Titre: Title:	Analyse fréquentielle intégratrice des débits projetés au Québec méridional
Auteur: Author:	Duy Anh Alexandre
Date:	2022
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Alexandre, D. A. (2022). Analyse fréquentielle intégratrice des débits projetés au Québec méridional [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/10358/

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Analyse fréquentielle intégratrice des débits projetés au Québec méridional

DUY ANH ALEXANDRE
Département de mathématiques et de génie industriel

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

Mai 2022

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Analyse fréquentielle intégratrice des débits projetés au Québec méridional

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

Julie CARREAU, présidente
Jonathan JALBERT, membre et directeur de recherche
David HUARD, membre

DÉDICACE

À mon papa,
sans qui rien de tout ceci
n'aurait été possible.

REMERCIEMENTS

Mes remerciements vont à tous ceux et celles qui m'ont aidé de près ou de loin dans la réalisation de ce projet.

D'abord, un grand merci à Jonathan Jalbert, mon directeur de recherche, pour son encadrement, sa disponibilité et ses conseils. Ton soutien pédagogique et ta constante bonne humeur m'ont grandement aidé à tirer le meilleur parti de cette expérience qu'est la maîtrise de recherche. Merci aussi pour m'avoir donné le goût de la recherche académique.

Le second remerciement va à Gabriel Gobeil, mon collègue de laboratoire et de bureau durant la majeure partie de ma maîtrise. Merci pour ta bonne humeur, ton aide à tous les égards, les discussions du dîner... et simplement ta présence, qui a rendu toutes ces heures de travail bien plus agréables.

De manière plus formelle, je tiens à remercier les scientifiques et ingénieurs de la DEH et d'Ouranos, avec qui j'ai eu la chance de collaborer sur ce projet et d'avoir des échanges enrichissants. J'ai toujours exprimé un vif intérêt à participer aux réunions mensuelles qui m'ont permis d'en apprendre un peu plus sur les domaines de l'hydrologie et de la climatologie. Je remercie particulièrement Simon Lachance-Cloutier et Gabriel Rondeau-Genesse pour votre réactivité et vos réponses rapides à mes diverses questions. Je remercie également Julie Carreau et David Huard pour avoir accepté de faire partie du jury de mon mémoire de maîtrise.

Plus globalement, j'aimerais remercier toutes les personnes qui ont contribué à rendre mon séjour au Canada mémorable : mes colocataires pour tous les bons moments passés ensemble, la gang, la team Shakti, Paul, Léo, les nouveaux comme les anciens... Merci !

RÉSUMÉ

Les inondations exceptionnelles de 2017 et de 2019 au Québec ont mis en lumière plusieurs défis en matière d'adaptation aux aléas climatiques. Pour mieux prévoir les risques d'inondation future, le ministère de l'Environnement et de la Lutte contre les changements climatiques a mis en place en avril 2018 le projet INFO-Crue, qui vise à produire une cartographie prévisionnelle des zones inondables dans une grande partie du Québec méridional. Ce projet y contribue en étudiant l'évolution des débits extrêmes dans les tronçons de rivières jaugés et non jaugés du Québec jusqu'à l'horizon 2100. Les estimations de débits futurs seront par la suite utilisées pour calculer les niveaux de crue attendus et cartographier les zones à risque.

L'étude des débits s'appuie sur deux ensembles de données fournis par la Direction de l'Expertise Hydrique (DEH), les pseudo-observations aux tronçons non jaugés et les simulations hydroclimatiques. Les pseudo-observations proviennent d'une méthode d'interpolation utilisant des observations des stations environnants et des simulations du modèle hydrologique Hydrotel. Elle permet d'estimer les débits dans les tronçons de rivière non jaugés (sans appareils de mesure) jusqu'à 2020. Les simulations proviennent d'un ensemble de modèles climatiques couplés au modèle hydrologique Hydrotel, simulant les débits futurs sous différents scénarios d'émission de gaz à effet de serre jusqu'à 2100. Ces simulations prennent en compte les changements climatiques.

Deux modèles hiérarchiques bayésiens ont été développés, un pour les pseudo-observations et un pour les simulations hydroclimatiques. Les débits maximaux annuels sont extraits et modélisés dans le cadre de la théorie des valeurs extrêmes. Les modèles hiérarchiques bayésiens sont particulièrement adaptés pour structurer les données similaires provenant de plusieurs sources. Ils permettent en outre une estimation des débits extrêmes qui prend en compte les sources d'incertitudes dans un cadre statistique cohérent. Les sorties du modèle statistique pour les simulations sont biaisées par rapport aux sorties du modèle statistique pour les pseudo-observations. Une méthode de post-traitement statistique est alors adaptée pour corriger le biais entre les deux modèles et estimer les débits extrêmes futurs.

La méthodologie proposée a été appliquée aux données de 234 tronçons de la rivière Chaudière au Québec. Les prévisions indiquent généralement une évolution à la baisse en amont et à la hausse en aval de la rivière pour le niveau de retour 100 ans. Les résultats de ce projet sont directement utilisables par la DEH pour cartographier les risques d'inondations actuels et futurs.

ABSTRACT

The exceptional floods of 2017 and 2019 in Quebec have shed light on several challenges regarding adaptation to climate hazards. To better foresee future flood risks, the Ministère de l'Environnement et de la Lutte contre les changements climatiques started the INFO-Crue project in April 2018, which aims to produce a predictive map of flood-prone areas in the majority of southern Quebec. This project contributes to INFO-Crue by studying the evolution of extreme flows in gauged and ungauged river sections in Quebec, until 2100. Future flow estimations will then be used by another research team to calculate expected flood levels and map areas at risk.

The statistical study is based on two datasets provided by the Direction de l'Expertise Hydrique (DEH): the pseudo-observations at ungauged sections and the hydroclimatic simulations. The pseudo-observations result from an interpolating method using observations from surrounding stations and simulations from the Hydrotel run-off model. It allows for estimations of riverflows in ungauged sections (without measuring devices) until 2020. The simulations come from a multi-model climate ensemble, coupled to Hydrotel to simulate future flows under different greenhouse gas emission scenarios until 2100. These simulations take climate change into account.

Two hierarchical bayesian models were developed, one for the pseudo-observations and one for the hydroclimatic simulations. Annual maximum flows are extracted and modeled within the extreme value theory framework. Hierarchical bayesian models are particularly suited for structuring similar data from several sources. In addition, they allow for estimations of extreme riverflows which account for uncertainty within a coherent statistical framework. The model output for the flow simulations is biased compared to the model output for the pseudo-observations. A statistical downscaling method is used to correct biases between the two models and estimate future extreme flows.

The proposed methodology was applied to flow data from 234 sections of the Chaudière River in Quebec. The forecast suggests a decreasing trend at sections located upstream and an increasing trend at sections located downstream, for the 100-year return level. The results of this project will be directly used by the DEH to map current and future flood risks.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

ANOVA	Analyse de la variance
DEH	Direction de l'Expertise Hydrique
GCM	Modèle global du climat
GES	Gaz à effet de serre
GEV	Valeur extrême généralisée
GIEC	Groupe d'experts intergouvernemental sur l'évolution du climat
HBM	Modèle hiérarchique bayésien
i.i.d	indépendants et identiquement distribués
MCMC	Monte-Carlo par chaînes de Markov
QQ-plot	Diagramme quantile-quantile
RCM	Modèle régional du climat
REA	Reliability Ensemble Averaging
RFA	Analyse fréquentielle régionale

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

Récemment, le Québec a été touché en 2017 et 2019 par deux inondations printanières massives ayant causé des dégâts matériaux et humains importants. En 2019, plus de 9000 résidences québécoises ont été inondées et plus de 13 000 personnes ont dû être évacuées. Cette année là, les inondations le long de la rivière des Outaouais furent reconnues comme l'événement météorologique le plus important de 2019 au Canada, par Environnement et Changement climatique Canada [1].

C'est dans ce contexte qu'en avril 2018, le ministère de l'Environnement et de la Lutte contre les changements climatiques a démarré le projet INFO-Crue. Il vise à développer et consolider les connaissances sur l'évolution des zones à risque d'inondation dans une grande partie du Québec méridional pour la production des cartes des zones inondables. Afin de contribuer à la réalisation de ces cartes, le but de ce projet consiste à estimer les débits extrêmes futurs pour l'ensemble des rivières du Québec méridional, y compris les rivières pour lesquelles aucune observation n'est disponible. Ce projet est piloté conjointement par la Direction de l'expertise hydrique (DEH) et Ouranos, consortium sur la climatologie régionale et l'adaptation aux changements climatiques. Les résultats obtenus seront exploités par un autre groupe de travail d'INFO-Crue pour calculer les débordements du lit, permettant de tracer la carte des zones inondables pour divers horizons temporels.

Pour estimer les débits futurs, une méthodologie scientifique s'appuyant sur les sciences statistique et physique et tenant compte des changements climatiques est nécessaire. En effet, il y a aujourd'hui un consensus dans la littérature sur le fait que le réchauffement climatique a des répercussions sur le cycle de l'eau, avec des impacts importants sur l'intensité et la fréquence des événements extrêmes comme les précipitations intenses et, indirectement, les inondations [2]. Les changements climatiques sont une conséquence directe de l'augmentation des gaz à effet de serre dans l'atmosphère.

Le Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC) a développé quatre scénarios d'émission des gaz à effet de serre (GES) d'ici la fin du siècle : RCP2.6, RCP4.5, RCP6.0 et RCP8.5, où les valeurs 2.6, 4.5, 6.0 et 8.5 correspondent au forçage radiatif en W/m^2 en 2100 (figure 1.1). Ils s'appuient sur un ensemble cohérent d'hypothèses sur les facteurs qui impactent les émissions d'origine humaine : les changements technologiques, la croissance démographique et le développement socio-économique. Ces quatre scénarios couvrent un large éventail de possibilités, correspondant aux efforts plus ou moins grands pour réduire les émissions mondiales de GES. Le scénario RCP2.6 suppose une stabilisation

mondiale des émissions, avec un point culminant avant 2050 (scénario optimiste). Le scénario RCP8.5, souvent dénommé *business as usual*, suppose une augmentation continue des émissions au rythme actuel (scénario pessimiste).

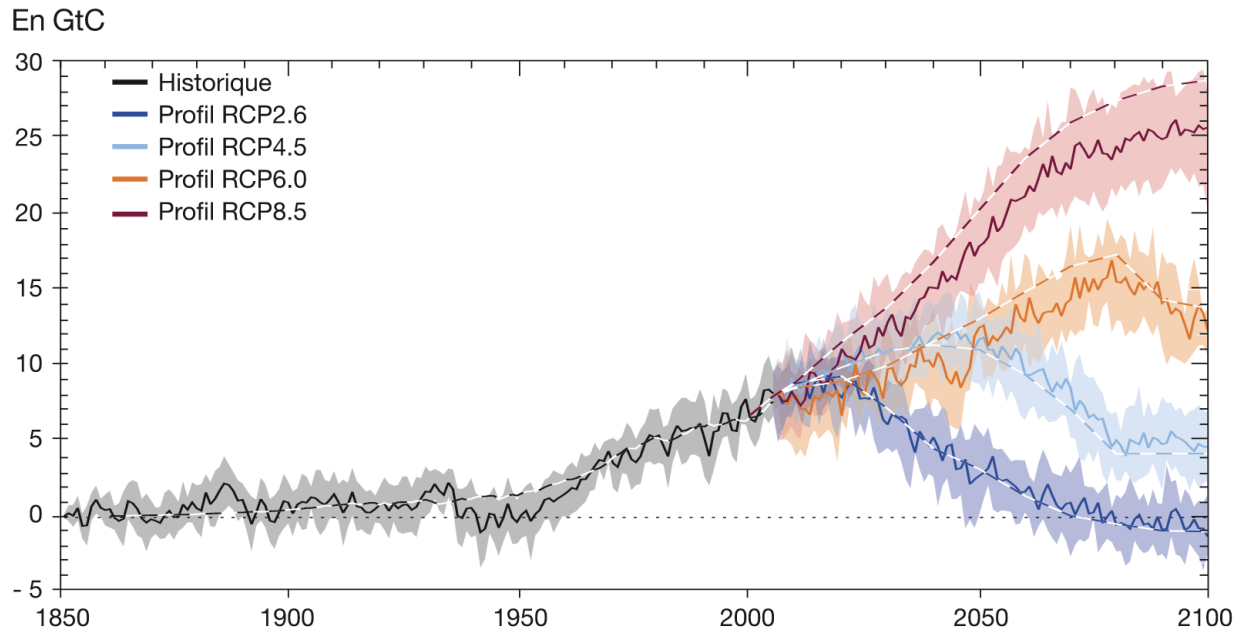


FIGURE 1.1 Quatre scénarios d'émission de GES, en gigatonne de carbone, par rapport à l'année de référence 1850. Source : GIEC, 2013.

Les scénarios d'émission de GES sont utilisés dans les modèles climatiques pour simuler les changements associés du système terrestre. Un modèle climatique est une représentation numérique de la planète et des interactions entre ses différentes composantes : l'atmosphère, l'océan, les glaciers et la biosphère continentale. Ces modèles sont indispensables pour obtenir des estimations des variables représentatives du climat futur (température, précipitation, etc.). Ils s'appuient sur la résolution des équations mathématiques qui décrivent les processus physiques du système climatique. On distingue généralement les modèles climatiques globaux (GCM) des modèles climatiques régionaux (RCM). Les premiers simulent le climat à l'échelle planétaire, avec une résolution spatiale limitée par la puissance du calcul informatique, de l'ordre d'une centaine de kilomètres. Les derniers ont été développés pour étudier le système climatique sur un plus petit domaine (une région de la planète), ce qui permet d'augmenter sa résolution spatiale, généralement de l'ordre du dizaine de kilomètres. Un RCM doit être piloté par un GCM, qui définit les phénomènes météorologiques synoptiques sur le domaine de simulation, ainsi que les conditions aux limites de ce domaine. Une telle association sera dénommée par la suite un couple GCM-RCM. Pour l'étude des débits extrêmes au Québec,

l'utilisation des modèles régionaux du climat est judicieuse car un modèle climatique global est incapable de modéliser certains phénomènes locaux dont l'échelle est plus petite que sa résolution. En particulier, les GCM ne peuvent pas représenter certaines configurations de circulation atmosphérique à l'origine des événements hydrologiques extrêmes [3].

À partir des conditions initiales du système terrestre à un moment donné et d'un scénario d'émission, la résolution des équations permet de simuler le climat jusqu'à des siècles dans le futur. En pratique, plusieurs modèles climatiques sont développés par différents laboratoires à travers le monde. Ils se distinguent par le choix des phénomènes physiques pris en compte dans les équations physiques, de la façon de les paramétrer, de la grille spatiale utilisée, etc. Pour chaque modèle climatique, plusieurs membres de simulation peuvent être disponibles. Les membres sont les simulations climatiques d'un même modèle. Dans la plupart des cas, ils se distinguent par des conditions initiales légèrement différentes. Ils rendent compte de la variabilité interne du climat, qui sont des fluctuations climatiques naturelles de nature chaotique ou causées par des phénomènes plus ou moins cycliques. Un ensemble de simulations provenant de plusieurs modèles climatiques permet alors de dresser les possibilités du climat futur. Il a été démontré que l'utilisation d'un ensemble multi-modèle pour la prévision météorologique et climatique est plus judicieuse et génère des résultats plus performants que si un seul modèle climatique était considéré [4].

Les sorties des modèles climatiques (température, précipitation, humidité, etc.) simulées à chaque point de la grille et à chaque intervalle de temps peuvent ensuite être utilisées par des modèles hydrologiques pour simuler le routage de l'eau dans les bassins versants. Un modèle hydrologique est une représentation numérique de la relation pluie-débit. Il permet de transformer des séries temporelles décrivant le climat dans un bassin versant donné en une série de débits. Selon le modèle hydrologique utilisé, la relation pluie-débit peut s'appuyer sur une modélisation empirique (basée sur l'observation, sans description du processus sous-jacent) ou physique (en utilisant les lois physiques régissant les processus hydrologiques). Forcés par les sorties des modèles climatiques, les modèles hydrologiques permettent de simuler les futurs débits dans les rivières au Québec, jusqu'à l'horizon 2100. Cependant, ces simulations hydroclimatiques ne sont pas utilisables directement car biaisées par rapport aux vraies observations de débits. Une comparaison avec des données observées sur une période historique est alors nécessaire pour corriger le biais entre les simulations et les observations.

Usuellement, les débits mesurés dans les tronçons jaugés sont utilisés comme référence pour la correction de biais des simulations hydroclimatiques. Cependant, la grande majorité des tronçons au Québec ne sont pas jaugés. Ainsi, les observations de débit n'y sont pas disponibles. Pour obtenir une estimation des débits dans ces tronçons, la DEH a développé une

méthode d'interpolation utilisant les observations des tronçons avoisinants et les débits simulés par modèles hydrologiques [5]. Ces débits interpolés sont appelés des *pseudo-observations*. Leur analyse est cruciale pour que la cartographie des zones inondables soit la plus complète possible spatialement.

L'approche développée dans ce projet pour estimer les débits extrêmes futurs permet une quantification intégrée de l'incertitude des pseudo-observations et des projections climatiques. Cette quantification correspond au niveau de confiance accordé à chaque quantité estimée, aspect d'autant plus important que les sources d'incertitudes sont nombreuses pour les débits extrêmes. Les simulations hydroclimatiques utilisées sont incertaines dû aux simplifications des phénomènes physiques et hydrologiques sous-jacents et à la variabilité interne du climat. Ainsi, nous n'avons pas accès à un seul scénario de climat futur, qui se réalisera, mais à un ensemble de projections climatiques. Les pseudo-observations issues de modèles hydrologiques s'accompagnent invariablement d'incertitudes. Par ailleurs, l'estimation des débits extrêmes est intrinsèquement incertaine, car la rareté des événements extrêmes résulte en une diminution des données exploitables. L'effet des changements climatiques sur les débits extrêmes ajoute à cette variabilité, notamment avec deux effets opposés attendus dans plusieurs régions du Québec : l'augmentation des précipitations printanières et la diminution de la couverture neigeuse.

Nous proposons donc de développer une méthodologie d'analyse fréquentielle non-stationnaire intégrant les sources d'incertitudes pour les débits projetés à chacun des tronçons surveillés du Québec méridional. Deux modèles statistiques bayésiens sont développés séparément pour analyser les pseudo-observations et les simulations hydroclimatiques. Ils sont adaptés à la prise en compte de l'incertitude avec l'accès naturel à des régions de confiance pour les estimateurs obtenus. Une méthode probabiliste de correction de biais est ensuite utilisée pour faire la jonction entre le modèle pour les simulations hydroclimatiques et le modèle pour les pseudo-observations. Ainsi, on parvient à obtenir la distribution future des débits extrêmes avec leurs intervalles de crédibilité. Les principaux objectifs du projet sont :

1. Développer un modèle statistique pour l'analyse des pseudo-observations dans les bassins non jaugés.
2. Développer un modèle statistique pour l'analyse des débits simulés par un ensemble de simulations hydroclimatiques.
3. Adapter une méthode de correction de biais pour faire la jonction des deux modèles et estimer les débits extrêmes futurs.

Les modèles développés sont codés dans le langage de programmation Julia, dans une librairie à l'intention de la DEH. Les aspects logiciels de ce projet, qui consistent à traduire en code

opérationnel les modèles statistiques développés, ont demandé un investissement important. En particulier, les algorithmes de simulation par chaînes de Markov Monte-Carlo utilisés ont été entièrement recodés.

La suite du mémoire est organisée ainsi : les principaux résultats théoriques exploités dans les modèles développés sont énoncés et expliqués au chapitre 2. Le chapitre 3 dresse une revue de littérature complète, avec pour cœur du sujet les modèles hiérarchiques bayésiens appliqués à la prédiction des extrêmes climatiques. Le chapitre 4 est consacré à la description des données. Le chapitre 5 est consacré au modèle pour les pseudo-observations. Le chapitre 6 décrit et valide le modèle pour les simulations. Le chapitre 7 décrit la méthode de post-traitement utilisée pour faire la jonction des deux modèles précédents et estimer les débits futurs. Les résultats globaux sont présentés et commentés au chapitre 8. Une discussion générale sur la méthodologie et les difficultés rencontrées est présentée au chapitre 9. Enfin, le chapitre 10 conclut en resituant le projet dans son contexte et en suggérant les pistes d'améliorations futures.

CHAPITRE 2 CADRE THÉORIQUE

2.1 Théorie des valeurs extrêmes

Pour une introduction accessible et complète à la théorie des valeurs extrêmes, le lecteur ou la lectrice intéressé.e peut se référer au livre de Coles [6].

2.1.1 Motivations et théorème fondamental

La théorie des valeurs extrêmes, développée originellement par Fisher et Tippett (1928) [7], est une branche de la statistique qui s'intéresse aux queues de distribution. Elle constitue le cadre théorique nécessaire à l'étude des quantiles élevés d'une variable aléatoire. En effet, la rareté des données observées pour les phénomènes d'amplitude extrême rend les techniques inférentielles standards peu performantes.

Soit $(X_1, X_2, \ldots, X_n)$ une suite de variables aléatoires indépendantes et identiquement distribuées (i.i.d) de fonction de répartition F. Pour n fixé, nous nous intéressons à la loi du maximum $M_n = \max(X_1, X_2, \ldots, X_n)$. Alors :

$$\begin{aligned} P(M_n \leq z) &= P(X_1 \leq z, X_2 \leq z, \ldots, X_n \leq z) \\ &= P(X_1 \leq z)^n \\ &= F^n(z) \end{aligned}$$

Une approche naïve pour estimer la loi de M_n est d'utiliser l'estimateur empirique de F :

$$P(M_n \leq z) \approx \hat{F}^n(z) \text{ où } \hat{F}(z) = \frac{card\ \{i : X_i \leq z\}}{n}.$$

Cependant, toute erreur d'approximation de $F(z)$ par $\hat{F}(z)$ serait amplifiée avec le passage à la puissance n. Ce problème est exacerbé lorsque z se rapproche de la limite supérieure du support de la loi sous-jacente. La théorie des valeurs extrêmes permet de caractériser M_n asymptotiquement (pour n grand) de manière paramétrique, sans même connaître la forme analytique de F. Le théorème fondamental de la théorie des valeurs extrêmes s'énonce ainsi [6] :

Théorème 1 (Fisher-Tippett-Gnedenko) *Soit $(X_1, X_2, \ldots, X_n)$ une suite de variables aléatoires i.i.d, et soit $M_n = \max(X_1, X_2, \ldots, X_n)$. S'il existe des suites de constantes $\{a_n >$*

$0\}$ *et* $\{b_n\}$ *telles que*

$$P\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \to G(z) \text{ pour } n \to +\infty$$

pour une distribution G *non dégénérée, alors* G *appartient à la famille de loi des Valeurs Extrêmes Généralisées (GEV) :*

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

définie sur $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$*, où* $\sigma > 0, \mu \in \mathbb{R}$ *et* $\xi \in \mathbb{R}$*.*

La loi GEV est une distribution paramétrique à trois paramètres : le paramètre de localisation μ, le paramètre d'échelle σ et le paramètre de forme ξ. Le théorème 1 fournit un résultat asymptotique sur la loi du maximum M_n normalisé par les constantes a_n et b_n. Cette normalisation est nécessaire car $P\{M_n \leq z\} \to 0$ quand $n \to +\infty$ pour tout $z < z^+$, la limite supérieure du support de F. On dit alors que la distribution M_n est asymptotiquement dégénérée. Les constantes a_n et b_n sont inconnues, mais ceci ne pose pas de problème en pratique car elles peuvent être absorbées dans les paramètres de la loi GEV à estimer. En effet, si le maximum normalisé converge vers une loi GEV de paramètres (μ, σ, ξ), le maximum non normalisé converge aussi vers une loi GEV :

$$\begin{aligned} \lim_{n\to+\infty} P\{M_n \leq z\} &= \lim_{n\to+\infty} P\left\{\frac{M_n - b_n}{a_n} \leq \frac{z - b_n}{a_n}\right\} \\ &= G\left(\frac{z - b_n}{a_n}\right) = G^*(z) \end{aligned}$$

où G^* est la loi GEV de paramètres $\mu^* = a_n\mu + b_n, \sigma^* = a_n\sigma$ et $\xi^* = \xi$. Il est donc possible d'estimer directement la distribution de M_n, sans calculer les constantes de normalisation.

Le signe du paramètre de forme ξ caractérise l'épaisseur de la queue de la distribution sous-jacente. Lorsque $\xi < 0$, la loi limite G est de type Weibull, avec une queue de distribution à droite à support fini. Lorsque $\xi = 0$, la loi est de type Gumbel, avec une queue de distribution légère (décroissance exponentielle). Lorsque $\xi > 0$, la loi est de type Fréchet, avec une queue de distribution épaisse, correspondant à une plus forte propension des valeurs extrêmes. Il est possible aussi de paramétrer une loi GEV avec le logarithme du paramètre d'échelle $\phi = \log(\sigma)$, qui prend alors ses valeurs dans $\mathbb{R}$.

En climatologie et hydrologie, la théorie des valeurs extrêmes est fréquemment utilisée pour estimer le risque associé aux événements rares. C'est ce qui est communément appelé *l'ana-*

lyse fréquentielle : à un événement extrême d'intensité donnée, on cherche à associer une probabilité de dépassement. Le lecteur ou la lectrice intéressé.e peut consulter l'article de Katz [8] pour une introduction à la théorie des valeurs extrêmes appliquée à l'hydrologie.

2.1.2 Modèle des maxima par bloc

Le modèle des maxima par bloc est une application directe du théorème 1. Il est utilisé comme la base des modèles développés dans ce projet. Il consiste à partitionner les données dans des blocs de taille n suffisamment grande et d'en extraire une série de maxima, pour laquelle la loi GEV peut être une approximation raisonnable.

Par exemple, si les données sont des observations journalières et que la taille d'un bloc correspond à une année ($n = 365$), le modèle génère la suite suivante de maxima annuels :

$$M_i = \max(X_{i1}, \ldots, X_{in}) \quad \text{pour } i = 1, \ldots, T$$

où

- T est le nombre d'années,
- M_i est le maximum annuel de l'année i,
- X_{ij} est l'observation du jour j de l'année i.

Les maxima $(M_i, \ldots, M_T)$ sont alors modélisés comme des réalisations i.i.d d'une distribution GEV, d'après le théorème 1. Ici, il est important que la taille du bloc soit assez grande pour que l'approximation asymptotique du théorème 1 soit raisonnable, mais pas trop pour limiter la perte de donnée. En effet, la quantité d'observations exploitables est inversement proportionnelle à la taille du bloc. Dans le cas des variables climatiques mesurées sur une base journalière, la taille de bloc usuelle est une année.

Le théorème énoncé précédemment s'applique à des variables aléatoires indépendantes. Cette hypothèse est peu réaliste pour des situations réelles. En effet, les données environnementales sont souvent corrélées : les précipitations intenses peuvent s'étaler sur plusieurs jours, les fortes températures arrivent par vague. Cependant, il est raisonnable de supposer que deux observations sont indépendantes si elles sont suffisamment éloignées dans le temps. Cette condition de dépendance temporelle à courte portée, appelée condition $D(u_n)$, est en fait suffisante pour que le résultat du théorème 1 demeure valide [6].

Nous terminons en énonçant la formule pour calculer le quantile d'ordre $1 - p$ d'une loi GEV, noté q_{1-p}. Il est aussi appelé le niveau de retour correspondant à la période de retour $T = 1/p$. C'est le niveau dépassé en moyenne une fois toutes les T unités de temps, ou de manière équivalente, avec probabilité p. Il s'obtient facilement par inversion de la distribution

GEV du théorème 1 :

$$q_{1-p} = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{-\log(1-p)\}^{-\xi}\right] & \text{si} \quad \xi \neq 0, \\ \mu - \sigma \log\{-\log(1-p)\} & \text{si} \quad \xi = 0, \end{cases} \tag{2.1}$$

Dans le cas des maxima annuels, q_{1-p} est le niveau dépassé en moyenne une fois toutes les T années. Une formulation équivalente est que pour une année donnée, la probabilité que la valeur de la variable sous-jacente excède q_{1-p} est de p.

2.1.3 Extension non-stationnaire

Le modèle des maxima par bloc est à l'origine développé pour une série stationnaire, c'est-à-dire une série où la loi des variables aléatoires est indépendante de l'indice temporel.

Définition 2.1.1 *Un processus aléatoire $X_1, X_2, ...$ est dit stationnaire si, étant donné un ensemble d'entiers $i_1, \ldots, i_k$ et un entier m, la distribution conjointe de $(X_{i_1}, \ldots, X_{i_k})$ est identique à celle de $(X_{i_{1+m}}, \ldots, X_{i_{k+m}})$*

Il est possible d'étendre le modèle des maxima par bloc à des séries non-stationnaires. En théorie des valeurs extrêmes, une approche simple et efficace consiste à intégrer cette non-stationnarité dans les paramètres de la loi GEV caractérisant les variables aléatoires.

Dénotons $(M_1, \ldots, M_T)$ les maxima annuels comme dans la section précédente. Pour $t = 1, \ldots, T$, soit $\boldsymbol{u}_t$ le vecteur des variables explicatives pour l'année t. Dans un modèle GEV non-stationnaire, les paramètres GEV décrivant chaque maximum annuel sont alors fonction de l'année en question :

$$M_t \sim \mathcal{G}EV\left\{\mu(\boldsymbol{u}_t), \sigma(\boldsymbol{u}_t), \xi(\boldsymbol{u}_t)\right\} \qquad \text{pour } t = 1, \ldots, T$$

Ce modèle est assez flexible, car la relation de dépendance entre les paramètres GEV et les variables explicatives peut être choisie en fonction de l'application. Par exemple, une relation linéaire est couramment utilisée pour exprimer la dépendance entre les paramètres de localisation et d'échelle de la loi GEV et les variables explicatives. Le choix d'imposer le paramètre de forme ξ constant est aussi une pratique usuelle. En effet, l'estimation de ce paramètre est notoirement difficile et une non-stationnarité sur ξ augmenterait considérablement l'incertitude d'estimation, en plus d'engendrer des problèmes d'identifiabilité pour l'ensemble des paramètres à estimer.

La non-stationnarité des paramètres de la loi GEV entraine celle du niveau de retour associé

à la période de retour $T = 1/p$. Ainsi, le niveau de retour *effectif* [8] pour l'année t s'écrit :

$$q_{1-p,t} = \begin{cases} \mu(\boldsymbol{u}_t) - \frac{\sigma(\boldsymbol{u}_t)}{\xi(\boldsymbol{u}_t)} \left[1 - \{-\log(1-p)\}^{-\xi(\boldsymbol{u}_t)}\right] & \text{si} \quad \xi(\boldsymbol{u}_t) \neq 0, \\ \mu(\boldsymbol{u}_t) - \sigma(\boldsymbol{u}_t) \log\{-\log(1-p)\} & \text{si} \quad \xi(\boldsymbol{u}_t) = 0, \end{cases} \tag{2.2}$$

C'est le niveau associé à la période de retour T à l'année t, si les conditions de l'année t se poursuivaient indéfiniment.

Après qu'un modèle des valeurs extrêmes soit défini, l'estimation des paramètres peut ensuite se faire par la méthode du maximum de la vraisemblance ou par inférence bayésienne.

2.2 Théorie de l'inférence bayésienne

2.2.1 Inférence bayésienne

En statistique, on désigne par le terme *inférence* l'ensemble des techniques permettant d'induire les caractéristiques inconnues d'une population à partir d'un échantillon de données observées issu de cette population. Lorsqu'un modèle statistique est paramétrique, l'inférence consiste à estimer les paramètres de la distribution utilisée pour modéliser la population. Une classe importante de méthodes inférentielles largement utilisées s'appuie sur la vraisemblance.

