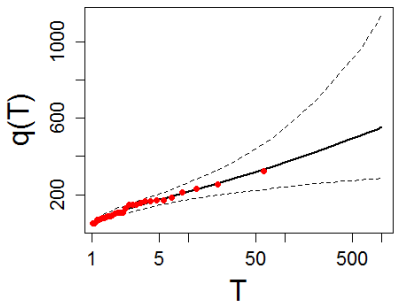
Procédures statistiques de la banque HYDRO 3: Évaluation et vérification des codes de calcul

Benjamin Renard

Procédures statistiques de la banque HYDRO 3

Evaluation et vérification 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 5

II. RAPIDE DESCRIPTION DES CODES 7

II.1 GENERALITES 7
II.2 PRINCIPALES OPTIONS DISPONIBLES 8
II.2.1 DISTRIBUTIONS 8
II.2.2 METHODES D'ESTIMATION ET DE QUANTIFICATION DES INCERTITUDES 9

III. EVALUATION EN UTILISANT DES DONNEES SIMULEES 10

III.1 PROCEDURE D'EVALUATION 10
III.2 CHOIX D'UNE METHODE D'ESTIMATION 13
III.2.1 DISTRIBUTIONS CLASSIQUES 13
III.2.2 DISTRIBUTIONS POUR LES VALEURS SUPERIEURES A UN SEUIL 15
III.2.3 DISTRIBUTIONS POUR LES MAXIMA 17
III.2.4 DISTRIBUTIONS POUR LES MINIMA 21
III.2.5 RECOMMANDATIONS 23
III.3 CHOIX D'UNE METHODE DE QUANTIFICATION DES INCERTITUDES 24
III.3.1 RESULTATS 24
III.3.2 RECOMMANDATION 26
III.4 APPLICABILITE A DE PETITS ECHANTILLONS 27
III.4.1 RESULTATS 27
III.4.2 RECOMMANDATION 29

IV. EVALUATION EN UTILISANT DES DONNEES REELLES 30

IV.1 PROCEDURE D'EVALUATION 30
IV.2 RESULTATS 31
IV.2.1 CORRECTION DE BUGS 31
IV.2.2 AJUSTEMENTS 31
IV.2.3 TESTS STATISTIQUES 34

V. CONCLUSIONS ET RECOMMANDATIONS 36

I. Introduction

L'opération HYDRO 3 a pour objectif de moderniser le système d'information de la prévision des crues et de l'hydrométrie. Parmi les nombreuses actions menées figure la modernisation des calculs permettant d'estimer des débits caractéristiques (et leur incertitude). Pour une station hydrométrique donnée, sur laquelle une série de données brutes est disponible, ces calculs comprennent les étapes suivantes :

1. **Extraction de la variable à analyser** : on peut par exemple extraire les minima/maxima/moyennes annuels, éventuellement après changement de pas de temps (par exemple du pas de temps variable au pas de temps journalier) et/ou application d'un filtre (par exemple moyenne mobile sur 7 jours).
2. **Ajustement d'une distribution** : en fonction de la variable à analyser, on choisit une distribution adéquate dont il faut estimer les paramètres. Cette estimation s'accompagne d'une quantification de l'incertitude d'échantillonnage, i.e. de l'incertitude liée à la taille réduite de la chronique.
3. **Calcul des quantiles** : à partir de la distribution estimée, on peut calculer un quantile de période de retour donnée, assorti de son incertitude.
4. **Tests statistiques** : la variable analysée peut également être soumise à plusieurs tests pour vérifier sa stationnarité (tests de tendance et de rupture) et vérifier la cohérence entre les données observées et la distribution estimée (test d'adéquation).

Dans le cadre de la convention SCHAPI-Irstea 2015, Irstea a fourni un code de calcul qui implémente les étapes 2 à 4 (mais qui ne traite pas de l'étape 1). Ce code de calcul modernise l'existant (implémenté dans le logiciel HYDRO2) en proposant de nouvelles distributions, de nouvelles méthodes d'estimation et de nouveaux outils (tests statistiques en particulier). De plus le code est implémenté sous une forme modulaire, et il est ouvert, largement commenté et documenté : ceci devrait faciliter la maintenance et les évolutions futures du code (nouveaux outils).

Le code de calcul est intégralement implémenté dans le langage R, et sera piloté par l'HydroPortail (le portail web de la banque Hydro 3), comme illustré dans la Figure 1 : l'HydroPortail extraira de la série brute la variable à analyser, puis transmettra cette variable au code R, accompagnée des spécifications sur la distribution à estimer (type de distribution, méthode d'estimation, etc.). A l'issue du calcul, le code R renverra les résultats (paramètres estimés, quantiles, intervalles de confiance, résultat des tests statistiques, etc.) à l'HydroPortail, qui les exploitera (graphiques, sauvegarde des valeurs calculées, etc.). La communication entre l'HydroPortail et le code R se fera via des fichiers d'échange normalisés au format JSON.

Le fait d'avoir implémenté ces outils sous le langage R présente l'avantage qu'un utilisateur averti pourra piloter directement le code de calcul, sans nécessairement passer par l'HydroPortail.

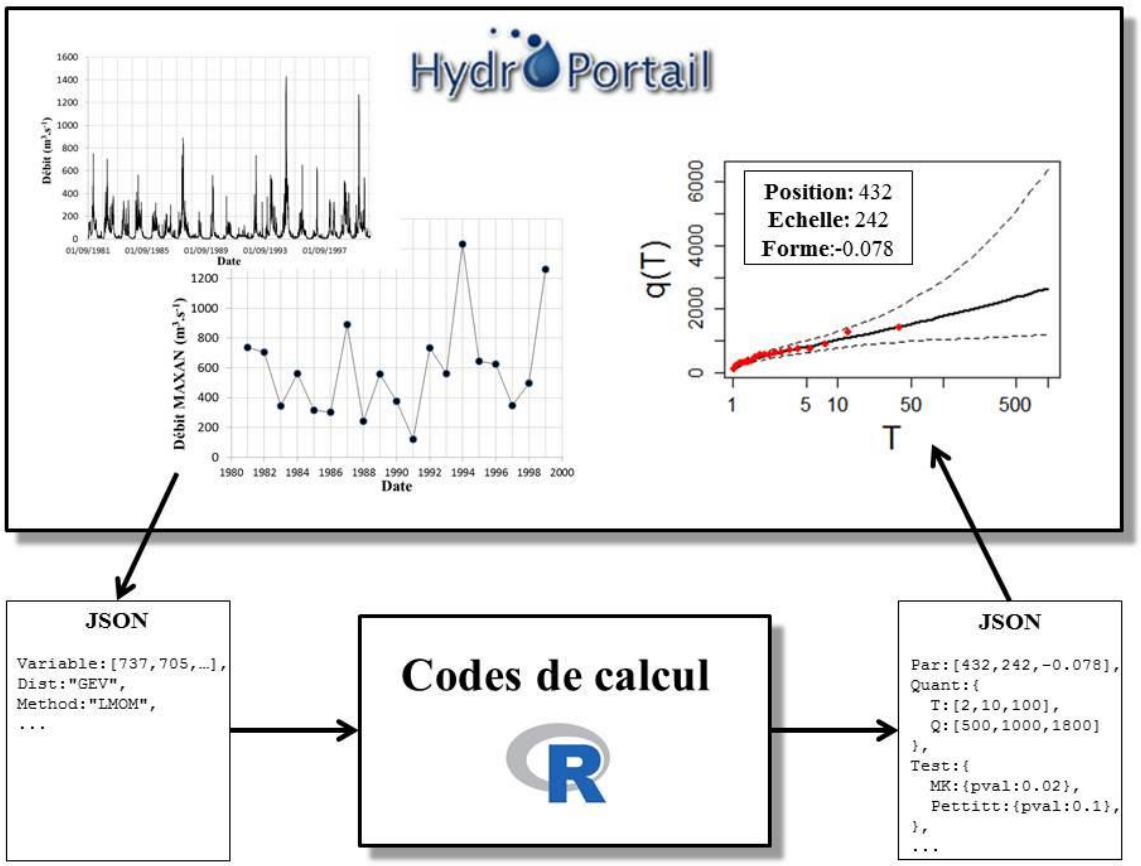


Figure 1. Répartition des tâches et communication entre l'HydroPortail et les codes de calcul sous R.

L'objectif de ce rapport est double : il s'agit tout d'abord de documenter les méthodes statistiques qui sont implémentées dans les codes R. Cet objectif est atteint via l'annexe à ce rapport, qui décrit en détail les formules et algorithmes utilisés. Le second objectif est de vérifier l'implémentation des codes, en particulier :

- Détecter et corriger les bugs éventuels ;
- Tester les codes sur un grand nombre de données simulées, afin d'évaluer les propriétés des différentes options disponibles, et de proposer des choix par défaut adéquats ;
- Tester les codes sur un grand nombre de données réelles sur toute la France Métropolitaine, afin de vérifier que les estimations effectuées sont réalistes d'un point de vue hydrologique, et de confirmer que les codes sont aptes à être appliqués en routine sur toutes les stations de la Banque Hydro 3.

Ce rapport est organisé comme suit : dans un premier temps les codes R sont brièvement décrits, en focalisant sur les options qui sont étudiées dans ce rapport (section II). L'évaluation basée sur des données simulées est ensuite décrite en section III, ainsi que les recommandations qui en découlent. La section 2 décrit l'application des codes à plus de 3000 stations de la banque HYDRO. Enfin, la section V résume les principales conclusions et recommandations qui émanent de ce travail.

II. Rapide description des codes

II.1 Généralités

Les codes se présentent comme un ensemble de scripts R qui définissent toutes les fonctions et objets constituant le code de calcul. Cependant, l'interface qui sera effectivement pilotée par l'HydroPortail est extrêmement simple puisqu'elle n'est constituée que d'une unique fonction, qui a la forme suivante :

```
h3 <- Hydro3_Estimation(y,dist,   # données, distribution
          Emeth,Umeth,options,    # méthodes et options
          Prior,mcmcoptions,      # options bayésiennes
          do.KS,do.MK,do.Pettitt) # application des tests ?
```

La fonction `Hydro3_Estimation` possède 2 arguments d'entrée obligatoires ainsi que 8 arguments optionnels qui sont décrits ci-dessous :

1. `y` [obligatoire] : vecteur d'observations (sans valeurs manquantes).
2. `dist` [obligatoire] : distribution à ajuster (cf. section II.2.1 pour la liste des distributions disponibles).
3. `Emeth` [optionnel] : méthode d'estimation des paramètres (cf. section II.2.2). Le choix de la valeur par défaut fait partie des objectifs de ce rapport.
4. `Umeth` [optionnel] : méthode de quantification des incertitudes (cf. section II.2.2). Le choix de la valeur par défaut fait partie des objectifs de ce rapport.
5. `options` [optionnel] : diverses options d'estimation. Des valeurs par défaut sont définies dans le code.
6. `Prior` [optionnel] : distributions *a priori*, uniquement utilisées pour l'estimation Bayésienne. Par défaut des distributions *a priori* non informatives sont utilisées.
7. `mcmcoptions` [optionnel] : diverses options pour le réglage de l'algorithme MCMC utilisé pour l'estimation Bayésienne. Des valeurs par défaut sont définies dans le code.
8. `do.KS,do.MK,do.Pettitt` [optionnel] : valeurs logiques (vrai/faux) indiquant s'il faut appliquer le test de Kolmogorov-Smirnov (test d'adéquation), le test de Mann-Kendall (détection de tendance) et le test de Pettitt (détection de rupture). Vrai par défaut pour les trois tests.

En sortie, la fonction `Hydro3_Estimation` renvoie un objet `h3` qui contient tous les résultats du calcul (paramètres estimés, quantiles, résultats des tests, etc.). Le contenu détaillé de cet objet, ainsi que de tous les arguments d'entrée de la fonction, fait l'objet d'une modélisation UML (Figure 2) afin de normaliser les échanges entre l'HydroPortail et le code R.

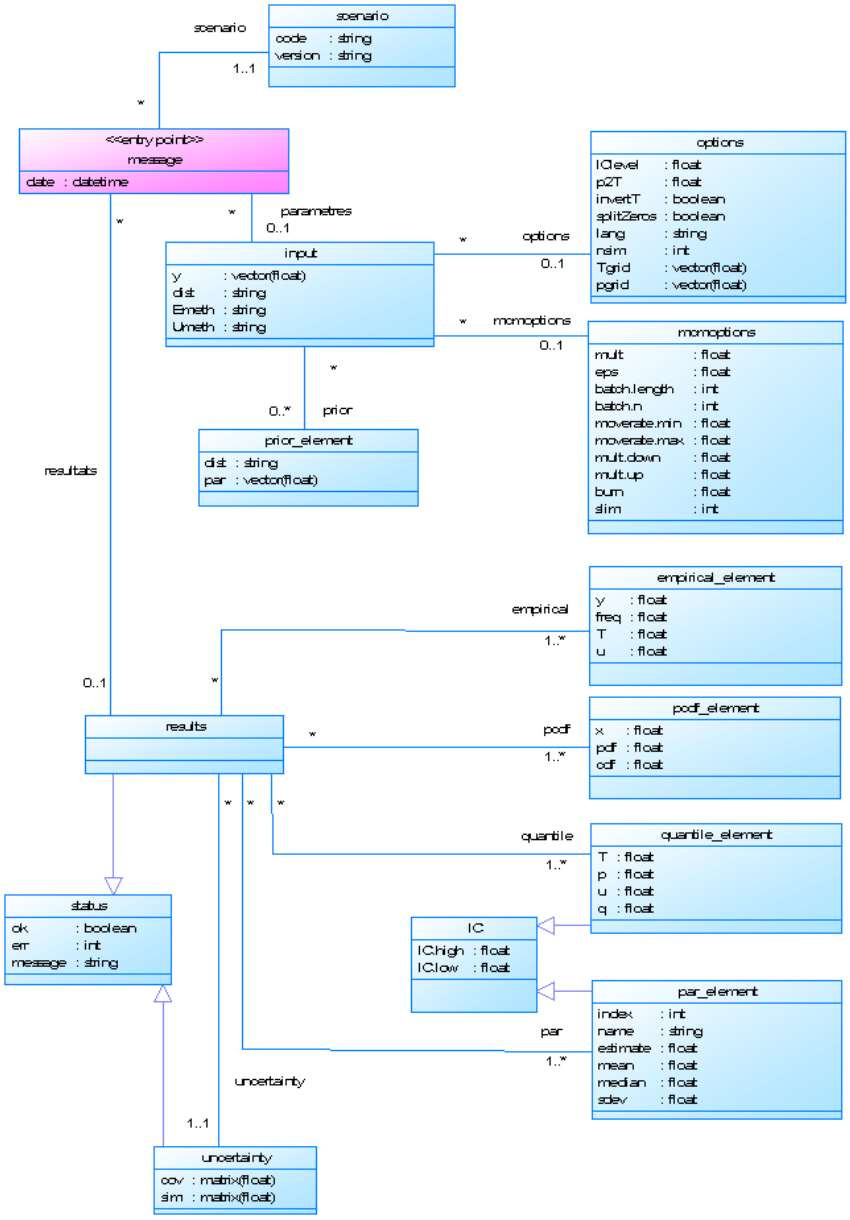


Figure 2. Diagramme de classes UML décrivant les arguments d'entrée et de sortie du code R. Version 1.0.1 non définitive datée du 02/05/2016.

II.2 Principales options disponibles

II.2.1 Distributions

Les distributions disponibles sont résumées dans le Tableau 1. Le choix d'une distribution adéquate dépend de la variable étudiée, et sera de la responsabilité de l'utilisateur. L'HydroPortail pourra néanmoins guider ce choix en proposant par défaut des distributions recommandées en fonction du type de variable étudiée (dernière colonne du tableau). Précisons que les deux dernières distributions de ce tableau sont implémentées dans le code R mais n'ont pas vocation à être interfacées dans l'HydroPortail (pour le moment en tout cas). Notons également que la loi de Poisson est particulière car elle ne peut être utilisée que pour des valeurs entières positives. De plus toutes les méthodes d'estimation sont équivalentes pour cette distribution : elle ne sera donc pas incluse dans les analyses menées dans ce rapport.

