European Geosciences Union
General Assembly 2010

Vienna, Austria, 02 - 07 April 2010

Session: NP3.4
Geophysical Extremes: Scaling representations and their applications

On the tail of the daily rainfall probability distribution: Exponential-type, power-type or something else?

S.M. Papalexiou and D. Koutsoyiannis
Department of Water Resources and Environmental Engineering, National Technical University of Athens, Greece (www.itia.ntua.gr)

Contact info: sp@itia.ntua.gr
Available online at: http://www.itia.ntua.gr/en/docinfo/977/

1. Abstract

While the traditional choice to describe the wet-day daily rainfall is the two-parameter Gamma distribution, many other distributions have been proposed and used, e.g., the two- and three-parameter Log-Normal, the Generalized Logistic, the Pearson Type III, the Pareto and the Generalized Pareto, the three- and four-parameter Kappa distributions, and many more. The asymptotic behaviour of the upper tails of these probability distributions may be generally categorized in two families: the exponential-type and the power-type tail families, where the latter family does not have all moments finite. However, there are exceptions such as the Log-normal family and the so-called stretched-exponential-tail family, which are generally acknowledged to be heavy-tailed, yet all their moments exist. The upper tail of the distribution governs both the magnitude and the frequency of the extreme events with the exponential-type distribution tails generating more "mild" and infrequent extremes compared to the power-type tails. This emphasizes the importance to assess correctly the tail behaviour and also to theoretically justify it. In general, the exponential-type distribution tails are the most commonly assumed; however, in the last years there is a shift towards the power-type-tail distributions. In this study, we investigate the assumption that the tail belongs to the stretched-exponential-type family that seems to be the "middle" way between exponential- and power-type tails. Finally, we use real-world daily rainfall datasets to examine this assumption empirically and to compare the performance of each tail family.

2. Notes on the distribution tail

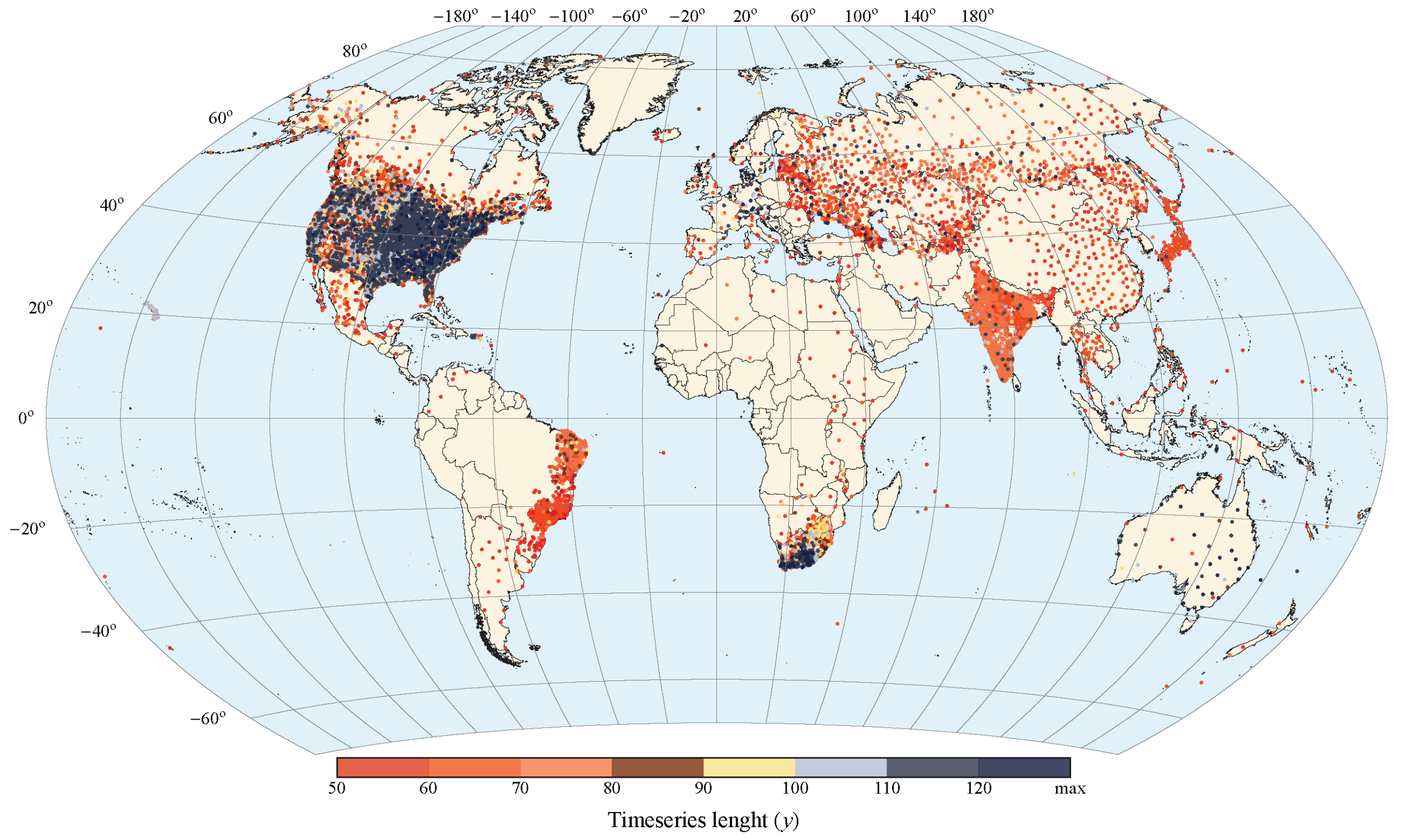
- The upper tail of the distribution is its most important part as it rules both the magnitude and the frequency of the extreme events. The heavy-tailed distributions, whose probability density function goes to zero less rapidly than the exponential, result in more frequent and more extreme events compared to exponential distributions.
- The typical procedure of selecting a distribution to describe a random variable like rainfall, is (a) to select a priori some of the many parametric families of distributions, (b) estimate the parameters, and (c) choose the best fitted. Nevertheless, this procedure does not guarantee that the selected distribution will model adequately the tail behavior. On the contrary, as only a very small portion of the empirical data belongs to the tail (unless a very large sample is available), all fitting methods will be "biased" against the tail, i.e., the estimated parameters will result in a distribution that tries to best describe the largest portion of the data. Additionally, the shortage of data belonging to the tail, makes almost impossible to reveal an ill-fitted tail.
- Clearly, an ill-fitted tail may result in a very serious error with severe consequences in the hydrological design. For example, the magnitude of the 1000-year precipitation may be seriously underestimated if it is calculated from an exponential distribution rather than a heavy-tailed distribution.
- Especially for daily rainfall, the most commonly used models, even today, belong to the exponential family, e.g., the Gamma distribution. However, several studies suggest that heavy tailed distributions may be more suitable. For example, a pioneering study by Milke [1973] proposed the use the Kappa distribution, a heavy tailed distribution, to describe daily rainfall.

3. Common distributions and their tails

Distribution	Tail function $\bar{F}_X(x)$	Comments
Pareto	$\bar{F}_X(x) = \left(1+\gamma\frac{x}{\beta}\right)^{-\frac{1}{\gamma}}$	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	$\bar{F}_X(x) = \exp\left(-\frac{x^{\gamma}}{\beta}\right)$	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.
Log-Normal	$\bar{F}_X(x) = \frac{1}{2}\operatorname{erfc}\left(\frac{\ln x-\gamma}{\sqrt{2}\beta}\right)$	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].
Gamma	$\bar{F}_X(x) = \frac{\Gamma\left(\gamma,\frac{x}{\beta}\right)}{\Gamma(\gamma)}$	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.

