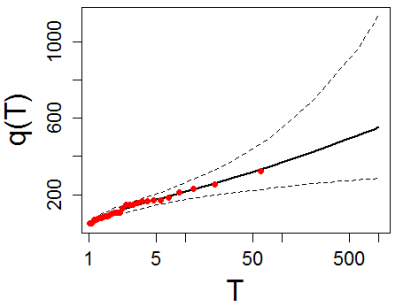
Procédures statistiques de la banque HYDRO 3

Annexe : documentation des codes de calcul



Septembre 2016

Benjamin Renard
Irstea Lyon-Villeurbanne
UR Hydrologie - Hydraulique

Institut national de recherche en sciences et technologies pour l'environnement et l'agriculture

Table des matières

I. Introduction

L'objectif de cette annexe est de documenter les méthodes statistiques qui sont implémentées dans les codes de calcul Hydro 3. Dans la première partie (section II), les principes généraux des méthodes d'estimation utilisées sont brièvement décrits. La seconde partie (section III) est un formulaire pour toutes les distributions disponibles dans Hydro 3. Cette seconde partie décrit également l'algorithme MCMC utilisé pour l'inférence Bayésienne, ainsi que trois tests statistiques (détection d'une tendance, d'une rupture à date inconnue, et test d'adéquation).

II. Principes généraux

II.1 Notations et hypothèses de base

Les données sont notées $\boldsymbol{y} = (y_1, ..., y_n)$ et sont considérées comme des réalisations indépendantes et identiquement distribuées (*iid*) d'une variable aléatoire Y. La distribution de cette variable aléatoire est paramétrée par un vecteur de paramètres $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$, qui sont inconnus et qu'il faut donc estimer. La densité de probabilité associée à cette distribution, calculée en une valeur z, est notée $f_Y(z;\boldsymbol{\theta})$, et la fonction de répartition est notée $F_Y(z;\boldsymbol{\theta})$. La fonction quantile calculée en une valeur p (comprise entre 0 et 1) est notée $F_Y^{-1}(p;\boldsymbol{\theta})$ (car c'est l'inverse de la fonction de répartition).

Les distributions disponibles dans Hydro 3 sont les suivantes : loi normale, loi log-normale, loi exponentielle, loi de Gumbel, loi de Gumbel pour les minima, loi de Pareto généralisée (GPD), loi généralisée des valeurs extrêmes (GEV), loi GEV pour les minima, loi de Pearson III, loi de Log-Pearson III et loi de Poisson.

Pour chacune de ces distributions, plusieurs méthodes d'estimation des paramètres sont disponibles, ainsi que plusieurs méthodes de quantification des incertitudes. Les sections qui suivent décrivent les principes généraux de ces méthodes.

II.2 Méthodes d'estimation des paramètres

II.2.1 Méthode des moments (MOM)

Le principe de la méthode des moments est d'estimer les paramètres de sorte que les moments théoriques de la distribution soient égaux aux moments empiriques calculés sur les données. Pour une distribution à k paramètres, on devra égaler k moments théoriques et empiriques. Les moments théoriques et empiriques d'ordre j sont définis de la façon suivante :

$$\begin{aligned} m_j^{(th)} &= E\left[Y^j\right] \\ m_j^{(emp)} &= \frac{1}{n}\sum_{i=1}^{n} y_i^j \end{aligned} \qquad (1)$$

L'application de la méthode des moments réclame donc de résoudre un système de k équations à k inconnues. La résolution de ces équations peut parfois être explicite, mais réclame parfois de recourir à des solveurs numériques (cf. le formulaire pour plus de détails).

II.2.2 Méthode des L-moments (LMOM)

Le principe est similaire à celui de l'estimation par la méthode des moments, mais on substitue aux moments les L-moments, qui sont définis de la façon suivante (on ne donne que les 3 premiers L-moments empiriques):

$$
\begin{aligned}
l_j^{(th)} &= j^{-1}\sum_{i=0}^{j-1}(-1)^j \binom{j-1}{i} E\left[Y_{(j-i:j)}\right] \\
l_1^{(emp)} &= \binom{n}{1}^{-1} \sum_{i=1}^{n} y_{(i)} \\
l_2^{(emp)} &= \frac{1}{2}\binom{n}{2}^{-1} \sum_{i=1}^{n}\left(\binom{i-1}{1} - \binom{n-i}{1}\right) y_{(i)} \\
l_3^{(emp)} &= \frac{1}{3}\binom{n}{3}^{-1} \sum_{i=1}^{n}\left(\binom{i-1}{2} - 2\binom{i-1}{1}\binom{n-i}{1} + \binom{n-i}{2}\right) y_{(i)}
\end{aligned}
\qquad (2)
$$

Dans ces équations, la notation $y_{(i)}$ signifie « la i^{eme} donnée triée par ordre croissant ».

Comme pour la méthode des moments, la résolution des équations peut parfois être explicite, mais réclame parfois de recourir à des solveurs numériques (cf. le formulaire pour plus de détails).

