Course : Hydrological Analysis and Modelling

Chapter 1: Frequency Analysis and Statistical Applications in Hydrology

1.1. Introduction

Frequency analysis is a statistical prediction method that involves studying past events characteristic of a given process (hydrological or otherwise) in order to define their probabilities of future occurrence. This prediction relies on the definition and implementation of a frequency model, which is an equation describing the statistical behavior of a process. These models describe the probability of occurrence of an event with a given value.

The choice of the number of years of observations is crucial for the reliability of the results; a minimal threshold suggests that at least 10 years of data are necessary to justify a statistical analysis within the framework of the annual maxima method.

With frequency analysis, we aim to model the function:

$F(x) = P(X > x) = 1 - F(x)$ where F is the cumulative distribution function of X. We do not seek to model it in its entirety, but only in the tail of the distribution, that is, when $x > X$. Two approaches are possible:

The first relies on partitioning the data into blocks, whose maxima are assumed to be distributed according to a GEV (Generalized Extreme Value) distribution. This is the approach typically used by hydrologists.

The second models the distribution of values exceeding a given threshold. It is commonly referred to as the POT (Peaks-Over-Threshold) method.

Purpose of Frequency Analysis

The purpose of frequency hydrology is to forecast the occurrence of a hydrological event in the future by calculating quantiles, based on the interpretation of the historical data of hydrological events, which highlights the importance of the availability and reliability of information. Four essential steps can be distinguished for frequency calculation:

- Selection of a data sample that meets certain statistical criteria;
- Identification of the best theoretical probability distribution reflecting the sample;
- Statistical inference to verify the adequacy of the chosen distribution;
- Use of the appropriate distribution to form a set of possible future scenarios.

Cumulative Distribution Function

The cumulative distribution function aims to completely describe the random variable under study by providing the probability that a realization of this variable is less than a value x. It represents the proportion of the considered population whose value is less than x.

The probability density function represents the derivative of the cumulative distribution function.

Choice of the Fitting Model

The fitting of a frequency model to a given sample aims to consolidate the information within that sample into a single function and its parameters. The choice of the model in frequency analysis is undoubtedly the most critical step, introducing the greatest uncertainties. Adopting a frequency model to study and describe hydrological phenomena is therefore a decision; it provides guidance for selecting the appropriate theoretical distribution:

In a given climatic region, a specific hydrological variable generally follows the same distribution at all observation sites, highlighting the importance of systematic studies and awareness of previous research.

In the absence of regional information, a graphical representation of the observed points on Gaussian paper can be attempted, which allows for considering a normal distribution as a skewed distribution.

The choice of the frequency model is crucial for the validity of the results of a frequency analysis; however, there is no universal formula for selecting a model. For extreme values, the GEV and Gumbel distributions are the most suitable. The Gumbel distribution is less likely to apply to extreme values of hydrological variables, while the Type II extreme value distribution (EV2) is a more appropriate choice.

1.3. Introduction and Reminder of Basic Concepts

Arithmetic Mean

The arithmetic mean is the first measure of central tendency, denoted as $\overline{x}$:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Where:

n: is the number of observations in the sample;

$\mathbf{X_i}$: is the value of an observation.

The Range of a Sample

The range of a sample is the difference between the two extreme values.

$$A = X_{max} - X_{min}$$

- Standardized Variable:

$$U = \frac{x-\bar{x}}{\sigma}$$

Variance and standard deviation

$$Var = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$\boldsymbol{\sigma} = \sqrt{\text{Var}}$$

With $\boldsymbol{\sigma}$ and $\bar{x}$ being the standard deviation and the mean of the hydrological series, respectively.

$$\sigma = \sqrt{\sum \frac{(x-\bar{x})^2}{n}}$$

With **n**: the size of the series.

- Coefficient of Variation:

$$Cv = \frac{\sigma}{\bar{x}}$$

This definition is valid for data with a non-zero mean.

- **The Median**: is the quantile at 50%.
- **Measure of Skewness**, also known as the coefficient of asymmetry. Pearson suggests using the coefficient.

With $\mathbf{m_3}$, the standardized central moment of order 3, it is equal to:

$$\mathbf{m_3} = \frac{n}{(n-1)(n-2)} \sum (x - \bar{x})^3$$

Three cases for β_1:

1. If β_1 is purely positive, the distribution is spread to the right; we observe the sequence mode, mean, and median, and the skewness is said to be positive.
2. If β_1 is purely negative, the distribution is spread to the left; we observe the sequence mean, median, and mode, and the skewness is said to be negative.
3. If β_1 is zero, in this case, the skewness is not necessarily symmetric.

- Measure of Kurtosis:

Pearson suggests using the coefficient β_2:

$$\beta_2 = \frac{m_4}{\sigma^4} - 3$$

With $\mathbf{m_4}$, the standardized central moment of order 4, it is equal to:

$$\mathbf{m_4} = \frac{1}{n} \sum (x - \bar{x})^4$$

Three cases for β_2:

1. If β_2 is positive, the distribution is more peaked than the normal distribution, indicating that the distribution is leptokurtic.
2. If β_2 is negative, the distribution is flatter than the normal distribution, indicating that the distribution is platykurtic.
3. If β_2 is zero, the kurtosis is the same as for the normal distribution, indicating that the distribution is mesokurtic.

1.4. Forecasting and Prediction

The least squares line between two variables "X" and "Y" is given by the expression:

$$\mathbf{Y = aX + b}$$

$$\sum Y = b.\, n + a \sum X$$

$$\sum XY = b.\sum X + a\sum X^2$$

With **n**: the size of the pair (x, y).

After estimating the linear model, it can be used to produce forecasts of the dependent variable for periods following those of the sample used for estimation.

We differentiate between short-term and long-term forecasting.

The efficiency **E** of the forecast is:

$$E = 1 + (1 - \frac{k}{n}).\frac{1-(k-2).r^2}{k-3}$$

With **K**: the number of years of the pair (x, y).

The effective number of years **n'** is equal to:

$$\mathbf{n'} = \frac{k}{E}$$

r: regression coefficient of the pair (x_n, y_n) which is equal to:

$$\mathbf{r} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

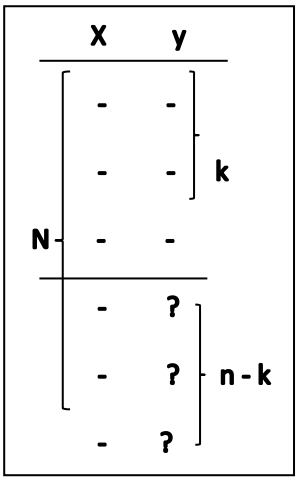
1.5. Filling in Gaps Using Linear Regression Method

To fill in the gaps in the series with discontinuous observations, we used the linear regression method.

This filling of a non-homogeneous discontinuous series is ensured by another homogeneous discontinuous series. For the method to be effective, the regression must be linear and the variables compared should follow a normal distribution. We estimate the variable (Y) based on the variable (X) by:

$$\mathbf{Y = a\ X + b}$$

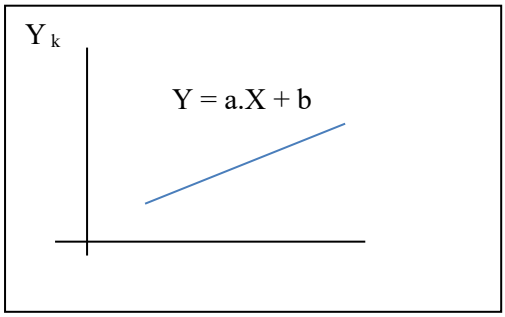
With **n**: size of the series (x) and **k**: size of the series (y).



$\mathbf{r_k}$: regression coefficient of the pair (x_k, y_k)

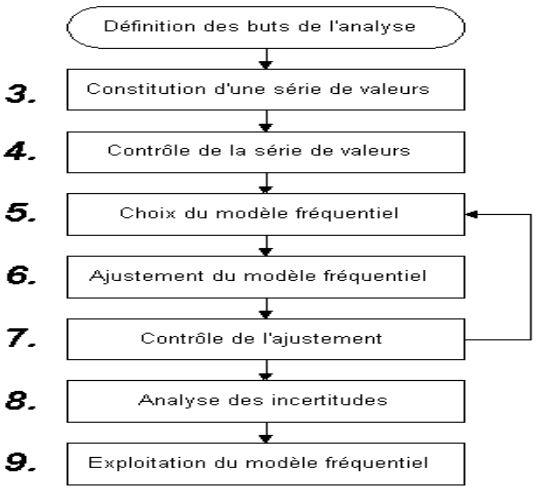
$$a = r_k.\frac{S_{k\,(y)}}{S_{k\,(x)}}, \quad b = \sum \frac{(x-\bar{x})(y-\bar{y})}{(x-\bar{x})^2}$$

With: $S_{k\,(y)} = \sigma_y^2$



1.4. The Principle of Frequency Analysis

The various steps can be outlined as follows:



1.5. Use of Frequency Models

1.5.1. Normal Distribution

The normal distribution function, also known as the Laplace-Gauss distribution.

