ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Ecole Doctorale ABIES
Agriculture
Agriculture
Alimentation
Food
Biologie
Biology
Environnement
Environment
Santé
Health

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

L'Institut des Sciences et Industries du Vivant et de l'Environnement

(AgroParisTech)

Spécialité : Statistiques

présentée et soutenue publiquement par

Aurélien BECHLER

le 11 décembre 2014

Modélisation et simulation conditionnelle de processus max-stables pour le downscaling des extrêmes

Directrice de thèse : **Liliane BEL**
Co-directeur de la thèse : **Mathieu VRAC**

Jury

M. Etienne LEBLOIS, Ingénieur-Chercheur, IRSTEA	Rapporteur
M. Carlo GAETAN, Associate professor, Università Ca' Foscari	Rapporteur
M. Denis ALLARD, Directeur de recherche, INRA Avignon	Examinateur
M. Jean-Noël BACRO, Professeur, Université Montpellier II	Examinateur
Mme Sophie BASTIN, Chargée de recherche, LATMOS-IPSL	Examinatrice
Mme Petra FRIEDERICHS, Senior Lecturer, University of Bonn	Examinatrice
Mme. Liliane BEL, Professeure, AgroParisTech/INRA	Directrice de thèse
M. Mathieu VRAC, Chargé de recherche, LSCE-IPSL/CNRS	Directeur de thèse

AgroParisTech
INRA UMR 518
16 rue Claude Bernard, 75005, Paris

The consequences of our actions are always so complicated, so diverse, that predicting the future is a very difficult business indeed.

Pr. Albus Dumbledore

Remerciements

Durant ces trois ans, j'ai eu la chance de rencontrer et de côtoyer des gens remarquables, que cela soit en tant que scientifiques ou en tant que personnes. Je tiens donc à vous remercier toutes et tous pour avoir rempli ces années de souvenirs aussi forts qu'indélébiles.

Pour débuter, un grand merci à mes deux encadrants qui m'ont épaulé et qui ont fait en sorte que l'on vive ces trois années ensemble dans les meilleures conditions. Merci à Liliane pour le soutien dont elle a fait preuve, son écoute et sa disponibilité tout au long de ces trois ans. Il reste néanmoins vrai que j'ai été victime d'une censure artistique sans nom au sujet de mon thème violet-crème, couleurs qui reviennent pourtant très forts dans les collections printemps/été 2015 des Beamers. Merci également à Mathieu pour son soutien et sa bonne humeur. Même si nous ne nous voyions que de manière espacée, c'était toujours un plaisir de discuter et de partager avec toi.

De manière générale, je tiens à remercier toutes les personnes du projet McSim qui m'ont apporté leur aide, chacune à leur manière. En particulier, merci à Jean-Noël qui m'a suivi pendant ces trois années, de nos aventures avec les contrôleurs de trains allemands jusqu'à la présidence de mon jury de thèse. Merci également à Christian pour son expertise, ses conseils avisés ainsi que pour l'entrain et la passion qu'il dégage. C'est quand même toi, qui la première fois est venu me décrire ce sujet thèse, des mois avant que je me lance, en griffonnant des processus Poissoniens sur un bout de papier (que j'ai encore). Des remerciements aussi pour Thomas et Marco pour leur précieuse aide et leur bonne humeur.

Ensuite vient la grande famille que constitue notre laboratoire à l'Agro. Je vais ordonner mes remerciements par le biais de l'agencement des bureaux, enjeu majeur de ces trois années (et sans doute de celles qui précèdent et suivront).

Une pensée pour ma première coloc de bureau, Aurélie, avec qui j'ai partagé beaucoup de choses (des bons moments, des opinions, des amis...). Beaucoup de très bons moments passés avec toi au boulot (le billard ça compte?) et en dehors. Et j'espère que cela continuera malgré le nouveau défi professionnel dans lequel tu te lances.

A la première réorganisation des bureaux, Jean-Baptiste est venu emménager et j'ai

découvert un garçon charmant, attentionné et qui a beaucoup compté pour moi. Beaucoup de moments partagés qui peuvent déjà paraître loin mais qui restent bien présents dans ma mémoire. Et puis j'ai aussi découvert ta chérie, Noémie, tout aussi adorable que toi. Je pense m'être fait deux amis que j'espère revoir, même si à l'heure actuelle de nombreux kilomètres et (au moins) un océan nous séparent (vous sentez le gars qui veut se faire inviter à Vancouver?).

Un merci également à Pierre Colin que j'ai retrouvé et appris à redécouvrir à l'Agro et en dehors lors d'une collaboration improvisée autour d'un projet sportif commun, et qui a squatté quelque temps la table de notre bureau.

Et puis il y a Fred, même si arrivé et reparti un petit peu après moi de l'Agro, en grande partie contemporain, mais tout sauf comptant pour rien à mes yeux. Merci à toi, mon Fredo, d'avoir égayé ces trois années, à commencer par tes débats sans queue ni tête (encore une fois : non Mufasa n'est pas un dictateur qui exile les hyènes car elles sont différentes. C'est un dirigeant qui fait de son mieux pour préserver la paix et l'équité au royaume des lions). Beaucoup de souvenirs ensemble, les soirées JV, les quiz foot (tu n'as toujours pas battu le singe), les midis DBZ (euh pardon Emir Kusturica). Mais aussi des discussions plus sérieuses auxquelles nous n'étions pas forcément habitués au début mais nous nous y sommes fait. Heureusement que tes anecdotes venaient alléger l'atmosphère et nous ont permis de réviser la géographie française et de nous rendre compte de la difficulté de l'obtention du permis de conduire. De manière générale, nous avons noué des liens d'amitié et créé une complicité que je ne pensais pas trouver en arrivant en thèse. Ta présence a permis, en plus de faire vivre les petits restaurateurs du quartier, d'apporter de la joie et de la bonne humeur à chacun des moments que j'ai passés dans ce bureau. Pour cela, je te remercie. Et puis sache que, même si dans l'avenir nos chemins divergent, il y a aura toujours un verre ou une console qui attendra que l'on se réunisse autour pour parler du bon vieux temps en riant aux éternelles mêmes blagues.

En dehors de mon bureau, j'ai croisé, au cours de ces trois ans, beaucoup de personnes qui m'ont marqué et dont je vais essayer de dresser une liste que j'espère être la plus complète possible.

A commencer par Pierre (alias le Barbillon), désigné comme étant mon alter-ego vestimentaire, ma projection sur l'étage du haut. Ces remerciements sont "à toi"qui m'a donné la force de continuer à croire en l'être humain, alors que la police de la mode fermait les yeux sur les agissements de certaines personnes dans un bar fameux de Bruxelles. A toi aussi qui m'a appris que la tektonic pouvait avoir de très belles heures devant elle. A toi qui m'a montré qu'être parisien était plus une question d'état d'esprit que de code postal. Pour tout ça et plein d'autres choses, même si je ne suis pas le premier à le faire, je te dis de tout coeur merci.

La dernière année, Anna et Loïc sont venus compléter notre bande (notre meute de deux, composée de Fred et moi) pour former les 4 fanstatistiques qui auront à leur manière marqué la vie de ce labo. Certains diront qu'ils se sont séparés trop tôt, au

sommet de leur art, protecteurs de la bonne humeur, chevaliers noirs de la blague vaseuse, mais sans doute “vaut-il mieux mourir en héros que vivre assez longtemps pour se voir endosser le mauvais rôle”. Au contraire, d’autres trouveront qu’ils ont accompli leur mission en combattant sans relâche l’ennui et la morosité pouvant apparaître dans la quotidien du monde du travail et que des fois il vaut mieux se retirer, en espérant que l’on retiendra de nous que nous fûmes les quatre fondateurs d’une nouvelle génération de doctorants, succédant à une génération dorée de thésards. Le 69 rue Mouffetard fut notre 12 square Grimault à nous, et si notre ordre était moins celui du Phoenix que celui de l’éléphant rose, cela fait longtemps que ce secret n’est plus à garder. La relève est désormais assurée et la bonne humeur du labo semble être en de très bonnes mains.

Merci à Marie-Pierre pour sa bonne humeur et joie de vivre qui la caractérise. Tu es pour beaucoup dans la bonne ambiance qui règne dans ce labo. Merci également à Emilie pour son sourire et pour son émerveillement perpétuel (même devant le Zénith de Caen, c’est dire...). Merci à Stéphane d’être toujours abordable et souriant. Un grand merci également à Eric que j’avais connu à l’ENSAI mais que j’ai redécouvert en tant que collègue, chaleureux, passionné et toujours prêt à aider. Merci à Maud, Sarah, Julien (toujours pas reçu d’invitation pour le mariage, satanés postiers!), Tristan, Paul... Un mot pour les anciens également qui ont marqué mes premiers mois à l’Agro : Stevenn, Aurore, Ophélie, Jean-Benoit, Antoine, Caroline, Alice...

De gros remerciements aux personnes que j’ai pu croiser en séminaire. A commencer par toutes les personnes qui ont vécu l’une et/ou l’autre des aventures à Rochebrune. Car aventures est le mot, je ne vais en mentionner que deux. Un grand merci à Nicolas Eckert pour son aide lors de l’épopée “restitution de chaussures de location” et une remarque d’ordre général : attention, la couleur rouge n’est pas une clé primaire dans la base de données des chaussures de ski, et si le lendemain d’avoir enflammé le dancefloor de la boom sur du YMCA vous ne trouvez pas mieux, demandez de l’aide... Ce même Nicolas était également présent deux ans après lors de l’opération “A la recherche du téléphone perdu dans la neige, en mode vibreur et avec une coque grise”(sans doute bientôt adaptée au cinéma). Même si cette opération (**Spoiler Alert**) ne fut pas couronnée de succès, la Team Priscilla m’a apporté une aide qui m’a fait chaud au coeur. Je profite d’ailleurs de cette tribune qui m’est offerte pour passer un message : si jamais vous avez trouvé ce téléphone, merci de me le rendre.

Merci à Marc que j’ai croisé quelques fois, mais cela a suffi pour bien sympathiser et pour voir en toi un supporter du PSG qui ne s’assume pas. Merci également à Romain : #thxmate, #laurens, #audessuscestleciel, #entreemaritine, #hashtag. Et sache qu’on ne t’en veut pas pour la finale de la coupe Davis.

En vrac des remerciements pour Francine P., Sophie B., Irène D., Philippe N., Pradeebane V. , Ara A., Pierre G., Sylvain L., Merlin K., Sophie A, Guillaume D., les petits

anges, les amis Alsaciens et Zlatan.

Un grand merci également à la team ENSAI, qui a toujours été là, dans les bons comme dans les moins bons moments. Cela fait déjà trois ans que nous avons quitté l'école (il parait que c'est un cap), difficile à imaginer tant la passion semble intacte et avoir encore de longues années devant elle. Merci à tous pour ce que vous êtes et ce que vous m'apportez, ne changez surtout pas. Des remerciements aussi pour les professeurs de l'ENSAI, en particulier Myriam Vimond, qui, après toutes ces années, est encore à l'écoute et toujours prête à nous proposer son aide.

De gros remerciements aussi pour ma famille élargie et en particulier mes parents qui m'ont toujours soutenu dans tout ce que j'entreprenais, sans jamais questionner mes choix, et qui sont pour beaucoup dans chacune de mes réussites. Pour cela je ne pourrais jamais vous remercier assez.

Et puis, bien sûr, le mot de la fin pour Delphine, mon amour, auprès de qui je me sens si bien. Merci pour tout ce que tu fais pour moi, pour nous, depuis toutes ces années et pour celles qui nous attendent. Durant ces trois ans, il est arrivé que les méandres de la vie aient mis notre couple à rude épreuve, mais nous avons su nous en relever et en ressortir plus fort, plus que jamais prêts à faire face aux événements que nous réserve l'avenir. Aussi extrêmes soient-ils...

Table des matières

I Introduction **11**

1 Contexte et objectifs **13**
1.1 Contexte et état de l'art 13
1.2 Objectifs 16

2 Données de précipitations **19**
2.1 Modélisation des précipitations 19
2.2 Région d'étude : le Sud-Est de la France 20

3 Théorie des valeurs extrêmes **25**
3.1 Cadre univarié 25
3.2 Cadre multivarié 28
3.3 Processus max-stables 31

4 Scores et critères d'évaluation **37**
4.1 Scores propres 38
4.2 Scores univariés 38
4.3 Critères de cohérence spatiale 40
4.4 Comparaison de distributions 41
4.5 Skill-Scores 42

II Simulation conditionnelle **43**

5 Introduction **45**

6 Simulation de processus max-stables (Article) **49**

7 Algorithme de tirage des partitions **79**
7.1 Méthodologie 79
7.1.1 Échantillonneur de Gibbs 79
7.1.2 Modèle des feuilles mortes 80

7.2 Application à des données simulées 85
7.2.1 Méthodologie 85
7.2.2 Validation du tirage des partitions 85
7.3 Conclusion 88

8 Conclusions de la partie II **89**

III Downscaling spatial des extrêmes **91**

9 Introduction **93**

10 Ajout d'information à grande échelle **97**
10.1 Présentation des données 98
10.2 Modélisation 98
10.2.1 Résumer l'information disponible 100
10.2.2 Modèles de transformation des marginales 100
10.2.3 Régimes automnaux 102
10.3 Simulations conditionnelles 106
10.4 Simulations non conditionnelles 108
10.5 Conclusion 109

11 Méthode hybride de downscaling spatial (Article) **113**

12 Conclusions de la partie III **143**

IV Conclusions et perspectives **145**

13 Conclusions **147**

14 Perspectives **151**

Bibliographie **155**

Annexes **167**

Résumé **173**

Abstract **173**

Première partie

Introduction

Chapitre 1

Contexte et objectifs

1.1 Contexte et état de l'art

Depuis l'Antiquité, l'être humain n'a eu de cesse que d'essayer de comprendre les phénomènes météorologiques. De par leurs conséquences directes sur la vie quotidienne des personnes, ces manifestations de la nature ont toujours été une préoccupation importante. Ces phénomènes inexpliqués ont été attribués au cours de l'histoire à diverses forces supérieures ou autres divinités. Les temps ont changé et les avancées scientifiques ont permis de perfectionner la compréhension du monde qui nous entoure. Néanmoins, encore au XXIème siècle, le rôle de ces phénomènes météorologiques sur nos vies reste capital. La compréhension et désormais la prévision de tels phénomènes est un objectif important de notre société. Ces événements ont des impacts majeurs sur l'agronomie, la consommation d'énergie, l'économie, les transports ou encore les performances sportives des athlètes, et leur modélisation apparaît comme un outil d'aide à la décision incontournable.

Toutes ces considérations sont encore amplifiées lorsque l'on s'intéresse aux événements climatiques extrêmes. De par leurs conséquences immédiates et leur dangerosité évidente, leur compréhension apparaît comme étant un enjeu sociétal et économique incontournable. Néanmoins, ces événements extrêmes sont par nature rares et inattendus et de ce fait la société est mal préparée à y faire face. Ainsi, des efforts grandissants sont menés pour comprendre et prévoir de tels phénomènes. D'un point de vue probabiliste, le cadre de travail utilisé depuis des décennies repose sur la théorie statistique des valeurs extrêmes. Si ces outils sont adaptés à la modélisation de processus univariés stationnaires (Beirlant et al., 2004), ils deviennent beaucoup moins appropriés en dehors de ce cadre de travail. Ainsi, depuis quelques années, des outils sont développés pour modéliser des extrêmes multivariés dans un contexte où la non stationnarité, spatiale et temporelle, devient plus la norme que l'exception (e.g., Panagoulia et al., 2014; Jonathan et al., 2014). En effet, l'hypothèse de stationnarité spatiale ou temporelle, qui suppose que les

caractéristiques du phénomène restent insensibles aux déplacements respectivement dans l'espace et dans le temps, n'est pas adaptée lorsque l'on s'intéresse à des régions aux spécificités hétérogènes ou lorsque l'on se place dans un cadre de changement climatique. En particulier dans ce contexte, imaginer que les modèles stationnaires calibrés sur le passé puissent être performants pour la projection future à moyen terme est très souvent illusoire (IPCC, 2013). Ce constat semble encore plus évident, lorsque l'on cherche à modéliser des précipitations extrêmes, phénomènes qui nous intéressent exclusivement tout au long de ce travail.

L'étude du climat et de ses extrêmes doit évidemment reposer pour une part sur des observations. Si la couverture des stations météorologiques est de plus en plus complète, elle ne sera jamais suffisamment dense pour fournir des données de précipitations en tous points de la carte. Or, dans de nombreux domaines, comme par exemple en entrée des modèles hydrologiques d'impact ou encore lorsque la collecte d'information est longue et coûteuse (e.g. en paléoclimat), il y a un vrai besoin d'obtenir des valeurs à des endroits où il n'y a pas forcément de point de mesure. Ainsi, des outils sont nécessaires pour interpoler les données observées et obtenir des valeurs aux emplacements géographiques vides d'observation. Ce genre de problématiques, très répandu en géostatistiques, peut être résolu par des techniques d'interpolation linéaire (Chilès and Delfiner, 1999). Néanmoins, ces méthodes ont tendance à lisser les valeurs obtenues et à trop se rapprocher des données observées, ce qui peut être problématique dans notre contexte. En effet, pour fournir des séries d'extrêmes réalistes, on aimerait autoriser une grande variabilité dans les valeurs obtenues, afin de pouvoir représenter des événements de faible probabilité. Pour rendre compte de cette variabilité, les simulations conditionnelles semblent plus adaptées (Lantuéjoul, 2002). Ces techniques permettent de générer par simulations des champs de réalisations, ayant une certaine structure spatiale dépendant du modèle choisi, et respectant les observations aux points conditionnants. Ces méthodes sont détaillées dans la partie II.

Cependant, pour pouvoir utiliser ces techniques de simulation conditionnelle, il faut avoir des observations disponibles. Or dans plusieurs cas de figure, il est utile de pouvoir générer des valeurs en absence d'observation. Par exemple, que cela soit pour reconstituer le passé lointain (e.g. paléoclimat) ou dans un contexte de projections futures, le conditionnement par des observations n'est pas possible. Néanmoins, d'autres outils restent disponibles dans ce cadre de travail : les sorties de modèles climatiques. En effet, les modèles de circulation générale (GCMs) sont des modèles déterministes basés sur des résolutions d'équations différentielles de la dynamique atmosphérique. Ils fournissent des valeurs pour toutes sortes de variables climatiques (température, pression, précipitation, etc...) à des horizons de temps pouvant être très lointains dans le passé (e.g., Joussaume et al., 1999; Dee et al., 2013) et dans le futur (IPCC, 2013). Ils sont

utilisés pour décrire les tendances générales possibles des décennies à venir, en fonction de différents scénarios de changement climatique. Le problème majeur des GCM est la résolution spatiale très faible des sorties ($\approx$ 200km) qui ne permet pas d'avoir des informations assez précises sur le comportement local des phénomènes étudiés, pour, par exemple, leur utilisation en entrée des modèles d'impact.

Afin de pallier à ce problème et d'obtenir des valeurs à une résolution plus fine, des techniques dites de « descente d'échelle » ou « downscaling » peuvent être appliquées aux sorties des GCMs. Le principe général du downscaling est de générer des valeurs climatiques ou météorologiques à une échelle locale, à partir d'informations provenant d'une échelle plus grande. L'objectif est ainsi d'obtenir les conséquences locales des différents scénarios fournis par les sorties de modèles climatiques globaux. Pour ce faire, il existe deux approches distinctes de descente d'échelle, le downscaling dynamique et le downscaling statistique.

Un downscaling dit « dynamique » est appliqué à ces sorties pour construire des Regional Climate Models (RCMs), qui peuvent être vus comme la version régionalisée des GCMs (e.g., Laprise, 2008; Rummukainen, 2010). Ces modèles régionaux sont forcés aux bords de leur domaine par les valeurs des GCMs, mais ont leur propre dynamique à l'intérieur de ces contraintes. Une version dite « nudgée » de ces modèles admet en plus des contraintes de rappel au centre du domaine de forçage, pour éviter que les RCMs ne s'éloignent trop des GCMs dont ils sont issus. Avec les RCM, la résolution spatiale a été grandement améliorée (de 50km à 5km en fonction des modèles), mais peut s'avérer encore trop faible lorsque l'on cherche à modéliser des phénomènes pouvant être très localisés (e.g. des orages). De plus, les grandes ressources informatiques et le temps de calcul requis limitent l'utilisation de ces techniques. Néanmoins, des efforts grandissants sont menés dans ce sens avec notamment l'initiative CORDEX[1] (Coordinated Regional Climate Downscaling Experiment) qui vise à inter-comparer différents modèles de RCMs et à standardiser leur évaluation (e.g. Vaittinada Ayar et al., 2014).

Depuis quelques années, de nouvelles techniques de downscaling, dites « statistiques » , sont de plus en plus utilisées (Maraun et al., 2010). Elles consistent à établir un lien statistique entre des variables ayant des résolutions spatiales différentes. Comme ces techniques sont applicables pour n'importe quelle résolution spatiale, dans la pratique cette relation est souvent établie entre des sorties de modèles climatiques (GCMs ou RCMs) et une variable observée localement. Ce terme de downscaling statistique regroupe un grand nombre de modèles puisque la plupart des outils statistiques classiques reliant deux variables entre elles peut être utilisée. Ces méthodes statistiques ont pour avantage d'être en général rapides d'utilisation et de permettre de fournir une incertitude sur les estimations obtenues. Les grandes lignes de ces méthodes sont abordées dans la

[1] http ://wcrp-cordex.ipsl.jussieu.fr/

partie III.

1.2 Objectifs

Différents objectifs sont poursuivis tout au long de ce travail. Dans un premier temps, nous cherchons à évaluer les risques à certains endroits pour lesquels il n'y a pas de mesures, en y générant des distributions de valeurs extrêmes en tenant compte de l'information donnée par les observations environnantes. Ainsi dans la pratique, on cherche à fournir un algorithme de simulation conditionnelle de processus max-stables, afin de pouvoir générer des distribution de maxima en des points vides d'observations. Par la suite, nous cherchons à fournir des distributions de valeurs extrêmes dans des régions où il n'y a pas d'observations disponibles (par exemple dans le futur mais pas nécessairement) mais de l'information provenant de scénarios à une très faible résolution spatiale. Dans cette optique, un premier pas est effectué en incorporant de l'information à grande échelle dans la procédure de simulation conditionnelle. En prenant en compte des variables climatiques à grande échelle qui évoluent dans le temps, cette approche peut permettre de rendre évolutif le processus simulé. On peut parler alors de simulation conditionnelle non stationnaire temporellement.

Par la suite, après avoir calibré le modèle sur des observations passées, on cherche, dans un contexte de downscaling, à pouvoir simuler un processus max-stable non stationnaire pour des projections futures, en se basant uniquement sur des sorties de modèles climatiques. On peut parler ainsi de simulations non conditionnelles, dans le sens où elles ne sont pas conditionnées par des observations, mais par de l'information à grande échelle.

Un autre objectif important est, une fois les buts précédemment énoncés remplis, de fournir à la communauté scientifique, sous forme d'un package R, les programmes permettant d'y parvenir pour rendre ces techniques utilisables par le plus grand nombre.

Le reste de cette introduction présente succinctement certains outils qui seront utilisés tout au long de ce travail. Ainsi dans le chapitre 2, la présentation de la nature des données de précipitations montre la nécessité d'avoir recours au cadre statistique de la théorie des valeurs extrêmes. Par la suite, quelques notions de ce cadre de travail sont rappelées chapitre 3 avec des illustrations des différents outils univariés et multivariés disponibles pour modéliser des événements extrêmes. Pour finir cette introduction, le chapitre 4 décrit les différents outils d'évaluation des méthodes qui sont utilisés tout au long de ce travail.

La suite du document se décompose en trois grandes parties. Dans la partie II, des outils de simulation conditionnelle sont présentés chapitre 5. Par la suite, l'article présenté dans le chapitre 6 détaille un algorithme de simulation conditionnelle de processus max-

stables, qui est étudié et appliqué à des données simulées puis à données réelles de précipitations. La partie II s'achève chapitre 7 avec une méthode alternative pour l'étape la plus compliquée et coûteuse de l'algorithme de simulation conditionnelle et leur comparaison sur des données simulées.

La partie III introduit des notions d'information issues de modèles climatiques et de downscaling au chapitre 9. Le chapitre 10 décrit une méthode permettant d'incorporer de l'information grande échelle dans l'algorithme de simulation conditionnelle afin de rendre la procédure utilisable également pour des prédictions futures d'extrêmes de précipitations. Par la suite, l'article figurant chapitre 11, une approche hybride combinant downscaling statistique, dynamique et procédure de simulation conditionnelle est présentée et appliquée sur les données de précipitations du sud-est de la France.

Enfin, la partie IV résume les différents résultats obtenus et énonce des perspectives futures d'application et d'amélioration des diverses méthodes proposées et utilisées tout au long de ce travail. Les annexes A détaillent certains scores obtenus chapitre 11 et les annexes B certains calculs particuliers.

Chapitre 2

Les précipitations : des données pas comme les autres

2.1 Modélisation des précipitations

Dans tout ce travail, nous nous intéressons aux données de précipitations. Ces données ont des caractéristiques qui leurs sont propres et qui rendent leur modélisation différente de celle d'autres variables climatiques. L'un des aspects particuliers de ces observations est leur grande proportion de zéros (voire de valeurs très faible), du moins pour des pas de temps assez courts (e.g. journaliers). Néanmoins, lorsque des événements pluvieux apparaissent, d'importantes précipitations peuvent être observées en très peu de temps. Ainsi, la distribution des observations en un site donné est fortement asymétrique, avec une queue de distribution pouvant être épaisse.

D'un point de vue multivarié, ces données admettent une variabilité spatiale très élevée. A l'inverse, par exemple, de données telles que la température qui sont relativement continues et régulières, même sur plusieurs dizaines de kilomètres, les précipitations peuvent être très hétérogènes même à courte distance. Cette variabilité spatiale est encore exacerbée lorsque l'on s'intéresse aux événements extrêmes puisque les pluies accompagnant les épisodes orageux peuvent être d'une très grande intensité et très localisées. Ainsi dans tout ce travail, la notion de résolution spatiale sera capitale et des efforts seront produits pour fournir des outils permettant une modélisation spatiale la plus précise possible.

Pour décrire des précipitations journalières y à un site donnée, Williams (1998) propose une formulation de la densité pour modéliser simultanément l'occurrence et l'intensité des précipitations :

$$\phi(y, \psi_o, \alpha) = (1 - \alpha)\delta(y) + \alpha\phi_0(y; \psi_o) \tag{2.1}$$

avec α la probabilité d'occurrence le jour étudié, δ la distribution de Dirac et ϕ_0 une

densité de probabilité de paramètre ψ_o décrivant l'intensité de la pluie. Dans cette écriture, le premier terme a pour rôle de modéliser les jours secs alors que le second s'occupe des jours pluvieux.

Initialement appliquée avec une loi Gamma (Williams, 1998), le choix de la distribution ϕ_0 est déterminant dans cette modélisation. En effet, d'autres fonctions de densité sont possibles comme par exemple une loi log-normale ou une loi Gaussienne (Carreau and Vrac, 2011). Néanmoins, pour modéliser des précipitations extrêmes, ce type de lois à queue légère peut s'avérer inapproprié. Ainsi, cette modélisation a été adaptée à des lois à queue lourde telles que des mélanges d'exponentielles, Gamma-GPD ou normale-GPD (e.g. Vrac and Naveau, 2007; Hundecha et al., 2009; Carreau and Vrac, 2011).

Cependant, ce genre de modèle ne sera pas utilisé dans ce travail, puisque nous nous intéressons aux maxima automnaux de précipitations dans une zone où les pluies intenses sont régulières. Ainsi, la possibilité de modéliser simultanément les occurrences et les intensités des précipitations ne présente pas un grand intérêt pour notre problématique.

2.2 Région d'étude : le Sud-Est de la France

Toutes les applications de ce travail se concentreront sur les précipitations extrêmes dans le sud de la France. Ainsi, nous nous intéressons aux précipitations couvrant une partie des régions du Languedoc-Roussillon et du Gard.

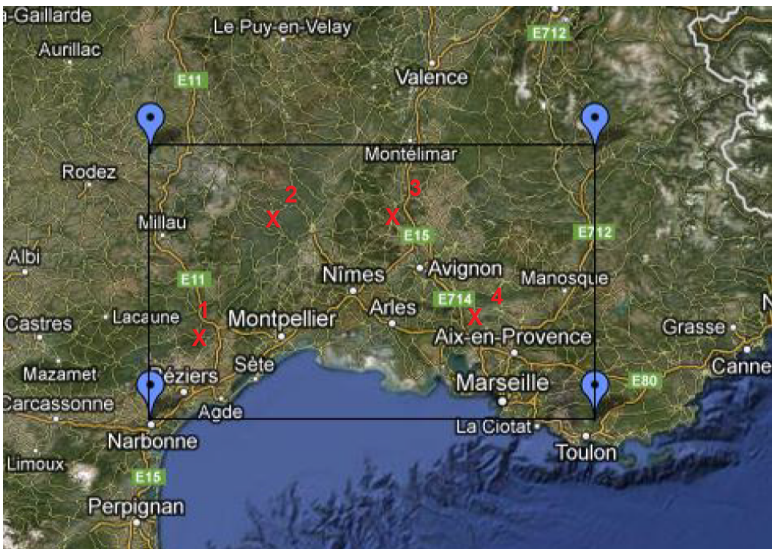


FIGURE 2.1 – Description de la zone d'étude.

La zone d'étude [3.4°E, 6.0°E] x [43.2°N, 44.5°N], de la Figure 2.1, est une région où apparaissent les événements dits cévenols entraînant de très fortes précipitations, le

plus souvent en période automnale. En effet, le massif des Cévennes est propice à la création d'orages violents qui éclatent sur notre région d'étude, entraînant de très fortes pluies et des crues. Cette région est couramment utilisée pour l'étude des précipitations extrêmes (e.g. Carreau et al., 2013; Flaounas et al., 2013).

La Figure 2.2 présente les histogrammes des précipitations journalières en quatre sites choisis arbitrairement dans cette zone géographique (et notés avec des croix rouges sur la Figure 2.2). Ces précipitations sont issues des réanalyses SAFRAN (Quintana-Seguí et al., 2008; Vidal et al., 2010) qui représentent les précipitations journalières sur la période $1960 - 2007$, gridées à une résolution spatiale de 8 km x 8 km. Dans tout ce travail, nous utilisons ces données comme des observations.

Par souci de lisibilité, les valeurs nulles qui sont fortement majoritaires ne sont pas présentées Figure 2.2.

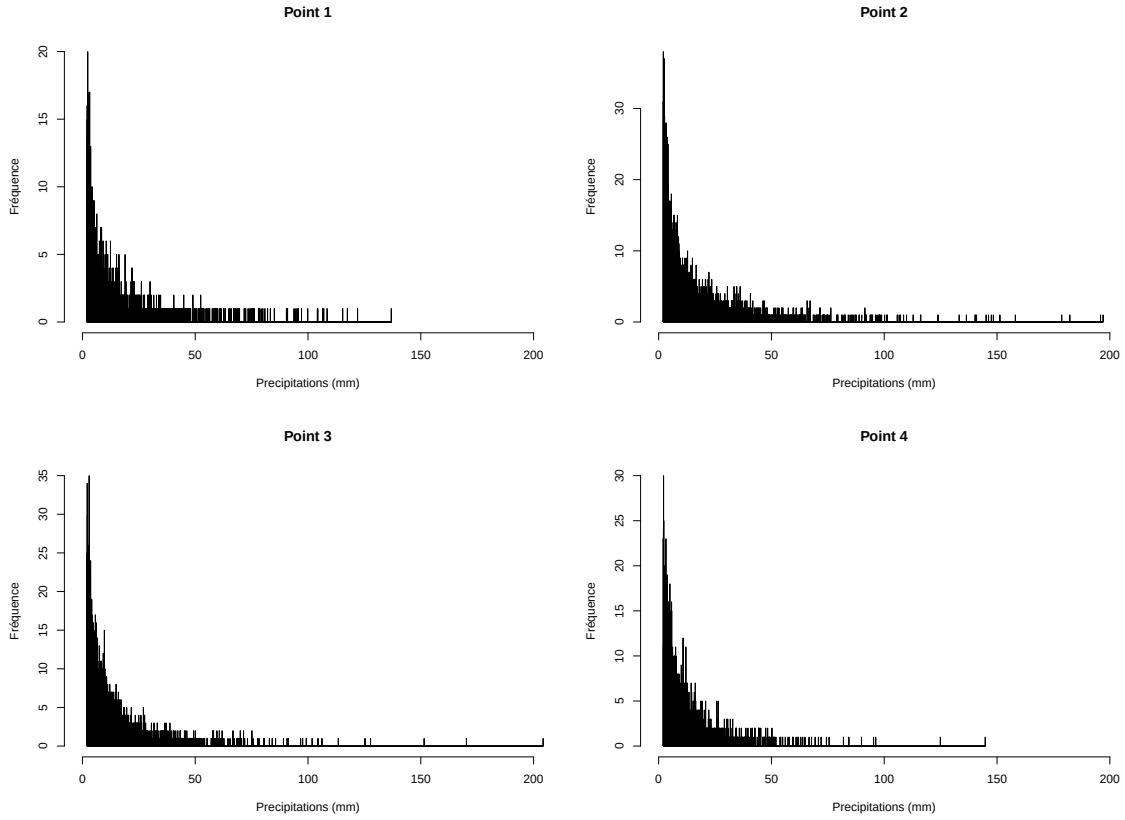


FIGURE 2.2 – Histogrammes des précipitations en quatre points de la zone d'étude.

Il apparaît que les données de précipitations journalières présentent un grand nombre de très faibles valeurs. Néanmoins, on constate également que les queues de distribution de ces histogrammes sont épaisses et autorisent ainsi des valeurs pouvant être très élevées.

Dans tout ce travail, nous nous intéressons aux extrêmes et plus particulièrement aux maxima annuels de précipitations. Dans cette optique, la Figure 2.3 montre les distributions des maxima annuels (sur la période $1960 - 2007$) aux mêmes sites que précédemment.

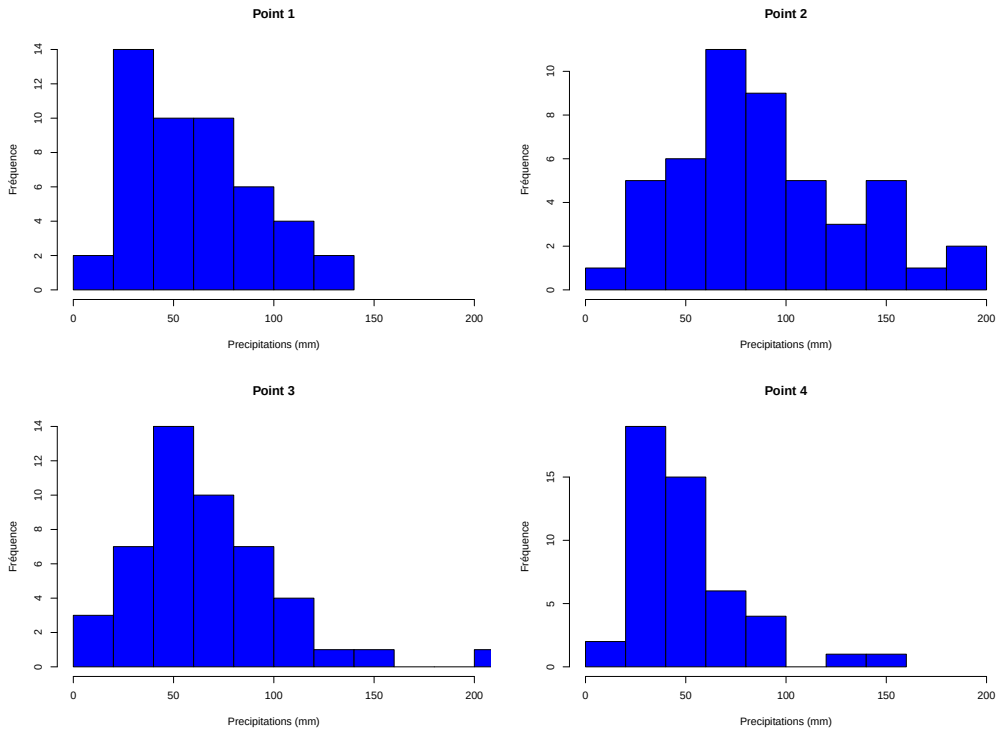


FIGURE 2.3 – Distribution des précipitations maximales annuelles en quatre points de la zone d'étude.

Cette figure construite à partir des maxima met en exergue l'épaisseur de la queue de distribution et l'importance d'en tenir compte dans la modélisation de ce phénomène. Il nous faut utiliser des lois ayant des queues de distributions à droite pouvant s'éloigner beaucoup de la valeur médiane de la loi. Ainsi, le cadre statistique classique ne semble pas adapté à ce genre de problématique et on a recours dans ce travail à la théorie statistique des valeurs extrêmes, dont certaines notions sont rappelées dans le prochain chapitre.

De plus, cette figure met également en évidence les différences dans les distributions des marginales qu'il peut exister d'un point à l'autre de la région. Cette non stationnarité spatiale sera également à prendre en compte dans le choix des outils de modélisation.

Pour illustrer la dépendance spatiale qui existe entre les maxima d'une même année, la Figure 2.4 présente la carte des maxima de l'année 1990. On peut identifier certaines structures de dépendance à moyenne distance. Il est donc important de choisir dans

la modélisation des précipitations extrêmes des outils capables de retranscrire cette dépendance.

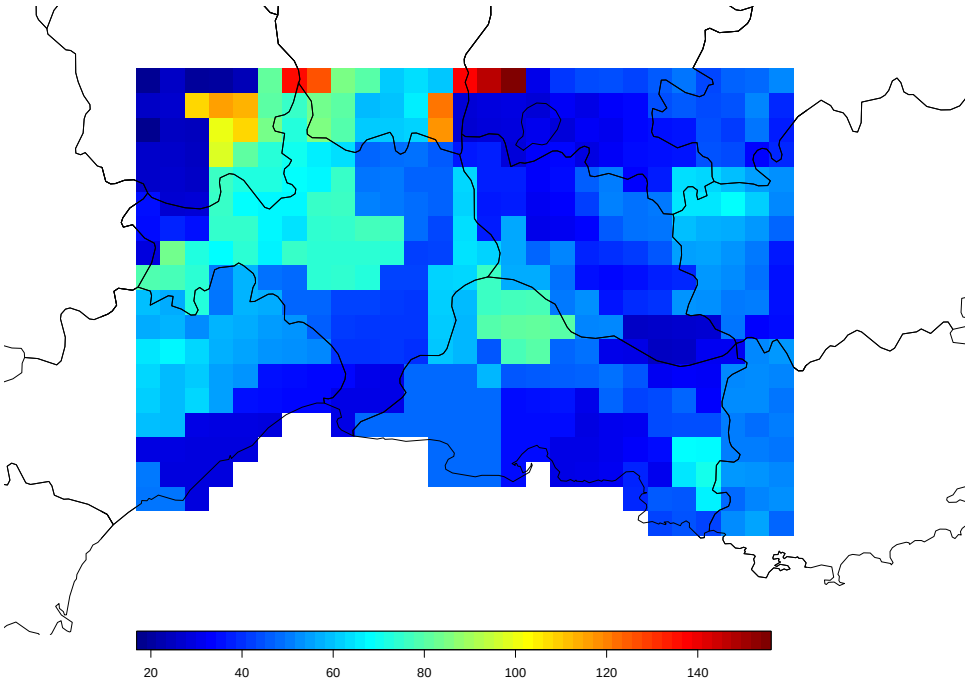


FIGURE 2.4 – Carte des maxima de l'année 1990 (données SAFRAN).

Chapitre 3

Théorie des valeurs extrêmes

L'introduction de la théorie des valeurs extrêmes est devenue une nécessité dans de nombreux domaines. En effet, lorsque l'on cherche à modéliser des événements extrêmes, plusieurs problèmes se posent avec le cadre probabiliste habituel qui est en grande partie développé sous l'hypothèse gaussienne. Le premier problème vient du fait que dans le monde des extrêmes, les phénomènes décrits n'admettent plus forcément de moments finis. Ainsi de nombreux outils statistiques habituels deviennent inutilisables, en particulier ceux reposant sur le théorème centrale-limite ou la loi forte des grands nombres. La deuxième difficulté majeure vient du fait que les événements extrêmes sont également rares. De cette rareté vient la nécessité de modéliser des événements avec des probabilités très faibles et il peut arriver que certains événements à modéliser n'aient jamais été observés. Ainsi avec les outils classiques, l'estimation de la probabilité d'un événement plus extrême que ceux déjà observés risque très souvent d'être nulle. En fonction de la nature des événements extrêmes considérés, cette sous-estimation peut s'avérer très dangereuse. Il y a donc la nécessité d'utiliser des modèles capables d'extrapoler et pas seulement de reproduire les phénomènes observés.

Les outils issus de la théorie probabiliste des valeurs extrêmes permettent de répondre à ces problèmes et sont adaptés à la modélisation de ces phénomènes particuliers. Ce chapitre permet d'en rappeler quelques notions.

3.1 Cadre univarié

En univarié, modéliser des événements extrêmes revient à chercher la distribution des quantiles extrêmes d'une variable aléatoire X. Plus précisément lorsque l'on s'intéresse aux maxima d'un certain nombre de réalisations, cela équivaut à trouver la distribution de $X_{(n)} = \max_{i=1,\dots,n} X_i$, avec $X_1, \dots, X_n$ des réalisations indépendantes de X. Les lois max-stables sont des outils souvent utilisés pour ce genre de problématiques. Ces lois ont la particularité d'être invariantes à une transformation affine près par le passage au

max. En d'autres termes, la loi du maximum, correctement normalisée, de plusieurs lois max-stables restera de même loi.

Une définition plus formelle possible des lois max-stables est donnée par

Définition 1. *La loi $\mathcal{H}_0$ est dite max-stable si pour tout $n \geq 2$, $(X_1, \ldots, X_n)$ étant des variables aléatoires indépendantes de loi $\mathcal{H}_0$, il existe deux suites $(a_n)_n \in \mathbb{R}$ et $(b_n)_n \in \mathbb{R}^+$, tels que $\frac{X_{(n)}-a_n}{b_n}$ suit la loi $\mathcal{H}_0$.*

En outre, si $\frac{X_{(n)}-a_n}{b_n}$, avec X_i de loi quelconque, converge en loi vers une limite non triviale, c'est-à-dire vers une variable aléatoire non constante, alors la loi limite est une loi max-stable. Le théorème suivant permettant d'identifier toutes les lois max-stables et la propriété qui en découle, est le théorème fondamental de la théorie des valeurs extrêmes.

Théorème 1. *Si la suite $\frac{X_{(n)}-a_n}{b_n}$ converge vers une loi limite non triviale, alors à une translation et un changement d'échelle près, la fonction de répartition de la limite est de la forme suivante :*

$$\textit{Loi de Weibull :}\ \Psi_\alpha(x) = \begin{cases} e^{-(-x)^\alpha} & \textit{si } x \leq 0 \\ 1 & \textit{si } x > 0 \end{cases} \quad \textit{pour } \alpha \geq 0$$

$$\textit{Loi de Gumbel :}\ \Lambda(x) = e^{-e^{-x}}\ \textit{pour } x \in \mathbb{R}$$

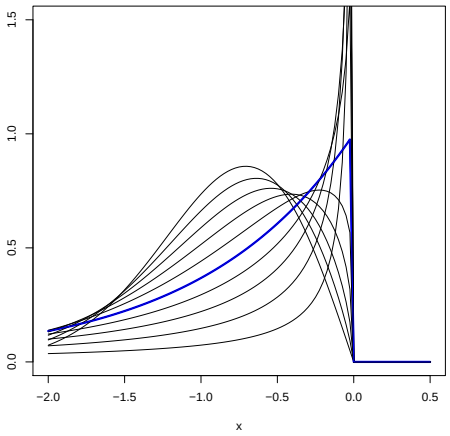
$$\textit{Loi de Fréchet :}\ \Phi_\alpha(x) = \begin{cases} 0 & \textit{si } x \leq 0 \\ e^{-x^{-\alpha}} & x > 0 \end{cases} \quad \textit{pour } \alpha \geq 0.$$

L'ensemble des lois limites s'obtient en considérant $cX + d$ avec X une loi de Weibull, Gumbel ou Fréchet. Toutes les lois ainsi formées sont des lois max-stables. Une autre écriture de ces lois max-stables est donnée par la propriété suivante :

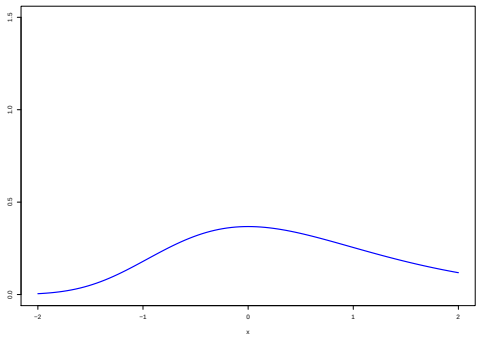
Propriété 1. *La fonction de répartition limite du **Théorème 1** est nécessairement une distribution « Generalized Extreme Value » (GEV) qui s'écrit :*

$$H(x) = \begin{cases} exp\Big(-\Big[1+\xi\Big(\dfrac{x-\mu}{\sigma}\Big)\Big]^{-1/\xi}\Big) & \textit{si } \xi \neq 0 \\ exp\Big(-exp\Big(\dfrac{x-\mu}{\sigma}\Big)\Big) & \textit{si } \xi = 0. \end{cases} \tag{3.1}$$

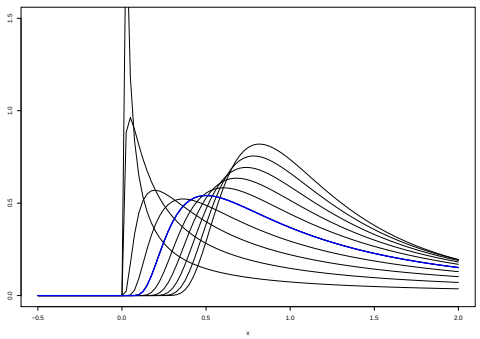
Si $\xi < 0$, alors H est dans le domaine d'attraction de Weibull, si $\xi = 0$, le domaine d'attraction de Gumbel et si $\xi > 0$, celui de Fréchet.



(a) Distributions de Weibull pour différents paramètres de formes ($\xi \in [-5, -0.5]$; en bleu $\xi = -1$).



(b) Distribution de Gumbel ($\xi = 0$).



(c) Distributions de Fréchet pour différents paramètres de formes ($\xi \in [0.5, 5]$; en bleu $\xi = 1$).

FIGURE 3.1 – Distributions GEV standards en fonction du paramètre de forme.

Ainsi, la loi max-stable est en partie déterminée par le paramètre ξ dit de forme, plus sa valeur sera élevée et plus la queue de distribution sera épaisse. La Figure 3.1 montre la forme des trois types de lois GEV standards (avec $\mu = 0$ et $\sigma = 1$) en fonction des différentes valeurs du paramètre de forme. On remarque que pour $\xi < 0$ la distribution est bornée supérieurement, pour $\xi = 0$ la queue de distribution est légère et pour $\xi > 0$ elle est lourde.

A noter que dans ce travail, nous nous intéressons aux maxima sur une période donnée et c'est pour cette raison que nous utilisons des loi max-stables et que nous obtenons des distributions appartenant à la famille GEV. L'autre axe majeur d'étude de la théorie des valeurs extrêmes consiste à s'intéresser à la loi des dépassements. Dans cette optique, ce ne sont plus des maxima sur des blocs qui sont considérés mais les observations dépassant un seuil donné. La loi ainsi obtenue appartient à la famille des « Generalized Pareto distributions » (GPD) qui permet de calculer des quantiles extrêmes et des niveaux de retour. Ce type de loi ne sera pas abordé dans ce travail. Pour plus de détails voir par exemple Beirlant et al. (2004).

Toutes les notions évoquées dans cette section sont relatives à des lois univariées. Or, notre objectif étant de fournir des champs d'extrêmes ayant une certaine structure spatiale, ce cadre univarié n'est pas suffisant. La prochaine section détaille des lois multivariées d'extrêmes dont les marginales appartiennent aux familles de lois univariées décrites précédemment.

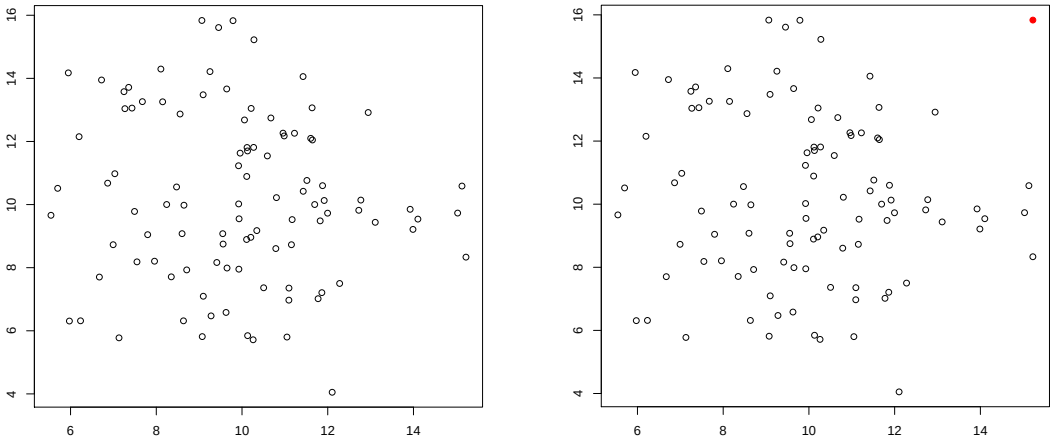
3.2 Cadre multivarié

Dans le cadre univarié, le but est de caractériser le quantile extrême d'une variable aléatoire donnée. La théorie des extrêmes multivariés repose toujours sur ce principe, mais désormais nous nous intéressons au vecteur des maxima par composant. Ainsi, dans le cas bivarié, en notant $X_{(n)} = \max_{i=1,\dots,n} X_i$, avec $X_1, \dots, X_n$ des réalisations indépendantes de X et $Y_{(n)} = \max_{i=1,\dots,n} Y_i$, avec $Y_1, \dots, Y_n$ des réalisations indépendantes de Y, la théorie des extrêmes multivariés est basée sur l'étude de $M_{(n)} = (X_{(n)}, Y_{(n)})$. Il est important de remarquer que le i pour lequel la composante X atteint son maximum n'est pas forcément la même que celui de la composante Y. Ainsi, une réalisation du vecteur $M_{(n)}$ n'est pas nécessairement un couple de points observés simultanément.

