Projet ANR-08-RISK-03-01

Prédétermination des valeurs extrêmes de pluies et de crues » (EXTRAFLO)

Programme RISKNAT 2008

Tâche II : Mise au point d'une stratégie commune d'inter-comparaison et de validation

Rapport II.1 « *Méthodologie de comparaison d'approches probabilistes d'estimation des valeurs extrêmes* »

Date : Mars 2010

Rapport réalisé par :

(1) Irstea, Centre de Lyon HHLY

Avec la participation de :

(2) Météo-France, Direction de la Climatologie

(3) HydroSciences Montpellier

(4) EDF/DTG

(5) Irstea, Centre d'Aix-en-Provence, OHAX

Auteurs :

B. Renard [1], M. Lang [1], J.M. Soubeyroux [2], L. Neppel [3], F. Garavaglia [4], P. Arnaud [5]

Sommaire

1 Introduction

1.1 Prédétermination des pluies ou des débits : un foisonnement d'implémentations

La prédétermination des pluies et des débits est une étape essentielle dans l'analyse et la gestion des risques hydrologiques. De très nombreuses méthodes de prédétermination ont été développées, classables dans plusieurs grandes familles d'approches :

- Les approches statistiques locales : choix d'une distribution particulière et estimation de ses paramètres en utilisant exclusivement les données du site cible ;
- Les approches utilisant l' information climatique : c'est par exemple le cas de la méthode SCHADEX-pluies, qui effectue ses estimations conditionnellement au type de temps [*Paquet et al.*, 2006; *Garavaglia et al.*, 2010] ;
- Les approches utilisant l'information historique [*Naulet*, 2002; *Naulet et al.*, 2005; *Payrastre*, 2005; *Neppel et al.*, 2010; *Payrastre et al.*, 2011] ;
- Les approches statistiques régionales, qui visent à combiner les données du site cible (si présentes) et les données issues de bassins versants similaires à celui du site cible [*Neppel et al.*, 2007; *Cipriani et al.*, 2012] ;
- Les approches par simulation, qui utilisent un générateur de pluie et/ou un modèle pluie-débit : par exemple la méthode SHYPRE [*Arnaud and Lavabre*, 1999, 2002] ou la méthode SCHADEX-débits [*Paquet et al.*, 2006].

A cette diversité d'approches se superpose un foisonnement de variantes à l'intérieur de chaque famille, différant par exemple dans le choix d'une distribution, d'une méthode d'estimation, d'un modèle pluie-débit, etc. Pour éviter toute ambiguïté, nous utiliserons systématiquement le terme « implémentation » pour désigner une variante particulière.

1.2 Différentes approches de comparaison

L'objectif principal du projet ExtraFlo est de comparer un ensemble d'implémentations appartenant aux grandes familles décrites ci-dessus. La méthodologie de comparaison mise en place doit pouvoir s'adapter à la grande variété des implémentations candidates. Parmi les approches de comparaison qui sont classiquement utilisées, on peut citer :

- Les approches par simulations Monte-Carlo : le principe est de générer des données dont on contrôle les caractéristiques (distribution, taille, etc.). Ceci permet de comparer diverses implémentations sur la base de critères objectifs puisque la « vraie » distribution des observations est connue. Malheureusement, ce type d'approche se prête assez mal au contexte de ce projet, caractérisé par une grande diversité d'implémentations : il est en effet difficile de mettre au point une simulation Monte-Carlo qui permette de comparer de manière équitable des approches aussi différentes que (par exemple) SHYPRE ou l'estimation d'une loi de Gumbel par la méthode des moments. Même si une telle simulation était mise en place, la portée des résultats dépendrait fortement du « réalisme » des données simulées, ce qui peut donner lieu à des discussions interminables.
- L'utilisation de tests statistiques : le principe est d'évaluer si la distribution supposée est cohérente avec les observations utilisées pour son estimation. Dans le cas contraire, on aura tendance à rejeter la distribution supposée. Cette approche est bien adaptée aux implémentations relativement simples comme les approches statistiques locales [*Laio*, 2004], mais est encore une fois problématique étant donnée la grande diversité des

implémentations que nous souhaitons comparer : pour la plupart de ces implémentations, des tests statistiques ne sont tout simplement pas disponibles.

Etant données ces difficultés, nous optons dans le cadre du projet ExtraFlo pour une troisième approche, basée sur la décomposition du jeu de données en sous-ensembles de calage et de validation. Les données issues du sous-ensemble de calage sont utilisées pour estimer les paramètres des implémentations candidates. Les estimations issues des implémentations sont ensuite comparées aux données de validation. De manière cruciale, la décomposition en calage/validation permet de comparer toutes les implémentations sur la base *des mêmes données de validation*, et ce en dépit de la diversité des implémentations candidates. De plus, une telle décomposition correspond étroitement au contexte opérationnel du dimensionnement, qui est une des applications principales de la prédétermination. En effet, lorsqu'on souhaite dimensionner un ouvrage, on estime la distribution des pluies/crues sur la base des données disponibles jusque là (correspondant aux données de calage). Mais l'ouvrage construit sur la base de ces estimations (potentiellement entachées d'erreurs) devra faire face à de nouveaux événements (correspondant aux données de validation).

1.3 Justesse et stabilité

Dans le contexte de la décomposition en calage-validation, la qualité des implémentations candidates sera jugée sur la base de deux critères : la justesse et la stabilité.

- La justesse est la capacité d'une implémentation à délivrer des estimations cohérentes avec les observations. Les observations peuvent provenir :
    - De la période de calage. Une implémentation qui ne serait pas juste en calage a peu de chance de le devenir en mode prédictif (validation). Néanmoins, une implémentation apparaissant juste en calage ne l'est pas forcément en validation (ex. : implémentation sur-paramétrées, reproduisant bien les observations de calage mais au pouvoir prédictif limité).
    - D'une période de validation. C'est sur ces données (et sur ces données seulement) que l'on peut juger le pouvoir prédictif d'une implémentation.
- La stabilité est la capacité d'une implémentation à fournir des estimations proches lorsque deux périodes de calage C1 et C2 sont utilisées.

Précisons que ces deux critères ne jouent pas le même rôle : en effet, une implémentation peut être parfaitement stable mais n'avoir aucune justesse. On tendra donc à évaluer d'abord la justesse des implémentations candidates, ce qui permettra dans un premier temps de rejeter celles n'atteignant pas une justesse acceptable. Dans un second temps, la stabilité pourra permettre de départager des implémentations (entre deux implémentations de justesse comparable, on préfèrera la plus stable).

1.4 Comparaison des incertitudes estimées

La majorité des implémentations considérées dans ce projet fournissent une estimation des incertitudes liées à l'estimation des paramètres. Néanmoins, il faut garder à l'esprit que ces incertitudes ne sont elles-mêmes que des estimations : en effet, la quantification des incertitudes nécessite d'émettre un certain nombre d'hypothèses (par exemple sur la distribution parente). Si ces hypothèses s'avèrent irréalistes, la quantification des incertitudes peut s'avérer inadéquate : en particulier, il est possible de sous-estimer l'incertitude.

Une des originalités de la comparaison que nous nous proposons de réaliser dans le projet ExtraFlo est de reconnaître cet état de fait, et d'inclure également une comparaison des incertitudes estimées. Cette comparaison pourra être mise en place grâce à l'utilisation de distributions prédictives (cf. section 5).

1.5 Notations et vocabulaire

Cette section introduit les notations qui seront utilisées tout au long de ce rapport. Soit $\boldsymbol{x} = \left(x_k^{(i)}\right)_{i=1:N_{site},k=1:n^{(i)}}$ le jeu de données qui sera utilisé pour effectuer la comparaison. L'exposant $^{(i)}$ désigne le site, tandis que l'indice $_k$ désigne le pas de temps. Typiquement, $x_k^{(i)}$ représentera le maximum annuel de pluie (ou de débit) au site *i* pour la k^{e} année (remarquons que le nombre d'années disponibles n'est pas forcément identique pour tous les sites). De manière similaire, la notation $\boldsymbol{c}$ désigne le sous-ensemble de $\boldsymbol{x}$ utilisé pour le calage des implémentations, et $\boldsymbol{v}$ le sous-ensemble (disjoint de $\boldsymbol{c}$) utilisé pour la validation. Lorsqu'aucune distinction n'est nécessaire, nous utiliserons la notation générique $\boldsymbol{d}$ pour désigner soit $\boldsymbol{c}$ soit $\boldsymbol{v}$.

La fonction de répartition de la « vraie » distribution (ou distribution « parente ») ayant généré les données $\boldsymbol{x}^{(i)}$ au site *i* est notée $F^{(i)}$. Cette distribution parente est évidemment inconnue, et l'objectif de la prédétermination est de l'estimer.

Chaque implémentation *M* fait une hypothèse quant à cette distribution : nous noterons $F_M^{(i)}(y \mid \boldsymbol{\theta})$ la distribution supposée par l'implémentation *M*. Dans cette notation, *y* est la valeur à laquelle la fonction de répartition est calculée et $\boldsymbol{\theta}$ représente un vecteur de paramètres inconnus qu'il faut estimer.

