METHODES BAYESIENNES ET MODELISATION DES RISQUES GEOPHYSIQUES EXTREMES

Eric Parent

Jacques Bernier

École Nationale du Génie Rural, des Eaux et Forêts
Laboratoire Gestion du Risque en Environnement et Sciences de l'Eau
15 Avenue du Maine - 75732 PARIS cedex 15

Résumé :

Les estimations de risques induits par les événements extrêmes tels que les crues des rivières sont soumises aux larges incertitudes de l'extrapolation des distributions des variables en jeu comme les débits des rivières. En effet les événements dont il faut se protéger sont généralement beaucoup plus rares que ceux déjà systématiquement enregistrés. Il existe une littérature très abondante sur le choix des modèles, les erreurs d'échantillonnage et les moyens de réduire partiellement les incertitudes induites. Mais, par respect d'une soi disant « objectivité scientifique », les hydrologues statisticiens classiques hésitent à introduire les connaissances a priori des experts dans leurs analyses. La mise en œuvre de l'approche bayésienne dans le cadre d'un modèle de dépassements (POT) classique Poisson-Pareto généralisé, permet la prise en compte rationnelle des opinions d'experts. La méthode, utilisant systématiquement les techniques de simulation dites MCMC, est appliquée au cas des crues de la Garonne et montre le gain notable en précision obtenu.

Mots clés : *Analyse bayésienne ; Méthodes de Monte Carlo ; Modèles de dépassements (POT) ; ; Théorie des valeurs extrêmes ; Dimensionnement des protections contre les crues ; Quantification des expertises a priori ; Distribution a priori semi-conjuguée.*

1. Introduction

L'extrapolation des courbes de distribution vers les grandes valeurs d'un côté et l'approche bayésienne de l'autre sont des méthodes également rejetées par nombre de statisticiens. Certains adoptent l'opinion qu'il n'est pas scientifiquement raisonnable d'extrapoler des modèles statistiques au-delà des données observées. D'autres ou les mêmes rejettent les procédures bayésiennes parce qu'elles sont intrinsèquement subjectives ([COL96b]). Ces arguments nous semblent toutefois de portée limitée car l'extrapolation a toujours été un instrument de découverte scientifique essentiel et tout travail de modélisation a une part de subjectivité importante de la part d'un chercheur même en statistique classique. Nous invitons le lecteur à considérer avec nous que les difficultés de l'extrapolation des distributions ne sont pas d'ordre philosophique mais d'ordre pratique : de nombreuses incertitudes entachent toute extrapolation. Prendre en compte ces incertitudes et toute information pertinente pour en réduire l'incidence, devient alors le travail essentiel du chercheur. L'approche bayésienne, appuyée par les méthodes de calcul par « simulation Monte Carlo en chaînes de Markoff » dites MCMC, paraît être l'outil adéquat pour traiter ce problème.

Ce qui suit peut être considéré comme une illustration sur des problèmes d'ingénieur des concepts bayésiens déjà présentés dans la même revue par [LEC97].

Une des tâches importantes en hydrologie opérationnelle est d'estimer « les crues de projets » de rivières en vue d'assurer la protection des structures hydrauliques ou de collectivités riveraines. Ces crues sont généralement des événements beaucoup plus rares que ceux qui ont été systématiquement enregistrés. Ainsi l'estimation recherchée appartient à ce domaine de l'extrapolation des courbes de distribution et est donc sujette à un haut niveau d'incertitude. Les hydrologues statisticiens classiques ont consacré à ce thème, un très grand nombre de travaux portant notamment sur le choix des modèles, les erreurs d'échantillonnage et sur les moyens de réduire au moins partiellement les incertitudes ([RAS94]). Ils ont été toujours réticents pour introduire dans leurs estimations les connaissances a priori des experts au nom du « respect de l'objectivité scientifique ». Cependant de telles expertises peuvent améliorer de façon significative la capacité d'un modèle probabiliste à extrapoler vers les événements extrêmes. L'approche bayésienne offre des outils cohérents pour quantifier et contrôler les connaissances a priori des experts qui permettent ainsi de réduire l'incertitude de façon significative.

2. Intérêt des connaissances a priori pour l'analyse des valeurs extrêmes

Pour estimer les probabilités des événements rares, une des manières usuelles est donc d'extrapoler largement les distributions des événements fréquents comme le montre la **figure 1** de façon qualitative.

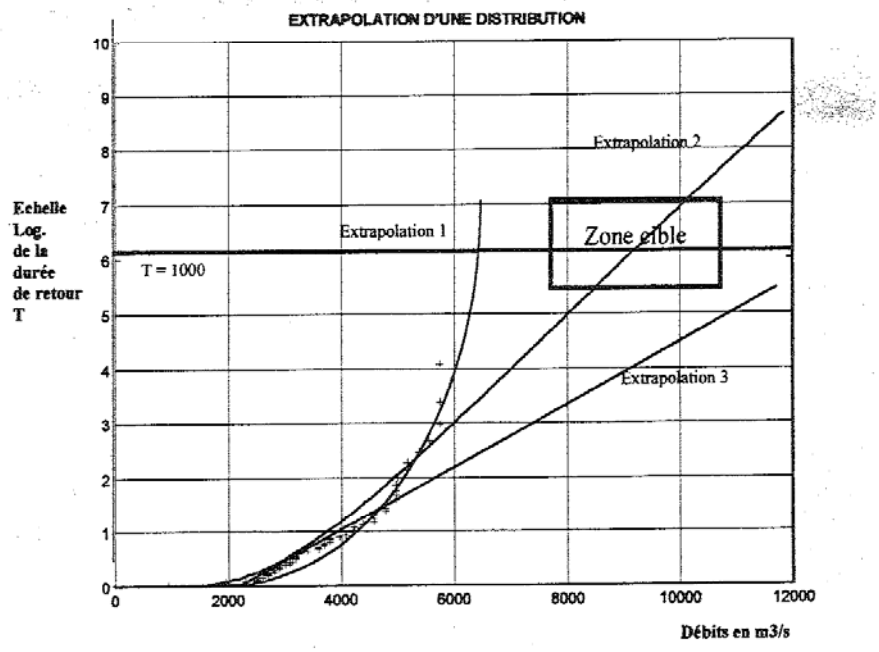


Fig. 1 : ***Les incertitudes de l'extrapolation des distributions***

La zone cible concerne la forme de la « queue » du modèle de distribution, en fait des probabilités d'ordre de grandeur 10^{-3} ou 10^{-4}, zone qui ne peut être atteinte que par extrapolation. Comme nous l'avons rappelé en introduction, les statisticiens de l'École Classique présentent des arguments de principe à l'encontre de l'utilisation des méthodes d'extrapolation statistique des distributions de valeurs extrêmes. En effet le concept classique de fréquence, à la base de l'application du calcul des probabilités, ne peut être raisonnablement justifié dans le cas de tels événements rares. Le concept subjectif de probabilité issu du paradigme bayésien ne souffre pas d'une telle limitation étant donné ses fondements en termes de paris décisionnels. Aussi abstrait ce concept puisse-t-il être, son

interprétation physique n'a rien à faire avec la nécessaire propriété de stationnarité à grande échelle de temps habituellement invoquée pour interpréter les probabilités des événements rares géophysiques. Cette stationnarité indispensable au paradigme de l'Ecole Classique est loin d'être une hypothèse réaliste.