La Figure 3.2 montre un exemple de la manière dont est construit le vecteur $M_{(n)}$ bivarié. On remarque que dans ce cas, M_n ne correspond pas à un point observé.

A ce stade, la normalisation de $M_{(n)}$ s'effectue comme en une dimension et le Théorème 1 peut être adapté en multivarié :

Théorème 2. *Si il existe deux suites $(a_n)_n \in \mathbb{R}$ et $(b_n)_n \in \mathbb{R}^+$, tels que $\frac{M_{(n)} - a_n}{b_n}$ converge en loi vers une limite non triviale $\mathcal{G}_0$, alors $\mathcal{G}_0$ est max-stable. De plus chacune de ses*



(a) Réalisations de (X, Y). (b) Ajout de $M_{(n)}$ en rouge.

FIGURE 3.2 – Exemple de construction de M_n en deux dimensions.

marginales est une loi univarié des extrêmes, c'est-à-dire qu'elles admettent chacune une distribution GEV (voir Propriété 1) mais pas nécessairement les mêmes.

On peut facilement transformer les marginales d'une distribution extrême bivariée en Fréchet, pour se ramener à un cadre de travail couramment utilisé. Ainsi, on note $\mathcal{G}_0^*$ une loi bivariée extrême de marginales Fréchet unité. Il apparaît que $\mathcal{G}_0^*$ admet une écriture générale donnée dans le théorème suivant.

Théorème 3. *Soit (X_i, Y_i) des vecteurs aléatoires dont les marges sont Fréchet units et $\left(X_{(n)}, Y_{(n)}\right)$ définis comme précédemment. Si :*

$$\mathbb{P}\left(X_{(n)} \leq x, Y_{(n)} \leq y\right) \xrightarrow{\mathcal{L}} \mathcal{G}_0^*(x, y) \tag{3.2}$$

où $\mathcal{G}_0^*$ est une distribution non dégénérée de marginales Fréchet unité, alors $\mathcal{G}_0^*$ est de la forme :

$$\mathcal{G}_0^*(x, y) = \exp\left(-V(x, y)\right), \quad x > 0, y > 0, \tag{3.3}$$

où $V(x, y) = \int_0^1 \max\left(\frac{\omega}{x}, \frac{1-\omega}{y}\right) dH(\omega)$ et H est une mesure finie sur $[0, 1]$ vérifiant les contraintes :

$$\int_0^1 \omega dH(\omega) = \int_0^1 (1 - \omega) dH(\omega) = 1 \tag{3.4}$$

Ainsi, toutes les lois bivariées d'extrêmes peuvent s'écrire de cette façon. A l'inverse du cadre univarié pour lequel il y a trois familles de lois d'extrêmes, la famille des lois

bivariées extrême est une famille non paramétrique qui ne dépend que des fonctions $V(.,.)$

L'autre notion fondamentale des lois multivariées est celle de la dépendance des extrêmes, à savoir comment les extrêmes de différentes variables sont liés entre eux. Deux indicateurs mesurant cette dépendance sont présentés dans la suite.

• Le « Upper Tail Dependence parameter » (Sibuya, 1959) noté χ est un premier paramètre caractérisant la dépendance asymptotique. Il s'écrit :

$$\chi = \lim_{x \to x_{F_1}} \mathbb{P}\left(Y > x | X > x\right) \tag{3.5}$$

avec (X, Y) de mêmes marginales F_1 et $x_{F_1} = \sup\{x : F_1(x) < 1\}$.

Ainsi, $\chi \in [0, 1]$, si $\chi = 0$ on dit que les variables X et Y sont asymptotiquement indépendantes et si $\chi > 0$ on dit que les variables X et Y sont asymptotiquement dépendantes.

De plus, même si les (X, Y) ont des marginales différentes, notées respectivement F_X et F_Y, le paramètre χ peut s'écrire

$$\chi = \lim_{u \to 1} \mathbb{P}\left(F_Y(Y) > u | F_X(X) > u\right). \tag{3.6}$$

On voit que χ est une mesure de la dépendance extrêmale de (X, Y) indépendante des marginales F_X et F_Y. La fonction $\chi(u)$ peut être interprétée comme une mesure de la dépendance au niveau du quantile d'ordre u. En particulier, le signe de $\chi(u)$ indique si les variables sont associées positivement ou négativement au niveau du quantile d'ordre u. Dans le cas de l'indépendance complète, $\chi(u) = 0$, alors que dans le cas de la dépendance parfaite, $\chi(u) = 1$.

De manière générale, si $\mathcal{G}_0$ est une distribution bivariée des extrêmes, la fonction $\chi(u)$ est constante. Pour une distribution bivariée des extrêmes avec dépendance on a :

$$\chi = 2 - V(1, 1) = \chi(u); \forall u \tag{3.7}$$

avec $V(.,.)$ désignant la même fonction que dans l'équation (3.3).

• L'autre outil permettant de mesurer la dépendance des extrêmes est le coefficient extrêmal (Schlather and Tawn, 2003). En prenant (X, Y) admettant des marginales identiquement distribuées, la fonction du coefficient extrêmal θ est définie telle que

$$\mathbb{P}(X < u, Y < u) = \left[\mathbb{P}(X < u)\right]^{\theta}. \tag{3.8}$$

$\theta \in [1, 2]$ mesure la force de la dépendance asymptotique, $\theta = 1$ correspond au cas de complète dépendance alors que $\theta = 2$ correspond au cas d'indépendance asymptotique.

Dans la suite, nous utiliserons exclusivement ce paramètre. Plus de détails quant à son utilisation sont fournis en fin de chapitre.

Dans tout ce travail, nous ne nous intéressons pas directement aux lois multivariées d'extrêmes mais à des processus d'extrêmes qui peuvent être vus comme des loi multivariées de dimension infinie. Les notions évoquées dans cette section sont cependant directement généralisables aux processus. La prochaine section présente quelques notions sur les processus d'extrêmes et détaille certains des plus couramment utilisés.

3.3 Processus max-stables

Soit $\{Z(x)\}_{x\in\mathbb{R}^d}$ ayant une distribution donnée $\mathcal{H}_0$ et $(Z_1,\ldots,Z_n)$ des copies indépendantes de Z. La Définition 1 peut être généralisée au cas des processus :

Définition 2. *Si il existe des suites $a_n(x)\in\mathbb{R}$ et $b_n(x)\in\mathbb{R}^+$ telles que pour tout $n\in\mathbb{N}$,*

$$\left\{\frac{\max_{i=1,\ldots,n} Z_i(x)-a_n(x)}{b_n(x)}\right\}_{x\in\mathbb{R}^d} \tag{3.9}$$

suit la distribution $\mathcal{H}_0$, alors $\{Z(x)\}_{x\in\mathbb{R}^d}$ est un processus max-stable (De Haan, 1984).

De plus, De Haan (1984) montre que tout processus max-stable peut s'écrire

$$\{Z(x)\}=\max_{n\geq 1}\{\xi_n W_n(x)\}. \tag{3.10}$$

avec W_n des champs aléatoires continus i.i.d. et ξ_n un processus de Poisson sur $\mathbb{R}^+$ qui admet comme intensité $t^{-(2)}dt$.

L'idée de cette formule est de considérer qu'en chaque point x, le processus max-stable est obtenu en prenant le maximum du produit entre W_n et ξ_n. La corrélation spatiale dépendra ainsi en grande partie de celle du champ aléatoire, alors que l'intensité, en particulier les événements extrêmes, est engendrée par des grandes valeurs du processus de Poisson. L'interprétation de cette formule en une dimension spatiale se retrouve Figure 3.3 sur laquelle chaque couleur est associé à un champ $W_n(x)$ et où le maximum de ces courbes, $\{Z(x)\}$, apparaît en gras. Pour chaque valeur de x, on prend le maximum du produit des valeurs des champs $W_n(x)$ et de celles des processus de Poisson. Il est important de remarquer que les ξ_i du processus de Poisson n'ont rien à voir avec le paramètre ξ de la distribution GEV introduite précédemment. Dans la suite, ces deux quantités ne sont plus évoquées simultanément pour éviter tout risque de confusion.

A ce stade, en fonction de la nature des champs aléatoires W_n, toute une batterie de processus max-stables peut être obtenue, dont certains vont être détaillés par la suite.

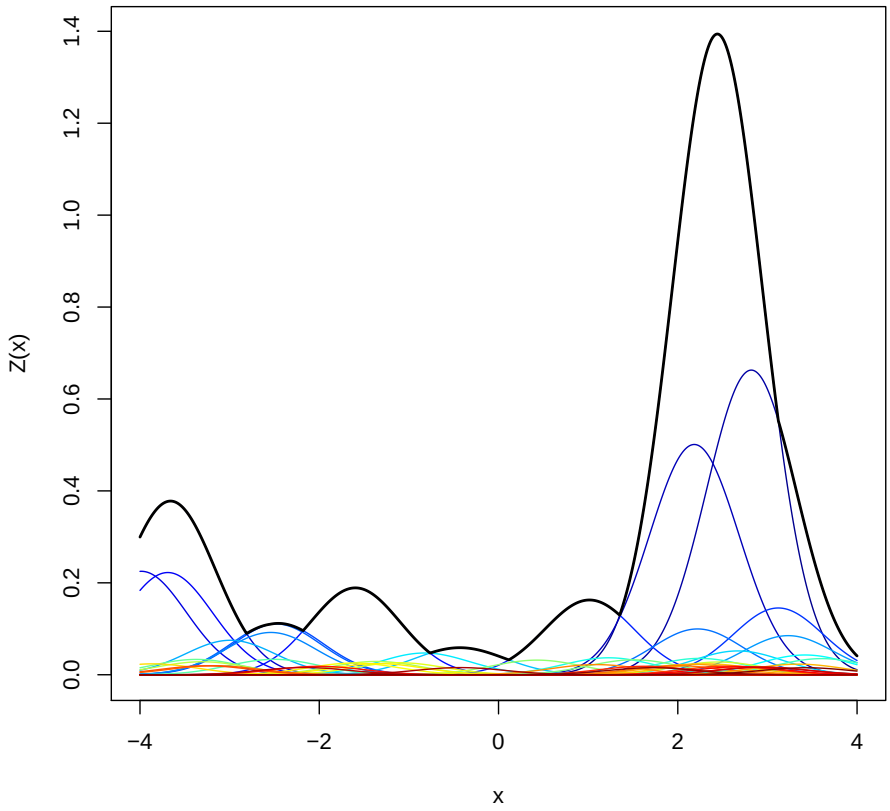


FIGURE 3.3 – Construction d'un processus max-stable en une dimension.

Processus de Smith

Le processus de Smith ou processus tempête (Smith, 1990), est défini par l'équation

$$\{Z(x)\} = \max_{i \geq 1} \{\xi_i f(x - U_i)\},\ x \in \mathcal{X} \subset \mathbb{R}^d, \text{ avec } \mathcal{X} \text{ un Borélien}, \tag{3.11}$$

avec ξ_i un processus de Poisson sur (0,+∞) d'intensité $t^{-2}dt$, U_i les réalisations d'un processus de Poisson homogène sur $\mathcal{X}$ et f la densité d'une loi normale centrée sur $\mathcal{X}$. Ce processus est appelé tempête car il peut être interprété de la façon suivante : les ξ_i définissent l'intensité de la tempête, U_i son centre et f sa forme. La fonction f est déterministe mais l'aléa provient de la translation des centres U_i.

La Figure 3.4 montre deux simulations de processus tempête dans $\mathbb{R}^2$. Les centres des tempêtes apparaissent clairement sur ces illustrations. Toutes les simulations de cette partie ont été obtenues grâce au package R `SpatialExtreme` (Ribatet, 2013b).

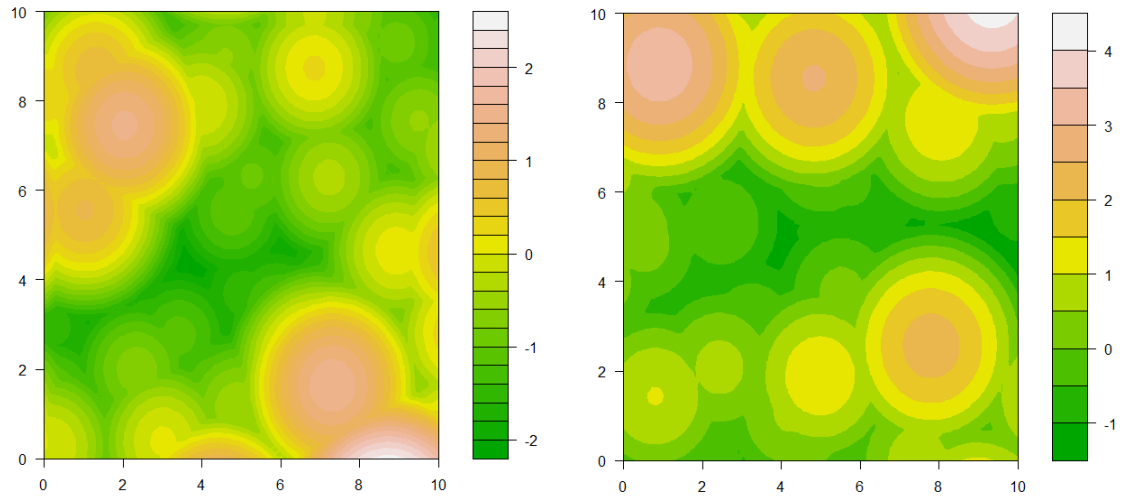


FIGURE 3.4 – Deux exemples de simulation de processus de Smith. Pour une meilleure visibilité, une échelle logarithmique a été utilisée.

Processus de Schlather

Le processus de Schlather (Schlather, 2002), également appelé Gaussien Extremal, est défini par l'équation

$$\{Z(x)\} = \sqrt{2\pi} \max_{i \geq 1} \{\xi_i W_i(x)\},\ x \in \mathcal{X} \tag{3.12}$$

avec ξ_i un processus de Poisson sur (0,+∞) d'intensité $t^{-2}dt$ et W_i des copies indépendantes d'une gaussienne centrée et stationnaire sur $\mathbb{R}^d$. Sous ces conditions, $\{Z(x)\}$ est un processus max-stable stationnaire de marginales Fréchet unité.

La Figure 3.5 montre des champs moins réguliers que ceux du processus tempête.

Processus *t* extrêmal

Le processus *t* extrêmal (Padoan, 2011) appartient également à la famille des processus max-stables. Opitz (2013) a proposé une construction spectrale de ce processus qui mène à l'écriture générale, similaire à celle de De Haan (1984) de l'équation 3.3, suivante :

$$\{Z(x)\} = \left\{m_\alpha^{-1/\alpha} \max_{i \geq 1} \xi_i W_i(x)\right\},\ \alpha > 0,\ x \in \mathcal{X} \tag{3.13}$$

avec $m_\alpha = \sqrt{\pi}^{-1} 2^{(\alpha-2)/2} \Gamma((\alpha+1)/2)$, ξ_i un processus de Poisson sur (0,+∞) avec comme mesure d'intensité $d\Lambda(t) = \alpha t^{-(\alpha+1)} dt$ et W_i des copies indépendantes d'un processus Gaussien standard sur $\mathbb{R}^d$. $\{Z(x)\}$ est ainsi stationnaire et admet des marginales

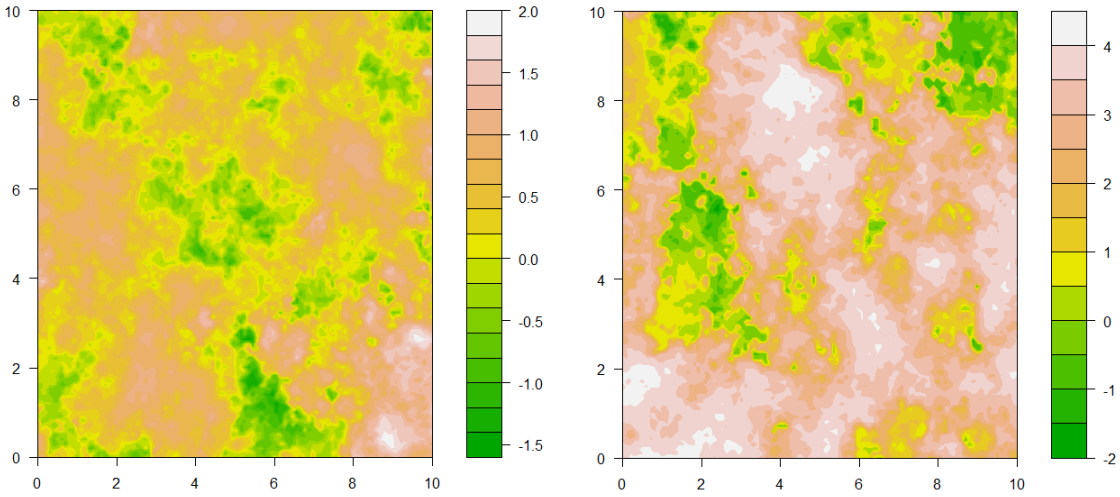


FIGURE 3.5 – Deux exemples de simulation de processus de Schlather. Pour une meilleure visibilité, une échelle logarithmique a été utilisée.

α-Fréchet. Ce processus est une généralisation du processus de Schlather, en admettant $\alpha > 0$ et non pas uniquement $\alpha = 1$. Les motifs obtenus Figure 3.6 sont proches de ceux du processus de Schlather. C'est le paramètre α qui détermine le degré de dépendance des extrêmes du processus.

Processus de Brown-Resnick

En prenant pour $W_i(x)$ les copies indépendantes d'un mouvement Brownien standard, Brown and Resnick (1977) propose le processus dit de Brown-Resnick :

$$\{Z(x)\} = \max_{i \geq 1} \xi_i \exp\left(W_i(x) - |x|/2\right\}, \; x \in \mathcal{X} \tag{3.14}$$

où ξ_i est un processus de Poisson sur (0,+∞) d'intensité $t^{-2}dt$.

Par la suite, Kabluchko et al. (2009) propose une généralisation de ce processus utilisant la notion de variogramme (Chilès and Delfiner, 1999). Ainsi, le processus de Brown-Resnick généralisé s'écrit :

$$\{Z(x)\} = \max_{i \geq 1} \xi_i \exp\left(W_i(x) - \gamma(x)\right\}, \; x \in \mathcal{X} \tag{3.15}$$

où ξ_i est un processus de Poisson sur (0,+∞) d'intensité $t^{-2}dt$ et W_i un processus gaussien centré d'accroissements stationnaires, de variogramme $\gamma(.)$ et tel que $W(0) = 0$ presque sûrement. Cette dernière contrainte rend très difficile d'obtenir des simulations non conditionnelles correctes. En effet, comme dans la pratique il est impossible de

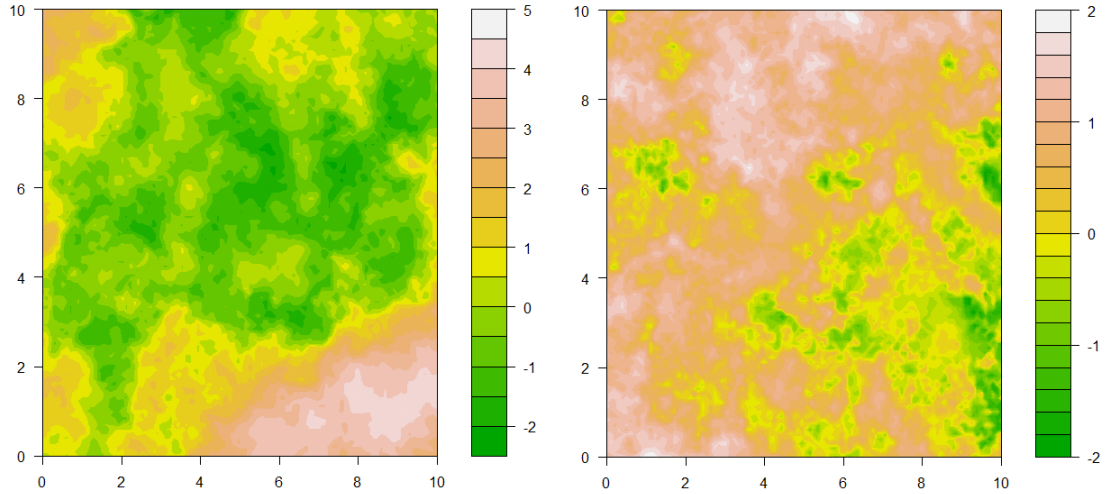


FIGURE 3.6 – Deux exemples de simulation de processus *t* extrêmal (pour $\alpha = 4$). Pour une meilleure visibilité, une échelle logarithmique a été utilisée.

prendre le maximum d'une infinité de réalisations, le fait de passer par l'origine entraîne l'apparition d'une non stationnarité des simulations de $\{Z(x)\}$. Des techniques consistant à translater l'origine sont possibles pour éviter cet écueil (e.g. Oesting et al., 2012).

Coefficient extrêmal

Le coefficient extrêmal est un outil majeur pour modéliser la dépendance de processus max-stables dans les extrêmes. Pour un processus max-stable Z admettant des marginales identiquement distribuées, la fonction du coefficient extrêmal $\theta(\cdot)$ est définie telle que

$$\mathbb{P}(Z(x) < u, Z(x+h) < u) = \left[\mathbb{P}(Z(x) < u)\right]^{\theta(h)}. \tag{3.16}$$

Dans la pratique, l'estimation de $\theta(\cdot)$ peut s'effectuer en utilisant la fonction madogramme $\nu(h) = \frac{1}{2}\mathbb{E}(|Z(x+h) - Z(x)|)$ couramment utilisée en géostatistique pour caractériser la structure spatiale d'un processus $Z(.)$ à une distance h. En utilisant $F(.)$ la fonction de répartition de $Z(.)$, Cooley et al. (2006) définissent le F-madogramme $\nu_F(h) = \frac{1}{2}\mathbb{E}(|F(Z(x+h)) - F(Z(x))|)$ et obtiennent la relation :

$$\hat{\theta}(h) = \frac{1 + 2\hat{\nu_F}(h)}{1 - 2\hat{\nu_F}(h)} \tag{3.17}$$

permettant d'obtenir une estimation de la fonction du coefficient extrêmal. Une généralisation de cette méthode d'estimation a également été proposée par Bel et al. (2008).

Parallèlement, Schlather and Tawn (2003) ont proposé une méthode basée sur la maximisation de la vraisemblance censurée, qui ne sera pas détaillée ici.

Par la suite, en notant Φ la fonction de répartition d'une Gaussienne centrée réduite, on peut obtenir les fonctions du coefficient extrêmal des processus décrits dans cette partie :

- $\theta(x_1 - x_2) = 2\Phi\left(\frac{\sqrt{(x_1-x_2)^T\Sigma^{-1}(x_1-x_2)}}{2}\right)$ pour le processus de Smith, avec Σ la matrice de covariance de la densité f (équation (3.11)).
- $\theta(x_1 - x_2) = 1 + \sqrt{\frac{1-\rho(x_1-x_2)}{2}}$ pour le processus de Schlather.
- $\theta(x_1 - x_2) = 2t_{\alpha+1}(\ 1 \mid \rho(x_1 - x_2),\ (\alpha+1)^{-1}(1-\rho(x_1-x_2)^2))$ pour le processus *t* extrêmal, avec $t_\nu(.|\mu, \Sigma)$ la fonction de répartition d'une distribution de Student de paramètres μ, Σ et ν (respectivement paramètres de position, d'échelle et degré de liberté).
- $\theta(x_1 - x_2) = 2\Phi\left(\sqrt{\frac{\gamma(x_1-x_2)}{2}}\right)$ pour le Brown-Resnick généralisé.

De plus, le coefficient extrêmal θ_∞, défini comme la limite de $\theta(h)$ quand h tend vers ∞, peut être calculé pour chaque processus décrit précédemment :

- $\theta_\infty = 2$ pour le processus de Smith.
- $\theta_\infty = 1 + \sqrt{2}/2$ pour le processus de Schlather.
- $\theta_\infty \in]1.5, 2[$ en fonction de la valeur de α pour le processus *t* extrêmal.
- $\theta_\infty = 2$ pour le processus de Brown-Resnick.
- $\theta_\infty \in]1.5, 2[$ en fonction de la nature du variogramme pour le Brown-Resnick généralisé.

Ainsi, le processus de Brown-Resnick généralisé et le processus *t* extrêmal autorisent une flexibilité dans le coefficient extrêmal. Cette flexibilité est paramétrable soit par la nature du variogramme $\gamma(.)$ soit par la valeur de α. En estimant ces éléments à partir de données observées, ces processus pourront donc reproduire la structure réelle de dépendance des extrêmes.

Les différentes notions et processus évoqués dans ce chapitre sont utilisés dans la suite de ce travail afin d'aider à la compréhension, la modélisation et la simulation d'extrêmes multivariés ayant une structure spatiale donnée.

Chapitre 4

Scores et critères d'évaluation

Tout au long de ce travail des efforts sont fournis pour chercher des outils d'évaluation, objectifs et variés, des différentes méthodes développées par la suite. En effet, dans un but de comparaison de nouvelles méthodes entre elles ou avec d'autres déjà existantes, de tels outils sont indispensables.

L'une des particularités des méthodes de simulations utilisées dans ce travail est qu'elles fournissent des distributions de valeurs en chaque point de la carte. Ainsi, les scores doivent être capables le plus souvent de comparer des distributions de valeurs simulées à une seule valeur observée. A noter également que les variables utilisées dans ce travail sont des variables continues, et que, par conséquent, nous ne nous intéressons pas à la modélisation et à l'évaluation des occurrences de précipitations. Ainsi, les scores qui sont à utiliser sur des variables catégorielles, comme par exemple le Brier Score (Brier, 1950) ou sa généralisation non binaire le Multicategory Ranked Probability Score (Epstein, 1969), ne sont pas adaptés.

Dans ce chapitre, différentes méthodes d'évaluation sont détaillées, chacune s'intéressant à divers aspects de la qualité des simulations, comme leurs caractéristiques univariées, multivariées, leurs propriétés spatiales ou encore le comportement de certains de leurs quantiles. Ces aspects sont évalués individuellement chacun par un score. Aucune méthode dite objet comme SAL (Structure, Amplitude, Location ; Wernli et al., 2008) permettant de mesurer ces différentes performances simultanément n'est utilisée.

A noter que dans toute la suite de ce travail, on écrit $F_{i,n}$ la fonction de répartition prédictive construite à partir des n simulations conditionnelles $\{\hat{y}_i^{(j)}\}_{j=1,\dots,n;i=1,\dots,m}$ et y_i la valeur observée au point i. De plus, nous appelons score la fonction qui compare les valeurs prédites aux observations. Ainsi, la plupart du temps, les scores utilisés dans ce travail sont positifs et doivent être minimisés.

4.1 Scores propres

Lorsque l'on utilise des scores pour évaluer les performances d'un modèle, il est important de savoir si les scores en question ont les bonnes propriétés. En particulier, la notion de « score propre » est très largement discutée dans la littérature (e.g., Parmigiani and Inoue, 2009; Gneiting and Raftery, 2007; Thorarinsdottir et al., 2013).

Définition 3. *Un score S est dit propre, si et seulement si, en notant $\mathcal{F}$ l'ensemble des distributions prédictives possibles et F la distribution réelle, on a*

$$\inf_{f \in \mathcal{F}} S(f) = S(F). \tag{4.1}$$

De plus, si F est la seule distribution qui minimise le score, alors il est dit « strictement propre » .

Cette notion, assez naturelle, s'intéresse au fait que le score utilisé mesure bien ce qui nous importe, et en particulier si il est minimisé par l'observation. Dans les conditions dans lesquelles nous les utilisons, la plupart des scores mentionnés dans cette partie sont des scores propres.

4.2 Scores univariés

Plusieurs scores univariés sont utilisés dans ce travail. Chaque score univarié est calculé individuellement en chacune des observations puis une moyenne spatiale est ensuite effectuée pour obtenir un score global de la méthode sur toute la région. Le premier score utilisé pour comparer la qualité des simulations obtenues avec les différentes méthodes est le Continuous Ranking Probability Score (Gneiting and Raftery, 2007), appelé CRPS par la suite.

$$\text{CRPS}_i(F_{i,n}; y_i) = \int_{-\infty}^{+\infty} \left[F_{i,n}(t) - \mathbb{1}_{\{t > y_i\}}\right]^2 dt, \tag{4.2}$$

avec $\mathbb{1}_{\{.\}}$ la fonction indicatrice. La Figure 4.1 montre une illustration du CRPS. Ce dernier varie dans le même sens qu le carré de la partie hachurée en rouge.

On remarque que plus la fonction de répartition des simulations est « piquée » autour de la valeur observée, meilleur est le score. Un bon CRPS est donc un CRPS proche de 0.

De plus, on peut réécrire ce score de la manière suivante (Hersbach, 2000; Gneiting and Raftery, 2007) :

$$\text{CRPS}_i(F_{i,n}; y_i) = \sum_{j=1}^{n/2} |\hat{y}_i^{(j)} - y_i| - \frac{1}{2} \sum_{j=1}^{n/2} |\hat{y}_i^{(j)} - \hat{y}_i^{(j+n/2)}|, \tag{4.3}$$

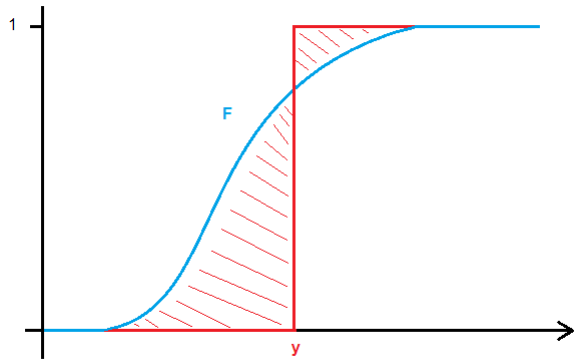


FIGURE 4.1 – Illustration of the CRPS.

où $\hat{y}_i^{(j)}$ et $\hat{y}_i^{(j+n/2)}$ sont deux échantillons indépendants de taille $n/2$ obtenus à partir de $\{\hat{y}_i^{(j)}\}$. Cette écriture permet de séparer le CRPS en deux termes étroitement liés, le premier pouvant être interprété comme un terme de fiabilité et le second comme un terme de résolution.De plus, on remarque que le CRPS s'exprime dans l'unité de la variable y_i et est donc directement interprétable. Une autre décomposition du CRPS est également possible, en trois termes mesurant respectivement la résolution, la fiabilité et l'incertitude d'une méthode (e.g., Candille and Talagrand, 2005; Tödter and Ahrens, 2012).

Le CRPS est couramment utilisé dans des applications en climat (e.g., Bröcker, 2012; Friederichs and Thorarinsdottir, 2012; Oesting and Schlather, 2014), mais des doutes ont été émis quant à sa capacité à évaluer des méthodes s'intéressant aux valeurs extrêmes. En effet, Naveau et al. (2014) observe que lorsque la valeur observée y_i est extrême et donc très éloignée des valeurs simulées, l'utilisation de ce score n'a plus de sens. Dans ce cas, la comparaison de deux distributions par le CRPS ne dépend plus de l'observation y_i, ce qui est problématique. Néanmoins, dans notre travail, il est important de noter que ce cas de figure ne se présente pas ou très rarement. En effet, du fait que nous modélisons et simulons des événements extrêmes, nous générons des distributions de maxima et même si une valeur observée est très élevée, on attend du nuage de valeurs obtenues par simulations conditionnelles qu'il encadre cette valeur observée.

La qualité des simulations conditionnelles sera également évaluée par le quantile score (Friederichs and Hense, 2007) défini comme :

$$\mathrm{QS}_{i,p}(F;y) = \nu_p(y_i - F_{i,n}^{-1}(p)) \tag{4.4}$$

avec p la probabilité d'intérêt et $\nu_p(u) = pu$ si $u \geq 0$ et $\nu_p(u) = (p-1)u$ sinon. $\nu_p(.)$ est la fonction appelée "check function" dans le cadre de la régression quantile (Koenker, 2000). Comme montré par Friederichs and Thorarinsdottir (2012), le CRPS

est proportionnel à l'intégrale du QS_p sur toutes les probabilités p :

$$\text{CRPS}_i(F_{i,n}; y_i) = 2 \int_0^1 \nu_p(y_i - F_{i,n}^{-1}(p))\text{d}p. \tag{4.5}$$

Ainsi, un bon QS_p est à nouveau un QS_p proche de 0. Dans ce travail, en choisissant une probabilité p élevée, on s'intéressera plus particulièrement aux performances prédictives dans la queue de distribution (Friederichs and Hense, 2007). Ce score a pour objectif de pénaliser les méthodes qui ne détectent pas les extrêmes. Cependant, il n'est pas forcément adapté à des distributions d'extrêmes, mais nous l'utilisons comme un complément du CRPS, pour nous donner des renseignements sur la capacité des différentes méthodes à détecter les événements « très » extrêmes.

4.3 Critères de cohérence spatiale

En complément de ces scores univariés, deux scores multivariés sont considérés. En effet, comme nous utilisons des méthodes spatialisées, il est important d'évaluer si les simulations respectent la structure spatiale des données.

Ainsi, nous proposons le Multivariate Score (MS) :

$$\text{MS}_d(\hat{y}_1, \ldots, \hat{y}_n; y_1, \ldots, y_m) = \sum_{I \in \mathcal{P}_m(d)} \sum_{j=1}^{n} ||\hat{y}_I^{(j)} - y_I||_d \tag{4.6}$$

avec $||.||_d$ la norme euclidienne dans $\mathbb{R}^d$, $\mathcal{P}_m(d)$ l'ensemble des vecteurs d'indices de longueur d dans $\{1, \ldots, m\}$, $y_I = \{y_i | i \in I\}$ et $\hat{y}_I^{(j)} = \{\hat{y}_i^{(j)} | i \in I\}$. Dans la pratique, ce score revient à construire un indicateur multivarié de l'erreur en se basant sur des combinaisons de d-uplets.

De plus, une version multivariée du CRPS, appelée Energy Score (Gneiting and Raftery, 2007; Gneiting et al., 2008), a pour écriture :

$$ES(\hat{\mathbf{y}}^{(1)}, \ldots, \hat{\mathbf{y}}^{(n)}, \mathbf{y}) = \frac{1}{n} \sum_{j=1}^{n} ||\hat{\mathbf{y}}^{(j)} - \mathbf{y}|| - \frac{1}{2n^2} \sum_{j'=1}^{n} \sum_{j=1}^{n} ||\hat{\mathbf{y}}^{(j)} - \hat{\mathbf{y}}^{(j')}|| \tag{4.7}$$

avec $\hat{\mathbf{y}}^{(j)} = (\hat{y}_1^{(j)}, \ldots, \hat{y}_m^{(j)})$ et $\mathbf{y} = (y_1, \ldots, y_m)$.

On retrouve une décomposition similaire à celle obtenue pour le CRPS (équation 4.3).

Ces deux scores permettent d'évaluer la cohérence spatiale des simulations, ce dont les scores univariés sont incapables. En plus de ces scores, un critère d'évaluation de la dépendance spatiale par rapport aux observation a été construit. En géostatistique, un outil

communément utilisé pour décrire la dépendance spatiale d'un champ d'observations est le variogramme (Chilès and Delfiner, 1999). Son rôle est de modéliser les variances des accroissements existant entre les points de l'espace en fonction de la distance les séparant. Le variogramme empirique $\gamma^*(h)$ peut donc s'écrire :

$$\gamma^*(h) = \frac{1}{2N(h)} \sum_{|s(i)-s(i')|\approx h} \big(y_i - y_{i'}\big)^2, \tag{4.8}$$

avec N(h) le nombre de couples de points distants d'approximativement *h.*

En comparant le variogramme des observations à ceux calculés à partir des simulations obtenues avec les différentes méthodes, on peut construire un critère évaluant la dépendance spatiale des simulations. En notant V le variogramme empirique calculé à partir des observations et V^j celui construit à partir du $j_{\text{ème}}$ champ de simulations, on utilise un critère RMSE (Root Mean Square Error) calculé à partir des distances entre les variogrammes :

$$\text{RMSE}_v(V^j, \ldots, V^n; V) = \frac{1}{n} \sum_{j=1}^{n} ||V^j - V||. \tag{4.9}$$

Ce critère permet donc de comparer la dépendance spatiale associée à chaque méthode.

4.4 Comparaison de distributions

La dernière catégorie de critères présentée ici est basée sur des comparaisons de distributions. Nous utilisons la statistique de test de Kolmogorov-Smirnov, s'écrivant :

$$\text{K-S}_i(F_{i,n}, F_i) = \sup_x |F_{i,n}(x) - F_i(x)|. \tag{4.10}$$

Cette statistique est basée sur l'écart maximal entre la fonction de répartition construite à partir des simulations $F_{i,n}$ et celle, F_i, construite à partir des observations au point i.

La statistique de test de Cramér-von Mises est également examinée :

$$\text{CVM}_i(F_{i,n}, F_i) = \int_{-\infty}^{\infty} [F_{i,n}(x) - F_i(x)]^2 \, \mathrm{d}x. \tag{4.11}$$

Cette fois, c'est l'écart quadratique moyen entre les deux fonctions de répartition qui est considéré.

Dans la pratique, les distributions observées F_i sont obtenues en considérant les valeurs observées sur plusieurs années consécutives. De plus, un critère global est obtenu en prenant la moyenne de chaque critère sur tous les points i .

4.5 Skill-Scores

Ces différents scores s'expriment chacun dans des unités différentes. Il est donc difficile de les comparer entre eux. En l'état, ils permettent simplement d'établir une hiérarchie entre différentes techniques, mais difficilement de quantifier l'amélioration d'une méthode par rapport à l'autre. Pour ce faire, Gneiting and Raftery (2007) propose de calculer un indicateur, appelé Skill-Score, à partir du score S considéré. Ainsi, en notant S^{new}, S^{ref} et S^{opt}, respectivement les scores de la nouvelle méthode, de la méthode de référence et le meilleur score possible, le Skill-Score de la méthode s'écrit :

$$S^{skill} = \frac{S^{new} - S^{ref}}{S^{opt} - S^{ref}}. \tag{4.12}$$

Pour les méthodes qui apportent une amélioration par rapport au modèle de référence, S^{skill} appartient à l'intervalle $[0, 1]$ et 1 correspond au cas optimal. Il est également intéressant de remarquer, que tous les scores proposés dans cette partie peuvent être considérés comme des distances. Ainsi, pour tous ces scores, on a $S^{opt} = 0$ et donc tous les Skill-Scores correspondants peuvent être réécrits

$$S^{skill} = \frac{S^{ref} - S^{new}}{S^{ref}}, \tag{4.13}$$

ce qui revient à calculer le pourcentage d'amélioration d'une méthode par rapport à la méthode de référence et qui est donc très facilement interprétable. Néanmoins, de manière générale les Skill-Scores ne sont pas des scores propres, même si les scores sous-jacents le sont (Gneiting and Raftery, 2007).

L'utilisation de Skill-Scores (relatifs aux différents scores présentés ici) permet donc d'évaluer directement l'amélioration d'une nouvelle méthode par rapport à d'autres méthodes déjà existantes. Ainsi, dans le reste de ce travail, nous utiliserons presque exclusivement des Skill-Scores construits à partir de scores adaptés aux objectifs poursuivis.

Deuxième partie

Simulation conditionnelle

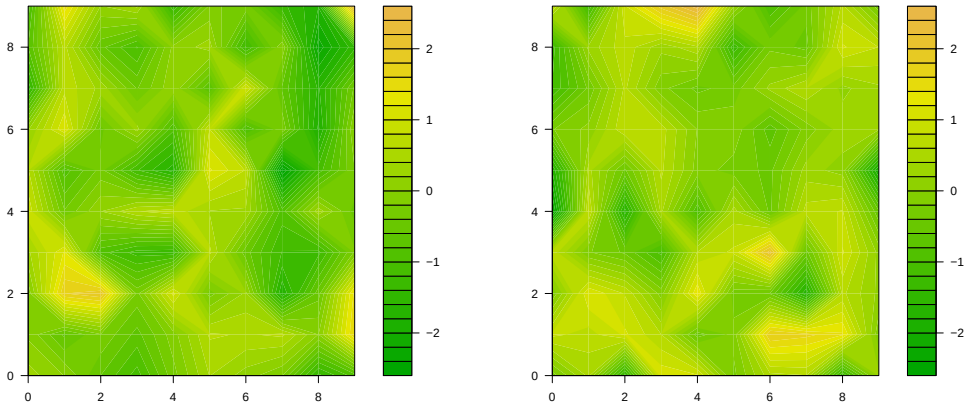
Chapitre 5

Introduction

Comme les extrêmes de précipitations admettent une dépendance spatiale, il est important que les valeurs que nous générons puissent reproduire cette structure. Ainsi, dans cette partie, on cherche à obtenir des simulations qui prennent des valeurs cohérentes tout en respectant la structure de dépendance spatiale du phénomène étudié. Afin de reproduire cette structure, il est important de simuler des champs et pas seulement des réalisations ponctuelles indépendantes. A l'étape de simulation, la structure de covariance des données est supposée connue. Dans la pratique, les paramètres caractérisant cette structure sont estimés à partir des observations disponibles. Ainsi, les simulations suivent la loi des observations et peuvent reproduire leur structure spatiale. Un autre avantage des techniques de simulation vient du côté stochastique des résultats obtenus qui permet de restituer la variabilité du phénomène modélisé. Cette variabilité peut être observée Figure 5.1 où deux simulations de champs Gaussiens ayant les mêmes paramètres sont représentées. Ces champs ont été obtenus par la méthode des bandes tournantes (Mantoglou and Wilson, 1982).

Différentes techniques de simulations existent, le plus souvent sous l'hypothèse gaussienne. Une approche souvent utilisée consiste à effectuer une anamorphose Gaussienne des données pour se ramener dans ce cadre de travail et ainsi pouvoir appliquer ces techniques de simulation. Pour plus de détails sur les procédures de simulation voir Chilès and Delfiner (1999) et Lantuéjoul (2002).

De plus, il peut être utile de non seulement utiliser les observations disponibles pour calibrer le modèle mais également comme contraintes à respecter lors de sa simulation. De la sorte, les simulations ont la bonne distribution et en plus coïncident avec la réalité aux points d'observation. C'est une procédure dite de simulation conditionnelle. Ce type de techniques est utilisé pour décrire des phénomènes physiques en intégrant des observations et en fournissant des intervalles de confiance aux estimations obtenues. Dans la pratique, ces simulations conditionnelles sont utilisées entre autres pour décrire la configuration des gisements de pétrole, la teneur des exploitations minières (Chilès and



(a) (b)

FIGURE 5.1 – Exemples de simulations gaussiennes.

Delfiner, 1999) et de plus en plus pour des problématiques environnementales (Mariethoz et al., 2009).

Ainsi, à l'inverse des simulations non conditionnelles pour lesquelles les champs **Z** simulés suivent une distribution cible $\mathcal{L}$ i.e.

$$\textbf{Z :} \text{ réalisation d'une loi } \mathcal{L},$$

les simulations conditionnelles permettent en plus de rajouter les contraintes formulées par les observations, i.e.

$$\begin{cases} \textbf{Z :} \text{ réalisation d'une loi } \mathcal{L}, \\ Z(x_i) = z_i, \ \forall z_i \text{ observations aux points } x_i. \end{cases}$$

Dans le cadre de certains modèles, comme par exemple les modèles Gaussiens ou Booléens, il existe des formules explicites pour obtenir ces simulations conditionnelles (Lantuéjoul, 2002).

A titre d'exemple, la Figure 5.2 présente 20 points tirés aléatoirement sur une grille. Les valeurs de ces points ont été obtenues en simulant un champ Gaussien.

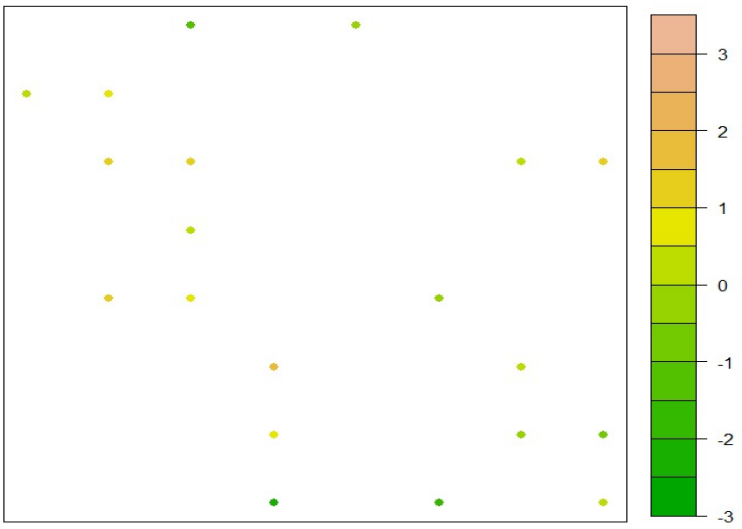
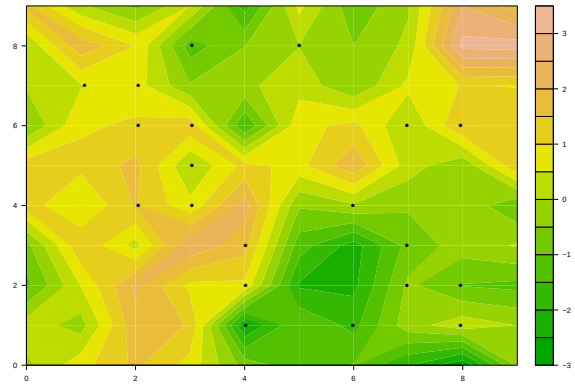


FIGURE 5.2 – Valeurs des points conditionnants

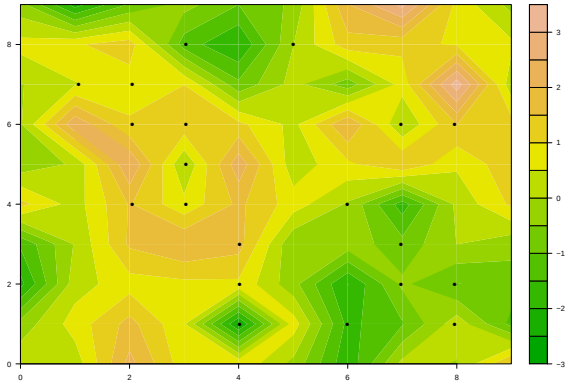
Le but est d'obtenir des valeurs simulées en tout point de la grille, mais également que les champs simulés respectent les contraintes imposées par les observations.Les Figures 5.3 ont été obtenues par simulation séquentielle gaussienne qui, sous l'hypothèse gaussienne, découle directement des techniques de krigeage (Alabert et al., 1990).

Ainsi, ces outils de simulation conditionnelle permettent d'obtenir des champs de valeurs respectant la structure spatiale des données et les valeurs aux points d'observations. De plus, on constate une certaine variabilité d'une simulation à une autre, qui va permettre, en effectuant plusieurs simulations, d'obtenir pour chaque nouveau point une distribution de probabilité conditionnelle. Ainsi, ces méthodes permettront d'obtenir différentes statistiques d'intérêt telles que des quantiles ou des indices de dispersion.

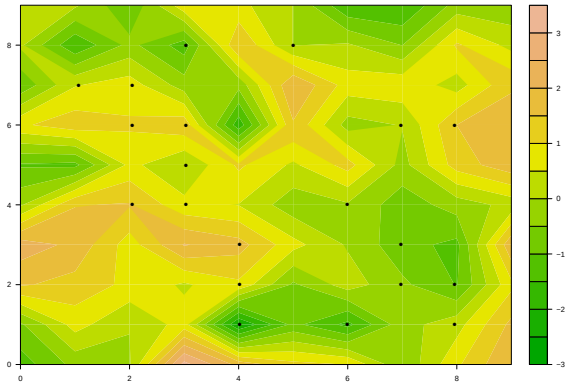
Ces réalisations sont donc d'une très grande utilité, lorsque que l'on désire obtenir des réalisations en tout point de notre région d'intérêt. Néanmoins, pour des modèles plus complexes, comme les processus max-stables qui nous intéressent dans ce travail, la procédure de simulation conditionnelle est beaucoup moins directe. Ainsi, dans le prochain chapitre, un algorithme de simulation conditionnelle de processus max-stables est présenté et appliqué à des données de précipitation.



(a) Une simulation conditionnelle



(b) Une simulation conditionnelle



(c) Une simulation conditionnelle

FIGURE 5.3 – Exemples de simulations conditionnelles.

Chapitre 6

Simulation conditionnelle de processus max-stables

Dans ce chapitre, nous allons fournir des outils pour permettre d'obtenir des distributions de maxima tout en tenant compte des observations disponibles. Usuellement, pour modéliser des maxima, le cadre statistique utilisé est celui de la théorie des valeurs extrêmes. Plus particulièrement, lorsque l'on s'intéresse à des phénomènes spatiaux, les outils utilisés sont les processus max-stables (voir chapitre 3). De plus, les simulations conditionnelles permettent de fournir des distributions en chaque point de la carte, tout en tenant compte des contraintes émanant des données.

Cependant, ces techniques ne sont pas applicables directement pour des processus max-stables. Ainsi, nous utilisons un algorithme de simulation conditionnelle proposé par Dombry et al. (2013), en l'adaptant à une famille de processus max-stable particulière : les processus t extrêmal. De plus, certaines modifications de l'algorithme permettent de l'améliorer, en particulier en autorisant un plus grand nombre de points conditionnants.

Une étude sur des données simulées permet de mettre en avant le rôle des différentes étapes de cet algorithme et celui des paramètres du modèle choisi. Une application sur des données réelles dans le sud-est de la France compare les résultats de cet algorithme avec ceux obtenus par d'autres méthodes.

Ces résultats sont encourageants et montrent des performances intéressantes vis à vis de critères univariés, multivariés et de la capacité de la méthode à reproduire les tendances générales des observations.

L'article suivant a été soumis et révisé pour publication dans la revue « Spatial Statistics ».

Conditional simulations of the extremal t process: application to fields of extreme precipitation

Aurélien Bechler[a,b,c], Liliane Bel[a,b], Mathieu Vrac[c]

[a] *INRA, UMR 518 Math. Info. Appli., F-75005 Paris, France*
[b] *AgroParisTech, UMR 518 Math. Info. Appli., F-75005 Paris, France.*
[c] *LSCE-IPSL, Centre d'Etudes de Saclay, Orme des Merisiers, Bat. 701 91191 Gif-sur-Yvette, France*

Abstract

The last decade has seen max-stable processes emerge as a powerful tool for the statistical modeling of spatial extremes and there are increasing works using them in a climate framework. One recent utilization of max-stable processes in this context is for conditional simulations that provide empirical distribution of a spatial field conditioned by observed values in some locations. In this work conditional simulations are investigated for the extremal t process taking benefits of its spectral construction.