II.2.3 Méthode du maximum de vraisemblance (ML pour Maximum Likelihood)

Dans le cadre d'observations *iid* comme supposé en section II.1, la fonction de vraisemblance, que l'on notera $L(\boldsymbol{\theta}; \boldsymbol{y})$, est définie de la façon suivante :

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{n} f_Y\left(y_i; \boldsymbol{\theta}\right) \qquad (3)$$

Le calcul de la vraisemblance consiste donc simplement à effectuer le produit des densités de probabilité évaluées en chacune des observations. L'estimateur du maximum de vraisemblance est alors défini comme le vecteur de paramètres qui maximise la vraisemblance :

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{ArgMax}\left\{L(\boldsymbol{\theta}; \boldsymbol{y})\right\} \qquad (4)$$

Dans certains cas, il est possible de calculer explicitement cet estimateur (cf. formulaire). En l'absence de résolution explicite, on aura recours à une optimisation numérique.

II.2.4 Méthode Bayésienne (BAY)

L'estimation Bayésienne est également basée sur la fonction de vraisemblance, mais utilise en plus une autre source d'information, nommée la distribution *a priori*. Cette distribution, dont la densité est notée $p(\boldsymbol{\theta})$, contient toute connaissance sur les paramètres à estimer qui peut être mobilisée sans utiliser les données au site d'étude. Une telle connaissance peut provenir par exemple d'expertise ou d'information régionale.

La distribution *a priori* et la fonction de vraisemblance sont combinées d'après le théorème de Bayes pour calculer la distribution *a posteriori* des paramètres, dont la densité est notée $p(\boldsymbol{\theta} \mid \boldsymbol{y})$:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto L(\boldsymbol{\theta}; \boldsymbol{y})\, p(\boldsymbol{\theta}) \qquad (5)$$

L'équation (5) stipule donc que l'on obtient la densité a posteriori en multipliant simplement la densité *a priori* et la fonction de vraisemblance (cf. équation (3)). Précisons que le symbole

'$\propto$' signifie '*est proportionnel à*' : la densité *a posteriori* n'est donc connue qu'à une constante multiplicative près. Ceci n'est pas problématique puisque l'estimateur utilisé correspond au vecteur de paramètres qui maximise la densité *a posteriori* (ce maximum étant inchangé lorsque la densité *a posteriori* est multipliée par une constante) :

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{ArgMax}\left\{ p(\boldsymbol{\theta} \mid \boldsymbol{y}) \right\} \tag{6}$$

En général il n'est pas possible de calculer explicitement cet estimateur, on a donc recours à une méthode d'optimisation numérique.

II.3 Méthodes de quantification des incertitudes

II.3.1 Méthode du Bootstrap (BOOT)

La méthode du Bootstrap est une méthode de ré-échantillonnage dont l'algorithme peut être décrit de la manière suivante :

Pour $i = 1 : N_{sim}$

1. **Ré-échantillonnage** : Tirer au hasard et avec remise n valeurs dans les données disponibles $\boldsymbol{y} = (y_1, ..., y_n)$. Certaines valeurs apparaîtront donc plusieurs fois, d'autres seront absentes.
2. **Estimation** : calculer l'estimateur $\hat{\boldsymbol{\theta}}^{(i)}$ sur ces données ré-échantillonnées.

Fin

L'ensemble des valeurs $\left(\hat{\boldsymbol{\theta}}^{(i)}\right)_{i=1:N_{sim}}$ représente ainsi l'incertitude sur les paramètres estimés. L'incertitude résultante sur les quantiles peut aisément être quantifiée en appliquant la fonction quantile de la distribution à l'ensemble de ces paramètres.

II.3.2 Méthode du Bootstrap paramétrique (PBOOT)

La méthode du Bootstrap paramétrique est également une méthode de ré-échantillonnage. La différence avec le Bootstrap « classique » est liée à la manière dont les échantillons simulés sont produits, comme décrit dans l'algorithme ci-dessous :

Calculer l'estimateur $\hat{\boldsymbol{\theta}}^{(0)}$ sur les données observées $\boldsymbol{y} = (y_1, ..., y_n)$.

Pour $i = 1 : N_{sim}$

1. **Ré-échantillonnage** : Simuler n valeurs dans la distribution estimée (de paramètres $\hat{\boldsymbol{\theta}}^{(0)}$).
2. **Estimation** : calculer l'estimateur $\hat{\boldsymbol{\theta}}^{(i)}$ sur ces données ré-échantillonnées.

Fin

Comme pour le Bootstrap « classique », l'ensemble des valeurs $\left(\hat{\boldsymbol{\theta}}^{(i)}\right)_{i=1:N_{sim}}$ représente l'incertitude sur les paramètres estimés, qui peut aisément être propagée aux quantiles. La différence entre le Bootstrap « classique » et le Bootstrap paramétrique vient de la méthode de

ré-échantillonnage : alors que le premier se contente de re-simuler des valeurs qui ont effectivement été observées, le second génère de nouvelles valeurs, différentes des observations.

