Université du Québec
Institut National de la Recherche Scientifique
Centre Eau Terre Environnement

ESTIMATION DES ÉVÉNEMENTS EXTRÊMES PAR UN MODÈLE GEV-B-SPLINES. Étude de cas : PRÉCIPITATIONS EXTRÊMES À LA STATION RANDSBURG EN CALIFORNIE

Par

Bouchra Nasri

Mémoire présenté pour l'obtention du grade de
Maître ès sciences (M.Sc.) en science de l'eau

Jury d'évaluation

Président du jury et Examinateur interne	Fateh Chebana INRS-ETE
Examinateur externe	Debbie Dupuis HEC-Montréal
Codirecteur de recherche	Salaheddine El Adlouni Université de Moncton
Directeur de recherche	Taha BMJ Ouarda INRS-ETE

© Droits réservés de Nasri Bouchra, 2013

AVANT-PROPOS

Ce mémoire de maîtrise par article est composé de deux chapitres. Le premier chapitre intitulé «Synthèse», fait état de la problématique et de la pertinence de mon sujet de recherche ainsi que de ma contribution à ce travail de recherche. Les détails sur la méthode théorique utilisée ainsi que les résultats obtenus sont présentés dans l'article au chapitre II de ce document. L'article a été soumis au journal «Open Journal of Statistics» en décembre 2012 et accepté en février 2013. Le choix de ce journal vient du fait que la théorie présentée dans cet article peut être utilisée dans des domaines autres que l'hydrologie et la climatologie.

Le titre et les auteurs de l'article sont les suivants :

Titre: Bayesian Estimation of the GEV-B-Splines Model

Auteurs : Bouchra Nasri, Salaheddine El Adlouni, Taha B.M.J Ouarda

Les contributions des auteurs sont divisées comme suit :

Bouchra Nasri : Contribution à l'élaboration de la méthodologie, la revue de littérature, le fondement théorique, la programmation informatique, la production et l'interprétation des résultats, et à la rédaction de l'article

Salaheddine El Adlouni : Contribution à l'élaboration de la méthodologie, la programmation informatique, l'interprétation des résultats et à la rédaction de l'article.

Taha B.M.J Ouarda : Contribution à l'élaboration de la méthodologie, l'interprétation des résultats et à la rédaction de l'article.

REMERCIEMENTS

La première personne que je tiens à remercier est mon directeur de recherche M. **Taha Ouarda**, pour l'orientation, la confiance, la patience qui ont constitué un apport considérable sans lequel ce travail n'aurait pas pu être mené au bon port. Qu'il trouve dans ce travail un hommage vivant à sa haute intelligence, sagesse et personnalité.

Mes remerciements vont également à M. **Salaheddine El Adlouni**, professeur à l'Université de Moncton, professeur invité à l'INRS-ETE et mon codirecteur de recherche, qui a déployé tant d'efforts pour la réussite de ma maîtrise. Sa sympathie et sa disponibilité à rendre service ont constitué en outre, des ingrédients nécessaires au bon déroulement de ma maîtrise.

Je tiens à exprimer ma gratitude à tous les membres du jury pour avoir accepté de juger mon travail.

RÉSUMÉ

Les deux dernières décennies ont vu un développement très important dans la modélisation statistique des événements hydrologiques et climatiques extrêmes. L'analyse fréquentielle (AF) est une des méthodes statistiques les plus utilisées pour l'estimation de l'occurrence et de l'intensité de ces événements. L'AF consiste à étudier les événements passés afin de définir les probabilités d'apparition futures. La détermination de la loi de probabilité des événements extrêmes est l'élément clé de la procédure de l'AF. La loi généralisée des valeurs extrêmes (GEV) est flexible et a fait l'objet de plusieurs études théoriques et des applications pour la modélisation des débits, des précipitations et des vents extrêmes. Une des hypothèses de base de l'AF classique est la stationnarité. Toutefois cette hypothèse n'est pas toujours vérifiée. L'introduction des covariables au niveau des paramètres de la loi GEV permet de tenir compte des fluctuations interannuelles pour l'estimation du risque dynamique associé aux événements extrêmes. Généralement la dépendance se fait par le biais de fonctions polynomiales de forme linéaire ou quadratique.

L'objectif du présent travail est d'étudier le modèle GEV avec des fonctions de dépendance de type B-Splines (GEV-B-Splines). L'estimation des paramètres du modèle GEV-B-Splines est faite dans un cadre bayesian et l'estimation de la loi a posteriori est effectuée par un algorithme de Monte Carlo par Chaînes de Markov *(Monte Carlo Markov Chain)* (MCMC). Le modèle GEV-B-Splines est appliqué pour estimer les quantiles des précipitations maximales annuelles à la station Randsburg en Californie. Nous avons considéré comme covariables deux indices climatiques : l'indice de l'oscillation australe *(Southern Oscillation Index)* (SOI) et l'indice de l'oscillation décennale du Pacifique *(Pacific decadal oscillation)* (PDO). Les résultats

indiquent une bonne performance du modèle et de la méthode d'estimation proposée pour l'estimation des quantiles extrêmes.

Mots-clés: Loi généralisée des valeurs extrêmes (GEV), les fonctions B-Splines, non stationnarité, non linéarité, (PDO), (SOI).

TABLE DES MATIÈRES

AVANT-PROPOS ..... **III**

REMERCIEMENTS ..... **IV**

RÉSUMÉ ..... **V**

LISTE DES TABLEAUX ..... **IX**

LISTE DES FIGURES ..... **XI**

LISTE DES ABRÉVIATIONS ..... **XIII**

CHAPITRE I : SYNTHÈSE **1**

1. INTRODUCTION **2**

2. SITUATION DE LA CONTRIBUTION DE L'ÉTUDIANT **4**

3. CONTRIBUTION DE L'ÉTUDIANTE **6**

4. RÉSULTATS **7**

5. CONCLUSION **9**

RÉFÉRENCES **11**

CHAPITRE II : ARTICLE **14**

BAYESIAN ESTIMATION FOR GEV-B-SPLINES MODEL **15**

1. INTRODUCTION **17**

2. BAYESIAN GEV-B-SPLINES MODEL **19**

2.1. GEV DISTRIBUTION 19

2.2 THE GEV-B-SPLINES MODEL 21

2.3. THE GEV-B-SPLINES MODEL IN BAYESIAN FRAMEWORK 22

3. CASE STUDY **27**

3.1. DATASET 27

3.2. MODEL DEVELOPMENT 28

4. PARAMETER ESTIMATION COMPARISON **29**

5. CONCLUSION AND RECOMMENDATIONS **30**

RÉFÉRENCES **32**

LISTE DES TABLEAUX

TABLE 1 : CHOICE OF PARAMETERS OF THE B-SPLINE FUNCTION. 36
TABLE 2: BAYESIAN ESTIMATION OF THE PARAMETERS OF THE MODEL 37
TABLE 3: COMPARISON OF ESTIMATION METHODS 37

LISTE DES FIGURES

FIGURE 1: GEOGRAPHIC LOCATION OF THE RANDSBURG STATION 38

FIGURE 2 : VARIATION OF MAXIMUM ANNUAL RAINFALL 38

FIGURE 3 : ANNUAL MAXIMUM RAINFALL AGAINST SOI AND PDO INDEX 39

FIGURE 4 : GEV-B-SPLINES ESTIMATORS OF THE 2, 20 AND 50-YEAR RETURN PERIOD QUANTILES CONDITIONAL UPON SOI 39

FIGURE 5 : GEV-B-SPLINES ESTIMATORS OF THE 2, 20 AND 50-YEAR RETURN PERIOD QUANTILES CONDITIONAL UPON PDO 40

LISTE DES ABRÉVIATIONS

AF : Analyse fréquentielle

GEV : Loi généralisée des valeurs extrêmes

MM : Méthode des moments

MV : Maximum de vraisemblance

MMP : Méthode des moments pondérés

M-H : Metropolis Hasting

MCMC : Monte Carlo par Chaîne de Markov

SOI : Indice de l'oscillation australe

PDO : Indice de l'oscillation décennale du Pacifique

RMSE : Racine carrée de l'erreur quadratique moyenne *(Root mean square error)*

MAR : Maximum annuel des précipitations

TVE : Théorie des valeurs extrêmes

CHAPITRE I : SYNTHÈSE

1. INTRODUCTION

Le travail de recherche présenté dans cette thèse de maîtrise concerne la modélisation statistique des événements extrêmes et l'estimation de leurs probabilités d'apparition. Les deux dernières décennies ont connu le développement de plusieurs approches pour améliorer l'efficacité des estimateurs et tenir compte des problèmes du non stationnarité. Un des nombreux exemples d'applications des approches basées sur la théorie des valeurs extrêmes est la modélisation des variables hydroclimatiques pour des fins de conceptions des ouvrages et de gestion des événements extrêmes.

