International Precipitation Conference (IPC10)

Coimbra, Portugal, 23 - 25 June 2010

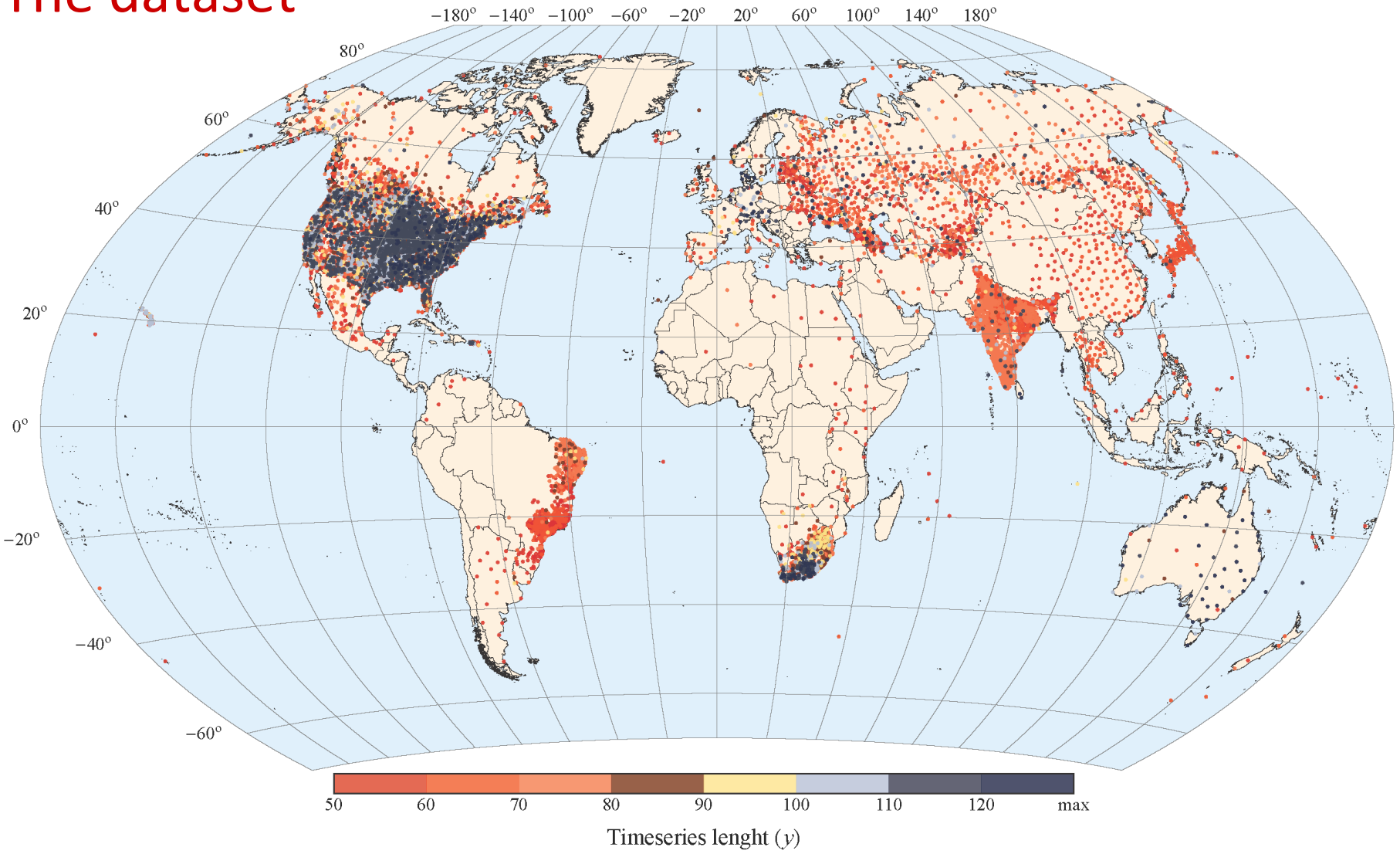
Topic 1 Extreme precipitation events:
Physics- and statistics-based descriptions

A world-wide investigation of the probability distribution of daily rainfall

S.M. Papalexiou and D. Koutsoyiannis

Department of Water Resources & Environmental Engineering
National Technical University of Athens, Greece

The dataset



A subset of the Global Historical Climatology Network-Daily database: stations with daily rainfall time series with length over 50 years (a total of 11 519 stations with very few missing values).

Entropy and maximum entropy distributions

- The Boltzmann-Gibbs-Shannon (BGS) entropy for a non-negative random variable X with probability density function $f(x)$ is

$$\Phi(X) = -\int_0^\infty f(x) \ln f(x) dx$$

- The Havrda-Charvát-Tsallis (HCT) entropy, a generalization of the BGS entropy that has gained popularity in the last decades, was originally introduced axiomatically by Havrda and Charvát [1967] and re-introduced by Tsallis [1988]; this is defined by

$$\Phi_\alpha(X) = \frac{1}{\alpha - 1}\left\{1 - \int_0^\infty \left[f(x)\right]^\alpha dx\right\}$$

- Application of the principle of maximum entropy (ME) [Jaynes, 1957a, b], for a given set of macroscopic constrains expressed in the form $E[g_i(X)] = c_i$, $i = 1,\ldots,m$, for the BGS and the HCT entropies using Lagrange multipliers λ_i results, respectively, in density functions

$$f_X(x) = \exp[-\lambda_0 - \lambda_1 g_1(x) - \ldots - \lambda_m g_m(x)], \quad x \geq 0$$

$$f_X(x) = \{1 + (1 - \alpha)[\lambda_0 + \lambda_1 g_1(x) + \ldots + \lambda_m g_m(x)]\}^{-1/(1-\alpha)}, \quad x \geq 0$$

The most common maximum entropy distributions

- The most common constraints in entropy optimization assume known first and second moments.
- The resulting probability distributions from these constraints for positive variables are:

  (a) for the BGS entropy, the normal distribution truncated at zero, and,

  (b) for the HCT entropy, a symmetric bell-shaped distribution with power-type tails truncated at zero, called the Pearson type VII distribution (introduced by Pearson in 1916; now also called the Tsallis distribution).
- These distributions do not adequately describe various properties of the empirical daily rainfall distribution, e.g.,

  (a) the truncated normal seems to fail to describe both the right and left tail of fine time-scale rainfall [Koutsoyiannis, 2005],

  (b) the truncated Tsallis distribution performs better but seems to fail to describe the left tail [Papalexiou and Koutsoyiannis, 2008a], and,

  (c) both distributions have limitations in the shapes they can form, as their skewness is only due to truncation.

Seeking a better entropy-based distribution for rainfall

- We seek a probability distribution resulting from the application of the principle of maximum entropy that is capable of describing the daily rainfall at all locations, if possible.
- It seems reasonable to generalize the constraints, from the first two moments (orders 1 and 2) to moments of unspecified order p, i.e. $E(X^p) = m_p$.
- Another simple constrain that seems reasonable for non-negative positively skewed variables is the geometric mean, i.e., $E(\ln X) = \ln \mu_G$.
- Using only two constrains (μ_G, m_p) and BGS entropy, the ME density is

$$f_X(x) = \exp[-\lambda_0 - \lambda_1 \ln x - \lambda_2 x^p], \quad x \geq 0$$

which after algebraic manipulation and parameter renaming becomes

$$f_X(x) = \frac{\gamma_2}{\beta \Gamma(\gamma_1/\gamma_2)} \left(\frac{x}{\beta}\right)^{\gamma_1 - 1} \exp\left[-\left(\frac{x}{\beta}\right)^{\gamma_2}\right]$$

Generalized Gamma GG(x; β, γ_1, γ_2)

where β *is a* scale parameter and γ_1 and γ_2 are shape parameters

- This includes as special cases many common distributions, e.g., the Gamma, Weibull, and Exponential distributions.

