



Titre: Estimation et interpolation des courbes d'intensité-durée-fréquence
des précipitations en Colombie-Britannique

Auteur: Léo Guillon
Author:

Date: 2023

Type: Mémoire ou thèse / Dissertation or Thesis

Référence: Guillon, L. (2023). Estimation et interpolation des courbes d'intensité-durée-
Citation: fréquence des précipitations en Colombie-Britannique [Mémoire de maîtrise,
Polytechnique Montréal]. PolyPublie. <https://publications.polymtl.ca/54867/>

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

URL de PolyPublie: <https://publications.polymtl.ca/54867/>
PolyPublie URL:

**Directeurs de
recherche:** Jonathan Jalbert
Advisors:

Programme: Maîtrise recherche en mathématiques appliquées
Program:

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

**Estimation et interpolation des courbes d'intensité-durée-fréquence des
précipitations en Colombie-Britannique**

LÉO GUILLON

Département de mathématiques et de génie industriel

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

Août 2023

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

**Estimation et interpolation des courbes d'intensité-durée-fréquence des
précipitations en Colombie-Britannique**

présenté par **Léo GUILLON**

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

Luc ADJENGUE, président

Jonathan JALBERT, membre et directeur de recherche

Elmira HASSANZADEH, membre

DÉDICACE

Aux produits scalaires.

REMERCIEMENTS

Ce mémoire est le fruit d'un travail de longue haleine, qui n'aurait pas été possible sans la présence et le soutien d'un certain nombre de personnes que je souhaite remercier.

Je tiens en premier lieu à remercier chaleureusement mon directeur de recherche Jonathan JALBERT de m'avoir accordé sa confiance et sa patience de m'aider à mener à bien ce projet et d'avancer tant bien que mal, et ce malgré les imprévus que mon projet a subi.

Je souhaite également remercier plus largement à remercier l'équipe Extremes de Polytechnique Montréal : Gabriel & Duy Anh pour leur expertise et conseils, mais aussi Paul, et les nouveaux : Auguste, Takumi, Anne, ...

Dans un cadre formel, je tiens également à remercier les institutions sans lesquels mon projet n'aurait pas pu aboutir : Environnement et Changement Climatique Canada, le projet CaSPAr pour avoir fourni ses données précieuses ; Jean GUÉRIN pour m'avoir permis de me donner le goût de l'enseignement et de la découverte ; Polytechnique Montréal et l'École Polytechnique plus généralement pour avoir permis de financer ce projet du début jusqu'à la fin ; et enfin Mme. Elmira HASSANZADEH et M. Luc ADJENGUE de faire partie de mon jury d'évaluation.

Un grand merci aussi aux multiples personnes qui m'auront permis de profiter à fond de l'expérience Montréalaise : Omar, Gabriel & Pierre, la coloc' Garnier, l'équipe PolyRad, les Magical People, et tant d'autres que j'oublie sûrement..

Je tenais enfin à remercier ma famille, qui malgré la distance d'un océan entier nous séparant, a su rester présente à mes côtés.

Merci à tous !

RÉSUMÉ

Les événements de précipitations extrêmes représentent un enjeu d’une importance croissante, dans un contexte de changement climatique. Les risques de crues et d’inondations augmentent d’année en année, à mesure que le climat se dérègle. Le Canada, de part sa grande superficie et ses importantes réserves en eau douce dont dépendent notamment la production hydro-électrique, se voit donc de plus en plus exposé à ces aléas hydriques. L’étendue du pays, sa faible densité de population et sa disparité d’infrastructures par ailleurs font qu’il est plus difficile d’estimer ces précipitations dans des zones moins peuplées, dépourvues de relevés. Les courbes d’intensité-durée-fréquence (IDF) sont un outil permettant aux ingénieurs de dimensionner les infrastructures civiles. Actuellement, ce sont les courbes d’Environnement Changement Climatique Canada (ECCC) qui sont utilisées par les ingénieurs, mais elles ne sont pour l’instant que calculées en leur stations de relevés météorologiques et présentes plusieurs hypothèses simplificatrices.

L’objectif de ce projet est donc d’une part de proposer un modèle d’estimation spatiale de ces courbes IDF, et d’autre part de se focaliser sur la Colombie-Britannique. Pour rendre compte des particularités de cette région, plusieurs variables environnementales sont incorporées simultanément dans l’estimation spatiale des précipitations extrêmes, avec notamment comme fer de lance d’une part une réanalyse des précipitations journalières, et d’autre part l’élévation par rapport au niveau de la mer, pour traduire les disparités de relief à travers la côte ouest canadienne. Ensuite, pour rendre compte de la dépendance spatiale des extrêmes, un modèle bayésien hiérarchique combiné à un champ markovien gaussien intrinsèque (iGMRF) d’ordre 1 a été mis en oeuvre. Le calcul numérique à partir des données fournies d’une part par ECCC, et par une réanalyse simulée de *Canadian Surface Prediction Archive* (CaSPAr), appelée Système Régional de Représentation Déterministe (*Region Deterministic Reforecasting System*, RDRS), a été effectué à l’aide des méthodes Monte-Carlo par Chaînes de Markov (*Monte-Carlo Markov Chains*, MCMC), et ce sur différentes associations de variables environnementales.

Par validation croisée des résultats, on en déduit que l’altitude et la réanalyse RDRS permettent d’interpoler les courbes IDF sur tout le territoire, sont des variables significatives pour les précipitations extrêmes, et permettent ainsi de rendre compte plus efficacement des différences entre les zones côtières et les zones montagneuses de la Colombie-Britannique. Un dashboard interactif a également été mis en place, pour présenter à la fois les données et les résultats de manière intuitive.

ABSTRACT

Extreme rainfall events represent a growing matter issue, in our current context of climate change. Flood risks are getting worse from year to year, as forth the deregulation of climate is going on. Canada, because of its large superfcy and relying its hydroelectric production on massive fresh water reserves, therefore becomes more and more exposed to these water hazards. As it's a large country, and having a scarce population density outside of the American border and a lack of infrastructures makes it so the rainfalls are harder to estimate in less populated zones. Then, intensity-duration-frequency (IDF) curves are a tool used be engineers to size civilian infrastructures. Currently, Environment Climate Change Canada (ECCC) is the main source providing these curves. However, their current way of being computed still overshadow some simplifications, and are computed only in weather stations.

This project has the goal of introducing a spatial model in order to estimate IDF curves throughout the country, and also to focus on a province somehow left behind by the laboratory "Extremes" from Polytechnique Montreal : British Columbia. In order to acknowledge the features of this province in front of the Eastern Coast, several environmental covariates are integrated in the spatial model ; on one hand, a new reanalysis from the precipitations across British Columbia, and on the other hand, the elevation, allowing to introduce the influence of the Rocky Mountains on extreme rainfalls. Then, a combination of a Hierarchical Bayesian Model (HBM) with an intrisinc Gaussion Markov Random Field (iGMRF) of first order was put together to create the spatial model. The computation, yielded by both data provided by both ECCC and a reanalysis from Canadian Surface Prediction Archive (CaSPAr), named Regional Deterministic Reforecasting System (RDRS), was implemented with Markov Chains Monte Carlo (MCMC), with different combinations of covariates.

Following a cross-validation of results, we show the importance of elevation and the RDRS reanalysis for extreme rainfalls, allowing to interpolate IDF curves across the entire province, thus highlighting the differences between coastal zones and mountain zones of British Columbia. An interactive dahsboard was also created, to explain datasets and results in a intuitive way.

TABLE DES MATIÈRES

DÉDICACE	iii
REMERCIEMENTS	iv
RÉSUMÉ	v
ABSTRACT	vi
TABLE DES MATIÈRES	vii
LISTE DES TABLEAUX	ix
LISTE DES FIGURES	x
LISTE DES ANNEXES	xii
LISTE DES SIGLES ET ABRÉVIATIONS	xiii
CHAPITRE 1 INTRODUCTION	1
CHAPITRE 2 CADRE THÉORIQUE	5
2.1 Théorie des valeurs extrêmes	5
2.1.1 Théorème principal	5
2.1.2 Modèle des maxima par blocs et niveaux de retour	8
2.2 Inférence bayésienne et modèles hiérarchiques	9
2.2.1 Inférence bayésienne	9
2.2.2 Modèle bayésien hiérarchique	10
2.2.3 Méthode Monte-Carlo par Chaînes de Markov	12
2.3 Champs markoviens gaussiens aléatoires	13
2.3.1 Théorie de base	14
2.3.2 Champs intrinsèques	15
CHAPITRE 3 REVUE DE LITTÉRATURE	17
CHAPITRE 4 DONNÉES ET ANALYSE EXPLORATOIRE	22
4.1 Données des stations	22
4.1.1 Carte des stations d'ECCC	22

4.1.2	Séries temporelles et estimation GEV simple	24
4.1.3	Stations et valeurs aberrantes	27
4.2	Grille des données réanalysées de RDRS 2.1	30
4.2.1	Description des données	30
4.2.2	Extraction et cartes des simulations	30
4.2.3	Carte de l'élévation	32
4.3	Jonction des deux ensembles de données	34
CHAPITRE 5 MODÉLISATION		36
5.1	Niveau des données	36
5.2	Niveau latent	37
5.3	Niveau des hyperparamètres	39
5.4	Modèle complet	39
5.5	Implémentation d'un algorithme MCMC	40
CHAPITRE 6 RÉSULTATS		43
6.1	Modèles unidimensionnels	43
6.1.1	Modèle utilisant la moyenne comme variable explicative	43
6.1.2	Modèle utilisant le quantile comme variable explicative	44
6.1.3	Conclusions sur les modèles unidimensionnels	44
6.2	Modèles multidimensionnels	49
6.2.1	Modèle complet et variables pertinentes	49
6.2.2	Modèle partiel	50
6.2.3	Modèle adapté	52
6.3	Comparaison des modèles	56
6.3.1	Validation croisée	56
6.3.2	Résultats finaux et discussion	58
CHAPITRE 7 CONCLUSION ET RECOMMANDATIONS		61
7.1	Synthèse du projet	61
7.2	Limitations et ouvertures	62
RÉFÉRENCES		63
ANNEXES		67

LISTE DES TABLEAUX

Tableau 6.1	Tableau des résultats de la MSE après validation croisée	59
Tableau 6.2	Tableau des résultats de la CVM après validation croisée	59

LISTE DES FIGURES

Figure 1.1	Courbes IDF d'ECCC à la station Vancouver UBC	2
Figure 2.1	Comparaison des modèles de dépendance	11
Figure 4.1	Histogramme du nombre d'années de relevé pour les stations d'ECCC en Colombie-Britannique	23
Figure 4.2	Carte de l'ensemble des stations d'ECCC en Colombie Britannique et de leur altitude respective	23
Figure 4.3	Cartes des paramètres de l'estimation naïve issue des données des sta- tions d'ECCC de la Colombie Britannique	25
Figure 4.4	Graphes de relation entre l'élévation et les paramètres GEV des stations d'ECCC	26
Figure 4.5	Histogramme du paramètre de forme ξ estimé pour les stations d'ECCC en Colombie-Britannique	27
Figure 4.6	Graphes de diagnostic pour les stations d'ECCC ayant un $\xi > 1$. . .	28
Figure 4.7	Détails de l'aberration à l'île de Moresby	29
Figure 4.8	Zone couverte par la réanalyse RDRS 2.1	31
Figure 4.9	Cartes des variables environnementales obtenues grâce à la réanalyse RDRS 2.1	32
Figure 4.10	Carte de l'élévation sur la zone réanalysée par CaSPAr	33
Figure 4.11	Carte des quatre stations d'ECCC choisies arbitrairement pour la mo- délisation	35
Figure 5.1	Schéma du modèle bayésien hiérarchique	40
Figure 6.1	Cartes du quantile à $T = 20$ années de retour de la GEV estimée par les modèles unidimensionnels	45
Figure 6.2	Traces et intervalles de crédibilité empiriques à 95% pour les coefficients de régression des modèles unidimensionnels	46
Figure 6.3	Traces des stations témoins pour le modèle M_1	47
Figure 6.4	Traces des stations témoins pour le modèle M_2	48
Figure 6.5	Quantile à 20 années de retour pour le modèle complet M_3	49
Figure 6.6	Intervalles de crédibilité empiriques pour les coefficients de régression du modèle complet M_3	51
Figure 6.7	Quantile à 20 années de retour pour le modèle partiel M_4	52
Figure 6.8	Intervalles de confiance empiriques pour les coefficients de régression du modèle partiel M_4	53

Figure 6.9	Quantile à 20 années de retour pour le modèle adapté M_5	54
Figure 6.10	Intervalle de confiance empiriques pour les coefficients de régression du modèle adapté M_5	55
Figure 6.11	Schéma résumé de la validation croisée	57
Figure A.1	Traces et intervalles de crédibilité des stations d'intérêt pour le modèle M_3 (Modèle complet)	67
Figure A.2	Traces et intervalles de crédibilité des stations d'intérêt pour le modèle M_4 (Modèle partiel)	68
Figure A.3	Traces et intervalles de crédibilité des stations d'intérêt pour le modèle M_5 (Modèle adapté)	69
Figure B.1	Déroulé du procédé de construction des IDF schématisé.	70

LISTE DES ANNEXES

Annexe A	Graphes pour les stations d'intérêt	67
Annexe B	Organigramme de la construction des courbes IDF	70

LISTE DES SIGLES ET ABRÉVIATIONS

GIEC	Groupe d'experts Intergouvernemental sur l'Évolution du Climat
GES	Gaz à effet de serre
ECCC	Environment Climate Change Canada
IDF	Intensité-Durée-Fréquence
i.i.d.	indépendantes identiquement distribuées
GEV	Generalized Extreme Values
GPD	Generalized Pareto Distribution
TCL	Théorème Central Limite
RFA	Regional Frequency Analysis
CaSPAr	Canadian Surface Project Archive
RDRS	Regional Deterministic Reforecast System
CaPA	Canadian Precipitation Analysis
CaLDAS	Canadian Land Data Assimilation System
MCMC	Markov Chains Monte Carlo
GMRF	Gaussian Markov Random Field
iGMRF	intrinsic Gaussian Markov Random Field
LOOCV	Leave One Out Cross-Validation
MSE	Mean Squared Error
CVM	Cramer-Von Mises
MALA	Metropolis-Adjusted Langevin Algorithm

CHAPITRE 1 INTRODUCTION

Il est de notoriété publique désormais que le changement climatique est une réalité tangible, bien plus palpable ces dernières années. Ce bouleversement apporte avec lui son lot d'évènements extrêmes, dont les fréquences se font de plus en plus pressantes sur les infrastructures et les populations. Parmi ces évènements, on peut notamment citer les pics de chaleurs au Canada causant de grands feux de forêt [1], des vagues de chaleurs en Espagne résultant en une désertification [2] [3], longue période sans pluie en France, vagues de chaleur mortelles en Inde [4], canicules marines en Mer Méditerranée [5], etc...

En particulier, les précipitations extrêmes sont une manifestation particulièrement courante de ce changement climatique. Des épisodes de pluie intense peuvent être source de crues, de débordements d'égouts, d'inondations et de dommages sérieux aux infrastructures exposées. L'estimation de ces précipitations extrêmes représente donc un enjeu certain pour les sociétés exposées à ces risques, aussi bien social qu'économique.

Un outil fréquemment utilisé pour dimensionner les infrastructures exposées aux précipitations sont les courbes d'Intensité-Durée-Fréquence (IDF). L'utilisation de ces courbes sont obligatoires dans divers domaines du génie pour dimensionner en fonction des des niveaux de pluies extrêmes auxquels on peut s'attendre sur un certain nombre d'années. À l'aide de ces tables, les ingénieurs peuvent concevoir plusieurs infrastructures telles les digues, les réseaux d'aqueduc, les canalisations ainsi que définir les zones inondables de petits bassins versants.

En pratique, l'utilisation d'une courbe IDF résulte en l'obtention d'une intensité de précipitation excédée en moyenne tous les T années, permettant ainsi de dimensionner notamment la hauteur des infrastructures concernées. Cette grandeur T est appelée la période de retour, et est donc variable suivant la politique de protection et de sauvegarde des bâtiments impactés.

Une série de courbes IDF se présente sous la forme des informations croisées entre plusieurs périodes de retour (habituellement 2, 5, 10, 20, 25, 50 ou 100 ans), les durées d'accumulation et les intensités de précipitations. En fonction de l'infrastructure, la durée d'accumulation sera fixée, et les intensités correspondant aux différentes périodes de retour peuvent être récupérées sur ces courbes. La construction d'une courbe est faite en la calculant en plusieurs points précis : en choisissant la durée d'accumulation D , on extrait les maxima annuels des précipitations pour cette même durée à la station étudiée. Ces maxima permettent ensuite d'ajuster une loi de probabilité, dont on va extraire les quantiles. Ces quantiles permettent ensuite de déduire l'intensité I_T pour chaque période de retour T . Le graphe résultant sera donc formé des points (D, I_T) , que l'on va interpoler par une courbe distincte pour chaque

période de retour T . Pour obtenir des courbes rectilignes, il est d'usage de les tracer sur des échelles logarithmiques.

Au Canada, les courbes IDF utilisées principalement sont celles d'Environnement Changement Climatique Canada (ECCC). Ces courbes sont disponibles en ligne [6] pour un ensemble de stations météorologiques répartis sur le territoire canadien. Un exemple de courbes fournies par ECCC, pour la station de l'université de Colombie-Britannique à Vancouver (UBC) est visible en figure 1.1. L'axe des abscisses correspond aux différentes durées, tandis que l'axe des ordonnées dénote l'intensité en mm/h, les deux axes étant en échelle logarithmique. Les croix indiquent l'intensité calculée pour une durée et un niveau de retour déterminés. Par-dessus les croix, on fait apparaître les courbes IDF, qui sont interpolées pour une période de retour fixée sur toutes les croix, de durée différente.

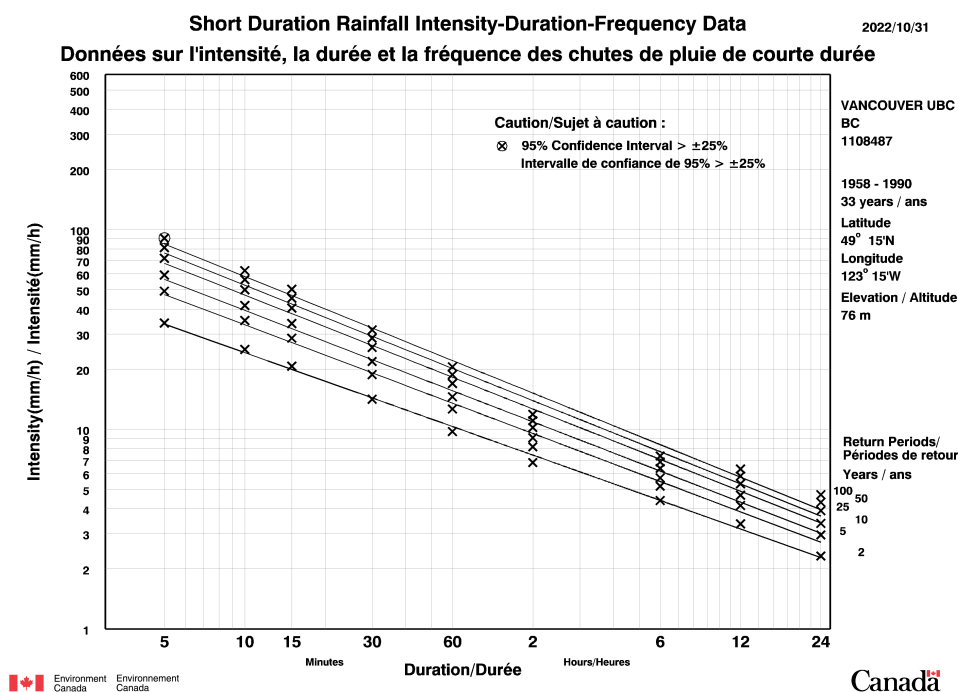


FIGURE 1.1 Courbes IDF d'ECCC à la station Vancouver UBC

Actuellement, les courbes IDF ne sont disponibles qu'aux sites des stations météorologiques exploitées par ECCC. En pratique, les ingénieurs utilisent les courbes de la station la plus proche du site de l'infrastructure planifiée. Parfois, plusieurs centaines de kilomètres séparent les deux sites et cette approche néglige l'hétérogénéité spatiale des précipitations extrêmes. Elles souffrent également d'une importante hypothèse simplificatrice : les intensités correspondant aux différentes périodes de retour sont estimées avec la loi de Gumbel, qui est une loi à queue légère alors que les précipitations sont généralement caractérisées par une queue

lourde. Cette approche n'est pas cohérente avec les récents développements en théorie des valeurs extrêmes.

Plusieurs études récentes se sont concentrées sur l'interpolation des précipitations extrêmes. Les modèles utilisés requièrent généralement l'introduction de variables explicatives et de fonctionnalités spécifiques à la région d'intérêt. Ceci est d'autant plus important pour le Canada, d'une part pour l'immensité et la diversité de son territoire, et, d'autre part, pour la répartition éparse des stations météorologiques. En effet, la grande majorité des stations se trouvent au sud du pays dans les zones les plus peuplées. Or, les régions nordiques reculées sont toutes aussi importantes pour l'estimation des courbes IDF, car un grand nombre d'infrastructures s'y trouvent, notamment les barrages hydroélectriques. Ainsi, le fait de pouvoir estimer les courbes IDF sur l'ensemble d'une province donnée et non pas juste à quelques points spécifiques relève d'une grande importance, spécialement pour un pays aussi étendu que le Canada, et c'est sur cet aspect que va porter principalement mon projet de recherche.

L'objectif général de ce projet de recherche consiste à estimer les courbes IDF pour les précipitations de la Colombie-Britannique. Actuellement, l'interpolation des courbes a été effectuée pour l'est du Canada, mais pas pour cette région bordée par l'océan Pacifique à l'ouest et les montagnes Rocheuses à l'est. Un modèle hiérarchique bayésien permettant l'interpolation des caractéristiques des précipitations extrêmes sera adapté spécifiquement pour cette région. Les principaux objectifs de ce projet sont donc :

1. Développer un modèle statistique spatial pour l'estimation des précipitations extrêmes en Colombie-Britannique.
2. Déterminer les variables environnementales pertinentes pour l'interpolation des caractéristiques des précipitations extrêmes.
3. Mettre en oeuvre un processus de validation croisée des résultats pour comparer les différents modèles retenus.

Au terme de ce travail, les niveaux de retour attendus pour la période de retour de 20 ans, sur la durée de relevé de 24 heures, seront obtenus. La réitération de la méthode présentée, avec les autres périodes de retour et durées de relevé, permettra à terme d'obtenir les courbes IDF en tout lieu de la Colombie-Britannique, conjointement avec celles déjà disponibles pour l'est du Canada.

Ce mémoire est structuré comme suit : un portrait large mais nécessaire de la théorie utilisée dans ce projet sera dressé dans le chapitre 2 ; ensuite, après une revue de littérature effectuée au cours du chapitre 3, les deux jeux de données principaux employés seront présentés à travers le chapitre 4, conjointement avec une analyse exploratoire. Le chapitre 5, lui, est dédié à expliquer le modèle spatial général développé dans le cadre de la Colombie Britannique. Le

chapitre 6 décrit finalement les résultats obtenus ainsi que leur validation croisée, permettant ainsi de trancher entre les différentes variables étudiées. Finalement, le chapitre 7 résume le projet dans son ensemble, les conclusions que l'on peut en tirer ainsi que de multiples pistes de développements futurs.

Pour effectuer ce travail, le langage `Julia` a été principalement utilisé, notamment pour implémenter les algorithmes Monte-Carlo par chaînes de Markov. Le langage `R` a quant à lui été utilisé pour développer un dashboard interactif présentant les jeux de données utilisés, ainsi qu'une synthèse des résultats.

CHAPITRE 2 CADRE THÉORIQUE

Dans ce chapitre, l'objectif est d'effectuer un tour d'horizon général des différentes théories mathématiques qui seront utiles pour l'estimation et l'interpolation des courbes IDF. Dans un premier temps, la théorie des valeurs extrêmes sera présentée, la branche des statistiques formant le fondement de ce projet.

