Développement d'un modèle statistique non stationnaire et régional pour les précipitations extrêmes simulées par un modèle numérique de climat

Jonathan Jalbert

UNIVERSITÉ GRENOBLE ALPES

THÈSE en cotutelle

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Océan, Atmosphère, Hydrologie**

Arrêté ministériel : 7 août 2006

Présentée par

Jonathan Jalbert

Thèse dirigée par **Anne-Catherine Favre** et **Claude Bélisle**
codirigée par **Jean-François Angers**

préparée au sein du **Laboratoire d'étude des Transferts en Hydrologie et Environnement (LTHE, UMR 5564, CNRS - Grenoble INP - IRD - UJF)**
dans l'École Doctorale **Terre Univers Environnement**

Développement d'un modèle statistique non stationnaire et régional pour les précipitations extrêmes simulées par un modèle numérique de climat

Thèse soutenue publiquement le **30 octobre 2015**,
devant le jury composé de :

Robert Guénette
Professeur titulaire, Université Laval (Canada), Président
Philippe Naveau
Directeur de recherche, CNRS, Rapporteur
Thierry Duchesne
Professeur titulaire, Université Laval (Canada), Rapporteur
Stéphane Girard
Directeur de recherche, INRIA Grenoble Rhône-Alpes, Membre
Anne-Catherine Favre
Professeure, ENSE3 -GINP, LTHE, Grenoble, Directrice de thèse
Claude Bélisle
Professeur titulaire, Université Laval (Canada), Directeur de thèse
Jean-François Angers
Professeur titulaire, Université de Montréal (Canada), Co-Directeur de thèse

Développement d'un modèle statistique non stationnaire et régional pour les précipitations extrêmes simulées par un modèle numérique de climat

Thèse en cotutelle
Doctorat de l'Université Laval en mathématiques
Doctorat de l'Université de Grenoble, spécialité : Océan, Atmosphère, Hydrologie

Jonathan Jalbert

Université Laval
Québec, Canada
Philosophiæ doctor (Ph.D.)

et

Université de Grenoble
Thèse effectuée au sein du
Laboratoire d'étude des Transferts en Hydrologie et Environnement (LTHE, UMR 5564, CNRS - Grenoble INP - IRD - UJF)
de l'École Doctorale Terre Univers Environnement
Grenoble, France
Philosophiæ doctor (Ph.D.)

Résumé

Les inondations constituent le risque naturel prédominant dans le monde et les dégâts qu'elles causent sont les plus importants parmi les catastrophes naturelles. Un des principaux facteurs expliquant les inondations sont les précipitations extrêmes. En raison des changements climatiques, l'occurrence et l'intensité de ces dernières risquent fort probablement de s'accroître. Par conséquent, le risque d'inondation pourrait vraisemblablement s'intensifier. Les impacts de l'évolution des précipitations extrêmes sont désormais un enjeu important pour la sécurité du public et pour la pérennité des infrastructures. Les stratégies de gestion du risque d'inondation dans le climat futur sont essentiellement basées sur les simulations provenant des modèles numériques de climat. Un modèle numérique de climat procure notamment une série chronologique des précipitations pour chacun des points de grille composant son domaine spatial de simulation. Les séries chronologiques simulées peuvent être journalières ou infra-journalières et elles s'étendent sur toute la période de simulation, typiquement entre 1961 et 2100. La continuité spatiale des processus physiques simulés induit une cohérence spatiale parmi les séries chronologiques. Autrement dit, les séries chronologiques provenant de points de grille avoisinants partagent souvent des caractéristiques semblables. De façon générale, la théorie des valeurs extrêmes est appliquée à ces séries chronologiques simulées pour estimer les quantiles correspondants à un certain niveau de risque. La plupart du temps, la variance d'estimation est considérable en raison du nombre limité de précipitations extrêmes disponibles et celle-ci peut jouer un rôle déterminant dans l'élaboration des stratégies de gestion du risque. Par conséquent, un modèle statistique permettant d'estimer de façon précise les quantiles de précipitations extrêmes simulées par un modèle numérique de climat a été développé dans cette thèse. Le modèle développé est spécialement adapté aux données générées par un modèle de climat. En particulier, il exploite l'information contenue dans les séries journalières continues pour améliorer l'estimation des quantiles non stationnaires et ce, sans effectuer d'hypothèse contraignante sur la nature de la non-stationnarité. Le modèle exploite également l'information contenue dans la cohérence spatiale des précipitations extrêmes. Celle-ci est modélisée par un modèle hiérarchique bayésien où les lois *a priori* des paramètres sont des processus spatiaux, en l'occurrence des champs de Markov gaussiens. L'application du modèle développé à une simulation générée par le Modèle régional canadien du climat a permis de réduire considérablement la variance d'estimation des quantiles en Amérique du Nord.

Table des matières

Liste des tableaux

Liste des figures

Se tromper est le privilège naturel de l'homme par rapport à tous les autres organismes. Cela conduit à la vérité ! Je suis homme parce que je déraisonne. On n'est jamais arrivé à une vérité sans avoir quatorze fois erré et peut-être cent quarante fois, et c'est d'ailleurs encore honorable.

Fiodor Dostoïevski dans Crime et châtiment, troisième partie, chapitre 1. Publié en 1866.

Remerciements

Je tiens à remercier chaleureusement mon *triumvirat* de directeurs, composé d'Anne-Catherine Favre, Claude Bélisle et Jean-François Angers. À vous d'inférer pour déterminer lequel joue le rôle de Crassus, César ou Pompée. ◡̈ Pour ma thèse, ils ont amalgamé enthousiasme, pragmatisme et rigueur. Je les remercie pour leur soutien inconditionnel dans les multiples situations qui se sont présentées. Anne-Catherine, merci d'avoir su attiser ma motivation lorsque nécessaire. Claude, merci d'avoir accepté de diriger un ingénieur-physicien converti et d'avoir persévéré devant mes limites en analyse. Jean-François, merci de m'avoir si bien dirigé pour ma première charge de cours. Je n'aurais certainement pas aussi bien réussi autrement.

Je remercie aussi mes *prefs* d'Ouranos (Dominique, Élyse, Jacinthe, Mathieu, Valérie) sans qui j'aurais certainement souffert d'une carence en caféine. Ils ne s'en doutent peut-être pas, mais ma thèse a avancé grâce à nos diverses conversations. C'est extrêmement enrichissant d'intégrer une équipe multidisciplinaire.

La thèse s'étant effectuée en cotutelle, elle a été ponctuée par plusieurs séjours à Grenoble. Je remercie les *thésards* du LTHE (Baptiste, Gérémy, Jérémy, Olivier, Stéphanie), qui m'ont immédiatement accueilli et intégré. Thibault, merci pour ton hospitalité, mais surtout pour ton amitié. Audrey ! Ma chère amie depuis l'INRS, quelle chance que nos parcours se soient croisés à Grenoble.

Plus formellement, je souhaite remercier le consortium Ouranos pour avoir financé mon projet de doctorat. Il n'aurait pas été possible de le réaliser sans cet appui important. Je remercie également l'Institut des sciences mathématiques pour l'octroi d'une bourse de recherche et le professeur Louis-Paul Rivest de l'Université Laval pour le versement d'une bourse de recherche pour ma dernière année.

Finalement, je tiens à rendre hommage à ma femme, Justine la Vertueuse, sans qui ce projet n'aurait jamais abouti. Ce travail est aussi le tien. Tu as agi en tant que muse et véritable mécène de la recherche. Merci aussi à ma famille pour votre support et votre confiance inébranlable.

Avant-propos

Ce manuscrit s'articule autour de trois articles scientifiques insérés tels quels. Je suis le principal auteur de ces articles et également le principal contributeur au contenu. Le premier article intitulé *Canadian RCM projected transient changes to precipitation occurrence, intensity and return level over North America* a été publié dans *Journal of Climate*[1]. Les deuxième et troisième articles sont en préparation pour une soumission imminente. Chacun des articles comporte une introduction et une revue de littérature détaillée. L'introduction du manuscrit demeure par conséquent assez générale en établissant notamment le contexte socio-politique, géographique et environnemental. Il est à noter que les références qui sont citées dans un chapitre sont fournies à la fin de celui-ci. Cette structure semblait appropriée pour une thèse par articles.

1. Jalbert, J., A.-C. Favre, C. Bélisle, J.-F. Angers, et D. Paquin (2015). Canadian RCM projected transient changes to precipitation occurrence, intensity and return level over North America. *Journal of Climate*, vol. 28, n°17, p. 6920–6937.

Introduction

Les inondations constituent les catastrophes naturelles les plus fréquentes au monde. Généralement, on parle d'inondation lorsque le niveau d'écoulement d'un cours d'eau devient tel qu'il menace des vies humaines et l'intégrité des infrastructures. En 2010, plus de 178 millions de personnes ont été touchées par les inondations et les pertes économiques excèdent 40 milliards de dollars américains (Jha et collab., 2012). Un recensement a permis de montrer que les inondations sont en augmentation, surtout au cours des 20 dernières années. Le nombre de victimes, les coûts d'indemnisation ainsi que les dégâts économiques sont également en augmentation. En zone rurale, le risque d'inondation est tributaire des caractéristiques du bassin versant. L'impact des précipitations abondantes sur le risque d'inondation augmente à mesure que le ratio entre l'écoulement occasionné par les précipitations et ces dernières s'accentue. En zone urbaine, les inondations sont la plupart du temps causées par des événements de précipitations dont l'intensité dépasse la capacité de drainage des réseaux d'évacuation des eaux pluviales.

La vulnérabilité des sociétés face aux inondations est réduite par un certain nombre de mesures gouvernementales. D'une part, une mesure simple consiste à limiter l'urbanisation dans les zones inondables. Typiquement, les zones inondables sont délimitées aux régions qui possèdent un certain risque d'être inondées à chaque année. Ce risque est fixé par les décideurs et il est fonction de l'utilisation des terres et des conséquences d'une inondation. D'autre part, les infrastructures essentielles telles les ponts, les barrages et les réseaux d'évacuation des eaux pluviales doivent être dimensionnées de façon à ce que leur risque de défaillance face à une inondation soit inférieur à un certain seuil. Encore une fois, ce risque est fonction de l'infrastructure et de l'ampleur des conséquences d'une défaillance. La définition des zones inondables et le dimensionnement des infrastructures nécessitent la connaissance du niveau d'écoulement pour lequel la probabilité annuelle qu'il soit dépassé corresponde au risque fixé. La plupart du temps, les niveaux d'écoulement sont estimés à l'aide d'un modèle hydrologique (Jiménez-Cisneros et collab., 2014). Une méthodologie répandue consiste à utiliser un modèle hydrologique pour transformer l'intensité des précipitations qui correspond au niveau de risque fixé en débit d'écoulement.

Le risque d'occurrence des précipitations abondantes causant les inondations est impossible à

prévoir de façon déterministe sur la durée de vie des installations. En effet, la nature chaotique des équations décrivant les processus générateurs des précipitations en empêche la prévision déterministe à long terme. Par conséquent, l'étude des précipitations s'effectue généralement sur une base probabiliste. La modélisation probabiliste consiste à caractériser l'occurrence et l'intensité des précipitations d'une série observée par une loi de probabilité. Généralement, les séries disponibles sont plutôt courtes, inférieures à une centaine d'années pour la grande majorité d'entre elles. Dans la plupart des cas, les niveaux de risque choisis sont de faibles à très faibles. Par exemple, les municipalités peuvent donner l'autorisation d'urbaniser une zone inondable que si le risque annuel d'inondation y est inférieur à 1/100. Pour les barrages hydroélectriques, ce risque peut être réduit à moins de 1/1000. Par conséquent, les intensités des précipitations qui correspondent à ces faibles risques ne sont habituellement pas observées. L'estimation du niveau correspondant à ce risque se retrouve donc extrapolée par le modèle probabiliste utilisé. Celui-ci doit s'appuyer sur de solides bases mathématiques pour que la confiance en l'extrapolation soit satisfaisante. C'est pourquoi la théorie des valeurs extrêmes (*cf.* chapitre 1) est préconisée pour l'étude des précipitations intenses. Cette théorie fournit entre autre les lois de probabilité asymptotiques pour les valeurs extrêmes. Celles-ci constituent donc les bases privilégiées pour le développement d'un modèle de valeurs extrêmes requis pour l'extrapolation des précipitations intenses.

Jusqu'à tout récemment, les stratégies de gestion du risque associé aux inondations sur la durée de vie des installations supposaient la stationnarité du climat. Or, une telle hypothèse est désormais reconnue comme étant inadéquate en raison des changements climatiques anticipés, d'autant plus que la fréquence et l'intensité des événements de précipitations intenses risquent fort probablement d'augmenter (IPCC, 2012). En fait, les changements climatiques de nature anthropique accentuent les facteurs de stress hydriques existants tels que la croissance démographique, l'urbanisation et le changement d'affectation des terres (IPCC, 2013). En raison d'une augmentation de sa concentration en gaz à effet de serre, l'atmosphère intercepte davantage de chaleur réémise par la Terre causant ainsi son réchauffement. Ce réchauffement a un impact entre autre, sur les précipitations car une atmosphère plus chaude peut contenir plus de vapeur d'eau. Ce n'est cependant qu'un des nombreux facteurs qui influencent les processus générateurs des précipitations. Celles-ci conditionnent la disponibilité de la ressource en eau et leur évolution pourrait avoir des impacts significatifs dans plusieurs secteurs essentiels tels que l'approvisionnement en eau potable, la production alimentaire, le maintien des écosystèmes naturels, la production énergétique et la qualité de l'eau. L'évolution du climat comprendra fort probablement une augmentation de la fréquence et de l'intensité des événements de précipitations intenses (IPCC, 2012). Cette augmentation occasionne également un accroîssement du risque d'inondation (Hirabayashi et collab., 2013). Il devient donc nécessaire d'inclure les changements climatiques afin d'établir des stratégies adéquates de gestion du risque pour les installations qui seront soumises au climat futur. Pour établir de telles stratégies, des projections quantitatives du climat futur sont nécessaires et en particulier des

projections de précipitations.

La complexité du système climatique terrestre empêche la prédiction simple de l'évolution des précipitations suite à une augmentation de la concentration des gaz à effet de serre. Selon le Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC), seuls les modèles numériques de climat sont en mesure de fournir des prédictions quantitatives des précipitations pour le climat futur. Un modèle numérique de climat constitue une représentation mathématique simplifiée de la physique du climat. Les équations décrivant le climat découlent directement des processus physiques de conservation de la quantité de mouvement, de l'énergie et de la masse. La résolution de ces équations s'effectue sur une grille tridimensionnelle couvrant l'atmosphère de la Terre. Typiquement, les modèles globaux possèdent une résolution horizontale comprise entre 250 et 600 km et comprennent entre 10 et 20 couches d'atmosphère. La puissance de calcul informatique limite la résolution spatiale de ces modèles. D'ailleurs, la résolution spatiale des modèles globaux est trop grossière pour simuler adéquatement certains phénomènes à plus petite échelle comme la formation des précipitations convectives. Ces modèles globaux ne sont donc généralement pas appropriés pour l'étude des précipitations intenses (Prudhomme et collab., 2002; Christensen et Christensen, 2003, 2007).

Une amélioration de la résolution des modèles, tant spatiale que temporelle, est une condition essentielle à une meilleure simulation des processus en jeu (Barrow et collab., 2004). Afin d'améliorer la résolution spatiale des modèles de climat globaux, les modèles de climat régionaux ont été développés. Ceux-ci sont très similaires aux modèles globaux à l'exception que leur domaine de simulation est limité à une région particulière du globe. En limitant le domaine, les résolutions spatiale (typiquement inférieure à 50 km) et temporelle du modèle peuvent s'accroître de façon à ce que les processus à plus petite échelle puissent être simulés. En particulier, les modèles régionaux reproduisent raisonnablement leurs caractéristiques justement grâce à leur capacité de simuler les phénomènes à plus petite échelle (Rasmussen et collab., 2012). C'est pourquoi les modèles régionaux sont la plupart du temps utilisés pour les études d'impacts. Par ailleurs, puisque les modèles régionaux ne couvrent pas tout le globe, les conditions aux frontières doivent être spécifiées par un modèle global.

Une simulation générée par un modèle de climat peut être interprétée comme une météorologie probable, autrement dit, une réalisation probable du climat, sur la période de simulation. Dans ce contexte, la modélisation probabiliste des variables simulées permet d'inférer sur le climat, soit sur l'ensemble des météorologies probables. De plus, pour bon nombre d'applications tel le dimensionnement des barrages hydroélectriques, le risque acceptable de défaillance est fixé si faible que la probabilité qu'un événement correspondant à un tel risque se produise sur la période de simulation est extrêmement mince. Alors, la théorie des valeurs extrêmes est encore nécessaire pour estimer l'intensité des précipitations correspondant à ce niveau de risque. Par conséquent, la modélisation probabiliste est tout aussi pertinente dans le contexte

des données simulées.

Les modèles numériques de climat occupent un rôle prépondérant dans les études d'impacts et d'adaptation aux changements climatiques. La modélisation probabiliste des variables simulées devrait être spécialement adaptée à leur structure afin d'exploiter le maximum d'information contenu dans leurs simulations. Cela est d'autant plus pertinent pour les événements extrêmes, telles les précipitations intenses, puisque la rareté de ces événements induit généralement une grande incertitude sur leur prédiction. L'incertitude sur les prédictions des précipitations intenses simulées peut être réduite en exploitant les caractéristiques des séries simulées, soit de longues séries chronologiques continues pour chacun des points de grille du domaine spatial. Le fait que les séries soient longues implique qu'elles sont susceptibles d'être non stationnaires en raison des changements climatiques. Pour exploiter toute la longueur de la série, cette non-stationnarité doit être modélisée. Aussi, les précipitations simulées sont spatialement corrélées dans l'espace. Une modélisation spatiale est nécessaire pour tirer profit de l'information provenant de cette cohérence spatiale.

Dans un premier temps, le modèle probabiliste de valeurs extrêmes devrait inclure la non-stationnarité des précipitations intenses simulées. Traditionnellement, l'évolution des précipitations intenses était évaluée en comparant les prédictions pour deux sous-périodes (Christensen et Christensen, 2007; Mearns et collab., 2009; Christensen et collab., 2013), par exemple les sous-périodes $[1961, 1990]$ et $[2071, 2100]$. Une telle démarche affranchissait la modélisation de la non-stationnarité car il était supposé que celle-ci était négligeable pour de tels intervalles. Toutefois, le fait que les années intermédiaires ne soient pas considérées constitue une importante perte d'information et occasionne également une grande incertitude sur les prédictions. La modélisation de la non-stationnarité permet d'exploiter l'information de toute la période de simulation. En ce sens, quelques modèles de valeurs extrêmes ont été développés pour les séries non stationnaires. Toutefois, la non-stationnarité dans ces modèles est généralement estimée en utilisant l'information contenue dans une série partielle d'événements intenses. L'information contenue dans le reste de la série n'est pas prise en compte. En somme, la prédiction des précipitations intenses dans le climat futur bénéficierait sans aucun doute du développement d'un modèle non stationnaire de valeurs extrêmes adapté aux longues séries chronologiques simulées par les modèles de climat.

Dans un deuxième temps, la cohérence spatiale des précipitations intenses simulées par un modèle numérique de climat devrait également être exploitée. En effet, les valeurs de précipitations simulées aux différents points de grille sont interdépendantes puisque, d'une part, les points de grille voisins sont interreliés par les conditions de continuité des processus physiques et, d'autre part, les conditions météorologiques de points de grille voisins peuvent être semblables. L'utilisation d'un modèle probabiliste multidimensionnel pour modéliser conjointement tous les points de grille permettrait d'exploiter l'information contenue dans la cohérence spatiale des précipitations intenses parmi les points de grille. En somme, la modélisation

probabiliste conjointe de tous les points de grille permettrait une exploitation plus complète de l'information disponible, améliorant ainsi les prédictions sur les précipitations intenses.

Par conséquent, le but du projet de doctorat consiste à développer un modèle probabiliste spécialement adapté aux précipitations intenses simulées par un modèle numérique de climat. Celui-ci devra inclure la modélisation de la non-stationarité des longues séries de précipitations intenses et leur cohérence spatiale sur le domaine de simulation. Dans le cadre de cette thèse, les précipitations simulées par le Modèle régional canadien du climat sur l'Amérique du Nord seront étudiées sur la période $[1961, 2100]$. Le premier objectif du projet consiste à introduire la non-stationnarité climatique dans un modèle probabiliste de valeurs extrêmes pour caractériser les précipitations intenses. Le second objectif du projet concerne la prise en compte de la dépendance spatiale dans le modèle d'extrêmes non stationnaires. Le modèle probabiliste non stationnaire et régional ainsi développé permettra d'améliorer les prédictions concernant les précipitations intenses du climat futur.

La thèse est composée de quatre chapitres, où trois d'entre eux constituent des articles scientifiques insérés tels quels. Le chapitre 1, intitulé *Cadre théorique*, présente les principaux éléments mathématiques utilisés dans les articles. Les chapitres 2 et 3, présents sous forme d'articles, remplissent le premier objectif du projet de thèse. Une méthode permettant d'inclure la non-stationnarité dans un modèle de valeurs extrêmes pour les précipitations simulées par un modèle de climat y est développée. Cette méthode peut être interprétée comme un prétraitement des données dont le motif consiste à stationnariser les séries avant d'appliquer un modèle de valeurs extrêmes pour les séries stationnaires. Le chapitre 4, présent également sous forme d'article, satisfait le second objectif du projet de thèse. Un modèle multidimensionnel de valeurs extrêmes y est développé dans le but de modéliser la dépendance spatiale des précipitations intenses simulées par un modèle de climat. Ce modèle probabiliste constitue un modèle hiérarchique bayésien (*cf.* chapitre 1) où la loi *a priori* constitue un champ de Markov gaussien intrinsèque (*cf.* chapitre 1). Cette forme de la loi *a priori* permet d'inclure la cohérence spatiale dans le modèle probabiliste. Finalement, le dernier chapitre contient la conclusion générale des travaux et quelques idées pour de futurs projets.

Références

Barrow, E., B. Maxwell et P. Gachon. 2004, *La variabilité et le changement climatique au Canada : le passé, le présent et le futur*, DSCA Séries d'évaluations scientifique, Service météorologique du Canada, Toronto, Canada, 114 pages.

Christensen, J. H. et O. B. Christensen. 2003, «Climate modelling : Severe summertime flooding in Europe», *Nature*, vol. 421, n° 6925, p. 805–806.

Christensen, J. H. et O. B. Christensen. 2007, «A summary of the PRUDENCE model projec-

tions of changes in European climate by the end of this century», *Climatic Change*, vol. 81, n° S1, p. 7–30.

Christensen, J. H., K. K. Kumar, E. Aldrian, S.-I. An, I. F. A. Cavalcanti, M. de Castro, W. Dong, P. Goswami, A. Hall, J. K. Kanyanga, A. Kitoh, J. Kossin, N.-C. Lau, J. Renwick, D. B. Stephenson, S.-P. Xie et T. Zhou. 2013, «Climate Phenomena and their Relevance for Future Regional Climate Change», dans *Climate Change 2013 : The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, chap. 14, Cambridge University Press, Cambridge, United Kingdom and New York, USA, p. 1217–1308.

Hirabayashi, Y., R. Mahendran, S. Koirala, L. Konoshima, D. Yamazaki, S. Watanabe, H. Kim et S. Kanae. 2013, «Global flood risk under climate change», *Nature Climate Change*, vol. 3, n° 9, p. 816–821.

IPCC. 2012, *Managing the risks of extreme events and disasters to advance climate change adaptation. A special report of Working Groups I and II of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge, United Kingdom and New York, USA, 582 pages.

IPCC. 2013, *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 University Press, Cambridge, United Kingdom and New York, USA, 1535 pages.

Jha, A. K., R. Bloch et J. Lamond. 2012, *Cities and flooding : a guide to integrated urban flood risk management for the 21st century*, World Bank Publications, Washington, USA, 632 pages.

Jiménez-Cisneros, B. E., T. Oki, N. W. Arnell, G. Benito, J. G. Cogley, P. Döll, T. Jiang et S. S. Mwakalila. 2014, «Freshwater resources», dans *Climate Change 2014 : Impacts, Adaptation, and Vulnerability. Part A : Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, chap. 3, Cambridge, United Kingdom and New York, USA, p. 229–269.

Mearns, L. O., W. Gutowski, R. Jones, R. Leung, S. McGinnis, A. Nunes et Y. Qian. 2009, «A Regional Climate Change Assessment Program for North America», *Eos, Transactions American Geophysical Union*, vol. 90, n° 36, p. 311.

Prudhomme, C., N. Reynard et S. Crooks. 2002, «Downscaling of global climate models for flood frequency analysis : where are we now ?», *Hydrological Processes*, vol. 16, n° 6, p. 1137–1150.

Rasmussen, S. H., J. H. Christensen, M. Drews, D. J. Gochis et J. C. Refsgaard. 2012, «Spatial-scale characteristics of precipitation simulated by regional climate models and the implications for hydrological modeling», *Journal of Hydrometeorology*, vol. 13, p. 1817–1835.

Chapitre 1

Cadre théorique

1.1 Introduction à la théorie des valeurs extrêmes

Le théorie des valeurs extrêmes est une discipline de la statistique qui concerne spécifiquement les valeurs dans les queues de distributions. Elle est largement utilisée en ingénierie, en actuariat et en hydrologie, pour estimer le risque associé aux événements extrêmes. Dans ce chapitre, certains éléments de cette théorie seront présentés dans le but de les appliquer à l'étude des précipitations extrêmes.

1.1.1 Le cas indépendant et identiquement distribué

Soit $X_1, X_2, \ldots$ une suite de variables aléatoires indépendantes et identiquement distribuées selon la fonction de répartition F. Notons par M_n le maximum des n premières variables aléatoires

$$M_n = \max\{X_1, X_2, \ldots, X_n\}. \tag{1.1}$$

La distribution de ce maximum se calcule de la façon suivante :

$$\mathbb{P}[M_n \leq x] = \mathbb{P}\left[(X_1 \leq x) \bigcap \ldots \bigcap (X_n \leq x)\right] = \prod_{i=1}^{n} \mathbb{P}[X_i \leq x] = F^n(x). \tag{1.2}$$

En pratique, la fonction de répartition F est généralement inconnue. On pourrait être tenté de remplacer F par une estimation de celle-ci mais l'erreur d'estimation ferait exploser l'erreur d'estimation de F^n. L'alternative préconisée par la théorie des valeurs extrêmes consiste à identifier une distribution limite directement pour F^n sans connaître F ni même l'estimer d'abord.

Une difficulté immédiate se présente dans la recherche de la loi asymptotique du maximum. Puisque

$$\lim_{n\to\infty} F^n(x) = \begin{cases} 1 & \text{si } F(x) = 1; \\ 0 & \text{si } F(x) < 1; \end{cases} \tag{1.3}$$

la loi asymptotique du maximum est dégénérée. Pour contourner cette difficulté, il est nécessaire d'effectuer une transformation du maximum (si elle existe) afin que le passage à la limite n'induise pas la dégénérescence de sa distribution. Fisher et Tippett (1928) ont considéré une transformation linéaire dans l'énoncé du théorème suivant où la démonstration formelle a été effectuée par Gnedenko (1943) :

Théorème 1.1 (Fisher et Tippett, 1928; Gnedenko, 1943)**.** *S'il existe des suites de constantes normalisantes* $(a_n > 0 : n \geq 1)$ *et* $(b_n : n \geq 1)$ *pour lesquelles la suite* $\left(\frac{M_n - b_n}{a_n} : n \geq 1\right)$ *converge en loi vers une distribution* G *non dégénérée, alors* G *appartient à l'une des trois familles de lois suivantes :*

$$\textit{Type I} : G(x) = \exp\left[-\exp\left\{-\left(\frac{x-\mu}{\sigma}\right)\right\}\right], x \in \mathbb{R} \tag{1.4}$$

$$\textit{Type II} : G(x) = \begin{cases} 0, & x \leq \mu \\ \exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^{-\kappa}\right\}, & x > \mu \end{cases} \tag{1.5}$$

$$\textit{Type III} : G(x) = \begin{cases} \exp\left[-\left\{-\left(\frac{x-\mu}{\sigma}\right)\right\}^{\kappa}\right], & x < \mu \\ 1, & x \geq \mu \end{cases} \tag{1.6}$$

où $\mu \in \mathbb{R}$, $\sigma > 0$ *et* $\kappa > 0$.

Ce résultat est fondamental en théorie des valeurs extrêmes puisqu'il fournit les seules distributions asymptotiques possibles pour le maximum normalisé. En ce sens, c'est l'équivalent pour le maximum du théorème limite central.

La famille de type I est généralement connue sous la dénomination loi de Gumbel, tandis que les familles de type II et de type III sont généralement nommées respectivement loi de Fréchet et loi de Weibull. Chacune de ces familles de loi de probabilité possède un paramètre de localisation μ et un paramètre d'échelle σ. Les familles de types II et III possèdent en plus un paramètre de forme κ.

Jenkinson (1955) a généralisé en une seule famille de lois les trois types de lois des valeurs extrêmes précédents, la loi des extrêmes généralisée (*Generalized Extreme Values*, GEV).

Définition 1.1 (La loi des extrêmes généralisée)**.** *La loi GEV avec les paramètres de localisation* $\mu \in \mathbb{R}$, *d'échelle* $\sigma > 0$ *et de forme* $\xi \in \mathbb{R}$, *dénotée par* $\mathcal{GEV}(\mu, \sigma, \xi)$, *est la loi de probabilité possédant la fonction de répartition* $\mathcal{GEV}(x|\mu, \sigma, \xi)$ *donnée par :*

$$\mathcal{GEV}(x|\mu, \sigma, \xi) = \exp\left[-\left\{1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right\}^{-\frac{1}{\xi}}\right]; \tag{1.7}$$

pour tous les x *tels que* $1 + \xi(x - \mu)/\sigma > 0$. *Le cas* $\xi = 0$ *est défini par continuité. Plus précisément,*

- *Le cas* $\xi = 0$ *:*

$$\mathcal{GEV}(x|\mu,\sigma,0) = \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\}; \tag{1.8}$$

- *Le cas* $\xi > 0$ *:*

$$\mathcal{GEV}(x|\mu,\sigma,\xi) = \begin{cases} 0 & si\ x < \mu - \sigma/\xi; \\ \exp\left[-\left\{1+\xi\frac{(x-\mu)}{\sigma}\right\}^{-1/\xi}\right] & si\ x > \mu - \sigma/\xi; \end{cases} \tag{1.9}$$

- *Le cas* $\xi < 0$ *:*

$$\mathcal{GEV}(x|\mu,\sigma,\xi) = \begin{cases} \exp\left[-\left\{1+\xi\frac{(x-\mu)}{\sigma}\right\}^{-1/\xi}\right] & si\ x < \mu - \sigma/\xi; \\ 1 & si\ x > \mu - \sigma/\xi. \end{cases} \tag{1.10}$$

On peut montrer que la loi GEV se simplifie en une loi de type I lorsque $\xi = 0$, en une loi de type II lorsque $\xi < 0$ et en une loi de type III lorsque $\xi > 0$. Le théorème fondamental de la théorie des valeurs extrêmes peut être réécrit en utilisant la loi GEV.

Théorème 1.2 (Fisher et Tippett, 1928; Gnedenko, 1943)**.** *S'il existe des suites de constantes normalisantes* $(a_n > 0 : n \geq 1)$ *et* $(b_n : n \geq 1)$ *pour lesquelles la suite* $\left(\frac{M_n - b_n}{a_n} : n \geq 1\right)$ *converge en loi vers une distribution* G *non dégénérée, alors* G *appartient à la famille de la loi GEV. Autrement dit,*

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{M_n - b_n}{a_n} \leq x\right] = \mathcal{GEV}(x|\mu,\sigma,\xi),\ \forall x \in \mathbb{R}. \tag{1.11}$$

Ce résultat est très utile en pratique lorsque l'on modélise le maximum d'une distribution inconnue pour laquelle les constantes de normalisation sont elles aussi inconnues. En effet, ce résultat permet de ne considérer qu'une seule famille de lois pour le maximum, la famille de la loi GEV, plutôt que de devoir considérer les trois types puis de déterminer *a posteriori* le plus approprié. Notons qu'en termes de la fonction de répartition F, l'équation (1.11) peut s'exprimer de la façon suivante :

$$\lim_{n\to\infty} F^n(a_n x + b_n) = \mathcal{GEV}(x|\mu,\sigma,\xi),\ \forall x \in \mathbb{R}. \tag{1.12}$$

La loi GEV possède la propriété de linéarité qui est présentée dans la proposition suivante :

Proposition 1.1. *Soit la variable aléatoire* X *telle que*

$$X \sim \mathcal{GEV}(\mu,\sigma,\xi)$$

et deux réels $c > 0$ *et* d*. Alors, on a que*

$$cX + d \sim \mathcal{GEV}(c\mu + d, c\sigma, \xi).$$

Démonstration.

$$\mathbb{P}[cX + d \leq x] = \mathbb{P}\left[X \leq \frac{x-d}{c}\right] = \exp\left\{-\left[1+\xi\left(\frac{\frac{x-d}{c}-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$
$$= \exp\left\{-\left[1+\xi\left(\frac{x-c\mu-d}{c\sigma}\right)\right]^{-1/\xi}\right\} = \mathcal{G}EV(x|c\mu+d, c\sigma, \xi).$$

□

Définition 1.2. *Le* domaine d'attraction *de la loi $\mathcal{G}EV(0,1,\xi)$ pour $\xi \in \mathbb{R}$ est l'ensemble des lois de probabilité F pour lesquelles il existe des constantes normalisantes $a_n > 0$ et b_n telles que*

$$\lim_{n\to\infty} F^n(a_n x + b_n) = \mathcal{G}EV(x|0,1,\xi), \ \forall x \in \mathbb{R}. \tag{1.13}$$

En termes d'une suite de variables aléatoires indépendantes et identiquement distribuées selon la loi F, la condition exprimée par l'équation précédente est équivalente à la suivante :

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{M_n - b_n}{a_n} \leq x\right] = \mathcal{G}EV(x|0,1,\xi), \ \forall x \in \mathbb{R}. \tag{1.14}$$

Le domaine d'attraction de la loi extrême de type I contient notamment la loi exponentielle, la loi normale et toutes transformations monotones de la loi normale. Quant au domaine d'attraction de la loi extrême de type II, il contient par exemple les lois de Pareto et de Cauchy. Le domaine d'attraction de la loi extrême de type III contient en autre les lois uniforme et exponentielle tronquée. Il arrive également qu'aucune constante de normalisation n'existe pour certaines lois. C'est notamment le cas pour la loi géométrique et pour la loi de Poisson.

La théorie des valeurs extrêmes fournit également la distribution limite de l'excédent au-dessus d'un seuil suffisamment élevé u. Définissons d'abord la loi de Pareto généralisée.

Définition 1.3. *La* loi de Pareto généralisée *avec les paramètres d'échelle $\sigma > 0$ et de forme $\xi \in \mathbb{R}$, dénotée par $\mathcal{G}P(\sigma,\xi)$, est la loi de probabilité ayant la fonction de répartition $\mathcal{G}P(x|\sigma,\xi)$ donnée par :*

$$\mathcal{G}P(x|\sigma,\xi) = 1 - \left(1+\xi\frac{x}{\sigma}\right)^{-1/\xi}; \tag{1.15}$$

pour tous les $x > 0$ si $\xi \geq 0$ et pour tous les $x \in (0, -\sigma/\xi)$ si $\xi < 0$. Le cas $\xi = 0$ est défini par continuité. Plus précisément,

- *Le cas $\xi = 0$:*

$$\mathcal{G}P(x|\sigma,0) = \begin{cases} 0 & si\ x \leq 0; \\ 1 - \exp\left(-\frac{x}{\sigma}\right) & si\ x > 0; \end{cases} \tag{1.16}$$

- *Le cas $\xi > 0$:*

$$\mathcal{G}P(x|\sigma,\xi) = \begin{cases} 0 & si\ x \leq 0; \\ 1 - (1+\xi\frac{x}{\sigma})^{-1/\xi} & si\ x > 0; \end{cases} \tag{1.17}$$

- *Le cas $\xi < 0$:*

$$\mathcal{G}P(x|\sigma,\xi) = \begin{cases} 0 & si\ x \leq 0; \\ 1 - \left(1 + \xi\frac{x}{\sigma}\right)^{-1/\xi} & si\ x \in (0, -\sigma/\xi); \\ 1 & si\ x \geq -\sigma/\xi. \end{cases} \tag{1.18}$$

La proposition suivante est l'analogue de la proposition 1.1. La démonstration élémentaire est omise.

Proposition 1.2. *Soit la variable aléatoire X telle que*

$$X \sim \mathcal{G}P(\sigma,\xi)$$

et un réel $c > 0$. Alors, on a que

$$cX \sim \mathcal{G}P(c\sigma,\xi).$$

Le théorème suivant met en évidence le lien fondamental entre la loi de Pareto généralisée et la loi des valeurs extrêmes généralisée.

Théorème 1.3 (Pickands III, 1975)**.** *Soit $\xi \in \mathbb{R}$. Soit F une fonction de répartition sur $\mathbb{R}$. Soit x_+ le suprémum du support de la distribution F, c'est-à-dire $x_+ = \sup\{x \in \mathbb{R} : F(x) < 1\}$. Soit X une variable aléatoire distribuée selon la loi F. Alors, les deux conditions suivantes sont équivalentes :*

(i) La loi F appartient au domaine d'attraction de la loi $\mathcal{G}EV(0,1,\xi)$.

(ii) Il existe une fonction positive $c(u)$ telle que

$$\lim_{u\uparrow x_+} \mathbb{P}\left[\frac{X-u}{c(u)} \leq y \middle| X > u\right] = \mathcal{G}P(y|1,\xi),\ \forall y \in \mathbb{R}. \tag{1.19}$$

1.1.2 Le cas non indépendant

Dans la plupart des applications pratiques, les conditions sur l'indépendance et la stationnarité des suites de variables aléatoires sont trop contraignantes pour que le résultat du théorème 1.2 puisse s'appliquer. Cette section concerne la distribution limite du maximum d'une suite de variables aléatoires stationnaires.

Définition 1.4. *La suite de variables aléatoires $X_1, , X_2 \ldots$, est dite* stationnaire *si pour tout $j \geq 1$, la distribution du vecteur aléatoire $(X_n, X_{n+1}, \ldots, X_{n+j})$ ne dépend pas de n.*

Remarque :

L'extension au cas non indépendante est très important dans l'étude des précipitations. Par exemple, le soulèvement lent et à grande échelle d'une masse d'air chargée d'humidité peut générer un événement de précipitations dites stratiformes qui dure généralement quelques jours. Ce processus induit la plupart du temps une corrélation temporelle de quelques jours sur les précipitations journalières.

Il sera exposé que le résultat du théorème de la loi des valeurs extrêmes généralisée tient encore pour des suites de variables aléatoires strictement stationnaires qui possèdent une dépendance à courte portée et limitée dans la queue de la distribution. Pour énoncer cette condition sur la dépendance, le concept de *strong mixing* introduit par Rosenblatt (1956) est utile :

Définition 1.5 (Strong mixing)**.** *On dit de la suite de variables aléatoires $X_1, X_2, \ldots$, qu'elle satisfait la condition de* strong mixing *s'il existe une fonction $g(k)$ qui tend vers 0 lorsque $k \to \infty$ et qui est telle que*

$$|\mathbb{P}[A \cap B] - \mathbb{P}[A]\mathbb{P}[B]| \leq g(k); \tag{1.20}$$

pour $A \in \mathscr{F}(X_1, \ldots, X_p)$ et $B \in \mathscr{F}(X_{p+k+1}, \ldots)$ où $\mathscr{F}(\cdot)$ dénote la tribu générée par les variables aléatoires indiquées et pour tout entiers positifs p et k.

Le concept de *strong mixing* est souvent utilisé pour restreindre la dépendance d'une suite de variables aléatoires stationnaires. Lorsqu'une suite de variables aléatoires satisfait cette condition, tout événement A basé sur les p premières variables aléatoires est approximativement indépendant de tout événement B basé sur les variables suivant X_{p+k} lorsque k est grand. L'indépendance correspond au cas où l'équation (1.20) est satisfaite avec $g(k) = 0$ pour tout $k \geq 0$.

La condition de *strong mixing* est uniforme dans le sens où elle ne dépend pas des événements A et B considérés. Or, les événements d'intérêt dans la théorie des valeurs extrêmes sont les événements $(X_i \leq u)$ puisque l'on s'intéresse particulièrement à l'événement $(M_n \leq u) = \cap_{i=1}^{n}(X_i \leq u)$. La condition de *strong mixing* peut donc être amoindrie dans le but de restreindre davantage la condition sur la dépendance en ne considérant que ce type d'événements. Leadbetter et collab. (1983) dénomment cette condition modifiée la *condition D*, qu'ils énoncent comme suit :

Définition 1.6 (La condition D)**.** *On dit que la condition D est satisfaite si pour n'importe quels entiers*

$$1 \leq i_1 < \ldots < i_p < j_1 < \ldots < j_q < \infty$$

tels que $j_1 - i_p > k$ et pour tout réel u, on a

$$\left| \mathbb{P}\left[\bigcap_{l=1}^{p} (X_{i_l} \leq u) \cap \bigcap_{m=1}^{q} (X_{j_m} \leq u) \right] - \mathbb{P}\left[\bigcap_{l=1}^{p} (X_{i_l} \leq u) \right] \times \mathbb{P}\left[\bigcap_{m=1}^{q} (X_{j_m} \leq u) \right] \right| \leq g(k), \tag{1.21}$$

où $g(k) \to 0$ lorsque $k \to \infty$.