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

4. The datasets



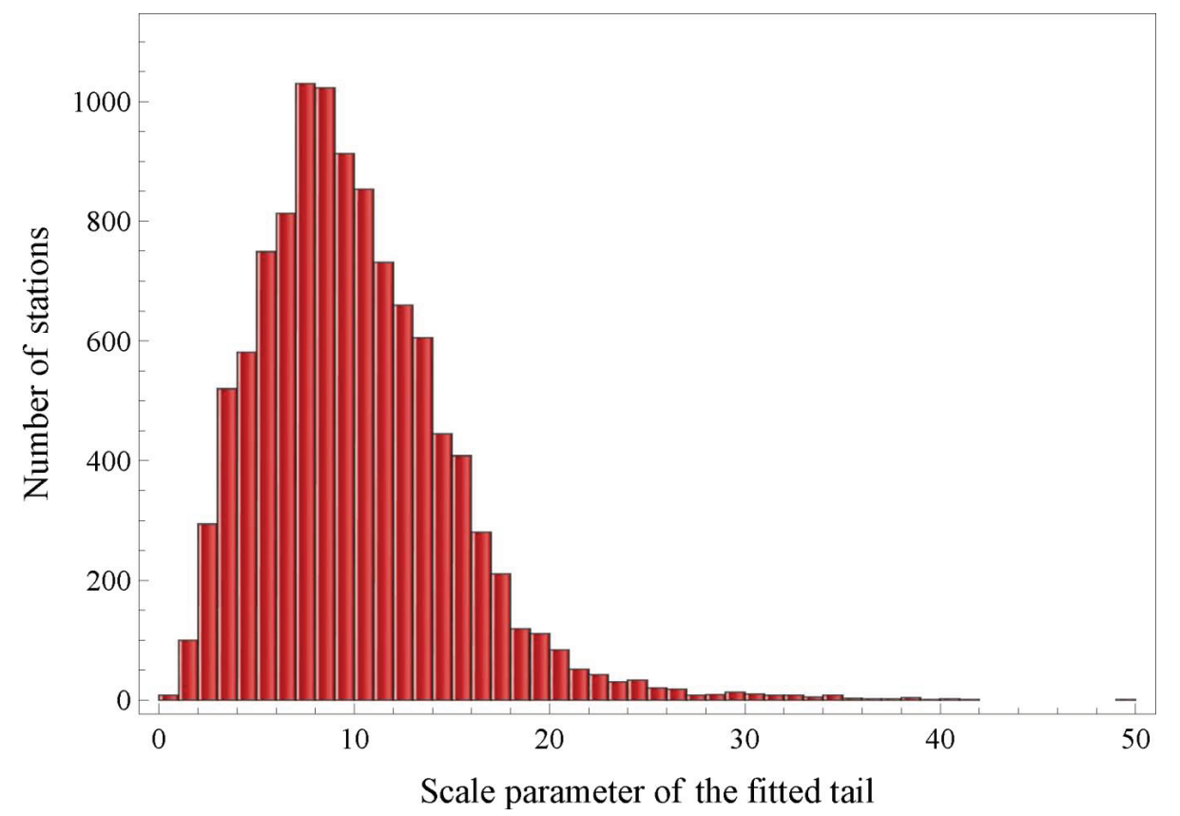
The figure depicts the station locations used in this study. All have daily rainfall time series with length greater than 50 years. The stations are a subset of the Global Historical Climatology Network-Daily database, that contains observations of several meteorological variables at more than 40 000 stations that are distributed across all continents.

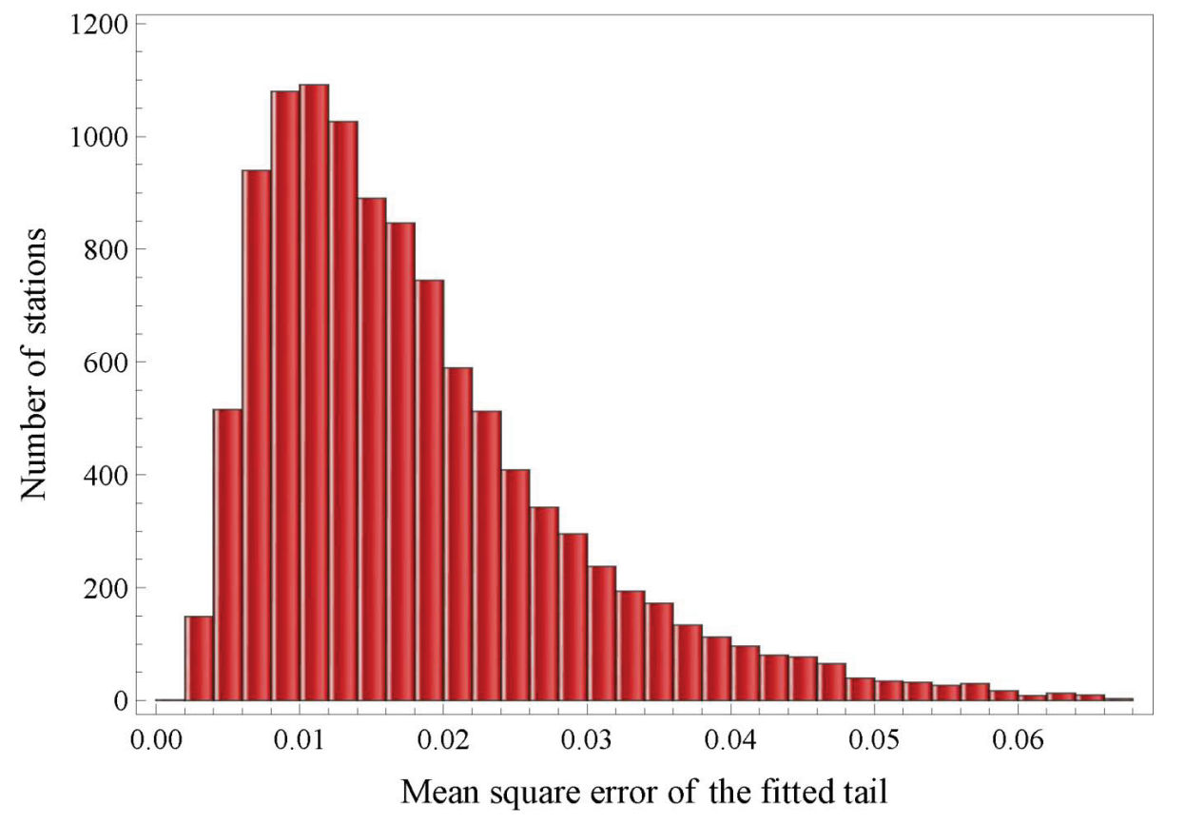
5. Fitting distribution tails to empirical data