Tableau 1. Description des distributions disponibles.

Distribution	Identifiant dans le code R	# par.	Exemples d'utilisation*
Distributions utilisables par HydroPortail			
Loi Normale	`"Normal"`	2	QA
Loi log-normale	`"LogNormal"`	2	QA, QN
Loi de Gumbel	`"Gumbel"`	2	QX
Loi généralisée des valeurs extrêmes	`"GEV"`	3	QX
Loi de Pearson III	`"PearsonIII"`	3	QX
Loi de log-Pearson III	`"LogPearsonIII"`	3	QX
Loi exponentielle	`"Exponential2"`	2	QS
Loi de Pareto généralisée	`"GPD3"`	3	QS
Loi de Gumbel pour les minima	`"Gumbel_min"`	2	QN
Loi généralisée des valeurs extrêmes pour les minima	`"GEV_min"`	3	QN
Loi de Poisson	`"Poisson"`	1	N
Distributions disponibles dans le code, sans utilisation par HydroPortail			
Loi exponentielle à 1 paramètre (seuil nul)	`"Exponential1"`	1	QS
Loi de Pareto généralisée à 1 paramètre (seuil nul)	`"GPD2"`	2	QS

* QA = débit annuel, QN = débit minimal, QX = débit maximal, QS = débit supérieur à un seuil, N = comptage.

II.2.2 Méthodes d'estimation et de quantification des incertitudes

Quatre méthodes d'estimation des paramètres ont été implémentées : méthodes des moments (MOM), des L-moments (LMOM), du maximum de vraisemblance (ML) et estimation Bayésienne (BAY). Une description plus approfondie de ces méthodes est consultable dans l'annexe à ce rapport.

De plus, cinq méthodes de quantification des incertitudes sont disponibles : Bootstrap (BOOT), Bootstrap paramétrique (PBOOT), maximum de vraisemblance (ML), approche Bayésienne (BAY) et aucune quantification des incertitudes (NONE). Comme précédemment, les détails techniques sont disponibles dans l'annexe.

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 bien particulière (typiquement, ML et BAY). Les combinaisons possibles sont résumées dans le Tableau 2.

Tableau 2. Combinaisons possibles entre méthode d'estimation des paramètres et méthode de quantification des incertitudes.

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

III. Evaluation en utilisant des données simulées

L'objectif d'une évaluation utilisant des données simulées est d'évaluer et comparer la performance de différentes méthodes dans un cas idéal où la vraie valeur de la quantité à estimer est connue. Ceci permet de calculer des critères de performance et de hiérarchiser les méthodes de manière objective.

III.1 Procédure d'évaluation

La procédure d'évaluation peut être résumée par le pseudo-algorithme suivant.

Algorithme 1. Procédure de simulation.

0. A fixer : nombre de simulations N_{sim}, distribution parente $D(\theta)$, taille de la série simulée *N*, période de retour *T* du quantile à estimer.
1. Répéter pour $i = 1 : N_{sim}$
   a. Simuler une série de *N* réalisations issues de la distribution parente.
   b. Estimer les paramètres et leur incertitude.
   c. Calculer le quantile $q_T^{(i)}$ et son intervalle de confiance à 90% $I_T^{(i)}$.
2. Calculer les indices de performances :
   a. Temps de calcul pour effectuer les N_{sim} simulations.
   b. Taux d'échec sur les N_{sim} simulations.
   c. Biais (à minimiser) : $\frac{1}{N_{sim}} \sum_{i=1}^{N_{sim}} q_T^{(i)} - q_T^{(\text{vrai})}$
   d. Erreur-type (à minimiser) : $\sqrt{\frac{1}{N_{sim}} \sum_{i=1}^{N_{sim}} \left(q_T^{(i)} - q_T^{(\text{vrai})} \right)^2}$
   e. Taux de recouvrement (devrait être proche de 90%) : $100 \times \frac{\sum_{i=1}^{N_{sim}} 1\left\{ q_T^{(\text{vrai})} \in I_T^{(i)} \right\}}{N_{sim}}$

En pratique, nous réalisons $N_{sim} = 1000$ simulations, ce qui permet d'obtenir des résultats fiables tout en conservant un temps de calcul raisonnable. Nous faisons varier les autres paramètres de cet algorithme de la façon suivante :

- Distribution parente $D(\theta)$: toutes les distributions du Tableau 1 sont évaluées tour à tour. Pour certaines distributions, plusieurs paramétrages sont évalués afin de faire varier les propriétés de la distribution parente (notamment son asymétrie et sa capacité à générer des extrêmes). Toutes les distributions parentes utilisées dans cet exercice sont représentées par leur densité dans la Figure 3.
- Taille de la série simulée N : varie entre 20 et 80, ce qui balaie la gamme d'ancienneté de la majorité des stations de la banque HYDRO. Un exercice spécifique est également réalisé pour les petits échantillons (section III.4) : dans ce cas N varie entre 3 et 15.
- Période de retour T du quantile à estimer : nous utilisons les valeurs 10, 50 et 100.

Les indices de performance utilisés peuvent être commentés comme suit :

- Le temps de calcul est une indication importante en pratique. En effet, les codes ont vocation à être lancés en batch sur un grand nombre de stations (pour la réalisation des fiches dites « synthèse » notamment). Un temps de calcul trop long serait donc rédhibitoire. Les temps de calcul indiqués dans ce rapport sont des temps cpu obtenus sur un ordinateur portable munis de processeurs Intel i7-4800MQ@2.70GHz, avec 16Go de mémoire vive.
- Le taux d'échec est également important en pratique. Certaines méthodes d'estimation réclament d'utiliser des méthodes numériques (pour résoudre une équation implicite ou optimiser une fonction) qui peuvent parfois ne pas converger. Dans ce cas le code R renvoie un message d'erreur, mais il convient de s'assurer que cette situation reste rare.
- Le biais mesure si *en moyenne sur toutes les simulations*, le quantile estimé est proche du vrai quantile. Un biais proche de zéro est évidemment souhaitable.
- L'erreur-type mesure l'amplitude de l'erreur typiquement effectuée sur les simulations. Cet indice est également appelé l'erreur quadratique moyenne. L'erreur-type doit être aussi faible que possible.
- Le taux de recouvrement correspond à la fréquence avec laquelle le vrai quantile se trouve à l'intérieur de l'intervalle de confiance à 90%. Si l'incertitude est correctement quantifiée, ce taux de recouvrement devrait être proche de 90%.

Notre stratégie d'évaluation se décompose en trois étapes :

Etape 1. Choix d'une méthode d'estimation. L'objectif est de sélectionner la méthode d'estimation à proposer par défaut, parmi MOM (moments), LMOM (L-moments) et ML (maximum de vraisemblance). La méthode BAY (estimation Bayésienne) est directement exclue comme choix par défaut, pour deux raisons : (i) son temps de calcul est beaucoup plus long que les autres méthodes (quelques dizaines de secondes typiquement, ce qui est acceptable pour une utilisation sur une station, moins pour une application en batch sur toutes les stations de la banque HYDRO) ; (ii) ses performances dépendent des distributions *a priori* spécifiées par l'utilisateur.

Etape 2. Choix d'une méthode de quantification des incertitudes. Ayant sélectionné une méthode d'estimation, les méthodes de quantification des incertitudes applicables à cette méthode d'estimation seront comparées.

Etape 3. Choix d'une taille minimale d'échantillon. Le couple méthode d'estimation / méthode de quantification des incertitudes sélectionné sera appliqué à de très petits échantillons, afin d'évaluer s'il est pertinent d'imposer une taille minimale de série.

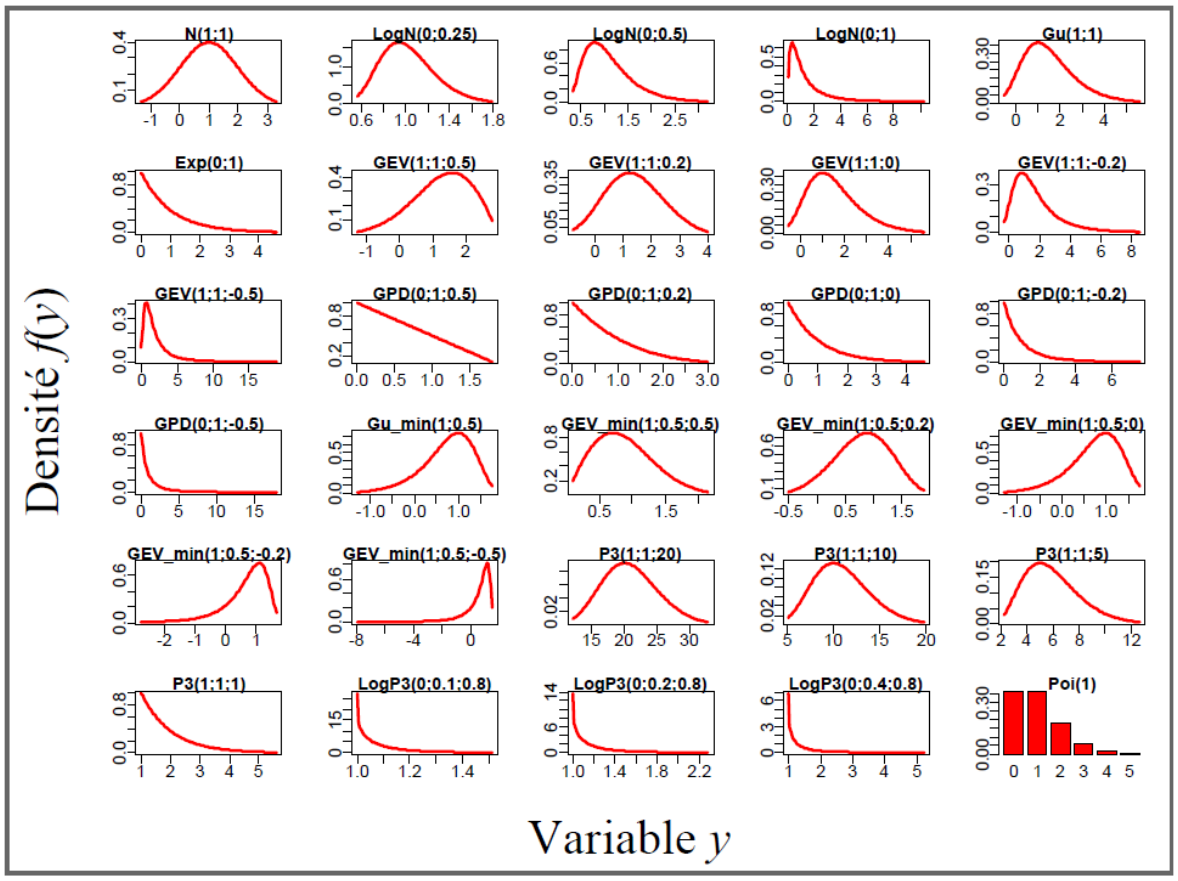


Figure 3. Densités de probabilité des distributions parentes utilisées dans cette analyse.

III.2 Choix d'une méthode d'estimation

Cette première étape se focalise sur le choix d'une méthode d'estimation. Nous reportons donc, pour chaque distribution, les indices de performance obtenus avec les méthodes MOM, LMOM et ML. L'indice de performance « taux de recouvrement » n'est pas évalué ici car il est spécifique aux incertitudes.

III.2.1 Distributions classiques

Loi Normale

Les trois méthodes sont similaires en termes de temps de calcul, de taux d'échec (toujours nul) et d'erreur-type. La méthode LMOM n'est pas biaisée, alors que les deux autres méthodes présentent un faible biais pour les petits échantillons (environ -2% pour $N = 20$).

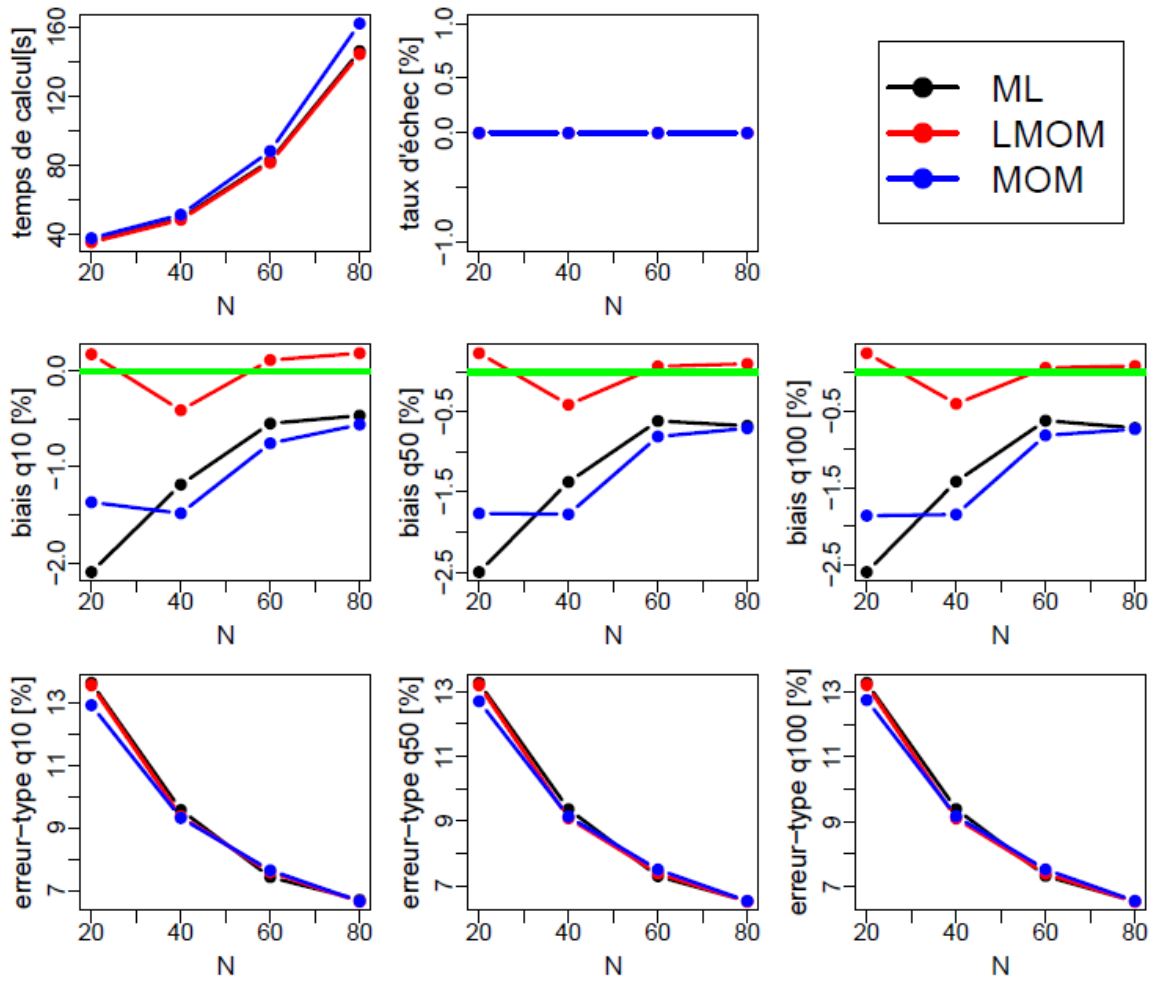


Figure 4. Indices de performance des trois méthodes d'estimation des paramètres, en fonction de la taille de l'échantillon. Vraie distribution : N(1,1).

Loi Log-normale

Comme pour la loi normale, les trois méthodes sont similaires en termes de temps de calcul, de taux d'échec (toujours nul) et d'erreur-type. La méthode LMOM semble légèrement moins biaisée, mais dans tous les cas les biais restent faibles.

Les résultats sont similaires pour d'autres valeurs des paramètres (non illustrés ici). Néanmoins, pour une loi log-normale très asymétrique ($LogN(0;1)$, voir Figure 3), la méthode ML n'a aucun biais, alors que LMOM a un léger biais positif d'environ 5% pour $N = 20 - 40$.