Ensuite, nous nous pencherons sur les bases de la théorie de l'inférence bayésienne, afin d'introduire les modèles bayésiens hiérarchiques, les algorithmes Monte-Carlo par chaînes de Markov (MCMC) et leur utilisation en pratique dans la suite du projet.

Enfin, nous introduirons rapidement la théorie des champs gaussiens markoviens (*Gaussian Markov Random Fields*, GMRF), servant à la représentation spatiale de nos données sous forme de grille.

2.1 Théorie des valeurs extrêmes

La théorie des valeurs extrêmes est une branche des probabilités et de la statistique qui s'intéresse aux valeurs rares d'une variable aléatoire. Si cet intérêt pour les valeurs extrêmes a émergé durant le 20e siècle [7], ce champ fut proprement théorisé au milieu des années 50. Cette théorie trouve de nombreuses applications, aussi bien en ingénierie qu'en climatologie ou en finance.

L'essentiel de la théorie présentée ici est principalement tirée de l'ouvrage de référence de Coles (2001) [8], que le lecteur intéressé pourra consulter pour davantage de détails.

2.1.1 Théorème principal

Posons tout d'abord un cadre mathématique pour l'étude des valeurs extrêmes. Formellement, si on note $X_1, \dots, X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées (i.i.d.), chacune de fonction de densité f_X et de fonction de répartition F_X , et si on pose la variable aléatoire M_n définie comme : $M_n = \max_{1 \leq i \leq n} X_i$, alors par indépendance des $(X_i)_{1 \leq i \leq n}$, on peut écrire M_n comme ceci :

$$\mathbb{P}(M_n \leq x) = \mathbb{P}(X_1 \leq x, X_2 \leq x, \dots, X_n \leq x) \quad (2.1)$$

$$= \prod_{i=1}^n \mathbb{P}(X_i \leq x) \quad (2.2)$$

$$= F_X^n(x). \quad (2.3)$$

Or, dans la pratique, la fonction de répartition des $(X_i)_{1 \leq i \leq n}$, F_X , est inconnue. Une première solution est d'estimer F_X au vu des données, mais de part les incertitudes que l'on peut avoir sur cette estimation, cette incertitude s'amplifie d'autant plus que notre taille d'échantillon n est grande, de part le passage à la puissance dans $\mathbb{P}(M_n \leq x) = F_X^n(x)$.

Heureusement, c'est ici où intervient la théorie des valeurs extrêmes : quand n est grand, le théorème fondamental de la théorie nous donne une caractérisation asymptotique de F_X^n , et ce indépendamment de F_X . Plus exactement, ce théorème se présente comme suit :

Théorème 2.1.1 (Théorème de Fisher-Tippett-Gnedenko). *Soit $(X_i)_{1 \leq i \leq n}$ une suite de variables aléatoires i.i.d., et soit M_n la variable aléatoire associée au maximum de cette suite, i.e. $M_n = \max_{1 \leq i \leq n} X_i$.*

Si l'on dispose de deux suites $a \in (\mathbb{R}_+^*)^{\mathbb{N}}$ et $b \in (\mathbb{R}^{\mathbb{N}})$ telles que :

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

où G est une distribution non-dégénérée, alors G est nécessairement de la forme :

$$G(z) = \begin{cases} \exp \left\{ - \left[1 + \xi \left(\frac{z - \mu}{\sigma} \right) \right]^{-1/\xi} \right\} & \text{si } \xi \neq 0 \\ \exp \left\{ - \exp \left(\frac{z - \mu}{\sigma} \right) \right\} & \text{si } \xi = 0. \end{cases} \quad (2.4)$$

Cette famille de distributions est appelée GEV, pour Generalized Extreme Value. Les trois paramètres (μ, σ, ξ) sont communément appelés respectivement paramètre de localisation, paramètre d'échelle et paramètre de forme. Si les deux premiers paramètres sont analogues à la moyenne et la variance, sans être exactement égaux à la vraie espérance ou variance, (cf. plus bas dans cette section), le paramètre de forme ξ quant à lui va dicter le poids de la queue de la distribution G . En effet, on peut classer les lois GEV en trois catégories suivant leur paramètre de forme ξ , à savoir les familles de Weibull (à queue droite finie), Gumbel (à

queue légère) & Fréchet (à queue lourde) :

$$\text{Weibull } (\xi < 0) : G(z) = \begin{cases} \exp\left(-\left(\frac{z-a}{b}\right)^{-\alpha}\right) & z \geq a \\ 0 & z < a \end{cases} \quad (2.5)$$

$$\text{Gumbel } (\xi = 0) : G(z) = \exp\left\{-\exp\left(\frac{z-\mu}{\sigma}\right)\right\}. \quad (2.6)$$

$$\text{Fréchet } (\xi > 0) : G(z) = \begin{cases} \exp\left(-\left(\frac{z-a}{b}\right)^\alpha\right) & z \geq a \\ 0 & z < a \end{cases}, \quad (2.7)$$

où l'on pose les normalisations $a = \mu - \sigma/\xi$, $b = \sigma/\xi$ et $\alpha = |\xi|$ dans le cas des familles de Weibull et Fréchet. La famille de Weibull a un support fini à droite, ce qui résulte en une queue de distribution elle aussi finie, et de moins en moins importante au fur et à mesure que ξ sera négatif. La famille de Gumbel est définie sur tout $\mathbb{R}$, et a une queue de distribution à décroissance exponentielle, tandis que la famille de Fréchet a un support fini à gauche et est une famille dite "à queue lourde".

En pratique en hydrologie, la loi Gumbel est souvent utilisée, mais c'est une mauvaise pratique, car *a priori* le paramètre de forme ξ n'est pas nécessairement nul, quand bien même la loi de Gumbel a l'avantage d'être définie sur tout $\mathbb{R}$, contrairement aux deux autres familles. Par ailleurs, même lorsque ξ est non nul, il est courant de supposer que $|\xi| \leq 0.5$ [9], car sinon on en vient à avoir une variance non définie. En effet, on a, pour $X \sim \text{GEV}(\mu, \sigma, \xi)$:

$$\mathbb{E}[X] = \mu + \frac{\mu}{\sigma}(g_1 - 1), \quad (2.8)$$

$$\text{Var}(X) = \frac{\sigma^2}{\xi^2}(g_2 - g_1^2), \quad (2.9)$$

où $g_k = \Gamma(1 - k\xi)$, pour $k \in \mathbb{N}^*$.

Enfin, en pratique, les suites (a_n) et (b_n) du théorème de Fisher-Tippett-Gnedenko permettent de normaliser la variable aléatoire M_n . Elles sont inconnues, mais ce n'est pas vraiment un problème car elles sont absorbées dans les paramètres à estimer. En effet, on a :

$$\mathbb{P}(M_n \leq z) = \mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq \frac{z - b_n}{a_n}\right) \quad (2.10)$$

$$= G\left(\frac{z - b_n}{a_n}\right) \quad (2.11)$$

$$= G^*(z), \quad (2.12)$$

où G^* est également membre de la famille des GEV [8].

2.1.2 Modèle des maxima par blocs et niveaux de retour

Le modèle de maxima par blocs est l'approche la plus courante en pratique pour l'application du théorème fondamental de la théorie des valeurs extrêmes. L'idée est de partitionner notre série de variables aléatoires en blocs successifs de taille identique n , et d'obtenir le maximum de chaque bloc, afin d'avoir une série de M_n pour estimer les paramètres de la loi GEV correspondante. Dans ce cas, l'estimation de G consiste à estimer directement F_X^n sans passer par F_X .

Un enjeu du modèle maxima par blocs est de justement avoir une taille de blocs adaptée. Si celle-ci est trop petite, alors l'estimation de la loi GEV sera de piètre qualité par manque de données pour chaque bloc, et mènera à un biais important ; inversement, si celle-ci est trop grande, les maxima seront alors potentiellement trop peu nombreux, ce qui mènera à une variance importante lors de l'estimation des paramètres.

En sciences environnementales, la taille de bloc considérée est souvent égale à une année, ce qui est le cas dans ce projet. Pour estimer les paramètres de la GEV compte tenu de notre série de données, on va recourir à l'estimation par maximum de vraisemblance. Si l'on note $Y_1, \dots, Y_m$ notre série de maxima, que l'on suppose indépendants, la log-vraisemblance de l'échantillon est donnée par :

$$l(\theta, \mathbf{y}) = \begin{cases} -m \ln(\sigma) - \sum_{i=1}^m \left[\left(1 + \frac{1}{\xi}\right) \ln \left(1 + \xi \frac{y_i - \mu}{\sigma}\right) + \left(1 + \xi \frac{y_i - \mu}{\sigma}\right)^{-1/\xi} \right] & \text{si } \xi \neq 0. \\ -m \ln(\sigma) - \sum_{i=1}^m \left[\frac{y_i - \mu}{\sigma} + \exp\left(-\frac{y_i - \mu}{\sigma}\right) \right] & \text{si } \xi = 0. \end{cases} \quad (2.13)$$

Malheureusement, cette log-vraisemblance est trop compliquée pour être calculée analytique-

ment ; elle est donc obtenue via des méthodes numériques, pour la plupart déjà préalablement implémentées dans la librairie `Extremes.jl` [10].

Enfin, une fois que les paramètres de la GEV sont obtenus numériquement, notre principale grandeur pour visualiser d'un coup nos réalisations de GEV est effectuée via ses quantiles. Par inversion de la forme de la GEV, on obtient :

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi} (1 - (-\ln(1-p))^{-\xi}) & \text{si } \xi \neq 0 \\ \mu - \sigma \ln(-\ln(1-p)) & \text{si } \xi = 0. \end{cases} \quad (2.14)$$

Le quantile z_p est associé au nombre d'années $T = 1/p$, et est aussi appelé niveau de retour correspondant à la période de retour de T années : cette grandeur va représenter environ le niveau dépassé par le maximum d'une année avec une probabilité p , ou plus simplement le niveau dépassé en moyenne tous les T années. C'est ainsi une grandeur permettant efficacement de comparer d'un seul coup les extrêmes d'une région, simplement en modulant la valeur T pour visualiser les maxima auxquels on peut s'attendre tous les ans, tous les 5 ans, ou tous les 20 ans par exemple.

2.2 Inférence bayésienne et modèles hiérarchiques

2.2.1 Inférence bayésienne

L'approche par inférence bayésienne est une méthode d'inférence statistique distincte de la méthode utilisant le maximum de vraisemblance ; en inférence, on se sert du théorème de Bayes pour estimer un paramètre θ à partir d'une population. Dans cette approche, le paramètre est vu comme une variable aléatoire, dont on va déterminer la distribution conditionnellement par rapport à nos données.

Plus formellement, le théorème de Bayes s'énonce comme suit :

Théorème 2.2.1 (Théorème de Bayes, version continue). *Soit X et Y deux variables aléatoires réelles, de distributions marginales respectives f_X et f_Y . La distribution conditionnelle de X en fonction de $Y = y$ peut s'écrire :*

$$f_{(X|Y=y)}(x) = \frac{f_{(Y|X=x)}(y)f_X(x)}{\int f_{(Y|X=x)}(y)f_X(x)dx}.$$

ici Ainsi, ce théorème permet en quelque sorte « d'inverser la flèche du temps », en reversant l'ordre de conditionnement entre deux variables aléatoires. Dans le cadre de l'inférence sta-

tistique, on va donc avoir nos données y et notre paramètre θ , et le théorème de Bayes va permettre de relier la loi *a posteriori* du paramètre $f_{(\theta|Y=y)}(\theta)$ et sa loi *a priori* $f_{\theta}(\theta)$:

$$f_{(\theta|Y=y)}(\theta) = \frac{f_{(Y|\theta)}(y)f_{\theta}(\theta)}{\int_{\theta} f_{(Y|\theta)}(y)f_{\theta}(\theta)d\theta}. \quad (2.15)$$

L'inférence bayésienne se base donc sur l'utilisation de cette loi *a posteriori* à partir de la loi *a priori*, et qui diffère de l'estimation classique où l'on a ici une distribution complète et non une valeur estimée ponctuelle. À partir de cette distribution, on peut en déduire d'autres grandeurs, telles que la densité prédictive *a posteriori*, pouvant servir à obtenir la densité pour une nouvelle population y' :

$$f_{(Y'|Y=y)}(y') = \int_{\theta} f_{(Y|\theta)}(y')f_{\theta}(\theta)d\theta. \quad (2.16)$$

Le choix de la loi *a priori* est une étape préliminaire importante lors d'une étude en inférence bayésienne, et de nombreux choix sont possibles. Concrètement, il existe trois cas pour choisir la loi *a priori* :

subjective : lorsque nous disposons d'assez d'informations sur θ pour construire nous-même entièrement la loi, ce qui arrive rarement.

partiellement informative : nécessite de connaître le modèle échantionnal afin de construire la loi *a priori* comme loi conjuguée.

objective : lorsque nous disposons de peu ou pas d'informations sur θ , on peut utiliser des lois de références : de nombreuses sont possibles, les plus communes étant la loi *a priori* de Laplace (uniforme), loi *a priori* de Jeffreys (obtenue à partir de la quantité d'information de Fisher pour θ).

Ces différents cas sont traités en détail dans le livre de Roberts [11].

2.2.2 Modèle bayésien hiérarchique

Les modèles bayésiens hiérarchiques sont une application, ou un prolongement, de la théorie de l'inférence bayésienne. Concrètement, cela consiste à ajouter une ou plusieurs couches supplémentaires de paramètres dans un modèle statistique donné. La plupart du temps, il y aura une seule couche, qui sera appelée la *couche latente*. Cette manière de construire un modèle est utile lorsqu'on a une famille de paramètres que l'on souhaite rassembler derrière

ces hyper-paramètres ; la structure hiérarchique représente une manière élégante d'avoir des paramètres différents spatialement, tout en étant unifiés via cette couche d'hyper-paramètres.

Plus formellement, si l'on conserve la notation θ pour les paramètres et que l'on rajoute les hyper-paramètres en les notant ψ , on peut écrire, par loi de Bayes conjuguée, la loi *a posteriori* comme suit :

$$f_{(\theta, \psi | Y=y)}(\theta, \psi) \propto f_{(Y|\theta)}(y) f_{\theta|\psi}(\theta) f_{\psi}(\psi), \quad (2.17)$$

où :

- $f_{(Y|\theta)}$ est toujours la loi des données (ou couche des données) ;
- $f_{(\theta|\psi)}$ est la loi conditionnelle des paramètres en conditionnellement aux hyper-paramètres (couche latente) ;
- $f_{\psi}(\psi)$ est la loi *a priori* des hyper-paramètres (couche des hyper-paramètres).

On reconnaît la loi de Bayes vue précédemment, mais agglomérée en deux couches distinctes et successives. On remarque d'ailleurs qu'en intégrant les deux côtés de cette relation par les hyperparamètres ψ , on retrouve la relation de Bayes décrite à la section précédente. Un résumé schématique de ce type de modèle est présenté en figure 2.1, en comparaison avec les modèles d'indépendance totale ou dépendance totale des données entre elles.

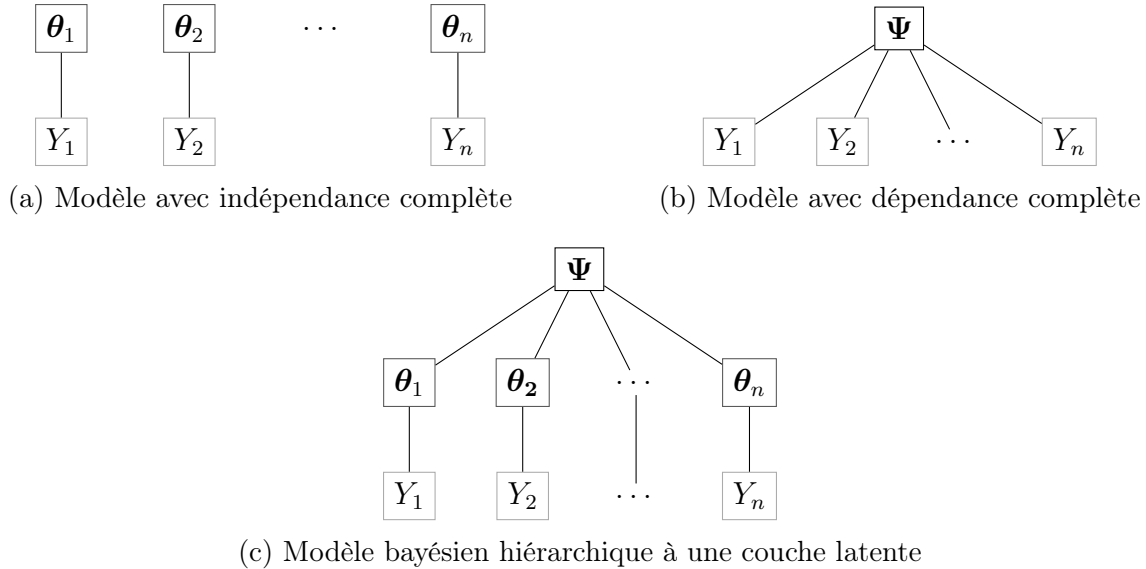


FIGURE 2.1 Comparaison des modèles de dépendance

Il est possible d'étendre à plus de couches latentes successives, mais alors cela augmente

drastiquement la complexité du modèle. Dans la plupart des cas, y compris le nôtre, une seule couche est amplement suffisante, et permet de modéliser cette interaction globale entre tous les paramètres répartis spatialement.

2.2.3 Méthode Monte-Carlo par Chaînes de Markov

Un des problèmes avec la mise en oeuvre en pratique de l'inférence bayésienne, en particulier en grande dimension, c'est de l'impossibilité de calculer la loi *a posteriori*. Une méthode utilisée pour contourner le problème et simuler cette loi repose dans les méthodes Monte-Carlo par chaînes de Markov (*Monte Carlo Markov Chains*, MCMC). Le principe de cette méthode est de construire une chaîne de Markov ayant pour loi stationnaire la loi que l'on cherche à obtenir. De plus, cette méthode ne nécessite pas d'avoir la constante de normalisation de cette loi, qui peut être impossible à calculer.

Les deux algorithmes principalement utilisés dans le domaine des MCMC sont l'échantillonnage de Gibbs et l'algorithme de Métropolis-Hastings. Le principe de ces algorithmes est de construire pas à pas cette chaîne de Markov ayant pour loi stationnaire notre loi cible ; à l'issue de son exécution, nous obtenons donc une chaîne $(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$, où n va être le nombre d'itérations effectué, choisi arbitrairement mais devant être assez important pour s'assurer une bonne exploration de l'espace des possibilités et une certaine convergence (usuellement plusieurs dizaines de milliers d'itérations suffisent). On s'assure également d'enlever une partie des itérations initiales, avant que la chaîne ne se stabilise. On parle de « temps de chauffe » m ; on obtient donc une série de valeurs $(x_{m+1}, \dots, x_n)$, qui peut être vu comme un tirage de la loi *a posteriori* que l'on cherche à simuler. Ainsi, pour estimer les paramètres θ , il suffit d'appliquer le théorème centrale limite (TCL) et moyenner toutes ces valeurs :

$$\mathbb{E}[\theta] \simeq \frac{1}{n - m} \sum_{i=m+1}^n x_i. \quad (2.18)$$

Pour commencer, l'échantillonnage de Gibbs consiste à simuler, pour une densité cible multidimensionnelle, ses paramètres un à un, conditionnellement aux autres successivement. Si l'on note $f_{\theta}(\theta_1, \dots, \theta_p)$, alors la méthode va utiliser les lois conditionnelles complètes successives $f_{(\theta_i|\theta_{-i})}$ pour $1 \leq i \leq p$. Une itération de l'algorithme de Gibbs est présentée dans l'algorithme 1.

Quant à l'algorithme de Métropolis-Hastings, il représente une forme de généralisation des algorithmes MCMC (notamment de l'échantillonnage de Gibbs). L'idée est qu'à chaque itéra-

Algorithme 1 : Itération i de l'échantillonnage de Gibbs

```

 $\theta \leftarrow (\theta_1^i, \dots, \theta_p^i);$ 
;
/* $i$ -ème élément de la chaîne de Markov*/
pour  $1 \leq k \leq p$  faire
| simuler  $\theta_k^{i+1} \sim f_{(\theta_k | \theta_1^{i+1}, \dots, \theta_{k-1}^{i+1}, \theta_{k+1}^i, \dots, \theta_p^i)}$ ;
fin
 $\theta \leftarrow (\theta_1^i, \dots, \theta_p^i);$ 

```

tion où l'on simule une valeur pour θ , on va l'accepter ou non comme nouvel élément de la chaîne de Markov. Ce taux d'acceptation, noté α , va dépendre d'une distribution instrumentale q choisie au préalable. À noter que cette distribution dépend conditionnellement de θ , elle sera donc noté $q(\cdot | \theta)$. Une itération de cet algorithme est présentée dans l'algorithme 2.

Algorithme 2 : Itération i de l'algorithme Métropolis-Hastings

```

 $\theta \leftarrow \theta^i;$ 
simuler  $\tilde{\theta} \sim q(\cdot | \theta^i);$ 
/*candidat de la  $i$ -ème itération*/
 $\alpha \leftarrow \min \left\{ 1, \frac{f_{\theta}(\tilde{\theta})q(\theta^i | \tilde{\theta})}{f_{\theta}(\theta^i)q(\tilde{\theta} | \theta^i)} \right\};$ 
/*Taux d'acceptation*/
simuler  $r \sim \mathcal{U}[0, 1];$ 
si  $\alpha > r$  alors
|  $\theta \leftarrow \tilde{\theta};$ 
/*On accepte le candidat*/
sinon
|  $\theta \leftarrow \theta^i;$ 
/*On rejette le candidat*/
fin

```

À noter qu'il existe beaucoup de variations des algorithmes MCMC autour des deux présentés ci-dessus. Le lecteur intéressé pourra se référer à l'ouvrage de Roberts [11].

2.3 Champs markoviens gaussiens aléatoires

Dans le cadre de ce projet de recherche, puisque les variables explicatives sont disponible sur une grille, les champs markoviens gaussiens (*Gaussian Markov Random Fields*, GMRF) se sont imposés naturellement comme modèle pour la couche latente. L'idée derrière cette théorie est de proposer une forme de vecteur aléatoire normal, dont la matrice de covariance dispose d'une structure calquée sur un graphe non-orienté. Ainsi, comme nous le verrons dans la partie modélisation, ce champ peut modéliser nos erreurs spatialement, en supposant évidemment leur normalité.

Le lecteur intéressé par plus de détails quant à la théorie pourra se référer à l'ouvrage très complet de Rue & Held [12], avec notamment plusieurs études de cas de modèles bayésiens

hiérarchiques dans leur chapitre 4.

2.3.1 Théorie de base

Tout d'abord, il convient de rappeler la définition et les notations d'un graphe :

Définition 2.3.1. *Un graphe non-orienté $\mathcal{G}$ est un couple $(\mathcal{V}, \mathcal{E})$ où :*

- $\mathcal{V}$ est un ensemble fini de taille n , appelé l'ensemble des sommets.
- $\mathcal{E}$ est un sous-ensemble de $V \times V$, vérifiant les propriétés suivantes :
 1. $\forall (x, y) \in \mathcal{V} \times \mathcal{V}, (x, y) \in \mathcal{E} \Leftrightarrow (y, x) \in \mathcal{E}$ (propriété de non-orientation).
 2. $\forall x \in V, (x, x) \notin \mathcal{E}$ (aucun sommet n'est relié à lui-même).

Dans la suite de cette partie, comme on travaille en vecteurs aléatoires dans $\mathbb{R}^n$, on aura par défaut notre ensemble de sommets $\mathcal{V} = \{1, \dots, n\}$. Par ailleurs, à la place d'utiliser la matrice de covariance de ces vecteurs, habituellement notée Σ , on préfère employer la matrice de précision Q , définie simplement comme l'inverse de la matrice de covariance, i.e. $Q = \Sigma^{-1}$.

