Analyse propabiliste régionale des précipitations : prise en compte de la variabilité et du changement climatique

Xun Sun

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Terre Univers Environnement**

Arrêté ministériel : 7 août 2006

Présentée par

Xun SUN

Thèse dirigée par **Michel LANG** et **Benjamin RENARD**
codirigée par **Mark THYER**

préparée au sein du **Irstea Lyon, France**
en collaboration avec **the University of Adelaide, Australia**

dans **l'École Doctorale Terre Univers Environnement**

Analyse probabiliste régionale des précipitations:

prise en compte de la variabilité et du changement climatique

Présentée et soutenue publiquement le **28 Octobre 2013**,
devant le jury composé de :

MME. Anne-Laure Fougères
Univ. Claude Bernard, Lyon I – Présidente
M. Bruno Merz
GFZ et Univ. Potsdam – Rapporteur
M. Henrik Madsen
DHI et Univ. Copenhagen – Rapporteur
M. Philippe Naveau
LSCE Gif-sur-Yvette – Examinateur
M. Michel Lang
Irstea, Lyon - Directeur de thèse
M. Benjamin Renard
Irstea, Lyon – Co-Directeur de thèse
M. Emmanuel Paquet
EDF/DTG Grenoble – Invité

Université Joseph Fourier / Université Pierre Mendès France /
Université Stendhal / Université de Savoie / Grenoble INP

Regional frequency analysis of precipitation accounting for climate variability and change

THESE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE DE GRENOBLE

Spécialité: Terre Univers Environnement

Préparée dans l'Unité de Recherche Hydrologie-Hydraulique

Irstea Lyon

Présentée par

Xun SUN

Soutenance le 28 Octobre 2013

Directeurs de thèse

Michel LANG, Irstea Lyon

Benjamin RENARD, Irstea Lyon

Collaborateur

Mark THYER

University of Adelaide

JURY

MME. Anne-Laure Fougères	Univ. Claude Bernard, Lyon I	Présidente
M. Bruno Merz	GFZ et Univ. Potsdam	Rapporteur
M. Henrik Madsen	DHI et Univ. Copenhagen	Rapporteur
M. Philippe Naveau	LSCE Gif-sur-Yvette	Examinateur
M. Michel Lang	Irstea, Lyon	Directeur de thèse
M. Benjamin Renard	Irstea, Lyon	Co- Directeur de thèse
M. Emmanuel Paquet	EDF/DTG Grenoble	Invité

Abstract

Extreme precipitations and their consequences (floods) are one of the most threatening natural disasters for human beings. In engineering design, Frequency Analysis (FA) techniques are an integral part of risk assessment and mitigation. FA uses statistical models to estimate the probability of extreme hydrological events which provides information for designing hydraulic structures. However, standard FA methods commonly rely on the assumption that the distribution of observations is identically distributed. However, there is now a substantial body of evidence that large-scale modes of climate variability (e.g. El-Niño Southern Oscillation, ENSO; Indian Ocean Dipole, IOD; etc.) exert a significant influence on precipitation in various regions worldwide. Furthermore, climate change is likely to have an influence on hydrology, thus further challenging the "identically distributed" assumption. Therefore, FA techniques need to move beyond this assumption. In order to provide a more accurate risk assessment, it is important to understand and predict the impact of climate variability/change on the severity and frequency of hydrological events (especially extremes).

This thesis provides an important step towards this goal, by developing a rigorous general climate-informed spatio-temporal regional frequency analysis (RFA) framework for incorporating the effects of climate variability on hydrological events. This framework brings together several components (in particular spatio-temporal regression models, copula-based modeling of spatial dependence, Bayesian inference, model comparison tools) to derive a general and flexible modeling platform. In this framework, data are assumed to follow a distribution, whose parameters are linked to temporal or/and spatial covariates using regression models. Parameters are estimated with a Monte Carlo Markov Chain method under the Bayesian framework. Spatial dependency of data is considered with copulas. Model comparison tools are integrated. The development of this general modeling framework is complemented with various Monte-Carlo experiments aimed at assessing its reliability, along with real data case studies.

Two case studies are performed to confirm the generality, flexibility and usefulness of the framework for understanding and predicting the impact of climate variability on hydrological events. These case studies are carried out at two distinct spatial scales:

- Regional scale: Summer rainfall in Southeast Queensland (Australia): this case study analyzes the impact of ENSO on the summer rainfall totals and summer rainfall maxima. A regional model allows highlighting the asymmetric impact of ENSO: while La Niña episodes induce a significant increase in both the summer rainfall totals and maxima, the impact of El Niño episodes is found to be not significant.
- Global scale: a new global dataset of extreme precipitation including 11588 rainfall stations worldwide is used to describe the impact of ENSO on extreme precipitations in the world. This is achieved by applying the regional modeling framework to 5x5 degrees cells covering all continental areas. This analysis allows describing the pattern of ENSO impact at the global scale and quantifying its impact on extreme quantiles estimates. Moreover, the asymmetry of ENSO impact and its seasonal pattern are also evaluated.

Résumé

Les événements de pluies extrêmes et les inondations qui en résultent constituent une préoccupation majeure en France comme dans le monde. Dans le domaine de l'ingénierie, les méthodes d'analyse probabiliste sont pratiquement utilisées pour prédire les risques, dimensionner des ouvrages hydrauliques et préparer l'atténuation. Ces méthodes sont classiquement basées sur l'hypothèse que les observations sont identiquement distribuées. Il y a aujourd'hui de plus en plus d'éléments montrant que des variabilités climatiques à grande échelle (par exemple les oscillations El Niño – La Niña, cf. indice ENSO) ont une influence significative sur les précipitations dans le monde. Par ailleurs, les effets attendus du changement climatique sur le cycle de l'eau remettent en question l'hypothèse de variables aléatoires "identiquement distribuées" dans le temps. Il est ainsi important de comprendre et de prédire l'impact de la variabilité et du changement climatique sur l'intensité et la fréquence des événements hydrologiques, surtout les extrêmes.

Cette thèse propose une étape importante vers cet objectif, en développant un cadre spatio-temporel d'analyse probabiliste régionale qui prend en compte les effets de la variabilité climatique sur les événements hydrologiques. Les données sont supposées suivre une distribution, dont les paramètres sont liés à des variables temporelles et/ou spatiales à l'aide de modèles de régression. Les paramètres sont estimés avec une méthode de Monte-Carlo par Chaînes de Markov dans un cadre Bayésien. La dépendance spatiale des données est modélisée par des copules. Les outils de comparaison de modèles sont aussi intégrés. L'élaboration de ce cadre général de modélisation est complétée par des simulations Monte-Carlo pour évaluer sa fiabilité.

Deux études de cas sont effectuées pour confirmer la généralité, la flexibilité et l'utilité du cadre de modélisation pour comprendre et prédire l'impact de la variabilité climatique sur les événements hydrologiques. Ces cas d'études sont réalisés à deux échelles spatiales distinctes:

- Echelle régionale: les pluies d'été dans le sud-est du Queensland (Australie). Ce cas d'étude analyse l'impact de l'oscillation ENSO sur la pluie totale et la pluie maximale d'été. En utilisant un modèle régional, l'impact asymétrique de l'ENSO est souligné: une phase La Niña induit une augmentation significative sur la pluie totale et maximale, alors qu'une phase El Niño n'a pas d'influence significative.
- Echelle mondiale: une nouvelle base de données mondiale des précipitations extrêmes composée de 11588 stations pluviométriques est utilisée pour analyser l'impact des oscillations ENSO sur les précipitations extrêmes mondiales. Cette analyse permet d'apprécier les secteurs où ENSO a un impact sur les précipitations à l'échelle mondiale et de quantifier son impact sur les estimations de quantiles extrêmes. Par ailleurs, l'asymétrie de l'impact ENSO et son caractère saisonnier sont également évalués.

Acknowledgements

At the beginning of this thesis, I want to deliver my sincere thanks to lots of people. Without their help, the thesis would not have been possible.

Many thanks must firstly be delivered to my supervisors, Michel Lang, Benjamin Renard of Irstea Lyon and Mark Thyer of the University of Adelaide. I have been very fortunate to be advised by them with a great project between France and Australia.

During these three years, Michel has provided me with numerous important suggestions, remarks for guiding this thesis. His idea of submitting our paper to *Nature* has greatly encouraged me for further research. Publishing in *Nature* is a dream. How great the dream is, and I hope it be realized in the further. Thanks Michel.

Special thanks should go to Benjamin. In composing this thesis, he has taught me lots of new knowledge of hydrology and statistics with great patience. Thanks to his tutorial on making presentation and writing report, my communication and writing skills are made huge progress. He has contributed many important thoughts to the achievement of this thesis. I also appreciate his efforts in organizing my visit to Australia. In my second year, I have been very fortunate to have a whole year's academic visit to Australia.

During my stay in Australia, Mark has helped me a lot not only in academics, but also in personal life, which facilitated my study there. The paper writing techniques that he taught me are very useful for my later papers. It was a worthwhile experience.

I also want to express my thanks to Seth Westra and Dmitri Kavetski of the University of Adelaide. Seth has offered a very complete database of global observed extreme precipitation for our case study and has contributed lots of ideas to the paper. Without him, the dream of publishing in *Nature* has not been existed yet. Dmitri has supported the FORTRAN library DMSL, which brought lots of convenience into model coding. The special "conditional" invitation of "lobster dinners" by Dmitri always makes me to think making good and "special" presentation in conferences. Although "condition" of presenting 50 slides within 20 minutes has not been reached, I'll try my best to do if so.

I have had many colleagues in the research unit, HHLY, whom I wish to thank: Jean-Philippe Vidal, Eric Sauquet and Thomas Cipriani have given me many supports in the dataset, catchment information and the usage of graphical tools. Jean-Philippe has also a great contribution on fixing the title of the thesis. Jean-Baptise Faure has helped me a lot in coding. Thanks all my colleagues who brought the home-made cakes and crêpes at coffee time. Dessert and sweets give me additional motivations for good jobs. I would also like to thank my officemate Mériem Labbas.

My thesis committee has given me numerous useful remarks, suggestions for this thesis. I would like to show my gratitude to Julie Carreau, Nicolas Eckert, Emmanuel Paquet, Mathieu Ribatet and Anne-Catherine Favre for their participation in my thesis committee.

I also wish to sincerely thank the juries for evaluating my PhD thesis: Bruno Merz, Henrik Madsen, Anne-Laure Fougères and Philippe Naveau.

The financial support of this thesis was provided by Irstea DRI and EDF (French electricity producer). The travel fund to Australia was supported by the University of Adelaide, Irstea DRI and region Rhône-Alpes (Explora'doc).

In the end, I wish to thank my parents and my fiancée Yichen. Thanks for their support and encouragement, as well as thanks Yichen for flying across continents several times to meet me, and making the unforgettable desserts for me. Their trust encourages me to continue my study and to live my dream.

Résumé étendu

1 Contexte général

Le rapport *IPCC* [*2007a*] du Groupe d'experts intergouvernemental sur les changements climatiques indique clairement que le climat a changé au cours du 20ème siècle. La température de surface moyenne mondiale a augmenté de 0,74 °C pendant les 100 dernières années (1906-2005), le niveau de la mer a augmenté en moyenne de 1,8 mm par an entre 1961 et 2003, les glaciers de montagne et la couverture neigeuse dans l'hémisphère nord ont également diminué de manière significative, avec une baisse de près de 7% depuis 1900 de la superficie maximale saisonnière du sol gelé, etc. En conséquence, ces modifications peuvent avoir un impact global sur le cycle de l'eau et les systèmes de circulation océaniques et atmosphériques, et donc influencer les précipitations et les écoulements fluviaux.

Par ailleurs, il existe de nombreux éléments qui montrent que les oscillations climatiques à grande échelle exercent également une influence significative sur les variables hydrologiques dans différentes régions du monde [*Henley et al., 2011*]. Par exemple, l'oscillation australe ENSO est l'un des modes de variabilité climatique les plus importants et a un impact global sur les variables hydro-météorologiques [*Hoerling et al., 1997*]; l'oscillation Nord Atlantique (NAO) contrôle le système de vents d'ouest et les trajectoires des tempêtes dans l'Atlantique Nord vers l'Europe [*Barnston and Livezey, 1987*], et l'oscillation de l'Océan Indien (IOD) a un lien significatif avec la saisonnalité et la variation des précipitations sur la région de l'océan Indien et affecte aussi la mousson asiatique [*Saji et al., 1999*].

Jusqu'à récemment, les études hydrologiques étaient souvent basées sur l'hypothèse de «stationnarité», en considérant que les observations du passé récent pouvaient être exploitées pour la gestion des ressources en eau et des phénomènes extrêmes des prochaines décennies. Or de plus en plus d'éléments suggèrent que cette hypothèse devrait être reconsidérée, en prenant en compte les effets attendus du changement climatique et en s'intéressant à la variabilité climatique des variables hydrologiques (par exemple, les précipitations extrêmes). Récemment, certains hydrologues ont même déclaré *«stationnarity is dead»* et ont suggéré d'abandonner l'hypothèse de stationnarité dans l'ingénierie liée à l'eau [*Milly et al., 2008*]. Bien que l'abandon du concept de stationnarité soit toujours en discussion [*Lins and Cohn, 2011*], un consensus se dégage sur le fait qu'il est important d'intégrer les impacts possibles des changements climatiques et/ou de la variabilité climatique pour donner des prévisions fiables sur les variables de précipitation et de débit [*Stedinger and Griffis, 2011*].

Cette thèse propose une étape importante vers cet objectif, en élaborant un cadre flexible d'analyse probabiliste qui permet de modéliser la variabilité temporelle des variables hydrologiques, que ce soit vis à vis de la variabilité climatique ou du changement climatique.

2 Modèles d'analyse fréquentielle

En hydrologie opérationnelle, les méthodes d'analyse probabiliste permettent d'estimer la probabilité associée au dépassement d'une valeur hydrologique de référence, ou inversement

de dimensionner un ouvrage par rapport à une probabilité cible de défaillance. De nouveaux cadres probabilistes ont été développés pour intégrer l'impact de la variabilité climatique et du changement climatique sur l'intensité et la fréquence des événements hydrologiques, en particulier pour les extrêmes.

2.1 Approche locale, en supposant que les variables sont identiquement distribuées

Les méthodes standard d'analyse probabiliste utilisent les observations pour estimer les paramètres d'une distribution prédéterminée. Plus formellement, étant donné un échantillon d'observations $Y_1,...,Y_n$ qui sont supposées identiquement distribuées dans la plupart des cas, les paramètres de la distribution sont estimés avec une méthode d'estimation particulière (par exemple, maximum de vraisemblance, méthode des moments ou des moments pondérés, estimation bayésienne). Une multitude de recherches a été menée dans ce cadre au cours des dernières décennies. La plupart des études ont porté sur le choix de la distribution parente et de la méthode d'estimation (par exemple, *Durrans et Tomic* [2001]; *Lui et Valeo* [2009]; *Hosking et al.* [1985]; *Kroll et Stedinger* [1996]; *Lang et al*. [1999]; *Madsen et al.* [1997a]; *Meshgi et Khalili* [2009]; *Ribatet et al.* [2007]; *Sankarasubramanian et Srinivasan* [1999]), ou la quantification de l'incertitude (par exemple *Chowdhury et al.* [1991]; *Cohn et al*. [2001]; *Kysely* [2008]; *Stedinger* [1983]; *Stedinger et Tasker* [1985]; *Stedinger et al.* [2008]).

En complément de ces méthodes basées sur l'estimation d'une distribution prédéfinie, il existe des approches basées sur la simulation hydrologique, en s'intéressant aux précipitations [*Arnaud et Lavabre*, 1999] et aux inondations [*Boughton et Droop*, 2003; *Hundecha et Merz*, 2012]. D'autres méthodes ont été développées pour intégrer des informations antérieures aux observations issues des réseaux d'observation systématique, à partir de recherches documentaires sur les crues historiques ou d'études paléo-hydrologiques sur les dépôts de sédiments de crue (par exemple *Naulet et al.* [2005]; *Neppel et al.* [2010]; *O'Connell et al*. [2002]; *Payrastre et al.* [2011]; *Reis et Stedinger* [2005]; *Stedinger et Cohn* [1986]).

2.2 Approche locale avec un paramétrage temporel

Dans un contexte non stationnaire, *Renard et al.* [2006a] et *Ouarda et El-Adlouni* [2011] ont proposé une adaptation de l'analyse probabiliste, avec l'introduction de paramètres de la distribution qui varient dans le temps. Avec une structure similaire, *Rust et al.* [2009] ont discuté de la saisonnalité des précipitations extrêmes au Royaume-Uni, *Kysely et al.* [2010] ont décrit les tendances dans la température journalière et *Nogaj et al.* [2006] ont analysé l'amplitude et la fréquence des températures extrêmes. *Khaliq et al.* [2006] ont réalisé une revue bibliographique des méthodes probabilistes locales avec un paramétrage temporel. Des informations climatiques/météorologiques ont également été intégrées à l'analyse : par exemple, *Micevski et al.* [2006] ont utilisé l'indice IPO sur l'oscillation du Pacifique pour caractériser le risque d'inondation en Australie, *Tramblay et al.* [2011] ont retenu différentes covariables climatiques pour analyser les pluies fortes dans le sud de la France, et *Garavaglia et al.* [2010], [2011]; *Paquet et al.* [2013] ont exploité le type du temps pour quantifier le risque des précipitations.

Bien que les méthodes probabilistes locales permettent d'inclure des informations sur la variabilité ou le changement climatique, ces modèles restent toujours limités sur deux aspects :

(1) l'analyse locale ne peut être appliquée aux sites non jaugés ;

(2) l'incertitude des estimations des paramètres est très large, du fait de la taille réduite de l'échantillon d'observations dans un modèle local. L'usage de modèles plus complexes, intégrant des co-variables climatiques ou temporelles, rend encore plus sensible ce type d'approche à l'incertitude d'échantillonnage. Les observations sont le plus souvent insuffisantes pour identifier correctement les paramètres [*Thyer et al*., 2006].

Ces difficultés expliquent le développement de méthodes d'analyse probabiliste régionale, qui utilisent les informations disponibles sur plusieurs sites et permettent d'augmenter la taille des échantillons.

2.3 Méthode probabiliste régionale en supposant que les variables sont identiquement distribuées

L'approche régionale classique consiste à exploiter les informations de plusieurs sites pour effectuer l'inférence des paramètres, et permet d'obtenir des estimations plus précises. Le principe des méthodes régionales est de supposer que certains paramètres sont communs pour tous les sites dans une région homogène, et que d'autres paramètres peuvent être prédits à partir d'une régression sur les caractéristiques du site, par exemple pour les précipitations, à partir de l'altitude, la distance à la mer, etc.

De nombreux travaux ont été réalisés sur les approches régionales en contexte stationnaire : par exemple *Durrans et Kirby* [2004]; *Overeem et al.* [2008]; *Yu et al*, [2004]; *Cooley et al*. [2007]; *Madsen et Rosberg* [1997a], [1997b]; *Madsen et al*. [1997b]; [2002] et *Ghosh et Mallick* [2011]. Une comparaison entre les approches régionales et locales sur les précipitations extrêmes a été effectuée par exemple par *Kysely et al.* [2011].

2.4 Méthode probabiliste régionale avec des co-variables temporelles

Cunderlik and Burn [2003] et *Leclerc et Ouarda* [2007] ont proposé des modèles régionaux non-stationnaires pour les crues, et *Hanel et al.* [2009] ont adapté la méthode de l'indice de crue avec un paramétrage temporel pour l'analyse des précipitations extrêmes. Récemment, plusieurs auteurs (*Aryal et al*. [2009]; *Lima et Lall* [2010]; *Maraun et al.* [2010]; *Maraun et al.* [2011]; *Sang et Gelfand* [2009]) ont commencé à développer des modèles spatio-temporels. *Gregersen et al.* [2013] ont utilisé un modèle de régression Poissonnien pour décrire la fréquence des précipitations extrêmes dans l'espace et le temps. Une difficulté commune de toutes ces approches est le traitement de la dépendance spatiale entre les données. Ce point fait l'objet d'un développement spécifique dans cette thèse.

3 Contributions principales de la thèse

La contribution principale de la thèse est de construire un cadre rigoureux d'analyse probabiliste régionale, avec un paramétrage qui prend en compte la variabilité spatio-

temporelle des variables hydrologiques. Ce cadre permet d'identifier et de quantifier d'éventuelles tendances temporelles et impacts de la variabilité climatique sur l'intensité et la fréquence des événements hydrologiques. Il intègre dans sa conception plusieurs composantes, avec des modèles de régression spatio-temporels, une modélisation de la dépendance spatiale avec une copule, une approche bayésienne pour l'estimation des paramètres et la comparaison de modèles. Il vise à obtenir une plate-forme de modélisation générale et flexible.

3.1 Objectifs

Les objectifs précis de la thèse sont décrits ci-dessous :

1. Développement du modèle, avec des outils pour l'inférence et la comparaison. Dans le modèle, les régressions avec les co-variables spatiales et temporelles sont utilisées pour décrire la variabilité des paramètres. Dans l'ajustement du modèle, la dépendance spatiale entre données est prise en compte et l'estimation des paramètres est effectuée dans un cadre bayésien, ce qui permet d'avoir directement une estimation des incertitudes. De plus, les différents modèles de régression liés au climat peuvent être comparés (par exemple, une régression linéaire et non-linéaire). Cela permet d'identifier la régression qui semble la mieux adaptée pour représenter la variabilité climatique et la variabilité spatiale.

2. Évaluation du modèle. Elle est réalisé à partir de cas d'études synthétiques avec des données simulées, et à l'aide de jeux de données observées: (i) évaluation de la cohérence et de la différence entre modèles variant dans le temps et modèles supposant que les variables sont identiquement distribuées par sous-période ; (ii) évaluation de l'intérêt de considérer la dépendance spatiale ; (iii) comparaison des différentes structures de dépendance spatiale (copule vs processus maximum stable) en termes d'estimation des probabilités conjointes et conditionnelles.

3. Applications du modèle. Deux cas d'études visant à quantifier l'impact ENSO sur les précipitations sont illustrés.

a) Quantification de l'impact de l'oscillation ENSO sur les pluies totales et maximales d'été dans le Sud-Est du Queensland, en Australie. La flexibilité du cadre permet de tester plusieurs hypothèses, qui sont associées aux questions suivantes :
   i) Y-a-t-il un impact de l'oscillation ENSO sur les pluies maximales journalières d'été ?
   ii) Est-ce que l'impact de l'oscillation ENSO sur les pluies maximales journalières d'été est asymétrique (c.a.d avec des effets différents suivant les phases El Niño et La Niña) ?

b) Evaluation de l'impact de l'oscillation ENSO sur l'intensité des précipitations extrêmes dans le Monde. Cette analyse n'est pas basée sur les données moyennées sur une maille, mais sur une nouvelle base de données mondiale d'observations (HadEX2). La thèse se concentre sur l'analyse des extrêmes observés au droit des stations, avec le modèle régional mis au point pendant la thèse. Il y a trois objectifs :
   i) Identifier les régions touchées par l'oscillation ENSO et quantifier son impact sur les quantiles de précipitations extrêmes (par exemple, précipitations centennales) ;

ii) Evaluer l'asymétrie possible de l'impact de l'oscillation ENSO ;

iii) Décrire le caractère saisonnier de l'impact de l'oscillation ENSO.

3.2 Intérêt opérationnel de la thèse pour l'ingénierie

Les ingénieurs chargés du dimensionnement des structures hydrauliques utilisent des valeurs hydrologiques de référence associées à une probabilité de défaillance admissible. Cette question est particulièrement complexe dans le contexte du changement climatique. Évidemment, il est dangereux d'extrapoler une tendance ajustée sur les observations du passé, car les résultats dépendent fortement de la formulation de la tendance [*Cooley*, 2013], et celle-ci n'a aucune raison particulière de se prolonger à l'identique dans le futur. Une approche commune est d'exploiter les sorties des modèles climatiques GCM/RCM, avec un modèle probabiliste stationnaire, où l'on considère que les données sont identiquement distribuées sur des sous-périodes élémentaires (e.g. *Brigode* [2013]; *Madsen et al.* [2009]). Le choix de la longueur des sous-périodes résulte d'un compromis entre une période assez courte pour que l'hypothèse de stationnarité reste acceptable et une période suffisamment longue pour avoir une incertitude d'estimation raisonnable. Le modèle développé dans cette thèse permet une approche plus complète en exploitant la totalité de la série simulée, avec un paramétrage temporel continu.

Par ailleurs, l'occurrence des événements extrêmes est liée à la variabilité climatique. Certaines phases des oscillations climatiques sont davantage propices à l'occurrence d'événements exceptionnels que d'autres (par exemple, phases El Niño ou La Niña). Il est alors possible, avec le modèle développé dans la thèse, d'estimer une distribution conditionnelle, sous hypothèse de phase climatique. Une meilleure planification des interventions d'urgence, et aussi potentiellement une amélioration des règles de fonctionnement des réservoirs peuvent être envisagées pour mieux contrôler les inondations et réduire l'impact des événements, dans certaines conditions particulières du climat.

4 Résultats de la thèse

Dans cette thèse, un cadre spatio-temporel d'analyse probabiliste régional a été développé, orienté vers l'analyse et la prédiction des risques hydrologiques, qui permet d'inclure les tendances temporelles et les effets de la variabilité climatique et du changement du climat.

4.1 Le développement du cadre de modélisation régional (Chapitres 2 et 4)

Le développement du modèle a été réalisé en deux étapes : (i) la construction du modèle local avec un paramétrage uniquement temporel (Chapitre 2), et (ii) la construction du modèle régional, avec un paramétrage spatial et temporel (Chapitre 4). Le premier modèle peut être considéré comme un cas particulier du modèle régional. Ce cadre spatio-temporel (Figure I) établit une plateforme très flexible pour l'analyse des variables hydrologiques en utilisant trois types de co-variables : temporelles, spatiales et spatio-temporelles. En particulier, le cadre

fournit un choix libre sur la distribution des variables, que ce soient des distributions discrètes ou continues. La relation entre les co-variables temporelles (ou spatio-temporelles) et les données est modélisée par une régression temporelle où les paramètres de la distribution sont des fonctions de co-variables temporelles. La sélection des co-variables temporelles est également flexible. Toutes les co-variables déterministes (e.g. le temps) et aléatoires (e.g. ENSO) peuvent être utilisées. Les effets spatiaux sur les paramètres sont pris en compte à l'aide d'une fonction de régression spatiale entre les variables spatiales et les paramètres.

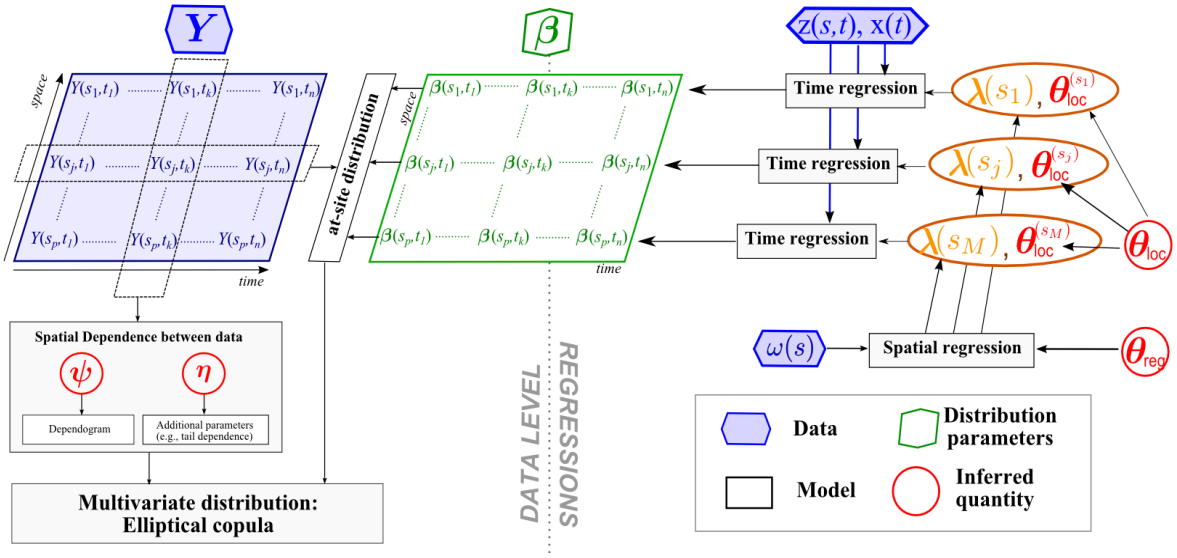


Figure I - Le schéma du modèle régional

Le modèle régional comporte un volet sur la modélisation de la dépendance entre données d'une même région. Dans la thèse, nous avons utilisé des copules elliptiques pour modéliser la dépendance spatiale des précipitations. Par rapport aux modèles qui ignorent la dépendance spatiale, ce modèle permet de mieux estimer les incertitudes des paramètres, ce qui conduit à un diagnostic plus réaliste sur la détection de tendance temporelle et/ou d'effet du climat.

Le modèle spatio-temporel est complété avec d'autres outils, pour faciliter le traitement des valeurs manquantes dans les données, pour réaliser l'estimation des paramètres et effectuer une comparaison et un diagnostic entre différentes implémentations du modèle. L'algorithme de calcul de la fonction de vraisemblance a été adapté pour pouvoir traiter des données non manquantes à chaque pas de temps, sans gaspiller aucune donnée ni procéder au comblement des lacunes. Par ailleurs, l'analyse bayésienne permet d'exploiter une information a priori sur les paramètres et de quantifier les incertitudes facilement et naturellement. Dans cette thèse, un nouvel algorithme MCMC est développé, qui combine plusieurs algorithmes existants (méthode adaptative de bloc de Metropolis, méthode adaptative de Metropolis-Hastings et méthode classique de Metropolis-Hastings). Cet algorithme permet d'estimer les paramètres rapidement et efficacement dans un cadre bayésien. Avec les variables variant dans le temps, le test d'adéquation est dans un premier temps réalisé à l'aide d'un graphique PP plot. Le critère DIC est utilisé pour comparer différents modèles dans un cadre bayésien, en testant différentes hypothèses de distribution ou différentes fonctions de régression. Avec

tous ces outils, le cadre spatio-temporel développé dans cette thèse fournit un moyen flexible et pratique pour mieux identifier les tendances temporelles et les impacts de la variabilité climatique sur les événements hydrologiques, ainsi que le risque hydrologique induit.

4.2 L'évaluation du cadre de modélisation (Chapitres 3 et 5)

Plusieurs cas d'étude sont présentés pour évaluer l'intérêt du cadre de modélisation. Au niveau local, avec des données issues de simulations par modèles GCM sur les précipitations du 21e siècle dans le bassin versant de la Durance, nous démontrons la flexibilité en termes de choix de distribution et de fonctions de régression. Ceci est très utile lorsque l'on analyse des variables pour lesquelles aucune expertise a priori ne permet de statuer sur le paramétrage du modèle. Parmi les variables étudiées, une tendance temporelle est détectée sur la variable "premier jour de neige". Cette variable est également analysée en utilisant la même distribution avec un modèle invariant dans le temps avec deux sous-périodes (1970-1999 et 2035-2064). Il y a une nette évolution de la valeur moyenne et des quantiles entre les deux sous-périodes (Figure II). Ce résultat est cohérent avec celui trouvé par le modèle non stationnaire, avec une évolution temporelle continue. Ce modèle permet en outre d'avoir des estimations avec de plus faibles incertitudes d'estimation. Les probabilités de défaillance sont également évaluées avec ces deux modèles, et les résultats montrent que le modèle développé dans la thèse est plus adapté pour évaluer le risque sur une longue durée.

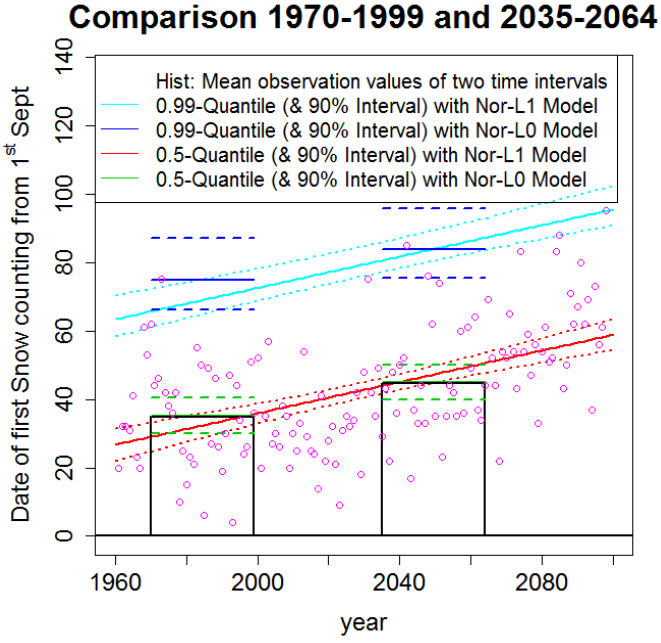


Figure II – Estimation de quantiles sur la variable « premier jour de neige », avec approche stationnaire par sous-période ou approche non stationnaire

La flexibilité du modèle est aussi discutée sur un jeu régional de précipitations, en zone méditerranéenne française (Figure III). Les tendances temporelles et les effets des oscillations NAO sur la précipitation journalière maximale annuelle sont analysés à l'échelle locale. Avec une distribution GEV, six modèles de régressions sont testés : stationnarité, tendance

uniquement temporelle ou avec un effet de l'oscillation NAO. A l'échelle locale, il n'y a généralement aucune indication forte sur l'existence d'une tendance temporelle ou d'un effet de l'oscillation NAO.

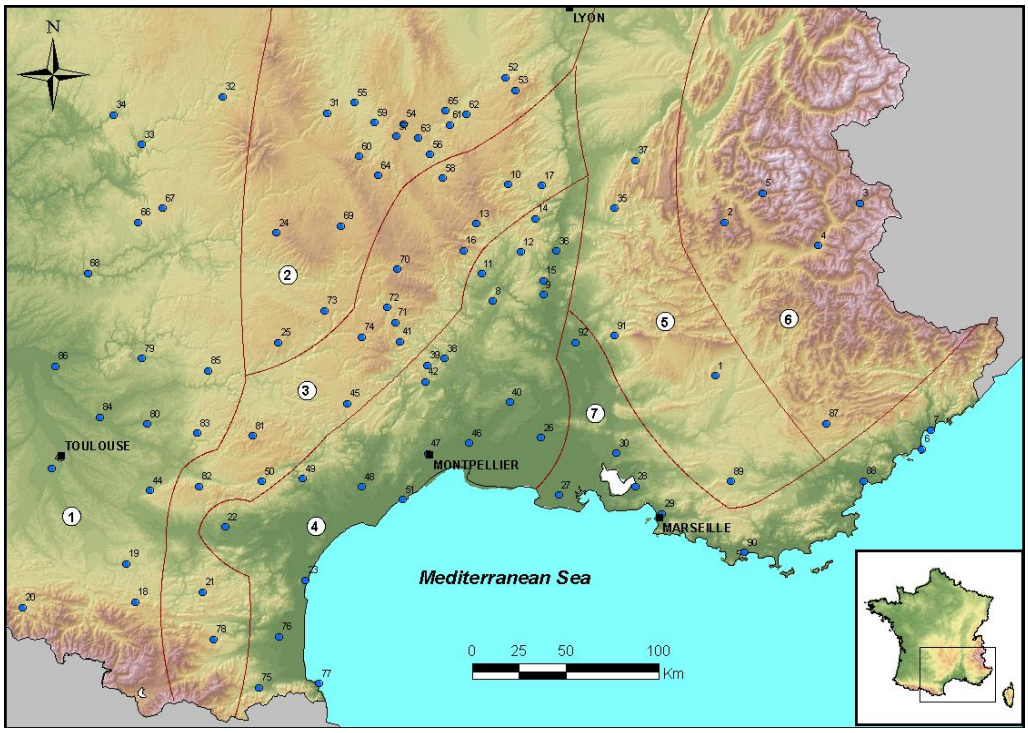


Figure III – Les 92 stations en France Méditerranée. Les sept régions homogènes ont été classifiées par Pujol et al., [2007].

L'analyse se poursuit à une échelle régionale. La deuxième évaluation consiste à évaluer l'intérêt de considérer une dépendance spatiale, en utilisant une copule elliptique. Deux groupes de données sont simulés avec une loi GEV qui varie dans le temps. L'estimation des paramètres est effectuée avec ou sans prise en compte de la dépendance spatiale. Ignorer la dépendance spatiale conduit à sous-estimer l'incertitude des paramètres (Figure IV). Une autre simulation est basée sur des données simulées à partir de processus maximum stable, avec des estimations réalisées avec une copule Gaussienne. Lorsque les données sont simulées avec une dépendance modérée, une amélioration significative a été trouvée avec l'estimation qui prend en compte la dépendance spatiale. Par contre, les résultats ne sont pas aussi convaincants avec une dépendance élevée : l'utilisation d'une copule Gaussienne donne une quantification réaliste sur l'incertitude des paramètres de position et la tendance, mais elle surestime le paramètre de la forme.

La troisième évaluation consiste à comparer la copule Gaussienne avec un processus maximum stable pour modéliser la dépendance spatiale sur les données extrêmes. Les probabilités conjointes et conditionnelle d'un événement dépassant un seuil entre deux sites sont comparées. Les résultats montrent que le modèle max-stable de Schlather surestime systématiquement les probabilités conjointes et conditionnelles, du fait de limitations pour représenter les dépendances à grande distance. D'autre part, le modèle max-stable Smith et la copule Gaussienne conduisent à de grosses différences sur l'estimation des probabilités

conditionnelles, ce qui peut s'expliquer par le fait que le modèle Smith est asymptotiquement dépendant, tandis que la copule gaussienne est asymptotiquement indépendante. En général, ces résultats suggèrent que même si une copule Gaussienne peut donner des résultats acceptables en termes d'estimation des paramètres marginaux, le calcul des probabilités de dépassement jointes ou conditionnelles est beaucoup plus sensible à la représentation de la dépendance spatiale. L'utilisation inadaptée d'une copule Gaussienne peut conduire à la sous-estimation de la probabilité de dépassement conditionnelle.

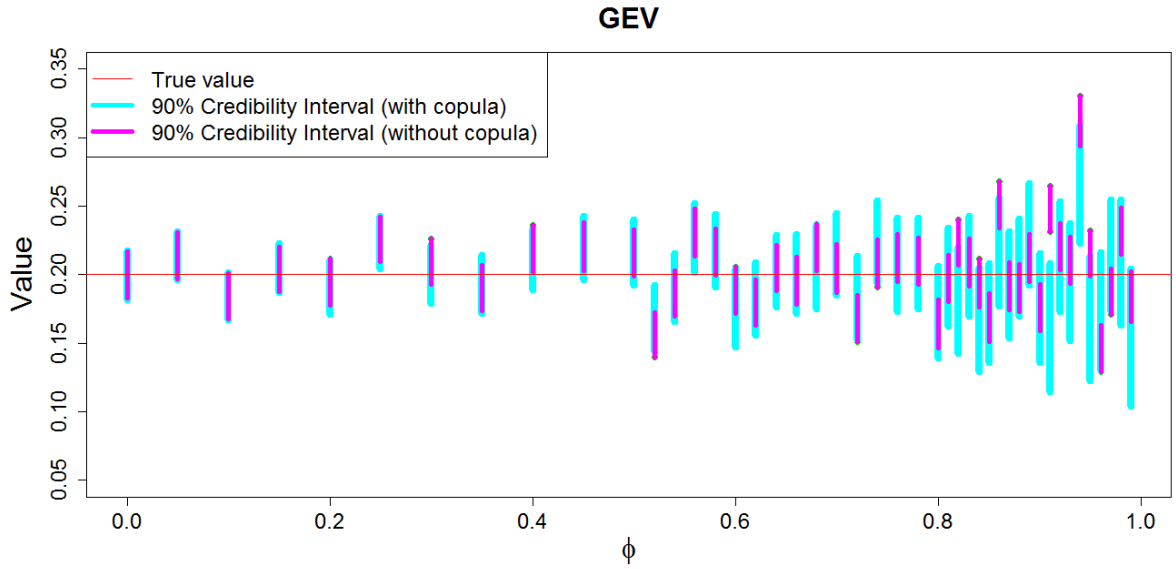


Figure IV – Comparaison des intervalles de crédibilité à 90% sur le paramètre de la dépendance obtenus en ignorant la dépendence ou en la modélisant avec une copule Gaussienne. Le paramètre ϕ contrôle le degré de dépendance spatiale.

4.3 L'impact des oscillations ENSO sur les précipitations globales

L'une des principales contributions de cette thèse est la quantification de l'impact des oscillations ENSO sur les précipitations. Nous présentons un premier cas, centré sur l'impact d'ENSO sur les pluies journalières maximales d'été dans le sud-est du Queensland en Australie. Le second cas d'étude concerne une analyse globale de l'impact d'ENSO sur les précipitations extrêmes dans le Monde. La co-variable temporelle utilisée dans le modèle est l'indice climatique SOI et l'impact d'ENSO est quantifié en fonction de la valeur de l'indice SOI.

Précipitations d'été dans le Sud-Est du Queensland (Chapitre 6)

Dans le cas du Sud-Est du Queensland, l'impact asymétrique de l'ENSO a été évalué sur la pluie totale d'été. Une fonction de régression linéaire asymétrique qui sépare les différentes phases de l'ENSO est utilisée sur la moyenne d'une loi log-Normale. Les résultats montrent qu'une phase La Niña exerce une influence importante dans la région, alors que l'impact d'El Niño n'est pas significatif. Le phénomène est cohérent avec la revue de la littérature. Ensuite, l'analyse se tourne vers les valeurs maximales de précipitation, ce qui est plus difficile à analyser que le cumul de la pluie d'été (cf. valeurs maximales). Grâce à la flexibilité du cadre

développé, les différentes hypothèses peuvent être comparées dans cette étude, qui utilise les paramètres locaux et régionaux pour exprimer les hypothèses de stationnarité, l'impact symétrique ou asymétrique de l'oscillation ENSO (Figure V). En utilisant le critère DIC de comparaison, les modèles régionaux permettent une meilleure identification de l'impact des oscillations ENSO sur les précipitations extrêmes dans le Sud-Est du Queensland. Les oscillations ENSO ont un impact asymétrique : la phase La Niña a une forte incidence sur les pluies maximales d'été, tandis que la phase El Niño n'a pas d'effet significatif. Les résultats sont cohérents avec les conclusions sur les pluies totales d'été.

Ainsi, la pluie centennale en phase La Niña peut être de 20 à 50% plus élevée que celle estimée sous hypothèse de stationnarité. Ceci fournit des informations utiles pour les planificateurs pour améliorer la gestion opérationnelle et organiser des interventions d'urgence pour une année en phase forte La Niña.

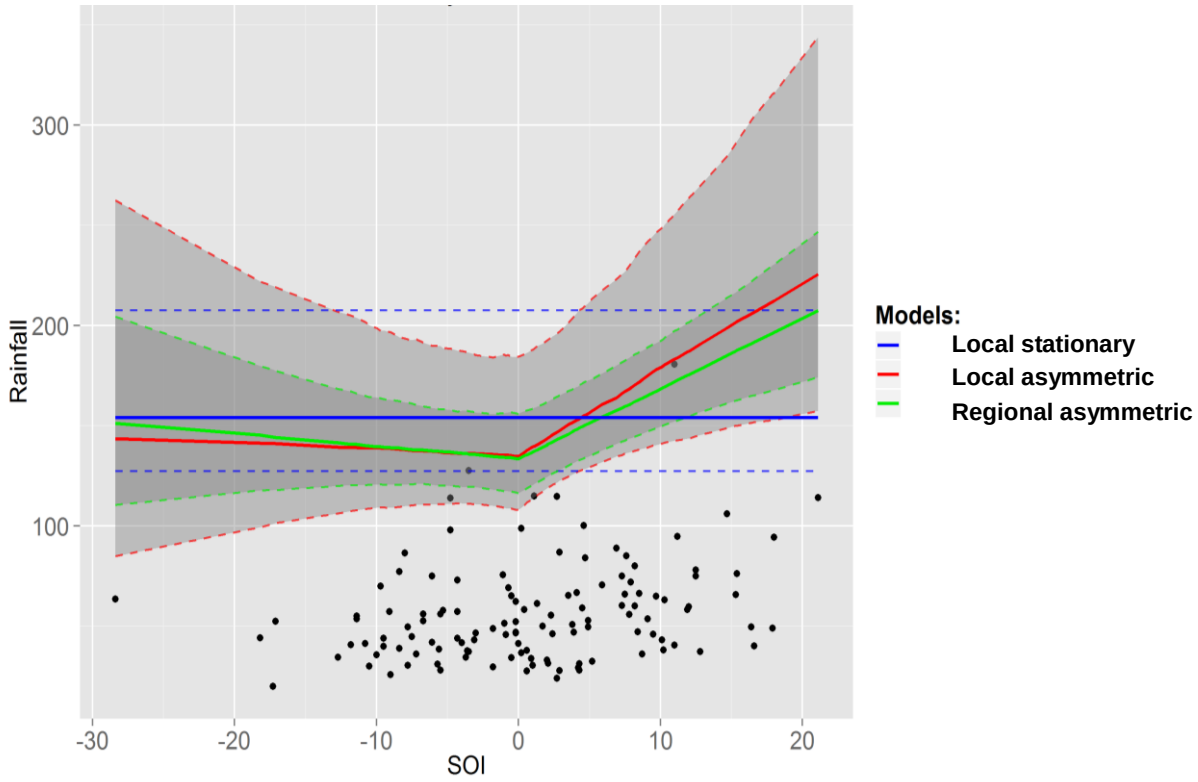


Figure V – Estimation de la pluie centennale journalière d'été sur un site en fonction de l'indice climatique SOI. L'analyse est effectuée avec trois modèles : stationnaire local, asymétrique local et asymétrique régional.

