POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Modélisation statistique de l'intensité, de la durée et de la fréquence des vagues de chaleur simulées par les modèles numériques de climat

PHILÉMON GAMET
Département de mathématiques et de génie industriel

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

Février 2021

© Philémon Gamet, 2021.

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Modélisation statistique de l'intensité, de la durée et de la fréquence des vagues de chaleur simulées par les modèles numériques de climat

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

Charles AUDET, président
Jonathan JALBERT, membre et directeur de recherche
Debbie J. DUPUIS, membre

DÉDICACE

A toutes les personnes qui m'ont
laissé de ces années au Canada
de si bons souvenirs ...

REMERCIEMENTS

Je tiens à remercier chaleureusement par ces quelques mots mon directeur de recherche, Jonathan Jalbert, qui m'a apporté une aide précieuse au cours de ces deux années passées sous sa supervision. Ses conseils avisés et sa bienveillance continue m'auront permis de tirer le meilleur de cette expérience et de mener à bien ce projet.

Je voudrais également remercier Éloïse et Gabriel que j'ai eu la chance de côtoyer au cours de ma maîtrise, et avec qui il a toujours été intéressant d'échanger, que ce soit sur des sujets scientifiques ou non.

Je remercie Philippe qui a été un acteur assidu des réunions de groupe hebdomadaires, et dont l'expertise en *Julia* a été par moments très précieuse. Son expérience sur les questions climatiques ainsi que ses relectures auront aussi été d'une grande aide.

Enfin, je remercie toutes les personnes que j'ai eu la chance de croiser ici à Montréal et qui auront participé, d'une manière ou d'une autre, à l'avancement de cette maîtrise.

RÉSUMÉ

Les vagues de chaleur sont considérées comme les catastrophes naturelles les plus mortelles dans le monde et ont un impact sérieux sur l'ensemble de notre société : économie, santé, infrastructures entre autres. D'après le dernier rapport du GIEC, elles seront plus intenses et plus fréquentes dans les prochaines années en raison des changements climatiques. Il est donc essentiel d'estimer le risque lié aux vagues de chaleur auquel nous seront exposés afin de mettre en place dès aujourd'hui des mesures d'adaptation adéquates : normes de construction des bâtiments, plans d'alerte canicule... Le but de ce projet de recherche consiste à modéliser l'évolution de la fréquence d'occurrence, de l'intensité et de la durée des vagues de chaleur.

Les ensembles de simulations climatiques permettent de faire des projections quantitatives sur l'évolution du climat et nous nous baserons sur ces derniers pour analyser les vagues de chaleur jusqu'à l'horizon 2100. La fréquence d'occurrence sera modélisée à l'aide d'un processus de Poisson non-homogène, la température maximale atteinte lors des vagues de chaleur sera modélisée par une loi de Pareto généralisée étendue et enfin la durée des vagues de chaleur sera modéliser de manière conditionnelle à l'intensité. L'inclusion des concentrations annuelles de gaz à effet de serre permettra d'obtenir un modèle statistique non-stationnaire pour les vagues de chaleur. L'approche bayésienne utilisée pour l'inférence des paramètres du modèle permet d'obtenir des intervalles de crédibilité pour l'ensemble des estimations réalisées.

En se basant sur l'ensemble de simulation ClimEx, l'évolution des niveaux de retour bidimensionnels du couple durée-intensité des vagues de chaleur est déterminée au point de grille de la ville de Montréal. A titre d'exemple, la période de retour des vagues de chaleur de plus d'une semaine atteignant une température maximale de 42.2°C, qui est de 100 ans en 2020, est réduite à 45 ans en 2030 et à un peu plus de 7 ans en 2050. Enfin, les régions du Québec et de l'Europe les plus touchées par les changements climatiques sont déterminées en se basant sur l'ensemble CESM. Il est notable que toutes les régions étudiées présentent une augmentation du nombre de vagues de chaleur extrêmes dans les prochaines années. Il apparaît que le sud-est du Québec, sous influence océanique est la région la plus exposée, tandis qu'en Europe, la fréquence d'occurrence des vagues de chaleur les plus extrêmes devrait particulièrement augmenter dans le bassin méditerranéen ainsi qu'en Europe de l'est.

ABSTRACT

Heatwaves have tremendous impact on society. Considered the deadliest natural disasters, they also affect economy, energy supply, infrastructures, etc. According to the IPCC latest assessment report, both their intensity and their frequency is expected to increase in the next decades due to climate change. As a consequence, it is essential to properly estimate the risk they induce in the near future, in order to design, starting today, adaptation and mitigation measures. The aim of this research is to provide policymakers an accurate description of heatwave evolution in terms of frequency, intensity and duration in the years to come and ultimately increase the resilience of our society in the face of climate change.

Climate simulations are the only available tools to make quantitative projections of the future climate. Evolutions of the heatwaves are determined based on those climate simulations up to 2100 under the RCP8.5 emission scenario assumption. The occurence frequency of heatwaves will be modeled with a non-homogeneous Poisson process, their intensity will be modeled using an extended generalized Pareto distribution and the duration of heatwaves will be modeled conditionnaly to the intensity with a negative binomial distribution. Atmospheric carbon concentration will be included as a covariate in order to build a transient model for heatwaves, allowing a study of climate change. Finally, the Bayesian approach will allow to build credibility intervals for each of our estimates.

The ClimEx simulation ensemble will be used to assess the evolution of bidimentional effective return levels of the intensity-duration pair at the location of Montreal. It was found that the return period of heatwaves reaching 42.2°C and during more than one week, which is of 100 years in 2020, will be reduced to 45 years by 2030 and will fall to 7 years by 2050. Then, the CESM ensemble was used to find the regions the most affected by climate change, both in the North-East quarter of Northern America and in Europe. It is noteworthy that all studied regions show an increase in the frequency of extreme heatwaves, but Southern Quebec under oceanic influence is the most exposed area for the American region while the mediterranean basin and Eastern Europe are the most exposed areas in Europe.

TABLE DES MATIÈRES

DÉDICACE iii

REMERCIEMENTS iv

RÉSUMÉ v

ABSTRACT vi

TABLE DES MATIÈRES vii

LISTE DES TABLEAUX x

LISTE DES FIGURES xi

LISTE DES SIGLES ET ABRÉVIATIONS xiv

CHAPITRE 1 INTRODUCTION 1

CHAPITRE 2 REVUE CRITIQUE DE LA LITTÉRATURE 3

CHAPITRE 3 DÉMARCHE DE L'ENSEMBLE DU TRAVAIL DE RECHERCHE ET ORGANISATION GÉNÉRALE DU DOCUMENT INDIQUANT LA COHÉRENCE DES ARTICLES PAR RAPPORT AUX OBJECTIFS DE LA RECHERCHE . . . 5

CHAPITRE 4 DONNÉES 6

4.1 Ensembles de simulations 6

4.1.1 ClimEx 6

4.1.2 CESM 6

4.2 Hypothèses 7

CHAPITRE 5 CADRE THÉORIQUE 8

5.1 Théorie des valeurs extrêmes 8

5.1.1 Théorème de Fisher-Tippett-Gnedenko 8

5.1.2 Théorème de Pickands–Balkema–de Haan 10

5.1.3 Validité pour des séries dépendantes 10

5.2 Méthodes d'estimation classiques 12

5.2.1 Estimateur du maximum de vraisemblance 12

5.2.2 Calcul des incertitudes associées à l'estimation 13
5.2.3 Choix parmi plusieurs modèles estimés par l'EMV 14
5.3 Méthodes d'estimation Bayésiennes 14
5.3.1 Cadre Bayésien 14
5.3.2 Méthodes MCMC 15

CHAPITRE 6 ARTICLE 1: EXTENDED GENERALIZED PARETO FOR SUBASYMPTOTIC TAIL ANALYSIS WITH APPLICATION TO HEATWAVE INTENSITIES 16
6.1 Introduction 17
6.2 Extended generalized Pareto distributions 20
6.2.1 Set of conditions for a new EGPD 20
6.2.2 Truncated normal EGPD 21
6.3 Simulation study 22
6.3.1 Experiments framework 22
6.3.2 Intrinsic quality of the EGPD model 22
6.3.3 Comparison with other EGPD models 24
6.4 Application to heatwave intensities 28
6.4.1 Data 29
6.4.2 Method 29
6.4.3 Results 31
6.5 Conclusion 33

CHAPITRE 7 MODÉLISATION DE LA DURÉE DES VAGUES DE CHALEUR . 39
7.1 Modélisation conditionnellement à l'intensité 39
7.2 Analyse exploratoire 40
7.3 Paramétrisation du modèle 41
7.4 Choix d'un modèle à l'aide du BIC 42
7.4.1 Application à ClimEx 42
7.5 Ajustement du modèle par MCMC 43
7.5.1 Application à ClimEx 44

CHAPITRE 8 MESURES DE RISQUE EN CLIMAT FUTUR 48
8.1 Modélisation des occurrences 48
8.1.1 Estimation par l'EMV 48
8.1.2 Estimation bayésienne 48
8.2 Utilisation des lois *a posteriori* pour simuler des vagues de chaleur 50
8.3 Procédure Monte Carlo 51

8.3.1 Estimation du taux d'occurrence 51
8.3.2 Estimation d'un niveau de retour 52
8.3.3 Augmentation du taux d'occurrence 53
8.4 Résultats 54
8.4.1 Ajustement bayésien en un point 54
8.4.2 Dynamiques spatiales 55

CHAPITRE 9 DISCUSSION GÉNÉRALE 59

CHAPITRE 10 CONCLUSION ET RECOMMANDATIONS 61

RÉFÉRENCES 62

LISTE DES TABLEAUX

Table 6.1 Parameter estimates of the truncated normal EGPD for heatwave intensities on ClimEx. 32
Table 6.2 Parameter estimates of the truncated normal EGPD for heatwave intensities on CESM. 33
Tableau 7.1 Résultats de la maximisation de la log-vraisemblance du modèle pour la durée des vagues de chaleur avec les données de ClimEx. 43
Tableau 7.2 Paramètres du meilleur modèle estimé par l'EMV pour les données de ClimEx. 43

LISTE DES FIGURES

Figure 6.1 Estimate of the scale (left), shape (center) and precision (right) parameters of the truncated normal EGPD model fitted on *Weibull*-tail simulated data using varying thresholds. The mean value is indicated by a solid line, while the range of the estimates is shaded to show the uncertainty in the estimation. 23

Figure 6.2 Quantile-quantile plots and densities of the fitted truncated normal EGPD on *Weibull* tails obtained with several thresholds. 23

Figure 6.3 Estimates of the scale (left), shape (center) and precision (right) parameters of the truncated normal EGPD model fitted on *Fréchet*-tail simulated data using varying thresholds. The mean value is indicated by a solid line, while the range of the estimates is shaded to show the uncertainty in the estimation. 24

Figure 6.4 Quantile-quantile plots and densities of the fitted truncated normal EGPD on *Fréchet* tails obtained with several thresholds.. 25

Figure 6.5 Comparison of the shape parameter estimates of the GPD (dashed), power EGPD (dot-dashed) and truncated normal EGPD (solid) fitted on *Fréchet*-tails (left) and *Weibull*-tails (right) using thresholds ranging from 0 to 4. 26

Figure 6.6 Standardized BIC of the GPD (dashed), power EGPD (dot-dashed) and truncated normal EGPD (solid) fitted on *Fréchet*-tails (left) and *Weibull*-tails (right) using thresholds ranging from 0 to 4. . 26

Figure 6.7 RMSE on extreme quantiles estimated with the GPD (dashed), power EGPD (dot-dashed) and truncated normal EGPD (solid) fitted on excesses of 100 samples of size 4000 with *Fréchet*-tails (left) and *Weibull*-tails (right) using thresholds ranging from 0 to 4. 28

Figure 6.8 RMSE on extreme quantiles estimated with the GPD (dashed), power EGPD (dot-dashed) and truncated normal EGPD (solid) fitted on excesses of 100 samples of size 1000 with *Fréchet*-tails (left) and *Weibull*-tails (right) using thresholds ranging from 0 to 4. 29

Figure 6.9 Number of heatwaves per year at the grid point containing the city of Montreal extracted from the 50 runs of the ClimEx ensemble. 31

Figure 6.10 Effective return levels of periods of 10 (dashed), 20 (dot-dashed) and 100 years (dotted) as a function of the time estimated in Montreal with ClimEx. 33
Figure 6.11 Correspondence between the return period of 2020 and the one of 2050 estimated in Montreal with ClimEx. 33
Figure 6.12 Comparison of estimated densities and ClimEx data (Montreal) 34
Figure 6.13 Quantile-quantile plot, ClimEx data (Montreal) 35
Figure 6.14 Number of extracted heatwaves in the grid point containing New York using the 30 runs of the CESM ensemble. 35
Figure 6.15 Effective return levels of 10 (dashed), 20 (dot-dashed) and 100 year periods (dotted) as a function of the time estimated in New York with the CESM ensemble. 36
Figure 6.16 Correspondence between the return period of 2020 and the one of 2050 estimated in New York with the CESM ensemble. 36
Figure 6.17 Comparison of estimated densities and CESM data (New York) 36
Figure 6.18 Quantile-quantile plot, CESM data (New York) 38
Figure 7.1 Distribution des couple durée-intensité des vagues de chaleur (ClimEx). 39
Figure 7.2 Moyenne et écart-type observés de la durée des vagues de chaleur. 40
Figure 7.3 Distribution de la durée en fonction de l'intensité. 41
Figure 7.4 Écart moyenne-variance pour la durée fonction de l'intensité. 41
Figure 7.5 Moyenne et écart-type estimés de la durée en fonction de l'intensité. . 43
Figure 7.6 Convergence du MCMC pour le modèle des durée (ClimEx). . 44
Figure 7.7 Diagnostic du modèle pour les durées. 46
Figure 7.8 Durée moyenne des vagues de chaleur estimées par MCMC (ClimEx). . 47
Figure 7.9 Moyenne et écart-type estimés de la durée en fonction de l'intensité. 47
Figure 8.1 Évolution de l'intensité du processus d'occurrence (EMV). 49
Figure 8.2 Évolution de l'intensité du processus d'occurrence (bayésien). 49
Figure 8.3 Diagnostic de la modélisation des occurrences par une loi de Poisson 49

Figure 8.4 Estimation par MCMC du modèle pour l'intensité des vagues de chaleur pour l'ensemble ClimEx. Chaînes de Markov montrant la convergence de l'algorithme et densité de la loi *a posteriori* pour chaque paramètre. 54
Figure 8.5 Évolution des niveaux de retour bidimensionnels calculés en 2020 en fonction de la période de retour. Intervalles de crédibilité à 95% mesurés par bootstrap. 55
Figure 8.6 Évolution des niveaux de retour bidimensionnels de période de retour 100 ans en fonction du temps. Intervalles de crédibilité à 95% mesurés par bootstrap. 55
Figure 8.7 Quantiles d'ordre 0.95 des maxima journaliers simulés de mai à octobre de l'ensemble CESM, utilisés pour définir les vagues de chaleur. 56
Figure 8.8 Niveaux de retour associés à une période de 100 ans en 2020 pour les vagues de chaleur d'au moins une semaine. 57
Figure 8.9 Niveaux de retour associés à une période de 100 ans en 2020 pour les vagues de chaleur d'au moins une semaine après avoir retranché la valeur du seuil. 58
Figure 8.10 Facteur d'augmentation entre 2020 et 2050 du taux d'occur-rence des vagues de chaleur de plus d'une semaine ayant une intensité supérieure au niveau de retour associé à la période de retour de 100 ans en 2020. 58

LISTE DES SIGLES ET ABRÉVIATIONS

BIC	*Bayesian Information Criterion*
EGPD	*Extended Genralized Pareto Distribution*
EMV	Estimateur du maximum de vraisemblance
GES	Gaz à effet de serre
GEV	*Generalized Extreme Value*
GIEC	Groupe d'experts intergouvernemental sur l'évolution du climat
GPD	*Generalized Pareto Distribution*
MCMC	Monte Carlo par chaîne de Markov
POT	*Peak Over Threshold*
RCP	*Representative Concentration Pathway*
TVE	Théorie des valeurs extrêmes

CHAPITRE 1 INTRODUCTION

L'augmentation globale de la température à la surface de la Terre entraîne avec elle une multiplication des évènements climatiques extrêmes (GIEC, 2018). Les épisodes de pluie extrême, d'inondation, de sécheresse ou encore de vague de chaleur sont amenés à se répéter de manière plus fréquente et plus intense avec les changements climatiques. Les vagues de chaleur en particulier impactent fortement l'activité humaine. Elles sont considérées comme étant l'une des catastrophes naturelles les plus mortelles, touchant aussi bien les pays développés que ceux émergents. À titre d'exemple, Robine et al. (2008) estime à 70 000 morts la surmortalité associée à la vague de chaleur qui a frappé l'Europe en 2003. Selon Mazdiyasni et al. (2017), un réchauffement global entraînerait une forte hausse de la mortalité liée aux vagues de chaleur, notamment dans les pays émergents situés à basses latitudes. Par ailleurs, le secteur de l'agriculture est aussi fortement touché et de Bono et al. (2004) estime les pertes liées à la vague de chaleur de 2003 à près de 13 milliard d'euros. Les vagues de chaleur sont également responsables de l'augmentation du nombre de feux de forêt (Parente et al., 2018). Les conséquences touchent aussi d'autres secteurs tels que la production ainsi que la demande en électricité (Añel et al., 2017), la qualité de l'air, le transport aérien, les réserves d'eau et la biodiversité, etc.

L'impact des évènements climatiques extrêmes peut être réduit au moyen de mécanismes d'adaptation : construction d'infrastructures ou mise en place de plans d'alertes. A la manière des canaux d'irrigation permettant de faire face à des épisodes de sécheresse ou des barrages permettant de maîtriser les crues de rivières lors d'épisodes de fortes pluies, il est possible de limiter l'impact des vagues de chaleur en créant de nouvelles normes de construction ou en augmentant la canopie urbaine pour réduire l'effet des îlots de chaleur urbains (Imran et al., 2019).

Afin de dimensionner les mesures d'adaptation, il est nécessaire de caractériser de manière précise les vagues de chaleur en climat présent et futur en tenant compte des changements climatiques. Ces changements futurs sont incertains et dépendent notamment des émissions de gaz à effet de serre (GES) présentes et futures (e.g. Matthews et al., 2009; Leduc et al., 2016). Plusieurs scénarios d'émissions de GES ont été établis par le groupe d'experts intergouvernemental sur l'évolution du climat (GIEC). Chaque scénario permet d'obtenir une évolution probable du climat en prenant pour hypothèse une trajectoire d'émission. En se basant sur des modèles physiques décrivant l'atmosphère, les océans, la Terre et leurs interactions, et en imposant un forçage radiatif dépendant du scénario *Representative Concentration*

Pathway (RCP) choisi, il est possible de simuler le climat futur. Bien que dépendantes du scénario RCP et du modèle climatique utilisé, ces simulations du climat permettent de produire des projections quantitatives des changements climatiques, ce que l'utilisation de données observées ne permet pas. D'une part les changements climatiques ne sont pas linéaires et il serait hasardeux de prévoir les changements futurs uniquement à partir des tendances passées observées. D'autre part, le nombre de données observées est relativement restreint pour l'estimation des extrêmes climatiques. Il est ambitieux d'espérer obtenir une estimation précise d'un évènement dont la période de retour serait du même ordre de grandeur que la période d'observation, alors qu'en analysant des données simulées provenant d'un ensemble de simulation (plus de détails en Section 4), l'étude des évènements rares devient possible.

Il n'existe pas de définition universelle des vagues de chaleur. Néanmoins, les définitions incluent toujours une succession de jours au cours desquels la température demeure haute, typiquement au-dessus d'un seuil défini à partir d'un quantile de la distribution des températures journalières en un lieu donné. Certaines définitions prennent également en compte la température minimale journalière. En effet, lorsque l'on s'intéresse aux impacts de la chaleur sur la santé, les organismes sont d'autant plus affectés par la chaleur lorsque la température demeure élevée la nuit empêchant ainsi l'organisme de récupérer. Nous adopterons donc une définition utilisant deux seuils proposée par Keellings and Waylen (2014), en définissant une vague de chaleur comme une succession de jours où la température maximale ainsi que la température minimale dépassent simultanément des seuils prédéfinis. Une vague de chaleur se termine si, chaque jour au cours d'une succession de r jours de répit, au moins l'un des deux seuils n'est pas franchi. L'interprétation d'une telle période de répit pour l'application à la santé est qu'il est nécessaire pour les organismes d'avoir au moins r jours de répit afin de récupérer des premières journées de forte chaleur. Si ce n'est pas le cas, on peut considérer que les températures atteintes lors des premiers jours de l'épisode ont encore un impact lorsqu'une nouvelle succession de jours chauds arrive et on ne considère donc qu'un seul évènement.

CHAPITRE 2 REVUE CRITIQUE DE LA LITTÉRATURE

La définition des vagues de chaleur donnée par Keellings and Waylen (2014) permet de caractériser les vagues de chaleur en terme de fréquence d'occurrence, de durée et d'intensité. Nous définirons l'intensité comme la température maximale atteinte au cours de la vague de chaleur comme Keellings and Waylen (2014), même si certaines études utilisent aussi la température moyenne au cours de la vague de chaleur (e.g. Mazdiyasni et al., 2019).

La fréquence d'occurrence des vagues de chaleur est souvent modélisée par un processus de Poisson non-homogène (e.g. Keellings and Waylen, 2014; Abadie et al., 2018; Furrer et al., 2010). Les caractéristiques de chaque évènement, à savoir la durée et l'intensité, peuvent ensuite être modélisées séparément. Cependant la durée et l'intensité de la vague de chaleur ne sont pas indépendantes et la modélisation doit en tenir compte. Plusieurs approches existent dans la littérature. Des modèles dit génératifs sont proposés par Furrer et al. (2010) et Shaby et al. (2014). Ces approches visent à modéliser l'intégralité de la vague de chaleur et donc permettent entre autre de tenir compte de la dépendance entre la durée de la vague de chaleur et la température maximale atteinte. Ces approches sont souvent complexes à mettre en œuvre puisqu'il est nécessaire de modéliser l'ensemble du processus au lieu de se contenter de décrire les caractéristiques d'intérêt du processus. Ainsi, de nombreuses études se contentent d'une approche descriptive en modélisant uniquement la durée et l'intensité des vagues de chaleur. Keellings and Waylen (2014) propose une approche descriptive mais ne tient pas compte de la dépendance entre durée et intensité. Mazdiyasni et al. (2019) propose lui d'utiliser une copule pour modéliser la dépendance entre la durée et l'intensité et ainsi obtenir la loi conjointe du couple à partir des deux lois marginales. Dans notre cas où la durée est une variable discrète, la mise en œuvre de cette approche est rendue plus délicate due à la non-unicité de la copule lorsqu'une marginale est discrète (Zilko and Kurowicka, 2016). La seconde approche que nous privilégierons consiste à passer par la distribution conditionnelle d'une des deux variables. Soit $D \in \mathbb{N}^*$ la durée d'une vague de chaleur et $Z \in \mathbb{R}_+$ son intensité. La densité du couple (D, Z) pour $d \in \mathbb{N}^*$ et $z \geq 0$ s'écrit :

$$f_{(D,Z)}(d, z) = \mathbb{P}(D = d \mid Z = z) \times f_Z(z). \tag{2.1}$$

Il suffit donc de caractériser la loi marginale de l'intensité des vagues de chaleur puis la loi de la durée conditionnellement à l'intensité. La dépendance entre les deux variables aléatoires sera alors contenue dans l'expression de cette loi conditionnelle.

