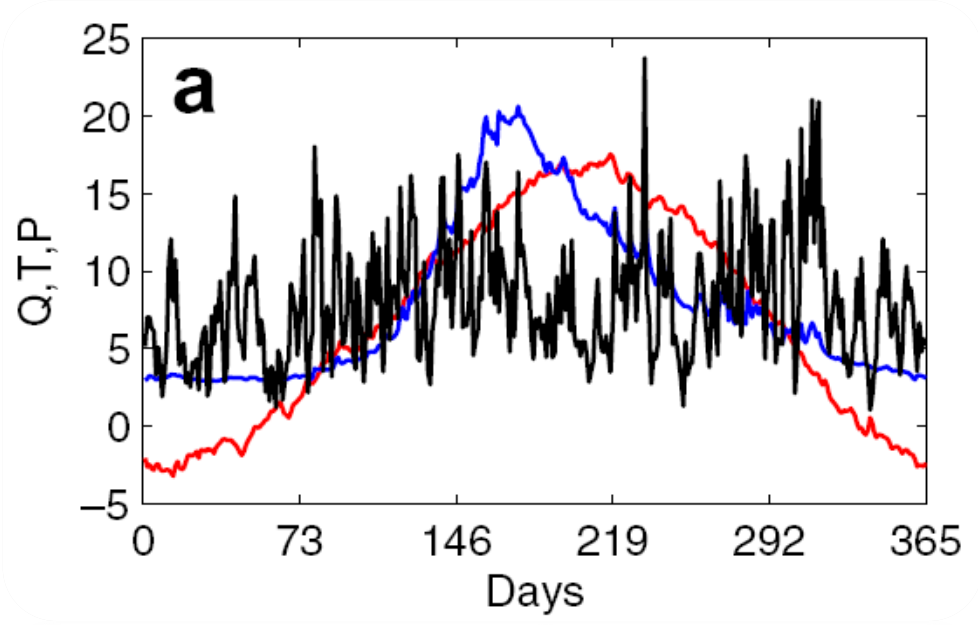
Water Resources Engineering and Management

(CIVIL-466, A.Y. 2024-2025)

5 ETCS, Master course

Prof. P. Perona
Platform of hydraulic constructions



Lecture 6-2: Data analysis, determinism vs stochasticity

Seasonal sample statistics

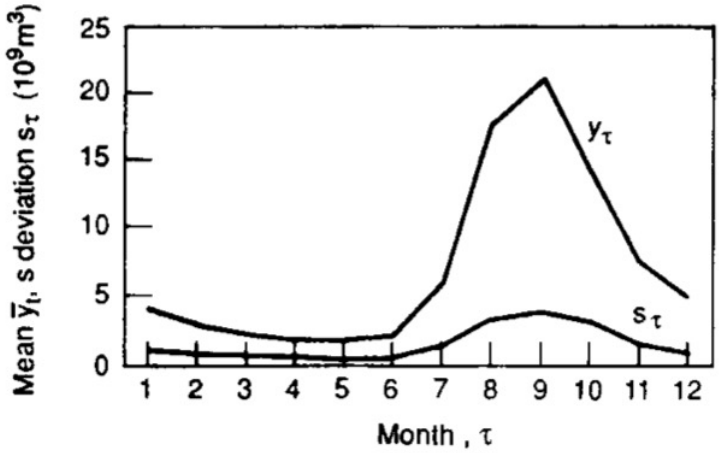
$$\bar{y}_\tau = \frac{1}{N}\sum_{\nu=1}^{N} y_{\nu,\tau} \qquad \tau = 1, \ldots, \omega$$

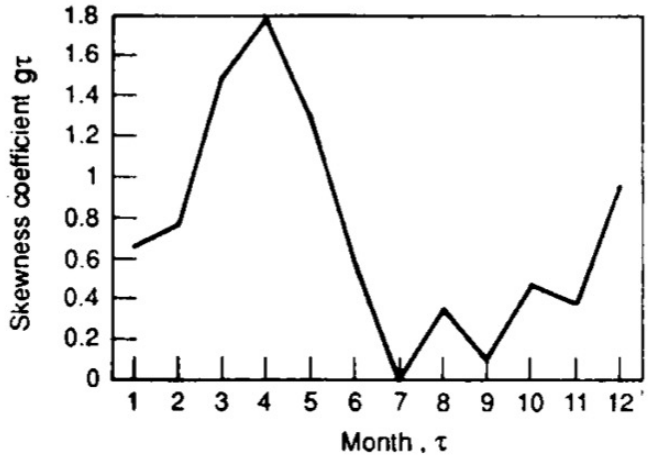
Seasonal mean

ν is the year; τ is the season

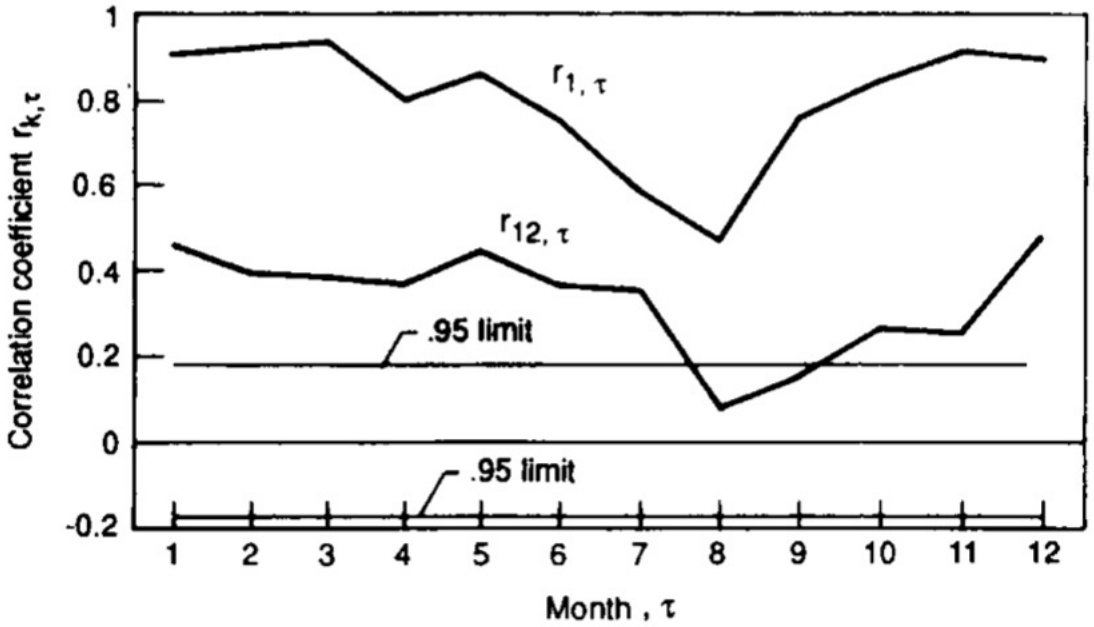
$$r_{k,\tau} = \frac{c_{k,\tau}}{(c_{0,\tau}\, c_{0,\tau-k})^{1/2}}$$

