Université du Québec
INRS-Eau, Terre et Environnement

DÉVELOPPEMENT DU MODÈLE LOG-NORMAL NON STATIONNAIRE ET APPLICATION À L'ESTIMATION DES EXTRÊMES

Par
Ilham Aissaoui Fqayeh

Mémoire présenté pour l'obtention
du grade de Maîtrise ès sciences (M.Sc.)
en Sciences de l'eau et de l'environnement

Jury d'évaluation

Examinateur externe	Pr. Saad Bennis École de Technologie Supérieure
Examinateur interne	Pr. Alain Maillot INRS-Eau, Terre et Environnement
Codirecteurs de recherche	Dr. Salah El Adlouni Pr. André St-Hilaire INRS-Eau, Terre et Environnement
Directeur de recherche	Pr. Taha B.M.J Ouarda INRS-Eau, Terre et Environnement

TABLE DES MATIÈRES

RÉSUMÉ ..... **V**

SECTION 1 : SYNTHÈSE ..... **VII**

REMERCIEMENTS ..... **VII**

INTRODUCTION **1**

MISE EN CONTEXTE **1**

PROBLÉMATIQUE **1**

OBJECTIFS DU MÉMOIRE **3**

CHAPITRE 1 : CONTRIBUTION DE L'ÉTUDIANTE **5**

CHAPITRE 2 : SITUATION DE LA CONTRIBUTION **7**

RÉFÉRENCES **7**

SECTION 2 : ARTICLE 1 ..... **11**

1 INTRODUCTION **13**

2 MODÈLE LOG-NORMAL NON-STATIONNAIRE **15**

2.1 Caractéristiques statistiques 15

2.1.1 Distribution Log-normale 15

2.1.2 Modèle Log-normal Non-stationnaire 17

2.2 Estimation des paramètres par la méthode du maximum de vraisemblance 19

2.3 Estimation des quantiles 21

3 COMPARAISON DES MODÈLES LN ET GEV NON-STATIONNAIRES **22**

3.1 Plan de simulation 22

3.2 Résultats 25

4 APPLICATION : PRÉCIPITATIONS À LA STATION TEHACHAPI **29**

5. CONCLUSIONS **34**

RÉFÉRENCES **35**

ANNEXE A **49**

ANNEXE B **51**

SECTION 3 : ARTICLE 2 ..... **53**

1 INTRODUCTION **53**

2 RAPPEL SUR LE MODÈLE LOG-NORMAL NON-STATIONNAIRE 53
2.1. Test sur la tendance: 55
2.2. Le modèle Log-normal non-stationnaire 56
3 CARACTÉRISTIQUES STATISTIQUES DES SÉRIES ÉTUDIÉES 57
3.1 Résultats du test de Mann-Kendall 57
3.2 Caractéristiques statistiques 58
4 ESTIMATION DES PARAMÈTRES ET DE LA MÉDIANE 58
5 DISCUSSION DES RÉSULTATS 59
RÉFÉRENCES 61
LISTE DE TABLEAUX 65
LISTE DE FIGURES 65
ANNEXE A 73
ANNEXE B 77
ANNEXE C 81
SECTION 4 : CONCLUSION 83

RÉSUMÉ

La connaissance des débits de crue est nécessaire pour la conception des aménagements des cours d'eau, le dimensionnement des ouvrages hydrauliques et la protection des zones urbaines. Des études statistiques spécifiques sont nécessaires pour la prédétermination des débits de crue. L'analyse fréquentielle est un des outils privilégiés pour l'estimation des débits de crue. Les approches classiques d'analyse fréquentielle nécessitent l'hypothèse de stationnarité des séries d'observations. Cette hypothèse est de plus en plus mise en doute dans le contexte des changements climatiques observés.

L'objectif principal de cette étude est de présenter le modèle Log-normal non-stationnaire pour l'estimation des séries non-stationnaires et de le comparer au modèle GEV non-stationnaire. L'estimation des paramètres est effectuée par la méthode du maximum de vraisemblance. Ce genre de modèle permet d'introduire des covariables au niveau du premier paramètre de la loi Log-normale ou le temps dans le cas de tendance temporelle. Trois formes différentes du modèle Log-normal non stationnaire sont investiguées: modèle classique, modèle avec tendance linéaire, et modèle avec tendance quadratique. Ces modèles ont été comparés, par simulation de Monte Carlo, aux modèles généralisés des valeurs extrêmes (GEV) non-stationnaires, qui ont été bien étudiés dans la littérature. Les modèles non-stationnaires, d'une manière générale, représentent des outils efficaces pour tenir compte de l'effet des changements climatiques en modélisant l'évolution des valeurs des paramètres en fonction du temps ou d'autres covariables. L'utilité de ces modèles est illustrée par l'étude de l'effet d'un indice climatique (Southern Oscillation Index, SOI) sur des données hydro-climatiques de la Californie, USA. L'effet de l'indice SOI sur les précipitations maximales annuelles, enregistrées à la station Tehachapi de la Californie est étudié

dans le premier article. Le deuxième article traite de l'effet de la tendance sur des données de débit de pointe dans plusieurs stations à travers le monde.

