

**DÉVELOPPEMENT DE L'ESTIMATEUR GLM-ML  
POUR LE MODÈLE LOG-NORMAL NON  
STATIONNAIRE ET APPLICATION À DES  
PRÉCIPITATIONS EXTRÊMES**

*Rapport de recherche No R-860*

*Mai 2006*

**DÉVELOPPEMENT DE L'ESTIMATEUR GLM-ML POUR LE  
MODÈLE LOG-NORMAL NON STATIONNAIRE ET APPLICATION  
À DES PRÉCIPITATIONS EXTRÊMES**

Par

I. Aissaoui-Fqayeh

S. El-Adlouni

T. B.M.J. Ouarda

A. St-Hilaire

Chaire en hydrologie statistique (Hydro-Québec / CRSNG)  
Chaire du Canada en estimation des variables hydrologiques  
INRS-ETE, Université du Québec  
490, rue de la Couronne (Québec), Canada G1K 9A9

**Rapport de Recherche N° R-860**

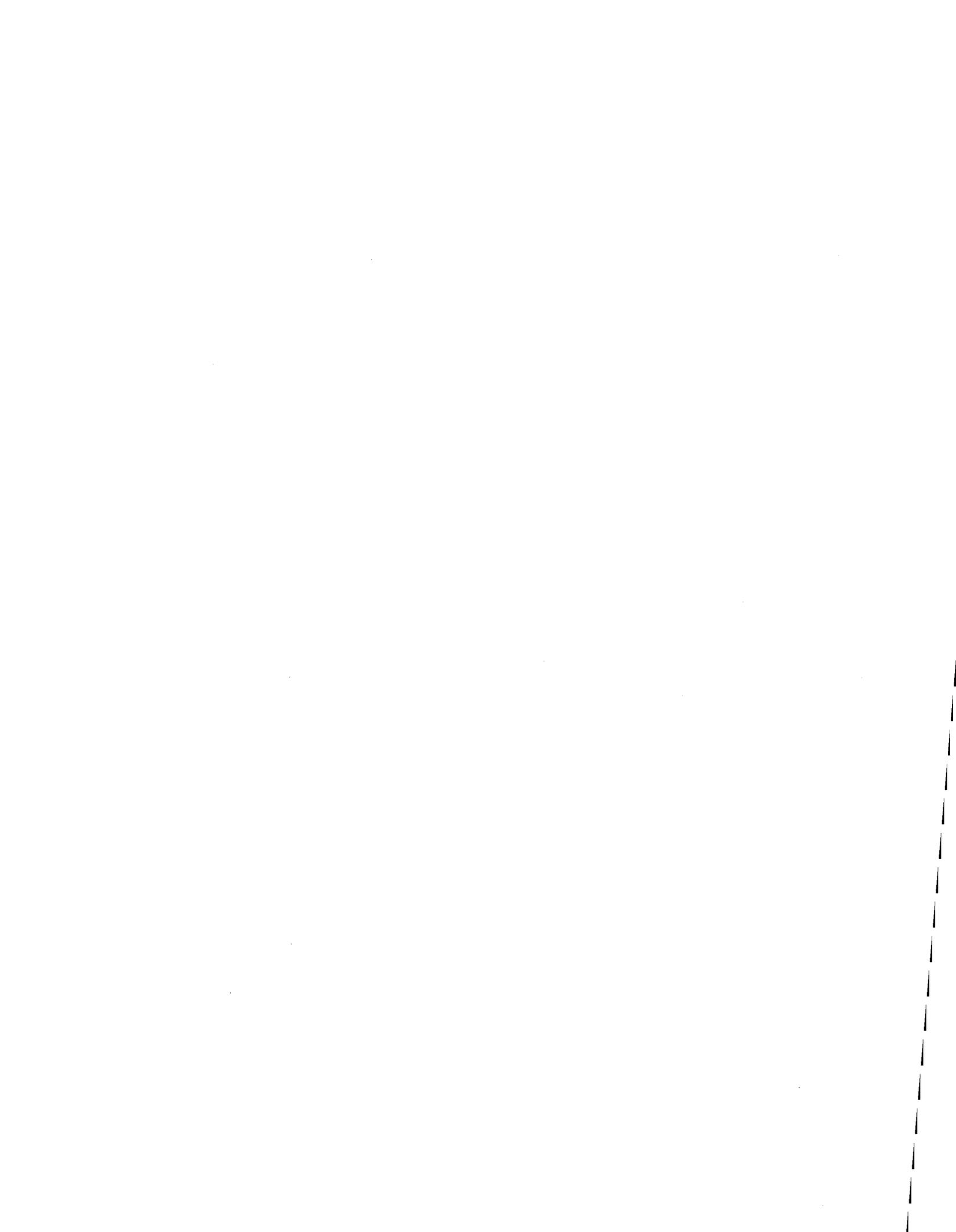
Mai 2006

ISBN : 2-89146-344-7

© INRS-ETE 2006

## TABLE DES MATIÈRES

1	Introduction .....	1
2	La méthode du maximum de vraisemblance pour le modèle Log-normal non-stationnaire .....	5
2.1	Le modèle Log-normal non-stationnaire .....	5
2.2	Estimation du maximum de vraisemblance pour le modèle LN non-stationnaire .....	9
3	Utilisation des GLM pour l'estimation des paramètres du modèle LN non-stationnaire .....	13
3.1	Les modèles linéaires généralisés (GLM) .....	13
3.2	Lien entre GLM et LN non-stationnaire.....	15
3.3	L'estimation des paramètres du modèle LN non-stationnaire par la méthode du maximum de vraisemblance pour les GLM .....	17
3.4	Estimation des quantiles .....	21
4	Propriétés des estimateurs de GLM-ML .....	23
4.1	Méthodologie de l'étude de simulation .....	23
4.2	Résultats .....	25
5	Choix du modèle .....	33
6	Application : Précipitation à la station Tehachapi.....	35
7	Conclusions .....	43
	Références .....	47
	Annexe A.....	51
	Annexe B.....	53
	Annexe C.....	55



## LISTE DES FIGURES

Figure 1. Situation géographique de la station Tehachapi de la Californie.....	35
Figure 2. Précipitations maximales annuelles observées et les valeurs des SOI correspondantes .....	36
Figure 3. Nuage de points des précipitations maximales annuelles observées et les valeurs des SOI correspondantes.....	37
Figure 4. La médiane Conditionnelle et ses intervalles de confiance estimés par les modèles $LN_{(0)}$ et $LN_{(1)}$ .....	40
Figure 5. La médiane Conditionnelle et ses intervalles de confiance estimés par les modèles $LN_{(0)}$ et $LN_{(2)}$ .....	41
Figure 6. Les estimateurs de la médiane Conditionnellement a la covariable SOI, pour les modèles $LN_{(0)}$ , $LN_{(1)}$ et $LN_{(2)}$ .....	42



## LISTE DES TABLEAUX

Tableau 1 : Paramètres $\mu$ et $\sigma$ correspondants aux différentes valeurs de la moyenne et du coefficient de variation( $C_v$ ). .....	23
Tableau 2 : RBIAIS des estimateurs GLM-ML et ML des quantiles de probabilité au non-dépassement 50%, 90%, et 99%, conditionnellement aux valeurs minimale, moyenne, et maximale de la covariable, pour le modèle $LN_{(1)}$ . .....	26
Tableau 3 : RRMSE des estimateurs GLM-ML et ML des quantiles de probabilité au non-dépassement 50%, 90%, et 99%, conditionnellement aux valeurs minimale, moyenne, et maximale de la covariable, pour le modèle $LN_{(1)}$ . .....	27
Tableau 4 : RBIAIS des estimateurs GLM-ML et ML des quantiles de probabilité au non-dépassement 50%, 90%, et 99%, conditionnellement aux valeurs minimale, moyenne, et maximale de la covariable, pour le modèle $LN_{(2)}$ . .....	29
Tableau 5 : RRMSE des estimateurs GLM-ML et ML des quantiles de probabilité au non-dépassement 50%, 90%, et 99%, conditionnellement aux valeurs minimale, moyenne, et maximale de la covariable, pour le modèle $LN_{(2)}$ . .....	30
Tableau 6 : Les estimateurs des paramètres par GLM-ML et les valeurs de log-vraisemblance correspondantes.....	38
Tableau 7. Les estimateurs des quantiles de probabilité au non-dépassement 50% conditionnellement aux valeurs min, mean, et max de SOI.....	39



# 1 INTRODUCTION

---

L'étude des processus et événements hydrologiques extrêmes à l'échelle locale et régionale passe par la mesure et l'analyse de différentes variables hydro-climatiques (précipitations, températures, débits, etc.). Cela nécessite des outils statistiques adaptés qui tiennent compte de l'interaction entre les différentes variables. Les approches classiques d'analyse fréquentielle supposent la stationnarité des séries d'observations, ainsi que l'indépendance et l'homogénéité. Autrement dit, les observations doivent être indépendantes et identiquement distribuées (i.i.d). En hydrologie, la loi de probabilité des crues peut changer avec le temps (existence de non-stationnarité) comme conséquence des activités humaines ou à cause des changements climatiques (El Adlouni et al., 2006, Zhang et al. (2001)). Il est donc nécessaire de développer des approches d'analyse fréquentielle qui tiennent compte de la non-stationnarité des séries des données hydro-climatiques. Ce genre de modèles permet d'inclure l'effet de différentes covariables sur la variabilité et l'évolution de la série observée.

Il est donc nécessaire de vérifier les hypothèses d'homogénéité et de stationnarité des séries observées. Plusieurs tests permettent de vérifier ces conditions. Dans ce travail on s'intéresse au problème de non-stationnarité. Pour tester cette dernière on utilise les tests de Mann, Kruskal-Wallis (Faucher et al., 1997) et Mann-Kendall (Önöz et Bayazit, 2003). Récemment de nouveaux tests ont été développés tel que le test de segmentation (Chen et Rao, 2002), et le test de Kwiatkowski-Phillips-Schmidt-Shin (le test KPSS) (Hobijn, 2004).

En analyse fréquentielle, plusieurs méthodes ont été développées pour l'estimation des paramètres des distributions d'intérêt. La méthode des moments de probabilités pondérées (PWM) et la méthode des L-moments (LM) (Hosking 1990) sont parmi les méthodes les plus utilisées, en raison de leur simplicité pour la majorité des distributions. Dans le cas non-stationnaire (existence d'une non-stationnarité temporelle ou de dépendance de covariables) les estimateurs des L-moments sont compliqués à implémenter d'une manière explicite. Dans ce cas, on considère souvent l'estimation par la méthode du maximum de vraisemblance (ML) qui permet, par le biais de la résolution numérique du système d'équations, d'estimer les paramètres même dans des cas très complexes (Coles, 2001). Le problème qui se pose avec la méthode du ML concerne l'efficacité des estimateurs dans le cas des échantillons de petites tailles et de leurs sensibilités au choix des points initiaux considérés dans la résolution numérique. La méthode du ML peut par conséquent conduire à des solutions physiquement inadmissibles dans le cas des séries hydrologiques. Comme solution à ce problème, El Adlouni et al., (2006) ont donné une extension de la méthode du maximum de vraisemblance généralisée (GML) introduite par Martins et Stedinger (2000) dans le cas classique, au cas de présence de covariables pour le modèle GEV non-stationnaire. La méthode GML permet d'intégrer une loi a priori sur le paramètre de forme de la loi GEV et de réduire l'espace des solutions de la méthode du ML à des valeurs admissibles dans le cas des séries hydrologiques.

La majorité des travaux sur les modèles non-stationnaires (ou en présence de covariables) sont liés au modèle généralisé des valeurs extrêmes (GEV) (Coles 2001 and Clarke 2002a,b). Katz et al. (2002), ont utilisé ce modèle pour traiter le cadre local ainsi que régional. En effet, la loi GEV est très utilisée en hydrologie et a été jugée dans plusieurs études comme adéquate pour l'ajustement des séries des maximums (Faucher et al. 1997). Cependant dans certaines

régions, d'autres distributions ont été suggérées pour l'étude des séries hydrométéorologiques : tel que la loi des valeurs extrêmes généralisée (GEV) en Grande Bretagne, Log-normale (LN) en Chine, Log-Pearson type III (LPIII) aux États-Unis d'Amérique (USWRC, 1982), ou la loi Log-normale à deux paramètres (LN) au Québec (Kouider, 2003).