En complément à l'aléa physique des phénomènes étudiés (incertitude naturelle), de nombreuses autres incertitudes peuvent entacher l'estimation des probabilités d'événements extrêmes et peuvent retentir sur les processus de décision :

- les incertitudes de mesure des données qui souvent interfèrent avec la limitation du nombre d'observations,

- les incertitudes de modélisation engendrées par l'incapacité du chercheur à choisir des hypothèses ou un modèle corrects sinon même réalistes. Particulièrement en hydrologie des extrêmes, de nombreux modèles différents ont été proposés ([RAS94].) sans justifications claires et complètement convaincantes pour la plupart,

- les incertitudes d'échantillonnage sur les paramètres, dues à la limitation de l'information disponible. Une littérature considérable a été consacrée à cette sorte d'incertitude ([FOR97]). Ces études ont été conduites principalement selon les principes et critères de la Statistique Classique mettant l'accent sur les critères « de biais et d'erreur quadratique moyenne » de quantiles ou autres estimateurs de paramètres.

Toute source d'information vaut la peine d'être introduite dans l'analyse pour réduire l'incertitude :

Prendre en compte ces incertitudes est une très importante tâche pour le chercheur s'il veut donner un fondement à son extrapolation. Il doit essayer de contrôler et réduire leurs effets sur les décisions subséquentes. La façon rationnelle est d'utiliser toute information disponible et pertinente quelle qu'elle soit.

Pour les événements extrêmes les séries de données systématiques utilisées (généralement les maxima annuels de débits des rivières) sont souvent très courtes et ne permettent pas des estimations fiables des risques induits par les événements rares. Heureusement ces séries, généralement considérées site par site, ne constituent pas les seules informations utiles.

Information systématique supplémentaire

La méthode la plus courante consiste à sélectionner les séries de valeurs maximales annuelles (dites AMS). Mais cette sélection, censée respecter certaines hypothèses d'indépendance inter annuelle, réduit indûment l'information significative pour le problème à une donnée par an. Au contraire, les méthodes POT (peak over threshold souvent appelées méthodes de renouvellement en France), sélectionnent tous les événements au-dessus d'un seuil et supposent l'indépendance et une même distribution pour ces valeurs. Le cas échéant une sélection complémentaire permet de ne garder que des épisodes hydrologiques différents ce qui assure l'indépendance.

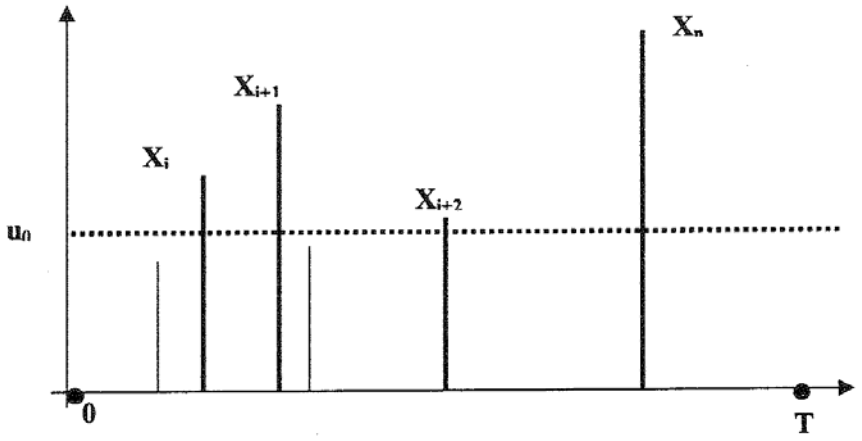


Fig. 2 : *Processus de dépassements schématique*

Les modèles POT ont été abondamment étudiés notamment par [PIC75] sur le plan théorique et notablement dans les applications hydrologiques (cf. [RAS94], pour une revue). Nous soulignons d'abord certains de leurs avantages théoriques. Considérons le modèle particulier décrit par l'hypothèse de Poisson du processus d'occurrence des dépassements d'un seuil u_0, chaque dépassement étant associé à des marques distribuées selon la loi de Pareto généralisée (voir ci après). Pickands a démontré que, sous des conditions générales, ce processus marqué

était la limite asymptotique du processus des valeurs extrêmes pour u_0 suffisamment grand. Il en résulte alors une cohérence asymptotique assurant une estimation rationnelle des paramètres du modèle comme les quantiles quel que soit le seuil. Cette cohérence réduit les erreurs d'estimation en général. Sans que, dans les applications concernées, ces justifications aient la force de certains résultats probabilistes pour d'autres problèmes où est censé s'appliquer le théorème de la limite centrale par exemple, le choix de ce modèle POT peut être vu comme un moyen de limiter les incertitudes de modélisation. Nous adopterons ce modèle dans la suite. Au-delà de ce choix et bien qu'elles n'aient pas encore été utilisées à ce propos pour le modèle POT général, les méthodes bayésiennes fournissent un cadre utile pour l'analyse des incertitudes de modélisation.

Information régionale

Aucun événement géophysique local (à un site), hydrométéorologique ou non, ne peut être séparé de son environnement régional. Les phénomènes générateurs de ces événements (tels que les averses et les perturbations atmosphériques) agissent généralement à grande échelle d'espace. De plus des ressemblances entre sites peuvent permettre des transpositions d'information d'un site à l'autre. Décrire ces ressemblances a été l'objectif de nombreux modèles ([GRE96]). On citera également le travail de [COL96a] dont l'intérêt est de régionaliser les modèles POT dans le domaine connexe des précipitations extrêmes. Ces travaux ressortent cependant de l'approche fréquentiste classique à base de modèles non linéaires généralisés à effets aléatoires. Du point de vue bayésien, de tels modèles appelés hiérarchiques, sont présentés dans les exemples de [GEL95].