$$c_{k,\tau} = \frac{1}{N}\sum_{\nu=1}^{N} (y_{\nu,\tau} - \bar{y}_\tau)(y_{\nu,\tau-k} - \bar{y}_{\tau-k})$$





Notice: lag 12 monthly autocorrelation is still significant, which indicates the complex temporal relationship between monthly flow discharges of the Nile river basin



Wiener-Khinchine theorem

The link between autocovariance and spectrum is given by the Wiener–Khinchin theorem;

$$\int_{-\infty}^{\infty} \rho(\tau) e^{i\omega\tau} \mathrm{d}\tau = A^2(\omega), \quad \longleftrightarrow \quad \int_{-\infty}^{\infty} A^2(\omega) e^{-i\omega\tau} \mathrm{d}\omega = \rho(\tau),$$

Autocovariance function

Energy spectrum

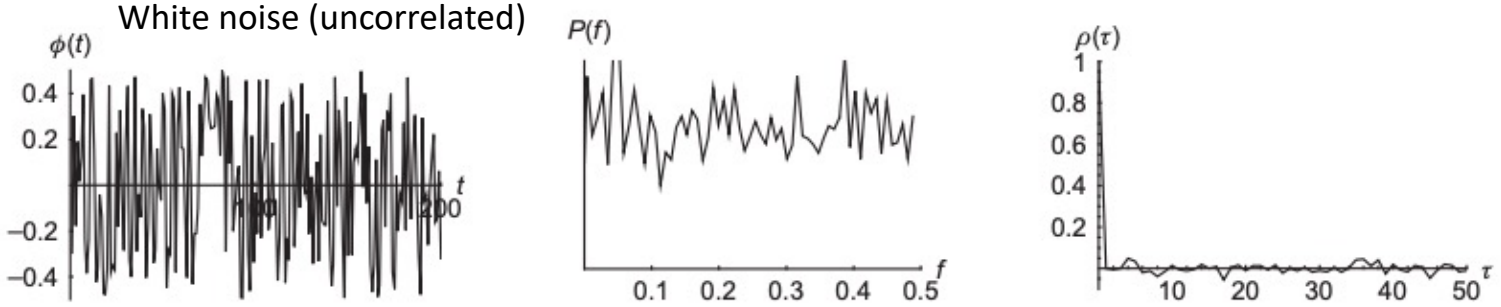
or, in terms of autocorrelation

$$\int_{-\infty}^{\infty} \overline{\rho}(\tau) e^{i\omega\tau} \mathrm{d}\tau = P(\omega), \quad \longleftrightarrow \quad \int_{-\infty}^{\infty} P(\omega) e^{-i\omega\tau} \mathrm{d}\omega = \overline{\rho}(\tau).$$

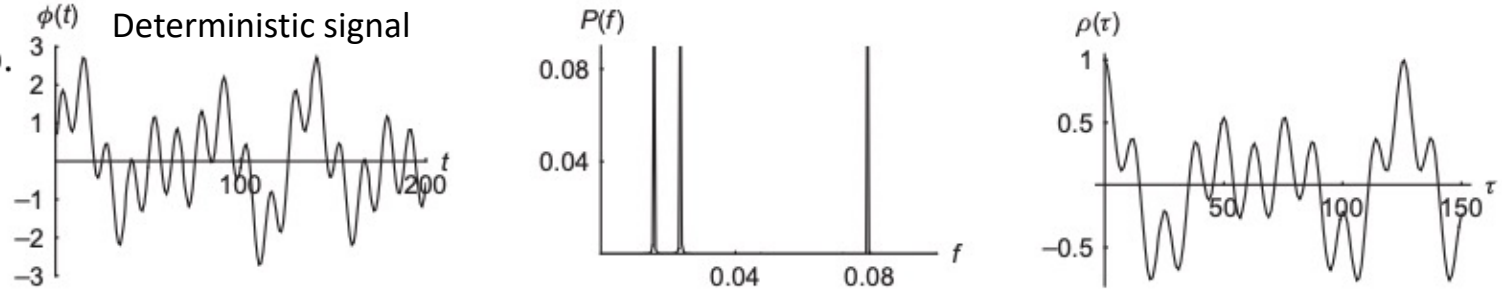
Autocorrelation function

Power spectrum

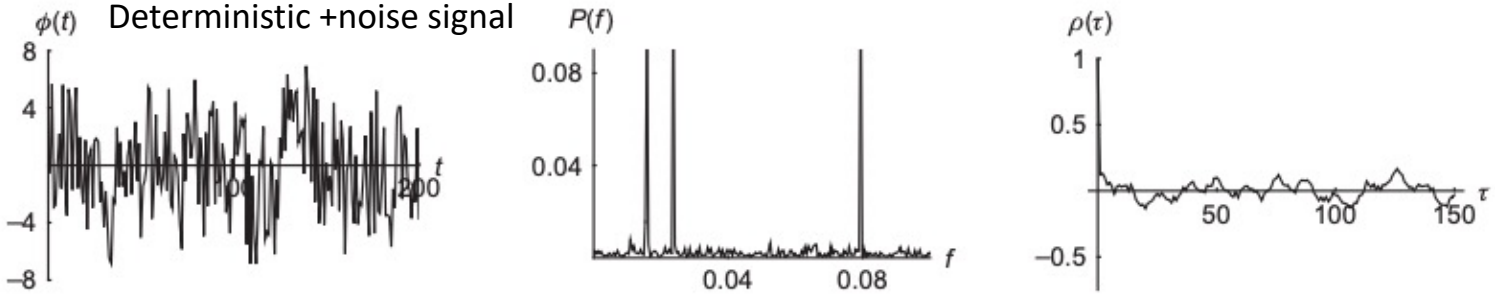
White noise (uncorrelated)



