, when p = n, with $b_{nnn} = 1$. Furthermore, if p > n then $b_{inp} = 0$ for all i, 1 ≤ i ≤ n, and therefore estimation becomes impossible.

Illustration that high-order K-moments are preferable to low-order K-moments

For the sake of illustration we intentionally choose the simplest and blatantly unsuitable model, the 1-parameter exponential distribution, $F(x) = 1 - \mathrm{e}^{-x/\lambda}$.

One moment suffices to estimate the single (scale) parameter λ—but which moment to choose?

The theoretical K-moments are: $K_{p1} = (H_p - 1)\lambda$, $K_{p2} = \left(\left(H_{p-1} - 1\right)^2 + H_{p-1}^{(2)}\right)\lambda^2$, $K_{pp} = \mu_p = (!\,p)\lambda^p$, where H_p is the pth harmonic number and $H_p^{(2)}$ is the pth harmonic number of order 2.



Fit by the methods of:
Classical moments
L-moments
Maximum likelihood

The moment order *p* affects the fitting dramatically.

The scale parameter λ increases with increasing *p*, *q*.

If we wish to model maxima, it is better to fit based on the 1000th K-moment than on the 1st!

What return period should the maximum value be assigned?

Option 1: $T/d = 1/\left(1 - \mathrm{E}\left[F\left(\underline{x}_{(n)}\right)\right]\right) = n + 1$.

Option 2: $T/d = 1/\left(1 - F\left(\mathrm{E}\left[\underline{x}_{(n)}\right]\right)\right) = 1/\left(1 - F(K_{n1})\right)$.

Assumptions for **illustration**:

- Exponential distribution $F(x) = 1 - \exp(-x/\lambda)$ with $\lambda = 1$.
- $n = 2$, $\max(x_1, x_2) = x_{(2)} = \hat{K}_{21}$
- For the two options we compare the standard error of the estimates of x_T and F_T at $T = n + 1 = 3$; for convenience a time unit $d = 1$ is assumed.

True vales: Distribution parameter $\lambda = 1$, $F_T = 2/3$, $x_T = \ln 3$

Option 1: $\hat{F}\left(x_{(2)}\right) = 2/3$, $F_{x_{(2)}}(x) = (F(x))^2$, $\hat{x}_T = \underline{x}_{(2)}$, $\mathrm{E}\left[F\left(\underline{x}_{(2)}\right)\right] = 2/3$ (unbiased)

$$\mathrm{E}\left[\left(\underline{\hat{x}}_T - x_T\right)^2\right] = \mathrm{E}\left[\left(\underline{x}_{(2)} - \ln 3\right)^2\right] = 7/2 + (\ln 3)^2 - 3\ln 3 = 1.411\,,$$

$$\mathrm{E}\left[\left(F\left(\underline{x}_{(2)}\right) - 2/3\right)^2\right] = \mathrm{E}\left[\left(1 - \exp\left(-\underline{x}_{(2)}\right)\ - 2/3\right)^2\right] = 1/18 = 0.0556$$

Option 2: $\hat{\lambda} = (x_1 + x_2)/2$, $\hat{K}_{21}\ = x_{(2)} = \max(x_1, x_2)$

$\hat{x}_T = \hat{\lambda} \ln 3 = (x_1 + x_2) \ln 3\,/2$, $\mathrm{E}\left[\left(\left(\underline{x}_1 + \underline{x}_2\right) \ln 3\,/2 - \ln 3\right)^2\right] = (\ln 3)^2/2 = 0.603 <\ 1.411$