L'analyse fréquentielle (AF) est une des approches statistiques les plus utilisées pour la modélisation des valeurs extrêmes. Elle consiste à étudier les événements passés afin de définir les probabilités d'apparition future. Elle repose sur la définition et la mise en œuvre d'un modèle fréquentiel, qui permet d'associer à chaque événement une probabilité d'apparition en tenant compte de l'intensité et des fréquences des événements observés. La détermination de la loi de probabilité des événements extrêmes est l'élément clé de la procédure de l'AF. Parmi les lois de probabilités qui ont fait l'objet de plusieurs études théoriques et des applications pour la modélisation des événements hydro-climatologiques extrêmes, on trouve la loi généralisée des valeurs extrêmes *(Generelized Extreme Value)*(GEV) introduite par Jenkinson (1955). La distribution GEV est caractérisée par trois paramètres : un paramètre de position, un paramètre d'échelle et un paramètre de forme. Elle regroupe trois distributions de probabilités dépendamment de la valeur du paramètre de forme : Gumbel (paramètre nul), Fréchet (paramètre de signe positif) et Weibull (paramètre de signe négatif). L'utilisation de la distribution GEV en analyse fréquentielle repose sur un résultat théorique solide, qui est le théorème de Fisher-

Tippett, qui sous certaines conditions, assure la convergence de la loi du maximum vers une loi GEV quelle que soit la distribution des observations initiales.

Pour l'estimation des paramètres de la loi GEV, parmi les méthodes les plus utilisées on cite : la méthode des moments (MM) (Thiele, 1903; Fisher, 1929), la méthode du maximum de vraisemblance (MV) (Smith, 1985), la méthode du maximum de vraisemblance généralisée (MVG) (Martins and Stedinger, 2000 ; El Adlouni et al. 2007), la méthode des moments pondérés (MMP) (Hosking, 1990), la méthode bayésienne (El Adlouni et Ouarda, 2009).

Dans l'AF, les observations doivent être indépendantes et identiquement distribuées; ce qui signifie que les observations doivent être homogènes et stationnaires. Ceci n'est pas toujours vérifié car les données hydro-climatiques peuvent être caractérisées par des formes de non stationnarité (Rao et al. 2003, El Adlouni et al. 2007, 2008) comme c'est le cas dans les contextes environnementaux. Il existe plusieurs tests pour vérifier la stationnarité d'une série de données : d'une part, au nombre des tests paramétriques, on peut citer le test de Dickey et Fuller (1979), le test de Phillips-Perron (1988), le test KPSS proposé par Kwiatkowski et al. (1992), le test de ERS introduit par Elliott et al. (1996) et d'autres tests basés sur le rapport de vraisemblance (cas du test LR : Coles, 2001 ; Mestre, 2003 ; Zhang et al. 2004; Parey et al. 2007). D'autre part, on a le groupe des tests non paramétriques dont le plus cité est le test de la non- stationnarité de Kendall (Önöz et Bayazit, 2003).

2. SITUATION DE LA CONTRIBUTION DE L'ÉTUDIANT

En hydrologie, la non-stationnarité est prise en compte en considérant des covariables au niveau des paramètres de la loi GEV. Ce qui permet de représenter le caractère extrême de la variable étudiée et de tenir compte des effets des covariables à travers les paramètres. L'introduction des covariables peut être effectuée au niveau de n'importe quel paramètre ou même au niveau de deux ou de trois paramètres à la fois dans la loi de probabilité. L'effet d'une covariable peut être pris en compte dans une forme polynomiale.

La plupart des travaux (Coles 2001, El Adlouni et al. 2007, 2008, Canon 2010) se sont concentrés sur le cas d'une dépendance polynomiale (linéaire ou quadratique) entre la covariable et les paramètres. Toutefois, la dépendance entre la covariable et les paramètres peut prendre différentes structures de dépendance. Chavez-Demoulin et Davison (2005) ont suggéré l'utilisation des fonctions semi-paramétriques telles que les splines de lissage pour décrire et estimer la relation entre les paramètres et les covariables.

La contribution du présent travail de recherche consiste à étudier le modèle GEV avec des fonctions de dépendance de type B-Splines (GEV-B-Splines). Les fonctions B-Splines sont des fonctions polynomiales par morceaux qui ont certains avantages. Un lissage à base B-spline est indépendant de la variable réponse et dépend seulement des informations suivantes: (i) l'étendue de la variable explicative ; (ii) le nombre et la position des nœuds et (iii) le degré du B-spline. Ces avantages en font une option appropriée pour l'utilisation dans le modèle GEV avec covariables afin d'estimer les quantiles conditionnels d'une variable de réponse donnée. Pour pouvoir estimer les quantiles, il faut tout d'abord estimer les paramètres de la loi GEV ainsi que les paramètres des fonctions B-Splines.

Dans cette étude l'estimation des paramètres du modèle GEV-B-Splines est faite dans un cadre bayésien et l'estimation de la loi a posteriori est effectuée en utilisant l'algorithme de Metropolis Hasting (M-H).

Dans l'approche bayésienne, les paramètres ne sont pas des valeurs constantes inconnues mais des variables aléatoires admettant une distribution a priori. Toute l'inférence bayésienne est basée sur la loi a posteriori des paramètres et donc des quantiles, qui combine l'information tirée des données à travers la vraisemblance et celle de la loi a priori.

L'estimation de la loi a posteriori est effectuée en utilisant l'algorithme de Metropolis-Hasting (M-H) qui est un algorithme de Monte Carlo par Chaîne de Markov (MCMC). En effet, dans toute la famille de méthodes MCMC, la plus générale est sans doute l'algorithme M-H dans le sens qu'il impose moins de conditions sur la densité cible. Cet algorithme fut d'abord publié par Metropolis et al. (1953) puis généralisé par Hasting (1970). À partir de la densité cible, on choisit une densité instrumentale conditionnelle à partir de laquelle il est assez facile de simuler.

Dans cette étude, le modèle GEV-B-Splines est appliqué pour l'estimation des quantiles de retour des précipitations maximales annuelles à la station Randsburg en Californie.

Les covariables considérées dans ce cadre sont : l'indice de l'oscillation australe (SOI) et l'indice de l'oscillation décennale du Pacifique (PDO). L'oscillation australe et l'oscillation décennale du Pacifique sont des variations de la température de surface de la mer dans les bassins de l'océan Pacifique. L'indice climatique SOI représente la différence de pression entre Tahiti et Darwin tandis que l'indice PDO représente les anomalies de température de surface dans le bassin pacifique.

3. CONTRIBUTION DE L'ÉTUDIANTE

Comme étape préliminaire, j'ai effectué durant l'hiver 2011 une revue de littérature concernant l'analyse fréquentielle non stationnaire, les lois de probabilités utilisées dans ce cadre et les méthodes d'estimation des paramètres de ces lois. Les résultats de cette recherche sont présentés dans la section «Introduction» de l'article au chapitre II du présent document.

Durant l'été 2011, j'ai travaillé sur le développement des équations mathématiques pour l'estimation bayésienne des paramètres du modèle GEV-B-Splines. Ces développements sont présentés dans la partie «Bayesian GEV-B-Splines model» de l'article au chapitre II.

Durant l'automne 2011, j'ai appliqué mon modèle sur une étude de cas qui est le maximum annuel des précipitations de la station Randsburg. Le choix de cette étude de cas est justifié par le fait qu'il y a plusieurs études antérieures qui ont utilisé les mêmes séries de données (El Adlouni et al., 2008, 2009; Cannon 2010). Ceci m'a permis de faire des comparaisons entre les modèles utilisés dans ces précédents travaux avec le modèle choisi dans le présent travail. Ensuite j'ai évalué la performance de la méthode d'estimation utilisée en la comparant à d'autres méthodes d'estimation comme la méthode du maximum de vraisemblance et la méthode des moments et ce, au moyen du biais et de la racine carrée de l'erreur quadratique moyenne *(Root mean square error)*(RMSE). Ensuite, j'ai analysé et interprété les résultats. Ces travaux de recherche sont présentés dans les sections « Case Study » et « Parameter estimation comparison » de l'article au chapitre II de la thèse.

Durant l'hiver 2012, j'ai entamé la rédaction de l'article scientifique et j'ai fait en parallèle mon séminaire de maîtrise.

4. RÉSULTATS

Dans cette étude, le modèle GEV-B-Splines est appliqué aux maximums annuels des précipitations de la station Randsburg (MAR) de la Californie, en utilisant les indices SOI et PDO comme covariables. Nous avons considéré des séries chronologiques de 70 années.

Tout d'abord, avant d'utiliser un modèle d'analyse fréquentielle, il faut commencer par vérifier les conditions d'homogénéité et de stationnarité, puis tester la dépendance entre la variable MAR et les deux covariables climatiques.

Ensuite, nous avons modélisé la série MAR par le modèle GEV-B-Splines. Pour choisir le nombre de nœuds et le degré des fonctions B-Splines utilisées dans cette étude, nous avons comparé plusieurs combinaisons de degrés et de nœuds en utilisant le maximum de vraisemblance puis nous avons choisi la combinaison optimale qui est (3,3) (voir Table 1). Ensuite, nous avons estimé les paramètres du modèle GEV-B-Splines en utilisant la méthode bayésienne. Les lois a priori choisies dans ce cas sont la loi normale multivariée pour les paramètres de la fonction B-Spline, la loi non-informative de Jeffrey qui est $1/\sigma$ pour le paramètre d'échelle et la loi normale pour le paramètre de la forme. Les lois a posteriori sont calculées à partir de l'algorithme M-H. Le modèle est appliqué pour la série MAR en utilisant les deux covariables SOI et PDO. Ceci nous a aidé à calculer les quantiles des précipitations pour des périodes de retour de 2, 20 et 50 ans.

