Un modèle de précipitations printanières en contexte d'inondation

Fateh Chebana [*] Hélène Cossette [†] Alexandre Lepage [†‡] Étienne Marceau [†]

13 avril 2021

Résumé

Dans cet ouvrage, nous partons de [Jalbert et al., 2019] pour proposer une modèle alternatif de prédiction des pluies printanières totales dans un contexte de modélisation des inondations saisonnières. Notre approche bonifie les modèles DDF (*Dept, Duration, Frequency*) communément utilisés en pratique en ajoutant les notions de processus de renouvellement alterné avec récompense tel que présenté dans [Salvadori and De Michele, 2006] et [Salvadori and De Michele, 2007]. Comme les changements climatiques sont d'actualité, on considère la possibilité que ce processus soit non-stationnaire. De plus, nous faisons une analyse exhaustive de la dépendance entre les différentes variables en jeux à l'aide de la théorie des copules telle que décrite dans [Joe, 1997] et [Nelsen, 2006].

Keywords— Processus de renouvellement alterné avec récompense, processus non stationnaire, dépendance, copules, valeurs extrêmes, précipitations, distribution Pareto généralisée, excès de seuil.

1 Introduction

Autant en assurance qu'en hydrologie, le sujet des inondations suscite l'intérêt des chercheurs qui tentent de modéliser ces événements afin de mieux se préparer à d'éventuelles catastrophes. Entre autres, le débordement du lac Champlain en 2011 a suscité l'intérêt de [Riboust and Brissette, 2016] qui a chercher à connaître les éléments déclencheurs d'une telle catastrophe. Lors de son étude, il arriva à la conclusion que, bien que la fonte des neiges soit une variable explicative importante, c'est l'accumulation de précipitations extrêmes dans la période de mars à juin qui a été la cause principale dans ce cas.

Parler de [Kao and Govindaraju, 2007]. Lire aussi [Vandenberghe et al., 2010]

Suite à la publication de [Riboust and Brissette, 2016], [Jalbert et al., 2019] a cherché à prédire l'accumulation des pluies printanières du lac Champlain afin d'estimer l'amplitude maximale qu'un tel événement aurait pu avoir et de calculer l'espérance du temps qui s'écoulera avant qu'un incident d'une telle ampleur survienne à nouveau.

Le modèle ainsi conçu sépare la modélisation des pluies en deux composantes. Une portion régulière qui représente la quantité de pluie totale tombée selon les normes saisonnières et une portion extrême où on considère

[*]Institut national de recherche scientifique, Québec, Qc, Canada
[†]École d'actuariat, université Laval, Québec, Qc, Canada
[‡]Tel.: 581-984-3704, E-mail: alexandre.lepage.3@ulaval.ca

les jours où la quantité de pluie tombée dépasse un certain seuil. Tandis que la première est bien modélisée avec une loi normale, la seconde nécessite plus de travail. En effet, afin de modéliser la quantité quotidienne de pluie tombée dans les cas extrêmes, [Jalbert et al., 2019] utilise la loi Pareto généralisée, aussi connue sous l'appellation *Peaks-Over-Threshold*(POT). Cependant, l'estimateur du paramètre de forme de la loi POT calculé sur les valeurs quotidiennes faisaient en sorte que la distribution avait un support fini et la probabilité qu'un événement de l'envergure de 2011 se produise était quasiment nulle. Conséquemment, [Jalbert et al., 2019] proposa une extension du modèle POT où il divise la quantité de pluie tombée lors d'une journée de pluie extrême par la proportion que cette quantité représente sur l'ensemble des précipitations tombées au cours d'une période de pluie continue. Se faisant, il n'y a plus de problème de borne supérieure à la distribution de la sévérité.

Le modèle que nous proposons est une alternative à celui de [Jalbert et al., 2019]. Il reprend le concept de regroupement des jours en périodes de pluie continue, mais il adopte une approche légèrement plus intuitive en s'inspirant grandement de la méthodologie proposée dans [Zhang and Singh, 2019] et [Shaw et al., 2010]. Comme il est expliqué, entre autre, au chapitre 9.6 de [Shaw et al., 2010], les modèles de précipitations utilisés dans la pratique, communément appelés modèles DDF (Dept, Duration, Frequency), sous-tendent les variables de sévérité de durée et de fréquence, lesquelles ont toutes leur importance. Or, [Jalbert et al., 2019] se concentre principalement sur la sévérité. La fréquence est modélisée avec une loi de Poisson conditionnelle-gamma, mais l'aspect de la durée est négligé. Notre approche, quant à elle, s'intéresse à toutes ces variables aléatoires et aux relations qui les unissent conformément à [Zhang and Singh, 2019]. Nous incorporons des notions des processus de renouvellement aux concepts communément utilisés en hydrologie afin d'améliorer la précision du modèle de fréquence et les liens de dépendance sont modélisés à l'aide de la théorie des copules. Également, une attention particulière est apportée à la non-stationnarité des distributions de probabilité afin de prendre en compte le contexte des changements climatiques.

L'article est découpé comme suit : la section 2 présente le modèle proposé. Puis, deux études de cas sont décrites dans la section 3. Une analyse de résultats est réalisée dans la section 3.3. Un algorithme de simulation destiné à étudier le modèle proposé est présenté dans l'annexe C. Finalement, des éléments de discussion sont abordés dans la section 4.

2 Le modèle proposé

L'intuition derrière le modèle proposé est de considérer les périodes de précipitations continues comme des événements uniques en agrégeant la quantité de pluie tombée lors de celles-ci. Se faisant, non seulement on considère les jours où le volume d'eau tombé sort de l'ordinaire, mais on considère également les séquences où le nombre de jours de pluie continue est anormalement élevé. Pour se faire, les processus de renouvellement alternés sont utilisés pour modéliser le nombre d'événements extrêmes. Également, on s'intéressera à la présence de dépendance unissant les v.a. en jeu.

2.1 Description du modèle

Pour une année $m, m \geq 0$, on définit la suite de v.a. $\underline{Y}^{(m)} = \left\{Y_l^{(m)}, l \in \{1, \ldots, 91\}\right\}$, où $Y_l^{(m)}$ représente la quantité, en mm de pluie, tombée lors des 91 jours du printemps, c.-à-d. du 1$^{\text{er}}$ avril au 30 juin. Selon [Riboust and Brissette, 2016] et [Jalbert et al., 2019], cette période de temps est la plus critique lors des crues printanières.

Soit $\underline{\mathcal{C}}^{(m)} = \left\{\mathcal{C}_j^{(m)}, j \in \{1, \ldots, n_{\mathcal{C}}^{(m)}\}\right\}$, la suite des $n_{\mathcal{C}}^{(m)}$ clusters regroupant les jours de pluie continue pour l'année m. En concordance avec [Jalbert et al., 2019], on définit un cluster comme une séquence de jours de

pluie suivie par au moins une journée d'ensoleillement (0 mm de pluie). On définit la suite de v.a. $\underline{K}^{(m)} = \left\{K_j^{(m)}, j \in \{1, \ldots, n_{\mathcal{C}}^{(m)}\}\right\}$, où $K_j^{(m)}$ représente la quantité de pluie totale tombée lors de la période $\mathcal{C}_j^{(m)}$ telle que $K_j^{(m)} = \sum_{l \in \mathcal{C}_j^{(m)}} Y_l^{(m)},\ j = 1, \ldots, n_{\mathcal{C}}^{(m)},\ \forall m$.

Exemple 1. *Supposons que, pour une année m, on ait* $Y_1^{(m)} = 30,\ Y_2^{(m)} = 0,\ Y_3^{(m)} = 2,\ Y_4^{(m)} = 1,\ Y_5^{(m)} = 3,\ Y_6^{(m)} = 0,\ Y_7^{(m)} = 0,\ Y_8^{(m)} = 18,\ Y_9^{(m)} = 3,\ Y_{10}^{(m)} = 0$. *Alors* $\mathcal{C}_1 = \{1\},\ \mathcal{C}_2 = \{3, 4, 5\},\ \mathcal{C}_3 = \{8, 9\}$. *De plus,* $K_1^{(m)} = Y_1^{(m)} = 30,\ K_2^{(m)} = Y_3^{(m)} + Y_4^{(m)} + Y_5^{(m)} = 6,\ K_3^{(m)} = Y_8^{(m)} + Y_9^{(m)} = 21$.

Pour une année m et pour un seuil critique u, on sépare les événements en deux sous-ensembles $\mathcal{Z}^{(m)} = \{j : K_j^{(m)} < u\}$ et $\mathcal{X}^{(m)} = \{j : K_j^{(m)} \geq u\}$. On définira alors $\underline{X}^{(m)} = \left\{X_i^{(m)}, i \in \{1, \ldots, \text{card}(\mathcal{X}^{(m)})\}\right\} = \{K_j^{(m)} : j \in \mathcal{X}^{(m)}\}$ comme étant une suite de v.a. où $X_i^{(m)}$ représente la quantité de pluie tombée lors du i-ème cluster extrême de l'année m.

Exemple 1, suite. *On fixe un seuil $u = 18$. L'ensemble $\mathcal{X}^{(m)}$ correspond à $j \in \{1, 3\}$. On a donc $X_1^{(m)} = K_1^{(m)} = 30$, et $X_2^{(m)} = K_3^{(m)} = 21$. Puis, pour ce qui est des événements non-extrêmes, l'ensemble $\mathcal{Z}^{(m)}$ correspond au complément de l'ensemble $\mathcal{X}^{(m)}$. On a alors $\mathcal{Z}^{(m)} = \{2\}$.*

Soit $\underline{T}^{(m)} = \left\{T_i^{(m)}, i \in \{0, 1, \ldots, \text{card}(\mathcal{X}^{(m)})\}\right\}$ une suite de v.a. où $T_i^{(m)}$ représente la première journée du cluster $\{\mathcal{C}_j^{(m)}, j \in \mathcal{X}^{(m)}\}$ associé à $X_i^{(m)}$. On défini que $T_0^{(m)}$ correspond au 1er avril de l'année m. Afin de calculer $T_i^{(m)}\ \forall\ i, m$, on définit une date d'origine qui correspond au jour zéro (p.ex. $T_0 = 1^{\text{er}}$ avril 1900), puis, on converti cette date en format numérique selon le nombre de jours écoulé depuis cette date de référence. Cette opération se réalise automatiquement en enchaînant les fonctions `as.Date` avec `as.numeric` dans le langage de programmation `R`.

Soit $\underline{D}^{(m)} = \left\{D_i^{(m)}, i \in \{1, \ldots, \text{card}(\mathcal{X}^{(m)})\}\right\}$, une suite de v.a. où $D_i^{(m)}$ représente la durée, en nombre de jours, du i-ème événement extrême de l'année m. Concrètement, cela signifie que $\{D_i^{(m)}, i \in \{1, \ldots, \text{card}(\mathcal{X}^{(m)})\} = \{\text{card}(C_j^{(m)}) : j \in \mathcal{X}^{(m)}\}$.