Définition 2.2.1 *Soit $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ un échantillon aléatoire de taille n, i.i.d selon la densité f de paramètre $\boldsymbol{\theta} \in \Theta$ où Θ est un compact de $\mathbb{R}^p$. Alors conditionnellement à $(\boldsymbol{Y} = \boldsymbol{y})$, la fonction de $\boldsymbol{\theta}$ définie par*

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = f_{(\boldsymbol{Y}|\boldsymbol{\theta})}(\boldsymbol{y}) = \prod_{i=1}^{n} f_{(Y_i|\boldsymbol{\theta})}(y_i)$$

est appelée la fonction de vraisemblance.

L'inférence bayésienne est une approche d'estimation de $\boldsymbol{\theta}$ qui lui associe une distribution de probabilité sur l'espace Θ. Ainsi, dans tout ce qui suit, il est important de se rappeler que le paramètre à inférer $\boldsymbol{\theta}$ est considéré comme une **variable aléatoire**. L'inférence bayésienne s'appuie sur la règle de Bayes, énoncée ci-dessous.

Lemme 2.2.1 *Soit X et Y deux variables aléatoires de distributions conditionnelle $f_{(Y|X=x)}(y)$ et marginale $f_X(x)$, alors la distribution de X sachant $Y = y$ peut s'écrire*

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

La règle de Bayes sert à relier la loi *a posteriori* du paramètre, notée $f_{(\boldsymbol{\theta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\theta})$, à la vraisemblance $L(\boldsymbol{\theta}|\boldsymbol{y})$ et à la loi *a priori* du paramètre, noté $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$:

$$f_{(\boldsymbol{\theta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\theta}) = \frac{f_{(\boldsymbol{Y}|\boldsymbol{\theta})}(\boldsymbol{y}) f_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} f_{(\boldsymbol{Y}|\boldsymbol{\theta})}(\boldsymbol{y}) f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) d\boldsymbol{\theta}} = \frac{L(\boldsymbol{\theta}|\boldsymbol{y}) f_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{y}) f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) d\boldsymbol{\theta}} \propto L(\boldsymbol{\theta}|\boldsymbol{y}) f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \tag{2.3}$$

La loi *a priori* est une distribution de probabilité sur $\boldsymbol{\theta}$ qui représente l'information disponible sur $\boldsymbol{\theta}$ avant d'avoir vu les données. La loi *a posteriori* combine la loi *a priori* et la vraisemblance, qui dépend des données. On dit alors que la loi *a posteriori* est la distribution actualisée de $\boldsymbol{\theta}$, après que l'information apportée par les données ait été prise en compte : elle provient d'une part de l'information *a priori* disponible, d'autre part de l'information apportée par les données via la fonction de vraisemblance. En inférence bayésienne, c'est la loi *a posteriori* qui résume toute l'information disponible sur $\boldsymbol{\theta}$. Elle se distingue de l'inférence classique qui produit un estimateur ponctuel et non une distribution de probabilité.

À partir de la loi *a posteriori*, plusieurs estimateurs ponctuels peuvent être déduits, tels que l'espérance :

$$E(\boldsymbol{\theta}|\boldsymbol{Y} = \boldsymbol{y}) = \int_{\boldsymbol{\theta}} \boldsymbol{\theta} f_{(\boldsymbol{\theta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

ou le mode :

$$\hat{\boldsymbol{\theta}}_{MAP} = \sup_{\boldsymbol{\theta} \in \boldsymbol{\theta}} f_{(\boldsymbol{\theta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\theta})$$

L'information apportée par les données existantes $\boldsymbol{Y} = \boldsymbol{y}$ peut être utilisée pour prédire une nouvelle donnée inconnue Y' issue de la même population. La distribution de Y' s'appelle alors la loi *a posteriori* prédictive :

$$f_{(Y'|\boldsymbol{Y}=\boldsymbol{y})}(y') = \int_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|y') f_{(\boldsymbol{\theta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

Le choix de la loi *a priori* est crucial en inférence bayésienne. Si de l'information préalable est disponible sur les paramètres, le modélisateur peut choisir une loi *a priori* informative. Il est usuel d'utiliser une loi dite conjuguée, qui facilite les calculs.

Définition 2.2.2 *Une famille $\mathcal{F}$ de lois sur $\boldsymbol{\theta} \in \Theta$ est dite conjuguée si, pour toute loi a priori $f_{\boldsymbol{\theta}}$ appartenant à cette famille, la loi a posteriori $f_{(\boldsymbol{\theta}|\boldsymbol{Y}=\boldsymbol{y})}$ appartient également à celle-ci.*

A titre d'exemple, lorsque la vraisemblance est modélisée par une loi normale de variance connue, une famille de loi conjuguée pour le paramètre de la moyenne noté θ est celle des lois normales $\mathcal{N}(\mu, \tau^2)$. Si l'on dispose de l'information préalable que θ est proche de 3, on pourrait

alors choisir $\mu = 3$. Si le modélisateur n'accorde pas beaucoup de poids à cette information, il peut choisir d'utiliser une loi *a priori* vague ou peu informative. Une façon usuelle de procéder est de régler les paramètres d'une loi conjuguée. Par exemple, si l'information précédente est incertaine, une loi *a priori* plausible est $\mathcal{N}(3, 10)$. Ainsi, on exprime la croyance que le paramètre est autour de 3, mais n'exclut pas la possibilité des valeurs loin de 3.

Quand aucune information préalable est disponible, il est préférable d'utiliser une loi *a priori* non informative, qui n'introduit pas de biais de la part du modélisateur. Dans ce cas, l'inférence de $\boldsymbol{\theta}$ repose entièrement sur l'information apportée par les données observées. Une loi non informative usuellement utilisée est la loi de Jeffreys :

Définition 2.2.3 *La loi a priori de Jeffreys est donnée par*

$$f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \propto \sqrt{\det(I(\boldsymbol{\theta}))}$$

où $I(\theta)$ est la matrice d'information de Fischer définie par

$$I_{ij}(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i\theta_j}\log L(\boldsymbol{\theta}|\boldsymbol{Y})\right], \quad 1 \leq i, j \leq p$$

Ici, l'espérance est définie par rapport à la distribution de la population $\boldsymbol{Y}$. La loi de Jeffreys est attrayante pour sa neutralité et d'autres propriétés comme l'invariance par reparamétrisation des paramètres. Cependant, elle peut être impropre : l'intégrale de sa densité est infinie. Ceci ne pose pas de problème à condition que la loi *a posteriori* résultante est propre (l'intégrale de sa densité est finie). Dans le cas contraire, une approche alternative est d'utiliser une loi *a priori* propre mais très peu informative, comme par exemple $\mathcal{N}(0, 10^2)$. La loi *a posteriori* résultante sera nécessairement propre, donc une vérification n'est pas nécessaire.

Le principal avantage de l'approche bayésienne par rapport aux méthodes classiques est la quantification naturelle de l'incertitude sur les paramètres à estimer. L'obtention d'une distribution sur $\boldsymbol{\theta}$ et non une valeur ponctuelle fournit des estimations par intervalles sans effort supplémentaire. L'incertitude contenue dans la distribution *a posteriori* inférée peut être propagée à des modèles d'impact ultérieurs s'ils dépendent des paramètres estimés. Ainsi, l'inférence bayésienne est très utile pour la prise de décision en contexte d'incertitude. Pour le lecteur ou la lectrice intéressé.e, une très bonne référence sur le sujet est [9].

2.2.2 Modèle hiérarchique bayésien

Un modèle hiérarchique bayésien (HBM) s'appuie sur la théorie de l'inférence bayésienne et introduit (généralement) un niveau supplémentaire dans le modèle, appelé la couche latente. Il est particulièrement adapté lorsque les données présentent des variations qui peuvent être structurées par couches, pour modéliser de manière simple des dépendances complexes entre les observations.

Dans un modèle hiérarchique bayésien, on peut distinguer les paramètres, notés $\boldsymbol{\theta}$, dont dépendent directement les données, des hyperparamètres, notés $\boldsymbol{\Psi}$, qui sont tous les autres paramètres des niveaux supérieurs. La loi *a posteriori* s'écrit alors :

$$f_{(\boldsymbol{\theta},\boldsymbol{\Psi}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\theta},\boldsymbol{\Psi}) \propto f_{(\boldsymbol{Y}|\boldsymbol{\theta})}(\boldsymbol{y}) f_{(\boldsymbol{\theta}|\boldsymbol{\Psi})}(\boldsymbol{\theta}) f_{\boldsymbol{\Psi}}(\boldsymbol{\Psi}) \tag{2.4}$$

où

- $f_{(\boldsymbol{\theta},\boldsymbol{\Psi}|\boldsymbol{Y}=\boldsymbol{y})}$ est la loi *a posteriori* des paramètres $\boldsymbol{\theta}$ et hyperparamètres $\boldsymbol{\Psi}$, à inférer,
- $f_{(\boldsymbol{Y}|\boldsymbol{\theta})}$ est la vraisemblance (couche des données),
- $f_{(\boldsymbol{\theta}|\boldsymbol{\Psi})}$ est la loi conditionnelle qui structure la couche latente,
- $f_{\boldsymbol{\Psi}}$ est la loi *a priori* sur les hyperparamètres.

Une représentation schématique d'un modèle hiérarchique bayésien est présentée à la figure 2.1.

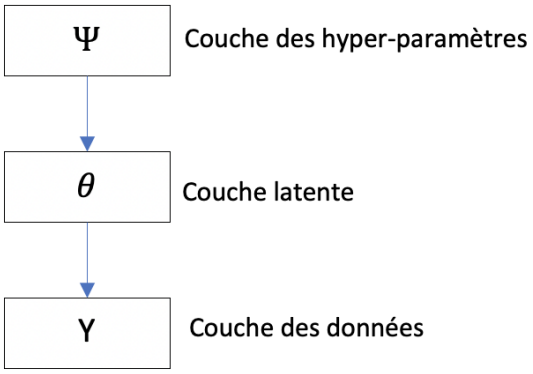


FIGURE 2.1 Schéma du modèle hiérarchique bayésien.

La couche latente, composante cruciale dans le modèle, décrit la structure de dépendance des paramètres utilisés dans la couche des données, en s'aidant des hyperparamètres. La couche des hyperparamètres décrit les lois *a priori* choisies pour tous les hyperparamètres du modèle. En fait, ces deux couches résultent de la décomposition de la loi *a priori* d'un modèle bayésien classique. En effet, en intégrant les deux côtés de l'équation 2.4 par rapport

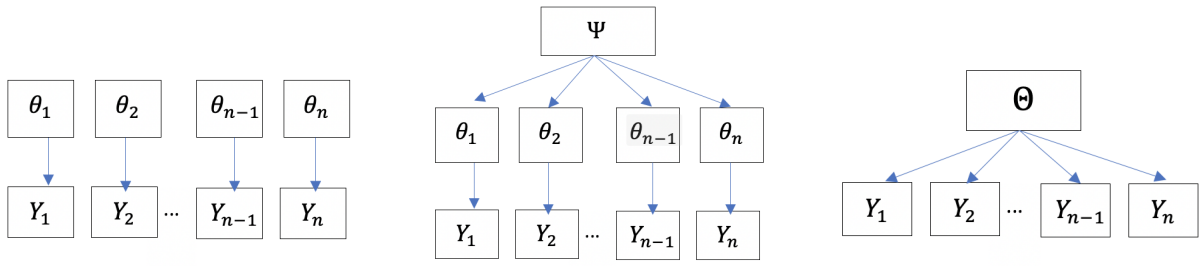
à $\boldsymbol{\Psi}$ et en notant $f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \int f_{(\boldsymbol{\theta}|\boldsymbol{\Psi})}(\boldsymbol{\theta}) f_{\boldsymbol{\Psi}}(\boldsymbol{\Psi}) d\boldsymbol{\Psi}$, on obtient :

$$f_{(\boldsymbol{\theta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\theta}) \propto f_{(\boldsymbol{Y}|\boldsymbol{\theta})}(\boldsymbol{y}) f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$$

Nous reconnaissons ici la règle de Bayes pour l'inférence bayésienne.

Un des avantages majeurs du modèle hiérarchique bayésien est la possibilité de mutualiser l'information pour augmenter la précision des estimations. Lorsque les données ont des caractéristiques en commun, leur structuration permet de tirer profit de toute l'information disponible. Par exemple, des données météorologiques sont mesurées à deux sites proches A et B, où A dispose de plus d'observations que B. Dans ce cas, la couche latente peut faire le lien entre l'information des deux sites. Ainsi, l'information provenant du site A profite aussi au site B, et l'estimation des quantités d'intérêt au site B devient plus précise.

Le modèle hiérarchique bayésien suppose l'hypothèse d'indépendance conditionnelle des données : conditionnellement à la structure de $\boldsymbol{\theta}$ décrite dans la couche latente par $f_{(\boldsymbol{\theta}|\boldsymbol{\Psi})}(\boldsymbol{\theta})$, les données sont indépendantes. Cette hypothèse est une façon simplifiée de modéliser la dépendance entre les données. Elle facilite l'écriture de la loi *a posteriori*, car la vraisemblance dans la couche des données s'écrit comme un produit de vraisemblances pour chaque observation. Une structure de dépendance plus complexe rendrait l'inférence et l'interprétation des résultats beaucoup plus difficiles. Ainsi, on peut dire que l'hypothèse d'indépendance conditionnelle est un bon compromis entre une structure de dépendance totale, réaliste mais complexe, et une structure d'indépendance totale, naïve mais simple à mettre en oeuvre. Une comparaison schématique des trois hypothèses est représentée à la figure 2.2.



(a) Indépendance totale (b) Indépendance conditionnelle (HBM) (c) Dépendance totale

FIGURE 2.2 Graphe directif de trois hypothèses de dépendance des données.

La liberté du choix de la loi conditionnelle dans la couche latente rend le modèle hiérarchique bayésien très flexible, pouvant décrire des situations simples ou plus complexes. En théorie, la loi *a priori* $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ peut être décomposée en une succession de k couches conditionnelles

séquentielles, en fonction de la complexité du problème à traiter. Dans ce cas, si l'on note $\boldsymbol{\Phi} = (\boldsymbol{\theta}, \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k)$, on peut écrire alors :

$$f_{\boldsymbol{\Phi}}(\boldsymbol{\Phi}) = f_{(\boldsymbol{\theta}|\boldsymbol{\psi}_1)}(\boldsymbol{\theta}) f_{(\boldsymbol{\psi}_1|\boldsymbol{\psi}_2)}(\boldsymbol{\psi}_1) \ldots f_{(\boldsymbol{\psi}_{k-1}|\boldsymbol{\psi}_k)}(\boldsymbol{\psi}_{k-1}) f_{\boldsymbol{\psi}_k}(\boldsymbol{\psi}_k)$$

En pratique, un modèle hiérarchique bayésien à trois niveaux ($k = 1$) est amplement suffisant dans la plupart des problèmes de modélisation. Comparé à d'autres modèles fréquentistes, un modèle hiérarchique bayésien possède une structure qui facilite la compréhension des liens entre les données et l'interprétation des sorties du modèle. Pour l'inférence, la modélisation à l'aide des lois conditionnelles est particulièrement adaptée aux méthodes de simulation Monte-Carlo par chaîne de Markov. Ces avantages contribuent à la popularité des modèles hiérarchiques et encouragent des recherches dans ce domaine.

2.3 Méthodes Monte-Carlo par chaîne de Markov

2.3.1 Théorie

Les méthodes Monte-Carlo par chaîne de Markov (MCMC) sont une classe de méthodes de simulation massivement utilisée en inférence bayésienne. Elles permettent de contourner les difficultés liées à la simulation des lois *a posteriori*, généralement impossibles à simuler directement. C'est la plupart du temps le cas pour la loi *a posteriori* d'un modèle hiérarchique bayésien. Ces méthodes génèrent un échantillon aléatoire de la loi *a posteriori* en construisant une chaîne de Markov avec cette loi comme loi stationnaire. Elles ne nécessitent pas la connaissance de la constante de normalisation dans la règle de Bayes (équation 2.3), qui est souvent très difficile à calculer.

Une exposition complète de la théorie derrière les chaînes de Markov serait trop longue, nous rappelons ici seulement sa définition :

Définition 2.3.1 *Une suite $(X_n)_{n\geq 0}$ de variables aléatoires est une chaîne de Markov (d'ordre 1) si et seulement si pour tout $k \in \mathbb{N}$, pour tout $(x_0, ..., x_{k+1})$ tels que $P(X_k = x_k, \ldots, X_0 = x_0) > 0$,*

$$P(X_{k+1} = x_{k+1} | X_k = x_k, \ldots, X_0 = x_0) = P(X_{k+1} = x_{k+1} | X_k = x_k)$$

La convergence d'une chaîne de Markov (X_n) vers sa loi stationnaire π (si elle existe) signifie que (X_n) converge en loi vers π. Des conditions supplémentaires sur la chaîne (X_n) sont requises pour garantir l'existence d'une loi stationnaire et la convergence de la chaîne vers cette loi. Elles ne sont pas explicitées ici, mais le lecteur ou la lectrice intéressé.e peut se référer au chapitre 6 de [10].

Les deux principaux algorithmes de simulation MCMC fréquemment utilisés sont l'échantillonnage de Gibbs et l'algorithme de Metropolis-Hastings. Ces algorithmes construisent une chaîne de Markov avec des propriétés particulières qui garantissent la convergence en loi vers la distribution d'intérêt. En pratique, chaque exécution d'un algorithme MCMC produit une réalisation d'une chaîne de Markov $(X_1 = x_1, \ldots, X_n = x_n)$, où n est généralement choisi très grand. Une période de préchauffage est définie par un entier $1 \leq m < n$, où l'on considère que la chaîne a convergé vers sa loi stationnaire à partir de $X_{m+1} = x_{m+1}$. Alors $(x_{m+1}, \ldots, x_n)$ peut être vu comme un tirage aléatoire selon la loi d'intérêt. En inférence bayésienne, l'ensemble $\{x_{m+1}, \ldots, x_n\}$ représente la loi *a posteriori*, et peut être utilisé pour calculer toutes les quantités d'intérêt. Si l'on note $\boldsymbol{\theta}_{pos}$ les paramètres *a posteriori* et g une fonction déterministe, alors $\{g(x_{m+1}), \ldots, g(x_n)\}$ est un tirage aléatoire de $g(\boldsymbol{\theta}_{pos})$. Un estimateur de $E\left[g(\boldsymbol{\theta}_{pos})\right]$ est alors simplement

$$E\left[g(\boldsymbol{\theta}_{pos})\right] \approx \frac{1}{n-m} \sum_{i=m+1}^{n} g(x_i) \tag{2.5}$$

Échantillonage de Gibbs La méthode d'échantillonage de Gibbs est particulièrement adaptée pour la simulation de lois multidimensionnelles, à partir des lois conditionnelles associées. Notons $f_{\boldsymbol{\theta}}(\theta_1, \ldots, \theta_d)$ la densité à simuler. Chaque θ_i peut être unidimensionnel ou multidimensionnel. La méthode construit une chaîne de Markov en simulant séquentiellement selon les lois conditionnelles complètes

$$f_{(\theta_1|\theta_2,\ldots,\theta_d)}, f_{(\theta_2|\theta_1,\theta_3,\ldots)}, \ldots, f_{(\theta_d|\theta_1,\ldots,\theta_{d-1})}$$

Elle suppose que l'on arrive à simuler plutôt facilement ces lois conditionnelles complètes, qui sont plus simples que $f_{\boldsymbol{\theta}}$ car de plus petite dimension.

Nous décrivons ci-dessous une étape i de l'algorithme :

Algorithme 1 : Algorithme de Gibbs

$\boldsymbol{\theta} \leftarrow (\theta_1^i, \ldots, \theta_d^i)$
simuler $\theta_1^{i+1} \sim f_{(\theta_1|\theta_2^i,\ldots,\theta_d^i)}$
simuler $\theta_2^{i+1} \sim f_{(\theta_2|\theta_1^{i+1},\theta_3^i\ldots,\theta_d^i)}$
...
simuler $\theta_d^{i+1} \sim f_{(\theta_d|\theta_1^{i+1},\ldots,\theta_{d-1}^{i+1})}$
$\boldsymbol{\theta} \leftarrow (\theta_1^{i+1}, \ldots, \theta_d^{i+1})$

Algorithme Metropolis-Hastings L'algorithme Metropolis-Hastings utilise une loi de proposition conditionnelle de densité arbitraire (avec néanmoins certaines propriétés requises pour assurer la convergence de l'algorithme) pour générer un nouvel état de la chaîne avec une certaine probabilité. Dénotons $q(-|\boldsymbol{\theta})$ cette loi de proposition et $f_{\boldsymbol{\theta}}$ la densité à simuler. Nous décrivons ci-dessous une étape i de l'algorithme Metropolis-Hastings.

Algorithme 2 : Algorithme Metropolis-Hastings

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^i$
simuler $\boldsymbol{\theta}' \sim q(-|\boldsymbol{\theta}^i)$
$\alpha \leftarrow \min\left\{\dfrac{f_{\boldsymbol{\theta}}(\boldsymbol{\theta}')q(\boldsymbol{\theta}^i|\boldsymbol{\theta}')}{f_{\boldsymbol{\theta}}(\boldsymbol{\theta}^i)q(\boldsymbol{\theta}'|\boldsymbol{\theta}^i)}, 1\right\}$
si $\alpha > rand()$ **alors**
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}'$
sinon
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^i$
fin

Algorithme Metropolis-dans-Gibbs L'algorithme Metropolis-dans-Gibbs combine les deux méthodes d'échantillonage décrites précédemment pour construire une chaîne de Markov convergeant vers la loi d'intérêt. C'est essentiellement un algorithme de Gibbs, où une ou plusieurs étapes de l'algorithme Metropolis-Hastings sont utilisées pour simuler une ou plusieurs composantes du paramètre multi-dimensionnel d'intérêt. Ces opérations sont nécessaires lorsque les lois conditionnelles complètes correspondantes ne sont pas simples à simuler directement. Il s'avère que la chaîne de Markov ainsi construite garde les même propriétés désirables, notamment la convergence vers une loi stationnaire qui est la loi *a posteriori* à estimer [10, 11].

À titre d'exemple, nous décrivons dans l'algorithme 3 une étape i de l'algorithme Metropolis-dans-Gibbs, pour une densité bidimensionnelle $f_{\boldsymbol{\theta}}(\theta_1, \theta_2)$. Nous supposons que $f_{(\theta_1|\theta_2)}$ peut être simulée directement alors que $f_{(\theta_2|\theta_1)}$ est difficile à simuler. Une étape Metropolis-Hastings est alors utilisée pour simuler selon $f_{(\theta_2|\theta_1)}$. On note $q(-|\theta)$ la loi de proposition.

Cette méthode sera utilisée pour inférer les modèles développés dans ce projet.

2.3.2 Algorithme MCMC adaptif

L'algorithme Metropolis-Hastings suppose d'accepter ou de rejeter une valeur simulée selon la loi de proposition, à chaque état de la chaîne MCMC. Le taux d'acceptation (empirique)

Algorithme 3 : Algorithme Metropolis-dans-Gibbs

$\boldsymbol{\theta} \leftarrow (\theta_1^i, \theta_2^i)$
simuler $\theta_1^{i+1} \sim f_{(\theta_1|\theta_2^i)}$
simuler $\theta' \sim q(-|\theta_2^i)$
$\alpha \leftarrow \min \left\{ \dfrac{f_{(\theta_2|\theta_1^{i+1})}(\theta') q(\theta_2^i|\theta')}{f_{(\theta_2|\theta_1^{i+1})}(\theta_2^i) q(\theta'|\theta_2^i)}, 1 \right\}$
si $\alpha > rand()$ **alors**
  $\theta_2^{i+1} \leftarrow \theta'$
sinon
  $\theta_2^{i+1} \leftarrow \theta_2^i$
fin
$\boldsymbol{\theta} \leftarrow (\theta_1^{i+1}, \theta_2^{i+1})$

de l'algorithme peut être défini comme la proportion des valeurs simulées acceptées tout au long de la chaîne. Ce taux dépend intrinsèquement de la loi de proposition choisie.

Traitons d'abord le cas unidimensionnel. Il est courant d'utiliser comme loi de proposition une marche aléatoire avec un pas normal :

$$\theta' = \theta_i + \epsilon \quad \text{où } \epsilon \sim \mathcal{N}(0, \delta_\theta^2)$$

- θ_i est l'état actuel du paramètre à simuler, à l'étape i,
- θ' est le candidat proposé pour l'état suivant du paramètre,
- ϵ est le pas aléatoire.

Le nouvel état de la chaîne résulte alors d'une exploration locale autour de l'état actuel. Pour que l'algorithme MCMC converge rapidement vers sa loi stationnaire, il est important que le pas de la marche aléatoire δ_θ soit convenablement choisi. Un pas trop petit rend l'exploration limitée : la chaîne n'explore pas assez l'espace possible des paramètres. On dit que la marche aléatoire se déplace trop lentement sur la surface de la loi à simuler [10]. À l'inverse, un pas trop grand peut entraîner un taux de rejet important des valeurs candidates. L'algorithme perd alors en efficacité. Le principe de la méthode MCMC adaptive [10, 12] vise à maintenir un taux d'acceptation empirique autour d'une valeur prédéterminée et jugée optimale, en ajustant régulièrement le pas δ_θ. Rosenthal a montré que cette valeur est 0.44 pour le cas unidimensionnel, et approximativement 0.234 lorsque la dimension du paramètre tend vers l'infini [12]. En pratique, une valeur entre 0.3 et 0.5 est adaptée dans la plupart des cas.

Les résultats énoncés plus haut s'appliquent aussi pour l'algorithme Metropolis-dans-Gibbs, où la définition du taux d'acceptation s'étend naturellement. Dans les modèles hiérarchiques bayésiens, il est commun que la loi d'intérêt soit de haute dimension ($d > 100$). Dans ce cas,

si l'on simule les composantes unidimensionnelles l'une après l'autre dans l'étape Metropolis-Hastings, la chaîne peut avoir du mal à converger rapidement [10]. Ceci peut être dû à des corrélations importantes entre les composantes, un problème qui arrive souvent en haute dimension. Par exemple, pour une loi GEV, le paramètre de forme ξ est très sensible aux valeurs des deux autres paramètres, en particulier le paramètre d'échelle [13]. Pour accélérer la convergence de l'algorithme Metropolis-dans-Gibbs, on peut simuler plusieurs composantes du paramètre d'intérêt simultanément. Dans ce cas, une loi de proposition courante pour l'étape Metropolis-Hastings est la marche aléatoire multidimensionnelle de pas normale. Un résultat d'optimalité dans le cas d'une loi stationnaire normale [12] motive le choix suivant pour la matrice de covariance de la marche aléatoire :

$$\Sigma_{\boldsymbol{\theta}} = \delta\Sigma \quad \delta > 0$$

où Σ est la vraie matrice de covariance des composantes à simuler. Cette matrice étant inconnue en pratique, nous l'approximons par la matrice de covariance empirique.

Ainsi, à l'étape $i+1$ de l'algorithme Metropolis-dans-Gibbs, on simule un nouveau candidat :

$$\boldsymbol{\theta}' = \boldsymbol{\theta}_i + \boldsymbol{\epsilon} \quad \text{où } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \delta_i \hat{\Sigma}_i)$$

- $\boldsymbol{\theta}_i$ est l'état actuel des composantes du paramètre à simuler,
- $\hat{\Sigma}_i = \mathrm{Cov}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_i)$ est la matrice de covariance empirique de ces composantes, calculée avec les i premières itérations MCMC,
- δ_i est le pas de la marche aléatoire à ajuster selon la méthode adaptive décrite dans [12], afin d'avoir un taux d'acceptation proche de 0.44.

Notons que le fait de changer le pas de la marche aléatoire dans la loi de proposition modifie les propriétés de la chaîne de Markov originelle, a priori. Néanmoins, l'algorithme MCMC adaptif construit dans le cadre de ce projet satisfait les deux conditions suffisantes qui préservent la convergence de la chaîne MCMC vers sa loi limite. Une discussion plus détaillée sur ces conditions se trouve dans [12].

2.4 Méthodes de post-traitement statistique

2.4.1 Méthode delta, ajustement par quantile

Le post-traitement statistique vise à corriger le biais d'une variable aléatoire par rapport à une autre qui sert de référence. Pour l'étude du climat, ceci consiste à corriger les variables en sortie des modèles climatiques pour qu'elles reflètent plus fidèlement les caractéristiques du

climat réellement observé [14]. Cette correction est nécessaire car la simplification de certains processus physiques et la résolution spatiale limitée dans les modèles climatiques engendrent des différences entre les quantités simulées et observées à l'échelle locale. Plusieurs méthodes de post-traitement existent. La plus simple est la méthode delta, qui consiste à modifier la série d'observations en appliquant un delta de changement climatique déduit à partir des données simulées. Mathématiquement similaires mais conceptuellement différentes, les méthodes de correction additive et multiplicative (*linear scaling*) consistent à appliquer à la série simulée une translation (addition) ou mise à l'échelle (multiplication) par un facteur qui représente le biais entre les séries simulée et observée. Ces méthodes ajustent seulement la moyenne de la série simulée. Pour ajuster les moments d'ordre supérieur (variance, asymétrie), des approches de post-traitement plus flexibles sont nécessaires.

L'ajustement par quantile (*quantile mapping*) consiste à faire correspondre la fonction de répartition de la variable aléatoire à corriger à celle de la variable aléatoire de référence. Il est connu pour mieux corriger les fréquences des variables d'intérêt que les approches précédentes. Dans le cas d'une simulation climatique, une valeur simulée (qui est un quantile de la distribution à corriger) est remplacée par le quantile de la distribution des observations correspondant à la même probabilité :

$$\begin{aligned} F_Y(\tilde{x}) &= F_X(x) \\ \tilde{x} &= F_Y^{-1}\{F_X(x)\} \end{aligned}$$

— F_Y est la fonction de répartition des observations,
— F_X est la fonction de répartition des simulations,
— x est la valeur à corriger,
— $\tilde{x}$ est la valeur corrigée.

Les distributions F_X et F_Y peuvent s'obtenir soit par une approche empirique soit en ajustant un modèle paramétrique.

2.4.2 Méthode CDF-*transform*

Initialement développée pour la mise à l'échelle des données simulées par des modèles climatiques, la méthode CDF-*transform* est une extension à la méthode d'ajustement par quantiles qui prend en compte la non-stationnarité due aux changements climatiques [15, 16]. Contrairement à l'ajustement par quantiles, elle fournit en sortie une fonction de répartition, ce qui peut présenter un intérêt pour les études d'impact ultérieures.

Posons :

- F_{Y_p} la fonction de répartition des variables observées pour la période de projection, inconnue.
- F_{Y_c} la fonction de répartition des variables observées pour la période de calibration.
- F_{X_c} la fonction de répartition des variables simulées pour la période de calibration.
- F_{X_p} la fonction de répartition des variables simulées pour la période de projection, connue.

La méthode CDF-*transform* calcule alors F_{Y_p} à partir des trois autres fonctions de répartition :

$$F_{Y_p}(x) = F_{Y_c}\left[F_{X_c}^{-1}\left\{F_{X_p}(x)\right\}\right]$$

Cette équation peut être vue comme une double opération d'ajustement par quantile. La première ajuste les quantiles simulés entre la période de calibration et de projection. La deuxième transpose le quantile simulé au quantile observé.

Comme pour l'ajustement par quantile, les fonctions de répartitions F_{Y_c}, F_{X_c} et F_{X_p} peuvent s'obtenir de manière empirique ou en ajustant un modèle paramétrique. D'autre part, comme la plupart des méthodes de post-traitement, CDF-*transform* suppose que la relation de correspondance entre la variable simulée et la variable observée est stationnaire, c'est-à-dire qu'elle ne dépend pas du temps. Pour bien comprendre son mécanisme, la figure 2.3 présente une illustration de cette méthode avec des distributions GEV. Nous voyons que le biais entre la distribution des variables simulées (courbe orange) et observées (courbe rouge) en période de calibration est le même qu'en période de projection. Ce biais, calculé à l'aide des distributions en période de calibration, est corrigé de la distribution de la variable simulée en période future (courbe verte), pour obtenir la distribution des observations en période future (courbe bleue).