L'impact des oscillations ENSO sur les précipitations extrêmes dans le monde (Chapitre 7)

Ce premier constat d'un impact asymétrique des oscillations ENSO sur les pluies d'été du Queensland en Australie nous conduit à nous intéresser maintenant à l'impact ENSO sur les précipitations extrêmes mondiales. Cette analyse est basée sur une nouvelle base de données mondiale des précipitations extrêmes composée de 11588 stations pluviométriques, dont environ 7000 stations avec plus de 40 ans d'observation. La carte du monde est

quadrillée suivant une grille de 5°*5° (latitude, longitude). Chaque carré est considéré comme une région homogène. Bien qu'elle varie avec la latitude, la surface de chaque cellule est suffisante pour effectuer une analyse régionale sur chacune des mailles de la grille.

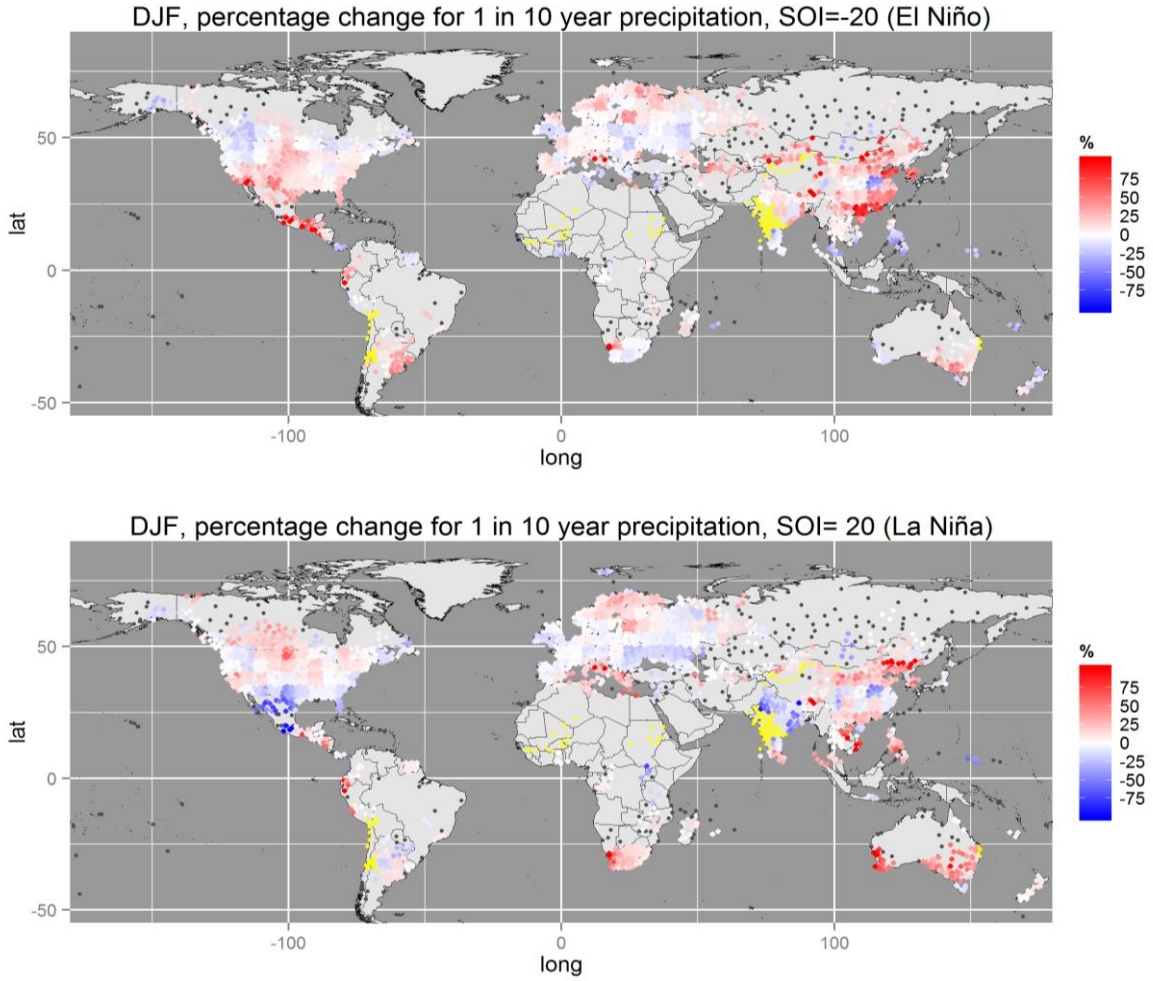


Figure VI – Pourcentage de changement de la pluie décennale, entre une phase forte de El Niño (haut) ou La Niña (bas) par rapport à une phase neutre.

Les résultats montrent pendant l'hiver boréal (DJF), lors de la phase El Niño, une augmentation des précipitations extrêmes en Amérique du Nord (le sud-ouest), en Amérique du Sud (le sud), la Chine (la côte sud-est) et l'Europe du Nord, alors que les précipitations extrêmes sont diminuées en Amérique du Nord et au Nord-Ouest de l'Afrique du Sud (faiblement). Pendant la phase La Niña, les précipitations extrêmes augmentent dans le nord de l'Amérique du Nord, en Afrique du Sud, en Australie et en Europe du Nord, alors que les précipitations extrêmes sont diminuées dans le sud de l'Amérique du Nord et en Inde du nord. La pluie décennale calculée pendant une phase forte El Niño (SOI = -20) ou La Niña (SOI = 20) est significativement différente, dans certaines régions, de celle calculée pendant une phase neutre (SOI = 0) (Figure VI). Par exemple, lors d'une phase forte El Niño, la pluie décennale peut augmenter de 50% en Amérique centrale, 40% sur la côte sud de la Chine, et de près de 20 % au centre de l'Amérique du Nord et au sud-est de l'Amérique du Sud. Une

diminution d'environ 15% est observée dans le nord-ouest des Etats-Unis et de 20 % aux Philippines. Lors d'un fort épisode de La Niña, la pluie décennale augmente d'environ 15% au nord de l'Amérique du Nord, en Europe du Nord et sur la région méditerranéenne, de 10% à 40 % (d'est en ouest) en Afrique du Sud, de plus de 20% dans le nord de la Chine, de plus de 40% dans l'est de l'Australie et de 60% dans l'ouest de l'Australie. Une diminution de 50 % est également observée au Mexique et de près de 25% dans le nord de l'Inde.

L'impact de l'oscillation ENSO se trouve être asymétrique dans beaucoup de régions, comme par exemple pendant la saison DJF, à l'ouest de l'Amérique du Nord, le sud-est de l'Amérique du Sud, l'est de la Chine, l'Australie et faiblement dans l'Europe du nord et le centre de l'Asie.

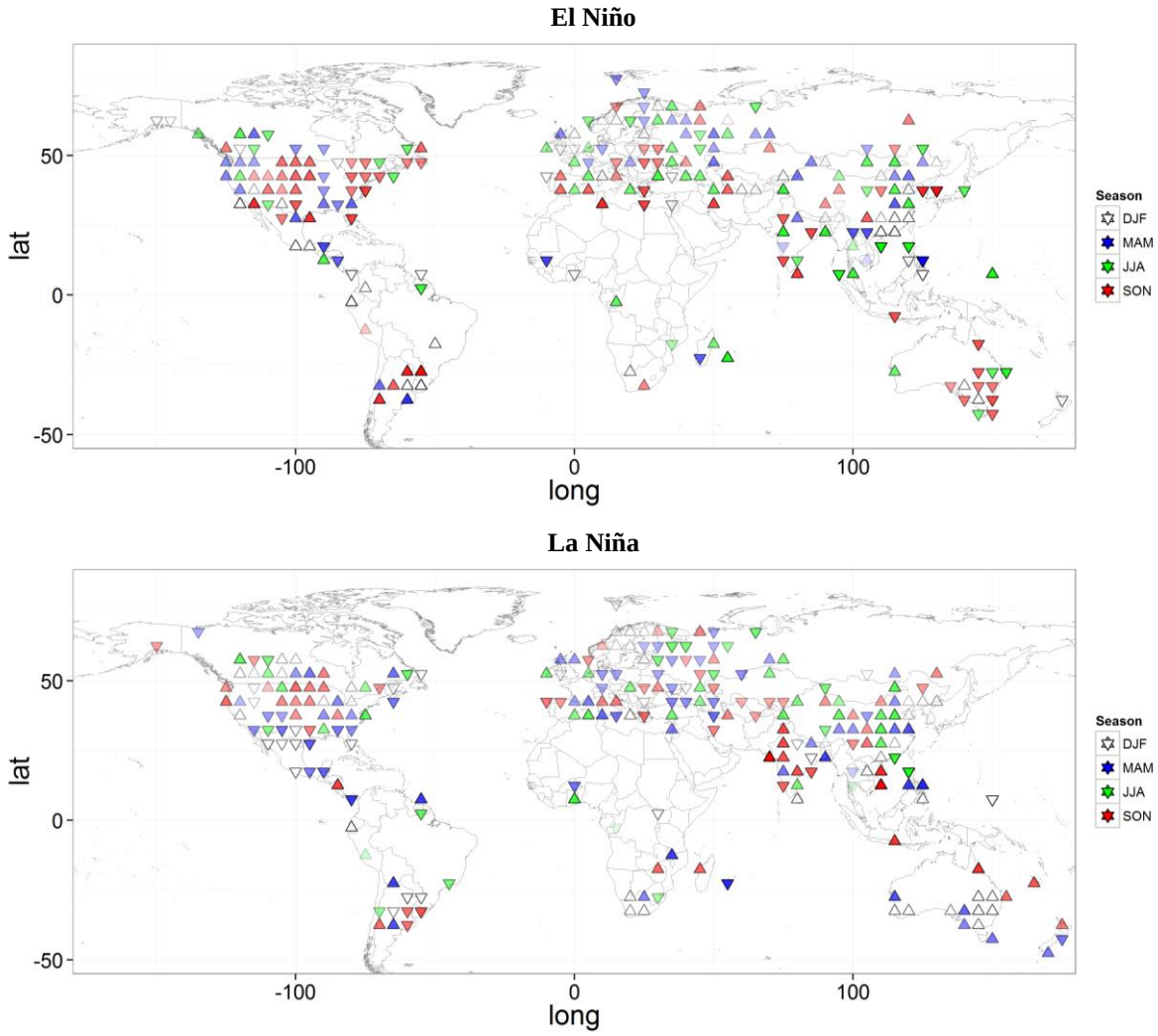
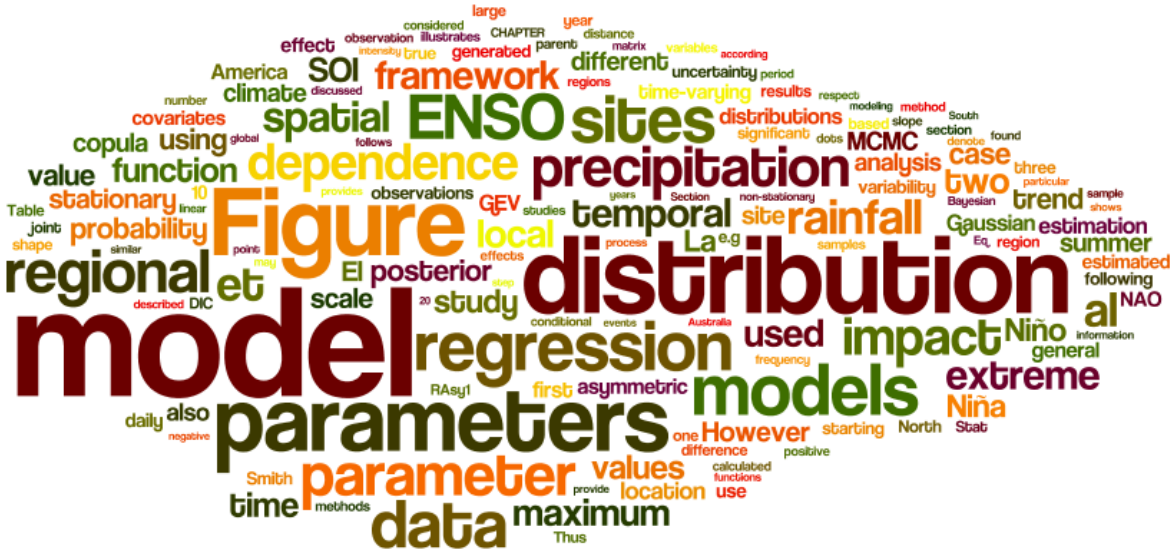


Figure VII–Carte des saisons pour lesquelles l'impact d'une phase El Niño (haut) ou La Niña (bas) est le plus fort.

Enfin, l'effet de l'oscillation ENSO est fortement variable suivant la saison considérée (Figure VII). En phase El Niño, la saison DJF est celle qui a le plus d'impact en Amérique du Sud et dans le sud-est de la Chine. La saison SON est la saison avec le plus fort impact ENSO en Amérique du Nord (augmentation des précipitations dans la partie est et diminution des précipitations à l'ouest), au sud de l'Amérique du Sud et dans la partie orientale de l'Australie

(diminution). Lors de la phase La Niña, le plus fort impact pendant la saison DJF est observé dans le nord de l'Amérique du Nord (diminution), le sud de l'Amérique du Sud (diminution), l'Afrique du Sud et l'Australie, alors que la saison MAM est la plus impactante au sud de l'Amérique du Nord (diminution) et au sud-est de la Chine. Dans la partie centrale de l'Amérique du Nord, le sud de l'Amérique du Sud (diminution) et le nord de l'Inde, la saison SON est celle qui a le plus d'effet. En général, l'impact de l'oscillation ENSO est faible pendant la saison JJA.


model
Figure
parameters
distribution
data
regression
regional
parameter
ENSO
sites
precipitation
dependence
framework
models
impact
rainfall
temporal
local
spatial
function
value
probability
stationary
copula
using
covariates
climate
America
SOI
effect
observation
illustrates
CHAPTER
considered
large
year
distance
matrix
parent
variables
according
generated
true
intensity
different
uncertainty
period
regions
time-varying
results
respect
modeling
method
South
section
found
denote
distributions
based
slope
significant
dobs
MCMC
case
three
particular
two
trend
sample
shows
Bayesian
analysis
variability
Gaussian
estimation
summer
region
Eq
estimated
following
NAO
al
Niño
general
information
extreme
Niña
North
Stat
starting
However
one
difference
positive
calculated
functions
use
location
values
Smith
provide
time
methods
maximum
Thus
daily
also
negative
RAsy1
first
asymmetric
used
frequency
conditional
events
Australia
study
scale
described
DIC
et
El
posterior
effects
process
La
e.g
samples
site
observations
provides
years
Section
non-stationary
GEV
studies
Table
joint
shape
similar
point
may
linear
10
20
Thus

Contents

Abbreviation

Abbreviation	Meaning
AIC	Akaike information criterion
AICc	Modified Akaike information criterion
BIC	Bayesian information criterion
cdf	Cumulative distribution function
DIC	Deviance information criterion
D-parameter	Distribution parameter
ENSO	El Niño Southern Oscillation
FA	Frequency Analysis
GCM	Global climate model
GEV	Generalized extreme value (distribution)
GP	Generalized Pareto (distribution)
iid	Independent and identically distributed
IOD	Indian Ocean Dipole
MCMC	Markov chain Monte Carlo
NAO	North Atlantic oscillation
pdf	Probability density function
POT	Peaks over threshold
pp-plot	probability-probability plot
RFA	Regional Frequency Analysis
R-parameter	Regression parameters
SEQ	Southeast Queensland
SOI	Southern Oscillation Index

Tables

Figures

CHAPTER 1 Introduction

1 General background

As reported by the Intergovernmental Panel on Climate Change *IPCC* [2007b], the climate system has changed since the 20st century. The global average surface temperature increased by 0.74°C during the last 100 years (1906-2005); the global average sea level raised with an average of 1.8 mm per year from 1961 to 2003; Northern Hemisphere mountain glaciers and snow cover also significantly decreased, with the maximum areal extent of seasonally frozen ground decreasing by nearly 7% since 1900, etc. As a consequence, those changes may impact the global water evaporation and circulation system, and hence precipitation and streamflow regimes.

Moreover, besides climate change, there is a substantial body of evidence that large-scale modes of climate variability also exert a significant influence on hydrological variables in various regions worldwide [*Henley et al.*, 2011]. For instance, the El Niño Southern Oscillation (ENSO) is one of the prominent modes of climate variability and has a global impact on hydro-meteorological variables [*Hoerling et al.*, 1997]; the North Atlantic Oscillation (NAO) controls the system of westerly winds and storm tracks across the North Atlantic to Europe [*Barnston and Livezey*, 1987]; and the Indian Ocean Dipole (IOD) is associated with significant temporal and rainfall variation over the Indian Ocean region and affects the Asian monsoon as well [*Saji et al.*, 1999].

Until recently, hydrologic studies were often based on the assumption of "stationarity"[1]. However, as will be reviewed subsequently, more and more evidence suggests that this assumption should be reconsidered in light of the influence of climate change/variability on hydrological variables (e.g. extreme precipitation). Recently, some hydrologists even declared that "*stationarity is dead*" and suggested abandoning the stationarity assumption in water-related design [*Milly et al.*, 2008]. Although the "death of stationarity" is still debated [*Lins and Cohn*, 2011], a common agreement is that providing reliable predictions of precipitation and streamflow variables requires incorporating the possible impacts of climate change and/or variability [*Stedinger and Griffis*, 2011].

This thesis provides an important step towards this goal, by developing a flexible frequency analysis framework that allows modeling the temporal variability of hydrologic variables, be it the consequence of climate variability or climate change.

1.1 Evidence of changes in precipitation

In hydrology, one of the main concerns with climate change is the change in the intensity or frequency of precipitation, especially the extreme precipitation, as well as their

[1]The precise definition of "stationarity" will be discussed in further depth in Section 2.

consequences (floods). During the 20th century, inter-decadal variations were observed for the global annual land mean precipitation, which also revealed a small upward trend of about 1.1 mm per decade (Figure 1.1, [*IPCC*, 2007a]).

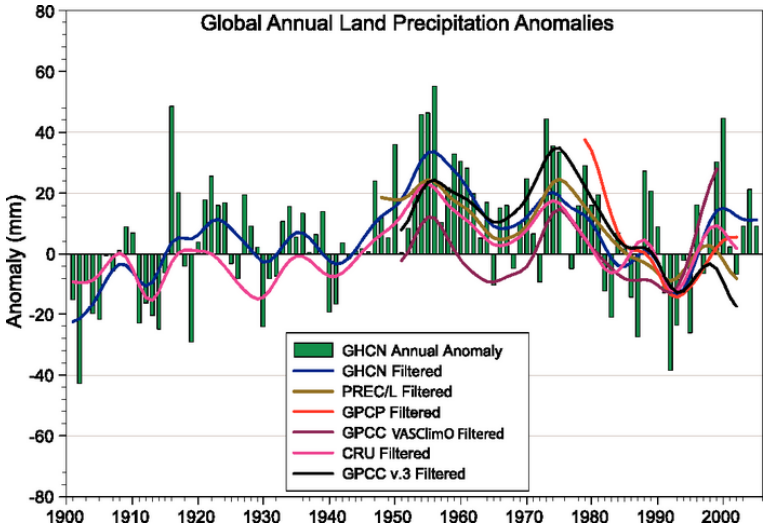


Figure 1.1-Annual global land precipitation anomalies (mm) for 1900 to 2005. The bar plot is with the Global Historical Climatology Network dataset. The smooth curves show decadal variations. (Figure source IPCC [2007b], figure 3.12)

Moreover, during the last decades, an increase in heavy rainfall has been reported in many regions worldwide. For example, the increase in extreme precipitation intensity was found in India [*Goswami et al.*, 2006], western China [*Zhai et al.*, 2005], north-eastern Italy [*Brunetti et al.*, 2001], the Czech Republic [*Kysely*, 2009], some parts of Australia [*Suppiah and Hennessy*, 1998], etc. In the U.S., *Kunkel et al.* [2010] described that heavy precipitation associated with tropical storms significantly increased. More recently, *Westra et al.* [2012] described that there are global increasing trends in annual maximum daily precipitation in most parts of the world based on the at-site observed extremes. This trend is likely to continue in the 21st century according to the regionally averaged precipitation of GCM projections (Figure 1.2, [*IPCC*, 2012]). Note however the strong discrepancy in the meaning of the word "extreme" between trend studies based on observations and projection studies: in the former case, "extreme precipitation" typically refers to annual maxima of station data, while in the latter case, it refers to precipitation spatially averaged over continental areas.

1.2 Evidence of impacts of climate variability on hydrologic variables

The impact of climate variability on hydrology is widespread. There exist many modes of climate variability that may influence hydrologic variables. However, in the following, we will restrict to reviewing some impacts of ENSO and NAO on hydrology.

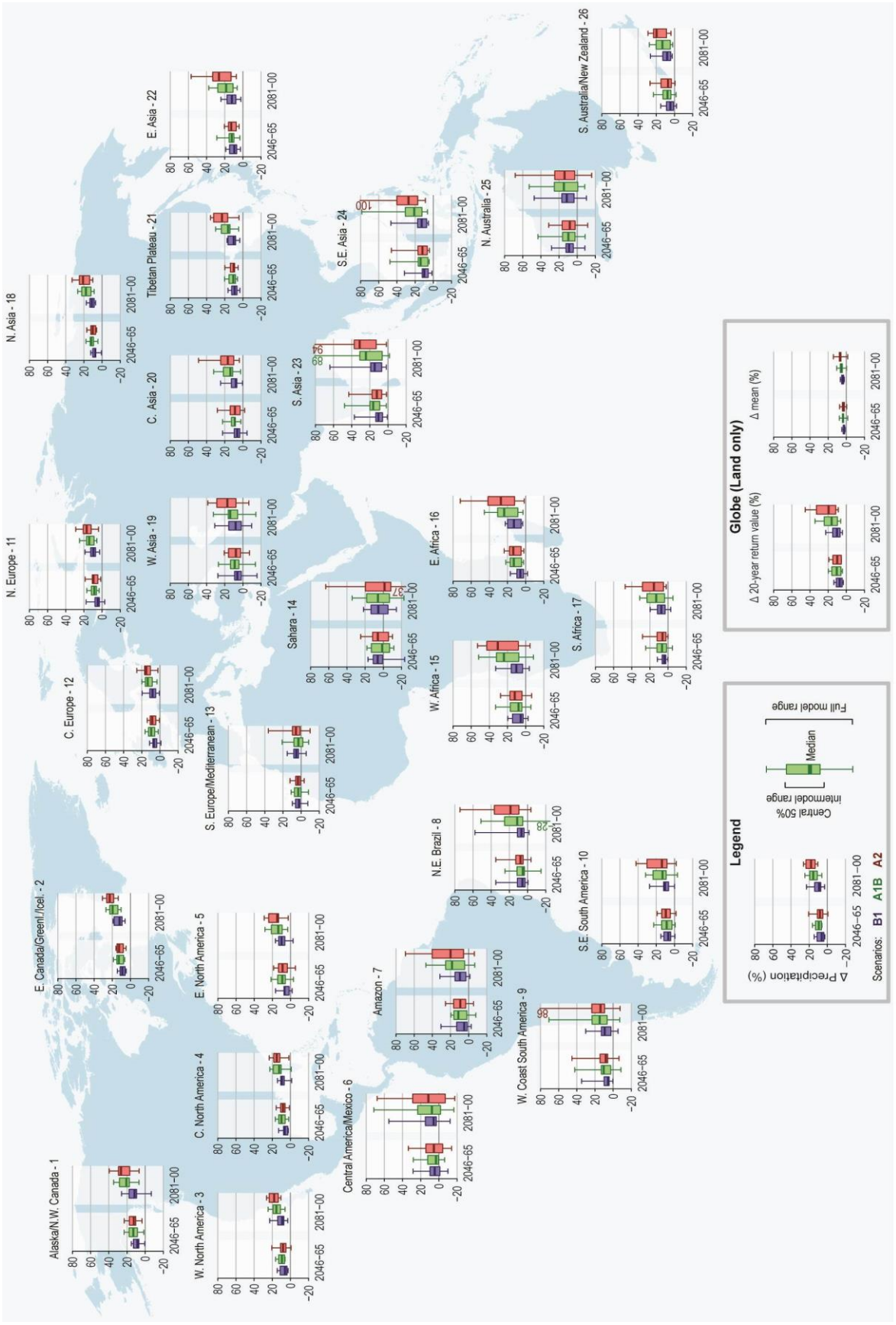


Figure 1.2-Projected changes (%) in 20-year return values of annual maximum 24-hour precipitation rates. Figure source: IPCC[2012], fig3-7a.

The ENSO is a climatic phenomenon in the tropical Pacific, which describes the variation of sea surface temperature (SST) anomalies in the tropical Eastern Pacific Ocean (see Figure 1.3). ENSO affects the atmospheric and oceanic circulations in the whole Pacific basin [*Kousky et al.*, 1984]. For example, during El Niño phase, warm humid air spreads from western Pacific to Eastern Pacific, which causes more precipitation in Eastern Pacific, while drought in Western Pacific. Figure 1.4 illustrates the impact of ENSO on the North America weather anomalies.

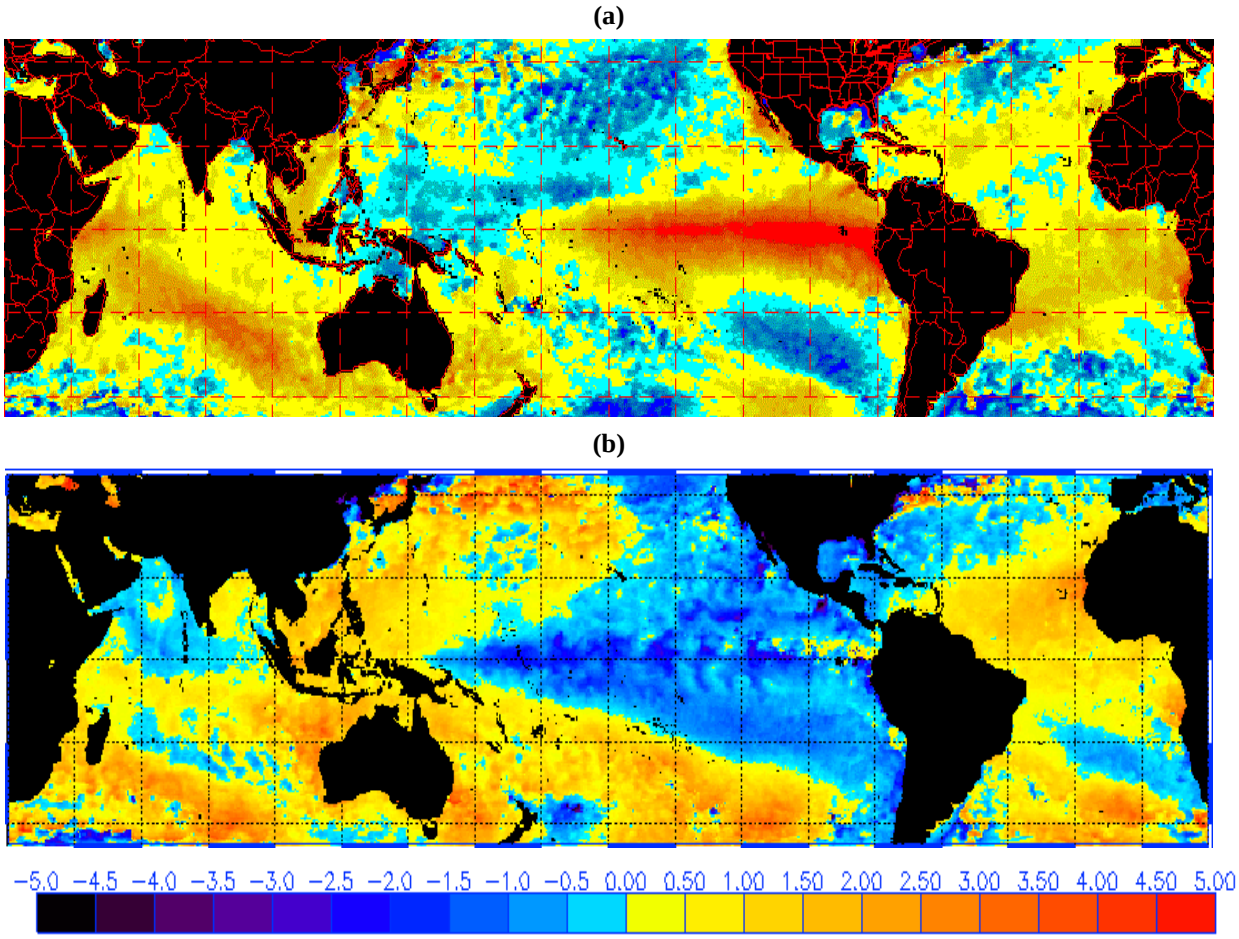


Figure 1.3-SST anomalies in degree Celsius (°C) during (a) El Niño 1998 and (b) La Niña 2010. Red (blue) denotes positive (negative) anomalies. Figure source: http://www.ospo.noaa.gov/Products/ocean/sst/anomaly/

ENSO is also considered as the most influential climate phenomenon producing global extremes of precipitation [*Dai et al.*, 1997], which leads to large precipitation anomalies in tropical area and influences precipitation patterns over Pacific, India and Atlantic Oceans [*Dai and Wigley*, 2000]. For example, in boreal winter, the impact of ENSO on precipitation variables was found in western U.S [*Castello and Shelton*, 2004; *Cayan et al.*, 1999; *Meehl et al.*, 2007], Southern South America [*Grimm and Tedeschi*, 2009], South Africa [*Kruger*, 1999; *Vanheerden et al.*, 1988], Southern China [*Wu et al.*, 2003] and Southeast Queensland, Australia [*Cai et al.*, 2010].

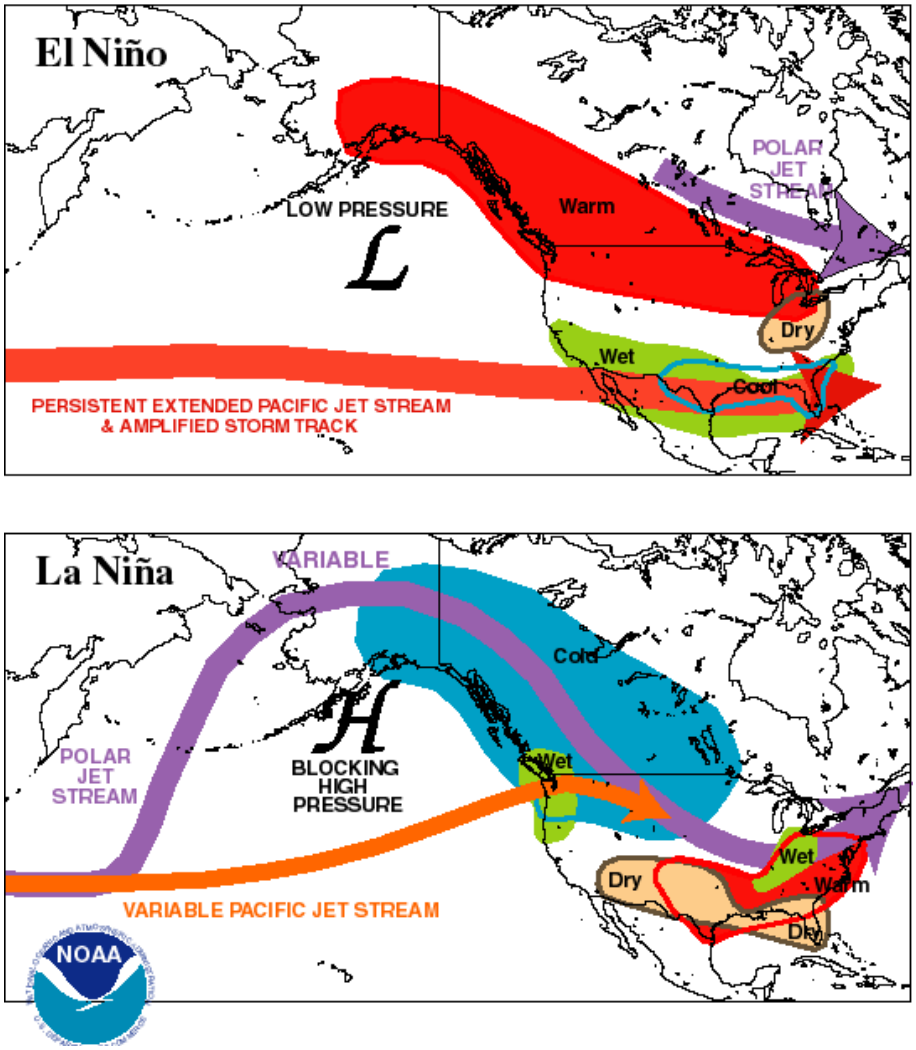


Figure 1.4-January-March weather anomalies during moderate to strong El Niño and La Niña. Figure source: http://www.srh.noaa.gov/tbw/?n=tampabayelninopage

The NAO is a climatic phenomenon in the North Atlantic Ocean, which describes the strength of the atmospheric pressure difference between the Icelandic low and the Azores anticyclone. The NAO is associated with the changes of strength and trajectory of North Atlantic storms [*Hurrell*, 1995] (see Figure 1.5) , and hence leads to changes in the pattern of temperature and precipitation in North America and Europe [*Hurrell and VanLoon*, 1997; *van Loon and Rogers*, 1978]. The NAO was found to affect the streamflow as well, for example, in Iceland [*Jónsdóttir et al.*, 2004], in Central Europe [*Kaczmarek*, 2003; *Limanówka et al.*, 2002; *Pociask-Karteczka et al.*, 2003], in Northern Europe [e.g. *Kiely*, 1999; *Kingston et al.*, 2006; *Stahl et al.*, 2001; *Wilby et al.*, 1997], and in the Iberian peninsula [*Trigo et al.*, 2004; *Vicente-Serrano and Cuadrat*, 2007].

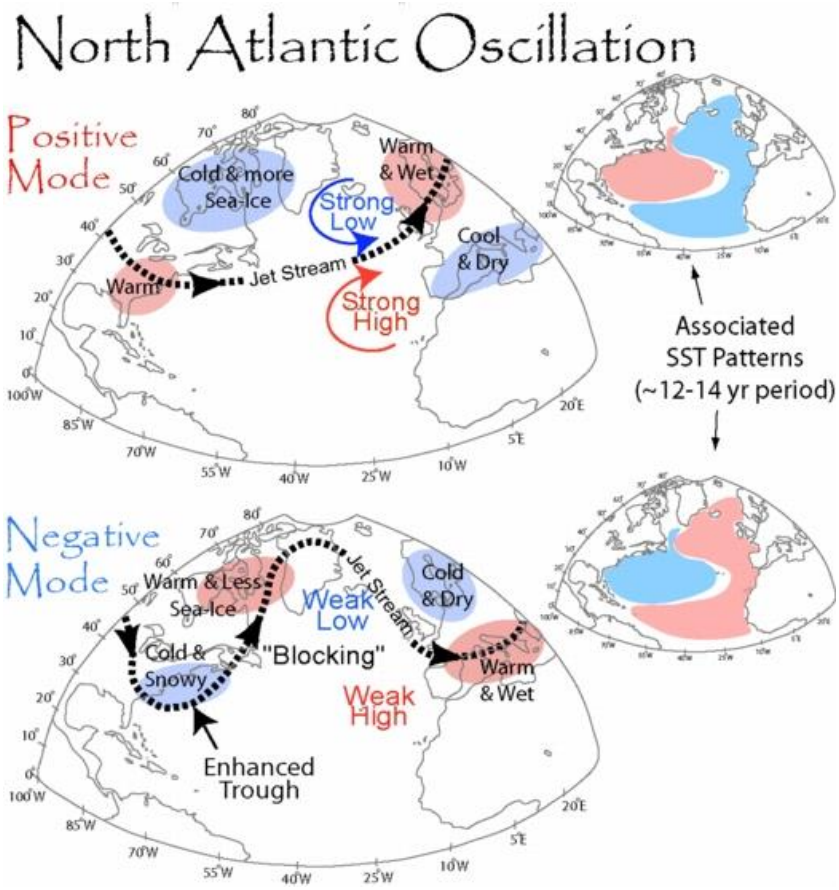


Figure 1.5-NAO weather pattern. Figure source: http://www.newx-forecasts.com/nao.html

2 On the notion of stationarity

The preceding sections illustrated that we should expect the distribution of hydrologic variables to vary with time, either as an effect of climate change or climate variability. In a probabilistic model, whether or not the distribution of random variables depends on time is related to the concept of stationarity. However, the precise definition of this term is sometimes garbled. This section therefore aims at clarifying the vocabulary that will be consistently used throughout this thesis, and in particular, to make the distinction between non-stationary and non-identically distributed variables.

According to *Brockwell and Davis* [2006], a random variable is strictly stationary if its distribution does not vary with time. However, in the context where a random variable Y may depend on the realization of some other random variable X, this definition is not precise enough: we need to specify *which* distribution this definition refers to. Note that this context is typically the one we will be interested in throughout this thesis: Y could for instance describe some precipitation variable, whose distribution may depend on some ENSO index X.

In a first step, we provide some simple illustrations using simulated data. Consider independent and identically distributed (iid) samples generated from a Normal distribution[2]

[2] Parameters μ and σ in the notation $N(\mu,\sigma)$ refer to the mean and the standard deviation.

$N(30,10)$, as illustrated in Figure 1.6(a). The parameters of the parent distribution are constant: they do not depend on time, either directly or indirectly through a time-varying covariate *X*. The samples in Figure 1.6(a) are therefore stationary and identically distributed.

Consider now Figure 1.6(b), which shows independent samples generated from the following Normal distribution:

$$Y_i \sim N(\mu_i, 10); \text{ where } \mu_i \sim N(30,5) \tag{1.1}$$

The samples in Figure 1.6(b) are not identically distributed, since the mean μ_i of the parent Normal distribution is different at each time step. However, Figure 1.6(b) suggests that they are realizations from a stationary distribution, in the sense that they do not display any deterministic trend in time.

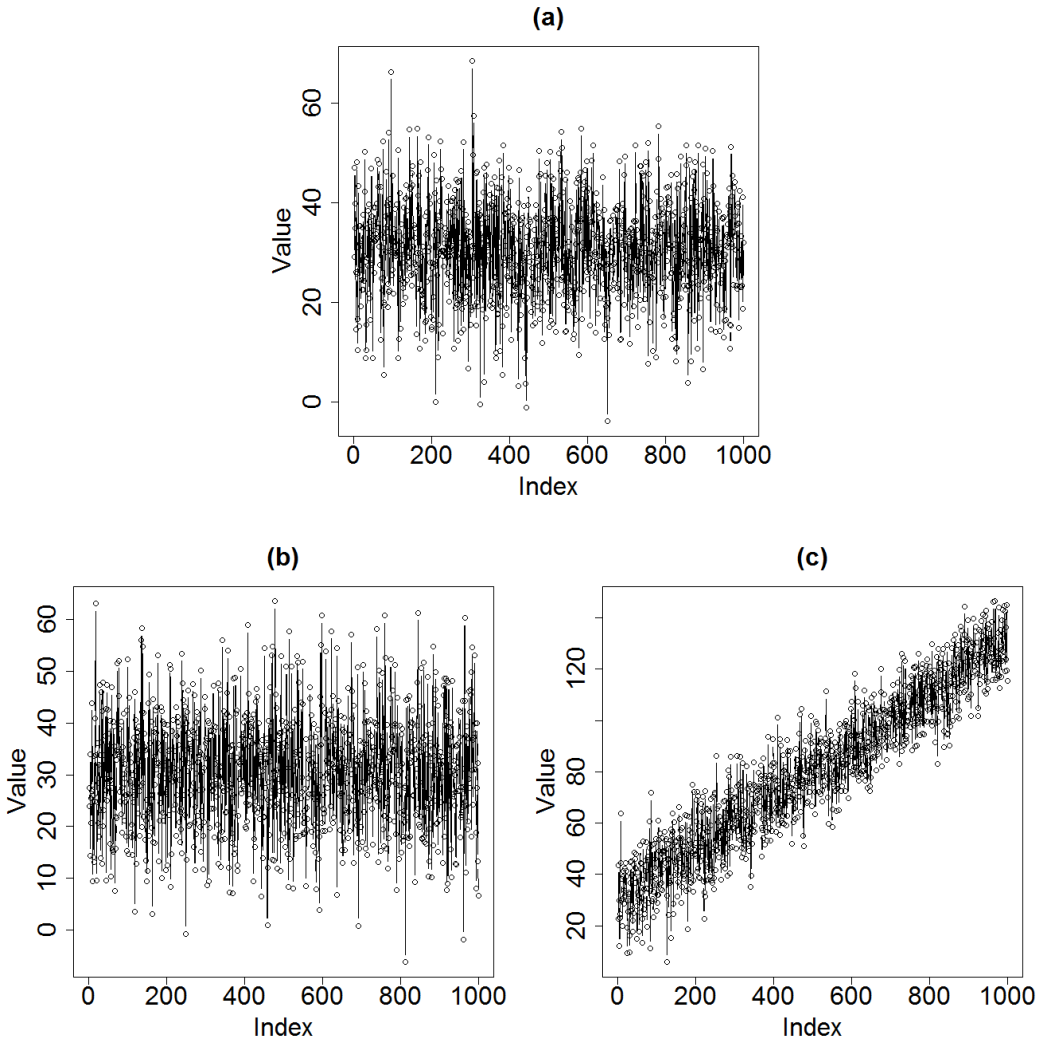


Figure 1.6-(a)Independent and identically distributed samples generated from a Normal distribution N(30,10). (b) Independent samples generated from Normal distributions N(μi,10), with μi generated from N(30,5). (c) Same as (b), but with μi generated from N(30+0.1i,5).

In order to formalize this distinction between non-identical distribution and stationarity, one needs to make a distinction between the *conditional* and the *marginal* distributions. Let $\boldsymbol{\beta}(t)$ denote the distribution parameters of the random variable $Y(t)$. The probability density function (pdf) of $Y(t)$ conditional on $\boldsymbol{\beta}(t)$ is denoted by $P_t(y|\boldsymbol{\beta})$. This distribution is termed the *conditional* distribution. In the example of Figure 1.6(b), this corresponds to the time-varying Normal distribution $N(\mu_i,10)$. The non-identically-distributed nature of samples in Figure 1.6(b) is (implicitly) a property of this conditional distribution.

By contrast, the marginal distribution is defined by integrating out the conditioning variable (or "un-conditioning"): if $\boldsymbol{\beta}$ is a random variable, the pdf of $\boldsymbol{\beta}$ at time *t* is denoted by $f_t(\boldsymbol{\beta})$. The *marginal* distribution can then be defined as:

$$P_t(y) = \int P_t(y|\boldsymbol{\beta}) f_t(\boldsymbol{\beta}) d\boldsymbol{\beta} \tag{1.2}$$

The property of stationarity is then related to the marginal distribution: the variable $Y(t)$ is stationary if its marginal distribution does not depend on time *t*. In the example of equation (1.1), some elementary algebra shows that the marginal distribution is a Normal distribution $N\left(30,\sqrt{5^2+10^2}\right)$: it indeed does not depend on time, which confirms that the samples in Figure 1.6(b), although non-identically-distributed (with respect to the conditional distribution), are realizations from a stationary (marginal) distribution.

