STOCHASTIC METHODS IN WATER RESOURCES

Unit 2: Hydrological statistics and extremes
Lecture 6: Probability distributions of extremes, distribution selection

Luis Alejandro Morales, Ph.D.

Universidad Nacional de Colombia
Department of Civil and Agriculture Engineering

August 16, 2025

Gumbel distribution

The **Gumbel distribution** is equivalent to the to asymptotic distribution type I and result from a basic exponential distribution. The Gumbel distribution has two parameters: the scale parameter α and the position parameter b, and is given for maximum values represented by X, where:

$$f_X(x) = \frac{1}{\alpha} e^{\left[-\frac{x-b}{\alpha} - e^{\left(-\frac{x-b}{\alpha}\right)}\right]}, \text{ for } -\infty < x < \infty$$

The cdf is:

$$F_X(x) = e^{\left[-e^{\left(-\frac{x-b}{\alpha}\right)}\right]}$$

and $\mu_X = b + 0.5772\alpha$ and $\sigma_X^2 = \frac{\pi^2\alpha^2}{6}$ and the asymmetry coefficient $\gamma_1 = 1.1396$. The quantile of this distribution is:

$$x(F) = b - \alpha \ln(-\ln F)$$

where $x(F)$ is the value of X for which the cdf value is F. This equation is equivalent to:

$$x_T = b - \alpha \ln\left[-\ln\left(1 - \frac{1}{T}\right)\right]$$

where T is the return period and x_T is the magnitude of the extreme event associated to this.

Gumbel distribution

The parameter estimation in the Gumbel distribution can be performed using different methods such as the Moments and the Maximum-Likelihood. The frequency factor K_T for this distribution when $n \to \infty$, converge asyntotically to:

$$K_T = -\frac{\sqrt{6}}{\pi}\left[0.5772 + \ln\left(\ln\left(1 - \frac{1}{T}\right)\right)\right]$$

On the other hand, for a value of the annual series, the frequency factor as a function of the size series, can be estimated as:

$$K_m = \frac{y_m - \bar{y}}{s_y}$$

where:

$$y_m = -\ln\left[-\ln\left(\frac{n+1-m}{n+1}\right)\right]$$

Note that the data are sorted in descending order; from the largest $(m = 1)$ to the smallest $(m = n)$. $\bar{y} = \frac{1}{n}\sum_{m=1}^{n} y_m$ and $s_y^2 = \frac{1}{n}\sum_{m=1}^{n}(y_m - \bar{y})^2$. For a return period T, the frequency factor is:

$$K_T = \frac{y_T - \bar{y}}{s_y}$$

where $y_T = -\ln\left[-\ln\left(1 - \frac{1}{T}\right)\right]$. Comparing the two equations to compute K_T the later equations provide larger results of K_T than the former one. This value is progressively larger as much as n decrease. This means that the shorter the series the larger the values of x_T.

Gumbel distribution

Estimation of the standard deviation s_T using different methods:

- Method of Moments

$$s_T^2 = \frac{s_X^2}{n}\left[1 + K_T\gamma_1 + \frac{K_T^2}{4}(\gamma_2 - 1)\right]$$

  where $\gamma_1 = 1.1396$ and $\gamma_2 = 5.4002$

- Method of Maximum-Likelihood

$$s_T^2 = \frac{\hat{\alpha}^2}{n}\left[1 + \frac{6}{\pi^2}(1 - 0.5772 + y_T)^2\right]$$

- Method of MPP

$$s_T^2 = \frac{\hat{\alpha}^2}{n}\left[1.1128 + 0.4574y_T + 0.8046y_T^2\right]$$