II.3.3 Méthode spécifique au maximum de vraisemblance (ML)

Dans le cas de l'estimation par maximum de vraisemblance, il existe une théorie asymptotique bien développée qui fournit une approximation sur la distribution d'échantillonnage des estimateurs. En effet, on peut montrer que lorsque *n* (la taille de l'échantillon) tend vers l'infini, la distribution d'échantillonnage de l'estimateur du maximum de vraisemblance $\hat{\boldsymbol{\theta}}$ est Gaussienne, de moyenne $\boldsymbol{\theta}_0$ (la vraie valeur du paramètre : absence de biais) et de matrice de covariance $V(\hat{\boldsymbol{\theta}})$, où $V(\boldsymbol{\theta})$ est la matrice d'information de Fisher :

$$V(\boldsymbol{\theta}) = \begin{pmatrix} -\frac{\partial^2}{\partial\theta_1^2}\ln\left(L(\boldsymbol{\theta};\boldsymbol{y})\right) & \dots & -\frac{\partial^2}{\partial\theta_1\partial\theta_k}\ln\left(L(\boldsymbol{\theta};\boldsymbol{y})\right) \\ \vdots & \ddots & \vdots \\ -\frac{\partial^2}{\partial\theta_k\partial\theta_1}\ln\left(L(\boldsymbol{\theta};\boldsymbol{y})\right) & \dots & -\frac{\partial^2}{\partial\theta_k^2}\ln\left(L(\boldsymbol{\theta};\boldsymbol{y})\right) \end{pmatrix} \qquad (7)$$

On représente ainsi l'incertitude sur les paramètres estimés par un loi Normale de moyenne $\hat{\boldsymbol{\theta}}$ et de matrice de covariance $V(\hat{\boldsymbol{\theta}})$. Cette incertitude sur les paramètres peut être propagée aux quantiles, par exemple en utilisant une approche Monte-Carlo (c'est ce qui est fait dans les codes Hydro 3).

Précisons que dans l'équation (7), les dérivées partielles sont approchées numériquement avec un schéma de différences finies.

II.3.4 Méthode spécifique à l'approche Bayesienne (BAY)

Dans l'approche Bayesienne, la distribution *a posteriori* (équation (5)) fournit directement une quantification de l'incertitude. Néanmoins, l'utilisation directe de cette distribution est délicate pour deux raisons : (i) il s'agit d'une distribution multi-dimensionnelle ; (ii) elle n'est connue qu'à une constante près.

Pour contourner cette difficulté, on fait appel à des simulateurs de Monte Carlo par Chaînes de Markov (MCMC). Les algorithmes MCMC désignent une famille de méthodes qui permettent de simuler des réalisations à partir d'une densité de probabilité arbitraire, connue éventuellement seulement à une constante près : ceci correspond exactement au contexte de la distribution *a posteriori*. L'ensemble des valeurs de paramètres simulées par MCMC peut ainsi être utilisé pour dériver aisément des incertitudes sous la forme d'intervalles par exemple.

Les détails techniques de l'algorithme utilisé dans Hydro 3 sont donnés dans le formulaire, en section III.2. Il s'agit du même algorithme que celui implémenté dans les logiciels JBay[1] et BaRatinAGE[2].

II.4 Combinaisons méthode d'estimation des paramètres / méthode de quantification des incertitudes

Toutes les combinaisons possibles ne sont pas autorisées, car certaines méthodes de quantification des incertitudes n'ont de sens que dans le cadre d'une méthode d'estimation

[1] https://forge.irstea.fr/projects/thebay/files
[2] https://forge.irstea.fr/projects/baratinage_v2/files

bien particulière (typiquement, ML et BAY). Le tableau ci-dessous résume les combinaisons possibles (la méthode « NONE » consistant simplement à ne pas quantifier l'incertitude).

Incertitudes / Estimation	BOOT	PBOOT	ML	BAY	NONE
MOM	✓	✓			✓
LMOM	✓	✓			✓
ML	✓	✓	✓		✓
BAY				✓	✓

III. Formulaire

III.1 Distributions et estimateurs

Dans les tableaux de cette section nous utilisons les notations suivantes. La moyenne, l'écart-type et le coefficient d'asymétrie des données $\boldsymbol{y} = (y_1, ..., y_n)$ sont respectivement notés m_y, s_y et k_y :

$$
\begin{aligned}
m_y &= \frac{1}{n}\sum_{i=1}^{n} y_i \\
s_y &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - m_y\right)^2} \\
k_y &= \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - m_y\right)^3}{s_y^3}
\end{aligned}
\qquad (8)
$$

Les L-moments empiriques d'ordre 2 et 3 sont respectivement notés $l_y^{(2)}$ et $l_y^{(3)}$ (cf. équation (2)). L'indice utilisé dans ces notations est important : ainsi, la notation $m_{\ln(y)}$ désigne la moyenne du logarithme népérien des données, i.e. la moyenne de $\ln(\boldsymbol{y}) = (\ln(y_1), ..., \ln(y_n))$.

Les tableaux de cette section donnent les propriétés de base de chaque distribution (paramètres, support, densité, fonction de répartition et fonction quantile) ainsi que les estimateurs des moments (MOM), des L-moments (LMOM) et du maximum de vraisemblance (ML). L'estimateur bayésien (BAY) n'apparait pas car il n'y a pas de formule explicite (on a recours aux simulations MCMC). De même, il n'y a pas de formules explicites pour les incertitudes (on a recours à des approches numériques pour toutes les méthodes).

Pour être tout à fait correct, certains estimateurs dans ces tableaux (repérés par le symbole $^{(*)}$) ne sont pas à strictement parler les estimateurs des moments ou des L-moments. C'est le cas par exemple des estimateurs des L-moments pour les distributions LogNormal et LogPearsonIII : on calcul en fait l'estimateur des L-moments de ln(*y*), ce qui n'est pas équivalent mathématiquement parlant, mais correspond à l'estimateur décrit dans la littérature.