Soit $\underline{W}^{(m)} = \left\{W_i^{(m)}, i \in \{1, \ldots, \text{card}(\mathcal{X}^{(m)})\}\right\}$, une suite de v.a. où $W_i^{(m)}$ représente le temps écoulé depuis la fin du dernier événement extrême telle que

$$W_i^{(m)} := T_i^{(m)} - T_{i-1}^{(m)} - D_{i-1}^{(m)} = T_i^{(m)} - T_{i-1}^{\star(m)},$$

où $T_i^{\star(m)} := T_0^{(m)} + \sum_{k=0}^{i} W_k^{(m)} + D_k^{(m)}$. Par convention, on a $W_0^{(m)} = D_0^{(m)} = 0,\ \forall m$.

Exemple 1, suite. *Si on attribut la valeur numérique 0 au 1er avril de l'année m, alors on a* $T_0^{(m)} = 0,\ T_1^{(m)} = T_0^{(m)} = 0,\ T_2^{(m)} = 7$. *On a également* $D_1^{(m)} = \text{card}(C_1^{(m)}) = 1,\ D_2^{(m)} = \text{card}(C_3^{(m)}) = 2$ *et* $W_1^{(m)} = T_1^{(0)} = 0, W_2^{(m)} = T_2^{(m)} - T_1^{(m)} - D_1^{(m)} = 7 - 0 - 1 = 6$.

Soit $\boldsymbol{N}^{(m)} = \{N_s^{(m)}(t),\ s, t \geq 0\} = \{N^{(m)}(s,\ s+t),\ s, t \geq 0\}$, un processus de renouvellement non-stationnaire alterné où un accroissement $N_s^{(m)}(t)$ permet de modéliser le nombre d'événements extrêmes ($\text{card}(\mathcal{X})$) survenus lors de l'intervalle de temps $[s, s+t]$. Dans le contexte des changement climatiques, aucune hypothèse n'est faite sur la

stationnarité du processus. Par définition, on a

$$N_{T_0}^{(m)}(t) := \sum_{i=1}^{\infty} \mathbb{1}\{T_i^{\star(m)} \leq T_0^{(m)} + t\} = \sum_{i=1}^{\infty} \mathbb{1}\{T_i^{(m)} + D_i^{(m)} \leq T_0^{(m)} + t\} = \inf\{i : T_i^{(m)} + D_i^{(m)} \leq T_0^{(m)} + t\}. \quad (1)$$

Soit $\boldsymbol{V}^{(m)} = \{V_s^{(m)}(t),\ s, t \geq 0\} = \{V^{(m)}(s,\ s+t),\ s, t \geq 0\}$, un processus de renouvellement non-stationnaire alterné avec récompenses où un accroissement $V_s^{(m)}(t)$ permet de modéliser la quantité totale d'eau accumulée lors des événements extrêmes sur un intervalle de temps $[s, s+t]$. Ainsi, on a

$$V_{T_0}^{(m)}(t) := \sum_{i=1}^{N_{T_0}^{(m)}(t)} X_i^{(m)} = \sum_{i=1}^{\infty} X_i^{(m)}\, \mathbb{1}\{T_i^{(m)} + D_i^{(m)} \leq T_0^{(m)} + t\}. \quad (2)$$

Les processus de renouvellement alternés de même que leur homologue avec récompenses sont utilisés dans la littérature en hydrologie, notamment dans [Salvadori and De Michele, 2006], [Salvadori and De Michele, 2007] et dans [Small and Morgan, 1986].

Soit $Z^{(m)}$, la v.a. de la quantité de pluie non-extrême totale tombée lors de la saison printanière de l'année m. On a $Z^{(m)} := \sum_{j \in \mathcal{Z}^{(m)}} K_j^{(m)}$. La suite de v.a. $\underline{Z} = \{Z^m, m \in \mathbb{N}\}$ n'est pas présumée stationnaire étant donné le contexte des changements climatiques. La quantité de pluie totale tombée au cours du printemps d'une année m est donc modélisée avec $S_{T_0}^{(m)}(91) = Z^{(m)} + V_{T_0}^{(m)}(91)$.

Exemple 1, suite. *Pour revenir à l'exemple 1, on a* $N_0^{(m)}(10) = \text{card}(\mathcal{X}) = 2$ *et* $V_0^{(m)}(10) = X_1^{(m)} + X_2^{(m)} = 30 + 21 = 51$. *Puis,* $Z^m = K_2^{(m)} = 6$. *Finalement* $S_0^{(m)}(10) = 51 + 6 = 57$.

Pour un exemple plus complet utilisant des données réelles provenant de la rivière Clearwater en Alberta et contenant plus d'une année, voir l'annexe A.

2.2 Hypothèses du modèle

En ce qui concerne l'hypothèse d'indépendance séquentielle, nous considérons qu'il est raisonnable de présumer que $Y_1^{(m)}$ est indépendante de $Y_{91}^{(m-1)}$. De ce fait, il faudrait réaliser le test d'indépendance proposé par [Genest and Rémillard, 2004] sur chacune des années de façon indépendante. Cependant, comme la méthode proposée réduit grandement le nombre d'observations disponibles pour chacune des années, on se retrouverait avec au plus une vingtaine d'observations par année pour les variables résultantes de l'agrégation. Conséquemment, les résultats du test risquent d'être instables et peu concluants. Pour cette raison et par souci de simplicité, nous supposons l'hypothèse d'indépendance séquentielle pour chacune des v.a. du modèle.

Pour ce qui est de l'hypothèse de stationnarité, considérant le contexte des changements climatiques, aucune présomption n'est faite à ce sujet. Un test de Mann-Kendall est effectué sur chacune des v.a. afin de vérifier l'hypothèse.

2.3 Distributions marginales

Avec la théorie des valeurs extrêmes, comme on cherche à modéliser l'ensemble des précipitations qui sortent des normales saisonnières, la méthode POT est tout indiquée pour modéliser les excédents de seuil (voir p.ex. [Hosking

and Wallis, 1987], [Klugman et al., 2013]). Alors on pose $(X_i - u | X_i \geq u) \sim \text{GPD}(\xi, \sigma)$ telle que la fonction de répartition s'exprime comme

$$F_u(x) := \begin{cases} 1 - (1 + \xi x/\sigma)^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp(-x/\sigma), & \xi = 0, \end{cases} \tag{3}$$

où $x \geq 0$, $\sigma > 0$, $\xi \in \mathbb{R}$ et $1 + \xi x/\sigma > 0$, pour un seuil $u > 0$. À noter que la paramétrisation de la loi est grandement influencée par la valeur attribuée au paramètre u, lequel peut être obtenu par optimisation tel que présenté dans [Bader et al., 2016] et [Bader et al., 2018] ou encore avec [Northrop et al., 2015]. Pour faire suite à [Jalbert et al., 2019] et pour des fins de simplicité considérant le temps imparti pour ce stage, le seuil utilisé est fixe. Cependant, il est d'intérêt de mentionner que plusieurs auteurs tels que [Kyselỳ et al., 2010], [Beguería et al., 2011] et [Cheng et al., 2014] préconisent pour un seuil non-stationnaire. Cette approche sera considérée dans une prochaine version de ce travail.

En ce qui concerne à la modélisation de la durée des périodes de pluie et les temps inter-occurrences, considérant que l'échelle de temps utilisée est en jours, les lois envisagées sont discrètes. De plus, considérant que le domaine des v.a. en jeux n'est pas fermé, la loi binomiale est écartée. Conséquemment, les lois retenues sont la loi binomiale négative, son homologue à un paramètre, la loi géométrique et la loi de Poisson. Également, étant donnée sa grande flexibilité, la loi Weibull discrétisée est également retenue.

Afin de paramétrer la distribution de W, on pose la fonction de vraisemblance (4).

$$L(\boldsymbol{\theta} | \boldsymbol{w}, \boldsymbol{t}^\star) = \prod_k f_W(w_k | \boldsymbol{\theta}, t_k^\star), \tag{4}$$

où $\boldsymbol{\theta}$ correspond au vecteur des paramètres de la loi de W, $\boldsymbol{t}^\star = \{t_k^\star, k \in \{1, \ldots, \text{card}(\boldsymbol{t})\}\}$ correspond à un vecteur d'observations où $t_k^\star$ est la k-ème observation de $T^\star$ dans les données disponibles. Puisque l'accroissement du processus de renouvellement est définit sur un intervalle de temps fini, on ajoute un terme pour tenir compte du temps qu'il reste à la fin du processus, pour chacune des années. Soit $w_{\text{res}}^{(m)} = 91 - t_N^{\star(m)}$ où $t_N^{\star(m)}$ correspond au temps de fin du dernier événement de l'année m et $w_{\text{res}}^{(m)}$ est le temps résiduel du processus. Alors (4) devient

$$L'(\boldsymbol{\theta} | \boldsymbol{w}, \boldsymbol{t}^\star) = L(\boldsymbol{\theta} | \boldsymbol{w}, \boldsymbol{t}^\star) \times \prod_m (1 - F_W(w_{\text{res}}^{(m)} | \boldsymbol{\theta}, t_N^{\star(m)})). \tag{5}$$

Du point de vue des précipitations non-extrêmes, [Jalbert et al., 2019] recommande l'utilisation la la loi normale. Cependant, comme cette dernière admet des valeurs négatives, on considérera aussi la loi gamma comme alternative.

2.4 Modélisation de la dépendance

La dépendance unissant les v.a. en jeu est modélisée avec la théorie des copules décrite dans [Joe, 1997] et [Nelsen, 2006]. Dans le contexte particulier des précipitations, on s'intéressera particulièrement à [Zhang and Singh, 2019] qui utilise les copules archimédiennes et elliptiques pour modéliser la dépendance entre la durée et l'intensité des précipitations. On s'intéresse également à [Salvadori and De Michele, 2006] qui défini un processus de renouvellement alterné avec structure de dépendance unissant les variables de durée des épisodes secs, de durée des périodes humides et d'une variable d'intensité pour modéliser les orages. Cette dépendance est modélisée à l'aide d'une copule en vigne tri-variée.

Dans le contexte des valeurs extrêmes, on s'intéresse à [Gudendorf and Segers, 2010] qui recommande, entre-autre, les copules de Gumbel, de Tawn et de Galambos lorsque les marginales impliquées sont de lois extrêmes.

Les copules elliptiques sont appréciées dans la littérature en hydrologie étant donnée leur flexibilité et la facilité avec laquelle on peut les paramétrer. Cependant, dans le contexte de la modélisation de valeurs extrêmes, nous les écartons, conformément aux conclusions de [Renard and Lang, 2007], puisque cette famille de copule tend à sous-estimer la dépendance lorsque les marginales sont constituées à la fois de v.a. extrêmes et de v.a. non-extrêmes.