The expression for the density function of the normal distribution is:

$$f(x) = \frac{1}{\sigma.\sqrt{2\pi}}.e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} \qquad (*)$$

With $\bar{\mathbf{x}}$ as the mean of the series.

We can replace $\frac{(x-\bar{x})^2}{\sigma^2}$ by z^2

With **z**: the standardized normal variable. The equation (*) becomes as follows:

$$f(z) = \frac{1}{\sigma.\sqrt{2\pi}}.e^{-\frac{z^2}{2}} \text{ et } F(z) = \frac{1}{\sigma.\sqrt{2\pi}}\int_{-\infty}^{z} e^{\frac{z^2}{2}}.dz$$

The values of **F(z)** are provided by the Gaussian integral table.

1.5.2. Log-Normal Distribution

The cumulative distribution function of the log-normal distribution (also known as Galton's law) is written as:

$$F(x) = \frac{1}{\sigma.\sqrt{2\pi}}\int_{-\infty}^{z} e^{\frac{z^2}{2}}.dz$$

With $z = a.\log(x - x_0) + b$

Where z follows a normal distribution while the distribution of x is called log-normal.

a, b, and x_0 are parameters, with x_0 being the location parameter.

1.5.3. Gumbel Distribution

Cumulative distribution function of the Gumbel distribution is:

$$F(x) = e^{-e^{-\alpha(x-x_0)}}$$

F(x): frequency of non-exceedance (FND).

α, x_0 : : adjustment coefficients.

By changing the variable $y = \alpha (x- x_0)$

The Gumbel law is written as:

$$F(x) = F(y) = e^{-e^{-y}}$$

y: Gumbel reduced variable

F(y): non-exceedance frequency

L'équation $y =\alpha(x - x_0)$

represented in the form $x = \frac{1}{\alpha} y + x_0$

An approximation of the values $\frac{1}{\alpha}$ and x_0 given by:

$\frac{1}{\alpha} = 0{,}78\,\sigma$, $x_0 = \bar{x} - \frac{0577}{\alpha}$ and $y = - \ln (- \ln F(x))$

1.5.4. Generalized Extreme Value (GEV) Law

The expression for the density function of the GEV distribution is:

$$f(x) = \frac{1}{s}\left(1 - \frac{k(x-x_0)}{s}\right)^{\frac{1}{k}-1} e^{-\left(1-\frac{k(x-x_0)}{s}\right)^{\frac{1}{k}}}$$

The distribution function of the GEV distribution is:

$$F(x) = e^{-\left(1- \frac{k(x-x_0)}{s}\right)^{\frac{1}{k}}}$$

a) Calculation of parameters: x_o , s and k

These parameters x_o, s and k, depend on the first three weighted momentsb 0 , b1, and b2 that needs to be calculated.

First, we will estimate b 0 , b1, and b2:

$$b_0 = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$b_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{i-1}{n-1} x_i$$

$$b_2 = \frac{1}{n}\sum_{i=1}^{n} \frac{(i-1)}{(n-1)}\frac{(i-2)}{(n-2)} x_i$$

n: the sample size and i: the rank in the sample sorted in ascending order.

It can be shown that these first three weighted moments b0, b1, and b2 are related to the three parameters xo, s, and k by the following relationships.

$$\frac{3b_2 - b_0}{2b_1 - b_0} = \frac{1 - 3^{-k}}{1 - 2^{-k}}$$

$b_0 = x_0 + \frac{s}{k}(1- \Gamma (k + 1))$

The Gamma function is: $\Gamma(1+ k) = k\,\Gamma(k)$

In the general case, where $- 0.5< k < 0.5$

We can evaluate k explicitly using the following relationship:

$$2b_1 - b_0 = \frac{s}{k}\,\Gamma\,(k + 1)(\,1 - 2^{-k})$$

The Gamma function is: : Γ(1+ k) = k Γ(k)

In the general case, where - 0.5< k < 0.5

We can explicitly evaluate k using the following relationship:

$$k = 2{,}9554\ c^{2} + 7{,}8590\ c$$

With $c = \frac{2b_1 - b_0}{3b_2 - b_0} - \frac{\ln 2}{\ln 3}$

We can immediately deduce the values of s and x_0 by:

$$S = \frac{(2b_1 - b_0)k}{(1-2^{-k})\,\Gamma\,(k + 1)} \text{ et}$$

$$x_0 = b_0 + s\ \frac{\Gamma\,(k + 1)-1}{k}$$

b) Calculating quantiles

To calculate quantiles, simply apply the formula:

$X_{p\,\%} = x_0 + \frac{s}{k}\,(1-(-\,\text{Ln}\,F(x))^{k})$ for each return period.

1.6. Estimation of frequency model parameters

Determination of the parameters of the chosen theoretical distribution law (Adjustment):

In frequency analysis, several methods have been developed for estimating the parameters of distributions of interest. These include:

- Maximum likelihood;
- The method of moments;
- Others

1.6.1. Maximum likelihood method

The maximum likelihood method can be used to estimate the parameters m1 and m2. The method stipulates that the values of m1 and m2 should be those that maximize the probability of obtaining the observed values for the variable x. Thus, the maximum likelihood estimation procedure first requires the definition of a function of m1 and m2, known as the likelihood function, which describes the probability of obtaining the observed values of " x " followed by the maximization of this function for m1 and m2.

1.6.2. The Method of Moments

The idea behind this method is as follows: if the parameters are estimated correctly, then there should be a match between the observed (or empirical) characteristics and the theoretical characteristics. We will seek this match using moments.

1- Adjusting the parameters of Galton's law

$$U = a.\ \log_{10}(x - x_0) + b$$

***) According to the method of moments:**

$$\frac{\sigma^4}{m_3} = \frac{(\bar{x} - x0)^3}{\sigma^2+3(\bar{x}-x_0)^2}$$

$$a = \frac{1.517}{\sqrt{\log_{10}\left(1+\frac{\sigma^2}{(\bar{x}-x_0)^2}\right)}}$$

The value of x 0 is found by successive approximation.

***) According to the maximum likelihood method:**

$$2{,}3026\left(\sum\frac{1}{xi-x_0}\right)\left(\frac{1}{n}\sum log^2(xi-x_0)-\frac{1}{n^2}\left(\sum\log(xi-x_0)\right)^2\right)=\frac{1}{n}\left(\sum\log(xi-x_0)\right)\left(\sum\frac{1}{xi-x_0}\right)-\sum\frac{\log(xi-x_0)}{xi-x_0}$$

$$a^2=\frac{1}{\frac{1}{n}\sum log^2(xi-x_0)-\frac{1}{n^2\left(\sum\log(xi-x_0)\right)^2}}$$

$$b=-\frac{a\sum\log(xi-x_0)}{n}$$

2- Adjustment of Gumbel law parameters

*) According to the method of moments:

$s = 0{,}78\ \boldsymbol{\sigma}_x$, $\quad x_0=\bar{x}-0{,}577s \quad \boldsymbol{\sigma}_x=\sqrt{\sum\frac{(x-\bar{x})^2}{n-1}}$

*) According to the maximum likelihood method:

$$\bar{x}=s+\frac{\sum xe^{\frac{-x}{s}}}{\sum e^{\frac{-x}{s}}} \qquad x_0=-s.\ln\left(\frac{\sum e^{\frac{-x}{s}}}{n}\right)$$

1.6.3. Confidence intervals

The confidence interval has three properties. Its amplitude is greater when:

- The degree of confidence (probability that the true value lies within the interval) chosen is high;
- The dispersion;
- The sample size N is small.

The choice of confidence level depends on the risk the project manager is willing to accept.

The higher the level of security sought, the higher the confidence level chosen.

Commonly accepted values are:

95% for projects that are economically important and/or require a high level of security.

70% for projects of lesser economic importance and/or that do not require a very high level of security.

Note:

In a 95% confidence interval, there is a 95% chance of finding the estimated parameter value, but there remains a 5% chance of finding it outside the interval, 2.5% for this value to exceed the upper limit of the confidence interval, and 2.5% for it to be below the lower limit of this confidence interval.