Dans cette étude on présente le modèle Log-normal non-stationnaire à deux paramètres. Ce modèle, tel qu'il a été déjà mentionné, s'ajuste bien à plusieurs séries de données hydrologiques. En plus, il est important de disposer de modèles non-stationnaires de distributions à deux paramètres à cause du principe de parcimonie. En effet, le nombre des paramètres augmente avec le nombre de covariables d'où l'avantage de considérer une distribution avec le moins de paramètres possible, surtout en hydrologie où la taille des séries d'observations est souvent faible.

Le deuxième objectif du présent travail est de développer une approche, pour l'estimation des paramètres du modèle Log-normal non-stationnaire, qui combine les techniques utilisées dans les modèles linéaires généralisés et la méthode du maximum de vraisemblance. Les modèles linéaires généralisés (GLM) représentent une classe plus large qui a été étendue à partir des modèles linéaires classiques, pour combler leurs faiblesses au niveau de certaines hypothèses limitatives (Nelder et McCullagh 1983, Clarke 2002a,b). Yang et Al., (2005) ont utilisé le modèle linéaire généralisé (GLM) pour étudier des données des précipitations et pour incorporer la structure appropriée de la dépendance dans le cadre régional. Ils ont constaté que les GLM fournissent un environnement puissant et flexible pour explorer des rapports entre des variables climatiques.

Dans le présent travail, et grâce au lien démontré entre la distribution Log-normale et les modèles linéaires généralisés à travers la distribution normale, on obtient une estimation des paramètres ainsi que des quantiles du modèle Log-normal non-stationnaire. Cette estimation va être accomplie par la combinaison de la méthode du maximum de vraisemblance classique, et des formules analytiques obtenues par la méthode du maximum de vraisemblance pour les modèles linéaires généralisés.

Après avoir rappelé certaines propriétés statistiques de la loi Log-normale on présentera le modèle Log-normal non-stationnaire et certains des cas particuliers avec la méthode du maximum de vraisemblance pour l'estimation des paramètres. La contribution originale du présent travail, consiste à développer une nouvelle approche pour l'estimation des paramètres du modèle Log-normal non-stationnaire en établissant un lien avec les modèles linéaires généralisés. Dans la deuxième section, nous présentons le modèle Log-normal non-stationnaire et l'approche d'estimation des paramètres de ML. L'approche proposée qui combine entre GML et ML, est développée dans la section 3. Les propriétés de ces estimateurs (ML and GLM-ML) seront examinées et comparées, par la simulation de Monte Carlo, dans la section 4. Une application qui illustre la flexibilité et l'efficacité du modèle proposé et de la nouvelle approche pour l'estimation des paramètres sera présentée dans la section 6. Les conclusions et la discussion générale sont présentées dans la section 7.

## 2 LA MÉTHODE DU MAXIMUM DE VRAISEMBLANCE POUR LE MODÈLE LOG-NORMAL NON-STATIONNAIRE

---

### 2.1 Le modèle Log-normal non-stationnaire

On présente ci-dessous, quelques propriétés statistiques de la loi Log-normale (Leurent, 1998). Une variable  $X$  est distribuée suivant une loi Log-normale si son logarithme suit une loi normale. Soient  $\mu$  et  $\sigma$ , la moyenne et l'écart-type de la loi du logarithmes, qui est gaussienne. La variable  $Z = (\ln(X) - \mu) / \sigma$  est distribuée suivant une loi normale centrée réduite. Soit  $\Phi$  la fonction de répartition de la loi normale centrée réduite et  $F$  celle de la loi Log-normale. On a donc les égalités suivantes :

$$F(x_0) = pr(\{X \leq x_0\}) = pr(\{Z \leq z_0\}) = \Phi(z_0) = \Phi((\ln(x_0) - \mu) / \sigma) \quad (1)$$

$$F^{-1}(p) = \exp(\mu + \sigma \cdot \Phi^{-1}(p)) \quad (2)$$

La fonction de répartition peut être écrite sous la forme :

$$F(x_0) = \int_0^{x_0} \frac{1}{\sigma t \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{\log(t) - \mu}{\sigma}\right)^2\right) dt \quad (3)$$

La fonction de densité de probabilité  $f(x)$  est donnée par :

$$f(x) = \frac{1}{\sigma x \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{\log(x) - \mu}{\sigma}\right)^2\right), \quad (4)$$

où  $x > 0$ ,  $\sigma > 0$  et  $\mu$  est un réel.

Les moments d'ordre 1 et d'ordre 2 sont donnés par:

$$E[X] = \bar{X} = \exp(\mu + \sigma^2 / 2) \quad (5)$$

$$\begin{aligned} E[X^2] &= \exp(2\mu + 2\sigma^2) \\ &= (E[X])^2 \times \exp(\sigma^2) \end{aligned} \quad (6)$$

La variance s'écrit donc de la façon suivante:

$$\sigma_x^2 = \exp(2\mu + \sigma^2) \cdot (\exp(\sigma^2) - 1) = \bar{X}^2 \cdot (\exp(\sigma^2) - 1) \quad (7)$$

Le coefficient de variation (Cv) et le coefficient d'asymétrie (Cs) sont donnés par :

$$Cv = (e^{\sigma^2} - 1)^{\frac{1}{2}} \quad \text{et} \quad Cs = (e^{\sigma^2} - 1)^{\frac{1}{2}} (e^{\sigma^2} + 2).$$

La distribution normale est atteinte en prenant le logarithme naturel des données ( $\log(x)$ ). À mesure que  $x$  augmente, la valeur de  $f(x)$  diminue. Cependant, pour toutes les valeurs positives de  $x$ ,  $f(x)$  est toujours plus grande que zéro (Ricciardi et al., 2005).

On peut démontrer que pour une variable distribuée suivant une loi Log-normale la moyenne  $\xi$  et la variance  $\tau$  sont reliées par la relation:  $\tau = \xi^2 \times (\exp(\sigma^2) - 1)$ .

Après une transformation logarithmique, on obtient des variables aléatoires normalement distribuées de moyenne  $\mu$  et de variance  $\sigma^2$  données par les formules ci-dessous (annexe A):

$$\mu = \log(\xi) - \frac{1}{2} \log\left(1 + \frac{\tau}{\exp(2 \log(\xi))}\right),$$

$$\sigma^2 = \sqrt{\log\left(1 + \frac{\tau}{\exp(2 \log(\xi))}\right)} \quad (8)$$

Les modèles fréquentiels classiques peuvent être généralisés pour intégrer la notion de non-stationnarité dans le processus de modélisation en intégrant le temps ou d'autres covariables au niveau des paramètres. De telles covariables pourraient incorporer des tendances, des cycles, ou des variables physiques réelles (par exemple, des modèles qui tiennent compte à grande échelle de la circulation océanique et du couplage avec l'atmosphère tels que le phénomène EL-Nino, Coles 2001, Katz 2002). Par conséquent ces modèles peuvent être appliqués pour étudier l'effet des changements climatiques, ainsi que l'effet de la variabilité climatique sur la fréquence et l'amplitude des événements extrêmes.

Le modèle log-normal non-stationnaire correspond au processus  $X_t$ , tel que pour des valeurs fixes de temps ou/et de covariables il est distribué suivant une loi Log-normal. Ainsi les paramètres sont fonction de temps ou de covariables. Même si (dans le cas des covariables dépendantes du temps) le processus peut être stationnaire, nous employons le même nom de tous ces processus et seront dénotés « distribution LN non-stationnaire  $X \sim LN2(\mu_t, \sigma_t)$  ». Différents genres de dépendance peuvent être considérés en représentant des paramètres comme fonctions des covariables. Cependant, quand deux modèles donnent un bon ajustement, le modèle avec peu de paramètres sera préféré.

L'introduction des covariables peut être effectuée dans n'importe quel paramètre, même appliquée à deux ou trois paramètres à la fois. Sankarasubramanian et Lall (2003) ont

employé le modèle LN non-stationnaire pour faire la comparaison entre une approche non paramétrique (régression de quantile) et une approche semiparamétrique (Davison et Ramesh, 2000) en employant la vraisemblance locale. Deux cas ont été étudiés. Dans le premier cas le paramètre  $\mu$  est fonction linéaire des covariables. Dans le deuxième cas les deux paramètres sont des fonctions de la covariable, toutefois le coefficient de variation est constant. Beaucoup de scientifiques suggèrent d'associer la variation de la variance à la moyenne afin de garder le coefficient de variation constant. Cependant, dans le cas de variable distribuée suivant une loi Log-normale, le coefficient de variation ( $Cv$ ) est fonction seulement du paramètre  $\sigma$ . Ainsi, garder le coefficient  $Cv$  constant implique seulement que le paramètre  $\sigma$  est constant. Ainsi dans le présent travail nous considérons que seulement le paramètre  $\mu$  est fonction des covariables.

Pour étudier l'exactitude des estimations des paramètres, trois modèles sont considérés dans le travail actuel :

- $LN_{(0)}(\mu, \sigma)$ . qui est un modèle classique où tous les paramètres sont constants  $\mu_t = \mu, \sigma_t = \sigma$ .
- $LN_{(1)}(\mu_t = \beta_1 + \beta_2 Y_t, \sigma)$ . Le paramètre  $\mu$  est fonction linéaire d'une covariable  $Y_t$ .
- $LN_{(2)}(\mu_t = \beta_1 + \beta_2 Y_t + \beta_3 Y_t^2, \sigma)$ . Le paramètre  $\mu$  est une fonction quadratique d'une covariable temporelle  $Y_t$  et l'autre paramètre est constant.

Notons que, pour la distribution Log-normale, le changement du paramètre  $\mu$  mène à un changement de la moyenne et de la variance de la variable, ce qui correspond aux changements observés des séries hydro-météorologiques.

## 2.2 Estimation du maximum de vraisemblance pour le modèle LN non-stationnaire

Les paramètres du modèle Log-normal non-stationnaire peuvent être estimés par la méthode du maximum de vraisemblance, par maximisation du logarithme de la fonction de vraisemblance. La fonction de vraisemblance s'écrit:

$$\begin{aligned}
 L_n(\underline{x} | \mu_t, \sigma_t) &= \prod_{t=1}^n f(x_t; \mu_t, \sigma_t) \\
 &= \prod_{t=1}^n \frac{1}{\sigma_t x_t \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{\log x_t - \mu_t}{\sigma_t}\right)^2\right) \\
 &= \frac{1}{\prod_{t=1}^n \sigma_t \prod_{t=1}^n x_t (2\pi)^{n/2}} \prod_{t=1}^n \exp\left(-\frac{1}{2} \left(\frac{\log x_t - \mu_t}{\sigma_t}\right)^2\right);
 \end{aligned} \tag{9}$$

où  $n$  est le nombre des observations.

Posons  $\sigma_t = \sigma$ , puisque la non-stationnarité considérée dans le présent travail est liée au paramètre  $\mu$ . Dans ce cas la fonction de log-vraisemblance  $l_n$  est donnée par :

$$l_n(\underline{x}; \mu_t, \sigma) = (-n \log(\sigma) - \sum_{t=1}^n \log(x_t) - \frac{n}{2} \log(2\pi)) + \sum_{t=1}^n \left(-\frac{1}{2} \left(\frac{\log x_t - \mu_t}{\sigma}\right)^2\right) \tag{10}$$

Par définition, les estimateurs par maximum de vraisemblance sont les solutions du système d'équation des dérivées partielles de  $l_n$  par rapport à chacun des paramètres.