A l'issue de l'étape d'estimation des paramètres, on obtient un vecteur de paramètres estimés $\hat{\boldsymbol{\theta}}$, conduisant à une distribution estimée, dont la fonction de répartition est notée $\hat{F}_M^{(i)}(y) = F_M^{(i)}(y \mid \hat{\boldsymbol{\theta}})$.

Il est important de bien faire la différence entre les différentes fonctions de répartition utilisées ici :

1. La distribution parente ($F^{(i)}$), qui est inconnue.
2. La distribution supposée ($F_M^{(i)}$) par l'implémentation *M*, dont les paramètres sont inconnus et doivent être estimés.
3. La distribution estimée ($\hat{F}_M^{(i)}$) par l'implémentation *M*, qui correspond à la distribution supposée avec pour paramètres les paramètres estimés.

2 Justesse

Cette section présente les outils utilisés pour évaluer la justesse des implémentations candidates (cf. section 1.3). Ces outils sont construits sur la base d'indices de justesse calculés sur chaque site (section 2.1). La distribution de ces indices sur l'ensemble des sites composant le jeu de données est utilisée pour mettre en place des outils graphiques (section 2.2) et pour associer à chaque implémentation un score (section 2.3). L'interprétation de ces outils est discutée en section 2.4.

2.1 Indices

2.1.1 Pval

Le premier indice de justesse vise à évaluer l'adéquation globale entre la distribution estimée ($\hat{F}_M^{(i)}$) et les observations $d_k^{(i)}$ (l'indice peut être appliqué soit aux données de calage soit aux données de validation). Pour un site *i* et un pas de temps *k*, l'indice *pval* est défini comme suit:

$$pval_k^{(i)} = \hat{F}_M^{(i)}(d_k^{(i)}) \quad (1)$$

Si l'estimation est juste ($\hat{F}_M^{(i)} = F^{(i)} \ \forall i$), on peut montrer que les valeurs $pval_k^{(i)}$ sont des réalisations d'une loi uniforme sur chaque site *i*: $pval_k^{(i)} \sim U[0;1] \ \forall i$ [pour une preuve formelle, cf. *Renard et al.*, 2013]. En d'autres termes, une estimation juste correspond à des probabilités de non-dépassements uniformément réparties entre 0 et 1.

L'hypothèse de justesse ($\hat{F}_M^{(i)} = F^{(i)} \ \forall i$) est utilisée ici comme une hypothèse de travail, au même titre que l'hypothèse H0 d'un test statistique : l'objectif est d'évaluer si les données montrent des signes d'incompatibilité avec cette hypothèse (ce qui se matérialiserait ici par des valeurs de *pval* non uniformément distribuées entre 0 et 1). Dans l'affirmative, on aurait tendance à rejeter l'hypothèse de justesse. Dans la négative, on ne pourrait pas prouver que l'implémentation évaluée est juste, mais on pourrait au moins en conclure que les données n'apportent pas la preuve du contraire.

En pratique, la comparaison de différentes implémentations sera effectuée en comparant la distribution des valeurs *pval* obtenues sur l'ensemble des années et l'ensemble des sites (en calage ou en validation). Des outils graphiques seront utilisés à cet effet, et seront décrits en détail en section 2.2.

Le critère *pval* décrit l'adéquation entre la distribution estimée et l'ensemble des observations (en calage ou validation), sans focaliser particulièrement sur les extrêmes. Il est donc plutôt adapté à détecter des biais systématiques (e.g. la distribution estimée est systématiquement plus basse que la distribution empirique), mais n'est que peu informatif sur le comportement des méthodes en extrapolation.

2.1.2 N_T

Le second indice de justesse s'intéresse à la justesse pour une période de retour *T* donnée. Pour cela, l'indice N_T consiste simplement à dénombrer le nombre de dépassements du quantile estimé $\hat{q}_T^{(i)}$ [*Interagency Advisory Committee on Water Data*, 1982; *Gunasekara and Cunnane*, 1992; *Garavaglia et al.*, 2010]. Formellement :

$$N_T^{(i)} = \sum_{k=1}^{n^{(i)}} 1_{[\hat{q}_T^{(i)};+\infty)}(d_k^{(i)})$$
$$\text{Avec } 1_A(x) = \begin{cases} 1 \text{ si } x \in A \\ 0 \text{ sinon} \end{cases} \quad (2)$$

Si l'estimation est juste ($\hat{q}_T^{(i)} = q_T^{(i)}$), on peut montrer qu'en chaque site *i*, la valeur $N_T^{(i)}$ est une réalisation d'une loi binomiale avec probabilité de succès $1/T$ et nombre d'essais $n^{(i)}$ (avec $n^{(i)}$ égal au nombre de données $d_k^{(i)}$ au site *i*) : $N_T^{(i)} \sim Bin(n^{(i)}, 1/T)$ [pour une preuve formelle, cf. *Renard et al.*, 2013]. Comme pour l'indice *pval*, la comparaison entre différentes implémentations sera basée sur les outils graphiques décrits en section 2.2.

2.1.3 FF

Le dernier indice de justesse, *FF*, consiste simplement à calculer la probabilité de non-dépassement de la valeur maximale observée $d_{\max}^{(i)}$ (en calage ou en validation) [*England et al.*, 2003; *Garavaglia et al.*, 2011]:

$$FF^{(i)} = \hat{F}_M^{(i)}(d_{\max}^{(i)}) \quad (3)$$

Si l'estimation est juste ($\hat{F}_M^{(i)} = F^{(i)}$), on peut montrer qu'en chaque site *i*, la valeur $FF^{(i)}$ est une réalisation d'une distribution de Kumaraswamy de paramètres ($n^{(i)}$;1): $FF^{(i)} \sim K[n^{(i)};1]$ [pour une preuve formelle, cf. *Renard et al.*, 2013]. Cette distribution a pour fonction de répartition (avec $n^{(i)}$ égal au nombre de données $d_k^{(i)}$ au site *i*):

$$F_K(t) = t^{n^{(i)}}, 0 \le t \le 1 \quad (4)$$

Le critère *FF* est égal au critère *pval* calculé uniquement sur la plus forte observation. Il ne renseigne donc aucunement sur la justesse des méthodes pour décrire le corps des observations, mais renseigne par contre sur sa justesse dans les extrêmes.

2.2 Représentations graphiques

Pour chacun des trois indices décrits précédemment (*pval*, N_T et *FF*), on connaît la distribution théorique (sous hypothèse de justesse) que devraient suivre les indices calculés sur chaque site. La comparaison entre différentes implémentations consiste donc à évaluer l'adéquation entre la distribution théorique et les valeurs des indices calculés sur chaque site. Si une implémentation conduit à une meilleure adéquation pour un indice donné, on aura tendance à la considérer comme « plus juste » pour cet indice. En pratique, cette adéquation est basée sur un ensemble d'outils graphiques décrits ci-après.

2.2.1 Graphiques probabilité-probabilité (pp-plot)

Pour un site *i* donné avec $n^{(i)}$ observations, notons $z^{(i)}$ un des indices définis en section 2.1 (par exemple, $FF^{(i)}$), et notons $H^{(i)}$ la fonction de répartition de la distribution théorique sous hypothèse de justesse (par exemple, la fonction de répartition d'une loi de Kumaraswamy $K(n^{(i)};1)$ comme décrite dans l'équation (4)). De très nombreux outils sont disponibles pour comparer un échantillon (ici, les valeurs $z^{(i)}$ calculées sur l'ensemble des sites) à une distribution théorique. Malheureusement, il existe une difficulté technique dans le cas présent car la distribution théorique varie d'un site à l'autre : elle dépend en effet du nombre d'observations $n^{(i)}$ (par exemple pour $FF^{(i)}$, la loi théorique au site *i* est une $K(n^{(i)};1)$).

Il est possible de contourner cette difficulté en utilisant un graphique en probabilité-probabilité (pp-plot) : sur chaque site, on transforme les valeurs de l'indice $z^{(i)}$ en probabilités en leur appliquant la fonction de répartition $H^{(i)}$ de la loi théorique. Les valeurs ainsi transformées, $w^{(i)} = H^{(i)}(z^{(i)})$, peuvent alors être triées et tracées contre les fréquences empiriques f_i, pour $i = 1{:}N_{site}$ (on utilisera ici la formule de Hazen, $f_i = (i\text{-}0.5)/N_{site}$). Une courbe proche de la diagonale sera le signe d'une bonne adéquation entre les indices $z^{(i)}$ calculés sur chaque site et leur distribution théorique sous hypothèse de justesse, $H^{(i)}$.