III.1.1 Loi Normale

Paramètres	Moyenne μ ; Ecart-type $\sigma > 0$
Support	$z \in]-\infty;+\infty[$
Densité	$f_Y(z;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)$
Fonction de répartition	Pas d'expression analytique (utilisation de la fonction *pnorm* de R).
Fonction quantile	Pas d'expression analytique (utilisation de la fonction *qnorm* de R).
MOM	$\hat{\mu} = m_y$ $\hat{\sigma} = s_y$
LMOM	$\hat{\mu} = m_y$ $\hat{\sigma} = \sqrt{\pi} \times l_y^{(2)}$
ML	$\hat{\mu} = m_y$ $\hat{\sigma} = s_y$

III.1.2 Loi Log-normale

Paramètres	Moyenne-log μ ; Ecart-type-log $\sigma > 0$
Support	$z \in]0;+\infty[$
Densité	$f_Y(z;\mu,\sigma) = \frac{1}{z\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(z)-\mu)^2}{2\sigma^2}\right)$
Fonction de répartition	Pas d'expression analytique (utilisation de la fonction *plnorm* de R).
Fonction quantile	Pas d'expression analytique (utilisation de la fonction *qlnorm* de R).
MOM	$\hat{\mu} = \ln(m_y) - 0.5w$ $\hat{\sigma} = \sqrt{w}$ avec $w = \ln\left(1 + \frac{s_y^2}{m_y^2}\right)$
LMOM (*)	$\hat{\mu} = m_{\ln(y)}$ $\hat{\sigma} = \sqrt{\pi} \times l^{(2)}_{\ln(y)}$
ML	$\hat{\mu} = m_{\ln(y)}$ $\hat{\sigma} = s_{\ln(y)}$

III.1.3 Loi exponentielle à 1 paramètre

Paramètres	Echelle $\sigma > 0$
Support	$z \in [0;+\infty[$
Densité	$f_Y(z;\sigma) = \frac{1}{\sigma}\exp\left(-\frac{z}{\sigma}\right)$
Fonction de répartition	$F_Y(z;\sigma) = 1-\exp\left(-\frac{z}{\sigma}\right)$
Fonction quantile	$F_Y^{-1}(p;\sigma) = -\sigma\ln(1-p)$
MOM	$\hat{\sigma} = m_y$
LMOM	$\hat{\sigma} = m_y$
ML	$\hat{\sigma} = m_y$

III.1.4 Loi exponentielle à 2 paramètres (seuil inconnu)

Paramètres	Seuil μ ; Echelle $\sigma > 0$
Support	$z \in [\mu; +\infty[$
Densité	$f_Y(z;\mu,\sigma) = \frac{1}{\sigma}\exp\left(-\frac{z-\mu}{\sigma}\right)$
Fonction de répartition	$F_Y(z;\mu,\sigma) = 1-\exp\left(-\frac{z-\mu}{\sigma}\right)$
Fonction quantile	$F_Y^{-1}(p;\mu,\sigma) = \mu - \sigma\ln(1-p)$
MOM	$\hat{\sigma} = s_y$ $\hat{\mu} = m_y - \hat{\sigma}$
LMOM	$\hat{\sigma} = 2 \times l_y^{(2)}$ $\hat{\mu} = m_y - \hat{\sigma}$
ML	$\hat{\mu} = \min(y)$ $\hat{\sigma} = m_y - \hat{\mu}$

III.1.5 Loi de Gumbel

Paramètres	Position μ ; Echelle $\sigma > 0$
Support	$z \in]-\infty;+\infty[$
Densité	$f_Y(z;\mu,\sigma) = \frac{1}{\sigma}\exp\left(-\frac{z-\mu}{\sigma} - \exp\left(-\frac{z-\mu}{\sigma}\right)\right)$
Fonction de répartition	$F_Y(z;\mu,\sigma) = \exp\left(-\exp\left(-\frac{z-\mu}{\sigma}\right)\right)$
Fonction quantile	$F_Y^{-1}(p;\mu,\sigma) = \mu - \sigma\ln\left(-\ln(p)\right)$
MOM	$\hat{\sigma} = \frac{\sqrt{6}}{\pi}s_y$ $\hat{\mu} = m_y - \gamma\hat{\sigma}$, où $\gamma = 0.5772$
LMOM	$\hat{\sigma} = l_y^{(2)} / \ln(2)$ $\hat{\mu} = m_y - \gamma\hat{\sigma}$
ML	Pas de formule explicite, optimisation numérique.

III.1.6 Loi de Gumbel pour les minima

Paramètres	Position μ ; Echelle $\sigma > 0$
Support	$z \in]-\infty;+\infty[$
Densité	$f_Y(z;\mu,\sigma) = \frac{1}{\sigma}\exp\left(\frac{z-\mu}{\sigma} - \exp\left(\frac{z-\mu}{\sigma}\right)\right)$
Fonction de répartition	$F_Y(z;\mu,\sigma) = 1 - \exp\left(-\exp\left(\frac{z-\mu}{\sigma}\right)\right)$
Fonction quantile	$F_Y^{-1}(p;\mu,\sigma) = \mu + \sigma\ln\left(-\ln(1-p)\right)$
MOM	$\hat{\sigma} = \frac{\sqrt{6}}{\pi} s_{-1\times y}$ $\hat{\mu} = -1\times\left(m_{-1\times y} - \gamma\hat{\sigma}\right)$, où $\gamma = 0.5772$
LMOM	$\hat{\sigma} = l^{(2)}_{-1\times y} / \ln(2)$ $\hat{\mu} = -1\times\left(m_{-1\times y} - \gamma\hat{\sigma}\right)$
ML	Pas de formule explicite, optimisation numérique.