3 Études de cas

Afin d'appliquer le modèle proposé dans la section 2, deux bases de données sont étudiées. La première fait suite à l'étude réalisée par [Jalbert et al., 2019] et concerne la ville de Burlington dans le Vermont. La seconde touche la rivière Clearwater en Alberta. Dans les deux cas, l'hypothèse de non-stationnarité est vérifiée pour chacune des v.a. en jeux. Différentes distributions sont testées afin de les modéliser et le critère de l'AIC est utilisé pour sélectionner les loi marginales offrant la meilleure adéquation. Cette approche s'appuie sur [Chebana and Ouarda, 2021] et sur [Khaliq et al., 2006]. Par la suite, les copules en vignes sont utilisées afin de modéliser la structure de dépendance unissant les différentes composantes du modèle proposé. La sélection de la copule la plus adéquate s'appuie sur [Dissmann et al., 2013] et les fonction `R` de la librairie `VineCopula`.

Pour les prochaines sections, afin de simplifier la notation, on pose simplement

$$K = \{K_j^{(m)},\ j \in \{1, \ldots, n_{\mathcal{C}}^{(m)}\}, \forall m\}, \quad X = \{X_i^{(m)},\ i \in \{1, \ldots, \text{card}(\mathcal{X}^{(m)})\}, \forall m\},$$

$$W = \{W_i^{(m)},\ i \in \{1, \ldots, \text{card}(\mathcal{X}^{(m)})\}, \forall m\}, \quad D = \{D_i^{(m)},\ i \in \{1, \ldots, \text{card}(\mathcal{X}^{(m)})\}, \forall m\}$$

$$\text{et } Z = \{Z^{(m)},\ \forall m\}.$$

3.1 Lac Champlain

La première étude de cas fait suite à celle réalisée par [Jalbert et al., 2019] et concerne la ville de Burlington dans l'état du Vermont aux États-Unis. Les données utilisées proviennent du site de la National Oceanic and Atmospheric Administration (NOAA) [1]. La station `USC00431072` a débuté ses opérations en 1884. Cependant, elle comporte quelques données manquantes. Afin d'interpoler sur celles-ci, l'approche recommandée par [Shaw et al., 2010] est de moyenner les observations obtenues sur les stations avoisinantes. De plus, comme cette station a fermé le 3 juin 1943, nous avons compensé, comme indiqué par [Jalbert et al., 2019], avec la station de l'aéroport de Burlington, soit la station `USW00014742`, laquelle a commencé ses opérations le 1er décembre 1940. Les données obtenues comporte de l'information jusqu'en 2020, soit 137 ans (12 467 observations printanières).

Dans un premier temps, l'étude des données débute avec une analyse de stationnarité réalisée avec le test de Mann-Kendall sur chacune des v.a. en jeu. Les résultats du test sont résumés dans le tableau 1. On y voit que, avec un seuil u qui est fixe, les v.a.de l'excédent de seuil, de la durée des événements extrêmes et des précipitations saisonnières non-extrêmes ne démontrent pas de signe de non-stationnarité, alors que les suites de v.a. W et K présentent de fortes évidences. Possiblement que le fait d'avoir utilisé un seuil fixe a fait en sorte de briser l'effet de non-stationnarité dans les distributions de $X - u|X \geq u$ et de Z. Cette hypothèse sera testée pour la prochaine version de ce travail.

1. https://www.ncdc.noaa.gov/cdo-web/search?datasetid=GHCND

Variable	Statistique	*p-value*
K	0.0401	0.0028
$X - u \vert X \geq u$	0.0228	0.5329
Z	0.0379	0.5124
W	-0.1010	0.0060
D	0.0461	0.2366

TABLEAU 1 – Résultats du test de Mann-Kendall calculé sur les différentes v.a. du modèle pour le jeu de données du lac Champlain.

Pour ce qui est de la détermination du seuil u servant à séparer les événements extrêmes des saisonniers, on applique l'approche proposée par [Bader et al., 2018] pour trouver un seuil optimal de 26.67mm de pluie. Cette quantité correspond au 86.39^{e} percentile des observations empiriques de K et permet de conserver 340 observations en excédent de seuil.

Suite au calcul des réalisations de $X - u|X \geq u$, les tests d'Anderson-Darling et de Cramer-von Mises sont effectués selon la méthode suggérée par [Choulakian and Stephens, 2001] pour vérifier l'adéquation des excès de seuil avec la famille de loi GP. Ceux-ci offrent des *p-value* de 0.132 et 0.130. Conséquemment, on ne peut rejeter l'hypothèse nulle que les excès de seuil sont issues de la loi de Pareto généralisée au seuil de 5%.

En ce qui a trait à la paramétrisation de la loi des excédents de seuil, considérant que le paramètre de forme de la loi GP obtenue de façon préliminaire avec la sélection de seuil automatique est de 0.06788321. La méthode utilisée pour trouver les estimateurs de la loi est la méthode des moments, conformément aux recommandations de [Hosking and Wallis, 1987]. Les estimateurs finaux ainsi obtenus sont donc $\hat{\xi} = 0.08056533$ et $\hat{\sigma} = 15.26847280$ selon la paramétrisation définie en (3). L'illustration 10a atteste de l'adéquation de la loi ainsi paramétrée.

Du côté de la modélisation de Z, comme il a été soulevé avec le tableau 1, les observations relatives aux précipitations non-extrêmes ne présentent pas d'évidence contre l'hypothèse de stationnarité lorsque le seuil u utilisé est fixe. En conformité avec l'observation de [Jalbert et al., 2019], une loi normale stationnaire est testée pour modéliser les précipitations non-extrêmes. Cependant, comme le test de Shapiro-Wilk offre une *p-value* de 0.03, la loi normale est écartée en faveur de la loi Gamma. L'illustration 10b atteste de l'adéquation de la loi avec les données pour les estimateurs des paramètres $\hat{\alpha}^{(Z)} = 14.47781$ et $\hat{\beta}^{(Z)} = 132.8195$, selon la paramétrisation que $\mathbb{E}[Z] = \alpha^{(Z)}/\beta^{(Z)}$. Les tests de Kolmogorov-Smirnov et d'Anderson-Darling donnent des *p-values* de 0.708 et de 0.817. Conséquemment, on peut déduire que la loi gamma est adéquate dans ce contexte.

Du point de vue des temps inter-occurrences et de la durée des événements extrêmes, L'illustration 7 présente la fonction de masses de probabilités observée pour chacune de ces v.a. On voit alors que possiblement que les loi géométrique et Weibull discrétisée aurait une belle adéquation avec la v.a. W, tandis que les lois de Poisson, binomiale négative et Weibull discrétisée seraient intéressantes à tester pour la v.a. D.

Afin de modéliser la tendance présentée dans l'illustration 9, la variance des séries temporelles semble *a priori* être stable dans les deux cas. Seule la moyenne semble avoir une tendance. Pour cette raison, on considère que l'espérance de la v.a. peut être modélisée à l'aide d'une fonction linéaire, soit $\mathbb{E}[W_i^{(m)}|t] = \mu(t) = a + bt,\ t \in \mathbb{R}$. Puis, le paramètre de forme de chacune des lois testées devient une fonction de cette moyenne.

Exemple 2. *Soit la v.a. W telle que $W \sim \text{Geo}(p)$, $p \in [0, 1]$, avec espérance $\mathbb{E}[W] < \infty$ définie comme $\mathbb{E}[W] = 1/p$.*

Pour paramétrer la loi de W en tenant compte de la tendance, on pose $p(t) = 1/\mu(t),\ t \in \mathbb{R}$.

Exemple 3. *Soit la v.a. W telle que $W \sim$ Weibull$(\alpha, \beta),\ \alpha, \beta > 0$, avec espérance $\mathbb{E}[W] < \infty$ définie comme $\mathbb{E}[W] = \frac{1}{\beta}\Gamma(1 + 1/\alpha)$. Pour paramétrer la loi de W en tenant compte de la tendance, on pose $\beta(t) = \mu(t)/\Gamma(1 + 1/\alpha),\ \alpha > 0,\ t \in \mathbb{R}$.*

Comme l'atteste le tableau 2, on trouve que la loi qui minimise l'AIC pour modéliser la v.a. W est la loi géométrique non-stationnaire, tandis que, du côté de la v.a. D, c'est la loi Weibull discrétisée non-stationnaire qui performe le mieux. En ce qui a trait aux paramètres des lois, on a $\hat{a}^{(W)} = 32.48488$, $\hat{b}^{(W)} = -0.000410909$ et $\hat{\alpha}^{(D)} = 2.104619$, $\hat{a}^{(D)} = 3.94734$, $\hat{b}^{(D)} = 1.746127e-05$.

Un fait intéressant à observer est que, malgré que le test de Mann-Kendall n'offre aucune évidence contre la non-stationnarité de la v.a. D, l'AIC suggère de tenir compte d'une tendance dans la distribution marginale de cette v.a. tout de même. Afin de démystifier cela, l'illustration 9b permet de constater qu'il semble y avoir une légère tendance ascendante.

Les illustrations 10c et 10d présentent l'adéquation graphique des lois paramétrées pour les v.a. W et D. Il serait pertinent de faire un test du chi-2.

V.a.	Tendance	AIC des lois testées		
		Géometrique	Binomiale nég.	Weibull disc.
W	**Avec**	**3066.589**	3068.48	3093.457
	Sans	3152.568	3076.167	3091.457
		Binomiale nég.	Poisson	**Weibull disc.**
D	**Avec**	1391.907	1389.905	**1378.473**
	Sans	1393.333	1391.352	1383.088

TABLEAU 2 – Comparaison de l'AIC des lois de probabilités testées pour modéliser les v.a. W et D selon qu'il y ait ou non une tendance qui soit tenue en compte.

Soit $\boldsymbol{x} = \{(X_i^{(m)} - u | X_i^{(m)} \geq u), \forall i, m\}$, la suite des observations où $(X_i^{(m)} - u | X_i^{(m)} \geq u)$ représente l'excédent de seuil observé pour le i-ème événement de l'année m. Soit $\boldsymbol{w} = \{W_i^{(m)}, \forall i, m\}$, la suite des observations où $W_i^{(m)}$ représente le temps écoulé entre les i-ème et $(i-1)$-ème événements de l'année m. Soit $\boldsymbol{d} = \{D_i^{(m)}, \forall i, m\}$, la suite des observations où $D_i^{(m)}$ représente la durée du i-ème événement extrême lors de l'année m. Soit $\boldsymbol{t} = \{T_i^{(m)}, \forall i, m\}$, la suite des observations où $T_i^{(m)}$ représente le temps de survenance du i-ème événement extrême lors de l'année m. Soit $u^{(X)} = \mathbb{P}(X - u \leq \boldsymbol{x} | X > u)$, le vecteur des uniformes générés en évaluant la fonction de répartition marginale estimée pour les excédents de seuil. Soit $u^{(W)} = \mathbb{P}(W \leq \boldsymbol{w}|\boldsymbol{t})$, le vecteur des uniformes générés en évaluant la fonction de répartition marginale estimée pour les temps inter-occurrences. Soit $u^{(D)} = \mathbb{P}(D \leq \boldsymbol{d}|\boldsymbol{t})$, le vecteur des uniformes générés en évaluant la fonction de répartition marginale estimée pour la durée de chacun des événements observés.