SECTION 1 : SYNTHÈSE

REMERCIEMENTS

J'exprime mes profonds remerciements à mon directeur de thèse, le professeur Taha B.M.J Ouarda pour l'aide compétente qu'il m'a apportée, pour sa patience et son encouragement à finir ce travail commencé il y a longtemps. Son oeil critique m'a été très précieux pour structurer le travail et pour améliorer la qualité des différentes sections.

Mes deux co-directeurs de thèse Salah El Adlouni et André St-Hilaire qui ont beaucoup contribué à mettre en forme mon français et mon anglais très approximatifs. Sans eux, j'aurais sans doute été découragée par ces langues que j'aime bien.

Je remercie mon mari, Abdelkader, pour l'amour qu'il me porte et pour son soutien et sa patience pendant que je travaillais souvent presque toute la journée et parfois le soir aussi.

D'autres personnes m'ont encouragé à finir ce travail qui sont : ma mère que j'aime beaucoup, mon père, mes sœurs et mes frères.

Mon adorable bébé Adam qui a 11 mois maintenant, et qui était très gentil avec moi durant la période de grossesse, il m'a laissé travailler sans me poser des problèmes connus chez la plupart des femmes enceintes.

"Last but not least", j'exprime ma gratitude aux membres de mon jury de thèse: Taha B.M.J Ouarda, Salah El Adlouni, André St-Hilaire, Saad Bennis et Alain Maillot.

INTRODUCTION

Mise en contexte

Ce mémoire « par Article » se divise en deux parties. La première partie qui représente le premier article, fait état de l'introduction du modèle Log-normal non-stationnaire, et de la méthode du maximum de vraisemblance pour l'estimation des paramètres. Ce modèle a été ensuite comparé au modèle GEV non-stationnaire pour évaluer les erreurs liées à la méthode d'estimation et celles correspondant à un mauvais choix du modèle. On y décrit la problématique, les objectifs, la méthodologie et les résultats de l'application de ces deux familles de modèles à des précipitations maximales annuelles journalières, enregistrées à la station Tehachapi de la Californie.

La deuxième partie du mémoire est constituée d'un deuxième article où on applique la nouvelle approche aux débits de pointe dans différentes stations répartis à travers le monde.

Problématique

Les valeurs extrêmes sont souvent les maximums d'une certaine quantité sur une période donnée. La théorie des valeurs extrêmes montre que la distribution de ces maxima est représentée par une des trois lois des valeurs extrêmes (GEV). Les conditions théoriques de l'utilisation de la loi GEV ne sont pas souvent vérifiées et d'autres lois sont considérées pour représenter les différentes formes d'extrêmes observés. Notons également que les approches classiques d'analyse fréquentielle qui supposent la stationnarité des séries d'observations, ainsi que l'indépendance et l'homogénéité ne sont pas toujours vérifiées. En effet, 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.* 2007, Zhang *et al.* 2001). Il faudrait donc développer des outils qui tiennent compte de la non-stationnarité des séries hydro-climatiques ou de dépendance d'autres covariables. Ce genre de modèles permet d'inclure l'effet de différentes covariables sur la variabilité et l'évolution de la série observée.

Dans le cadre d'une analyse fréquentielle, les hypothèses d'indépendance, d'homogénéité et de stationnarité des séries observées doivent être vérifiées. Plusieurs tests permettent de vérifier ces conditions. On s'intéresse principalement, dans le cadre de cette étude, à l'hypothèse de non-stationnarité. Plusieurs tests permettent de tester cette hypothèse comme par exemple le test 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).

L'estimation des quantiles en analyse fréquentielle dépend des propriétés des estimateurs des paramètres de la distribution à ajuster. Les méthodes d'estimation les plus utilisées sont la méthode des moments de probabilités pondérées (PWM) et la méthode des L-moments (LM) (Hosking 1990) en raison de leur simplicité pour la majorité des distributions. Cependant, dans un cadre non-stationnaire, les estimateurs des L-moments ne peuvent pas être obtenus d'une manière directe. 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).