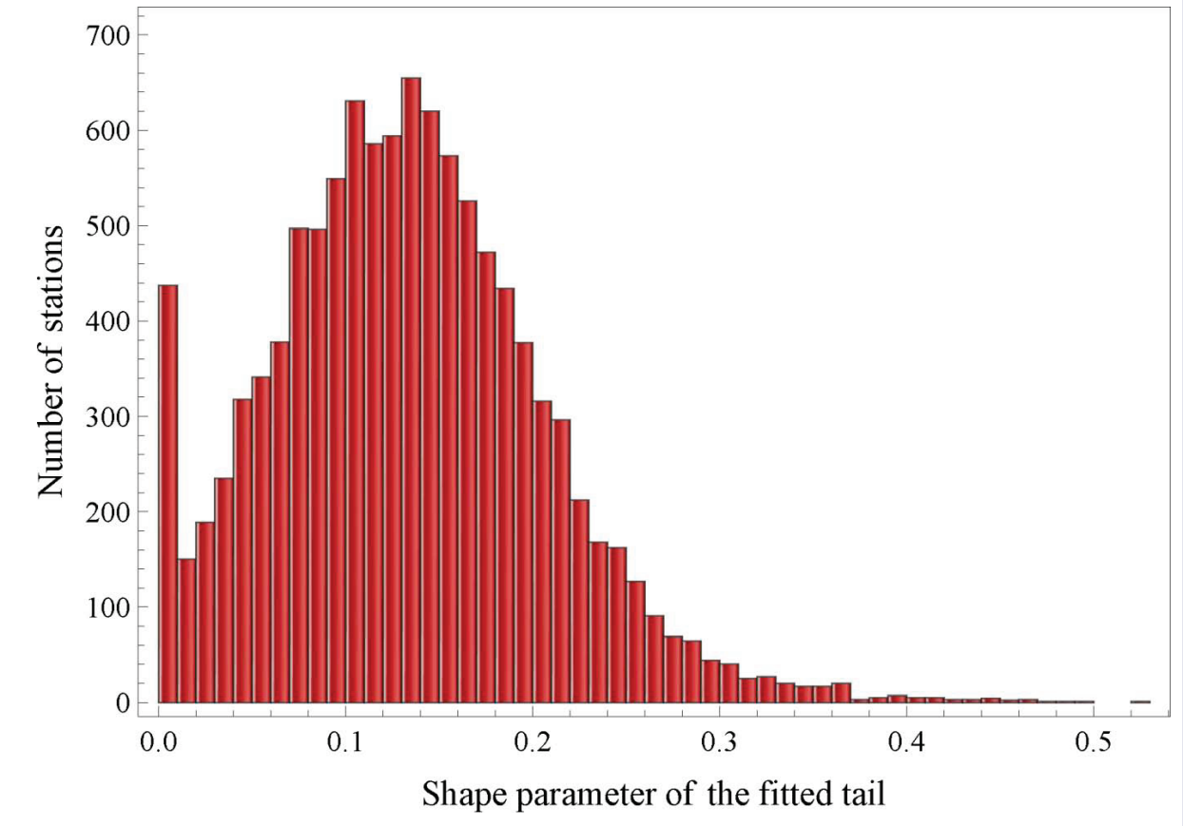
- The daily precipitation data of a total of 10 817 stations with records larger than 50 years and matching some quality criteria (e.g., very small percentage of missing values) were investigated regarding their distribution tail.
- The probability distribution of the daily rainfall is of mixed type, i.e., one part of the distribution is discrete (a finite probability of a dry day), and one part is continuous, describing the magnitude of the wet-day rainfall. Thus, the theoretical tail, denoted as $\bar{F}_X(x)$, is the upper part of the continuous distribution.
- Each station has a record length of N years and a total number n of non-zero values depending on the probability dry. We defined the empirical distribution tail, $\bar{F}_n(x_i)$, as the empirical probability (according to the Weibull plotting position) of the N largest non-zero rainfall values, i.e., $\bar{F}_n(x_i) = r(x_i)/(n+1)$, 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 distribution tails were chosen and fitted to the empirical tails, the Pareto, the Weibull, the Log-Normal and the Gamma.
- The theoretical tails were fitted to the empirical tails by minimizing a modified mean square error (MSE) norm defined as $\mathrm{MSE} = (\bar{F}_X(x_i)/\bar{F}_n(x_i) - 1)^2/N$. This norm is superior to the classic MSE norm as it treats equally each data point.

