Supplement of Hydrol. Earth Syst. Sci., 27, 349–361, 2023
https://doi.org/10.5194/hess-27-349-2023-supplement

Hydrology and Earth System Sciences
Open Access

Supplement of

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

François Ritter

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

Part I: outliers

1) **Determination of the C value in $\alpha(n) = A\ log(n) + B + C/n$**

The value of C has been parametrized on the Pearson family for $n = 9$ (m_* would break with a single outlier for $n = 8$) so that the percentage of flagged outliers is equal to 0.1%, which corresponds to the order of magnitude of the theoretical threshold (from $\sim 0.033\%$ to 0.1% for small samples)

The optimum value of $C = 36$ has been determined with the following Monte-Carlo procedure: for a given C, the percentage of flagged outliers is estimated over 100 generations. Each generation is composed of 100 random distributions of the Pearson family, each distribution generates 1000 random samples of size $n = 9$. The total number of points per generation is therefore $100 \times 1000 \times 9 = 9 \times 10^5$.

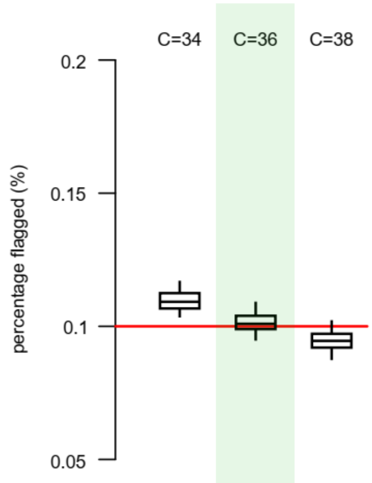

Fig. S1. Impact of the C value on the percentage of outliers flagged in small samples ($n = 9$) with the Boxplot rule using $\alpha = A\log(n) + B + \frac{C}{n}$ with A and B known. Each boxplot has been constructed on 100 generations of 10^5 random samples of size $n = 9$ from the Pearson family.

Part II: the ctbi procedure

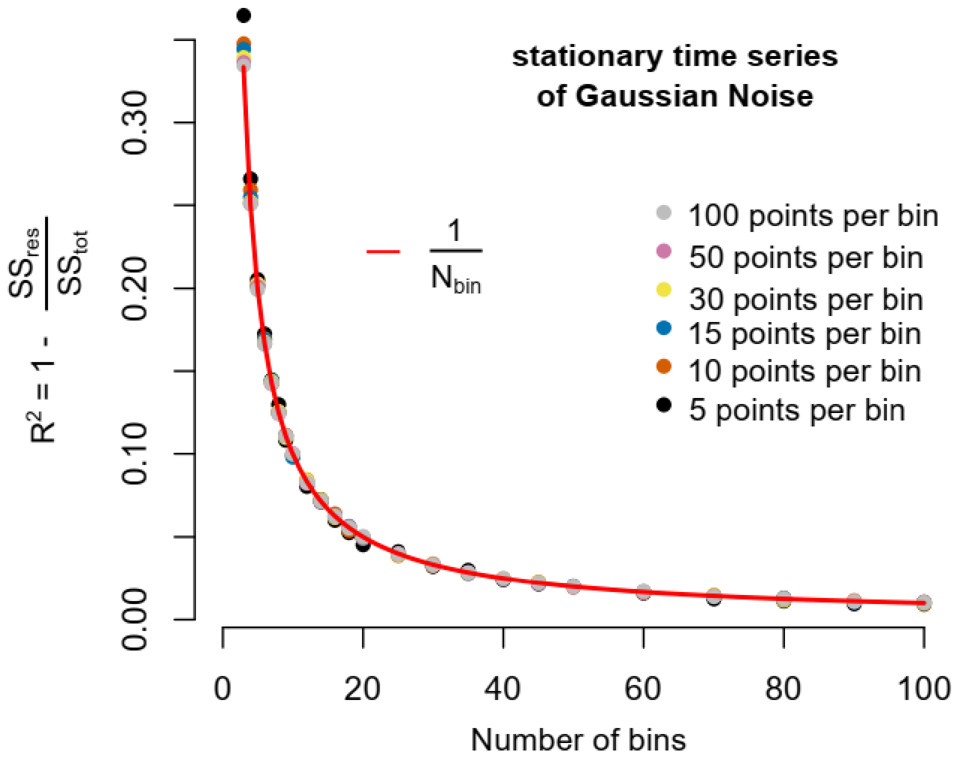

Fig. S2. The coefficient of determination ($R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$) has been calculated for multiple stationary time series of Gaussian noise y, with $SS_{res} = \sum(y_i - S_i)^2$, $SS_{tot} = \sum(y_i)^2$ and S the cyclic component calculated with the **ctbi** procedure.

1) **Definition of the Stacked Cycles Index**
Considering Fig. S2, an inverse relationship appears between the coefficient of determination calculated on a pure Gaussian noise and the number of bins used (related to the sample size). This relationship is independent from the number of points per bin (illustrated by different colors). Theoretically, a stationary timeseries has a null cyclicity ($R^2 = 0$). While this is observed for a large number of bins ($N_{bin} \gg 100$), a bias of $N_{bin}{}^{-1}$ exists at a smaller amount and needs to be corrected. This justifies the definition of the stacked cycles index as $SCI = R^2 - N_{bin}^{-1}$.