Pour le modèle  $LN_{(1)}$ , ou on a  $\mu_t = \beta_1 + \beta_2 Y_t$ , où  $Y_t$  est le vecteur des covariables, les estimateurs du maximum de vraisemblance des paramètres  $\beta_1, \beta_2$  et  $\sigma$  seront solutions du système d'équations suivant:

$$\begin{cases} \frac{1}{\sigma^2} \sum_{t=1}^n (\log x_t - \beta_1 - \beta_2 y_t) = 0 \\ \frac{1}{\sigma^2} \sum_{t=1}^n y_t (\log x_t - \beta_1 - \beta_2 y_t) = 0 \\ -\frac{n}{\sigma} + \left( \sum_{t=1}^n \frac{(\log x_t - \beta_1 - \beta_2 y_t)^2}{\sigma^3} \right) = 0 \end{cases} \quad (11)$$

Si on pose

$$\begin{cases} k_1 = \sum_{t=1}^n \log x_t \\ k_2 = \sum_{t=1}^n y_t \\ k_3 = \sum_{t=1}^n y_t \log x_t \\ k_4 = \sum_{t=1}^n y_t^2 \\ k_5 = \sum_{t=1}^n (\log x_t)^2 \end{cases} \quad (12)$$

Le système d'équations du maximum de vraisemblance est :

$$\begin{cases} \frac{1}{\sigma^2}(k_1 - n\beta_1 - k_2\beta_2) = 0 \\ \frac{1}{\sigma^2}(k_3 - k_2\beta_1 - k_4\beta_2) = 0 \\ -\frac{n}{\sigma} + \frac{1}{\sigma^3}(k_5 + n\beta_1^2 + k_4\beta_2^2 + 2k_2\beta_1\beta_2 - 2k_1\beta_1 - 2k_3\beta_2) = 0 \end{cases} \quad (13)$$

Ce système peut être résolu par une méthode numérique comme par exemple l'algorithme de Newton-Raphson. Cependant, l'efficacité et la vitesse de convergence (le nombre d'itérations nécessaires) vers les estimateurs du maximum de vraisemblance de cet algorithme dépendent du choix des valeurs initiales.

D'une manière analogue on obtient un système équivalent au système précédent pour le modèle  $LN_{(2)}$ , avec une quatrième équation correspondante au paramètre  $\beta_3$  (Annexe B).



### 3 UTILISATION DES GLM POUR L'ESTIMATION DES PARAMÈTRES DU MODÈLE LN NON-STATIONNAIRE

---

#### 3.1 Les modèles linéaires généralisés (GLM)

Bien que des modèles linéaires traditionnels soient employés intensivement dans l'analyse de données statistiques, il existe certaines situations pour lesquelles ils ne sont pas appropriés.

On rappelle la forme générale des modèles linéaires classiques :

$$X = \underline{Y}^T \beta + \varepsilon \quad (14)$$

avec

$$E(X) = \underline{Y}^T \beta \quad (15)$$

où  $X$  est un vecteur d'observations,  $\underline{Y}$  est une matrice des covariables, les  $\varepsilon = (\varepsilon_1 \dots \varepsilon_n)^T$  représentent les erreurs qui sont indépendantes et normalement distribuées, avec une moyenne zéro et une variance  $\sigma^2$  constante.  $\beta$  est le vecteur des paramètres du modèle.

Dobson (2002) a affirmé que les progrès dans la théorie statistique et les outils informatiques ont permis d'employer des méthodes analogues à celles développées pour les modèles linéaires. Un de ces derniers progrès a été l'identification de plusieurs propriétés de la distribution normale qui sont partagées par une classe plus large de distributions appelées la famille des distributions exponentielles dont la distribution de probabilité est de la forme :

$$f(x; \theta) = s(x)t(\theta)e^{a(x)b(\theta)} \quad (16)$$

où  $a$ ,  $b$ ,  $s$  et  $t$  sont des fonctions connues.

Notons que la symétrie entre  $x$  et  $\theta$  est soulignée si l'équation (1) est écrite sous la forme :

$$f(x; \theta) = \exp[a(x)b(\theta) + c(\theta) + d(x)] \quad (17)$$

où  $s(x) = \exp d(x)$  et  $t(\theta) = \exp c(\theta)$ . Si  $a(x) = x$ , la distribution serait en forme canonique et  $b(\theta)$  s'appelle le paramètre naturel de la distribution. S'il y a d'autres paramètres, en plus du paramètre d'intérêt  $\theta$ , ils sont considérés comme des paramètres de nuisances. Les paramètres de nuisances font partie des fonctions  $a$ ,  $b$ ,  $c$ , et  $d$ , et ils sont traités comme s'ils sont connus.

Le deuxième avantage des modèles linéaires généralisés sur les modèles classiques, est l'extension des méthodes numériques pour estimer les paramètres du modèle linéaire dans le cas où il y a une certaine fonction non linéaire reliant ( $E(X)$ ) à la composante linéaire ( $\underline{Y}^T \beta$ ), de la forme :

$$\eta = g(E[X]) = \underline{Y}^T \beta \quad (18)$$

La fonction  $g$  s'appelle la fonction de lien pour les GLM, l'utilisation de cette fonction donne une certaine flexibilité pour résoudre un large éventail de problèmes de modélisation statistique. On appelle donc un modèle linéaire généralisé tout modèle linéaire qui vérifie les trois propriétés suivantes (Nelder and McCullagh, 1983) :

1. Les variables de réponse sont distribuées selon la même distribution de famille exponentielle de forme canonique.
2. Il possède un vecteur de paramètres  $\beta$  et une variable explicative  $Y$  sous la forme matricielle suivante :

$$\underline{Y}^T = \begin{bmatrix} \underline{Y}_1^T \\ \vdots \\ \underline{Y}_N^T \end{bmatrix} = \begin{bmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & & \\ y_{N1} & & y_{Np} \end{bmatrix} \quad (19)$$

3. Il existe une fonction de lien différentiable et monotone  $g$  telle que:

$$g(E[X]) = \underline{Y}^T \beta$$

### 3.2 Lien entre GLM et LN non-stationnaire

Nous construisons un modèle linéaire généralisé en choisissant une variable de réponse et des variables explicatives pour nos données et en choisissant une fonction de lien appropriée et une distribution de probabilité de réponse. Les variables explicatives peuvent être n'importe quelle combinaison de variables continues, des variables de classification ou des interactions. Ceci va permettre d'établir un lien important entre les modèles linéaires généralisés et les modèles non-stationnaires. En effet, les variables explicatives correspondent aux covariables introduites dans les paramètres du modèle non-stationnaire, la fonction de lien est la même considérée dans le modèle non-stationnaire entre le paramètre  $\mu$  du modèle non-stationnaire et les covariables alors que la distribution des erreurs est celle considérée dans le modèle d'ajustement. Une approche semblable a été employée par Clarke (2002a) pour la distribution de Gumbel en choisissant une fonction appropriée de lien. La distribution log-normale appartient à la famille exponentielle. En effet, si  $X$  est une variable aléatoire telle que  $X \sim LN(\mu, \sigma^2)$ , alors la fonction de densité de probabilité peut être écrite comme suit :

$$f(x; \mu) = \exp \left[ -\frac{(\log(x))^2}{2\sigma^2} + \frac{\mu \log(x)}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2 x^2) \right] \quad (20)$$

ou  $\log(2\pi\sigma^2 x^2) = \log(2\pi\sigma^2) + \log(x^2)$ , et si on prend  $a(x) = \log(x)$ ,  $b(\mu) = \frac{\mu}{\sigma^2}$ ,

$$c(\mu) = -\frac{\mu^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2), \text{ et } d(x) = -\frac{(\log(x))^2}{2\sigma^2} - \log(x).$$

On constate qu'il s'agit d'une distribution de la famille exponentielle généralisée puisqu'elle vérifie (17).

Les expressions de l'espérance et la variance de la variable  $a(X)$  d'une distribution de la famille exponentielle sont données par les formules suivantes (Dobson, 2002) :

$$E[a(X)] = -c'(\theta) / b'(\theta) \quad (21)$$

$$\text{var}[a(X)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3} \quad (22)$$

D'après ces formules, on a donc dans un cadre non-stationnaire et pour une valeur fixée de la covariable ( $Y = y_i$ ):

$$E[\log(X) | Y = y_i] = - \left( \frac{\left( \frac{-\mu_i}{\sigma^2} \right)}{\frac{1}{\sigma^2}} \right) = \mu_i \quad ; \quad \eta_i = g(\mu_i)$$

$$\text{et } \text{Var}(\log X | Y = y_i) = \frac{0 \times \left( \frac{-\mu_i}{\sigma^2} \right) - \left( -\frac{1}{\sigma^2} \right) \times \frac{1}{\sigma^2}}{\left( \frac{1}{\sigma^2} \right)^3} = \sigma^2$$

En effet,  $c'(\mu) = -\frac{\mu}{\sigma^2}$ ,  $b'(\mu) = \frac{1}{\sigma^2}$ ,  $b''(\mu) = 0$  et  $c''(\mu) = -\frac{1}{\sigma^2}$ .

Donc si on a une suite de variables aléatoires qui proviennent d'une distribution Log-normale, les logarithmes de ces variables permettent de construire une suite de variables aléatoires normalement distribuées de moyenne  $\mu$  et de variance  $\sigma^2$ . Comme le modèle normal est un modèle linéaire généralisé, toutes les propriétés des modèles linéaires généralisés, en particulier les propriétés des modèles linéaires normaux, vont s'appliquer à cette suite des logarithmes. On a donc ces mêmes propriétés dans le cas de la loi Log-normale.

### 3.3 L'estimation des paramètres du modèle LN non-stationnaire par la méthode du maximum de vraisemblance pour les GLM

Soit  $(x_i)_{i=1, \dots, n}$ , un échantillon de taille  $n$ . On peut démontrer que les estimateurs des paramètres  $\beta$  (équation (18)) pour un modèle linéaire généralisé sont donnés par un algorithme itératif. À la  $m^{\text{ième}}$  itération les estimateurs  $b^{(m)}$  sont donnés par la formule suivante (Dobson, 2002):

$$\underline{Y}^T W \underline{Y} b^{(m)} = \underline{Y}^T W z \quad (23)$$

où  $\underline{Y}^T$  est la matrice des variables explicatives (matrice des covariables) formée par les colonnes  $\underline{Y}_i^T$  et  $W$  est une matrice carrée ( $n \times n$ ) diagonale définie par :

$$w_{ii} = \frac{1}{\text{var}(X_i)} \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2 \quad i=1, \dots, n \quad (24)$$

$z$  est un vecteur dont les éléments sont solution du système itératif suivant :

$$z_i = \sum_{k=1}^p y_{ik} b_k^{(m-1)} + (x_i - \mu_i) \left( \frac{\partial \mu_i}{\partial \eta_i} \right) \quad (25)$$

où  $p$  est le nombre de covariables. Les facteurs  $\mu_i$  et  $\frac{\partial \mu_i}{\partial \eta_i}$  sont évalués en  $b^{(m-1)}$ , et  $b^{(m)}$  est le

vecteur des estimateurs des paramètres  $\beta_1, \dots, \beta_p$  à la  $m^{\text{ième}}$  itération.

L'équation (23) à la même forme que l'équation normale des modèles linéaires obtenue par la méthode des moindres carrés, qui doit être résolue par une méthode numérique (pour plus de détails voir Dobson, 2002).

Pour une variable  $X$  distribuée suivant une loi Log-normale pour une valeur donnée  $y_i$  de la covariable  $Y$  ( $X|Y = y_i \square LN(\mu_i, \sigma^2)$ ), le logarithme de cette variable suit une loi normale ( $\log(X)|Y = y_i \square N(\mu_i, \sigma^2)$ ) et  $E[\log(X)|Y = y_i] = \mu_i = \beta_1 + \beta_2 y_i$  (cas du modèle  $LN_{(1)}$ ). Les deux paramètres  $\beta_1$  et  $\beta_2$  sont estimés par la méthode d'estimation du maximum de vraisemblance pour les modèles linéaires généralisés (On fait de même pour l'estimation des paramètres du modèle  $LN_{(2)}$ ).

Cette approche est plus simple, puisque les formules sont explicites.