6. 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

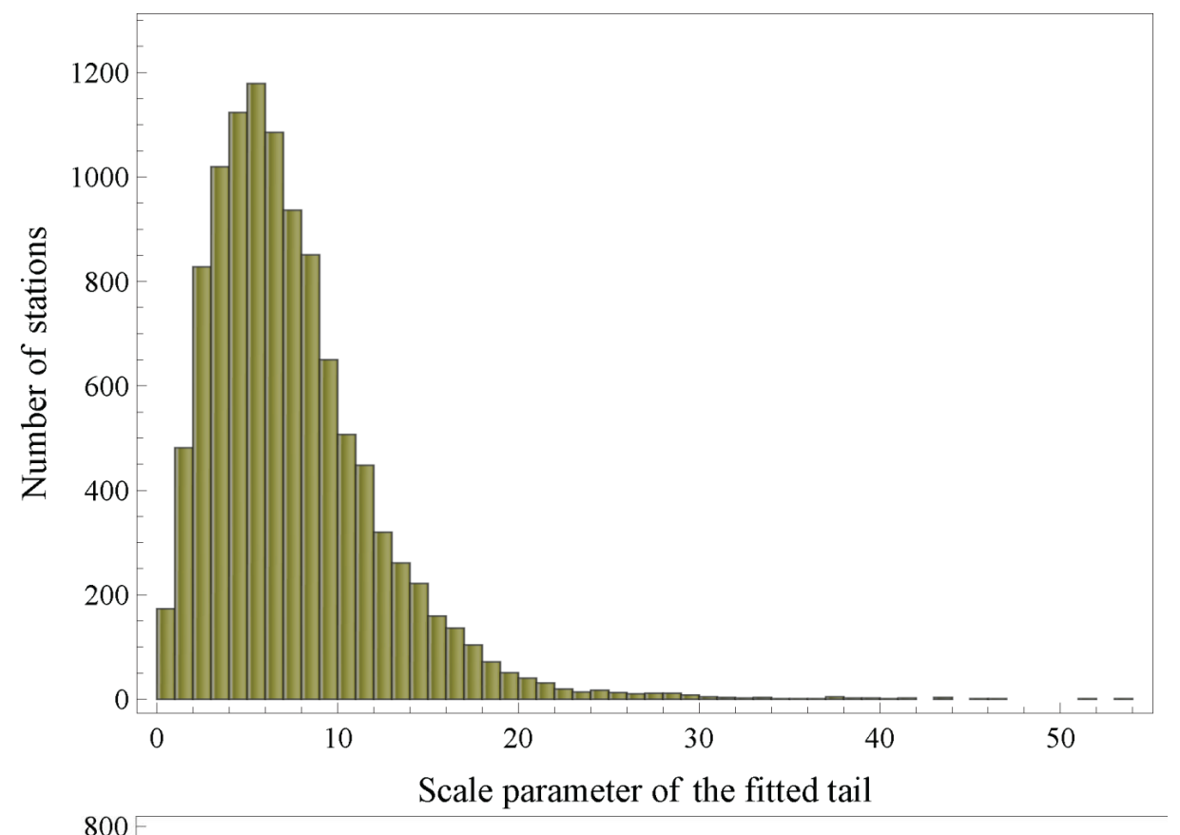


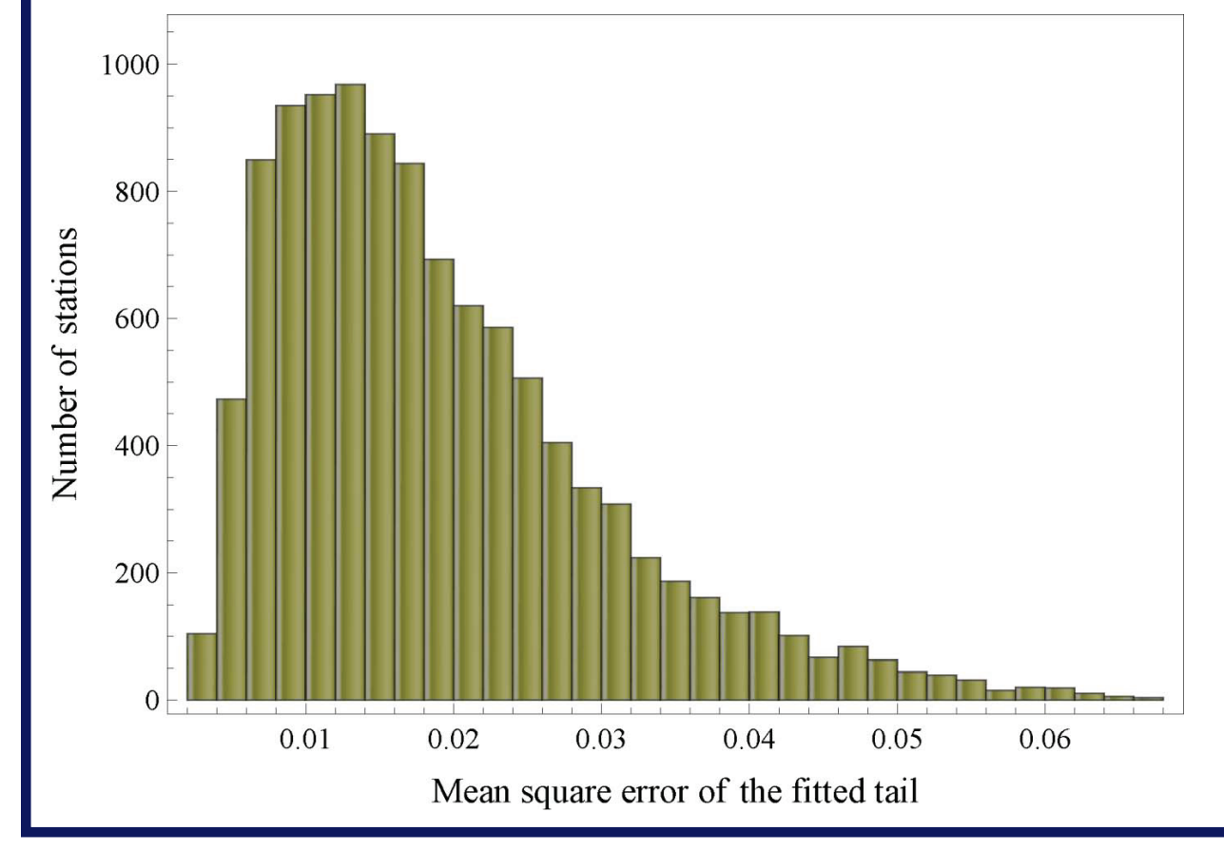


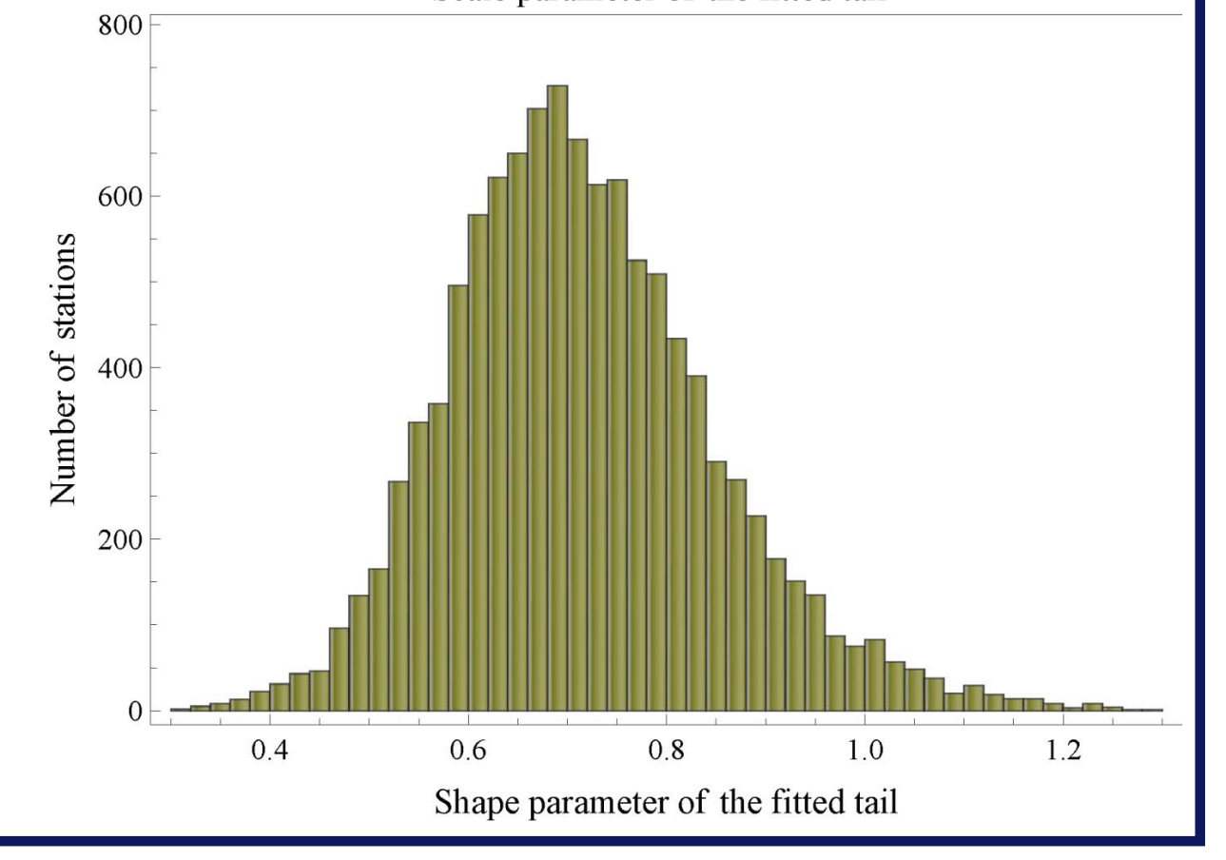


7. Fitting results: the Weibull tail