Consider now the samples in Figure 1.6(c), generated from the following distribution:

$$Y_i \sim N(\mu_i,10); \text{ where } \mu_i \sim N(30+0.1*i,5) \tag{1.3}$$

For the same reasons as previously, those samples are not identically distributed. But this time, they are not stationary either: indeed, the marginal distribution at time *i* is a Normal distribution $N\left(30+0.1*i,\sqrt{5^2+10^2}\right)$. This distribution does depend on time, as illustrated by the upward trend visible in Figure 1.6(c).

As discussed in Section 1.2, hydrological events are affected by climate variability. In a probabilistic model, the observations can hence be considered as realizations from a time-varying distribution conditional on some climate (temporal) covariates. This describes the concept of climate-informed models. However, in the real life, it is difficult to know in advance whether or not such climate-informed models would be stationary, because the existence of a temporal trend in the climate temporal covariate is uncertain. For example, Figure 1.7 illustrates the NAO index in January for the period 1825-2010. In this period, there is no apparent temporal trend. Conversely, changes in the NAO pattern may appear in the future. It is therefore difficult to determine whether or not the marginal distribution described in Eq(1.2) depends on time, because in general the distribution of the conditioning variable of Eq(1.2) is unknown (and is not necessary to make conditional predictions). Therefore, as non-identically distributed variables are not necessarily non-stationary, we will favor the terminology "time-varying" or "climate-informed", rather than non-stationary models.

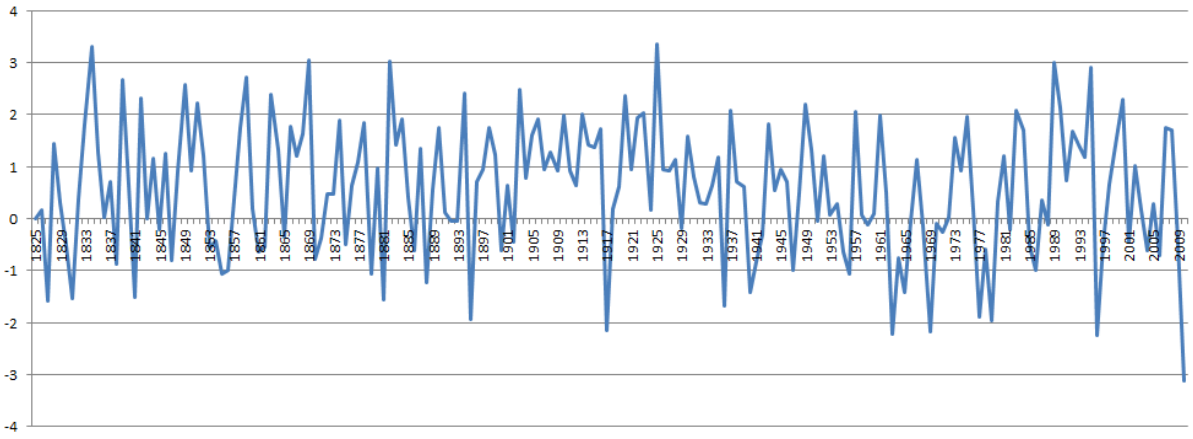


Figure 1.7-NAO index in January of period 1825-2010.

3 Frequency analysis models

In engineering design, Frequency Analysis (FA) techniques are an integral part of risk assessment and mitigation. FA uses statistical models to estimate the probability of hydrological events, which provides information for designing hydraulic structures. As reviewed subsequently, it is more and more important to understand the impact of climate variability/change on the severity and frequency of hydrological events, especially extremes.

3.1 Context of developed FA models

3.1.1 At-site methods with identically distributed variables

The standard at-site FA uses the at-site observations to estimate parameters from a pre-specified distribution. More formally, given a sample of observations $Y_1,...,Y_n$ that are assumed iid in most cases, the parameters $\boldsymbol{\beta}$ of a given distribution $D(\boldsymbol{\beta})$ are estimated using a particular estimation method (e.g. maximum likelihood, moment-based approaches, Bayesian estimation). A wealth of research has been carried out within this context during the last decades. Most studies focused on the choice of the parent distribution and of the estimation approach (e.g., *Durrans and Tomic* [2001]; *He and Valeo* [2009]; *Hosking et al.* [1985]; *Kroll and Stedinger* [1996]; *Lang et al.* [1999]; *Madsen et al.* [1997a]; *Meshgi and Khalili* [2009]; *Ribatet et al.* [2007]; *Sankarasubramanian and Srinivasan* [1999]), or the quantification of uncertainty (e.g. *Chowdhury et al.* [1991]; *Cohn et al.* [2001]; *Kysely* [2008]; *Stedinger* [1983]; *Stedinger and Tasker* [1985]; *Stedinger et al.* [2008]).

Besides the basic at-site FA based on estimating a pre-specified distribution, some model-based FA methods were also developed, based on models reproducing the main characteristic of hydrological variables, such as rainfall [*Arnaud and Lavabre*, 1999] and flood [*Boughton and Droop*, 2003; *Hundecha and Merz*, 2012]. Moreover, since a common problem of the at-site analysis is the relatively small data length, additional documentary sources on historical flood events, or data obtained from sediment deposits can be used to extent the data period. Historical and paleoflood data analysis were developed to deal with

these additional data (e.g *Naulet et al.* [2005]; *Neppel et al.* [2010]; *O'Connell et al.* [2002]; *Payrastre et al.* [2011]; *Reis and Stedinger* [2005]; *Stedinger and Cohn* [1986]).

3.1.2 At-site methods with time-varying variables

In the time-varying context, *Renard et al.* [2006a] and *Ouarda and El-Adlouni* [2011] discussed some new FA models by estimating time-varying parameters from a pre-specified distribution. With similar structures, *Rust et al.* [2009] discussed the seasonality of extreme precipitation in UK, *Kysely et al.* [2010] described the trends on daily temperature and *Nogaj et al.* [2006] analyzed the amplitude and frequency of extreme temperature. More generally, *Khaliq et al.* [2006] reviewed time-varying at-site FA methods. In addition to the analysis on the temporal variation, climate/weather information were also integrated to the analysis: for instance, *Micevski et al.* [2006] used the Inter-decadal Pacific Oscillation (IPO) to characterize the flood hazard in Australia; *Tramblay et al.* [2011] used various climate covariates to analyze the heavy rainfall in Southern France; and *Garavaglia et al.* [2010]; [2011]; *Paquet et al.* [2013] incorporated weather type information to quantify the rainfall hazard.

While at-site FA methods enabling the inclusion of climate information or non-stationarity are becoming common, such at-site models remain limited by two important drawbacks:

(1) Local analysis cannot be applied to ungauged sites.

(2) Uncertainty in parameter estimates (and hence predictive estimates) tends to be very large due to the limited number of observations in a local model. In addition, if climate information is included and more complex models are proposed, these observations may not be sufficient to identify the parameters [*Thyer et al.*, 2006].

This motivates the development of regional frequency analysis (RFA) models that use information from multiple sites to overcome these shortcomings.

3.1.3 RFA methods with identically distributed variables

In classical RFA methods, information from multiple sites is used to perform the inference, which may provide more precise estimations. More precisely, the basis of most RFA methods is to assume that some parameters are common to all sites within a given homogeneous region, or that the parameters can be predicted from a regression with site characteristics such as (for precipitation) elevation, distance to sea, etc.

Under the strict stationarity assumption, many RFA methodologies have been developed over the years (e.g. *Durrans and Kirby* [2004]; *Overeem et al.* [2008]; *Yu et al.* [2004]; *Cooley et al.* [2007]; *Madsen and Rosbjerg* [1997a]; [1997b]; *Madsen et al.* [1997b]; [2002] and *Ghosh and Mallick* [2011]). A comparison between regional and at-site approaches for extreme rainfall was performed by e.g. *Kysely et al.* [2011].

3.1.4 RFA methods with time-varying variables

In order to move beyond the assumption of strict stationarity, *Cunderlik and Burn* [2003] and *Leclerc and Ouarda* [2007] described non-stationary RFA models for flood analysis, and *Hanel et al.* [2009a] introduced a time-varying index-flood model for extreme precipitation. Recently, several authors (*Aryal et al.* [2009]; *Lima and Lall* [2010]; *Maraun et al.* [2010]; *Maraun et al.* [2011]; *Sang and Gelfand* [2009]) started investigating spatio-temporal models. In the same vein, *Gregersen et al.* [2013] also used Poisson regression models to describe the frequency of extreme rainfall in both space and time. A common difficulty for all these approaches is the treatment of the spatial dependence existing between data. This is also one of the main topics that will be discussed in this thesis.

3.2 FA methods for the extremes

With FA methods, extreme data can be characterized through appropriate distributions. In general, there are two different ways to extract extreme data [*Coles*, 2001], which correspond to two distribution families:

1. Block maxima: In each block of data (e.g. 1 year), only the maximum value is extracted. According to the extreme value theory [*Fisher and Tippett*, 1928], the corresponding distribution is the generalized extreme value distribution (GEV).
2. Peaks-over-threshold: All values exceeding a certain threshold are extracted. The corresponding distribution is the generalized Pareto distribution (GP) [*Balkema and De Haan*, 1974; *Pickands III*, 1975].

3.2.1 Block maximum

In a block of n iid random variables $\left(Z_i\right)_{i=1,n}$, the maximum is denoted by $Y_n = \max_{i=1,n} Z_i$.

The extreme value theory [*Fisher and Tippett*, 1928] provides an asymptotic distribution for Y_n, when n tends to infinity. This distribution is called the generalized extreme value (GEV) distribution, whose cumulative distribution function (cdf) is given by:

$$G(y \mid \mu, \sigma, \xi) = \exp\left(-\left[1 - \xi\left(\frac{y-\mu}{\sigma} \right) \right]^{1/\xi} \right) \tag{1.4}$$

where μ, σ and ξ denote the location, scale and shape parameters respectively. There are three different families in terms of the shape parameter ξ known as, respectively, Gumbel, Weibull and Fréchet families for zero, positive and negative values of ξ[3]. These cdfs are shown below:

[3] In the hydrologic literature, the shape parameter ξ has generally the opposite sign as in the statistical literature. The hydrologic convention is used throughout this thesis.

Gumbel distribution (ξ = 0):

$$G(y \mid \mu,\sigma) = \exp\left\{-\exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\right\} \quad \text{for } y \in \boldsymbol{R} \tag{1.5}$$

Weibull-type distribution ($\xi > 0$):

$$G(y \mid a,b,\xi) = \begin{cases} \exp\left\{-\left[-\left(\frac{y-b}{a}\right)\right]^{1/\xi}\right\} & y < b \\ 1 & y \geq b \end{cases} \tag{1.6}$$

where $a = \frac{\sigma}{\xi}$ and $b = \mu + \frac{\sigma}{\xi}$.

Fréchet-type distribution ($\xi < 0$):

$$G(y \mid a,b,\xi) = \begin{cases} 0 & y \leq b \\ \exp\left\{-\left(\frac{y-b}{a}\right)^{1/\xi}\right\} & y > b \end{cases} \tag{1.7}$$

where $a = -\frac{\sigma}{\xi}$ and $b = \mu + \frac{\sigma}{\xi}$.

3.2.2 Peaks over threshold

Let $Z_1, Z_2, \ldots$ be a sequence of iid random variables. Suppose that the block maxima $Y_n = \max_{i=1,n} Z_i$, satisfies Eq(1.4) for large *n*. Then for a large enough threshold *u*, the cdf of the threshold exceedances *Z-u*, conditional on *Z>u* is asymptotically (when *u* tends to infinity):

$$\mathrm{P}\left(Z-u > x \mid Z > u\right) = \begin{cases} 1-\left(1-\frac{\xi x}{\tilde{\sigma}}\right)^{1/\xi} & \text{for } \xi \neq 0 \\ 1-\exp\left(-\frac{x}{\tilde{\sigma}}\right) & \text{for } \xi = 0 \end{cases} \tag{1.8}$$

where ξ is the same as in Eq(1.4) and $\tilde{\sigma} = \sigma - \xi(u-\mu)$.

This family of distribution is called the generalized Pareto [*Balkema and De Haan*, 1974; *Pickands III*, 1975].

3.2.3 The notion of return period in the time-varying context

In risk analysis and assessment, the rarity of an extreme event is generally quantified through a return period, rather than in terms of a probability of exceedance. In the context of identically distributed variables, the "*T* year event" corresponds to the event with non-exceedance probability $p = 1 - 1/T$. A 10-year event has therefore a probability 1/10 to be exceeded every year. However, in the time-varying context, such simple equivalence does not hold any more, because the distribution varies with time. *Renard* [2006] and *Cooley* [2013] discussed two definitions of the return period in the time-varying context: (i) the return period associated with a value *Q* is the expected waiting time between two successive exceedances of level *Q*, and (ii) the return period associated with a value *Q* is computed so that the expected number of exceedances of level *Q* in the next *T* years is equal to one. While both definition are equivalent in an identical distribution context, and indeed correspond to an annual probability of exceedance equal to 1/T, this is not the case if a temporal trend exists. The interpretation of "*T* year events" is much less direct than under the strict stationarity assumption. An alternative quantification of the rarity of an event uses the concept of "failure probability". The failure probability associated with a level *Q* and a duration *n* is defined as the probability that at least one event exceeds *Q* during the next *n* years. Typically, *n* can be considered as, the lifetime of a hydraulic structure. This concept is much easier to interpret than return periods outside of the stationary context. Moreover, it allows investigating questions such as "what should be the capacity of a dam to ensure that the probability of exceeding the dam capacity during its lifetime of *n* years is less than 0.01?" A more thorough discussion of return periods and failure probabilities can be found in *Salas and Obeysekera* [2013].

The concept of return period is not easier to handle with climate-informed models. Indeed, as explained in Section 2, the stationarity of climate-informed models is generally unknown. Moreover, even if the marginal distribution turned out to be stationary, we may still be interested in computing conditional probabilities of exceedance. It may be interesting to know how different the extreme events are during different climate conditions. For instance, if the main resource of a river is snowmelt, the probability of observing a large snowmelt flood will surely be different if one knows that temperatures are particularly high or low. This refers to the return level based on the conditional probability. More precisely, if the cdf of a random variable *Y* is $F(y \mid \boldsymbol{\beta})$, then the "*T* year event" conditional on $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ corresponds to the event with non-exceedance probability $p = 1 - 1/T = F(y \mid \boldsymbol{\beta}_0)$. A 10-year event conditional on $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ has therefore a conditional probability 1/10 to be exceeded every year for which the climate condition satisfies $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

Despite these difficulties in interpreting return periods in a non-identically-distributed context, their use is so deeply ingrained in engineering practice that most users expect a quantification in terms of return period, even in a non-stationary context. As an illustration, the IPCC quantifies projected changes in extreme precipitation by computing the future return period of today's 20-year quantiles [*IPCC*, 2012]. Consequently, some results in this thesis will sometimes be reported in terms of return period, and they should be interpreted in terms of exceedance probability.

4 Main contributions

The main contribution of this thesis is the construction of a rigorous regional spatio-temporal framework that extends the usage of FA techniques to the time-varying context. In addition, this framework enables the quantification of temporal trends and impacts of climate variability on the severity/frequency of hydrological events. This framework provides a general and flexible modeling platform by integrating several separately developed components, such as spatio-temporal regression models, copula-based modeling of spatial dependence, Bayesian inference, model comparison tools. .

4.1 Objectives

The precise objectives of this thesis can be described as follows:

1. Model Development, inference and comparison: the construction of the model, using regressions with spatial and temporal covariates to describe the spatio-temporal variability of the parameters, is described. Inference accounts for spatial dependence between data and uses a Bayesian framework, thereby enabling a direct quantification of estimation and predictive uncertainty. In addition, within this general framework, different climate-informed regression models can be compared (for instance, linear vs. non-linear regression). This helps identifying the most suitable regression to link climate variability and spatio-temporal hydrological variability.

2. Model Assessment: the usefulness of the modeling framework is assessed using synthetic and real-world case studies aimed at: (i) assessing the consistency and difference between time-varying FA models and their identically-distributed counterpart. (ii) Evaluating the importance of considering spatial dependence. (iii) Comparing different spatial dependence structures (copula vs. maximum stable process) in terms of joint and conditional probabilities estimation.

3. Model Applications: Two case studies aimed at quantifying the ENSO impact on precipitation are illustrated to highlight the usefulness of the framework.

   a) Quantify the ENSO impact on the summer total and extreme rainfall in Southeast Queensland (SEQ), Australia. The flexibility of the framework enables several competing hypotheses to be rigorously compared, thereby addressing the following questions:
      i) Does ENSO have an impact on summer maximum daily rainfall?
      ii) Is the impact of ENSO on summer maximum daily rainfall asymmetric (i.e., distinct impacts during El Niño and La Niña episodes)?

   b) Assess the impact of ENSO on the intensity of global seasonal extreme precipitation. This analysis is not based on a gridded dataset, but instead on a new global high

quality observation dataset (HadEX2). We focus on analyzing the at-site extremes by using a climate-informed regional frequency analysis (RFA) framework. There are three objectives in the study:

i) to identify the regions affected by ENSO and quantify its impact on extreme precipitation quantiles (e.g., 10 or 100-year precipitation)
ii) to evaluate the possible asymmetry of the impact of ENSO
iii) to describe the seasonality of ENSO impacts.

4.2 Values of the thesis from an engineering perspective

A common question facing engineers before building some hydraulic structure is to design it in order to ensure that its failure probability remains acceptably low. This question is especially complex in the context of climate change. Evidently, it is dangerous to extrapolate any trend fitted on past observations, because results strongly depend on the formulation of the trend [*Cooley*, 2013]. Instead, a common approach is to use GCM/RCM outputs, and then to use standard iid models for different present/future sub-periods [*Brigode*, 2013; *Madsen et al.*, 2009]. However, choosing the length of the sub-periods is a limitation: it results from a tradeoff between a period short enough to make the stationarity approximation acceptable and a period long enough to have reasonable uncertainty. The time-varying framework developed in this thesis provides a reasonable solution for this tradeoff, which enables providing more precise information for future hydraulic constructions.

On the other hand, the occurrence of many extreme events is linked to the climate variability. When a strong link exists between the climate state and hydrologic events (e.g. heavy rainfall), it is more likely to have exceptional events under some particular climate condition (e.g. strong El Niño or La Niña). One advantage of a time-varying framework is that a quantitative link can be established between the hydrologic event and the climate state. Even if hydraulic constructions were built under a strictly stationarity assumption, planners/water resource managers can at least know how much the probability of large events will be increased under a certain climate condition. Armed with this knowledge, planners/water resource managers would be able to undertake better planning of emergency response, and potentially improve reservoir operating rules to better control floods, and reduce the impact of the events during some particular climate condition.

5 Organization of the thesis

This thesis is organized as follows: Chapter 2 illustrates the construction of time-varying models at the local scale. Chapter 3 presents the usage of the general framework developed in Chapter 2. Two case studies are discussed in this chapter. The first one focuses on GCM-predicted hydrological variables in the Durance catchment, and compares a continuously time-varying model with an iid model for present/future sub-periods. The second case study analyzes extreme precipitation in the French Mediterranean area, and assesses the existence of trends and NAO impacts. Chapter 4 then describes the construction of time-varying models at

the regional scale. The advantage of using regional parameters for detecting weak signals is highlighted. In Chapter 5, the importance of considering spatial dependence is discussed. In particular, copulas and maximum stable processes are compared in this chapter. Chapter 6 and Chapter 7 present the applications of the regional time-varying framework to quantify the impact of ENSO on precipitation in Australia (Chapter 6) and at the global scale (Chapter 7). A brief summary of the main results and some avenues of further extensions are described at the end of the thesis.

Part I Time-varying frequency analysis framework: Local model

CHAPTER 2 Development of a general time-varying modeling framework at the local scale

The objective of this chapter is to develop a general frequency analysis framework that allows considering temporal trends and/or effects of climate variability indices. The framework intends to provide a very flexible platform to consider different hypotheses for various dataset. The parent distribution of data can be any discrete or continuous distribution. Both deterministic variables (e.g. time) and stochastic variables (e.g. climate indices) can be used as covariates. Based on different hypotheses, flexible regression functions are used to establish the relationship between covariates and parent distribution parameters. Moreover, the generality of the framework is highlighted for the tools around the model as well. Parameter inference tools, diagnostic tools and model selection tools are all integrated in this framework. In this chapter, we focus on at-site (local) models as the first step of a more general framework.

1 Local model construction

In the standard FA context, observations are assumed to be identically distributed, i.e. they follow a common distribution with constant parameters. However, in both the climate-informed and non-stationary contexts, the assumption of identical distribution does not hold any more. Therefore, we assume that, in this framework, observations follow a time-varying distribution, in which time-varying covariates are used to induce temporal variations of the parameters.

Figure 2.1 illustrates the general principle of the local model. Observations $\boldsymbol{Y}$ follow a distribution with time-varying parameters $\boldsymbol{\beta}$. These time-varying parameters are modeled with temporal regressions which are functions of time-varying covariates $\boldsymbol{x}$ and regression parameters $\boldsymbol{\theta}$. This construction provides us with a flexible and convenient framework to model the effect of climate variability and/or temporal trends on the observations. The following sections describe each building block of the local model in more details.

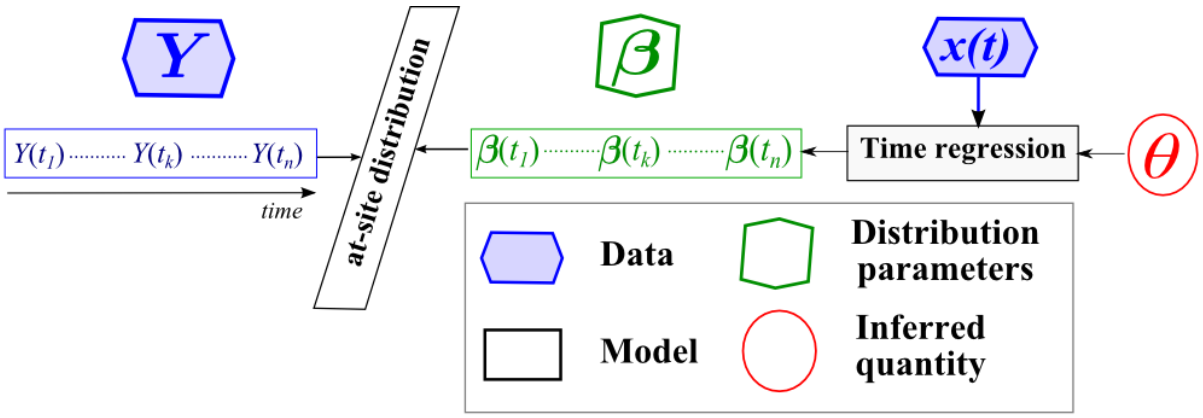


Figure 2.1- Schematic of the Local model.

1.1 Parent distribution for local model

Let $Y(t)$ denote the observation at time t and $\boldsymbol{Y} = \left(Y(t_1), Y(t_2), ..., Y(t_n)\right)$ [4] denote the collection of observations at n time steps. A local model starts by defining a parent distribution for the observations:

$$Y(t) \sim D(\boldsymbol{\beta}(t)) \tag{2.1}$$

where D is the assumed distribution of Y and $\boldsymbol{\beta}(\boldsymbol{t}) = \left(\beta_1(t), \beta_2(t), \ldots, \beta_m(t)\right)$ is the collection of all m distribution parameters at time t ($m=2$ for a Gaussian distribution, m=3 for a *GEV* distribution, etc.). Note that there is no particular restriction on the choice of the distribution D: in particular, both continuous and discrete distributions can be considered.

[4] In this thesis, bold letters denote vectors.

1.2 Regression with temporal covariates

The parameters $\boldsymbol{\beta}$ directly characterize the parent distribution *D*, such as its location, scale and shape. These parameters may depend on some time-varying covariates, like time, atmospheric pressure or some climate indices. Thus a regression function is defined for each component of the parameter vector $\boldsymbol{\beta}$ as follows:

$$\beta_i(t) = l_i^{-1}\left(h_i(\boldsymbol{x}(t);\boldsymbol{\theta}_i)\right) \quad i = \{1, 2, ..., m\} \tag{2.2}$$

where h_i is the regression function for the i^{th} component $\beta_i(t)$, $\boldsymbol{x}(t)$ is the collection of temporal covariates and $\boldsymbol{\theta_i}$ is the collection of all parameters used in the regression function h_i. l_i^{-1} is the inverse link function, which establishes a one-to-one mapping between $\beta_i(t)$ and the regression function.

To avoid confusion with the *D*-parameters $\boldsymbol{\beta}(\boldsymbol{t})$, we call $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_m})$, the parameters we are going to estimate, the regression parameters (R-parameters). Note that in Figure 2.1, the inverse link function is made implicit in the time regression model.

There is usually no particular restriction on the regression functions. The most common choice will be to assume that a *D*-parameter linearly depends on the covariates. Such linear form is a very simple and straightforward hypothesis, whereas real cases are certainly more complex. One simple example of a linear model is the linear trend on time: $\beta_i(t) = \theta_1 + \theta_2 t$. Moreover, a linear model could also be used to describe the effect of climate variability. Instead of time *t*, the covariate could be some climate index, for example the Southern Oscillation Index (*SOI*), which is a descriptor of ENSO. The linear model would therefore become $\beta_i(t) = \theta_1 + \theta_2 * SOI(t)$. Lastly, different covariates could be considered together. For instance, it could be assumed that a *D*-parameter is affected by both a temporal trend and the effect of *SOI*, yielding: $\beta_i(t) = \theta_1 + \theta_2 t + \theta_3 SOI(t)$

Non-linear regression models can also be used. For instance, a cosine regression model could describe a periodicity in the *D*-parameter (due to seasonality for instance), such as $\beta_i(t) = \theta_1 \cos(\theta_2 t)$. Of course, the time *t* in the formula could also be replaced by any other time-varying covariate.

The last building block of the regression model is the inverse-link function. It is mostly used to set some restrictions on the range of parameters. For example, in order to ensure that a standard deviation parameter is positive, an exponential function is often applied. In order to obtain values between 0 and 1 (e.g. the success probability parameter for a binomial distribution), an inverse-logit function is often used. Table 2.1 lists some common inverse-link functions.

The choice of these models for different parameters is strongly influenced by the nature of these parameters. For instance, for a *GEV* distribution, a linear model may operate well for the location parameter, but it may not be advisable to apply it to the shape parameter due to the uncertainties.

Table 2.1-Standard inverse link functions

Inverse link function	Formula	Usage
Identity function	$l^{-1}(x) = x$	
Exponential function	$l^{-1}(x) = exp(x)$	Ensure positivity of a scale parameter (e.g. standard deviation σ in a Gaussian distribution)
Inverse function	$l^{-1}(x) = \frac{1}{x}$	Switch from intensity to rate (e.g. λ in a Poisson distribution)
Inverse-logit function	$l^{-1}(x) = \frac{exp(x)}{exp(x) + 1}$	set probability parameters between 0 and 1 (e.g. p in a Binomial distribution)

1.3 An illustration of local model construction

Assume that we want to analyze the seasonal maximum of daily precipitation data $\boldsymbol{Y} = (Y(t_1), Y(t_2), ..., Y(t_n))$. As suggested by *Katz et al.* [2002] and *Coles et al.* [2003], we assume that $\boldsymbol{Y}$ follows a *GEV* distribution with time-varying parameters $\boldsymbol{\beta(t)} = (\beta_1(t), \beta_2(t), \beta_3(t)) = (\mu(t), \sigma(t), \xi(t))$.

For the location parameter $\mu(t)$, we want to assess the existence of a temporal trend and an effect of *SOI* (Figure 2.2). This can be implemented by assuming a linear regression for $\mu(t)$ (with the identity as inverse link function). For the scale and shape parameters, in this example, they are assumed to be constant in time. We use an exponential inverse link function for the scale parameter, thus the R-parameter θ_4 is defined from $-\infty$ to $+\infty$. Therefore, the regression parameters to be estimated in this model are $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$.

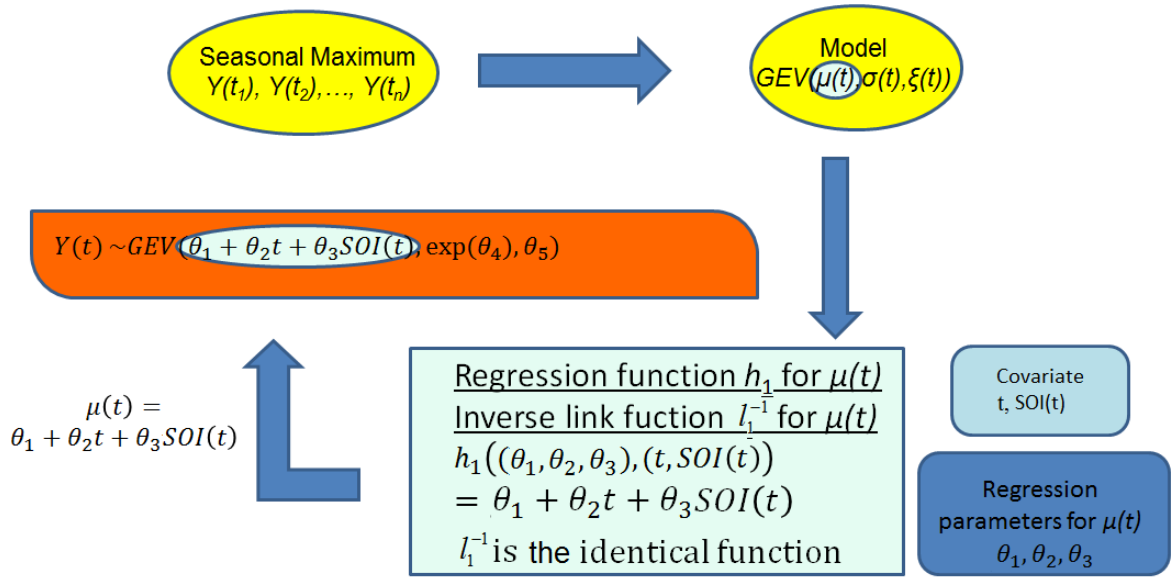


Figure 2.2-Schematic for the construction of regressions.

1.4 Relationship with other modeling frameworks

Other modeling frameworks are very similar to the regression models introduced in this section, in particular Generalized linear model (GLM) [*Nelder and Wedderburn*, 1972] and Generalized Additive model (GAM) [*Hastie and Tibshirani*, 1986]. The GLM is a generalization of the classical linear model, in which the distribution of data is not restricted to the Normal distribution. However, this model is still restricted to linear regression functions, whereas non-linear regressions can be introduced without difficulty in the framework developed here.

In the GAM, the data $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ are assumed to be realizations from a parent distribution whose mean satisfy $E(Y) = l^{-1}(\theta + h_1(x_1) + h_2(x_2) + \ldots + h_m(x_m))$, where $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$ are the covariates. The regression functions $h_i(x_i)$ could be estimated by both parametric and non-parametric means. The regression model used in this thesis could be considered as a particular case of GAM, because we restrict to parametric settings for two reasons: (i) parametric models are easier to regionalize than non-parametric ones; (ii) parametric models are well-suited to Bayesian inference, which will be favored in this thesis. Also note that in the GAM, the regression only applies to the expectation of the random variable. This has been generalized by introducing GAM for location, scale and shape parameters (GAMLSS) [*Rigby and Stasinopoulos*, 2005].

2 Posterior distribution and parameter inference

In this section, we are going to discuss statistical tools for parameter inference integrated in the framework. Parameters are estimated in the Bayesian framework with the help of MCMC methods. Once parameters have been estimated, quantiles and related uncertainties can be calculated.

2.1 Parameter estimation methods

Several methods can be used to estimate the parameters, such as the maximum likelihood method [*Lecam*, 1990], L-moments method [*Hosking*, 1990] or the Bayesian method [*Berger*, 1985]. The Bayesian method is chosen in this thesis for the following reasons [*Renard et al.*, 2013].

1. The Bayesian framework is a general framework that can be easily applied to estimate parameters for any parent distribution and regression functions.
2. Prior information of parameters is involved, which could provide additional information for parameter estimation.
3. The uncertainty (credibility interval) is naturally and directly obtained through the posterior distribution, thus there is no need for asymptotic approximations of the sampling distribution of estimates.

4. Once the Markov chain Monte Carlo (MCMC) methods are properly and carefully implemented, parameter estimation in the Bayesian framework is very convenient. They can be applied to a wide range of problems.

2.1.1 Bayesian Inference

The Bayesian method is a statistical inference procedure that is used to estimate the posterior distribution of parameters in the proposed model. Instead of assuming parameter have deterministic but unknown values, the Bayesian method assumes that parameters are random variables following a "prior" distribution. This prior information is then updated based on the information brought by the data, yielding the posterior distribution of parameters.

More precisely, denote by $\boldsymbol{y}$ the collection of observations and $\boldsymbol{\theta}$ the collection of all parameters. The likelihood function of $\boldsymbol{y}$ conditional on $\boldsymbol{\theta}$ is $f(\boldsymbol{y}|\boldsymbol{\theta})$. Then the posterior pdf is computed by:

$$f(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\boldsymbol{y})} \tag{2.3}$$

where $f(\boldsymbol{y}) = \int f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}$.

Usually it is difficult to compute ***f(y)***, since there doesn't always exist explicit formulas to compute the integral for large dimensional $\boldsymbol{\theta}$. However, as the observations $\boldsymbol{y}$ are all fixed values, ***f(y)*** is a constant with respect to parameters $\boldsymbol{\theta}$. Consequently, the Bayes theorem is often written with an unnormalized posterior density as follows:

$$f(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{\theta})f(\boldsymbol{y}|\boldsymbol{\theta}) \tag{2.4}$$

where $\propto$ is the symbol for proportionality.

Compared with other approaches for parameter estimation, the Bayesian method can directly obtain estimation uncertainties, whereas other approaches require additional assumption (e.g. asymptotic normality) on the distribution of estimated parameters to provide uncertainties.

2.1.2 Markov chain Monte Carlo (MCMC) method

The Markov chain Monte Carlo (MCMC) method is an iterative method that is often applied in the Bayesian framework to generate samples from the posterior distribution of parameters. These samples are then subsequently used to approximate the posterior distributions and derived quantities (using e.g. histograms of MCMC samples, posterior mean/median/variance etc.). In general, the MCMC method can help to efficiently obtain an asymptotic estimation for the target distribution when facing the following difficulties, which are also typically present in the Bayesian framework:

1. The target distribution is multidimensional. It is hard to find an implicit formula for the pdf or cdf.

2. The target distribution is un-normalized.
3. The target distribution does not belong to a standard distribution family (e.g. Gaussian, Exponential, etc.).

The idea of MCMC is to generate random walks to obtain target distribution samples. The general procedure is described as follows:

1. Choose a starting point.
2. Generate another sample according to some proposal distribution that is conditioned on the previous sample.
3. Accept or reject the new sample using some acceptance rule.
4. Redo the previous two steps and collect a large number of samples.
5. Use the collected samples to approximate the posterior distribution.

Figure 2.3 illustrates parameter estimation with the MCMC method in a two-dimensional parameter space. The starting point of $\boldsymbol{\theta}^{(0)} = \left(\theta_1^{(0)}, \theta_2^{(0)}\right)$ is (10, 5), and the chain converges around the value of (0, 2.5) in the end. The points in the convergence region are used to estimate the posterior distribution of these two parameters. The points generated prior to convergence are considered belonging to a burn-in period, which will be removed from the simulated chain used to approximate the posterior distribution.

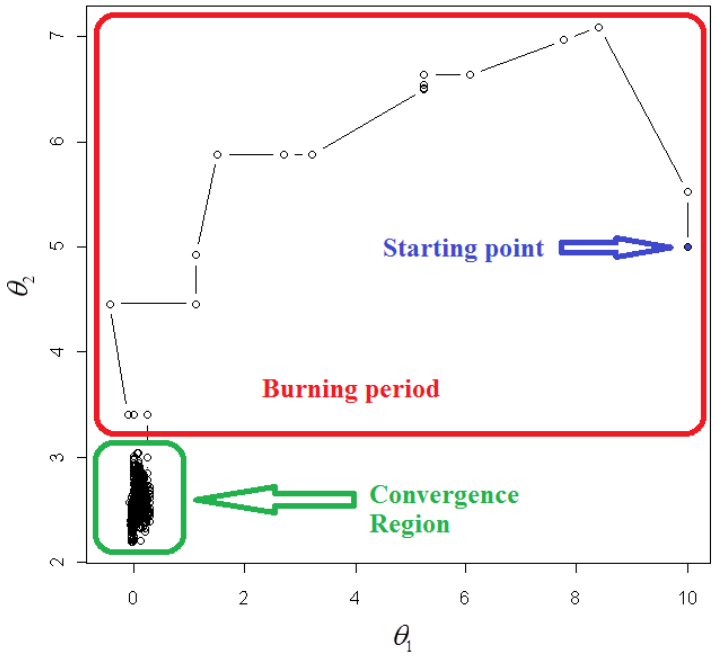


Figure 2.3-Illustration of parameter estimation with the MCMC method in a two-dimensional parameter space.

Classical MCMC method: Metropolis-Hastings

Many algorithms are available to implement the procedure described previously. The Metropolis-Hastings method is one of the most classical MCMC algorithms.

Let $f(\boldsymbol{x})$ denote the target distribution. For a multi-dimensional distribution f, $\boldsymbol{x}$ is hence a vector. In general, it is difficult to obtain the samples generated from f directly. The idea of MCMC is to use another distribution to generate samples (one by one), which is called the jump distribution. According to the property of Markov chains, the following sample does only depend on the previous one. Based on well-selected acceptance rules, the generated samples provide an asymptotic approximation of the target distribution.

More precisely, assumed that $\boldsymbol{x}$ is the current sample, the next sample $\boldsymbol{x}^{(*)}$ is generated according to a jump distribution $J(\boldsymbol{z}|\boldsymbol{x})$, which is conditional on $\boldsymbol{x}$. A common choice for the jump distribution is a multi-dimensional Normal distribution with its center on $\boldsymbol{x}$ and a covariance matrix $\boldsymbol{\Sigma}$. The selection of the covariance matrix will be discussed in each specific method. Thus, $J(\boldsymbol{z} \mid \boldsymbol{x}) = f_{Normal}(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\Sigma})$.

A classical method known as the Metropolis-Hastings method [*Metropolis and Ulam*, 1949] is summarized as follows.

Algorithm 1. Metropolis-Hastings

- Choose a starting point $\boldsymbol{x}^{(0)}$
- For $i=1,N_{sim}$ (N_{sim} is the number of MCMC simulation)
  - Generate a candidate sample $\boldsymbol{x}^{(*)}$ according to the jump distribution $J(\boldsymbol{z}/\boldsymbol{x}^{(i-1)})$
  - Calculate the acceptance ratio $\tau = \dfrac{f(\boldsymbol{x}^{(*)})}{f(\boldsymbol{x}^{(i-1)})} \dfrac{J(\boldsymbol{x}^{(i-1)} \mid \boldsymbol{x}^{(*)})}{J(\boldsymbol{x}^{(*)} \mid \boldsymbol{x}^{(i-1)})}$
  - If $\tau>1$ then accept the candidate sample ($\boldsymbol{x}^{(i)}=\boldsymbol{x}^{(*)}$); otherwise, accept it with probability τ (If accepted, $\boldsymbol{x}^{(i)}=\boldsymbol{x}^{(*)}$; if not, $\boldsymbol{x}^{(i)}=\boldsymbol{x}^{(i-1)}$).
- End for i

N_{sim} needs to be determined according to the convergence speed, which depends on the shape of the target distribution and its dimension. Questions about how to monitor the convergence will be discussed later. For a small dimensional distribution, MCMC chains converge quickly, but for large dimensional distributions, the MCMC chains often take a long time to reach convergence.

Block Metropolis-Hastings sampler

An alternative method described in the following is called the Block Metropolis method (see *Marshall et al.* [2004] for a thorough description and evaluation). The general idea is to update only a part of the parameter vector at each iteration, instead of the whole parameter vector as described in *Algorithm 1*. A particular Block Metropolis sampler is the "one-at-a-time sampler", which corresponds to using blocks of size one: a single dimension of the parameter is updated at each iteration, the rest remaining at the current value. This strategy is helpful to derive 'good' jump distributions, because by updating a single component at a time,

only one-dimensional jump distributions are needed. However, this algorithm also has a larger computational complexity. For the same simulation numbers, the "one-at-a-time" sampler will have N_{dim} times more computation than the classical Metropolis-Hastings methods, where N_{dim} is the dimension of the target distribution.

Algorithm 2. Block Metropolis

- Choose a starting point $\boldsymbol{x}^{(0)} = \left(x_j^{(0)}\right)_{j=1,N_{\dim}}$
- For $i=1,N_{sim}$ (N_{sim} is the number of MCMC simulation)
  - Set $\boldsymbol{x}^{(i*)} = \boldsymbol{x}^{(i-1)}$
  - For $j=1,N_{dim}$
    - Generate a candidate sample $x_j^{(*)}$ according to the jump distribution $J(z \mid x_j^{(i-1)})$.
    - Calculate the acceptance ratio τ (see *Algorithm 1*), where $\boldsymbol{x}^{(*)}$ is equal to $\boldsymbol{x}^{(i*)}$ except the j^{th} component is equal to $x_j^{(*)}$.
    - If τ>1 then accept $x_j^{(*)}$ as the j^{th} component of $\boldsymbol{x}^{(i*)}$ ($x_j^{(i*)} = x_j^{(*)}$); otherwise, accept it with probability τ (If accepted, $x_j^{(i*)} = x_j^{(*)}$; if not, $x_j^{(i*)} = x_j^{(i-1)}$).
  - End for j
  - Update $\boldsymbol{x}^{(i)} = \boldsymbol{x}^{(i*)}$
- End for i

A well-selected starting point (e.g. located in a high-density area of the target distribution) can help to limit the length of the burn-in period. However, it is difficult to foresee the high-density area prior to sample the posterior distribution. One possibility to obtain a starting point close to the high-density area is to use the prior mean. Another possibility is to use optimization algorithms.

The jump distribution is obviously the most important part of MCMC methods, since it is directly linked to the algorithm efficiency. In general, a 'good' jump distribution should be 'similar' to the target distribution in terms of its size and shape. For a Gaussian jump distribution, such character is controlled by the covariance matrix. Non-Gaussian jump distributions are also interesting for some specific target distributions (e.g. Student distribution). However, as our framework is designed for all kinds of distribution, we use Gaussian jump distributions. In the following, we are going to describe adaptive methods for controlling the covariance matrix.

Adaptive MCMC algorithms

It is difficult to find an adequate covariance matrix (or variance for uni-variate Normal distribution) for the jump distribution prior to sample the posterior distribution. More precisely, if the variance for each dimension of the multivariate Normal jump distribution is large, a candidate point will jump 'far away' from the original point. This point is consequently very likely to be rejected. In contrast, if the variance is small, the candidate

point will be 'close to' the original point. This point will hence be very likely to be accepted. A possible way to overcome this drawback is to use an adaptive method, which will learn from the existent samples to 'automatically' adjust the covariance values. Such automatic adaptation are particularly easy to implement in the "one-at-a-time" sampler, because the jump distribution is a univariate Normal distribution, hence only requiring a variance specification (as opposed to a full covariance matrix). More precisely, if the acceptance rate is high, then one can increase the variance to avoid remaining stuck into the same region for many iterations. Conversely, if the acceptance rate is low, one can decrease the variance to avoid too large jumps. More details can be found in *Renard et al.* [2006b].

The adaptive technique used in this thesis is summarized as follows. A scale parameter δ is multiplied to the covariance matrix or the variance of the jump distribution. For a fixed iteration number N_{iter}, if the acceptance rate ρ is higher than a predefined rate ρ_H, multiply δ by an increasing rate ρ_{inc} (ρ_{inc} >1). Conversely, if the acceptance rate ρ is lower than a predefined rate ρ_L, multiply δ by a decreasing rate ρ_{dec} (0< ρ_{dec} <1). With this notation, the Adaptive Metropolis-Hastings methods and Adaptive Block Metropolis methods are summarized as follows.