La fonction de lien  $g$  pour ce spécial cas est la fonction identité car :

$$\eta_i = g(\mu_i) = Y_i' \beta = \mu_i \quad (26)$$

Ce qui implique que

$$\frac{\partial \mu_i}{\partial \eta_i} = 1 \quad (27)$$

D'après l'équation (24) on a

$$w_{ii} = \frac{1}{\text{var}(\log X | Y = y_i)} = \frac{1}{\sigma^2} \quad (28)$$

Si on pose  $b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$  l'estimateur de  $\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$ , l'équation (25) devient

$$z_i = b_1 + b_2 y_i + \log x_i - b_1 - b_2 y_i = \log x_i \quad (29)$$

Notons que  $z_i$  ne dépend pas de  $b$  et l'équation (23) peut être résolue directement en inversant

la matrice  $\underline{Y}^T W \underline{Y}$  :

ce qui donne

$$\underline{Y} = \begin{pmatrix} 1 & y_1 \\ 1 & y_2 \\ \vdots & \\ \vdots & \\ 1 & y_n \end{pmatrix} \quad \text{et} \quad \underline{Y}^T = \begin{pmatrix} 1 & \dots \dots 1 \\ y_1 & \dots y_n \end{pmatrix} \quad \text{et} \quad W = \begin{pmatrix} 1/\sigma^2 & \dots 0 \dots \dots 0 \\ 0 \\ \vdots \\ 0 & & & 1/\sigma^2 \end{pmatrix}$$

donc

$$\underline{Y}^T W \underline{Y} = \begin{pmatrix} n/\sigma^2 & 1/\sigma^2 \sum_{i=1}^n y_i \\ 1/\sigma^2 \sum_{i=1}^n y_i & 1/\sigma^2 \sum_{i=1}^n y_i^2 \end{pmatrix} \quad (30)$$

d'autre part on a

$$\begin{aligned}
 Y^T W_Z &= \begin{pmatrix} 1 & \dots & 1 \\ y_1 & \dots & y_n \end{pmatrix} \begin{pmatrix} 1/\sigma^2 \dots 0 \dots \dots 0 \\ 0 \\ \vdots \\ 0 & & 1/\sigma^2 \end{pmatrix} \begin{pmatrix} \log x_1 \\ \vdots \\ \vdots \\ \log x_n \end{pmatrix} \\
 &= \begin{pmatrix} 1/\sigma^2 \sum_{i=1}^n \log x_i \\ 1/\sigma^2 \sum_{i=1}^n y_i \log x_i \end{pmatrix}. \tag{31}
 \end{aligned}$$

Donc l'estimateur du maximum de vraisemblance (MLE) des paramètres  $\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$  est obtenu par

inversion de la matrice  $\underline{Y}^T W \underline{Y}$  :

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = (\underline{Y}^T W \underline{Y})^{-1} \begin{bmatrix} 1/\sigma^2 \sum_{i=1}^n \log x_i \\ 1/\sigma^2 \sum_{i=1}^n y_i \log x_i \end{bmatrix} \tag{32}$$

Cette approche est équivalente à celle présentée à la section (2.2) pour le modèle Log-normal non-stationnaire. L'estimation des paramètres par la méthode du maximum de vraisemblance correspond à la résolution d'un système d'équations des dérivées de la fonction de log-vraisemblance par rapport à chacun des paramètres. L'estimation des paramètres dans le cas des modèles linéaires généralisés permet d'éviter les problèmes de convergence des méthodes numériques ainsi que la sensibilité de la solution aux valeurs initiales dans le cas d'un modèle plus compliqué (cas quadratique ou cas où on a plus d'un seul covariable).

Cependant, pour les modèles  $LN_{(1)}(\mu_i = \beta_1 + \beta_2 Y_i, \sigma)$  (resp.  $LN_{(2)}(\mu_i = \beta_1 + \beta_2 Y_i + \beta_3 Y_i^2, \sigma)$ ), la méthode d'estimation utilisée pour les modèles linéaires généralisés n'est pas suffisante, car elle ne nous donne que l'estimation des paramètres  $\beta_1$  et  $\beta_2$  (resp.  $\beta_1, \beta_2$  et  $\beta_3$ ). C'est pourquoi on suggère une approche qui combine les deux méthodes d'estimation, celle du maximum de vraisemblance classique, et celle utilisée pour les modèles linéaires généralisés pour pouvoir estimer le paramètre  $\sigma$ . Notons cette méthode la méthode de GLM-ML pour le modèle LN non-stationnaire. On considère donc les estimateurs obtenus par les modèles linéaires généralisés pour  $\beta_1$  et  $\beta_2$  (resp.  $\beta_1, \beta_2$  et  $\beta_3$ ) comme point de départ de l'estimation du maximum de vraisemblance du modèle Log-normal non-stationnaire. Une résolution numérique, par l'algorithme de Newton-Raphson permet donc d'obtenir les estimateurs de tous les paramètres dans l'équation (13) ( $LN_{(1)}$ ) (un système équivalent pour le  $LN_{(2)}$  sera donné dans l'annexe B).

Les propriétés statistiques de ces estimateurs seront présentées à la section 4.

### 3.4 Estimation des quantiles

Dans les itérations de la méthode du maximum de vraisemblance, on prend comme conditions initiales des paramètres de la moyenne  $\mu$  (sous forme linéaire ou quadratique), les estimateurs obtenus de ces paramètres par la méthode du maximum de vraisemblance des modèles linéaires généralisés. On obtient donc l'estimateur de la variance  $\sigma$  d'une façon sûre et après une seule itération seulement.

Les quantiles de probabilité au non-dépassement  $p$  par rapport à une valeur observée  $y_0$  de la covariable  $Y$ , qui correspondent au vecteur des paramètres  $(\mu_{y_0}, \sigma)$ , sont données par la formule suivante démontrée en Annexe C :

$$x_{p,y_0} = \exp(\mu_{y_0} + \sigma \Phi^{-1}(p)). \quad (33)$$

Où  $\mu_{y_0}$  est le paramètre de position du modèle Log-normal non-stationnaire conditionnellement à la valeur particulière  $y_0$  de la covariable. Ce paramètre change selon les différents modèles, pour le modèle  $LN_{(0)}$ ,  $\mu_{y_0} = \mu$ , pour le modèle  $LN_{(1)}$  il prend la forme :  $\mu_{y_0} = \beta_1 + \beta_2 y_0$ , et finalement pour le modèle  $LN_{(2)}$  il prend la forme  $\mu_{y_0} = \beta_1 + \beta_2 y_0 + \beta_3 y_0^2$ .

En remplaçant ces paramètres par leurs estimateurs, on obtient les estimateurs des quantiles.

## 4 PROPRIÉTÉS DES ESTIMATEURS DE GLM-ML

---

### 4.1 Méthodologie de l'étude de simulation

Dans cette partie on évalue par simulation la méthode d'estimation (GLM-ML) qui combine les deux approches. Le choix des paramètres des modèles simulés est effectué sur la base de la moyenne et du coefficient de variation de 40 séries hydrologiques du Québec étudiées par Kouider et al. (2003). Les résultats obtenus à partir de cette étude permettent d'obtenir les intervalles de variation des deux paramètres de la loi LN. Le tableau (1) ci-dessous représente neuf cas tirés de la combinaison de trois valeurs particulières de la moyenne et du coefficient de variation ainsi que les paramètres de la loi LN correspondant. Ces neuf couples de paramètres couvrent toute la gamme de la moyenne et du coefficient de variation  $C_v$  observés pour les 40 stations étudiées.

**Tableau 1 : Paramètres  $\mu$  et  $\sigma$  correspondants aux différentes valeurs de la moyenne et du coefficient de variation ( $C_v$ ).**

<i>Moyenne</i> / $C_v$	<b>0.1</b>	<b>0.2</b>	<b>0.3</b>
<b>50</b>	$\sigma=0.098$ $\mu=1.694$	$\sigma=0.193$ $\mu=1.68$	$\sigma=0.283$ $\mu=1.659$
<b>500</b>	$\sigma=0.098$ $\mu=2.694$	$\sigma=0.193$ $\mu=2.68$	$\sigma=0.283$ $\mu=2.659$
<b>1500</b>	$\sigma=0.098$ $\mu=3.171$	$\sigma=0.193$ $\mu=3.157$	$\sigma=0.283$ $\mu=3.136$

Les valeurs de l'asymétrie, pour une variable distribuée selon une loi LN, peuvent être déduites de celles du coefficient de variation à partir de la formule suivante :

$$C_s = C_v(C_v^2 + 3) \quad (34)$$

Les valeurs du coefficient d'asymétrie pour le cas étudié sont 0.3, 0.6 et 0.9 et ils correspondent aux valeurs respectives 0.1, 0.2 et 0.3 du coefficient de variation.

Pour les deux modèles  $LN_{(1)}$  et  $LN_{(2)}$ , et pour chaque valeur du  $C_v$  et de la moyenne tirée du tableau 1, on génère  $M=5000$  échantillons d'une loi Log-normale de taille  $n$ , (on choisi une taille  $n=50$  à titre d'exemple qui représente en moyenne les tailles des séries rencontrées en hydrologie). Les valeurs théoriques des paramètres  $\mu$  et  $\sigma$  pour le modèle  $LN_{(0)}$  ( $\beta_1$  et  $\sigma$  pour les models  $LN_{(1)}$  et  $LN_{(2)}$ ) sont calculées à partir de celles de la moyenne et du coefficient de variation correspondant. Les paramètres de la tendance dans les deux modèles sont convenablement choisis de telle sorte que la tendance soit faible et ressemble aux cas réellement observés en hydrologie. Nous avons considéré les paramètres suivants pour le cas linéaire ( $LN_{(1)}$ )  $\beta_2 = -1$ , et pour le cas quadratique ( $LN_{(2)}$ )  $\beta_2 = -0.2$ ,  $\beta_3 = 0.1$ . La covariable  $Y_i$  est générée à partir d'une loi normale de moyenne 0 et de variance 1. Les quantiles sont estimés en fonction des valeurs particulières de la covariable (les valeurs : minimale (min), maximale (max) et moyenne (moy)). Pour comparer les deux méthodes GLM-ML et ML, on évalue les quantiles de probabilité au non-dépassement 50%, 90% et 99% pour en déduire le biais relatif (RBIAIS) et la racine de l'erreur quadratique moyenne relative (RRMSE). RBIAIS et RRMSE sont donnés par les formules suivantes (Kim et al., 2002) :

$$RBIAIS = \frac{1}{M} \sum_{i=1}^M \frac{|\hat{Q}_{i,p} - Q_p|}{Q_p} \quad (35)$$

$$\text{RRMSE} = \left[ \frac{1}{M} \sum_{i=1}^M \left( \frac{\hat{Q}_{i,p} - Q_p}{Q_p} \right)^2 \right]^{1/2} \quad (36)$$

Où  $\hat{Q}_{i,p}$  et  $Q_p$  sont respectivement les estimateurs des quantiles et les quantiles théoriques de probabilité au non-dépassement  $p$  et  $M$  est le nombre de simulations.

## 4.2 Résultats

La comparaison entre les deux méthodes ML et GLM-ML était faite pour les 9 cas présentés dans le tableau 1. Les résultats montrent que pour les cas où le coefficient de variation ne change pas les valeurs de RBIAIS et RRMSE restent presque les mêmes. Le tableau 2 (respectivement tableau 3) représente le RBIAIS (respectivement le RRMSE) des estimateurs des quantiles obtenus par la méthode de GLM-ML et la méthode de ML pour le modèle  $LN_{(1)}$ . Tous les quantiles sont calculés de la distribution conditionnelle de la variable  $X$  sachant  $Y = y_0$  ( $y_0$  prend trois valeurs différentes : minimale, moyenne et maximale).