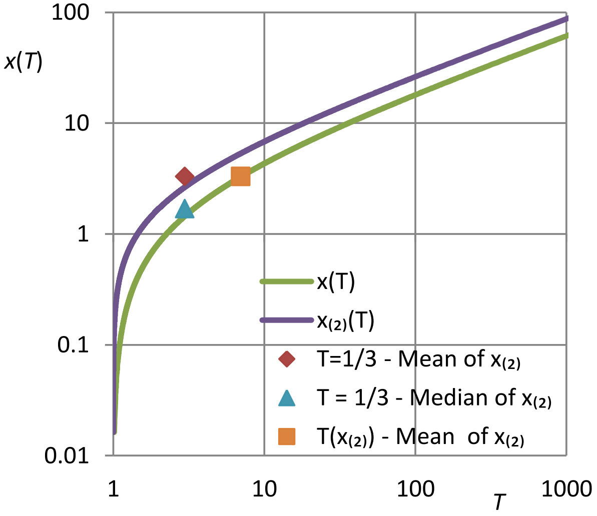
$\hat{F}\left(x_{(2)}\right) = F\left(\hat{K}_{21}\right) = 1 - \exp\left(-2x_{(2)}/\left(x_{(1)} + x_{(2)}\right)\right)$, $\mathrm{E}\left[\hat{F}\left(\underline{x}_{(2)}\right)\right] = \mathrm{E}\left[F\left(\underline{\hat{K}}_{21}\right)\right] = 0.7675 \neq 2/3$ (bias)

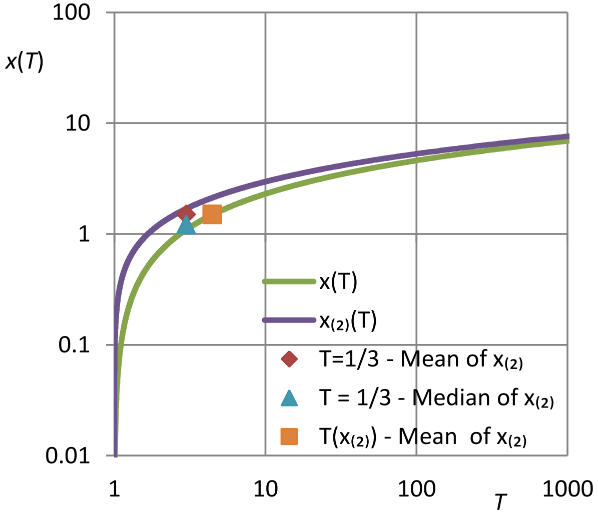
$\mathrm{E}\left[\left(F\left(\underline{x}_{(2)}\right) - 2/3\right)^2\right] = \mathrm{E}\left[\left(F\left(\underline{\hat{K}}_{21}\right)\ - 2/3\right)^2\right] = 0.0146 < 0.0556$. (Note: the numerical results were calculated by numerical integration).

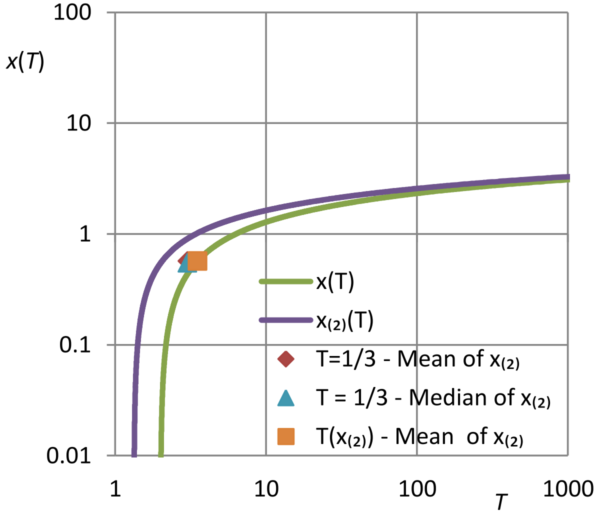
Result: **Option 2 is preferable** for estimating both x and F even though it entails bias for the latter.