Étant donné que les impacts des vagues de chaleur ne sont pas linéaires avec leur sévérité

(e.g. Schlenker and Roberts, 2006), il est capital de modéliser correctement les valeurs extrêmes correspondant aux plus gros risques. La théorie des valeurs extrêmes fournit un cadre théorique solide pour modéliser les extrêmes d'une distribution. La loi de Pareto généralisée (GPD) est largement utilisée dans la littérature pour modéliser l'intensité des vagues de chaleur (e.g. Furrer et al., 2010; Shaby et al., 2014; Keellings and Waylen, 2014). L'usage de la GPD n'est cependant justifié qu'asymptotiquement si le seuil de définition des vagues de chaleur est suffisamment élevé. Si cela est le cas en climat passé et présent, où la définition d'une vague de chaleur coïncide avec un extrême statistique, le nombre de vague de chaleur est amené à considérablement croître au cours des prochaines années rendant le seuil utilisé pour définir les vagues de chaleur sous-asymptotique au sens de la théorie des valeurs extrêmes. Papastathopoulos and Tawn (2013) proposent l'usage de la loi de Pareto généralisée étendue (EGPD) afin d'étendre la GPD à des valeurs sous-asymptotiques (plus de détails en Section 6). Nous utiliserons ce type de modèle pour modéliser l'intensité des vagues de chaleur lorsque le seuil de définition est sous-asymptotique. Cette approche peut aussi permettre d'étendre la méthodologie *Peak Over Threshold* (POT) à des seuils sous-asymptotiques.

CHAPITRE 3 DÉMARCHE DE L'ENSEMBLE DU TRAVAIL DE RECHERCHE ET ORGANISATION GÉNÉRALE DU DOCUMENT INDIQUANT LA COHÉRENCE DES ARTICLES PAR RAPPORT AUX OBJECTIFS DE LA RECHERCHE

Ce mémoire est organisé de la manière suivante. Les données utilisées dans le mémoire seront décrites en détail dans la section 4. La section 5 présentera ensuite les bases théoriques nécessaires à la compréhension de la suite du mémoire. Les principaux résultats de la théorie des valeurs extrêmes seront donnés et les méthodes d'estimation classiques et bayésiennes que nous utiliserons par la suite seront détaillées. La section 6, se présentant sous la forme d'un article, traitera de la modélisation de l'intensité des vagues de chaleur à l'aide d'un modèle EGPD. La modélisation conditionnelle de la durée des vagues de chaleur sera présenté dans la section 7. Enfin, la section 8 utilisera les modèles développés dans les sections précédentes pour décrire les vagues de chaleur en climat futur. Des estimations ponctuelles dans un cadre bayésien permettront de mettre l'accent sur la détermination des incertitudes, tandis que des estimations par estimateur du maximum de vraisemblance (EMV) permettront de dégager des tendances spatiales sur l'évolution des vagues de chaleur afin de déterminer les régions les plus touchées par les changements climatiques.

CHAPITRE 4 DONNÉES

4.1 Ensembles de simulations

Nous ferons usage dans ce mémoire de deux ensembles de simulations climatiques : ClimEx et CESM.

4.1.1 ClimEx

Présenté en détail dans Leduc et al. (2019), ClimEx est un ensemble de simulations climatiques visant à étudier les extrêmes climatiques, notamment en hydrologie, afin de mieux contrôler la gestion des eaux en Bavière et au Québec. Il comporte 50 membres d'un modèle climatique régional (CRCM5) ayant une haute résolution horizontale d'environ 11km. Chaque membre du modèle régional est pilotée par une simulation d'un modèle climatique de circulation global (CanESM2) ayant lui une résolution plus faible (200km). Chaque membre de l'ensemble CanESM2 correspond à des conditions initiales légèrement perturbées. Cela permet de rendre compte de la variabilité naturelle du climat. Les simulations sont conduites pour chaque membre de 1950 à 2100, en prenant le scénario RCP8.5 pour le forçage radiatif lié aux gaz à effets de serre. Le forçage radiatif de la Terre peut être vu comme le bilan entre l'énergie absorbée et rayonnée par la Terre. Les gazs à effet de serre participe positivement au forçage radiatif de la Terre ce qui conduit à un réchauffement de l'atmosphère. Le scénario RCP8.5 est l'un des quatre scénarios de trajectoire du forçage radiatif à l'horizon 2300 établis par le GIEC pour son cinquième rapport (IPCC, 2013). Ce scénario est le plus pessimiste et correspond à un forçage radiatif $+8.5\mathrm{W.m}^{-2}$ en 2100, il n'est cependant pas irréaliste compte-tenu des émissions de gaz à effet de serre actuelles.

4.1.2 CESM

L'ensemble CESM-LE (Community Earth System Model Large Ensemble), présenté dans Kay et al. (2015), est lui un ensemble de simulations du modèle climatique global Community Earth System Model. En conséquence, sa résolution est moins élevée, de l'ordre d'un degré de résolution horizontale. Cela empêche de rendre compte d'effets très localisés, cependant cela permet bien de déceler des dynamiques spatiales à plus grandes échelles. Tout comme l'ensemble ClimEx, il utilise le scénario d'émission RCP8.5. Il comporte 30 membres et les simulations que nous analyserons vont de 2006 à 2080.

4.2 Hypothèses

Pour chaque ensemble de simulations, nous ferons l'hypothèse que chacun des membres constitue une réalisation équiprobable du climat, et indépendante des autres conditionnellement au modèle climatique. L'étude de l'ensemble des membres doit permettre de rendre compte de la variabilité interne du modèle climatique étudié. De plus, l'étude de tous les membres permet plus d'observations et plus particulièrement plus d'observations d'extrêmes climatiques. Nous ferons donc l'hypothèse, réaliste et étudiée dans Leduc et al. (2016) pour ClimEx, que les différentes réalisations du climat sont conditionnellement indépendantes et identiquement distribuées.

CHAPITRE 5 CADRE THÉORIQUE

5.1 Théorie des valeurs extrêmes

5.1.1 Théorème de Fisher-Tippett-Gnedenko

La théorie des valeurs extrêmes est une branche des statistiques s'intéressant aux valeurs extrêmes des distributions de probabilités, développée originellement par Emil Julius Gumbel (Gumbel, 1958). Elle nous donne le cadre théorique nécessaire pour étudier la distribution des extrêmes d'une variable aléatoire donnée. Un des résultats principaux de la théorie des valeurs extrêmes est le théorème de Fisher-Tippett-Gnedenko (Fisher and Tippett, 1928). Il énonce un résultat permettant de caractériser la loi du maximum d'un échantillon de variables aléatoires indépendantes et identiquement distribuées. De manière analogue au théorème central limite pour la moyenne, il énonce une convergence en loi du maximum normalisé par deux suites de constantes vers une famille de lois limite lorsque la taille de l'échantillon augmente. Soit $X_1, X_2, ...$ des variables aléatoires indépendantes et identiquement distribuées selon une loi inconnue F. Nous souhaitons inférer la loi du maximum M_n de n de ces variables aléatoires défini par

$$M_n = \max\left(X_1, \ldots, X_n\right).$$

Théorème 1 (Fisher-Tippett-Gnedenko) *S'il existe des suites déterministes (a_n) réelle positive et (b_n) réelle telles que*

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

où G est une distribution non dégénérée, alors G appartient à l'une des familles suivantes : Fréchet, Gumbel ou Weibull. La fonction de répartition de chacune de ces familles de lois peut s'écrire sous la même forme paramétrique que l'on appelle loi loi d'extremum généralisée (GEV) :

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \tag{5.1}$$

où $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_+^*$ et $\xi \in \mathbb{R}$ sont respectivement appelés paramètres de localisation, d'échelle et de forme, et où $z \in \{z \in \mathbb{R}, 1 + \xi\left(\frac{z-\mu}{\sigma}\right) \geq 0\}$.

Le signe du paramètre de forme indique la famille de la distribution limite. Ainsi $\xi > 0$ correspond à la famille de Fréchet, $\xi \to 0$ correspond à la famille de Gumbel, et $\xi < 0$ correspond à

la famille de Weibull. Le paramètre de forme caractérise le poids de la queue de distribution de la loi. Plus le paramètre de forme est grand et plus la queue de la distribution est lourde. Dans le cas des lois de la famille de Gumbel, la décroissance de la fonction de densité de la distribution est exponentielle. Dans le cas des lois de Fréchet, la queue de distribution est plus lourde, tandis que pour les lois de Weibull, le domaine de définition où la fonction de densité est strictement positive est borné à droite. Ainsi les valeurs prises sont bornées par la borne $z_{\max} = \mu - \sigma/\xi$ dès lors que $\xi < 0$.

Le théorème 1 fournit un résultat asymptotique sur la loi du maximum M_n normalisé par les constantes a_n et b_n. Cependant nous pouvons déduire de ce résultat la loi du maximum de l'échantillon M_n qui est aussi une loi des valeurs extrêmes généralisée. En effet, supposons que

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

pour n suffisamment grand, alors de manière équivalente

$$\begin{aligned} \mathbb{P}\left(M_n \leq z\right) &\approx G\left\{\left(z - b_n\right)/a_n\right\} \\ &= G^*(z) \end{aligned}$$

où G^* est un autre membre de la famille GEV. Les paramètres de G et de G^* sont reliés par le système suivant

$$\begin{cases} \mu^* &= a_n\mu + b_n \\ \sigma^* &= a_n\sigma \\ \xi^* &= \xi \end{cases}$$

Nous estimerons par la suite pour un échantillon de taille n directement les paramètres de la loi du maximum annuel sans nous soucier des constantes de normalisation que nous intégrerons directement aux paramètres de la loi estimée.

Il est important de noter que contrairement au théorème central limite qui s'applique sur n'importe quelle loi ayant une variance finie, le théorème de Fisher-Tippet-Gnedenko énonce que si la loi d'un maximum normalisé converge, alors sa limite ne peut être qu'un certain type de loi. En outre, ce théorème n'énonce pas que la loi d'un maximum normalisé converge. Nous devrons par la suite dans notre étude faire l'hypothèse que la loi du maximum normalisé converge puis vérifier *a posteriori* cette hypothèse.

5.1.2 Théorème de Pickands–Balkema–de Haan

Nous avons caractérisé la loi du maximum d'un échantillon indépendant et identiquement distribué de variables aléatoires, nous aimerions maintenant pouvoir quantifier la probabilité de dépassement de seuils extrêmes. Cela ne peut en général pas être fait de manière traditionnelle en comptant la proportion d'observations au-dessus du seuil donné puisque ces seuils étant extrêmes, on n'a que peu voir aucune observations au-dessus de ce seuil dans notre échantillon. C'est pourquoi la théorie des valeurs extrêmes fournit un second résultat central : le théorème de Pickands-Balkema-de Haan (Balkema and de Haan, 1974; Pickands, 1975), qui nous donne la distribution conditionnelle au-delà d'un seuil u fixé.

Théorème 2 (Pickands-Balkema-de Haan) *Soit $X_1, X_2, ...$ des variables aléatoires indépendantes et identiquement distribuées selon une loi F inconnue. Supposons que F satisfait le théorème de Fisher-Tippett-Gnedenko, au sens où pour n assez grand $\mathbb{P}(M_n \leq z) = G(z)$ où G est une distribution des valeurs extrêmes généralisée pour $\mu, \sigma > 0$ et ξ donnés. Alors pour u assez grand,*

$$\mathbb{P}\left(X - u \leq y | X > u\right) \simeq 1 - \left(1 + \xi \frac{y}{\sigma_u}\right)^{-\frac{1}{\xi}}. \tag{5.2}$$

La distribution ainsi trouvée est appelée GPD.

Il est important de noter que dans ce résultat le paramètre de forme ξ est le même que pour celui de la loi GEV du maximum M_n. En revanche le paramètre d'échelle σ_u est variable et dépend du seuil u selon l'équation

$$\sigma_u = \sigma + \xi(u - \mu) \tag{5.3}$$

Dans le cadre de notre étude, nous devons faire face à un problème avant de pouvoir utiliser les théorèmes 1 et 2 avec nos données : nos observations journalières ne sont pas indépendantes.

5.1.3 Validité pour des séries dépendantes

Dans le cas des températures, la série des températures maximales journalières estivales est très autocorrélée (autocorrélation de 0.84 avec un lag d'un jour, 0.46 avec un lag d'une semaine). La théorie des valeurs extrêmes fournit également des résultats permettant d'utiliser les théorèmes 1 et 2 lorsque les données ne sont pas indépendantes, sous certaines conditions néanmoins.

Définition 1 *Soit $X_1, X_2 \ldots$ un processus stationnaire. On dit qu'il satisfait la condition*

$D(u_n)$ si pour tout $i_1 < \cdots < i_p < j_1 < \cdots < j_p$ avec $j_1 - i_p > l$,

$$|\mathbb{P}\left\{X_{i_1} \leq u_n, \ldots, X_{i_p} \leq u_n, X_{j_1} \leq u_n, \ldots, X_{j_p} \leq u_n\right\} \\ - \mathbb{P}\left\{X_{i_1} \leq u_n, \ldots, X_{i_p} \leq u_n\right\} \cdot \mathbb{P}\left\{X_{j_1} \leq u_n, \ldots, X_{j_p} \leq u_n\right\}| \leq \alpha(n, l)$$

où $\alpha(n, l_n) \to 0$ pour une certaine suite l_n telle que $l_n/n \to 0$.

Pour une série stationnaire indépendante, la condition $D(u_n)$ est toujours vérifiée puisque la différence de probabilité intervenant dans la définition est toujours nulle. Cette condition servira à exprimer le fait que pour deux ensembles de variables suffisament séparées temporellement, bien que non nulle la différence de probabilité avec le cas indépendant est suffisamment petite pour ne pas affecter la loi limite des extrêmes. Cette idée est explicitée par le théorème suivant.

Théorème 3 *Soit $X_1, X_2 \ldots$ un processus stationnaire et $M_n = \max\{X_1 \ldots X_n\}$. S'il existe des suites de constantes $a_n > 0$ et b_n telles que*

$$\mathbb{P}\left\{(M_n - b_n)/a_n \leq z\right\} \to G(z),$$

où G est une fonction de répartition non-dégénérée, et que la condition $D(a_n z + b_n)$ est vérifiée pour tout réel z, alors G appartient à la famille GEV.

Ce résultat précise que pour des séries stationnaires ayant des dépendances temporelles limitées aux niveaux extrêmes (au sens précisé par la condition $D(u_n)$), leur maxima suivent une distribution limite de la même famille que celles de séries indépendantes. Pour une même distribution marginale, le lien entre la loi limite pour un processus indépendant et pour un processus vérifiant simplement $D(u_n)$ est explicité par le théorème suivant.

Théorème 4 *Soit $X_1, X_2 \ldots$ un processus stationnaire vérifiant une condition $D(u_n)$, et $X_1^*, X_2^* \ldots$ un processus stationnaire indépendant ayant la même distribution marginale. En définissant M_n et M_n^* les maxima d'échantillons de taille n de ces deux séries,*

$$\mathbb{P}\left(M_n^* \leq z\right) \to G_1(z)$$

où G_1 est une fonction de répartition non-dégénérée, si et seulement si

$$\mathbb{P}\left(M_n \leq z\right) \to G_2(z)$$

où

$$G_2(z) = G_1^{\theta}(z)$$

pour une constante θ *vérifiant* $0 < \theta \leq 1$.

Ce théorème nous dit que si le maximum d'une série stationnaire converge (ce qui est le cas par le théorème 3 si la série vérifie une condition de type $D(u_n)$), alors sa distribution limite est liée à celle d'une série indépendante avec la même distribution marginale. Le seul effet de la dépendance est de remplacer la distribution limite G_1 par G_1^{θ}, ce qui est cohérent avec le théorème 3, puisqu'alors G_1^{θ} appartient toujours à la famille GEV. En effet, si $\xi \neq 0$

$$\begin{aligned} G_1^{\theta}(z) &= \left(\exp\left\{-\left[1+\xi\left(\tfrac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}\right)^{\theta} \\ &= \exp\left\{-\theta\left[1+\xi\left(\tfrac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \\ &= \exp\left\{-\left[1+\xi\left(\tfrac{z-\mu^*}{\sigma^*}\right)\right]^{-1/\xi}\right\} \end{aligned}$$

où

$$\mu^* = \mu - \frac{\sigma}{\xi}\left(1-\theta^{-\xi}\right) \quad \text{et} \quad \sigma^* = \sigma\theta^{\xi}.$$

Le paramètre de forme est lui inchangé. De plus, le même type de relations existe dans le cas de la famille de Gumbel. Le paramètre θ est appelé indice extrémal. Il quantifie la tendance qu'on les extrêmes de la série à se produire en groupe. Plus θ est petit et plus les extrêmes auront tendance à se produire en groupe, et plus la distribution limite sera éloignée de la distribution limite qu'aurait la série stationnaire indépendante de même distribution marginale.

Pour résumer, selon certaines conditions plus faibles que l'indépendance de la série, les mêmes théorèmes sont valables pour étudier les extrêmes d'une série stationnaire. Bien que la condition $D(u_n)$ soit en pratique difficile à vérifier, ces résultats donnent des arguments théoriques nous permettant d'utiliser les théorèmes 1 et 2 pour notre étude sur les températures extrêmes.

5.2 Méthodes d'estimation classiques

5.2.1 Estimateur du maximum de vraisemblance

L'ensemble des modèles que nous développerons au cours de ce mémoire sont des modèles paramétriques. C'est-à-dire que nous fixons un ensemble de fonctions de densité de probabilité possibles que l'on peut définir à partir d'un ensemble de paramètres $\Theta \subset \mathbb{R}^p$ où p est le nombre

de degrés de liberté du modèle.

$$\mathcal{M} = \{f(x;\theta)|\theta \in \Theta\}.$$

A partir d'un ensemble d'observations $(x_1, \ldots, x_n)$, il s'agit de sélectionner le vecteur de paramètre maximisant la vraisemblance du modèle considéré. Nous cherchons à maximiser la vraisemblance

$$L(x,\theta) = \prod_{i=1}^{n} f(x_i;\theta). \tag{5.4}$$

Pour des raisons numériques il est souvent plus efficace de considérer de manière équivalente la log-vraisemblance $\ell(x,\theta)$. Si $(x_1, \ldots, x_n)$ sont des réalisations indépendantes d'une variable aléatoire ayant la fonction de densité $f(x;\theta_0)$, nous définissons l'EMV $\hat{\theta}_0$ comme la valeur de θ maximisant la fonction de vraisemblance définie précédemment. Il convient de garder à l'esprit que le choix d'un modèle engendre généralement un biais puisque la véritable fonction de densité de nos observations n'appartient en général pas au modèle $\mathcal{M}$ fixé préalablement.

5.2.2 Calcul des incertitudes associées à l'estimation

Nous voudrons généralement obtenir des intervalles de confiance pour les paramètres que nous estimerons par l'EMV. Le théorème suivant nous permet d'obtenir l'incertitude liée à l'EMV avec un échantillon d'observations de taille n.

Théorème 5 *Soit $(x_1 \ldots x_n)$ un échantillon d'observations indépendantes d'une variable aléatoire ayant une fonction de densité dans une famille paramétrique $\mathcal{M}$ de dimension p. Alors pour n assez grand*

$$\hat{\theta}_0 \sim N_p(\theta_0, I_E(\theta_0)^{-1}), \tag{5.5}$$

où I_E est l'espérance de l'opposé de la matrice héssienne de la log-vraisemblance, c'est-à-dire l'espérance de la matrice d'information.

Comme nous ne connaissons pas en général θ_0, nous remplacerons I_E par la matrice d'information observée pour obtenir des intervalles de confiance pour nos estimations. Nous aurons ensuite grandement recours à la méthode delta.

Théorème 6 *Soit $g : \mathbb{R}^p \to \mathbb{R}^d$ une fonction différentiable et $\hat{\theta}_0$ l'EMV, de matrice de covariance V_θ. Alors si $\phi = g(\theta)$, l'EMV de $\phi_0 = g(\theta_0)$ satisfait*

$$\hat{\phi}_0 \sim N_d(\phi_0, V_\phi), \tag{5.6}$$

où

$$V_\phi = \nabla g(\hat{\theta}_0)^T V_\theta \nabla g(\hat{\theta}_0) \tag{5.7}$$

5.2.3 Choix parmi plusieurs modèles estimés par l'EMV

Lors d'une estimation utilisant l'EMV, un choix capital à faire est celui de la sélection du modèle initial. Plus celui-ci est complexe, plus la vraisemblance maximale sera élevée. Cela ne signifie pas pour autant que le meilleur modèle sera le plus complexe. Il y a un compromis à faire afin de ne pas avoir un modèle trop complexe.

Critère d'information bayésien

Afin de sélectionner un modèle parmi un ensemble de plusieurs modèles $\mathcal{M}_i$ de dimensions respectives k_i, nous utiliserons le critère d'information Bayésien (BIC) défini par

$$BIC_i = \ell_i(x, \hat{\theta}_0) - \frac{k_i}{2} \log n.$$

où n est la taille de l'échantillon. Nous retiendrons le modèle qui maximisera ce critère.

Cas particulier de modèle imbriqués

Si $\mathcal{M}_0 \subset \mathcal{M}_1$, on peut également quantifier la préférence pour un modèle plutôt qu'un autre en comparant la différence des log-vraisemblances à une loi du χ^2. Si $2(\ell_1(\hat{\theta}_0) - \ell_0(\hat{\theta}_0)) > c_\alpha$ où c_α est le quantile d'ordre $1-\alpha$ d'une loi χ^2_k où k est le nombre de degrés de liberté de différence entre $\mathcal{M}_1$ et $\mathcal{M}_0$, alors on rejettera $\mathcal{M}_0$ au niveau α de significativité.

5.3 Méthodes d'estimation Bayésiennes

Cette section vise à décrire l'approche bayésienne utilisée pour estimer les paramètres d'un modèle.

5.3.1 Cadre Bayésien

Considérons un vecteur $\boldsymbol{y} = (y_1, \ldots, y_n)$ de réalisation d'une variable aléatoire Y distribuée selon la densité $f(\boldsymbol{y}|\boldsymbol{\theta})$ où f est connue et le vecteur de paramètres $\boldsymbol{\theta}$ est une variable aléatoire dont on cherche à déterminer la loi. En l'absence d'observations $\boldsymbol{y}$, on peut supposer que $\boldsymbol{\theta}$ suit la loi *a priori* $f_\Theta(\boldsymbol{\theta})$. Cette loi peut être informative si l'on veut inclure de l'information *a priori* sur la distribution des paramètres, ou non-informative si l'on ne dispose

pas de connaissances particulières sur la distribution attendue. La loi *a priori* peut être une loi impropre, *i.e.* d'intégrale infinie (*e.g. une loi constante sur un domaine non borné*). Le théorème de Bayes permet de déterminer la loi *a posteriori* des paramètres, compte-tenu des observations :

$$f_{(\boldsymbol{\Theta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\theta}) = \frac{f_{(\boldsymbol{Y}|\boldsymbol{\Theta}=\boldsymbol{\theta})}(\boldsymbol{y})\, f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}{\int f_{(\boldsymbol{Y}|\boldsymbol{\Theta}=\boldsymbol{\theta})}(\boldsymbol{y})\, f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}}. \tag{5.8}$$

En particulier, on peut écrire

$$f_{(\boldsymbol{\Theta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\theta}) \propto f_{(\boldsymbol{Y}|\boldsymbol{\Theta}=\boldsymbol{\theta})}(\boldsymbol{y})\, f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}). \tag{5.9}$$

5.3.2 Méthodes MCMC

Les algorithmes Monte Carlo par chaîne de Markov (MCMC) sont des méthodes permettant de déterminer à partir d'observations la loi *a posteriori* d'un vecteur de paramètres. Plus précisément, ils permettent de générer un échantillon de la loi *a posteriori* à partir d'une chaîne de Markov ayant la loi *a posteriori* comme loi stationnaire. Présentons l'un des algorithmes de MCMC les plus connus, que nous utiliserons dans ce mémoire : l'échantillonnage de Gibbs (Casella and George, 1992). Soit $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ un vecteur de paramètres initiaux, et $\boldsymbol{\delta}$ un vecteur de réels strictement positifs. A chaque itération de l'algorithme, un candidat $\boldsymbol{\theta}^{(i+1)}$ est construit à partir du vecteur de paramètres courant $\boldsymbol{\theta}^{(i)}$. La spécificité de l'échantillonnage de Gibbs par rapport à l'algorithme de Metropolis-Hasting (Hastings, 1970) est que chacune des composantes du vecteur $\boldsymbol{\theta}^{(i+1)}$ est construite séparément selon la procédure décrite par l'algorithme 1. Pour des tailles de pas $\boldsymbol{\delta}$ bien choisies, la loi du vecteur $\boldsymbol{\theta}^{(i)}$ converge vers la

Algorithme 1 Itération de l'échantillonnage de Gibbs

$\boldsymbol{\theta}^{(i+1)} \leftarrow \boldsymbol{\theta}^{(i)}$
for $1 \leq j \leq d$ **do**
  $\sigma_j \sim \mathcal{N}(0,1)$
  $\theta_j \leftarrow \theta_j^{(i)} + \delta_j \sigma_j$
  $\boldsymbol{\theta}^c \leftarrow \left(\theta_1^{(i+1)}, \ldots, \theta_{j-1}^{(i+1)}, \theta_j, \theta_{j+1}^{(i)}, \ldots, \theta_d^{(i)}\right)$
  $\alpha \leftarrow \log f\left(\boldsymbol{y} \mid \boldsymbol{\theta}^c\right) - \log f\left(\boldsymbol{y} \mid \boldsymbol{\theta}^{(i+1)}\right)$
  $u \sim \mathcal{U}\left([0,1]\right)$
  if $\alpha > u$ **then**
    $\boldsymbol{\theta}^{(i+1)} \leftarrow \boldsymbol{\theta}^c$
  end if
end for
return $\boldsymbol{\theta}^{(i+1)}$

loi *a posteriori* et les itérations de l'algorithme nous fournisse un échantillon de cette loi.