Remarque 1. *L'utilisation des fonctions de répartitions empiriques $F_n(x) = \text{rank}(x)/(n+1)$ sous-tend l'hypothèse que les observations disponibles sont représentatives des minimums et maximums des distributions réelles des données. Hors, comme la v.a. des excédents de seuil fait partie de la famille des lois à valeurs extrêmes, cette hypothèse doit être rejetée. C'est pourquoi, il est mieux d'utiliser les fonctions de répartitions estimées pour produire les pseudo-observations plutôt que d'utiliser une méthode basée sur les rangs comme le suggère généralement la littérature sur la théorie des copules (voir p.ex. [Genest and Favre, 2007]).*

On peut étudier la dépendance entre les différentes v.a. du modèle en débutant par le calcul du coefficient de corrélation de Pearson sur les vecteurs d'uniformes observées pour chacune des v.a. De cette façon, on trouve la matrice des ρ de Pearson (6).

$$\boldsymbol{\rho}_P(u^{(X)}, u^{(D)}, u^{(W)}) = \begin{pmatrix} 1.0000 & 0.1358 & 0.3689 \\ 0.1358 & 1.0000 & 0.0082 \\ 0.3689 & 0.0082 & 1.0000 \end{pmatrix}. \tag{6}$$

Un test de Mantel-Haensel (voir [Mantel, 1963]) ne permet pas de rejeter l'hypothèse nulle d'indépendance entre les variables D et W. Cependant, considérant que la dépendance est significative pour les autres paires de v.a., alors on considère tout de même une copule tri-variée pour modéliser la dépendance entre ces v.a. L'illustration 1 présente les nuages de points ainsi qu'une estimation des copules représentant la dépendance entre les paires de v.a.

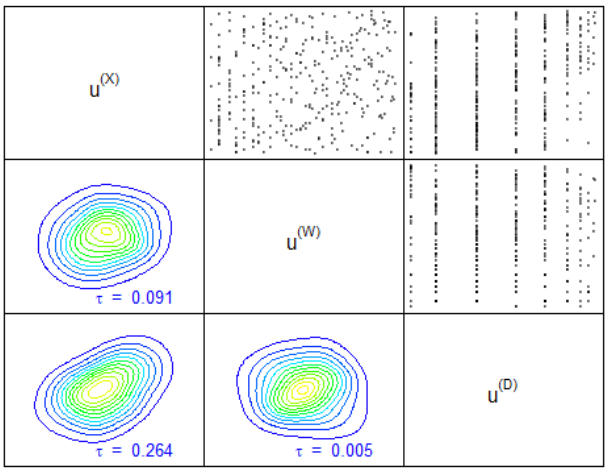


Illustration 1 – Triangle supérieur de la matrice : Nuage de points des paires de v.a. Triangle inférieur de la matrice : graphiques de contours pour les estimations des copules empiriques selon une approche par noyau gaussien.

La structure de copule en vigne utilisée est représentée par la matrice C-vine (7)(voir [Joe et al., 2010], [Dissmann et al., 2013] ou la documentation de la librairie `VineCopula` [2] sur `CRAN`)

$$M = \begin{pmatrix} 1 & & \\ 2 & 2 & \\ 3 & 3 & 3 \end{pmatrix}, \tag{7}$$

où les indices 1,2 et 3 réfèrent respectivement à $u^{(X)}, u^{(W)}, u^{(D)}$. La matrice 7 peut être exprimée sous la forme de la structure C-Vine présentée dans l'illustration 2.

2. https://cran.r-project.org/web/packages/VineCopula/VineCopula.pdf

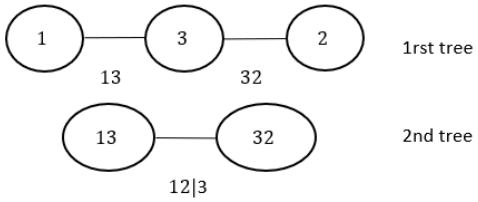


ILLUSTRATION 2 – Structure C-Vine de la copule tri-variée modélisant la dépendance entre (1) $u^{(X)}$, (2) $u^{(W)}$ et (3) $u^{(D)}$.

Pour ce qui est du choix des copules bi-variées composant la copule en vigne, les copules à valeur extrêmes de Tawn de type 1 ainsi que celle de Gumbel sont prises en compte. De plus, on considère la famille des copules archimédiennes telles que celle de Frank, de Clayton ainsi que les copules BB1-BB3 et BB6-BB7 décrites dans [Joe, 1997]. Finalement, on considère également des rotations à 180° de ces copules ; dans ce cas, on parle alors de copules de survie. Le tableau 3 présente les résultats obtenus pour les 5 copules offrant la meilleure adéquation avec les couples d'observations selon les critères de l'AIC et du BIC. En ce qui a trait à la dépendance unissant le couples

Variables	Copules	$\boldsymbol{\theta}$	log-vrais.	AIC	BIC
$(u^{(X)}, u^{(D)})$	Gumbel	1.29	24.77	-47.55	-43.72
	Frank	2.33	24.27	-46.55	-42.72
	BB8	[2.45, 0.75]	**25.83**	**-47.65**	-40.00
	Clayton (180°)	**0.49**	24.81	-47.61	**-43.78**
	BB1 (180°)	[0.41,1.05]	25.29	-46.58	-38.92
$(u^{(X)}, u^{(W)}\|u^{(D)})$	**Indépendance**	$\emptyset$	0.00	0.00	**0.00**
	Gumbel	1.1	2.12	-2.25	1.58
	Frank	0.81	2.80	**-3.60**	0.23
	Clayton (180°)	0.18	2.26	-2.52	1.31
	Tawn type 1 (180°)	[1.34, 0.20]	**3.57**	-3.13	4.52

TABLEAU 3 – Résultats de l'estimation des copules candidates.

$(u^{(W)}, u^{(D)})$, comme il a été sus-mentionné que la dépendance n'était pas significative, la copule d'indépendance est utilisée. Selon le critère du BIC, la copule retenue pour modéliser la dépendance entre les excédents de seuil et les temps inter-occurrences est la copule de survie de Clayton. Pour ce qui est du deuxième arbre de la copule en C-vine, on observe qu'en conditionnant sur les pseudo-observations de D, la force de dépendance entre les deux autres v.a. a significativement diminué, laissant percevoir que la dépendance *a priori* du couple de v.a. $(u^{(x)}, u^{(W)})$ est fortement influencée par celle observée avec $u^{(D)}$. L'hypothèse d'indépendance conditionnelle devient alors une hypothèse plausible. Conséquemment, on peut résumer le modèle de dépendance à une copule bi-variée unissant les v.a. des excédents de seuil et de la durée des événements extrêmes.

Remarque 2. *Afin de tenir compte du fait que deux des trois v.a. impliquées ont un support discret, [Genest and Nešlehová, 2007] suggère de ne pas utiliser la méthode des moments pour paramétrer les copules puisque cela*

insérerait un biais. La méthode de paramétrisation s'appuie donc sur la méthode de la pseudo-vraisemblance (voir [Kim et al., 2007]).

Pour ce qui est de la dépendance unissant les réalisations de Z et de V, on trouve que le rho de Spearman empirique est de -0.244. Néanmoins, comme l'ajout d'une copule pour modéliser cette dépendance dégrade les résultats de prédiction dû à la complexité du modèle, aucune copule n'est utilisée dans ce cas. Cette décision est cohérente avec l'approche adoptée par [Jalbert et al., 2019].

Finalement, afin de pouvoir étudier le comportement de $S_{T_0}^{(m)}(t)$, on procède par simulation. À cet effet, l'algorithme 1 propose une procédure pour obtenir une matrice de valeurs simulées sur un horizon de temps variable. L'illustration 4 compare les quantiles observés aux quantiles simulés avec l'algorithme 1. Une analyse plus approfondie des résultats obtenus est réalisée dans la section 3.3.

3.2 Rivière Clearwater

La rivière Clearwater se trouve en Alberta, au Canada. Les données utilisées proviennent de la station `07CD001` qui est située près de Fort McMurray, aux coordonnées GPS suivantes : (56°41'06" N, 111°15'18" W). L'illustration 6 présente l'emplacement de la station. Les données utilisées couvrent la période de 1960 à 2013 (53 ans ; 4823 observations printanières). Celles-ci sont disponibles sur le site des Relevés hydrologiques du Canada[3]. Aucune donnée manquante n'est recensée dans cette base de données.

Pour cette deuxième étude de cas, on reprend la méthodologie décrite dans la section 3.1. Du point de vue de la stationnarité, le test de Mann-Kendall appliquée aux différentes v.a. observées dans le jeu de données de la Rivière Clearwater offre les résultats du tableau 4. Dans celui-ci, on a le même constat quant aux v.a. des

Variable	Statistique	*p-value*
K	-0.0621	0.0056
$X-u\|X\geq u$	0.1010	0.1388
Z	-0.0587	0.5357
W	0.1710	0.0123
D	0.2690	0.0002

TABLEAU 4 – Résultats du test de Mann-Kendall calculé sur les différentes v.a. du modèle pour le jeu de données de la rivière ClearWater

excès de seuil et des précipitations non-extrêmes que dans la section 3.1. C.-à-d. que l'utilisation d'un seuil fixe semble briser la non-stationnarité qui existe pour la v.a. K, mais qui n'est plus présente pour $X-u|X>u$ et Z. Par la suite, il est plus clair dans ce cas-ci que la v.a. D est non-stationnaire par rapport à l'étude de cas précédente.

En ce qui a trait au seuil de valeurs extrêmes calculé avec la méthode de [Bader et al., 2018], on trouve $u = 18.29828$. Celui-ci correspond au 88.74^{e} percentile empirique et permet de conserver 100 observations pour l'étude des événements extrêmes. Le test d'Anderson-Darling et de Cramer-von Mises d'adéquation à la famille GP offrent des seuils observés de 0.452 et de 0.725. L'adéquation est donc excellente comme en atteste l'illustration 11a.

3. https://eau.ec.gc.ca/search/historical_f.html

Les estimateurs des paramètres de la loi selon la méthode des moment pondérés (voir [Hosking and Wallis, 1987]) sont $\hat{\xi} = 0.1317715$ et $\hat{\sigma} = 14.2908019$.

Du point de vue des précipitations non extrêmes, un test de Shapiro-Wilk offre un seuil observé de 0.53576. On ne peut donc pas rejeter l'hypothèse nulle que la distribution empirique provient de la loi normale. Une comparaison de l'AIC d'une loi normale avec la loi gamma vient confirmer que la première est plus vraisemblable pour modéliser les données. Les estimateurs des paramètres trouvés sont alors $\hat{\mu}^{(Z)} = 63.54724$, $\hat{\sigma}^{(Z)} = 19.46997$. L'illustration 11b compare les quantiles observés à ceux estimés.