Information historique

Dans les pays habités depuis longtemps, des données historiques sont souvent disponibles. Même si ces données sont éparses et quelque peu imprécises en comparaison avec les relevés systématiques plus récents, elles peuvent donner une information valable sur le comportement des distributions vers le domaine extrême. Ces données particulières peuvent être valorisées

au moyen de modèles spécifiques et apportent une réduction significative des intervalles de crédibilité des paramètres d'intérêt. Ce type d'approche avait déjà été développé en termes statistiques classiques par [STE86] qui l'ont étendu à l'utilisation de données paléohydrologiques (pouvant exister en deçà des périodes historiques). Cependant la conception fréquentiste de la probabilité reste particulièrement mal adaptée au traitement des données historiques pour les échantillons desquelles la répétabilité à l'identique justifiant le concept fréquentiste de risques d'erreurs n'a guère de sens. Nous avons entrepris le traitement des données historiques dans le cadre bayésien mieux fondé. Ces travaux, utilisant les techniques de calcul « d'augmentation des données » de [TAN87], ont été exposés aux journées de statistique bayésienne de la SFdS à l'institut Henri Poincaré et figurent aux compte rendus de ces journées (2000).

Information recueillie auprès d'experts du domaine.

Enfin on ne saurait négliger l'information fournie par les connaissances des experts du domaine (O' Hagan, 1998). Les « croyances d'expert » constituent peut être la source d'information la plus communément disponible mais aussi la plus fréquemment négligée parce que les études fréquentistes considèrent leur usage comme une faute contre l'objectivité scientifique. Dans le contexte bayésien cependant, l'utilisation rationnelle de l'expertise a priori subjective peut être faite et le présent article veut montrer comment on peut incorporer une telle information dans l'analyse statistique par quantification (on emploiera le franglais « élicitation » plus parlant) des paris a priori sur les paramètres d'un modèle POT.

Pourquoi l'expertise a priori devrait elle être négligée ?

L'analyse statistique fréquentiste est généralement recommandée comme la seule procédure objective permettant d'évaluer des probabilités d'événements. Ces probabilités sont considérées comme des quantités vraies appartenant au monde réel, objectif qui devrait être le seul objet d'étude d'un scientifique. En rapport avec cette attitude philosophique et eu égard

aux difficultés d'évaluation des probabilités a priori, l'approche bayésienne peut être considérée comme inutile et rejetée sur le plan des principes. A l'opposé, le paradigme bayésien suppose que tout objet conceptuel, modélisé, et toute probabilité est un tel objet, ne peut avoir qu'une signification subjective, c'est à dire celle d'une construction de l'esprit du chercheur. Par exemple l'incertitude sur un paramètre θ est représentée par un modèle de variable aléatoire et le principe classique d'objectivité est remplacé par le principe de cohérence entre les idées a priori sur θ et le jugement a posteriori conditionnel aux données et que supportent les probabilités a posteriori. Le « processeur d'information » entre jugements a priori et a posteriori est le théorème de Bayes fondé sur la fonction de vraisemblance du modèle phénoménologique et dont la **figure 3** illustre le principe.

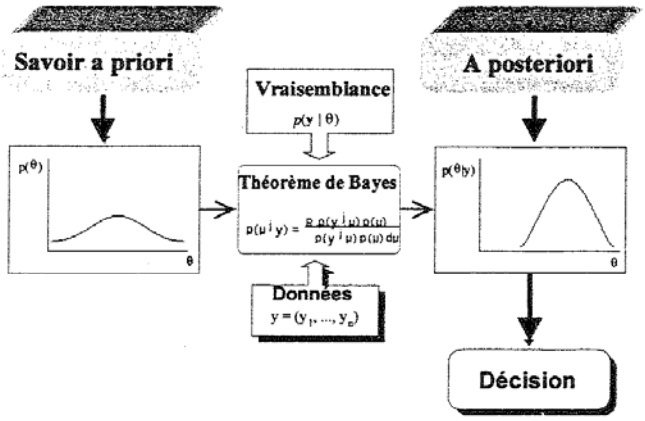


Fig.3 : ***Diagramme du raisonnement conditionnel bayésien.***

L'attention des statisticiens et des praticiens a été récemment renouvelée sur le problème de «l'élicitation des priors» c'est à dire la traduction en termes quantitatifs des opinions d'experts, souvent exprimées en termes qualitatifs, en prenant convenablement en compte leurs propres incertitudes. Des discussions approfondies sur le sujet ont été présentées dans «The Statistician» ([KAD98]). On notera que ces discussions ont maintenant abandonné les considérations philosophiques (en supposant scientifiquement admis l'usage des estimations a priori) pour se focaliser sur les aspects pratiques de « l'élicitation des priors ». Une contribution cruciale dans cette voie a été le travail de [COL96a] utilisant une expertise a priori pour estimer des quantiles de distributions des précipitations extrêmes.

Le processus d'élicitation peut se heurter à des difficultés. Les psychologues expérimentaux Tversky et Kahnemann ([PAL95]) ont montré que face à des événements incertains, tout individu, et même les experts scientifiquement entraînés, utilise des « heuristiques psychologiques » très éloignés des règles et concepts usuels du calcul des probabilités. Les écarts qu'ils baptisent « biais cognitifs » concernent les jugements d'experts et impliquent un manque de cohérence avec les estimations mathématiques et probabilistes. Ces difficultés n'empêchent pas l'élicitation mais doivent être gardées en mémoire pour une saine utilisation de la procédure. En premier lieu celle ci doit être clairement comprise et acceptée par l'expert de façon à fournir des résultats fiables. La quantification directe de paramètres statistiques doit être évitée quand ceux-ci n'ont pas un sens compréhensible direct. Pour un expert hydrologue, le concept de quantile associé à des événements annuels de durées de retour pas trop grandes (jusqu'à l'événement centennal par exemple) peut être aisément perçu comme le mentionne [OHA98]. Par contre les paramètres des modèles POT d'extrêmes sont non observables mais des paramètres de signification aisée comme des moyennes ou des quantiles peuvent être soumis aux experts pour élicitation avant d'en déduire les estimations de paramètres plus fondamentaux.

Le présent article illustre l'application d'une méthode d'élicitation des priors adaptée au modèle POT des extrêmes utilisé pour décrire les risques de crue de la Garonne à Agen.