Algorithm 3. Adaptive Metropolis-Hastings

- Choose a starting point $\boldsymbol{x}^{(0)}$ and starting scale parameter δ
- Subdivide the N_{sim} iterations into blocks, with each block containing N_{iter} iterations. The total number of iterations is still equal to N_{sim}
- In each block of N_{iter} iterations:
  - Generate a candidate sample $\boldsymbol{x}^{(*)}$ according to the jump distribution $J(\boldsymbol{z}/\boldsymbol{x}^{(i-1)})$, in which the covariance matrix of J is $\delta\boldsymbol{\Sigma}$.
  - Calculate the acceptance ratio τ (see *Algorithm 1*)
  - If τ>1 then accept the candidate sample ($\boldsymbol{x}^{(i)}=\boldsymbol{x}^{(*)}$); otherwise, accept it with probability τ (If accepted, $\boldsymbol{x}^{(i)}=\boldsymbol{x}^{(*)}$; if not, $\boldsymbol{x}^{(i)}=\boldsymbol{x}^{(i-1)}$).
- Compute the acceptance rate ρ (number of $\boldsymbol{x}^{(*)}$ accepted in the last N_{iter} iterations divided by N_{iter})
- Update the scale parameter δ for the next N_{iter} iterations: if ρ <ρ_L, then δ=δ*ρ_{dec}; if ρ >ρ_H, then δ= δ *ρ_{inc}.
- Move to next block.

Algorithm 4. Adaptive Block Metropolis

- Choose a starting point $\boldsymbol{x}^{(0)} = \left(x_j^{(0)}\right)_{j=1,N_{\dim}}$ and starting scale parameters $\boldsymbol{\delta} = \left(\delta_j\right)_{j=1,N_{\dim}}$, where N_{dim} is the dimension of the target distribution
- Subdivide the N_{sim} iterations into blocks, with each block containing N_{iter} iterations.
- In each block of N_{iter} iterations:
  - Set $\boldsymbol{x}^{(i*)} = \boldsymbol{x}^{(i-1)}$
  - For $j=1,N_{dim}$
    - Generate a candidate sample $x_j^{(*)}$ according to the jump distribution $J(z \mid x_j^{(i-1)})$, in which the variance for J is $\delta_j\sigma^2$.
    - Calculate the acceptance ratio τ (see *Algorithm 1*), where $\boldsymbol{x}^{(*)}$ equals to $\boldsymbol{x}^{(i*)}$ except the j^{th} component equals to $x_j^{(*)}$.
    - If $\tau>1$ then accept $x_j^{(*)}$ as the j^{th} component of $\boldsymbol{x}^{(i*)}$ ($x_j^{(i*)} = x_j^{(*)}$); otherwise, accept it with probability τ (If accepted, $x_j^{(i*)} = x_j^{(*)}$; if not, $x_j^{(i*)} = x_j^{(i-1)}$).
  - End for j
  - Update $\boldsymbol{x}^{(i)} = \boldsymbol{x}^{(i*)}$
- Compute the acceptance rate $\boldsymbol{\rho} = \left(\rho_{(j)}\right)_{j=1,N_{\dim}}$ for each dimension (number of $x_j^{(*)}$ accepted in the last N_{iter} iterations divided by N_{iter})
- Update the scale parameters $\boldsymbol{\delta} = \left(\delta_j\right)_{j=1,N_{\dim}}$: for all dimension $j = 1,N_{dim}$: if $\rho_{(j)} < \rho_L$, then $\delta_j = \delta_j * \rho_{dec}$; if $\rho_{(j)} > \rho_H$, *then* $\delta_j = \delta_j * \rho_{inc}$.
- Move to next block.

MCMC sampler used in the thesis

In this thesis, we use a combination of the adaptive Metropolis-Hastings and Block Metropolis methods described previously.

The general idea behind this combination is the following:

- In Stage 1, we use the "one-at-a-time" sampler for its ease of adaptation (since it only involves one-dimensional jump distributions), and its robustness to poorly-chosen starting values. This allows generating a first set of samples that can be used to estimate the jump covariance matrix.
- Stage 2 uses an adaptive Metropolis-Hastings sampler, which is much faster than the "one-at-a-time" sampler. The objective of this stage is to adapt the scale factor of the covariance matrix.
- Stage 3 is the "production stage", where the samples that will be used to approximate the target distribution are generated using a standard, non-adaptive Metropolis-Hastings sampler.

Algorithm 5. Combined MCMC sampler

1. Choose a starting point $\boldsymbol{x}^{(0)} = \left(x_j^{(0)}\right)_{j=1,N_{\dim}}$ and starting scale parameters $\boldsymbol{\delta} = \left(\delta_j\right)_{j=1,N_{\dim}}$, where N_{dim} is the dimension of target distribution
2. **Stage 1**: Run N_{sim_1} iterations with the Adaptive "one-at-a-time" Metropolis method (*Algorithm 4*). The last sample is denoted by $\boldsymbol{x}^{(N_{sim_1})}$.
3. Burn the first $\rho_{burn} N_{sim_1}$ samples, where ρ_{burn} is the burn-in rate. Compute the sample covariance matrix $\boldsymbol{\Sigma_1}$ with the remaining samples from step 2.
4. **Stage 2**: Run N_{sim_2} iterations with the Adaptive Metropolis-Hastings method (*Algorithm 3*), with $\boldsymbol{x}^{(N_{sim_1})}$ as starting point and $\delta_0\boldsymbol{\Sigma_1}$ as covariance matrix for the jump distribution. At the end of this stage, the scale factor δ_0 has been updated to a final value δ.
5. **Stage 3**: Randomly select N_{chain} samples from the N_{sim_2} samples generated in step 4 as starting points. Run N_{chain} parallel chains for N_{sim_3} iterations with the classical Metropolis-Hastings method (*Algorithm 1*). The covariance matrix for the jump distribution is $\delta\boldsymbol{\Sigma_1}$.
6. Verify the convergence of parallel chains.

If the parallel chains converged, each chain obtained from step 5 provides an estimation for the target distribution. The number of iterations $N_{sim_1}, N_{sim_2}, N_{sim_3}$ in steps 2, 4, 6 is case-specific. Since *Algorithm 4* requires N_{dim} more computation than *Algorithm 3*, N_{sim_1} is generally smaller than N_{sim_2} and N_{sim_3}. Through the first two stages (steps 1-4), we solely intend to obtain 'good' starting points for each chain and a 'good' covariance matrix for the jump distribution. Thus N_{sim_1} and N_{sim_2} don't need to be very large.

Monitoring convergence

As illustrated in Figure 2.3, the MCMC samples provide an estimate for the target distribution only after convergence. This raises the question of detecting that the MCMC chain has converged. In this thesis, we run several parallel chains with different starting points, and based on these chains, compute the Gelman-Rubin (GR) index [*Gelman and Rubin*, 1992] to monitor convergence. As described by *Gelman and Rubin* [1992], we consider the parallel chains converged when the GR index is smaller than 1.2.

For m parallel MCMC sequences $\left(x_{ij}\right)_{i=1,n;\, j=1,m}$, the GR index for m chains with size n (already burned-in) is calculated as follows:

$$GRindex = \sqrt{\frac{n-1}{n} + \frac{m+1}{nm}\frac{B}{W}} \tag{2.5}$$

where B and W are the between- and within-chain variances and are calculated as follows:

$$B = n\sum_{j=1}^{m}\frac{(\bar{x}_{.j}-\bar{x}_{..})^2}{m-1}$$
$$W = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{(x_{ij}-\bar{x}_{.j})^2}{n-1} \quad (2.6)$$

where $\bar{x}_{.j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ and $\bar{x}_{..} = \frac{1}{m}\sum_{j=1}^{m}\bar{x}_{.j}$

2.2 Posterior distribution

Given the pre-specified distribution *D*, the regression functions $\boldsymbol{h}$ and the link functions $\boldsymbol{l}$, regression parameters $\boldsymbol{\theta}$ can be estimated in a Bayesian framework with MCMC methods. The posterior pdf of the regression parameters is computed as follows:

$$f(\boldsymbol{\theta}\,|\,\boldsymbol{Y}) \propto f(\boldsymbol{Y}\,|\,\boldsymbol{\theta}) f(\boldsymbol{\theta}) \quad (2.7)$$

where $f(\boldsymbol{\theta})$ is the prior pdf of regression parameters and $f(\boldsymbol{Y}|\boldsymbol{\theta})$ is the likelihood function:

$$\begin{aligned} f(\boldsymbol{Y}\,|\,\boldsymbol{\theta}) &= \prod_{t=t_1}^{t_n} f\left(Y(t)\,|\,\beta_1(t),\beta_2(t),...,\beta_m(t)\right) \\ &= \prod_{t=t_1}^{t_n} f\left(Y(t)\,|\,l_1^{-1}\left(h_1(\boldsymbol{x}_t,\boldsymbol{\theta}_1)\right), l_2^{-1}\left(h_2(\boldsymbol{x}_t,\boldsymbol{\theta}_2)\right),...,l_m^{-1}\left(h_m(\boldsymbol{x}_t,\boldsymbol{\theta}_m)\right)\right) \end{aligned} \quad (2.8)$$

In equation (2.8), a temporal independence assumption is applied: $\forall t_1 \neq t_2$, $Y(t_1)$ is independent of $Y(t_2)$.

It is difficult to give general guidelines for prior specification in the context of this framework, because the meaning of the inferred R-parameters depends on the chosen distribution and the regression model. Consequently, the use of standard priors (e.g. default Jeffreys' priors which have the property of being invariant under re-parameterization) is not straightforward. Thus, in the absence of any strong prior knowledge on R-parameters, we use flat improper priors in most of our studies. Particular prior distributions and distribution/regression models will be described in each case study independently.

2.3 Quantile computation based on the posterior distribution

For a fixed probability *p* and time t_0, the *p*-quantile for a parent distribution *D* (with its cdf *F*) is calculated by its inverse cdf with *D*-parameter $\boldsymbol{\beta}(t_0)$. As $\boldsymbol{\beta}(t_0)$ depends on R-parameters $\boldsymbol{\theta}$, we denote the *p*-quantile $F_{t_0}^{-1}(p\,|\,\boldsymbol{\theta})$.

In the Bayesian framework, R-parameters $\boldsymbol{\theta}$ are estimated with MCMC methods. We denote $\left(\boldsymbol{\theta}^{(k)}\right)_{k=1,N_{sim}}$ the N_{sim} MCMC generated R-parameters. Thus a posterior sample for the p-quantile at time t_0 is given by $\left(F_{t_0}^{-1}(p\,|\,\boldsymbol{\theta}^{(k)})\right)_{k=1,N_{sim}}$.

3 Model diagnosis and selection

In this section, we present graphical tools for model diagnosis. Then model comparison tools are discussed for selecting and comparing specific competing models.

3.1 Diagnostic tools

Graphical evaluation provides an easy way to check the goodness-of-fit. Compared with statistical tests, the graphical evaluation methods provide a simpler qualitative judgment, although they are not able to provide a quantitative judgment for fit. In this framework, we intend to make a quick check of model fit prior to starting further model comparison analyses. Thus graphical evaluation methods are integrated in the framework for their convenient usage.

One of the most commonly used graphical methods is the Quantile-Quantile plot [*Wilk and Gnanadesikan*, 1968]. For a parent distribution D with constant parameters $\boldsymbol{\beta}$ (not varying with time) and cdf F, the empirical quantiles (sorted observations $(y_i)_{i=1,m}$) are plotted against the theoretical quantiles $\left(F^{-1}(\frac{i}{m+1}\,|\,\boldsymbol{\beta})\right)_{i=1,m}$. If the model has a good fit, the plot of empirical quantiles versus model quantiles should be close to the diagonal. However, this plot cannot be used in a climate-informed or non-stationary context. Indeed, the derivation of this QQ plot supposes that data are identically distributed, which is not the case here, since the parameters vary according to the temporal covariates.

An alternative graphical method called the Probability-Probability plot (PP plot) is therefore used in a climate-formed or non-stationary context. The idea of the PP plot is that if $(Y_i)_{i=1,m}$ is a random variable with distribution $(D_i)_{i=1,m}$, whose cdf are $(F_i)_{i=1,m}$, then $(F_i(Y_i))_{i=1,m}$ are identically distributed according to a uniform [0,1] distribution. Therefore sorted values of $F_1(y_1), F_2(y_2), ..., F_m(y_m)$ are plotted against $\left(\frac{i}{m+1}\right)_{i=1,m}$. If the fit is good, this plot should be close to the diagonal. More explanations and usage of PP plot in a non-identically-distributed context can be found in *Coles* [2001, page 110-114].

3.2 Model comparison tools

The general framework allows analyzing the impact of different covariates on hydrologic data by using distinct regression models. Moreover, distinct parent distributions may also be tested and compared. This gives rise to a potentially large number of competing model formulations. Thus, a comparison tool is introduced to judge the performance of these

competing models. Questions like whether the impact of ENSO on the precipitation is linear or non-linear could be answered with these tools. In the literature, several criteria are proposed. The principle of these criteria is based on the tradeoff between goodness-of-fit and model complexity. This section proposes a short review of these criteria.

3.2.1 Criteria based on point-estimates of the parameters

The Akaike Information Criterion (*AIC*) [*Akaike*, 1974], its modified version *AICc* [*Hurvich and Tsai*, 1989] and the Bayesian information criterion (*BIC*) [*Schwarz*, 1978] are three commonly used criteria based on point-estimates of the parameters (maximum likelihood estimation). The performance of these criteria in different conditions (e.g. sample size, parent distribution, number of inferred parameters) was discussed by *Burnham and Anderson* [2002]. For instance, these authors suggested using *AICc* rather than *AIC* for a larger number of parameters *k* or a small number of observations *n*.

These three criteria are computed as follows:

$$AIC = 2k - 2\ln(L) \tag{2.9}$$

$$AICc = AIC + \frac{2k(k+1)}{n-k-1} \tag{2.10}$$

$$BIC = -2\ln(L) + k\ln(n) \tag{2.11}$$

where *k* is the number of inferred parameters in the model, *L* is the maximized value of the likelihood function and *n* is the sample size. In practice, models with small value of *AIC*, *AICc* and *BIC* are preferred. In the formula of these three criteria, a common term is *-2ln(L)*, which is the term for the goodness-of-fit. Thus the difference among these three criteria is the penalization for the number of inferred parameters.

3.2.2 Criteria based on the posterior distribution of parameters

The use of the criteria *AIC*, *AICc* or *BIC* may seem at odds with the Bayesian context we adopted in our modeling framework. Indeed, all three criteria are based on the maximum-likelihood estimates, while Bayesian inference is rather based on the full posterior distribution (which may markedly differ from the likelihood with precise priors and/or short samples). Moreover, all three criteria are based on point-estimates of the parameters, which somehow discards the benefit of using a full posterior distribution for inference. Two additional criteria overcome these limitations and are discussed here: Bayes Factors and the Deviance Information Criterion (DIC).

Bayes Factors

Bayesian model selection (BMS) technique is a method for model selection in the Bayesian framework. *Kass and Raftery* [1995] provide a guideline for the development, usage

and interpretation of such technique, and *Frost* [2004] provides further discussion and interpretation of the BMS tools.

Assume that $(M_i)_{i=1,q}$ is the collection of competing models and $(y_i)_{i=1,m}$ are the observations. $f(M_1),...,f(M_q)$ denote the prior probabilities that data are generated from model $M_1,...,M_q$. Consequently, $\sum_{i=1}^{q} f(M_i) = 1$.

The choice between model M_i and M_j can be made by computing the Bayes Factor, which is defined as follows:

$$BF(M_i, M_j) = \frac{f(\boldsymbol{y} \mid M_i)}{f(\boldsymbol{y} \mid M_j)} \tag{2.12}$$

In this term, the marginal likelihood of observations $f(\boldsymbol{y} \mid M_i)$ is defined as follows:

$$f(\boldsymbol{y} \mid M_i) = \int_{\boldsymbol{\theta} \in \Theta} f(\boldsymbol{Y} \mid \boldsymbol{\theta}^{(M_i)}, M_i) f(\boldsymbol{\theta}^{(M_i)} \mid M_i) d\boldsymbol{\theta}^{(M_i)} \tag{2.13}$$

where $f(\boldsymbol{\theta}^{(M_i)} \mid M_i)$ is the prior distribution of $\boldsymbol{\theta}^{(M_i)}$ conditional on model M_i. One can recognize in this equation the denominator of the Bayes theorem.

The direct calculation of the marginal likelihood of observations $f(\boldsymbol{y} \mid M_i)$ is difficult in most cases. Several asymptotic approximation methods are proposed by *Kass and Raftery* [1995]. However, a limitation of Bayes factors is that they require using proper prior distributions, which is often not the case in our studies. Therefore, the BMS approach is not used in our framework.

DIC

The Deviance information criterion (*DIC*) introduced by *Spiegelhalter et al.* [2002] is an alternative method for model selection in the Bayesian framework. We use *DIC* in our framework for the following two reasons:

1. It is based on the full posterior distribution and hence takes into account parameter uncertainty for judging different models
2. It accounts for the effect of prior information when available but remains usable with non-informative or improper priors (provided that the posterior distribution is proper).

The deviance for one given point $\boldsymbol{\theta}$ in parameter space is defined as follow:

$$Dev(\boldsymbol{\theta}) = -2\ln(f(y \mid \boldsymbol{\theta})) \tag{2.14}$$

The DIC criterion is then computed by:

$$DIC = \overline{Dev} + p_{Dev} \tag{2.15}$$

where $\overline{Dev} = E^{\boldsymbol{\theta}}\left[Dev(\boldsymbol{\theta})\right]$ is the expectation of the deviance (with respect to the posterior distribution) and $p_{Dev} = \overline{Dev} - Dev(\bar{\boldsymbol{\theta}})$ is the model complexity penalty. Models with small DIC values are preferred. Similar to *AIC*, *AICc* and *BIC*, the *DIC* also uses the term *-2ln(L)* for the goodness-of-fit. The difference is that it is now averaged with respect to the posterior distribution, rather than computed at the maximum likelihood value.

The *DIC* can be easily computed with the MCMC samples. Let $\left(\boldsymbol{\theta}^{(k)}\right)_{k=1,N_{sim}}$ be the N_{sim} MCMC generated points in parameter space. The calculation of *DIC* uses the following algorithm:

- For each $\boldsymbol{\theta}^{(k)}$, calculate the deviance $Dev(\boldsymbol{\theta}^{(k)})$ using equation (2.14), then get its average $\overline{Dev}$.
- Get the average point $\bar{\boldsymbol{\theta}} = \dfrac{1}{N_{sim}}\sum_{k=1}^{N_{sim}}\boldsymbol{\theta}^{(k)}$, then calculate $Dev(\bar{\boldsymbol{\theta}})$ using equation (2.14).
- Calculate *DIC* according to Eq (2.15).

4 Synthetic case studies

In this section, we intend to verify the numerical implementation of the modeling framework and to quantitatively assess the extent to which temporal variations can be detected with local models. Two synthetic datasets sampled from different time-varying models are used in the sequel.

4.1 Synthetic study 1

In the first test, a non-stationary *GEV* model is proposed with a seasonality component for the location parameter and a trend in time for the scale parameter. The test data $\boldsymbol{y} = (y_1, y_2, ..., y_{500})$ are sampled from the following distribution (Figure 2.4):

$$\text{For } t = 1, 2, ..., 500,\ Y_t \sim GEV(10\cos(0.05t), 10 + 0.01t, -0.1) \tag{2.16}$$

We assume the following *GEV(μ(t),σ(t),ξ(t))* model for the observations $\boldsymbol{y}$:

$$\begin{aligned} \mu(t) &= \theta_1 \cos(\theta_2 t) \\ \sigma(t) &= \theta_3 + \theta_4 t \\ \xi(t) &= \theta_5 \end{aligned} \tag{2.17}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$ are the regression parameters to be estimated. The starting point for the MCMC sampler is chosen as $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)}, \theta_4^{(0)}, \theta_5^{(0)}) = (10, 0.1, 8, 0.1, 0)$.

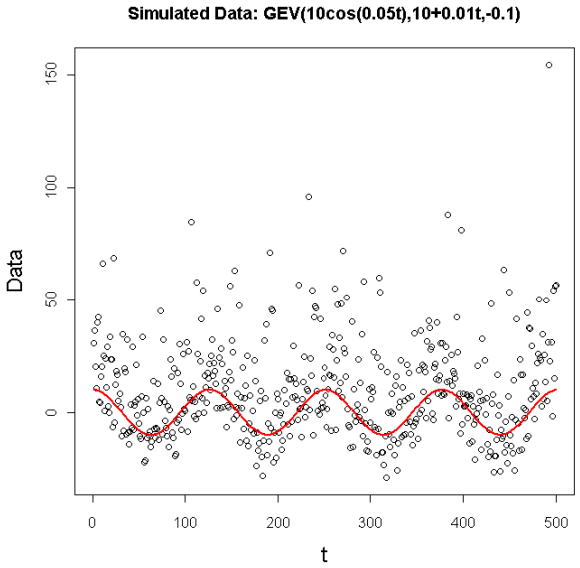


***Figure 2.4-Simulated data for synthetic study 4.1. The red curve represents the curve of function f(t)=10cos(0.05t) in the location parameter of Eq* (2.16).**

Figure 2.5 illustrates the MCMC sequences for the five R-parameters: (a),(b) and (c) are respectively for the regression parameters in location, scale and shape parameter of the parent *GEV* distribution. Results show that values of these regression parameters converge to the true parameters even if the starting values are poor. In fact, the farther starting points are from the true values, the more iteration is needed to reach convergence.

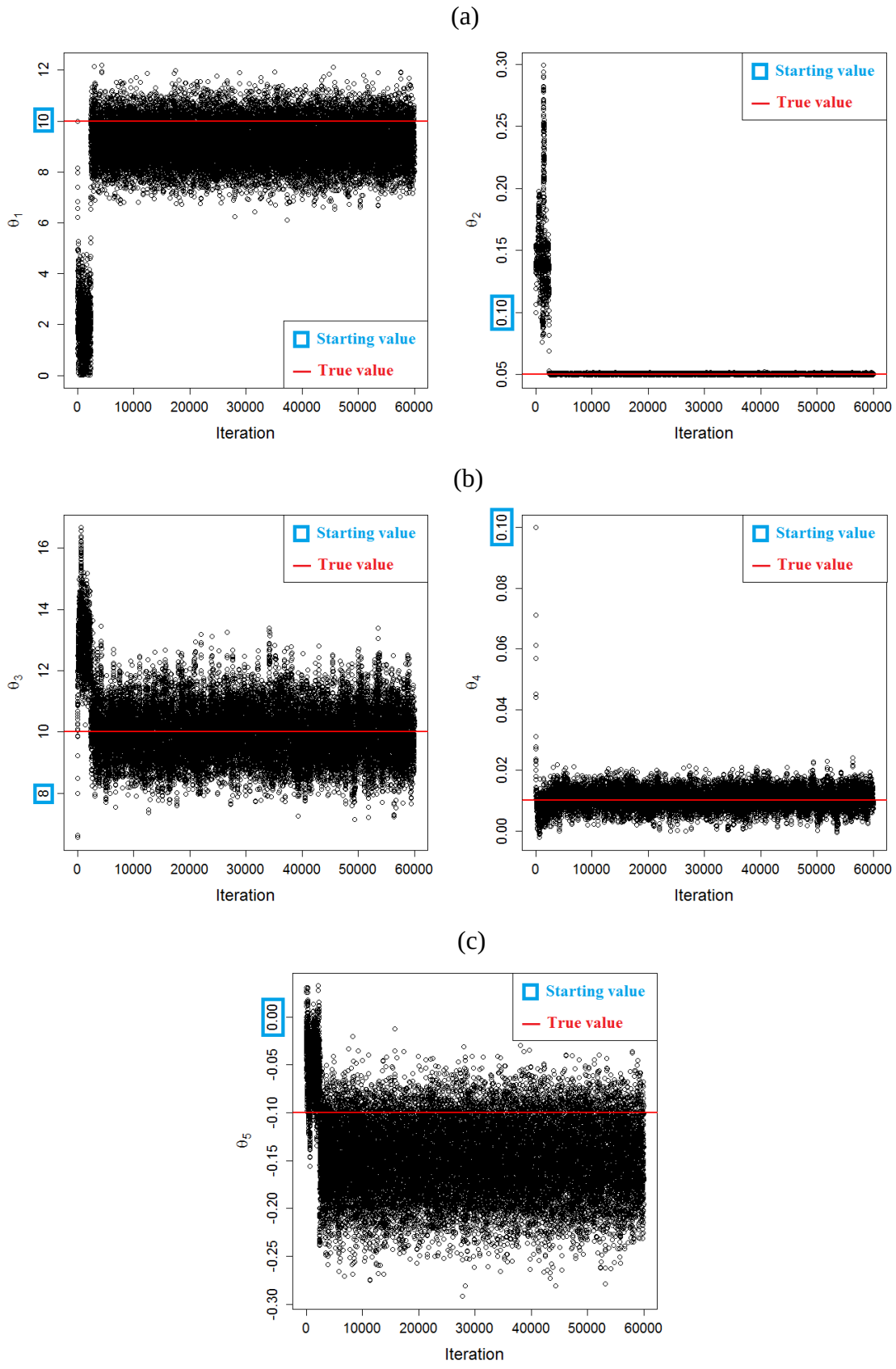


Figure 2.5- MCMC sequences for the five regression parameters: (a) parameters θ_1 and θ_2 for the location parameter; (b) parameters θ_3 and θ_4 for the scale parameter; (c) parameter θ_5 for the scale parameter. The red lines represent the true values.

4.2 Synthetic study 2

The second test is based on a *GEV* distribution with both location and scale parameters described as linear functions of time. The synthetic data $\boldsymbol{y} = (y_1, y_2, ..., y_{500})$ are sampled from the following distribution (Figure 2.6):

$$\text{For } t = 1, 2, ..., 500,\ Y_t \sim GEV(50 + 0.5t, 10 + 0.01t, -0.1) \tag{2.18}$$

Similarly to the first study, the following *GEV(μ(t),σ(t),ξ(t))* model is used to describe the observations ***y***:

$$\begin{aligned} \mu(t) &= \theta_1 + \theta_2 t + \theta_3 t^2 \\ \sigma(t) &= \theta_4 + \theta_5 t \\ \xi(t) &= \theta_6 \end{aligned} \tag{2.19}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6)$ are the regression parameters that need to be estimated. The starting point for the MCMC sampler is chosen as $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)}, \theta_4^{(0)}, \theta_5^{(0)}, \theta_6^{(0)}) = (10, 0.1, 0.1, 8, 0.1, 0)$. We add θ_3 in the regression function for location parameter to characterize the trend on t^2, which is used to test whether the redundant regression parameters will bias the estimation result.

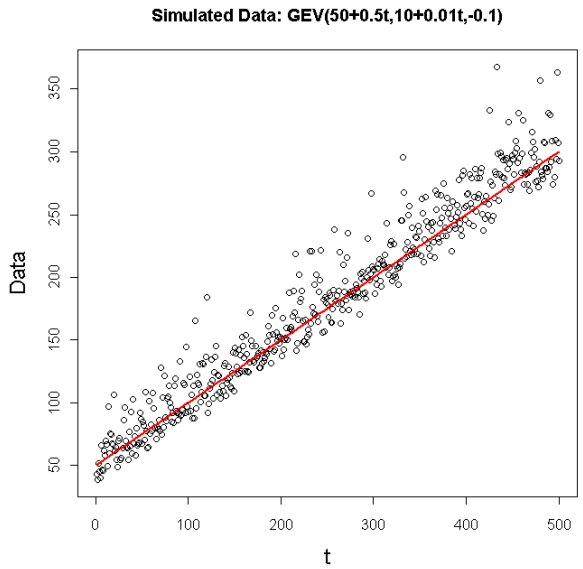


Figure 2.6- Simulated data for synthetic study 4.2.

Two estimations using respectively 50 and 500 observations are performed. The objective is to assess the difference between these two estimations. It is expected that both estimation results remain consistent with the true parameter values. However, our interest is also in assessing the difference in terms of estimation uncertainties.

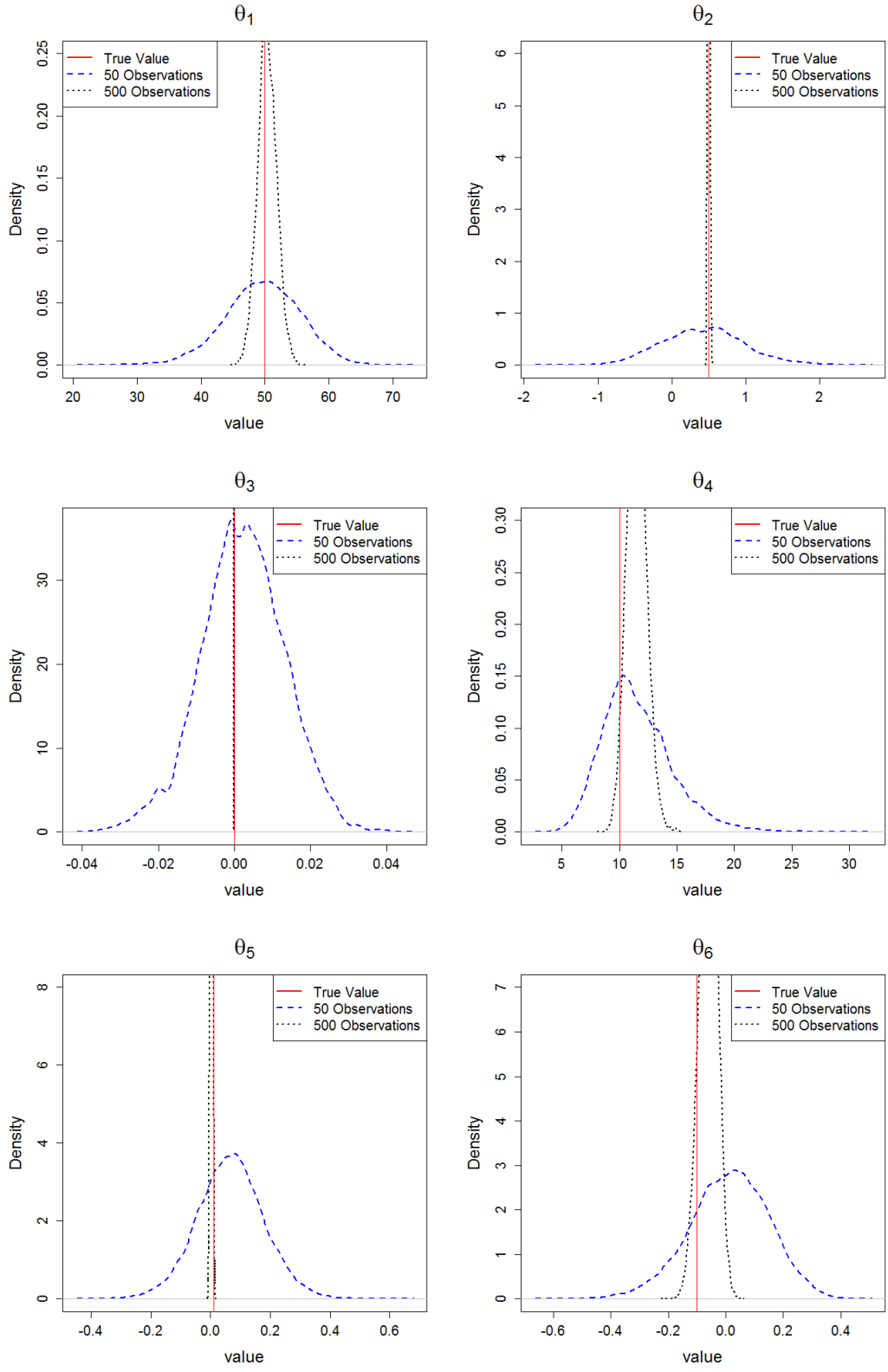


Figure 2.7-Posterior distributions of the six regression parameters using 50 and 500 observations.

Figure 2.7 illustrates the posterior distributions of the six regression parameters using respectively 50 and 500 observations. Not surprisingly, it is found that the posterior distributions based on 500 observations are much more concentrated around the true values than those based on 50 observations. Consequently, a relatively "small" number of observations (which actually corresponds to the typical length we should expect from hydrologic data) will lead to very large uncertainties. The burn-in period (Figure 2.8) is a little bit longer when 500 observations are used. This is because the uncertainty with 50 observations is larger, thus the starting values are almost already in the convergence region. By contrast, with 500 observations, the starting point is not in the convergence region, thus it still needs some time for the chain to converge.

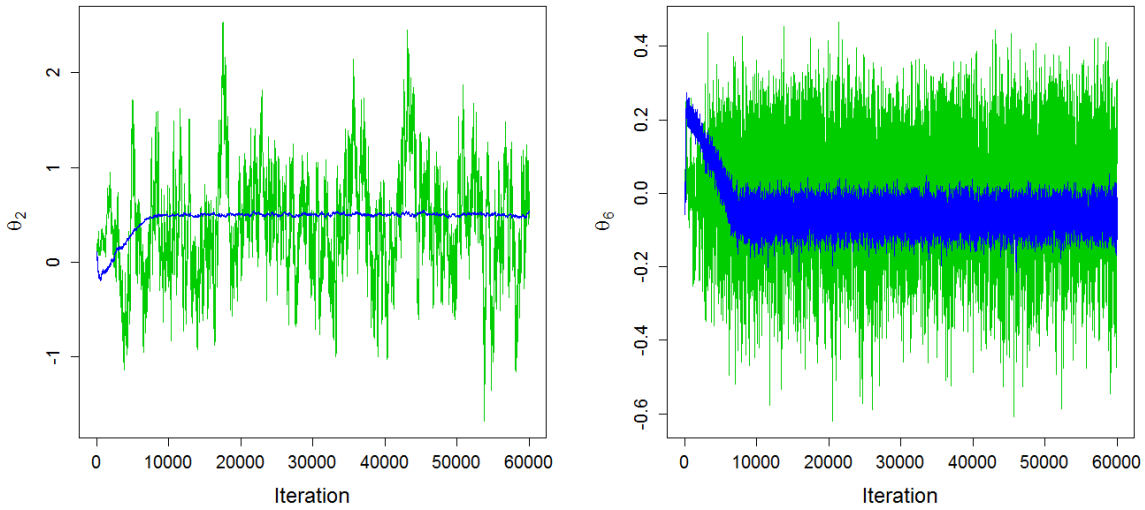


Figure 2.8- MCMC sequences for regression parameters θ_2 and θ_6 with 50 (green line) and 500 (blue line) observations.

5 Conclusion on the local climate-informed framework

In this chapter, we built a general climate-informed frequency analysis model, geared towards detecting and quantifying the effect of climate variability on hydrological variables. This is undertaken by using temporal regression models where the parameters of the parent distribution are a function of time-varying covariates (e.g. ENSO). A major objective was to keep this framework as general as possible. In particular, it enables free choices on the parent distribution and covariates. Moreover the selection of linear or non-linear regression functions is also convenient and flexible. We also implemented several tools for facilitating the use of this framework, in particular: (i) an inference framework based on Bayesian estimation and MCMC sampling; (ii) graphical diagnostics to evaluate goodness-of-fit; (iii) model comparison criteria, enabling the comparison of different competing hypotheses for the parent distribution and/or the regression models.

This framework was implemented as a flexible computer code (in FORTRAN), whose reliability was assessed through several synthetic case studies. We verified that this general

framework is able to detect and quantify temporal variations that may be due to climate variability or non-stationarity. These synthetic case studies also drew our attention to the very large uncertainties affecting parameter estimates with relatively short samples, which unfortunately correspond to typical sample sizes available in Hydrology. In order to improve the precision of the estimation, the most direct way would be to increase the sample sizes. This is however easier said than done, since the historical observation data are limited at each station. Seeking alternative solutions to overcome this shortage becomes especially important. It is also the motivation for moving from a local to a regional analysis, in which observations from different sites are taken into account together. Compared with the local analysis, this might help reducing estimation uncertainty, at the cost of additional assumptions on the regional variability of parameters. This will be discussed in more details in Part II.

CHAPTER 3 Case studies with local time-varying models

The objective of this chapter is to illustrate the use of the general FA framework at the local scale. Two cases will be studied in this chapter.

The first one describes temporal trends for various hydrological variables in the Durance catchment, as projected by a GCM coupled with a downscaling method under a given scenario of future greenhouse gases emission. The first specific objective of this case study is to highlight the generality and flexibility of the modeling framework in terms of parent distributions, since several distributions – both continuous and discrete – are considered. The second objective is to illustrate how the assumption made to describe the temporal evolution of hydrological variables impacts the estimation of failure probabilities.

The second case study aims to detect and quantify temporal trends and *NAO* effects on annual maximum daily precipitation in the Mediterranean area. The specific objective of this second case study is to consider several hypotheses regarding the temporal variability of rainfall extremes and to illustrate the use of model comparison tools to evaluate the relevance of those hypotheses.

1 Projected changes in the precipitation regime of the Durance catchment

In this section, we aim to verify that the general framework is applicable with a variety of parent distributions. More precisely, we are going to assess the existence of temporal trends on hydrological variables from the Durance catchment, for which the parent distribution is not evident to choose. Different parent distributions, both discrete and continuous, will be compared with various regression models.

The second objective is to evaluate the impact of the model used to describe non-stationarity on risk analysis. To this aim, we will compare failure probabilities based on various stationary and non-stationary hypotheses.

The data used in this section are 21st century downscaled climate projections. We stress that the aim of this section is just to probabilistically describe the outcome of this particular model, rather than predict the uncertainties in future changes, which would require using various climate models, emission scenarios and downscaling methods.

1.1 Data

1.1.1 The Durance catchment

The Durance River is located in southeast France with its source in *Montgenèvre*, *Hautes Alpes* at an altitude of 2300 meters. The river is 302 kilometers long and flows through the region of *Provence-Alpes-Côte d'Azur* to join the Rhône River at the border of *Bouches du Rhône* and *Vaucluse* departments.

The catchment area is 14,250 km² (Figure 3.1). Melting snow and rainfall during spring, autumn and winter contribute as the main resources of water for the catchment. Various climate forcings influence the runoff regimes of the catchment: mountainous condition for the upper reaches of the river and Mediterranean condition for the lower reaches. The maximum runoff usually occurs during the snowmelt period (May, June). However, historical records show that most violent floods were occurring in autumn.

1.1.2 Simulated data

Daily precipitation data come from the projection of the Sea Atmosphere Mediterranean Model (SAMM), a coupling of the ARPEGE atmospheric model [*Gibelin and Déqué*, 2003] and the regional ocean circulation model for the Mediterranean (OPAMED) [*Somot et al.*, 2008], under a given scenario of current and future (A2) greenhouse gases emission [*Nakicenovic et al.*, 2000]. This projection has been further statistically downscaled based on weather types and conditional resampling [*Boe et al.*, 2006; *Pagé et al.*, 2008] from SAFRAN near-surface atmospheric reanalysis [*Vidal et al.*, 2010]. The simulated daily data are available from 1960 to 2100, with a decomposition of total precipitation into rainfall and snowfall with a spatial resolution of 8km*8km. The daily catchment precipitation is estimated as the

weighted average of gridded precipitation according to the surface of the catchment contained in each cell.

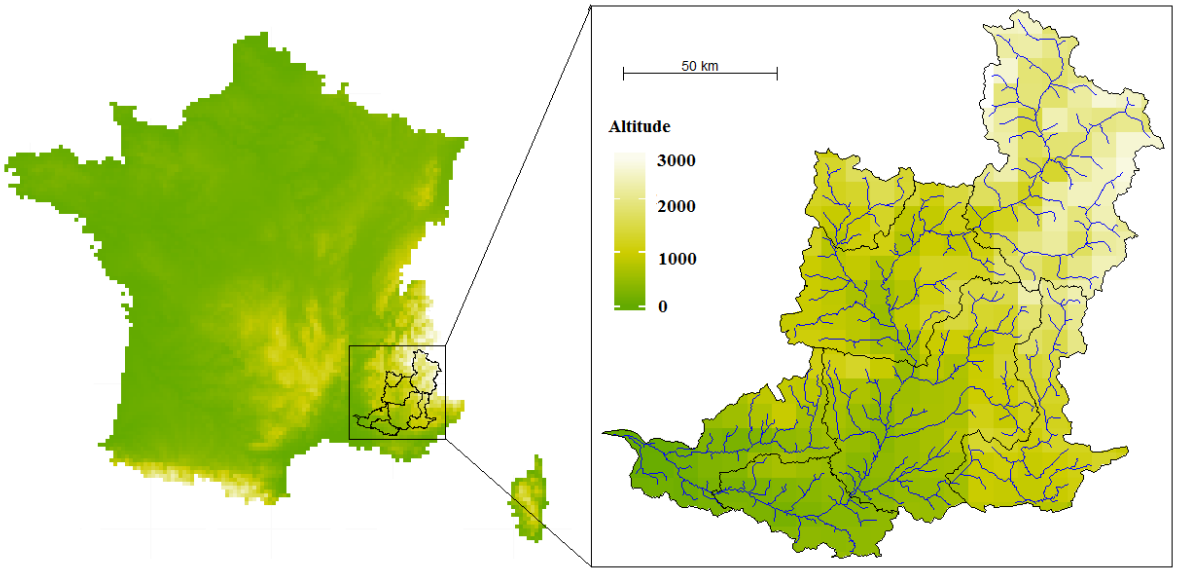


Figure 3.1-The Durance Catchment

1.2 Precipitation variables

With the projected daily precipitation dataset, we extracted several variables, such as seasonal/annual maximum and total precipitation, seasonal/annual non-precipitation days, annual maximum consecutive dry days, date of first snow, etc.

Figure 3.2 gives an example of the extracted variables. Most variables don't show any significant temporal trend, as exemplified in Figure 3.2 with the annual maximum daily precipitation and the annual non-precipitation days.

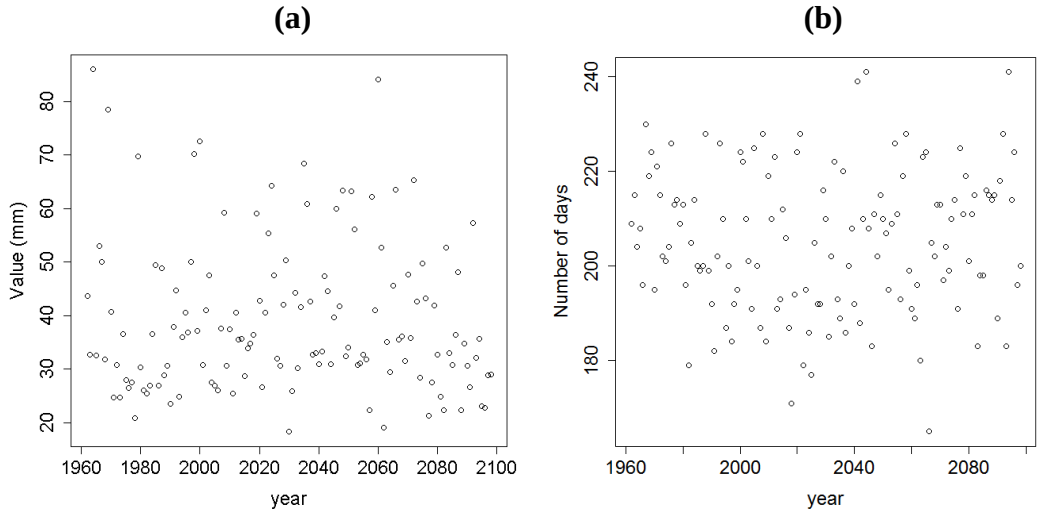


Figure 3.2-Annual maximum daily precipitation (a) and annual non-precipitation days (b) in the Durance catchment

However, the first snowy day shows a significant increasing trend (Figure 3.3). This variable counts the number of days from 1st September before a snowfall higher than 0.5 mm is observed. A simple linear regression applied to the whole series indicates that the first snowy day will be delayed by about 2.3 days every 10 years.

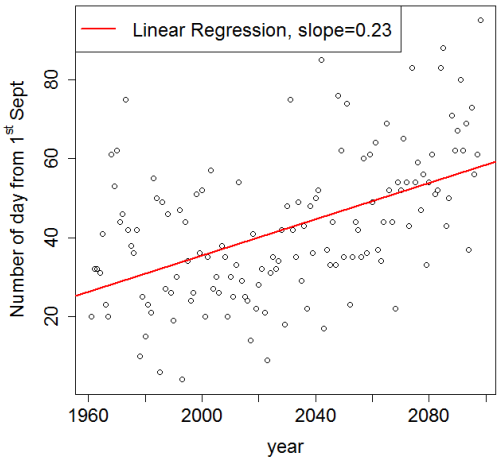


Figure 3.3-Date of first snow counting from 1st September every year

In order to provide a probabilistic analysis of this variable, we are going to apply the general FA framework built in the previous chapter to the variable 'first snowy day".

1.3 Parent distribution selection

The first step in applying the framework is to select the parent distribution. Both discrete and continuous distributions are considered here. Table 3.1 lists the tested distributions and a qualitative assessment of goodness-of-fit on the whole data under the stationary hypothesis.