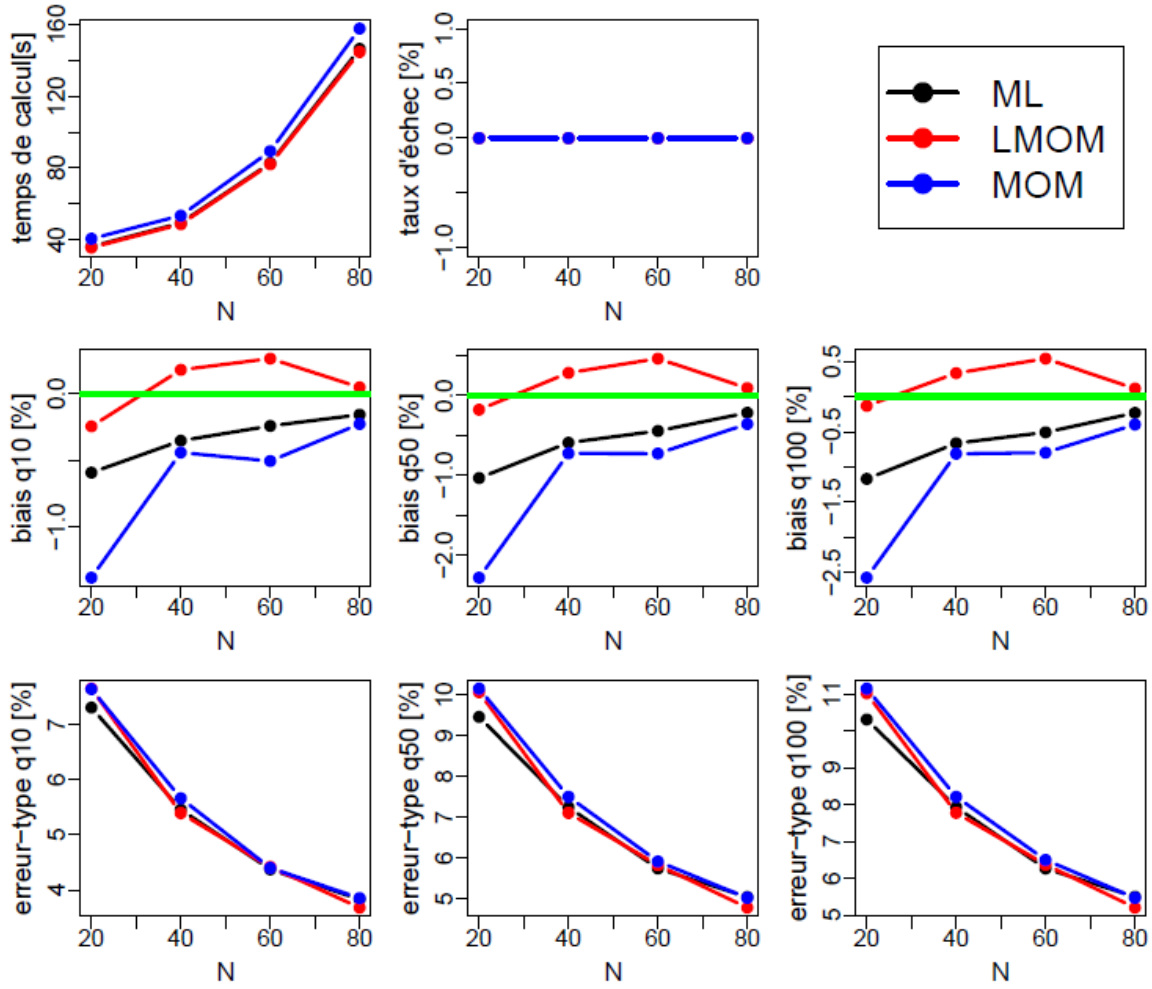


Figure 5. Indices de performance des trois méthodes d'estimation des paramètres, en fonction de la taille de l'échantillon. Vraie distribution : LogN(0,0.25).

III.2.2 Distributions pour les valeurs supérieures à un seuil

Loi exponentielle

Les trois méthodes sont indissociables en termes de temps de calcul et de taux d'échec (toujours nul). Les biais sont faibles mais la méthode MOM semble la plus affectée, suivie de ML puis de LMOM. En termes d'erreur-type, les méthodes sont encore assez similaires, avec un très léger avantage pour ML sur LMOM.

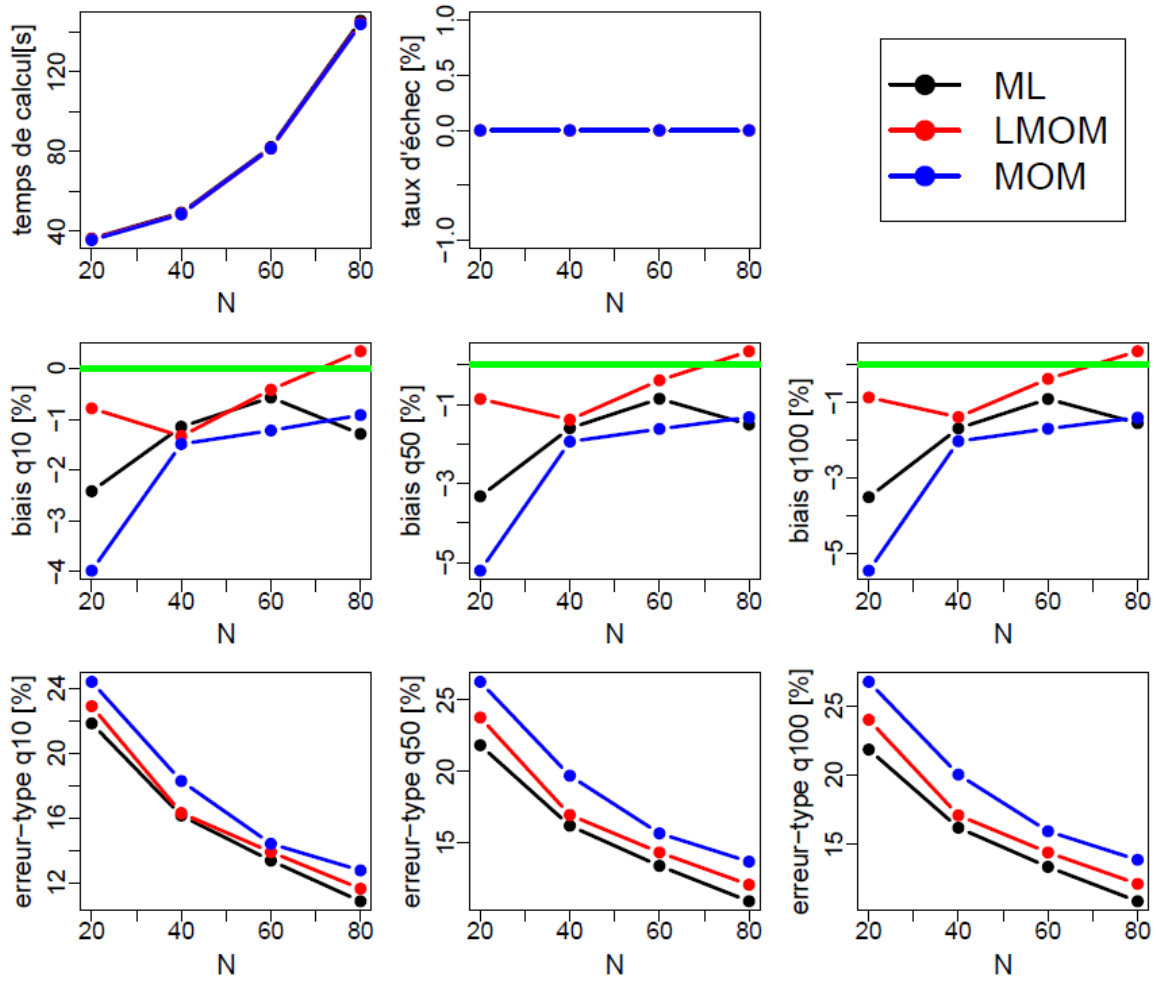


Figure 6. Indices de performance des trois méthodes d'estimation des paramètres, en fonction de la taille de l'échantillon. Vraie distribution : Exp(0,1).

Loi GPD

La méthode ML se démarque sur les deux premiers critères, puisque son temps de calcul est environ 2 fois plus important et son taux d'échec, quoique faible, n'est pas nul (environ 2-3%). Ceci est dû au fait que la méthode ML requiert une optimisation numérique, alors que les deux autres méthodes conduisent à des formules explicites, plus rapides à calculer et plus fiables.

En termes de biais, la méthode LMOM est assez nettement la plus performante. Les biais sont ici non négligeables, et peuvent dépasser 10% (en valeur absolue) pour le quantile centennal dans le cas des méthodes ML et MOM. La méthode ML peut présenter une erreur-type beaucoup plus forte que les autre méthodes pour de petits échantillons ($N = 20$).

Les résultats sont similaires pour d'autres valeurs des paramètres (non illustrés ici). Les performances de la méthode ML se dégradent très fortement quand le paramètre de forme est très négatif ($\xi = -0.5$, ce qui correspond à une distribution produisant des extrêmes très forts, voir Figure 3).

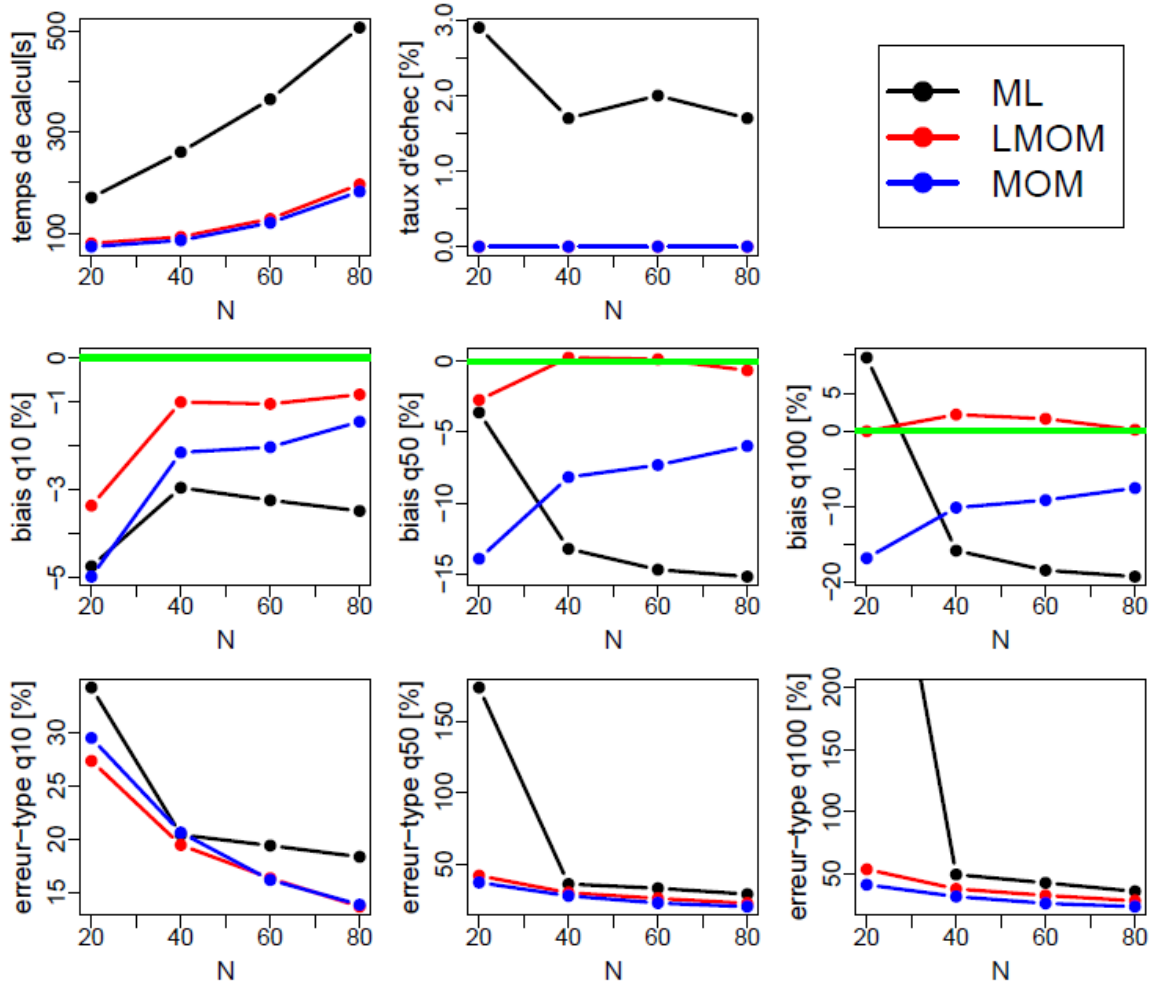


Figure 7. Indices de performance des trois méthodes d'estimation des paramètres, en fonction de la taille de l'échantillon. Vraie distribution : GPD(0,1,-0.2).

III.2.3 Distributions pour les maxima

Loi de Gumbel

La méthode ML a un temps de calcul légèrement supérieur aux autres méthodes, car elle fait appel à une méthode d'optimisation numérique. Néanmoins le taux d'échec reste cette fois nul.

En termes de biais, la méthode LMOM est encore une fois la plus performante. Les biais restent cependant modérés dans tous les cas (pas plus de 4% en valeur absolue).

Les trois méthodes sont similaires en termes d'erreur-type, avec un léger avantage pour ML devant LMOM puis MOM.

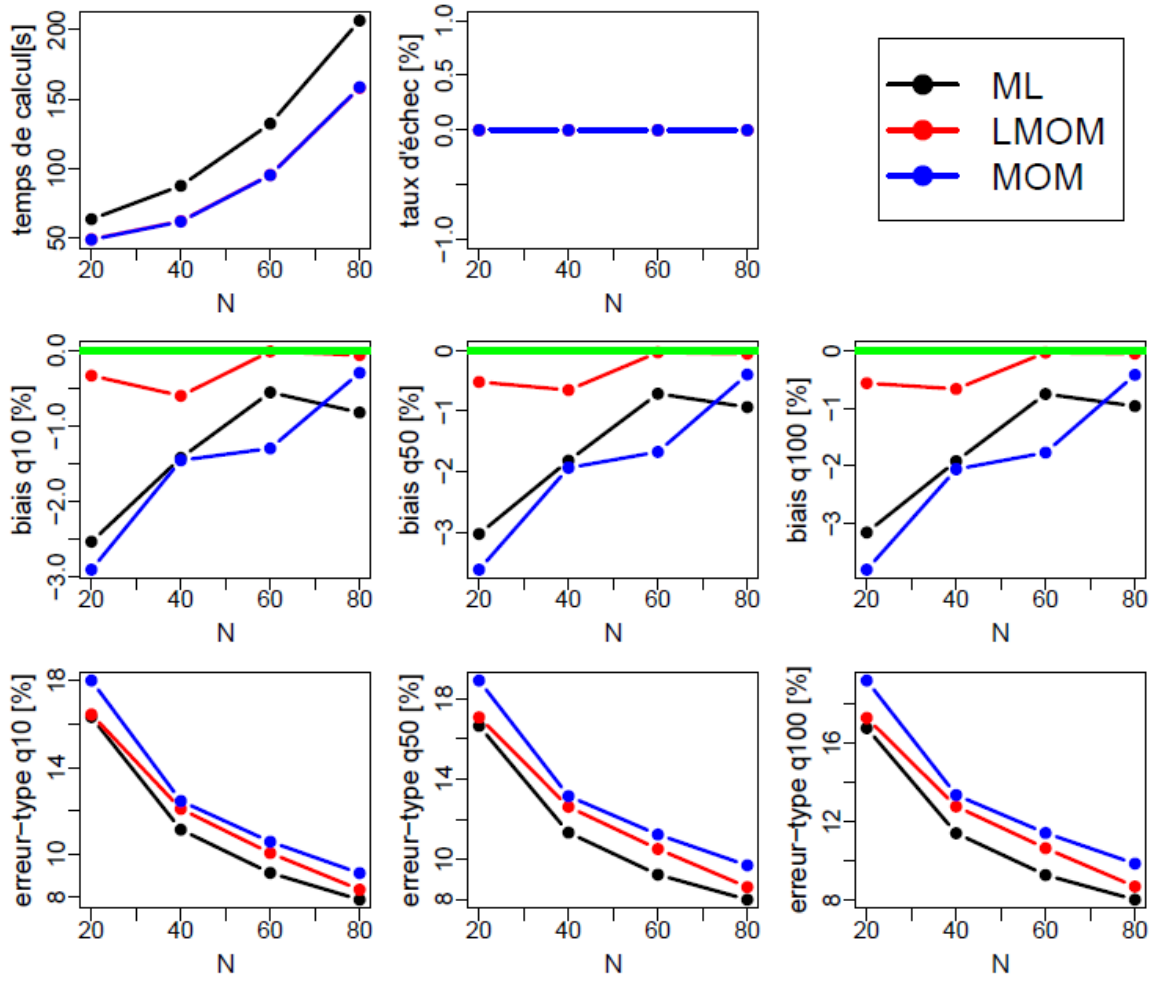


Figure 8. Indices de performance des trois méthodes d'estimation des paramètres, en fonction de la taille de l'échantillon. Vraie distribution : Gum(1,1).