Bien que la condition D diminue les restrictions sur la dépendance par rapport au *strong mixing*, celles-ci peuvent l'être davantage en considérant non pas tout réel u mais seulement certaines valeurs u_n suffisamment élevées. En effet, la théorie des valeurs extrêmes s'intéresse

aux événements $(X_i \leq u : 1 \leq i \leq n)$ pour u très grand. Cette nouvelle condition sur la dépendance, dénommée *condition* $D(u_n)$ par Leadbetter et collab. (1983, chap. 3), ne considère que certaines suites $(u_n : n \geq 1)$ particulières. Elle s'énonce comme suit :

Définition 1.7 (La condition $D(u_n)$)**.** *Soit* $(u_n : n \geq 1)$ *une suite de nombres réels. On dit que la condition* $D(u_n)$ *est satisfaite si pour n'importe quels entiers*

$$1 \leq i_1 < \ldots < i_p < j_1 < \ldots < j_q \leq n$$

tels que $j_1 - i_p > k$*, on a*

$$\left| \mathbb{P}\left[\bigcap_{l=1}^{p} (X_{i_l} \leq u_n) \cap \bigcap_{m=1}^{q} (X_{j_m} \leq u_n) \right] - \mathbb{P}\left[\bigcap_{l=1}^{p} (X_{i_l} \leq u_n) \right] \times \mathbb{P}\left[\bigcap_{m=1}^{q} (X_{j_m} \leq u_n) \right] \right| \leq \alpha_{n,k_n}, \tag{1.22}$$

où $\alpha_{n,k_n} \to 0$ *lorsque* $n \to \infty$*, pour une certaine suite* $(k_n : n \geq 1)$ *telle que* $k_n/n \to 0$ *lorsque* $n \to \infty$*.*

La condition $D(u_n)$ sera utilisée pour des seuils u_n qui augmenteront avec n. En ce sens, cette condition adoucit les restrictions sur la dépendance des événements $(X_i \leq u)$ à une dépendance à portée limitée dans la queue de la distribution F, c'est-à-dire à une dépendance à portée limitée des événements $(X_i \leq u_n)$ pour u_n suffisamment grand. Il est à noter que la condition de *strong mixing* implique la condition D et que cette dernière implique la condition $D(u_n)$.

Le théorème fondamental de la théorie des valeurs extrêmes peut être généralisé aux séries stationnaires pour lesquelles les variables aléatoires qui les composent ne sont pas indépendantes. Le théorème suivant constitue cette généralisation.

Théorème 1.4 (Leadbetter et collab., 1983, chap. 2) **.** *Soit* $X_1, X_2, \ldots$ *une suite de variables aléatoires stationnaire et soit* $M_n = \max\{X_1, \ldots, X_n\}$ *le maximum des n premières variables. S'il existe des suites de constantes normalisantes* $(a_n > 0 : n \geq 1)$ *et* $(b_n : n \geq 1)$ *pour lesquelles la suite* $(\frac{M_n - b_n}{a_n})$ *converge en loi vers une distribution G non dégénérée et si la condition* $D(u_n)$ *est satisfaite avec* $u_n = a_n x + b_n$ *pour tout* $x \in \mathbb{R}$*, alors G appartient à la famille de la loi GEV.*

Remarque :

> *Puisque les précipitations saisonnières ne sont pas indépendantes, le théorème précédent est très utile car il stipule que le maximum saisonnier est tout de même distribué selon la loi GEV. En effet, il est naturel de considérer que la condition* $D(u_n)$ *soit satisfaite pour les précipitations étant donné que le processus générateur n'induit une dépendance que de quelques jours. Plus particulièrement, on*

suppose généralement que les observations séparées d'au moins quelques jours, disons m jours, sont indépendantes :

$$X_i \perp X_j, \text{ pour } |i-j| > m,$$

ce qui satisfait la condition de strong mixing *et ainsi les conditions D et* $D(u_n)$.

Dans le cas non-stationnaire, les caractéristiques de la suite de variables aléatoires peuvent changer au cours du *temps*. Il n'existe cependant pas de théorie générale qui s'applique aux suites non stationnaires contrairement au cas non indépendant.

1.1.3 La modélisation statistique des maxima

Cette section illustre comment les résultats des théorèmes 1.2 et 1.4 s'utilisent en pratique. Dans la majorité des applications, la distribution sous-jacente F des données est inconnue. Il est donc impossible de déterminer si les constantes normalisantes a_n et b_n existent. Cependant, afin d'appliquer les résultats de la théorie des valeurs extrêmes, il est nécessaire de supposer leur existence. Selon le théorème 1.2, si n est grand, l'approximation suivante est raisonnable :

$$\mathbb{P}\left[\frac{M_n - b_n}{a_n} \leq x\right] \overset{\mathcal{L}}{\approx} \mathcal{G}EV(x|0,1,\xi). \tag{1.23}$$

Selon cette dernière approximation, on peut approximer la distribution du maximum non normalisé par la distribution suivante :

$$\mathbb{P}\left[M_n \leq x\right] = \mathbb{P}\left[\frac{M_n - b_n}{a_n} \leq \frac{x - b_n}{a_n}\right] \overset{\mathcal{L}}{\approx} \mathcal{G}EV\left(\frac{x-b_n}{a_n}\middle|0,1,\xi\right) = \mathcal{G}EV\left(x\,|b_n,a_n,\xi\right). \tag{1.24}$$

Le cas stationnaire

Soit $X_1,\ldots,X_m$ une suite de variables aléatoires stationnaire pour laquelle les conditions du théorème 1.4 sont satisfaites et où $m = K \times L$ pour deux entiers positifs K et L. La suite des variables aléatoires peut être partitionnée en K blocs de même taille L : $(X_{kl} : 1 \leq l \leq L), 1 \leq k \leq K$. Si la taille L de ces K blocs est suffisamment grande, alors le maximum de chacun des blocs, dénoté par

$$M_k = \max\{X_{k1},\ldots,X_{kL}\},\ 1 \leq k \leq K; \tag{1.25}$$

peut être supposé approximativement distribué selon la loi GEV en vertu du théorème 1.4. Puisque la condition $D(u_n)$ restreint la dépendance à longue portée dans la queue de la distribution des X_i, il est raisonnable de supposer que la suite des maxima est indépendante si la taille des blocs est suffisamment grande. L'estimation des paramètres de la loi GEV peut se faire notamment en maximisant la fonction de vraisemblance suivante :

$$\mathcal{L}(\mu,\sigma,\xi) = \prod_{k=1}^{K} gev(m_k|\mu,\sigma,\xi); \tag{1.26}$$

où l'expression $gev(x|\mu, \sigma, \xi)$ dénote la densité de la loi GEV de paramètres (μ, σ, ξ) évaluée en x.

La taille des blocs joue un rôle important dans la méthodologie. Des blocs trop étroits peuvent invalider l'approximation asymptotique de la loi du maximum par la loi GEV tandis que des blocs trop larges limitent le nombre de maxima disponibles pour l'estimation des paramètres.

Remarque :

Dans l'étude des précipitations et plus généralement pour la plupart des variables liées au climat, l'effet du cycle annuel de la Terre autour du Soleil suggère une partition annuelle des observations. En ce sens, un bloc est composé de l'ensemble des observations d'une année. L'approximation d'indépendance de la collection de maxima échantillonnés sur des années différentes est généralement considérée tout à fait raisonnable dans ce contexte.

Le cas non stationnaire

Soit $X_1, \ldots, X_m$ une suite de variables aléatoires non stationnaire partionnées en K blocs de taille L : $(X_{kl} : 1 \leq l \leq L), 1 \leq k \leq K$, où $m = K \times L$. Chacune des sous-suites de variables aléatoires composant les K blocs satisfait les conditions du théorème 1.4. La non-stationnarité se manifeste d'un bloc à l'autre. En supposant que L est suffisamment grand, la loi du maximum de chacun des blocs est approximée par une loi GEV. Or, comme le processus aléatoire est non stationnaire, les paramètres de la loi GEV peuvent être différents d'un bloc à l'autre. L'approche statistique proposée par Coles (2001) consiste à modéliser les paramètres de la GEV en fonction du bloc k à partir duquel le maximum a été extrait. La vraisemblance peut s'écrire de la façon suivante :

$$\mathcal{L}\{(\mu_k, \sigma_k, \xi_k), 1 \leq k \leq K\} = \prod_{k=1}^{K} \mathcal{G}EV(m_k|\mu_k, \sigma_k, \xi_k). \tag{1.27}$$

Remarque :

La non-stationnarité dans les séries de précipitations peut se manifester au cours d'une même année dû à la saisonnalité. Par exemple, durant les mois hivernaux, très peu de précipitations convectives sont générées en raison d'un manque d'énergie solaire incidente alors que durant les mois estivaux, ce processus est prépondérant. Par conséquent, il est astucieux de partitionner à nouveau les précipitations annuelles en précipitations saisonnières pour enlever cet effet des saisons. Les précipitations saisonnières d'une même année peuvent ainsi être supposées stationnaires. La non-stationnarité dans les séries de précipitations peut également se manifester à plus long terme au cours des années, causée entre autres par les changements climatiques. Cette non-stationnarité dans la série de maxima peut

être considérée en modélisant les paramètres de la GEV comme proposé dans la présente section.

1.1.4 La modélisation statistique des excédents au-dessus d'un seuil

Cette section illustre comment le résultat du théorème 1.3 s'utilise en pratique. À l'instar de la modélisation des maxima, la fonction de répartition F est la plupart du temps inconnue dans les applications où l'on modélise les excédents au-dessus d'un seuil u. Il est donc impossible de déterminer s'il existe une fonction positive $c(u)$ telle que l'équation (1.19) soit satisfaite. Il est alors nécessaire de supposer son existence, ce qui revient à faire l'hypothèse que F appartient au domaine d'attraction de la loi $GEV(0, 1, \xi)$. Pour une taille d'échantillon n suffisamment grande et un seuil u suffisamment élevé, le résultat du théorème 1.3 implique que l'approximation suivante est raisonnable :

$$\mathbb{P}\left[\frac{X-u}{c(u)} \leq y | X > u\right] \overset{\mathcal{L}}{\approx} \mathcal{G}P(y|1, \xi). \tag{1.28}$$

Alors, la distribution de l'excédent non normalisé peut être approximé par :

$$\mathbb{P}\left[X - u \leq y | X > u\right] = \mathbb{P}\left[\frac{X-u}{c(u)} \leq \frac{y}{c(u)}\middle| X > u\right] \overset{\mathcal{L}}{\approx} \mathcal{G}P\left(\frac{y}{c(u)}\middle| 1, \xi\right) = \mathcal{G}P\left\{y|c(u), \xi\right\}. \tag{1.29}$$

Le cas stationnaire

Soit $X_1, \ldots, X_n$ une suite de variables aléatoires stationnaire pour laquelle les conditions du théorème 1.4 sont satisfaites. Pour un seuil u suffisamment élevé, les observations qui dépassent ce seuil sont récupérées $(x_1, \ldots, x_k)$. Puisque la dépendance dans la suite $(X_i : 1 \leq i \leq n)$ pourrait induire plusieurs dépassements groupés (consécutifs ou trop rapprochés), Coles (2001) propose de filtrer les dépassements groupés en ne conservant que le maximum du groupe de façon à ne pas inclure la dépendance dans la fonction de vraisemblance des excédents. Les dépassements ainsi filtrés $(x_1, \ldots, x_{k'})$ étant suffisamment séparés dans le temps peuvent par la suite être supposés approximativement indépendants par la condition $D(u_n)$. Le résultat du théorème 1.3 permet d'écrire la vraisemblance de ces dépassements comme suit afin d'estimer les paramètres de la loi de Pareto généralisée :

$$\mathcal{L}(\sigma, \xi) = \prod_{j=1}^{k'} gp(x_j - u|\sigma, \xi); \tag{1.30}$$

où $gp(x|\sigma, \xi)$ dénote la densité de la loi $\mathcal{G}P(\sigma, \xi)$.

Le choix du seuil est crucial dans cette méthodologie. D'une part, un seuil trop bas entraînerait l'invalidité de l'approximation asymptotique consitant à modéliser les excédents par la loi de Pareto généralisée. Dautre part, un seuil trop élévé limiterait le nombre d'excédents disponibles pour l'estimation des paramètres.

Pour être complet, le modèle statistique doit aussi inclure une modélisation pour l'occurrence des dépassements filtrés. Le modèle naturel consiste à supposer que pour chacune des variables aléatoires X_t, la probabilité que X_t dépasse le seuil est égale à p. Ceci est une conséquence de la stationnarité. On a que $\mathbb{P}[X_i > u]$ ne dépend pas de i mais uniquement de u. Dans le cas indépendant, le nombre total de dépassements de seuil suit une loi binomiale avec le nombre d'essais égal à T et la probabilité de succès égale à p. L'estimation de la probabilité de succès peut se faire en maximisant la fonction de vraisemblance suivante :

$$\mathcal{L}(p) = \mathcal{B}inomiale(k|T, p). \tag{1.31}$$

Dans le cas non indépendant, Leadbetter et collab. (1983) démontrent que sous certaines conditions le nombre de dépassements peut être approximé par un processus de Poisson. L'intensité τ de ce processus peut être estimée en maximisant la fonction de vraisemblance suivante :

$$\mathcal{L}(\tau) = \mathcal{P}oisson(k|\tau). \tag{1.32}$$

Ce modèle statistique qui caractérise à la fois l'occurrence des dépassements et l'intensité de ceux-ci est généralement dénommé par *Peaks-Over-Threshold* (POT) dans la littérature.

Remarque :

> *Pour l'étude des précipitations extrêmes, le modèle POT est utilisé lorsque l'on suspecte que plusieurs événements attribuables au domaine des extrêmes peuvent survenir dans une même saison. Le modèle POT permet d'intégrer tous ces événements contrairement au modèle GEV n'utilisant que le maximum saisonnier.*

Le cas non stationnaire

Soit $X_1, \ldots, X_n$ une suite de variables aléatoires non stationnaire. En se basant sur l'approximation (1.29), Davison et Smith (1990) ont proposé de modéliser les excédents non stationnaires au-dessus d'un seuil constant u de la façon suivante :

$$\mathbb{P}[X_i - u \leq y | X_i > u] = \mathcal{G}P(y|\sigma_i, \xi_i). \tag{1.33}$$

Il est supposé que le seuil u est suffisamment élevé pour toutes les variables aléatoires X_i. Dans le cas contraire, l'approximation de la distribution de l'excédent par la loi de Pareto généralisée ne serait pas appropriée. Dans le cas non indépendant, les dépassements groupés doivent être filtrés de façon à simplifier l'expression de la vraisemblance. Soit les excédents filtrés $(x_1 - u, \ldots, x_{k'} - u)$ survenus aux temps $(t_1, \ldots, t_{k'})$, les paramètres de la loi de Pareto généralisée peuvent être estimés en maximisant la fonction de vraisemblance suivante :

$$\mathcal{L}\left\{(\sigma_{t_j}, \xi_{t_j}), 1 \leq j \leq k'\right\} = \prod_{j=1}^{k'} gp(x_{t_j} - u|\sigma_{t_j}, \xi_{t_j}). \tag{1.34}$$

L'occurrence des dépassements doit également être modélisée. Puisque la suite est non stationnaire, la probabilité de dépassement du seuil, $\mathbb{P}[X_i > u]$, peut évoluer avec l'indice i. Le nombre de dépassements peut être modélisé par un processus de Poisson non homogène.

On peut trouver dans la littérature deux autres approches populaires pour modéliser les excédents de seuil d'une suite non stationnaire. La première consiste à considérer un seuil évolutif u_t de façon à stabiliser les paramètres de la loi de Pareto généralisée pour l'ensemble des excédents. Cette méthode repose sur l'approximation implicite suivante

$$\mathbb{P}[X_t \leq u_t + y | X_t > u_t] \approx 1 - \left(1 + \frac{\xi}{\sigma} y\right)^{-1/\xi}. \tag{1.35}$$

Il est cependant assez difficile en pratique de définir un seuil pour lequel les paramètres de la loi de Pareto généralisée se stabilisent. La seconde approche est plus rigoureuse. Elle modélise conjointement la probabilité de dépassement du seuil et l'intensité de celui-ci par un processus ponctuel marqué. Une introduction sur cette dernière approche se trouve dans Coles (2001, chapitre 7).

1.2 La notion de graphe simple

Dans le cadre de cette thèse, le domaine spatial étudié est composé de 180×200 points d'observations disposés sur une grille régulière couvrant l'Amérique du Nord. Puisque nous n'étudions que les précipitations simulées sur les 12 570 points de grille terrestres du domaine spatial, la grille résultante est interprétée comme une grille régulière mais dont les valeurs aux sites de mer sont manquantes. Cette grille peut être représentée par un graphe simple $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, où $\mathcal{V}$ constitue l'ensemble des sommets du graphe et $\mathcal{E}$ constitue l'ensemble des arêtes du graphe. Les points de grille jouent le rôle des sommets. Une arête dans un graphe simple est constituée d'un couple de sommets distincts. Deux sommets i et j sont dits *connectés* si l'arête $\{i, j\}$ constitue un élément de l'ensemble $\mathcal{E}$. Les arêtes permettent de définir le voisinage d'un sommet. Le voisinage du sommet $i \in \mathcal{V}$, dénoté δ_i, est défini comme étant l'ensemble des sommets $j \in \mathcal{V}$ qui y sont connectés :

$$\delta_i = \{j \in \mathcal{V} : \{i, j\} \in \mathcal{E}\}. \tag{1.36}$$

Le nombre de voisins d'un sommet i, dénoté n_i, est défini comme étant la cardinalité de son voisinage :

$$n_i = \text{Card}(\delta_i). \tag{1.37}$$

Généralement la structure du graphe est utile pour définir la structure de dépendance qui existe entre les différents points de grille. Plusieurs définitions de voisinage peuvent être considérées sur une grille régulière, par exemple le voisinage de premier ordre qui comprend les quatre points de grille les plus rapprochés ou celui de deuxième ordre qui comprend les 8 points de grille les plus rapprochés.

1.3 Introduction aux modèles hiérarchiques bayésiens

Un modèle hiérarchique bayésien est un modèle statistique bayésien $\left(f_{[X|\boldsymbol{\theta}=\theta]}(x), f_{\boldsymbol{\theta}}(\theta)\right)$ composé d'une vraisemblance $f_{[X|\boldsymbol{\theta}=\theta]}(x)$ et d'une loi *a priori* $f_{\boldsymbol{\theta}}(\theta)$ pour le vecteur de paramètres $\boldsymbol{\theta}$ où la loi *a priori* est décomposée en une succession de sous-modèles conditionnels de la façon suivante :

$$f_{\boldsymbol{\theta}}(\theta) = \int_{\phi_1,\ldots,\phi_m} f_{[\boldsymbol{\theta}|\boldsymbol{\phi}_1=\phi_1]}(\theta) \times f_{[\boldsymbol{\phi}_1|\boldsymbol{\phi}_2=\phi_2]}(\phi_1) \times \ldots \times f_{[\boldsymbol{\phi}_{m-1}|\boldsymbol{\phi}_m=\phi_m]}(\phi_{m-1}) \times f_{\boldsymbol{\phi}_m}(\phi_m)\, d\phi_1 \ldots d\phi_m.$$

Le processus de conditionnement peut se poursuivre autant de fois que nécessaire mais il est rarement utile, en pratique, de considérer plus de deux niveaux. En effet, plutôt que d'ajouter des niveaux supplémentaires, et par le fait même des paramètres supplémentaires, il est préférable la plupart du temps de considérer la loi *a priori* qui intègre les niveaux supérieurs.

Les modèles hiérarchiques bayésiens permettent de construire des modèles complexes en utilisant une succession de modèles conditionnels généralement plus simples. La structure hiérarchique permet dans plusieurs cas de faciliter l'interprétation du modèle. Banerjee et collab. (2004) fournissent une multitude d'exemples d'application où les modèles hiérarchiques peuvent être avantageusement utilisés. Dans le cadre de la thèse, les modèles hiérarchiques bayésiens seront utilisés pour modéliser la dépendance spatiale des paramètres marginaux d'un modèle de valeurs extrêmes.

Remarque :

> *Les précipitations générées par le Modèle régional canadien du climat sur la grille couvrant l'Amérique du Nord sont spatialement dépendantes. Il en résulte que les paramètres marginaux d'un modèle de valeurs extrêmes ajusté à chacun des points de grille montreront également une structure de dépendance spatiale. La modélisation de la dépendance spatiale des paramètres marginaux par un modèle hiérarchique bayésien permettra une amélioration de l'estimation de ceux-ci.*

1.3.1 Application des modèles hiérarchiques bayésiens pour des données spatiales

Soit le graphe simple $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ illustrant le domaine spatial. Soit $X = \{X_i : i \in \mathcal{V}\}$ un ensemble de vecteurs aléatoires $X_i = (X_{i,1}, \ldots, X_{i,n})$ de taille n pour lesquels la densité est $f_{[X_i|\boldsymbol{\theta}_i=\theta_i]}(x_i)$. S'il existe une dépendance spatiale des $\{X_i : i \in \mathcal{V}\}$, il est alors avantageux de modéliser conjointement l'ensemble des vecteurs aléatoires $\{X_i : i \in \mathcal{V}\}$ dans le but d'améliorer l'estimation des paramètres par rapport à une estimation marginale pour chacun des sommets i. Généralement il est très difficile de modéliser directement cette dépendance de l'ensemble

des vecteurs aléatoires. Une approche raisonnable consiste à utiliser un modèle hiérarchique bayésien pour modéliser la dépendance spatiale de l'ensemble des paramètres marginaux $\boldsymbol{\theta} = \{\boldsymbol{\theta}_i : i \in \mathcal{V}\}$ de chacun des sommets. Cette approche repose sur l'hypothèse que l'ensemble des paramètres $\boldsymbol{\theta}_i$ qui caractérisent X contient également une partie de l'information concernant leur dépendance spatiale. La distribution conjointe utilisée pour modéliser la dépendance spatiale des $\boldsymbol{\theta}_i$ peut être une distribution avec laquelle il est facile de travailler. Généralement on utilise des lois multidimensionnelles pour lesquelles les densités possèdent une expression analytique simple, notamment la loi normale et la loi de Student.

Dans les modèles hiérarchiques, il est généralement supposé que les données localisées aux différents sommets $i \in \mathcal{V}$ sont conditionnellement indépendantes sachant les paramètres spatiaux θ (Banerjee et collab., 2004). Cette hypothèse d'indépendance conditionnelle permet d'écrire la vraisemblance de l'ensemble des données observées de la manière suivante :

$$f_{[X|\boldsymbol{\theta}=\theta]}(x) = \prod_{i \in \mathcal{V}} f_{[X_i|\boldsymbol{\theta}=\theta]}(x_i). \tag{1.38}$$

Le modèle hiérarchique bayésien peut donc s'écrire de la façon suivante :

$$f_{(X,\boldsymbol{\theta},\boldsymbol{\phi})}(x,\theta,\phi) = f_{[X|\boldsymbol{\theta}=\theta,\boldsymbol{\phi}=\phi]}(x) \times f_{[\boldsymbol{\theta}|\boldsymbol{\phi}=\phi]}(\theta) \times f_{\boldsymbol{\phi}}(\phi); \tag{1.39}$$

où $\boldsymbol{\phi}$ est le vecteur de paramètres modélisant la loi conjointe des paramètres marginaux $\boldsymbol{\theta}$ et $f_{\boldsymbol{\phi}}(\phi)$ la loi *a priori* de ces paramètres. L'hypothèse d'indépendance conditionnelle permet d'écrire :

$$f_{[X|\boldsymbol{\theta}=\theta,\boldsymbol{\phi}=\phi]}(x) = f_{[X|\boldsymbol{\theta}=\theta]}(x). \tag{1.40}$$

1.4 Introduction aux champs de Markov gaussiens

Dans le cadre de cette thèse, les champs de Markov gaussiens sont utilisés dans un modèle hiérarchique bayésien pour modéliser la dépendance spatiale des paramètres marginaux d'un modèle de valeurs extrêmes. Dans cette section, les principales propriétés de ces modèles statistiques seront présentées et il sera aussi montré comment la propriété de Markov simplifie le modèle statistique par rapport aux modèles gaussiens plus généraux pour des données disposées sur une grille.

1.4.1 Champs gaussiens

En géostatistique, les champs aléatoires gaussiens sont largement utilisés, notamment en raison des simplifications mathématiques que procurent l'utilisation de la loi normale. En plus de posséder une expression analytique simple pour le cas multidimensionnel, toutes les lois conditionnelles et marginales qui en découlent sont elles aussi normales. En géostatistique, une nette disctinction est faite entre les champs gaussiens et les champs de Markov gaussiens (Cressie, 2003), bien que ces derniers ne soient que des cas particuliers de champs gaussiens.

Généralement les champs gaussiens sont utilisés lorsque les données sont disposées irrégulièrement dans l'espace mais où elles peuvent en principe être échantillonnées en tout point d'un domaine continu. Les champs de Markov gaussiens sont plutôt utilisés dans un contexte où les données sont organisées spatialement sous la forme d'une grille. Les deux approches se distinguent par la façon dont la matrice de covariance est modélisée. Dans le cas des champs gaussiens, la matrice de covariance est généralement modélisée par un variogramme qui correspond à une mesure de variance en fonction de la distance entre les sites. Dans le cas des champs de Markov, une hypothèse d'indépendance conditionnelle est plutôt utilisée pour modéliser l'inverse de la matrice de covariance.

1.4.2 Champs de Markov gaussiens

Un champ de Markov gaussien défini sur un graphe $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ est un vecteur aléatoire $X = (X_i, i \in \mathcal{V})$ dont la densité est une loi normale multidimensionnelle avec une matrice de covariance particulière. La propriété de Markov est héritée de l'hypothèse d'indépendance conditionnelle posée sur les sommets du graphe. L'indépendance conditionnelle pour trois vecteurs de variables aléatoires est définie ci-dessous. La proposition 1.3 énonce une conséquence de cette définition.

Définition 1.8 (Indépendance conditionnelle)**.** *Soit trois vecteurs de variables aléatoires X, Y et Z ayant pour densité conjointe la densité $f_{XYZ}(x, y, z)$. On dit que X et Y sont conditionnellement indépendants à Z, et on note*

$$X \perp Y \mid Z;$$

lorsque la densité conditionnelle de X et de Y sachant $Z = z$ peut s'écrire de la façon suivante

$$f_{[(X,Y)|Z=z]}(x, y) = f_{[X|Z=z]}(x) \times f_{[Y|Z=z]}(y).$$

Proposition 1.3 (Rue et Held, 2005, chap. 2)**.** *Il existe des fonctions g et h telles que*

$$X \perp Y \mid Z \quad \Leftrightarrow \quad f_{(X,Y,Z)}(x, y, z) = g(x, z) \times h(y, z);$$

pour $\{z : f_Z(z) > 0\}$.

Démonstration.
Par hypothèse, on a que

$$f_{(X,Y,Z)}(x, y, z) = f_{[X|Z=z]}(x) \times f_{[Y|Z=z]}(y) \times f_Z(z). \tag{1.41}$$

On peut définir de la façon suivante g et h pour démontrer l'implication directe :

$$g(x, z) = f_{[X|Z=z]}(x); \tag{1.42}$$

$$h(y, z) = f_{[Y|Z=z]}(y) \times f_Z(z). \tag{1.43}$$

Pour l'implication inverse, on a que

$$f_{(X,Y,Z)}(x,y,z) = g(x,z) \times h(y,z). \tag{1.44}$$

Alors,

$$f_{(X,Z)}(x,z) = \int_y g(x,z) \times h(y,z)\ dy = g(x,z) \int_y h(y,z)\ dy;$$

et

$$f_{[X|Z=z]}(x) = \frac{g(x,z)}{f_Z(z)} \int_y h(y,z)\ dy. \tag{1.45}$$

De façon équivalente, on a que

$$f_{[Y|Z=z]}(y) = \frac{h(y,z)}{f_Z(z)} \int_y g(x,z)\ dx. \tag{1.46}$$

On peut également écrire que

$$f_Z(z) = \int_x \int_y g(x,z) \times h(y,z)\ dydx = \int_x g(x,z)\ dx \int_y h(y,z)\ dy. \tag{1.47}$$

En utilisant les trois dernières équations, on obtient finalement que

$$f_{(X,Y,Z)}(x,y,z) = f_{[X|Z=z]}(x) \times f_{[Y|Z=z]}(y) \times f_Z(z); \tag{1.48}$$

ce qui démontre l'implication inverse. □

Dans la définition des champs de Markov, une hypothèse d'indépendance conditionnelle est supposée selon la structure du graphe $\mathcal{G}$. Cette hypothèse s'exprime de la façon suivante :

$$f_{[X_i|X_{-i}=x_{-i}]}(x_i) = f_{[X_i|X_{\delta_i}=x_{\delta_i}]}(x_i),\ \forall i \in \mathcal{V}; \tag{1.49}$$

où X_{-i} dénote l'ensemble des variables aléatoires à l'exception de X_i et $X_{\delta_i} = (X_j : j \in \delta_i)$. Par conséquent, l'ensemble des arêtes du graphe $\mathcal{E}$ définit la structure d'indépendance conditionnelle des variables aléatoires.

Avant d'introduire formellement les champs de Markov gaussiens, il est utile de définir une paramétrisation différente de celle couramment utilisée pour la loi normale multidimensionnelle.

Définition 1.9 (Loi normale multidimensionnelle)**.** *Soit $\mu = (\mu_1, \ldots, \mu_n)^T \in \mathbb{R}^n$ et Q une matrice de précision symétrique définie positive de taille $n \times n$. On dit du vecteur aléatoire $X = (X_1, \ldots, X_n)^T$ de densité*

$$f_X(x) = \frac{1}{(2\pi)^{n/2}} |Q|^{1/2} \exp\left\{-\frac{1}{2}(x-\mu)^T Q (x-\mu)\right\} \tag{1.50}$$

qu'il est distribué selon une loi normale multidimensionnelle d'espérance μ et de matrice de précision Q. On note cette densité comme suit :

$$f_X(x) = \mathcal{N}_n^*(x|\mu, Q).$$

La densité de l'équation (1.50) ne constitue qu'une paramétrisation différente de la loi normale usuelle dénotée par $\mathcal{N}_n(x|\mu,\Sigma)$, où Σ est une matrice de covariance symétrique définie positive et où la densité s'exprime par

$$f_X(x) = \frac{1}{(2\pi)^{n/2}}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}.$$

Plus précisément, on a que

$$\mathcal{N}_n^*(x|\mu,Q) = \mathcal{N}_n(x|\mu,Q^{-1}) \tag{1.51}$$

La paramétrisation utilisant la matrice de précision Q possède une propriété intéressante concernant les variables aléatoires conditionnellement indépendantes. Cette propriété est présentée dans la proposition suivante.

Proposition 1.4 (Rue et Held, 2005, chap. 2)**.** *Soit le vecteur aléatoire $X = (X_1,\ldots,X_n)^T$ ayant pour densité la loi normale multidimensionnelle $f_X(x) = \mathcal{N}_n^*(x|\mu,Q)$ de moyenne $\mu = (\mu_1,\ldots,\mu_n)^T$ et de matrice de précision $Q = (q_{ij} : 1 \leq i \leq n, 1 \leq j \leq n)$. Pour tout $i \neq j$, on a*

$$X_i \perp X_j \mid X_{-i,-j} \quad \Leftrightarrow \quad q_{ij} = 0;$$

où $X_{-i,-j} = (X_l : l \notin \{i,j\})$, l'ensemble du vecteur aléatoire à l'exception des composantes i et j.

Démonstration inspirée de Rue et Held (2005, page 32).
Pour simplifier l'écriture, posons $Y = X - \mu$ et $\mathcal{V} = \{1,2,\ldots,n\}$. Fixons $i \neq j$, tous deux dans $\mathcal{V}$. On a alors

$$\begin{aligned}
f_Y(y) &\propto \exp\left(-\frac{1}{2}y^TQy\right);\\
&= \exp\left(-\frac{1}{2}\sum_{k=1}^{n}\sum_{l=1}^{n}y_kq_{kl}y_l\right);\\
&= \exp\Bigg(-\frac{1}{2}y_iq_{ii}y_i - \frac{1}{2}y_jq_{jj}y_j - \frac{1}{2}y_iq_{ij}y_j - \frac{1}{2}y_jq_{ji}y_i\\
&\qquad -\frac{1}{2}\sum_{k\in\mathcal{V}\setminus\{i,j\}}y_kq_{kj}y_j - \frac{1}{2}\sum_{l\in\mathcal{V}\setminus\{i,j\}}y_iq_{il}y_l - \frac{1}{2}\sum_{k\in\mathcal{V}\setminus\{i,j\}}y_kq_{ki}y_i\\
&\qquad -\frac{1}{2}\sum_{l\in\mathcal{V}\setminus\{i,j\}}y_jq_{jl}y_l - \frac{1}{2}\sum_{(k,l)\in\{\mathcal{V}\setminus\{i,j\}\}^2}y_kq_{kl}y_l\Bigg).
\end{aligned}$$

Or, la matrice Q est symétrique, *i.e.* $q_{kl} = q_{lk}$ pour tout k et l. Donc,

$$\begin{aligned}
f_Y(y) \propto \exp\Bigg(&-\frac{1}{2}y_iq_{ii}y_i - \frac{1}{2}y_jq_{jj}y_j - y_iq_{ij}y_j\\
&-y_j\sum_{k\in\mathcal{V}\setminus\{i,j\}}q_{kj}y_k - y_i\sum_{l\in\mathcal{V}\setminus\{i,j\}}q_{il}y_l - \frac{1}{2}\sum_{(k,l)\in\{\mathcal{V}\setminus\{i,j\}\}^2}y_kq_{kl}y_l\Bigg).
\end{aligned}$$

Comme le dernier terme ne dépend pas de y_i ni de y_j, on a que

$$\begin{aligned} f_Y(y) &\propto \exp\left(-y_i q_{ij} y_j\right) \times \exp\left(-\frac{1}{2} y_i q_{ii} y_i - y_i \sum_{l \in \mathcal{V} \setminus \{i,j\}} q_{il} y_l\right) \\ &\quad \times \exp\left(-\frac{1}{2} y_j q_{jj} y_j - y_j \sum_{k \in \mathcal{V} \setminus \{i,j\}} q_{kj} y_k\right); \\ &= \exp\left(-y_i q_{ij} y_j\right) \times g(y_i, y_{-i,-j}) \times h(y_j, y_{-i,-j}). \end{aligned}$$

On peut écrire $f_Y(y)$ sous la forme de la proposition 1.3 si et seulement si $q_{ij} = 0$. □

Cette propriété devient intéressante lorsque la structure de dépendance conditionnelle est telle que la matrice de précision devient creuse, ce qui est généralement le cas pour des données disposées sur une grille. La manipulation numérique de cette matrice peut donc s'effectuer par les algorithmes dédiés aux matrices creuses afin d'améliorer l'efficacité computationnelle. Pour des grilles de grande taille, le gain en efficacité numérique devient encore plus important. Il est à noter que si la matrice de précision Q est creuse, il en est autrement de la matrice de covariance $\Sigma = Q^{-1}$ qui est généralement dense.

La définition d'un champ de Markov gaussien peut maintenant être formellement introduite.

Définition 1.10. *Un* champ de Markov gaussien *défini sur le graphe simple $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ est un vecteur aléatoire $X = (X_i : i \in \mathcal{V})^T$ dont la distribution est la loi normale multidimensionnelle pour laquelle la matrice de précision $Q = (q_{ij} : i \in \mathcal{V}, j \in \mathcal{V})$ respecte la condition suivante :*

$$q_{ij} \neq 0 \text{ si et seulement si ou bien } i = j, \text{ ou bien } j \in \delta_i.$$

Selon la définition précédente, l'ensemble des arêtes du graphe $\mathcal{E}$ dicte la structure des éléments non nuls de la matrice de précision Q. De plus, puisque la loi conjointe est normale multidimensionnelle, toutes les lois conditionnelles et marginales le sont également. La proposition suivante résume quelques propriétés intéressantes des lois conditionnelles d'un champ de Markov gaussien.

Proposition 1.5 (Rue et Held, 2005, chap. 2)**.** *Soit le champ de Markov gaussien $X = (X_i : i \in \mathcal{V})^T$, de moyenne μ et de matrice de précision Q, défini sur le graphe simple $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Alors on a*

(i) $\mathbb{E}[X_i | X_{-i} = x_{-i}] = \mu_i - \frac{1}{q_{ii}} \sum_{j \in \delta_i} q_{ij}(x_j - \mu_j)$;

(ii) $\mathrm{Prec}[X_i | X_{-i} = x_{-i}] = q_{ii}$;

(iii) $\mathbb{C}or[X_i, X_j | X_{-i,-j} = x_{-i,-j}] = -\frac{q_{ij}}{\sqrt{q_{ii} q_{jj}}}$.

Démonstration inspirée de Rue et Held (2005, page 23).
Pour simplifier l'écriture, posons $Y = X - \mu$. Alors,

$$f_{[Y_i|Y_{-i}=y_{-i}]}(y_i) = \frac{f_Y(y)}{\int f_Y(y)\ dy_{-i}};$$
$$\propto \exp\left(-\frac{1}{2}y^T Q y\right) = \exp\left(-\frac{1}{2}\sum_{(k,l)\in\mathcal{V}^2} y_k q_{kl} y_l\right).$$

En ne conservant que les termes dépendant de y_i et comme $q_{ik} = 0$ si $k \notin \delta_i$, on obtient :

$$\begin{aligned} f_{[Y_i|Y_{-i}=y_{-i}]}(y_i) &\propto \exp\left(-\frac{1}{2}y_i q_{ii} y_i - y_i \sum_{k\in\mathcal{V}\setminus\{i\}} q_{ik} y_k\right); \\ &= \exp\left(-\frac{1}{2}y_i q_{ii} y_i - y_i \sum_{j\in\delta_i} q_{ij} y_j\right); \\ &= \exp\left(-\frac{q_{ii}}{2}\left(y_i^2 + 2y_i \frac{1}{q_{ii}}\sum_{j\in\delta_i} q_{ij} y_j\right)\right); \\ &= \exp\left\{-\frac{q_{ii}}{2}\left(y_i^2 + 2y_i c\right)\right\}; \end{aligned}$$

où $c = \frac{1}{q_{ii}}\sum_{j\in\delta_i} q_{ij} y_j$. Ansi,

$$\begin{aligned} f_{[Y_i|Y_{-i}=y_{-i}]}(y_i) &= \exp\left\{-\frac{q_{ii}}{2}(y_i + c)^2 + \frac{q_{ii}}{2}c^2\right\}; \\ &\propto \exp\left\{-\frac{q_{ii}}{2}\left(y_i + \frac{1}{q_{ii}}\sum_{j\in\delta_i} q_{ij} y_j\right)^2\right\}. \end{aligned}$$

On reconnaît la forme fonctionnelle de la loi normale. Donc,

$$f_{[Y_i|Y_{-i}=y_{-i}]}(y_i) = \mathcal{N}\left(y_i \middle| -\frac{1}{q_{ii}}\sum_{j\in\delta_i} q_{ij} y_j, \frac{1}{q_{ii}}\right).$$

En transformant pour X, on trouve que

$$f_{[X_i|X_{-i}=x_{-i}]} = \mathcal{N}\left(x_i \middle| \mu_i - \frac{1}{q_{ii}}\sum_{j\in\delta_i} q_{ij}(x_j - \mu_j), \frac{1}{q_{ii}}\right);$$

ce qui démontre (i) et (ii). Pour démontrer le point (iii), on effectue à nouveau la transformation $Y = X - \mu$. Puisque la loi conjointe Y est une normale multidimensionnelle, la loi conditionnelle du couple (Y_i, Y_j) sera une loi normale bidimensionnelle. Il suffit de trouver la

matrice de covariance appropriée :

$$f_{[(Y_i,Y_j)|Y_{-i,-j}=y_{-i,-j}]}(y_i,y_j) \propto \exp\left\{-\frac{1}{2}y^TQy\right\}$$
$$\propto \exp\left\{-\frac{1}{2}(y_i\ ,\ y_j)\begin{pmatrix} q_{ii} & q_{ij} \\ q_{ji} & q_{jj}\end{pmatrix}\begin{pmatrix} y_i \\ y_j\end{pmatrix} + \begin{array}{c}\text{termes linéaires} \\ \text{en } y_i \text{ et } y_j\end{array}\right\}$$

On a que la loi normale bivariée de matrice de covariance Σ peut s'écrire de la façon suivante :

$$f_{(Y_i,Y_j)}(y_i,y_j) \propto \exp\left\{-\frac{1}{2}(y_i\ ,\ y_j)\begin{pmatrix} \varsigma_{ii} & \varsigma_{ij} \\ \varsigma_{ji} & \varsigma_{jj}\end{pmatrix}^{-1}\begin{pmatrix} y_i \\ y_j\end{pmatrix} + \begin{array}{c}\text{termes linéaires} \\ \text{en } y_i \text{ et } y_j\end{array}\right\}$$

Alors, on doit avoir que

$$\begin{pmatrix} q_{ii} & q_{ij} \\ q_{ji} & q_{jj}\end{pmatrix}^{-1} = \begin{pmatrix} \varsigma_{ii} & \varsigma_{ij} \\ \varsigma_{ji} & \varsigma_{jj}\end{pmatrix}.$$

Ceci montre que la distribution conditionnelle de (Y_i, Y_j) sachant $Y_{-i,-j} = y_{-i,-j}$ est une loi normale bivariée avec matrice de covariance

$$\Sigma = \begin{pmatrix} q_{ii} & q_{ij} \\ q_{ji} & q_{jj}\end{pmatrix}^{-1} = \begin{pmatrix} q_{jj}/\Delta & -q_{ij}/\Delta \\ -q_{ji}/\Delta & q_{ii}/\Delta\end{pmatrix}; \tag{1.52}$$

où $\Delta = q_{ii}q_{jj} - q_{ij}^2$. On conclut que

$$\mathbb{C}or[Y_i, Y_j|Y_{-i,-j} = y_{-i.-j}] = \frac{-q_{ij}}{\sqrt{q_{ii}q_{jj}}}. \tag{1.53}$$

Par linéarité, on a également que

$$\mathbb{C}or[X_i, X_j|X_{-i,-j} = x_{-i.-j}] = \frac{-q_{ij}}{\sqrt{q_{ii}q_{jj}}}. \tag{1.54}$$

□

De cette dernière proposition, il est facile de retrouver l'égalité suivante :

$$f_{X_i|X_{-i}=x_{-i}}(x_i) = f_{X_i|X_{\delta_i}=x_{\delta_i}}(x_i); \tag{1.55}$$

qui constitue la propriété de Markov appliquée aux graphes.