Jusqu'à maintenant, l'adéquation des v.a. étudiées pour ce jeu de données était quasi-parfaite. Cependant pour ce qui est de la distribution de la v.a. W, les choses se corsent. En effet, l'illustration 8b présente l'histogramme de la fonction de masses de probabilités empiriques pour cette v.a. On y voit que la loi sous-jacente à la v.a. W possède une queue de distribution très lourde. Par ailleurs, considérant que les données sont censurées à 91, et que la distribution est non stationnaire, la modélisation de cette v.a. représente un défi. Pour y arriver, on considère deux familles de loi avec une queue de distribution très lourde : la loi Weibull et la famille GPD. Dans le cas le la loi Weibull, la méthode du maximum de vraisemblance standard fonctionne bien. En revanche, pour la loi GP, celle-ci ne peut être paramétrée avec cette méthode. C'est pourquoi la méthode du maximum de vraisemblance généralisée proposée par [El Adlouni et al., 2007] est utilisée comme alternative. L'algorithme utilisé est celui de Metropolis-Hastings qui est implémentée dans la fonction `Metro_Hastings`, disponible dans la librairie `MHadaptive` dans le langage de programmation `R`. Soit $\xi^{(W)}$ et $\sigma^{(W)}$, les paramètres de forme et d'échelle de la loi GP tels qu'utilisés dans (3). Comme la méthode du maximum de vraisemblance généralisée est une approche bayésienne, il faut définir des distributions *a priori* pour les paramètres du modèle. Dans le cas stationnaire, on a

$$\xi^{(W)} \sim \mathcal{N}(0, 0.25), \quad \sigma^{(W)} \sim \Gamma(5, 1). \tag{8}$$

Pour ce qui est du cas non-stationnaire, on pose $\xi_t^{(W)} = \xi^{(W)}$ ainsi que $\log(\sigma_t^{(W)}) = a^{(W)} + b^{(W)} \times t,\ t \geq 0$. Puis, on pose

$$\xi^{(W)} \sim \mathcal{N}(0, 0.25), \quad a^{(W)} \sim \mathcal{N}(\tilde{a}, 3), \quad b^{(W)} \sim \mathcal{N}(0, 1), \tag{9}$$

où $\tilde{a}$ correspond au paramètre initial de l'optimisation. Celui-ci peut être grossièrement estimé en calculant le logarithme naturel de la moyenne empirique des observations de W. En utilisant 15 000 itérations dont 8 000 sont *brûlées* afin de faire converger la chaîne de Markov, conformément à la méthodologie proposée par [El Adlouni et al., 2007]. Paramètres de la loi de W : -0.5056498 58.5460087
D : 2.2442893 1.0591309 0.7522425

En ce qui a trait à la loi de D, la méthodologie utilisée est la même que pour l'étude de cas précédente. Les AIC ainsi calculés, autant pour W que pour D sont présentés dans le tableau 5.

V.a.	Tendance	AIC des lois testées		
		Weibull disc.	**GPD disc.**	
W	**Avec**	960.9686	**952.082**	
	Sans	960.6018	952.8248	
		Binomiale nég.	Poisson	**Weibull disc.**
D	**Avec**	413.2777	411.2829	**407.7906**
	Sans	430.4946	430.4746	429.1043

TABLEAU 5 – Comparaison de l'AIC des lois de probabilités testées pour modéliser les v.a. W et D selon qu'il y ait ou non une tendance qui soit tenue en compte.

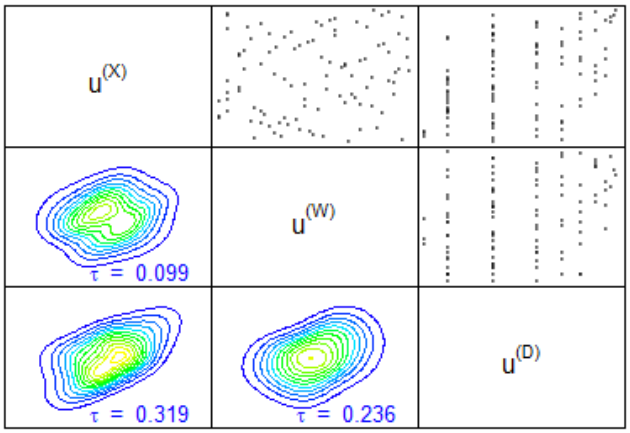
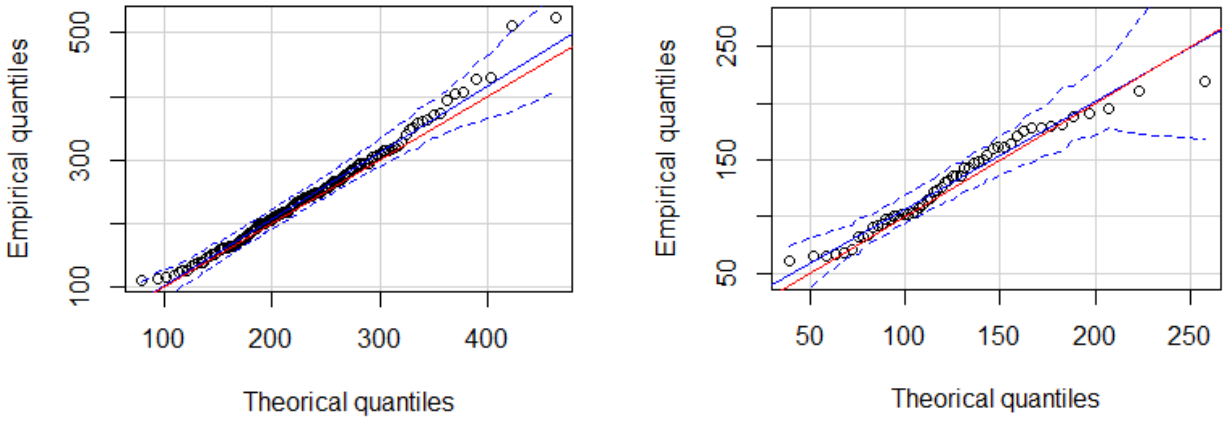


ILLUSTRATION 3 – Triangle supérieur de la matrice : Nuage de points des paires de v.a. Triangle inférieur de la matrice : graphiques de contours pour les estimations des copules empiriques selon une approche par noyau gaussien.

3.3 Résultats



(a) Lac Champlain

(b) Rivière Clearwater

ILLUSTRATION 4 – Diagrammes quantile-quantile de $S_{T_0}^{(m)}(91)$: La ligne bleue présente la droite de tendance des quantiles. Les pointillés présentent l'intervalle de confiance au seuil de 5% pour cette droite et la ligne rouge est la diagonale d'adéquation parfaite. L'objectif est que la diagonale rouge se situe dans l'intervalle de confiance.

4 Discussion

Limites : Dans la méthodologie utilisée, mentionnons que le test statistique pour valider l'indépendance séquentielle des données utilisée est expérimentale. Également, la quantité de données ne permet pas d'avoir une appréciation fiable de la stationnarité de la dépendance entre les variables conformément à [Chebana and Ouarda, 2021]. Ceci étant dit, les hypothèses de stationnarité de la dépendance et de l'indépendance séquentielle des variables aléatoires réduit considérablement la complexité du modèle. En ce qui a trait à la principale faiblesses du modèle par rapport à ceux n'agrégeant pas les jours de pluie est que cette agrégation réduit considérablement le nombre d'observations disponibles pour entraîner le modèle. Malgré tout, les résultats obtenus sont excellents et le fait de considérer des périodes de plusieurs jours fait en sorte que la loi GP ainsi entraînée obtient un estimateur de forme lui permettant d'avoir un support non fini.

4.1 Pour aller plus loin

Considérer [Cheng et al., 2014] pour le calcul de la période de retour dans le cas non-stationnaire.

La prochaine étape du projet serait de considérer un seuil qui sont non-stationnaire conformément à [Kyselỳ et al., 2010] et [Beguería et al., 2011]. De plus, il restera à comparer les résultats du modèle proposé avec ceux obtenus avec le modèle de [Jalbert et al., 2019] et possiblement d'autres articles concurrents qui n'utilisent pas de regroupement des observations.

Pour aller plus loin dans la construction du modèle, on pourrait considérer, non seulement le temps, mais aussi l'espace dans le modèle. En effet, [Baigorria et al., 2007] démontre qu'il existe une forte corrélation entre les stations avoisinantes. À cette fin, certains auteurs utilisent des méthodes de régression telles que [Hession and Moore, 2011] et [Beguería and Vicente-Serrano, 2006] afin de réaliser un lissage spatial des paramètres de leur modèle de précipitations. Cependant, une telle approche sous-tend l'hypothèse que la dépendance unissant les différentes stations est linéaire. Or, comme l'a démontré [Zhang and Singh, 2019] cette relation n'est pas linéaire puisque la dépendance tend à être plus forte dans les ailes de distributions conjointes. C'est pourquoi ce dernier propose l'utilisation de copules en vignes. Cependant, cette dernière approche ajoute une grande complexité au modèle et ne permet pas de faire un lissage à grande échelle. Ainsi, pour aller plus loin, on pourrait tester une approche utilisant les réseaux de neurones convolutifs récurrents telle que présentée dans [Liu et al., 2016].

Finalement, pour revenir à la motivation initiale de [Jalbert et al., 2019] et de notre modèle, l'objectif ultime serait de prédire les inondations de manière exhaustive. Pour se faire, en s'inspirant de [Riboust and Brissette, 2016], quelqu'un pourrait proposer un modèle tri-varié où les variables en jeu seraient les précipitations hivernales ainsi que la température et les précipitations printanières. Ainsi, on aurait un modèle complet pour prédire la quantité totale d'eau accumulée lors du printemps.

Remerciements

L'auteur aimerait remercier l'INRS ainsi que la Chaire d'actuariat de l'Université Laval pour avoir financé ce projet. Il aimerait également remercier les professeurs Étienne Marceau et Hélène Cossette pour leurs conseils et commentaires, de même que Fateh Chebana pour avoir supervisé le projet.

Références

[Bader et al., 2016] Bader, B., Yan, J., and Zhang, X. (2016). Automated threshold selection for extreme value analysis via goodness-of-fit tests with application to batched return level mapping. *arXiv preprint arXiv :1604.02024*.

[Bader et al., 2018] Bader, B., Yan, J., Zhang, X., et al. (2018). Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate. *The Annals of Applied Statistics*, 12(1) :310–329.

[Baigorria et al., 2007] Baigorria, G. A., Jones, J. W., and O'Brien, J. J. (2007). Understanding rainfall spatial variability in southeast usa at different timescales. *International Journal of Climatology : A Journal of the Royal Meteorological Society*, 27(6) :749–760.

[Beguería et al., 2011] Beguería, S., Angulo-Martínez, M., Vicente-Serrano, S. M., López-Moreno, J. I., and El-Kenawy, A. (2011). Assessing trends in extreme precipitation events intensity and magnitude using non-stationary peaks-over-threshold analysis : a case study in northeast spain from 1930 to 2006. *International Journal of Climatology*, 31(14) :2102–2114.

[Beguería and Vicente-Serrano, 2006] Beguería, S. and Vicente-Serrano, S. M. (2006). Mapping the hazard of extreme rainfall by peaks over threshold extreme value analysis and spatial regression techniques. *Journal of applied meteorology and climatology*, 45(1) :108–124.

[Chebana and Ouarda, 2021] Chebana, F. and Ouarda, T. B. (2021). Multivariate non-stationary hydrological frequency analysis. *Journal of Hydrology*, 593 :125907.