3. Méthode

3.1. Le modèle POT des événements extrêmes ([PIC75])

La distribution des dépassements sur une période T

Considérons une séquence $X_1, X_2 \ldots$ de variables indépendantes et identiquement distribuées. Et soit un seuil u_0. Pickands a démontré que, sous des conditions très générales, le comportement limite pour u_0.grand des dépassements de la séquence sur la demi-droite ouverte $]u_0,+\infty[$ (voir la **figure 2**) est celui d'un processus de Poisson marqué dont les marques sont distribuées selon la loi de Pareto généralisée, c'est à dire :

$$\Pr(X \le x | X \ge u_0) = G(x|\rho,\beta,u_0) = \begin{cases} 1-(1-\beta(x-u_0)^{\frac{\rho}{\beta}} \text{ pour } \beta \neq 0 \\ 1-\exp(-\rho(x-u_0)) \text{ pour } \beta = 0 \end{cases}$$

$$\Pr(K(X \ge u_0) = k | \text{sur T années}) = \frac{(\mu T)^k \exp(-\mu T)}{k!} \qquad (1)$$

K est le nombre aléatoire de dépassements (ici sur T années)

L'important est l'indépendance des marques au-dessus du seuil. Dans la suite nous considérons le seuil assez haut pour que l'approximation asymptotique soit réaliste et adoptons les équations (1) comme un modèle POT pour les maxima au-dessus du seuil. Ce modèle n'est pas nouveau en hydrologie ([SMI84]). Les X_i sont les hauts débits successifs d'une rivière, qui sont dépendants à courte échelle de temps mais le modèle s'applique aux dépassements d'un seuil assez haut, dépassements interprétés comme des événements de crue indépendants.

Une conséquence immédiate de ce modèle est la distribution du maximum sur T années. On retrouve directement les trois formes limites de la loi des maxima d'un échantillon :

$$\Pr(\max(X_1, X_2, .. \le x | T \text{ années}, x \ge u_0) = \begin{cases} \exp((-\mu T)(1-\beta(x-u_0))^{\frac{\rho}{\beta}} \text{ pour } \beta \neq 0 \\ \exp((-\mu T)\exp(-\rho(x-u_0)) \text{ pour } \beta = 0 \end{cases} \qquad (2)$$

Cette distribution, tronquée inférieurement par u_0, est caractérisé par un vecteur θ de paramètres à 3 dimensions : $\theta = \{\mu, \beta, \rho\}$. Seul β peut être négatif.

Notons que μT est le nombre moyen de dépassements du seuil sur T années et $k = \beta / \rho$ est un paramètre hydrologique dont l'interprétation facilitera l'élicitation du prior.

La vraisemblance

Avec ce modèle il est aisé de calculer la vraisemblance de $\theta = \{\mu, \beta, \rho\}$.pour une série observée x de n dépassements indépendants $x_1, x_2, \ldots x_n$ du seuil pendant T années.

$$L(x;\mu,\beta,\rho)=\left[\frac{(\mu T)^n \exp(-\mu T)}{n!}\right]\left[\rho^n \exp((\rho-\beta)\right]S_n(x,\beta)$$
$$avec \quad S_n(x,\beta)=\frac{1}{\beta}\sum_{i=1}^{n}\log(1-\beta(x_i-u_0)) \qquad (3)$$

Le cas particulier $\beta = 0$ peut être obtenu par continuité et on a :

$$S_n(x,0)=-\sum_{i=1}^{n}(x_i-u_0)$$

Nous utiliserons $\theta = \{\mu, \beta, \rho\}$ comme la paramétrisation naturelle du modèle dont nous tirerons avantage de la structure semi-conjuguée (conjuguée naturelle pour β fixé) du prior correspondant (voir [ROB92]).

3.2. Le modèle de prior

Dans l'approche bayésienne, le choix du modèle de prior fait partie de la tâche du modélisateur. Nous supposerons que les croyances a priori sur θ sont représentées par la famille de densités :

$$\pi(\theta)=\frac{\lambda^\nu}{\Gamma(\nu)}\mu^{\nu-1}\exp(-\lambda\mu)\rho^{\gamma-1}\exp(-\varphi\rho)\pi_0(\beta) \qquad (4)$$

densité conjointe définie sur le domaine : $\mu > 0, \beta \in]-\infty,+\infty[, \rho > 0$.

En mots la densité marginale de β est $\pi_0(\beta)$; $\pi_0(\beta)$ est arbitraire et sa forme fonctionnelle doit permettre à l'expert d'encoder ses connaissances a priori. Conditionnellement à ces croyances sur β la densité de ρ est représentée par celle d'une distribution Gamma caractérisée par les « hyperparamètres φ, γ » qui seront fonction de β pour représenter une dépendance entre β et ρ. On suppose de plus que μ et le doublet (β,ρ) sont indépendants. Les croyances a priori sur μ sont représentées par la conjuguée naturelle du modèle de Poisson c'est à dire une distribution Gamma avec « hyper paramètres λ, ν » Ces choix sont fondés à la fois sur des argument pratiques (souplesse nécessaire pour représenter les croyances de l'expert) et théoriques (modèle partiellement conjugué, pour β fixé, adapté à la forme partiellement exponentielle de la vraisemblance).

3.3. Densité a posteriori de θ

La densité a posteriori conjointe du vecteur θ est donnée par l'application de la formule de Bayes en combinant les équations (3) et (4). Elle a la même structure conditionnelle que la densité a priori :

* μ reste indépendant du couple (β, ρ) et est distribué selon une Gamma de paramètres $(\nu + n, \lambda + T)$,

* conditionnellement à β, ρ est aussi distribué selon une Gamma de paramètres

$(\nu + n, \varphi - S_n(x, \beta))$,:

* la forme analytique de la densité marginale a posteriori de β est connue à une constante de normalisation près.

$$\pi(\beta|x) \propto [\frac{\Gamma(\gamma+n)}{\Gamma(\gamma)}\varphi^{\gamma+n}]\frac{\pi_0(\beta)\exp(-\beta S_n(x,\beta))}{(\varphi - S_n(x,\beta))^{\gamma+n}} \quad (5)$$

* β et ρ sont a posteriori dépendants comme le montre la loi a posteriori de β connaissant ρ :

$$\pi(\beta|\rho,x) \propto [\frac{\varphi^{\gamma}}{\Gamma(\gamma)}]\rho^{\gamma-1}\exp(-\varphi\rho)\exp((\rho-\beta)S_n(x,\beta))\pi_0(\beta) \quad (6)$$

Il est donc très facile de simuler des tirages Monte Carlo dans la loi a posteriori $\pi(\beta|x)$, notamment en discrétisant la distribution sur une grille de valeurs de β.