III.1.7 Loi GPD à 2 paramètres

Paramètres	Echelle $\sigma > 0$; Forme ξ Quand $\xi \to 0$, la loi GPD tend vers la loi exponentielle.
Support	Si $\xi < 0$, $z \in [0; +\infty[$; Si $\xi > 0$, $z \in [0; \sigma/\xi[$;
Densité	$f_Y(z;\sigma,\xi) = \frac{1}{\sigma}\left(1-\xi\frac{z}{\sigma}\right)^{\frac{1}{\xi}-1}$
Fonction de répartition	$F_Y(z;\sigma,\xi) = 1-\left(1-\xi\frac{z}{\sigma}\right)^{\frac{1}{\xi}}$
Fonction quantile	$F_Y^{-1}(p;\sigma,\xi) = \frac{\sigma}{\xi}\left(1-(1-p)^{\xi}\right)$
MOM	$\hat{\sigma} = 0.5 \times m_y \times (w+1)$ $\hat{\xi} = 0.5 \times (w-1)$ avec $w = \frac{m_y^2}{s_y^2}$
LMOM	$\hat{\xi} = \frac{m_y}{l_y^{(2)}} - 2$ $\hat{\sigma} = \left(1+\hat{\xi}\right) \times m_y$
ML	Pas de formule explicite, optimisation numérique.

III.1.8 Loi GPD à 3 paramètres (seuil inconnu)

Paramètres	Seuil μ ; Echelle $\sigma > 0$; Forme ξ Quand $\xi \to 0$, la loi GPD tend vers la loi exponentielle.
Support	Si $\xi < 0$, $z \in [\mu; +\infty[$; Si $\xi > 0$, $z \in [\mu; \mu + \sigma/\xi[$;
Densité	$f_Y(z;\mu,\sigma,\xi) = \frac{1}{\sigma}\left(1 - \xi\frac{z-\mu}{\sigma}\right)^{\frac{1}{\xi}-1}$
Fonction de répartition	$F_Y(z;\mu,\sigma,\xi) = 1 - \left(1 - \xi\frac{z-\mu}{\sigma}\right)^{\frac{1}{\xi}}$
Fonction quantile	$F_Y^{-1}(p;\mu,\sigma,\xi) = \mu + \frac{\sigma}{\xi}\left(1-(1-p)^{\xi}\right)$
MOM (*)	$\hat{\mu} = \min(\boldsymbol{y})$ $\hat{\sigma} = 0.5 \times (m_y - \hat{\mu}) \times (w+1)$ $\hat{\xi} = 0.5 \times (w-1)$ avec $w = \frac{(m_y - \hat{\mu})^2}{s_y^2}$
LMOM	$\hat{\xi} = \frac{1 - 3l_y^{(3)}}{1 + l_y^{(3)}}$ $\hat{\sigma} = \left(1+\hat{\xi}\right) \times \left(2+\hat{\xi}\right) \times l_y^{(2)}$ $\hat{\mu} = m_y - \left(2+\hat{\xi}\right) \times l_y^{(2)}$
ML	Pas de formule explicite, optimisation numérique.

III.1.9 Loi GEV

Paramètres	Position μ ; Echelle $\sigma > 0$; Forme ξ Quand $\xi \to 0$, la loi GEV tend vers la loi de Gumbel.
Support	Si $\xi < 0$, $z \in]\mu + \sigma/\xi ; +\infty[$; Si $\xi > 0$, $z \in]-\infty ; \mu + \sigma/\xi[$;
Densité	$f_Y(z;\mu,\sigma,\xi) = \frac{1}{\sigma}\left(1-\xi\frac{z-\mu}{\sigma}\right)^{\frac{1}{\xi}-1} \exp\left(-\left(1-\xi\frac{z-\mu}{\sigma}\right)^{\frac{1}{\xi}}\right)$
Fonction de répartition	$F_Y(z;\mu,\sigma,\xi) = \exp\left(-\left(1-\xi\frac{z-\mu}{\sigma}\right)^{\frac{1}{\xi}}\right)$
Fonction quantile	$F_Y^{-1}(p;\mu,\sigma,\xi) = \mu + \frac{\sigma}{\xi}\left(1-(-\ln(p))^{\xi}\right)$
MOM	$-\frac{\hat{\xi}}{\lvert\hat{\xi}\rvert}\frac{\left[\Gamma(3\hat{\xi}+1)-3\Gamma(\hat{\xi}+1)\Gamma(2\hat{\xi}+1)+2\Gamma^3(\hat{\xi}+1)\right]}{\left[\Gamma(2\hat{\xi}+1)-\Gamma^2(\hat{\xi}+1)\right]^{3/2}} = k_y$ (à résoudre numériquement) $\hat{\sigma} = \lvert\hat{\xi}\rvert s_y\left[\Gamma(2\hat{\xi}+1)-\Gamma^2(\hat{\xi}+1)\right]^{-1/2}$ $\hat{\mu} = m_y - \frac{\hat{\sigma}}{\hat{\xi}}\left[1-\Gamma(\hat{\xi}+1)\right]$
LMOM	$\hat{\xi} = 7.8590w + 2.9554w^2$, avec $w = \frac{2}{3+l_y^{(3)}} - \frac{\ln(2)}{\ln(3)}$ $\hat{\sigma} = \left(\hat{\xi}\times l_y^{(2)}\right)/\left((1-2^{-\hat{\xi}})\Gamma(1+\hat{\xi})\right)$ $\hat{\mu} = m_y + \left(\hat{\sigma}/\hat{\xi}\right)\left(1-\Gamma(1+\hat{\xi})\right)$
ML	Pas de formule explicite, optimisation numérique.