[Cheng et al., 2014] Cheng, L., AghaKouchak, A., Gilleland, E., and Katz, R. W. (2014). Non-stationary extreme value analysis in a changing climate. *Climatic change*, 127(2) :353–369.

[Choulakian and Stephens, 2001] Choulakian, V. and Stephens, M. A. (2001). Goodness-of-fit tests for the generalized pareto distribution. *Technometrics*, 43(4) :478–484.

[Dissmann et al., 2013] Dissmann, J., Brechmann, E. C., Czado, C., and Kurowicka, D. (2013). Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis*, 59 :52–69.

[El Adlouni et al., 2007] El Adlouni, S., Ouarda, T. B., Zhang, X., Roy, R., and Bobée, B. (2007). Generalized maximum likelihood estimators for the nonstationary generalized extreme value model. *Water Resources Research*, 43(3).

[Genest and Favre, 2007] Genest, C. and Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of hydrologic engineering*, 12(4) :347–368.

[Genest and Nešlehová, 2007] Genest, C. and Nešlehová, J. (2007). A primer on copulas for count data. *ASTIN Bulletin : The Journal of the IAA*, 37(2) :475–515.

[Genest and Rémillard, 2004] Genest, C. and Rémillard, B. (2004). Test of independence and randomness based on the empirical copula process. *Test*, 13(2) :335–369.

[Gudendorf and Segers, 2010] Gudendorf, G. and Segers, J. (2010). Extreme-value copulas. In *Copula theory and its applications*, pages 127–145. Springer.

[Hession and Moore, 2011] Hession, S. L. and Moore, N. (2011). A spatial regression analysis of the influence of topography on monthly rainfall in east africa. *International journal of climatology*, 31(10) :1440–1456.

[Hosking and Wallis, 1987] Hosking, J. R. and Wallis, J. R. (1987). Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, 29(3) :339–349.

[Jalbert et al., 2019] Jalbert, J., Murphy, O. A., Genest, C., and Nešlehová, J. G. (2019). Modelling extreme rain accumulation with an application to the 2011 lake champlain flood. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 68(4) :831–858.

[Joe, 1997] Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press.

[Joe et al., 2010] Joe, H., Cooke, R. M., and Kurowicka, D. (2010). Regular vines : generation algorithm and number of equivalence classes. In *Dependence Modeling : Vine Copula Handbook*, pages 219–231. World Scientific.

[Kao and Govindaraju, 2007] Kao, S.-C. and Govindaraju, R. S. (2007). A bivariate frequency analysis of extreme rainfall with implications for design. *Journal of Geophysical Research : Atmospheres*, 112(D13).

[Khaliq et al., 2006] Khaliq, M. N., Ouarda, T. B., Ondo, J.-C., Gachon, P., and Bobée, B. (2006). Frequency analysis of a sequence of dependent and/or non-stationary hydro-meteorological observations : A review. *Journal of hydrology*, 329(3-4) :534–552.

[Kim et al., 2007] Kim, G., Silvapulle, M. J., and Silvapulle, P. (2007). Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics & Data Analysis*, 51(6) :2836–2850.

[Klugman et al., 2013] Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2013). *Loss models : Further topics*. John Wiley & Sons.

[Kyselỳ et al., 2010] Kyselỳ, J., Picek, J., and Beranová, R. (2010). Estimating extremes in climate change simulations using the peaks-over-threshold method with a non-stationary threshold. *Global and Planetary Change*, 72(1-2) :55–68.

[Liu et al., 2016] Liu, Y., Racah, E., Correa, J., Khosrowshahi, A., Lavers, D., Kunkel, K., Wehner, M., Collins, W., et al. (2016). Application of deep convolutional neural networks for detecting extreme weather in climate datasets. *arXiv preprint arXiv :1605.01156*.

[Mantel, 1963] Mantel, N. (1963). Chi-square tests with one degree of freedom ; extensions of the mantel-haenszel procedure. *Journal of the American Statistical Association*, 58(303) :690–700.

[Nelsen, 2006] Nelsen, R. B. (2006). An introduction to copulas. springer, new york. *MR2197664*.

[Northrop et al., 2015] Northrop, P., Attalides, N., and Jonathan, P. (2015). Cross-validatory extreme value threshold selection and uncertainty with application to ocean storm severity. *arXiv preprint arXiv :1504.06653*.

[Renard and Lang, 2007] Renard, B. and Lang, M. (2007). Use of a gaussian copula for multivariate extreme value analysis : some case studies in hydrology. *Advances in Water Resources*, 30(4) :897–912.

[Riboust and Brissette, 2016] Riboust, P. and Brissette, F. (2016). Analysis of lake champlain/richelieu river's historical 2011 flood. *Canadian Water Resources Journal/Revue canadienne des ressources hydriques*, 41(1-2) :174–185.

[Salvadori and De Michele, 2006] Salvadori, G. and De Michele, C. (2006). Statistical characterization of temporal structure of storms. *Advances in Water Resources*, 29(6) :827–842.

[Salvadori and De Michele, 2007] Salvadori, G. and De Michele, C. (2007). On the use of copulas in hydrology : theory and practice. *Journal of Hydrologic Engineering*, 12(4) :369–380.

[Shaw et al., 2010] Shaw, E. M., Beven, K. J., Chappell, N. A., and Lamb, R. (2010). *Hydrology in practice*. CRC press.

[Small and Morgan, 1986] Small, M. J. and Morgan, D. J. (1986). The relationship between a continuous-time renewal model and a discrete markov chain model of precipitation occurrence. *Water Resources Research*, 22(10) :1422–1430.

[Vandenberghe et al., 2010] Vandenberghe, S., Verhoest, N., and De Baets, B. (2010). Fitting bivariate copulas to the dependence structure between storm characteristics : A detailed analysis based on 105 year 10 min rainfall. *Water resources research*, 46(1).

[Zhang and Singh, 2019] Zhang, L. and Singh, V. P. (2019). *Copulas and their applications in water resources engineering*. Cambridge University Press.

A Exemple illustrant le modèle

Pour illustrer les différentes composantes du modèle décrit dans la section 2.1, le tableau 6 utilise les 15 premières observations printanières de la base de données Clearwater River, pour 5 années consécutives.

l	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$Y_l^{(2009)}$	-	-	-	-	-	0.50	-	1.99	0.50	-	0.50	3.48	0.50	-	-
$Y_l^{(2010)}$	0.49	-	-	0.49	-	-	-	2.05	43.78	3.42	3.91	-	-	0.49	0.49
$Y_l^{(2011)}$	-	0.77	0.13	-	-	-	-	0.13	-	-	2.57	1.16	-	-	-
$Y_l^{(2012)}$	0.13	1.54	-	0.13	0.13	0.13	-	-	-	0.39	-	0.13	3.47	11.97	0.26
$Y_l^{(2013)}$	-	-	-	0.71	0.20	-	-	0.20	-	1.62	2.94	-	0.51	-	0.20

TABLEAU 6 – Observations de $Y_l^{(m)}$, pour $l = 1, \ldots, 15$ et $m = 2009, \ldots, 2013$, dans la base de données Clearwater River. Cette variable est représentée par la colonne `Precip. (mm)` et la période couvre du 1er au 15 avril des années 2009 à 2013.

Avec les observations du tableau 6, on peut réaliser le clustering de manière à trouver les résultats du tableau 7. Puis, si on fixe un seuil $u = 1$, on obtient $\mathcal{X}^{(2009)} = \{2, 3\}$, $\mathcal{X}^{(2010)} = \{3\}$, $\mathcal{X}^{(2011)} = \{3\}$, $\mathcal{X}^{(2012)} = \{1, 4\}$, $\mathcal{X}^{(2013)} = \{3\}$. On trouve alors les résultats des tableaux 8 et 9.

j	1	2	3	4	5
$\mathcal{C}_j^{(2009)}$	6	8,9	11,12,13	-	-
$\mathcal{C}_j^{(2010)}$	1	4	8,9,10,11	14,15	-
$\mathcal{C}_j^{(2011)}$	2,3	8	11,12	-	-
$\mathcal{C}_j^{(2012)}$	1,2	4,5,6	10	12,13,14,15	-
$\mathcal{C}_j^{(2013)}$	4,5	8	10,11	13	15
$K_j^{(2009)}$	0.50	2.49	4.48	-	-
$K_j^{(2010)}$	0.49	0.49	53.2	0.98	-
$K_j^{(2011)}$	0.90	0.13	3.73	-	-
$K_j^{(2012)}$	1.67	0.39	0.39	15.83	-
$K_j^{(2013)}$	0.91	0.20	4.56	0.51	0.20

TABLEAU 7 – Clusterisation des observations du tableau 6.

i	1	2	3
X_i^{2009}	2.49	4.48	-
X_i^{2010}	53.2	-	-
X_i^{2011}	3.73	-	-
X_i^{2012}	1.67	15.83	-
X_i^{2013}	4.56	-	-
D_i^{2009}	2	3	-
D_i^{2010}	4	-	-
D_i^{2011}	2	-	-
D_i^{2012}	2	4	-
D_i^{2013}	2	-	-
W_i^{2009}	7	1	≥ 2
W_i^{2010}	7	≥ 4	-
W_i^{2011}	10	≥ 3	-
W_i^{2012}	0	9	-
W_i^{2013}	9	≥ 4	-

TABLEAU 8 – Construction des v.a. relatives aux événements excédents le seuil $u = 1$, à partir du tableau 7.

Format date			
i	0	1	2
$T_i^{(2009)}$	2009-04-01	2009-04-08	2009-04-11
$T_i^{(2010)}$	2010-04-01	2010-04-08	-
$T_i^{(2011)}$	2011-04-01	2011-04-11	-
$T_i^{(2012)}$	2012-04-01	2012-04-01	2012-04-12
$T_i^{(2013)}$	2013-04-01	2013-04-10	-
Format numérique			
i	0	1	2
$T_i^{(2009)}$	0	7	10
$T_i^{(2010)}$	365	372	-
$T_i^{(2011)}$	730	740	-
$T_i^{(2012)}$	1096	1096	1107
$T_i^{(2013)}$	1461	1470	-

TABLEAU 9 – Temps d'arrivée des événements décrits dans le tableau 8.

Avec les tableaux 7 et 8, on peut calculer les résultats présentés dans le tableau 10.

m	2009	2010	2011	2012	2013
$N_{T_0}^{(m)}(15)$	2	1	1	2	1
$V_{T_0}^{(m)}(15)$	6.97	53.2	3.73	17.5	4.56
$Z^{(m)}$	0.5	1.96	1.03	0.78	1.82
$S_{T_0}^{(m)}(15)$	7.47	55.16	4.76	18.28	6.38

TABLEAU 10 – Représentation de la quantité de pluie totale tombée pendant les 15 premiers jours du printemps des années 2009 à 2013 selon la notation présentée dans la section 2.