Définition 2.3.2. *Un vecteur aléatoire $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ est un GMRF de graphe $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, de moyenne $\boldsymbol{\mu}$ et de matrice de précision Q si et seulement si sa densité de probabilité est de la forme :*

$$f_X(\mathbf{x}) = \frac{1}{(2\pi)^n} |Q|^{\frac{1}{2}} \exp \left\{ -(\mathbf{x} - \boldsymbol{\mu})^\top Q (\mathbf{x} - \boldsymbol{\mu}) \right\}, \quad (2.19)$$

et :

$$\forall (i, j) \in 1, \dots, n^2, i \neq j \Rightarrow (Q_{ij} \neq 0 \Leftrightarrow (i, j) \in \mathcal{E}). \quad (2.20)$$

Ainsi, la matrice de précision, qui correspond à l'inverse de la matrice de covariance, résulte en la forme du graphe, et inversement. Par ailleurs, si un vecteur aléatoire gaussien dispose d'une matrice de covariance symétrique définie positive, alors ce vecteur est un GMRF, et inversement.

La définition des GMRF permet donc d'obtenir les moments de x_i conditionnellement à tous les autres (ce qu'on note x_{-i}) :

Théorème 2.3.1. *Soit x un GMRF lié au graphe $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, de moyenne μ et de matrice de précision $Q := \Sigma^{-1}$. On a :*

$$\mathbb{E}(x_i|x_{-i}) = \mu_i - \frac{1}{Q_{ii}} \sum_{j|(i,j) \in \mathcal{E}} Q_{ij}(x_j - \mu_j); \quad (2.21)$$

$$Prec(x_i|x_{-i}) = Q_{ii}; \quad (2.22)$$

$$Cor(x_i, x_j|x_{-ij}) = -\frac{Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}} \text{ où } i \neq j. \quad (2.23)$$

Pour ce qui est de l'aspect numérique, il existe plusieurs algorithmes basés sur des théorèmes d'algèbre linéaire (décomposition de Cholesky notamment) permettant la simulation numérique des GMRF. Le lecteur intéressé pourra trouver de nombreux détails dans les parties 2.3 & 2.4 de Rue & Held [12].

2.3.2 Champs intrinsèques

Dans les définitions générales des GMRF, une composante importante est la matrice de précision (ainsi que la matrice de covariance, de part leur relation). Pour que la matrice de précision existe, on doit pouvoir inverser la matrice de covariance. Néanmoins, à quelques ajustements près de la définition d'origine, on peut en définir une version « impropre » :

Définition 2.3.3. (*GMRF impropre*) Soit Q une matrice symétrique semi-définie positive, de rang $n - k > 0$. $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top$ est un GMRF impropre de paramètres (μ, Q) associé au graphe $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ si sa densité est de la forme :

$$f_X(\mathbf{x}) = (2\pi)^{-\frac{n-k}{2}} (|Q|^*)^{1/2} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \mu)^\top Q(\mathbf{x} - \mu) \right\}, \quad (2.24)$$

et :

$$\forall (i, j) \in 1, \dots, n^2, i \neq j \Rightarrow (Q_{ij} \neq 0 \Leftrightarrow (i, j) \in \mathcal{E}). \quad (2.25)$$

Ici, le nombre k est appelé "*rank deficiency*", ou bien « défaillance de rang » : il dénote à quel point l'on s'éloigne du matrice de précision inversible, et donc d'un GMRF non impropre. À noter que l'on ne parle plus de moyenne pour μ ni de de précision pour Q relativement à x , car ces dernières n'existent pas.

Ce que l'on va utiliser principalement en pratique dans ce projet, c'est les iGMRF, où le « i » signifie « intrinsèque ». Il s'agit simplement d'un cas particulier des GMRF impropres, où l'on fixe la déficience de rang. On peut classer ces iGMRF en fonction de leur ordre, qui

correspond simplement à la déficience de rang qu'on leur attribue, associée à une contrainte linéaire simple :

Définition 2.3.4 (iGMRF du premier ordre). *Un iGMRF du premier ordre est un GMRF avec une déficience de rang $k = 1$ vérifiant $Q\mathbb{1}_n = 0$,*

où l'on pose $\mathbb{1}_n = (1, 1, \dots, 1)^\top \in \mathbb{R}^n$. La condition $Q\mathbb{1}_n = 0$ peut aussi s'écrire $\sum_j Q_{ij} = 0$ pour tout i .

Usuellement, on introduit également une grandeur κ , appelée *paramètre de précision*. κ sera une constante d'homotéthisme appliquée sur toute la matrice Q qui va permettre de paramétrer la variance conditionnelle entre les cellules : plus κ est grand et plus le champ sera lisse, et inversement.

Cette sous-catégorie de GMRF permet d'avoir une matrice de structure avec des contraintes linéaires, permettant de représenter les liens entre sommets voisins. L'exemple le plus simple de matrice de structure est donné, pour une marche linéaire allant de 1 à n (avec donc chaque sommet ayant deux voisins, à part les extrémités qui n'en ont qu'un) :

$$R = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & (0) \\ & -1 & 2 & -1 & \\ & & \ddots & \ddots & \ddots \\ (0) & & & -1 & 2 & -1 \\ & & & & -1 & 1 \end{pmatrix}.$$

Dans notre cas, où l'on est en deux dimensions, chaque case de la grille aura 4 voisins, chacun dans les 4 points cardinaux, à l'exception des cases aux bordures qui n'auront que 3 voisins, et les cases aux coins n'en auront que 2.

De la même manière, on peut définir des iGMRF d'ordre supérieur, mais la définition se complexifie légèrement, car elle fait intervenir une matrice S_k de rang k pour laquelle un iGMRF d'ordre k doit vérifier $QS_k = 0$.

En pratique, la génération des iGMRF sera effectuée à l'aide de la librairie Julia `GMRF.jl` [13], où tout est orienté vers les iGMRF de rang 1 ou 2, et utilisant la plupart des algorithmes présents dans Rue & Held [12].

CHAPITRE 3 REVUE DE LITTÉRATURE

Cela fait plusieurs années maintenant, avec l’impulsion de l’accélération du changement climatique, que plusieurs études portant sur les précipitations extrêmes voient le jour, et ce dans un contexte de modélisation spatiale. La plupart de ces méthodes s’appuient sur des théories vues au chapitre précédent, à savoir les modèles bayésiens hiérarchiques, la théorie des valeurs extrêmes, mais également les champs markoviens gaussiens.

Dans ce chapitre, nous effectuerons une revue de littérature des études récentes de ce domaine, et leur influence sur le déroulé et la mise en oeuvre de ce projet de recherche. Dans un premier temps, nous aborderons la méthode d’analyse fréquentielle régionale, et plusieurs articles ayant utilisé cette approche ; puis, nous continuerons sur l’approche avec des modèles bayésiens hiérarchiques, avec une composante de dépendance spatiale. Nous nous attarderons également sur d’autres études localisées dans le monde. Enfin, nous nous pencherons sur le modèle climatique utilisé dans ce projet, RDRS v2.1.

Pour commencer, présentons la méthode d’analyse fréquentielle régionale (Regional Frequential Analysis, RFA). Elle a été développée par Hosking & Wallis en 1997 [14], et elle permet de partager l’information sur les extrêmes entre les sites. L’idée est de séparer la région étudiée en plusieurs sous-régions homogènes du point de vue des extrêmes. Cette façon d’aborder les valeurs extrêmes est tout à fait légitime lorsque les hypothèses sont raisonnables, et permet d’obtenir des résultats légèrement meilleurs qu’en considérant une indépendance totale des stations de relevés. En revanche, elle comporte plusieurs limites, qui s’avèrent pertinentes dans le contexte de ce projet ; tout d’abord, les sous-régions délimitant l’espace total sont choisies de manière arbitraire, ce qui peut s’avérer complexe dans des régions étendues et manquant de repères géographiques pour se séparer. C’est en partie le cas ici avec la Colombie-Britannique : à part les Rocheuses, il n’est pas évident quelle séparation adopter. Une autre limite est l’impossibilité d’ajouter des variables explicatives avec l’approche RFA [15, 16]. Ces deux raisons ont ainsi encore plus orienté le projet vers l’utilisation de l’approche bayésienne, via les modèles bayésiens hiérarchiques. Un exemple d’utilisation de l’approche RFA est effectuée dans l’article de Mailhot *et al.* en 2007 [17], qui l’applique aux séries temporelles de précipitations de mai à octobre sur une période de 30 ans, dans la région du sud du Québec, sur une grille de 45×45 km. Un autre exemple se trouve dans l’article de Piras *et al.* paru en 2016 [18], qui étudie les précipitations extrêmes d’un bassin versant sur l’île de la Sardaigne.

Revenons à l’approche par inférence bayésienne. Elle voit son utilisation augmenter significativement au cours des années 90, avec l’impulsion par exemple du travail de Coles en collaboration avec Powell en 1996 [19]. Cette approche convient déjà mieux à ce projet, de par la possibilité d’ajouter des variables significatives, ainsi que la flexibilité pour construire la hiérarchie dans les problèmes. Un autre intérêt sous-jacent à cette structure de modèle est de pouvoir mettre en commun les paramètres d’un même modèle, et les estimer globalement plus efficacement. Par ailleurs, l’amélioration de la puissance de calcul numérique de ces dernières décennies a rendue possible la mise en oeuvre d’algorithmes demandant en ressources informatiques pour l’inférence bayésienne, via les algorithmes MCMC.

La source principale d’inspiration pour l’élaboration du modèle général de ce projet provient des articles récents de Jalbert *et al.*, parus en 2017 [20] et en 2022 [21]. Dans ces deux études, un modèle bayésien hiérarchique spatial est mis en place pour modéliser les précipitations extrêmes respectivement sur toute l’Amérique du Nord, ou seulement l’est du Canada. De plus, pour ces modèles, une structure de iGMRF est introduite pour caractériser la dépendance spatiale entre les cellules. La différence est que cette structure est soit utilisée pour modéliser les paramètres de la GEV dans l’article de 2017, soit pour modéliser les erreurs par régression dans l’article de 2022. Le modèle climatique régional canadien (*Canadian Region Climate Model*, CRCM) a été utilisé comme variable environnementale pour les deux études. La facilité d’accès aux ressources quant à ce modèle spatial (aussi bien numériques que physiques), ainsi que son adaptabilité, a donc naturellement orienté le projet de recherche dans le sens de cette construction.

Un autre exemple récent intéressant de modèle spatial provient de l’article de Sharkey & Winter (2019) [15]. Dans cet article, un modèle bayésien hiérarchique est construit pour estimer les précipitations extrêmes de la Grande-Bretagne. Une différence principale de ce modèle avec les articles précédents est l’utilisation d’une loi de Pareto généralisée (*Generalized Pareto Distribution*, GPD) à la place d’une GEV. Cette loi sert à modéliser une approche légèrement différente que les maxima par blocs : à la place, on va isoler les valeurs dépassant un certain seuil, et les modéliser selon cette loi GP ; logiquement, cette approche est nommée "*peaks over thresholds*". Le reste du modèle décrit dans l’article est assez similaire à ce projet de recherche, à part que la grille de la Grande-Bretagne dispose de cellules assez grande, opérant ainsi un découpage assez peu fin. Néanmoins, leur façon d’approcher la dépendance spatiale entre les cellules est très proche de celle de Jalbert *et al.* (2022) [21] : les paramètres de la GEV pour chaque case sont linéaires en fonction de variables, et les erreurs sont modélisés à l’aide d’un iGMRF, et suivent une loi normale conditionnée en fonction de leurs voisins respectifs. Le niveau des hyper-paramètres, lui, va un peu loin et utilise des lois Wishart-Inverse pour les matrices de structure des erreurs. Leur conclusion est que la dépendance

spatiale permet de diminuer les singularités spatiales dans les paramètres de la GEV ; en outre, le paramètre de forme est également inclus dans la dépendance spatiale, et varie à travers la Grande-Bretagne.

On peut également citer le travail de Rohrbeck & Tawn paru en 2021 [22], qui adopte une approche légèrement différente de la dépendance spatiale des extrêmes. S'ils utilisent toujours la méthode des *"peaks over thresholds"* couplée à une loi GPD, cette fois la paramétrisation spatiale va se faire par classification : leur espace d'étude sera découpé en J clusters, et la dépendance spatiale va se matérialiser sous la forme d'un conditionnement entre les seuils de dépassement. Ce conditionnement prend la forme d'une loi bêta, dont l'espérance conditionnellement aux clusters respectifs de deux stations sera exponentiellement décroissante vis-à-vis de leur distance, avec un facteur dans l'exponentielle dépendant de si les deux stations sont dans le même cluster ou non : ils définissent ainsi une famille de paramètres γ_i , pour $1 \leq i \leq J$, avec γ_0 le paramètre pour la loi conditionnelle entre deux stations situées dans deux clusters différents. Pour la pratique, ils emploient un algorithme de saut MCMC réversible. Leurs régions géographique d'étude sont les précipitations du Sud de la Norvège d'une part, et les débits de rivières au Royaume-Uni. Leurs résultats montrent notamment que pour ce deuxième jeu de données que les clusters obtenus au cours de l'algorithme, diffèrent des réseaux de rivières. En guise de conclusion, leur approche a permis de mieux appréhender la répartition spatiale des extrêmes.

Dans un registre similaire, on dispose également de l'article de Reich & Shaby [23], qui effectue une classification spatiale semblable à celle effectuée par Rohrbeck & Tawn [22], mais sur une grille fixée plutôt que stations par stations, comme dans l'étude de Sharkey & Winter [15]. Leur analyse de dépendance s'ajuste à l'aide de deux paramètres, α et ϕ . Elle porte sur toute l'Amérique du Nord, et leur modèle climatique utilisé est CLIMDEX-GHCNDEX [24].

D'autres modélisations spatiales des extrêmes peuvent se trouver dans la littérature récente, appliquées à divers zones géographiques : Wi *et al.*, en 2016 [25], se sont intéressés à modéliser les précipitations extrêmes de la Corée du Sud de manière non-stationnaires, via des lois GEV ou GPD, et regroupées en clusters géographiques. Cette année, Santinaque *et al.* [26] ont appliqué une classification spatiale par dendrogramme aux stations de relevé météorologiques de l'Uruguay, couplé à des ajustement de lois GEV classiques.

Pour ce qui est de la zone à étudier, si l'est du pays a été étudié à maintes reprises par d'autres membres du laboratoires Extremes à Polytechnique Montréal [20, 21, 27], ou dans d'autres instituts et universités basées à Montréal [17, 28], la côte Ouest l'a moins été. La province de la Colombie Britannique, notamment, représente l'entièreté de la côte, qu'elle ne partage qu'avec les états de l'Alaska et du Washington frontaliers. Par ailleurs, la présence

de la chaîne montagneuse des Rocheuses présente une différence significative par rapport au Québec et aux provinces voisines : la région dispose ainsi d’une plus importante disparité de relief.

Un élément que l’on peut utiliser dans l’élaboration d’un modèle spatial réside dans les copules. Sans rentrer dans les détails, une copule est un objet mathématique permettant de caractériser les dépendances entre coordonnées d’un vecteur aléatoire : c’est donc un objet qui se prête bien aux modèles spatiaux. Un exemple d’utilisation des copules se trouve dans l’article de Davidson *et al.* (2012) [29], qui emploie un modèle bayésien hiérarchique basé sur une copule pour la couche latente, visant à créer spatialement une grille des précipitations extrêmes de la Suisse. On peut aussi les retrouver dans l’étude de Palacios-Rodriguez *et al.*, publiée tout récemment en 2023 [28]. Elle se concentre sur l’utilisation de copules pour estimer les précipitations extrêmes du sud-est du Canada, sur une zone incluant le Québec, le Nouveau-Brunswick et l’Île du Prince-Édouard. Toutefois, malgré l’intérêt des copules, il a été décidé de ne pas les employer pour ce projet, de part leur complexité théorique apparente ; elles ne sont pas nécessaires, bien qu’elles représentent une base intéressante. Elles seront potentiellement utilisées dans des développements ou applications futurs du projet.

Une autre possibilité pour aller plus loin dans la modélisation des IDF est de prendre en compte les dépendances inter-durées entre les précipitations extrêmes. C’est la motivation principale du modèle dit de d -GEV présenté dans l’article de Koutsoyiannis (1998) [30], qui a servi de base à cette théorie. Les d -GEV ont par exemple été utilisées dans l’article de Lehmann *et al.* (2016) [31], dans un cadre non-stationnaire prenant en compte l’influence du changement climatique, le tout appliqué à la région côtière de Sydney. Toutefois, les dépendances inter-durées ne sont pas l’objectif de ce projet de recherche, mais c’est un axe de développement potentiel à considérer pour le futur.

La réanalyse servant de base à ce projet a été choisi pour être le modèle récent produit par la *Canadian Surface Prediction Archive* (CaSPAr), la version 2.1. du système régional de prévision déterministe (*Regional Deterministic Reforecasting System*, RDRS). Il est présenté de manière détaillée dans l’article de Gasset *et al.* paru en 2021 [32]. La production de ce modèle provient directement d’ECCC, qui souhaitait disposer d’un modèle climatique régional sur une grille de 10km, avec un découpage de simulation allant jusqu’à l’heure précise, et s’étendant sur une période de 1980 à 2018. L’article détaille notamment assez précisément le couplage avec les observations de surface, permettant ainsi à la fois l’assimilation au modèle ainsi que sa vérification une fois les simulations produites.

La base du système pour produire et analyser les précipitations réside au sein de la *Canadian Precipitation Analysis* (CaPA), un modèle produit et décrit par Mahfouf *et al.* en 2007 [33] et analysé par Lespinas *et al.* en 2015 [34]. Ce modèle, bien qu’il produit une analyse sur grille des précipitations, reste toujours par tranche horaire de 6 ou 24 heures. Cela servait toutefois d’un point de départ pour produire par la suite, couplé à un système de données de terrain nommé CaLDAS (pour *Canadian Land Data Assimilation System*), de produire RDRS. Les versions successives de ce dernier s’étendent simplement sur des périodes temporelles de plus en plus longues : la version 1 couvrait seulement les années 2010 à 2014, tandis que la version actuelle, 2.1, s’entend sur presque quarante ans, de 1980 à 2018 [35].

Depuis la publication de cette réanalyse par ECCC en 2017, RDRS a été utilisé dans plusieurs études. On peut notamment citer l’article d’Abaza *et al.*, paru en 2020 [36], qui modélise les prévisions hydrologiques pour le Lac Champlain et la rivière Richelieu, et emploie le RDRS à une résolution de 15 km. On peut également mentionner le travail imposant de Mai *et al.* en 2022 [37], qui cette fois produit des simulations hydrologiques pour l’ensemble des Grands Lacs.

Cependant, ces études utilisent RDRS principalement pour des études centrées sur des bassins hydrologiques, et n’a pas encore vu d’utilisation dans le domaine des valeurs extrêmes. Ce projet représente donc une bonne opportunité pour essayer ce nouveau modèle dans un contexte d’estimation de courbes IDF, et encore plus sur une province moins étudiée dans ce domaine, la Colombie-Britannique.

CHAPITRE 4 DONNÉES ET ANALYSE EXPLORATOIRE

Pour mettre en oeuvre le modèle permettant d’interpoler les courbes d’intensité-durée-fréquence de la Colombie Britannique, deux jeux de données ont été utilisés : d’une part des données observées issues de relevés de stations météorologiques, afin d’avoir des informations réelles sur le terrain ; d’autre part, des données issues de la réanalyse RDRS disponible sur une grille couvrant tout le territoire à l’étude, afin d’avoir des informations pour l’ensemble du territoire et interpoler plus efficacement les données réelles.

4.1 Données des stations

4.1.1 Carte des stations d’ECCC

Les données des stations ont été obtenues via ECCC (Environnement et Changement Climatique Canada), et se présentent sous la forme suivante : 134 stations réparties sur le territoire, dont on dispose comme variables environnementales les coordonnées (latitude & longitude) ainsi que l’élévation par rapport au niveau de la mer. Pour chacune de ces stations, on dispose d’un fichier .csv contenant les maximas annuels pour cette station, pour différentes durées de relevés : 5 minutes, 10 minutes, 15 minutes, 30 minutes, 1 heure, 2 heures, 6 heures, 12 heures et 24 heures ; ces maximas annuels sont en intensité, plus précisément en mm/h.

Le nombre d’années de relevé disponibles varie d’une station à l’autre : on peut avoir de 10 à 65 années de données enregistrées, comme visible dans l’histogramme en figure 4.1. La plupart des séries de données commencent autour de 1960 et 1970, et terminent en 2018/2019.

Comme la plupart des stations ont des relevés constants pour la durée 24 heures, et qu’il s’agit de la norme pour étudier les courbes IDF, l’analyse exploratoire se concentrera sur cette durée dans la suite. Les influences entre les différentes durées des courbes IDF n’est pas le sujet de ce projet de recherche, mais a été effectué cette année au sein du laboratoire Extremes.

La carte des stations a été créée sous la forme d’un Dashboard R, consultable sur Github à l’adresse suivante : <https://github.com/LeoGuillon/project-dashboard> ; la carte est également présentée à la figure 4.2.

On peut observer que la répartition des stations sur le territoire est assez inégale : on relève une grande concentration de stations au niveau de Vancouver et ses alentours (île de Victoria

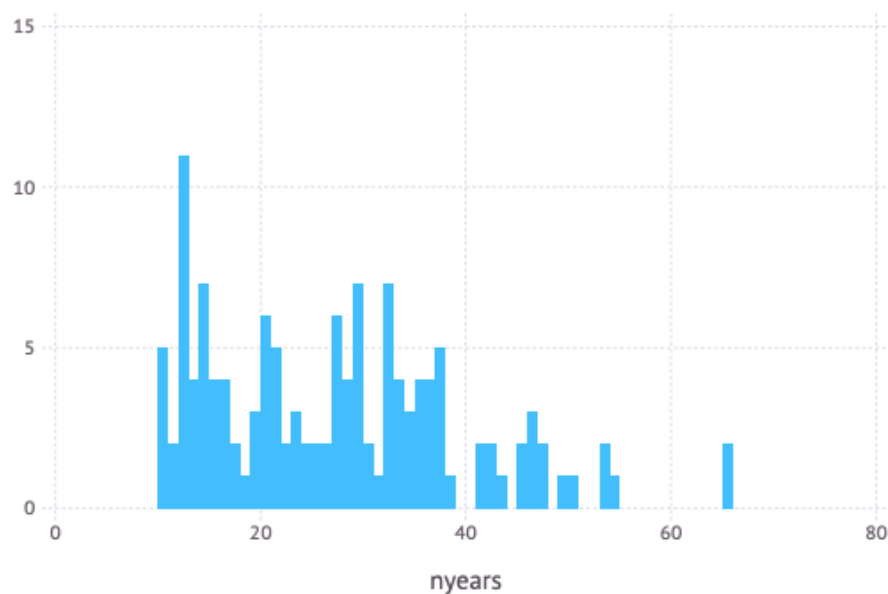


FIGURE 4.1 Histogramme du nombre d'années de relevé pour les stations d'ECCC en Colombie-Britannique

et sud des Rocheuses), et une répartition plus éparse à mesure que l'on progresse vers le nord de la province. L'élévation dûes à la chaîne des Rocheuses est par ailleurs clairement visible, et sera un point important dans la modélisation spatiale des précipitations.

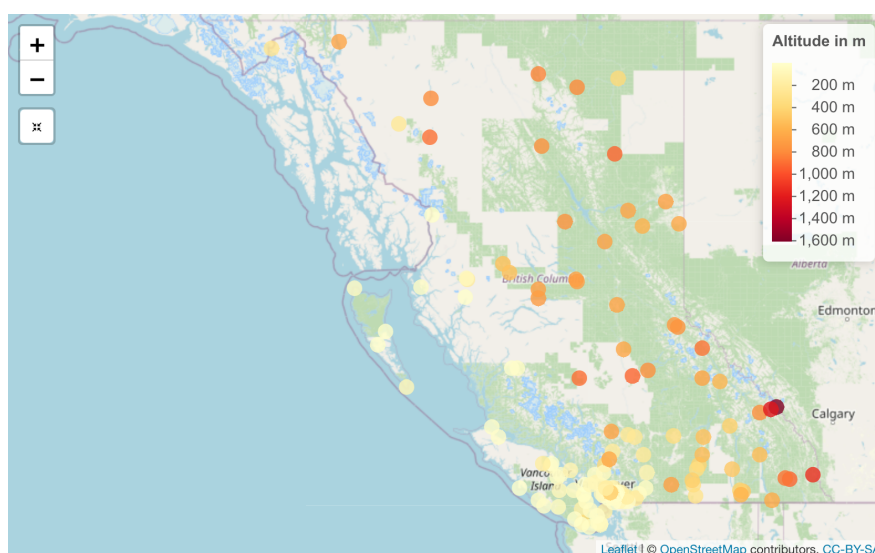


FIGURE 4.2 Carte de l'ensemble des stations d'ECCC en Colombie Britannique et de leur altitude respective

4.1.2 Séries temporelles et estimation GEV simple

Ensuite, les maxima annuels ont été extraits pour chacune des 134 stations, de sorte à avoir les séries temporelles annuelles pour les maxima à 24h de relevé. Pour chaque série temporelle, la librairie Julia `Extremes.jl` a été employé pour estimer directement les paramètres de la GEV (qui utilise la méthode de maximum de vraisemblance détaillée dans le chapitre Théorie). Cette estimation est effectuée dans le but d’avoir une base de paramètres, bien qu’elle soit naïve dans le sens où les stations sont estimées indépendamment les unes des autres.