3.4. Elicitation a priori de la distribution de θ

Importance des paramètres hydrologiquement interprétables

Il importe que le processus d'élicitation soit bien compris par l'expert (c'est à dire un expert dans son propre domaine mais pas nécessairement en statistiques). L'élicitation directe des paramètres naturels du modèle POT précédent n'a pas de sens pour l'expert. Une méthode

pratique doit distinguer entre le modèle phénoménologique (Poisson– Pareto POT) d'un côté et les hypothèses plus simples (désignées ici par le terme « prior d'élicitation ») utilisées pour encoder les opinions de l'expert de l'autre côté. On distinguera donc plus précisément :

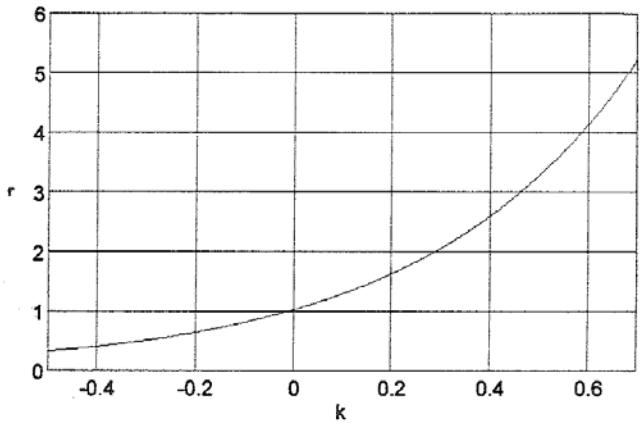
Les paramètres directement élicités : ce sont des quantités hydrologiquement significatives telles que des quantiles ou des valeurs moyennes. Ici nous utiliserons :

* l'espérance μ du nombre annuel de dépassements au-dessus du seuil,
* les différences de quantiles $q_2 = Q_{100} - Q_{10}$, $q_3 = Q_{1000} - Q_{100}$ entre les débits de crues survenant en moyenne tous les 10, 100 et 1000 ans,
* Conjointement à ces quantités, il faut noter que l'accroissement relatif des quantiles r :

$$r = \frac{q_3}{q_2}$$

est hydrologiquement interprétable et est relié fonctionnellement au paramètre $k = \beta / \rho$ du modèle POT.

Fig. 4 : *Relation fonctionnelle entre les paramètres r et k.*



$$r = \frac{(-\log(0.99))^k - (-\log(0.999))^k}{(-\log(0.9))^k - (-\log(0.99))^k}$$

Une valeur particulière intéressante est $r = 1$ correspondant à $k = 0$ qui caractérise le cas particulier de la distribution de Gumbel pour les valeurs maximales dans la famille POT

« Poisson – Pareto généralisée ». Quand $r > 1$, la distribution du maximum appartient au domaine de Weibull et les quantiles relatifs de débits de crue croissent de façon plus importante quand ils passent de l'événement centennal au millennal que lorsqu'ils passent du décennal au centennal.

Ces interprétations facilitent la tâche d'élicitation de l'expert.

Le modèle d'élicitation

Ce modèle contient toutes les hypothèses nécessaires pour encoder les connaissances de l'expert. Il n'est pas possible de demander à celui-ci son opinion sur chaque paramètre μ, β, ρ. Il est encore bien plus difficile de lui faire quantifier une forme de densité et les propriétés de dépendance a priori de ces paramètres. On ne peut demander à un expert d'évaluer ce qu'il ne connaît pas. Du point de vue statistique cependant de telles appréciations a priori sont nécessaires. En suivant [COL96a], nous supposerons que μ, q_2, q_3 sont indépendants en probabilité et distribués selon des lois Gamma. Notons qu'il existe des cas où les hypothèses constituant le modèle d'élicitation sont directement imposées par la structure conjuguée du modèle phénoménologique (exemple des modèles normaux).

La distribution a priori modélisée

Nous désignons ainsi le « prior » associé avec le modèle phénoménologique (POT), choisi ici pour des raisons théoriques et des facilités de calcul apportées par la structure semi-conjuguée de la distribution a priori.

L'utilisation d'un tel modèle a priori paramétrique comme le conjugué associé au modèle POT contraint nécessairement l'opinion de l'expert. Ainsi les quantiles Q_{10}, Q_{100}, Q_{1000} sont très fortement liés. Mais quel que soit le choix des priors, il est possible d'accepter une distance entre prior modélisé et prior élicité pourvu que les écarts calculés avec les évaluations directes de l'expert et les hypothèses nécessairement introduites soient jugés acceptables par celui ci.

3.5. Procédure d'encodage des connaissances d'expert dans le prior modélisé.

La procédure d'encodage des croyances de l'expert n'est pas directe. Nous en détaillons maintenant les étapes.

La première étape consiste à demander à l'expert de se livrer à 2 paris θ_p tels que :

$p = 0.5$ (c'est à dire avec 1 chance contre 1 que θ soit plus grand que la valeur donnée) et $p = 0.9$ (1 chance contre 9) pour chacun des paramètres μ, q_2, r. La table ci dessous montre les réponses de l'expert consulté pour notre cas d'étude (crues maximales de la Garonne près d'Agen avec un seuil u_0 = 2500 m3/s).

θ élicité	Médiane ($p = 0.5$)	Décile supérieur ($p = 0.9$)
μ	1,7	2,1
q_2	1000 m3/s	1600 m3/s
r	2	3.5

Ces évaluations semblent raisonnables sur le plan hydrologique et leur précision mesurée par l'écart : Décile – Médiane, indique une certaine confiance de l'expert au moins pour les paramètres μ et $\tilde{q}_2$. En nous référant à la **figure 4** la valeur supérieure 3,5 correspond à la valeur k = *0.5* couramment admise comme proche d'une borne supérieure dans de nombreuses études de crues.

Avec nos hypothèses d'élicitation, ces valeurs ont été encodées dans notre prior semi-conjugué en suivant les étapes successives suivantes.

* Supposant μ, q_2, q_3 indépendamment distribués selon des gamma de paramètres respectifs $(b_\mu, a_\mu)\,(b_2, a_2)\,(b_3, a_3)$ ces divers paramètres (b, a) sont déterminés de façon à satisfaire les contraintes du tableau ; on obtient :

θ élicité	a	b
μ	34,5	0.0498
q_2	6,5	162,08
q_3	100	20, 067

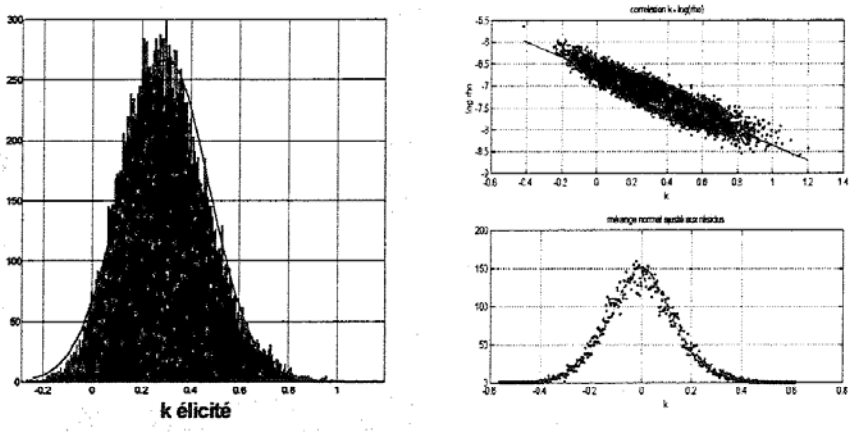
* On procède ensuite à une simulation de Monte Carlo classique de 20000 échantillons indépendants de μ, q_2, q_3 et r selon ces distributions gamma. Des estimations des distributions des valeurs des paramètres μ, β, ρ sont alors obtenues par transformations de variables à partir des grandeurs μ, k, log(ρ). Des calculs analytiques sont possibles à ce niveau mais la méthode de Monte Carlo apparaît plus souple et permet de dégager aisément les propriétés utiles des distributions a priori. C'est ainsi que la **figure 5** montre l'histogramme du paramètre k et son ajustement par la loi normale $N(m_k, \sigma_k)$ avec $m_k = 0.30$ et $\sigma_k = 0.18$.

