Data from processes with mixed-type marginals cannot be treated using continuous marginals

Simon Michael Papalexiou

Department of Civil and Environmental Engineering, University of California, Irvine, CA, USA (simon@uci.edu)

The paper of (Ye et al., 2018) entitled "***The Probability Distribution of Daily Precipitation at the Point and Catchment Scales in the United States***" deals with an important topic, i.e., the identification of probability distributions to describe the daily rainfall both at station level but also at large catchments. The paper has a nice and clear logical structure, it is easy to read (quite a rare quality) and it is the first study as far as I know that deals with a large number of records at the catchment level. Clearly, there is potential in this study, but unfortunately in my opinion there is a fundament issue that needs to be addressed, i.e., the part that uses the whole record of precipitation values including zeros.

Apples with oranges

It is well-known that many processes in nature, including precipitation, are intermittent processes, and therefore their marginal distribution is of mixed-type, i.e., it has both probability mass (pmf) to express concentration at zero and probability density (pdf) to express the nonzero values. Of course the expressions of the distribution function $F_X(x)$, the pdf $f_X(x)$ and the quantile function $Q_X(u)$ can be related to the conditional expressions for $X|X>0$. Thus, if p_0 is the probability dry, then the cdf, pdf (it is not actually pdf, it is pmf and pdf at the same time: dirac delta notation can be used to unify to pdf) and quantile functions of X are given by

$$F_X(x) = (1-p_0)F_{X|X>0}(x) + p_0 \quad x \geq 0 \tag{1}$$

$$f_X(x) = \begin{cases} p_0 & x = 0 \\ (1-p_0)f_{X|X>0}(x) & x > 0 \end{cases} \tag{2}$$

$$x_u = Q_X(u) = \begin{cases} 0 & 0 \leq u \leq p_0 \\ Q_{X|X>0}\left(\dfrac{u-p_0}{1-p_0}\right) & p_0 < u \leq 1 \end{cases} \tag{3}$$

Now, this affects profoundly the expressions of moments, as the q-th raw moment is given by

$$m(q) = (1 - p_0) \int_0^{\infty} x^q f_{X|X>0}(x) \mathrm{d}x = (1 - p_0) m_{X|X>0}(q) \tag{4}$$

and of course using the well-known formulas that relate the central moments to raw moments we can find easily the expressions of the mean, variance, skewness, kurtosis etc. For example, the mean, variance, and the third and fourth central moments are given by

$$\mu_X = (1 - p_0) \mu_{X|X>0} \tag{5}$$

$$\sigma_X^2 = (1 - p_0) \sigma_{X|X>0}^2 + p_0 (1 - p_0) \mu_{X|X>0}^2 \tag{6}$$

$$\mu(3) = 2m(1)^3 - 3m(1)m(2) + m(3) \tag{7}$$

$$\mu(4) = -3m(1)^4 + 6m(1)^2 m(2) - 4m(1)m(3) + m(4) \tag{8}$$

where of course the raw moments in Eqs (7)-(8) should be replaced using Eq (4).

I show these expressions using product moments as they are analytical to stress how summary statistics are affected by the presence of zeros. For example, if product moment ratio-plots were used to identify an appropriate distribution, using empirical statistics of the whole record would be valid only if compared with the corresponding theoretical curves that express the mixed-type distribution.

The situation with L-moments is the same. Particularly, we can define the L-moments for the mixed-type marginal, if I am not mistaken, as

$$\lambda_1 = \int_{p_0}^{1} Q_{X|X>0} \left(\frac{u - p_0}{1 - p_0} \right) \mathrm{d}u \tag{9}$$

$$\lambda_2 = \int_{p_0}^{1} Q_{X|X>0} \left(\frac{u - p_0}{1 - p_0} \right) (2u - 1) \mathrm{d}u \tag{10}$$

$$\lambda_3 = \int_{p_0}^{1} Q_{X|X>0} \left(\frac{u - p_0}{1 - p_0} \right) (6u^2 - 6u + 1) \mathrm{d}u \tag{11}$$

$$\lambda_4 = \int_{p_0}^{1} Q_{X|X>0} \left(\frac{u - p_0}{1 - p_0} \right) (20u^3 - 30u^2 + 12u - 1) \mathrm{d}u \tag{12}$$

for which analytical expressions can be derived for some specific distributions.

The authors here are presenting in their L-ratios plot a comparison of summary statistics estimated from the whole record (mixed-type data) with the theoretical curves or points of the continuous distributions and not of the mixed-type distributions which can be derived from the equation I previously presented.