Gumbel distribution

So far, we have analysed annual series of maximum values using the Gumbel distribution. However, this distribution can be used to analyzed minimum values Z so the pdf is:

$$f_Z(z) = \frac{1}{\alpha} e^{\left[\frac{z-b}{\alpha} - e^{\left(\frac{z-b}{\alpha}\right)}\right]}, \text{ for } -\infty < z < \infty$$

and the cdf is:

$$F_Z(z) = 1 - e^{\left[-e^{\left(\frac{x-b}{\alpha}\right)}\right]}$$

The $\mu_Z = b - 0.5772\alpha$, $\sigma_Z^2 = \frac{\pi^2\alpha^2}{6}$ and the asymmetry coefficient is $\gamma_1 = -1.1396$. The quantile is:

$$z(F) = b + \alpha \ln(-\ln(1-F))$$

where $z(F)$ is the value of Z for which the cdf value is F. This equation is equivalent to:

$$z_T = b + \alpha \ln\left[-\ln\left(1 - \frac{1}{T}\right)\right]$$

Similarly, the parameter of the distribution can be estimated using different methods.

Gumbel distribution

When $n \to \infty$ asymptotically, the frequency factor K_T converge to :

$$K_T = \frac{\sqrt{6}}{\pi}\left[0.5772 + \ln\left(-\ln\left(1-\frac{1}{T}\right)\right)\right]$$

On the other hand, for a value of the annual series, the frequency factor as a function of the size series, can be estimated as:

$$K_m = \frac{\bar{y} - y_m}{s_y}$$

where:

$$y_m = \ln\left[-\ln\left(\frac{m}{n+1}\right)\right]$$

Note that the data are sorted in ascending order; from the smallest $(m = 1)$ to the largest $(m = n)$. $\bar{y} = \frac{1}{n}\sum_{m=1}^{n} y_m$ and $s_y^2 = \frac{1}{n}\sum_{m=1}^{n}(y_m - \bar{y})^2$.
For a return period T, the frequency factor is:

$$K_T = \frac{\bar{y} - y_T}{s_y}$$

where $y_T = -\ln\left[-\ln\left(-\frac{1}{T}\right)\right]$.

Pearson type III and Log-Pearson type III distributions

The pdf of the **Pearson type III distribution** is:

$$f_X(x) = \frac{1}{a\Gamma(b)} \left(\frac{x-c}{a}\right)^{b-1} e^{\left[-\left(\frac{x-c}{a}\right)\right]}$$

where the parameters a, b and c are the scale, the shape and the position, respectively. Note that $b > 0$ and that $c < x < \infty$. If $c = 0$, the pdf is reduced to the Gamma distribution. Despite a can be either positive or negative, if $a < 0$ the pdf is bounded above, this means that $x < c$ and it is not suitable to analyse maximum extreme events. Using the methods of Moments:

- Mean

$$\mu_X = c + ba$$

- Variance

$$\sigma_X^2 = ba^2$$

- Asymmetry coefficient

$$\gamma_1 = sgn(a)\frac{2}{\sqrt{b}}$$

  where *sgn* is the sign function.

Sometimes, the annual series are transformed using the logarithm function and the data is fitted to the Pearson type III distribution, which is equivalent to the **Log-Pearson type III distribution**. This distribution is commonly used to analyse maximum streamflows and requires that $b > 1$ and $a > 0$. The equations to estimate the frequency factor are given in textbook tables.

Weibull distribution

The **Weibull distribution** is equivalent to the asymptotic type III distribution, and is popular to the analysis of annual series of minimum values Z. The pdf is defined as:

$$f_Z(z) = \left(\frac{\kappa}{\alpha}\right)\left(\frac{z-b}{\alpha}\right)^{\kappa-1} e^{\left[-\left(\frac{z-b}{\alpha}\right)^{\kappa}\right]}$$

where $z \geq b$, $\alpha > b$ and $\kappa > 0$. The cdf is:

$$F_Z(z) = 1 - e^{\left[-\left(\frac{z-b}{\alpha}\right)^{\kappa}\right]}$$

Some expressions are:

- Mean

$$\mu_Z = b + \alpha\Gamma\left(1 + \frac{1}{\kappa}\right)$$

- Variance

$$\sigma_Z^2 = \alpha^2\left[\Gamma\left(1+\frac{2}{\kappa}\right) - \Gamma^2\left(1+\frac{1}{\kappa}\right)\right]$$

- Asymmetry coefficient

$$\gamma_1 = \frac{\Gamma\left(1+\frac{3}{\kappa}\right) - 3\Gamma\left(1+\frac{1}{\kappa}\right)\Gamma\left(1+\frac{2}{\kappa}\right) + 2\Gamma^3\left(1+\frac{1}{\kappa}\right)}{\left[\Gamma\left(1+\frac{2}{\kappa}\right) - \Gamma^2\left(1+\frac{1}{\kappa}\right)\right]^{1.5}}$$

where $\Gamma(.)$ is the Gamma function.