La Figure 1 illustre la réalisation de ces pp-plots. Chaque point de la courbe correspond à une station *i* particulière. En abscisses sont reportées les valeurs triées de $w^{(i)} = H^{(i)}(z^{(i)})$ calculées sur tous les sites $i = 1{:}N_{site}$, tandis que les fréquences empiriques $f_i = (i\text{-}0.5)/N_{site}$ sont reportées en ordonnées. La Figure 1 illustre également quelques formes typiques qui seront observées dans les actions de comparaison :

a. Les courbes représentées dans la Figure 1a correspondent à une tendance pour l'implémentation *M* à la sur- ou à la sous-estimation. Notons qu'avec la convention

utilisée (valeurs transformées de l'indice en abscisses et fréquences empiriques en ordonnées), la sur-estimation correspond à une courbe au-dessus de la diagonale (et inversement pour la sous-estimation).

b. La courbe représentée en Figure 1b ne peut s'observer que lorsque les indices sont calculés sur les données de calage, et dénote une implémentation *M* sur-paramétrée. En effet, la forme en *S* de la courbe indique que les valeurs observées de l'indice sont *moins* variables que ce que l'on observerait avec la vraie distribution parente. En d'autres termes, l'implémentation *M* a tendance à trop « coller » aux données de calage. Ceci s'accompagne généralement d'une dégradation de la justesse sur les données de validation.

c. La courbe représentée en Figure 1c est typique d'un manque de justesse sur les données de validation. On peut noter en particulier le comportement de la courbe dans le coin supérieur droit, qui suggère que l'on observe trop souvent des valeurs élevées de l'indice : en d'autres termes, l'implémentation *M* a tendance à sur-évaluer le caractère « extrême » des données de validation.

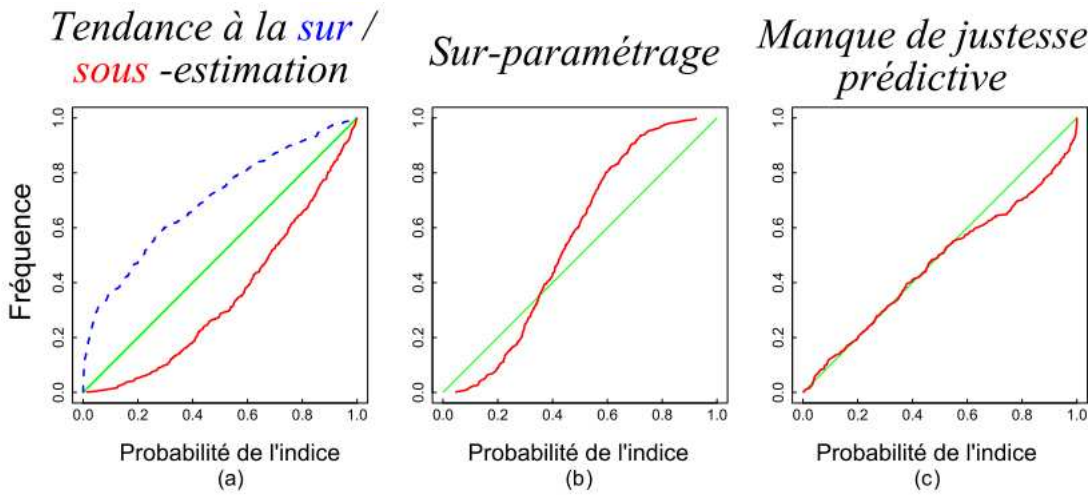

Figure 1. Illustration de quelques courbes typiques pour les pp-plots. L'indice peut désigner *pval*, *FF* ou N_T. La ligne verte représente la première bissectrice ($y = x$).

Précisons que l'indice N_T demande un traitement particulier : en effet, du fait du caractère discret de la distribution théorique de N_T sous hypothèse de justesse (loi binomiale), une procédure de randomisation des valeurs $w^{(i)} = H^{(i)}(z^{(i)})$, doit être appliquée avant la réalisation du pp-plot. Cette procédure est décrite en annexe (section 6).

2.2.2 Graphiques quantile-quantile (qq-plot)

Il peut s'avérer utile d'effectuer un changement de repère dans la Figure 1 afin de focaliser l'attention sur certaines zones du graphiques (par exemple le coin supérieur droit où le manque de justesse est souvent le plus visible). Etant donné que les valeurs en abscisses et en ordonnées dans la Figure 1 sont des probabilités comprises entre 0 et 1, cela peut être aisément implémenté en transformant ces probabilités en quantiles, à l'aide d'une fonction quantile du choix de l'utilisateur : on obtient ainsi un graphique quantile-quantile, ou qq-plot. En particulier la fonction quantile d'une loi de Gumbel (de paramètres (position;échelle) = (0;1)) s'est avérée bien adaptée pour « zoomer » sur les valeurs fortes de l'indice.

La Figure 2(d-f) illustre la réalisation de ces qq-plots, et les compare aux représentations en pp-plot (Figure 2(a-c)). Pour obtenir ces qq-plots, les valeurs en abscisses et en ordonnées dans les Figure 2(a-c) ont simplement été transformées via la fonction quantile

$Q(p) = -\log(-\log(p))$. La Figure 2 illustre que la représentation en qq-plot permet de mieux mettre en évidence le comportement de l'indice pour les valeurs les plus fortes, mais a tendance à « écraser » les courbes pour les valeurs courantes. Les représentations en pp-plot et qq-plot sont donc complémentaires et pourront être utilisées au choix en fonction des caractéristiques du cas d'étude.

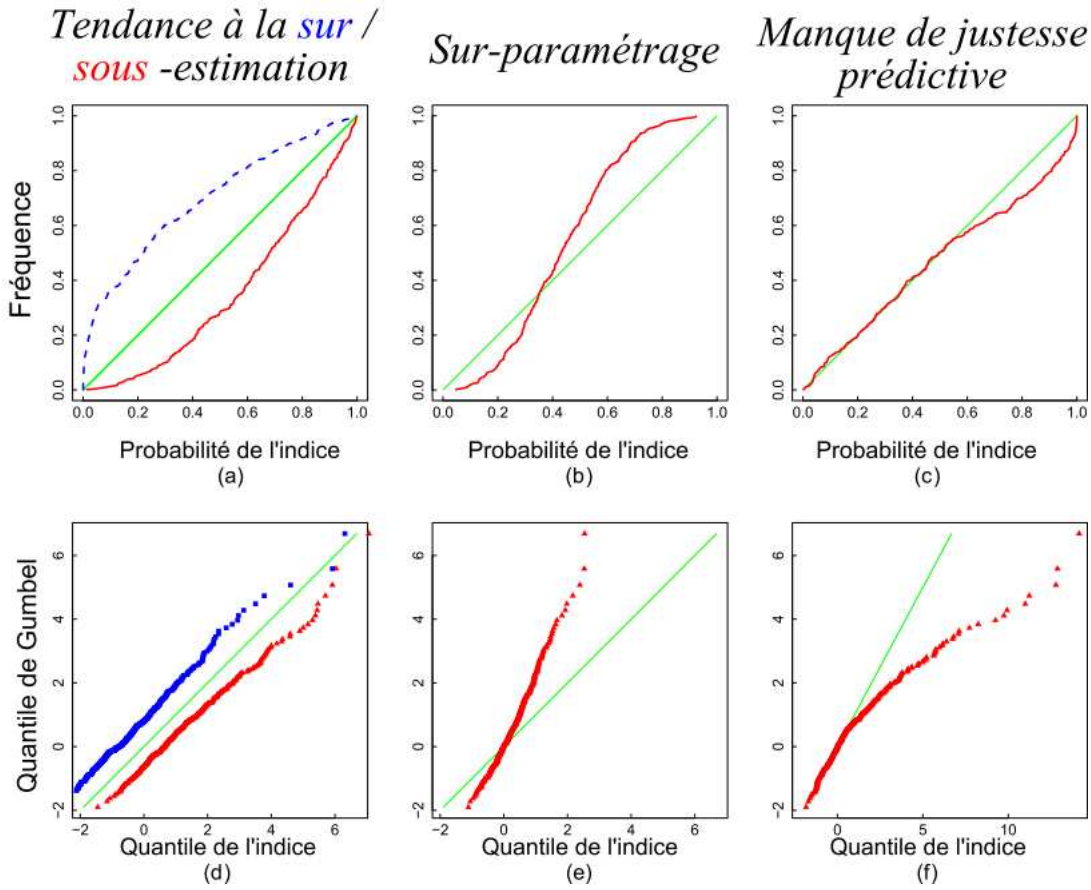

Figure 2. Illustration de quelques courbes typiques pour les pp-plots (a-b-c) et les qq-plots dans un repère de Gumbel (d-e-f). L'indice peut désigner *pval*, *FF* ou N_T. La ligne verte représente la première bissectrice ($y = x$).

