Dear Anke Hildebrandt and dear Jens Schumacher,

I managed to update the Logbox method so that the outlier cutting threshold is adapted to the shape and size of the data. For the past two years, I have thought that this was not feasible due to a large risk of overfitting at small/moderate samples sizes. However, the results are surprisingly good (I used data from 6307 century-old weather stations to gain confidence in the method) and I have to thank Dr. Schumacher for encouraging me in this attempt. Part I has been entirely reshaped, and part II has been positively impacted by these changes. Table 1 disappeared, all figures have been updated, the supplementary material as well. The manuscript is clearer and better.

I explain the major steps that led to this update in the following:

1) The family of distributions required to parametrize the model. The Pearson family represents light-tailed samples (Beta has been excluded due to a bounded support). For heavy-tailed samples (ignored in the original manuscript), I chose the Generalized Extreme Value family (Weibull, Fréchet & Gumbel) usually used to model the behavior of extrema. I have access to the quantile function Q of each distribution with high accuracy.

2) The framework. In the original manuscript, I picked $\pm 3/4/5\sigma$ cutting thresholds and associated them a sample size of $10^2/10^4/10^6$ to fit a line and find $\alpha = \mathrm{k}\log(n) + 1$ (the intercept was forced to be 1 to simplify the model). As Jens Schumacher suggested, I replaced this discrete framework with $\alpha(n) = \frac{Q\left(1-\frac{f(n)}{2n}\right)-Q(0.75)}{Q(0.75)-Q(0.25)}$, with f a continuous function that gives the number of erroneously flagged outliers: $f(n) = 0.001\sqrt{n}$. For example, for a sample size of 10^6 I am expecting to cut $f(10^6) = 1$ point. For each distribution, I computed $\alpha(n)$ versus n for five samples ($n = 10^2, 10^3, 10^4, 10^5, 10^6$) and found that $\alpha(n) = A\log(n) + B$ is an accurate model for both the Pearson & GEV family ($r^2 = 0.994 \pm 0.005$ & $r^2 = 0.99 \pm 0.01$). Now there are two known parameters (A,B) for each distribution, and you can notice that $f(n)$ is **not** a flat number, otherwise alpha would become arbitrary large.

3) Case of small sample sizes. The original manuscript ignored biases emerging in small samples. This has been fixed by adding $\frac{C}{n}$ to α in a similar manner than Carling et al. (2000): $\alpha(n) = A\log(n) + B + \frac{C}{n}$. The constant value of $C = 36$ had been determined with small random samples ($n = 9$) to limit type I errors to 0.1%.

4) Model to determine A and B on an unknown sample. The centered Moors $m = m_- + m_+ - 1.23$ is a predictor of the kurtosis excess (Moors 1988, Kim & White 2004) using $m_- = (E_3 - E_1)/(E_6 - E_2)$ and $m_+ = (E_7 - E_5)/(E_6 - E_2)$ with $E_i = q(i/8)$ the sample octile. This study introduces a modified version defined as $m_* = \max(m_-, m_+) - 0.6165$ because m_* is more appropriate than m to determine if a sample is light-tailed or heavy-tailed. For example, a Gaussian distribution ($m_- = m_+ \approx 0.6165$; $m = m_* \approx 0$) and a right-skewed distribution with one heavy tail ($m_- = 0.1$ and $m_+ = 1.13$) will share identical m but different m_*. This difference between m and m_* explains why my attempts to construct a model adapted to the shape of the residuals failed in the past. Figure 1e,f shows the relationship A versus m_* and B versus m_* for all distributions, with their best fit, $g_A(m_*) = 0.2294e^{2.9416m_* - 0.0512{m_*}^2 - 0.0684{m_*}^3}$ ($r^2 = 0.999$) and $g_B(m_*) = 1.0585 + 15.6960m_* - 17.3618{m_*}^2 + 28.3511{m_*}^3 - 11.4726{m_*}^4$ ($r^2 = 0.999$).

5) Logbox model. The Logbox model is finally $\alpha(n) = g_A(m_*)\log(n) + g_B(m_*) + \frac{36}{n}$ for $n \geq 9$. For $3 \leq n \leq 8$, the cutting thresholds are computed using the MAD (safer breakdown point, see method).

6) Testing Logbox and comparing it with other methods in the literature. In the original manuscript, the model was tested on the same theoretical distributions used to parametrize it, without connection with real residuals obtained in Earth Science (this was a critic of the reviewer #1). I changed this and downloaded all the data available for the oldest weather stations on Earth (6307 stations with more than 100 years of daily precipitation and temperature). The residuals have been extracted and the suspicious values flagged by the NOAA have been discarded. The observed percentage of flagged outliers is shown in Fig. 2, and I can compare it with the theoretical percentage: $p_{theo} = f(n) \times \frac{100}{n} = \frac{0.1}{\sqrt{n}}\%$. Results went beyond my most optimistic expectation for both small and moderate samples.

7) Changes in part II. There only is one change in the outlier level used in the precipitation. Now that the behavior of heavy-tailed distributions has been explored in Part I, it appears that the outlier level formerly

chosen for the precipitation dataset (30 years of daily data) was too *low*: $y_{outlier} = y_{max} + \frac{1}{2}(y_{max} - \mu)$ with y_{max} and μ respectively the maximum and mean of the dataset. These $y_{outlier}$ values are in fact statistically plausible, and coincidently correspond to the cutting threshold computed by the new Logbox procedure (it has been proved in part I that the cutting threshold was correctly produced by Logbox). In order to choose a less arbitrary outlier level, I applied the following procedure: For each station with daily precipitation over 100 years, I considered that a precipitation event 20% above the century maximum is "impossible" : $y_{outlier} = 1.2 \times (y_{max})_{100\ years}$. Then I randomly selected 30 years within each station i to compute $\lambda_i = \frac{y_{outlier}}{(y_{max})_{30\ years}}$. The mean value for all stations is $\lambda = 1.6 \pm 0.4$, leading to $y_{outlier} = 1.6 \times y_{max}$ that is less arbitrary than $y_{outlier} = y_{max} + \frac{1}{2}(y_{max} - \mu)$. This change only affects 1 false negative for the Logbox procedure, but does not affect the *tsoutlier* function which fails at capturing outliers anyway.