En pratique, l'application des champs de Markov gaussiens est contrainte par la limite de corrélation marginale que peut induire ce champ aléatoire. Puisque la matrice de précision Q doit être symétrique et définie positive, la corrélation marginale d'un couple de variables aléatoires s'en retrouve limitée. Pour une grille régulière infinie, Besag et Kooperberg (1995) ont montré que la corrélation marginale maximale d'un couple de variables aléatoires était d'environ 0,75. Dans la situation où une corrélation marginale supérieure est requise, l'application des champs de Markov gaussiens intrinsèques peut s'avérer astucieuse, ce qui correspond à l'objet de la prochaine section.

Remarque :

Pour les précipitations extrêmes simulées par le Modèle régional canadien du climat, un champ de Markov gaussien n'était pas en mesure de modéliser toute la dépendance spatiale des paramètres marginaux du modèle de valeurs extrêmes. Autrement dit, les paramètres marginaux possédaient une corrélation spatiale supérieure à ce que le champ de Markov pouvait modéliser.

1.4.3 Champs de Markov gaussiens intrinsèques

Les champs de Markov gaussiens intrinsèques s'apparentent aux champs de Markov gaussiens à la différence que leur matrice de précision Q n'est pas de plein rang (Rue et Held, 2005). Cette matrice étant singulière, le déterminant est nul, invalidant ainsi la densité conjointe définie à l'équation (1.50). Il est néanmoins possible de définir un champ aléatoire dit *impropre* pour lequel la densité conjointe est non-normalisée.

Définition 1.11. *Soit le graphe simple $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Soit $n = \text{card}(\mathcal{V})$, le nombre de sommets du graphe et soit k, un entier tel que $1 \leq k \leq n$. Un* champ de Markov gaussien impropre de rang $(n-k)$ *défini sur le graphe $\mathcal{G}$ est un vecteur aléatoire $X = (X_i : i \in \mathcal{V})^T$ pour lequel la densité conjointe est impropre et admet la repésentation suivante :*

$$f_X(x) \propto |Q|^{*1/2} \exp\left\{-\frac{1}{2}(x-\mu)^T Q (x-\mu)\right\}. \tag{1.56}$$

Ici, $|\ |^*$ dénote le déterminant généralisé, c'est-à-dire le produit des valeurs propres non nulles et Q est une matrice symétrique semi-définie positive de rang $n-k$ pour laquelle la condition suivante est respectée :

$$q_{ij} \neq 0 \text{ si et seulement si ou bien } i = j, \text{ ou bien } i \in \delta_i.$$

Le fait que la densité conjointe soit impropre n'est généralement pas problématique dans la mesure où les champs de Markov gaussiens impropres sont utilisés comme loi *a priori* dans un modèle hiérarchique bayésien. Sous le paradigme bayésien, la loi *a posteriori* peut être valide même si la loi *a priori* utilisée est impropre. Dans le contexte où la fonction de vraisemblance est $f_{[Z|\boldsymbol{\theta}=\theta]}(z)$ et où la loi *a priori* admet une densité impropre $f_{\boldsymbol{\theta}}(\theta)$, la loi *a posteriori* admettra une densité propre si la condition suivante est satisfaite :

$$\int_\theta f_{[Z|\boldsymbol{\theta}=\theta]}(z) f_{\boldsymbol{\theta}}(\theta)\ d\theta < \infty. \tag{1.57}$$

Généralement, une classe particulière de champs de Markov impropre est considérée, la classe des champs de Markov intrinsèques.

Définition 1.12. *Un* champ de Markov gaussien intrinsèque d'ordre k *défini sur le graphe simple $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ est un champ de Markov impropre de rang $(n-k)$ pour lequel la densité impropre est invariante sous l'addition d'un polynôme de degré $(k-1)$.*

Par exemple, pour un graphe constitué d'une ligne, la densité d'un champ de Markov intrinsèque de rang $(n-k)$ défini sur ce graphe est invariante à l'addition du polynôme d'ordre $(k-1)$:

$$f_X(x) = f_X\left(x + \sum_{i\in\mathcal{V}} a_0 + a_1 x_i + \ldots + a_{k-1}x_i^{k-1}\right); \tag{1.58}$$

pour $a_0, a_1, \ldots, a_{k-1} \in \mathbb{R}$. La notation devient plus compliquée pour les graphes simples généraux. En effet, il est alors nécessaire de considérer les coefficients des produits mixtes $x_i^{k_1}x_j^{k_2}$ pour tout k_1 et k_2 tels que $k_1 + k_2 = k-1$ et pour tout i et j dans $\mathcal{V}$.

Remarque :

> *En pratique, le rang de la matrice Q est fixé par le statisticien en fonction du lissage spatial qu'il souhaite obtenir. Il construit ensuite cette matrice en fonction du graphe.*

Champs de Markov gaussiens intrinsèques d'ordre 1

Soit le graphe simple $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Rappelons que $n_i = \text{card}(\delta_i)$ dénote le nombre de voisins du sommet $i \in \mathcal{V}$. Soit κ un paramètre positif. Un champ de Markov intrinsèque d'ordre 1 peut être construit sur le graphe $\mathcal{G}$ avec la matrice de précision $Q = (q_{ij} : i \in \mathcal{V}, j \in \mathcal{V})$ définie de la façon suivante :

$$q_{ij} = \begin{cases} \kappa n_i & \text{si } j = i; \\ -\kappa & \text{si } j \in \delta_i; \\ 0 & \text{sinon.} \end{cases} \tag{1.59}$$

On note que $Q\underline{1} = \underline{0}$, où $\underline{1}$ et $\underline{0}$ sont respectivement des vecteurs colonnes de 1 et de 0. Autrement dit que la somme de chaque ligne de Q est nulle. On note également que Q est de rang $(n-1)$. La densité conjointe du champ admet alors la représentation :

$$f_X(x) \propto \kappa^{(n-1)/2} \exp\left(-\frac{1}{2}x^T Q x\right). \tag{1.60}$$

Bien que la densité conjointe soit impropre, les lois conditionnelles sont bien définies comme démontré dans la proposition suivante.

Proposition 1.6 (Besag et Kooperberg, 1995)**.** *Soit $X = (X_i : i \in \mathcal{V})^T$ le champ de Markov gaussien intrinsèque d'ordre 1 défini sur le graphe simple $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ ayant pour paramètres $\mu = 0$ et la matrice Q telle que définie à l'équation (1.59). Alors on a*

$$f_{[X_i|X_{-i}=x_{-i}]}(x_i) = \mathcal{N}\left(x_i \middle| \frac{1}{n_i}\sum_{j\in\delta_i} x_j, \frac{1}{\kappa n_i}\right). \tag{1.61}$$

Démonstration.
Sans perte de généralité, considérons le cas $i = 1$. On a alors que

$$
\begin{aligned}
&f_{[X_1|X_{-1}=x_{-1}]}(x_1) \\
&\qquad \propto \frac{\exp\left(-\frac{1}{2}x^T Q x\right)}{f_{X_{-1}}(x_{-1})}; \\
&\qquad \propto \exp\left(-\frac{1}{2}x^T Q x\right); \\
&\qquad = \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} x_i q_{ij} x_j\right); \\
&\qquad = \exp\left\{-\frac{1}{2}\left(x_1 q_{11} x_1 + \sum_{j=2}^{n} x_1 q_{1j} x_j + \sum_{i=2}^{n} x_i q_{i1} x_1 + \sum_{i=2}^{n}\sum_{j=2}^{n} x_i q_{ij} x_j\right)\right\}; \\
&\qquad \propto \exp\left\{-\frac{1}{2}\left(x_1 q_{11} x_1 + \sum_{j=2}^{n} x_1 q_{1j} x_j + \sum_{i=2}^{n} x_i q_{i1} x_1\right)\right\}; \\
&\qquad = \exp\left\{-\frac{\kappa}{2}\left(n_1 x_1^2 - 2x_1 \sum_{j\in\delta_1} x_j\right)\right\}; \\
&\qquad \propto \exp\left\{-\frac{\kappa n_1}{2}\left(x_1 - \frac{1}{n_1}\sum_{j\in\delta_1} x_j\right)^2\right\}.
\end{aligned}
$$

De la dernière équation, on reconnaît la forme fonctionnelle de la loi normale de moyenne $\frac{1}{n_1}\sum_{j\in\delta_1} x_j$ et de variance $(\kappa n_1)^{-1}$. Cette loi conditionnelle est une densité de probabilité puisque

$$
\int_{-\infty}^{\infty} \exp\left\{-\frac{\kappa n_1}{2}\left(x_1 - \frac{1}{n_1}\sum_{j\in\delta_1} x_j\right)^2\right\} dx_1 = \sqrt{\frac{2\pi}{\kappa n_1}} < \infty. \tag{1.62}
$$

□

Pour un champ de Markov gaussien intrinsèque d'ordre 1, l'espérance conditionnelle au sommet i sachant l'ensemble des valeurs aux autres sommets est la moyenne de ses voisins. Une telle expression était impossible pour les champs de Markov gaussiens en raison des contraintes imposées à Q. Pour le champ intrinsèque, la variance conditionnelle dépend de la précision du champ κ et du nombre de voisins considérés. Plus le nombre de voisins considérés est grand, plus la variance conditionnelle de X_i sera petite. L'équation (1.61) permet également de donner une interprétation au paramètre κ. Plus κ est grand, plus la variance conditionnelle sera petite et, par conséquent, plus le champ induit sera lisse. Le champ intrinsèque procure une paramétrisation conditionnelle intuitive et facilement interprétable, ce qui n'était pas le cas pour les champs de Markov gaussiens.

De façon marginale, les variables aléatoires qui composent un champ de Markov intrinsèque gaussien d'ordre 1 possèdent une moyenne indéfinie et une variance infinie. Or, Besag et Kooperberg (1995) ont montré que tout contraste $c^T X$ de variables aléatoires composant ce champ possède une loi de probabilité bien définie, où $c \neq \underline{0}$ et $c^T \underline{1} = 0$. En particulier, la différence entre les variables aléatoires de deux sommets voisins est distribuée selon la loi suivante :

$$f_{X_i - X_j}(y) = \mathcal{N}(y|0, \kappa^{-1}), \text{ pour } j \in \delta_i \text{ et } i \in \mathcal{V}. \tag{1.63}$$

Cette équation permet de donner une interprétation alternative au paramètre κ. Plus κ est grand, plus la différence marginale entre les valeurs de deux sommets voisins sera petite. Donc, plus κ est grand, plus le champ induit sera lisse dans l'espace.

Pour ce champ de Markov gaussien intrinsèque d'ordre 1, on a que

$$f_X(x) = f_X(x + c\underline{1}), \quad \text{pour } c \in \mathbb{R}, \tag{1.64}$$

puisque $Q\underline{1} = \underline{0}$. La densité impropre du champ est invariante à l'addition de n'importe quelle constante ajoutée à l'ensemble des variables aléatoires. Utilisée comme loi *a priori*, cette densité impropre est non informative concernant une valeur centrale générale pour l'ensemble des variables aléatoires. Elle est cependant informative pour les contrastes $X_i - X_j$, tel que décrit par l'équation (1.63). Utiliser un champ de Markov intrinsèque d'ordre 1 comme loi *a priori* revient à modéliser la connaissance *a priori* sur la texture locale du champ, autrement dit sur $\{X_i - X_j : \{i, j\} \in \mathcal{E}\}$ plutôt que sur $\{X_i - \mu : i \in \mathcal{V}\}$ pour un certain niveau $\mu \in \mathbb{R}$. D'ailleurs, la densité conjointe peut s'exprimer en fonction de ces contrastes grâce à l'identité suivante :

Lemme 1.1.

$$x^T Q x = \sum_i q_{i+} x_i^2 - \sum_{i<j} q_{ij}(x_i - x_j)^2; \tag{1.65}$$

où q_{i+} *dénote* $\sum_j q_{ij}$.

Démonstration.

$$\begin{aligned}
x^T Q x &= \sum_i \sum_j x_i q_{ij} x_j; \\
&= \sum_i q_{ii} x_i^2 + 2 \sum_{i<j} x_i q_{ij} x_j; \\
&= \sum_i \sum_j q_{ij} x_i^2 - 2 \sum_{i<j} x_i q_{ij} x_i + 2 \sum_{i<j} x_i q_{ij} x_j; \\
&= \sum_i q_{i+} x_i^2 - \sum_{i<j} q_{ij} x_i^2 - \sum_{i<j} q_{ij} x_j^2 + 2 \sum_{i<j} x_i q_{ij} x_j; \\
&= \sum_i q_{i+} x_i^2 - \sum_{i<j} q_{ij}(x_i^2 - 2 x_i x_j + x_j^2); \\
&= \sum_i q_{i+} x_i^2 - \sum_{i<j} q_{ij}(x_i - x_j)^2.
\end{aligned}$$

□

Pour un champ de Markov gaussien intrinsèque d'ordre 1, $q_{i+} = 0$ et $q_{ij} = -\kappa$ pour $\{i, j\} \in \mathcal{E}$. Donc, la densité conjointe exprimée à l'équation (1.60) peut aussi s'écrire de la façon suivante :

$$f_X(x) \propto \kappa^{(n-1)/2} \exp\left(\frac{\kappa}{2} \sum_{\{i,j\}\in\mathcal{E}} (x_i - x_j)^2\right). \tag{1.66}$$

Remarque :

Il existe des formulations pour des champs de Markov gaussiens intrinsèques d'ordres supérieurs. Toutefois, celles-ci conviennent pour des grilles régulières infinies, c'est-à-dire sans effet de bord ni données manquantes. Adapter ces formulations pour la portion terrestre du domaine spatial du Modèle régional canadien du climat serait très compliqué. Par conséquent, la présentation de cette section et les analyses statistiques de cette thèse se limitent aux champs intrinsèques d'ordre 1.

Références

Banerjee, S., A. E. Gelfand et B. P. Carlin. 2004, *Hierarchical modeling and analysis for spatial data*, 2e éd., Chapman & Hall, Boca Raton, USA, 562 pages.

Besag, J. et C. Kooperberg. 1995, «On conditional and intrinsic autoregressions», *Biometrika*, vol. 82, no 4, p. 733–746.

Coles, S. 2001, *An introduction to statistical modeling of extreme values*, Springer, London, United Kingdom, 209 pages.

Cressie, N. A. 2003, *Statistics for spatial data*, Wiley, New York, USA, 900 pages.

Davison, A. C. et R. L. Smith. 1990, «Models for exceedances over high thresholds (with Discussion)», *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 52, no 3, p. 393–442.

Fisher, R. A. et L. H. C. Tippett. 1928, «Limiting forms of the frequency distribution of the largest or smallest member of a sample», dans *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 24, Cambridge University Press, p. 180.

Gnedenko, B. 1943, «Sur la distribution limite du terme maximum d'une série aléatoire», *Annals of mathematics*, vol. 44, no 3, p. 423–453.

Jenkinson, A. F. 1955, «The frequency distribution of the annual maximum (or minimum) values of meteorological elements», *Quarterly Journal of the Royal Meteorological Society*, vol. 81, no 348, p. 158–171.

Leadbetter, M. R., G. Lindgren et H. Rootzén. 1983, *Extremes and related properties of random sequences and processes*, Springer, New York, USA, 336 pages.

Pickands III, J. 1975, «Statistical inference using extreme order statistics», *The Annals of Statistics*, vol. 3, n° 1, p. 119–131.

Rosenblatt, M. 1956, «A central limit theorem and a strong mixing condition.», *Proceedings of the National Academy of Sciences of the United States of America*, vol. 42, n° 1, p. 43–47.

Rue, H. et L. Held. 2005, *Gaussian Markov random fields : theory and applications*, CRC Press, Boca Raton, USA, 259 pages.

Chapitre 2

Canadian RCM projected transient changes to precipitation occurrence, intensity and return level over North America

Jonathan Jalbert, Anne-Catherine Favre, Claude Bélisle,

Jean-François Angers et Dominique Paquin.

Cet article a été publié dans *Journal of Climate*. La référence complète de celui-ci est la suivante :

Jalbert, J., A.-C. Favre, C. Bélisle, J.-F. Angers, et D. Paquin (2015). Canadian RCM projected transient changes to precipitation occurrence, intensity and return level over North America. *Journal of Climate*, vol. 28, n°17, p. 6920–6937.

Résumé

L'augmentation de la concentration des gaz à effet de serre dans l'atmosphère influence les processus générateurs des précipitations, ce qui occasionne des changements dans l'occurrence et dans l'intensité de celles-ci. Il devient primordial d'estimer ces changements pour en prévoir les conséquences sur l'environnement. Plusieurs études ont été publiées concernant l'évolution des précipitations simulées par un modèle numérique de climat. Ces études sont majoritairement basées sur la comparaison de deux sous-intervalles, généralement d'une durée

de trente années, extraits de la période de simulation. Cette approche simplifie la modélisation statistique puisqu'il est communément supposé que la non-stationnarité est négligeable pour ces sous-intervalles de courtes durées. L'analyse des précipitations devrait toutefois être plus précise si la période complète de simulation était considérée. Par conséquent, le but de cet article est de développer un modèle statistique continu pour les précipitations journalières simulées par le Modèle régional canadien du climat. Le modèle statistique permet de décrire l'évolution continue de l'occurrence et de l'intensité des précipitations journalières simulées pour l'Amérique du Nord. Les résultats montrent que l'occurrence et l'intensité des précipitations augmentent au fil du temps au Canada et diminuent au Mexique. Pour les États-Unis, les résultats démontrent une augmentation en hiver et une diminution en été. Le modèle statistique est utilisé pour étudier l'évolution des précipitations extrêmes dans une méthodologie d'analyse fréquentielle non stationnaire.

Abstract

As a consequence of the increase in atmospheric greenhouse gas concentrations, potential changes in both precipitation occurence and intensity may lead to several consequences for earth's environment. It is therefore relevant to estimate these changes in order to anticipate their consequences. Many studies have been published on precipitation changes based on climate simulations. These studies are almost always based on time slices; precipitation changes are estimated in comparing two 30-years windows. To this extent, it is commonly assumed that non-stationary processes are not significant on such a 30-years slice. Thus, it frees the investigator to statistically model non-stationary processes. However using transient runs instead of time slices surely leads to more accurate analysis since more data are taken into account. Therefore, the aim of the present study was to develop a transient probabilistic model for describing simulated daily precipitation from the Canadian Regional Climate Model (`CRCM`) in order to investigate precipitation evolution over North America. Precipitation changes to both occurrence and intensity are then assessed from a continuous time period. Extreme values are also investigated with the transient run; a methodology using the models for precipitation occurrence and intensity was developed for achieving non-stationary frequency analysis. Our results show an increase in both precipitation occurrence and intensity for most part of Canada while a decrease is expected over Mexico. For the continental U.S., a decrease in both occurrence and intensity is expected in summer while an increase is expected in winter.

2.1 Introduction

Precipitations play a central role for the ecosystems, the human infrastructures, the freshwater supplies, etc. Many infrastructures were built to resist up to a particular water level. Accord-

ing to climate change, the precipitation generating process might evolve in time, leading to changes in its occurrence and intensity. The potential impact on ecosystems and infrastructures is a major concern. To address this issue, temporal precipitation evolution according to climate change must be described. Even though the general rainfall generating mechanisms are fairly known, the evolution of rainfall in a non-stationary climate is intractable. Indeed, the non-linear and chaotic nature of the equations describing rainfall mechanisms prevent simple predictions for precipitation. Therefore, numerical simulations from climate models may be the best way to provide quantitative prediction of precipitations in a non-stationary climate (IPCC, 2012). Regional Climate Models (RCMs) are possibly better tools for studying climate change impact on precipitation than Global Climate Models (GCMs). The resolution of RCMs (from 50 down to 10 km) enables a good representation of some important features in the formation of precipitation (mainly orography, surface and hydrology). The spatial-scale characteristics of precipitation is reasonably represented in RCMs on a length scale of roughly 100 km (Rasmussen et al., 2012).

Most climate studies are based on the comparison of two standard time windows. In IPCC AR5 (Christensen et al., 2013), GCMs projected climate changes are assessed from two 20-years time slices: $[1986, 2005]$ stands for the present period while $[2081, 2100]$ stands for the future climate period. For RCMs projected changes, it is often two 30-years time slices that are investigated: $[1961, 1990]$ and $[2071, 2100]$. This is the case for the PRUDENCE experiment [1] (Christensen and Christensen, 2007), the ENSEMBLE project [2] (van der Linden and Mitchell, 2009) over Europe and the NARCCAP program [3] (Mearns et al., 2009) over North America. This approach was used, for a certain time, because regional climate models were too expensive to run transient simulation. Even though transient runs are now available, time-sliced analyzes are still largely performed, notably to avoid the modeling of non-stationary processes. Indeed, it is generally assumed that a 30-years window is long enough for climate to dominate natural variability and short enough for non-stationary processes to be negligible. But, from a statistical point of view, so much information is lost in the omitted intermediate years, an unfortunate scenario given the high cost of such simulations. The intermediate years should be used in order to describe the non-stationary process on the whole period $[1961, 2100]$. It may help distinguishing climate change from natural variability (Kay and Jones, 2012). To our knowledge, only Goderniaux et al. (2011) have modeled daily RCM precipitation on the continuous time period $[2010, 2085]$ on a probabilistic basis. Their probabilistic model was then used to supply a hydrological model with several proper precipitation scenarios.

In studies of extreme values, where only few observations per year are available, the need to consider transient simulations in order to increase the number of observations is crucial. For this reason, many recent papers used transient run to adjust a time-varying Generalized

1. http://prudence.dmi.dk/main.html
2. http://www.ensembles-eu.org
3. http://www.narccap.ucar.edu

Extreme Value (GEV) distribution to describe series of annual maxima (see for example Kharin and Zwiers, 2005; Hanel et al., 2009; Kay and Jones, 2012) or a time-varying Peaks Over Threshold (POT) model to describe series of threshold exceedances (see for example Sugahara et al., 2009; Mailhot et al., 2013). Such percentile estimation corresponding to a given return level is still challenging, even if the entire time period is used, because the non-stationary characteristics have to be assessed with only few annual observations. For example, three non-stationary GEV parameters have to be estimated with only 140 annual maxima, those from the period $[1961, 2100]$.

On the other hand, regional frequency analysis stands as an alternative to assess extreme precipitation using 30-years time slices. This approach compensates the lack of observational records by incorporating regional information given a set of homogeneous grid points. It has been applied for precipitation extremes (see for instance Fowler and Ekström, 2009; Mladjic et al., 2011; Mailhot et al., 2012; Roth et al., 2014). However finding such homogeneous set of points is the most challenging part of this approach. It is not yet completely solved for North America.

This paper focuses on the assessment of transient changes to precipitation occurrence, intensity and 100-years return level over North America for the continuous time period $[1961, 2100]$, using a climate simulation from the Canadian RCM (`CRCM`, Caya and Laprise, 1999). Several papers have been published on `CRCM` precipitation changes between two sub-periods (see for example Mladjic et al., 2011; Mailhot et al., 2012; Monette et al., 2012; Paquin et al., 2014), but none on transient change. At this stage, only one simulation is investigated. The proposed methodology aims to extract more reliable information from a single simulation. Eventually, the methodology should be extended for an ensemble of simulations in order to give better estimates of climate change. The first step of the methodology is to fit several parametric models to CRCM simulated precipitation occurrence and intensity. Thereafter, only the best model, according to the BIC criterion, will be kept. Finally, an approach to estimate non-stationary return levels will be described.

The paper is organized as follows. Section 2.2 is devoted to the description of the data and the notation used throughout the paper. The methodology is developed in Section 2.3. Section 2.4 describes the results on North America and finally the discussion and the conclusion are exposed in Section 2.5.

2.2 Data from Canadian Regional Climate Model

2.2.1 `CRCM` Simulations Description

The regional simulation under investigation was performed by the Canadian Regional Climate Model `CRCM` 4.2.3, based on `CRCM` 4.2 (Music and Caya, 2007; de Elìa and Côté, 2010). It

was run over the North American domain on a polar stereographic grid with a horizontal resolution of 45 km true at 60°N (totaling 200 by 192 grid points), using 29 vertical levels, a time step of 15 minutes and an archival period of 6 hours. The free domain, exclusive of sponge zone, is composed of 182×174 grid points. The regional model was driven by the outputs from the third version of the Canadian Global Coupled Model (`CGCM3` Scinocca et al., 2008) which performed a transient simulation following IPCC "observed 20th century" scenario for years 1961-2000 and "SRES A2" (Nakicenovic et al., 2000) scenario for years 2001-2100. The simulation is driven by the 4^{th} member from an ensemble of five 20^{th} century and SRES A2 simulations.

`CRCM` convective precipitation is parametrized following Bechtold-Kain-Fritsch (Bechtold et al., 2001) while large-scale (stratiform) precipitation is parametrized as a simple supersaturation-based condensation scheme. `CRCM`'s sensitivity of precipitation simulations to the choices of parameterization scheme can be found in Paquin et al. (2002), Jiao and Caya (2006) and Christensen et al. (2008). `CRCM`, as many other RCMs, overestimates occurrence of small precipitation amount (Christensen et al., 2008). In order to avoid a large artificial occurrence probability, we assumed that under a threshold of 1 mm, no precipitation occurs. The 1 mm threshold follows the recommandation of the European Statistical and Regional Dynamical Downscaling of Extremes for European Seasons (STARDEX; http://www.cru.uea.ac.uk/projects/stardex). We also reduced the corresponding amount of 1 mm when it occurs. Removing 1 mm frees us from the needs to work with truncated statistical distribution for modeling daily intensity, otherwise distribution supports would be $[1, \infty)$. If small to moderate precipitation amounts are important for an investigator, it would be possible to add 1 mm when post-processing the results. Nevertheless, such a reduction does not influence the results for large precipitation events.

`CRCM` simulation domain covers North America, as shown in Figure 2.1. We restrict the study to the 12,570 grid points that represent land points. The sea grid points are not investigated.

2.2.2 Notations

In this paper, the precipitation description is performed by seasons. We define the winter season as December, January and February, the spring season as March, April and May, the summer season as June, July and August and the autumn season as September, October and November.

Let us introduce the notation used throughout the paper. For any given grid point and season, let Y_{kl} be the amount of daily precipitations in mm (after the bias reduction of 1 mm) on the

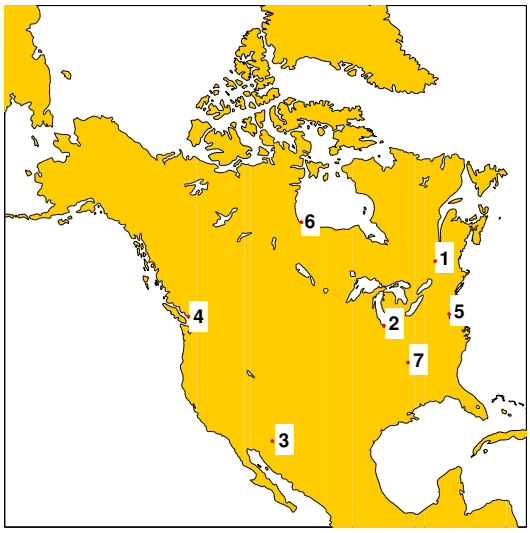


Figure 2.1 – CRCM simulation domain. 1-Montreal. 2-Chicago. 3-Tucson. 4-Vancouver. 5-Washington D.C. 6-Churchill. 7-Nashville.

l^{th} day of year index k, where

$$k \in \{0, 1, \ldots, 139\};$$
$$l \in \{1, 2, \ldots, L\}.$$

The index k indicates the number of years elapsed since 1961. Thus, $k = 0$ stands for the year 1961 while $k = 139$ indicates the last year (2100). The index l stands for days in a particular season. For instance, assume that we are interested in the summer season. In this case, index l goes from 1 to 92, $l = 1$ stands for June 1^{st} and $l = 92$ stands for August 31^{st}.

In order to describe precipitation occurrence, we define a dichotomous random variable X_{kl} that indicates if precipitations occur on the l^{th} day of year k

$$X_{kl} = \begin{cases} 1 & \text{if } Y_{kl} > 0; \\ 0 & \text{if } Y_{kl} = 0. \end{cases} \tag{2.1}$$

Thus, the number of days during which precipitation occurred in a given season and a given year k is

$$N_k = \sum_{l=1}^{L} X_{kl}. \tag{2.2}$$

2.3 Methodology

Cooley and Sain (2010) claimed that "climate is what you expect and weather is what you actually get". In this paper, we tried to describe climate with one of its possible realizations: a RCM simulation. To infer from weather to climate, parametric statistical distributions

are fitted to precipitation data. Moreover, statistical models are needed for establishing the significance of the non-stationarity. Without such models, results can only be descriptive. Also, the model's parameters are the best way to extract and summarize the information and to quantify the sampling uncertainty.

Precipitation are subject to seasonal fluctuations that are much larger than climate variations. Therefore, it is a better strategy to partition the calendar year into seasons and then study each part separately. Within a season, we assumed that both daily precipitation occurrence and intensity are independent and identically distributed. This assumption seems more suitable for convective than stratiform precipitation since the later might be auto-correlated over days. Nevertheless, we made this assumption first to simplify the parametric models. Actually, this approach frees us from the needs to statistically model seasonal fluctuation, which could be very difficult and lead to large uncertainties.

In this paper, we developed several statistical models for describing daily precipitation occurrence, intensity and extremes. At every grid point, all these models have been estimated but only the best model, according to the Bayesian Information Criterion (BIC), was kept. The next sections are devoted to the description of the considered statistical distribution for modeling precipitation occurrence, intensity and extremes.

2.3.1 Parametric models for occurrence

For any given day l of year k, either precipitation occurs or it does not. The natural probability model for this dichotomous variable is the Bernoulli distribution. Let p_k be the probability that precipitation occurs on a given day of year k, then

$$\begin{cases} \mathbb{P}[X_{kl} = 1] = p_k; \\ \mathbb{P}[X_{kl} = 0] = 1 - p_k. \end{cases} \tag{2.3}$$

Assume that this probability does not evolve over time, i.e. $p_k = p, \forall k \in \{0, \ldots, 139\}$. This leads to the first parametric model for precipitation occurrence,

$$\mathcal{P}_1 : N_k \sim \mathcal{B}\text{inomial}(L, p). \tag{2.4}$$

Suppose now that this probability could evolve with time, this gives the second parametric model:

$$\mathcal{P}_2 : N_k \sim \mathcal{B}\text{inomial}(L, p_k); \tag{2.5}$$

where

$$p_k = \frac{\exp(\zeta_0 + \zeta_1 k)}{1 + \exp(\zeta_0 + \zeta_1 k)} \tag{2.6}$$

is the inverse logit link function widely used in logistic regression. Note that we always have $0 < p_k < 1$, with p_k close to 1 if $\zeta_0 + \zeta_1 k$ is a large positive number and close to 0 if $\zeta_0 + \zeta_1 k$ is a large negative number. Note that the first model is stationary while the second is non-stationary over years.

2.3.2 Parametric models for intensity

For precipitation intensity, we only modeled strictly positive precipitation: the zero values are ignored. For a given year index k and conditional on days where precipitation occured, eight different parametric distributions were considered to model $Y_{kl}|X_{kl} = 1$:

$$
\begin{aligned}
&\mathcal{M}_1 : Y_{kl}|X_{kl} = 1 \sim \mathcal{E}\text{xp}(\theta_1);\\
&\mathcal{M}_2 : Y_{kl}|X_{kl} = 1 \sim \mathcal{G}\text{amma}(\lambda, \theta_1);\\
&\mathcal{M}_3 : Y_{kl}|X_{kl} = 1 \sim (1-w)\ \mathcal{E}\text{xp}(\theta_0) + w\ \mathcal{E}\text{xp}(\theta_1);\\
&\mathcal{M}_4 : Y_{kl}|X_{kl} = 1 \sim \mathcal{E}\text{xp}\left\{\theta_1(k)\right\};\\
&\mathcal{M}_5 : Y_{kl}|X_{kl} = 1 \sim \mathcal{G}\text{amma}\left\{\lambda, \theta_1(k)\right\};\\
&\mathcal{M}_6 : Y_{kl}|X_{kl} = 1 \sim (1-w)\ \mathcal{E}\text{xp}\left\{\theta_0(k)\right\} + w\ \mathcal{E}\text{xp}\left\{\theta_1(k)\right\};\\
&\mathcal{M}_7 : Y_{kl}|X_{kl} = 1 \sim (1-w(k))\ \ \mathcal{E}\text{xp}\left(\theta_0\right) + w(k)\ \mathcal{E}\text{xp}\left(\theta_1\right);\\
&\mathcal{M}_8 : Y_{kl}|X_{kl} = 1 \sim (1-w(k))\ \ \mathcal{E}\text{xp}\left\{\theta_0(k)\right\} + w(k)\ \mathcal{E}\text{xp}\left\{\theta_1(k)\right\};
\end{aligned}
\tag{2.7}
\tag{2.8}
$$

where

$$
\begin{aligned}
\theta_0(k) &= \alpha_0\alpha_1^k;\\
\theta_1(k) &= \beta_0\beta_1^k;\\
w(k) &= \frac{\exp\left(\gamma_0 + \gamma_1 k\right)}{1+\exp\left(\gamma_0 + \gamma_1 k\right)}.
\end{aligned}
\tag{2.9}
$$

and

$$
\begin{aligned}
&\theta_0 > 0,\ \theta_1 > 0,\ \lambda > 0, p \in (0,1),\\
&\alpha_0 > 0,\ \alpha_1 > 0,\ \beta_0 > 0,\ \beta_1 > 0,\\
&(\gamma_0, \gamma_1) \in \mathbb{R}^2.
\end{aligned}
\tag{2.10}
$$

Model $\mathcal{M}_3$, as models $\mathcal{M}_6$, $\mathcal{M}_7$ and $\mathcal{M}_8$, is a mixture model and more specifically a mixture of exponentials distributions. Such a model conditional on $X_{kl} = 1$ can be interpreted as follows: a particular realization y_{kl} of the random variable Y_{kl} has a probability $(1-w)$ of being a realization of the first mixture component $\mathcal{E}\text{xp}(\theta_0)$ and a probability w of being a realization of the second mixture component $\mathcal{E}\text{xp}(\theta_1)$. See Titterington et al. (1985) for more details.

The temporal functions in Equation (2.9) have been chosen in order to insure that $\theta_0(k)$ and $\theta_1(k)$ are strictly positive for any value of k. Thus, the different models based on exponential and gamma distributions are always definite, no matter the value of k. Other functions like $\theta_0(k) = \alpha_0 + \alpha_1 \times k$ may be tricky to use in order to have $\theta_0(k) > 0$ for $k \in (0, \infty)$ because parameters constraints would be $\alpha_0 > 0$ and $\alpha_1 \in \mathbb{R}$. An exponential transformation could be appropriate though.

Table 2.1 – Expectations and variances as a function of the year index k for the occurrence models.

Model	ν_k	τ_k^2
$\mathcal{P}_1$	Lp	$Lp(1-p)$
$\mathcal{P}_2$	$Lp(k)$	$Lp(k)\{1-p(k)\}$

Table 2.2 – Expectations and variances as a function of the year index k for the intensity models.

Model	ν_k	τ_k^2
$\mathcal{M}_1$	$\frac{1}{\theta_1}$	$\frac{1}{\theta_1^2}$
$\mathcal{M}_2$	$\frac{\lambda}{\theta_1}$	$\frac{\lambda}{\theta_1^2}$
$\mathcal{M}_3$	$\frac{1-w}{\theta_0} + \frac{w}{\theta_1}$	$\frac{2(1-w)}{\theta_0^2} + \frac{2w}{\theta_1^2} - \left(\frac{1-w}{\theta_0} + \frac{w}{\theta_1}\right)^2$
$\mathcal{M}_4$	$\frac{1}{\theta_1(k)}$	$\frac{1}{\theta_1^2(k)}$
$\mathcal{M}_5$	$\frac{\lambda}{\theta_1(k)}$	$\frac{\lambda}{\theta_1^2(k)}$
$\mathcal{M}_6$	$\frac{1-w}{\theta_0(k)} + \frac{w}{\theta_1(k)}$	$\frac{2(1-w)}{\theta_0^2(k)} + \frac{2w}{\theta_1^2(k)} - \left\{\frac{1-w}{\theta_0(k)} + \frac{w}{\theta_1(k)}\right\}^2$
$\mathcal{M}_7$	$\frac{1-w(k)}{\theta_0} + \frac{w(k)}{\theta_1}$	$\frac{2(1-w(k))}{\theta_0^2} + \frac{2w(k)}{\theta_1^2} - \left\{\frac{1-w(k)}{\theta_0} + \frac{w(k)}{\theta_1}\right\}^2$
$\mathcal{M}_8$	$\frac{1-w(k)}{\theta_0(k)} + \frac{w(k)}{\theta_1(k)}$	$\frac{2(1-w(k))}{\theta_0^2(k)} + \frac{2w(k)}{\theta_1^2(k)} - \left\{\frac{1-w(k)}{\theta_0(k)} + \frac{w(k)}{\theta_1(k)}\right\}^2$

Table 2.1 and Table 2.2 summarize the expectations and the variances as a function of the year index k respectively for the occurrence models and the intensity models. The expectations and variances of the occcurence models are $L \times p$ and $L \times p \times (1-p)$ for Model $\mathcal{P}_1$ and $L \times p_k$ and $L \times p_k \times (1-p_k)$ for Model $\mathcal{P}_2$.

2.3.3 Model estimation

Every model has been adjusted using maximum likelihood estimation. The maximum likelihood estimation of the parameters in models $\mathcal{P}_1$, $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_4$ and $\mathcal{M}_5$ is quite straightforward. For Model $\mathcal{P}_2$, we used the well-known algorithms for estimating parameters of a logistic regression. For the mixture models $\mathcal{M}_3$, $\mathcal{M}_6$, $\mathcal{M}_7$ and $\mathcal{M}_8$, we had to use the EM algorithm (Dempster et al., 1977) to maximize the likelihood function.

2.3.4 Statistical model selection

For every grid points, two models were fitted for precipitation occurrence and eight for precipitation intensity. Then, we used the Bayesian Information Criterion (BIC) to select the best model among those fit. The BIC is defined as

$$BIC = -2\log(\mathcal{L}) + \kappa \log(n); \tag{2.11}$$

where $\mathcal{L}$ is the likelihood of the parametric model, κ is the number of parameters of the model and n is the number of observations. The best model, for a given grid point, is the one with the smallest BIC. The BIC has a penalty term for the number of parameters which allows to select better model in the trade-off between ajustment and number of parameters.

2.3.5 Return level

Extreme Value Theory (EVT) is generally considered to be the best tool for estimating return levels from a time series (Coles, 2001). One of the central results of EVT tells us that, under appropriate conditions, the distribution of the maximum of a large number of random variables can be approximated by the Generalized Extreme Value (GEV) distribution. For our application, the seasonal maximum is extracted from a series of 90 observations. In general, it is unclear whether a sequence of 90 observations is sufficient to invoke the GEV distribution. However, in the context of precipitation maxima, a period of 90 days is usually assumed to be suitable (see for example Fowler and Ekström, 2009; Katz, 2013).

Seasonal precipitation series are subject to non-stationarity due to climate changes. The standard approach for addressing non-stationarity in maxima series is to model the GEV parameters as a function of time (e.g. Coles, 2001; Katz, 2013). Examples of applications can be found for instance in Katz (2010), Gilleland and Katz (2011) and Cheng et al. (2014). A common pragmatic approach is to consider the shape parameter invariant in time because a large uncertainty is associated to its estimation even in the stationary case. Let M_k^* be the seasonal maximum for the year k at a particular grid point. The usual approach consists in modeling M_k^* as follows:

$$M_k^* \approx GEV(\mu_k, \sigma_k, \xi); \tag{2.12}$$

for different link functions σ_k and μ_k. For example, one could model those as

$$\mu_k = \beta_1 + \beta_2 \cdot k, \tag{2.13}$$

$$\sigma_k = \exp\left(\beta_3 + \beta_4 \cdot k\right); \tag{2.14}$$

and estimate the parameter vector $[\xi, \beta_1, \beta_2, \beta_3, \beta_4]$ with the vector of maxima $(M_k^* : k \in \{0, \dots 139\})$. In most cases, these parameters have to be estimated with only a few maxima, leading to a large estimation variance.