2.3 Scores

Dans une optique de comparaison de nombreuses implémentations il peut être utile de résumer les représentations graphiques de la Figure 2 sous la forme d'une valeur numérique qui quantifie l'écart de la courbe à la diagonale. Pour ce faire, on peut associer à chaque courbe (dans la représentation en pp-plot) un score, basé sur l'aire entre la courbe et la diagonale. Cette aire sera normalisée afin de varier entre 0 (faible justesse) et 1 (justesse parfaite). Pour un indice z quelconque (*pval*, *FF* ou N_T), transformé en probabilité via la transformation $w^{(i)} = H^{(i)}(z^{(i)})$, on peut définir ce score de la manière suivante :

$$score = 1 - 2 * Aire(courbe, diagonale)$$

$$= 1 - \frac{2}{N_{site} + 1} \sum_{i=1}^{N_{site}} \left| w^{(i)} - f_i \right| \qquad (5)$$

La Figure 3 illustre les scores de justesse associés à quelques courbes dans une représentation en pp-plot.

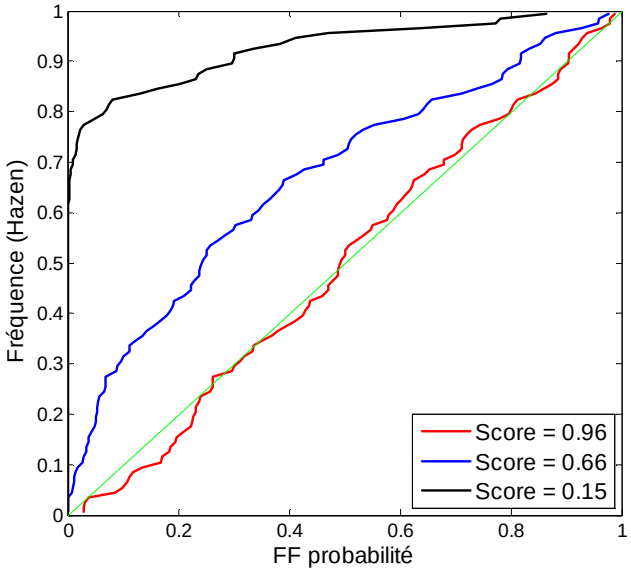

Figure 3. Illustration des scores de justesse.

2.4 Discussion sur l'interprétation des graphiques et des scores

Les représentations graphiques présentées dans les sections précédentes doivent être interprétées avec précaution. Afin d'illustrer l'utilisation de ces diagnostics graphiques, nous proposons ci-dessous une application à des données simulées.

Des données ont été simulées sur 30 années et 500 sites à partir d'une distribution parente GEV, identique pour tous les sites, et de paramètres (100, 50, -0.2) (position, échelle, forme). Nous nous proposons d'évaluer la justesse des 3 implémentations suivantes :

1. L'implémentation M1 conduit à utiliser une loi de Gumbel de paramètres (100, 50) sur tous les sites. Il s'agit donc d'une sous-estimation systématique.
2. L'implémentation M2 conduit à utiliser une loi GEV de paramètres (100, 50, -0.3) sur tous les sites. Il s'agit donc d'une sur-estimation systématique.
3. L'implémentation M3 conduit à utiliser une loi GEV de paramètres de position = 100, d'échelle = 50, et dont le paramètre de forme varie uniformément entre –0.5 et 0.1 sur l'ensemble des sites. L'implémentation M3 commet donc des erreurs aléatoires en fonction des sites : sous-estimation sur certains sites, sur-estimation sur d'autres.

La Figure 4 illustre les graphiques de justesse obtenus pour ces 3 implémentations, avec les indices *FF*, N_{10} et N_{10000}, sur la base de représentations en pp-plot. Les principales observations que l'on peut faire sont les suivantes :

- Comme attendu, l'implémentation M1 conduit à des courbes sous la diagonale (sous-estimation), tandis que M2 conduit à des courbes au-dessus de la diagonale (sur-estimation) ;
- Les courbes correspondant à l'implémentation M3 croisent la diagonale, illustrant le fait que cette implémentation conduit parfois à des sous-estimations, parfois à des sur-estimations ;
- On observe néanmoins que les courbes M3 sont globalement plus proches de la diagonale que les courbes M1 et M2 : ceci est dû au fait que les erreurs aléatoires commises par l'implémentation M3 sont beaucoup plus difficiles à détecter que les erreurs systématiques commises par les implémentations M1 et M2.

Signalons de plus quelques dangers dans l'interprétation de ces graphiques, à travers des exemples de « mauvais diagnostics » que l'on pourrait être tenté d'émettre :

- « *La courbe M2 est proche de la diagonale pour l'indice N_{10000}, on peut donc en conclure que l'implémentation M2 est juste pour l'estimation de la crue décamillénale* ». Ce n'est évidemment pas le cas : dans ce cas d'étude simulé, le quantile décamillénal estimé par M2 vaut environ 2600, alors que le vrai quantile décamillénal de la distribution parente vaut en réalité 1400, soit une sur-estimation de plus de 80% ! Le fait que la courbe M2 reste proche de la diagonale signifie seulement que les données ne suffisent pas à mettre en évidence un net manque de justesse pour ce quantile : en effet, le nombre de dépassement du quantile décamillénal estimé vaut zéro sur tous les sites… ce qui n'est pas anormal pour une période d'observation de 30 ans sur 500 sites ! Cet exemple illustre que les courbes de justesse doivent être interprétées de la façon suivante :
    i. Une courbe s'écartant de la diagonale signale une implémentation qui manque de justesse ;
    ii. Mais l'inverse n'est pas vrai : une courbe proche de la diagonale ne prouve pas qu'une implémentation est juste !
    iii. Les diagnostics graphiques ont avant tout une valeur comparative : si une courbe M1 est proche de la diagonale et qu'une autre M2 s'en écarte, on pourra dire que M1 est plus juste que M2.
- « *Les courbes sont globalement plus proches de la diagonale pour l'indice N_{10000} que pour l'indice N_{10}, on peut donc en conclure que les estimations décamillénales sont plus justes que les estimations décennales* ». Cette comparaison entre différents indices n'a pas vraiment de sens, car la puissance à détecter un manque de justesse varie fortement d'un indice à l'autre. En l'occurrence, le fait que les courbes soient globalement plus proches de la diagonale pour l'indice N_{10000} que pour l'indice N_{10} reflète seulement le fait qu'il est beaucoup plus difficile de détecter un manque de justesse pour une estimation décamillénale que pour une estimation décennale. En conséquence, s'il est tout à fait possible de comparer plusieurs implémentations pour un indice donné, on s'abstiendra en revanche de comparer plusieurs indices pour une implémentation donnée.

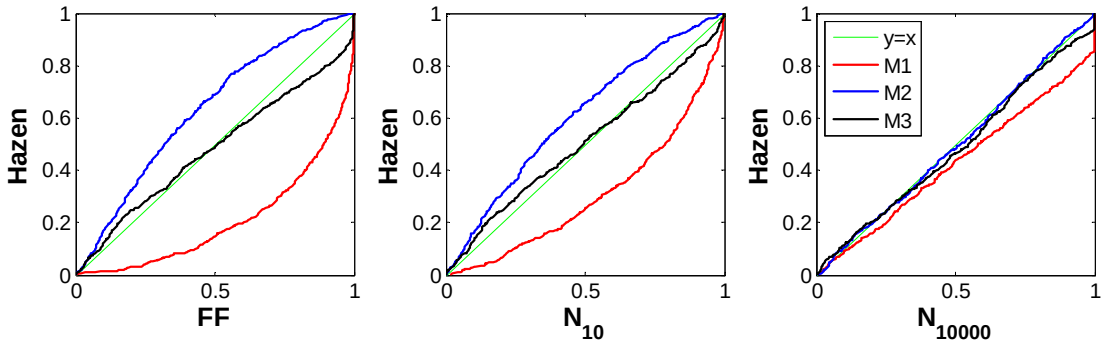

Figure 4. Diagnostics de justesse (pp-plots) pour le cas d'étude simulé.

Signalons pour finir que l'utilisation de scores, comme proposé en section 2.3, conduit à une perte d'information importante et ne peut donc pas suffire à émettre des conclusions définitives. A titre d'illustration, la Figure 5 montre deux pp-plots conduisant à des scores quasiment identiques, mais reflétant des comportements bien distincts : alors que la courbe

bleue dénote une tendance à la sur-estimation systématique, la courbe rouge reflète plutôt des erreurs aléatoires (à la fois des sous- et des sur-estimations).

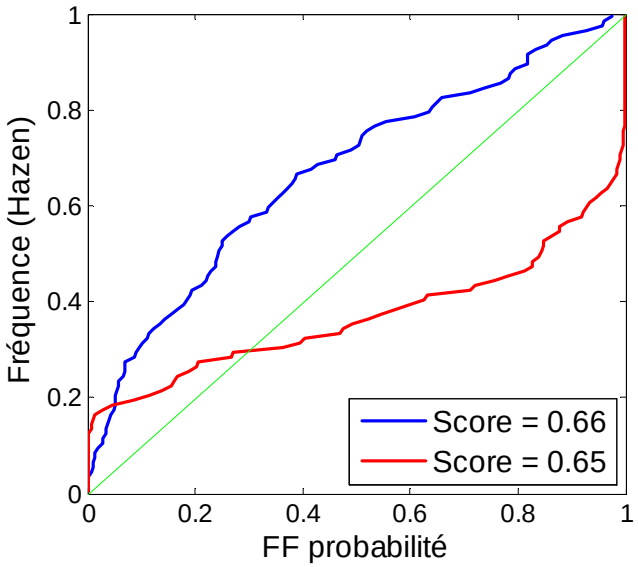

Figure 5. Exemple illustrant deux courbes en pp-plot conduisant à des scores similaires mais dénotant des comportements bien distincts.