CHAPITRE 6 ARTICLE 1: EXTENDED GENERALIZED PARETO FOR SUBASYMPTOTIC TAIL ANALYSIS WITH APPLICATION TO HEATWAVE INTENSITIES

Cet article est actuellement en révision à la revue *Journal of the Royal Statistical Society Series C*.

Philémon Gamet and Jonathan Jalbert,
Département de mathématiques et de génie industriel, Polytechnique Montréal

For both financial and environmental applications, tail distributions often correspond to extreme risks, and accurate modeling is mandatory. The peaks-over-threshold methodology is a classic way to model extreme values using the generalized Pareto distribution. However, for some applications, the choice of a threshold is challenging and the asymptotic conditions are not satisfied. We propose the use of extended generalized Pareto models to enhance the peaks-over-threshold methodology to subasymptotic thresholds. These existing extensions have either infinite or null density at the origin, which makes them particularly adapted to model an overall distribution but poorly suited to model tails. A novel extension of the generalized Pareto distribution is developed to model subasymptotic tails with a nonzero and finite density at the threshold. Then, the proposed method is tested on simulated data to assess its accuracy. The proposed extension provides a better estimate of extreme quantiles than existing models, especially for small samples. Finally, this new distribution is used to model nonstationary heatwave intensity from large climate simulation ensembles.

Keywords: extreme value theory, extended generalized Pareto distribution, peaks-over-threshold, heatwave, climate change

6.1 Introduction

Climate extremes such as heatwaves are of great interest for policy planners and policy makers because of their critical impacts on health, agriculture and infrastructure. For example, the 2003 heatwave over Europe caused over 70,000 deaths and cost more than 13 billion euros in agricultural losses. In the context of global warming, the frequency and severity of heatwaves are expected to increase in the coming decades, as reported in GIEC (2018). Thus, adaptation and mitigation measures must be made in line with future risks, which must be assessed by reliable statistical modeling of heatwave characteristics.

While the definition of a heatwave depends on the climate at a given location, it generally refers to a streak of consecutive days during which the temperature exceeds a given threshold. Heatwave intensity can be defined as the maximum temperature reached during the series of hot days. If the threshold is sufficiently high, standard extreme value models, such as the peaks-over-threshold (POT) methodology (Davison, 1984; Smith, 1984), can be used to model the heatwave intensity (see Keellings and Waylen, 2014; Furrer et al., 2010, for example). This methodology, described in detail by Coles (2001, chap. 4), relies on the *Pickands-Balkema-de Haan* theorem (Balkema and de Haan, 1974; Pickands, 1975), which ensures, under some technical assumptions, that for any random variable X of unknown distribution F_X and sufficiently high threshold u, the conditional tail distribution can be approximated as follows

$$\mathbb{P}\left(X - u \leq z | X > u\right) \approx 1 - \left(1 + \xi \frac{z}{\sigma}\right)_{+}^{-1/\xi}, \text{ for } z \geq 0. \tag{6.1}$$

This limiting distribution is referred to as the generalized Pareto distribution (GPD), and the parameters $\sigma > 0$ and ξ are, respectively, the scale and shape parameters.

The choice of the threshold u is critical. A threshold that is too low would lead to a biased estimate as the GPD is only asymptotically justified, and a threshold that is too high would leave insufficient data for inference and, thus, lead to considerable variance. Several diagnostics, such as shape stability plot or mean residual life plot, are available for choosing an adapted threshold (see Coles, 2001, chap. 4).

As mentioned by Stein (2017), the conditions for using the extreme value theory (EVT) for studying extreme temperatures are not always satisfied. For instance, the threshold used to define heatwaves is often not sufficiently high for Eq. (6.1) to be a good approximation. It is especially the case in the global warming context: the threshold is increasingly exceeded as the years progress. The fact that the heatwave definition cannot be used as the threshold in

Eq. (6.1) is inconvenient because the heatwave definition is easy to interpret by the scientific and nonscientific community.

Various approaches exist in the literature to address this subasymptotic modeling issue. One method to investigate the distribution of a random variable X while preserving the GPD asymptotically is to use a mixture model for the probability density function (pdf) f_X:

$$f_X(x) = \begin{cases} h_X(x) & \text{if } x \leq u \\ \{1 - H_X(u)\}\, g_{(X-u|X>u)}(x-u) & \text{if } x > u \end{cases} \tag{6.2}$$

where $h_X(\cdot)$ and $H_X(\cdot)$ denote respectively the probability density function and the cumulative distribution function of X and $g_{(X-u|X>u)}(x-u)$ denotes the GPD density of the tail $(X-u|X>u)$. This choice of mixture distribution is motivated by the difficulty of estimating both the bulk and the tail distribution. The mixture may be discontinuous, as in McDonald et al. (2011), even though h_X and $g_{(X-u|X>u)}$ are continuous. However, by imposing conditions on the estimator or on the parameters of the model, we can ensure the continuity of f_X at u (Carreau and Bengio, 2009; Furrer et al., 2010). Nevertheless, these additional constraints may seriously reduce the flexibility of the model. Another approach to guarantee the continuity of f_X is to use a smooth transition function between both h_X and $g_{(X-u|X>u)}$ densities, as in Frigessi et al. (2002):

$$f_X(x) = \{1 - p(x)\}\; h_X(x) + p(x)\, g_{(X-u|X>u)}(x-u) \tag{6.3}$$

where $p(\cdot)$ is an increasing function with values in $[0,1]$. Defining such a function requires additional parameters that add complexity to the model and make the inference numerically more difficult. Mixture models have the advantage of automatically defining a threshold u above which the asymptotic framework of the EVT holds. However, the subasymptotic estimate h_X in Eq. (6.2) is often a nonparametric kernel density estimate, and such an approach is difficult to implement in a nonstationary framework, which is often required in climatic or financial applications where many exogenous covariates are at stake. Moreover, if one prefers a parametric approach for the estimate of h_X, it requires considerable prior information, which may depend on the applications.

An alternative method based on probability integral transform was developed by Papastathopoulos and Tawn (2013) to take into account subasymptotic thresholds without the disadvantages of mixture models. They proposed several parametric distribution families that they referred to as *extended generalized Pareto distributions* (EGPDs). These families preserve the asymptotic properties of the GPD while allowing greater flexibility in the bulk

of the distribution. Each EGPD is constructed through a probability integral transform $X = G^{-1}(V; \boldsymbol{\theta})$, where $G(\cdot; \boldsymbol{\theta})$ is the generalized Pareto cdf of parameters $\boldsymbol{\theta} = (\sigma, \xi)$ and V is a random variable taking values in $(0, 1)$ with distribution $F_V(\cdot; \boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ is a parameter vector. Using the notation of Papastathopoulos and Tawn (2013), the cdf of X is given by

$$K(x; \boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathbb{P}\left\{G^{-1}(V; \boldsymbol{\theta}) \leq x\right\} = F_V\left\{G(x; \boldsymbol{\theta}); \boldsymbol{\lambda}\right\}, \tag{6.4}$$

for all $x \in (0, \infty)$, and the corresponding pdf is given by

$$k(x; \boldsymbol{\theta}, \boldsymbol{\lambda}) = f_V\left\{G(x; \boldsymbol{\theta}); \boldsymbol{\lambda}\right\} \times g(x; \boldsymbol{\theta}), \tag{6.5}$$

for all $x \in (0, \infty)$. When V is not a uniform random variable, the resulting distribution K differs from the GPD distribution. To ensure that the EGPD has the same tail behavior as G, Papastathopoulos and Tawn (2013) imposed that the following limiting ratio $\lim_{x \to x^*} \bar{K}(x; \boldsymbol{\theta}, \boldsymbol{\lambda}) / \bar{G}(x; \boldsymbol{\theta})$ is positive and finite. This EGPD approach provides smooth extensions of the GPD that require neither many additional parameters nor *prior* knowledge of the subasymptotic distribution, in contrast to mixture models.

Papastathopoulos and Tawn (2013) proposed the three following distributions F_V satisfying the tail equivalence constraint:

1. $F_V(v; \kappa) = v^\kappa, \ \kappa > 0$
2. $F_V(v; \kappa) = \beta\left\{1 - (1 - v)^{|\xi|}, \kappa, |\xi|^{-1}\right\}, \ \kappa > 0, \ \xi \in \mathbb{R}$
3. $F_V(v; \kappa) = 1 - \gamma\left\{-\log(1 - v), \kappa\right\}, \ \kappa > 0$

where β and γ stand for the regularized incomplete beta and regularized lower incomplete gamma functions respectively. These transformations of the uniform distribution do not affect the tail of the GPD, which retains the same tail index ξ, while the additional parameter κ enables the GPD to be modified subasymptotically. Another important property of these models is that they all reduce to the GPD when $\kappa = 1$.

These three families of EGPDs corresponding to the choice of F_V all have either a null or an infinite density at the lower end point of the support when $\kappa \neq 1$, *i.e.* when the EGPD does not simplify to the GPD. This is a practical constraint since in many application where the POT methodology is used, the density at the origin is positive and finite. For instance, let Y be a random variable with continuous density f_Y defined on the support interval I. The conditional density above the threshold $u \in I$ evaluated at u is as follows:

$$f_{(Y|Y \geq u)}(u) = f_Y(u) / \bar{F}_Y(u), \tag{6.6}$$

where $\bar{F}_Y(\cdot)$ denotes the survival function of Y. For a threshold u in the open interval I and for common random variables, this last quantity is strictly positive.

In the present paper, we develop a new EGPD family whose density can take any nonnegative value at the origin. The properties of this family are studied on several simulated datasets with different tail behavior and compared to the original EGPD proposed by Papastathopoulos and Tawn (2013) in Section 6.3. This new EGPD model is particularly suitable for modeling heatwave intensities, as shown in Section 6.4. We believe that the model is also advantageous in many other applications where subasymptotic thresholds are defined.

Our contribution in the present paper is twofold. On the one hand, we add a new EGPD family to the list proposed by Papastathopoulos and Tawn (2013). EGPDs are already used in various contexts. For example, they are used in state-of-the-art hydrological studies to model the distribution of overall rainfall (*e.g.* see Naveau et al., 2016; Blanchet et al., 2018; Evin et al., 2018; Tencaliec et al., 2020; Taillardat et al., 2019; Taillardat and Mestre, 2020), river discharge (*e.g.* see Li et al., 2020) and daily temperature maxima (*e.g.* see Stein, 2020). Given that EGPDs are already quite popular, we believe that this contribution will be of interest in several domains of environmental statistics. On the other hand, the application for modeling heatwave intensity is innovative, as well as scientifically and practically relevant.

6.2 Extended generalized Pareto distributions

6.2.1 Set of conditions for a new EGPD

To develop a new EGPD with a positive conditional density at the origin (see Eq. 6.6), we establish a set of 3 conditions that the model should satisfy.

Condition 1: The EGPD density at the origin should take any nonnegative value.

This condition results from Eq. (6.6). In the probability integral transform framework proposed by Papastathopoulos and Tawn (2013), this condition is equivalent to the following (see Eq. 6.5):

$$f_V(0; \boldsymbol{\lambda}) > 0. \tag{C1}$$

The distributions proposed by Papastathopoulos and Tawn (2013) do not respect Condition 1, as their density is either zero or infinite at the origin outside the special case of the GPD. This last property makes these EGPDs particularly adapted to model the whole distribution of positive variables as Naveau et al. (2016) proposed, as in this case, the density at the lower bound is often zero or infinite. For this reason, the aforementioned EGPDs are often used in hydrology to model the whole distribution. Nevertheless, Condition 1 is desirable

in standard applications of the POT methodology when exceedances above a subasymptotic threshold are considered.

Condition 2: Tail equivalence.

To ensure that the EGPD has the same tail behavior as the GPD family, Papastathopoulos and Tawn (2013) characterized the class of valid transition functions f_V using the Von Mises condition. This condition is also maintained by Naveau et al. (2016), as follows:

$$\frac{\bar{K}(x;\boldsymbol{\theta},\boldsymbol{\lambda})}{\bar{G}(x;\boldsymbol{\theta})} = \frac{\bar{F}_V(1-v;\boldsymbol{\lambda})}{v} \to a > 0 \quad \text{when } v \to 0. \tag{C2}$$

This last condition is equivalent to the following:

$$\lim_{v\to 1^-} f_V(v;\boldsymbol{\lambda}) = a > 0. \tag{C2-b}$$

Condition 3: Asymptotic behavior.

To ensure that the EGPD does not disturb the GPD behavior far in the tail, the following condition may be desirable:

$$\lim_{v\to 1^-} f'_V(v;\boldsymbol{\lambda}) = 0. \tag{C3}$$

Therefore, at high values, V is close to the uniform distribution, and the EGPD is practically equal to the GPD.

6.2.2 Truncated normal EGPD

The three conditions (C1) to (C3) lead us to propose the normal distribution with mean 1 and precision $\tau > 0$ truncated on the $(0,1)$ interval:

$$f_V(v;\tau) = \frac{2\tau}{2\Phi(\tau)-1}\phi\left\{\tau(v-1)\right\}, \tag{6.7}$$

where ϕ (resp. Φ) is the pdf (resp. cdf) of the standard normal distribution. On the basis of this distribution, the EGPD cdf (Eq. 6.4) is given by

$$K(x;\sigma,\xi,\tau) = \frac{2}{2\Phi(\tau)-1}\left[\Phi\left\{\tau\left(G(x;\sigma,\xi)-1\right)\right\} - \left\{1-\Phi(\tau)\right\}\right], \tag{6.8}$$

for any $x \geq 0$ and the EGPD pdf (Eq. 6.5) is given by

$$k(x;\sigma,\xi,\tau) = \frac{2\tau}{2\Phi(\tau)-1}\phi\left[\tau\left\{G(x;\sigma,\xi)-1\right\}\right] \times g(x;\sigma,\xi) \tag{6.9}$$

for any $x \geq 0$.

It can be verified (see Appendix A) that conditions (C1) to (C3) are satisfied by the truncated normal distribution expressed in Eq. (6.7). Moreover, the truncated normal unit expectation comes from Condition (C3): the density derivative at 1 is null when the mean is 1. Additionally, as τ tends to 0, the distribution f_V converges to the uniform law and the GPD is recovered. Thus, we can extend the definition of Eq. (6.7) with the uniform density when $\tau \to 0$. More details on the construction of F_V are available in Appendix A. Before modeling high temperatures with this distribution, the properties and the estimation procedure are studied in the following section.

6.3 Simulation study

6.3.1 Experiments framework

To experimentally validate the truncated normal model described above, $M = 100$ samples of size $n = 4000$, are drawn from several known distributions F_X with different shape parameters. Then, the model is fitted by maximum likelihood to each of these samples for a range of thresholds u. For each threshold u, the mean and the standard deviation over the M estimates are considered. This Monte Carlo procedure enables to compare the estimated parameters to the real ones and also assess the precision of the fit.

6.3.2 Intrinsic quality of the EGPD model

According to the EVT, the estimate of the shape should not depend on the threshold. For distributions whose shape parameter ξ^* is known (*Fréchet*, *Weibull*, *Exponential*...), the absolute error of the estimated shape parameter is a good indicator of the model's ability to estimate the extremes of the distribution. The evolution of this bias with the threshold u will be investigated. The estimated precision parameter τ is expected to decrease toward 0 as the threshold u increases and becomes suitable for the GPD. The scale parameter is expected to evolve linearly with u for large thresholds, with a slope depending on the sign of the shape (see Coles (2001) page 75 for more details).

Weibull data

Samples from a *Weibull* distribution with shape parameter $\xi^* = -0.1$ are considered in this section. The estimates of the truncated normal EGPD parameters are presented in Figure 6.1. The scale parameter σ_u is decreasing as the threshold u increases. This is consistent with the

EVT, which ensures that the scale must decrease linearly with the threshold asymptotically. The shape appears to be slightly underestimated for subasymptotic thresholds; however, it converges to ξ^* for high thresholds. Finally, the precision parameter τ decreases toward 0 with the threshold value, which is also consistent with theory, as the GPD model becomes justified asymptotically.

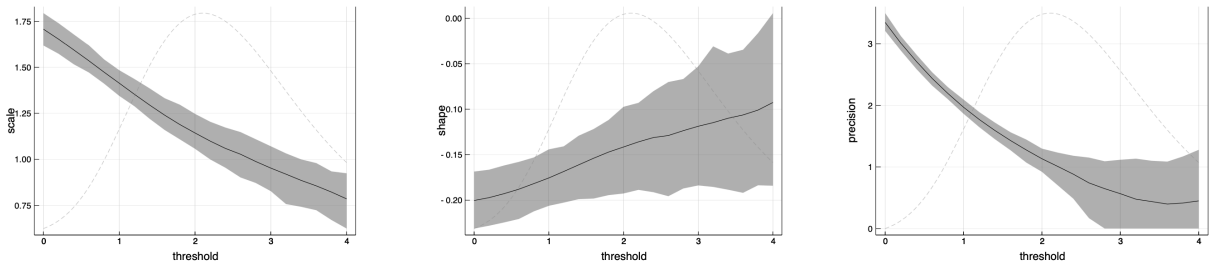

Figure 6.1 Estimate of the scale (left), shape (center) and precision (right) parameters of the truncated normal EGPD model fitted on *Weibull*-tail simulated data using varying thresholds. The mean value is indicated by a solid line, while the range of the estimates is shaded to show the uncertainty in the estimation.

The overall ability of the model to fit the tails is assessed in Figure 6.2 which shows the quantile-quantile plot and estimated density of the EGPD model for several thresholds on a particular arbitrarily chosen sample. The proposed EGPD is able to model a wide variety of excesses over a subasymptotic threshold.

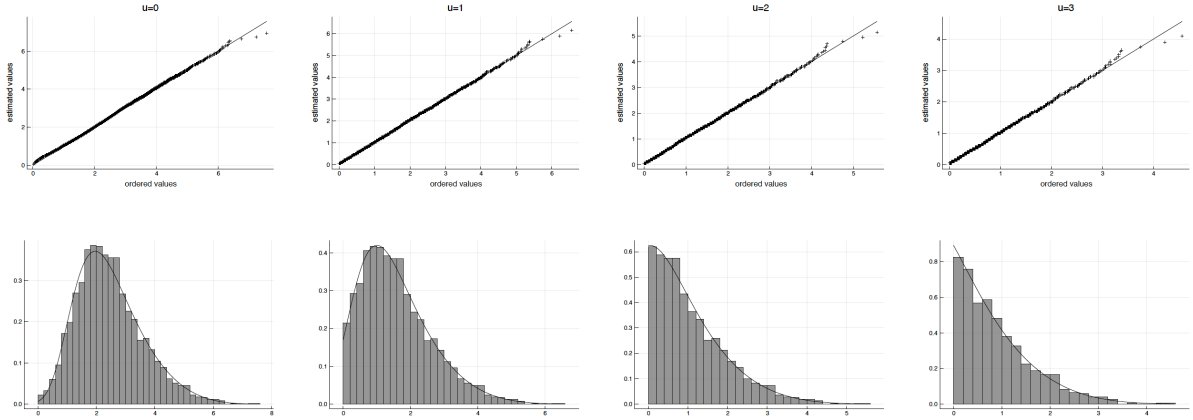

Figure 6.2 Quantile-quantile plots and densities of the fitted truncated normal EGPD on *Weibull* tails obtained with several thresholds.

Fréchet data

Samples from a *Fréchet* distribution with shape parameter $\xi^* = 0.1$ are now considered. As shown in Figure 6.3, the scale variation is not monotonous with the threshold. The scale is decreasing for subasymptotic thresholds and slightly increasing for higher thresholds. This result makes sense, as the scale is supposed to increase asymptotically when the shape is positive. As for the *Weibull* distribution, the shape is underestimated for subasymptotic thresholds. However, it converges to the appropriate value as the threshold increases. Finally, the precision parameter τ behaves similarly to that in the case of *Weibull* tails.

The overall quality of the fit is assessed in Figure 6.4, which presents a quantile-quantile plot and estimated densities for a particular sample. Again, the EGPD captures a wide variety of excess distributions, even though the underestimation of the shape for subasymptotic thresholds leads to underestimation of the higher quantiles, which is visible in the first quantile-quantile plots of Figure 6.4.

6.3.3 Comparison with other EGPD models

The developed truncated normal EGPD model is compared to the EGPD described in Papastathopoulos and Tawn (2013), which uses a power function $f_V(v) = v^{\kappa-1}$. Both models have the same number of parameters. These EGPD models are also compared to the GPD model which has one less parameter, which should be asymptotically better. Comparisons with the two other models proposed by Papastathopoulos and Tawn (2013) are not shown as they yielded very similar results.

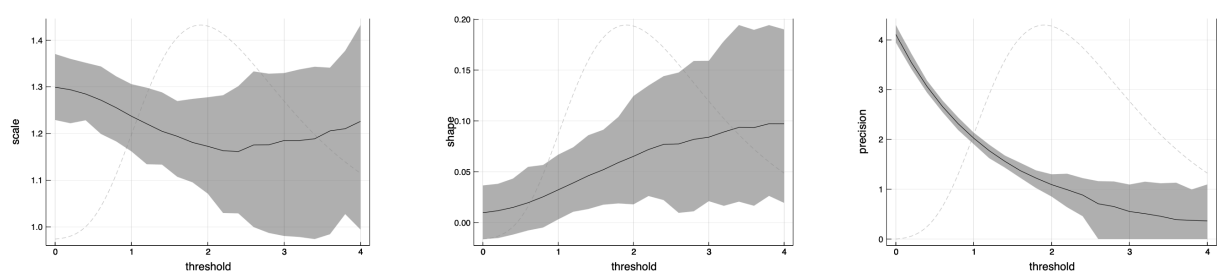

Figure 6.3 Estimates of the scale (left), shape (center) and precision (right) parameters of the truncated normal EGPD model fitted on *Fréchet*-tail simulated data using varying thresholds. The mean value is indicated by a solid line, while the range of the estimates is shaded to show the uncertainty in the estimation.

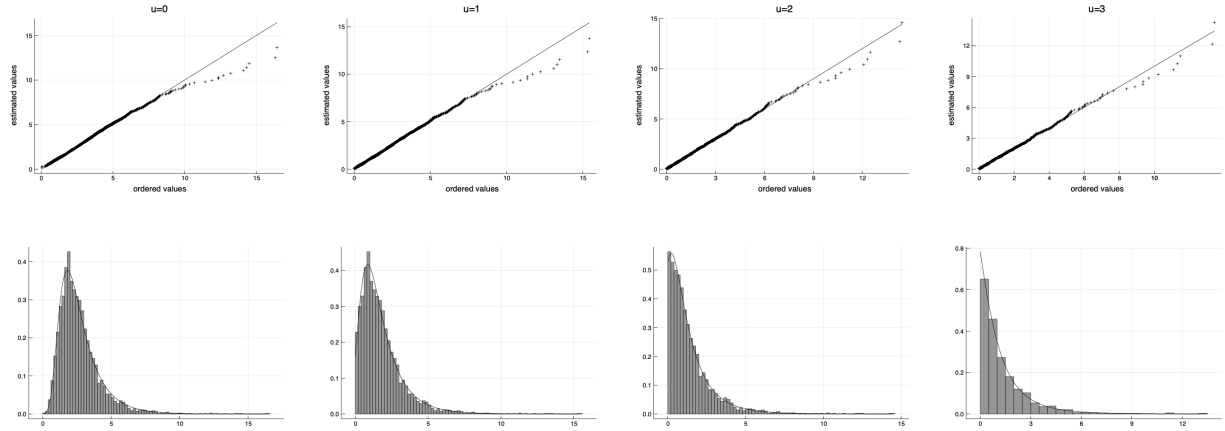

Figure 6.4 Quantile-quantile plots and densities of the fitted truncated normal EGPD on *Fréchet* tails obtained with several thresholds.