Enfin, nous avons comparé la méthode d'estimation bayésienne aux méthodes d'estimation classiques que sont la méthode du maximum de vraisemblance et la méthode des moments.

La comparaison est effectuée en utilisant le biais et le RMSE des quantiles estimés aux probabilités de non-dépassements : 0.5, 0.8, 0.9 et 0.99. Ci-dessous, on présente les résultats obtenus dans ce travail.

Les résultats ont montré la non-stationnarité de la série MAR à un niveau de signification de 1% ainsi que la dépendance significative entre la variable MAR et les covariables SOI et PDO au niveau de 5%. Ces résultats préliminaires sont bien développés dans le paragraphe 3.1 «Dataset» au chapitre II.

Les résultats de l'estimation des paramètres du modèle GEV-B-Splines sont indiqués dans la table 2. Les figures 4 et 5 montrent les quantiles de précipitations estimés à des temps de retour de 2, 20 et 50 ans. Nous avons constaté que, généralement, l'indice SOI a une corrélation négative avec les précipitations, tandis que l'indice PDO est positivement corrélé avec les précipitations. Les valeurs négatives du SOI et les valeurs positives de PDO coïncident avec les observations MAR relativement élevées. Les quantiles de précipitations augmentent lentement avec l'augmentation de PDO, ensuite ils augmentent de façon exponentielle quand les valeurs de PDO sont supérieures à 1. D'autre part, on a constaté des différents points d'inflexion dans la relation entre les quantiles de précipitations et l'indice SOI, ce qui indique peut être une relation plus complexe entre MAR et SOI qu'entre MAR et PDO.

Les résultats de la comparaison des méthodes d'estimation ont montré la performance supérieure de la méthode bayésienne. Les résultats de la comparaison sont détaillés dans le paragraphe 4 «parameter estimation comparison» au chapitre II.

5. CONCLUSIONS

L'étude de l'estimation des quantiles présente une grande importance dans le domaine de l'hydrologie, pour la raison qu'elle apporte de l'information sur les risques de crues. Les deux dernières décennies ont vu un développement de la modélisation statistique des valeurs extrêmes. La modélisation est passée du modèle hydrologique classique stationnaire (GEV0) aux modèles GEV-B-Splines, en cherchant toujours à améliorer les résultats.

Comme le modèle est très important à élaborer, la méthode avec laquelle on estime les paramètres est l'une des clés de l'optimisation des résultats. Ce travail présente une nouvelle approche pour estimer les paramètres du modèle GEV-B-Splines.

En résumant, ce travail montre dans un premier lieu les fondements théoriques de l'approche bayesienne, ensuite l'application à une série de précipitations avec deux covariables. Enfin, le travail est conclu par une comparaison de la méthode d'estimation bayesienne avec les deux autres méthodes d'estimation : la méthode des moments et la méthode de maximum de vraisemblance. La comparaison de ces méthodes a montré l'utilité du choix de la méthode d'estimation en fonction des différentes périodes de retour.

Le modèle GEV a été comparé avec d'autres types de modèles notamment avec le modèle log-normal (Ouarda et El Adlouni 2009). Cependant ces modèles présentent plusieurs difficultés d'estimations ainsi que d'interprétation des résultats, dans le cas de présence de plusieurs covariables.

Une alternative à ces modèles est la régression des quantiles (Buchinsky, 1998). La régression des quantiles a été présentée en climatologie par Jagger and James (2006) pour décrire la vitesse

du vent en fonction de plusieurs covariables climatiques. Les futurs travaux seront concentrés sur la comparaison des modèles de valeurs extrêmes avec la régression des quantiles pour distinguer l'utilité de l'utilisation de ces modèles dans différentes situations.

RÉFÉRENCES

Aissaoui-Fqayeh, I., S. El Adlouni, T. B. M. J. Ouarda et A. St-Hilaire (2009). Développement du modèle log-normal non-stationnaire et comparaison avec le modèle GEV non-stationnaire. Journal des sciences hydrologiques 54:6, 1141-1156.

Buchinsky, M. (1998). The dynamics of changes in the female wage distribution in the USA: a quantile regression approach journal of applied econometrics, Vol. 13.

Cannon, A. (2010). A flexible nonlinear modelling framework for nonstationary generalized extreme value analysis in hydroclimatology. Hydrological Process 24, 673–685.

Coles S. (2001). An Introduction to statistical modeling of extreme values. Springer: London.

Chavez-Demoulin, V. et A. Davison (2005). Generalized additive modeling of sample extremes. Applied Statistics 54: 207–222.

Dickey, D.A. et W.A. Fuller. (1979). Distribution of the estimators for autoregressive time series with a unit root. J. Am. Stat. Assoc. 74, 423–431.

El Adlouni, S., T.B.M.J. Ouarda, X. Zhang, R. et. Roy et B. Bobee (2007). Generalized maximum likelihood estimators for the nonstationary generalized extreme value model. Water Resources Research 43: W03410.

El Adlouni, S. et T.B.M.J. Ouarda (2008). Comparison of methods for estimating the parameters of the Non-Stationary GEV Model. Revue des Sciences de l'Eau 21(1): 35-50. ISSN: 1718-8598.

El Adlouni, S. et T.B.M.J. Ouarda (2009). Joint Bayesian Model Selection and Parameter Estimation of the Generalized Extreme Value Model With Covariates Using Birth-Death Markov Chain Monte Carlo. Water resources research 45:W06403.

El Adlouni, S., F.Chebana et B. Bobée (2010). Generalized Extreme Value vs. Halphen System: An exploratory study. Journal of Hydrologic Engineering, Vol 15: 2, pp. 79-89.

Fisher, R.A. (1929). Moments and Product Moments of Sampling Distributions. Proceedings of the London Mathematical Society 30, pp199-238.

Jagger, T. H. and James, B. E. (2006). Climatology Models for Extreme Hurricane Winds near the United States. J. Climate, 19, pp: 3220–3236.

Jenkinson A. F (1955). The frequency distribution of the annual maximum (or minimum) of meteorological elements, Quarterly journal of the royal meteorological society., pp.158– 171.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov Chains and their applications, Biometrika, Vol.57, No.1,97-109.

Hosking J. (1990). L-moments: analysis and estimation of distributions using linear combinations of order statistics. Journal of the Royal Statistical Society, Series B 52, pp:105 124.

Kwiatkowski, D., P.C.B. Phillips, P. Schmidt et Y. Shin (1992). Testing the null of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? J. Econometrics 54, 159–178.

Parey, S., F. Malek, C. Laurent et D. Dacunha-Castelle (2007). Trends and climate evolution: statistical approach for very high temperatures in France. Climatic Change n°81, pp 331-352

Martins, E.S. and Stedinger, J.R. (2000). Generalized maximum likelihood GEV quantile estimators for hydrologic data. Water Resources Research., 36, 737-744.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E.(1953) Equation of state calculations by fast computing machines, The Journal of Chemical Physics, Vol.21,No.6.pp.1087-1092.

Smith, R.L. (1985). Maximum Likelihood Estimation in a Class of Non-Regular Cases. Biometrika 72 pp. 67-92.

Thiele, T.N. (1903). Theory of Observations. C. and E. Layton, London. Annals of Mathematical Statistics 2, pp.165-308.

Zhang, X. B., F. W. Zwiers et G. L. Li (2004). Monte Carlo experiments on the detection of trends in extreme values. J. Clim., 17, pp. 1945-1952.

Önöz, B. et M. Bayazıt (2003). The Power of Statistical Tests for Trend Detection. Turkish J. Eng.Env. Sci., 247-251.

CHAPITRE II: ARTICLE

BAYESIAN ESTIMATION OF THE GEV-B-SPLINES MODEL

B.Nasri[1], S. El Adlouni[2], T.B.M.J Ouarda[1, 3]

[1] Canada Research Chair on the estimation of hydrometeorological variables, INRS-ETE, 490 DE La Couronne, Québec, QC, Canada.

[2] Université de Moncton, Département de mathématique et de statistique NB, Canada E1A 3E9

[3] Masdar Institue of Science and Technology, P.O.Box 54224, Abu Dhabi,UAE

Tel: +1 418 654 3842, Fax: +1 418 654 2600

E-mail: Bouchra.nasri@ete.inrs.ca or touarda@masdar.ac.ae

or salah-eddine.el.adlouni@umoncton.ca

2012

Abstract

The stationarity hypothesis is essential in hydrological frequency analysis and statistical inference. This assumption is often not fulfilled for large observed datasets, especially in the case of hydro-climatic variables. The Generalized Extreme Value distribution with covariates allows to model data in the presence of non-stationarity and/or dependence on covariates. Linear and non-linear dependence structures have been proposed with the corresponding fitting approach. The objective of the present study is to develop the GEV model with B-Spline in a Bayesian framework. A Markov Chain Monte Carlo (MCMC) algorithm has been developed to estimate quantiles and their posterior distributions. The methods are tested and illustrated using simulated data and applied to meteorological data. Results indicate the better performance of the proposed Bayesian method for rainfall quantile estimation according to the BIAS and RMSE criteria especially for high return period events.