Toujours est-il qu’en effectuant ce processus, on obtient les paramètres pour toutes les stations, à l’exception de deux d’entre elles où le nombre trop faible d’années de relevés ne permet pas l’estimation par maximum de vraisemblance, sortant ainsi une erreur.

Pour se donner une idée de la représentation de ces paramètres spatialement, on peut déjà les afficher sur la carte (sur le dashboard, il suffit de changer le paramètre affiché). Les cartes pour chaque paramètre sont visibles en figure 4.3.

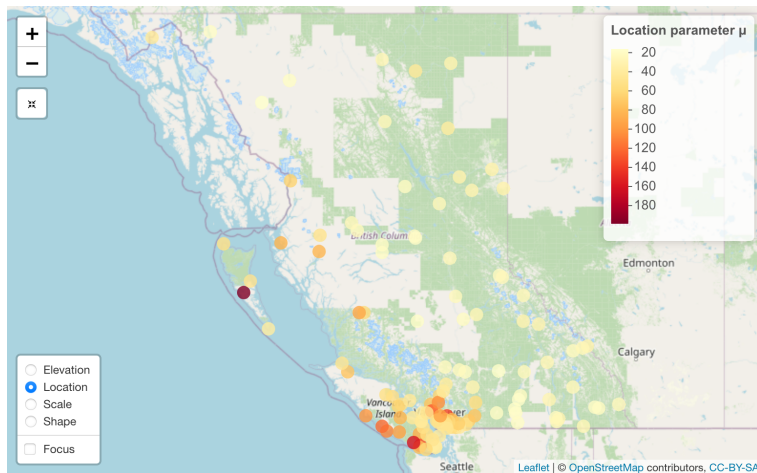
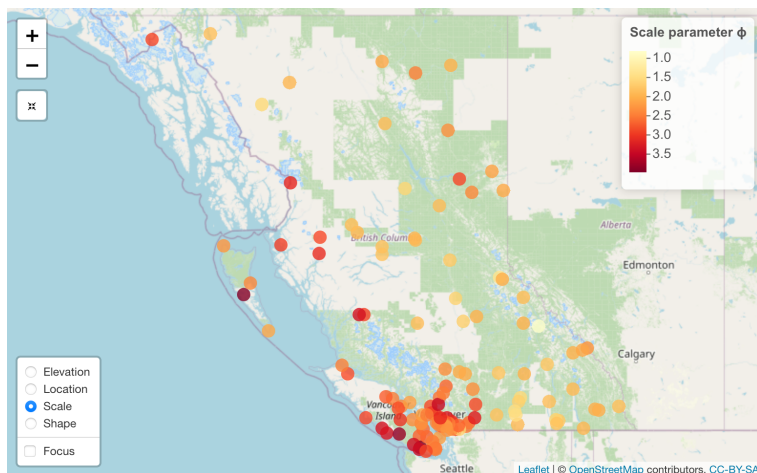
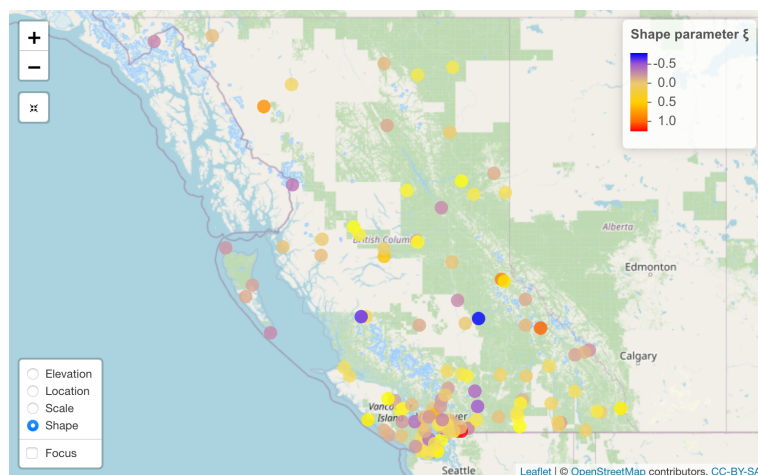
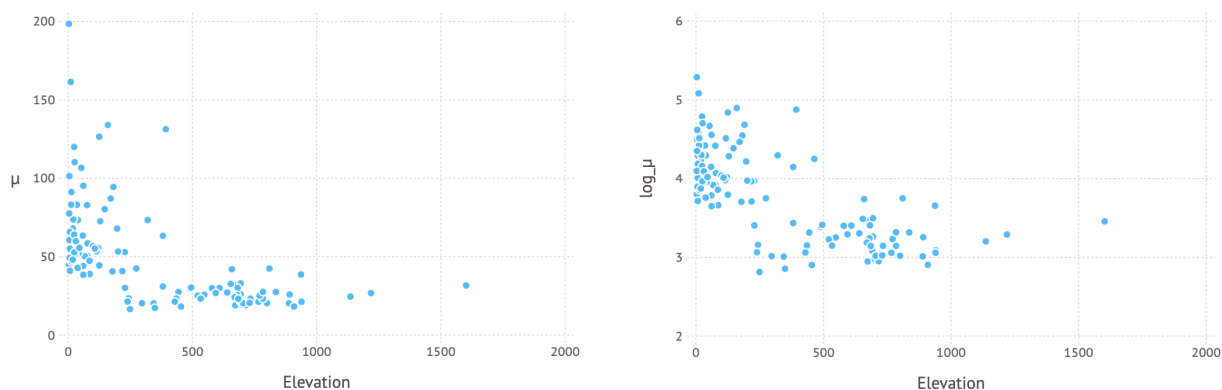
(a) Carte du paramètre de localisation μ (b) Carte du paramètre d'échelle logarithmique $\phi := \ln(\sigma)$ (c) Carte du paramètre de forme ξ

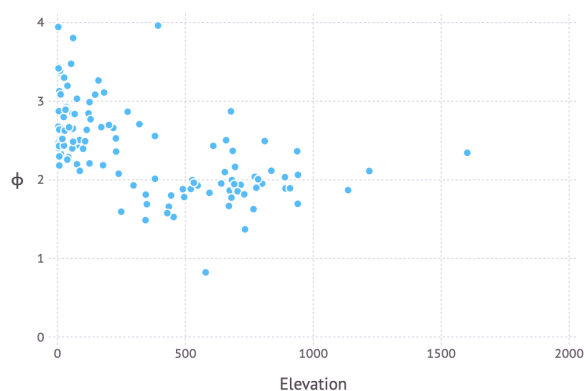
FIGURE 4.3 Cartes des paramètres de l'estimation naïve issue des données des stations d'ECCC de la Colombie Britannique

On peut déjà observer que pour les paramètres de localisation μ et d'échelle ϕ , les stations dans l'entourage de Vancouver et Victoria sont plus élevées que sur le reste du territoire ; par ailleurs, la zone située au sud-est de la province, couverte par les Rocheuses, semble avoir des paramètres de localisation et d'échelle plus faibles, de même que pour toute l'étendue dans l'est de la province, bordant l'Alberta. Ceci pourrait laisser sous-entendre que l'élévation (ou la distance à la mer si on lorgne du côté des variables environnementales auxquelles on n'a pas encore accès) soit une variable significative pour estimer les paramètres de la GEV spatialement. Pour confirmer cette impression, on peut afficher l'élévation en fonction des paramètres estimés, visible en figure 4.4.



(a) Graphe du paramètre de localisation μ en fonction de l'élévation

(b) Graphe du logarithme du paramètre de localisation μ en fonction de l'élévation



(c) Graphe du paramètre d'échelle ϕ en fonction de l'élévation

FIGURE 4.4 Graphes de relation entre l'élévation et les paramètres GEV des stations d'ECCC

On observe effectivement une relation linéaire ou logarithmique entre le paramètre de locali-

sation μ et l'élévation ; quant au paramètre d'échelle ϕ , une relation ne semble pas apparaître directement de manière visible, mais ce sera à discuter plus tard lors de la modélisation approfondie.

Pour revenir aux cartes, on peut observer que le paramètre de forme ξ ne semble pas avoir de répartition claire dans la province, au mieux on peut considérer une forme de continuité dans le sud de la province et à quelques autres endroits éparses. Ceci peut justifier d'avoir un paramètre de forme ξ unifié pour l'ensemble du territoire.

4.1.3 Stations et valeurs aberrantes

Pour le paramètre de forme, il fallait aussi vérifier que ses valeurs se trouvent dans l'intervalle $[-1, 1]$, et dans le meilleur des cas dans l'intervalle $[-0.5, 0.5]$: en effet, de part la forme de la loi GEV, afin qu'elle ait une moyenne et une variance, son paramètre de forme doit être dans cet intervalle. L'histogramme du paramètre de forme est visible en figure 4.5.

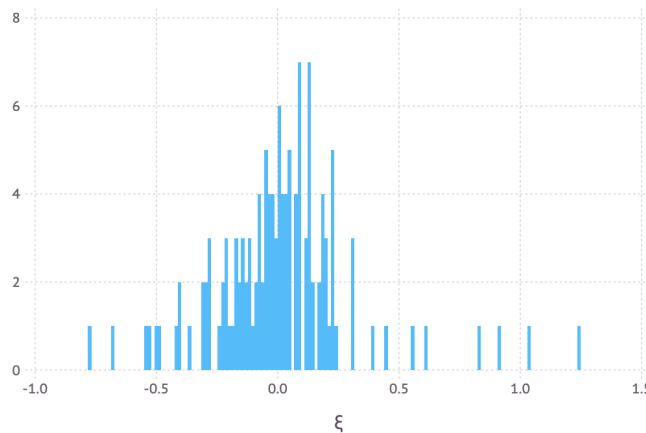
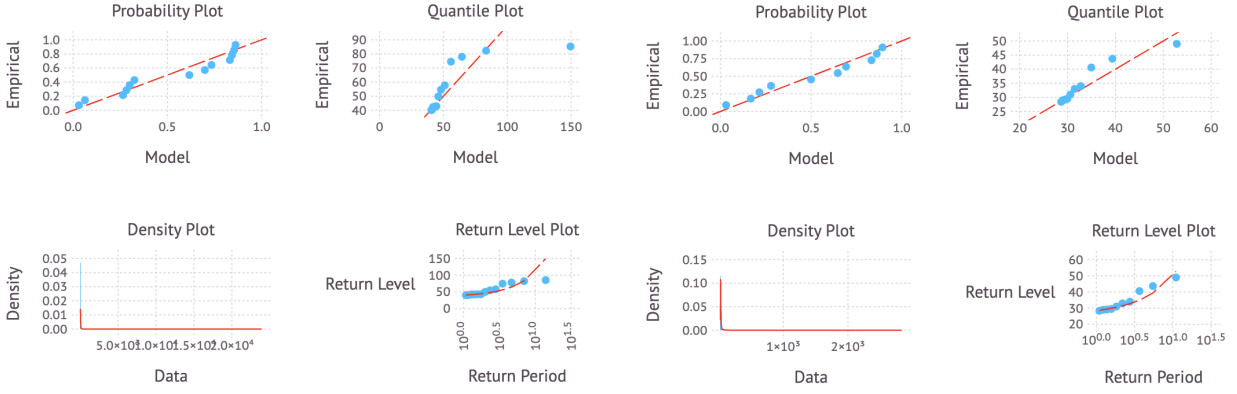


FIGURE 4.5 Histogramme du paramètre de forme ξ estimé pour les stations d'ECCC en Colombie-Britannique

On observe qu'il y a notamment deux stations qui ont un paramètre de forme supérieur à 1 : *Huntington Vye Road* et *Mica Dam*. On peut s'y intéresser de plus près, notamment en observant leurs graphes de diagnostic pour comprendre d'où vient cet écart. Les figures 4.6a et 4.6b sont composées de quatre éléments : les deux premiers sont les graphes probabilité-probabilité et quantile-quantile, pour comparer les fonctions de répartition ou les quantiles théoriques (comparée à la loi GEV estimée) avec leurs valeurs empiriques ; ensuite, on a un graphe de densité, et enfin le graphe des niveaux de retour.

Ce qui ressort sur ces graphes, c'est d'abord la présence de quantiles extrêmes s'écartant



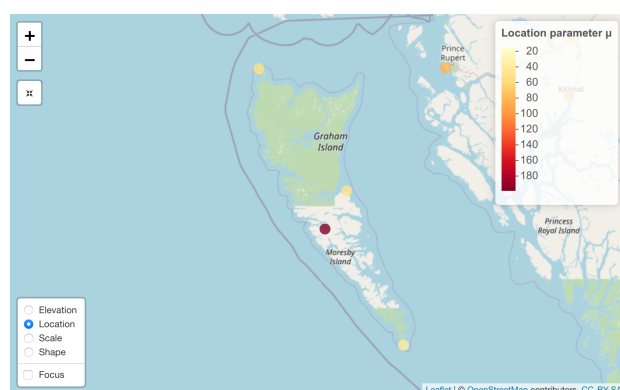
(a) Graphes de diagnostic pour la station "Huntington Vye Road"

(b) Graphes de diagnostic pour la station "Mica Dam"

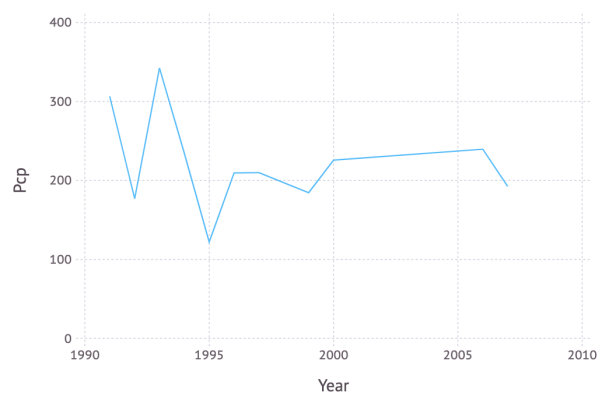
FIGURE 4.6 Graphes de diagnostic pour les stations d'ECCC ayant un $\xi > 1$

significativement de la droite, et résultant ainsi en un graphe de densité complètement déséquilibré avec une queue visiblement très (trop) lourde et avec un pic à peine visible. Il a donc été décidé de retirer ces deux stations des modélisations et analyses futures.

Par ailleurs, en observant plus en détail les cartes des paramètres, on peut relever une station qui a un paramètre de localisation μ bien plus élevé que les autres : la station *Moresby Island Mitchell Inlet* dispose en effet d'un $\mu \simeq 189$, ce qui est de loin la valeur la plus élevée de tout le territoire de la Colombie Britannique. Ce qui est curieux, c'est qu'elle est située non loin de trois autres stations, toutes sur les îles de Graham et de Moresby (en figure 4.7a), mais qu'elle a une série temporelle de maximas anormalement plus élevée que ses voisines et que même toutes les autres stations de la province, comme visible en figure 4.7b. Il a donc été décidé de retirer également cette station dans les modélisations et analyses futures.



(a) Carte des stations des îles de Graham & Moresby



(b) Série temporelle des maxima annuels de précipitation à la station Moresby Island Mitchel Inlet

FIGURE 4.7 Détails de l'aberration à l'île de Moresby

4.2 Grille des données réanalysées de RDRS 2.1

Bien qu’avec les données d’ECCC, on ait accès aux maximas aux stations de relevés, ces données ne concernent qu’une petite partie de l’entièreté de la province. Ainsi, pour avoir des covariables environnementales obtenues ou estimées partout ailleurs, un deuxième jeu de données bien plus massif a été employé, la réanalyse provenant de CaSPAR (*Canadian Surface Prediction Archive*) [38]. Leur produit de simulation des précipitations utilisé ici est RDRS, en version 2.1, permettant d’avoir des simulations allant de 1980 à 2018.

4.2.1 Description des données

Le jeu de données brut consiste en un ensemble de tables, une par journée, s’étalant du 1er janvier 1980 au 30 décembre 2018. Dans chacune des tables, on dispose de la grille des latitudes et longitudes de la zone étudiée, la période temporelle de la journée, ainsi que les simulations des précipitation en m/s sur chacun des points de la grille pour chaque heure de la plage horaire ; la particularité de cette dernière est qu’elle est glissante sur 24h, c’est à dire qu’elle va de midi d’une journée à 11AM du lendemain matin.

La zone étudiée pour la simulation ici est, sans surprise, la Colombie Britannique. Toutefois, comme il était plus pratique d’avoir une zone rectangulaire qui englobe l’entièreté de la province, il a fallu empiéter sur les régions voisines ; ainsi, on a dans la zone des légères portions de l’Alberta, du Yukon, des Territoires du Nord-Ouest, et même un peu des états d’Alaska et de Washington. Il en résulte une grille de 190 par 112 cases, chacune d’entre elle représentant un carré de 10 kilomètres de côté. La zone complète est présentée à la figure 4.8.

4.2.2 Extraction et cartes des simulations

Tout d’abord, il fallait remettre à l’échelle les données : elles sont fournies en m/s, mais il était plus convenable de les passer en mm/h : ainsi, il a seulement suffi de passer toutes les données par un facteur 3600.

Par la suite, il m’a fallu réduire ce gigantesque ensemble de données (de l’ordre de $190 \times 112 \times 24 \times 365 \times 38$) à une unique grille de covariables pouvant être utilisée d’un coup.

Ainsi, dans un premier temps, j’ai extrait la moyenne de précipitation pour chaque point de grille, sur l’ensemble des 38 années de simulation. Cette variable explicative est notamment employée dans le modèle de Jalbert et al. [21], et permet de rendre compte de l’influence des précipitations quotidiennes. L’influence des maximas, elle, peut-être incorporée par l’ajout

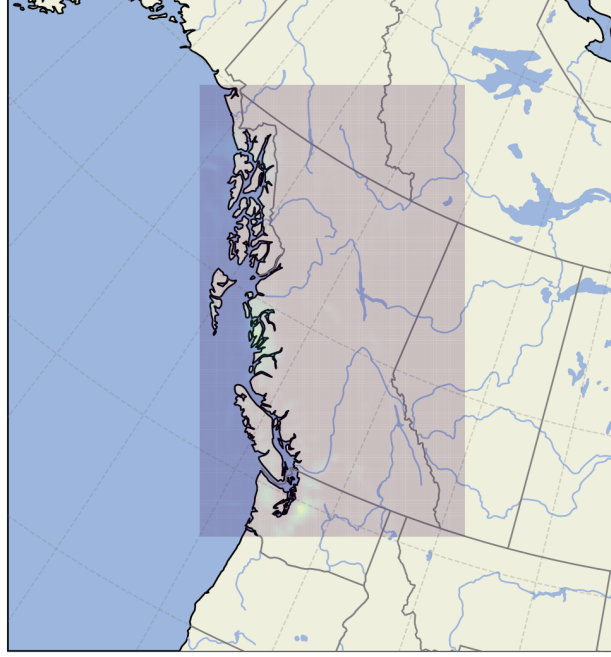
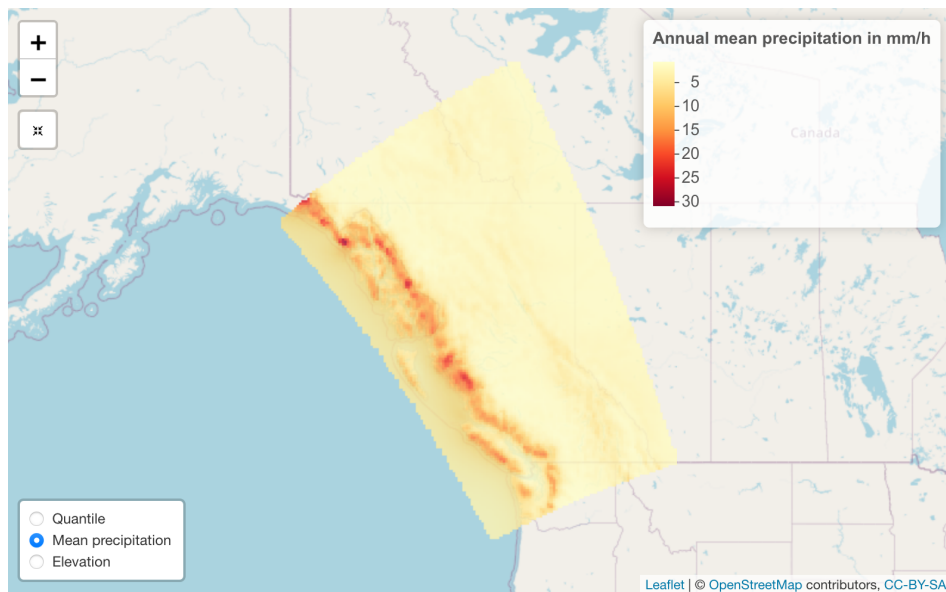


FIGURE 4.8 Zone couverte par la réanalyse RDRS 2.1

du maxima annuel pour chaque point de grille. Néanmoins, il peut être plus intéressant et plus précis de procéder directement à une estimation GEV pour chaque point de grille, indépendamment des autres.

On obtient ainsi, pour chaque point de grille $i \in [1, m]$, une série temporelle de maxima annuels $\mathbf{y}_i = (y_{i,t})_{1 \leq t \leq n_i}$ sur les 38 ans de simulation RDRS 2.1. On peut donc effectuer une estimation des paramètres GEV par maximum de vraisemblance pour chacune de ces séries. À l'issue de cette estimation, on obtient donc une nouvelle grille de variables, constituée des trois paramètres de la GEV ; j'ai ajouté une dernière variable, qui sera celle utilisée de préférence dans la suite : le niveau de retour à $T = 20$ ans, permettant ainsi de résumer en une seule variable l'information de la GEV. Le choix spécifique de la durée de 20 ans est motivé par sa nature de durée intermédiaire au sein de l'éventail des périodes de retour employées usuellement, s'échelonnant habituellement de 1 an à 100 ans.

Les cartes de ces deux valeurs sont présentées à la figure 4.9. Chacune de ces cartes est une couche d'informations, disponible dans l'onglet "Covariates" du dashboard interactif du projet.



(a) Grille de la moyenne de précipitation

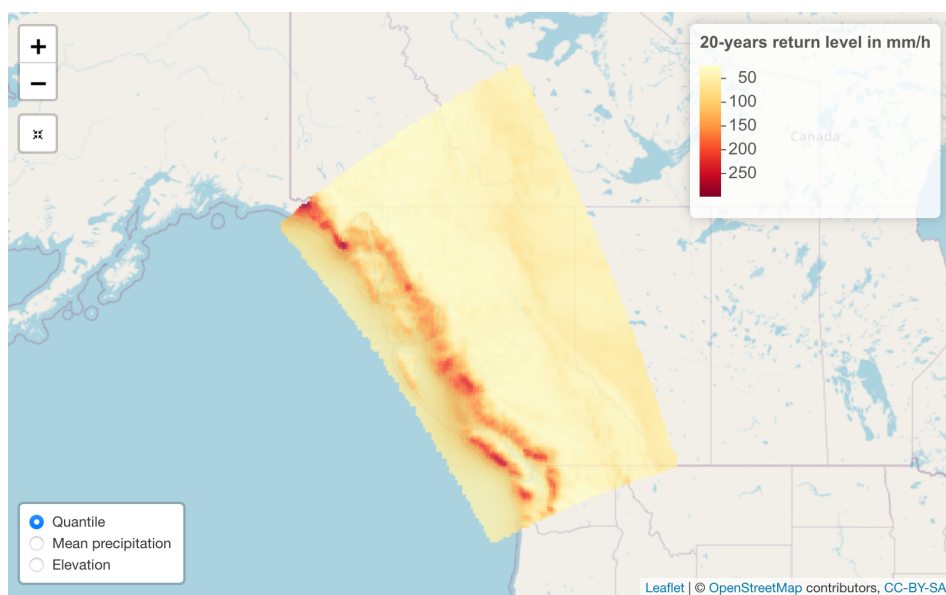
(b) Quantile à $T = 20$ années de retour

FIGURE 4.9 Cartes des variables environnementales obtenues grâce à la réanalyse RDRS 2.1

4.2.3 Carte de l'élévation

Une dernière variable estimée comme intéressante, suite à l'analyse exploratoire des stations d'ECCC, est l'altitude. Pour l'obtenir sur l'ensemble de la zone étudiée, j'ai utilisée une API d'obtention de cette grandeur, comme il en existe plusieurs disponibles sur le web. Celle choisie est *Open Meteo*, disponible en ligne [39]. En extrayant l'altitude en mètres pour

chaque point de grille, on obtient la carte présentée à la figure 4.10.

