

Titre: Modélisation conjointe des maxima annuels d'intensité de
précipitation

Auteur: Paul Mathivon

Date: 2023

Type: Mémoire ou thèse / Dissertation or Thesis

Référence: Mathivon, P. (2023). Modélisation conjointe des maxima annuels d'intensité de
précipitation [Master's thesis, Polytechnique Montréal]. PolyPublie.
Citation: <https://publications.polymtl.ca/54156/>

 **Document en libre accès dans PolyPublie**
Open Access document in PolyPublie

URL de PolyPublie: <https://publications.polymtl.ca/54156/>
PolyPublie URL:

**Directeurs de
recherche:** Jonathan Jalbert, & Arthur Charpentier
Advisors:

Programme: Maîtrise recherche en mathématiques appliquées
Program:

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

**Modélisation conjointe des maxima
annuels d'intensité de précipitation**

PAUL MATHIVON

Département de mathématiques et de génie industriel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Mathématiques appliquées

Juin 2023

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

**Modélisation conjointe des maxima
annuels d'intensité de précipitation**

présenté par **Paul MATHIVON**

en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Julie CARREAU, présidente

Jonathan JALBERT, membre et directeur de recherche

Christian GENEST, membre et codirecteur de recherche

Arthur CHARPENTIER, membre

REMERCIEMENTS

Je tiens à remercier mon directeur de recherche, Jonathan Jalbert, pour son accompagnement et ses conseils qui m'ont permis de développer le sens de la recherche et de la démarche scientifique. Ta bonne humeur et ton optimisme, Jonathan, m'ont aidé à surmonter les épreuves qui se sont présentées tout au long de ce projet de maîtrise. Je te remercie également pour les opportunités que tu m'as offertes. Par exemple, je me souviendrai longtemps de la chance que j'ai eue de pouvoir participer au Joint Statistical Meetings à Washington!

Je souhaite également remercier mon codirecteur de recherche, Christian Genest. Les échanges que nous avons pu avoir avec toi tout au long de la maîtrise nous ont permis à chaque fois de prendre du recul sur les travaux réalisés et nous ont permis d'explorer d'autres pistes lorsque cela était nécessaire. Ton sens du détail et de l'argumentation m'a permis de développer ma capacité à justifier mes choix.

Je veux dire merci aussi à mes collègues de laboratoire : Anne, Arian, Auguste, Duy, Gabriel et Léo. Merci pour votre bonne humeur et votre humour. Je garderai de bons souvenirs des journées passées au labo et des déjeuners à refaire le monde. Mention spéciale pour mes collègues québécois qui ont su me faire découvrir leur culture avec beaucoup d'humour et de passion. Promis, je ne me souviendrai pas seulement de la tradition du 5 à 7!

Merci aussi à tous les amis qui ont rendu ces deux années au Canada mémorables : les copains du 5060, ceux du 454 ou encore ceux du 901. Autant d'adresses où j'ai passé d'excellents moments et des soirées mémorables.

Enfin, je tiens à remercier Julie Carreau et Arthur Charpentier d'avoir accepté d'être membres du jury d'examen de ce mémoire.

RÉSUMÉ

Les courbes intensité-durée-fréquence (IDF) sont indispensables pour la prévention des incidents provoqués par les précipitations extrêmes. Au Canada, les courbes IDF officielles sont produites par Environnement et Changement Climatique Canada (ECCC). Actuellement, ECCC utilise la loi de Gumbel pour estimer les niveaux de retour d'intensité de précipitation. Cependant, cette modélisation est remise en question par les résultats les plus récents de la théorie des valeurs extrêmes et l'étude empirique des maxima de précipitation. De plus, la rareté des données incite à optimiser celles-ci le plus efficacement possible afin d'estimer les niveaux de retour d'intensité de précipitation plus précisément. Ce projet a pour objet d'étudier le modèle d-GEV. Celui-ci permet de combiner les observations de maxima d'intensité de précipitation pour l'ensemble des durées d'accumulation au sein d'un unique modèle paramétrique.

Le modèle d-GEV suppose que les maxima d'intensité de précipitation suivent une loi généralisée de valeurs extrêmes (GEV) pour chaque durée d'accumulation et que les paramètres de localisation et d'échelle de celles-ci soient fonction de la durée d'accumulation considérée. À l'aide des données recueillies à l'aéroport Trudeau à Montréal, on montre que le modèle d-GEV est prometteur en comparaison de l'estimation indépendante des paramètres des lois marginales GEV. Cependant, en raison de la corrélation entre les maxima d'intensité de précipitation pour différentes durées d'accumulation, la vraisemblance utilisée dans la modélisation d-GEV est impropre. En effet, celle-ci suppose que l'ensemble des observations soient indépendantes. En raison de cette vraisemblance impropre, les incertitudes associées aux niveaux de retour d'intensité de précipitation sont sous-estimées. Nous avons envisagé deux stratégies pour résoudre ce problème.

La première stratégie consiste à modéliser la corrélation inter-durée à l'aide d'une copule. On établit d'abord que pour les données étudiées, la corrélation peut être modélisée par une copule gaussienne dont les termes de la matrice de corrélation sont écrits à l'aide de la fonction de covariance de Matérn et une mesure de distance basée sur le logarithme des durées. Une étude de simulation permet de montrer que dans le cas où le modèle est convenablement spécifié, l'ajout d'une copule permet de corriger les incertitudes tout en réduisant celles-ci par rapport à l'estimation indépendante des lois GEV. Cependant, les incertitudes estimées avec ce modèle restent trop étroites pour les données de l'aéroport Trudeau. L'étude complète de la dépendance inter-durée des maxima d'intensité de précipitation nous a cependant permis d'aboutir à une modélisation satisfaisante de celle-ci qui peut être utile pour d'autres fins.

La seconde stratégie envisagée repose sur l'utilisation de méthodes robustes d'estimation des incertitudes. Ces méthodes sont adaptées lorsque la vraisemblance maximisée est impropre ou mal spécifiée. On montre, grâce à une étude de simulation, que la méthodologie d'estimation robuste permet de corriger partiellement les limitations du modèle d-GEV sans avoir à étudier la dépendance au sein des données. Cette approche peut ainsi être attractive pour de nombreux utilisateurs des courbes IDF.

Finalement, une comparaison des deux méthodes proposées est effectuée afin de guider l'utilisateur dans le choix d'une méthode d'estimation en fonction de ses besoins.

ABSTRACT

Intensity-duration-frequency curves (IDF curves) are essential for the prevention of incidents caused by extreme rainfall events. In Canada, official IDF curves are produced by Environment and Climate Change Canada (ECCC). However, their use of the Gumbel distribution to estimate the rainfall intensity return levels is challenged by the most recent developments in Extreme Value Theory (EVT). Moreover, data availability is limited, which encourages EVT practitioners to search for means of optimizing it. The purpose of this project is to study the ability of the duration GEV (d-GEV) model to represent rainfall intensity extremes jointly for all accumulation periods within a single parametric model.

In the d-GEV model, the marginal distribution of precipitation intensity annual maxima follows a Generalized Extreme Value (GEV) distribution for each accumulation period. The location and scale parameters of each GEV are written as a function of the accumulation period. Using the data from the Montréal Trudeau International Airport, we show that the fitted d-GEV model is promising compared to independently fitted GEV distributions. However, the likelihood of the d-GEV model is misspecified due to the correlation that is present in the maxima across accumulation periods. As a result of misspecification, estimated uncertainties associated with return levels are underestimated. Two solutions are considered to remedy this issue.

The first solution is to model inter-duration correlation with a copula. We show that the Gaussian copula with a Matérn covariance matrix is an adequate choice. A simulation study is carried out to show that this modeling strategy corrects the estimated uncertainties when the model is properly specified. However, the estimated uncertainties for the Trudeau Airport data remain too narrow.

Another option is to use robust methods based on Godambe's information matrix with improper composite likelihoods. A simulation study shows that this method corrects most shortcomings of the d-GEV model while avoiding the identification of the dependence structure. It might thus be considered an adequate approach for most IDF curve users.

A comparison of both strategies is offered to help practitioners choose the method which is best suited to their needs.

TABLE DES MATIÈRES

REMERCIEMENTS	iii
RÉSUMÉ	iv
ABSTRACT	vi
TABLE DES MATIÈRES	vii
LISTE DES TABLEAUX	x
LISTE DES FIGURES	xi
LISTE DES SIGLES ET ABRÉVIATIONS	xii
LISTE DES ANNEXES	xiii
CHAPITRE 1 INTRODUCTION	1
CHAPITRE 2 CADRE THÉORIQUE	4
2.1 La théorie des valeurs extrêmes	4
2.1.1 Intérêt d'une modélisation spécifique pour les maxima	4
2.1.2 Théorème fondamental de la théorie des valeurs extrêmes	5
2.1.3 Modèle statistique des maxima par blocs	7
2.1.4 Cas des séries temporelles et niveaux de retour	7
2.2 Modélisation de la dépendance entre plusieurs variables aléatoires	8
2.2.1 Intérêt des copules	8
2.2.2 Densité d'un modèle écrit avec une copule	9
2.2.3 Log-vraisemblance d'un modèle écrit avec une copule	10
2.2.4 Mesure de la dépendance	10
2.2.5 Copules max-stable	11
2.3 Estimation des incertitudes	12
2.3.1 Log-vraisemblance et fonction de score	12
2.3.2 Matrice d'information de Fisher	12
2.3.3 Normalité asymptotique des estimateurs du maximum de vraisemblance	13
2.3.4 Lien avec l'information de Fisher observée	13
2.3.5 Méthode Delta	14

2.4	Dérivation automatique	15
2.4.1	Objets duaux	15
2.4.2	Surcharge des opérateurs mathématiques	16
2.4.3	Règle de dérivation en chaîne	17
2.4.4	Lien entre les nombres duaux et la dérivation en chaîne	18
2.4.5	Dimensions supérieures	18
2.4.6	Pratique	19
CHAPITRE 3 REVUE DE LITTÉRATURE		20
3.1	Méthodologie actuelle d'estimation des courbes IDF	20
3.2	Modèles d'échelle pour l'estimation des courbes IDF	21
3.3	Modélisation de la structure de dépendance	23
CHAPITRE 4 APPLICATION INITIALE DU MODÈLE D-GEV ET EXPLORA- TION DE LA CORRÉLATION		25
4.1	Présentation des données	25
4.1.1	Origines des données	25
4.1.2	Transformation des données	25
4.2	Application du modèle d-GEV aux données de l'aéroport Trudeau	26
4.2.1	Méthodologie	26
4.2.2	Résultats	27
4.2.3	Conclusion sur l'application de la méthodologie d-GEV simple	27
4.3	Exploration de la corrélation dans les données	29
4.4	Conclusions de l'analyse exploratoire	31
4.4.1	Résultats de l'exploration	31
4.4.2	Stratégies de résolution	32
CHAPITRE 5 MODÉLISATION DE LA DÉPENDANCE ENTRE LES MAXIMA ANNUELS AVEC UNE COPULE		33
5.1	Proposition d'un modèle pour la corrélation	33
5.1.1	Modélisation de la corrélation	33
5.1.2	Choix d'une copule	34
5.1.3	Considérations sur la mesure de corrélation et la copule de Student	36
5.1.4	Choix de la forme paramétrique la corrélation	36
5.1.5	Modèle complet et méthodologie d'estimation des paramètres	39
5.1.6	Résultats	41
5.1.7	Conclusions intermédiaires	44

5.2	Étude de simulation : Taux de couverture des intervalles de confiance lorsque les données sont générées par le modèle proposé	45
5.2.1	Objectif de la démarche	45
5.2.2	Résultats attendus	46
5.2.3	Données synthétiques	46
5.2.4	Procédure d'estimation	47
5.2.5	Résultats	48
5.2.6	Conclusion de l'étude de simulation	51
CHAPITRE 6	MODÉLISATION DE L'ERREUR DE SPÉCIFICATION	53
6.1	Vraisemblance composite simple	53
6.2	Vraisemblance composite par paires	54
6.3	Estimation des incertitudes	54
6.4	Étude de simulation : Méthodes par vraisemblance composite	55
6.4.1	Génération des données	55
6.4.2	Présentation des résultats	56
6.4.3	Analyse des résultats	57
6.5	Résultats de l'application de la méthode composite simple aux données réelles	58
CHAPITRE 7	ESTIMATION DES COURBES IDF DES PRÉCIPITATIONS	60
7.1	Choix de la méthode	60
7.1.1	Avantage des deux méthodes par rapport à l'estimation indépendante des lois marginales GEV	60
7.1.2	Méthodologie avec copule	60
7.1.3	Méthodologie avec une vraisemblance composite	61
7.2	Estimation des courbes IDF pour la station de Montréal Trudeau	62
CHAPITRE 8	CONCLUSION ET RECOMMANDATIONS	64
8.1	Synthèse des travaux	64
8.2	Limites des modèles et améliorations envisagées	65
RÉFÉRENCES		66
ANNEXES		68

LISTE DES TABLEAUX

Tableau 5.1	Niveau de retour $T = 100$ estimés par les 3 méthodes et bornes des intervalles de confiance au niveau nominal 95% (mm/h)	44
Tableau 5.2	Taux de couverture des intervalles de confiance estimés au niveau nominal 95% et écart-type moyen estimé avec la méthode delta pour $\rho = 3$ et $\kappa = 1$	51
Tableau 5.3	Taux de couverture des intervalles de confiance estimés au niveau nominal 95% et écart-type moyen estimé avec la méthode delta pour $\rho = 1$ et $\kappa = 0.5$	51
Tableau 6.1	Résultats de l'étude de simulation pour les méthodes avec vraisemblance composite. $n = 75$	56
Tableau 6.2	Résultats de l'étude de simulation pour les méthodes avec vraisemblance composite. $n = 200$	56
Tableau A.1	Taux de couverture des intervalles de confiance estimés au niveau nominal 95% et écart-type moyen estimé avec la méthode Delta pour $\rho = 3$ et $\kappa = 1$	68
Tableau A.2	Taux de couverture des intervalles de confiance estimés au niveau nominal 95% et écart-type moyen estimé avec la méthode Delta pour $\rho = 1$ et $\kappa = 0.5$	68
Tableau B.1	Résultats de l'étude de simulation pour la méthode composite simple avec un niveau plus élevé de dépendance. $n = 75$	69

LISTE DES FIGURES

Figure 1.1	Courbes IDF officielles pour l'Aéroport Pierre-Eliot-Trudeau de Montréal	1
Figure 4.1	Diagrammes quantile-quantile obtenus pour chacune des méthodes pour 4 durées	28
Figure 4.2	Diagrammes de dispersion pour la série des maxima de 5 minutes contre une sélection de durées, Aéroport Trudeau	29
Figure 4.3	Diagrammes de dispersion pour la série des maxima de 1 heure contre une sélection de durées, Aéroport Trudeau	30
Figure 4.4	Corrélation mesurée par le tau de Kendall empirique entre les maxima annuels d'intensité de précipitation pour l'ensemble des durées d'accumulation	31
Figure 5.1	Représentation de $\sin(\pi\hat{\tau}/2)$ en fonction de h	37
Figure 5.2	Ajustement de la relation fonctionnelle de Matérn par la méthode des moindres carrés	38
Figure 5.3	Log-vraisemblance maximisée en fonction du nombre de degrés de liberté de la copule de Student	42
Figure 5.4	Diagrammes quantiles-quantiles obtenus pour le modèle d-GEV avec et sans copule pour une sélection de durées	42
Figure 5.5	Ajustements des corrélations obtenues par maximisation de la vraisemblance et par minimisation des erreurs quadratiques	43
Figure 5.6	Fonctions de Matérn utilisées pour l'étude de simulation	47
Figure 5.7	Diagrammes en boîte des résultats de l'étude de simulation pour une sélection de durées	48
Figure 6.1	Comparaison des résultats de l'étude de simulation pour les méthodes composites et la méthode d-GEV avec copule	57
Figure 6.2	Diagrammes Quantile-Quantile et intervalles de confiance pour la méthodologie d-GEV composite et classique	59
Figure 7.1	Courbes IDF obtenues avec la méthodologie de la vraisemblance composite simple	62

LISTE DES SIGLES ET ABRÉVIATIONS

ACN	Association canadienne de normalisation
ECCC	Environnement et Changement Climatique Canada
GEV	Loi Généralisée de Valeurs Extrêmes
IDF	Intensité-Durée-Fréquence
OMM	Organisation Mondiale de Météorologie

LISTE DES ANNEXES

Annexe A	Résultats de l'étude de simulation pour des jeux de données plus grands	68
Annexe B	Résultats complémentaires pour l'étude de simulation portant sur les méthodes composite	69

CHAPITRE 1 INTRODUCTION

Les précipitations extrêmes ont un impact majeur sur les activités humaines. Elles peuvent notamment endommager les infrastructures, provoquer des glissements de terrain ou encore entraîner des inondations printanières comme celles survenues en 2019 au Québec. La caractérisation de ces événements extrêmes en termes de magnitude et de fréquence d'occurrence est ainsi primordiale afin de convenablement protéger les infrastructures sensibles et dimensionner les systèmes de retenue d'eau ou d'écoulement des eaux de pluie.

L'Association canadienne de normalisation (CSA) impose l'usage de courbes Intensité-Durée-Fréquence (IDF) à ces fins. Les courbes IDF permettent de décrire, pour une période de retour T donnée, la relation entre la durée d'accumulation et le niveau de retour associé, c'est-à-dire l'intensité de précipitation qui est dépassée en moyenne toutes les T années. Au Canada, les courbes IDF faisant figure d'autorité sont produites par Environnement et Changement Climatique Canada (ECCC). Elles représentent les niveaux de retour pour des périodes de retour allant de 2 à 100 ans et des durées d'accumulation allant de 5 minutes à 24 heures. La figure 1.1 illustre les courbes IDF produites par ECCC des précipitations enregistrées à l'Aéroport international Pierre-Elliott-Trudeau de Montréal.

Actuellement, les courbes IDF produites par ECCC sont estimées indépendamment pour chacune des durées. Les maxima annuels correspondant à une durée d'accumulation donnée sont modélisée par une loi de Gumbel ajustée par la méthode des l-moments (Hogg, 1989) et les quantiles extrêmes d'intérêts sont estimés. Ensuite, une interpolation linéaire dans l'espace

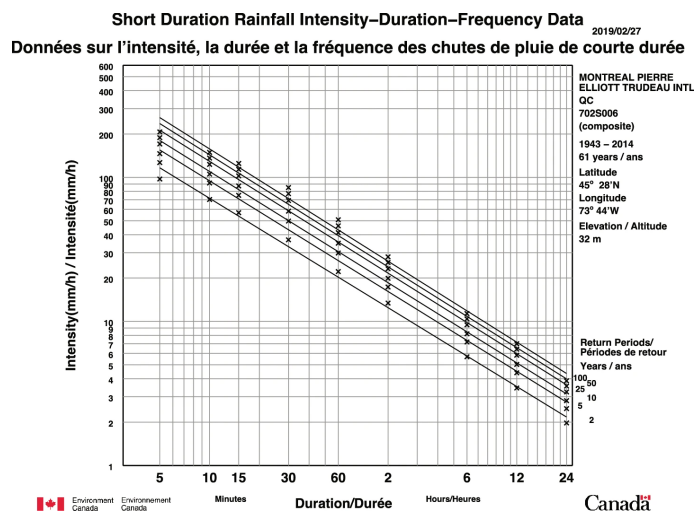


FIGURE 1.1 Courbes IDF officielles pour l'Aéroport Pierre-Elliott-Trudeau de Montréal

logarithmique relie tous les quantiles du même ordre des différentes durées d'accumulation. Cette procédure n'utilise pas les récents développements en théorie des valeurs extrêmes et modélisation de la dépendance. D'une part, l'utilisation de la loi de Gumbel n'est pas cohérente avec la théorie des valeurs extrêmes et l'estimation indépendante pour chacune des durées d'accumulation ne tire pas profit de toute l'information disponible pour estimer les quantiles extrêmes.

En théorie des valeurs extrêmes, l'usage de la loi Généralisée de Valeurs Extrêmes (GEV) permet une estimation rigoureuse des précipitations intenses correspondant à des périodes de retour élevées. En particulier, la capacité de cette loi à extrapoler la distribution des intensités de précipitation extrêmes au-delà des quantités observées est un atout majeur. Cependant, l'estimation des paramètres de la loi GEV nécessite une quantité importante de données. Cela représente un facteur limitant dans le cas des précipitations extrêmes car les stations gérées par ECCC comptabilisent en moyenne 25 années de données.

La faible quantité de données disponible entraîne une variance importante pour les estimateurs des paramètres de la loi GEV. Ceux-ci ne sont pas biaisés grâce à l'usage de la loi GEV mais pour autant, les estimations des niveaux de retour sont entourées d'une grande incertitude. De plus, pour une période de retour donnée, il existe une relation d'ordre entre les niveaux de retour pour les différentes durées. Plus la durée considérée est courte, plus le niveau de retour en intensité est grand. Des estimations de la loi GEV menées indépendamment pour chaque durée d'accumulation ne permettent pas de s'assurer que cette contrainte physique soit respectée.

Par ailleurs, cette contrainte physique signifie qu'il existe des informations partagées par les maxima d'intensité de précipitation pour les différentes durées d'accumulation. Cela est particulièrement visible sur la figure 1.1, où l'on observe que pour une période de retour donnée, les niveaux de retour pour les différentes durées sont distribués de part et d'autre d'une droite. Étant donné qu'une échelle logarithmique a été utilisée pour les deux axes, cela signifie que les niveaux de retour décroissent selon une loi de puissance selon la durée. Une telle décroissance peut être décrite en imposant que les paramètres des lois GEV soient fonction de la durée d'accumulation. On appelle les modèles associés à ces relations fonctionnelles des modèles d'échelles. Le modèle d-GEV que l'on étudiera dans ce mémoire est l'un de ces modèles. L'avantage escompté d'utiliser un modèle conjoint est de combiner les informations contenues dans les séries de maxima d'intensité afin de réduire les incertitudes d'estimation des niveaux de retour. Le niveau de confiance dans les courbes IDF obtenues serait alors plus important.

Cependant, contrairement à la loi GEV qui a un caractère asymptotiquement exact pour les maxima d'intensité de précipitation grâce à la théorie des valeurs extrêmes, le modèle d-GEV est un modèle dont il faut vérifier l'adéquation pour l'estimation des niveaux de retour d'intensité de précipitation. Il existe, en effet, un risque que celui-ci ne soit pas une représentation adéquate de la réalité. Il est donc primordial d'évaluer l'aptitude du modèle d-GEV à décrire la loi des maxima d'intensité de précipitation et à estimer convenablement les incertitudes associées.

Pour une durée d'accumulation donnée, on émet généralement l'hypothèse que les maxima annuels d'intensité sont indépendants. Cette hypothèse semble raisonnable car les événements climatiques qui génèrent les précipitations extrêmes sont éloignés dans le temps et ne correspondent pas aux mêmes systèmes météorologiques. Une hypothèse d'indépendance n'est cependant pas raisonnable si l'on considère les maxima d'intensité de précipitation des différentes durées survenus la même année. En effet, l'événement climatique qui génère les maxima d'intensité pour deux durées différentes peut fort bien être le même. Ainsi, si l'on décide d'utiliser un modèle conjoint pour l'ensemble des durées, il est nécessaire de s'interroger sur la dépendance qui peut éventuellement être présente dans les données et de la modéliser. La contribution de ce mémoire est de mesurer cette dépendance, de la modéliser et d'évaluer ses effet sur l'estimation des niveaux de retour et les incertitudes associées.

