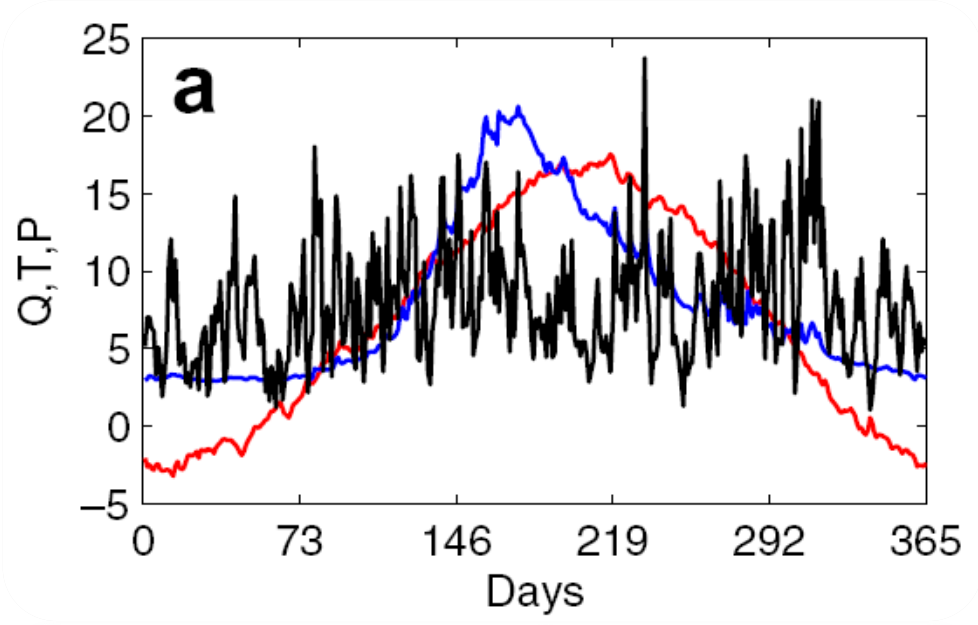
Water Resources Engineering and Management

(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

Statistical properties of time series

OVERALL SAMPLE STATISTICS

$$\bar{y} = \left(\frac{1}{N}\right) \sum_{t=1}^{N} y_t$$

Sample aritmetic mean, with N being the sample size

$$s^2 = \frac{1}{N-1} \sum_{t=1}^{N} (y_t - \bar{y})^2$$

Sample variance

$$c_v = s/\bar{y}$$

Coefficient of variation

NOTICE: the sample statistics are estimators of the population statistics μ, σ^2, γ and ρ_k

$$g = \frac{N \sum_{t=1}^{N} (y_t - \bar{y})^3}{(N-1)(N-2)s^3}$$

Skewness coefficient

$$r_k = \frac{c_k}{c_0}$$

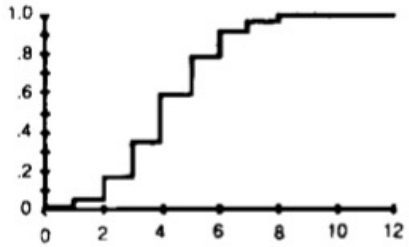
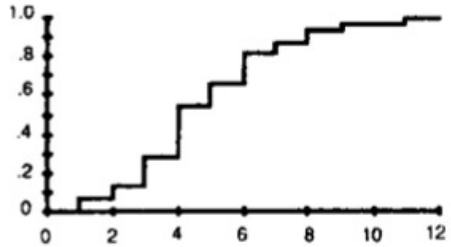
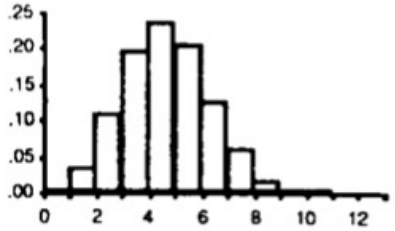
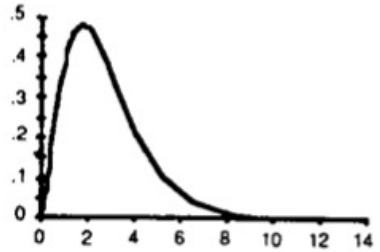
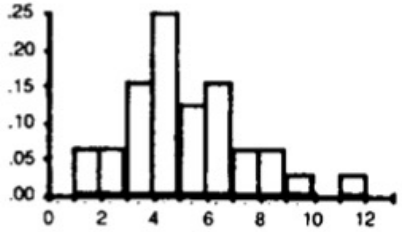
Sample autocorrelation function

$$c_k = \left(\frac{1}{N}\right) \sum_{t=1}^{N-k} (y_{t+k} - \bar{y})(y_t - \bar{y}) \qquad k \geq 0$$

Sample autocorrelation coefficient at lag k

Source: Handbook of hyrology, Maidment

Parameters of statistical populations and samples

Concept	Population value, discrete case	Population value, continuous case	Sample value
Cumulative distribution function (cdf)	 Describes the probability that a random variable is less than or equal to a specified value x	 Describes the probability that a random variable is less than or equal to a specified value x	 Empirical distribution function (edf): describes the observed frequency of a random variable being less than or equal to a specified value x
Probability mass function (pmf) and probability density function (pdf)	 pmf: the probability that X is equal to k	 pdf: first derivative of the cumulative distribution function $f(x) = \frac{dF(x)}{dx}$	 Histogram: observed frequency with which random variable X falls into the assigned ranges
Mean, average, or expected value	$\mu = \sum_{i=1}^{\infty} P(X = x_i)x_i$	$\mu = \int_{-\infty}^{\infty} xf(x)\,dx$	$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$

Source: Handbook of hydrology, Maidment

Parameters of statistical populations and samples

Variance	$\sigma^2 \equiv \sum_{i=1}^{\infty} P(X = x_i)(x_i - \mu)^2$	$\sigma^2 \equiv \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx$	$S^2 \equiv \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$
kth central moment	$M_k \equiv \sum_{i=1}^{\infty} P(X = x_i)(x_i - \mu)^k$	$M_k \equiv \int_{-\infty}^{\infty} (x - \mu)^k f(x)\, dx$	$\tilde{M}_k \equiv \sum_{i=1}^{n} \frac{(X_i - \bar{X})^k}{n}$
Standard deviation	$\sigma \equiv \sqrt{\sigma^2}$		$S \equiv \sqrt{S^2}$
Coefficient of variation or relative standard deviation (if $\mu \neq 0$)	$\mathrm{CV} \equiv \frac{\sigma}{\mu}$		$\mathrm{CV} \equiv \frac{S}{\bar{X}}$
Coefficient of skew (a measure of asymmetry)	$\gamma \equiv \frac{M_3}{\sigma^3}$		$G \equiv \frac{\tilde{M}_3}{S^3}$
Quantiles	x_p is any value of X that has the properties that $P(X < x_p) \leq p$, $P[X > x_p] \leq 1 - p$		$\hat{X}_p$ is the pth quantile of EDF
Median (useful for describing central tendency regardless of skewness)	$x_{0.5}$ Any value of X that has the property that $P[X < x_p] \leq 0.5$, $P[X > x_p] \leq 0.5$		$\hat{X}_{0.5}$ The middle observation in a sorted sample, or the average of the two middle observations if the sample size is even.
Upper quartile, lower quartile, and hinges	Upper quartile $\equiv x_{0.75}$, Lower quartile $\equiv x_{0.25}$		Upper hinge $\equiv \hat{X}_{0.75}$ This is an approximation to the sample upper quartile; it is defined as the median of all sample values of $X \geq x_{0.50}$. The lower hinge, $\hat{X}_{0.25}$, is defined analogously.
Interquartile range (useful for describing spread of data regardless of symmetry)	$x_{0.75} - x_{0.25}$ Width of central region of population containing probability of 0.5		$\hat{X}_{0.75} - \hat{X}_{0.25}$ Width of central region of data set encompassing approximately half the data

Source: Handbook of hyrology, Maidment

Calculate autocorrelation or serial correlation

$$r_k = \frac{c_k}{c_0}$$

$$c_k = \left(\frac{1}{N}\right) \sum_{t=1}^{N-k} (y_{t+k} - \bar{y})(y_t - \bar{y}) \qquad k \geq 0$$

The autocorrelation tells how much of the variance of the data at a given time t is explained by the variance of data at previous lags