Deterministic signal



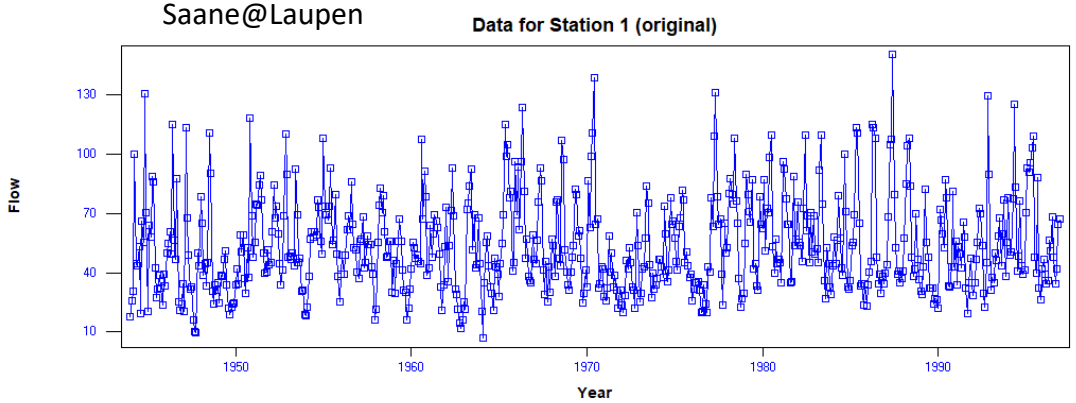
Deterministic +noise signal

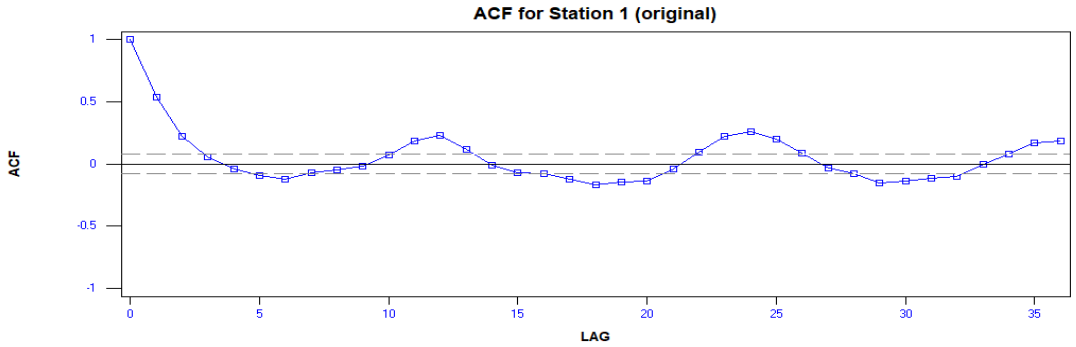


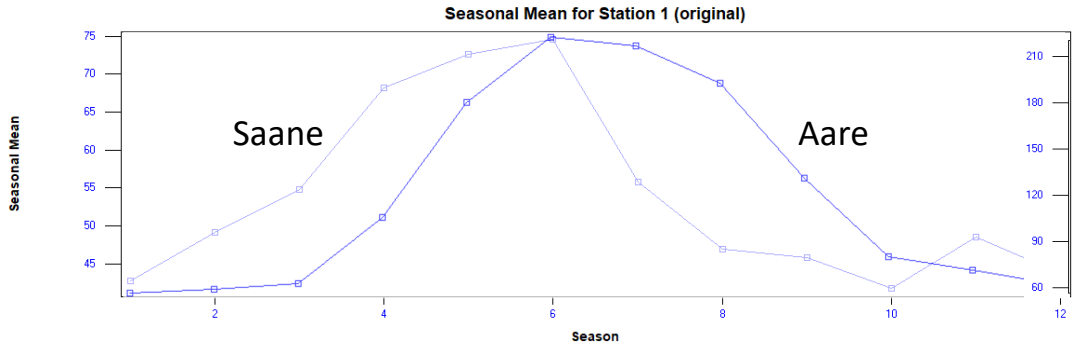
Source: Ridolfi et al., 2011

Example: Monthly discharges

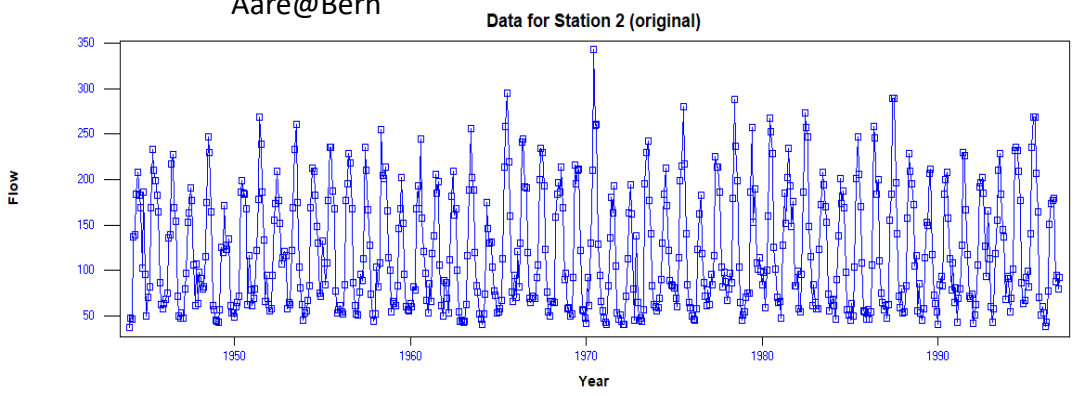
Saane@Laupen