It should be apples with apples and oranges with oranges. Thus, if the authors want to use the whole record they have to construct the corresponding curves for the mixed-type cases. So, the fact that the P3 seems a good choice for the whole records it is just an artifact, as well as the nice and neat concertation of points. It is the changes in probability dry that dominate the statistics. And since the domination comes from the probability dry I would guess that if the authors construct the corresponding curves (for fixed p_0; otherwise they form an area) for the mixed-type case they will find that for high p_0 values these curves for different conditional distributions are very similar.

This can be easily also verified by empirical points using simulations. In the Fig. 1 I generated synthetic precipitation having the same correlation structure, the same probability dry, i.e., 90%, but two very different marginals (for a method on how to generate precipitation with any marginal, and any correlation structure and preserving intermittency see Papalexiou (2018)). In Fig.1a is precipitation from a Pareto II with tail exponent 0.2 and in Fig. 1b is from an exponential (light tail). One hundred samples were generated for each case and the L-ratio points were estimated (red and blue dots correspond to Pareto and Exponential cases, respectively). As we see in Fig 1.c the L-ratios for the whole sample (including zeros) are essentially the same for the two distributions forming a linear line (note the narrow range, e.g., in skewness from 0.87 to 0.93 and the huge overlapping). On the other hand, the L-ratios in Fig.1d referring only to the nonzero sample they are quite different (see the large range and insignificant overlapping).

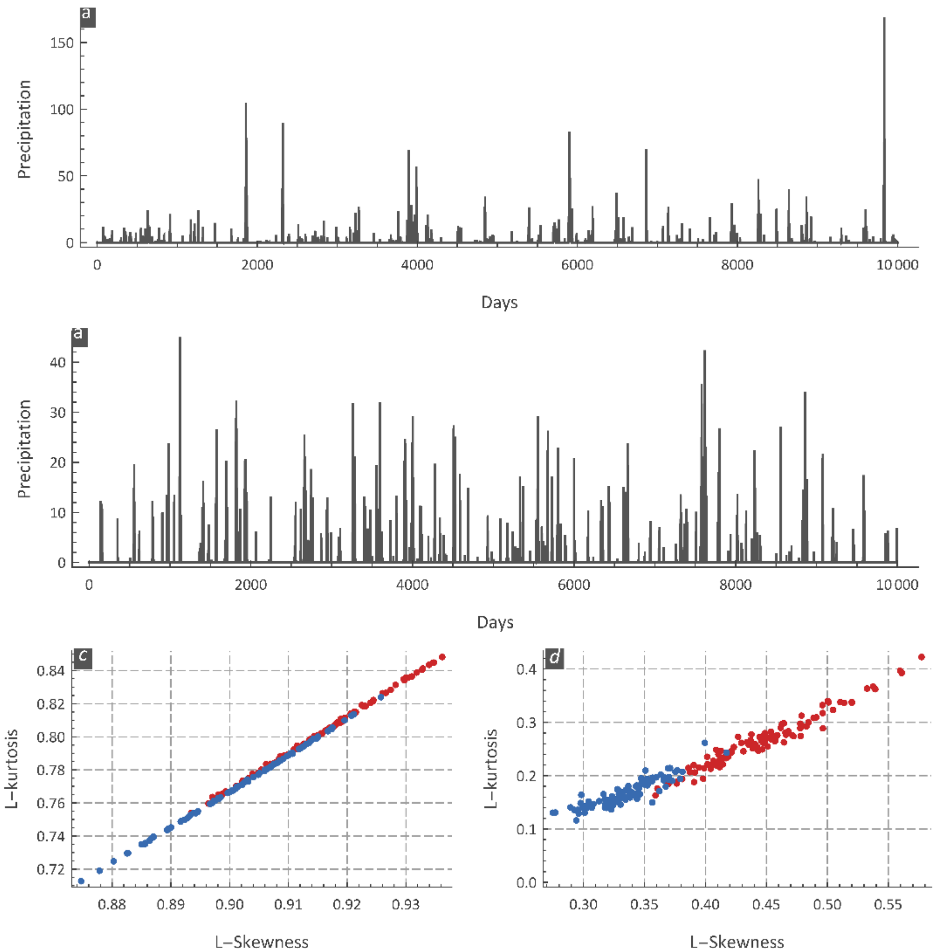

Figure 1: Synthetic precipitation having the same autocorrelation structure and probability dry (90%) but different marginal distributions, that is (a) Pareto II and (b) Exponential. Sample L-ratio points for one hundred generated samples from each case for the whole samples (c) and the nonzero samples (d).

Thus all parts that refer to the whole record as well as the conclusions drawn from the comparisons with the nonzero samples have to be modified in my opinion.