To avoid estimating temporal trends with only a few maximum data, the information extracted from the daily series could be used instead. The idea is to model the entire precipitation distribution evolving over the whole period and then transform the data into a stationary series. This could be achieved with the intensity models described in Section 2.3.2. Let ν_k and τ_k^2 be respectively the annual conditional mean and conditional variance of the chosen

model for a certain grid point:

$$\nu_k = \mathbb{E}[Y_{kl}|X_{kl} = 1] \tag{2.15}$$

$$\tau_k^2 = \mathbb{V}ar[Y_{kl}|X_{kl} = 1]. \tag{2.16}$$

We defined Z_{kl}, the standardized amount of precipitation on day l of year k, as:

$$Z_{kl} = \frac{Y_{kl} - \nu_k}{\tau_k}. \tag{2.17}$$

Both the expectation and the standard deviation in the last equation can evolve with time depending on the model chosen for intensity. The standardized random variable $Z_{kl}|X_{kl} = 1$ has the following properties: $\mathbb{E}[Z_{kl}|X_{kl} = 1] = 0$ and $\mathbb{V}ar[Z_{kl}|X_{kl} = 1] = 1$ for all k and l. The standardized series is thus second-order stationary while its two first moments are time invariant. However, let us assume that the transformed series is strictly stationary. We will see further in the present section that this assumption is equivalent to considering a constant shape parameter. The seasonal maxima series of the transformed series is defined as

$$M_k = \max_{l \in L} Z_{kl}; \tag{2.18}$$

where L denotes the set of days in a given season. Since we have assumed that after transformation the corresponding daily precipitation is stationary, the series of maxima may be considered as independent and identically GEV distributed:

$$M_k \approx \mathcal{G}EV(\mu, \sigma, \xi), \text{ for } 0 \leq k \leq 139. \tag{2.19}$$

A classical frequency analysis can thus be performed on the transformed series. Every theoretical result and definition, such as the return period, are directly applicable to the transformed series. Once the GEV is fitted, the quantile $z_{[1-1/T]}$ corresponding to a return period of T-years can be estimated. The quantile $z_{[1-1/T]}$ is such that

$$\mathbb{P}[M_k > z_{(1-1/T)}] = 1/T, \text{ for } 0 \leq k \leq 139. \tag{2.20}$$

The corresponding return level in the original precipitation space is obtained with the following inverse transformation

$$R_k^T = z_{(1-1/T)} \times \tau_k + \nu_k. \tag{2.21}$$

Since $\boldsymbol{Y}_k$ could be non-stationary, a different return level is computed for each year in order to keep the exceedance probability constant in time. Katz et al. (2002) called these time evolving quantiles "effective return levels". In this paper, we define the return level R^T for a certain subset of years K as the maximum effective return level on this subset:

$$R^T = \max_{k \in K} R_k^T. \tag{2.22}$$

This definition of a T-years non-stationary return level ensures that the estimate in the original precipitation space is at least a T-years return level in all years of the period:

$$\mathbb{P}[M_k^* > R^T] \leq 1/T, \ \forall k \in K. \tag{2.23}$$

Despite the fact that there exist more appropriate alternatives to the definition of return period in a non-stationary context (see Rootzén and Katz, 2013), we chose the maximum over the period because we believe that is easier to understand and to interpret. However, for practical purposes, we recommend to use a more advanced approach.

Hereafter, the maximum distributions of the original series and the standardized series are compared. The usual approach consists in assuming that the seasonal maximum M_k^* from the original series $\boldsymbol{Y}_k$ for a year index k at a certain grid point is approximately GEV distributed:

$$M_k^* \approx \mathcal{G}EV(\mu_k, \sigma_k, \xi); \tag{2.24}$$

where ξ is assumed time invariant. Following this, the distribution of the standardized maximum M_k can be computed as:

$$\begin{aligned}
\mathbb{P}[M_k \leq z] &= \mathbb{P}\left[\frac{M_k^* - \nu_k}{\tau_k} \leq z\right] = \mathbb{P}\left[M_k^* \leq \tau_k \cdot z + \nu_k\right] \\
&= \exp\left\{-\left[1 + \xi\left(\frac{\tau_k \cdot z + \nu_k - \mu_k}{\sigma_k}\right)\right]^{-1/\xi}\right\} \\
&= \exp\left\{-\left[1 + \xi\left(\frac{z - \left(\frac{\mu_k - \nu_k}{\tau_k}\right)}{\sigma_k/\tau_k}\right)\right]^{-1/\xi}\right\} \\
&= \exp\left\{-\left[1 + \xi\left(\frac{z - \mu}{\sigma}\right)\right]^{-1/\xi}\right\};
\end{aligned} \tag{2.25}$$

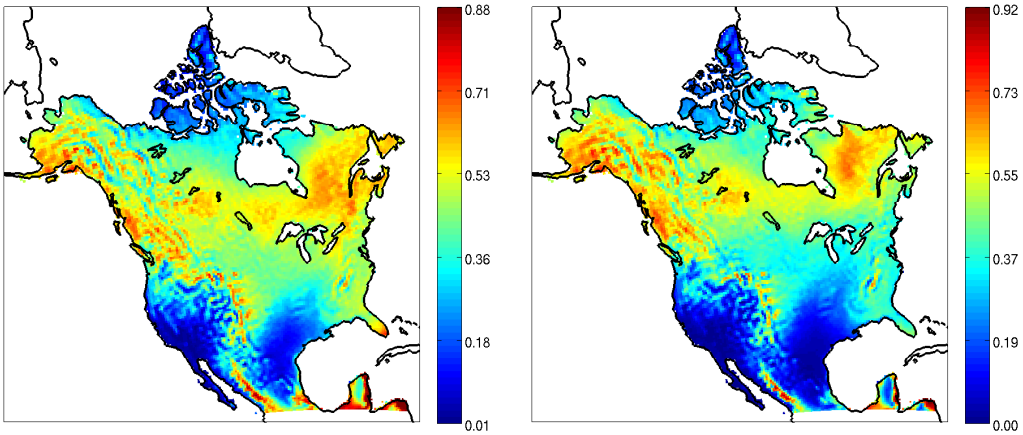
with

$$\begin{aligned}
\sigma &= \sigma_k/\tau_k; \\
\mu &= (\mu_k - \nu_k)/\tau_k.
\end{aligned} \tag{2.26}$$

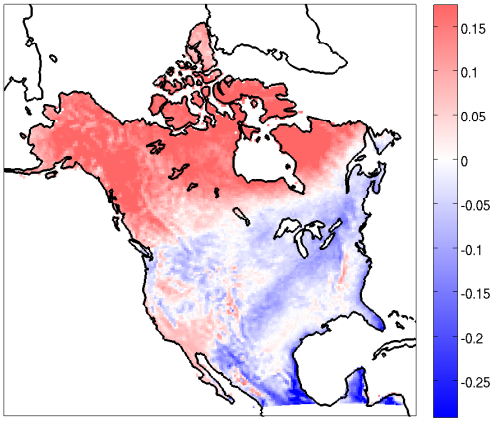
The standardized maximum is also GEV distributed and is characterized with the same shape parameter. Actually, the standardization is equivalent to the usual model (Equation 2.12) but using different link functions for both the scale and the location parameters. The difference lies in the fact that the parameters of these link functions are estimated with the daily precipitations and not only with seasonal maxima. Since more data are taken into account, the estimation variance should be reduced compared to the first approach.

2.4 Results

In the following section we expose the results obtained for summer precipitation while the next one shows the results for the winter, spring and fall precipitation. We recall that we adjusted both occurrence models independently to every grid points and then we selected the best one according to the BIC. Every intensity model has been fit independently of the occurrence models and independently to every grid point. Again, the best model has been chosen according to the BIC. We show in this section the results for every studied grid point.



(a) 1961 (b) 2100



(c) Difference

Figure 2.2 – Model expectation of summer total precipitation occurrence for (a) 1961, (b) 2100 and (c) for the difference between 2100's and 1961's occurrences.

2.4.1 Results for summer precipitation

Figure 2.2 shows the occurrence expectation of summer total precipitation for the 1961 climate, for the 2100 climate and for the difference between both climates. We see that summer precipitation occurrence increases over time in the northern part of the domain (Canada and Alaska) while it globally decreases south of the Great Lakes.

Figure 2.3 shows the chosen model for the precipitation intensity in summer according to the BIC. Model $\mathcal{M}_1$ was never chosen and model $\mathcal{M}_8$ was only chosen for 0.7% of the grid points. The most popular models were $\mathcal{M}_2$ (21%) and $\mathcal{M}_3$ (46%). Since these two models are stationary, summer precipitation intensity may be stationary for 67% of the grid points. As shown in Figure 2.4, intensity seems to increase in the northern part of Canada and Alaska and largely decreases in Mexico. For southern Canada and continental U.S., no changes are

Table 2.3 – Chosen model for describing summer precipitation at six given grid points.

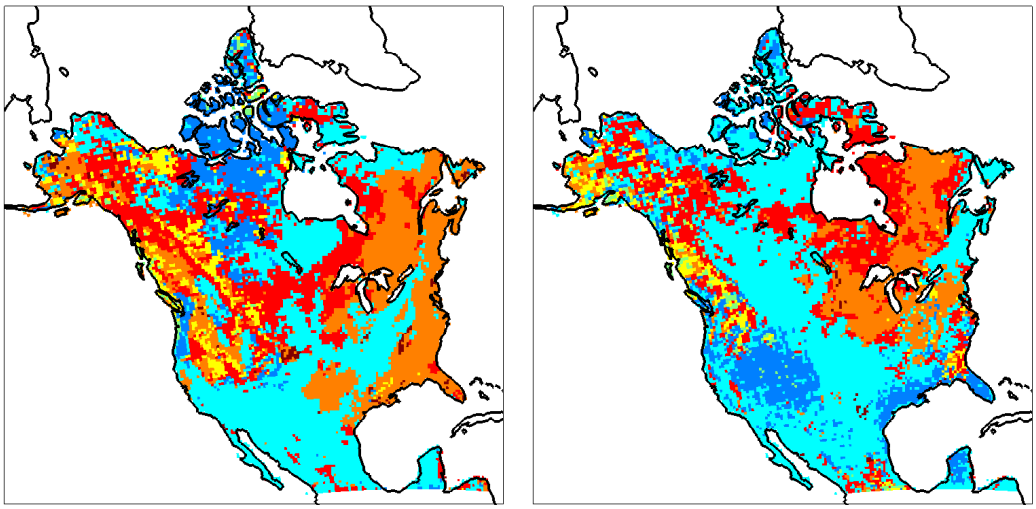
City	Chosen Models	
Montreal	$\mathcal{P}_2$	$\mathcal{M}_3$
Chicago	$\mathcal{P}_2$	$\mathcal{M}_3$
Churchill	$\mathcal{P}_2$	$\mathcal{M}_7$
Vancouver	$\mathcal{P}_2$	$\mathcal{M}_3$
Washington D.C.	$\mathcal{P}_2$	$\mathcal{M}_2$
Nashville	$\mathcal{P}_2$	$\mathcal{M}_3$

expected in summer. Increase in both precipitation occurrence and intensity in northern regions for the summer season come from an increase of convective precipitation, both in occurrence and intensity, as shown on Figure 2.5.

To show that the chosen models for intensity fit the data very well, we present the QQplots for 6 grid points subject to different meteorologies (see Figure 2.6). These grid points under consideration are the ones containing the cities of Montreal, Chicago, Churchill, Vancouver, Washington D.C. and Nashville (see Figure 2.1). The chosen theoretical models (see Table 2.3) fit the data very well up to approximately the 99^{th} quantile. For instance, the theoretical model fits the 6541 observations below the 99^{th} quantile. Upon this point, the theoretical models underestimate intense precipitation values. This is why we did not use these models to estimate intense precipitation. Instead, the upper tail of the distribution is modeled by the GEV distribution. Nonetheless, it seems that the estimation of time-varying expectation and variance with the considered models are quite suitable.

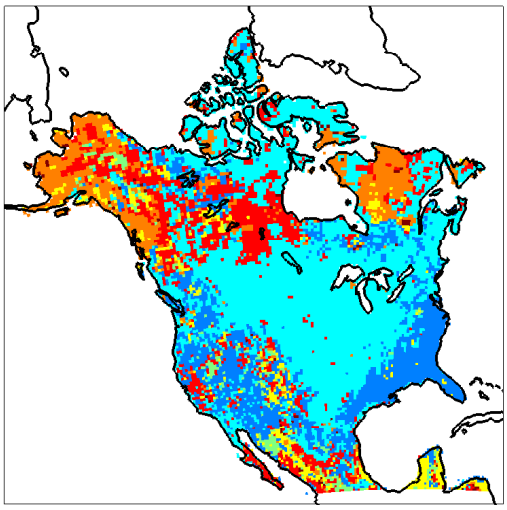
On the other hand, we modeled intense precipitation with the GEV distribution. Figure 2.7 shows the QQPlots for the fitted GEV on the 140 transformed annual maxima, *i.e* the vector $\boldsymbol{M}$, for the six given grid points subject to different meteorologies. From this last figure, we see that the GEV fits the data very well, at least for the six grid points considered. We have no reason to believe that it could be different for the remaining grid points, except perhaps for a small number of them. In the potential scenario where the fit would be poor, the return level estimation variance would be larger but we have to live with it; when the sample size from which the maximum is extracted is large enough, the GEV distribution is the only distribution with reasonable mathematical foundation to model maxima.

For non-stationary models, the 100-years return level estimation was the maximum of the 100-years return level computed for the 140 years of the period. The estimated 100-years return levels for the entire domain are shown in Figure 2.8. We also plot in this last figure the difference between 100-years return level for 2100 and 1961 climates. Note that the comparison does not concern two years, but the difference of model means evaluated at these specific years. It is rather a difference between the climates of 2100 and 1961. Besides, the transient statistical models that we developed provide a climatology for every year in the

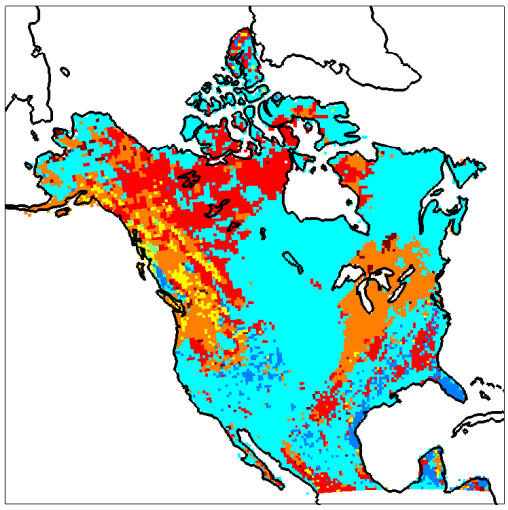


(a) Winter

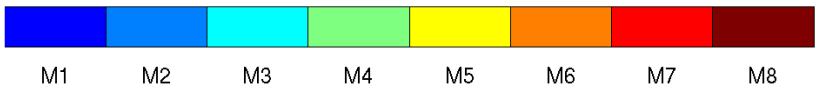
(b) Spring



(c) Summer

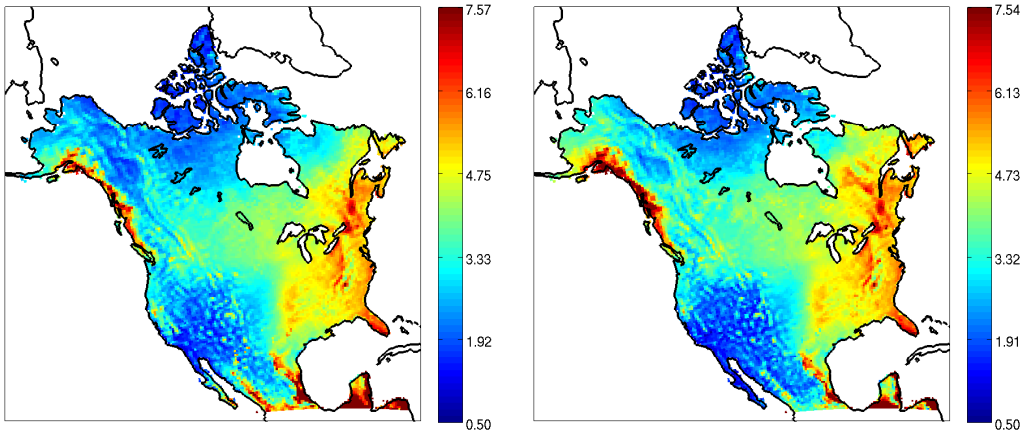


(d) Falls

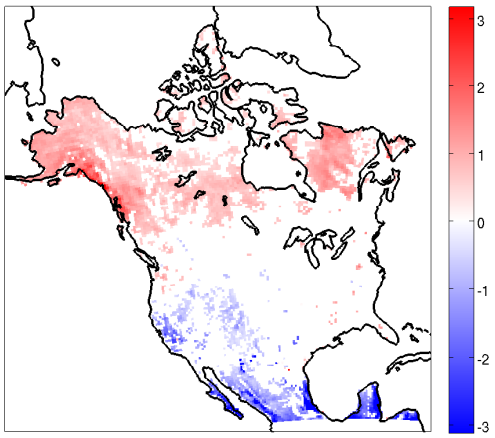


(e)

Figure 2.3 – Chosen model for describing winter, spring, summer and fall precipitation intensity.



(a) 1961 (b) 2100



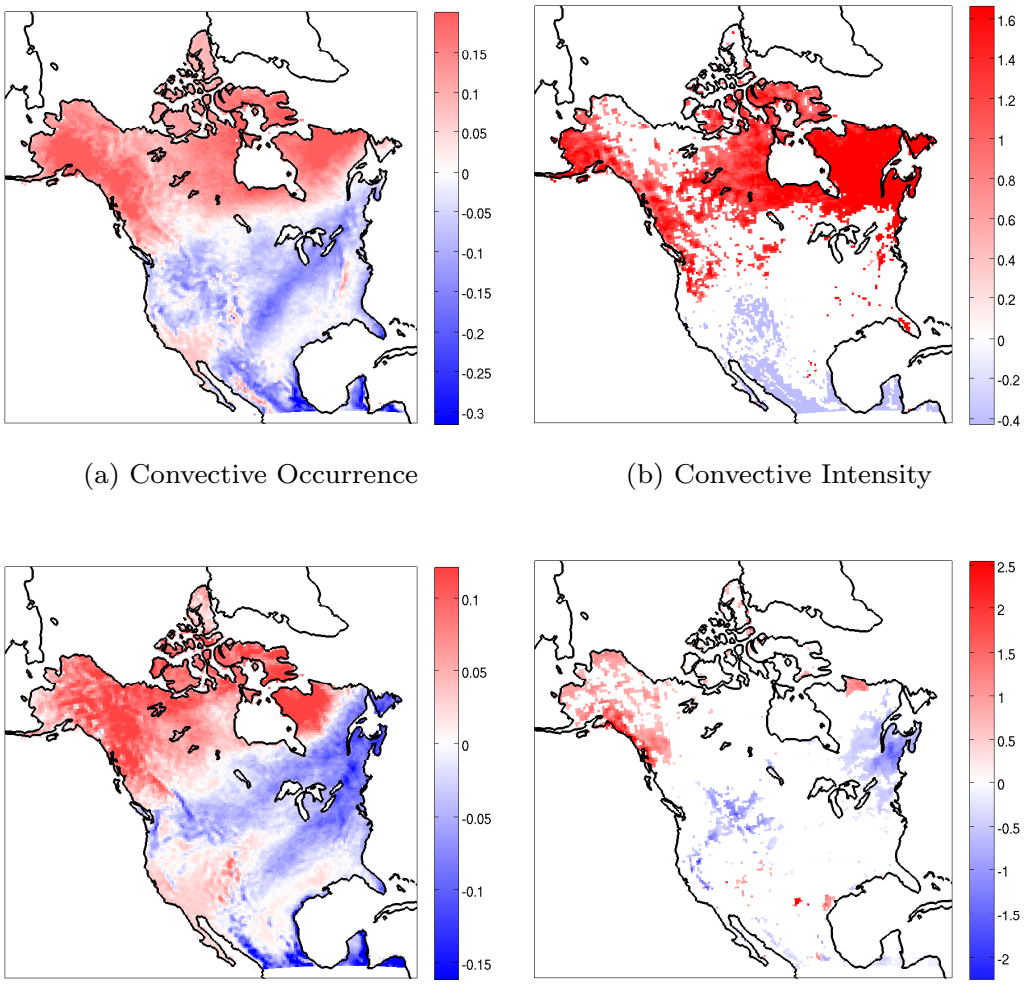
(c) Difference

Figure 2.4 – Model expectation of summer total precipitation intensity in *mm* for (a) 1961, (b) 2100 and (c) for the difference between 2100's and 1961's intensities.

studied interval $[1961, 2100]$. It can be seen in Figure 2.8 that 100-years precipitation levels are stationary for continental United-States, while it increases for Canada and it decreases for Mexico.

2.4.2 Results for the other seasons

Figure 2.9 shows the difference in precipitation between 2100 and 1961 climate for the winter, spring and fall seasons. For the winter season, precipitation increases over most regions in Canada and in United States for both occurrence and intensity. It is only in Mexico that winter precipitation occurrence largely decreases. The spring and fall precipitation are somehow a transition between summer and winter. Changes in winter are mostly due to an increase in both convective and stratiform precipitation in north of Mexico (see Figure 2.10).



(a) Convective Occurrence

(b) Convective Intensity

(c) Large-scale Occurrence

(d) Large-scale Intensity

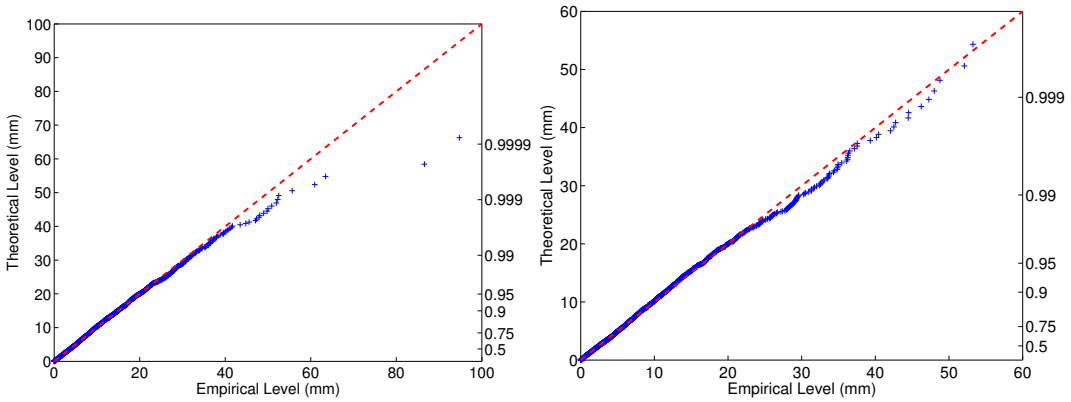
Figure 2.5 – Changes to occurence and intensity of both summer convective and large-scale precipitation.

Figure 2.11 shows the 100-years return level for winter, spring and fall precipitation, also the differences between 2100 and 1961 climate. The increase of winter 100-years precipitation level is mostly located on coastal areas. For spring and fall, the increase of 100-years precipitation level are mostly on the Pacific coast and the eastern part of continental America. In Mexico, no change in 100-years precipitation level is expected.

2.5 Discussion and Conclusion

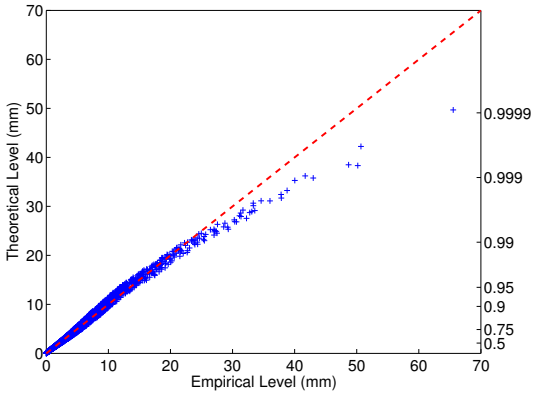
2.5.1 Results comparison

Our results on a CRCM transient simulation are consistent with those from de Elìa and Côté (2010). Their study investigates changes in precipitation and temperature over North America

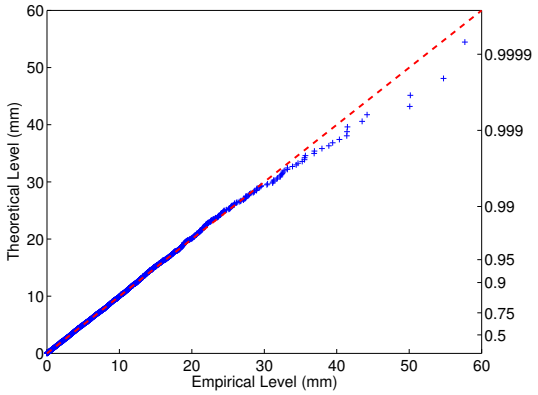


(a) Montreal

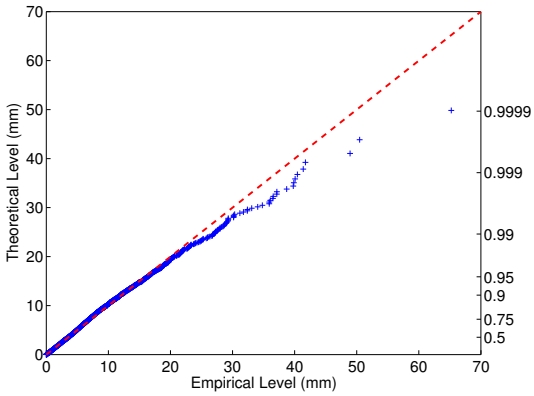
(b) Chicago



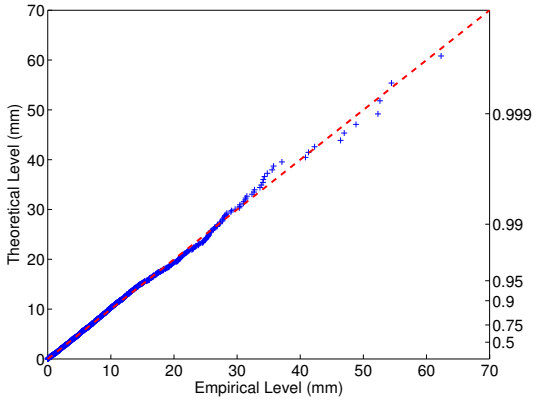
(c) Churchill



(d) Vancouver

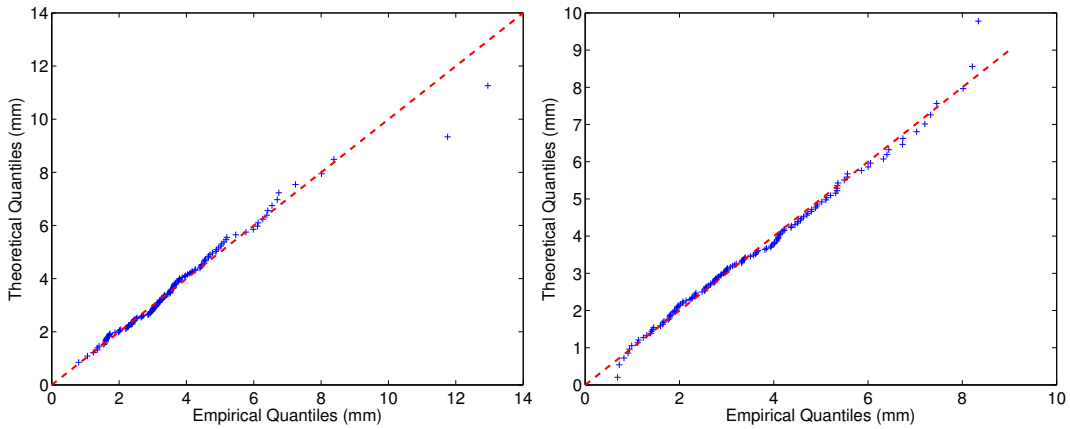


(e) WashingtonDC



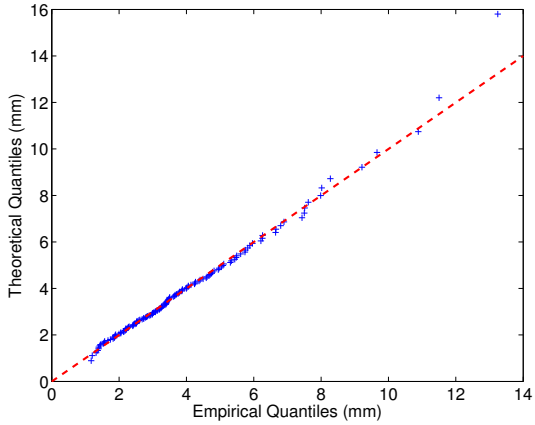
(f) Nashville

Figure 2.6 – QQplots between empirical summer total precipitation and the chosen model.

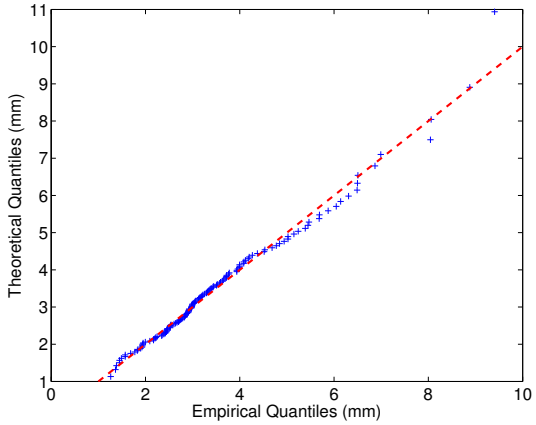


(a) Montreal

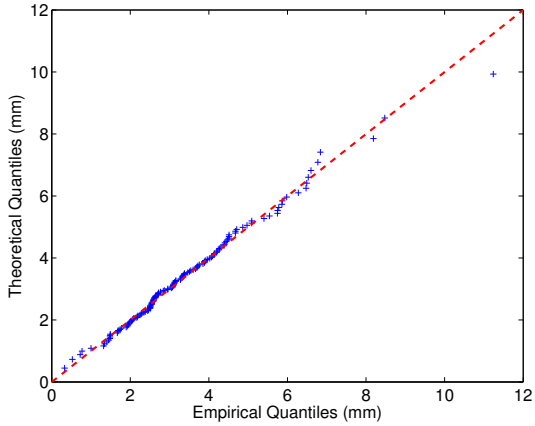
(b) Chicago



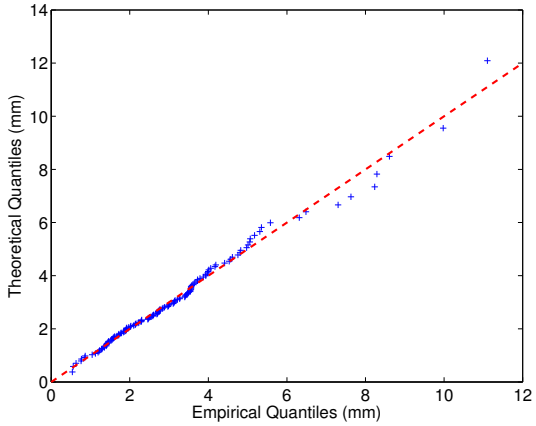
(c) Churchill



(d) Vancouver

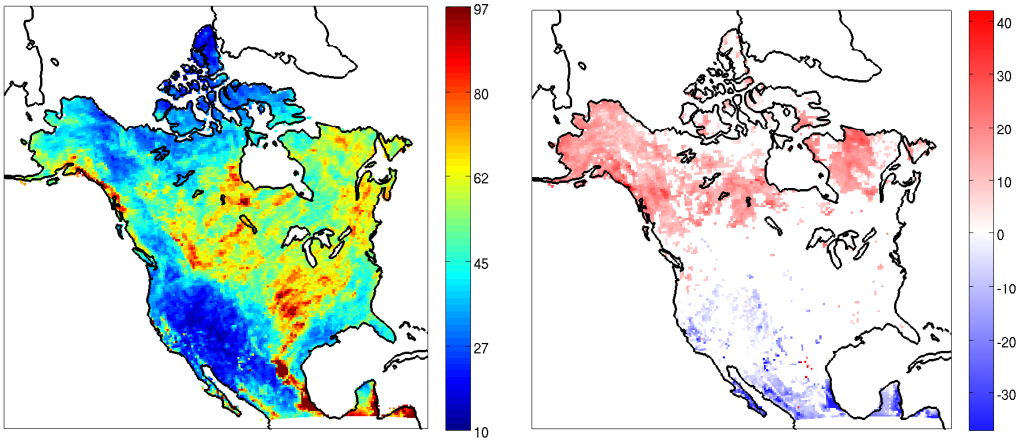


(e) WashingtonDC



(f) Nashville

Figure 2.7 – QQplots between yearly empirical summer total precipitation maxima and the estimated GEV.



(a) 100-years return level

(b) Difference between 2100 and 1961 climates

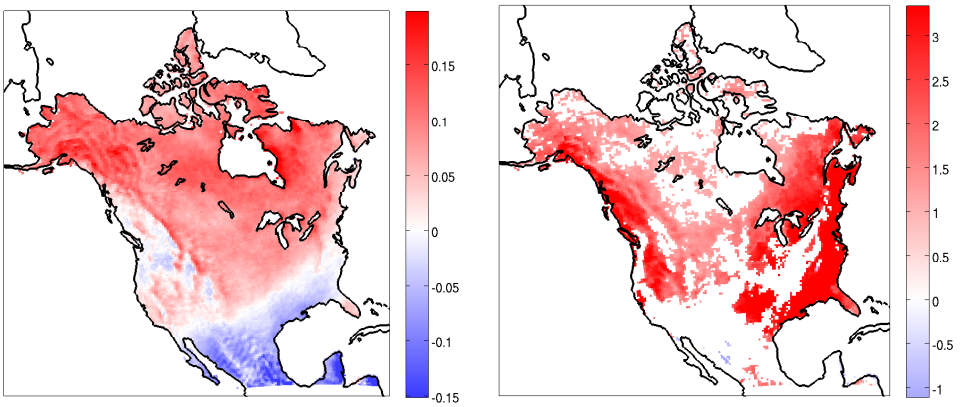
Figure 2.8 – 100-years return level for summer precipitation.

for an ensemble of 18 RCM simulations. Changes were assessed using two 30-years period: $[1961, 1990]$ and $[2041, 2070]$. They showed that summer precipitation are expected to increase in Northern Canada while they are expected to decrease in Mexico and continental U.S. In the same way, our results are consistent with those from Wehner (2012) for a time-sliced analysis on an ensemble of 8 simulations. Moreover, the increase of winter convective precipitation over the Gulf of Mexico, the Great Lakes and the Atlantic coast is also consistent with the study of Paquin et al. (2014).

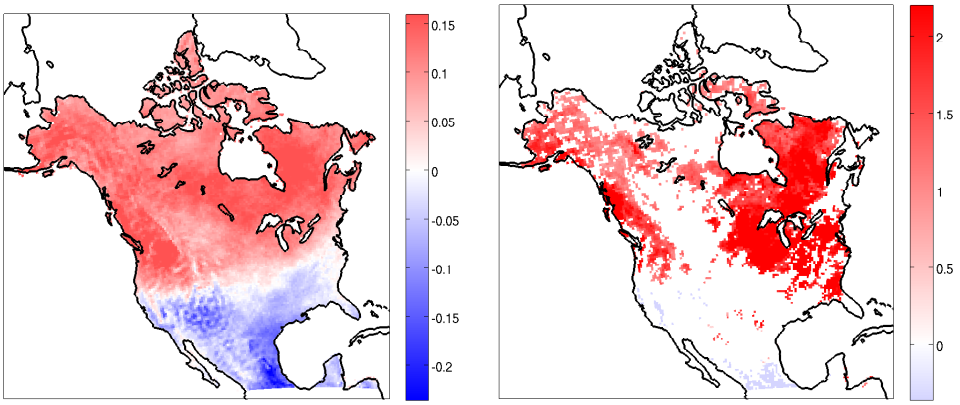
For 100-years return level, our results over the United States are consistent with those from Bukovsky and Karoly (2011) who used two time slices, $[1990, 1999]$ and $[2090, 2099]$, from one simulation of the Weather Research and Forecasting model. An increase of the 100-years precipitation intensity is expected for almost all grid points in the U.S. It may come from any season but summer. For Canada, `CRCM` projections suggest a 100-years precipitation increase, which is consistent with the studies of Mailhot et al. (2012) based on 4 simulations and Monette et al. (2012). It is expected for all seasons but it is much larger in summer for Northern Canada.

2.5.2 Contributions and limitations

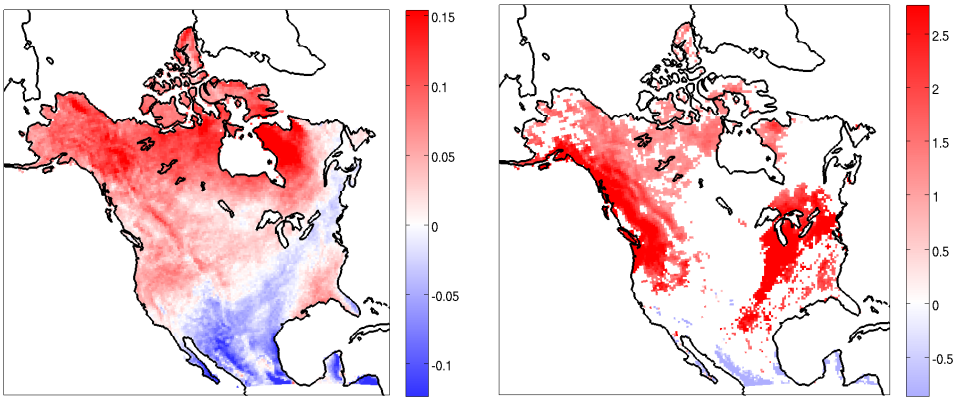
This study has provided a quantitative description of transient precipitation series generated by a RCM. We believe that using a transient simulation leads to a more accurate and complete description; changes estimation are more accurate and non-stationary processes are described. Even if the description of daily precipitation is better, the major gain consists in achieving a non-stationary frequency analysis to assess return level in a changing climate. Non-stationary processes are not estimated with only few data but with the entire dataset. This methodol-



(a) Winter occurrence (b) Winter intensity

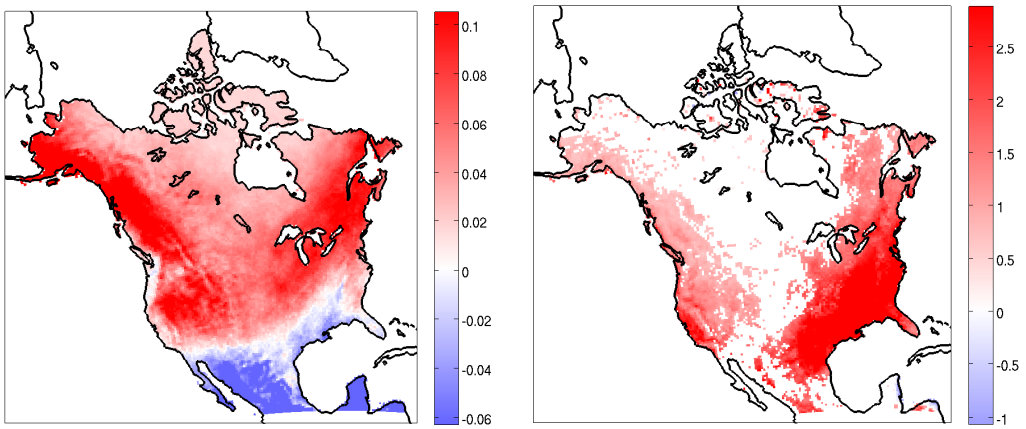


(c) Spring occurrence (d) Spring intensity

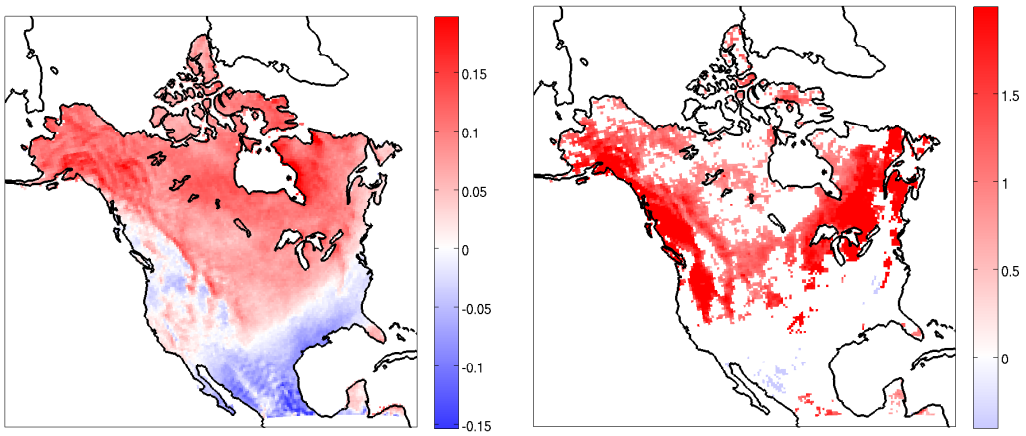


(e) Fall occurrence (f) Fall intensity

Figure 2.9 – Differences between precipitation of 2100 and 1961 for both occurrence and intensity.



(a) Convective Occurrence (b) Convective Intensity

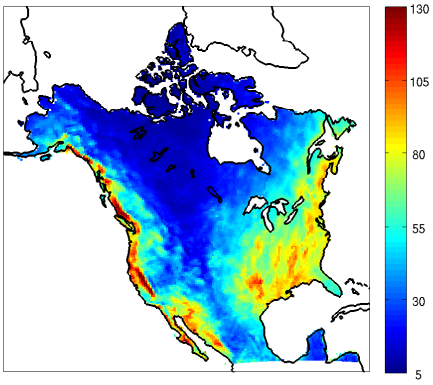


(c) Large-scale Occurrence (d) Large-scale Intensity

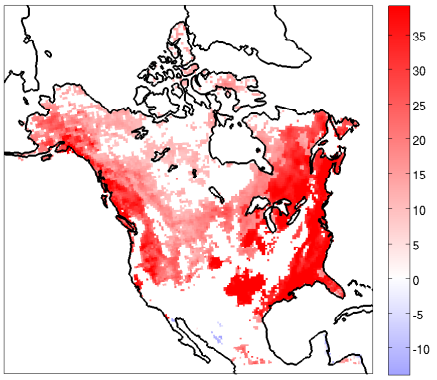
Figure 2.10 – Changes to occurence and intensity of both winter convective and large-scale precipitation.

ogy lies on the assumption stating that non-stationary processes that drive extreme values are consistent with those that drive daily precipitation. Since we statistically modeled the distribution of precipitation, the last assumption was straightforward. Non-stationary statistical distributions were useful to transform the original precipitation series into a stationary series. Thenceforth, a classical frequency analysis was achieved on the transformed series. Thus, every theoretical result and definition, like return period, are directly applicable on the transformed series. We believe that we gain in interpretation since classical frequency analysis is well known and understood.

