Hydrol. Earth Syst. Sci. Discuss., 9, C4773–C4782, 2012
www.hydrol-earth-syst-sci-discuss.net/9/C4773/2012/

Hydrology and Earth System Sciences Discussions

Interactive comment on "How extreme is extreme? An assessment of daily rainfall distribution tails" *by* S. M. Papalexiou et al.

S. M. Papalexiou et al.

smp@itia.ntua.gr

Received and published: 14 October 2012

RC: Reviewer's comments
AC: Authors' comments

We honestly thank R. Deidda for providing us with a detailed review. Below we answer to all of his comments but in general we disagree with his main suggestion to use the Generalized Pareto distribution, while as it will be apparent below, our study does not contain any incorrect interpretation of the shape parameters. Yet, we found some of his comments very useful. We believe that there should be no ambiguity regarding to what this study is about; both the Introduction section and the Methods section make clear that this study regards the tails and not the whole distribution.

RC1. *"the heavier, the better" (page 5766, line 15) and other similar sentences (in several parts of the manuscript) can be misleading. Indeed, distributions fitted to the highest rainfall values can be characterized by a wide range of shape parameters and degrees of skewness. Thus, in my opinion, it is not that important to catch the highest tails, but the distribution that reliably fits heavy tailed as well as exponentially distributed records, as it is the case in this and other studies. Using rainfall records restricted to a limited region, I made some analyses on the tails of daily rainfalls (see reference below) and found that distributions are often heavy tailed but, in some cases, they can also be exponential. I have also found that the shape parameters can display patterns depending on orography (manuscript in preparation). Wilson and Toumi (2005) used the stretched exponential distribution (i.e. the Weibull in eq.(6) provided by the Authors) on a worldwide daily rainfall database and showed a dependence of the shape parameter on the geographic location. In some cases, the shape parameter of the stretched exponential was equal to 1, indicating an exponential tail. Thus, if the region of interest lies where the shape parameter is very close to 1, I would say "the exponential, the better"! In conclusion, I strongly suggest to reformulate some sentences in the manuscript to avoid misleading emphasis and, also, better convey the message "rainfall can display a wide range of more or less tailed extremes, the XYZ distribution can fit better whatever the shape parameter is".*

AC1. "The heavier, the better" resulted as a general rule. Fig. 4 makes clear that there is a significant percentage of records that are better described by a light tail, like the Gamma tail. Yet, the reviewer may be right and we should stress in the manuscript that the general rule that the heavier tail is the better one does not apply in all stations. In Papalexiou and Koutsoyiannis (2012a) we performed a worldwide analysis of the shape parameter of the GEV distribution and indeed there is a geographical dependence of the shape parameter value; consequently thus, there should be a dependence in the stretched exponential parameter as it quantifies the extreme behaviour of the rainfall. We note though that here we did not investigate the geographical variation of the parameters. However, it is something we were thinking to do in a future study. In any case, we will reformulate some sentences in a revised manuscript.

RC2. *Why not Generalized Pareto distribution? On Page 5764, lines 7-11, the four distributions used for comparisons are classified as follows: "The Pareto and the Lognormal distributions belong to the sub-exponential class and are considered heavy-tailed distributions. The Weibull and the Gamma distribution, depending on the values of the shape parameter, can belong to both classes, but in general their tails are lighter than the Pareto or the Lognormal". Why the Authors avoid using the Generalized Pareto distribution (GPD), which also includes the Pareto type II distribution used by the Authors? The Generalized Pareto distribution has the advantage to describe heavy tailed (subexponential) distributions for positive values of the shape parameter, the exponential distribution (shape parameter equal to zero) and hyper-exponential distributions (negative shape parameter values). Recent studies by Begueria (2005), Deidda and Puliga (2006), and Begueria et al. (2009) reported strong evidence [based on L-moment ratio diagrams (Hosking, 1990)] that GPD is the best candidate for daily rainfall series. In addition, there are theoretical arguments to substantiate the use of GPD in fitting the excesses above proper thresholds (see e.g. Coles, 2001, Deidda, 2010 and references therein), while the adoption of Lognormal, Weibull or Gamma distribution models is not supported by extreme value theory. Concluding, I strongly suggest the use of a GPD rather than a Pareto type II model. The latter is included in the GPD family, corresponding to positive values of the shape parameter.*