Table 3.1-Goodness-of-fit of different distributions for the date of first snow

Distribution type	Distribution Name	Parameter number	Goodness-of-fit
Discrete	Geometric	1	Bad Fit
	Poisson	1	Bad Fit
	Binomial	2	Bad Fit
	Negative Binomial	2	Good Fit
Continuous	Normal	2	Good Fit
	GEV	3	Good Fit

The two-parameter negative binomial and normal distributions and the three-parameter GEV distribution provide a good fit for this variable. We will focus on the two-parameter distributions, i.e. negative binomial and Normal distributions, as the parent distributions.

Indeed, the GEV distribution is adapted for extreme data defined as block maxima, which is not the case of the variable "date of first snow" studied here.

1.4 Regression models

Two competing hypotheses are both applied to the selected parent distributions, either with stationary parameters or with a temporal trend on the mean parameter. Table 3.2 lists the four competing models, where $\mu(t)$ and $\sigma(t)$ are the mean and standard deviation of the Normal distribution, and $\mu(t)$ and $r(t)$ are the mean and the pre-fixed lost number of the negative binomial distribution. In the latter case, if p denotes the probability of success, the Negative Binomial distribution is re-parameterized so that $\mu = \frac{pr}{1-p}$. Flat priors are used for all regression parameters.

Table 3.2-Competing regression models for selected parent distributions

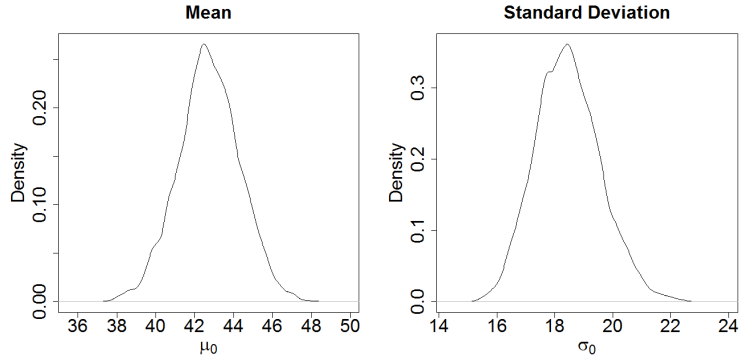
Model Name	Hypothesis	Model
		$N\left(\mu(t),\sigma(t)\right)$
Nor-L_0	Stationary	$\mu(t)=\mu_0,\ \sigma(t)=\sigma_0$
Nor-L_1	Non-stationary	$\mu(t)=\mu_0+\mu_1 t,\ \sigma(t)=\sigma_0$
		$Neg\left(\mu(t),r(t)\right)$
NB-L_0	Stationary	$\mu(t)=\mu_0,\ r(t)=r_0$
NB-L_1	Non-stationary	$\mu(t)=\mu_0+\mu_1 t,\ r(t)=r_0$

1.5 Posterior distribution of regression parameters

Figure 3.4 illustrates the posterior distribution of regression parameters in *Nor-L_0* and *Nor-L_1* models. The mean value μ_0 of *Nor-L_0* is significantly different from the intercept μ_0 of *Nor-L_1*. This is because a significant positive slope μ_1 is detected for *Nor-L_1* model. Moreover, the standard deviation is slightly smaller in *Nor-L_1*, suggesting that a part of the data variability is due to the existence of a trend. Similar results are found for the negative binomial distribution.

The slope parameter μ_1 in both *Nor-L_1* and *NB-L_1* models shows a significantly positive value (Figure 3.5). Moreover, the posterior distributions are similar in both models. This consistency shows that the temporal trend on the mean can be detected with two different parent distributions.

(a) Nor-L_0



(b) Nor-L_1

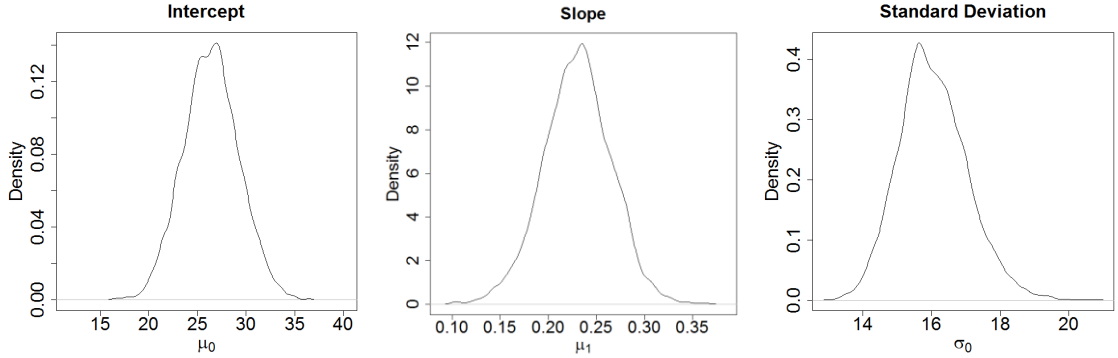


Figure 3.4- Posterior distribution of regression parameters in (a) Nor-L_0 and (b) Nor-L_1 models

(a) Nor-L_1 (b) NB-L_1

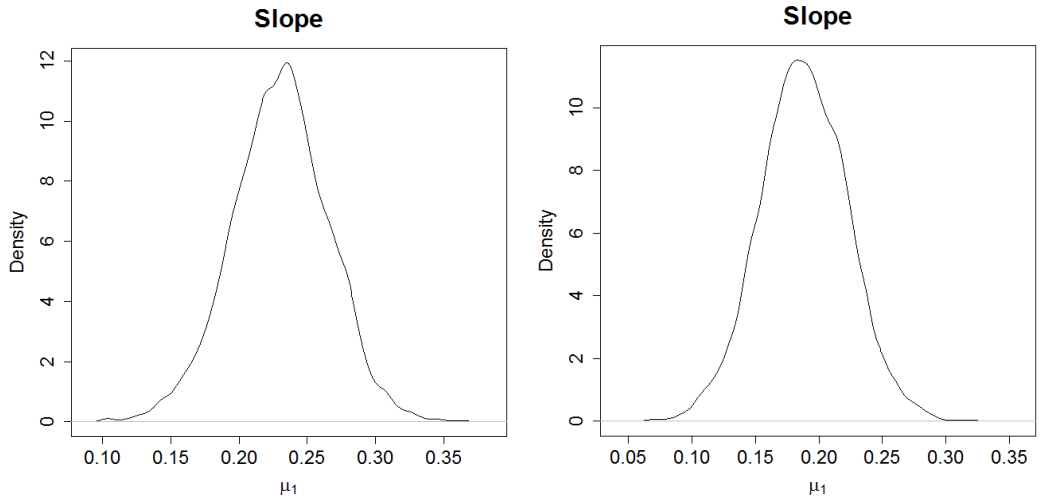
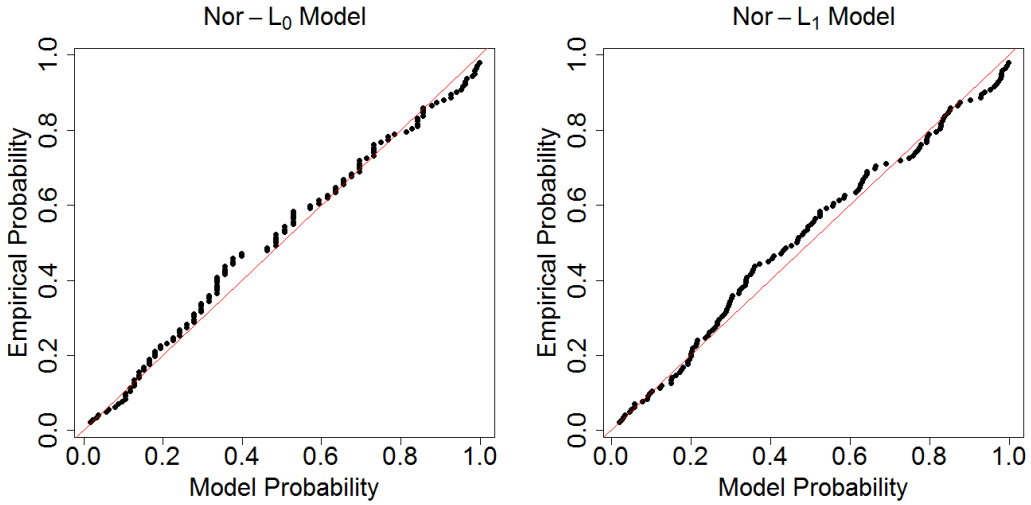


Figure 3.5-Posterior distribution of the slope parameter in (a) Nor-L_1 and (b) NB-L_1 models

1.6 Goodness-of-fit

Figure 3.6 shows the PP plot of the models in Table 3.2. For all four models, the PP plots are close to the diagonal. Although this suggests all four models have an acceptable fit, it is difficult to judge the model performance solely based on this PP plot.

(a) Normal distribution



(b) Negative binomial distribution

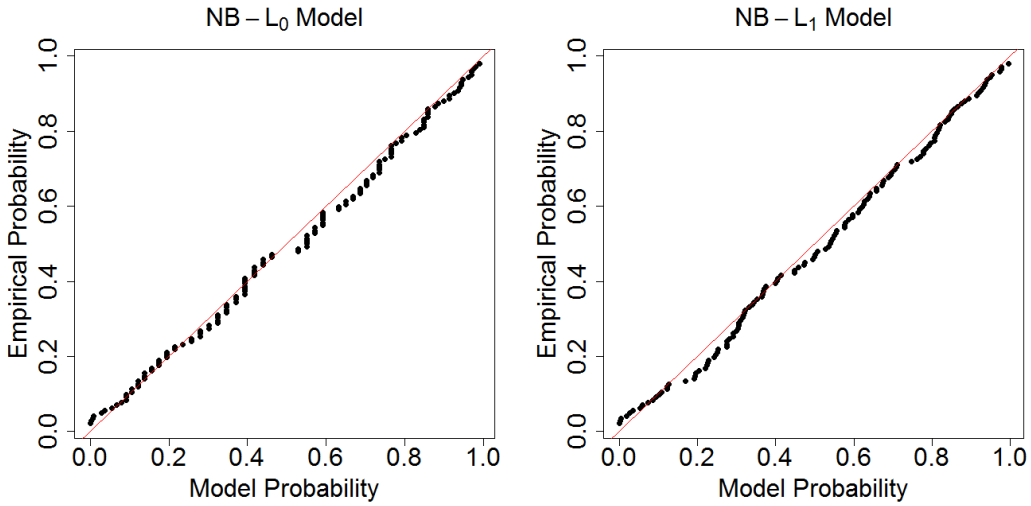


Figure 3.6-PP plot of the four competing models

1.7 Model comparison

Table 3.3 lists the *AICc*, *BIC* and *DIC* values for the four models. We can find that the *AICc*, *BIC* and *DIC* values for non-stationary models (*Nor-L_1* and *NB-L_1*) are much smaller than their stationary counterparts. It indicates that the temporal regression parameter is not a negligible component in the regression model. All these three criteria are also in agreement that *Nor-L_1* is preferred to *NB-L_1*.

Table 3.3-AICc, BIC and DIC values for the four models

Model	Nor-L_0	Nor-L_1	NB-L_0	NB-L_1
AICc	1197.01	1157.92	1193.87	1165.07
BIC	1202.86	1166.70	1199.73	1173.85
DIC	1194.26	1154.27	1176.67	1159.11

For all three indices, the negative binomial distribution seems better in a stationary model, but the Normal distribution becomes better in a non-stationary model, which suggests that the choice of the parent distribution interacts with the choice of the regression model. Thus the performance of parent distributions could not be discussed without regressions functions.

1.8 Accounting for non-stationarity in GCM projections: stationary sub-periods vs. continuous trend

As discussed in Chapter 1 (Section 4.2), many previous studies of GCM outputs rely on the study of two sub-periods (representing present and future climate conditions), which are assumed to be strictly stationary. The evolution is then simply described as the difference (or the ratio) between the estimates of both periods. This can be explained by the fact that earlier generations of GCM did not provide "transient" runs, but only runs on several sub-periods. However, current GCM generally provide transient runs, thus enabling a continuous description of changes.

In this section, we aim to verify whether the results from an analysis assuming strictly stationary sub-periods are consistent with the non-stationary trend model for a long period. Two "present" and "future" periods are selected for this comparison: end of 20st century (1970-1999) and middle of 21nd century (2035-2064).

As the previous section shows that *Nor-L_1* is the preferred model, we use a Normal distribution as the parent distribution for the stationary analysis. Results of stationary FA analyses on these two sub-periods are compared with the full non-stationary analysis.

1.8.1 Stationary sub-periods model

Two 30-years periods data (1970-1999 and 2035-2064) are fit with a normal distribution $N(\mu,\sigma)$. Figure 3.7 presents the posterior distribution of μ and σ. The standard deviation of these two periods is remarkably similar. However, there is a marked difference between the mean values of these two periods. This result confirms once again that there exists an evolution in the first snowy day.

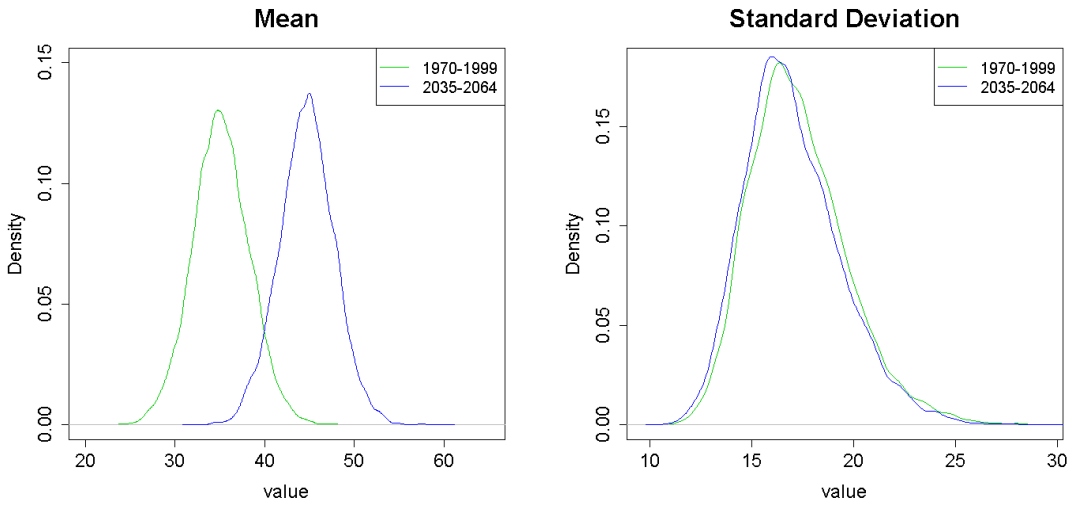


Figure 3.7-Posterior distribution of μ and σ with a Normal parent distribution

1.8.2 Stationary sub-periods model vs. non-stationary trend model

Figure 3.8 illustrates the results of the stationary sub-periods model and the non-stationary trend model. The 0.5-quantile and 0.99-quantile computed with the non-stationary trend model get across the credibility intervals of the stationary sub-periods model during the two sub-periods. It indicates that the prediction with the non-stationary model is consistent with stationary model results for each sub-period.

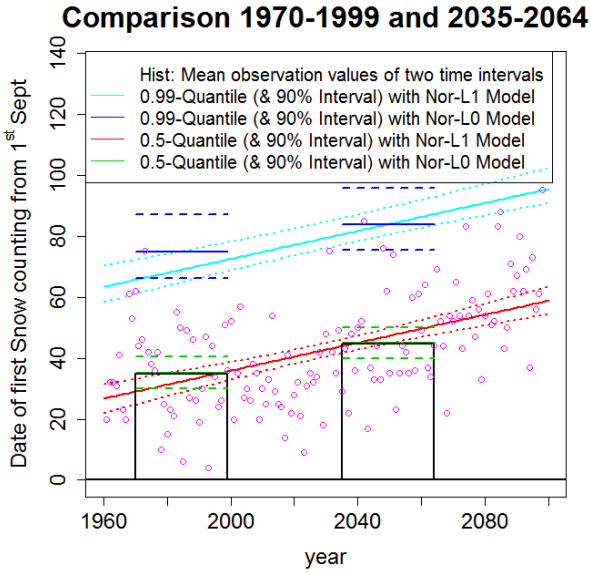


Figure 3.8-Results of the stationary sub-periods model and the non-stationary trend model. Pink circles are GCM-projected data. The histogram is the mean observation during 1970-1999 and 2035-2064. Green (blue) lines are 0.5-quantiles (0.99-quantiles) with 90% credibility interval (dashed lines) for the two sub-periods with the stationary model. Red (light blue) lines are 0.5-quantiles (0.99-quantiles) for the whole period with the non-stationary trend model.

For computational reasons, it is easier to use stationary models within two sub-periods than a time-varying model. However, if the sub-periods are too short, the estimation may be much uncertain. This is illustrated in Figure 3.8 by the credibility intervals for the 0.99-quantile being twice as large for the stationary sub-periods model as for the time-varying model. On the other hand, if the sub-periods are too long, the stationary hypothesis may become inadequate. There is no such restriction for a continuously time-varying model.

1.8.3 Failure probabilities

In this section, we are interested in the probability that the first snow will arrive later than a fixed threshold at least once during the next *n* years. This can be interpreted as a "failure probability". As an illustration, consider a ski station whose opening date is scheduled *D* days after the 1st of September. The first snow happening more than *D* days after the 1st of September will lead to a failure to open the station on due time. The concept of failure probability is more general and is central in risk assessment, for instance for designing civil structures such as dams. A quantity of interest is the probability that the volume of water will exceed the limit of a dam at least once during the next *n* years. This information can be used to decide the capacity of the dam prior to its construction. For a given threshold *D*, and making the assumption of temporal independence, the failure probability is calculated as follows:

$$\begin{aligned} p_D &= \mathrm{P}\left(\{Y_1 > D\} \cup \ldots \cup \{Y_n > D\}\right) \\ &= 1 - \mathrm{P}\left(\{Y_1 < D\} \cap \ldots \cap \{Y_n < D\}\right) \\ &= 1 - \prod_{i=1}^{n} \mathrm{P}\left(Y_i < D\right) \end{aligned} \tag{3.1}$$

In a stationary model, $(Y_i)_{i=1,n}$ have the same distribution. In a time-varying model, the distribution of Y_i depends on the covariates (time in this case study). Figure 3.9 presents the failure probability for the first snow happening later than 60, 70, 80 and 90 days after 1st Sept at least once during the *n* years following 2013. The failure probability is computed in three distinct ways: (i) using the stationary estimates from the period 1970-1999; (ii) using the stationary estimates from the period 2035-2064; (iii) using the non-stationary estimates from the trend model. The failure probabilities are calculated with modal regression parameters (that correspond to the maximum posterior values). For all four thresholds, the failure probability of the stationary estimates from the period 1970-1999 is always the smallest. This is because the observations in this period are markedly smaller than during the period 2035-2064. Compared with the stationary models, the time-varying model increases much faster with respect to the duration. For durations larger than about 45 years, the failure probability exceeds that derived from the stationary model estimated on the period 2035-2064.

Figure 3.10 illustrates the failure probability for a fixed duration *n*. If a ski station wants to decide of its opening date so that the failure probability is less than 0.2 in the following 60 years, the answer lies between $D \approx 80$ ($\approx 20^{th}$ November, stationary estimates from 1970-1999) and $D \approx 90$ ($\approx 30^{th}$ November, stationary estimates from 2035-2064 and trend model).

The length of 90% credibility intervals is about 20 days for the stationary models and 10 days for the non-stationary model. This provides preliminary information for analyzing this problem, based on which relevant adaptation procedures could be developed, for example designing reservoirs for artificial snow production, etc.

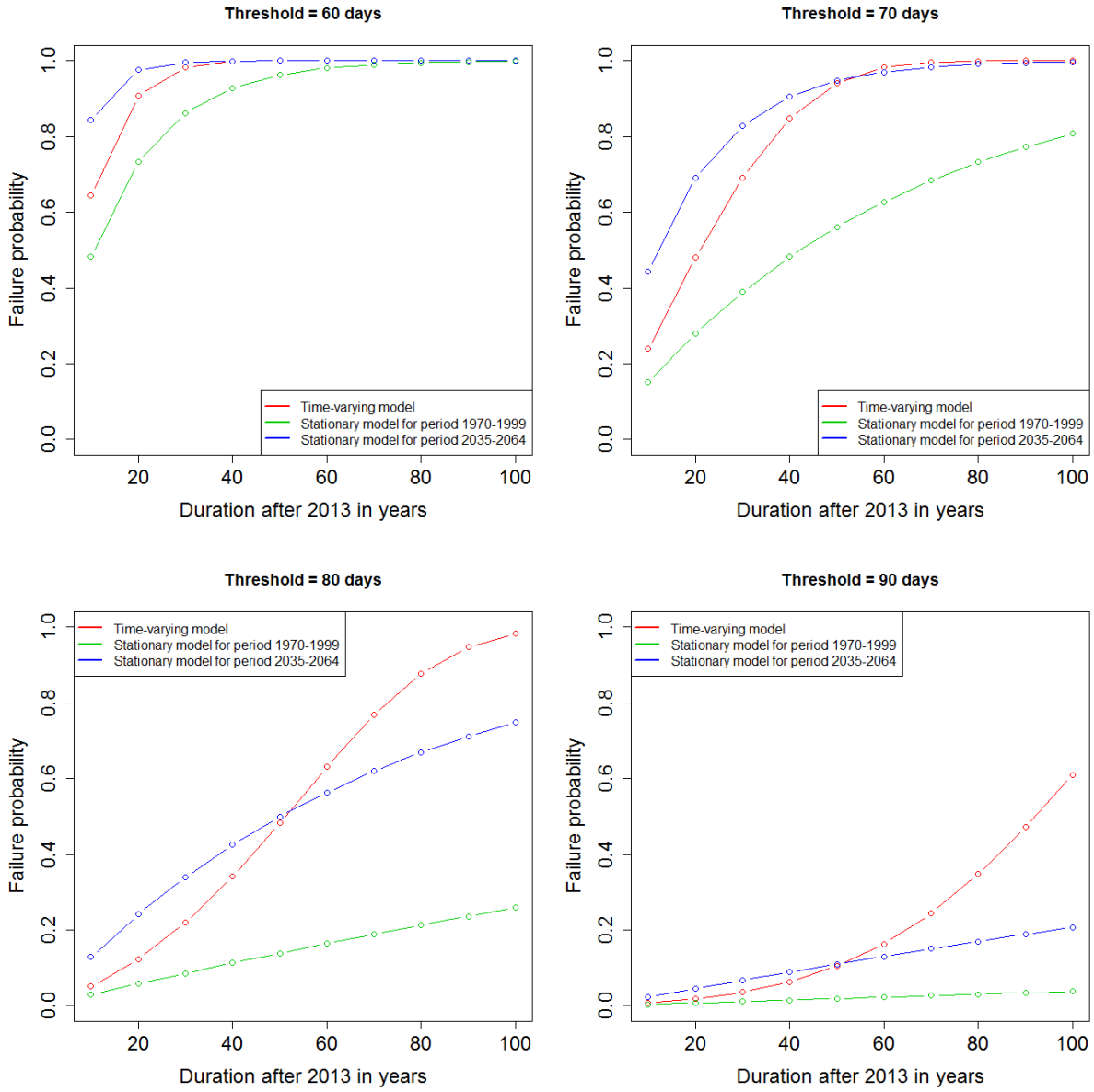


Figure 3.9-Failure probability for the first snow happening later than 1st Sept + 60, 70, 80 and 90 days at least once during the n years following 2013.

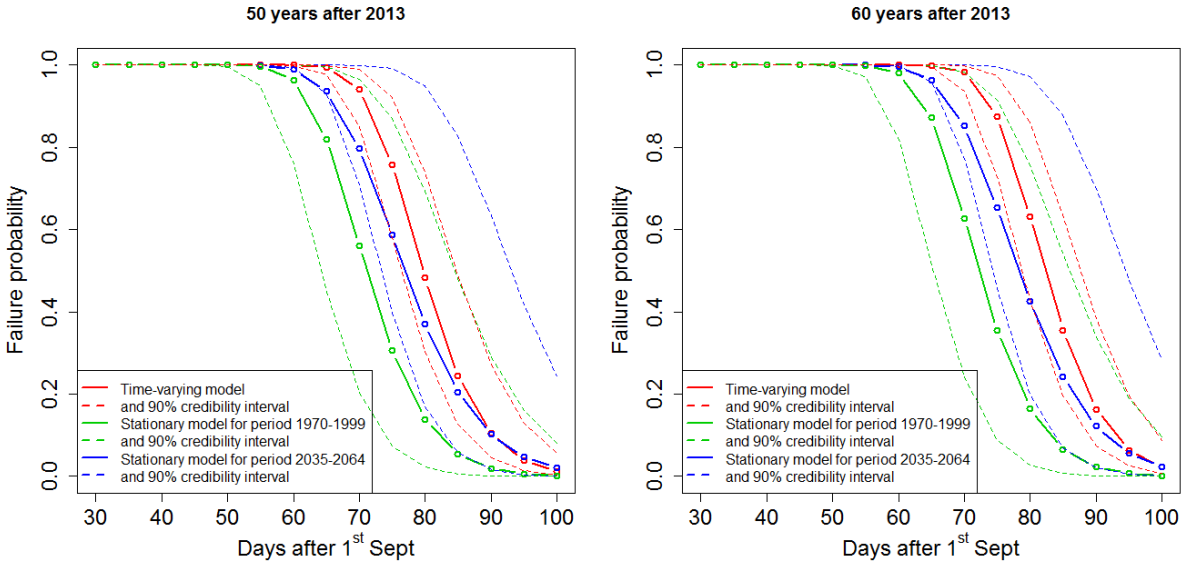


Figure 3.10- Failure probability for the first snow happening Q days later than 1st Sept at least once during the 50 (left) and 60 (right) years following 2013.

1.9 Conclusion and discussion

In this case study, we demonstrate the usage of the general time-varying framework in a context where no strong guidance exists to select the parent distribution. Both discrete and continuous parent distributions are hence trialed. As a result, the negative binomial distribution and the Normal distribution both provide a good description for the variable "first snowy day". Thus, the flexibility of the framework to choose a parent distribution is highlighted.

Moreover, the existence of a temporal trend is assessed through a linear regression model with time as covariate. The flexibility in the choice of the regression functions provides a convenient way to evaluate stationary and non-stationary hypotheses by using the model selection tools. In this study, all three criteria (*AICc*, *BIC* and *DIC*) are in agreement to suggest that a non-stationary model is more adequate than the stationary model for the variable "first snowy day".

Projected quantiles are estimated with both a stationary sub-periods model and the non-stationary trend model. Although the results of both models are consistent, choosing the length of the sub-periods is a limitation: it results from a tradeoff between a period short enough to make the stationarity approximation acceptable and a period long enough to have reasonable uncertainty.

The comparison of failure probabilities evaluated under the two hypotheses also highlights the importance of the model used to describe the evolution of the studied variable. The failure probabilities calculated with the non-stationary trend model increases faster than the stationary model with respect to the duration. This probability based on the stationary model is still useful for a limited duration. However, for longer durations, a continuously-varying model is more adapted to the computation of failure probabilities.

2 *NAO* effects and temporal trends in extreme precipitation in Mediterranean France

In this section, we analyze temporal trends and the effects of *NAO* on the intensity of extreme precipitation in Mediterranean France by using the general climate-informed FA framework. Six regression models under three competing hypotheses regarding the temporal variability of extreme precipitations are studied. Throughout this section, some randomly selected sites (e.g. sites 13 and 60) are used to illustrate the main steps involved in the construction of a local model and its use for prediction: specification of the building blocks of the model, exploration of the posterior distribution, goodness-of-fit evaluation, model comparison and prediction, etc.

2.1 Data

2.1.1 Precipitation data

The precipitation dataset comprises daily precipitation series from 92 precipitation gauges located in Mediterranean France (Figure 3.11). The record starting years among these sites are ranging from 1887 to 1949, and most of them finish in 2004 (yielding record lengths between 57 years and 118 years). The median and average record lengths are 65 and 70 years, respectively.

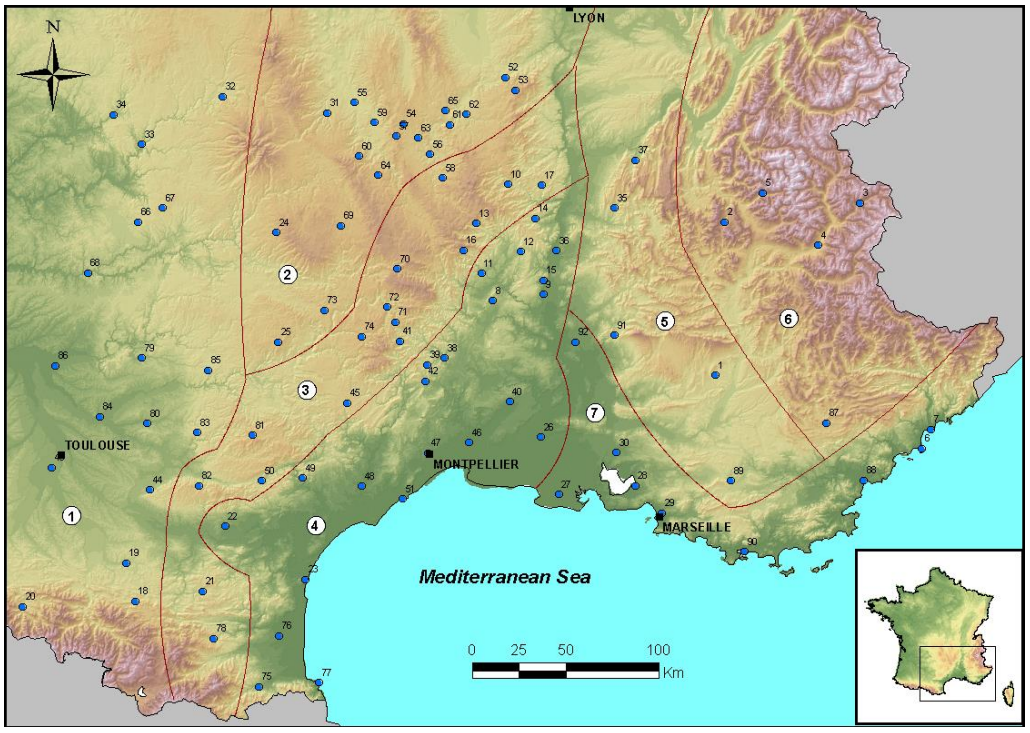


Figure 3.11-Location of the 92 precipitation gauges (blue dots) and homogeneous regions (numbers in white circles) as defined by Pujol et al. [2007].

These 92 stations are classified into seven homogenous zones as defined by *Pujol et al.* [2007]. Each zone contains a different number of observation stations. The precipitation regime within each zone is considered as homogeneous for the extreme events (although the distribution of extremes may vary from site to site due to elevation or exposition effects).

In this study, we focus on the annual maximum daily precipitation (from January to December), which are extracted from the daily precipitation series.

2.1.2 Covariate: the *NAO* index

The North Atlantic Oscillation (*NAO*) is one of the major large-scale modes affecting the climate variability in the Northern Hemisphere. The *NAO* index is defined by the difference of the normalized sea level pressure between Gibraltar and Southwest Iceland [*Jones et al.*, 1997]. The *NAO* index used in this study is obtained from the Climatic Research Unit of the University of East Anglia, in which monthly *NAO* values are available (http://www.cru.uea.ac.uk/~timo/datapages/naoi.htm). In this database, the minimum value of the *NAO* index is about -3.5, and the maximum is about 4.

In this study, we are interested in the annual maximum daily precipitation, thus we use the annual average *NAO* index as covariate (Figure 3.12), which is computed from the monthly *NAO* series. Besides the *NAO* index, another covariate is time, which is used to assess the existence of temporal trends in extreme precipitation.

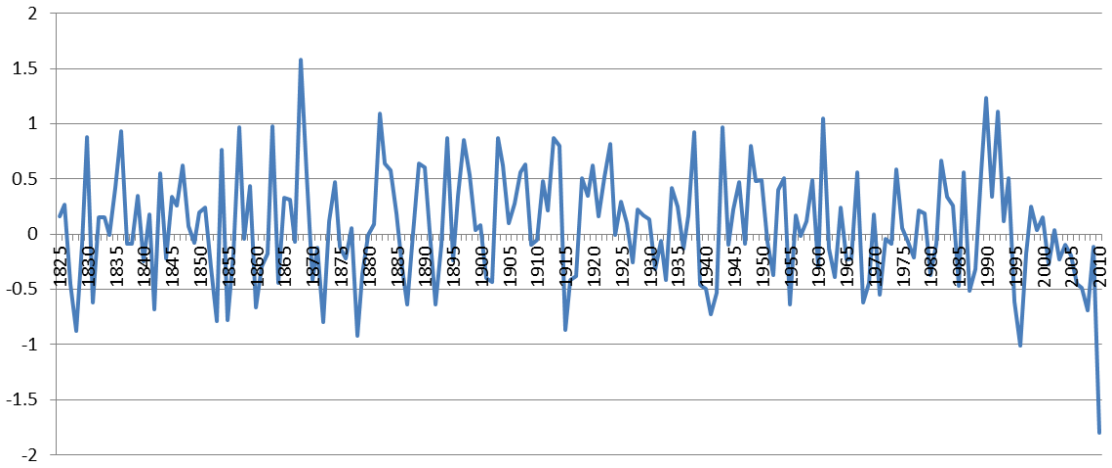


Figure 3.12-Annual average of the NAO index

2.2 Regression models under three competing hypotheses

Following the model construction introduced in Chapter 2, we assume a GEV distribution for the annual maximum daily precipitation [*Coles et al.*, 2003; *Katz et al.*, 2002]. For analyzing the temporal trend and the impact of *NAO*, six competing regression models associated with different hypotheses are introduced. In these six models, the shape parameter is always assumed to be constant. This is because of the well-recognized difficulty in estimating this parameter using relatively short local series.

Table 3.4 lists the six competing regression models. Linear regression models are applied to the location or the scale parameter or both. An exponential inverse link function is applied for the scale parameter. The subscript on the top right (resp. bottom right) of the name "*GEV*" denotes the model for the location (resp. scale) parameter. "*0*" means stationary, "*t*" means linear with respect to time and "*t,NAO*" means linear with respect to both time and *NAO*.

A model selection tool will be used to judge which hypothesis is the most suitable. Under the same hypothesis, we can also evaluate which model is better for presenting the temporal effect on the extreme precipitation in Mediterranean France.

In this study, flat priors are used for the regression parameters.

Table 3.4-Six competing regression models for the extreme precipitation in Mediterranean France

Name	Model	Hypothesis
GEV_0^0	$GEV\left(\mu_0, \exp(\sigma_0), \xi\right)$	Stationary
GEV_0^t	$GEV\left(\mu_0 + \mu_1 t, \exp(\sigma_0), \xi\right)$	Temporal trend
GEV_t^0	$GEV\left(\mu_0, \exp(\sigma_0 + \sigma_1 t), \xi\right)$	
GEV_t^t	$GEV\left(\mu_0 + \mu_1 t, \exp(\sigma_0 + \sigma_1 t), \xi\right)$	
$GEV_0^{t,NAO}$	$GEV\left(\mu_0 + \mu_1 t + \mu_2 NAO(t), \exp(\sigma_0), \xi\right)$	Temporal trend and *NAO* effect
$GEV_{t,NAO}^{t,NAO}$	$GEV\left(\mu_0 + \mu_1 t + \mu_2 NAO(t), \exp(\sigma_0 + \sigma_1 t + \sigma_2 NAO(t)), \xi\right)$	

2.3 Posterior distribution of regression parameters

Regression parameters are estimated using a MCMC sampler in the Bayesian framework, as detailed in Chapter 2 (Section 2.1.2). As an illustration, Figure 3.13 shows the posterior distribution of seven regression parameters in model $GEV_{t,NAO}^{t,NAO}$ for site 60. From the posterior distribution, the temporal trend and the effects of *NAO* are close to zero for both the location and the scale parameters. The uncertainties of the regression parameters are very large. This is because it is hard to provide a good estimation in such a complex model with only 60 observations. Thus weak temporal trend and *NAO* effects are more likely to be masked.

2.4 Goodness-of-fit

As explained in Chapter 2 (Section 3.1), the goodness-of-fit can be evaluated graphically using a PP plot. The modal parameters (maximizing the posterior pdf) are used to compute the theoretical probability in the PP plot.

Figure 3.14 shows the PP plot for site 60 as an illustration. The PP plot of most sites is close to the diagonal for all six regression models, which indicates that all six models provide a good description of the observations. Although comforting at first sight, this observation

also highlights a strong limitation of such a PP plot diagnostic: its power to distinguish between competing model hypotheses is very low and does not enable a conclusive assessment of the most suitable model. This will be improved by considering model comparison tools in subsequent Section 2.8.

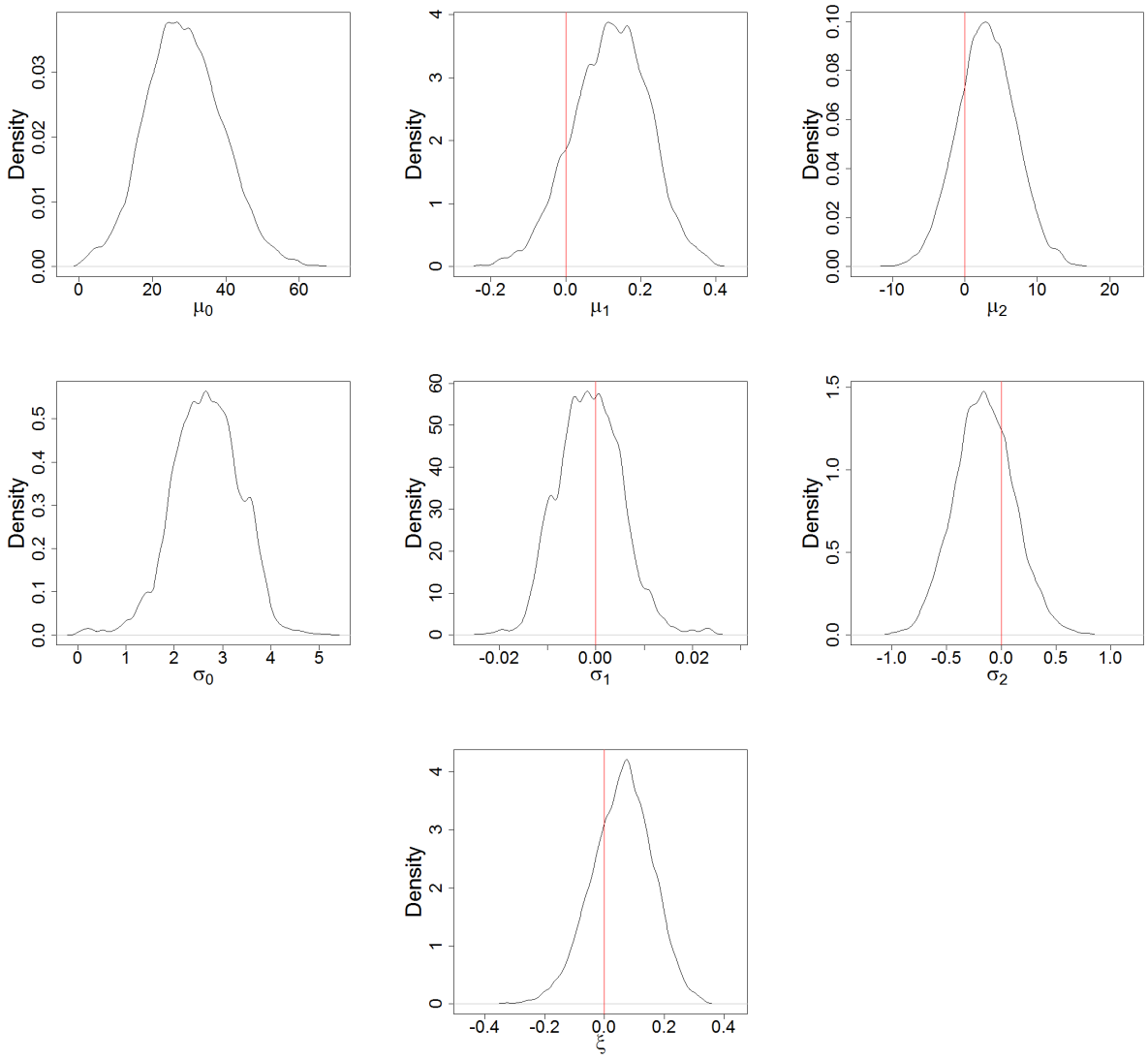
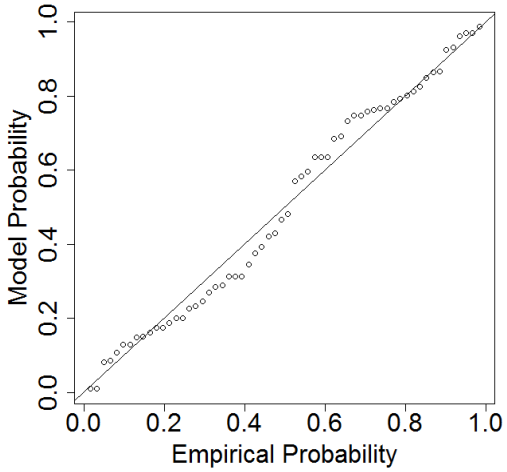
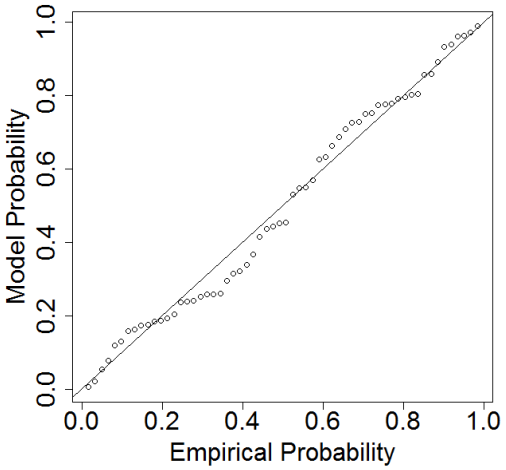


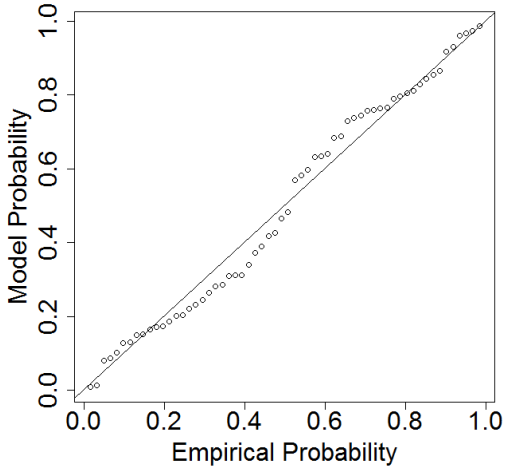
Figure 3.13-Posterior distribution of seven regression parameters in $GEV_{t,NAO}^{t,NAO}$ ***model for site 60***



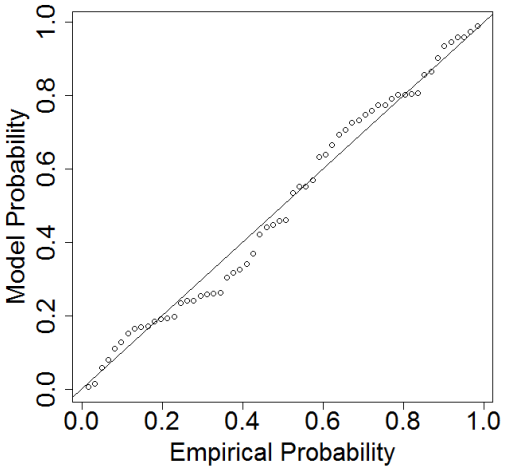
(a) GEV_0^0



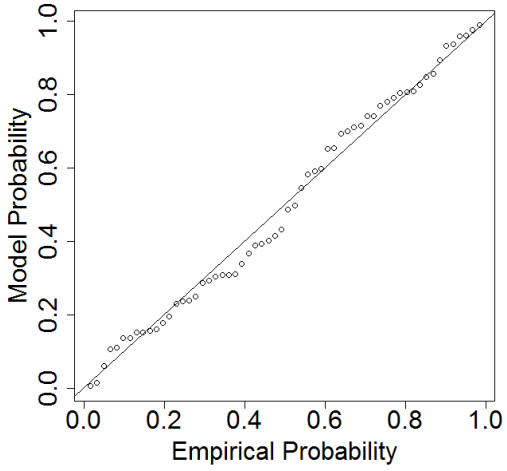
(b) GEV_0^t



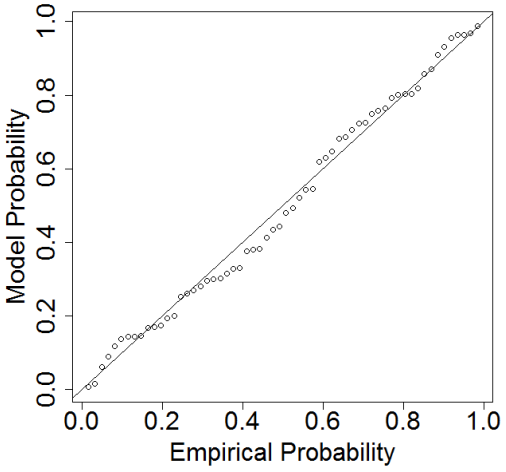
(c) GEV_t^0



(d) GEV_t^t



(e) $GEV_0^{t,NAO}$



(f) $GEV_{t,NAO}^{t,NAO}$

Figure 3.14-Probability-probability plot of the six regression models for site 60

2.5 Conditional predictions

For each of the three hypotheses we aim to evaluate (stationarity, temporal trend, temporal trend + NAO effect), we restrict to the most complete models (GEV_0^0, GEV_t^t and $GEV_{t,NAO}^{t,NAO}$) to illustrate the predictions that can be made from the general FA framework.