What return period should the maximum value be assigned? (2)

The graphs show results of Monte Carlo simulations for the three indicated distributions again for the maximum of two variables, $x_{(2)} = \max(x_1, x_2)$







Pareto distribution, κ =0.5

$$F(x) = 1 - (1 + \kappa x)^{-\frac{1}{\kappa}}$$

$$x(F) = \frac{(1-F)^{-\kappa} - 1}{\kappa}$$

Exponential distribution

$$F(x) = 1 - \exp(-x)$$

$$x(F) = -\ln(1 - F)$$

Normal distribution

$$f(x) = \exp(-x^2/2)/\sqrt{2\pi}$$

The graphs show that:

- Differences resulting from Options 1 and 2 can be substantial. Particularly for the Pareto distribution the average of $x_{(2)}$ for Options 1 and 2 corresponds to $T = 3$ and T = 7.2, respectively. The value $T = 3$ looks unrealistically low, except for the normal distribution.
- Practically, the median of $x_{(2)}$ corresponds to $T = 3$ for all distributions.

Assigning return periods to K-moments of any order

- The non-central K-moment for q = 1 is:

$$K'_{p1} = p\mathrm{E}\left[\left(F\left(\underline{x}\right)\right)^{p-1} \underline{x}\right]$$

- By definition, it represents the expected value of the maximum of p realizations of $\underline{x}$.
- To determine the theoretical return period $T(K'_{p1})$ we introduce the ratio Λ_p which happens to vary only slightly with p:

$$T\left(K'_{p1}\right) = \frac{d}{1-F\left(K'_{p1}\right)}, \quad \Lambda_p := \frac{T(K'_{p1})}{d\,p} = \frac{1}{p\left(1-F\left(K'_{p1}\right)\right)}$$

- The slight variation of Λ_p with p can be very well approximated if we first accurately determine the specific values Λ_1 and Λ_∞. For the approximation of Λ_p we use one of the following relationships:

$$\Lambda_p \approx \Lambda_\infty \left(\frac{\Lambda_1}{\Lambda_\infty}\right)^{\frac{1}{p^c}}, \quad \Lambda_p \approx \Lambda_\infty + (\Lambda_1 - \Lambda_\infty)\frac{1}{p^c}$$

where c is a constant depending on the distribution function with default value c = 1. The former approximation is more accurate, but the latter more convenient as for c = 1 it yields a linear relationship between the return period T and the K-moment order p:

$$\frac{T(K'_{p1})}{d} = p\Lambda_p \approx \Lambda_\infty p + (\Lambda_1 - \Lambda_\infty)$$

Exact and approximate relationships between p and T

- For given p and distribution function $F(x)$, K'_{p1} is theoretically determined; then $T(K'_{p1})$ is determined from the exact relationship. In absence of an analytical solution, we can establish an exact relationship between p and T by doing numerical calculations for several p.
- Alternatively, we can make exact calculations only for $\mathit{\Lambda}_1$ and $\mathit{\Lambda}_\infty$ (the latter as a limit for large p) and use the approximate equations for any p. The table below provides such results.