A generalization scheme

- The following generalizations of the exp and ln functions are used (compatible with the definition of the exp function by Euler, 1748)

$$\exp_n(x) := (1+nx)^{1/n}, \text{ so that } \exp(x) = \lim_{n\to 0} \exp_n(x)$$

$$\ln_n(x) := (x^n - 1)/n, \text{ so that } \ln(x) = \lim_{n\to 0} \ln_n(x)$$

- Using these functions, the ME distribution for the HCT entropy is

$$f_X(x) = \exp_{a-1}[-\lambda_0 - \lambda_1 g_1(x) - \ldots - \lambda_n g_n(x)], \quad x \geq 0$$

- Using constraints $E(X^p) = m_p$ and $E(\ln_n X) = \ln_n \mu_{Gn}$, the ME distribution is

$$f_X(x) = \exp_n[-\lambda_0 - \lambda_1 \ln_n x - \lambda_2 x^p], \quad x \geq 0$$

- After algebraic manipulations and assuming small n, and after parameter renaming, this is approximated as:

$$f_X(x) = \frac{\gamma_2 \gamma_3^{\gamma_1/\gamma_2}}{\beta B(-\gamma_1/\gamma_2 + 1/\gamma_3, \gamma_1/\gamma_2)} \left(\frac{x}{\beta}\right)^{\gamma_1 - 1} \left[1 + \gamma_3 \left(\frac{x}{\beta}\right)^{\gamma_2}\right]^{-1/\gamma_3}$$

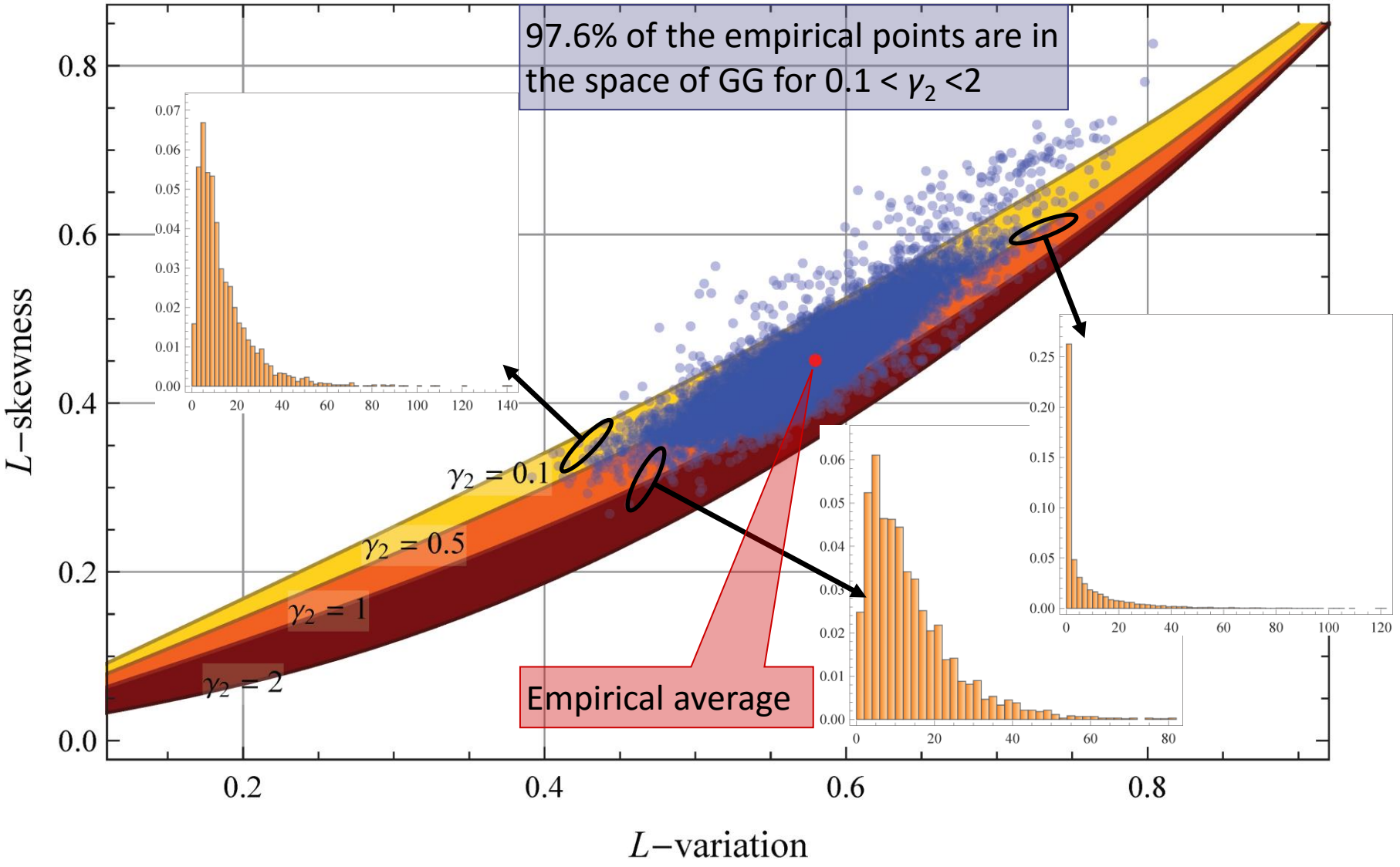
Generalized Beta of 2nd kind
GBII(x; β, γ_1, γ_2, γ_3)

- The Generalized Gamma is a special case of GBII; another three-parameter special case of GBII is the Burr type XII distribution, introduced in 1942:

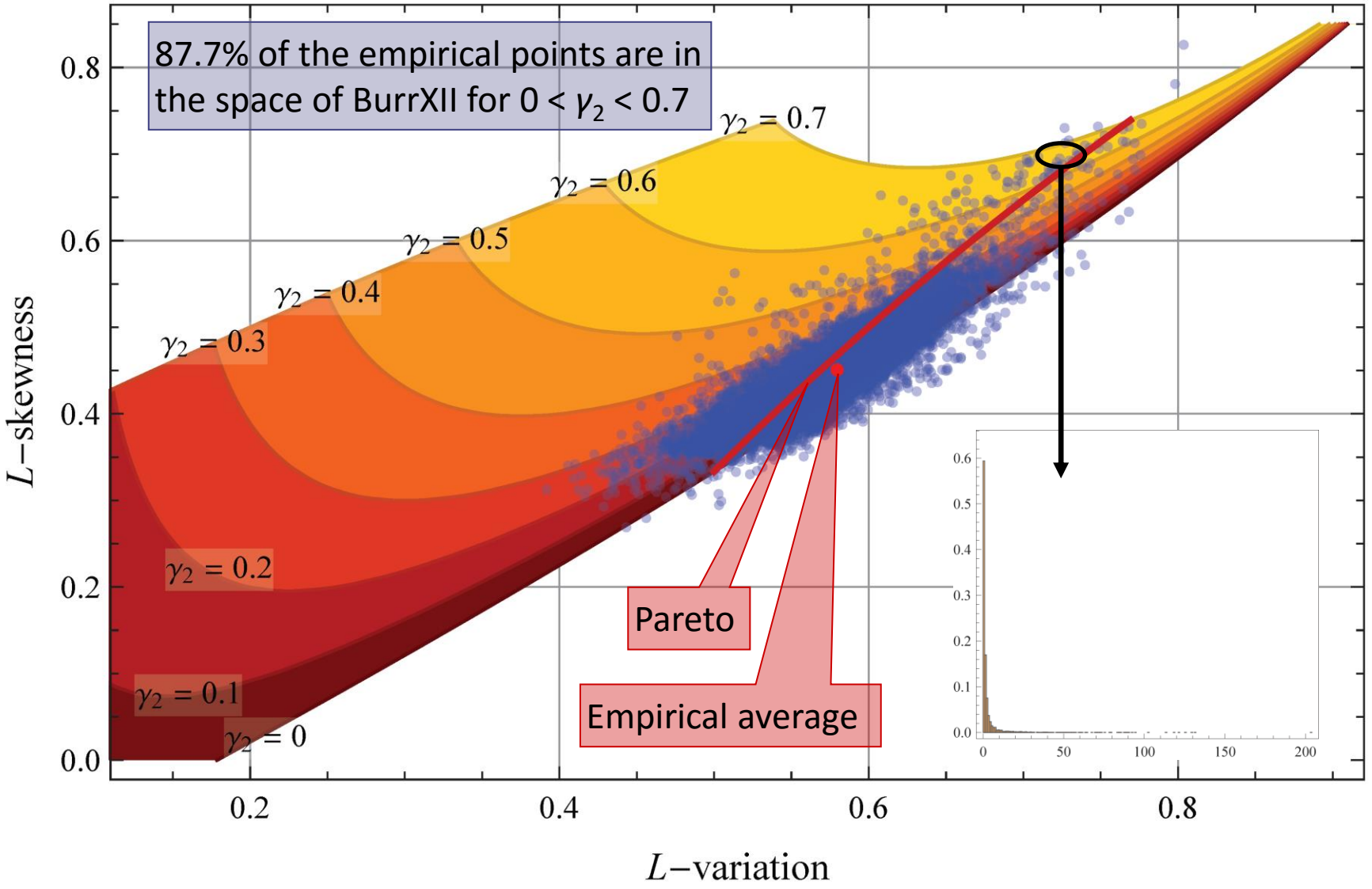
$$f_X(x) = \frac{\gamma_1}{\beta} \left(\frac{x}{\beta}\right)^{\gamma_1 - 1} \left[1 + \gamma_1 \gamma_2 \left(\frac{x}{\beta}\right)^{\gamma_1}\right]^{-1 - \frac{1}{\gamma_1 \gamma_2}}$$

Burr type XII
BurrXII(x; β, γ_1, γ_2)

Empirical points in the Generalized Gamma *L*-moments space



Empirical points in the Burr type XII L-moments space



From the body to the tail of the distribution

- The upper tail of the distribution is its most important part as it rules the magnitude and the frequency of the extreme events.
- The heavy-tailed distributions, whose probability density function goes to zero less rapidly than exponentially, result in more frequent and more intense extreme events compared to distributions of exponential type.
- The most commonly used models for daily rainfall belong to the exponential family (e.g. Gamma). However, several studies suggest that heavy tailed distributions may be more suitable (e.g. a pioneering study by Milke, 1973, which proposed the Kappa distribution, a heavy tailed distribution, for daily rainfall).
- As only a very small portion of the empirical data belongs to the tail, fitting a simple (e.g. two-parameter) distribution function using all data will be "biased" against the tail (the estimated parameters will result in a fitting that best describes the largest portion of the data, i..e the body rather than the tail).
- An ill-fitted tail may result in serious errors with severe consequences in hydrological design. For example, the magnitude of the 1000-year precipitation may be seriously underestimated if it is calculated from an exponential tail rather than a heavy tail.
- It may be assumed that a simple, two-parameter distribution can give an acceptable fit on the tail only, i.e. over a certain threshold.

Two-parameter special cases and their tails

Distribution	Survival function $F_X^*(x)$	Comments
Pareto BurrXII(x, β, 1, γ)	$F_X^*(x)=\left(1+\gamma\frac{x}{\beta}\right)^{-\frac{1}{\gamma}}$	The Pareto distribution is the simplest heavy-tailed distribution and depending on the shape parameter γ may produce very extreme events. Other distributions, tail-equivalent with Pareto, are the Burr [*Tadikamalla*, 1980], the Kappa [*Mielke*, 1973] and the Log-Logistic [e.g. *Ahmad et al.*, 1988].
Weibull GG(x, β, γ, γ)	$F_X^*(x)=\exp\left(-\frac{x^\gamma}{\beta}\right)$	The Weibull distribution, a common model in hydrology [e.g. *Heo et al.*, 2001], can be considered as a generalization of the exponential distribution, and for a shape parameter $\gamma < 1$, results in a heavier tail compared to that of the standard exponential distribution tail.
Gamma GG(x, β, γ, 1)	$F_X^*(x)=\frac{\Gamma\left(\gamma,\frac{x}{\beta}\right)}{\Gamma(\gamma)}$	The Gamma distribution is probably the most popular model for describing daily rainfall [e.g. *Buishand*, 1978]. Asymptotically, it behaves like the standard exponential distribution.
Log-Normal	$F_X^*(x)=\frac{1}{2}\mathrm{erfc}\left(\frac{\ln x-\gamma}{\sqrt{2}\beta}\right)$	The Log-Normal distribution is a very common distribution in hydrology that may approximate power-law distributions for a large portion of the body of the distribution [*Mitzenmacher*, 2004].

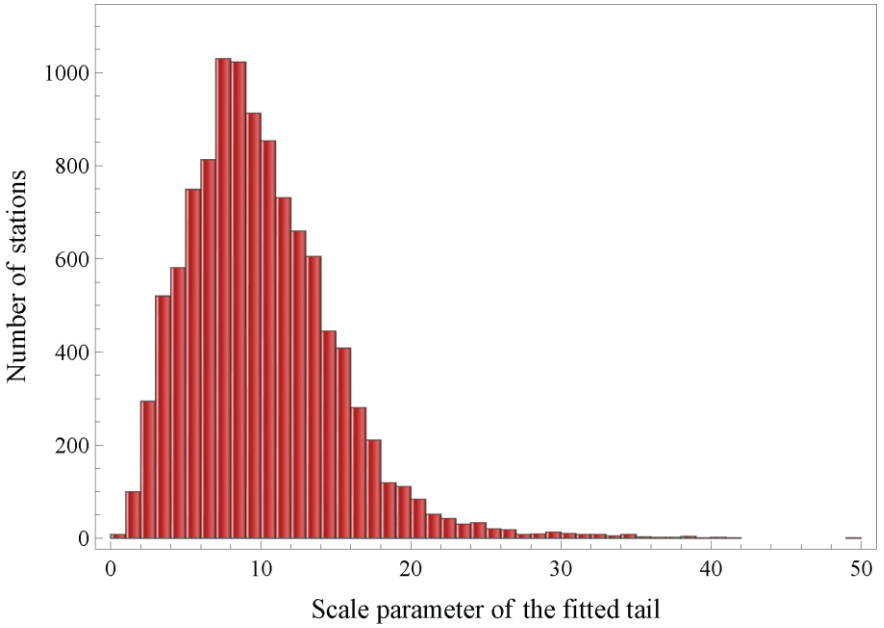
Notes: $F_X^*(x)=1-F_X(x)$, $\mathrm{erfc}(x)=\frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}\,dt$, $\Gamma(s,x)=\int_x^\infty t^{s-1}e^{-t}\,dt$ and $\Gamma(s)=\int_0^\infty t^{s-1}e^{-t}\,dt$