AC2. The difference of the Generalized Pareto(GP) and the Pareto II (PII) is that the former has an additional location parameter α. This means that the GP is defined for $x \geq \alpha$ which is theoretically inconsistent with the nature of rainfall, while the PII is defined for $x \geq 0$ which is consistent. If the shape parameter values of the GP as well as of the PII are negative then the distribution is bounded above implying thus an upper limit in the rainfall which we believe is a major fallacy (see our reply to A. Clauset). The

reviewer refers to some studies based on L-moments claiming that they provide strong evidence for the adoption of the GP for daily rainfall. Nevertheless, these studies are based on the analysis of a small number of records. We wish to refer the reviewer to our other study (Papalexiou and Koutsoyiannis, 2012b) where we analysed more than 10000 daily rainfall records also using L-moments diagrams and the results do not indicate that the GP distribution is a good model for the daily rainfall in general. Additionally, an ongoing study where we analysed daily rainfall on a monthly basis (approximately in 200000 time series) verifies these results. An additional reason for using PII rather than GP is that we wanted to make a fair comparison of the different distribution and therefore we chose the two-parameter versions for all of them. Finally, there is a common problem in the theoretical background that allegedly supports the use of the GP distribution. The results obtained by Balkema and de Haan (1974) and Pickands III (1975) are limiting laws, i.e., valid assuming that the threshold tends to infinity. If the original distribution is power type or exponential then indeed there is fast convergence and indeed values above a threshold can be modelled by the GP or the exponential distribution, respectively. However, the convergence for many other cases, e.g., Lognormal, stretched exponential etc. is too slow. For some examples exhibiting the slow convergence to the classical limiting distributions of EVT (the same philosophy applies for Balkema-de Haan-Pickands theorem) the reviewer can see slide no. 3 in Papalexiou and Koutsoyiannis (2012a) (presentation of a study under review) and Fig. 1 in Koutsoyiannis (2004). On the contrary, our approach, which consists of fitting directly four very common tails and comparing the fitting results is free of this problem.

RC3. *The first subplot in Figure 6 (top left), displays the shape parameter estimates for the Pareto type II distribution model. The irregularity in the unexpected large number of records filling the first bin, should be due to the use of a Pareto type II distribution model. More precisely, I suppose that the numerical algorithms for parameter estimation were bounded to provide ONLY positive shape parameter values and to avoid de-*

generacy of eq.(6)) when the shape parameter approach zero (exponential tail). From Table 1, one concludes that the lower bound set by the Authors should be equal to 0.001. Following my suggestion to fit a Generalized Pareto rather than a Pareto type II distribution model (see above), the Authors should also find shape parameter values equal to zero (exponential distribution) or smaller than zero (hyper-exponential distributions). I am quite confident that this new result will be coherent with those presented in the bottom subplots of the same Figure for the Weibull and the Gamma distribution models. For those models, shape parameters larger than one are associated with hyper-exponential distributions. How to interpret these (few) estimates characterizing hyper-exponential distributions? On the basis of my modest opinion, after a visual inspection of the survival function (see next point for plotting suggestions), in most of the cases negative shape parameter values should be due to statistical variability (i.e. estimation variance) and the data could be reliably described using an exponential distribution model. However, this is just my personal opinion, based on experience from rainfall data originating from a limited geographical region and, hence, I do not expect to be generally accepted. Anyway, using the Generalized Pareto distribution, a negative shape parameter characterizes an upper bounded distribution.

AC3. Indeed, we allowed only positive shape parameter values for the Pareto II because negative shape parameter values correspond to a distribution with an upper bound which is physically inconsistent (see our reply to A. Clauset regarding the upper bound). Indeed, the lower limit we set, for numerical reasons, is 0.001 which is practically zero and essentially corresponds to an exponential distribution; it is trivial to verify this just by plotting an exponential distribution and a Pareto II distribution with shape parameter 0.001. The Generalized Pareto proposed by the reviewer differs from the Pareto II only in one additional location parameter and thus all the other mathematical properties are the same. So, mathematically it would be possible to allow the shape parameter to take on negative values, but as we explained this is physically inconsistent. We do not agree with the reviewer regarding the interpretation of the shape

parameter of the Gamma distribution as we explain in AC9 below. There is no need for some special interpretation of hyperexponential distributions; they are distributions with lighter tails than the exponential tail which, from a practical point of view, “translates” to mild and infrequent extremes.