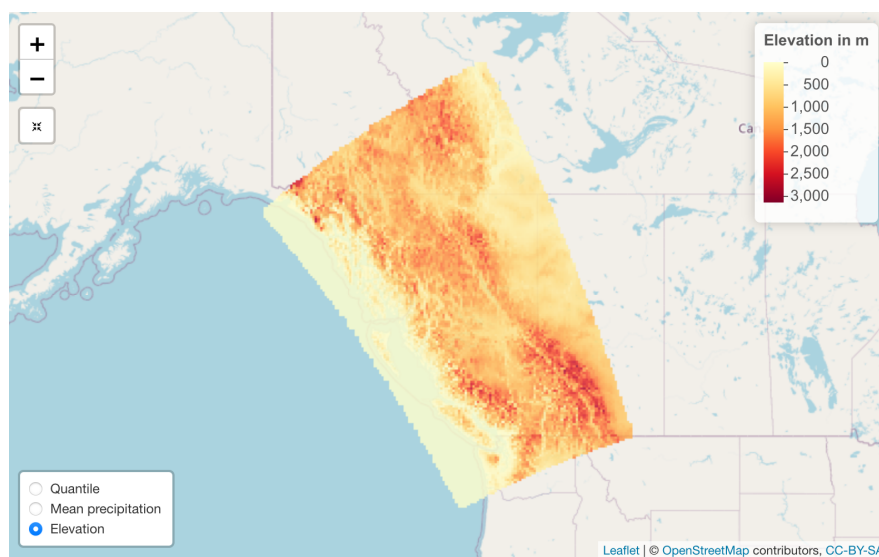


FIGURE 4.10 Carte de l'élévation sur la zone réanalysée par CaSPAr

4.3 Jonction des deux ensembles de données

Enfin, il fallait opérer une restriction sur les stations d'ECCC pour joindre les deux ensembles de données : en effet, certaines zones de la Colombie britannique étant particulièrement denses en stations, il fallait s'assurer qu'il n'y en ait pas plus d'une par cellule de la grille de CaSPAR.

Tout d'abord, pour savoir à quelle cellule appartient chaque station, on effectue simplement une recherche de plus proche voisin, les points de grille étant au centre de chaque cellule. Une fois que l'on a obtenu les cellules pour chaque station, on peut donc s'intéresser à savoir s'il y a des cellules avec au moins 2 stations à l'intérieur.

On trouve effectivement qu'il y a 8 cellules avec deux stations ou plus (il y en a une seule sur les huit qui dispose de trois stations, et aucune qui en dispose de quatre ou plus) ; le critère pour choisir les données utilisées sera simple : comme il y a une relativement grande disparité du nombre d'années de relevés d'une station à l'autre, ce sera donc suivant la taille des séries temporelles que les stations seront choisies sur une même cellule de grille.

Ainsi, avec sept cellules avec deux stations et une cellule à trois stations, on diminue le nombre de stations total à 120.

Il reste enfin une ultime étape : lors de notre modélisation, il faudra s'intéresser à quelques stations choisies arbitrairement, afin de consulter l'état des algorithmes en des points précis de notre zone étudiée. J'ai décidé d'en choisir quatre, chacune répartie (presque littéralement) aux quatres coins de la Colombie Britannique :

- *Vancouver UBC*, située sur la péninsule de la métropole de Vancouver, au milieu de la zone dense en stations du Sud-Ouest de la province ;
- *Duncan Lake Dam*, située dans les Rocheuses, au Sud-Est ;
- *Germansen Landing*, située en plein centre de la Colombie Britannique, dans une région bien moins dense en stations ;
- *Pleasant Camp*, située à l'extrême Nord-Ouest, au contact de l'Alaska.

Ces quatre stations peuvent être mises en valeur sur le dashboard en ajoutant la surcouche "*Focus*", visible également en figure 4.11.

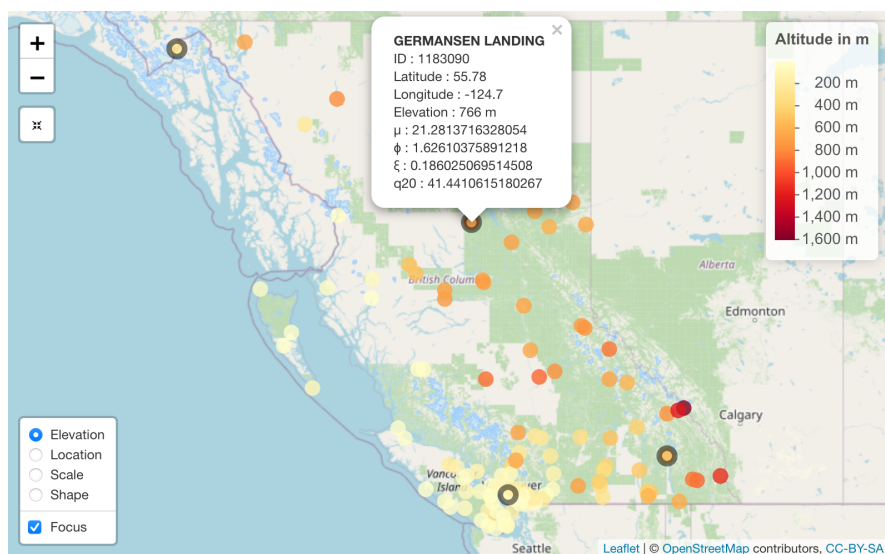


FIGURE 4.11 Carte des quatre stations d'ECCC choisies arbitrairement pour la modélisation

CHAPITRE 5 MODÉLISATION

Ce chapitre décrit le modèle statistique permettant d’interpoler les caractéristiques des précipitations extrêmes afin d’estimer les courbes IDF sur tout le territoire de la Colombie-Britannique. Dans un premier temps, la construction du modèle hiérarchique bayésien sera présentée, où la loi GEV sera utilisée pour la couche des données et le iGMRF dans la couche latente pour la modélisation de la dépendance spatiale. Ensuite, l’implémentation d’un algorithme Monte-Carlo par chaînes de Markov (*Monte-Carlo Markov Chains*, MCMC), brièvement évoqués au chapitre précédent, sera présenté afin de générer un échantillon aléatoire de la loi *a posteriori* des paramètres.

La construction de ce modèle hiérarchique bayésien est principalement adaptée des modèles proposés par Jalbert *et al.* (2022) [21] et Sharkey & Winter (2019) [15]. L’idée principale est de considérer que les paramètres de la loi GEV à chaque point de grille suivent une dépendance linéaire avec des variables explicatives et où l’erreur est spatialement dépendante. Marginalement, les erreurs sont distribuées selon la loi normale et conjointement par un champ de Markov gaussien intrinsèque.

5.1 Niveau des données

Tout d’abord, le premier niveau de ce modèle est consacrée à la modélisation des maxima annuels des intensités enregistrées aux 120 stations météorologiques. Puisque les variables explicatives se trouvent sur une grille, le modèle complet sera construit sur cette grille. Il faut alors attribuer observations aux stations aux cellules correspondantes de la grille. On dénote cette grille par la graphe simple $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, où $\mathcal{V}$ dénote l’ensemble des cellules et $\mathcal{E}$ dénote l’ensemble des arrêtes définissant le voisinage.

Comme on ne dispose d’observations empiriques qu’aux cellules disposant de stations de relevé, on va donc attribuer un sous-ensemble de ces cellules à des observations de précipitation. Au vu de la taille de la grille, le nombre de cellules disposant d’observations est dérisoire par rapport aux cellules sans observations : 120 stations retenues sur les 134 de départ, par rapport aux 190×112 cellules au total. Par ailleurs, comme détaillé dans le chapitre précédent, si une cellule contient plusieurs stations, on ne conserve que les observations de la station disposant du plus grand nombre d’années de relevé. Ce choix s’explique par la qualité de l’estimation de la loi GEV est d’autant meilleure que l’on a beaucoup d’années de relevés.

Dénotons par Y_{ij} le maximum annuel observé à la cellule $i \in \mathcal{V}$ et à l’année $1 \leq i \leq n_i$, où

n_i dénote le nombre d'années d'observation. Pour chaque cellule $i \in \mathcal{V}$, on suppose que le maximum Y_{ij} est distribué selon la loi GEV :

$$Y_{ij} \sim \mathcal{GEV}(\mu_i, \sigma_i, \xi); \quad (5.1)$$

où les paramètres de localisation et d'échelle dépendent de la cellule mais où le paramètre de forme est considéré spatialement constant. Ce choix est commun dans la littérature et les études préliminaires ne semblaient suggérer une variation spatiale de ce paramètre.

Dans ce modèle, la loi *a priori* des paramètres est décomposée en deux niveaux : le niveau latent et le niveau des hyperparamètres. C'est dans le niveau latent que la dépendance spatiale est modélisée.

5.2 Niveau latent

Pour le niveau latent, on effectue l'hypothèse que les paramètres de localisation et d'échelle suivent une loi linéaire en des variables explicatives. Plus précisément, on pose que, pour tout point de grille $i \in \mathcal{V}$:

$$\begin{cases} \mu_i &= \mathbf{x}_{ui}\beta_u + U_i \\ \phi_i &= \mathbf{x}_{vi}\beta_v + V_i \end{cases} \quad (5.2)$$

où $\mathbf{x}_{ui}$ (resp. $\mathbf{x}_{vi}$) correspond au vecteurs des variables explicatives pour le paramètre de localisation (resp. pour le paramètre d'échelle) de la cellule $i \in \mathcal{V}$, β_u & β_v sont les vecteurs de régression, et U_i & V_i sont les erreurs aléatoires spatialement dépendantes.

La différence avec les modèles précédemment développés par Jalbert [20] [21] ou Sharkey & Winter [15], est que l'on prend en compte *plusieurs* variables explicatives, contrairement au modèle de Jalbert *et al* [21]. Il sera possible de prendre des variables explicatives différentes pour les deux paramètres de la GEV.

Les erreurs aléatoires sont modélisées conjointement par un iGMRF d'ordre 1, afin d'incorporer la dépendance spatiale au modèle. Concrètement, l'ensemble des arêtes du graphe $\mathcal{E}$ relient les cellules immédiatement adjacentes de la grille d'étude décrite précédemment. Formellement, on peut écrire :

$$E = \{(i, j) \in \mathcal{V}^2 \mid \|i - j\| = 1\}.$$

C'est une structure canonique pour les iGMRF d'ordre comme décrit dans l'ouvrage Rue &

Held [12] et dans le chapitre précédent.

À partir de ce graphe, on peut donc déduire la matrice de structure Q du iGMRF, qui sera simplement définie de la façon suivante :

$$Q_{i,j} = \begin{cases} d_i & \text{si } i = j \\ -1 & \text{si } i \neq j \text{ et } (i,j) \in E \\ 0 & \text{sinon,} \end{cases}$$

pour tout $(i,j) \in \mathcal{V}^2$ et où d_i est le nombre de voisins de la cellule d'indice $i \in \mathcal{V}$. Le nombre de voisins est 4 pour les cellules intérieures mais est plus petit au bordure de la grille. La densité conjointe des erreurs de régression $\mathbf{U} = (U_i : i \in \mathcal{V})$ s'exprime sous la forme suivante :

$$f_{(\mathbf{U}|\kappa_u)}(\mathbf{u}) \propto \kappa_u^{(m-1)/2} \exp\left(-\frac{\kappa_u}{2} \mathbf{u}^\top Q \mathbf{u}\right),$$

où κ_u est le paramètre de précision du champ aléatoire. Pour ce modèle, on suppose que le paramètre de précision est constant pour toute la zone étudiée, cela supposant que la dépendance spatiale entre cellules n'est due qu'à leur position relative, i.e. si elles sont voisines, et non pas due à leur localisation dans la province (qu'elles soient en région côtière ou montagneuse par exemple). Ce point peut être un axe de développement futur.

Conditionnellement aux erreurs observées aux autres cellules ($\mathbf{U}_{-i} = \mathbf{u}_{-i}$) des autres cellules, la densité de l'erreur à la cellule $i \in \mathcal{V}$ est la suivante :

$$f_{(U_i|\mathbf{U}_{-i}=\mathbf{u}_{-i})}(u_i) = \mathcal{N}\{u_i|\bar{u}_i, (\kappa_u W_{ii})^{-1}\},$$

où $\mathcal{N}(x|\mu, \sigma^2)$ dénote la valeur de la fonction de densité de la loi normale de moyenne μ et de variance σ^2 évaluée à x et où

$$\bar{u}_i = -\frac{1}{W_{ii}} \sum_{i \neq j} W_{ij} u_j = -\frac{1}{W_{ii}} \sum_{(i,j) \in E} u_j.$$

L'erreur conditionnelle ($U_i | \mathbf{U}_{-i} = \mathbf{u}_{-i}$) est donc centrée à la moyenne des erreurs des cellules voisines.

On écrit des expressions similaires pour les V_i , avec leur propre paramètre de précision κ_v .

5.3 Niveau des hyperparamètres

On dispose de plusieurs hyperparamètres dont il faut préciser les lois de probabilités, et ainsi compléter le modèle bayésien hiérarchique.

Tout d'abord, pour les paramètres de précision κ_u et κ_v , on choisit d'utiliser comme loi *a priori* la loi gamma vague suivante :

$$f_{\kappa_u}(\kappa_u) = \mathcal{G}(\kappa_u|1, 1/100),$$

où $\mathcal{G}(x|\alpha, \beta)$ dénote la valeur de la fonction de densité d'une loi gamma de paramètres α et β évaluée en x [40].

Pour le paramètre de forme, que pour rappel, on considère comme commun à toute la zone étudiée, on utilise la loi *a priori* proposée par Martins & Stedinger (2000) [9] :

$$f_{\xi}(\xi) = \mathcal{B}(\xi + 0.5|6, 9),$$

où $\mathcal{B}(x|a, b)$ est la fonction de densité de la loi bêta de paramètres (a, b) évaluée en x [41]. Le choix de cette distribution s'explique par le fait qu'on souhaite avoir le paramètre de forme compris entre les valeurs ± 0.5 afin que la loi GEV aient une moyenne et une variance finies.

Enfin, pour les paramètres de régression β_u et β_v , on fait le choix d'une loi *a priori* uniforme :

$$\begin{cases} f_{\beta_u}(\beta_u) & \propto 1 \\ f_{\beta_v}(\beta_v) & \propto 1 \end{cases}$$

Le choix de ces distributions finalement n'importe qu'assez peu, au vu du nombre de stations constituant le modèle global final, ce qui va fortement diluer l'influence de ces lois *a priori*, pour le niveau des hyper-paramètres.

5.4 Modèle complet

En combinant toutes les lois *a priori* présentées précédemment, on obtient finalement les lois conditionnelles complètes suivantes, pour $i \in \mathcal{V}$ contenant une station de relevé :

$$\begin{cases} f_{(\mu_i|\cdot)}(\mu_i) & \propto [\Pi_{j=1}^{n_i} \text{GEV}(M_{ij}|\mu_i, e^{\phi_i}, \xi)] \times \mathcal{N}(\mu_i|\bar{\mu}_i, (\kappa_u W_{ii})^{-1}) \\ f_{(\phi_i|\cdot)}(\phi_i) & \propto [\Pi_{j=1}^{n_i} \text{GEV}(M_{ij}|\mu_i, e^{\phi_i}, \xi)] \times \mathcal{N}(\phi_i|\bar{\phi}_i, (\kappa_v W_{ii})^{-1}), \end{cases}$$

où, par ajustement des lois des paramètres avec la normalité des erreurs, on a :

$$\begin{cases} \bar{\mu}_i &= X_{ui}\beta_u - \frac{1}{W_{ii}} \sum_{(i,j) \in E} u_j \\ \bar{\phi}_i &= X_{vi}\beta_v - \frac{1}{W_{ii}} \sum_{(i,j) \in E} v_j \end{cases}$$

La forme de ces lois conditionnelles complètes ont été calculées en associant d'une part les lois des données, supposées suivre des GEV, puis les lois des paramètres, associées avec les lois Normales des erreurs. Les variables comprises dans le symbole $\cdot$ sont donc $(\mu_{-i}, \phi_{-i}, \xi, \beta_u, \beta_v, \kappa_u, \kappa_v)$.

Finalement, voici un résumé plus visuel de ce modèle :

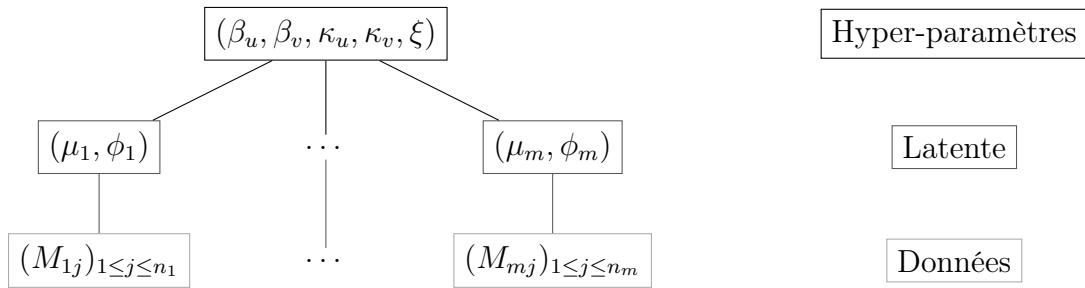


FIGURE 5.1 Schéma du modèle bayésien hiérarchique

5.5 Implémentation d'un algorithme MCMC

En pratique, nous allons utiliser une combinaison de Gibbs et Métropolis-Hastings pour notre modèle ; en effet, nous allons pour l'ensemble des stations de relevé, générer des candidats pour les paramètres de localisation et d'échelle à la manière de Métropolis-Hastings, mais nous allons les simuler conditionnellement l'un après l'autre, comme pour Gibbs ; nous allons également simuler le paramètre de forme à la manière de Métropolis-Hastings. Les hyper-paramètres $\beta_u, \beta_v, \kappa_u, \kappa_v$ seront calculés à chaque itération. Le nombre d'itérations choisi est de $NITER = 50000$ avec un temps de chauffe de $WARMUP = 5000$. Pour alléger l'affichage des chaînes et faciliter leur traitement, elles ont également été réduites (appelé en anglais "*thinning*" de la chaîne), en ne gardant qu'un élément sur 10. L'algorithme complet est présenté dans l'encadré Algorithme 3.

Une fois cet algorithme effectué, nous avons donc des chaînes de Markov pour nos paramètres et hyperparamètres, mais seulement pour les stations de relevé météorologiques d'ECCC. Il faut donc terminer par l'étape d'interpolation, permettant d'avoir les paramètres sur toute la grille. Cette étape d'interpolation s'effectue à l'aide de l'algorithme décrit dans l'ouvrage de

Rue & Held [12], se basant sur une décomposition de Cholesky. Une fois ceci effectué, on peut enfin obtenir les lois GEV sur l'ensemble du territoire, et en déduire ainsi leurs quantiles, et enfin obtenir les courbes IDF.

Algorithme 3 : Algorithme complet en pratique

Données : Matrices des covariables significatives $\mathbf{X}_u$ & $\mathbf{X}_v$, maximas de précipitations aux stations $\mathbf{Y}$

Résultat : Chaînes de Markov des $(U_i, V_i, \beta_u, \beta_v, \kappa_u, \kappa_v, \xi)$

```

;
Calcul des valeurs initiales des paramètres  $(\mu_i, \phi_i)$  en fittant une loi Gumbel sur
chaque station ;
Calcul des valeurs initiales des  $(\beta_u, U_i, \beta_v, V_i)$  par régression linéaire ;
Génération des valeurs initiales de  $(\kappa_u, \kappa_v)$  ;
Création des iGMRF  $F_u$  &  $F_v$  de taille  $n_{lat} \times n_{lon}$ , de paramètre de précision
respectif  $\kappa_u, \kappa_v$  ;
pour  $j \in [2, NITER]$  faire
    ;
    ;
    Génération du candidat  $\tilde{U}$  ;
    Calcul du vecteur  $\tilde{\mu}$  ;
    Calcul par régression linéaire de  $\beta_u^j$  ;
    Mise à jour du iGMRF latent  $F_u$  avec le candidat  $\tilde{u}$ , conditionnellement à  $\phi^{j-1}$ 
    et  $\xi^{j-1}$  ;
    Calcul et ajout aux chaînes des nouvelles valeurs  $U_i, \beta_u$  &  $\kappa_u$  ;
    ;
    ;
    Génération du candidat  $\tilde{V}$  ;
    Calcul du vecteur  $\tilde{\phi}$  ;
    Calcul par régression linéaire de  $\beta_v^j$  ;
    Mise à jour du iGMRF latent  $F_v$  avec le candidat  $\tilde{V}$ , conditionnellement à  $\mu^j$  et
     $\xi^{j-1}$  ;
    Calcul et ajout aux chaînes des nouvelles valeurs  $V_i, \beta_v$  &  $\kappa_v$  ;
    ;
    ;
    Génération du candidat  $\tilde{\xi}$  ;
    Calcul du taux d'acceptation  $\alpha$  comme différence des log-vraisemblance,
    conditionnellement aux  $\mu^j$  &  $\phi^j$  ;
    Simuler  $r \leftarrow \mathcal{U}[0, 1]$  ;
    si  $\alpha > r$  alors
        |  $\xi^j \leftarrow \tilde{\xi}$  ;
    sinon
        |  $\xi^j \leftarrow \xi^{j-1}$  ;
    fin
fin

```

CHAPITRE 6 RÉSULTATS

Ce chapitre présente et compare plusieurs variations du modèle hiérarchique bayésien décrit au chapitre précédent. Les variations concernent les choix des variables explicatives dans le niveau latent du modèle (équations 5.2). Le choix des variables explicatives pertinentes sera basé sur celles qui possèdent le plus grand pouvoir prédictif en interpolation. Nous verrons dans un premier temps ce que donnent les modèles unidimensionnels, afin d’observer l’intérêt d’utiliser la réanalyse RDRS ; dans un deuxième temps, les modèles multidimensionnels seront estimés, pour enfin dresser plusieurs critères de comparaison entre ces modèles afin de dégager la ou les variables significatives les plus pertinente(s) dans le cadre de cette étude.

Comme décrit dans le chapitre sur les données, les trois variables explicatives potentielles retenues pour ce projet sont, pour chaque point de grille :

- la moyenne des précipitations de la réanalyse RDRS 2.1 ;
- le quantile associé à la période de retour de 20 ans estimée à partir des des précipitations de la réanalyse RDRS 2.1 en ajustant une loi GEV à la série des maxima ;
- l’élévation par rapport au niveau de la mer (obtenue via l’API Open Meteo [39]).

D’autres variables explicatives auraient aussi pu être étudiée, notamment des quantiles associé à d’autres périodes de retour, la distance à la côte (qui a déjà été utilisée en Norvège par Dyrørdal *et al.* [42] par exemple). Or, cette dernière variable est moins facilement accessible que l’altitude et finalement assez similaire dans le cas précis de la Colombie-Britannique.

6.1 Modèles unidimensionnels

Dans cette section, l’utilité de la réanalyse RDRS 2.1 comme variable explicative spatiale est évaluée. Le modèle spatial sera ajusté en utilisant la moyenne des précipitations réanalysées à chaque cellule et ensuite le quantile correspondant à la période de retour de 20 ans. Le modèle utilisant la moyenne est dénoté par M_1 et celui utilisant le quantile par M_2 .