The inverse of the Weibull cdf is:

$$z(F) = b + \alpha\left[-\ln\left(1-F\right)\right]^{1/\kappa}$$

Note that when $\kappa = 1$ this distribution becomes the exponential distributio. Other properties of the Weibull distribution are given in textbooks tables.

Fréchet distribution

The **Fréchet distribution** is equivalent to the asymptotic type II distribution. The pdf is defined as:

$$f_X(x) = \frac{\theta}{a}\left(\frac{a}{\theta}\right)^{\theta+1} e^{\left[-\left(\frac{a}{x}\right)^{\theta}\right]}$$

where $x > 0$, and the scale and shape parameters are $a > 0$ and $\theta > 0$, respectively. The cdf is:

$$F_X(x) = e^{\left[-\left(\frac{a}{x}\right)^{\theta}\right]}$$

Some expressions are:

- Mean

$$\mu_X = a\Gamma\left(1 - \frac{1}{\theta}\right)$$

  valid when $\theta > 1$

- Variance

$$\sigma_X^2 = a^2\left[\Gamma\left(1 - \frac{2}{\theta}\right) - \Gamma^2\left(1 - \frac{1}{\theta}\right)\right]$$

$$V_X = \left[\frac{\Gamma(1 - 2/\theta)}{\Gamma^2(1 - 1/\theta)} - 1\right]^{0.5}$$

  where the functions for σ_X^2 and V_X are valid for $\theta > 2$.

The quantile for the Fréchet pdf is given:

$$x(F) = a(-\ln F)^{-\theta}$$

Regarding to the exponential positive shape for $\theta > 0$, $x(F)$ increases faster than the Gumbel distribution when F increase. This means that Fréchet distribution produces a frequency curve of larger magnitudes. These two distributions are related through the logarithm transform, because when X follow a Fréchet pdf with parameters a and θ, $Y = \ln(X)$ follow a Gumbel pdf with parameters $\alpha = 1/\theta$ and $b = \ln(a)$. Other properties of the Fréchet distribution are given in textbooks tables.

GEV distribution

The **Generalized Extreme Value (GEV) distribution** is used for extreme maximum values, usually meteoroligical data. The pdf is:

$$f_X(x) = \frac{1}{a}\left[1 - \frac{b}{a}(x - c)\right]^{(1-b)/b} e^{-\left[1-\frac{b}{a}(x-c)\right]^{1/b}}$$

and the cdf is:

$$F_X(x) = e^{-\left[1-\frac{b}{a}(x-c)\right]^{1/b}}$$

where $a > 0$, b and c are the parameters of scale, shape and position, respectively. If $b > 0$ the distribution is bounded above representing a type III distribution which is useful for the analysis of extreme mimimum events. In contrast, if $b < 0$ the distribution is bounded below equivalent to a type II distribution suitable for extreme maximum events. If $b = 0$, the GEV distribution is transformed to a Gumbel distribution.

The fist two moments are:

- Mean

$$\mu_X = c + \frac{a}{b}[1 - \Gamma(1 + b)]$$

- Variance

$$\sigma_X^2 = \frac{a^2}{b^2}[\Gamma(1 + 2b) - \Gamma^2(1 + b)]$$

where $\Gamma(r)$ is the Gamma function evaluated in r. Note that while the equation for μ_X is valid for $b > -1$, the equation for σ_X^2 is valid for $b > -0.5$. A quantile is computed as:

$$x(F) = c + \frac{a}{b}\left[1 - (-\ln F)^b\right]$$

Other properties of the GEV distribution are given in textbooks tables.