The methodology of conditional simulations proposed by Dombry et al. (2013) for Brown-Resnick and Schlather models is adapted for the extremal t process with some original improvements which enlarge the possible number of conditional points. A simulation study enables to highlight the role of the different parameters of the model and to emphasize the importance of the steps of the algorithm.

An application is performed on precipitation data in the south of France where extreme precipitation events (Cevenol) may generate major floods. This shows that the model and the algorithm perform well provided the stationary assumptions are fulfilled.

Keywords:
Max-stable processes, extremal t processes, spatial extremes, conditional simulation, extreme precipitation.

1. Introduction

Extreme events are one of the major demonstrations of complex environmental systems. As they are inherently rare and often unexpected, society is not well prepared to face them. One important feature of extreme events is that their probability decreases rapidly with increasing intensity, while damage increases strongly as does the cost of protection against such events. Both analyses of observational data and projections from climate models suggest that the occurrence and magnitudes of some extreme events will increase in the future (IPCC, 2011). A potential consequence is an increasing number of catastrophes such as flooding, storms or crop failures. In addition to their direct effects and costs in terms of human lives, building destruction or food security, these events will also have indirect impacts such as increasing costs for building

or strengthening infrastructures, in addition to higher insurance premiums, and before even mentioning the consequences of global population displacements (IPCC, 2011). A better understanding and modeling of extreme events is therefore a crucial need in many economic aspects and societal contexts for both public and private decision makers.

From a mathematical point of view, the suitable probabilistic framework for modeling the frequency and intensity of those special events is the statistical theory of extreme values. The extreme value theory is well established for univariate stationary series. There is now a growing effort to develop methods and models to deal with complex data involved in applications where space-time problems and non-stationarity are the rule rather than the exception.

Multivariate extreme distributions have been used in climate and environmental applications in order to take into account extremal dependence (Coles and Tawn, 1994; De Haan and De Ronde, 1998; Schlather and Tawn, 2003; Fawcett and Walshaw, 2012). For the study of extremes in a spatial context, max-stable processes are a relevant framework for modeling and inference. Introduced by De Haan (1984) different parametric models have been later proposed by Smith (1990), Schlather (2002) or Kabluchko et al. (2009). The extremal t process (Padoan, 2011; Opitz, 2013) is a model of that family that extends the Schlather model. It has a supplementary parameter that lets the extremal coefficient function vary in a wider range. Max-stable processes have been used in applications for modeling rainfall (Buishand et al., 2008; Davison et al., 2012), temperature (Davison and Gholamrezaee, 2012; Fuentes et al., 2013), snowfall (Gaume et al., 2013) or snow depth (Blanchet and Davison, 2011). A main objective of these works is to provide return periods of rare events for prevention purposes even in locations where no observations are available.

The aim of the present paper is to further investigate conditional simulations of extreme precipitation. Conditional simulations provide empirical distributions of a random field knowing the values observed at some given locations. Consequently, they allow us to statistically generate extreme values at locations without any recording, in agreement with surrounding observations. This capacity to generate time series at observation-free locations is of major importance for some impact models, such as hydrological models that require input data, e.g. precipitation values, correctly spatially distributed. The spatial coverage of the meteorological inputs of such models has to be as precise as possible. Furthermore, when other techniques provide individual values or indicators, conditional simulations provide distributions at every new locations which allows us to get numerous statistical quantities as e.g. quantile, dispersion indicators, etc... The conditional simulation techniques are then of major importance.

For Gaussian and Boolean fields, algorithms for conditional simulations are well established (Lantuéjoul, 2002). They have been used for representing physical phenomena by integrating heterogeneous data and assigning confidential limits to the estimates. For example, they have been used for predicting the geometry of oil reservoirs or the grade of mining blocks (Chilès and Delfiner, 1999) and are used more and more in the environmental sciences (Mariethoz et al., 2009). Concerning max-stable processes, Wang and Stoev (2011) consider max-linear random fields, a special case of max-stable processes,

and provide an exact expression of the conditional distributions as well as an efficient algorithm for the simulation. In a further work, Dombry and Eyi-Minko (2013) derive the regular conditional distribution for the very general class of max-infinitely divisible processes which encompasses the max-stable processes. This result has been applied then to provide a methodology and an algorithm for Brown-Resnick and Schlather processes (Dombry et al., 2013). However, for processes which are not regular, that is the exponent measure is not absolutely continuous with respect to Lebesgue measure, the conditional distribution is not explicite. Oesting and Schlather (2014) discuss the frameworks where Wang and Stoev (2011) and Dombry et al. (2013) can be applied, and, in case of non regular models, propose an algorithm based on martingale arguments.

In this paper we adapt the methodology of Dombry et al. (2013) to the extremal t process. In this case all the calculations of the algorithm can be derived explicitely, nevertheless some steps still present computational issues. In particular, one of the steps of the algorithm involves an acceptance/rejection procedure for which the rate of rejection increases very fast with the number of conditioning points. This limiting point makes the algorithm inapplicable when the number of conditioning points is greater than approximately 30. To bypass the issue, we propose some original modifications of the algorithm to replace the acceptance/rejection step by a Gibbs sampler which speeds up the calculations and enables a large number of conditioning points. Despite this improvement, some difficulties persist in the algorithm, mostly due to the calculation of high dimensional integrals that have to be approximated. Then to get a better understanding of the algorithm mechanisms, a simulation study is led to assess the influence of the parameters from the extremal t process and to evaluate the importance of the most problematic and time-consuming step of the algorithm.

We focus on the French Mediterranean basin, which is a key spot of occurrences of meteorological extremes such as heavy precipitation (e.g., Ducrocq et al., 2008; Lebeaupin et al., 2006; Beaulant et al., 2011; Vrac and Yiou, 2010; Vrac et al., 2012). Indeed, major extreme precipitation events are regularly observed in this region near the Cevennes mountains. The modeling and the forecast of this extreme precipitation – the so-called "Cevenol events" – are of major importance for hydrological studies in this complex terrain since they often trigger major floods in this region (e.g. Ducrocq et al., 2008).

The structure of this article is the following. In Section 2, a short reminder on max-stable processes is provided and the extremal t process is described. Section 3 contains theoretical considerations on conditional simulations of an extremal t process, as well as the associated algorithm and its improvement. An extensive simulation study is carried out in Section 5 to evaluate the performances of the algorithm. The use of the continuous ranking probability score and the quantile score (Gneiting and Raftery (2007)) enables us to assess the accuracy of our methodology and to compare it to other models. This evaluation allows us to quantify the importance of the model parameters and of the algorithm steps. Section 6 displays an application to maximum precipitation

data in the south of France where extreme precipitation may cause major floods. Some comments and conclusions are given in section 7.

2. Extreme Value Theory

2.1. Max-Stable Process

We will note vectors in plain text and vectors of vectors in bold. Let us consider a stationary random field $\mathbf{Z} = \{Z(x)\}_{x\in\mathbb{R}^d}$ with a given distribution $\mathcal{H}_0$ and $(Z_1, \ldots, Z_n)$ independent copies of Z. If there exist sequences $a_n(x) \in \mathbb{R}$ and $b_n(x) \in \mathbb{R}^+$ such that for all $n \in \mathbb{N}$,

$$\left\{ \frac{\max_{i=1,\ldots,n} Z_i(x) - a_n(x)}{b_n(x)} \right\}_{x\in\mathbb{R}^d} \tag{1}$$

follows the distribution $\mathcal{H}_0$, then $\mathbf{Z} = \{Z(x)\}_{x\in\mathbb{R}^d}$ is a max-stable process (De Haan, 1984).

As a consequence, all finite dimensional marginal distributions are max-stable, and in particular, the univariate margins $Z(x)$ distribution belong to the Generalized Extreme Value (GEV) family:

$$\mathbb{P}(Z(x) \leq z) = \begin{cases} \exp\left(-\left[1 + \xi(x)\big(\frac{z-\mu(x)}{\sigma(x)}\big)\right]^{-1/\xi(x)} \right) & \text{if } \xi(x) \neq 0, \\ \exp\left(-\exp\big(\frac{z-\mu(x)}{\sigma(x)}\big) \right) & \text{if } \xi(x) = 0, \end{cases} \tag{2}$$

where $\mu(x)$, $\sigma(x)$ and $\xi(x)$ are respectively the location, scale, and shape parameters at site x. Furthermore, the sign of $\xi(x)$ will determine the family distribution $Z(x)$ belongs to among Fréchet ($\xi > 0$), Gumbel ($\xi = 0$) or Weibull ($\xi < 0$), and consequently the distribution tail characteristics.

Different max-stable models have been proposed and used in the past by several authors, e.g. Brown and Resnick (1977), Smith (1990), Schlather (2002) or Kabluchko et al. (2009).
Their margins are Fréchet distributed, i.e.

$$\mathbb{P}(Z(x) \leq z) = e^{-z^{-1}}. \tag{3}$$

2.2. The extremal t process: definition and properties

The extremal t process (Padoan, 2011) belongs to the max-stable process family. Opitz (2013) proposed a spectral construction of the extremal t process which leads to the general expression

$$\mathbf{Z} = \{Z(x)\} = \big\{ m_\alpha^{-1/\alpha} \max_{i\geq 1} \xi_i W_i(x) \big\},\ \alpha > 0,\ x \in \mathcal{X} \subset \mathbb{R}^d,\ \text{with } \mathcal{X} \text{ a Borel set}, \tag{4}$$

with $m_\alpha = \sqrt{\pi}^{-1} 2^{(\alpha-2)/2}\Gamma((\alpha+1)/2)$, ξ_i a Poisson point process on $(0,+\infty)$ with intensity measure $d\Lambda(t) = \alpha t^{-(\alpha+1)}dt$ and W_i independent copies of a stationary standard Gaussian process on $\mathbb{R}^d$ with correlation function $\rho(\cdot)$. $\mathbf{Z}$ margins are α-Fréchet distributed, i.e.

$$\mathbb{P}(Z(x) \leq z) = e^{-z^{-\alpha}}. \tag{5}$$

Unlike a common representation of max-stable processes with unit-Fréchet margins, we choose the α-Fréchet representation to be consistent with the notation of Opitz (2013). It has to be noted that for $\alpha = 1$, this representation leads to a Schlather process and hence the extremal t process can be considered as a generalization of the Schlather process, with a greater flexibility.

The parameter α and the correlation function $\rho(\cdot)$ determine the dependence structure. Furthermore, as will be seen in the next section, α characterizes the measure of spatial dependence in extreme values.

2.3. Extremal coefficient

For a max-stable process such that all margins are equally distributed the extremal coefficient function $\theta(\cdot)$ (Schlather and Tawn, 2003) is defined as

$$\mathbb{P}(Z(x) < u, Z(x+h) < u) = \left[\mathbb{P}(Z(x) < u)\right]^{\theta(h)}. \tag{6}$$

$\theta(h) \in [1,2]$ measures the strength of dependence in extremes values, $\theta(h) = 1$ corresponds to the complete dependence case, $\theta(h) = 2$ corresponds to the independence case. The extremal coefficient of the extremal t process depends on α and is given by (Opitz, 2013)

$$\theta(h) = 2t_{\alpha+1}(\ 1 \mid \rho(h),\ (\alpha+1)^{-1}(1-\rho(h)^2)), \tag{7}$$

with $t_\nu(.|\mu,\Sigma)$ the cdf (cumulative distribution function) of a Student distribution with the parameters μ, Σ and ν (respectively location, shape and degree of freedom parameters).

In particular for the extremal t process with $\alpha \geq 0$, θ_∞, the limit of $\theta(h)$ when h tends to ∞, belongs to the interval $]1.5, 2[$, while $\theta_\infty = 1 + \sqrt{2}/2$ (≈ 1.7) for the Schlather process.

Consequently, thanks to this supplementary parameter α, the extremal t process enables us to model a greater variety of situations for dependence in extremes. Furthermore, this flexibility of the behavior of the dependence in extremes values is different from the one given by the class of general Brown-Resnick processes (Kabluchko et al., 2009).

3. Conditional simulations

Conditional simulations provide simulations of random fields at locations without observation, given observations in fixed locations, according to a model. The simulated fields are required to fit the observed values at the fixed locations. In contrast to simple simulations for which simulated values follow a target distribution – i.e. $\mathbf{Z}$ is a

realization from a distribution $\mathcal{L}$ –, conditional simulations take into account constraints expressed by the observations, i.e.

$$\begin{cases} \mathbf{Z} \text{ is a realization from a distribution } \mathcal{L}, \\ Z(x_i) = z_i, \ \forall z_i \text{ observations at the locations } x_i, i = 1, \ldots, k. \end{cases}$$

Conditional simulations are powerful tools which provide not only individual values at new locations, but probabilistic distributions at those locations, given a chosen model. Those distributions allow the calculation of many statistical quantities as for example quantiles, dispersion indicators, etc. . . .
For some models, such as Gaussian or Boolean, algorithms for performing conditional simulations are quite straightforward and based on kriging techniques (Lantuéjoul, 2002). Dombry et al. (2013) proposed a general framework for conditional simulations of max-stable processes. The max-stable process $\mathbf{Z}$ is written as the point-wise maximum over an infinity of continuous functions

$$Z(x) = \max_{i \geq 1} \phi_i(x). \tag{8}$$

Let $\mathbf{x} = (x_1, x_2, \ldots, x_k) \in \mathcal{X}^k$ the k conditioning points with $\mathcal{X} \subset \mathbb{R}^d$, $z = (z_1, z_2, \ldots, z_k) \in \mathbb{R}^k$ such that $Z(x_i) = z_i$ and $\mathbf{s} = (s_1, s_2, \ldots, s_m) \in \mathcal{X}^m$ the m locations the process $\mathbf{Z}$ is to be simulated. For each conditioning point x_i, an unique function ϕ_j almost surely exists, called extremal function and denoted ϕ_j^+ such that $\phi_j^+(x_i) = z_i$. One extremal function may be extremal for several conditioning points. Let $\tau = (\tau_1, \ldots, \tau_l)$ be a partition of the set $\{x_1, \ldots, x_k\}$ with $l = |\tau|$ $(\leq k)$ such that

$$\phi_j^+(x_i) \ = \ z_i \text{ if } x_i \in \tau_j \ \text{ and } \ \phi_j^+(x_i) \ < \ z_i \text{ if } x_i \notin \tau_j, \quad (i = 1, \ldots, k; j = 1, \ldots, l). \tag{9}$$

The partition τ is called a hitting scenario (Wang and Stoev, 2011). The conditional simulation algorithm is then divided into three steps. The first one consists in drawing a partition of $\{x_1, \ldots, x_k\}$ such that each class of the partition gathers conditioning points that have the same extremal function. Then, for each class in the drawn partition, each extremal function ϕ_j^+ is drawn according to the constraints (9) given by the conditional observations. Finally a number N sufficiently large of non extremal functions denoted ϕ_j^- are drawn such that for all conditioning points they verify

$$\phi_j^-(x_i) \ < \ z_i \text{ for all } x_i \quad (i = 1, \ldots, k; j = 1, \ldots N). \tag{10}$$

The simulated values on the set $\mathbf{s}$ are obtained by taking the point-wise maximum of $\{\phi_j^-(\mathbf{s}), \phi_j^+(\mathbf{s})\}$. To perform the three steps of the algorithm, the conditional distribution of the partition, the $\{\phi^+\}$ and the $\{\phi^-\}$ are needed. Dombry et al. (2013) derived the expression of the conditional distributions for the Schlather and the Brown-Resnick process. The conditional distribution for the extremal t process is obtained thanks to a slight adaptation.

3.1. Conditional distribution for the extremal t process

Let $\mathbf{Z}$ be an extremal t process. It can be written as

$$Z(x) = \max_{i \geq 1} \phi_i^{(\alpha)}(x) \tag{11}$$

with

$$\phi_i^{(\alpha)}(x) = m_\alpha^{-1/\alpha} \xi_i W_i(x), \quad i = 1, 2 \ldots . \tag{12}$$

$\Phi^{(\alpha)} = \{\phi_i^{(\alpha)}\}_{i \geq 1}$ is a Poisson point process on $\mathcal{C}_0$, the space of positive continuous function on $\mathcal{X} \subset \mathbb{R}^d$. The intensity measure $\Lambda_{\mathbf{x}}^{(\alpha)}$ of the point process $\Phi^{(\alpha)}(\mathbf{x}) = \{\phi_i^{(\alpha)}(\mathbf{x})\}_{i \geq 1}$ on $(0, +\infty)^k$ is

$$\Lambda_{\mathbf{x}}^{(\alpha)}(A) = \int_0^{+\infty} \mathbb{P}[\Phi^{(\alpha)}(\mathbf{x}) \in A] \alpha \xi^{-(\alpha+1)} \mathrm{d}\xi, \quad A \subset (0, +\infty)^k \text{ a Borel set.} \tag{13}$$

The extremal t process fulfills the conditions necessary for $\Lambda_{\mathbf{x}}^{(\alpha)}$ being absolutely continuous with respect to the Lebesgue measure (Dombry and Eyi-Minko, 2013). The density (Radon-Nikodym derivative) $\lambda_{\mathbf{x}}^{(\alpha)}(z)$ such that $\Lambda_{\mathbf{x}}^{(\alpha)}(dz) = \lambda_{\mathbf{x}}^{(\alpha)}(z)dz$ and the conditional density

$$\lambda_{\mathbf{s}|\mathbf{x},z}^{(\alpha)}(u) = \frac{\lambda_{\mathbf{s},\mathbf{x}}^{(\alpha)}(u, z)}{\lambda_{\mathbf{x}}^{(\alpha)}(z)}, \quad (\mathbf{s}, \mathbf{x}) \in \mathcal{X}^{m+k}, z \in (0, +\infty)^k, u \in \mathbb{R}^m. \tag{14}$$

are given by :

$$\lambda_{\mathbf{x}}^{(\alpha)}(z) = \frac{\alpha \pi^{\frac{\alpha-1}{2}}}{w_{k+\alpha-2}} \, |\Sigma_{\mathbf{x}}|^{-1/2} \, [z^T \Sigma_{\mathbf{x}}^{-1} z]^{-(\frac{k+\alpha}{2})} \Gamma((\alpha+1)/2)^{-1} \tag{15}$$

and

$$\lambda_{\mathbf{s}|\mathbf{x},z}^{(\alpha)}(u) = \frac{w_{k+\alpha-2}}{w_{k+m+\alpha-2}} \sqrt{|\Sigma_{\mathbf{s},\mathbf{x}}|^{-1} |\Sigma_{\mathbf{x}}|} \frac{[z^T \Sigma_{\mathbf{x}}^{-1} z]^{(k+\alpha)/2}}{[(u,z)^T \, \Sigma_{\mathbf{s},\mathbf{x}}^{-1} \, (u,z)]^{(k+m+\alpha)/2}}, \tag{16}$$

with $w_d = \pi^{d/2}/\Gamma(d/2+1)$, $|\,.\,|$ the determinant operator and $\Sigma_{\mathbf{x}}$, $\Sigma_{\mathbf{s},\mathbf{x}} = \begin{pmatrix} \Sigma_{\mathbf{s}} & \Sigma_{\mathbf{s}:\mathbf{x}} \\ \Sigma_{\mathbf{x}:\mathbf{s}} & \Sigma_{\mathbf{x}} \end{pmatrix}$ the covariance matrices of respectively $W(\mathbf{x})$ and $W(\mathbf{s}, \mathbf{x})$. For the details of the calculations see Appendix A.

For $\alpha = 1$, we note that the expressions (15) and (16) are the ones obtained for the Schlather process (Dombry et al., 2013). They are equivalent to those obtained by Ribatet (2013a).

Conditionally to $Z(\mathbf{x}) = z$, the Poisson point process $\Phi^{(\alpha)}$ can be decomposed into two independent Poisson processes, such that $\Phi^{(\alpha)} = \Phi_{\mathbf{x},z}^{(\alpha)+} \cup \Phi_{\mathbf{x},z}^{(\alpha)-}$, with

$$\begin{aligned} \Phi_{\mathbf{x},z}^{(\alpha)-} &= \{\phi^{(\alpha)} \in \Phi^{(\alpha)} |\ \forall i = 1, \ldots, k,\ \phi^{(\alpha)}(x_i) < z_i\} = \{\phi_j^{(\alpha)-}\}_{j \geq 1}, \\ \Phi_{\mathbf{x},z}^{(\alpha)+} &= \{\phi^{(\alpha)} \in \Phi^{(\alpha)} |\ \exists i = 1, \ldots, k,\ \phi^{(\alpha)}(x_i) = z_i\} = \{\phi_j^{(\alpha)+}\}_{j \geq 1}. \end{aligned} \tag{17}$$

3.2. Algorithm

The algorithm proposed by Dombry et al. (2013), performed for the conditional simulation of Brown-Resnick and Schlather processes, can also be used for extremal t processes. It is based on the density $\lambda_{\mathbf{x}}^{(\alpha)}$ and the decomposition of $\Phi^{(\alpha)}$. The following notations have been chosen to be consistent with notations of Dombry et al. (2013). Let us note $x_{\tau_j} = \{x_i \mid x_i \in \tau_j, i = 1, \ldots, k\}$, $z_{\tau_j} = \{z_i \mid Z(x_{\tau_j}) = z_i, i = 1, \ldots, k\}$ and τ_j^c the complementary of τ_j in $\{x_1, \ldots, x_k\}$.

Step 1 Draw a random partition τ of the set $(x_1, \ldots, x_k)$ conditionally to $Z(\mathbf{x}) = z$, according to the distribution:

$$\mathbb{P}[T = \tau \mid Z(\mathbf{x}) = z] \propto \prod_{j=1}^{|\tau|} \lambda_{\mathbf{x}_{\tau_j}}^{(\alpha)}(z_{\tau_j}) \int_{\{\mathrm{u}_j < z_{\tau_j^c}\}} \lambda_{\mathbf{x}_{\tau_j}^c | \mathbf{x}_{\tau_j}, z_{\tau_j}}^{(\alpha)}(\mathrm{u}_j)\mathrm{du}_j. \tag{18}$$

Step 2 Given a partition $\tau = (\tau_1, \ldots, \tau_l)$, draw l independent extremal functions $(\phi_1^{(\alpha)+}, \ldots, \phi_l^{(\alpha)+})$ at $\mathbf{s}$ with finite dimensional distribution

$$\mathbb{P}[\phi_j^{(\alpha)+}(\mathbf{s}) \in \mathrm{dv} \mid Z(\mathbf{x}) = z, T = \tau] \propto \left\{ \int \mathbb{1}_{\{\mathrm{u}_j < z_{\tau_j^c}\}} \lambda_{(\mathbf{s}, \mathbf{x}_{\tau_j}^c) | \mathbf{x}_{\tau_j}, z_{\tau_j}}^{(\alpha)}(\mathrm{v}, \mathrm{u}_j)\mathrm{du}_j \right\} \mathrm{dv}. \tag{19}$$

Set $Z^+(s_j) = \max(\phi_1^{(\alpha)+}(s_j), \ldots, \phi_l^{(\alpha)+}(s_j))$ for $j \in 1, \ldots, m$.

Step 3 Independently of the two previous steps, draw a number N sufficiently large of $\phi^{(\alpha)-}$ at $(\mathbf{s}, \mathbf{x})$ such that $\phi^{(\alpha)-}(\mathbf{x}) \leq z$ and set

$$Z^-(s_j) = \max_{i \geq 1} \{\phi_i^{(\alpha)-}(s_j)\} \text{ for } j \in 1, \ldots, m, \tag{20}$$

Then $\tilde{Z}(\mathbf{s}) = \max\{Z^+(\mathbf{s}), Z^-(\mathbf{s})\}$ is distributed according to the conditional distribution of $Z(\mathbf{s})$ given $Z(\mathbf{x}) = z$.

Step 1 is the most challenging one because of the tremendous number of partition candidates. Indeed the number of possible partitions is equal to the Bell number (Spivey, 2008). This number grows very fast with the number of conditioning points. For instance, for $k = 10$, it is already equal to 115 975. Direct calculation of the probability in equation (18) is therefore not possible for a realistic number of conditioning points. In order to bypass this computational issue, Dombry et al. (2013) propose drawing the partition using a Gibbs sampler (Casella and George, 1992) which changes the class partition of one observed value at a time. Then once the transition probabilities involved simplified, the chain runs through the space of all possible partitions and converges to the target law described in equation (18).

Nevertheless, there are still some issues that need to be solved. The major problem is that the integrals involved in the transition distribution (see equation (18)) have to

be calculated at each step of the Gibbs sampler. The dimension of these integrals can be as high as $(k-1)$. The values of the integrals can be estimated by classic Monte-Carlo methods or quasi Monte-Carlo methods (Caflisch, 1998) but in both cases, the estimation requires a huge number of points to get satisfying results and the Gibbs sampling can become computationally heavy.

For Step 2, it is necessary to simulate from the distribution $\lambda^{(\alpha)}_{\mathbf{s}|\mathbf{x},z}$. Using the same arguments as in Dombry et al. (2013), it can be shown that $\lambda^{(\alpha)}_{\mathbf{s}|\mathbf{x},z}(u)$, $u \in \mathbb{R}^m$ can be rewritten as

$$\lambda^{(\alpha)}_{\mathbf{s}|\mathbf{x},z}(u) = \pi^{-m/2}(k+\alpha)^{-m/2}|\tilde{\Sigma}|^{-1/2}\left\{1+\frac{(u-\mu)^T\tilde{\Sigma}^{-1}(u-\mu)}{k+\alpha}\right\}^{\frac{-(m+k+\alpha)}{2}}\frac{\Gamma(\frac{k+m+\alpha}{2})}{\Gamma(\frac{k+\alpha}{2})} \quad (21)$$

with

$$\mu = \Sigma_{\mathbf{s}:\mathbf{x}}\Sigma_{\mathbf{x}}^{-1}z, \quad (22)$$

and

$$\tilde{\Sigma} = \frac{a_{\mathbf{x}}(z)}{k+\alpha}(\Sigma_{\mathbf{s}} - \Sigma_{\mathbf{s}:\mathbf{x}}\Sigma_{\mathbf{x}}^{-1}\Sigma_{\mathbf{x}:\mathbf{s}}) \text{ with } \Sigma_{\mathbf{s},\mathbf{x}} = \begin{pmatrix}\Sigma_{\mathbf{s}} & \Sigma_{\mathbf{s}:\mathbf{x}}\\ \Sigma_{\mathbf{x}:\mathbf{s}} & \Sigma_{\mathbf{x}}\end{pmatrix} \text{ and } a_{\mathbf{x}}(z) = z^T\Sigma_{\mathbf{x}}^{-1}z. \quad (23)$$

Then, it appears that $\lambda^{(\alpha)}_{\mathbf{s}|\mathbf{x},z}(u)$, $u \in \mathbb{R}^m$ is the density of a multivariate Student distribution with $(k+\alpha)$ degrees of freedom, with location parameter μ and scale matrix $\tilde{\Sigma}$.

Therefore $\mathbb{1}_{\{\mathrm{u}_j < z_{\tau_j^c}\}}\lambda^{(\alpha)}_{(\mathbf{s},\mathbf{x}^c_{\tau_j})|\mathbf{x}_{\tau_j},z_{\tau_j}}(\mathrm{v},\mathrm{u}_j)$ involved in the integral of (19) is a truncated multivariate Student distribution and step 2 results in a simulation under constraints. A procedure to perform this simulation is proposed and detailed in the following section.

Step 3 is the most accessible one. It consists in simulating values for an extremal t process at every conditioning and unconditioning location and check whether they respect the constraints given by the conditioning points using an acceptance/rejection procedure. In this context, the rejection rate of the algorithm is not very high even for a large number of conditioning points. Indeed, it rarely happens that one of the $\phi_i^{(\alpha)-}$ becomes greater than the conditioning value, because in this model, the conditioning values are meant to be, by construction, extreme realizations of these functions.

To simulate Z^-, especially the Poisson point process ξ, we adapted the simulation techniques described in Schlather (2002) for the Schlather process by noting that

$$\xi = \left\{\left(\sum_{i=1}^{n} e_i\right)^{-1/\alpha}\right\}_{n\in\mathbb{N}}, \quad (24)$$

with $e_i \sim \mathcal{E}(1)$ a standard exponential distribution, is a Poisson point process with intensity measure $\alpha t^{-(\alpha+1)}dt$.

By construction the sequence ξ is decreasing and for i large enough $\xi_i W_i(.)$ will not

participate to the maxima Z^- for any $x \in \mathcal{X}$ provided that $W_i(x) < C$ for all $x \in \mathcal{X}$. C can be set to 5 and we choose the number N of sub-extremal functions such that $5 * \xi_N < Z^-(x)$ for all $x \in \mathcal{X}$. This number depends on α and increases with it.

3.3. Multivariate Student under constraints

A straightforward way to sample from a truncated multivariate Student distribution is to simulate from a multivariate Student distribution and add an acceptance/rejection procedure. Unfortunately, the rate of acceptance of this algorithm becomes very low (approximatively $< 10^{-7}$ but depending on the parameters) for a number of conditioning points higher than 30 and this prevents its use in this case. We propose using instead the general methodology developed by Geweke [1991] for sampling a p-dimensional vector y from a multivariate Student distribution with p linear constraints i.e.

$$y \sim t_\nu(\mu, \Sigma), \text{with } a \leq Dy \leq b,$$

with D a full rank $p \times p$ matrix and some elements of the p-dimensional vectors a and b can be equal respectively to $-\infty$ and $+\infty$. The algorithm takes advantage on the construction of a Student variable as a ratio between a normal variable and a normalized χ^2 variable (or a normalized Gamma variable if the degree of freedom is not an integer). The method involves two steps simulating alternatively the denominator and the numerator of the multivariate Student random vector.
At the very first step a multivariate normal vector is drawn. Then the denominator is simulated as a constrained χ^2 variable via an acceptance/rejection procedure (on the ratio between the multivariate normal realization and the χ^2 or Gamma variable) and each component of the numerator is updated sequentially according to a Gibbs sampler. Indeed, Geweke [1991] gives the distribution of each component of the multivariate normal vector given the current values of the other components. This distribution is actually based on a truncated multivariate normal distribution, which fulfils the constraints. The whole procedure (drawn of the χ^2 variable and update of every component of the multivariate normal distribution) is repeated until convergence.

It has to be noted that the rejection rate of the χ^2 or Gamma variable simulation is low because the constraint is univariate, and that the Gibbs sampler avoids using an acceptance/rejection procedure on the multivariate part. In addition, it can be observed that whatever the number of iterations of the Gibbs sampler each random vector simulated this way respects by construction the constraints. By increasing the number of iterations, we intend to achieve the convergence of the distribution to the truncated Student distribution. The adaptation of this methodology to the step (2) (equation (19)) of the algorithm is straightforward setting for each $j \in \{1, \ldots, l\}$, all a components to $-\infty$, D to the $(m + |\tau_j^c|)$-dimensional identity matrix , the first m components of b to $+\infty$ and the $|\tau_j^c|$ last ones to $z_{\tau_j^c}$. Performing this way the number of conditioning points that can reasonably be handled is increased up to several hundreds.

4. Comparison criteria

For assessing the quality of the results, criteria comparing distributions resulting from the conditional simulations and a value corresponding to an observation have to be used.

The accuracy of conditional simulations is assessed using first the continuous ranking probability score crps (Gneiting and Raftery, 2007). Letting F be the predictive cumulative distribution function derived from the n conditional simulations $\{\hat{y}^{(j)}\}_{j=1,\dots,n}$ and y the test observation, the continuous ranking probability score is written

$$\text{crps}(F; y) = \int_{-\infty}^{+\infty} \big[F(t) - H(t-y)\big]^2 dt, \tag{25}$$

where $H(u)$ is the Heavyside function which is equal to 0 for $u < 0$ and 1 otherwise. The steeper the cdf is around the real value, the better. A good score is a score close to 0. It also appears that the calculation of the crps will result in a value with the same unit as the variable of interest.

Generally, there is no closed-form expression for this integral. Nevertheless, when F denotes a GEV distribution, a closed-form expression exists (Friederichs and Thorarinsdottir, 2012).

Moreover, the conditional simulations are also evaluated based on the quantile score (qs) defined as

$$\text{qs}_p(F; y) = \nu_p(y - F^{-1}(p)) \tag{26}$$

where p is the probability of interest and $\nu_p(u) = pu$ if $u \geq 0$ and $\nu_p(u) = (p-1)u$ otherwise. As shown by Friederichs and Thorarinsdottir (2012), the crps is proportional to the integral of the qs_p over all possible probabilities p. Then a good qs_p is a qs_p close to 0 as well. By choosing a high probability p, the quantile score focuses on the predictive performance in the upper tail (Gneiting and Raftery, 2007).

Then, overall scores are built from equations (25) and (26), by averaging on all points:

$$\begin{aligned} \overline{\text{CRPS}}((F_1, \dots, F_m); (y_1, \dots, y_m)) &= \frac{1}{m} \sum_{i=1}^{m} \text{crps}(F_i; y_i). \\ \overline{\text{QS}_p}((F_1, \dots, F_m); (y_1, \dots, y_m)) &= \frac{1}{m} \sum_{i=1}^{m} \text{qs}_p(F_i; y_i). \end{aligned} \tag{27}$$

In addition to these two univariate scores, which assess the marginal qualities of the conditional simulation, we propose a multivariate score:

$$\text{MS}_d(\hat{y}_1, \dots, \hat{y}_n; y_1, \dots, y_m) = \sum_{I \in \mathcal{P}_m(d)} \sum_{j=1}^{n} ||\hat{y}_I^{(j)} - y_I||_d \tag{28}$$

with $||.||_d$ the euclidean norm in $\mathbb{R}^d$, $\mathcal{P}_m(d)$ the ensemble of indices of length d in $\{1,\ldots,m\}$, $y_I = \{y_i | i \in I\}$ and $\hat{y}_I^{(j)} = \{\hat{y}_i^{(j)} | i \in I\}$. This score enables us to assess the spatial coherency of the conditional simulations.

In order to compare models, skill scores can be built (Gneiting and Raftery, 2007):

$$S^{skill} = \frac{S^{mod} - S^{ref}}{S^{opt} - S^{ref}} \tag{29}$$

with S^{mod}, S^{ref} and S^{opt} the score of the current model, the score of a reference model and the optimal score respectively. For models that improve on the reference model, S^{skill} belongs to $[0,1]$ with 1 correspond to the optimal case. For $\overline{\text{CRPS}}$, $\overline{\text{QS}}_p$ and MS_d, $\overline{\text{CRPS}}^{opt} = \overline{\text{QS}}_p^{opt} = \text{MS}_d^{opt} = 0$ and the corresponding skill scores can be written for these three scores as

$$\text{CRPSS} = \frac{\overline{\text{CRPS}}^{ref} - \overline{\text{CRPS}}^{mod}}{\overline{\text{CRPS}}^{ref}}, \text{QSS}_p = \frac{\overline{\text{QS}}_p^{ref} - \overline{\text{QS}}_p^{mod}}{\overline{\text{QS}}_p^{ref}}, \text{MSS}_d = \frac{\text{MS}_d^{ref} - \text{MS}_d^{mod}}{\text{MS}_d^{ref}}. \tag{30}$$

which corresponds to the percentages of $\overline{\text{CRPS}}$, $\overline{\text{QS}}_p$ or MS_d improvement brought by the current model with respect to the reference model.

5. Simulation Study

In order to check the properties of the conditional simulations provided by this algorithm, simulations according to extremal t processes are drawn, and conditional simulations at non conditioning points (test sample) are compared to their real values. 100 points are sampled at random in a unit square. 20 are kept as conditioning points, and the other 80 are used as test values. For all simulations the covariance function is the exponential with scale parameter 1.

5.1. Influence of the partition

As stated in the section 3, the first and hardest step for conditional simulation is to draw a partition of the conditioning points according to the distribution given in equation (18). Indeed, it appears that the Gibbs sampler converges very slowly to the target law– in the sense that the number of iterations required to reach the objective distribution may be very high– and then it is likely that we do not draw the partition from the target distribution. We investigate how the resulting partition distribution, obtained at different convergence stages of the Gibbs sampler, influences the performance of the whole algorithm. For $\alpha = 2$, we simulate extremal t fields and compare the performance of the algorithm when (i) the partition which is used to build the data is known and has probability one (ii) when it is drawn from a Gibbs sampler (10 000 iterations) according to the distribution (18) and (iii) when the partition is drawn randomly. Table 1 gives the CRPSS calculated by taking the random partition model as the reference. Furthermore, this table also shows the absolute bias ($\frac{1}{nm}\sum_{i,j} |\overline{\widehat{y}_i}^{(j)} - y_i|$)

and variance ($\frac{1}{nm}\sum_{i,j}(\widehat{y}_i^{(j)} - \overline{\widehat{y}_i}^{(j)})^2$), both calculated on the margins. For each method, 50 simulations are performed.

	Known Part.	Gibbs sampler Part.	Random Part.
CRPSS	**9.5%**	5.0%	0%
Bias	**0.95**	1.33	1.77
Variance	**2.50**	4.16	4.42

Table 1: Influence of the partition distribution: values of the CRPSS, bias and variance for $\alpha = 2, (i)$ when the partition with the highest probability is used, (ii) when the partition is drawn from a Gibbs sampler (10 000 iterations) according to the distribution (18) and (iii) when the partition is drawn randomly.

As expected, there is an improvement of the quality of the conditional simulations when the partition is drawn in the distribution giving probability one to the partition used for building the data. It is important to see that, in addition to the expected variance reduction there is also an improvement of the bias. This demonstrates the strong gain to be had by drawing close to the right distribution and confirms that the efforts made for this step of the algorithm are inevitable and should even be developed. Indeed, the gain in drawing the partition with the Gibbs sampler is not very high compared to using a random partition. We can think that drawing randomly the partition instead of performing a Gibbs sampler, for time-saving reasons for instance, would not change considerably the results especially when the number of conditioning points is high.

It also has to be noted that the impact of drawing the partition in the right distribution instead of randomly largely depends on the configuration of the data. Indeed, if the observed values are of the same order of magnitude, not very large relative to the unit Fréchet scale, partitions with numerous classes have similar weights, and will lead to similar results as random partitions. Contrarily when some of the observed values are rather large, then partitions with few classes will have significant weight and will lead to very different results than random partitions.

5.2. Influence of α

We investigate how the value of the parameter α influences the quality of the conditional simulations. For each value of α, 15 simulations of an extremal t process are drawn and 20 conditioning points are sampled for each simulation. Then 50 conditional simulations are performed according to each conditioning point set. The partition which is used to build the data is known, just as α and $\rho(.)$ which are not estimated.

The conditional simulations of the extremal t process are compared to conditional simulations of a Schlather process (Schlather, 2002), a Brown-Resnick process (Brown and Resnick, 1977), and a Gaussian process with the same conditioning points and unconditional simulations of the extremal t process. Ordinary Kriging (Chilès and Delfiner, 1999) is set as the reference prediction for the CRPSS calculation. The conditional simulations of the Schlather and Brown-Resnick processes are performed

according to the algorithm of Dombry et al. (2013) and with `SpatialExtremes` (Ribatet, 2013b). For the Gaussian process conditional simulations, we apply first a Gaussian margin transformation and then perform the methodology described in Lantuéjoul (2002) for Gaussian processes (even if it is only marginally Gaussian). For the ordinary kriging and the Gaussian simulations, the covariance function is estimated from the data. Furthermore, for fair comparison purposes, each simulated dataset is marginally transformed into unit-Fréchet. Hence, the different values of the parameter α have only an impact on the spatial dependence of the simulated data.

Figure 1 describes the evolution of the different CRPSS with α. As expected since it is based on the right model, the extremal t model gives better CRPSS than do the other models. The other models present results which are not as good as the ones with the extremal t model, even if every CRPSS are positive which means that every tested models brings some improvement compared to the kriging model. It also appears that the unconditional extremal t simulations become more and more efficient as α increases, especially compared to the Schlather model. It is worth noticing that the right spatial dependence structure enables the model to provide good results even without conditioning points. The good performances of the unconditioning simulations may also be explained by the fact that when α increases, the dependence among extremes becomes weaker and then the influence of the conditioning values becomes less important.

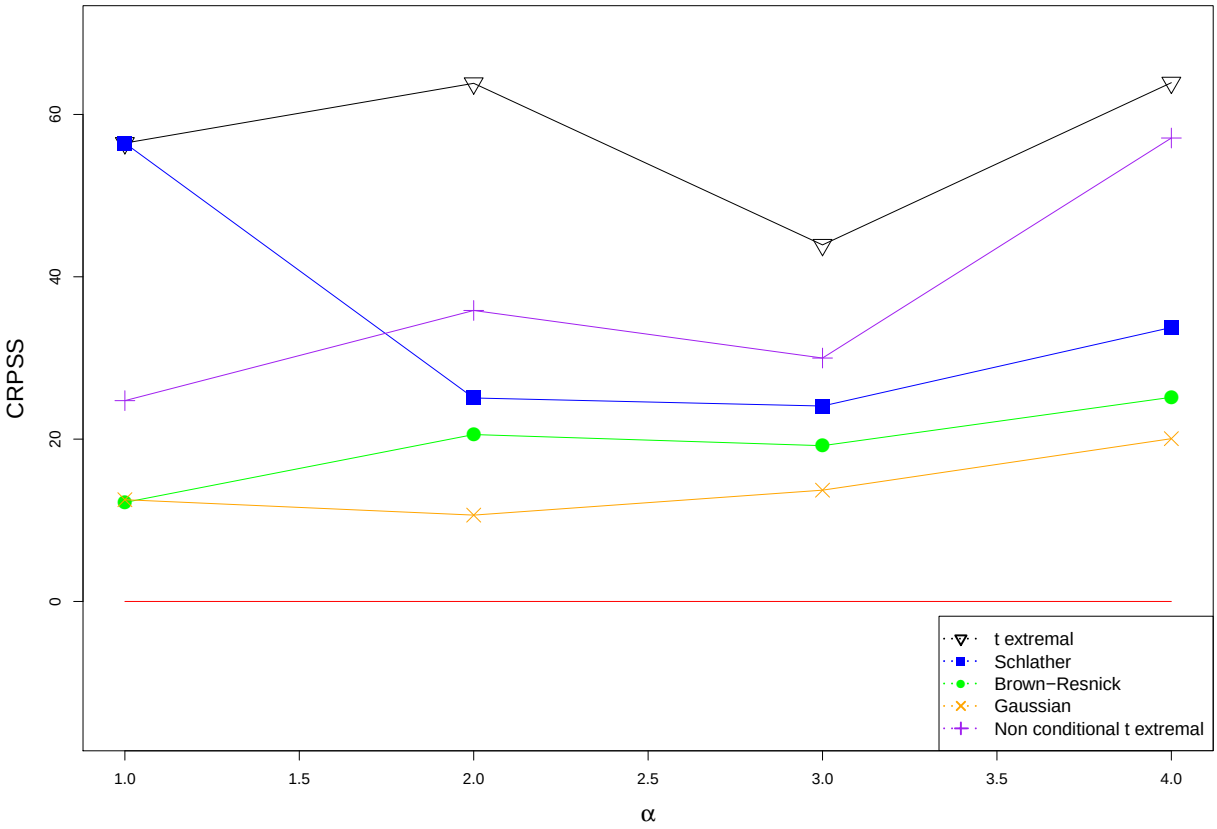


Figure 1: Evolution of the CRPSS (in %) with the α with ordinary kriging as reference.

5.3. Asymptotic dependence properties

We now investigate a notion related to the asymptotic dependence for the conditional simulations. As mentioned by Dombry et al. (2013), $Z(.)|\{Z(\mathbf{x}) = z\}$ is not max-stable and cannot be asymptotically dependent. However, the process integrated on all the conditional events recovers the max-stability property, and its asymptotic dependence can be characterized by the madogram (Cooley et al., 2006). Using the madogram we compare the pairwise extremal coefficient estimates θ_{est} to the theoretical extremal coefficient function $\theta_{th}(h)$ at a distance h in order to check that the asymptotic dependence behavior of the conditional simulations is similar to that of the theoretical one.

For each $\alpha \in \{2, 4, 6\}$, we simulate 1 000 extremal t fields at the same 20 conditional points and estimate the extremal coefficient function from the remaining 80 points with the `R` package `SpatialExtremes` (Ribatet, 2013b). Figure 2 compares the theoretical extremal function $\theta_{th}(h)$ and the estimated one $\theta_{est}(h)$ calculated from the madogram of the integrated conditional simulations.

For the three values of α, the conditional simulations respect the expected asymptotic dependence behavior. Indeed, the theoretical extremal coefficient function fits relatively well the pairwise extremal coefficient estimates, although it tends to overestimate them when α becomes high.

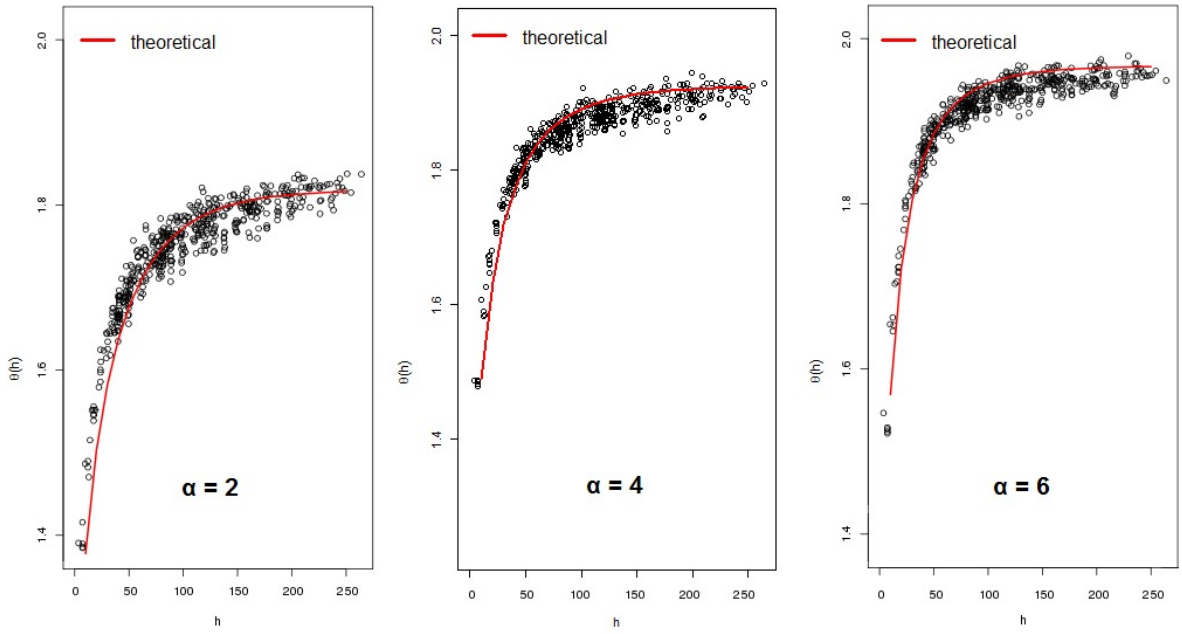


Figure 2: Extremal coefficient: comparison of $\theta_{th}(h)$ and $\theta_{est}(h)$ for $\alpha \in \{2, 4, 6\}$.

6. Application to precipitation data

6.1. Description and margin modeling

To check the behavior of the model and the algorithm on real data, we extracted daily precipitation data from the SAFRAN analyses (Quintana-Seguí et al., 2008) in the southeast of France covering the region [3.4°E, 6.0°E] x [43.2°N, 44.5°N] – as represented by the red square in Figure 3 – with a spatial resolution of 8 km $\times$ 8 km, which corresponds to 457 grid cells. Then, for each cell, we computed the autumnal maximum daily precipitation (from 91 days between September 1^{st} and November 30^{th}) for $1960 - 2007$: hence, for each grid cells, we have 48 autumnal maximum daily precipitation values. 50 grid cells have been randomly selected as conditioning points. Conditional simulations are then performed on the 407 remaining grid cells given the observation of the 50 conditioning spots.

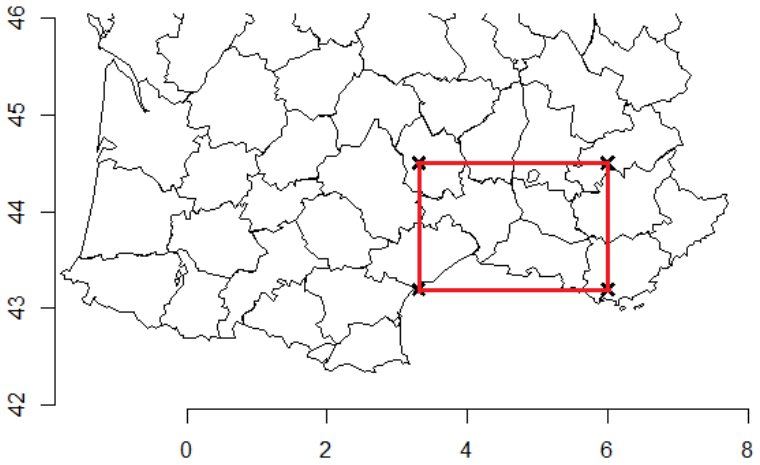


Figure 3: Study area of SAFRAN data subset.

The choice of the autumnal period allows us to make the temporal stationarity assumption, as it is commonly accepted that rainfall data can be considered as stationary within a given season.

In order to perform the conditional simulations presented in section 3 the margins have first to be transformed into an α-Fréchet distribution. If Z is distributed according to a GEV(μ, σ, ξ), then Y defined as

$$Y = \left(1 + \frac{\xi(Z-\mu)}{\sigma}\right)_+^{1/\alpha\xi} \tag{31}$$

follows an α-Fréchet distribution (Beirlant et al., 2004).

To avoid the estimation of the parameters (μ, σ and ξ) of the GEV for all points and to deal with spatial non-stationarity, these parameters are modeled as functions of the spatial coordinates LAT and LON except ξ that is assumed to be constant for the whole region:

$$\begin{aligned} \mu &= \exp\left[\mu_0 + \mu_1 * LAT + \mu_2 * LON\right] \\ \sigma &= \exp\left[\sigma_0 + \sigma_1 * LAT + \sigma_2 * LON\right] \\ \xi &= \xi_0 \end{aligned} \tag{32}$$

It can be observed that the exponential formulation is employed for μ and σ as σ has to be positive for model definition reasons and μ is chosen to be positive because of the nature of the observed phenomenon (precipitation).