Shape parameter estimate

The shape parameter estimates of the three models are plotted to evaluate their asymptotic bias. Figure 6.5 reveals that all the models underestimate the shape for subasymptotic thresholds, both for *Fréchet* and *Weibull* tails. However, the bias is smaller for the truncated normal EGPD than for the other two. Moreover, the truncated normal EGPD appears to converge to the real value ξ^*, while the two others continue to underestimate the shape for high thresholds.

Bayesian information criterion

The Bayesian information criterion (BIC) provides an overview of the global quality of the fit while taking into account the complexity of the model. As the absolute value of the BIC depends on the number of observations and thus the threshold value, Figure 6.6 compares the BIC of the three models after normalizing the BIC values with the mean and standard deviation of the BIC values of the three models. The results of these experiments on *Weibull* and *Fréchet* tails are similar. For subasymptotic thresholds, both EGPD models are much better than the GPD model in terms of BIC, while asymptotically, the GPD outperforms EGPDs because the additional parameter (τ or κ) is unnecessary. Notably, if both EGPDs have similar BIC for asymptotic thresholds, the truncated normal model outperforms the power model for subasymptotic thresholds. It appears that the gap between both EGPDs is the largest for thresholds near the mode of the underlying distribution f_X.

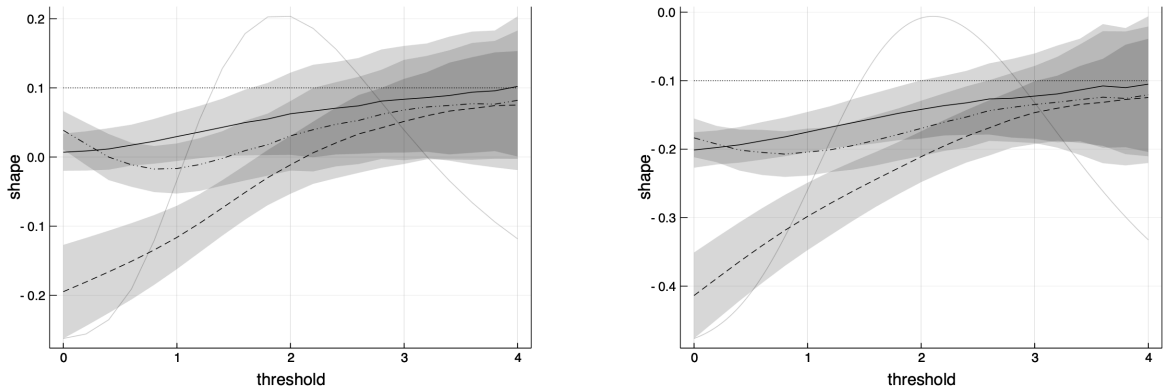

Figure 6.5 Comparison of the shape parameter estimates of the GPD (dashed), power EGPD (dot-dashed) and truncated normal EGPD (solid) fitted on *Fréchet*-tails (left) and *Weibull*-tails (right) using thresholds ranging from 0 to 4.

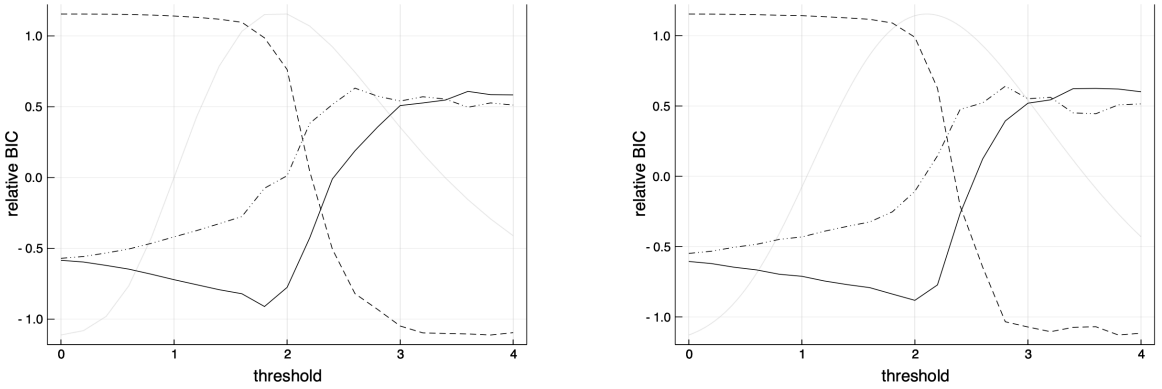

Figure 6.6 Standardized BIC of the GPD (dashed), power EGPD (dot-dashed) and truncated normal EGPD (solid) fitted on *Fréchet*-tails (left) and *Weibull*-tails (right) using thresholds ranging from 0 to 4.

RMSE on extreme quantiles

Working with simulated data allows one to benefit from knowledge of the underlying distribution. The real quantiles of the distribution f_X are compared to the estimates of these quantiles from the EGPD models and the GPD. Let $p \in [0,1]$ be a probability and $u \in [0, x^*]$ be a threshold. We denote by z_p the quantile of order $1-p$ of f_X defined by $F_X(z_p) = 1-p$, and $\zeta_u = \mathbb{P}(X > u)$. Then,

$$\mathbb{P}(X \leq z_p) = 1 - \zeta_u \mathbb{P}(X \geq z_p \mid X \geq u) \tag{6.10}$$

and

$$\mathbb{P}(X \leq z_p \mid X \geq u) = 1 - p\zeta_u^{-1} \tag{6.11}$$

which leads finally to

$$\mathbb{P}(X - u \leq z_p - u \mid X \geq u) = 1 - p\zeta_u^{-1}. \tag{6.12}$$

Therefore, the quantile of order $1-p$ of F_X equals the quantile of order $1 - p\zeta_u^{-1}$ of the tail distribution above the threshold u plus the value of u. Let us define z_p^u such that $\mathbb{P}\left(X - u \leq z_p^u \mid X \geq u\right) = 1 - p$. The quantile of order $1-p$ of the underlying distribution of X is

$$z_p = u + z^u_{p\zeta_u^{-1}} \tag{6.13}$$

which allows us to derive from this expression an estimator $\hat{z}_p$ using the estimated tail distribution. The RMSE between a set of extreme quantiles z_p of orders $1-p$ with $p \in \mathcal{P}$ and their corresponding estimate $\hat{z}_p$ is defined by

$$\mathrm{RMSE}(u, \mathcal{P}) = \sqrt{\frac{1}{|\mathcal{P}|} \sum_p (z_p - \hat{z}_p)^2}. \tag{6.14}$$

The quantiles of order $1-p$ of the truncated normal EGPD are obtained using Eq. (6.8), which leads to

$$z_p = G^{-1}_{\sigma,\xi}\left\{1 + \frac{1}{\tau}\Phi^{-1}\left(\frac{1}{2}\left[1 - p\left(2\Phi(\tau) - 1\right)\right]\right)\right\}. \tag{6.15}$$

Figure 6.7 plots the RMSEs for the set of quantiles $\mathcal{P} = \left\{\frac{1}{10\times 2^k}, 2 \leq k \leq 10\right\}$ for *Weibull* and *Fréchet* tails. This plot corresponds to return periods ranging from 40 to 10240 years, assuming that there is one observation of the random variable X per year. This particular choice of logarithmic scale for the return period set was made to assess the quality of the fit at various levels of extremes. These figures both show that the truncated normal EGPD model provides a better estimate of extreme quantiles, reaching a slightly lower RMSE than

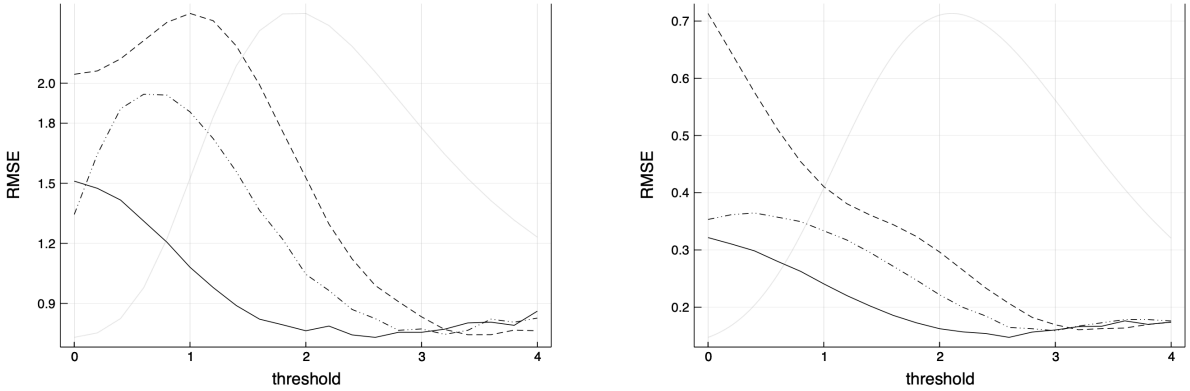

Figure 6.7 RMSE on extreme quantiles estimated with the GPD (dashed), power EGPD (dot-dashed) and truncated normal EGPD (solid) fitted on excesses of 100 samples of size 4000 with *Fréchet*-tails (left) and *Weibull*-tails (right) using thresholds ranging from 0 to 4.

the two other models. Moreover, this minimum RMSE is reached for a lower threshold. It emphasizes the fact that even when the interest is in extreme levels, an EGPD may give better results than the GPD, benefiting from subasymptotic values for inference. It appears, on these examples, that the optimal threshold is approximately at the mode of the underlying distribution, but no general conclusions can be made without further investigations.

Furthermore, this optimal threshold for the truncated normal EGPD is lower when the sample length is reduced (see Figure 6.8, where a sample size of 1000 was used). Indeed, when the amount of data is reduced, it is even more important to keep as much data as possible for inference and, thus, to reduce the threshold using an appropriate EGPD.

6.4 Application to heatwave intensities

The study of heatwaves has generated considerable interest recently. Severe heatwave events have raised serious concerns and led to projections for the years to come (see for example Lau and Nath, 2012; Habeeb et al., 2015; Chavaillaz et al., 2019). Many of the studies on heatwaves tend to characterize mean values or trends; only a few attempt to fully describe the distribution of heatwave characteristics, such as intensity, duration and frequency. Furrer et al. (2010) proposed a generative approach to model heatwaves, while Keellings and Waylen (2014) used a descriptive approach, modeling only the intensity of heatwaves and not all the daily maximum temperatures. In both approaches, the GPD is used to model the heatwave intensity. Both approaches examined the presence of trends, but while Keellings and Waylen (2014) divided the ensemble of observations in two equal time periods, Furrer et al. (2010)

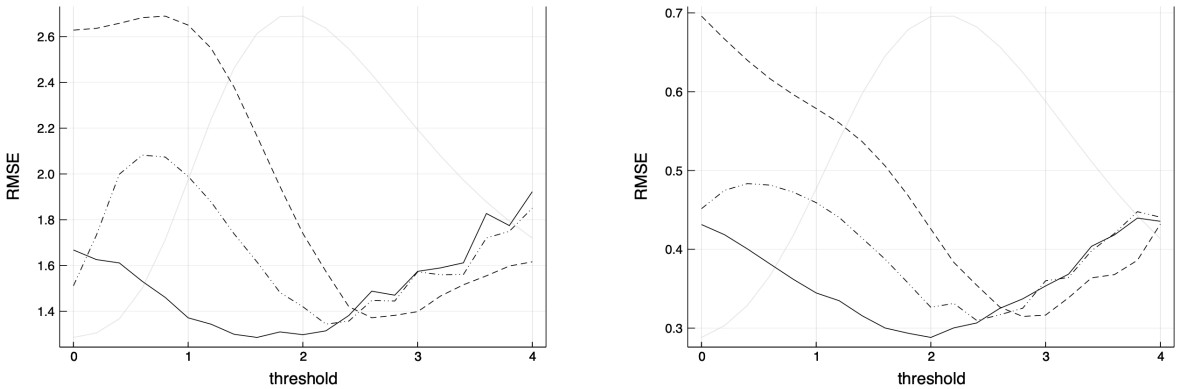

Figure 6.8 RMSE on extreme quantiles estimated with the GPD (dashed), power EGPD (dot-dashed) and truncated normal EGPD (solid) fitted on excesses of 100 samples of size 1000 with *Fréchet*-tails (left) and *Weibull*-tails (right) using thresholds ranging from 0 to 4.

opted for a transient modeling, introducing covariates in the parameters of the model.

6.4.1 Data

Such studies on heatwaves are generally based on climate model simulations since they represent the only tools for providing physically based quantitative projections of future climate. The simulated temperature of two ensembles of climate simulations are analyzed in this paper: the ClimEx ensemble (Leduc et al., 2019) and the CESM ensemble (Kay et al., 2015).

The ClimEx ensemble consists of a 50-member set of hourly climate simulations ranging from 1955 and to 2100 performed under the RCP8.5 greenhouse gases emission scenario and driven by the CanESM2 global model. It covers the northeastern part of North American and Europe with a regular grid of an approximate resolution of 12 km.

The CESM ensemble consists of a set of 30 climate simulations performed with the Community Earth System Model. The simulations range from 2006 to 2080 and were obtained using the RCP8.5 emission scenario. Unlike ClimEx, CESM is a global model; consequently, its spatial resolution is larger (approximately 1° horizontal resolution).

6.4.2 Method

The POT model is adapted to the context of heatwaves. First, the declustering scheme is tailored for the heatwave definition. Second, the cluster intensity is modeled with the EGPD developed in Section 6.2. Last, the models for the occurrences and intensities are extended

to the nonstationary context. Notably, the thresholds remain constant for the whole period, as does the heatwave definition.

Declustering

For many applications, such as those in health sciences, both maximum and minimum daily temperatures are considered to define a heatwave. This leads to a two-threshold definition: a heatwave is a streak of days during which the maximum temperature exceeds the higher threshold and the minimum temperature exceeds the lower threshold. In addition, a relief period is often considered. This period corresponds to a minimum time gap between two heatwaves during which at least one of the thresholds is not exceeded. This generic definition of heatwaves corresponds to the one given by Keellings and Waylen (2014) and adapts to many applications by defining different threshold values and relief period.

In this application, heatwaves are defined as consecutive days during which the temperature exceeds 33.5°C and stays above 21°C, using a relief period of 3 days. This definition is motivated by the study of Giroux et al. (2017), which showed a strong correlation between excess mortality and heatwaves defined in such a manner in the province of Quebec.

Occurrence model

The occurrence of a heatwave is modeled as a Poisson point process separately for each simulated year. Since the model is fitted on a large ensemble of simulations, many realizations of a given year are available to estimate the occurrence of a heatwave: 50 realizations in the case of ClimEx and 30 in the case of CESM. This approach frees us from explicitly modeling nonstationarity.

Intensity model

Heatwave intensity is defined by the maximum temperature reached during the heatwave. Since the higher threshold could be subasymptotic or become subasymptotic as the global temperature increases, the truncated normal EGPD described in Section 6.2 is proposed to model the heatwave intensity. To take into account the nonstationarity, the temporal covariate ($C_t : 1955 \leq t \leq 2100$) is introduced to the scale, shape and precision parameters:

$$\sigma_t = \exp(\sigma_0 + \sigma_1 C_t) \tag{6.16}$$

$$\xi_t = \xi_0 + \xi_1 C_t \tag{6.17}$$

$$\tau_t = \exp(\tau_0 + \tau_1 C_t). \tag{6.18}$$

The shape parameter is often modeled as stationary to avoid nonidentifiability issues. However, considering the large amount of data available for inference due to the use of the simulated ensemble, we can consider more complex models with multiple trends.

The time-varying covariate C_t consists of the atmospheric equivalent CO2 concentration in year t given by the RCP8.5 scenario – the emission scenario under which the simulations were generated. The equivalent CO2 is directly related to the long-term warming of the atmosphere (see for example Matthews et al., 2009; Leduc et al., 2016).

6.4.3 Results

ClimEx Ensemble

Heatwaves simulated by the ClimEx ensemble are modeled with the proposed model for the grid point containing the city of Montreal. Figure 6.9 shows the number of heatwaves per year for the 50 members. Clearly, the higher threshold of 33.5°C is crossed more and more often as the years advance.

The parameter estimates of the truncated normal EGPD for the heatwave intensities are given in Table 6.1. These estimates were obtained via the maximum likelihood method. According

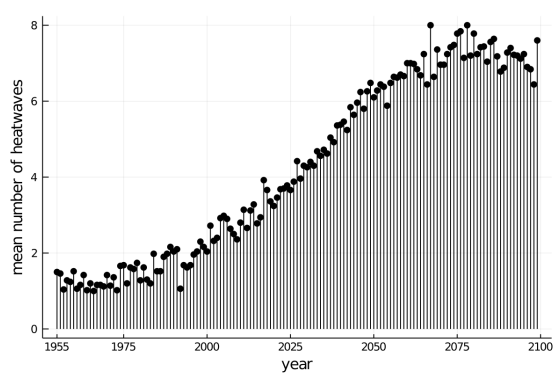

Figure 6.9 Number of heatwaves per year at the grid point containing the city of Montreal extracted from the 50 runs of the ClimEx ensemble.

to the BIC, the nonstationary parameters for all parameters are significant. The precision parameter τ increases over time, so the EGPD differs from the GPD as the global temperature increases. Thus, the heatwave intensities are less statistically extreme in the future climate. Moreover, the scale parameter increases significantly, which moves the intensities to more extreme temperatures.

Figure 6.12 shows the histograms of heatwave intensities and the corresponding estimated density for the whole time period partitioned into 5 and 10-year time slices. The density was evaluated with the parameters corresponding to the mid-interval. With such short 10-year time slices, it is reasonable to assume that the climate is approximately stationary. The estimated densities fits the data well. Furthermore, even in the beginning of the simulation period, a GPD could not have been adapted to model the heatwave intensities as the higher threshold is not sufficiently high. A quantile-quantile plot for the same time slices is shown in Figure 6.13. The fit is fairly good, and no systematic bias is observed in the tail, giving credit to the shape parameter estimate.

Figure 6.10 shows the effective return levels (see Katz et al., 2002) of heatwave intensities computed by combining the intensity modeling with an occurrence model (see Appendix B). Figure 6.11 matches the return period of 2020 to the one of 2050 (see Equation 18 of Appendix B). One can see, for instance, that a 100-year heatwave in 2020 occurs on average once every 8 years in 2050.

CESM ensemble

For the CESM ensemble, the model is fitted to the grid point containing the city of New York. Heatwaves are extracted using the same thresholds, and the truncated normal EGP model is fitted to their intensities. Figure 6.14 shows the number of heatwaves per year.

The truncated normal EGPD parameter estimates are given in Table 6.2. The scale and precision parameters are both increasing, while the shape parameter is decreasing, and according to the BIC, these trends are all significant. At the beginning of the period, the truncated normal EGPD is not far from the GPD as $\tau_{2006} = 0.57$. However, as years pass, the heatwave intensity differs more and more from the GPD.

Table 6.1 Parameter estimates of the truncated normal EGPD for heatwave intensities on ClimEx.

σ_0	σ_1	ξ_0	ξ_1	τ_0	τ_1
1.10	0.32	-0.22	-0.03	0.17	0.04

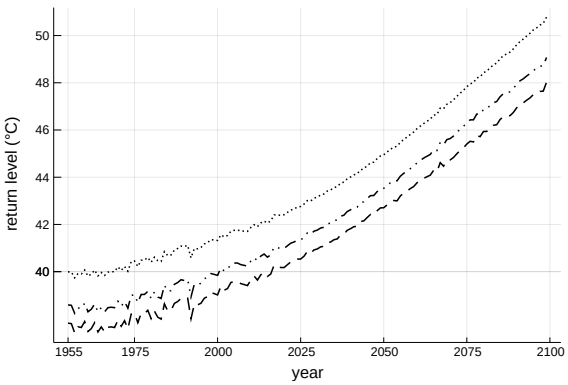

Figure 6.10 Effective return levels of periods of 10 (dashed), 20 (dot-dashed) and 100 years (dotted) as a function of the time estimated in Montreal with ClimEx.

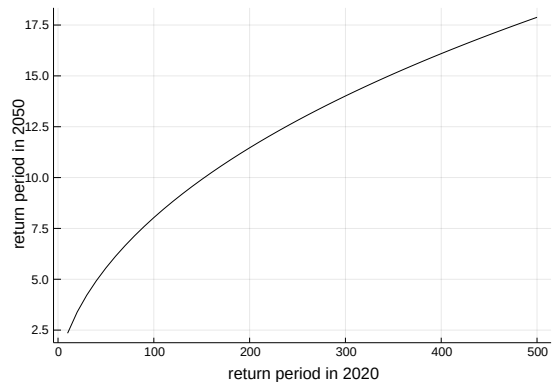

Figure 6.11 Correspondence between the return period of 2020 and the one of 2050 estimated in Montreal with ClimEx.

Table 6.2 Parameter estimates of the truncated normal EGPD for heatwave intensities on CESM.

σ_0	σ_1	ξ_0	ξ_1	τ_0	τ_1
0.34	0.31	-0.10	-0.04	-0.04	0.36

Density and quantile-quantile plots are given in Figures 6.17 and 6.18 for 10-year time slices spanning the simulation period. The GPD is suitable at the beginning of the simulation period but becomes unsuitable as the density mode begins to shift. Again, the quantile-quantile plot shows no systematic bias. Figure 6.15 shows the increasing 10, 20 and 100 year effective return levels in New York between 2006 and 2080. Figure 6.16 matches the return period of 2020 to the one of 2050. Here, we expect the 100 year return level of 2020 to occur more often than once every 4 years in 2050 and the 500 year return level of 2020 to occur once a decade in 2050.