Les travaux sur les modèles d'analyse fréquentielle dans un cadre non-stationnaires (ou en présence de dépendance de covariables) sont liés au modèle GEV (Coles 2001 et Clarke 2002a,b). En effet, la loi GEV est la loi la plus utilisée en hydrologie pour la modélisation des extrêmes. C'est dû principalement aux propriétés théoriques et à la simplicité de la fonction de densité de la loi GEV et sa fonction quantile. La loi GEV a été, également jugée dans plusieurs études comme adéquate pour l'ajustement des séries des maximums (Faucher *et al.* 1997). Dans d'autres régions, d'autres distributions ont été suggérées pour l'étude des séries hydrométéorologiques : 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). Notons que la loi LN a un comportement au niveau des extrêmes qui est très semblable à celui des distributions à queues lourdes (Ouarda *et al.*, 1994). Il est souvent très difficile de discriminer entre la loi LN et les distributions de type puissance (comme par exemple la loi GEV et la loi Log-Pearson type III).

Objectifs du mémoire

Dans le premier article on présente le modèle Log-normal non-stationnaire à deux paramètres. Ce modèle s'ajuste bien à plusieurs variables hydrométéorologiques. Il est également 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 de l'article 1 est de comparer, sur la base de simulations Monte Carlo, les propriétés des estimateurs des quantiles obtenus par la méthode du maximum de vraisemblance à partir des modèles LN non-stationnaires et GEV non-stationnaires pour évaluer les erreurs liées à la taille de l'échantillon et à l'efficacité de la méthode d'estimation ainsi que les erreurs générées par un mauvais choix du modèle. Une application des deux familles de modèles non-stationnaires LN et GEV à des données de précipitations maximales annuelles conditionnellement à l'indice climatique SOI a permis de comparer leurs résultats, et d'illustrer la flexibilité de ces modèles pour l'ajustement des séries hydrométéorologiques.

L'objectif principal de l'article 2 est d'appliquer le modèle Log-normal non stationnaire à des séries de débit maximum annuel de 17 stations tirées du site de l'UNESCO et qui correspondent à différentes rivières du monde. Les critères de sélection de ces séries sont : (1) la taille de la série (supérieure à 70 ans d'observations), (2) les tests d'ajustement montrent que la loi Log-normale représente bien la série étudiée. On considère ensuite le modèle Log-normal non-stationnaire pour l'estimation des tendances temporelles qui pourraient exister dans ces séries de débit de crue.

CHAPITRE 1 : CONTRIBUTION DE L'ÉTUDIANTE

La problématique et l'approche globale ont été proposées par mon directeur de recherche. J'ai réalisé une revue de littérature sur les modèles non-stationnaires et les tests de détection de non-stationnarité durant l'hiver 2005. Les résultats de cette revue sont présentés dans le chapitre «Introduction » de l'article 1. J'ai ensuite travaillé, sous la supervision de mes directeurs de recherche, au développement des équations mathématiques pour l'estimation des paramètres par la méthode du maximum de vraisemblance (ML) pour le modèle Log-normal non-stationnaire et la comparaison par simulation des modèles LN et GEV non-stationnaires. Ces développements sont présentés respectivement dans les sections 2 et 3 de l'article 1. J'ai ensuite analysé les résultats des simulations avec la contribution des mes directeurs de recherche.

J'ai colligé les banques de données des débits maximums annuels de plusieurs rivières du monde à partir du site de l'UNESCO pour l'étude de la non-stationnarité et l'application des modèles étudiés. Les résultats de cette application sont présentés dans l'article 2.

J'ai entièrement programmé les codes Matlab pour tester la stationnarité des séries étudiées, les codes pour les simulations Monte Carlo ainsi que les codes pour le calcul des estimateurs des paramètres et des quantiles sous la supervision du Dr. El Adlouni.

J'ai rédigé les deux articles avec la contribution de mes directeurs de recherche. J'ai ensuite rédigé le présent mémoire qui a été corrigé par mes directeurs de recherche.

CHAPITRE 2 : SITUATION DE LA CONTRIBUTION

Dans le premier travail (article1), après avoir présenté les modèles LN non-stationnaires les résultats des simulations ont permis d'évaluer les erreurs liées à la taille de l'échantillon et la méthode d'estimation ainsi que les erreurs qui sont dues à un mauvais choix du modèle.