6.2. Inference

The parameters of the model, i.e.:

1. μ_0, μ_1, μ_2, σ_0, σ_1, σ_2, ξ, the parameters for the margin model,
2. α, the extremal t parameter,
3. and a the scale parameter of the exponential correlation function of W,

have to be estimated from the data. Nevertheless, for dimensions higher than 2, the multivariate density of the extremal t process is not tractable and usual likelihood maximization methods are not available. Alternatively, composite likelihood (Lindsay, 1988; Varin et al., 2011) can be used. It consists in calculating bivariate densities for each pair of points and maximizing their weighted sum. Let us note $\{z_{m,x_i}\}$ the autumnal maximum on $m = 1, \ldots, M$ years for a given x_i in $(x_1, \ldots, x_k)$. The composite log-likelihood is

$$\mathcal{LC}(\Psi; \mathbf{z}) = \sum_{m=1}^{M} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} w_{ij} l_{ij}(\Psi) \tag{33}$$

with $l_{ij}(\Psi) = \log f(z_{m,x_i}, z_{m,x_j}; \Psi)$ the bivariate log-density for points x_i and x_j with Ψ the vector of parameters to estimate, and w_{ij} the corresponding weights. These weights w_{ij} have been chosen to be decreasing (exponentially) with the distance between x_i and x_j. The maximization of this composite likelihood leads to consistent estimations of the target parameters (Padoan et al., 2010; Xu and Reid, 2011).

6.3. Results

The inference has been driven by the data over the whole time period. The maximization of the composite likelihood gives $\alpha = 2.2$ (with $\theta_\infty = 1.83$) which means that even for long distances there is still some asymptotic dependence. Furthermore, the estimated range of the exponential covariance function is equal to 360 km.The estimated parameters and the confidence intervals are presented in Table B.1.

The validation is performed only for the year 2007 with 100 conditional simulations. The most represented partition in the history of the Gibbs sampler is composed of 5 classes. However, there is no notable spatialization of these classes (not shown).

Figure 4 shows the individual crps scores (as defined in equation (25)) computed for the year 2007 taking the observed values as test observations (y_i) and the empirical distributions of the conditional simulations as the predictive cdf (F_i).

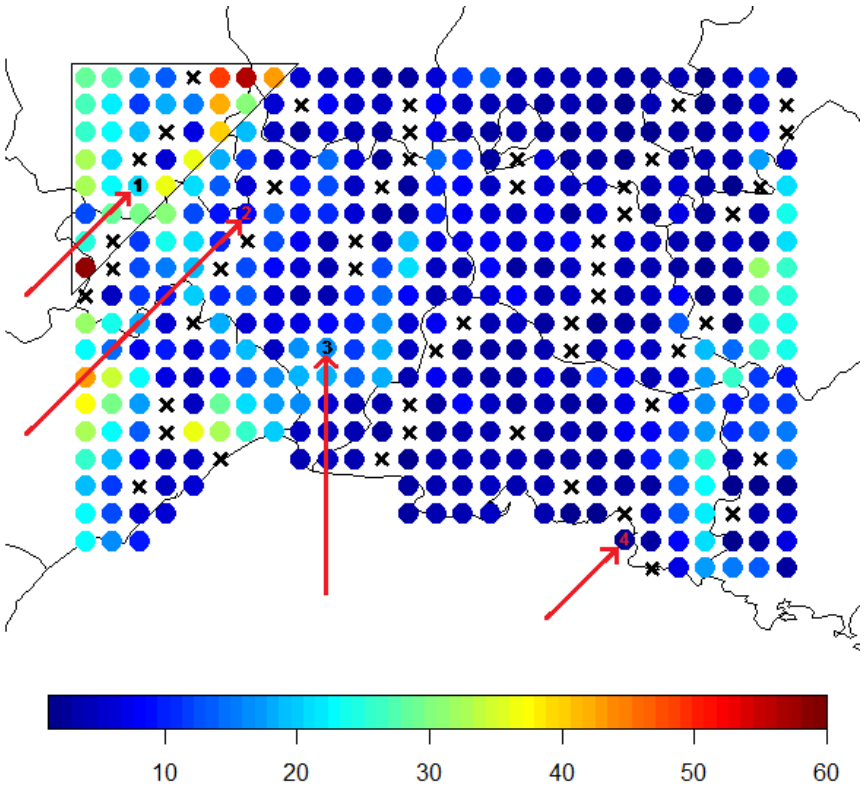


Figure 4: Map of the crps (in mm) for 2007. For further illustrations, four grid cells have been arbitrarily selected and are indicated here (from 1 to 4).

In addition, Figure 5 describes the distributions of the conditional simulations at four arbitrarily chosen grid cells specified on Figure 4. It appears naturally that the closer to conditioning points the new location is, the better the simulation (e.g. points 2 and 4 are close to conditioning points whereas point 3 is far). However, this is clearly not the only important criterion. Indeed, point 1 is quite close to conditioning points yet its simulations are not as good as expected.

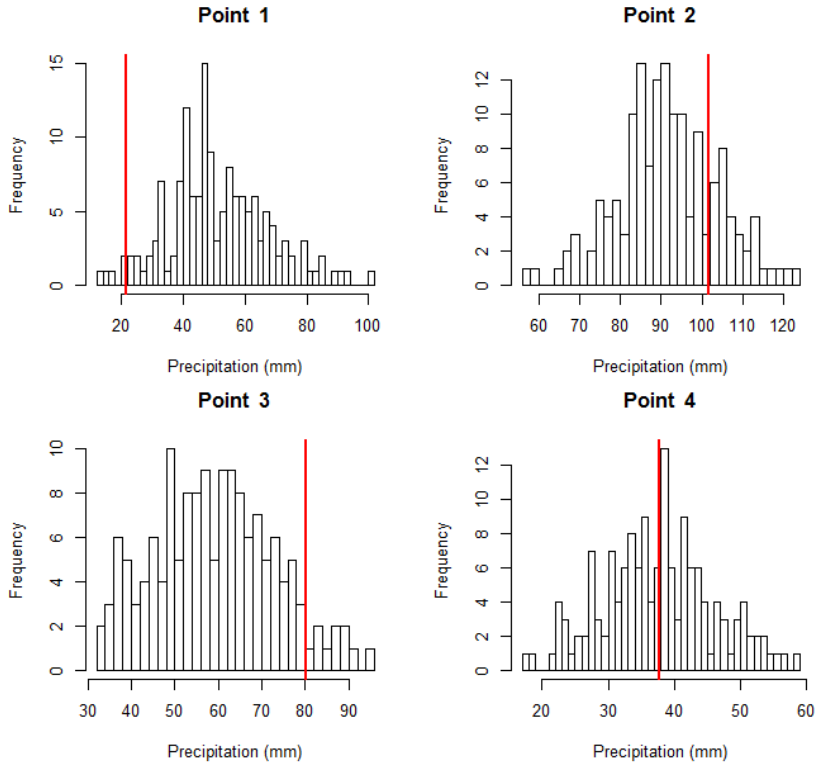


Figure 5: Histograms of the conditional simulations at 4 locations, with the real value (in red).

In order to evaluate the quality of these conditional simulations, other models and algorithms are considered and compared:

(A) conditional simulation of extremal t model, with Gibbs sampler,

(B) conditional simulation of extremal t , with random partitions,

(C) analog model,

(D) historical model,

(E) unconditional simulation of extremal t model.

For simulations based on the analog model (C) (e.g., Barnett and Preisendorfer, 1978; Zorita and Von Storch, 1999; Pascal et al., 2013), at each non conditioning location we consider the nearest conditioning grid cell. Then we look at the history of this point (1960-2006) to find the year with the closest observed value: the value observed this year at the non conditioning location is taken as the estimate value for the non conditioning point. For the so-called "historical" model (D), we randomly draw a value in the history of a non conditioning location and consider it as the value of the simulation.

Table 2 gives the $\overline{\text{CRPS}}$, CRPSS, $\overline{\text{QS}_{95}}$, QSS_{95}, MS_2 and MSS_2 for each method with the historical model (D) as reference.

Method	$\overline{\text{CRPS}}$	CRPSS	$\overline{\text{QS}_{95}}$	QSS_{95}	MS_2	MSS_2
(A) Cond. Sim. Gibbs	10.05	-4.1%	**1.95**	**34.8%**	27.55	31.3%
(B) Cond. Sim. Random	10.96	-13.1%	2.02	32.4%	28.99	27.7%
(C) Analog	10.53	-9.1%	5.68	-90.0%	**18.41**	**54.1%**
(D) Historical	**9.65**	**0%**	2.99	0%	40.11	0%
(E) Uncond. Sim.	12.71	-31.6%	3.33	-11.4%	48.07	-19.9%

Table 2: $\overline{\text{CRPS}}$, CRPSS, $\overline{\text{QS}_{95}}$, QSS_{95}, MS_2 and MSS_2 for the methods (A), (B), (C), (D) (as reference) and (E).

The historical model appears as the best one regarding the $\overline{\text{CRPS}}$ score. In term of MS_2, the model analog, which respects by construction the spatial coherency, gives the best results. Model (A) shows good results in terms of $\overline{\text{QS}_{95}}$ and generally appears to be the most versatile method adapted to different purposes (for each score, it gets at worst the second best value). It is also important to note that method (A) is better than method (B): there is a real gain in running a Gibbs sampler to estimate the partitions.

Generally, models (A)-(B) are the best to catch high quantile behavior (see $\overline{\text{QS}_{95}}$). Nevertheless, as crps is the overall score of qs_p integrated over all probabilities p, it indicates that for lower probabilities, model (A) is not as good as the others. It appears that for $p \geq 0.75$, $\overline{\text{QS}_p}$ of model (A) calculated over the whole region is the best.

Furthermore, by deepening the analysis of the results, it appears that for some points the error (in terms of crps) is fairly high, especially for model (A). Those points are located in the upper left corner of Figure 4 (i.e., the northwest of the region), corresponding to the higher region(s) of the Cevennes mountains. The climate characteristics in this area are very different from the rest of the region and it is likely that the chosen covariates (LAT and LON) are not sufficient to overcome the inhomogeneity of this region.

Then, for comparison purposes inside the stationary framework, we removed these points for the calculation of the scores. Table 3 shows the new scores.

Method	$\overline{\text{CRPS}}$	CRPSS	$\overline{\text{QS}_{95}}$	QSS_{95}	MS_2	MSS_2
(A) Cond. Sim. Gibbs	**8.85**	**4.4%**	**1.78**	**37.2%**	30.14	21.8%
(B) Cond. Sim. Random	9.46	-2.1%	1.80	36.5%	30.79	20.0%
(C) Analog	10.14	-9.4%	5.55	-95.7%	**18.15**	**52.9%**
(D) Historical	9.27	0%	2.83	0%	38.51	0%
(E) Uncond. Sim.	11.13	-20.3%	3.09	-9.1%	47.24	-22.7%

Table 3: $\overline{\text{CRPS}}$, CRPSS, $\overline{\text{QS}_{95}}$, QSS_{95}, MS_2 and MSS_2 for the methods (A), (B), (C), (D)(as reference) and (E) with removal of the outliers area.

The CRPSS is now better for method (A), with a gain of 4.4% with respect to the reference model and both QSS_{95} and MSS_2 remain good. It also appears that the $\overline{\text{CRPS}}$

and CRPSS have been improved for all methods.

It is also interesting to check whether the conditional simulations follow the temporal variations of the real values. To this aim, we have simulated 50 conditional fields on 50 grid cells from the SAFRAN database, on the period $2000 - 2007$ with the same estimated parameters. As running the Gibbs sampler for each year with 50 conditional points would be too time-consuming, the partitions are drawn randomly for each year. Figure 6 shows the temporal evolutions of the means of the conditional simulations at the four previously chosen grid cells.

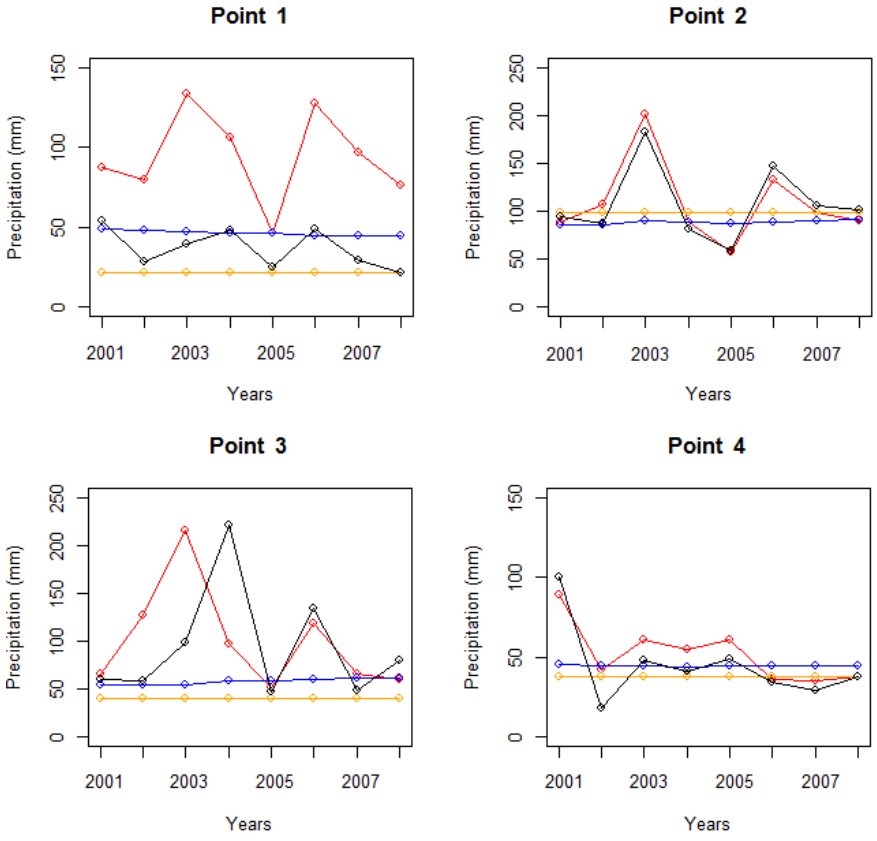


Figure 6: Comparison of the evolution of the mean of the conditional simulations (in red), the historical model (in blue), analog model (in orange) and the test values (in black).

For points 2 and 4, the temporal evolutions of the conditional simulations are almost the same as for the test values, whereas for point 3, there is a big discrepancy in the first period and a good match in the second period. A constant significant bias is noted for point 1, which is a location in the already mentioned mountain region, and it can be suspected that it behaves in a quite different way than do the conditioning points even though it approximately follows the temporal variations.

The scatterplot of the simulated values vs. the test values is represented in Figure

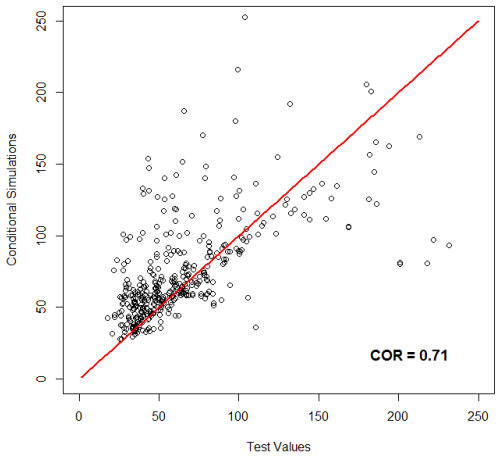


Figure 7: Correlation between the simulated values and the test values.

7. The Pearson correlation coefficient is equal to 0.71. It appears that the conditional simulations roughly follow the intrinsic variations of the data over the years. Indeed, a correlation coefficient of 0.71 is a satisfying result which indicates that the method is able to reproduce the general evolution of the observed phenomenon. In comparison, the correlation coefficient is 0.47 for the historical model and 0.33 for the analog model.

7. Discussion and conclusions

We have investigated conditional simulations for maxima precipitation using the extremal t model. Thanks to its parametrization, this model offers a great flexibility in modeling the asymptotic dependence. The conditional simulation algorithm designed by Dombry et al. (2013) for Brown-Resnick and Schlather models has been adapted to the extremal t process and the second step, consisting in simulating from a multivariate Student distribution under constraints, has been improved by a two-state Gibbs sampler proposed by Geweke (1991) enabling us to handle a greater number of conditioning points.

A simulation study highlighted that the methodology of conditional simulations for the extremal t process is effective for a large range of α ($\in [1, 6]$). Moreover, when the value of α is too high, its estimation becomes more and more uncertain. In such cases, we would advise the use of the conditional simulations of Gaussian processes with transformed Gaussian margins (Oesting and Schlather, 2014; Lantuéjoul, 2002) which are faster than the general Brown-Resnick processes (Kabluchko et al., 2009) which show similar properties to those of an extremal t process for high values α.

A major aspect that needs to be improved is the way to generate the partitions of the conditioning points. Indeed, the Gibbs sampler requires too many computational resources to provide better results. Some alternative methods have to be considered

in order to avoid the calculations of the multiple integrals involved in the transition probabilities.

The application performed on a dataset of rainfall in the south of France showed that the model is relevant and that the algorithm performs well, provided that the spatial and temporal assumptions of stationarity are fulfilled. The margins modelling and transformation (equations (33) and (34)) intended to make the transformed process spatially stationary. However, the sole use of the coordinates (LAT+LON) is likely not sufficient to remove all the spatial heterogeneity of the process especially at the boundaries or where some other major climatic factors have to be taken into account. For this example, a small sub-region with different climate characteristics has been identified where the spatial coordinates do not seem to be sufficient to get satisfying results. Hence, the introduction of other covariates, such as altitude or information from climate or meteorological situations, could help to improve the simulations.

However, for simulations of fields of maxima precipitation in a future climate change context, no conditioning points (i.e., really observed at locations) are available. The only information stems from Global Climate Models (GCM) or Regional Climate Models (RCM) driven by various greenhouse gas emission scenarios. Those models provide many data but at low spatial resolution (from about 300 km down to about 10 km at best, so far). Hence, the best those models can give will be mean precipitation values over relatively large regions. Those data are clearly not adapted to feeding hydrological, ecological or, more generally, environmental impact models, which require high-resolution input climate data. Hence, the use of a max-stable process to simulate data for the future and provide those simulations to environmental impact models needs to be based on a change of spatial resolution. In other words, the max-stable process will not have "station" values available as conditioning points but only large-scale climate conditioning information. Therefore, one important need is to be able to condition max-stable processes by those large-scale climate data, which corresponds to performing a "statistical downscaling" to generate high-resolution fields of extreme precipitation. This downscaling in a spatial modeling framework will enable us to statistically generate fields and times series that will be useful for many impact studies trying to anticipate the potential consequences of extreme precipitation events in a climate change context.

Appendix A. Calculation of the density and the conditional density

Proof. According to equation (13), we have

$$\begin{aligned} \Lambda_{\mathbf{x}}^{(\alpha)}(A) &= \int_0^{+\infty} \mathbb{P}(m_\alpha^{-1/\alpha}\xi W(\mathbf{x}) \in A)\alpha\xi^{-(\alpha+1)}\mathrm{d}\xi \\ &= \int_0^{+\infty} \int_{\mathbb{R}^k} \mathbb{1}_{\{m_\alpha^{-1/\alpha}\xi z \in A\}} g_{\mathbf{x}}(z)dz\alpha\xi^{-(\alpha+1)}\mathrm{d}\xi \end{aligned} \tag{A.1}$$

with for $z \in \mathbb{R}^k$,

$$g_{\mathbf{x}}(z) = (2\pi)^{-(k/2)}|\Sigma_{\mathbf{x}}|^{-1/2}\exp[-\frac{1}{2}z^T\Sigma_{\mathbf{x}}^{-1}z].$$

Setting $y = m_\alpha^{-1/\alpha}\xi z$, we get

$$\Lambda_{\mathbf{x}}^{(\alpha)}(A) = \int_0^{+\infty}\int_A g_{\mathbf{x}}\Big(\frac{y}{\xi m_\alpha^{-1/\alpha}}\Big)\frac{\alpha}{m_\alpha^{-k/\alpha}}\xi^{-(k+1+\alpha)}\mathrm{d}y\mathrm{d}\xi \tag{A.2}$$

and then

$$\begin{aligned}\lambda_{\mathbf{x}}^{(\alpha)}(y) &= \int_0^{+\infty} g_{\mathbf{x}}\Big(\frac{y}{\xi m_\alpha^{-1/\alpha}}\Big)m_\alpha^{k/\alpha}\alpha\xi^{-(k+\alpha+1)}\mathrm{d}\xi. \\ &= \alpha(\sqrt{2\pi})^{-k}|\Sigma_{\mathbf{x}}|^{-1/2}m_\alpha^{k/\alpha}\int_0^{+\infty}\exp\Big[-\frac{y^T\Sigma_{\mathbf{x}}^{-1}y}{2m_\alpha^{-2/\alpha}\xi^2}\Big]\xi^{-(k+\alpha+1)}\mathrm{d}\xi \end{aligned}\tag{A.3}$$

By noting that for $l \geq 2$,

$$\int_0^{+\infty}\exp\Big[-\frac{z^T\Sigma_{\mathbf{x}}^{-1}z}{2m_\alpha^{-2/\alpha}\xi^2}\Big]\xi^{-l}\mathrm{d}\xi = 2^{(l-3)/2}\ [z^T\Sigma_{\mathbf{x}}^{-1}z]^{-(\frac{l-1}{2})}\ m_\alpha^{-(l-1)/\alpha}\ \Gamma\big(\frac{l-1}{2}\big) \tag{A.4}$$

$\lambda_{\mathbf{x}}^{(\alpha)}(\mathbf{z})$ can be written:

$$\begin{aligned}\lambda_{\mathbf{x}}^{(\alpha)}(z) &= \alpha\pi^{-\frac{k-1}{2}}\ |\Sigma_{\mathbf{x}}|^{-1/2}\ [z^T\Sigma_{\mathbf{x}}^{-1}z]^{-(\frac{k+\alpha}{2})}\ \Gamma\big(\frac{k+\alpha}{2}\big)\Gamma((\alpha+1)/2)^{-1} \\ &= \frac{\alpha\pi^{\frac{\alpha-1}{2}}}{w_{k+\alpha-2}}\ |\Sigma_{\mathbf{x}}|^{-1/2}\ [z^T\Sigma_{\mathbf{x}}^{-1}z]^{-(\frac{k+\alpha}{2})}\Gamma((\alpha+1)/2)^{-1} \end{aligned}\tag{A.5}$$

and

$$\lambda_{\mathbf{s}|\mathbf{x},z}^{(\alpha)}(u) = \frac{w_{k+\alpha-2}}{w_{k+m+\alpha-2}}\sqrt{|\Sigma_{\mathbf{s},\mathbf{x}}|^{-1}|\Sigma_{\mathbf{x}}|}\frac{[z^T\Sigma_{\mathbf{x}}^{-1}z]^{(k+\alpha)/2}}{[(u,z)^T\ \Sigma_{\mathbf{s},\mathbf{x}}^{-1}\ (u,z)]^{(k+m+\alpha)/2}}, \tag{A.6}$$

with $w_d = \pi^{d/2}/\Gamma(d/2+1)$ the volume of the unit ball in d dimensions. □

Appendix B. Estimated parameters

Parameters	Estimation	CI (95%)
μ_0	-3.0565	[-3.0942 ; -3.0187]
μ_1	-0.2227	[-0.2233 ; -0.2221]
μ_2	0.1804	[0.1803 ; 0.1805]
σ_0	-4.3431	[-4.3732 ; -4.3129]
σ_1	-0.3665	[-0.3675 ; -0.3654]
σ_2	0.2084	[0.2084 ; 0.2085]
ξ	0.1145	[0.1128 ; 0.1163]
α	2.2077	[2.0354 ; 2.3945]
a	1.2262	[1.1304 ; 1.3301]

Table B.1: Estimated parameters (and CI) with composite likelihood maximization.

References

Barnett, T., Preisendorfer, R., 1978. Multifield analog prediction of short-term climate fluctuations using a climate state vector. Journal of Atmospheric Sciences 35, 1771–1787.

Beaulant, A.-L., Joly, B., Nuissier, O., Somot, S., Ducrocq, V., Joly, A., Sevault, F., Deque, M., Ricard, D., 2011. Statistico-dynamical downscaling for mediterranean heavy precipitation. Quarterly Journal of the Royal Meteorological Society 137 (656), 736–748.

Beirlant, J., Goegebeur, Y., Teugels, J., Segers, J., Waal, D. D., Ferro, C., 2004. Statistics of extremes. Theory and applications. Hoboken, NJ: John Wiley & Sons.

Blanchet, J., Davison, A. C., 2011. Spatial modeling of extreme snow depth. The Annals of Applied Statistics 5 (3), 1699–1725.

Brown, B. M., Resnick, S. I., 1977. Extreme values of independent stochastic processes. Journal of Applied Probability 14 (4), 732–739.

Buishand, T., de Haan, L., Zhou, C., 2008. On spatial extremes: with application to a rainfall problem. The Annals of Applied Statistics, 624–642.

Caflisch, R. E., 1 1998. Monte carlo and quasi-monte carlo methods. Acta Numerica 7, 1–49.

Casella, G., George, E. I., 1992. Explaining the Gibbs sampler. The American Statistician 46 (3), 167–174.

Chilès, J.-P., Delfiner, P., 1999. Geostatistics: Modeling Spatial Uncertainty. Wiley, New-York.

Coles, S., Tawn, J., 1994. Statistical methods for multivariate extremes - an application to structural design. Journal of the Royal Statistical Society - Series C: Applied Statistics 43 (1), 1–48.

Cooley, D., Naveau, P., Poncet, P., 2006. Variograms for spatial max-stable random fields. In: Dependence in probability and statistics. Springer, pp. 373–390.

Davison, A. C., Gholamrezaee, M. M., 2012. Geostatistics of extremes. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science 468 (2138), 581–608.

Davison, A. C., Padoan, S. A., Ribatet, M., 05 2012. Statistical modeling of spatial extremes. Statistical Science 27 (2), 161–186.

De Haan, L., 1984. A spectral representation for max-stable processes. The Annals of Probability 12 (4), 1194–1204.

De Haan, L., De Ronde, J., 1998. Sea and wind: multivariate extremes at work. Extremes 1 (1), 7–45.

Dombry, C., Eyi-Minko, F., 2013. Regular conditional distributions of continuous max-infinitely divisible random fields. Electron. J. Probab 18 (7), 1–21.

Dombry, C., Éyi-Minko, F., Ribatet, M., 2013. Conditional simulation of max-stable processes. Biometrika 100 (1), 111–124.

Ducrocq, V., Nuissier, O., Ricard, D., Lebeaupin, C., Thouvenin, T., 2008. A numerical study of three catastrophic precipitating events over southern france. ii: Mesoscale triggering and stationarity factors. Quarterly Journal of the Royal Meteorological Society 134 (630), 131–145.

Fawcett, L., Walshaw, D., 2012. Estimating return levels from serially dependent extremes. Environmetrics 23 (3), 272–283.

Friederichs, P., Thorarinsdottir, T. L., 2012. Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. Environmetrics 23 (7), 579–594.

Fuentes, M., Henry, J., Reich, B., 2013. Nonparametric spatial models for extremes: application to extreme temperature data. Extremes 16 (1), 75–101.

Gaume, J., Eckert, N., Chambon, G., Naaim, M., Bel, L., 2013. Mapping extreme snowfalls in the french alps using max-stable processes. Water Resources Research 49 (2), 1079–1098.

Geweke, J., 1991. Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities. In: Computing science and statistics: Proceedings of the 23rd symposium on the interface. Citeseer, pp. 571–578.

Gneiting, T., Raftery, A. E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102 (477), 359–378.

IPCC, 2011. Managing the risks of extreme events and disasters to advance climate change adaptation. Tech. rep., Intergovernmental Panel of Climate Change (IPCC).

Kabluchko, Z., Schlather, M., de Haan, L., 09 2009. Stationary max-stable fields associated to negative definite functions. The Annals of Probability 37 (5), 2042–2065.

Lantuéjoul, C., 2002. Geostatistical simulation: models and algorithms. Springer.

Lebeaupin, C., Ducrocq, V., Giordani, H., 2006. Sensitivity of torrential rain events to the sea surface temperature based on high-resolution numerical forecasts. Journal of Geophysical Research: Atmospheres 111 (D12).

Lindsay, B. G., 1988. Composite likelihood methods. Contemporary Mathematics 80 (1), 221–39.

Mariethoz, G., Renard, P., Cornaton, F., Jaquet, O., 2009. Truncated plurigaussian simulations to characterize aquifer heterogeneity. Ground Water 47 (1), 13–24.

Oesting, M., Schlather, M., 2014. Conditional sampling for max-stable processes with a mixed moving maxima representation. Extremes 17 (1), 157–192.

Opitz, T., 2013. Extremal t-processes: Elliptical domain of attraction and a spectral representation. Journal of Multivariate Analysis 122, 409 – 413.

Padoan, S. A., 2011. Multivariate extreme models based on underlying skew-t and skew-normal distributions. Journal of Multivariate Analysis 102, 977–991.

Padoan, S. A., Ribatet, M., Sisson, S. A., 2010. Likelihood-based inference for max-stable processes. Journal of the American Statistical Association 105 (489), 263–277.

Pascal, Y., Tamara, S., Philippe, D., Laurent, M., Robert, V., Mathieu, V., 2013. Ensemble reconstruction of the atmospheric column from surface pressure using analogues. Climate dynamics 41 (5-6), 1333–1344, eng.

Quintana-Seguí, P., Le Moigne, P., Durand, Y., Martin, E., Habets, F., Baillon, M., Canellas, C., Franchisteguy, L., Morel, S., 2008. Analysis of near-surface atmospheric variables: Validation of the SAFRAN analysis over France. Journal of applied Meteorology and Climatology 47 (1), 92–107.

Ribatet, M., 2013a. Spatial extremes: Max-stable processes at work. Journal de la Société Francaise de Statistique 154 (2), 156–177.

Ribatet, M., 2013b. Spatialextremes : Modelling spatial extremes.

Schlather, M., 2002. Models for stationary max-stable random fields. Extremes 5 (1), 33–44.

Schlather, M., Tawn, J. A., 2003. A dependence measure for multivariate and spatial extreme values: Properties and inference. Biometrika 90 (1), 139–156.

Smith, R. L., 1990. Max-stable processes and spatial extreme, unpublished.

Spivey, M. Z., 2008. A generalized recurrence for bell numbers. Journal of Integer Sequences 11 (2), Article ID 08.2.5, 3 p., electronic only–Article ID 08.2.5, 3 p., electronic only.

Varin, C., Reid, N. M., Firth, D. D., 2011. An overview of composite likelihood methods. Statistica Sinica Vol.21 (No.1), 5–42.

Vrac, M., Drobinski, P., Merlo, A., Herrmann, M., Lavaysse, C., Li, L., Somot, S., 2012. Dynamical and statistical downscaling of the french mediterranean climate: uncertainty assessment. Natural Hazards and Earth System Science 12 (9), 2769–2784.

Vrac, M., Yiou, P., 2010. Weather regimes designed for local precipitation modeling: Application to the mediterranean basin. Journal of Geophysical Research: Atmospheres 115 (D12), 2156–2202.

Wang, Y., Stoev, S. A., 2011. Conditional sampling for spectrally discrete max-stable random fields. Advances in Applied Probability 43 (2), 461–483.

Xu, X., Reid, N., 2011. On the robustness of maximum composite likelihood estimate. Journal of Statistical Planning and Inference 141 (9), 3047–3054.

Zorita, E., Von Storch, H., 1999. The analog method as a simple statistical downscaling technique: comparison with more complicated methods. Journal of climate 12 (8), 2474–2489.

Chapitre 7

Algorithme de tirage des partitions

Dans cette partie, nous revenons sur l'étape 1 de la procédure de simulation conditionnelle (équation (18) du chapitre 6). Cette étape, consistant à regrouper les points en différentes classes pour former une partition tirée selon une loi cible, est très importante (voir Table 1) mais également complexe à mettre en œuvre. En effet, le nombre de partitions possibles, i.e. $|\mathcal{P}_k|$ avec k le nombre de points conditionnants, correspond au nombre de Bell. Ainsi, le cardinal de cet ensemble devient vite très grand et cela même pour un nombre limité de points conditionnants. Par exemple, pour $k = 10$, il y a déjà $115\ 975$ possibilités. Il est donc impossible de parcourir toutes les partitions possibles pour un nombre réaliste de points d'observations. De ce fait, même si on peut calculer le poids de chaque partition individuellement, il est très difficile d'obtenir la loi complète des partitions.

7.1 Méthodologie

7.1.1 Échantillonneur de Gibbs

Pour résoudre la difficulté computationnelle engendrée par le tirage des partitions selon la loi désirée, un algorithme MCMC (Monte-Carlo Markov Chain) est envisagé. Plus précisément, Dombry et al. (2013) proposent d'utiliser un échantillonneur de Gibbs, adapté à ce type de problèmes. Avec les mêmes notations que dans le chapitre 6, cet algorithme est basé sur la probabilité

$$\mathbb{P}[\theta \in . \mid \theta_{-j} = \tau_{-j}], \tag{7.1}$$

avec $\theta \in \mathcal{P}_k$ une partition aléatoire qui suit la loi cible $\pi_{\mathbf{x}}(\mathbf{z}, .)$, τ la partition courante de la chaine de Markov et τ_{-j} la partition τ restreinte à l'ensemble $\{x_1, \ldots, x_k\} \backslash \{x_j\}$.

Ainsi, pour tout $\tau^* \in \mathcal{P}_k$ tels que $\tau^*_{-j} = \tau_{-j}$, on peut écrire

$$\mathbb{P}[\theta = \tau^* \mid \theta_{-j} = \tau_{-j}] = \frac{\pi_{\mathbf{x}}(\mathbf{z}, \tau^*)}{\sum_{\tilde{\tau} \in \mathcal{P}_k} \pi_{\mathbf{x}}(\mathbf{z}, \tilde{\tau}) \mathbb{1}_{\{\tilde{\tau}_{-j} = \tau_{-j}\}}} \propto \frac{\prod_{j=1}^{|\tau^*|} \omega_{\tau^*,j}}{\prod_{j=1}^{|\tau|} \omega_{\tau,j}} \tag{7.2}$$

avec

$$\omega_{\tau,j} = \lambda_{\mathbf{x}_{\tau_j}}(\mathbf{z}_{\tau_j}) \int_{\{\mathbf{u} < \mathbf{z}_{\tau^c_j}\}} \lambda_{\mathbf{x}^c_{\tau_j} | \mathbf{x}_{\tau_j}, \mathbf{z}_{\tau_j}}(\mathbf{u}) \mathrm{d}\mathbf{u} \tag{7.3}$$

Ainsi les quantités utilisées dans l'équation (7.2) ont une expression explicite. Néanmoins, le calcul de l'intégrale dans les $\omega_{\tau,j}$ est possible par méthode de Monte-Carlo, mais les temps de calcul deviennent vite très conséquents. Une solution pour diminuer les ressources computationnelles nécessaires est d'utiliser des méthodes de quasi Monte-Carlo (Caflisch, 1998), qui consistent à estimer la fonction d'intérêt en certains points du domaine, de manière à remplir plus vite les « vides » qui peuvent apparaître avec les méthodes Monte-Carlo classiques. Ce genre de techniques permet de réduire le nombre d'itérations nécessaires à la bonne estimation de l'intégrale dans (7.3).

Ainsi, l'échantillonneur de Gibbs peut être lancé, en mettant à jour les poids ω (équation (7.2)) à chaque itération. Puis dans la même itération, la classe d'un élément de la partition τ va être ou non modifiée. A chaque itération, l'élément cible va être choisi de manière aléatoire. La distribution de la sélection des éléments suit donc une loi discrète uniforme, i.e. $p = (k^{-1}, \ldots, k^{-1})$. D'autres méthodes de sélection sont disponibles, mais celle-ci est couramment utilisée et est suffisamment performante. Avec cette méthode, le tirage d'une partition peut être obtenu suivant la loi souhaitée.

7.1.2 Modèle des feuilles mortes

Cette section présente une alternative à l'échantillonneur de Gibbs pour obtenir le tirage des partitions selon nos lois d'intérêt.

Définition

Le modèle des feuilles mortes (Matheron, 1968; Jeulin, 1979; Lantuéjoul, 2002) repose, comme son nom l'indique, sur l'analogie avec des feuilles qui tombent en se recouvrant partiellement et qui recouvrent progressivement tout l'espace de définition. L'algorithme repose sur quelques propriétés :

- Les feuilles sont indépendantes et sont des sous-ensembles compacts aléatoires de $\mathbb{R}^d$ notés F_i.

- Les feuilles sont réparties uniformément dans l'espace et tombent selon un processus ponctuel de Poisson d'intensité λ dans $\mathbb{R}^d \times [0, +\infty]$.
- A chaque feuille est associée une variable catégorielle appelée *couleur* et notée i. L'ensemble des couleurs est fini et différentes feuilles ont des couleurs indépendantes.
- Toutes les feuilles de même couleur suivent la même distribution (qui ne prend pas de valeur nulle). La taille de chacune de ces feuilles dépendra de cette distribution.

Il vient de ces propriétés que n'importe quel point de l'espace de travail sera, à l'infini, recouvert par une feuille. L'algorithme s'arrête dès qu'une feuille recouvre le point d'intérêt choisi préalablement et c'est la couleur de cette feuille spécifique que retourne l'algorithme.

Ainsi, en notant $C(y)$ la couleur de la première feuille qui recouvre le point d'intérêt $y \in \mathbb{R}^d$ et $a_i = p_i \lambda \mathbb{E}(F_i)$ avec p_i la probabilité de tirer une feuille de couleur i, on peut écrire :

$$\mathbb{P}(C(y) = i) = \frac{a_i}{\sum_i a_i}. \tag{7.4}$$

Il apparaît de manière naturelle que la probabilité que la couleur obtenue par l'algorithme soit i est proportionnelle à la probabilité p_i d'associer cette couleur à une feuille et à la taille $\mathbb{E}(F_i)$ de la feuille de couleur i.

Algorithme des feuilles mortes appliqué à la simulation de partitions

Le modèle des feuilles mortes peut être appliqué pour obtenir une partition suivant la distribution espérée. La distribution cible des partitions est :

$$\begin{aligned} \pi_{\mathbf{x}}(\mathbf{z}, \tau) = \mathbb{P}[\theta = \tau \mid Z(\mathbf{x}) = \mathbf{z}] &= \frac{1}{C(\mathbf{x}, \mathbf{z})} \prod_{j=1}^{|\tau|} \lambda_{\mathbf{x}_{\tau_j}}(\mathbf{z}_{\tau_j}) \int_{\{\mathbf{u}_j < \mathbf{z}_{\tau_j^c}\}} \lambda_{\mathbf{x}^c_{\tau_j} | \mathbf{x}_{\tau_j}, \mathbf{z}_{\tau_j}}(\mathbf{u}_j) \mathrm{d}\mathbf{u}_j, \\ &= \frac{\eta(\tau)}{\sum_{\tau \in \mathcal{P}_k} \eta(\tau)}. \end{aligned} \tag{7.5}$$

Il est alors possible d'adapter l'algorithme des feuilles mortes (Kendall and Thönnes, 1999) en définissant la couleur d'une feuille comme la partition τ qui lui est attachée et sa taille étant égale à $\eta(\tau)$. Le critère d'arrêt choisi dans cet algorithme correspond à la chute d'une feuille sur l'origine du repère (i.e. $y = 0$). Il peut être démontré que la partition τ relative à cette feuille spécifique suit la distribution définie par l'équation (7.5) (Lantuéjoul, 2002).

Démonstration. Dans ce cas de figure, on se place en dimension $d = 1$ et avec des feuilles ayant comme forme des intervalles de $\mathbb{R}^+$. Ainsi, la feuille relative à la partition τ sera définie comme $F_\tau = [x - \eta(\tau); x]$ de manière à avoir $\mathbb{E}(F_\tau) = \eta(\tau)$. De plus, en choisissant les probabilités de tirer une partition τ toutes identiques i.e. $p_\tau = 1/|\mathcal{P}_k|$, on peut réécrire l'équation (7.4) :

$$\mathbb{P}(C(0) = \tau) = \frac{a_\tau}{\sum_{\tau \in \mathcal{P}_k} a_\tau} = \frac{\lambda \eta(\tau)/|\mathcal{P}_k|}{\sum_{\tau \in \mathcal{P}_k} \lambda \eta(\tau)/|\mathcal{P}_k|} = \frac{\eta(\tau)}{\sum_{\tau \in \mathcal{P}_k} \eta(\tau)} = \pi_{\mathbf{x}}(\mathbf{z}, \tau) \tag{7.6}$$

Ainsi, prendre la partition correspondant à la première feuille recouvrant l'origine revient à tirer la partition dans notre loi cible. □

L'algorithme des feuilles mortes pour le tirage des partitions peut donc se formuler de la façon suivante :

Algorithm 1 Feuilles Mortes

Initialiser $w_1 = 1,\ \eta(\tau) = 0,\ w_2 = 0$
while $w_1 > \eta(\tau)$ **do**
  Générer un point $u = (w_1, w_2)$ avec
    $w_1 \sim \mathcal{U}_{[0,\eta_{max}]}$ (point de chute de la feuille)
    $w_2 = w_2 + e$ avec $e \sim \mathcal{E}(\lambda)$ (temps d'arrivée de la feuille)
  Attribuer à ce point une partition aléatoire τ (tirée de manière uniforme).
  Calculer $\eta(\tau)$
end while
Retourner τ

Il est important de remarquer que, à l'inverse des algorithmes MCMC tels que l'échantillonneur de Gibbs qui approximent la distribution cible, l'algorithme des feuilles mortes simule de manière exacte dans la distribution d'intérêt. Dans la pratique l'application de cet algorithme présente encore quelques difficultés. On remarque que pour une partition τ fixée, on peut approximer $\eta(\tau)$ par $\widehat{\eta}(\tau)$ en estimant les intégrales de l'équation (7.5) par méthodes de Monte-Carlo. Néanmoins, le tirage uniforme de partitions ou le fait de trouver un majorant $\eta_{\max}$ aux $\widehat{\eta}(\tau)$ restent des problèmes importants qui seront traités dans la prochaine partie.

En guise d'illustration, la Figure 7.1 montre une réalisation de cet algorithme en deux dimension dans le cas général (issue de Lantuéjoul (2002)). La Figure 7.2 montre une réalisation en une dimension, comme il sera utilisé dans notre cas de figure. Dans cet exemple, c'est la partition correspondant au label bleu qui sera conservée. Dans ces deux exemples, les couleurs n'ont aucune signification particulière.

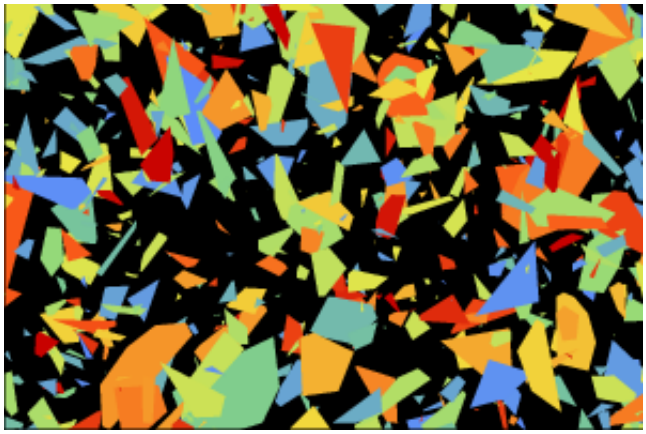


FIGURE 7.1 – Représentation de l'algorithme des feuilles mortes en deux dimensions

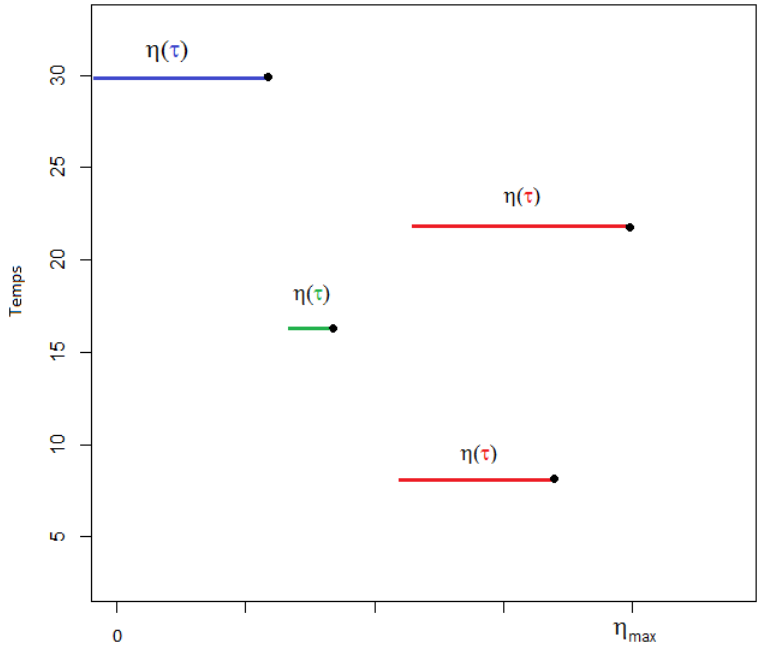


FIGURE 7.2 – Représentation de l'algorithme des feuilles mortes en une dimension

Dans la prochaine partie, on va chercher à résoudre les problèmes relatifs au tirage uniforme des partitions et au choix du majorant $\eta_{\max}$.

Tirage uniforme et majorant

Le premier enjeu de cette section est de parvenir à tirer de manière uniforme une partition, c'est-à-dire faire en sorte que chaque partition ait la même chance d'être choisie. Dans le

cas où le nombre n de points est très petit, il est possible de stocker toutes les partitions possibles et d'en choisir une au hasard. Néanmoins, on rappelle que le nombre de partitions que l'on peut former à partir de n points est égal au nombre de Bell $\mathcal{B}_n$. Pour $n > 5$, il semble déjà compliqué de stocker toutes les partitions possibles, il faut donc trouver une autre méthode.

De manière naturelle, on pourrait penser à attribuer de manière uniforme à chacun des n points une classe. Mais les partitions ainsi formées ne seront pas toutes équiprobables. En effet, on peut regarder par exemple la probabilité de choisir la partition constituée d'une seule classe :

$$\mathbb{P}((1,\dots,1)) = \frac{1}{n^{n-1}} \neq \frac{1}{\mathcal{B}_n}, \forall n > 3. \tag{7.7}$$

Cette technique ne fonctionne donc pas pour faire un tirage équiprobable de partition.

Une solution donnée par Pitman (1997) consiste à d'abord tirer selon une distribution spécifique M un majorant m du nombre de classes de la partition, puis de répartir de manière aléatoire les n points dans ces m classes. En effet, il montre qu'il faut avoir une distribution de M de la forme :

$$\mathbb{P}(M = m) = \frac{m^n p_m}{\mathcal{B}_n},\ \forall m \in [\![1,\dots,n]\!], \tag{7.8}$$

pour une distribution de probabilité p_m sur $\mathbb{N}$ qui a ses n premiers moments factoriels égaux à 1. Ainsi dans la pratique le majorant M sera obtenu par tirage dans la distribution

$$\mathbb{P}(M = m) \propto \frac{m^n}{m!} \sum_{i=0}^{n-m} \frac{(-1)^i}{i!},\ \forall m \in [\![1,\dots,n]\!]. \tag{7.9}$$

On notera que tous les termes de l'équation (7.9) sont facilement calculables puisqu'elle ne fait plus intervenir le nombre de Bell $\mathcal{B}_n$. Pitman (1997) démontre qu'en tirant M de cette façon, puis en attribuant de manière aléatoire à chacun des points une classe parmi $[\![1,\dots,M]\!]$, la partition obtenue est tirée selon une loi uniforme.

En ce qui concerne le majorant des $\widehat{\eta}(\tau)$, on se contentera pour l'instant d'une approximation qui peut être effectuée de manière empirique pour obtenir une valeur de $\eta_{\max}$. Ainsi, des poids choisis aléatoirement sont calculés et conduisent à une majoration de l'ensemble des valeurs de $\eta(\tau)$. Dans la suite, il faudra réfléchir à une méthode plus rigoureuse pour obtenir une valeur de $\eta_{\max}$, en passant par exemple par une majoration des densités $\lambda_{\mathbf{x}_{\tau_j}}$ et $\lambda_{\mathbf{x}^c_{\tau_j}|\mathbf{x}_{\tau_j},\mathbf{z}_{\tau_j}}$.

7.2 Application à des données simulées

7.2.1 Méthodologie

La méthodologie de simulations conditionnelles a été appliquée à un jeu de données simulées. Un même processus de Schlather a été simulé en plusieurs points d'une grille (package R `SpatialExtreme` de Ribatet (2013b)). Dans ce processus, un modèle de variogramme est utilisé pour construire la relation de covariance entre les différents points de la grille. Certains de ces points seront considérés comme étant conditionnants, et les autres serviront à valider les simulations conditionnelles. La Figure 7.3 montre la configuration des points retenus.

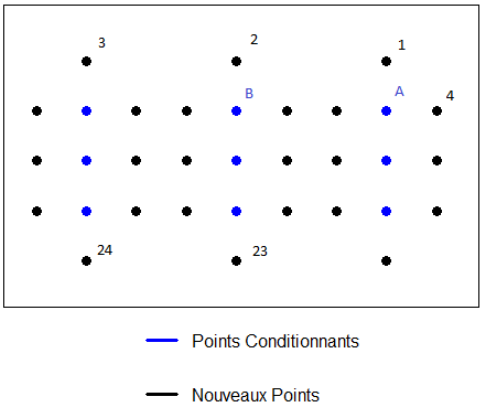


FIGURE 7.3 – Configuration des points conditionnants et des nouveaux points.

L'objectif est de simuler conditionnellement avec nos méthodes des valeurs pour les nouveaux points. Par la suite, il faudra comparer ces nouvelles valeurs à celles construites par les simulations du processus de Schlather.

Par souci de cohérence, le variogramme utilisé ici est le même que celui qui a servi à construire ces données. Une fois le partitionnement obtenu, il sera alors possible de construire $\tilde{Z} = \max\{Z^+, Z^-\}$ en chacun des nouveaux points.

7.2.2 Validation du tirage des partitions

Dans cette partie, on cherche à comparer, sur un exemple simple, les probabilités théoriques de chaque partition avec celles obtenues avec l'échantillonneur de Gibbs et par l'algorithme des feuilles mortes.