As discussed in Chapter 2 (Section 2.3), in a stationary case, the GEV quantile y_p associated with an exceedance probability *1-p* is calculated as follows:

$$y_p = \frac{\sigma}{\xi} K_p + \mu \tag{3.2}$$

where $K_p = 1 - \left(-\log(p)\right)^{\xi}$

In a time-varying case, μ and σ are varying with time according to the specific regression models. For example, in the $GEV_{t,NAO}^{t,NAO}$ model, the quantile y_p in formula (3.2) becomes:

$$\begin{aligned} y_p &= \frac{\sigma_0 + \sigma_1 t + \sigma_2 NAO(t)}{\xi} K_p + \mu_0 + \mu_1 t + \mu_2 NAO(t) \\ &= \left(\mu_0 + \frac{\sigma_0}{\xi} K_p\right) + \left(\mu_1 + \frac{\sigma_1}{\xi} K_p\right) t + \left(\mu_2 + \frac{\sigma_2}{\xi} K_p\right) NAO(t) \end{aligned} \tag{3.3}$$

With N_{sim} MCMC samples, quantiles $\left(y_p^{(k)}\right)_{k=1,N_{sim}}$ are calculated. Thus the credibility interval is obtained directly for these quantiles.

Figure 3.15 shows the annual maximum precipitation data, the median and the 0.99-quantile for site 60 according to these three models. Dashed lines correspond to 90% credibility intervals for the median and the 0.99-quantile. With GEV_t^t model, there is a slight temporal trend on the median and the 0.99-quantile. This trend still holds with $GEV_{t,NAO}^{t,NAO}$ model, and a fluctuation appears according to the values taken by the *NAO* index. In general, the predication with these three models is similar. This is because the temporal trend and *NAO* effects are weak compared with the natural variability of annual maximum precipitation, thus the variation of prediction between these models is limited. When looking at the 0.99-quantile, the uncertainties are still large.

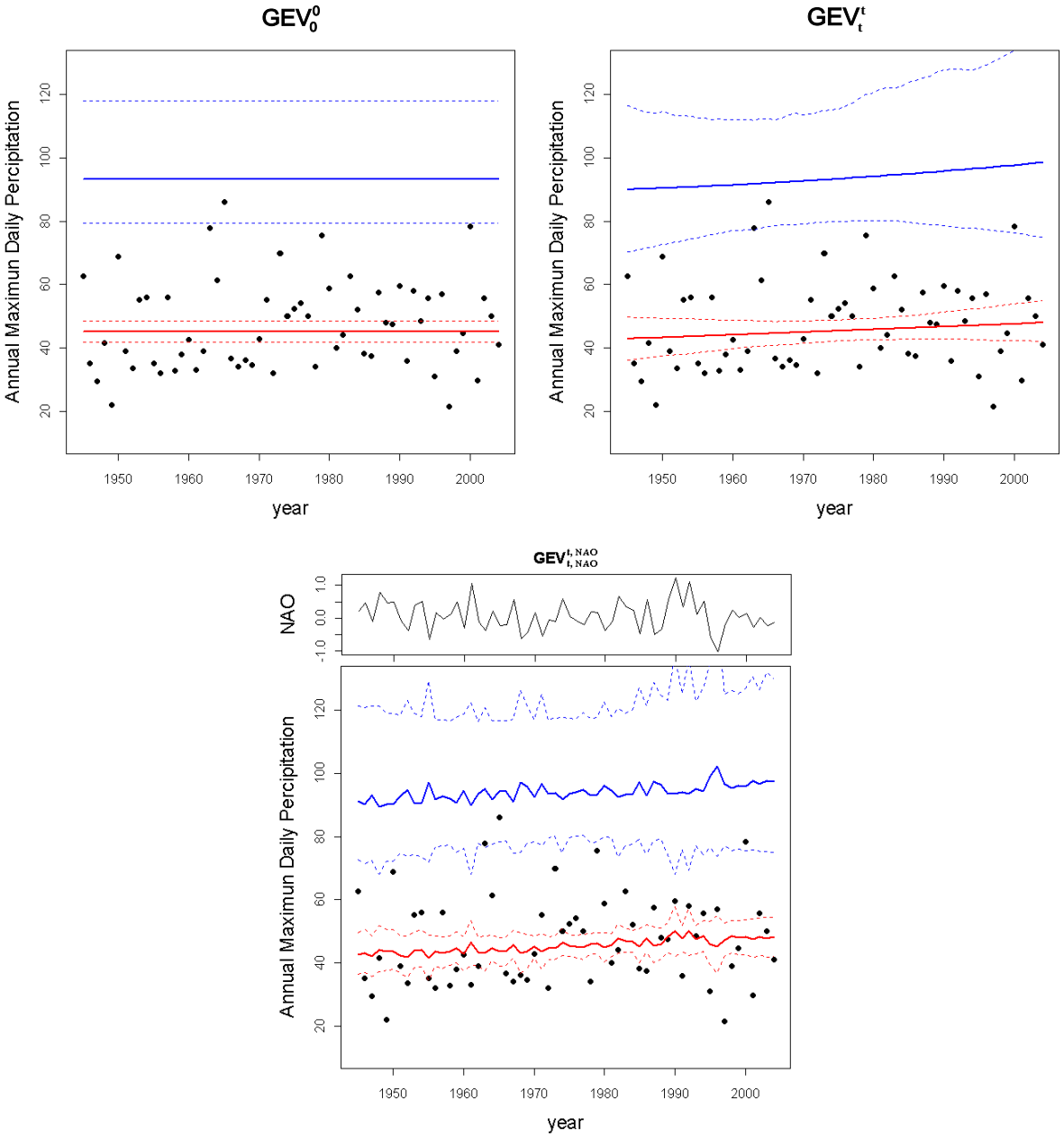


Figure 3.15- Observation, median and 0.99-quantile for site 60 according to three regression models. Black dots represent the observation of annual maximum precipitation. Red and blue lines are respectively 0.5 and 0.99 quantile. The dashed lines correspond to 90% credibility intervals.

Figure 3.16 provides an alternative representation of the predictions by the three models, by showing the posterior distribution of the 0.9-quantile according to the three models for site 13. The 0.9-quantiles are computed at t equal to 2012. In order to highlight the difference between predictions for different values of NAO, we use two extreme values NAO = -1 and NAO = 1. The figure shows that uncertainty increases with the complexity of the model: the stationary model provides the prediction with the smallest uncertainty, while $GEV_{t,NAO}^{t,NAO}$ model yields the largest uncertainty. For site 13, the temporal trend is not clear since the posterior distribution for model GEV_t^t is similar to that of the stationary model. On the other hand,

when *NAO* is considered ($GEV_{t,NAO}^{t,NAO}$ model), a stronger difference appears between the distributions of 0.9-quantiles conditioned on the two extreme *NAO* values. However, the prediction uncertainty associated with $GEV_{t,NAO}^{t,NAO}$ model remains too large to yield an unambiguous conclusion regarding the difference between the two NAO-conditional predictions.

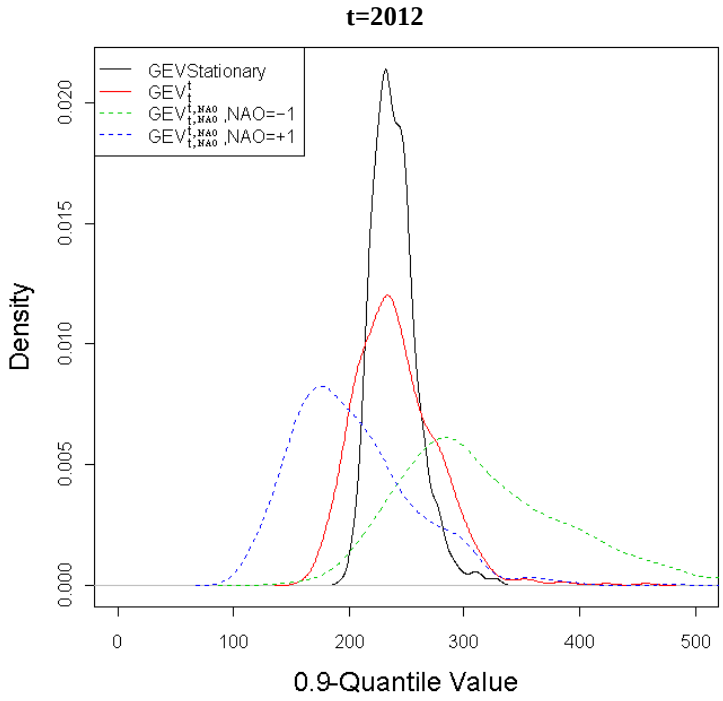


Figure 3.16-Posterior distribution of the 0.9-quantile based on three models for site 13

2.6 Temporal trend and *NAO* impact for all 92 sites

One of the objectives of this case study is to evaluate whether the temporal trend and the *NAO* effect are significant for extreme precipitations in Mediterranean France. In the models proposed in Table 3.4, the regression parameters μ_1 and σ_1 characterize the temporal trend, while μ_2 and σ_2 characterize the effect of *NAO*. If the effect is significant, these parameters will be significantly different from 0. In Figure 3.13, the posterior distributions of these four parameters are spread around 0, which indicates that the temporal trend and the *NAO* effect are not significant for site 60.

In order to extend this assessment to the whole dataset of 92 sites, Figure 3.17 presents the boxplots of the posterior distributions of μ_1 and μ_2 for all sites (reorganized by region as symbolized by the colored dots in the Figure). Although the boxplots tend to be preferentially positive for most stations, we still cannot make any definitive conclusion for the regional trend since most boxplots are not far away from zero. This can be mostly explained by the very large uncertainty affecting the estimation of parameter μ_1 and μ_2.

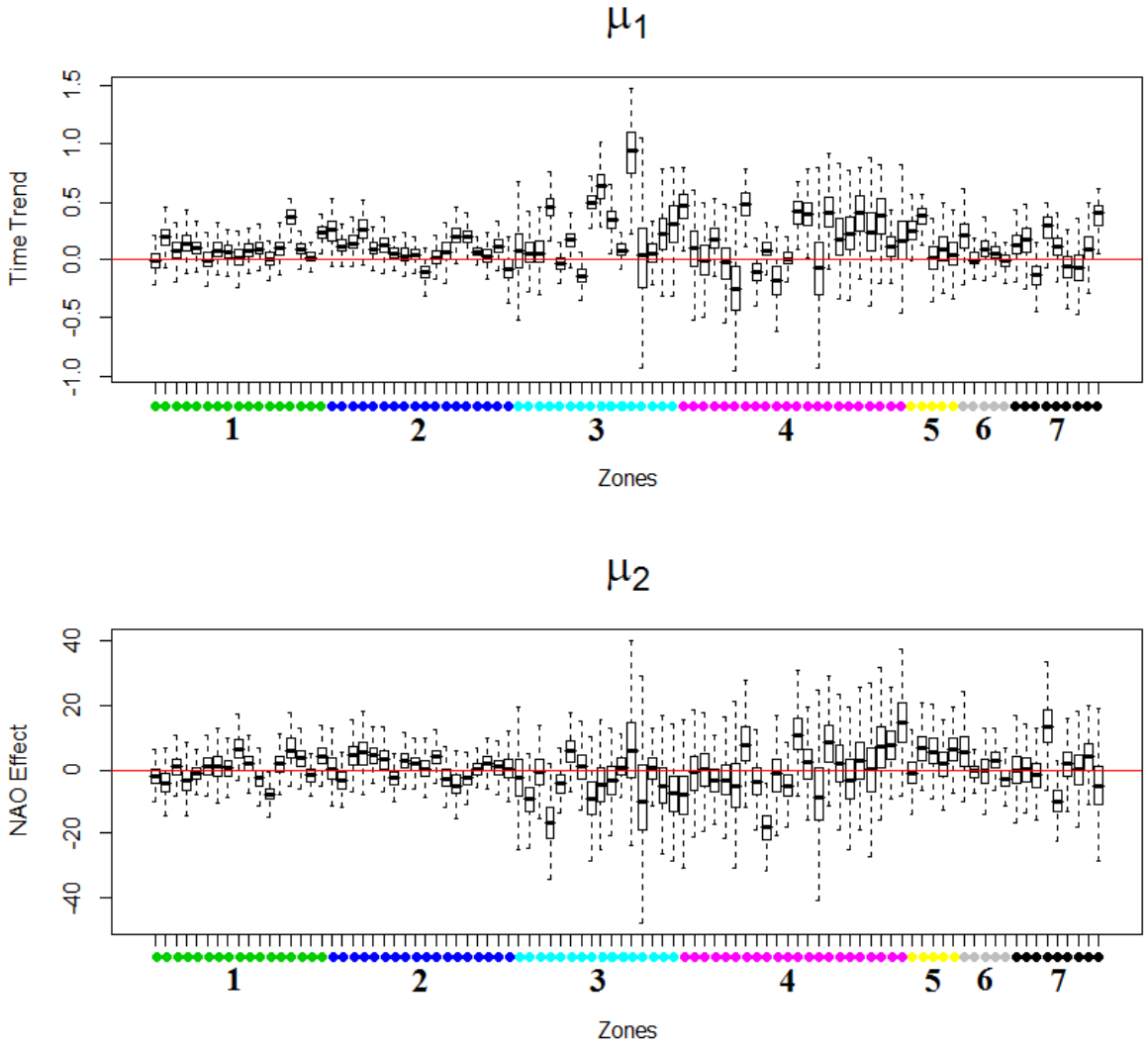


Figure 3.17-Boxplot of the posterior distribution of parameter μ_1 and μ_2 for all 92 sites in the model $GEV_{t,NAO}^{t,NAO}$ **. Colored dots denote the homogeneous regions the stations belong to.**

2.7 Conditional quantiles for all 92 sites and their uncertainty

In Section 2.5, we observed that the uncertainty in predicted quantiles increases with the complexity of the model. In this section, we give a general view of the prediction for all sites under three hypotheses. Figure 3.18 shows the boxplot of 0.9-quantile for all sites with GEV_0^0, GEV_t^t and $GEV_{t,NAO}^{t,NAO}$ models. The number of parameters in these models is respectively three, five and seven. Sites with same color are in the same climatic zone. Covariates are fixed to t = 2004 and NAO = -0.1. In general, the boxes size increases with the number of parameters. For example, at the site that provides the biggest 0.9-quantile value in the stationary model (the fourth site from the right in the third zone), the posterior distribution of the 0.9-quantile lies approximately between 200 and 300 with GEV_0^0 model. Based on GEV_t^t model the

distribution goes up to 350. It reaches more than 500 with $GEV_{t,NAO}^{t,NAO}$ model. Once again, this illustrates that the use of more complex models to describe temporal variability comes at the cost of much larger uncertainties, at least within the local estimation framework considered here.

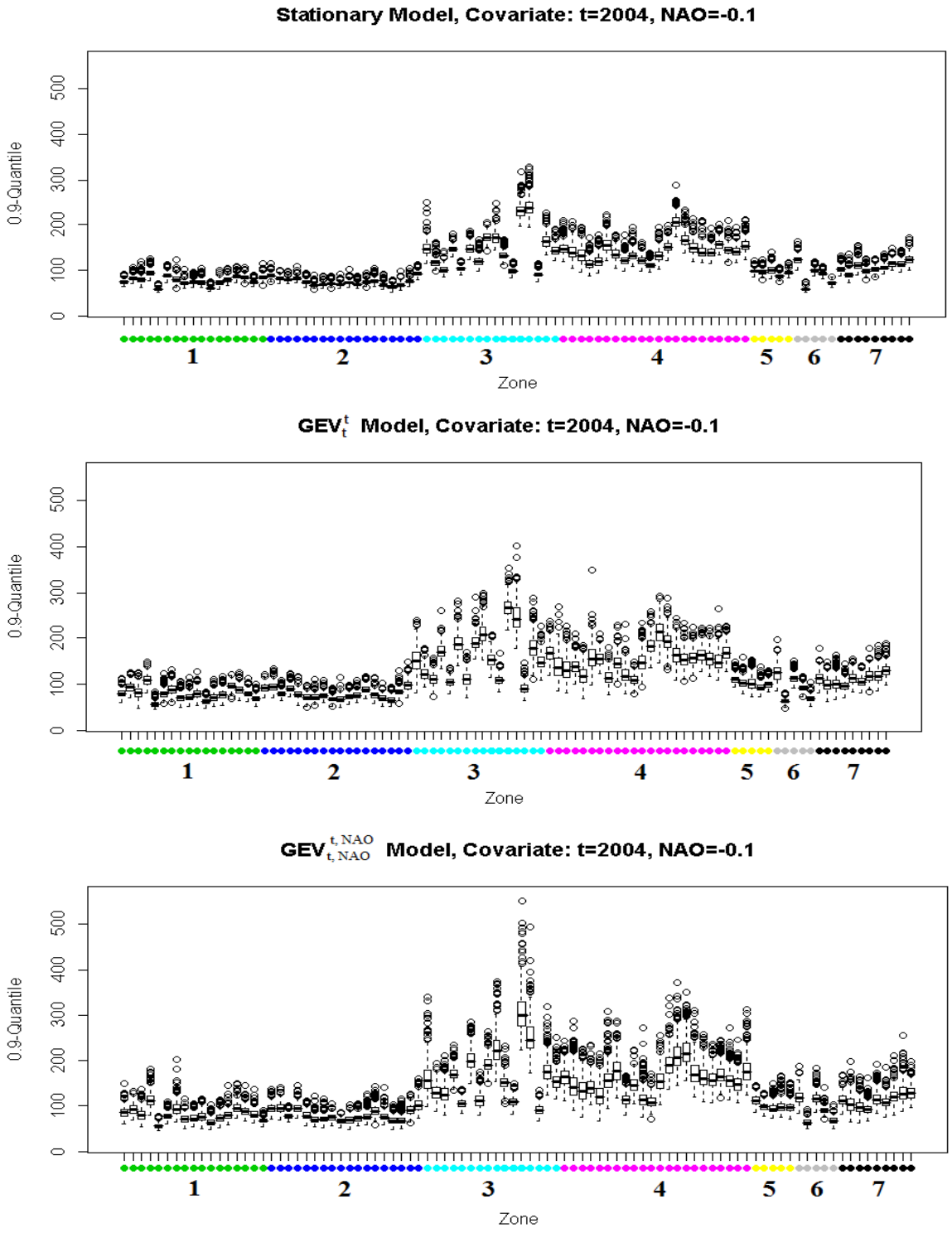


Figure 3.18-Boxplot of 0.9-quantiles for all sites within seven zones (denoted by colored dots). Covariates are chosen as time t=2004 and NAO=-0.1 (relatively weak).

2.8 Model comparison

AICc, *BIC* and *DIC* values are computed for each model. Table 3.5 gives the maximum log-likelihood, *AICc*, *BIC* and *DIC* values for site 60. In general, the maximum likelihood increases with the model complexity, since complex model contains more degree of freedom and hence better fit the observations. For site 60, the log-likelihood value of the stationary (GEV_0^0) model is the smallest, as theoretically expected. However, it also gives the smallest value for all three criteria. This confirms our finding in Section 2.6 that neither the temporal trend nor the *NAO* effect is detected for extreme precipitation in site 60.

Table 3.5-Maximum log-likelihood, AICc, BIC and DIC values for site 60

Site 60	GEV_0^0	GEV_0^t	GEV_t^0	GEV_t^t	$GEV_0^{t,NAO}$	$GEV_{t,NAO}^{t,NAO}$
MaxlogL	-241.532	-241.059	-241.505	-241.046	-240.796	-240.654
AICc	489.065	490.118	491.011	493.093	492.592	497.308
BIC	495.348	498.495	499.388	502.564	502.064	509.968
DIC	488.544	489.815	490.308	491.710	490.963	494.358

Table 3.6 summarizes this model comparison exercise for all 92 sites by counting the number of sites for which each model is ranked as "best" based on the *AIC*, *BIC* and *DIC* criteria. Among all 92 sites, GEV_0^0 and GEV_0^t are two most frequently selected models. In particular, the stationary (GEV_0^0) model seems to be the most adapted for the majority of sites. In general, complex models with a large number of parameters are not frequently selected. However, it should be reminded that time trends and *NAO* effects may exist in some sites, but be masked because of the large estimation uncertainties.

Table 3.6-Number of sites for which each model is ranked as "best" based on AICc, BIC and DIC criteria

	GEV_0^0	GEV_0^t	GEV_t^0	GEV_t^t	$GEV_0^{t,NAO}$	$GEV_{t,NAO}^{t,NAO}$
AICc	49	19	8	6	7	3
BIC	74	15	2	1	0	0
DIC	47	17	7	8	7	6

Compared with *AICc* and *DIC*, *BIC* is known as the most penalizing criterion for models with many parameters [*Spiegelhalter et al.*, 2002], which corresponds to the results of Table 3.6. Moreover, there are still respectively 3 and 6 sites for which the $GEV_{t,NAO}^{t,NAO}$ model is the best according to *AICc* and *DIC*. It turns out that the three sites where *AICc* select this model are included in the six sites where *DIC* selects it. It means that the results of these two criteria are in agreement. Figure 3.19 maps the sites where *NAO*-accounting models ($GEV_0^{t,NAO}$, $GEV_{t,NAO}^{t,NAO}$) are selected by *AICc* and *DIC*. These sites are located in different

zones, which seem surprising since we would expect *NAO* effects to show some spatial consistency. Consequently, this local analysis is not conclusive regarding the impact of *NAO* on extreme precipitations in the Mediterranean area.

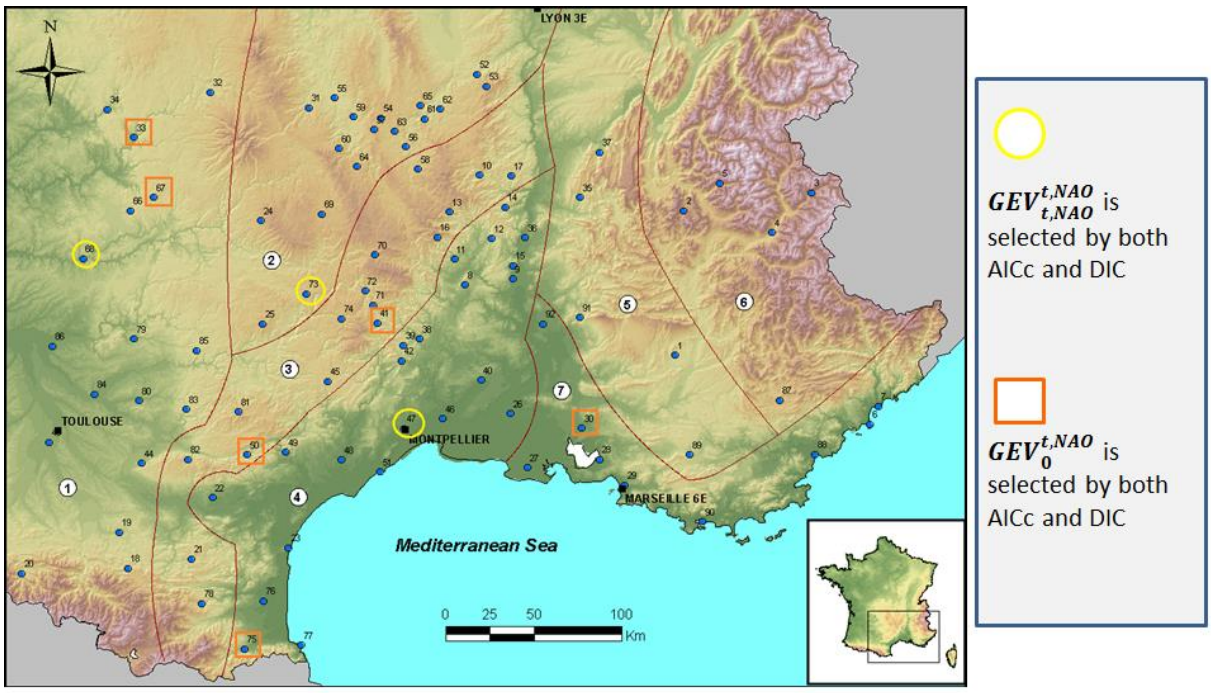


Figure 3.19-Map of sites where NAO-accounting models are selected by AICc and DIC criteria.

2.9 Discussion and conclusion

In this section, we analyzed the existence of temporal trends and *NAO* effects on annual maximum daily precipitation in the Mediterranean area. Through this study, we illustrated the usage of the local modeling framework for describing precipitation extremes, and highlighted its flexibility and generality.

With a pre-specified GEV parent distribution, we compared six regression models under three hypotheses of stationarity, temporal trend only, and both temporal trend and *NAO* effect. Comparing the posterior distributions of parameters and using model selection criteria, we found no conclusive evidence of an impact of *NAO*. Although some evidence of a temporal trend may exist, it is still difficult to identify it at the local scale due to the insufficient observations and large uncertainties.

This case study could be further developed in several ways. Firstly, annual average values of *NAO* are used in this study. This temporal resolution may be too coarse to detect a significant effect of *NAO*: the *NAO* effect may be strong during some periods, but weak during other periods. Thus, studying the effect of *NAO* at the seasonal scale may yield different conclusions.

Secondly, since the sample sizes are not sufficient to provide a precise estimation for each regression parameter, making an extrapolation for a high return period will lead to large

uncertainties. One solution to increase the sample size is to extract data by the peak over threshold (POT) method [*Lang et al.*, 1999]. However, the problem might more fundamentally reflect the inherent limitation of identifying temporal variations at the local scale, especially for extremes. Another more promising solution would therefore be to move from a local to a regional analysis.

When looking at the boxplots of different parameters (e.g. trend parameters: Figure 3.17, shape parameter: Figure 3.20), the boxes in the same zone have similar values. For instance, in Figure 3.17, boxes for sites in zone 1 (green) are close to 0 and are similarly ranging from -0.5 to 0.5. The shape parameter (Figure 3.20) for sites within the same zone also reveals consistency. It indicates that such parameters could be assumed identical for all sites within each region. Thus, in a regional model, we could set common parameters for different sites, which may yield a more precise estimation on these parameters, and hence obtain a more robust identification of temporal signals.

In the following chapter, we will introduce a regional modeling framework to implement such regional analyzes.

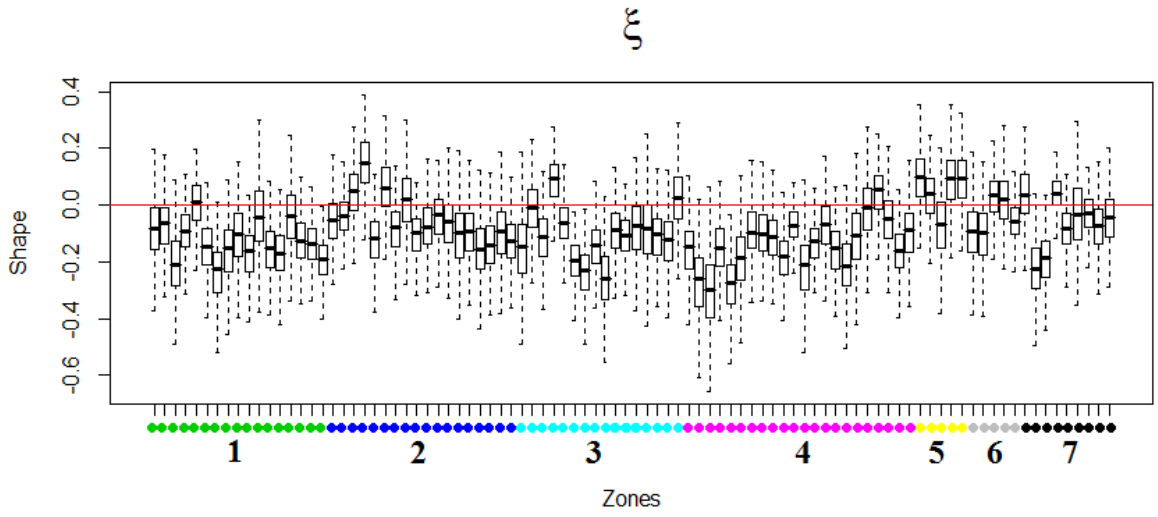


Figure 3.20-Boxplot of the shape parameter for all 92 sites in the model $GEV_{t,NAO}^{t,NAO}$ **.**
Each color denotes one zone.

Part II Time-varying frequency analysis framework: Regional model

CHAPTER 4 Development of a general spatio-temporal regional frequency analysis framework

The case studies performed with local models highlighted that large uncertainties affect parameter inference and model predictions. To overcome this limitation, we extend the local modeling framework to the regional scale. In the regional modeling framework, data of all sites are clustered together to perform the inference. Spatial effects can also be considered at this regional scale. This chapter describes in details the building blocks of such spatio-temporal RFA framework: spatio-temporal regressions, spatial dependence modeling as well as the parameter inference tools, diagnostic tools and model selection tools.

1 Regional model construction

The objective of this section is to develop a general regional FA framework. Compared with the local model, the main advantage of the regional model is that spatial effects are considered. More precisely, in the regional model, two types of covariates are involved in addition to the temporal covariates used in Chapter 2: spatial and spatio-temporal covariates. Spatial effects can thus be integrated through regression functions linked to these two types of covariates. Moreover, the spatial dependence of data is considered with elliptical copulas in this framework.

Figure 4.1 illustrates the general principle of the regional model. Observations $\boldsymbol{Y}$ of all sites at different time steps are used together. Through the local regression parameters $\boldsymbol{\theta}_{loc}$ and regional regression parameters $\boldsymbol{\theta}_{reg}$, both site-specific parameters and regionalized parameters can be expressed conveniently. This construction provides us with a flexible and convenient framework to model both the temporal variation and the associated spatial effects. The following sections describe in more details each building block of the regional model.

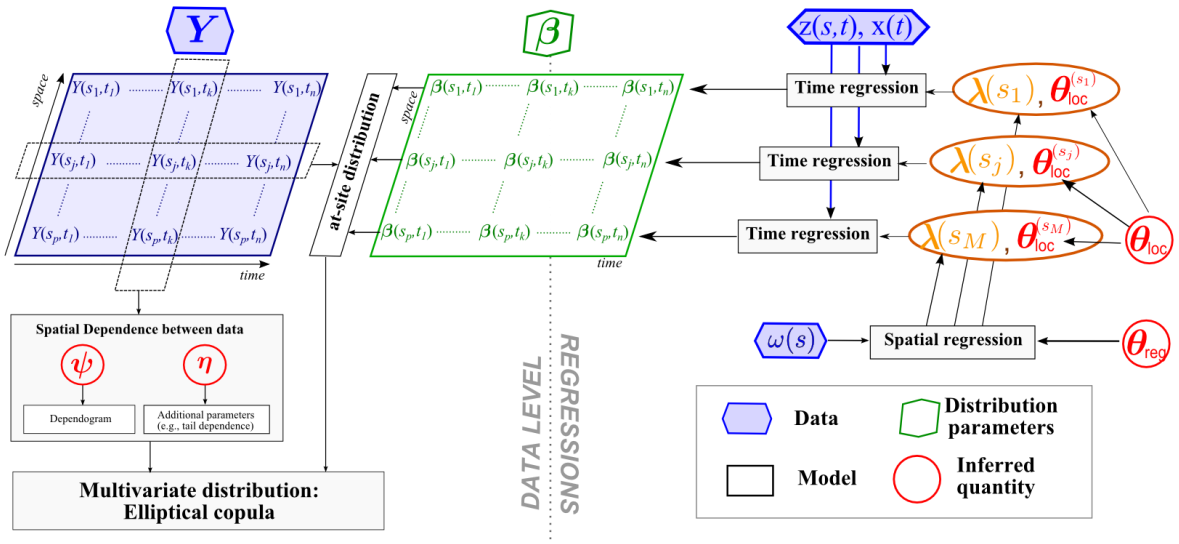


Figure 4.1- Schematic of the Regional model

1.1 Parent distribution

In a regional context, data from several sites are used together. The notation in Chapter 2 is hence modified to introduce spatial variations. Let $Y(s,t)$ be the observation at site s and time t and $\boldsymbol{Y} = \left(Y(s_j,t_k),\, j=\{1,2,...,p\}, k=\{1,2,...,n\}\right)$ be the collection of observed data at all p observation sites for n time steps. Similarly to the local model, a common distribution D is assumed for all sites, but with parameters varying in both space and time:

$$Y(s,t) \sim D(\boldsymbol{\beta}(s,t)) \tag{4.1}$$

where $\boldsymbol{\beta}(s,t)=\left(\beta_i(s,t), i=\{1,2,...,m\}\right)$ is the collection of all distribution parameters: m is the number of distribution parameters of D; $\beta_i(s,t)$ is the i^{th} distribution parameter at time t and site s.

1.2 Spatio-temporal regression

1.2.1 Three type of covariates

Similarly to the local model, regressions are used to describe spatio-temporal variations in the parameters $\beta_i(s,t)$. However, the temporal variations of the local model are extended in the regional model to include three different kinds of covariates:

- Temporal covariates $\boldsymbol{x}(\boldsymbol{t})$: e.g. time, *SOI* (Southern Oscillation Index), *NAO* (North Atlantic Oscillation);
- Spatial covariates $\boldsymbol{\omega}(\boldsymbol{s})$: e.g. altitude, coordinates;
- spatio-temporal covariates $\boldsymbol{z}(\boldsymbol{s},\boldsymbol{t})$: e.g. temperature.

Temporal covariates only change over time (but are common to all sites), and spatial covariates only change over sites (but do not change in time). Spatio-temporal covariates change over both these two dimensions.

1.2.2 Construction of regression models: a two-step approach

The regional regression is established in two steps: specification and spatialization. The first step establishes site-specific regressions with temporal and spatio-temporal covariates. The second step establishes a spatial model with the spatial covariates (see Figure 4.1):

1. **Specification step**: specify the time model using at-site regressions for a distribution parameter $\beta_i(s,t)$:

   For a given site s:

$$\beta_i(s,t)=l_i^{-1}\left(h_i(\boldsymbol{x}(\boldsymbol{t}),\boldsymbol{z}(\boldsymbol{s},\boldsymbol{t});\boldsymbol{\theta}(\boldsymbol{s}))\right) \tag{4.2}$$

   where l_i^{-1} is the inverse link function, h_i is the regression function, $\boldsymbol{x}$ and $\boldsymbol{z}$ are temporal and spatio-temporal covariates, and $\boldsymbol{\theta}(\boldsymbol{s})$ are the regression parameters. This step is exactly the same as the regression with temporal covariates for a local model, as introduced in Chapter 2 (Section 1.2).

2. **Spatialization step**: regression parameters $\boldsymbol{\theta}(\boldsymbol{s})$ are split into two groups: $\boldsymbol{\theta}(\boldsymbol{s})=\{\boldsymbol{\theta}_{\boldsymbol{loc}}^{(s)};\boldsymbol{\lambda}(\boldsymbol{s})\}$, where $\boldsymbol{\theta}_{\boldsymbol{loc}}^{(s)}$ is the collection of purely local parameters, whose value remains specific to each site s, and $\boldsymbol{\lambda}(\boldsymbol{s})$ represents all the parameters waiting to be

spatialized. For each of its component $\lambda(s)$, we apply a spatial regression function. This spatial regression is time-invariant: neither spatial regression parameters nor covariates change over time. Hence, at this step, only regional parameters and spatial covariates are used. Thus a spatial regression function *g* is introduced:

$$\lambda(s) = l^{-1}\left(g(\boldsymbol{\omega}(\boldsymbol{s}); \boldsymbol{\theta}_{\boldsymbol{reg}})\right) \tag{4.3}$$

where l^{-1} is the inverse link function, $\boldsymbol{\omega}(\boldsymbol{s})$ is a vector of spatial covariates and $\boldsymbol{\theta}_{\boldsymbol{reg}}$ is a vector of regional regression parameters (identical for all sites). For abbreviation, $\boldsymbol{\theta}_{\boldsymbol{loc}}^{(\boldsymbol{s})}$ and $\boldsymbol{\theta}_{\boldsymbol{reg}}$ are called local and regional R-parameters, respectively.

This two-step mechanism is very general and corresponds to a standard regionalization reasoning. As an illustration, consider the following "trend analysis" situation: the mean of some hydrologic variable is assumed to be a linear function of time (step 1, specification). Then, the slope of this trend may be allowed to vary across sites according to elevation (step 2, spatialization). This two-step mechanism also provides a simple way to develop a hierarchical model in future work, in which a random spatial term could be added directly into step 2.

The following specification illustrates this situation in more details. We assume that the random variables $Y(s,t)$ follow a non-stationary Poisson distribution (one-parameter distribution) with time *t* as the temporal covariate and elevation *el(s)* as the spatial covariate. We are going to establish a model in which the distribution varies with *t* and the slope on *t* varies with the elevation *el(s)* of each site. The two-step procedure is as follows:

$$Y(s,t) \sim Pois(\beta(s,t)) \tag{4.4}$$

Step 1: $$\beta(s,t) = \exp\left(\theta_{loc}^{(s)} + \lambda(s) * t\right) \tag{4.5}$$

Step 2: $$\lambda(s) = \theta_{reg_1} + \theta_{reg_2} * el(s) \tag{4.6}$$

Therefore, the spatio-temporal regression model for $\beta(s,t)$ is $\beta(s,t) = \exp\left(\theta_{loc}^{(s)} + \left(\theta_{reg_1} + \theta_{reg_2} * el(s)\right) * t\right)$, where $\theta_{reg_1}, \theta_{reg_2}$ and $\boldsymbol{\theta}_{loc}^{(s)} = \left(\theta_{loc}^{(s_1)}, \theta_{loc}^{(s_2)}, ..., \theta_{loc}^{(s_p)}\right)$ are the regression parameters that need to be estimated. The exponential inverse link function is used since the parameter of a Poisson distribution should be positive.

1.2.3 Example of spatio-temporal regression models

According to the two-step procedure, the general framework enables establishing regression functions for various specific hypotheses. Table 4.1 lists some examples of spatio-temporal models that can be constructed in this way.

Table 4.1-Example of spatio-temporal regressions.

	Regression model	Two steps	Hypothesis
1	$\beta(s,t) = \theta_{loc}^{(s)} \ \forall t$	*i*) $\beta(s,t) = \theta_{loc}^{(s)} \ \forall t$ *ii*)	Stationary, site-specific model
2	$\beta(s,t) = \theta_{reg} \ \forall s,t$	*i*) $\beta(s,t) = \lambda_1(s) \ \forall t$ *ii*) $\lambda_1(s) = \theta_{reg} \ \forall s$	Stationary, purely regional model
3	$\beta(s,t) = \theta_{loc_1}^{(s)} + \theta_{loc_2}^{(s)} x(t)$	*i*) $\beta(s,t) = \theta_{loc_1}^{(s)} + \theta_{loc_2}^{(s)} x(t)$ *ii*)	Time-varying, purely local model
4	$\beta(s,t) = \theta_{reg_1} + \theta_{reg_2} \omega(s)$	*i*) $\beta(s,t) = \lambda_1(s)$ *ii*) $\lambda_1(s) = \theta_{reg_1} + \theta_{reg_2} \omega(s)$	Stationary model with trend in space
5	$\beta(s,t) = \theta_{reg_1} + \theta_{reg_2} \omega(s) + \theta_{reg_3} x(t)$	*i*) $\beta(s,t) = \lambda_1(s) + \lambda_2(s) x(t)$ *ii*) $\lambda_1(s) = \theta_{reg_1} + \theta_{reg_2} \omega(s)$ $\lambda_2(s) = \theta_{reg_3}$	Separable space and time effects
6	$\beta(s,t) = \theta_{loc_1}^{(s)} + [\theta_{reg_1} + \theta_{reg_2} \omega(s)] x(t)$	*i*) $\beta(s,t) = \theta_{loc_1}^{(s)} + \lambda_1(s) x(t)$ *ii*) $\lambda_1(s) = \theta_{reg_1} + \theta_{reg_2} \omega(s)$	Trend in time, with slope varying in space
7	$\beta(s,t) = \theta_{reg_1} + \theta_{reg_2} z(s,t)$	*i*) $\beta(s,t) = \lambda_1(s) + \lambda_2(s) z(s,t)$ *ii*) $\lambda_1(s) = \theta_{reg_1}$ $\lambda_2(s) = \theta_{reg_2}$	Space and time effects through a spatio-temporal covariate

For implementation convenience, we opted for the use of this 2-step procedure in the thesis, rather than directly implementing the regression models in the first column of Table 4.1. In fact, with quite simple building blocks, it becomes possible to build fairly complex spatio-temporal regressions, which is easier than re-writing a new spatio-temporal regression function at each case study.

1.3 An illustration of the regional model

We use the same example as introduced in Chapter 2 (Section 1.3). But this time, we assume that the temporal trend in the region is site-specific and the impact of *SOI* depends on the distance to the sea *SeaDist*(*s*) of each site *s*. Figure 4.2 illustrates the schematic for modeling the location parameter $\mu(s,t)$. Step 1 gives the temporal regression using the temporal covariates *t* and *SOI*. Step 2 specifies the spatial regression on the parameter quantifying the effect of *SOI*, using the spatial covariate *SeaDist*.

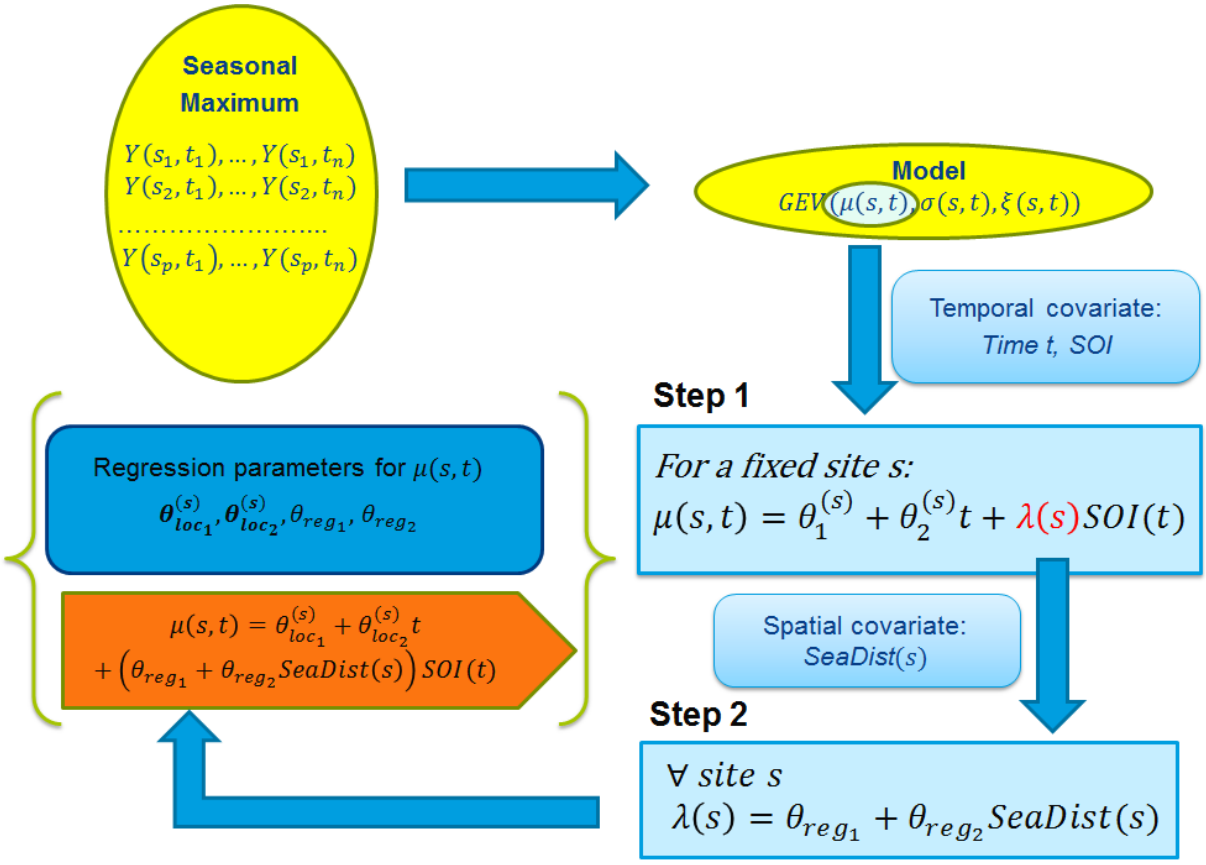


Figure 4.2-Schematic for the construction of spatio-temporal regressions.