**Tableau 2 : RBIAIS des estimateurs GLM-ML et ML des quantiles de probabilité au non-dépassement 50%, 90%, et 99%, conditionnellement aux valeurs minimale, moyenne, et maximale de la covariable, pour le modèle  $LN_{(1)}$ .**

Modèle $LN_{(1)}$	Valeur de la covariable	RBIAIS % (GLM-ML)			RBIAIS % (ML)		
		50%	90%	99%	50%	90%	99%
$\sigma = 0,098$ $\mu = 1,694$	Min	2,74	2,95	3,29	2,80	2,90	3,29
	Moy	1,11	1,48	2,13	1,11	1,52	2,15
	Max	2,74	2,96	3,32	2,75	2,93	3,37
$\sigma = 0,098$ $\mu = 2,694$	Min	2,74	2,92	3,30	2,74	2,89	3,32
	Moy	1,09	1,49	2,16	1,10	1,46	2,15
	Max	2,79	2,92	3,39	2,72	2,85	3,26
$\sigma = 0,098$ $\mu = 3,171$	Min	2,79	2,98	3,30	2,79	2,96	3,33
	Moy	1,10	1,50	2,17	1,10	1,50	2,18
	Max	2,74	2,95	3,35	2,78	2,96	3,37
$\sigma = 0,193$ $\mu = 1,68$	Min	5,53	5,77	6,62	5,38	5,79	6,43
	Moy	2,15	2,92	4,30	2,19	2,99	4,19
	Max	5,54	5,78	6,56	5,36	5,78	6,58
$\sigma = 0,193$ $\mu = 2,68$	Min	5,43	5,85	6,50	5,38	5,76	6,49
	Moy	2,17	2,91	4,23	2,18	2,95	4,26
	Max	5,45	5,70	6,60	5,46	5,75	6,50
$\sigma = 0,193$ $\mu = 3,157$	Min	5,38	5,78	6,33	5,53	5,73	6,53
	Moy	2,16	2,97	4,24	2,19	2,99	4,28
	Max	5,49	5,78	6,50	5,47	5,74	6,69
$\sigma = 0,283$ $\mu = 1,659$	Min	7,96	8,43	9,53	7,91	8,47	9,49
	Moy	3,19	4,38	6,15	3,21	4,27	6,13
	Max	7,96	8,57	9,60	8,02	8,54	9,29
$\sigma = 0,283$ $\mu = 2,659$	Min	7,99	8,47	9,59	8,05	8,35	9,83
	Moy	3,15	4,29	6,21	3,21	4,37	6,28
	Max	7,97	8,36	9,59	7,97	8,46	9,51
$\sigma = 0,283$ $\mu = 3,136$	Min	7,96	8,39	9,56	7,90	8,58	9,35
	Moy	3,19	4,33	6,23	3,16	4,34	6,12
	Max	7,88	8,48	9,53	7,91	8,63	9,47

Tableau 3 : RRMSE des estimateurs GLM-ML et ML des quantiles de probabilité au non-dépassement 50%, 90%, et 99%, conditionnellement aux valeurs minimale, moyenne, et maximale de la covariable, pour le modèle  $LN_{(1)}$ .

Modèle $LN_{(1)}$	Valeur de la covariable	RRMSE % (GLM-ML)			RRMSE %(ML)		
		50%	90%	99%	50%	90%	99%
$\sigma = 0,098$ $\mu = 1,694$	Min	3,46	3,75	4,13	3,51	3,67	4,13
	Moy	1,39	1,85	2,65	1,38	1,90	2,69
	Max	3,51	3,75	4,16	3,51	3,68	4,21
$\sigma = 0,098$ $\mu = 2,694$	Min	3,46	3,70	4,14	3,48	3,66	4,17
	Moy	1,38	1,86	2,66	1,39	1,84	2,68
	Max	3,54	3,70	4,25	3,45	3,61	4,10
$\sigma = 0,098$ $\mu = 3,171$	Min	3,56	3,76	4,15	3,52	3,72	4,19
	Moy	1,39	1,87	2,71	1,38	1,88	2,72
	Max	3,48	3,70	4,19	3,51	3,70	4,21
$\sigma = 0,193$ $\mu = 1,68$	Min	7,04	7,30	8,36	6,83	7,31	8,08
	Moy	2,71	3,46	5,32	2,73	3,73	5,23
	Max	7,02	7,28	8,19	6,81	7,29	8,22
$\sigma = 0,193$ $\mu = 2,68$	Min	6,91	7,36	8,19	6,82	7,28	8,14
	Moy	2,72	3,66	5,23	2,76	3,69	5,30
	Max	6,92	7,21	8,32	6,97	7,30	8,18
$\sigma = 0,193$ $\mu = 3,157$	Min	6,86	7,28	7,97	6,98	7,20	8,19
	Moy	2,71	3,72	5,25	2,74	3,74	5,33
	Max	6,95	7,29	8,15	7,03	7,31	8,36
$\sigma = 0,283$ $\mu = 1,659$	Min	10,15	10,76	11,96	10,08	10,70	11,95
	Moy	4,02	5,48	7,66	4,03	5,37	7,68
	Max	10,16	10,86	12,10	10,23	10,82	11,79
$\sigma = 0,283$ $\mu = 2,659$	Min	10,22	10,69	12,09	10,25	10,57	12,35
	Moy	3,96	5,39	7,72	4,02	5,48	7,85
	Max	10,10	10,66	12,11	10,19	10,68	12,10
$\sigma = 0,283$ $\mu = 3,136$	Min	10,10	10,75	11,95	10,07	10,80	11,76
	Moy	4,00	5,41	7,76	3,95	5,39	7,60
	Max	10,08	10,69	12,09	10,00	10,94	11,87

Les résultats montrent que pour le modèle  $LN_{(1)}$  (Tableau 2 et 3), les estimateurs de GLM-ML et ceux de ML ont le même RBIAIS et RRMSE pour tous les cas. Conditionnellement à la valeur centrale de la covariable, on remarque que le RBIAIS et le RRMSE sont plus petits que ceux obtenus pour les valeurs extrêmes de SOI, et le RBIAIS ainsi que le RRMSE croissent avec le coefficient de variation  $C_v$ . Le tableau 4 (respectivement le tableau 5) donnent les mêmes résultats qu'ils étaient présentés dans les tableaux 2 et 3 pour le modèle  $LN_{(1)}$ , mais cette fois pour le modèle  $LN_{(2)}$ .

Tableau 4 : RBIAIS des estimateurs GLM-ML et ML des quantiles de probabilité au non-dépassement 50%, 90%, et 99%, conditionnellement aux valeurs minimale, moyenne, et maximale de la covariable, pour le modèle  $LN_{(2)}$ .

Modèle $LN_{(2)}$	Valeur de la covariable	RBIAIS %(GLM-ML)			RBIAIS % (ML)		
		50%	90%	99%	50%	90%	99%
$\sigma = 0,098$ $\mu = 1,694$	Min	4,59	4,57	4,86	19,62	17,61	15,75
	Moy	1,37	1,80	2,37	5,52	8,30	10,73
	Max	4,48	4,66	4,82	19,79	17,75	15,86
$\sigma = 0,098$ $\mu = 2,694$	Min	4,48	4,61	4,84	19,83	17,74	15,99
	Moy	1,39	1,77	2,43	5,57	8,32	10,73
	Max	4,53	4,60	4,86	19,64	17,57	15,77
$\sigma = 0,098$ $\mu = 3,171$	Min	4,57	4,60	4,87	19,85	17,64	15,83
	Moy	1,39	1,72	2,34	5,52	8,45	10,75
	Max	4,51	4,60	4,82	19,69	17,62	15,83
$\sigma = 0,193$ $\mu = 1,680$	Min	8,80	9,03	9,54	19,41	18,78	18,37
	Moy	2,75	3,37	4,53	5,76	6,70	7,94
	Max	9,02	9,08	9,41	19,43	18,83	18,48
$\sigma = 0,193$ $\mu = 2,680$	Min	9,07	9,16	9,45	19,65	18,96	18,31
	Moy	2,79	3,38	4,64	5,72	6,79	8,07
	Max	8,98	9,12	9,33	19,63	18,95	18,12
$\sigma = 0,193$ $\mu = 3,157$	Min	8,97	8,95	9,72	19,54	18,92	18,29
	Moy	2,77	3,48	4,70	5,76	6,74	7,95
	Max	8,98	8,73	9,53	19,57	18,66	18,49
$\sigma = 0,283$ $\mu = 1,659$	Min	13,21	13,26	13,84	19,56	19,53	20,00
	Moy	4,09	5,02	6,68	6,03	6,58	8,09
	Max	13,21	13,11	13,95	19,79	19,78	20,02
$\sigma = 0,283$ $\mu = 2,659$	Min	13,02	13,27	13,94	19,47	19,66	19,99
	Moy	4,03	4,90	6,76	6,05	6,55	7,88
	Max	13,19	13,26	13,61	19,72	19,80	19,71
$\sigma = 0,283$ $\mu = 3,136$	Min	12,74	13,18	13,73	19,63	19,92	19,96
	Moy	4,03	4,97	6,81	6,14	6,72	7,97
	Max	12,90	13,11	13,77	19,69	19,59	20,00

Tableau 5 : RRMSE des estimateurs GLM-ML et ML des quantiles de probabilité au non-dépassement 50%, 90%, et 99%, conditionnellement aux valeurs minimale, moyenne, et maximale de la covariable, pour le modèle  $LN_{(2)}$ .

Modèle $LN_{(2)}$	Valeur de la covariable	RRMSE %(GLM-ML)			RRMSE %(ML)		
		50%	90%	99%	50%	90%	99%
$\sigma = 0,098$ $\mu = 1,694$	Min	5,90	5,81	6,17	20,61	18,71	16,99
	Moy	1,73	2,24	2,95	5,89	8,84	11,59
	Max	5,75	5,91	6,11	20,80	18,83	17,09
$\sigma = 0,098$ $\mu = 2,694$	Min	5,74	5,82	6,11	20,83	18,84	17,24
	Moy	1,74	2,21	3,00	5,94	8,85	11,59
	Max	5,77	5,86	6,14	20,67	18,66	17,01
$\sigma = 0,098$ $\mu = 3,171$	Min	5,84	5,87	6,18	20,85	18,76	17,08
	Moy	1,74	2,15	2,91	5,88	8,99	11,58
	Max	5,80	5,92	6,14	20,71	18,77	17,10
$\sigma = 0,193$ $\mu = 1,68$	Min	11,28	11,64	12,15	21,13	20,55	20,46
	Moy	3,48	4,20	5,66	6,56	7,82	9,49
	Max	11,57	11,53	12,03	21,10	20,67	20,53
$\sigma = 0,193$ $\mu = 2,68$	Min	11,67	11,66	12,12	21,29	20,82	20,42
	Moy	3,49	4,23	5,78	6,53	7,88	9,68
	Max	11,49	11,59	11,96	21,33	20,77	20,20
$\sigma = 0,193$ $\mu = 3,157$	Min	11,56	11,60	12,31	21,24	20,76	20,34
	Moy	3,48	4,33	5,86	6,58	7,84	9,58
	Max	11,39	11,17	12,07	21,19	20,47	20,50
$\sigma = 0,283$ $\mu = 1,659$	Min	17,22	16,89	17,66	21,98	22,16	22,89
	Moy	5,09	6,26	8,28	7,26	8,16	10,30
	Max	17,38	16,89	17,80	22,32	22,32	22,81
$\sigma = 0,283$ $\mu = 2,659$	Min	17,13	17,24	17,92	22,00	22,29	22,79
	Moy	5,07	6,15	8,38	7,28	8,15	9,97
	Max	17,08	17,22	17,28	22,15	22,37	22,47
$\sigma = 0,283$ $\mu = 3,136$	Min	16,56	16,95	17,76	22,09	22,49	22,75
	Moy	5,09	6,20	8,44	7,36	8,27	10,12
	Max	16,78	16,85	17,72	22,15	22,13	22,86

Pour le modèle  $LN_{(2)}$ , les résultats donnés par la méthode de GLM-ML sont meilleurs que ceux obtenus par la méthode de ML. La différence est claire en particulier pour le RBIAIS. En fait, pour la méthode de ML et pour un modèle plus compliqué, le biais augmente. Pour ce cas étudié, les résultats montrent que la méthode proposée GLM-ML est une bonne alternative à la méthode de ML classique.



## 5 CHOIX DU MODÈLE

---

La qualité de l'ajustement de GLM est mesurée par la déviance  $D$ , qui est liée à la comparaison des fonctions de vraisemblance de deux modèles emboîtés. L'expression algébrique pour la déviance  $D$  diffère pour chaque membre de la famille exponentielle. Pour le cas spécial où la distribution de  $Y_i$  est normale et la fonction de lien est la fonction identité, la déviance devient simplement la somme des carrés des résidus (RSS) de la régression multiple (Clarke, 2002).

Pour le modèle linéaire normal, Dobson (2002) a donné cette déviance par :

$$D = \frac{1}{\sigma^2} \sum_{i=1}^N (y_i - \hat{\mu}_i)^2 \quad (37)$$

(Avec  $\mu_i$  définie de la même façon que pour les modèles linéaires généralisés).