Les probabilités théoriques sont données par la formule

$$\pi_{\mathbf{x}}(\mathbf{z},\tau) = \mathbb{P}[\theta = \tau \mid Z(\mathbf{x}) = \mathbf{z}] = \frac{1}{C(\mathbf{x},\mathbf{z})} \prod_{j=1}^{|\tau|} \lambda_{\mathbf{x}_{\tau_j}}(\mathbf{z}_{\tau_j}) \int_{\{\mathbf{u}_j < \mathbf{z}_{\tau_j^c}\}} \lambda_{\mathbf{x}^c_{\tau_j}|\mathbf{x}_{\tau_j},\mathbf{z}_{\tau_j}}(\mathbf{u}_j)\mathrm{d}\mathbf{u}_j,$$

où la constante de normalisation $C(\mathbf{x},\mathbf{z})$ est donnée par

$$C(\mathbf{x},\mathbf{z}) = \sum_{\tilde{\tau}\in\mathcal{P}_k} \prod_{j=1}^{|\tilde{\tau}|} \lambda_{\mathbf{x}_{\tilde{\tau}_j}}(\mathbf{z}_{\tilde{\tau}_j}) \int_{\{\mathbf{u}_j < \mathbf{z}_{\tilde{\tau}_j^c}\}} \lambda_{\mathbf{x}^c_{\tilde{\tau}_j}|\mathbf{x}_{\tilde{\tau}_j},\mathbf{z}_{\tilde{\tau}_j}}(\mathbf{u}_j)\mathrm{d}\mathbf{u}_j,$$

avec τ une partition de $\{x_1,\dots,x_n\}$, $|\tau|$ le cardinal de la partition, et les notations $v_{\tau_j} = (v_i \mid i \in I_j)$ et $v_{\tau_j^c} = (v_i \mid i \notin I_j)$, avec $I_j = \{i \mid x_i \in \tau_j\}$ pour tout k-uplet v de $\mathcal{X}$ ou $\mathbb{R}^k$.

Par la suite, l'échantillonneur de Gibbs a été effectué avec 2000 itérations et l'historique de la chaîne a été observé pour les 1600 dernières itérations (400 itérations pour le burn-in).

Le jeu de données choisi est constitué de 4 points conditionnants (les 4 premiers points de la Figure 7.3). Ce nombre de points est volontairement très faible pour permettre le calcul des probabilités théoriques de chaque partition (ici il y en a 15 au total).

En complément l'algorithme des feuilles mortes a été effectué $1\,000$ fois, et le tableau 7.1 résume la fréquence des partitions obtenues.

Partition	**Probabilité Théorique**	**Prob. Empirique Gibbs**	**Prob. Empirique FM**
(1, 1, 1, 1)	34.2%	30.3%	32.2%
(1, 2, 2, 1)	22.3%	20.8%	25.0%
(1, 2, 2, 3)	25.2%	26.3%	27.7%
(1, 2, 2, 2)	4.1%	5.1%	3.2%
(1, 2, 3, 4)	2.0%	2.6%	2.0%
Autres	12.2%	14.9%	9.9%

Tableau 7.1 – Probabilités théoriques et empiriques de chaque partition.

Il apparaît que les résultats obtenus par nos deux algorithmes sont proches de ceux attendus, c'est-à-dire des probabilités théoriques.

Il faut toutefois noter, que les fréquences de l'échantillonneur de Gibbs sont celles observées sur l'historique d'une seule chaîne. Les temps de calcul étant trop conséquents,

il est impossible de lancer cet algorithme un grand nombre de fois. En revanche, l'algorithme des feuilles mortes a lui été appliqué $1\ 000$ fois de manière indépendante. Dans ce cas, lorsque le nombre total de partitions est peu élevé, la supériorité en temps de calcul de l'algorithme de feuilles mortes est très nette. Il reste toutefois à voir si ce résultat demeure lorsque le nombre de points conditionnants est plus élevé. En effet, la probabilité de tomber sur une partition d'un poids significatif en tirant des partitions de manière uniforme, va décroître très rapidement avec le nombre de points conditionnants. Cet algorithme risque donc d'être trop lent pour un nombre de points plus élevé. De plus, la recherche d'un majorant $\eta_{\max}$ risque aussi de devenir plus compliquée avec un grand nombre de points conditionnants.

Des simulations ont été effectuées pour mesurer le temps de calcul nécessaire pour une réalisation de l'algorithme des feuilles mortes en fonction du nombre de points conditionnants. Cette procédure a été répétée 20 fois. La Figure 7.4 décrit ces résultats.

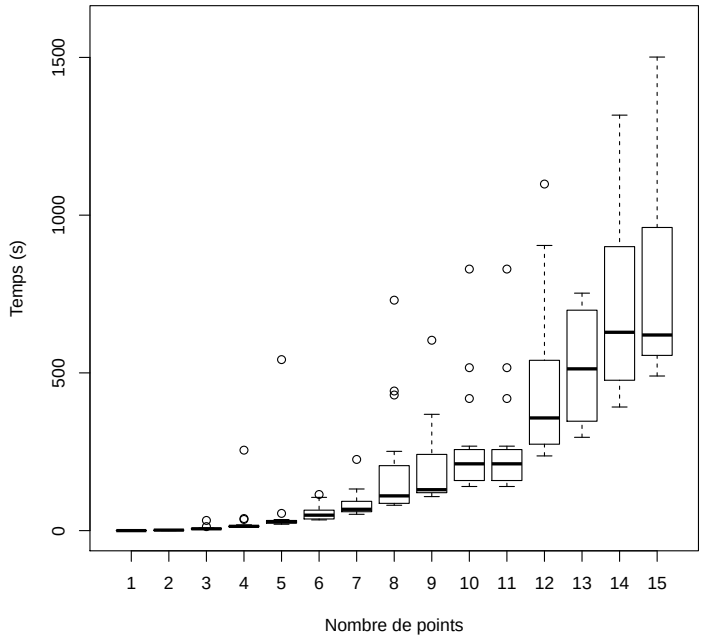


FIGURE 7.4 – Temps de calcul nécessaire, en fonction du nombre de points conditionnants, pour une réalisation de l'algorithme des feuilles mortes.

Il apparaît que le temps de calcul nécessaire pour obtenir une réalisation de l'algorithme des feuilles mortes augmente rapidement avec le nombre de points conditionnants. De plus, la variabilité du temps nécessaire augmente également très rapidement. Ainsi, pour plus d'une quinzaine de points, utiliser cet algorithme semble irréaliste. De plus,

il faut obtenir une nouvelle partition pour chaque nouvelle simulation, ce qui implique d'exécuter cet algorithme un grand nombre de fois (à l'inverse de l'échantillonneur de Gibbs pour lequel on utilise l'historique de la chaîne de Markov construite).

7.3 Conclusion

Dans ce chapitre, l'algorithme des feuilles mortes a été proposé comme alternative à l'échantillonneur de Gibbs pour tirer des partitions dans la loi cible. Cet algorithme très efficace et performant lorsque le nombre de points conditionnants est inférieur à environ 15, devient par la suite inapplicable. En effet, le nombre de partitions possibles augmentant, la probabilité de tirer aléatoirement une partition avec un poids non négligeable devient rapidement très faible. Ce problème vient du fait que chaque partition candidate est tirée aléatoirement et indépendamment de la précédente. En parallèle, les problèmes pour obtenir empiriquement un majorant aux poids des partitions s'amplifient avec le nombre de points conditionnants.

Ces résultats permettent néanmoins de confirmer que ces deux approches sont adaptées pour obtenir des partitions suivant les lois désirées. L'échantillonneur de Gibbs apparaît comme étant la meilleure solution pour un nombre de points élevé, alors que dans le cas contraire, l'algorithme des feuilles mortes est rapide et plus performant.

Chapitre 8

Conclusions de la partie II

Dans cette première partie, nous avons proposé une méthodologie permettant d'obtenir des distributions de maxima en des points vides d'observations. Grâce aux simulations conditionnelles de processus max-stables, les distributions obtenues sont à la fois spatialement cohérentes entre elles et respectent les valeurs observées.

Néanmoins, la simulation conditionnelle de tels processus nécessite l'utilisation d'algorithmes complexes et gourmands en temps de calcul. Des efforts ont été faits pour comprendre et adapter un algorithme déjà existant proposé par Dombry et al. (2013). En effet, nous avons adapté cet algorithme à une famille de processus max-stable particulière : le processus t extrêmal. Ce processus peut être vu comme une généralisation du processus de Schlather avec l'ajout d'un paramètre supplémentaire. Ce paramètre a un rôle déterminant dans la dépendance des extrêmes du processus. Ainsi, le fait de pouvoir estimer ce paramètre sur les données offre au processus une flexibilité très intéressante. L'utilisation des simulations conditionnelles de ce processus sur des données simulées montre de bons résultats.

De plus, de nombreux tests ont été effectués (chapitre 6) pour comprendre et analyser l'importance de chacune des trois étapes de cet algorithme. La première étape de l'algorithme, la plus coûteuse, consiste à tirer une partition des points conditionnants dans la bonne loi. Cela peut vite se révéler très compliqué lorsque le nombre de points conditionnants augmente. Ainsi, nous avons utilisé un échantillonneur de Gibbs, comme préconisé par Dombry et al. (2013). Par la suite, nous avons également proposé une approche basée sur l'algorithme des feuilles mortes permettant également de tirer une partition dans la loi cible. Si cette approche se révèle très efficace et rapide pour un faible nombre de points conditionnants, elle devient inapplicable dans le cas contraire. Cette méthode peut néanmoins se révéler utile et offrir une alternative intéressante dans certains cas de figure.

Un autre ajout de ce travail est la modification de la deuxième étape de l'algorithme qui permet désormais de considérer un grand nombre de points conditionnants. Précédemment une étape d'acceptation/rejet était utilisée. Il apparaît que le taux d'acceptation devient extrêmement faible à mesure que le nombre de points conditionnants augmente et rend ainsi l'algorithme inutilisable. Nous avons proposé de rajouter un échantillonneur de Gibbs, qui est rapide et qui permet d'obtenir des résultats pour un nombre de points élevé.

Par la suite, cette méthode a été testée sur des données de précipitations dans le sud-est de la France. Les résultats sont encourageants et montrent que cette approche est performante, pour différents critères, comparée à plusieurs approches usuellement utilisées. Il apparaît toutefois que cette méthode rencontre des difficultés pour modéliser les précipitations extrêmes dans certaines zones. En effet, on constate que dans les régions où l'altitude est élevée, la qualité des simulations est moins bonne. Ainsi, l'ajout de cette variable dans la modélisation des marginales, donc dans la non stationnarité spatiale, peut s'avérer être une solution.

Néanmoins, à ce stade de nombreuses améliorations doivent être effectuées. L'un des enjeux majeur de ce travail est de parvenir à fournir ce genre de distributions, mais pour des années futures. Dans ce cas-là, il n'y a plus d'observations disponibles et donc ce genre de méthodes ne peut plus être utilisé tel quel. La seule information disponible qui provient de modèles climatiques n'est pas utilisable directement du fait notamment de leur faible résolution spatiale. Il faudra donc adapter les techniques présentées dans cette partie à ce nouveau type de données pour pouvoir obtenir des distributions de maxima de précipitations en n'importe quel point de la région, et ce même pour des années futures. De plus, en utilisant les sorties de modèles climatiques correspondants aux différents scénarios de changements climatiques, il serait également possible grâce à ces techniques de voir les conséquences locales de ces différents scénarios.

Troisième partie

Downscaling spatial des extrêmes

Chapitre 9

Introduction

Dans un contexte de projections sous contraintes de changement climatique, les sorties de modèles climatiques (GCMs) sont des outils majeurs pour générer des valeurs cohérentes de variables climatiques (IPCC, 2013). Cependant, leur faible résolution ($\approx$ 200km) rend difficile leur utilisation directe dans des modèles d'impact. Une étape de descente d'échelle (downscaling) est donc nécessaire pour obtenir des valeurs plus locales. Il existe deux approches distinctes de descente d'échelle, le downscaling dynamique et le downscaling statistique.

Les modèles régionaux (RCMs) de downscaling dynamique des sorties de GCM peuvent être considérés comme une version régionalisée des GCMs se basant également sur la modélisation physique des dynamiques atmosphériques. Ces RCMs résolvent directement une partie des phénomènes climatiques impliqués, à une échelle plus fine que celle des GCMs, tout en fournissant des sorties cohérentes avec ces derniers. La résolution spatiale est donc meilleure, de 50km à 5km selon les modèles (e.g. Laprise, 2008; Rummukainen, 2010), mais peut rester toutefois insuffisante en fonction des objectifs poursuivis. De plus, les coûts de calcul requis pour obtenir ces RCMs sont très conséquents et limitent de ce fait leur cadre d'utilisation à certaines régions ou certaines périodes temporelles (Baron et al., 2005). De cette complexité algorithmique découle également la grande difficulté de faire de la modélisation d'ensemble de RCMs, qui consiste à réaliser un grand nombre de simulations et à évaluer la variabilité des sorties obtenues. Ces différents écueils rendent les sorties de RCM difficilement utilisables directement pour décrire le caractère local d'un phénomène.

Ainsi, afin d'obtenir des valeurs à plus haute résolution spatiale, le downscaling statistique des sorties de modèles climatiques peut être utilisé. Ce terme englobe une grande variété de méthodes ayant pour principe d'établir une relation entre des variables à grande échelle (issues de GCM ou RCM) et des variables plus localisées, pouvant être des observations. On peut diviser le downscaling statistique en plusieurs grandes familles

présentées ci-après. Pour plus de détails sur ces différentes familles de downscaling statistique voir par exemple Maraun et al. (2010) et Vrac (2012).

• Le downscaling basé sur les **fonctions de transfert** cherche à établir un lien statistique déterministe entre les variables à haute et basse résolution. Ce lien peut être modélisé par différentes méthodes statistiques comme par exemple à l'aide de régressions paramétriques linéaires (e.g. Wigley et al., 1990; Huth, 2002; Busuioc et al., 2008; Nicholas and Battisti, 2012) ou non linéaires (e.g. Sailor and Xiangshang, 1999), de régressions non paramétriques (e.g. Vrac et al., 2007a; Levavasseur et al., 2011) ou encore par réseau de neurones (e.g. Snell et al., 2000; Ghosh and Mujumdar, 2008; Fernández-Ferrero et al., 2009).

• Les méthodes appelées **générateurs stochastiques de temps** qui simulent la variable locale conditionnellement à la variable à grande échelle en estimant sa densité de probabilité conditionnelle. Ainsi, différents modèles de densité peuvent être choisis, et ce sont les paramètres intervenant dans la densité qui dépendent de l'information issue de sorties de modèles climatiques (e.g. Wilks, 1998; Bellone et al., 2000; Hingray and Mezghani, 2007; Carreau and Vrac, 2011). Les variations dans ces variables à grande échelle peuvent donc être transférées à la densité de probabilité de la variable à l'échelle locale et faire évoluer les simulations dans le temps (e.g. Vrac and Naveau, 2007).

• Les méthodes basées sur les **régimes de temps** ou **types de temps** cherchent à classifier de manière non supervisée les conditions atmosphériques de la période étudiée (jour, mois, année etc...) en fonction de variables climatiques à grande échelle. Après cette étape de pré-traitement, les techniques de downscaling de type fonction de transfert (e.g. Huth et al., 2008) ou générateur de temps (e.g. Schnur and Lettenmaier, 1998; Vrac and Naveau, 2007; Vrac et al., 2007b) peuvent être appliquées, en les conditionnant par le régime de temps dans lequel se trouve la période ciblée. Ces méthodes reposent sur l'hypothèse que dans le futur, les conditions météorologiques locales restent associées aux mêmes régimes de temps obtenus eux à partir des variables à grande échelle (e.g. Huth, 2001; Boé et al., 2006).

• Le downscaling reposant sur la **correction de biais** ou de **Model Output Statistics** qui consiste à corriger la fonction de répartition de la variable à grande échelle pour la faire « ressembler » à celle de l'échelle locale (e.g. Kalnay, 2003; Wilks, 2006; Déqué, 2007; Piani et al., 2010; Vrac et al., 2012; Gudmundsson et al., 2012). Une fois la modélisation effectuée, la correction de biais est réalisée selon une approche de type quantile-quantile (Déqué, 2007).

La plupart de ces techniques de downscaling statistique s'effectuent de manière

univariée. Dans notre cadre de travail, cela peut s'avérer insuffisant puisque l'on cherche à générer des valeurs spatialement cohérentes entre elles. Cet aspect multivarié peut être obtenu de différentes manières comme par exemple avec des méthodes de classification ou par downscaling multi-sites (e.g. Wilks, 2012). De plus, ces techniques permettent de générer des valeurs à haute résolution mais uniquement aux points où le modèle peut être calibré, c'est-à-dire aux emplacements où des observations sont disponibles (dans le passé).

Malgré ces écueils, cette étape de downscaling statistique peut se révéler utile dans la construction d'une méthode de projection future basée sur la méthodologie de simulation conditionnelle. En effet, les techniques utilisées dans la partie II nécessitent d'avoir des observations pour appliquer l'algorithme de simulation conditionnelle. Ces techniques ne sont plus applicables de la même manière puisque l'information conditionnante locale disparaît lorsque l'on cherche à simuler des événements pour des années futures.

Ainsi dans cette partie III, nous enrichissons progressivement les techniques de simulation conditionnelle présentées précédemment. Le chapitre 10 décrit une première tentative d'incorporation d'information à grande échelle dans les modèles de simulations conditionnelles qui est effectuée pour modéliser la non stationnarité temporelle. Dans la suite de ce même chapitre, nous proposons d'utiliser des techniques de simulation non conditionnelle (i.e. sans observations locales) qui incluent l'information à grande échelle pour se placer dans un contexte de downscaling. Enfin, dans le chapitre 11, nous cherchons à conditionner les modèles de simulation conditionnelle non plus par des observations mais par des sorties de modèles climatiques, en ayant recours à des techniques de downscaling statistique.

Pour conclure cette introduction, il est important de comprendre que nous cherchons à développer des outils applicables dans un contexte de changement climatique, mais qu'à aucun moment dans ce travail nous n'appliquons ces techniques à des projections futures. En effet, dans une optique de d'évaluation de la qualité des méthodes présentées, nous effectuons les projections sur le passé proche afin de pouvoir les comparer avec les vraies valeurs observées.

Chapitre 10

Comment ajouter de l'information à grande échelle ?

Le but de ce chapitre est de parvenir à ajouter de l'information issue de variables climatiques dans les méthodes de modélisation des extrêmes. Pour y parvenir, l'idée est d'utiliser l'algorithme de simulation conditionnelle présenté précédemment et d'inclure l'information atmosphérique dans la transformation des marginales. On a vu chapitre 6 que cette transformation des marginales (équation (31) du chapitre 6) était nécessaire pour pouvoir modéliser les données par le processus max-stable choisi. Ainsi, la simulation du processus max-stable s'effectue dans le monde des données transformées avec des marginales α-Fréchet. De plus, cette transformation des marginales permet également d'incorporer de la non stationnarité spatiale dans notre modélisation. Précédemment, cette non stationnarité ne dépendait que des coordonnées géographiques, mais désormais le but est d'incorporer des variables climatiques pour enrichir la modélisation et ajouter une composante de non stationnarité temporelle.

Deux approches distinctes de simulation de ces processus non stationnaires sont développées dans ce chapitre. Dans un premier temps, on s'intéresse à savoir si l'incorporation de la non stationnarité temporelle améliore la qualité des simulations conditionnelles. Puis, une première démarche de downscaling est présentée, dans laquelle nous évaluons la qualité des simulations non stationnaires et non conditionnelles en nous affranchissant des points conditionnants. Il est par ailleurs important de noter que dans tout ce chapitre, nous parlons de simulation non conditionnelle lorsqu'elle ne repose pas sur des observations ponctuelles, même si elle est conditionnée par de l'information à grande échelle. En effet dans ce cas, les techniques de simulation utilisées sont non conditionnelles mais de l'information à grande échelle y est incorporée. Même sans être une « pure » méthode de downscaling, cette approche peut être considérée comme une technique de la famille des générateurs stochastiques de temps.

Parallèlement, suite aux conclusions du chapitre 6, l'intérêt de considérer l'altitude comme covariable dans la transformation des marginales est également étudié dans cette

partie.

10.1 Présentation des données

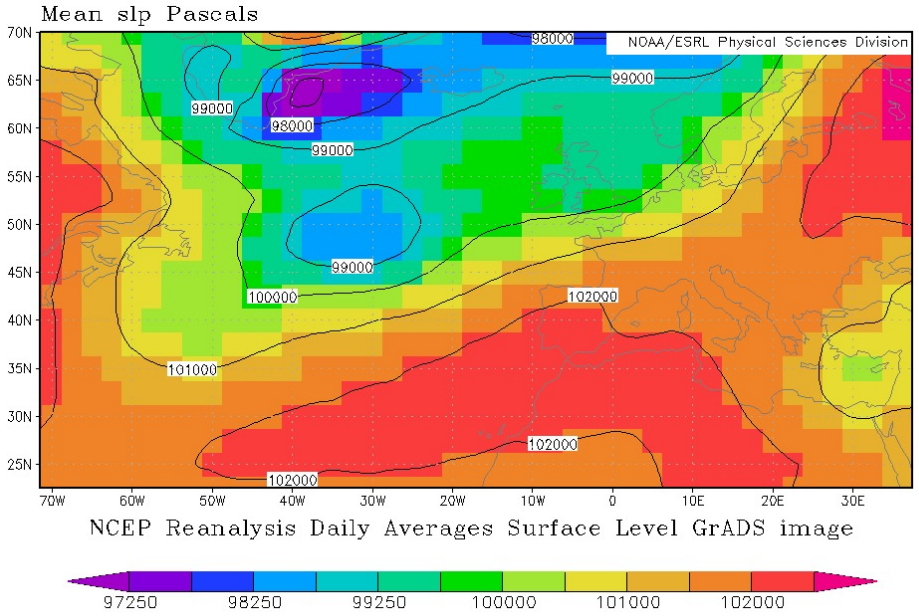
Les variables climatiques utilisées sont issues des réanalyses NCEP/NCAR[1] (National Centers for Environmental Predictions/National Center for Atmospheric Research ; Kalnay et al., 1996), données assez communément utilisées (e.g. Zolina et al., 2004; Mannshardt-Shamseldin et al., 2010). Plus particulièrement, nous nous intéressons à la variable « Sea Level Pressure » (SLP) qui mesure la pression atmosphérique ramenée au niveau de la mer. Cette variable apparaît comme pouvant avoir des conséquences directes sur plusieurs phénomènes climatiques et particulièrement les précipitations extrêmes (e.g. Goodess and Jones, 2002). Les réanalyses sont gridées et nous considérons deux régions géographiques différentes pour voir si celles-ci influent sur la qualité des simulations. Les deux régions considérées sont donc celles présentées Figure 10.1a et une zone géographique plus petite, présentée Figure 10.1b. Les données SLP de ces zones sont souvent utilisées pour expliquer les précipitations dans l'Europe de l'ouest (e.g. Yiou and Nogaj, 2004; Vrac and Yiou, 2010).

Pour expliquer les précipitations issues de données SAFRAN (Figure (3) et (4) du chapitre 6), nous utilisons les réanalyses journalières des jours d'automne sur la période $1960 - 2007$. Comme dans le chapitre 6, la période $1960 - 1999$ sera la période d'apprentissage et $2000 - 2007$ la période test.

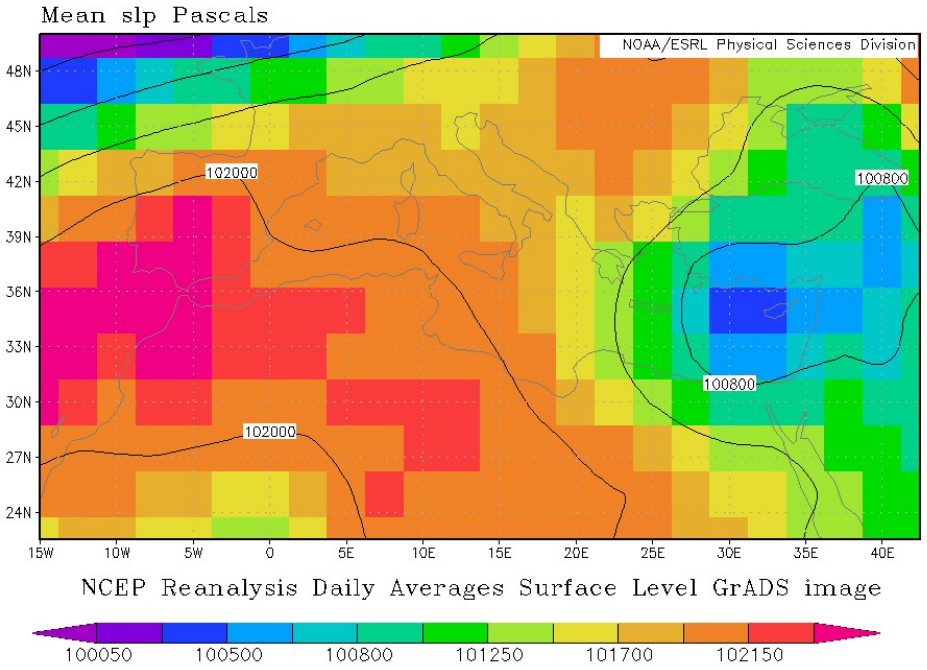
10.2 Modélisation

L'objectif principal de ce chapitre est de proposer un modèle qui combine la méthodologie de simulation conditionnelle de processus max-stable exposée précédemment et l'information à grande échelle comme les données SLP. Dans cette optique, deux directions principales sont explorées. Les deux consistent à ajouter l'information SLP lors de l'étape de transformation des marginales (équation (31) du chapitre 6). Néanmoins dans nos zones d'étude, les données SLP fournissent respectivement 940 et 288 cellules, le tout sur une période de 91 jours par année. Le nombre de variables disponibles est ainsi trop élevé pour modéliser les extrêmes automnaux. De plus, expliquer ces précipitations par les 91 jours des valeurs de SLP ne semble pas être un choix très pertinent. Dans la suite, nous allons donc chercher à résumer cette information et trouver des statistiques plus parcimonieuses pour ces variables.

[1] http ://www.ncep.noaa.gov/



(a)



(b)

FIGURE 10.1 – Carte des données SLP pour le 1^{er} Janvier 1960 sur la zone large (a) et la zone restreinte (b).

10.2.1 Résumer l'information disponible

Une analyse en composantes principales (e.g. Mardia et al., 1979) a été menée sur les points de grilles SLP, dans le but de mettre en exergue un comportement général et de résumer l'information disponible. Ainsi en fonction de la zone géographique considérée, chaque composante principale est une combinaison linéaire des 940 ou 288 valeurs SLP journalières. Pour chaque année, chaque composante principale prend 91 valeurs sur la période automnale. Pour les deux premières composantes principales nous nous intéressons à deux statistiques particulières : la moyenne et l'écart-type des 91 valeurs journalières prises par ces composantes principales. Ces indicateurs fournissent des informations annuelles sur le niveau de pression moyen et sa variabilité qui peuvent être directement reliés aux précipitations extrêmes dans notre zone d'intérêt.

De manière générale, il est important de remarquer que la qualité des simulations est dépendante de la qualité de l'estimation des paramètres. Cette estimation par maximisation de la vraisemblance composite (par optimisation numérique) est de moins en moins efficace à mesure que le nombre de paramètres augmente. Ainsi, augmenter le nombre de paramètres à estimer augmente les difficultés à le faire et ne conduira pas nécessairement à une amélioration de la qualité des simulations. Il nous faut donc trouver un modèle à la fois explicatif et parcimonieux.

10.2.2 Modèles de transformation des marginales

Dans cette section, deux méthodes de transformation des marginales vont être présentées. Les deux reposent sur une modélisation des paramètres de la GEV.

Le modèle de transformation $T1$ propose la transformation des marginales suivante :

$$\begin{aligned}
\mu(s,t) &= \mu_0 + \mu_1 * LAT(s) + \mu_2 * LON(s) + \mu_3 * MEAN_{CP1}(t) + \mu_4 * SD_{CP1}(t) + \\
&\quad \mu_5 * MEAN_{CP2}(t) + \mu_6 * SD_{CP2}(t) \\
\sigma(s,t) &= \sigma_0 + \sigma_1 * LAT(s) + \sigma_2 * LON(s) + \sigma_3 * MEAN_{CP1}(t) + \sigma_4 * SD_{CP1}(t) + \\
&\quad \sigma_5 * MEAN_{CP2}(t) + \sigma_6 * SD_{CP2}(t) \\
\xi &= \xi_0
\end{aligned}$$

avec s décrivant les sites de la région et t les années. LAT désigne la latitude, LON la longitude, et $MEAN_{CPi}$ et SD_{CPi} représentent respectivement la moyenne des valeurs de la i_{me} composante principale et leur écart type.

Grâce au grand nombre de paramètres, ce modèle permet une grande flexibilité. Néanmoins, l'estimation de ces paramètres (les μ_i, le σ_i et ξ_0), dans le cadre de la simulation conditionnelle de processus max-stable se révèle très compliquée. En effet, la maximisation de la vraisemblance composite rencontre beaucoup de difficultés pour aboutir à des estimations plausibles des paramètres. C'est également pour cette raison que l'on

a choisi de ne retenir que deux composantes principales, qui n'expliquent pourtant que respectivement 36% et 59% de la variance globale (en fonction de la zone choisie).

Le modèle de transformation $T2$ propose une autre manière d'incorporer les variables SLP dans la transformation des marginales qui consiste à considérer que la période d'apprentissage est composée de différent régimes de temps. Selon le centre national français de recherches météorologiques, *la circulation atmosphérique de grande échelle aux latitudes tempérées est caractérisée par des fluctuations des courant-jets entre différents états quasi-stationnaires qu'on appelle les régimes de temps.* En d'autres mots, l'idée est de classer les années d'apprentissage en différentes classes homogènes en se basant sur de l'information apportée par des variables atmosphériques. Pour nous différencier des régimes de temps qui sont généralement associés à des jours, nous parlerons de régimes automnaux puisque nous cherchons à attribuer un régime à une année. En regroupant les années d'apprentissage en différents régimes, on peut adapter la transformation des marginales en fonction de cette classification. La variable utilisée pour la classification est à nouveau SLP et la même analyse en composantes principales est utilisée. Les seules différences sont que l'on ne garde plus uniquement les 2 principales composantes mais autant que nécessaire pour expliquer 90% de la variance (dans ce cas, on en conservera 16). La moyenne et l'écart-type sont à nouveau utilisés, mais cette fois comme variables de classification.

Le modèle $T2$ propose la transformation des marginales suivante :

$$\begin{aligned}
\mu(s,t) &= \sum_{i=1}^{|R|} \left[\mu_{0R_i} + \mu_{1R_i} * LAT(s) + \mu_{2R_i} * LON(s)\right] \mathbb{1}_{\{Y_t=R_i\}} \\
\sigma(s,t) &= \sum_{i=1}^{|R|} \left[\sigma_{0R_i} + \sigma_{1R_i} * LAT(s) + \sigma_{2R_i} * LON(s)\right] \mathbb{1}_{\{Y_t=R_i\}} \\
\xi(t) &= \sum_{i=1}^{|R|} \left[\xi_{0R_i}\right] \mathbb{1}_{\{Y_t=R_i\}}
\end{aligned}$$

avec $|R|$ le nombre total de régimes automnaux et Y_t le régime de l'année t.

L'estimation des jeux de paramètres peut être effectuée séparément pour chaque régime, ce qui diminue le nombre de paramètres à estimer à chaque fois et facilite la maximisation de la vraisemblance composite. Néanmoins, le nombre d'années disponibles pouvant être faible pour un régime donné, le jeu de paramètres correspondant est estimé sur très peu de données.

Il reste ensuite à trouver une règle d'attribution pour savoir à quel régime une nouvelle année correspond. On a choisi d'utiliser un algorithme des k plus proches voisins. Cette méthode non paramétrique de classification non supervisée consiste à

attribuer une classe à un individu (ici une année) en regardant ses k plus proches voisins au sens d'une distance choisie. Dans notre cas, la moyenne et l'écart-type de chaque composante principale annuelle seront les variables qui serviront à mesurer les distances entre les années. Dans la pratique, pour éviter de choisir arbitrairement un nombre k, nous effectuons l'algorithme des plus proches voisins pour différentes valeurs de k. Chacun de ces algorithmes nous désigne un régime comme étant le plus proche de la nouvelle année considérée. Par la suite, on attribue le régime qui est choisi le plus souvent comme étant le plus proche.

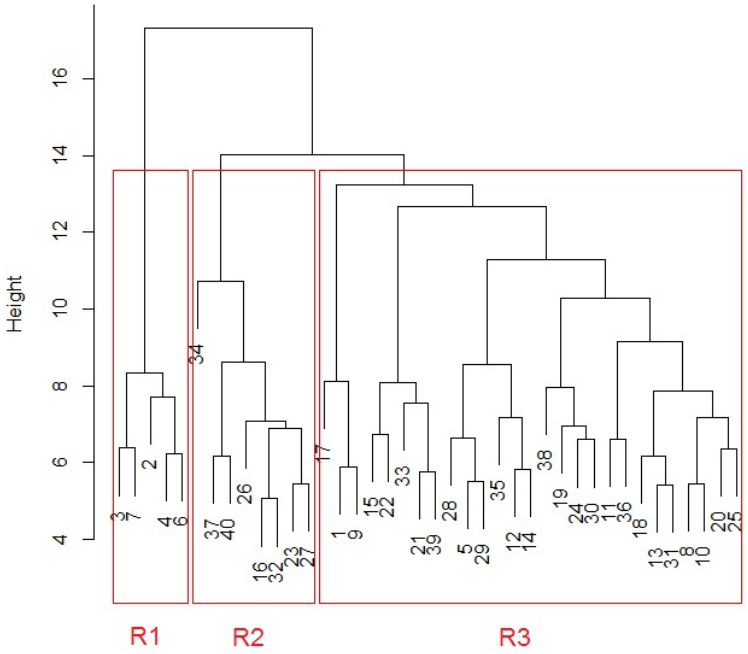
10.2.3 Régimes automnaux

Pour le modèle $T2$, une classification ascendante hiérarchique (e.g. Gordon, 1999) est utilisée pour identifier différents régimes automnaux. Les résultats sont présentés Figure 10.2a et 10.2b. A noter que les années sont numérotées par ordre croissant avec l'année 1 correspondant à 1960.

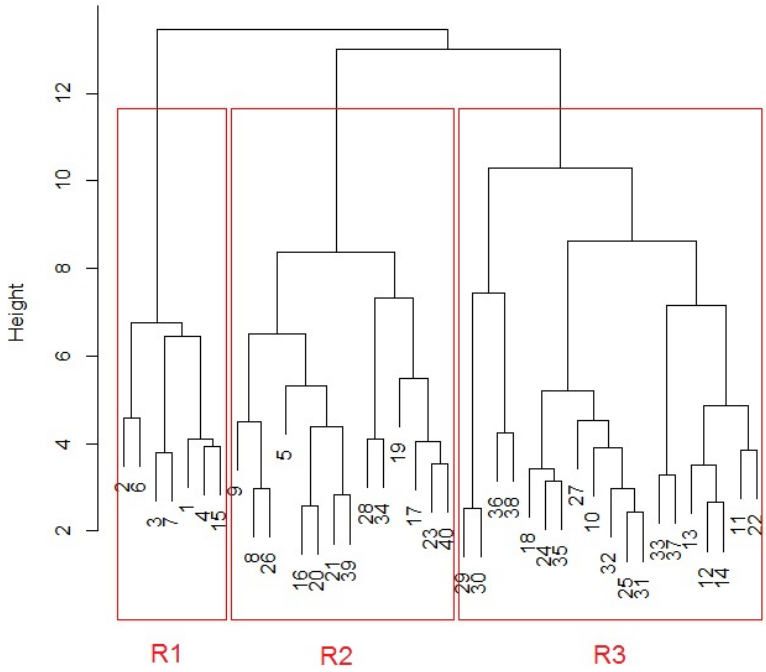
Au vu des résultats, on a choisi de conserver 3 classes dans un souci de ne pas avoir des classes avec trop peu d'individus pour ne pas entraîner des difficultés dans l'étape d'estimation des paramètres. De plus, on remarque que dans les deux cas, la première classe est constituée en grande partie des années les plus anciennes, montrant de grandes différences avec les autres et pour lesquelles la qualité des données peut être discutable. Pour les deux zones géographiques, aucune des années sur lesquelles les simulations vont être évaluées $(2000 - 2007)$ n'appartient au régime $R1$. C'est pour cette raison que nous avons choisi de considérer deux autres classes en plus de celle-ci.

De plus, la proportion d'années qui restent dans le même régime d'une classification à l'autre est de 65%. Il y a donc certaines différences entre les régimes obtenus à partir des deux zones géographiques.

Par la suite, les Figures 10.3 et 10.4 montrent les cartes des moyennes et des écart-types de la variable SLP pour chacun des 3 régimes. Sur ces cartes, chaque point est construit comme la moyenne de la variable SLP parmi tous les jours appartenant au régime en question. Pour les deux zones géographiques, certaines différences apparaissent au niveau des moyennes entre les régimes, spécialement entre le régime 1 et les deux autres, alors que les cartes des écart-types semblent similaires.



(a)



(b)

FIGURE 10.2 – Dendogram sur les 40 années d'apprentissage en considérant l'information issue la zone large (a) et celle de la zone restreinte (b).

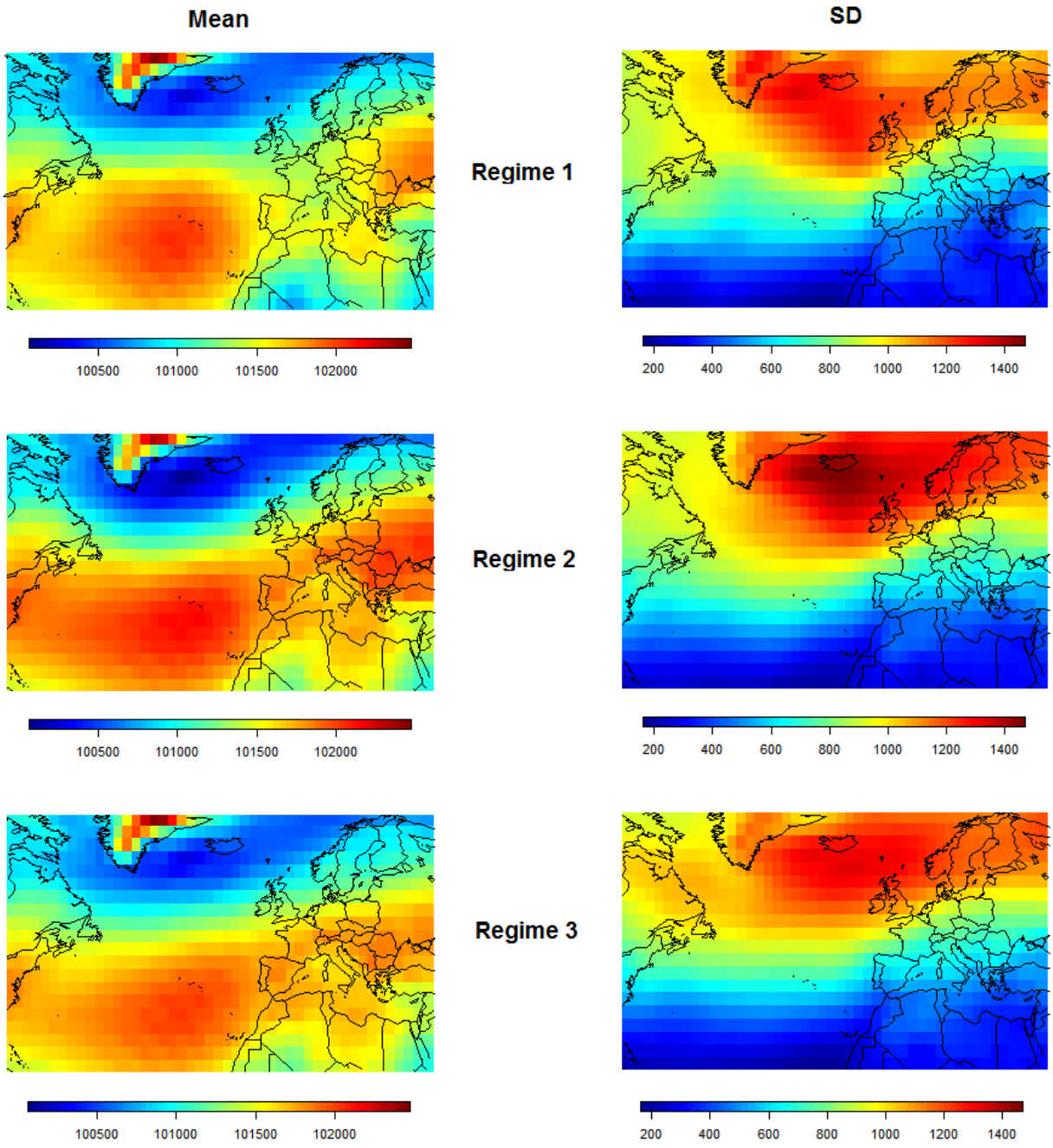


FIGURE 10.3 – Comparaison des cartes de SLP (Moyenne et écart-type) entre les 3 régimes.

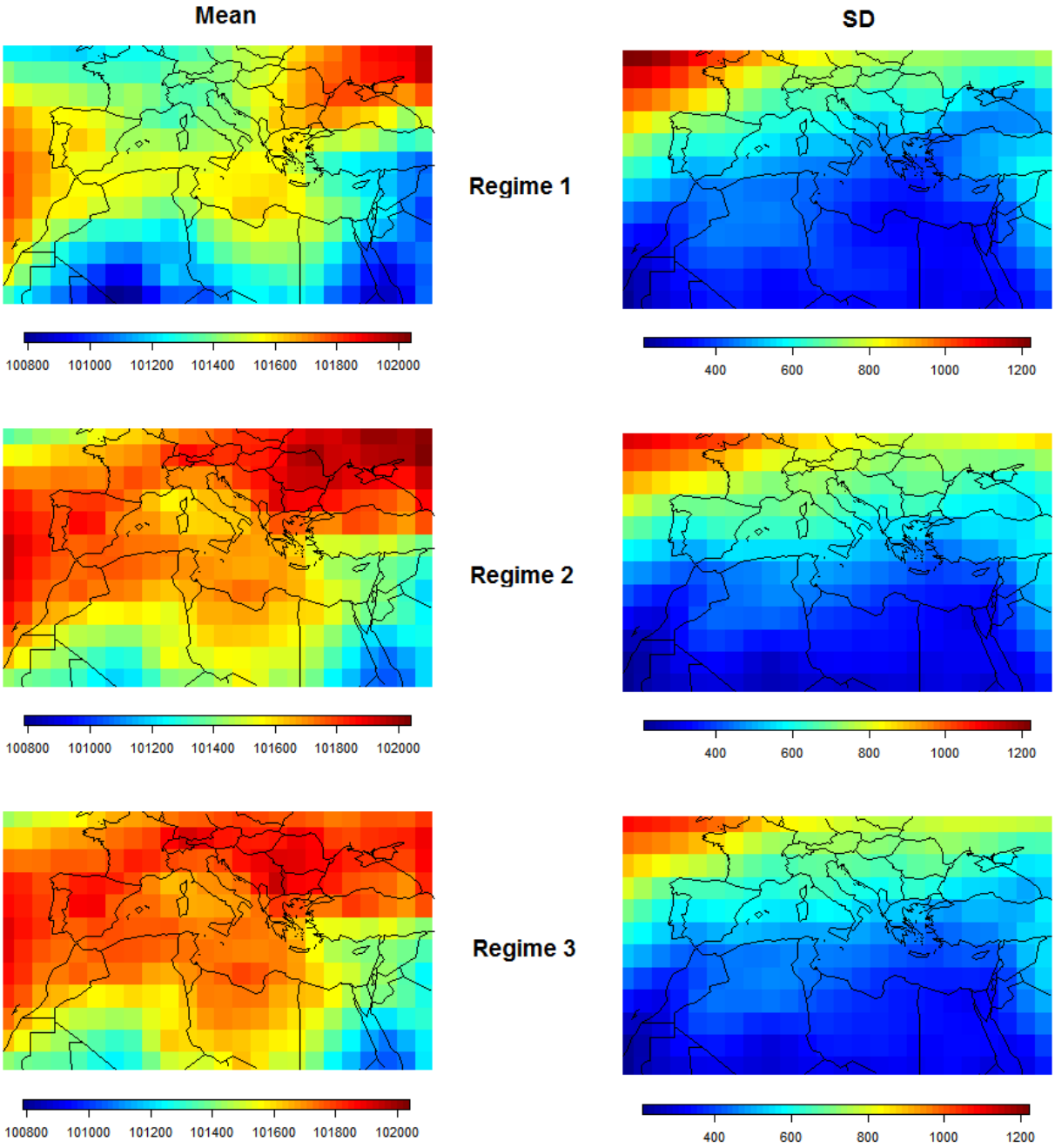


FIGURE 10.4 – Comparaison des cartes de SLP (Moyenne et écart-type) entre les 3 régimes dans la zone restreinte.

A ce stade, les deux approches $T1$ et $T2$ peuvent être appliquées aux deux régions considérées.

10.3 Simulations conditionnelles

Désormais, les deux modèles de transformation de marginales sont testés dans le cadre de la méthodologie de simulation conditionnelle du processus t extrêmal présentée chapitre 6. Cette étape supplémentaire incorpore une composante de non stationnarité temporelle à la méthode. L'application s'effectue dans la même zone que précédemment, le sud de la France ([3.4°E, 6.0°E] x [43.2°N, 44.5°N]), et l'on considère une nouvelle fois les 50 même points conditionnants issus du jeu de données SAFRAN (voir chapitre 6). L'estimation des différents paramètres est effectuée par maximisation numérique de la vraisemblance composite sur les années $1960 - 1999$. Différentes déclinaisons des modèles $T1$ et $T2$ sont considérés avec notamment l'ajout ou non de l'altitude dans les covariables. Les $\overline{\text{CRPS}}$ et CRPSS sont calculés pour comparer les résultats. Les CRPSS sont calculés en prenant le modèle $T0$ le plus simple (LAT+LON) en référence. Ce modèle stationnaire en temps, que l'on cherche à améliorer, est exactement le même que celui utilisé chapitre 6. Il est également important de noter que dans tout ce chapitre, on utilise des partitions tirées aléatoirement dans la première étape de l'algorithme (équation (18) du chapitre 6), pour toutes les méthodes et cela pour des raisons de temps de calcul. De plus, comme ces tirages sont effectués de la sorte pour toutes les méthodes que nous comparons, cela n'aura pas d'impact dans la comparaison des performances. Les tableaux 10.1 et 10.2 décrivent les résultats obtenus à partir de 100 simulations de chaque modèle.

	Modèles	$\overline{\text{CRPS}}$	CRPSS
Modèle $T0$	LAT+LON	15.45	0%
	LAT+LON+ALT	16.37	-5.94%
Modèle $T1$	LAT+LON+SLP	15.68	-1.49%
	LAT+LON+ALT+SLP	16.30	-5.50%
Modèle $T2$	LAT+LON	15.94	-3.17%
	LAT+LON+ALT	15.49	-0.26%

Tableau 10.1 – $\overline{\text{CRPS}}$ et CRPSS obtenus avec 100 simulations conditionnelles des différentes versions des modèles $T1$ et $T2$ (avec modèle $T0$ 'LAT+LON' en référence).

	Modèles	$\overline{\text{CRPS}}$	CRPSS
Modèle $T0$	LAT+LON	15.45	0%
	LAT+LON+ALT	16.37	-5.94%
Modèle $T1$	LAT+LON+SLP	14.67	5.01%
	LAT+LON+ALT+SLP	14.58	5.63%
Modèle $T2$	LAT+LON	16.70	-8.09%
	LAT+LON+ALT	14.54	5.89%

Tableau 10.2 – $\overline{\text{CRPS}}$ et CRPSS obtenus avec simulations conditionnelles des différentes versions des modèles $T1$ et $T2$ (avec modèle $T0$ 'LAT+LON' en référence) avec les données SLP sur la zone restreinte.

Pour la zone géographique large, il apparaît qu'aucun des modèles qui incorporent l'information SLP ne présente de meilleurs résultats que le modèle stationnaire $T0$ de référence. Au vu des résultats, les différents modèles semblent être équivalents en terme de CRPS.

Pour la zone restreinte, en ne rajoutant que l'altitude à la modélisation des marginales, les résultats ne sont pas meilleurs que le modèle avec uniquement la latitude et la longitude. En revanche, certaines versions du modèle parviennent à une amélioration de près de 5% du $\overline{\text{CRPS}}$ ce qui est encourageant. Les versions complètes (avec le maximum de descripteurs) des modèles $T1$ et $T2$ ont des performances similaires, ce qui montre que les deux approches peuvent s'avérer intéressantes.

On peut également formuler l'hypothèse que lorsque la région de variable SLP est trop large, elle n'explique pas suffisamment les extrêmes de précipitations dans notre zone d'étude, ce qui justifierait la différence de performances des deux zones géographiques.

De manière générale, le gain induit par l'incorporation de ces descripteurs de la variable SLP dans la transformation des marginales n'est pas aussi conséquent qu'espéré. Il est possible que les valeurs aux points conditionnants dirigent en grande partie les valeurs simulées. Ainsi, la transformation des marginales n'aurait qu'un faible effet comparé aux valeurs conditionnantes et donc l'approche développée dans ce chapitre ne pourrait pas améliorer grandement les performances de l'algorithme de simulation conditionnelle. Ces conclusions restent néanmoins à pondérer puisque nous n'avons testé qu'une seule covariable à grande échelle et uniquement un résumé de cette information (les deux premières composantes principales). L'ajout d'autres variables pourrait enrichir ce modèle et apporter une amélioration en terme de $\overline{\text{CRPS}}$.

Dans un cadre de projections futures, il est important d'observer que la notion de points conditionnants n'existe plus étant donné qu'il n'y a plus d'observations ponctuelles. Les techniques présentées dans cette partie ne peuvent donc pas être utilisées directement.

Néanmoins, l'information à grande échelle (ici les données SLP) est toujours disponible par le biais de modèles climatiques (e.g. RCM).