De plus cette figure montre également que le modèle conditionnel de *log,* pour *k* fixé, est bien ajusté par une régression linéaire en *k* dont les résidus sont représentés par un mélange normal à 2 composantes :

$$\lambda = \log \rho = a + bk + \varepsilon$$
$$\varepsilon = (p)N(0,\sigma_1) + (1-p)N(0,\sigma_2)$$

avec : a = -6,7 b = -1,6
p = 0,5 $\sigma_1 = 0,14$ $\sigma_1 = 0,08$

Fig. 5 ***: Distribution de 20000 réplicats de k et des résidus de la régression de logρ***



* Adoptant ce modèle bidimensionnel pour k et log(ρ), on peut en déduire analytiquement la distribution du couple β,ρ.

En utilisant différents paris a priori on a vérifié que cette méthode semi-analytique, semi-Monte Carlo, était assez stable et robuste pour déterminer la distribution a priori du couple β,ρ et notamment la marginale $\pi_0(\beta)$ et la conditionnelle $\pi_0(\rho|\beta)$

* L'étape finale consiste à utiliser une gamma $(a(\beta), b(\beta))$ comme approximation de la conditionnelle élicitée $\pi_0(\rho|\beta)$, calculée à l'étape précédente. Pour cela on égale espérances et variances conditionnelles des deux distributions :

$$a_m(\beta) = \frac{(\tilde{E}_0(\rho|\beta))^2}{\tilde{V}_0(\rho|\beta)}$$

$$b_m(\beta) = \frac{\tilde{V}_0(\rho|\beta)}{\tilde{E}_0(\rho|\beta)}$$

Les $Gamma(b_\mu, a_\mu)$, marginale $\pi_0(\beta)$ et conditionnelle Gamma($a_m(\beta), b_m(\beta)$) sont finalement prises comme caractérisant le prior modélisé ayant la structure semi-conjuguée désirée.

Toutes ces distributions sont directement calculées sur une grille de valeurs de β prédéfinie.

4. Résultats dans l'étude des crues de la Garonne

65 ans (1913-1977) de données de crues de la Garonne près d'Agen (bassin versant de superficie égale à 52000km^2) ont été utilisées. Les « pointes de crues » supérieures à 2500 m3/s ont été sélectionnées soit 151 crues en 65 ans. Les compétences hydrologiques du second auteur ont été utilisées comme base de connaissances d'un expert. C'est pour nous l'occasion de mettre en garde le lecteur contre un argument classique de la logique « usuelle » de la règle de Bayes qui imposerait la complète indépendance du prior et des données. Cette indépendance théorique ne peut être qu'une fiction en pratique car tout expert, reconnu comme tel dans le domaine, a réfléchi sur les données objectives. L'important est d'éviter la circularité du raisonnement bayésien. Ici les hypothèses pratiques du processus d'élicitation ont été choisies aussi simples que possible et les paris a priori ont été fixés aussi honnêtement et raisonnablement que possible dans ce contexte. Toutefois au titre de l'analyse de sensibilité, plusieurs essais furent effectués et donnèrent des résultats cohérents. Bien entendu

ceci est un exercice théorique permettant de valider la méthode avant le dialogue prévu avec de « vrais experts ».

Validations du processus d'élicitation

La **figure 6** présente trois histogrammes et un diagramme de corrélation pour les 20000 simulations des paramètres POT obtenus avec le modèle d'élicitation (c'est à dire μ et les écarts de quantiles indépendamment distribués comme des lois gammas).

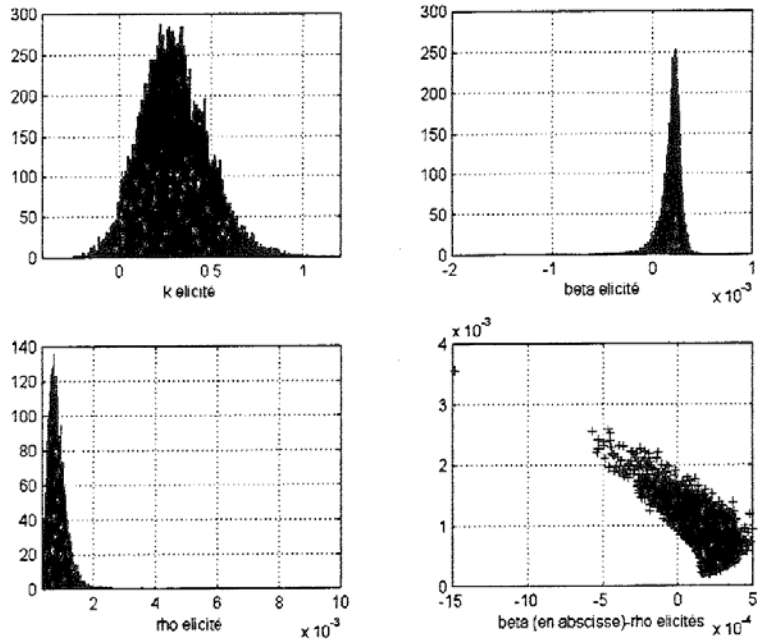


Fig.6 : ***Distributions de 20000 réplicats de k, β,ρ***

Pour valider ces calculs nous avons effectué 5000 simulations « en retour » à partir du prior semi-conjugué modélisé estimé. La figure parallèle à (5), reconstruite à partir de ce prior semi-conjugué, en est très proche qualitativement et n'est pas donnée ici.

Un comparaison synthétique est illustrée dans les tableaux suivants : donnant, en parallèle les résultats élicités (20000 réplicats) et validés (5000 réplicats) la médiane et la limite supérieure crédible à 90% pour les paramètres μ, q_2, q_3 et r d'un côté et les crues décennale, centennale et millénnale *Q10, Q100, Q1000* de l'autre côté.

