Technical note: A procedure to clean, decompose and aggregate time series

François Ritter[1]

~~[1]Fonds de Dotation O, 9 rue Duperré, 75009 Paris, France.~~

[1]Laboratoire des Sciences du Climat et de l'Environnement, LSCE/IPSL, CEA-CNRS-UVSQ, Université Paris-Saclay, Gif-sur-Yvette 91191, France

Correspondence to: François Ritter (ritter.francois@gmail.com)

Abstract. Errors, gaps and outliers complicate and sometimes invalidate the analysis of time series. While most fields have developed their own strategy to clean the raw data, no generic procedure has been promoted to standardize the pre-processing. This lack of harmonization makes the inter-comparison of studies difficult, and leads to screening methods that ~~are usually ambiguous~~can be arbitrary or case-specific. This study provides a generic pre-processing procedure implemented in R (**ctbi**, for cyclic/trend decomposition using bin interpolation) dedicated to univariate time series. **Ctbi** is based on data binning and decomposes the time series into a long-term trend and a cyclic component (quantified by a new metric, the Stacked Cycles Index) to finally aggregate the data. Outliers are flagged with an enhanced ~~Boxplot rule called *LogBox*.~~boxplot rule called Logbox that corrects biases due to the sample size and that is adapted to non-Gaussian residuals. Three different Earth Science datasets (contaminated with gaps and outliers) are successfully cleaned and aggregated with **ctbi**. This illustrates the robustness of this procedure that can be valuable to any discipline.

1 Introduction

In any discipline, raw data need to be evaluated during a pre-processing procedure before performing the analysis. Errors are removed, values that deviate from the rest of the population are flagged (outliers, see Aguinis et al., 2013), in some cases gaps are filled. Because the raw data are altered, pre-processing is a delicate and time-consuming task that can be neglected due to cognitive biases deflecting our understanding of reality ('I see what I want to see'), or due to our impatience to obtain results. The fate of extreme values is crucial as they usually challenge scientific or economic theories (Reiss et al., 1997).

Time series are particularly difficult to pre-process (Chandola et al., 2009). A value can or cannot be considered as an outlier just depending on its timestamp (e.g., a freezing temperature in summer), large data gaps are common, abrupt changes can occur and a background noise covers the true signal. In Earth Science, in-situ or remote measurements routinely produce time

series that first need to be visually inspected. The expert-knowledge of the researcher, technician or engineer is essential to flag suspicious periods of possible instrument failure (e.g., a rain gauge blocked by snowflakes), violation of the experimental conditions (e.g., a passing car during CO_2 measurements in a forest), or human error (e.g., calibration of the wrong sensor). Once these suspicious periods have been flagged, a pre-processing algorithm is necessary to evaluate the quality of the remaining portion of the measurements. However, there currently is no consensus on which procedure to use even in the simple univariate case: a recent review (Ranjan et al., 2021) covered more than 37 preprocessing methods for univariate time series, and Aguinis et al. (2013) listed 14 different outlier definitions that are mutually exclusive. Despite this (overwhelming) abundance of methods and conventions, there are surprisingly few R packages that offer a pre-processing function. It is worth mentioning *hampel* (package **pracma**, Borchers, 2021) that applies a Hampel filter (Pearson, 2002) to time series and flags outliers based on the Mean Absolute Deviation (MAD)**,** which is a robust approximation of the standard deviation defined as $MAD(x) = 1.4826 \times \mathrm{M}(|x - \mathrm{M}(x)|)$ with M the median operator. However, the *hampel* function is not robust to missing values and the scaling factor of 1.4826 is not adapted to non-Gaussian residuals. Another option is the function *tsoutliers* (package **forecast**, Hyndman et al., 2020) that applies a Seasonal and Trend decomposition using Loess (STL, Cleveland et al., 1990) to data showing a seasonal pattern, complemented by a smoothing function to estimate the trend of non-seasonal time series (Friedman's super smoother, Friedman, 1984). The residuals obtained can be transformed to follow a Gaussian distribution (Cox-Box method, Box & Cox, 1964), and then outliers are flagged using the ~~Boxplot~~boxplot rule (Tukey, 1977). This method will be proved in this study to work well with data associated with nearly-Gaussian residuals, but to show poor performance otherwise.

This study offers an alternative pre-processing procedure (implemented in R) called **ctbi** for cyclic/trend decomposition using bin interpolation. The time series is divided into a sequence of non-overlapping time intervals of equal period (called bins), and outliers are flagged with an enhanced version of the ~~Boxplot~~boxplot rule (called ~~*LogBox*~~Logbox) that is adapted to non-Gaussian data ~~and insensitive to the~~for different sample ~~size~~sizes. **Ctbi** fulfils four purposes:

i) Data cleaning: bins with insufficient data are discarded, and outliers are flagged in the remaining bins. If there is a cyclic pattern within each bin, missing values can be imputed as well.

ii) Decomposition: the time series is decomposed into a long-term trend and a cyclic component.

iii) Cyclicity analysis: the mean cycle of the stacked bins is calculated, and the strength of the cyclicity is quantified by a novel index, the Stacked Cycles Index.

iv) Aggregation: data are averaged (or summed) within each bin.

This procedure is particularly adapted to univariate time series that are *messy*, with outliers, data gaps or irregular timesteps. The inputs offer a large flexibility in terms of imputation level or outlier cutoff, but also in the timestamp of the bins: a day

does not necessarily start at midnight or a year the 1st of January. The timeline is not limited to daily or monthly data but can vary from milliseconds to millenaries. The outputs keep track of the changes brought to the data: contaminated bins are flagged, as well as outliers and imputed data points.

This paper is divided into two distinct parts. The first part describes the ~~*LogBox*~~Logbox method and compares its performance with four other outlier detection methods in the literature~~, including the MAD~~ based on daily precipitation & temperature data extracted from 6307 century-old weather stations. The second part describes the **ctbi** procedure, and then applies it to three datasets that have been contaminated beforehand to show the efficiency of the algorithm. A comparison with *tsoutliers* is performed, and, finally, limitations and good practice recommendations are discussed.

2 Part I, outliers

2.1 Context

This first part is dedicated to the detection of outliers present in univariate datasets (without the time component). The boxplot (or Tukey's) rule is a commonly used method to flag outliers below a lower boundary l and above an upper boundary u (Tukey, 1977):

$$\begin{cases} l = q(0.25) - \alpha \times \big(q(0.75) - q(0.25)\big) \\ u = q(0.75) + \alpha \times \big(q(0.75) - q(0.25)\big) \end{cases}$$

With q the sample quantile (e.g., $q(0.5)$ is the median) and $\alpha = 1.5$ a constant that corresponds to 99.3% of Gaussian data falling within $[l, u]$. This method is simple and robust to the presence of a maximum of 25% of outliers in the dataset (known as the breakdown point). ~~However, two issues emerge from this rule~~When a real data point falls outside the $[l, u]$ range, it is considered as an erroneously flagged outlier (or type I error). Conversely, a type II error occurs when a real outlier is not flagged. The type I error is more common for three reasons:

(i) For ~~a~~small Gaussian ~~population,~~ samples ($n < 30$), up to 8.6% of data (Hoaglin et al., 1986) can be cut due to the inaccuracy of the sample quantile for small n.

~~(i)~~(ii) For large Gaussian samples ($n > 10^3$), $\alpha = 1.5$ is inappropriate ~~for large sample sizes ($n \geq 10^3$),~~ because the number of ~~points~~ erroneously flagged ~~as~~ outliers increases linearly with n ~~(~~due to the 99.3% of data captured by $[l, u]$~~).~~.

~~(ii)~~(iii) For ~~a~~ non-Gaussian ~~population~~populations, $\alpha = 1.5$ is generally too restrictive. For example, ~4.8 % of data following an Exponential distribution would be ~~erroneously flagged as outliers.~~cut.

~~Two~~While studies have ~~attempted to address~~corrected biases in the ~~second~~ detection of outliers in small samples (see Carling, 2000; Schwertman et al., 2004), they have ignored those emerging in large samples. Concerning the last issue ~~(~~, Kimber~~,~~ (1990~~;~~) and Hubert & Vandervieren, (2008) ~~by adjusting~~have adjusted α to the skewness ~~*S* (third standardized moment (~~related

to the asymmetry of a distribution) ~~while ignoring the~~but did not consider the kurtosis (related to the tail weight) that will be proven to be a key variable in this study. Therefore, there currently is no generic procedure that can be used when the population is non-Gaussian and/or the sample size is large.

To understand how to address this problem, two sets of common distributions with known skewness S, kurtosis excess ~~kurtosis~~ κ_{ex} ~~(fourth standardized moment related to the tail weight). Other studies have corrected biases emerging at small sample sizes (Carling, 2000; Schwertman et al., 2004); however, none have designed a method based on the boxplot rule that can handle outliers present in large sample sizes.~~

~~A more generic method (called *LogBox*) has been developed in this study to assign~~and quantile function Q are used (Fig. 1). The first set is the Pearson family composed of light-tailed distributions that represent any theoretically possible residuals with moderate S & κ_{ex}. Pearson originally worked to create distributions that cover the entire (S,κ_{ex}) space (Pearson, 1895; 1901 & 1916), but they took their modern names later on (Gamma, Inverse-gamma, Beta prime, Student, Pearson IV). The second set is the Generalized Extreme Value family composed of the Gumbel, Weibull and Fréchet that are heavy-tailed distributions (high S & κ_{ex}) used in Extreme Value Theory to model the behavior of extrema (Jenkinson, 1955). Based on this framework, this study finds that $\alpha = k(n) = A\log(n) + B + \frac{C}{n}$ reasonably addresses all previously mentioned issues, with C fixed as a ~~positive number that corresponds~~constant ($C = 36$). The two parameters A and B correspond to the nature of the distribution ~~(e.g., $k = 0.16$ for Gaussian data; $k = 0.8$ for Exponential data). A default value of $k = 0.6$ has been determined with an ensemble of non-Gaussian distributions (the Pearson family) that represent univariate datasets with moderate S and κ_{ex} (Fig. 1).~~and are estimated based on a predictor of the kurtosis modified from Moors (1988). A comparison ~~with~~ between this procedure (called Logbox) and four other existing models is performed ~~to test the resistance of each method to different types of distributions and different sample sizes~~on residuals obtained from 6307 weather stations with more than 100 years of daily temperature and precipitation measurements (Fig. 2). Finally, ~~*LogBox*~~Logbox is implemented ~~(with the value of k left to the user) in the aggregation procedure described~~ in part II to clean the residuals obtained after fitting the univariate time series with a robust and nonparametric method.

2.2 Method

2.2.1 ~~Definition of α^- and α^+~~Distributions

Residuals with ~~the 3σ, 4σ and 5σ convention~~

~~Let D_X be a probability distribution of a single random variable X associated with the population quantile function Q. Two strictly positive functions α^- and α^+ attached to D_X are defined for $0.75 < p < 1$:~~

$$\alpha^{-}(p) = \frac{Q(0.25) - Q(1-p)}{Q(0.75) - Q(0.25)}$$

$$\alpha^{+}(p) = \frac{Q(p) - Q(0.75)}{Q(0.75) - Q(0.25)}$$

The boxplot rule can now be expressed with α^{-} and α^{+}:

$$\begin{cases} l = q(0.25) - \alpha^{-}(p_{lim}) \times \left(q(0.75) - q(0.25)\right) \\ u = q(0.75) + \alpha^{+}(p_{lim}) \times \left(q(0.75) - q(0.25)\right) \end{cases}$$

With q the sample quantile and p_{lim} related to the percentage of data falling within $[Q(1-p_{lim}), Q(p_{lim})]$, independent from the nature of the distribution. In order to set a framework consistent with the Gaussian case, we derive three p_{lim} values ($p_{3\sigma}, p_{4\sigma}$ and $p_{5\sigma}$) expressed as $p_{j\sigma} = \Phi(j)$ with Φ the cumulative distribution function of the standard Normal distribution $\mathcal{N}(0,1)$ and $j = \{3,4,5\}$ implicit throughout the study. These p_{lim} values are associated with the percentage of Gaussian data captured by $\pm j\sigma$ (known as the "sigma Rule"), with σ the standard deviation of the Gaussian. The corresponding $\alpha_{j\sigma}^{\mathcal{N}} = \alpha^{+}(p_{j\sigma})$ values are computed in the Gaussian case with $Q = \Phi^{-1}$ (Table 1), and $\alpha^{-}(p_{j\sigma})$ is discarded in this study because only symmetrical or right-skewed distributions are considered (supplementary material).

j sigma rule	$p_{j\sigma} = \Phi(j)$	$2p_{j\sigma} - 1$ (% of data captured)	$\alpha^{+}(p_{j\sigma})$ (Gaussian)	$M(\alpha^{+}(p_{j\sigma}))$ (Pearson family)	Suggested Sample size
"$\pm 3\sigma$"	0.99865	99.73%	$\alpha_{3\sigma}^{\mathcal{N}} = 1.7$	$\alpha_{3\sigma}^{\mathcal{P}} = 3.8$	$n_{3\sigma} \sim 10^2$
"$\pm 4\sigma$"	0.9999683	99.994%	$\alpha_{4\sigma}^{\mathcal{N}} = 2.5$	$\alpha_{4\sigma}^{\mathcal{P}} = 6.7$	$n_{4\sigma} \sim 10^4$
"$\pm 5\sigma$"	0.9999997	99.99994%	$\alpha_{5\sigma}^{\mathcal{N}} = 3.2$	$\alpha_{5\sigma}^{\mathcal{P}} = 9.4$	$n_{5\sigma} \sim 10^6$

Table 1. Values of $\alpha^{+}(p_{j\sigma})$ for $j = \{3,4,5\}$ associated with the Gaussian distribution (4th column, $\alpha_{j\sigma}^{\mathcal{N}}$) and the distributions from the Pearson Family (5th column, median $\alpha_{j\sigma}^{\mathcal{P}}$). Φ is the cumulative distribution of the Standard Normal distribution $\mathcal{N}(0,1)$ and M is the median. The sample sizes $n_{j\sigma}$ correspond to less than 1 erroneously flagged outlier (based on the percentage of data captured).