3 Stabilité

Cette section présente les outils utilisés pour évaluer la stabilité des implémentations candidates (cf. section 1.3). La démarche est similaire à celle adoptée pour la justesse : définition d'un indice de stabilité (section 3.1), et utilisation d'outils graphiques (section 3.2) et de scores (section 3.3) pour comparer plusieurs implémentations.

3.1 Indice

La stabilité est quantifiée sur la base d'un indice décrivant la différence relative entre les quantiles estimés sur deux jeux de calage distincts $\boldsymbol{c}_1$ et $\boldsymbol{c}_2$. L'indice $SPAN_T$ proposé par *Garavaglia et al*. [2011] est utilisé. Soit $\hat{q}_T^{(i)}$ le quantile de période de retour T estimé au site i, à partir de la distribution estimée $\hat{F}_M^{(i)}$. Pour un site i donné, $SPAN_T$ est formellement défini comme suit:

$$SPAN_T^{(i)} = \frac{\left|\hat{q}_T^{(i)}(\boldsymbol{c}_1) - \hat{q}_T^{(i)}(\boldsymbol{c}_2)\right|}{\frac{1}{2}\left(\hat{q}_T^{(i)}(\boldsymbol{c}_1) + \hat{q}_T^{(i)}(\boldsymbol{c}_2)\right)} \qquad (6)$$

L'indice $SPAN_T$ varie entre 0 et 2 : pour une implémentation parfaitement stable, $\hat{q}_T^{(i)}(\boldsymbol{c}_1) = \hat{q}_T^{(i)}(\boldsymbol{c}_2)$ et $SPAN_T = 0$. Inversement, $SPAN_T$ tend vers la valeur 2 lorsqu'un des quantiles est beaucoup plus grand que l'autre.

3.2 Représentation graphique

La comparaison entre plusieurs implémentations candidates est effectuée simplement en comparant la fonction de répartition empirique des valeurs $SPAN_T^{(i)}$ calculées sur l'ensemble des sites $i = 1{:}N_{site}$. La Figure 6 illustre la réalisation de ces courbes de stabilité. Chaque point de la courbe correspond à une station i particulière. En abscisses sont reportées les fréquences

empiriques $f_i = (i\text{-}0.5)/N_{site}$, tandis que les valeurs triées de $SPAN_T^{(i)}$ calculées sur tous les sites ($i = 1{:}N_{site}$) sont reportées en ordonnées. Avec ces conventions, l'implémentation dont la courbe $SPAN_T$ reste la plus proche de l'axe des abscisses est la plus stable : dans le cas de la Figure 6, l'implémentation rouge apparaît ainsi plus stable que l'implémentation bleue, qui est elle-même plus stable que l'implémentation verte.

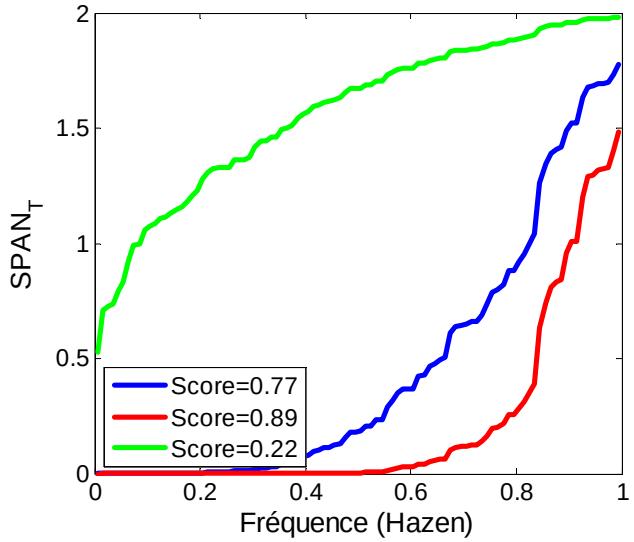

Figure 6. Illustration des courbes et des scores de stabilité.

3.3 Score

Comme c'était le cas pour la justesse, il peut être utile de résumer la représentation de la Figure 6 sous la forme d'un score de stabilité. Pour ce faire, on peut associer à chaque courbe un score basé sur l'aire sous la courbe. Cette aire sera normalisée afin de varier entre 0 (faible stabilité) et 1 (stabilité parfaite). Etant donné que les valeurs $SPAN_T^{(i)}$ varient entre 0 et 2, l'aire sous les courbes de la Figure 6 varie également entre 0 et 2, ce qui conduit au score suivant :

$$score = 1 - 0.5 * Aire(courbe, abscisses)$$

$$= 1 - \frac{1}{2N_{site}} \sum_{i=1}^{N_{site}} SPAN_T^{(i)} \qquad (7)$$

4 Autres représentations combinant justesse et stabilité

Une vue d'ensemble des scores à la fois de justesse et de stabilité peut être obtenue en représentant ces scores (qui sont tous normalisés entre 0 et 1) dans un graphique en étoile. A titre d'illustration, la Figure 7 montre une comparaison pour trois implémentations. Ce graphique nous indique que sur la base des scores (qui, rappelons le, ne fournissent qu'une vue incomplète de la justesse et de la stabilité, cf. section 2.4), l'implémentation B est plus performante que l'implémentation A à la fois en justesse (scores *FF*, N_{10} et N_{100}) et en stabilité (scores $SPAN_T$). L'implémentation C quant à elle est nettement la plus stable, mais présente par contre une justesse bien plus médiocre que les deux autres implémentations. Sur la base de ce graphique, on aurait donc tendance à préférer l'implémentation B qui constitue le meilleur compromis entre justesse et stabilité.

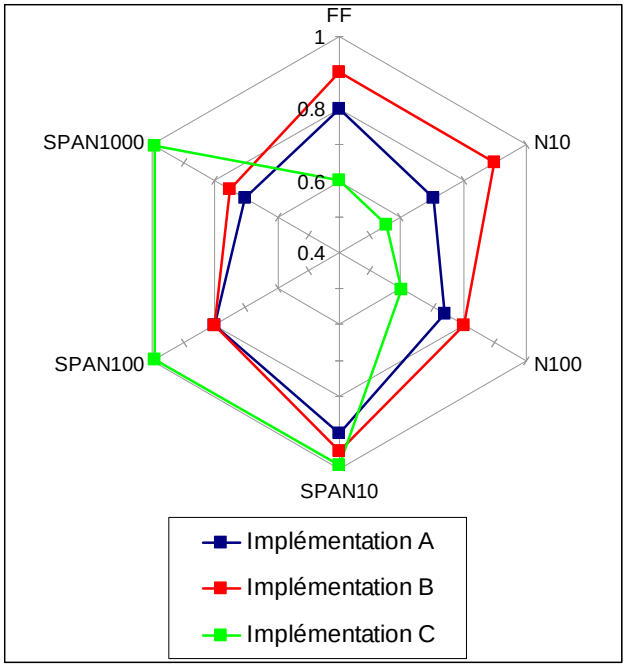

Figure 7. Exemple de représentation en étoile des scores de justesse et de stabilité.

Une autre représentation intéressante, combinant des éléments liés à la justesse et à la stabilité, est présentée en Figure 8. Ce graphique est obtenu de la manière suivante : on commence par calculer un score de justesse (*FF* par exemple) sur les données de validation. Puis l'implémentation évaluée est recalée sur ces mêmes données de validation, qui deviennent donc des données de calage, et le score est recalculé à partir de cette implémentation recalée. On peut alors comparer les scores en mode calage et validation, qui sont calculés sur exactement les mêmes données, mais diffèrent à cause du statut de ces données vis-à-vis de l'implémentation (données utilisées pour le calage ou non).

L'analyse de l'histogramme en mode validation donne la même information que le test *FF* avec l'indication sur la justesse (SHYPRE2 disqualifié du à l'asymétrie de l'histogramme, et fréquence importante de 0 et de 1 avec « GEV bornée », due à de nombreuses valeurs jugées improbables). La dispersion des résultats entre $FF_{\text{validation}}$ et FF_{calage} donne une information sur la stabilité : en effet, une implémentation stable devrait donner des *FF* similaires puisque les données sont les mêmes (seul le mode de calage change). En l'occurrence, les deux implémentations SHYPRE sont beaucoup plus stable que l'implémentation « GEV bornée ». Enfin, la forme de l'histogramme en mode calage permet de repérer des cas de sur-paramétrage (cf. courbe fortement sous-dispersée pour « GEV bornée »).