	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

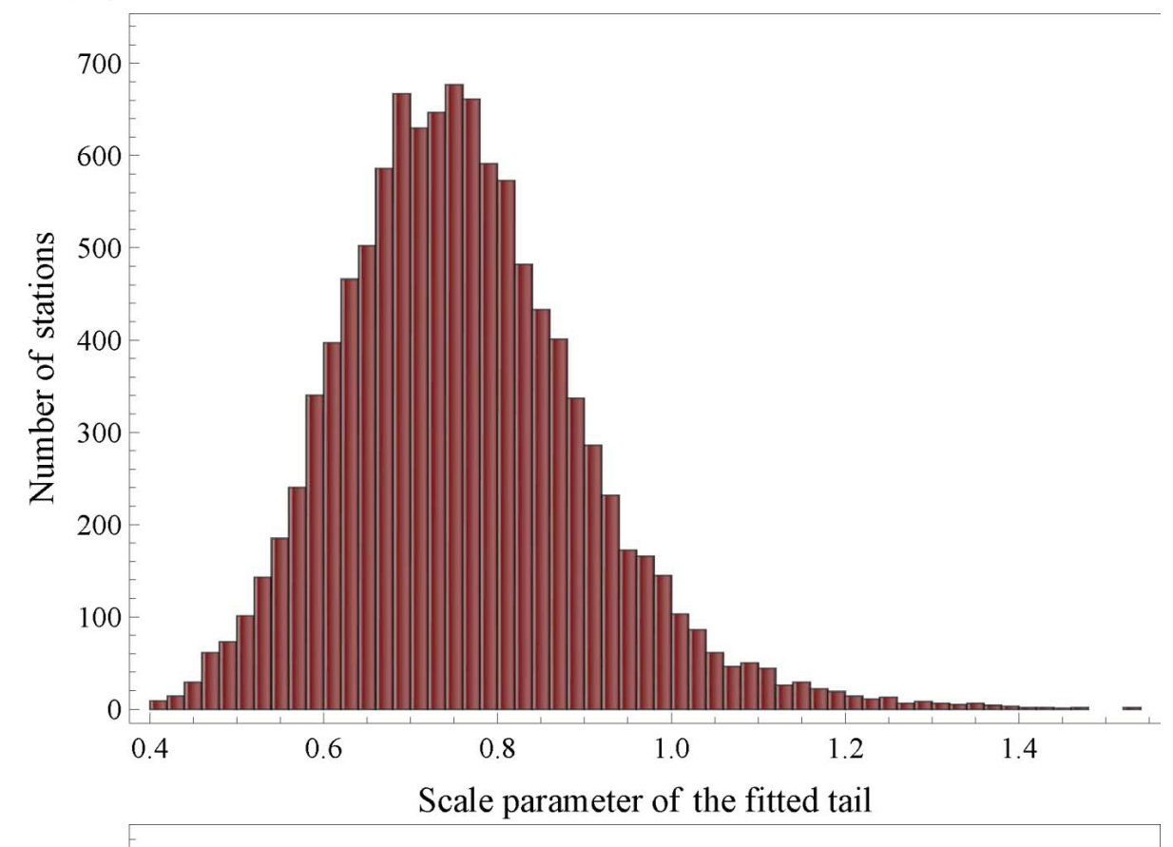


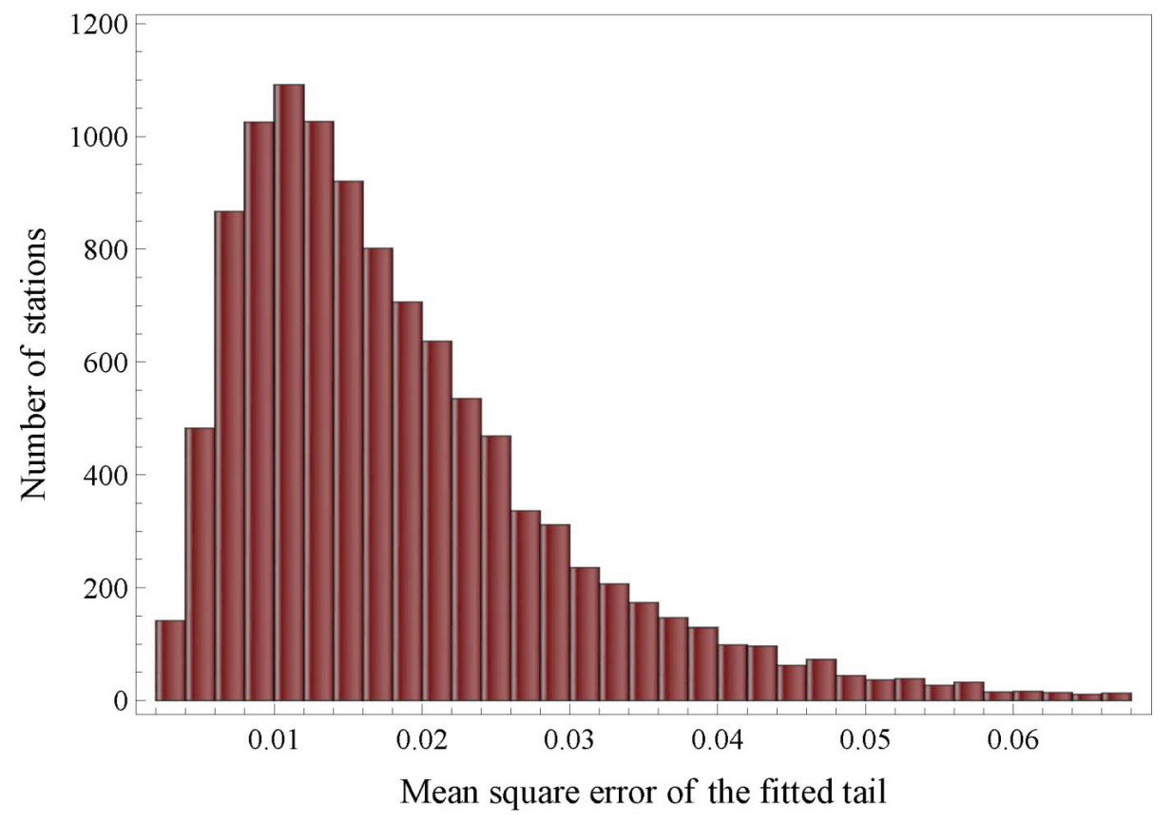


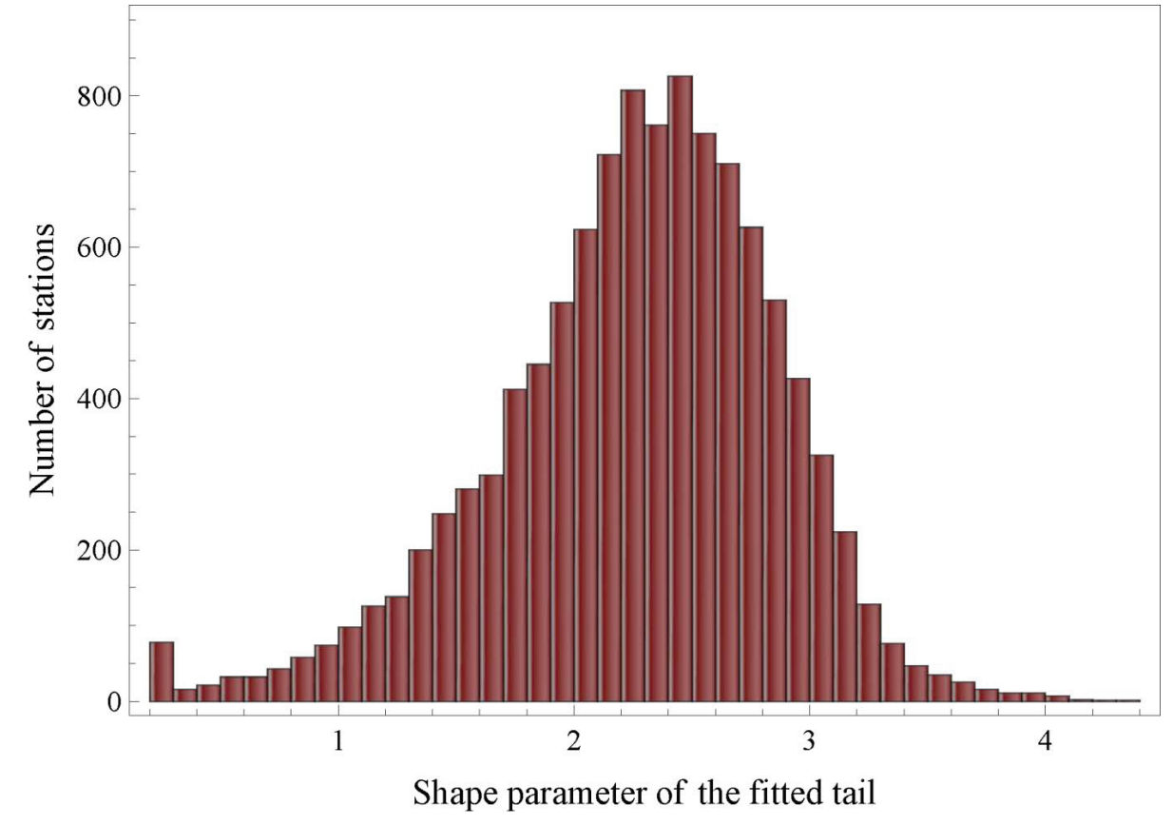


8. Fitting results: the Log Normal tail