Comparé à l'ajustement par quantile, CDF-*transform* a l'avantage de prendre en compte le signal des changements climatiques en utilisant deux distributions pour les variables simulées. Lorsque ces distributions sont ajustées à partir d'un même modèle paramétrique non-stationnaire, il est théoriquement possible d'extrapoler F_{X_p} à n'importe quelle période future, puis d'obtenir la distribution F_{Y_p} correspondante. Par contraste, l'ajustement par quantile ne peut pas corriger le biais en dehors de la période où les observations sont disponibles.

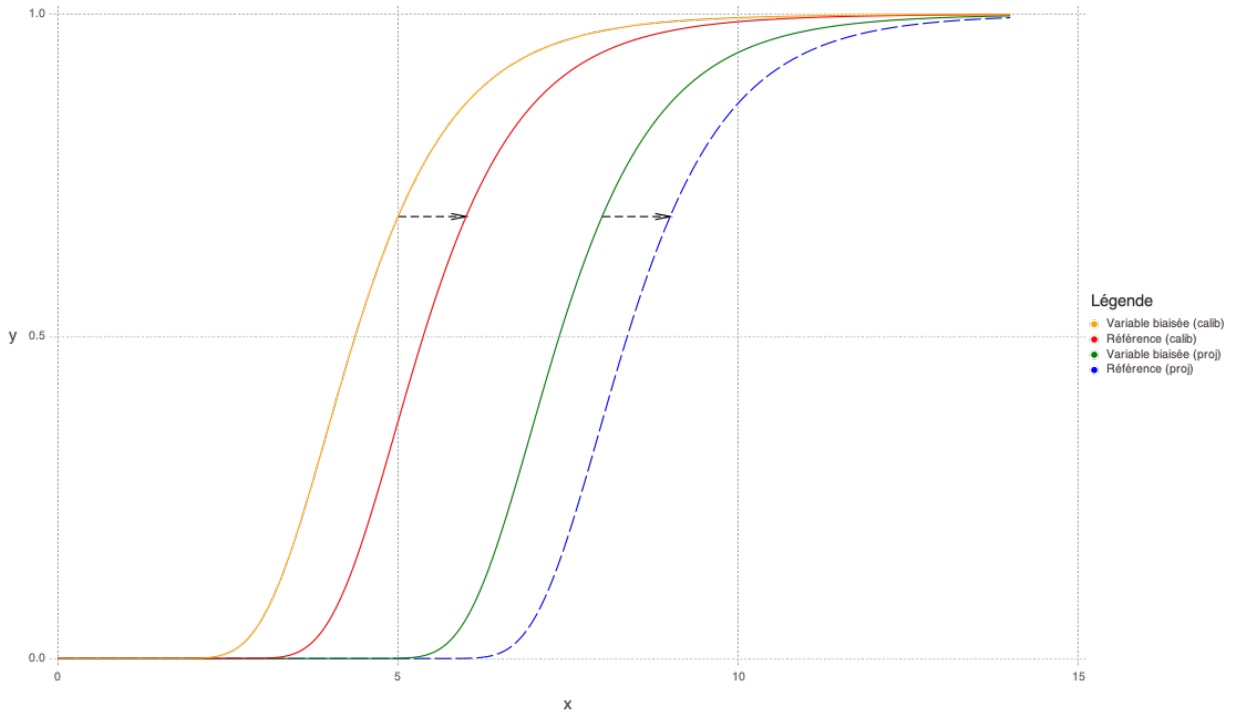


FIGURE 2.3 Transfert du biais entre le quantile simulé et observé, de la période de calibration à la période future, par la méthode CDF-*transform*. Les distributions illustrées sont GEV, de mêmes paramètres d'échelle et de forme.

CHAPITRE 3 REVUE DE LITTÉRATURE

Dans ce chapitre, une revue de littérature des méthodes d'analyse fréquentielle simultanée de plusieurs débits est présentée. Ces méthodes s'appuient sur la théorie des valeurs extrêmes. Les débits étudiés sont généralement des observations de plusieurs stations de mesure dans une même région ou des simulations provenant de plusieurs modèles climatiques. L'approche multivariée permet d'extraire un maximum d'information en exploitant les structures entre les données. Une petite section sera consacrée à la revue des méthodes de post-traitement statistique.

3.1 Analyse fréquentielle de plusieurs débits observés

Dans cette section, les méthodes d'analyse fréquentielle simultanée de plusieurs débits observés sont décrites. En modélisant la dépendance entre les extrêmes observés, elles parviennent à combiner les sources d'information.

La méthode d'analyse fréquentielle régionale (RFA) développée par Hosking et Wallis [17], consiste à partitionner la région d'étude en sous-régions homogènes et indépendantes, au sein desquelles les données sont analysées conjointement. Jusqu'à récemment, cette méthode est la plus utilisée en analyse fréquentielle hydrologique en raison de sa simplicité. Les extrêmes sont d'abord mis à l'échelle par une constante propre à chaque site, appelée en hydrologie *index flood*. Le modèle suppose alors que les extrêmes standardisés d'une même sous-région ont la même distribution. À partir de cette distribution régionale commune, on retrouve les caractéristiques de chaque site en effectuant une mise à l'échelle inverse. Lorsque les hypothèses du modèle sont raisonnables, la méthode RFA présente des estimations de quantiles plus justes qu'une analyse fréquentielle indépendante à chaque station. Néanmoins, elle comporte plusieurs limites [8, 18]. La délimitation des sous-régions est généralement faite de manière arbitraire. Même lorsqu'elle résulte de techniques statistiques de classification, l'existence de frontières entre les sous-régions provoque une discontinuité spatiale des estimations (même si la méthode dite de *régions d'influence* issue de l'approche RFA permet d'y remédier). De plus, la méthode RFA ne permet pas de calculer l'incertitude d'estimation des quantités d'intérêt. À titre d'exemple, cette approche régionale est utilisée dans [19] pour étudier les extrêmes de précipitation et de débits provenant d'une chaîne de modélisation hydroclimatique dans le bassin Méditerranéen. Elle est adaptée dans ce contexte car l'estimation régionale ne constitue pas le cœur de l'étude. Dans le présent projet, la méthode RFA ne sera pas utilisée, au profit des modèles hiérarchiques bayésiens.

Depuis la fin des années 1990, l'inférence bayésienne est de plus en plus utilisée en théorie des valeurs extrêmes. La revue de littérature de S. Coles et E. Powell [20] donne un aperçu des premières applications bayésiennes en statistiques des extrêmes. L'inférence bayésienne offre un cadre statistique cohérent et complet, avec plusieurs avantages : la prise en compte directe de l'incertitude d'estimation, la possibilité d'incorporer naturellement l'information *a priori* supplémentaire [8,13]. Les modèles hiérarchiques constituent une classe de modèle utile pour l'analyse d'extrêmes multivariés en utilisant des structures de dépendance simples et adaptables. Dans le cadre d'une étude lichénométrique, Cooley *et al.* [21] souligne la capacité d'un modèle hiérarchique à mettre en commun les informations provenant de différentes sources selon des structures flexibles. Cette mutualisation de l'information permettrait d'améliorer la précision d'estimation des paramètres, notamment le paramètre de forme de la loi GEV, usuellement difficile à estimer. La popularité croissante de ces modèles est rendue possible grâce aux développements récents en théorie bayésienne computationnelle (méthodes de simulation MCMC) et à la puissance de calcul grandissante des ordinateurs.

Pour les débits observés, la recherche bibliographique effectuée montre que les modèles hiérarchiques bayésiens sont la plupart du temps utilisés pour modéliser spatialement les extrêmes. Lima *et al.* [18] ont développé un modèle hiérarchique spatial où un processus latent gaussien est utilisé pour structurer les paramètres GEV décrivant les maxima de débits à différentes localisations du bassin Méditerranéen. L'auteur montre que l'incertitude est considérablement réduite par rapport à la méthode du maximum de la vraisemblance, grâce au partage d'information entre les différents sites. Ceci est d'autant plus marquant pour les sites avec peu d'observation et pour les niveaux de retour associés à des grandes périodes de retour. Un modèle hiérarchique bayésien similaire est utilisé dans [22] pour l'analyse fréquentielle spatiale des extrêmes de débits annuels dans le bassin du lac Dongting en Chine. Les auteurs montrent que les niveaux de retour inférés sont fournis avec des intervalles de crédibilité plus fiables que pour la méthode du maximum de la vraisemblance. L'approche bayésienne permet aussi de se dispenser des hypothèses plutôt restrictives de la méthode RFA [22]. Thordis *et al.* [23] utilise un modèle hiérarchique bayésien pour analyser des débits extrêmes, où la couche latente relie les données de différentes stations et incorpore des covariables géographiques (surface de drainage, présence de lac, longitude et latitude, etc). Ces études soulignent toutes la capacité du modèle hiérarchique à combiner l'information de manière cohérente et à produire des intervalles de crédibilité fiables pour les quantités d'intérêt.

Des méthodes d'analyse des extrêmes multivariés plus complètes existent, dans la mesure où elles cherchent à décrire la loi conjointe de toutes les variables en jeu : par exemple, les modèles utilisant des copules de valeurs extrêmes et les processus max-stables, dont une description peut se trouver dans [24]. Cependant, le grand nombre de paramètres dans ces modèles

les rend difficiles à implémenter en pratique pour des données de grande dimension. En particulier pour un processus max-stable, la densité conjointe des données observées ne s'écrit pas généralement sous une forme analytique en raison du grand nombre de termes. Même quand c'est le cas, l'estimation de cette densité serait computationnellement très intensive [24]. Enfin, une prise en compte explicite de la structure de dépendance des extrêmes est moins utile dans un contexte où la quantité d'intérêt ne dépend que des lois marginales, comme c'est le cas de ce projet.

Enfin, il est à noter que la littérature est inexistante pour ce qui est de l'analyse fréquentielle des pseudo-observations de débits. La nouveauté de la méthodologie pour générer ces pseudo-observations [5] et la spécificité de ces données peuvent expliquer cette absence de référence dans la littérature scientifique.

3.2 Utilisation des modèles climatiques pour l'analyse fréquentielle de débits futurs

Une abondance d'études démontre que les extrêmes de variables hydrologiques et de débits en particulier présentent des tendances non-stationnaires liées aux changements climatiques [2, 8, 25, 26]. Généralement, ces changements impactent la moyenne et la dispersion des débits extrêmes. Ils n'exhibent pas nécessairement une tendance à la hausse, étant liés à des phénomènes complexes d'évolution incertaine comme les précipitations intenses et la couverture neigeuse [8]. De plus, la grande variabilité naturelle des extrêmes peut rendre difficile l'identification d'une tendance. Étant donnée que l'évolution de la moyenne (cœur de la distribution) peut être de nature très différente de celle des extrêmes, il est recommandé de modéliser la non-stationnarité au sein même de la théorie des valeurs extrêmes [25, 26]. Une façon usuelle de procéder est de faire dépendre les paramètres des modèles de valeurs extrêmes de covariables qui dépendent du temps [8, 25, 27]. Une vaste partie de la littérature traitant de la non-stationnarité des extrêmes environnementaux utilise le temps comme variable explicative, voir par exemple [25, 27, 28]. Pourtant, lorsque les données de covariables ayant un lien direct avec les changements climatiques sont disponibles, il est préférable de les utiliser à la place de la variable temps. En effet, il est hasardeux de supposer que les tendances observées dans le passé sont les mêmes que dans un climat futur [29, 30]. Dans un article de Sraj *et al.* [30], les débits extrêmes de deux bassins de la Slovénie ont été analysés avec des distributions GEV non-stationnaires, en utilisant le temps et les précipitations annuelles comme covariables. Les auteurs trouvent qu'un modèle non-stationnaire s'ajuste mieux aux données, et que le modèle utilisant les précipitations annuelles exhibe le meilleur ajustement. Une comparaison intéressante entre l'inférence bayésienne MCMC et l'estimation du maxi-

mum de la vraisemblance montre aussi que la méthode bayésienne produit des intervalles de crédibilité plus précis. Dans un article de Das *et al.* [31], les anomalies de température globales et locales sont utilisées comme covariables dans un modèle GEV non-stationnaire, appliqué aux maxima annuels de débits simulés dans un bassin versant indien. Dans le présent projet, la variable explicative choisie est la concentration de GES dans l'atmosphère, en lien direct avec les changements climatiques. Dans la littérature, ce choix a rarement été fait pour l'étude des phénomènes hydrologiques extrêmes.

Le recours aux modèles climatiques (couplés avec des modèles hydrologiques) est nécessaire pour étudier les extrêmes de débits futurs en contexte de changements climatiques. En effet, contrairement aux données observées qui sont limitées par le temps, les débits simulés sont disponibles jusqu'à l'horizon 2100 et plus. Il est maintenant établi dans la communauté scientifique qu'aucun modèle climatique ne peut être considéré comme le meilleur et qu'il est important d'exploiter simultanément plusieurs modèles pour l'étude des extrêmes [32, 33]. Le développement de plusieurs modèles climatiques à travers le monde, découlant sur une multitude de scénarios futurs, rend possible l'étude des débits à partir d'ensembles de simulations. Une telle étude parviendrait à prendre en compte la grande incertitude du système climatique à plusieurs niveaux.

La revue de littérature de Katz [34] explique en détail les différentes techniques d'analyse d'incertitude dans l'étude des changements climatiques. L'analyse de sensibilité et l'analyse de scénarios sont des méthodes simples encore massivement utilisées aujourd'hui en sciences du climat. L'analyse de scénarios calcule les résultats d'un modèle en l'appliquant à partir de différents scénarios de départ, jugés comme des alternatives probables. L'ensemble des résultats rend compte de la variabilité due à l'incertitude sur les scénarios. Une telle approche est parfois accompagnée d'une analyse de la variance (ANOVA), qui permet de comparer l'influence de différents facteurs sur la variabilité de la quantité à l'étude [35–37]. Décrite dans [35], l'ANOVA consiste à décomposer la variance totale de la variable d'intérêt en une somme additive de variations provenant de différentes sources. À titre d'exemple, Zhang *et al.* [37] effectue des prédictions de débits extrêmes dans trois bassins fluviaux en Chine, en utilisant comme scénarios des combinaisons de trois scénarios d'émission GES, quatre GCM, quatre modèles hydrologiques et quatre méthodes de mise à l'échelle. La méthode ANOVA est ensuite utilisée pour quantifier la contribution de chaque source à l'incertitude totale sur les débits.

Katz [34] préconise tout de même que seule une analyse probabiliste complète des simulations multi-ensembles, s'appuyant sur des techniques statistiques modernes, peut rendre compte de toute l'incertitude du système climatique. L'auteur souligne l'intérêt du paradigme bayé-

sien et notamment des modèles hiérarchiques bayésiens dans l'atteinte de ce but. La revue de littérature de Tebaldi et Knutti [4] présente des méthodes statistiques pour combiner les sorties de modèles climatiques. Ces méthodes peuvent s'appliquer de même aux sorties de modèles hydroclimatiques, pour l'analyse des débits extrêmes. D'abord, le moyennage d'ensemble consiste à moyenner les résultats en associant un poids à chaque modèle. Les modèles peuvent être équiprobables ou le poids d'un modèle peut refléter d'une certaine manière sa performance. Plusieurs méthodes existent pour calculer les poids de modèle, comme le moyennage de modèles bayésiens ou la méthode de *Reliability Ensemble Averaging* (REA) développée par Giorgi et Mearns [38], fondée sur les critères dits de *biais* et de *convergence* de modèle. Une critique parfois faite au moyennage pondéré d'ensembles est que le calcul du poids s'appuie souvent sur des métriques de performance quelque peu arbitraires [4, 39].

Une approche alternative d'analyse multi-ensemble est le modèle hiérarchique bayésien. Tebaldi *et al.* [40] développe un modèle hiérarchique bayésien qui combine les sorties de différents modèles climatiques présents et futurs, ainsi que les données observées, sous des hypothèses gaussiennes. Plus précisément, les projections des modèles climatiques sont considérées comme des réalisations gaussiennes autour du vrai signal climatique, à inférer de manière bayésienne. S'inspirant des travaux de Tebaldi, Le Vine [39] développe un modèle hiérarchique pour l'analyse fréquentielle intégrée des débits extrêmes observés et simulés par 22 membres hydroclimatiques provenant d'un même RCM. La couche latente du modèle relie les différentes données sous des hypothèses gaussiennes, et ajoute un biais aléatoire qui relie les données simulées et observées. Le modèle hiérarchique bayésien dans Giuntoli *et al.* [41] combine les débits extrêmes simulés par la combinaison de neuf modèles hydrologiques et de cinq modèles globaux du climat, dans la région de l'est des États-Unis. Un modèle à effet aléatoire est utilisé dans la couche latente pour structurer les maxima de débits annuels simulés par les différents modèles hydroclimatiques. L'inférence bayésienne est ensuite utilisée pour obtenir la distribution *a posteriori* du changement de magnitude des débits extrêmes entre deux périodes de référence. Dans le cadre de notre projet, une approche bayésienne hiérarchique sera utilisée pour combiner l'information des membres d'un même modèle climatique. L'intégration des résultats sur tous les modèles climatiques s'effectuera par moyennage d'ensemble.

3.3 Post-traitement statistique des simulations climatiques

En climatologie, le biais d'un modèle climatique peut être défini comme la différence systématique entre une statistique simulée et la statistique correspondante réellement observée [14]. Les techniques de correction de biais dites de post-traitement statistique (*Model Output Statistics*, MOS) visent à établir une relation statistique entre la distribution d'une variable

observée et celle de la même variable simulée par un modèle climatique. La méthode d'ajustement par quantile, développée par Panofsky et Brier [42], consiste à faire correspondre la fonction de répartition de la variable simulée à celle de la variable observée pour une période de calibration donnée. L'ajustement est généralement effectué entre les quantiles empiriques des deux distributions, ou entre les quantiles de lois paramétriques ajustées aux données. Cependant, une estimation empirique est en général inadaptée pour les quantiles élevés, dû à la rareté des observations extrêmes. De plus, les lois paramétriques usuelles comme la loi gamma sont adaptées pour décrire le cœur des distributions mais ajustent mal la queue des distributions. Pour ces raisons, l'ajustement par quantile est peu performant pour corriger le biais dans les queues de distribution [3], ce qui nécessite des techniques de correction de biais conçues spécifiquement pour les extrêmes.

Kallache *et al.* [16] proposent la méthode XCDF-*transform*, une extension de la méthode d'ajustement par quantile où les fonctions de répartition sont issues de la théorie des valeurs extrêmes. En se basant sur la méthode CDF-*transform* de Michelangeli *et al.* [15], les auteurs établissent le lien entre la fonction de répartition de modèles aux valeurs extrêmes ajustée sur les observations et celle ajustée sur les simulations, à l'aide d'une fonction de transfert. Cette fonction de transfert est estimée sur une période de calibration et utilisée pour corriger le biais en période de projection future, tenant ainsi compte des changements climatiques. La méthode XCDF-*transform* a été utilisée à plusieurs reprises avec succès. Laflamme *et al.* [43] l'utilise pour le post-traitement des précipitations extrêmes simulées par un RCM et le calcul des précipitations extrêmes projetées en Nouvelle-Angleterre. Comme dans Kallache *et al.* [16], des distributions Pareto généralisée sont ajustées aux séries de précipitations simulées et observées avant d'appliquer la correction de biais. Dans un contexte d'étude océanographique, Towe *et al.* [44] propose des extensions à la méthode XCDF-*transform*, pour tenir compte simultanément du cœur et de la queue de distribution. Leur modification permet aussi d'appliquer la fonction de transfert à des variables qui sont sur des échelles de mesure différentes. Leur approche permet alors de mettre en relation les données de vent simulées par un GCM et les données de hauteurs de vague observées localement, pour prédire les hauteurs de vagues extrêmes futures en mer du Nord. Dans le présent projet, une variante de la méthode XCDF-*transform* avec des fonctions de répartition GEV est utilisée pour faire la jonction entre le modèle statistique pour les simulations et le modèle statistique pour les pseudo-observations.

CHAPITRE 4 DONNÉES

Les données décrites ci-dessous sont disponibles pour 234 tronçons de la rivière Chaudière au Québec. Le bassin de la rivière Chaudière s'étend sur 6713 km^2, de la frontière américaine au sud jusqu'à la ville de Québec au nord. Il totalise 78 municipalités et 179 000 habitants. La carte de la zone d'étude est présentée à la figure 4.1. Sur les 234 tronçons, 211 ont une surface de drainage supérieure à 50 km^2. Les modèles développés sont testés et validés sur ces 211 tronçons. Les 23 tronçons restants présentent des débits trop faibles pour poser un risque d'inondations. La méthodologie développée sera éventuellement appliquée à l'ensemble des bassins versants du Québec méridional, comprenant plus de 18 000 tronçons.

Pour chaque tronçon de rivière, la Direction de l'Expertise Hydrique (DEH) met à disposition deux bases de données : les données de pseudo-observations et les données de simulations hydroclimatiques.

4.1 Pseudo-observations

Les débits pseudo-observés sont des données journalières s'étendant sur la période 1961 à 2020. Ils ont été obtenus avec une méthode d'interpolation statistique utilisant des modèles hydrologiques, décrite dans [5]. C'est une méthode d'assimilation des données qui permet d'estimer les débits dans les tronçons non jaugés en combinant les observations dans les tronçons jaugés environnants et les débits simulés par modèle hydrologique.

Le modèle hydrologique utilisé pour l'interpolation des pseudo-observations est Hydrotel. C'est un modèle semi-distribué utilisant des données géoréférencées telles que l'altitude, l'utilisation des terres, le type de sol et les réseaux de rivières et lacs. Il a déjà été utilisé avec succès dans d'autres études climatiques au Québec méridional [5]. Six configurations différentes du modèle Hydrotel sont utilisées, ce qui génère six versions des pseudo-observations. Ces configurations sont identifiées par la DEH sous les acronymes : LN24HA, MG24HA, MG24HI, MG24HK, MG24HQ et MG24HS.

La méthode d'interpolation des débits précédemment décrite introduit de l'incertitude de modélisation, représentée par une distribution d'erreur log-normale autour de la valeur la plus probable de chaque débit. Les paramètres de ces distributions log-normales ne sont pas fournis, mais nous avons accès aux 19 quantiles d'ordres $0.05, 0.1, \ldots, 0.95$: $\{Q_{0.05\times k} : k = 1, \ldots, 19\}$. Il s'agirait de la meilleure estimation disponible des séries de débits historiques. Il est à noter que ces données sont en cours d'amélioration par une autre équipe de recherche

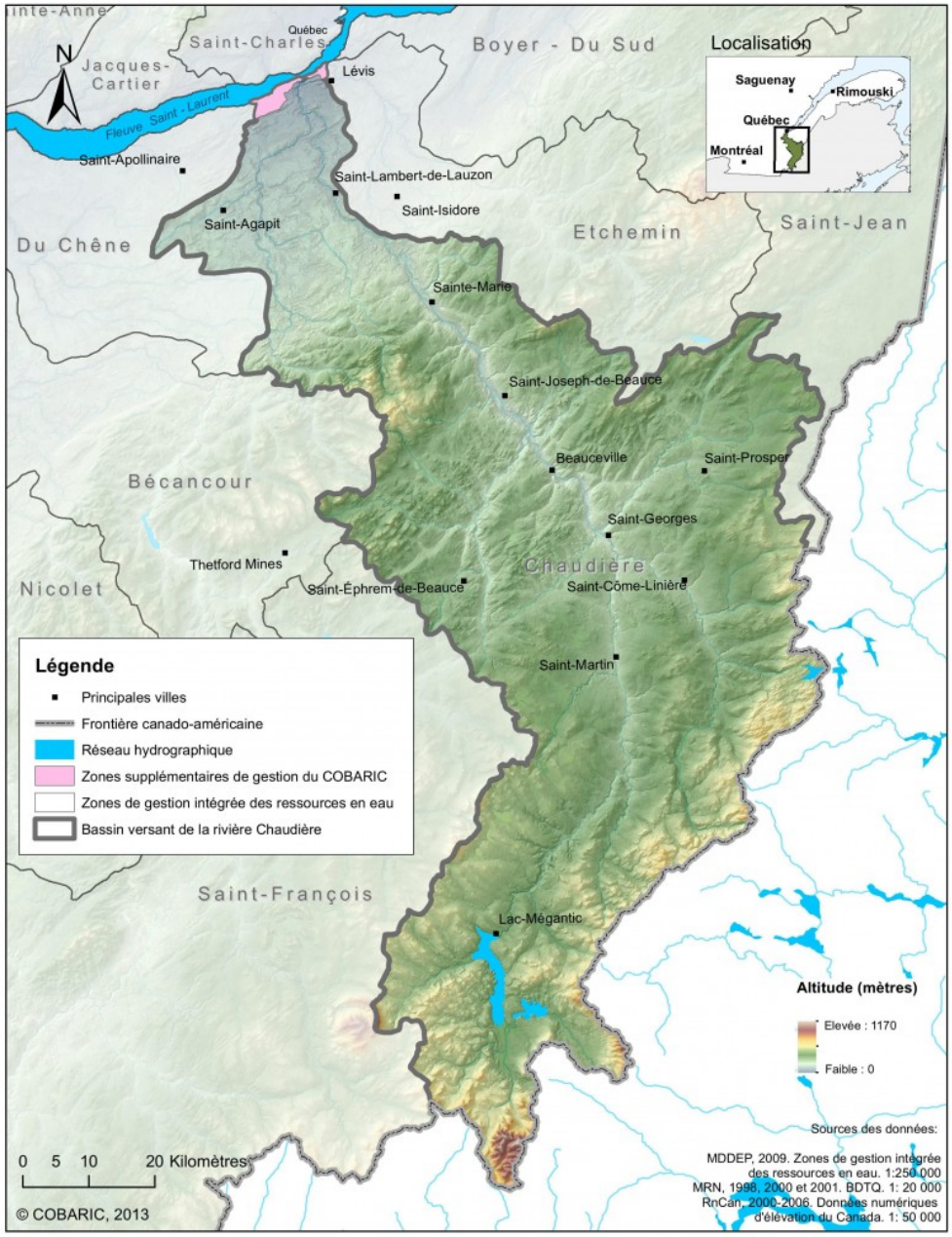


FIGURE 4.1 Bassin de la rivière Chaudière. Source : site du Comité de bassin de la rivière Chaudière.

du projet INFO-Crue.

À partir de ces quantiles, nous déduisons les paramètres des distributions d'erreur log-normales associées à chaque valeur de débit journalier. Les deux quantiles d'ordre 0.25 et 0.75 ont été utilisés pour le calcul des paramètres des lois log-normales. On obtient alors la distribution marginale de chaque pseudo-observation du jour j simulée avec la configuration i. La base de données complète est alors l'ensemble des couples de paramètres log-normales :

$$\{(\eta_{ij}, \zeta_{ij}) : i = 1, \ldots, S; j = 1, \ldots, D\}$$

où S est le nombre de configurations hydrologiques et D est le nombre de jours de données disponibles.

4.2 Débits simulés par un ensemble de simulations

4.2.1 Ensemble de simulations climatiques

Les données de 180 simulations climatiques sont disponibles pour chaque tronçon. Chaque simulation climatique est identifiée par la combinaison unique d'un couple GCM-RCM, d'un scénario d'émission de GES et d'un membre. Pour faciliter l'écriture, nous considérons dans le reste du mémoire qu'un GCM est un couple GCM-RCM, dont le RCM est vide. Les membres d'un même couple GCM-RCM se distinguent par de légères perturbations dans les conditions initiales du modèle qui les pilote. Ces perturbations forcent les simulations à se différencier et représentent la variabilité naturelle du climat. Deux scénarios d'émission sont considérés dans l'ensemble de simulations : RCP4.5 et RCP8.5. Certains couples GCM-RCM ont été forcés avec ces deux scénarios d'émission, mais la plupart ont été forcés avec le scénario RCP8.5 seulement. La totalité des données provient de trois ensembles de simulations : CORDEX, CMIP5 et CLIMEX.

CORDEX [45] est un projet du Programme mondial de recherches sur le climat qui cherche à évaluer la performance des modèles climatiques régionaux. Les 31 simulations appartenant à l'ensemble CORDEX s'appuient sur des modèles régionaux du climat. Les simulations de la cinquième version du Modèle Régional Canadien du Climat (CRCM5), forcées par le GCM CanESM2, comportent 5 membres. Pour toutes les autres simulations, un seul membre est disponible.

CMIP5, la cinquième version du Projet d'intercomparaison des modèles couplés [46], est un projet du Programme mondial de recherches sur le climat. D'ampleur mondiale, il a pour but de rassembler les centres climatiques du monde entier pour réaliser des simulations

climatiques de façon coordonnée entre les différents groupes de recherche. Les 99 simulations provenant de l'ensemble CMIP5 sont toutes basées sur des modèles globaux du climat, de résolution plus faible. Le nombre de membres varie de 1 à 9, en fonction du modèle.

Enfin, CLIMEX [47] est un projet visant à étudier les extrêmes climatiques et ses conséquences pour la gestion hydrique en Bavière et au Québec. Les 50 simulations de l'ensemble CLIMEX correspondent à 50 membres du couple GCM-RCM CanESM2/CRCM5-Ouranos. Elles supposent toutes le scénario d'émission RCP8.5.

4.2.2 Modèle hydrologique et débits simulés

Les sorties de simulations climatiques (données journalières de précipitation et température) ont été post-traitées puis utilisées comme intrants au modèle hydrologique Hydrotel pour simuler les débits journaliers de 234 tronçons de la rivière Chaudière, sur la période 1955 à 2100. Ce travail a été effectué par la DEH. Comme pour les pseudo-observations, les mêmes configurations d'Hydrotel sont utilisées, identifiées sous les acronymes LN24HA, MG24HA, MG24HI, MG24HK, MG24HQ et MG24HS. Chaque simulation climatique produit alors six séries de débits simulés, correspondant aux six configurations du modèle hydrologique.

À titre d'illustration, la figure 4.2 présente deux séries de maxima annuels extraites des débits simulés pour le tronçon SLSO00003 de la rivière Chaudière. Les deux simulations hydroclimatiques représentées utilisent la configuration LN24HA d'Hydrotel, le membre 1 du couple GCM-RCM CanESM2/CRCM5-Ouranos (en bleu) ou le membre 1 du GCM IPSL-CM5A-LR (en rouge). On observe que les maxima annuels simulés par IPSL-CM5A-LR ont généralement des valeurs supérieures à ceux simulés par CanESM2/CRCM5-Ouranos.

FIGURE 4.2 Maxima annuels extraits de la série de débits simulés du tronçon SLSO00003. Les simulations utilisent la configuration LN24HA d'Hydrotel et le membre 1 de l'ensemble CLIMEX (en bleu) ou le membre 1 du GCM IPSL-CM5A-LR (en rouge).

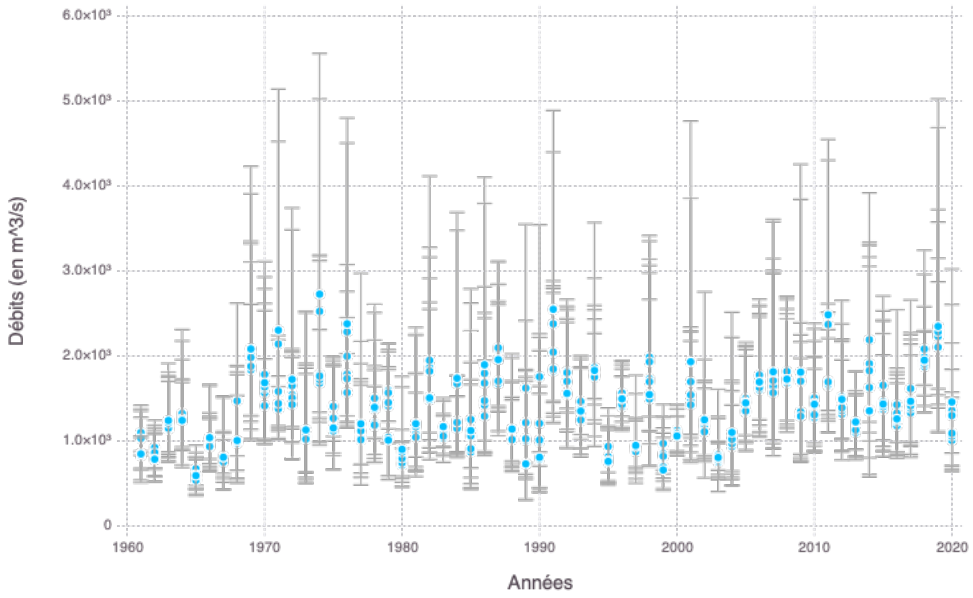


FIGURE 4.3 Distributions des maxima annuels de débits extraites des pseudo-observations du tronçon SLSO00003. Chaque barre correspond à la distribution log-normale d'un débit pseudo-observé. Les points et la hauteur des barres représentent respectivement la médiane, les quantiles d'ordre 0.25 et 0.75.