La contribution originale du présent travail est de présenter un nouveau modèle pour l'ajustement des séries à des lois de probabilité dont les paramètres sont fonctions de covariables. Le choix du modèle Log-normal est justifié par deux principales raisons : (1) la loi Log-normale est très utilisée pour la modélisation des séries hydrométéorologiques et possède des propriétés très intéressantes pour l'ajustement des extrêmes (queue de distribution); (2) elle a seulement deux paramètres, ce qui lui permet d'avoir une flexibilité pour la modélisation d'une large gamme de variation et d'asymétrie en respectant le principe de parcimonie. Une application qui illustre la flexibilité et l'efficacité du modèle proposé et de la nouvelle approche pour l'estimation des paramètres est aussi présentée.

Dans l'article2 on étudie le débit de pointe de 17 stations tirées du site de l'UNESCO et qui correspondent à différentes rivières du monde. Les critères de sélection de ces séries sont : (1) la taille de la série (supérieure à 70 ans d'observations), (2) les tests d'ajustement montrent que la loi Log-normale semble être adéquate pour représenter la série étudiée. Le modèle Log-normal non-stationnaire est considéré pour l'estimation des débits de crue, pour les 17 stations. Ensuite, le choix du modèle le plus adéquat (parmi les trois différents modèles utilisés : classique et deux types de modèle non-stationnaire) pour chaque station est effectué en utilisant la statistique de déviance.

RÉFÉRENCES

Clarke, R. T. (2002a). Estimating time trends in Gumbel-distributed data by means of generalized linear models, Water Resources Research, *38*(7), 1111, doi:10.1029/2001WR000917.

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.

El Adlouni S., Ouarda T. B.M.J., Zhang X., Roy R. et Bobée B. (2007). Generalized maximum likelihood estimators for the non-stationary GEV model. Water Resources Research, sous presse.

Faucher, D., Ouarda, T.B.M.J. et B. Bobée (1997). Revue bibliographique des tests de stationnarité. INRS-Eau, rapport de Recherche no R-499, 66 pages.

Hosking, J.R.M. (1990). L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. Newspaper of the Royal Statistical Society, 52:105-124.

Katz R.W., Parlange M.B. et Naveau P. (2002). Statistics of extremes in hydrology, Advances in Water Resources, vol. 25, pp. 1287-1304.

Kouider A. (2003). Analyse fréquentielle locale des crues au Québec. M. Sc. Thesis, INRS-ETE, University of Quebec, Canada.

Martins E.S. et Stedinger J.R. (2000). Generalized Maximum Likelihood GEV Quantile Estimators for Hydrologic Data. Water Resources Research, vol. 36. No. (3). pp.737-744.

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

Zhang X., K. D. Harvey, W. D. Hogg, et T. R. Yuzyk (2001). Trends in Canadian streamflow. Water Resources Research, Vol. 37, No. 4, pp. 987–999.

SECTION 2 : ARTICLE 1

Développement du modèle Log-normal non-stationnaire et comparaison avec le modèle GEV non-stationnaire

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

Soumis au Journal des Sciences Hydrologiques

Mars 2007

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.*, 2007, 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.

Avant de procéder à une analyse fréquentielle il faut vérifier les hypothèses d'homogénéité et de stationnarité des séries observées. Plusieurs tests permettent de vérifier ces hypothèses. 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, de plus la méthode des L-moments (LM) introduit un faible biais pour les faibles échantillons. 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). El Adlouni *et al.*, (2007) ont donné une extension de la méthode du maximum de vraisemblance généralisée (GML) introduite par Martins et Stedinger (2000), 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 hydrométéorologiques.

La majorité des travaux sur les modèles non-stationnaires sont liés au modèle généralisé des valeurs extrêmes (GEV) (Coles 2001 et 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 : 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 *et al.*, 2002). Notons que la loi LN a un comportement au niveau des extrêmes très semblable à celui des distributions à queues lourdes (heavy tailed distributions) (Ouarda *et al.* 1994). Il est souvent très difficile de discriminer entre la loi LN et les distributions de type puissance (comme par exemple la loi GEV et la loi Log-Pearson type III).

Dans cette étude on présente le modèle Log-normal non-stationnaire à deux paramètres. Ce modèle, comme il a été déjà mentionné, s'ajuste bien à plusieurs séries de données hydrologiques et respecte le 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 est de comparer, sur la base de simulations Monte Carlo, les propriétés des estimateurs de quantiles par la méthode du maximum de vraisemblance obtenus pour les modèles LN non-stationnaire et GEV non-stationnaire.