Pareto distribution

The **Pareto distribution** is commonly used to analysed variables with long, straight tails. This distribution was initially implemented to analysed the wealth distribution in society where a low percentage concentrate the largest wealth. The pdf classic Pareto distribution or Pareto I distribution is:

$$f_x(x) = a_l c_l^{a_l} x^{a_l - 1}$$

where $x > c_l$ and $a_l > 0$. c_l and a_l are the position and shape parameters, respectively. The cdf is:

$$F_X(x) = 1 - \left(\frac{c_l}{x}\right)^{a_l}$$

The fist two moments are:

- Mean

$$\mu_X = \frac{c_l a_l}{a_l - 1}$$

  Note that this is valid when $a_l > 1$, and ∞ on the contrary.

- Variance

$$\sigma_X^2 = \left(\frac{cl}{a_l - 1}\right)^2 \frac{a_l}{a_l - 2}$$

  when $a_l > 2$. If $1 < a_l \leq 2$, $\sigma_X^2 = \infty$, and not exist when $a_l \leq 1$.

Pareto distribution

The Pareto I distribution is embeded in a more general Pareto II distribution. The pdf of this distribution is:

$$f_X(x) = \begin{cases} \frac{1}{b}\left(1 - a\frac{x}{b}\right)^{\frac{1}{a}-1} & \text{for } a \neq 0 \\ \frac{1}{b}e^{\left(-\frac{x}{b}\right)} & \text{for } a = 0 \end{cases}$$

The cdf is:

$$F_X(x) = \begin{cases} 1 - \left(1 - a\frac{x}{b}\right)^{\frac{1}{a}} & \text{for } a \neq 0 \\ 1 - e^{\left(-\frac{x}{b}\right)} & \text{for } a = 0 \end{cases}$$

where b is the scale parameter. Including a third parameter c, the position parameter, the pdf of the Pareto III distribution is:

$$f_X(x) = \begin{cases} \frac{1}{b}\left(1 - a\frac{x-c}{b}\right)^{\frac{1}{a}-1} & \text{for } a \neq 0 \\ \frac{1}{b}e^{\left(-\frac{x-c}{b}\right)} & \text{for } a = 0 \end{cases}$$

The cdf is:

$$F_X(x) = \begin{cases} 1 - \left(1 - a\frac{x-c}{b}\right)^{\frac{1}{a}} & \text{for } a \neq 0 \\ 1 - e^{\left(-\frac{x-c}{b}\right)} & \text{for } a = 0 \end{cases}$$

Note that when $c = 0$, the Pareto III distribution becomes the Pareto II distribution. Also, when $a < 0$ and $c = 0$ becomes the Pareto I distribution with $a_I = 1/a$ and $c_I = -b/a$.

Frequency curves for minimum events

- The commonly implemented probability distributions to analysed annual series of minimum events are: Weibull, Gumbel, Lognormal, Pearson type III and GEV. In the case of the Gumbel distribution, which is a unbounded distribution, can cause problem because negative values of the analysed variable can arise that are not realistic as is the case of some hydrological variables.
- These distribution can be fitted to the annual series of minimum values $z_i, i = 1, \cdots, n$ using a data transformation, that consist in multiplying the data by -1, which means that $x_i = -z_i$. This means that low values becomes high values and vice-versa. Thus, the fit of the probability distribution is made based on a annual series of maximum values x_i. At the end, the quantiles estimated for the frequency curve must be multiplied by -1 to return to the domain of Z.

Fitting process

The following steps summarize the process of fitting a probability distribution to an annual series.