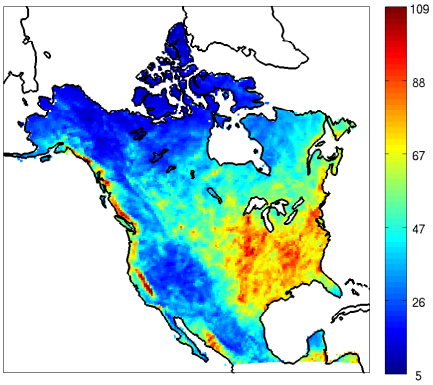
From a meteorological point of view, large-scale precipitation may be correleted on successive days. Nevertheless, the effects of such temporal dependence of large-scale precipitation is intractable in total precipitation. From a statistical point of view, the independence assump-



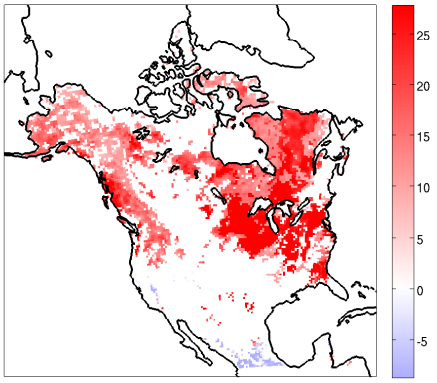
(a) Winter 100-years return level



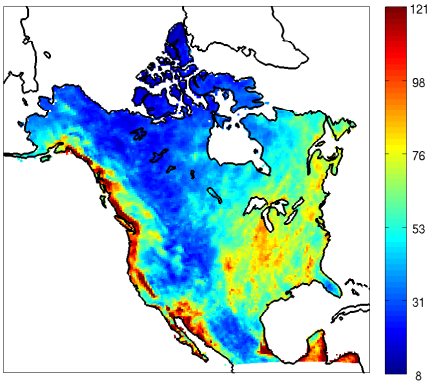
(b) Winter difference between 2100 and 1961 climates



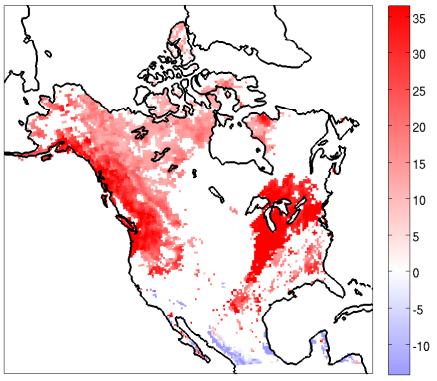
(c) Spring 100-years return level



(d) Spring difference between 2100 and 1961 climates



(e) Fall 100-years return level



(f) Fall difference between 2100 and 1961 climates

Figure 2.11 – 100-years return level for winter, spring and fall precipitation.

tion that we have made is quite difficult to test for both occurrence and intensity of total precipitation because these quantities are non normal and non-stationary. A way to verify it, and to handle it if so, would be to introduce a Markovian dependence (see for example Evin et al., 2011) in the occurrence and intensity models. If such models are chosen, dependence would be significative and it would be taken into account. However, such Markovian modeling for all the considered statistical models is beyond the scope of the present paper. Considering the dependence could reduced the estimation variance. On the other hand, the chosen independence models considered in the present manuscript fit the data very well as shown in Figure 2.6. Therefore, we believe that the consequence of such an assumption is quite limited.

However, we could further improve the methodology using regionalization. Such an approach takes advantage of spatial correlation to improve estimates, provided that spatial correlation can be modeled. For our approach, we are trying to develop regional models based on the occurrence and intensity models presented in this paper. This is an issue for which we are still working.

Despite the fact that the results of this study are consistent with those in the literature, they are based on only one `CRCM` simulation. The results could be substantially different for another simulation. It is generally recognized that considering a large number of models should minimize the structural uncertainty which climate models suffer (Hagedorn et al., 2005). In that sense, our study does not cover a large range of possible outcomes for precipitation, which could be achieved by considering an ensemble of different climate models. We recommend that this be done. Nevertheless, we believe that the approach presented here is a significant contribution to the study of transient simulations.

Acknowledgments

The CRCM data has been generated and supplied by Ouranos. We acknowledge Professor Alain Mailhot for his helpful comments and suggestions. We also acknowledge the two anonymous referees for their suggestions for improving the paper.

References

Bechtold, P., Bazile, E., Guichard, F., Mascart, P., and Richard, E. (2001). A mass flux convection scheme for regional and global models. *Quarterly Journal of the Royal Meteorological Society*, 127(573):869–886.

Bukovsky, M. S. and Karoly, D. J. (2011). A regional modeling study of climate change

impacts on warm-season precipitation in the central United States. *Journal of Climate*, 24(7):1985–2002.

Caya, D. and Laprise, R. (1999). A Semi-Implicit Semi-Lagrangian Regional Climate Model: The Canadian RCM. *Monthly Weather Review*, 127(3):341–362.

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

Christensen, J., Boberg, F., Christensen, O., and Lucas-Picher, P. (2008). On the need for bias correction of regional climate change projections of temperature and precipitation. *Geophysical Research Letters*, 35(20):L20709.

Christensen, J. H. and Christensen, O. B. (2007). A summary of the PRUDENCE model projections of changes in European climate by the end of this century. *Climatic Change*, 81(S1):7–30.

Christensen, J. H., Kumar, K. K., Aldrian, E., An, S.-I., Cavalcanti, I. F. A., de Castro, M., Dong, W., Goswami, P., Hall, A., Kanyanga, J. K., Kitoh, A., Kossin, J., Lau, N.-C., Renwick, J., Stephenson, D. B., Xie, S.-P., and Zhou, T. (2013). Climate Phenomena and their Relevance for Future Regional Climate Change. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, chapter 14, pages 1217–1308. Cambridge University Press, Cambridge, United Kingdom and New York, USA.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer, London, United Kingdom.

Cooley, D. and Sain, S. R. (2010). Spatial hierarchical modeling of precipitation extremes from a regional climate model. *Journal of Agricultural, Biological, and Environmental Statistics*, 15(3):381–402.

de Elìa, R. and Côté, H. (2010). Climate and climate change sensitivity to model configuration in the Canadian RCM over North America. *Meteorologische Zeitschrift*, 19(4):325–339.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Evin, G., Merleau, J., and Perreault, L. (2011). Two-component mixtures of normal, gamma, and Gumbel distributions for hydrological applications. *Water Resources Research*, 47(8).

Fowler, H. J. and Ekström, M. (2009). Multi-model ensemble estimates of climate change impacts on UK seasonal precipitation extremes. *International Journal of Climatology*, 29(3):385–416.

Gilleland, E. and Katz, R. W. (2011). New software to analyze how extremes change over time. *Eos*, 92(2):13–14.

Goderniaux, P., Brouyére, S., Blenkinsop, S., Burton, A., Fowler, H. J., Orban, P., and Dassargues, A. (2011). Modeling climate change impacts on groundwater resources using transient stochastic climatic scenarios. *Water Resources Research*, 47(12).

Hagedorn, R., Doblas-Reyes, F. J., and Palmer, T. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. *Tellus A*, 57(3):219–233.

Hanel, M., Buishand, T. A., and Ferro, C. A. T. (2009). A non-stationary index-flood model for precipitation extremes in transient regional climate model simulations. *Journal of Geophysical Research*, 114(D15):D15107.

IPCC (2012). *Managing the risks of extreme events and disasters to advance climate change adaptation. A special report of Working Groups I and II of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, USA.

Jiao, Y. and Caya, D. (2006). An Investigation of Summer Precipitation Simulated by the Canadian Regional Climate Model. *Monthly Weather Review*, 134(3):919–932.

Katz, R. W. (2010). Statistics of extremes in climate change. *Climatic Change*, 100(1):71–76.

Katz, R. W. (2013). Statistical methods for nonstationary extremes. In *Extremes in a changing climate*, chapter 2, pages 15–37. Springer, New York, USA.

Katz, R. W., Parlange, M. B., and Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in Water Resources*, 25(8):1287–1304.

Kay, A. L. and Jones, D. A. (2012). Transient changes in flood frequency and timing in Britain under potential projections of climate change. *International Journal of Climatology*, 32(4):489–502.

Kharin, V. V. and Zwiers, F. W. (2005). Estimating Extremes in Transient Climate Change Simulations. *Journal of Climate*, 18(8):1156–1173.

Mailhot, A., Beauregard, I., Talbot, G., Caya, D., and Biner, S. (2012). Future changes in intense precipitation over Canada assessed from multi-model NARCCAP ensemble simulations. *International Journal of Climatology*, 32(8):1151–1163.

Mailhot, A., Lachance-Cloutier, S., Talbot, G., and Favre, A.-C. (2013). Regional estimates of intense rainfall based on the Peak-Over-Threshold (POT) approach. *Journal of Hydrology*, 476:188–199.

Mearns, L. O., Gutowski, W., Jones, R., Leung, R., McGinnis, S., Nunes, A., and Qian, Y. (2009). A Regional Climate Change Assessment Program for North America. *Eos, Transactions American Geophysical Union*, 90(36):311.

Mladjic, B., Sushama, L., Khaliq, M. N., Laprise, R., Caya, D., and Roy, R. (2011). Canadian RCM projected changes to extreme precipitation characteristics over Canada. *Journal of Climate*, 24(10):2565–2584.

Monette, A., Sushama, L., Khaliq, M. N., Laprise, R., and Roy, R. (2012). Projected changes to precipitation extremes for northeast Canadian watersheds using a multi-RCM ensemble. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 117(D13).

Music, B. and Caya, D. (2007). Evaluation of the hydrological cycle over the Mississippi river basin as simulated by the Canadian Regional Climate Model (CRCM). *Journal of Hydrometeorology*, 8(5):969–988.

Nakicenovic, N., Alcamo, J., and Davis, G. (2000). *Emissions scenarios*. Cambridge University Press, Cambridge, United Kingdom and New York, USA.

Paquin, D., Caya, D., and Laprise, R. (2002). Treatment of moist convection in the Canadian Regional Climate Model. Technical report, Ouranos, Available from Paquin.Dominique@ouranos.ca, Montréal, Canada.

Paquin, D., de Elìa, R., and Frigon, A. (2014). Change in North American Atmospheric Conditions Associated with Deep Convection and Severe Weather using CRCM4 Climate Projections. *Atmosphere-Ocean*, pages 1–16.

Rasmussen, S. H., Christensen, J. H., Drews, M., Gochis, D. J., and Refsgaard, J. C. (2012). Spatial-scale characteristics of precipitation simulated by regional climate models and the implications for hydrological modeling. *Journal of Hydrometeorology*, 13:1817–1835.

Rootzén, H. and Katz, R. W. (2013). Design life level: quantifying risk in a changing climate. *Water Resources Research*, 49(9):5964–5972.

Roth, M., Buishand, T., Jongbloed, G., Klein Tank, A., and van Zanten, J. (2014). Projections of precipitation extremes based on a regional, non-stationary peaks-over-threshold approach: A case study for the Netherlands and north-western Germany. *Weather and Climate Extremes*, 4:1–10.

Scinocca, J. F., McFarlane, N. A., Lazare, M., Li, J., and Plummer, D. (2008). The CCCma third generation AGCM and its extension into the middle atmosphere. *Atmospheric Chemistry and Physics Discussions*, 8(2):7883–7930.

Sugahara, S., da Rocha, R. P., and Silveira, R. (2009). Non-stationary frequency analysis of extreme daily rainfall in Sao Paulo, Brazil. *International Journal of Climatology*, 29(9):1339–1349.

Titterington, D. M., Smith, A. F., and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Wiley, New York, USA.

van der Linden, P. and Mitchell, J. (2009). *ENSEMBLES: Climate Change and its Impacts: Summary of research and results from the ENSEMBLES project*. Met Office Hadley Centre, Exeter, United Kingdom.

Wehner, M. F. (2012). Very extreme seasonal precipitation in the NARCCAP ensemble: model performance and projections. *Climate Dynamics*, 40(1-2):59–80.

Chapitre 3

Extremes of non-stationary processes : reconciling the standard non-stationary models with the preprocessing approach.

Jonathan Jalbert, Claude Bélisle, Anne-Catherine Favre et

Jean-François Angers.

Cet article est en préparation pour soumission.

Résumé

Plusieurs modèles statistiques ont récemment été développés pour les valeurs extrêmes d'une suite non stationnaire de variables aléatoires. Ces modèles peuvent être divisés en deux groupes. Le premier groupe est composés des approches consistant à inclure la non-stationnarité par la modélisation des paramètres de la loi des valeurs extrêmes généralisée (GEV) pour les maxima de la suite et des paramètres de la loi de Pareto généralisée (GP) pour les excès au-dessus d'un seuil. En pratique, le paramètre de forme de ces lois est supposé invariant pour que le modèle demeure identifiable. Ces approches sont dénommées «modèles standard» dans cet article. Le deuxième groupe concerne les approches de prétraitement qui visent à retirer la non-stationnarité de la suite avant de modéliser les extrêmes de la suite transformée par un modèle stationnaire. Dans les approches standard, la non-stationnarité est estimée en fonction des précipitations extrêmes uniquement. Celles-ci constituent habituellement un échantillon

de taille limitée, ce qui induit une grande variance d'estimation. À l'opposé, le prétraitement des données utilise la série complète pour estimer la non-stationnarité. Néanmoins, supposer que le prétraitement de la série ait complètement retiré la non-stationnarité, ce qui inclut également la non-stationnarité des valeurs extrêmes, contraint généralement son application puisqu'il est très difficile de vérifier cette hypothèse forte en pratique. Dans cet article, une méthode de prétraitement est proposée pour l'analyse des valeurs extrêmes d'une suite non stationnaire. Cette méthode de prétraitement implique la simplification des modèles standard, soit de considérer le paramètre de forme invariant. Il est montré que la méthode de prétraitement développée permet une meilleure estimation des valeurs extrêmes non stationnaires pour des hypothèses équivalentes. L'approche est illustrée pour deux jeux de données réels : les niveaux marins journaliers à San-Francisco (États-Unis d'Amérique) et les précipitations journalières à Manjimup (Australie).

Abstract

Several statistical models have been developed recently for the analysis of extreme values of non-stationary sequences. These statistical models can be divided into two groups of approaches. The first one includes the standard stationary extreme value models, such as the block maxima and the peaks-over-threshold, but where their parameters are allowed to evolve with a vector of covariates explaining the non-stationarity. They are referred to as the standard non-stationary approaches in the present paper. The second group of approaches contains the preprocessing approaches, that consist in removing the non-stationarity from the data before applying a stationary extreme value model. The advantage of preprocessing over the standard approaches lies in the use of the complete dataset to better estimate the non-stationarity before removing it. However, verifying that the preprocessing has completely removed the non-stationnarity is difficult to verify in practice, particularly for the extremal data. In the present paper, a preprocessing methodology that implies the usual assumption of the standard non-stationary models is proposed. It is shown that the proposed preprocessing method gives more accurate estimations than the standard models. The relative efficiency of the proposed approach over the standard ones is illustrated with a simulation study and both methodologies are also compared on two real non-stationary datasets: the sea-levels at San Francisco (USA) and the winter precipitation at Manjimup (Australia).

3.1 Introduction

During the last couple of decades, the extreme value theory has been widely applied to environmental time series for describing meteorological extremes. The extreme value theory provides the asymptotic distributions for the maximum and for the excess over a large threshold of a sequence of independent and identically distributed random variables. Those distributions are

respectively the Generalized Extreme Value (GEV) distribution and the Generalized Pareto (GP) distribution.

Definition 1. *The* Generalized Extreme Value *distribution with location parameter* $\mu \in \mathbb{R}$*, scale parameter* $\sigma > 0$ *and shape parameter* $\xi \in \mathbb{R}$*, denoted by* $\mathcal{GEV}(\mu, \sigma, \xi)$*, is the probability distribution with the following cumulative distribution function denoted by* $\mathcal{GEV}(x|\mu, \sigma, \xi)$*:*

$$\mathcal{GEV}(x|\mu, \sigma, \xi) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}; \tag{3.1}$$

for all x *such that* $1 + \xi(x-\mu)/\sigma > 0$*.*

Definition 2. *The* Generalized Pareto *distribution with scale parameter* $\sigma > 0$ *and shape parameter* $\xi \in \mathbb{R}$*, denoted by* $\mathcal{GP}(\sigma, \xi)$*, is the probability distribution with the following cumulative distribution function denoted by* $\mathcal{GP}(x|\sigma, \xi)$*:*

$$\mathcal{GP}(x|\sigma, \xi) = 1 - \left(1 + \xi\frac{x}{\sigma}\right)^{-1/\xi}; \tag{3.2}$$

for all $x > 0$ *if* $\xi \geq 0$ *and for all* $x \in (0, -\sigma/\xi)$ *if* $\xi < 0$*.*

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with distribution function F. Let the maximum among the n first variables be denoted as follows:

$$M_n = \max\{X_1, \ldots, X_n\}. \tag{3.3}$$

The central result of the extreme value theory is presented in Theorem 3.1.

Theorem 3.1 Fisher and Tippett, 1928, Gnedenko, 1943**.**
If there exist sequences of constants $(a_n > 0 : n \geq 1)$ *and* $(b_n : n \geq 1)$ *such that the sequence* $\left(\frac{M_n - b_n}{a_n} : n \geq 1\right)$ *converges in distribution to a non degenerate distribution, then this distribution belongs to the GEV family. In other words, there exists constants* μ*,* σ *and* ξ *such that*

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{M_n - b_n}{a_n} \leq x\right] = \mathcal{GEV}(x|0, 1, \xi), \forall x \in \mathbb{R}. \tag{3.4}$$

We say that the distribution F belongs to the domain of attraction of the $\mathcal{GEV}(0, 1, \xi)$ if there exist some constants $(a_n > 0 : n \geq 1)$ and $(b_n : n \geq 1)$ such that Eq. (3.4) is fullfilled.

Another important result from the extreme value theory concerns the asymptotic distribution of the excess over a large threshold u:

Theorem 3.2 Pickands III, 1975**.** *Let* $x_+ = \sup\{x : F(x) < 1\}$ *be the right end point of the support of* F*. Then, the two following statements are equivalent:*

(i) The distribution F *belongs to the domain of attraction of* $\mathcal{GEV}(0, 1, \xi)$*.*

(ii) There exists a positive function $c(u)$ such that

$$\lim_{u \uparrow x_+} \mathbb{P}\left(\frac{X-u}{c(u)} \leq y \middle| X > u \right) = \mathcal{G}P(y|1,\xi). \tag{3.5}$$

The previous theorems lead to different modeling strategies for extreme value analysis of actual dataset (see Coles, 2001, Chap. 3–4). Practical use of Theorem 3.1 consists in dividing the independent and identically distributed data into K homogeneous blocks of length L, such as $(X_{kl} : 1 \leq l \leq L), 1 \leq k \leq K$. When the study is related to environmental processes, the data naturally arises in blocks corresponding for example to one year or one season. Let $M_k = \max\{X_{k1}, \ldots, X_{kL}\}$ be the maximum among the data of block k. If the block size L is large enough, then Theorem 3.1 suggests that the distribution for any block maximum M_k can be approximated by the GEV distribution:

$$\frac{M_k - b_L}{a_L} \overset{\mathcal{L}}{\approx} \mathcal{G}EV(0,1,\xi). \tag{3.6}$$

Denoting b_L by μ_L and a_L by σ_L, the model can be written as follows:

$$M_k \overset{\mathcal{L}}{\approx} \mathcal{G}EV(\mu_L, \sigma_L, \xi), 1 \leq k \leq K. \tag{3.7}$$

This model, referred to as block maxima, considers only one value per block. A waste of information occurs when more than one value per block can be attributed to the extreme behavior. The Peaks-Over-Threshold (POT) approach, originally popularized by Davison and Smith (1990), was developed to include those potential additional extreme values in the statistical model. The approach is based on the following approximation for the distribution of the excesses over a sufficiently large threshold u as suggested by Theorem 3.2:

$$\mathbb{P}\left(\frac{X-u}{c(u)} \leq y \middle| X > u \right) \overset{\mathcal{L}}{\approx} \mathcal{G}P(y|1,\xi). \tag{3.8}$$

Denoting $c(u)$ by σ_u, the model can be written as follows:

$$\mathbb{P}\left(X - u \leq y | X > u\right) \overset{\mathcal{L}}{\approx} \mathcal{G}P(y|\sigma_u, \xi). \tag{3.9}$$

In environmental datasets, time series commonly exhibit temporal dependence and temporal non-stationarity. In the present context, temporal non-stationarity refers to the changes through time of the process characteristics. While the extreme value theory is still applicable under short-range dependence of large values (see Chap. 3 of Leadbetter et al., 1983, for the formal description), there exists no general theory for extremes of non-stationary sequences. Non-stationarity in extreme values are, in most cases, handled with statistical modeling of the asymptotic distribution parameters (see Coles, 2001, Chap. 6): the GEV distribution for the maxima and the GP distribution for the excesses over a large threshold. These parameters are modeled as functions of covariates explaining the non-stationarity. It is common though

to consider an invariant shape parameter for both the GEV and the GP distributions (see for example Katz, 2013) notably because its estimation variance is very large. Also, Renard et al. (2013) showed that identifiability issues may occur when the shape parameter as well as the remaining parameters are modeled with time as a covariate, which makes inference impossible. Generally, the estimations of such non-stationary parameters have a large variance due to the usual small sample sizes. An alternative approach to the modeling of the extreme value distribution parameters is to remove the non-stationarity in the sequence before applying a stationary model for the extremes. Such an approach is commonly referred to as data preprocessing in time series analysis (Chatfield, 2004). An example of this strategy in the context of extreme values is provided by Eastoe and Tawn (2009). They applied a Box-Cox location-scale transformation to the entire dataset before applying an extreme value model for the excesses over a large invariant threshold. Critical to this approach is that the model for the full dataset has to remove as much as possible the non-stationarity in the extremes. But within the limit of this assumption, using the full dataset can provide a more precise description of the non-stationarity than using only the extremal data.

Despite the fact that Eastoe and Tawn (2009) argued that preprocessing gives more efficient estimates for describing non-stationary extremes, some practitioners are reluctant to use preprocessing. They notably believe that information in the bulk distribution has little value in the tail behavior. In the present paper, we suggest a methodology that has the advantages of both approaches. We propose a simple preprocessing procedure for the complete data that implies the simplification generally used in the standard non-stationary extreme models, say an invariant shape parameter. Therefore, the non-stationary model can be seen as a standard non-stationary model but where the non-stationarity in the remaining parameters is estimated with the complete dataset. The remainder of the paper is as follows. Section 3.2 introduces the notation used throughout this paper. Section 3.3 presents the standard statistical approches for modeling non-stationary extremes. Section 3.4 describes our preprocessing methodology. Section 3.5 is devoted to a simulation study. Section 3.6 contains several case studies with environmental dataset. The conclusion is included in Section 3.7.

3.2 Notation and assumptions

Let K be a positive integer and let $(X_{kl} : 1 \leq k \leq K, 1 \leq l < \infty)$ be a rectangular array of random variables. For each index k, let the sequence $(X_{kl} : l \geq 1)$ be identically distributed with the distribution function F_k. The maximum among the l random variables of a given block k is denoted by:

$$M_{kl} = \max\{X_{k1}, X_{k2}, \ldots, X_{kl}\}, \text{ for } 1 \leq l < \infty. \tag{3.10}$$

Throughout the paper, we assume the existence of some constants $a_{kl} > 0$, $b_{kl} \in \mathbb{R}$ and $\xi_k \in \mathbb{R}$ such that for $1 \leq k \leq K$:

(A1) the $D(u_l)$ *condition* specified by Leadbetter et al. (1983, Chap. 3) on the dependence of the sequence $(X_{kl} : l \geq 1)$ is satisfied with $u_l = a_{kl}x + b_{kl}$ for all $x \in \mathbb{R}$;

(A2) the sequence $\left(\frac{M_{kl}-b_{kl}}{a_{kl}}\right)$ converges in distribution to the GEV distribution when $l \to \infty$, *i.e.*:

$$\frac{M_{kl} - b_{kl}}{a_{kl}} \xrightarrow{\mathcal{L}} \mathcal{G}EV(0, 1, \xi_k) \text{ when } l \to \infty. \tag{3.11}$$

Assumption (A1) limits the range of the dependence for the extreme values. In the independent case, Assumption (A2) can be written in terms of the distributions functions F_k:

$$\lim_{l\to\infty} F_k^l\,(a_{kl}x + b_{kl}) = \mathcal{G}EV(0, 1, \xi_k),\ \forall x \in \mathbb{R},\ \text{for } 1 \leq k \leq K. \tag{3.12}$$

Assumption (A2) also implies (see Theorem 3.2) that for every $1 \leq k \leq K$, there exists a positive function $c_k(u_k)$ such that

$$\lim_{u\uparrow x_+} \mathbb{P}\left(\frac{X_{kl} - u_k}{c_k(u_k)} \leq y \middle| X_{kl} > u_k\right) = \mathcal{G}P(y|1, \xi_k),\ \text{for } 1\ \leq l < \infty. \tag{3.13}$$

Also, suppose that there exists a set of covariates $(\boldsymbol{x}_k : 1 \leq k \leq K)$ explaining the non-stationarity of the sequence $(X_{kl} : l \geq 1), 1 \leq k \leq K$, and/or of the sequence $(M_{kl} : 1 \leq k \leq K)$ depending of the context. Note that the bold font is used to distinguish the covariates $\boldsymbol{x}_k$ from the possible realizations x_{kl} of the random variable X_{kl}.

3.3 Non-stationary statistical approaches for extremes

The present section describes three approaches commonly used for dealing with the non-stationarity in extreme value analysis. The first two consists in the non-stationary block maxima model and the non-stationary POT model. Both of them are referred to as the "standard models" in this paper. The last methodology concerns preprocessing; we present the most popular one for extreme value analysis.

3.3.1 Non-stationary block maxima

If the block size L if sufficiently large, then Eq. (3.11) suggests the following approximation for $1 \leq k \leq K$:

$$\frac{M_{kL} - b_{kL}}{a_{kL}} \overset{\mathcal{L}}{\approx} \mathcal{G}EV(0, 1, \xi_k). \tag{3.14}$$

For notation simplicity, let us drop the block size index L and let us introduce the parameters μ_k and σ_k for the unnormalized maximum M_k:

$$M_k \overset{\mathcal{L}}{\approx} \mathcal{G}EV(\mu_k, \sigma_k, \xi_k). \tag{3.15}$$

The non-stationary block maxima model, as proposed by Coles (2001, Chap. 6), consists in handeling the non-stationarity in the maxima sequence $(M_k : 1 \leq k \leq K)$ by modeling the

GEV parameters as functions of the covariates $\boldsymbol{x}_k$, where it is common though to consider an invariant shape parameter:

$$\mu_k = f_\mu(\boldsymbol{x_k}) \tag{3.16}$$

$$\sigma_k = f_\sigma(\boldsymbol{x_k}) \tag{3.17}$$

$$\xi_k = \xi. \tag{3.18}$$

3.3.2 Non-stationary excesses over a large threshold

For a given k, Eq. (3.13) suggests the following approximation for the distribution of the excess over a sufficiently large threshold u_k:

$$\mathbb{P}\left(\frac{X_{kl} - u_k}{c_k(u_k)} \leq y \middle| X_{kl} > u_k\right) \overset{\mathcal{L}}{\approx} \mathcal{G}P(y|1, \xi_k). \tag{3.19}$$

Denoting $c_k(u_k)$ by σ_{ku_k}, the approximated distribution for the unnormalized excess can be written as:

$$\mathbb{P}\left(X_{kl} - u_k \leq y | X_{kl} > u_k\right) \overset{\mathcal{L}}{\approx} \mathcal{G}P(y|\sigma_{ku_k}, \xi_k). \tag{3.20}$$

The non-stationary POT model, as proposed by Coles (2001, Chap. 6), consists in modeling the GP parameters as functions of $\boldsymbol{x_k}$. Again, it is common to consider an invariant shape parameter:

$$u_k = f_u(\boldsymbol{x}_k); \tag{3.21}$$

$$\sigma_{ku_k} = f_{\sigma_k}(u_k, \boldsymbol{x}_k); \tag{3.22}$$

$$\xi_k = \xi. \tag{3.23}$$

As a consequence of the sequence dependence (see Assumption A1), the excesses might occur in clusters. When it comes to parameter estimation of the POT model, this dependence is difficult to take into account in the likelihood function of the excesses. Coles (2001, Chap. 5) proposes to filter the excesses by retaining only the maximum of a cluster. Then, the sequence of the filtered excesses can be assumed approximately independent.

3.3.3 Data preprocessing

A strategy for handling the non-stationarity in a sequence consists in removing it before applying a suitable model for a stationary series. The method developed by Eastoe and Tawn (2009) for a sequence of non-stationary random variables is presented. The notation has been modified for the rectangular array of random variables as defined in Section 3.2. Eastoe and Tawn (2009) proposed the following Box-Cox location-scale transformation:

$$\frac{X_{kl}^{\lambda(\boldsymbol{x}_k)} - 1}{\lambda(\boldsymbol{x}_k)} = \omega(\boldsymbol{x}_k) + \varphi(\boldsymbol{x}_k) Z_{kl}; \tag{3.24}$$

to obtain the approximate stationary sequence $(Z_{kl} : l \geq 1), 1 \leq k \leq K$, from the non-stationary sequence $(X_{kl} : l \geq 1), 1 \leq k \leq K$, where λ, ω and $\log \varphi$ are linear functions of the covariates. Eastoe and Tawn (2009) argued that most, if not all, of the non-stationarity has been removed in the sequence Z_{kl}. However, they did not model the excesses over a large threshold u with a stationary model. They instead used the non-stationary model described in Section 3.3.2 but with an invariant threshold.

Inference for this model can be conducted with a two steps procedure: first, estimate the Box-Cox location-scale transformation parameters $(\lambda, \omega, \varphi)$ and second, estimate the non-stationary POT parameters. Eastoe and Tawn (2009) estimated the Box-Cox location-scale transformation parameters assuming that the underlying distribution is Gaussian. Therefore, the likelihood function is known and easy to work with. They also argued that such an inference procedure is robust to observations in the tails even if the underlying distribution is non-normal.

Note that this method could also be adapted for block maxima. The block maxima

$$(\max\{Z_{k1}, \ldots, Z_{kl}\} : 1 \leq k \leq K)$$

can be modeled with the GEV distribution.

3.4 Preprocessing based on standardization

In the present section, we propose a preprocessing methodology for studying the extreme values of a non-stationary sequences as defined in Section 3.2. It takes the advantages of the preprocessing approach, say to remove the non-stationarity using more valuable data than only with the extreme values, but only requiring the usual assumption of the standard models, say an invariant shape parameter.

3.4.1 Standardized extremes

Let $(\nu_k : 1 \leq k \leq K)$ and $(\tau_k > 0 : 1 \leq k \leq K)$ be some sequences of constants. We consider the following transformation for the random variables composing the rectangular array defined in Section 3.2:

$$X'_{kl} = \frac{X_{kl} - \nu_k}{\tau_k}. \tag{3.25}$$

Let M'_{kl} denotes the maximum of the l transformed random variables of a given block k:

$$M'_{kl} = \max\{X'_{k1}, X'_{k2}, \ldots, X'_{kl}\}, \text{ for } l = 1, 2, 3, \ldots \tag{3.26}$$

It is worth noting that

$$M'_{kl} = \frac{M_{kl} - \nu_k}{\tau_k}; \tag{3.27}$$

and combined with Eq. (3.11),

$$\frac{M'_{kl} - b'_{kl}}{a'_{kl}} \xrightarrow{\mathcal{L}} \mathcal{G}EV(0, 1, \xi_k), \text{ when } l \to \infty; \tag{3.28}$$

with

$$a'_{kl} = \frac{a_{kl}}{\tau_k} \quad \text{and} \quad b'_{kl} = \frac{b_{kl} - \nu_k}{\tau_k}. \tag{3.29}$$

The following proposition concerns the choice of the sequences of constants $(\nu_k : 1 \leq k \leq K)$ and $(\tau_k > 0 : 1 \leq k \leq K)$:

Proposition 3.1. *If the sequence* $(X'_{kl} : l \geq 1)$, $1 \leq k \leq K$, *is identically distributed, then for any* k_1 *and* k_2 *in* $\{1, 2, \ldots, K\}$*:*

$$\lim_{l\to\infty} \frac{a_{k_1 l}/\tau_{k_1}}{a_{k_2 l}/\tau_{k_2}} = 1; \qquad \lim_{l\to\infty} \frac{\frac{b_{k_1 l} - \nu_{k_1}}{\tau_{k_1}} - \frac{b_{k_2 l} - \nu_{k_2}}{\tau_{k_2}}}{a_{k_1 l}/\tau_{k_1}} = 0; \qquad \xi_{k_1} = \xi_{k_2}; \tag{3.30}$$

or in other terms:

$$\lim_{l\to\infty} \frac{a'_{k_1 l}}{a'_{k_2 l}} = 1; \qquad \lim_{l\to\infty} \frac{b'_{k_1 l} - b'_{k_2 l}}{a'_{k_1 l}} = 0; \qquad \xi_{k_1} = \xi_{k_2}. \tag{3.31}$$

The proof is given in Appendix 3.A.

3.4.2 In practice

It is possible to develop a preprocessing methodology based on Proposition 3.1. Let us define ν_k and τ_k as follows:

$$\nu_k = \mathbb{E}(X_{kl}) \quad \text{and} \quad \tau_k^2 = \mathbb{V}ar(X_{kl}). \tag{3.32}$$

Then, the standardized random variables $(X'_{kl}, 1 \leq k \leq K, 1 \leq l \leq L)$, have by construction a null expectation and a unit variance. The transformation of Eq. (3.25) with those defined ν_k and τ_k thus corresponds to a standardization. If the standardized random variables $(X'_{kl}, 1 \leq k \leq K, 1 \leq l \leq L)$, are assumed identically distributed, then the result of Proposition 3.1 suggests that the normalizing constants a'_{kL} and b'_{kL} do not depend on k and so does the shape parameter. Therefore, for a large enough block size L, the distribution of the standardized maxima can be approximated by the following distribution as suggested by Eq. (3.28):

$$M'_{kL} \overset{\mathcal{L}}{\approx} \mathcal{G}EV(\mu_L, \sigma_L, \xi); \tag{3.33}$$

where $\mu_L \in \mathbb{R}$, $\sigma_L > 0$ and $\xi \in \mathbb{R}$. The approximate distribution for the unstandardized maxima can then be obtained by the inverse transformation:

$$M_{kL} \overset{\mathcal{L}}{\approx} \mathcal{G}EV(\tau_k \mu_L + \nu_k, \tau_k \sigma_L, \xi). \tag{3.34}$$

In order to use the result of Proposition 3.1, we assumed that the random variables $(X'_{kl} : 1 \leq k \leq K, 1 \leq l \leq L)$ were identically distributed. At least, all the expectations and the

variances are equal but it is not necessarily the case for the remaining moments (providing they exist). Assuming that the standardized random variables are identically distributed is a strong assumption. Nevertheless, this assumption implies the standard non-stationary model for block maxima. Indeed, Eq. (3.34) and (3.15) with $\xi_k = \xi$ are equivalent. Therefore, the stationarity of the first two moments is sufficient to meet the common simplification used in the standard non-stationary extreme models, say an invariant ξ. In general, critical to the use of preprocessing is that the transformation have to remove completely the non-stationarity of the original data. However, this stationarity assumption can be relaxed with our proposed preprocessing methodology in so far as a invariant shape parameter is considered.

The preprocessing procedure for a block maxima analysis can be resumed as follows:

1. Standardize the maxima:
$$M'_k = \frac{M_k - \nu_k}{\tau_k}.$$
2. Estimate the GEV parameters with the standardized maxima:
$$M'_k \overset{\mathcal{L}}{\approx} \mathcal{G}EV(\mu, \sigma, \xi).$$
3. Do the inverse transformation to obtain the distribution of the orginal maxima:
$$M_k \overset{\mathcal{L}}{\approx} \mathcal{G}EV(\tau_k\mu + \nu_k, \tau_k\sigma, \xi).$$

Equivalently for the excesses over a sufficiently large threshold u, Theorem 3.2 and Proposition 3.1 suggest that if the sequence $(X'_{kl}, 1 \leq l \leq L), 1 \leq k \leq K$, is assumed strictly stationary, then
$$\mathbb{P}\left(X'_{kl} - u \leq y | X'_{kl} > u\right) \overset{\mathcal{L}}{\approx} \mathcal{G}P(y|\sigma_u, \xi). \tag{3.35}$$
The distribution for the unstandardized excesses is given by:
$$\mathbb{P}\left(X_{kl} - u_k \leq y | X_{kl} > u_k\right) \overset{\mathcal{L}}{\approx} \mathcal{G}P(y|\sigma_{ku}, \xi); \tag{3.36}$$
where
$$u_k = \tau_k u + \nu_k \qquad \text{and} \qquad \sigma_{uk} = \tau_k \sigma_u. \tag{3.37}$$

The preprocessing procedure for an excesses over a large threshold analysis can be resumed as follows:

1. Standardize the original data:
$$X'_{kl} = \frac{X_{kl} - \nu_k}{\tau_k}.$$
2. Define a sufficiently large threshold u.
3. Estimate the GP parameters with the excesses
$$\mathbb{P}\left(X'_{kl} - u \leq y | X'_{kl} > u\right) \overset{\mathcal{L}}{\approx} \mathcal{G}P(y|\sigma, \xi).$$
4. Do the inverse transformation to obtain the distribution of the orginal excesses:
$$\mathbb{P}\left(X_{kl} - u_k \leq y | X_{kl} > u_k\right) \overset{\mathcal{L}}{\approx} \mathcal{G}P(y|\sigma_{uk}, \xi);$$
where u_k and σ_{u_k} are defined as in Eq. (3.37).

3.4.3 Estimating the constants for the standardization

In a general context, the block means and the block standard deviations can be estimated using nonlinear regression analysis (see notably Bates and Watts, 1988). The mean ν_k, view as a function of the block index k and its associated covariates $\boldsymbol{x}_k$, is modeled using a given parametric regression function f with unknown parameters θ associated to the data with the following relation:

$$X_{kl} = f(\boldsymbol{x}_k, \theta) + \varepsilon_{kl}; \tag{3.38}$$

where $\mathbb{E}[\varepsilon_{kl}] = 0$. Because the random sequence $(X_{kl} : 1 \leq l \leq L)$ for any given k may be dependent and also because the variance can evolve with the covariate vector $(\boldsymbol{x}_k : 1 \leq k \leq K)$, the parameter estimation is based on weighted least squares for taking into account both the dependence and the heteroscedasticity. Several parametric models can be considered and the one with the smallest mean squared error may be chosen.

For the variance estimation, sample block estimates can be obtained in the following way:

$$S_k^2 = \frac{1}{L}\sum_{l=1}^{L}\left(X_{kl} - f(\boldsymbol{x}_k, \hat{\theta})\right)^2, 1 \leq k \leq K; \tag{3.39}$$

where $\hat{\theta}$ is the estimation of θ obtained by minimizing the weighted squares. Again, nonlinear regression can be used to model the series of variance estimations $(s_k^2 : 1 \leq k \leq K)$ as a function of the covariate $\boldsymbol{x}_k$:

$$S_k^2 = g(\boldsymbol{x}_k, \phi) + \epsilon_{kl}; \tag{3.40}$$

where $\mathbb{E}[\epsilon_{kl}] = 0$, g is the parametric regression function with parameters ϕ.

3.5 Simulation study

In the present section, the standard non-stationary model for block maxima is compared with our preprocessing model. Generally, the purpose for applying the extreme value theory consists in estimating large quantiles which are, most of the time, referred to as return levels. Therefore, the comparison is based on the estimation efficiency of such return levels. The simulation design used for this purpose was developed by Eastoe and Tawn (2009) for evaluating the efficiency of their own preprocessing model over the standard non-stationary peaks-over-threshold model. We believe that this simulation design is also suitable for studying the performance of our approach. For a given sample size K and a given block size L, 10,000 replicates of the rectanguar array $(X_{kl} : 1 \leq k \leq K, 1 \leq l \leq L)$ were generated with the following distribution:

$$X_{kl} \sim \mathcal{G}umbel\left\{10 + \frac{k}{50}, \exp\left(\frac{7}{10} + \frac{k}{500}\right)\right\}; \tag{3.41}$$

where $\mathcal{G}umbel(\alpha, \beta)$ denotes a random variable with the following distribution function:

$$F(x) = \exp\left\{-\exp\left(\frac{x-\alpha}{\beta}\right)\right\}. \tag{3.42}$$

The expectation and the variance as functions of index k are expressed as follows:

$$\mathbb{E}[X_{kl}] = 10 + \frac{k}{50} + \left(\frac{7}{10} + \frac{k}{500}\right)\gamma; \tag{3.43}$$

$$\mathbb{V}ar[X_{kl}] = \frac{\pi^2}{6}\exp\left\{2\left(\frac{7}{10} + \frac{k}{500}\right)\right\}; \tag{3.44}$$

where γ is the Euler-Mascheroni constant.