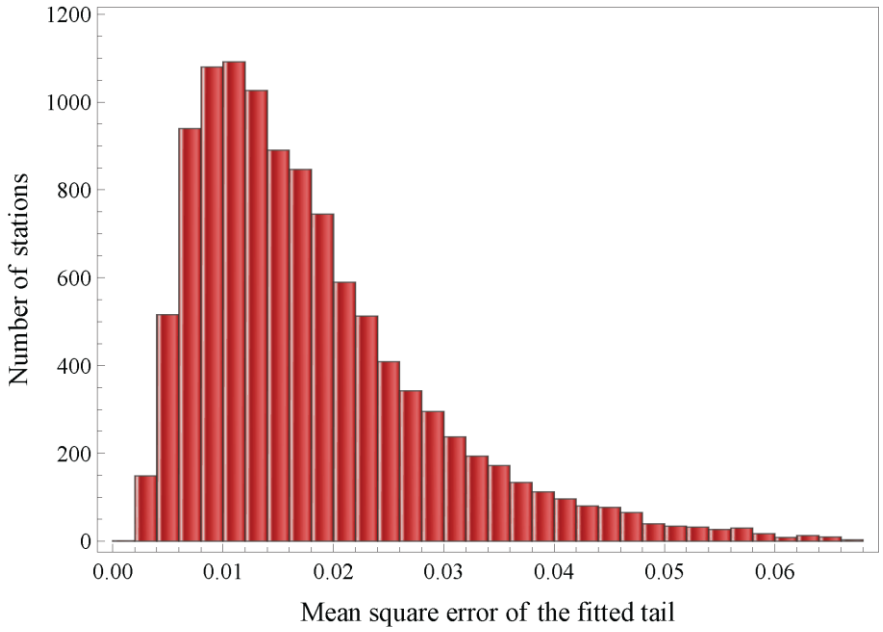
Fitting distribution tails to empirical data

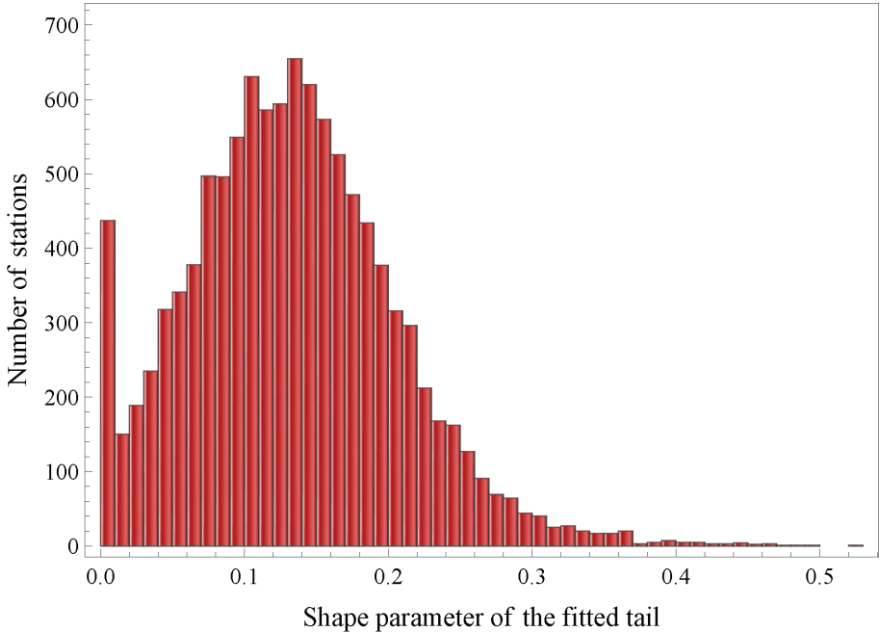
- The upper tail is assessed through its survival function $F_X^*(x) := 1 - F_X(x) = P\{X > x \,|\, x > 0\}$ for large x.
- For each station with record length of N years and a total number n of non-zero values, we derived the empirical survival function at the tail, $F_n^*(x_i)$, as the empirical probability (according to the Weibull plotting position) of the N largest non-zero rainfall values, i.e., $F_n^*(x_i) = r(x_i)/(n + 1)$, with $r(x_i)$ being the rank of the value x_i, i.e., the position of x_i in the ordered sample $x_{(1)} \geq,\ldots, \geq x_{(n)}$.
- Four different theoretical distributions were chosen and fitted to the empirical tails: Pareto, Weibull, Log-Normal and Gamma.
- The theoretical survival functions were fitted to the empirical ones by minimizing a modified mean square error (MSE) norm defined as $\text{MSE} = (F_X^*(x_i) / F_n^*(x_i) - 1)^2/N$; this is superior to the classical MSE norm (more balanced).

Fitting results: the Generalized Pareto tail

	MSE	Scale parameter	Shape parameter
Min	0.0019	0.73	0.00
Mean	0.0179	9.99	0.13
Median	0.0153	9.30	0.13
Max	0.0679	50.00	0.53
Standard Deviation	0.0107	4.92	0.07