Loi GEV

Les résultats sont qualitativement similaires à ceux de la distribution GPD, qui est aussi une distribution d'extrêmes à 3 paramètres :

- Temps de calcul et taux d'échec plus important pour ML (dû à une optimisation numérique).
- Quasi-absence de biais pour LMOM, alors que les deux autres méthodes présentent des biais pouvant être importants (plus de 20% en valeur absolue).
- Erreurs-types comparables pour LMOM et MOM, mais nettement plus fortes pour ML.
- Résultats similaires pour d'autres valeurs des paramètres (non illustrés ici), avec des performances qui se dégradent pour MOM et surtout ML lorsque le paramètre de forme est très négatif ($\xi = -0.5$, voir Figure 3).

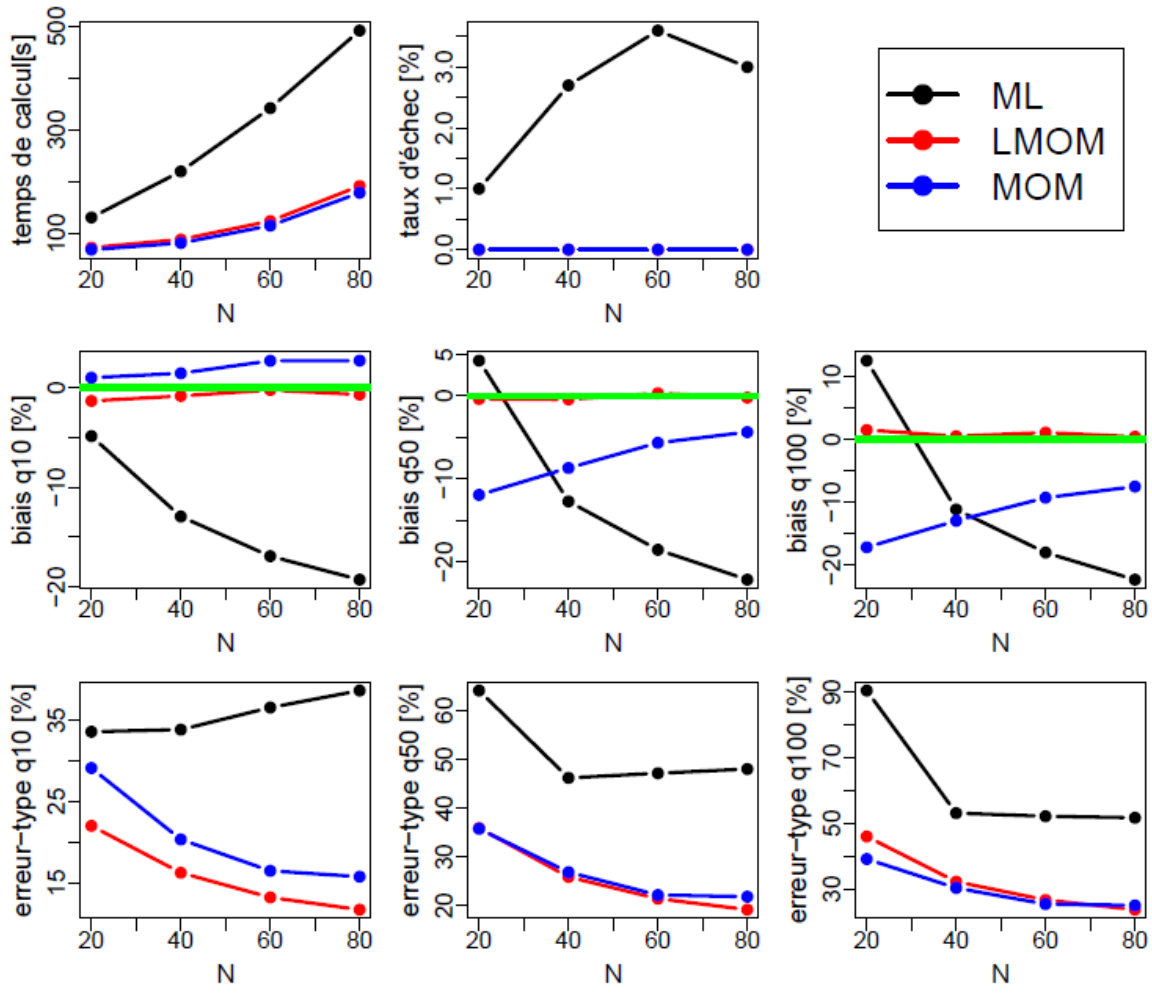


Figure 9. Indices de performance des trois méthodes d'estimation des paramètres, en fonction de la taille de l'échantillon. Vraie distribution : GEV(1,1,-0.2).

Loi de Pearson III

Le temps de calcul est beaucoup plus important pour ML que pour les autres méthodes, et le taux d'échec est très important (20-30%). Notons que MOM a également un taux d'échec d'environ 10% pour des petits échantillons.

Les biais sont assez faibles pour toutes les méthodes, la méthode LMOM semblant encore une fois la moins biaisée. Les erreurs-types sont comparables pour les trois méthodes.

Les résultats sont qualitativement similaires pour d'autres valeurs des paramètres (non illustrés ici). On note cependant les tendances suivantes :

- Le taux d'échec de la méthode MOM augmente lorsque le troisième paramètre (paramètre de forme) augmente (correspondant à une distribution plus symétrique, voir Figure 3).
- Les biais sont plus importants lorsque le paramètre de forme diminue.
- Pour une distribution très asymétrique, le taux d'échec de la méthode ML peut atteindre 60%.

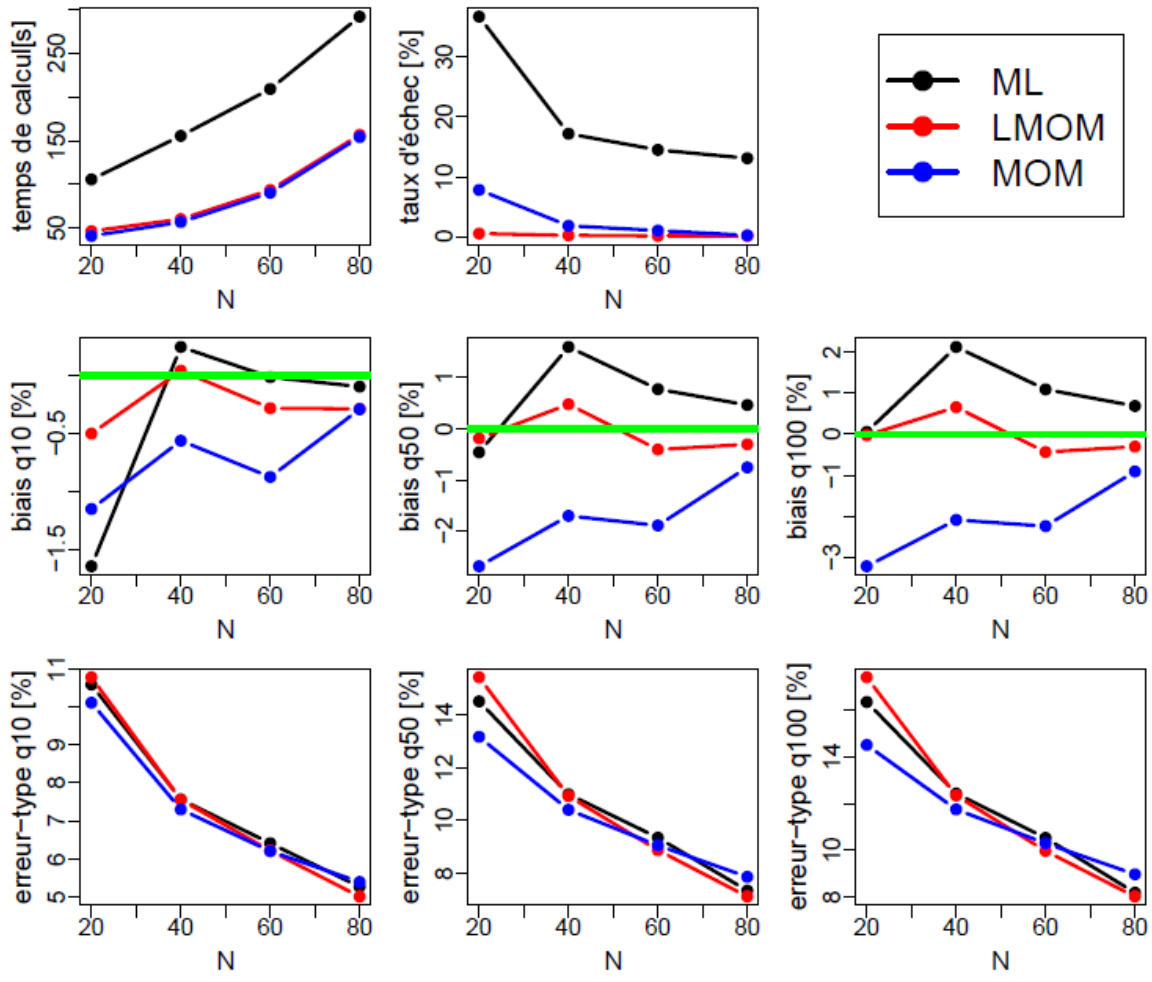


Figure 10. Indices de performance des trois méthodes d'estimation des paramètres, en fonction de la taille de l'échantillon. Vraie distribution : PIII(1,1,5).

Loi de Log-Pearson III

La méthode ML semble inappropriée pour cette distribution, puisque son taux d'échec atteint les 80%. Les deux autres méthodes (MOM et LMOM) se valent globalement, et les résultats sont très similaires avec d'autres valeurs des paramètres.

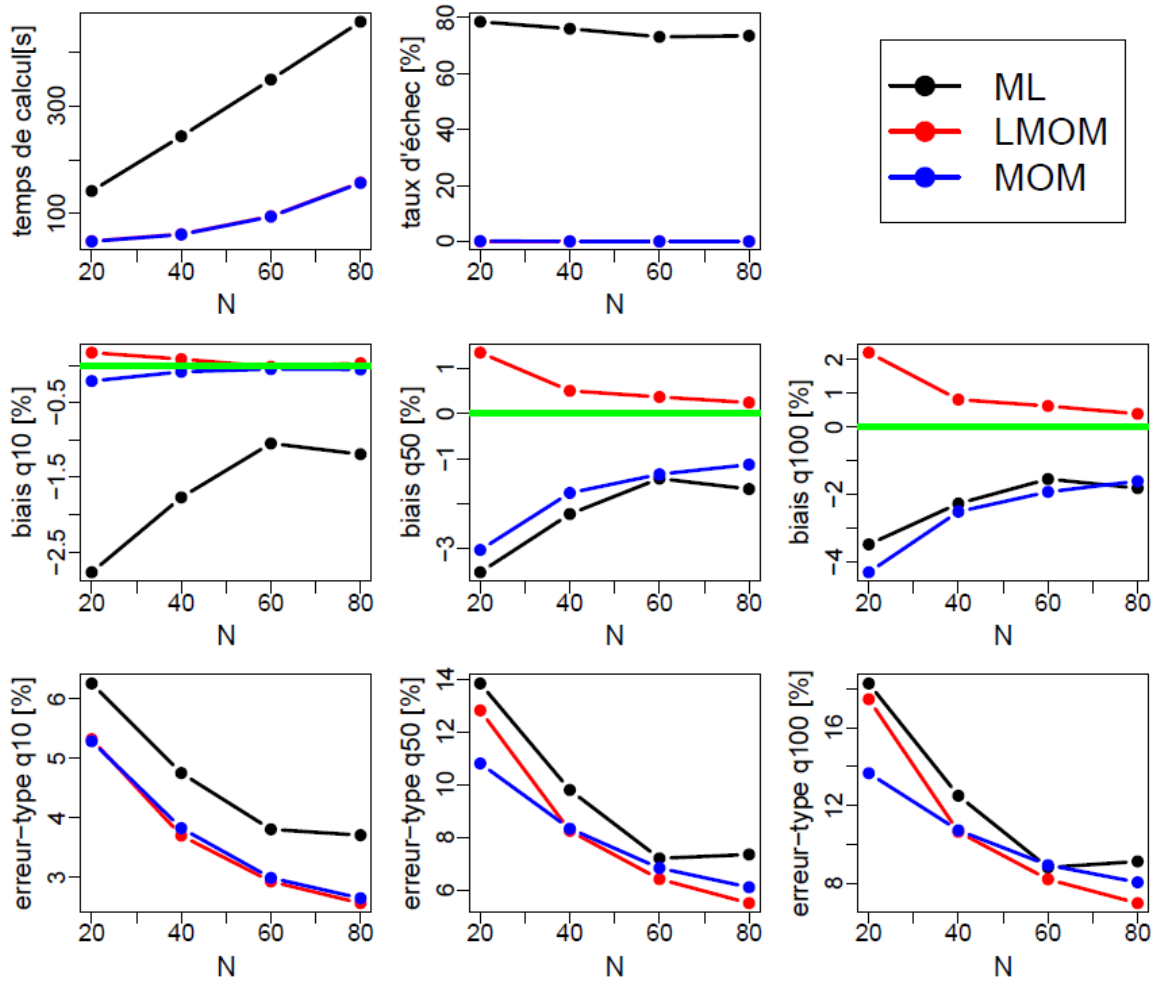


Figure 11. Indices de performance des trois méthodes d'estimation des paramètres, en fonction de la taille de l'échantillon. Vraie distribution : LogPIII(0,0.1,0.8).

III.2.4 Distributions pour les minima

Loi de Gumbel pour les minima

Les résultats sont qualitativement similaires à ceux trouvés pour la loi de Gumbel classique :

- La méthode ML a un temps de calcul légèrement supérieur aux autres méthodes, mais son taux d'échec reste nul.
- La méthode LMOM est encore une fois la moins biaisée. Les biais peuvent être importants pour les autres méthodes (plus de 20% en valeur absolue).
- Les trois méthodes sont similaires en termes d'erreur-type, avec un léger avantage pour ML devant LMOM puis MOM.