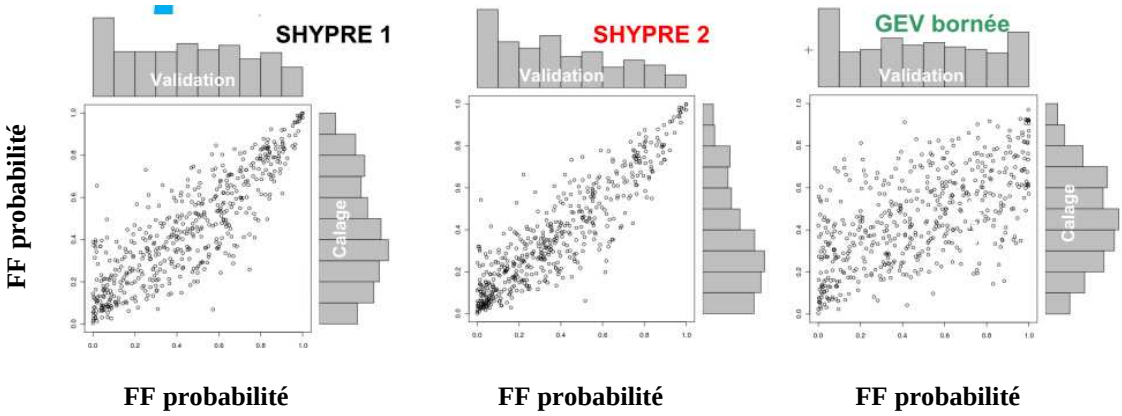

Figure 8. Exemple de représentation combinant justesse et stabilité.

5 Comparaison des incertitudes : utilisation de distributions prédictives

L'évaluation de la justesse et de la stabilité des incertitudes estimées bute sur un certain nombre de problèmes pratiques. Par exemple, il serait tentant de définir un critère consistant à dénombrer le nombre de points inclus dans une enveloppe d'incertitude au niveau 1-α, et vérifier que ce nombre représente bien une proportion proche de (1-α) des observations (Figure 9). Néanmoins, cette évaluation serait biaisée, car elle ignore l'incertitude liée au calcul des fréquences empiriques. Pour s'en convaincre, il suffit d'imaginer la méthode « parfaite », qui estime systématiquement la vraie distribution, sans incertitude. Le nombre de points dans l'intervalle de confiance serait alors nul, alors que la quantification des incertitudes est parfaitement juste !

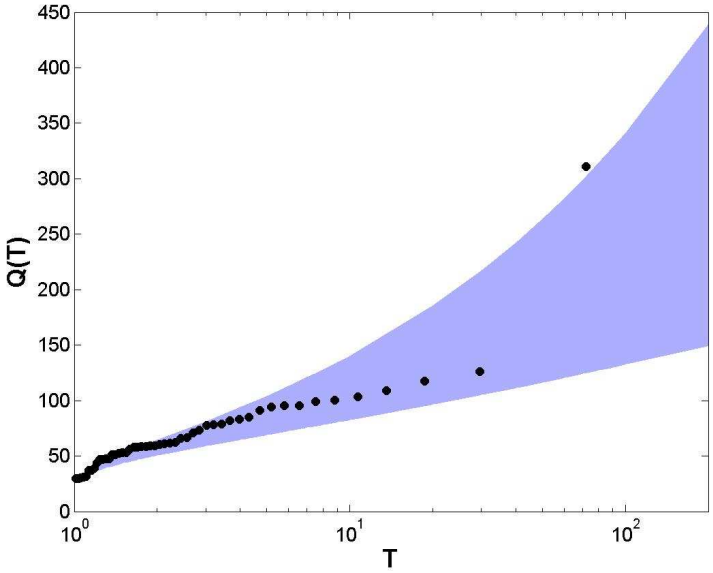

Figure 9. Illustration d'une enveloppe d'incertitude au niveau 1-α.

Une approche alternative consiste à ne pas directement utiliser une enveloppe d'incertitude pour évaluer la justesse des incertitudes, mais à intégrer ces incertitudes dans une nouvelle distribution : la distribution prédictive.

5.1 Principe d'une distribution prédictive

Le concept de distribution prédictive est central en statistiques bayésiennes, mais peut également être étendu aux statistiques classiques. Des définitions précises et rigoureuses peuvent être trouvées dans [*Renard et al.*, 2013]. Nous privilégions ici une description plus empirique, afin d'expliquer le principe sous-jacent.

La distribution estimée $\hat{F}_M^{(i)}(y) = F_M^{(i)}(y \mid \hat{\boldsymbol{\theta}})$, utilisée jusqu'ici pour évaluer justesse et stabilité des implémentations candidates (sections et 2 et 3), correspond à la distribution supposée par l'implémentation M ($F_M^{(i)}$) pour une valeur particulière des paramètres ($\hat{\boldsymbol{\theta}}$). Cette valeur $\hat{\boldsymbol{\theta}}$ correspond au "meilleur" estimateur, où la notion d'optimalité dépend de la méthode d'estimation utilisée. Par exemple, pour une estimation par la méthode des moments, $\hat{\boldsymbol{\theta}}$ est optimal au sens où il permet de retrouver les moments empiriques calculés sur les données de calage. Pour une estimation par maximum de vraisemblance, $\hat{\boldsymbol{\theta}}$ est optimal au sens où il maximise la vraisemblance des données de calage.

En dépit de cette « optimalité », l'estimateur $\hat{\boldsymbol{\theta}}$ reste entaché d'incertitudes d'estimation. Ainsi, une autre valeur $\hat{\boldsymbol{\theta}}^* \neq \hat{\boldsymbol{\theta}}$, conduisant à une distribution estimée $F_M^{(i)}(y \mid \hat{\boldsymbol{\theta}}^*)$, conduit peut-être à une meilleure approximation de la distribution parente inconnue. Pour aller plus loin, l'incertitude d'estimation des paramètres $\hat{\boldsymbol{\theta}}$ peut être représentée par la *distribution d'échantillonnage* de l'estimateur $\hat{\boldsymbol{\theta}}$, qui représente (schématiquement) la distribution de l'ensemble des valeurs $\hat{\boldsymbol{\theta}}^*$ qui sont plausibles au vu des données de calage.

Imaginons alors le procédé de simulation suivant, illustré en Figure 10 : soit $\hat{\boldsymbol{\theta}}^*$ une valeur « plausible » des paramètres. En utilisant cette valeur, la distribution estimée est $F_M^{(i)}(y \mid \hat{\boldsymbol{\theta}}^*)$ (courbe rouge dans la Figure 10). En simulant un grand nombre de valeurs à partir de cette distribution, on génère une réalisation de ce que les futures crues (ou pluies) pourraient être, *sous hypothèse que la valeur $\hat{\boldsymbol{\theta}}^*$ conduit à la vraie distribution parente*. Or, à cause de l'incertitude d'estimation, cette valeur $\hat{\boldsymbol{\theta}}^*$ n'est qu'une valeur plausible parmi d'autres: on pourrait également considérer une autre valeur $\hat{\boldsymbol{\theta}}^{**}$, et simuler un grand nombre de valeurs à partir de la distribution estimée $F_M^{(i)}(y \mid \hat{\boldsymbol{\theta}}^{**})$ (courbe bleue en Figure 10).

Si l'on répète ce procédé de simulation pour l'ensemble des valeurs « plausibles » $\hat{\boldsymbol{\theta}}^*$ (en d'autres termes, pour un ensemble de $\hat{\boldsymbol{\theta}}^*$ tirés au hasard dans la distribution d'échantillonnage de l'estimateur), on génère ainsi des réalisations de ce que les futures crues (ou pluies) pourraient être, *en considérant l'ensemble des valeurs $\hat{\boldsymbol{\theta}}^*$ qui sont plausibles au vu des données de calage*. La distribution de l'ensemble de réalisations futures est appelée la distribution prédictive (histogramme gris en Figure 10).