CHAPITRE 5 MODÈLE STATISTIQUE POUR LES PSEUDO-OBSERVATIONS

5.1 Extraction des maxima annuels

Pour les pseudo-observations, une série chronologique journalière de couples de paramètres log-normales est disponible pour chacune des six configurations du modèle hydrologique. Pour chaque configuration, nous considérons que les maxima annuels correspondent aux jours de l'année où la médiane de la loi log-normale est la plus grande. On note (η_{ij}, ζ_{ij}) le couple de paramètres de la loi log-normale correspondant au débit maximal de l'année j simulé par la configuration i. Pour l'année j, les couples $\{(\eta_{ij}, \zeta_{ij}) : i = 1, \ldots, S\}$ résument alors l'incertitude sur le vrai maximum annuel. La figure 4.3 illustre cette incertitude pour les maxima annuels de débits du tronçon SLSO00003, situé à l'exutoire de la rivière Chaudière, où elle rejoint le fleuve Saint-Laurent près de la ville de Québec.

Le critère retenu pour l'extraction des maxima est la médiane des distributions d'erreur log-normales, mais d'autres choix auraient été possibles. Par exemple, la moyenne des lois log-normales aurait pu être choisie pour extraire le couple de paramètres caractérisant le maximum annuel. Cependant, dans le cas de la loi log-normale, la moyenne dépend également du paramètre d'échelle. Une grande incertitude d'interpolation pourrait occasionner une grande moyenne, ce qui rend cette mesure moins robuste que la médiane.

Il aurait aussi été possible de simuler indépendamment chaque pseudo-observation journalière selon sa loi marginale log-normale, puis d'extraire les maxima annuels pour chacune des configurations. Cependant, avec cette méthode, la série temporelle obtenue ne refléterait pas la réalité puisqu'elle ne prend pas en compte la dépendance temporelle entre les débits consécutifs. De plus, l'incertitude associée à chaque pseudo-observation étant grande (voir la figure 4.3), les dates des maxima annuels varieraient beaucoup d'un tirage à l'autre.

5.2 Modèle statistique

Nous développons un modèle statistique pour les pseudo-observations qui tire profit de toute l'information disponible provenant des six configurations hydrologiques. Il permet d'obtenir une estimation de la série des maxima annuels la plus probable permettant l'analyse fréquentielle des extrêmes. Une approche séquentielle qui estimerait d'abord les maxima annuels non observés puis ensuite leur distribution ignorerait l'incertitude d'estimation des maxima. Un modèle bayésien est alors développé pour estimer conjointement la série des vrais maxima

annuels non observés ainsi que la loi de ces maxima dans un même modèle statistique cohérent de façon à prendre en compte l'incertitude sur les maxima. Il est à noter que le modèle n'est pas un modèle hiérarchique classique, car nous n'avons pas accès directement aux débits mais à des distributions d'erreur sur ces derniers.

5.2.1 Analyse fréquentielle si les maxima annuels étaient connus

Soit Y_j le vrai maximum annuel de l'année j qui n'est pas observé pour un tronçon non jaugé et soit $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ la série des n maxima annuels non observés. S'ils était réellement observés, une analyse fréquentielle classique pourrait être effectuée en supposant que les maxima sont distribués selon la loi GEV :

$$Y_j \sim \mathcal{G}EV(\mu, e^{\phi}, \xi) \quad \text{pour } j = 1, \ldots, n$$

où $\sigma = e^{\phi}$.

L'estimation des paramètres de la loi GEV pourrait se faire sous le paradigme bayésien en considérant la loi *a priori* partiellement informative suivante :

$$f_{(\mu,\phi,\xi)}(\mu, \phi, \xi) \propto \mathcal{B}eta(\xi + 0.5 \mid 9, 6)$$

où $\mathcal{B}eta(x \mid 9, 6)$ dénote la densité de la loi bêta de paramètres $(9, 6)$ évaluée en x. La loi est non-informative pour μ et ϕ mais informative pour le paramètre de forme ξ, utilisant la loi informative définie par Martins et Stedinger [48] pour le paramètre de forme de la loi GEV dans un contexte d'extrêmes environnementaux.

Si la série des maxima exhibait une non-stationnarité, il serait possible de la modéliser avec la variable explicative $\boldsymbol{u} = (u_1, \ldots, u_n)$ intégrée dans le modèle statistique de valeurs extrêmes de la façon suivante :

$$Y_j \sim \mathcal{G}EV\left\{\mu_0 + \mu_1 u_j, \exp(\phi_0 + \phi_1 u_j), \xi\right\} \text{ pour } j = 1, \ldots, n.$$

La loi *a priori* partiellement informative pour ce modèle serait la suivante :

$$f_{(\mu_0,\mu_1,\phi_0,\phi_1,\xi)}(\mu_0, \mu_1, \phi_0, \phi_1, \xi) \propto \mathcal{B}eta(\xi + 0.5 \mid 9, 6)$$

D'après la règle de Bayes, la loi *a posteriori* s'écrirait alors :

$$f_{(\mu_0,\mu_1,\phi_0,\phi_1,\xi|\boldsymbol{Y}=\boldsymbol{y})}(\mu_0,\mu_1,\phi_0,\phi_1,\xi) \propto \mathcal{B}eta(\xi{+}0.5 \mid 9,6) \prod_{j=1}^{n} \mathcal{G}EV(y_j|\mu_0{+}\mu_1 u_j, \exp(\phi_0{+}\phi_1 u_j), \xi) \tag{5.1}$$

où $\mathcal{G}EV(y|\mu,\sigma,\xi)$ dénote la densité de la loi GEV de paramètres (μ,σ,ξ) évaluée en y.

5.2.2 Loi des maxima de débit en fonction des pseudo-observations

Dans ce projet, comme les maxima annuels de débit ne sont pas connus, nous utilisons les lois log-normales fournies pour les inférer. Dans le but de prendre en compte l'incertitude sur le maximum annuel Y_j de l'année j, l'information apportée par les $S = 6$ différentes configurations hydrologiques est utilisée de la façon suivante :

$$f_{(Y_j|\eta_{ij},\zeta_{ij})}(y_j) \propto \prod_{i=1}^{S} \mathcal{L}N(y_j|\eta_{ij},\zeta_{ij}) \text{ pour } j = 1,\ldots,n \tag{5.2}$$

où

- $\mathcal{L}N(y|\eta,\zeta)$ dénote la densité de la loi log-normale de paramètres (η,ζ) évaluée en y,
- (η_{ij},ζ_{ij}) est le couple de paramètres de la loi log-normale pour le débit maximum annuel de l'année j avec la configuration i,
- S est le nombre de configurations,
- n est le nombre d'années d'observation.

La densité du maximum annuel Y_j de l'année j est proportionnelle au produit des S densités log-normales associées à chacune des configurations du modèle hydrologique. Chaque paramètre $\{\eta_{ij} : i = 1,\ldots,S\}$ apporte de l'information à l'estimation de Y_j, dont le poids est pondéré par ζ_{ij}, l'incertitude sur les pseudo-débits. Pour mieux visualiser ceci, on remarque que si le débit maximal $Y_j = y_j$ était connu, le modèle revient à supposer que les paramètres de localisation des lois log-normales sont centrés sur le logarithme du débit :

$$f_{(\eta_{ij}|Y_j=y_j,\zeta_{ij})}(\eta_{ij}) = \mathcal{N}(\eta_{ij}|\log y_j, \zeta_{ij}^2) \quad \text{pour } i = 1,\ldots,S.$$

En l'absence d'indication supplémentaire sur les configurations hydrologiques, elles sont supposées indépendantes. Cette procédure est répétée pour chaque année $j \in \{1,\ldots,n\}$ pour obtenir la densité conjointe des n maxima annuels $\boldsymbol{Y} = (Y_1,\ldots,Y_n)$.

5.2.3 Modèle statistique complet

Le modèle statistique pour les pseudo-observations combine les équations 5.1 et 5.2. Il s'écrit alors :

$$f_{(Y_j|\eta_{ij},\zeta_{ij})}(y_j) \propto \prod_{i=1}^{S} \mathcal{L}N(y_j|\eta_{ij},\zeta_{ij}), \quad j = 1,\ldots,n$$

$$f_{(\mu_0,\mu_1,\phi,\xi|\boldsymbol{Y}=\boldsymbol{y})}(\mu_0,\mu_1,\phi,\xi) \propto \prod_{j=1}^{n} \mathcal{G}EV(y_j|\mu_0+\mu_1 u_j, \exp(\phi), \xi) \times \mathcal{B}eta(\xi+0.5|6,9)$$

où u_j est la concentration de GES en CO_2 équivalent à l'année j.

Une loi GEV non-stationnaire est utilisée pour l'ajustement des maxima annuels. En effet, une analyse exploratoire sur les 211 tronçons de la rivière Chaudière porte à croire que les maxima de débits de plusieurs tronçons ne sont pas stationnaires sur la période historique. Sur les tronçons étudiés, 52 présentent au moins deux séries de maxima annuels (correspondant à deux configurations hydrologiques) avec une non-stationnarité significative par le test de Mann-Kendall. Le test de tendance de Mann-Kendall est un test statistique non paramétrique pour détecter la stricte monotonie d'une série temporelle. Ici, la non-stationnarité est modélisée pour le paramètre de localisation de la loi GEV seulement. Nous choisissons de ne pas ajouter une relation non-stationnaire pour le paramètre d'échelle. En effet, dû à la courte durée d'observation, l'utilisation d'un modèle avec plus de paramètres conduirait à une grande incertitude d'estimation. Le choix de la concentration de GES comme variable explicative est cohérent avec le modèle développé pour les simulations (voir le chapitre 6), même si nous sommes conscients que les changements climatiques ne constituent sûrement pas le seul facteur à l'origine des potentielles tendances observées. Les données du scénario RCP8.5 sont utilisées. C'est un choix arbitraire mais raisonnable, car la différence avec le scénario RCP4.5 est minime sur la période historique.

Pour la loi *a priori* des paramètres GEV non-stationnaire, le paramètre de forme est contraint par la loi de Martins et Stedinger à prendre des valeurs entre -0.5 et 0.5, avec un mode *a priori* attendu autour de -0.12, une valeur raisonnable pour les extrêmes de débit. Compte tenu de la grande incertitude sur les pseudo-observations obtenues par la méthode d'interpolation précédemment décrite, une telle contrainte est nécessaire pour éviter des estimations instables et irréalistes du paramètre de forme. À part le paramètre ξ, les autres paramètres GEV suivent une loi non informative et uniforme sur l'espace (μ_0, μ_1, ϕ). Ce choix de loi *a priori* pour les paramètres GEV garantit l'existence de la constante de normalisation pour la loi *a posteriori* lorsque $n \geq 2$ [49]. La densité $f_{(\mu_0,\mu_1,\phi,\xi|\boldsymbol{Y}=\boldsymbol{y})}$ est donc bien définie.

Il est à noter qu'une écriture du modèle sous forme hiérarchique aurait été possible, mais

nous préférons la description ci-dessus qui nous parait plus naturelle et compréhensible. Une représentation du modèle bayésien pour les pseudo-observations sous forme de graphe directif est présentée à la figure 5.1.

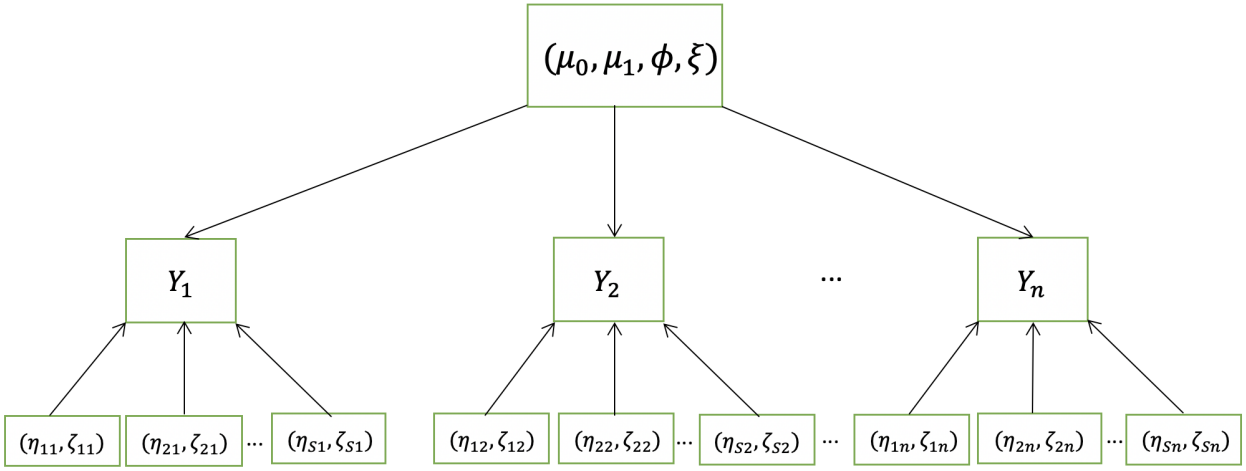


FIGURE 5.1 Représentation schématique du modèle bayésien pour les pseudo-observations.

5.3 Estimation des paramètres

Les paramètres du modèle à inférer sont : $(\boldsymbol{Y}, \mu_0, \mu_1, \phi, \xi)$ où $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ est la série des maxima annuels de débits non observée. La loi *a posteriori* complète s'écrit alors :

$$
\begin{aligned}
&f_{(\mu_0,\mu_1,\phi,\xi,\boldsymbol{Y}|\eta_{ij},\zeta_{ij})}(\mu_0, \mu_1, \phi, \xi, y_1, \ldots, y_n) \\
&\qquad = f_{(\mu_0,\mu_1,\phi,\xi|\boldsymbol{Y}=\boldsymbol{y})}(\mu_0, \mu_1, \phi, \xi) \times \prod_{j=1}^{n} f_{(Y_j|\eta_{ij},\zeta_{ij})}(y_j) \\
&\qquad \propto \mathcal{B}eta(\xi + 0.5|6, 9) \prod_{j=1}^{n} \mathcal{G}EV(y_j|\mu_0 + \mu_1 u_j, \exp(\phi), \xi) \times \prod_{j=1}^{n} \prod_{i=1}^{S} \mathcal{L}N(y_j|\eta_{ij}, \zeta_{ij})
\end{aligned}
$$

Les lois conditionnelles complètes sont :

$$
\begin{aligned}
&f_{(Y_j|-)}(y_j) \propto \prod_{i=1}^{S} \mathcal{L}N(y_j|\eta_{ij}, \zeta_{ij}) \times \mathcal{G}EV(y_j|\mu_0 + \mu_1 u_j, \exp(\phi), \xi), \quad j = 1, \ldots, n \\
&f_{(\mu_0|-)}(\mu_0) = f_{(\mu_1|-)}(\mu_1) = f_{(\phi|-)}(\phi) \propto \prod_{j=1}^{n} \mathcal{G}EV(y_j|\mu_0 + \mu_1 u_j, \exp(\phi), \xi) \\
&f_{(\xi|-)}(\xi) \propto \prod_{j=1}^{n} \mathcal{G}EV(y_j|\mu_0 + \mu_1 u_j, \exp(\phi), \xi) \times \mathcal{B}eta(\xi + 0.5|6, 9)
\end{aligned}
$$

Un algorithme Metropolis-dans-Gibbs est utilisé pour générer un échantillon de la loi *a posteriori*. Les lois conditionnelles complètes ne s'expriment pas sous la forme de lois connues. Chacune d'elle sera simulée par une étape Metropolis-Hastings avec exploration locale, utilisant une marche aléatoire avec un pas distribué selon la loi normale. Les paramètres ont été simulés dans l'algorithme MCMC dans l'ordre où ils ont été écrits.

La chaîne de Markov est initialisée avec des valeurs probables des paramètres. Les valeurs initiales pour $\boldsymbol{Y}$ sont $y_j = \frac{1}{S}\sum_{i=1}^{S}\exp(\eta_{ij})$, la moyenne empirique des médianes des lois log-normales. Ensuite, les paramètres GEV initiaux sont les estimateurs du maximum de la vraisemblance pour la série $(y_1, \ldots, y_n)$. Ces valeurs constituent une approximation raisonnable à utiliser pour les conditions initiales de l'algorithme MCMC. Les pas des marches aléatoires sont ajustés toutes les 50 itérations selon la méthode adaptive décrite dans la section 2.3.2, afin de se rapprocher du taux d'acceptation moyen optimal de 0.44.

La visualisation de la trace ainsi que l'utilisation de deux tests usuels de convergence, le test de Gelman-Rubin et celui de Geweke, indiquent une bonne convergence de l'algorithme après 5000 itérations vers l'état stationnaire. Pour s'assurer de la convergence sur tous les tronçons, la période de préchauffage est portée à 50 000 itérations. Les chaînes MCMC produites pour les quatre paramètres GEV non-stationnaire sont présentées à la figure 5.2.

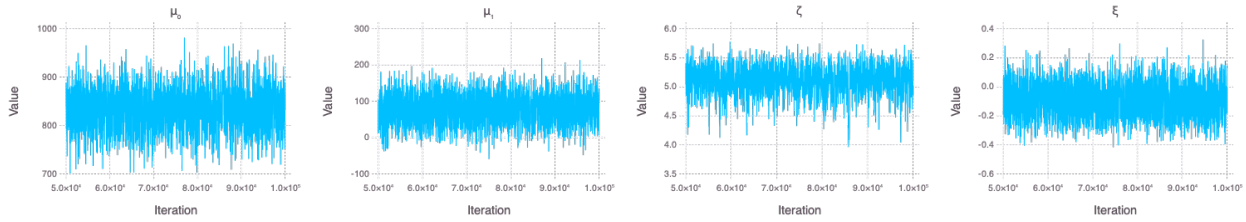


FIGURE 5.2 Chaînes MCMC correspondant à la loi *a posteriori* marginale des quatre paramètres de la loi GEV non-stationnaire $(\mu_0, \mu_1, \phi, \xi)$ du modèle pour les pseudo-observations, pour le tronçon SLSO00003.

5.4 Sélection de modèle

On souhaite vérifier si les maxima annuels de débits sont stationnaires sur la période historique. Pour ce faire, deux modèles ont été testés et comparés : le modèle GEV stationnaire $(\mu_1 = 0)$ et le modèle GEV non-stationnaire à 4 paramètres décrit ci-dessus. Généralement, un modèle plus sophistiqué s'ajuste mieux aux données, mais le prix à payer est la plus grande incertitude d'estimation lorsque le nombre de paramètres dans le modèle augmente [30]. Ceci est d'autant plus vrai que la période des pseudo-observations est assez courte pour l'étude

des valeurs extrêmes.

La sélection de modèle se fait avec le DIC (Deviance Information Criterion). Le DIC, introduit par Spiegelhalter *et al.* [50], peut être considéré comme la version bayésienne du AIC pour la méthode du maximum de la vraisemblance, particulièrement adapté lorsqu'un échantillon de la loi *a posteriori* est obtenu par simulation MCMC. Il s'écrit :

$$DIC = D(\bar{\boldsymbol{\theta}}) + 2p_D$$

où $D(\boldsymbol{\theta}) = -2\log L(\boldsymbol{\theta}|\boldsymbol{Y})$ est un terme proportionnel à la log-vraisemblance du modèle et $p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}})$ est un terme de pénalisation équivalent à celui dans l'AIC. La barre horizontale dénote l'estimateur de la moyenne empirique, qui se calcule en moyennnant sur les itérations MCMC en sortie du modèle. Plus précisément, si l'on note $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$ les valeurs des paramètres des M itérations MCMC, on a :

$$\bar{\boldsymbol{\theta}} = \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{\theta}_i$$

et

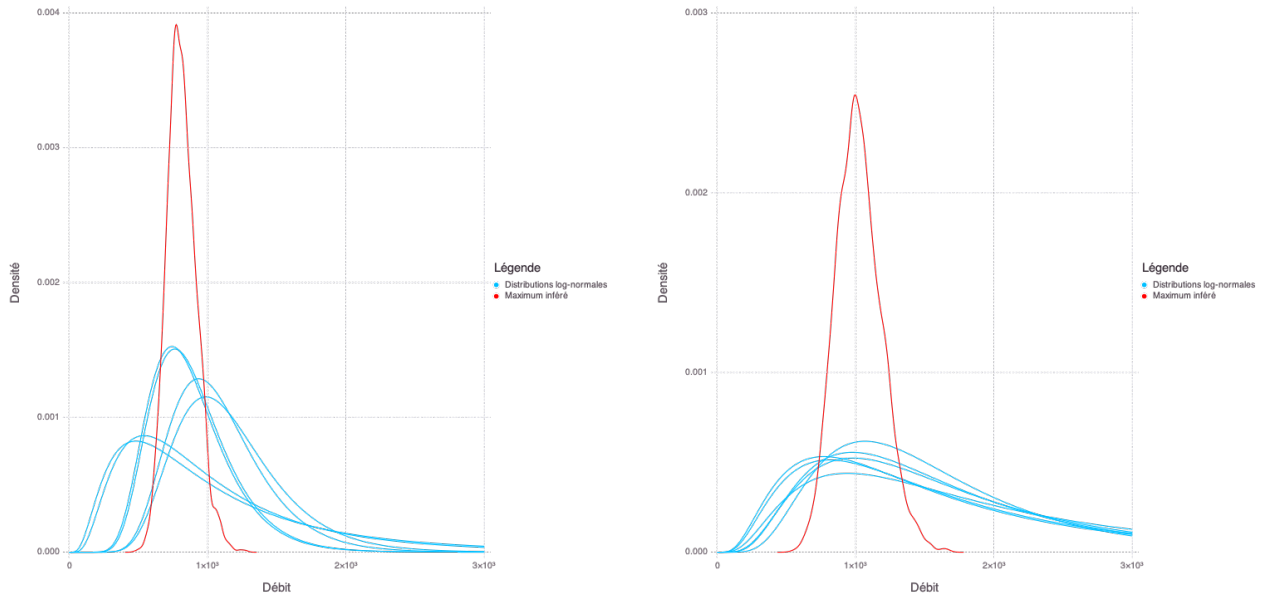
$$\overline{D(\boldsymbol{\theta})} = -\frac{1}{M}\sum_{i=1}^{M} 2\log L(\boldsymbol{\theta}_i|\boldsymbol{Y})$$

Un modèle est considéré meilleur qu'un autre lorsque la valeur du DIC correspondant est plus faible. Pour le tronçon SLSO00003, les valeurs du DIC sont 3142 pour le modèle stationnaire et 3137 pour le modèle non-stationnaire. Le modèle non-stationnaire est alors plus adapté et sera illustré dans la suite.

5.5 Validation du modèle et résultats

La figure 5.3 présente la densité *a posteriori* de Y_1 et Y_{57}, les maxima de débits des années 1961 et 2017, ainsi que les distributions d'erreur log-normales correspondantes. On observe que la densité inférée du maximum annuel non observé combine bien l'information provenant des différentes configurations hydrologiques, et permet de réduire considérablement l'incertitude de départ. La loi du maximum annuel inféré est alors concentrée autour des modes des distributions log-normales, avec une dispersion beaucoup plus réduite. Ce comportement est observé pour toute la série de maxima annuels de débits.

Pour vérifier le bon ajustement des paramètres GEV par rapport à la série des maxima inférés, nous utilisons un diagramme quantile-quantile (QQ-plot) standardisé, voir le chapitre 6 de [6]. Notons $(\hat{\mu}_0 + \hat{\mu}_1 u_j, \exp(\hat{\phi}), \hat{\xi})$ le triplet de paramètres GEV inféré pour l'année j. Alors si le



(a) 1961 (b) 2017

FIGURE 5.3 Densité inférée pour les maxima annuels de (a) 1961 et (b) 2017, ainsi que les distributions d'erreur d'interpolation log-normales correspondantes, pour le tronçon SLSO00003.

modèle s'ajuste bien aux maxima inférés $(\hat{y}_j)$, les variables standardisées

$$\tilde{y}_j = \frac{1}{\hat{\xi}} \log \left\{ 1 + \hat{\xi} \frac{\hat{y}_j - (\hat{\mu}_0 + \hat{\mu}_1 u_j)}{\exp(\hat{\phi})} \right\}, \quad j = 1, \ldots, n$$

devraient suivre la distribution Gumbel unitaire de fonction de répartition

$$F(y) = \exp\{-e^{-y}\}$$

Le QQ-plot est alors constitué des paires de points :

$$\left\{ \left(-\log\left(-\log \frac{j}{n+1} \right), \tilde{y}_{(j)} \right) : j = 1, \ldots, n \right\} \tag{5.3}$$

où $\tilde{y}_{(k)}$ est la statistique d'ordre k de l'ensemble $\{\tilde{y}_j : j = 1, \ldots, n\}$. Un modèle bien ajusté aux données aura un nuage de points proche de la diagonale $y = x$.

Pour le calcul de $\tilde{y}_j$, nous calculons $\{\tilde{y}_{jk} : k = 1, \ldots, K\}$ avec les valeurs inférées des para-

mètres à chaque itération MCMC k. La moyenne

$$\tilde{y}_j = \frac{1}{K} \sum_{k=1}^{K} \tilde{y}_{jk}$$

est utilisée pour tracer le QQ-plot. L'avantage de cette méthode est la possibilité d'inclure l'intervalle de crédibilité bayésien à 95%, en utilisant les quantiles empiriques d'ordre 0.025 et 0.975 de $\{\tilde{y}_{jk} : k = 1, \ldots, K\}$. Le QQ-plot pour le tronçon SLSO00003 est présenté à la figure 5.4. L'ajustement semble très lisse, mais ce n'est pas anormal car ici les maxima ne sont pas donnés mais inférés. Même en combinant les configurations hydrologiques, l'incertitude sur les maxima annuels reste élevée, et ils peuvent alors s'ajuster plus facilement à la loi GEV.

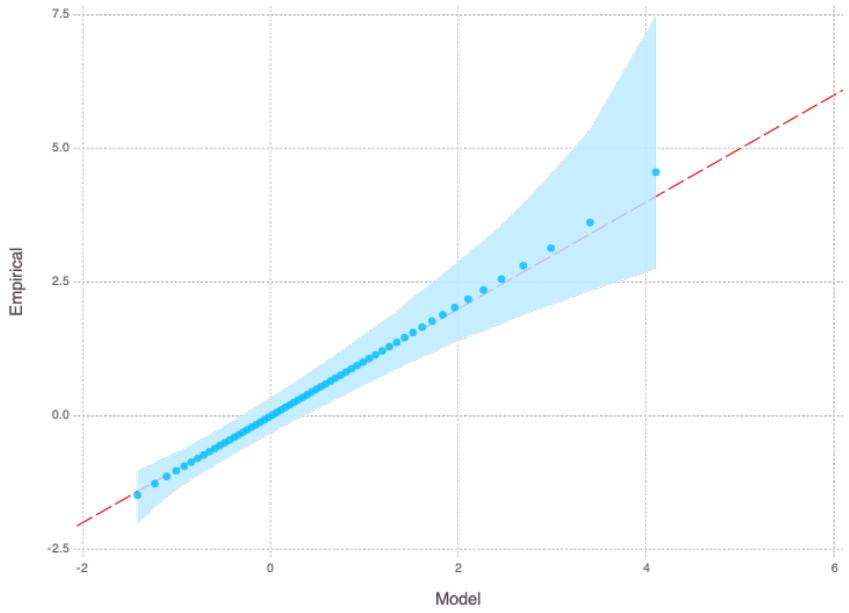


FIGURE 5.4 QQ-plot standardisé pour l'ajustement GEV du modèle pour les pseudo-observations, tronçon SLSO00003.

La figure 5.5 illustre les résultats du modèle en fonction du temps. Chaque point en gris correspond au mode d'une loi log-normale représentant l'incertitude sur le débit d'une configuration du modèle hydrologique. L'espérance des lois *a posteriori* est utilisée comme estimateur ponctuel. La série inférée des maxima de débits (Y_j) ainsi que les intervalles de crédibilité à 95% sont représentés en rouge. L'espérance et les quantiles d'ordre 0.025 et 0.975 des lois GEV inférées sont représentés avec le trait plein orange et le ruban orange, respectivement. L'incertitude sur les données de départ n'est pas représentée, mais elle est très grande, dû à l'erreur d'interpolation des débits dans les bassins non jaugés. Le modèle parvient ici à ré-

duire l'incertitude sur les vrais maxima non observés, en combinant l'information provenant des six configurations hydrologiques et des maxima des autres années. La tendance positive des maxima annuels paraît cohérente.

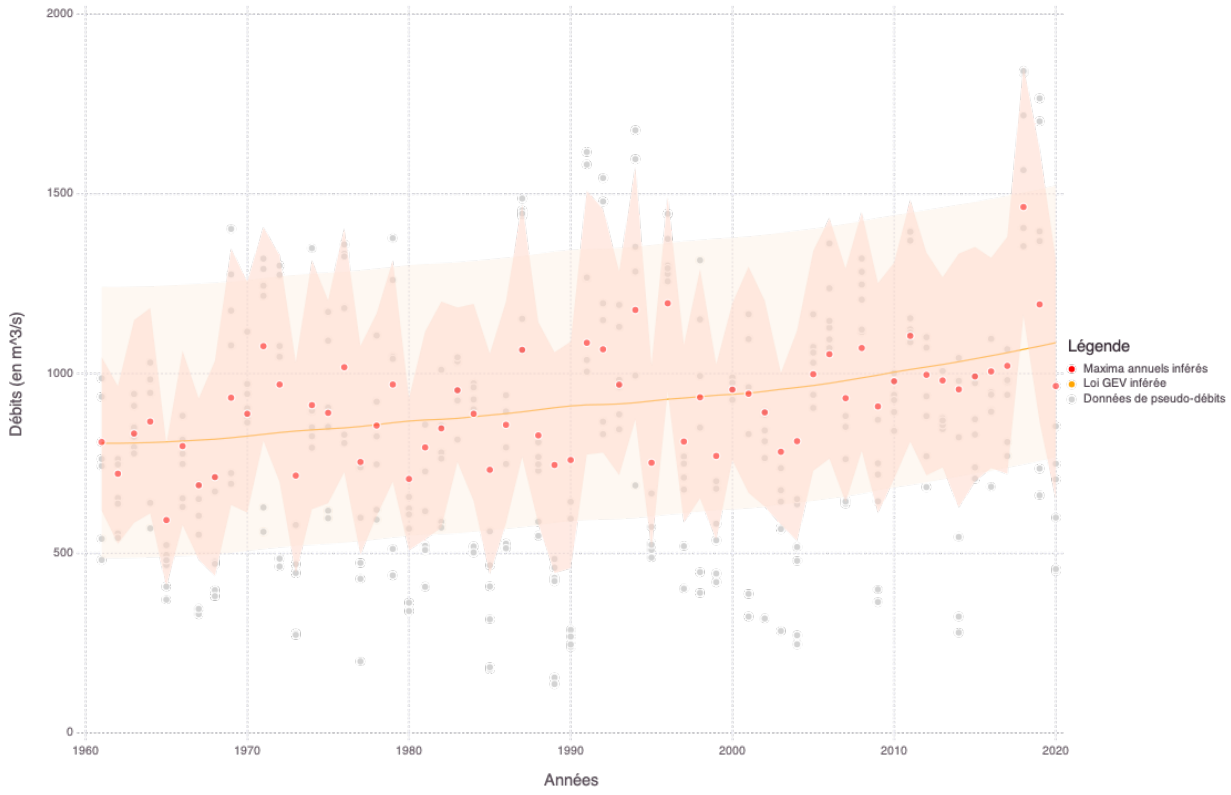


FIGURE 5.5 Pour le tronçon SLSO00003, comparaison des données en entrée (en gris), de la série inférée des maxima avec l'intervalle de crédibilité à 95% (en rouge) et des lois GEV inférées (en orange). Chaque point gris correspond au mode d'une distribution log-normale. Le trait plein orange représente l'espérance des lois GEV. Le ruban orange est délimité par les quantiles d'ordre 0.025 et 0.975 des lois GEV.