10.4 Simulations non conditionnelles

Si dans un cadre prédictif, l'absence d'observation rend l'algorithme de simulation conditionnelle inapplicable en tant que tel, il est toujours possible d'effectuer des simulations non conditionnelles. Le processus t extrêmal à simuler est estimé sur les observations dans le passé et ainsi les simulations non conditionnelles respectent une cohérence spatiale. De plus, la transformation initiale des marginales permet de fournir des valeurs réalistes. Ces simulations non conditionnelles sont obtenues en utilisant la technique de simulation présentée équation (23) du chapitre 6. De même que dans le cadre de la simulation conditionnelle, on peut inclure de l'information à grande échelle dans la modélisation des marginales pour ce modèle de simulation non conditionnelle. Ainsi, ce modèle non stationnaire pourra fournir des distributions de maxima en tout point de la carte et cela pour des années futures tout étant adaptatif et en tenant compte de l'information à grande échelle.

Les modèles $T1$ et $T2$ sont réutilisés dans ce cadre de simulations non conditionnelles. Comme précédemment, les transformations de marginales sont effectuées en utilisant la variable SLP globale (figure 10.1a) et celles de la zone restreinte (figure 10.1b). Les tableaux 10.3 et 10.4 présentent les résultats respectifs.

	Modèles	$\overline{\text{CRPS}}$	CRPSS
Modèle $T0$	LAT+LON	20.13	0%
	LAT+LON+ALT	19.74	1.9%
Modèle $T1$	LAT+LON+SLP	19.72	2.0%
	LAT+LON+ALT+SLP	19.48	3.2%
Modèle $T2$	LAT+LON	23.67	-17.6%
	LAT+LON+ALT	23.20	-15.3%

Tableau 10.3 – $\overline{\text{CRPS}}$ et CRPSS obtenus avec simulations non conditionnelles des différentes versions des modèles $T1$ et $T2$ (avec modèle $T0$ 'LAT+LON' en référence) avec les données SLP sur la zone large.

	Modèles	$\overline{\text{CRPS}}$	CRPSS
Modèle $T0$	LAT+LON	20.13	0%
	LAT+LON+ALT	19.74	1.9%
Modèle $T1$	LAT+LON+SLP	21.33	-6.0%
	LAT+LON+ALT+SLP	20.00	0.6%
Modèle $T2$	LAT+LON	22.46	-11.6%
	LAT+LON+ALT	22.26	-10.6%

Tableau 10.4 – $\overline{\text{CRPS}}$ et CRPSS obtenus avec simulations non conditionnelles des différentes versions des modèles $T1$ et $T2$ (avec modèle $T0$ 'LAT+LON' en référence) avec les données SLP sur la zone restreinte.

Dans les deux cas, il apparaît que la version complète du modèle $T1$ conduit à une amélioration du $\overline{\text{CRPS}}$ même si cette amélioration est légère. En outre, le modèle $T2$ n'est pas du tout efficace pour ce qui est de la simulation non conditionnelle. Cela peut être expliqué par le fait que dans ce cas de figure, la qualité des simulations est entièrement dépendante de la qualité de l'estimation des paramètres, puisqu'il n'y a plus de valeurs conditionnantes pour contrôler les simulations. Et pour ce modèle $T2$, cette estimation se révèle moins précise dans le cadre des régimes automnaux, car certains régimes peuvent comporter peu d'années et donc peu de données.

Il faut également remarquer que le fait que les 8 années tests soient proches des années d'apprentissage peut expliquer que le modèle $T0$ sans autres covariables que LAT et LON soit aussi performant. Néanmoins, il est naturel de penser que cette qualité de simulation du modèle stationnaire temporellement va décroître à mesure que l'éloignement par rapport à la période d'apprentissage va augmenter. Cette constatation classique des séries temporelles pourrait être accentuée par une tendance dans les années futures due au changement climatique. Dans ce cas de figure, le modèle $T0$ se révélerait inadapté pour prédire des phénomènes éloignés dans le futur. En revanche, l'ajout de covariables telles que les variables SLP, issues de modèles climatiques à grande échelle, permet au modèle de s'adapter et de suivre les tendances futures. Ainsi, si pour des années proches les performances des modèles $T0$ et $T1$ sont similaires, il est possible que les projections futures du modèle avec covariables soient plus performantes.

10.5 Conclusion

Ce chapitre a permis d'incorporer de l'information à grande échelle, issue de modèles climatiques, au sein de la méthodologie de simulation conditionnelle. Plus particulièrement, les données de SLP ont été ajoutées dans l'étape de transformation des marginales dans le but d'améliorer la qualité des simulations. En testant plusieurs manières de résumer le

grand nombre de variables fournies par les données SLP et en ciblant différentes zones d'études, nous sommes parvenus à obtenir des descripteurs acceptables pour expliquer les précipitations dans le sud est de la France. Deux familles de techniques ont été utilisées dans la transformation des marginales, à savoir l'ajout de covariables directement dans cette transformation ou une transformation dépendant des régimes automnaux auxquels appartiennent les années.

Dans le cadre de la simulation conditionnelle, les deux transformations présentent de bons résultats, avec une amélioration par rapport au modèle initial. On peut toutefois remarquer que cette amélioration n'est pas aussi conséquente qu'espérée. Il est probable que les valeurs conditionnantes aient un rôle prépondérant dans la qualité des simulations conditionnelles.

Par la suite, dans le cadre de prédictions futures, l'incorporation de variables climatiques à grande échelle a été effectuée pour modéliser la non stationnarité temporelle. Ainsi, les mêmes modèles de transformations de marginales ont été adaptés à la simulation non conditionnelle (car en absence d'observation). Une fois encore, le modèle avec inclusion de covariables dans la modélisation des marginales conduit à de meilleurs résultats. A l'inverse, les modèles de régime automnaux se montrent inefficaces en fournissant de moins bon résultats que le modèle sans covariables. Cette détérioration des résultats est sans doute due à une mauvaise estimation des paramètres du modèle. En effet, certains jeux de paramètres sont estimés sur très peu d'années, car certains régimes de temps sont minoritaires. De plus, l'absence de valeurs conditionnantes donne une grande liberté aux simulations et un plus gros poids à la phase d'estimation des paramètres, ce qui explique les résultats moyens de cette méthode.

Cette étude a également permis d'inclure l'altitude dans la modélisation. Or les apports en termes de qualité de simulations ne sont pas vraiment notables. Outre la difficulté accrue dans la phase de maximisation de la vraisemblance, une explication peut venir de la nature de la variable altitude qui présente une grande variabilité. Ainsi, la qualité de la calibration du modèle dépend en grande partie de l'hétérogénéité des altitudes des points dans l'échantillon de calibration et comme les points à forte altitude sont rares, ils seront sous-représentés dans cette étape. Il n'y a sans doute pas assez de points à forte altitude dans l'échantillon de calibration, ce qui entraîne des difficultés à bien modéliser ce genre de région. Dans la suite, la variable altitude ne sera plus incluse dans la modélisation.

De manière générale, différentes pistes d'améliorations apparaissent. Dans un premier temps, il serait sans doute utile de considérer d'autres variables à grande échelle, comme par exemple la température maximale (pour les systèmes convectifs), les hauteurs de géopotentiel à différents niveaux de pression ou encore l'occurrence et l'intensité des précipitations (à grande échelle). Dans cette optique, il est important de comprendre qu'aucune analyse de sensibilité ou étude de sélection de modèles n'a été effectuée dans

ce travail. L'objectif n'était pas d'obtenir le meilleur modèle possible mais un modèle donnant des résultats encourageants, montrant que l'approche est viable. C'est donc un point d'amélioration possible.

Un autre aspect important de cette modélisation vient de l'incorporation de la non stationnarité temporelle par ajout de covariables dans la transformation des marginales. Ainsi, cela permet de modifier les valeurs du processus t extrêmal simulé en fonction de variables qui évoluent dans le temps. Cependant, la structure de dépendance spatiale de ce processus reste inchangée et identique à celle qui est estimée sur les observations passées. Ajouter une composante temporelle sous forme d'une variable à grande échelle dans la modélisation de la structure spatiale du processus (e.g. dans dans la fonction de corrélation du W_i apparaissant équation (4) du chapitre 6) peut être une solution pour aller plus loin dans la non stationnarité temporelle du processus.

De manière générale, cette partie a permis d'incorporer de l'information à grande échelle dans les approches de simulations existantes et de fournir des résultats encourageants. Cependant, ayant la sensation d'être arrivés au bout de ce modèle sans apport de variable supplémentaire, ce qui est difficile pour l'estimation des paramètres, nous avons décidé de changer notre démarche en contexte de « pur » downscaling.

Dans le chapitre suivant, cet objectif d'incorporer de l'information à grande échelle est approfondi, en proposant une approche hybride reposant à la fois sur des techniques de downscaling statistiques et sur la méthodologie générale de simulation conditionnelle de processus max-stables.

Chapitre 11

Méthode hybride de downscaling spatial

Dans le chapitre précédent, nous avons constaté que l'ajout de covariables dans la modélisation des marginales ne conduisait qu'à une légère amélioration des résultats, dans le cas de la simulation non conditionnelle. Or l'objectif de cette partie est d'obtenir des réalisations de maxima, lorsqu'il n'y a plus d'observations disponibles. Pour cela, des méthodes de downscaling statistique sont présentées et utilisées dans la suite.

De plus, une approche hybride proposant l'utilisation à la fois des méthodes de downscaling statistique classiques et l'algorithme de simulation conditionnelle présenté dans la première partie est introduite. Cette approche permet d'effectuer de la descente d'échelle d'extrêmes de manière à ce que les valeurs obtenues respectent la structure spatiale des observations.

Une fois encore, l'application est effectuée sur les mêmes données de précipitations que précédemment, à savoir celles du sud-est de la France. Différentes versions de cette approche hybride sont comparées à des méthodes de downscaling statistique usuelles. Pour cela, plusieurs critères sont utilisés, évaluant les caractéristiques univariées, multivariées, de distributions et de variabilités spatiales. En complément des tableaux de scores qui apparaîssent dans ce chapitre, des résultats complémentaires sont présenté chapitre A.

Ce travail a fait l'objet d'un article soumis à la revue « Journal of Geophysical Research ».

A spatial hybrid approach for downscaling of extreme precipitation

Aurélien Bechler[a,b,c], Mathieu Vrac[c], Liliane Bel[a,b]

[a] *INRA, UMR 518 Math. Info. Appli., F-75005 Paris, France*
[b] *AgroParisTech, UMR 518 Math. Info. Appli., F-75005 Paris, France.*
[c] *LSCE-IPSL, Centre d'Etudes de Saclay, Orme des Merisiers, Bat. 701 91191 Gif-sur-Yvette, France*

Abstract

For a few decades, climate models are used to provide future scenarios of precipitation with increasingly higher spatial resolution. However, this resolution is not yet sufficient to describe efficiently what happens at local scale. Dynamical and statistical methods of downscaling have been developed and allow us to make the link between two levels of resolution and enable us to get values at a local scale based on large-scale information from global or regional climate models. Nevertheless, both the extreme behavior and the spatial structures are not well described by these downscaling methods. We propose a two-step methodology, called spatial hybrid downscaling (SHD), to solve this problem. The first step consists in applying a univariate (i.e., 1-dimensional) statistical downscaling to link the high- and low-resolution variables at some given locations. Once this 1d-link is performed, a conditional simulation algorithm of max-stable processes adapted to the extremal t process enables us to get conditional distributions of extreme precipitation at any point of the region. Application is performed on precipitation data in the south of France where extreme (Cevenol) events have major impacts (e.g. floods). Different versions of the SHD approach are tested. Most of them show particularly good results regarding univariate and multivariate criteria and overcome classical downscaling techniques. Furthermore, these conclusions are robust to the choice of the 1d-link functions tested, and to the choice of the conditioning points to drive the conditional local-scale simulations performed by the SHD approach.

Keywords: Max-stable processes, spatial model, conditional simulation, extreme precipitation, downscaling.

1. Introduction

Future climate projections from Global Climate models (GCMs) offer a way to investigate future trends of climatic events. One main issue comes from the spatial resolution of these model outputs as GCMs give values on a grid of about 200 km of resolution (IPCC, 2011). Hence, when the process of interest is very localized, as for example for extreme precipitation, GCM outputs are not relevant. Furthermore, most of the GCMs are built to characterize the mean behavior of some variables and thus are not able to produce realistic rare and very intense (i.e., extreme) local events. For these reasons, GCM outputs can not be used directly as inputs in impact models (e.g., agriculture, water resources, pollution, etc...) that need local-scale data. A commonly used solution consists in simulating more local values by performing downscaling techniques.

There are two main downscaling families. The first one is a dynamical approach based on regional climate models (RCMs) (e.g., Laprise, 2008; Rummukainen, 2010) and may be seen as a regional version of the global models, aiming to solve the physical equations of the atmosphere dynamics, but at a higher spatial resolution than the GCM (from 50km to 5km). Thus, these RCMs are able to solve explicitly a part of the physical processes involved at a more local scale-grid and are consistent from a meteorological and hydrological point of view. Nevertheless, even with the increasing computational means, the complexification due to the improvement of these regional equations leads to the possibility of using RCM only over limited regions and time periods. Moreover, these RCMs may still be at a resolution too rough to be used in a local framework by impact models, and need to be themselves downscaled.

More recently, another approach called statistical regionalization, has emerged as an efficient alternative. This approach does not explicitly solve the equations of the dynamics of the atmosphere but is based on statistical modeling of the existing relations between large-scale variables (coming from GCM or RCM) and variables observed at a more local scale. Hence, due to its statistical formulation, this approach enables fast simulations of local variables and the low computational resources requested makes possible the modeling of the uncertainty associated to the local projections. In practice, statistical downscaling gathers a high diversity of very different methods based on transfer functions (e.g., Friederichs and Hense, 2007; Vrac et al., 2007; Levavasseur et al., 2011), weather generator (e.g., Richardson et al., 1998; Wilks, 2012), clustering-based techniques (Huth, 2001; Boé et al., 2006) and bias correction approaches (e.g., Déqué, 2007; Piani et al., 2010; Vrac et al., 2012) .

Some of these statistical downscaling techniques are spatialized, i.e. the generated values are spatially correlated and respect the spatial structure of the observations. This spatial characteristic is of huge importance to generate fields of spatially coherent values. This spatial coherency can be achieved by different ways, as for example clustering or multisite downscaling (see e.g., Wilks, 2012, for a recent review), and is of huge importance for various impact models (Fowler and Ekström, 2009). This is also true for extreme events for which downscaling requires to adapt the existing tools (e.g., Fowler et al., 2010; Kallache et al., 2011) by using extreme statistical tools (e.g., Vrac and Naveau, 2007; Carreau and Vrac, 2011).

More generally, the understanding and forecast of extreme events are major challenges. Indeed, as they are inherently rare and often unexpected, society is not ready to face them. One of the particularities of extreme events is that their probability decreases rapidly when the intensity increases, and so do the damages and the cost of protection against such events. Both analyses of the last decade observations and future projections from climate models tend to suggest that the occurrence and the intensity of such events may increase in the following years (IPCC, 2011). In that case, an increasing number of catastrophes such as storms, flooding or crop failures may be expected. In addition to their direct effects in terms of human lives or building destruction, extreme events will also have indirect impacts such as increasing costs for building specific infrastructures or higher insurance premiums. Additionally, these events may also have consequences on global population displacements (IPCC, 2011). Then, a better understanding and modeling of extreme events is a crucial need for both economical and societal points of view. From a mathematical point of view, the statistical theory of ex-

treme values appears to be the suitable probabilistic framework for modeling their frequency and intensity. The extreme value theory (EVT) is well established for univariate stationary series (Beirlant et al., 2004) and there is now a growing effort to develop EVT models able to reproduce spatial variability and consistency and to deal with non-stationary problems (e.g., Panagoulia et al., 2014; Jonathan et al., 2014). Multivariate extreme distributions appear in climate and environmental applications in order to take into account the dependence among extremes (e.g., Coles and Tawn, 1994; De Haan and De Ronde, 1998; Schlather and Tawn, 2003; Fawcett and Walshaw, 2012). For the study of extremes in a spatial context (i.e., not only multivariate), max-stable processes are suitable tools for modeling and inference. Initially introduced by De Haan (1984), different parametric models have later been proposed by Smith (1990), Schlather (2002), Kabluchko et al. (2009) or Padoan (2011). Max-stable processes have been used in many climatic applications such as for modeling rainfall (e.g., Buishand et al., 2008; Davison et al., 2012; Bechler et al., 2014), temperature (e.g., Davison and Gholamrezaee, 2012; Fuentes et al., 2013), snowfall (e.g., Gaume et al., 2013) or snow depth (e.g., Blanchet and Davison, 2011).

In this work, we propose a new spatial downscaling approach by explicitly modeling the spatial structure of the data, focusing on extreme events. This approach is called "spatial hybrid downscaling" (SHD) since that it takes advantage of both the physical consistency of the regional climate model outputs and the statistical spatial properties of a conditional simulations procedure to provide distribution of local-scale variables at any location in a given region. Indeed, conditional simulations allow us to statistically generate values at locations without any recording, in agreement with surrounding observations. Furthermore, while this SHD approach is developed in a general way, it is adapted in the present application to downscaling of extreme events by performing conditional simulations of max-stable processes. The application and tests of this SHD model focuses on the French Mediterranean basin, which is a key spot in term of occurrences of meteorological extreme events such as heavy precipitations (e.g., Ducrocq et al., 2008; Lebeaupin et al., 2006; Beaulant et al., 2011; Vrac and Yiou, 2010; Vrac et al., 2012). Indeed, major extreme rainfall are regularly observed in this region close to the Cévennes mountains. Modeling and forecast of such extreme precipitation, called "Cevenol events", are major challenges for example for the hydrological studies in this area because those very intense precipitations often lead to severe floods in this region (e.g., Ducrocq et al., 2008).

The structure of this article is the following. After the presentation of the two datasets (local observations and larger-scale RCM outputs) in section 2, the general methodology of the SHD approach, its adaptation to extremes events and all the theoretical considerations are described in section 3. Then, section 4 displays a comparison of the results of the SHD approach with those from other methods. Different versions of the SHD approach are also discussed in this section. Some conclusions and discussions are finally given in section 5.

2. Large- and local-scale data

As for any statistical downscaling technique, the calibration of our SHD approach requires a reference local-scale dataset of the variable of interest, as well as a larger-scale dataset to

be downscaled.

2.1. Local-scale data: the SAFRAN precipitation

Daily precipitation data have been extracted from the SAFRAN dataset (Quintana-Seguí et al., 2008; Vidal et al., 2010) in the southeast of France covering the region [3.4°E, 6.0°E] x [43.2°N, 44.5°N] – as represented by the square in Figure C.1 – with a spatial resolution of 8 km × 8 km, which corresponds to 457 grid cells. Then, for each cell, we computed the autumnal maximum daily precipitation (from 91 days between September 1^{st} and November 30^{th}) for $1960 - 2007$: hence, for each grid-cell, we have 48 autumnal maximum daily precipitation values. The choice of the autumnal period allows us to assume temporal stationarity, as it is commonly accepted that rainfall data can be considered as stationary within a given season.

2.2. Larger-scale data: RCM outputs

The larger-scale data to be downscaled in this study are the outputs of the RCM named WRF (Weather Research and Forecasting), version 3.1.1. The WRF model is a limited-area model developed by the National Center for Atmospheric Research (NCAR) (Skamarock et al., 2008). More precisely, the fall (SON) daily precipitation outputs of the model IPSL-WRF311 for 1989-2007 have been downloaded. This corresponds to WRF applied at a 0.44^o resolution over the Mediterranean region of the CORDEX[1] exercise to dynamically downscale the ERA-Interim reanalyses (Dee et al., 2011) within the hindcast evaluation context of the Med-CORDEX[2] initiative (Ruti et al., 2014). The IPSL-WRF311 model – hereafter referred to as WRF-MED – is one of the few nudged RCMs involved in Med-CORDEX. Nudging provides one way to constrain the RCM simulations and keep them from diverging too far from the coarse-scale driving fields (e.g., von Storch et al., 2000). For our region of interest (see Figure C.1), 12 WRF-MED grid-cells are needed. On each of these grid-cells and for each year, we computed the autumnal maximum of the daily precipitation. The Figure C.1 shows the autumnal maximum of precipitation extracted from the WRF-MED data for the year 1989.

Hereafter, the RCM outputs are considered as large-scale data while SAFRAN data are referred to as local-scale data. For both datasets, the years before 2 000 serve as learning period and the years 2000-2007 are used for the evaluation.

3. Spatial Hybrid Downscaling Approach

3.1. General Methodology

This method is meant to use both climate model outputs and statistical advantages of the conditional simulations to provide distribution of local-scale variables at any location of a given region, by using both large-scale and local-scale information. Performing conditional simulation needs conditioning points at which observed values are available. When only large-scale climate outputs are available instead of observed values, the idea is to simulate data at some local-scale conditioning points based on univariate downscaling of the climate

[1] http://wcrp-cordex.ipsl.jussieu.fr/
[2] http://www.medcordex.eu/

model outputs. Accordingly, the methodology of the proposed SHD method can be divided into two steps:

1. **Calibration.**
   1-a Establish a univariate statistical link (called "1d-link function"), calibrated on the calibration period, between the large-scale information and some selected local-scale conditioning points.
   1-b Choose a spatial model.
   1-c Model the marginal (i.e., univariate) distributions with covariates including latitudes and longitudes and estimate the parameters of the spatial model based on the selected conditioning points.

2. **Simulation.**
   2-a Based on the 1d-link functions, generate downscaled-projections (DS-PRJ) at the selected conditioning locations from the large-scale outputs on the evaluation period.
   2-b Perform a conditional simulation algorithm of the variable of interest according to the spatial model with the DS-PRJ as conditioning values.

This approach enables us to provide simulated data for any location within the region, even if we do not have observations for those locations. More generally, by repeating several times those conditional simulations, it can also be estimated the empirical distribution of the variable of interest at any location, even in the future when no observation is available but only climate models outputs.

Many aspects of this method have to be detailed. First, as in this work we are interested in extreme precipitation events, in step **1-c**, a max-stable model has been chosen, more precisely an extremal t process (Padoan, 2011). It has to be noted that this step adds some extra-information in comparison to the classical downscaling techniques. Indeed, the parameters of the simulated max-stable process are estimated from observations (see section 3.3). This brings information about the spatial characteristics of the phenomenon. The gain compared to more classical statistical downscaling methods is that this approach takes explicitly into account both the spatial and the extreme components in the downscaling procedure. The question of the choice of the 1d-link function is treated section 3.2. Descriptions of the max-stable extremal t process and of the margins modeling (including latitudes and longitudes as covariates) are then provided section 3.3. The algorithm used to perform conditional simulations is presented section Appendix B. The selection of the number and the location of the points will be discussed section 4.1.

3.2. 1d-link functions

Two 1d-link models commonly used in downscaling context have been tested here in steps **1-a** and **2-a**. The first method, called "Cumulative Distribution Function - transform" (CDF-t) and initially proposed by Michelangeli et al. (2009), consists in comparing the CDF of the local-scale variable Z_{LS} and the one of the large-scale variable Z_{HS}. Let us note $F_{LS,c}$ the CDF of the predictand and $F_{HS,c}$ the CDF of the predictor on the calibration period. The CDF-t approach looks for a transformation T from $[0,1]$ to $[0,1]$ such that:

$$T(F_{HS,c}(x)) = F_{LS,c}(x) \qquad x \in \mathbb{R}. \tag{1}$$

The underlying assumption is that Equation 1 remains valid when $F_{LS,c}$ and $F_{HS,c}$ are replaced by respectively $F_{LS,f}$ and $F_{HS,f}$ the CDFs of the predictand and the predictor in the future. A solution consists in choosing T as

$$T(u) = F_{LS,c}(F^{-1}_{HS,c}(u)) \qquad u \in [0,1], \tag{2}$$

and then we obtain the relationship

$$F_{LS,f}(x) = F_{LS,c}(F^{-1}_{HS,c}(F_{HS,f}(x))) \tag{3}$$

in replacing u by $F_{HS,f}(x)$ in Equation 2. Every CDF is empirically modeled from the historical observations, and from the large-scale data of the calibration or evaluation time periods. Then the transformation (equation (3)) is performed for each margin and gives us one DSPRJ for the selected locations by applying a quantile-mapping based technique between large- and local-scale distributions of the evaluation period. More theoretical and technical details can be found in Vrac et al. (2012).

The other tested 1d-link function consists in a linear regression at each location where the large-scale precipitation is used as an explanatory variable of the local scale. In our application, after different tests (not shown here), the best model consists in explaining the local-scale precipitation by the mean of the large-scale precipitation over the nine RCM grid-cells around the SAFRAN location of interest as shown by equation (4).

$$Z_{LS} = a \times \overline{Z_{HS}}^{9cells}. \tag{4}$$

As each 1d-link function is univariate, they are different from one conditioning location to another (i.e., for example, the parameter a is estimated for each location).

3.3. Max-stable processes and conditional simulations

The spatial model chosen for steps **1-c** and **2-b** is described here. In order to model fields of extremes, max-stable processes (De Haan, 1984) are common tools. Extremal t process (Padoan, 2011) is a particular max-stable process. Opitz (2013) proposed a spectral construction of this process which leads to the general expression

$$\mathbf{Z} = \{Z(x)\} = \left\{m_\alpha^{-1/\alpha} \max_{i\geq 1} \xi_i W_i(x)\right\},\ \alpha > 0,\ x \in S \subset R^2,\ \text{with } S \text{ a Borel set}, \tag{5}$$

with $m_\alpha = \sqrt{\pi}^{-1} 2^{(\alpha-2)/2}\Gamma((\alpha+1)/2)$, ξ_i a Poisson point process on (0,+∞) with intensity measure $d\Lambda(t) = \alpha t^{-(\alpha+1)}dt$ and W_i independent copies of a stationary standard Gaussian process on $\mathbb{R}^2$ with correlation function ρ. $\mathbf{Z}$ margins are α-Fréchet distributed. It can be remarked that for $\alpha = 1$, the representation in Equation (5) leads to a Schlather process. Hence the extremal t process can be considered as a generalization of the Schlather process, with a greater flexibility. The correlation function $\rho(\cdot)$ determines the spatial dependence structure and the parameter α characterizes the spatial asymptotic dependence, that is the dependence in extreme values.

Before being modeled by an extremal t process, data must be stationarized in space by transforming the margins into α-Fréchet distributions. As we are interested in maxima of precipitation values, margin distributions are in the GEV family. To be able to perform the reverse transformation for any location in the region and to deal with spatial non-stationarity, the parameters (μ, σ and ξ) of the GEV are modeled as functions of the spatial coordinates latitudes and longitude (hereafter referred to as LAT and LON) except ξ that is assumed to be constant over the whole region:

$$\begin{aligned} \mu &= \exp\big[\mu_0 + \mu_1 * LAT + \mu_2 * LON\big] \\ \sigma &= \exp\big[\sigma_0 + \sigma_1 * LAT + \sigma_2 * LON\big] \\ \xi &= \xi_0 \end{aligned} \tag{6}$$

It can be observed that the exponential formulation is employed for μ and σ as σ has to be positive for model definition reasons and μ is chosen to be positive because of the nature of the observed phenomenon (precipitation). The parameters of the model, i.e.:

1. μ_0, μ_1, μ_2, σ_0, σ_1, σ_2, ξ, the parameters for the margin model,
2. α, the extremal t parameter,
3. and θ, the parameters of the correlation function ρ,

have to be estimated from the data. Nevertheless, for a dimension higher than 2, the multivariate density of the extremal t process is not tractable and usual likelihood maximization methods are not available. In this paper we used a full likelihood maximization based on partial censoring of components above a high margin threshold proposed by Thibaud and Opitz (2013). Furthermore, the correlation function chosen here is the exponential one.

In step **2-b**, in order to provide distributions at any location in the region of interest, conditional simulations are used. These tools provide simulations of random fields at locations without observation, given observations in fixed (i.e., conditioning) locations, according to a chosen spatial model. The simulated fields are required to fit the observed values at the fixed locations. In contrast to simple simulations for which simulated values follow a target distribution conditional simulations take into account constraints expressed by the observations. Furthermore, when other techniques provide individual values or indicators, conditional simulations provide distributions at every new location, which allows us to get numerous statistical quantities as e.g., quantile, dispersion indicators, etc... The conditional simulation techniques are then powerful tools. For some models, such as Gaussian or Boolean, conditional simulations algorithms based on kriging techniques are quite straightforward (Lantuéjoul, 2002) but for max-stable models, this kind of techniques is not applicable. Then, Dombry et al. (2013) proposed a general framework for conditional simulations of max-stable processes. This algorithm, originally used on Schlather and Brown-Resnick processes (Dombry et al., 2013) can be applied to extremal t processes (Ribatet, 2013; Bechler et al., 2014). In this paper, we use a version of the algorithm proposed by Bechler et al. (2014), which is computationally heavy but shows good results and allows to consider a large number of conditioning points.

4. Results

Different versions of the SHD approach are described here and compared to some basic downscaling techniques.

- As downscaling techniques, *Interpolation* is a bilinear interpolation of the RCM outputs values taken at the center of each grid-cells and *Linear downscaling + Interpolation* consists in performing a linear regression for the conditioning points (same as in section 3.2) and then a bilinear interpolation.

- We test four versions of the SHD approach. Two of them use *CDF-t* and *linear regression* (described section 3.2) as 1d-link function. For comparison purposes, we perform two extreme cases of 1d-link function. In the *raw* method, no 1d-link function is used (i.e., the RCM outputs are taking directly as conditioning values). In the *optimal* method, the real values from the SAFRAN dataset are taking as conditioning values. It is the virtual case where the 1d-link function is the most accurate and is able to rebuild exactly the observations. This *optimal* method should not be overcome by any other SHD methods. For the four versions, the conditioning points are chosen by co-localization techniques, i.e. by taking the points which are at the center of each grid-cells.

Then, different scores are proposed to compare the performances of these methods (for details about these scores, see Appendix Appendix C). In addition to the Continuous Rank Probability Score (CRPS) and the 95% Quantile Score (QS_{95}) (Gneiting and Raftery, 2007), the multivariate energy score (ES) which generalizes the CRPS (Gneiting et al., 2008) is computed to assess the spatial coherency of the simulations. Each of the marginal scores is calculated at every location and global scores are obtained by taking the mean of the individual scores over the region. ES is a multivariate score calculated for each simulated field. Global score is obtained by taking the mean over every simulations. The lower these scores, the better.

Then, Kolmogorov-Smirnov test statistics is computed as well to compare the concatenated distribution of the real values over the eight test years to the distribution obtained by the different methods. At last, the Cramer-Von Mises criterion, which assesses the quadratic mean gap between the cdfs (in contrary to the K-S which evaluates the biggest gap between the distributions), is used. For these two tests, a statistic close to 0 leads to the acceptation of the H_0 hypothesis which corresponds to the fact that the two samples come from the same distribution.

Then, we compute the variograms on the simulations obtained with the different methods. We assess the spatial variability coherency by comparing, for each method, the variogram of the numerous simulated fields to the one from the real observations. The Root Mean Squared Error (RMSE) between the variogram of the observations and the ones from the simulations is calculated every year for each methods.

Table C.1 shows for all the scores the respective skill-scores which correspond to the percentage of improvement of the method in comparison to a reference method (here the *Linear downscaling + Interpolation*). The mathematical formulation is given in Appendix Appendix C.

It appears that regarding marginal distribution (i.e. CRPS, QS, K-S and CVM scores), the different versions of the SHD approach present better results than the reference one (only downscaling). More precisely, the linear regression 1d-link function seems to be the more relevant choice. This 1d-link function remains also good regarding spatial coherency (ES) and spatial variability (RMSE_v). Figure C.2, displaying the different variograms for year 2003 shows that the SHD approach is not able to reproduce the whole spatial variability of the data. We can also see on that Figure that CDF-t 1d-link function is the technique which preserves the most the variability of the data which is consistent with the fact that it has the lowest RMSE_v. The two variograms of the basic downscaling methods are far from the observed one.

Nevertheless, Figure C.3, displaying annual variograms obtained with the optimal method, for several years, shows big differences of variability between the years. It appears that the optimal method presents similar behaviors and is able to capture the global changes between years.

In addition, Figure C.4 shows the evolution of the mean over the region of the simulations for the different models, compared to the mean of the observations. The four versions of the SHD approach are able to reproduce the general variations of the observations and also the good magnitude even if *Raw* method seems to under-estimate the real values. *Interpolation* and *Linear downscaling + Interpolation* follow the variations but provide very biased values. The performances of each method are evidently consistent with the results from table C.1.

The correlation coefficient between observed spatial means, and simulated spatial means has also been calculated for each method. Results (given in Figure C.4) are globally the same between the SHD approach and basic downscaling techniques, except for the *Raw* method which is not as good as the others. The absence of 1d-link function has major consequences on the quality of the conditional simulations. It also appear that the downscaling techniques are able to reproduce well the general trend of the observations but underestimate it, even after the 1d-link function. Nevertheless, the high correlation is a positive result about the quality of the information brought by the RCM.

4.1. Importance of conditioning points

Until now, to determine the SAFRAN dataset locations used as conditioning points, we have chosen the point which is the closest to the center of each RCM grid-cell. Thus, the number of conditioning points is necessarily the same as the number of grid-cells above our area (e.g., 12 for WRF-MED dataset). However, one can wonder if other techniques would give better results.

Then, to investigate this question, we applied a clustering technique on the SAFRAN dataset to exhibit different classes and in each one a representative location which is kept as conditioning point. The main advantage of this approach is that the choice of conditioning points is performed independently of the RCM. Once the selection done, tests with different RCMs can be performed without any modification. Bernard et al. (2013) proposed a Partitioning Around Medoids (PAM) algorithm for maxima based on the F-madogram (Cooley et al., 2006). PAM is a partitioning algorithm which represents each cluster center by a representative point called medoid (when the k-means algorithm gives the cluster means as representative point). In order to determine the optimal number of cluster, we use a silhouette criterion (Rousseeuw, 1987). The silhouette coefficient for the point i after the PAM

algorithm runs with K clusters is defined as

$$s_i(K) = 1 - \frac{d_{i,k(i)}}{\gamma_{i,-k(i)}}, \tag{7}$$

where $d_{i,k(i)}$ represents the intra-cluster distance between medoid $k(i)$ and point i and $\gamma_{i,-k(i)}$ corresponds to the smallest distance between point i and all the other medoids but $k(i)$. If the silhouette coefficient is around 1 the point i is considered to be well classified. The mean over every points i of this coefficient enables to choose the optimal number of conditioning points, which is 57 in our case. Figure C.5 shows the global silhouette coefficient, which is the mean of $s_i(K)$ over every i, according to the number of clusters and displays the selected medoids for the optimal criterion.

We are now able to perform conditional simulations with the 57 medoids as conditioning points. It appears (results not shown) that the general ranking of the different methods is more or less the same (similar skill-scores). But the results are not as good as expected since there is no improvement (except for the optimal model) in the different scores. We see in Figure C.5 that the medoids are chosen to be in high proportion the uncommon points near the boundaries. Some of them are not representative of an area and have been chosen by the PAM algorithm because of their atypicalness (and may be alone in a class). These characteristics lead to a difficult calibration of the 1d-link function and then to non accurate conditional simulations. It has also to be noted that, the partition draw step (by Gibbs sampler) in the conditional simulation algorithm becomes inaccurate and very time consuming when the number of conditioning points is so high. Thus the parsimony in the number of conditioning points may be an important criterion.

In order to limit the number of conditioning points, we use the PAM algorithm on maxima to find some representative points among the SAFRAN dataset by limiting the search to one representative point per grid-cell. Figure C.6 shows the conditioning points found by the algorithm.

Table C.1 shows that again, the different versions of the SHD approach give better results than the basic downscaling methods, but this time it is the CDF-t method which is the most accurate, see Table C.2. Nevertheless, the scores are sometimes not as good as the ones with the co-localization techniques. It seems that in this case the 1d-link function has some difficulties to build the relationship between the two scales. That explains the fact that the scores of the optimal method are very good (better than the co-localization techniques) and that it is not always true in the other cases. The advantage of choosing these particular points is lost in the 1d-link function modeling. In general it appears that the choice of the conditioning points has a major impact on the accuracy of the 1d-link function and on the performances of the algorithm. In the next section, we try to avoid this choice of points by randomly changing them for each simulation.

In this last test, the conditioning points will not be set by co-localization techniques or by PAM algorithm but they will change at each simulation. More precisely, for each simulation, one point into each RCM grid-cell is drawn randomly and these locations will be used as conditioning points for one conditional simulation. This methodology, called Stochastic Spa-

tial Hybrid Downscaling (stochastic SHD), is coherent with the fact that a RCM value is the mean over a grid-cell. Nevertheless, it raises two major issues. First, it is computationally impossible to estimate the model parameters with likelihood maximization for each drawn ensemble of 12 points. Then, the parameters are estimated only once by maximization of the full likelihood built with all available observations (i.e. at the 457 locations). The same estimated parameter set is used for each margins transformation (Equation A.4) no matter the drawn conditioning points. Second, in order to use the conditional simulations algorithm, we need to draw a partition ensemble of 12 points. It is very unlikely to run the Gibbs sampler for each simulation. Therefore, we perform a Data Augmentation algorithm proposed by Lantuéjoul and Bel (2014), a deteriorated but still accurate algorithm. It can be run much faster that a Gibbs Sampler. This makes the stochastic SHD method possible. For the score calculation of the location used as conditioning points, the downscaled value are used. Hence, the quality of the Optimal method is a little bit overrated even if the proportion of conditioning points at each simulation is low (12/457).

Table C.3 describes the results of the different versions of the stochastic SHD compared to other methods. With this way of choosing points, every score is improved and the hybrid approach remains the most accurate method. In addition, CDF-t seems to be the best 1d-link function. Moreover, this method prevents us from performing a case-by-case selection of conditioning points and provides distribution at every points of the map, compared to the other techniques that only gives the downscaled value as estimation at the conditioning locations.

It is interesting to note that for each choice of the conditioning locations, the SHD approach with 1d-link function (i.e. except the *Raw* method) presents very good results. Thus, the superiority of this approach seems to be robust to the choice of the conditioning points.

Finally, in order to see if the changes of conditioning points improve the quality of the simulations, Table C.4 summarizes the skill-scores obtained with the different methods for the choice of the conditioning points. This time, we only compare SHD approaches and the reference is the co-localization method for the choice of points. Based on the previous results, the 1d-link function considered for each method is CDF-t.

It appears that PAM into a grid-cell does not necessarily improve the quality of the conditional simulation and that stochastic SHD presents generally the best results. Nevertheless, depending on the criterion, co-localization method provide quite proper results which are not outcome by more complicated methods. Then, the choice of the conditioning points will largely depend on the density of the dataset available. If it does not allow the use of stochastic SHD, co-localization techniques can still be used and present good results.

The variograms of the optimal SHD approaches with the different choices of points are presented Figure C.7. It appears that the PAM approach and even stochastic SHD allows more variability in the conditional simulations. This variability largely depends on the conditioning values used in the conditional simulation algorithm and on their locations. High conditioning values will also bring more variability in the simulations. This aspect is not discussed here because it depends on the data and no generalization is possible after our study. For the location of the conditioning points, co-localization technique leads to uniformly spaced points whereas other techniques allow wider space without conditioning points. That can explain partially the low variability of the simulations with co-localization technique. Nevertheless, as it appears in Table C.4 for PAM method, enabling too much variability in the simulations

does not lead to better results.

4.2. Results with other RCM outputs

In order to test if the performances of our SHD approach are robust, another RCM dataset has also been tested. The WRF-IPSL-INERIS44 (e.g., Gobiet et al., 2012; Vautard et al., 2013), hereafter referred as WRF-EURO data are RCM outputs available from 1989 to 2007. The main difference with the previous RCM (WRF-MED) is that the WRF-EURO is not nudged. We use the expected daily mean precipitation over a grid with a resolution of about 50km. Given this spatial resolution, the WRF-EURO data cover our region of interest with 15 grid-cells. On each of these grid-cells, we build the autumnal maximum of the daily mean precipitation. The Figure C.8 shows the autumnal maximum of precipitation built extracted from the WRF-EURO data. We consider the years 1989-1999 as learning period and the years 2000-2007 are used once again for the evaluation. We choose the conditioning points by co-localization techniques, i.e. by taking the points which are the closest to the center of each grid-cell. We then obtain 15 conditioning points.

The skill-scores results are given in Table C.5. Even if the results and conclusions are globally similar as these for the nudged WRF-MED RCM and that the SHD approach remains relevant, its accuracy in term of spatial coherency is not as good as before. It may be explained by the fact that this RCM is not nudged. This has some consequences on the results of the whole algorithm.

5. Conclusions and discussion

We have proposed and investigated a new hybrid approach which combines downscaling techniques and statistical conditional simulations to get distributions of extremes at any location. Performing downscaling technique only for a given subset of locations and then using these downscaled values as conditioning values for the conditional simulation algorithm enable to provide spatially coherent simulations of extreme fields, even for locations with no observations available.

The application of this approach on real datasets and in realistic conditions shows great improvements compared to basic downscaling techniques. Furthermore, different versions of this SHD approach have been tested to highlight the role of the different components of the method. First, 1d-link functions manage to reduce the bias existing in the RCM outputs. Second, thanks to the inference of the max-stable process from the real observations, the conditional simulation step brings additional information and spatial coherency compared to the univariate downscaling. It appears that both statistical downscaling and conditional simulation steps are essential for the quality of the SHD approach. It is also important to note, that the quality of the results from this SHD approach largely depends on the quality of the climate model outputs. The bias of these outputs may be corrected by the downscaling step, but the trend remains driven by the climate model. This intuitive statement has been confirmed by the results of the SHD approach on an other RCM.

Beside, the choice of the number and the location of the conditioning points is a major issue as well. Different selection strategies have been developed and it appears that the increasing number of conditioning points does not necessary lead to an improvement of the

results. An other way of selection has been proposed, consisting in choosing randomly one conditioning point into each grid-cell and to change it at each simulation. It seems to provide the best results even if it presents more technical difficulties. In practice, this choice will depend on the number of observations available, but it has to be kept in mind that the choice of the conditioning points has to be done very carefully. Nevertheless, it is important to note that the superiority of the SHD approach compared to classical downscaling techniques seems to be robust to the choice of conditioning points.

Among the various possible extensions and perspectives studies, this choice of points constitutes one potential improvement. We can try to change the clustering algorithm to an algorithm which deals with outlier classes (prevent point from being alone in a class) as for example the super-paramagnetic clustering algorithm (Blatt et al., 1996). Other technical studies can be performed, for example a comparative study between the conditional simulation algorithm presented here and the data augmentation one performed for time-save reasons when applying the stochastic SHD method.

Furthermore, the downscaling techniques performed in this work are not specific to extreme modeling. Some bias correction techniques adapted to extremes may improve the quality of the downscaling step. For example, an extension of the CDF-t method has been proposed by Kallache et al. (2011). This bias correction technique, called XCDF-t, based on the extreme value theory could be incorporated quite straightforward to the SHD approach in future work to see if there is an improvement in the simulation quality. Nevertheless, the general methodology of the SHD approach is not specific to extreme events and could be applied to more "regular" event with classical statistical spatial models.

One of the major perspectives of this work is to provide prediction of future local-scale variables, based on climate model outputs. Thus, we could describe the local consequences of the different large-scale scenarios provided by (IPCC, 2011). It may also be interesting to compare the results of the SHD approach presented here to those obtained by taking directly the GCM outputs instead of the RCM. Indeed, it may appear that statistical downscaling techniques on GCM are more efficient than those on RCM (Eden et al., 2014).

Nevertheless, this SHD approach will enable us to generate fields and times series meant to be useful for many impact studies aiming to anticipate the consequences of extreme precipitation events in a climate change context.

Appendix A. Max-stable processes

Commonly used tools to model extreme values and specially observation of maxima are max-stable processes. Let us consider a stationary random field $\{Z(x)\}_{x\in\mathbb{R}^d}$ with a given distribution $\mathcal{H}_0$ and $(Z_1,\dots,Z_n)$ independent copies of Z. If there exists sequences $a_n(x)\in\mathbb{R}$ and $b_n(x)\in\mathbb{R}^+$ such that for all $n\in\mathbb{N}$,

$$\left\{\frac{\max_{i=1,\dots,n} Z_i(x) - a_n(x)}{b_n(x)}\right\}_{x\in\mathbb{R}^d} \tag{A.1}$$

follows the distribution $\mathcal{H}_0$, then $\{Z(x)\}_{x\in\mathbb{R}^d}$ is a max-stable process (De Haan, 1984).
As a consequence, all finite dimensional marginal distributions are max-stable, and in particular, the univariate marginal $Z(x)$ distribution belongs to the Generalized Extreme Value

(GEV) family:

$$\mathbb{P}(Z(x) \leq z) = \begin{cases} \exp\left(-\left[1 + \xi(x)\left(\frac{z-\mu(x)}{\sigma(x)}\right)\right]^{-1/\xi(x)}\right) & \text{if } \xi(x) \neq 0, \\ \exp\left(-\exp\left(\frac{z-\mu(x)}{\sigma(x)}\right)\right) & \text{if } \xi(x) = 0, \end{cases} \tag{A.2}$$

where $\mu(x)$, $\sigma(x)$ and $\xi(x)$ are respectively the location, scale, and shape parameters at site x. Furthermore, the sign of $\xi(x)$ will determine the family distribution $Z(x)$ belongs to among Fréchet ($\xi > 0$), Gumbel ($\xi = 0$) or Weibull ($\xi < 0$), and consequently the distribution tail characteristics.

Different max-stable models have been proposed and used in the past by several authors, e.g. Brown and Resnick (1977), Smith (1990), Schlather (2002) or Kabluchko et al. (2009). Their margins are Fréchet distributed, i.e.

$$\mathbb{P}(Z(x) \leq z) = e^{-z^{-1}}. \tag{A.3}$$

If Z is distributed according to a GEV(μ, σ, ξ), then Y defined as

$$Y = \left(1 + \frac{\xi(Z-\mu)}{\sigma}\right)_+^{1/\alpha\xi} \tag{A.4}$$

follows an α-Fréchet distribution (Beirlant et al., 2004).

Appendix B. Conditional Simulation algorithm

Dombry et al. (2013) proposed a general framework for conditional simulations of max-stable processes. The algorithm is divided in three steps:

- **Step 1:** Generate a partition of the conditioning points.
- **Step 2:** For each class of this partition, generate one extremal function which is equal to the observed value at the points of this class and which is lower to the observed values at the other locations.
- **Step 3:** Independently of the two first steps, generate a high number of sub-extremal functions which are lower to the observed values at all conditioning points.

This algorithm, originally used on Schlather and Brown-Resnick processes (Dombry et al., 2013) can be applied to extremal t processes (Ribatet, 2013; Bechler et al., 2014). In this paper, the first step is performed thanks to a Gibbs sampler as in Dombry et al. (2013). For the second step, the general methodology of simulations under constraints detailed by Bechler et al. (2014) is applied and the simulation of sub-extremal functions is performed via an acceptance-rejection step.

Appendix C. Scores

Letting F_n be the predictive cumulative distribution function derived from the n conditional simulations $\{\hat{y}^{(j)}\}_{j=1,\dots,n}$ and y the test observation, the continuous ranking probability score is written

$$\mathrm{crps}(F_n; y) = \int_{-\infty}^{+\infty} \left[F_n(t) - H(t-y)\right]^2 dt, \tag{C.1}$$

where $H(u)$ is the Heavyside function which is equal to 0 for $u < 0$ and 1 otherwise. The steeper the cdf is around the real value, the better, and a good score is a score close to 0. It also appears that the calculation of the crps will result in a value with the same unit as the variable.

The quantile score (qs) is defined as

$$\mathrm{qs}_p(F_n; y) = \nu_p(y - F_n^{-1}(p)) \tag{C.2}$$

where p is the probability of interest and $\nu_p(u) = pu$ if $u \geq 0$ and $\nu_p(u) = (p-1)u$ otherwise. As shown by Friederichs and Thorarinsdottir (2012), the crps is proportional to the integral of the qs_p over all possible probabilities p. Then a good qs_p is a qs_p close to 0 as well. By choosing a high probability p, the quantile score focuses on the predictive performance in the upper tail (Gneiting and Raftery, 2007).

In addition to these two univariate scores, which assess the marginal qualities of the conditional simulation, we propose to use the energy score (Gneiting et al., 2008) which may be written as:

$$ES(\hat{\mathbf{y}}^{(1)}, \dots, \hat{\mathbf{y}}^{(n)}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} ||\hat{\mathbf{y}}^{(i)} - \mathbf{y}|| - \frac{1}{2n^2}\sum_{j=1}^{n}\sum_{i=1}^{n} ||\hat{\mathbf{y}}^{(i)} - \hat{\mathbf{y}}^{(j)}|| \tag{C.3}$$

where $\hat{\mathbf{y}}^{(i)}$ is one of the n simulated fields and $\mathbf{y}$ is the field of observations. This score enables us to assess the spatial coherency of the simulations.

For comparison of distribution purposes, we also use the test statistic of the Kolmogorov-Smirnov test:

$$\text{K-SS} = \sup_x |F_n(x) - F(x)| \tag{C.4}$$

and

$$\text{CVMS} = \int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 \, \mathrm{d}x \tag{C.5}$$

where F is the empirical cumulative distribution function from the observations.

In order to compare models, Skill scores can be built (Gneiting and Raftery, 2007):

$$S^{skill} = \frac{S^{mod} - S^{ref}}{S^{opt} - S^{ref}} \tag{C.6}$$

with S^{mod}, S^{ref} and S^{opt} the score of the current model, the score of a reference model and the optimal score respectively. For models that improve on the reference model, S^{skill} belongs

to $[0,1]$ and 1 corresponds to the optimal case. For the five scores presented here, we have $S^{opt}=0$ and then the corresponding skill scores can be written for these three scores as

$$\text{CRPSS}=\frac{\overline{\text{CRPS}}^{ref}-\overline{\text{CRPS}}^{mod}}{\overline{\text{CRPS}}^{ref}},\text{QSS}_p=\frac{\overline{\text{QS}}_p^{ref}-\overline{\text{QS}}_p^{mod}}{\overline{\text{QS}}_p^{ref}},\text{ESS}=\frac{\text{ES}^{ref}-\text{ES}^{mod}}{\text{ES}^{ref}}. \tag{C.7}$$

$$\text{K-S SS}=\frac{\text{K-SS}^{ref}-\text{K-SS}^{mod}}{\text{K-SS}^{ref}},\text{CVM SS}=\frac{\text{CVMS}^{ref}-\text{K-SS}^{mod}}{\text{CVMS}^{ref}} \tag{C.8}$$

which corresponds to the percentages of the scores improvement brought by the current model with respect to the reference model.