Distribution	Distribution definition	$\mathit{\Lambda}_1$	$\mathit{\Lambda}_\infty$	Exact $\mathit{\Lambda}_p$	Approximate $\mathit{\Lambda}_p$
Normal	$f(x) = \frac{\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}{\sqrt{2\pi}\,\sigma}$	$\mathit{\Lambda}_1 = 2$	$\mathrm{e}^\gamma = 1.781$		$\mathit{\Lambda}_\infty + (\mathit{\Lambda}_1 - \mathit{\Lambda}_\infty)/p$
Exponential	$f(x) = \mathrm{e}^{-x/\lambda+\psi}/\lambda$	$\mathrm{e} = 2.718$	$\mathrm{e}^\gamma = 1.781$	e^{H_p}/p	$\mathit{\Lambda}_\infty + (\mathit{\Lambda}_1 - \mathit{\Lambda}_\infty)/p$
Gamma	$f(x) = \frac{x^{\xi-1}\mathrm{e}^{-x/\lambda}}{\lambda^{1/\kappa}\,\Gamma(\xi)}$	$\frac{\Gamma(\xi)}{\Gamma_\xi(\xi)}$	$\mathrm{e}^\gamma = 1.781$		$\mathit{\Lambda}_\infty + (\mathit{\Lambda}_1 - \mathit{\Lambda}_\infty)/p$
Weibull	$F(x) =$ $1 - \exp\left(-\left(\frac{x}{\lambda} - \psi\right)^\xi\right)$	$\mathrm{e}^{(\Gamma(1+\kappa))^{\frac{1}{\kappa}}}$	$\mathrm{e}^\gamma = 1.781$		$\mathit{\Lambda}_\infty \left(\frac{\mathit{\Lambda}_1}{\mathit{\Lambda}_\infty}\right)^{\frac{1}{p^c}}$ $c = \mathit{\Lambda}_\infty / \mathit{\Lambda}_1$
Pareto	$F(x) =$ $1 - \left(1 + \kappa\left(\frac{x}{\lambda} - \psi\right)\right)^{-\frac{1}{\kappa}}$	$\left(\frac{1}{1-\kappa}\right)^{\frac{1}{\kappa}}$	$(\Gamma(1-\kappa))^{\frac{1}{\kappa}}$	$\frac{\left((p+1-\kappa)\,\mathrm{B}(1-\kappa, p+1)\right)^{\frac{1}{\kappa}}}{p}$	$\mathit{\Lambda}_\infty + (\mathit{\Lambda}_1 - \mathit{\Lambda}_\infty)/p$
Lognormal	$f(x) = \frac{\exp\left(-\frac{(\ln(x/\lambda-\psi))^2}{2\sigma^2}\right)}{\sqrt{2\pi}\,\sigma\,x}$	$\frac{2}{\mathrm{erfc}\left(\sigma/2^{3/2}\right)}$	$\mathrm{e}^\gamma = 1.781$		$\mathit{\Lambda}_\infty \left(\frac{\mathit{\Lambda}_1}{\mathit{\Lambda}_\infty}\right)^{\frac{1}{p^c}}$ $c = 0.12(1 + \mathit{\Lambda}_\infty / \mathit{\Lambda}_1)$

Explanations of symbols: γ = 0.5772 the Euler constant; Γ( ) the gamma function; Γ_α( ) the incomplete gamma function; B(,) is the beta function; erfc( ) the complementary error function; $H_p := \sum_{i=1}^{p} 1/i$ is the pth harmonic number.

Practical considerations for the relationships between *p* and *T*

- Depending on the required degree of accuracy, we can determine the moment order p for a specified return period T by any of the following relationships (with the first being very rough and the last being exact).

$$p^{(\mathrm{A})} = \frac{T}{2d}, \quad p^{(\mathrm{B})} = \frac{T}{\Lambda_\infty d} - \frac{\Lambda_1}{\Lambda_\infty} + 1, \quad p^{(\mathrm{C})} = p(T/d, \boldsymbol{\alpha})$$

  where $\boldsymbol{\alpha}$ is a vector of the shaper parameters of the distribution function (the location and scale parameters should not affect the relationship between p and T).

- A justification of $p^{(\mathrm{A})}$ is that the table in the previous page supports a rough approximation (for preliminary estimates) of $\Lambda_p \approx \Lambda_1 \approx 2$. Furthermore, any symmetric distribution for $p = 1$ will give exactly $\Lambda_1 = 2$ because K'_{11} is the mean, which in this case equals the median and thus has a return period of $2d$.
- A justification for $p^{(\mathrm{B})}$ is that it is readily derived from the last approximate equation of p. 13.