1.6.3.1. Confidence intervals of the Gaussian distribution

Confidence interval for the mean:

$u_{\frac{1-\alpha}{2}} = 1.96$ for a 95% confidence interval $\bar{x} - u_{\frac{1-\alpha}{2}}.\frac{s}{\sqrt{n}} < \bar{x} < \bar{x} + u_{\frac{1-\alpha}{2}}.\frac{s}{\sqrt{n}}$

Confidence interval for standard deviation:

$$s - u_{\frac{1-\alpha}{2}}.\frac{s}{\sqrt{2n}} < \sigma < s + u_{\frac{1-\alpha}{2}}.\frac{s}{\sqrt{2n}}$$

Confidence interval for a quantile xp

$$x_p - u_{\frac{1-\alpha}{2}}.\frac{s}{\sqrt{2n}}\sqrt{2+u_p^2} < xp < x_p + u_{\frac{1-\alpha}{2}}.\frac{s}{\sqrt{2n}}\sqrt{2+u_p^2}$$

Where $\bar{x}$, σ, and s" are the mean, standard deviation, and variance of the series, respectively.

Confidence intervals of the log-normal distribution for the mean

$$\text{Ln}\,\bar{x} - u_{\frac{1-\alpha}{2}} \frac{s}{\sqrt{2n}} \sqrt{2+u_p^2} < \text{Ln}\,\bar{x} < \text{Ln}\,\bar{x} + u_{\frac{1-\alpha}{2}} \frac{s}{\sqrt{2n}} \sqrt{2+u_p^2}$$

1.6.3.2. Confidence intervals for the Gumbel distribution

Quantile:

$$X_F - h_1\, S_x < X_F < X_F + h_2\, S_x$$

F: the frequency of the α value.

h1 and h2 are parameters that depend on the sample size "n."

$$h_1\,,\, h_2 = \frac{(\frac{t_\alpha}{n^{0.5}})(1+1.13t_F+1.1\,t_F^2)^{0.5} \pm \frac{t_\alpha^2}{n}(1.1\,t_F+0.57)}{1-1.1\,t_\alpha^2/n}$$

and $t_F = \frac{-ln\left(-ln(F)\right)-0.577}{1.28}$

t_α = reduced Gaussian variable corresponding to the FND =1 – (1-α)/2

Example for IC = 90%

FND = 1 – (1- α)/2 = 1- (1- 0.9)/2 = 0.95

Where $\mathbf{t}_\alpha$ = 1.6448 (Gaussian table)

$$t_F = \frac{-ln\left(-ln(F)\right)-0.577}{1.28} = \frac{-ln\left(-ln(0.9)\right)-0.577}{1.28} = 1.31$$

1.7. Confidence interval (linear regression)

In the case of linear regression Y = a X +b

Calculation of the confidence interval:

$$I_{p\%}(Y/X) = a.\,x + b \pm Z(p)\,\boldsymbol{\sigma}_\varepsilon$$

$I_{p\%}(Y/X)$: confidence interval of variable Y related to variable X.

Z (p): Gaussian variable (see Gaussian table);

$\boldsymbol{\sigma}_\varepsilon$: the standard deviation of the residual (ε), which gives the length of the confidence band in the y-direction.

$$\boldsymbol{\sigma}_\varepsilon^{\,2} = \boldsymbol{\sigma}_y^{\,2}\,(1-r^2).$$

r : regression coefficient between y and x.

Another method for calculating ε :

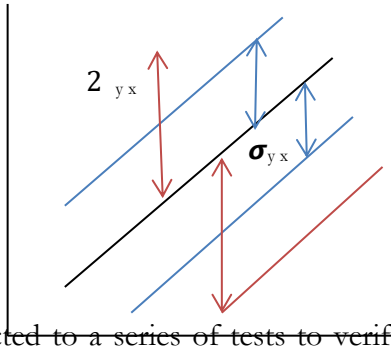
$\varepsilon = \boldsymbol{\sigma}_{y\,x} = \sqrt{\frac{(y\ existant - y\ estimé)^2}{n}}$ It is the standard deviation of Y related to X.

ε : "Is called the mean squared error."

The mean square error has the same properties as $\boldsymbol{\sigma}$.

$\boldsymbol{\sigma}_{y\,x}$: in the figure corresponds to the first confidence band

2.$\boldsymbol{\sigma}_{y\,x}$: in the figure corresponds to the second confidence band.



1.8. Adequacy tests (model validation tests)

Once the adjustment model has been chosen, it must be subjected to a series of tests to verify its adequacy for the chosen sample. The most commonly used tests include the χ 2 law and the Kolmogorov-Smirnov test. These are non-parametric tests that allow the H0 hypothesis to be tested, according to which the observed data are generated by a model involving a probability law or a probability family. Their principles are as follows:

1.8.1. Chi-square test

A measure of the difference between the observed frequencies and the theoretical frequencies is given by the χ 2 statistics:

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_1 - e_1)^2}{e_1} + \ldots\ldots + \frac{(o_1 - e_1)^2}{e_1} = \sum_{j=1}^{k} \frac{(o_{1j} - e_j)^2}{e_j}$$

Where N is the number of observations

1.8.2. Kolmogorov-Smirnov test

This is a goodness-of-fit test that compares a distribution of observed values to a theoretical distribution. This test plays the same role as the chi-square test.

With the Kolmogorov-Smirnov test, we seek the maximum value of the absolute value of the difference between the empirical distribution function FN(x) of a sample of N values and the corresponding theoretical distribution function F(x).:

$$DN = D\ max = max\ |F_n(x) - F(x)|$$

The table below shows all the steps to follow for the Kolmogorov-Smirnov test.

Column 1 indicates the order number i = 1,2,3,……..

Column 2 shows the rainfall data sorted in ascending order.

In column 3, we calculated the frequency at which the experimental value was not exceeded:

$$FND = \frac{i - 0.5}{N}$$

Column 4 indicates the reduced variable Z i -1 = $\frac{X_{I-1} - \bar{X}}{\sigma}$

Column 5 gives the theoretical FND taken from the Gauss table for each value;

Column 6 indicates the difference DN = D max = max $|F_n(x) - F(x)|$

1	2	3	4	5	6
ordre	pluies triées	fréquences expérimentales	variable réduite	fréquence théorique	différences absolues
		Fe	z	Fz	Fe - Fz

1.8.3. Anderson Darling Test

The Anderson-Darling statistic determines the extent to which data follows a specific distribution law. For a specific data set and distribution law, the better the law fits the data, the lower this statistic will be.

The Anderson-Darling test consists of comparing the theoretical distribution $F_0(x)$ to the experimental distribution F(x).

The standard Anderson-Darling case corresponds to the following weighting function:

By modifying the weighting function to w (x) = $\frac{1}{F_0(x)(1-F_0(x))}$

w (x) : non-negative weighting function.

The table below shows all the steps to follow in the Anderson Darling test.

1	2	3	4	5	6	7	8	9	10	11
initial data	data sorted in ascending order	i	2 i - 1	$z_i = \frac{x_i - \bar{x}}{\sigma}$	f(z_i) FND	Ln (f(z_i))	z_{n-i+1}	f(z_{n-i+1}) FND	Ln(1- f(z_{n-i+1}))	(4) x (7+10)
										Somme : (**S**)
										= - n - $\frac{1}{n}$.**S**

Anderson's test statistic:

$A^2 = - n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)(\ln F(z_i) + \ln(1 - F(z_{n-i+1})))$

F (z_i): is the theoretical frequency of the centered and reduced normal distribution law associated with the standard value.

Or $z_i = F(\frac{x_i - \bar{x}}{\sigma})$, with F distribution function of a centered-reduced normal distribution. Reduced to observation i, F(xi) corresponds to the theoretical frequency (FND).

The " x i " are the data (ordinates) and "n" is the sample size.

A correction is recommended for small sample sizes; this corrected statistic is also used to calculate the p-value.

$A_m = A^2(1 + \frac{0.75}{n} + \frac{2.25}{n^2})$

The A2 test statistic can be compared with a theoretical table.

This correction should only be made if n ≤ 40.