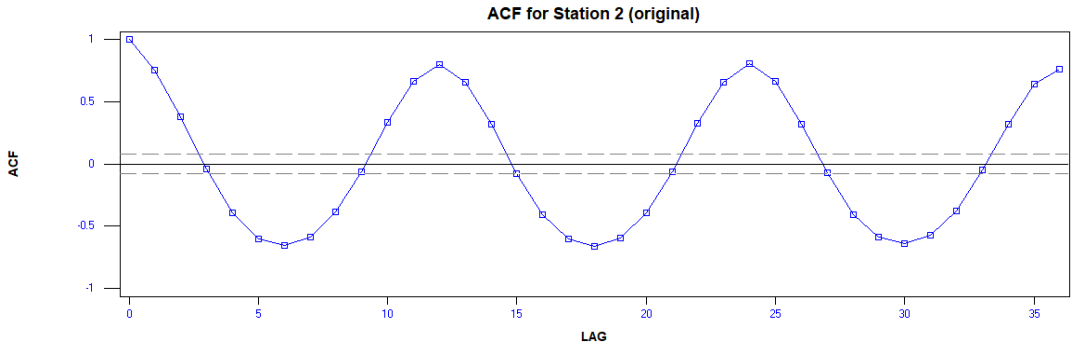


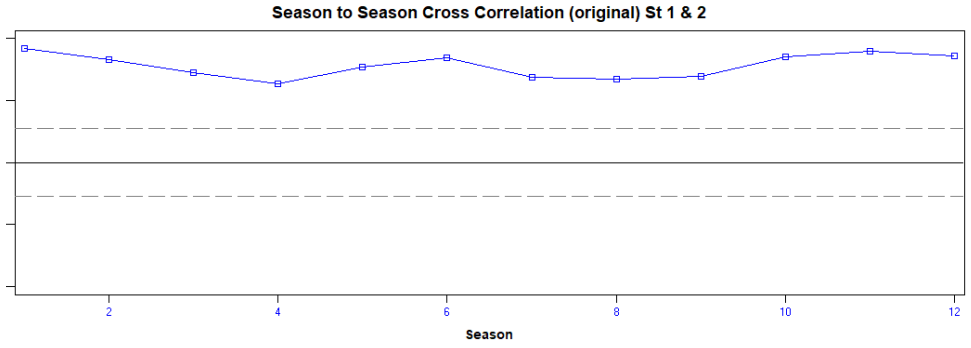




Aare@Bern

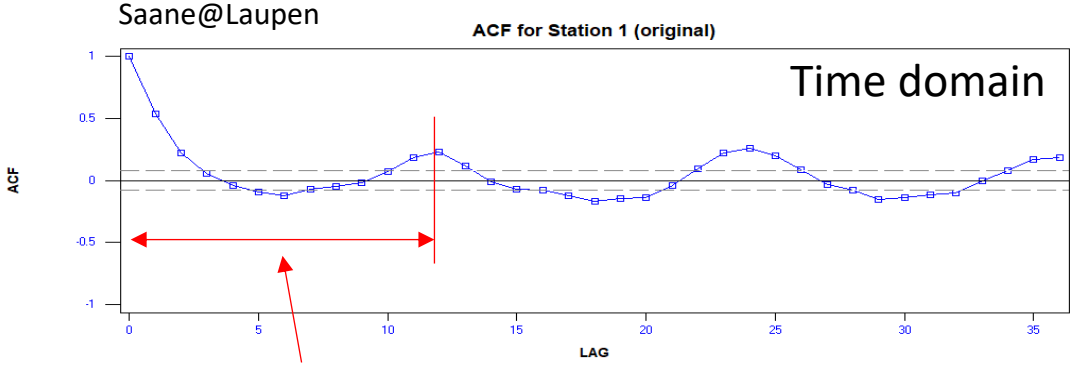




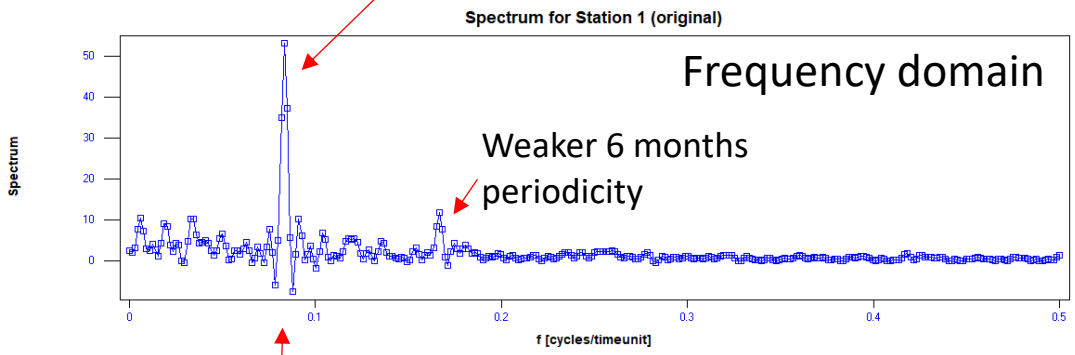


Power spectrum vs autocorrelation

Saane@Laupen

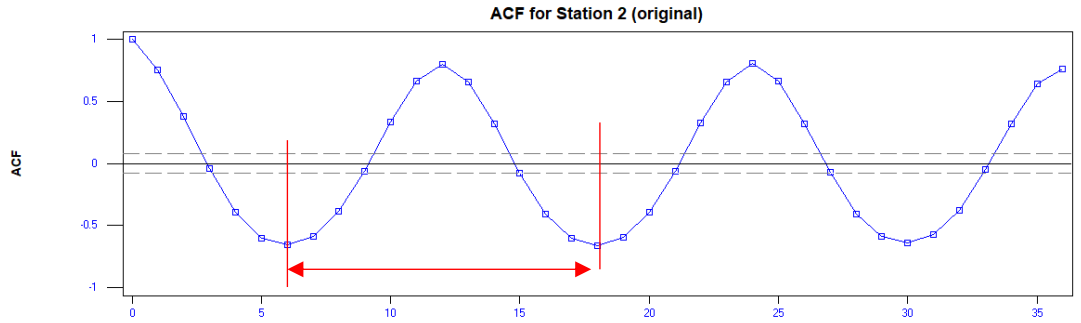


12 months lag (periodicity)