6.5 Conclusion

A new extended generalized Pareto model based on the truncated normal distribution is developed. This model is particularly suited to model excess over a subasymptotic threshold. It extends the POT framework to subasymptotic levels, complementing the work from Papastathopoulos and Tawn (2013), who introduced the EGPD models. We proved in Section 6.2 that our model has all the required properties to extend the GPD and is sufficiently flexible to fit a wide variety of tails. We highlighted in Section 6.3 the accuracy of our model on simulated data. The proposed model performed better than other EGPD models on sub-

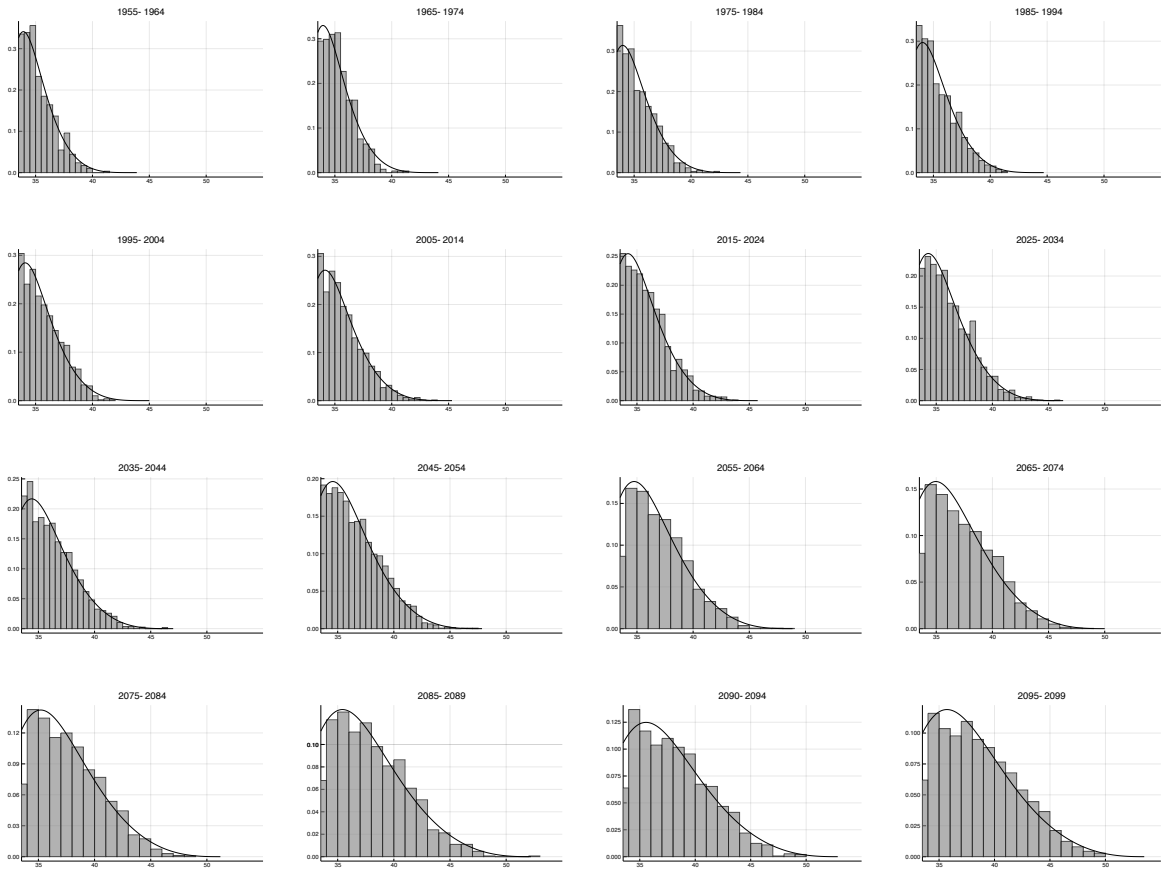

Figure 6.12 Comparison of estimated densities and ClimEx data (Montreal)

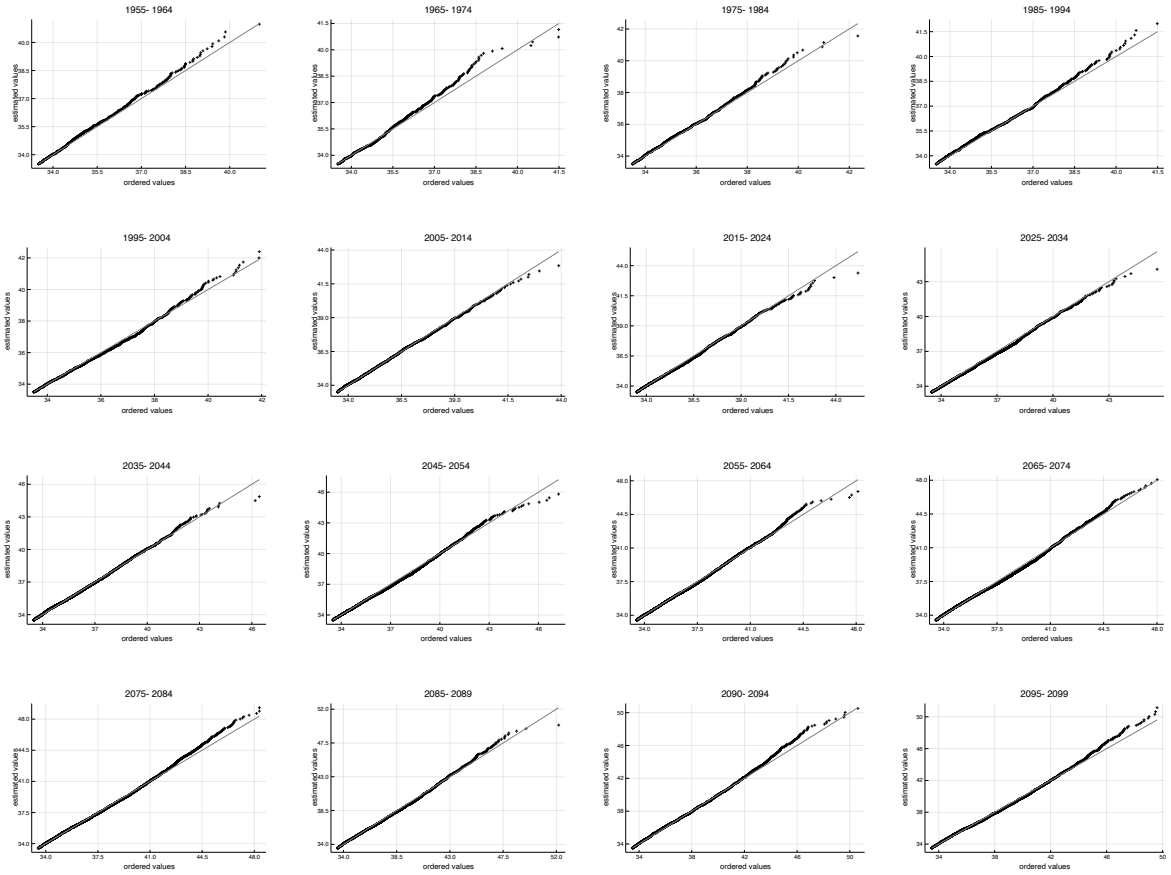

Figure 6.13 Quantile-quantile plot, ClimEx data (Montreal)

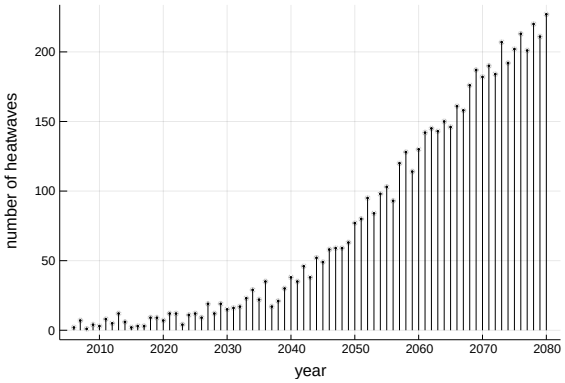

Figure 6.14 Number of extracted heatwaves in the grid point containing New York using the 30 runs of the CESM ensemble.

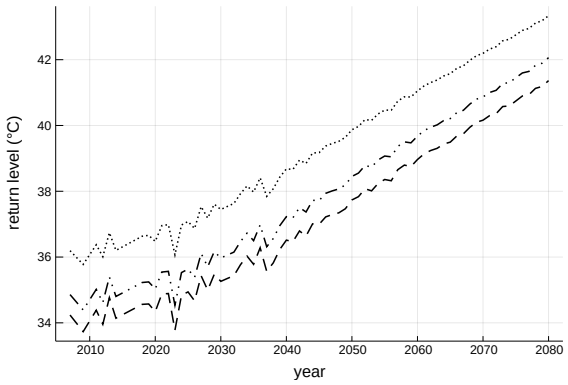

Figure 6.15 Effective return levels of 10 (dashed), 20 (dot-dashed) and 100 year periods (dotted) as a function of the time estimated in New York with the CESM ensemble.

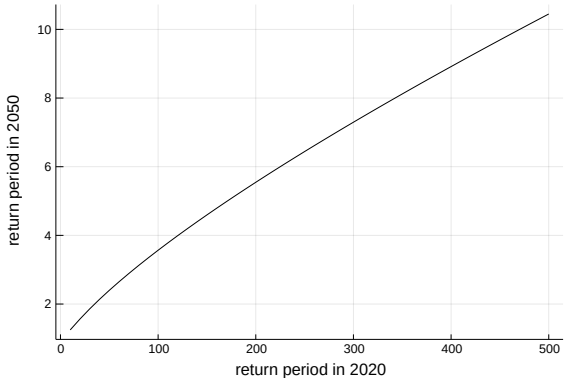

Figure 6.16 Correspondence between the return period of 2020 and the one of 2050 estimated in New York with the CESM ensemble.

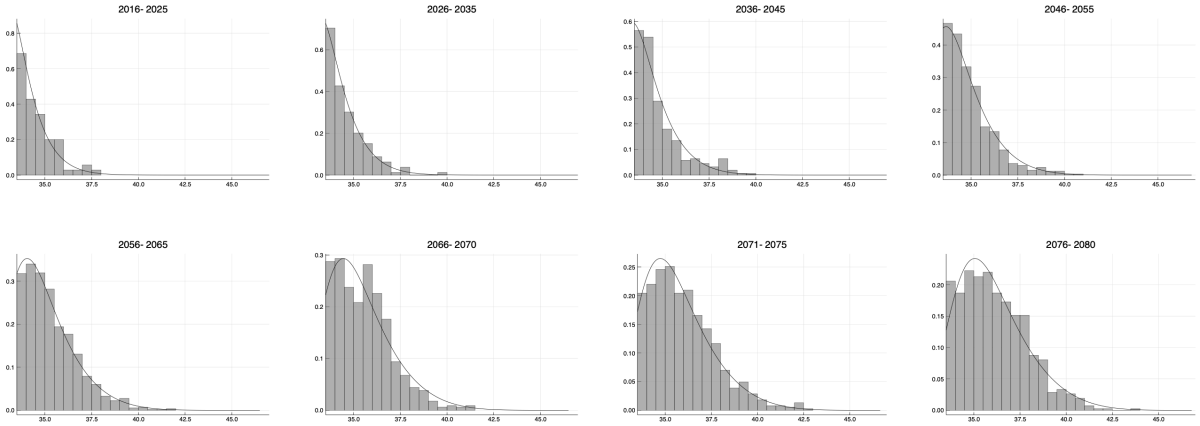

Figure 6.17 Comparison of estimated densities and CESM data (New York)

asymptotic thresholds. Moreover, even though we are interested in extreme levels, the use of EGPD models with subasymptotic thresholds may lead to a better estimate of the tail, especially when we have limited data. Finally the proposed truncated normal EGPD was used to model the intensities of heatwaves in a nonstationary context. The developed EGPD model was particularly suited to this application and produced good transient estimates of heatwave intensities.

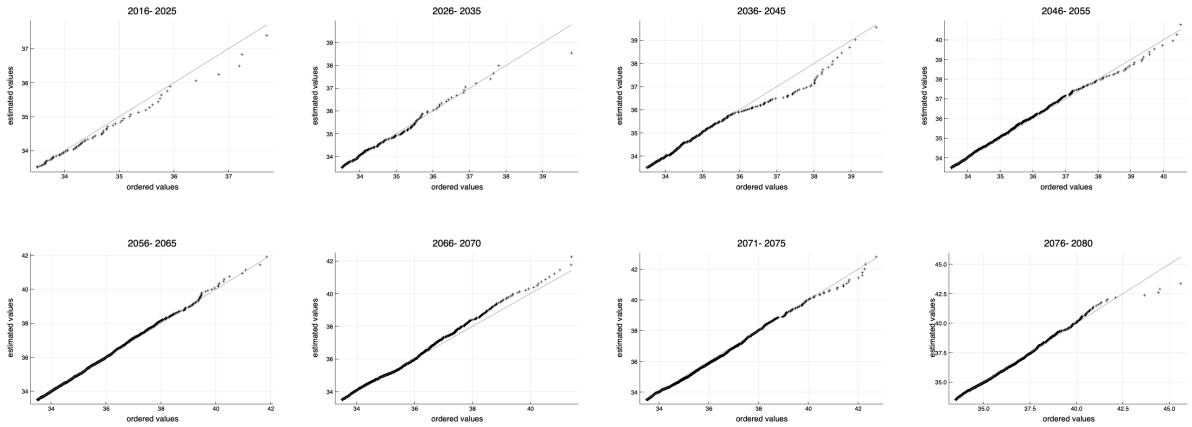

Figure 6.18 Quantile-quantile plot, CESM data (New York)

CHAPITRE 7 MODÉLISATION DE LA DURÉE DES VAGUES DE CHALEUR

7.1 Modélisation conditionnellement à l'intensité

L'intensité des vagues de chaleur a été modélisée au chapitre 6 dans un contexte non-stationnaire. La sévérité d'une vague de chaleur dépend aussi de la durée de cette dernière. Il convient donc de modéliser la loi conjointe du couple durée-intensité. La figure 7.1 illustre la répartition des couples durée-intensité des vagues de chaleur simulées de l'ensemble ClimEx. Il apparaît que ces deux variables sont grandement corrélées et que cette dépendance doit par conséquent impérativement être modélisée. Une première approche consiste à modéliser indépendamment les lois marginales de l'intensité (variable continue) et la durée (variable discrète) des vagues de chaleur et d'utiliser ensuite une fonction copule pour modéliser la dépendance. Cependant cette approche est problématique lorsqu'une des deux marginales est une loi discrète puisque la fonction copule n'est alors pas unique. Nous modéliserons donc plutôt la durée conditionnellement à l'intensité. Une fois les modèles pour l'intensité M et pour la durée conditionnellement à l'intensité $D \mid \{M - u\}$ ajustés, la loi conjointe pour le couple (M, D) est donnée pour $z > 0$ et $d \in \mathbb{N}^*$ par

$$f_{(M,D)}(u + z, d) = \mathbb{P}\left(D = d \mid M - u = z\right) k(z) \tag{7.1}$$

où k est la densité de l'intensité de la forme décrite au Chapitre 6.

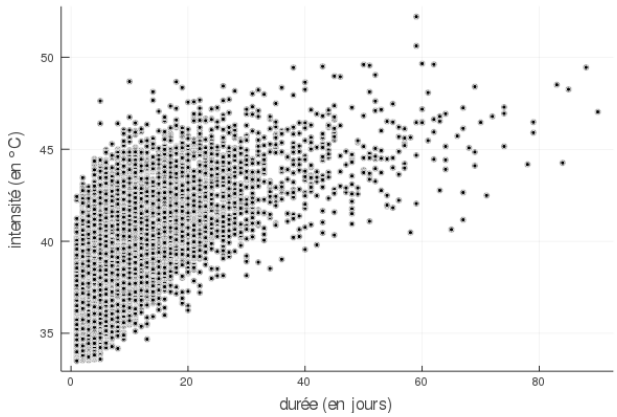

Figure 7.1 Distribution jointe des couples durée-intensité pour l'ensemble des vagues de chaleur simulées de l'ensemble ClimEx

7.2 Analyse exploratoire

La durée des vagues de chaleur doit être modélisée par une loi discrète dont les paramètres dépendent à la fois du temps grâce à une covariable C_t et de l'intensité de la vague de chaleur. Il existe de nombreuses lois discrètes (e.g. géométrique, binomiale, binomiale négative...) et une analyse exploratoire des données de ClimEx est nécessaire pour en sélectionner une.

Afin de voir l'influence de l'intensité sur la distribution de la durée, on fixe des périodes temporelles assez courtes sur lesquelles il est raisonnable de supposer le climat stationnaire. La moyenne et l'écart-type de la durée des vagues de chaleur de chaque période temporelle considérée sont ensuite calculés par fenêtre de 1°C. La figure 7.2 illustre l'évolution de la moyenne et de l'écart-type de la durée fonction de l'intensité des vagues de chaleur pour les périodes 1990-2014, 2045-2064 et 2080-2099. Il apparaît que la moyenne comme l'écart-type augmentent avec l'intensité, et cette augmentation est de plus en plus grande en climat futur. Ces remarques suggèrent d'une part que la distribution de $D \mid \{M - u\}$ n'est pas stationnaire, et d'autre part que pour une année t fixée, l'espérance conditionnelle $\mathbb{E}_t\left(D \mid M - u = z\right)$ et la variance conditionnelle $\mathbb{V}_t\left(D \mid M - u = z\right)$ sont des fonctions croissantes de z.

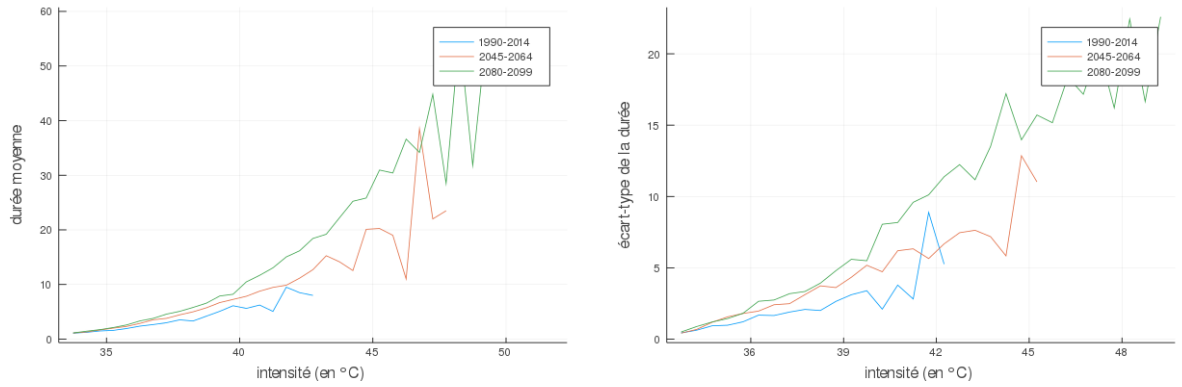

Figure 7.2 Moyenne (gauche) et écart-type (droite) de la durée des vagues de chaleur fonction de leur intensité sur trois périodes temporelles distinctes.

L'étude des deux premiers moments de la distribution n'est pas suffisant pour déterminer la forme de la distribution conditionnelle de la durée. Afin d'obtenir plus d'information, la figure 7.3 montre la densité empirique de la durée pour différentes intensités sur une période temporelle donnée. Il apparaît que la distribution s'étire vers la droite lorsque l'intensité augmente confirmant l'étude des deux premiers moments. De plus, pour des petites intensités le mode de la distribution est 1 tandis que pour des plus grandes intensités le mode se décale vers des plus grandes valeurs. Cette observation élimine toutes les distributions dont le mode a une valeur fixée comme la loi géométrique. Enfin, la figure 7.4 montre l'écart moyenne-

variance fonction de l'intensité pour diverses période temporelles. Il apparaît que pour des grandes intensité cet écart diverge. Cette observation élimine la loi de Poisson pour laquelle la variance est égale à la moyenne. La loi binomiale négative est une alternative intéressante

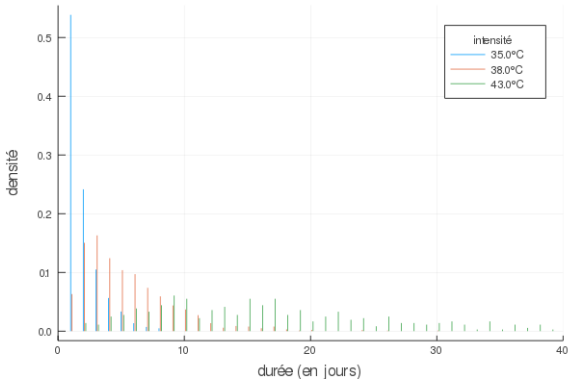

Figure 7.3 Distribution empirique de $D|\{M = m\}$ pour $m \in \{35, 38, 43\}$ entre 2075 et 2099.

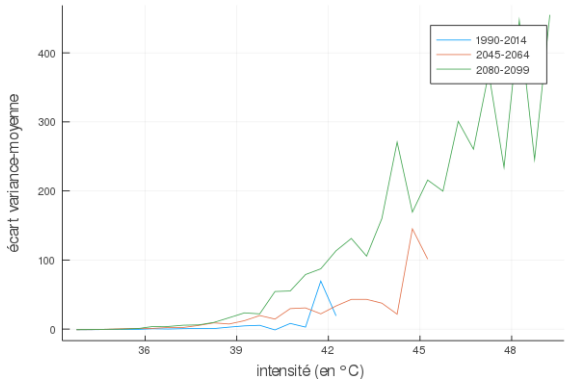

Figure 7.4 Écart moyenne-variance pour la durée fonction de l'intensité des vagues de chaleur.

à la loi de Poisson pour des données discrètes, non-bornées et dont la variance est plus grande que la moyenne. De plus, la loi binomiale négative englobe également le cas d'une loi géométrique et d'une loi de Poisson asymptotiquement puisque $\mathcal{G}(p) = NegBin(1, p)$ et $\mathcal{P}(\lambda) = \lim_{r\to\infty} NegBin\left(r, \frac{r}{\lambda+r}\right)$.

7.3 Paramétrisation du modèle

Une première analyse exploratoire suggère d'utiliser la loi binomiale négative pour modéliser la durée des vagues de chaleur conditionnellement à leur intensité. Il apparaît aussi que cette distribution doit dépendre à la fois du temps et de la valeur de l'intensité de la vague de chaleur. Ces dépendances seront modélisées à travers les deux paramètres de la loi binomiale négative. Le support de la loi binomiale négative est $\mathbb{N}$ alors que les durées D sont à valeurs dans $\mathbb{N}^*$, la variable $D - 1$ sera plutôt considérée. Ainsi nous posons pour tout $z > 0$:

$$D - 1 \mid \{M - u = z\} \sim NegBin\left\{\rho_t(z), \theta_t(z)\right\}, \tag{7.2}$$

où $\rho_t(z) > 0$, et $\theta_t(z) \in]0, 1[$.

Il apparaît sur la figure 7.2 que toute la masse de la distribution est en 1 lorsque l'excédent z au-dessus du seuil u tend vers 0. Cela signifie en effet que la probabilité d'observer plusieurs jours consécutifs où la température reste tout juste au dessus du seuil tend vers 0. On peut

donc poser

$$\rho_t(z) = \rho_t z, \tag{7.3}$$

ainsi on obtient bien $\lim_{z\to 0} \mathbb{E}_t\left(D \mid M - u = z\right) = 1$. Afin de modéliser l'évolution temporelle et de respecter la contrainte de positivité, nous poserons

$$\rho_t = \exp(\rho^0 + \rho^1 C_t), \tag{7.4}$$

où ρ^0 et ρ^1 sont deux paramètres réels à déterminer. Pour le second paramètre $\theta_t(z)$ nous poserons

$$\theta_t(z) = \frac{1}{2}\left\{1 + \tanh\left(\chi_t + \psi_t z\right)\right\}, \tag{7.5}$$

où $\chi_t = \chi^0 + \chi^1 C_t$, $\psi_t = \psi^0 + \psi^1 C_t$ et la fonction $(1+\tanh)/2$ garantit d'obtenir un paramètre dans l'intervalle $]0, 1[$.

7.4 Choix d'un modèle à l'aide du BIC

Le modèle décrit à la section ci-dessus est très général et certains paramètres peuvent être fixés à 0 afin de réduire le nombre de paramètres du modèle. Pour cela, plusieurs sous-modèles sont ajustés aux données et celui ayant le meilleur BIC est retenu. Pour simplifier les notations, posons $\boldsymbol{\lambda}_* = (\rho^0, \rho^1, \chi^0, \chi^1, \psi^0, \psi^1)$ et $\boldsymbol{\lambda}_{-i}$ le vecteur $\boldsymbol{\lambda}_*$ où la i-ième composante est remplacée par un 0. Pour des observations $(\boldsymbol{d}_t, \boldsymbol{m}_t)_{1\le t\le T}$, considérons la fonction de log-vraisemblance suivante

$$\ell\left(\boldsymbol{\lambda}_*; \boldsymbol{d}, \boldsymbol{m}\right) = \sum_{t=1}^{T}\sum_{i=1}^{n_t} \log f\left(d_{t,i} - 1 \mid m_{t,i} - u, \boldsymbol{\lambda}_*\right) \tag{7.6}$$

où $f\left(k \mid z, \boldsymbol{\lambda}\right)$ dénote la fonction de masse évaluée en k de la loi binomiale négative de paramètres $\rho = \rho_t z$ et $\theta = \frac{1}{2}\left\{1 + \tanh\left(\chi_t + \psi_t z\right)\right\}$.

