Dear Anonymous Referee #1,

Thank you for your detailed feedback. I am sorry to learn that, overall, you did not find a particular interest in this work. I probably miscommunicated the importance of pre-processing, the improvement of the boxplot rule and the description of the strength of a cyclicity in a signal. I hope this response to your comments will clarify some points.

First of all, I would like to apologize concerning the package availability (I have had trouble uploading it on Github). It is now publicly available (https://github.com/fritte2/ctbi) as well as the code used in the study (https://github.com/fritte2/hess-2021-609). Due to name conflicts, the name **past** has been replaced with **ctbi**, which stands for "Cyclic-Trend decomposition using Bin Interpolation". It can be installed with the following command: "library(devtools) ; install_github("fritte2/ctbi") ; library(ctbi)".

In the following, the main points of your review (in grey) are highlighted in bold.

i) The structure and aim of the manuscript do not fit the HESS journal

One main concern is that the structure of the manuscript (length, aim, introduction, content) is not adapted to a HESS Technical note.

1) The title says it is a technical note, but the structure does not really fit that. According to instructions, submissions of technical notes should be only a few pages, while this submission is substantially longer. Also, in the abstract it is stated that the submission wants to propose a standardized way to pre-process time series, which is not really something that can be done in a technical note. However, to be a full research article several important parts are missing in the submission (see comments further down)

2) The article lacks references and introduction in the field of HESS. The introduction is quite general on time-series and R packages, but does not discuss what is commonly used in hydrology or earth sciences and to which extend there is a need for additional improvement of outlier detection and gap imputation within these areas. [...]

[…]

Potentially the submission could be resubmitted as a purely technical note describing the R package and discussing the possible inputs to the function (k, …) and with some examples of different choices on the output. Such a submission should more intuitively describe which effect a change on k has, rather than rely on a simulation study that is difficult to relate to in practice. For example, describing how it might work for a normal and a log-normal distribution often met in earth science.

I agree that the manuscript is difficult to put in a specific category. It does not discuss any physical mechanism encountered in Hydrology or even Earth Science (this rejects the manuscript as an Earth Science research paper); however, the package and the framework are dedicated to all researchers dealing with temporal measurements coming from Earth Science (this rejects the manuscript as a purely statistical paper). I decided to submit it to HESS (and not a statistical journal) because the need for this R package originally came from my field (hydrology & ecosystem), where in-situ or remote measurements are often of poor quality. Additionally, there is a clear gap between statisticians and Earth scientists, as these two separate communities do not read the same journals.

This manuscript was originally submitted as a research paper (which explains the length), but the editor suggested to change it as a technical note. I can adapt the introduction to add a

paragraph entirely dedicated to Earth Science protocols on how to deal with outliers or missing values, but this paragraph will be related to what was already described (e.g., boxplot rule, STL, Hampel filter...), or will be completely specific to a field (e.g., extreme value theory used in return event of 100-year floods). In any case, this will increase the length of the final article.

Concerning the potential resubmission as a purely technical note describing the package, the simulation based on the Pearson family is fundamental in the problem of flagging outliers and justifies the creation of the *Logbox* method. The logarithmic law of $\alpha = k\log(n) + 1$ is solely based on the Pearson family. I therefore disagree on cutting it.

The discussion about how the *LogBox* method works for a Normal distribution is already present in the manuscript (L65-66, L75, L95, L171-175 & Table 1).

ii) The framework developed in this manuscript is unnecessary

Another concern is that, overall, the study does not focus on essential problems in Earth Science. Gaps, outliers, univariate timeseries have been handled by each community separately so far, why do we need change things?

3) Generally, in environmental and earth science time series can have a large variety of different structures and the questions to be investigated vary widely. Depending on which statistical analysis is to be made, the filling of gaps or outlier detection can be more or less important. For many approaches, outliers or gaps are not a crucial problems, but can be handled intrinsically. It is, thus, not obvious that a standardized way to preprocess is desirable. Obviously, when several series are within the same academic study they should be handled similarly, but no examples of this being a real problem at present is given. Also, in earth sciences there are few situations where only single time series need to be handled. Either there are several variables observed at the same time point, which can be used to identify if there is something wrong with the sample altogether one variable specifically, or there are nearby stations available that can be used to identify outliers or fill gaps.