III.1.10 Loi GEV pour les minima

Paramètres	Position μ ; Echelle $\sigma > 0$; Forme ξ Quand $\xi \to 0$, la loi GEV pour les minima tend vers la loi de Gumbel pour les minima
Support	Si $\xi > 0$, $z \in]\mu - \sigma/\xi ; +\infty[$; Si $\xi < 0$, $z \in]-\infty ; \mu - \sigma/\xi[$;
Densité	$f_Y(z;\mu,\sigma,\xi) = \frac{1}{\sigma}\left(1+\xi\frac{z-\mu}{\sigma}\right)^{\frac{1}{\xi}-1} \exp\left(-\left(1+\xi\frac{z-\mu}{\sigma}\right)^{\frac{1}{\xi}}\right)$
Fonction de répartition	$F_Y(z;\mu,\sigma,\xi) = 1 - \exp\left(-\left(1+\xi\frac{z-\mu}{\sigma}\right)^{\frac{1}{\xi}}\right)$
Fonction quantile	$F_Y^{-1}(p;\mu,\sigma,\xi) = \mu - \frac{\sigma}{\xi}\left(1-(-\ln(1-p))^{\xi}\right)$
MOM	$-\frac{\hat{\xi}}{\lvert\hat{\xi}\rvert}\frac{\left[\Gamma(3\hat{\xi}+1) - 3\Gamma(\hat{\xi}+1)\Gamma(2\hat{\xi}+1) + 2\Gamma^3(\hat{\xi}+1)\right]}{\left[\Gamma(2\hat{\xi}+1) - \Gamma^2(\hat{\xi}+1)\right]^{3/2}} = k_{-1\times y}$ (à résoudre numériquement) $\hat{\sigma} = \lvert\hat{\xi}\rvert s_{-1\times y}\left[\Gamma(2\hat{\xi}+1) - \Gamma^2(\hat{\xi}+1)\right]^{-1/2}$ $\hat{\mu} = -1\times\left(m_{-1\times y} - \frac{\hat{\sigma}}{\hat{\xi}}\left[1-\Gamma(\hat{\xi}+1)\right]\right)$
LMOM	$\hat{\xi} = 7.8590w + 2.9554w^2$, avec $w = \frac{2}{3+l^{(3)}_{-1\times y}} - \frac{\ln(2)}{\ln(3)}$ $\hat{\sigma} = \left(\hat{\xi}\times l^{(2)}_{-1\times y}\right) / \left((1-2^{-\hat{\xi}})\Gamma(1+\hat{\xi})\right)$ $\hat{\mu} = -1\times\left(m_{-1\times y} + \left(\hat{\sigma}/\hat{\xi}\right)\left(1\text{-}\Gamma(1+\hat{\xi})\right)\right)$
ML	Pas de formule explicite, optimisation numérique.

III.1.11 Loi de Pearson III

Paramètres	Position μ ; Echelle $\sigma \neq 0$; Forme $\xi > 0$
Support	Si $\sigma > 0$, $z \in]\mu; +\infty[$; Si $\sigma < 0$, $z \in]-\infty; \mu[$;
Densité	$f_Y(z;\mu,\sigma,\xi) = \dfrac{\left(\dfrac{z-\mu}{\sigma}\right)^{\xi-1} \exp\left(-\dfrac{z-\mu}{\sigma}\right)}{\lvert\sigma\rvert\,\Gamma(\xi)}$
Fonction de répartition	Pas d'expression analytique (utilisation de la fonction *pgamma* de R).
Fonction quantile	Pas d'expression analytique (utilisation de la fonction *qgamma* de R).
MOM	$\hat{\xi} = 4 / k_y^2$ $\hat{\sigma} = \text{signe}\left(k_y\right) \times \dfrac{s_y}{\sqrt{\hat{\xi}}}$ $\hat{\mu} = m_y - \hat{\sigma} \times \hat{\xi}$
LMOM	$\hat{\mu} = m_y - 2(\tau/\eta)$ $\hat{\sigma} = 0.5\,\tau\eta$ $\hat{\xi} = 4/\eta^2$ où les valeurs τ et η sont définies de la façon suivante : $\eta = \left(2/\sqrt{a}\right) \times signe\left(l_y^{(3)}\right)$ $\tau = l_y^{(2)} \sqrt{\pi a}\, \dfrac{\Gamma(a)}{\Gamma(a+0.5)}$ avec : si $\left\lvert l_y^{(3)}\right\rvert < 1/3$: $w = 3\pi (l_y^{(3)})^2$ $a = \dfrac{1 + 0.2906w}{w + 0.1882w^2 + 0.0442w^3}$ sinon: $w = 1 - \left\lvert l_y^{(3)}\right\rvert$ $a = \dfrac{0.36067w - 0.59567w^2 + 0.25361w^3}{1 - 2.78861w + 2.56096w^2 - 0.77045w^3}$
ML	Pas de formule explicite, optimisation numérique.