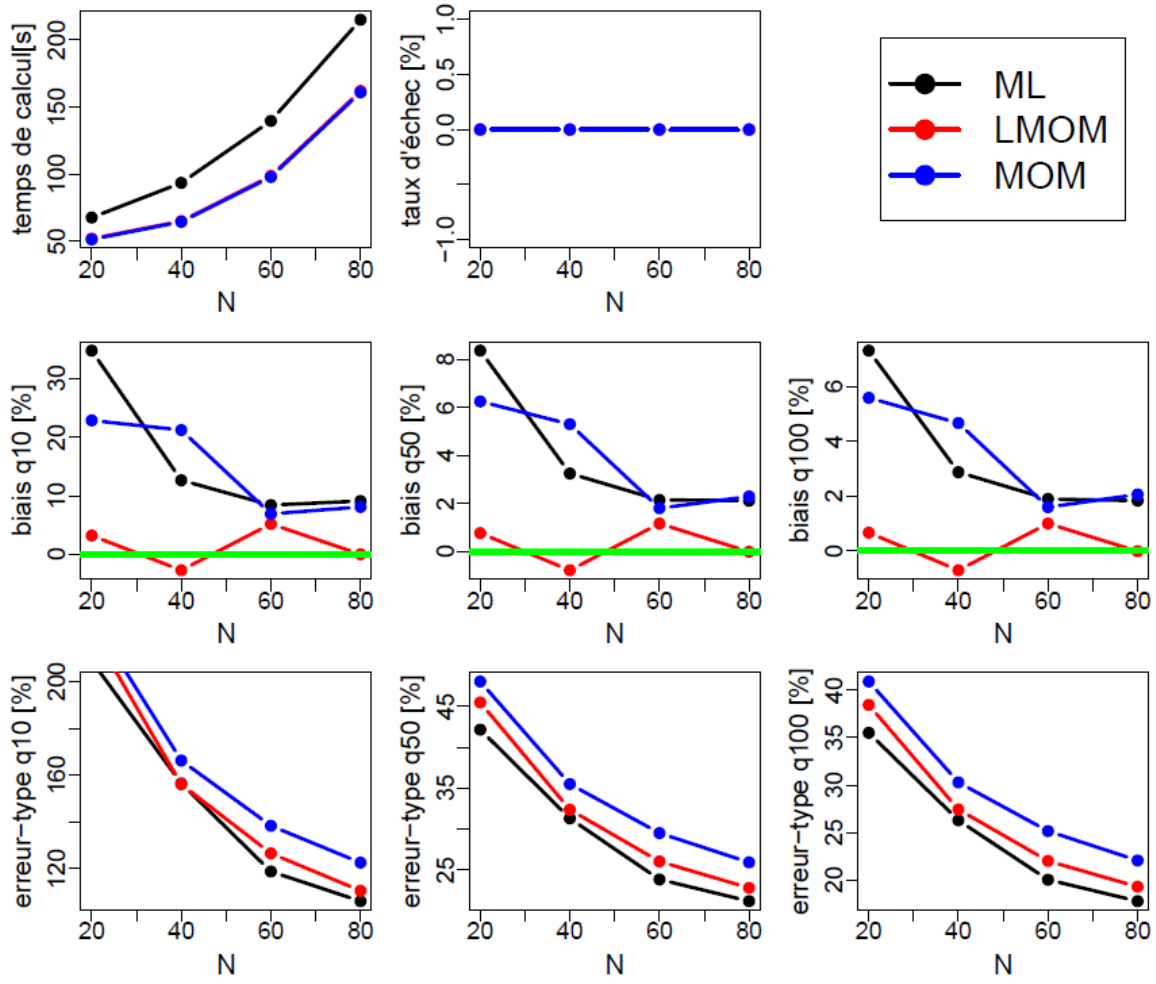


Figure 12. Indices de performance des trois méthodes d'estimation des paramètres, en fonction de la taille de l'échantillon. Vraie distribution : Gum_min(1,0.5).

Loi GEV pour les minima

Les résultats sont qualitativement similaires à ceux de la distribution GEV classique :

- Temps de calcul plus important pour ML (mais taux d'échec nul).
- Biais plus faible pour LMOM, alors que les deux autres méthodes présentent des biais pouvant être importants (plus de 30% en valeur absolue).
- Erreurs-types comparables pour les trois méthodes, avec un léger avantage pour MOM.
- Résultats assez similaires pour d'autres valeurs des paramètres (non illustrés ici), mais les performances de MOM et ML se dégradent lorsque le paramètre de forme est très positif ou très négatif ($\xi = 0.5$ ou $\xi = -0.5$, voir Figure 3).

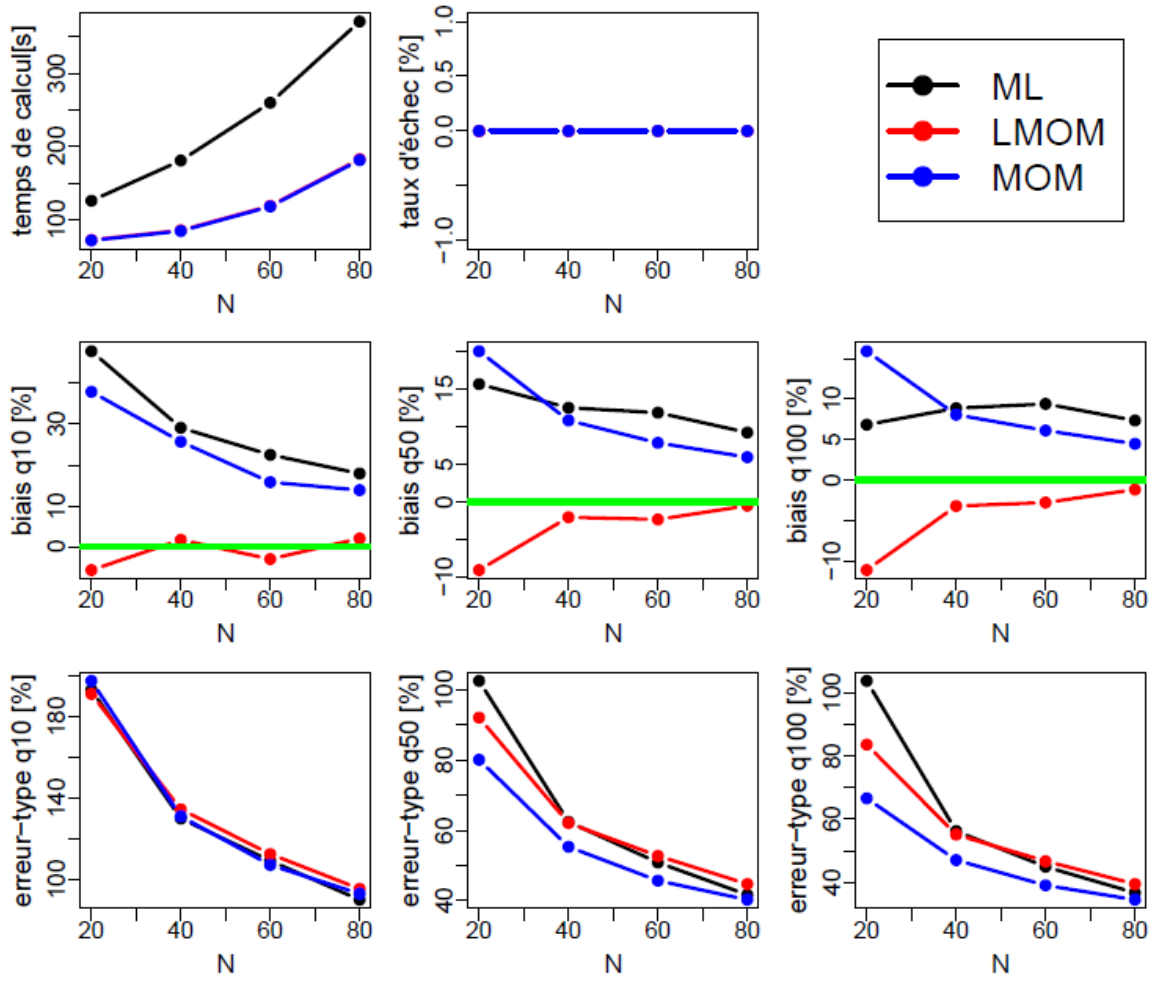


Figure 13. Indices de performance des trois méthodes d'estimation des paramètres, en fonction de la taille de l'échantillon. Vraie distribution : GEV_min(1,0.5,0.2).

III.2.5 Recommandations

Quelques tendances générales se dégagent des résultats de cette première étape :

- La méthode LMOM est en général non biaisée, alors que les méthodes ML et MOM présentent des biais non négligeables pour certaines distributions.
- Dans la plupart des cas, les erreurs-types sont comparables pour les trois méthodes.
- Pour plusieurs distributions, le temps de calcul de la méthode ML est nettement plus important.
- Pour les distributions Pearson III et Log-Pearson III, le taux d'échec de la méthode ML est très fort.

En résumé, la méthode LMOM est parfois bien meilleure que ses compétitrices, mais elle n'est jamais bien pire. Il semble donc naturel de la proposer comme méthode d'estimation des paramètres par défaut pour toutes les distributions. Ceci a l'avantage de la simplicité, et n'a pas réellement de conséquence néfaste puisque la plus-value des autres méthodes pour certaines distributions et certains critères est dans tous les cas assez faible.

Pour les distributions Pearson III et Log-Pearson III, il est également recommandé de ne pas proposer la méthode d'estimation ML, étant donné son fort taux d'échec.

III.3 Choix d'une méthode de quantification des incertitudes

La méthode d'estimation LMOM étant choisie, il s'agit à présent de sélectionner la méthode de quantification des incertitudes parmi celles applicables à LMOM, c'est-à-dire parmi BOOT (Bootstrap) et PBOOT (Bootstrap paramétrique).

III.3.1 Résultats

La procédure de simulation de l'Algorithme 1 est encore une fois appliquée à toutes les distributions. Il n'y a pas de grande différence en termes de temps de calcul entre BOOT et PBOOT (non illustré ici) : en guise d'ordre de grandeur, il faut environ 300s (5 minutes) pour réaliser 1000 estimations. Le temps de calcul dépend légèrement de la distribution, et augmente modérément en fonction de la taille de l'échantillon (augmentation d'environ 90s (1 minute 30) pour 1000 simulations entre N = 20 et N = 80). Pour les deux méthodes BOOT et PBOOT associées à la méthode d'estimation LMOM, le taux d'échec est nul pour toutes les distributions sauf pour la distribution PearsonIII (taux d'échec faible entre 0 et 0.7% en fonction du paramétrage).

Etant donné que ni le temps de calcul ni le taux d'échec ne montrent de réelles différences entre les méthodes BOOT et PBOOT, le choix doit être fait sur la base du taux de recouvrement, qui est donc l'unique critère de performance qui est montré dans les figures ci-dessous.

Le Tableau 3 regroupe les taux de recouvrement calculés pour toutes les distributions. Les résultats sont très clairs : la méthode PBOOT est bien plus performante que la méthode BOOT, puisque son taux de recouvrement est toujours plus proche du taux cible de 90%. Pour certaines distributions (notamment les distributions d'extrêmes à 3 paramètres GEV et GPD3), les intervalles de confiance à 90% calculés avec BOOT ne contiennent la vraie valeur du quantile que dans 60% des cas, ce qui correspond à une sous-estimation assez marquée des incertitudes. En comparaison, la méthode PBOOT parvient à maintenir un taux de recouvrement supérieur à 80%. Des résultats similaires avaient déjà été signalés dans la littérature scientifique[1].

De manière générale, on observe également que la sous-estimation des incertitudes tend à diminuer lorsque la taille N de l'échantillon augmente, et que cette sous-estimation est plus marquée pour les lois à 3 paramètres que pour les lois à 2 paramètres.

[1] Kysely, J. (**2008**), A Cautionary Note on the Use of Nonparametric Bootstrap for Estimating Uncertainties in Extreme-Value Models, *Journal of Applied Meteorology and Climatology*, 47(12), 3236-3251.

Tableau 3. Taux de recouvrement des méthodes BOOT (bleu) et PBOOT (rouge) en fonction de la taille de l'échantillon pour différentes distributions parentes.

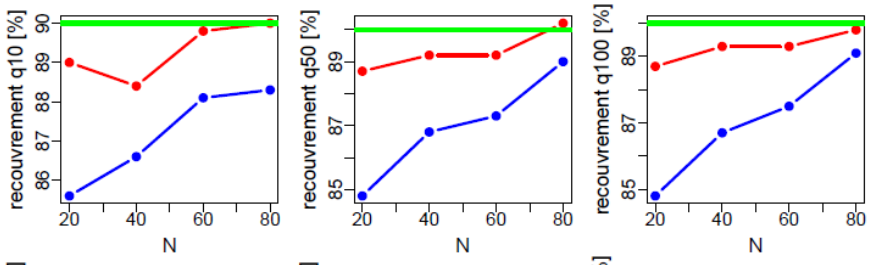
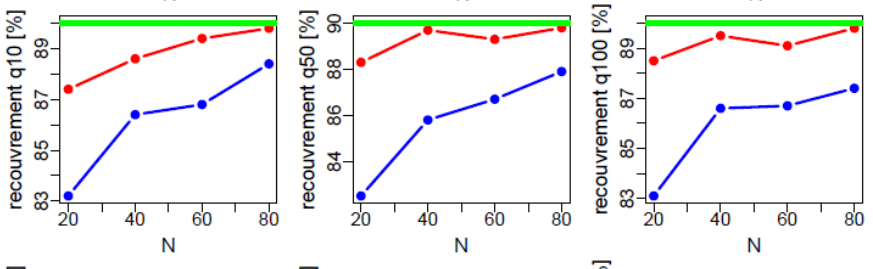
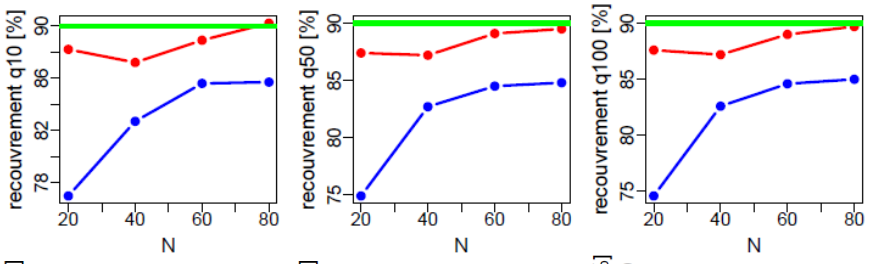
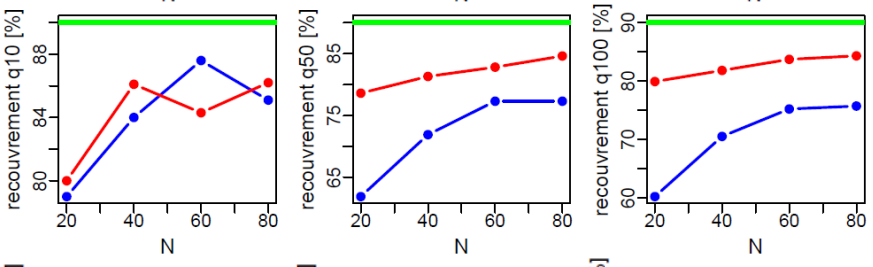
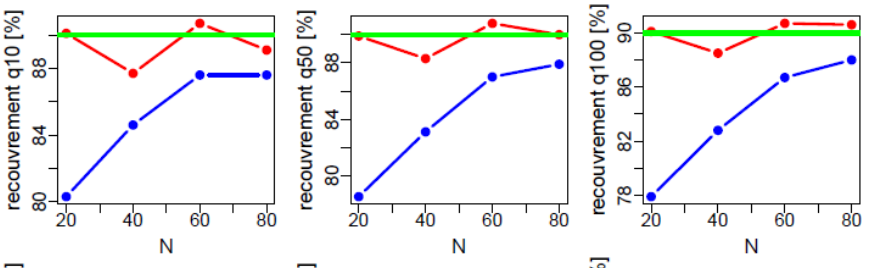
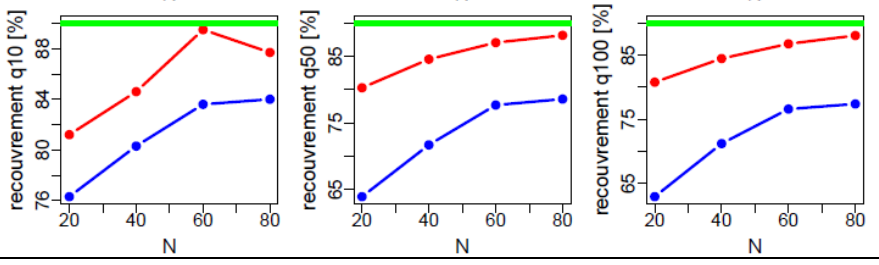
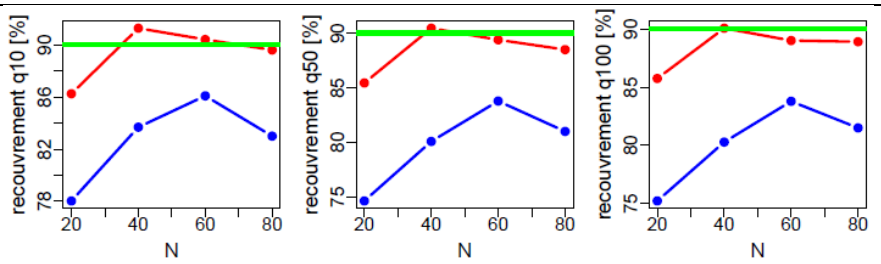
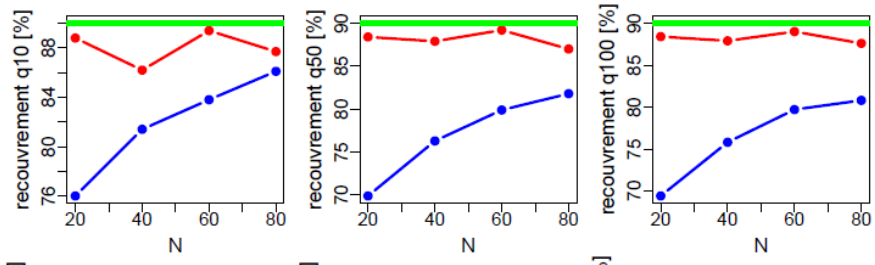
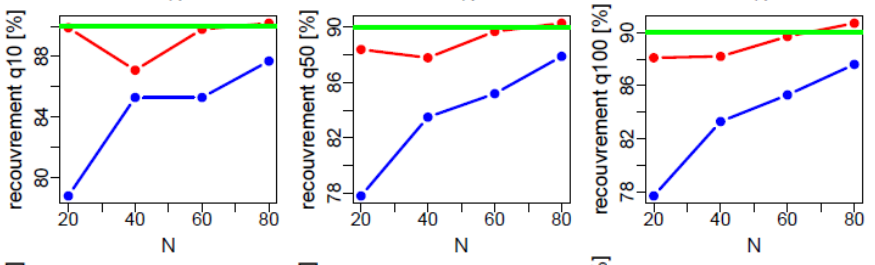
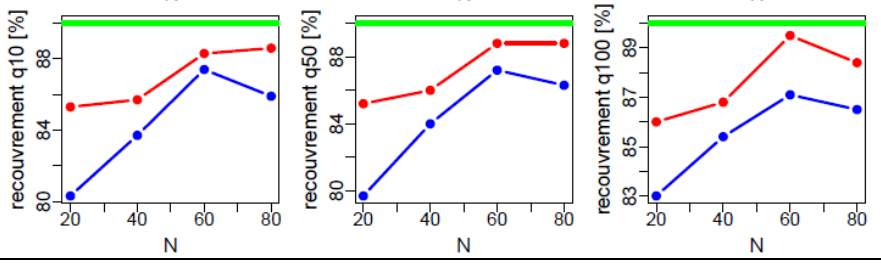
Distribution	Résultats
Normal	
LogNormal	
Exponential2	
GPD3	
Gumbel	
GEV	