1. Establishment of the annual series after applying statistical tests to detect outliers, and verify independence, homogeneity and stationarity in the series. Recall, that the size of the series affect in the estimation of the frequency curve and in the associated uncertainty. The literature thus recommends that the number of values of the series must be 25-30.
2. Selection of the pdf $f_X(x)$
3. Selection of the methods for parameter estimation $\hat{\theta}$
4. Estimation of magnitudes of events for selected return periods T, which depends upon the cdf $F_X(x)$ and the estimated parameters $\hat{\theta}$. Note that the T values should be less than 3-4 times the length (number of years) of the series. For maximum events:
$$\hat{X}_T = F_X^{-1}\left[1 - \frac{1}{T}\right]$$
and for minimum events:
$$\hat{X}_T = F_X^{-1}\left[\frac{1}{T}\right]$$
5. Estimation of the confidence limits for X_T for a significance level α. This means, confidence limits for the frequency curve.
6. Check through Q-Q plots and plots of the frequency curve if the chosen distribution is compatible with the annual series. Also, the use of the moments can verify the correcness of the chosen distribution with respect to the annual series. For instance, when γ_1 is close to zero, the Normal distribution could be considered. On the contrary, other distribution could be more appropriate. Also, hypothesis tests can be applied such as the Chi-square, Kolmogorov-Smirnov and Anderson-Darling, to verify the goodness of fit.

Anderson-Darling test

The **Anderson-Darling test** to verify the goodness of fit of a distribution is more adequate for the analysis of extreme events, because the test give more weight to the distribution tails. To perform this test on the following distributions: Gumbel, Fréchet, Normal, Lognormal, GEV, Gamma, Pearson type III and Log-Pearson type III, the first requirement is that the parameters estimators must be asymptotically efficient, usually estimated using the maximum likelihood method. For the Fréchet, Lognormal and Log-Pearson type III the annual series must be transformed taking logarithms and then fitted the transformed series to the Gumbel (max), Normal and Pearson type III distributions, respectively. After adjusting the data to any of these distributions using the maximum likelihood method the test statistics A^2 is estimated using:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln(F_0(x_i)) + \ln(1 - F_0(x_{n-i+1}))\right]$$

where x_i are sorted in ascending order and $F_0(x_i)$ is the cdf evaluated in x_i. This is followed by the estimation of the parameter ω:

$$\omega = 0.116\left(\frac{A^2 - \xi_n}{\beta_n}\right)^{1.1751\eta_n} + 0.0403 \qquad \text{for } \xi_n \leq A^2$$

$$\omega = \left[0.116\left(\frac{0.2\xi_n}{\beta_n}\right)^{1.1751\eta_n} + 0.0403\right]\left[\frac{A^2 - 0.2\xi_n}{\beta_n}\right] \qquad \text{for } \xi_n > A^2$$

where the parameters ξ_n, β_n and η_n are computed using the expressions in the following table for different distributions and according to the size n of the annual series.