Après avoir rappelé certaines propriétés statistiques de la loi Log-normale on présente le modèle Log-normal non-stationnaire et certains des cas particuliers (sous trois formes particulières) avec la méthode du maximum de vraisemblance pour l'estimation des paramètres. Nous présentons ensuite dans la section 3 une comparaison par simulation Monte Carlo entre le modèle LN non-stationnaire et le modèle GEV non-stationnaire. Une application qui illustre la flexibilité et l'efficacité du modèle proposé sera présentée dans la section 4. Les conclusions et la discussion générale sont présentées dans la section 5.

2 MODÈLE LOG-NORMAL NON-STATIONNAIRE

2.1 Caractéristiques statistiques

2.1.1 Distribution Log-normale

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 μ et σ, la moyenne et l'écart-type de la transformée logarithmique, $\log(X)$. La variable $Z=(\log(X)-\mu)/\sigma$ est distribuée suivant une loi normale centrée réduite. Soit Φ la fonction de répartition de la loi normale centrée réduite et F celle de la loi Log-normale. On a donc les égalités suivantes :

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

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

La fonction de répartition peut être écrite sous la forme :

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

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

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

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

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

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

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

La variance s'écrit donc de la façon suivante:

$$\sigma_X^2=\exp(2\mu+\sigma^2).(\exp(\sigma^2)-1)=E[X]^2.(\exp(\sigma^2)-1) \tag{7}$$

Le coefficient de variation (C_v) et le coefficient d'asymétrie (C_S) sont donnés par :

$$C_v = \left(e^{\sigma^2} - 1\right)^{\frac{1}{2}} \quad \text{et} \quad C_S = \left(e^{\sigma^2} - 1\right)^{\frac{1}{2}} \left(e^{\sigma^2} + 2\right).$$

Après une transformation logarithmique, on obtient des variables aléatoires normalement distribuées de moyenne μ et de variance σ^2 données par les formules ci-dessous:

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

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

2.1.2 Modèle Log-normal Non-stationnaire

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 indices climatiques tels que les indices d'oscillations climatiques de basse fréquence (Coles 2001, Katz *et al.* 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 les paramètres sont fonction du temps ou d'autres covariables. Ainsi, pour des valeurs fixes du temps ou/et de covariables, X_t est distribué selon une loi Log-normale. Différents types de dépendances peuvent être considérés en représentant des paramètres comme fonctions des covariables. Cependant, quand deux modèles donnent un bon ajustement, le modèle avec le nombre de paramètres le plus faible sera préféré.

L'introduction des covariables peut être effectuée au niveau de n'importe quel paramètre, ou même à deux ou trois paramètres à la fois. Sankarasubramanian et Lall (2003) ont comparé trois méthodes pour l'estimation des quantiles conditionnels. La première méthode est paramétrique et est basée sur une transformation d'une loi Normale dont les paramètres dépendent de covariables. Les deux autres méthodes correspondent à une approche non paramétrique (Régression des quantiles, Koenkar et Bassett, 1978) et une approche semi-paramétrique (Modèle de vraisemblance locale, Davison et Ramesh, 2000).

L'introduction de covariables au niveau des paramètres implique des changements des caractéristiques statistiques de la distribution ajustée. Plusieurs auteurs suggèrent d'associer un changement dans la variance à celui de la moyenne pour que le coefficient de variation (Cv) reste constant. Cependant, dans le cas d'une variable distribuée suivant une loi Log-normale, le coefficient Cv est fonction seulement du paramètre σ. Ainsi, garder le coefficient Cv constant implique que le paramètre σ est constant. Par conséquent, dans le présent travail seul le paramètre μ est fonction des covariables.

Pour étudier les propriétés des estimateurs des paramètres, trois modèles sont considérés dans le présent travail :

$LN_{(0)}(\mu,\sigma)$. 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 μ est une fonction linéaire d'une covariable Y_t.

$LN_{(2)}\left(\mu_t = \beta_1 + \beta_2 Y_t + \beta_3 Y_t^2, \sigma\right)$. Le paramètre μ 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 μ mène à un changement de la moyenne et de la variance de la variable, ce qui correspond souvent aux changements observés dans des séries hydrométéorologiques.

2.2 Estimation des paramètres par la méthode du maximum de vraisemblance

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:

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

où n est la taille de la série à ajuster.

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

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

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

Pour le modèle $LN_{(1)}$, où $\mu_t = \beta_1 + \beta_2 Y_t$ et Y_t est le vecteur des covariables, les estimateurs du maximum de vraisemblance des paramètres β_1, β_2 et σ sont solutions du système d'équations suivant:

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

Si on pose

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

Le système d'équations du maximum de vraisemblance est :

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

Ce système peut être résolu directement par substitution ou à partir d'une écriture matricielle équivalente.