Key Words: GEV, Bayesien, B-Spline, Nonlinearity, Covariate, Non-stationarity

1. Introduction

Many fields of modern science and engineering have to deal with rare events with significant consequences. Extreme value theory (EVT) provides the basis for the statistical modeling of such extremes. The main result of EVT shows that the maxima, of independent and identically distributed (iid) events, are asymptotically Generalized Extreme Value (GEV) distributed (Jenkinson, 1955). A number of methods have been proposed to estimate the parameters of the GEV such as the method of moments (MM) (Thiele, 1903; Fisher, 1929), maximum likelihood (ML) (Smith, 1985) and the method of probability weighted moments (Hosking, 1990).

The stationarity assumption is essential to carry out a classical statistical frequency analysis. However, in many fields, such as hydroclimatology, observed data series are not stationary (Dupuis, 2012; Milly et al., 2008). For hydrological datasets, two main types of non-stationarity have been observed due to temporal trends or cycles corresponding to the effect of other covariates. In hydrology, the second kind of non-stationarity, has been largely studied during the last decade through the GEV model with covariates for local frequency analysis (e.g Olsen et al., 1999; Coles, 2001; Cunderlik et al., 2007; El Adlouni et al., 2007; Hundecha et al., 2008; El Adlouni and Ouarda, 2009; Cannon, 2010) and for regional analysis (e.g Cunderlik and Ouarda, 2006; and Leclerc and Ouarda, 2007).

Taking into account the effect of a covariate can be considered in a polynomial form (e.g. Coles, 2001; El Adlouni et al., 2007 and 2008). These polynomial forms for estimating the GEV parameters were developed by the introduction of covariates in a linear or quadratic function. However, the dependence between covariates and variables of interest can have different structures.

Chavez-Demoulin and Davison (2005) suggested the use of semi-parametric functions such as smoothing splines to estimate the relationship between the parameters and the covariates. The smoothing splines are based on the minimization of the penalized sum of the squared errors and the choice of the smoothing parameter (De Boor, 2001). The main disadvantages of this type of function are that inference, often through the confidence bands, is not straightforward and that a smoothing parameter needs to be specified at the beginning (Müller and Wang, 2007). A smoothing-based B-spline function resolves these problems and presents several other advantages.

The B-spline functions are linear combinations of non negative piecewise-polynomial real functions. A B-spline function does not depend on the response variable, or the variable of interest, but depends only on: (i) the support of the covariates, (ii) the number and position of knots and (iii) the degree of B-Spline function (De Boor, 2001). The above advantages of B-Spline functions make it an appropriate option to be used in the GEV model with covariates to estimate the quantiles conditionally on given factors. The GEV model with B-spline called mixed GEV-B-Splines model, is rigorous and flexible and allows the fitting of a large number of dependence structures (e.g. Chavez-Demoulin and Davison, 2005; Padoan and Wand, 2008). Chavez-Demoulin and Davison (2005) describe smooth non-stationary generalized additive modeling for sample extremes, in which spline smoothers are incorporated into models for exceedances over high thresholds with the Generalized Pareto distribution. They developed the maximum penalized likelihood estimation approach with uncertainty assessed by using differences of deviances and bootstrap simulation.

The main objective of the present study is to develop the GEV model with covariates where the dependence structure is represented by B-spline functions in a Bayesian framework. Prior

distributions are proposed and the posterior distribution is simulated through the Metropolis-Hasting (MH) algorithm of the Monte Carlo Markov Chain (MCMC) method.

In the next section, the theoretical development of the Bayesian method of parameter estimation is discussed. A case study is then presented in Section 3. A comparison between the proposed Bayesian approach and classical estimation methods such as the method of moments and the maximum likelihood method is presented in the fourth section. The last section corresponds to the conclusions and recommendations for future work.

2. Bayesian GEV-B-Splines model

2.1. GEV distribution

The extreme value theory introduced by Fisher and Tippett (1928) shows that the limiting distribution of the maximum is one of the following distributions: Gumbel, Fréchet or Weibull. These three distributions can be grouped in a single Generalized Extreme Value (GEV) distribution:

$$F(y,\mu,\sigma,\xi) = \exp\left[-\left(1+\xi\left(\frac{y-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}}\right] \qquad \xi \neq 0$$
$$F(y,\mu,\sigma,\xi) = \exp\left[-\exp\left(-\frac{y-\mu}{\sigma}\right)\right] \qquad \xi = 0 \tag{1}$$

- The Gumbel distribution has two parameters defined on $\mathbb{R}$, the distribution function is obtained by letting $\xi \to 0$;

- The Fréchet distribution has three parameters defined on the interval $\left]-\frac{\sigma}{\xi}+\mu,+\infty\right[$, obtained for $\xi > 0$;
- The Weibull distribution has three parameters defined on the interval $\left]-\infty,\mu+\frac{\sigma}{\xi}\right[$, obtained for $\xi < 0$.

We consider a random variable Y that follows the GEV distribution and t the time before the event $Y > y_T$. Then t is distributed according to a Geometric distribution with a parameter $p = P(Y > y_T)$.

Let $(Y_1, Y_2, ..., Y_n)$ be i.i.d. random variables from the GEV distribution. The probability that $t = k$ *(k > 0)* is given by:

$$\begin{aligned} P(t=k) &= P\left(\bigcap_{i<k} Y_i < y_T\right) P(Y_k > y_T); \qquad k = 1,2,3,... \\ &= (1-p)^{k-1} p \end{aligned} \tag{2}$$

With:

$$T = E(t) = \frac{1}{p} \tag{3}$$

Since the variable Y follows the GEV distribution with distribution function F, equation (3) becomes:

$$E(t) = \frac{1}{P(Y > y_T)} = \frac{1}{1-F(y_T)} \tag{4}$$

So the quantile y_T of the GEV distribution is:

$$y_T = F^{-1}\left(1-\frac{1}{E[t]}\right) = \mu - \frac{\sigma}{\xi}\left[1-\left(\log\left(1-\frac{1}{T}\right)\right)^{-\xi}\right] \qquad (5)$$

In the non-stationary case, the parameters of the GEV are functions of time or other covariates. Consequently, the quantile y_T depends on these covariates. In the present study, the parameters σ and ξ are supposed constant. Let Y be a random variable that follows the $GEV(\mu_{\underline{x}}, \sigma, \xi)$, and $\underline{X} = (X_1, X_2, ..., X_q)$ be a vector of covariates. Let the location parameter of the GEV model be a function of covariates:

$$\mu_{\underline{x}} = \sum_{i=1}^{q} f_i(X_i) = f_1(X_1) + f_2(X_2) + \cdots + f_q(X_q) \qquad (6)$$

where f_i represents the function that describes the relationship between the parameter and the covariate X_i.

In the classical GEV model with covariates, dependence is represented through polynomial functions of linear or quadratic forms. In the following paragraph, the dependence structure in the GEV model with covariates is given by B-Splines. This model will be called GEV-B-Splines.

2.2. The GEV-B-Splines model

The function f_i can be decomposed in the form of basic spline functions:

$$f_i(x_i) = \beta_0 + \sum_{j=1}^{m} \beta_j B_{j,d}(x_i) \qquad (7)$$

where

$$\begin{cases} B_{j,d}(x) = \dfrac{x - x_j}{x_{j+d} - x_j} B_{j,d-1}(x) + \dfrac{x_{j+d+1} - x}{x_{j+d+1} - x_{j+1}} B_{j+1;d-1}(x) \quad for\ j = 0, m-d-2 \\ B_{j,0}(x) = \begin{cases} 1 & if\ x_j \le x < x_{j+1} \\ 0 & otherwise \end{cases} \quad for\ j = 0, m-2 \end{cases} \tag{8}$$

$B_{j,d}(x)$ is a polynomial of degree d on each interval and m is the number of control points.

Hence, equation (6) becomes

$$\mu_{\underline{x}} = \sum_{i=1}^{q} f_i(x_i) = \sum_{i=1}^{q} \left(\beta_{i,0} + \sum_{j=1}^{m} \beta_{i,j} B_{j,d}(x_i) \right) \tag{9}$$

The matrix form of equations (8) and (9) gives

$$\beta_0 = \begin{pmatrix} \beta_{1,0} \\ \vdots \\ \beta_{q,0} \end{pmatrix}_{(q,1)} \qquad \beta = \begin{pmatrix} \beta_{1,1} \\ \vdots \\ \beta_{q,m} \end{pmatrix}_{(q \times m,1)} \qquad B = \left(B_{1,1}(x_1), \ldots, B_{p,m}(x_q) \right)_{(1,q \times m)} \tag{10}$$

$$\mu_{\underline{x}} = \sum_{j} (1 \quad B) * \begin{pmatrix} \beta_0 \\ \beta \end{pmatrix} \tag{11}$$

where 1 is the unit vector of size q.