Étant donné les problématiques relevées à propos du modèle d-GEV, les objectifs spécifiques de ce mémoire sont donc les suivants :

1. Modéliser les maxima annuels d'intensité de précipitation à l'aide du modèle d-GEV.
2. Modéliser la dépendance entre les maxima d'intensité de précipitation par une copule.
3. Proposer une méthode d'estimation des incertitudes associées au modèle d-GEV.

CHAPITRE 2 CADRE THÉORIQUE

2.1 La théorie des valeurs extrêmes

Dans de nombreuses applications, on s'intéresse à la queue de distribution d'une variable aléatoire dont la distribution sous-jacente est inconnue. Comme nous allons le voir, les outils statistiques usuels sont souvent mal adaptés pour l'estimation des quantiles les plus élevés. La théorie des valeurs extrêmes a pour vocation d'apporter une réponse à cette limitation. Pour de plus amples détails, on pourra se référer à l'ouvrage de référence de Coles (2001).

2.1.1 Intérêt d'une modélisation spécifique pour les maxima

Soit X une variable aléatoire d'intérêt ayant F_X pour fonction de répartition absolument continue. Imaginons que l'on dispose de n réalisations indépendantes et identiquement distribuées (*iid*) de X , dénotées $X_1, \dots, X_n$. Si la loi F_X est connue, on peut immédiatement caractériser la loi du maximum $M_n = \max(X_1, \dots, X_n)$ en calculant, pour tout $x \in \mathbb{R}$,

$$F_{M_n}(x) = \mathbb{P}(M_n \leq x) = \mathbb{P}(X_1 \leq x) \times \dots \times \mathbb{P}(X_n \leq x).$$

Or, comme les réalisations sont *iid*, on a que pour tout $x \in \mathbb{R}$, $\mathbb{P}(M_n \leq x) = \{\mathbb{P}(X_1 \leq x)\}^n$. Il s'ensuit que, pour tout $x \in \mathbb{R}$,

$$F_{M_n}(x) = F_X^n(x).$$

Dans la pratique, la loi F_X n'est pas connue et on doit l'estimer pour utiliser cette approche. Deux cas de figures se présentent alors.

- Premier cas : On suppose que X appartient à une certaine famille paramétrique de lois, caractérisée par un vecteur de paramètre $\theta \in \Theta$. Alors, par certaines méthodes, comme par exemple la maximisation de la vraisemblance, on peut obtenir un estimateur $\hat{\theta}$ et un estimateur de la fonction de répartition $\hat{F}_{X|\hat{\theta}}$.
- Second cas : on peut utiliser l'estimateur empirique non paramétrique de F , à savoir

$$\hat{F}_X(x) = \frac{1}{n} \text{Card}(i : X_i \leq x),$$

où n dénote le nombre d'observations utilisées.

Dans le premier cas, on a $\hat{F}_{M_n} = \hat{F}_{X|\hat{\theta}}^n$. Le problème vient alors du fait que toute erreur dans l'estimation de θ est exacerbée par le passage à la puissance n , ce qui rend l'estimation très instable. Si l'appartenance de X à la famille paramétrique n'est pas avérée, le biais provoqué par la mauvaise spécification peut lui aussi être exacerbé.

Dans le second cas, si x est supérieur au maximum m_n observé alors on a $\hat{F}(x) = 1$ et $\hat{F}_{M_n}(x) = 1$, ce qui rend la procédure inefficace lorsqu'on cherche à connaître la probabilité associée à des événements qui n'ont pas été observés.

En réponse à ces limitations, la théorie des valeurs extrêmes permet de s'affranchir d'estimer la fonction de répartition de la variable aléatoire sous-jacente. En effet, la théorie des valeurs extrêmes permet d'estimer directement F_{M_n} à l'aide d'une famille paramétrique sans passer par l'estimation de F_X .

2.1.2 Théorème fondamental de la théorie des valeurs extrêmes

Le théorème de von Mises et Jenkinson va nous permettre d'associer une distribution paramétrique à M_n lorsque n tend vers l'infini. Son énoncé est le suivant.

Théorème 1 (*von Mises et Jenkinson*)

Soit $X_1, \dots, X_n$ une suite de variables aléatoires mutuellement indépendantes et identiquement distribuées (iid) de fonction de répartition F_X . Soit $M_n = \max(X_1, \dots, X_n)$.

S'il existe deux suites de réels $(a_n)_{n \in \mathbb{N}}$ et $(b_n)_{n \in \mathbb{N}}$ telles que $a_n > 0$ et

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\frac{M_n - b_n}{a_n} \leq z \right) = G(z),$$

où G est une fonction de répartition non-dégénérée, alors G appartient à la famille des lois généralisées de valeurs extrêmes (GEV). Cette famille est définie par sa fonction de répartition, soit

$$G^*(z) = \exp \left[- \left\{ 1 + \xi \left(\frac{z - \mu^*}{\sigma^*} \right) \right\}^{-1/\xi} \right]$$

où $\mu^* \in \mathbb{R}$, $\sigma^* > 0$ et $\xi \in \mathbb{R} \setminus \{0\}$. Une forme limite est possible pour G^* dans le cas où $\xi \rightarrow 0$.

Quelques remarques peuvent être faites sur ce résultat. En effet, on remarque que l'utilisation de ce théorème est conditionnée par l'existence de deux suites (a_n) et (b_n) qui nous permettent de renormaliser le maximum M_n afin que la fonction de répartition limite ne soit pas dégénérée. Si une telle renormalisation n'est pas possible, il n'existe pas de résultats sur la loi limite. On constate aussi que ce résultat est un résultat asymptotique, c'est-à-dire qu'il n'est valide que dans la limite $n \rightarrow \infty$. L'utilisation de la loi GEV pour des données réelles finies représente donc toujours une approximation et doit être sujette à caution.

Lorsque nous travaillerons sur des données réelles, on supposera qu'il est effectivement possible de trouver des constantes de renormalisation et que la loi limite non-dégénérée est donc une loi GEV. De plus, en supposant que n soit suffisamment grand pour faire l'approximation de la loi GEV, on aura

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq z\right) \approx G^*(z),$$

soit encore

$$\mathbb{P}(M_n \leq z) \approx G^*\{(z - b_n)/a_n\}.$$

Or on a

$$G^*\{(z - b_n)/a_n\} = \exp\left[-\left\{1 + \xi \left(\frac{\frac{z-b_n}{a_n} - \mu^*}{\sigma^*}\right)\right\}^{-1/\xi}\right].$$

En posant $\sigma = \sigma^* \times a_n$ et $\mu = \mu^* \times a_n + b_n$, on trouve

$$G^*\{(z - b_n)/a_n\} = \exp\left[-\left\{1 + \xi \left(\frac{z - \mu}{\sigma}\right)\right\}^{-1/\xi}\right].$$

Finalement, on trouve

$$\mathbb{P}(M_n < z) \approx G(z),$$

où G appartient à la famille GEV avec paramètres μ , σ et ξ . Nous n'aurons donc pas à nous soucier dans la pratique de trouver les constantes de renormalisation a_n et b_n .

La loi GEV constitue ainsi une famille paramétrique avec trois paramètres. On appelle μ le paramètre de localisation, σ le paramètre d'échelle et ξ le paramètre de forme. La valeur de ξ est particulièrement déterminante pour l'estimation des quantiles supérieurs. En fonction de la loi de X , la variable aléatoire sous-jacente, le paramètre ξ peut appartenir à trois domaines :

(i) $\xi < 0$ lorsque la queue de distribution de X est bornée à droite.

(ii) $\xi = 0$ lorsque la queue de distribution décroît exponentiellement. C'est le cas, par

exemple, de la loi normale. On parle alors de queue légère.

- (iii) $\xi > 0$ lorsque la queue de distribution décroît selon une loi de puissance. On parle alors de queue lourde.

2.1.3 Modèle statistique des maxima par blocs

Pour l'estimation des paramètres GEV, nous allons utiliser la technique des maxima par blocs et la propriété de max-stabilité de la loi GEV. Définissons tout d'abord cette dernière propriété.

Définition 1 *Max-stabilité*

On dit qu'une loi est max-stable si, pour tout entier $n \in \mathbb{N}$, il existe des constantes $\alpha_n > 0$ et β_n telles que

$$G^n(\alpha_n x + \beta_n) = G(x).$$

Cela signifie que si $X \sim G$, où G est max-stable, alors il existe une renormalisation telle que $M_n = \max(X_1, \dots, X_n)$ suive aussi la loi G .

Supposons que l'on a $n = k \times m$, où k , le nombre de blocs, et m la taille des blocs, sont des entiers positifs. On peut donc écrire $M_n = \max(M_{m,1}, \dots, M_{m,k})$, où $M_{m,i} = \max(X_{(i-1)m+1}, \dots, X_{im})$. On a $F_{M_n} = F_{M_{m,1}}^k$ car les variables sont *iid*. En supposant n assez grand et par extension m assez grand, on a $M_{m,i} \sim \text{GEV}(\mu, \sigma, \xi)$ et l'on dispose de k échantillons pour estimer μ , σ et ξ en utilisant, par exemple, la méthode du maximum de vraisemblance. Comme la loi GEV est max-stable (c'est d'ailleurs la seule loi max-stable), on pourra retrouver à l'aide d'une transformation les paramètres GEV pour M_n .

2.1.4 Cas des séries temporelles et niveaux de retour

Dans le cadre de cette maîtrise, nous serons amenés à travailler avec des séries temporelles de précipitation. Soit $(X_t)_{t \geq 0}$ une série temporelle. Les pas successifs X_{t_i} d'une série temporelle ne sont généralement pas indépendants et identiquement distribués, ce qui ne correspond donc plus au cadre présenté précédemment. Cependant, des théorèmes plus avancés de théorie des valeurs extrêmes nous permettent d'utiliser malgré tout l'approximation GEV pour $M_n = \max(X_{t_1}, \dots, X_{t_n})$ pour n suffisamment grand. En général, on travaille avec les maxima annuels de la série.

Les niveaux de retour sont une mesure courante utilisée pour l'analyse des risques. Soit T une durée en année que l'on appelle la période de retour. Le niveau de retour associé à T

correspond au quantile qui est dépassé en moyenne une fois toutes les T années. De manière équivalente, il s'agit du quantile qui a, chaque année, une probabilité $1/T$ d'être dépassé. Par exemple, le niveau de retour de précipitation associé à $T = 100$ ans est la quantité de précipitation qui a 1% de chance d'être excédée chaque année. La forme des niveaux de retour pour la loi GEV basée sur les maximas annuels est donnée par

$$Q_T = \mu + \frac{\sigma}{\xi} \left[\left\{ -\ln \left(1 - \frac{1}{T} \right) \right\}^{-\xi} - 1 \right] \quad \text{si } \xi \neq 0,$$

$$Q_T = \mu - \sigma \times \ln \left[\left\{ -\ln \left(1 - \frac{1}{T} \right) \right\} \right] \quad \text{si } \xi = 0.$$

2.2 Modélisation de la dépendance entre plusieurs variables aléatoires

Dans le cadre de ce projet de maîtrise, nous serons amenés à modéliser la dépendance entre les maxima annuels d'intensité de précipitation pour différentes durées d'accumulation. Pour cela, il est donc nécessaire de s'intéresser aux techniques de modélisation de la dépendance entre plusieurs variables aléatoires corrélées.

2.2.1 Intérêt des copules

Les copules sont des fonctions de répartition multidimensionnelles définies pour des vecteurs aléatoires dont les marginales sont uniformes sur l'intervalle $[0, 1]$. Les copules permettent d'expliciter la dépendance entre plusieurs variables aléatoires réelles indépendamment de leurs lois marginales en transformant les données à l'aide de leur fonction de répartition.

Définition 2 (*Copule*) : Une copule, notée C , est une fonction de $[0, 1]^d \rightarrow [0, 1]$, où d représente la dimension, qui possède les caractéristiques suivantes :

$$(i) \ C(u_1, \dots, u_d) = 0 \text{ s'il existe } u_i = 0 ;$$

$$(ii) \ C(1, \dots, u_i, 1, \dots, 1) = u_i ;$$

$$(iii) \ C \text{ est } d\text{-croissante.}$$

Cette définition est suffisante pour caractériser la fonction de répartition d'un vecteur aléatoires de dimension d à marginales uniformes sur $[0, 1]$. Pour faire le lien entre les variables aléatoire réelles continues et les marges uniformes sur $[0, 1]$ on utilise le théorème de Sklar.

Théorème 2 (Sklar)

Soit F la fonction de répartition conjointe des variables aléatoires absolument continues $X_1, \dots, X_d$ de fonctions de répartition marginales $F_{X_1}, \dots, F_{X_d}$. Il existe une unique copule C telle que

$$C(u_1, \dots, u_d) = F\{F_{X_1}^{-1}(u_1), \dots, F_{X_d}^{-1}(u_d)\},$$

où $F_{X_i}^{-1}$ représente la fonction quantile de X_i donnée par $F_i^{-1}(u) = \inf\{x : F_{X_i}(x) \geq u\}$. On dit alors que C est la copule associée à $(X_1, \dots, X_d)$.

Inversement, la fonction de répartition de $(X_1, \dots, X_d)$ est telle que

$$F_{(X_1, \dots, X_d)}(x_1, \dots, x_d) = C\{F_1(x_1), \dots, F_d(x_d)\}. \quad (2.1)$$

Le théorème de Sklar permet ainsi de séparer la structure de dépendance, qui n'apparaît plus que dans les marges uniformes, et les lois marginales. Pour générer des observations du vecteur aléatoire $(X_1, \dots, X_d)$, il suffit de générer des observations de $(U_1, \dots, U_d)$, le vecteur aléatoires de marges uniformes et de fonction de répartition C , puis d'appliquer la relation

$$(X_1, \dots, X_d) = (F_{X_1}^{-1}(U_1), \dots, F_{X_d}^{-1}(U_d)).$$

La difficulté restante peut alors être de générer un vecteur aléatoire de fonction de répartition C et de calculer les inverses des fonctions de répartitions marginales.

2.2.2 Densité d'un modèle écrit avec une copule

À partir de la formule de la fonction de répartition donnée à l'équation (2.1), on peut établir la densité $f_{(X_1, \dots, X_d)}$ d'un vecteur aléatoire $(X_1, \dots, X_d)^\top$. Il faut pour cela dériver partiellement $F_{(X_1, \dots, X_d)}(x_1, \dots, x_d)$ par rapport à l'ensemble des composantes x_1 à x_d . On a ainsi

$$f_{(X_1, \dots, X_d)}(x_1, \dots, x_d) = \frac{\partial^d (x_1, \dots, x_d)}{\partial x_1 \dots \partial x_d} F_{(X_1, \dots, X_d)}.$$

On remplace par l'expression de l'équation (2.1), ce qui donne

$$f_{(X_1, \dots, X_d)}(x_1, \dots, x_d) = \frac{\partial^d}{\partial x_1 \dots \partial x_d} C\{F_1(x_1), \dots, F_d(x_d)\}.$$

Comme il s'agit désormais d'une fonction composée, on applique la dérivation en chaîne.

Pour cela on pose $u_i = F_{X_i}(x_i)$ pour tout $i \in \{1, \dots, d\}$ et l'on obtient

$$f_{(X_1, \dots, X_d)}(x_1, \dots, x_d) = \frac{\partial^d}{\partial u_1 \dots \partial u_d} C(u_1, \dots, u_d) \times \frac{\partial F_{X_1}(x_1)}{\partial x_1} \times \dots \times \frac{\partial F_{X_d}(x_d)}{\partial x_d}.$$

En notant f_{X_i} la densité de la variable aléatoire X_i pour tout $i \in \{1, \dots, d\}$ et en notant c la densité de la copule C , on trouve ainsi

$$f_{(X_1, \dots, X_d)}(x_1, \dots, x_d) = c \{F_{X_1}(x_1), \dots, F_{X_d}(x_d)\} \times f_{X_1}(x_1) \times \dots \times f_{X_d}(x_d). \quad (2.2)$$

2.2.3 Log-vraisemblance d'un modèle écrit avec une copule

On se positionne dans cette section dans un cadre paramétrique. On dénote $\boldsymbol{\theta} \in \Theta$ le vecteur de paramètres qui contient l'ensemble des paramètres des lois marginales de X_1 à X_d et $\boldsymbol{\rho}$ le vecteur des paramètres de dépendance de la copule. L'ensemble des densités et des fonctions de répartition peuvent ainsi s'écrire conditionnellement à $\boldsymbol{\theta}$ et $\boldsymbol{\rho}$. On utilise la densité obtenue à l'équation (2.2) pour écrire la log-vraisemblance du modèle pour une observation $(x_1, \dots, x_d)^\top$, soit

$$\ln L(\boldsymbol{\theta}, \boldsymbol{\rho} \mid (x_1, \dots, x_d)^\top) = \ln c_{U_1, \dots, U_d \mid \boldsymbol{\rho}} \{F_{X_1 \mid \boldsymbol{\theta}}(x_1), \dots, F_{X_d \mid \boldsymbol{\theta}}(x_d)\} + \sum_{i=1}^d \ln f_{X_i \mid \boldsymbol{\theta}}(x_i). \quad (2.3)$$

On constate que l'on a d'une part la log-vraisemblance de la copule et d'autre part la somme des log-vraisemblances des composantes marginales. C'est pour cela, entre autres, que l'on dit que l'utilisation d'une copule permet de séparer la spécification des lois marginales de la spécification de la dépendance. Cependant, on note que les arguments de la fonction c sont définis à l'aide des fonctions $F_{X_1 \mid \boldsymbol{\theta}}, \dots, F_{X_d \mid \boldsymbol{\theta}}$. Ainsi le vecteur de paramètres marginaux $\boldsymbol{\theta}$ a une incidence sur la log-vraisemblance de la copule. Dans le cadre d'une maximisation globale de la log-vraisemblance, l'ajout d'une copule peut avoir une incidence sur l'estimation des paramètres marginaux.

2.2.4 Mesure de la dépendance

Les copules sont invariantes à l'application de transformations croissantes. Ainsi, la copule qui modélise la dépendance entre deux variables aléatoires réelles X et Y est la même que celle qui modélise la dépendance entre $g(X)$ et $h(Y)$, où g et h sont deux transformations croissantes. La mesure de dépendance choisie doit donc elle aussi être invariante à l'application de transformations croissantes.

La corrélation de Pearson, est une mesure couramment employée qui mesure la corrélation affine entre deux variables aléatoires X et Y , soit

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y},$$

où σ_X et σ_Y sont les écarts-types de X et Y . Dans le cas général, on a $\rho(X, Y) \neq \rho\{g(X), h(Y)\}$. Ainsi la corrélation de Pearson n'est pas adaptée à l'utilisation d'une copule car elle n'est pas une mesure invariante aux transformations croissantes.

Deux mesures de dépendance courantes possèdent la propriété souhaitée :

- (i) Le tau de Kendall mesure la concordance entre deux paires indépendantes (X, Y) et $(\tilde{X}, \tilde{Y})$: $\tau = \mathbb{E}[\text{sign}\{(X - \tilde{X})(Y - \tilde{Y})\}]$. Son écriture pour une copule bivariee C est donnée par

$$\tau = -1 + 4 \int_1^1 \int_1^1 C(u_1, u_2) dC(u_1, u_2).$$

- (ii) Le rho de Spearman mesure la corrélation de Pearson entre $F_X(X)$ et $F_Y(Y)$. Son écriture pour une copule C à deux dimensions est donnée par

$$\rho_S = -3 + 12 \int_1^1 \int_1^1 uv dC(u, v).$$

On privilégiera donc ces deux mesures à la corrélation de Pearson.

2.2.5 Copules max-stable

Dans le cadre de ce projet, on souhaite modéliser la dépendance entre les maxima annuels d'intensité de précipitation pour plusieurs durées d'accumulation. Une problématique qui survient alors est que ces maxima ne se produisent généralement pas aux mêmes dates. En effet, le maxima d'intensité de précipitation pour une durée de cinq minutes n'a généralement pas lieu le même jour que le maxima d'intensité associé à une durée de 24 heures. Ainsi la structure de dépendance des maxima annuels ne peut être déduite de la structure de dépendance des observations quotidiennes et elle doit être analysée séparément.

De manière équivalente au cas unidimensionnel des valeurs extrêmes, on pourrait espérer obtenir une fonction de répartition multidimensionnelle pour les maxima annuels d'intensité qui ait la propriété de max-stabilité et qui possède une justification asymptotique.

La max-stabilité est cependant une propriété contraignante qui, mis à part pour des copules simples comme la copule de Gumbel qui sera présentée dans le projet, rend l'estimation des paramètres du modèle compliquée. De plus, la convergence asymptotique vers le modèle de valeurs extrêmes multidimensionnel peut nécessiter des blocs plus grands qu'une année pour être valable.

2.3 Estimation des incertitudes

Dans cette section, on s'intéresse à une variable aléatoire, ou un vecteur aléatoire, $\mathbf{X}$ de densité $f_{\mathbf{X}|\boldsymbol{\theta}}$ paramétrée par un vecteur $\boldsymbol{\theta} \in \Theta$.

2.3.1 Log-vraisemblance et fonction de score

Pour une série d'observations $(\mathbf{x}_i)$ mutuellement indépendantes et identiquement distribuées, la log-vraisemblance associée est donnée par :

$$\ell_n\{\boldsymbol{\theta} \mid (\mathbf{x}_i)\} = \sum_{i=1}^n \ln\{f_{\mathbf{X}|\boldsymbol{\theta}}(\mathbf{x}_i)\}.$$

Pour une seule observation, le score est donné par

$$\Psi(\mathbf{x}, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ln\{f_{\mathbf{X}|\boldsymbol{\theta}}(\mathbf{x})\}.$$

Ainsi, Ψ est donc un vecteur dont la dimension est celle de $\boldsymbol{\theta}$. On note k cette dimension. On cherche alors le maximum de vraisemblance $\hat{\boldsymbol{\theta}}_n$ comme une solution de l'équation de vraisemblance

$$\sum_{i=1}^n \Psi(\mathbf{x}_i, \boldsymbol{\theta}) = \mathbf{0}_k.$$

2.3.2 Matrice d'information de Fisher

La matrice d'information de Fisher est définie par

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} \left\{ \Psi(\mathbf{X}, \boldsymbol{\theta}) \Psi(\mathbf{X}, \boldsymbol{\theta})^\top \right\}.$$

Il s'agit donc d'une matrice carrée de dimension k . En supposant que les dérivées partielles du second ordre puissent être passées sous signe intégral, on montre que la matrice d'information

de Fisher peut également s'écrire

$$\mathcal{I}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}}\{\dot{\Psi}(\mathbf{X}, \boldsymbol{\theta})\}.$$

où $\dot{\Psi}(\mathbf{x}, \cdot)$ est la hessienne de la log-vraisemblance. C'est-à-dire la matrice carrée qui contient les dérivées partielles secondes de la log-vraisemblance par rapport aux paramètres pour une observation $\mathbf{x}$ fixée.

2.3.3 Normalité asymptotique des estimateurs du maximum de vraisemblance

On utilise le théorème de Cramér pour caractériser la normalité asymptotique des estimateurs du maximum de vraisemblance.