2.2.2 The Pearson family

Univariate datasetsmoderate κ_{ex} & S are represented in this study with 97024999 light-tailed distributions from the Pearson family (Pearson, 1895; 1901 & 1916) composed of the Gaussian, Gamma (196 distributions, including the Exponential), Inverse gamma (170), Beta (4703), Beta prime (1135), Pearson IV (3377) and Student (120) distributions (Fig. 1a). These

distributions cover the entire (κ_{ex}, S^2) space without overlap. ~~Their quantile functions are already implemented in *R* to compute with high precision $\alpha^+(p_{fg})$, and their *shape* parameters~~ , except for the Beta distribution that has been discarded due to a bounded support (unrealistic residuals). The shape parameters of each distribution have been chosen to produce regularly-spaced points ~~in the (κ_{ex}, S^2) space without overlap~~ and with a mean distance of 0.05 in the (κ_{ex}, S^2) space and with a range between ~~them (Fig.~~ the Gaussian and the Exponential: $\kappa_{ex} \in [0,6]$ and $S \in [0,2]$. Heavy-tailed residuals are represented with 368 distributions from the Generalized Extreme Value (GEV) family (Fig. 1d) composed of the Gumbel, Weibull (244~~1a).~~ ~~The range of excess kurtosis and squared skewness for all~~ distributions ~~has been picked as~~) and Fréchet (123). Their shape parameters cover a larger range: $\kappa_{ex} \in [0,6500]$ and ~~$S^2 \in [S \in [0, \frac{4}{5}(\kappa_{ex}+2)]$ (supplementary material).~~ , 15].

Mis

Mis

2.2.~~3 Models~~

2 The ~~general procedure developed in this study (*LogBox*) is based on the~~ Logbox model

Based on the boxplot rule ~~and replaces the original constant $\alpha = 1.5$ with~~ , α can be defined as:

$$\alpha = k \log(n) + 1, \text{ with } (n) = \frac{Q\left(1-\frac{f(n)}{2n}\right) - Q(0.75)}{Q(0.75) - Q(0.25)}$$

With n the sample size ~~and $k = 0.6$ the default value. This relationship is established based on the median values $\alpha_{fg}^{\mathcal{P}} = \{3.8, 6.7, 9.4\}$ from the Pearson family, and the three sample sizes $n_{fg} = \{10^2, 10^4, 10^6\}$.~~ Q the population quantile function and f a function that ~~correspond on average to less than one~~ gives the number of erroneously flagged outliers. In the original boxplot rule, $Q = \Phi^{-1}$ (with Φ the cumulative distribution function of the Gaussian) and $f(n) = 0.007 \times n$ which leads to $\alpha = 1.5$. As explained in the introduction, this choice of f is not valid for large sample sizes due to the linear dependence on n. A flat number of erroneously flagged outliers ($f(n) = b$) or a logarithmic relationship ($f(n) = b \log(n)$) would not be appropriate either, because $\alpha(n)$ could take arbitrary large values as $1 - \frac{f(n)}{2n}$ would approach 1 too rapidly ($Q(1) = \infty$). This study suggests instead $f(n) = 0.001\sqrt{n}$ as a compromise. For example, for a sample of size $n = 10^2$, 10^4 or 10^6; respectively 0.01, 0.1 or 1 point would be erroneously flagged as outlier (instead of 0.7, 70 or 7000 points with the original boxplot rule). To characterize the relationship $\alpha(n)$ versus n, $\alpha(n)$ is derived with high accuracy (Q implemented in *R*) for each distribution of the Pearson and GEV family for 5 sample sizes ($n_i = 10^i$ with $i \in [2,6]$). It appears that $\alpha(n) = A \log(n) + B$ is an accurate model for both the Pearson family (mean of $r^2 = 0.994 \pm 0.005$) and the GEV family ($r^2 = 0.99 \pm 0.01$). To account for biases emerging at small sample size, an additional constant term is added following Carling (2000): $\alpha(n) = A \log(n) + B + \frac{C}{n}$. The parameter $C = 36$ has been numerically determined with a Monte-Carlo simulation on the distributions of the Pearson family to ensure that the percentage of erroneously flagged outliers corresponds to ~0.1% for

Mis

Mis

Mis

$n = 9$ (supplementary material). For the particular case of the Gaussian (ubiquitous in nature), $A = 0.08$ and $B = 2$ ($r^2 = 0.999$).

2.2.3 The Logbox procedure

Let (m_-, m_+) be two positive functions defined as $m_- = (E_3 - E_1)/(E_6 - E_2)$ and $m_+ = (E_7 - E_5)/(E_6 - E_2)$ with $E_i = q(i/8)$ the sample octile. The centered Moors $m = m_- + m_+ - 1.23$ is a robust predictor of the kurtosis excess with a breakdown point of 12.5% (Moors 1988, Kim & White 2004). This study introduces a modified version defined as $m_* = \max(m_-, m_+) - 0.6165$. The parameter 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_*. The Logbox procedure is the following for an unknown sample of size n:

(i) If $n \in [3,8]$, the outlier threshold $[l, u]$ is $l = q(0.50) - \beta \times MAD$ and $u = q(0.50) + \beta \times MAD$ with $\beta = 12.5$.

(ii) If $n \geq 9$, m_* is computed (bounded by $[0,2]$) and the boxplot rule is used with $\alpha(n) = g_A(m_*)\log(n) + g_B(m_*) + \frac{36}{n}$.

For very small sample sizes (case i), the median and MAD are preferred (Leys et al., 2013) over the boxplot rule because a single outlier would break the estimator m_*. The value of $\beta = 12.5$ has been numerically determined with a Monte-Carlo simulation and corresponds to ~0.1% of erroneously flagged outliers in the Pearson family for $n = 9$ (supplementary material). For larger sample sizes (case ii), $g_A(x) = 0.2294e^{2.9416x - 0.0512x^2 - 0.0684x^3}$ ($r^2 = 0.999$) and $g_B(x) = 1.0585 + 15.6960x - 17.3618x^2 + 28.3511x^3 - 11.4726x^4$ ($r^2 = 0.999$). Each function has been parametrized based on the Pearson and GEV family together (Fréchet has been excluded due to a different behavior, see ~~Table 1 &~~ Fig. ~~1d). *LogBox*~~1e,f). The coefficients have been determined with a Monte-Carlo simulation that minimizes the root-mean square error ($N \sim 10^8$).

2.2.4 Former models

Logbox is compared with four other models (Kimber, 1990; Hubert & Vandervieren, 2008; Schwertman et al., 2004; Leys et al., 2013). The first two models (Kim. and Hub.) adjust the boxplot method with respect to the skewness:

$$\begin{cases} l_{Kim.} = q(0.25) - 3 \times \big(q(0.50) - q(0.25)\big) \\ u_{Kim.} = q(0.75) + 3 \times \big(q(0.75) - q(0.50)\big) \end{cases}$$

And

$$\begin{cases} l_{Hub.} = q(0.25) - 1.5 \times f(-MC) \times \big(q(0.75) - q(0.25)\big) \\ u_{Hub.} = q(0.75) + 1.5 \times f(MC) \times \big(q(0.75) - q(0.25)\big) \end{cases}$$

$$\begin{cases} l_{Hub.} = q(0.25) - 1.5 \times h(-MC) \times \big(q(0.75) - q(0.25)\big) \\ u_{Hub.} = q(0.75) + 1.5 \times h(MC) \times \big(q(0.75) - q(0.25)\big) \end{cases}$$

With the function fh defined as $fh(MC) = e^{4MC}$ for $MC < 0$ and $fh(MC) = e^{3MC}$ for $MC \geq 0$. The Medcouple MCMC $\in [-1,1]$ is a robust estimator of SS, with an algorithm complexity of $O(n\ log\ n)$ and a breakdown point of 25% (Brys et al., 2004). The third model (Sch.) constructs the lower and upper boundary around the median:

$$\begin{cases} l_{Sch.} = q(0.50) - \dfrac{Z}{k_n} \times 2\big(q(0.50) - q(0.25)\big) \\ u_{Sch.} = q(0.50) + \dfrac{Z}{k_n} \times 2\big(q(0.75) - q(0.50)\big) \end{cases}$$

With k_nk$_n$ a function of the sample size nn to adjust for small samples (given as a table in Schwertman et al., 2004) and ZZ a constant related to the percentage of data captured by $[l_{Sch.}, u_{Sch.}]$, here picked as $Z = 3$ (Gaussian case for the $\pm 3\sigma$ window). Finally, the last model (Ley.) uses the MAD around the median:

$$\begin{cases} l_{Ley.} = q(0.50) - 3 \times MAD \\ u_{Ley.} = q(0.50) + 3 \times MAD \end{cases}$$

2.2.45 Comparison between models

The comparison between models is performed on two sets of residuals obtained from weather stations part of the Global Historical Climatology Network (GHCN-daily) with at least 100 years of daily temperature (2693 stations, 9.4×10^7 days) or daily precipitation (6277 stations, 5.8×10^7 wet days, dry days are excluded). Because this network is used to calibrate products that are remote-sensing based and because suspicious values are routinely flagged (Menne et al., 2012; Xungang et al. 2012), the risk of errors in these century-old stations can be considered small. The residuals are extracted with the robust method described in part II based on non-overlapping bins (bins with less than 80% of data are discarded). To reduce the impact of the extraction method on the analysis, three bin intervals (5, 10 and 20 days) are used to obtain three replicas for each station. The sensitivity of each outlier detection method to the sample size has also been estimated. For each station and for each sample size $l_i = 10i$ (i varying from 1 to 10), $N_i = \frac{100}{i}$ samples are randomly selected and the number of flagged outliers is summed over all the N_i samples (the total number of points is constant, $N_i \times l_i = 1000$).

For the four models (Ley., Hub., Kim., Sch.), the percentage of flagged outliers is computed for each station, and then the mean (± 1 SD) is calculated over all stations. For the Logbox model, this method is not appropriate because the expected number of erroneously flagged outlier per station is less than one ($0.001\sqrt{\sim 10^4} \sim 0.1$). Instead, the percentage of flagged

outliers is calculated over the total number of points: $\rho = (\Sigma n_j^{flagged}) \times 100/(\Sigma n_j)$ with j a station. The variability is estimated by subsampling the total number of stations N_s: $\sqrt{N_s}$ sets of $\sqrt{N_s}$ random stations are selected without replacement. The parameter ρ is computed for each set, and the associated variability is calculated on all ρ values ($\pm 1\ SD$ in Fig. 2f and quantiles in Fig. 2c).