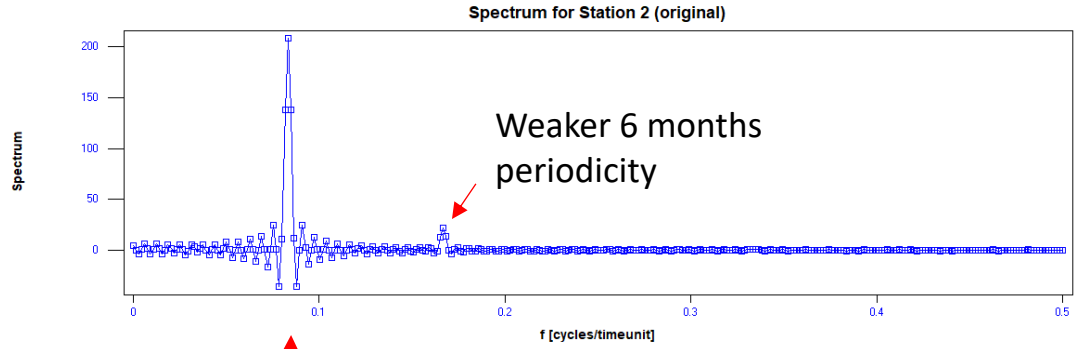


$f_{main}=1/12=0.083\ y^{-1}$

Aare@Bern



12 months lag (periodicity)

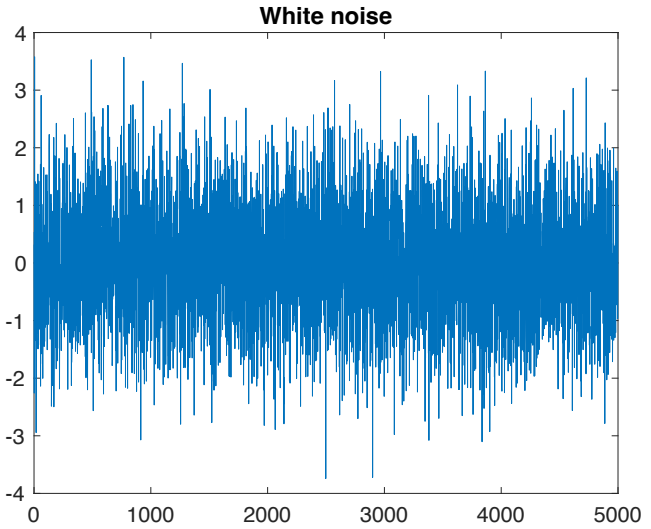


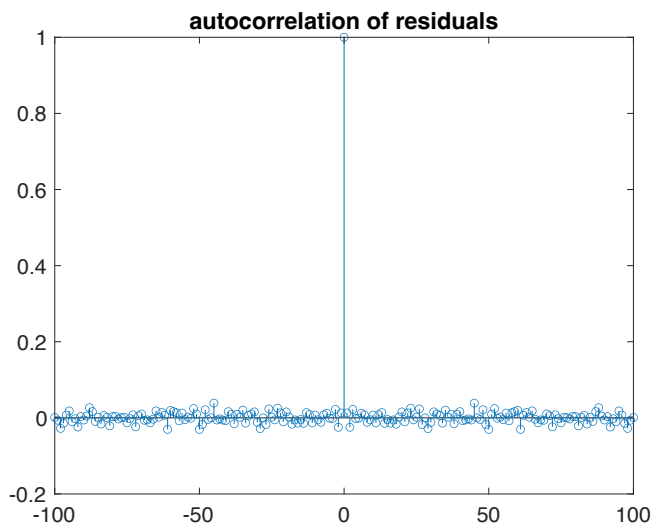
$f_{main}=1/12=0.083\ y^{-1}$

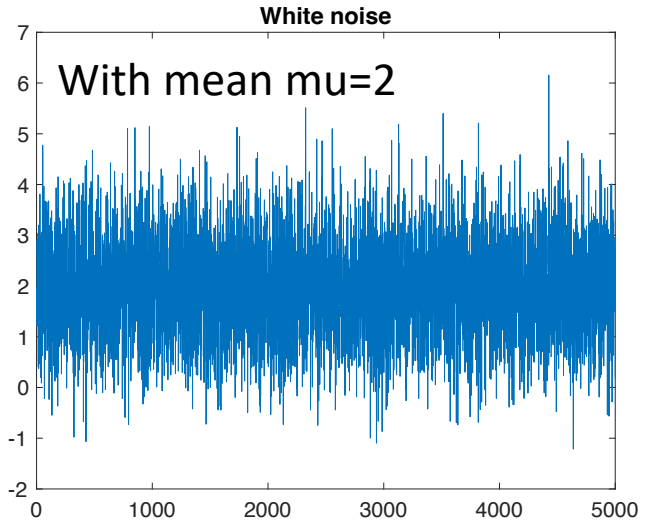
More about autocorrelation

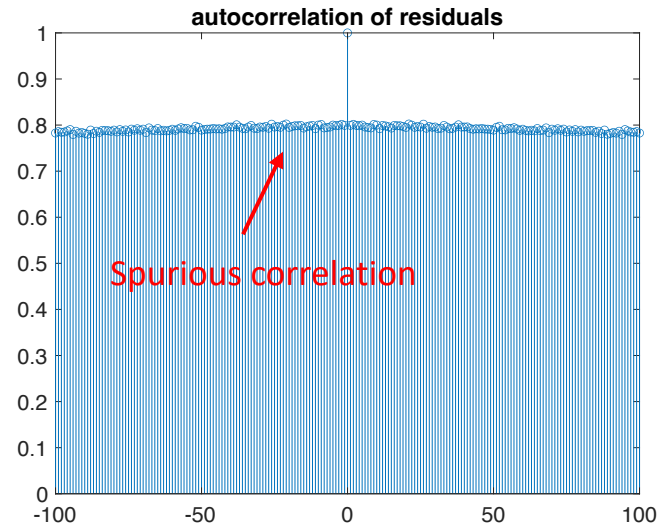
Always remove the mean of the signal when calculating ACF. Not removing the mean will introduce a spurious correlation, which by no means should be interpreted as long term memory of the process!