Let $(M_k : 1 \leq k \leq K)$ be the sequence of block maxima, where

$$M_k = \max\{X_{k1}, \ldots, X_{kL}\}. \tag{3.45}$$

In a non-stationary context, the distribution quantiles are evolving because the probability may change according to the different indexes. For instance, the quantiles of the distribution of Eq. (3.41) are function of index k. Katz et al. (2002) called these time evolving quantiles "effective return levels". Let $q_{k,T}$ be the quantile of order $1-1/T$ for the random variable X_{kl} for any $1 \leq l \leq L$. In term of years, $q_{k,T}$ is the effective T-year return level for the index k. Let $\hat{q}^{std}_{k,T}$ be the estimation of $q_{k,T}$ obtained with the standard non-stationary block maxima model and let $\hat{q}^{pre}_{k,T}$ be the estimation obtained with the proposed preprocessing model. The relative efficiency of $\hat{q}^{pre}_{k,T}$ over $\hat{q}^{std}_{k,T}$ can be defined as follows:

$$\text{eff}(\hat{q}^{pre}_{k,T}, \hat{q}^{std}_{k,T}) = \frac{\mathbb{E}[(\hat{q}^{std}_{k,T} - q_{k,T})^2]}{\mathbb{E}[(\hat{q}^{pre}_{k,T} - q_{k,T})^2]}. \tag{3.46}$$

If $\text{eff}(\hat{q}^{pre}_{k,T}, \hat{q}^{std}_{k,T}) > 1$, then the estimation obtained with the preprocessing model is preferable. Note that a relative efficiency value is defined for every $1 \leq k \leq K$. Let us define a global measure of relative efficiency for a sequence:

$$r_T = \frac{1}{K}\sum_{k=1}^{K} \text{eff}(\hat{q}^{pre}_{k,T}, \hat{q}^{std}_{k,T}). \tag{3.47}$$

Therefore, if $r_T > 1$, then the estimations obtained with the preprocessing model are in general preferable.

3.5.1 Standard non-stationary block maxima model

The non-stationary model for the block maxima sequence $(M_k : 1 \leq k \leq K)$ is the following:

$$M_k \overset{\mathcal{L}}{\approx} \mathcal{G}EV(\mu_k, \sigma_k, \xi_k); \tag{3.48}$$

where

$$\begin{aligned}
\mu_k &= \mu_0 + \mu_1 k; \\
\sigma_k &= \exp(\sigma_0 + \sigma_1 k); \\
\xi_k &= \xi.
\end{aligned}$$

Parameter estimates were obtained with maximum likelihood.

3.5.2 Preprocessing model

Let $\bar{X}_k$ be the sample mean of block k:

$$\bar{X}_k = \frac{1}{L}\sum_{l=1}^{L} X_{kl}. \tag{3.49}$$

The considered model for the block means is the following:

$$\bar{X}_k = \nu_0 + \nu_1 k + \epsilon_k, 1 \leq k \leq K. \tag{3.50}$$

Let S_k^2 be the sample variance of block k:

$$S_k^2 = \frac{1}{L-1}\sum_{l=1}^{L} \left(Y_{kl} - \nu_0 - \nu_1 k\right)^2. \tag{3.51}$$

The nonlinear regression model considered for the sequence of sample variances is the following:

$$S_k^2 = \exp\left(\tau_0 + \tau_1 k\right) + \varepsilon_k, 1 \leq k \leq K. \tag{3.52}$$

Let M'_k be the standardized maximum of block k:

$$M'_k = \frac{M_k - \nu_0 - \nu_1 k}{\exp\left(\frac{\tau_0 + \tau_1 k}{2}\right)}. \tag{3.53}$$

As proposed in Section 3.4, the standardized maxima are assumed identically distributed with the following distribution:

$$M'_k \overset{\mathcal{L}}{\approx} \mathcal{GEV}(\mu', \sigma', \xi'), 1 \leq k \leq K. \tag{3.54}$$

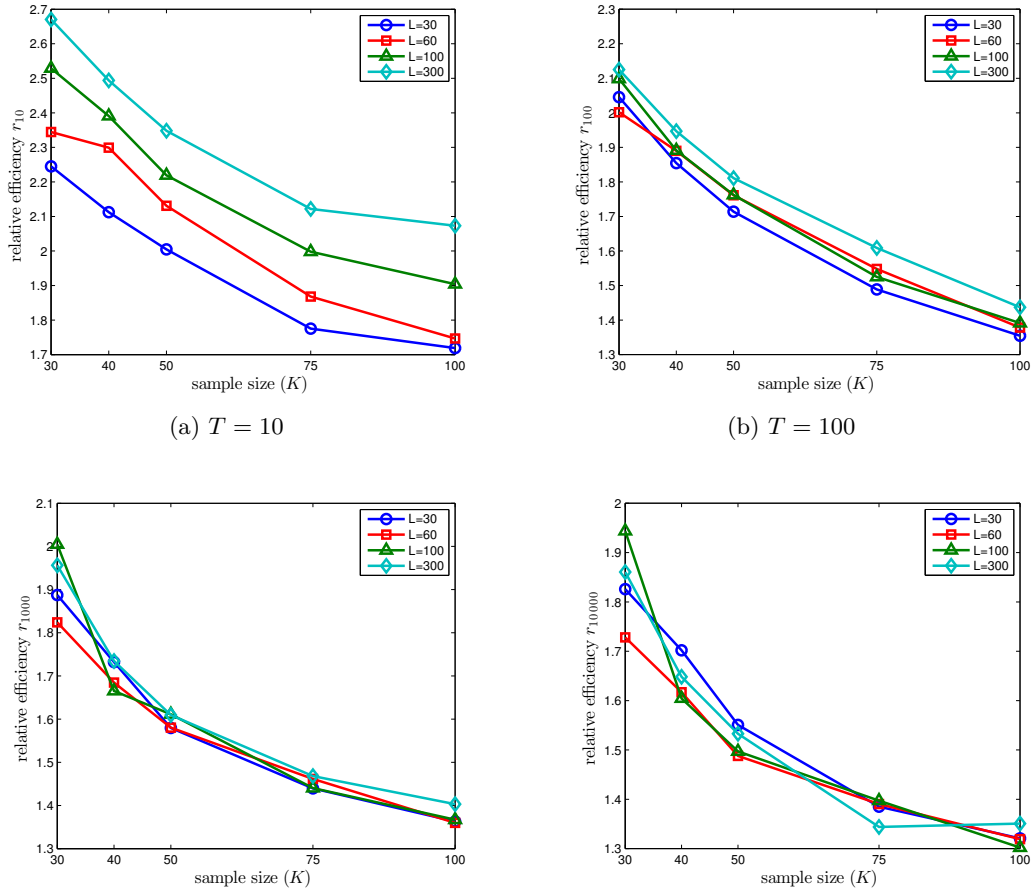
Therefore, the model for the unstandardized maxima can be written as follows:

$$M_k \overset{\mathcal{L}}{\approx} \mathcal{GEV}\left\{\mu' \exp\left(\frac{\tau_0 + \tau_1 k}{2}\right) + \nu_0 + \nu_1 k, \sigma' \exp\left(\frac{\tau_0 + \tau_1 k}{2}\right), \xi'\right\}. \tag{3.55}$$

Estimations of (μ', σ', ξ') were obtained using Eq. (3.54) and the maximum likelihood method. The least square method was used to estimate (ν_0, ν_1) of Eq. (3.50) and to estimate (τ_0, τ_1) of Eq. (3.52).

3.5.3 Results

For several combinations of K and L and for 10,000 replicates for each of them, the relative efficiency r_T for several return periods T are shown in Figure 3.1. For the given simulation design, the return levels extracted from the model based on standardization are more precise than those obtained with the standard approach since $r_T > 1$ for every combination of K, L and T. For $T = 20$ and $T = 100$, the relative efficiency increases as the block size L increases.



(a) $T = 10$ (b) $T = 100$

(c) $T = 1000$ (d) $T = 10000$

Figure 3.1 – Relative efficiency of the proposed T-year return level estimator based on the preprocessing model over the estimator based on the standard non-stationary model.

This is mostly due to the improved estimations of both ν_k and τ_k in the regression analysis. As the sample size becomes larger, this advantage seems to decreases as the estimation of the non-stationarity with the maxima series becomes more precise. Nevertheless, the estimators obtained with the preprocessing approach are still more efficient.

For large return periods, the estimation variance of ξ has a larger influence on the corresponding estimated return level variance. As a consequence, the block size L seems not to be a factor in the relative efficiency for large return periods (see Figure 3.1). In other words, the improved estimations of ν_k and τ_k as the block size increases have a small impact on the efficiency because these estimations are used in the location and the scale parameters of the GEV distribution. Nevertheless, the return level estimator based on our preprocessing model is still more effective. It is probably due to the better estimation of ξ when the functions for μ_k and σ_k are based on the standardization approach.

3.5.4 Discussion

In the given simulation design, the random variables $(X_{kl} : 1 \leq k \leq K, 1 \leq l \leq L)$ were independent. If short-range temporal dependence was introduced among the sequences $(X_{kl} : 1 \leq l \leq L), 1 \leq k \leq K$, as defined in Assumption (A2), the relative efficiency of return level estimations would be reduced. Indeed, the *effective* block size L available for estimating both ν_k and τ_k would be deceased. Since the relative efficiencies are larger than one, even for the smallest block size $L = 30$, we believe that the efficiencies should remain larger than one. Therefore, for the given simulation design, a short-range temporal dependence should not change the conclusion that the return level estimations based on the standardization approach are more precise.

Criticisms on preprocessing come mostly from the argument stating that information in the bulk distribution has little value in the tails. The example often given concerns a mixture of distributions where the extremes come from another distribution than the bulk of the data. This scenario is studied in Appendix 3.B and the results show that the proposed preprocessing model is robust enough to this situation.

In order to keep the ongoing section parsimonious, we only studied the non-stationary block maxima approach. The results should be similar for the non-stationary excesses over an appropriate large threshold as the methodology is also based on Theorem 3.1 and Proposition 3.1.

3.6 Case studies

In the present section, the proposed methodology is applied to two non-stationary environmental dataset: the daily sea levels at San Francisco (USA) and the daily winter precipitation at Manjimup (Australia). These datasets were selected because of their long and high quality time series and because they have been previously studied using standard non-stationary extreme value models. Except for the block means and the block standard deviations ν_k and τ_k, the statistical models have been fitted using the Bayesian procedure described by Renard et al. (2013). In particular, non-informative prior distributions have been used for both the GEV and the GP parameters and a sample from the posterior distribution is obtained by performing a MCMC algorithm.

3.6.1 Sea levels at San Francisco

The present version of the dataset consists in a time series of the daily sea level maximum in meters at San Francisco (United States) from August 1897 to January 2015. The data are available from the Sea Level Center website (`http://uhslc.soest.hawaii.edu`) of the University of Hawaii. The time series exhibits seasonality along the year. In San Francisco, the sea level is higher during winter mostly due to the astronomical tide (Méndez et al., 2007).

We therefore restrict the analysis on the sea levels from September 1st to March 31th. The sea level maximum of day l of year k is denoted by X_{kl}, where $1 \leq l \leq 211$ and $1 \leq k \leq 118$. The index $l = 1$ stands for September 1st and the index $k = 1$ stands for the year 1897. Figure 3.2 shows the daily winter sea levels at San Francisco as function of the date. From this last figure, a global rise in the sea levels is observed. Douglas (1991) provided physical explanations for this rise. The goal of the present section is to estimate the *effective* T-year return levels for the observed period with the proposed preprocessing model and to compare it with the estimations obtained with the standard non-stationary block maxima model.

The standard non-stationary model for the seasonal sea level maxima is the following:

$$M_k \overset{\mathcal{L}}{\approx} \mathcal{G}EV\left(\mu_0 + \mu_1 k, \sigma, \xi\right), 1 \leq k \leq 118; \tag{3.56}$$

Note that the usual function $\sigma_k = \exp\left(\sigma_0 + \sigma_1 k\right)$ for modeling the shape parameter as function of time is not significative in this case.

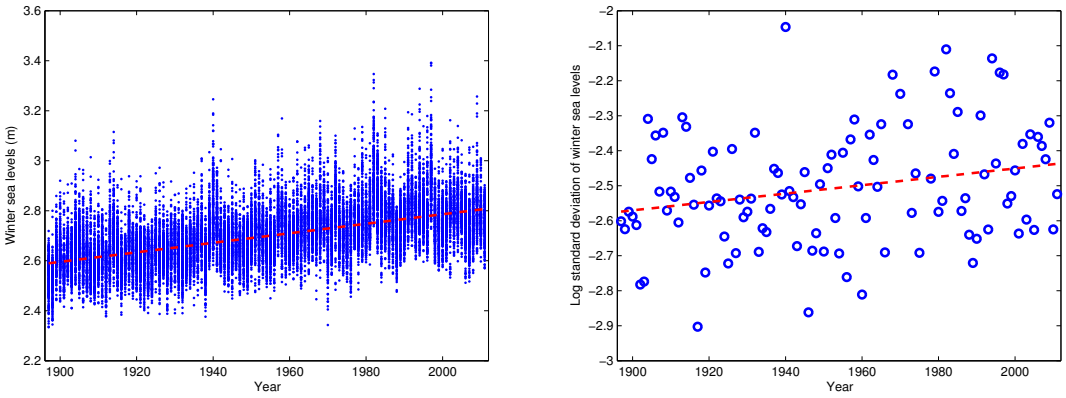
For the application of our proposed preprocessing model, both the seasonal means and the seasonal standard deviations has to be modeled as functions of year index k. The following regression functions were used to model respectively the means and the standard deviations:

$$\begin{aligned} \nu_k &= \nu_0 + \nu_1 k; \\ \tau_k &= \exp\left(\tau_0 + \tau_1 k\right). \end{aligned} \tag{3.57}$$

Estimation of the parameters in Eq. (3.57) were perfomed with regression analysis as described in Section 3.4. The sea levels were autocorrelated along three days. Therefore, we only retained one of every three observations for estimating the seasonal sample variance (see Eq. 3.39). It corresponds to a simple filtering method for obtaining an approximate independent sample. Both regression functions of Eq. (3.57) were significative. Figure 3.2 shows the fits of the following estimated regression functions:

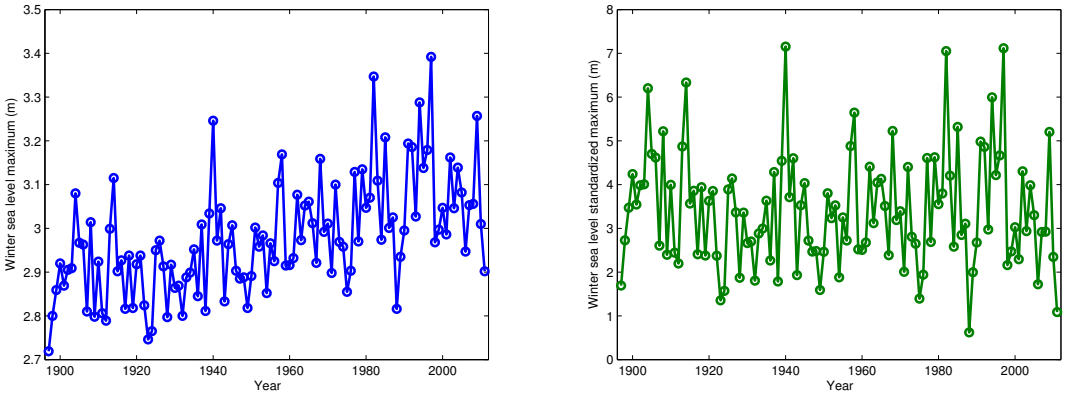
$$\begin{aligned} \hat{\nu}_k &= 2.59 + 2.02 \times 10^{-3} k; \\ \hat{\tau}_k &= \exp\left(-2.57 + 1.22 \times 10^{-3} k\right). \end{aligned} \tag{3.58}$$

The block mean estimates, $(\hat{\nu}_k : 1 \leq k \leq K)$ and the block standard deviation estimates $(\hat{\tau}_k : 1 \leq k \leq K)$ were used to standardize the original winter sea levels according to Eq. (3.25). Figure 3.3(a) shows the original maxima series Figure 3.3(b) and the standardized maxima series. The standardized maxima series seems to be stationary, or sufficiently close to, for applying the standard stationary block maxima model. This assumption was validated with a likelihood ratio test. The model with no trend in the location and the scale parameter was not rejected at the 5% level of significance. Therefore, the standard stationary model for block maxima has been fitted to the standardized maxima. Note that the stationarity assumption could also be verified with the non-parametric approach proposed by Naveau et al. (2014), which was developed specifically for extreme data.



(a) Seasonal mean ν_k (b) Seasonal standard deviation τ_k

Figure 3.2 – Characteristics of the winter sea levels at San Francisco (USA).



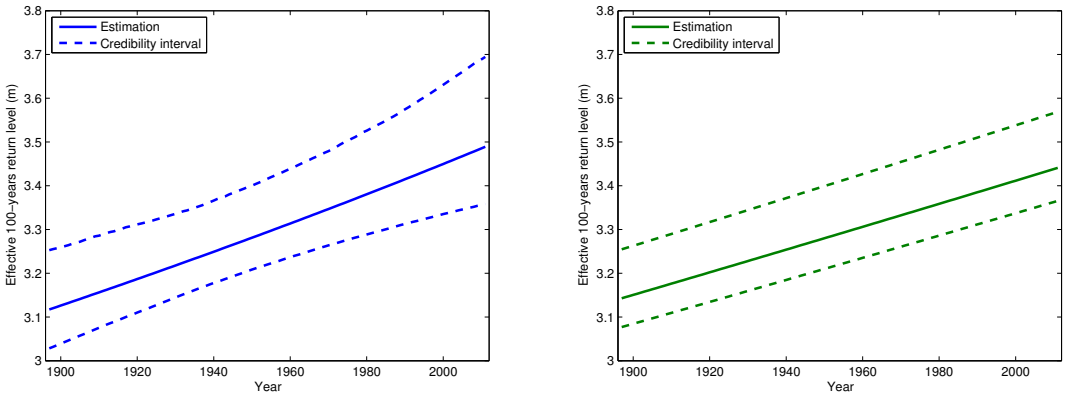
(a) Unstandardized maxima (b) Standardized maxima

Figure 3.3 – Maxima and standardized maxima of the winter sea levels at San Francisco (USA).

Table 3.1 compiles the estimated parameters of the standard non-stationary model and the preprocessing model. Note that the credibility interval of the shape parameter is thinner when it is estimated with the preprocessing model. Recalling that the standardization does not influence the shape parameter, a direct comparison of theses estimations is possible. Figures 3.4 illustrates the effective 100-year return level estimations obtained using both approaches. The 100-year effective return level estimations are quite similar, but the credibility interval width associated with the preprocessing model is reduced compared to the standard model. It is consistent with the results of our simulation study that our preprocessing approach gives more accurate estimations. It shows that our proposed preprocessing model for block maxima enhances the estimation precision of the sea level extremes.

3.6.2 Daily precipitation at Manjimup

Li et al. (2005) investigated the decrease in winter rainfall extremes in south west Australia



(a) Standard approach

(b) Preprocessing approach

Figure 3.4 – 100-year return level for the winter sea levels at San Francisco (USA).

Table 3.1 – Parameter estimations for the standard model and the proposed preprocessing model.

Standard model		
Parameter	Estimation	Credibility interval (95%)
μ_0	2.82	$(2.78, 2.85)$
μ_1	0.0020	$(0.0014, 0.0025)$
σ	1.10	$(1.08, 1.11)$
ξ	-0.072	$(-0.188, 0.059)$

Preprocessing model		
Parameter	Estimation	Credibility interval (95%)
μ	2.85	$(2.63, 3.09)$
σ	1.14	$(1.00, 1.30)$
ξ	-0.084	$(-0.198, 0.042)$

for five weather stations. In their paper, Li et al. (2005) put the emphasis on the weather station Wilgarrup, numbered 9619, located at Manjimup. The data is available online from the Australian Government Bureau of Meteorology website (`http://www.bom.gov.au/climate/data-services/`). Katz (2013) also studied this dataset on the same period. Figure 3.5 illustrates the time series of winter (May-October) daily rainfall from 1930 to 2004 at Manjimup. Note that the observations beyond 2004 were ignored in this paper in order to directly compare our results with those from Li et al. (2005) and Katz (2013). The precipitation amount in mm at day l of year k is denoted by X_{kl}, where $1 \leq l \leq 183$, $1 \leq k \leq 74$. The index $l = 1$ stands for May 1^{st} and $k = 1$ stands for October 31^{th}.

Using the Mann-Kendall-Pettitt stationarity test, Li et al. (2005) detected a change-point in the annual maximum winter precipitation series at the year 1965, which consists in an abrupt change in the annual maxima distribution prior to 1966. According to Li et al. (2005), the sub-series of both sub-periods can be assumed stationary. They thus fitted a stationary

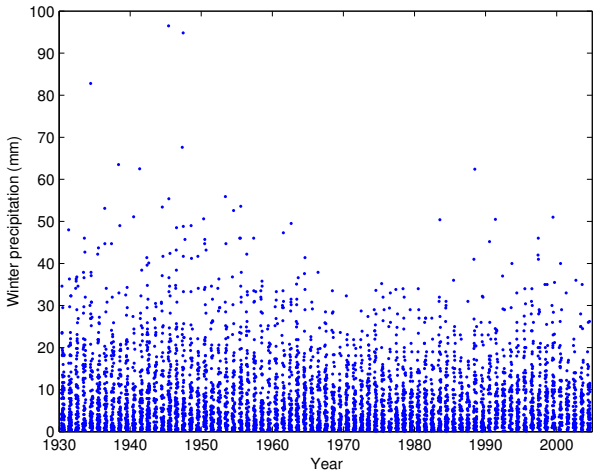


Figure 3.5 – Winter precipitation series at Manjimup (Australia).

POT model independently for both sub-periods. They defined a threshold of 30 mm for the first sub-period (prior to 1965) and a threshold of 22 mm for the second sub-period. Katz (2013) considered a single non-stationary POT model for the whole period. He considered many parametrizations for the threshold and the scale parameter but the shape parameter was assumed constant for the whole period. According to a likelihood ratio test procedure, the evidence was also in favor of the change point model for the threshold and the scale parameters. Note that Katz (2013) used the threshold defined by Li et al. (2005) for the change point parametrizations. The difference with the previous approach is that Katz (2013) considered that both sub-period have the same shape parameter while Li et al. (2005) fitted two POT models independently, thus permitting a different shape parameter for both sub-periods.

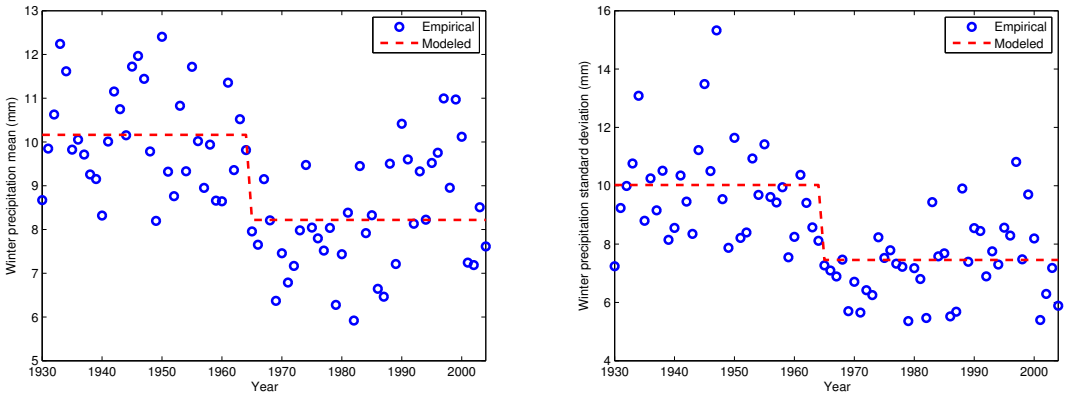
The proposed preprocessing model was also fitted on this dataset. The first step is to model both the winter means and the winter standard deviations as functions of year index k. Several models were considered for the means: the constant, the linear, the change-point, the quadratic, the exponential and the Weibull. The best according to the Mean Squared Error (MSE) was in favor of a change-point in 1963:

$$\hat{\nu}_k = \begin{cases} 10.2 & \text{if } k \leq 1963; \\ 8.2 & \text{if } k > 1963. \end{cases} \tag{3.59}$$

For the standard deviations, the change-point model was also the best among the constant and the linear:

$$\hat{\tau}_k = \begin{cases} 10.0 & \text{if } k \leq 1963; \\ 7.5 & \text{if } k > 1963. \end{cases} \tag{3.60}$$

The location of the change points were forced to be the same for the mean sequence and the standard deviation sequence. Figure 3.6 shows the corresponding regression functions along with the data at Manjimup. Using the whole dataset, the change-point location differs



(a) Seasonal mean ν_k (b) Seasonal tandard deviation τ_k

Figure 3.6 – Winter precipitation at Manjimup (Australia).

slightly from the one estimated using the maxima series: 1965 for Li et al. (2005) while it is estimated at 1963 with the complete seasonal dataset.

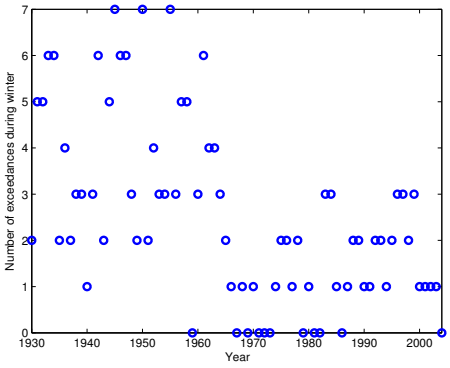
The threshold for the whole period of the standardized data have been set to $u = 1.98$, which corresponds to the threshold of 30 mm defined by Li et al. (2005) for the first sub-period. Figure 3.7 shows the threshold exceedance rate as well as the excess magnitudes as functions of the year for both the original series and the standardized series. It is particularly evident that non-stationarity has been removed for the number of threshold exceedances per year. The stationarity assumption of the standardized excesses has been verified with a likelihood ratio test. A GP distribution for with a linear trend for the scale parameter was compared with the alternative model with no trend. The model with no trend was not rejected at the 5% level. Therefore, it is reasonable to apply the stationary POT model on the standardized series.

Table 3.2 compiles the parameter estimations for the standard models and for the preprocessing model. As mentioned in Section 3.4.2, the threshold for the standardized series can be traced back to the unstandardized data. This leads to the following varying threshold into the original space:

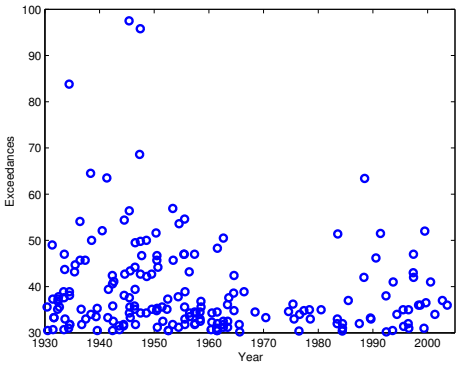
$$u_k = \begin{cases} 30.0 & \text{if } k \leq 1963; \\ 23.0 & \text{if } k > 1963. \end{cases} \tag{3.61}$$

Recalling that we set the invariant threshold u for the standardized series as the corresponding value of the 30 mm threshold defined in Li et al. (2005) and Katz (2013) for the first sub-period, the inverse transformation of the invariant threshold is very close to the 22 mm threshold defined by these authors for the second sub-period. Therefore, our threshold definition using preprocessing is consistent with the one used by Li et al. (2005) and Katz (2013).

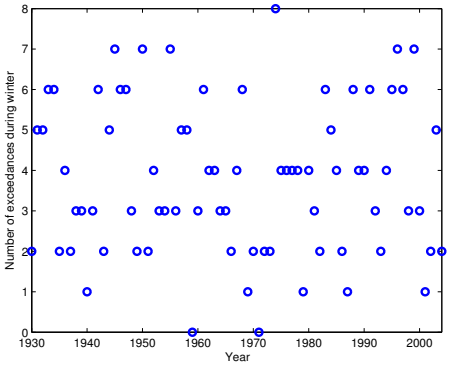
The different models are compared on the basis of the 100-year return level estimations. Figure 3.8 shows the effective return levels for the model of Li et al. (2005), the model of



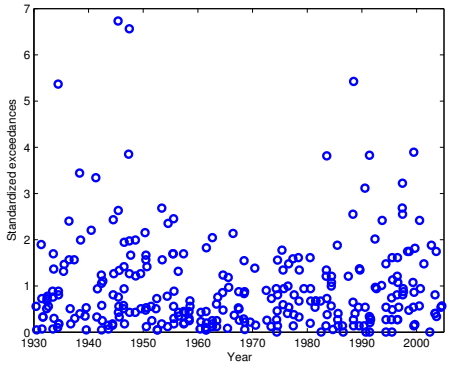
(a) Number of exceedances above a 30mm threshold



(b) Exceedances above a 30mm threshold



(c) Number of exceedances above a standardized threshold



(d) Exceedances above a standardized threshold

Figure 3.7 – Winter intense precipitation at Manjimup (Australia).

Katz (2013) and the proposed preprocessing model. The resulting estimations obtained with the approach of Li et al. (2005) are quite different from the other methodologies. It is mostly due to the stationary shape ξ assumption in the approach of Katz (2013) and in the proposed preprocessing approach. While the threshold and the scale parameter are both varying in the original precipitation space, the shape parameter is constant so it is estimated with the entire series of exceedances. In the approach of Li et al. (2005), the two different stationary peaks-over-threshold models were independently adjusted for both sub-periods, including two independent estimations for the shape parameters. The estimation variance for this parameter is large and leads to a large uncertainty of the return level estimations. As shown in Table 3.2, the credibility intervals of the shape parameters of both period overlaps, which suggests a single value for the whole period. The model of Katz (2013) and our preprocessing model assumed a single parameter for the whole period. Therefor, the estimation of the shape parameter is based on a larger sample size, which reduces the uncertainty in return level estimation.

The estimated 100-year return levels are very similar between the model developed by Katz

Table 3.2 – Parameter estimations for the standard models and the proposed preprocessing model.

Standard model of Li et al. (2005)			
Time period	Parameter	Estimation	Credibility interval (95%)
$[1930, 1965]$	u	30 mm	-
	σ	8.7	$(6.9, 11.0)$
	ξ	0.12	$(-0.03, 0.30)$
$[1966, 2004]$	u	22 mm	-
	σ	8.0	$(6.4, 9.8)$
	ξ	-0.06	$(-0.17, 0.07)$

Standard model of Katz (2013)			
Time period	Parameter	Estimation	Credibility interval (95%)
$[1930, 1965]$	u	30 mm	-
	σ	8.2	$(6.9, 9.6)$
$[1966, 2004]$	u	22 mm	-
	σ	7.2	$(5.9, 8.5)$
$[1930, 1965]$	ξ	0.08	$(-0.02, 0.19)$

Preprocessing model		
Parameter	Estimation	Credibility interval (95%)
u	1.98	-
σ	0.91	$(0.77, 1.05)$
ξ	0.05	$(-0.05, 0.19)$

(2013) and the proposed preprocessing model. There is a small difference in the credibility interval width. The credibility intervals are a bit thinner with the preprocessing approach, which is consistent with the previous simulation study of Section 3.5. In summary, the preprocessing approach enhances the analyis of the extreme precipitations at Manjimup.

3.7 Conclusion

In the present paper, a methodology was developed for the analysis of non-stationary extreme values that reconciles the preprocessing approach with the standard approaches. The proposed methodology used the standard non-stationary models but but with the parameters estimated with a preprocessing approach. As argued by Eastoe and Tawn (2009), the non-stationarity in extreme values is therefore better captured using the whole dataset because the physical processes involved for non-stationarity are often intricately tied up with the general generating mechanism. We proposed a data transformation based on the whole dataset that removes the non-stationarity in the first two moments. It was shown that this transformation is sufficient to meet the usual simplication stated in the standard non-stationary model, say an invariant shape parameter.

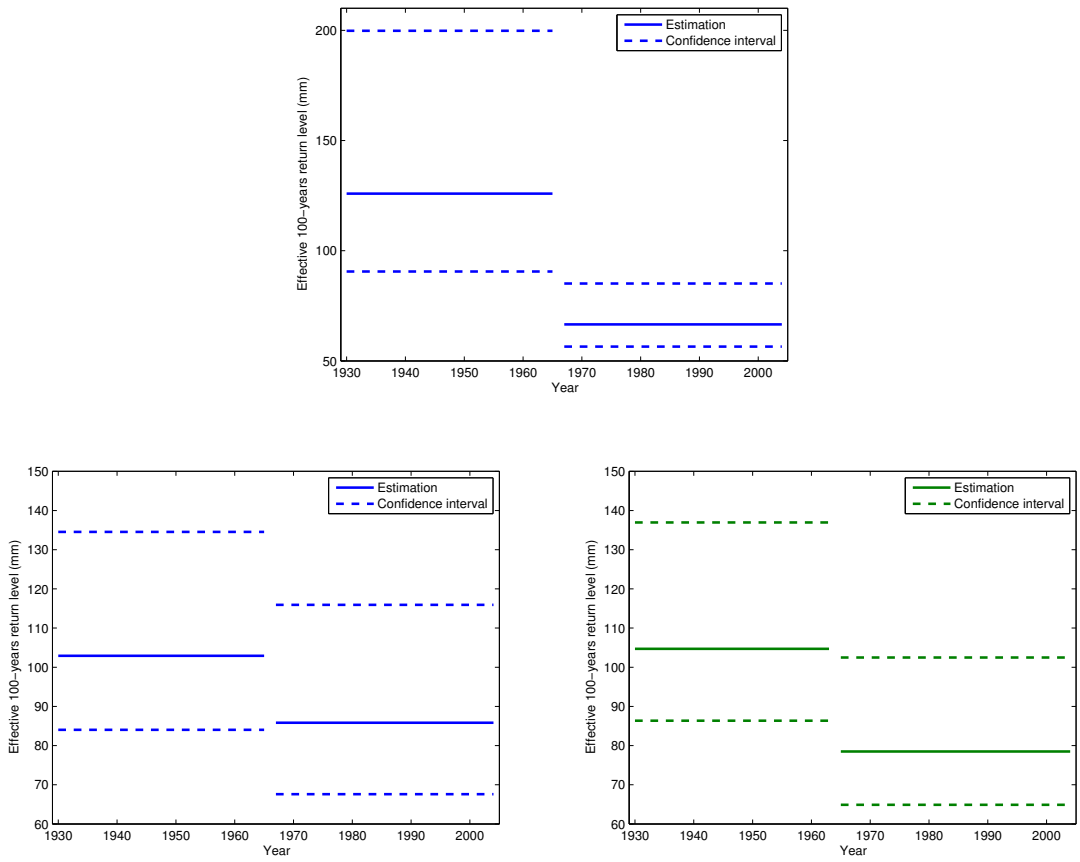


Figure 3.8 – 100-year estimated return level for the daily winter precipitation at Manjimup (Australia).

We have shown that the statistical model based on the proposed preprocessing method for studying non-stationary extreme values gives more accurate T-year return level estimations. From the simulation study of Section 3.5, it has been observed that the return level estimators were more efficient than those based on the standard non-stationary models. Our proposed methodology was also applied on two real datasets that have been already studied by the standard models in the litterature. The comparison showed no major difference in the results, but the credibility intervals according to our preprocessing approach were thinner. For both case studies, the time series were long, which is quite uncommon in an environmental framework. As it has been shown in the simulation study, the gain in precision in return level estimation increases as the sample size decreases. Thus for common size datasets, usually under 50 years of observations, our method based on preprocessing should be even more efficient than the standard approaches. In summary, if an investigator is willing to assume an invariant shape parameter, then our approach may be more accurate for studying non-stationary extreme values.

Future work may be necessary to study the impact of nonlinear regression analysis that we

used to estimate the time-varying expectation and variance in Section 3.4. It was a generic approach introduced because it was suitable for the majority of cases. However, more appropriate formulations may exist for some specific problems. Also, it would be interesting to include additionnal covariates for explaining the non-stationarity and see how our preprocessing approach performs compared with the standard approaches.

References

Bates, D. M. and Watts, D. G. (1988). *Nonlinear regression: iterative estimation and linear approximations*. Wiley.

Chatfield, C. (2004). *The analysis of time series – an introduction*. Chapman & Hall, Boca Raton, USA, 6 edition.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer, London, United Kingdom.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds (with Discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(3):393–442.

Douglas, B. C. (1991). Global sea level rise. *Journal of Geophysical Research: Oceans (1978–2012)*, 96(C4):6981–6992.

Eastoe, E. F. and Tawn, J. A. (2009). Modelling non-stationary extremes with application to surface level ozone. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 58(1):25–45.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, page 180. Cambridge University Press.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of mathematics*, 44(3):423–453.

Katz, R. W. (2013). Statistical methods for nonstationary extremes. In *Extremes in a changing climate*, chapter 2, pages 15–37. Springer, New York, USA.

Katz, R. W., Parlange, M. B., and Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in Water Resources*, 25(8):1287–1304.

Leadbetter, M. R., Lindgren, G., and Rootzén, H. (1983). *Extremes and related properties of random sequences and processes*. Springer, New York, USA.

Li, Y., Cai, W., and Campbell, E. P. (2005). Statistical modeling of extreme rainfall in southwest Western Australia. *Journal of Climate*, 18(6):852–863.

Méndez, F. J., Menéndez, M., Luceño, A., and Losada, I. J. (2007). Analyzing monthly extreme sea levels with a time-dependent GEV model. *Journal of Atmospheric and Oceanic Technology*, 24(5):894–911.

Naveau, P., Guillou, A., and Rietsch, T. (2014). A non-parametric entropy-based approach to detect changes in climate extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(5):861–884.

Pickands III, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119–131.

Renard, B., Sun, X., and Lang, M. (2013). Bayesian methods for non-stationary extreme value analysis. In *Extremes in a changing climate*, chapter 3, pages 39–95. New York, USA.

3.A Demonstration of Proposition 1

To prove Proposition 3.1, we need the following proposition :

Proposition 3.2. *Suppose that*

$$\frac{M_l - b_l}{a_l} \xrightarrow{\mathcal{L}} \mathcal{G}EV(0,1,\xi), \text{ when } l \to \infty; \tag{3.62}$$

then

$$\frac{M_l - b_l^*}{a_l^*} \xrightarrow{\mathcal{L}} \mathcal{G}EV(0,1,\xi^*), \text{ when } l \to \infty; \tag{3.63}$$

if and only if

$$\lim_{l\to\infty} \frac{a_l}{a_l^*} = 1; \qquad \lim_{l\to\infty} \frac{b_l - b_l^*}{a_l} = 0; \qquad \xi = \xi^*. \tag{3.64}$$

The proof is a direct application of Khintchine's theorem (see Leadbetter et al., 1983, p.7).

Proof of Proposition 3.1. Since the K sequences $(X'_{kl} : l \geq 1)$ have the same distribution, so do the K sequences $(M'_{kl} : l \geq 1)$. Therefore, we have for $1 \leq k \leq K$

$$\frac{M'_{1l} - b'_{kl}}{a'_{kl}} \xrightarrow{\mathcal{L}} \mathcal{G}EV(0,1,\xi_k). \tag{3.65}$$

Thus,

$$\frac{M'_{1l} - b'_{1l}}{a'_{1l}} \xrightarrow{\mathcal{L}} \mathcal{G}EV(0,1,\xi_1) \qquad \text{and} \qquad \frac{M'_{1l} - b'_{2l}}{a'_{2l}} \xrightarrow{\mathcal{L}} \mathcal{G}EV(0,1,\xi_2). \tag{3.66}$$

The application of Proposition 3.2 with $M_l = M'_{1l}$, $a_l = a'_{1l}$, $b_l = b'_{1l}$, $a_l^* = a'_{2l}$ and $b_l^* = b'_{2l}$ gives

$$\lim_{l\to\infty} \frac{a'_{1l}}{a'_{2l}} = 1; \qquad \lim_{l\to\infty} \frac{b'_{1l} - b'_{2l}}{a'_{1l}} = 0; \qquad \xi_1 = \xi_2. \tag{3.67}$$

Applying $(K-2)$ more times Proposition 3.2 gives the result.

□

3.B Supplementary simulations

Let K and L be some positive integers. Consider the rectangular array of independent random variables $(X_{kl} : 1 \leq k \leq K, 1 \leq l \leq L)$, with

$$X_{kl} \sim (1-\lambda) \times \mathcal{E}xp(1) + \lambda \times \mathcal{G}P\left(\sigma_k, \xi\right); \tag{3.68}$$

where $0 < \lambda < 1$, $\sigma_k > 0$ and $\xi < 1/2$.

For a given L, the effective T-year return level $q_{k,T}$ is the solution of the following equation:

$$(1-\lambda)\left(1-e^{-q_{k,T}}\right) + \lambda\left\{1-\left(1+\frac{\xi}{\sigma_k}q_{k,T}\right)^{-1/\xi}\right\} = \left(1-\frac{1}{T}\right)^{1/L}. \tag{3.69}$$

Let $\lambda = 1/20$, $\sigma_k = 8/5 + k/125$ and $\xi = 1/5$. With this set of parameters, the expectation for $k = 0$ and $k = 100$ are respectively

$$\begin{array}{rl} \mathbb{E}(X_{0,l}) & = \frac{19}{20} + \frac{1}{20}\frac{8/5}{1-1/5} = \frac{21}{20}, 1 \leq l \leq L; \\ \mathbb{E}(X_{100,l}) & = \frac{19}{20} + \frac{1}{20}\frac{8/5+100/125}{1-1/5} = \frac{22}{20}, 1 \leq l \leq L. \end{array} \tag{3.70}$$

The effective 100-year return level for $k = 0$ and for $k = 100$ are as respectively

$$\begin{array}{rl} q_{0,100} & = 26.5; \\ q_{100,100} & = 39.8. \end{array} \tag{3.71}$$

Note that the bulk of the distribution mostly comes from the exponential distribution while the extreme values mostly comes from the heavy tailed generalized Pareto distribution. Therefore, the extreme values are more affected by non-stationarity than the mean values. This model was used in a simulation design to verify if the proposed model based on standardization is still suitable in this scenario, say where non-stationarity of extreme values are not intricately tied up with mean values.

