Dear reviewers and dear editor,

Thank you for your feedbacks that were complementary. Major revisions of the manuscript have been made to take into account your suggestions, as well as to respect the structure and size of a "Technical note". It was decided to keep part I within the study to give elements of context to future readers (*LogBox* method), but also to shorten it as HESS is not a purely statistical journal.

In the following, you will find a short summary of the changes, then a point-by-point response to your suggestions. Most of these points already received a long individual response (interactive discussion), so in this note I simply pointed out the changes in the manuscript that correspond to specific issues, and otherwise referred to my previous response (in blue).

Update summary:

In the R package:

1) **past** has been replaced with **ctbi,** due to name conflict. This package is currently under review on CRAN.

2) Some minor updates in the **ctbi** R package have slightly changed the number of false positives/false negatives in Table 2, these do not impact the figures, discussion or conclusion in the study. The updates in the **ctbi** are:

(i) $n_{bin\ min} = ceiling((1 - f_{NA}) \times n_{bin})$ instead of $n_{bin\ min} = floor((1 - f_{NA}) \times n_{bin})$ with $n_{bin\ min}$ the minimum number of points for a bin to be accepted ($n_{bin\ min}$ is now rounded up instead of being rounded down, otherwise it was counter-intuitive).

(ii) The left side of the first bin and the right side of the last bin of a time series were treated as NA values, this has been removed.

3) The output $n_{bin\ min}$ is now available, otherwise the user did not know the minimum number of points for a bin to be accepted.

In the manuscript:

1) The *tsoutliers* pre-processing alternative has now an important role. It has recently been updated on CRAN (https://robjhyndman.com/hyndsight/tsoutliers/) to include a smoothing function for non-seasonal time series and a Cox-Box method to transform the residuals into a Gaussian (mentioned by Reviewer #1). *tsoutliers* is now better described in the introduction, and the false positives/negatives can be compared with **ctbi** in Table 2, as well as in the discussion. However, the long-term trend and cyclic component of *tsoutliers* are not available, which limits the comparison.
2) Part I has been shortened: most of the unnecessary elements of the method have been moved to the supplementary material.
3) The limits of **ctbi** concerning signals of residuals with non-stationary variance has been recognized in a new section: **Limits & recommendations.** These signals can be handled following a protocol applied to the soil respiration dataset MIGLIAVACCA of reviewer #2 in the supplementary material.
4) The importance of expert-knowledge in pre-processing has been recognized, particularly for periods of instrument failure or human errors where most algorithms usually fail.

Step by step response:

Reviewer #1

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)

- The format of the article has been shortened to fit as much as possible the structure of a technical note. The use of α^+ instead of α^-, the choice of the boundaries of the Pearson family, the choice of the subset of 600 random distributions, the Generalized Lambda Distribution (GLD) system and their explanations have been moved to the supplementary material. The wording of the discussion has been simplified as well.

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. 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.

[…]

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

- the *tsoutliers* function is now mentioned in the introduction (L35-41), detailed in the method (L335-338), its performance is available in Table 2, and the discussion includes a comparison with **ctbi** (L373-377). This is the most general pre-processing option present in the field of HESS (*tsoutliers* is part of the package **forecast** which is extensively used for time series analysis in Earth Science).

I have also mentioned common problems related to measurements in Earth Science (L23-28) in response to Reviewer #2, but I did not dive into details as this is not the goal of this study.

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.

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.

- Responses to these concerns have been given in the editorial decision. Additionally, from my previous response:

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 and they constitute half of the manuscript (part II).

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.

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 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 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.

[…]

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.

- To avoid redundancy, I have removed the terms type I and type II errors and have replaced them with false positives and false negatives throughout the study. Two concerns were that the cutting threshold of *LogBox* is too high, and that the outliers used in part II were unrealistic. The comparison with *tsoutliers* shows that it is not the case:

i. *tsoutliers* flags false negatives in all three datasets, proving that the outlier level is not unrealistic.

ii. *LogBox* does not flag any false negatives, showing that the threshold of *LogBox* is not too high.