Ex. See white-noise signal aside









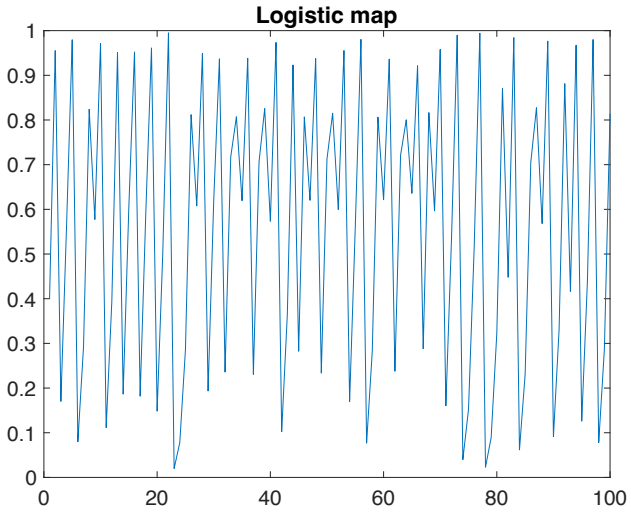
More about autocorrelation

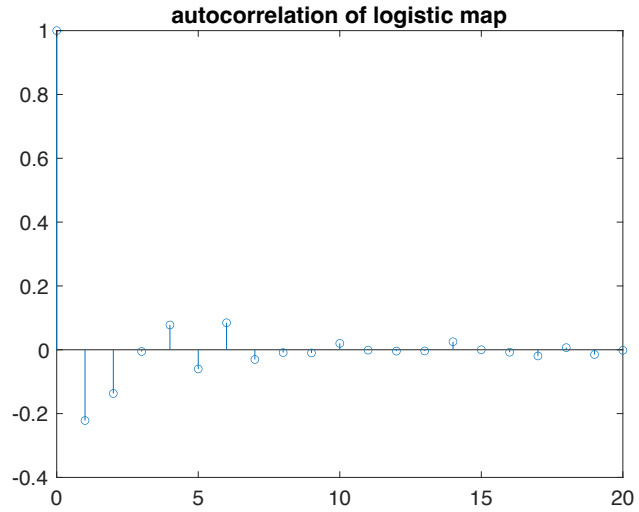
1. ACF Captures Only Linear Dependence

•The **ACF measures linear correlations** between time-lagged values.

•However, some processes exhibit **nonlinear dependencies** that the ACF cannot detect.

•Example: A **chaotic system** can have low or zero autocorrelation but still exhibit long-term memory through deterministic structures. The system is still unpredictable due to sensitivity to initial conditions.





2. ACF Does Not Always Capture Long-Range Dependence

•Some processes (e.g., **fractional Brownian motion (fBm)** with H>0.5) have **long-range dependence** where past values influence the future over very long time horizons.

•ACF might decay slowly in these cases, but other methods (e.g., the **Hurst exponent**) are needed to properly characterize the memory.

- Stochastic processes may show H>0.5 at short range and asymptotically yet tend to H=0.5. This means that the Hurst analysis associating short or long range memory is a meaningful concept only in an asymptotic sense.

Autocorrelation for stationary and non-stationary processes

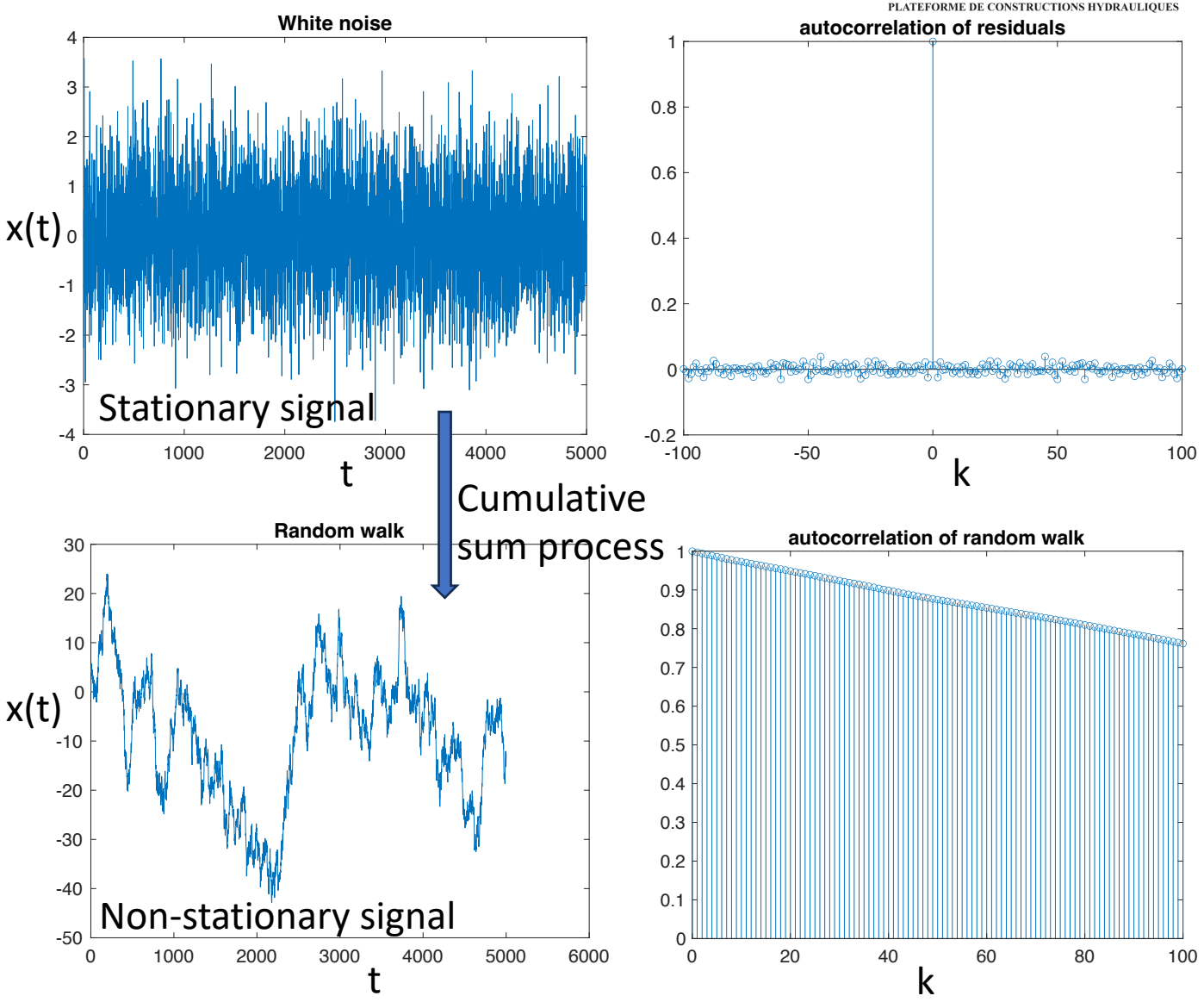
PLATEFORME DE CONSTRUCTIONS HYDRAULIQUES

3. Stationary vs. Non-Stationary Processes

- **For stationary processes**, ACF is more useful because memory effects do not change over time.
- **For non-stationary processes** (like a random walk), the ACF does not decay in a typical way, making it misleading as a memory measure.
- Alternative methods like **variance analysis, rescaled range (R/S) analysis (or Hurst), or DFA (Detrended Fluctuation Analysis)** are needed to assess true memory.