The sequence of r_k for varying k is called the autocorrelation function

Purely random data (e.g., white noise) are serially uncorrelated and the autocorrelation function is a Dirac-delta distribution, i.e. c_k=1 only per k=0 and c_k=0 per k≠0

k=0	
x(t)	x(t-0)
0.5	0.5
0.52203365	0.52203365
-0.1808619	-0.1808619
-0.0761155	-0.0761155
-0.9058361	-0.9058361
-1.1259897	-1.1259897
-1.8979435	-1.8979435
-1.9881003	-1.9881003
-1.1741327	-1.1741327
-0.4695902	-0.4695902
0.08045628	0.08045628
-0.46581	-0.46581
-0.4449673	-0.4449673
-1.0086056	-1.0086056
-0.2618347	-0.2618347
0.33297074	0.33297074
-0.130812	-0.130812
-0.5008106	-0.5008106
0.4618072	0.4618072
0.70869675	0.70869675
-0.3454617	-0.3454617

c_0=0.56

r_k=r_0=1

k=1	
x(t)	x(t-1)
0.5	
0.52203365	0.5
-0.1808619	0.52203365
-0.0761155	-0.1808619
-0.9058361	-0.0761155
-1.1259897	-0.9058361
-1.8979435	-1.1259897
-1.9881003	-1.8979435
-1.1741327	-1.9881003
-0.4695902	-1.1741327
0.08045628	-0.4695902
-0.46581	0.08045628
-0.4449673	-0.46581
-1.0086056	-0.4449673
-0.2618347	-1.0086056
0.33297074	-0.2618347
-0.130812	0.33297074
-0.5008106	-0.130812
0.4618072	-0.5008106
0.70869675	0.4618072
-0.3454617	0.70869675

c_0=0.544

r_k=r_1=0.664

k=2	
x(t)	x(t-2)
0.5	
0.52203365	
-0.1808619	0.5
-0.0761155	0.52203365
-0.9058361	-0.1808619
-1.1259897	-0.0761155
-1.8979435	-0.9058361
-1.9881003	-1.1259897
-1.1741327	-1.8979435
-0.4695902	-1.9881003
0.08045628	-1.1741327
-0.46581	-0.4695902
-0.4449673	0.08045628
-1.0086056	-0.46581
-0.2618347	-0.4449673
0.33297074	-1.0086056
-0.130812	-0.2618347
-0.5008106	0.33297074
0.4618072	-0.130812
0.70869675	-0.5008106
-0.3454617	0.4618072

c_0=0.52

r_k=r_1=0.265

Meaning of autocorrelation

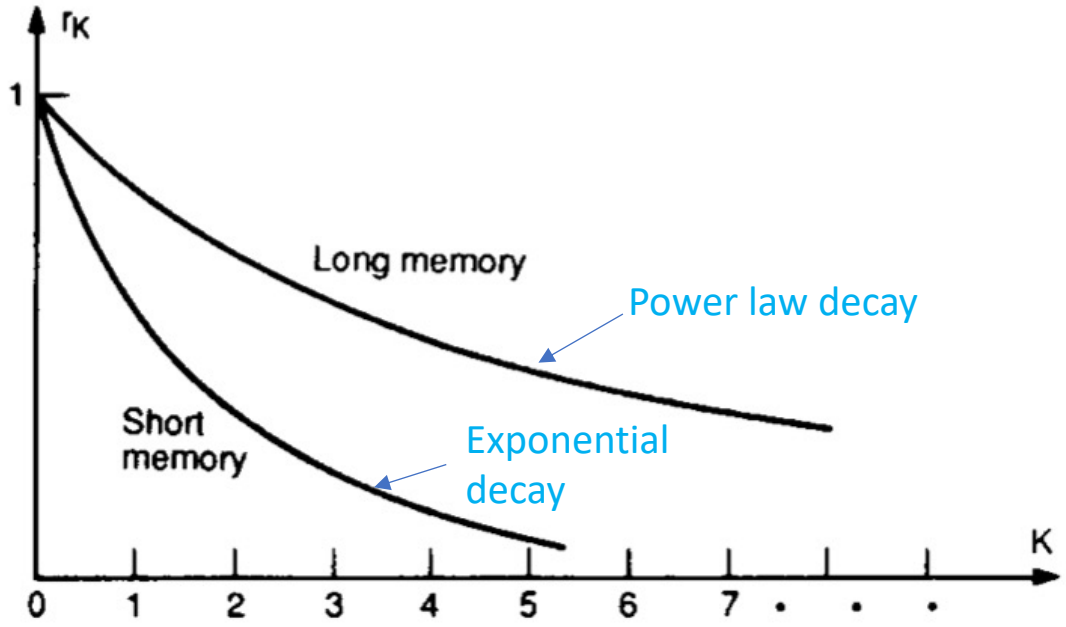


FIGURE 19.2.1 Schematic representation of a correlogram with short and long memory.

Time series with power law decaying serial correlation exibit long-term memory, where time series with faster decaying (e.g., exponential) serial correlation show only short-term memory or zero (if pure noise)

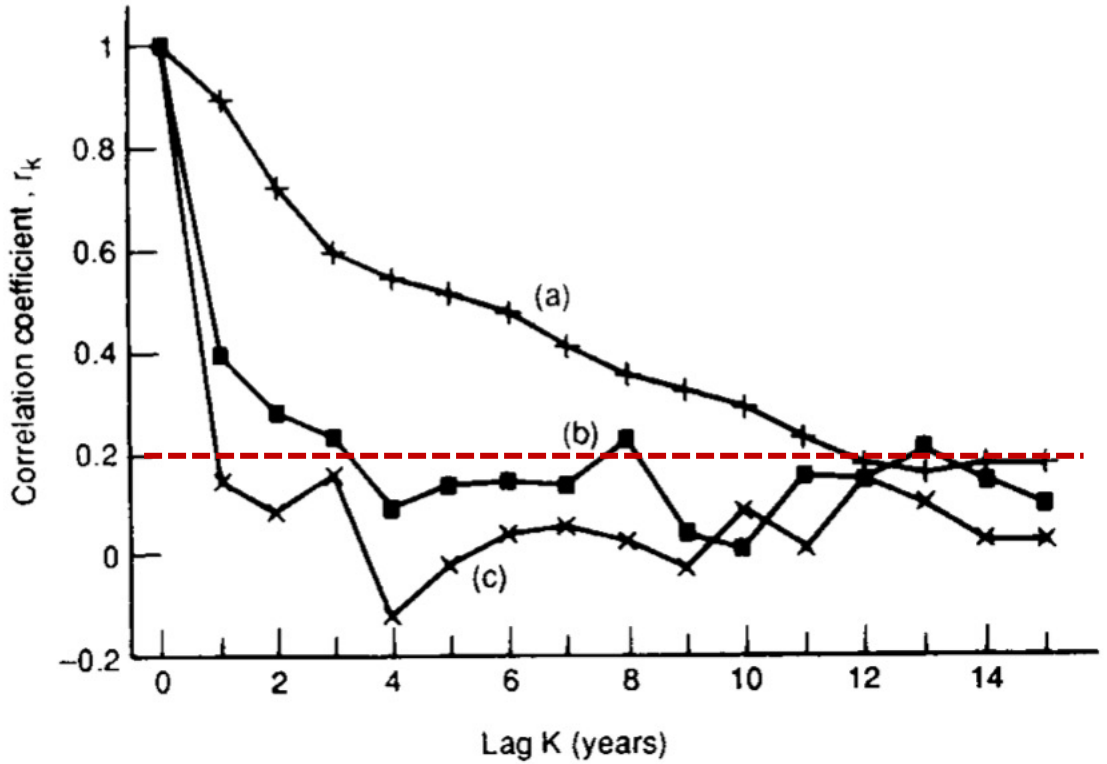
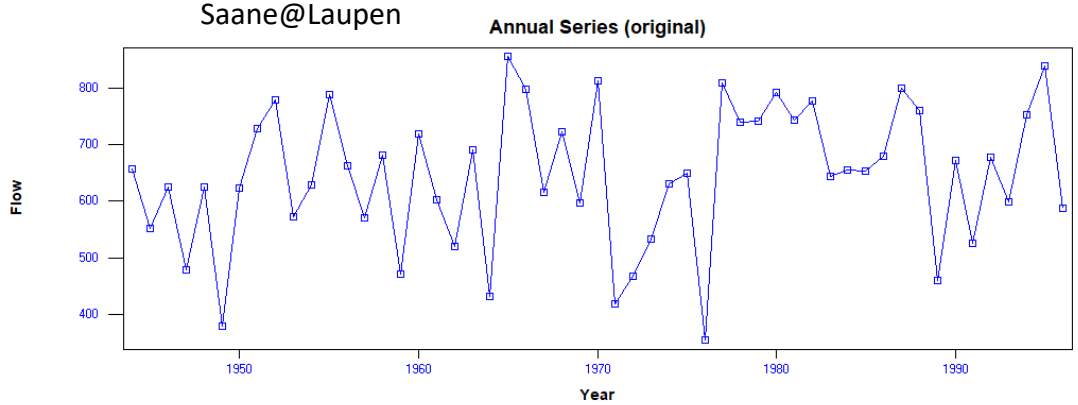
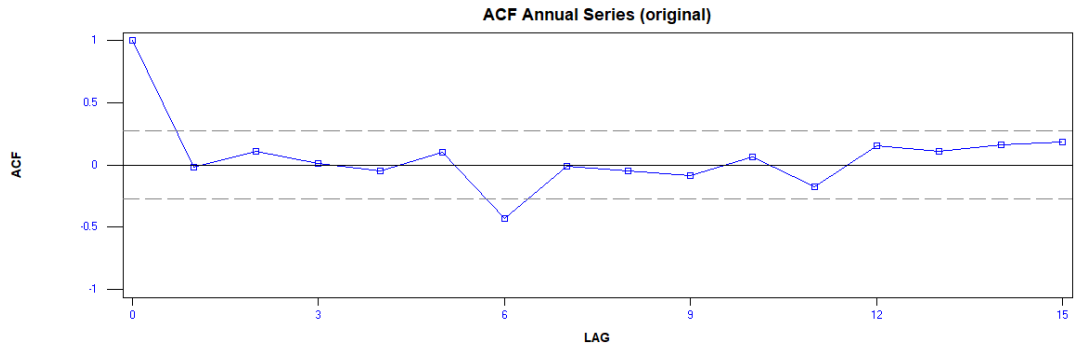


FIGURE 19.2.2 Correlogram of annual flows of (*a*) the White Nile River at Mongalla (1914–1983), (*b*) the Nile River at Aswan (1871–1989), and (*c*) the Blue Nile River at Khartoum (1912–1989).