RC4. *Plotting of survival functions. A very useful diagnostic plot to identify different tail behaviors is that of the logarithm of the survival function versus recorded values. Linear behavior corresponds to an exponential tail, convexity characterizes subexponential (heavy-tailed) distributions, whereas concavity characterizes hyperexponential (bounded) distributions. Since characterization of distribution tails is one of the main scopes of the paper (see e.g. the sentence/definition on page 5760 lines 3-4, which I like a lot: “Here, we use the term “heavy tail” in an intuitive and general way, i.e. to refer to tails approaching zero less rapidly than exponential tails”), I suggest to use this kind of plots.*

AC4. We are aware of these plots; yet, we do not see any useful way to use them. We analysed more than 15000 stations here, so, creating more than 15000 graphs, albeit possible, seems meaningless. These graphs could only be interpreted by visual inspection, as it is not trivial to “quantify” convexity or concavity in empirical data.

RC5. *Threshold selection. This is still an unresolved issue. There are methods to cope with the uncertainty in determining exactly the optimum threshold (see e.g. Deidda 2010), but the authors skipped this issue and decided to consider a number of highest values equal to the number of years of observation. It is my opinion that an optimum threshold would allow inclusion of more values with consequent reduction of estimation variance, but I understand the Authors’ choice since they analyze a large amount of stations. In such a way they are almost sure the distribution of the excesses belongs to the domain of attraction of the Generalized Pareto distribution, but not necessary a Pareto type II distribution.*

AC5. In his paper the reviewer assumes that values above a threshold can be described by the Generalized Pareto and, based on this, he proposes a method for threshold selection. We do not make this assumption here and thus we cannot apply his method. Actually, we do not make any kind of assumption; we just fit four common distribution tails to the record's *N* largest values (the reviewer is also referred to our reply to A. Clauset for this issue). Additionally, we have derived some analytical relationships regarding the exceedence probability of the threshold resulting by the method we follow and we will add them in the revised manuscript (see also our reply to S. Begueria). Finally, we note that the dataset is open access and free and thus the reviewer and anyone interested can apply his method and derive his own results.

RC6. *Page 5763 lines 9-12: "On the contrary, the norm given in Eq. (3) treats each data point equally as it considers the relative error between the theoretical and the empirical values which is independent of the absolute values". This sentence is theoretically incorrect: please remove or reformulate this sentence. Weights, such as that introduced in eq.(3), are sometimes applied to goodness of fit statistics for tails: indeed the CDF is usually S-shaped thus even a very small difference between empirical and fitted CDFs would imply a large error in the quantile. Anyway, the only theoretically consistent approach to treat each data point equally, is by building a norm on quantiles, as suggested by another reviewer.*

AC6. We do not understand why this sentence is "theoretically incorrect". Anyway, we will try to rephrase it as in this part we will add some additional information regarding the norm we used. The reviewer is referred to our reply to F. Laio, where we describe a Monte Carlo scheme we performed, which actually verifies that the norm we use is far better than the classical one (and perhaps advocated by this reviewer) based on quantiles.

RC7. *Last but not least, as already suggested, Authors should make it clear, especially in the abstract, introduction and conclusion which are the objectives of the paper. I*

believe that the paper can contribute to identify the distribution which can best fit a wide range of tail types as those observed in rainfall time series (including exponential). I would remark that we cannot simply say "we need a heavy tailed distribution", since, for instance, if we are performing a regional analysis we cannot apply a Lognormal in a station and a Pareto Distribution in a close station: usually we have to make a choice and use the same distribution over the study area.