One way to calculate the p-value associated with Am is the following algorithm:

If $A_m < 0.2$ then $p = 1 - e^{-13..436 + 101.14\, A_m - 223.73\, A_m^2}$

If $0.2 \leq A_m < 0.34$ then $p = 1 - e^{-8.318 + 42.796\, A_m - 59.938\, A_m^2}$

If $0.34 \leq A_m < 0.6$ then $p = e^{0.9177 - 4.279A_m - 1.38\,A_m^2}$

If $0.6 \leq A_m < 10$ then $p = e^{1.2937 - 5.709A_m + 0.0186\,A_m^2}$

If $A_m \geq 10$ then $p < 0.0001$

The null hypothesis H0 is: The variable follows a normal distribution F x = FL N

1.8.4. Comparison of models (Akaike information criterion (AIC) and Bayesian information criterion (BIC)).

The statistical distribution that best fits the samples is selected using two criteria: the Akaike criterion and the Bayesian information criterion (BIC).

These two criteria allow the best-fitting distribution to be chosen, taking into account estimation error and parsimony.

a) Akaike's criterion (AIC):

The expression for the AIC criterion is as follows:

$$AIC = -2 \log (L) + 2k$$

With L: is the likelihood of the sample, also known as LL (likelihood).

$$\mathbf{L\ L =} \text{Log } L(\mu|X) = -\frac{n}{2} \log (2\pi\, \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

The likelihood of the sample relative to the data is the probability that the observation fits the model.

k: the number of parameters.

b) Bayesian criterion (BIC):

The BIC criterion is used in a Bayesian context for selecting a probabilistic model.

The BIC model maximizes the posterior distribution of models, i.e., the most plausible model based on the data. The Bayesian criterion is expressed as follows:

$$BIC = -2 \ln L + k \ln N$$

Where: L is the sample likelihood,

N: the sample size

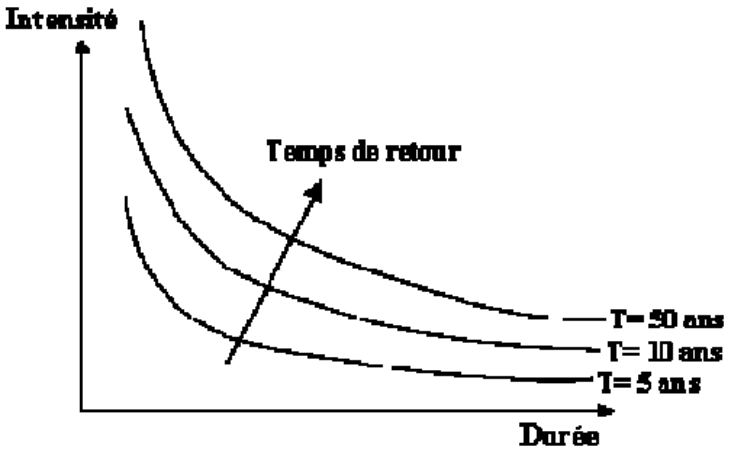
k: the number of parameters.

1.9. Use of a frequency model to construct IDF curves.

IDF curves represent rainfall intensity "i" as a function of rainfall duration and return period T.

Among the descriptive elements of precipitation and flow rates, Intensity-Duration-Frequency curves (IDF curves) are particularly useful as tools for designing hydraulic structures. They link the maximum average intensities of rainfall events to their duration and return period. Their methods of development, in which they are considered either from a series of spot measurements (analysis at a single station) or from a network of recording rain gauges (regional analysis).

The concept of frequency is expressed by the concept of return period.



1.10. Methods for constructing local intensity-duration-frequency curves

Either according to Montana's law or Talbot's law.

I (mm/min) = $\frac{a}{t^{b}}$ Montana formula

Where t is the intensity of the rainfall in (mm/min), a and b are adjustment parameters depending on local rainfall as a function of time, and t is the duration of the rainfall in minutes.

The estimation of Montana's parameters a and b is simplified by taking the logarithm of this formula to obtain a linear relationship:

Ln (i_t) = ln(a) – b. ln (t)

$I_{max} = \frac{a}{(b+t)^{b}}$

For T =1/FD given or t = Δt duration of the shower.

$I_{max} = \frac{P_{\Delta t\,(mm)x\,60(\min)}}{\Delta T\,(\min)}$

1.11. Different formulas for calculating empirical frequency:

Each distribution requires its own empirical probability formula, but there are no exact analytical formulas for calculating empirical frequencies. Formulas based on the median are independent of the parent distribution of the samples and are used in a standard manner.

The values are sorted in ascending or descending order, which gives us the frequencies of non-exceedance and exceedance, respectively, because in descending order, the rank value r is always exceeded by the rank value "i -1"; the same reasoning applies to frequencies of non-exceedance.

Once the values are sorted, the general formula for the empirical frequency is as follows:

F (i) = $\frac{i-a}{N+1-2a}$

The most commonly used empirical frequency formulas:

- California formula: This was one of the first formulas proposed and is expressed as follows:

F(xi) = $\frac{i_r}{N}$

With: i r: rank assigned to frequencies sorted in ascending or descending order.

N: sample size

- For a = 0.5, we find Hazen's formula: F(xi) = $\frac{i-0.5}{N}$

- For a = 0, we find the Weibull formula: F (r) = $\frac{i}{N+1}$

- For a = 0.3, we find Chegodayev's formula: F (r) = $\frac{i-0.3}{N+0.4}$

- For a = 0.4, we find Cunnane's formula: F (r) = $\frac{i-0.4}{N+0.2}$

- For a = 0.44, we find Gringorten's formula: F (r) = $\frac{i-0.44}{N+0.12}$

Chapter 2: Correlations and data analysis

2.1. Definitions

A time series, or chronological series, is a sequence of observations ordered in time, usually at equal intervals. Chronological series can be annual, quarterly, monthly, weekly, or daily.

2.2. Characteristic values of a chronological series

It is mathematically defined by the values y 1, y 2, ... of a variable y at times t 1, t 2, ...

The analysis of a hydrological series consists of describing the movements that compose it.

The decomposition of a time series is done in order to distinguish its evolution, i.e., the (general) trend, the seasonal variations that repeat each year, and the unpredictable accidental variations.

- **The trend**

Corresponds to the long-term evolution of the series, the fundamental evolution of the series (the estimation of the trend by least squares or by moving average).

- **Cyclical variations**

Seasonal variations are periodic fluctuations within the year, which recur more or less permanently from one year to the next (using the percentage of the average method, ratio to the trend, ratio to the moving average, etc.).

Once the seasonal data has been adjusted, it can be further adjusted according to the trend. A moving average taken over just a few months smooths out irregular variations and retains only cyclical variations.

- Accidental variations: accidental variations are irregular and unpredictable fluctuations. They are generally assumed to be of low amplitude.

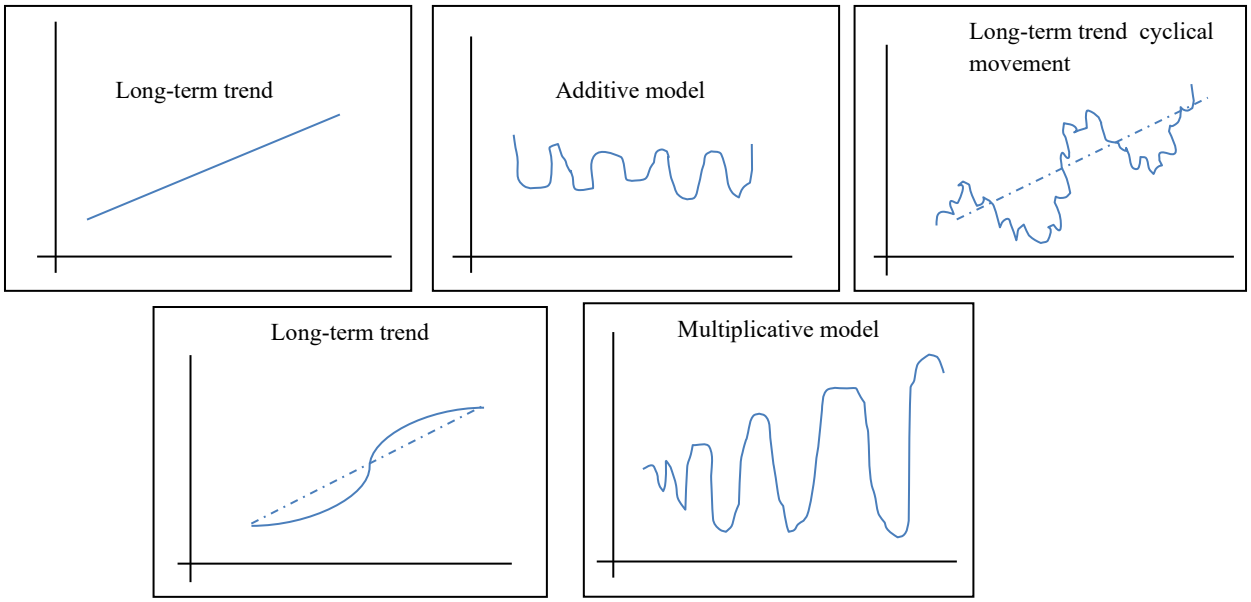


Figure 1: Movements in a time series

The main steps in processing a time series are as follows:

• Data correction

• Observation of the series

• Modeling

• Analysis of the series based on these components

• Prediction (= forecasting).