Le tableau 5.1 présente les valeurs des paramètres GEV inférés avec les intervalles de crédibilité à 95%. Le paramètre de forme est négatif, ce qui est généralement le cas pour les débits. Cependant, l'intervalle de crédibilité à 95% contient 0, donc il n'est pas exclu que le paramètre de forme soit positif. Le paramètre de non-stationnarité sur la localisation μ_1 est quant à lui strictement positif avec un seuil de crédibilité élevé, indiquant une augmentation significative des maxima annuels.

Enfin, la figure 5.6 présente la densité prédictive du niveau de retour associé à une période de retour 100 ans pour l'année 2020. La meilleure estimation du niveau de débit correspondant à la période de retour 100 ans en 2020 est 1646 m^3/s pour le tronçon SLSO00003.

TABLEAU 5.1 Valeurs des paramètres GEV inférés (avec l'intervalle de crédibilité 95%) du modèle pour les pseudo-observations du tronçon SLSO00003.

μ_0	μ_1	ϕ	ξ
833 (756, 915)	80 (4, 157)	5.13 (4.59, 5.53)	-0.11 (-0.31, 0.12)

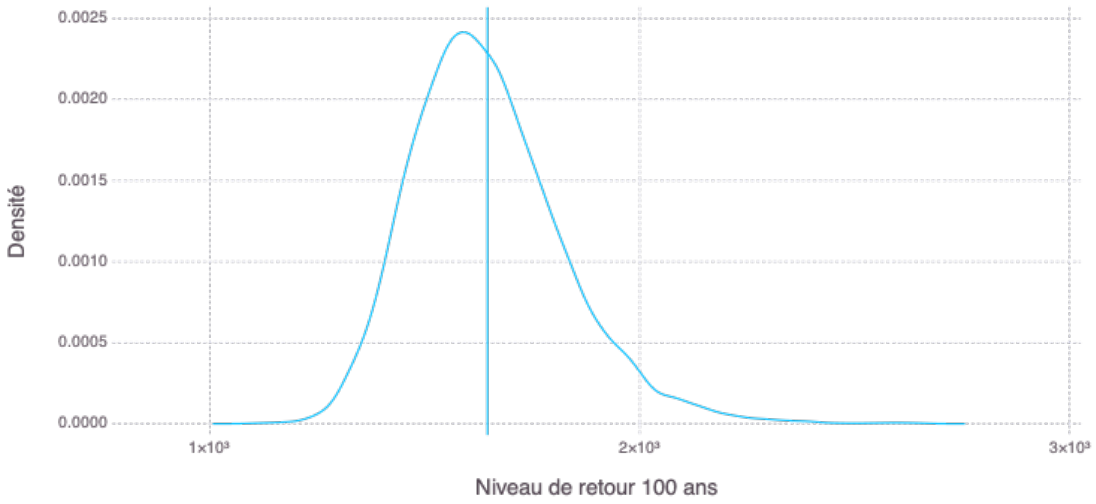


FIGURE 5.6 Pour le tronçon SLSO00003, densité prédictive et espérance du niveau de retour 100 ans en 2020.

CHAPITRE 6 MODÈLE STATISTIQUE POUR LES SIMULATIONS

6.1 Choix préliminaires

L'ensemble des simulations climatiques disponibles constitue une large base de données de débits simulés. Pour rappel, chaque couple GCM-RCM de l'ensemble peut comporter un ou plusieurs membres de simulation, représentant la variabilité naturelle du climat. Un membre peut avoir été forcé avec un ou deux scénarios d'émission (RCP4.5 et RCP8.5). Aussi, chaque membre climatique est utilisé comme entrée pour les six configurations du modèle hydrologique Hydrotel, générant ainsi six membres hydroclimatiques. Pour l'analyse des extrêmes des débits simulés, et en accord avec la DEH et Ouranos, le choix a été fait de considérer les simulations provenant de deux couples GCM-RCM distincts comme indépendantes. Aussi, les simulations d'un même couple GCM-RCM avec des résolutions spatiales différentes ne sont pas considérées comme indépendantes. C'est notamment le cas pour le couple GCM-RCM CanESM2/CRCM5-Ouranos, dont 50 membres ont une résolution de 12 km et 5 membres ont une résolution de 24 km.

Ainsi, le modèle statistique développé sera appliqué aux membres hydroclimatiques provenant d'un même couple GCM-RCM, quelle que soit la résolution. Ces simulations de débits analysées conjointement peuvent avoir été réalisées avec des conditions initiales différentes, une configuration hydrologique différente ou un scénario d'émission différent. Ultérieurement, la détermination des poids accordés à chaque couple GCM-RCM (non étudiée dans ce projet) permettra d'intégrer les résultats sur l'ensemble des simulations, par une approche de moyennage pondéré d'ensemble.

Pour tenir compte des changements climatiques, des lois GEV non-stationnaires sont utilisées pour modéliser les maxima annuels des débits simulés. La concentration annuelle de GES en CO_2 équivalent est utilisée comme variable explicative, celle-ci étant la cause directe des changements climatiques. Ce choix nous permet d'analyser conjointement les débits simulés par le même couple GCM-RCM mais assumant deux scénarios d'émission différents.

Dans ce mémoire, le modèle statistique pour les simulations sera illustré pour deux couples GCM-RCM : CanESM2/CRCM5-Ouranos et IPSL-CM5A-LR, pour le tronçon SLSO00003 situé à l'exutoire de la rivière Chaudière. Il peut être étendu sans difficulté à tous les couples GCM-RCM. Le couple GCM-RCM CanESM2/CRCM5-Ouranos comporte 50 membres provenant de l'ensemble CLIMEX, avec une résolution spatiale de 12 km et 5 membres provenant de l'ensemble CORDEX, avec une résolution spatiale de 24 km. Les membres de CLIMEX

supposent le scénario d'émission RCP8.5, alors que ceux de CORDEX supposent les deux scénarios RCP4.5 et RCP8.5. Le GCM IPSL-CM5A-LR appartient à l'ensemble CMIP5 et comporte 4 membres, simulés sous les deux scénarios d'émission. Sa résolution est d'environ 200 km dans le sens sud-nord et 400 km dans le sens est-ouest ($1.9° \times 3.75°$).

6.2 Modèle statistique

Nous décrivons ci-dessous les trois niveaux du modèle hiérarchique bayésien développé pour les simulations de débits : la couche des données, la couche latente et la couche des hyperparamètres.

6.2.1 La couche des données

Les données d'un même couple GCM-RCM sont constituées de m séries journalières de débits correspondant aux m membres hydroclimatiques, où un membre hydroclimatique est une combinaison unique membre de simulation - configuration hydrologique - scénario d'émission. Pour chaque membre hydroclimatique, la série des maxima annuels est extraite. Soit X_{ij} la variable aléatoire représentant le maximum de débit simulé par le membre i de l'année j. Nous supposons que conditionnellement à l'appartenance au même membre, les maxima de débits sont des observations indépendantes suivant une loi GEV :

$$X_{ij} \sim \mathcal{G}EV(\mu_{0i} + \mu_{1i}u_{ij}, \exp(\phi_{0i} + \phi_{1i}u_{ij}), \xi_i) \quad \text{pour } i = 1, \ldots, m; j = 1, \ldots, n;$$

où

- m est le nombre de membres hydroclimatiques,
- n est le nombre d'années de simulations,
- u_{ij} est la concentration de GES dans l'atmosphère à l'année j du scénario d'émission utilisé dans le membre i.

Dans ce projet, deux modèles plus simples que le modèle non-stationnaire complet sont aussi testés :

- Le modèle stationnaire à 3 paramètres : $X_{ij} \sim \mathcal{G}EV(\mu_{0i}, \exp(\phi_{0i}), \xi_i)$
- Le modèle non-stationnaire à 4 paramètres : $X_{ij} \sim \mathcal{G}EV(\mu_{0i} + \mu_{1i}u_{ij}, \exp(\phi_{0i}), \xi_i)$

Le modèle stationnaire et non-stationnaire à 4 paramètres étant des sous-modèles du modèle non-stationnaire complet, nous nous restreignons à décrire exhaustivement ce dernier. Le paramètre de localisation et le logarithme du paramètre d'échelle sont des fonctions linéaires de la variable explicative. Les paramètres μ_{1i} et ϕ_{1i} modélisent la non-stationnarité

des maxima de débits. Enfin, ξ_i est supposé constant parce qu'il est difficile de détecter une non-stationnarité de ce paramètre [6].

6.2.2 La couche latente

La couche latente décrit la structure de dépendance entre les m membres hydroclimatiques. Ces membres ne sont pas indépendants car ils simulent les débits d'un même tronçon et s'appuient sur le même modèle climatique. Nous supposons alors que les m vecteurs des paramètres de la loi GEV décrivant ces membres sont reliés selon un modèle à effets aléatoires :

$$\begin{aligned}
\mu_{0i} &\sim \mathcal{N}(\nu_0, \tau_0^2) \quad i = 1, \ldots, m \\
\mu_{1i} &\sim \mathcal{N}(\nu_1, \tau_1^2) \quad i = 1, \ldots, m \\
\phi_{0i} &\sim \mathcal{N}(\nu_2, \tau_2^2) \quad i = 1, \ldots, m \\
\phi_{1i} &\sim \mathcal{N}(\nu_3, \tau_3^2) \quad i = 1, \ldots, m \\
\xi_i &\sim \mathcal{N}(\nu_4, \tau_4^2) \quad i = 1, \ldots, m
\end{aligned}$$

En mettant en lien les données provenant des membres hydroclimatiques différents, nous tirons profit de toute l'information disponible. Une représentation du modèle hiérarchique bayésien pour les simulations sous forme de graphe acyclique directif est présentée à la figure 6.1.

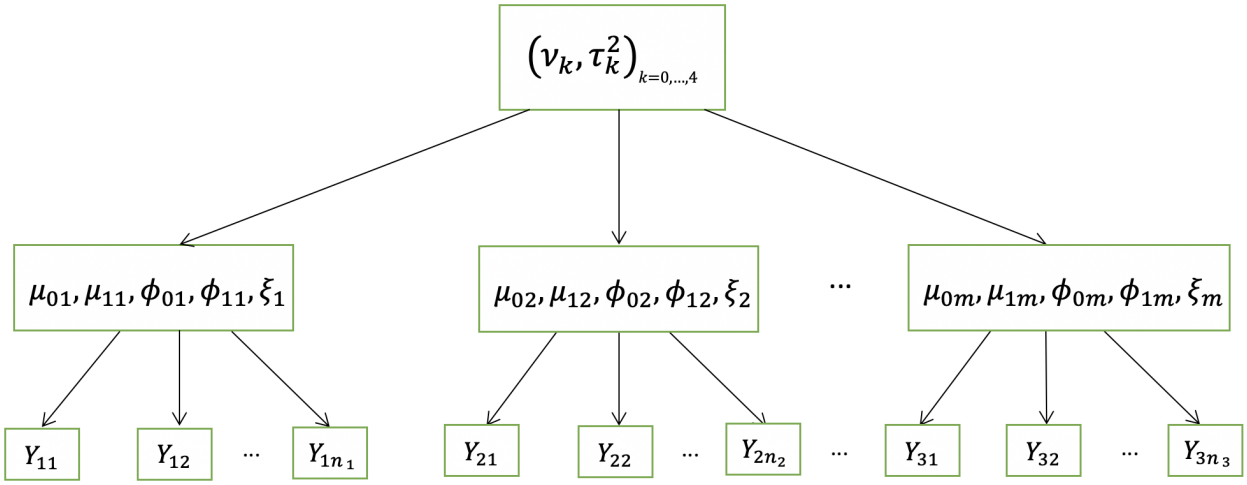


FIGURE 6.1 Représentation schématique du modèle hiérarchique bayésien pour les simulations, avec de haut en bas : la couche des hyperparamètres, la couche latente, la couche des données.

6.2.3 Couche des hyperparamètres

Des lois *a priori* conjuguées peu informatives sont utilisées pour les hyperparamètres $\{(\nu_k, \tau_k^2) : k = 0, \ldots, 4\}$, pour ne pas introduire d'information supplémentaire à celle contenue dans les données. Un tel choix permet aussi de facilement opter pour une loi informative si besoin, en modifiant les valeurs des paramètres de la loi *a priori*, ce qui ne changera pas la structure du modèle.

Une loi conjuguée pour la loi normale $\mathcal{N}(\nu, \tau^2)$ est la loi normale-inverse-gamma. Cette loi, de paramètres $(\rho, \lambda, \alpha, \beta)$, correspond au produit $\mathcal{N}(\nu|\rho, \tau^2/\lambda) \times \mathcal{I}nvGamma(\tau^2|\alpha, \beta)$. Il s'avère que c'est généralement le paramètre de la variance qui est à la source du caractère impropre de la loi *a posteriori* dans un modèle hiérarchique bayésien [51]. Dans la limite $\lambda \to 0$, on a $\nu \propto 1$. On choisit alors comme loi sur les hyperparamètres :

$$f_{(\nu,\tau^2)}(\nu, \tau^2) = \mathcal{I}nvGamma(\tau^2|\alpha, \beta)$$

où $\mathcal{I}nvGamma(\tau^2|\alpha, \beta)$ dénote la densité de la loi inverse-gamma de paramètres (α, β) évaluée en τ^2.

Les données sont en quantité suffisante pour permettre l'utilisation d'une loi *a priori* peu informative : $\alpha = 0.01, \beta = 0.01$. Elle a déjà été utilisée dans des modèles hiérarchiques bayésiens similaires, voir par exemple [22]. La preuve de la propriété de la loi *a posteriori* se trouve en annexe A.

Il est à noter que la loi *a priori* non informative de Jeffreys pour les hyperparamètres $\{(\nu_k, \tau_k^2), k = 0, \ldots, 4\}$:

$$f_{(\nu_k,\tau_k^2)}(\nu_k, \tau_k^2) \propto \frac{1}{\tau_k^2}$$

ne résulte pas en une loi *a posteriori* propre pour ce modèle (voir annexe A). Malgré ses avantages, elle n'était donc pas une option envisageable.

6.3 Estimation des paramètres

La loi *a posteriori* $f_{(\boldsymbol{\theta}|\boldsymbol{X}=\boldsymbol{x})}$ où $\boldsymbol{\theta} = \{(\mu_{0i}, \mu_{1i}, \phi_{0i}, \phi_{1i}, \xi_i)_{1\leq i\leq m}, (\nu_k, \tau_k^2)_{0\leq k\leq 4}\}$ est le vecteur des paramètres à inférer s'écrit alors :

$$\begin{aligned} f_{(\boldsymbol{\theta}|\boldsymbol{X}=\boldsymbol{x})}(\boldsymbol{\theta}) \propto & \prod_{i=1}^{m}\prod_{j=1}^{n} \mathcal{G}EV(X_{ij}|\mu_{0i}+\mu_{1i}u_{ij}, \exp(\phi_{0i}+\phi_{1i}u_{ij}), \xi_i) \\ & \times \prod_{i=1}^{m} \mathcal{N}(\mu_{0i}|\nu_0, \tau_0^2)\mathcal{N}(\mu_{1i}|\nu_1, \tau_1^2)\mathcal{N}(\phi_{0i}|\nu_2, \tau_2^2)\mathcal{N}(\phi_{1i}|\nu_3, \tau_3^2)\mathcal{N}(\xi_i|\nu_4, \tau_4^2) \\ & \times \prod_{k=0}^{4} \mathcal{I}nvGamma(\tau_k^2|\alpha, \beta) \end{aligned}$$

où $\mathcal{N}(x|\nu, \tau^2)$ dénote la densité normale de paramètres (ν, τ^2) évaluée en x.

Les lois conditionnelles complètes sont :

$$\begin{aligned} & f_{(\mu_{0i},\mu_{1i},\phi_{0i},\phi_{1i},\xi_i|-)}(\mu_{0i}, \mu_{1i}, \phi_{0i}, \phi_{1i}, \xi_i) \propto \mathcal{N}(\mu_{0i}|\nu_0, \tau_0^2)\mathcal{N}(\mu_{1i}|\nu_1, \tau_1^2) \\ & \times \mathcal{N}(\phi_{0i}|\nu_2, \tau_2^2)\mathcal{N}(\phi_{1i}|\nu_3, \tau_3^2)\mathcal{N}(\xi_i|\nu_4, \tau_4^2) \times \prod_{j=1}^{n} \mathcal{G}EV(X_{ij}|\mu_{0i}+\mu_{1i}u_{ij}, \exp(\phi_{0i}+\phi_{1i}u_{ij}), \xi_i) \\ & f_{(\nu_k|-)}(\nu_k) = \mathcal{N}\left(\nu_k \middle| \frac{1}{m}\sum_{i=1}^{m}\theta_i, \frac{\tau_k^2}{m}\right) \\ & f_{(\tau_k^2|-)}(\tau_k^2) = \mathcal{I}nvGamma\left(\tau_k^2 \middle| \frac{m}{2}+\alpha, \sum_{i=1}^{m}\frac{(\nu_k-\theta_i)^2}{2}+\beta\right) \end{aligned}$$

où $\theta = \mu_0/\mu_1/\phi_0/\phi_1/\xi$ si $k = 0/1/2/3/4$, respectivement.

Comme pour le modèle pour les pseudo-observations, un algorithme Metropolis-dans-Gibbs est utilisé pour obtenir un échantillon de la loi *a posteriori*. Les lois conditionnelles pour les hyperparamètres sont des lois usuelles facilement simulables. Cependant, pour les paramètres GEV, les lois conditionnelles complètes ne s'expriment pas sous la forme de lois connues. Une étape Metropolis-Hastings est utilisée pour les simuler simultanément selon la loi conditionnelle complète multidimensionnelle $f_{(\mu_{0i},\mu_{1i},\phi_{0i},\phi_{1i},\xi_i|-)}(\mu_{0i}, \mu_{1i}, \phi_{0i}, \phi_{1i}, \xi_i)$. La loi de proposition est une loi normale multivariée, correspondant à une exploration locale autour de la valeur actuelle des paramètres.

Les paramètres GEV sont initialisés avec les valeurs des estimateurs du maximum de la vraisemblance obtenus par une analyse indépendante de chaque membre hydroclimatique. Les valeurs initiales de $\{(\nu_k, \tau_k^2) : k = 0, \ldots, 4\}$ sont les moyennes et écart-types empiriques des estimateurs précédents. Pour chaque membre, la matrice de covariance de la loi de proposition GEV est initialisée comme la matrice de covariance empirique issue de l'estimation

par maximum de la vraisemblance précédente. Elle est mise à jour toutes les 50 itérations MCMC.

Finalement, nous effectuons une standardisation préalable à la fois de la variable d'intérêt (les maxima annuels de débit) et de la variable explicative (la concentration de GES). Cette pratique usuelle (voir par exemple [52, 53]) rend l'algorithme MCMC plus stable numériquement et évite certains comportements pathologiques de non convergence. Elle est adaptée aux lois GEV car les paramètres inférés peuvent être remis à l'échelle par une simple opération linéaire, permettant de retrouver facilement les paramètres de la loi *a posteriori* pour les données de départ.

La convergence des chaines MCMC est évaluée grâce à la visualisation des traces et aux tests de convergence de Gelman-Rubin et de Geweke. Ils indiquent une convergence de l'algorithme après 5000 itérations. Pour s'assurer de la convergence pour tous les tronçons, la période de préchauffage est portée à 20 000 itérations.

6.4 Sélection de modèle

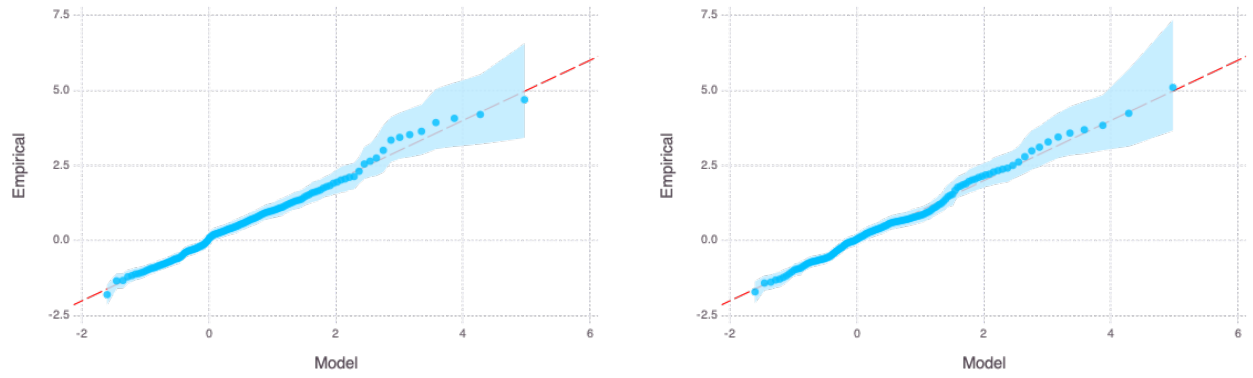
La sélection de modèle s'est faite avec le DIC, comme pour le modèle pour les pseudo-observations. Pour le tronçon SLSO00003, les valeurs du DIC pour les trois modèles testés sont représentées dans le tableau 6.1, pour le couple GCM-RCM CanESM2/CRCM5-Ouranos et le GCM IPSL-CM5A-LR. Dans les deux cas, le meilleur modèle sélectionné est le modèle non-stationnaire complet. La dernière colonne du tableau indique le pourcentage de membres hydroclimatiques où une non-stationnarité significative de la série des maxima annuels est détectée par le test de Mann-Kendall. Un faible pourcentage ne signifie pas forcément que la série est stationnaire, car le test de Mann-Kendall ne détecte que la stricte monotonie d'une série et n'est pas adapté à d'autres comportements non-stationnaires.

Tableau 6.1 Valeurs du DIC des modèles pour les simulations, pour le couple GCM-RCM CanESM2/CRCM5-Ouranos et le GCM IPSL-CM5A-LR. La dernière colonne indique le pourcentage de membres où une non-stationnarité significative est détectée par le test de Mann-Kendall.

	Stat.	Non stat. I	Non stat. II	Mann-Kendall
CanESM2/CRCM5-Ouranos	63799	63551	**63510**	29%
IPSL-CM5A-LR	9362	9349	**9342**	29%

6.5 Validation du modèle et résultats

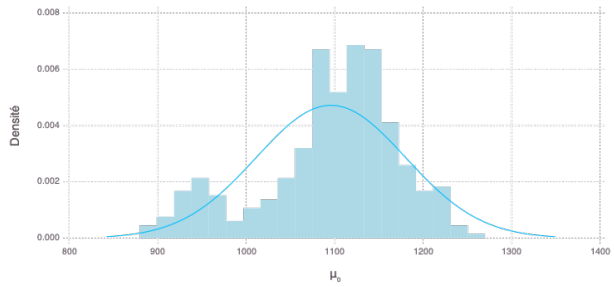
Des QQ-plots standardisés sont utilisés pour valider l'ajustement GEV du modèle, comme pour le modèle pour les pseudo-observations. Ils ont été tracés en utilisant l'espérance *a posteriori* des paramètres GEV comme estimateurs ponctuels. Les quantiles *a posteriori* d'ordre 0.025 et 0.975 sont utilisés pour tracer l'intervalle de crédibilité à 95%. Les QQ-plots pour les simulations du membre 1 de CanESM2/CRCM5-Ouranos et du membre 1 de IPSL-CM5A-LR sont présentés à la figure 6.2. Le résultat témoigne d'un bon ajustement du modèle aux données de simulation.



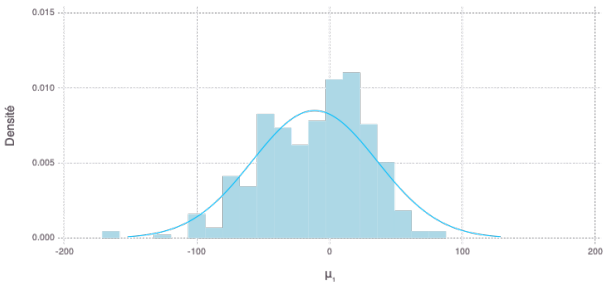
(a) CanESM2/CRCM5-Ouranos (b) IPSL-CM5A-LR

FIGURE 6.2 Pour le tronçon SLSO00003, QQ-plots pour les variables standardisés du modèle pour les simulations, pour le membre 1 de CanESM2/CRCM5-Ouranos et le membre 1 de IPSL-CM5A-LR.

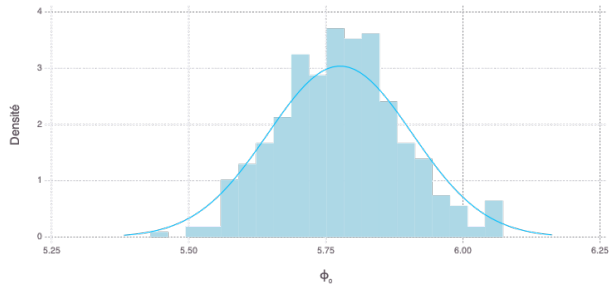
Les valeurs inférées des paramètres du modèle sont présentées à la figure 6.3 pour CanESM2/CRCM5-Ouranos et à la figure 6.4 pour IPSL-CMA5-LR, où l'espérance *a posteriori* est utilisée comme estimateur ponctuel. Les histogrammes résument l'information sur les paramètres GEV inférés pour tous les membres hydroclimatiques. Les courbes sont les densités normales de paramètres $\{(\nu_k, \tau_k^2) : k = 0, \ldots, 4\}$. Pour les deux couples GCM-RCM, les estimations des paramètres de tendance sur la localisation μ_{1j} sont réparties symétriquement autour de 0, indiquant une tendance linéaire incertaine globalement. Les résultats indiquent aussi une dispersion plus importante des débits futurs avec les changements climatiques, comme le montre les valeurs plutôt positives des paramètres ϕ_{1j}. Si un modèle stationnaire était utilisé, pour une partie des membres hydroclimatiques exhibant une non-stationnarité significative, l'ajustement serait de moins bonne qualité. C'est pour cette raison que le DIC tend à favoriser le modèle non-stationnaire complet ici. Pour CanESM2/CRCM5-Ouranos, les estimations des paramètres de forme ξ_j sont plutôt également réparties autour de 0, alors que



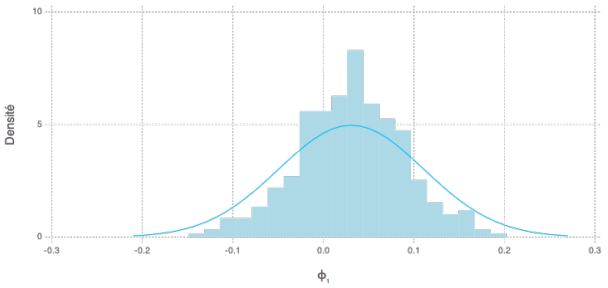
(a) Histogramme de $(\hat{\mu}_{0i} : 1 \leq i \leq m)$ et densité de la loi $\mathcal{N}(\hat{\nu}_0, \hat{\tau}_0^2)$



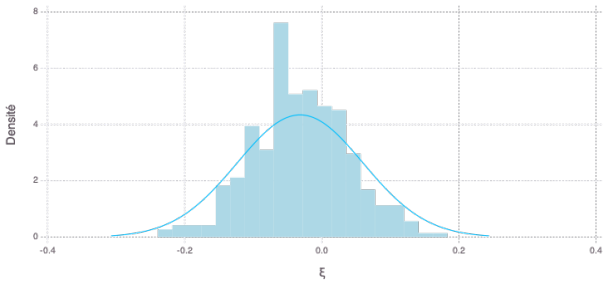
(b) Histogramme de $(\hat{\mu}_{1i} : 1 \leq i \leq m)$ et densité de la loi $\mathcal{N}(\hat{\nu}_1, \hat{\tau}_1^2)$



(c) Histogramme de $(\hat{\phi}_{0i} : 1 \leq i \leq m)$ et densité de la loi $\mathcal{N}(\hat{\nu}_2, \hat{\tau}_2^2)$

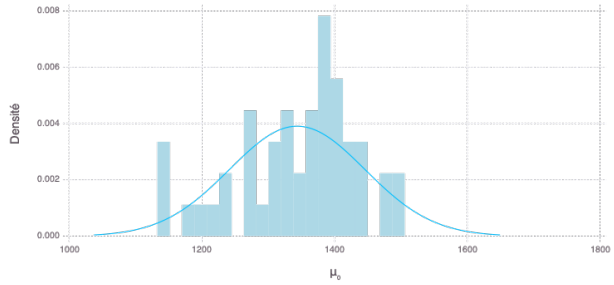


(d) Histogramme de $(\hat{\phi}_{1i} : 1 \leq i \leq m)$ et densité de la loi $\mathcal{N}(\hat{\nu}_3, \hat{\tau}_3^2)$

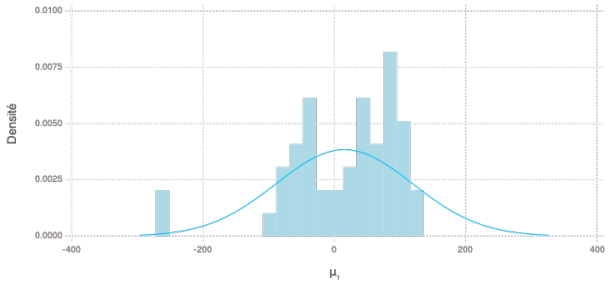


(e) Histogramme de $(\hat{\xi}_i : 1 \leq i \leq m)$ et densité de la loi $\mathcal{N}(\hat{\nu}_4, \hat{\tau}_4^2)$

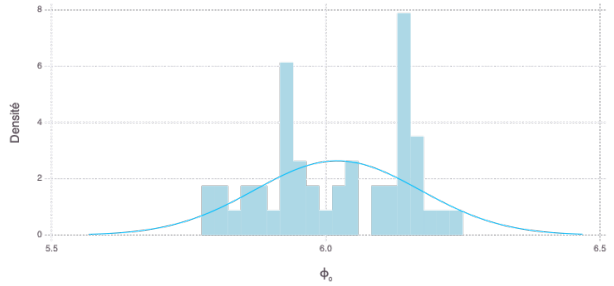
FIGURE 6.3 Pour le tronçon SLSO00003, représentation des valeurs inférées pour tous les paramètres du modèle pour les simulations du couple GCM-RCM CanESM2/CRCM5-Ouranos. Les histogrammes représentent les estimations des paramètres de la loi GEV. Les courbes sont des gaussiennes de paramètres les valeurs inférées des hyperparamètres.