III.1.12 Loi de Log-Pearson III

Paramètres	Position-log μ ; Echelle-log $\sigma \neq 0$; Forme-log $\xi > 0$
Support	Si $\sigma > 0$, $z \in]e^{\mu}; +\infty[$; Si $\sigma < 0$, $z \in]0; e^{\mu}[$;
Densité	$f_Y(z;\mu,\sigma,\xi) = \dfrac{\left(\dfrac{\ln(z)-\mu}{\sigma}\right)^{\xi-1} \exp\left(-\dfrac{\ln(z)-\mu}{\sigma}\right)}{\lvert\sigma\rvert\,\Gamma(\xi)} - \ln(z)$
Fonction de répartition	Pas d'expression analytique (utilisation de la fonction *pgamma* de R).
Fonction quantile	Pas d'expression analytique (utilisation de la fonction *qgamma* de R).
MOM (*)	$\hat{\xi} = 4 / k^2_{\ln(y)}$ $\hat{\sigma} = \text{signe}\left(k_{\ln(y)}\right) \times \dfrac{s_{\ln(y)}}{\sqrt{\hat{\xi}}}$ $\hat{\mu} = m_{\ln(y)} - \hat{\sigma} \times \hat{\xi}$
LMOM (*)	$\hat{\mu} = m_{\ln(y)} - 2(\tau/\eta)$ $\hat{\sigma} = 0.5\,\tau\eta$ $\hat{\xi} = 4/\eta^2$ où les valeurs τ et η sont définies de la façon suivante : $\eta = \left(2/\sqrt{a}\right) \times signe\left(l^{(3)}_{\ln(y)}\right)$ $\tau = l^{(2)}_{\ln(y)} \sqrt{\pi a}\, \dfrac{\Gamma(a)}{\Gamma(a+0.5)}$ avec : si $\left\lvert l^{(3)}_{\ln(y)}\right\rvert < 1/3$: $w = 3\pi (l^{(3)}_{\ln(y)})^2$ $a = \dfrac{1 + 0.2906w}{w + 0.1882w^2 + 0.0442w^3}$ sinon: $w = 1 - \left\lvert l^{(3)}_{\ln(y)}\right\rvert$ $a = \dfrac{0.36067w - 0.59567w^2 + 0.25361w^3}{1 - 2.78861w + 2.56096w^2 - 0.77045w^3}$
ML	Pas de formule explicite, optimisation numérique.

III.1.13 Loi de Poisson

Paramètres	Taux λ
Support	$k \geq 0$ entier
Fonction de masse	$f_Y(k;\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$
Fonction de répartition	$F_Y(z;\lambda) = e^{-\lambda}\sum_{i=0}^{\lfloor z \rfloor}\frac{\lambda^i}{i!}$ (note : $\lfloor z \rfloor$ désigne la partie entière de z)
Fonction quantile	Par convention, $F_Y^{-1}(p;\lambda)$ est le plus petit entier k tel que $F_Y(k;\lambda) \geq p$
MOM	$\hat{\lambda} = m_y$
LMOM	$\hat{\lambda} = m_y$
ML	$\hat{\lambda} = m_y$

III.2 Algorithme MCMC

L'algorithme MCMC implémenté dans les codes d'Hydro 3 est décrit ci-dessous.

0. Choisir un point de départ $\boldsymbol{\theta}^{(0)} = \left(\theta_1^{(0)},...,\theta_k^{(0)}\right)$ et des écarts-types de saut $\boldsymbol{v} = (v_1,...,v_k)$.
1. $c = 0$ # Initialisation du compteur
2. Répéter pour $i = 1 : N_{cycles}$ # A l'issue de chaque cycle on adaptera l'écart-type du saut
   a. Répéter pour $j = 1 : N_{adapt}$ # Boucle sans adaptation des écarts-types de saut
      i. $c = c + 1$; # Incrément du compteur
      ii. Répéter pour $d = 1 : k$ # Boucle sur chaque composante du vecteur $\boldsymbol{\theta}$
         - Générer un candidat $\theta_d^{(*)}$ à partir d'une distribution gaussienne de moyenne $\theta_d^{(c-1)}$ et d'écart-type v_d ;
         - Calculer le ratio entre la nouvelle et l'ancienne valeur de la densité a posteriori :
           $$\tau = \frac{p\left(\theta_1^{(c)},...,\theta_{d-1}^{(c)},\theta_d^{(*)},\theta_{d+1}^{(c-1)}...,\theta_p^{(c-1)} \mid \boldsymbol{y}\right)}{p\left(\theta_1^{(c)},...,\theta_{d-1}^{(c)},\theta_d^{(c-1)},\theta_{d+1}^{(c-1)}...,\theta_p^{(c-1)} \mid \boldsymbol{y}\right)}$$;
         - Accepter le candidat ($\theta_d^{(c)} = \theta_d^{(*)}$) avec une probabilité égale à $\min(1;\tau)$; sinon, rejeter le candidat ($\theta_d^{(c)} = \theta_d^{(c-1)}$).
   b. Mise à jour des écarts-types de saut
      i. Répéter pour $d = 1 : k$ # Boucle sur chaque composante du vecteur $\boldsymbol{\theta}$
         - Calculer le taux d'acceptation α_d pour la composante d ;
         - Si $\alpha_d \le \alpha_{\min}$, $v_d = \phi^- \times v_d$ # Le taux d'acceptation est trop faible, donc on diminue l'écart-type de saut
         - Si $\alpha_d \ge \alpha_{\max}$, $v_d = \phi^+ \times v_d$ # Le taux d'acceptation est trop grand, donc on augmente l'écart-type de saut
         - Sinon conserver v_d # Le taux d'acceptation est entre les bornes souhaitées $[\alpha_{\min};\alpha_{\max}]$, on conserve donc l'écart-type de saut.