	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

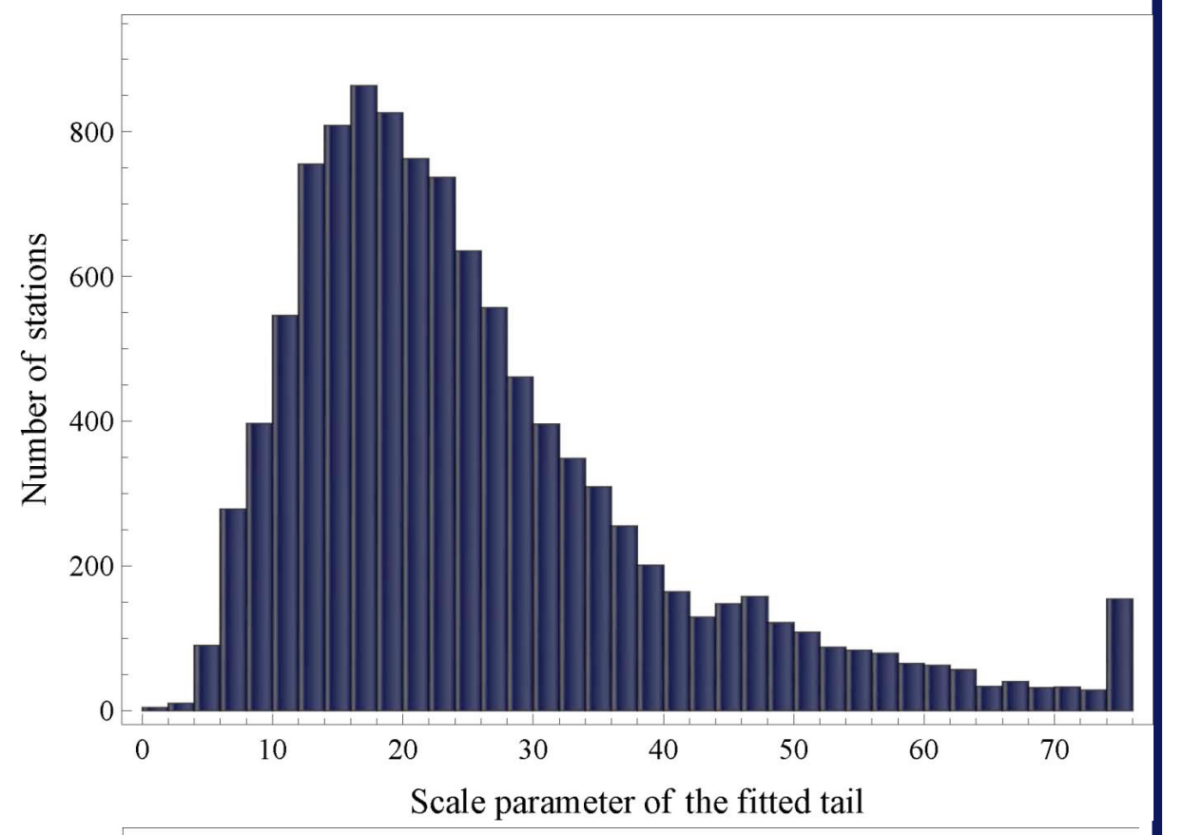


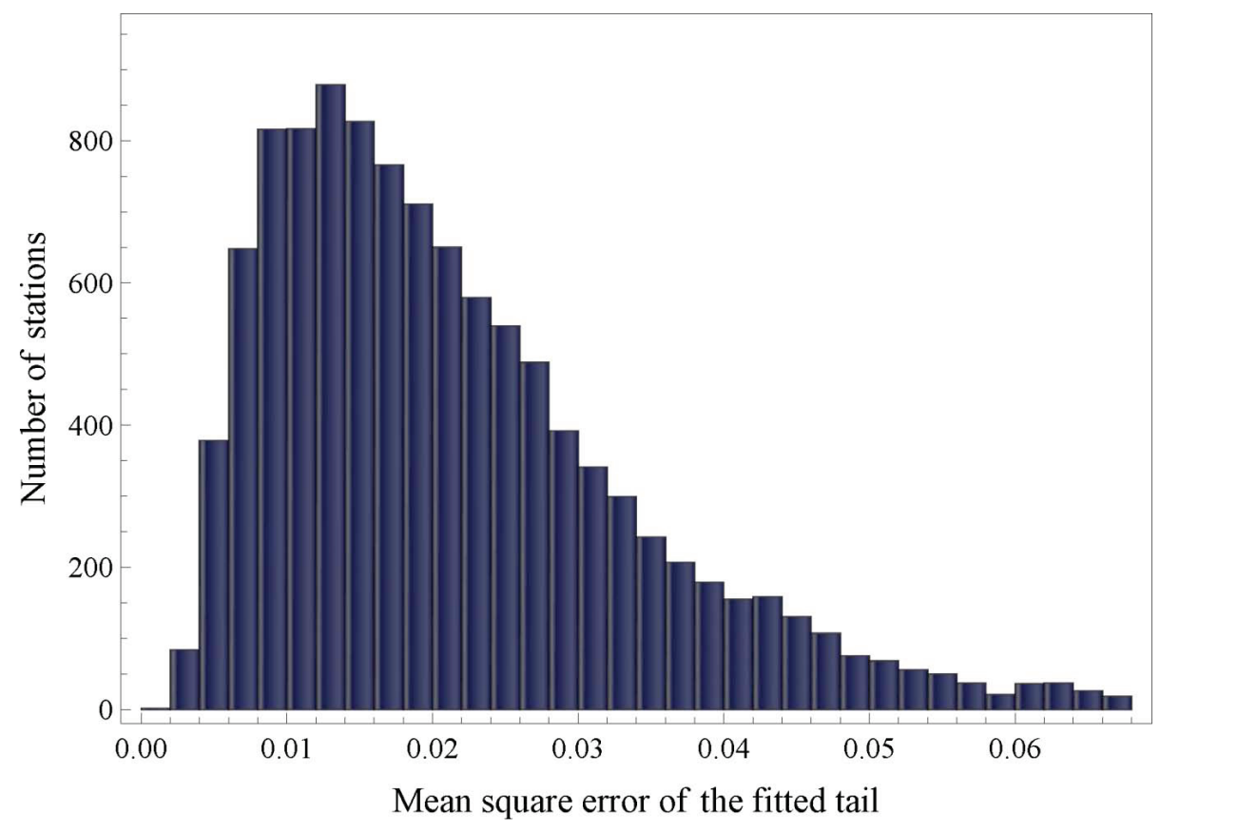


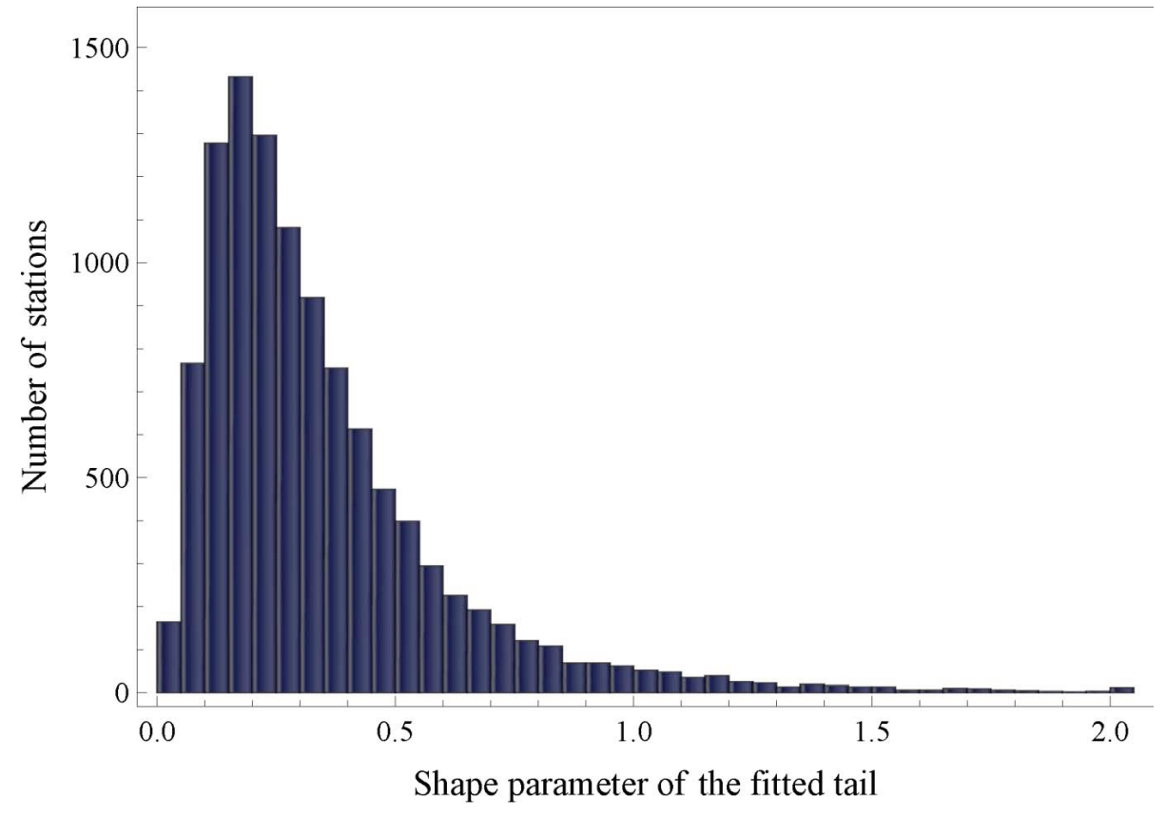


9. Fitting results: the Gamma tail