Comparaison de $\mu, \tilde{q}_2(m3/s), \tilde{q}_3(\mathrm{m3/s})$ et r élicités et validés

	μ_5	μ_9	$q2_5$	$q2_9$	$q3_5$	$q3_9$	r_5	r_9
Elicité	1,69	2,10	1001	1609	2002	2265	2,00	3,51
Validé	1.71	2,11	987	1654	490	1349	2,02	3,30

Comparaison des quantiles élicités et validés

	$Q10_5$	$Q10_9$	$Q100_5$	$Q100_9$	$Q1000_5$	$Q1000_9$
Elicité	4987	5550	6062	6708	6596	7842
Validé	5001	5473	6028	6755	6515	7962

Ces tableaux montrent des résultats élicités et validés très comparables (différences relatives inférieures à 10%) à l'exception de **q3** dont les écarts sont importants.

Ce résultat concernant **q3** illustre une certaine incompatibilité entre les hypothèses analytiques élicitées : distributions a priori gamma et indépendance a priori des écarts de quantile d'un côté et le modèle semi conjugué associé au processus POT de Poisson-Pareto de l'autre côté. Toutefois les différences sont beaucoup plus faibles en termes de quantiles eux-mêmes et devraient être acceptables pour un expert indépendant.

5. Analyse a posteriori – Comparaison avec une distribution a priori non informative.

Pour l'analyse a posteriori du modèle POT de Poisson Pareto généralisée, nous avons développé un programme écrit en langage MATLAB et qui peut traiter toute distribution a priori « propre » aussi bien que « les priors impropres non informatifs » modélisant

l'ignorance a priori. Les techniques de calculs utilisées appartiennent à la boite à outils maintenant classique des méthodes de simulation de Monte Carlo par chaînes de Markoff (MCMC). Elles sont aujourd'hui largement mises en œuvre et ne sont pas décrites ici : le lecteur en recherche d'initiation à ces techniques est renvoyé à [ROB98]. Ce sont essentiellement ici les méthodes d'échantillonnage de Gibbs (facilitées par la structure semi-conjuguée du modèle phénoménologique) mais dont on utilise une version hybride intégrant une étape de Metropolis Hastings dans le cas du traitement des données historiques qui ne fait pas partie du présent article toutefois. Il est ici opportun de rappeler que ces techniques fournissent des évaluations qu'on peut qualifier d'exactes dans la mesure où les résultats du calcul bayésien ne sont pas tributaires de simplifications ou hypothèses ad hoc, d'approximations asymptotiques ou autres. Bien sûr les estimations peuvent être entachées d'approximations de calcul cependant contrôlables puisque celles ci sont liées à la limitation pratique des trajectoires des algorithmes markoviens utilisés. Ici 10000 réplicats de chaînes de Markoff ont été simulés dont un sur dix des 2000 derniers supposés indépendants des valeurs initiales ont été utilisés pour l'analyse statistique. Les **figures 7** et **8** ci-après présentent les valeurs de distributions (densités et quantiles) lissées soit par des méthodes d'estimations à noyaux normaux soit par les techniques dites de Rao-Blackwell ([GEL90]).

La **figure 7** présente les estimations de densités a posteriori des quantiles décennaux, centennaux et millennaux prenant en compte les observations, avec :

- le prior non informatif,
- le prior élicité, présenté précédemment.

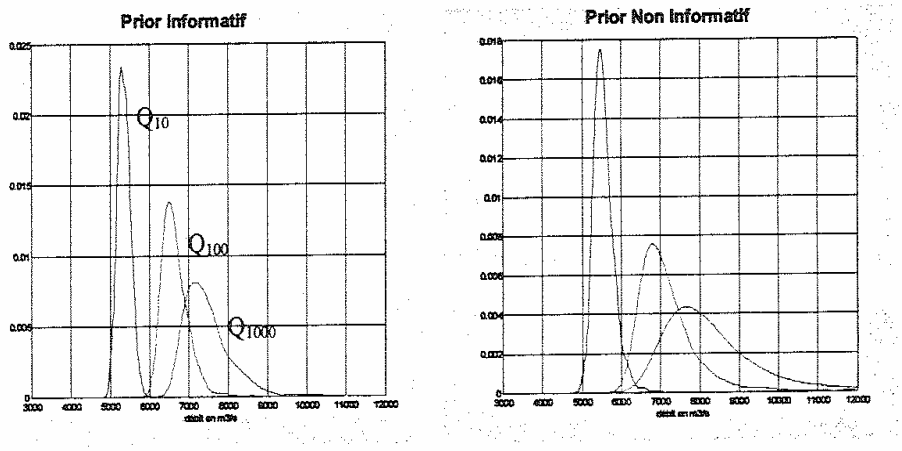


Fig.7 : ***Estimation MCMC des densités de quantiles avec deux priors différents***

La **figure 8** montre la moyenne a posteriori des quantiles Q_P ainsi que leurs limites de crédibilité à 0,05 et 0,95% pour une gamme de durées de retour T(Q) allant jusqu'à 10^3 ans:

$$T(Q) = \frac{1}{\Pr ob(X_{Max\ Annuel} \geq Q)}$$

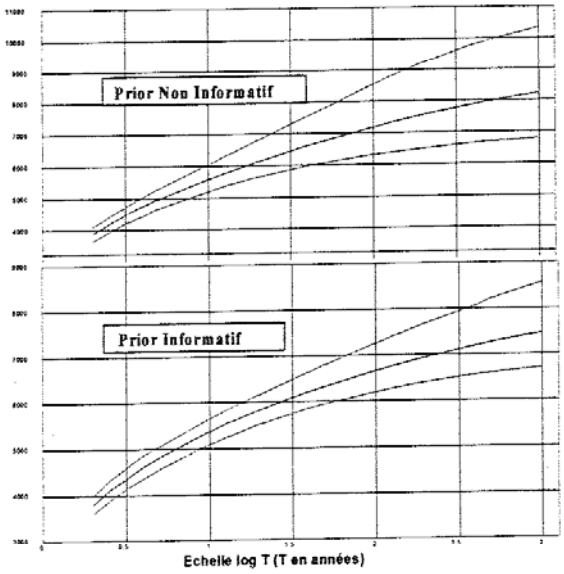


Fig.9 : ***Variation de la médiane et limites crédibles (5% et 95%) des quantiles en fonction de leurs durées de retour (en années)***

Sur toutes ces figures on voit clairement l'effet important de réduction des incertitudes apporté par la prise en compte des connaissances a priori de l'expert vis à vis du prior non informatif usuel. Plus quantitativement, il est utile de considérer la crue centennale Q100 par exemple.

Résultats a posteriori sur *Q100*

	Médiane a post.	.05 credible lim.	.95 credible lim.	Δ
Non inform.	7000	6290	8500	2210
Inform. prior	6590	6090	7240	1050

Δ est l'étendue de l'intervalle de crédibilité à 90% (contrepartie bayésienne de l'intervalle de confiance classique de signification différente toutefois, cf [LEC97].