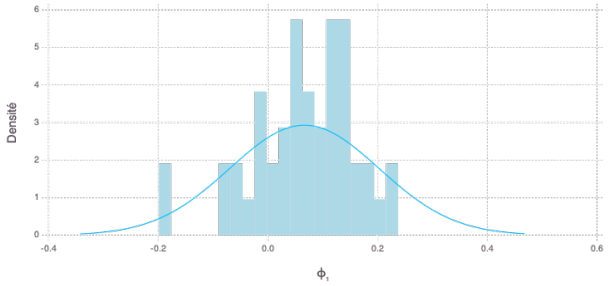
(a) Histogramme de $(\hat{\mu}_{0i} : 1 \leq i \leq m)$ et densité de la loi $\mathcal{N}(\hat{\nu}_0, \hat{\tau}_0^2)$



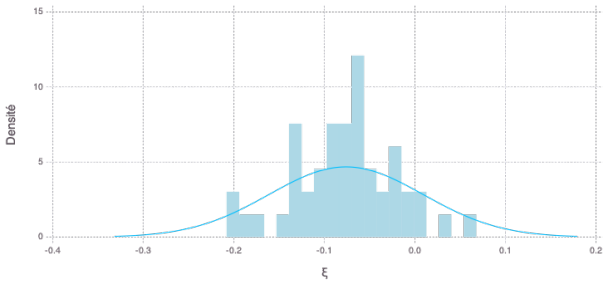
(b) Histogramme de $(\hat{\mu}_{1i} : 1 \leq i \leq m)$ et densité de la loi $\mathcal{N}(\hat{\nu}_1, \hat{\tau}_1^2)$



(c) Histogramme de $(\hat{\phi}_{0i} : 1 \leq i \leq m)$ et densité de la loi $\mathcal{N}(\hat{\nu}_2, \hat{\tau}_2^2)$



(d) Histogramme de $(\hat{\phi}_{1i} : 1 \leq i \leq m)$ et densité de la loi $\mathcal{N}(\hat{\nu}_3, \hat{\tau}_3^2)$



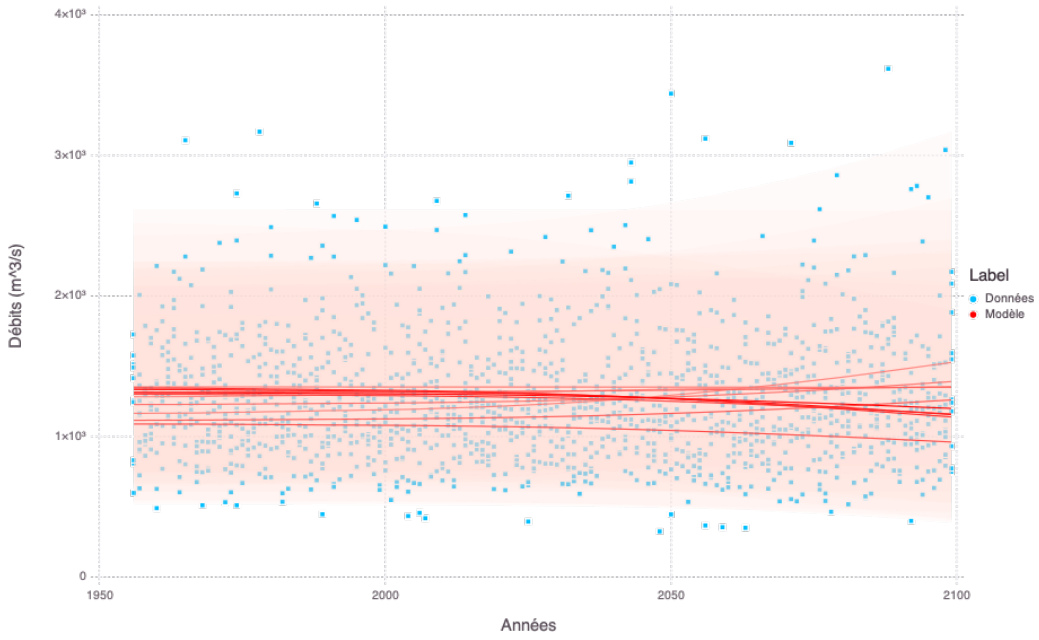
(e) Histogramme de $(\hat{\xi}_i : 1 \leq i \leq m)$ et densité de la loi $\mathcal{N}(\hat{\nu}_4, \hat{\tau}_4^2)$

FIGURE 6.4 Pour le tronçon SLSO00003, représentation des valeurs inférées pour tous les paramètres du modèle pour les simulations du GCM IPSL-CM5A-LR. Les histogrammes représentent les estimations des paramètres de la loi GEV. Les courbes sont des gaussiennes de paramètres les valeurs inférées des hyperparamètres.

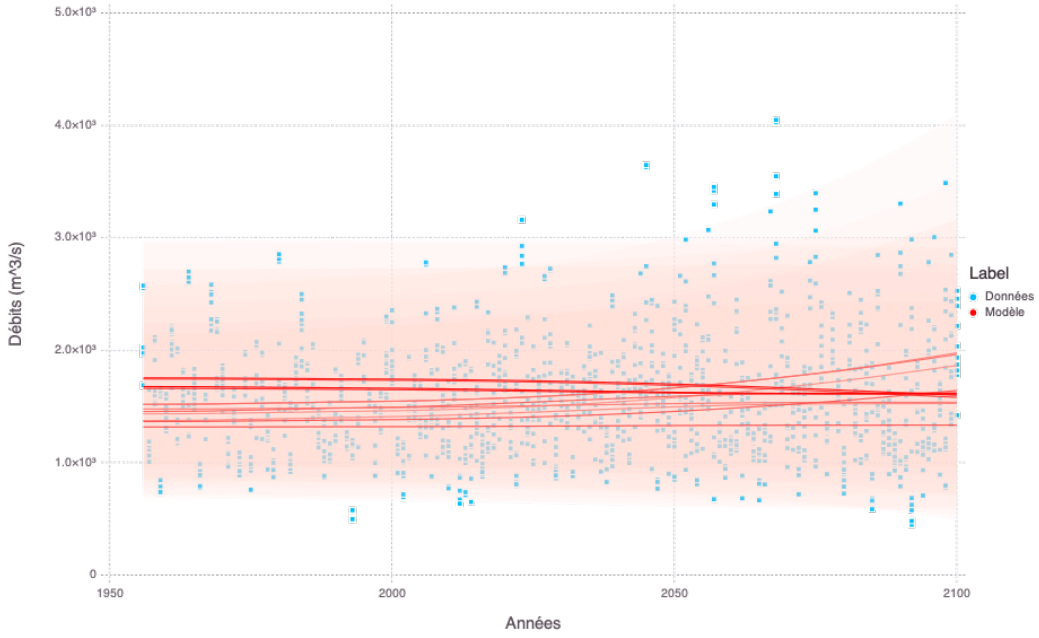
pour IPSL-CMA5-LR, elles ont tendance à être négatives.

Sur les figures 6.3 et 6.4, on s'attend à ce que l'histogramme des paramètres de la loi GEV soit cohérent avec la densité de la loi normale de paramètres les hyperparamètres correspondants. La modélisation de la couche latente avec la loi normale est plutôt adaptée pour CanESM2/CRCM5-Ouranos. Pour IPSL-CMA5-LR, le nombre de membres étant plus petit, l'ajustement normal est de moins bonne qualité. Pour certains paramètres comme les (μ_{1i}) et (ϕ_{0i}), on observe un caractère bimodal qui peut s'expliquer par l'existence de deux scénarios d'émission dans ce modèle climatique (RCP4.5 et RCP8.5). Nous avons essayé de tenir compte de cette hétérogénéité en introduisant la concentration de GES comme variable explicative, mais il apparaît que cela n'est peut-être pas suffisant pour expliquer toute la variabilité liée aux scénarios d'émission. C'est une des limites du modèle développé pour les simulations.

Pour mieux cerner la non-stationnarité de ces modèles, nous présentons l'évolution temporelle des lois GEV inférées pour quelques membres de simulation du couple GCM-RCM CanESM2/CRCM5-Ouranos (figure 6.5a) et du GCM IPSL-CMA5-LR (figure 6.5b). Chaque trait plein rouge correspond à l'espérance de la loi GEV inférée pour un membre. Le ruban autour délimite les quantiles d'ordre 0.025 et 0.975 de la loi GEV. Pour faciliter la lecture, seuls 10 membres hydroclimatiques sont représentés. Les données en entrée du modèle sont présentées en bleu, pour comparaison. On observe d'abord un bon ajustement des modèles aux données. Une tendance linéaire significative est indétectable à première vue, mais on remarque une dispersion légèrement plus importante des débits maximaux en 2100 par rapport à 1955, comportement qui semble être bien capturé par le modèle hiérarchique bayésien dans les deux cas.



(a) CanESM2/CRCM5-Ouranos (10 membres)



(b) IPSL-CM5A-LR (10 membres)

FIGURE 6.5 Comparaison des données de simulation et du modèle inféré, pour le tronçon SLSO00003. Les points bleus sont les données de maxima annuels de débits. Les traits pleins en rouge sont la moyenne des 10 lois GEV inférés (10 membres). Le ruban rouge correspond à l'écart entre le quantile d'ordre 0.025 et celui d'ordre 0.975 des lois GEV.

CHAPITRE 7 MODÈLE STATISTIQUE POUR LA PRÉDICTION DES DÉBITS FUTURS

7.1 Post-traitement statistique pour la jonction des modèles

Les simulations climatiques sont des outils indispensables pour étudier l'évolution des phénomènes climatiques et hydrologiques. Néanmoins, ces simulations comportent généralement un biais par rapport au climat réellement observé [14]. Pour le tronçon SLSO00003, la figure 7.1 présente les densités des lois GEV inférées décrivant les maxima de débits pseudo-observés (bleu), simulés par le membre 1 de CanESM2/CRCM5-Ouranos (rouge) et simulés par le membre 1 d'IPSL-CM5A-LR (orange) en 2020. Comparées aux pseudo-observations, les simulations surestiment la moyenne et la dispersion des débits.

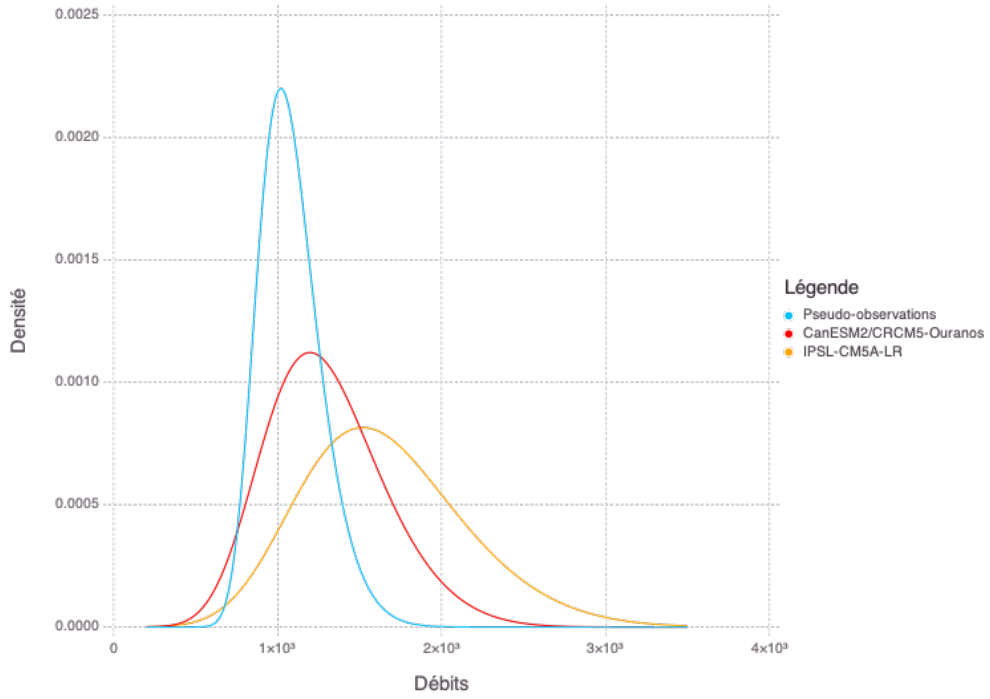


FIGURE 7.1 Pour le tronçon SLSO00003, densités des lois GEV inférées pour les pseudo-observations, les simulations du membre 1 de CanESM2/CRCM5-Ouranos et les simulations du membre 1 d'IPSL-CM5A-LR, en 2020.

Pour corriger ce biais et faire la jonction entre le modèle pour les pseudo-observations et le modèle pour les simulations, nous appliquons la méthode CDF-*transform* décrite dans

la section 2.4.2. Elle est adaptée à notre problème d'analyse fréquentielle, en fournissant la fonction de répartition des débits futurs.

Soit le membre hydroclimatique i d'un couple GCM-RCM. En reprenant les notations de la section 2.4.2, F_{Y_c} est ici la fonction de répartition GEV aux paramètres estimés avec le modèle statistique pour les pseudo-observations. F_{X_c} et F_{X_p} sont les fonctions de répartition GEV aux paramètres estimés avec le modèle statistique pour les simulations, pour le membre i. Puisque la modélisation des simulations est continue et non-stationnaire, elle ne fait pas la distinction entre une période de calibration et une période de projection. Il est alors nécessaire de choisir une année de calibration c et une année de projection p. Nous faisons le choix de ne pas interpoler les débits à une année p plus tard que la dernière année de données simulées, soit 2100. En effet, le système climatique étant hautement complexe et comportant des boucles de rétroaction non linaires, il serait hasardeux de prédire les débits à des dates futures en se basant uniquement sur la concentration de GES, sans accès à des simulations climatiques.

En reprenant les notations des chapitres 5 et 6

$$\begin{aligned} Y_c &\sim \mathcal{G}EV(\mu_0 + \mu_1 u_c, \exp(\phi), \xi) \\ X_c &\sim \mathcal{G}EV(\mu_{0i} + \mu_{1i} u_{ic}, \exp(\phi_{0i} + \phi_{1i} u_{ic}), \xi_i) \\ X_p &\sim \mathcal{G}EV(\mu_{0i} + \mu_{1i} u_{ip}, \exp(\phi_{0i} + \phi_{1i} u_{ip}), \xi_i) \end{aligned}$$

et $(\mu_0, \mu_1, \phi, \xi, \mu_{0i}, \mu_{1i}, \phi_{0i}, \phi_{1i}, \xi_i)$ sont les paramètres GEV issus des modèles pour les pseudo-observations et les simulations. Les variables u_t et u_{it} sont la concentration de GES à l'année t et celle à l'année t pour le membre i, respectivement. Un calcul rapide montre alors que la fonction de répartition des débits projetés F_{Y_p} est aussi GEV :

$$Y_p \sim \mathcal{G}EV(\eta_*, \exp(\phi_*), \xi_*) \tag{7.1}$$

$$\begin{aligned} \eta_* &= \mu_{0i} + \mu_{1i} u_{ip} + (\mu_0 + \mu_1 u_c - \mu_{0i} - \mu_{1i} u_{ic}) \exp(\phi_{1i}(u_{ip} - u_{ic})) \\ \phi_* &= \phi + \phi_{1i}(u_{ip} - u_{ic}) \\ \xi_* &= \xi \end{aligned}$$

On observe que le paramètre de forme de la distribution des débits projetés est identique à celui pour les débits pseudo-observés historiques. Ceci résulte de l'hypothèse de stationnarité du paramètre de forme pour la série des maxima annuels de débits simulés. Ici, la fonction de transfert utilise le signal des changements climatiques provenant du modèle pour les simulations pour calibrer les changements attendus sur les observations. Pour les

pseudo-observations, ce transfert correspond à une transformation affine des débits observés de l'année de calibration à l'année de projection :

$$x \mapsto \frac{\sigma_p}{\sigma_c} x + \left(\mu_p - \frac{\sigma_p}{\sigma_c} \mu_c \right)$$

$$\begin{aligned} \mu_c &= \mu_{0i} + \mu_{1i} u_{ic} \\ \mu_p &= \mu_{0i} + \mu_{1i} u_{ip} \\ \sigma_c &= \exp(\phi_{0i} + \phi_{1i} u_{ic}) \\ \sigma_p &= \exp(\phi_{0i} + \phi_{1i} u_{ip}) \end{aligned}$$

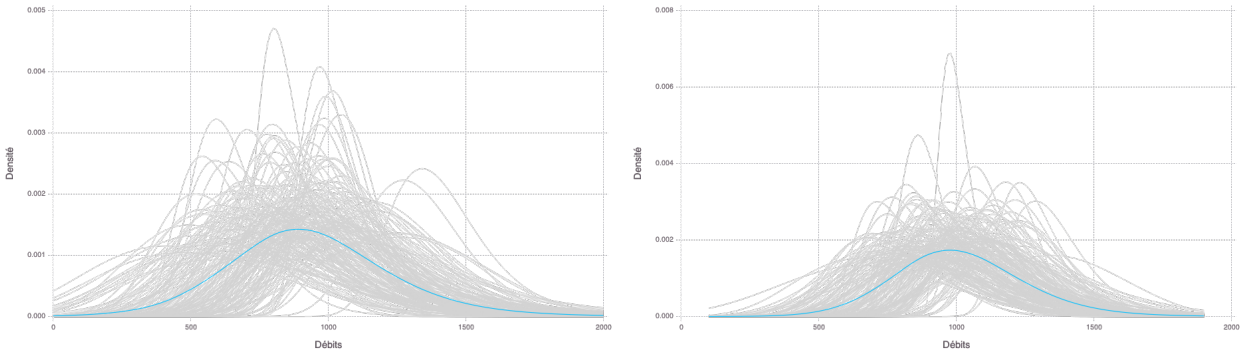
Si le modèle pour les simulations est non-stationnaire à 4 paramètres, on a $\sigma_c = \sigma_p$. Le transfert par méthode CDF-*transform* est alors une translation $x \mapsto x + \mu_p - \mu_c$. Si ce modèle est stationnaire, $\sigma_c = \sigma_p$ et $\mu_p = \mu_c$. Le transfert correspond alors à l'identité. Autrement dit, le débit prédit à l'année de projection est le même que celui pseudo-observé à l'année de référence c. Nous comprenons alors que la prise en compte de la non-stationnarité dans le modèle pour les simulations est cruciale et impacte directement les résultats obtenus.

Pour quantifier l'incertitude sur les débits projetés, nous calculons la loi prédictive des débits futurs de façon bayésienne. Pour cela, l'équation 7.1 est calculée avec les valeurs des paramètres GEV de chaque itération MCMC en sortie des modèles pour les simulations et les pseudo-observations. La figure 7.2 présente la densité de la loi prédictive du débit en 2099 en bleu pour le tronçon SLSO00003, pour le membre 1 de CanESM2/CRCM5-Ouranos et le membre 1 de IPSL-CM5A-LR. Les courbes en gris sont les densités de Y_p pour chaque itération MCMC. L'année de calibration est 2020. Par la suite, la méthode CDF-*transform* est appliquée à chaque membre hydroclimatique d'un couple GCM-RCM, et l'intégration des résultats sur tous les membres représente toute l'incertitude sur la chaîne de modélisation.

De la même façon, la loi prédictive des niveaux de retour futurs doit être calculée de manière bayésienne, en utilisant les lois *a posteriori* des paramètres GEV issus des modèles pour les simulations et les pseudo-observations. On a alors accès facilement à l'intervalle de crédibilité à 95 % de la loi prédictive, ainsi qu'à tous les estimateurs ponctuels associés.

7.2 Débits projetés pour un tronçon

La figure 7.3 illustre la différence entre les sorties du modèle pour les simulations (pour 10 membres, en rouge) et les sorties du modèle pour les pseudo-observations (en bleu). Pour les deux couples GCM-RCM, les simulations hydroclimatiques surestiment les débits, si l'on



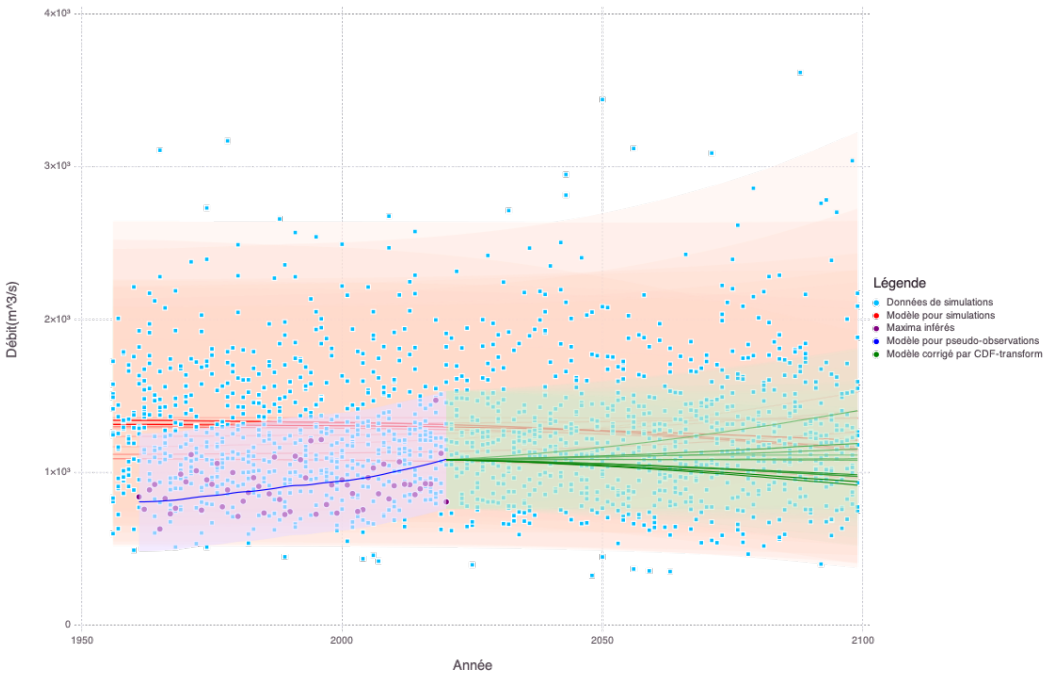
(a) Membre 1, CanESM2/CRCM5-Ouranos (b) Membre 1, IPSL-CM5A-LR

FIGURE 7.2 Loi prédictive des débits du tronçon SLSO00003 en 2099 (courbe bleu). Les densités de Y_p pour chaque itération MCMC sont en gris. L'année de calibration est 2020.

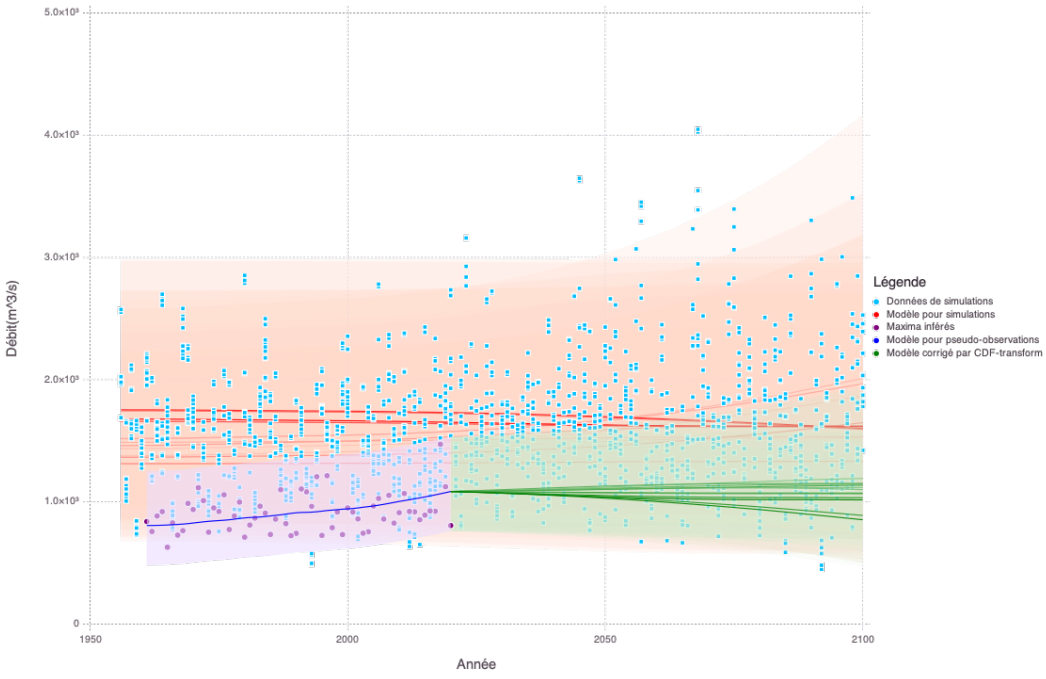
compare avec les pseudo-observations sur la période historique. Ce phénomène est plus marqué pour IPSL-CM5A-LR. Étant donné que les simulations d'IPSL-CM5A-LR utilisent un modèle global du climat, on s'attend à ce que les données soient de moins bonne qualité que pour CanESM2/CRCM5-Ouranos.

Pour le tronçon SLSO00003, la tendance positive des pseudo-observations est bien plus marquée que celles des débits simulés. Pour les deux couples GCM-RCM à l'étude, la borne supérieure de l'intervalle de crédibilité à 95% est plus petite pour les pseudo-observations que pour les simulations, notamment pour deux raisons. D'une part, la valeur inférée du paramètre de forme est plus grande pour les simulations que pour les pseudo-observations. D'autre part, les modèles climatiques comportent plusieurs membres pour rendre compte de la variabilité interne du climat, ce qui augmente l'incertitude sur les débits extrêmes. Ce biais entre les distributions simulées et observées est corrigé par la jonction des modèles.

La figure 7.3 présente également les sorties du modèle prédictif des débits futurs pour 10 membres hydroclimatiques après la correction de biais, en vert. L'année de calibration est ici la dernière année des pseudo-observations, 2020. On observe que les caractéristiques non-stationnaires des débits simulés sont transférés à la série des débits pseudo-observés (plus grande dispersion future, tendance linéaire incertaine), et que le biais a été corrigé. Les distributions obtenues combinent les caractéristiques des pseudo-observations (moyenne, épaisseur de la queue de distribution) et le signal de changements climatiques contenu dans les simulations. Les distributions prédictives des débits futurs ont alors une plus grande dispersion que la distribution des pseudo-observations historiques. Aussi, la tendance en moyenne est moins marquée que pour les pseudo-observations historiques.

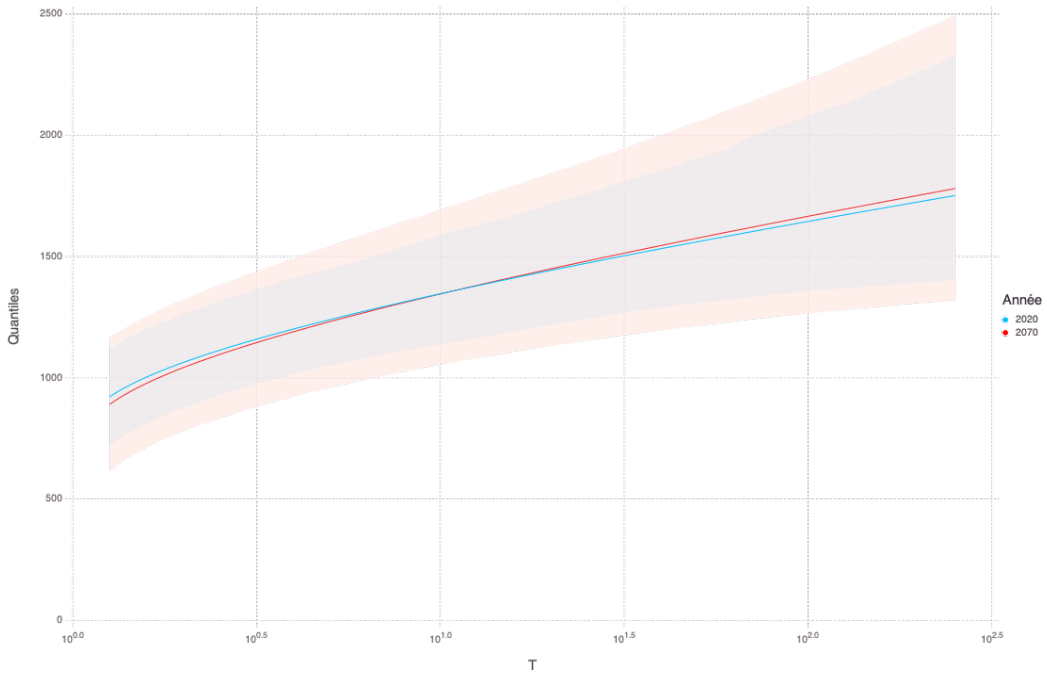


(a) CanESM2/CRCM5-Ouranos (10 membres)

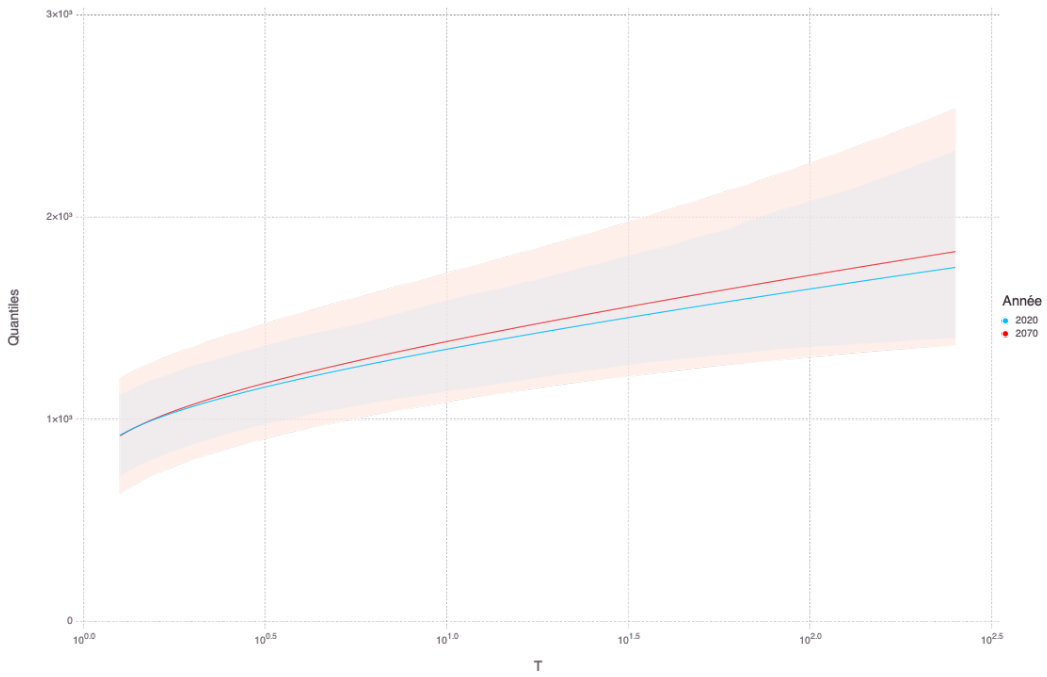


(b) IPSL-CMA5-LR (10 membres)

FIGURE 7.3 Résultats du modèle pour les simulations (en rouge), du modèle pour les pseudo-observations (en bleu) et de la jonction des deux modèles (en vert), pour le tronçon SLSO00003. Les traits pleins correspondent à l'espérance des lois GEV inférées. Les rubans sont délimités par les quantiles d'ordre 0.025 et 0.975 des lois GEV inférées.



(a) CanESM2/CRCM5-Ouranos



(b) IPSL-CMA5-LR

FIGURE 7.4 Pour le tronçon SLSO00003, estimation des niveaux de retour en fonction de la période de retour en 2020 et 2070, avec l'intervalle de crédibilité à 95 %.

La figure 7.4 présente les niveaux de retour prédits en fonction des périodes de retour, en 2020 (bleu) et 2070 (rouge), avec leurs intervalles de crédibilité à 95 %. Comme la tendance temporelle moyenne des débits projetés n'est pas certaine, l'écart entre les traits rouge et bleu, qui exprime le signal des changements climatiques entre 2020 et 2070, n'est pas significatif. On observe tout de même une légère tendance à la hausse des niveaux de retour pour le GCM IPSL-CMA5-LR en 50 ans. Cette hausse peut résulter de la plus grande dispersion des débits extrêmes qui a tendance à accroître la valeur des quantiles les plus élevés. Pour les deux couples GCM-RCM, l'intervalle de crédibilité 95% s'élargit avec la période de retour, un comportement usuel en théorie des valeurs extrêmes qui traduit l'incertitude grandissante des prévisions quand l'événement devient plus rare. Il est aussi plus grand pour l'année 2070 que 2020, indiquant que les prévisions deviennent plus incertaines pour un horizon temporel lointain. Enfin, la concavité des courbes de niveaux de retour est liée à la négativité du paramètre de forme pour les pseudo-observations.