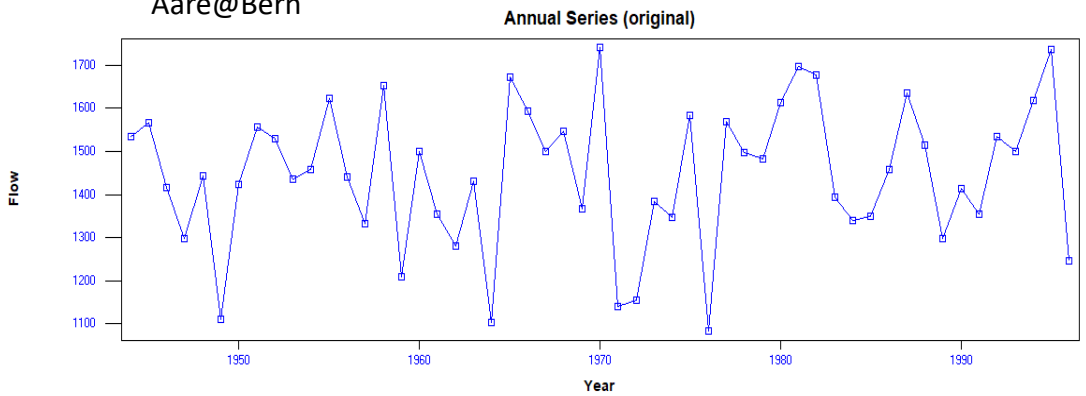
Examples: Annual discharges

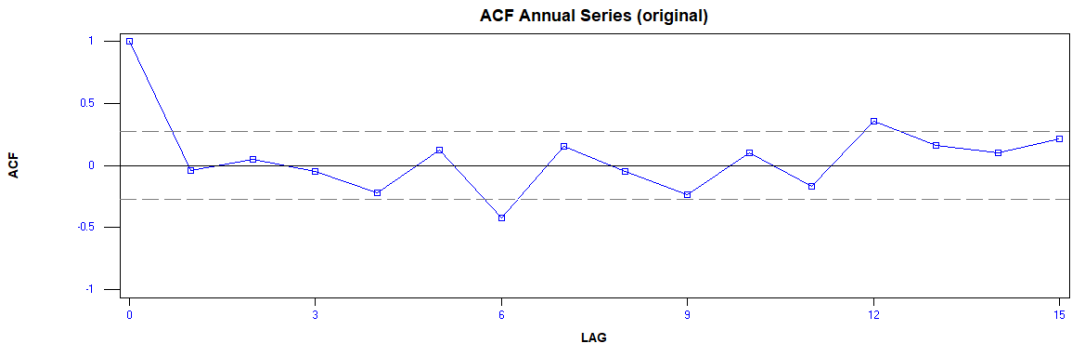
Saane@Laupen





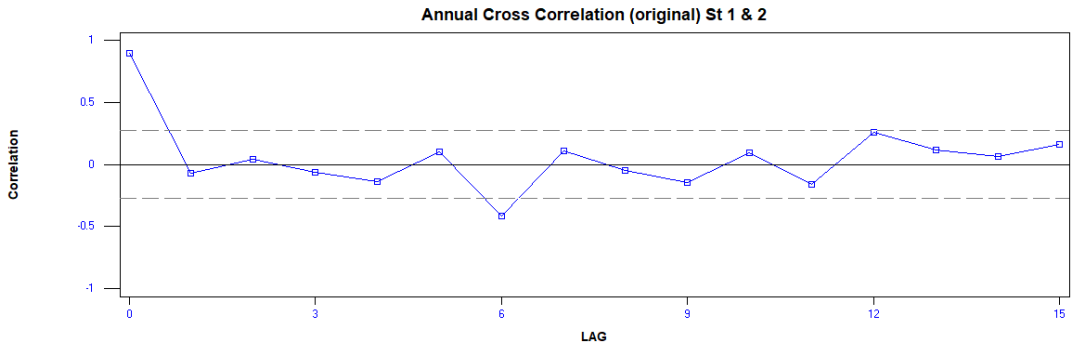
Aare@Bern





Station 1: SAANE_@_LAUPEN
Station 2: AARE_@_BERN

St_NUM	Mean	StDev	CV	Skewness	Min	Max	acf(1)	acf(2)
ST(1):	644.7	122.0	0.1893	-0.4261	354.4	856.8	-0.0174	0.1074
ST(2):	1448.	163.9	0.1132	-0.3880	1082.	1742.	-0.0426	0.0465



Partitioning of the time-series structure

Hydrologic time series, in various degrees, exhibit trends, shifts, jumps, seasonality, autocorrelation and non-normality.
These attributes are referred to as components.
Therefore, a time series can be decomposed (partitioned) into his components

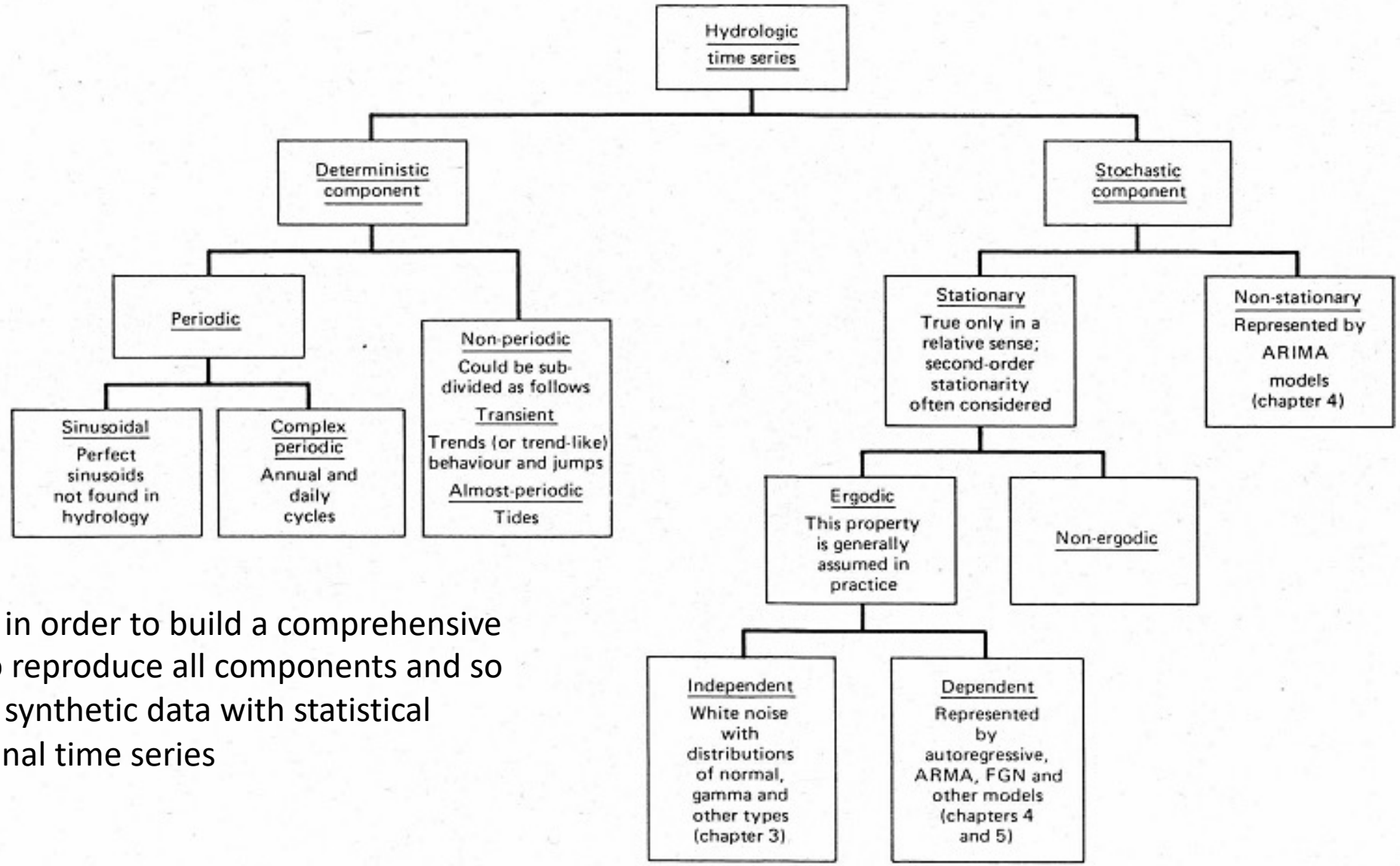
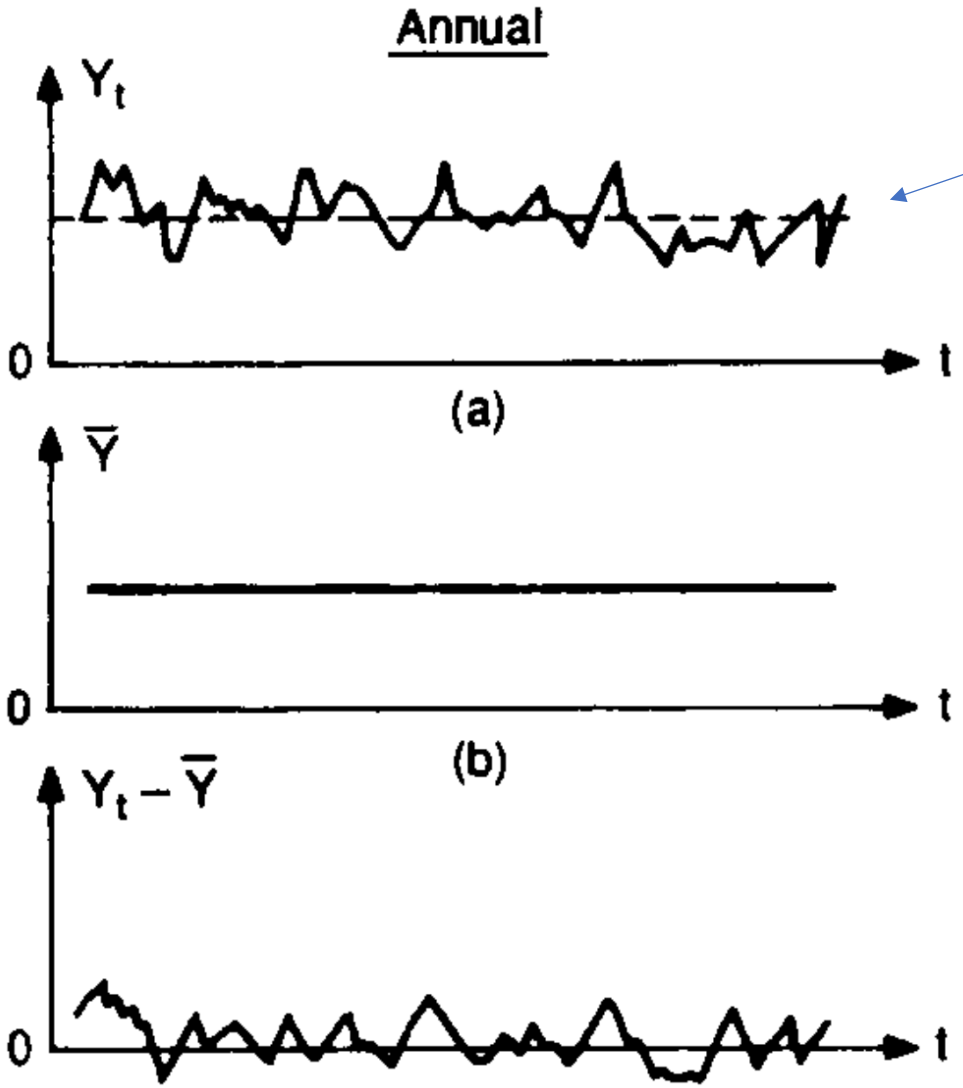