7.4.1 Application à ClimEx

La log-vraisemblance (7.6) est maximisée avec les données de ClimEx en prenant en compte l'ensemble des paramètres possibles $\boldsymbol{\lambda}_*$ puis en fixant chacun des paramètres tour à tour. Les valeurs des log-vraisemblances obtenues ainsi que des BIC correspondants sont regroupées dans le tableau 7.1. L'analyse du tableau 7.1 indique d'une part que le modèle où le paramètre ψ^1 est fixé à 0 ($\boldsymbol{\lambda}_{-6}$) donne le meilleur BIC. De plus, on peut conclure qu'il s'agit du meilleur de tous les sous-modèles selon le BIC. En effet la log-vraisemblance maximisée de tous les modèles $\boldsymbol{\lambda}_{-i}$ pour $1 \le i \le 5$ est plus petite que le BIC du modèle $\boldsymbol{\lambda}_{-6}$ ce qui veut dire que tout modèle où l'un des 5 premiers paramètres est retiré aurait nécessairement un moins bon

Tableau 7.1 Résultats de la maximisation de la log-vraisemblance du modèle pour la durée des vagues de chaleur avec les données de ClimEx.

	$\boldsymbol{\lambda}_*$	$\boldsymbol{\lambda}_{-1}$	$\boldsymbol{\lambda}_{-2}$	$\boldsymbol{\lambda}_{-3}$	$\boldsymbol{\lambda}_{-4}$	$\boldsymbol{\lambda}_{-5}$	$\boldsymbol{\lambda}_{-6}$
ℓ	-56275	-56528	-56315	-56310	-56405	-56707	-56279
BIC	-56306	-56553	-56340	-56335	-56430	-56732	**−56304**

BIC. Le tableau 7.2 donne les paramètres estimés pour le modèle retenu. La figure 7.9 montre

Tableau 7.2 Paramètres du meilleur modèle estimé par l'EMV pour les données de ClimEx.

ρ^0	ρ^1	χ^0	χ^1	ψ^0
-0.61	-0.17	0.14	-0.21	-0.06

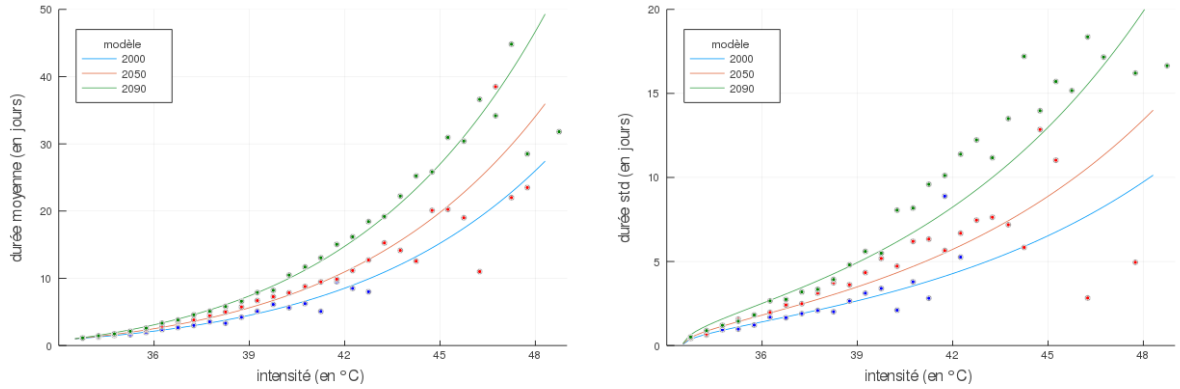

Figure 7.5 Moyenne (gauche) et écart-type (droite) de la durée des vagues de chaleur fonction de leur intensité estimés par le modèle en 2000, 2050 et 2090. Les points représentent les valeurs empiriques calculé sur des périodes temporelles de 20 ans centrées sur l'année d'estimation.

l'estimation de la moyenne et de la variance de la durée des vagues de chaleur en fonction de leur intensité, obtenue à partir des paramètres du tableau 7.2. Bien que la variance des estimations empiriques soit très grande pour des intensités élevées à cause du faible nombre d'observation, le modèle estimé est assez fidèle et reproduit bien les tendances observées sur la figure 7.2.

7.5 Ajustement du modèle par MCMC

Reprenons le modèle sélectionné par le BIC et estimons les paramètres de ce modèle en utilisant cette fois un algorithme MCMC. L'intérêt d'un MCMC par rapport à l'EMV est qu'il permet d'obtenir la loi *a posteriori* des paramètres du modèle. La loi *a posteriori* permet

d'obtenir des intervalles de confiances sur les paramètres du modèle, mais aussi sur toute quantité dépendant de ces paramètres. Les valeurs des paramètres de l'EMV sont prises comme valeurs initiales pour l'algorithme MCMC. Les tailles de pas sont ensuite ajustées afin d'obtenir un taux d'acceptation entre 0.4 et 0.7.

7.5.1 Application à ClimEx

La convergence de l'algorithme MCMC ainsi que les distributions *a posteriori* des paramètres $\boldsymbol{\lambda} = (\rho^0, \rho^1, \chi^0, \chi^1, \psi^0)$ du modèle sont illustrées par la Figure 7.6. Le taux d'acceptation pour

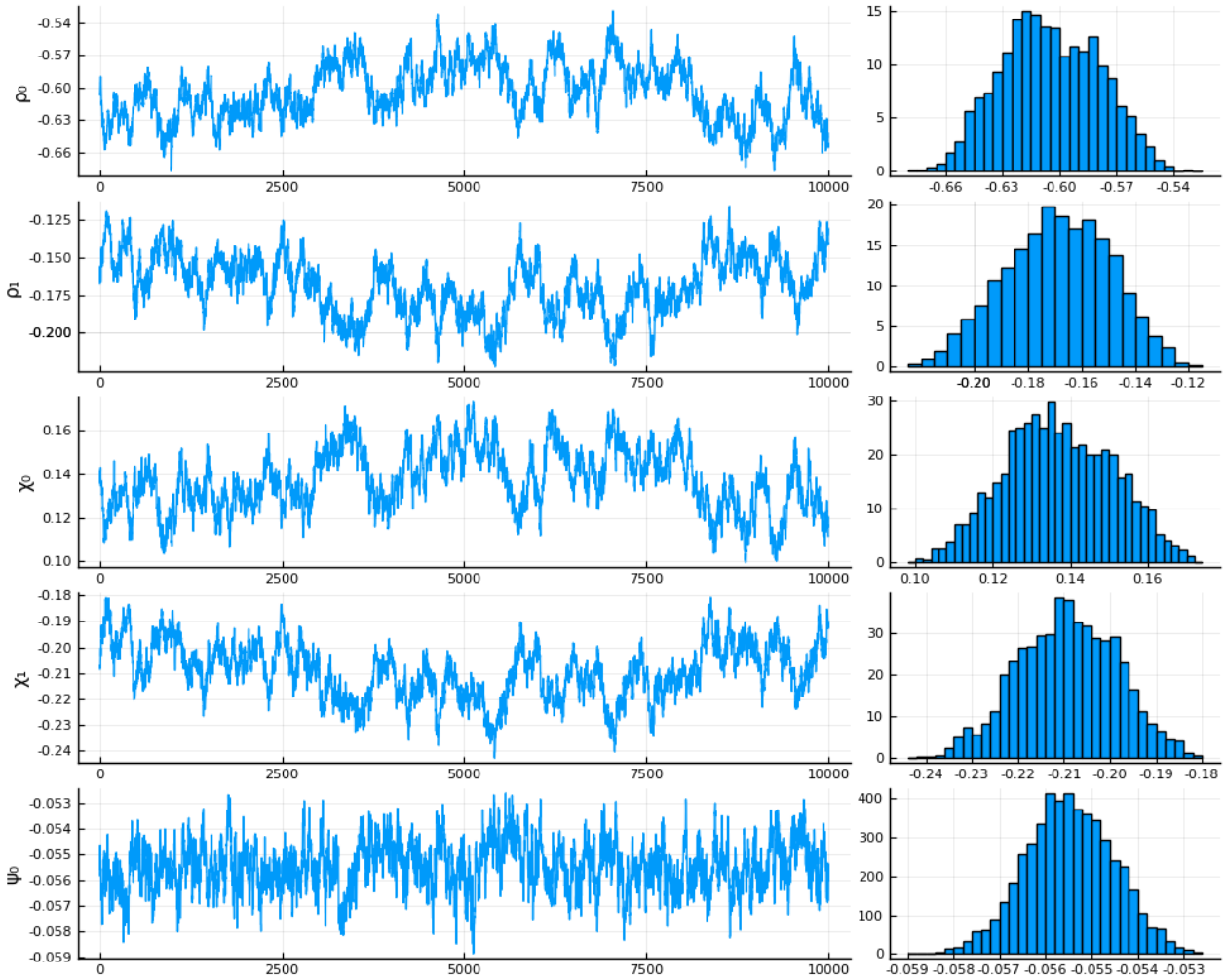

Figure 7.6 Estimation par MCMC du modèle pour les durée pour ClimEx. Chaines de Markov montrant la convergence de l'algorithme et densité de la loi *a posteriori* pour chaque paramètre.

chacun des paramètres sur l'ensemble des 10000 itérations effectuées est bien compris entre 0.4 et 0.7. Les histogrammes des échantillons de la loi *a posteriori* obtenus pour chacun

des paramètres du modèle permet de déterminer des intervalles de confiance basés sur les quantiles de ces échantillons. La qualité de l'estimation est vérifiée en comparant les densités conditionnelles des durées observées aux densités estimées par le modèle. La figure 7.7 montre les densités empiriques et estimées pour différentes périodes et différentes intensités. Il apparaît que le modèle est assez fidèle aux observations et parvient à modéliser des distributions assez variées.

A partir de l'échantillon obtenu par MCMC de la loi *a posteriori* nous pouvons obtenir la loi *a posteriori* de la moyenne à une année pour une certaine intensité. La figure 7.8 donne la durée moyenne des vagues de chaleur estimée à partir du MCMC, avec un intervalle de confiance à 95% basé sur les quantiles de la distribution *a posteriori*.

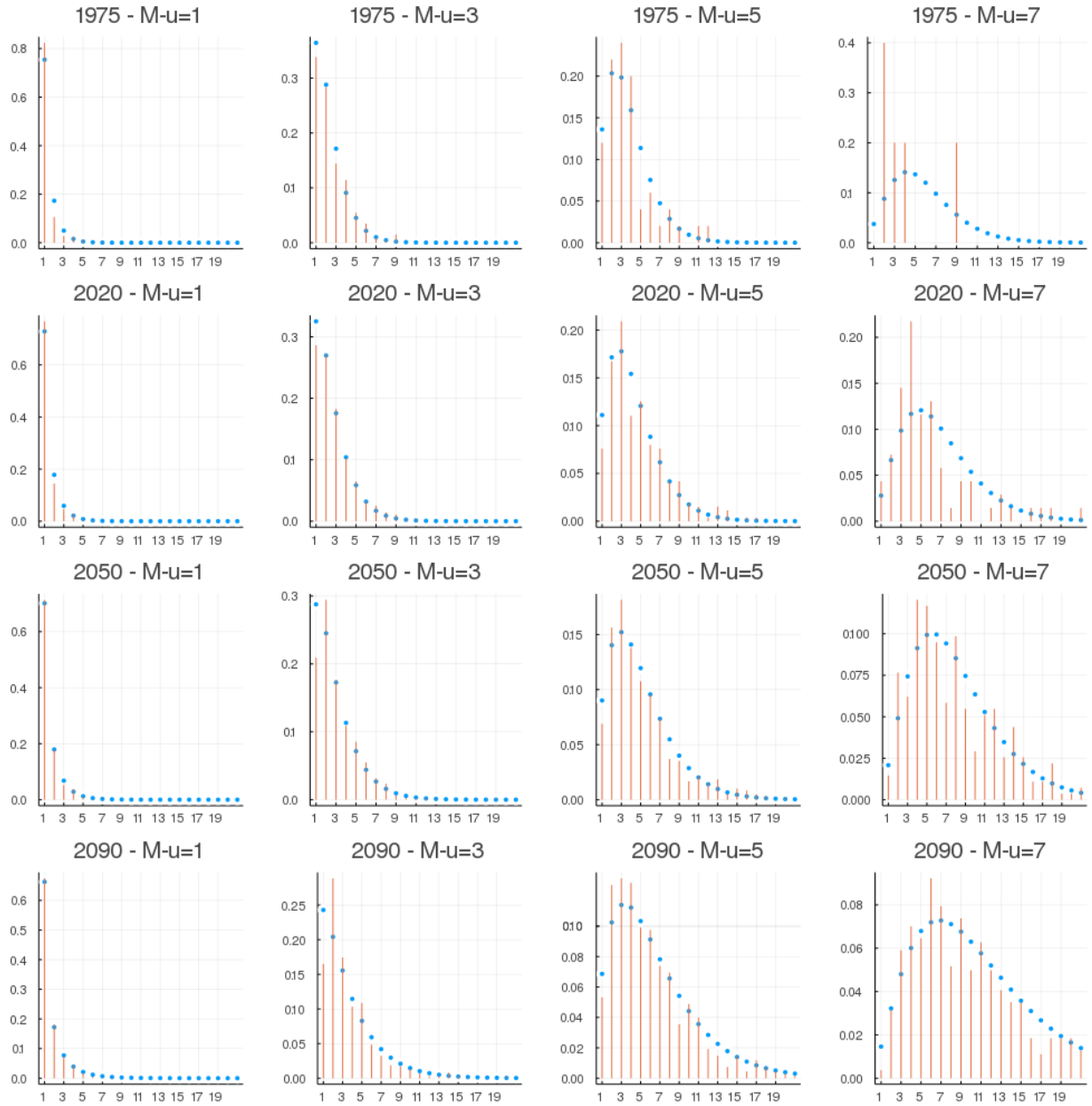

Figure 7.7 Diagnostic du modèle pour les durées estimé par MCMC. Comparaison des estimations (points) avec les données observées (barres verticales) sur des périodes de ±10 ans à une intensité donnée à ±0.5°C.

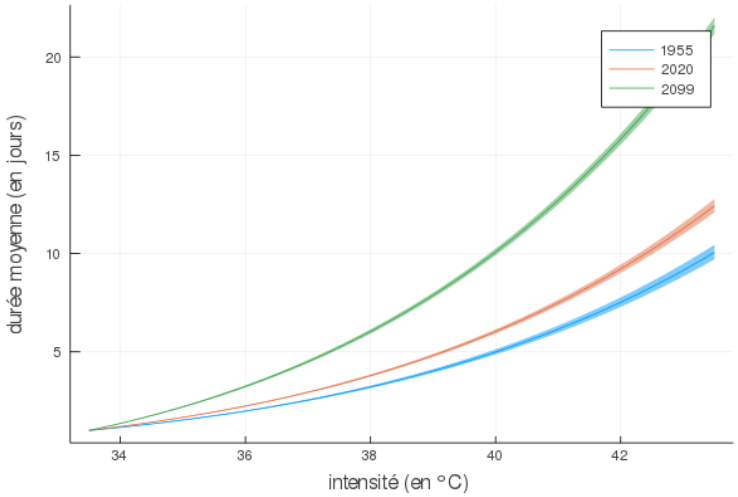

Figure 7.8 Évolution la durée moyenne des vagues de chaleur en fonction de leur intensité obtenu par MCMC avec un intervalle de confiance à 95%.

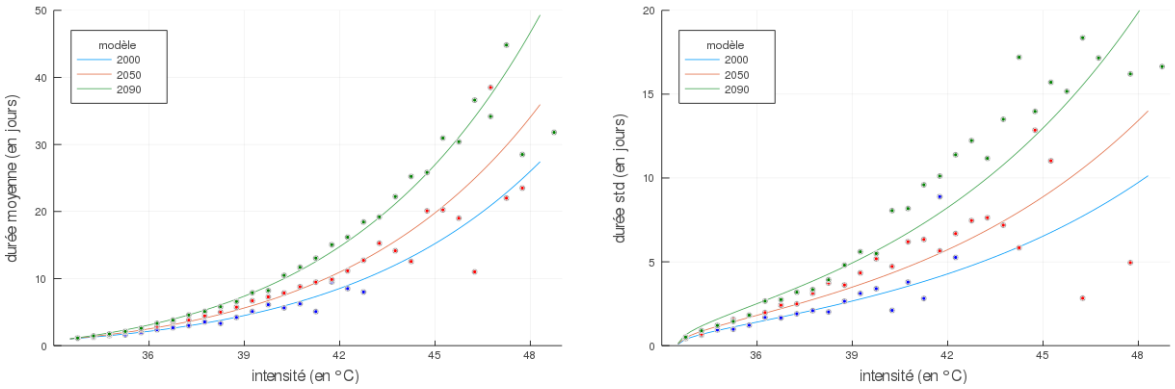

Figure 7.9 Moyenne (gauche) et écart-type (droite) de la durée des vagues de chaleur en fonction de leur intensité estimés par le modèle en 2000, 2050 et 2090. Les points représentent les valeurs empiriques calculées sur des périodes temporelles de 20 ans centrées sur l'année d'estimation.

CHAPITRE 8 MESURES DE RISQUE EN CLIMAT FUTUR

8.1 Modélisation des occurrences

Le nombre de vagues de chaleur par année de l'ensemble ClimEx est visiblement non-stationnaire (voir figure 6.9). Nous modélisons le nombre d'occurrence d'une année donnée N_t par une loi de Poisson de paramètre μ_t :

$$N_t \sim \mathcal{P}\left(\mu_t\right). \tag{8.1}$$

Pour chaque année, nous disposons d'une observation de cette variable aléatoire par simulation de l'ensemble ce qui nous permet d'estimer l'intensité μ_t directement sans modéliser la non-stationnarité.

8.1.1 Estimation par l'EMV

Il est possible d'utiliser l'EMV pour déterminer l'intensité de la loi de Poisson. L'intensité μ_t est donnée par la moyenne sur l'ensemble de simulation du nombre d'occurrence à l'année t, *i.e.*

$$\hat{\mu}_t = \frac{1}{K}\sum_{i=1}^{K} n_{t,i} \tag{8.2}$$

où K est le nombre de simulations de l'ensemble et $n_{t,i}$ le nombre de vagues de chaleur observée à l'année t pour la i-ème simulation. La Figure 8.1 montre l'évolution de l'intensité du processus d'occurrence en fonction du temps. On observe une augmentation de l'intensité d'une valeur proche de 1 en 1955 à près de 8 en 2100.

8.1.2 Estimation bayésienne

Il est également possible d'estimer l'intensité du processus de Poisson de manière bayésienne. Considérons les K observations $(n_{t,1}, \ldots, n_{t,K})$ de la variable aléatoire N_t. Si l'on considère l'intensité μ_t comme une variable aléatoire suivant une loi *a priori* gamma

$$\mu_t \sim \text{Gamma}(\alpha, \beta), \tag{8.3}$$

alors conditionnellement à μ_t, les observations sont bien distribuées comme une Poisson, *i.e.*

$$N_t \mid \mu_t \sim \mathcal{P}(\mu_t) \tag{8.4}$$

et la loi *a posteriori* du paramètre μ_t est

$$\mu_t \mid n_{t,1}, \ldots, n_{t,K} \sim \text{Gamma}\left(\alpha + \sum_{i=1}^{K} n_{t,i}, \beta + K\right). \tag{8.5}$$

Nous prenons $\alpha = \beta = 0$, ce qui correspond à une loi *a priori* non-informative et impropre. La figure 8.2 donne la densité de la loi *a posteriori* du paramètre d'intensité du processus d'occurrence pour différentes années couvrant la période d'analyse. On observe une augmentation de l'intensité avec le temps ainsi qu'une augmentation de l'incertitude sur l'estimation du paramètre μ_t.

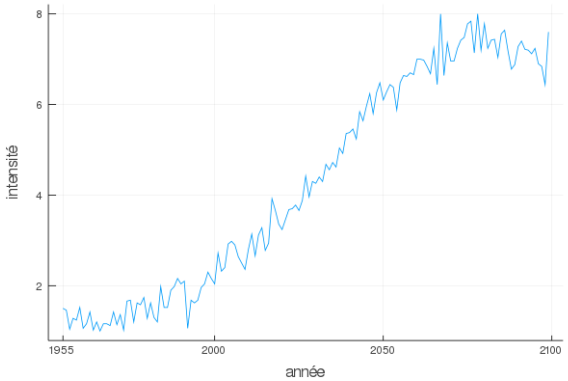

Figure 8.1 Évolution de l'intensité du processus d'occurrence (EMV).

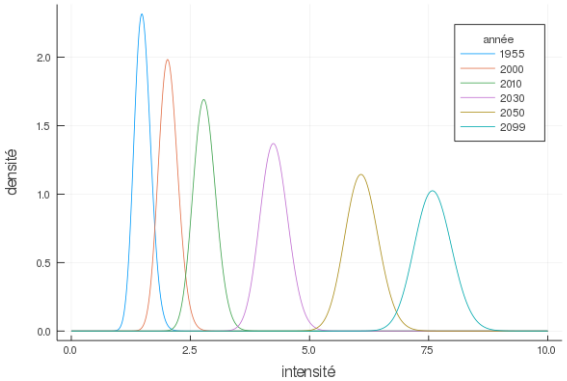

Figure 8.2 Évolution de l'intensité du processus d'occurrence (bayésien).

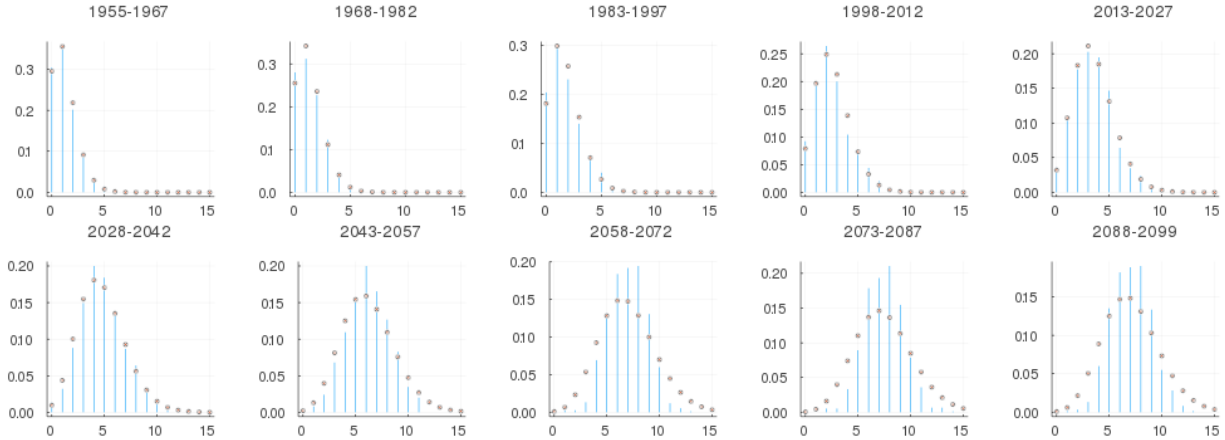

Figure 8.3 Diagnostic de la modélisation des occurrences par une loi de Poisson. La densité estimée par Monte Carlo à partir de la loi *a posteriori* (points rouges) est comparée à la densité empirique (barres bleues).

La figure 8.3 compare les densités estimées par le modèle Poisson aux observations de ClimEx.

Le modèle semble très bon jusqu'en 2050 puis semble légèrement sur-dispersé ensuite bien que la valeur moyenne reste fidèle.

8.2 Utilisation des lois *a posteriori* pour simuler des vagues de chaleur

L'étude des niveaux de retour est un moyen simple pour caractériser l'évolution des extrêmes d'une distribution non-stationnaire. Nous avons vu à la section 6 qu'en couplant l'étude des quantiles de la distribution étudiée à un modèle d'occurrence il était possible d'obtenir les niveaux de retour effectifs, c'est-à-dire le niveau effectivement susceptible de revenir avec une période de retour de T années en tenant compte du fait qu'il peut y avoir de multiples évènements une même année. Dans le cas d'un modèle d'occurrence suivant une loi de Poisson, nous avons vu qu'il était possible de calculer analytiquement les niveaux de retour effectifs.

La tâche devient plus complexe si l'on veut caractériser les extrêmes d'évènements bidimensionnels comme les couples intensité-durée des vagues de chaleur. Nous nous heurtons à deux difficultés principales : il n'existe pas de définition universelle de ce qu'est un extrême bidimensionnel, et ici l'une des deux distributions est discrète.