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 breakdown points of the outlier detection methods are not given.

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.

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?

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 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 definition of an outlier is established 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 $\alpha+$ that will replace α. The value of $\alpha+$ 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). The values of $\alpha+$(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 $\alpha+$(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 for the Gaussian (Table 1, column 4 versus column 6, $k = 0.16$), but also for 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, $k = 0.8$). *LogBox* is simply an extension of these three cases.

Details justifying the choice of the 600 distributions are now given in the supplementary material, as well as the reason why α^+ is computed on the 9702 distributions instead of the subset.

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?

[…]

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

- The *tsoutliers* function gives the user the possibility to apply the Cox-Box method to the residuals, which transforms them into a Gaussian. Table 2 shows that this method does not properly work for heavy-tailed residuals, which makes the problem of flagging outliers non-trivial and also justifies this study. Additionally, the *tsoutliers* function uses the STL procedure.

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.

- This has been clarified L213.

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

- I am still not sure to understand this comment. There are no elements of the methods in the results, however figures are referenced in both sections. Please let me know the specific figures or lines where this problem appears.

Also sections called context and method are not clearly divided.

- Fixed.

Reviewer #2

(a) clarify and discuss the assumptions on the data series.

(a) The basic assumption is that the dataset is an additive signal of a long-term trend, a periodic anomaly (termed "cylce" in the manuscript) that does not change with time, observational noise and outliers. As demonstrated, the method is already useful for such cases. However, of to be even more useful, the authors should think about about extending the method to infer or take into account changes of the anomaly with time. At least they need to give the possibility to the user to supply a mask slicing the time series into chunks where the anomaly can be assumed constant, e.g., stacking winter/spring/summer into different stacks.

In the first of the three examples, the daily cycle of temperature (luckily) does not vary. But what about synoptic cycles? During clear-sky weeks the daily temperature cycle will be larger than on cloudy weeks. For signals influenced by vegetation, the cycles will differ with phenology, etc.

I tried applying the method to several soil respiration time series of the publicly available COSORE dataset. I got it to work technically, but was not able to properly detect outliers and aggregate to annual values. For some series, there was probably too few data within periods (Vern series: 4hourly measurement within a daily period), for others the properties of the signal changed too strongly with season (Migliavacca series).

The authors need to better clarify the assumption and limitations of the method. The method is not as general as claimed in the first version of the manuscript.

- The limits of **ctbi** have been acknowledged in the section "**Limits & recommendations**", L408-429. Following my former comment on this point:
  (i) Residuals showing non-stationary variance (complex seasonality) can reasonably be well handled by **ctbi** if the original timeseries is segmented into bins of similar variability (quantified by the MAD). This has been performed on the MIGLIAVACCA dataset and added to the supplementary material.
  (ii) Concerning cycles that stack on each other (daily cycle + weekly cycle + annual cycle + decadal cycle), the whole idea of **ctbi** was to use the aggregation as a tool to progressively remove high frequencies in the original signal. Successive aggregations will unfold these different periodicities.
  (iii) Another limit is related to your next comment (b), which is the periods of instrument failure or human errors. **Ctbi** is not capable to make a distinction between a physically consistent signal and a random noise.

(b) The application of case-specific outlier-detection and aggregation is discussed as being a thing one wants to avoid. However, usually researchers know their data quite well and know their distributions, stability over time, problematic periods, changes in measurement equipment etc. It needs a more balanced discussion on the value of consistency for meta-analysis and usage of expert-knowledge.

- the importance of expert-knowledge has been acknowledged in the introduction (L24-27), limits & recommendations (L408-410), and the conclusion (L444) in order to balance the discussion.

Outlook: many observational time series come in replicates. Can you think of ways to extend the method to use information across the replicates?

- It is difficult to design a generic script that can achieve that because most data are not standardized (different timestamps for the replica for example). I think it is much safer to let the user code the program to handle replicas by using **ctbi** as a tool. My advice would be the following:

- For each time step, calculate the median of all replicas.

- Run ctbi on all the medians.

Thank you!

François Ritter