Dans l'algorithme ci-dessus, les valeurs de N_{cycles}, N_{adapt}, $\alpha_{\min}$, $\alpha_{\max}$, ϕ^- et ϕ^+ peuvent être modifiées pour régler les propriétés de l'échantillonneur MCMC. Les valeurs par défaut définies dans Hydro 3 ($N_{cycles} = 100$, $N_{adapt} = 100$, $\alpha_{\min} = 0.1$, $\alpha_{\max} = 0.5$, $\phi^- = 0.9$ et $\phi^+ = 1.1$) devraient néanmoins convenir dans la grande majorité des cas.

Pour finir, les simulations brutes issues de l'algorithme ci-dessus sont post-traitées de la manière suivante :

- Brûlage : on efface la première partie des simulations car les premières simulations sont parfois peu représentatives de la distribution cible. Dans Hydro 3, le facteur de brûlage utilisé par défaut est égal à 0.5 (c'est-à-dire qu'on efface la première moitié des simulations) ;

- Affinage : sur les simulations restantes, on ne conserve qu'une valeur toutes les $N_{affinage}$. Dans Hydro 3, la valeur par défaut est $N_{affinage} = 5$. La perte d'information qui en résulte est limitée car les simulations MCMC brutes sont très autocorrélées.

Ainsi, avec ces valeurs par défaut, l'algorithme MCMC générera 10 000 simulations, mais après post-traitement, seules 1000 simulations seront conservées, ce qui est largement suffisant dans la plupart des cas et permet de diminuer à la fois le temps de calcul et les volumes de stockage.

III.3 Tests statistiques

III.3.1 Détection d'une tendance : test de Mann-Kendall

Conditions d'application : échantillon $(y_1,...,y_n)$ de valeurs indépendantes.

Soit $S=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} signe(y_j-y_i)$.

On peut calculer la variance de S comme suit : $Var(S)=n(n-1)(2n+5)/18$

En cas de valeurs ex-aequo dans l'échantillon, cette variance est modifiée comme suit :

$$Var(S)=\left[n(n-1)(2n+5)-\sum_{k=2}^{n}t_k k(k-1)(2k+5)\right]\Big/18$$

où t_k désigne le nombre d'égalités impliquant k valeurs

La statistique de test est finalement:

$$Z=\begin{cases}\dfrac{S-1}{\sqrt{Var(S)}} \text{ si } S>0\\ 0 \text{ si } S=0\\ \dfrac{S+1}{\sqrt{Var(S)}} \text{ si } S<0\end{cases}$$

En l'absence de tendance (hypothèse H_0), la distribution de la statistique de test peut être approchée par une loi normale centrée réduite.

III.3.2 Détection d'une rupture à date inconnue : test de Pettitt

Conditions d'application : échantillon $(y_1,...,y_n)$ de valeurs indépendantes.

Statistique de test : $S=\max_k(|U(k)|)$, où $U(k)=\sum_{i=1}^{k}\sum_{j=k+1}^{n} signe(y_i-y_j)$.

Une estimation du point de rupture est donnée par $\hat{r}=\operatorname{Arg}\max_k(|U(k)|)$.

En l'absence de rupture (hypothèse H_0), on peut approcher la distribution de la statistique de test comme ceci : $\Pr(S\geq s_0)=2\exp\left(\dfrac{-6s_0^2}{n^3+n^2}\right)$.

III.3.3 Test d'adéquation : test de Kolmogorov-Smirnov

Note : le code utilise la fonction existante `ks.test` dans R.

Conditions d'application : échantillon $(y_1,...,y_n)$ de valeurs indépendantes. On souhaite tester l'hypothèse H_0 que ces valeurs sont des réalisations d'une distribution connue, dont la fonction de répartition est $F(y)$.

Soit $F_n(y)$ la fonction de répartition empirique (également appelée courbe des fréquences cumulées) calculée à partir de l'échantillon : $F_n(y)=\frac{1}{n}\sum_{i=1}^{n}1\{y_i\leq y\}$.

La statistique de test est : $S=\sup_y(|F_n(y)-F(y)|)$.

La distribution de S sous l'hypothèse H_0 est tabulée ou peut être approchée par un algorithme spécifique[3].

[3] Marsaglia,G., Tsang, W.W and Wang, J. (**2003**), Evaluating Kolmogorov's distribution. *Journal of Statistical Software*, 8/18.

Irstea – centre de Lyon-Villeurbanne
UR Hydrologie-Hydraulique
5 rue de la Doua – BP 32108
69616 Villeurbanne Cedex
tél. +33 (0)4 72 20 87 87
www.irstea.fr