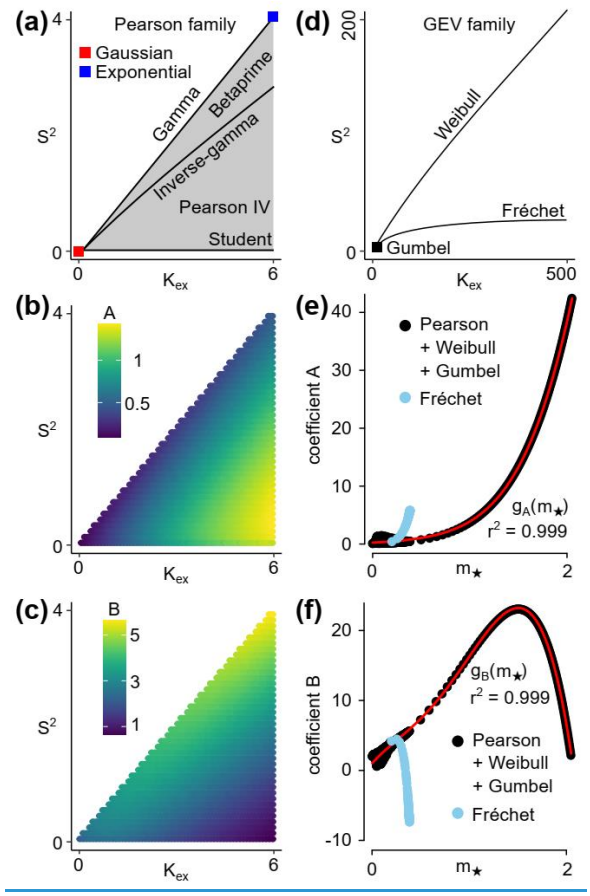


Fig. ~~The comparison between models is performed on a subset of the Pearson Family (600 random distributions, with 100 random distributions per type, see supplementary material). For a given model, a given distribution and a given sample size n, the following procedure is performed to calculate the percentage of data captured by the model (with $m = 0$ initially):~~

~~Step 1: generate random deviates (of sample size n)~~

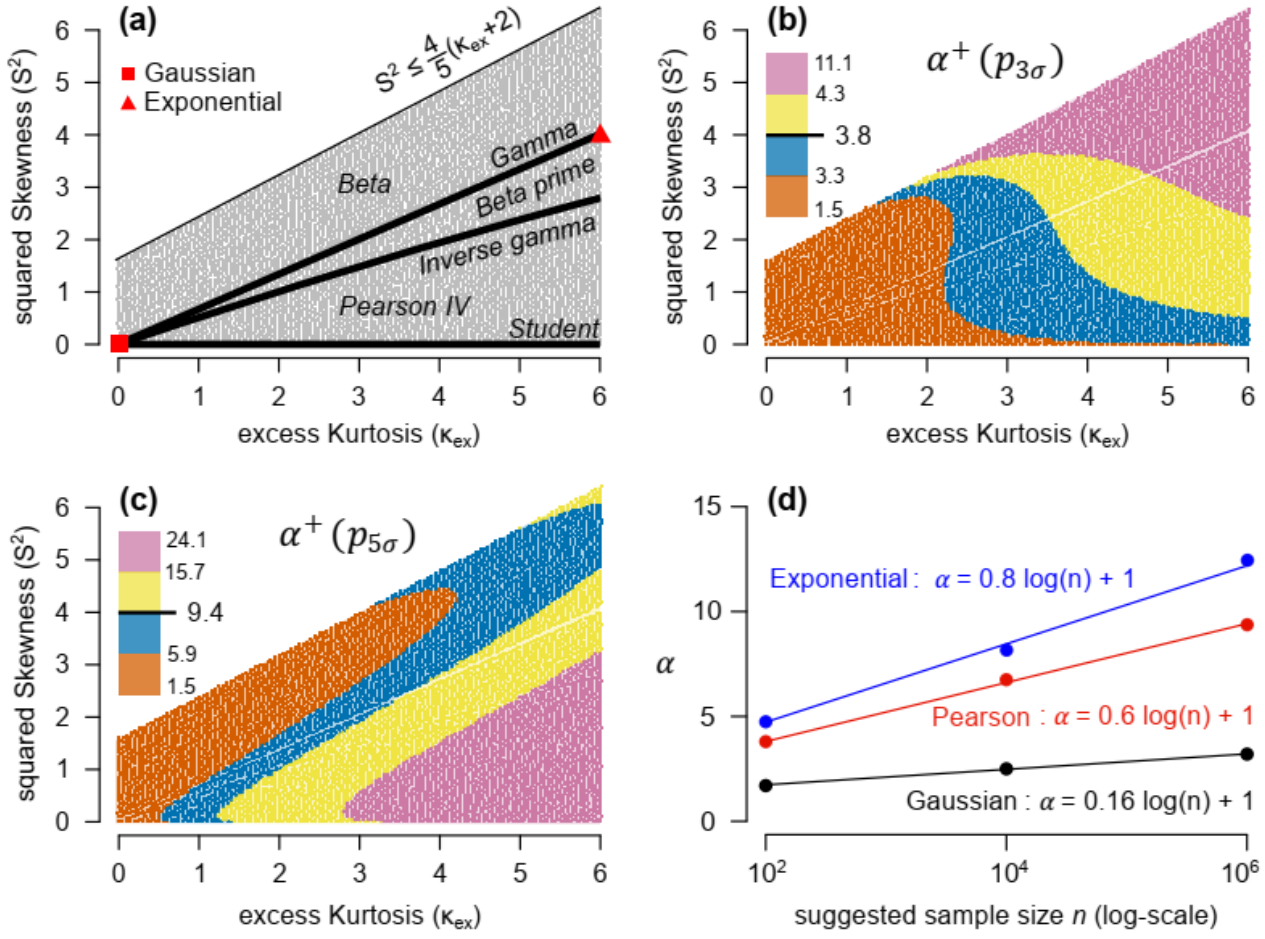
~~Step 2: calculate l and u~~

~~Step 3: Let m_{within} be the number of points falling within $[l, u]$~~

~~$m \leftarrow m + m_{within}$~~

~~Step 4: Repeat N times Step 1 to Step 3, with $N = ceiling\left(\frac{10^6}{n}\right)$~~

Finally, the percentage of data captured by the model for the given distribution and given sample size is $\frac{m}{(N+1)n}$. This percentage is computed for all of the distributions, and the median value of the population of 600 percentages is defined as M_i, associated with a sample size $n_i = 2^i$ varying from $n_4 = 16$ to $n_{14} = 16384$ (Fig. 2).



~~Fig~~ 1. Location of the ~~9702~~4999 light-tailed distributions of the Pearson family (panel **a**) and the 368 heavy-tailed distributions of the GEV family (panel **d**) in the ~~(κ_{ex}, S^2)~~(κ_{ex}, S^2) space (~~panel **a**). Values of $\alpha^+(p_{j\sigma})$~~kurtosis excess, squared skewness). The coefficients A and B correspond to $\alpha(n) = A \times \log(n) + B + \frac{36}{n}$ used to replace $\alpha = 1.5$ in the boxplot rule. For the Pearson family, they are shown ~~for $j = 3$ (99.73% of data captured, panel~~in the (κ_{ex}, S^2) space (panels **b**~~)~~ and ~~$j = 5$ (99.99994%, panel~~ **c**). ~~The five numbers indicated in the legend~~For the GEV family (panels **e** and **f**), they are ~~the minimum, the three quartiles and the maximum. Relationships (R^2 =~~

~~$\mathbf{0.99}$) between $\boldsymbol{\alpha^{+}(p_{J\sigma})}$ and~~ shown against a predictor of $\boldsymbol{\kappa_{ex}}$ defined as $\boldsymbol{m_* = (E_7 - E_5)/(E_6 - E_2) - 0.6165}$ for right-skewed distributions ($\boldsymbol{E_i = q(i/8)}$ the sample ~~size $\boldsymbol{n_{J\sigma}}$ for the Gaussian, Exponential and Pearson family (panel **d**, see also Table 1~~octile).

2.3 Results and discussion

The parameter $\alpha = 1.5$ used in the boxplot rule is sensitive to the sample size n, and the relationship $\alpha(n) = A\log(n) + B + \frac{C}{n}$ corrects for this effect for both light-tailed distributions (Pearson family, Fig. 1a) and heavy-tailed distributions (GEV family, Fig. 1d). The value of A, B and C depends on the outlier threshold level and the nature of the distribution. The convention in this study is to set the expected number of erroneously flagged outliers to $f(n) = 0.001\sqrt{n}$, which leads to homogeneous A and B values among the Pearson family ($A = 0.8 \pm 0.3$, $B = 3 \pm 1$, Fig. 1b,c) used to numerically determine $C = 36$ (supplementary material). Because the value of A and B rapidly diverges for heavy-tailed distributions, a model adapted to the shape of the residuals is required (Fig. 1e,f). To keep this model simple, the asymmetry of a distribution (i.e., the skewness) is ignored in this study in order to only focus on the weight of the heavier tail. Possible outliers might not be flagged on the light tail of an asymmetric distribution (risk of type II error), but residuals with strong asymmetry are usually produced when the range of possible values is semi-bounded (e.g., precipitation in $[0, +\infty[$), which makes the detection of errors trivial (negative precipitation). For this purpose, the parameter m_* is a robust predictor of the kurtosis excess (breakdown point of 12.5%) that has been slightly modified from Kim & White (2004). Finally, $\alpha(n) = g_A(m_*)\log(n) + g_B(m_*) + \frac{36}{n}$ for $n \geq 9$ and $m_* \in [0,2]$ with the functions g_A and g_B parametrized on both families (Fig. 1e,f). The Fréchet distribution has been excluded because its tails are decaying too rapidly (the A and B coefficients are bounded despite an extreme kurtosis).

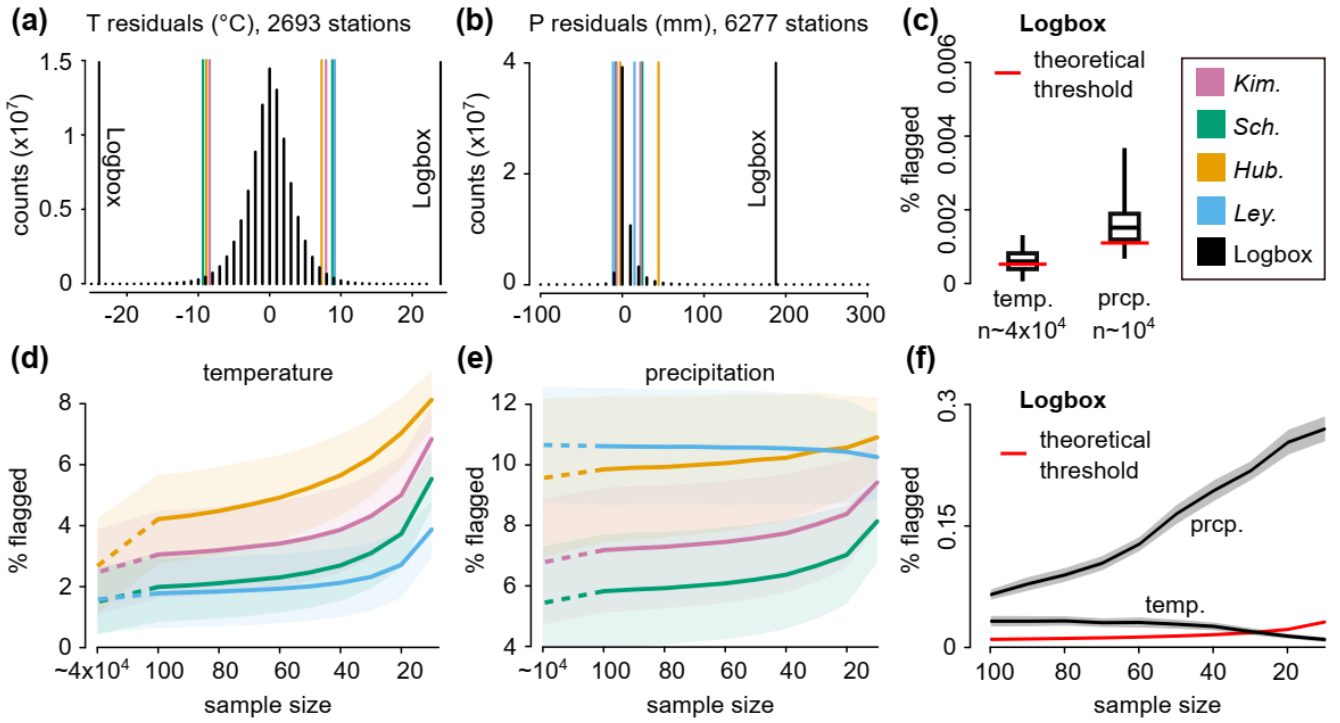


Fig. ~~The original boxplot rule captures 99.3% of a Gaussian population using the constant α = 1.5. In Table 1, this constant is shown to take larger values when considering wider windows (above 99.3%) and non Gaussian distributions (Pearson family, Fig.1a). In the Gaussian case, α ranges from $\alpha_{3\sigma}^{\mathcal{N}} = 1.7$ to $\alpha_{5\sigma}^{\mathcal{N}} = 3.2$. Both values are similar to the α = 1.5 and α = 3 originally used by Tukey (1977) to describe "outside" and "far out" outliers. These criterions are nonetheless too restrictive for non-Gaussian data: the median of the $\alpha^{+}(p_{j\sigma})$ values from the Pearson family ranges from $\alpha_{3\sigma}^{\mathcal{P}} = 3.8$ (therefore above $\alpha_{5\sigma}^{\mathcal{N}}$!) to $\alpha_{5\sigma}^{\mathcal{P}} = 9.4$.~~

~~**The difference in α values among the distributions of the Pearson family is interesting to visualize in the (κ_{ex}, S^2) space (**~~2. Comparison between five outlier detection methods performed on two sets of residuals (temperature and precipitation) obtained from weather stations with daily measurements over at least 100 years. The two histograms (panels **a** and **b**) represent aggregated residuals from all stations (for visualization purpose only) and show counts with at least 100 daily occurrences, with the median of the lower/upper threshold displayed for each method. For the methods Kim. (Kimber, 1990), Sch. (Schwertman et al., 2004), Hub. (Hubert & Vandervieren, 2008) and Ley. (Leys et al., 2013), the mean percentage ($\pm 1\ SD$) of flagged data is shown for sample sizes varying from 10 to 100, and for all available points per station ($n \sim 4 \times 10^4$ for the temperature and $n \sim 10^4$ for the precipitation, panels **d** and **e**). For Logbox (panels **c** and **f**), this percentage is calculated by pooling all points, and the variability is estimated with a random resampling of stations (see method). The theoretical threshold is the expected percentage of erroneously flagged outliers ($p_{theo} = f(n) \times \frac{100}{n} = \frac{0.1}{\sqrt{n}}\%$).