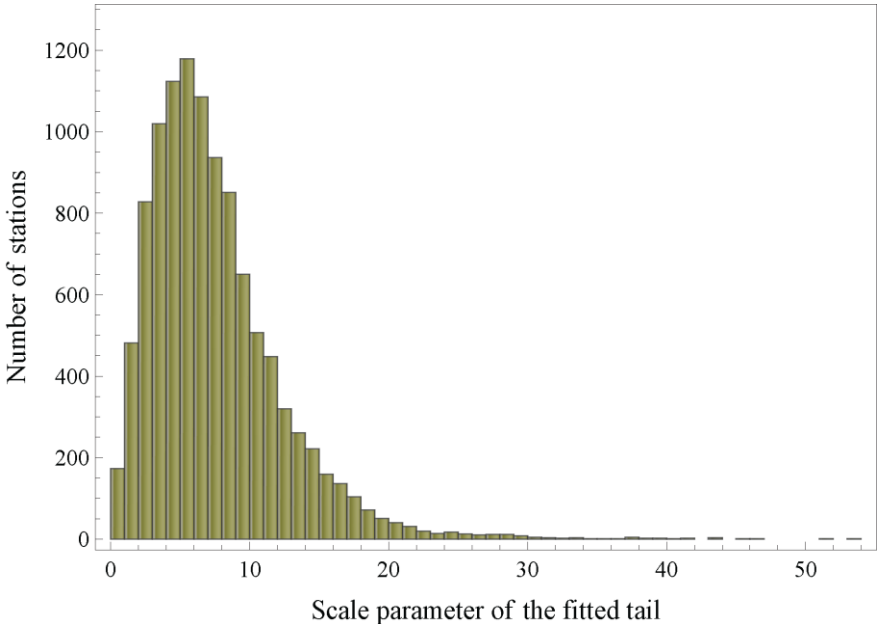


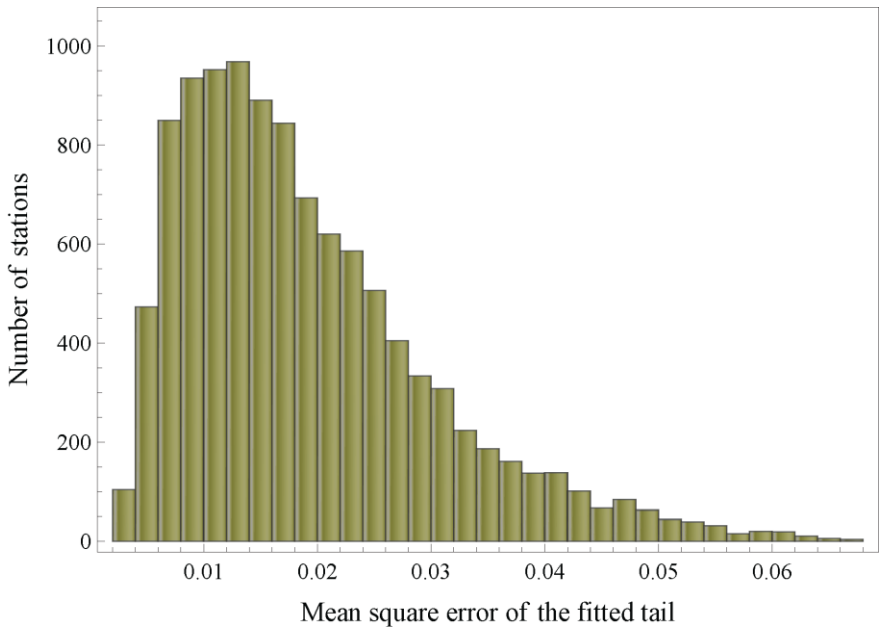


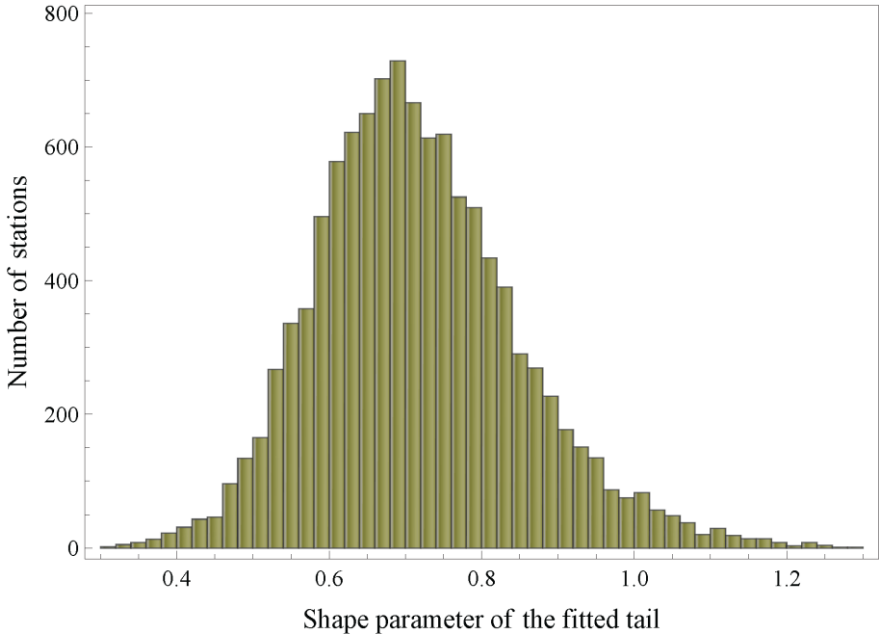


Fitting results: the Weibull tail

	MSE	Scale parameter	Shape parameter
Min	0.0020	0.17	0.31
Mean	0.0191	7.49	0.72
Median	0.0166	6.55	0.70
Max	0.0664	53.17	1.30
Standard Deviation	0.0111	4.79	0.13