2) **The outlier level for the precipitation dataset**
Because daily precipitation data follow a heavy-tailed distribution, it is difficult to determine an outlier level that seems "reasonable" for a 30-year time series. The outlier level is defined as $y_{outlier} = \lambda y_{max}$ and the constant λ is determined using all century-old weather stations. The procedure is the following for a station i:
(i) Compute $(y_{max})_{100\ years}$. An "impossible" event is defined as occurring above $1.2 \times (y_{max})_{100\ years}$ (20% above the century maximum)
(ii) Randomly select 30 continuous years, and compute $(y_{max})_{30\ years}$
(iii) Compute and store $\lambda_i = 1.2 \frac{(y_{max})_{100\ years}}{(y_{max})_{30\ years}}$

The mean value for all stations is $\lambda = 1.6 \pm 0.4$

3) Complex seasonality

Signals showing residuals with non-stationary variance need to be split into several parts of similar variance. This can be achieved using bins with similar MADs. This operation is illustrated on a soil respiration dataset (MIGLIAVACCA) from the COSORE database (Bond-Lamberty et al., 2020).

The MIGLIAVACCA dataset consists of 8 sensors (or "ports") performing measurements of CO_2 flux from 2015 to 2020. Each port performs a measurement every 32 minutes, and there is a 4-minute gap between two successive ports. For each port, the following protocol is applied (Fig. S3):

(i) Apply **ctbi** with the median every day (bin.period = '1 day'), do not flag outliers ($coeff.outlier = NA$) or impute data ($SCI_{min} = NA$).
(ii) Split the raw data into data.low (bins with low MAD) and data.high (bins with high MAD).
(iii) Apply **ctbi** separately to data.low and data.high, and flag outliers ($coeff.outlier = 'auto'$)
(iv) Merge data.low and data.high
(v) Repeat steps (i) to (iv) with bin.period = '1 month'

The comparison between $coeff.outlier = 'auto'$ and $coeff.outlier = NA$ is shown in Fig. **S3**. While obvious periods of instrument failure are still present (in September 2017, August-October 2018 for port 4 or March 2019 for port 1), this procedure proves that most outliers are correctly flagged (all ports were treated independently) when compared to an aggregation without pre-processing ($coeff.outlier = NA$ in Fig. S3).

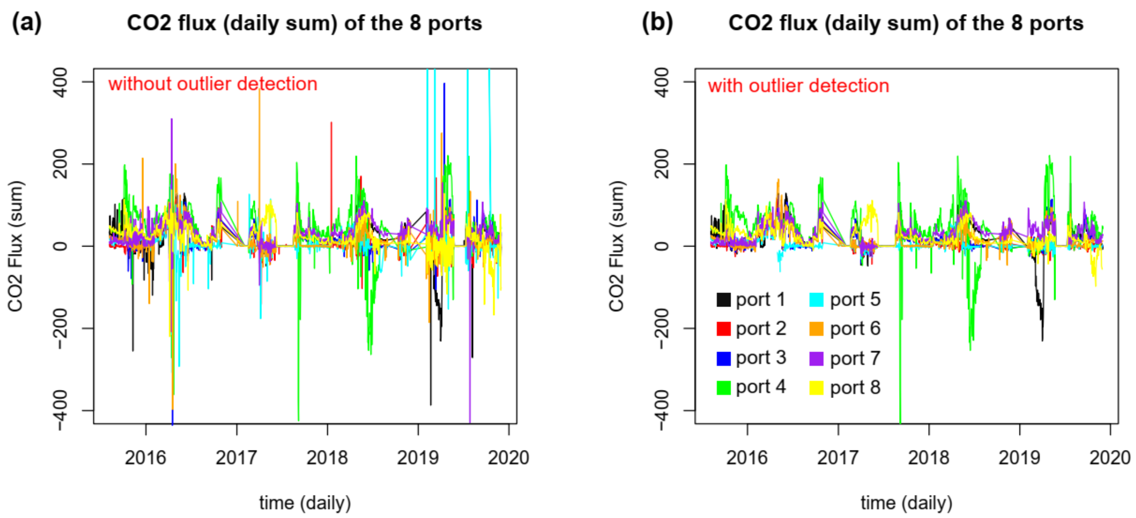

Fig. S3. Soil respiration (daily flux) for the MIGLICAVACCA dataset, with the value of $coeff.outlier = NA$ (no outlier detection, panel **a**) and $coeff.outlier = 'auto'$ (Logbox procedure, panel **b**).

References

Bond-Lamberty, Ben, Danielle S. Christianson, Avni Malhotra, Stephanie C. Pennington, Debjani Sihi, Amir AghaKouchak, Hassan Anjileli et al. "COSORE: A community database for continuous soil respiration and other soil-atmosphere greenhouse gas flux data." *Global change biology* 26, no. 12 (2020): 7268-7283.