D'une manière analogue on obtient un système équivalent au système précédent pour le modèle $LN_{(2)}$, avec une quatrième équation pour le paramètre β_3. Les équations correspondant à l'estimation des paramètres du modèle $LN_{(2)}$ par la méthode du maximum de vraisemblance sont présentées dans l'annexe A.

2.3 Estimation des quantiles

Le quantile de probabilité au non-dépassement p conditionnellement à une valeur observée y_0 de la covariable Y, est obtenu à partir d'une loi Log-normale de paramètres $\left(\mu_{y_0}, \sigma\right)$. Ce quantile peut être présenté sous forme explicite :

$$x_{p,y_0} = \exp\left(\mu_{y_0} + \sigma\Phi^{-1}(p)\right). \tag{14}$$

Où μ_{y_0} est le premier paramètre 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 est fonction linéaire de la covariable: $\mu_{y_0} = \beta_1 + \beta_2 y_0$, et pour le modèle $LN_{(2)}$ il s'écrit sous la forme : $\mu_{y_0} = \beta_1 + \beta_2 y_0 + \beta_3 y_0^2$.

En remplaçant ces paramètres par leurs estimateurs dans l'équation (14), on obtient les estimateurs des quantiles.

3 COMPARAISON DES MODÈLES LN ET GEV NON-STATIONNAIRES

L'objectif de cette section est de comparer le modèle LN non-stationnaire au modèle GEV non-stationnaire. Les modèles GEV non-stationnaires considérés dans ce travail ont été décrits en détail par El Adlouni *et al.* (2007) et sont présentés dans l'annexe B. Pour étudier la différence entre les deux modèles, nous avons considéré la même méthode d'estimation à savoir la méthode du maximum de vraisemblance. Nous avons considéré plusieurs types de non-stationnarité qui peuvent être observées dans des séries hydrométéorologiques. Dans tous les cas, la covariable est distribuée suivant une loi Normale centrée réduite.

3.1 Plan de simulation

Dans cette partie on compare par simulation les modèles LN et GEV non-stationnaires (Annexe B) dans l'estimation des extrêmes conditionnellement aux valeurs de covariables. Le choix des paramètres des modèles simulés est effectué sur la base de la moyenne et du coefficient de variation de 40 séries de débit maximum annuel du Québec, étudiées par Kouider *et al.* (2002). Les résultats obtenus à partir de cette étude nous ont permis d'obtenir les intervalles de variation des deux paramètres de la loi LN. Les coefficients de variation pour ces 40 séries appartiennent à l'intervalle $[0.15; 0.35]$. Nous avons considéré pour cette étude deux coefficients de variation : $C_v = 0.2$ et $C_v = 0.3$. La moyenne a été fixée à $E[X] = 2000$. Les valeurs de l'asymétrie, pour une variable distribuée suivant une loi LN, peuvent être déduites de celles du coefficient de variation à partir de la formule suivante :

$$C_s = C_v(C_v^2 + 3) \tag{15}$$

Les coefficients d'asymétrie correspondant aux deux coefficients de variation $C_v = 0.2$ et $C_v = 0.3$ sont $Cs = 0.6$ et $Cs = 0.9$, respectivement. Comme conséquence au choix de ces deux cas, les valeurs du paramètre σ pour la loi LN et le paramètre de forme, κ, pour la loi GEV, sont fixés pour chaque asymétrie (Tableau 1).

Comme nous l'avons déjà mentionné, le modèle $LN_{(1)}$ peut représenter (1) le cas où la moyenne dépend d'une covariable et (2) la moyenne et la variance sont fonction de la covariable. Nous avons considéré les trois cas suivants :

Cas 1 : on compare M1 = $LN_{(1)}$ et M2 = $GEV_{(10)}$ (le paramètre de position est fonction linéaire de la covariable).

Cas 2 : on compare M1 = $LN_{(1)}$ et M2 = $GEV_{(11)}$ (les paramètres de position et d'échelle sont fonctions linéaires de la covariable).

Cas 3 : comparaison des modèles M1 = $LN_{(2)}$ et M2 = $GEV_{(21)}$ (le paramètre de position est une fonction quadratique et le paramètre d'échelle est fonction linéaire de la covariable).

Pour les trois cas, la covariable Y_t est distribuée selon une loi normale centrée réduite (de moyenne 0 et de variance 1). Les paramètres de dépendance de la covariable, pour les deux modèles, ont été choisis convenablement pour que les échantillons générés à partir des deux modèles aient les mêmes caractéristiques de tendance et de changement d'échelle. Le tableau 1 présente les paramètres des modèles pour les trois cas considérés et pour les deux coefficients de variation.