Théorème 3 (Cramér) *Soit $\mathbf{X}_1, \dots, \mathbf{X}_n$ des vecteurs aléatoires i.i.d. de densité $f_{\mathbf{X}|\boldsymbol{\theta}}$, où $\boldsymbol{\theta} \in \Theta$. Soit $\boldsymbol{\theta}_0$ la vraie valeur du vecteur de paramètres. Supposons les conditions suivantes vérifiées :*

1. Θ est un sous ensemble ouvert de $\mathbb{R}_k$.
2. Les dérivées partielles de second ordre de f par rapport à $\boldsymbol{\theta}$ existent, sont continues et peuvent être passées sous le signe intégral ;
3. Il existe une fonction K telle que $\mathbb{E}_{\boldsymbol{\theta}}\{K(X)\} < \infty$ et $|\dot{\Psi}(\mathbf{x}, \boldsymbol{\theta})|$ est bornée uniformément par $K(x)$ dans un voisinage de $\boldsymbol{\theta}_0$.
4. $\mathcal{I}(\boldsymbol{\theta}_0)$ est définie positive.
5. Si $f_{\mathbf{X}|\boldsymbol{\theta}} = f_{\mathbf{X}|\boldsymbol{\theta}_0}$ presque partout, alors $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Alors il existe une série de racines $\hat{\boldsymbol{\theta}}_n$ de l'équation de vraisemblance telle que, quand $n \rightarrow \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \rightsquigarrow \mathcal{N}[\mathbf{0}_k, \mathcal{I}^{-1}(\boldsymbol{\theta}_0)].$$

Ce théorème permet d'établir la normalité asymptotique à l'aide de l'information de Fisher pour une série de racines de l'équation de vraisemblance et pas pour l'estimateur du maximum de vraisemblance. Néanmoins, lorsque qu'il existe, pour n fixé, une unique racine de l'équation de vraisemblance, alors cette racine est l'estimateur du maximum de vraisemblance et elle est asymptotiquement normale.

2.3.4 Lien avec l'information de Fisher observée

La matrice d'information de Fisher est définie comme $\mathcal{I}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}}\{\dot{\Psi}(\mathbf{X}, \boldsymbol{\theta})\}$. Comme elle est définie à l'aide d'une espérance, l'information de Fisher s'écrit sous la forme d'une intégrale,

soit

$$\mathcal{I}(\boldsymbol{\theta}) = - \int \dot{\Psi}(\mathbf{x}, \boldsymbol{\theta}) \times f_{\mathbf{X}|\boldsymbol{\theta}}(\mathbf{x}) d\mathbf{x}.$$

Le calcul explicite de l'intégrale étant complexe, voire infaisable, on substitue généralement à la matrice d'information de Fisher la matrice d'information de Fisher observée, qui s'écrit

$$\mathcal{I}_n(\boldsymbol{\theta}) = -\frac{1}{n} \sum_{i=1}^n \dot{\Psi}(\mathbf{x}_i, \boldsymbol{\theta}).$$

L'information observée demande seulement d'être en mesure de calculer les dérivées partielles d'ordre 2 de la log-vraisemblance en fonction des paramètres. En notant toujours $\boldsymbol{\theta}_0$ le vecteur des vrais paramètres, la loi faible des grands nombres appliquée à $\dot{\Psi}(\cdot, \boldsymbol{\theta}_0)$, qui est continue presque sûrement, nous permet de montrer que, lorsque $n \rightarrow \infty$,

$$\mathcal{I}_n(\boldsymbol{\theta}_0) \xrightarrow{\mathbb{P}} \mathcal{I}(\boldsymbol{\theta}_0).$$

Or, on peut montrer que $\hat{\boldsymbol{\theta}}_n \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_0$ et que $\mathcal{I}_n(\hat{\boldsymbol{\theta}}_n) \xrightarrow{\mathbb{P}} \mathcal{I}_n(\boldsymbol{\theta}_0)$. Cette dernière convergence en probabilité requiert l'équicontinuité de la famille de fonctions $\{I_n : n \in \mathbb{N}\}$. On peut finalement montrer que, quand $n \rightarrow \infty$,

$$\mathcal{I}_n(\hat{\boldsymbol{\theta}}_n) \xrightarrow{\mathbb{P}} \mathcal{I}(\boldsymbol{\theta}_0).$$

On peut alors substituer $\mathcal{I}_n(\hat{\boldsymbol{\theta}}_n)$ à $\mathcal{I}(\boldsymbol{\theta}_0)$ dans la formule du théorème de Cramér. On note cependant que cette substitution repose sur de multiples convergences en probabilité et que le régime de normalité asymptotique peut être long à atteindre. Il est donc toujours important de s'interroger sur la validité des méthodes asymptotiques avant de les utiliser.

2.3.5 Méthode Delta

Un autre théorème dû à Cramér nous permet ensuite de caractériser la distribution d'une fonction des paramètres $\boldsymbol{\theta}$.

Théorème 4 (*Méthode Delta*) Soit $g : \mathbb{R}^k \rightarrow \mathbb{R}^l$ telle que $\dot{g}$ est continue dans un voisinage de $\boldsymbol{\theta}_0$. Alors si $\mathbf{X}_n$ est une série de vecteur aléatoire telle que $\sqrt{n}(\mathbf{X}_n - \boldsymbol{\theta}_0) \rightsquigarrow \mathbf{X}$, alors

$$\sqrt{n}\{g(\mathbf{X}_n) - g(\boldsymbol{\theta}_0)\} \rightsquigarrow \dot{g}(\boldsymbol{\theta}_0)\mathbf{X}.$$

En combinant le théorème 4 et les propriétés que nous avons obtenues avec le théorème 3 et la matrice de Fisher observée, on a finalement que, lorsque $n \rightarrow \infty$,

$$\sqrt{n} \{g(\hat{\theta}_n) - g(\theta_0)\} \rightsquigarrow \mathcal{N} \left[\mathbf{0}_k, \dot{g}(\theta_0) \mathcal{I}^{-1}(\theta_0) \dot{g}(\theta_0)^\top \right].$$

Cette propriété s'avérera utile pour obtenir la distribution asymptotique de quantités d'intérêt s'écrivant comme une fonction des paramètres du modèle. Par exemple, on peut s'en servir pour estimer les intervalles de confiance pour les quantiles de la loi généralisée des valeurs extrêmes.

2.4 Dérivation automatique

Les méthodes d'estimation des incertitudes qui ont été présentées à la section 2.3 nécessitent de pouvoir calculer les dérivées partielles d'ordre 2 de la log-vraisemblance pour obtenir la matrice d'information de Fisher observée, ou encore les dérivées partielles d'ordre 1 des fonctions d'intérêt pour pouvoir appliquer la méthode delta. Plusieurs possibilités s'offrent alors à l'utilisateur. On peut entre autres calculer explicitement l'ensemble des dérivées à la main ou à l'aide d'un logiciel de dérivation symbolique. Cependant, cette stratégie peut vite devenir complexe lorsque les fonctions que l'on souhaite dériver sont des fonctions compliquées constituées de plusieurs rangs de fonctions composées. Par ailleurs, elle présente aussi un manque de flexibilité car le moindre changement dans le modèle nécessite de recalculer l'ensemble des dérivées. Une autre possibilité est d'utiliser une approximation des dérivées à l'aide des formules de différences finies. Toutefois, cette méthode peut être instable numériquement.

La stratégie qui a été retenue dans ce projet a été d'utiliser la dérivation automatique. Il s'agit de dériver exactement l'ensemble des quantités d'intérêt. Il est important de noter que le langage de programmation `Julia` permet de prendre en charge presque nativement la dérivation automatique. C'est-à-dire que moyennant quelques ajustements, il est toujours possible d'estimer les dérivées d'une fonction codée en `Julia`, y compris si celle-ci fait appel à de nombreuses fonctions composées. Cette propriété du langage informatique a été un atout dans le cadre de cette maîtrise. On présente donc dans cette section quelques notions qui permettent de mieux appréhender la dérivation automatique.

2.4.1 Objets duaux

La dérivation automatique repose sur l'utilisation de nombres duaux. Ceux-ci sont construits à partir de l'algèbre des nombres réels. On adjoint à un réel x un autre nombre réel x' .

Lorsqu'une fonction sera appliquée au réel x , une opération correspondant à la dérivation sera appliquée à x' . Le nombre dual formé s'écrit alors

$$x + \epsilon x'.$$

Ici, ϵ est une quantité abstraite non réelle. On pourrait la comparer au nombre imaginaire i des nombres complexes. Deux propriétés de cette quantité sont : $\epsilon \neq 0$ et $\epsilon^2 = 0$. La nécessité d'imposer la seconde propriété sera expliquée dans la partie 2.4.2. On peut également utiliser la notation plus concise $\langle x, x' \rangle$.

2.4.2 Surcharge des opérateurs mathématiques

On sait que la dérivée de la somme de deux fonctions est égale à la somme des dérivées de chaque fonction. On peut appliquer ce principe pour débiter la définition d'une arithmétique des nombres duaux. On a donc

$$(x + \epsilon x') + (y + \epsilon y') = (x + y) + \epsilon(x' + y').$$

L'opérateur de sommation des nombres duaux peut alors être noté $+_\epsilon$. En utilisant la notation concise, on a alors $\langle x, x' \rangle +_\epsilon \langle y, y' \rangle = \langle x + y, x' + y' \rangle$. On peut aisément appliquer ce même principe à la soustraction. On peut désormais s'intéresser à la multiplication. On a

$$(x + \epsilon x') \times (y + \epsilon y') = xy + (x'y + xy')\epsilon + x'y'\epsilon^2.$$

Or on sait que la dérivée d'une fonction dérivable $f \times g$ est $f'g + g'f$. Cela explique donc pourquoi il est nécessaire que $\epsilon^2 = 0$. On peut donc définir l'opérateur de multiplication des nombres duaux $\times_\epsilon$ comme étant l'opérateur : $\langle x, x' \rangle \times_\epsilon \langle y, y' \rangle = \langle xy, xy' + yx' \rangle$. On peut enfin appliquer ce principe à la division en multipliant de part et d'autre par le conjugué du dénominateur :

$$\frac{x + \epsilon x'}{y + \epsilon y'} = \frac{(x + \epsilon x') \times (y - \epsilon y')}{(y + \epsilon y') \times (y - \epsilon y')},$$

soit

$$\frac{x + \epsilon x'}{y + \epsilon y'} = \frac{xy + \epsilon(x'y - y'x)}{y^2},$$

soit enfin

$$\frac{x + \epsilon x'}{y + \epsilon y'} = \frac{x}{y} + \epsilon \frac{x'y - y'x}{y^2}.$$

On peut donc également définir l'opérateur $\div_{\epsilon}$ comme étant l'opérateur qui effectue l'opération

$$\langle x, x' \rangle \div_{\epsilon} \langle y, y' \rangle = \left\langle \frac{x}{y}, \frac{xy' - yx'}{y^2} \right\rangle.$$

En plus des quatre opérations de base, il est possible de surcharger les opérateurs pour les nombres duaux à partir des dérivées usuelles. Par exemple, pour la fonction logarithme on peut définir l'opérateur dual associé : $\ln_{\epsilon} \langle x, x' \rangle = \langle \ln x, x'/x \rangle$ pour $x > 0$. La majorité des opérateurs arithmétique réels dont on connaît la dérivée peuvent ainsi être surchargés pour définir l'arithmétique dual. Dans le cas où l'opérateur surchargé correspondant à une fonction ne peut pas être défini, il est toujours possible de dériver le développement de Taylor à n'importe quel ordre de la fonction correspondante pour obtenir la dérivée avec la précision souhaitée. Il est donc parfois nécessaire de définir à la main le développement de Taylor d'une fonction si on souhaite la dériver automatiquement.

2.4.3 Règle de dérivation en chaîne

Lorsque la fonction que l'on souhaite dériver est composée, on doit appliquer la règle de la chaîne. Supposons que la fonction d'intérêt est $h = f \circ g \circ z$, où les fonctions f , g et z sont des fonctions de $\mathbb{R}$ dans $\mathbb{R}$. On note :

$$\begin{aligned}\omega_0 &= x; \\ \omega_1 &= z(\omega_0); \\ \omega_2 &= g(\omega_1); \\ y &= \omega_3 = f(\omega_2).\end{aligned}$$

Alors la règle de dérivation en chaîne donne

$$\frac{dy}{dx}(x) = \frac{d\omega_3}{d\omega_2} \frac{d\omega_2}{d\omega_1} \frac{d\omega_1}{d\omega_0}(x).$$

On peut donc, pour un nombre arbitraire de compositions de fonctions, définir la dérivée de la quantité ω_i par rapport à x récursivement à l'aide de la relation

$$\frac{d\omega_i}{dx} = \frac{d\omega_i}{d\omega_{i-1}} \frac{d\omega_{i-1}}{dx}.$$

On appelle cette formule la formule de propagation avant ou mode avant. Il est également

possible de définir une formule de propagation arrière sous la forme

$$\frac{dy}{d\omega_i} = \frac{dy}{d\omega_{i+1}} \frac{d\omega_{i+1}}{d\omega_i}.$$

2.4.4 Lien entre les nombres duaux et la dérivation en chaîne

On peut désormais se demander comment appliquer la règle de dérivation en chaîne à un nombre dual. La solution à ce problème peut être appréhendée par l'utilisation d'un développement de Taylor. Soit une fonction $f : \mathbb{R} \rightarrow \mathbb{R}$. On applique le développement de Taylor autour de x , soit

$$f(x + \epsilon x') = f(x) + f'(x) \times (x + \epsilon x' - x) + f''(x) \times \frac{(x + \epsilon x' - x)^2}{2} + \mathcal{O}\{(\epsilon x')^3\}.$$

Or comme on a $\epsilon^2 = 0$, on a aussi $\epsilon^k = 0$ pour tout entier $k \geq 2$. Ainsi on a seulement $f(x + \epsilon x') = f(x) + \epsilon f'(x) \times x'$.

On peut désormais composer la fonction f par une fonction g en appliquant le même principe et en appliquant le développement de Taylor autour de $f(x)$, viz.

$$g \circ f(x + \epsilon x') = g \{f(x) + \epsilon f'(x) \times x'\} = g \circ f(x) + g' \{f(x)\} \times \{f(x) + \epsilon f'(x) \times x' - f(x)\}.$$

On trouve alors

$$g \circ f(x + \epsilon x') = g \{f(x) + \epsilon f'(x) \times x'\} = g \circ f(x) + \epsilon g' \{f(x)\} \times f'(x) x'.$$

En appliquant itérativement ce principe, on retrouve le principe de la dérivation en chaîne. Lorsque x est l'inconnue au rang le plus bas de la hiérarchie de composition de fonction, on pose $x' = 1$.

2.4.5 Dimensions supérieures

Le principe de dérivation automatique peut être généralisé à des fonctions multivariées $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$ en appliquant le même principe à des vecteurs duaux. La dérivée première est alors remplacée par la jacobienne de f . Des règles de composition pour appliquer la dérivation en chaîne sont aussi définies dans le cas multidimensionnel. Un cas particulier intéressant est celui des fonctions de $\mathbb{R}^m$ dans $\mathbb{R}$ pour lesquelles la jacobienne est un gradient. On peut alors augmenter le concept de nombres duaux pour inclure des termes correspondant aux dérivées partielles d'ordre 2 pour obtenir la hessienne d'une fonction de $\mathbb{R}^m$ dans $\mathbb{R}$. Cela

sera notamment utile lorsqu'on cherchera à calculer l'information de Fisher observée.

2.4.6 Pratique

Les nombres duaux, qui ont été inventés en 1973 par William Clifford, sont revenus sur le devant de la scène avec l'avènement de l'apprentissage machine. En effet, lorsqu'on leur applique les règles de dérivation en chaîne, ces nombres permettent d'obtenir aisément et rapidement les gradients de presque n'importe quelle fonction continue et dérivable.

En pratique, on utilise rarement explicitement les nombres duaux pour dériver automatiquement les fonctions. Ils sont encapsulés dans des structures appelées tenseurs. De plus, la dérivation en chaîne est gérée automatiquement par le logiciel utilisé.

CHAPITRE 3 REVUE DE LITTÉRATURE

On présente dans cette revue de littérature plusieurs méthodes d'estimation pour les courbes Intensité-Durée-Fréquence (IDF). On présentera d'abord la méthode actuellement employée par Environnement et Changement Climatique Canada (ECCC) et les améliorations qui peuvent y être apportées grâce à l'usage de la loi Généralisée de Valeurs Extrêmes (GEV). Ensuite, on présentera les modélisations s'appuyant sur une structure fonctionnelle pour les paramètres GEV et les avantages que ces modélisations peuvent apporter. Enfin, on s'intéressera aux propositions qui ont été faites pour tenir compte de la dépendance entre les séries de données.

3.1 Méthodologie actuelle d'estimation des courbes IDF

La méthode utilisée par ECCC consiste à estimer les niveaux de retour associés à des périodes de retour allant de 2 à 100 ans indépendamment pour chaque durée d'accumulation à l'aide de la loi de Gumbel et de la méthode des moments appliquées aux maxima annuels d'intensité de précipitation (Hogg, 1989). L'usage de la loi de Gumbel pour l'estimation des précipitations extrêmes a cependant été remis en question (Koutsoyiannis, 2003) car les queues de distribution des maxima d'intensité de précipitation, en particulier pour des durées courtes, semblent plus lourdes que celles que la loi de Gumbel peut modéliser. La loi de Gumbel pourrait ainsi biaiser les estimations des niveaux de retour. Par ailleurs, même dans le cas d'une distribution sous-jacente qui serait dans le domaine d'attraction de la loi de Gumbel, il est préférable d'utiliser une loi GEV afin d'éviter de sous-estimer les incertitudes inhérentes à l'estimation d'un modèle de valeurs extrêmes (Coles, 2001). La loi de Gumbel étant un sous-cas de la loi GEV, privilégier la loi GEV n'entraîne pas de biais. Cependant, l'estimation du paramètre de forme pose une difficulté supplémentaire et une quantité importante de données peut être nécessaire pour l'estimer adéquatement.

Jalbert *et al.* (2022) ont proposé de modéliser les précipitations extrêmes à l'aide de lois GEV estimées indépendamment pour chaque durée d'accumulation. Leur modèle incluait des covariables spatiales et s'appuyait sur un modèle bayésien hiérarchique permettant le partage d'information entre les stations du réseau météorologique de l'est du Canada. L'objectif était de pouvoir interpoler spatialement les courbes IDF dans l'est du Canada pour compenser le faible nombre de stations et leur répartition spatiale inégale. D'autres auteurs avaient déjà expérimenté l'utilisation de la loi GEV pour la modélisation des maxima journaliers de précipitation et avaient mis en avant les avantages de la loi GEV par rapport à la loi de

Gumbel (Papalexiou et Koutsoyiannis, 2013).

3.2 Modèles d'échelle pour l'estimation des courbes IDF

Une stratégie de valorisation des données s'appuie sur l'utilisation d'une relation d'échelle pour les paramètres des lois GEV pour les maxima d'intensité de précipitation. De tels modèles ont, par exemple, été proposés par Koutsoyiannis *et al.* (1998) ou encore Blanchet *et al.* (2016). Les modèles d'échelle sont basés sur la décroissance en loi de puissance des niveaux de retour avec la durée qui est observée dans les courbes IDF et qui apparaît dans le tracé des droites dans la figure 1.1. Un unique modèle est ainsi estimé conjointement pour l'ensemble des durées d'accumulation, ce qui a pour effet de multiplier le nombre d'observations utilisées sans multiplier le nombre de paramètres. Une telle optimisation des données est particulièrement souhaitable dans le contexte d'un faible nombre d'observations (Koutsoyiannis *et al.*, 1998).

Le modèle proposé par Koutsoyiannis *et al.* (1998) est appelé le modèle d-GEV. Celui-ci sera étudié tout au long de ce mémoire. Ce modèle s'appuie sur des lois marginales GEV dont les paramètres de localisation μ_d et d'échelle σ_d sont fonction de la durée d'accumulation d considérée. On considère le paramètre de forme ξ commun à l'ensemble des durées. Il s'agit d'une hypothèse courante dans la littérature ; par ailleurs le paramètre ξ est souvent difficile à identifier, de sorte que combiner l'ensemble des données pour l'estimer permet d'éviter ces difficultés.

Pour aboutir au modèle d-GEV, les auteurs sont partis de l'hypothèse que le niveau de retour pour la période de retour T et la durée d s'écrit comme le ratio de deux fonctions séparables en T et d , à savoir

$$Q_T(d) = a(T)/b(d).$$

La fonction $a(T)$ est positive. Comme on s'appuie sur une loi de puissance, la fonction b au dénominateur peut s'écrire

$$b(d) = (d + \theta)^\eta,$$

où

$\eta \in [0, 1]$ représente l'exposant de la loi de puissance ;

$\theta \in \mathbb{R}^+$ est un paramètre permettant de rajouter de la concavité à la relation en loi de puissance.

Le paramètre θ offre un degré de liberté supplémentaire à la loi de puissance. Son signe positif permet d'ajouter de la concavité aux courbes IDF. L'effet de θ est principalement visible pour les durées courtes et l'on peut constater qualitativement sur la figure 1.1 que les niveaux de retour empiriques présentent effectivement une légère concavité. L'intervalle de définition de η permet de s'assurer la décroissance des niveaux de retour avec d et le respect de la condition de non-croisement des courbes IDF. C'est-à-dire que pour $T > T'$ et pour tout entier d , on doit avoir $Q_T(d) > Q_{T'}(d)$.

Pour aboutir à une forme fonctionnelle pour σ_d et μ_d , on utilise le fait que $a(T) = Q_T(d) \times b(d)$. Ainsi $Q_T(d) \times b(d)$ est une fonction indépendante de d . Or

$$Q_T(d) \times b(d) = \left[\mu_d + \frac{\sigma_d}{\xi} \times \left[-1 + \{-\ln(1 - 1/T)\}^{-\xi} \right] \right] \times b(d),$$

soit

$$Q_T(d) \times b(d) = \mu_d \times (d + \theta)^\eta + \frac{\sigma_d \times (d + \theta)^\eta}{\xi} \times \left[-1 + \{-\ln(1 - 1/T)\}^{-\xi} \right].$$

L'expression précédente est indépendante de d si l'on pose $\mu_d = \mu_0 \times (d + \theta)^{-\eta}$ et $\sigma_d = \sigma_0 \times (d + \theta)^{-\eta}$. Pour simplifier l'écriture de la fonction de répartition GEV, on pose ensuite $\mu_d = \tilde{\mu} \times \sigma_d$ où $\tilde{\mu} = \mu_0/\sigma_0$. On a $\tilde{\mu} \in \mathbb{R}$ et $\sigma_0 > 0$. On pose $\sigma_0 = \exp(\phi_0)$, où $\phi_0 \in \mathbb{R}$ pour contraindre le moins possible le modèle pour l'optimisation.

La fonction de répartition de la loi GEV pour une durée d s'écrit alors

$$G(x_d \mid \tilde{\mu}, \phi_0, \xi, \theta, \eta) = \begin{cases} \exp \left[- \left\{ 1 + \xi \left(\frac{x_d}{e^{\phi_0} \times (d + \theta)^{-\eta}} - \tilde{\mu} \right) \right\}^{-1/\xi} \right] & \text{si } \xi \neq 0, \\ \exp \left[- \exp \left(- \frac{x_d}{e^{\phi_0} \times (d + \theta)^{-\eta}} - \tilde{\mu} \right) \right] & \text{si } \xi = 0. \end{cases}$$

L'aptitude de la modélisation d-GEV à modéliser convenablement les données a été testée par Ulrich *et al.* (2020). Dans cet article, des covariables spatiales ont été utilisées dans les paramètres de la d-GEV pour modéliser les courbes IDF dans une région de l'ouest de l'Allemagne. Les auteurs ont noté des améliorations dans l'estimation des courbes IDF grâce à l'usage de la loi d-GEV mais ont aussi souligné des limitations concernant les durées inférieures à 30 minutes. À notre connaissance, la loi d-GEV n'a jamais été utilisée pour estimer les courbes IDF au Canada.