The Logbox procedure is tested and compared with four other models on daily precipitation and temperature residuals from century-old weather stations (Fig. 2). It is firstly visually striking that the outlier threshold from the four traditional methods

cut too many data points not only for the precipitation but also for the temperature residuals (Fig. 2a,b). The percentage of flagged data points per station varies around $2 \pm 0.5\%$ for the temperature (Fig. 2d, median of $36634 \approx 4 \times 10^4$ days per station), and from 5.5% (Sch.) to 10.5% (Ley.) for the precipitation (Fig. 2e, median of $8352 \approx 10^4$ wet days per station). The reason for the large discrepancy between observed and expected percentage of flagged outliers ($\sim 0.7\%$ based on the boxplot rule) is that these four methods have been designed for *nearly*-Gaussian residuals. Even daily temperatures are diverging from normality because the fitting model used to extract residuals from the time series minimizes the root-mean-square-error. The anomalies are therefore more concentrated around 0 than those produced by a Gaussian, but with larger extremes (Fig. 2a, leptokurtic distribution). For small samples of temperature residuals, the type I error is even higher due to the inaccuracy of the quantiles: from 2.5% (Sch.) to 7% (Hub.) of points are cut for $n = 20$ (Fig. 2d). The precipitation is less affected by the sample size effect because the type I error was already high in large samples. This analysis proves that none of the traditional methods is suitable to the outlier detection in non-Gaussian residuals.

In comparison, the Logbox procedure shows a percentage of flagged outliers close to the expected values for large sample sizes (Fig. 2c), with 0.0006 ± 0.0003 % for the temperature (expected value of 0.0005%) and $0.0017 \pm 0.0009\%$ for the precipitation (expected value of 0.001%). These results are *surprisingly* accurate knowing that 12.5% of the extreme values are disregarded for robustness reasons (m_*), and also knowing that Logbox has only been parametrized on theoretical distributions (Pearson & GEV family). For smaller sample sizes ($n < 30$ in Fig. 2f), the precipitation residuals are cut too frequently ($0.2 - 0.3\%$) compared to the expected threshold ($\sim 0.03\%$), but the temperature residuals are not cut enough. The constant parameter used to correct for a sample size effect ($C = 36$) is only adapted to nearly-Gaussian residuals, and it cannot be better estimated because any predictor (such as m_*) becomes inaccurate at smaller sample sizes. However, the percentage of flagged outliers remains within one order of magnitude of the expected threshold, which is a reasonable compromise between type I errors (precipitation) and type II errors (temperature).

~~Fig. 1b,c). A non-linear relationship is observed between α and (κ_{ex}, S^2), and the direction of this relationship depends on the outlier threshold. For the $\pm 3\sigma$ convention (Fig 1b), α increases with both κ_{ex} and S^2. For the $\pm 5\sigma$ convention (Fig 1c), α increases with κ_{ex} but decreases with S^2, which shows that a higher skewness does not necessary lead to more extreme events. Different families of distributions were also investigated (Jones, 2015), such as the Generalized Lambda Distribution (GLD) system (Carling, 2000). However, all distributions from this family have a bounded support and they therefore produce unrealistic datasets ($\alpha_{3\sigma} = \alpha_{4\sigma} = \alpha_{5\sigma}$, see supplementary material).~~

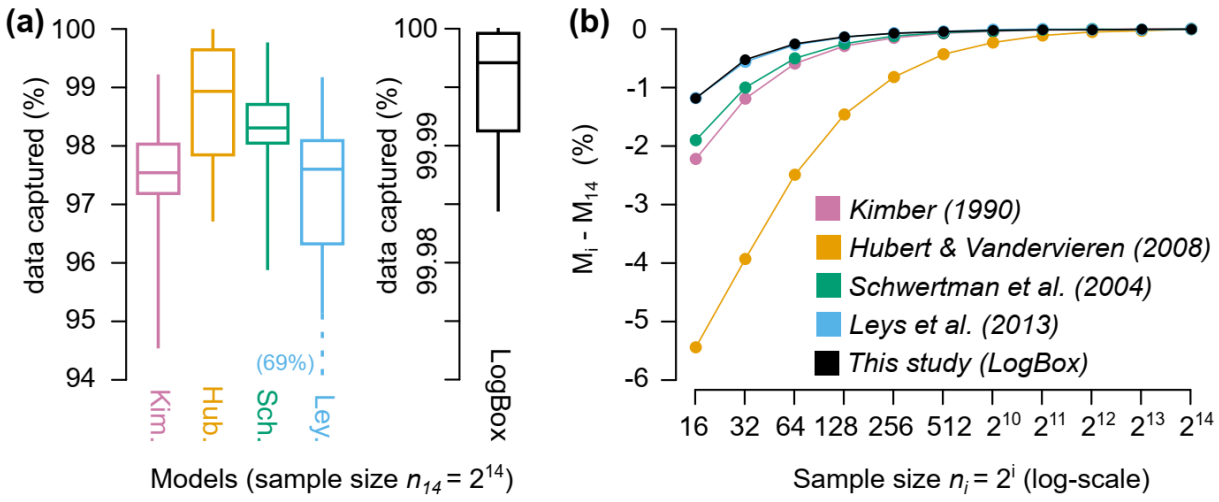


Fig. 2. Comparison between the five models performed on a subset of 600 distributions from the Pearson family (100 random distributions from the Gamma, Inverse gamma, Beta, Betaprime, Pearson IV and Student). Percentage of data captured by each model for a large sample size ($\boldsymbol{n_{14} = 2^{14} = 16384}$, panel **a**). Deviation of the median value $\boldsymbol{M_i}$ (associated with a sample size $\boldsymbol{n_i = 2^i}$) from $\boldsymbol{M_{14}}$ (panel **b**). The boxplots show 95% of the 600 distributions (the whiskers are the 2.5% and 97.5% quantiles).

The comparison between models is described in the following. For a purely Gaussian distribution, *Kim.* and *Hub.* theoretically capture 99.3% of the data (boxplot rule) while *Sch.* and *Ley.* capture 99.73% ($\pm 3\sigma$ convention). For non-Gaussian distributions with a large sample size, these values are found to be lower: 97.5%, 98.9%, 98.3%, and 97.6%, respectively (Fig. 2a). These four models therefore produce a high number of erroneously flagged outliers (false positives), which is not the case of *LogBox* that captures 99.997% of data (less than one outlier flagged on average in the sample size of 2^{14}). A concern could be that the cutting threshold of *LogBox* is too high, and, therefore, a large number of outliers would not be flagged in real datasets (false negatives). The comparison with *tsoutliers* in part II will demonstrate that this is not the case (Table 2).

Models also show different sensitivity to the type of distribution encountered (error bars in Fig. 2a). The most stable model is *LogBox* (95% of the distributions fall between 99.98% and 100% of data captured) while *Ley.* is the most sensitive with a percentage of data captured varying from 69.3% to 99.2%. This poor performance can be explained by the use of *MAD*, which contains a scaling factor parametrized on the standard deviation of the Gaussian (Leys et al., 2013).

Finally, the sensitivity of the models to the sample size is tested (Fig. 2b). All models show a negative bias in the percentage of data captured compared to the large sample size. This bias is minimal with *LogBox* (−1.2%) but important with *Hub.*

~~(−5.4% for $n = 16$). This comes from the complexity of their model: the Medcouple is a remarkable estimator of the skewness but it requires a large sample size to reach convergence.~~

To summarize, Logbox ~~is a simple method inspired by the Boxplot~~enhances the boxplot rule ~~to flag outliers~~ by considering the sample size effect and by adapting the cutting thresholds to the data. This method has been implemented in ~~univariate datasets. It has been adapted to non-Gaussian data with a large sample size and shows good performances when compared to other methods~~ the function *ctbi.outlier* **(**in the ~~literature. This justifies its implementation~~R package **ctbi)** that will be used to flag potential outliers in the residuals obtained by the aggregation procedure described in part II.

3 Part II, the ctbi procedure

3.1 Context

This second part is dedicated to the pre-processing, partial imputation and aggregation of univariate time series. In order to flag outliers, one first needs to produce residuals that represent the variability around the *signal*. In its simplest form, the time series y_t is represented with the following additive decomposition (Hyndman & Athanasopoulos, 2018): $y_t = T_t + S_t + \epsilon_t$, with T_t a long-term trend, S_t a cyclic component (originally, *seasonal component* but the term cyclic is preferred here as it is more generic) with period τ ($\forall\, t, S_t = S_{t+\tau}$) and ϵ_t the residuals~~.~~ that are considered to be stationary. A popular algorithm that performs this decomposition is the Seasonal and Trend decomposition using Loess (or **STL**, Cleveland et al., 1990), that is robust to the presence of outliers. The enhanced version of the algorithm, **STLplus** (Hafen, 2016), is also robust to the presence of missing values and data gaps. Unfortunately, there are three major drawbacks to using **STLplus** in the general case: (i) This algorithm has specifically been designed for signals showing seasonal patterns, which makes it less relevant for other types of data; (ii) The long-term trend based on loess needs several input parameters (s.window, s.degree,..) and the decomposition is therefore not unique; (iii) The algorithm has a complexity of $O(n^2)$ due to the loess, which is resource intensive and not adapted to long time series ($n > 10^7$). In particular, the first point explains why the function *tsoutliers* needs to use a smoothing function (Friedman, 1984) to complement the **STL** procedure.

A new robust and nonparametric procedure (**ctbi**) is proposed instead to calculate T_t and S_t using non-overlapping bins. Outliers are flagged in the residuals ϵ_t with the ~~*LogBox*~~Logbox method described in part I, and imputation is performed using $T_t + S_t$ if the cyclic pattern is strong enough, which is quantified by a new index introduced in this study (the Stacked Cycles Index or SCI). Bins with sufficient data can finally be aggregated, while other bins are discarded. The procedure is simple (entirely described in Fig. 3), the long-term trend T_t is unique and non-parametrized (based on linear interpolations crossing each bin), the cyclic component S_t is simply the mean stack of bins using detrended data (equivalent to STL for periodic time series). The algorithm complexity is of the order of $O(n\, log(n))$ because the loess is not necessary anymore. In the following,

the procedure is first described more in details and then applied to three case studies (a temperature, precipitation and methane dataset) that have been contaminated with outliers, missing values and data gaps. Comparison with the raw data demonstrates the reliability of the **ctbi** procedure, whose performance is compared to *tsoutliers*.

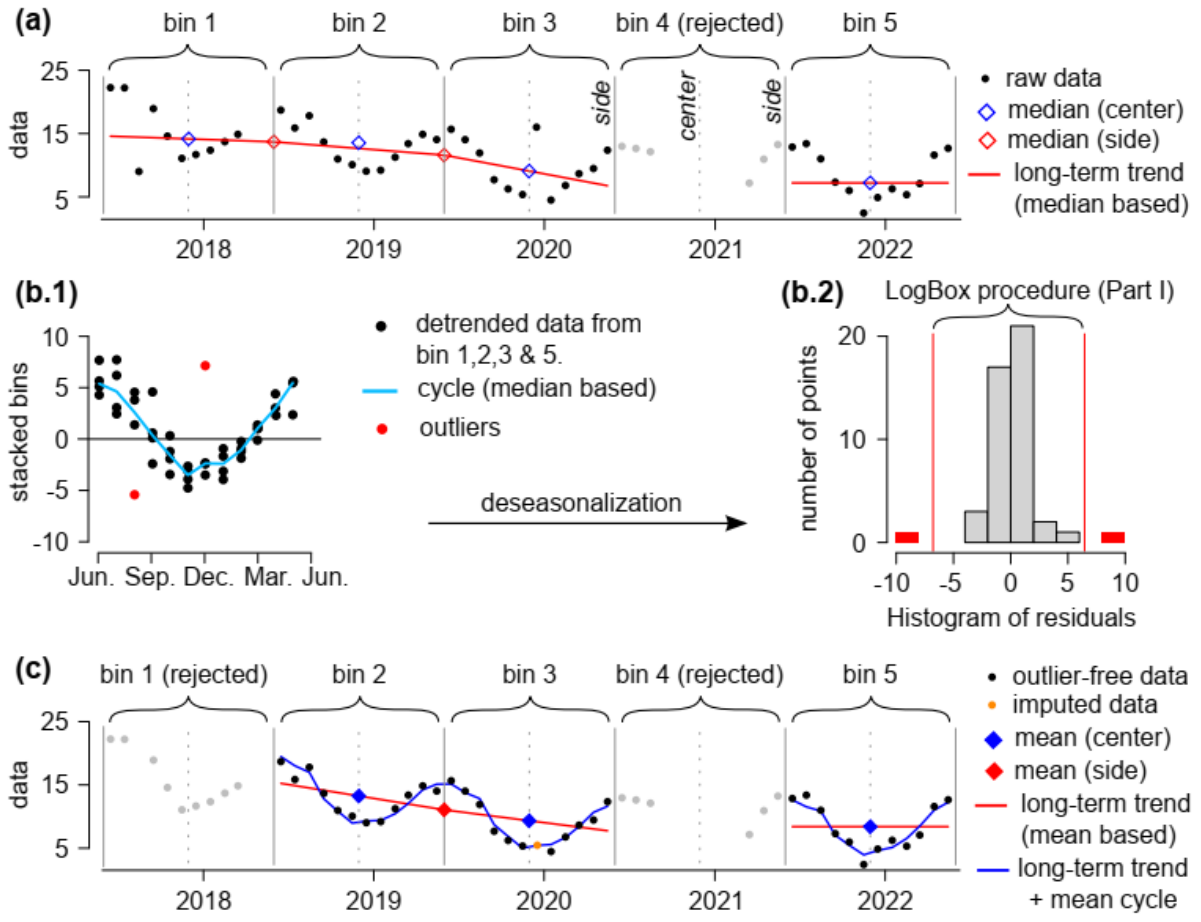
3.2 Method

3.2.1 Definitions

Bin: a *bin* is a time window characterized by a left *side* (inclusive), a right *side* (exclusive), a *center* and a *period* (e.g., 1 year in Fig. 3a). Any univariate time series can be decomposed in a sequence of non-overlapping bins, with the first and last data point contained in the first and last bin, respectively (Fig. 3a). The *bin size* n_{bin} is the rounded median of the number of ~~data~~ points (including NA values) present in each non-empty bin ~~of the sequence~~. A bin is *accepted* when its number of non-NA data points is above $n_{bin}(1 - f_{NA})$ with $f_{NA} \in [0,1]$ the maximum fraction of NA values per bin (input left to the user). Otherwise, the bin is *rejected* and all its data points are set to NA (Fig. 3a, bin 4).