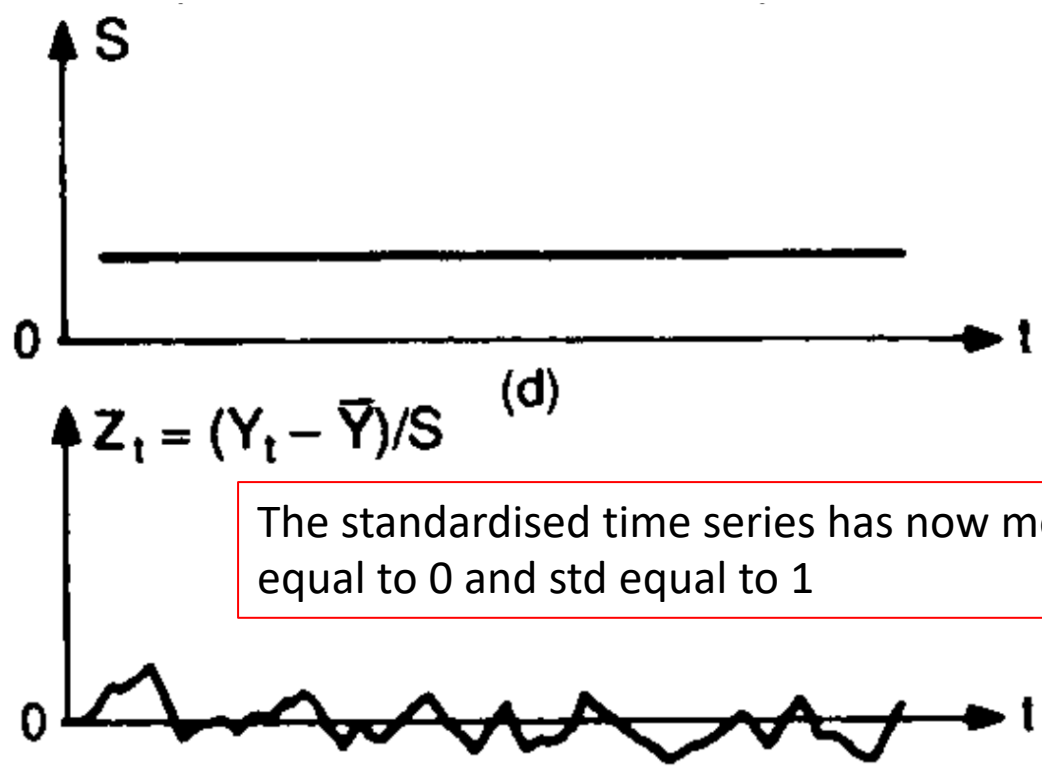
Figure 2.3 Components and subdivisions of a hydrological time series

Partitioning should be done in order to build a comprehensive mathematical model able to reproduce all components and so later be used to reconstruct synthetic data with statistical properties equal to the original time series

Standardisation



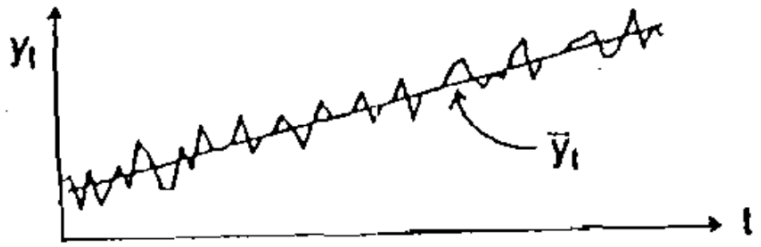
The univariate (single) time series is a sequence of annual total river discharges, for example. The original time series has (constant) mean different from zero and (constant) std different from 1.



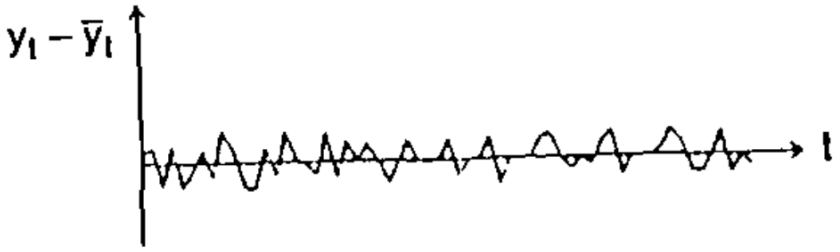
The standardised time series has now mean equal to 0 and std equal to 1

Removing trends

y_t is the variable affected by the linear or nonlinear trend $\bar{y}_t$



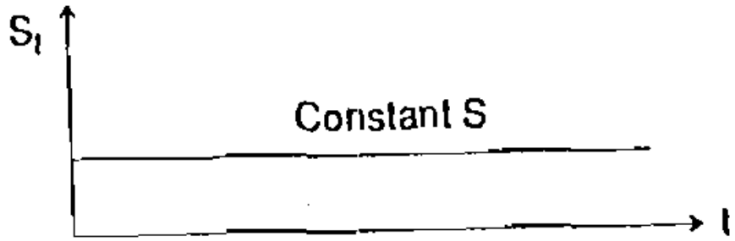
(a)



(b)

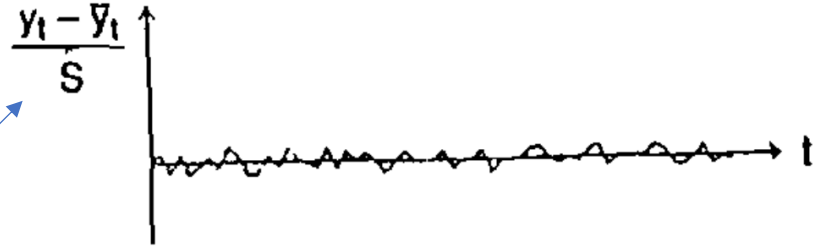
The detrended variable has now zero mean

The detrended variable constant (or nonconstant) standard deviation



(c)

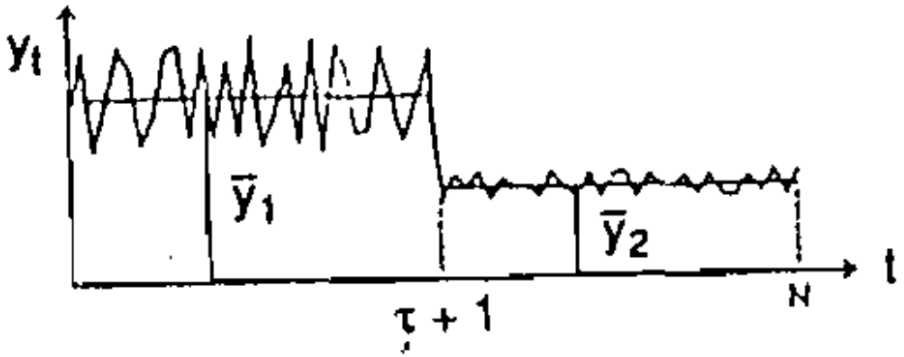
Standardisation



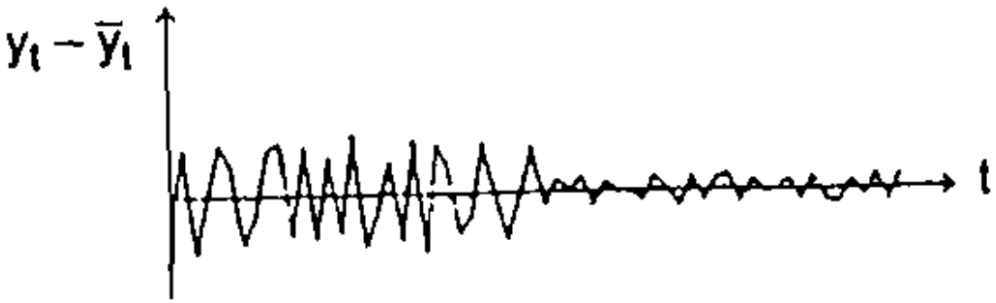
(d)