3.3 Modélisation de la structure de dépendance

Le modèle d-GEV semble encourageant pour l'estimation des courbes IDF. La log-vraisemblance utilisée pour estimer les paramètres d-GEV s'écrit

$$\ln\{L(\boldsymbol{\theta} \mid \mathbf{x})\} = \sum_{i=1}^n \sum_{d \in D} \ln \{g(x_{i,d} \mid d, \boldsymbol{\theta})\}, \quad (3.1)$$

où

$\boldsymbol{\theta} = (\tilde{\mu}, \phi_0, \xi, \theta, \eta)^\top$ est le vecteur des paramètres d-GEV ;

$g(\cdot \mid d, \boldsymbol{\theta})$ est la densité de la loi marginale GEV du modèle d-GEV pour la durée d ;

n est le nombre d'années d'observations ;

D dénote l'ensemble des durées d'accumulation ;

$\mathbf{x} = (x_{i,d}) : 1 < i < n, d \in D$ dénote l'ensemble des observations d'intensité de précipitation.

Le choix de cette log-vraisemblance suppose que les observations sont indépendantes d'une durée et d'une année à l'autre. Lorsque l'on considère des observations mesurées lors d'années différentes, cette hypothèse n'est pas contraignante. Cependant pour des observations obtenues la même année pour différentes durées, on ne peut pas a priori supposer quelles soient indépendantes. Elles ne le sont d'ailleurs pas pour des raisons physiques. En effet, on sait que pour $d > d'$ on a nécessairement $x_d \leq x_{d'}$ car le maximum d'intensité pour une durée plus courte est toujours plus grand ou égal à celui d'une durée plus longue. Par ailleurs, pour des années proches, les maxima d'intensité annuels peuvent avoir été mesurés lors du même événement climatique.

Jurado *et al.* (2020) ont modélisé la dépendance entre les maxima d'intensité de précipitation de différentes durées à l'aide d'un processus max-stable de Brown–Resnick. Cependant, ils n'ont pas estimé les incertitudes associées à leur modèle. La paramétrisation de la dépendance choisie par Jurado *et al.* (2020) est inspirée de l'article de Van de Vyver et Van den Bergh (2018), qui proposent de modéliser la dépendance entre les précipitations accumulées à l'aide de la fonction de covariance de Matérn en utilisant la distance logarithmique entre les durées d'accumulation. Cependant, ces auteurs utilisent ce modèle pour mesurer la dépendance entre les précipitations accumulées sur des durées variant de 1 mois à 1 an. Ce n'est donc pas une étude sur les maxima de précipitation ou d'intensité de précipitation et les échelles de temps considérées sont très différentes de celles que nous considérons dans ce projet. Cependant, on pourra essayer d'adapter les méthodes proposées par Van de Vyver et Van den Bergh (2018) à notre cas d'étude.

À notre connaissance, la modélisation d-GEV n'a jamais été associée à une modélisation de la dépendance dans les données de maxima d'intensité de précipitation. La dépendance n'a pas d'incidence sur l'estimation ponctuelle des niveaux de retour. Cependant, elle affecte l'estimation des incertitudes les entourant. En effet, des données corrélées ont moins de pouvoir explicatif que des données indépendantes. Il existe donc un risque de sous-estimer ces incertitudes si l'on ne considère pas la corrélation présente dans les données.

CHAPITRE 4 APPLICATION INITIALE DU MODÈLE D-GEV ET EXPLORATION DE LA CORRÉLATION

4.1 Présentation des données

4.1.1 Origines des données

Les données utilisées dans cette maîtrise sont fournies par Environnement et Changement Climatique Canada (ECCC). Les stations gérées par ECCC sont assujetties aux normes imposées par l'Organisation Mondiale de Météorologie (OMM), ce qui suggère un degré de confiance élevé pour l'application des méthodes de valeurs extrêmes utilisées. Le réseau de stations de l'est du Canada contient 336 stations météorologiques avec, en moyenne, 25 années de données.

Le jeu de données des maxima de précipitation annuels mesurés à l'aéroport Pierre-Eliot-Trudeau à Montréal constituera le fil conducteur de cette maîtrise pour l'application et la validation des modèles proposés. Les données enregistrées correspondent aux maxima annuels de précipitation pour neuf durées allant de 5 minutes à 24 heures. Elles couvrent la période de 1943 à 2021. Les années 1994, 1995 et 2003 manquent. Il y a donc en tout 76 années de données. Les maxima pour les différentes durées d'accumulation ont été reconstruits à partir de la série temporelle de précipitation mesurée par intervalles de 5 minutes. Le choix de ce jeu de données est particulièrement adapté à notre problématique de par sa longueur et par l'absence de valeurs manquantes. Cette quantité importante de données permettra ainsi de faire émerger des aspects intéressants du problème.

Afin d'appliquer les méthodes de la théorie des valeurs extrêmes, il est nécessaire de vérifier la stationnarité des séries de maxima annuels. Jalbert *et al.* (2022) ont utilisé un test de Mann–Kendall au niveau nominal 5% qui n'a pas permis de rejeter l'hypothèse nulle de stationnarité. Nous supposons donc que les séries temporelles des maxima sont stationnaires.

4.1.2 Transformation des données

La méthodologie d-GEV a pour vocation d'être utilisée avec les maxima annuels d'intensité de précipitation. Il est donc nécessaire de transformer les données d'accumulation en intensités. Pour cela on applique la formule

$$x_{i,d} = Q_i(d)/d,$$

où

$x_{i,d}$ est l'intensité maximale de précipitation associé à la durée d pour l'année i ;

$Q_i(d)$ est le maximum de précipitation associé à la durée d pour l'année i ;

d est la durée d'accumulation considérée, exprimée en heures.

4.2 Application du modèle d-GEV aux données de l'aéroport Trudeau

4.2.1 Méthodologie

Nous allons appliquer le modèle d-GEV aux données de l'aéroport Trudeau. Pour cela, on utilise une méthodologie fréquentiste de maximisation de la vraisemblance. La méthode de maximisation de la vraisemblance va nous permettre d'estimer les incertitudes entourant les estimateurs à l'aide des propriétés de normalité asymptotique. La log-vraisemblance du modèle est donnée dans l'équation (3.1). Comme pour le modèle GEV, il n'existe pas de solution analytique pour estimer le maximum de vraisemblance de la d-GEV. On utilise donc des méthodes d'optimisation numériques. Ces méthodes sont déjà implantées dans le langage de programmation `Julia`.

Une fois l'optimisation effectuée, une première méthode de validation visuelle du modèle peut être employée à l'aide de diagrammes quantiles-quantiles ("QQ plots" ou droites de Henri). Pour cela on affiche les points de coordonnées $(\hat{F}_d^{-1}\{i/(n+1)\}, \tilde{x}_{i,d})$, où $\hat{F}_d^{-1}$ est l'estimation de la fonction de quantile associée à la durée d . Elle est donnée par

$$\hat{F}_d^{-1}(p) = \begin{cases} \hat{\mu}_d - \hat{\sigma}_d \times \ln\{-\ln(p)\} & \text{si } \hat{\xi} = 0, \\ \hat{\mu}_d + \frac{\hat{\sigma}_d}{\hat{\xi}} \times [\{-\ln(p)\}^{-\hat{\xi}} - 1] & \text{si } \hat{\xi} \neq 0, \end{cases} \quad (4.1)$$

avec $\hat{\sigma}_d = e^{\hat{\phi}_0} \times (d + \hat{\theta})^{-\hat{\eta}}$ et $\hat{\mu}_d = \hat{\mu} \times \hat{\sigma}_d$. La série $(\tilde{x}_{i,d})$ correspond aux observations du maxima annuel d'intensité de précipitation pour la durée d triées par ordre croissant.

L'abscisse de chaque point représente le quantile théorique associé à la probabilité $p_i = i/(n+1)$. L'ordonnée est le quantile observé. En cas d'adéquation parfaite entre le modèle et les données, c'est-à-dire lorsque les quantiles réels et théoriques sont égaux, les points sont alignés le long d'une droite de pente égale à 1. L'utilisation de la méthode Delta permet par ailleurs d'obtenir des intervalles de confiance pour vérifier l'adéquation du modèle.

4.2.2 Résultats

On présente sur la figure 4.1 les QQ-plots pour plusieurs durées d'accumulation. Ceux sur la colonne de gauche ont été obtenus avec la méthode d-GEV. Ceux sur la colonne de droite ont été obtenus à l'aide de lois GEV estimées indépendamment pour chacune des durées. Les QQ-plots sont assortis des intervalles de confiance au niveau de confiance nominal 95%. Ces intervalles de confiance ont été estimés à l'aide de la méthode Delta. La matrice d'information de Fisher observée et le gradient de la fonction quantile sont estimés par dérivation automatique. Les graphes pour une durée donnée sont mis à la même échelle pour les deux modèles afin de pouvoir les comparer plus aisément.

Qualitativement, on constate que le modèle d-GEV offre un niveau d'adéquation aux données acceptable. L'ajustement est comparable pour les deux modèles et la disposition des points autour de la droite de pente 1 est proche d'un modèle à l'autre. On constate que l'ajustement pour la durée de 30 minutes est assez mauvais dans les deux cas, mais le saut d'environ 10 mm/h constaté dans les données peut difficilement être modélisé par une distribution classique. On remarque, sur le graphe associé à la durée d'accumulation $d = 6$ heures, que le modèle d-GEV semble être équivalent au modèle GEV avec un léger décalage. Il est possible que le fait d'imposer la forme fonctionnelle des paramètres de localisation et d'échelle biaise les résultats. Il est difficile d'évaluer l'effet du paramètre ξ commun à toutes les durées pour le modèle d-GEV car il n'y a pas de différences notables entre les deux modèles pour l'ajustement de la queue de distribution.

Ce qui distingue davantage les deux modèles sont les intervalles de confiance. En effet, pour le modèle d-GEV, on utilise les données pour toutes les durées afin d'estimer les cinq paramètres du modèle d-GEV. En revanche, pour l'estimation des lois GEV indépendantes, seules les données de la durée considérée permettent d'estimer les trois paramètres GEV. Étant donné qu'il y a neuf durées pour l'aéroport Trudeau, les niveaux de confiance obtenus pour l'aéroport Trudeau sont beaucoup plus étroits pour le modèle d-GEV. On constate que les intervalles de confiance au niveau 95% du modèle d-GEV semblent contenir une proportion significativement inférieure à 95% des données observées. Il existe donc un risque de sous-estimer l'incertitude réelle pour les niveaux de retour en utilisant le modèle d-GEV.

4.2.3 Conclusion sur l'application de la méthodologie d-GEV simple

Les résultats de cette première application de la méthodologie d-GEV nous permettent de constater que la d-GEV semble être une approche prometteuse car l'ajustement marginal

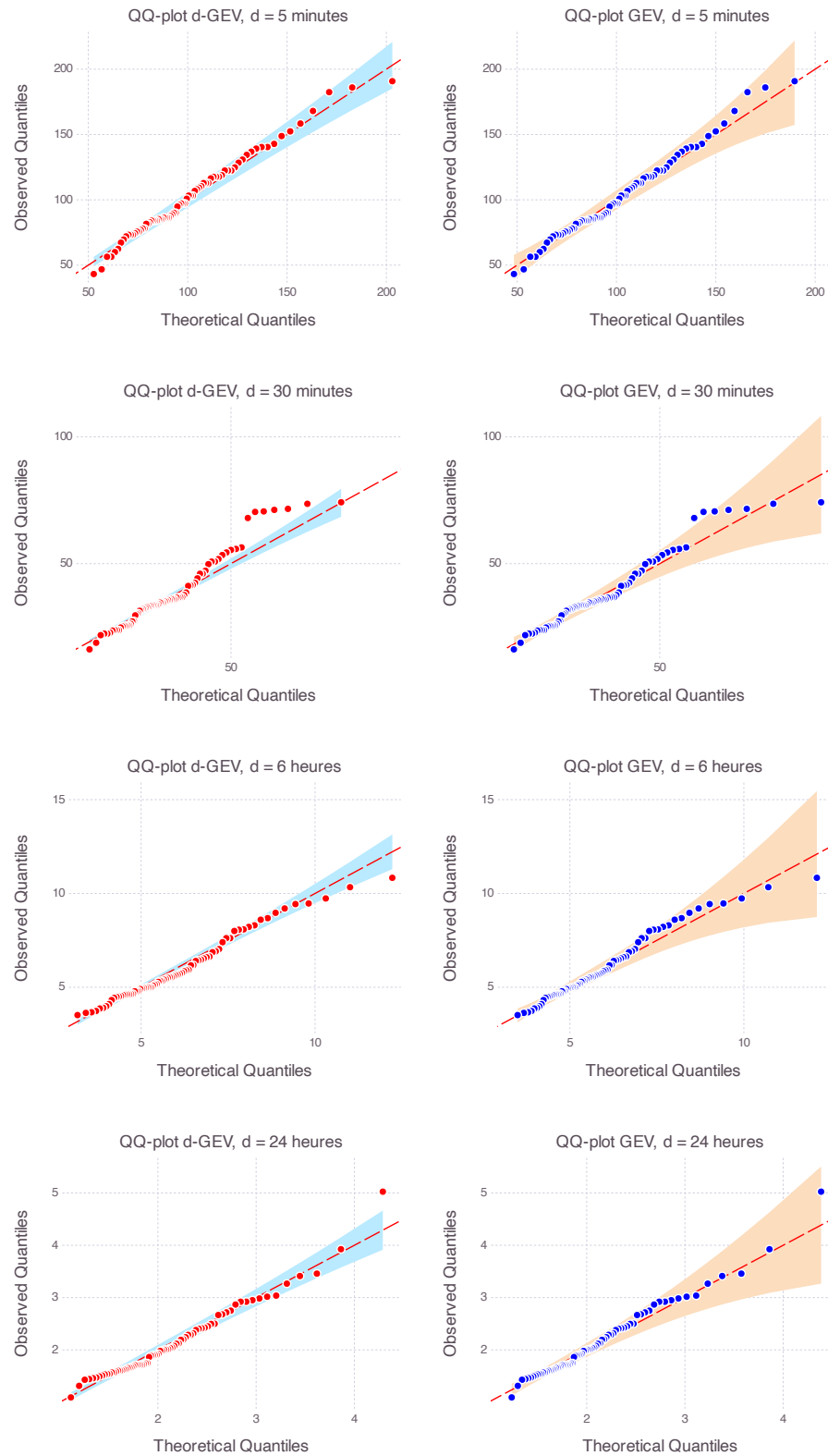


FIGURE 4.1 Diagrammes quantile-quantile obtenus pour chacune des méthodes pour 4 durées

de celle-ci est bon et semble très similaire à celui obtenu en effectuant des estimations indépendantes pour chaque série de maxima. L'estimation est par ailleurs aisée et permet de regrouper l'ensemble des séries en un seul modèle.

Cependant on peut aussi mettre en avant une limite importante au modèle d-GEV. Le modèle sous-estime les incertitudes lorsqu'on utilise la méthode Delta pour estimer des intervalles de confiance associés aux quantiles. Étant donné l'importance d'avoir une estimation convenable des incertitudes pour les précipitation extrêmes, il est nécessaire de traiter cette problématique.

4.3 Exploration de la corrélation dans les données

Les maxima annuels de précipitation pour différentes durées d'accumulation sont corrélés. La figure 4.2, pour la durée de référence $d = 5$ minutes, permet de constater que la série des maxima d'intensité de précipitation associée à la durée de 5 minutes est corrélée aux autres séries de maxima d'intensité. La force de cette corrélation décroît au fur et à mesure que l'on considère des durées lointaines de 5 minutes. La corrélation n'est presque plus perceptible lorsqu'on considère la durée de 12h. Sur la figure 4.3, on constate aussi un motif visible de corrélation lorsque l'on considère les maxima associés à la durée 1h et ceux d'autres durées.

Deux remarques qualitatives supplémentaires peuvent être faites. Premièrement le degré de corrélation entre les séries de maxima de 1 heure et 30 minutes semble être d'un degré comparable à celle pour les maxima de 1h et 2h. Deuxièmement, le niveau de corrélation visible dans ces deux graphes ne semble pas plus faible que celui observé lorsque l'on considère le graphe 5 minutes vs 10 minutes. Ces deux observations semblent indiquer que le niveau de dépendance ne dépend pas de la durée absolue qui sépare les deux durées d'accumulation, qui est de seulement 5 minutes dans un cas et jusqu'à une heure dans l'autre.

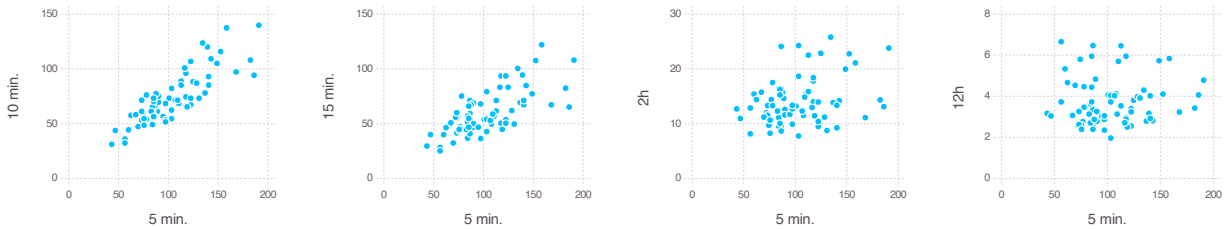


FIGURE 4.2 Diagrammes de dispersion pour la série des maxima de 5 minutes contre une sélection de durées, Aéroport Trudeau

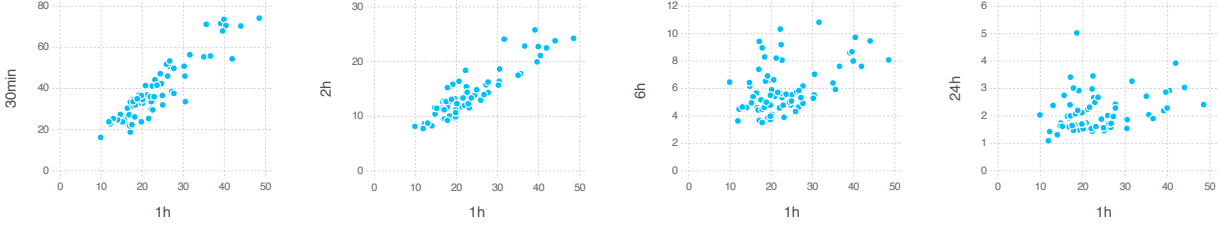


FIGURE 4.3 Diagrammes de dispersion pour la série des maxima de 1 heure contre une sélection de durées, Aéroport Trudeau

La figure 4.4a montre la corrélation empirique, mesurée par le tau de Kendall, entre les séries de maxima d'intensité pour l'ensemble des durées d'accumulation. On constate, à nouveau, que le degré de corrélation ne dépend pas de la différence absolue entre les durées d'accumulation. En effet, la valeur de la corrélation entre la série de 12h et la série de 24h est semblable à celle qui sépare les séries pour 5 minutes et 10 minutes.

À partir de ces observations, on peut penser que le ratio entre les durées d'intérêt peut être une métrique utile pour caractériser l'évolution de la dépendance. On peut par exemple suivre la proposition faite par Van de Vyver et Van den Bergh (2018) et proposer la mesure de distance logarithmique entre deux durées $d > 0$ et $d' > 0$, à savoir

$$h(d, d') = |\ln(d/d')|,$$

soit encore

$$h(d, d') = |\ln(d) - \ln(d')| = |\ln(d') - \ln(d)|. \quad (4.2)$$

Le graphe de la figure 4.4b montre la corrélation telle que mesurée par le tau de Kendall, $\tau(d, d')$, en fonction de $h(d, d')$ pour les données de l'aéroport Trudeau. La fonction $h(\cdot, \cdot)$ permet naturellement de définir une distance entre les durées. On a en particulier que $h(d, d) = 0$ et que h est symétrique. Le choix d'utiliser tau de Kendall comme mesure de corrélation, plutôt que le rho de Pearson, sera justifié ultérieurement. Il se dégage de ce graphe un motif de décroissance de la corrélation avec h .

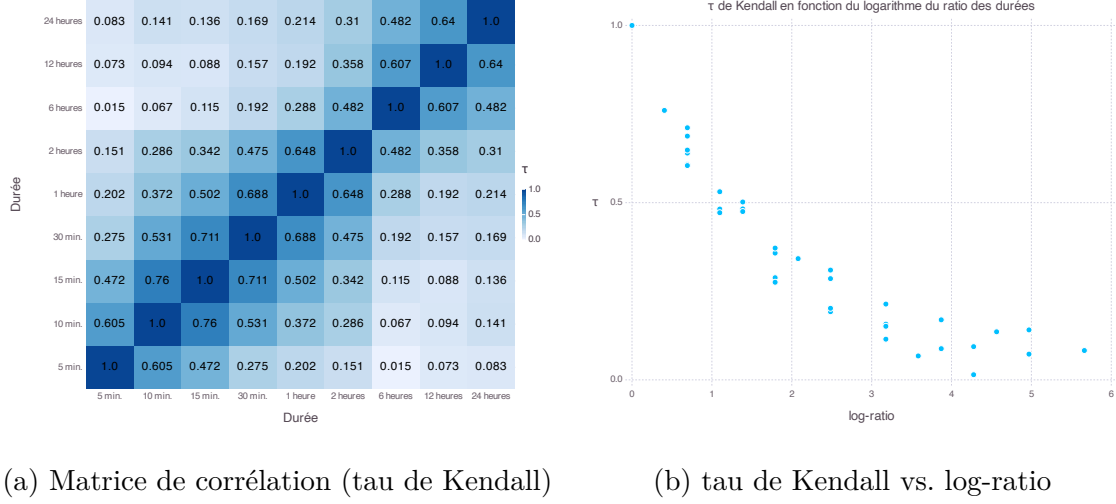


FIGURE 4.4 Corrélation mesurée par le tau de Kendall empirique entre les maxima annuels d'intensité de précipitation pour l'ensemble des durées d'accumulation

4.4 Conclusions de l'analyse exploratoire

4.4.1 Résultats de l'exploration

Cette analyse exploratoire des données nous a permis de constater que la méthodologie d-GEV est une alternative prometteuse à l'estimation indépendante des lois GEV marginales pour chaque durée d'accumulation. En effet, la qualité de l'ajustement de la d-GEV est comparable à celui des lois GEV indépendantes. En tirant parti de la forme fonctionnelle des paramètres d'échelle et de localisation en fonction de la durée, on peut espérer obtenir une estimation plus précise des incertitudes pour les niveaux de retour.

Cependant, nous avons pu constater que, d'une durée à l'autre, les données sont corrélées et que la force de cette corrélation dépend de la distance logarithmique h entre les durées. Ainsi la vraisemblance exprimée à l'équation (3.1), qui suppose l'indépendance des observations d'une durée et d'une année à l'autre, est impropre. En raison de cette vraisemblance impropre, les incertitudes estimées à l'aide de la méthode delta sont réduites de manière irréaliste. En effet, des observations corrélées ont moins de pouvoir informatif que des observations indépendantes. La figure 4.1 permet ainsi de constater que les bandes d'incertitudes sont trop étroites.