B Illustrations

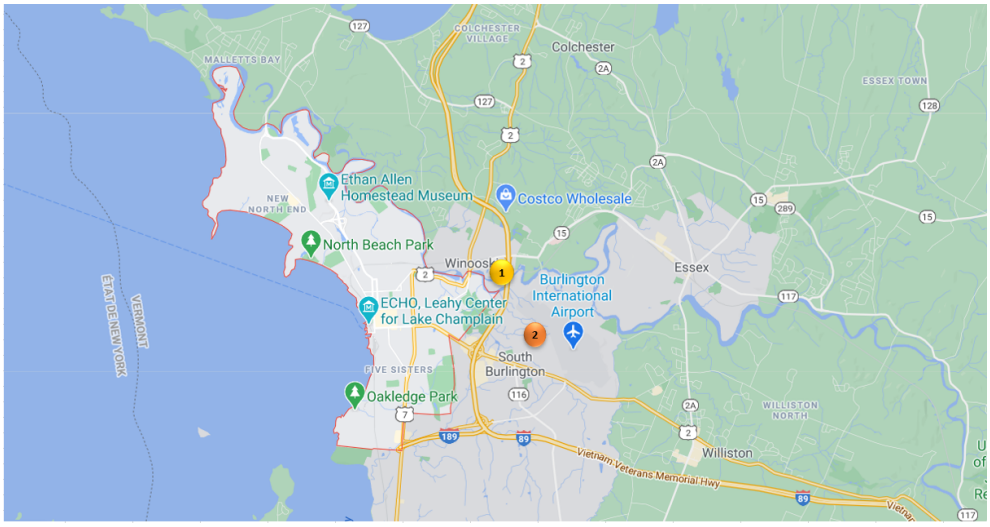


Illustration 5 – Emplacement des stations (1) `USC00431072` et (2) `USW00014742` pour les données de Burlington, au Vermont, USA. (Image tirée de Google Map)

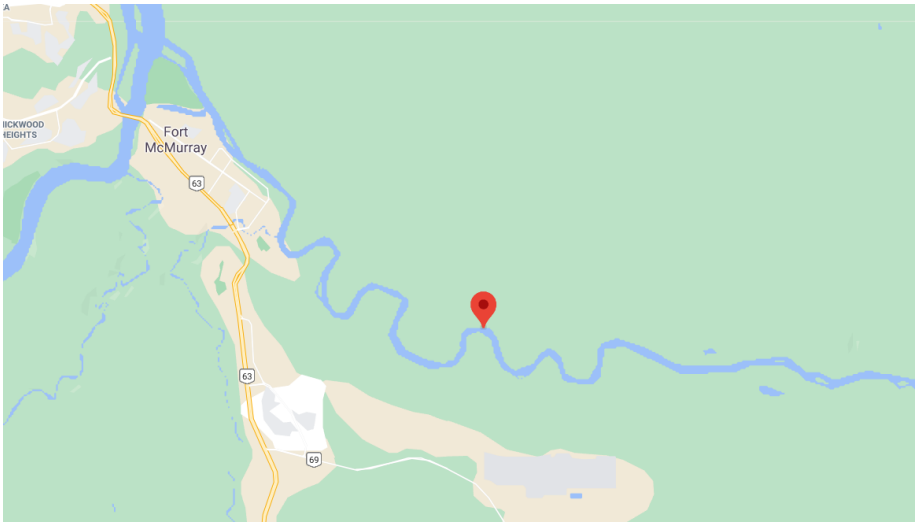
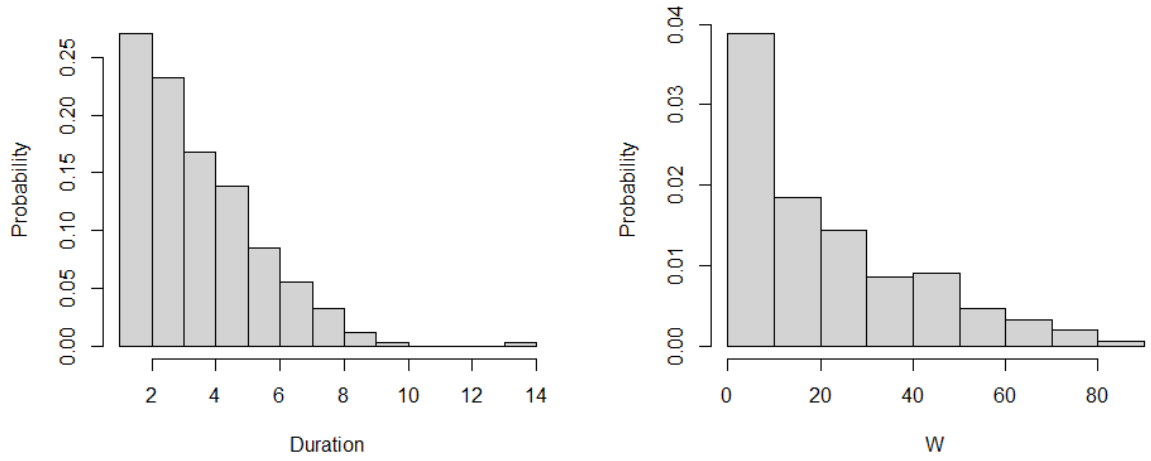
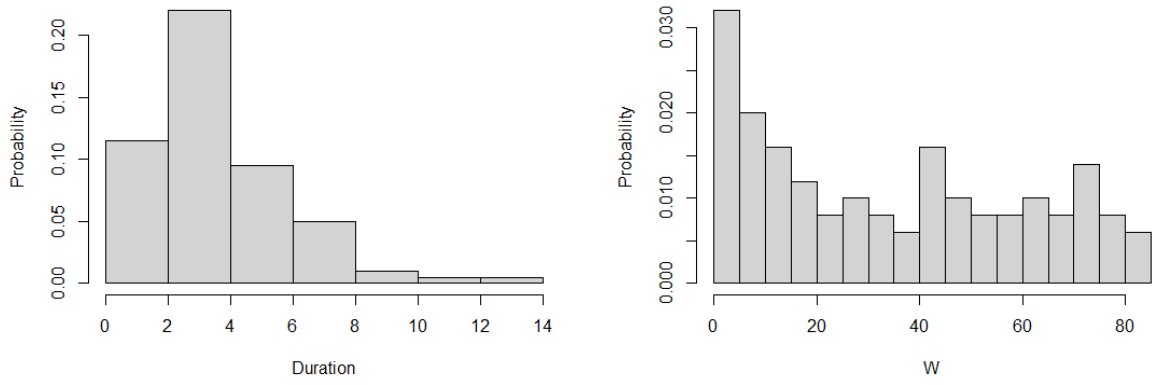


Illustration 6 – Emplacement de la station `07CD001` où ont été collectées les données pour l'exemple de la rivière Clearwater. Les coordonnées de la stations sont Longitude -111.2554 et Latitude 56.68528. (Image tirée de Google Map)



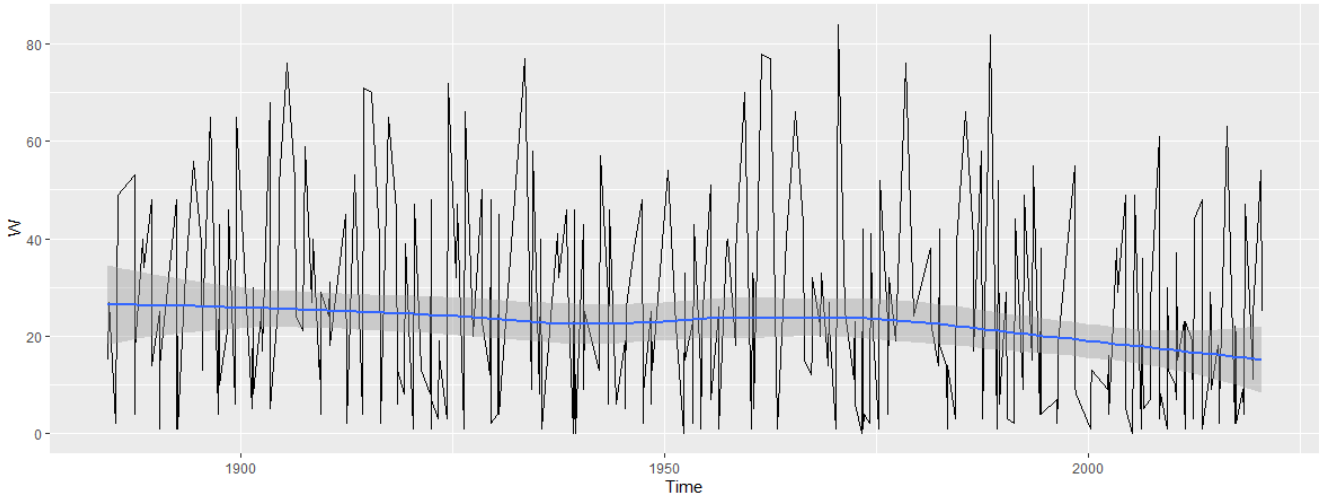
(a) D (b) W

Illustration 7 – Fonctions de masses de probabilités empiriques des v.a. W et D pour l'étude de cas de la section 3.1.

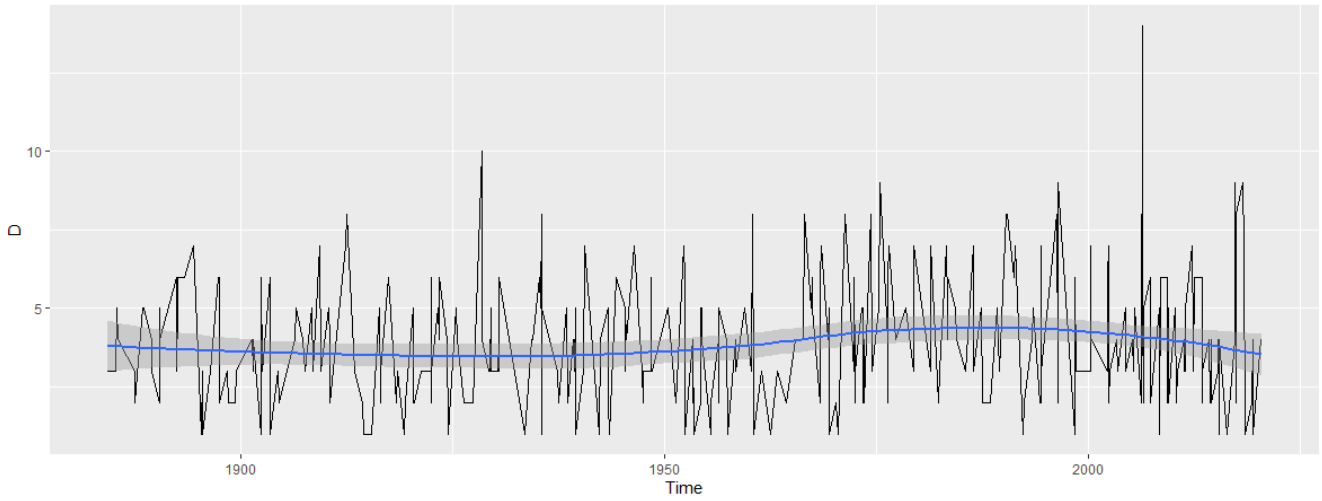


(a) D (b) W

Illustration 8 – Fonctions de masses de probabilités empiriques des v.a. W et D pour l'étude de cas de la section 3.2.