Firstly, the R package developed in this study is not entirely dedicated to flagging outliers or imputing missing values. It can be used without flagging outliers (*k.outliers = Inf* in **ctbi**) or imputing data (*SCI.min = Inf*), but simply as a tool to aggregate data in a flexible manner and quantify the strength of a cyclic pattern in a generic way. These two problems are non-trivial, they constitute half of the manuscript (part II), and they are not mentioned in your review.

Secondly, any discipline that produces measurements (in-situ or remote) has to handle gaps and outliers. Even the most standardized products (GHCN, MODIS, NEON,..) contain spikes or suspicious values. Before the creation of this package, I had no tools on how to properly preprocess them before performing the analysis. For example, the *tsoutliers* function (R package *forecast*) fails when data are non-gaussian. I therefore had to pre-process my raw data “manually”, by removing errors “visually”, in a procedure that is case-specific. This method is not only time-consuming; it is completely arbitrary. Two scientists pre-processing the same raw data would end up with two different preprocessed datasets. Additionally, the pre-processing method detailed in published articles is sometimes cryptic to anyone who did not manipulate the raw data.

I hope these points prove how much the non-standardization of preprocessing goes against the scientific rule of reproducibility. I have difficulties to understand how you never encountered

these issues when dealing with measurements, and how this problem can be considered as '*not crucial'*.

As you mentioned it, the procedure developed in the manuscript concerns univariate timeseries for a single location, and not multivariate timeseries recorded at different spatial locations. However, the procedure can be applied to each variable individually (with their own parametrization) to merge them into a standardized aggregated dataset. For example, if someone measures climatic variables (precipitation, temperature, humidity, …) at a different time resolution, **ctbi** will be able to aggregate them at the same resolution. The copyleft license (GPL-3) and the open-source nature of the package leave room for future improvements, for example spatial considerations.

iii) The comparison with existing methods has not been performed

2) [...]The only part that connects to the journal are the case studies, which are relevant, but as the results are not compared with other approaches or articles that have used these series before in a different way conclusions are difficult to make.

[…]

- At least one of the case studies has a seasonal pattern, which would allow a comparison to STL or Stlplus

[…]

- References to the use of outlier detection methods in earth sciences are missing.

The second part of the manuscript did not include any comparison with existing methods because I did not find a popular referenced method (or R package) that can handle the three contaminated cases together as they are so different from a statistical point of view (temperature, precipitation & methane). **Stlplus** does not handle the temperature and methane datasets, and the function **hampel** that applies a Hampel filter to flag outliers does not work on data with missing values. The function **tsoutlier** can be applied to the three datasets but it will cut too many data points (Gaussian assumption). I could compare my results to it, but, again, the article will be significantly longer and the goal of the article is not an inter-comparison of outlier detection methods.

Additionally, I have indicated L240 that the cyclic component of **ctbi** is identical to the cyclic component of **stlplus** for periodic time series. If someone is interested in comparing both decompositions, I can show the long-term trend of **stlplus** and the long-term trend of **ctbi** in a single graphic in the supplementary material.

- It is argumented that Stlplus has severe disadvantages compared to the proposed method. For example, it is said that the trend modelled with loess needs to be parametrized. No reference is given and it is unclear what is meant by this, as loess is a non-parametric regression methods and does not need a parametrization.

stlplus is a non-parametric method: it does not assume a particular distribution of the raw data. However, there are several input parameters that will change the shape of the LOESS. Namely, s.degree; t.window; t.degree; fc.degree; l.window; l.degree in the function **stlplus** of the R package **stlplus**. I can rephrase this sentence to clarify the difference between a non-parametric method and the input parameters required to use it.

iv) The treatment of outliers is unclear

iv.a) Definition of an outlier

- It is not discussed which definition of outlier is used in this context, and especially it would be important to define outliers in highly skewed distributions and how it would be possible to distinguish them from observations that belong to the distribution.

The definition of an outlier is established L58-62 within the framework of the boxplot rule: a value below a lower boundary *l* or above a higher boundary *u* is considered as an outlier. These two boundaries only depend on the value of α, originally chosen as $\alpha = 1.5$ to fit the Gaussian distribution for a small sample size ($n < 100$).