2.3. The GEV-B-Splines model in the Bayesian framework

The GEV-B-Splines is considered in a fully Bayesian framework. For a given parameter prior distribution, $\pi(\theta)$, Bayes theorem allows the definition of the posterior distribution :

$$f(\theta \mid y) = \frac{f(y \mid \theta) * \pi(\theta)}{f(y)} \tag{12}$$

where

$$\theta = (\mu_t, \sigma, \xi) = (\boldsymbol{\beta}, \sigma, \xi) \quad \text{where } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta \end{pmatrix}$$

is the vector of the parameters, and β_0 and β are the vector of the hyper-parameters of the location parameter.

Martins and Stedinger (2000) proposed the Beta (6, 9) distribution as a prior distribution for the shape parameter of a stationary GEV model, in order to avoid irrational estimations of the shape parameter. In the present study, we considered an equivalent prior for the shape parameter, it is the normal distribution with mean 0.1 and variance 0.12.

El Adlouni et al. (2007) adopted this prior distribution for the GEV model with covariates with polynomial dependence. Other studies have suggested adopting the normal distribution to model the hyper parameters of the location parameter for the GEV model with covariates and B-Spline dependence (e.g. Padoan and Wand, 2009; Neville et al., 2011).

$$\boldsymbol{\beta} \sim N(0, \Sigma_\beta \times \mathrm{I}) \tag{13}$$

For the scale parameter, we used a non informative prior distribution $1/\sigma$

The posterior distribution of θ is written as follows:

$$f(\theta \mid y) = \frac{f(y \mid \theta) * \pi(\theta)}{f(y)} = \frac{f(y \mid \theta) * \pi(\beta) * \pi(\xi) * \pi(\sigma)}{f(y)} \tag{14}$$

then

$$f(\theta \mid y) \propto \frac{1}{\sigma}\left\{1+\xi\left(\frac{y-\sum_{j}(1 \quad B)*\begin{pmatrix}\beta_0\\ \beta\end{pmatrix}}{\sigma}\right)\right\}^{-\frac{1}{\xi}-1} * \exp\left(-\left(1+\xi\left(\frac{y-\sum_{i}(1 \quad B)*\begin{pmatrix}\beta_0\\ \beta\end{pmatrix}}{\sigma}\right)\right)^{-\frac{1}{\xi}}\right) \quad (15)$$

$$*\left(2\pi \det\left(\Sigma_\beta\right)\right)^{-\frac{k}{2}} * \exp\left(-\frac{\|\beta\|^2}{2\sigma^2}\right) * \frac{1}{\sigma_\xi\sqrt{2\pi}} * \exp\left(-\frac{(\xi-0.1)^2}{2*\sigma_\xi^2}\right) * \frac{1}{\sigma}$$

The posterior distribution $f(\theta \mid y)$ is a function of the hyperparameters $\sigma, \Sigma_\beta, \sigma_\xi, \beta, \beta_0, \xi$.

Considering a simple case of one covariate and m = 1 and d = 1, equation (15) becomes:

$$f(\theta \mid y) \propto \frac{1}{\sigma}\left\{1+\xi\left(\frac{y-\left(\beta_{1,0}+B_{1,1}(x)*\beta_{1,1}\right)}{\sigma}\right)\right\}^{-\frac{1}{\xi}-1} * \exp\left(-\left(1+\xi\left(\frac{y-\left(\beta_{1,0}+B_{1,1}(x)*\beta_{1,1}\right)}{\sigma}\right)\right)^{-\frac{1}{\xi}}\right) \quad (16)$$

$$*\left(2\pi \det\left(\Sigma_\beta\right)\right)^{-\frac{k}{2}} * \exp\left(-\frac{\left(\beta_{1,0}^2+\beta_{1,1}^2\right)}{2\sigma^2}\right) * \frac{1}{\sigma_\xi\sqrt{2\pi}} * \exp\left(-\frac{(\xi-0.1)^2}{2*\sigma_\xi^2}\right) * \frac{1}{\sigma}$$

where

$$B_{1,1}(x) = \frac{x-x_1}{x_2-x_1} * B_{1,0}(x) + \frac{x_3-x}{x_3-x_2} * B_{2,0}(x) \quad (17)$$

σ and ξ are the parameter set by the prior distribution. To estimate the above function, initial values of the parameters $\Sigma_\beta, \sigma_\xi, \boldsymbol{\beta}$ then should be given in order to simulate their joint posterior distribution by a MCMC algorithm. The marginal distributions of the parameters can be deduced by integrating equation (15), with respect to the rest of the parameter vector:

$$f(\theta \mid y) \propto \int f(y \mid \theta) * f(\theta) \partial\theta \quad (18)$$

The following section presents the details of the proposed MCMC algorithm to estimate the GEV-B-Splines parameter and quantile distributions.

2.4. MCMC algorithm for the GEV-B-Splines model

The MCMC method constitutes an alternative to the numerical methods, especially in Bayesian statistical analysis. The basic idea of the MCMC method is, for each parameter, to construct a Markov chain with the posterior distribution being a stationary and ergodic distribution. After running the Markov chain, of size N, for a given burn-in period N_0, one obtains a sample from the posterior distribution $f(\theta|y)$. One popular method for constructing a Markov chain is via the Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953 and Hastings, 1970). For the Generalized maximum likelihood (GML) method, we simulated the realizations from the posterior distribution by way of a single-component MH algorithm (Gilks et al., 1996). Each parameter was updated using a random-walk Metropolis algorithm with a Gaussian proposal density centered at the current state of the chain. Some methods to assess the convergence of the MCMC methods make it possible to determine the length of the chain and the burn-in time such as the Raftery & Lewis diagnostic (Raftery and Lewis, 1992; 1995) and subsampling methods (El Adlouni et al. 2006). In all cases, the convergence methods indicated that the Markov chains converged within a few iterations. In this study, we considered chains of size $N = 15000$ and a burn-in period of $N_0 = 8000$ runs. In every case, a sample of $N - N_0 = 7000$ values is collected from the posterior of each of the elements of θ. The GML corresponds to the mode of the empirical posterior distribution obtained from the $N - N_0$ values generated by the MCMC algorithm.

The MCMC algorithm produces also the conditional quantile distribution for an observed value x_0, of the covariate X_t. Indeed, for each iteration i of the MCMC algorithm $i = 1,\ldots,N$, the quantiles with non-exceedance probability $1-p$, $x_{p,x_0}^{(i)}$ corresponding to the parameter vector $\left(\mu_{x_0}^{(i)},\sigma^{(i)},\xi^{(i)}\right)$, are computed using the inverse of the cumulative distribution function of the GEV distribution:

$$y_{p,x_0}^{(i)} = \mu_{x_0}^{(i)} - \frac{\sigma}{\xi^{(i)}}\left[1-\left(\log(1-p)\right)^{-\xi^{(i)}}\right] \tag{19}$$

Where $\mu_{x_0}^{(i)}$ is the position parameter conditional on the particular value x_0 of X_t. Several statistical characteristics of the conditional quantile distribution can be determined from the values $x_{p,x_0}^{(i)}, i = N_0,\ldots,N$, such as the mean, the mode or the confidence interval. The main steps of the MH algorithm can be summarized as follows (El Adlouni and Ouarda. 2009):

(1) Choose a proposal distribution q,

(2) Given the current state u, generate u^* from $q(.\,|\,u)$,

(3) Accept u^* with probability:

$$\rho\left(u,u^*\right) = \min\left\{1,\frac{\pi\left(u^*\right)}{\pi(u)}\frac{q\left(u\,|\,u^*\right)}{q\left(u^*\,|\,u\right)}\right\}$$

3. Case Study

3.1. Dataset

The proposed model is considered to model the maximum annual rainfall (MAR) at Randsburg station (047253), California for the period of 1938-2007. The Randsburg station is located in the south east of the state of California (35.37°N, 117.65°W). Figure 1 illustrates the geographic location of the Randsburg station. Figure 2 shows the 70-year variation of MAR at Randsburg Station.

Figure 1

Figure 2

We consider the 70-year Southern Oscillation Index (SOI) and Pacific Decadal Oscillation (PDO) time series as covariates for MAR non-stationary quantile estimation. The SOI and PDO describe the pressure and temperature anomalies over the Pacific Ocean and have a clear impact on water systems in North America (e.g Brown and Comrie, 2002; Canon, 2010). By using SOI and PDO as covariates in estimating the parameters of the GEV-B-Splines model, we will take into account the effect of multiannual climate fluctuations on extreme rainfall events. We first apply the Mann Kendall test to examine the existence of non-stationarity (Trend) in MAR time series. The result shows that the MAR is not stationary at 1% significant level. The Spearman's rho correlation coefficient between the covariates and MAR is -0.52 and 0.51 for SOI and PDO respectively. These values are significant at the 5% level. Figure 3 shows the variation of maximum annual rainfall against SOI and PDO.