Acknowledgments

This work has been partially supported by the ANR-project McSIM, the ANR-project StaRMIP, the VW-project PLEIADES and the ANR-project REMEMBER. This work is part of the Med-CORDEX[3] and EURO-CORDEX[4] initiative.

All computations have been made in R. An R package containing functions for the conditional simulations procedure based on the extremal t processes should soon be made available on the CRAN website[5] or upon request to the authors.

IPSL-WRF311 dataset is available on http://www.medcordex.eu/medcordex.php, WRF-IPSL-INERIS44 dataset on http://www.euro-cordex.net/EURO-CORDEX-Data.2613.0.html and SAFRAN dataset is available upon request to Météo-France.

[3]http://www.medcordex.eu/
[4]http://www.euro-cordex.net/
[5]http://cran.r-project.org/

Figures

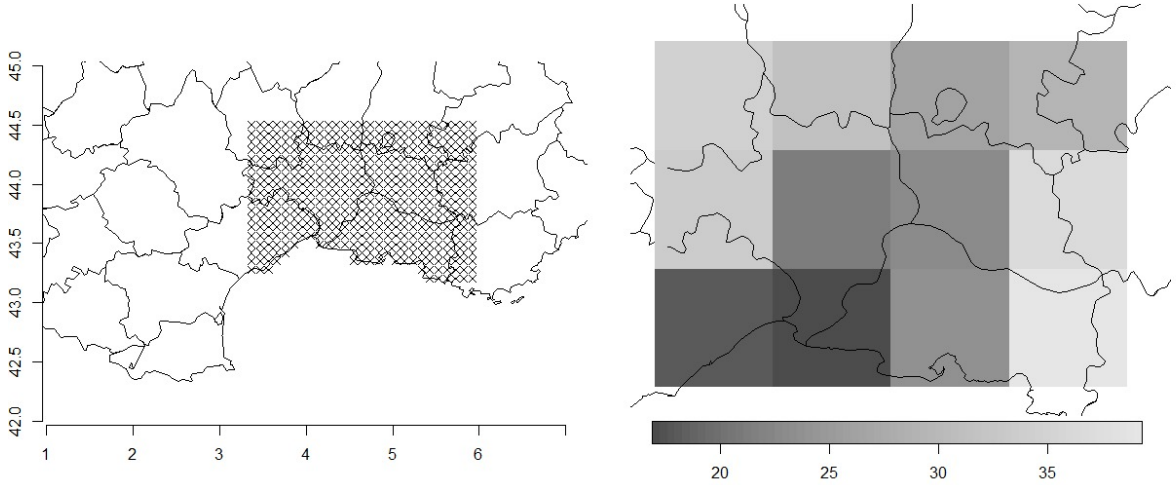


Figure C.1: Left: Study area of SAFRAN data subset. Right: WRF-MED maximum precipitation data for the year 1989 (in mm).

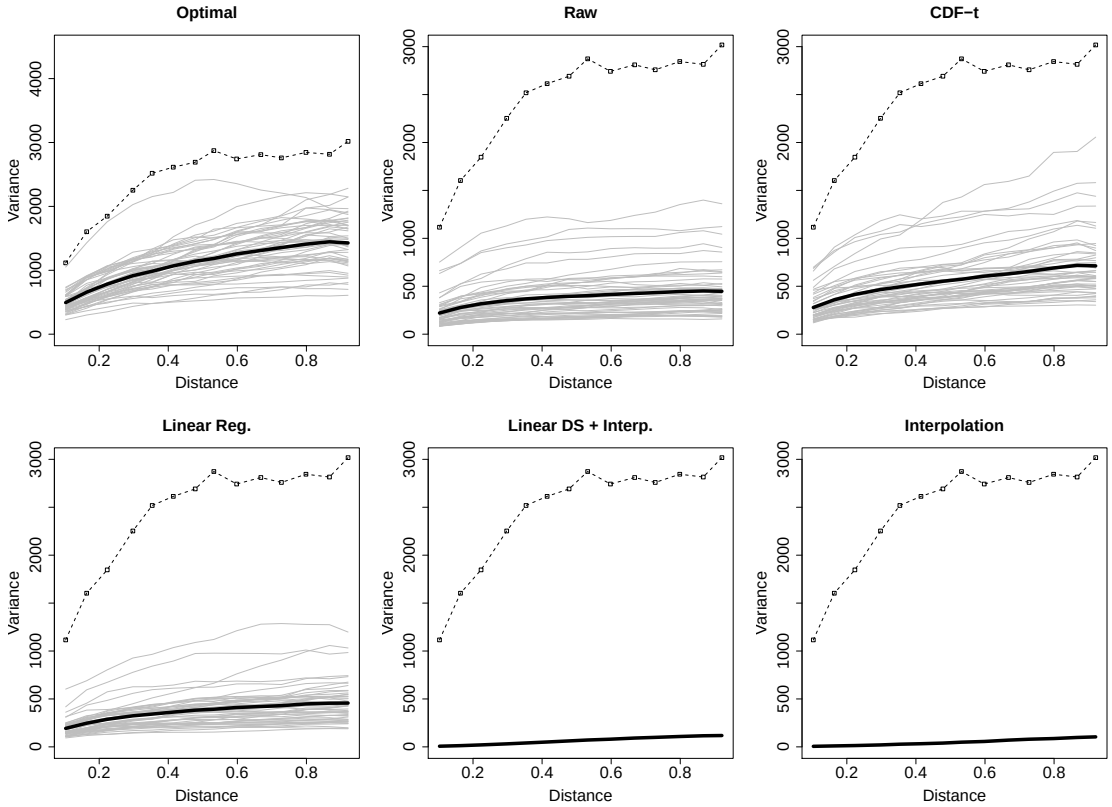


Figure C.2: Variograms obtained by the six methods. In dotted line, the real one, in grey the ones of the different conditional simulations and in black, the mean of every conditional simulations (for the year 2003).

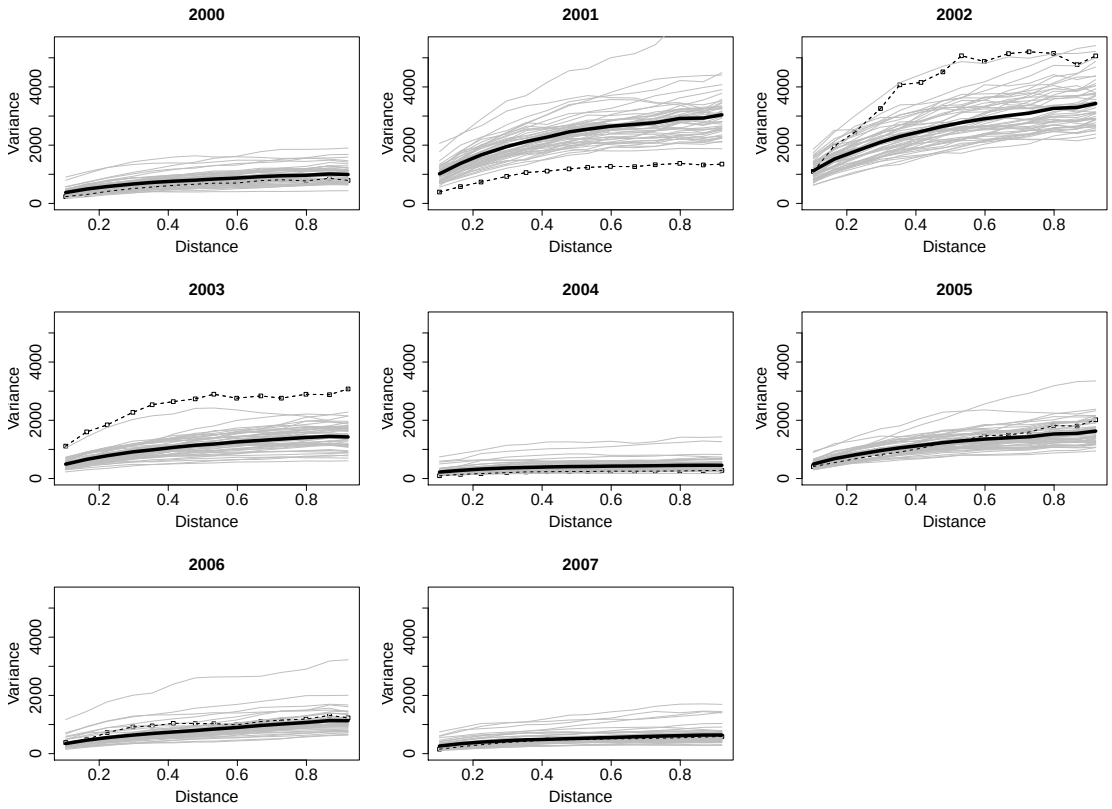


Figure C.3: Variograms obtained for the eight years by the optimal method. The real one in dotted line, in grey the ones of the different conditional simulations and in black, the mean of every conditional simulations.

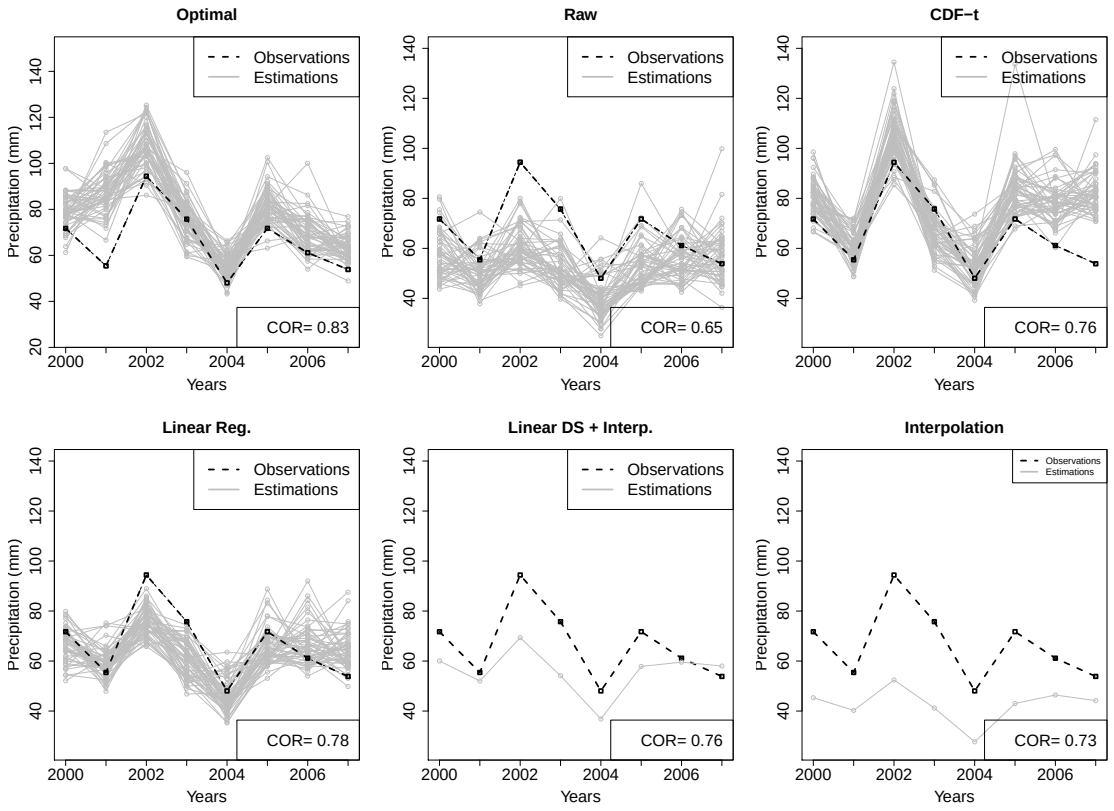


Figure C.4: Spatial annual means obtained by the six methods. The observed in black and each trajectory in grey is one conditional simulation.

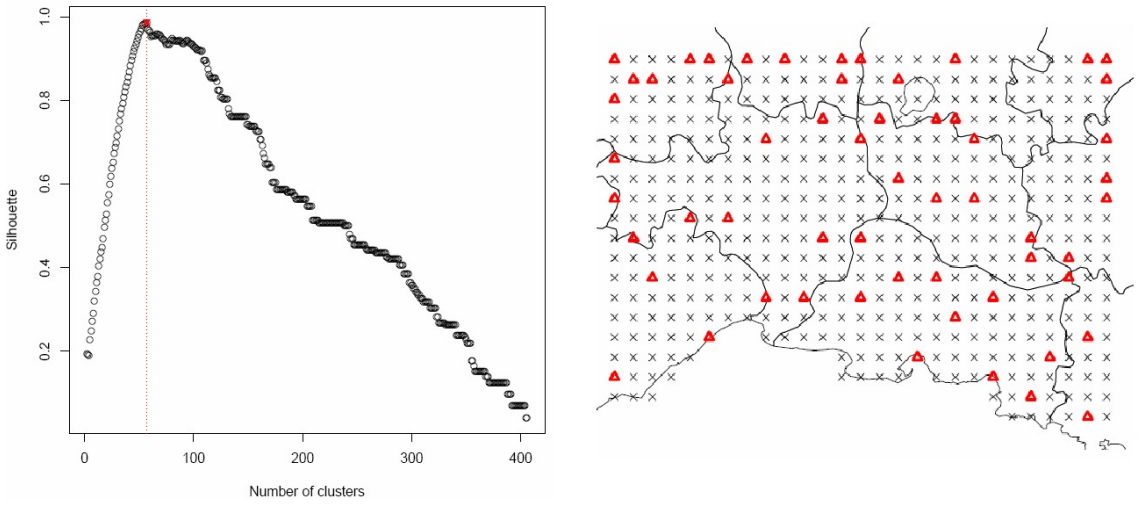


Figure C.5: Left: Silhouette criterion for choosing the number of clusters in PAM algorithm. Right: Description of the 57 medoids (triangles) among the SAFRAN dataset.

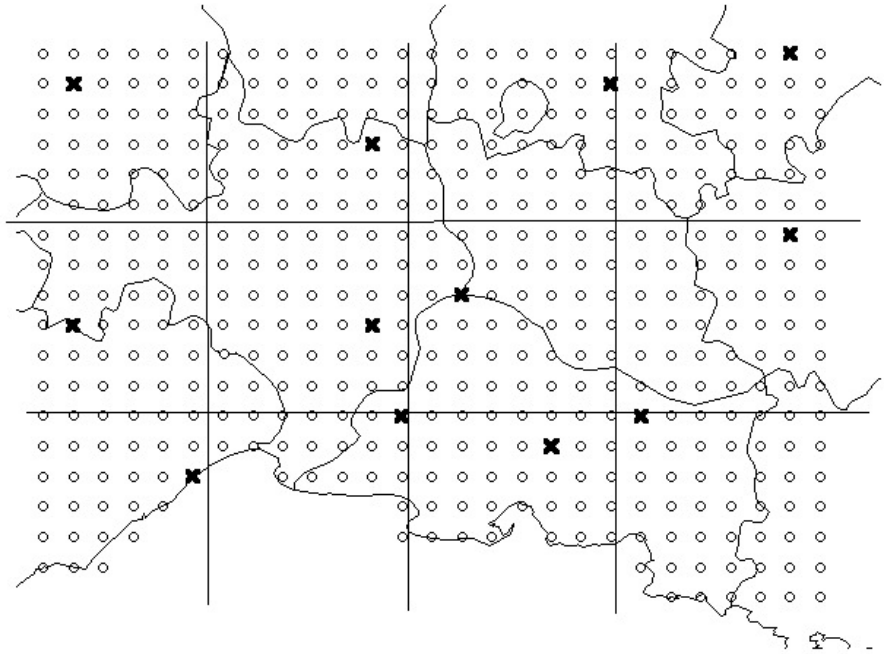


Figure C.6: Locations of the 12 conditioning points obtained by a PAM algorithm into each RCM grid-cell.

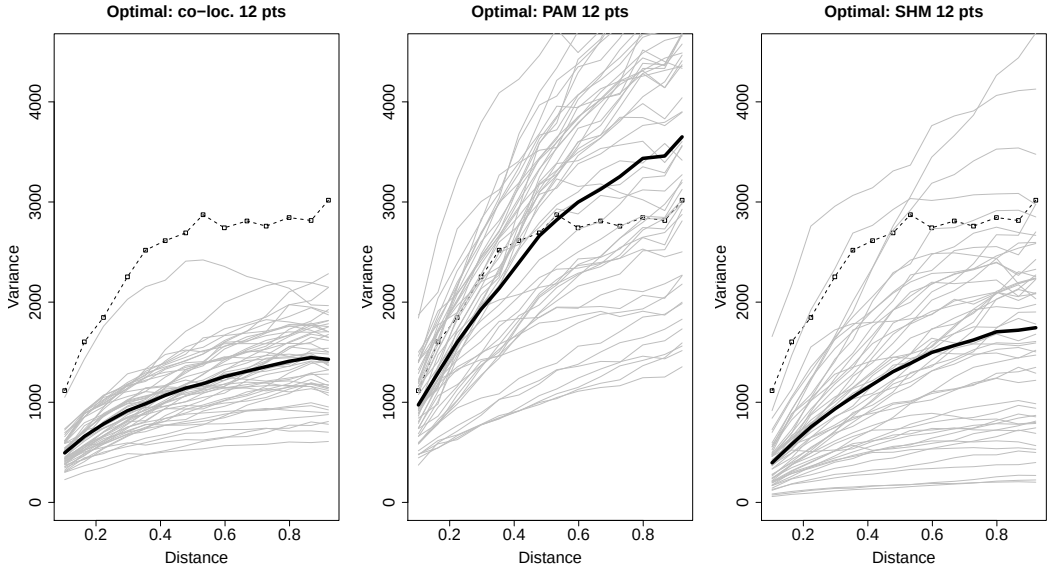


Figure C.7: Variograms obtained with the optimal method for the different methods for the choice of conditioning points. The real one in dotted line, in grey the ones of the different conditional simulations and in black, the mean of every conditional simulations (year 2003).

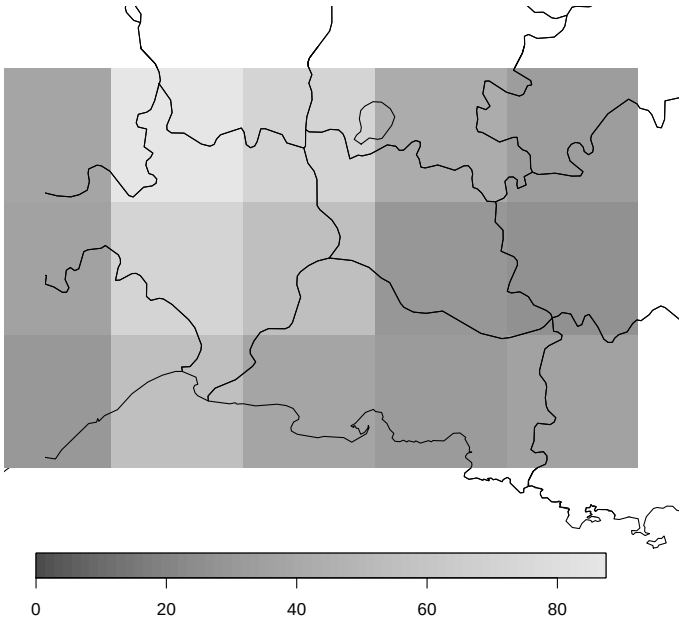


Figure C.8: WRF-EURO maximum precipitation data for the year 1989 (in mm).

Tables

Table C.1: Skill-Scores (with *Linear downscaling + Interp.* as reference) of the SHD algorithm with the different methods for building the DS-PRJ from the WRF-MED dataset. 12 conditioning points chosen by co-localization techniques.

	Models	CRPSS	QSS_{95}	ESS	K-S SS	CVM SS	$RMSE_v$ SS
No simulations	Interpolation	-20.1%	-54.8%	-17.7%	-19.8%	-64.4%	-0.4%
	Linear DS + Interp.	0%	0%	0%	0%	0%	0%
Conditional Simulations	Raw	14.0%	44.4%	14.6%	10.4%	-11.7%	20.2%
	CDF-t	14.1%	**60.6%**	17.0%	12.3%	-8.9%	**40.4%**
	Linear Regression	**20.1%**	53.4%	**22.8%**	**15.0%**	**1.0%**	22.5%
	Optimal	24.2%	70.7%	29.5%	18.7%	3.9%	52.9%

Table C.2: Skill-Scores (with *Linear downscaling* as reference) of SHD algorithm with the different methods for building the DS-PRJ observations from the WRF-MED dataset with 12 conditioning points chosen by PAM algorithm into each grid-cell.

	Models	CRPSS	QSS_{95}	ESS	KS SS	CVM SS	$RMSE_v$ SS
No simulations	Interpolation	-6.5%	-22.6%	-4.7%	-19.8%	-20.2%	1.9%
	Linear downscaling	0%	0%	0%	0%	0%	0%
Conditional Simulations	Raw	9.6%	26.9%	12.2%	-7.7%	-9.4%	12.4%
	CDF-t	**22.5%**	**58.9%**	**26.8%**	**17.6%**	**35.7%**	**34.2.%**
	Linear Regression	18.4%	38.8%	18.9%	4.6%	15.2%	11.8%
	Optimal	34.6%	72.4%	39.3%	23.6%	43.3%	61.5%

Table C.3: Skill-Scores (with *Linear downscaling* as reference) of stochastic SHD algorithm with the different methods for building the DS-PRJ observations from the WRF-MED dataset with 12 randomly chosen and switching conditioning points.

	Models	CRPSS	QSS_{95}	ESS	KS SS	CVM SS	$RMSE_v$ SS
No simulations	Interpolation	-29.2%	-78.4%	-26.4%	-27.4%	-88.1%	-3.2%
	Linear downscaling	0%	0%	0%	0%	0%	0%
Conditional Simulations	Raw	-8.3%	-4.9%	-6.1%	-9.5%	44.9%	7.8%
	CDF-t	**12.2%**	**39.8%**	**14.0%**	**15.8%**	**50.4%**	**23.8%**
	Linear Regression	10.1%	31.6%	11.6%	8.8%	35.5%	14.2%
	Optimal	36.7%	75.8%	37.3%	18.1%	56.2%	67.6%

Table C.4: Skill-Scores (with Co-localization as reference) of SHD algorithm with the different methods for the choice of the conditioning points. For each method, the CDF-t technique is used as 1d-link function.

Models	CRPSS	QSS_{95}	ESS	KS SS	CVM SS	$RMSE_v$ SS
Co-localization	0%	**0%**	0%	0%	0%	**0%**
PAM into a grid-cell	-0.2%	-27.6%	0.2%	9.8%	18.9%	-16.1%
Stochastic SHD	**9.7%**	-15.9%	**5.0%**	**15.8%**	**20.9%**	-29.1%

Table C.5: Skill-Scores (with *Linear downscaling* as reference) of SHD algorithm with the different methods for building the DS-PRJ observations from the WRF-EURO dataset with 15 conditioning points.

	Models	CRPSS	QSS_{95}	ESS	KS SS	CVM SS	$RMSE_v$ SS
No simulations	Interpolation	-35.8%	-62.4%	-29.7%	-25.4%	-72.5%	4.6%
	Linear downscaling	0%	0%	0%	0%	**0%**	0%
Conditional Simulations	Raw	18.9%	50.7%	20.3%	**16.6%**	8.3%	25.6%
	CDF-t	10.8%	**62.7%**	19.6%	10.0%	-3.0%	35.6%
	Linear Regression	**24.9%**	55.1%	**25.3%**	13.6%	**2.6%**	**52.9%**
	Optimal	36.3%	73.5%	34.8%	20.0%	15.1%	60.5%

References

Beaulant, A.-L., Joly, B., Nuissier, O., Somot, S., Ducrocq, V., Joly, A., Sevault, F., Deque, M., Ricard, D., 2011. Statistico-dynamical downscaling for mediterranean heavy precipitation. Quarterly Journal of the Royal Meteorological Society 137 (656), 736–748.

Bechler, A., Bel, L., Vrac, M., 2014. Conditional simulations of extremal t process: Application to fields of extreme precipitation, under review, http://hal-agroparistech.archives-ouvertes.fr/hal-01053605.

Beirlant, J., Goegebeur, Y., Teugels, J., Segers, J., Waal, D. D., Ferro, C., 2004. Statistics of extremes. Theory and applications. Hoboken, NJ: John Wiley & Sons.

Bernard, E., Naveau, P., Vrac, M., Mestre, O., Oct. 2013. Clustering of Maxima: Spatial Dependencies among Heavy Rainfall in France. Journal of Climate 26, 7929–7937.

Blanchet, J., Davison, A. C., 2011. Spatial modeling of extreme snow depth. The Annals of Applied Statistics 5 (3), 1699–1725.

Blatt, M., Wiseman, S., Domany, E., Apr 1996. Superparamagnetic clustering of data. Phys. Rev. Lett. 76, 3251–3254.

Boé, J., Terray, L., Habets, F., Martin, E., 2006. A simple statistical-dynamical downscaling scheme based on weather types and conditional resampling. Journal of Geophysical Research: Atmospheres 111 (D23).

Brown, B. M., Resnick, S. I., 1977. Extreme values of independent stochastic processes. Journal of Applied Probability 14 (4), 732–739.

Buishand, T., de Haan, L., Zhou, C., 2008. On spatial extremes: with application to a rainfall problem. The Annals of Applied Statistics, 624–642.

Carreau, J., Vrac, M., 2011. Stochastic downscaling of precipitation with neural network conditional mixture models. Water Resources Research 47 (10).

Coles, S., Tawn, J., 1994. Statistical methods for multivariate extremes - an application to structural design. Journal of the Royal Statistical Society - Series C: Applied Statistics 43 (1), 1–48.

Cooley, D., Naveau, P., Poncet, P., 2006. Variograms for spatial max-stable random fields. In: Dependence in probability and statistics. Springer, pp. 373–390.

Davison, A. C., Gholamrezaee, M. M., 2012. Geostatistics of extremes. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science 468 (2138), 581–608.

Davison, A. C., Padoan, S. A., Ribatet, M., 05 2012. Statistical modeling of spatial extremes. Statistical Science 27 (2), 161–186.

De Haan, L., 1984. A spectral representation for max-stable processes. The Annals of Probability 12 (4), 1194–1204.

De Haan, L., De Ronde, J., 1998. Sea and wind: multivariate extremes at work. Extremes 1 (1), 7–45.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kallberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., Vitart, F., 2011. The ERA-interim reanalysis: configuration and performance of the data assimilation system. Quarterly Journal of the Royal Meteorological Society 137 (656), 553–597.

Déqué, M., 2007. Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: model results and statistical correction according to observed values. Global and Planetary Change 57 (1), 16 – 26, extreme Climatic Events.

Dombry, C., Éyi-Minko, F., Ribatet, M., 2013. Conditional simulation of max-stable processes. Biometrika 100 (1), 111–124.

Ducrocq, V., Nuissier, O., Ricard, D., Lebeaupin, C., Thouvenin, T., 2008. A numerical study of three catastrophic precipitating events over southern france. ii: Mesoscale triggering and stationarity factors. Quarterly Journal of the Royal Meteorological Society 134 (630), 131–145.

Eden, J., Widmann, M., Maraun, D., Vrac, M., 2014. Comparison of GCM- and RCM-simulated precipitation following stochastic postprocessing. Journal of Geophysical Research - AtmospheresTo appear.

Fawcett, L., Walshaw, D., 2012. Estimating return levels from serially dependent extremes. Environmetrics 23 (3), 272–283.

Fowler, H. J., Cooley, D., Sain, S. R., Thurston, M., 2010. Detecting change in UK extreme precipitation using results from the climateprediction. net BBC climate change experiment. Extremes 13 (2), 241–267.

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

Friederichs, P., Hense, A., 2007. Statistical downscaling of extreme precipitation events using censored quantile regression. Monthly Weather Review 135, 2365–2378.

Friederichs, P., Thorarinsdottir, T. L., 2012. Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. Environmetrics 23 (7), 579–594.

Fuentes, M., Henry, J., Reich, B., 2013. Nonparametric spatial models for extremes: application to extreme temperature data. Extremes 16 (1), 75–101.

Gaume, J., Eckert, N., Chambon, G., Naaim, M., Bel, L., 2013. Mapping extreme snowfalls in the french alps using max-stable processes. Water Resources Research 49 (2), 1079–1098.

Gneiting, T., Raftery, A. E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102 (477), 359–378.

Gneiting, T., Stanberry, L., Grimit, E., Held, L., Johnson, N., 2008. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. Test 17 (2), 211–235.

Gobiet, A., Jacob, D., Euro-Cordex Community, Apr. 2012. A new generation of regional climate simulations for Europe: The EURO-CORDEX Initiative. In: Abbasi, A., Giesen, N. (Eds.), EGU General Assembly Conference Abstracts. Vol. 14 of EGU General Assembly Conference Abstracts. p. 8211.

Huth, R., 2001. Disaggregating climatic trends by classification of circulation patterns. International Journal of Climatology 21 (2), 135–153.

IPCC, 2011. Managing the risks of extreme events and disasters to advance climate change adaptation. Tech. rep., Intergovernmental Panel of Climate Change (IPCC).

Jonathan, P., Ewans, K., Randell, D., 2014. Non-stationary conditional extremes of northern North Sea storm characteristics. Environmetrics 25 (3), 172–188.

Kabluchko, Z., Schlather, M., de Haan, L., 09 2009. Stationary max-stable fields associated to negative definite functions. The Annals of Probability 37 (5), 2042–2065.

Kallache, M., Vrac, M., Naveau, P., Michelangeli, P.-A., 2011. Non-stationary probabilistic downscaling of extreme precipitation. Journal of Geophysical Research: Atmospheres 116 (D5).

Lantuéjoul, C., 2002. Geostatistical simulation: models and algorithms. Springer.

Lantuéjoul, C., Bel, L., 2014. Simulation conditionnelle du processus de schlather. In: 46èmes Journées de Statistique de la SFdS.

Laprise, R., 2008. Regional climate modelling. Journal of Computational Physics 227 (7), 3641–3666.

Lebeaupin, C., Ducrocq, V., Giordani, H., 2006. Sensitivity of torrential rain events to the sea surface temperature based on high-resolution numerical forecasts. Journal of Geophysical Research: Atmospheres 111 (D12).

Levavasseur, G., Vrac, M., Roche, D., Paillard, D., Martin, A., Vandenberghe, J., 2011. Present and LGM permafrost from climate simulations: contribution of statistical downscaling. Climate of the Past Discussions 7, 1647–1692.

Michelangeli, P.-A., Vrac, M., Loukos, H., 2009. Probabilistic downscaling approaches: Application to wind cumulative distribution functions. Geophysical Research Letters 36 (11).

Opitz, T., 2013. Extremal t-processes: Elliptical domain of attraction and a spectral representation. Journal of Multivariate Analysis 122, 409 – 413.

Padoan, S. A., 2011. Multivariate extreme models based on underlying skew-t and skew-normal distributions. Journal of Multivariate Analysis 102, 977–991.

Panagoulia, D., Economou, P., Caroni, C., 2014. Stationary and nonstationary generalized extreme value modelling of extreme precipitation over a mountainous area under climate change. Environmetrics 25 (1), 29–43.

Piani, C., Haerter, J., Coppola, E., 2010. Statistical bias correction for daily precipitation in regional climate models over Europe. Theoretical and Applied Climatology 99 (1-2), 187–192.

Quintana-Seguí, P., Le Moigne, P., Durand, Y., Martin, E., Habets, F., Baillon, M., Canellas, C., Franchisteguy, L., Morel, S., 2008. Analysis of near-surface atmospheric variables: Validation of the safran analysis over france. Journal of applied meteorology and climatology 47 (1), 92–107.

Ribatet, M., 2013. Spatial extremes: Max-stable processes at work. Journal de la Société Francaise de Statistique 154 (2), 156–177.

Richardson, C., Barrow, E., Semenov, M., Brooks, R., 1998. Comparison of the WGEN and LARS-WG stochastic weather generators for diverse climates. Climate Research 10 (2), 95–107.

Rousseeuw, P. J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53 – 65.

Rummukainen, M., 2010. State-of-the-art with Regional Climate Models. Wiley Interdisciplinary Reviews: Climate Change 1 (1), 82–96.

Ruti, P., Somot, S., Dubois, C., Calmanti, S., Ahrens, B., Alias, A., Aznar, R., Bartholy, J., Bastin, S., Béranger, K., Brauch, J., Calvet, J.-C., Carillo, A., Decharme, B., Dell'Aquila, A., Djurdjevic, V., Drobinski, P., Elizalde-Arellano, A., Gaertner, M., Galán, P., Gallardo,

C., Giorgi, F., Gualdi, S., Harzallah, A., Herrmann, M., Jacob, D., Khodayar, S., Krichak, S., Lebeaupin, C., L'Heveder, B., Li, L., Liguro, G., Lionello, P., Onol, B., Rajkovic, B., Sannino, G., Sevault, F., 2014. MED-CORDEX initiative for mediterranean climate studies, in revision.

Schlather, M., 2002. Models for stationary max-stable random fields. Extremes 5 (1), 33–44.

Schlather, M., Tawn, J. A., 2003. A dependence measure for multivariate and spatial extreme values: Properties and inference. Biometrika 90 (1), 139–156.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, M., Duda, K. G., Huang, X. Y., Wang, W., Powers, J. G., 2008. A description of the Advanced Research WRF Version 3. Tech. rep., National Center for Atmospheric Research.

Smith, R. L., 1990. Max-stable processes and spatial extreme, unpublished.

Thibaud, E., Opitz, T., 2013. Efficient inference and simulation for elliptical pareto processes. arXiv preprint arXiv:1401.0168.

Vautard, R., Gobiet, A., Jacob, D., Belda, M., Colette, A., Déqué, M., Fernández, J., García-Díez, M., Goergen, K., Guttler, I., Halenka, T., Karacostas, T., Katragkou, E., Keuler, K., Kotlarski, S., Mayer, S., van Meijgaard, E., Nikulin, G., Patarčić, M., Scinocca, J., Sobolowski, S., Suklitsch, M., Teichmann, C., Warrach-Sagi, K., Wulfmeyer, V., Yiou, P., 2013. The simulation of european heat waves from an ensemble of regional climate models within the EURO-CORDEX project. Climate Dynamics 41 (9-10), 2555–2575.

Vidal, J.-P., Martin, E., Franchistéguy, L., Baillon, M., Soubeyroux, J.-M., 2010. A 50-year high-resolution atmospheric reanalysis over France with the Safran system. International Journal of Climatology 30 (11), 1627–1644.

von Storch, H., Langenberg, H., Feser, F., 2000. A spectral nudging technique for dynamical downscaling purposes. Monthly weather review 128 (10), 3664–3673.

Vrac, M., Drobinski, P., Merlo, A., Herrmann, M., Lavaysse, C., Li, L., Somot, S., 2012. Dynamical and statistical downscaling of the french mediterranean climate: uncertainty assessment. Natural Hazards and Earth System Science 12 (9), 2769–2784.

Vrac, M., Marbaix, P., Paillard, D., Naveau, P., 2007. Non-linear statistical downscaling of present and LGM precipitation and temperatures over europe. Climate of the Past 3 (4), 669–682.

Vrac, M., Naveau, P., 2007. Stochastic downscaling of precipitation: From dry events to heavy rainfalls. Water Resources Research 43 (7).

Vrac, M., Yiou, P., 2010. Weather regimes designed for local precipitation modeling: Application to the mediterranean basin. Journal of Geophysical Research: Atmospheres 115 (D12), 2156–2202.

Wilks, D. S., 2012. Stochastic weather generators for climate-change downscaling, part ii: multivariable and spatially coherent multisite downscaling. Wiley Interdisciplinary Reviews: Climate Change 3 (3), 267–278.

Chapitre 12

Conclusions de la partie III

Dans cette partie III, nous nous sommes efforcés d'étendre les méthodes de simulation conditionnelle présentées dans la partie II en y incorporant de l'information à grande échelle et ainsi les rendre plus adaptatives et applicables lorsqu'il n'y a plus d'observations disponibles.

Dans un premier temps, l'adaptation à cette nouvelle problématique s'est effectuée en incorporant de l'information à grande échelle dans la procédure de simulation conditionnelle (i.e. lorsqu'il y a encore des observations). Cette information, qu'on a choisi d'utiliser sous la forme de la variable SLP, est issue de sorties de modèles climatiques. Ces modèles déterministes de modélisation de la dynamique atmosphérique permettent de produire des conditions climatiques réalistes.

Différents descripteurs issus de ces variables ont donc été introduits dans la méthodologie de simulation conditionnelle de processus max-stable, en particulier dans l'étape de transformation des marginales. Cette étape peut être vue comme un moyen de gérer la non stationnarité spatiale et temporelle du phénomène modélisé, en fonction de composantes purement géographiques (latitude, longitude, altitude) et de composantes climatiques qui évoluent dans le temps (ici SLP). Néanmoins, cet enrichissement de la modélisation entraîne d'autres problèmes, comme par exemple l'augmentation du nombre de paramètres du modèle à estimer et donc plus de difficultés à obtenir de bonnes estimations. Dans la pratique, on observe que ces nouvelles versions de l'algorithme de simulation conditionnelle améliorent légèrement les scores finaux obtenus par rapport aux modèles stationnaires en temps.

L'étape suivante a été de réutiliser ces modèles incorporant l'information SLP, mais dans une procédure de simulation non conditionnelle (i.e. sans observations). En effet, dans un objectif de projections futures, seule cette version non conditionnelle de la procédure est applicable. A noter que les variables climatiques restent disponibles dans le futur. Des conclusions similaires ont été observées, avec un léger gain en termes de

qualité des simulations pour les versions de la procédure de simulation non conditionnelle utilisant l'information SLP. De plus, on peut supposer ce type de modèle plus capable de s'adapter aux conditions futures et il est donc possible que l'écart entre les versions avec et sans information SLP se creuse au fur et à mesure que l'on s'éloigne des années de calibration du modèle.

Par la suite, la prise en compte d'information à grande échelle a été effectuée en utilisant des méthodes de downscaling statistique. Ces techniques variées consistant à établir des liens statistiques entre deux variables avec des résolutions spatiales différentes permettent d'obtenir des valeurs locales à partir d'information à grande échelle. Néanmoins, peu de techniques permettent de générer des valeurs en différents points de l'espace avec une cohérence spatiale, qui plus est lorsque l'on cherche à modéliser des extrêmes (e.g. Frost, 2007; Chandler et al., 2011; Ambrosino, 2012)

Pour pallier à ces problèmes, une méthode hybride appelée « Spatial Hybrid Downscaling » (SHD) a été proposée et vise à combiner les techniques de downscaling statistique avec l'algorithme de simulation conditionnelle de processus max-stables. L'idée est d'utiliser le downscaling pour obtenir des valeurs en certains points, puis d'utiliser ces projections comme valeurs conditionnantes de l'algorithme de simulation conditionnelle. En procédant de la sorte, il est possible de construire des champs de maxima spatialement cohérents uniquement à partir de sorties de modèles climatiques et d'avoir des séries de valeurs en tout point de la carte. De plus, cette approche permet de tirer avantage de la cohérence physique des sorties de modèles climatiques et des propriétés statistiques spatiales de la simulation conditionnelle.

L'application sur données réelles montre que cette approche fournit de bons résultats et que cela reste le cas pour les différentes versions du modèle qui ont été testées (choix des points conditionnants et des techniques de downscaling). Cette robustesse de l'approche est un point très important, en particulier pour le choix des points conditionnants puisque ce dernier dépend en grande partie des données disponibles. Ainsi, dans la pratique, il peut arriver que ce choix soit grandement contraint. Néanmoins, il est apparu qu'un nombre de points conditionnants plutôt parcimonieux pouvait être un bon choix. Dans ce cas de figure, l'utilisation de l'algorithme des feuilles mortes (chapitre 7) plutôt que l'échantillonneur de Gibbs est envisageable pour accélérer la procédure.

Pour résumer, cette méthode hybride permet donc de générer des séries de valeurs ponctuelles à partir de sorties de modèles climatiques. Ainsi, on est capable de générer des séries de valeurs locales issues de contraintes à grande échelle dictées par différents scénarios climatiques. Dans le cadre du changement climatique, ces séries à haute résolution obtenues peuvent être utilisées en entrée de nombreux modèles d'impact dans le but d'anticiper les conséquences de futures précipitations extrêmes.

Quatrième partie

Conclusions et perspectives

Chapitre 13

Conclusions

Lors de ces travaux de thèse, nous nous sommes particulièrement intéressés à la modélisation de précipitations extrêmes et aux méthodes permettant de fournir des champs spatialisés d'extrêmes à partir de différentes sources d'information.

Dans un premier temps, nous nous sommes placés dans un contexte où, ayant des observations de maxima annuels à certains points de la carte, nous voulons obtenir des distributions de probabilité en des emplacements vides d'observation. Nous nous sommes intéressés aux outils permettant de générer des champs de maxima tout en respectant les contraintes émanant des observations, c'est-à-dire aux outils permettant de faire de la simulation conditionnelle de processus max-stables. En se basant sur la méthodologie générale développée par Dombry et al. (2013), nous avons étendu son cadre d'utilisation. En effet, nous avons vu chapitre 6 que cet algorithme complexe en trois étapes pouvait être adapté à un processus max-stable intéressant par sa flexibilité : le processus *t* extrêmal. Ce processus est une généralisation du processus de Schlather avec un paramètre supplémentaire α ayant un impact sur le coefficient extrêmal du processus, plus particulièrement permettant d'obtenir une gamme de valeurs plus large du coefficient extrêmal. Pouvoir estimer ce paramètre à partir des données permet ainsi d'obtenir un processus dont la dépendance des extrêmes est proche de celle des observations.

De plus, nous avons proposé une modification de la seconde étape de l'algorithme de simulation conditionnelle permettant désormais un plus grand nombre de points conditionnants. En effet, en utilisant un échantillonneur de Gibbs à la place de l'algorithme d'acceptation/rejet initialement proposé, le temps de calcul nécessaire augmente mécaniquement avec le nombre de points conditionnants, mais l'algorithme reste cependant applicable, même pour un nombre de points élevé (> 100).

Par la suite, une étude sur données simulées a permis d'établir pour quelles valeurs de α, il est pertinent d'utiliser l'algorithme de simulation conditionnelle. Cette étude a également mis en avant l'importance de la première étape de l'algorithme. Cette étape, qui est la plus coûteuse d'un point de vue algorithmique, consiste à répartir les

points conditionnants en partitions selon une loi cible. En comparant l'efficacité de l'algorithme avec une partition aléatoire ou la vraie partition (dans un cas où elle est disponible facilement), nous avons mis en avant le rôle capital de cette étape. De plus, il est également apparu que le fait d'utiliser un échantillonneur de Gibbs pour le tirage des partitions était efficace mais loin d'être optimal à cause de la vitesse de convergence très faible de cet algorithme.

Une application sur données réelles a permis de comparer l'algorithme par simulations conditionnelles avec d'autres approches. Les résultats obtenus, au vu de critères univariés et multivariés sont globalement meilleurs. On constate cependant que certaines zones de la région sont mal modélisées par notre méthode, en particulier celles en altitude. C'est une variable qui n'entre en effet pas en compte dans la modélisation des marginales du processus, ce qui peut expliquer cette difficulté.

Dans le chapitre 7, nous nous sommes focalisés sur la première étape de l'algorithme de simulation conditionnelle, en proposant une approche par algorithme des feuilles mortes à la place de l'échantillonneur de Gibbs pour la simulation des partitions. Cet algorithme a l'avantage d'être une méthode de simulation exacte, alors que l'échantillonneur de Gibbs est une approche par MCMC. Il apparaît ainsi que pour un nombre de points conditionnants qui reste relativement faible (< 15), cette nouvelle approche est performante, que ça soit en terme de qualité des simulations ou en temps de calcul et offre une très bonne alternative à l'échantillonneur de Gibbs. Cependant, pour un nombre de points conditionnants supérieur, l'algorithme des feuilles mortes devient inutilisable.

Ainsi, cette procédure de simulations conditionnelles est applicable à des cas concrets, tout en proposant certaines alternatives d'utilisation en fonction des situations qui se présentent.

Dans un second temps, nous nous sommes intéressés à l'ajout d'information à grande échelle dans la modélisation. En particulier, chapitre 10, l'étape de transformation des marginales de l'algorithme de simulation conditionnelle permet d'incorporer de l'information climatique (ici la variable SLP). Ainsi, en plus des variables latitude et longitude qui permettent de rendre compte de la non stationnarité spatiale du phénomène étudié, l'ajout de SLP donne la possibilité de modéliser la non stationnarité temporelle. Cette modification qui ne change pas l'algorithme en tant que tel, permet donc de pouvoir faire de la simulation conditionnelle non stationnaire spatialement et temporellement. Des applications ont été effectuées en proposant deux types de transformation de marginale. La première incorpore directement un résumé de la variable SLP dans la modélisation alors que la deuxième utilise cette information comme variable de classification pour des régimes automnaux homogènes. Chaque régime entraîne une transformation différente des marginales.

L'application de ces algorithmes de simulation conditionnelle non stationnaire sur

des données réelles montre que l'ajout de la non stationnarité temporelle sous cette forme permet d'améliorer légèrement les résultats, par rapport aux méthodes qui n'incorporent pas la variable SLP. Dans la configuration de nos tests, il est possible que les valeurs conditionnantes aient un trop grand impact dans la qualité des simulations conditionnelles pour que ces transformations aient un poids.

Par la suite, nous nous sommes placés dans un cadre de projections futures, c'est-à-dire lorsqu'il n'y a plus d'observation disponible mais uniquement de l'information climatique à grande échelle. Des processus max-stables, dont les paramètres sont estimés sur les observations passées, peuvent être simulés mais cette fois-ci non conditionnellement. Cependant, comme précédemment, la non stationnarité temporelle peut être atteinte par l'ajout de la variable SLP dans la modélisation des marginales. Les résultats montrent que cet ajout d'information peut dans certain cas améliorer légèrement la qualité des simulations non conditionnelles. Les performances peuvent sans doute être améliorées en utilisant plus de variables à grande échelle dans la modélisation et en faisant une étude de sélection de modèles. Cependant, cette approche a l'avantage d'être applicable pour générer des projections futures non stationnaires de maxima de précipitations.

Dans le chapitre 11, nous avons eu l'idée de combiner l'algorithme de simulation conditionnelle avec une étape de downscaling statistique de sorties de RCM. Cette méthode consiste à générer des valeurs locales par downscaling statistique des sorties de RCM, puis de les utiliser comme valeurs conditionnantes pour l'algorithme de simulation conditionnelle. Ainsi, cette approche SHD (Spatial Hybrid Downscaling) utilise les avantages statistiques de la simulation conditionnelle, la flexibilité des techniques de downscaling statistique et la cohérence physique des RCMs.

L'application a été effectuée sur les données réelles en utilisant les sorties de précipitations des RCM. Nous avons comparé différentes versions de cette approche hybride, notamment en modifiant les techniques de downscaling utilisées ou alors les emplacements et le nombre de points conditionnants choisis. Pour la majeure partie des paramétrisations testées, l'approche SHD présente de meilleurs résultats que des techniques de downscaling statistiques usuelles. De plus, même si les méthodes de downscaling statistiques sont relativement simples, les résultats obtenus ne sont pas très éloignés de ceux du modèle optimal qui correspond au cas irréaliste où la fonction de downscaling statistique est parfaite. En perfectionnant ce lien reliant les deux échelles, il peut être possible d'améliorer les performances de l'algorithme, mais la qualité actuelle des simulations obtenues est encourageante.

En outre, cette qualité est dépendante de la qualité du RCM utilisé. Ainsi, en testant avec un autre modèle de downscaling dynamique, les résultats diffèrent naturellement mais l'ordre entre les différentes méthodes testées est inchangé, signe de la robustesse de l'approche.

Dans une optique de transmission des connaissances et pour rendre ces méthodes accessibles à un grand nombre d'utilisateurs, il est également important de noter qu'un package R (`MCSIM`) a été développé et regroupe tous les codes nécessaires pour utiliser les méthodes techniques présentées tout au long de ce travail. Ce package devrait être prochainement rendu disponible sur le site du CRAN[1] et sur demande.

Pour conclure, les différents objectifs ont été remplis en fournissant à la fois des méthodes de simulations conditionnelles flexibles incorporant de la non stationnarité spatiale et temporelle, et par la suite la possibilité d'utiliser uniquement des sorties de modèles climatiques pour générer des champs spatialisés de précipitations extrêmes dans un cadre de projections sous contraintes de changement climatique (passées ou futures).

[1] http ://cran.r-project.org/

Chapitre 14

Perspectives

Malgré tous ces points positifs et encourageants, de nombreuses pistes restent à explorer, que cela soit pour améliorer les techniques déjà développées, élargir leurs possibilités d'utilisation ou bien pour les tester dans certaines situations.

Du point de vue technique, quelques améliorations sont possibles, notamment pour l'étape de tirage des partitions dans l'algorithme de simulation conditionnelle présenté chapitre 6. Par exemple, il serait intéressant de voir dans quelles mesures l'algorithme des feuilles mortes présenté et développé chapitre 7 peut être perfectionné pour être utilisable pour un plus grand nombre de points conditionnants. Lorsque ce nombre de points augmente, la probabilité de tirer une partition avec un poids suffisamment élevé chute très rapidement.

Cette partition doit être tirée de manière aléatoire et équiprobable dans l'ensemble des partitions. Comme en pratique on constate seulement un faible nombre de partitions de poids non négligeable, peut-être serait-il judicieux de proposer une version approchée de cet algorithme qui ciblerait particulièrement ces partitions intéressantes. Nous pouvons imaginer un tirage des partitions qui garderait en mémoire les partitions ayant des poids trop faibles pour éviter de retomber dessus dans la suite. Ou pour aller plus loin dans cette optique d'algorithme des feuilles mortes à mémoire, nous pouvons même envisager un algorithme qui combinerait la méthodologie classique des feuilles mortes avec une étape de transition d'une partition à l'autre. Il serait intéressant de voir dans quelles mesures ce type d'algorithme devient utilisable en termes de temps de calcul et en quoi ses performances en termes de qualité des simulations sont détériorées.

Une autre idée vient du constat que l'approche par feuilles mortes est très performante lorsque le nombre de points conditionnants est peu élevé. Ainsi, nous pouvons imaginer considérer des sous-ensembles des points conditionnants sur lesquels des algorithmes peuvent être utilisés très rapidement. Il faudrait ensuite trouver un moyen de construire la partition globale (celle qui nous intéresse) par méta-analyse à partir de toutes les sous-partitions obtenues. Du travail devrait également être fait dans le choix des sous-

ensembles de points pour éviter que des problèmes de combinatoire n'apparaissent.