4. The Case of the Random Walk

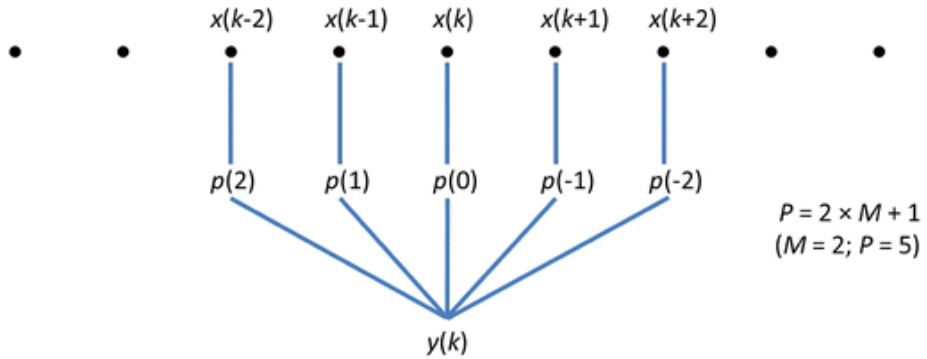
- A **random walk** has **strong autocorrelation in levels**, but **zero autocorrelation in differences (returns)**.
- This means that while past values influence present ones (persistence in levels), **there is no real long-term memory** in the sense of long-range dependence.



Other ways of removing trends and patterns

Moving average smoothing

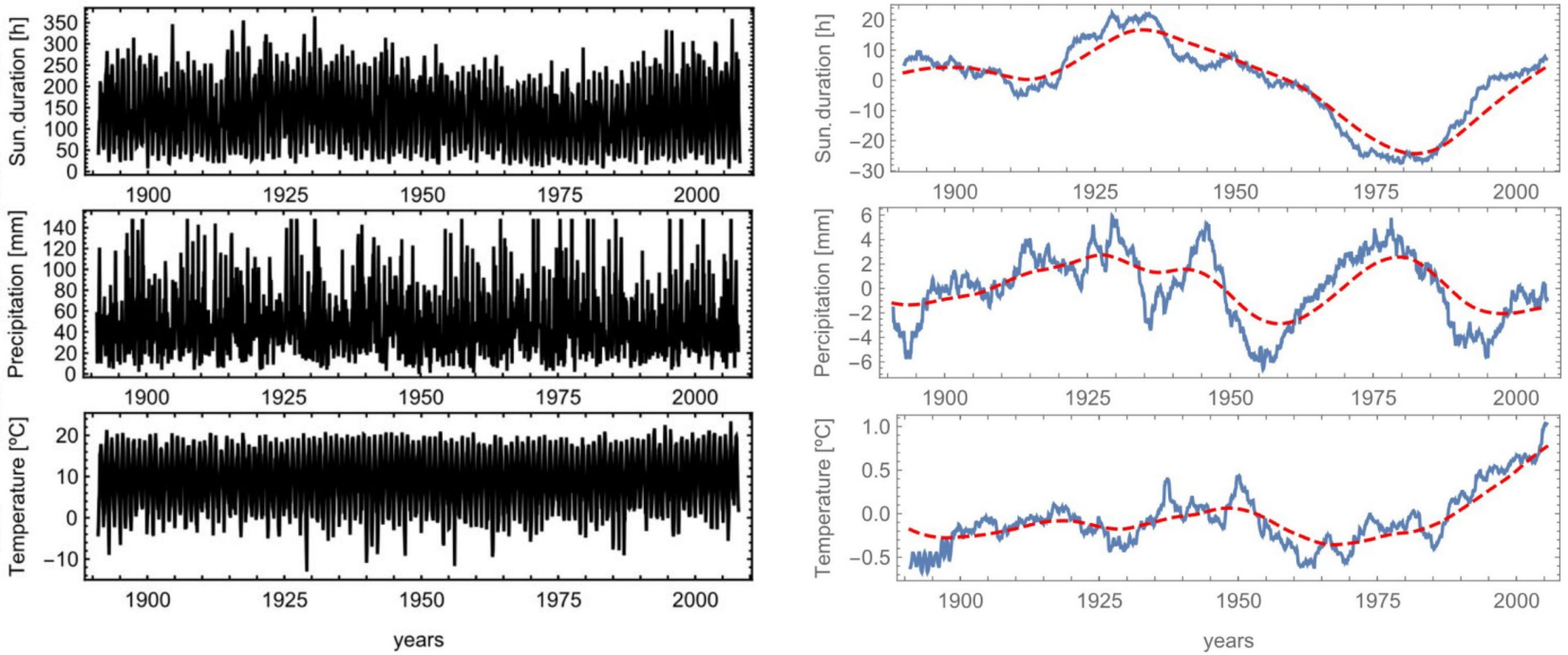
The moving average is a smoothing operation (low-pass filter) which can be used to extract the trend underlying time series. The weighted moving average is a mathematical mean centered (symmetric or non) and operated over a number of points before and after the centre one. Then the centre is moved and the mean recalculated (it results a sort of convolution). MA may introduce a lag shift in the smoothed data



$$y_k = \frac{1}{\sum p_m} \sum_{m=-M}^{M} p_m x_{k+m}$$

Other types of moving average exist, i.e. when all p_m=1, simple MA (cfr voice comments).