4.4.2 Stratégies de résolution

Malgré les limitations évoquées précédemment, nous souhaitons exploiter la relation fonctionnelle d-GEV en mettant en place une stratégie de correction des incertitudes. Deux possibilités sont envisagées dans ce mémoire.

La première stratégie envisagée consiste à modéliser la corrélation présente dans les données à l'aide d'une copule. On supposera que les lois marginales suivent la relation d-GEV et on identifiera une copule adaptée pour modéliser la corrélation dans les données. La vraisemblance du modèle complet exprimée dans l'équation (2.3) est alors propre et les propriétés de normalité asymptotique peuvent être exploitées pour estimer les incertitudes.

La seconde stratégie envisagée consiste à prendre en compte le fait que la vraisemblance (3.1) est impropre et de l'analyser sous l'angle de la théorie des vraisemblances composites. Avec cette méthodologie, on suppose que le modèle est mal spécifié, ce qui est effectivement le cas puisqu'on suppose l'indépendance des observations alors qu'elles ne le sont pas. La méthode d'estimation des incertitudes est adaptée pour prendre en compte cette erreur de spécification.

CHAPITRE 5 MODÉLISATION DE LA DÉPENDANCE ENTRE LES MAXIMA ANNUELS AVEC UNE COPULE

Le premier objectif de ce chapitre est d'identifier les raisons qui expliquent la mauvaise couverture des intervalles de confiance obtenus avec la méthode d-GEV. Davantage que le biais induit par le modèle d-GEV, le fait de ne pas prendre en compte la corrélation présente dans les données semble être un facteur déterminant pour expliquer ce mauvais recouvrement. Le second objectif de ce chapitre consiste à proposer une méthodologie pour corriger ce problème afin d'obtenir des intervalles de confiance ayant un taux de recouvrement adéquat.

5.1 Proposition d'un modèle pour la corrélation

5.1.1 Modélisation de la corrélation

Étant donné la corrélation présente entre les différentes séries de maxima, une modélisation conjointe des maxima pour l'ensemble des durées est nécessaire. On considère le vecteur aléatoire $\mathbf{X} = (X_{d_1}, \dots, X_{d_p})$ des maxima d'intensité de précipitation pour les p durées d'accumulation $d_1, \dots, d_p$, dont la fonction de répartition est

$$F_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(X_{d_1} \leq x_{d_1}, \dots, X_{d_p} \leq x_{d_p}).$$

Cette fonction de répartition multidimensionnelle est difficile à manipuler et on utilise le théorème de Sklar pour écrire la fonction de répartition à l'aide d'une copule C . Cela permet de séparer la structure de la dépendance et les lois marginales. On a alors

$$F_{\mathbf{X}}(\mathbf{x}) = C\{F_{X_{d_1}}(x_{d_1}), \dots, F_{X_{d_p}}(x_{d_p})\}, \quad (5.1)$$

où $F_{X_{d_i}}$ dénote la fonction de répartition marginale de X_{d_i} . À partir de cette écriture, il est possible de conserver marginalement la structure de la d-GEV. C'est-à-dire que pour chaque durée, on aura $F_{X_{d_i}}(x) = \text{GEV}(x \mid \mu_d, \sigma_d, \xi)$, où $\sigma_d = e^{\phi_0} \times (d + \theta)^{-\eta}$ et $\mu_d = \tilde{\mu} \times \sigma_d$. Il faut donc désormais identifier la copule C adéquate pour modéliser la dépendance entre les maxima.

5.1.2 Choix d'une copule

Plusieurs critères doivent être pris en compte pour identifier une copule adaptée à la modélisation de la dépendance entre les extrêmes d'intensité de précipitation. Tout d'abord, on souhaite utiliser une copule qui nous permet d'exprimer la décroissance de la corrélation comme une fonction de la distance h proposée à l'équation (4.2). Ensuite, on souhaite utiliser un modèle simple qui ne nous force pas à multiplier les paramètres. Dans son ouvrage de référence le plus récent, Joe (2014) propose une revue détaillée des familles de copules communément utilisées en analyse multivariée. On va s'intéresser à deux familles couramment utilisées.

La famille des copules archimédiennes

Malgré leur facilité de construction et de mise en œuvre, on doit exclure toute la classe des copules archimédiennes des candidats potentiels. En effet, ces copules ont comme propriété l'échangeabilité des composantes marginales. De ce fait, toutes les composantes ont le même niveau de dépendance. Or nous souhaitons que le degré de dépendance varie comme cela est observée empiriquement. Par exemple, la dépendance entre les maxima d'intensité pour $d = 5$ minutes et $d = 10$ minutes est plus grande que celle entre les maxima d'intensité pour $d = 5$ minutes et $d = 24$ heures. Ce n'est donc pas compatible avec ce que l'on souhaite modéliser.

Les copules archimédiennes hiérarchiques permettent d'obtenir un niveau de dépendance différent entre les marginales à l'aide d'une représentation hiérarchique. Cependant elles ne sont pas non plus adaptées à notre contexte. Prenons en exemple la copule de Gumbel multidimensionnelle, qui est la seule copule archimédienne vérifiant une propriété de max-stabilité, ce qui la rend adaptée à l'étude de maxima corrélés. La fonction de répartition de la copule de Gumbel s'écrit

$$G_\delta(\mathbf{x}) = \exp \left\{ - \left(\sum_{j=1}^p x_j^{-\delta} \right)^{-1/\delta} \right\}.$$

On constate que toutes les composantes marginales sont échangeables, ce qui comme on l'a vu, n'est pas souhaitable. Prenons maintenant la copule archimédienne hiérarchique construite à partir de la copule de Gumbel. On se limitera au cas tridimensionnel pour démontrer le

problème. On a, en notant $x_j = -\ln(u_j)$,

$$C(u_1, u_2, u_3, \delta_1, \delta_2) = \exp \left\{ - \left[\left(x_1^{\delta_1} + x_2^{\delta_1} \right)^{\delta_2/\delta_1} + x_3^{\delta_2} \right]^{1/\delta_2} \right\}.$$

On peut alors écrire la copule sous sa notation archimédienne hiérarchique. On note C_δ la copule de Gumbel bidimensionnelle de paramètre δ , soit

$$C(u_1, u_2, u_3, \delta_1, \delta_2) = C_{\delta_2} \{ C_{\delta_1}(u_1, u_2), u_3 \}.$$

On constate que les composantes u_1 et u_2 sont échangeables et ont le même niveau de dépendance à la troisième composante u_3 . Or nous souhaitons pouvoir avoir un niveau de dépendance différent pour chaque composante. Ce problème survient avec l'ensemble des copules archimédiennes hiérarchiques car au sein d'un même niveau hiérarchique, les composantes sont échangeables.

Familles des copules elliptiques

Pour exploiter la décroissance de la corrélation en fonction de la distance $h(\cdot, \cdot)$ entre les durées, la stratégie naturelle consiste à choisir une famille de copules qui utilise une matrice pour caractériser la dépendance entre les composantes. Le niveau de dépendance entre deux composantes marginales peut alors être une fonction des durées qui leur sont associées. Bien que d'autres choix soient possibles, le choix d'une copule elliptique est le plus courant. C'est par exemple le choix qui a été fait par Van de Vyver et Van den Bergh (2018), qui ont utilisé une copule gaussienne. En particulier, notre objectif étant de corriger les incertitudes estimées par la méthode d-GEV et pas de modéliser très exactement la dépendance entre chaque variable, un autre choix de copule ajouterait un niveau de complexité dans la modélisation qui ne serait pas forcément souhaitable.

Les deux copules elliptiques les plus courantes sont la copule de Student et la copule gaussienne. Elles sont toutes les deux construites à partir de la distribution multidimensionnelle du même nom. On privilégiera la copule de Student car celle-ci permet de modéliser n'importe quel niveau de dépendance dans les queues de distribution marginales alors que la copule gaussienne ne contient pas de dépendance de queue (Joe, 2014). Par ailleurs, la copule de Student tend vers la copule gaussienne lorsque le nombre de degrés de liberté augmente. Le choix d'une copule de Student n'est donc pas limitant et ne nécessite pas de poser d'hypothèse d'indépendance dans les queues de distribution.

La copule de Student à ν degrés de liberté s'écrit

$$C_{\Sigma, \nu}(u_1, \dots, u_p) = t_{\Sigma, \nu}\{t_{\nu}^{-1}(u_1), \dots, t_{\nu}^{-1}(u_p)\},$$

où $t_{\Sigma, \nu}$ est la fonction de répartition d'une loi de Student multidimensionnelle de matrice d'échelle Σ et ν degrés de liberté et t_{ν}^{-1} est la fonction quantile d'une distribution de Student univariée à ν degrés de liberté.

5.1.3 Considérations sur la mesure de corrélation et la copule de Student

Pour la copule de Student, si $\rho_{d, d'}$ est le terme de la matrice de corrélation des durées d et d' , alors le tau de Kendall empirique s'exprime sous la forme $\tau = 2/\pi \times \arcsin(\rho_{d, d'})$; voir Demarta et McNeil (2005). On a donc aussi $\rho_{d, d'} = \sin(\tau\pi/2)$. La même formule s'applique à la copule gaussienne, ce qui est attendu puisque la copule gaussienne est la limite de la copule de Student dans le cas $\nu \rightarrow \infty$.

Il n'existe pas de formule simple pour le rho de Spearman. Par ailleurs, la mesure du rho de Spearman dépend du nombre de degrés de liberté ν . On notera néanmoins qu'une formule a été établie par Heinen et Valdesogo (2020). On s'appuiera donc sur le tau de Kendall, qui a l'avantage de présenter une formule simple qui soit à la fois utilisable avec la copule de Student et la copule gaussienne. On peut estimer ce paramètre au moyen de la formule suivante :

$$\hat{\tau}_n = \frac{2}{n(n-1)} \sum_{i < j} \text{sign}\{(x_i - x_j)(y_i - y_j)\}.$$

Si l'on souhaite utiliser une copule dont les termes de la matrice de corrélation sont fonction de h , il est nécessaire de vérifier qu'il existe bien une relation entre $\sin(\tau_{d, d'} \times \pi/2)$ et $h(d, d')$. La figure 5.1 permet de vérifier qu'une telle relation existe effectivement.

5.1.4 Choix de la forme paramétrique la corrélation

La copule de Student permet de modéliser la corrélation au sens du tau de Kendall comme une fonction de h . De plus, elle permet de modéliser tout niveau de dépendance dans la queue de distribution. Elle est donc adaptée à la modélisation de la dépendance des maxima annuels de précipitation. Par conséquent, on a fait le choix d'une copule de Student.

Étant donné que la copule de Student est construite à partir de la loi de Student multidimensionnelle, on doit lui associer une matrice d'échelle Σ . Pour une loi de Student multidimensionnelle, la matrice d'échelle est liée à la matrice de covariance du vecteur aléatoire par

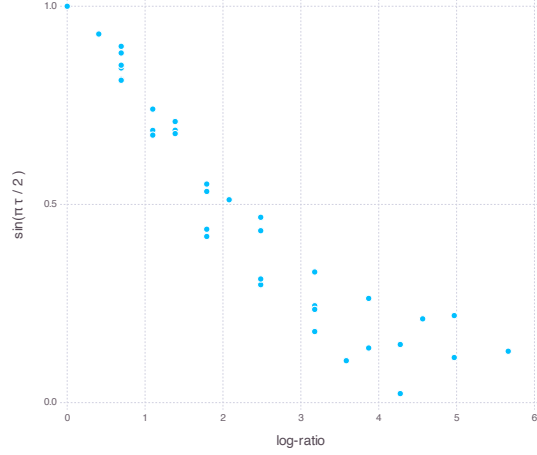


FIGURE 5.1 Représentation de $\sin(\pi\hat{\tau}/2)$ en fonction de h

la relation

$$\text{cov}(\mathbf{X}) = \frac{\nu}{\nu - 2} \Sigma.$$

Cependant, nous n'aurons pas à nous inquiéter en pratique de ce facteur multiplicatif car, comme le précisent Demarta et McNeil (2005), la copule de Student est invariante à toute transformation strictement croissante de ses composantes et en particulier à la standardisation des composantes. La copule de Student peut donc être définie à l'aide de la matrice de corrélation car celle-ci est une transformation linéaire strictement croissante de la matrice d'échelle. Par souci de clarté, on utilisera toujours P la matrice de corrélation induite dont les coefficients $P_{i,j}$ correspondent à la corrélation de Pearson entre les séries d'observations pour les durées d_i et d_j lorsque ces séries sont ramenées à l'échelle d'une distribution de Student unidimensionnelle. On a par définition $P_{i,i} = 1$ pour tout entier $i \in \{1, \dots, p\}$.

On doit désormais choisir une forme paramétrique pour les termes de P en fonction de h . La première contrainte est que les termes diagonaux de P soient égaux à 1. Il faut aussi que la matrice de corrélation soit symétrique et définie positive. Un choix courant en statistique spatiale est d'utiliser une matrice de covariance de Matérn (Cressie et Wikle, 2011). C'est par exemple le choix qui a été fait par Van de Vyver et Van den Bergh (2018). Or, on a défini une distance h entre les durées d'accumulation, ce qui nous permet de considérer le problème comme un problème spatial. La covariance de Matérn peut être paramétrée ainsi :

$$M(h) = \sigma^2 \times \frac{2^{1-\kappa}}{\Gamma(\kappa)} \times \left(\sqrt{2\kappa} \frac{h}{\rho} \right)^\kappa \times K_\kappa \left(\sqrt{2\kappa} \frac{h}{\rho} \right),$$

où

Γ est la fonction Gamma d'Euler ;

K_κ est la fonction de Bessel modifiée de seconde espèce d'ordre κ ;

$\rho > 0$ est un paramètre d'échelle ;

$\kappa > 0$ est un paramètre de forme ;

σ^2 est la variance marginale des composantes.

Comme on souhaite une matrice de corrélation, on a $\sigma^2 = 1$. Ainsi, les termes de la matrice P s'écrivent

$$P_{i,j} = \frac{2^{1-\kappa}}{\Gamma(\kappa)} \times \left(\sqrt{2\kappa} \frac{h_{i,j}}{\rho} \right)^\kappa \times K_\kappa \left(\sqrt{2\kappa} \frac{h_{i,j}}{\rho} \right) \quad h_{i,j} = h(d_i, d_j). \quad (5.2)$$

La matrice P de la copule de Student est donc fonction de deux paramètres ρ et κ , ainsi que des durées. Comme nous souhaitons pouvoir différencier la vraisemblance automatiquement pour obtenir des intervalles de confiance, il faut une implantation de P qui soit différentiable par rapport aux paramètres κ et ρ . La fonction K_κ a été implantée de façon compatible avec la différentiation automatique par Geoga *et al.* (2022).

On observe sur la figure 5.2 les ajustements pour trois stations de la forme paramétrique proposée par rapport à $\sin(\pi\tau/2)$: les aéroports Trudeau (à Montréal), Jean-Lesage (à Québec), et la station du lac Témiscamingue. Les stations choisies sont celles qui contiennent le plus

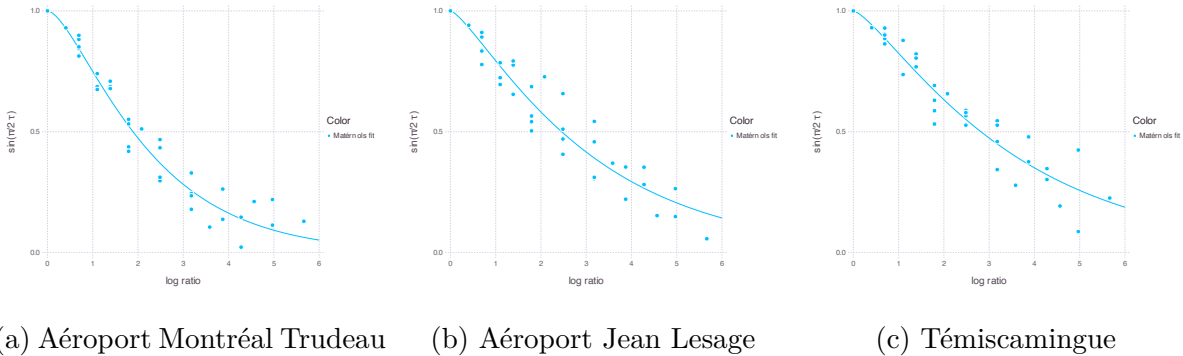


FIGURE 5.2 Ajustement de la relation fonctionnelle de Matérn par la méthode des moindres carrés

de données et qui ont donc une estimation plus précise du tau de Kendall. Il est important de noter que les ajustements illustrés sont obtenus par la méthode des moindres carrés. Dans la suite du travail, on utilisera des méthodes de maximisation de la vraisemblance pour estimer les paramètres ρ et κ . Il sera intéressant de comparer les courbes obtenues par régression avec celles obtenues par maximisation de la vraisemblance.

5.1.5 Modèle complet et méthodologie d'estimation des paramètres

Modèle

Le modèle que l'on va estimer est composé de marges qui suivent chacune une loi GEV et dont les paramètres de localisation et d'échelle sont liés par la relation fonctionnelle d-GEV. Elles sont associées à une structure de dépendance utilisant une copule de Student dont les termes de la matrice de corrélation sont définis à l'aide de la fonction de covariance de Matérn. L'utilisation d'une copule de Student implique de devoir choisir le nombre ν de degrés de liberté. Ce paramètre pourrait être considéré comme un paramètre du modèle à optimiser, mais il est préférable de prévoir une procédure distincte pour estimer ce paramètre que l'on choisit généralement entier.

Si l'on note ν le nombre de degrés de liberté de la copule de Student, la log-vraisemblance du modèle peut être écrite comme la somme de la log-vraisemblance du modèle marginal et de celle de la copule. On dénote $x_{i,d}$ le maxima d'intensité observé l'année i pour la durée d et $u_{i,d} = F_{(X_d|\boldsymbol{\theta})}(x_{i,d})$ les observations ramenées à l'échelle uniformes sur $[0, 1]$, où $\boldsymbol{\theta} = (\tilde{\mu}, \phi_0, \xi, \theta, \eta)$ dénote le vecteur de paramètres du modèle d-GEV. On a

$$\ell(\boldsymbol{\theta}, \rho, \kappa \mid \mathbf{x}) = \ell(\boldsymbol{\theta} \mid \mathbf{x}) + \ell_{C_\nu}(\rho, \kappa \mid \mathbf{u}), \quad (5.3)$$

où

$$\mathbf{x} = (x_{i,d_j} : 1 \leq i \leq n, 1 \leq j \leq p);$$

$$\mathbf{u} = (u_{i,d_j} : 1 \leq i \leq n, 1 \leq j \leq p);$$

(ρ, κ) sont les paramètres de la fonction de covariance de Matérn;

$\ell(\boldsymbol{\theta} \mid \mathbf{x})$ est la log-vraisemblance du modèle d-GEV évaluée en $\boldsymbol{\theta}$;

$\ell_{C_\nu}(\rho, \kappa \mid \mathbf{u})$ est la log-vraisemblance de la copule de Student à ν degrés de liberté évaluée en (ρ, κ) .

La log-vraisemblance de la copule ℓ_{C_ν} est donnée par

$$\ell_{C_\nu}\{\rho, \kappa \mid (u_{i,d})_{i,d}\} = \sum_{i=1}^n \ln\{f_{\nu, P_{\rho, \kappa}}(\mathbf{u})\} - \sum_{i=1}^n \sum_{d \in D} \ln\{f_\nu(u_{i,d})\},$$

où f_ν est la densité d'une loi de Student à ν degrés de liberté et $f_{\nu, P}$ est la densité de la loi de Student multidimensionnelle à ν degrés de liberté de matrice d'échelle $P_{\rho, \kappa}$. Les termes de $P_{\rho, \kappa}$ sont donnés par l'équation (5.2).

Il est important de noter que les observations uniformisées $u_{i,d} = F_{X_d|\boldsymbol{\theta}}(x_{i,d})$ dépendent du vecteur de paramètre d-GEV $\boldsymbol{\theta}$. Ainsi, la vraisemblance de la copule de Student dépend également de $\boldsymbol{\theta}$. Dans le cadre d'une maximisation de la vraisemblance du modèle complet, l'ajout de la copule peut donc avoir une influence sur l'estimation des paramètres d-GEV.

Méthodologie d'estimation des paramètres

Le modèle choisi contient ainsi huit paramètres : cinq pour les marginales d-GEV et trois pour la copule de Student. Joe (2005) note que dans le cadre de lois marginales GEV, l'estimation du modèle global peut être difficile et aboutir à un maximum local de vraisemblance. Par ailleurs, le maximum global peut être mal défini si la courbure de la log-vraisemblance au niveau du maximum local est aplatie. Il suggère alors de procéder à une estimation en deux étapes. On estime d'abord les paramètres marginaux, puis on estime ensuite les paramètres de dépendance en fixant les paramètres marginaux à leurs valeurs estimées. Cette procédure a l'avantage de ne pas modifier l'ajustement marginal, qui était déjà assez bon. Cependant, son inconvénient principal est que les propriétés de normalité asymptotique des paramètres estimés sont modifiées par l'estimation en deux étapes (Joe, 2005). On ne peut ainsi pas simplement faire l'estimation en deux étapes puis utiliser la matrice hessienne du modèle complet pour obtenir les intervalles de confiance via la méthode Delta.

La solution proposée par Joe (2005) pour obtenir des propriétés de normalité asymptotique avec l'estimation en deux étapes suppose que les lois marginales aient été estimées indépendamment. Le cas de la d-GEV, pour laquelle toutes les marginales sont estimées conjointement, n'est donc pas conforme au cadre posé dans l'article. Une autre possibilité pour éviter ces difficultés est de procéder à l'estimation en deux étapes, puis d'utiliser les valeurs des paramètres obtenus comme valeurs de départ pour l'estimation globale. C'est la procédure que l'on favorisera.

5.1.6 Résultats

Choix du nombre de degrés de liberté

On utilise à nouveau les données de précipitation mesurées à l'aéroport Trudeau comme exemple pour illustrer la méthode. On maximise la vraisemblance du modèle complet à l'aide de la méthodologie présentée dans la section précédente. Pour le choix du nombre de degrés de liberté de la copule de Student, on effectue la maximisation de la vraisemblance avec une large plage de nombres de degrés de liberté différents. Les résultats montrés sur la figure 5.3 montrent que la log-vraisemblance estimée est une fonction croissante du nombre de degrés de liberté pour les trois stations étudiées. Ce résultat sous-entend que la copule gaussienne, qui peut être envisagée comme la limite de la copule de Student dans le cas limite $\nu \rightarrow \infty$, serait adéquate pour modéliser la dépendance dans les données. Cela signifie également l'absence de dépendance caudale asymptotique dans les données.

Malgré ce résultat qui semble prescrire l'usage de la copule gaussienne, l'étude de la copule de Student nous a permis de ne pas avoir à poser d'hypothèses a priori sur la dépendance asymptotique dans les données. Le degré de confiance que nous avons dans le modèle avec une copule gaussienne est ainsi plus grand que si nous l'avions utilisé d'emblée, sans tester d'abord la copule de Student. En accord avec les résultats, on utilisera une copule gaussienne pour modéliser les données d'intensité de précipitation.