Standardisation transforms the process variable into another one having zero mean and unitary variance

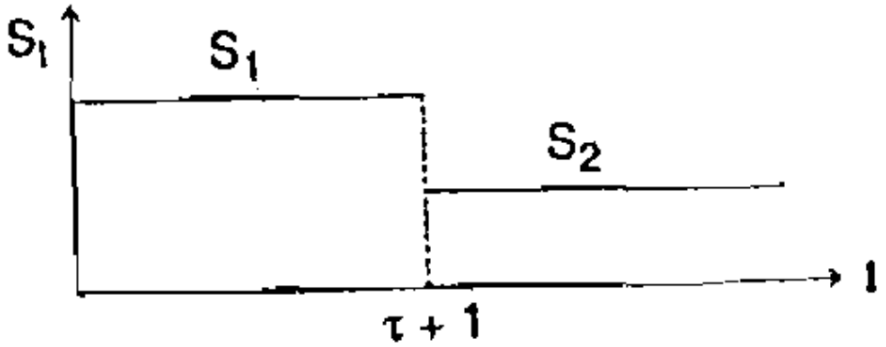
Removing shifts



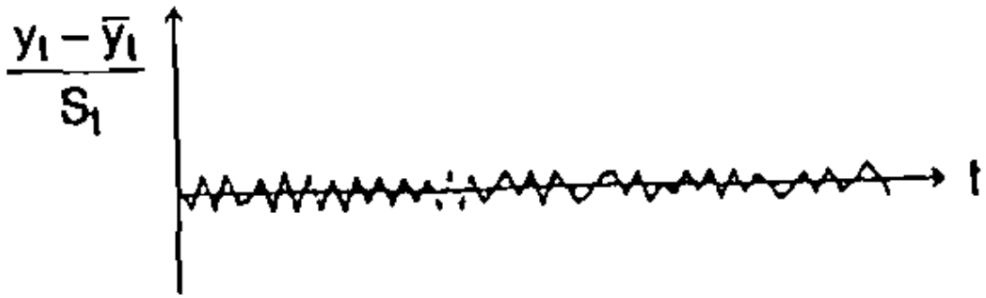
(a')



(b')



(c')



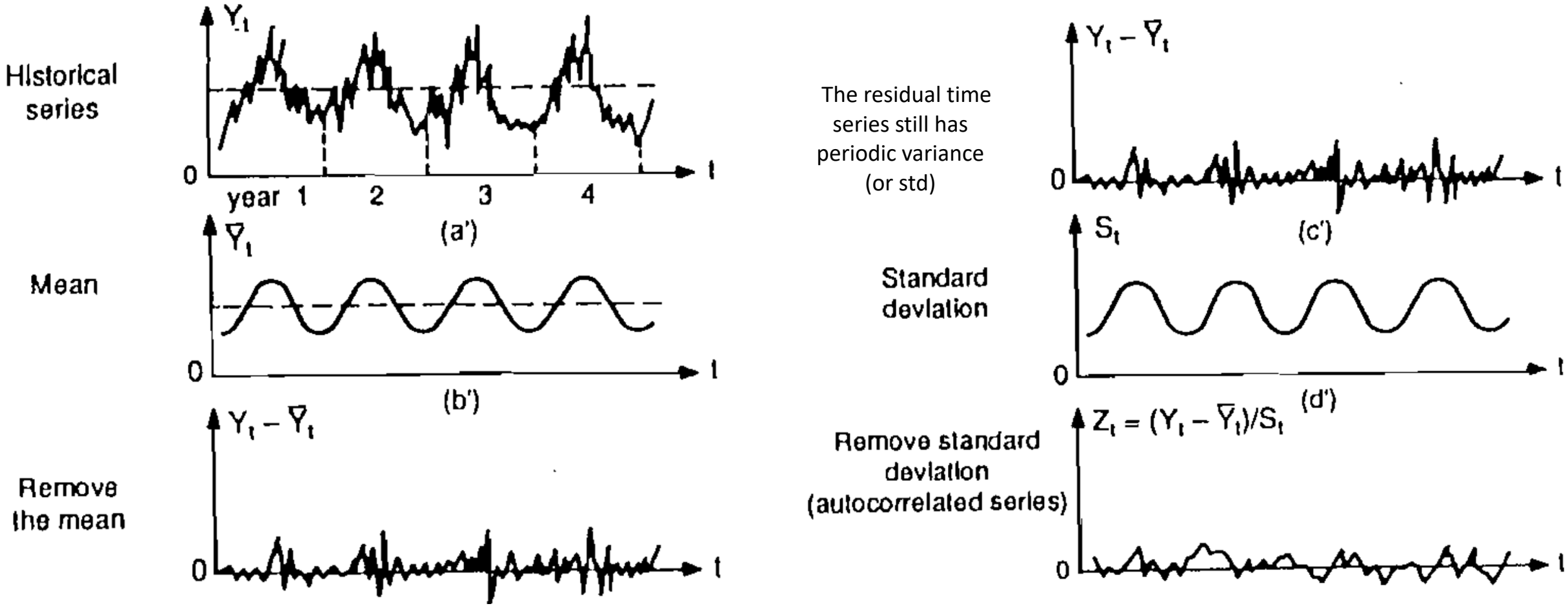
(d')

Shifts can be removed by translating the time series as for trends. The residual series may still possess time dependent variance caused by the shift

Standardisation may be used to obtain a transformed time series having homogeneous moments, yet containing temporal autocorrelation

Removing time dependent trends (e.g., seasonality)

Hydrological and meteorological variables often exhibit almost periodic patterns at the monthly (seasonality) and the daily (synoptic) time scales. These can be removed as well from the moments

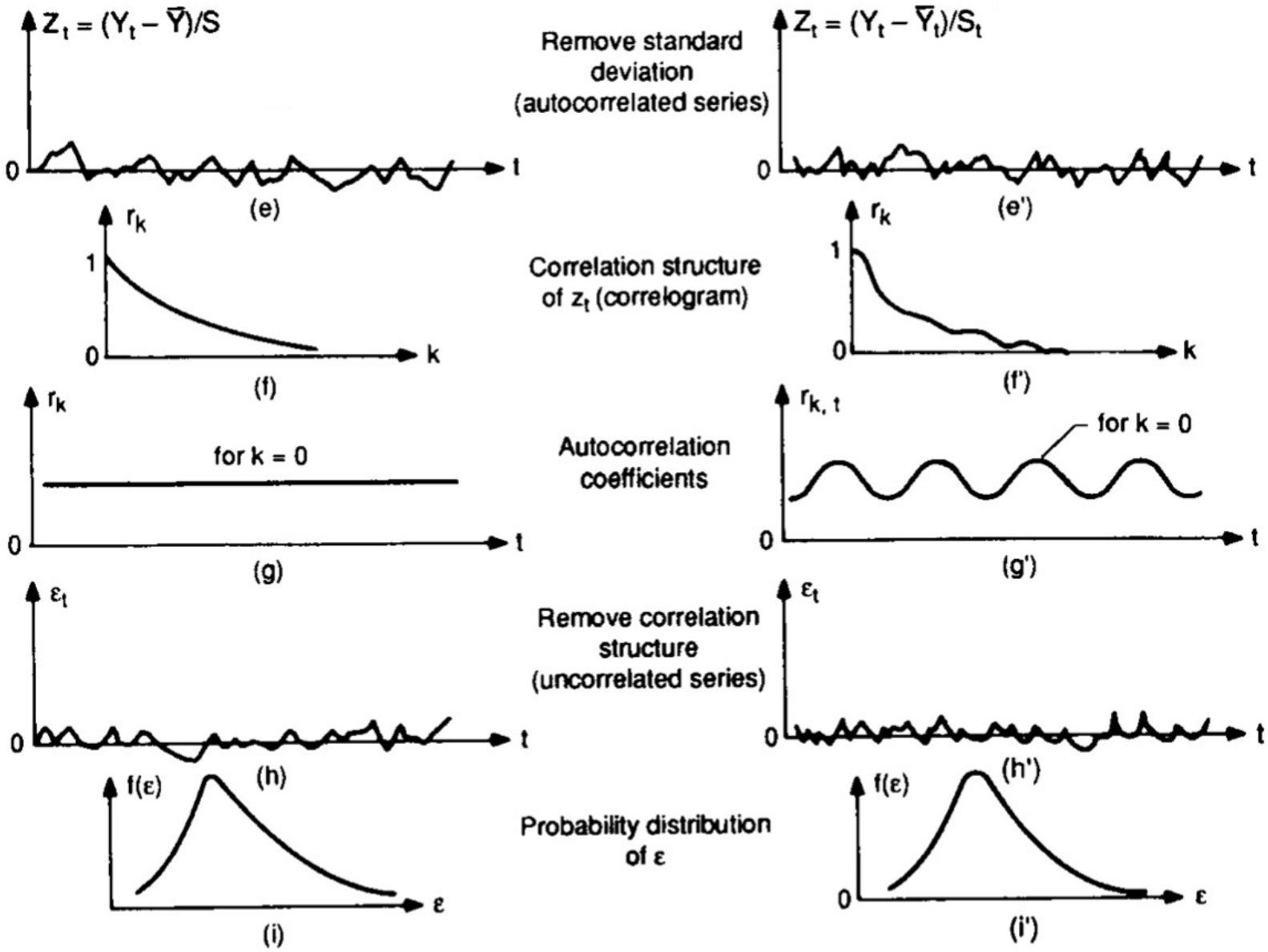


Seasonal standardisation (deseasonalisation)