$$y_k = \frac{1}{2M+1} \sum_{m=-M}^{M} x_{k+m}$$



Source: Glogowski et al, 2021

Exponential smoothing

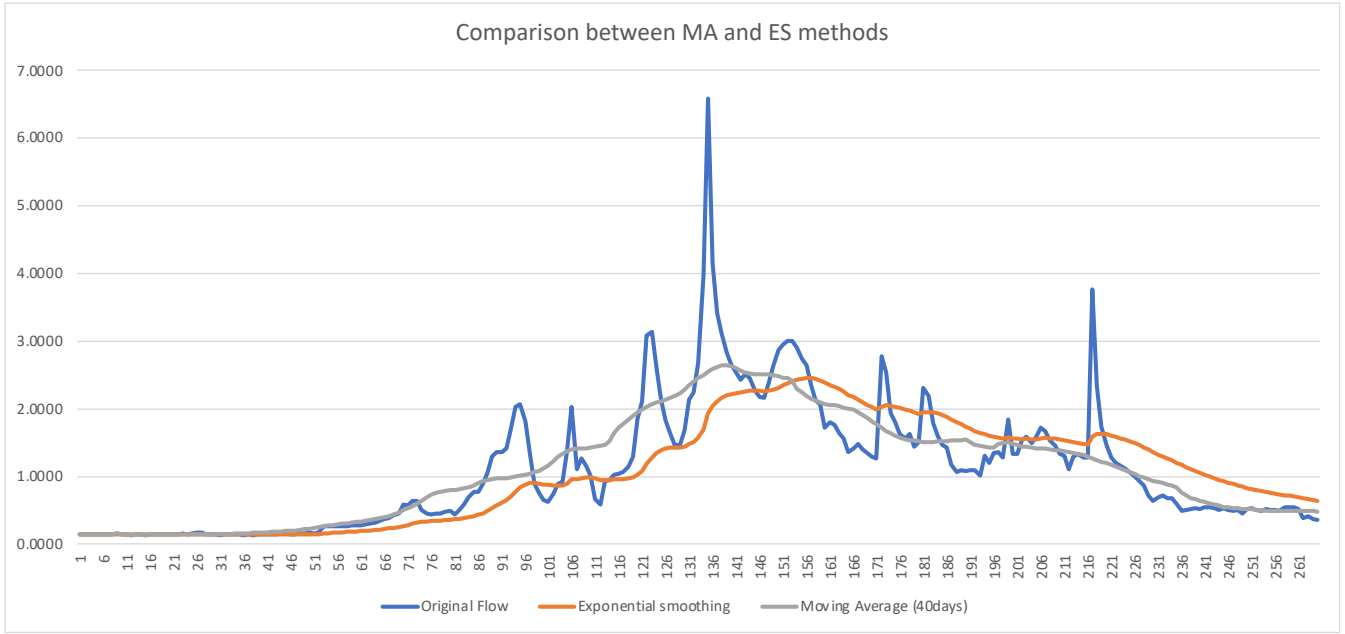
The exponential smoothing is a technique for smoothing fluctuations in a time series based on a weighted average of past values, where the weighting functions are indeed exponential functions of the lag.

$$s_0 = x_0$$
$$s_t = \alpha x_t + (1-\alpha) s_{t-1} \qquad \text{for } t>0$$

$0 \leq \alpha \leq 1$, smoothing factor

The same relationship can also be rewritten as

$$s_0 = x_0$$
$$s_t = s_{t-1} + \alpha(x_t - s_{t-1}) \qquad \text{for } t>0$$



As for the moving average, the exponential smoothing will introduce a lag shift in the data, whose correction is not obvious (cfr., voice comments)

Correction for bias

An estimator of a population statistics is biased when its sample value is different from the population value. Several corrections have been proposed. Such methods often use a function of the sample size N as correction factor

$$\hat{\rho}_1 = \frac{r_1 N + 1}{N - 4}$$

For autocorrelated data ($r_k \neq 0$), the bias is downward, i.e. sample statistics are smaller than the population ones.

$$\hat{\sigma}^2 = \frac{(N-1)\, s^2}{N - K}$$

For uncorrelated data, the sample variance is an unbiased estimator of the population variance. Otherwise, the variance can be corrected as stated

$$K = \frac{[N(1-\hat{\rho}_1^2) - 2\,\hat{\rho}_1\,(1-\hat{\rho}_1^N)]}{[N(1-\hat{\rho}_1)^2]}$$

Correction for the sample variance

Statistical tests and transformation to normal

Statistical tests exist in both parametric and non-parametric forms. In hydrology some important tests are

- Determining the significance of trends, e.g. t-student with linear regression model, Mann-Kendall non parametric
- Determining the significance of shift or jumps, e.g., t-Test for shift in the mean, Mann-Whitney
- Determining the presence of seasonality, t- test or the extended one-way analysis of variance
- Determining the degree of normality, e.g. the Chi-square, the Kolmogorov-Smirnov test

A widely used technique in hydrology to obtain normally distribute data is through variable transformation. Two most popular ones are:

- Logarithmic transformation of a lognormally distributed data x_t

$$y_t = \log (x_t - c)$$

- Power transformation

$$y_t = (x_t - c)^b$$

- Box-Cox transformation

$$y_t = \begin{cases} \dfrac{(x_t^\lambda - 1)}{\lambda} & \lambda \neq 0 \\ \ln (x_t) & \lambda = 0 \end{cases}$$

Take home message from these three lectures

PLATEFORME DE CONSTRUCTIONS HYDRAULIQUES

- L6.1 I can explain the differences between the the type of data sequences (e.g., continuous, discrete, etc.)
- L6.1 I understand and can explain the effects of shifts, trends, periodicity, etc.
- L6.1 I understand the meaning of serial and cross correlation in data sequences

- L6.2 I know how to calculate basic moment statistics for the sample and the population
- L6.2 I know how to calculate the serial (auto)correlation of data sequences and understand its meaning
- L6.2 I know how to standardize a time series and what does this mean
- L6.2 I can explain how partitioning works and I know what to do at each step

- L6.3 I know how to calculate seasonal sample statistics
- L6.3 I can explain (but do not need to remember) the relationship between time and frequency domains
- L6.3 I know how to perform a moving average and can write the general formula
- L6.3 I know how to perform and write the relationship of the exponential smoothing
- L6.3 I understood (but not need to remember) formulas to correct for bias and transform to normal a given data series