[Tableau 1]

Pour comparer ces modèles (deux à deux), on génère $N = 5000$ échantillons de taille $n = 50$ à partir du modèle M_i ($i = 1, 2$) et on estime les quantiles de probabilité au non-dépassement 50%, 90% et 99% à partir des deux modèles M1 et M2. Les quantiles sont estimés conditionnellement à des valeurs particulières de la covariable (les valeurs : minimale (min), moyenne (moy) et maximale (max)). On calcule ensuite le biais relatif (BR) et la racine de l'erreur quadratique moyenne relative (REQMR). Le BR et la REQMR sont donnés par les formules suivantes :

$$\text{BR}= \frac{1}{M}\sum_{i=1}^{M}\frac{\left|\hat{Q}_{i,p}-Q_p\right|}{Q_p} \tag{16}$$

$$\text{REQMR} = \left[\frac{1}{M}\sum_{i=1}^{M}\left(\frac{\hat{Q}_{i,p}-Q_p}{Q_p}\right)^2\right]^{1/2} \tag{17}$$

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.

Cette comparaison permettra d'évaluer : (1) les erreurs reliées à la méthode d'estimation des paramètres pour une faible taille de l'échantillon ($n = 50$) dans le cas où on effectue un bon choix du modèle ainsi que (2) les erreurs reliées au mauvais choix du modèle (choisir un modèle LN au lieu de GEV et vice-versa).

3.2 Résultats

Les résultats de la comparaison entre les modèles LN et GEV non-stationnaires, pour les 6 cas considérés sont présentés dans les tableaux 2 à 7. Dans chaque cas on présente le BR et le REQMR des quantiles conditionnellement à la valeur minimale, moyenne et maximale de la covariable (distribuée suivant une loi Normale). Notons que la différence entre les deux modèles et relativement importante dans le cas des quantiles extrêmes (probabilité au non-dépassement 99%) conditionnellement à la valeur minimale ou maximale de la covariable. Cependant, l'analyse des résultats va être faite principalement pour le quantile de probabilité au non-dépassement 99% conditionnellement à la valeur maximale de la covariable.

Les modèles $LN_{(1)}$ et $GEV_{(10)}$

Pour le premier cas, on compare les modèles $LN_{(1)}$ et $GEV_{(10)}$ avec une faible asymétrie (le coefficient de variation $C_v = 0.2$). Les résultats sont présentés dans le tableau 2.

[Tableau 2]

On remarque que lorsqu'on génère à partir du modèle $GEV_{(10)}$ (Bloc de ligne correspondant au modèle $GEV_{(10)}$) le BR du quantile (de probabilité au non-dépassement 99% conditionnellement à la valeur maximale de la covariable) estimé à partir du modèle $LN_{(1)}$ est de l'ordre de -9% alors qu'il est très faible pour $GEV_{(10)}$ (-0.75%). Cependant, la différence en terme de la REQMR des quantiles estimés à partir de ces deux modèles, n'est pas aussi importante que celle obtenue pour le biais. La REQMR est de l'ordre de 12% pour le modèle $GEV_{(10)}$ et de 14% pour le modèle $LN_{(1)}$ (pour le même quantile, 99%, et la même valeur de la covariable qui est le maximum).

Dans le cas où les échantillons sont générés à partir d'un modèle $LN_{(1)}$, (Bloc de ligne correspondant au modèle $LN_{(1)}$) le BR maximum est de -0.17% pour le modèle $LN_{(1)}$ et 21% pour le modèle $GEV_{(10)}$. On remarque que cette différence est très importante pour le biais et elle l'est aussi pour l'erreur quadratique moyenne (REQMR est de l'ordre de 10% pour $LN_{(1)}$ et 32% pour $GEV_{(10)}$).

Le tableau 3 présente les résultats de la comparaison de ces deux modèles, $LN_{(1)}$ et $GEV_{(10)}$, pour le coefficient de variation $C_v = 0.3$.

[Tableau 3]

Les mêmes conclusions sont valables pour le deuxième cas ($C_v = 0.3$), avec des valeurs du BR et REQMR relativement plus élevées que celles obtenus dans le cas de faible asymétrie (Tableau 3). Notons que les REQMR des quantiles estimés à partir du modèle $LN_{(1)}$ sont presque les mêmes quel que soit la population mère (générer à partir du modèle $GEV_{(10)}$ ou $LN_{(1)}$). Alors que, lorsque l'ajustement est effectué avec le modèle $GEV_{(10)}$ la différence est très grande dépendamment de la population mère (12% quand on génère à partir de $GEV_{(10)}$ et 32% dans le cas de valeurs simulées à partir de $LN_{(1)}$).