2.3. Seasonal coefficients

Seasonal coefficients allow seasonal variations to be taken into account in the forecast. Seasonal coefficients can be calculated from monthly, quarterly, or semi-annual data.

Seasonal coefficient $C_s = = \frac{moyenne\ du\ trimestre}{moyenne\ générale} = \frac{totale\ du\ trimestre}{total\ général}$

Corrected series : $y'_i = \frac{Y_i}{C_s}$ (case of the multiplicative model)

$y'_i = Y_i - C_s$ (additive model case)

The objective of seasonal adjustment (seasonally adjusted series) is to detect and estimate the effects. The resulting series is then said to be seasonally adjusted and comprises only the trend-cycle and the irregular component.

2.4. Forecasting based on extrapolation of past data

Analysis of historical data allows trends to be identified that can be used to extrapolate future data.

2.4.1. Smoothing methods (simple, exponential, and Winter)

The moving average method is a data smoothing technique. Its principle is to substitute a series of observed values with their average. This average is calculated by taking, for example, three values (we will say that these are moving averages of order 3), four values (moving averages of order 4), etc. Let us illustrate the principle of this method with the following example:

The moving averages of order 3,

denoted MM3, are calculated as follows

x_i	y_i
x_1	y_1
x_2	y_2
$x3$	y_3
x_4	y_4
x_5	y_5
x_6	y_6
x_7	y_7
x_8	y_8
x_9	y_9

x_i	$MM3$
x_1	
x_2	$\frac{y_1+y_2+y_3}{3}$
$x3$	$\frac{y_2+y_3+y_4}{3}$
x_4	$\frac{y_3+y_4+y_5}{3}$
x_5	$\frac{y_4+y_5+y_6}{3}$
x_6	$\frac{y_5+y_6+y_7}{3}$
x_7	$\frac{y_6+y_7+y_8}{3}$
x_8	$\frac{y_7+y_8+y_9}{3}$
x_9	

2.4.2. Multiplicative model and additive model

Observation of time series allows us to distinguish between two main types of series: those that conform to the multiplicative model and those that conform to the additive model. In the additive model, variations

around the trend remain within a fairly constant range (Fig. 1). In the multiplicative model, on the other hand, variations around the trend amplify (Fig. 2).

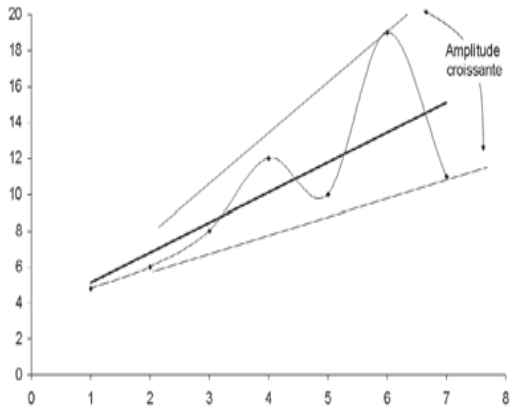


Figure 1. Multiplicative model

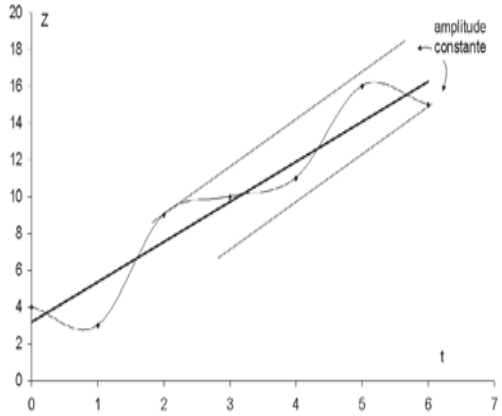


Figure 2. Additive model

In an additive model graph, the amplitude of the variations is constant around the trend and the two lines drawn are roughly parallel to each other.

Conversely, in a multiplicative model graph, the amplitude of the variations is not constant around the trend and the two lines drawn are not parallel to each other.

2.4.3. Exponential smoothing

Exponential smoothing methods are empirical methods for forecasting time series. They have the advantage of being easy to understand, and their recursive implementation makes them an effective tool for processing large volumes of data or in embedded systems with limited memory

- Simple exponential smoothing
- Double exponential smoothing (Holt);
- Holt-Winters exponential smoothing.

2.4.3.1 Simple exponential smoothing

Simple exponential smoothing can be used to make forecasts for time series with a constant trend and no seasonality. The basic idea is to calculate the forecast for time t+1 as the weighted average of:

- The last available observation (y t)
- The last calculated forecast ($\hat{y}$ t-1)

Using the following formula:

$$\hat{y}_t = \alpha^* y_t + (1- \alpha) \hat{y}_{t-1}$$

With $\hat{y} = y_1$ for an initial forecast. The smoothing constant α must be chosen from between 0 and 1, which gives a forecast that best fits the weighted least squares method.

The name "smoothing factor" given to α is misleading in that smoothing decreases as α increases and, in the limiting case where $\alpha = 1$, the smoothed series is identical to the raw series.

Values of α close to 1 reduce the impact of the past and give more weight to recent values; conversely, values close to 0 increase smoothing and reduce the impact of recent values.

2.4.3.2. Double exponential smoothing

Double exponential smoothing generalizes the idea of simple exponential smoothing to cases where the series can be fitted by a straight line in the vicinity of T. In this case, we seek a forecast at horizon h, Y(h), of the form:

$$\hat{y}_{T+h} = a_{0\,t} + a_{1\,t}\,h$$

The above formulas can be used to calculate a forecast for stationary time series without a trend. As its name suggests, the double smoothing technique involves smoothing a series that has already been smoothed.

$$S_t = a\,Y_t + (1 - a)\,s_{t-1}$$

$$S\,s_t = a\,S_t + (1- a)\,s\,s_{t-1}$$

$$a_{1\,t} = a/1\text{-}a^*(s_t - ss_t)$$

$$a_{0\,t} = 2\,s_t - ss_t$$

That is to say: $S_t = \hat{y}_t$: est la prévision par lissage simple

Ss_t : est le terme de lissage double.

2.4.3.3. Holt-Winters exponential smoothing

The double exponential smoothing method can be used to process series that show a linear trend but no seasonality.

This approach is a generalization of double smoothing, which allows the following models to be proposed, among others:

• local linear trend

• local linear trend + seasonality (additive model)

• local linear trend * seasonality (multiplicative model)

2.4.4. Linear correlation

A model is linear if its parameters and variables are linear.

The correlation coefficient (r) measures the degree of dependence between two random variables. The absence of correlation does not imply the absence of dependence.

Since (r) only evaluates linear dependence, a close curvilinear relationship is not necessarily reflected in a high value of (r). Conversely, the existence of a correlation between two variables does not necessarily imply that they are linked by a cause-and-effect relationship.

Let (x) and (y) be two variables under consideration. Using a scatter plot in a rectangular coordinate system, we can locate (x, y).

The equation for simple linear correlation is given by the following expression:

$$Y = a.\,X + b$$

With "n" sizes from series (x) and (y).

r: regression coefficient of the pair (x n, y n) which is equal to:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$a = r_n \cdot \frac{S_{n(y)}}{S_{n(x)}}$, $b = \sum \frac{(x-\bar{x})(y-\bar{y})}{(x-\bar{x})^2}$ Avec : $S_{n(y)} = \sigma_y^2$

Interpretation of the simple linear correlation coefficient

- If |r| is greater than 0.8, the correlation can be considered good, i.e., there is a causal link between the two variables. The closer "r" is to 1, the better the correlation and the stronger the link between the two phenomena.

- If 0.5 < |r| < 0.8, the correlation is moderate.

- If |r| < 0.5, the correlation is poor, there is no relationship between the two variables studied, and the closer r is to 0, the weaker the causal link.