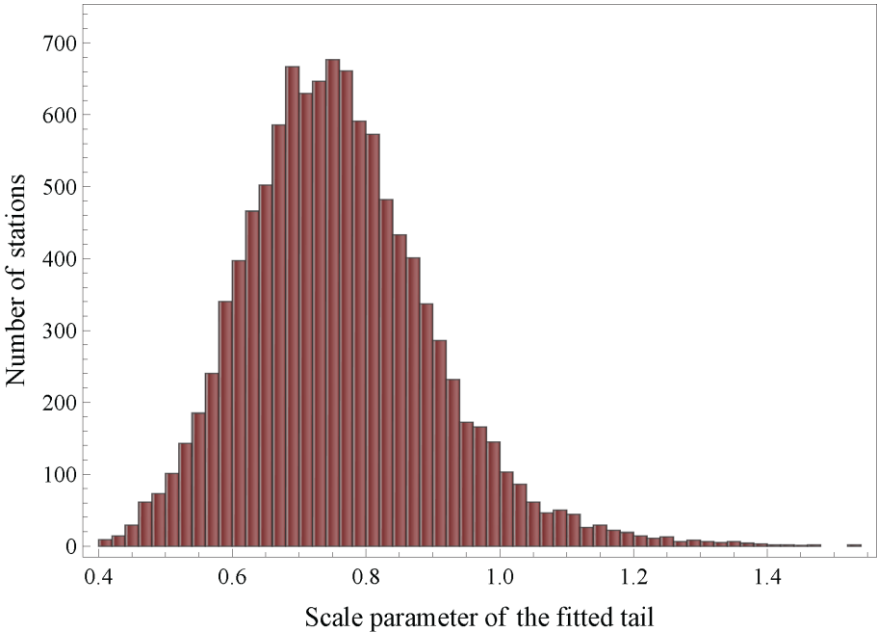


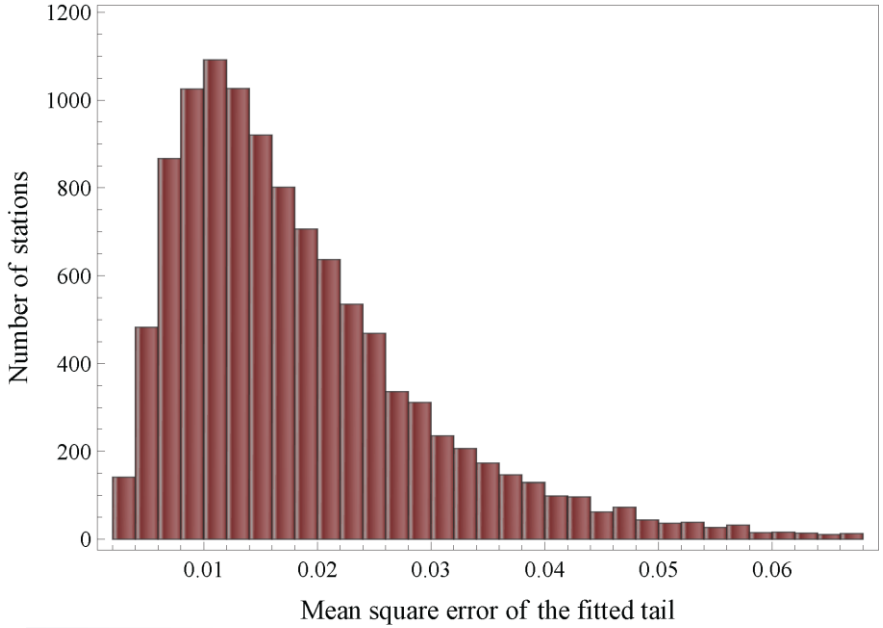


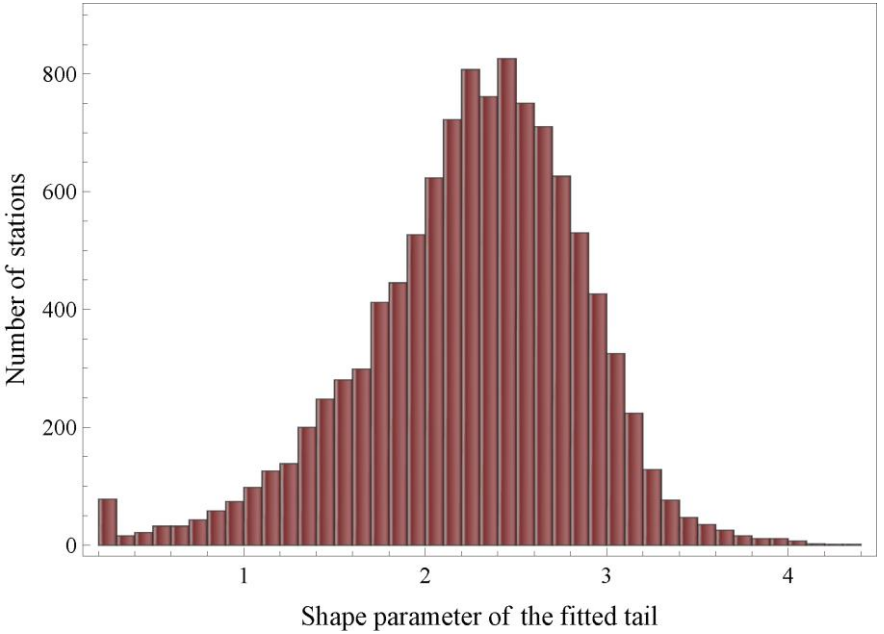


Fitting results: the Log Normal tail

	MSE	Scale parameter	Shape parameter
Min	0.0021	0.40	0.20
Mean	0.0183	0.76	2.26
Median	0.0157	0.75	2.32
Max	0.0680	1.53	4.34
Standard Deviation	0.0110	0.14	0.60