Example with comparison of the three options for Pareto distribution with shape parameter κ = 0.15.

	d = 10 min			d = 1 h			d = 1 d		
	$p^{(\mathrm{A})}$	$p^{(\mathrm{B})}$	$p^{(\mathrm{C})}$	$p^{(\mathrm{A})}$	$p^{(\mathrm{B})}$	$p^{(\mathrm{C})}$	$p^{(\mathrm{A})}$	$p^{(\mathrm{B})}$	$p^{(\mathrm{C})}$
T = 2 months	4 383	4 307	4307	731	717	717	30	29	29
T = 1 year	26 298	25 842	25 842	4 383	4 307	4 307	183	179	179
T = 2 years	52 596	51 684	51 684	8 766	8 614	8 614	365	358	358
T = 100 years	2 629 800	2 584 212	2 584 212	438 300	430 702	430 702	18 263	17 945	17 945

Final fitting on K-moments for orders *p* from ~100 to 10 000 (*T* = ~2 to 200 years)

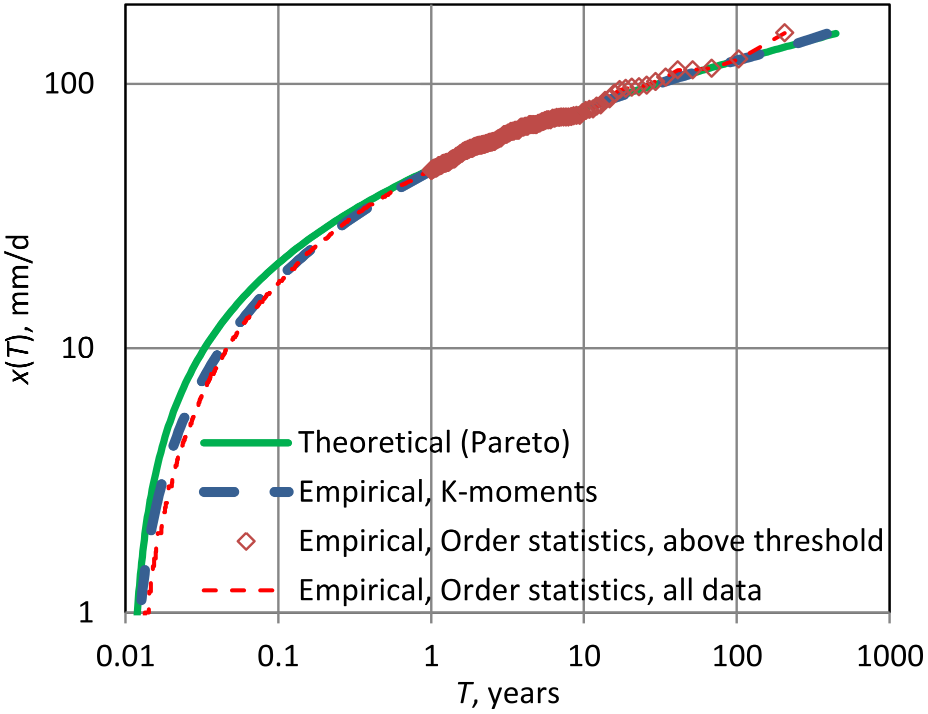
We assume Pareto distribution with zero lower bound (for physical consistency):

$F(x) = 1 - (1 + \kappa\, x/\lambda)^{-\frac{1}{\kappa}}$ or

$\frac{T(x)}{d} = (1 + \kappa\, x/\lambda)^{\frac{1}{\kappa}}$

The estimated K-moments have return period:

$$\frac{\hat{T}(\hat{K}'_{p1})}{d} = p\Lambda_p = \left(\frac{1}{\kappa} + (p + 1 - \kappa)\,\mathrm{B}(1 - \kappa, p + 1)\right)$$



We estimate the parameters by minimizing the mean square error of the logarithms of the empirical $\hat{T}(\hat{K}'_{p1})$ from the theoretical $T(\hat{K}'_{p1})$. We calculate the error for a range of *T* from 2 to 200 years. The fitted parameters are $\kappa = 0.096$, $\lambda = 8.37$ mm/d.