The standard non-stationary model is the following:

$$M_k \overset{\mathcal{L}}{\approx} \mathcal{G}EV(\mu_0 + \mu_1 k, \sigma_0 + \sigma_1 k, \xi), 1 \leq k \leq K. \tag{3.72}$$

The preprocessing model is the following:

$$M_k \overset{\mathcal{L}}{\approx} \mathcal{G}EV(\tau_k \mu + \nu_k, \tau_k \sigma, \xi), 1 \leq k \leq K; \tag{3.73}$$

where

$$\begin{array}{rl} \nu_k & = \nu_0 + \nu_1 k; \\ \tau_k & = \tau_0 + \tau_1 k; \\ M'_k & \overset{\mathcal{L}}{\approx} \mathcal{G}EV(\mu, \sigma, \xi). \end{array} \tag{3.74}$$

For $L = 100$, the relative efficiency is larger than one for any return period T. For instance, $r_{100} = 1.18$. Therefore, the T-year return level estimate obtained with the preprocessing approach is more efficient than the estimate obtained with the standard model. The reason why the preprocessing model is suitable lies on the fact that the increase of extreme values also grows the variance, which is taken into account in the standardization.

Chapitre 4

A non-stationary and regional model for annual precipitation maxima simulated by a climate model ; with an application for estimating the 100-year return levels in North America.

Jonathan Jalbert, Anne-Catherine Favre, Claude Bélisle et Jean-François Angers.

Cet article est en préparation pour soumission.

Résumé

Les précipitations extrêmes constituent un facteur important dans les événements d'inondations. Le risque d'inondation devrait s'accroître avec les changements climatiques puisqu'il est attendu que l'occurrence et l'intensité des précipitations extrêmes augmenteront. L'étude de l'évolution des précipitations extrêmes devient un enjeu important pour la sécurité du public et pour la pérennité des infrastructures. Les modèles numériques de climat sont les seuls outils disponibles pour obtenir des prédictions quantitatives du climat futur. La plupart

du temps, la théorie des valeurs extrêmes est utilisée pour estimer les quantiles des précipitations simulées correspondant à différentes périodes de retour. De façon générale, la variance d'estimation de ces quantiles est considérable en raison notamment du nombre limité de précipitations extrêmes disponibles. L'incertitude sur l'estimation peut jouer un rôle déterminant dans l'élaboration des stratégies de gestion du risque. Par conséquent, un modèle statistique pour les précipitations extrêmes simulées par un modèle de climat est développé dans cet article dans le but de réduire la variance d'estimation. Pour ce faire, le modèle développé est spécialement adapté aux données générées par un modèle de climat. Il prend en compte la non-stationnarité des séries ainsi que la cohérence spatiale des précipitations. En particulier, la non-stationnarité est retirée des séries par une méthode de prétraitement qui exploite l'information contenue dans les séries de précipitations journalières entières. Ensuite, la cohérence spatiale est modélisée par un modèle hiérarchique bayésien employant des champs de Markov gaussiens intrinsèques comme lois *a priori*. Le modèle est utilisé pour estimer les quantiles correspondant à la période de retour centennale en Amérique du Nord à partir d'une simulation générée par le Modèle régional canadien du climat. Les résultats montrent une importante réduction de la variance d'estimation.

Abstract

Extreme precipitations play a major role in flooding events and their occurrence as well as their intensity are expected to increase. It is therefore important to anticipate the impacts of such an increase to ensure the public safety and the infrastructure's sustainability. Since climate models are the only tools for providing quantitative projections of precipitation, flood risk management for the future climate may be based on their simulation. Most of the time, extreme value theory is used to estimate extreme precipitations from a climate simulation, such as the T-year return levels. The variance of the estimations are generally large notably because the sample size of the maxima series is small. Such variance could have a significant impact for flood risk management. It is therefore relevant to reduce the estimation variance of simulated return levels. For this purpose, the aim of this paper is to develop a non-stationary and regional statistical model specifically suited for climate models that estimates precipitation extremes. At first, the non-stationarity is removed by a preprocessing approach. Thereafter, the spatial correlation is modeled by a Bayesian hierarchical model including an intrinsic Gaussian Markov random field. The model has been used to estimate the 100-year return levels over North America from a simulation by the Canadian Regional Climate Model. The results show a large estimation variance reduction when using the regional model.

4.1 Introduction

According to climate change, precipitation are expected to evolve in the coming decades. It is therefore relevant to investigate such changes in precipitation extremes because they play a major role in flooding events. Although general rainfall generating mechanisms are well understood, the evolution of rainfall in a non-stationary climate is intractable. Indeed, the non-linear and chaotic nature of the equations describing rainfall mechanisms prevent simple predictions of precipitation. Because of that, numerical simulations from climate models may be the best method for providing quantitative predictions of precipitation in non-stationary climates (IPCC, 2012). Regional Climate Models (RCMs) are potentially better suited for studying the impacts of climate change on precipitation than Global Climate Models (GCMs) because the spatial resolution of RCMs (from 50 down to 10 km) allows some important features in the formation of precipitation to be well represented. The spatial-scale characteristics of precipitation are reasonably represented in RCMs at a length scale of roughly 100 km (Rasmussen et al., 2012).

In flood risk management, conceptual designs of infrastructures are flood resistant up to the level that arises on average once every T years; such level is referred to as T-year return level. Estimation of return levels is generally done with the methodology called frequency analysis (see Meylan et al., 2012, for an introduction), where the Extreme Value Theory (EVT, see Coles, 2001, for an introduction) is most of the time applied. The EVT framework provides under some conditions the asymptotic distribution for the maximum of a sequence of independent and identically distributed random variables. This distribution is called the Generalized Extreme Value (GEV) distribution and its cumulative distribution function is as follows:

$$\mathcal{G}EV(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \text{ for } 1+\xi\left(\frac{z-\mu}{\sigma}\right) > 0; \tag{4.1}$$

where μ, σ and ξ are respectively the location, the scale and the shape parameters. Let Y_{kl} be the daily precipitation amount on day l of year k. Suppose that K years are available and let the annual maxima be denoted for any year $0 \leq k \leq (K-1)$ by

$$M_k = \max_{1 \leq l \leq 365} Y_{kl}. \tag{4.2}$$

For a given k, if the sequence of daily precipitation $(Y_{kl} : 1 \leq l \leq 365)$ is independent and identically distributed, then the corresponding maximum can be approximated by the GEV distribution. However, this sequence is not independent nor identically distributed. Nevertheless, the dependence is not a problem. For daily precipitation data, it is generally assumed that precipitation events are independent if they are separated by a certain number of days. For such a structure of dependence, the GEV approximation for the maximum still holds (Leadbetter et al., 1983, Chap. 3). Also, precipitation is subject to seasonal fluctuations that violates the assumption that the $(Y_{kl} : 1 \leq l \leq 365)$ have a common distribution.

Nevertheless, it is often assumed, for practical reasons, that the maximum still converges to a GEV distribution (Coles, 2001, Chap. 3). The T-year return level is simply the quantile of order $1 - 1/T$ of the fitted GEV.

To account for the non-stationarity induced by climate changes in the maxima analysis, the standard approach consists in allowing the GEV distribution's parameters to be modeled as functions of time (e.g. Coles, 2001):

$$M_k \sim \mathcal{G}EV(\mu_k, \sigma_k, \xi_k). \tag{4.3}$$

It is common though to consider a time invariant shape parameter (see for example Katz, 2013) because its estimation variance can be very large. Actually, it is even the case in a stationary scenario. Also, Renard et al. (2013) showed that identifiability issues occur when the shape parameter is modeled with time as well as the other parameters; which makes inference impossible.

Generally, the estimations of return levels have a large variance due to the lack of extremal data available to fit with precision the extreme value model. Moreover, considering the non-stationarity in the maxima sequence $(M_k : 0 \leq k \leq K-1)$ increases the estimation variance. Indeed, the same amount of data is used to estimate more parameters since the GEV parameters are then functions of k. To counter this increase of variance in the non-stationary context, an approach that consists in preprocessing the data before taking the maximum was developed by Eastoe and Tawn (2009). They modeled the non-stationary ozone concentration extremes using a Box-Cox transformation to stationarize the data before applying a stationary extreme model. Eastoe and Tawn (2009) argued that the preprocessing gives a better description of the non-stationarity involved in the extreme value as the non-stationarity of the generating process was estimated. In a similar way, Jalbert et al. (2015) developed a preprocessing approach equivalent to considering time varying GEV parameters but using link funtions that relate the daily precipitation to time. Unlike the Box-Cox transformation proposed by Eastoe and Tawn (2009), the transformation proposed by Jalbert et al. (2015) is tractable in the GEV parametrization, which improves the model interpretation.

Also in order to reduce the estimation variance in extreme value analysis, regional statistical models have recently been developed to combine the data from several sites. Such models allow information to be shared across locations. The paper by Davison et al. (2012) provides a review of these models. Two groups of approaches have been developed: one that models directly spatial dependence of the data and one that models spatial dependence of the marginal parameters. The spatial distribution of the data can be modeled with copulas (see Genest and Favre, 2007, for an introduction). Copulas have been applied to model precipitation extremes on multiple sites, notably by Sang and Gelfand (2010) and Ghosh and Mallick (2011). Modeling with copulas for spatial purposes is difficult because their formulation in high dimensions is intractable. Even if some formulation that uses pairwise dependence ex-

ists (see Varin, 2008, for example), copulas are computationally inefficient for large datasets. Max-stable processes (de Haan and Ferreira, 2007) are another approach to model spatial extremes. Their practical application is also difficult since their likelihood function is unknown in its closed form for more than two locations. Nevertheless, Smith and Stephenson (2009), Thibaud et al. (2013) and Huser and Davison (2014) used a pairwise composite likelihood to fit such a model for precipitation extremes. However, when considering many locations, max-stable processes are computationally too heavy to fit. In summary, the first group of approaches is not yet suitable for analyzing large datasets like RCM outputs. The second group, that concerns spatial dependence of marginal parameters, includes the Regional Frequency Analysis approach (RFA, Hosking and Wallis, 2005) and the Bayesian Hierarchical Models approach (BHMs, Banerjee et al., 2004). RFA suffers from a limitative constraint, which is the existence assumption of a scaling constant and of a mean flood index over a defined homogeneous region. On one hand, the definition of such homogeneous region is not straightforward and on the other, the existence of such index flood is not certain (Katz et al., 2002). Hierarchical modeling, however, does not require the definition of homogeneous regions as the marginal local parameters are the results of a spatial stochastic process. It links over space the marginal parameters. The consequence of this resides in the fact that estimation at a particular site benefits from information provided by other sites. Cooley et al. (2007), Sang and Gelfand (2009) and Dyrrdal et al. (2015) applied hierarchical models on observed extreme precipitation data, Schliep et al. (2009) and Cooley and Sain (2010) used these models for precipitation extremes simulated by RCMs. Also, Geirsson et al. (2015) applied these models to the precipitation generated by a meteorological model. Because it models the marginal parameters, a BHM cannot provide estimations of precipitation extremes simultaneously at several sites in opposition to max-stable processes and copulas.

The goal of this paper is to provide a non-stationary and regional statistical model especially suited for the extreme precipitations generated by a climate model. The way non-stationarity is handled differs from the usual approach. Instead of estimating temporal trend with maxima series only, we used the daily precipitation series as proposed by Jalbert et al. (2015). Our methodology can be seen as a data preprocessing step before applying a regional statistical model, but actually it is equivalent to modeling the GEV parameters as functions of daily precipitation, not as functions of the maxima or the excesses. The non-stationarity is then better captured and, combined with the regional approach, the estimation of return levels is improved. A Bayesian hierarchical model that incorporates an intrinsic Gaussian Markov random field as prior to account for spatial dependency of the local GEV parameters is developed. Despite the fact that BHM does not model the data directly, it nevertheless enables to reproduce marginal behavior, which is useful for describing extremes. An application of the model, which consists in estimating the 100-year return level is presented. The return levels are estimated for North America with the developed model fitted with a simulation from the Canadian Regional Climate Model (`CRCM`). The remainder of the paper is as follows:

Section 4.2 introduces the data produced by the Canadian Regional Climate Model and the notation used throughout the paper, Section 4.3 describes the statistical model, Section 4.4 is devoted to the description of the results, Section 4.5 contains the discussion and Section 4.6 concerns the conclusion.

4.2 Data

4.2.1 CRCM description

The regional simulation under investigation was performed by the Canadian Regional Climate Model 4.2.3, based on `CRCM` 4.2 (Music and Caya, 2007; de Elìa and Côté, 2010). The regional model was driven by the outputs from the third version of the Canadian Global Coupled Model (`CGCM3` Scinocca et al., 2008), which performed a transient simulation following the IPCC "observed 20^{th} century" scenario for 1961-2000 and the "SRES A2" (Nakicenovic et al., 2000) scenario for 2001-2100. The `CRCM` simulation domain covers North America with a regular grid of size 180×200 and each grid point contains a vector of daily precipitation during the period $[1961, 2100]$. We restricted our analysis to the 12,570 grid points that represent land areas, which implies that the data lies on a regular grid but where the sea grid points can be seen as missing values.

4.2.2 Notation

The precipitation data are indexed with three subscripts i, k and l. The first index i spans the set of grid points, denoted $\mathcal{V}$, composing the `CRCM` spatial domain. The index k spans the set of simulated years and the index l spans the set of days in a year. For instance, Y_{ikl} denotes the amount of precipitation (in mm) on the l^{th} day of year index k at grid point i, where

$$\begin{aligned} &i \in \mathcal{V}; \\ &0 \leq k \leq 139; \\ &1 \leq l \leq 365. \end{aligned}$$

Note that the `CRCM` does not contain leap years. The index $k = 0$ stands for the year 1961 and $k = 139$ stands for the year 2100.

We use the standard convention to distinguish between random variables and their possible realizations: upper case letters for the random variables and lower case letters for their possible realizations. Since the proposed methodology will be developed under the Bayesian paradigm, the model parameters, which are usually denoted by Greek letters, are also random variables. With Greek letters, we use bold font for the random variables and regular font for their possible realizations.

Some basic notions of graph theory will be useful to define an intrinsic Gaussian Markov random field on the CRCM spatial domain. Let us introduce them. A simple graph $\mathcal{G}$ is composed with nodes and connecting edges. For our purposes, the nodes consist in the CRCM land grid points and the edges connect a pair of adjacent land grid points in the cardinal directions. The graph is usually denoted by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of edges. If an edge exists between the nodes $i \in \mathcal{V}$ and $j \in \mathcal{V}$, then $\{i, j\} \in \mathcal{E}$. Two nodes are neighbors if they are connected with an edge. The neighborhood of a node $i \in \mathcal{V}$, denoted by δ_i, is the set of neighbors of i and it is defined as

$$\delta_i = \{j \in \mathcal{V} : \{i, j\} \in \mathcal{E}\}.$$

The total number of nodes will be denoted n and the number of neighbors of node i will be denotes n_i. Thus,

$$n = \text{Card}(\mathcal{V}) = 12,570; \tag{4.4}$$

$$n_i = \text{Card}(\delta_i). \tag{4.5}$$

Note that $n_i = 4$ for inner nodes and $n_i = 1, 2, 3$ for boundary nodes.

4.3 Regional model for non-stationary maxima

In this section, the developed statistical model used to estimate the CRCM's extreme precipitation is described. The model includes three hierarchical levels. The first level consists in modeling the maxima series marginally for every grid point with the GEV distribution. The second level models the spatial dependence of the GEV marginal parameters with an intrinsic Gaussian Markov random field. The third level provides prior information for the parameters. In Bayesian Hierarchical Models (BHMs), the different levels are linked with conditional relationships (Banerjee et al., 2004).

4.3.1 Marginal model for grid point maxima

This layer of the hierarchical model provides the conditional model for the annual maxima series at each grid point. Possible non-stationarity in these series has to be taken into account. To avoid estimating temporal trends with only a few maximum data as in the model of Eq. (4.3), we instead model the GEV parameter using link functions related to daily precipitation. This methodology relies on the assumption that the non-stationary processes that drive extreme values are consistent with those that drive daily precipitation. As pointed out by Eastoe and Tawn (2009), the non-stationarity is better captured from the daily series because the complete result of the non-stationary underlying process is taken into account, not only the result on extremal data. We use a simple approach that requires preprocessing of the data. Suppose for now that daily precipitation are successfully modeled using the year index

k and the day index l as the covariates for each grid point i (Section 4.3.2 is devoted to this topic). Let ν_{ikl} and τ_{ikl} be the corresponding theoretical mean and standard deviation. Note that both of these quantities can evolve with time, *i.e.* with the index k and l. We define the standardized amount of precipitation Z_{ikl} on day l of year k at the grid point i, as:

$$Z_{ikl} = \frac{Y_{ikl} - \nu_{ikl}}{\tau_{ikl}}. \tag{4.6}$$

By construction, the standardized random variable Z_{ikl} has the following properties: $\mathbb{E}[Z_{ikl}] = 0$ and $\mathbb{V}ar[Z_{ikl}] = 1$ for all k, l and i. Note that the standardized series $(Z_{ikl} : 1 \leq l \leq 365), 0 \leq k \leq 139$, at grid point i is second-order stationary as its first two moments are time invariant. If we assume that $(Z_{ikl} : 1 \leq l \leq 365), 0 \leq k \leq 139$, is strictly stationary, then a stationary model for extremes can be applied. It has been shown by Jalbert et al. (2015) that this working hypothesis implies the standard non-stationary extreme value models. It follows that the standardized seasonal maximum at grid point i for year k, denoted by

$$M_{ik} = \max_{1 \leq l \leq 365} Z_{ikl}; \tag{4.7}$$

can be assumed GEV distributed from EVT:

$$f_{[M_{ik}|(\boldsymbol{\mu_i},\boldsymbol{\sigma_i},\boldsymbol{\xi_i})=(\mu_i,\sigma_i,\xi_i)]}(m_{ik}) = \mathcal{G}EV(m_{ik}|\mu_i, \sigma_i, \xi_i). \tag{4.8}$$

The GEV parameters are stationary in time since we suppose that the underlying process Z_{ikl} is, but they are functions of the grid point's location.

4.3.2 Statistical models for daily precipitation

Suitable statistical models are needed for daily precipitation amount to estimate both ν_{ikl} and τ_{ikl} required for the standardization defined in Eq. (4.6). Precipitation is subject to seasonal fluctuations that are much larger than climate variations. Therefore, partitioning the calendar year into seasons to study each of them separately is a good strategy. We assumed that within a season the precipitation amount are identically distributed. For the set of index l corresponding to a season, the last assumption leads to

$$\nu_{ikl} = \nu_{ik}; \tag{4.9}$$

$$\tau_{ikl} = \tau_{ik}. \tag{4.10}$$

We define winter as December, January and February, spring as March, April and May, summer as June, July and August and autumn as September, October and November. For a given grid point, the daily precipitation amounts are modeled separately for each season. We fitted the models proposed in Jalbert et al. (2015), some of which are stationary and some are not. These models are based on the exponential distribution, the gamma distribution and the exponential mixture distribution (see Appendix 4.A for details). The eight models were fitted but only the best was chosen to standardize the seasonal series of the grid point

under consideration. Besides, one could use any other suitable model for daily precipitation to estimate both the non-stationary mean and the non-stationary standard deviation.

When the annual maxima M_{ik} is extracted from the standardized series $(Z_{ikl}, 1 \leq l \leq 365)$ for a given year k at grid point i, the season from wich the maximum comes from is retained with the following random variable:

$$I_{ik} = \begin{cases} 1 & \text{if } M_{ik} \text{ occurs during winter;} \\ 2 & \text{if } M_{ik} \text{ occurs during spring;} \\ 3 & \text{if } M_{ik} \text{ occurs during summer;} \\ 4 & \text{if } M_{ik} \text{ occurs during fall.} \end{cases} \tag{4.11}$$

4.3.3 Regional prior

To account for spatial dependence of the marginal GEV parameters, an intrinsic Gaussian Markov Random Field (iGMRF, Besag and Kooperberg, 1995; Rue and Held, 2005) can be used as a *prior*. Such a tool has been widely applied for modeling underlying dependence in Bayesian hierarchical models. In precipitation extremes, only Cooley and Sain (2010) applied this kind of model to describe precipitation extremes simulated from a RCM. Intrinsic Gaussian Markov random fields are presented in this section with the generic random variable $X = (X_i : i \in \mathcal{V})$ but keep in mind that the GEV parameters $(\boldsymbol{\xi}, \boldsymbol{\sigma}, \boldsymbol{\mu})$ will be modeled independently with this random field.

Définition 4.1. *Let Q be a positive semi-definite matrix of size $n \times n$ with rank $(n-1)$ and with $Q\underline{1} = \underline{0}$, where $\underline{1}$ and $\underline{0}$ denote column vectors of ones and zeros of length n. A first order centered intrinsic Gaussian Markov Random Field (iGMRF) with the precision matrix Q on a finite simple graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a random vector $X^T = (X_i : i \in \mathcal{V})$ with the following improper density*

$$f_X(x) \propto |Q|^* \exp\left\{-\frac{1}{2} x^T Q x\right\}, \; x \in \mathbb{R}^n, \tag{4.12}$$

where the matrix Q satisfies the following property: $q_{ij} \neq 0$ if and only if either $i = j$ or $\{i, j\} \in \mathcal{E}$. The notation $|\cdot|^*$ corresponds to the generalized determinant, i.e. *the product of the non-null eigenvalues.*

The density is improper because the precision matrix Q is singular, which leads to an infinite normalizing constant.

The fact that the joint density is improper does not represent a major problem in our context. Employed as a prior distribution in the Bayesian paradigm, an improper prior $f_{\boldsymbol{\theta}}(\theta)$ could lead to a proper posterior distribution according to the likelihood $f_{[Y|\boldsymbol{\theta}=\theta]}(y)$ if

$$\int_\theta f_{[Y|\boldsymbol{\theta}=\theta]}(y) f_{\boldsymbol{\theta}}(\theta) d\theta < \infty. \tag{4.13}$$

Thus, we will have to verify that this condition is fullfilled with our complete statistical model. Moreover, the impropriety of iGMRFs leads to

$$f_X(x) = f_X(x + c\underline{1}), \text{ for } c \in \mathbb{R};$$

because $Q\underline{1} = 0$. The iGMRF density is thus invariant to the addition of any constant level. So the definition of iGMRF without a constant location parameter is general. Only the precision matrix Q has to be specified.

Even if the joint density is improper, the conditional densities $X_i|X_{-i}$ are proper, where X_{-i} denotes $(X_l : l \neq i)$. It can be shown (see Rue and Held, 2005, Chap. 3) that

$$f_{[X_i|X_{-i}=x_{-i}]}(x_i) = \mathcal{N}\left(x_i \middle| -\frac{1}{q_{ii}} \sum_{j \in \delta_i} q_{ij} x_j, \frac{1}{q_{ii}}\right), \forall i \in \mathcal{V}. \tag{4.14}$$

The conditional density at grid point i can be interpreted as a measure of deviation from a local level defined as a weighted mean of the neighboring sites.

A particular scaled precision matrix Q has been used for spatial modeling, which is defined by $Q = \kappa W$ (Rue and Held, 2005), where $0 < \kappa < \infty$ is the precision parameter and W is a structure matrix known from the graph structure. The structure matrix $W = (w_{ij} : i \in \mathcal{V}, j \in \mathcal{V})$ is defined as

$$w_{ij} = \begin{cases} n_i & \text{if } j = i; \\ -1 & \text{if } j \in \delta_i; \\ 0 & \text{otherwise.} \end{cases} \tag{4.15}$$

For such a matrix, we have that W is of rank $n - 1$ and $W\underline{1} = \underline{0}$. We also have that the conditional distribution of X_i knowing all other sites is centered at the arithmetic mean of the neighbors and the variance is smaller as the number of neighbors increases:

$$f_{[X_i|X_{-i}=x_{-i}]}(x_i) = \mathcal{N}\left(x_i \middle| \frac{1}{n_i} \sum_{j \in \delta_i} x_j, \frac{1}{n_i \kappa}\right). \tag{4.16}$$

From the last equation, it can be seen that the parameter κ affects the smoothness of the field. A large precision parameter induces a smooth spatial field while a small value induces a coarse field. Defining a constant κ implicitly assumes that the dependence structure is invariant through space; *i.e.* the grid points i and j have the same dependance to their neighbors wherever i and j are located. We recall here that if one (or more) of the four nearest grid points is not a land point, it is not accounted as a neighbor. So n_i is the number of land points among the four nearest grid points.

For first order iGMRFs, the marginal distribution of any X_i is improper, leading to an undefined mean and an infinite variance (Besag and Kooperberg, 1995). However, it can be shown that any contrast $c^T X$, such that $c^T\underline{1} = 0$, has a proper Gaussian distribution (Besag and

Kooperberg, 1995). For the scaled precision matrix defined in Eq. (4.15), the distribution of $(X_i - X_j)$ for $j \in \delta_i$ and $i \in \mathcal{V}$ is

$$f_{[X_i - X_j]}(y) = \mathcal{N}\left(y \,\middle|\, 0, \kappa^{-1}\right). \tag{4.17}$$

The last equation gives an equivalent interpretation for the parameter κ controling the smoothness of the field.

4.3.4 Hyperpriors

In the last section, we defined the matrix Q as the product between a precision parameter κ and the structure matrix W related to the graph $\mathcal{G}$ representing the spatial domain of our data. Since the structure matrix is known, only the precision parameter has to be estimated. We use the gamma distribution to model the precision parameter:

$$f_{\boldsymbol{\kappa}}(\kappa) = \frac{\beta^\alpha}{\Gamma(\alpha)} \kappa^{\alpha-1} \exp\left(-\beta\kappa\right) = \mathcal{G}amma(\kappa|\alpha, \beta), \text{ for } \kappa > 0, \alpha > 0 \text{ and } \beta > 0. \tag{4.18}$$

The parameters specification for this hyperprior distribution can be facilitated with Eq. (4.17). We put a prior knowledge on the precision of the difference of two neighbors, *i.e.* on the precision of $X_i - X_j$. In other words, we put a prior knowledge on the smoothness of the field. If such a knowledge is not available, Sørbye and Rue (2014) proposed an approach to define appropriate informative hyperprior based on the gamma distribution that can be adapted to the present context.

4.3.5 Implementation for the regional model

From the assumptions stated in Section 4.3.1, the likelihood of the standardized annual maxima sequences $(M_{ik} : 0 \leq k \leq 365)$ at grid point i is

$$f_{[M_i|(\boldsymbol{\mu_i},\boldsymbol{\sigma_i},\boldsymbol{\xi_i})=(\mu_i,\sigma_i,\xi_i)]}(m_i) = \prod_{k=0}^{139} \mathcal{G}EV(m_{ik}|\mu_i, \sigma_i, \xi_i). \tag{4.19}$$

Using the common assumption in BHMs of conditional independence given the parameters (Banerjee et al., 2004), the joint conditional likelihood for the set of grid points is given by:

$$f_{[M|(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\xi})=(\mu,\sigma,\xi)]}(m) = \prod_{i\in\mathcal{V}} \prod_{k=0}^{139} \mathcal{G}EV\left(m_{ik}|\mu_i, \sigma_i, \xi_i\right). \tag{4.20}$$

The spatial dependence in each GEV parameters is modeled in the upper level with a first order intrinsic Gaussian Markov random field:

$$f_{[\boldsymbol{\mu}|\boldsymbol{\kappa_\mu}=\kappa_\mu]}(\mu) \propto \exp\left(-\frac{\kappa_\mu}{2}\mu^T W \mu\right); \tag{4.21}$$

$$f_{[\boldsymbol{\phi}|\boldsymbol{\kappa_\phi}=\kappa_\phi]}(\phi) \propto \exp\left(-\frac{\kappa_\phi}{2}\phi^T W \phi\right); \tag{4.22}$$

$$f_{[\boldsymbol{\xi}|\boldsymbol{\kappa_\xi}=\kappa_\xi]}(\xi) \propto \exp\left(-\frac{\kappa_\xi}{2}\xi^T W \xi\right); \tag{4.23}$$

where $\boldsymbol{\phi} = \log \boldsymbol{\sigma}$. The hyperprior distribution for the parameters $(\boldsymbol{\kappa}_\mu, \boldsymbol{\kappa}_\phi, \boldsymbol{\kappa}_\xi)$ defining the smoothness of the iGMRFs is the following informative density:

$$f_{(\boldsymbol{\kappa}_\mu, \boldsymbol{\kappa}_\phi, \boldsymbol{\kappa}_\xi)}(\kappa_\mu, \kappa_\phi, \kappa_\xi) \propto \kappa_\mu^{\alpha_\mu - 1} \exp\left(-\beta_\mu \kappa_\mu\right) \kappa_\phi^{\alpha_\phi - 1} \exp\left(-\beta_\phi \kappa_\phi\right) \kappa_\xi^{\alpha_\xi - 1} \exp\left(-\beta_\xi \kappa_\xi\right). \quad (4.24)$$

The corresponding posterior distribution is proportional to the product of Eq. (4.20) to Eq. (4.24). It can be shown that the posterior distribution is proper as long as $n > 1$ (see Appendix 4.B). However, the normalizing constant is intractable but a sample can be obtained using the Gibbs sampler algorithm. The Gibbs sampler is a Markov chain Monte Carlo procedure that uses the complete conditional distributions to generate a sample from the joint distribution. For our model, the complete conditional distributions are quite easy to obtain, thanks to the conditional independence assumption in BHMs. A sample of size m from the parameter distribution given the data is obtained by iterating the Gibbs sampler algorithm given in Appendix 4.C.

4.3.6 Return level estimation

For most practical purposes, the GEV parameter estimations are not the quantity of interest but the estimated T-year return level is. For the standardized series Z_{ikl}, the T-year return level at grid point i corresponds to the $(1 - \frac{1}{T})$-quantile of the distribution $\mathcal{G}EV(\mu_i, \sigma_i, \xi_i)$, denoted by $m_i^{(T)}$. A simple inversion of the cumulative distribution function of the GEV distribution shows that this quantile is given by

$$m_i^{(T)} = \mu_i - \frac{\sigma_i}{\xi_i} \left\{ 1 - \left[-\log\left(1 - \frac{1}{T}\right) \right]^{-\xi_i} \right\}. \quad (4.25)$$

We estimate this quantity using the MCMC sample from the posterior distribution of $(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\xi})$.

The unstandardized T-year return levels in the precipitation space, denoted $r_{ik}^{(T)}$ can be defined as a function of the season s and the year index k as:

$$[r_{ik}^{(T)} | S_i = s] = \tau_{ik}^s \times m_i^{(T)} + \nu_{ik}^s; \quad (4.26)$$

where ν_{ik}^s and τ_{ik}^s denotes respectively the mean and the standard deviation of the (X_{ikl} : for the set of l that corresponds to season s). Note that a different return level can be computed for each season and each year in order to keep the exceedance probability constant in time. Katz et al. (2002) called these time evolving quantiles "effective return levels". For simplicity, a single T-year return level can be defined for the whole period at each grid point i, denoted $R_i^{(T)}$, for which the probability of exceedance is at most $1/T$ for all years of the period. It is simply the maximum among the year index k:

$$[R_i^{(T)} | S_i = s] = \max_{0 \leq k \leq 139} \tau_{ik}^s \times m_i^{(T)} + \nu_{ik}^s; \quad (4.27)$$

The unconditional estimations of the unstandardized annual T-year return levels can be computed by conditional expectation:

$$\mathbb{E}\left(R_i^{(T)}\right) = \sum_{s=1}^{4} \mathbb{E}\left(R_i^{(T)} | S_i = s\right) \hat{\mathbb{P}}\left(S_i = s\right); \tag{4.28}$$

where

$$\hat{\mathbb{P}}\left(S_i = s\right) = \frac{\text{card}\{k : I_{ik} = s\}}{140}. \tag{4.29}$$

4.4 Results

The Bayesian hierarchical model described in the present paper was fitted with a Markov chain generated with a parallel Monte Carlo procedure. A sample of the parameters posterior distribution of size $110,000$ was generated. The first $10,000$ iterations were discarded as the burn-in period and only one in every 10 iterations was kept to reduce the dependence between the iterations. Bayes estimates were defined as the posterior mean. It took a total of 18 hours of computation time on a 2.53 GHz processor.

We also fitted a local model to compare the results with the regional one. The local model consisted in the model expressed in Eq. (4.19), where the parameter estimation procedure was performed independently at each grid point by maximum likelihood. Those estimates were used as initial values for the MCMC procedure required to fit the regional model.

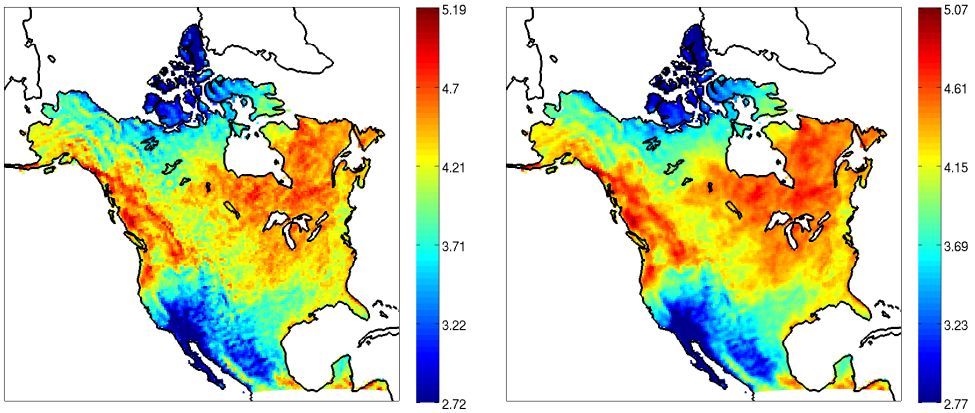
The initial values for the spatial parameters $\boldsymbol{\xi}$, $\boldsymbol{\sigma}$ and $\boldsymbol{\mu}$ were the local GEV maximum likelihood estimates.

4.4.1 Local vs regional estimations

The GEV local estimates for the standardized precipitation annual maxima are shown in Figure 4.1. As shown in Figure 4.1, a strong spatial correlation is apparent for the three parameters. The modeling of this spatial correlation could allow information to be shared across grid points, enabling a better estimation accuracy. The regional model introduced in the present paper was developed especially for this purpose. The regional estimates, which come from our Bayesian hierarchical model, are also shown in Figure 4.1. These regional estimates are smoother than the corresponding local estimates. It is a consequence of the information sharing across locations.

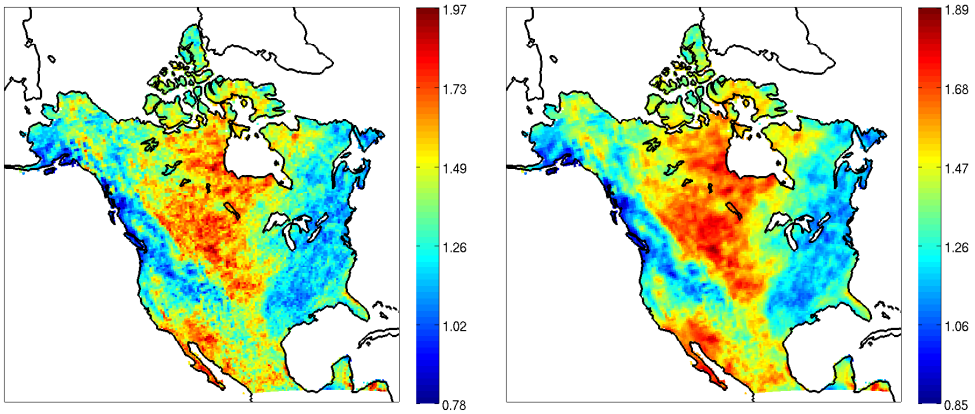
4.4.2 Goodness of fit

Figure 4.2 shows the quantile–quantile plots between the standardized annual maxima and the regionally fit stationary GEV for six grid points subject to different meteorologies. The grid points under consideration are the ones containing the cities of Montreal (QC), Chicago (IL), Churchill (MA), Vancouver (BC), Washington D.C. and Nashville (TN). The adjusted



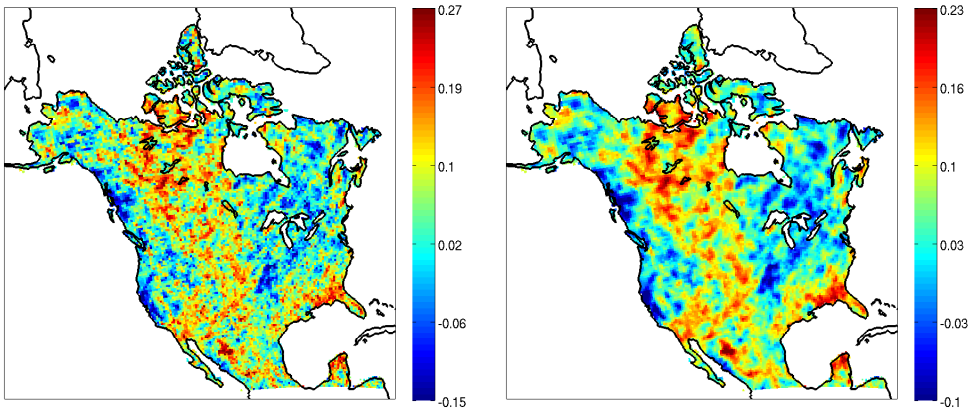
(a) Pointwise μ estimates

(b) Regional μ estimates



(c) Pointwise σ estimates

(d) Regional σ estimates



(e) Pointwise ξ estimates

(f) Regional ξ estimates

Figure 4.1 – GEV parameter estimates for the standardized maxima series at every grid point. On the left, local estimates obtained with local maximum likelihood. On the right, regional estimates obtained with our regional model.

stationary GEV fits the standardized annual maxima series very well. On one hand, it validates our choice to preprocess the series for removing the non-stationarity. On the other hand, it shows that the regional GEV parameter estimations are still well suited for marginal description. In fact, a local model should better fit the data because it is not subject to external constraints like the regional one that we imposed with the iGMRF in the upper level. Nevertheless, the fit with the regional model is quite satisfactory.

4.4.3 Return level estimations

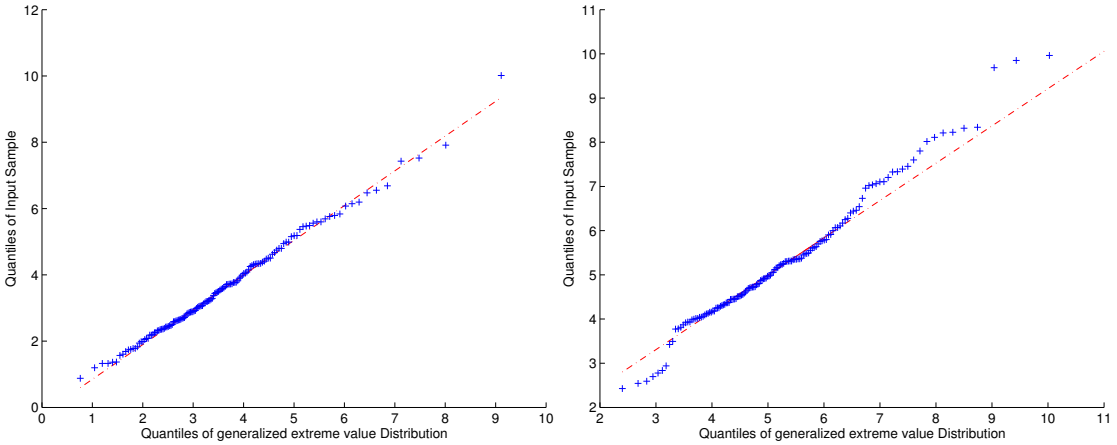
The 100-year unconditional return level estimations, as defined in Eq. (4.28), were computed for every grid point with the local model and our regional model. The local estimations were obtained using the marginal GEV maximum likelihood estimations in Eq. (4.25). The local and the regional estimations of the 100-year return level are shown in Figure 4.3. The local and the regional estimations are quite similar. However, the standard deviation corresponding to the 100-year return level regional estimations is quite smaller than the standard deviation of the local estimations. We compared the standard deviations of the quantity Eq. (4.25) obtained with the local model and the regional model. Let σ_i^L be the standard deviation at grid point i for the local model and σ_i^R be the standard deviation at grid point i for the regional model. Figure 4.4 shows the ratios $\left(\sigma_i^L/\sigma_i^R : i \in \mathcal{V}\right)$. The standard deviations of the regional model are between 15% and 50% of the corresponding ones from the local model. Such a reduction comes from the information sharing across locations in the regional model. Therefore, for similar 100-year return level estimations, the uncertainty is quite reduced when using the regional model. The regional model could thus be very useful for instance in a risk-based strategy for flood management.

4.5 Discussion

4.5.1 Non-stationarity

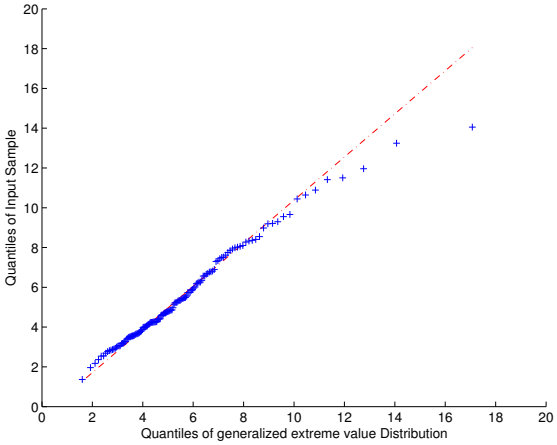
In the proposed approach, the data were preprocessed for removing the non-stationarity, or at least simplifying it enough. The standardized maxima series were then modeled with a stationary regional model. The first mentioned argument in favor of preprocessing was related to the better estimation of the non-stationary processes using the complete data instead of using only the maxima series. Another argument could be the difficulty to elaborate a regional model for non-stationary maxima. For instance, if a non-stationary GEV was used as the marginal model for the maxima series. Suppose that this particular function is used for the location parameter at grid point i: $\mu_{ik} = \mu_i^0 + \mu_i^1 k$. Then the multivariate vector (μ^0, μ^1) would have to be jointly modeled. This would lead to a much more complex multivariate modeling.