6.1.1 Modèle utilisant la moyenne comme variable explicative

Le modèle utilisant la moyenne des précipitations réanalysées comme variable explicative pour les paramètres de localisation et d’échelle a été ajusté. L’échantillonnage de Gibbs a été utilisé pour générer un échantillon aléatoire de la loi *a posteriori* des paramètres. Une chaîne de 50 000 itérations a été générées où les 5000 premières itérations de la phase de chauffe ont été retirées, et un élagage ("*thinning*") de 10 a été effectué pour alléger les chaînes. La

figure 6.3 illustre les traces des paramètres de régressions des chaînes obtenues et des erreurs spatiales des quatre stations témoins. On constate que la chaîne exhibe une stationnarité et un bon mixing, ce qui est souhaité. On constate également que la moyenne des précipitations réanalysées est significative car la valeur 0 n'est pas incluse dans les intervalles de crédibilité à 95% des coefficients de régression β_u et β_v , comme présenté en figure 6.2a.

Les figures 6.1a et 6.1b illustrent la carte des quantiles associés à la période de 20 ans estimés avec le modèle spatiale utilisant la moyenne des précipitations réanalysées, en échelle linéaire ou logarithmiques. On constate une grande dépendance spatiale des estimations et des valeurs très élevées.

6.1.2 Modèle utilisant le quantile comme variable explicative

À l'instar de la section précédente, la modèle spatial a été ajusté en utilisant le quantile associé à la période de retour de 20 ans des précipitations réanalysées. La figure 6.4 illustre la trace des coefficients de régressions et des erreurs spatiales aux cellules des quatre stations témoins. Encore une fois, la chaîne exhibe une stationnarité et un bon mixing.

Le quantile semble être significative comme variables explicative spatiale puisque la valeur 0 n'est pas incluse dans les intervalles de crédibilité à 95%, comme illustré en figure 6.2b. La figure 6.1c illustre les estimations des quantiles associés à la période de retour de 20 ans. On constate la dépendance spatiale semblable au modèle précédent mais les estimations diffèrent quelque peu. Cela peut être causée par l'incertitude supérieure de l'estimation du quantile de la variable explicative par rapport à l'estimation de la moyenne. Aussi, les deux variables explicatives ne mesurent pas la même chose. La moyenne est une mesure de localisation et le quantile associé à la période de retour de 20 ans mesure les extrêmes.

6.1.3 Conclusions sur les modèles unidimensionnels

Les deux modèles unidimensionnels ont permis de constater que les précipitations de la réanalyse RDRS 2.1 peuvent être utilisées comme variable explicative spatiale. Il s'agit maintenant de déterminer la combinaison de ces variables qui maximise le pouvoir prédictif d'interpolation des caractéristiques des précipitations extrêmes observées.

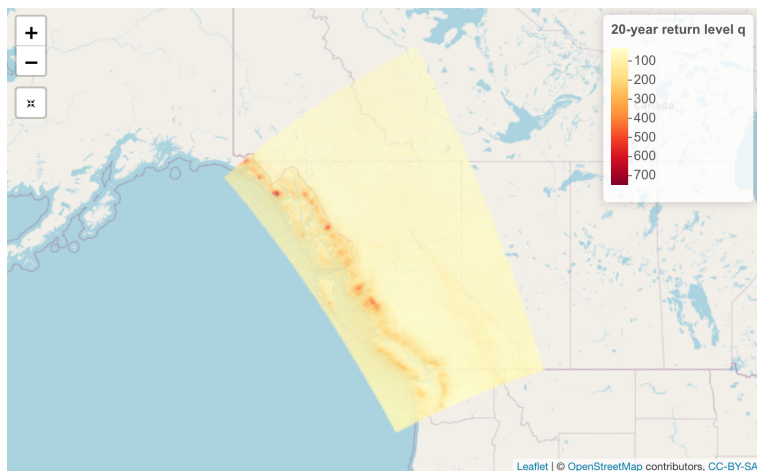
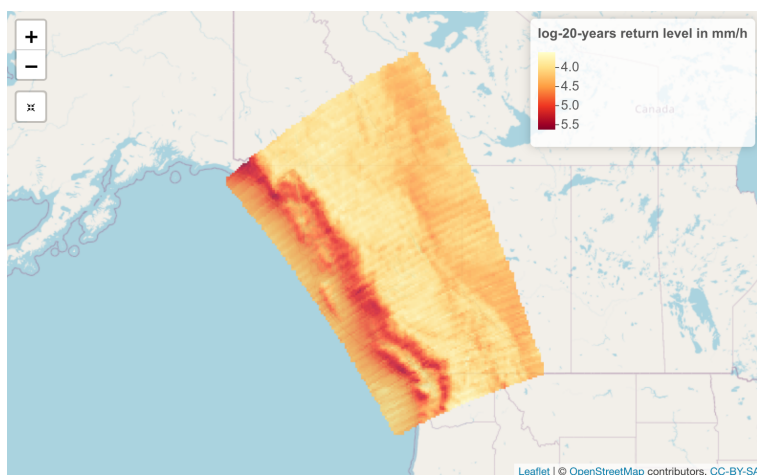
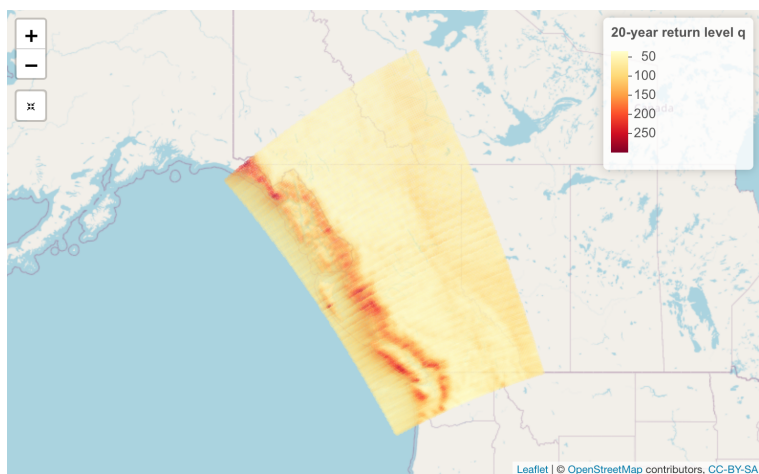
(a) Modèle M_1 (b) Modèle M_1 , échelle logarithmique(c) Modèle M_2

FIGURE 6.1 Cartes du quantile à $T = 20$ années de retour de la GEV estimée par les modèles unidimensionnels

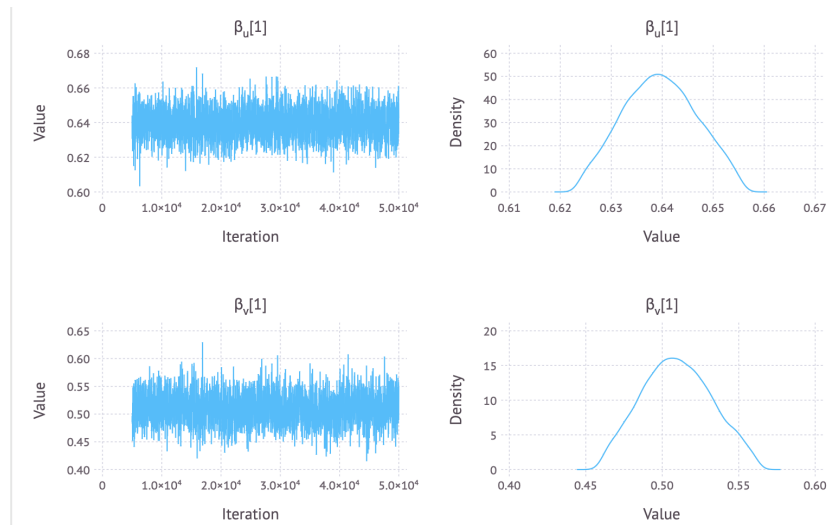
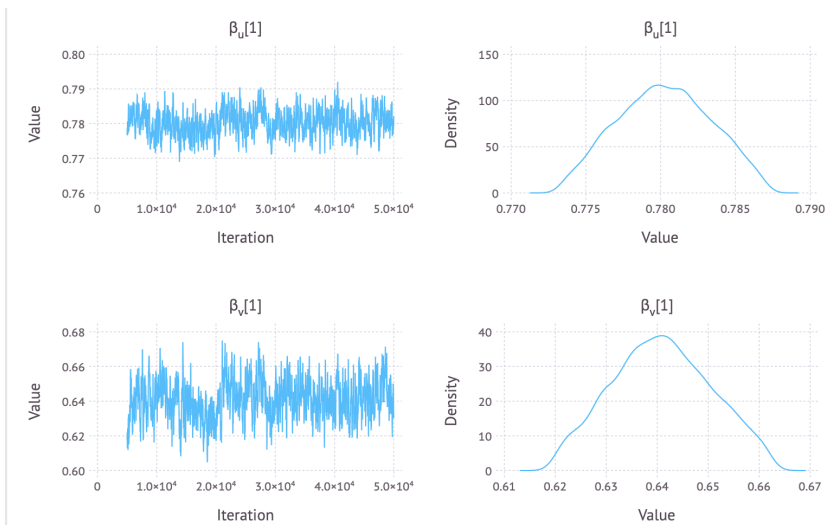
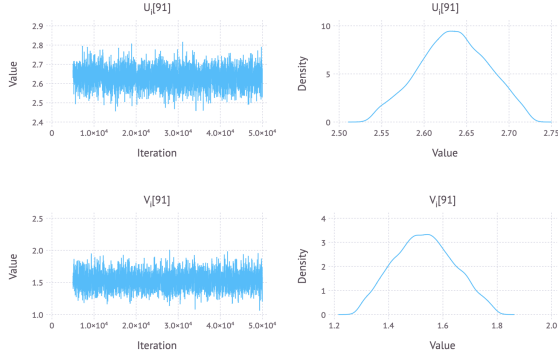
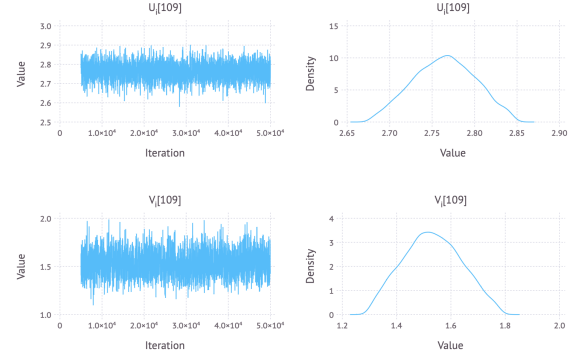
(a) Modèle M_1 (b) Modèle M_2

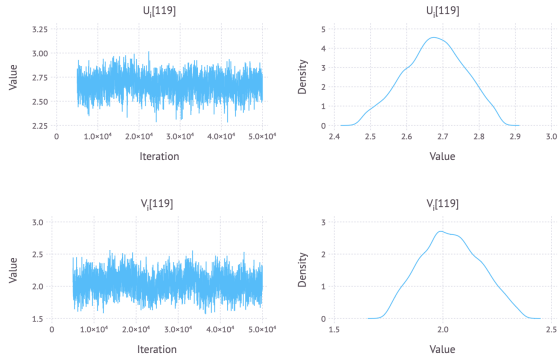
FIGURE 6.2 Traces et intervalles de crédibilité empiriques à 95% pour les coefficients de régression des modèles unidimensionnels



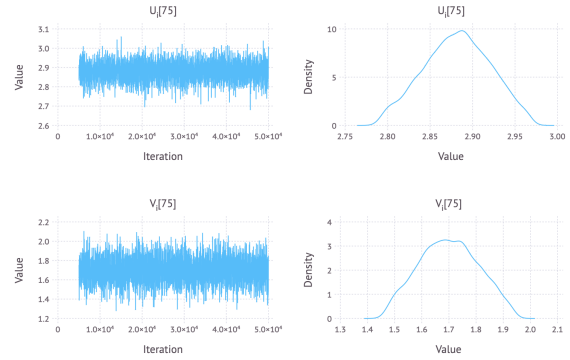
(a) Duncan Lake Dam



(b) Germansen Landing

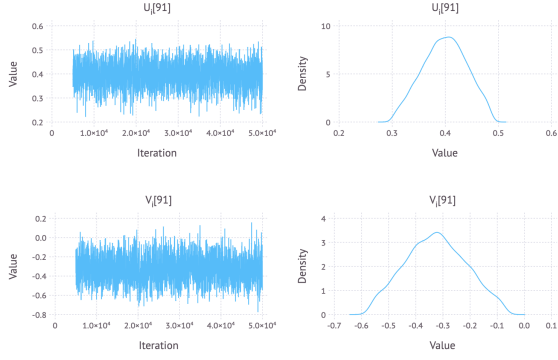


(c) Pleasant Camp

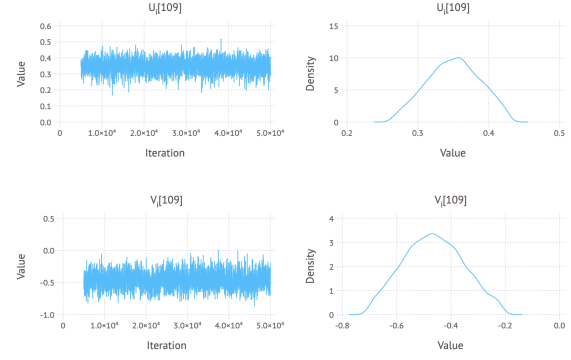


(d) Vancouver UBC

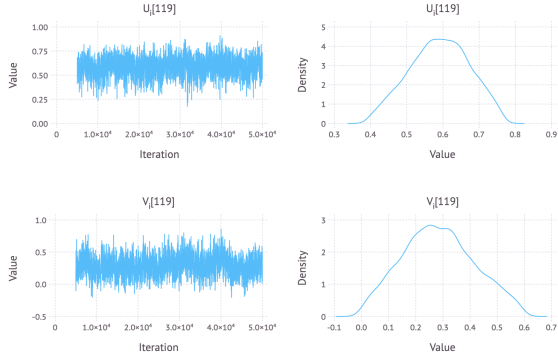
FIGURE 6.3 Traces des stations témoins pour le modèle M_1



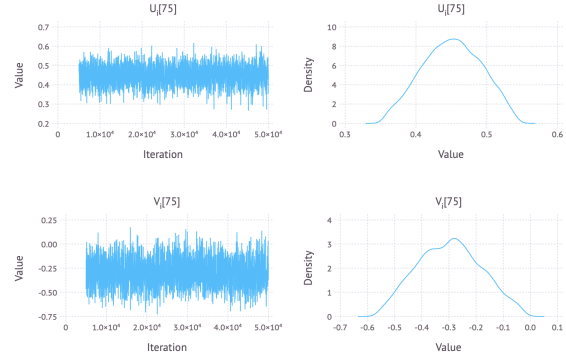
(a) Duncan Lake Dam



(b) Germansen Landing



(c) Pleasant Camp



(d) Vancouver UBC

FIGURE 6.4 Traces des stations témoins pour le modèle M_2

6.2 Modèles multidimensionnels

Dans cette section, les variables explicatives ayant le plus grand pouvoir prédictif sont recherchées. Une procédure de validation croisée sera utilisée pour identifier ces variables explicatives.

6.2.1 Modèle complet et variables pertinentes

Le modèle complet, noté M_3 dans la suite, inclut donc les trois variables disponibles : la moyenne des précipitations réanalysées, le quantile associé à la période de retour de 20 ans de la GEV estimée avec les maxima annuels des précipitations réanalysées et l'altitude moyenne de la cellule. La moyenne de la loi prédictive du quantile associé à la période de retour de 20 ans est affichée pour toutes les cellules du territoire à la figure 6.5. Les traces pour les stations témoins sont disponibles quant à elles en annexe.

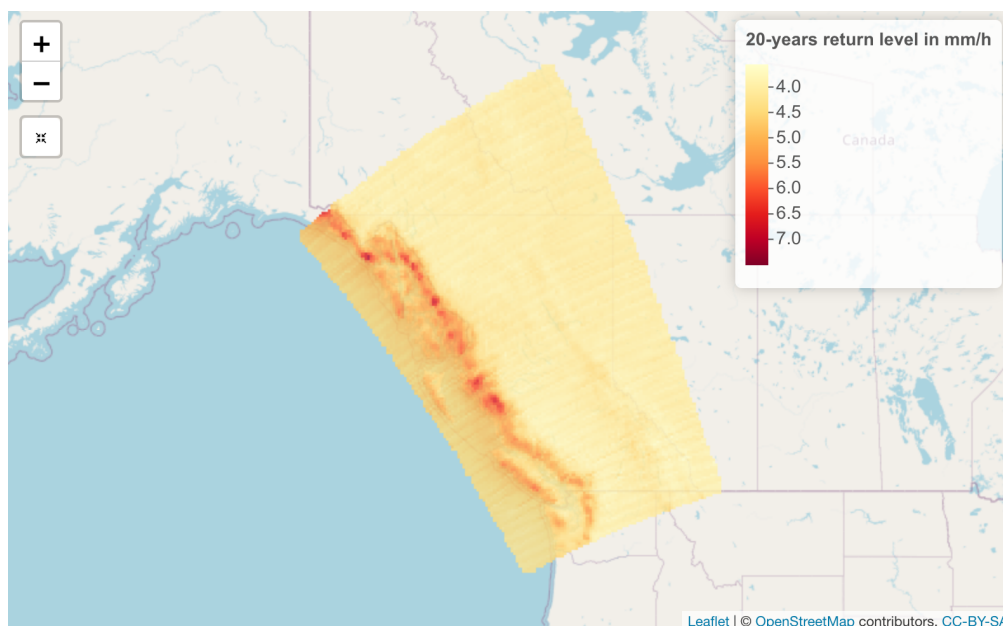


FIGURE 6.5 Quantile à 20 années de retour pour le modèle complet M_3

On remarque, par rapport aux cartes des modèles unidimensionnels, une similarité assez grande avec la carte du modèle unidimensionnel M_2 . Les estimations semblent globalement plus homogènes à l'intérieur du territoire, et la chaîne des Rocheuses ressort encore moins.

L'influence de chacune des trois variables peut être illustrée par leur intervalles de crédibilité empirique de niveau 95%, visibles en figure 6.6. Les intervalles sont, dans l'ordre, pour le quantile de la GEV issue de la réanalyse, la moyenne de précipitation, et l'altitude. Pour

rappel, β_u est le coefficient de régression pour le paramètre de localisation, alors que β_v est le coefficient pour le paramètre d'échelle.

Pour le paramètre de localisation μ , tous les intervalles semblent largement exclure 0. Même pour l'altitude, où les valeurs sont très faibles (de l'ordre de 10^{-5}), la valeur 0 demeure en dehors de l'intervalle de crédibilité. Ces graphes suggèrent que toutes ces variables sont bien significatives pour le paramètre de localisation μ de la loi GEV.

Quant au paramètre d'échelle σ , le quantile de la GEV estimée a un intervalle de confiance quasiment centré en 0, tandis que pour l'altitude, 0 est dans la partie droite de l'intervalle. À partir de ces observations, qui concordent avec les premières remarques que l'on pouvait relever des cartes des données, on pourrait conclure que l'altitude ainsi que le quantile de la GEV n'est pas ou peu significative dans la régression du paramètre d'échelle σ . Toutefois, cette dernière remarque est à nuancer, car tout dépend de notre taille d'intervalle de confiance. En effet, le choix s'est fait avec des intervalles à 95%, mais si on prenait des intervalles de confiance à 90%, alors l'altitude, selon ce critère, serait significative.

Ceci nous donne donc deux nouveaux modèles potentiels, en gardant toutes les variables pour le paramètre de localisation, puis soit en gardant la moyenne des précipitations et l'altitude pour le paramètre d'échelle, soit en ne gardant que la moyenne des précipitations. Ces deux nouveaux modèles seront appelés respectivement modèle partiel, et modèle adapté, et seront notés M_4 et M_5 , dans la suite logique du modèle complet.

6.2.2 Modèle partiel

D'abord, l'ajustement du modèle dit « partiel », noté dans la suite M_4 , est effectué : c'est le modèle où l'on conserve, pour le paramètre d'échelle, à la fois la moyenne de la précipitation, mais également l'altitude, malgré le fait qu'elle puisse être non significative.

La moyenne de la loi prédictive du quantile associé à la période de retour de 20 ans est illustré à la figure 6.7. La figure 6.8 illustre les traces des chaînes et leur intervalle de crédibilité empirique respectif.

Sur la carte, on n'observe pas de différence majeure par rapport au modèle complet ; de même, les intervalles de confiance restent cohérents avec ce qu'on a observé avec le modèle complet. Reste donc à voir si l'on a une différence plus marquée avec le troisième et dernier modèle.

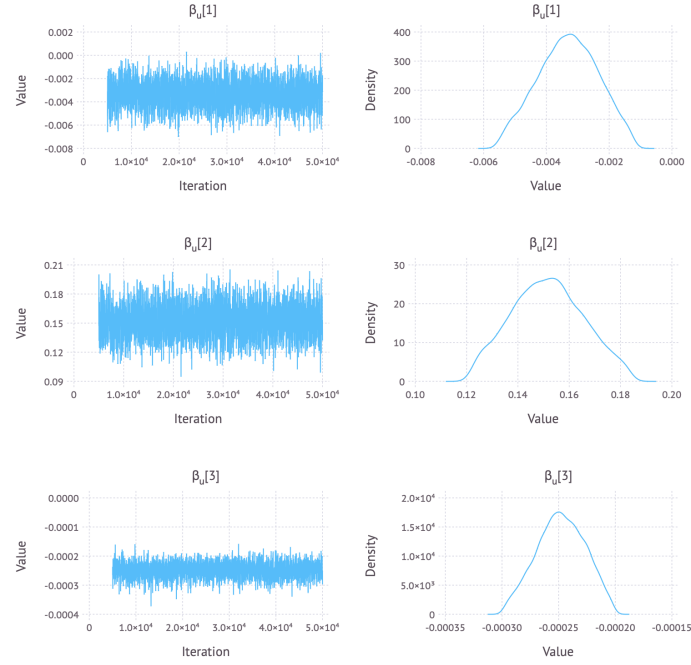
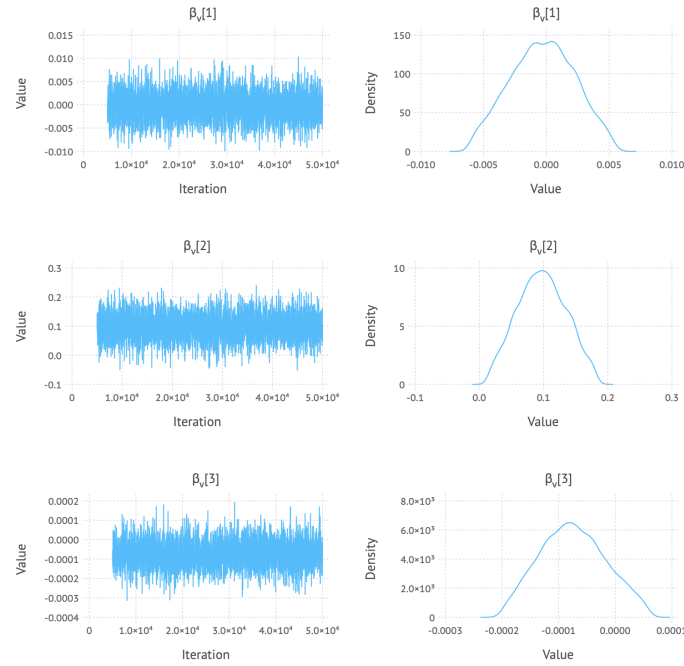
(a) β_u (b) β_v

FIGURE 6.6 Intervalles de crédibilité empiriques pour les coefficients de régression du modèle complet M_3

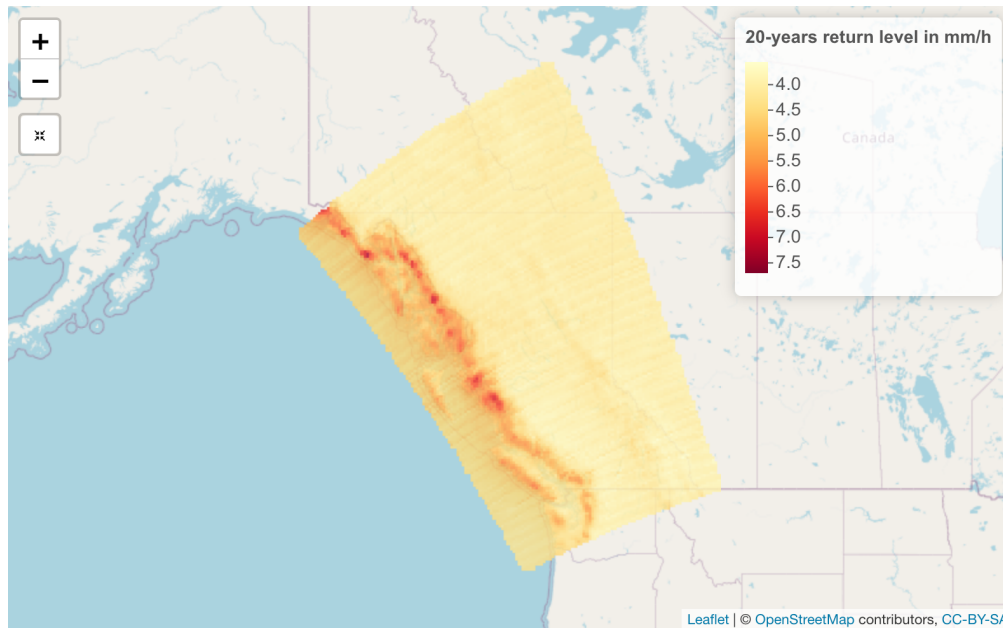


FIGURE 6.7 Quantile à 20 années de retour pour le modèle partiel M_4

6.2.3 Modèle adapté

Ce nouveau modèle va donc être mis en oeuvre, en adaptant les variables pris en compte pour chaque paramètre : pour le paramètre de localisation, tout est conservé, tandis que pour le paramètre d'échelle, on n'utilise que la moyenne de la précipitation. Le MCMC donne les résultats présentés en figure 6.10.