Figure 3

3.2. Model development

For model development, the following function is first fitted:

GEV-B-Splines ($MAR \sim GEV\left(f_1(SOI) + f_2(PDO), \sigma, \xi\right)$)

f_1, f_2 are independent spline functions, for which the degree and the number of nodes should be determined. In this application the number of nodes and the degree of the function are both chosen to take the value 3 (see table 1).

Table 2 shows the GEV-B-Splines parameters fitted to SOI and PDO time series using a Bayesian method. Figures 4 and 5 show the estimated 2, 20 and 50-year return period maximum rainfall quantiles as function of the covariates (SOI and PDO). It can be seen that, generally, the SOI has a negative correlation with precipitation, while PDO is positively correlated with precipitation. The negative values of SOI (e.g. El Nino phase) and positive values of PDO (Warm Phase of PDO) coincide with the relatively high MAR observations. MAR quantiles increase slowly with increasing PDO values and then increase exponentially for PDO values greater than 1. On the other hand, different inflexion points, in the relationship between SOI and MAR are observed (for example at SOI= -1.5, SOI=0 and SOI= 1.5), indicating a more complex relationship between SOI and MAR than between PDO and MAR.

Table 2

Figure 4

Figure 5

4. Parameter estimation comparison

In this section, we propose a comparison of the Bayesian parameter estimation method for the GEV-B-Splines model (BAYES) and other estimation methods such as the conventional method of moments (MM) and the method of maximum likelihood (ML). The theoretical background of these two methods for the GEV-B-Splines model is presented in Appendices 1 and 2, respectively. The comparison of these methods is carried out based on a simulation of MAR-SOI relationship only. The quantile with a non-exceedence probability $1-p$ is computed for the maximum SOI using the parameters given by the Bayesian method. The objective is to compare the quantile estimation methods for the quantiles estimated from 1000 samples of size n = 70 generated from each estimation method. The parameter values chosen for simulation are $\left((\beta_0 = -44.8, \beta_1 = -2.6, \beta_2 = 81.6, \beta_3 = 53.6, \beta_4 = 54.4); \sigma = 7.2; \xi = -0.17\right)$.To compare the Bayesian method with other methods, we consider the parameters of GEV-B-Splines for the MAR with SOI values as the covariate. The comparison is carried out using the BIAS and the root mean square error (RMSE) of quantile estimations at non-exceedance probabilities, $1-p$ = 0.5, 0.8, 0.9, 0.99 corresponding to return periods of 2, 5, 10, 100. The results are given in Table 3.

Table 3

Results show that the Bayesian estimation for the GEV-B-Splines model in all cases represents the best results. However, this estimation method requires large time-consuming numerical calculations and does not meet a convergence point easily. For our case, the MCMC method details, such as the choice of numerical method burning period and number of iterations are the key points to the convergence of the MCMC algorithm. On the other hand, even if the method of

moments is the easiest method to implement, the corresponding results are largely unsatisfactory. The method of maximum likelihood, however, is a compromise between the other two methods. It is interesting to note that for the case of low return periods, i.e. T = 2, T = 5 and T = 10 years, the maximum likelihood method gives almost comparable results with the Bayesian estimation. However, the error of the ML method increases rapidly with the increase in the return period and the method becomes increasingly less effective. Therefore, the Bayesian method leads to a superior performance for the estimation of the extreme rainfall quantiles for all return periods. The Bayesian method offers also a general framework to combine observed and subjective information and the possibility to estimate the entire predictive distribution of the parameters and quantiles.

5. Conclusions and Recommendations

Statistical risk assessment is of great importance in hydrology and many other fields of applied statistics. The last two decades have witnessed the development of a number of statistical modeling approaches for extreme values in the presence of non-stationarity or dependence on covariates. The GEV-B-Splines model which takes into account the non-stationarity and nonlinearity offers a great flexibility and takes into account the heavy tailed character of the extreme distribution. The present study proposes a Bayesian estimation framework of the GEV-B-splines model for hydro-meteorological variables. The Bayesian approach is general, flexible and connected with the decision theory. It combines observed and prior information and estimates the entire posterior distribution of the parameters and quantiles.

Results of the simulated data show the advantage of the proposed method for quantile estimation of an extreme variable such as maximum rainfall especially for high return periods.

The evaluation for the quantile uncertainty using BIAS and RMSE criteria also indicated the superiority of the proposed method in comparison to other estimation methods, especially for high return period quantile estimation. However, the uncertainty of quantile estimation of low return periods does not show a significant difference between the bayesian and the maximum likelihood method. On the other hand, one can see that the numerical calculation is the main disadvantage of these types of models when the number of covariates increases which may lead to divergence problem. The quantile regression model can be a good alternative to overcome this problem (Buchinsky, 1998; Jagger and Elsner, 2006). Therefore, future work can focus on the comparison of extreme value models with regression quantiles in order to use different covariates in quantile estimations.

REFERENCES

Brown, D. P., and Comrie, A. C. (2002). Sub-regional seasonal precipitation linkages to SOI and PDO in the Southwest United States. Atmospheric Science Letters, 3(2-4), 94-102. doi:10.1006/asle.2002.0057

Buchinsky, M. (1998). The dynamics of changes in the female wage distribution in the USA: a quantile regression approach. Journal of Applied Econometrics, Vol. 13.

Cannon, A. (2010). A flexible nonlinear modeling framework for nonstationary generalized extreme value analysis in hydroclimatology. Hydrological Process, 24, 673–685.

Chavez-Demoulin, V. and Davison, A. (2005). Generalized additive modeling of sample extremes. Applied Statistics 54: 207–222.

Cunderlik, J.M. and Ouarda, T.B.M.J. (2006). Regional Flood- Duration-Frequency Modeling in a Changing Environment. Journal of Hydrology 318:276-291.

Cunderlik, J.M., Jourdain, V., Ouarda, T.B.M.J. and Bobée, B. (2007). Local NonStationary Flood-Duration-Frequency Modeling. Canadian Water Resources Journal, 32(1):43-58.

Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values. Springer: London.

De Boor, C. (2001). A practical guide to spline. Springer Ser. Stat., 208 pp., Springer, London. USA.

Dupuis, D.J. (2012). Modeling waves of extreme temperature: The changing tails of four cities.Journal of the American Statistical Association, 107, 24-39.

El Adlouni, S., Favre, A.C. and Bobée, B. (2006). Comparison of methodologies to assess the convergence of Markov Chain Monte Carlo methods. Computational Statistics and Data Analysis, 50(10): 2685-2701.

El Adlouni, S., Ouarda ,T.B.M.J., Zhang, X., Roy, R. and Bobee, B. (2007). Generalized maximum likelihood estimators for the nonstationary generalized extreme value model. Water Resources Research 43: W03410.

El Adlouni, S. and Ouarda, T.B.M.J. (2008). Comparison of Methods for Estimating the Parameters of the Non-Stationary GEV Model. Revue des Sciences de l'Eau 21(1): 35-50. ISSN: 1718-8598.

El Adlouni, S. and Ouarda, T.B.M.J. (2009). Joint Bayesian Model Selection and Parameter Estimation of the Generalized Extreme Value Model With Covariates Using Birth-Death Markov Chain Monte Carlo. Water Resources Research 45:W06403.

Fisher, R.A., Tippett, L.H.C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proceedings of the Cambridge Philosophical Society, V. 24, 180-190.

Fisher, R.A. (1929). Moments and Product Moments of Sampling Distributions. Proceedings of the London Mathematical Society 30:199-238.

Gilks, W. R., Richardson, S., and Spiegelhalter. D. J. (1996). Markov chain Monte Carlo in practice. Chapman and Hall, London, UK.

Greenwood, J.A., Landwehr, J.M, Matalas, N.C. et Wallis, J.R. (1979). Probability weighted moments: definition and relation to parameters of several distributions expressible in inverse form. Water Ressource. Res., 15, 1049-1054.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov Chains and their applications, Biometrika, Vol.57, No.1,97-109.

Hosking, J. (1990). L-moments: analysis and estimation of distributions using linear combinations of order statistics. Journal of the Royal Statistical Society, Series B 52: 105 124.

Hundecha, Y., Ouarda, T.B.M.J. and Bárdossy, A. (2008). Regional estimation of parameters of a rainfall–runoff model at ungauged watersheds using the spatial structures of the parameters within a canonical physiographic-climatic space. Water Resources Research, 44, W01427. doi:10.1029/2006WR005439.

Jagger, T. H. and James, B. E. (2006). Climatology Models for Extreme Hurricane Winds near the United States. J. Climate, 19: 3220–3236.

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) of meteorological elements, Q. J. R. Meteorol. Soc., 81: 158– 171.

Leclerc, M. and Ouarda, T.B.M.J. (2007). Non-Stationary Regional Flood Frequency Analysis at Ungauged Sites. Journal of Hydrology 343:254-265, doi: 10.1016/j.jhydrol.2007.06.021.

Martins, E.S. and Stedinger, J.R. (2000). Generalized maximum likelihood GEV quantile estimators for hydrologic data. Water Resour. Res., 36, 737-744.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E.(1953) Equation of state calculations by fast computing machines, The Journal of Chemical Physics, Vol.21,No.6.pp.1087-1092.