Nous voyons que l'utilisation du prior de l'expert (quelque peu confiant dans ses évaluations) réduit l'étendue de quelques 50%. Dans un autre article (comptes rendus Journées SFdS stat.bayes. IHP 2000) des résultats analogues sont présentés avec cette fois un prior non informatif mais en utilisant, malgré leurs imprécisions dont les effets peuvent être pris en compte, 12 extrêmes historiques réellement observés sur la Garonne au cours de 141 années antérieures à 1913.

6. Discussion et conclusions

[COL96a] ont présenté un exercice d'élicitation semblable en compagnie d'un expert (le professeur D. Reed de l'Institute of Hydrology) dans le domaine voisin de la météorologie. Ils n'ont pas fait usage de la distinction entre prior d'élicitation et prior modélisé mais introduisirent directement le premier dans le calcul de la règle de Bayes. Conceptuellement cette méthode est plus aisée mais implique peut être des calculs plus complexes avec emploi systématique d'algorithmes hybrides : Metropolis-Hastings - échantillonnage de Gibbs. L'utilisation d'une distribution a priori semi conjuguée a quelques avantages :

Les premiers avantages concernent le calcul des distributions a posteriori conditionnelles complètes des paramètres et qui sont partiellement analytiquement connues. les techniques MCMC sont simples et faciles ici.

Aussi des développements importants des modèles ont été ou pourront être effectués de façon plus aisée :

- Introduction de données historiques (avec leurs propres imprécisions) et d'autres informations complémentaires en utilisant les techniques « d'augmentation des données » ([TAN87]) qui s'adaptent très bien à la génération séquentielle des chaînes de Markoff.

- Analyse prédictive des problèmes de décision (en matière de protection contre les extrêmes) prenant en compte l'ensemble des incertitudes, les risques induits et l'attitude des décideurs (aversion notamment) devant ces risques.

- Extension des modèles de risques à la prise en compte de la variabilité spatiale des phénomènes hydrologiques, etc....

On ne saurait oublier les avantages conceptuels de ce type d'approche. Avant tout ces problèmes peuvent être résolus en suivant un canevas rationnel unique éliminant toutes hypothèses approximatives sur le plan des calculs. Cependant l'utilisation des « modèles d'élicitation » soulève une question. Les hypothèses de Coles et Tawn, adoptées et adaptées par nous, ne sont pas uniques en tant que modèle d'élicitation. Quelle est la sensibilité des résultats à ces hypothèses ? L'introduction du prior modélisé montre que la question qui importe est la suivante :

Le prior modélisé semi conjugué est il assez souple pour prendre en compte n'importe quelle opinion quantifiée d'un expert ?

Deux modèles d'élicitation différents sont alors équivalents si leurs différents procédés pour encoder la même connaissance a priori conduisent à une « inférence semblable » et à des résultats opérationnels semblables via le modèle phénoménologique. L'exercice présenté ici veut démontrer la capacité du modèle de distribution a priori proposé à prendre en compte les connaissances de l'expert de façon acceptable en matière d'analyse de risques d'extrêmes hydrologiques. Cela est possible même si le prior de l'expert est particulièrement précis. Plusieurs autres essais nous ont montré la souplesse de la chaîne complète de traitement :

Expertise subjective a priori → *élicitation* → *calage du prior modélisé semi-conjugué* → *analyse a posteriori.*

En conclusion l'approche bayésienne permet de combiner rationnellement les données d'observation quantitatives et l'expertise professionnelle même qualitative et subjective. Dans le cas d'application de la Garonne les estimations et intervalles de crédibilité peuvent être notablement changés après introduction rationnelle de cette expertise.

En ce qui concerne les problèmes de risques extrêmes, le modèle phénoménologique classique POT/ Poisson – Pareto généralisé peut être couplé avec une méthode complète d'encodage d'une large gamme de connaissances a priori. Cette méthode est basée sur sa structure semi-conjuguée particulière. Cette approche peut être utilisée dans une grande variété de problèmes d'analyse de risques environnementaux ou industriels.

BIBLIOGRAPHIE

[COL96a] Coles S. G, Tawn J. A. (1996) : A Bayesian Analysis of Extreme Rainfall Data, Appl. Statist. 45 n°4.

[COL96b] Coles S. G, Powell E. A. (1996) : Bayesian Methods in Extreme Value Modelling, Intern. Statist. Rew. 64 n°1

[FOR97] Fortin V, Bernier J, Bobée B. (1997) : Simulation, Bayes and bootstrap in statistical hydrology ; Water Resources Research, vol.33, n°3

[GEL90] Gelfand, A. and Smith, A.F.M. (1990) : Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association}, 85,:398--409.

[GEL95] Gelman A, Carlin J B, Stern H S, Rubin D B : Bayesian Data Analysis, Chapman–Hall,1995.

[GRE96] GREHYS (1996) (Research Group in Statistical Hydrology) : Presentation and review of some methods for regional flood frequency analysis. - Journal of Hydrology : 186.

[KAD98] Kadane J. B, Wolson L.J., O'Hagan A., Craig et al. (1998). : Papers on 'Elicitation' and Discussions. The Statistician 47 n°1

[LEC97] Lecoutre B. (1997) : C'est bon à savoir ! Et si vous étiez un bayésien "qui s'ignore", La revue de Modulab n°18, INRIA (8)

[OHA98] O'Hagan (1998) : Elicitation expert beliefs in substancial practical applications - The Statistician 47 n°1

[PAL95] Palmerini M. P.(1995) : La réforme du jugement ou comment ne plus se tromper. O. Jacobs Paris.

[PIC75] Pickands J : (1975). Statistical Inference Using Extreme Order Statistics, Ann. Statist. 3.

[RAS94] Rasmussen P.F., Ashkar F., Rosbjerg D, Bobée B. (1994) : The POT method for flood estimation : a review. Stoch.Stat Methods in Hydrol Envir. Eng. (Hipel; Kluwer editors).

[ROB92] Robert C.P. (1992) : L'analyse statistique bayésienne; Economica, Paris.

[ROB98] Robert C.P., Casella G. (1998) : Monte Carlo Statistical Methods; Springer Verlag.

[SMI84] Smith R. L.(1984) : Threshold model for sample extreme in 'Statistical Extremes and Applications - de Oliveira - Heidel editors.

[STE86] Stedinger J R, Cohn T A (1986) : Flood frequency analysis with historical and paleoflood information. Water Resour. Res, 22(5).

[TAN87] Tanner M. A, Wong W. H. (1987) : The Calculation of Posterior Distributions by Data Augmentation, Amer. Statist. Ass. 82 n°398