Nous chercherons ici à estimer la probabilité d'occurrence pour une année donnée d'une vague de chaleur dépassant une intensité z avec une durée d'au moins d jours. Pour un couple $(z, d) \in \mathbb{R}_+ \times \mathbb{N}^*$, notons $\mu_t(z, d)$ le taux d'occurrences du processus de Poisson filtré caractérisant l'occurrence de ces évènements. Notons que le taux μ_t du processus complet définie en (8.1) est obtenu en prenant $(z, d) = (0, 1)$.

Analytiquement, l'expression du taux d'occurrence du processus filtré est

$$\mu_t(z, d) = \mu_t \cdot \mathbb{P}_t \left(M - u \geq z, D \geq d\right). \tag{8.6}$$

Pour obtenir une expression à partir des lois estimées pour l'intensité et la durée, il est possible de conditionner selon la valeur de l'intensité :

$$\mu_t(z, d) = \mu_t \cdot \frac{\int_z^\infty \sum_{k=d}^\infty f\left(k - 1 \mid z, \boldsymbol{\lambda}_t\right) k\left(z; \boldsymbol{\theta}_t\right) \mathrm{d}z}{\int_z^\infty k\left(z; \boldsymbol{\theta}_t\right) \mathrm{d}z} \tag{8.7}$$

où f est la densité de la binomiale négative définie en (7.6), $\boldsymbol{\lambda}_t$ est le vecteur de paramètre pour la binomiale négative, k la densité du modèle EGPD et $\boldsymbol{\theta}_t$ le vecteur de paramètre de cette distribution. L'équation obtenue ne peut pas être estimée exactement avec les paramètres obtenus par l'EMV, et le calcul de l'incertitude par la méthode delta est complexe. Dans un contexte bayésien, il faudrait intégrer cette équation selon la loi *a posteriori* des paramètres $\boldsymbol{\lambda}_t$ et $\boldsymbol{\theta}_t$ ce qui mène aussi à une impasse. Nous nous contenteront plutôt d'une estimation

par Monte Carlo du taux $\mu_t(z, d)$.

8.3 Procédure Monte Carlo

L'objectif de l'approche Monte Carlo est de simuler des étés avec leurs vagues de chaleur à partir des lois *a posteriori* des paramètres du modèle fréquence-intensité-durée développé précédemment. L'utilisation des loi *a posteriori* permet également de caractériser l'incertitude lors des estimations. Pour la simulation d'un été, l'algorithme 2 est utilisée. Le résultat de cet algorithme est une liste S contenant un nombre N de couple intensité-durée représentant l'ensemble des vagues de chaleur d'un été.

Algorithme 2 Simulation d'un été

tirer $\mu_t \sim \text{Gamma}\left(\sum_{i=1}^{K} n_{t,i}, K\right)$
tirer $N \sim \mathcal{P}\left(\mu_t\right)$
for $1 \leq i \leq N$ **do**
  tirer $(\sigma^0, \sigma^1, \xi^0, \xi^1, \tau^0, \tau^1) \in \boldsymbol{\theta}_{\text{MCMC}}$
    $\sigma_t \leftarrow \exp(\sigma^0 + \sigma^1 C_t)$
    $\xi_t \leftarrow \xi^0 + \xi^1 C_t$
    $\tau_t \leftarrow \exp(\tau^0 + \tau^1 C_t)$
  tirer $Z_i \sim K\left(\cdot; \sigma_t, \xi_t, \tau_t\right)$
  tirer $(\rho^0, \rho^1, \chi^0, \chi^1, \psi^0) \in \boldsymbol{\lambda}_{\text{MCMC}}$
    $\rho_t(Z_i) \leftarrow \exp(\rho^0 + \rho^1 C_t) \cdot Z_i$
    $\theta_t(Z_i) \leftarrow \frac{1}{2}\left\{1 + \tanh\left(\chi^0 + \chi^1 C_t + \psi^0 Z_i\right)\right\}$
  tirer $D_i - 1 \sim NegBin\left(\rho_t(Z_i), \theta_t(Z_i)\right)$
end for
return $S = \{(Z_i, D_i) \mid 1 \leq i \leq N\}$

8.3.1 Estimation du taux d'occurrence

Le taux d'occurrence $\mu_t(z, d)$ est estimé à partir du nombre moyen par été simulé de vague de chaleur dépassant les seuils d'intensité et de durée définis. Soit M et B deux entiers. La méthode du boostrap consiste à simuler M étés selon l'algorithme 2, puis ré-échantillonner B fois à partir de l'échantillon simulé. Le taux $\mu_t(z, d)$ est ensuite estimé à partir des B échantillons bootstrap, ainsi qu'un intervalle de confiance basé sur les quantiles de la distribution des B estimations bootstrap. L'algorithme du bootstrap est détaillé par l'algorithme 3.

Algorithme 3 Estimation du taux d'occurrence

for $1 \leq m \leq M$ **do**
  Simuler un été $S_m = \{(Z_i, D_i) \mid 1 \leq i \leq N\}$ avec l'algorithme 2
  $n_m \leftarrow \sum_{i=1}^{N} \mathbf{1}_{Z_i \geq z} \cdot \mathbf{1}_{D_i \geq d}$
end for
for $1 \leq b \leq B$ **do**
  Créer un échantillon bootstrap $(n_1^*, \cdots, n_M^*)$ à partir de l'échantillon $(n_1, \cdots, n_M)$
  $\hat{\mu}_b \leftarrow \frac{1}{M} \sum_{m=1}^{M} n_m^*$
end for
$\hat{\mu} \leftarrow \frac{1}{B} \sum_{b=1}^{B} \hat{\mu}_b$
return $\hat{\mu}$
Utiliser les quantiles empiriques d'ordres $\alpha/2$ et $1-\alpha/2$ de l'échantillon $(\hat{\mu}_1, \cdots, \hat{\mu}_B)$ pour un intervalle de confiance à $(1-\alpha)\%$.

8.3.2 Estimation d'un niveau de retour

Nous pouvons définir comme étant un niveau de retour aux T années, à une année t donnée, un couple (z, d) tel que le taux d'occurrence à l'année t d'une vague de chaleur dépassant les niveaux d'intensité et de durée (z, d) soit de $1/T$. Définissons $z_d^T(t)$ tel que

$$\mu_t \left(z_d^T(t), d \right) = \frac{1}{T}. \tag{8.8}$$

Ce niveau s'interprète comme l'intensité atteinte en moyenne une fois toutes les T années par les vagues de chaleur d'au moins d jours avec le climat de l'année t. Ce niveau de retour peut être estimé à partir de l'algorithme 4.

Algorithme 4 Estimation d'un niveau de retour

Require: période de retour T, durée d, année t.
for $1 \leq m \leq M$ **do**
  Simuler un été $S_m = \{(Z_i, D_i) \mid 1 \leq i \leq N\}$ avec l'algorithme 2
  $I_m \leftarrow \{Z_i \mid D_i \geq d, 1 \leq i \leq N\}$
end for
for $1 \leq b \leq B$ **do**
  Créer un échantillon bootstrap $(I_1^*, \cdots, I_M^*)$ à partir de l'échantillon $(I_1, \cdots, I_M)$
  Soit Z^* la concaténation des valeurs de $(I_1^*, \cdots, I_M^*)$ triée par ordre décroissant
  $\hat{z}_b \leftarrow Z^*_{\lfloor M/T \rfloor}$
end for
$\hat{z} \leftarrow \frac{1}{B} \sum_{b=1}^{B} \hat{z}_b$
return Niveau de retour $\hat{z}$
Utiliser les quantiles empiriques d'ordres $\alpha/2$ et $1-\alpha/2$ de l'échantillon $(\hat{z}_1, \cdots, \hat{z}_B)$ pour un intervalle de confiance à $(1-\alpha)\%$.

8.3.3 Augmentation du taux d'occurrence

Définissons le facteur d'augmentation du taux d'occurrence $\mu_t(z,d)$ entre les années t_0 et t_1 :

$$\delta_{t_0,t_1}(z,d) = \frac{\mu_{t_1}(z,d)}{\mu_{t_0}(z,d)}. \tag{8.9}$$

En prenant z égal au niveau de retour à l'année de référence $z_d^T(t_0)$, on obtient

$$\delta_{t_0,t_1}\left(z_d^T(t_0),d\right) = T \times \mu_{t_1}\left(z_d^T(t_0),d\right), \tag{8.10}$$

ce qui définit le facteur d'augmentation à l'année t_1 du taux d'occurrence des vagues de chaleur d'au moins d jours qui avaient une période de retour de T années à l'année de référence t_0. Pour simplifier les notations nous noterons $\delta_{t_0,d}^T(t_1)$ ce facteur d'augmentation. Ce facteur d'augmentation est estimé avec l'algorithme 5.

Algorithme 5 Estimation du facteur d'augmentation

Require: période de retour T, durée d, année de référence t_0, année t.
  for $1 \leq m \leq M$ **do**
    Simuler un été de l'année de référence $S_m^0 = \{(Z_i, D_i) \mid 1 \leq i \leq N\}$ avec l'algorithme 2
    $I_m \leftarrow \{Z_i \mid D_i \geq d, 1 \leq i \leq N\}$
  end for
  for $1 \leq m \leq M$ **do**
    Simuler un été de l'année t, $S_m = \{(Z_i, D_i) \mid 1 \leq i \leq N\}$ avec l'algorithme 2
  end for
  for $1 \leq b \leq B$ **do**
    Créer un échantillon bootstrap $(I_1^*, \cdots, I_M^*)$ à partir de l'échantillon $(I_1, \cdots, I_M)$
    Soit Z^* la concaténation des valeurs de $(I_1^*, \cdots, I_M^*)$ triée par ordre décroissant
    $\hat{z}_b \leftarrow Z^*_{\lfloor M/T \rfloor}$
    Créer un échantillon bootstrap $(S_1^*, \cdots, S_M^*)$ à partir de l'échantillon $(S_1, \cdots, S_M)$
    Calculer pour chaque été de l'échantillon bootstrap n_m^* le nombre de vagues de chaleur dépassant les seuils $(\hat{z}_b, d)$
    $\hat{\delta}_b \leftarrow T \times \frac{1}{M} \sum_{m=1}^{M} n_m^*$

  end for
  $\hat{\delta} \leftarrow \frac{1}{B} \sum_{b=1}^{B} \hat{\delta}_b$
  return Facteur d'augmentation $\hat{\delta}$
  Utiliser les quantiles empiriques d'ordres $\alpha/2$ et $1-\alpha/2$ de l'échantillon $\left(\hat{\delta}_1, \cdots, \hat{\delta}_B\right)$ pour un intervalle de confiance à $(1-\alpha)\%$.

8.4 Résultats

8.4.1 Ajustement bayésien en un point

Les modèles bayésiens pour l'intensité du processus d'occurrence et pour les durées des vagues de chaleur ont déjà été ajustés pour les données de ClimEx au point de grille de Montréal. La figure 8.4 montre la convergence de l'algorithme MCMC sur 10000 itérations avec ces mêmes données. Nous disposons maintenant des lois *a posteriori* de l'ensemble des paramètres du

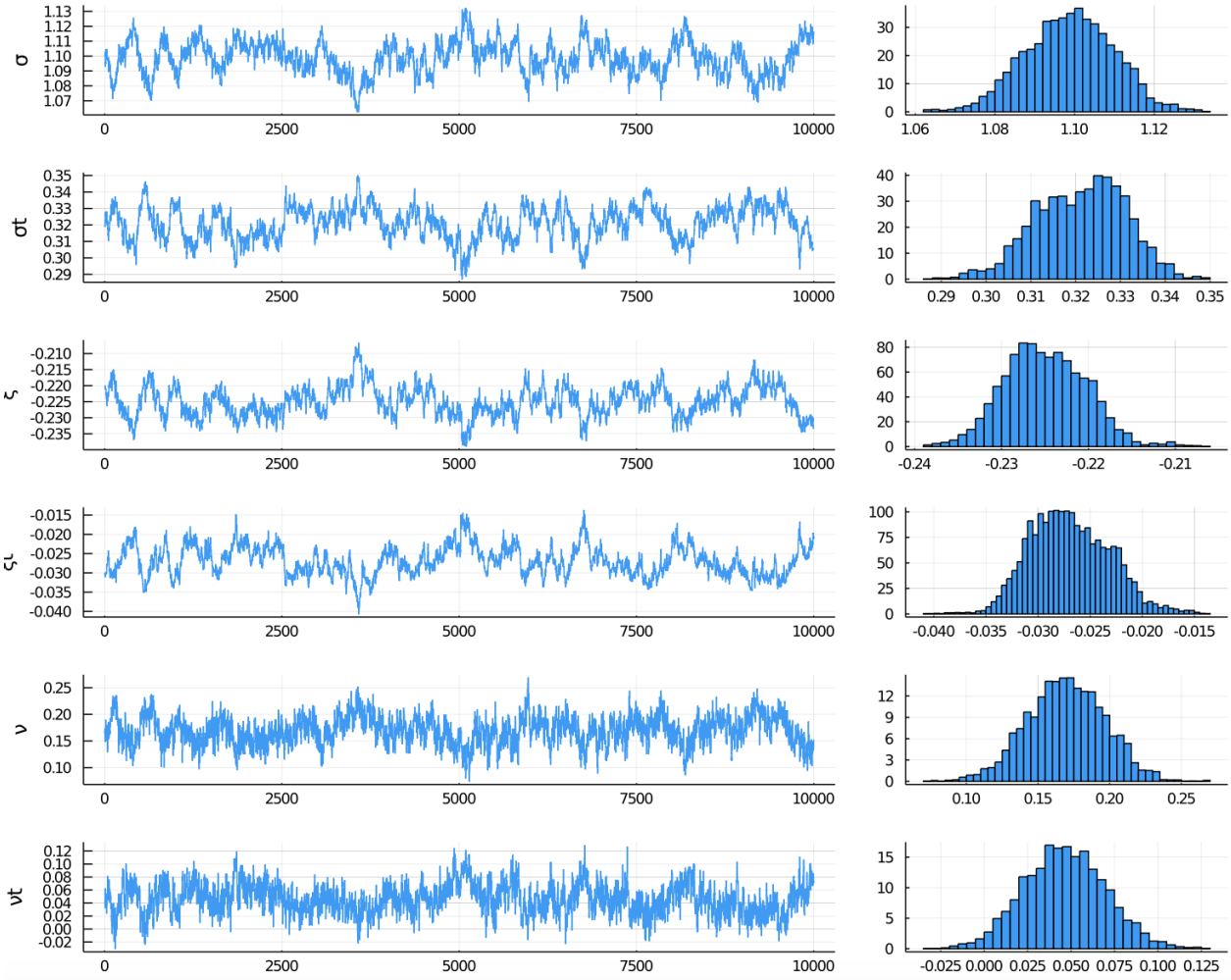

Figure 8.4 Estimation par MCMC du modèle pour l'intensité des vagues de chaleur pour l'ensemble ClimEx. Chaînes de Markov montrant la convergence de l'algorithme et densité de la loi *a posteriori* pour chaque paramètre.

modèle : μ_t pour les occurrences, $\boldsymbol{\theta}_t$ pour l'intensité et $\boldsymbol{\lambda}_t$ pour la durée. La figure 8.5 présente l'estimation bayésienne en 2020 des niveaux de retour bidimensionnels correspondant à des périodes de retour de 100, 200 et 500 ans. Des intervalles de crédibilité à 95% sont obtenus

par bootstrap. On remarque que l'intensité des niveaux de retour décroît avec leur durée, et cela d'autant plus vite que la période de retour considérée est faible. En effet, les vagues de chaleurs les plus longues sont plus rares et demandent donc une plus faible intensité pour une période de retour fixée. La figure 8.6 présente l'évolution des niveaux de retour de période 100 ans en fonction du temps. On présente l'évolution de 2020 à 2040. On peut voir sur ces

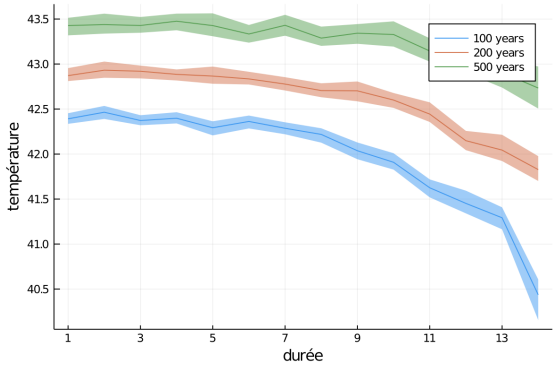

Figure 8.5 Évolution des niveaux de retour bidimensionnels calculés en 2020 en fonction de la période de retour. Intervalles de crédibilité à 95% mesurés par bootstrap.

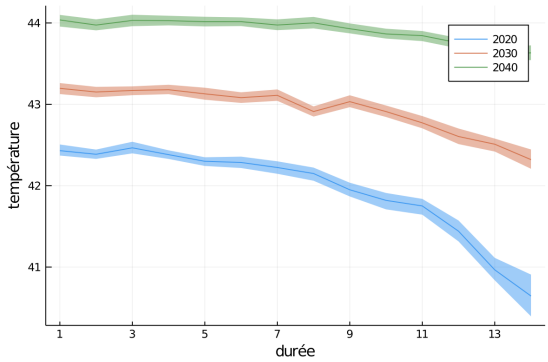

Figure 8.6 Évolution des niveaux de retour bidimensionnels de période de retour 100 ans en fonction du temps. Intervalles de crédibilité à 95% mesurés par bootstrap.

deux figures que les niveaux de retour centennal de 2030 sont assez proches de ceux de 2020 ayant une période de retour 200 ans, et que ceux de 2040 correspondent à peu de choses près aux niveaux de 2020 de période 500 ans.

8.4.2 Dynamiques spatiales

L'ajustement par l'EMV possède l'avantage d'être plus rapide et facile à mettre en oeuvre numériquement que l'ajustement par MCMC comme il ne nécessite que de donner un vecteur de valeurs initiales. Cela permet d'ajuster le modèle fréquence-intensité-durée en chaque point de grille d'un modèle climatique sans se soucier d'ajuster la taille des pas et de vérifier la convergence de la chaîne de Monte Carlo.

Dû à la diversité des climats à grande échelle spatiale nous ne pouvons pas avoir une même définition de ce qu'est une vague de chaleur en tout point de grille. Pour l'application qui suit en analysant l'ensemble de simulation CESM, une vague de chaleur en un point est défini avec pour seuils les quantiles d'ordre 0.95 de la distribution des maxima et minima journaliers sur les mois de mai à septembre de la première année de simulation (2006). La figure 8.7 montre les seuils élevés utilisés en chaque point de grille des deux domaines étudiés : le Québec et l'Europe. Pour le Québec et le nord-est des États-Unis, il existe un

gradient de température assez net du nord au sud, la région la plus chaude étant au sud des grands lacs, à l'intérieur des terres (Illinois, Indiana). En Europe, il y a plus de disparités liées au relief (Alpes, Pyrénées et Carpates) et aux différents climats (océanique, continental et méditerrannéen). Ainsi les régions les plus chaudes sont l'Andalousie, la côte adriatique italienne, ainsi que le nord-est de la Grèce. Les niveaux de retour associés à une période de

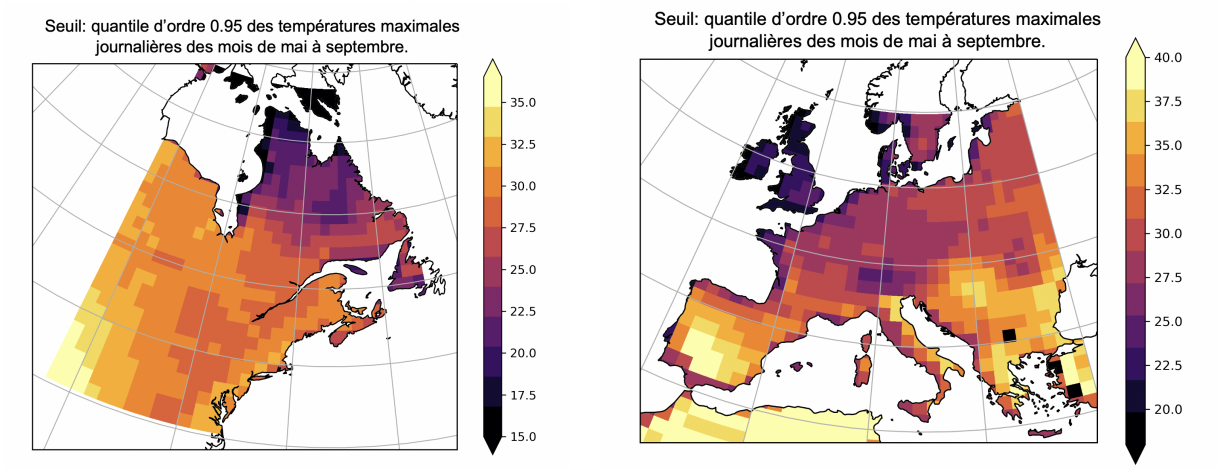

Figure 8.7 Quantiles d'ordre 0.95 des maxima journaliers simulés de mai à octobre de l'ensemble CESM, utilisés pour définir les vagues de chaleur.

100 ans en 2020 (année de référence) pour les vagues de chaleur d'au moins une semaine sont ensuite calculés (figure 8.8). Au Québec on observe les mêmes dynamiques spatiales que sur la figure 8.7. En Europe, on observe au sud-est de l'Europe, en Roumanie, Bulgarie et Hongrie notamment, des vagues de chaleur extrêmes d'intensité comparables à celles que l'on trouve au sud de l'Espagne. On observe également des niveaux assez élevés dans le sud-ouest de la France. Afin de mieux voir l'amplitude relative des vagues de chaleur en différents points où les seuils utilisés sont différents, la F figure 8.9 présente les mêmes niveaux de retour après avoir soustrait la valeur du seuil. L'intensité relative des vagues de chaleur extrêmes est importante en plus de l'intensité absolue puisque dans les régions où les maxima journaliers sont moins élevés les infrastructures sont généralement moins adaptées pour faire face à des vagues de chaleur de grandes intensités, ce qui amplifie les impacts de ces dernières. Il apparaît que relativement au seuil utilisé, représentant le coeur de la distribution des maxima journaliers, les vagues de chaleur les plus intenses se trouvent dans le nord de la France, au nord-est de l'Allemagne et en Pologne. Enfin, la figure 8.10 donne le facteur d'augmentation entre 2020 et 2050 du taux d'occurrence des vagues de chaleur de plus d'une semaine ayant une intensité supérieure au niveau de retour associé à la période de retour de 100 ans en 2020. On voit que l'occurence des vagues de chaleurs extrêmes de période de retour 100

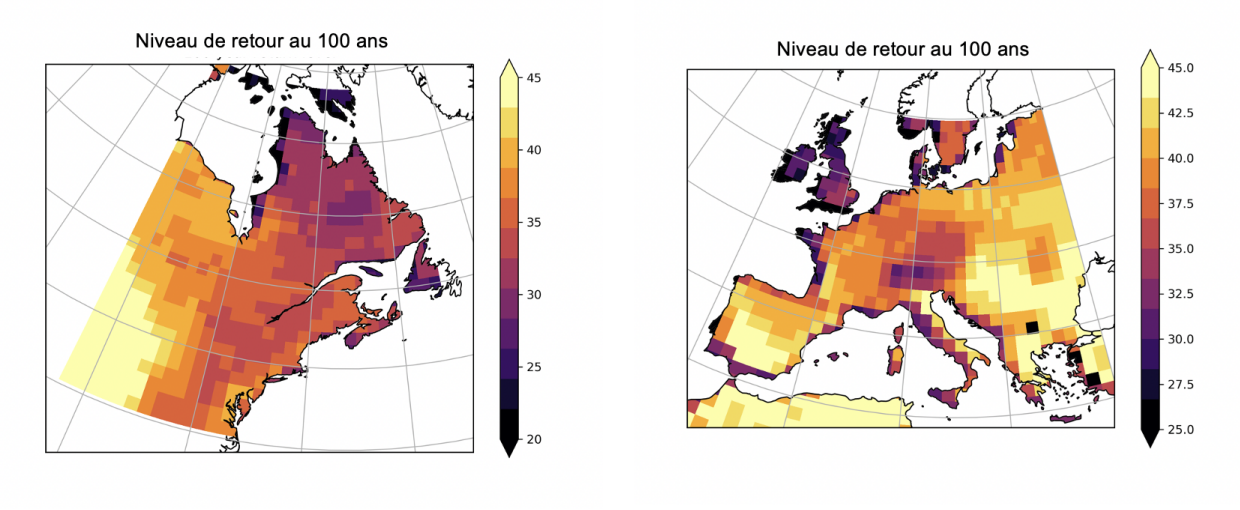

Figure 8.8 Niveaux de retour associés à une période de 100 ans en 2020 pour les vagues de chaleur d'au moins une semaine.

ans en 2020 augmente partout à horizon 2050. Au Québéc, l'augmentation la plus forte est autour de l'estuaire du Saint-Laurent, dans les provinces plus à l'est (Gaspésie, Nouveau-Brunswick Nouvelle-Écosse), ainsi que dans les états du nord-est des États-Unis (Maine, Vermont...), avec un facteur d'augmentation allant de $\times 20$ à plus de $\times 30$, ce qui correspond à des périodes de retour réduites à moins de 5 ans. En Europe, l'augmentation est la plus forte sur les côtes méditerranéennes de la France et de l'Italie, avec un facteur allant jusqu'à $\times 25$. L'augmentation est également assez forte à l'est de l'Europe ainsi qu'au nord-est de l'Espagne avec des facteurs allant de $\times 15$ à $\times 20$.