Milly, P.C.D., Betancourt, J., Falkenmark ,M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier, D.P. and Stouffer, R.J. (2008). Stationarity Is Dead: Whither Water Management? Science 319.pp. 573–574.

Müller, H.G. and Wang, J.L. (2007). Density and failure rate estimation. Encyclopedia of Statistics in Quality and Reliability. pp. 517-522.

Neville, S., Palmer, M.J. and Wand, M.P. (2011). Generalized extreme value geoadditive model analysis via mean field variational Bayes, Australian and New Zealand Journal of Statistics, 53(3): 305–330.

Olsen, J.R., Stedinger, J.R., Matalas, N.C., and Stakhiv, E.Z. (1999). Climate Variability and Flood Frequency Estimation for the Upper Mississippi and Lower Missouri Rivers. Journal of the American Water Resources Association 35(6):1509-1524.

Padoan, S.A. and Wand, M.P. (2008). Mixed model-based additive models for sample extremes. Statistics and Probability Letters 2850–2858

Raftery, A.E. and Lewis, S.M. (1992). One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. Statistical Science, 7, 493-497.

Raftery, A.E. and Lewis, S.M. (1995). The number of iterations, convergence diagnostics and generic Metropolis algorithms. In Practical Markov Chain Monte Carlo (W.R. Gilks, D.J. Spiegelhalter and S. Richardson, eds.). London, U.K.: Chapman and Hall.

Smith, R.L. (1985). Maximum Likelihood Estimation in a Class of Non-Regular Cases. Biometrika 72:67-92.

Thiele, T.N. (1903). Theory of Observations. C. and E. Layton, London. Annals of Mathematical Statistics 165-308.

List of Tables

Table 1 : Choice of parameters of the B-spline function.

Table 2: Bayesian estimation of the parameters of the model

Table 3: Comparison of estimation methods

Figure Captions

Figure 1: Geographic location of the Randsburg station

Figure 2: Variation of maximum annual rainfall

Figure 3: Annual maximum rainfall against SOI and PDO index

Figure 4: GEV-B-Splines estimators of the 2, 20 and 50-year return period quantiles conditional upon SOI

Figure 5: GEV-B-Splines estimators of the 2, 20 and 50-year return period quantiles conditional upon PDO

Table1: Choice of parameters of the B-spline function.

(Number of nodes , Degree)	Negative Maximum Likelihood
(1,1)	387.21
(1,2)	353.30
(1,3)	356.18
(1,4)	274.58
(2,1)	287.98
(2,2)	378.01
(2,3)	293.57
(2,4)	263.47
(3,1)	382.51
(3,2)	331.61
(3,3)	242.195
(3,4)	261.65
(4,1)	397.31
(4,2)	307.35
(4,3)	371.24

Table 2: Bayesian estimation of the parameters of the model

Parameter	Climate Index	
	SOI	PDO
$\beta_{i,0}$	-114.728	11.355
$\beta_{i,1}$	48.112	44.616
$\beta_{i,2}$	147.695	0
$\beta_{i,3}$	116.138	4.556
$\beta_{i,4}$	158.933	7.932
$\beta_{i,5}$	0	-30.110
σ	5.678	7.566
ξ	-0.124	0.145

For i= 1,2

Table 3: Comparison of estimation methods

Probability	BIAS			RMSE		
	BAYES	MM	ML	BAYES	MM	ML
0.5	0.020	-0.075	0.052	0.403	0.715	0.435

0.8	-0.060	-0.094	0.090	0.418	0.901	0.514
0.9	-0.148	0.249	-0.177	0.450	1.847	1.525
0.99	-0.182	-0.655	-0.448	0.826	3.128	2.879

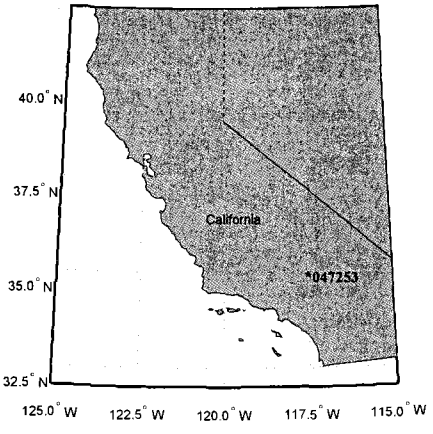

Figure 1: Geographic location of the Randsburg station

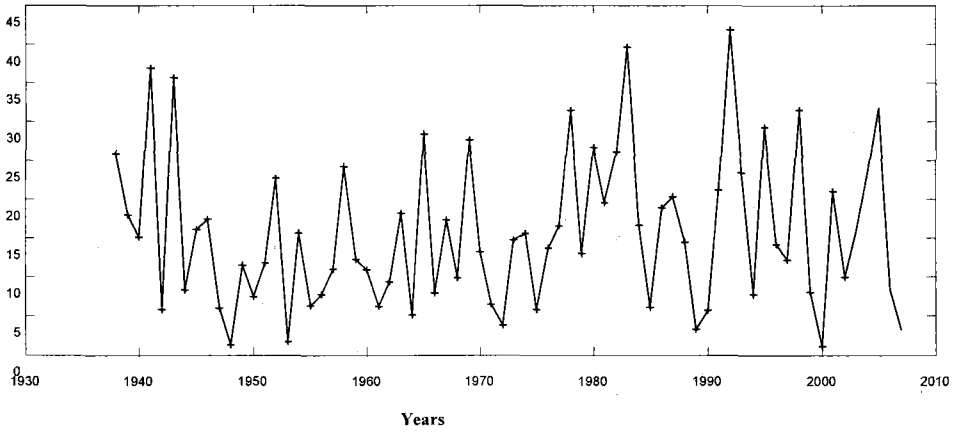

Figure 2 : Variation of maximum annual rainfall

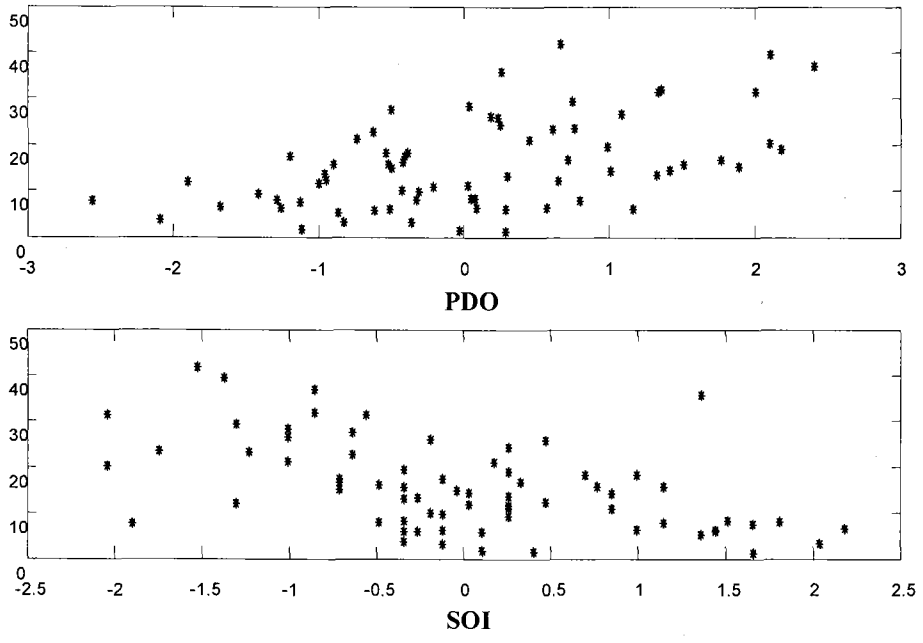

Figure 3: Annual maximum rainfall against SOI and PDO index

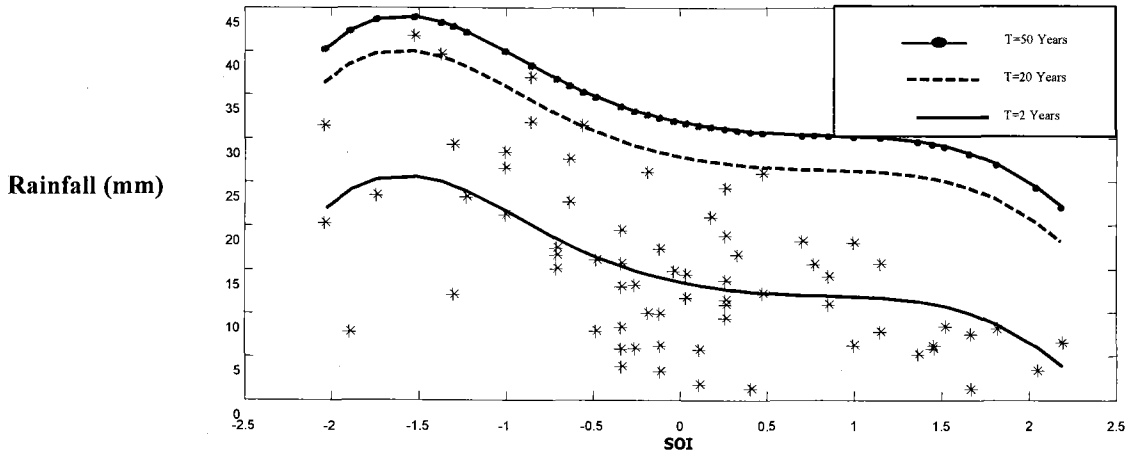

Figure 4: GEV-B-Splines estimators of the 2. 20 and 50-year return period quantiles conditional upon SOI