Résultats

On peut désormais utiliser le nouveau modèle, consistant à associer une copule gaussienne et le modèle d-GEV, pour tracer les diagrammes quantiles-quantiles pour les différentes durées d'accumulation. Les deux éléments principaux que nous souhaitons évaluer sont les suivants :

- (i) L'ajustement marginal du modèle suite à l'ajout de la copule.
- (ii) Le niveau d'incertitude obtenu à l'aide de la méthode delta en comparaison de la méthodologie d-GEV classique et de l'estimation GEV indépendante.

Les résultats de l'estimation du modèle pour les données de l'aéroport Trudeau sont présentés à la figure 5.4 sous forme de diagrammes quantile-quantile. La sélection de durées est la même qu'à la partie précédente. Ces résultats sont mis en miroir de ceux obtenus avec la méthode d-GEV seule. Pour des fins de comparaison, l'échelle utilisée est la même pour les deux

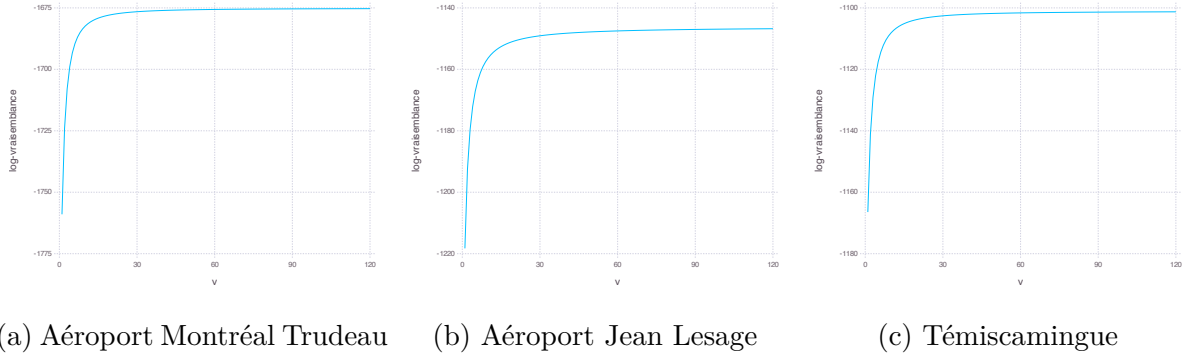


FIGURE 5.3 Log-vraisemblance maximisée en fonction du nombre de degrés de liberté de la copule de Student

méthodes.

On constate que l'ajustement marginal a été modifié par l'ajout de la copule à la modélisation. Selon les cas, l'ajustement est amélioré, comme c'est par exemple le cas pour la durée de 6h, ou il est détérioré ; c'est par exemple le cas pour la durée de 24 heures. Pour la durée $d = 24h$, on note dans les deux modèles que l'observation la plus extrême sort largement de l'intervalle de confiance. On remarque en outre que les bandes d'incertitudes au niveau nominal 95% sont légèrement élargies en comparaison du modèle d-GEV simple. Cependant cette

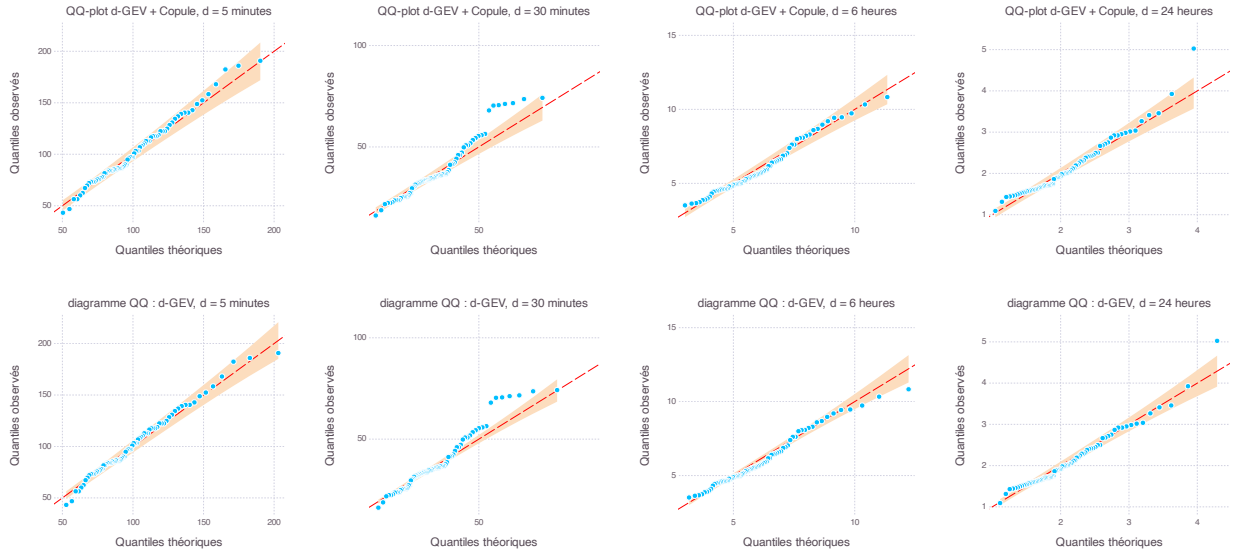


FIGURE 5.4 Diagrammes quantiles-quantiles obtenus pour le modèle d-GEV avec et sans copule pour une sélection de durées

augmentation ne semble pas significative. Pour la durée $d = 30\text{m}$, on constate que beaucoup d'observations sortent des bandes d'incertitudes pour les deux modèles. Ceci démontre que l'ajustement obtenu via un modèle multidimensionnel peut marginalement être mauvais. En effet, avec le modèle GEV estimé indépendamment, les bandes d'incertitudes recouvraient la majorité des observations.

Pour comparer les modèles, on peut aussi s'intéresser aux niveaux de retour à $T = 100$ ans ainsi qu'aux incertitudes associées. Les résultats sont présentés dans le tableau 5.1. Pour l'ensemble des durées, les intervalles de confiance obtenus avec le modèle GEV recouvrent entièrement les incertitudes des deux modèles d-GEV. On constate que les intervalles de confiance pour les deux modèles d-GEV sont de tailles comparables. Ceci n'était pas attendu, car nous devrions avoir des incertitudes plus petites pour le modèle d-GEV sans copule qui surestime l'information. Par ailleurs, les estimations sont plus basses pour le modèle d-GEV avec copule. Cela est lié au fait que le paramètre de forme ξ est positif pour le modèle d-GEV simple : $\hat{\xi}_{\text{dGEV}} = 0.043$ et négatif pour le modèle avec copule $\hat{\xi}_{\text{dGEV/Cop}} = -0.033$. Les queues de distribution estimées sont donc plus lourdes dans le cas de la d-GEV simple.

Ajustement de la corrélation

Les paramètres de dépendance estimés par la méthode de maximisation de la vraisemblance sont $(\hat{\rho}, \hat{\kappa}) = (2.245, 0.981)$. On peut ainsi comparer la corrélation de Matérn correspondante à celle obtenue par une méthode de minimisation des carrés. Cette comparaison est affichée à

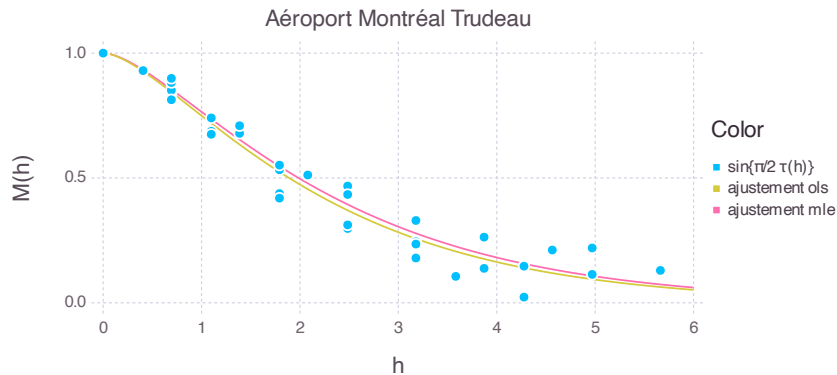


FIGURE 5.5 Ajustements des corrélations obtenues par maximisation de la vraisemblance et par minimisation des erreurs quadratiques

TABLEAU 5.1 Niveau de retour $T = 100$ estimés par les 3 méthodes et bornes des intervalles de confiance au niveau nominal 95% (mm/h)

Durée	GEV			d-GEV			d-GEV/Copule		
	inf.	$\hat{q}$	sup.	inf.	$\hat{q}$	sup.	inf.	$\hat{q}$	sup.
5 min.	159.3	195.0	230.8	191.2	210.7	230.2	176.4	195.9	215.4
15 min.	94.4	120.6	146.8	110.2	119.5	128.9	100.7	111.4	122.1
1 h.	37.9	51.3	64.7	43.7	47.5	51.3	39.9	44.2	48.4
6 h.	8.8	12.65	16.5	11.7	12.7	13.7	10.6	11.7	12.7
24h.	3.3	4.6	5.9	4.0	4.5	4.9	3.6	4.1	4.5

la figure 5.5. Les ajustements obtenus à l'aide des deux méthodes sont comparables et suivent le diagramme de dispersion de $\sin(\tau\pi/2)$ de manière précise. Il semblerait donc que le choix de la forme de corrélation soit adéquat.

5.1.7 Conclusions intermédiaires

La copule gaussienne avec covariance de Matérn semble être adéquate pour modéliser la dépendance dans les données. En effet, les paramètres de dépendance estimés par maximisation de la vraisemblance permettent de reproduire la forme du sinus du tau de Kendall empirique de manière fidèle. Ce résultat est important, car il permet de mettre en avant une structure de dépendance inter-durée pour les maxima d'intensité de précipitation. Cependant, l'estimation des incertitudes n'est pas satisfaisante car les intervalles de confiance estimés sont toujours trop étroits.

On peut expliquer ces résultats par le fait que le modèle utilisé est trop peu flexible. En effet, le modèle complet repose sur des postulats forts :

- (i) les observations marginales doivent suivre une loi GEV ;
- (ii) les paramètres de localisation et d'échelle des lois marginales doivent être régis par une relation fonctionnelle dépendant de la durée d'accumulation ;
- (iii) le paramètre de forme doit être commun à l'ensemble des durées ;
- (iv) la structure de dépendance doit être décrite par une copule gaussienne avec une fonction de covariance de Matérn.

Ces trois postulats sont résumés par un modèle contenant seulement sept paramètres, alors

que les phénomènes physiques qui régissent la dynamique des précipitations sont particulièrement complexes. Les données réelles ne peuvent donc pas être très exactement capturées par le modèle choisi. Or la méthode d'estimation des incertitudes avec la méthode Delta repose sur l'hypothèse que les données proviennent effectivement du modèle paramétrique utilisé. Dans certains cas, la modélisation décrit suffisamment bien le phénomène réel pour que les incertitudes estimées soient acceptables. Dans notre cas, le modèle est trop restrictif et trop peu flexible et les incertitudes estimées ne sont, hélas, pas adéquates.

Bien qu'il ne soit pas adapté à l'étude des incertitudes, le modèle offre un niveau d'ajustement marginal qui reste bon et peut être utilisé pour remplacer les méthodes d'ajustements indépendants des lois marginales. Par ailleurs, il permet de recueillir de l'information sur la structure de dépendance, ce qui peut être utile à d'autres fins.

5.2 Étude de simulation : Taux de couverture des intervalles de confiance lorsque les données sont générées par le modèle proposé

5.2.1 Objectif de la démarche

L'objectif de cette partie est d'étudier, indépendamment de possibles effets liés à une mauvaise spécification du modèle, les conséquences de la corrélation entre les composantes marginales lorsque celles-ci sont d-GEV. Pour cela, on utilisera des données synthétiques générées avec des paramètres connus pour comparer les résultats obtenus à l'aide des trois méthodes possibles :

- (i) l'estimation GEV indépendante ;
- (ii) l'estimation d-GEV sans prendre en compte la corrélation ;
- (iii) l'estimation de la d-GEV avec prise en compte de la corrélation.

Les aspects sur lesquels nous focaliserons sont les suivants :

- (a) le biais et la variance dans l'estimation des niveaux de retour ;
- (b) la couverture des intervalles de confiance obtenus à l'aide de la méthode Delta et la taille moyenne de ceux-ci.

5.2.2 Résultats attendus

Les trois méthodes étudiées correspondent à une spécification marginale exacte du modèle. En effet, les données seront générées à l'aide de composantes marginales GEV et les paramètres de localisation et d'échelle suivent une loi d-GEV. L'estimation des paramètres GEV ou d-GEV ne devrait donc pas être biaisée et on s'attend donc à ce que l'estimation des niveaux de retour ne présente pas un biais significatif.

On s'attend aussi à ce que la variance dans l'estimation des niveaux de retour soit plus grande lorsque l'on estime les marginales indépendamment que lorsque l'on utilise le modèle d-GEV. En effet, pour les modèles d-GEV, l'ensemble des données est utilisé pour estimer les paramètres.

Les intervalles de confiance au niveau 95% pour les niveaux de retour estimés avec la méthode d-GEV avec copule et ceux obtenus en estimant indépendamment les lois marginales GEV devraient contenir dans une proportion proche de 95% les niveaux de retour exacts. On note néanmoins que selon la taille des échantillons utilisés, le régime de normalité asymptotique peut ne pas être atteint, ce qui peut mener à une couverture différente de 95%. Ces intervalles de confiance devraient être plus larges lorsque l'on estime les marginales indépendamment plutôt qu'avec le modèle d-GEV avec copule. On s'attend à ce que l'omission de la dépendance dans le modèle d-GEV simple mène à des intervalles de confiance trop étroits et à un recouvrement plus faible que 95%.

5.2.3 Données synthétiques

Pour chaque scénario étudié et à chaque itération de l'étude de simulation, on génère un jeu de données synthétiques avec des paramètres connus. Les durées d'accumulation utilisées sont les mêmes que pour les données réelles soit : 5 minutes, 10 minutes, 15 minutes, 30 minutes, 1 heure, 2 heures, 6 heures, 12 heures et 24 heures.

On fixe la taille des échantillons à $n = 75$, ce qui est de l'ordre de grandeur des jeux de données les plus grands disponibles. Cela nous permettra de mettre en exergue les limites de nos modèles du point de vue de l'état de normalité asymptotique lorsque l'on utilise des données réelles. On discriminera les effets liés à la normalité asymptotique de ceux liés à une mauvaise spécification du modèle en faisant les mêmes expériences avec $n = 500$. Les résultats obtenus avec $n = 500$ sont consignés dans l'annexe A.

Les paramètres d-GEV utilisés pour générer les données seront fixés aux valeurs suivantes :

$$(\tilde{\mu}, \phi_0, \xi, \theta, \eta) = (5, 1.5, 0.1, 0.05, 0.75).$$

Bien qu'arbitraires, ces valeurs correspondent à celles qui pourraient être rencontrées dans des données réelles. On pourrait utiliser d'autres valeurs pour comparer les résultats, mais ici on souhaite principalement étudier les limitations de la méthode d-GEV lorsque l'on ne modélise pas la dépendance ; faire varier les paramètres d-GEV n'est donc pas aussi intéressant que de faire varier les paramètres de dépendance. On note que l'on évite d'utiliser des valeurs trop grandes pour ξ , car les propriétés des estimateurs à vraisemblance maximale ne sont pas assurées pour $|\xi| \geq 0.5$ (Coles, 2001).

Pour la dépendance, on utilise une copule gaussienne avec matrice de corrélation de Matérn. On va utiliser deux jeux de paramètres différents afin de mesurer l'impact de la dépendance. On pourra d'abord prendre $\rho = 3$ et $\kappa = 1$, ce qui représente une corrélation importante qui décroît, par exemple, moins vite avec h que la corrélation observée à l'aéroport Trudeau. On prendra aussi $\rho = 1$ et $\kappa = 0.5$, ce qui correspond à un cas avec une corrélation décroissant plus vite avec h qu'à l'aéroport Trudeau. Les deux motifs sont visibles à la figure 5.6.

5.2.4 Procédure d'estimation

Afin de pouvoir tirer des conclusions générales sur les différentes méthodes d'estimations, on va procéder à $N = 1000$ itérations pour chacun des cas considérés, c'est-à-dire pour les différentes valeurs des paramètres de dépendance et les différentes tailles d'échantillon. À chaque itération :

1. On génère un jeu de données de paramètres connus.
2. On obtient les estimateurs des paramètres pour chacun des modèles considérés en maximisant la log-vraisemblance correspondante.
3. On calcule les estimateurs du niveau de retour pour $T = 100$ ans pour chaque durée

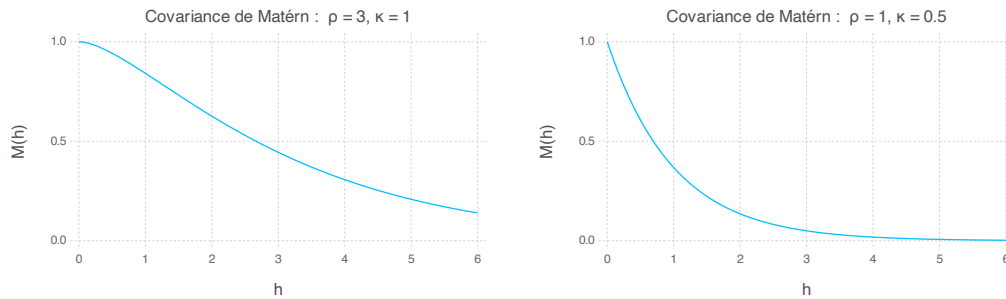


FIGURE 5.6 Fonctions de Matérn utilisées pour l'étude de simulation

et chaque modèle.

4. On calcule l'écart-type de l'estimateur à l'aide de la méthode Delta.
5. On calcule les intervalles de confiance au niveau nominal 95%.
6. On vérifie si le vrai niveau de retour est compris dans l'intervalle de confiance estimé.

5.2.5 Résultats

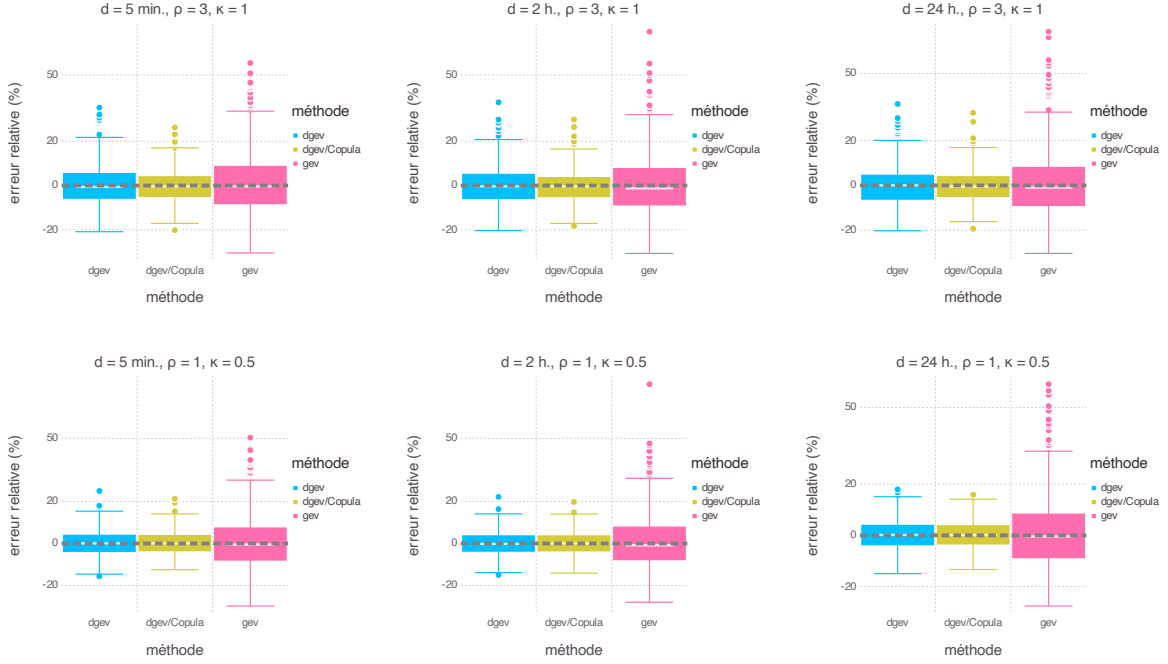


FIGURE 5.7 Diagrammes en boîte des résultats de l'étude de simulation pour une sélection de durées

Biais et dispersion : On s'intéresse d'abord au biais et à la dispersion des estimateurs de $\hat{q}_{100}$ le niveau de retour associé à la période $T = 100$ ans. Les résultats pour l'étude avec taille d'échantillon $n = 75$ sont présentés dans les diagramme en boîtes (ou hamacs) de la figure 5.7. Afin que les résultats soient comparables d'une durée à l'autre, on montre l'erreur relative, à savoir

$$\hat{e}_{q_{100}} = \frac{\hat{q}_{100} - q_{100}}{q_{100}}.$$

Par ailleurs, on a pris soin d'avoir une échelle commune pour un même niveau de dépendance. Les graphes de la ligne supérieure correspondent au niveau de dépendance élevé et ceux de la ligne inférieure au niveau de dépendance plus faible. On observe pour les deux niveaux

de dépendance qu'aucune des méthodes ne présente de biais systématique. On constate que pour l'ensemble des méthodes, les valeurs aberrantes estimées correspondent à un biais positif. Ceci est probablement lié au choix d'un paramètre de forme positif pour générer les données. Par ailleurs, on constate que pour une méthode donnée, la dispersion de l'erreur relative ne semble pas varier significativement d'une durée à l'autre. Ceci indique que la précision de la méthodologie est similaire pour toutes les durées d'accumulation. Cette dernière hypothèse a également été vérifiée numériquement.

Les niveaux de retour estimés à l'aide de la méthode GEV présentent une dispersion plus grande qu'avec les deux autres méthodes, ainsi que davantage de valeurs aberrantes. Le fait que les estimations marginales indépendantes utilisent une moins grande quantité de données pour estimer les trois paramètres GEV explique ces résultats. Pour cette méthode, il n'y a pas de différence significative entre les deux niveaux de dépendance, ce qui est normal car marginalement la dépendance n'a pas d'impact sur le pouvoir informatif des observations.

Les deux méthodes d-GEV présentent une dispersion plus faible que la méthode GEV. Ceci s'explique par l'utilisation simultanée de l'ensemble des données pour estimer les paramètres. On constate que la dispersion est plus faible pour la méthode d-GEV avec copule que pour la méthode d-GEV sans copule, et ce d'autant plus que la dépendance est forte. Ainsi, pour le niveau de dépendance élevé, la boîte est significativement plus fine pour la méthode d-GEV avec copule par rapport à la méthode d-GEV sans copule.

Couverture et largeur des intervalles de confiance : Les résultats de l'étude pour le niveau de dépendance élevé et le niveau de dépendance faible sont montrés respectivement sur les tableaux 5.2 et 5.3. On montre, d'une part, la couverture des intervalles de confiance estimés au niveau nominal 95%, c'est-à-dire la proportion de ces intervalles qui contiennent effectivement le niveau de retour réel. On peut faire les observations suivantes :

1. Pour le modèle GEV, la couverture ne varie pas significativement avec le niveau de dépendance et semble être aux alentours de 90%.
2. Pour la méthode d-GEV, la couverture est significativement plus faible lorsque le niveau de dépendance est plus grand.
3. Pour le modèle d-GEV avec copule, la couverture est légèrement inférieure à 95% dans les deux cas. Par ailleurs, la couverture semble être meilleure lorsque le niveau de dépendance est plus faible.