Comme autre point de vue sur l'influence des changements climatiques sur les niveaux de retours, la figure 7.5 présente la densité prédictive du niveau de retour 100 ans en 2070 utilisant le couple GCM-RCM CanESM2/CRCM5-Ouranos (en rouge) ou le GCM IPSL-CM5A-LR (en orange). Cette visualisation confirme les commentaires déjà faits. Pour les deux couples GCM-RCM, on observe un comportement très similaire de dispersion du débit centennal (niveau de retour 100 ans), alors qu'en moyenne le niveau projeté est quasiment le même qu'en période historique (2020). Cette plus grande dispersion résulte d'une part du comportement attendu des débits extrêmes, d'autre part de la prise en compte de l'incertitude du système climatique futur et l'incertitude de modélisation.

7.3 Débits projetés pour l'ensemble des simulations climatiques

Pour le tronçon SLSO00003, nous présentons les résultats sur les niveaux de retour pour l'ensemble des simulations climatiques, partitionné en 42 couples GCM-RCM. En annexe B, le lecteur ou la lectrice peut trouver les informations sur cette partition et le meilleur modèle GEV choisi pour chaque couple GCM-RCM. L'équivalence entre la codification GCM-RCM et les modèles climatiques correspondants se trouve dans le tableau B.2. La plupart des couples GCM-RCM ont 10 membres ou moins, et la plupart considère deux scénarios d'émission de GES (RCP4.5 et RCP8.5). Seuls quatre couples GCM-RCM comportent plus de 40 membres : le GCM IPSL-CM5A-LR (IAL) avec 40 membres, le GCM CanESM2 (CE2) avec 50 membres, le GCM CSIRO-Mk3-6-0 (CSI) avec 100 membres et le couple GCM-RCM CanEMS2/CRCM5-Ouranos (CE2/CO) avec 280 membres. Le meilleur modèle choisi par le DIC est stationnaire, non-stationnaire à 4 paramètres et non-stationnaire complet pour 8, 21

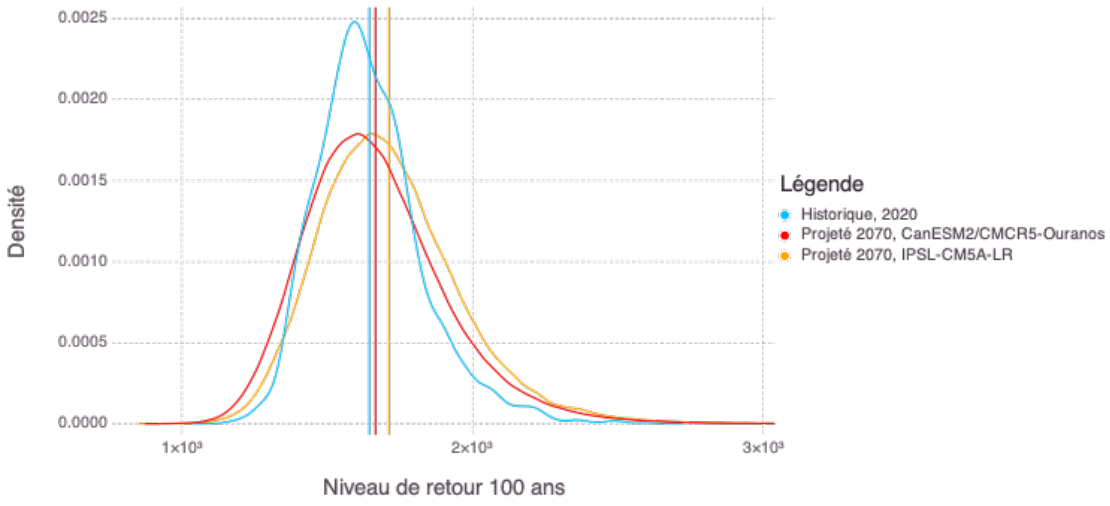
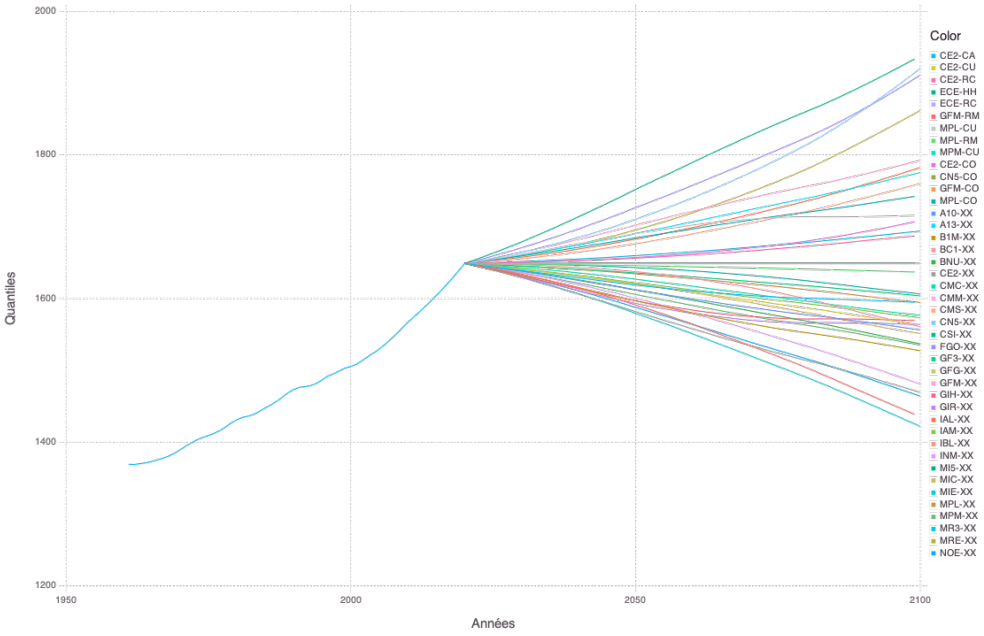


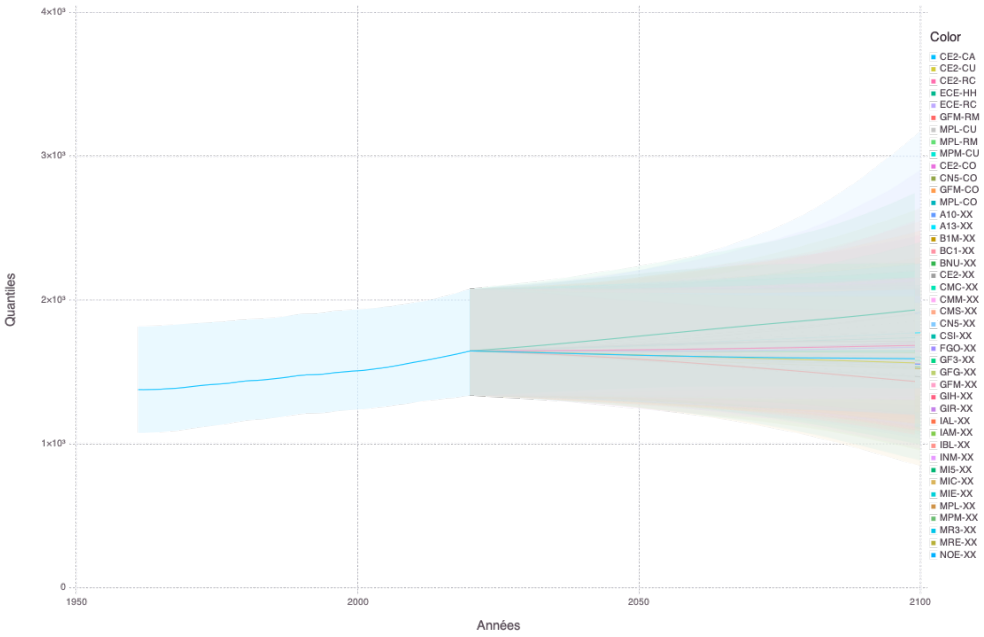
FIGURE 7.5 Pour le tronçon SLSO00003, densités prédictives et espérances du niveau de retour 100 ans en 2020 et projeté par les modèles CanESM2/CRCM5-Ouranos et IPSL-CM5A-LR en 2070.

et 13 couples GCM-RCM respectivement.

Pour le tronçon SLSO00003, la figure 7.6a présente le niveau de retour 100 ans projeté jusqu'en 2100 pour l'ensemble des simulations climatiques. L'incertitude sur chaque courbe de niveau de retour n'est pas présentée pour plus de lisibilité, mais elle est bien plus importante que la variabilité entre les couples GCM-RCM (voir figure 7.6b). L'année de calibration pour la jonction des modèles est la dernière année de pseudo-observations, 2020. On observe une grande variabilité des débits centennaux projetés en fonction du couple GCM-RCM, avec des tendances à la hausse comme à la baisse et des débits variant de 1400 m^3/s à 1900 m^3/s en 2100. Le nombre de couples GCM-RCM prédisant des niveaux de débit centennal à la hausse, à la baisse et constants est respectivement 13, 21 et 8. Par la suite, une pondération adéquate des différents couples GCM-RCM permettrait d'obtenir une courbe de niveau de retour projeté unique, avec son intervalle de crédibilité à 95%.



(a) Sans intervalles de crédibilité



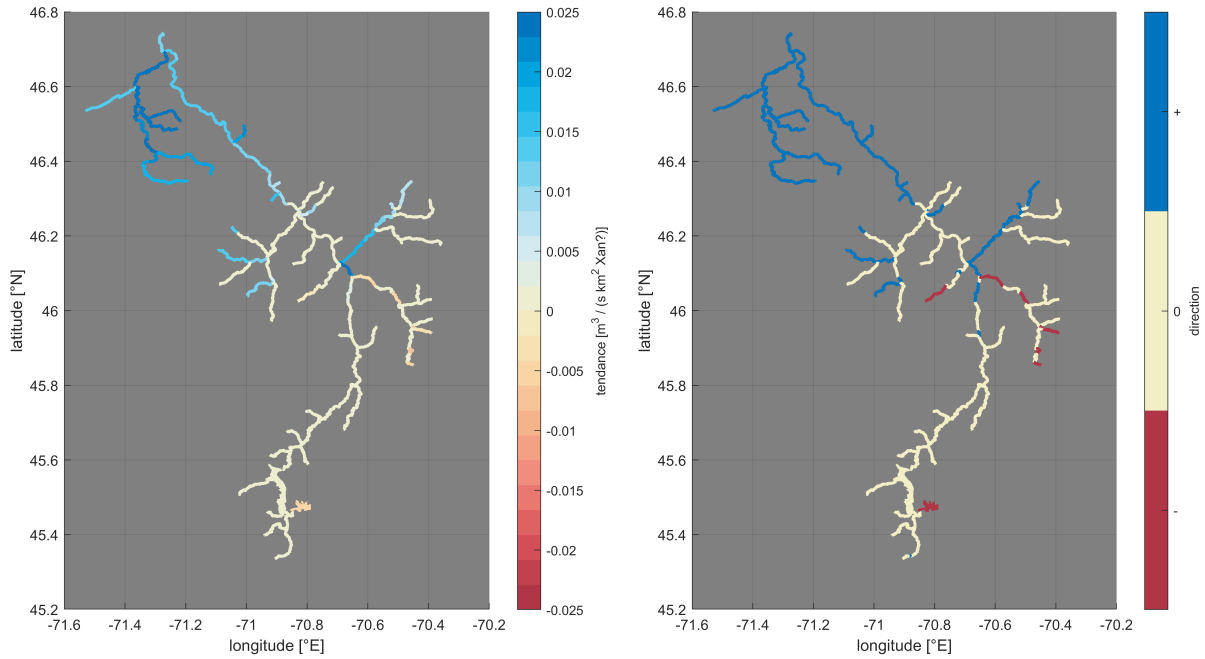
(b) Avec intervalles de crédibilité à 95 %

FIGURE 7.6 Pour le tronçon SLSO00003, estimations des niveaux de retour 100 ans projetés jusqu'en 2100 pour l'ensemble des simulations hydroclimatiques. L'année de calibration est 2020. Chaque trait plein correspond à un couple GCM-RCM. Les intervalles de crédibilité à 95 % sont inclus dans la figure (b).

CHAPITRE 8 RÉSULTATS GLOBAUX

Les résultats sur les niveaux de retour projetés sont présentés pour les 211 tronçons de la rivière Chaudière dont la surface de drainage dépasse 50 km^2, avec deux couples GCM-RCM : CanESM2/CRCM5-Ouranos et IPSL-CMA5-LR.

La sélection de modèle pour les pseudo-observations favorise le modèle stationnaire pour 111 tronçons et le modèle non-stationnaire pour 100 tronçons. Pour IPSL-CMA5-LR, le meilleur modèle pour les simulations est stationnaire, non-stationnaire à 4 paramètres et non-stationnaire complet pour 18, 16 et 177 tronçons respectivement. Quant à CanESM5/CRCM5-Ouranos, le meilleur modèle est non-stationnaire à 4 paramètres pour 11 tronçons et non-stationnaire complet pour 200 tronçons. Le modèle stationnaire n'est choisi pour aucun tronçon. Pour favoriser une cohérence spatiale sur l'ensemble des débits simulés, et puisque le modèle GEV non-stationnaire complet inclut les deux autres, ce dernier sera utilisé pour l'analyse fréquentielle des débits pour tous les tronçons de la rivière Chaudière.

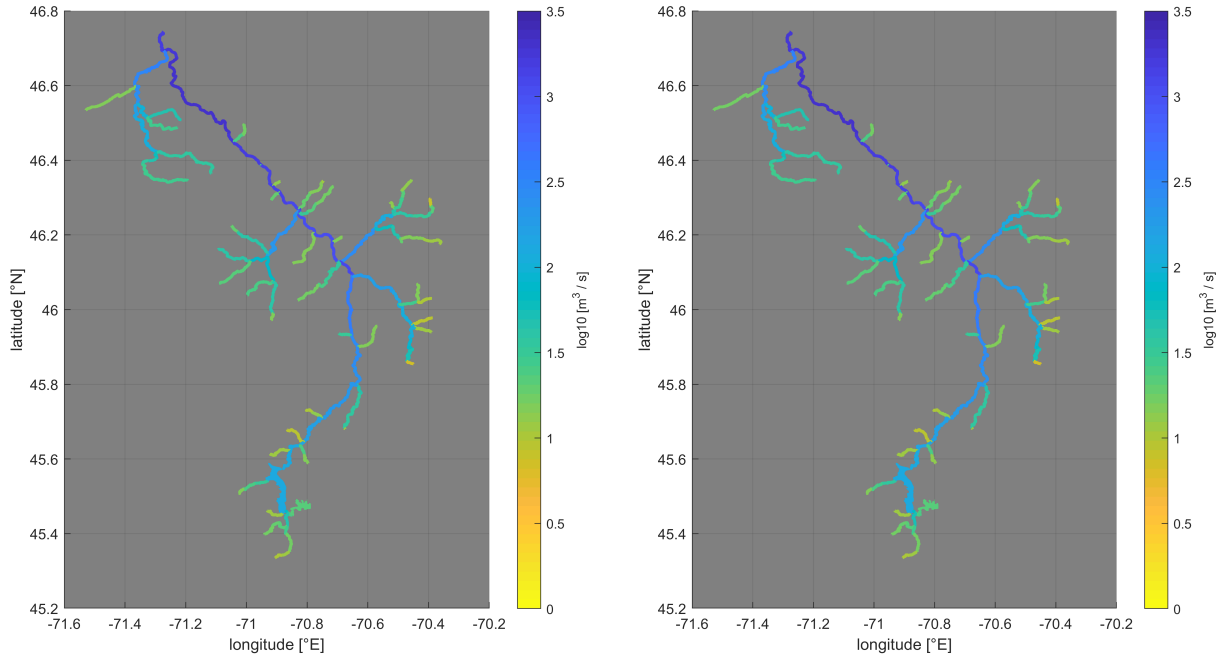


(a) Changement

(b) Signe du changement

FIGURE 8.1 Tendances annuelles des pseudo-observations normalisées par la surface de drainage, pour l'ensemble des tronçons de la rivière Chaudière. Elle vaut zéro si le modèle stationnaire est choisi. Les valeurs sont en unité m^3/(s $\times$ km$^2\times$ an).

La figure 8.1 présente les tendances annuelles des pseudo-observations normalisées par la surface de drainage, pour l'ensemble des tronçons de la rivière Chaudière. C'est une représentation des estimations du paramètre μ_1 du modèle statistique pour les pseudo-observations. On observe une cohérence spatiale, avec des tendances quasi-nulles en amont de la Chaudière et des tendances positives en aval. La section de rivière avec la tendance la plus marquée se trouve près de l'embouchure de la Chaudière, ce qui semble raisonnable. La carte des signes du changement montre en plus que très peu de tronçons affichent une tendance à la baisse sur la période historique.



(a) CanESM2/CRCM5-Ouranos

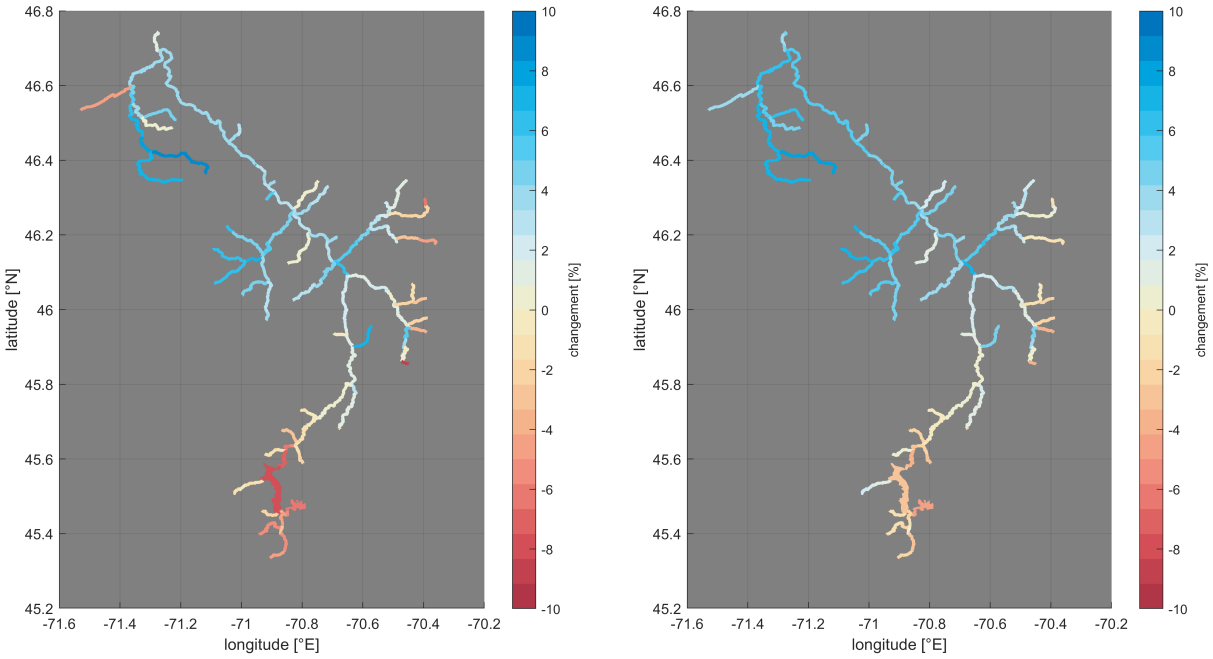
(b) IPSL-CM5A-LR

FIGURE 8.2 Débits centennaux en 2070 prédits par le modèle statistique complet, pour l'ensemble de la rivière Chaudière. Les valeurs affichées sont à l'échelle logarithmique (base 10).

La figure 8.2 présente une carte spatiale des niveaux de retour 100 ans projetés en 2070 par la chaîne de modélisation complète. On observe une cohérence spatiale des débits projetés, même quand les données n'ont pas été modélisées spatialement. Ceci pourrait s'expliquer par la qualité globale des données disponibles. Les sorties de modèles climatiques ont été post-traitées avant d'être utilisées par le modèle hydrologique Hydrotel, générant des débits simulés plus fidèles aux observations. La méthode d'interpolation des pseudo-débits intègre déjà d'une certaine manière la dimension spatiale, en utilisant les observations dans les stations les plus

proches. Les tronçons avec les forts débits correspondent à des branches principales de la rivière, avec un gradient croissant en direction de l'exutoire (coin en haut à gauche des cartes). Ce comportement est similaire pour les deux couples GCM-RCM considérés. À l'œil nu, les débits projetés sont très semblables pour CanEMS2/CRCM5-Ouranos et pour IPSL-CM5A-LR.

La figure 8.3 présente les changements relatifs en pourcentage des débits centennaux entre 2020 (dernière année de pseudo-observations historiques) et 2070, projetés par notre modèle. De nouveau ici, nous retrouvons une cohérence spatiale sur l'ensemble du bassin de la rivière Chaudière. Globalement, les tronçons en aval exhibent des tendances à la hausse (jusqu'à 10% de changement relatif), alors que les tronçons en amont exhibent des tendance à la baisse (jusqu'à -10% de changement relatif). En particulier pour la branche sud de la Chaudière au niveau du lac Mégantic (en bas des cartes), on constate une diminution significative des débits centennaux de 2020 à 2070. Cette diminution est plus marquée pour CanESM2/CRCM5-Ouranos que pour IPSL-CM5A-LR. On observe aussi des différences notables entre les deux couples GCM-RCM, où les changements projetés par IPSL-CM5A-LR semblent plus lisses spatialement que ceux projetés par CanESM2/CRCM5-Ouranos. Après discussion avec des hydrologues de la DEH, ceci pourrait s'expliquer par le fait que la résolution spatiale de IPSL-CM5A-LR est beaucoup plus grossière que celle de CanESM2/CRCM5-Ouranos, qui s'appuie sur un modèle régional du climat. Ainsi, contrairement à IPSL-CM5A-LR, le couple GCM-RCM CanESM2/CRCM5-Ouranos peut capturer des comportement hydrologiques locaux qui pourraient être à l'origine des hétérogénéités spatiales observées à la figure 8.3a.



(a) CanESM2/CRCM5-Ouranos

(b) IPSL-CM5A-LR

FIGURE 8.3 Changements relatifs (en %) des débits centennaux entre 2020 et 2070, pour l'ensemble de la rivière Chaudière.

CHAPITRE 9 DISCUSSION

La modélisation statistique des pseudo-observations a été une étape importante et ardue de ce projet. Ces données étaient difficiles à appréhender dans la mesure où elles correspondent à des distributions probabilistes d'erreur et non pas à des observations ponctuelles. Durant la phase de modélisation, plusieurs modèles statistiques ont été développés et testés, mais la plupart conduisaient à une surestimation des débits extrêmes, avec des valeurs observées irréalistes d'un point de vue hydrologique. Le modèle retenu parvient à réduire l'incertitude inhérente à la méthode d'interpolation utilisée pour générer les pseudo-observations, en combinant les six configurations hydrologiques et en incorporant une loi *a priori* informative sur le paramètre de forme de la loi GEV. C'est un des avantages majeurs de la méthodologie proposée. Nous rappelons enfin que les pseudo-observations fournies jusqu'à maintenant ne sont pas encore à leurs versions finales, et sont en cours d'amélioration par une autre équipe de recherche du projet INFO-Crue. Les modèles ont été développés en fonction de cette évolution et la méthodologie développée pourra s'appliquer directement à de nouvelles données produites par la DEH.

La prise en compte de la non-stationnarité des données historiques est un choix non conventionnel, car usuellement une non-stationnarité des séries de débits observés est rarement détectée. Pour les pseudo-observations fournies, les maxima annuels inférés des quatre dernières années (2017 à 2020) sont particulièrement élevés, ce qui peut être lié aux inondations exceptionnelles de 2017 et 2019 au Québec. Au lieu de censurer ces dernières observations, il nous a paru préférable d'exploiter toute l'information disponible et de modéliser une non-stationnarité globale sur la période historique. Après discussions avec les hydrologues et climatologues de l'équipe DEH-Ouranos, il s'avère qu'une tendance des débits pseudo-observés est possible mais doit être prise avec précaution. D'une part, une hétérogénéité temporelle peut exister pour certaines séries de débits, liée à l'ouverture de nouvelles stations de mesure. Pour les tronçons à très faible débit, la non-stationnarité constatée semble être des artifices plutôt que des changements hydrologiques, ce qui explique le choix de ne pas traiter les bassins dont la surface de drainage est inférieure à 50 km^2. D'autre part, une tendance historique, si elle existe, peut relever d'autres raisons que les changements climatiques, par exemple l'urbanisation du territoire. Néanmoins, en absence d'information supplémentaire, nous avons choisi la concentration de GES comme variable explicative pour modéliser la non-stationnarité, en étant conscient des limites de ce choix. Enfin, modéliser une non-stationnarité devient de plus en plus pertinente à mesure que la période historique s'allonge et qu'une potentielle tendance devient détectable dans l'avenir, ce qui constitue un argument en faveur de notre choix de

modélisation.

Une conséquence de la non-stationnarité des pseudo-observations est que la distribution des débits extrêmes diffère pour la première année et la dernière année de données historiques. La non-stationnarité et le choix de l'année de calibration pour la jonction des modèles ont une influence importante sur les valeurs des niveaux de retour projetés. Pour rappel, ce choix est nécessaire dans la mesure où notre modélisation est continue dans le temps, donc ne distingue pas une période de calibration historique et une période de projection. L'année de calibration est celle à partir de laquelle la distribution de débits est corrigée par la méthode CDF-*transform* pour suivre l'évolution simulée par les modèles climatiques. Comme les tendances historiques et simulées par les modèles climatiques sur la période historique peuvent différer, l'année de référence impacte directement les résultats obtenus, surtout les quantiles élevés qui y sont très sensibles, comme illustré dans la figure 9.1. On voit dans cette figure que pour le tronçon SLSO00003, si l'année de calibration est 1961 (première année de pseudo-observations), le débit centennal corrigé après la jonction des modèles n'augmente pas aussi fortement que celui inféré du modèle pour les pseudo-observations sur la période historique. Ainsi, le modèle qui simule la plus forte hausse du débit centennal prédit 1650 m^3/s en 2100, alors que cette valeur est plutôt 1900 m^3/s si l'année de référence est 2020 (voir figure 7.6a). Pour les analyses illustrées dans ce mémoire, 2020 a été choisie comme année de calibration, l'argument étant qu'on exploite l'information historique la plus récente, jugée plus fiable que celle produite par une simulation hydroclimatique pour la même date. Dans tous les cas, l'année de calibration peut être facilement changée par l'utilisateur du modèle, pour avoir le choix le plus adapté possible en fonction du tronçon et de l'application considérée dans l'analyse.

Dans ce projet, l'utilisation d'un ensemble de simulations climatiques permet de dresser un large éventail des futurs possibles. Elle est d'autant plus importante pour l'étude des extrêmes climatiques qui sont sujets à de grandes fluctuations, comme le montre la grande incertitude sur les débits centennaux inférés pour le tronçon SLSO00003 de la rivière Chaudière (voir la figure 7.6b). Le modèle hiérarchique bayésien pour les simulations permet de structurer les débits simulés provenant du même couple GCM-RCM et de quantifier l'incertitude sur les débits projetés par chaque membre hydroclimatique. La méthode utilisée attribue un poids égal à chaque membre du même couple GCM-RCM, définit comme la combinaison unique d'un membre de simulation, d'un scénario d'émission et d'une configuration hydrologique. La modélisation normale dans la couche latente du modèle pour les simulations est un choix simple qui peut être moins adapté lorsqu'un couple GCM-RCM suppose deux scénarios d'émission, ce qui peut engendrer une certaine hétérogénéité des paramètres GEV inférés pour chaque membre. Cette limite pourrait être surmontée avec un modèle à effets aléatoires plus raffiné,

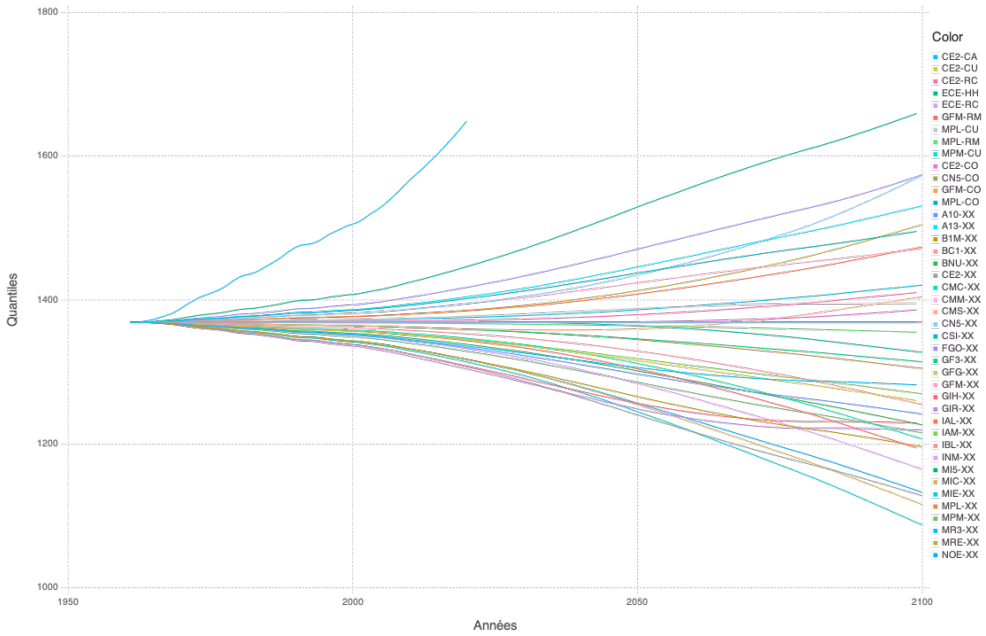


FIGURE 9.1 Pour le tronçon SLSO00003, estimation du niveau de retour 100 ans projeté jusqu'en 2100 pour l'ensemble des simulations hydroclimatiques. L'année de calibration est 1963. Chaque trait plein correspond à un couple GCM-RCM.

qui différencie les membres selon le scénario d'émission.

Enfin, la méthode de post-traitement bayésien utilisée pour inférer la loi prédictive des débits futurs parvient à transférer les incertitudes des pseudo-observations et des simulations aux débits futurs. Ainsi, nous parvenons à obtenir des estimations de débits extrêmes projetés avec leurs intervalles de crédibilité à 95%. Ces intervalles représentent l'incertitude sur les scénarios climatiques et configurations hydrologiques, l'incertitude d'interpolation des pseudo-observations ainsi que l'incertitude de modélisation statistique. Dans ce sens, nous effectuons bien une analyse fréquentielle *intégratrice* des débits extrêmes projetés.

CHAPITRE 10 CONCLUSION ET RECOMMANDATIONS

10.1 Synthèse des travaux

Dans ce projet, deux modèles statistiques bayésiens ont été développés et validés pour analyser les extrêmes de débits provenant d'un ensemble de simulations hydroclimatiques d'une part, des *pseudo-observations* de débits historiques d'autre part. Une méthode de post-traitement statistique a été adaptée pour faire la jonction des deux modèles précédents et faire l'analyse fréquentielle des débits projetés dans les différents bassins versants du Québec. La méthodologie proposée *intègre* les sources d'incertitudes sur toute la chaîne de modélisation. Elle produit à la fin des estimateurs accompagnés de leur intervalles de crédibilité à 95 %.

D'abord, la théorie des valeurs extrêmes, l'inférence bayésienne ainsi que les méthodes de simulation Monte-Carlo par chaînes de Markov ont été introduites pour poser le socle théorique des modèles statistiques développés. En s'inspirant des travaux déjà effectués dans la littérature scientifique, deux modèles bayésiens de valeurs extrêmes utilisant la loi GEV ont été développés. Pour les simulations, le modèle permet de modéliser l'évolution des débits simulés par les membres de chaque couple GCM-RCM de l'ensemble de simulations, utilisant des lois GEV non-stationnaires. L'ensemble des résultats rend compte des évolutions possibles des débits liées aux changements climatiques. Pour les pseudo-observations, le modèle développé combine l'information provenant des données et de la loi *a priori* pour réduire l'incertitude et estimer la série de maxima annuels historique la plus probable. Les lois *a posteriori* découlant de ces modèles ont été simulées par des algorithmes de simulation MCMC adaptifs, dont la convergence a été vérifiée sur les données de l'ensemble des tronçons de la rivière Chaudière. L'ajustement des modèles a été vérifié en examinant les QQ-plots et les représentations graphiques des lois *a posteriori*. La méthode CDF-*transform* décrite dans [16] a été adaptée au cadre bayésien pour lier les deux modèles précédemment décrits et estimer les débits extrêmes futurs. Elle réussit à transférer les évolutions identifiées dans le modèle pour les simulations aux pseudo-observations historiques, corriger le biais entre ces deux types de données et tenir compte de toutes les sources d'incertitude de la chaîne de modélisation.