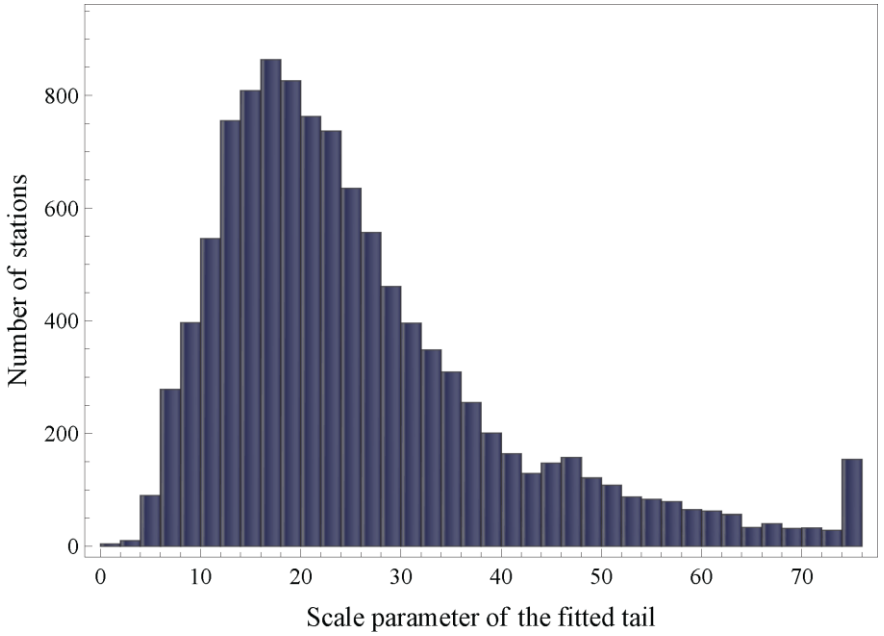


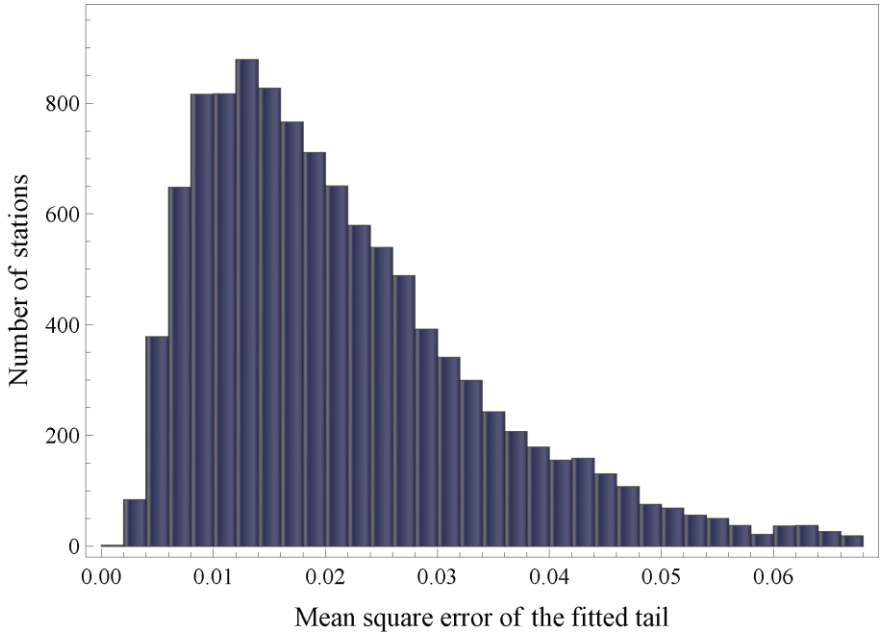


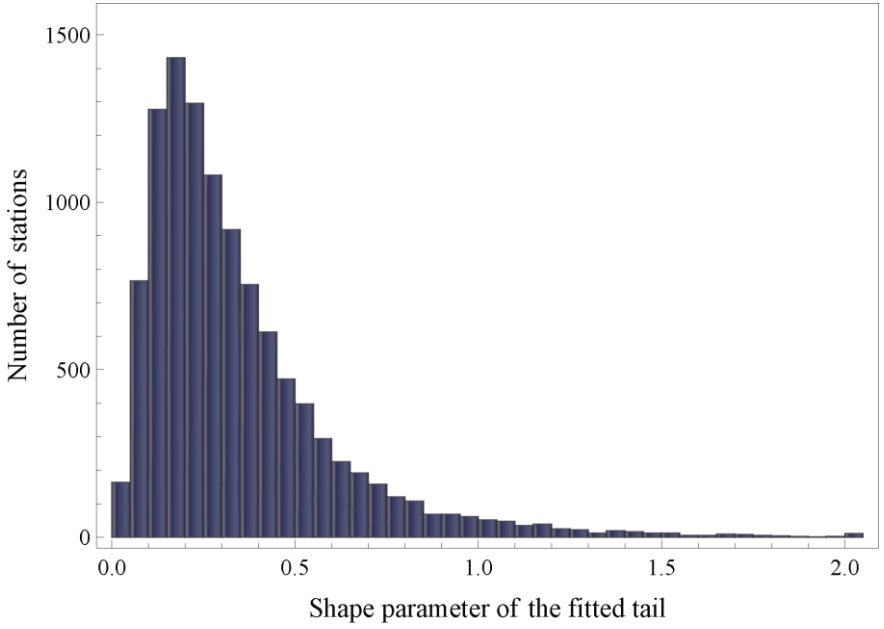


Fitting results: the Gamma tail

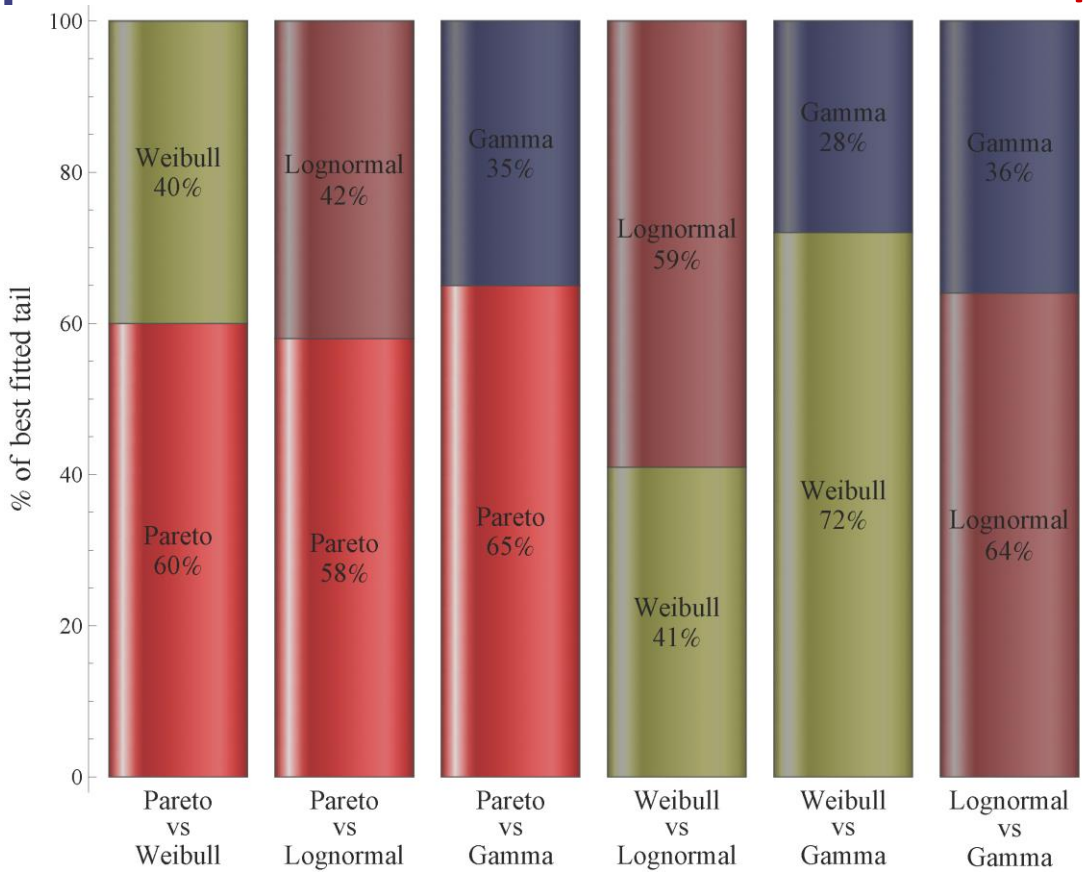
	MSE	Scale parameter	Shape parameter
Min	0.0020	0.82	0.01
Mean	0.0212	25.78	0.34
Median	0.0185	22.17	0.27
Max	0.0679	75.00	2.00
Standard Deviation	0.0123	14.57	0.26

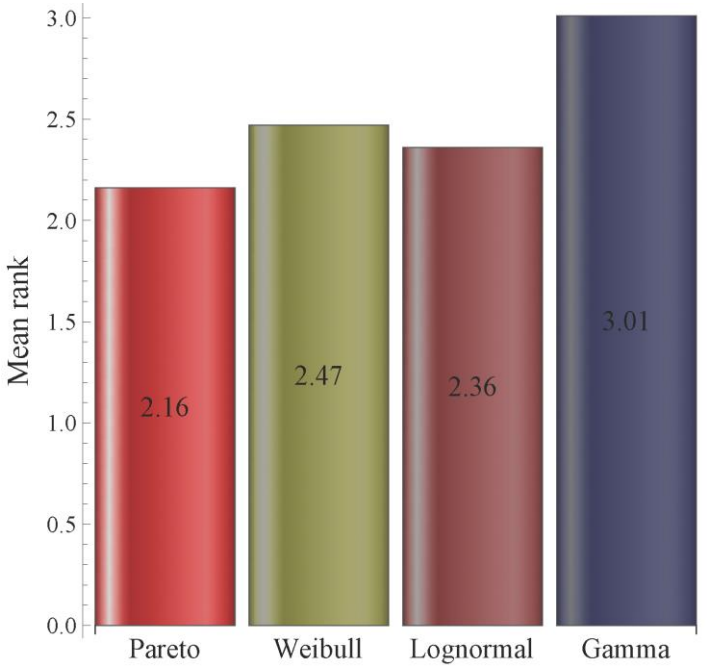






The fatter the better, the more popular the worse!