Long-term trend: the *long-term trend (median based)* is a linear interpolation of the median values associated with each side (calculated between two consecutive centers, see Fig. 3a). A side value is set as missing if the number of non-NA data points (between the two nearest consecutive centers) is below $n_{bin}(1 - f_{NA})$. To solve for boundaries issues and missing sides values, the interpolation is extended using the median value associated with each center (bin 1, 3 & 5 in Fig. 3a). Once the outliers have been quarantined, the *long-term trend (mean based)* will be calculated following the same method but using the mean instead of the median (Fig. 3c).

Cycle: the *cycle (median based)* is composed of n_{bin} points that are the medians of the stack of all accepted bins with the long-term trend (median based) removed (Fig. ~~3.b.1~~3b1). Once the outliers have been quarantined, the *cycle (mean based)* will be the mean stack of accepted bins with the long-term trend (mean based) removed (~~Fig. 4a;~~ bin 2, 3 & 5 in Fig. 4a). The cyclic component S_t is the sequence of consecutive cycles.


(a)
bin 1
bin 2
bin 3
bin 4 (rejected)
bin 5
data
side
center
side
raw data
median (center)
median (side)
long-term trend
(median based)
2018
2019
2020
2021
2022
(b.1)
stacked bins
detrended data from
bin 1,2,3 & 5.
cycle (median based)
outliers
deseasonalization
Jun. Sep. Dec. Mar. Jun.
(b.2)
LogBox procedure (Part I)
number of points
Histogram of residuals
(c)
bin 1 (rejected)
bin 2
bin 3
bin 4 (rejected)
bin 5
outlier-free data
imputed data
mean (center)
mean (side)
long-term trend
(mean based)
long-term trend
+ mean cycle

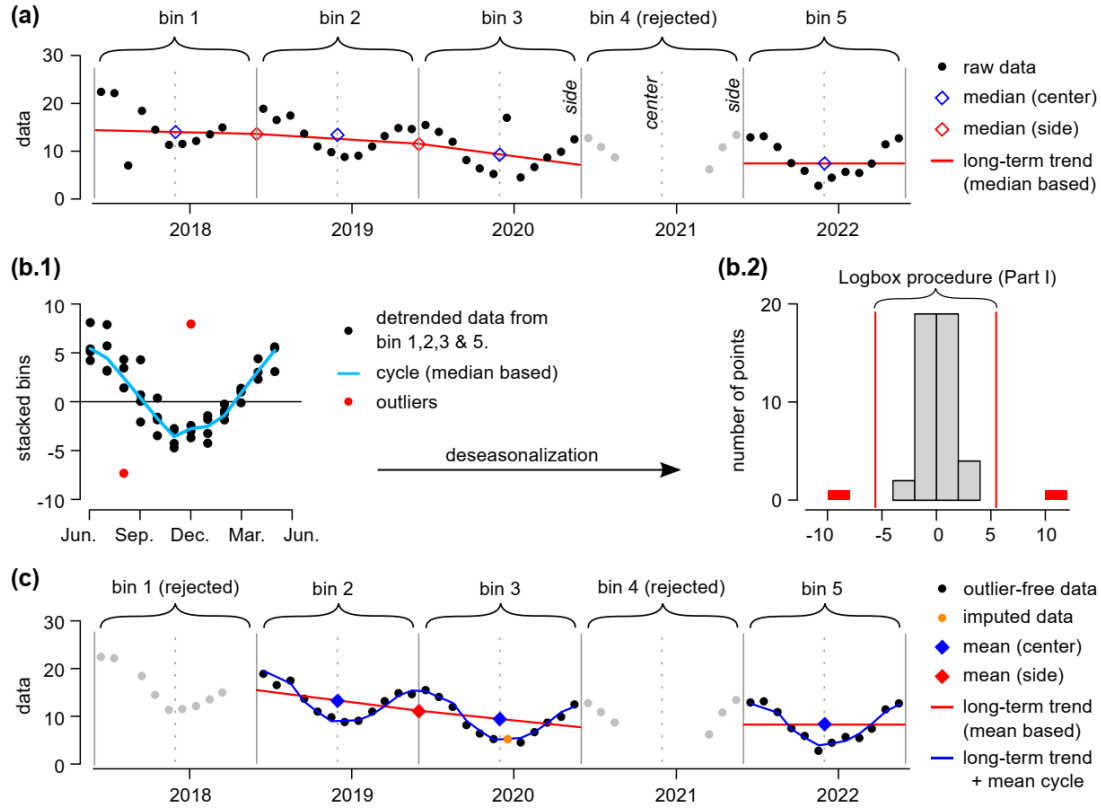


Fig. 3. Example of the aggregation procedure with the following inputs: bin side = 2020-06-01, bin period = 1 year, $f_{NA} = 0.2$ (minimum of 10 months of data for a bin to be accepted~~), $k = 0.6$ (outlier level~~) and $SCI_{min} = 0.6$ (cyclic imputation level). The bin 4 has been rejected because it contains only 6 months of data (panel **a**). Two outliers have been flagged in the residuals (detrended and deseasonalized data, panel **b.2**). After the outliers have been replaced with NA values, the bin 1 has been rejected (9 months of data), and the long-term trend and cycle have been updated using the mean instead of the median (panel **c**). A point in bin 3 has been imputed based on the cyclicity ($SCI_{min} \leq SCI = 0.61$).

Stacked Cycles Index: SCI ≤ 1 is an adimensional parameter quantifying the strength of a cycle based on the variability around the mean stack (Fig. 4). Its structure is similar to another index developed in a former study (Wang et al., 2006), however a factor of N_{bin}^{-1} has been added to correct for a bias emerging at a small number of bins (N_{bin} is the number of accepted bins). This correcting factor has been calculated based on stationary time series of Gaussian noise (with therefore a null cyclicity per definition, see supplementary material).

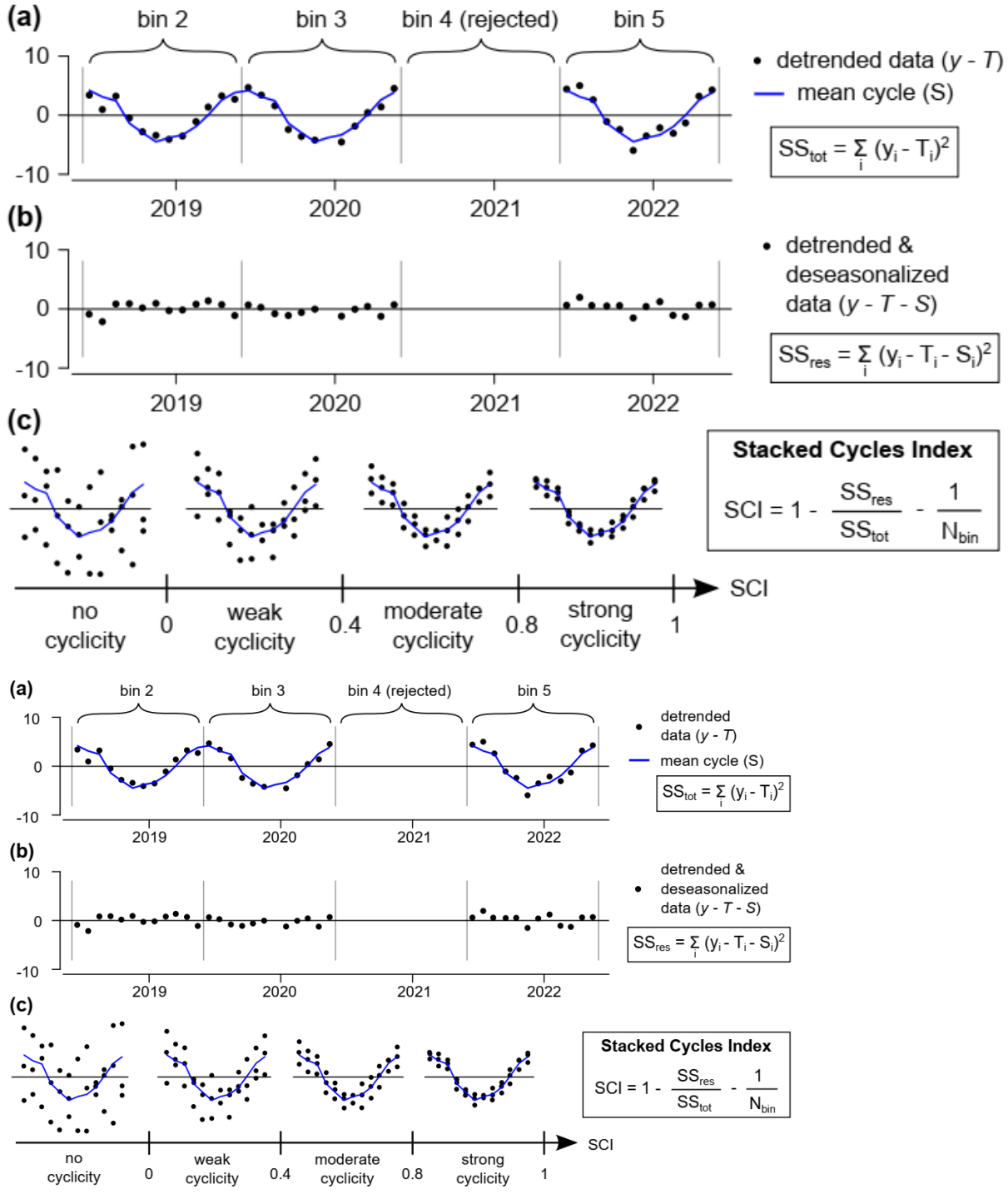


Fig. 4. The Stacked Cycles Index ($SCI \leq 1$) quantifies the strength of the cyclicity associated with the period of a bin ~~(data from Fig. 3c).~~ The long-term trend (mean based) is first removed to compute the total sum of squares (panel **a**). Then the cyclic component (mean based)

is also removed to compute the sum of squared residuals (panel **b**). SCI is the coefficient of determination minus N_{bin}^{-1} to correct for a bias emerging at a small number of bins, with N_{bin} the number of accepted bins (here $N_{bin} = 3$, panel **c**).

3.2.2 Ctbi procedure

Inputs.

1. The univariate time series (1st and 2nd column: time and raw data, respectively~~)~~).
2. One bin center or one bin side (e.g., 2020-06-01~~)~~).
3. The period of the bin (e.g., 1 year~~)~~).
4. The aggregation operator (mean, median or sum~~)~~).
5. The range of possible values (default value $y_{lim} \in$ ~~$]-\infty,+\infty[$~~$]-\infty,+\infty[$).
6. The maximum fraction of NA values per bin (default value $f_{NA} = 0.2$~~)~~).
7. The ~~*k* outlier level~~ A, B, C coefficients used in ~~*LogBox* (~~the Logbox method (automatically calculated by default ~~value $k = 0.6$~~), $coeff.outlier = 'auto'$).
8. The minimum SCI for imputation (default value $SCI_{min} = 0.6$~~)~~).

Outputs.