Tableau 3 (suite)

Distribution	Résultats
PearsonIII	
LogPearsonIII	
Gumbel_min	
GEV_min	

III.3.2 Recommandation

Les résultats très clairs de cette étude conduisent à une recommandation simple : la méthode PBOOT devrait être associée à la méthode d'estimation LMOM comme méthode de quantification des incertitudes par défaut, quelle que soit la distribution. De plus, étant donné que la méthode BOOT semble systématiquement moins performante que PBOOT, on pourra envisager de ne même pas la proposer dans l'HydroPortail.

III.4 Applicabilité à de petits échantillons

La dernière question abordée dans cette étude par simulation concerne l'opportunité d'imposer un nombre minimum de données avant d'estimer une distribution. Il existe en fait une limite mathématique à ce nombre : pour estimer une distribution ayant *p* paramètres, il faut au minimum *p* valeurs pour que les calculs soient faisables (par exemple, on ne peut pas calculer un écart-type avec une seule donnée !). Etant donné que certaines des distributions qui seront disponibles dans l'HydroPortail possèdent 3 paramètres, il semble déjà logique d'imposer un minimum absolu de 3 valeurs avant de tenter toute estimation.

Afin de déterminer s'il serait opportun d'imposer une limite plus drastique, au moins pour certaines distributions, la procédure de simulation de l'Algorithme 1 est appliquée à des échantillons de taille faible variant entre $N = 3$ et $N = 15$. Le couple LMOM+PBOOT est utilisé pour effectuer les estimations et quantifier les incertitudes.

III.4.1 Résultats

Le premier résultat important est que le taux d'échec est nul pour toutes les tailles d'échantillon, sauf pour les distributions PearsonIII et LogPearsonIII où l'on observe des taux d'échec faibles compris entre 0 et 0.7% (non illustrés ici). Autrement dit, même pour une taille d'échantillon égale à 3, le calcul est presque toujours faisable avec la combinaison LMOM+PBOOT, et le code de calcul renverra donc des résultats. Evidemment, ceci n'implique pas que les résultats sont suffisamment bons pour être utilisables : avec seulement 3 valeurs, il faut s'attendre à des erreurs d'estimation potentiellement très fortes et des incertitudes gigantesques.

Afin d'évaluer les résultats obtenus avec de petits échantillons de manière plus précise, le Tableau 4 compile les indices de performance obtenus pour chaque distribution. La première partie du tableau concerne les distributions à deux paramètres. Dans la plupart des cas, les biais sont faibles pour toutes les tailles d'échantillon, sauf dans le cas des distributions LogNormal et Gumbel_min, où un biais plus important est observé quand $N = 3$. L'erreur-type augmente évidemment lorsque la taille de l'échantillon diminue, et peut dépasser 100% pour les deux distributions précédentes. Enfin, les incertitudes sont globalement assez bien quantifiées puisque les taux de recouvrement sont toujours supérieurs à 80%, et atteignent une valeur proche du 90% cible assez rapidement.

La seconde partie du tableau concerne les distributions à trois paramètres. Comme précédemment, les biais sont globalement assez faibles dès que *N* dépasse 5, à l'exception de la distribution GEV_min. Ceci est néanmoins en partie un artefact lié à l'expression du biais en pourcentage, alors que les quantiles sont assez proches de zéro (puisqu'il s'agit d'une distribution pour les minima). L'erreur-type augmente également lorsque la taille de l'échantillon diminue. Néanmoins, contrairement à ce qui était observé pour les distributions à deux paramètres, les incertitudes sont assez fortement sous-estimées pour les tailles d'échantillon faibles. Par exemple, pour la distribution GEV, le taux de recouvrement est de seulement 50% quand $N = 3$, mais atteint rapidement 70% quand $N = 5$ et se rapproche des 80% quand $N = 7$ avant de croître plus lentement vers le taux cible de 90% pour les tailles d'échantillon supérieures.

Tableau 4. Indices de performance relatifs aux quantiles décennal (rouge), cinquantennal (bleu) et centennal (noir) pour de petits échantillons.

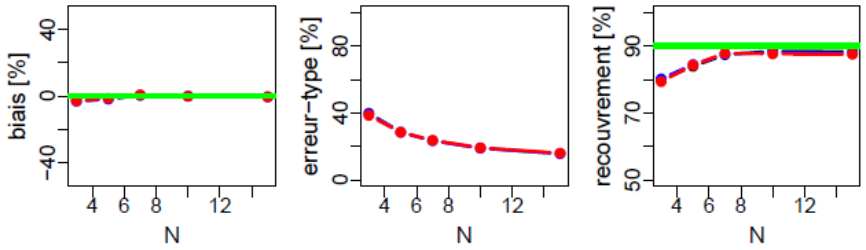
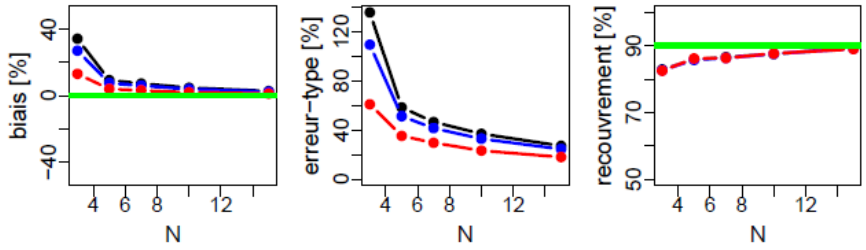
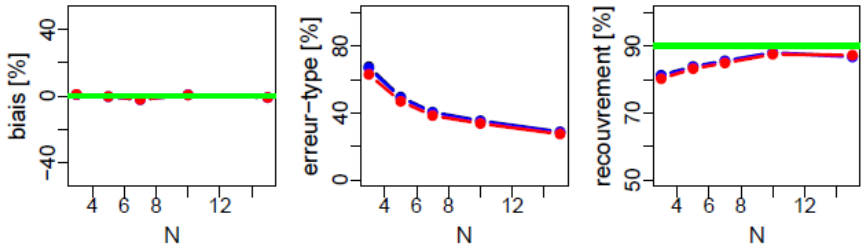
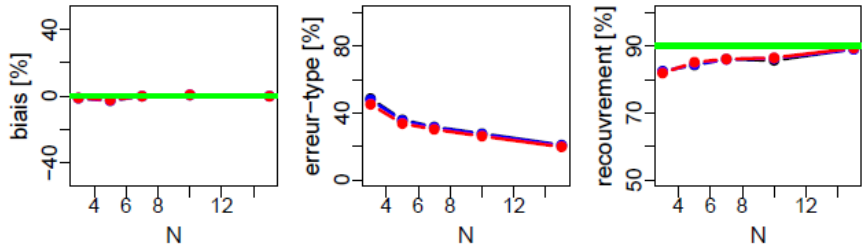
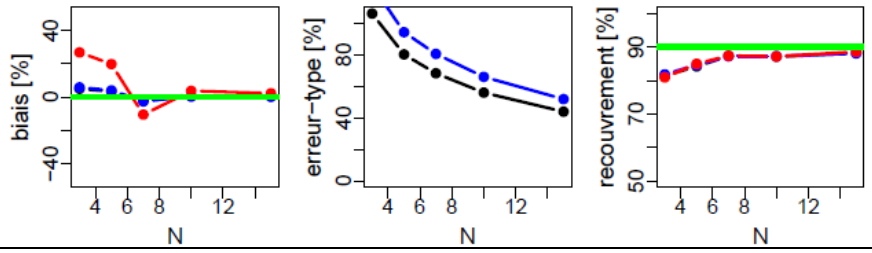
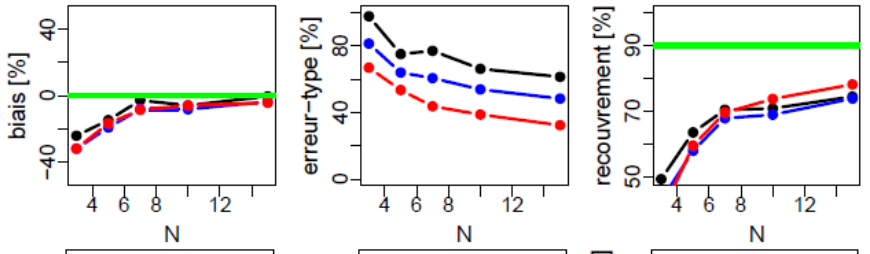
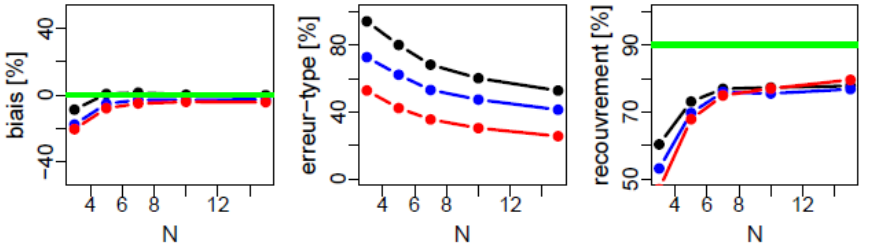
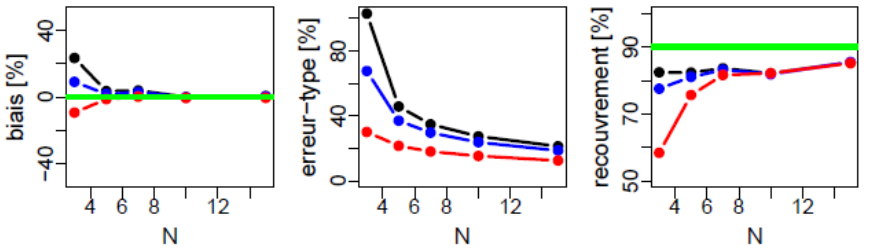
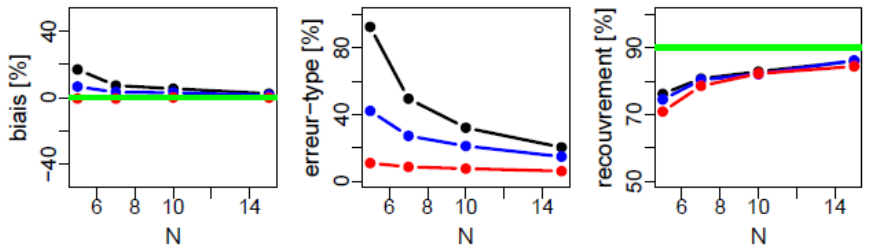
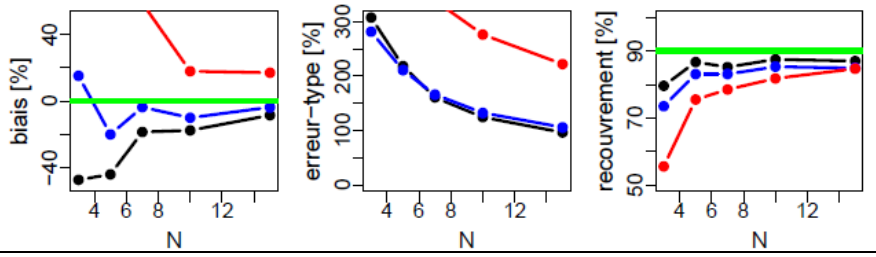
Distribution	Résultats
Normal	
LogNormal	
Exponential2	
Gumbel	
Gumbel_min	

Tableau 4 (suite)

Distribution	Résultats
GPD3	
GEV	
PearsonIII	
LogPearsonIII	
GEV_min	

III.4.2 Recommandation

Au vu des résultats précédents, deux différentes options peuvent être envisagées :

1. Etant donné que le calcul est presque toujours faisable dès $N = 3$, il est possible de fixer la limite basse à 3 observations (inclus). L'utilisateur recevra en retour des résultats assortis d'incertitudes gigantesques, et il sera de sa responsabilité d'interpréter cela comme un avertissement sur le manque de fiabilité de l'estimation.
2. On peut également considérer que le raisonnement précédent ne vaut que si les incertitudes sont correctement quantifiées, et les estimations non systématiquement biaisées. Or, des biais non négligeables existent pour les très petits échantillons, et de plus les lois à trois paramètres sous-estiment assez nettement les incertitudes. Il est donc également envisageable d'imposer des limites légèrement plus drastiques : $N = 5$ (inclus) pour les lois à deux paramètres et $N = 7$ (inclus) pour les lois à trois paramètres.