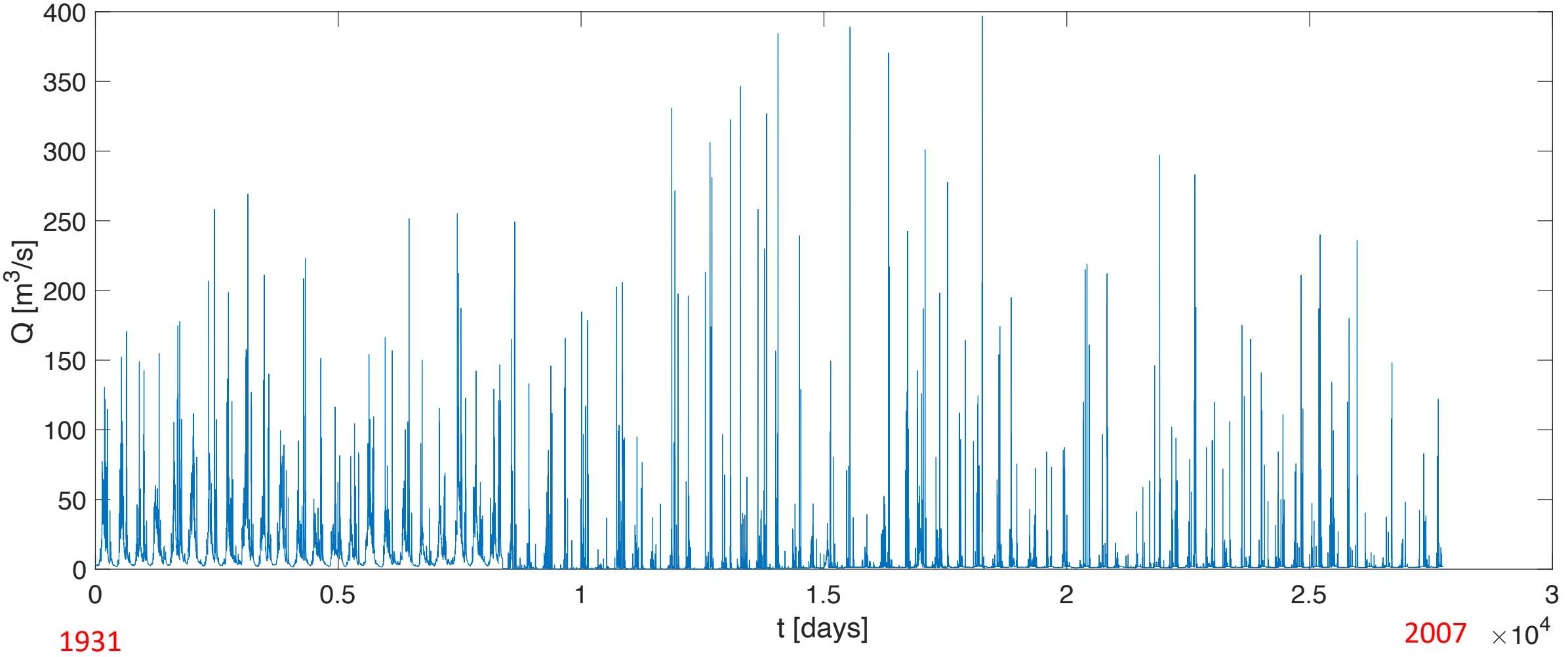
Removing correlation

The standardized time series may still present a temporal correlation structure (even periodic), which can be removed in order to separate the deterministic structure from the noise affecting the data

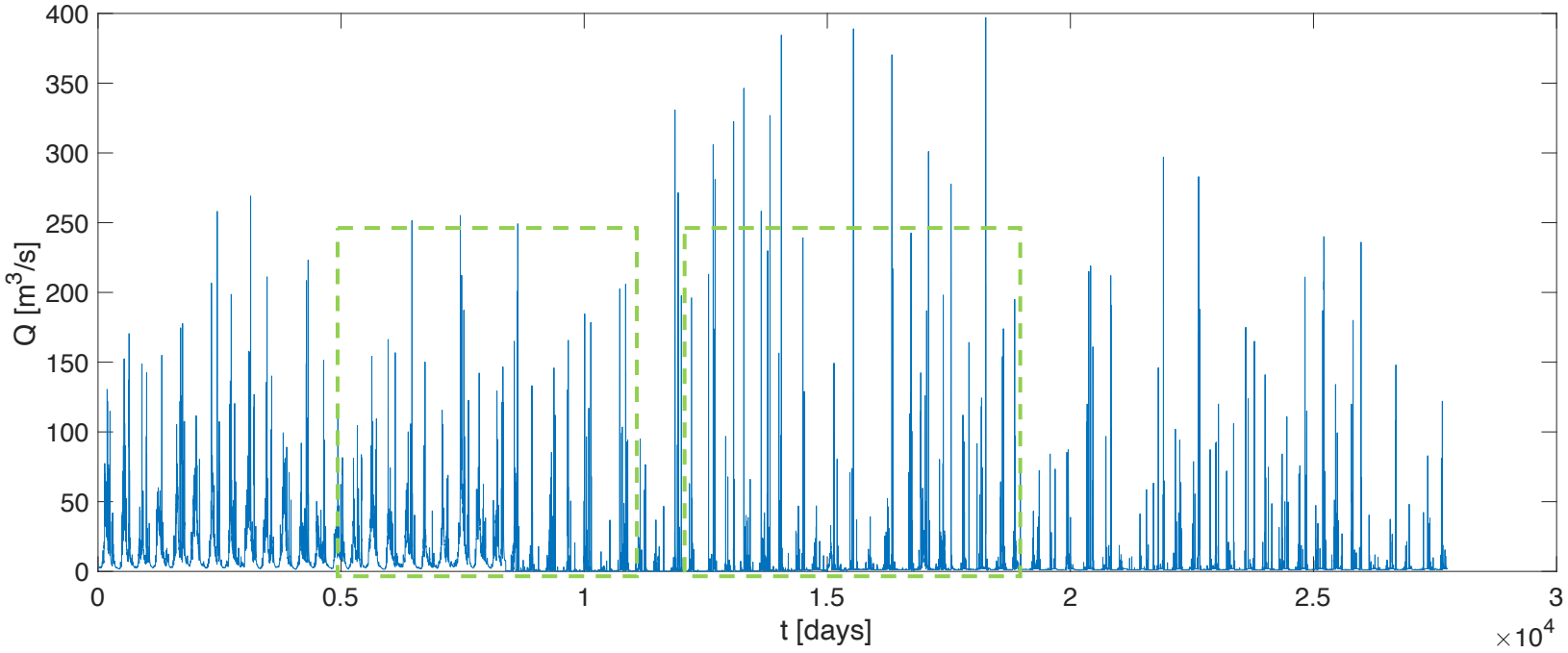
In the following, we will examine how all these operations can be done by means of deterministic and stochastic models