Les modèles $LN_{(1)}$ et $GEV_{(11)}$

Pour le deuxième cas, on considère des échantillons générés à partir des modèles $LN_{(1)}$ et $GEV_{(11)}$ pour tenir compte de l'effet d'une dépendance linéaire du paramètre d'échelle de la loi GEV en fonction de la covariable. Les résultats dans le cas de $C_v = 0.2$ sont présentés dans le tableau 4.

[Tableau 4]

On remarque que dans ce cas les BR des quantiles estimés à partir du modèle $GEV_{(11)}$, pour des échantillons générés à partir du modèle $LN_{(1)}$, sont relativement faibles par rapport aux résultats de comparaison des modèles $GEV_{(10)}$ et $LN_{(1)}$. Les BR obtenus sont du même ordre que ceux obtenus avec le modèle $LN_{(1)}$ lorsque les échantillons sont générés à partir du modèle $GEV_{(11)}$. Notons que ces deux modèles se ressemblent plus que ceux considérés dans la première comparaison $GEV_{(10)}$ et $LN_{(1)}$. On remarque également que :

les REQMR des quantiles estimés par le modèle $GEV_{(11)}$ sont approximativement les mêmes quelle que soit la population à partir de laquelle on génère ;

- les REQMR des quantiles obtenus à partir du modèle $LN_{(1)}$ sont faibles par rapport à ceux du modèle $GEV_{(11)}$ quel que soit le modèle à partir duquel on génère. Ceci est dû principalement au nombre de paramètres considérés dans chaque modèle (3 paramètres pour $LN_{(1)}$ et 5 paramètres pour $GEV_{(11)}$). En effet, les modèles avec plus de paramètres permettent d'avoir une bonne estimation au niveau du biais (relativement faible par rapport au biais de modèles avec moins de paramètres) alors que l'erreur d'estimation augmente avec le nombre de paramètres.

Les résultats correspondants à la comparaison de ces deux modèles dans le cas de $C_v = 0.3$ sont donnés dans le tableau 5.

[Tableau 5]

Les conclusions tirées pour le premier cas, faible asymétrie, restent valables pour $C_v = 0.3$ avec une légère augmentation du REQMR pour les deux modèles (tableau 5). Cependant, les REQMR des quantiles sont presque les mêmes si on considère le même modèle d'ajustement (càd indépendamment du modèle de génération). Les REQMR des quantiles obtenus par le modèle $LN_{(1)}$ restent toujours inférieurs à ceux du modèle $GEV_{(11)}$ quelle que soit la population mère (le modèle considéré pour générer les données, $GEV_{(11)}$ ou $LN_{(1)}$).

<u>Les modèles</u> $LN_{(2)}$ **<u>et</u>** $GEV_{(21)}$

Pour ce troisième cas on compare les modèles $LN_{(2)}$ et $GEV_{(21)}$ où les paramètres de position et d'échelle sont, respectivement, des fonctions quadratique et linéaire de la covariable. Ces modèles permettent de modéliser la dépendance des covariables avec plus de flexibilité. Les résultats de cette comparaison sont présentés dans le tableau 6 (cas où $C_v = 0.2$) et le tableau 7 (cas où $C_v = 0.3$).

[Tableau 6]

On remarque que les REQMR des deux modèles restent les mêmes quelle que soit la distribution de la population mère. Ceci peut être expliqué par la flexibilité de ces deux modèles pour représenter les deux populations considérés dans cette étude (générée à partir du modèle $GEV_{(21)}$ et du modèle $LN_{(2)}$).

Quand on génère à partir du modèle $GEV_{(21)}$, le modèle $LN_{(2)}$ performe mieux que le modèle $GEV_{(21)}$ en terme de REQMR, pour les deux asymétries. Ceci est dû principalement à la variance élevée causée par le nombre de paramètre qui est égale à 6 dans le cas du modèle $GEV_{(21)}$. Alors que pour le biais, le bon choix du modèle mène automatiquement à un faible biais.

[Tableau 7]

4 APPLICATION : PRÉCIPITATIONS À LA STATION TEHACHAPI

Nous avons appliqué les modèles GEV et LN non-stationnaires à une série $\underline{X} = (x_1, \ldots, x_n)$ de précipitations maximales annuelles journalières en mm, enregistrées à la station Tehachapi de la Californie (Station USGS #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 de la station d'étude. Située au sud de la Californie, les précipitations de cette région sont fortement modulées par le phénomène El Niño.

[Figure 1]