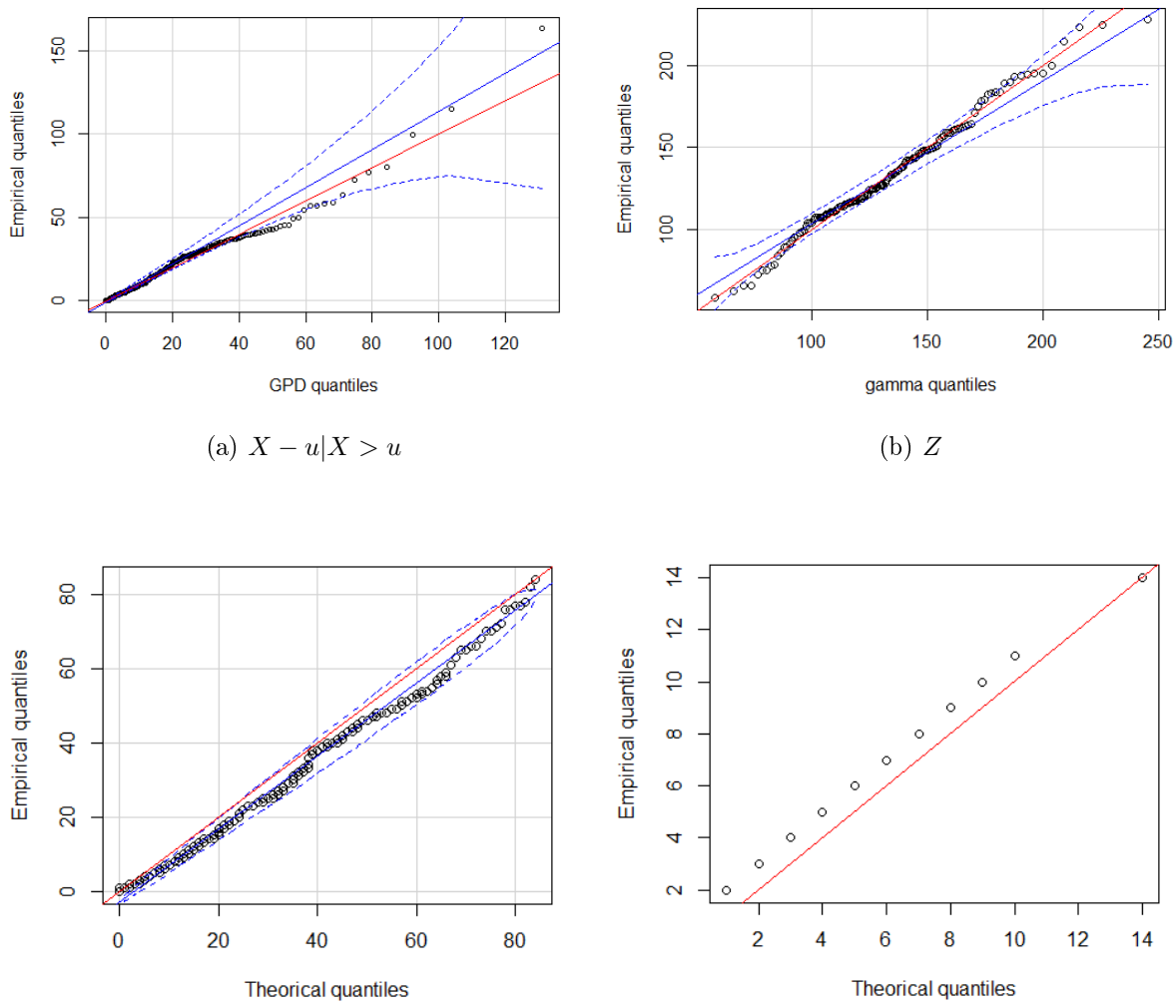


(a) $W_i^{(m)}$



(b) $D_i^{(m)}$

ILLUSTRATION 9 – Séries temporelles des v.a. $W_i^{(m)}$ et $D_i^{(m)}$, pour $i = 1, \ldots, 91$, $m = 1940, dots, 2020$.



(a) $X - u|X > u$

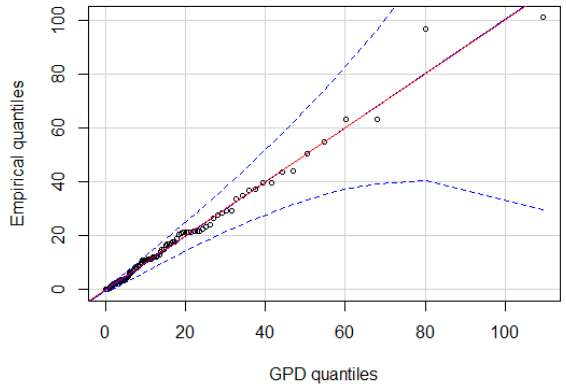
(b) Z

(c) W

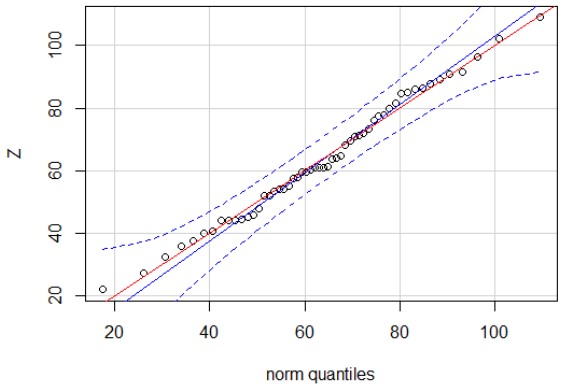
(d) D^*

ILLUSTRATION 10 – Diagrammes quantile-quantile pour l'étude de cas de la section 3.1 : La ligne bleue présente la droite de tendance des quantiles. Les pointillés présentent l'intervalle de confiance au seuil de 5% pour cette droite et la ligne rouge est la diagonale d'adéquation parfaite. L'objectif est que la diagonale rouge se situe dans l'intervalle de confiance.

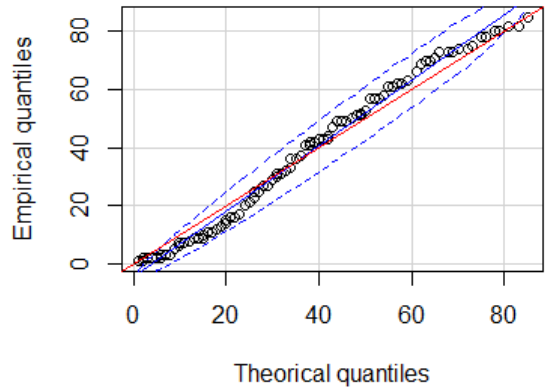
* : À noter qu'aucun intervalle de confiance n'est dessiné autour de la droite pour l'illustration 10d. La raison étant que la fonction `qqPlot` de la librairie `car` ne fonctionne pas bien lorsqu'il y a trop d'égalités dans les quantiles utilisés.



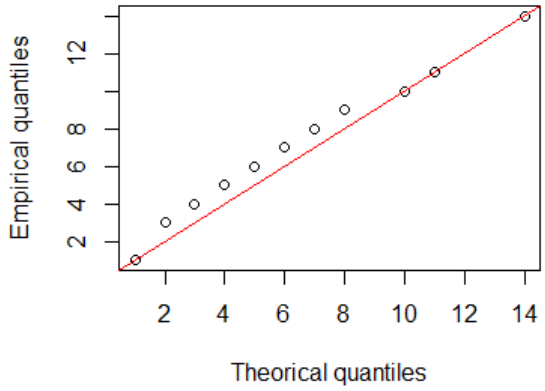
(a) $X - u | X > u$



(b) Z



(c) W



(d) D^*

ILLUSTRATION 11 – Diagrammes quantile-quantile pour l'étude de cas de la section 3.2 : La ligne bleue présente la droite de tendance des quantiles. Les pointillés présentent l'intervalle de confiance au seuil de 5% pour cette droite et la ligne rouge est la diagonale d'adéquation parfaite. L'objectif est que la diagonale rouge se situe dans l'intervalle de confiance.

* : À noter qu'aucun intervalle de confiance n'est dessiné autour de la droite pour l'illustration 11d. La raison étant que la fonction `qqPlot` de la librairie `car` ne fonctionne pas bien lorsqu'il y a trop d'égalités dans les quantiles utilisés.

C Algorithme de simulation

Soit $F_{X-u|X>u}$, F_W et F_D, les fonctions de répartition théoriques des v.a. $\{X-u|X>u\}$, W et D. Soit $F^{-1}_{X-u|X>u}$, F^{-1}_W et ${}^{-1}F_D$, les fonctions quantiles de ces mêmes v.a. La procédure de simulation permettant de générer des réalisations de $S(m)(91)$ est décrit dans l'algorithme 1.

Algorithm 1: Simuler un processus de renouvellement alterné avec récompense utilisant les copules pour modéliser la dépendance.

input : m (entiers) : vecteur des années pour lesquelles on désire simuler ;
t (entier) : la durée du processus (91 jours dans notre cas) ;
$N_{\max}$ (entier) : Le nombre maximum d'événements attendu dans une année (p. ex. 100) ;
n (entier) : nb de réalisations de la simulation (p. ex. $n = 10^3$).

output: matrice de dimension $n \times \text{card}(m)$ correspondant à des réalisations du processus $S^{(m)}_{T_0}(t)$.

foreach m **do**
  poser $T_0 =$ `as.numeric(as.Date("`m`-04-01"))` ;
  poser $t_{\max} = T_0 + t$;
  for $j \leftarrow 1$ **to** n **do**
    simuler $(u^{(X)}_l, u^{(W)}_l, u^{(D)}_l)$, $l = 1, \dots N_{\max}$, des réalisations de la copule G ;
    poser $i = 0$;
    do
      incrémenter $i = i + 1$;
      calculer $W_i = F^{-1}_W(u^{(W)}_i | T_{i-1})$, une réalisation de $W^{(m)}$;
      calculer $T_i = T_{i-1} + W_i$, une réalisation de $T^{(m)}$;
      calculer $D_i = F^{-1}_D(u^{(D)}_i | T_i)$, une réalisation de $D^{(m)}$;
      calculer $T_i = T_i + D_i$, une réalisation de $T^{\star(m)}$;
    while $T_i <= t_{\max}$
    calculer $N = i - 1$, une réalisation de $N^{(m)}_{T_0}(t)$;
    if $N = 0$ **then**
      poser $V = 0$;
      else
        for $i \leftarrow 1$ **to** N **do**
          calculer $X_i = u + F^{-1}_{X-u|X>u}(u^{(X)}_i)$, une réalisation de $X^{(m)}$;
        end
        calculer $V = \sum_{i=1}^{N} X_i$, une réalisation de $V^{(m)}_{T_0}(t)$;
      end
    end
    simuler Z, une réalisation de $Z^{(m)}$;
    calculer $S_{j,m} = V + Z$, une réalisation de $S^{(m)}_{T_0}(t)$;
  end
end
return S