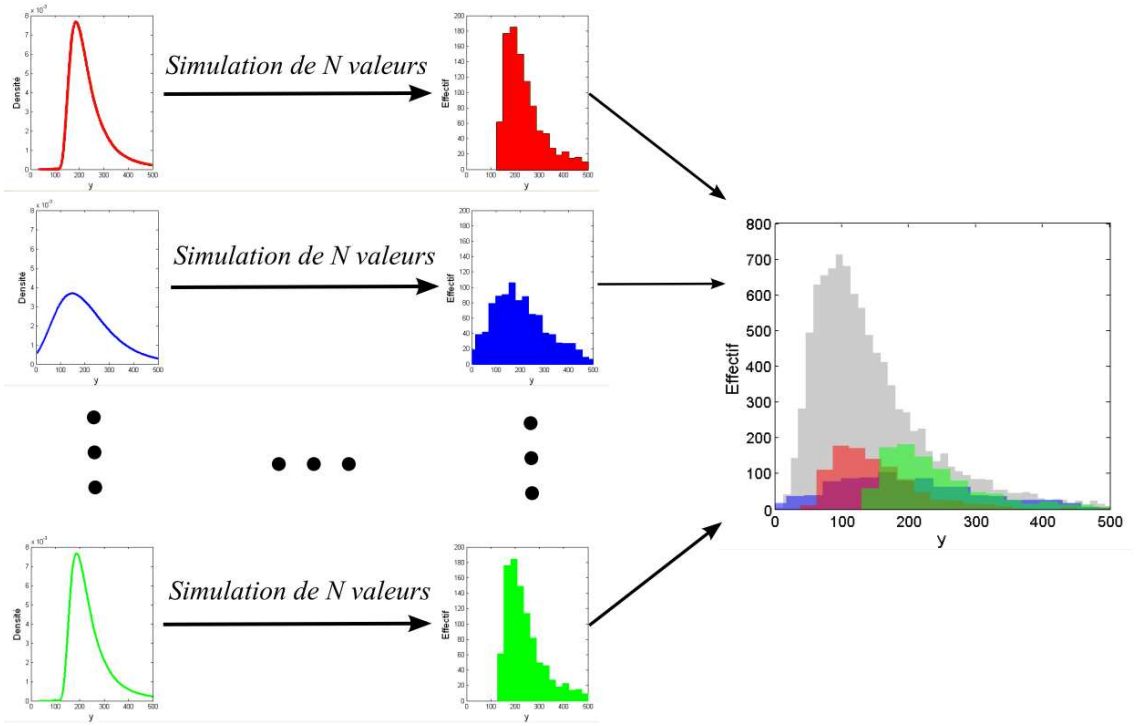

Figure 10. Illustration de la génération d'une distribution prédictive.

5.2 Définition de la distribution prédictive

Le procédé de simulation décrit ci-dessus peut être formalisé en termes mathématiques et permet de définir formellement la densité de probabilité de la distribution prédictive. Soit $f_M(y|\boldsymbol{\theta})$ la densité de la distribution supposée par l'implémentation *M*, et $I_M(\boldsymbol{\theta}|\boldsymbol{c})$ la densité de la distribution d'échantillonnage de l'estimateur $\hat{\boldsymbol{\theta}}$ (qui dépend des données de calage $\boldsymbol{c}$). La densité de la distribution prédictive est donnée par :

$$\pi_M(y) = \int f_M(y|\boldsymbol{\theta}) I_M(\boldsymbol{\theta}|\boldsymbol{c}) d\boldsymbol{\theta} \tag{8}$$

La distribution prédictive est obtenue en intégrant la distribution supposée des observations, $f_M(y|\boldsymbol{\theta})$, sur la distribution d'échantillonnage de l'estimateur, $I_M(\boldsymbol{\theta}|\boldsymbol{c})$, qui représente l'incertitude. Par opposition, la distribution estimée consiste à utiliser la distribution supposée pour *une* valeur particulière des paramètres. La Figure 11 illustre la différence, en termes de quantiles, entre la distribution prédictive (bleue) et la distribution estimée (rouge). En général, la distribution prédictive est plus variable que la distribution estimée (car elle intègre les incertitudes). Ceci est d'autant plus vrai dans les cas des quantiles de pluie ou crue que les intervalles d'incertitude sont typiquement très asymétriques, avec une borne supérieure plus éloignée que la borne inférieure de la distribution estimée.

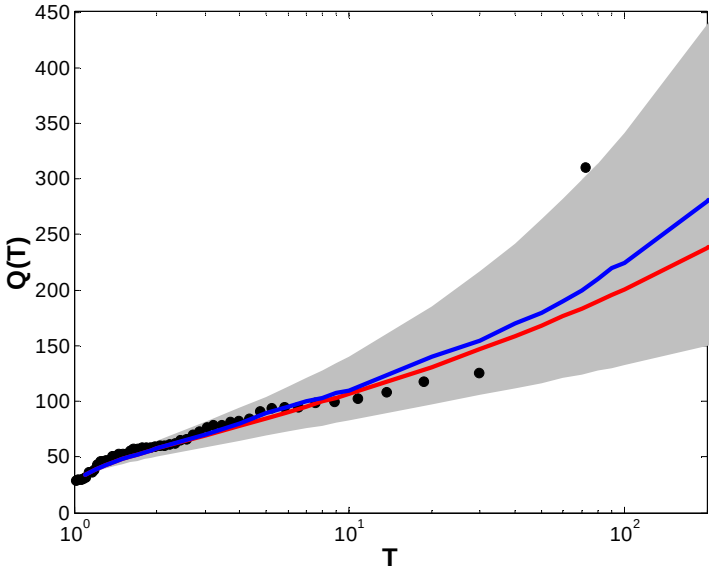

Figure 11. Comparaison entre distribution prédictive (bleue) et distribution estimée (rouge).

5.3 Obtention pratique de la distribution prédictive

En pratique, l'intégration figurant dans l'équation (8) n'est pas réalisée analytiquement, mais est basée sur des simulations Monte-Carlo. La procédure suit le schéma de simulation discuté en section 5.1, et peut être formalisée comme suit :

Do $j = 1 : N_{sim}$

- Générer $\hat{\boldsymbol{\theta}}_j^*$ à partir de la distribution d'échantillonnage de l'estimateur, $I_M(\boldsymbol{\theta} \mid \boldsymbol{x})$.
- Générer un échantillon de taille m, $\boldsymbol{y}_j^* = \left(y_{j,1}^*, ... y_{j,m}^*\right)$, à partir de la distribution estimée avec $\hat{\boldsymbol{\theta}}_j^*$, $F_M^{(i)}(y \mid \hat{\boldsymbol{\theta}}_j^*)$.

END

L'ensemble des $m*N_{sim}$ valeurs générées, $\left(\boldsymbol{y}_j^*\right)_{j=1:N_{sim}}$, est un échantillon issu de la distribution prédictive.

Notons que dans certains cas, l'incertitude est exprimée sur les quantiles plutôt que sur les paramètres. La procédure de simulation ci-dessus peut alors être modifiée de la façon suivante (voir aussi Figure 12) :

Do $j = 1 : N_{sim}$

- Générer u_j^* à partir d'une loi uniforme sur [0 ;1].
- Transformer u_j^* en période de retour, $T_j^* = 1/(1-u_j^*)$.
- Générer un échantillon de taille m, $\boldsymbol{y}_j^* = \left(y_{j,1}^*, ... y_{j,m}^*\right)$, à partir de la distribution d'échantillonnage du quantile $q(T_j^*)$.

END

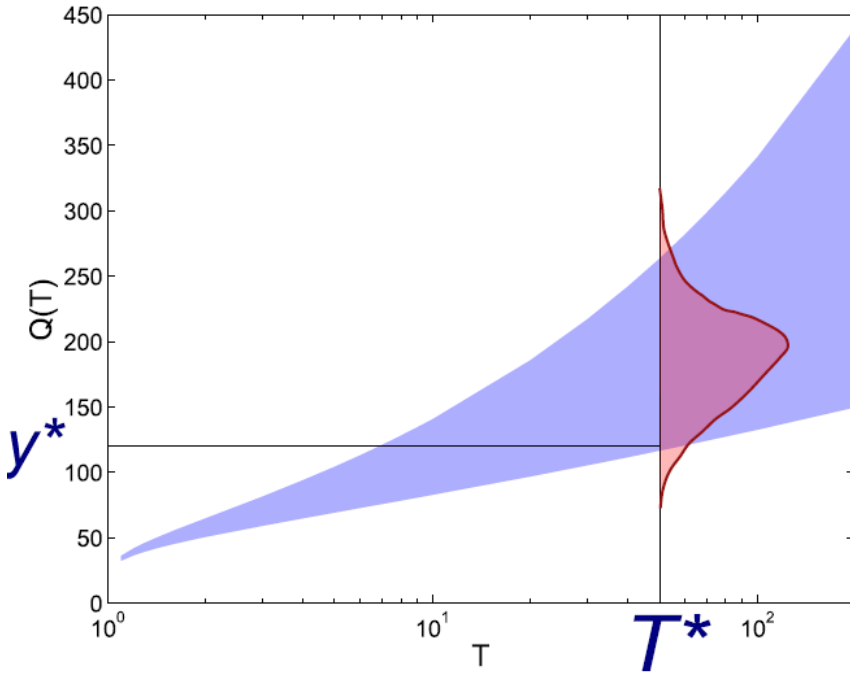

Figure 12. Illustration de la procédure de Monte-Carlo pour la génération de valeurs issues de la distribution prédictive lorsque l'incertitude est exprimée sur les quantiles.

5.4 Utilisation des distributions prédictives pour comparer les incertitudes estimées

Un des avantages majeurs d'utiliser la distribution prédictive pour évaluer indirectement la justesse et la stabilité des incertitudes estimées est que la distribution prédictive… est une distribution ! En conséquence, on peut définir sa fonction de répartition $\Pi_M(y)$ et appliquer l'ensemble des indices, graphiques et scores présentés en sections 2 et 3 en remplaçant simplement la distribution estimée $\hat{F}_M^{(i)}(y)$ par la distribution prédictive $\Pi_M(y)$.