On ne remarque pas directement de différences notables avec le modèle complet a première vue via la carte. De même, les traces pour les stations ou encore les traces et intervalles de crédibilité pour les coefficients de régression, ou le paramètre de forme, ne laissent pas sous-entendre de grandes différences avec le modèle complet, si ce n'est que les intervalles de crédibilité à 95% excluent tous 0, assurant ainsi que toutes nos variables sont significatives, comme visible en figure 6.10.

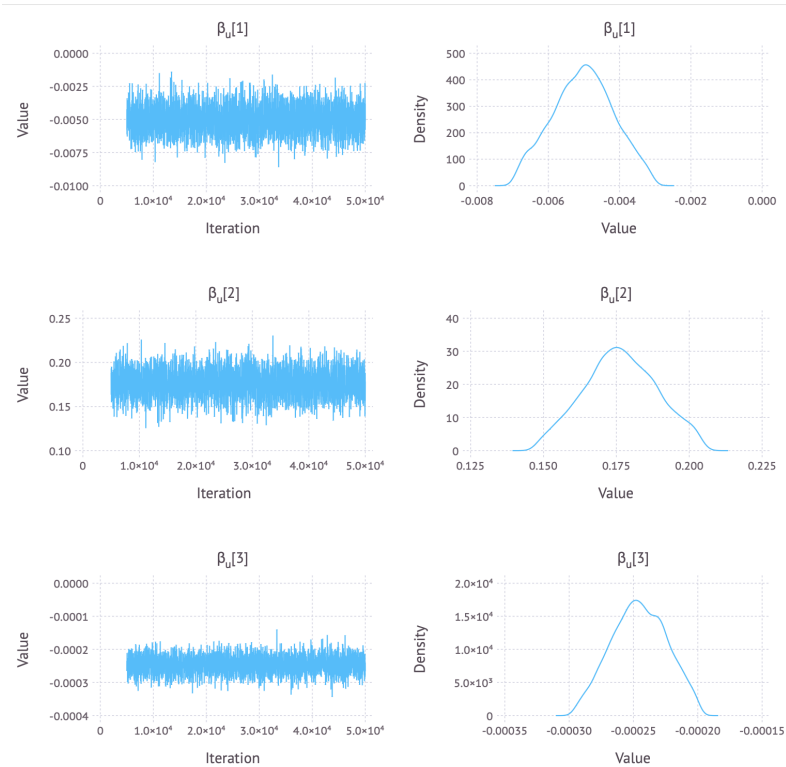
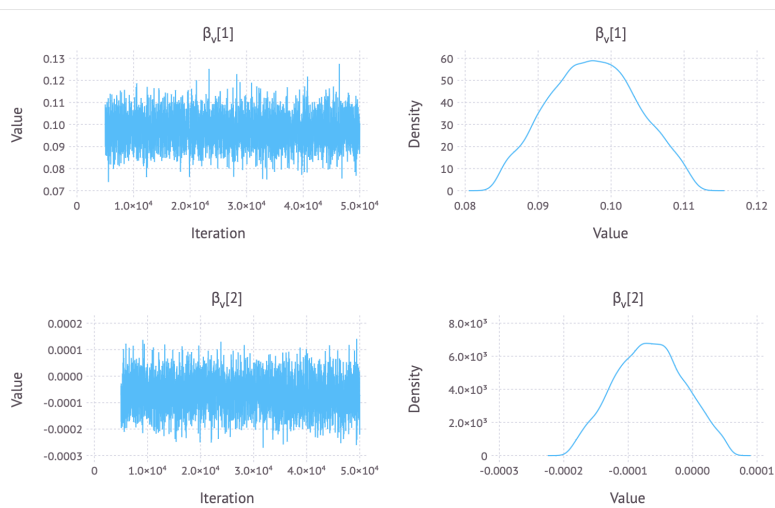
(a) β_u (b) β_v

FIGURE 6.8 Intervalles de confiance empiriques pour les coefficients de régression du modèle partiel M_4

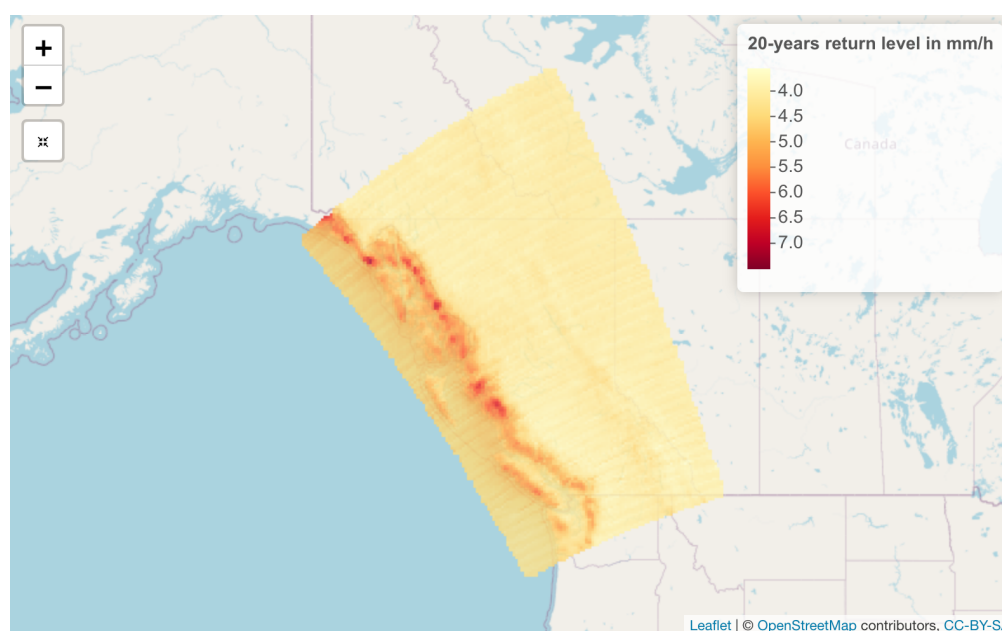


FIGURE 6.9 Quantile à 20 années de retour pour le modèle adapté M_5

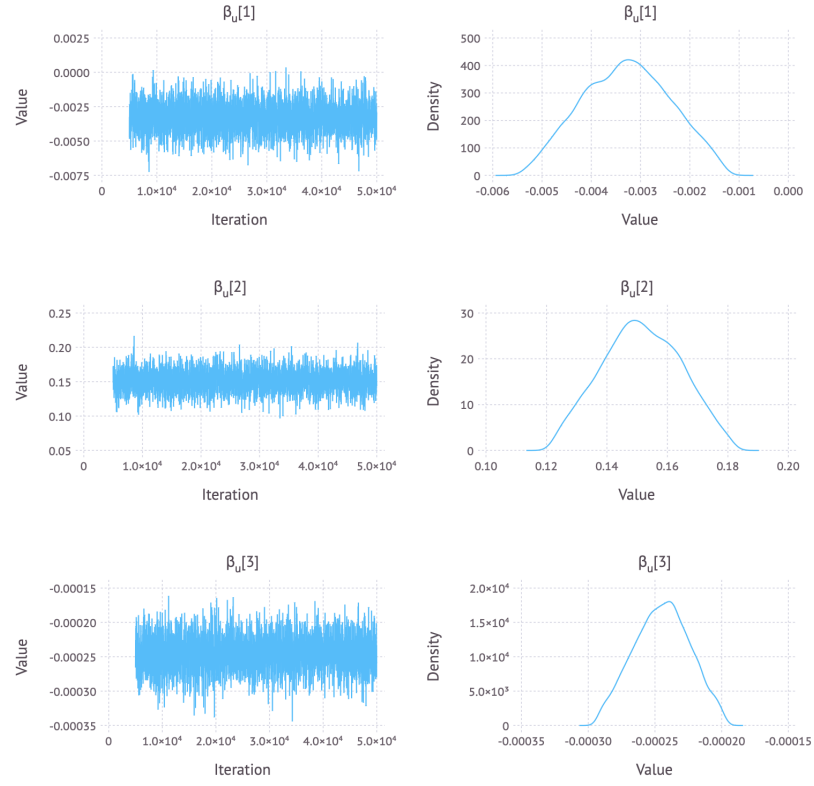
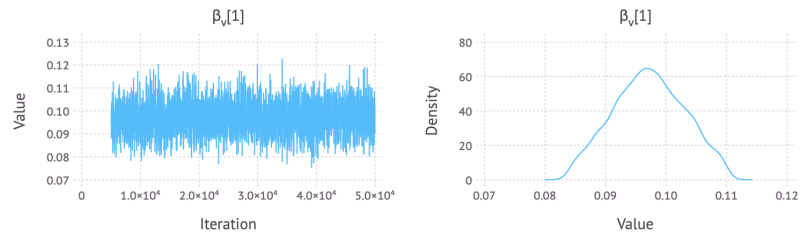
(a) β_u (b) β_v

FIGURE 6.10 Intervalles de confiance empiriques pour les coefficients de régression du modèle adapté M_5

6.3 Comparaison des modèles

Une autre méthode pour comparer ces différents modèles consiste à effectuer une validation croisée. Cette façon permet de comparer les modèles sur la base de leur pouvoir prédictif. L'erreur moyenne quadratique (*Mean Squared Error*, MSE), ainsi que le critère de Cramer-Von Mises (CVM) seront utilisés pour comparer les estimations à une cellule avec station en ayant auparavant retiré les données de cette cellule avec les données non utilisées.

Les trois modèles suivants sont comparés :

M_3 : Le modèle complet, c'est à dire qui inclut toutes les variables pour chacun des deux paramètres μ et σ .

M_4 : Le modèle partiel, où l'on enlève le quantile de la GEV mais on conserve l'altitude et la moyenne de précipitation pour le paramètre d'échelle.

M_5 : Le modèle adapté, où l'on ne garde que la moyenne des précipitations pour le paramètre d'échelle.

6.3.1 Validation croisée

La validation croisée est une méthode en apprentissage pour confirmer que nos résultats via des modèles ne sont pas issus d'un sur- ou sous-apprentissage. Elle consiste à comparer un morceau de notre population avec notre modèle, entraîné sur le reste de la population, et de répéter l'opération de sorte à balayer toute la population afin de mesurer le pouvoir prédictif du modèle.

La question de la taille du morceau est le point quant à la forme de validation croisée que l'on va effectuer : lorsque l'on prend point par point, on effectue une validation dite "*Leave One Out Cross Validation*" (LOOCV). Lorsqu'on en prend plusieurs points d'un coup, on effectue une validation dite "*K-fold*", où K est le nombre de blocs.

Pour effectuer la validation croisée, la méthode K -folds a été retenue plutôt que la simple LOOCV, d'autant plus qu'avec ce nombre relativement important de points à valider, à savoir les 120 stations, la division en folds sera plus adaptée. On note b_K la taille d'un bloc associée, de sorte à avoir $m = Kb_K$. Un schéma explicatif de la validation croisée par méthode K -folds est présentée en figure 6.11.

Ainsi, comme 120 est un nombre disposant de beaucoup de diviseurs, la validation a été effectuée avec des nombres de folds différents, permettant un meilleur balayage. J'ai donc répété la validation en utilisant succinctement des tailles de 5, 6, 8 et 10 stations.

Voici la procédure détaillée, adaptée à la forme du projet de recherche :

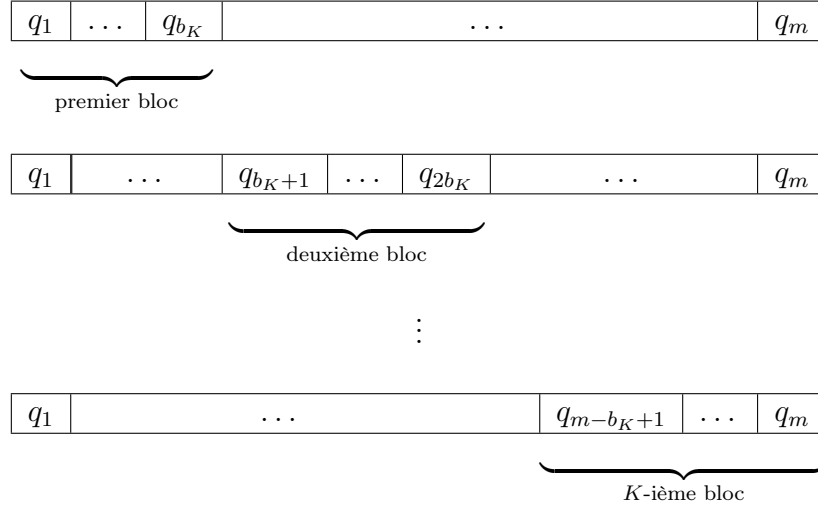


FIGURE 6.11 Schéma résumé de la validation croisée

1. On choisit notre modèle, c'est à dire les variables que l'on va inclure ou non pour tel ou tel paramètre.
2. On estime les paramètres en générant un échantillon de la loi *a posteriori* du MCMC afin d'effectuer l'interpolation spatiale, de manière similaire à la partie précédente (même nombre d'itérations, de réduction de chaîne, etc...)
3. Une fois les chaînes obtenues, on extrait les valeurs finales.
4. En fonction du nombre de folds K et de la taille de ces folds b_K , on partitionne aléatoirement les stations de relevé dans chacun de ces folds.
5. Pour chaque fold, on effectue une nouvelle MCMC avec la procédure suivante :
 - a) On crée une nouvelle structure de données, où les valeurs initiales de la grille sont les valeurs finales de la première chaîne, et où l'on enlève le fold de l'ensemble des stations interpolées.
 - b) On fait à nouveau tourner les chaînes, cette fois pour une durée plus réduite : seulement 1000 itérations, aucun affinage de la chaîne.
 - c) Une fois la chaîne terminée et l'interpolation effectuée sur l'ensemble du territoire, on récupère les paramètres de la GEV aux stations du fold (qui étaient, pour rappel, retirées initialement de la chaîne).
 - d) À partir des résultats obtenus aux stations du fold, on calcule les critères de comparaison pour ce fold.
6. Une fois que l'on a les critères pour tous les folds, on moyenne le tout pour obtenir nos critères pour un modèle et un nombre de folds donnée.

Cette procédure sera donc répétée pour tous les modèles présentés au-dessus, pour chaque

nombre de fold retenue.

Les critères de comparaison choisis sont au nombre de deux :

Tout d’abord, on va prendre l’erreur quadratique moyenne (Mean Squared Error, MSE), erreur où l’on va calculer les écarts au carré entre le quantile à 20 années de retour que l’on obtient soit via le modèle, soit empiriquement via les données d’ECCC. Le quantile modélisé est donc calculé après interpolation faite sur un fold : ensuite, on effectue la moyenne de la MSE pour chaque fold pour avoir la MSE globale, pour une taille de fold K et un modèle donné.

L’autre critère employé est le critère de Cramer-Von Mises ; il s’agit d’une mesure qui évalue la qualité d’une distribution par rapport à une distribution empirique, ou juste permet de comparer deux distributions empiriques entre elles. Le critère a originellement été développé par Cramer [43] et Von Mises [44] à la fin des années 20, et a été étendu à deux échantillons par Anderson (1962) [45].

Sa définition formelle est donnée par :

$$\omega^2 = \int_{-\infty}^{+\infty} [F_n(x) - F^*]^2 dF^*(x),$$

où F^* est la fonction de répartition que l’on compare à la fonction de répartition empirique F_n . En pratique, on utilise la statistique de test :

$$T = n\omega^2 = \frac{1}{12n} + \sum_{i=1}^n \left\{ \frac{2i-1}{2n} - F(x_i) \right\}^2,$$

où $(x_1, \dots, x_n)$ est notre échantillon de valeurs observées. Le lecteur intéressé par plus de détails et de variation autour de ce critère pourra consulter

6.3.2 Résultats finaux et discussion

En termes techniques, chaque option de la validation, c’est à dire pour un modèle et une taille de fold donnée, a été exécuté sur un ordinateur portable personnel, et chacune d’entre elles exigeait entre 4 et 6 heures de calculs. Ce temps significativement plus long que les premières MCMC s’explique par le le nombre nécessaire de chaînes à faire et refaire tourner, chacune pour un fold. Le temps d’exécution s’allongeait d’autant plus que la taille des folds diminuait, car plus les folds sont petits, plus leur nombre est important. C’est une autre raison quant à la faible taille des chaînes une fois la première grosse chaîne effectuée.

Les résultats pour la MSE sont visibles en table 6.1 tandis que ceux pour la CVM sont en table 6.2 ; la case mise en gras est le meilleur critère que l'on obtient pour une taille de fold fixée.

TABLEAU 6.1 Tableau des résultats de la MSE après validation croisée

Modèle \ b_K	5	6	8	10
M_3	1278	1353	1334	1333
M_4	1271	1340	1324	1321
M_5	1286	1356	1337	1330

TABLEAU 6.2 Tableau des résultats de la CVM après validation croisée

Modèle \ b_K	5	6	8	10
M_3	0.1694	0.1613	0.1744	0.1621
M_4	0.1693	0.1606	0.1762	0.1628
M_5	0.1693	0.1602	0.1733	0.1628

On remarque, en observant ces tables, que la MSE nous donne des résultats unanimes, à savoir que le modèle réduit M_4 est le meilleur : globalement, l'écart quadratique moyen est le plus faible. Par contre, pour le CVM, il n'y pas de modèle qui ressorte clairement, au mieux le modèle adapté M_3 dispose du plus de résultats en sa faveur.

Dans l'ensemble, les différences entre les critères d'un modèle à l'autre sont vraiment faibles : par endroits, les différences se sont même révélées inférieures au nombre de chiffres significatifs initialement choisi, comme pour $K = 5$ entre les modèles M_4 et M_5 pour le CVM. Tous ces résultats laissent entendre une compétition très proche entre les modèles M_4 et M_5 , avec un léger avantage pour le modèle M_4 . En revanche, le fait que l'altitude soit révélée comme non significative à 95% dans chaque modèle multidimensionnel est un argument de poids contre ce modèle M_4 , le seuil de 95% étant la norme usuelle en statistiques. Ainsi, cette concordance de résultats tendent à orienter la modélisation spatiale des GEV de la Colombie-Britannique avec le modèle M_5 .

Finalement, ces modèles permettent de conclure que l'élévation au niveau à la mer est une variable significative pour les maxima de précipitation annuels en Colombie-Britannique en plus de la réanalyse RDRS 2.1. Toute cette méthodologie ayant été effectuée avec la durée D de 24 heures et la période de retour $T = 20$ ans, il est possible de construire les IDF de la province en répétant la procédure en utilisant les données d'ECCC pour une durée différente,

et calculer les quantiles pour une période de retour différente. Ces calculs donneront donc les points (D, I_D) du graphe, que l'on pourra relier entre eux pour une même période de T et obtenir finalement les nouvelles courbes IDF de la Colombie-Britannique.

CHAPITRE 7 CONCLUSION ET RECOMMANDATIONS

7.1 Synthèse du projet

Au travers de ce projet, un modèle bayésien hiérarchique spatial a été mis en place afin d’estimer et d’interpoler les courbes IDF de la Colombie-Britannique, permettant ainsi d’analyser les données des précipitations extrêmes de la province, et en les combinant avec un modèle climatique simulé, de rendre compte de la significativité de l’élévation par rapport au niveau de la mer, dans une zone où elle a effectivement plus d’importance que dans des régions plus planes que le Québec. On obtient à la fin une carte des quantiles des lois de probabilités sur l’ensemble de la province pour la période de retour de 20 ans et la durée de relevé de 24 heures. La réitération de cette permettant ainsi de fournir à la suite les courbes IDF mises à jour sous ce nouveau modèle, comme détaillé en annexe.

Pour cela, plusieurs éléments théoriques ont été introduits. Tout d’abord, la théorie des valeurs extrêmes, essentielle au projet, mais également les modèles bayésiens hiérarchiques, les champs markoviens gaussiens. En s’inspirant de la littérature récente existante et du travail antécédent du laboratoire Extremes, un modèle spatial a été mis en place, en y ajoutant plusieurs variables environnementales. Une fois la structure du modèle principal présentée, les algorithmes MCMC pour le calcul numérique sont énumérés. Pour ce qui est des simulations, la significativité est déterminée grâce aux intervalles de confiance empirique à 90 ou 95%, permettant ainsi par la suite de comparer entre eux plusieurs variantes intégrant tout ou partie de ces variables. Le départage de ces variantes est ensuite effectué par validation croisée, grâce à des MCMC successives, puis permettant de calculer des critères de sélection entre les modèles.

Cette dépendance de l’altitude, et plus généralement cette régression de plusieurs variables environnementales, et ce appliqué à une nouvelle province du Canada, est une avancée pour le travail du laboratoire Extremes. Cette méthodologie pourra ainsi être appliquée à d’autres zones du Canada, notamment les provinces centrales et les territoires du nord, avec les adaptations nécessaires à ces zones.

7.2 Limitations et ouvertures

Cette étude comporte bien entendu de nombreuses limites, en partie due au déplacement vers la Colombie-Britannique que ce projet a subi. Tout d'abord, la plus évidente est la non-stationnarité des extrêmes, qui a été mentionné dans de nombreuses références lors de la revue de littérature, ainsi qu'abordé dans d'autres projets de recherche au sein du laboratoire Extremes [27, 46], d'autant plus que les ajustements à effectuer ne sont pas énormes, et les variables climatiques permettant d'étudier la non-stationnarité sont légion (à commencer par la concentration en GES dans l'atmosphère par année). Par ailleurs CaSPAr fournit beaucoup de données environnementales simulées en plus des précipitations, ce qui représenterait d'autres variables à ajouter, d'autant qu'il existe des modèles complexes faisant intervenir de nombreux paramètres [47].

Une étude des performances, en comparant à d'autres périodes de retour, aurait également amélioré notre aperçu sur la qualité de l'analyse menée. Pour la conclusion de ce projet et sa continuation, des analyses plus exhaustives seront menées.

Une autre possibilité de développement est d'introduire la dépendance inter-durées des extrêmes. En effet, pour l'instant on estime les paramètres des GEV indépendamment de la durée d'accumulation, mais ce n'est pas si évident que ça à négliger en réalité. Les travaux récents de Paul MATHIVON [48] au sein du laboratoire pourraient être employés dans cette optique.