IV. Evaluation en utilisant des données réelles

IV.1 Procédure d'évaluation

L'objectif de cette évaluation est de mettre les codes à l'épreuve des données réelles de la banque HYDRO. En effet, l'utilisation de données simulées est utile pour évaluer les méthodes implémentées, mais constitue un cadre idéalisé où la distribution à estimer est la vraie distribution parente. De plus, les données réelles peuvent être affectées par plusieurs problèmes qui n'apparaissent pas avec des données simulées (valeurs erronées, erreurs de conversion d'unité, etc.). L'application des codes de calcul à un grand nombre de stations de la banque HYDRO devrait donc constituer un « crash-test » réaliste se rapprochant des conditions opérationnelles dans lesquelles les codes sont appelés à être utilisés.

Quelques restrictions ont tout de même été imposées pour sélectionner les stations à utiliser : les stations virtuelles ou qualifiées de « bidon » ont été exclues, de même que quelques stations possédant une donnée pour un 29 février d'une année non-bissextile (ex : P1350010 ou P2380010 en février 1900). De plus, nous avons imposé un minimum de 4 années (validées ou non) et une localisation X/Y connue. Enfin, nous avons imposé que la différence entre les bassins versants hydrologique et topographique soit inférieure (en valeur absolue) à 1%. Cette dernière contrainte est due au fait que nous avons transformé tous les débits en lames d'eau (exprimées en mm) afin de pouvoir comparer les estimations entre bassins, et que nous souhaitions éviter des erreurs trop importantes liées à un bassin versant hydrologique trop mal connu. Au final, 3139 stations ont ainsi été sélectionnées, et sont représentées en Figure 14. On peut observer quelques erreurs de localisation que nous n'avons pas cherché à corriger : l'objectif de cette étude n'est pas de fournir une cartographie précise de l'hydrologie française, mais seulement de mettre les codes de calcul à l'épreuve. La Figure 15 résume également le jeu de données analysé, et montre qu'il couvre toute la gamme de taille de bassin, d'altitude et d'ancienneté que l'on peut trouver en France métropolitaine.

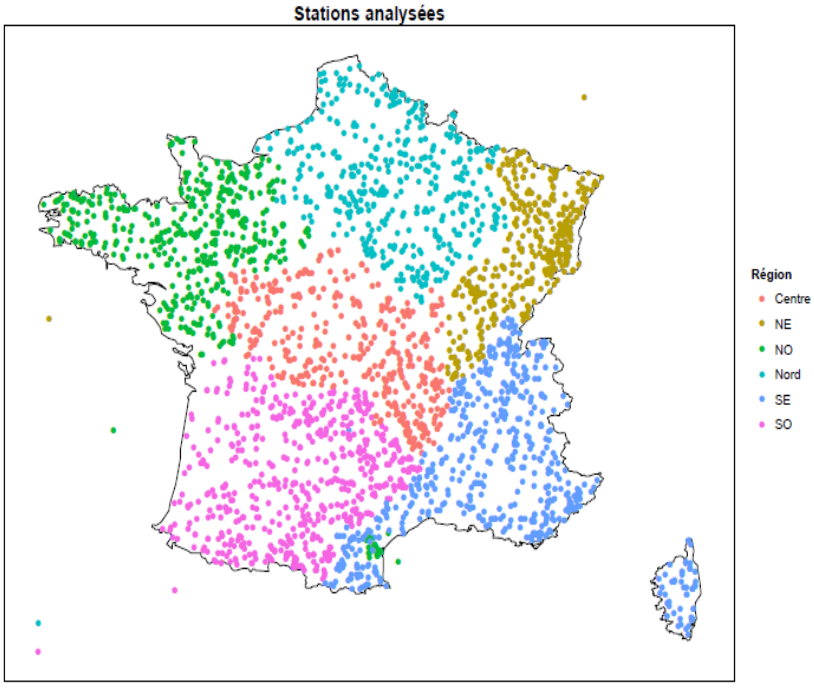


Figure 14. Carte des stations analysées et affectation à une région (basée sur la première lettre du code HYDRO).

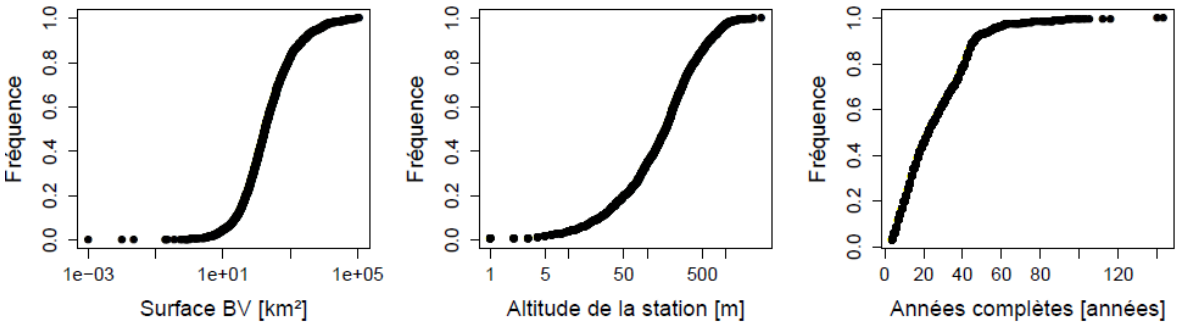


Figure 15. Propriétés des stations étudiées.

Pour chaque station analysée, trois variables hydrologiques ont ensuite été calculées à partir des séries de débits journaliers : le débit annuel (QA), le débit moyen sur 7 jours minimum annuel (VCN7) et le débit maximum annuel (MAXAN). Les années s'entendent ici en années hydrologiques, commençant pour chaque station au premier jour du mois le plus sec pour les variables QA et MAXAN, et au premier jour du mois le plus humide pour la variable VCN7. Seules les années hydrologiques complètes (sans données manquantes) ont été conservées.

Enfin, pour chaque station et chaque variable hydrologique, toutes les distributions disponibles (Tableau 1) ont été estimées en utilisant la combinaison recommandée par défaut LMOM+PBOOT. Evidemment, certaines distributions sont totalement inappropriées pour certaines variables : par exemple, les distributions Exponential2 et GPD3 sont spécifiques aux variables obtenues par échantillonnage SUP-SEUIL, et n'ont donc pas lieu d'être appliquées pour les trois variables étudiées ; de même, utiliser la loi Normale sur la variable MAXAN est très fortement déconseillé. Néanmoins, ces distributions inappropriées sont utilisées sciemment car elles permettent justement de vérifier la fiabilité des codes de calcul, et on peut de plus s'attendre à ce que les futurs utilisateurs testent toutes ces distributions disponibles. Encore une fois, l'objectif de cette étude n'est pas de dresser un tableau de l'hydrologie française, mais de mettre les codes de calcul à l'épreuve.

IV.2 Résultats

IV.2.1 Correction de bugs

Le premier résultat tangible de cette évaluation est qu'elle a permis de corriger un certain nombre de bugs dans le code, qui n'étaient pas apparus avec des données simulées. La plupart de ces bugs étaient assez mineurs et concernaient la capture d'erreurs de calcul qui n'avaient pas été anticipées. Par exemple, sur certaines stations, la variable VCN7 est constituée d'une unique valeur répétée plusieurs fois (qui correspond probablement à une limite de quantification ou à la résolution du limnimètre). Ceci conduit à une variance nulle, qui rend certains calculs impossibles. Un de ces bugs était plus problématique, puisqu'il concernait une formule incorrectement implémentée pour les distributions PearsonIII et LogPearsonIII.

Evidemment, l'intégralité des résultats présentés dans ce rapport (y compris ceux des sections précédentes) ont été obtenus en utilisant les codes corrigés de ces bugs.

IV.2.2 Ajustements

Les taux d'échec très faibles rapportés dans le Tableau 5 confirment que l'estimation est presque toujours possible avec la combinaison LMOM+PBOOT. Les valeurs de 0.03% apparaissant à de nombreuses reprises dans le tableau correspondent en fait à une unique station (N4015210, l'Hermitain à Romans) qui possède environ dix ans de débits journaliers nuls ! Il est évidemment impossible d'ajuster une distribution sur de telles données sans aucune variabilité. Les taux d'échecs sont plus élevés pour la variable VCN7 (autour de 2%) : pour la plupart, ces échecs correspondent au cas évoqué précédemment d'une unique valeur

répétée plusieurs fois. Encore une fois, ces échecs sont normaux puisqu'on ne peut pas ajuster une distribution sur des données ne possédant aucune variabilité.

Enfin, on remarque que les taux d'échec sont légèrement plus élevés pour les distributions PearsonIII et LogPearsonIII par rapport aux autres distributions (tout en restant faibles), ce qui avait déjà été observé avec des données simulées (voir sections III.3 et III.4).

Tableau 5. Taux d'échec en %.

Variable / Distribution	MAXAN	QA	VCN7
Normale	0.03	0.03	2.09
LogNormale	0.03	0.03	2.09
Gumbel	0.03	0.03	2.09
Exponentielle	0.03	0.03	2.09
GEV	0.03	0.03	2.09
GPD	0.07	0.03	2.16
Gumbel_min	0.03	0.03	2.09
GEV_min	0.03	0.03	2.09
Pearson III	0.31	0.28	2.44
Log-Pearson III	0.49	0.80	2.72

La Figure 16 cartographie le quantile décennal calculé pour chaque variable. Une unique distribution de référence est représentée pour chaque variable, cette distribution ayant été sélectionnée en considérant qu'elle correspondait à la pratique actuelle la plus répandue parmi les utilisateurs d'HYDRO2. Pour une variable donnée, les cartes sont en fait très similaires avec d'autres distributions car à cette échelle, les différences régionales (qui font varier les quantiles sur plusieurs ordres de grandeur) sont beaucoup plus importantes que les différences induites par différents choix de distribution.

Globalement, les débits moyens annuels décennaux varient entre 0.1 mm et 10 mm. Les valeurs les plus fortes sont observés dans les massifs montagneux, les valeurs les plus faibles dans le bassin Parisien, ce qui correspond bien aux grandes lignes de l'hydrologie française. La carte des débits maximum annuels décennaux est assez similaire, si ce n'est que l'arc cévenol ressort plus nettement, avec des valeurs qui peuvent dépasser les 100 mm. Enfin, la carte des VCN7 décennaux présente moins de cohérence spatiale. On peut tout de même remarquer des valeurs plus élevées dans les Alpes et les Pyrénées (potentiellement dues à l'effet régularisateur de la fonte des neiges) et des valeurs plus faibles dans la région des Pays de la Loire.

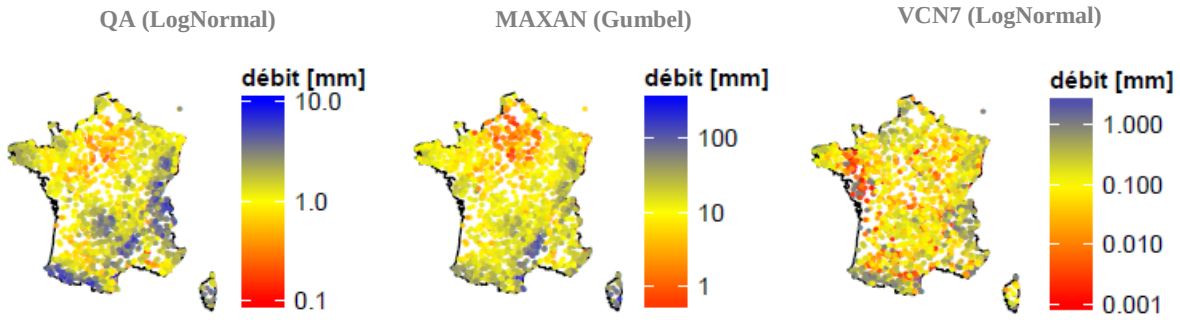


Figure 16. Carte des quantiles décennaux pour les trois variables étudiées (seules les stations disposant d'au moins 20 années complètes sont représentées). Les points gris pour la variable VCN7 correspondent à des débits nuls.

La Figure 17 et la Figure 18 montrent qu'en général, les quantiles les plus forts pour les variables QA et MAXAN sont associés à des petits bassins versants et/ou à des stations situées en altitude. On remarque également que la majorité de ces forts quantiles sont localisés dans les régions Sud-Ouest et Sud-Est. Pour la variable VCN7, la relation entre le quantile décennal et l'altitude ou la taille du bassin est beaucoup moins nette.

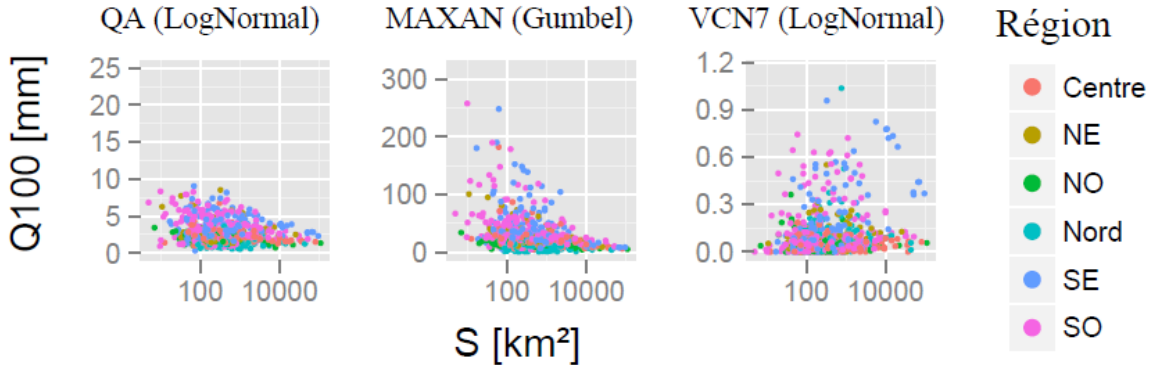


Figure 17. Estimation du quantile centennal en fonction de la taille du bassin versant (seules les stations disposant d'au moins 40 années complètes sont représentées).

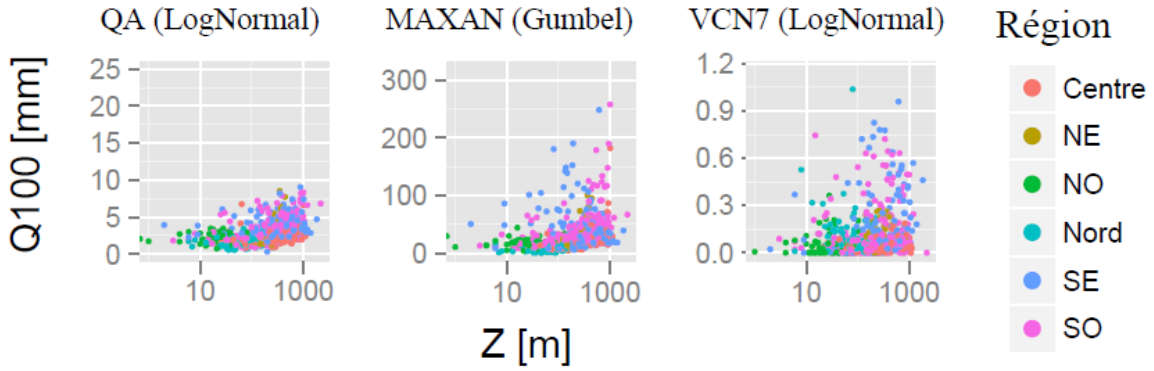


Figure 18. Estimation du quantile centennal en fonction de l'altitude de la station (seules les stations disposant d'au moins 40 années complètes sont représentées).

La Figure 19 décrit la dépendance de l'incertitude affectant le débit centennal (exprimée en %) à la taille de l'échantillon disponible pour l'estimation. Le résultat est illustré pour la variable MAXAN mais est qualitativement similaire pour les autres variables. Comme attendu, l'incertitude peut prendre des valeurs plus importantes pour les petits échantillons, quelle que soit la distribution. On remarque cependant que l'incertitude atteint des valeurs beaucoup plus importantes pour les distributions à trois paramètres que pour celles à deux paramètres, et que cette différence est plus forte pour les petits échantillons. La loi log-

normale constitue une exception à cette observation, puisqu'elle semble se comporter comme une loi à trois paramètres (alors qu'elle n'en possède que deux), atteignant parfois des incertitudes extrêmement fortes (dépassant les 200%) pour quelques stations.