La méthode d'analyse fréquentielle des pseudo-observations est novatrice et particulièrement adaptée aux données fournies par la DEH, qui sont des distributions d'erreur d'interpolation. Les débits projetés seront utilisés par la DEH pour calculer les risques de crue et tracer la carte des zones inondables.

10.2 Limitations et améliorations futures

La méthodologie développée dans ce projet comporte certaines limites. Le modèle pour les simulations peut devenir intensif computationnellement, si le nombre de membres considéré dépasse 300. Un modèle à effets aléatoires dans la couche latente qui incorpore des effets propres aux membres climatiques, aux scénarios d'émission et/ou aux configurations hydrologiques pourrait être intéressant à explorer. D'une part, il permettrait de quantifier l'influence intrinsèque à ces facteurs, d'autre part, il réduirait le nombre de paramètres à estimer. Par ailleurs, si on arrivait à considérer que les couples GCM-RCM de l'ensemble de simulations ne sont pas indépendants, le modèle hiérarchique développé peut être étendu à l'ensemble des modèles climatiques. Une telle approche, qui se rapproche des modèles décrits dans [41] et dans [54], peut être une piste à explorer dans le futur. Dans ce cas, la façon de prendre en compte le poids relatif accordé à chaque couple GCM-RCM devra être étudiée.

Pour l'analyse des pseudo-observations, la méthode actuellement utilisée pour choisir le débit maximal de chaque année est plutôt arbitraire, en prenant comme critère de sélection le maximum de la médiane des distributions d'erreur log-normales. Une analyse de sensibilité avec d'autres critères (le mode ou l'espérance des lois log-normales) pourrait être faite, pour évaluer l'influence de ce choix sur les résultats obtenus. Plus généralement, des méthodes alternatives pourraient être étudiées pour la sélection des maxima annuels. Par exemple, si des informations supplémentaires sur la corrélation temporelle des débits pseudo-observés étaient disponibles, un modèle de série temporelle pourrait être utilisé pour estimer ces maxima.

La méthode de post-traitement statistique utilisée dans ce projet nécessite le choix d'une année de calibration pour la jonction des modèles, ce qui impacte les résultats obtenus sur les niveaux de retour. Des améliorations de cette méthode pour contourner cette difficulté pourraient être étudiées dans le futur. D'autres méthodes de correction de biais pourraient aussi être explorées et leurs performances respectives comparées.

Finalement, la non-stationnarité des débits extrêmes a été modélisée en utilisant une relation linéaire entre les paramètres GEV et la concentration de gaz à effet de serre. D'autres choix de variables explicatives (émissions cumulatives, anomalie de température) et de relation de dépendance pourraient être explorés dans le futur, dans une analyse comparative, par exemple.

RÉFÉRENCES

[1] Gouvernement du Canada, "Les dix événements météorologiques les plus marquants au Canada en 2019," accédé le 10/02/2022. [En ligne]. Disponible : https://www.canada.ca/fr/environnement-changement-climatique/services/dix-evenements-meteorologiques-plus-marquants/2019.html

[2] B. Asadieh et N. Y. Krakauer, "Global change in streamflow extremes under climate change over the 21st century," *Hydrol. Earth Syst. Sci.*, vol. 21, n°. 11, p. 5863–5874, 2017. [En ligne]. Disponible : https://hess.copernicus.org/articles/21/5863/2017/

[3] D. Maraun, F. Wetterhall, A. M. Ireson, R. E. Chandler, E. J. Kendon, M. Widmann, S. Brienen, H. W. Rust, T. Sauter, M. Themeßl, V. K. C. Venema, K. P. Chun, C. M. Goodess, R. G. Jones, C. Onof, M. Vrac et I. Thiele-Eich, "Precipitation downscaling under climate change : Recent developments to bridge the gap between dynamical models and the end user," *Reviews of Geophysics*, vol. 48, n°. 3, 2010. [En ligne]. Disponible : https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2009RG000314

[4] C. Tebaldi et R. Knutti, "The use of the multi-model ensemble in probabilistic climate change projections," *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, vol. 365, p. 2053–75, 2007.

[5] S. Lachance-Cloutier, R. Turcotte et J.-F. Cyr, "Combining streamflow observations and hydrologic simulations for the retrospective estimation of daily streamflow for ungauged rivers in southern Quebec (Canada)," *Journal of Hydrology*, vol. 550, p. 294–306, 2017. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0022169417302986

[6] S. Coles, *An introduction to statistical modeling of extreme values*. Springer, 2001.

[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°. 2, p. 180–190, 1928.

[8] R. W. Katz, M. B. Parlange et P. Naveau, "Statistics of extremes in hydrology," *Advances in Water Resources*, vol. 25, n°. 8, p. 1287–1304, 2002. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0309170802000568

[9] C. Robert, *Le choix bayésien : Principes et pratique*. Springer Science et Business Media, 2005.

[10] C. Robert et G. Casella, *Monte Carlo statistical methods*. Springer, 2004.

[11] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari et D. B. Rubin, *Bayesian data analysis*. CRC press, 2013.

[12] J. S. Rosenthal, "Optimal proposal distributions and adaptive MCMC," *Handbook of Markov Chain Monte Carlo*, vol. 4, n°. 10.1201, 2011.

[13] S. G. Coles et J. A. Tawn, "A Bayesian analysis of extreme rainfall data," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 45, n°. 4, p. 463–478, 1996. [En ligne]. Disponible : http://www.jstor.org/stable/2986068

[14] D. Maraun, "Bias correcting climate change simulations - a critical review," *Current Climate Change Reports*, vol. 2, n°. 4, p. 211–220, 2016. [En ligne]. Disponible : https://doi.org/10.1007/s40641-016-0050-x

[15] P.-A. Michelangeli, M. Vrac et H. Loukos, "Probabilistic downscaling approaches : Application to wind cumulative distribution functions," *Geophysical Research Letters*, vol. 36, n°. 11, 2009. [En ligne]. Disponible : https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2009GL038401

[16] M. Kallache, M. Vrac, P. Naveau et P.-A. Michelangeli, "Nonstationary probabilistic downscaling of extreme precipitation," *Journal of Geophysical Research : Atmospheres*, vol. 116, n°. D5, 2011. [En ligne]. Disponible : https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2010JD014892

[17] J. R. M. Hosking et J. R. Wallis, *Regional Frequency Analysis : An Approach Based on L-Moments*. Cambridge University Press, 1997.

[18] C. H. R. Lima, U. Lall, T. Troy et N. Devineni, "A hierarchical Bayesian GEV model for improving local and regional flood quantile estimates," *Journal of Hydrology*, vol. 541, p. 816–823, 2016. [En ligne]. Disponible : http://dx.doi.org/10.1016/j.jhydrol.2016.07.042

[19] M. Piras, G. Mascaro, R. Deidda et E. R. Vivoni, "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.

[20] S. G. Coles et E. A. Powell, "Bayesian methods in extreme value modelling : A review and new developments," *International Statistical Review*, vol. 64, n°. 1, p. 119–136, 1996. [En ligne]. Disponible : http://www.jstor.org/stable/1403426

[21] D. Cooley, P. Naveau, V. Jomelli, A. Rabatel et D. Grancher, "A Bayesian hierarchical extreme value model for lichenometry," *Environmetrics*, vol. 17, n°. 6, p. 555–574, 2006. [En ligne]. Disponible : https://onlinelibrary.wiley.com/doi/abs/10.1002/env.764

[22] Y.-b. Wu, L.-q. Xue et Y.-h. Liu, "Local and regional flood frequency analysis based on hierarchical bayesian model in Dongting lake basin, China," *Water*

Science and Engineering, vol. 12, n°. 4, p. 253–262, 2019. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S167423701930122X

[23] T. L. Thorarinsdottir, K. H. Hellton, G. H. Steinbakk, L. Schlichting et K. Engeland, "Bayesian regional flood frequency analysis for large catchments," *Water Resources Research*, vol. 54, n°. 9, p. 6929–6947, 2018. [En ligne]. Disponible : https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2017WR022460

[24] A. C. Davison, S. A. Padoan et M. Ribatet, "Statistical Modeling of Spatial Extremes," *Statistical Science*, vol. 27, n°. 2, p. 161–186, 2012. [En ligne]. Disponible : https://doi.org/10.1214/11-STS376

[25] J. D. Salas, J. Obeysekera et R. M. Vogel, "Techniques for assessing water infrastructure for nonstationary extreme events : a review," *Hydrological Sciences Journal*, vol. 63, n°. 3, p. 325–352, 2018. [En ligne]. Disponible : https://doi.org/10.1080/02626667.2018.1426858

[26] R. W. Katz, "Statistics of extremes in climate change," *Climatic Change*, vol. 100, n°. 1, p. 71–76, 2010. [En ligne]. Disponible : https://doi.org/10.1007/s10584-010-9834-5

[27] L. Cheng, A. AghaKouchak, E. Gilleland et R. W. Katz, "Non-stationary extreme value analysis in a changing climate," *Climatic Change*, vol. 127, n°. 2, p. 353–369, 2014. [En ligne]. Disponible : https://doi.org/10.1007/s10584-014-1254-5

[28] S. Wi, J. B. Valdes, S. Steinschneider et T. W. Kim, "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, n°. 2, p. 583–606, 2016.

[29] Y. Tramblay, L. Neppel, J. Carreau et K. Najib, "Non-stationary frequency analysis of heavy rainfall events in southern France," *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*, vol. 58, n°. 2, p. 280–294, 2013.

[30] M. Sraj, A. Viglione, J. Parajka et G. Bloschl, "The influence of non-stationarity in extreme hydrological events on flood frequency estimation," *Journal of Hydrology and Hydromechanics*, vol. 64, n°. 4, p. 426–437, 2016.

[31] J. Das et N. V. Umamahesh, "Uncertainty and nonstationarity in streamflow extremes under climate change scenarios over a river basin," *Journal of Hydrologic Engineering*, vol. 22, n°. 10, p. 04017042, 2017.

[32] B. J. McAvaney, C. Covey, S. Joussaume, V. Kattsov, A. Kitoh, W. Ogana, A. Pitman, A. Weaver, R. Wood et Z.-C. Zhao, "Model evaluation," dans *Climate Change 2001 : The scientific basis. Contribution of WG1 to the Third Assessment Report of the IPCC (TAR)*. Cambridge University Press, 2001, p. 471–523.

[33] R. Knutti, R. Furrer, C. Tebaldi, J. Cermak et G. A. Meehl, "Challenges in combining projections from multiple climate models," *Journal of Climate*, vol. 23, n°. 10, p. 2739–2758, 2010.

[34] R. W. Katz, "Techniques for estimating uncertainty in climate change scenarios and impact studies," *Climate Research*, vol. 20, n°. 2, p. 167–185, 2002. [En ligne]. Disponible : https://www.int-res.com/abstracts/cr/v20/n2/p167-185/

[35] S. Yip, C. A. T. Ferro, D. B. Stephenson et E. Hawkins, "A simple, coherent framework for partitioning uncertainty in climate predictions," *Journal of Climate*, vol. 24, n°. 17, p. 4634–4643, 2011. [En ligne]. Disponible : https://journals.ametsoc.org/view/journals/clim/24/17/2011jcli4085.1.xml

[36] I. Giuntoli, G. Villarini, C. Prudhomme et D. M. Hannah, "Uncertainties in projected runoff over the conterminous United States," *Climatic Change*, vol. 150, n°. 3-4, p. 149–162, 2018.

[37] L. Zhang, F. Yuan, B. Wang, L. Ren, C. Zhao, J. Shi, Y. Liu, S. Jiang, X. Yang, T. Chen et S. Liu, "Quantifying uncertainty sources in extreme flow projections for three watersheds with different climate features in China," *Atmospheric Research*, vol. 249, 2021. [En ligne]. Disponible : http://dx.doi.org/10.1016/j.atmosres.2020.105331

[38] F. Giorgi et L. O. Mearns, "Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the "Reliability Ensemble Averaging" (REA) method," *Journal of Climate*, vol. 15, n°. 10, p. 1141–1158, 2002. [En ligne]. Disponible : https://journals.ametsoc.org/view/journals/clim/15/10/1520-0442_2002_015_1141_coaura_2.0.co_2.xml

[39] N. Le Vine, "Combining information from multiple flood projections in a hierarchical bayesian framework," *Water Resources Research*, vol. 52, n°. 4, p. 3258–3275, 2016. [En ligne]. Disponible : http://dx.doi.org/10.1002/2015WR018143

[40] C. Tebaldi, R. L. Smith, D. Nychka et L. O. Mearns, "Quantifying uncertainty in projections of regional climate change : A bayesian approach to the analysis of multimodel ensembles," *Journal of Climate*, vol. 18, n°. 10, p. 1524–1540, 2005. [En ligne]. Disponible : https://journals.ametsoc.org/view/journals/clim/18/10/jcli3363.1.xml

[41] I. Giuntoli, I. Prosdocimi et D. M. Hannah, "Going beyond the ensemble mean : Assessment of future floods from global multi-models," *Water Resources Research*, vol. 57, n°. 3, 2021. [En ligne]. Disponible : http://dx.doi.org/10.1029/2020WR027897

[42] H. A. Panofsky et G. W. Brier, *Some applications of statistics to meteorology*. Mineral Industries Extension Services, College of Mineral Industries, Pennsylvania State University, 1958.

[43] E. M. Laflamme, E. Linder et Y. Pan, “Statistical downscaling of regional climate model output to achieve projections of precipitation extremes,” *Weather and Climate Extremes*, vol. 12, p. 15–23, 2016.

[44] R. Towe, E. Eastoe, J. Tawn et P. Jonathan, “Statistical downscaling for future extreme wave heights in the north sea,” *The Annals of Applied Statistics*, vol. 11, n°. 4, p. 2375–2403, 2017.

[45] F. Giorgi, C. Jones et G. Asrar, “Addressing climate information needs at the regional level : The CORDEX framework,” *WMO Bull*, vol. 53, 11 2008.

[46] K. E. Taylor, R. J. Stouffer et G. A. Meehl, “An Overview of CMIP5 and the Experiment Design,” *Bulletin of the American Meteorological Society*, vol. 93, n°. 4, p. 485–498, 2012. [En ligne]. Disponible : https://journals.ametsoc.org/view/journals/bams/93/4/bams-d-11-00094.1.xml

[47] M. Leduc, A. Mailhot, A. Frigon, J.-L. Martel, R. Ludwig, G. B. Brietzke, M. Giguère, F. Brissette, R. Turcotte, M. Braun et J. Scinocca, “The ClimEx Project : A 50-Member Ensemble of Climate Change Projections at 12-km Resolution over Europe and Northeastern North America with the Canadian Regional Climate Model (CRCM5),” *Journal of Applied Meteorology and Climatology*, vol. 58, n°. 4, p. 663–693, 2019. [En ligne]. Disponible : https://journals.ametsoc.org/view/journals/apme/58/4/jamc-d-18-0021.1.xml

[48] E. S. Martins et J. R. Stedinger, “Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data,” *Water Resources Research*, vol. 36, n°. 3, p. 737–744, 2000. [En ligne]. Disponible : https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/1999WR900330

[49] P. J. Northrop et N. Attalides, “Posterior propriety in Bayesian extreme value analyses using reference priors,” *Statistica Sinica*, p. 721–743, 2016.

[50] D. J. Spiegelhalter, N. G. Best, B. P. Carlin et A. Van Der Linde, “Bayesian measures of model complexity and fit,” *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, vol. 64, n°. 4, p. 583–639, 2002.

[51] A. Gelman, “Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper),” *Bayesian Analysis*, vol. 1, n°. 3, p. 515–534, 2006.

[52] C. A. Love, B. E. Skahill, J. F. England, G. Karlovits, A. Duren et A. AghaKouchak, “Integrating climatic and physical information in a bayesian hierarchical model of extreme daily precipitation,” *Water*, vol. 12, n°. 8, p. 2211, 2020. [En ligne]. Disponible : https://www.mdpi.com/2073-4441/12/8/2211

[53] J. A. Garcia, J. Martin, L. Naranjo et F. J. Acero, “A bayesian hierarchical spatio-temporal model for extreme rainfall in Extremadura (Spain),” *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*, vol. 63, n°. 6, p. 878–894, 2018.

[54] P. J. Northrop et R. E. Chandler, “Quantifying sources of uncertainty in projections of future climate,” *Journal of Climate*, vol. 27, n°. 23, p. 8793–8808, 2014. [En ligne]. Disponible : https://journals.ametsoc.org/view/journals/clim/27/23/jcli-d-14-00265.1.xml

ANNEXE A PREUVE DE LA PROPRIÉTÉ DE LA LOI A POSTERIORI DU MODÈLE POUR LES SIMULATIONS

Loi a posteriori avec la loi a priori du modèle pour les simulations

Nous démontrons que la loi *a posteriori* dans le modèle statistique pour les simulations est propre, lorsque la loi *a priori* est uniforme sur les hyperparamètres ν_k et inverse-gamma sur les hyperparamètres τ_k^2. La preuve est donnée pour la loi GEV stationnaire. Les deux autres cas peuvent être démontrés de façon similaire. Le modèle s'écrit :

$$\begin{aligned}
X_{ij} &\sim \mathcal{G}EV(\mu_i, \exp(\phi_i), \xi_i), \quad i = 1, \ldots, m; j = 1, \ldots, n \\
\mu_i &\sim \mathcal{N}(\nu_0, \tau_0^2), \quad i = 1, \ldots, m \\
\phi_i &\sim \mathcal{N}(\nu_1, \tau_1^2), \quad i = 1, \ldots, m \\
\xi_i &\sim \mathcal{N}(\nu_2, \tau_2^2), \quad i = 1, \ldots, m \\
(\nu_k, \tau_k^2) &\sim \mathcal{I}nvGamma(\tau_k^2|\alpha_k, \beta_k), \quad \alpha_k > 0, \beta_k > 0, k = 0, \ldots, 2.
\end{aligned}$$

La loi *a priori* ici est impropre car $\nu_k \propto 1$.

La loi *a posteriori* s'écrit alors :

$$f_{((\mu_i,\phi_i,\xi_i)_{1\leq i\leq m},(\nu_k,\tau_k^2)_{0\leq k\leq 2}|\boldsymbol{X}=\boldsymbol{x})}((\mu_i,\phi_i,\xi_i)_{1\leq i\leq m},(\nu_k,\tau_k^2)_{0\leq k\leq 2}) =$$
$$\prod_{i=1}^{m}\prod_{j=1}^{n}\mathcal{G}EV(x_{ij}|\mu_i,\exp(\phi_i),\xi_i)\prod_{i=1}^{m}\mathcal{N}(\mu_i|\nu_0,\tau_0^2)\mathcal{N}(\phi_i|\nu_1,\tau_1^2)\mathcal{N}(\xi_i|\nu_2,\tau_2^2)\prod_{k=0}^{2}\mathcal{I}nvGamma(\tau_k^2|\alpha_k,\beta_k)$$

Un résultat de [49] dit que pour l'inférence bayésienne d'une loi GEV, la loi *a priori* uniforme sur (μ, ϕ, ξ) où $\phi = \log\sigma$ rend la loi *a posteriori* propre si le nombre d'observations $n \geq 4$. Un corollaire est qu'une loi *a priori* sur (μ, ϕ, ξ) bornée est suffisante pour que la loi *a posteriori* soit propre. Appliqué à notre cas, cela revient à montrer que la loi $\pi((\mu_i, \phi_i, \xi_i)_{i=1,\ldots,m})$ est bornée, où $\pi((\mu_i, \phi_i, \xi_i)_{i=1,\ldots,m})$ est égal à :

$$\int \cdots \int \left[\prod_{i=1}^{m} \mathcal{N}(\mu_i|\nu_0, \tau_0^2)\mathcal{N}(\phi_i|\nu_1, \tau_1^2)\mathcal{N}(\xi_i|\nu_2, \tau_2^2) \prod_{k=0}^{2} \mathcal{I}nvGamma(\tau_k^2|\alpha_k, \beta_k) \right] d\nu_k \, d\tau_k^2 \tag{A.1}$$

$$= \iint \prod_{i=1}^{m} \mathcal{N}(\mu_i|\nu_0, \tau_0^2)\mathcal{I}nvGamma(\tau_0^2|\alpha_0, \beta_0) \, d\nu_0 \, d\tau_0^2 \tag{A.2}$$

$$\times \iint \prod_{i=1}^{m} \mathcal{N}(\phi_i|\nu_1, \tau_1^2)\mathcal{I}nvGamma(\tau_1^2|\alpha_1, \beta_1) \, d\nu_1 \, d\tau_1^2 \tag{A.3}$$

$$\times \iint \prod_{i=1}^{m} \mathcal{N}(\xi_i|\nu_2, \tau_2^2)\mathcal{I}nvGamma(\tau_2^2|\alpha_2, \beta_2) \, d\nu_2 \, d\tau_2^2 \tag{A.4}$$

Nous faisons le calcul pour l'intégrale (A.2). Le calcul des intégrales (A.3) et (A.4) se fait de la même manière. Pour plus de lisibilité, nous omettons l'indice 0 pour $\nu_0, \tau_0^2, \alpha_0, \beta_0$. Comme nous voulons borner cette intégrale, le signe $\propto$ est utilisé pour signifier l'omission des termes constants qui ne dépendent pas de ν_0, τ_0^2 et des $\{\mu_i : i = 1, \ldots, m\}$.

Le terme à l'intérieur de l'intégrale (A.2) se réécrit

$$\begin{aligned} &(2\pi\tau^2)^{-m/2} \exp\left(-\frac{1}{\tau^2} \sum_{i=1}^{m} \frac{(\mu_i - \nu)^2}{2} \right) (\tau^2)^{-\alpha-1} \exp\left(-\frac{\beta}{\tau^2} \right) \\ \propto &(\tau^2)^{-\frac{m}{2}-\alpha-1} \exp\left[-\frac{1}{\tau^2} \left(\sum_{i=1}^{m} \frac{(\mu_i - \nu)^2}{2} + \beta \right) \right] \end{aligned}$$

qui est proportionnel à une densité $\mathcal{I}nverseGamma(\tau^2|\alpha', \beta')$ où $\alpha' = \frac{m}{2} + \alpha$ et $\beta' = \sum_{i=1}^{m} \frac{(\mu_i - \nu)^2}{2} + \beta$. En intégrant sur τ^2, on obtient

$$\Gamma\left(\frac{m}{2} + \alpha \right) \left(\sum_{i=1}^{m} \frac{(\mu_i - \nu)^2}{2} + \beta \right)^{-m/2-\alpha} \propto \left(\sum_{i=1}^{m} \frac{(\mu_i - \nu)^2}{2} + \beta \right)^{-m/2-\alpha} \tag{A.5}$$

Un peu de calcul donne

$$\sum_{i=1}^{m} \frac{(\mu_i - \nu)^2}{2} = \frac{m}{2} \left[(\nu - \bar{\mu})^2 + \bar{\mu}_2 - \bar{\mu}^2 \right] \quad \text{où } \bar{\mu}_2 = \frac{1}{m} \sum_{i=1}^{m} \mu_i^2 \text{ et } \bar{\mu} = \frac{1}{m} \sum_{i=1}^{m} \mu_i$$

.

La quantité (A.5) devient

$$\begin{aligned}&\left[\frac{m}{2}(\nu-\bar{\mu})^2+\frac{m}{2}(\bar{\mu}_2-\bar{\mu}^2)+\beta\right]^{-m/2-\alpha}\\=&\left[\frac{m}{2}(\bar{\mu}_2-\bar{\mu}^2)+\beta\right]^{-m/2-\alpha}\left[1+\frac{(\nu-\bar{\mu})^2}{(\bar{\mu}_2-\bar{\mu}^2)+\frac{2\beta}{m}}\right]^{-m/2-\alpha}\end{aligned}$$

qui est proportionnelle à une densité de Student généralisée :

$$f(\nu|\gamma,\mu,\sigma)=\frac{\Gamma(\frac{\gamma+1}{2})}{\Gamma(\frac{\gamma}{2})\sqrt{\pi\gamma}\sigma}\left[1+\frac{1}{\gamma}\left(\frac{\nu-\mu}{\sigma}\right)^2\right]^{-\frac{\gamma+1}{2}}$$

où $\mu=\bar{\mu},\gamma=m+2\alpha-1,\sigma=\sqrt{\frac{(\bar{\mu}_2-\bar{\mu}^2)+\frac{2\beta}{m}}{m+2\alpha-1}}$ qui est bien défini par inégalité de Jensen, dès que $m\geq 1$.

En intégrant par rapport à ν, on a que l'intégrale (A.2) est proportionnelle à

$$\begin{aligned}&\left(\frac{m}{2}(\bar{\mu}_2-\bar{\mu}^2)+\beta\right)^{-m/2-\alpha}\frac{\Gamma(\frac{\gamma}{2})\sqrt{\pi\gamma}}{\Gamma(\frac{\gamma+1}{2})}\sqrt{\frac{(\bar{\mu}_2-\bar{\mu}^2)+\frac{2\beta}{m}}{m+2\alpha-1}}\\=&\frac{\Gamma(\frac{\gamma}{2})\sqrt{\pi\gamma}}{\Gamma(\frac{\gamma+1}{2})}\left(\frac{m}{2}\right)^{-m/2-\alpha}\left[(\bar{\mu}_2-\bar{\mu}^2)+\frac{2\beta}{m}\right]^{-m/2-\alpha}\sqrt{\frac{(\bar{\mu}_2-\bar{\mu}^2)+\frac{2\beta}{m}}{m+2\alpha-1}}\\\propto&\left[(\bar{\mu}_2-\bar{\mu}^2)+\frac{2\beta}{m}\right]^{-m/2-\alpha+1/2}\leq\left(\frac{2\beta}{m}\right)^{-m/2-\alpha+1/2}\quad\text{dès que } m\geq 1\end{aligned}$$

Comme $\frac{2\beta}{m}>0$, nous avons réussi à borner l'intégrale (A.2), et par conséquence la loi *a priori* $\pi(\mu_i,\phi_i,\xi_{i,i=1,\ldots,m})$.

Loi a posteriori avec la loi a priori de Jeffreys

Le cas limite quand $\alpha_k\to 0$ et $\beta_k\to 0$ correspond à la loi *a priori* de Jeffreys. Remarquons que dans le cas limite $\alpha\to 0$ et $\beta\to 0$, l'intégrale (A.2) est proportionnelle à

$$(\bar{\mu}_2-\bar{\mu}^2)^{-m/2+1/2}$$

et la loi *a priori* $\pi((\mu_i,\phi_i,\xi_i)_{i=1,\ldots,m})$ met une masse infini sur les points où $\bar{\mu}_2-\bar{\mu}^2=0$. Or la vraisemblance GEV à ces points est non nulle. Ainsi, la loi *a posteriori* est forcément impropre, si la loi de Jeffreys est utilisée pour les hyperparamètres.

ANNEXE B ENSEMBLE DE SIMULATIONS CLIMATIQUES

TABLEAU B.1 Récapitulatif de l'ensemble des simulations climatiques. Les colonnes sont respectivement : la codification du GCM, la codification du RCM (XXX signifie l'absence de RCM), le modèle choisi pour les simulations pour le tronçon SLSO00003 (1 : stationnaire, 2 : non-stationnaire à 4 paramètres, 3 : non-stationnaire complet), le nombre de membres, le nombre de scénarios d'émission utilisés (1 : RCP8.5, 2 : RCP4.5 et RCP8.5).

	GCM	RCM	modèle	m	scén.
1	CE2	CO	3	280	2
2	CSI	XX	2	100	2
3	CE2	XX	2	50	2
4	IAL	XX	3	40	2
5	MI5	XX	1	30	2
6	MPL	XX	2	30	2
7	CE2	CA	2	20	2
8	GF3	XX	2	20	2
9	MPM	XX	2	20	2
10	CN5	CO	1	10	2
11	GFM	CO	1	10	2
12	MPL	CO	2	10	2
13	CE2	CU	2	10	2
14	MPM	CU	1	10	2
15	ECE	HH	3	10	2
16	CE2	RC	2	10	2
17	ECE	RC	1	10	1
18	GFM	RM	3	10	1
19	MPL	RM	2	10	1
20	A10	XX	2	10	2
21	A13	XX	2	10	2

	GCM	RCM	modèle	m	scén.
22	B1M	XX	2	10	2
23	BC1	XX	3	10	2
24	BNU	XX	2	10	2
25	CMM	XX	1	10	2
26	CMS	XX	3	10	2
27	CN5	XX	3	10	2
28	FGO	XX	3	10	2
29	GFG	XX	1	10	2
30	GFM	XX	3	10	2
31	GIR	XX	2	10	1
32	IAM	XX	2	10	2
33	IBL	XX	1	10	2
34	INM	XX	2	10	2
35	MIC	XX	3	10	2
36	MIE	XX	2	10	2
37	MR3	XX	2	10	2
38	NOE	XX	2	10	2
39	MPL	CU	3	5	1
40	CMC	XX	3	5	1
41	GIH	XX	2	5	1
42	MRE	XX	3	5	1

Tableau B.2 Codification GCM-RCM.

	GCM	RCM	Nom GCM	Nom RCM
1	CE2	CA	CanESM2	CanRCM4
2	CE2	CU	CanESM2	CRCM5-UQAM
3	CE2	RC	CanESM2	RCA4
4	ECE	HH	EC-EARTH	HIRHAM5
5	ECE	RC	EC-EARTH	RCA4
6	GFM	RM	GFDL-ESM2M	RegCM4
7	MPL	CU	MPI-ESM-LR	CRCM5-UQAM
8	MPL	RM	MPI-ESM-LR	RegCM4
9	MPM	CU	MPI-ESM-MR	CRCM5-UQAM
10	CE2	CO	CanESM2	CRCM5-Ouranos
11	CN5	CO	CNRM-CM5	CRCM5-Ouranos
12	GFM	CO	GFDL-ESM2M	CRCM5-Ouranos
13	MPL	CO	MPI-ESM-LR	CRCM5-Ouranos
14	A10	XX	ACCESS1-0	-
15	A13	XX	ACCESS1-3	-
16	B1M	XX	BCC-CSM1-1-m	-
17	BC1	XX	BCC-CSM1-1	-
18	BNU	XX	BNU-ESM	-
19	CE2	XX	CanESM2	-
20	CMC	XX	CMCC-CESM	-
21	CMM	XX	CMCC-CM	-
22	CMS	XX	CMCC-CMS	-
23	CN5	XX	CNRM-CM5	-
24	CSI	XX	CSIRO-Mk3-6-0	-
25	FGO	XX	FGOALS-g2	-
26	GF3	XX	GFDL-CM3	-
27	GFG	XX	GFDL-ESM2G	-
28	GFM	XX	GFDL-ESM2M	-
29	GIH	XX	GISS-E2-H	-
30	GIR	XX	GISS-E2-R_r6i1p1	-
31	IAL	XX	IPSL-CM5A-LR	-
32	IAM	XX	IPSL-CM5A-MR	-
33	IBL	XX	IPSL-CM5B-LR	-
34	INM	XX	INMCM4	-
35	MI5	XX	MIROC5	-
36	MIC	XX	MIROC-ESM-CHEM	-
37	MIE	XX	MIROC-ESM	-
38	MPL	XX	MPI-ESM-LR	-
39	MPM	XX	MPI-ESM-MR	-
40	MR3	XX	MRI-CGCM3	-
41	MRE	XX	MRI-ESM1	-
42	NOE	XX	NorESM1-M	-