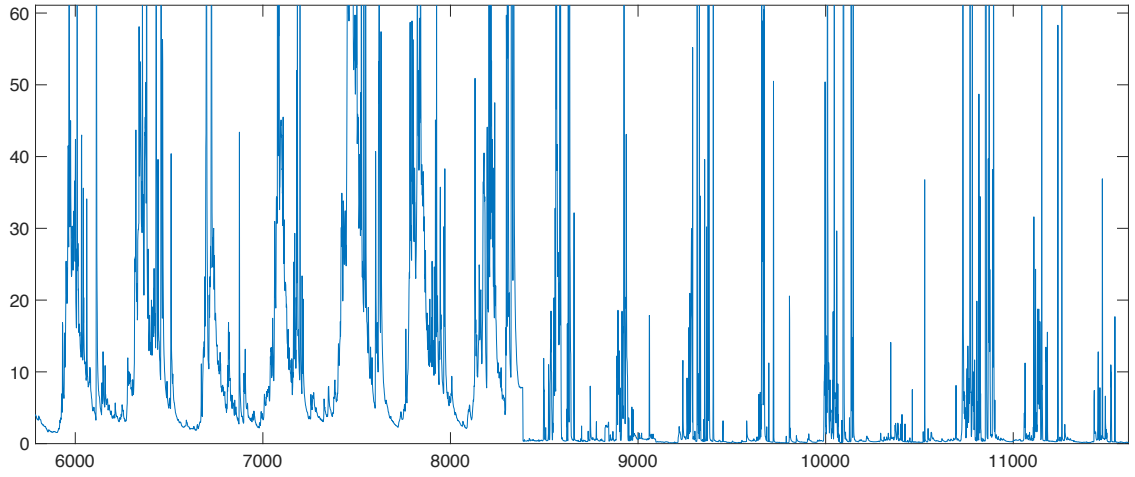


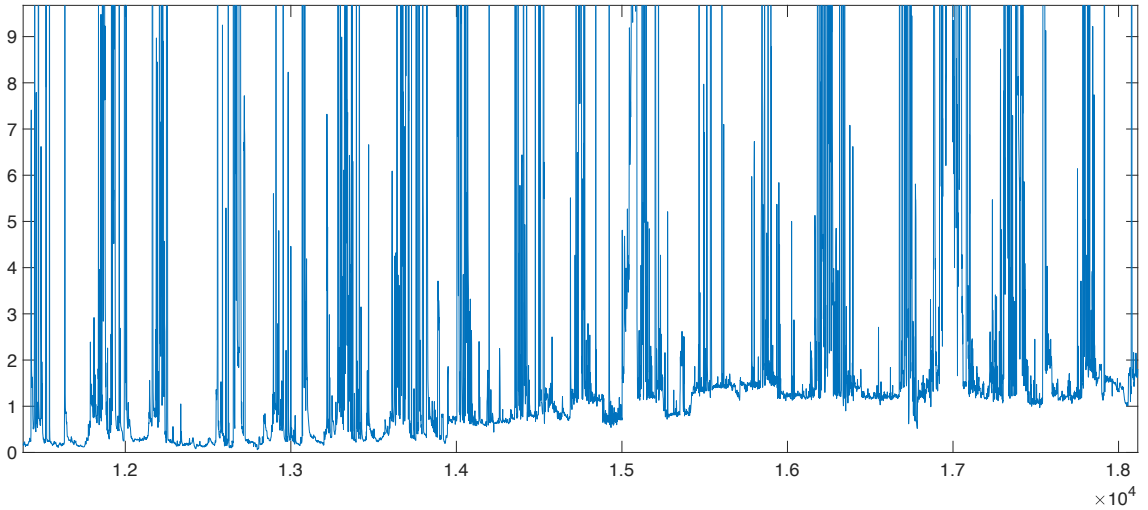
Example: diagnostic of a real river discharge (Maggia)



What do we see here?