Anderson-Darling test

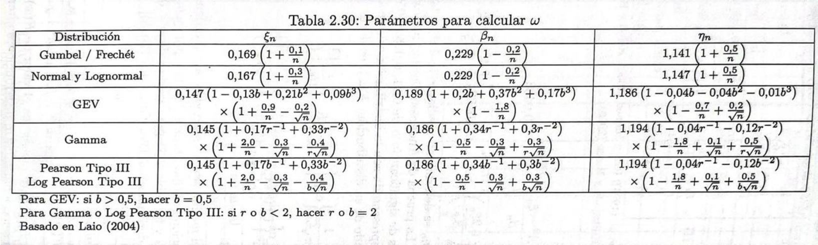

Tabla 2.30: Parámetros para calcular ω

Distribución	ξ_n	β_n	η_n
Gumbel / Frechét	$0{,}169\left(1+\frac{0{,}1}{n}\right)$	$0{,}229\left(1-\frac{0{,}2}{n}\right)$	$1{,}141\left(1+\frac{0{,}5}{n}\right)$
Normal y Lognormal	$0{,}167\left(1+\frac{0{,}3}{n}\right)$	$0{,}229\left(1-\frac{0{,}2}{n}\right)$	$1{,}147\left(1+\frac{0{,}5}{n}\right)$
GEV	$0{,}147\left(1-0{,}13b+0{,}21b^2+0{,}09b^3\right)$ $\times\left(1+\frac{0{,}9}{n}-\frac{0{,}2}{\sqrt{n}}\right)$	$0{,}189\left(1+0{,}2b+0{,}37b^2+0{,}17b^3\right)$ $\times\left(1-\frac{1{,}8}{n}\right)$	$1{,}186\left(1-0{,}04b-0{,}04b^2-0{,}01b^3\right)$ $\times\left(1-\frac{0{,}7}{n}+\frac{0{,}2}{\sqrt{n}}\right)$
Gamma	$0{,}145\left(1+0{,}17r^{-1}+0{,}33r^{-2}\right)$ $\times\left(1+\frac{2{,}0}{n}-\frac{0{,}3}{\sqrt{n}}-\frac{0{,}4}{r\sqrt{n}}\right)$	$0{,}186\left(1+0{,}34r^{-1}+0{,}3r^{-2}\right)$ $\times\left(1-\frac{0{,}5}{n}-\frac{0{,}3}{\sqrt{n}}+\frac{0{,}3}{r\sqrt{n}}\right)$	$1{,}194\left(1-0{,}04r^{-1}-0{,}12r^{-2}\right)$ $\times\left(1-\frac{1{,}8}{n}+\frac{0{,}1}{\sqrt{n}}+\frac{0{,}5}{r\sqrt{n}}\right)$
Pearson Tipo III Log Pearson Tipo III	$0{,}145\left(1+0{,}17b^{-1}+0{,}33b^{-2}\right)$ $\times\left(1+\frac{2{,}0}{n}-\frac{0{,}3}{\sqrt{n}}-\frac{0{,}4}{b\sqrt{n}}\right)$	$0{,}186\left(1+0{,}34b^{-1}+0{,}3b^{-2}\right)$ $\times\left(1-\frac{0{,}5}{n}-\frac{0{,}3}{\sqrt{n}}+\frac{0{,}3}{b\sqrt{n}}\right)$	$1{,}194\left(1-0{,}04r^{-1}-0{,}12b^{-2}\right)$ $\times\left(1-\frac{1{,}8}{n}+\frac{0{,}1}{\sqrt{n}}+\frac{0{,}5}{b\sqrt{n}}\right)$

Para GEV: si $b > 0{,}5$, hacer $b = 0{,}5$

Para Gamma o Log Pearson Tipo III: si r o $b < 2$, hacer r o $b = 2$

Basado en Laio (2004)

Anderson-Darling test

The Anderson-Darling test shows that if $\omega < \omega_*$, where $\omega_* = 0.743, 0.581, 0.461$ and0.347 for significant levels of $\alpha = 0.01, 0.025, 0.05$ and0.1, H_0 is accepted, which means that the annual series can be represented by the chosen distribution. Alternatively, with the value of ω, F is calculated as as:

$$F(\omega) = \frac{1}{\pi\sqrt{\omega}}\left[e^{\left(-\frac{1}{16\omega}\right)}K_{1/4}\left(\frac{1}{16\omega}\right)\right] + \frac{1}{\pi\sqrt{\omega}}\left[1.118e^{\left(-\frac{25}{16\omega}\right)}K_{1/4}\left(\frac{25}{16\omega}\right)\right] \text{ for } \omega < 1.2$$

$$F(\omega) = 1 \text{ for } \omega \geq 1.2$$

where $K_{1/4}[.]$ is the modified Bessel function of order $1/4$. If $F(\omega) < (1-\alpha)$, H_0 is accepted; the lesser $F(\omega)$ the more suitable is the distribution chosen.