1.4 Accounting for spatial dependence between sites

In a region, the observations from different stations are in general not completely independent. Two nearby stations are more likely to record the same rainfall episode than two stations farther away from each other: the dependence between close sites is hence higher.

There exist several ways to model this dependence. In this thesis, we opt for the use of copulas. Max-stable processes are an interesting alternative, especially in the context of extremes, but they are not considered in the proposed framework for the following reasons:

1. Max-stable processes are only suitable for extreme data, but the framework we propose is not restricted to extreme value distributions and leaves the choice of the marginal distribution open (see Eq (4.1)): in this respect, using max-stable processes would result in an important loss of generality.

2. Estimation of max-stable processes is challenging due to the difficulty of computing the whole likelihood. Pragmatic solutions based on the use of "composite likelihoods" have been proposed within a maximum-likelihood estimation context (see [*Padoan et al.*, 2010] for further discussion). In this thesis, we choose to use the Bayesian inference framework which enables a straightforward quantification of parameter and predictive

uncertainty. For max-stable processes, development of a Bayesian inference framework is very challenging and not yet available.

An analysis of the difference between the maximum stable process and a Gaussian copula on the extreme data will be discussed in the next chapter.

Copulas are used to build a joint distribution from a set of marginal distributions [*Sklar*, 1959]. For a *p*-dimensional multivariate random variable $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_p)$ with marginal cumulative distribution functions (cdf) $F_1, F_2, \ldots, F_p$, a copula is a function C :

$$\begin{aligned} C &: [0,1]^p \rightarrow [0,1] \\ &(F_1(y_1), F_2(y_2), ..., F_p(y_p)) \mapsto F(y_1, y_2, ..., y_p) \end{aligned} \tag{4.7}$$

where F is the joint cdf of the random variable $\boldsymbol{Y}$.

Sklar [1959] showed the existence of such a function and pointed out that if the marginal distributions are continuous, then the copula C is unique. Some further analyses of the usage of copula have been proposed by these authors, amongst many others: *Favre et al.* [2004], *Bardossy and Li* [2008], *Bardossy and Pegram* [2009] , *Renard and Lang* [2007] and *AghaKouchak et al.* [2010].

Due to the convenience in highly dimensional setups (which is typically the case with spatial datasets) [*Renard*, 2011], elliptical copula are favored in our framework. The elliptical copulas are linked to elliptical distributions [*Genest et al.*, 2007]. The two most commonly used are the Gaussian copula and the Student copula. In practice, these two copulas are very convenient since the modeling of spatial dependence is related to the properties of multivariate Gaussian and Student distributions, which are already well known. In particular, both copulas are parameterized by a symmetric matrix $\boldsymbol{\Sigma}$ representing pairwise dependence between sites.

The joint cdf using a Gaussian/Student copula is defined as follow:

$$F(y_1, y_2, ..., y_p) = G_{\boldsymbol{\Sigma}}(u_1, u_2, ..., u_p) \tag{4.8}$$

where

$u_i = \gamma^{-1}(F_i(y_i)), i = \{1, 2, ..., p\}$, with γ the cdf of a univariate standard Gaussian/Student distribution;

G_{Σ} is the cdf of a multivariate Gaussian /Student distribution with correlation matrix $\boldsymbol{\Sigma}$ (the degree of freedom of the Student distribution is made implicit in the notation).

The dependence matrix $\boldsymbol{\Sigma}$ represents pairwise dependence between sites: for any $s_i \neq s_j$, $\Sigma(s_i, s_j)$ quantifies the dependence between u_i and u_j . This dependence is modeled by a function of the distance between two sites:

$$\Sigma(s_i, s_j) = \Upsilon(\| s_i, s_j \|, \boldsymbol{\eta}) \tag{4.9}$$

where ||.|| is the distance function and Υ is the dependence model function whose variables are the distance and the dependence parameters $\boldsymbol{\eta}$.

The dependence calculated with Eq(4.9) is termed a "pseudo-correlation" between sites, because it corresponds to the correlation of the transformed variables $(u_i)_{i=1,p}$ in Eq(4.8), rather than the correlation of the raw observations $(y_i)_{i=1,p}$.

1.5 Parameter inference

Incorporating the elliptical copula described in Section 1.4, the joint pdf of the data at a fixed time t_k is computed as follows. The joint pdf is obtained by differentiating equation (4.7):

$$f\left(y(s_1,t_k), y(s_2,t_k), ..., y(s_p,t_k) \,|\, \left(\beta_i(s_j,t_k), i=\{1,2,...,m\},\, j=\{1,2,...,p\}\right), \boldsymbol{\eta}\right)$$

$$= \left(\frac{\prod_{j=1}^{p} f_j(y(s_j,t_k) \,|\, (\beta_i(s_j,t_k), i=\{1,2,...,m\}))}{\prod_{j=1}^{p} \phi(u_{j,k})} \right) \Phi_{\Sigma}(u_{1,k}, u_{2,k}, ..., u_{p,k} \,|\, \boldsymbol{\eta})$$

$$= \left(\frac{\prod_{j=1}^{p} f_j(y(s_j,t_k) \,|\, \boldsymbol{\theta}_{\boldsymbol{loc}}^{(j)}, \boldsymbol{\theta}_{\boldsymbol{reg}})}{\prod_{j=1}^{p} \phi(u_{j,k})} \right) \Phi_{\Sigma}(u_{1,k}, u_{2,k}, ..., u_{p,k} \,|\, \boldsymbol{\eta}) \tag{4.10}$$

where

$f_j(y(s_j,t_k) \,|\, (\beta_i(s_j,t_k), i=\{1,2,...,m\}))$ is the marginal pdf for site s_j at time t_k;

$u_{j,k} = \gamma^{-1}\left(F_j\left(y(s_j,t_k)\right)\right)$, with γ the cdf of a univariate standard Gaussian/Student distribution;

$\phi(u)$ is the standard Gaussian pdf or Student pdf with v degree of freedom (the latter being made implicit in the notation);

$\Phi(u_{1,k}, u_{2,k}, ..., u_{p,k})$ is the multivariate Gaussian pdf (with mean=0, correlation matrix $\boldsymbol{\Sigma}$) or multivariate Student pdf (with mean=0, correlation matrix $\boldsymbol{\Sigma}$ and degree of freedom v, the latter being made implicit in the notation).

The derivation of the full likelihood uses a time independence assumption: $\forall s, \forall t_1 \neq t_2$, $Y(s,t_1)$ is independent of $Y(s,t_2)$. Therefore, the full likelihood function $f(\boldsymbol{Y} \mid \boldsymbol{\theta}_{loc}, \boldsymbol{\theta}_{reg}, \boldsymbol{\eta})$ for all time steps and all sites is the product of equation (4.10) applied to all *n* time steps.

Similar to equation (2.4), the posterior pdf of the regression parameters is given as follows:

$$f(\boldsymbol{\theta}_{loc}, \boldsymbol{\theta}_{reg}, \boldsymbol{\eta} \mid \boldsymbol{Y}) \propto f(\boldsymbol{Y} \mid \boldsymbol{\theta}_{loc}, \boldsymbol{\theta}_{reg}, \boldsymbol{\eta}) f(\boldsymbol{\theta}_{loc}, \boldsymbol{\theta}_{reg}, \boldsymbol{\eta}) \tag{4.11}$$

where $f(\boldsymbol{\theta}_{loc}, \boldsymbol{\theta}_{reg}, \boldsymbol{\eta})$ is the prior pdf. The posterior pdf of $\boldsymbol{\theta}_{loc}, \boldsymbol{\theta}_{reg}, \boldsymbol{\eta}$ is estimated by the MCMC sampler [*Renard et al.*, 2006a] described in Chapter 2 (Section 2.1).

1.6 Missing values

In any kind of dataset, missing values are inevitable. Especially in a hydro-meteorological dataset, numerous missing values generally exist. For instance, extreme phenomena may cause measurement errors or make measuring equipment out of service. Moreover, using several stations together in a regional context generally leads to many missing values, because data availability varies from station to station. In data analysis, using appropriate methods to deal with the missing values is important to make the best of available data, while avoiding bias and improving the analysis quality.

In terms of implementation, the easiest way is to remove data to derive a missing-value-free dataset. If some missing values exist in some sites for a given time step, all data for this time step are rejected. The obvious drawback of this approach is that a lot of well-observed data are wasted: this approach is therefore not considered as an acceptable solution.

A possible solution to avoid wasting data is to fill in the missing ones. Four techniques are often used to fill in the missing data: (i) use constant values, e.g. the mean of local observation; (ii) use correlated stations; (iii) random generated values from some distribution; (iv) data augmentation: consider the missing values as unknown quantities, which are estimated together with the model regression parameters. The first three techniques may not be reliable when many missing values exist in the dataset, since the manually filled data may change the statistical property of the original data (for instance, systematically filling missing values with the mean results in a loss of variability). The fourth technique, data augmentation, is a good alternative to avoid modifying the statistical properties of the original data. However, if there are many missing values in the dataset, there will be many more unknown quantities to be estimated, resulting in a large increase of computation time.

In our framework, we decided to stick to the original dataset, rather than complete the data in some way, for the reasons of reliability and computation time. Thus we try to make the best use of the available dataset, without filling missing values or wasting non-missing ones. For a given time step, if missing values exist for some sites, we will just reject the missing sites, rather than all sites at this time step. This requires a careful implementation of the likelihood function computation. Our technique is to compute the likelihood function of all sites for each time step only through the non-missing sites. The missing sites at that time step will automatically be skipped. This implies that the number of sites used to compute the joint

distribution varies from time step to time step. Based on the time-independence hypothesis, the total likelihood is the product of the likelihood of each time step. For example, Figure 4.3 illustrates the availability of data for *10* observation sites. Gray area represents the missing values. The likelihood function for *time=0* to *50* will be computed using the last 5 sites only. For *time= 70* to 8*0*, it will be computed using site 1 to site 8, etc.

The difficulty in this approach is related to the inversion of the spatial dependence matrix of the copula for the non-missing data sites. Indeed, this inversion is the computational bottleneck of the likelihood computation: inverting the matrix at each time step would result in a huge loss of computational efficiency. To overcome this issue, we identify the blocs made of the same non-missing data sites. The matrix is inverted once for each bloc and then stocked in memory to avoid repeating the calculation for another time step consisting of the same available sites. With this approach, only 5 matrix inversions are required for the example in Figure 4.3 (5 sites on bloc 0-30, 2 sites on bloc 31-50, 7 sites on bloc 51-60, 8 sites on bloc 70-80, 10 sites on all other time steps). This is to be compared with the 200 matrix inversions that would be required with a naïve implementation.

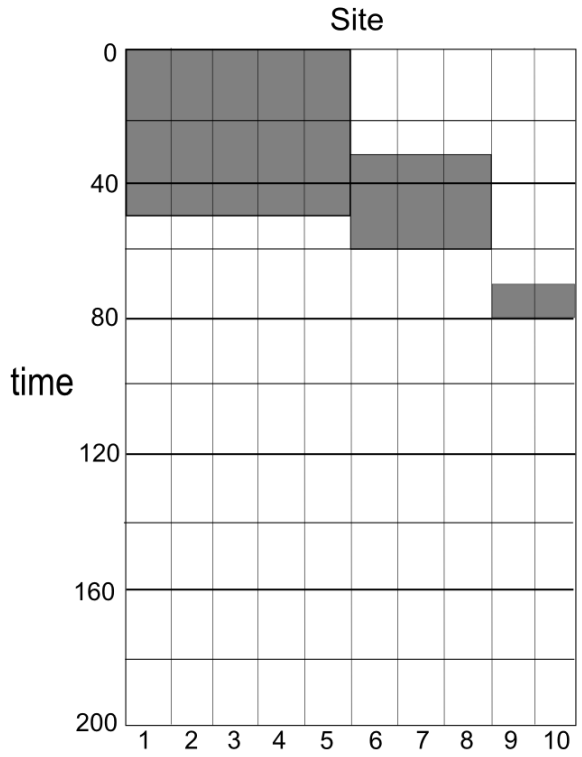


Figure 4.3-Illustration of data availability. Gray areas denote the missing data.

1.7 MCMC sampling, model diagnosis and model comparison

In the regional model, we use the same MCMC algorithms as introduced in Chapter 2 (Section 2.1.2). The model diagnostic tool and model selection tools are similar to that in the local model described in Chapter 2 (Section 3). PP plot is used to graphically evaluate goodness-of-fit. For each site, the theoretical cdf for each time step is calculated using both

local R-parameters and regional R-parameters. As in the local model, the plot of sorted cdf values $\left(F_i(y_i)\right)_{i=1,m}$ against$\left(\frac{i}{m+1}\right)_{i=1,m}$ should be close to diagonal for a good fit.

For model selection, the *DIC* value is computed for the whole region based on the joint pdf calculated in Eq(4.10). Hence, various hypotheses are investigated together at a regional scale, rather than for specific sites. The generality and convenience of the regional framework is highlighted once again.

2 Can the model detect spatio-temporal variations? Synthetic case studies

The objective of this section is to check the numerical implementation of the regional modeling framework and to quantitatively assess the extent to which temporal variations and spatial effects can be detected with regional models. Two synthetic studies are discussed in this section. In both of them, data are assumed to be temporally independent. Data are spatially independent in the first case, while spatial dependence exists in the second case. In the first case, we demonstrate that usage of regional parameters can improve the identification of weak signals. In the second case, regression parameters are estimated both considering and ignoring the spatial dependence, in a preliminary attempt to evaluate the importance of spatial dependence modeling.

2.1 Synthetic study 1

In this case study, we simulate a dataset containing 10 observation sites $\left(s_j\right)_{j=1,10}$ with 200 time steps $\left(t_k\right)_{k=1,200}$ for each site.

We also assume that data of sites 1 to 5 are missing for time *t*=1 to 50, data of sites 6 to 8 are missing for *t* =30 to 60 and data of sites 9 and 10 are missing for *t*=70 to 80 as illustrated in Figure 4.3. In total, about 10% of the data are assumed to be missing.

The altitude $\left(alt\left(s_j\right)\right)_{j=1,10}$ is used as a spatial covariate and is generated from a uniform Unif(0,100) distribution. The 10 independently generated altitude values are shown in Table 4.2.

Table 4.2-Simulated altitude values for 10 sites.

Site	1	2	3	4	5	6	7	8	9	10
Altitude	32.99	12.21	95.17	17.33	93.64	45.52	91.96	38.33	10.56	95.48

2.1.1 First simulation: low-frequency temporal variability

In this first simulation, the data are generated from a non-stationary $GEV\left(\mu(s,t),\sigma(s,t),\xi(s,t)\right)$ distribution with the following regression models:

$$\mu_1(s,t) = \left(20 + 0.2alt(s)\right)\cos(0.05t) \tag{4.12}$$

$$\sigma(s,t) = 10 \tag{4.13}$$

$$\xi(s,t) = -0.5 \tag{4.14}$$

Figure 4.4(a) shows the simulated data of site 9, and Figure 4.4(b) shows the data of all sites at time t = 61 to 65. The low-frequency temporal variability is clearly visible in the data (Figure 4.4) and should therefore be easily identifiable.

We fit the simulated data with the following $GEV\left(\tilde{\mu}(s,t),\tilde{\sigma}(s,t),\tilde{\xi}(s,t)\right)$ regression model:

Step 1: specification

$$\tilde{\mu}(s,t) = \lambda_1(s)\cos\left(\theta_{loc_1}^{(s)}t\right) \tag{4.15}$$

$$\tilde{\sigma}(s,t) = \theta_{loc_2}^{(s)} \tag{4.16}$$

$$\tilde{\xi}(s,t) = \lambda_2(s) \tag{4.17}$$

Step 2: spatialization

$$\lambda_1(s) = \theta_{reg_1} + alt(s) * \theta_{reg_2} \tag{4.18}$$

$$\lambda_2(s) = \theta_{reg_3} \tag{4.19}$$

In this regression model, $\boldsymbol{\theta}_{loc_1}^{(s)} = (\theta_{loc_1}^{(s_1)}, \theta_{loc_1}^{(s_2)}, ..., \theta_{loc_1}^{(s_{10})})$ and $\boldsymbol{\theta}_{loc_2}^{(s)} = (\theta_{loc_2}^{(s_1)}, \theta_{loc_2}^{(s_2)}, ..., \theta_{loc_2}^{(s_{10})})$ denote all the site-specific regression parameters, while θ_{reg_1}, θ_{reg_2} and θ_{reg_3} are regional regression parameters. Note that there is no model misspecification in this first case study: the model specified in equations (4.15)-(4.17) is consistent with the model used to generate the data (equation (4.12)-(4.14)). The only difference is that we assume that parameters controlling the frequency of the oscillation $\theta_{loc_1}^{(s)}$ are site-specific, while data have actually been generated by a unique regional parameter, which is equal to 0.05.

(a) Simulated data for site 9

(b) Simulated data for all sites at time steps *t=61 to 65*

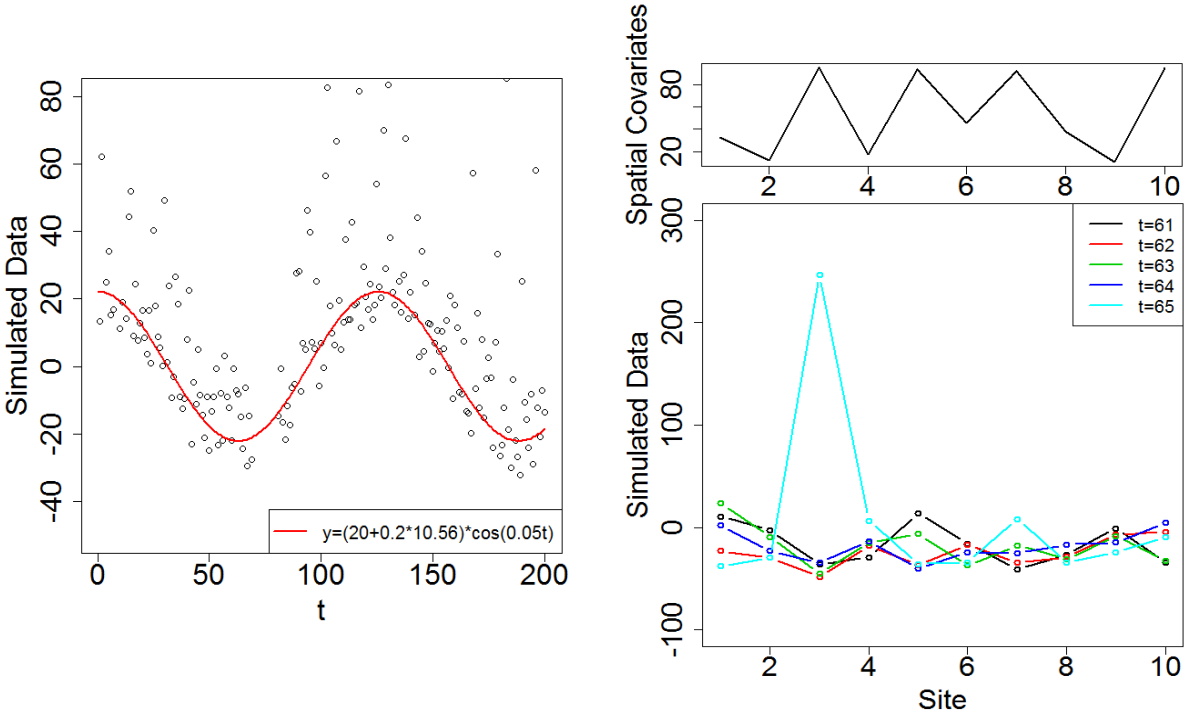


Figure 4.4-Illustration of the first simulated dataset

We estimate the posterior distribution of all regression parameters using MCMC sampling. Flat priors are used for all regression parameters. As the cosine function is symmetric around zero, the priors of $\boldsymbol{\theta}_{loc_1}^{(s)}$ are restricted to positive values. The starting value for $\boldsymbol{\theta}_{loc_1}^{(s)}, \boldsymbol{\theta}_{loc_2}^{(s)}$, $\theta_{reg_1}, \theta_{reg_2}$ and θ_{reg_3} are **0.01**, **20**, 10, 0.1 and 0, respectively. This correspond to very poor starting values given the true values of these parameters (**0.05**, **10**, 20, 0.2 and -0.5 respectively). 20000 MCMC iterations are performed.

Figure 4.5 shows the MCMC sequences for the three regional parameters θ_{reg_1} (intercept of the elevation regression), θ_{reg_2} (slope of the elevation regression) and θ_{reg_3} (regional shape parameter). Figure 4.6 and Figure 4.7 present the MCMC sequences for the local regression parameters $\boldsymbol{\theta}_{loc_1}^{(s)}$ (controlling the frequency of the cosine function) and $\boldsymbol{\theta}_{loc_2}^{(s)}$ (local scale parameters). In this first case study, MCMC simulations for all parameters immediately converge around the true values, despite the poorly-chosen starting points. Sometimes, the true value may be located at the edge of the posterior distribution (e.g. Site 8, Figure 4.7), which is due to the random variation of the parameters rather than the estimation problem. This positive result builds confidence into the correct implementation of the regional framework, and the efficiency of the MCMC sampler. However, as mentioned previously, the temporal variability was quite visible in the data (Figure 4.4(a)). The next simulation setup considers a more challenging situation, where the temporal signal is much less evident.

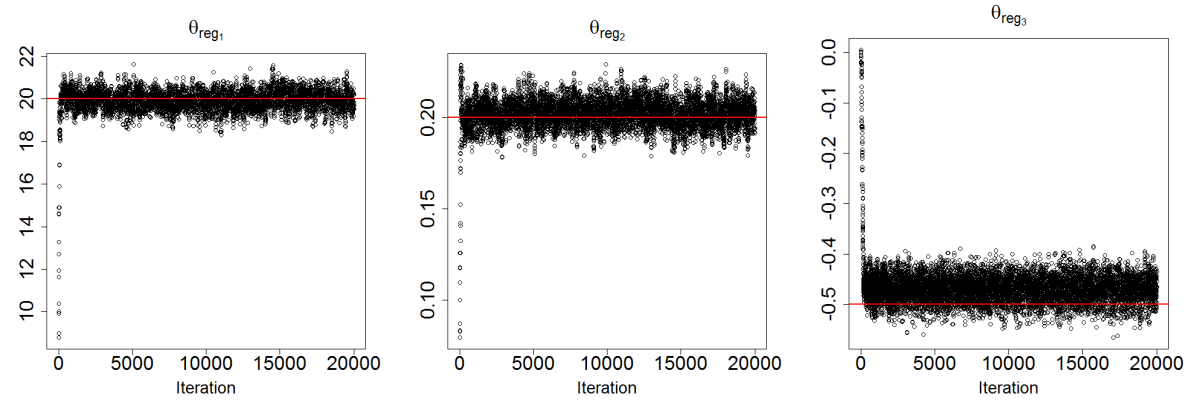


Figure 4.5-MCMC sequences for the three regional R-parameters. θ_{reg_1} is the intercept of the elevation regression, θ_{reg_2} is the slope of the elevation regression and θ_{reg_3} is the shape parameter.

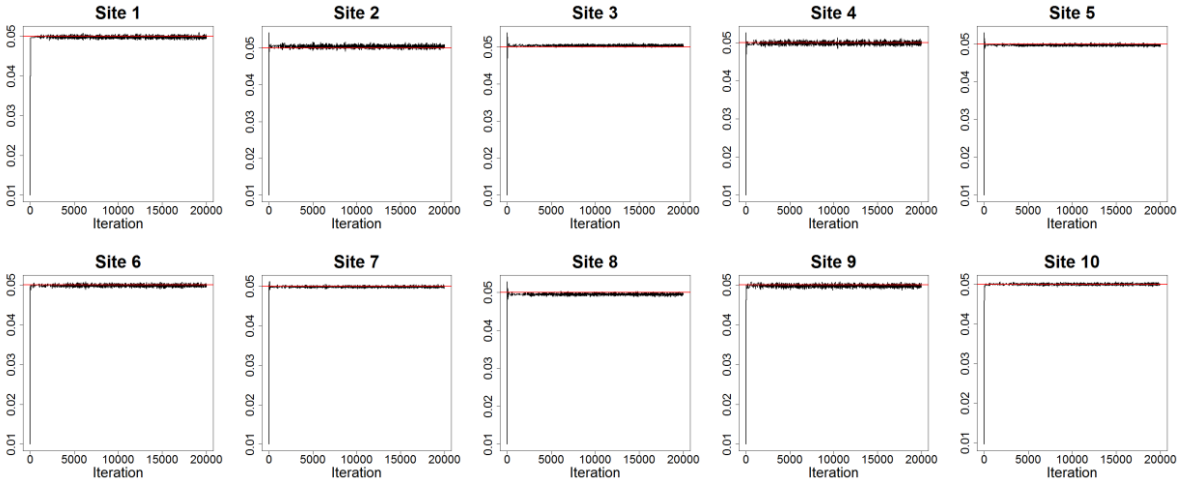


Figure 4.6-MCMC sequences for the local R-parameters $\theta_{loc_1}^{(s)}$ (controlling the frequency of the cosine function) at all 10 stations

2.1.2 Second simulation: high-frequency temporal variability

In this second test, we want to increase the frequency of the cosine function and verify whether the MCMC chains still converge. We replaced the parameter 0.05 of the cosine function of Equation (4.12) by 0.25, while the scale and shape *D*-parameters (Eq (4.13),(4.14)) remain the same. Thus the equation (4.12) becomes:

$$\mu_2(s,t) = \big(20 + 0.2alt(s)\big)\cos(0.25t) \tag{4.20}$$

Figure 4.8 shows the newly simulated data for site 9. The frequency is much increased compared with the first simulation (Figure 4.4). As a result, the underlying cosine temporal signal is much less visible in the data and should therefore be more difficult to identify.

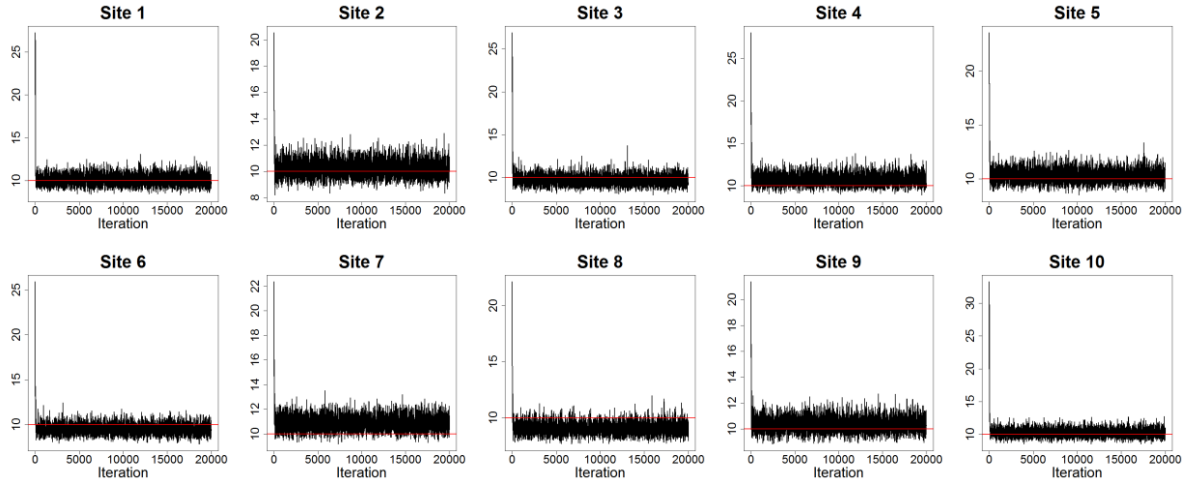


Figure 4.7-MCMC sequences for the local R-parameters $\theta_{loc_2}^{(s)}$ (scale parameters) at all 10 stations

Regression model 1

We start by using the same regression functions as in the previous simulation (equations (4.15)-(4.19)). The starting point of $\boldsymbol{\theta}_{loc_1}^{(s)}$ is set to **0.3**, and the other regression parameters use the same starting points as in the previous simulation. Figure 4.9 and Figure 4.10 present the MCMC sequences for the 20 local regression parameters $\boldsymbol{\theta}_{loc_1}^{(s)}$ and $\boldsymbol{\theta}_{loc_2}^{(s)}$. Most of them still converge to the true values. However, both parameters $\theta_{loc_1}^{(s_2)}$ and $\theta_{loc_2}^{(s_2)}$ of site s_2 do not converge to the true values. A significant gap exists between the true value (red line) and the MCMC sequence (black points). For the shape parameter θ_{reg_3} (Figure 4.11), the true value -0.5 is at the very limit of the distribution tail when estimated with a bad starting point. A clear convergence problem is raised in this simulation.

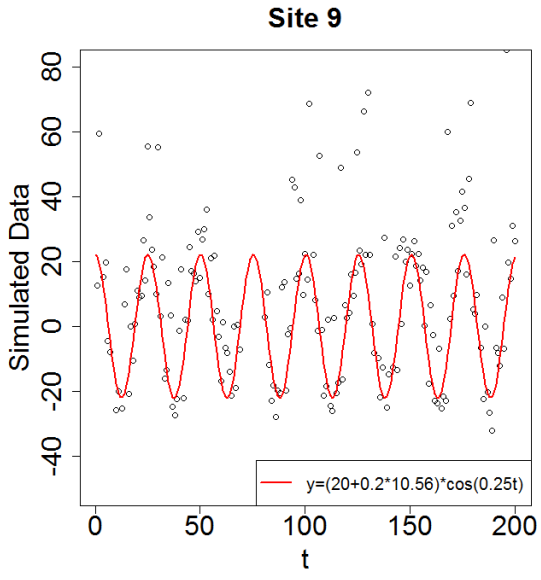


Figure 4.8-Illustration of the second simulated dataset.

Due to the large number of R-parameters in this model, the current MCMC sampler was not able to converge to the true values for all parameters. To verify it, we adjusted the starting points by setting all the regression parameters to the true values. Figure 4.9, Figure 4.10 (cyan lines) show that all the regression parameters converge immediately to their true values when starting with the good points.

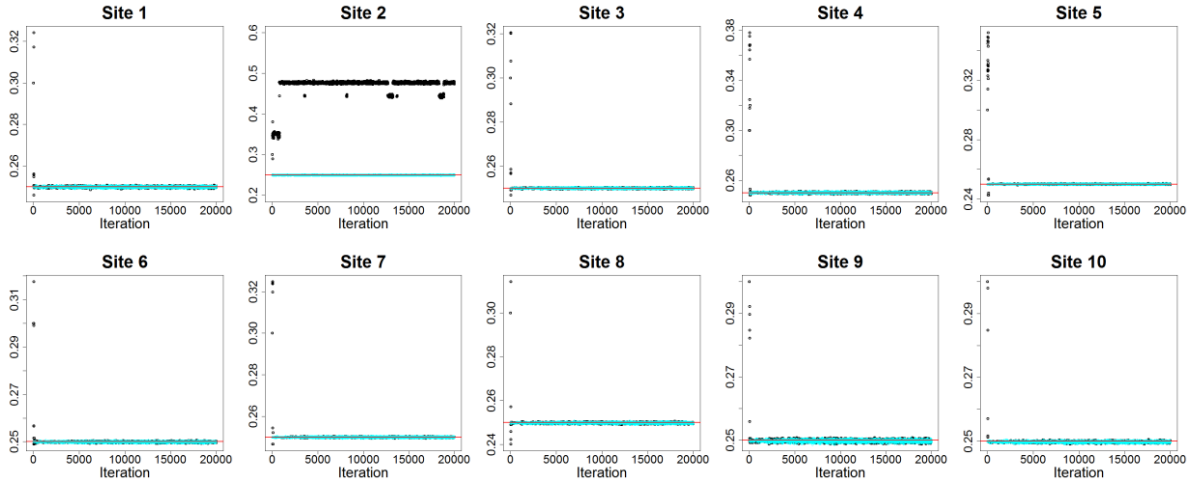


Figure 4.9-MCMC sequences for the local R-parameters $\theta_{loc_1}^{(s)}$ (controlling the frequency of the cosine function) at all 10 stations. The red line is the true value. The black points show the MCMC sequence with a bad starting point and the cyan line sow the MCMC sequence with the true value as starting point.

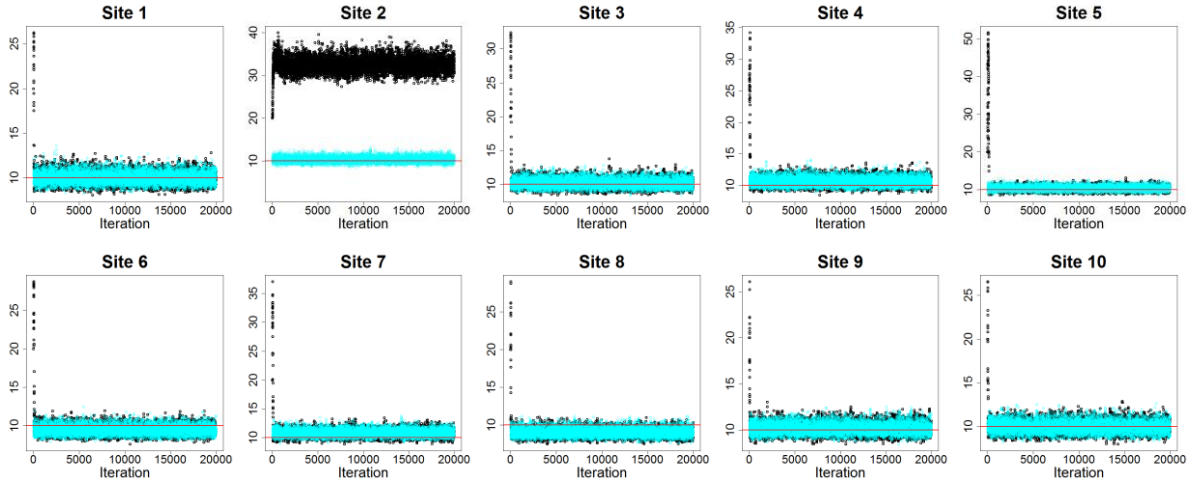


Figure 4.10-MCMC sequences for the local R-parameters $\theta_{loc_2}^{(s)}$ (scale parameters) at all 10 stations. The red line is the true value. The black points show the MCMC sequence with a bad starting point and the cyan line show the MCMC sequence with the true value as starting point.

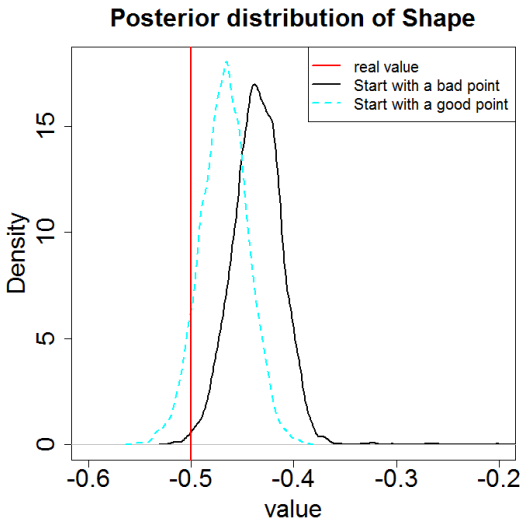


Figure 4.11-Posterior distribution of the shape parameter

To evaluate the influence of different starting points, we launch 4 parallel chains with different starting points. Figure 4.12 presents the local R-parameters θ_{reg_1} (left) and θ_{reg_2} (right) of site 2. Among the four chains, two converge to the true values (green and cyan are overlapped), and two (blue and black) do not. This example illustrates the benefit of using parallel chains in MCMC sampling: it increases the chance to detect a convergence issue. In this particular case, the convergence failure at site 2 would be easily detected by the Gelman-Rubin criterion.

There are several possible explanations for the false convergence or non-convergence of the MCMC sampler. The reason of such problems mainly comes from two aspects: (i) the MCMC sampler is not powerful enough; (ii) the posterior distribution has a very complex surface, including multimodality, many small local modes, etc. The latter possibility typically arises when the number of observation is not sufficient to support the identification of the regression model, or when the regression model is poorly specified.

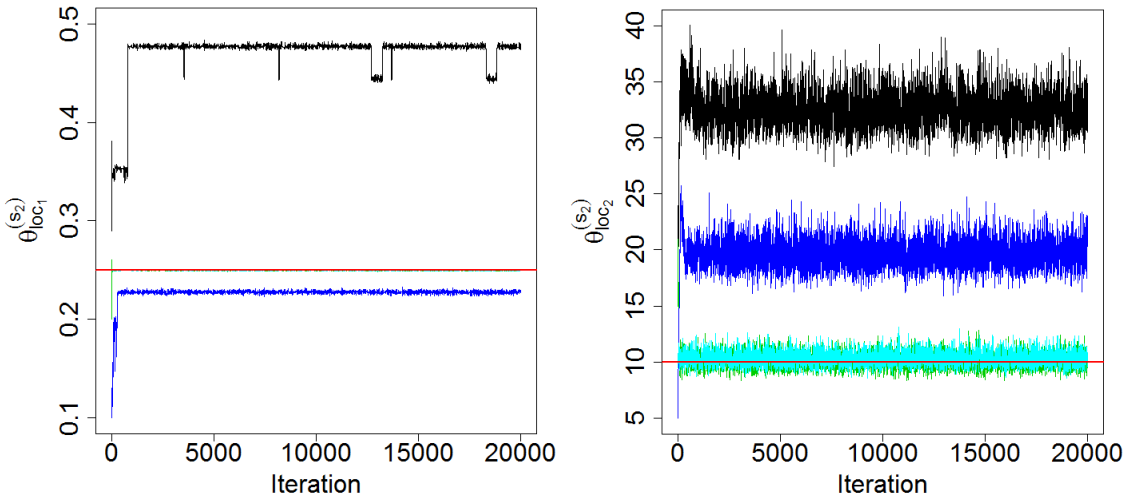


Figure 4.12-MCMC sequences for local R-parameters $\theta_{loc_1}^{s_2}$ (left) and $\theta_{loc_2}^{s_2}$ (right) of site 2

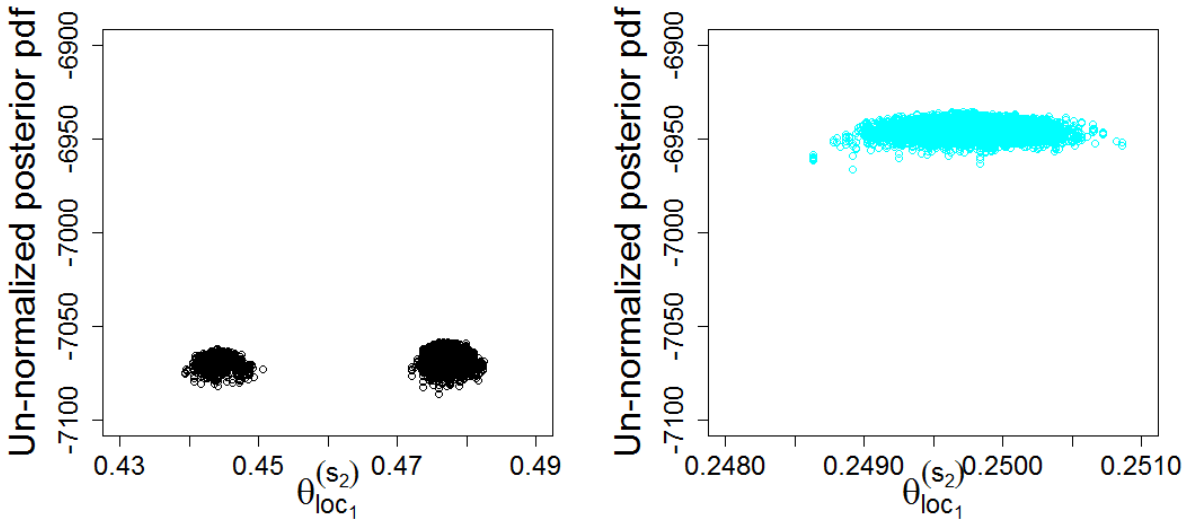


Figure 4.13- Scatterplots of the sampled $\theta_{loc_1}^{s_2}$ values of site 2 vs. the corresponding (un-normalized) posterior pdf. The left (right) panel corresponds to the black (cyan) chain in Figure 4.12.

In order to investigate the geometry of the posterior distribution, we report in Figure 4.13 the 'dotty plots' [*Wagener and Kollat*, 2007] for parameter $\theta_{loc_1}^{s_2}$, which are the scatterplots of the sampled $\theta_{loc_1}^{s_2}$ values vs. the corresponding (un-normalized) posterior pdf. It suggests that in this region of the parameter space, two modes having very similar heights exist. The right panel shows the convergent chain, and confirms that this region of the parameter space has a higher posterior value. Overall, Figure 4.13 confirms that the posterior surface is quite complex, including multiple secondary modes where MCMC chains with poor starting values can easily get trapped. This kind of posterior surface is typical of poorly-behaved inferences, and suggests that the model should probably be modified given the amount of available data.

In this case study, as the frequency of the cosine function increases, the temporal signal becomes weaker (Figure 4.4 and Figure 4.8) and is thus more difficult to identify at some sites. Therefore, it is not very adequate to use local R-parameters for presenting the frequency, while a regional effect exists (in this case, the true frequency is the same for all sites). In the following, we are going to use a regional R-parameter, instead of local parameters, for the frequency of the cosine function.

Regression model 2

The first regression model has two local regression parameters. The first one controls the frequency of the cosine function and the second one is the scale. In this regression model, we want to assess whether the parameter presenting the frequency can be better identified if the signal is assumed to be common for all sites.

Under this assumption, the ten local R-parameters for the location *D*-parameter are reduced to one parameter. Then the regression models become:

Step 1: specification

$$\tilde{\mu}(s,t) = \lambda_1(s)\cos\left(\lambda_3(s)t\right) \tag{4.21}$$

$$\sigma(s,t) = \theta_{loc_2}^{(s)} \tag{4.22}$$

$$\xi(s,t) = \lambda_2(s) \tag{4.23}$$

Step 2: spatialization

$$\lambda_1(s) = \theta_{reg_1} + alt(s) * \theta_{reg_2} \tag{4.24}$$

$$\lambda_2(s) = \theta_{reg_3} \tag{4.25}$$

$$\lambda_3(s) = \theta_{reg_4} \tag{4.26}$$

where $\theta_{reg_1}, \theta_{reg_2}, \theta_{reg_3}$ and θ_{reg_4} are regional regression parameters and only $\boldsymbol{\theta}_{loc_2}^{(s)} = (\theta_{loc_2}^{(s_1)}, \theta_{loc_2}^{(s_2)}, ..., \theta_{loc_2}^{(s_{10})})$ are site-specific regression parameters.

The starting points of $\boldsymbol{\theta}_{loc_2}^{(s)}$, $\theta_{reg_1}, \theta_{reg_2}$ and θ_{reg_3} remain the same, and the starting point of θ_{reg_4} is set to 0.3 as used for $\boldsymbol{\theta}_{loc_1}^{(s)}$. Figure 4.14 presents the MCMC sequences of the four regional R-parameters. Although we use the same starting points as in the previous case which did not converge, the R-parameters converge very quickly with this regression model. This illustrates the benefit of regionalizing in the case of weak temporal signals: such signals are difficult to identify locally, while they may become much clearer at the regional scale. Of course, this benefit comes at a cost: we made the additional assumption that the frequency of the temporal signal is identical for all sites. The relevance of this assumption should be thoroughly assessed once estimation has been performed.

2.2 Synthetic study 2

In this study, we simulate a dataset whose properties are similar to that of a real hydrologic dataset. We assume to have data from 20 sites in a region, with 50 data available for each site. The time series are denoted by $\left(Y(s_j,t_k)\right)_{j=1,20;k=1,50}$. 10% of the data are assumed to be missing (see Figure 4.15(a)).