1. The original dataset, with ~~8~~9 columns: (i) time; (ii) outlier-free and imputed data; (iii) index of the bins associated with each data points (the index is negative if the bin is rejected); (iv) long-term trend; (v) cyclic component; (vi) residuals (including the outliers); (vii) quarantined outliers; (~~vii~~viii) value of the imputed data points; (~~viii~~ix) relative position of the data points in their bins, between 0 (the point falls on the left side) and 1 (the point falls on the right side~~)~~).
2. The aggregated dataset, with 10 columns: (i) aggregated time (center of the bins); (ii) aggregated data; (iii) index of the bin (negative value if the bin is rejected); (iv) start of the bin; (v) end of the bin; (vi) number of points per bin (including NA values); (vii) number of NA values per bin, originally; (viii) number of outliers per bin; (ix) number of imputed points per bin; (x) variability associated with the aggregation (standard deviation for the mean, MAD for the median and nothing for the sum~~)~~).
3. The mean cycle, with 3 columns: (i) time boundary of the first bin with n_{bin} points equally spaced; (ii) the mean value associated with each point; (iii) the standard deviation associated with the mean value.
4. ~~The~~A summary of the bins: the Stacked ~~Cycles Index~~
5. ~~The~~Cycle index (SCI), the representative number of data points per bin~~, n_{bin}~~
6. ~~4.~~ ~~The~~ (n_{bin}) and the minimum number of data points ~~with available data~~ for a bin to be accepted~~, $n_{bin\ min}$~~ ($n_{bin\ min}$).
5. A summary of the Logbox output: the coefficients A, B and C, m_*, the number of points used, the lower/upper outlier threshold.

Step 1, data screening. The bin size n_{bin} is calculated; values above or below y_{lim} are set to NA; the number of accepted bins N_{bin} is assessed; all data points within rejected bins are set to NA; the long-term trend and cycle (both median based) are calculated (Fig. 3a,~~b.1~~b1).

Step 2, outliers. Outliers are flagged in the residuals (detrended and deseasonalized data) using ~~the *LogBox* procedure~~Logbox (Fig. ~~3b.2~~3b2); outliers are quarantined and their values are set to NA; the number of accepted bins N_{bin} is updated; all data points within newly rejected bins are set to NA (bin 1 in Fig. 3c~~, bin 1~~).

Step 3, long-term trend and cycle (mean based): The long-term trend and the cycle are calculated using the mean instead of the median (Fig. 3c); SCI is calculated (Fig. 4).

Step 4, imputation: If $SCI > SCI_{min}$, all NA values in accepted bins are imputed with the long-term trend + the mean cycle (imputation bounded by y_{lim}). Repeat Step 3 and Step 4 three times to reach convergence.

Step 5, aggregation: Accepted bins are aggregated around their center.

3.2.3 Case studies

Three univariate datasets are chosen to illustrate the potential of the aggregation procedure (Fig. 5, first column). The first dataset is an in-situ temperature (in °C) measured during summer in the canopy of an Oak woodland of California (month of August, temporal resolution of 5 min), and provided by the National Ecological Observatory Network (NEON 2021, site SJER). The second dataset is an in-situ daily precipitation record (in mm) measured at the station of Cape-Leeuwin (South westerly coast of Australia) from 1990 to 2020 and available on the Global Historical Climatology Network (Menne et al., 2012; Xungang et al. 2012). The last dataset is a Methane proxy record (in ppbv) published in Loulergue et al. (2008) that covers 800000 years with irregular timesteps (varying from 1 to 3461 years, with a median of 311 years). None of the datasets contain obvious outliers or large data gap.

3.2.4 Contamination of the datasets

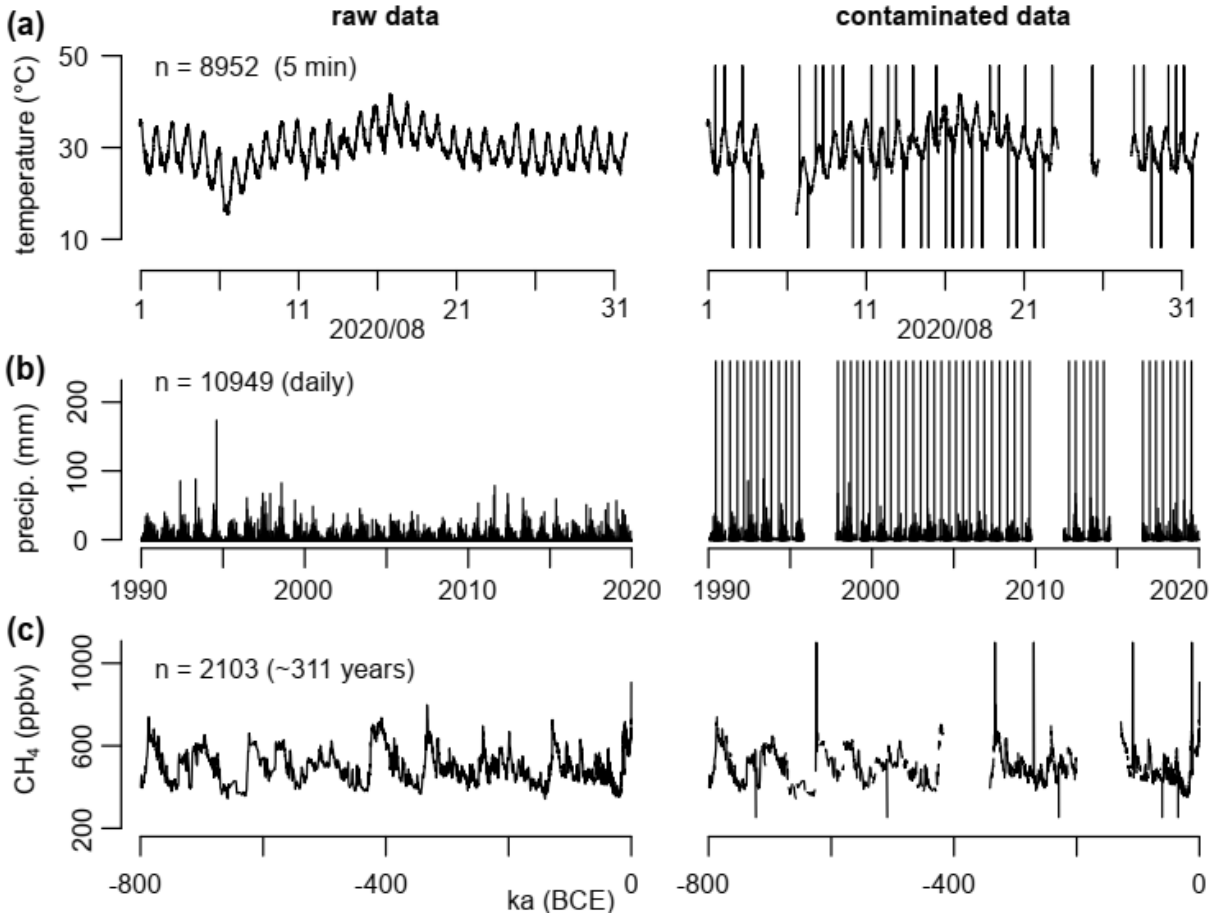
To test for the robustness of the aggregation procedure, the three raw datasets are contaminated by 30% (Fig 5, second column) with the use of three data gap (20% of the dataset), random NA values (9.5% of the dataset) and outliers (0.5% of the dataset). The three data gaps are picked with random length and position. The position of the outliers and the NA values follows a Poisson law. The value of the outliers is picked equal to $y_{min} - \frac{1}{2}(\mu - y_{min})$ or $y_{max} + \frac{1}{2}(y_{max} - \mu)$ with y_{min}, y_{max} and μ respectively the minimum, maximum and mean of the dataset~~. No negative outliers are set for the precipitation because these values are impossible.~~ (temperature and methane datasets). The precipitation is supposed to follow a heavy-tail distribution (extremes are more frequent), and negative values are impossible, which is why outlier values are set to $1.6 \times y_{max}$ instead (supplementary material).

3.2.5 Aggregation of the datasets

Each dataset (raw and contaminated version) is consecutively aggregated twice (Fig. 6). The temperature dataset is aggregated (using the mean) every hour ($n_{bin} = 12$) and then every day ($n_{bin} = 24$). The precipitation dataset is aggregated (using the sum) every month ($n_{bin} = 31$) and then every year ($n_{bin} = 12$). The methane dataset is aggregated (using the mean) every 2000 years ($n_{bin} = 4$) and then every 20000 years ($n_{bin} = 10$). For each dataset, the mean cycle of the second level of aggregation is shown in Fig. ~~5~~6 (second column).

The aggregation inputs are chosen as default values. The only exceptions are ~~$k = \infty$~~$coeff.outlier = NA$ and $SCI_{min} = $ ~~1~~NA for the raw data (outliers are not checked, data are not imputed), $f_{NA} = 1$ for the Methane dataset (bins with at least 1 non-NA data point are accepted due to the high irregularity in the sampling frequency) and $y_{lim} = [0, +\infty[$ for the precipitation dataset (negative precipitation are impossible).

The number of false positive (real data points flagged as outliers) and false negative (outliers that have not been flagged) are counted during the first level of aggregation (Table ~~2~~1), and compared with the *tsoutliers* function with λ = "auto", which means that the residuals have been transformed to follow a Gaussian with the Cox-Box method (Box & Cox, 1964), or λ = NULL, which means the original residuals are not transformed. The ~~Boxplot~~boxplot rule in *tsoutliers* uses $\alpha = 3$, and the long-term trend or cyclic component are not available for comparison.


(a)
raw data
contaminated data
temperature (°C)
50
30
10
n = 8952 (5 min)
1
11
21
31
2020/08
(b)
precip. (mm)
200
100
0
n = 10949 (daily)
1990
2000
2010
2020
(c)
CH4 (ppbv)
1000
600
200
n = 2103 (~311 years)
-800
-400
0
ka (BCE)