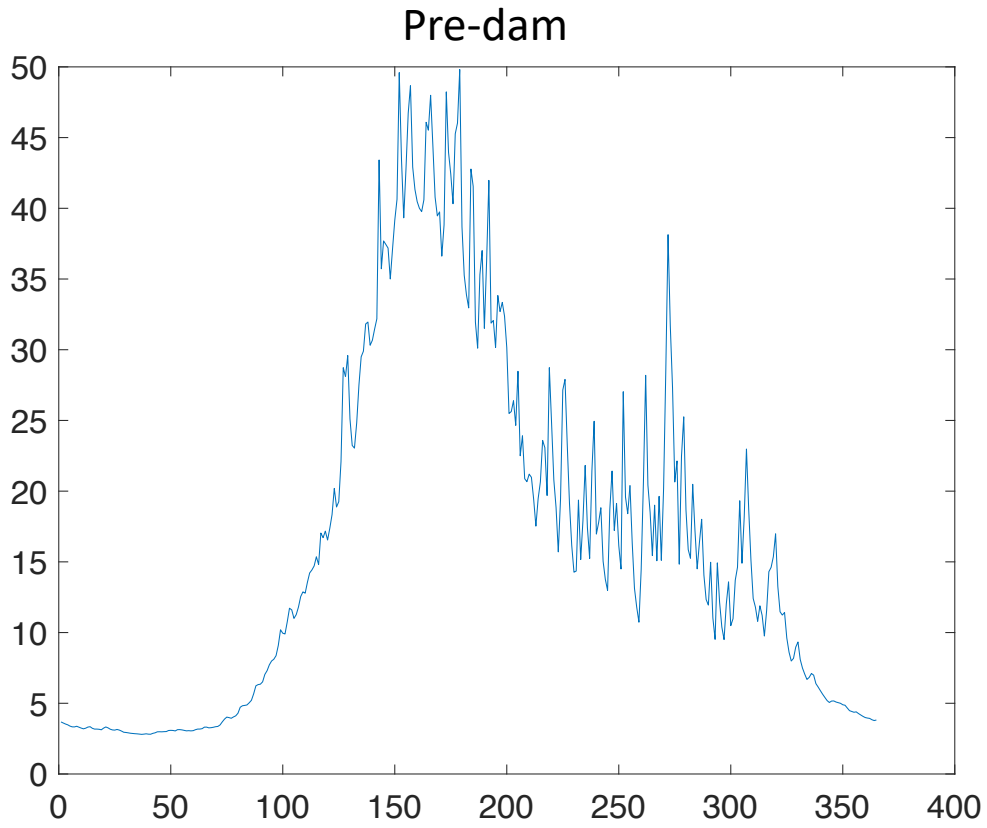


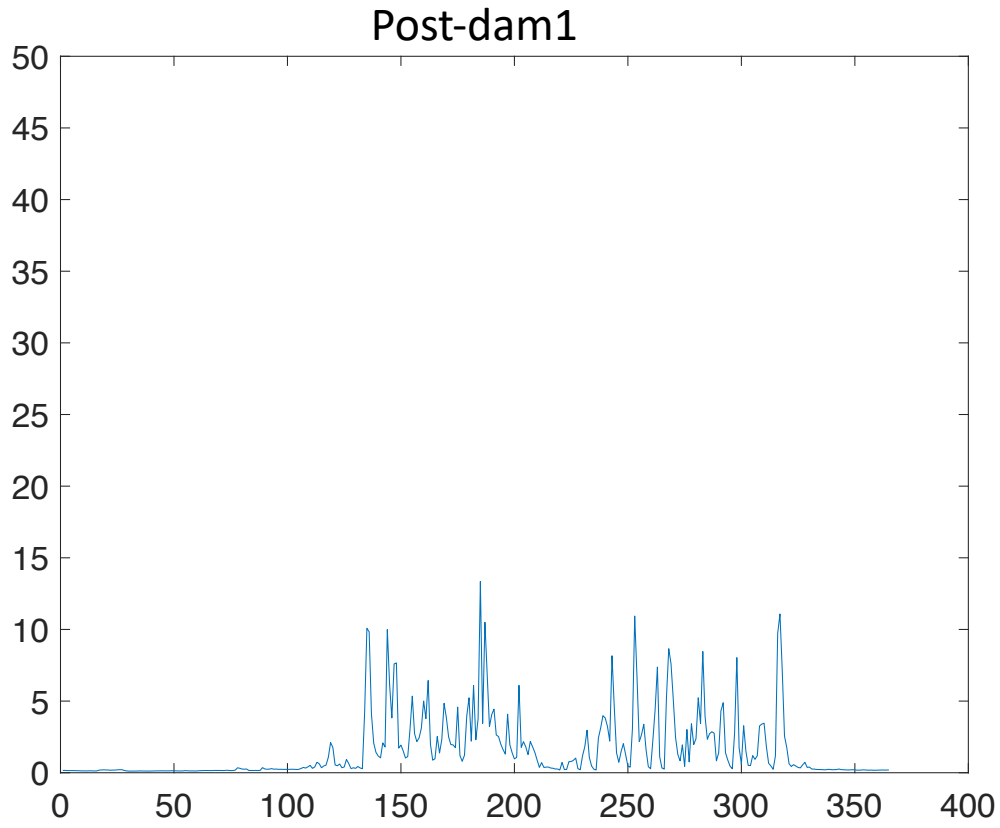


Vertical shift: onset of storage detention

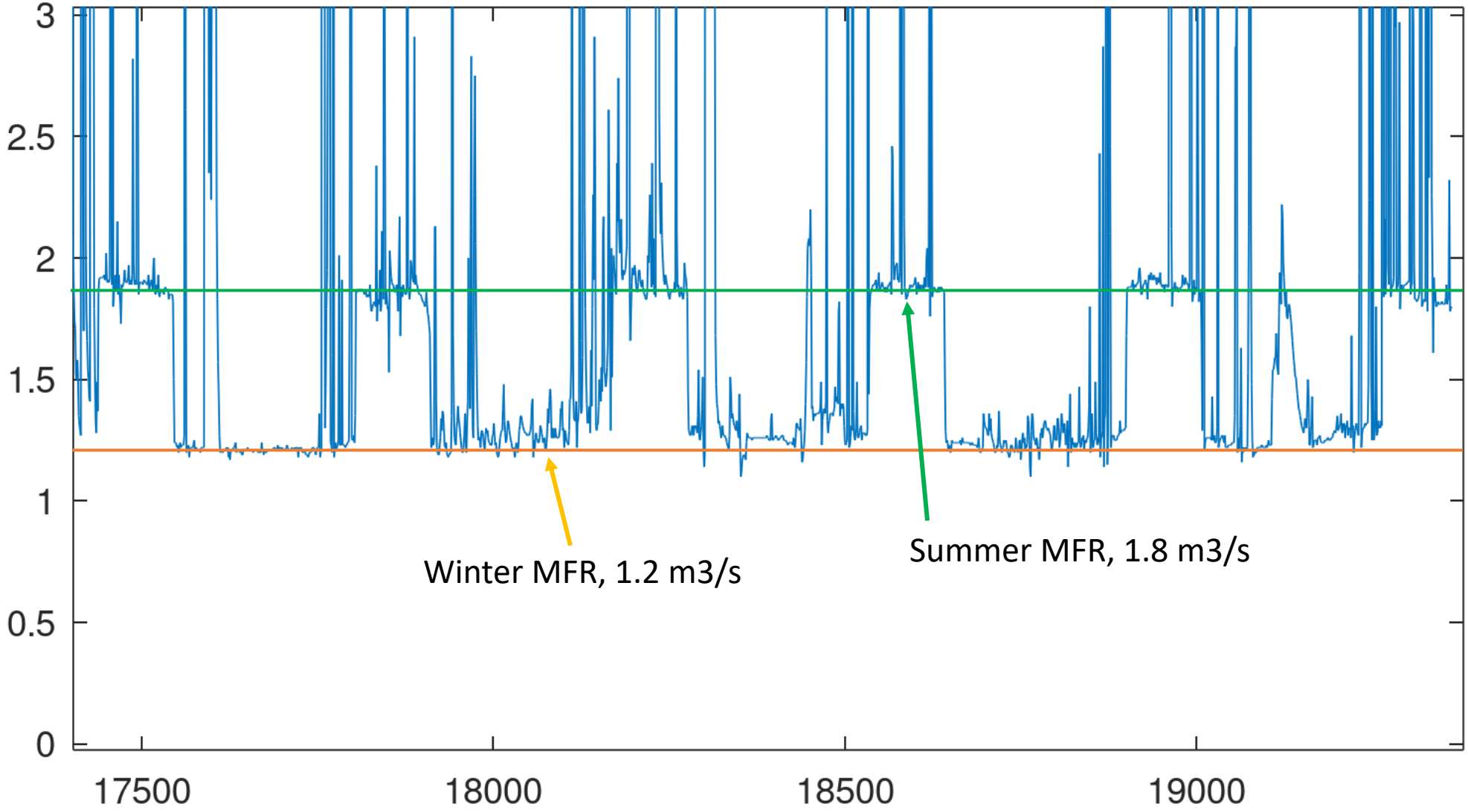
Gradual minimum increase: onset of MF policy

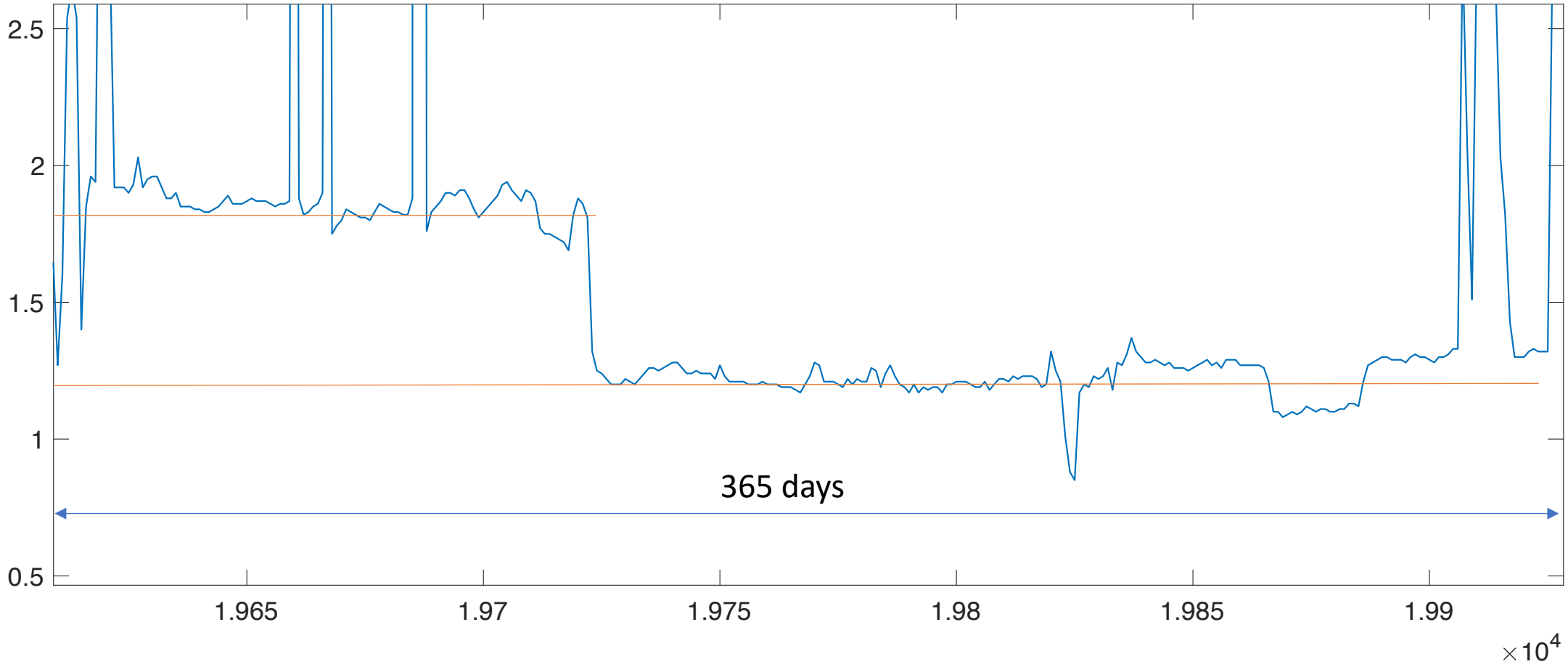
Average year in the pre- and post-dame periods

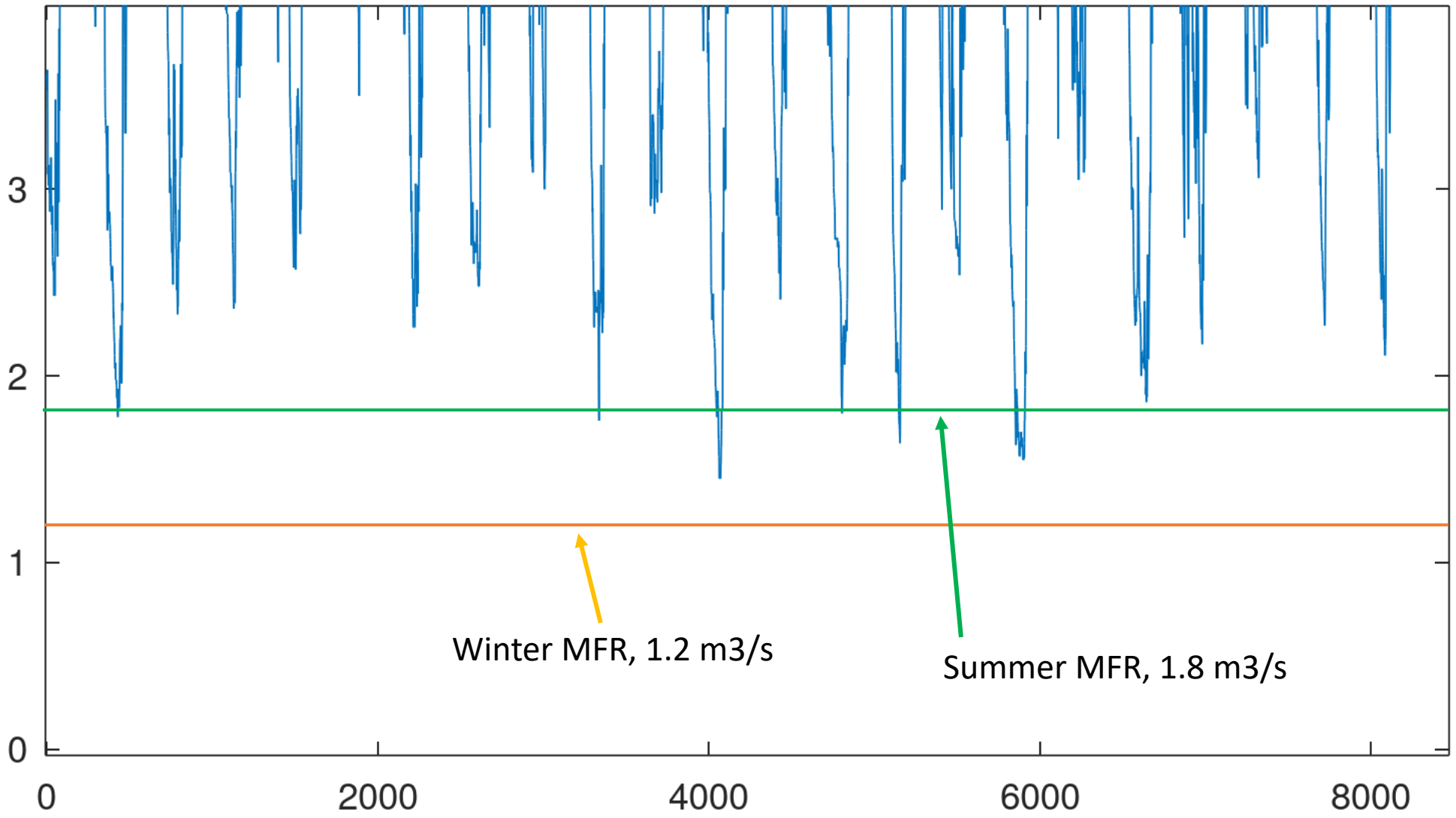




Minimal Flows: summer and winter values




2.5
2
1.5
1
0.5
365 days
1.965
1.97
1.975
1.98
1.985
1.99
×10^4


3
2
1
0
0
2000
4000
6000
8000
Winter MFR, 1.2 m3/s
Summer MFR, 1.8 m3/s