The graph shows a perfect fit of theoretical and empirical curves for *T* > 1 year (the two curves are indistinguishable).

For comparison, empirical curves for the order statistics are also plotted but these have not been used at any step of the fitting procedure.

Note: Minimizing the error of $\hat{K}'_{p1}$ with respect to K'_{p1}, without reference to *T*, is another possibility but presupposes exact relationships for K'_{p1}, which may be infeasible to derive for some distributions. In the Pareto case this is feasible and the resulting parameters are practically the same as above.

A note on the behaviour of the Λ factor

The exact relationship for Pareto distribution is:

$$\Lambda_p = \frac{\left((p+1-\kappa)\,\mathrm{B}(1-\kappa, p+1)\right)^{\frac{1}{\kappa}}}{p}$$

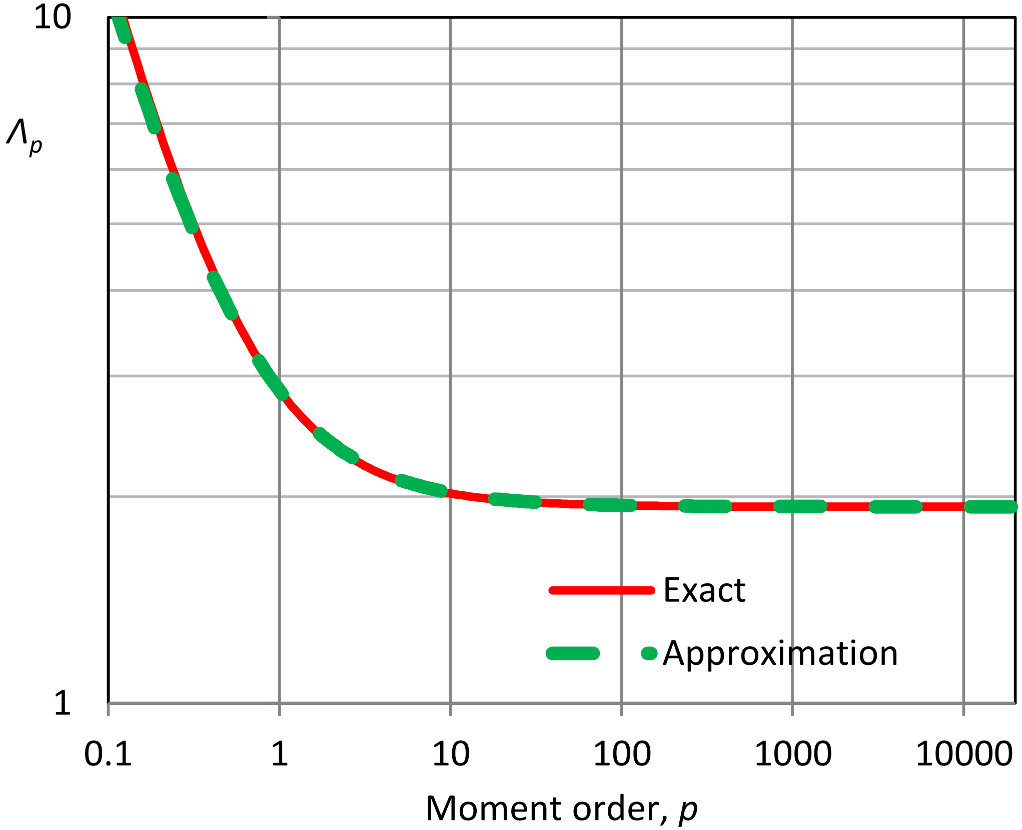
The approximate relationship is:

$$\Lambda_p = \Lambda_\infty + (\Lambda_1 - \Lambda_\infty)/p$$

with

$$\Lambda_1 = \left(\frac{1}{1-\kappa}\right)^{\frac{1}{\kappa}},$$

$$\Lambda_\infty = (\Gamma(1-\kappa))^{\frac{1}{\kappa}}$$



The graph shows that the approximation is perfect.