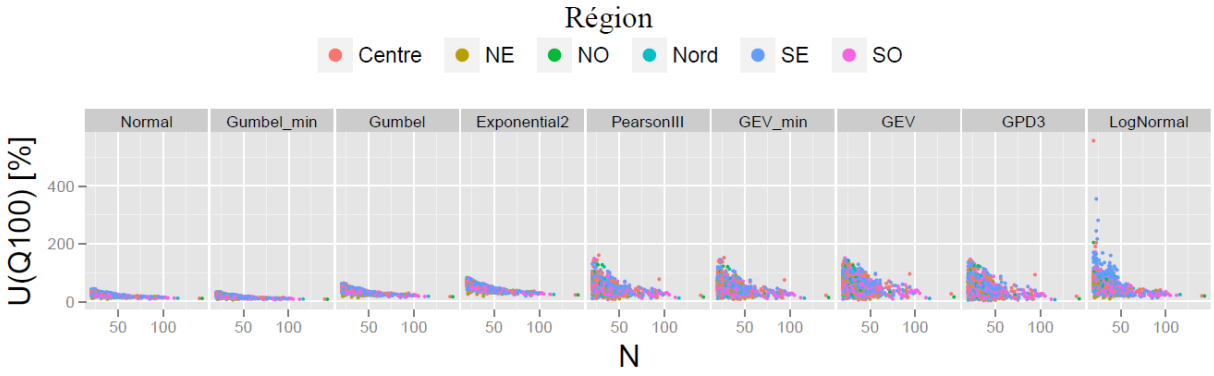


Figure 19. Incertitude d'estimation du quantile centennal en fonction du nombre de données disponibles pour la variable MAXAN (seules les stations disposant d'au moins 20 années complètes sont représentées).

IV.2.3 Tests statistiques

Les codes de calcul incluent la possibilité d'appliquer quelques tests statistiques à l'échantillon. La Figure 20 montre ainsi la carte des résultats du test de détection de tendance de Mann-Kendall. La carte pour le test de détection de rupture de Pettitt est très similaire, et n'est donc pas illustrée ici. Pour la variable MAXAN, environ 18% des stations présente une tendance (au risque 10%). Cependant, ces tendances ne montrent aucune cohérence spatiale, ce qui suggère une origine « locale » (par exemple une non-homogénéité d'origine métrologique) plutôt que « globale » (par exemple le climat). Les résultats sont similaires pour la variable QA (avec environ 20% de tendances significatives). Les tendances détectées sont plus nombreuses pour la variable VCN7 (28%), et parmi les changements significatifs, les baisses sont plus fréquentes que les hausses (62% contre 38%). Ces résultats sont cohérents avec des études de la littérature[2]. Il faut néanmoins rester vigilant sur l'interprétation du résultat de ce test : en effet, il suppose les données indépendantes, ce qui n'est pas forcément le cas, surtout pour les variables de basses eaux qui peuvent présenter une autocorrélation importante dans certains cas (inertie liée aux nappes d'accompagnement). Il est donc possible que certaines des tendances détectées ne soient qu'un artéfact induit par l'autocorrélation.

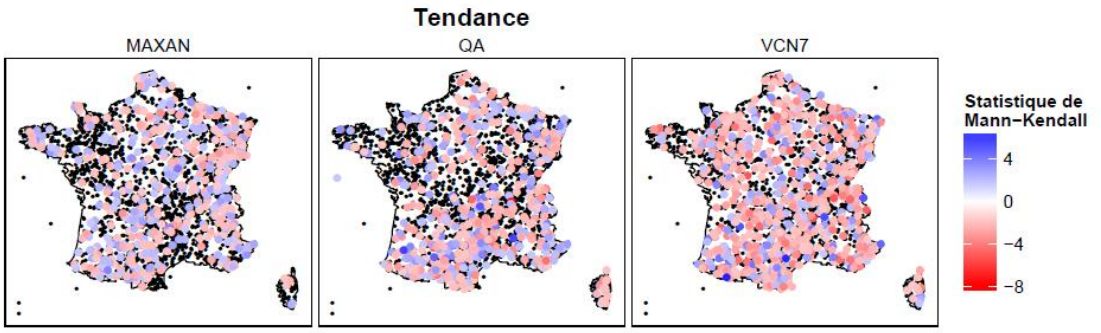


Figure 20. Test de détection de tendance : statistique de Mann-Kendall pour les trois variables étudiés. Les points noirs correspondent aux changements non-significatifs (pour lesquels la statistique de Mann-Kendall est inférieure à 1.64 en valeur absolue).

[2] Par exemple, Giuntoli, I., B. Renard, J. P. Vidal, and A. Bard (**2013**), Low flows in France and their relationship to large scale climate indices, *J. Hydrol.*, 482, 105-118.

Le Tableau 6 montre les pourcentages de rejet pour le test d'adéquation de Kolmogorov-Smirnov. La distribution de Gumbel pour les minima est très souvent rejetée, y compris pour la variable VCN7 qui est pourtant définie comme un minimum. Ceci peut s'expliquer par le fait que cette distribution n'est pas bornée, et tend à prendre des valeurs négatives avec une probabilité non négligeable, ce qui est évidemment problématique pour des débits. La loi exponentielle est également fréquemment rejetée, ce qui n'est pas étonnant en soi étant donné que cette distribution est spécifique aux variables de type SUP-SEUIL, ce qui n'est le cas d'aucune des trois variables analysées. La distribution normale est rejetée assez fréquemment pour les variables MAXAN et VCN7, mais beaucoup moins pour la variable QA, ce qui suggère qu'elle pourrait être un candidat possible pour cette variable sur certaines stations.

Pour toutes les autres distributions, les pourcentages de rejet sont bien inférieurs au risque α = 10%. En particulier, ces pourcentages sont très faibles pour les distributions à trois paramètres. Ceci illustre la limite de ce test d'adéquation, qui est dans ce cas très peu puissant, c'est-à-dire qui ne rejettera que très rarement des distributions possédant de nombreux paramètres. Il sera donc important d'insister, dans l'interface de l'HydroPortail ainsi que dans l'aide et les formations, sur les limites de ce test, afin d'éviter que les utilisateurs ne lui accordent trop d'importance. Des recommandations en ce sens sont proposées dans la section V.

Tableau 6. Pourcentage de rejet du test de Kolmogorov-Smirnov appliqué au risque α = 10% (seules les stations disposant d'au moins 40 années complètes sont prises en compte).

Variable / Distribution	MAXAN	QA	VCN7
Gumbel_min	56.20	35.18	58.67
Exponentielle	20.44	23.65	18.37
Normale	15.62	2.34	15.56
LogNormale	2.04	1.17	4.74
Gumbel	3.80	1.90	3.85
GEV	0.00	0.29	1.33
GPD	0.88	0.44	1.63
GEV_min	0.58	0.58	1.78
Pearson III	0.44	0.58	2.08
Log-Pearson III	0.00	0.15	0.74

V. Conclusions et recommandations

Dans le cadre de l'opération HYDRO 3, une modernisation des calculs permettant d'estimer des débits caractéristiques (et leur incertitude) a été initiée. Irstea a fourni un code de calcul à cet effet, intégralement implémenté dans le langage R. Le premier objectif de ce rapport était de documenter les méthodes statistiques qui ont été implémentées : ceci est fait dans l'annexe, qui recense toutes les formules et algorithmes qui constituent le code de calcul. Le second objectif était principalement de mettre ce code à l'épreuve, en le testant à la fois sur des données simulées et sur plus de 3000 stations de la banque HYDRO. Cette mise à l'épreuve s'est avérée très utile, puisqu'elle a permis de corriger quelques bugs et d'améliorer la gestion des erreurs. De plus, les résultats obtenus au cours de ces tests permettent de formuler un certain nombre de recommandations sur l'implémentation du menu « Statistiques » de l'HydroPortail, qui pilotera le code de calcul et constituera l'interface effectivement manipulée par les utilisateurs. Ces recommandations sont les suivantes :

- Parmi les méthodes d'estimation des paramètres et de quantification des incertitudes disponibles, le couple LMOM / PBOOT devrait être le choix par défaut.
- La méthode de quantification des incertitudes BOOT peut éventuellement être ignorée dans l'interface : elle n'apporte aucune plus-value par rapport à PBOOT.
- Pour les distributions PearsonIII et LogPearsonIII, la méthode d'estimation ML ne devrait pas être proposée, car elle ne converge pas dans de trop nombreux cas.
- Un nombre minimum de points doit être disponible pour pouvoir estimer une distribution. Un minimum de 3 valeurs est suffisant pour assurer que le code renverra un résultat dans la quasi-totalité des cas. Il est néanmoins possible d'imposer des limites légèrement plus strictes pour minimiser le biais des estimations et éviter une sous-estimation des incertitudes : 5 valeurs minimum pour les distributions à deux paramètres, et 7 valeurs minimum pour les distributions à trois paramètres.

En complément, les recommandations suivantes pourront être considérées. Elles ne découlent pas directement des travaux décrits dans ce rapport et concernent surtout l'ergonomie de l'interface proposée dans l'HydroPortail :

- Le choix de la distribution à utiliser dépend fortement du type de variable. A minima, il convient d'organiser les distributions proposées à l'utilisateur d'une manière qui l'incitera à utiliser préférentiellement les distributions recommandées. Une version plus stricte consisterait à ne pas proposer les distributions non recommandées pour certains types de variable. Le Tableau 7 propose une organisation possible.
- Pour améliorer l'ergonomie, le comportement par défaut de l'interface pourrait être de ne pas proposer de choix dans les méthodes d'estimation et de quantification des incertitudes. En effet, l'option par défaut par défaut LMOM / PBOOT devrait convenir à la majorité des utilisateurs. Un bouton « pour les experts » pourrait ouvrir une fenêtre ou un onglet permettant de modifier ce choix par défaut.
- Il est important qu'un message d'avertissement accompagne les résultats des tests statistiques (sous forme d'une info-bulle ou d'un point d'interrogation cliquable). Voici une proposition de textes :
    - Test de Mann-Kendall : « *Attention : ce test fait l'hypothèse que les données sont indépendantes. Si les données présentent une autocorrélation non négligeable, le test risque de détecter fréquemment une tendance qui n'existe pas.* »
    - Test de Pettitt : comme ci-dessus, en remplaçant « *tendance* » par « *rupture* ».
    - Test de Kolmogorov-Smirnov : « *Attention : ce test fait l'hypothèse que les données sont indépendantes. De plus, il est censé n'être applicable que pour*

comparer les données à une distribution CONNUE, et non ESTIMEE comme ici. L'interprétation des résultats du test devrait donc être la suivante : si une distribution est rejetée, alors elle est sûrement inadéquate ; par contre une distribution non-rejetée ne doit en aucun cas être interprétée comme une preuve de sa fiabilité ! »

- La loi de Poisson étant seulement définie pour des valeurs entières, elle ne devrait être proposée que pour les variables de type « durée » exprimées par exemple en nombre de jours.
- Le code de calcul propose une option `splitZeros` pour le traitement des valeurs nulles (voire négatives, ce qui pourrait arriver si le code est appliqué à des données de hauteur). Lorsque cette option est activée (`splitZeros=TRUE`), la distribution n'est estimée qu'en utilisant les valeurs strictement positives. En parallèle, la fréquence des valeurs inférieures ou égales à zéro est calculée et est utilisée dans le calcul des périodes de retour. En conséquence, dans le cas où l'échantillon comprend des valeurs inférieures ou égales à zéro, l'interface devrait proposer une case à cocher pour activer ou désactiver cette option (par défaut, il est recommandé de l'activer), dont l'intitulé pourrait être « *traiter les valeurs nulles et négatives à part* ». Si l'option est désactivée, alors les distributions `LogNormal` et `LogPearsonIII` devraient être exclues (car elles ne sont définies que pour des valeurs strictement positives).
- Le code de calcul propose également une option `invertT` qui définit la relation entre la période de retour T et la probabilité au non-dépassement p. Si `invertT=FALSE` alors les grandes périodes de retour correspondent aux valeurs fortes, et la relation est $T=1/(1-p)$. Ainsi un événement centennal a une probabilité de dépassement 1/100 (et donc une probabilité de non-dépassement de p = 99/100). Si `invertT=TRUE` alors les grandes périodes de retour correspondent aux valeurs faibles, et la relation devient $T=1/p$. C'est le cas par exemple pour la variable « minimum annuel » : un étiage dit centennal correspondra à de faibles valeurs du minimum annuel, et donc une probabilité de non-dépassement de p = 1/100. L'interface devrait donc proposer une case d'option pour activer ou désactiver cette option. L'intitulé pourrait être « *les grandes périodes de retour correspondent aux valeurs : fortes / faibles* ». Le choix par défaut dépend de la variable étudiée, et une proposition est faite dans le Tableau 8.
- Enfin, le code de calcul propose une option `p2T` qui représente le nombre moyen de valeurs échantillonnées par an (τ) et qui sert à pondérer la relation entre la période de retour et la probabilité au non-dépassement (la formule est $T = 1/[\tau(1-p)]$ si `invertT=FALSE`, et $T = 1/\tau p$ sinon). Cette option n'est utile que pour les variables de type SUP-SEUIL (qui peuvent conduire à 2 ou 3 événements par an en moyenne, par exemple), et devrait donc être interfacée uniquement pour ce type de variable. Pour tous les autres types de variables, l'échantillonnage est annuel et la valeur par défaut `p2T=1` peut donc être utilisée sans intervention de l'utilisateur.

Tableau 7. Recommandation sur le choix de la distribution à utiliser en fonction du type de variable. Une distribution est recommandée (☺) s'il existe une raison théorique pour la sélectionner (par exemple, théorème des valeurs extrêmes) ou si la pratique a montré qu'elle était souvent adéquate. Une distribution est déconseillée (☹) quand il existe des raisons bien identifiées pour ne pas l'utiliser (par exemple Gumbel et GEV sont spécifiques aux maxima ; Exponential2 et GPD3 sont spécifiques aux variables définies par rapport à un seuil). Les autres distributions sont possibles (😐).

Distribution / Type de variable	Normal	LogNormal	Gumbel	GEV	Gumbel_min	GEV_min	Exponential2	GPD3	PearsonIII	LogPearsonIII	Poisson
Moyenne	☺	☺	☹	☹	☹	☹	☹	☹	😐	😐	☹
Maximum	☹	☹	☺	☺	☹	☹	☹	☹	😐	😐	☹
Minimum	😐	☺	☹	☹	☺	☺	☹	☹	😐	😐	☹
SUP-SEUIL	☹	☹	☹	☹	☹	☹	☺	☺	☹	☹	☹
Durée sous un seuil	😐	😐	☹	☹	☹	☹	😐	😐	😐	😐	😐(*)
Valeur de fréquence *f*	😐	😐	😐	😐	😐	😐	☹	☹	😐	😐	☹
Centre de masse	😐	😐	☹	☹	☹	☹	☹	☹	😐	😐	☹

(*) Seulement si la durée est exprimée comme un nombre entier (nombre de jours par exemple)

Tableau 8. Valeurs par défaut recommandées pour l'option `invertT` permettant de déterminer si les grandes périodes de retour correspondent aux valeurs fortes ou faibles.

Type de variable	Les grandes périodes de retour correspondent aux valeurs…	Valeur par défaut proposée pour `invertT`
Moyenne	fortes pour un quantile « humide » faibles pour un quantile « sec »	`FALSE`
Maximum	fortes	`FALSE`
Minimum	faibles	`TRUE`
SUP-SEUIL	fortes	`FALSE`
Durée sous un seuil	fortes	`FALSE`
Valeur de fréquence *f*	fortes si $f > 0.5$, faibles sinon	`FALSE` si $f > 0.5$, `TRUE` sinon
Centre de masse	fortes pour un quantile « tardif » faibles pour un quantile « précoce »	`FALSE`

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