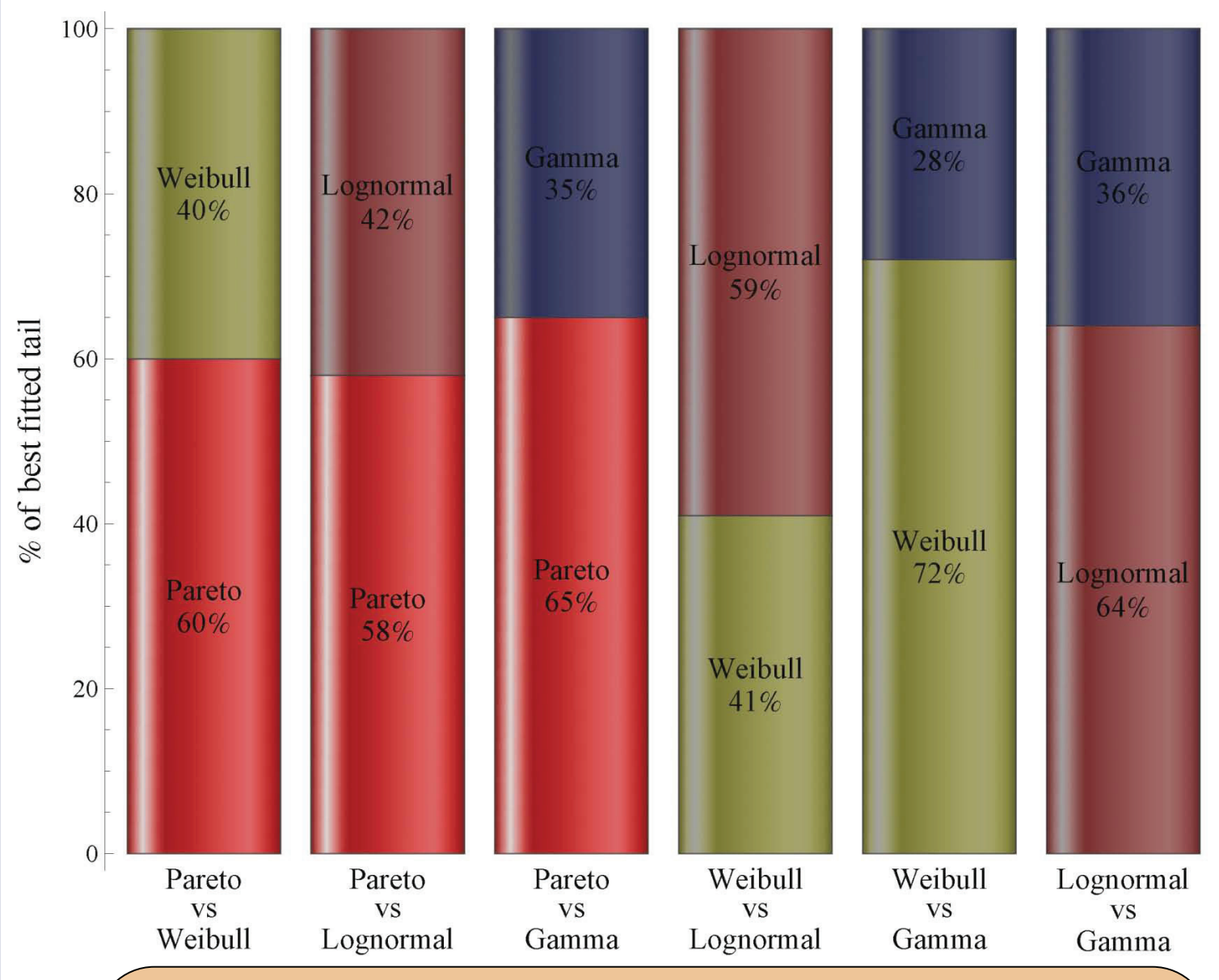
	MSE	Scale parameter	Shape parameter
Min	0.0020	0.82	0.01
Mean	0.0212	25.78	0.34
Median	0.0185	22.17	0.27
Max	0.0679	75.00	2.00
Standard Deviation	0.0123	14.57	0.26