Rainfall (mm)

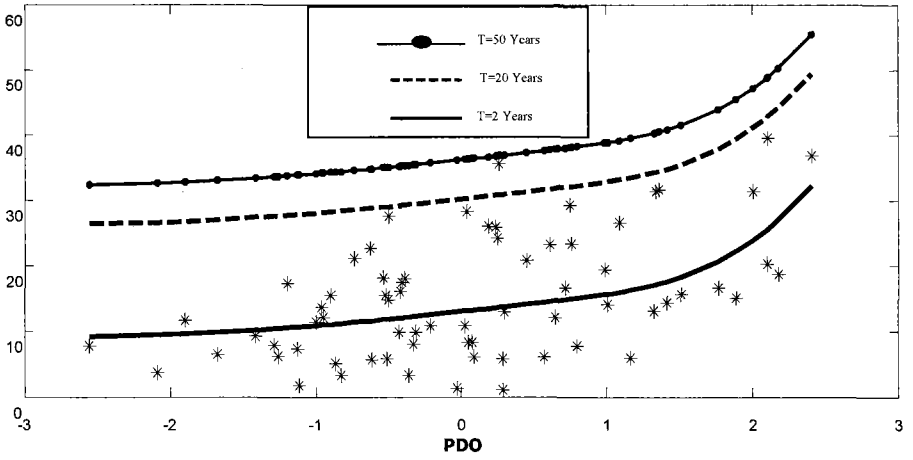

Figure 5: GEV-B-Splines estimators of the 2. 20 and 50-year return period quantiles conditional upon PDO

Appendix 1: GEV-B-Splines moment

Let Y be a random variable that follows a GEV distribution therefore:

$$Y \sim GEV(\mu_{\underline{x}}, \sigma, \xi) \tag{A1}$$

With $\mu_{\underline{x}} = f(X)$ is a parameter that depends on a covariate X.

$$f(x) = \sum_{i=0}^{n} \beta_i B_i(x) = \beta_0 + \sum_{i=1}^{n} \beta_i B_i(x) \tag{A2}$$

B is a spline basis function.

where

$$f(x) - \sum_{i=1}^{n} \beta_i B_i(x) = \beta_0 \tag{A3}$$

And thus

$$\mu_{\underline{x}} - \sum_{i=1}^{n} \beta_i B_i(x) = \beta_0 \tag{A4}$$

Then

$$Z = Y - \sum_{i=1}^{n} \beta_i B_i(x) \sim GEV(\beta_0, \sigma, \xi) \tag{A5}$$

The following equations are used to estimate the parameters $\beta_0 . \sigma . \xi$:

$$\hat{\xi} = 7{,}8590c + 2{,}9554c^2 \tag{A6}$$

$$\hat{\sigma} = \frac{l_2 \hat{\xi}}{\left(1 - 2^{-\hat{\xi}}\right)\Gamma(1+\hat{\xi})} \tag{A7}$$

$$\beta_0 = l_1 - \frac{\hat{\sigma}}{\hat{\xi}}\left\{1 - \Gamma\left(1+\hat{\xi}\right)\right\} \tag{A8}$$

With

$$c = \frac{2}{3+t_3} - \frac{\log(2)}{\log(3)}, \; t_3 = \sum_{i=1}^{n} \frac{c_i * y_{(i)} + \frac{\overline{y}}{n}}{l_2}, \; c_i = 6\frac{(i-1)(i-2)}{n(n-1)(n-2)} - 6\frac{(i-1)}{n(n-1)} \tag{A9}$$

$$l_1 = b_0 = \frac{1}{n}\sum_{j=1}^{n} y_j \; . \; l_2 = \frac{2}{n}\sum_{j=2}^{n} \frac{j-1}{(n-1)} y_i - \frac{1}{n}\sum_{j=1}^{n} y_i$$

See Greenwood et al. (1979). Hosking (1990). El Adlouni and Ouarda (2007).

The other β_i values are estimated using the linear regression between Y and the basis matrix B of the B-spline corresponding to the covariates.

Appendix 2: GEV-B-Splines Maximum likelihood

Let Y be a random variable that follows a GEV distribution therefore:

$$Y \sim GEV(\mu_{\underline{x}}, \sigma, \xi) \tag{A10}$$

With $\mu_{\underline{x}} = f(X)$ is a parameter that depends on a covariate X.

$$f(x) = \sum_{i=0}^{n} \beta_i B_i(x) = \beta_0 + \sum_{i=1}^{n} \beta_i B_i(x) \tag{A11}$$

B is a spline basis function.

The maximum likelihood (ML) function is written as

$$L_n\left(\underline{y} \mid \mu_{\underline{x}}, \sigma, \xi\right) = \prod_{t=1}^{n_1} \frac{1}{\sigma} \exp\left\{-\left[1-\xi\left(\frac{y_t-\mu_t}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} * \left[1-\xi\left(\frac{y_t-\mu_t}{\sigma}\right)\right]^{-\left(1-\frac{1}{\xi}\right)}$$
$$* \prod_{t=n_1+1}^{n} \frac{1}{\sigma} \exp\left\{-\left(\frac{y_t-\mu_t}{\sigma}\right)\right\} * \exp\left\{-\exp\left[-\left(\frac{y_t-\mu_t}{\sigma}\right)\right]\right\} \tag{A12}$$

n_1 is the number of observations when $\xi \neq 0$.

In the case of $\xi \neq 0$. the log-likelihood function is:

$$l_n\left(\underline{y}; \mu_{\underline{x}}, \sigma, \xi\right) = -\mathrm{n}\log(\sigma) - \sum_{t=1}^{n}\left[1-\xi\left(\frac{y_t-\mu_t}{\sigma}\right)\right]^{-\frac{1}{\xi}}$$
$$-\sum_{t=1}^{n}\left(1-\frac{1}{\xi}\right)\log\left[1-\xi\left(\frac{y_t-\mu_t}{\sigma}\right)\right] \tag{A13}$$

The ML estimators are the solution of an equation system formed by setting to zero the partial derivate of (A12) with respect to each parameter.

In the case of one covariate and m = 1. p = 1. we have 4 parameters to estimate $(\beta_0, \beta_1, \sigma, \xi)$.

$$l_n\left(\underline{y};\mu_{\underline{x}},\sigma,\xi\right) = -\mathrm{n}\,\log(\sigma) - \sum_{t=1}^{n}\left[1-\xi\left(\frac{y_t-\left(\beta_0+B_{1,1}(x_t)*\beta_1\right)}{\sigma}\right)\right]^{-\frac{1}{\xi}}$$
$$-\sum_{t=1}^{n}\left(1-\frac{1}{\xi}\right)\log\left[1-\xi\left(\frac{y_t-\left(\beta_0+B_{1,1}(x_t)*\beta_1\right)}{\sigma}\right)\right] \qquad \text{(A14)}$$

The ML estimators of the parameter $(\beta_0,\beta_1,\sigma,\xi)$ are the solution of the following system:

$$\begin{cases}
\dfrac{\partial l_n}{\partial \xi}=0 \Rightarrow \displaystyle\sum_{t=1}^{n}\left\{\ln(w_t)\left[1-\xi-w_t^{-\frac{1}{\xi}}\right]+\frac{1-\xi-w_t^{-\frac{1}{\xi}}}{w_t}\xi\left(\frac{y_t-\left(\beta_0+B_{1,1}(x_t)*\beta_1\right)}{\sigma}\right)\right\}=0 \\
\dfrac{\partial l_n}{\partial \sigma}=0 \Rightarrow -n+\displaystyle\sum_{t=1}^{n}\left[\frac{1-\xi-w_t^{-\frac{1}{\xi}}}{w_t}\left(\frac{y_t-\left(\beta_0+B_{1,1}(x_t)*\beta_1\right)}{\sigma}\right)\right]=0 \\
\dfrac{\partial l_n}{\partial \beta_0}=0 \Rightarrow \displaystyle\sum_{t=1}^{n}\frac{1-\xi-w_t^{-\frac{1}{\xi}}}{w_t}=0 \\
\dfrac{\partial l_n}{\partial \beta_1}=0 \Rightarrow \displaystyle\sum_{t=1}^{n}B_{1,1}(x_t)\frac{1-\xi-w_t^{-\frac{1}{\xi}}}{w_t}=0
\end{cases} \qquad \text{(A15)}$$

Where

$$w_t=\left[1-\xi\left(\frac{y_t-\left(\beta_0+B_{1,1}(x_t)*\beta_1\right)}{\sigma}\right)\right]^{-\frac{1}{\xi}} \qquad \text{(A16)}$$

Numerical methods such as Newton-Raphson must be used to solve the system (A15). The initial values correspond to the estimators given by the method of moments.