D'autre part, on peut faire les observations suivantes pour l'écart-type moyen calculé à l'aide de la méthode delta :

1. L'écart-type moyen le plus faible est obtenu avec la méthode d-GEV et ne varie pas de façon appréciable avec le niveau de dépendance.
2. Le modèle d-GEV avec copule présente un écart-type intermédiaire entre les deux méthodes qui évolue avec le niveau de dépendance. Plus la dépendance est forte, plus l'écart-type moyen est grand.
3. L'écart-type moyen obtenu avec la méthode GEV est beaucoup plus important qu'avec les deux autres méthodes et ne dépend pas du niveau de dépendance.

Pour le modèle GEV, le taux de couverture et l'écart-type qui ne varient pas d'un niveau de dépendance à l'autre sont cohérents avec le fait que la dépendance n'a pas d'incidence marginalement sur l'information apportée par les observations. La sous-couverture observée peut être attribuée au fait que le régime de normalité asymptotique n'est pas atteint avec seulement 75 observations. Lorsqu'on prend $n = 500$, le taux de couverture est très proche de 95%, ce qui tend à valider cette hypothèse (voir Annexe A). Malgré cette sous-couverture, l'écart-type moyen est presque le double de celui obtenu avec les autres méthodes.

Pour le modèle d-GEV, le taux de couverture dépend fortement du niveau de dépendance. Le fait de ne pas prendre en compte la dépendance entraîne une sous-estimation notable de l'écart-type du niveau de retour et donc une sous-couverture des intervalles de confiance. Cela confirme que l'information apportée par les observations est d'autant plus faible qu'elles sont corrélées et que le modèle d-GEV surestime l'information effective.

Pour le modèle d-GEV avec copule, on a le meilleur taux de couverture ainsi que des incertitudes intermédiaires entre le modèle GEV et le modèle d-GEV. Le fait que la couverture soit inférieure à 95% peut être attribué au fait que le régime de normalité asymptotique n'est pas atteint. Par ailleurs, cela explique aussi le fait que la couverture soit meilleure avec un niveau de dépendance plus faible. En effet, moins les observations sont corrélées, plus elles sont informatives et plus le régime de normalité asymptotique est atteint rapidement. Lorsque l'on prend des échantillons de taille $n = 500$ la couverture pour le modèle d-GEV avec copule tend vers 95% pour les deux niveaux de dépendance. (voir Annexe A).

On peut remarquer de façon annexe que pour des jeux de données avec peu d'années et un niveau faible de dépendance, les intervalles de confiance estimés avec la d-GEV pourraient avoir

TABLEAU 5.2 Taux de couverture des intervalles de confiance estimés au niveau nominal 95% et écart-type moyen estimé avec la méthode delta pour $\rho = 3$ et $\kappa = 1$

Modèle	$d = 5$ minutes		$d = 2$ heures		$d = 24$ heures	
	Couverture	$\bar{\sigma}_{q_{5min}}$	Couverture	$\bar{\sigma}_{q_{2h}}$	Couverture	$\bar{\sigma}_{q_{24h}}$
d-GEV	71.8%	10.0	68.3%	1.18	70%	0.195
d-GEV/Copule	92.8%	14.5	93.5%	1.85	93.6%	0.295
GEV	88.5%	28.9	90.7%	3.7	89.6%	0.589

un meilleur taux de couverture que ceux estimés avec la GEV du fait que la GEV nécessite davantage d'années d'observations pour atteindre le régime de normalité asymptotique.

5.2.6 Conclusion de l'étude de simulation

À l'aide de cette étude de simulation, on a montré que lorsque les données sont effectivement générées à l'aide de marginales corrélées dont les paramètres sont d-GEV, le modèle d-GEV à marginales indépendantes peut significativement sous-estimer les incertitudes en ne prenant pas en compte la baisse d'information provoquée par la corrélation. Les intervalles de confiance obtenus par la méthode Delta présentent en effet une couverture plus faible que celle recherchée.

On a aussi observé que la méthode GEV, bien qu'elle corresponde à une spécification adéquate pour estimer les lois marginales, souffre de deux limitations. Tout d'abord les intervalles de confiance obtenus par la méthode Delta sont bien plus larges que pour les autres méthodes et on a pu constater une dispersion très importante des niveaux de retour estimés. Ensuite, avec une quantité de données limitée, ce qui est souvent le cas avec des données réelles, le régime asymptotique peut ne pas être atteint, ce qui mène à un recouvrement trop faible des intervalles de confiance.

La modélisation avec copule permet d'obtenir les meilleurs résultats. La largeur des intervalles de confiance dépend du degré de corrélation entre les marginales mais tire néanmoins parti

TABLEAU 5.3 Taux de couverture des intervalles de confiance estimés au niveau nominal 95% et écart-type moyen estimé avec la méthode delta pour $\rho = 1$ et $\kappa = 0.5$

Modèle	$d = 5$ minutes		$d = 2$ heures		$d = 24$ heures	
	Couverture	$\bar{\sigma}_{q_{5min}}$	Couverture	$\bar{\sigma}_{q_{2h}}$	Couverture	$\bar{\sigma}_{q_{24h}}$
d-GEV	87.8%	9.9	87.1%	1.16	88.7%	0.193
d-GEV/Copule	94.9%	11.4	94.4%	1.38	95.1%	0.227
GEV	90.1%	28.5	90.6%	3.71	90%	0.595

de l'estimation conjointe de l'ensemble des données pour obtenir des intervalles de confiance plus fins qu'avec la GEV. On note cependant qu'en fonction du niveau de dépendance, le régime de normalité asymptotique peut ne pas être atteint malgré l'estimation conjointe. Ainsi l'intervalle de confiance au niveau nominal 95% peut contenir le niveau de retour réel avec une probabilité inférieure à 95%.

CHAPITRE 6 MODÉLISATION DE L'ERREUR DE SPÉCIFICATION

On propose ici de prendre en compte les problèmes de spécification du modèle en incluant l'hypothèse d'une mauvaise spécification dans l'estimation des incertitudes. Pour cela, on va s'appuyer sur les méthodes de vraisemblances composites dont un exposé est proposé par Varin *et al.* (2011). Quelques exemples sont proposés dans Joe (2014) pour des données présentant de la dépendance. Il s'agit de vraisemblances qui sont impropres car elles ne sont pas toujours associées à une densité normalisée. Leur utilisation repose sur la sommation des log-vraisemblances de vecteurs de dimensions inférieures à celle du problème complet. On se penchera sur le cas d'une modélisation par marginales unidimensionnelles et par marginales bidimensionnelles.

6.1 Vraisemblance composite simple

On écrit la vraisemblance composite simple comme la somme des vraisemblances pour les marginales unidimensionnelles. Soit la log-vraisemblance composite :

$$\ell_{uni.}\{\boldsymbol{\theta} \mid (y_{i,d})_{i,d}\} = \sum_{d \in D} \sum_{i=1}^n \ln\{f_{(Y_{i,d}|\boldsymbol{\theta})}(y_{i_d})\}. \quad (6.1)$$

Il s'agit très exactement de la log-vraisemblance que l'on avait utilisée dans le chapitre 4 lorsque la dépendance dans les données n'avait pas encore été étudiée. Cependant, la manière d'envisager cette vraisemblance composite est différente car l'utilisateur sait qu'il s'agit d'une log-vraisemblance qui n'est pas une spécification correcte des données. Il s'agit de la réduction d'un problème avec une log-vraisemblance multidimensionnelle à une somme de problèmes avec des vraisemblance unidimensionnelles. On peut également envisager cette vraisemblance composite simple comme une modélisation des données avec la copule d'indépendance. Dans les deux cas, il s'agit d'une spécification incorrecte du modèle. En conséquence, les méthodes d'estimation d'incertitudes que l'on mettra en place seront différentes.

L'utilisation de cette log-vraisemblance composite implique donc que l'on ne modélise pas la dépendance dans les données. Cependant, l'objectif principal est ici de corriger les incertitudes pour les niveaux de retour. Or les niveaux de retour dépendent des paramètres des lois marginales et pas des paramètres de dépendance. Il n'est donc pas forcément nécessaire de modéliser celle-ci. Par ailleurs, la modélisation de la dépendance peut être faite séparément après avoir estimé les lois marginales.

6.2 Vraisemblance composite par paires

Un autre choix possible, que l'on envisage à titre de comparaison puisqu'il ne permet pas d'éviter de modéliser la dépendance, est d'utiliser la log-vraisemblance composite par paires. Pour le modèle étudié, celle-ci s'exprime sous la forme

$$\ell_{bi.}\{\boldsymbol{\psi} \mid (y_{i,d})_{i,d}\} = \sum_{i=1}^n \left[\sum_{r=1}^{m-1} \sum_{s=r+1}^m \ln\{f_{(Y_r, Y_s|\boldsymbol{\psi})}(y_{i,r}, y_{i,s})\} \right], \quad \boldsymbol{\psi} = (\boldsymbol{\theta}, \rho, \kappa), \quad (6.2)$$

où

m est le nombre de durées d'accumulation dans le modèle ;

r et s sont les indices associés aux durées d_r et d_s — la mesure de distance est toujours $h(d_r, d_s)$;

$\ln\{f_{(Y_r, Y_s|\boldsymbol{\psi})}(\cdot, \cdot)\}$ est la log-vraisemblance d'une distribution multidimensionnelle de dimension 2 dont les marginales sont d-GEV avec paramètre $\boldsymbol{\theta} = (\tilde{\mu}, \phi_0, \xi, \theta, \eta)$ et liées par une copule gaussienne avec covariance de Matérn de paramètres (ρ, κ) .

La log-vraisemblance de chaque paire (Y_r, Y_s) correspond donc à la vraisemblance du modèle complet (d-GEV et copule) mais pour des paires d'observations seulement. La vraisemblance composite ainsi construite est impropre car certaines variables unidimensionnelles Y_i sont comptabilisées plus d'une fois.

6.3 Estimation des incertitudes

La méthodologie d'estimation des incertitudes est la même pour les deux vraisemblances composites étudiées. On définit d'abord pour chaque méthode le score composite pour une observation $\mathbf{y} = (y_{d_1}, \dots, y_{d_m})$ par $u_{uni.}(\boldsymbol{\theta}, \mathbf{y}) = \nabla_{\boldsymbol{\theta}} \ell_{uni.}(\boldsymbol{\theta}, \mathbf{y})$ pour l'une ou par $u_{bi.}(\boldsymbol{\psi}, \mathbf{y}) = \nabla_{\boldsymbol{\psi}} \ell_{bi.}(\boldsymbol{\psi}, \mathbf{y})$. Finalement, on définit pour chacune des vraisemblances la matrice de sensibilité empirique H_n , qui n'est autre que la matrice d'information de Fisher observée associée à la log-vraisemblance composite. On définit également la matrice de variabilité empirique J_n . Par exemple, pour la vraisemblance composite par paires, ces matrices sont données par

$$H_n(\boldsymbol{\psi}) = -\frac{1}{n} \sum_{i=1}^n \nabla_{\boldsymbol{\psi}} u(\boldsymbol{\psi}, \mathbf{y}_i), \quad J_n(\boldsymbol{\psi}) = \frac{1}{n} \sum_{i=1}^n u(\boldsymbol{\psi}, \mathbf{y}_i) u(\boldsymbol{\psi}, \mathbf{y}_i)^\top.$$

La matrice d'information de Godambe empirique pour la log-vraisemblance par paires est alors définie par

$$G_n(\boldsymbol{\psi}) = H_n(\boldsymbol{\psi})J_n(\boldsymbol{\psi})^{-1}H_n(\boldsymbol{\psi}).$$

Si on note $\hat{\boldsymbol{\psi}}$ l'estimateur empirique de $\boldsymbol{\psi}$ obtenu en maximisant la log-vraisemblance composite par paires, on a, quand $n \rightarrow \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}) \rightsquigarrow \mathcal{N}[\mathbf{0}, G_n(\hat{\boldsymbol{\psi}})^{-1}].$$

La procédure est exactement similaire pour la log-vraisemblance composite simple avec le vecteur de paramètres $\boldsymbol{\theta}$. Varin *et al.* (2011) mettent cependant en garde les utilisateurs que le régime de normalité asymptotique peut être long à atteindre. Il est ainsi préférable que $\dim(\boldsymbol{\psi}) = o(n)$. Dans notre cas on a $\dim(\boldsymbol{\psi}) = 7$ et $n = 76$. Cela peut être limitant pour considérer que le régime asymptotique soit atteint. Pour la vraisemblance composite simple, on a $\dim(\boldsymbol{\theta}) = 5$, ce qui peut considérablement accélérer la convergence. Cela sera vérifié par une étude de simulation. Des méthodes de ré-échantillonnage, comme le jackknife ou le bootstrap, peuvent également être employées pour estimer la variance des paramètres.

6.4 Étude de simulation : Méthodes par vraisemblance composite

On présente ici une étude de simulation menée pour tester la couverture des intervalles de confiance au niveau nominal 95% obtenus en utilisant les méthodes par vraisemblance composites et la matrice de Godambe.

6.4.1 Génération des données

Les données sont générées selon le modèle d-GEV avec copule gaussienne et matrice de covariance de Matérn. Les paramètres utilisés pour générer les données sont

$$(\tilde{\mu}, \phi_0, \xi, \theta, \eta, \rho, \kappa) = (3.5, 1.75, 0.05, 0.07, 0.75, 2., 1.).$$

Ces paramètres sont choisis parce qu'ils sont proches de ceux estimés avec les données de l'aéroport Trudeau. Les durées choisies sont les mêmes que pour les données réelles. On utilisera $n = 75$ et $n = 200$ années de données simulées. On pourra ainsi déterminer les limitations des méthodes composites en termes de normalité asymptotique. On procède à $N = 4000$ itérations pour chaque méthode et chaque valeur de n .

6.4.2 Présentation des résultats

On a procédé à l'estimation des niveaux de retour 100 ans et calculé le taux de recouvrement et la taille moyenne des intervalles de confiance au niveau nominal 95% associés. Les résultats sont présentés dans les tableaux 6.1, 6.2, ainsi qu'à la figure 6.1. Les résultats obtenus avec la vraisemblance du modèle exact sont aussi indiqués à titre de comparaison. La durée pour mener l'étude varie de quelques minutes pour la méthode simple à plusieurs heures pour la méthode utilisant la vraisemblance composite par paires.

TABLEAU 6.1 Résultats de l'étude de simulation pour les méthodes avec vraisemblance composite. $n = 75$

Durée	Composite Simple		Composite Paire		d-GEV/Copule	
	Couverture	$\bar{\sigma}_q$	Couverture	$\bar{\sigma}_q$	Couverture	$\bar{\sigma}_q$
5 minutes	90.9%	14.54	91.3%	13.7	93.2%	12.82
10 minutes	90.1%	10.01	90.3%	9.48	93.1%	9.00
15 minutes	89.9%	7.97	90.2%	7.56	92.8%	7.22
30 minutes	89.8%	5.30	90.4%	4.94	92.8%	4.72
1 heure	89.6%	3.22	90.4%	3.06	92.7%	2.92
2 heures	89.8%	1.94	90.2%	1.83	92.8%	1.75
6 heures	89.9%	0.86	90.4%	0.81	93.0%	0.77
12 heures	90.4%	0.52	90.8%	0.50	93.0%	0.47
24 heures	90.7%	0.33	90.7%	0.31	93.0%	0.29
Moyenne	90.1%	N/A	90.5%	N/A	92.9%	N/A

TABLEAU 6.2 Résultats de l'étude de simulation pour les méthodes avec vraisemblance composite. $n = 200$

Durée	Composite Simple		Composite Paire		d-GEV/Copule	
	Couverture	$\bar{\sigma}_q$	Couverture	$\bar{\sigma}_q$	Couverture	$\bar{\sigma}_q$
5 minutes	92.6%	9.14	93.1%	8.64	94.4%	7.84
10 minutes	92.0%	6.33	92.2%	6.00	93.9%	5.51
15 minutes	92.2%	5.04	92.4%	4.79	93.9%	4.42
30 minutes	91.9%	3.29	92.4%	3.13	93.9%	2.89
1 heure	91.9%	2.04	92.0%	1.94	93.9%	1.79
2 heures	91.9%	1.23	92.2%	1.16	94.2%	1.07
6 heures	92.1%	0.54	92.4%	0.51	93.9%	0.47
12 heures	92.7%	0.33	92.7%	0.31	93.9%	0.29
24 heures	93.1%	0.21	93.0%	0.19	94.0%	0.18
Moyenne	92.2%	N/A	92.5%	N/A	94.0%	N/A

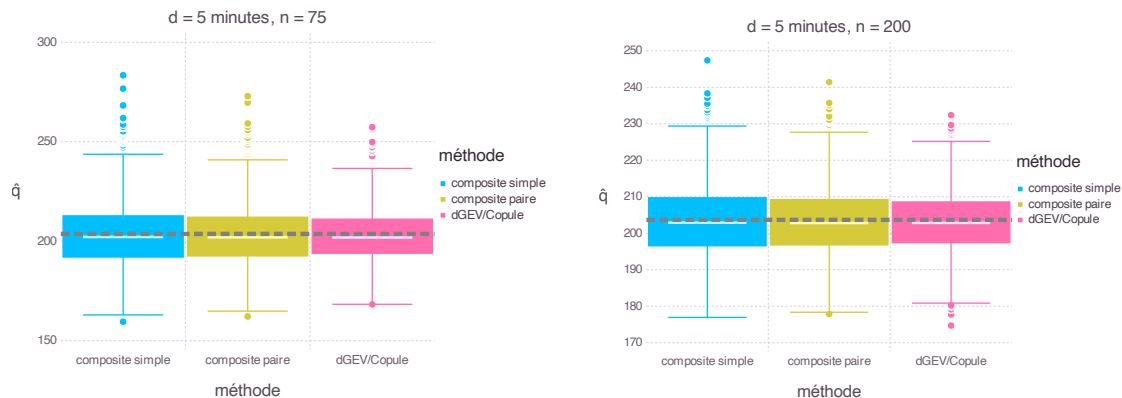


FIGURE 6.1 Comparaison des résultats de l'étude de simulation pour les méthodes composites et la méthode d-GEV avec copule

6.4.3 Analyse des résultats

On constate dans le tableau 6.1 que malgré un écart-type estimé en moyenne plus grand pour les niveaux de retour associés à la période de retour $T = 100$ ans, la méthode utilisant une vraisemblance composite simple a un taux de couverture des intervalles de confiance estimé au niveau nominal 95% plus faible qu'avec le modèle d-GEV avec copule. Par ailleurs, la méthode utilisant une vraisemblance composite par paires a un taux de recouvrement plus élevé que la méthode composite simple malgré un écart-type en moyenne plus faible. La figure 6.1 montre que la dispersion des valeurs estimées pour le niveau de retour $T = 100$ ans autour de la vraie valeur est plus faible lorsque l'on utilise le modèle complet plutôt que les modèles avec vraisemblances composites. La méthode avec vraisemblance composite simple présente la plus grande dispersion ; c'est donc la moins précise. Pour les trois méthodes, il n'y a pas de biais systématique.

Cette plus grande dispersion permet d'expliquer que malgré un écart-type estimé en moyenne plus petit, le taux de recouvrement est plus proche du niveau nominal souhaité pour la méthode utilisant la vraisemblance du modèle complet. Par ailleurs, cette observation est conforme aux résultats que nous avons obtenus dans l'étude de simulation de la section 5.2 qui montraient également que la dispersion par rapport au vrai niveau de retour était plus grande pour la d-GEV par rapport à la d-GEV avec copule. On constate dans le tableau 6.2 que le taux de recouvrement des intervalles de confiance estimés au niveau nominal 95% se rapproche du taux nominal lorsque n augmente. Cependant les observations faites précédemment, tant en terme de taux de couverture que d'écart-type, restent valables.

Malgré ces résultats, qui incitent à privilégier l'utilisation du modèle complet, on constate que l'utilisation de la matrice d'information de Godambe a permis d'améliorer le taux de couverture pour la vraisemblance d-GEV. En effet, les taux de recouvrement sont meilleurs que ceux qui avaient été observés dans l'étude de simulation menée à la section 5.2. Des résultats montrés dans l'annexe B montrent par ailleurs que le taux de couverture ne varie pas significativement lorsque le niveau de dépendance augmente. Les incertitudes s'adaptent alors à la baisse d'information entraînée par la dépendance dans les données. Ces résultats indiquent que l'utilisation de la matrice d'information de Godambe à la place de la matrice d'information de Fisher permet de corriger partiellement les limitations de la méthode d-GEV pour l'estimation des incertitudes et ce, sans avoir à modéliser la corrélation inter-durée. Les niveaux de couverture atteints sont acceptables.

Ce résultat est important car, d'une part, la modélisation de la structure de dépendance est complexe et, d'autre part, comme on l'a rappelé dans la conclusion en 5.1.7, le modèle proposé est trop peu flexible pour obtenir une estimation convenable des incertitudes lorsque l'on utilise les données réelles. On peut donc s'interroger sur la capacité de la vraisemblance composite simple à estimer convenablement les incertitudes. On note également que la méthode utilisant une vraisemblance composite par paires est plus performante lorsque l'on utilise des données simulées. Cependant, elle nécessite de modéliser la dépendance et reste moins performante que le modèle complet.

6.5 Résultats de l'application de la méthode composite simple aux données réelles

Les résultats de l'étude de simulation de la section précédente ont montré que la méthode utilisant une vraisemblance composite simple permet de corriger en partie les limitations de la méthodologie d-GEV pour l'estimation des incertitudes. De plus, elle ne nécessite pas une étude approfondie de la dépendance inter-durée présente dans les données. Par ailleurs, bien que nous ayons identifié que la copule gaussienne était un candidat acceptable pour modéliser la dépendance inter-durées entre les maxima d'intensité de précipitation, il est possible que celle-ci ne capture pas parfaitement les effets physiques complexes qui entraînent cette dépendance. Pour ces raisons, il peut être souhaitable d'appliquer la méthode d-GEV et d'estimer les incertitudes à l'aide de la matrice d'information de Godambe. Pour un nombre important d'utilisateurs de courbes IDF, cette méthodologie est plus accessible que la méthodologie reposant sur la modélisation de la dépendance.

Nous allons donc appliquer la méthodologie de la vraisemblance composite simple pour les données de l'aéroport Trudeau de Montréal. On maximise la log-vraisemblance composite

simple et on applique la méthode Delta en remplaçant la matrice d'information de Fisher par la matrice d'information de Godambe. Les résultats sont présentés à la figure 6.2 sous formes de diagrammes quantile-quantile. Les intervalles de confiance ainsi obtenus sont superposés à ceux que nous avons obtenus pour la méthode d-GEV sans prendre en compte la dépendance. Les points du diagrammes sont les mêmes pour les deux méthodes par construction.

On constate que les intervalles de confiance estimés au niveau nominal 95% sont plus larges que ceux estimés par la méthode d-GEV sans l'être autant que ceux obtenus par la méthodologie d'estimation indépendante GEV. Mis à part pour la durée $d = 30$ minutes, qui n'est pas encore adéquatement recouverte, l'ajustement des intervalles de confiance semble convenable.