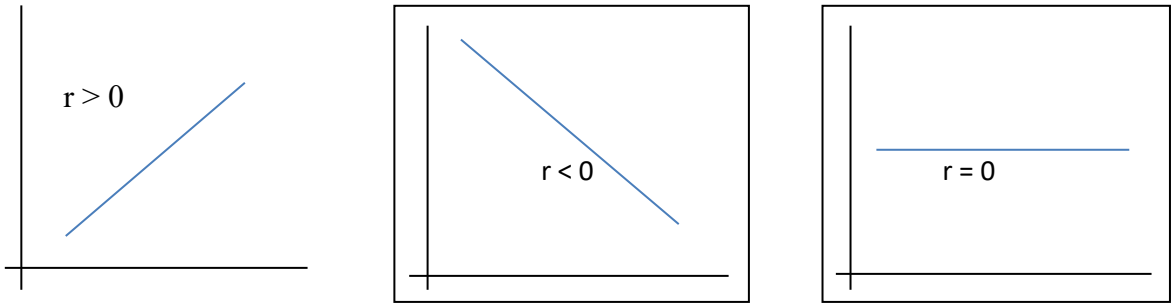


Figure 1: Different correlation coefficient values

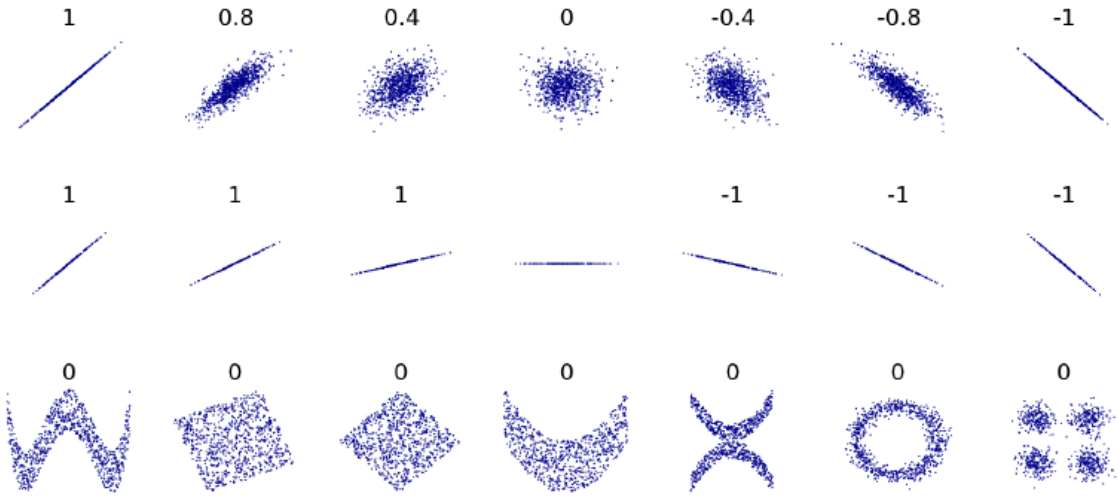


Figure 2. Examples of correlation coefficients: linear correlation for the first two lines, non-linear for the third line.

2.4.5. Multiple correlation

Multiple correlation refers to the correlation between three or more variables.

Let us consider a continuous variable and a set of P continuous explanatory variables denoted by X1, X2, X3, ….. X p. The multiple linear correlation coefficient between Y and X1, X2, X3, ….X p is defined by the maximum value of the (empirical) linear correlation "ϱ" between Y and a linear combination of the variables X 1, X 2, X 3,…..X p.

Applications

1- Reconstruction of missing data;

2- Forecasting models (low water levels, floods, etc.);

3- Data control.

Note 1:

Multiple regression is a generalization of the simple regression model when the explanatory variables are finite in number.

A linear regression model is defined by an equation:

$$\mathbf{Y}_{n+1} = \mathbf{X}_{n\,x\,p} + \beta_{p\,x\,1} + \varepsilon_{n\,x\,1}$$

That is to say:

$Y_i = \beta_0 + \beta_1 . X_{i1} + \beta_2 . X_{i2} + \ldots\ldots + \beta_n . X_{in} + \varepsilon_I$ (i = 1,2,…, n) it is a matter of finding the Y: is a random vector of dimension n.

X: is a known matrix of size (n x p), called the design matrix.

β: is the vector of dimension "p" of the unknown parameters of the model.

ε: is the centered vector of dimension (n) of the errors.

The matrix (x) of size (n X p) of the correlation coefficients: coefficients $\beta_0, \beta_1, \beta_2, \ldots\ldots \beta_n$ who to verify U $(\hat{\beta}_{0,}\beta_1) = \text{Min}\ \varepsilon\ (Y- (\beta_0 + \beta_1 . X_{i1} + \beta_2 . X_{i2}))^2$

$$X = R = \begin{bmatrix} 1 & r_{12\ldots\ldots} & r_{1p} \\ r_{21} & 1 & r_{2p} \\ \vdots & & \vdots \\ r_{n1} & \ldots\ldots & 1 \end{bmatrix}$$

Δ : the determinant of R.

Note:

$$Y_i = \beta_1 . X_{i1} + \beta_2 . X_{i2} + \ldots\ldots + \beta_n . X_{in} + \varepsilon_I$$

Matrix notation is:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} x_{11} & x_{12\ldots\ldots} & x_{1p} \\ x_{21} & x_{22\ldots\ldots} & \\ \vdots & & \\ x_{np} & \ldots\ldots & x_{nk} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_k \end{bmatrix}$$

The multiple linear regression coefficients for three variables **X 1**, **X 2**, and **X 3** is:

$$R_{1.\,23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2.r_{12}.r_{13}.r_{23}}{1 - r_{23}^2}}$$

2.5. Test of the multiple correlation coefficient

The Fisher or Student test is used to determine whether the multiple correlation coefficient is significantly different from zero.

Once the correlation coefficient (r) has been calculated, it can be examined using the Student test (t), where: t $= r\sqrt{\frac{N-2}{1-r^2}}$ where N: size of the hydrological series studied.

t: follows a Student's distribution with γ = N-2 degrees of freedom.

2.6. Different types of regressions (linear, power, exponential)

Exponential functions

1. Log-log or double log models

N Linear form (logarithmic transformation): on-linear form: $Q = AK^{a1}L^{a2}e^{u}$

Linear form (logarithmic transformation): $q = a0 + a1\ k + a2\ l + u$

With $q = \log(Q)$; $k = \log(K)$; $l = \log(L)$; $a0 = \log(A)$; $u = \log(e\ u)$; $a1\ et\ a2$ = Constants.

2. Nonlinear form: $D = A\ (\ /P\ )^{b}$

Linear form (logarithmic transformation): $\ln D = \ln A + b[\ln W - \ln P] \rightarrow$

$d = a + bw - bp$

With: $d = \ln$; $a = \ln A$; $w = \ln W$; $p = \ln P$

Power functions

$Y = a\ b^{x}$ ou $\log Y = \log a + X(\log b) = a_0 + a_1 X$

Polynomial functions

$Y = a_0 + a_1 X + a_2 X^2 + a_3 X^3$ cubic curve

2.7. Primary data quality analysis

The acquired data often requires primary processing in order to make it usable. It is necessary to convert the measurement to ⟦0.1⟧ and check the stationarity of the series.

Stationarity is examined using the Jarque-Bera test:

$$\gamma_{exp} = \frac{n}{6}\beta_1 + \frac{n}{24}(\beta_2 + 3)^2$$

Where: n is the size of the series, β1 is the skewness coefficient, and β2 is the kurtosis coefficient.

If $\gamma_{exp} > \gamma_{the}$ the, the experimental distribution follows the theoretical distribution (Student's table), meaning that the hydrological series follows the normal distribution (for α = 5%).

2.8. Normalization of variables and the concept of data transformation

That is, replacing a series of observations x 1, x 2, … x n with a transformed series y1, y2,…,y n.

With the same number of observers.

The choice of transformation is linked to the search for properties that are not immediately apparent. These properties include nonlinearity, symmetry, or variability.

2.8.1. Linear transformation

2.8.1. Transformation of origin

That is, we can replace the series of observations:

x_1 by $y_1 = x_1 \pm x_0$

x_2 by $y_2 = x_2 \pm x_0$

x_n by $y_n = x_n \pm x_0$

x_0: either the mean, mode, median, or others.

2.8.2. Unit conversion

Converting values in the series:

x_1 by $y_1 = \frac{x1}{x0}$

x_2 by $y_2 = \frac{x2}{x0}$

x_n by $y_n = \frac{xn}{x0}$

x_0 : either the mean, mode, median, or others.

2.8.3. Change of origin and unit

This refers to two transformations at the same time of origin and unit.

x_1 by $y_1 = \frac{x \pm x0}{\sigma}$

2.8.2. Functional transformation

To simplify the analysis by making the phenomenon more symmetrical.