Pour le choix du modèle le mieux adapté aux observations, Dobson (2002) suggère pour le modèle linéaire normal un test qui représente une comparaison entre le rapport  $F$  qui est la statistique du test basé sur les déviances des deux modèles à comparer, et la loi de cette statistique qui est la loi de Fisher  $F(p - q, N - p)$ . Cette statistique est donnée par :

$$F = \frac{D_0 - D_1}{p - q} / \frac{D_1}{N - p} = \frac{\left\{ \sum [y_i - \hat{\mu}_i(0)]^2 - \sum [y_i - \hat{\mu}_i(1)]^2 \right\} / (p - q)}{\sum [y_i - \hat{\mu}_i(1)]^2 / (N - p)} \quad (38)$$

où  $D_0$  est la déviance du modèle qui possède le moins des paramètres ( $q$ ), et  $D_1$  la déviance du modèle qui possède le plus des paramètres ( $p$ ),  $p > q$ .

Dans le cas où le premier modèle ne s'ajuste pas bien aux données on va se trouver dans la région de rejet de la loi  $F(p - q, N - p)$ .

De façon générale pour n'importe quelle distribution on prend  $\Delta D = D_0 - D_1 \sim \chi^2(p - q)$ . Si la valeur de  $\Delta D$  est conformée à la distribution  $\chi^2(p - q)$  nous choisirons généralement le premier modèle.

## 6. APPLICATION : PRÉCIPITATION DE LA STATION TEHACHAPI

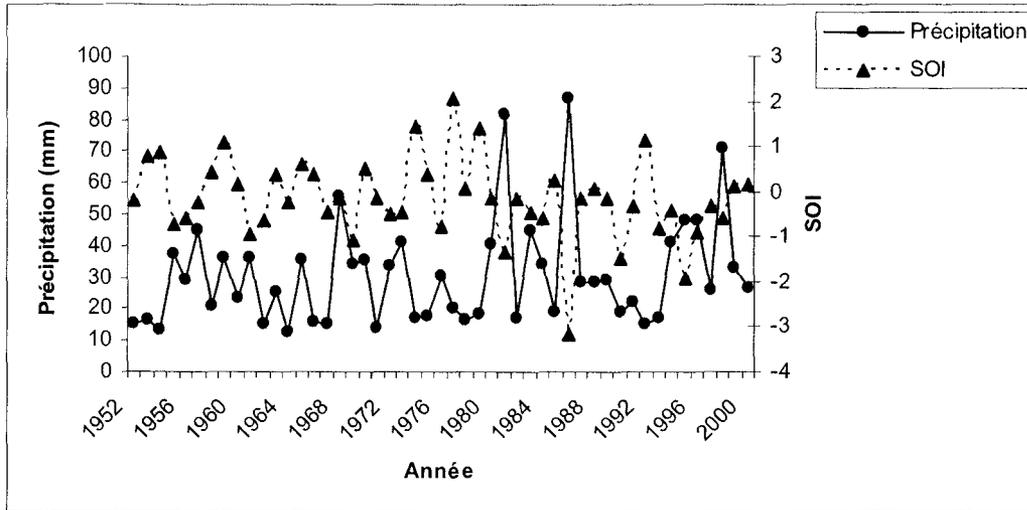
---

Nous avons appliqué les modèles présentés ci-dessus (section 2) à une série  $\underline{X} = (x_1, \dots, x_n)$  de précipitations maximales annuelles instantanées en mm, enregistrées à la station Tehachapi de la Californie (Station 048826 Latitude : 35.13 Longitude : -118.45 Période 1952-2000 et le nombre d'années d'enregistrement  $n = 49$ ). La figure 1 montre la localisation géographique de la station d'étude. Située au sud de la Californie, les précipitations de cette région sont fortement affectées par le phénomène El Niño.



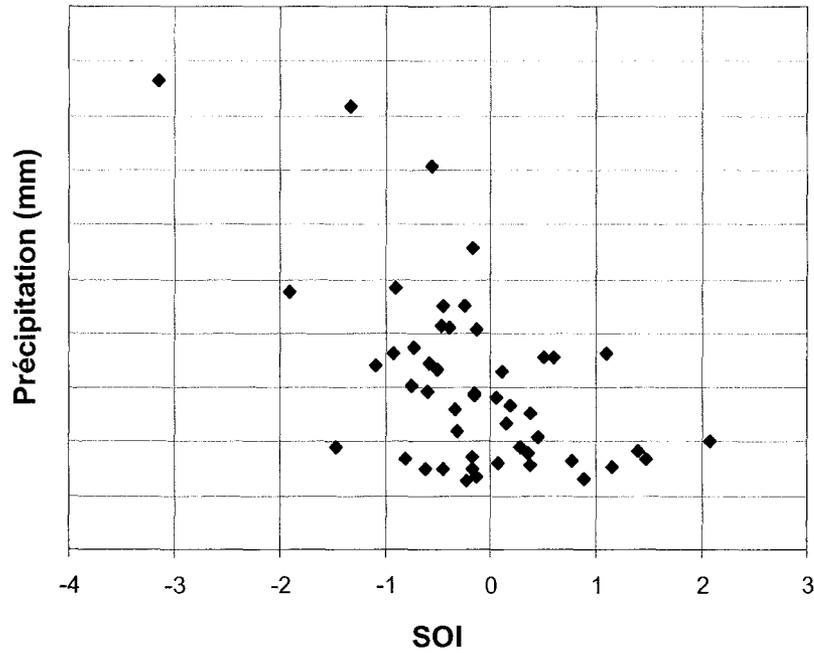
**Figure 1. Situation géographique de la station Tehachapi de la Californie.**

L'objectif de ce travail est d'étudier la distribution des précipitations maximales annuelles, en appliquant le modèle Log-normal non-stationnaire sous ces deux formes  $LN_{(1)}$  et  $LN_{(2)}$ , avec comme covariable l'indice climatique SOI □ *Southern Oscillation Index* □ . Le coefficient de corrélation entre la variable  $X$  et la covariable SOI est -0.57. Les deux figures 2 et 3 illustrent la relation entre ces deux variables.



**Figure 2. Précipitations maximales annuelles observées et les valeurs des SOI correspondantes**

La figure 3 montre qu'il y a une corrélation négative significative (coefficient de corrélation  $\rho = -0.57$ ) entre les précipitations et l'indice climatique SOI. On remarque d'après la figure 3 que les plus grandes valeurs des précipitations correspondent aux faibles valeurs de l'indice climatique SOI.



**Figure 3. Nuage de points des précipitations maximales annuelles observées et les valeurs des SOI correspondantes**

Le but de cette application est d'estimer les paramètres et les quantiles des trois modèles non-stationnaire  $LN_{(0)}$ ,  $LN_{(1)}$  et  $LN_{(2)}$  en utilisant l'approche de GLM-ML. Le tableau 6 présente les estimations des paramètres des trois modèles considérés pour cet exemple ainsi que les valeurs des log-vraisemblances correspondantes. On s'intéressera surtout à l'estimation de la médiane (quantile correspondant à la probabilité de non-dépassement 50%) par les trois modèles.

**Tableau 6 : Les estimateurs des paramètres par GLM-ML et les valeurs de log-vraisemblance correspondantes**

	$l_n^*$	$\mu_1$	$\mu_2$	$\mu_3$	$\sigma$
$LN_{(0)}$	<b>-195.69</b>	3,3006	-	-	0.4839
$LN_{(1)}$	<b>-187.94</b>	3,2560	-0,2870	-	0.4132
$LN_{(2)}$	<b>-187.34</b>	3,2261	-0,2591	0,0430	0.4081

Pour cette application, le choix du modèle va être fait en utilisant la déviance statistique, ce test est simple et très utilisé dans le cas où il y a des modèles qui sont des cas spéciaux des autres (Coles, 2001). Le modèle  $LN_{(0)}$  est un cas particulier du modèle  $LN_{(1)}$  et le modèle  $LN_{(1)}$  est un cas particulier du modèle  $LN_{(2)}$ . Quand la différence entre les deux modèles est non évidente, il est préférable d'utiliser le modèle le plus simple afin de respecter le principe de parcimonie. Une méthode simple pour comparer la validité d'un modèle  $M_1$  ou un modèle  $M_0$ , sachant que  $M_0 \subset M_1$ , est d'utiliser la déviance statistique défini par (Coles, 2001) :

$$D = 2 \{l_n^*(M_1) - l_n^*(M_0)\} \quad (39)$$

Où  $l_n^*(M)$  est la fonction du log-vraisemblance du modèle  $M$ . Une grande valeur de  $D$  indique que le modèle  $M_1$  est le plus adéquat pour ajuster les variations des données que le modèle  $M_0$ . La statistique  $D$  est distribuée suivant une loi de chi-deux ( $\chi^2_\nu$ ). Le paramètre  $\nu$  est la différence entre la dimension (le nombre de paramètres) du modèle  $M_1$  et la dimension du modèle  $M_0$ . Des valeurs de  $D$  qui sont plus grandes que les quantiles de la loi

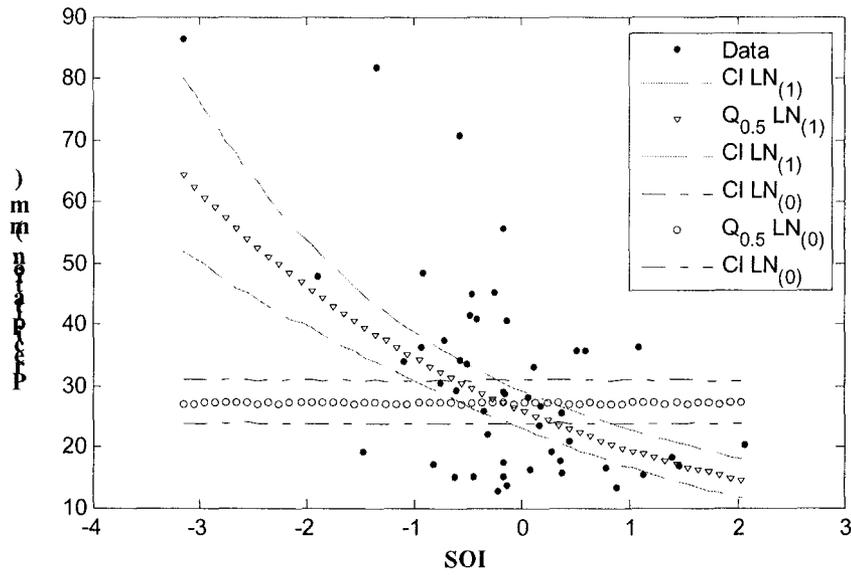
de  $\chi^2_\nu$  pour un niveau de confiance particulier  $\alpha = 5\%$ , signifie que le modèle  $M_1$  est meilleur que le modèle  $M_0$ .

Les résultats montrent qu'il y a une différence significative entre les modèles  $LN_{(1)}$  et  $LN_{(0)}$  puisque  $D = 15.49$  est cette valeur est plus grande que le quantile (95%) de la loi  $\chi^2_1$  ( $\Pr(\chi^2_1 \leq 15.49) = 0.9999$ ). D'autre part, pour les modèles  $LN_{(1)}$  et  $LN_{(2)}$ , le test statistique montre qu'il n'y a pas de différence significative entre ces deux modèles, car  $D = 1.20$  et  $\Pr(\chi^2_1 \leq 1.20) = 0.7273$ . Ainsi, pour le niveau de signification  $\alpha = 5\%$ , le modèle  $LN_{(1)}$  est le plus adéquat pour représenter la dépendance entre les précipitations maximales annuelles et la covariable SOI.

**Tableau 7. Les estimateurs des quantiles de probabilité au non-dépassement 50% conditionnellement aux valeurs min, mean, et max de SOI.**

	SOI=-3.1600	SOI=-0.1553	SOI= 2.0720
$LN_{(0)}$	27,1290	27,1290	27,1290
$LN_{(1)}$	64,2678	27,1290	14,3147
$LN_{(2)}$	87,7767	26,2438	17,7078

Les médianes sont calculées conditionnellement à trois valeurs spéciales de la covariable: le minimum, la moyenne et le maximum (Tableau7). Les figures 4 et 5 représentent la médiane et ses intervalles de confiance. La médiane est estimée conditionnellement aux valeurs de SOI pour les trois modèles.