Note that for $p > 10$, $\Lambda_p \approx 2$.

Note also that the procedure and the approximation work well even for $p < 1$.

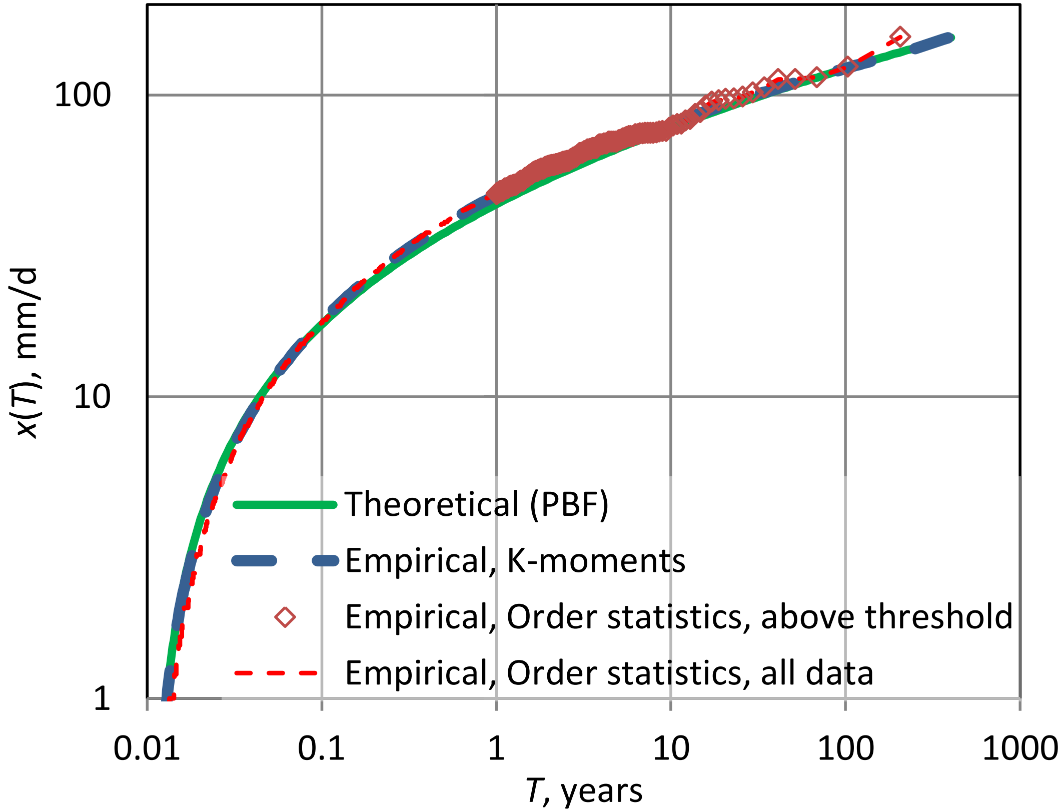
Slight improvement for a global fitting

By adding one parameter to the theoretical distribution function we can get a model applicable for the entire range of rainfall depth.

Namely, we use the Pareto-Burr-Fuller (PBF) distribution with zero lower bound (for physical consistency for rainfall):

$$F(x) = 1 - (1 + \kappa(x/\lambda)^c\ )^{-\frac{1}{c\kappa}}$$

We use the same estimation procedure as above but calculate the error on the entire range of values. However, in order not to distort the good fitting on the tail, we keep the tail index κ as estimated for the Pareto distribution (κ =0.096).



The estimated parameters are: κ = 0.096, c= 0.883, λ = 5.04 mm/d.

A perfect fit of the model (green continuous line) and empirical curve (blue dashed line) is seen for the entire range.