Une extension possible repose sur l'algorithme MCMC à utiliser. Dans ce projet, un Métropolis-Hastings combiné à un échantillonnage de Gibbs a été employé, ce qui représente une des versions les plus classiques du genre, mais d'autres solutions sont possibles. On peut notamment penser à l'algorithme Metropolis Adjusted Langevin Algorithm (MALA), qui permet une exploration de l'espace des possibilités plus efficace [49], ou encore le Monte Carlo hybride (Hamiltonian Monte Carlo) [50, 51], permettant un meilleur taux d'acceptation.

Par ailleurs, au vu de la disparité visible entre les zones côtières, il pourrait être intéressant de considérer une approche en clusters, comme déjà effectué précédemment [22], où les clusters seraient plus grands dans les zones montagneuses, et plus resserrés dans les zones côtières. Il faudrait néanmoins un découpage en cantons pour la province.

RÉFÉRENCES

- [1] H. Jouan, “D’Ouest en Est, le Canada est ravagé par des incendies hors normes,” *Le Monde*, Juin 2023. [En ligne]. Disponible : https://www.lemonde.fr/planete/article/2023/06/02/au-canada-une-saison-d-incendies-hors-norme_6175940_3244.html
- [2] R. Bourdillon, “L’Espagne affronte la désertification,” *Reporterre*, Février 2020. [En ligne]. Disponible : <https://reporterre.net/On-va-avoir-un-paysage-lunaire-L-Espagne-affronte-la-desertification>
- [3] C. Thibaud, “L’Espagne, entre la désertification et la destruction du littoral,” *Les Echos*, Novembre 2021. [En ligne]. Disponible : <https://reporterre.net/On-va-avoir-un-paysage-lunaire-L-Espagne-affronte-la-desertification>
- [4] S. Landrin, “Inde : une canicule précoce écrase le pays,” *Le Monde*, Avril 2022. [En ligne]. Disponible : https://www.lemonde.fr/planete/article/2022/04/27/l-inde-ecrasee-par-une-vague-de-chaaleur-precoce-et-extreme_6123864_3244.html
- [5] T. Wagner et R. Seguin, “Canicules marines : des incendies sous-marins aux conséquences alarmantes,” *Bon Pote*, Juin 2023. [En ligne]. Disponible : <https://bonpote.com/canicules-marines-des-incendies-sous-marins-aux-consequences-alarmantes>
- [6] Courbes IDF d’ECCC. [En ligne]. Disponible : https://climate.weather.gc.ca/prods_servs/engineering_e.html
- [7] R. A. Fisher et L. H. C. Tippett, “Limiting forms of the frequency distribution of the largest or smallest member of a sample,” *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 24, n^o. 2, p. 180–190, 1928.
- [8] S. Coles, *An Introduction to Statistical Modeling of Extreme Values*, 3^e éd. Springer, 2004.
- [9] E. S. Martins et J. R. Stedinger, “Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data,” *Water Resources Research*, vol. 36, n^o. 3, p. 737–744, 2000.
- [10] J. Jalbert. Librairie Julia **Extremes.jl**. [En ligne]. Disponible : <https://github.com/joal5/Extremes.jl/>
- [11] C. Robert, *Le choix bayésien : principe et pratique*. Springer, 2006.
- [12] H. Rue et L. Held, *Gaussian Markov Random Fields : Theory and Applications*. Taylor & Francis Group, 2005.
- [13] J. Jalbert et M.-L. Farmer. Librairie Julia **GMRF.jl**. [En ligne]. Disponible : <https://github.com/joal5/GMRF.jl>

- [14] J. R. M. Hosking et J. R. Wallis, *Regional frequency analysis*. Cambridge, 1997.
- [15] P. Sharkey et H. Winter, “A Bayesian spatial hierarchical model for extreme precipitation in Great Britain,” *Environmetrics*, vol. 30, p. e2529, 2019.
- [16] R. W. Katz *et al.*, “Statistics of extremes in hydrology,” *Advances in water resources*, vol. 25, n°. 8-12, p. 1287–1304, 2002.
- [17] A. Mailhot *et al.*, “Assessment of future change in intensity–duration–frequency (IDF) curves for Southern Quebec using the Canadian Regional Climate Model (CRCM),” *Journal of hydrology*, vol. 347, n°. 1-2, p. 197–210, 2007.
- [18] M. Piras *et al.*, “Impacts of climate change on precipitation and discharge extremes through the use of statistical downscaling approaches in a Mediterranean basin,” *Science of the Total Environment*, vol. 543, p. 952–964, 2016.
- [19] S. G. Coles et E. A. Powell, “Bayesian methods in extreme value modelling : a review and new developments,” *International Statistical Review/Revue Internationale de Statistique*, p. 119–136, 1996.
- [20] J. Jalbert *et al.*, “A spatiotemporal model for extreme precipitation simulated by a climate model, with an application to assessing changes in return levels over North America,” *Journal of the Royal Statistical Society Series C : Applied Statistics*, vol. 66, n°. 5, p. 941–962, 2017.
- [21] J. Jalbert *et al.*, “Interpolation of Precipitation Extremes on a Large Domain Toward IDF Curve Construction at Unmonitored Locations,” *IEEE Journal of Agricultural, Biological and Environmental Statistics*, vol. 27, p. 461–486, 2022.
- [22] C. Rohrbeck et J. A. Tawn, “Bayesian spatial clustering of extremal behavior for hydrological variables,” *Journal of Computational and Graphical Statistics*, vol. 30, n°. 1, p. 91–105, 2021.
- [23] B. J. Reich et B. A. Shaby, “A spatial Markov model for climate extremes,” *Journal of Computational and Graphical Statistics*, vol. 28, n°. 1, p. 117–126, 2019.
- [24] Jeux de données de CLIMDEX. [En ligne]. Disponible : <https://www.climdex.org/learn/datasets/>
- [25] S. Wi *et al.*, “Non-stationary frequency analysis of extreme precipitation in South Korea using peaks-over-threshold and annual maxima,” *Stochastic environmental research and risk assessment*, vol. 30, p. 583–606, 2016.
- [26] F. Santiñaque *et al.*, “Spatial clustering of extreme annual precipitation in Uruguay,” *Revista Brasileira de Meteorologia*, vol. 37, p. 419–426, 2023.
- [27] E. Nolet-Gravel, “Changements projetés des précipitations extrêmes au Québec,” Maîtrise de recherche, Polytechnique Montréal, 2019.

- [28] F. Palacios-Rodriguez *et al.*, “Smooth Copula-based Generalized Extreme Value model and spatial interpolation for sparse extreme rainfall in Central Eastern Canada,” *Environmetrics*, 2023.
- [29] A. C. Davison *et al.*, “Statistical Modeling of Spatial Extremes,” *Statistical Science*, vol. 27, n^o. 2, p. 161 – 186, 2012.
- [30] K. D. et al., “A mathematical framework for studying rainfall intensity-duration-frequency relationships,” *Journal of Hydrology*, vol. 206, n^o. 1-2, p. 118–135, 1998.
- [31] E. A. Lehmann *et al.*, “Spatial modelling framework for the characterisation of rainfall extremes at different durations and under climate change,” *Environmetrics*, vol. 27, n^o. 4, p. 239–251, 2016.
- [32] N. Gasset *et al.*, “A 10 km North American precipitation and land-surface reanalysis based on the GEM atmospheric model,” *Hydrology and Earth System Sciences*, 2021.
- [33] J.-F. Mahfouf *et al.*, “A Canadian precipitation analysis (CaPA) project : Description and preliminary results,” *Atmosphere-ocean*, vol. 45, n^o. 1, p. 1–17, 2007.
- [34] F. Lespinas *et al.*, “Performance evaluation of the Canadian precipitation analysis (CaPA),” *Journal of Hydrometeorology*, vol. 16, n^o. 5, p. 2045–2064, 2015.
- [35] Modèles disponibles et archivés produits par CaSPAr. [En ligne]. Disponible : https://github.com/julemai/CaSPAr/wiki/Available-products_fr
- [36] M. Abaza *et al.*, “Assessing 32-Day hydrological ensemble forecasts in the Lake Champlain–Richelieu River watershed,” *Journal of Hydrologic Engineering*, vol. 25, n^o. 10, p. 04020045, 2020.
- [37] J. Mai *et al.*, “The great lakes runoff intercomparison project phase 4 : the great lakes (GRIP-GL),” *Hydrology and Earth System Sciences*, vol. 26, n^o. 13, p. 3537–3572, 2022.
- [38] Site internet de CaSPAr. [En ligne]. Disponible : <https://caspar-data.ca>
- [39] Site internet de l’API Open Meteo. [En ligne]. Disponible : <https://open-meteo.com/en/docs/elevation-api>
- [40] “Loi gamma,” dans *Wikipédia*. [En ligne]. Disponible : https://fr.wikipedia.org/wiki/Loi_Gamma
- [41] “Loi bêta,” dans *Wikipédia*. [En ligne]. Disponible : https://fr.wikipedia.org/wiki/Loi_bêta
- [42] A. V. Dyrddal *et al.*, “Bayesian hierarchical modeling of extreme hourly precipitation in Norway,” *Environmetrics*, vol. 26, n^o. 2, p. 89–106, 2015.
- [43] H. Cramér, “On the composition of elementary errors,” *Scandinavian Actuarial Journal*, vol. 1928, n^o. 1, p. 13–74, 1928.
- [44] R. von Mises, *Statistik und Wahrheit*. Julius Springer, 1928.

- [45] T. W. Anderson, “On the Distribution of the Two-Sample Cramér-von Mises Criterion,” *The Annals of Mathematical Statistics*, vol. 33, n°. 3, p. 1148–1159, 1962.
- [46] D.-A. Alexandre, “Analyse fréquentielle intégratrice des débits projetés au Québec Méridional,” Maîtrise de recherche, Polytechnique Montréal, 2022.
- [47] M. Kallache *et al.*, “Nonstationary probabilistic downscaling of extreme precipitation,” *Journal of Geophysical Research : Atmospheres*, vol. 116, n°. D5, 2011.
- [48] P. Mathivon, “Modélisation conjointe des maxima annuels d’intensité de précipitation,” Maîtrise de recherche, Polytechnique Montréal, 2023.
- [49] T. Xifara *et al.*, “Langevin diffusions and the Metropolis-adjusted Langevin algorithm,” *Statistics & Probability Letters*, vol. 91, p. 14–19, 2014.
- [50] M. Betancourt et M. Girolami, “Hamiltonian Monte Carlo for hierarchical models,” *Current trends in Bayesian methodology with applications*, vol. 79, n°. 30, p. 2–4, 2015.
- [51] M. Betancourt, “A conceptual introduction to Hamiltonian Monte Carlo,” *arXiv preprint arXiv :1701.02434*, 2017.

ANNEXE A GRAPHES POUR LES STATIONS D'INTÉRÊT

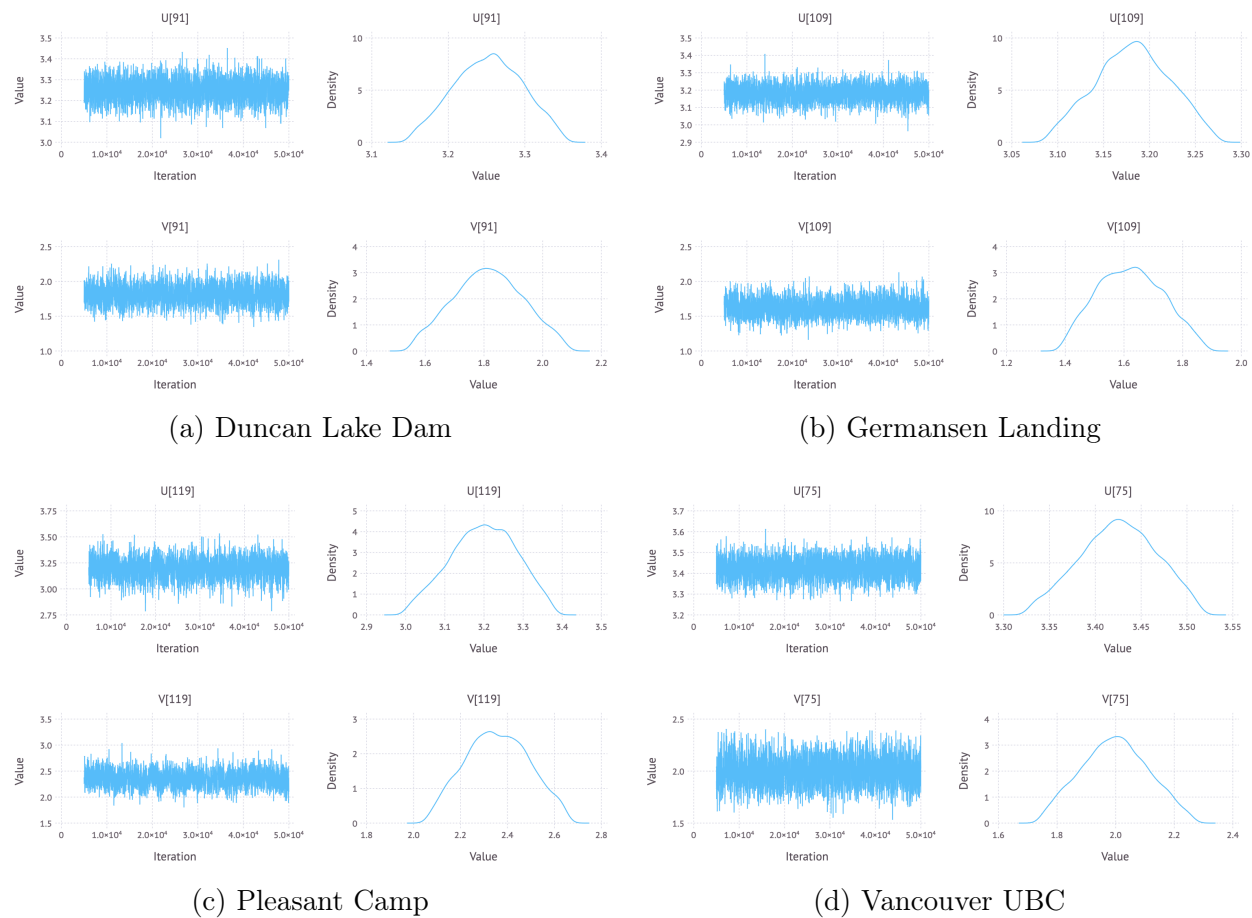


FIGURE A.1 Traces et intervalles de crédibilité des stations d'intérêt pour le modèle M_3 (Modèle complet)

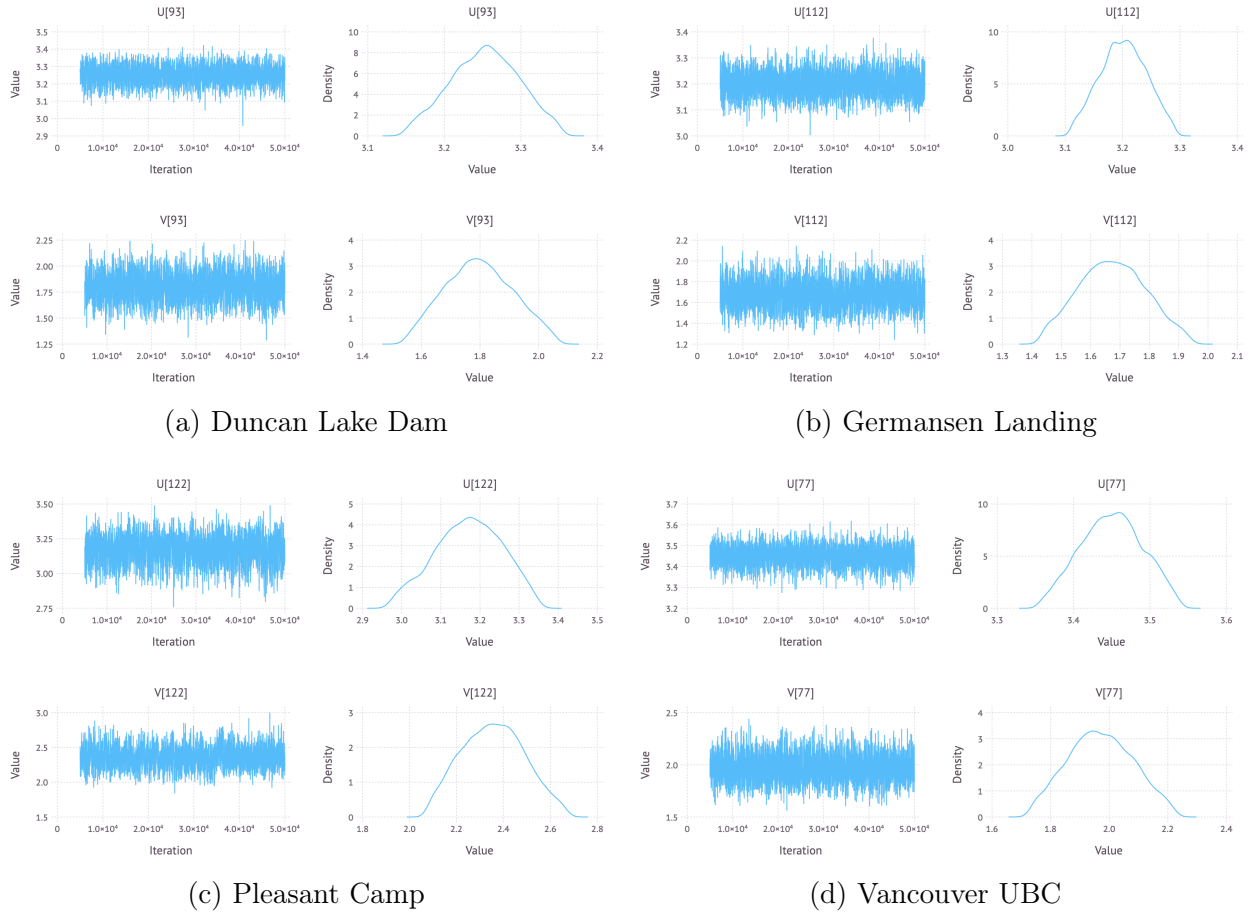


FIGURE A.2 Traces et intervalles de crédibilité des stations d'intérêt pour le modèle M_4 (Modèle partiel)

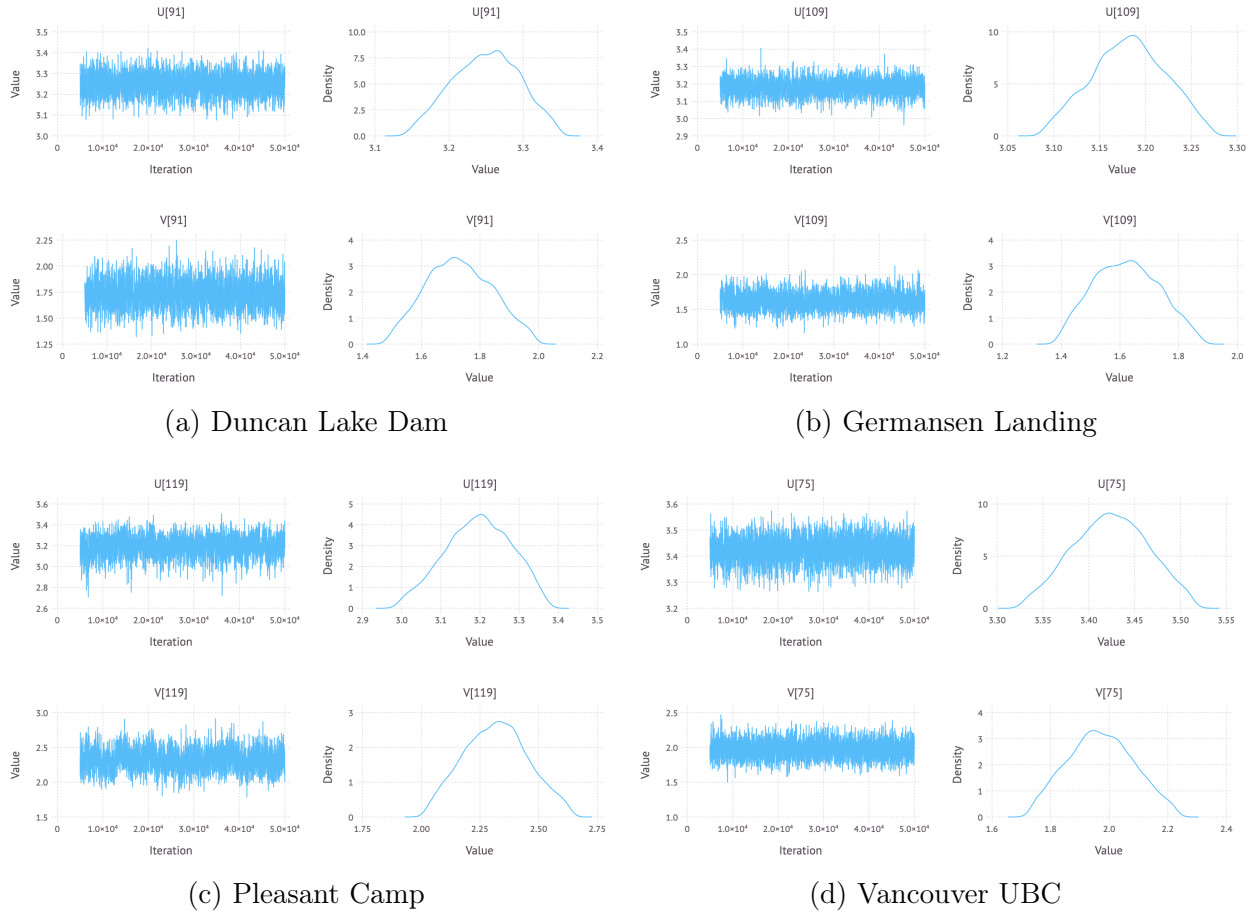


FIGURE A.3 Traces et intervalles de crédibilité des stations d'intérêt pour le modèle M_5 (Modèle adapté)

ANNEXE B ORGANIGRAMME DE LA CONSTRUCTION DES COURBES IDF

L'organigramme en figure B.1 présente la démarche permettant, à partir de la méthodologie développée dans ce mémoire, de produire les courbes IDF réactualisées au complet. Le procédé permet d'obtenir le quantile q_T pour une période de retour T et une durée de relevé D données, en un point de grille de la zone d'étude. Pour construire les courbes IDF au complet, il suffit de réitérer l'ensemble du processus pour les valeurs T (usuellement 1, 2, 5, 10, 20, 50 et 100 ans) et D (usuellement 5, 10, 30 minutes, 1, 2, 6, 12 et 24 heures).

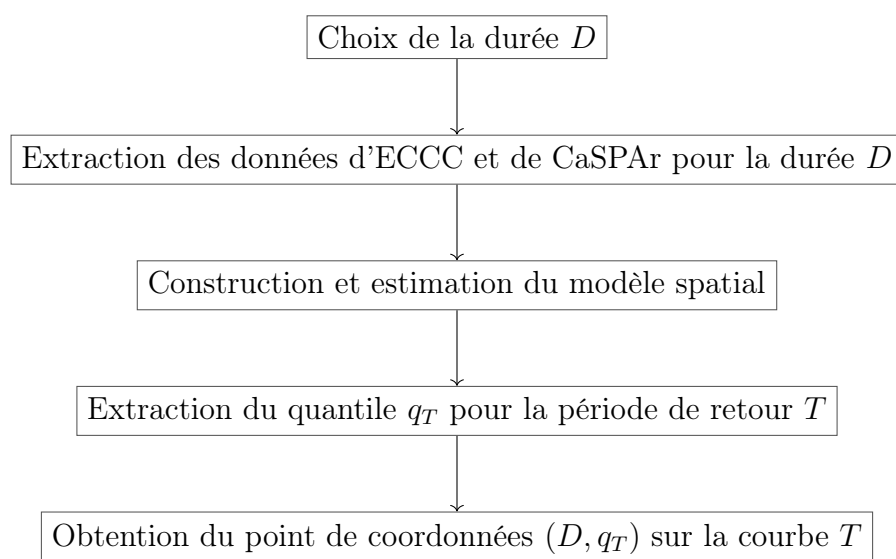


FIGURE B.1 Déroulé du procédé de construction des IDF schématisé.