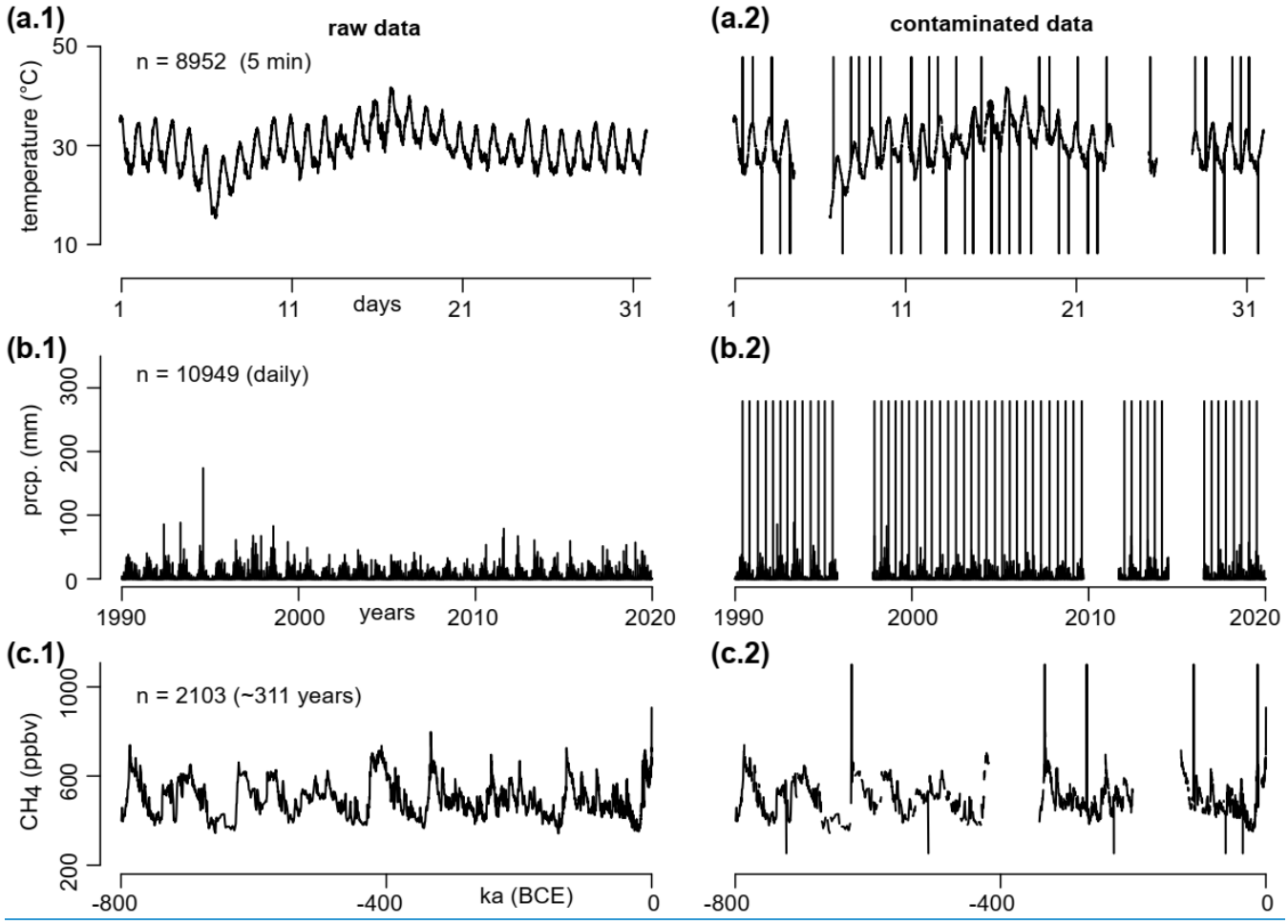
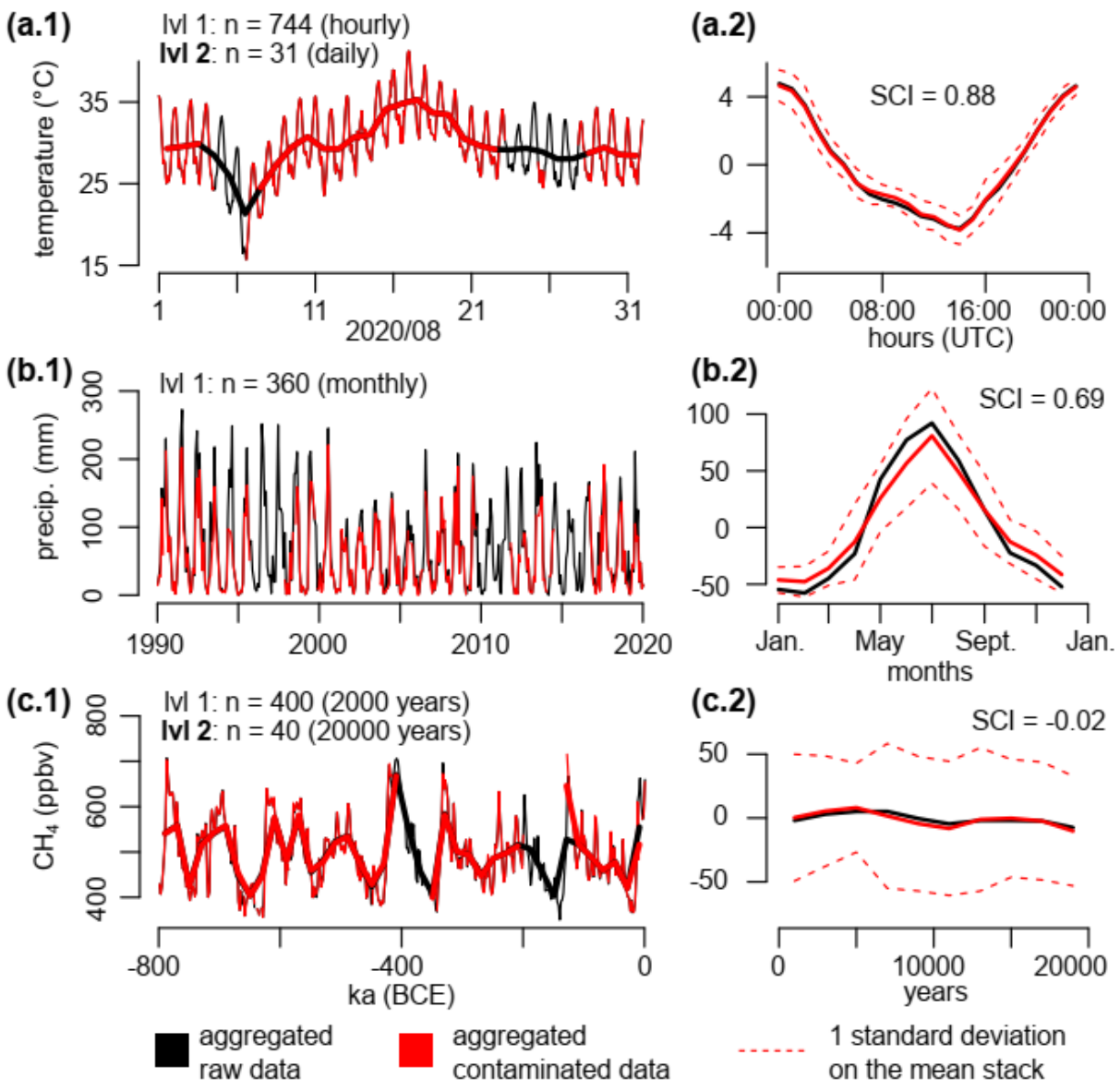


Fig. 5. Raw and contaminated versions of the three datasets used as case studies: temperature (panel **a**), precipitation (panel **b**) and methane (panel **c**). The sampling frequency is given in parenthesis. The contaminated versions contain three large data gaps (20% of the datasets), random missing values (9.5%) and random outliers (0.5%) set as a constant level ~~below the minimum value and above the maximum value~~.

3.3 Results and discussion

The three univariate time series have been chosen as case studies due to their various statistical characteristics that are commonly seen in the scientific or economic field (Fig. 5, 1st column). The long-term trend follows smooth or moderate variations in the temperature and precipitation datasets, but shows a much higher volatility in the methane dataset. The cyclic pattern varies from strong diurnal cycles (temperature) and moderate seasonal cycles (precipitation) to no apparent cyclicity over a period of 20000 years ~~period~~ (methane). The detrended and deseasonalized residuals follow distributions from ~~gaussian~~Gaussian (temperature) or seemingly ~~exponential~~Exponential (methane) to heavy-tailed (precipitation). Finally, the sampling frequency goes from sub-hourly (temperature) or daily (precipitation) to highly variable (1 to 3461 years, methane).

To test the limits of the aggregation procedure, these three datasets are *severely* contaminated by data gaps, outliers and missing values (Fig. 5, 2nd column).



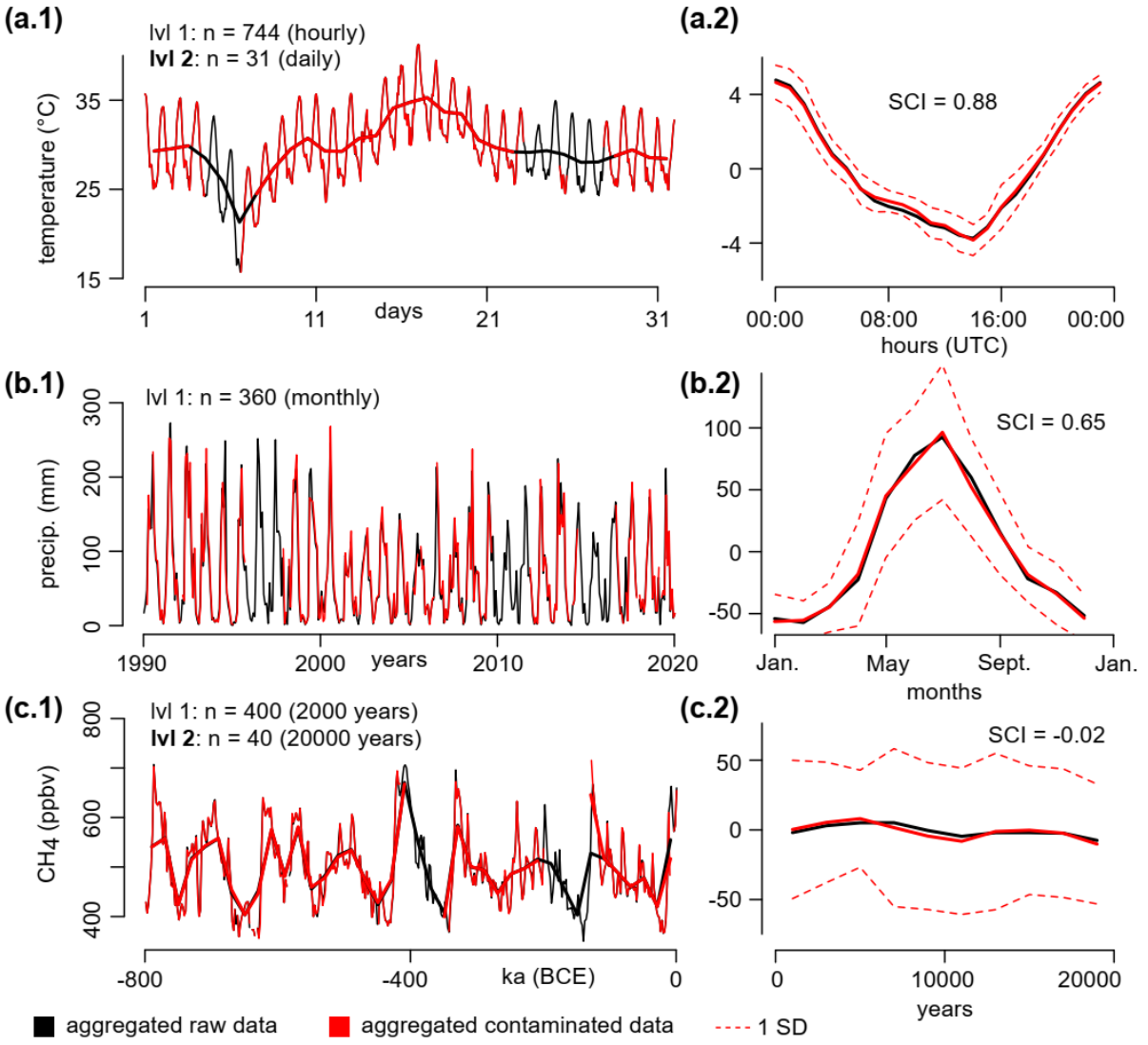


Fig. 6. Aggregation of the temperature (panel **a**), precipitation (panel **b**) and methane (panel **c**) in two consecutive levels: 1 (thin lines) and 2 (bold lines). Only the first level of aggregated precipitation is shown for clarity. Black and red colors are associated with the raw and contaminated datasets, respectively. The mean cycles of the second level of aggregation are shown in the second column, with their SCI displayed (the raw and contaminated versions share similar values).

The first level of aggregation recovers most of the destroyed signal with ~80% of the bins being accepted for all three datasets (Fig. 6). In these accepted bins, all outliers have been correctly flagged (Table ~~2, zero false negatives).~~1). The mean percentage of difference between the contaminated and raw aggregates (level 1) is virtually zero for the temperature (0 ± 0.1~~%, 1 standard deviation), small for~~%), the methane ($-0.1 \pm 2\%$) ~~but large for~~ and the precipitation ~~($-9 \pm 24\%$). This comes from the significant number of extreme precipitations events (67 days) that have been erroneously flagged as outliers (Table 2, false positives). Daily precipitation events are known to follow heavy tailed distributions (Wilks & Wilby, 1999), which is why the~~

~~default outlier level of $k = 0.6$ is insufficient here. A value of $k \sim 5$ is optimum in this case as it preserves the extreme events while cutting the outliers (the mean percentage of difference becomes~~ ($0 \pm 17\%$). For the Methane dataset, the ~~5 false positives (Table 2) come from the difficulty for~~ only false positive (Table 1) is located at the beginning of the time series (modern time), because the anthropogenic change in CH_4 is unprecedented when compared to the geological history (the long-term ~~trend to properly~~fit does not capture the abrupt ~~changes in CH_4 over few centuries. Again, this problem can be solved by increasing $k = 0.6$ to $k \sim 1$ without affecting the false negatives. The~~increase due to climate change). In comparison, the function *tsoutliers* successfully flags the outliers in the contaminated ~~Temperature~~temperature and ~~Methane~~methane datasets (with the Cox-Box method), however it fails with the contaminated daily precipitation dataset (Table ~~2~~1). This comes from the inability of ~~the procedure~~*tsoutliers* to handle heavy tailed distributions, creating 55 false negatives (all outliers have been missed) with the Cox-Box method and ~~1187~~1125 false positives ~~(real data points seen as outliers)~~ without it, due to the limitation of the ~~Boxplot~~boxplot rule ~~with~~using a constant $\alpha = 3$ (see part I).

Procedure	**ctbi**			*tsoutliers* (with/without Cox-Box)		
Datasets	T	P	CH_4	T	P	CH_4
Number of false positives	0	~~67~~0	~~5~~1	0 / 0	0 / ~~1187~~1125	0 / 3
Number of false negatives	0	0	0	~~1~~3 / 0	55 / 0	0 / 2

Table ~~2~~1. Number of false positives (real data points flagged as outliers, type I error) and false negatives (outliers that have not been flagged, type II error) for the contaminated Temperature (n=8952), Precipitation (n=10949) and Methane (n=2103) datasets shown in Fig. 5 with the **ctbi** procedure ~~(using $k = 0.6$ in the *LogBox* method)~~ and the *tsoutliers* function (with/without the Cox-Box method). ~~For the precipitation (respectively Methane) dataset, $k = 5$ (resp. $k = 1$) will minimize both the false positives and false negatives with **ctbi**.~~

The second level of aggregation has been performed to test for the cyclicity in the signal (Fig. 6, 2nd column) using the mean cycles and their associated Stacked Cycles Index (Fig. 4). The raw and contaminated mean cycles share similar magnitude within 1 standard deviation on the mean, and their SCI are the same: -0.02 for the methane (no apparent cycles of 20000 years period), 0.~~69~~65 for the precipitation (moderate seasonality) and 0.88 for the temperature (strong diurnal cycles). The SCI reveals itself being useful when comparing signals of different nature or periodicities, which is not possible for seasonal indices that only focuses on one field (e.g., hydrology) or data format. (e.g., monthly) such as the seasonality index of Feng et al. (2013). ~~Interestingly, the **ctbi** procedure manages to recover the seasonality of the precipitation dataset despite cutting most of the extreme events (Fig. 6, b2). This result illustrates the fact that climatic models are able to capture the mean trend while having difficulties to simulate exceptional events (Asadieh & Krakauer, 2015).~~

The cyclicity seen in the temperature and precipitation is strong enough to impute the missing data in all accepted bins, which further improves the reconstruction of the signal. ~~For example, 11 months have been imputed in 9 different years for the precipitation dataset. Using $k = 0.6$, the mean percentage of difference with the raw data went for these years from $-24 \pm 10\%$ (without imputation) to $-20 \pm 5\%$ (with imputation).~~ Because SCI has a similar structure than a coefficient of determination, imputations based on high SCI (> 0.6) are respecting the original signal, which is sometimes not the case with a linear interpolation. ~~Again, the choice of performing or ignoring the imputation is left to the user with the input parameter SCI_{min} that will be compared to SCI (see method).~~ These three case studies demonstrate that **ctbi** is capable of aggregating signals of poor quality that have a stationary variance in the residuals. The next section explains how to handle more complex time series.