In the proposed methodology, the non-stationarity in precipitation varies in space to account

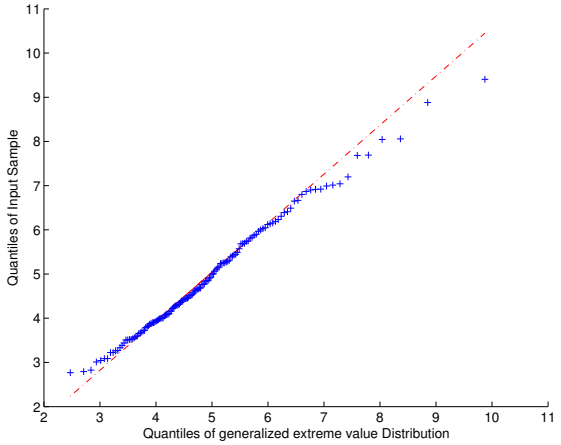


(a) Montreal

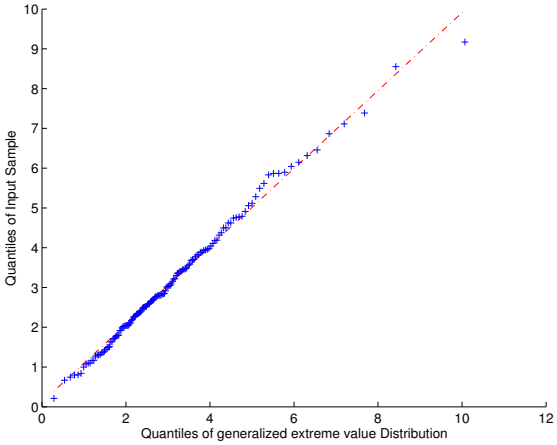
(b) Chicago



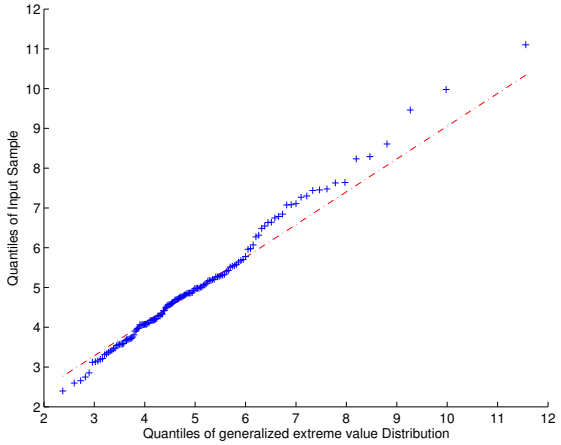
(c) Churchill



(d) Vancouver

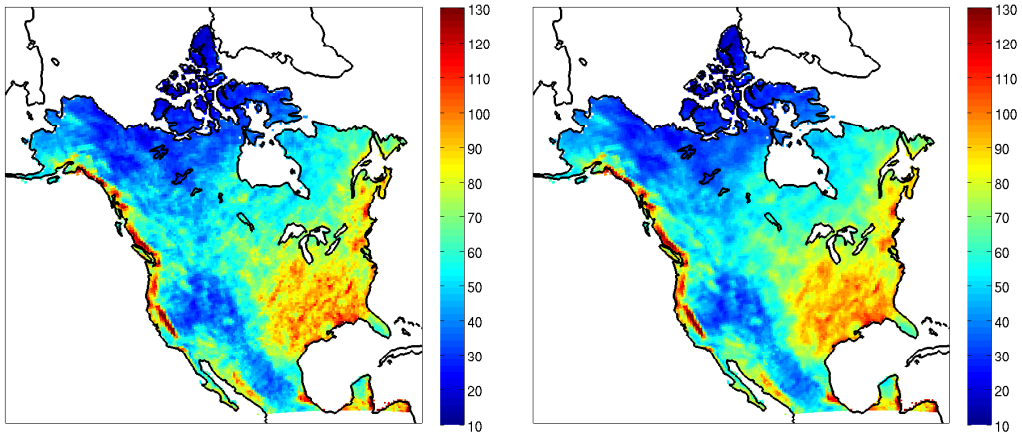


(e) WashingtonDC



(f) Nashville

Figure 4.2 – QQplots between the standardized maxima and the corresponding GEV distribution fit with the regional model.



(a) Local estimates (b) Regional estimates

Figure 4.3 – 100-years return levels estimates for the local model and our regional model at every grid point.

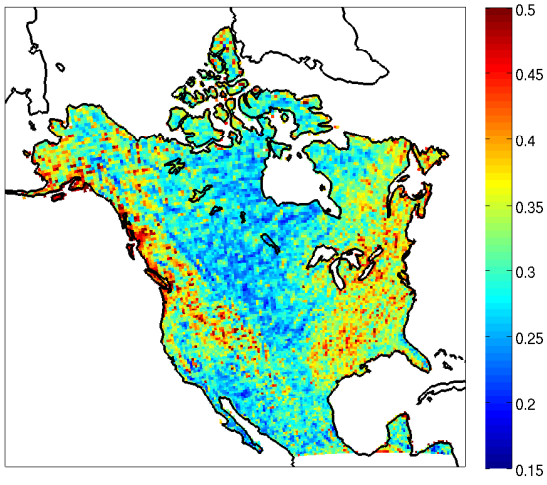


Figure 4.4 – Standard deviation reduction factor between the T-year return level estimations obtained with the regional approach over the local approach at every grid point.

for the different meteorologies over North America. However, a more general model could be developed for regionalizing the non-stationary parameters as well. In other words, the estimations of ν_{ikl} and τ_{ikl} should come from a global regional model. It is not clear if such a model would enhance the results because the number of daily observations required to estimate ν_{ikl} and τ_{ikl} is very large, but it could be verified.

4.5.2 Regional model

Because of the gridded structure of the RCM outputs, Gaussian Markov random fields are more appropriate than ordinary Gaussian fields to model the spatial dependency among the marginal GEV parameter. Gaussian Markov random fields are conveniently expressed in terms of the precision matrix which, given the neighbor structure of our grid, is a very sparse matrix. Sparseness of the precision matrix facilitates computer implementation. Furthermore, unlike Gaussian Markov random fields, ordinary Gaussian fields usually fail to exploit the conditional independence present in our data. Ordinary Gaussian fields are described via their covariance matrices and for the situation at hand such a matrix will not be sparse; the covariance between two grid points would decrease with distance but would not be null. Nonetheless, ordinary Gaussian fields may be useful for incorporating covariates in the statistical model. Such models have been used for RCM outputs on precipitation (Sain et al., 2011) and on runoff (Najafi and Moradkhani, 2014), but not on precipitation extremes.

4.5.3 Structure matrix

We also considered the neighborhood structure defined as the eight nearest neighbors. Such a neighborhood did not significantly influence parameter estimation nor the return level estimation. We therefore chose to retain the four neighbors structure.

Other definitions for the structure matrix W are also available. For instance, Rue and Held (2005) defined the structure matrix using distinct weights for different distances between neighbor sites. Yue and Speckman (2010) defined a structure matrix as:

$$w_{ij} = \begin{cases} 20 & \text{if } j = i; \\ -8 & \text{if } j \text{ is one of the 4 nearest neighbors;} \\ 2 & \text{if } j \text{ is one of the 4 first diagonal neighbors;} \\ 1 & \text{if } j \text{ is one of the second-order cardinal neighbors;} \\ 0 & \text{otherwise.} \end{cases} \tag{4.30}$$

which corresponds to a thin-plate spatial spline (Cressie, 2003). However, such approaches would be time consuming to implement on our grid because corrections should be applied for every grid point that does not have four neighbors, *i.e.* four adjacent land grid points.

Smoothness of the intrinsic Gaussian Markov random fields were assumed constant over the entire spatial domain; *i.e.* the $(\boldsymbol{\kappa_\xi}, \boldsymbol{\kappa_\phi}, \boldsymbol{\kappa_\mu})$ parameters were space invariant. Considering

non-stationary iGMRFs, where the smoothness parameters can change with location, may be an improvement for future work.

4.5.4 Extension to an ensemble of simulations

It is generally recognized that considering a large number of climate models should minimize the structural uncertainty that climate models suffer from (Hagedorn et al., 2005). For the moment, the methodology developed in the present paper can be applied independently on several simulations. However, the methodology could be extended to consider simultaneously many simulations. Through iGMRFs, it would be possible to connect grid points of different simulations as proposed by Sain et al. (2011). Considering more than one simulation opens other possibilities for model parametrization through the neighborhood structure. However, this would lead to a much more complex graph for the iGMRF model and, consequently, a significant increase in computation time as well as the possibility of parameter identifiability issues. These issues would be interesting to investigate in future work.

4.6 Conclusion

In this paper, a non-stationary and regional hierarchical model has been developed to describe extreme precipitation from a climate model. The primary interest in developing a regional statistical model was to improve the T-year return level estimations for the non-stationary precipitation simulated by a climate model. At first, non-stationary means and standard deviations were estimated, for every grid points, with the statistical models proposed by Jalbert et al. (2015). Then, we standardized the annual maxima according to Equation (4.6). Spatial process modeling with iGRMFs benefits from the gridded structure of the data to improve the computational efficiency. Non-stationarity in the seasonal maxima was assessed from the daily precipitation so that the whole transient series was considered. We believe that using transient daily precipitation results in a more accurate description of non-stationarity than using only seasonal maxima. Such an approach has never been implemented before with a regional model to describe the transient precipitation extremes of a climate model simulation.

Despite the fact that the description of extreme precipitation is improved with this model, we only investigated one climate model simulation. Thus, our study does not cover a large range of possible outcomes for precipitation, which could be achieved by considering an ensemble of different climate models. Nevertheless, we believe that the approach presented here constitutes a significant contribution to the study of transient simulations. Furthermore, we believe that with appropriate modifications our approach could be used to handle several climate simulations simultaneously.

References

Banerjee, S., Gelfand, A. E., and Carlin, B. P. (2004). *Hierarchical modeling and analysis for spatial data*. Chapman & Hall, Boca Raton, USA, 2 edition.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236.

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer, London, United Kingdom.

Cooley, D., Nychka, D., and Naveau, P. (2007). Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, 102(479):824–840.

Cooley, D. and Sain, S. R. (2010). Spatial hierarchical modeling of precipitation extremes from a regional climate model. *Journal of Agricultural, Biological, and Environmental Statistics*, 15(3):381–402.

Cressie, N. A. (2003). *Statistics for spatial data*. Wiley, New York, USA.

Davison, A. C., Padoan, S. A., and Ribatet, M. (2012). Statistical modeling of spatial extremes. *Statistical Science*, 27(2):161–186.

de Elìa, R. and Côté, H. (2010). Climate and climate change sensitivity to model configuration in the Canadian RCM over North America. *Meteorologische Zeitschrift*, 19(4):325–339.

de Haan, L. and Ferreira, A. (2007). *Extreme value theory: an introduction*. Springer, New York, USA.

Dyrrdal, A. V., Lenkoski, A., Thorarinsdottir, T. L., and Stordal, F. (2015). Bayesian hierarchical modeling of extreme hourly precipitation in Norway. *Environmetrics*, 26(2):89–106.

Eastoe, E. F. and Tawn, J. A. (2009). Modelling non-stationary extremes with application to surface level ozone. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 58(1):25–45.

Geirsson, O. P., Hrafnkelsson, B., and Simpson, D. (2015). Computationally efficient spatial modeling of annual maximum 24-h precipitation on a fine grid. *Environmetrics*, 26(5):339–353.

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

Ghosh, S. and Mallick, B. K. (2011). A hierarchical Bayesian spatio-temporal model for extreme precipitation events. *Environmetrics*, 22(2):192–204.

Hagedorn, R., Doblas-Reyes, F. J., and Palmer, T. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. *Tellus A*, 57(3):219–233.

Hosking, J. R. M. and Wallis, J. R. (2005). *Regional frequency analysis: an approach based on L-moments*. Cambridge University Press, Cambridge, United Kingdom.

Huser, R. and Davison, A. C. (2014). Space-time modelling of extreme events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):439–461.

IPCC (2012). *Managing the risks of extreme events and disasters to advance climate change adaptation. A special report of Working Groups I and II of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, USA.

Jalbert, J., Favre, A. C., Bélisle, C., Angers, J.-F., and Paquin, D. (2015). Canadian RCM projected transient changes to precipitation occurrence, intensity and return level over North America. *Journal of Climate*, 28(17):6920–6937.

Katz, R. W. (2013). Statistical methods for nonstationary extremes. In *Extremes in a changing climate*, chapter 2, pages 15–37. Springer, New York, USA.

Katz, R. W., Parlange, M. B., and Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in Water Resources*, 25(8):1287–1304.

Leadbetter, M. R., Lindgren, G., and Rootzén, H. (1983). *Extremes and related properties of random sequences and processes*. Springer, New York, USA.

Meylan, P., Favre, A.-C., and Musy, A. (2012). *Predictive hydrology: a frequency analysis approach*. CRC Press, Boca Raton, USA.

Music, B. and Caya, D. (2007). Evaluation of the hydrological cycle over the Mississippi river basin as simulated by the Canadian Regional Climate Model (CRCM). *Journal of Hydrometeorology*, 8(5):969–988.

Najafi, M. R. and Moradkhani, H. (2014). A hierarchical Bayesian approach for the analysis of climate change impact on runoff extremes. *Hydrological Processes*, 28(26):6292–6308.

Nakicenovic, N., Alcamo, J., and Davis, G. (2000). *Emissions scenarios*. Cambridge University Press, Cambridge, United Kingdom and New York, USA.

Rasmussen, S. H., Christensen, J. H., Drews, M., Gochis, D. J., and Refsgaard, J. C. (2012). Spatial-scale characteristics of precipitation simulated by regional climate models and the implications for hydrological modeling. *Journal of Hydrometeorology*, 13:1817–1835.

Renard, B., Sun, X., and Lang, M. (2013). Bayesian methods for non-stationary extreme value analysis. In *Extremes in a changing climate*, chapter 3, pages 39–95. New York, USA.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC Press, Boca Raton, USA.

Sain, S. R., Furrer, R., and Cressie, N. (2011). A spatial analysis of multivariate output from regional climate models. *The Annals of Applied Statistics*, 5(1):150–175.

Sang, H. and Gelfand, A. E. (2009). Hierarchical modeling for extreme values observed over space and time. *Environmental and ecological statistics*, 16(3):407–426.

Sang, H. and Gelfand, A. E. (2010). Continuous spatial process models for spatial extreme values. *Journal of agricultural, biological, and environmental statistics*, 15(1):49–65.

Schliep, E. M., Cooley, D., Sain, S. R., and Hoeting, J. A. (2009). A comparison study of extreme precipitation from six different regional climate models via spatial hierarchical modeling. *Extremes*, 13(2):219–239.

Scinocca, J. F., McFarlane, N. A., Lazare, M., Li, J., and Plummer, D. (2008). The CCCma third generation AGCM and its extension into the middle atmosphere. *Atmospheric Chemistry and Physics Discussions*, 8(2):7883–7930.

Smith, E. L. and Stephenson, A. G. (2009). An extended Gaussian max-stable process model for spatial extremes. *Journal of Statistical Planning and Inference*, 139(4):1266–1275.

Sørbye, S. H. and Rue, H. (2014). Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, 8(0):39–51.

Thibaud, E., Mutzner, R., and Davison, A. C. (2013). Threshold modeling of extreme spatial rainfall. *Water Resources Research*, 49(8):4633–4644.

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis*, 92(1):1–28.

Yue, Y. and Speckman, P. L. (2010). Nonstationary Spatial Gaussian Markov Random Fields. *Journal of Computational and Graphical Statistics*, 19(1):96–116.

4.A Parametric models for daily precipitation

For daily precipitation intensity, we only modeled the positive precipitation (after a bias reduction of 1 mm) and the zero values are ignored, see Jalbert et al. (2015) for more details. Eight different parametric distributions were considered to model Y_{ikl}:

$$
\begin{aligned}
&\mathcal{M}_1 : Y_{ikl} \sim \mathcal{E}\text{xp}(\theta_1);\\
&\mathcal{M}_2 : Y_{ikl} \sim \mathcal{G}\text{amma}(\lambda, \theta_1);\\
&\mathcal{M}_3 : Y_{ikl} \sim (1-w)\ \mathcal{E}\text{xp}(\theta_0) + w\ \mathcal{E}\text{xp}(\theta_1);\\
&\mathcal{M}_4 : Y_{ikl} \sim \mathcal{E}\text{xp}\left\{\theta_1(k)\right\};\\
&\mathcal{M}_5 : Y_{ikl} \sim \mathcal{G}\text{amma}\left\{\lambda, \theta(k)\right\};\\
&\mathcal{M}_6 : Y_{ikl} \sim (1-w)\ \mathcal{E}\text{xp}\left\{\theta_0(k)\right\} + w\ \mathcal{E}\text{xp}\left\{\theta_1(k)\right\};\\
&\mathcal{M}_7 : Y_{ikl} \sim (1-w(k))\ \mathcal{E}\text{xp}\left(\theta_0\right) + w(k)\ \mathcal{E}\text{xp}\left(\theta_1\right);\\
&\mathcal{M}_8 : Y_{ikl} \sim (1-w(k))\ \mathcal{E}\text{xp}\left\{\theta_0(k)\right\} + w(k)\ \mathcal{E}\text{xp}\left\{\theta_1(k)\right\};
\end{aligned}
$$

where

$$
\begin{aligned}
\theta_0(k) &= \alpha_0\alpha_1^k;\\
\theta_1(k) &= \beta_0\beta_1^k;\\
w(k) &= \frac{\exp\left(\gamma_0+\gamma_1 k\right)}{1+\exp\left(\gamma_0+\gamma_1 k\right)}.
\end{aligned}
\tag{4.31}
$$

and

$$
\begin{gathered}
\theta_0 > 0,\ \theta_1 > 0,\ \lambda > 0, p \in (0,1),\\
\alpha_0 > 0,\ \alpha_1 > 0,\ \beta_0 > 0,\ \beta_1 > 0,\\
(\gamma_0, \gamma_1) \in \mathbb{R}^2.
\end{gathered}
$$

Model $\mathcal{M}_3$, as models $\mathcal{M}_6$, $\mathcal{M}_7$ and $\mathcal{M}_8$, is a mixture of exponential distributions. Such a model can be interpreted as follows: a particular realization y_{ikl} of the random variable Y_{ikl} has a probability $(1-w)$ of being a realization of the first mixture component $\mathcal{E}\text{xp}(\theta_0)$ and a probability w of being a realization of the second mixture component $\mathcal{E}\text{xp}(\theta_1)$.

The temporal functions in Equation (4.31) were used to insure that $\theta_0(k)$ and $\theta_1(k)$ are positive for any value of k. Thus, the different models based on exponential and gamma distributions are always well defined, regardless of the value of k. Other functions, such as $\theta_0(k) = \alpha_0 + \alpha_1 \times k$, may be tricky to use to obtain $\theta_0(k) > 0$ for $k \in (0, \infty)$ because the parameter constraints would be $\alpha_0 > 0$ and $\alpha_1 \in \mathbb{R}$.

Table 4.1 summarizes the expectations and the variances as a function of the year index k for these models.

Table 4.1 – Expectations and variances as a function of the year index k for the intensity models.

Index i	Model	$\nu(k)$	$\tau^2(k)$
1	$\mathcal{M}_1$	$\frac{1}{\theta_1}$	$\frac{1}{\theta_1^2}$
2	$\mathcal{M}_2$	$\frac{\lambda}{\theta_1}$	$\frac{\lambda}{\theta_1^2}$
3	$\mathcal{M}_3$	$\frac{1-w}{\theta_0} + \frac{w}{\theta_1}$	$\frac{2(1-w)}{\theta_0^2} + \frac{2w}{\theta_1^2} - \left(\frac{1-w}{\theta_0} + \frac{w}{\theta_1}\right)^2$
4	$\mathcal{M}_4$	$\frac{1}{\theta_1(k)}$	$\frac{1}{\theta_1^2(k)}$
5	$\mathcal{M}_5$	$\frac{\lambda}{\theta_1(k)}$	$\frac{\lambda}{\theta_1^2(k)}$
6	$\mathcal{M}_6$	$\frac{1-w}{\theta_0(k)} + \frac{w}{\theta_1(k)}$	$\frac{2(1-w)}{\theta_0^2(k)} + \frac{2w}{\theta_1^2(k)} - \left\{\frac{1-w}{\theta_0(k)} + \frac{w}{\theta_1(k)}\right\}^2$
7	$\mathcal{M}_7$	$\frac{1-w(k)}{\theta_0} + \frac{w(k)}{\theta_1}$	$\frac{2(1-w(k))}{\theta_0^2} + \frac{2w(k)}{\theta_1^2} - \left\{\frac{1-w(k)}{\theta_0} + \frac{w(k)}{\theta_1}\right\}^2$
8	$\mathcal{M}_8$	$\frac{1-w(k)}{\theta_0(k)} + \frac{w(k)}{\theta_1(k)}$	$\frac{2(1-w(k))}{\theta_0^2(k)} + \frac{2w(k)}{\theta_1^2(k)} - \left\{\frac{1-w(k)}{\theta_0(k)} + \frac{w(k)}{\theta_1(k)}\right\}^2$

4.B Posterior propriety

We will show that the posterior density $f_{[(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\xi},\boldsymbol{\kappa_\mu},\boldsymbol{\kappa_\sigma},\boldsymbol{\kappa_\xi})|M=m]}(\mu,\sigma,\xi,\kappa_\mu,\kappa_\sigma,\kappa_\xi)$ is proper, *i.e.* the integration of the joint density

$$I = \int_\theta f_{[(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\xi},\boldsymbol{\kappa_\mu},\boldsymbol{\kappa_\sigma},\boldsymbol{\kappa_\xi},M)]}(\mu,\sigma,\xi,\kappa_\mu,\kappa_\sigma,\kappa_\xi,m)d\theta$$

is finite, where $\theta = (\mu,\sigma,\xi,\kappa_\mu,\kappa_\sigma,\kappa_\xi)$.

$$\begin{aligned} I = &\int_\theta \prod_{i\in\mathcal{V}}\prod_{k=0}^{139} \mathcal{GEV}(m_{ik}|\mu_i,\sigma_i,\xi_i) \\ &\times {\kappa_\mu}^{\frac{n-1}{2}} \exp\left\{-\frac{\kappa_\mu}{2}\mu^T W\mu\right\} \times {\kappa_\sigma}^{\frac{n-1}{2}} \exp\left\{-\frac{\kappa_\sigma}{2}\sigma^T W\sigma\right\} \times {\kappa_\xi}^{\frac{n-1}{2}} \exp\left\{-\frac{\kappa_\xi}{2}\xi^T W\xi\right\} \\ &\times \kappa_\mu^{\alpha_\mu-1}\exp\left(-\beta_\mu\kappa_\mu\right) \times \kappa_\sigma^{\alpha_\sigma-1}\exp\left(-\beta_\sigma\kappa_\sigma\right) \times \kappa_\xi^{\alpha_\xi-1}\exp\left(-\beta_\xi\kappa_\xi\right) d\theta. \end{aligned}$$

If $n \geq 2$, we have

$$\begin{aligned} I = &\int_\xi\int_\phi\int_\mu \prod_{i\in\mathcal{V}}\prod_{k=0}^{139} \mathcal{GEV}(m_{ik}|\mu_i,\sigma_i,\xi_i) \\ &\times \frac{\Gamma\left(\alpha_\mu+\frac{n-1}{2}\right)}{\left(\beta_\mu+\frac{\mu^T W\mu}{2}\right)^{\alpha_\mu+\frac{n-1}{2}}} \times \frac{\Gamma\left(\alpha_\sigma+\frac{n-1}{2}\right)}{\left(\beta_\sigma+\frac{\sigma^T W\sigma}{2}\right)^{\alpha_\sigma+\frac{n-1}{2}}} \times \frac{\Gamma\left(\alpha_\xi+\frac{n-1}{2}\right)}{\left(\beta_\xi+\frac{\xi^T W\xi}{2}\right)^{\alpha_\xi+\frac{n-1}{2}}} d\mu d\phi d\xi. \end{aligned}$$

Considering the following identity:

$$x^T Q x = \sum_i q_{i+}x_i^2 - \sum_{i<j} q_{ij}(x_i - x_j)^2; \tag{4.32}$$

where q_{i+} denotes $\sum_j q_{ij}$. For fisrt order iGMRFs, q_{i+} equals to zero. Therefore, $x^T Q x > 0$ and we have that

$$I < \frac{\Gamma\left(\alpha_\mu + \frac{n-1}{2}\right)}{(\beta_\mu)^{\alpha_\mu + \frac{n-1}{2}}} \times \frac{\Gamma\left(\alpha_\sigma + \frac{n-1}{2}\right)}{(\beta_\sigma)^{\alpha_\sigma + \frac{n-1}{2}}} \times \frac{\Gamma\left(\alpha_\xi + \frac{n-1}{2}\right)}{(\beta_\xi)^{\alpha_\xi + \frac{n-1}{2}}} \times \int_\xi \int_\phi \int_\mu \prod_{i \in \mathcal{V}} \prod_{k=0}^{139} \mathcal{G}EV(m_{ik}|\mu_i, \sigma_i, \xi_i) d\mu d\phi d\xi;$$

which is finite.

4.C MCMC procedure

A sample of size m from the parameter distribution given the data is obtained by iterating the Gibbs sampler algorithm given in Algorithm 1 and 2 from the initial starting vector $(\mu^0, \phi^0, \xi^0, \kappa^0)$.

Algorithm 1 Gibbs sampler algorithm for generating a sample of size m from the unnormalized density $f_{[(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\xi},\boldsymbol{\kappa_\mu},\boldsymbol{\kappa_\phi},\boldsymbol{\kappa_\xi})|M=m]}(\mu, \sigma, \xi, \kappa_\mu, \kappa_\phi, \kappa_\xi)$

for $t = 1$ to m **do**

Generate (μ^t, ϕ^t, ξ^t) from the unnormalized density $f_{[(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\xi})|M=m,\boldsymbol{\kappa}=(\kappa_\mu,\kappa_\phi,\kappa_\xi)]}(\mu, \sigma, \xi)$ (see Algorithm 2).

Generate κ_μ^t from the density

$$f_{[\boldsymbol{\kappa_\mu}|M=m,(\boldsymbol{\mu},\boldsymbol{\phi},\boldsymbol{\xi})=(\mu,\phi,\xi),\boldsymbol{\kappa_\phi}=\kappa_\phi,\boldsymbol{\kappa_\xi}=\kappa_\xi]}(\kappa_\mu) = \mathcal{G}amma\left(\kappa_\mu \middle| \frac{n-1}{2}, \frac{\mu^T W \mu}{2}\right).$$

Generate κ_ϕ^t from the density

$$f_{[\boldsymbol{\kappa_\phi}|M=m,(\boldsymbol{\mu},\boldsymbol{\phi},\boldsymbol{\xi})=(\mu,\phi,\xi),\boldsymbol{\kappa_\mu}=\kappa_\mu,\boldsymbol{\kappa_\xi}=\kappa_\xi]}(\kappa_\phi) = \mathcal{G}amma\left(\kappa_\phi \middle| \frac{n-1}{2}, \frac{\phi^T W \phi}{2}\right).$$

Generate κ_ξ^t from the density

$$f_{[\boldsymbol{\kappa_\xi}|M=m,(\boldsymbol{\mu},\boldsymbol{\phi},\boldsymbol{\xi})=(\mu,\phi,\xi),\boldsymbol{\kappa_\mu}=\kappa_\mu,\boldsymbol{\kappa_\phi}=\kappa_\phi]}(\kappa_\xi) = \mathcal{G}amma\left(\kappa_\xi \middle| \frac{n-1}{2}, \frac{\xi^T W \xi}{2}\right).$$

end for

In order to achieve the first step of this algorithm, a nested Gibbs sampler algorithm has to be performed. This step is divided in n sub-steps corresponding to the number of grid points. The conditional distribution for a given grid point i is obtained using Eq. (4.14) and is as

follows:

$$f_{[(\boldsymbol{\mu_i},\boldsymbol{\phi_i},\boldsymbol{\xi_i})|M=m,\boldsymbol{\kappa}=\kappa]}(\mu_i,\phi_i,\xi_i) \propto \prod_{k=0}^{139} \mathcal{G}EV\left(m_{ik}|\mu_i,\exp(\phi_i),\xi_i\right) \times \mathcal{N}\left(\mu_i \left| \frac{1}{n_i}\sum_{j\in\delta_i}\mu_j, \frac{1}{n_i\kappa_\mu}\right.\right) \times \mathcal{N}\left(\phi_i \left| \frac{1}{n_i}\sum_{j\in\delta_i}\phi_j, \frac{1}{n_i\kappa_\phi}\right.\right) \times \mathcal{N}\left(\xi_i \left| \frac{1}{n_i}\sum_{j\in\delta_i}\xi_j, \frac{1}{n_i\kappa_\xi}\right.\right). \tag{4.33}$$

The Metropolis-Hastings algorithm is used to generate a realization from the unnormalized density expressed in Eq. (4.33) at each grid point i (see Algorithm 2), where the candidates for $(\mu_i^t, \phi_i^t, \xi_i^t)$ were generated with a random walk.

Algorithm 2 Gibbs sampler algorithm for generating a single realization from the unnormalized density $f_{[(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\xi})|M=m,\boldsymbol{\kappa}=(\kappa_\mu,\kappa_\phi,\kappa_\xi)]}(\mu,\sigma,\xi)$

for all $i \in \mathcal{V}$ **do**
  Generate the candidate $(\mu_i', \sigma_i', \xi_i')$ from an instrumental density say g.
  Generate $u \sim \mathcal{U}niform(0,1)$.
  Compute

$$r = \frac{f_{[\boldsymbol{\mu_i},\boldsymbol{\sigma_i},\boldsymbol{\xi_i}|M=m,\boldsymbol{\kappa}=(\kappa_\mu,\kappa_\phi,\kappa_\xi)]}(\mu_i',\sigma_i',\xi_i') \times g\left(\mu_i^{t-1},\phi_i^{t-1},\xi_i^{t-1}\right)}{f_{[\boldsymbol{\mu_i},\boldsymbol{\sigma_i},\boldsymbol{\xi_i}|M=m,\boldsymbol{\kappa}=(\kappa_\mu,\kappa_\phi,\kappa_\xi)]}\left(\mu_i^{t-1},\phi_i^{t-1},\xi_i^{t-1}\right) \times g\left(\mu_i',\sigma_i',\xi_i'\right)}.$$

  if $r > u$ **then**
    $(\mu_i^t, \phi_i^t, \xi_i^t) = (\mu_i', \sigma_i', \xi_i')$.
  else
    $(\mu_i^t, \phi_i^t, \xi_i^t) = \left(\mu_i^{t-1}, \phi_i^{t-1}, \xi_i^{t-1}\right)$.
  end if
end for

Instead of performing 12,570 sub-steps, step one can be achieved in performing in parallel only two sub-steps using the *coding technique* described by Besag (1974). Such a procedure is possible because of the Markov property of iGMRFs.

The initial values required for the Gibbs sampler were the local GEV maximum likelihood estimates. The hyperparameters defining the prior information on the iGMRFs smoothness were set to

$$\alpha_\mu = \alpha_\phi = \alpha_\xi = 1; \tag{4.34}$$

$$\beta_\mu = \beta_\phi = \beta_\xi = 100; \tag{4.35}$$

which correspond to vague priors. The values of theses hyperparameters did not influence the final results on the estimation of $(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\xi})$, certainly because the number of grid points is very large.

Conclusion

Dans le cadre de cette thèse, un modèle probabiliste non stationnaire et régional a été développé pour les maxima de précipitations simulées par un modèle numérique de climat. La non-stationnarité a d'abord été retirée des séries avant de concevoir le modèle régional pour les maxima des séries transformées supposées stationnaires. Cette méthodologie, qui consiste à prétraiter les données, a permis d'utiliser l'ensemble des données pour estimer la non-stationnarité. En particulier, l'ensemble de la série chronologique des précipitations journalières a été utilisé. Dans les modèles standard pour les valeurs extrêmes non stationnaires, la non-stationnarité est estimée à partir des précipitations extrêmes seulement. Ces dernières constituent généralement un échantillon de taille très limitée. Le prétraitement a donc permis de prendre en compte beaucoup plus de données et d'estimer avec plus de précision la non-stationnarité. La méthode de prétraitement utilisée consistait à standardiser les séries journalières. Les variables aléatoires composant la série standardisée étaient donc centrées et réduites. Les paramètres de la standardisation dépendaient à la fois de la saison et de l'année. La standardisation ne permet que de retirer la non-stationnarité des deux premiers moments de la série journalière. La série transformée était par construction faiblement stationnaire, elle n'était pas nécessairement stationnaire au sens strict. Il a été démontré que supposer que la série standardisée est stationnaire au sens strict implique la simplification usuelle des modèles standard, soit de considérer un paramètre de forme ξ invariant (*cf.* chapitre 3). Par conséquent, pour les mêmes hypothèses, il était avantageux d'utiliser cette méthodologie par rapport à l'approche standard puisque la non-stationnarité était mieux estimée par le prétraitement. Il a d'ailleurs été montré que l'estimation des seuils correspondant au temps de retour centennal était plus précise en utilisant la méthode de prétraitement que la méthode standard (*cf.* chapitre 3).

Cette méthode de prétraitement pour la non-stationnarité a été appliquée aux séries de précipitations provenant d'une simulation effectuée par le Modèle régional canadien du climat. Ce modèle de climat découpe l'Amérique du Nord en une grille régulière composée de 12 570 points de terre. Pour un point de grille donné, tous les maxima annuels standardisés étaient supposés indépendants et identiquement distribués selon une loi GEV. Dans un premier temps, l'ajustement de la loi GEV à chacun des points de grille s'est effectué de façon indépendante. Les résultats ont montré que les estimations ainsi obtenues possédaient une

importante cohérence spatiale (*cf.* chapitre 2). Autrement dit, les estimations des paramètres de la GEV étaient similaires pour des points de grille avoisinants. Il est naturel qu'une telle cohérence spatiale existe, notamment en raison des conditions de continuité des processus physiques, mais l'effet de la standardisation sur celle-ci était inconnu. La standardisation affecte les paramètres de localisation et d'échelle de la GEV mais une cohérence spatiale a subsisté nonobstant la standardisation. Pour le paramètre de forme ξ, la méthode de prétraitement n'a aucun effet. La standardisation n'a donc pas affecté la cohérence spatiale de ce paramètre.

Un modèle hiérarchique bayésien a été développé pour modéliser la dépendance spatiale des paramètres marginaux de la GEV. Au premier niveau du modèle hiérarchique et pour un point de grille i élément de l'ensemble des points de grille $\mathcal{V}$, il a été supposé que les maxima annuels standardisés $(M'_{ik} : 0 \leq k \leq 139)$ de ce point de grille étaient indépendants et identiquement distribués selon la loi $\mathcal{GEV}(\mu_i, \sigma_i, \xi_i)$. De plus, l'hypothèse d'indépendance conditionnelle habituellement posée dans les modèles hiérarchiques a été employée. Elle stipulait que l'ensemble des suites de maxima pour tous les points de grille $(M'_{ik} : 0 \leq k \leq 139), \forall i \in \mathcal{V}$, était conditionnellement indépendants à l'ensemble des paramètres (μ, σ, ξ). Au second niveau du modèle hiérarchique, la cohérence spatiale des $(\boldsymbol{\mu_i}, \boldsymbol{\sigma_i}, \boldsymbol{\xi_i}), \forall i \in \mathcal{V}$ a été modélisée. Une simplification importante a été effectuée qui consistait à supposer que les processus aléatoires spatiaux pour $(\boldsymbol{\mu}_i : i \in \mathcal{V})$, $(\boldsymbol{\sigma}_i : i \in \mathcal{V})$ et $(\boldsymbol{\xi}_i : i \in \mathcal{V})$ étaient indépendants. Les processus spatiaux pour chacun de ces paramètres ont été modélisés par un champ de Markov gaussien intrinsèque d'ordre 1. Le troisième et dernier niveau du modèle hiérarchique était composé de l'information *a priori* sur les paramètres de lissage des champs spatiaux utilisés pour chacun des paramètres de la GEV. Un tel modèle a permis d'améliorer la précision d'estimation des paramètres des lois GEV marginales. En particulier, il a été montré que la variance d'estimation des seuils correspondant à un temps de retour centennal a été considérablement réduit (*cf.* chapitre 4). L'amplitude de cette réduction dépendait du point de grille mais elle était supérieure à 50%.

Le modèle probabiliste non stationnaire et régional a donc permis de réduire considérablement la variance d'estimation des paramètres des lois GEV marginales et par conséquent, des seuils correspondant à des temps de retour donnés. Le modèle développé est novateur puisqu'il incorpore à la fois la non-stationnarité et la cohérence spatiale des précipitations extrêmes simulées par un modèle de climat. Dans le cadre de cette thèse, l'approche de prétraitement développée a permis d'utiliser un modèle régional pour les extrêmes identiquement distribués dans le temps. Autrement, la modélisation spatio-temporelle aurait été beaucoup plus complexe. De plus, la taille du domaine spatial étudié constitue aussi un aspect novateur. L'efficacité numérique du modèle probabiliste a rendu possible l'étude d'un grand ensemble de points de grille. D'une part, les matrices de précision des champs de Markov gaussiens étaient creuses, facilitant la manipulation numérique de celles-ci. D'autre part, la propriété de Markov a permis de simplifier les lois conditionnelles et de paralléliser la procédure Monte-

Carlo nécessaire pour obtenir un échantillon de la loi *a posteriori*. Par ailleurs, le modèle est relativement facile à implémenter et les hypothèses sur lesquelles il repose sont classiques. La spécification des lois *a priori* informatives pour les paramètres de lissage des champs de Markov gaussiens pourrait paraître délicate. Or, ces lois n'ont pratiquement pas d'influence sur l'estimation des paramètres marginaux des lois GEV pour un tel domaine spatial. D'ailleurs, une loi *a priori* vague pourrait être appropriée dans ce cas-ci.

Le modèle probabiliste développé a modélisé la dépendance spatiale des paramètres marginaux des lois GEV. Par conséquent, il était impossible d'estimer les différentes quantités d'intérêt de façon conjointe pour un ensemble de points de grille, par exemple les points de grille composant un grand bassin versant comme celui des Grands Lacs. La capacité d'estimer un risque conjoint pour de grands bassins versants pourrait être très utile pour élaborer des stratégies de gestion du risque d'inondation pour de très grands bassins versants. Les modèles statistiques basés sur les processus max-stables et sur les copules de valeurs extrêmes pourraient être utilisés à cette fin. Toutefois, il n'existe pas de formulation multidimensionnelle simple pour ceux-ci. Dans le cas des processus max-stables, des formulations multidimensionnelles approximatives pour la vraisemblance d'un ensemble de variables aléatoires ont été développées en utilisant tous les couples de variables aléatoires ; la fonction de vraisemblance pour un couple étant connue. Cependant, l'implémentation de tels modèles sur le jeu de données étudié est une tâche complexe en raison du nombre important de points de grille étudiés. Peut-être serait-il possible de simplifier la formulation multidimensionnelle en faisant une hypothèse d'indépendance conditionnelle ? Néanmoins, le modèle développé dans cette thèse est bien adapté pour estimer le risque local associé aux précipitations extrêmes. Il peut par exemple servir à estimer le risque d'inondation pour les petits bassins versant et pour les zones urbaines.

Quelques aspects concernant l'amélioration du modèle mériteraient une investigation. D'abord, les paramètres servant à prétraiter les séries de précipitations journalières ont été estimés de façon indépendante pour chacun des points de grille. Un modèle général pour l'ensemble du domaine pourrait être approprié en raison de la cohérence spatiale des précipitations. La non-stationnarité serait peut-être mieux estimée par un modèle régional. Par ailleurs, il pourrait être profitable de considérer des champs de Markov gaussiens non-stationnaires dans l'espace. Le paramètre de lissage du champ pourrait être variable dans l'espace, par exemple il pourrait évoluer en fonction de la topographie. Le champ pourrait être plus lisse pour les régions sans grande variation de topographie et plus granuleux pour les régions montagneuses. Aussi, il pourrait être utile de considérer les champs de Markov gaussiens d'ordre deux. Un champ d'ordre deux offre davantage de lissage spatial notamment en considérant plus de voisins. Cette propriété pourrait être très utile spécialement pour améliorer la précision de l'estimation du paramètre de forme ξ de la loi GEV.

En conclusion, l'objectif de la thèse a été complété. Un modèle probabiliste non stationnaire et régional a été développé pour les précipitations extrêmes générées par un modèle de climat. La

perspective prépondérante pour des travaux futurs consiste à généraliser le modèle développé pour un ensemble de simulations climatiques. Une simulation générée par un modèle de climat doit être interprétée comme une réalisation probable du climat futur. Néanmoins, l'étude d'un ensemble de simulations provenant de différents modèles de climat devrait être privilégiée afin de contrebalancer l'incertitude structurelle inhérente aux modèles de climat. Un ensemble de simulations couvrirait ainsi un large spectre de réalisations probables du climat futur. Pour une étude d'impact impliquant les précipitations extrêmes, il serait recommandé d'étudier de façon indépendante plusieurs simulations climatiques avec le modèle probabiliste développé. Cela permettrait de couvrir différentes possibilités pour les précipitations extrêmes sans toutefois être en mesure de fournir une description conjointe pour l'ensemble des simulations. Les prédictions concernant les précipitations extrêmes seraient ainsi basées sur un ensemble de réalisations probables du climat futur. De plus, l'incertitude associée à ces prédictions pourrait être également estimée. Un tel modèle généralisé à un ensemble de simulations serait sans aucun doute opportun dans un premier temps pour prédire les précipitations extrêmes, puis pour les études d'impact.