Other issues

1. Lines 363-365: "*demonstrating that the parameter Gamma distribution cannot describe the tail behavior of full-record series of precipitation as has often been assumed in the past.*"
   These lines are just the opportunity for commenting on tail issues. Summary shape statistics are of course affected by the tail behaviour but they are not sufficient to reveal in a robust way the behaviour of the tail if the whole sample is used (I mean all nonzero values) and not values that belong to the tail. For example in the paper the authors cite (Papalexiou and Koutsoyiannis, 2016) after the fitting using L-moments various

measures were proposed in order to compare the fitting in the most extreme value, the largest extremes the whole sample etc. The author can see that the performance of distributions changed, still the GG performed better but the BrXII increased its performance too. I just want to say that indeed this approach can favour specific distributions and exclude others like the G2 the authors mention, yet this is based judging the whole distributional shape properties and it is not really robust to judge on the tail when using the whole nonzero sample. Also other global studies indicated the sub exponential nature of tails focusing on using only "tail" data (Papalexiou et al., 2013; Serinaldi and Kilsby, 2014); the latter was also applied in a seasonal basis, which by the way might be also a nice idea, i.e., the authors to explore seasonal variation.

2. The P3 distribution is just the two-parameter Gamma distribution (G2) with an additional location parameter which does not affect the shape characteristics and thus τ_3 and τ_4. So the curve of P3 shown in $\tau_4 - \tau_3$ ratio plots is the same as the G2. And obviously they have the same tail. The same holds for GPA and GP2 and for any other distribution that has one shape parameter and additional location parameters are added. Maybe to ease the reader, as different formulations can be found in the literature, it would be no harm to add a table of the distributions functions used.
3. The Weibull distribution could also be added in the analysis as a representative of distributions with stretched exponential tails.
4. When we use distributions with a location parameter to describe a positive variable like the nonzero precipitation this parameter might end far from zero or even negative indicating a lower bound. So, this distribution cannot be used in stochastic modelling of precipitation as it will result in inconsistent values. It would be interesting the authors to actually show some box plots of the estimated parameters.
5. The principle of parsimony should always be applied. If the authors, generate samples from a 4-parameter distribution like the kappa and try to estimate a posteriori the parameters, even for the sample sizes used here, they will find a huge variability that makes, in my opinion, the operational use of 4-parameter distributions quite risky. Of course a 4-parameter distribution like the kappa has a great flexibility, yet this does imply that a better fitting to an observed sample is a better choice to extrapolate values for large return periods.
6. The authors, since this is the first large scale study on catchment precipitation, could provide some analysis on the relation of catchment size and distributional shape. As the spatial averaging process will tend to make the distributions more bell-shaped and with thinner tails. This is the explanation of the performance decrease of the heavy-tailed distribution shown in Fig. 7 compared to Fig. 6 (commenting on the Wet-day; full-day results should be modified).

7. Also, some regions of the USA, mainly in Midwest, show quite intense changes (or maybe natural variability) on extremes. The authors could also comment on that or do a quick extra analysis on the daily precipitation.
8. Finally, I believe the literature should be updated with many other works, e.g., there are several papers that are using other distributions for daily precipitation, e.g., one that came to mind is the by Wilson and Toumi (2005).

References

Papalexiou, S.M., 2018. Unified theory for stochastic modelling of hydroclimatic processes: Preserving marginal distributions, correlation structures, and intermittency. Advances in Water Resources. https://doi.org/10.1016/j.advwatres.2018.02.013

Papalexiou, S.M., Koutsoyiannis, D., 2016. A global survey on the seasonal variation of the marginal distribution of daily precipitation. Advances in Water Resources 94, 131–145. https://doi.org/10.1016/j.advwatres.2016.05.005

Papalexiou, S.M., Koutsoyiannis, D., Makropoulos, C., 2013. How extreme is extreme? An assessment of daily rainfall distribution tails. Hydrol. Earth Syst. Sci. 17, 851–862. https://doi.org/10.5194/hess-17-851-2013

Serinaldi, F., Kilsby, C.G., 2014. Rainfall extremes: Toward reconciliation after the battle of distributions. Water Resour. Res. 50, 336–352. https://doi.org/10.1002/2013WR014211

Ye, L., Hanson, L.S., Ding, P., Wang, D., Vogel, R.M., 2018. The Probability Distribution of Daily Precipitation at the Point and Catchment Scales in the United States. Hydrol. Earth Syst. Sci. Discuss. 2018, 1–28. https://doi.org/10.5194/hess-2018-85

Wilson, P.S., Toumi, R., 2005. A fundamental probability distribution for heavy rainfall. Geophys. Res. Lett. 32, L14812. https://doi.org/10.1029/2005GL022465