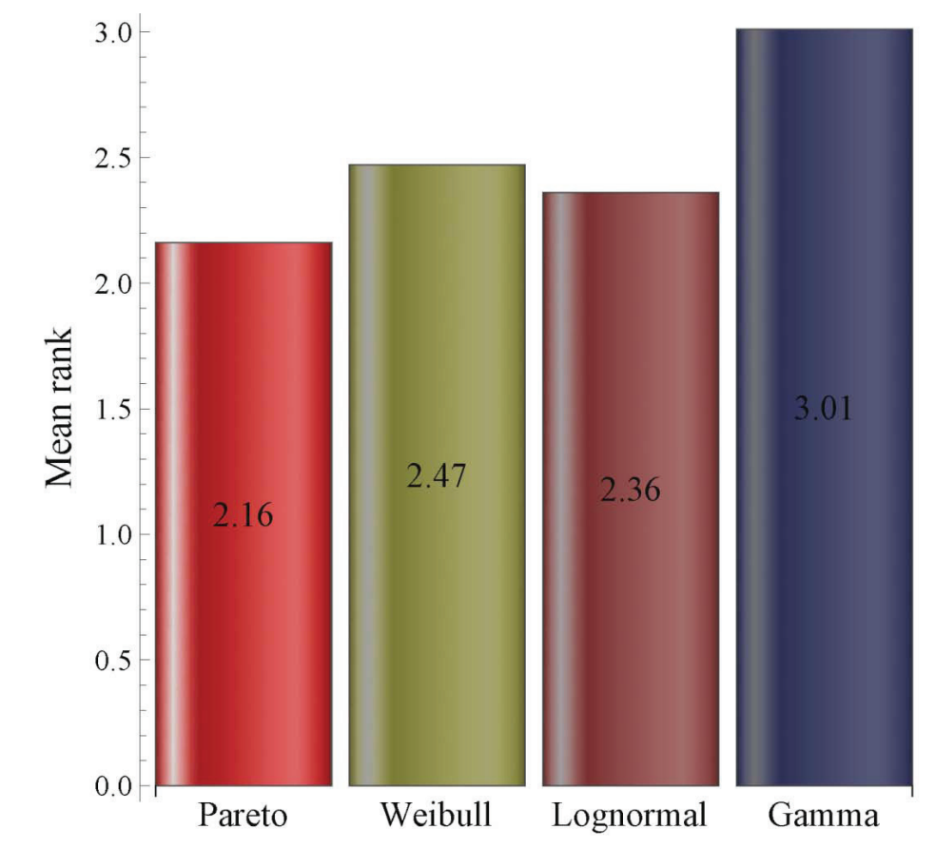




10. The fatter the better, the more popular the worse!

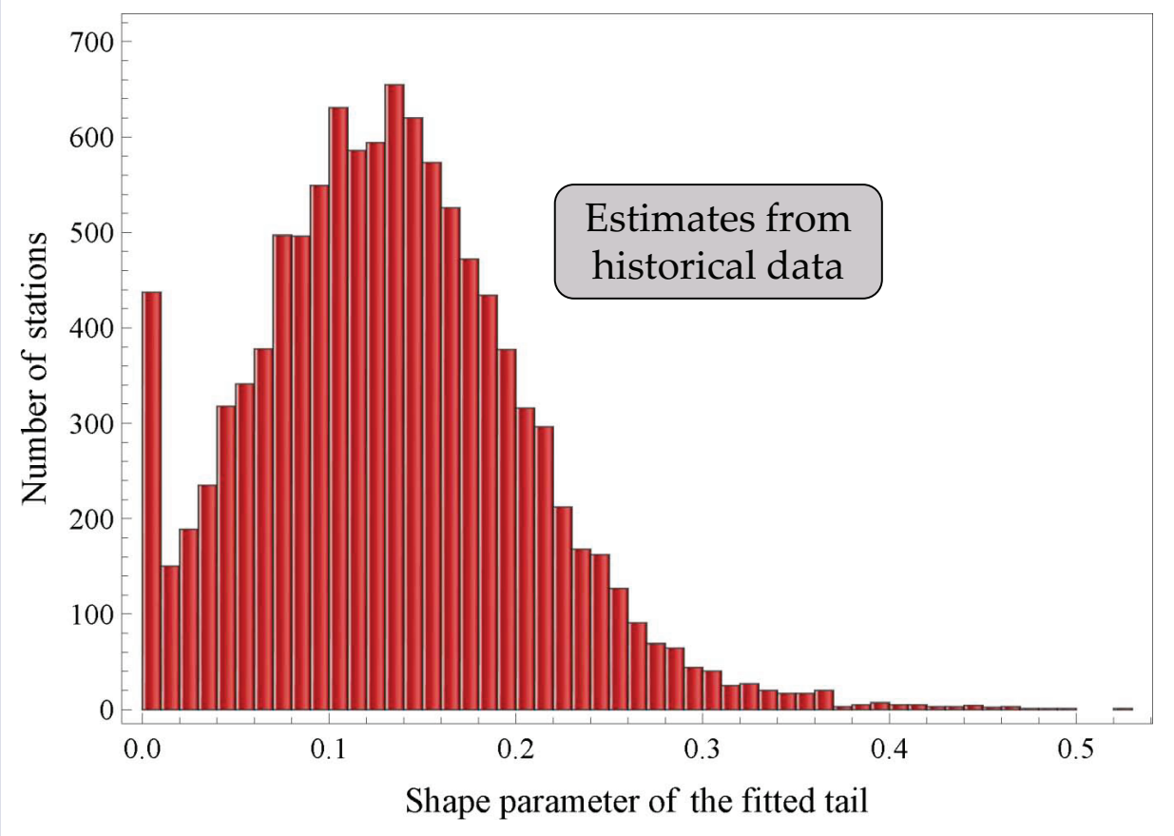


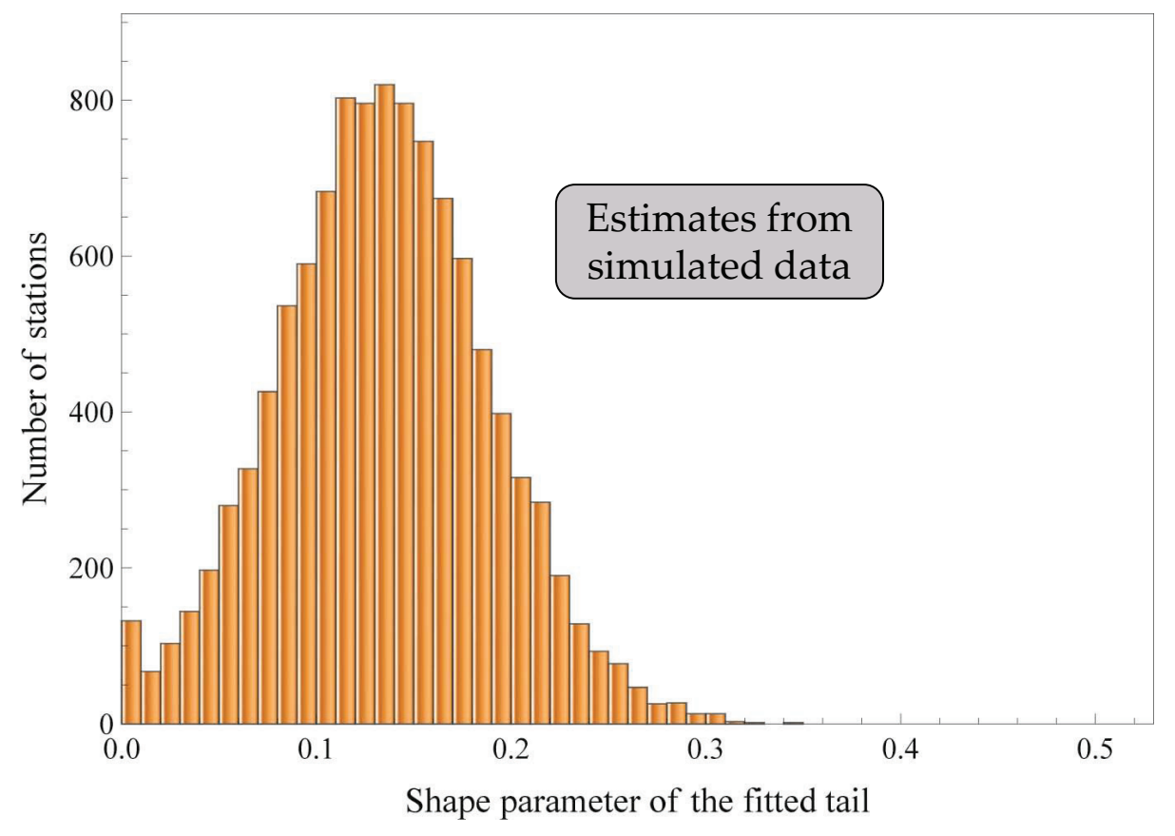
In order to assess the best fitted tail, the four tails were compared in couples by means of the resulted MSE, i.e., the tail with the smaller MSE is considered best fitted. Within the couples, the Pareto tail was better fitted in approximately 60% of the stations. Interestingly, the fatter tail of each 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.

11. 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

In order to check the assumption of a global tail index (Pareto shape parameter) [see e.g., *Koutsoyiannis*, 2004] we generated random samples from a Pareto distribution with tail index $\gamma = 0.13$. We generated 10817 samples, with lengths equal to the historical ones, and the same methodology was applied to estimate the shape parameter. The simulation revealed a high variability of the shape parameter that describes a large portion of the empirical distribution; however, the historical and simulated histograms are not identical. Nevertheless, it seems plausible that the parameter belongs to a narrow range e.g., from 0.1 to 0.2 but this needs to be further investigated.

12. Conclusions

- The daily rainfall data of a total number of 10 817 stations were investigated regarding their distribution tail.
- Four common models were fitted, the Pareto, the Weibull, the Log-Normal and the Gamma distribution with complete emphasis on the tail.
- The model that performed the best was the Pareto, which is a heavy-tailed, asymptotically power law distribution. Additionally, the comparison of the four distributions in pairs revealed that the one with the heavier tail performed better.
- The assumption of a global Pareto tail index was also investigated. A large portion of the empirical distribution of the Pareto tail index can be described if a global tail index is assumed equal with 0.13, however, a single (global) value fails to completely describe the whole empirical distribution . Nonetheless, it appears that a tail index with values in narrow range, i.e., from 0.1 to 0.2, would reproduce completely the empirical histogram of tail indices, but this needs further study.
- Finally, the most popular model, the Gamma distribution, performed the worst.

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.
- 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.
- 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.
- 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.
- Tadikamalla, P. R. (1980), A Look at the Burr and Related Distributions, *International Statistical Review / Revue Internationale de Statistique*, *48*(3), 337-344.