I have adapted this framework to *any* kind of distribution (with an arbitrary skewness or kurtosis) and *any* sample size to define the function α^+that will replace α (see L86-87). The value of α^+ depends on two parameters: 1) Q, the quantile function of the distribution that depends on the nature of the distribution (Gaussian, Exponential, etc.) and 2) *p* that determines the percentage of data captured within [*l*,*u*] (Table 1 columns 2 & 3).

A percentage of data captured is associated with a specific sample size (e.g., 99.73% of data captured is associated with a sample size of 100), see Table 1 columns 3 & 6. This will ensure that less than 1 outlier is flagged on average. This association led to the *Logbox* rule $\alpha^+(n) = k\log(n) + 1$, with k depending on the nature of the distribution (see Fig. 1d).

iv.b) The Pearson family

- It is also not clear how the 9702 distributions are defined and how they are chosen. At one place, a reference to the supplementary is given, but there is no info on distributions in the supplementary.

The Pearson family is a system of distributions (Fig. 1a) that covers most of the kurtosis-skewness space. Each distribution has a one or several shape parameters that will change the value of the skewness and the kurtosis. L115: “their *shape* parameters have been chosen to produce regularly-spaced points in the (κ, S^2) space without overlap and with a mean distance of 0.05 between them (Fig. 1).”. The value of the shape parameters is available on GitHub, these distributions are well-known (gamma, beta, beta-prime, student…) and I thought it was not necessary to detail all their characteristics.

Let’s consider you have detrended & deseasonalized your data to extract the residuals that follow an unknown distribution. You will not be able to calculate an estimate of the kurtosis and the skewness of this distribution (it would require an extremely high sample size and no outliers). For the same reason, the α value cannot be estimated based on this distribution either (although the 4 models in Fig. 2 have attempted - and failed - to do so). The reasoning is the following: if the unknown distribution is *slightly* non-gaussian (moderate skewness & kurtosis), it will be one of the distributions of the Pearson family (one point in the cloud of distributions in Fig. 1a). And the most representative α of this family is the median of the 9702 distributions.

- It is rather unclear how the value of k are determined. Are simulations in Figure 1a-1c made for several sample sizes and their medians are shown in panel d?

Fig. 1b,c does not include any sample size. Based on the definition of α^+ (L 87), you simply need the quantile function of a distribution (which is an analytical formula) and the value of *p* (given in Table 1) to calculate $\alpha^+(p)$. The values of α^+(p) in Fig. 1b,c are therefore *exact*, they have not been numerically determined with simulations. The median is calculated on the population of all α^+(p) of the 9702 distributions of the Pearson family for a given *p*.

How do I find the law of $\alpha = k \log(n) + 1$? This law works the Gaussian (Table 1, column 4 versus column 6, $k = 0.16$), but also the medians of all α values of the Pearson family (Table 1, column 5 versus column 6, $k = 0.6$), as well as for the Exponential (Fig. 1d, $k = 0.8$).

- For comparison between outlier detection method 600 distributions were selected to give the same weight to different types of distributions. For determining the value of alpha and k all 9702 distribution are used. It is unclear why.
- It is rather unclear how well the suggested values of 3.8 and 9.4 work in practice as they are the median of values achieved in the simulation. This means probably that these values work considerably worse for some specific distribution. No discussion is made about this.

Why introducing 600 random distributions, 100 per family, to compare models? Two reasons: (i) If I randomly pick into the 9702 distributions with equal weight, I will have 83% of chance to pick a Beta or Pearson IV (L151), and the comparison will be biased. For example, the other models (Ley., Sch., Hub., Kim.) might perform much better on a student or gamma distribution. (ii) *Logbox* has been parametrized on these 9702 distributions, it would be incorrect to use the exact same pool of distributions to compare models.

The performance of the *Logbox* method on distributions with α values different from 3.8 or 9.4 (For example, orange and pink areas in Fig. 1b and Fig. 1c) is shown in Fig. 2a. These distributions contribute to the variability seen in the percentage of data captured by *Logbox*, which shows an excellent performance to distributions with various skewness & kurtosis (L203-204).

iv.c) Type I error and type II errors are unclear

- A new boxplot rule is suggested and motivated by that using this rule leads to far less false positives, i.e. the type I error is improved. No mention is made on the type II error, which is typically increases, when the type I error decreases. Clearly, this is not easy to study as, in a univariate time series, only outliers above a certain threshold can be detected. [..]. In the recommendations it is stated that the value of k=0.6 will minimize the type I and type II errors, but it is very unclear how this determined and generally it is not possible to minimize type I and type II errors at the same time.