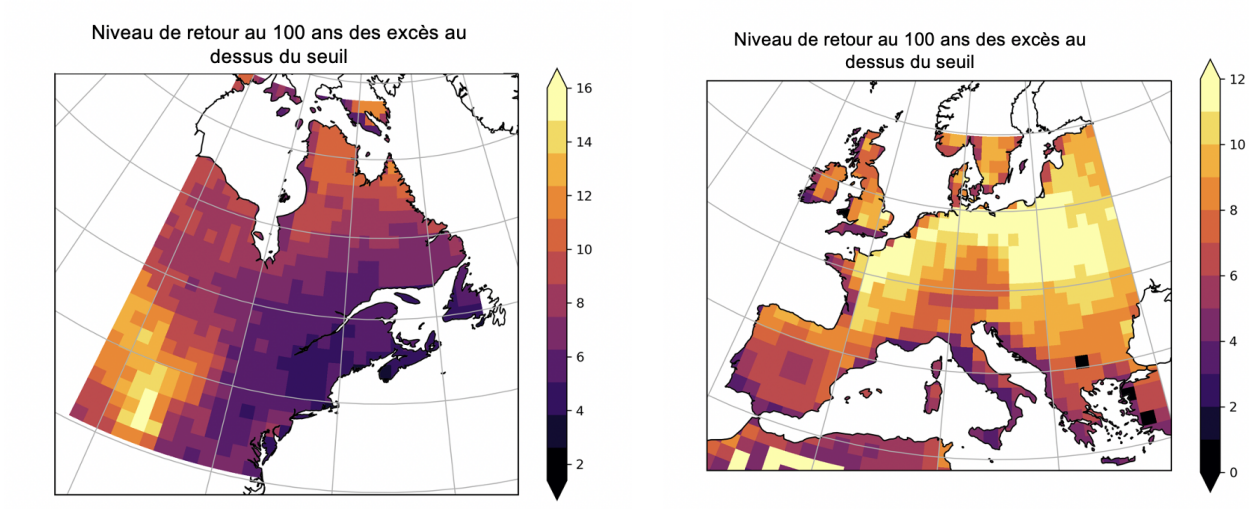

Figure 8.9 Niveaux de retour associés à une période de 100 ans en 2020 pour les vagues de chaleur d'au moins une semaine après avoir retranché la valeur du seuil.

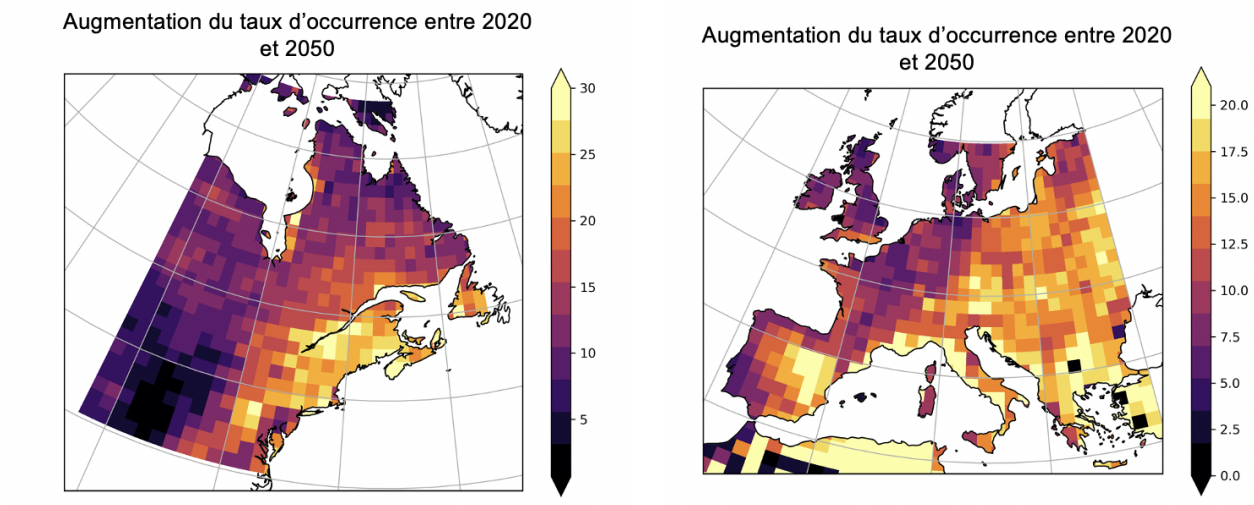

Figure 8.10 Facteur d'augmentation entre 2020 et 2050 du taux d'occurrence des vagues de chaleur de plus d'une semaine ayant une intensité supérieure au niveau de retour associé à la période de retour de 100 ans en 2020.

CHAPITRE 9 DISCUSSION GÉNÉRALE

Dans ce mémoire, un modèle bayésien a été développé pour étudier les vagues de chaleur en terme d'occurrence, d'intensité et de durée. L'intensité des vagues de chaleur a été modélisée de manière non-stationnaire par une extension de la loi de Pareto généralisée. La non-stationnarité est modélisée avec la concentration de gaz à effet de serre dans l'atmosphère du scénario utilisé dans la simulation des séries de température. La modélisation à l'aide d'une loi de Pareto généralisée étendue permet une estimation fiable des niveaux les plus extrêmes puisque celle-ci se base sur la théorie des valeurs extrêmes. En outre, nous avons montré que cette méthodologie peut aussi être utilisée pour étendre la méthode *peak-over-threshold* lorsque le seuil utilisé est sous-asymptotique au sens de la théorie des valeurs extrêmes. De plus la loi normale tronquée utilisée pour la transformation intégrale de probabilité est particulièrement bien adapté à la modélisation d'une distribution tronquée avec un seuil sous-asymptotique, ce qui est le cas pour les vagues de chaleur en climat futur.

La durée des vagues de chaleur est modélisée de manière non-stationnaire et conditionnellement à l'intensité. Encore une fois la non-stationnarité est modélisée avec la concentration en gaz à effet de serre. Cette approche permet de tenir compte de la dépendance entre l'intensité et la durée des vagues de chaleur sans introduire de copule. Il s'avère qu'une loi binomiale négative dont les paramètres dépendent à la fois du temps à travers une covariable et de l'intensité permet assez bien de décrire la distribution de la durée des vagues de chaleur.

Enfin l'occurrence des vagues de chaleur est modélisé par un processus de Poisson non-homogène. La modélisation des occurrences permet de calculer des niveaux de retour bidimensionnels pour la paire intensité-durée ainsi que leur évolution dans le temps.

A titre d'exemple, en se basant sur les simulations de l'ensemble ClimEx, on prévoit pour la ville de Montréal que les vagues de chaleur durant plus d'une semaine et atteignant une température maximale dépassant 42.2°C qui ont en 2020 une période de retour de 100 ans, auront d'ici 2030 une période de retour réduite à 45 ans et seulement 7 ans en 2050.

Le modèle a aussi été ajusté par l'EMV en différents points de grille de l'ensemble de simulations CESM afin de dégager des tendances spatiales pour les vagues de chaleur. Cela permet par exemple de voir les régions pour lesquelles les changements climatiques auront le plus d'impact sur les vagues de chaleur et aussi quantifier ces changements. Deux zones ont été considérées : la première allant de la région des grands lac nord-américains jusqu'au nord du Québec, la seconde étant le continent européen. L'étude révèle que la fréquence des vagues de chaleurs extrêmes augmentera dans l'ensemble des zones étudiées. Aussi, pour la

zone nord-américaine, la région pour laquelle les changements seront les plus marqués est le sud-est du Québec sous influence océanique, ainsi que la vallée du Saint-Laurent. Pour ces régions, la fréquence d'occurrence de vagues de chaleur ayant des périodes de retour de 100 ans en 2020 augmentera d'un facteur supérieur à 20 d'ici 2050. En d'autres termes, les vagues de chaleur aujourd'hui centennales se produiront en moyenne une fois tous les 5 ans en 2050. Pour l'Europe, les régions les plus touchées sont le sud-est et l'est de l'Europe ainsi que le nord-est de l'Espagne, et les côtes méditérranéennes françaises et italiennes. Là encore, les facteurs d'augmentation de l'occurrence des vagues de chaleur centennales en 2020 sont parfois supérieurs à 20 d'ici à 2050. Ces estimations quantitatives peuvent être de grande utilité pour prévoir les impacts du changement climatique et mettre en place des mesures d'adaptation appropriées région par région.

CHAPITRE 10 CONCLUSION ET RECOMMANDATIONS

L'étude proposée suppose que chaque membre de l'ensemble de simulations étudié est une réalisation d'un même climat et que toutes sont donc identiquement distribuées. Ceci constitue une hypothèse assez forte dans la mesure où certaines peuvent évoluer différemment que d'autres. Un modèle hiérarchique bayésien pourrait être développé afin de tenir compte de l'hétérogénéité entre les différentes simulations. De plus, le modèle proposé ne permet d'analyser qu'un ensemble de simulations à la fois, ce qui ne permet pas de tenir compte du possible biais d'un modèle climatique ou de l'incertitude sur le choix du modèle. Cela pourrait être étudié en adaptant la méthodologie pour considérer simultanément des simulations de plusieurs modèles climatiques. En particulier il pourrait être intéressant d'étudier la sensibilité des résultats à une modification de la trajectoire d'émission de gaz à effet de serre. Cela pourrait peut-être être fait en considérant des simulations d'un même modèle utilisant différents scénarios RCP.

D'autre part, le modèle a été ajusté en différents points de grille de manière indépendante alors que l'on pourrait tirer profit de la cohérence spatiale lors de l'inférence. En effet, un modèle spatiale pourrait être construit à l'aide de champs de Markov gaussiens intrinsèques. Cela permettrait de tenir compte de l'information contenue dans les points de grille voisins lors de l'inférence en un point donné.

Enfin, il serait intéressant de voir si la méthodologie utilisée peut s'appliquer dans le cadre d'une étude à très haute-résolution spatiale comme celle proposée par Ortiz et al. (2019). Cela permettrait une analyse d'impacts plus précise en étudiant notamment le phénomène des îlots de chaleur en ville.

RÉFÉRENCES

L. Abadie, A. Chiabai, et M. Neumann, “Stochastic diffusion models to describe the evolution of annual heatwave statistics : A three-factor model with risk calculations”, *The Science of the total environment*, vol. 646, pp. 670–684, 07 2018. DOI : 10.1016/j.scitotenv.2018.07.158

J. Añel, M. Fernández-González, X. Labandeira, X. Lopez-Otero, et L. de la Torre Ramos, “Impact of cold waves and heat waves on the energy production sector”, *Atmosphere*, vol. 8, p. 209, 10 2017. DOI : 10.3390/atmos8110209

A. A. Balkema et L. de Haan, “Residual life time at great age”, *Annals of Probability*, vol. 2, no. 5, pp. 792–804, 10 1974. DOI : 10.1214/aop/1176996548

J. Blanchet, E. Paquet, P. Ayar, et D. Penot, “An Objective Cross-Validation Framework for Mapping Rainfall Hazard Based on Rain Gauge Data”, *Hydrology and Earth System Sciences Discussions*, pp. 1–32, 03 2018. DOI : 10.5194/hess-2018-105

J. Carreau et Y. Bengio, “A hybrid Pareto model for asymmetric fat-tailed data : The univariate case”, *Extremes*, vol. 12, pp. 53–76, 2009. DOI : https://doi.org/10.1007/s10687-008-0068-0

G. Casella et E. I. George, “Explaining the Gibbs sampler”, *The American Statistician*, vol. 46, no. 3, pp. 167–174, 1992. DOI : 10.1080/00031305.1992.10475878

Y. Chavaillaz, P. Roy, A.-I. Partanen, L. Da Silva, E. Bresson, N. Mengis, D. Chaumont, et H. D. Matthews, “Exposure to Excessive Heat and Impacts on Labour Productivity Linked to Cumulative CO2 Emissions”, *Scientific Reports*, vol. 9, p. 13711, 2019. DOI : 10.1038/s41598-019-50047-w

S. Coles, *An Introduction to Statistical Modeling of Extreme Values*, série Springer Series in Statistics. London : Springer-Verlag, 2001. DOI : https://doi.org/10.1007/978-1-4471-3675-0

A. C. Davison, “Modelling excesses over high thresholds, with an application”, dans *Statistical Extremes and Applications*, J. T. de Oliveira, éd. Dordrecht : Springer, 1984, vol. 131. DOI : https://doi.org/10.1007/978-94-017-3069-3$_3$4

A. de Bono, G. Giuliani, S. Kluser, et P. Peduzzi, “Impacts of summer 2003 heat wave in Europe”, *UNEP/DEWA/GRID Eur. Environ. Alert Bull.*, vol. 2, pp. 1–4, 01 2004.

G. Evin, A.-C. Favre, et B. Hingray, “Stochastic Generation of Multi-Site Daily Precipitation Focusing on Extreme Events”, *Hydrology and Earth System Sciences*, vol. 22, no. 1, pp. 655–672, 2018. DOI : 10.5194/hess-22-655-2018. En ligne : https://hess.copernicus.org/articles/22/655/2018/

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, no. 2, p. 180–190, 1928. DOI : 10.1017/S0305004100015681

A. Frigessi, O. Haug, et H. Rue, “A dynamic mixture model for unsupervised tail estimation without threshold selection”, *Extremes*, vol. 5, pp. 219–235, 2002. DOI : https://doi.org/10.1023/A:1024072610684

E. Furrer, R. Katz, M. Walter, et R. Furrer, “Statistical modeling of hot spells and heat waves”, *Climate Research*, vol. 43, pp. 191–205, 10 2010. DOI : 10.3354/cr00924

GIEC, “Impacts of 1.5°C global warming on natural and human systems”, dans *Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty*, O. Hoegh-Guldberg, D. Jacob, M. Taylor, M. Bindi, S. Brown, I. Camilloni, A. Diedhiou, R. Djalante, K. L. Ebi, F. Engelbrecht, J. Guiot, Y. Hijioka, S. Mehrotra, A. Payne, S. I. Seneviratne, A. Thomas, R. Warren, et G. Zhou, éds. Press., 2018, book section 3, p. 189–191.

J.-X. Giroux, F. Chebana, P. Gosselin, et R. Bustinza, “Indicateurs et valeurs-seuils météorologiques pour les systèmes de veille-avertissement canicule pour le Québec : Mise à jour de l’etude de 2010 et développement d’un logiciel de calcul pour les systèmes d’alerte”, Institut National de la Recherche Scientifique, Rapp. tech., 2017.

E. J. Gumbel, “Statistics of extremes”, *Columbia University Press*, 1958.

D. Habeeb, J. Vargo, et B. Stone Jr., “Rising Heat Wave Trends in Large US Cities”, *Natural Hazards*, vol. 76, pp. 1651–1665, 04 2015. DOI : 10.1007/s11069-014-1563-z

W. K. Hastings, “Monte Carlo sampling methods using Markov chains and their applications”, *Biometrika*, vol. 57, no. 1, pp. 97–109, 04 1970. DOI : 10.1093/biomet/57.1.97.

En ligne : https://doi.org/10.1093/biomet/57.1.97

H. Imran, J. Kala, A. Ng, et S. Muthukumaran, “Effectiveness of vegetated patches as green infrastructure in mitigating urban heat island effects during a heatwave event in the city of Melbourne”, *Weather and Climate Extremes*, vol. 25, pp. 59 – 72, 2019. DOI : 10.1016/j.wace.2019.100217

IPCC, *Climate Change 2013 : The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom and New York, NY, USA : Cambridge University Press, 2013. DOI : 10.1017/CBO9781107415324. En ligne : www.climatechange2013.org

R. Katz, M. Parlange, et P. Naveau, “Statistics of Extremes in Hydrology”, *Advances in Water Resources*, vol. 25, pp. 1287–1304, 08 2002. DOI : 10.1016/S0309-1708(02)00056-8

J. E. Kay, C. Deser, A. Phillips, A. Mai, C. Hannay, G. Strand, J. Arblaster, S. Bates, G. Danabasoglu, J. Edwards, M. Holland, P. Kushner, J.-F. Lamarque, D. Lawrence, K. Lindsay, A. Middleton, E. Munoz, R. Neale, K. Oleson, L. Polvani, et M. Vertenstein, “The Community Earth System Model (CESM) Large Ensemble Project : A Community Resource for Studying Climate Change in the Presence of Internal Climate Variability”, *Bulletin of the American Meteorological Society*, 2015. En ligne : 10.1175/BAMS-D-13-00255.1

D. Keellings et P. Waylen, “Increased risk of heat waves in Florida : Characterizing changes in bivariate heat wave risk using extreme value analysis”, *Applied Geography*, vol. 46, pp. 90 – 97, 2014. DOI : https://doi.org/10.1016/j.apgeog.2013.11.008

N.-C. Lau et M. J. Nath, “A Model Study of Heat Waves Over North America : Meteorological Aspects and Projections for the Twenty-First Century”, *Journal of Climate*, vol. 25, no. 14, pp. 4761–4784, 07 2012. DOI : 10.1175/JCLI-D-11-00575.1. En ligne : https://doi.org/10.1175/JCLI-D-11-00575.1

M. Leduc, A. Mailhot, A. Frigon, J. Martel, R. Ludwig, G. 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*, 2019. En ligne : https://doi.org/10.1175/JAMC-D-18-0021.1.

M. Leduc, H. D. Matthews, et R. de Elía, “Regional estimates of the transient climate response to cumulative co2 emissions”, *Nature Climate Change*, vol. 6, pp. 474–478, 2016. DOI : 10.1038/nclimate2913

X. Li, C. Genest, et J. Jalbert, “A Self-Exciting Marked Point Process Model for Drought Analysis”, *Environmetrics*, 2020.

H. D. Matthews, N. P. Gillett, P. A. Stott, et K. Zickfeld, “The proportionality of global warming to cumulative carbon emissions”, *Nature*, vol. 459, pp. 829–832, 2009. DOI : 10.1038/nature08047

O. Mazdiyasni, A. AghaKouchak, S. Davis, S. Madadgar, A. Mehran, E. Ragno, M. Sadegh, A. Sengupta, S. Ghosh, D. C T, et M. Niknejad, “Increasing probability of mortality during Indian heat waves”, *Science Advances*, vol. 3, p. e1700066, 07 2017. DOI : 10.1126/sciadv.1700066

O. Mazdiyasni, M. Sadegh, F. Chiang, et A. AghaKouchak, “Heat wave intensity duration frequency curve : A multivariate approach for hazard and attribution analysis”, *Scientific Reports*, vol. 9, 12 2019. DOI : 10.1038/s41598-019-50643-w

A. McDonald, C. J. Scarrott, D. Lee, B. Darlow, M. Reale, et G. Russell, “A flexible extreme value mixture model”, *Computational Statistics and Data Analysis*, vol. 55, no. 6, pp. 2137 – 2157, 2011. DOI : https://doi.org/10.1016/j.csda.2011.01.005

P. Naveau, R. Huser, P. Ribereau, et A. Hannart, “Modeling Jointly Low, Moderate, and Heavy Rainfall Intensities Without a Threshold Selection”, *Water Resources Research*, vol. 52, no. 4, pp. 2753–2769, 2016. DOI : 10.1002/2015WR018552

L. Ortiz, J. Gonzalez, R. Horton, W. Lin, W. Wu, P. Ramamurthy, M. Arend, et R. Bornstein, “High-resolution projections of extreme heat in New York City”, *International Journal of Climatology*, vol. 39, 04 2019. DOI : 10.1002/joc.6102

I. Papastathopoulos et J. A. Tawn, “Extended generalised pareto models for tail estimation”, *Journal of Statistical Planning and Inference*, vol. 143, no. 1, pp. 131–143, 2013. DOI : https://doi.org/10.1016/j.jspi.2012.07.001

J. Parente, M. Pereira, M. Amraoui, et E. Fischer, “Heat waves in Portugal : Current regime, changes in future climate and impacts on extreme wildfires”, *Science of the Total Environment*, vol. 631-632, pp. 534 – 549, 2018. DOI : 10.1016/j.scitotenv.2018.03.044

J. Pickands, “Statistical inference using extreme order statistics”, *Annals of Statistics*, 1975.

J.-M. Robine, S. L. K. Cheung, S. Le Roy, H. Van Oyen, C. Griffiths, J.-P. Michel, et F. R. Herrmann, "Death toll exceeded 70,000 in Europe during the summer of 2003", *Comptes Rendus Biologies*, vol. 331, no. 2, pp. 171 – 178, 2008. DOI : https://doi.org/10.1016/j.crvi.2007.12.001

W. Schlenker et M. Roberts, "Nonlinear effects of weather on corn yields", *Review of Agricultural Economics*, vol. 28, pp. 391–398, 02 2006. DOI : 10.2307/3877185

B. Shaby, B. Reich, D. Cooley, et C. Kaufman, "A Markov-switching model for heat waves", *The Annals of Applied Statistics*, vol. 10, 05 2014. DOI : 10.1214/15-AOAS873

R. L. Smith, "Threshold methods for sample extremes", dans *Statistical Extremes and Applications*, J. T. de Oliveira, éd. Dordrecht : Springer, 1984, vol. 131, pp. 621–638. DOI : https://doi.org/10.1007/978-94-017-3069-3$_4$8

M. L. Stein, "Parametric Models for Distributions When Interest is in Extremes With an Application to Daily Temperature", *Extremes*, 06 2020. DOI : https://doi.org/10.1007/s10687-020-00378-z

M. Stein, "Should Annual Maximum Temperatures Follow a Generalized Extreme Value Distribution ?" *Biometrika*, vol. 104, pp. 1–16, 03 2017. DOI : 10.1093/biomet/asw070

M. Taillardat et O. Mestre, "From Research to Applications – Examples of Operational Ensemble Postprocessing in France Using Machine Learning", *Nonlinear Processes in Geophysics*, vol. 27, pp. 329–347, 05 2020. DOI : 10.5194/npg-27-329-2020

M. Taillardat, A.-L. Fougères, P. Naveau, et O. Mestre, "Forest-Based and Semiparametric Methods for the Postprocessing of Rainfall Ensemble Forecasting", *Weather and Forecasting*, vol. 34, no. 3, pp. 617–634, 06 2019. DOI : 10.1175/WAF-D-18-0149.1. En ligne : https://doi.org/10.1175/WAF-D-18-0149.1

P. Tencaliec, A.-C. Favre, P. Naveau, C. Prieur, et G. Nicolet, "Flexible Semiparametric Generalized Pareto Modeling of the Entire Range of Rainfall Amount", *Environmetrics*, vol. 31, no. 2, 2020. DOI : 10.1002/env.2582. En ligne : https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2582

A. A. Zilko et D. Kurowicka, "Copula in a multivariate mixed discrete–continuous model", *Computational Statistics and Data Analysis*, vol. 103, pp. 28 – 55, 2016. DOI : https://doi.org/10.1016/j.csda.2016.02.017