Use of the TUKEY or Cox-Box family of simple transformations.

x by $\frac{1}{x}$, $\frac{1}{x^2}$, $\frac{1}{\sqrt{x}}$, log x , $\sqrt{x}$, x^2 , x^3 , e^x , x log x……

Log(x- x0) with x0: constitutes a lower bound for the variable "x"; sometimes zero is given and sometimes not.

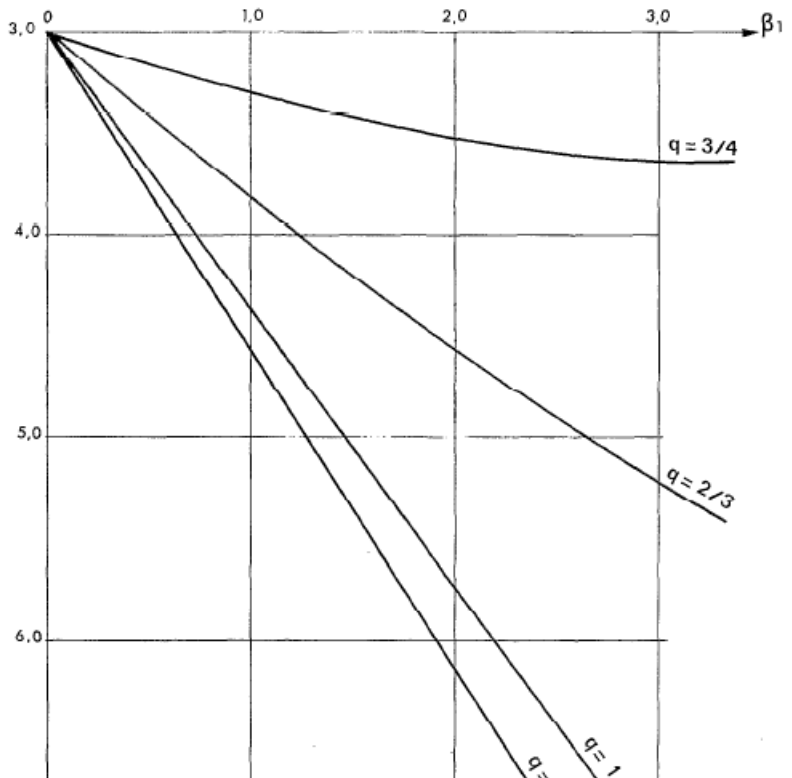
For X q: with X ≥ 0 and q being a number between 0 and 1.

The value of "q" can be determined based on the PEARSON coefficients β I and β 2 of variable X, using the chart established by Mr. MILU ROSENBERG based on theoretical considerations (see Chart).

Remember that β I and β 2 measure the degrees of asymmetry and flatness of a distribution, respectively.

Abacus: transformation by fractional

power q as a function of β I and β 2



2.9. Homogeneity Test

2.9.1. The Wilcoxon Test

This is a non-parametric test that uses the series of ranks of observations, rather than the series of their values, to verify the homogeneity of a series.

If the sample (of rain, for example) X comes from the same population Y, the sample X U Y (union of X and Y) also comes from it.

Let us consider a series of observations of length N from which we draw two samples X and Y: N1 and N2 are the sizes of these samples, respectively, with N = N1 + N2 and N1 ≤ N2.

Next, we classify the values in our series in ascending order. Subsequently, we will only be interested in the rank of each of the elements of the two samples in this series. If a value is repeated several times, we assign it the corresponding average rank.

We then calculate the sum W x of the ranks of the elements of the first sample in the common series:

W x = Σ Rank x.

Wilcoxon a constitute a homogeneous series, the quantity W x is between two limits W max and W min given by the following formulas:

$$W_{max} = \frac{(N_1 + N_2 + 1)}{N_1} - W_{min}$$

$$W_{min} = \frac{(N_1 + N_2 + 1).N_1 - 1}{2} - Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{N_1 N_2 (N_1 + N_2 + 1)}{12}}$$

$Z_{1-\frac{\alpha}{2}}$ Represents the value of the reduced centered variable of the normal distribution corresponding to $1-\frac{\alpha}{2}$

2.9.2. The median test (equality of the medians of two independent samples)

The median is the value that divides a series of data into two equal parts, so that half of the values are below it and half are above it.

This statistical test is used to compare the positions of two data sets; it is a non-parametric test, meaning that it does not assume that the data follow a given probability distribution.

The principle is to determine the median of all observations and to count the observations below and above this median for each of the two samples. This produces a 2 x 2 contingency table, from which either a corrected χ^2 test (known as YATE's continuity correction) can be performed.

The initial observations are therefore subdivided into two samples (1 and 2) relative to the overall median.

$$\chi^2_{corrigé} = \frac{N(|AD-BC|-{}^{N}/_{2})^2}{(A+B)(C+D)(A+C)(B+D)}$$

with degree of freedom equal to 1 (df = 1) ($\alpha = 0.01$)

N: overall median of the set of observations.

2 x 2 contingency table

	sample 1	sample 2	total
x < overall median	A	B	A+B
x ≥ overall median	C	D	C+D
Total	A+C	B+D	N

A: absolute frequency lower than the overall median for sample 1.

B: absolute frequency lower than the overall median for sample 2.

C: absolute frequency greater than or equal to the overall median for sample 1.

D: absolute frequency greater than or equal to the overall median for sample 2.

If $\chi^2_{corrigé} > \chi^2_{du\ tableau}$ From Khi 2, we accept the hypothesis that the medians of the two samples are equal.

Note:

There is a simple formula for calculating the degrees of freedom of a table.

dll = (number of rows - 1) x (number of columns – 1)

Or the number of degrees of freedom is $\gamma = k - 1 - r$

Where: k: number of classes, r: number of parameters that exactly define the theoretical distribution (normal distribution in our case) = 2.

2.10. Conformity test

2.10.1 Mann Whitney Z test (applied to independent samples)

This test is used to test the H0 hypothesis that a hydrological series is homogeneous.

We divide our sample into two subsets of size n1 (x) and n2 (y) with n2 > n1

The size of the original sample is n = n1 + n2

We then rank our values in ascending order from 1 to n and note the ranks R(x) of the elements of the first subset and those R(y) of the elements of the second subset.

We define K and S as follows:

$$K = L - \frac{n_1 (n_1+1)}{2}$$

And: S = n1. n2 – k

With $L = \sum_{i=1}^{n_1} R(x_i)$ sum of the ranks of the elements of sample 1 in the original sample

K: sum of the numbers of exceedances of each element of the second sample by those of the first sample.

S: sum of the numbers of exceedances of the elements of the first sample by those of the second.

We can test the hypothesis H0 that the two samples come from the same population at the significance level.

We compare the value "T" with the reduced centered normal variable with a probability of exceeding $\alpha/2$

$$\text{With } T = \left|\frac{K-\overline{K}}{S_k}\right|$$

And $\overline{K} = \frac{n1.n2}{2}$ named the average and $S_k = \frac{n1.n2}{12}(n1 + n2 + 1)$

1	2	3	4	5	6	7	8	9	10	11
ordr	Original series	sorted series	sample 1	Rank 1	sample 2	Rank 2	sample 1 sorted	Number exceeded	sample 2 sorted	Number exceeded
1										
2										

Column 1 corresponds to the ranks of the data (sorted);

Columns 2 and 3 indicate the original series and the series sorted in ascending order, respectively;

Columns 4 and 6 correspond to subsamples (x) and (y), respectively;

Columns 5 and 7 correspond to the ranks of subsamples (x) and (y) in the original series in column 2, respectively;

Column 8 shows the sorted values of subset 1;

Column 9 indicates the number of times each element of subset 1 is exceeded by the elements of subset 2;

Column 10 shows the values of subset 2 sorted;

Column 11 indicates the number of times each element of subset 2 is exceeded by the elements of subset 1.

2.10.2 Student's t-test

The Student's t-test is used to compare two independent samples.

Before performing parametric tests, you must:

- Ensure that the sample distribution is compatible with the assumption of Gaussian distribution of the variable (normality test). If not, you can try to make this distribution compatible with a Gaussian distribution by performing a transformation, for example a logarithmic transformation.
- Check the homogeneity of the variances of all samples;

2.11. Checking the homogeneity of variances.

A test of equality of variances allows you to check the equality of variances between populations. Many statistical methods, such as analysis of variance (ANOVA) and regression.

Chapter 3: Hydrological modeling

The term "hydrological system modeling" usually refers to the use of mathematical and logical expressions that define relationships between output characteristics and their conditional factors (inputs).