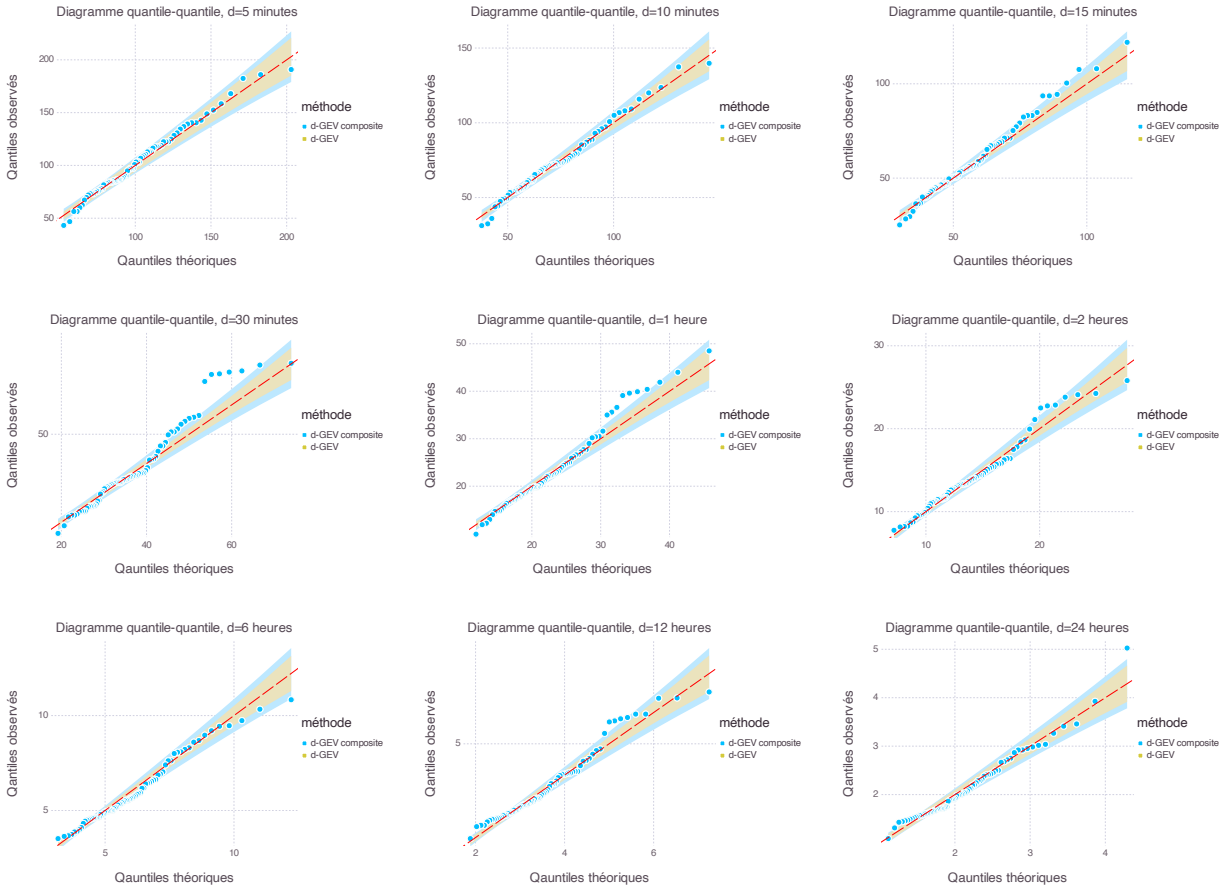


FIGURE 6.2 Diagrammes Quantile-Quantile et intervalles de confiance pour la méthodologie d-GEV composite et classique

CHAPITRE 7 ESTIMATION DES COURBES IDF DES PRÉCIPITATIONS

7.1 Choix de la méthode

Nous avons étudié deux approches pour estimer les incertitudes associées à l'estimation des niveaux de retour d'intensité de précipitation avec le modèle d-GEV :

- (i) La première méthode envisagée consiste à identifier une copule pour modéliser la dépendance inter-durée des maxima d'intensité de précipitation afin d'établir une vraisemblance propre dont on puisse exploiter les propriétés de normalité asymptotique.
- (ii) La seconde méthode envisagée est d'utiliser une vraisemblance composite impropre et d'adapter la méthodologie d'estimation des incertitudes en utilisant la matrice d'information de Godambe.

Ces deux méthodes présentent des avantages et des inconvénients qui sont énumérés dans ce qui suit.

7.1.1 Avantage des deux méthodes par rapport à l'estimation indépendante des lois marginales GEV

Ces deux méthodes exploitent la relation d'échelle qui lie les paramètres des lois GEV marginales. La contrainte physique de décroissance des niveaux de retour d'intensité de précipitation avec la durée d'accumulation est donc respectée par construction, ce qui n'est pas le cas lorsque les lois marginales sont estimées indépendamment.

Par ailleurs, dans ces deux modélisations, on prend en compte la relation d'échelle qui apparaît dans les courbes IDF en amont de l'estimation. Cela permet de combiner les données dans un modèle unique afin de réduire les incertitudes d'estimation. On note que les incertitudes ainsi obtenues sont toujours plus faibles que pour l'estimation indépendante des marginales GEV. La vérification de l'adéquation de ces incertitudes avec les données est donc nécessaire.

7.1.2 Méthodologie avec copule

La méthodologie avec copule est la plus complète et celle qui présente les meilleurs résultats d'un point de vue théorique. Toutefois, elle est plus sensible aux erreurs de spécification des

lois marginales et de la dépendance.

Du point de vue pratique, son utilisation nécessite une étude approfondie des données afin d'identifier une copule adaptée. Ce mémoire a établi qu'une copule gaussienne, associée à la fonction de covariance de Matérn avec une mesure de distance logarithmique entre les durées, est adaptée pour modéliser la dépendance pour plusieurs stations de l'est du Canada. Cependant, l'utilisateur précautionneux qui souhaiterait établir un résultat similaire devrait mener à nouveau une étude équivalente pour ses propres données. Il est important de souligner également que la modélisation des données à l'aide de copules apporte un degré de complexité supplémentaire qui n'est pas forcément envisageable pour certains utilisateurs des courbes IDF non-spécialisés en statistique.

Du point de vue théorique, cette stratégie est la meilleure car les études de simulation montrent qu'identifier la copule qui caractérise la dépendance permet d'atteindre une efficacité asymptotique accrue par rapport à la méthode d-GEV ou à la méthode d'estimation indépendante des lois marginales GEV. Par ailleurs, bien que les autres modèles ne soient pas biaisés marginalement, la maximisation de la vraisemblance du modèle complet avec copule engendre moins de dispersion dans l'estimation des niveaux de retour.

Les résultats pour les données de l'aéroport Trudeau présentés à la figure 5.4 montrent cependant que le taux de couverture des intervalles de confiance obtenus à l'aide de cette méthode reste trop faible par rapport au taux nominal de 95%. On suspecte que le modèle manque de flexibilité par rapport à la réalité physique sous-jacente et qu'il ne soit pas une modélisation exacte du processus à l'œuvre. En termes d'estimation ponctuelle, le modèle présente des résultats qui peuvent être utilisés pour produire des courbes IDF. Cependant, on déconseille d'utiliser ce modèle pour estimer les incertitudes car le modèle s'éloigne trop de la réalité pour que les propriétés de normalité asymptotique soient vérifiées.

7.1.3 Méthodologie avec une vraisemblance composite

La méthodologie avec vraisemblance composite permet de corriger en partie les limitations de la méthodologie d-GEV en prenant en compte l'erreur de spécification de la dépendance dans l'estimation des incertitudes. Son utilisation ne requiert pas d'étudier la dépendance dans les données, ce qui rend cette approche accessible à un plus grand nombre d'utilisateurs. Les incertitudes estimées sont plus conservatrices qu'avec les autres méthodes. Les résultats théoriques obtenus dans les études de simulation montrent de moins bonnes performances en terme d'estimation des niveaux de retour et des incertitudes associées que la méthodologie basée sur la vraisemblance du modèle complet. Cependant, en prenant en compte le fait qu'il n'est nécessaire de faire aucune hypothèse sur la structure de dépendance, les résultats

obtenus sont satisfaisants.

Les résultats qualitatifs pour les données de l'aéroport Trudeau présentés à la figure 6.2 montrent que les incertitudes estimées avec cette méthode recouvrent adéquatement les observations pour presque toutes les durées d'accumulation.

7.2 Estimation des courbes IDF pour la station de Montréal Trudeau

On peut, en considération des remarques précédentes, proposer d'utiliser la méthodologie avec vraisemblance composite simple pour produire les courbes IDF de la station de l'aéroport Trudeau à Montréal en incorporant les incertitudes estimées au niveau nominal 95%. Ces courbes sont montrées pour trois périodes de retour à la figure 7.1. On pourrait aisément inclure les courbes pour d'autres périodes de retour. L'extrapolation à des durées d'accumulation en deçà de 5 minutes et au-delà de 24 heures doit être sujette à caution. À cause de l'échelle logarithmique utilisée en abscisse et en ordonnée, la largeur des bandes d'incertitudes semble ne pas varier significativement avec la durée d'accumulation pour une période de retour donnée.

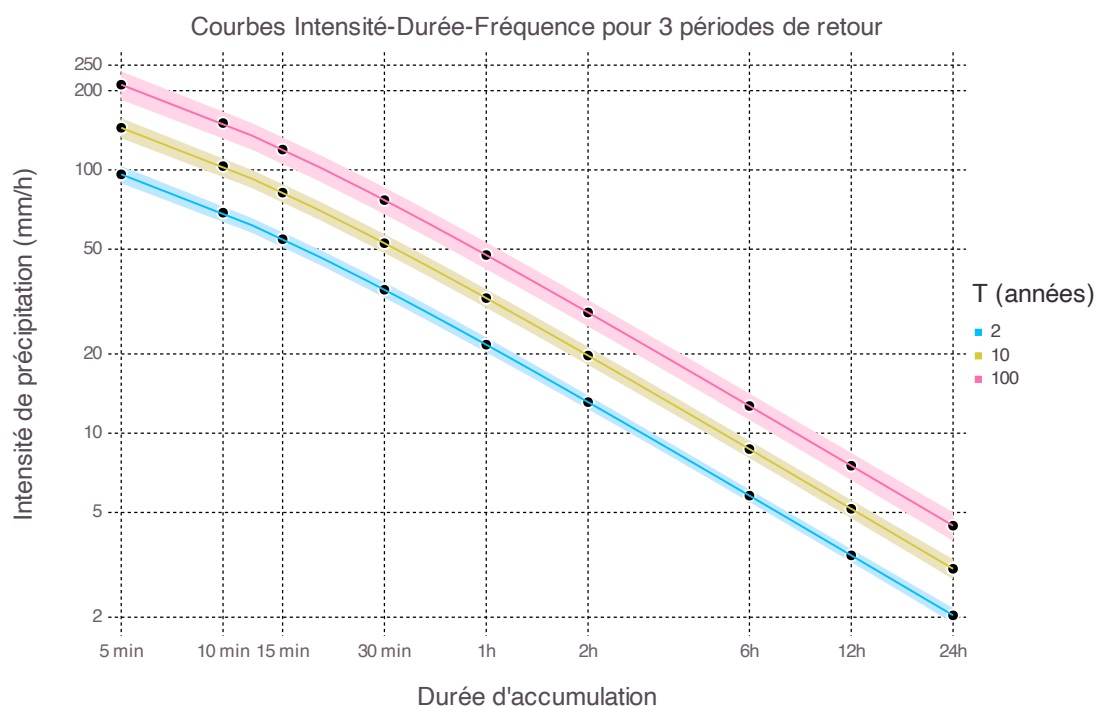


FIGURE 7.1 Courbes IDF obtenues avec la méthodologie de la vraisemblance composite simple

Cependant, il s'agit bien d'un effet de perspective lié à l'échelle logarithmique. Par exemple, pour la période de retour $T = 100$ ans, on a, au niveau nominal 95% :

$$\hat{q}_{5\text{min.}} = 211 \pm 26 \text{ mm/h pour la durée d'accumulation } d = 5 \text{ minutes ;}$$

$$\hat{q}_{24\text{h.}} = 4.45 \pm 0.55 \text{ mm/h pour la durée d'accumulation } d = 24 \text{ heures.}$$

Ces résultats peuvent être comparés à ceux que nous aurions obtenus en ajustant des lois GEV indépendamment pour chaque durée d'accumulation. Pour la période de retour $T = 100$ ans nous aurions ainsi eu, toujours au niveau nominal 95% :

$$\hat{q}_{5\text{min.}} = 195 \pm 36 \text{ mm/h pour la durée d'accumulation } d = 5 \text{ minutes ;}$$

$$\hat{q}_{24\text{h.}} = 4.60 \pm 1.28 \text{ mm/h pour la durée d'accumulation } d = 24 \text{ heures.}$$

On constate que les incertitudes de la méthodologie avec vraisemblance composite simple sont réduites par rapport à celles de l'estimation indépendante des lois GEV marginales. Si l'on juge que le modèle et la méthodologie sont convenables, ce qui à la vue des diagrammes quantile-quantile montrés à la figure 6.2 est une hypothèse plausible, cela signifie que notre méthodologie nous a permis de combiner les données d'intensité de précipitation de manière à obtenir des estimations plus précises des niveaux de retour.

CHAPITRE 8 CONCLUSION ET RECOMMANDATIONS

8.1 Synthèse des travaux

L'objectif de ce projet de maîtrise était de tester l'aptitude de la modélisation d-GEV à décrire adéquatement la distribution des maxima annuels d'intensité de précipitation pour des durées d'accumulation variant de 5 minutes à 24 heures.

On a d'abord établi que le modèle d-GEV s'ajustait convenablement aux données de précipitation à l'aide des données recueillies à l'aéroport Trudeau de Montréal. L'ajustement est qualitativement comparable à celui obtenu avec des lois GEV estimées indépendamment. La d-GEV permet ainsi d'inclure, en amont de l'estimation, la dépendance fonctionnelle des niveaux de retour en fonction de la durée d'accumulation. Cependant, la vraisemblance utilisée est impropre car elle ne tient pas compte de la dépendance inter-durée entre les maxima d'intensité de précipitation. Il en résulte une sous-estimation des incertitudes liées au modèle. On a envisagé deux méthodes de résolution à cette limitation du modèle d-GEV.

La première solution testée consistait à modéliser la corrélation inter-durée afin d'aboutir à une vraisemblance propre. Cette vraisemblance propre peut alors être utilisée pour obtenir des incertitudes à l'aide des propriétés de normalité asymptotique. Une étude de la corrélation présente dans les données a permis d'établir que le degré de corrélation était fonction de la distance entre les logarithmes des durées d'accumulation. En nous inspirant des résultats d'autres travaux, on a montré que la fonction de covariance de Matérn pouvait être utilisée pour décrire la décroissance de la corrélation avec la distance choisie. Ce résultat n'avait pas, à notre connaissance, été établi précédemment dans le contexte du modèle d-GEV. On a ensuite montré que la dépendance pouvait être modélisée à l'aide d'une copule gaussienne dont les termes de la matrice de corrélation sont écrits avec la fonction de covariance de Matérn. Pour chaque paire de maxima, les termes de la matrice de covariance peuvent s'écrire en fonction du sinus du tau de Kendall. Cela a permis de vérifier l'ajustement du modèle au tau de Kendall empirique. Malgré un ajustement ponctuel acceptable, on montre que le modèle proposé n'est pas une représentation assez fidèle de la réalité pour que les incertitudes estimées soient exploitables. La modélisation de la dépendance reste cependant une contribution utile, qui peut être exploitée dans d'autres travaux.

La seconde solution testée est l'utilisation de méthodes robustes d'estimation des incertitudes construite à partir de la matrice d'information de Godambe. La méthodologie d'estimation des incertitudes tient alors compte du fait que la vraisemblance utilisée est impropre. Cette

méthodologie est particulièrement attractive pour l'utilisateur souhaitant éviter de modéliser la dépendance. On a d'abord montré dans une étude de simulation que les méthodes robustes permettent de corriger de façon satisfaisante les incertitudes estimées à l'aide de la vraisemblance du modèle d-GEV, bien que le modèle soit théoriquement moins efficient que le modèle avec copule. L'utilisation de la méthodologie robuste sur les données de l'aéroport Trudeau montre un niveau d'ajustement convenable des incertitudes estimées. La méthode robuste d'estimation des incertitudes peut ainsi être une alternative facile à mettre en place pour obtenir une estimation des incertitudes.

8.2 Limites des modèles et améliorations envisagées

Le modèle d-GEV est une modélisation dont on ne peut négliger les atouts. Cependant, la relation fonctionnelle peut manquer de flexibilité et il serait intéressant d'étudier la possibilité de rendre le modèle plus flexible tout en exploitant la relation d'échelle. On note aussi que la contrainte physique qui impose une relation d'ordre pour les maxima d'intensité de précipitation annuels en fonction de la durée d'accumulation n'est pas explicitement imposée dans la vraisemblance utilisée. Cette relation d'ordre pourrait être imposée à l'aide d'une variable indicatrice, mais la vraisemblance résultante ne serait alors plus normalisée. Des méthodes d'échantillonnage bayésiennes pourraient permettre d'estimer malgré tout les paramètres du modèle sans avoir à calculer la constante de renormalisation de la vraisemblance.

Il faut également noter que nous avons limité notre stratégie d'identification d'une copule adaptée aux familles paramétriques les plus courantes. Ce choix reposait sur une volonté d'avoir un modèle interprétable avec un degré de complexité raisonnable. Il est toutefois possible d'envisager d'autres familles de copules qui pourraient améliorer la qualité de l'ajustement. Les copules non paramétriques pourraient également représenter une alternative à la modélisation paramétrique.

La stratégie d'estimation des incertitudes par des méthodes robustes est prometteuse car elle permet d'éviter de modéliser la dépendance inter-durée. Cependant, l'estimation de la matrice d'information de Godambe peut être instable lorsque la quantité de données disponibles est trop faible. Or, nous avons seulement testé ce modèle pour les stations avec le plus de données et nous n'avons pas étudié son comportement sur des échantillons plus petits. Des méthodes de ré-échantillonnage pourraient ainsi mieux convenir pour estimer les incertitudes. Cela pourrait être l'objet de travaux futurs.

RÉFÉRENCES

- BLANCHET, J., CERESSETTI, D., MOLINIÉ, G. et CREUTIN, J.-D. (2016). A regional GEV scale-invariant framework for intensity–duration–frequency analysis. *Journal of Hydrology*, 540:82–95.
- COLES, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer London.
- CRESSIE, N. et WIKLE, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley-Blackwell, Chichester, England.
- DEMARTA, S. et MCNEIL, A. J. (2005). The t copula and related copulas. *International Statistical Review / Revue Internationale de Statistique*, 73(1):111–129.
- GEOGA, C. J., MARIN, O., SCHANEN, M. et STEIN, M. L. (2022). Fitting Matérn smoothness parameters using automatic differentiation. *Statistics and Computing*.
- HEINEN, A. et VALDESOGO, A. (2020). Spearman rank correlation of the bivariate Student t and scale mixtures of normal distributions. *Journal of Multivariate Analysis*, 179:104650.
- HOGG, W. D. (1989). *Rainfall intensity-duration frequency values for Canadian locations*. CLI; 1-89. Environment Canada, Atmospheric Environment Service, Downsview, ON.
- JALBERT, J., GENEST, C. et PERREAULT, L. (2022). Interpolation of precipitation extremes on a large domain toward IDF curve construction at unmonitored locations. *Journal of Agricultural, Biological and Environmental Statistics*, 27(3):461–486.
- JOE, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2):401–419.
- JOE, H. (2014). *Dependence modeling with copulas*. CRC Press.
- JURADO, O. E., ULRICH, J., SCHEIBEL, M. et RUST, H. W. (2020). Evaluating the performance of a max-stable process for estimating intensity-duration-frequency curves. *Water*, 12(12):3314.
- KOUTSOYIANNIS, D. (2003). *On the appropriateness of the Gumbel distribution in modeling extreme rainfall*, pages 303–319. Unpublished.
- KOUTSOYIANNIS, D., KOZONIS, D. et MANETAS, A. (1998). A mathematical framework for studying rainfall intensity-duration-frequency relationships. *Journal of Hydrology*, 206(1-2):118–135.
- PAPALEXIOU, S. M. et KOUTSOYIANNIS, D. (2013). Battle of extreme value distributions : A global survey on extreme daily rainfall. *Water Resources Research*, 49(1):187–201.

- ULRICH, J., JURADO, O. E., PETER, M., SCHEIBEL, M. et RUST, H. W. (2020). Estimating IDF curves consistently over durations with spatial covariates. *Water*, 12(11):3119.
- VAN DE VYVER, H. et VAN DEN BERGH, J. (2018). The Gaussian copula model for the joint deficit index for droughts. *Journal of Hydrology*, 561:987–999.
- VARIN, C., REID, N. et FIRTH, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42.

ANNEXE A RÉSULTATS DE L'ÉTUDE DE SIMULATION POUR DES JEUX DE DONNÉES PLUS GRANDS

On consigne ici les résultats de l'étude de simulation présentée à la section 5.2 lorsque la taille d'échantillon est $n = 500$.

TABLEAU A.1 Taux de couverture des intervalles de confiance estimés au niveau nominal 95% et écart-type moyen estimé avec la méthode Delta pour $\rho = 3$ et $\kappa = 1$

Modèle	$d = 5$ minutes		$d = 2$ heures		$d = 24$ heures	
	Couverture	$\bar{\sigma}_{q_{5min}}$	Couverture	$\bar{\sigma}_{q_{2h}}$	Couverture	$\bar{\sigma}_{q_{24h}}$
d-GEV	71.5%	3.78	66.9%	0.445	67%	0.074
d-GEV/Copule	94.5%	5.64	94.6%	0.72	95.6%	0.115
GEV	93.5%	10.30	94.8%	1.32	94.7%	0.208

TABLEAU A.2 Taux de couverture des intervalles de confiance estimés au niveau nominal 95% et écart-type moyen estimé avec la méthode Delta pour $\rho = 1$ et $\kappa = 0.5$

Modèle	$d = 5$ minutes		$d = 2$ heures		$d = 24$ heures	
	Couverture	$\bar{\sigma}_{q_{5min}}$	Couverture	$\bar{\sigma}_{q_{2h}}$	Couverture	$\bar{\sigma}_{q_{24h}}$
d-GEV	88.8%	3.77	88.1%	0.44	89.8%	0.073
d-GEV/Copule	95.3%	4.34	95.8%	0.53	94.4%	0.087
GEV	95.3%	10.21	94.1%	1.31	95.6%	0.208

ANNEXE B RÉSULTATS COMPLÉMENTAIRES POUR L'ÉTUDE DE SIMULATION PORTANT SUR LES MÉTHODES COMPOSITE

Les paramètres d-GEV utilisés sont les mêmes que dans l'étude de simulation 6.4. Les paramètres de dépendance sont :

$$\rho, \kappa = 3, 1.5$$

TABLEAU B.1 Résultats de l'étude de simulation pour la méthode composite simple avec un niveau plus élevé de dépendance. $n = 75$.

Durée	Composite Simple	
	Couverture	$\bar{\sigma}_q$
5 minutes	89.1%	16.75
10 minutes	88.7%	11.8
15 minutes	88.5%	9.44
30 minutes	88.6%	6.13
1 heure	88.6%	3.80
2 heures	88.8%	2.30
6 heures	88.9%	1.03
12 heures	89.0%	0.62
24 heures	89.5%	0.38
Moyenne	88.8%	N/A