For comparison, empirical curves for the order statistics are also plotted but these have not been used at any step of the fitting procedure.

Is this the last word?

No.

Time dependence and in particular long-term persistence (Hurst-Kolmogorov behaviour) is present in the Bologna record—and in most of rainfall records.

Time dependence induces bias to estimators of K-moments. A K-moment is a characteristic of the marginal (first order) distribution of the process and therefore is not affected by the dependence structure. However, the estimator is. Thus, we need to quantify the bias of the estimator of K-moments and take it into consideration in the model fitting.

Naturally by considering the bias, the resulting model will give higher rainfall values for specified return periods.

Cyclostationarity (periodic seasonal variation and possibly diurnal variation) also affect the rainfall process and need to be considered in the model building.

These issues are covered in Koutsoyiannis (2020).

Conclusions

- The K-moments are powerful tools that unify other statistical moments (classical, L-, probability weighted) and order statistics, offering several advantages.
- In particular, they offer a sound basis for distribution fitting with emphasis on extremes.
- For independent identically distributed variables, K-moments offer unbiased, reliable and workable estimators for low and high orders *p*, up to the sample size *n*.
- For practical applications the following steps can be used for an extreme-oriented distribution fitting:
    - Use all available data without exceptions (not block maxima, not values over threshold).
    - Estimate noncentral K-moments $\hat{K}'_{p1}$ for a set of orders *p* up to the sample size *n*. (Do not hesitate to calculate moments or orders of several thousands or millions.)
    - Choose a distribution tail according to the theory of extreme value distributions (Pareto for rainfall and streamflow, exponential or normal for temperature, etc.)
    - For that tail, estimate the empirical return periods (T) of the estimated values $\hat{K}'_{p1}$.
    - Also determine the theoretical return periods of the estimated values $\hat{K}'_{p1}$ as functions of the distribution parameters.
    - Determine the parameters by numerically minimizing a fitting metric, such as the mean square error of (logarithms of) empirical and theoretical T for, say, $T > 2$ years.
- Time dependence influences the estimates and fitting but this topic has not been covered in this presentation.

References

David, H.A., and Nagaraja, H.N. (2004), *Order Statistics.* Wiley Online Library.

Koutsoyiannis, D. (2019), Knowable moments for high-order stochastic characterization and modelling of hydrological processes, *Hydrological Sciences Journal*, 64 (1), 19–33, doi:10.1080/02626667.2018.1556794.

Koutsoyiannis, D. (2020), *Stochastics of Hydroclimatic Extremes – A Cool Look at Risk* (in preparation).

Lombardo, F., Volpi, E., Koutsoyiannis, D., and Papalexiou, S.M. (2014) Just two moments! A cautionary note against use of high-order moments in multifractal models in hydrology, *Hydrology and Earth System Sciences*, 18, 243–255, doi:10.5194/hess-18-243-2014.

Volpi, E., Fiori, A., Grimaldi, S., Lombardo, F., and Koutsoyiannis, D. (2019) Save hydrological observations! Return period estimation without data decimation, *Journal of Hydrology*, doi:10.1016/j.jhydrol.2019.02.017.

Data sources

GHCN Version 3 data: retrieved on 2019-02-17 from https://climexp.knmi.nl/gdcnprcp.cgi?WMO=ITE00100550

Dext3r data: retrieved on 2019-02-17 from http://www.smr.arpa.emr.it/dext3r/

Accompanying poster papers – Thursday 11 April 2019, 16:15–18:00

A.107 | EGU2019-10348
N. Agkatheris, Th. Iliopoulou, P. Dimitriadis and D. Koutsoyiannis, Stochastic modelling of extreme rainfall using K-moments and K-climacogram

A.117 | EGU2019-10070
K.-G. Glynis, Th. Iliopoulou, P. Dimitriadis and D. Koutsoyiannis, Investigating the tail behaviour of surface temperature on global scale using K- moments