AC7. This is a study regarding the tails of rainfall distribution worldwide and this we believe is crystal clear from the beginning of the manuscript. The results indicate that a general rule we can say that in most cases a "heavy" type or subexponential type of tail is preferable. With this we do not mean that a power type distribution like Pareto is always the case (this is apparent in Fig. 4). Pareto, Lognormal, Weibull, etc. are all models, not reality. The problem, brought up by the reviewer, of regional analysis and the related choices about adjacent stations is a very interesting one indeed, but it is out of the scope of the present paper whose domain is the entire globe. We hope to be able to study the regional problem in one of our future works.

RC8. *Page 5764, line 18: "For $\gamma \to \infty$ it degenerates to the exponential tail .". This sentence is incorrect: eq.(4) tends to be an exponential distribution when approaches zero.*

AC8. The reviewer is obviously correct. Strangely, in the submitted manuscript we had it correct ("For $\gamma = 0$ it degenerates to the exponential tail ...") yet, we failed to spot this in the proof corrections.

RC9. *Page 5766, line 5: "for $\gamma > 1$ the distribution is sub-exponential and form and for $\gamma < 1$ hyper exponential." Apart from errors in English usage, this sentence is also incorrect: the Gamma distribution is sub-exponential (heavy tailed) when the shape parameter is smaller than 1. Fixing these two errors in shape parameters interpretation and introducing the GPD will make all the subplots in Figure 6 coherent each other and* *will help the readers to correctly draw their picture.*

AC9. We do not think that the reviewer is right here. The pdf of the Gamma distribution is proportional to $x^{\gamma-1}\exp(-x/\beta)$; thus, for $\gamma > 1$ the $x^{\gamma-1}$ is a monotonically increasing power function which when multiplied with the monotonically decreasing $\exp(-x/\beta)$ will "fatten" the tail, while for $\gamma < 1$, the $x^{\gamma-1}$ is monotonically decreasing and when multiplied with $\exp(-x/\beta)$ it will make the tail decrease faster. So, compared to the corresponding exponential tail, i.e., $\exp(-x/\beta)$, the Gamma tail is "heavier" for $\gamma > 1$ (although bell-shaped), and "lighter" for $0 < \gamma < 1$ (although J-shaped). However, we were not precisely correct in the text as we have written that for $\gamma > 1$ the Gamma tail is subexponential; the truth is that the Gamma distribution does not belong to the subexponential class of distributions (whatever the value of γ is) as it fails to pass particular tests of subexponentiality, e.g., the convolution property (see e.g., Goldie and Klüppelberg, 1998); it belongs to another class called $S(\gamma)$. We will add a paragraph correcting and clarifying this issue further.

RC310. *There are several errors in English usage, but I cannot be of any additional help, since my English is not that good.*

AC10. We assume that the reviewer is right and we do not pride ourselves for perfect English. We will try to spot those errors the reviewer implies.

References

Balkema, A. A. and de Haan, L.: Residual Life Time at Great Age, Ann. Probab., 2(5), 792–804, doi:10.1214/aop/1176996548, 1974.

Goldie, C. M. and Klüppelberg, C.: Subexponential distributions, in A Practical Guide to Heavy Tails: Statistical Techniques and Applications, edited by R. Adler, R. Feldman, and M. s. Taggu, pp. 435–459, Birkhäuser Boston., 1998.

Koutsoyiannis, D.: Statistics of extremes and estimation of extreme rainfall, 1, Theoretical in-

vestigation, Hydrological Sciences Journal, 49(4), 575–590, 2004.

Papalexiou, S. M. and Koutsoyiannis, D.: A global survey on the distribution of annual maxima of daily rainfall: Gumbel or Fréchet?, in European Geosciences Union General Assembly 2012, p. 10563. [online] Available from: http://www.itia.ntua.gr/en/docinfo/1207/ (Accessed 3 October 2012a), 2012.

Papalexiou, S. M. and Koutsoyiannis, D.: Entropy based derivation of probability distributions: A case study to daily rainfall, Advances in Water Resources, 45(0), 51–57, doi:10.1016/j.advwatres.2011.11.007, 2012b.

Pickands III, J.: Statistical inference using extreme order statistics, the Annals of Statistics, 119–131, 1975.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 9, 5757, 2012.