Pour l'heure, la méthode permettant d'obtenir le plus rapidement un tirage de partitions cohérent est la méthode par augmentation de données proposée par Lantuéjoul and Bel (2014). Néanmoins, en l'état c'est une méthode approchée qui n'est pas aussi performante qu'un échantillonneur de Gibbs, tout en fournissant des résultats très intéressants en beaucoup moins de temps de calcul. Il serait cependant nécessaire de regarder plus en détails l'éventuelle perte d'efficacité en termes de qualité des simulations obtenues engendrée par l'utilisation de cette méthode.

En ce qui concerne les considérations techniques relatives uniquement à l'approche SHD, des démarches peuvent être menées pour améliorer l'étape du choix des points conditionnants. Si différentes stratégies ont été proposées chapitre 11, aucune n'est vraiment apparue comme étant le meilleur choix. Une méthode qui semblait prometteuse consistait à choisir des représentants des points de l'échelle locale par clustering, indépendamment des RCMs, n'a pas donné les résultats escomptés. Une piste serait de remplacer l'algorithme de clustering utilisé (PAM) par un algorithme qui empêcherait que certains points atypiques se retrouvent seuls dans une classe (e.g. Super-Paramagnetic Clustering Algorithm ; Blatt et al., 1996).

Parallèlement, des efforts ont été effectués pour rendre l'algorithme de simulation conditionnelle applicable à un plus grand nombre de points conditionnants. Cependant, de cette démarche naissent (ou s'amplifient) les difficultés d'estimation des paramètres du modèle pouvant être quelques fois nombreux. Ainsi, à partir de la partie III de ce document, nous avons choisi d'utiliser une approche de maximisation de vraisemblance complète proposée par Thibaud and Opitz (2013) à la place de la maximisation de vraisemblance composite utilisée usuellement dans ce genre de problématiques. Les performances de cette approche semblent meilleures lorsque le nombre d'observations augmente, notamment en temps de calcul. Ces approches sont néanmoins complexes et très coûteuses en temps de calcul, et doivent être effectuées plusieurs fois car elles semblent être assez sensibles à l'initialisation des paramètres à estimer. Des méthodes d'optimisation plus complexes devraient être incluses à ce travail pour rendre cette étape plus performante.

Tout au long de ce travail, la gestion de la non stationnarité du processus a été un point central de toutes les démarches et elle peut encore être améliorée. Dans les différentes approches, nous avons tenu compte de la non stationnarité spatiale grâce à l'incorporation de la latitude et de la longitude dans l'étape de transformation des marginales. Si cette démarche est un premier pas dans la non stationnarité spatiale, elle ne permet cependant pas de modéliser autre chose qu'un gradient spatial linéaire. Tout autre schéma de non stationnarité ne peut pas être capturé par cette méthode. Une amélioration possible est de

considérer une fonction des coordonnées qui ne soit pas forcément linéaire.

En restant dans la transformation des marginales pour obtenir la non stationnarité spatiale, des efforts doivent être menés pour incorporer la variable altitude dans la modélisation. En effet, les résultats de l'étude du chapitre 6 laissent penser que l'ajout de cette information est utile et nécessaire pour parvenir à bien modéliser les zones de forte altitude présentant des caractéristiques très différentes des autres régions. Néanmoins les essais réalisés chapitre 10 ne se sont pas avérés concluants, la faute sans doute à la grande variabilité de la variable altitude et au peu de points en altitude dans notre zone d'étude. Une solution pourrait être de transformer la variable altitude en une variable catégorielle qui indiquerait si le point en question se trouve ou non au dessus d'un certain seuil (voire une variable à plusieurs facteurs). Malgré la perte d'information, cette démarche pourrait faciliter la procédure d'estimation des paramètres et ainsi la qualité des simulations dans ces zones en altitude.

Parallèlement, l'ajout de covariables à grande échelle qui évoluent dans le temps (SLP) a permis, chapitre 10, d'inclure un côté évolutif aux simulations conditionnelles et non conditionnelles en incorporant de la non stationnarité temporelle. Sur ce point, de nombreuses améliorations restent également possibles. Comme dans ce travail l'important n'était pas d'obtenir le meilleur modèle possible mais uniquement de chercher à savoir si une telle démarche pouvait se révéler utile, aucun autre prédicteur n'a été testé. En effet, même si le choix de la variable SLP semble être pertinent, d'autres prédicteurs pourraient être informatifs comme par exemple la température ou encore les hauteurs de géopotentiel à différent niveaux de pression. Des études de sensibilité ou de sélection de modèles pourraient être menées et conduiraient sans doute à une amélioration de la qualité des simulations. Cependant, dans le chapitre 11, l'approche SHD présentée n'utilise pas cette information à grande échelle dans la transformation des marginales. Or, il est possible que la relation entre les sorties de modèles RCMs et les observations ponctuelles soient amenées à évoluer dans le temps. Ainsi, l'ajout d'une composante de non stationnarité temporelle pourrait enrichir l'approche SHD.

En outre, il faut remarquer que, même si elles incorporent de la non stationnarité spatiale et temporelle, ces transformations de marginales ne modifient pas fondamentalement la structure de dépendance du processus. Si cette hypothèse conservative est raisonnable, il est toutefois possible que cette structure de dépendance puisse se modifier dans les décennies à venir. Il peut donc être nécessaire d'inclure cette non stationnarité directement dans la structure de dépendance du processus, par exemple dans la fonction de corrélation du processus Gaussien qui entre dans la construction du processus t extrêmal (équation (3.13)), sous la forme d'une variable à grande échelle (e.g., Paciorek and Schervish, 2006). Ce paramètre supplémentaire peut être estimé simultanément aux autres dans la maximisation de la vraisemblance.

Pour conclure ce paragraphe traitant de la gestion de la non stationnarité, il est

important de remarquer que dans ce travail, cette non stationnarité est atteinte par transformation des marginales, tout en simulant par la suite un processus stationnaire. Il serait donc intéressant de comparer ces résultats avec ceux obtenus en simulant directement un processus non stationnaire, reposant par exemple sur des techniques non paramétriques.

En ce qui concerne l'application de l'algorithme SHD, de nombreuses utilisations sont possibles. Dans un premier temps, il serait intéressant de tester d'autres techniques de downscaling statistique univarié en plus de la régression linéaire et de CDF-t. Même si les résultats sont satisfaisants avec ces techniques plutôt simples, il se peut que des améliorations puissent être obtenues en complexifiant cette relation entre les variables haute et basse résolution.

D'autre part, il est utile de remarquer que toutes les démarches de downscaling statistique univarié considérées dans ce travail ne sont pas spécifiques à la descente d'échelle d'extrêmes. Ainsi, en considérant des techniques de downscaling extrême (e.g. la méthode XCDF-t ; Kallache et al., 2011), il est peut-être possible d'améliorer la qualité des simulations. D'autre part, en considérant un processus non extrême (par exemple un processus Gaussien) couplé à une procédure de simulation conditionnelle plus classique, il est donc tout à fait possible d'adapter l'approche SHD pour simuler des réalisations non extrêmes. Pour rester dans notre problématique d'événements exceptionnels, nous pouvons même envisager d'obtenir des réalisations journalières non extrêmes par cette approche SHD, puis de reconstruire le maximum automnal à partir de celles-ci. Même si cette approche devrait avoir plus de difficultés à simuler des valeurs de précipitations très élevées, la comparaison serait intéressante.

De plus, toutes ces techniques de downscaling statistique univarié restent empiriques, dans le sens où nous utilisons des méthodes statistiques classiques pour établir un lien entre les deux échelles climatiques, sans réel cadre théorique. Il serait intéressant de considérer un downscaling « théorique » qui consisterait à établir la loi conditionnelle du processus spatial sachant une ou plusieurs conditions (de type maximum ou moyenne) sur une région qui pourrait être un point de grille d'un modèle à plus grande échelle. Oesting (2014) a obtenu des résultats dans ce sens pour quelques processus particuliers.

Par la suite, il serait pertinent de regarder la perte d'efficacité (si perte il y a) de la méthode SHD lorsque l'on utilise des GCMs à la place des RCMs pour fournir l'information à grande échelle. Il faudrait voir si la diminution du nombre de points de grille (d'une dizaine pour les RCMs à un seul pour les GCMs) de la variable à grande échelle utilisée pour expliquer les observations détériore la qualité des simulations.

Pour finir, l'application majeure qui reste à développer consiste à effectuer des projections futures des précipitations grâce à l'approche SHD. En utilisant l'information fournie par les différents scénarios à grande échelle disponibles, il serait très intéressant de voir les conséquences locales en termes de précipitations extrêmes de chacun de ces scénarios. De plus, il est désormais possible, en comparant les distributions obtenues à

partir de divers scénarios d'avenir possibles, d'évaluer les différences de conséquences locales d'un scénario à l'autre et d'évaluer la significativité de ces différences.

Bibliographie

Alabert, F., Massonnat, G., et al., 1990. Heterogeneity in a complex turbiditic reservoir : stochastic modelling of facies and petrophysical variability. In : SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers.

Ambrosino, C., 2012. STATISTICAL DOWNSCALING WITH GLIMCLIM :Generalised LInear Modelling of daily CLIMate sequences.

Barnett, T., Preisendorfer, R., 1978. Multifield analog prediction of short-term climate fluctuations using a climate state vector. Journal of Atmospheric Sciences 35, 1771–1787.

Baron, C., Sultan, B., Balme, M., Sarr, B., Traore, S., Lebel, T., Janicot, S., Dingkuhn, M., 2005. From GCM grid cell to agricultural plot : scale issues affecting modelling of climate impact. Philosophical Transactions of the Royal Society B : Biological Sciences 360 (1463), 2095–2108.

Beaulant, A.-L., Joly, B., Nuissier, O., Somot, S., Ducrocq, V., Joly, A., Sevault, F., Deque, M., Ricard, D., 2011. Statistico-dynamical downscaling for mediterranean heavy precipitation. Quarterly Journal of the Royal Meteorological Society 137 (656), 736–748.

Bechler, A., Bel, L., Vrac, M., 2014. Conditional simulations of extremal t process : Application to fields of extreme precipitation, under review, http ://hal-agroparistech.archives-ouvertes.fr/hal-01053605.

Beirlant, J., Goegebeur, Y., Teugels, J., Segers, J., Waal, D. D., Ferro, C., 2004. Statistics of extremes. Theory and applications. Hoboken, NJ : John Wiley & Sons.

Bel, L., Bacro, J. N., Lantuéjoul, C., 2008. Assessing extremal dependence of environmental spatial fields. Environmetrics 19 (2), 163–182.

Bellone, E., Hughes, J. P., Guttorp, P., 2000. A hidden markov model for downscaling synoptic atmospheric patterns to precipitation amounts. Climate research 15 (1), 1–12.

Bernard, E., Naveau, P., Vrac, M., Mestre, O., Oct. 2013. Clustering of Maxima : Spatial Dependencies among Heavy Rainfall in France. Journal of Climate 26, 7929–7937.

Biau, G., Zorita, E., von Storch, H., Wackernagel, H., 1999. Estimation of precipitation by kriging in the eof space of the sea level pressure field. Journal of Climate 12 (4), 1070–1085.

Blanchet, J., Davison, A. C., 2011. Spatial modeling of extreme snow depth. The Annals of Applied Statistics 5 (3), 1699–1725.

Blatt, M., Wiseman, S., Domany, E., Apr 1996. Superparamagnetic clustering of data. Phys. Rev. Lett. 76, 3251–3254.

Boé, J., Terray, L., Habets, F., Martin, E., 2006. A simple statistical-dynamical downscaling scheme based on weather types and conditional resampling. Journal of Geophysical Research : Atmospheres 111 (D23).

Brier, G. W., 1950. Verification of forecasts expressed in terms of probability. Monthly weather review 78 (1), 1–3.

Bröcker, J., 2012. Evaluating raw ensembles with the continuous ranked probability score. Quarterly Journal of the Royal Meteorological Society 138 (667), 1611–1617.

Brown, B. M., Resnick, S. I., 1977. Extreme values of independent stochastic processes. Journal of Applied Probability 14 (4), 732–739.

Buishand, T., de Haan, L., Zhou, C., 2008. On spatial extremes : with application to a rainfall problem. The Annals of Applied Statistics, 624–642.

Busuioc, A., Tomozeiu, R., Cacciamani, C., 2008. Statistical downscaling model based on canonical correlation analysis for winter extreme precipitation events in the emilia-romagna region. International Journal of Climatology 28 (4), 449–464.

Caflisch, R. E., 1 1998. Monte carlo and quasi-monte carlo methods. Acta Numerica 7, 1–49.

Candille, G., Talagrand, O., 2005. Evaluation of probabilistic prediction systems for a scalar variable. Quarterly Journal of the Royal Meteorological Society 131 (609), 2131–2150.

Carreau, J., Neppel, L., Arnaud, P., Cantet, P., 2013. Extreme rainfall analysis at ungauged sites in the south of france : Comparison of three approaches. Journal de la Société Française de Statistique 154 (2), 119–138.

Carreau, J., Vrac, M., 2011. Stochastic downscaling of precipitation with neural network conditional mixture models. Water Resources Research 47 (10).

Casella, G., George, E. I., 1992. Explaining the Gibbs sampler. The American Statistician 46 (3), 167–174.

Chandler, R. E., Bates, B. C., Charles, S. P., 2011. Rainfall trends in southwest western australia. Statistical Methods for Trend Detection and Analysis in the Environmental Sciences, 283–306.

Chilès, J.-P., Delfiner, P., 1999. Geostatistics : Modeling Spatial Uncertainty. Wiley, New-York.

Coles, S., Tawn, J., 1994. Statistical methods for multivariate extremes - an application to structural design. Journal of the Royal Statistical Society - Series C : Applied Statistics 43 (1), 1–48.

Cooley, D., Naveau, P., Poncet, P., 2006. Variograms for spatial max-stable random fields. In : Dependence in probability and statistics. Springer, pp. 373–390.

Davison, A. C., Gholamrezaee, M. M., 2012. Geostatistics of extremes. Proceedings of the Royal Society A : Mathematical, Physical and Engineering Science 468 (2138), 581–608.

Davison, A. C., Padoan, S. A., Ribatet, M., 05 2012. Statistical modeling of spatial extremes. Statistical Science 27 (2), 161–186.

De Haan, L., 1984. A spectral representation for max-stable processes. The Annals of Probability 12 (4), 1194–1204.

De Haan, L., De Ronde, J., 1998. Sea and wind : multivariate extremes at work. Extremes 1 (1), 7–45.

Dee, D., Balmaseda, M., Balsamo, G., Engelen, R., Simmons, A., Thépaut, J.-N., 2013. Toward a consistent reanalysis of the climate system. Bulletin of the American Meteorological Society.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kallberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., Vitart, F., 2011. The ERA-interim reanalysis : configuration and performance of the data assimilation system. Quarterly Journal of the Royal Meteorological Society 137 (656), 553–597.

Déqué, M., 2007. Frequency of precipitation and temperature extremes over France in an anthropogenic scenario : model results and statistical correction according to observed values. Global and Planetary Change 57 (1), 16 – 26, Extreme Climatic Events.

Dombry, C., Eyi-Minko, F., 2013. Regular conditional distributions of continuous max-infinitely divisible random fields. Electron. J. Probab 18 (7), 1–21.

Dombry, C., Éyi-Minko, F., Ribatet, M., 2013. Conditional simulation of max-stable processes. Biometrika 100 (1), 111–124.

Ducrocq, V., Nuissier, O., Ricard, D., Lebeaupin, C., Thouvenin, T., 2008. A numerical study of three catastrophic precipitating events over southern france. ii : Mesoscale triggering and stationarity factors. Quarterly Journal of the Royal Meteorological Society 134 (630), 131–145.

Eden, J., Widmann, M., Maraun, D., Vrac, M., 2014. Comparison of GCM- and RCM-simulated precipitation following stochastic postprocessing. Journal of Geophysical Research - Atmospheres.To appear.

Epstein, E. S., 1969. A scoring system for probability forecasts of ranked categories. Journal of Applied Meteorology 8 (6), 985–987.

Fawcett, L., Walshaw, D., 2012. Estimating return levels from serially dependent extremes. Environmetrics 23 (3), 272–283.

Fernández-Ferrero, A., Sáenz, J., Ibarra-Berastegi, G., Fernández, J., 2009. Evaluation of statistical downscaling in short range precipitation forecasting. Atmospheric Research 94 (3), 448–461.

Flaounas, E., Drobinski, P., Vrac, M., Bastin, S., Lebeaupin-Brossier, C., Stéfanon, M., Borga, M., Calvet, J.-C., 2013. Precipitation and temperature space–time variability and extremes in the mediterranean region : evaluation of dynamical and statistical downscaling methods. Climate dynamics 40 (11-12), 2687–2705.

Fowler, H. J., Cooley, D., Sain, S. R., Thurston, M., 2010. Detecting change in UK extreme precipitation using results from the climateprediction. net BBC climate change experiment. Extremes 13 (2), 241–267.

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

Friederichs, P., Hense, A., 2007. Statistical downscaling of extreme precipitation events using censored quantile regression. Monthly Weather Review 135, 2365–2378.

Friederichs, P., Thorarinsdottir, T. L., 2012. Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. Environmetrics 23 (7), 579–594.

Frost, A., 2007. Australian application of a statistical downscaling technique for multisite daily rainfall : Glimclim. In : MODSIM 2007 : International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, pp. 553–559.

Fuentes, M., Henry, J., Reich, B., 2013. Nonparametric spatial models for extremes : application to extreme temperature data. Extremes 16 (1), 75–101.

Gaume, J., Eckert, N., Chambon, G., Naaim, M., Bel, L., 2013. Mapping extreme snowfalls in the french alps using max-stable processes. Water Resources Research 49 (2), 1079–1098.

Geweke, J., 1991. Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities. In : Computing science and statistics : Proceedings of the 23rd symposium on the interface. Citeseer, pp. 571–578.

Ghosh, S., Mujumdar, P., 2008. Statistical downscaling of GCM simulations to streamflow using relevance vector machine. Advances in Water Resources 31 (1), 132 – 146.

Gneiting, T., Raftery, A. E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102 (477), 359–378.

Gneiting, T., Stanberry, L., Grimit, E., Held, L., Johnson, N., 2008. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. Test 17 (2), 211–235.

Gobiet, A., Jacob, D., Euro-Cordex Community, Apr. 2012. A new generation of regional climate simulations for Europe : The EURO-CORDEX Initiative. In : Abbasi, A., Giesen, N. (Eds.), EGU General Assembly Conference Abstracts. Vol. 14 of EGU General Assembly Conference Abstracts. p. 8211.

Goodess, C., Jones, P., 2002. Links between circulation and changes in the characteristics of iberian rainfall. International Journal of Climatology 22 (13), 1593–1615.

Gordon, A., 1999. Classification. 1999. Chapman&Hall, CRC, Boca Raton, FL.

Gudmundsson, L., Bremnes, J., Haugen, J., Engen-Skaugen, T., 2012. Technical note : Downscaling rcm precipitation to the station scale using statistical transformations–a comparison of methods. Hydrology and Earth System Sciences 16 (9), 3383–3390.

Hersbach, H., Oct. 2000. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. Weather and Forecasting 15 (5), 559–570.

Hingray, B., Mezghani, A., 2007. Utilisation des réanalyses NCEP pour la génération de scénarios météorologiques. application pour la génération de scénarios de crues pour le rhône à l'amont du Leman. La Houille Blanche 6, 104–110.

Hundecha, Y., Pahlow, M., Schumann, A., 2009. Modeling of daily precipitation at multiple locations using a mixture of distributions to characterize the extremes. Water resources research 45 (12).

Huth, R., 2001. Disaggregating climatic trends by classification of circulation patterns. International Journal of Climatology 21 (2), 135–153.

Huth, R., 2002. Statistical downscaling of daily temperature in central Europe. Journal of Climate 15 (13), 1731–1742.

Huth, R., Kliegrova, S., Metelka, L., 2008. Non-linearity in statistical downscaling : does it bring an improvement for daily temperature in Europe ? International Journal of Climatology 28 (4), 465–477.

IPCC, 2007. Contribution of working group i to the fourth assessment report of the intergovernmental panel on climate change. Tech. rep., Intergovernmental Panel of Climate Change (IPCC).

IPCC, 2011. Managing the risks of extreme events and disasters to advance climate change adaptation. Tech. rep., Intergovernmental Panel of Climate Change (IPCC).

IPCC, 2013. Contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change. Tech. rep., Intergovernmental Panel of Climate Change (IPCC).

Jeulin, D., 1979. Morphologie mathématique et propriétés physiques des agglomérés de minerais de fer et du coke métallurgique. Ph.D. thesis, Ecole des Mines de Paris.

Jonathan, P., Ewans, K., Randell, D., 2014. Non-stationary conditional extremes of northern North Sea storm characteristics. Environmetrics 25 (3), 172–188.

Joussaume, S., Taylor, K., Braconnot, P., Mitchell, J., Kutzbach, J., Harrison, S., Prentice, I., Broccoli, A., Abe-Ouchi, A., Bartlein, P., et al., 1999. Monsoon changes for 6000 years ago : results of 18 simulations from the paleoclimate modeling intercomparison project (pmip). Geophysical Research Letters 26 (7), 859–862.

Kabluchko, Z., Schlather, M., de Haan, L., 09 2009. Stationary max-stable fields associated to negative definite functions. The Annals of Probability 37 (5), 2042–2065.

Kallache, M., Vrac, M., Naveau, P., Michelangeli, P.-A., 2011. Non-stationary probabilistic downscaling of extreme precipitation. Journal of Geophysical Research : Atmospheres 116 (D5).

Kalnay, E., Dec. 2003. Atmospheric Modeling, Data Assimilation and Predictability, 1st Edition. Cambridge University Press.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., et al., 1996. The ncep/ncar 40-year reanalysis project. Bulletin of the American meteorological Society 77 (3), 437–471.

Kendall, W. S., Thönnes, E., 1999. Perfect simulation in stochastic geometry. Pattern Recognition 32 (9), 1569 – 1586.

Koenker, R., 2000. Quantile regression. International Encyclopedia of the Social Sciences.

Lantuéjoul, C., 2002. Geostatistical simulation : models and algorithms. Springer.

Lantuéjoul, C., Bel, L., 2014. Simulation conditionnelle du processus de schlather. In : 46èmes Journées de Statistique de la SFdS.

Laprise, R., 2008. Regional climate modelling. Journal of Computational Physics 227 (7), 3641–3666.

Lebeaupin, C., Ducrocq, V., Giordani, H., 2006. Sensitivity of torrential rain events to the sea surface temperature based on high-resolution numerical forecasts. Journal of Geophysical Research : Atmospheres 111 (D12).

Levavasseur, G., Vrac, M., Roche, D., Paillard, D., Martin, A., Vandenberghe, J., 2011. Present and LGM permafrost from climate simulations : contribution of statistical downscaling. Climate of the Past Discussions 7, 1647–1692.

Lindsay, B. G., 1988. Composite likelihood methods. Contemporary Mathematics 80 (1), 221–39.

Mannshardt-Shamseldin, E. C., Smith, R. L., Sain, S. R., Mearns, L. O., Cooley, D., et al., 2010. Downscaling extremes : A comparison of extreme value distributions in point-source and gridded precipitation data. The Annals of Applied Statistics 4 (1), 484–502.

Mantoglou, A., Wilson, J. L., 1982. The turning bands method for simulation of random fields using line generation by a spectral method. Water Resources Research 18 (5), 1379–1394.

Maraun, D., Wetterhall, F., Ireson, A., Chandler, R., Kendon, E., Widmann, M., Brienen, S., Rust, H., Sauter, T., Themeßl, M., et al., 2010. Precipitation downscaling under climate change : recent developments to bridge the gap between dynamical models and the end user. Reviews of Geophysics 48 (3).

Mardia, K. V., Kent, J. T., Bibby, J. M., 1979. Multivariate analysis. Academic press.

Mariethoz, G., Renard, P., Cornaton, F., Jaquet, O., 2009. Truncated plurigaussian simulations to characterize aquifer heterogeneity. Ground Water 47 (1), 13–24.

Matheron, G., 1968. Schéma booléen séquentiel de partition aléatoire. Paris School of Mines publications.

Michelangeli, P.-A., Vrac, M., Loukos, H., 2009. Probabilistic downscaling approaches : Application to wind cumulative distribution functions. Geophysical Research Letters 36 (11).

Naveau, P., de Fondeville, R., Cooley, D., 2014. Inference the continuous ranked probability score for finite samples. In : Summer School on Weather Extremes.

Nicholas, R. E., Battisti, D. S., 2012. Empirical downscaling of high-resolution regional precipitation from large-scale reanalysis fields. Journal of Applied Meteorology and Climatology 51 (1), 100–114.

Oesting, M., 2014. On the distribution of a max-stable process conditional on max-linear functionals, submitted.

Oesting, M., Kabluchko, Z., Schlather, M., 2012. Simulation of brown–resnick processes. Extremes 15 (1), 89–107.

Oesting, M., Schlather, M., 2014. Conditional sampling for max-stable processes with a mixed moving maxima representation. Extremes 17 (1), 157–192.

Opitz, T., 2013. Extremal t-processes : Elliptical domain of attraction and a spectral representation. Journal of Multivariate Analysis 122, 409 – 413.

Paciorek, C. J., Schervish, M. J., 2006. Spatial modelling using a new class of nonstationary covariance functions. Environmetrics 17 (5), 483–506.

Padoan, S. A., 2011. Multivariate extreme models based on underlying skew-t and skew-normal distributions. Journal of Multivariate Analysis 102, 977–991.

Padoan, S. A., Ribatet, M., Sisson, S. A., 2010. Likelihood-based inference for max-stable processes. Journal of the American Statistical Association 105 (489), 263–277.

Panagoulia, D., Economou, P., Caroni, C., 2014. Stationary and nonstationary generalized extreme value modelling of extreme precipitation over a mountainous area under climate change. Environmetrics 25 (1), 29–43.

Parmigiani, G., Inoue, L., May 2009. Decision theory : principles and approaches. Wiley.

Pascal, Y., Tamara, S., Philippe, D., Laurent, M., Robert, V., Mathieu, V., 2013. Ensemble reconstruction of the atmospheric column from surface pressure using analogues. Climate dynamics 41 (5-6), 1333–1344, eng.

Piani, C., Haerter, J., Coppola, E., 2010. Statistical bias correction for daily precipitation in regional climate models over Europe. Theoretical and Applied Climatology 99 (1-2), 187–192.

Pitman, J., Mar. 1997. Some probabilistic aspects of set partitions. The American Mathematical Monthly 104 (3), 201–209.

Quintana-Seguí, P., Le Moigne, P., Durand, Y., Martin, E., Habets, F., Baillon, M., Canellas, C., Franchisteguy, L., Morel, S., 2008. Analysis of near-surface atmospheric variables : Validation of the SAFRAN analysis over France. Journal of applied Meteorology and Climatology 47 (1), 92–107.

Ribatet, M., 2013a. Spatial extremes : Max-stable processes at work. Journal de la Société Francaise de Statistique 154 (2), 156–177.

Ribatet, M., 2013b. Spatialextremes : Modelling spatial extremes.

Richardson, C., Barrow, E., Semenov, M., Brooks, R., 1998. Comparison of the WGEN and LARS-WG stochastic weather generators for diverse climates. Climate Research 10 (2), 95–107.

Rousseeuw, P. J., 1987. Silhouettes : A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53 – 65.

Rummukainen, M., 2010. State-of-the-art with Regional Climate Models. Wiley Interdisciplinary Reviews : Climate Change 1 (1), 82–96.

Ruti, P., Somot, S., Dubois, C., Calmanti, S., Ahrens, B., Alias, A., Aznar, R., Bartholy, J., Bastin, S., Béranger, K., Brauch, J., Calvet, J.-C., Carillo, A., Decharme, B., Dell'Aquila, A., Djurdjevic, V., Drobinski, P., Elizalde-Arellano, A., Gaertner, M., Galán, P., Gallardo, C., Giorgi, F., Gualdi, S., Harzallah, A., Herrmann, M., Jacob, D., Khodayar, S., Krichak, S., Lebeaupin, C., L'Heveder, B., Li, L., Liguro, G., Lionello, P., Onol, B., Rajkovic, B., Sannino, G., Sevault, F., 2014. MED-CORDEX initiative for Mediterranean Climate studies, in revision.

Sailor, D. J., Xiangshang, L., 1999. A semi-empirical downscaling approach for predicting regional temperature impacts associated with climatic change. Journal of Climate 12, 103–114.

Schlather, M., 2002. Models for stationary max-stable random fields. Extremes 5 (1), 33–44.

Schlather, M., Tawn, J. A., 2003. A dependence measure for multivariate and spatial extreme values : Properties and inference. Biometrika 90 (1), 139–156.

Schnur, R., Lettenmaier, D. P., 1998. A case study of statistical downscaling in Australia using weather classification by recursive partitioning. Journal of Hydrology 212, 362–379.

Sibuya, M., 1959. Bivariate extreme statistics, i. Annals of the Institute of Statistical Mathematics 11 (2), 195–210.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, M., Duda, K. G., Huang, X. Y., Wang, W., Powers, J. G., 2008. A description of the Advanced Research WRF Version 3. Tech. rep., National Center for Atmospheric Research.

Smith, R. L., 1990. Max-stable processes and spatial extreme, unpublished.

Snell, S. E., Gopal, S., Kaufmann, R. K., 2000. Spatial interpolation of surface air temperatures using artificial neural networks : Evaluating their use for downscaling gcms. Journal of Climate 13 (5), 886–895.

Spivey, M. Z., 2008. A generalized recurrence for bell numbers. Journal of Integer Sequences 11 (2), Article ID 08.2.5, 3 p., electronic only–Article ID 08.2.5, 3 p., electronic only.

Thibaud, E., Opitz, T., 2013. Efficient inference and simulation for elliptical pareto processes. arXiv preprint arXiv :1401.0168.

Thorarinsdottir, T. L., Gneiting, T., Gissibl, N., 2013. Using proper divergence functions to evaluate climate models.

Tödter, J., Ahrens, B., 2012. Generalization of the ignorance score : Continuous ranked version and its decomposition. Monthly Weather Review 140 (6), 2005–2017.

Vaittinada Ayar, P., Vrac, M., Bastin, S., Carreau, J., 2014. Statistical regionalization models intercomparisons and hydrological impacts project (starmip) : First results. In : EGU General Assembly Conference Abstracts. Vol. 16. p. 44.

Varin, C., Reid, N. M., Firth, D. D., 2011. An overview of composite likelihood methods. Statistica Sinica Vol.21 (No.1), 5–42.

Vautard, R., Gobiet, A., Jacob, D., Belda, M., Colette, A., Déqué, M., Fernández, J., García-Díez, M., Goergen, K., Guttler, I., Halenka, T., Karacostas, T., Katragkou, E., Keuler, K., Kotlarski, S., Mayer, S., van Meijgaard, E., Nikulin, G., Patarcić, M., Scinocca, J., Sobolowski, S., Suklitsch, M., Teichmann, C., Warrach-Sagi, K., Wulfmeyer, V., Yiou, P., 2013. The simulation of european heat waves from an ensemble of regional climate models within the EURO-CORDEX project. Climate Dynamics 41 (9-10), 2555–2575.

Vidal, J.-P., Martin, E., Franchistéguy, L., Baillon, M., Soubeyroux, J.-M., 2010. A 50-year high-resolution atmospheric reanalysis over France with the Safran system. International Journal of Climatology 30 (11), 1627–1644.

von Storch, H., Langenberg, H., Feser, F., 2000. A spectral nudging technique for dynamical downscaling purposes. Monthly weather review 128 (10), 3664–3673.

Vrac, M., 2012. Modélisations statistiques à différentes échelles climatiques et environnementales. Habilitation à diriger des recherches en mathématiques et en informatique.

Vrac, M., Drobinski, P., Merlo, A., Herrmann, M., Lavaysse, C., Li, L., Somot, S., 2012. Dynamical and statistical downscaling of the french mediterranean climate : uncertainty assessment. Natural Hazards and Earth System Science 12 (9), 2769–2784.

Vrac, M., Marbaix, P., Paillard, D., Naveau, P., 2007a. Non-linear statistical downscaling of present and LGM precipitation and temperatures over europe. Climate of the Past 3 (4), 669–682.

Vrac, M., Naveau, P., 2007. Stochastic downscaling of precipitation : From dry events to heavy rainfalls. Water Resources Research 43 (7).

Vrac, M., Stein, M., Hayhoe, K., 2007b. Statistical downscaling of precipitation through nonhomogeneous stochastic weather typing. Climate Research 34 (3), 169.

Vrac, M., Yiou, P., 2010. Weather regimes designed for local precipitation modeling : Application to the mediterranean basin. Journal of Geophysical Research : Atmospheres 115 (D12), 2156–2202.

Wang, Y., Stoev, S. A., 2011. Conditional sampling for spectrally discrete max-stable random fields. Advances in Applied Probability 43 (2), 461–483.

Wernli, H., Paulat, M., Hagen, M., Frei, C., 2008. Sal-a novel quality measure for the verification of quantitative precipitation forecasts. Monthly Weather Review 136 (11), 4470–4487.

Wigley, T. M. L., Jones, P. D., Briffa, K. R., Smith, G., 1990. Obtaining sub-grid-scale information from coarse-resolution general circulation model output. Journal of Geophysical Research : Atmospheres 95 (D2), 1943–1953.

Wilks, D., 1998. Multisite generalization of a daily stochastic precipitation generation model. Journal of Hydrology 210 (1), 178–191.

Wilks, D. S., Dec. 2006. Statistical methods in the atmospheric sciences, 2nd Edition. Academic Press.

Wilks, D. S., 2012. Stochastic weather generators for climate-change downscaling, part ii : multivariable and spatially coherent multisite downscaling. Wiley Interdisciplinary Reviews : Climate Change 3 (3), 267–278.

Williams, P. M., 1998. Modelling seasonality and trends in daily rainfall data. Advances in Neural Information Processing Systems 10, 985–991.

Xu, X., Reid, N., 2011. On the robustness of maximum composite likelihood estimate. Journal of Statistical Planning and Inference 141 (9), 3047–3054.

Yiou, P., Nogaj, M., 2004. Extreme climatic events and weather regimes over the north atlantic : When and where ? Geophysical Research Letters 31 (7).

Zolina, O., Kapala, A., Simmer, C., Gulev, S. K., 2004. Analysis of extreme precipitation over europe from different reanalyses : a comparative assessment. Global and Planetary Change 44 (1), 129–161.

Zorita, E., Von Storch, H., 1999. The analog method as a simple statistical downscaling technique : comparison with more complicated methods. Journal of climate 12 (8), 2474–2489.

Annexe A

Scores du chapitre 11

Tableau A.1 – Scores of the SHD algorithm with the different methods for building the DS-PRJ from the WRF-MED dataset. 12 conditioning points chosen by co-localization techniques.

	Models	CRPS	QS_{95}	ES	K-S	CVM	$RMSE_v$
No simulations	Interpolation	28.58	25.76	907	0.544	0.457	1406
	Linear downscaling	23.80	16.64	771	0.454	**0.278**	1351
Conditional Simulations	Raw	20.47	9.26	658	0.407	0.456	1077
	CDF-t	20.44	**6.55**	640	0.398	0.303	**805**
	Linear Regression	**19.01**	7.76	595	**0.386**	0.395	1047
	Optimal	18.04	4.87	543	0.369	0.366	636

Tableau A.2 – Scores of SHD algorithm with the different methods for building the DS-PRJ observations from the MEDCORDEX dataset with 57 conditioning points.

	Models	CRPS	QS_{95}	ES	KS	CVM	$RMSE_v$
No simulations	Zone Interpol.	28.18	25.04	46.45	0.548	0.467	1394
	Linear downscaling	24.92	17.98	41.12	0.480	**0.326**	1317
Conditional simulations	Raw	24.58	15.68	47.59	0.499	0.752	1244
	CDF-t	24.07	12.65	47.45	**0.398**	0.436	1278
	Linear Regression	**19.91**	**10.32**	**40.71**	0.399	0.428	**1158**
	Optimal	12.39	3.18	29.00	0.310	0.226	126

Tableau A.3 – Skill-Scores (with *Linear downscaling* as reference) of SHD algorithm with the different methods for building the DS-PRJ observations from the MEDCORDEX dataset with 57 randomly chosen conditioning points.

	Models	CRPSS	QSS_{95}	MSS_2	KS SS	CVM SS	$RMSE_v$ SS
No simulations	Interpolation	-13.1%	-39.3%	-13.0%	-20.8%	-43.3%	-5.8%
	Linear downscaling	0%	0%	0%	0%	**0%**	0%
Conditional simulations	Raw	1.4%	12.8%	-15.7%	-14.2%	-130.7%	5.5%
	CDF-t	3.4%	29.6%	-15.4%	-4.0%	-33.8%	3.0%
	Linear Regression	**20.1%**	**42.6%**	**1.0%**	**17.1%**	-31.3%	**12.1%**
	Optimal	50.3%	82.3%	29.5%	35.4%	30.7%	90.4%

Tableau A.4 – Scores of SHD algorithm with the different methods for building the DS-PRJ observations from the WRF-MED dataset with 12 conditioning points chosen by PAM algorithm into each grid-cell.

	Models	CRPS	QS_{95}	ES	KS	CVM	$RMSE_v$
No simulations	Interpolation	28.12	24.93	914	0.544	0.459	1393
	Linear downscaling	26.41	20.33	873	0.454	0.382	1420
Conditional simulations	Raw	23.87	14.87	766	0.489	0.752	1244
	CDF-t	**20.48**	**8.36**	**639**	**0.374**	**0.245**	**935**
	Linear Regression	21.56	12.44	708	0.433	0.528	1253
	Optimal	17.26	5.61	530	0.347	0.309	547

Tableau A.5 – Scores of stochastic SHD algorithm with the different methods for building the DS-PRJ observations from the WRF-MED dataset with 12 randomly chosen and switching conditioning points.

	Models	CRPS	QS_{95}	ES	KS	CVM	$RMSE_v$
No simulations	Interpolation	27.18	22.50	893	0.549	0.929	1406
	Linear downscaling	21.03	12.61	707	0.431	0.490	1363
Conditional simulations	Raw	22.77	13.23	750	0.472	0.689	1257
	CDF-t	**18.46**	**7.59**	**608**	**0.363**	**0.239**	**1039**
	Linear Regression	18.90	8.62	625	0.393	0.405	1169
	Optimal	13.31	3.05	445	0.353	0.308	441

Tableau A.6 – Scores of the SHD algorithm with the different methods for building the DS-PRJ observations from the WRF-EURO dataset with 15 conditioning points.

	Models	CRPS	QS_{95}	ES	KS	CVM	$RMSE_v$
No simulations	Interpolation	35.85	28.18	1053	0.553	0.469	1276
	Linear downscaling	26.39	17.35	812	0.441	0.272	1338
Conditional simulations	Raw	21.39	8.55	647	0.368	0.380	995
	CDF-t	23.53	**6.47**	652	0.397	**0.410**	**862**
	Linear Regression	**19.81**	7.79	**606**	**0.381**	0.432	930
	Optimal	16.98	4.59	529	0.353	0.305	528

Annexe B

Densités bivariées du processus *t* extrêmal

Afin d'estimer les paramètres du modèle *t* extrêmal (voir section *6.2* du chapitre 6), nous avons recours à une technique de maximisation de vraisemblance composite (Lindsay, 1988; Varin et al., 2011) car l'écriture de la vraisemblance complète n'est pas explicite. Cette vraisemblance, basée sur des paires d'observations, s'écrit en fonction des densités bivariées du processus étudié. Ainsi, ce chapitre détaille les calculs requis pour cette étape, notamment celui de la densité bivariée du processus *t* extrêmal.

Pour des distributions admettant des marginales α-Fréchet, on définit la fonction de dépendance

$$M(\mathbf{z}) = -\log \mathbb{P}(\mathbf{Z}^{\alpha} \leq \mathbf{z}), \text{ avec } z \in \mathbb{R}^d. \tag{B.1}$$

Notons f_{t_μ} et F_{t_μ} respectivement la fonction de densité et de répartition d'une Student standard univariée à μ degrés de liberté. Pour un processus *t* extrêmal et avec $d = 2$, Opitz (2013) a montré que

$$V(z_1, z_2) = \frac{1}{z_1} F_{t_{\alpha+1}}\left(\frac{(\frac{z_2}{z_1})^{1/\alpha} - \rho}{\sqrt{1-\rho^2}}\sqrt{1+\alpha}\right) + \frac{1}{z_2} F_{t_{\alpha+1}}\left(\frac{(\frac{z_1}{z_2})^{1/\alpha} - \rho}{\sqrt{1-\rho^2}}\sqrt{1+\alpha}\right), \tag{B.2}$$

avec $\rho = Cor(z_1, z_2)$.
Comme nous recherchons la densité bivariée du processus Z, on pose :

$$V(z_1, z_2) = M(z_1^{\alpha}, z_2^{\alpha}) \tag{B.3}$$

On peut donc écrire la densité bivariée du processus *t* extrêmal :

$$f(z_1, z_2) = e^{-V(z_1,z_2)}\left[\frac{\partial V(z_1, z_2)}{\partial z_1}\frac{\partial V(z_1, z_2)}{\partial z_2} - \frac{\partial^2 V(z_1, z_2)}{\partial z_1 \partial z_2}\right]. \tag{B.4}$$

En posant, $c = \sqrt{\alpha+1}\sqrt{1-\rho^2}^{-1}$, $c_{12} = c((\frac{z_1}{z_2}) - \rho)$ et $c_{21} = c((\frac{z_2}{z_1}) - \rho)$, on peut expliciter les termes du calcul précédent. Ainsi,

- $\dfrac{\partial V(z_1,z_2)}{\partial z_1} = -\dfrac{\alpha}{z_1^{\alpha+1}} F_{t_{\alpha+1}}(c_{21}) - \dfrac{c}{z_1^{\alpha+2}} z_2 f_{t_{\alpha+1}}(c_{21}) + \dfrac{c}{z_2^{\alpha+1}} f_{t_{\alpha+1}}(c_{12})$
- $\dfrac{\partial V(z_1,z_2)}{\partial z_2} = -\dfrac{\alpha}{z_2^{\alpha+1}} F_{t_{\alpha+1}}(c_{12}) - \dfrac{c}{z_2^{\alpha+2}} z_1 f_{t_{\alpha+1}}(c_{12}) + \dfrac{c}{z_1^{\alpha+1}} f_{t_{\alpha+1}}(c_{21})$
- $\dfrac{\partial^2 V(z_1,z_2)}{\partial z_1 \partial z_2} = -c(\alpha+1)\left[z_1^{-(\alpha+2)} f_{t_{\alpha+1}}(c_{21}) + z_2^{-(\alpha+2)} f_{t_{\alpha+1}}(c_{12})\right]$
$$- c^2\left[z_1^{-(\alpha+3)} z_2 f'_{t_{\alpha+1}}(c_{21}) + z_2^{-(\alpha+3)} z_1 f'_{t_{\alpha+1}}(c_{12})\right] \quad \text{(B.5)}$$

Une fois cette densité connue, il reste un travail à effectuer sur les marginales du processus (similaire à Padoan et al., 2010). En effet, pour respecter les hypothèses du modèle processus *t* extrêmal, les marginales doivent être α-Fréchet. Ainsi, on considère la bijection $(Z_i, Z_j) = g(Y_i, Y_j)$ qui admet comme fonction inverse :

$$\begin{aligned} Y_i &= \left(1 + \frac{\xi_i(Z_i - \mu_i)}{\lambda_i}\right)_+^{1/\alpha\xi_i} \\ Y_j &= \left(1 + \frac{\xi_j(Z_j - \mu_j)}{\lambda_j}\right)_+^{1/\alpha\xi_j}. \end{aligned} \quad \text{(B.6)}$$

Pour chaque marginale Z, les constantes μ, λ et ξ assurent que Y suit une distribution α-Fréchet. La densité bivariée transformée est donc

$$f_\alpha(z_i, z_j) = f\left[g^{-1}(z_i, z_j)\right] \times |J(z_i, z_j)|, \quad \text{(B.7)}$$

avec $f(.)$ la densité bivariée calculée dans (B.4) et

$$|J(z_i,z_j)| = \frac{1}{\alpha^2 \lambda_i \lambda_j} \left(1 + \frac{\xi_i(Z_i-\mu_i)}{\lambda_i}\right)_+^{1/\alpha\xi_i - 1} \times \left(1 + \frac{\xi_j(Z_j-\mu_j)}{\lambda_j}\right)_+^{1/\alpha\xi_j - 1}. \quad \text{(B.8)}$$

Ainsi, les paramètres de la GEV, α et les ρ sont les paramètres à estimer par maximisation de la vraisemblance composite. Comme cette maximisation ne peut se faire de manière analytique, il faut effectuer une maximisation numérique.

ABSTRACT

Extreme events are one of the major demonstrations of complex environmental systems. As they are inherently rare and often unexpected, society is not well prepared to face them.We are interested in extremes occurring in a spatial context and we address two different but related issues. The first one aims at evaluating the risk at locations lacking of measurements by providing distributions at these sites taking into account information available at surrounding locations. The second one is to provide distributions of extreme values based on information derived from climate models available at a very low spatial resolution, in a region where there is no observations available (for instance in the future but not necessarily).

From a mathematical point of view, the suitable probabilistic framework for modeling the frequency and intensity of extreme events is the statistical theory of extreme values. This theory is well established for univariate stationary series. There is now a growing effort to develop methods and models to deal with complex data involved in applications where space-time problems and non-stationarity are the rule rather than the exception. In that way, the last decade has seen max-stable processes emerge as a powerful tool for the statistical modeling of spatial extremes and there are increasing works using them in a climate framework. One recent utilization of max-stable processes in this context is for conditional simulations that provide empirical distribution of a spatial field conditioned by observed values in some locations. In this work, we improve an existing conditional simulation algorithm and adapt it to a recently studied and flexible process: the extremal t process.

For the second objective, in future projections context, we want to add high scale information to drive the conditional simulation methods. Thus, these metholodologies can be applied in the case where there is no longer local observations available.

For a few decades, climate models are used to provide future scenarios of precipitation with increasingly higher spatial resolution. However, this resolution is not yet sufficient to describe efficiently what happens at local scale. Dynamical and statistical methods of downscaling have been developed and allow us to make the link between two levels of resolution and enable us to get values at a local scale based on large-scale information from global or regional climate models. Nevertheless, both the extreme behavior and the spatial structure are not always well described by these downscaling methods. We propose a two-step methodology, called spatial hybrid downscaling (SHD), to solve this problem. The first step consists in applying a univariate (i.e., 1-dimensional) statistical downscaling to link the high- and low-resolution variables at some given locations. Once this 1d-link is performed, a conditional simulation algorithm of max-stable processes adapted to the extremal t process enables us to get conditional distributions of extreme precipitation at any point of the region.

After some tests on simuled data to understand the behaviour of the conditional simulation algorithm, applications are performed on precipitation data in French Mediterranean basin which is a key spot of occurrences of meteorological extremes such as heavy precipitation.

RÉSUMÉ

Les événements extrêmes sont des démonstrations majeures de systèmes naturels complexes. Comme ils sont par nature rares et imprévisibles, la société est mal préparée à y faire face. Nous nous intéressons aux événements extrêmes ayant une structure spatiale et nous ciblons deux objectifs corrélés. Dans un premier temps, on cherche à évaluer les risques à certains endroits pour lesquels il n'y a pas de mesures, en y déterminant des distributions de valeurs extrêmes en tenant compte de l'information donnée par les observations environnantes. Dans un second temps, nous cherchons à fournir des distributions de valeurs extrêmes à partir de sorties de modèles climatiques à faible résolution spatiale, dans des régions où il n'y a pas d'observations disponibles (par exemple dans le futur mais pas nécessairement).

D'un point de vue mathématique, le cadre de travail le plus approprié pour modéliser la fréquence et l'intensité des événements extrêmes est celui de la théorie statistique des valeurs extrêmes. Cette théorie est très bien établie pour des séries stationnaires univariées. Il y a désormais de plus en plus d'efforts fournis pour adapter ces méthodes et modèles à des applications spatio-temporelles où la non stationnarité est très courante. Dans cette optique, les dernière années ont vu se développer les processus max-stables pour la modélisation spatiale d'extrêmes. Par ailleurs leur utilisation est de plus en plus présente dans des travaux appliqués à des problématiques climatiques. Dans ce contexte, les processus max-stables ont été récemment utilisé pour des simulations conditionnelles, qui consistent à fournir des distributions empiriques d'un champ spatialisé, conditionné par des observations en certains sites. Dans ce travail, nous améliorons un algorithme de simulation conditionnel existant et l'adaptons à un processus apparu récemment et très flexible : le processus t extrêmal.

En ce qui concerne le second objectif, dans un contexte de projections futures, on souhaite incorporer de l'information à grande échelle pour guider ces méthodes de simulation conditionnelle. De la sorte, ces méthodes deviennent applicables dans un cas où il n'y a plus d'observation locale disponible.

Depuis quelques dizaines d'années, les modèles climatiques sont utilisés pour fournir des scénarios de précipitations futures avec une résolution spatiale qui ne cesse d'augmenter. Cependant, cette résolution demeure encore insuffisante pour décrire efficacement le comportement de ces phénomènes à l'échelle locale. Des méthodes de descente d'échelle dynamiques et statistiques ont été développées et nous permettent d'établir un lien entre ces deux niveaux de résolution. Ainsi, ces méthodes permettent d'obtenir des valeurs à l'échelle locale à partir d'information provenant des modèles de climat globaux ou régionaux. Néanmoins, le comportement des extrêmes et la structure spatiale des précipitations ne sont tous deux pas toujours bien retranscrits par ces techniques de downscaling. Nous proposons une méthode en deux étapes, appelée "Spatial Hybrid Downscaling"(SHD), pour résoudre ces problèmes. La première étape consiste à appliquer un downscaling statistique univarié pour établir une relation entre la variable à grande échelle et la variable locale en certains sites. Une fois ce lien établi, l'algorithme de simulation conditionnelle de processus max-stable adapté au processus t extrêmal nous permet d'obtenir des distributions conditionnelles de précipitation extrême en tout point de la région.

Après plusieurs tests sur données simulées pour comprendre le comportement de l'algorithme de simulation conditionnelle, des applications sont effectuées sur des données de précipitations dans le bassin méditerranéen français, une région où les occurrences de précipitations extrêmes sont régulières.