Yes, a sentence on the difference between type I error (real data points are flagged as outliers) and type II error (real outliers are missed) and how they are related to the concept of "percentage of data captured by a model" is missing in the part I of the manuscript. This will be updated. Actually, the framework developed in part I with the theoretical distributions of the Pearson family takes both the type I and type II errors into account. The *Logbox* has been parametrized to flag, on average, less than 1 point as an outlier for a distribution with moderate skewness & kurtosis. Short example: if 1000 data points are generated following an Exponential distribution and the Logbox method cuts 1 point (it captures 99.9% of the data), it means (in this single case) that the Error of type I is 1 and error of type II is 0 (because it will cut any outliers above this threshold). If the percentage of data captured was 100%, there would be an issue of type II error because the threshold might be too high. But more than

~80% of the distributions of the Pearson family are cut below 100% (Fig 2a), which means there rarely is an issue of type II error.

The only problem is that the detrended & deseasonalized residuals might follow a distribution drastically different from those used in the Pearson family. This is for example the case of the daily precipitation (heavy tailed), with a large number of false positives. Users therefore need to make an educated guess on the nature of their data, from Gaussian (k=0.16) through Exponential (k=0.8) to heavy tailed (k > 5). The default value of k=0.6 concerns distributions departing from the Gaussian, with moderate skewness and moderate kurtosis.

- In this study this threshold is chosen to be very high, leading probably to situations where few (real) outliers are detected. This is also one of the reasons outlier detection methods flag rather many observations as outlies, giving the user the possibility to doublecheck the correctness of those and keep the ones that seem reasonable
- In the case studies, outliers are introduced and can be identified with the proposed method, but the outliers are completely unrealistic and could be identified by visual inspection only, no advanced methods are needed.

The threshold of $k = 0.6$ could actually be considered as not high enough. As you can observe in Table 2, there are false positives for the three datasets, including the temperature whose residuals are supposed to be Gaussian. This proves how much the classical boxplot rule ($\alpha = 1.5$) is outdated, and why some authors use $\alpha = 3$ instead (package **tsoutliers**, see https://robjhyndman.com/hyndsight/tsoutliers/).

How can someone determine if an outlier is "correct" once it is flagged? Is it through a visual inspection? More generally, it is difficult to say if a value is "realistic" based on "visual inspection only". In Earth Science, record breaking events (mega-droughts, mega-floods, etc..) are actually not realistic at all, which is exactly why models fail to capture them. A recent example: 202 mm of rain have fallen in 1 hour in Zhengzhou (China) on July 20th 2021. Without contextualization, a "visual inspection" would classify this event as impossible.

The level of outliers in part II has been chosen to illustrate the robustness of the method. If this level was too close to the raw signal, one could argue that these are not outliers, simply extreme events.

- The breakdown points of the outlier detection methods are not given.

I can provide the breakdown points, but they apply to the residuals that are not shown in the study.

- In many cases in environmental and earth science it is common to work with e.g. log-transformed values (or models that use a log-link) to account for single high values in skewed distributions. This would make it unnecessary to develop outlier detection methods for skewed distribution. Instead, conventional methods can be used on the symmetric transformed values. Has typical handling of skewed distributions in earth science been studied? Is there a need to find outliers in skewed distributions?

The log-transformation of data can be used in different contexts, such as producing Gaussian residuals (Cox-Box method) or extracting the different components of a signal in a particular case, from $y_t = S_t \times T_t \times \varepsilon_t$ to $\log(y_t) = \log(S_t) + \log(T_t) + \log(\varepsilon_t)$. A famous example is the AirPassengers data in R. In any case, these methods are case-specific and hazardous to

implement in a generic case. Also, a high *skewness* is not necessary a problem, but more often a high *kurtosis* leads to extreme events that seem impossible.

Comments on structure

- Referencing within the article is not clear. Often Figures are referenced already in the methods description, making the reading difficult. A better separation of method and results would be helpful
- Also sections called context and method are not clearly divided.

Can you please give me the specific lines and figures where these issues appear?

Thank you!