A model is a simplified representation of reality that aims to translate the mechanisms of the phenomenon under study and provide a better understanding of them. One model may be better than another at describing reality, of course.

3.1 General information on hydrological modeling

Regardless of the type of representation used, models are generally composed of different modules that are often independent. Each module is designed to take into account the main processes of the hydrological cycle, such as precipitation, evaporation, infiltration, surface runoff, subsurface runoff, and baseflow, as well as the volume of water flowing through the river at each time step and over each element of the watershed.

3.2 Different modeling approaches

3.2.1 Types of models (conceptual, empirical, physics-based, etc.)

a) Conceptual hydrological models

Conceptual models are simplified representations that often do not correspond to physical reality. They consider watersheds as a collection of several connected reservoirs. Changes over time in reservoir levels are due to interactions between the reservoirs and the atmosphere (evapotranspiration, precipitation), which makes it possible to determine the outflow (discharge at the watershed outlet). They attempt to reproduce basic processes such as interception, infiltration, evaporation, and surface runoff. They require calibration using measured flows in order to mimic hydrological behavior. Their parameters may have a conceptual interpretation, but often these parameters are not measurable and have no physical meaning. Conceptual models are used for flood forecasting.

b) Physically based hydrological models

Physically based hydrological models emerged in the 1980s; they are based on a detailed description of the entire hydrological functioning of a watershed. They are based on physical laws, such as Saint-Venant's equations for surface flow in watercourses, Darcy's law for saturated groundwater flow, and Richards' or Boussinesq's equation for unsaturated groundwater flow.

They are characterized by measurable parameters that describe the physical properties of the environment. From a practical standpoint, several constraints limit their use at the watershed scale, such as their very long calculation time, availability, the number and quality of the data required, and the large number of parameters needed.

Models can also be event-based or continuous, depending on the modeling objectives. In situations where the aim is to describe a given meteorological event without taking into account the entire hydrological process, event-based models are generally sufficient, unlike continuous integral models. Continuous models can be used to simulate the evolution of state variables over several hydrological years. The main difference from event-based models is that certain processes that can be neglected during an event cannot be neglected in a continuous simulation because they become predominant between events.

c) Empirical models

Empirical models are based on the observed relationships between inputs and outputs of the hydrological system under consideration. They express the relationship between the system's input and output variables (rain-runoff relationship) using a set of equations developed and adjusted on the basis of data obtained from the system.

An empirical model does not seek to describe the causes of the hydrological phenomenon under consideration or to explain how the system works; the system is considered a black box.

Note

The model has two main functions aimed at describing as accurately as possible the flow of water to the outlet of a watershed:

- The production function
- The transfer function.

The production function concerns the “vertical” flow of water, the main phenomena of which are:

• Precipitation

• Snowmelt

• Evapotranspiration

• Infiltration

• The interaction of surface and deep reserves.

The production function is calculated for each entire tile on a daily basis.

The transfer function concerns "horizontal" flow in the drainage network. The processes included in this section also take into account the influence of lakes, marshes, and artificial structures such as dams, diversions, etc. The transfer function is performed using partial tiles.

3.2.2 Production functions

The production function is used to determine the portion of gross rainfall (called net or effective rainfall) that will contribute to runoff and discharge at the outlet. It seeks to represent the interaction between the soil and the precipitated water layer in order to determine how much of the rainfall will be stored in the soil, how much will run off, and under what conditions.

The production functions commonly used in flow modeling are:

• The 4-parameter rural engineering function, GR4.

• The Soil Conservation Service SCS function, which depends on a CN parameter called Curve Number. This describes the infiltration capacity of the soil, i.e., its ability to create runoff.

• The Infiltration Variable, which assumes a variation in infiltration capacity depending on the topography of the terrain and changes in soil and vegetation.

During a rainy event, part of the watershed becomes saturated with water as the water table rises to the surface, thus limiting infiltration.

3.2.3 Transfer functions

Several methods exist for determining the transfer function of a basin. The most complex use Barre Saint Venant equations with diffusive wave or kinematic wave models.

The purpose of the transfer function is to transfer the effective rainfall available for runoff from each unit element to the watershed outlet. In the geomorphological approach, the shape of the watershed is taken into account to calculate this transfer. The travel time of rainwater for each unit element is estimated and depends on the total length of the water's path, the dimensions and shape of the section, and the slope of each section traveled. The travel times of all unit elements are used to determine the transfer function, which represents the travel time as a function of the weighted area.

3.2.4 Presentation of some watershed models (GR, HBV)

a) The HBV (Hydrologiska Byråns Vattenbalansavdelning) hydrological model

This is a comprehensive conceptual model developed by the Swedish Meteorological and Hydrological Institute. It is used to simulate lake water flow and levels, analyze river flow, and assess pollution.

It is based on the concept of storage reservoirs, which can operate on a daily or monthly basis. This model assumes that parameters do not change over space. The model structure is based on two superimposed reservoirs.

Time series of observed temperatures and precipitation are required for each stage of the model, as well as monthly estimates of potential evapotranspiration and altitude for spatial discretization in homogeneous areas of the basin.

Precipitation is treated as snow or temperature-dependent precipitation at each stage of the day.

Potential evapotranspiration is calculated by the proposed model. Hydrological processes are subdivided into four different main components; the first is related to snow accumulation and snowmelt; the second to effective precipitation and soil; the third to groundwater; and a fourth for calculating flood propagation (table).

Modèle	Paramètres libres	Min	Max	Unité
HBV	K0 : Coefficient puissance de vidange du réservoir supérieur	*0.05*	0.2	jour^{1}
	Béta : Paramètre puissance du modèle (infiltration)	*1*	7	Adimensionnel
	C : Coefficient de correction de l'évapotranspiration	*0.01*	0.07	1/°C
	DD : Quantité d'eau provenant de la neige accumulée pour 1°C au-dessus de 0°C	*3*	7	mm/jour/°C mm/jour/°mm
	FC : Seuil de saturation du sol	*100*	200	mm
	L : Seuil du ruissellement direct	*2*	5	mm
	K1 : Coefficient de vidange dans le réseau hydrographique	*0.01*	0.1	jour^{-1}
		0.01	0.05	jour^{-1}
	K2 : Coefficient de vidange dans le réseau	*0.01*	0.05	jour^{-1}
	Kp : Coefficient de percolation	*90*	180	mm
	PWP : Seuil à partir duquel l'évapotranspiration est potentielle			

b) The GR Hydrological Model of Rural Engineering

These models establish the link between the amount of precipitation on a watershed and its discharge at the outlet. They raise essential questions about how to represent the transformation of rainfall into discharge at the watershed scale. The GR model is a conceptual daily hydrological model with four free parameters. It uses a production reservoir and another routing reservoir to simulate the outgoing flow based on two input data: daily precipitation and potential evapotranspiration.

Potential evapotranspiration has been calculated using an empirical model based on daily average temperature and extraterrestrial radiation.

3.3 Calibration and Validation

The optimization of the model's performance and validity is typically carried out using statistical performance criteria that analyze the residual errors between observed and predicted values.

Model calibration involves selecting a representative calibration period and a set of parameters to calibrate, choosing an appropriate evaluation technique, and implementing an actual iterative procedure. Calibration can include one or more parameters and may be performed by applying objective functions. Studies have also shown that a smaller number of parameters generally provides more accurate model performance while increasing the information content of each parameter.

Quality and Optimization Criteria

Rainfall-runoff models are practical tools for research in hydrology and hydraulic engineering. Several evaluation criteria have been developed, which are either graphical or analytical. The most commonly used in hydrology include:

Nash-Sutcliffe Criterion

The most common quality criterion is the Nash-Sutcliffe efficiency (Nash & Sutcliffe, 1970). It is used for optimization and parameter identification.

$$\text{NS} = \left[1 - \frac{\sum_{i=1}^{n}(x_i^{obs} - x_i^{sim})^2}{\sum_{i=1}^{n}(x_i^{obs} - \bar{x}_i^{obs})^2}\right] x100$$

x_i^{obs} : Observed daily variable on day iii

x_i^{sim} : Calculated daily variable on day iii

$\bar{x}_i^{obs}$: Mean observed variable over the simulation period

N : Number of time steps in the simulation

The Root Mean Square Error (RMSE) provides information similar to that of the Nash-Sutcliffe criterion. The formula is as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i^{obs} - x_i^{sim})^2}$$

The mean error (ME) provides information about the model's balance. The formula is as

$$\text{ME} = \frac{1}{n}\sum_{i=1}^{n}(x_i^{obs} - x_i^{sim})$$