Point-by-point answer

Detection/removal of outliers is critical because – as the author correctly states – the raw data are altered thus impacting any subsequent statistical analysis. A widely accepted description of outliers as "an observation (or subset of observations) which APPEARS to be inconsistent with the remainder of that set of data" (Barnett&Lewis 1994) stresses the notorious difficulty to unambiguously identify a data point as an outlier and there is always the choice between correcting the data OR correcting the model. The problem is most eminent if there are few data – then there is a great chance that data are removed as outliers just because the assumed model is misspecified (and we must fear that this practice is wide-spread). The author considers the problem of outlier detection in the context of time series where the amount of data is typically relatively large. This opens the chance that we have a lot of information about the underlying distribution (the model) and we can exploit that knowledge to identify inconsistent observations. The approach chosen by the author addresses exactly this situation and the proposed modification of the original boxplot-idea seems to be very promising and robust. However, in the present form the implementation of the idea is not consistent. Roughly speaking, the author compares his approach with several older suggestions which are more or less inspired by the Gaussian case and he replaces these by a suggestion (k= 0.6) which is based on an average(median) of reasonable models. This will almost certainly lead to an improvement since the Gaussian case is an extreme case in the Skewness-Kurtosis plane. But in any particular application we don't have an average of distributions – we have one particular distribution and we should use an outlier detection method which is tailored to this particular distribution as good as possible. The author notes in his discussion of the precipitation use case: "… k=0.6 is insufficient here. A value of k ~ 5 is optimum in this case …". As the authors goal is to provide a "generic pre-processing procedure implemented in R" I would strongly advise to choose the value of k in a data-driven way instead of relying on a default value of k=0.6 thus introducing a quasi-standard even if there is the opportunity for the user to change this value. As far as I see, all the required knowledge is there. My idea would be to robustly estimate skewness and kurtosis from the data and then choose a value for k derived from the corresponding Pearson distribution. This would probably lead to a substantial improvement in robustness of the ctbi method and would also justify the publication of the LogBox approach in a more statistically oriented journal. I am not a specialist in robust estimation but Kim&White (2004) could probably provide some ideas on how to estimate the kurtosis.

[…]

• The argument with the 3σ, 4σ and 5σ convention seems relatively arbitrary. Only together with the requirement that the absolute number of erroneously flagged outliers should be restricted this makes sense. This important information is hidden only in the caption to Table 1. Perhaps you could argue the other way around starting from the expected number of erroneously flagged outliers. This would of course change the regression equations (If you go for exactly one erroneously flagged outlier in the Gaussian case the equation should roughly be -0.5 + 0.322 log(n))

I have entirely updated the manuscript to follow these suggestions (see above).

• The fact, that the Pearson family provides a distribution for any theoretically possible combination of skewness and kurtosis could be more explicitly stated. Also the fact that all the mentioned families (Beta,Gamma, …) are Pearson families and only later got their current names could be worth mentioning.

This has been updated in the introduction of Part I.

• I don't see the need to present the results for the generalized Gamma distribution as the bounded support precludes its use a priori.

The Generalized Gamma Distribution has been removed from the supplementary material and discussion. I also removed the Beta distribution from the Pearson family due to the same problems emerging with the bounded support.

• You are looking for methods which are adapted to particular distributions. A general procedure to deal with non-normal data, which is also popular in the time-series context, is the transformation (log, square root, Box-Cox, …) of the data before statistical analysis. It would be worth comparing the transformation approach with the LogBox approach when detecting outliers.

Data transformations such as the Box-Cox method have been used for comparison in part II.

• I personally find ctbi a bit cryptic. How about CTBin which at least makes the main ingredient – the bins – visible.

I unfortunately need to stick to the ctbi name as it has already been created on the CRAN.

Important details

- All the code has been updated in https://github.com/fritte2/ctbi_article
- The 6307 stations can be downloaded and the residuals can be extracted with this code, however this will take ~40 hours of computing. Instead, I can share the data with Jens Schumacher on a google drive link (~8 Gb).
- The new ctbi version has not been updated on the CRAN yet, but all the changes are available on https://github.com/fritte2/ctbi. I will upload the new ctbi version once I receive feedbacks on the manuscript.
- My affiliation has changed (for the last time!) : « Laboratoire des Sciences du Climat et de l'Environnement, LSCE/IPSL, CEA-CNRS-UVSQ, Université Paris-Saclay, Gif-sur-Yvette 91191, France »
- The high quality figures are available in .PDF. The figures shown in the manuscript are only poor quality snapshots (Microsoft Word has problem to incorporate PDF). Will the editorial staff be able to include the original PDF figures in the article? Thank you!

References

Carling, Kenneth. "Resistant outlier rules and the non-Gaussian case." Computational Statistics & Data Analysis 33, no. 3 (2000): 249-258.

Kim, Tae-Hwan, and Halbert White. "On more robust estimation of skewness and kurtosis." *Finance Research Letters* 1, no. 1 (2004): 56-73.

Moors, J. J. A. "A quantile alternative for kurtosis." *Journal of the Royal Statistical Society: Series D (The Statistician)* 37, no. 1 (1988): 25-32.