En pratique, la fonction de répartition $\Pi_M(y)$ est estimée empiriquement sur la base des $m*N_{sim}$ valeurs générées $\left(y_j^*\right)_{j=1:N_{sim}}$.

5.5 Stabilité des incertitudes : COVER$_T$

Un dernier indice visant à caractériser la stabilité des incertitudes peut être calculé sans passer par la notion de distribution prédictive. Cet indice est calculé en considérant le recouvrement des intervalles d'incertitude obtenus avec deux périodes de calage C_1 et C_2 [*Garavaglia et al.*, 2011]. Plus précisément, soient $a_{\alpha,i}$ et $b_{\alpha,i}$ les limites de l'intervalle de confiance de niveau α (en %) du quantile $\hat{q}_{T,i}$ pour une station *i*, avec deux échantillons :

Echantillon C_1 : $P\left[a_{\alpha,i}(C_1) < \hat{q}_{T,i}(C_1) < b_{\alpha,i}(C_1)\right] = \alpha$

Echantillon C_2 : $P\left[a_{\alpha,i}(C_2) < \hat{q}_{T,i}(C_2) < b_{\alpha,i}(C_2)\right] = \alpha$

Le croisement des deux intervalles de confiance donne les limites :

$a = Max(a_{\alpha,i}(C_1); a_{\alpha,i}(C_2))$ et $b = Min(b_{\alpha,i}(C_1); b_{\alpha,i}(C_2))$

La probabilité de recouvrement des deux intervalles vaut :

$$COVER_{T,i} = (1/\alpha^2)\, P\left[a < \hat{q}_{T,i}(C_1) < b\right] P\left[a < \hat{q}_{T,i}(C_2) < b\right] \quad (9)$$

La Figure 13 fournit une représentation graphique de ce calcul. Cet indice étant compris entre 0 et 1, un score peut aisément être calculé en effectuant la moyenne des indices obtenus sur tous les sites, $score = (1/N_{site}) \sum_{i=1}^{N_{site}} COVER_{T,i}$

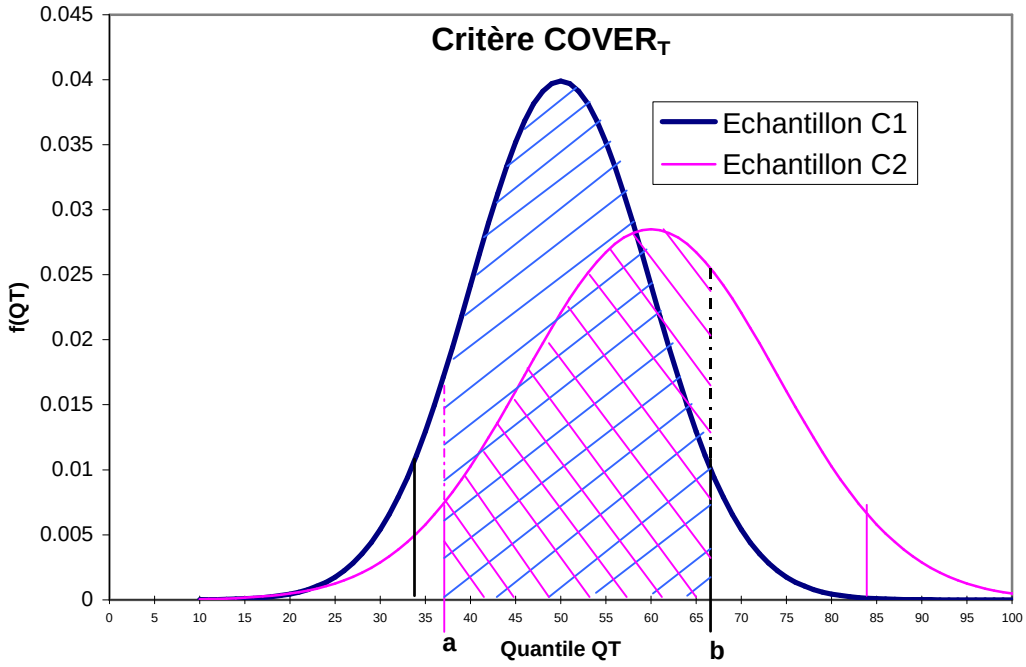

Figure 13. Illustration du calcul de l'indice COVER$_T$.

6 Appendice : pp-plot randomisé pour l'indice N_T

Soit $b(-1) = 0$ et $b(j) = \Pr(N \le j), j \ge 0$, où N est une variable aléatoire suivant une loi binomiale $Bin(n^{(i)}, 1/T)$. En un site i donné, sur lequel la valeur $N_T^{(i)}$ a été observée, la procédure classique de génération d'un pp-plot demande de calculer la valeur $w^{(i)} = \Pr(N \le N_T^{(i)}) = b\left(N_T^{(i)}\right)$. La procédure de randomisation consiste à remplacer cette valeur $w^{(i)}$ par une valeur $w^{(i)*}$ tirée aléatoirement à partir d'une distribution uniforme entre $b\left(N_T^{(i)} - 1\right)$ et $b\left(N_T^{(i)}\right)$.

Une justification du bien-fondé théorique de cette procédure peut être trouvée dans [*Renard et al.*, 2013].

7 Références

Arnaud, P., and J. Lavabre (1999), Using a stochastic model for generating hourly hyetographs to study extreme rainfalls, *Hydrological Sciences Journal*, *44*(3), 433-446

Arnaud, P., and J. Lavabre (2002), Coupled rainfall model and discharge model for flood frequency estimation, *Water Resources Research*, *38*(6)
Cipriani, T., T. Toilliez, and E. Sauquet (2012), Estimating 10 year return period peak flows and flood durations at ungauged locations in France, *La houille blanche*
England, J. F., R. D. Jarrett, and J. D. Salas (2003), Data-based comparisons of moments estimators using historical and paleoflood data, *J. Hydrol.*, *278*(1-4), 172-196
Garavaglia, F., J. Gailhard, E. Paquet, M. Lang, R. Garcon, and P. Bernardara (2010), Introducing a rainfall compound distribution model based on weather patterns sub-sampling, *Hydrol. Earth Syst. Sci.*, *14*(6), 951-964
Garavaglia, F., M. Lang, E. Paquet, J. Gailhard, R. Garcon, and B. Renard (2011), Reliability and robustness of a rainfall compound distribution model based on weather pattern sub-sampling, *Hydrology and Earth System Sciences.*, *15*(2), 519-532, doi: 10.5194/hess-15-519-2011.
Gunasekara, T. A. G., and C. Cunnane (1992), Split Sampling Technique for Selecting a Flood Frequency-Analysis Procedure, *J. Hydrol.*, *130*(1-4), 189-200
Interagency Advisory Committee on Water Data (1982), *Guidelines for determining flood-flow frequency: Bulletin 17B of the Hydrology Subcommittee*, U.S. Geological Survey, Reston, Va.
Laio, F. (2004), Cramer-von Mises and Anderson-Darling goodness of fit tests for extreme value distributions with unknown parameters, *Water Resources Research*, *40*(9)
Naulet, R. (2002), Utilisation de l'information des crues historiques pour une meilleure prédétermination du risque d'inondation. Application au bassin de l'Ardèche à Vallon Pont-d'Arc et St-Martin d'Ardèche, Ph.D. Thesis thesis, 322 pp, University Joseph Fourier / University of Québec / INRS / Cemagref, Lyon, France.
Naulet, R., M. Lang, T. B. M. J. Ouarda, D. Coeur, B. Bobee, A. Recking, and D. Moussay (2005), Flood frequency analysis on the Ardeche river using French documentary sources from the last two centuries, *J. Hydrol.*, *313*(1-2), 58-78
Neppel, L., P. Arnaud, and J. Lavabre (2007), Extreme rainfall mapping: Comparison between two approaches in the Mediterranean area, *C. R. Geosci.*, *339*(13), 820-830, doi: 10.1016/j.crte.2007.09.013.
Neppel, L., et al. (2010), Flood frequency analysis using historical data: accounting for random and systematic errors, *Hydrol. Sci. J.-J. Sci. Hydrol.*, *55*(2), 192-208, doi: 10.1080/02626660903546092.
Paquet, E., J. Gailhard, and R. Garcon (2006), Evolution de la méthode du gradex : approche par type de temps et modélisation hydrologique, *La houille blanche*, *5*, 80-90
Payrastre, O. (2005), Faisabilité et utilité du recueil de données historiques pour l'étude des crues extrêmes de petits cours d'eau. Etude du cas de quatre bassins versants affluents de l'Aude, Ph.D. Thesis thesis, 392 pp, ENPC, Marne la Vallée, France.
Payrastre, O., E. Gaume, and H. Andrieu (2011), Usefulness of historical information for flood frequency analyses: Developments based on a case study, *Water Resources Research*, *47*, doi: 10.1029/2010WR009812.
Renard, B., et al. (2013), Data-based comparison of frequency analysis methods: a general framework, *Water Resources Research, in press*