**Figure 4. La médiane Conditionnelle et ses intervalles de confiance estimés par les modèles  $LN_{(0)}$  et  $LN_{(1)}$ .**

Les intervalles de confiance sont calculés en générant 3000 ( $N=3000$ ) échantillons par la méthode de Bootstrap. Les résultats montre que les modèles  $LN_{(1)}$  et  $LN_{(2)}$  donnent le même estimateur de la médiane pour presque toutes les valeurs de SOI (Figure 6). D'autre part, la différence entre le cas stationnaire et les cas non-stationnaires est claire, en particulier pour les petites valeurs de SOI.

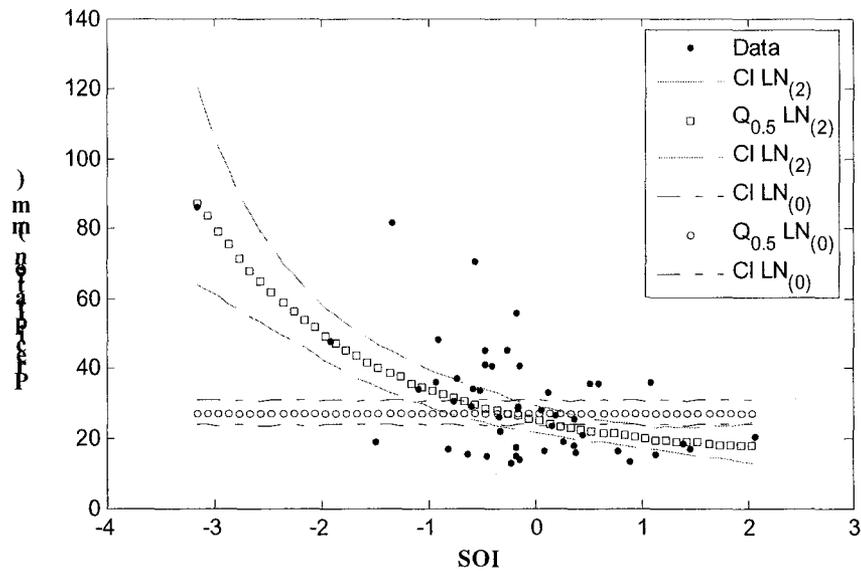
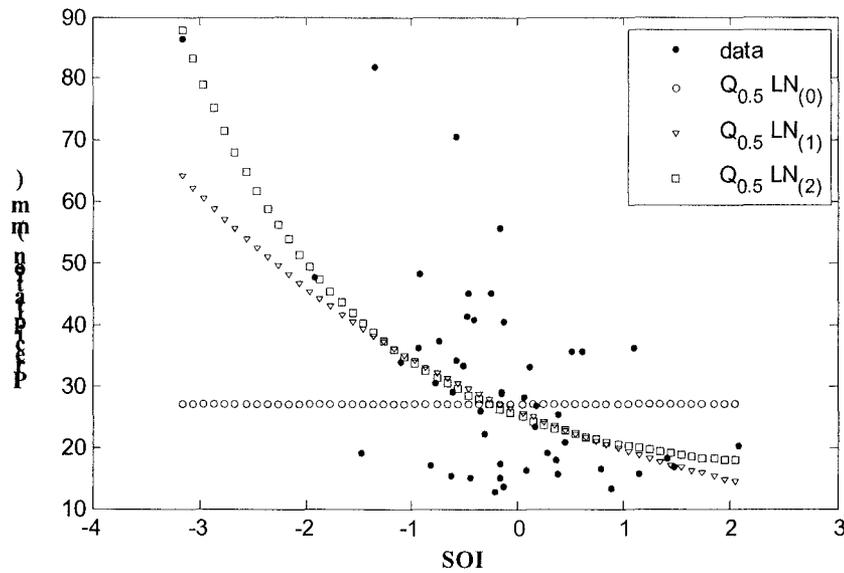


Figure 5. La médiane Conditionnelle et ses intervalles de confiance estimés par les modèles  $LN_{(0)}$  et  $LN_{(2)}$ .

Dans ce dernier cas, l'estimateur de la médiane par le modèle  $LN_{(2)}$  peut être trois fois plus grand que celui estimé par le modèle classique  $LN_{(0)}$  ((Figure 4 et 5). Les intervalles de confiance sont plus larges dans le cas des petites valeurs de la covariable. Cela est dû au nombre d'observations correspondantes aux précipitations extrêmes qui est petit, ainsi l'incertitude est plus grande pour les valeurs centrales et élevées de SOI.



**Figure 6. Les estimateurs de la médiane Conditionnellement a la covariable SOI, pour les modèles  $LN_{(0)}$ ,  $LN_{(1)}$  et  $LN_{(2)}$ .**

La figure 6 montre que les modèles  $LN_{(1)}$  et  $LN_{(2)}$  donnent des estimateurs équivalents pour la médiane, dans le cas des valeurs centrales et élevées de SOI. Notons également que la courbe du modèle  $LN_{(1)}$  a la même forme que celle du modèle  $LN_{(2)}$ . En effet, le fait que la variance de la loi Log-normal dépend du paramètre  $\mu$  produit une dépendance convexe de la covariable.

## 7 CONCLUSION

---

L'objectif principal de cette étude était de résoudre le problème d'estimation des paramètres du modèle Log-normal non-stationnaire. L'estimation se fait généralement par la méthode du maximum de vraisemblance et par le biais de méthodes numériques pour la résolution du système d'équations correspondant. Dans le cas non-stationnaire d'autres paramètres qui représentent la dépendance des paramètres du modèle stationnaire aux covariables s'ajoutent au problème d'estimation et rendent la résolution plus compliquée. Pour résoudre ce problème, nous avons remarqué que le modèle Log-normal non-stationnaire peut être présenté sous forme d'un modèle linéaire généralisé. Cette correspondance nous a permis d'employer les méthodes classiques et explicites d'estimation des paramètres pour les modèles linéaires généralisés. Cependant ces techniques ne permettent pas d'estimer le deuxième paramètre de la loi Log-normale qui correspond à l'écart-type de la transformation logarithmique de la variable étudiée. L'approche proposée GLM-ML considère les estimateurs des paramètres obtenus par les expressions des modèles linéaires généralisés, comme point de départ de l'algorithme de Newton-Raphson pour la résolution du système d'équations du maximum de vraisemblance.

L'approche GLM-ML est illustrée dans le présent travail pour les modèles  $LN_{(1)}$  et  $LN_{(2)}$  est générale et peut être appliquée à d'autre forme de non-stationnarité ou de dépendance des covariables. Ceci permet dans le cadre des études hydro-climatiques d'introduire les différents effets de covariables ou des changements climatiques.

Les deux méthodes, GLM-ML et ML, sont comparées par simulation de Monte Carlo, dans le cas de la dépendance linéaire et quadratique à l'égard de la covariable qui est normalement distribuée. La comparaison des deux méthodes se base sur le calcul du biais relatif (RBIAIS) et la racine de l'erreur quadratique moyenne relative (RRMSE).

Les résultats montre que la méthode proposée « GLM-ML » est meilleure que la méthode de ML, en particulier dans le cas de la dépendance quadratique. En effet, pour ce modèle, le paramètre  $\mu$  de la loi Log-normal est une fonction quadratique de la covariable, et le système de ML comporte plus d'équations à résoudre. Dans ce cas l'utilisation de la méthode de GLM-ML est plus simple et les solutions sont obtenues dans peu d'itérations.

L'application de la méthode de cette étude à des données de précipitations maximales annuelles, enregistrées à la station Tehachapi de la Californie, conditionnellement à l'indice climatique SOI, a été faite en utilisant les trois modèles. En fait, ce type de modèle permet de tenir compte des dépendances entre les variables aléatoires, et de faire introduire l'effet des changements climatiques sur la variabilité de l'évolution de la série observée. Les deux modèles  $LN_{(1)}$  et  $LN_{(2)}$  ont été utilisés pour estimer les quantiles par la méthode GLM-ML, et les intervalles de confiance ont été construits par la technique de Bootstrap. La comparaison entre les trois modèles, avec un test basé sur la déviance statistique des différents modèles, montre que le modèle  $LN_{(1)}$  est le plus convenable pour représenter la variance des données. En effet, la différence entre les modèles  $LN_{(1)}$  et  $LN_{(2)}$  est non significative, c'est pour cette raison qu'on a choisit le modèle qui a le moins de paramètres.

Le cas étudié, montre l'avantage du modèle LN non-stationnaire pour incorporer l'information additionnelle. En fait, pour les petites valeurs de la covariable SOI, le modèle classique sous-estime la médiane ce qui peut avoir des conséquences très dramatiques. L'utilisation des estimations conditionnelles a permis de déduire les valeurs des quantiles correspondants aux événements spéciales qu'on a définis par la covariable.



## RÉFÉRENCES

---

- Chen H. and Rao A. (2002).** Testing Hydrologic Time Series for Stationarity, Newspaper of Hydrologic Engineering, v. 7(2) : 129-136.
- Clarke R.T. (2002a).** Estimating time trends in Gumbel-distributed data by means of generalized linear models, Water Resources Research, vol. 38(7) : 16-1/16-11, 1111, 10.1029/2001 WR000917
- Clarke, R. T. (2002b).** Estimating trends in data from the Weibull and a generalized extreme value distribution, Water Resources Research, 38(6), 1089, doi:10.1029/2001WR000575.
- Coles G.S. (2001).** An Introduction to Statistical Modeling of Extreme Value, Springer, 208 p.
- Davison, A. C. and Ramesh, N. I. (2000).** Local likelihood smoothing of sample extremes. Journal of the Royal Statistical Society Series B, Statistics in Society, Volume 62, part 1, pp. 191-208.
- Dobson A. J. (2002).** An introduction to Generalized Linear Models, Chapman & Hall / CRC second Edition.
- El Adlouni S., Ouarda T. B.M.J., Zhang X., Roy R. and Bobée B. (2006).** Generalized maximum likelihood estimators of the non-stationary GEV model parameters. (under review, Water Resources Research).
- Faucher, D., Ouarda, T.B.M.J. et B. Bobée (1997).** Revue bibliographique des tests de stationnarité. INRS-Eau, rapport de recherche n° R-499, 66 pages.