3.4 Limits & recommendations

The **ctbi** procedure ~~should be used as a *complement* of an~~complements the expert-knowledge related to a dataset, but it does not ~~as a *replacement*.~~replace it. In particular, this procedure is not capable of ~~flagging~~detecting long periods of instrument failure or human error, and it is essential to flag them manually and/or visually before running **ctbi**. This procedure also presents difficulties to pre-process signals with a complex seasonality associated with residuals of non-stationary variance. A typical example is a daily precipitation record with a pronounced monsoon: several months of droughts (low variability in the signal) are followed by few weeks of severe floods (high variability). These two periods do not have the same statistical characteristics, and need to be treated separately. In this situation, two pools of bins can be created using the MAD as a robust indicator of variability within each bin. The procedure is the following: (i) apply **ctbi** with the median operator (do not flag outliers or impute data, ~~$k = \infty$~~ $coeff.outlier = NA$ and $SCI_{min} = ~~1~~NA$) so that each bin will be associated with a specific MAD; (ii) Flag bins with a low MAD ('dry' season) and a high MAD ('wet' season); (iii) split the raw data into two datasets of bins with a low and high MAD, respectively; (iv) apply **ctbi** separately to each dataset to flag outliers and/or impute data; (v) merge the two datasets. This procedure is successfully applied to a soil respiration dataset (supplementary material).

Other issues can usually be addressed by varying the inputs: period of the bin, maximum ratio of missing values per bin (f_{NA}), ~~outlier level (k)~~ and cyclic imputation level (SCI_{min}). It is recommended to pick the period of a bin so that it contains on average between 4 and ~50 data points. Below 4 would decrease the breakdown point to unsafe levels (1 outlier would be enough to contaminate the bin), and above 50 would produce a long-term trend that might not properly capture the variability in the signal. A maximum of 20% of the bin can be missing by default ($f_{NA} = 0.2$), but when data are sparse and irregularly distributed, a value of $f_{NA} = 1$ is possible (example ~~of~~with the Methane dataset: bins with only 1 data point were accepted). ~~An outlier level of $k = 0.6$ will work in most cases, but can vary up to $k \sim 10$ for time series with exceptional spikes.~~ Finally,

the imputation level (default of $SCI_{min} = 0.6$) can vary between 0 (forced imputation even without cyclic pattern) and 1 (no imputation).

4 Conclusion (Part I & II)

Although univariate time series are the simplest type of temporal data, this study reveals a lack of consensus in the literature on how to objectively ~~isolate~~flag outliers ~~from the signal~~ especially in raw data of poor quality. In part I, a comparison between outlier detection methods ~~for univariate datasets has shown that~~ is performed on daily residuals from century-old weather stations (precipitation & temperature data). All traditional outlier detection methods flag extreme events ~~are too often flagged~~ as outliers~~, especially~~ too frequently (type I error). The alternative procedure developed in ~~non-Gaussian populations with a large sample size. This led to a new method (*LogBox*) that~~this study (Logbox) improves the boxplot rule by replacing the original $\alpha = 1.5$ with $\alpha = ~~k~~A \log(n) + ~~1,~~B + \frac{36}{n}$, with ~~*n* the sample size~~A and ~~*k* left to the user (default value~~B determined with a predictor of ~~0.6).~~ the kurtosis excess (m_*). Logbox is parametrized on two families of distributions (Pearson & Generalized Extreme Value), and the theoretical percentage of type I error decreases with the sample size ($p_{theo} = \frac{0.1}{\sqrt{n}}\%$). Logbox therefore produces cutting thresholds that are tailored to the shape and size of the data, with a good match between observed and expected type I errors in the precipitation and temperature residuals.

In part II, a pre-processing procedure (**ctbi** for cyclic/trend decomposition using bin interpolation~~, implemented in R)~~) cleans, decomposes, imputes and aggregates time series based on data binning ~~has been proposed to clean, decompose, impute and aggregate signals~~. The strength of the cyclic pattern within each bin is assessed with a novel and adimensional index (~~SCI for~~the Stacked Cycles Index) inspired by the coefficient of determination.

The **ctbi** procedure is able to filter contaminated data by selecting bins with sufficient data points (input: f_{NA}) which are then cleaned from outliers (input: ~~*k*).~~$coeff.outlier$). The cyclic pattern within each bin is evaluated (SCI) and missing data are imputed in accepted bins if the cyclicity is strong enough (input: SCI_{min}). Most of the signal can be retrieved from univariate time series with diverse statistical characteristics, illustrated in this study with a temperature, precipitation and methane datasets that have been contaminated with gaps and outliers. Limits in the use of **ctbi** have been acknowledged for signals with a long-period of instrument failure, but also for signals presenting a complex seasonality. ~~However, **ctbi** is capable to handle the~~The last situation can be handled by splitting the raw data into two (or more) datasets containing bins with similar variability quantified by the Mean Absolute Deviation (MAD). The pre-processing procedure is then separately applied to each dataset to correctly identify outliers. ~~In any case, a prior knowledge of the data is essential to correctly choose f_{NA}, k and SCI_{min}, and it~~It is strongly recommended to examine the data before and after using ~~the procedure~~**ctbi** to ensure that rejected bins and flagged outliers seem reasonable, and to be transparent about the inputs used in your future study.

Author contribution

F.R.: Design, writing, coding.

Competing interests

The author declares no competing interests.

Acknowledgement

The author would like to warmly thank Dr. Rob Hyndman for his advice, as well as the ~~two~~three reviewers of this study. Dr. Jens Schumacher has particularly stimulated the improvement of the Logbox method. The author is also grateful to the Fonds de Dotations O that funded this project.

Data availability

The GHCN dataset is available on https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/. The Methane dataset is available on https://doi.org/10.1038/nature06950. The temperature dataset is available on https://doi.org/10.48443/2nt3-wj42.

Code availability

~~If the~~The **ctbi** package is ~~not~~ available ~~yet~~ on the comprehensive R Archive Network (CRAN~~), please use the version on https://github.com/fritte2/ctbi.~~). The code and data used in the study & the supplementary material ~~is~~are available on https://github.com/fritte2/ctbi_article.

References

Aguinis, Herman, Ryan K. Gottfredson, and Harry Joo. "Best-practice recommendations for defining, identifying, and handling outliers." Organizational Research Methods 16, no. 2 (2013): 270-301.

~~Asadieh, Behzad, and Nir Y. Krakauer. "Global trends in extreme precipitation: climate models versus observations." Hydrology and Earth System Sciences 19, no. 2 (2015): 877-891.~~

Borchers, Hans W., and Maintainer Hans W. Borchers. "Package 'pracma'." (2021).

Box, George EP, and David R. Cox. "An analysis of transformations." Journal of the Royal Statistical Society: Series B (Methodological) 26, no. 2 (1964): 211-243.

Brys, Guy, Mia Hubert, and Anja Struyf. "A robust measure of skewness." Journal of Computational and Graphical Statistics 13, no. 4 (2004): 996-1017.

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

Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." ACM computing surveys (CSUR) 41, no. 3 (2009): 1-58.

Cleveland, Robert B., William S. Cleveland, Jean E. McRae, and Irma Terpenning. "STL: A seasonal-trend decomposition." J. Off. Stat 6, no. 1 (1990): 3-73.
Feng, Xue, Amilcare Porporato, and Ignacio Rodriguez-Iturbe. "Changes in rainfall seasonality in the tropics." Nature Climate Change 3, no. 9 (2013): 811-815.
Friedman, Jerome H. *A variable span smoother*. Stanford Univ CA lab for computational statistics, 1984.
Hafen, Ryan. "Stlplus: enhanced seasonal decomposition of time series by Loess." R package version 0.5 1 (2016).
Hoaglin, David C., Boris Iglewicz, and John W. Tukey. "Performance of some resistant rules for outlier labeling." *Journal of the American Statistical Association* 81, no. 396 (1986): 991-999.
Hubert, Mia, and Ellen Vandervieren. "An adjusted boxplot for skewed distributions." Computational statistics & data analysis 52, no. 12 (2008): 5186-5201.
Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed on 2021/10/13.
Hyndman, Rob J., George Athanasopoulos, Christoph Bergmeir, Gabriel Caceres, Leanne Chhay, Mitchell O'Hara-Wild, Fotios Petropoulos, Slava Razbash, and Earo Wang. "Package 'forecast'." [Online] https://cran. r-project. org/web/packages/forecast/forecast. pdf (2020).
~~Jones, M. C. "On families of distributions with shape parameters." International Statistical Review 83, no. 2 (2015): 175-192.~~
Jenkinson, Arthur F. "The frequency distribution of the annual maximum (or minimum) values of meteorological elements." *Quarterly Journal of the Royal Meteorological Society* 81, no. 348 (1955): 158-171.
Kimber, A. C. "Exploratory data analysis for possibly censored data from skewed distributions." Journal of the Royal Statistical Society: Series C (Applied Statistics) 39, no. 1 (1990): 21-30.
Kim, Tae-Hwan, and Halbert White. "On more robust estimation of skewness and kurtosis." *Finance Research Letters* 1, no. 1 (2004): 56-73.
Leys, Christophe, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median." Journal of experimental social psychology 49, no. 4 (2013): 764-766.
Loulergue, Laetitia, Adrian Schilt, Renato Spahni, Valérie Masson-Delmotte, Thomas Blunier, Bénédicte Lemieux, Jean-Marc Barnola, Dominique Raynaud, Thomas F. Stocker, and Jérôme Chappellaz. "Orbital and millennial-scale features of atmospheric CH 4 over the past 800,000 years." Nature 453, no. 7193 (2008): 383-386.
Menne, Matthew J., Imke Durre, Russell S. Vose, Byron E. Gleason, and Tamara G. Houston. "An overview of the global historical climatology network-daily database." Journal of atmospheric and oceanic technology 29, no. 7 (2012): 897-910.
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.
NEON (National Ecological Observatory Network). Single aspirated air temperature, RELEASE-2021 (DP1.00002.001). https://doi.org/10.48443/2nt3-wj42. Dataset accessed from https://data.neonscience.org on October 13, 2021
Pearson, Karl. "X. Contributions to the mathematical theory of evolution.—II. Skew variation in homogeneous material." Philosophical Transactions of the Royal Society of London.(A.) 186 (1895): 343-414.
Pearson, Karl. "XI. Mathematical contributions to the theory of evolution.—X. Supplement to a memoir on skew variation." Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 197, no. 287-299 (1901): 443-459.
Pearson, Karl. "IX. Mathematical contributions to the theory of evolution.—XIX. Second supplement to a memoir on skew variation." Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 216, no. 538-548 (1916): 429-457.
Pearson, Ronald K. "Outliers in process modeling and identification." IEEE Transactions on control systems technology 10, no. 1 (2002): 55-63.
Ranjan, Kumar Gaurav, B. Rajanarayan Prusty, and Debashisha Jena. "Review of preprocessing methods for univariate volatile time-series in power system applications." Electric Power Systems Research 191 (2021): 106885.
Reiss, Rolf-Dieter, Michael Thomas, and R. D. Reiss. Statistical analysis of extreme values. Vol. 2. Basel: Birkhäuser, 1997.
Schwertman, Neil C., Margaret Ann Owens, and Robiah Adnan. "A simple more general boxplot method for identifying outliers." Computational statistics & data analysis 47, no. 1 (2004): 165-174.
Tukey, John W. Exploratory data analysis. Vol. 2. 1977.

Wang, Xiaozhe, Kate Smith, and Rob Hyndman. "Characteristic-based clustering for time series data." Data mining and knowledge Discovery 13, no. 3 (2006): 335-364.

~~Wilks, Daniel S., and Robert L. Wilby. "The weather generation game: a review of stochastic weather models." Progress in physical geography 23, no. 3 (1999): 329-357.~~

Xungang Yin, Steven Anthony, Ron Ray, Russell S. Vose, Byron E.Gleason, and Tamara G. Houston (2012): Global Historical Climatology Network - Daily (GHCN-Daily), Version 3. NOAA National Climatic Data Center. doi:10.7289/V5D21VHZ [2021/10/13]