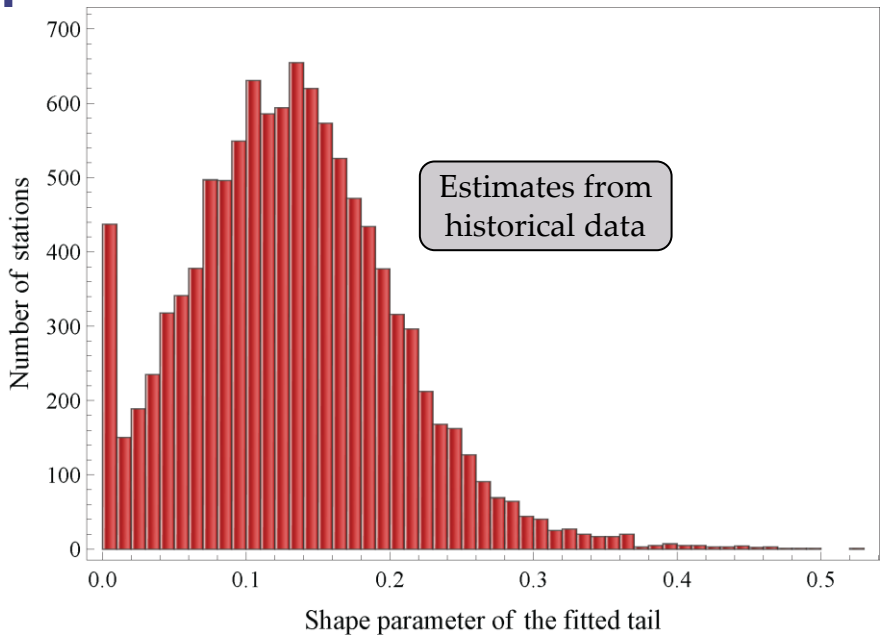


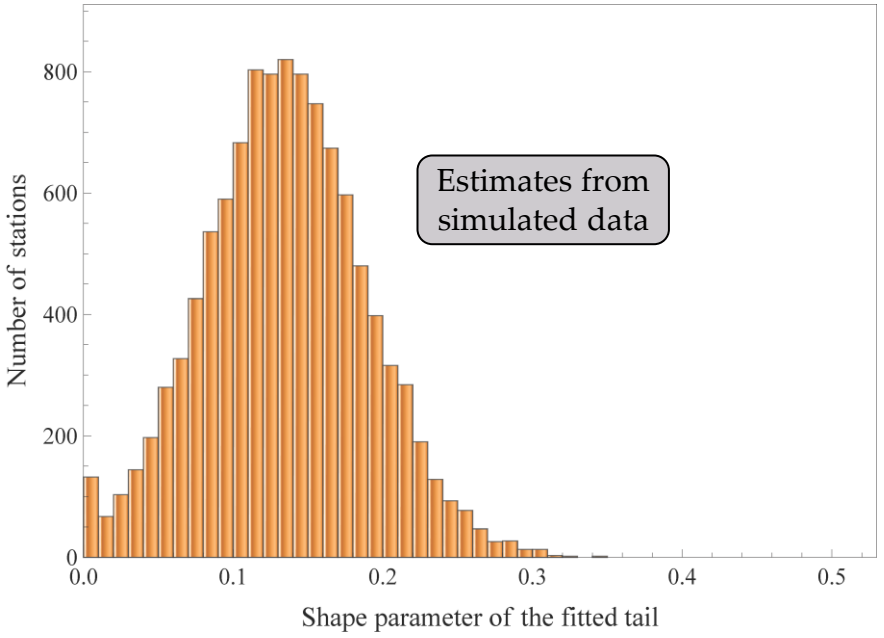


The four tails were compared in couples by means of the resulting MSE, i.e., the tail with the smaller MSE is considered best fitted. Within the couples, the Pareto tail was better fitted in approximately 60% of the stations. Interestingly, the fatter tail of each couple, in all cases, was better fitted in a higher percentage of the stations, i.e., the fatter, the better!

The four fitted tails in each station were ranked according to their MSE. The tail with the smaller MSE was ranked as 1 and the one with larger as 4. The figure depicts the mean rank of all stations. The Pareto is the best fitted, while the most common model, the Gamma, is the worst.

Can we assume a global tail index?





	Historical shape parameter	Simulated shape parameter	Historical MSE	Simulated MSE
Min	0.001	0.001	0.002	0.002
Mean	0.133	0.134	0.018	0.017
Median	0.130	0.134	0.015	0.014
95% CI	(0.017, 0.255)	(0.045, 0.224)	(0.006, 0.040)	(0.005, 0.036)
Max	0.530	0.349	0.068	0.124
Standard Deviation	0.072	0.054	0.011	0.010

To test the assumption of a global tail index (Pareto shape parameter) [e.g., *Koutsoyiannis*, 2004] we generated random samples from a Pareto distribution with tail index $\gamma = 0.13$ and lengths equal to the historical ones, and applied the same methodology to estimate the shape parameter. The sampling variability is as high as in the real world data—but the historical and simulated histograms are not identical.

Conclusions

- A four-parameter distribution, the Generalized Beta of the 2nd kind, derived by maximum-entropy considerations, can describe daily rainfall at all 11 500 examined locations.
- The same distribution can also describe rainfall at a wide range of time scales, from hourly to annual [Papalexiou and Koutsoyiannis, 2008b].
- Three-parameter special cases of this distribution, i.e. the Generalized Gamma distribution and the Burr type XII distribution can describe very large portions of the entire daily rainfall data set (97.6% and 87.7%, respectively).
- Two-parameter special cases of this distribution, i.e., the Pareto, the Weibull, and the Gamma, along with the widely used Lognormal distribution were tested for their ability to describe the distribution tails, and the following results were obtained:
  - In comparisons in pairs, the distribution with the heavier tail performed better.
  - Overall, the Pareto distribution performed best.
  - The most popular model, the Gamma distribution, performed worst.
- Overall, the investigation supports the general conclusions that:
  - Distributions bounded from above should be excluded.
  - Distributions with light tails (in particular, Gamma) are generally not appropriate.

References

- Ahmad, M., C. Sinclair, and A. Werritty (1988), Log-logistic flood frequency analysis, *Journal of Hydrology*, *98*(3-4), 205-224, doi:10.1016/0022-1694(88)90015-7.
- Buishand, T. A. (1978), Some remarks on the use of daily rainfall models, *Journal of Hydrology*, *36*(3-4), 295-308, doi:10.1016/0022-1694(78)90150-6.
- Euler, L., (1748), *Introductio in Analysin Infinitorum*
- Havrda, J., and F. Charvát (1967), Concept of structural a-entropy, *Kybernetika*, *3*, 30–35.
- Heo, J. H., J. D. Salas, and D. C. Boes (2001), Regional flood frequency analysis based on a Weibull model: Part 2. Simulations and applications, *Journal of Hydrology*, *242*(3-4), 171–182.
- Jaynes, E. T. (1957a), Information Theory and Statistical Mechanics, *Phys. Rev.*, *106*(4), 620, doi:10.1103/PhysRev.106.620.
- Jaynes, E. T. (1957b), Information Theory and Statistical Mechanics. II, *Phys. Rev.*, *108*(2), 171, doi:10.1103/PhysRev.108.171.
- Koutsoyiannis, D. (2004), Statistics of extremes and estimation of extreme rainfall, 2, Empirical investigation of long rainfall records,*Hydrological Sciences Journal*, *49*(4), 591–610.
- Koutsoyiannis, D. (2005), Uncertainty, entropy, scaling and hydrological stochastics, 1, Marginal distributional properties of hydrological processes and state scaling, *Hydrological Sciences Journal*, *50*(3), 381-404.
- Mielke, P. W. (1973), Another Family of Distributions for Describing and Analyzing Precipitation Data, *Journal of Applied Meteorology*, *12*(2), 275-280.
- Mitzenmacher, M. (2004), A brief history of generative models for power law and lognormal distributions, *Internet mathematics*, *1*(2), 226–251.
- Papalexiou, S., and D. Koutsoyiannis (2008a), Ombrian curves in a maximum entropy framework, in *European Geosciences Union General Assembly 2008*, p. 00702. (www.itia.ntua.gr/en/docinfo/851/).
- Papalexiou, S.M., and D. Koutsoyiannis (2008b), Probabilistic description of rainfall intensity at multiple time scales, *IHP 2008 Capri Symposium: "The Role of Hydrology in Water Resources Management"*, Capri, Italy, UNESCO, International Association of Hydrological Sciences (www.itia.ntua.gr/en/docinfo/884/).
- Tadikamalla, P. R. (1980), A Look at the Burr and Related Distributions, *International Statistical Review / Revue Internationale de Statistique*, *48*(3), 337-344. Brazauskas, V. (2002), Fisher information matrix for the Feller-Pareto distribution, *Statistics & Probability Letters*, *59*(2), 159-167, doi:10.1016/S0167-7152(02)00143-8.
- Tsallis, C. (1988), Possible generalization of Boltzmann-Gibbs statistics, *Journal of Statistical Physics*, *52*(1), 479-487