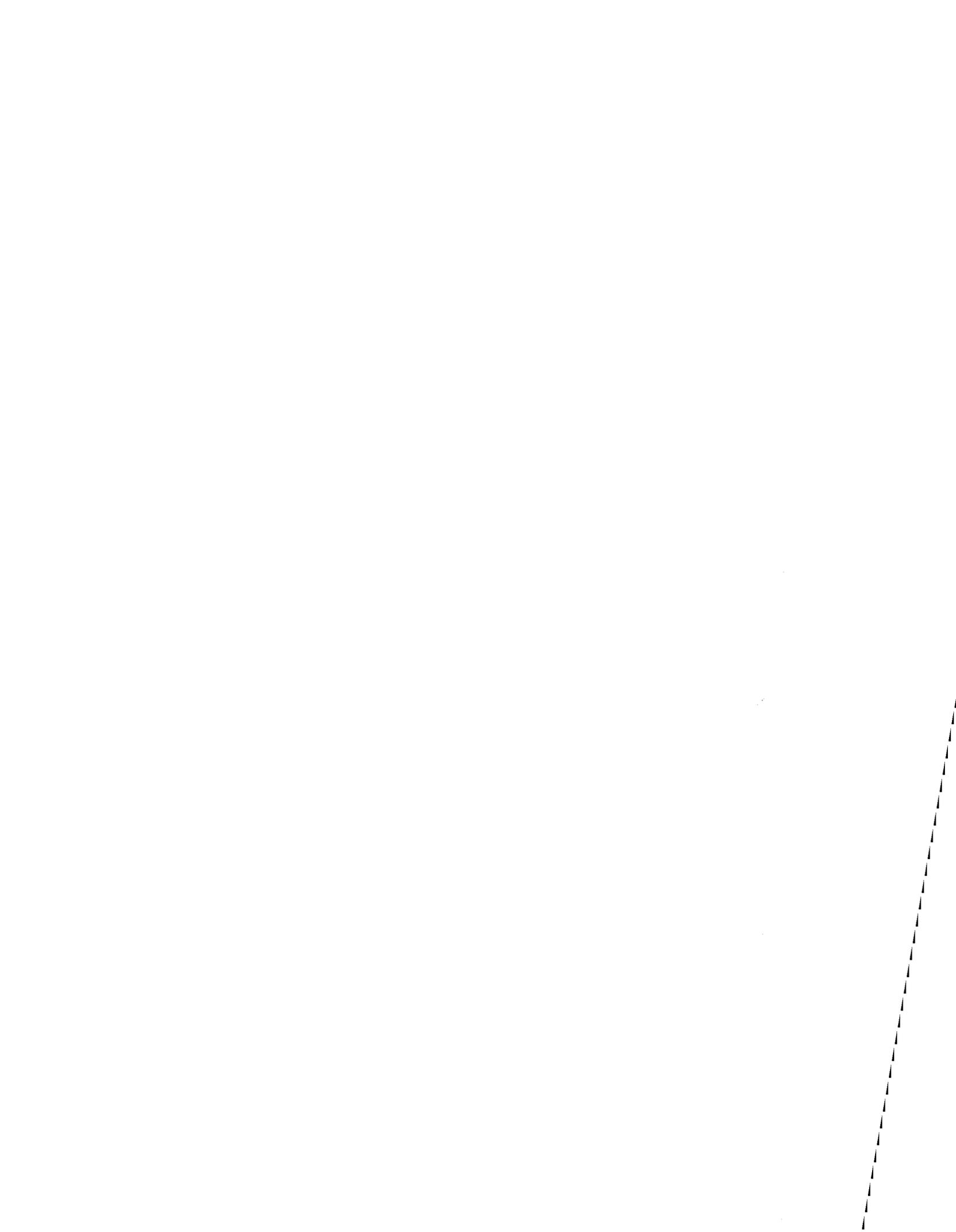
- Hobijn B. (2004).** Generalizations of the KPSS-test for Stationarity, *Statistica Neerlandica*, vol. 58(4) : 483-502.
- Hosking, J.R.M. (1990).** L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. *Newspaper of Royal Statistical Society*, 52 : 105-124.
- Katz R.W., Parlange M.B. and Naveau P. (2002).** Statistics of extremes in hydrology, *Advances in Water Resources*, vol. 25 : 1287-1304.
- Kim K.-D. and Heo J.-H (2002).** Comparative study of flood quantiles estimation by non parametric models. *Journal of Hydrology*. vol 260 : 176-193.
- Kouider A. (2003).** Analyse fréquentielle locale des crues au Québec. Memory for the obtaining of the rank of Mastery in science (M.Sc).
- Leurent F. (1998).** Les valeurs du temps des automobilistes à Marseille en 1995: The distribution of car drivers' values-of-time in Marseilles (1995), *Recherche - Transports – Sécurité*, vol. 60 : 19-38.
- Martin E.S and Stedinger J.R. (2000).** Generalized Maximum Likelihood GEV Quantile Estimators for Hydrologic Data. *Water Resources Research*, vol. 36(3) : 737-744.
- Nelder J.A . and McCullagh P.M. (1983).** *Generalized Linear Models*. Chapman and Hall edition.
- Önöz B. and Bayazit M (2003).** The Power of Statistical Tests for Trend Detectio. *Turkish J. Eng. Env. Sci.* 27 : 247 - 251.
- Ricciardi K. L., Pinder G. F. and Belitz K. (2005).** Comparison of the lognormal and beta distribution functions to describe the uncertainty in permeability. *Newspaper of Hydrology*, pp. 1-9.

**Sankarasubramanian, A. and U. Lall (2003).** Flood Quantiles and Changing Climate: Seasonal forecasts and causal relations. *Water Resources Research*, 39(5) : 1134.

**USWRC (1982).** Guidelines for Determining Flood Flow Frequencial. Bulletin 17B, Hydrology Committee, (U.S. Water Resources Council):Washington, DC.

**Yang C., Chandler R.E. and Isham S. (2005).** Spatial-temporal rainfall simulation using generalized linear models. *Water Resources Research*, vol. 41 : 1-13.

**Zhang X., K. D. Harvey, W. D. Hogg, and T. R. Yuzyk (2001).** Trends in Canadian streamflow. *Water Resources Research*, Vol. 37(4) : 987–999.



## Annexe A

Soit  $Y \square LN(\mu, \sigma^2)$ ,

Si on pose  $\xi = E(Y)$  et  $\tau = Var(Y)$ , on a le système suivant

$$\begin{cases} \xi = \exp\left(\mu + \frac{\sigma^2}{2}\right) \\ \tau = \exp(\sigma^2 + 2\mu)(\exp(\sigma^2) - 1) \end{cases} \Leftrightarrow \begin{cases} \log(\xi) = \mu + \frac{\sigma^2}{2} \\ \tau = \exp(\sigma^2 + 2\log(\xi) - \sigma^2)(\exp(\sigma^2) - 1) \end{cases}$$

$$\Leftrightarrow \begin{cases} \mu = \log(\xi) - \frac{\sigma^2}{2} \\ \tau = \exp(2\log(\xi))(\exp(\sigma^2) - 1) \end{cases} \Leftrightarrow \begin{cases} \mu = \log(\xi) - \frac{\sigma^2}{2} \\ \exp(\sigma^2) = 1 + \frac{\tau}{\exp(2\log(\xi))} \end{cases} \Leftrightarrow$$

$$\begin{cases} \mu = \log(\xi) - \frac{\sigma^2}{2} \\ \sigma = \sqrt{\log\left(1 + \frac{\tau}{\exp(2\log(\xi))}\right)} \end{cases} \Leftrightarrow \begin{cases} \mu = \log(\xi) - \frac{1}{2}\log\left(1 + \frac{\tau}{\exp(2\log(\xi))}\right) \\ \sigma = \sqrt{\log\left(1 + \frac{\tau}{\exp(2\log(\xi))}\right)} \end{cases}$$

Donc  $\log(Y)$  est une variable aléatoire normalement distribuée de

moyenne  $\log(\xi) - \frac{1}{2}\log\left(1 + \frac{\tau}{\exp(2\log(\xi))}\right)$ , et de variance  $\sqrt{\log\left(1 + \frac{\tau}{\exp(2\log(\xi))}\right)}$ .



## Annexe B

La fonction de log-vraisemblance  $l_n$  pour le modèle LN non-stationnaire est donnée par :

$$l_n(\underline{x}; \mu_t, \sigma) = (-n \log(\sigma) - \sum_{t=1}^n \log(x_t) - \frac{n}{2} \log(2\pi)) + \sum_{t=1}^n \left( -\frac{1}{2} \left( \frac{\log x_t - \mu_t}{\sigma} \right)^2 \right) \quad (\text{B1})$$

Par définition, les estimateurs par maximum de vraisemblance sont les solutions du système d'équation des dérivées partielles de  $l_n$  par rapport à chacun des paramètres.

Pour le modèle  $LN2_{(2)}$  ou on a  $\mu_t = \mu_1 + \mu_2 Y_t + \mu_3 Y_t^2$

où  $Y_t$  est le vecteur des covariables, les estimateurs du maximum de vraisemblance des paramètres  $\mu_1, \mu_2, \mu_3$  et  $\sigma$ , seront solutions du système d'équations suivant:

$$\begin{cases} \frac{1}{\sigma^2} \sum_{t=1}^n (\log x_t - \mu_1 - \mu_2 Y_t - \mu_3 Y_t^2) = 0 \\ \frac{1}{\sigma^2} \sum_{t=1}^n Y_t (\log x_t - \mu_1 - \mu_2 Y_t - \mu_3 Y_t^2) = 0 \\ \frac{1}{\sigma^2} \sum_{t=1}^n Y_t^2 (\log x_t - \mu_1 - \mu_2 Y_t - \mu_3 Y_t^2) = 0 \\ -\frac{n}{\sigma} + \left( \frac{1}{\sigma^3} \sum_{t=1}^n (\log x_t - \mu_1 - \mu_2 Y_t - \mu_3 Y_t^2)^2 \right) \end{cases}$$

Si on pose

$$\left\{ \begin{array}{l} k_1 = \sum_{t=1}^n \log x_t \\ k_2 = \sum_{t=1}^n Y_t \\ k_3 = \sum_{t=1}^n Y_t \log x_t \\ k_4 = \sum_{t=1}^n Y_t^2 \\ k_5 = \sum_{t=1}^n (\log x_t)^2 \\ k_6 = \sum_{t=1}^n Y_t^3 \\ k_7 = \sum_{t=1}^n Y_t^4 \\ k_8 = \sum_{t=1}^n Y_t^2 \log x_t \end{array} \right. \quad (\text{B2})$$

le système d'équations du maximum de vraisemblance est :

$$\left\{ \begin{array}{l} \frac{1}{\sigma^2} (k_1 - n\mu_1 - k_2\mu_2 - k_4\mu_3) = 0 \\ \frac{1}{\sigma^2} (k_3 - k_2\mu_1 - k_4\mu_2 - k_6\mu_3) = 0 \\ \frac{1}{\sigma^2} (k_8 - k_4\mu_1 - k_6\mu_2 - k_7\mu_3) = 0 \\ -\frac{n}{\sigma} + \frac{1}{\sigma^3} (k_5 + n\mu_1^2 + k_4\mu_2^2 + 2k_2\mu_1\mu_2 - 2k_1\mu_1 - 2k_3\mu_2 - 2\mu_3k_8 + 2k_4\mu_1\mu_3 + 2k_6\mu_2\mu_3 + k_7\mu_3^2) = 0 \end{array} \right.$$

Ce système peut être résolu par une méthode numérique comme par exemple l'algorithme de Newton-Raphson.

## Annexe C

Les quantiles sont donnés par l'équation (2).

La fonction de répartition d'une loi normale centrée réduite pour une variable aléatoire  $Z$  ( $Z \sim N(0,1)$ ) est donnée par :

$$\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \quad (\text{C1})$$

La fonction de répartition d'une loi normale non centrée réduite d'une variable aléatoire  $X$  ( $X \sim N(\mu, \sigma^2)$ ) s'écrit :

$$F(x) = \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{t-\mu}{\sigma}\right)^2/2} dt \quad (\text{C2})$$

$$\text{En posant } z = \frac{t-\mu}{\sigma} \Rightarrow dz = \frac{dt}{\sigma}$$

On a :

$$\begin{aligned} F(x) &= \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{t-\mu}{\sigma}\right)^2/2} dt \\ &= \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right) \end{aligned} \quad (\text{C3})$$

Donc

$$F(x) = \Phi(z) = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt \quad (C4)$$

on pose  $y = \frac{t}{\sqrt{2}} \Rightarrow dy = \frac{dt}{\sqrt{2}}$

On a :

$$\begin{aligned} p = \Phi(z) &= \int_{-\infty}^{\frac{x-\mu}{\sigma\sqrt{2}}} \frac{e^{-y^2}}{\sqrt{\pi}} dy \\ &= \int_{-\infty}^0 \frac{e^{-y^2}}{\sqrt{\pi}} dy + \int_0^{\frac{x-\mu}{\sigma\sqrt{2}}} \frac{e^{-y^2}}{\sqrt{\pi}} dy \\ &= \frac{1}{2} \left( 1 + \operatorname{erf} \left( \frac{x-\mu}{\sigma\sqrt{2}} \right) \right) \end{aligned}$$

où la fonction  $\operatorname{erf}$  est définie par :

$$\operatorname{erf}(x) = 2 \int_0^x \frac{e^{-t^2}}{\sqrt{\pi}} dt \quad (C5)$$

Et 
$$\int_{-\infty}^{+\infty} \frac{e^{-y^2}}{\sqrt{\pi}} dy = 1 \quad (C6)$$

D'où

$$\Phi^{-1}(p) = z = \frac{x-\mu}{\sigma} = [\operatorname{erf}^{-1}((2 \times p) - 1)] \times \sqrt{2} \quad (C7)$$

Où la fonction  $\operatorname{erf}^{-1}$  est la fonction inverse de la fonction  $\operatorname{erf}$ .

Donc d'après la formule (2) et (C7) on a :

$$H^{-1}(p) = \exp(\mu + \sigma \times \sqrt{2} \times \operatorname{erfinv}((2 \times p) - 1)) \quad (\text{C8})$$

Ce qui équivaut à

$$H^{-1}(p) = \exp(\mu + \sigma \Phi^{-1}(p)) \quad (\text{C9})$$

La formule (C9) représente la formule des quantiles de probabilité au non-dépassement  $p$ .