Distribution fitting

Annual series of maximum streamflow at Banco station, Magdalena River

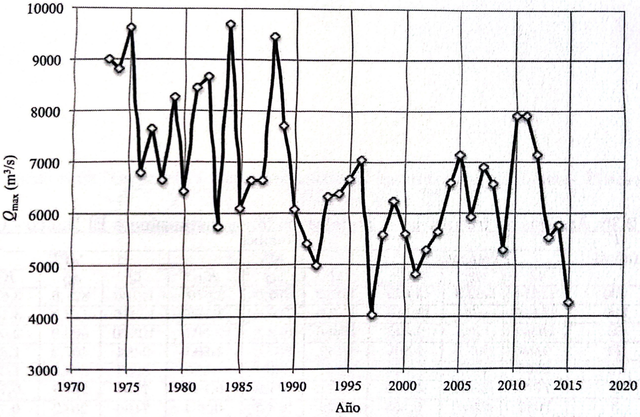

Figura 2.5: Serie de tiempo de la serie anual de caudales máximos instantáneos en El Banco

Distribution fitting

Annual series of maximum streamflow at Banco station, Magdalena River

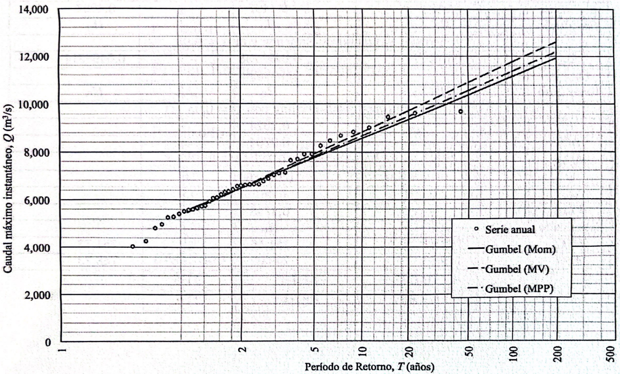

Figura 2.6: Curvas de frecuencia estimadas con Gumbel para la serie anual de caudales máximos instantáneos en El Banco

Distribution fitting

Annual series of maximum streamflow at Banco station, Magdalena River

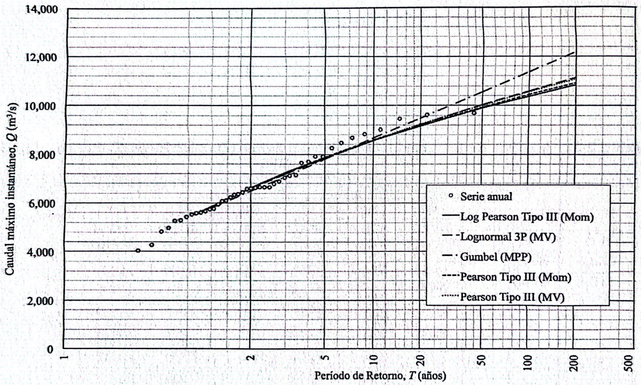

Figura 2.7: Curvas de frecuencia estimadas con Pearson y Log Pearson Tipo III y Lognormal 3P para la *serie* anual de caudales máximos instantáneos en El Banco

Distribution fitting

Annual series of maximum streamflow at Banco station, Magdalena River

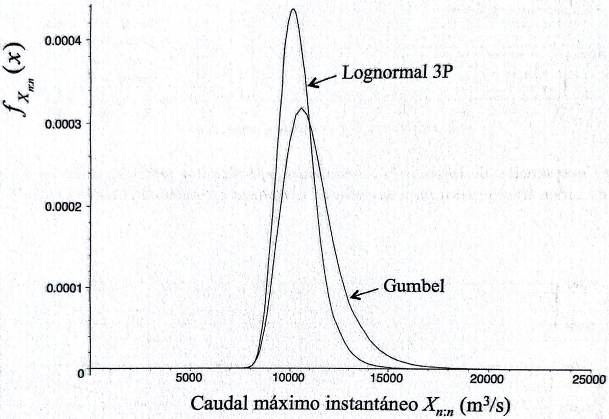

Figura 2.15: Funciones de densidad de probabilidad $f_{X_{n:n}}(x)$ del máximo caudal en 43 años en El Banco, a partir de las curvas de frecuencia ajustadas con Gumbel (MPP) y Lognormal 3P (MV)

Generalities

The processes described above can be applied to any type of distribution using the different parameter-estimation methods. HERERE