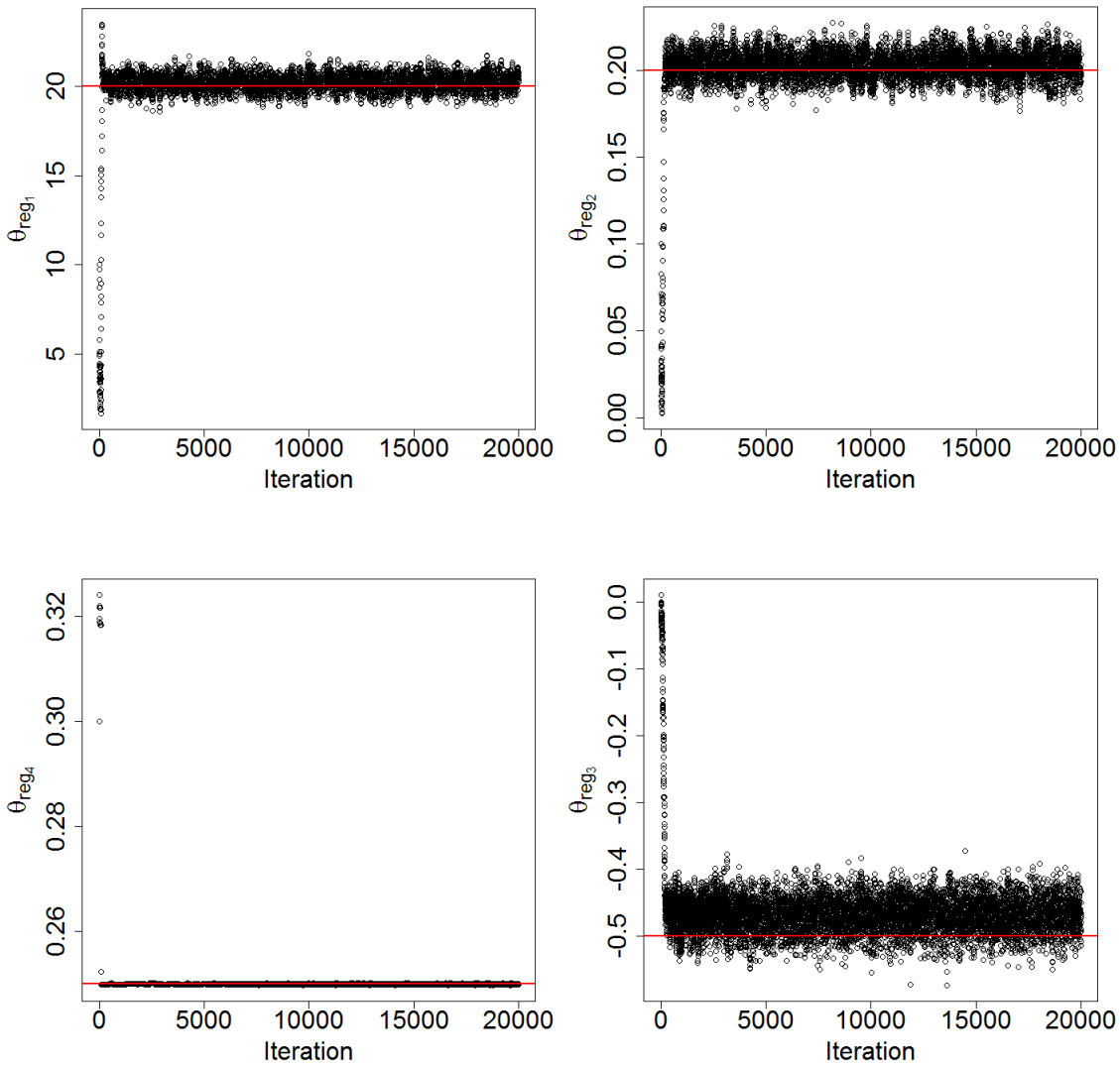


Figure 4.14-MCMC sequences of the four regional R-parameters. The red line is the true value.

2.2.1 Data generation

We start by generating the spatial covariate, which is the distance to the sea coast *SeaDist*(*s*). The distance to the sea coast is calculated with the coordinates $\left(x_{s_j}, y_{s_j}\right)_{j=1,20}$ of each site. The *x* and *y* coordinates of the first 16 sites are independently generated from a uniform Unif(0,30) distribution. The coordinates of the other 4 sites are generated from a uniform Unif(0,200) distribution, which locates those four sites far away from the 16 first sites. The idea behind this simulation setup is to generate a bunch of highly (spatially) dependent data, and a few nearly-independent additional series. The coastline is supposed to be located at the equation -0.5x+y+60=0. Figure 4.15(b) illustrates the location of the stations. The distance to the sea coast of each site is listed in Table 4.3.

Table 4.3-Distance to sea

Site	1	2	3	4	5	6	7	8	9	10
Distance	55.8	77.8	46.7	65.3	58.5	47.9	56.0	53.3	72.6	55.2
Site	11	12	13	14	15	16	17	18	19	20
Distance	52.7	70.4	69.2	48.9	76.3	57.9	54.0	87.0	80.8	146.2

(a) Data availability. Gray areas represent the missing values.

(b) Location of the stations

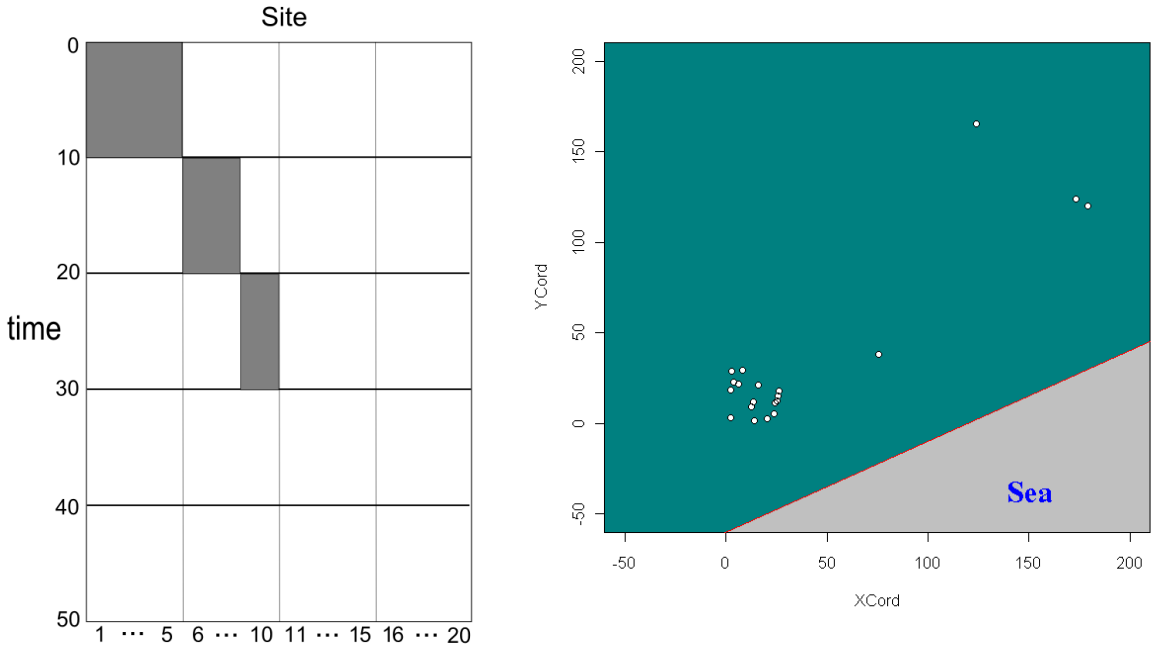


Figure 4.15-Data availability and location of the stations

A Gaussian copula is used to build the inter-site dependence. The spatial dependence between sites s_i and s_j is computed by the formula:

$$\Sigma(s_i, s_j) = \exp\left(-0.05 * \| s_i, s_j \|\right) \tag{4.27}$$

where $\| s_i, s_j \|$ is the distance between s_i and s_j. Figure 4.16 shows the dependence function and the pseudo-correlation between the 20 sites.

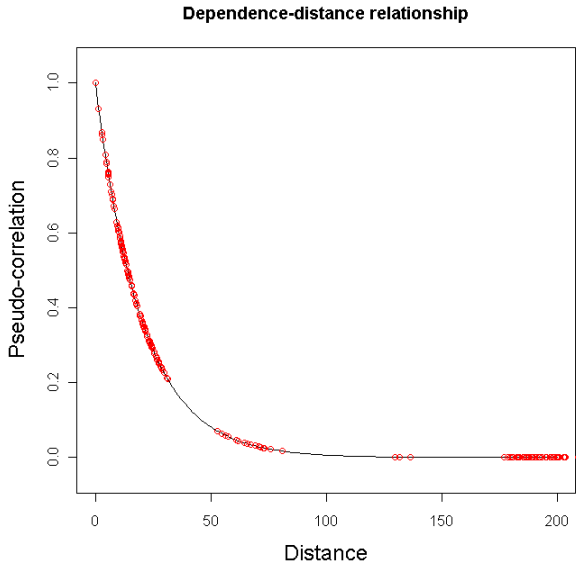


***Figure 4.16-Dependence-distance function. Red dots are the pseudo-correlations between all pairs of sites calculated by equation* (4.27).**

Data of these 20 sites are generated with a non-stationary $GEV\left(\mu(s,t),\sigma(s,t),\xi(s,t)\right)$ distribution. We assume that the temporal trend depends on the distance to the sea coast (Figure 4.17), which is shown through $\mu(s,t)$:

$$\mu(s,t)=50+\exp\left(-0.01*SeaDist(s)\right)*t \tag{4.28}$$

$$\sigma(s,t)=10 \tag{4.29}$$

$$\xi(s,t)=-0.1 \tag{4.30}$$

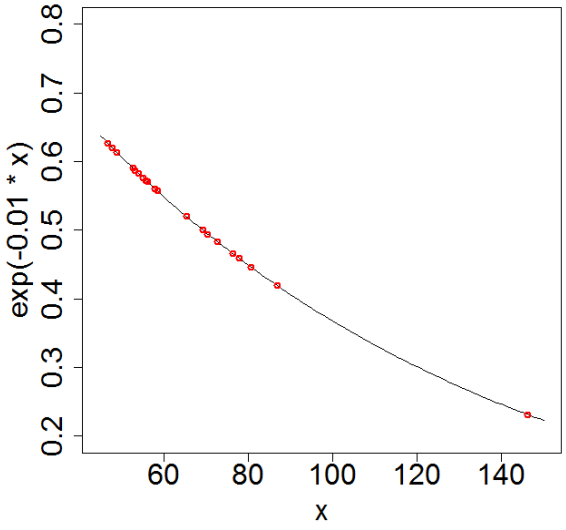


Figure 4.17-Temporal trend with respect to the distance to the sea. Red dots denote the twenty sites.

2.2.2 Regression and parameter estimation

Estimation is performed with two distinct models. Spatial dependence is considered in the first model using a Gaussian Copula. However spatial dependence is ignored in the second model. Both estimations use the following $GEV\left(\tilde{\mu}(s,t),\tilde{\sigma}(s,t),\tilde{\xi}(s,t)\right)$ regression model:

Step 1: specification

$$\tilde{\mu}(s,t)=\theta_{loc_1}^{(s)}+\lambda_1(s)t \tag{4.31}$$

$$\tilde{\sigma}(s,t)=\theta_{loc_2}^{(s)} \tag{4.32}$$

$$\tilde{\xi}(s,t)=\lambda_2(s) \tag{4.33}$$

Step 2: spatialization

$$\lambda_1(s)=\theta_{reg_1}*\exp\left(\theta_{reg_2}*SeaDist(s)\right) \tag{4.34}$$

$$\lambda_2(s)=\theta_{reg_3} \tag{4.35}$$

where $\boldsymbol{\theta}_{loc_1}^{(s)}=(\theta_{loc_1}^{(s_1)},\theta_{loc_1}^{(s_2)},...,\theta_{loc_1}^{(s_{20})})$ and $\boldsymbol{\theta}_{loc_2}^{(s)}=(\theta_{loc_2}^{(s_1)},\theta_{loc_2}^{(s_2)},...,\theta_{loc_2}^{(s_{20})})$ are local regression parameters. $\theta_{reg_1},\theta_{reg_2}$ and θ_{reg_3} are regional regression parameters.

For the first estimation, a Gaussian copula is used to describe the spatial dependence. The dependence-distance function is assumed to have the following form:

$$\Sigma(s_i,s_j)=\eta_1*\exp\left(-\eta_2*\|s_i,s_j\|\right) \tag{4.36}$$

where η_1 and η_2 are two parameters that need to be estimated. Flat priors are used for all parameters in both models.

Figure 4.18 present the MCMC sequence of the two parameters controlling spatial dependence. It shows that these parameters can be identified quite precisely from the data.

The posterior distribution of the local (Figure 4.20, Figure 4.21) and regional (Figure 4.19) regression parameters shows that whether or not spatial dependence is considered, the MCMC sequences are consistent with the true values. However, our expectation was that the estimation without spatial dependence would tend to under-estimate the uncertainties, because data from close sites are potentially over-weighted if spatial dependence is ignored. As an illustration, consider the following situation: there are three sites in a region, two are very close to each other and the third is far away from them; if the spatial dependence is considered, there are only two effective data (those who are close to each other should be considered as one), otherwise, three data will be used in the estimation. Since the data are potentially reused while spatial dependence is not considered, the uncertainty is thus under-estimated. However, in Figure 4.19, Figure 4.20 and Figure 4.21, the uncertainty of the two estimations does not show significant differences. If anything, a slight shift can be observed at some sites for the local R-parameters $\boldsymbol{\theta}_{loc_1}^{(s)}$ and $\boldsymbol{\theta}_{loc_2}^{(s)}$. The precise reasons for this behavior are

unclear at this stage, but the following causes could be considered: (i) the current GEV model has too many parameters, in which the uncertainties are shared by all these parameters. The difference of uncertainty between considering and ignoring spatial dependence is therefore not visible through one particular parameter; (ii) as the GEV distribution is sensitive to the shape parameter, a slight difference on the shape parameter can therefore mask such difference of uncertainty on the location and scale parameters.

In this section, we are not going to provide further investigation on the reason why the difference of uncertainty between the two estimations is not visible. It motivates us to make it clear that whether our expectation holds in our RFA framework. We are going to answer this question in the next Chapter.

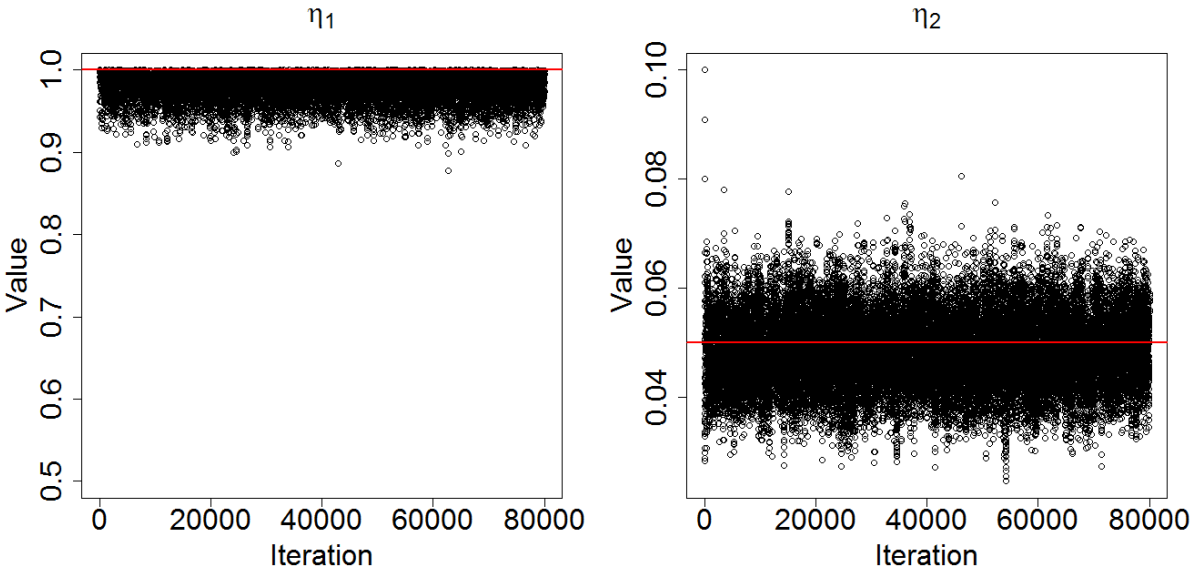


Figure 4.18-MCMC sequences of the dependence function parameters. The red line denotes the true value.

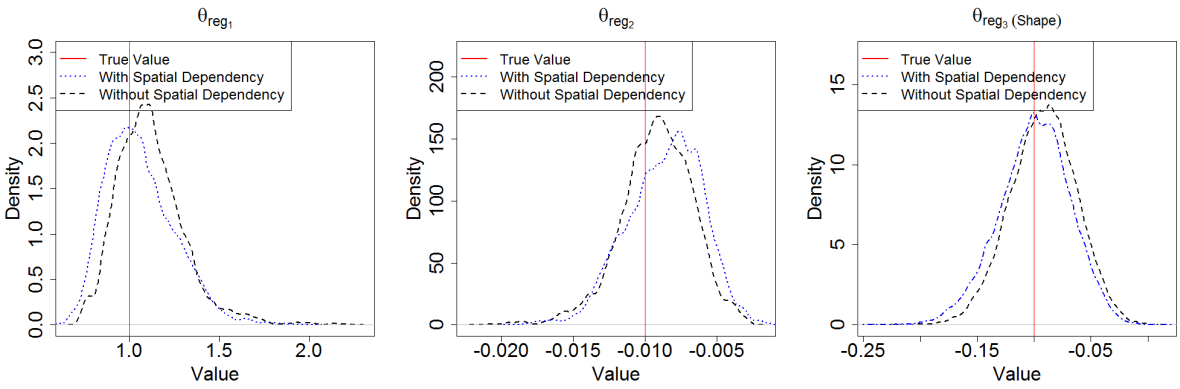


Figure 4.19-MCMC sequences of the regional R-parameters

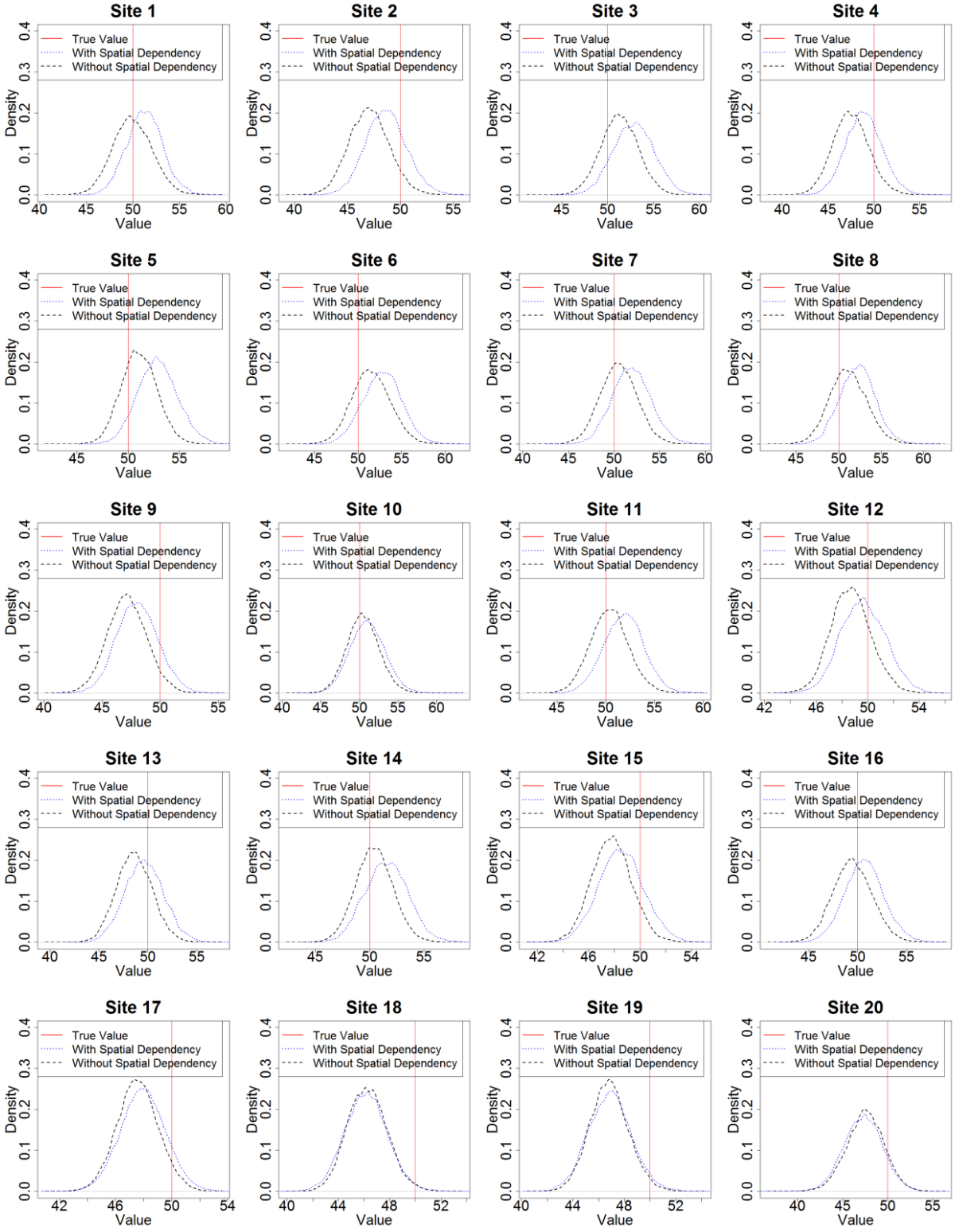


Figure 4.20-Posterior distribution of 20 local R-parameters $\theta_{loc_1}^{(s)}$

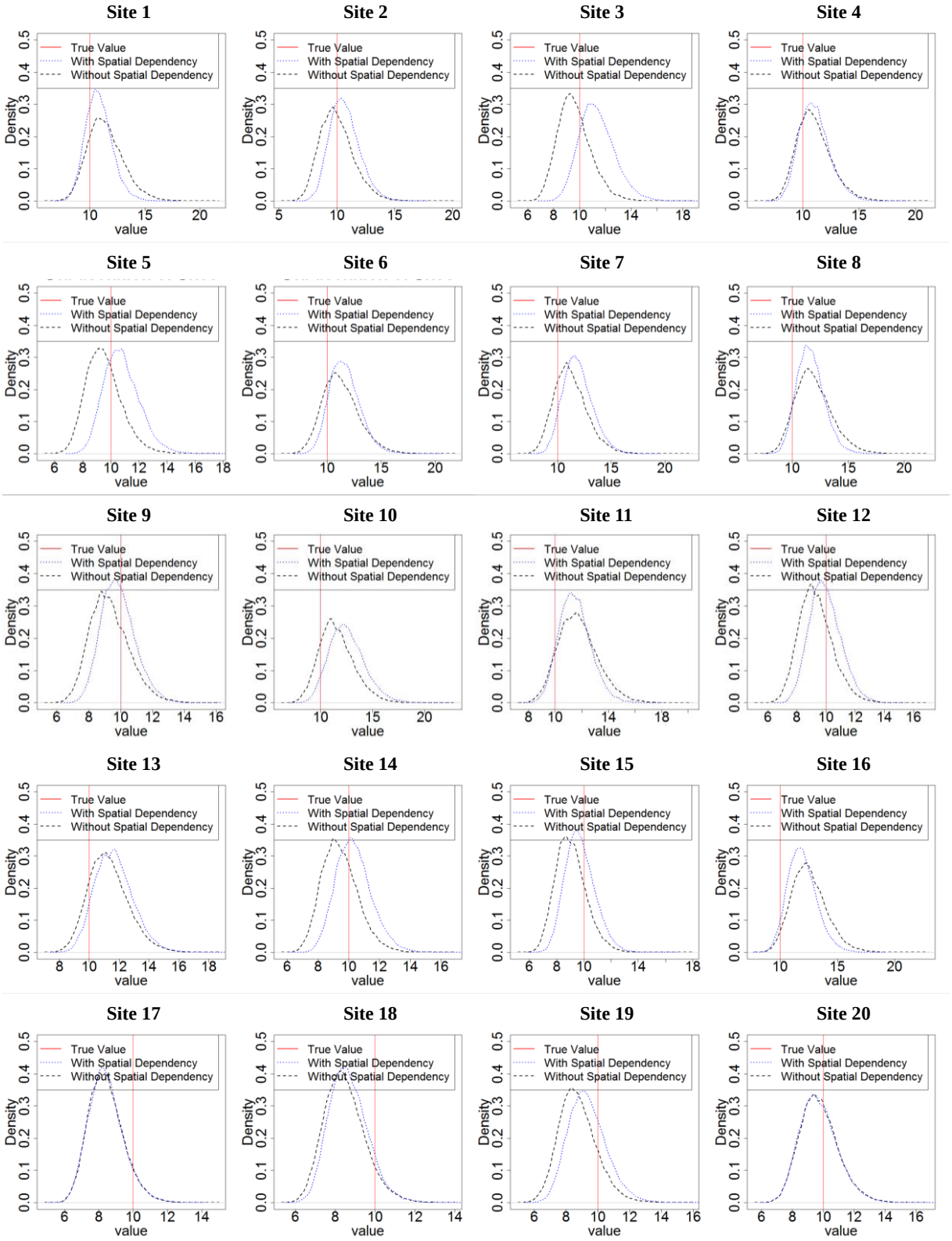


Figure 4.21-Posterior distribution of 20 local R-parameters $\theta_{loc_2}^{(s)}$

3 Conclusion on the spatio-temporal regional model

In this chapter, we describe a general spatio-temporal regional frequency analysis framework, geared towards detecting and quantifying the effect of climate variability and/or temporal trends on hydrological variables. Compared with the covariates in the local framework (only varying with time), the covariates in the regional framework are of two other types: they may vary in space or in both time and space. Thus three types of covariates are introduced in the regional model: temporal, spatial and spatio-temporal. These three types of covariates play different roles in the regression functions, especially the spatial covariates which are used to describe the spatial effects by means of a spatial regression function.

The flexibility of the framework provides a convenient way to compare various models and to select the most relevant relationship between covariates and hydrological data. For the regional analysis, spatial dependence is incorporated using elliptical copulas and a Bayesian approach is used for inference to enable uncertainties to be easily quantified. The use of a Bayesian regional framework provides the opportunity to assess the value of regional information in better identifying the impact of climate variability and temporal trends on hydrological variables (extremes in particular).

Two synthetic case studies were performed to illustrate the implementation of the regional model. The advantage of setting regional parameters is highlighted through the first case study by comparing the estimation results for the same regression model with respectively local parameters (regression model 1 of Section 2.1.2) and regional parameters (regression model 2 of Section 2.1.2). The advantage of using regional regression parameters to identify a weak signal was highlighted.

In the second synthetic study, missing values and spatial dependence are taken into account in the regional model. Models considering or ignoring spatial dependence are compared. The advantage of considering spatial dependence in such a regional model still needs to be discussed, and will be the topic of the next chapter.

Evidently, the regional model provides an estimation with less uncertainty than in the local model, since data from different sites are clustered to make the analysis. However, a bad choice for the regional parameter will also lead to unreliable predictions. Therefore it is also very important to determine which parameter in the regression model should remain local and which one can be spatialized.

CHAPTER 5 On the treatment of spatial dependence

This chapter focuses on the treatment of spatial dependence in the modeling framework described in Chapter 4. It aims at investigating two questions through two synthetic case studies.

In the first case study, we investigate whether ignoring the spatial dependence lead to a demonstrable under-estimation of uncertainties in the estimation of marginal parameters. To this aim, we compare the estimations obtained with a model using a Gaussian copula and a model assuming spatially independent data.

The second case study focuses on extremes and compares a copula-based modeling of spatial dependence and the use of maximum stable processes. More precisely, we investigate the following questions:

1. Is it acceptable to use a Gaussian copula to model data generated from a max-stable process if ones solely focus on the estimation of marginal parameters?

2. What are the differences in terms of estimating joint or conditional probabilities of exceedance of high values for several sites?

1 Does ignoring spatial dependence leads to an under-estimation of uncertainties?

As discussed in Chapter 4 (Section 2.2), ignoring spatial dependence is expected to lead to the underestimation of uncertainties. The objective of this section is to confirm that this indeed holds in our regional framework, although this was hardly observed in Chapter 4 (Section 2.2).

To investigate this question, we estimate only one marginal parameter to avoid interactions with other parameters. In the following study, we consider 20 sites containing 50 data without missing values.

1.1 First simulation with a Gaussian parent distribution

In the first simulation, data are generated with a non-stationary multi-dimensional Gaussian distribution $N\left(\boldsymbol{\mu}(t), \boldsymbol{\Sigma}(t)\right)$, where $\boldsymbol{\mu}(t) = \left(\mu_1(t), ..., \mu_{20}(t)\right)$ is the mean and $\boldsymbol{\Sigma}(t)$ is the covariance matrix. For simplification, we assume that these 20 sites are located on a line with equal distance between two sites. Thus the correlation between two sites s_i and s_j is assumed to be:

$$D_{ij}(\phi) = \phi^{|i-j|} \tag{5.1}$$

where $0 \le \phi < 1$ is a constant.

The dependence between sites will increase with ϕ. Thus for different value of ϕ, we are going to evaluate the difference between considering and ignoring spatial dependence.

We further assume that $\mu(t)$ is the same for all sites and $\boldsymbol{\Sigma}(t)$ is stationary. These parameters are specified as follows:

$$\forall i \in \{1, 2, ..., 20\},\ \mu_i(t) = 20 + 0.1t \tag{5.2}$$

$$\forall i, j \in \{1, 2, ..., 20\},\ \Sigma_{ij}(\phi) = 25 D_{ij}\left(\phi\right) = 25\phi^{|i-j|} \tag{5.3}$$

where *t* is the temporal covariate, which takes values from 1 to 50.

For each value ϕ, the generated data $\boldsymbol{Y}_\phi(s,t)$ are modeled with a Gaussian distribution with following regression functions:

$$Y_\phi(s,t) \sim N\left(20 + \theta_1 t, 25\right) \tag{5.4}$$

where θ_1 is the only *R*-parameter to be estimated. We focus on this trend parameter because it corresponds to a major motivation behind the derivation of the regional modeling framework: we wish to improve the identification of trends or climate variability effects.

Spatial dependence is ignored in the first estimation, thus θ_1 is the only parameter that need to be estimated. In the second estimation, spatial dependence is taken into account through a Gaussian copula with distance-dependence function is Eq(5.1). Two parameters θ_1 and ϕ_{est} therefore need to be estimated. Parameters are estimated using MCMC method under the Bayesian framework. In both estimations, flat priors are used.

Figure 5.1 presents boxplots of the posterior distribution samples of ϕ_{est} with respect to the true value ϕ. The boxes are located on the diagonal, hence, true values of ϕ can be accurately estimated. Furthermore, the uncertainty decreases as ϕ increases, which means that when the dependence between sites is weak, the uncertainty of the dependence parameter becomes larger.

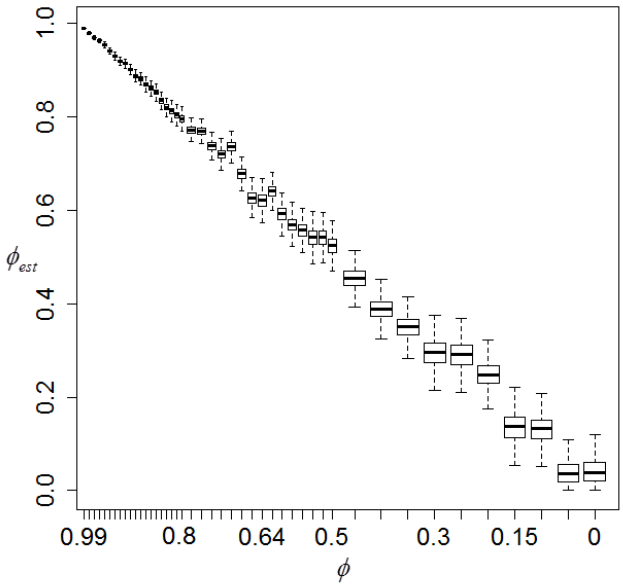


Figure 5.1-Boxplot of the estimated ϕ_{est} with respect to the true ϕ for a Gaussian parent distribution

Figure 5.2 presents the posterior variance of the trend parameter θ_1, estimated from the posterior distribution samples, as a function of the true values of ϕ. For the estimation ignoring spatial dependence, the posterior variance is almost constant, illustrating that the estimated uncertainty does not depend on the amount of spatial dependence existing in the data. Conversely, for the estimation considering spatial dependence, the posterior variance increases with the true dependence parameter ϕ: the estimated uncertainty is larger for highly dependent datasets, which is consistent with our expectation.

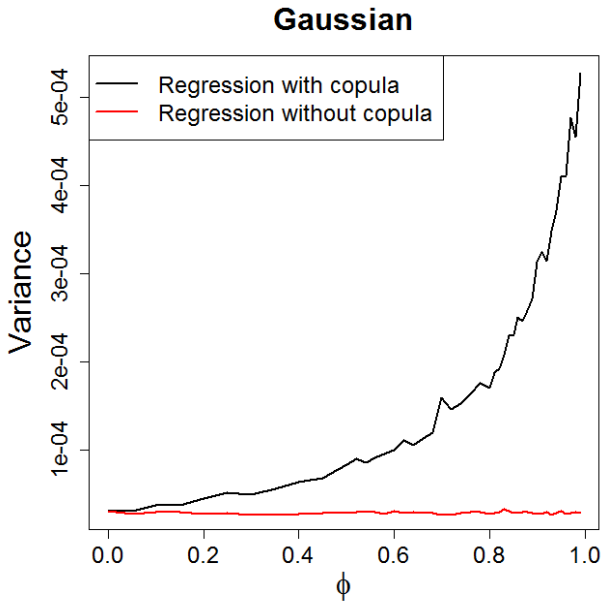


Figure 5.2-Posterior variance estimated from the MCMC samples of θ_1.

Figure 5.3 shows 90% credibility intervals of θ_1. For the estimation ignoring spatial dependence, the size of 90% credibility interval remains almost the same, while for the estimation with copula, this size increases with the true dependence. This is fully consistent with the behavior of the posterior variance previously described. However, when the value of ϕ increases, the 90% credibility interval of θ_1 often does not comprises the true value of θ_1 when spatial dependence is ignored. More precisely, out the 45 simulations carried out, only 25 (~56%) are able to recover the true value θ_1. By contrast, when spatial dependence is accounted for, the true value of θ_1 is included in the 90% credibility intervals for all 45 simulations but 3 (yielding a coverage of 93%, which is much more consistent with a 90% credibility interval). This confirms our expectation expressed at the beginning of this section that ignoring spatial dependence will lead to an underestimation of uncertainties.

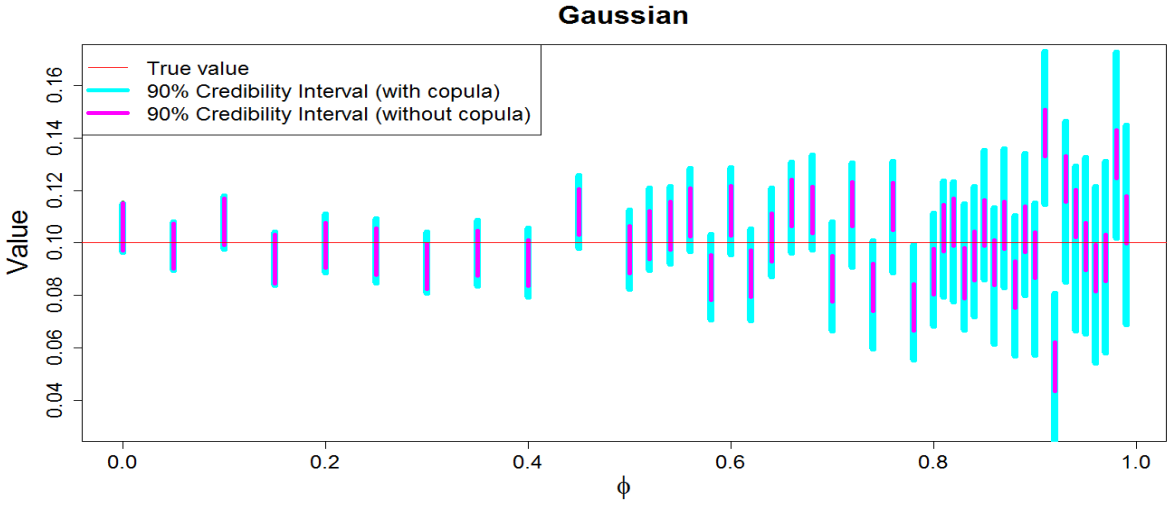


Figure 5.3-90% credibility interval of estimated θ_1 with respect to ϕ.

1.2 Second simulation with a GEV parent distribution

In this second simulation, we investigate whether the differences in terms of uncertainties discussed in Section 1.1 can still be observed for samples arising from a different distribution. Thus the Gaussian distribution is replaced by a GEV distribution in this section.

Data are generated with a non-stationary $GEV\left(\mu(s,t),\sigma(s,t),\xi(s,t)\right)$ distribution, coupled with a Gaussian copula with dependence-distance relationship introduced in Eq(5.1). The GEV parameters are specified as follows:

$$\forall s, \mu(s,t) = 25 + 0.2t \tag{5.5}$$

$$\forall s,t, \sigma(s,t) = 10 \tag{5.6}$$

$$\forall s,t, \xi(s,t) = -0.1 \tag{5.7}$$

where t is the temporal covariate, which takes its values from 1 to 50.

For each value of the dependence parameter ϕ, the generated data $\boldsymbol{Y}_{\phi}(s,t)$ are modeled with a GEV distribution with the following regression functions:

$$Y_{\phi}(s,t) \sim GEV\left(25 + \theta_2 t, 10, -0.1\right) \tag{5.8}$$

where the trend parameter θ_2 is the only R-parameter to be estimated.

Similarly as in the Section 1.1, two estimations are realized. One considers spatial dependence and the other ignores it. The distance-dependence function is still given by Eq(5.1). Thus, there are one (θ_2) and two (θ_2 and ϕ_{est}) parameters to be estimated in these two estimations.

Figure 5.4 presents the boxplots of the posterior distribution samples of ϕ_{est} with respect to the true value ϕ. This figure is similar to Figure 5.1: the true value ϕ is well estimated and the uncertainty grows when the value of ϕ decreases.

Figure 5.5 presents the posterior variance of the trend parameter θ_2 estimated from the posterior distribution samples. As in Figure 5.2, the posterior variance is almost constant when dependence is ignored, while it increases with ϕ when dependence is modeled with the copula. However, the increase in variance is less regular (less smooth) than in the previous case, especially for large ϕ.

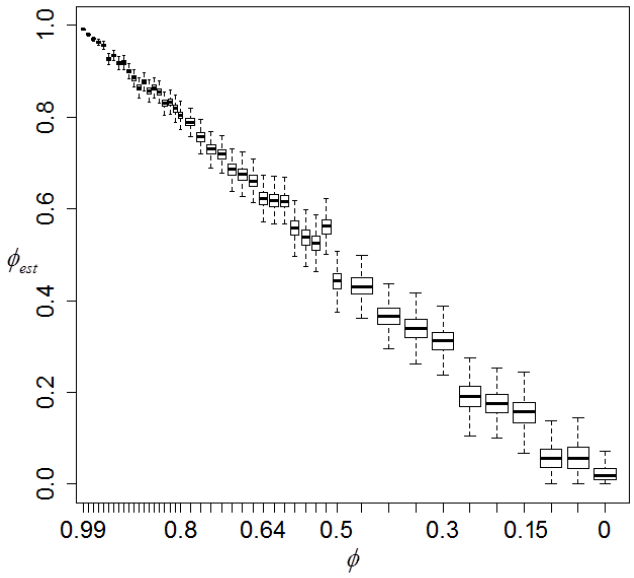


Figure 5.4-Boxplots of the estimated ϕ_{est} with respect to the true ϕ for a GEV parent distribution

Figure 5.6 shows the 90% credibility interval of θ_2. The main conclusions are similar to that of Figure 5.3: ignoring spatial dependence leads to an underestimation of uncertainties, with the true value of ϕ being often outside the 90% interval. More precisely, the coverage is around 60% when spatial dependence is ignored, against 93% when dependence is accounted for. Interestingly, the 90% credibility interval without the copula is sometimes not completely included in the interval obtained with the copula, and sometimes they are completely disjoint.

In general, the results in both simulations match our expectation. Since the GEV-simulated data have larger variability, skew and kurtosis than the Gaussian-simulated ones, the variation between the estimations is larger as well (they are less regular).

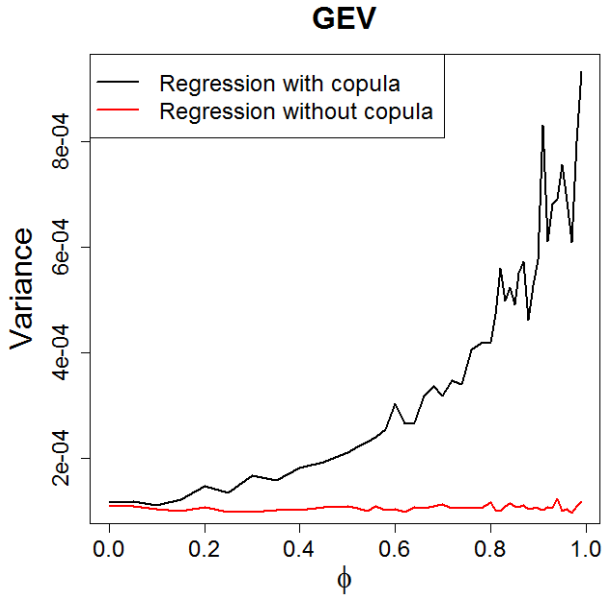


Figure 5.5-Posterior variance estimated from the MCMC samples of θ_2

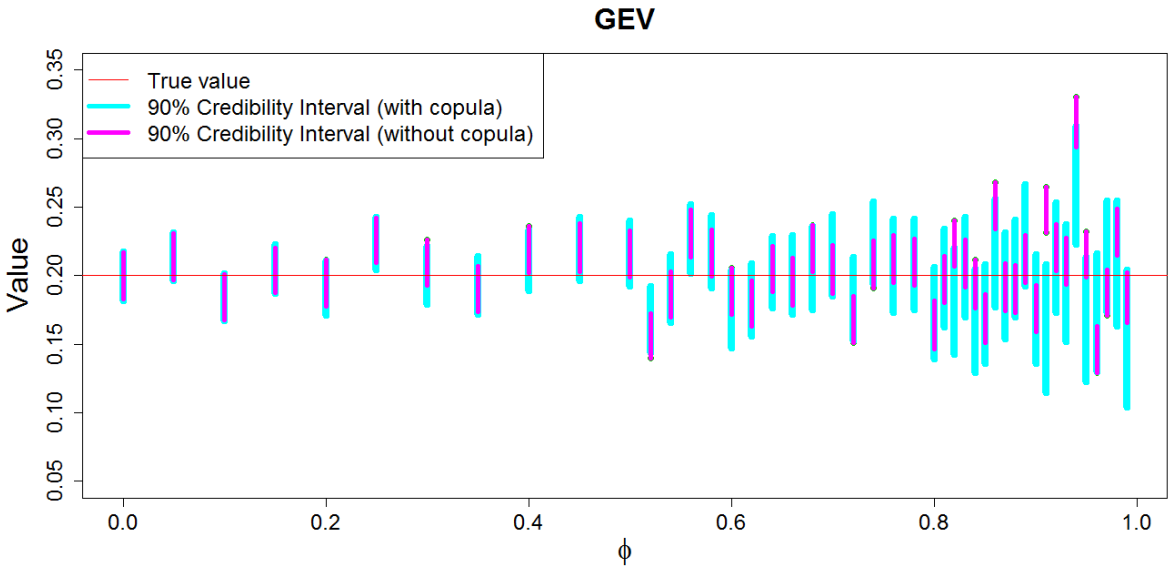


Figure 5.6-90% credibility interval of estimated θ_2 with respect to ϕ.

1.3 Conclusion

In this synthetic case study, we evaluated the difference between the estimations considering and ignoring spatial dependence. Consistently with our expectation, ignoring spatial dependence leads to an under-estimation of the parameter uncertainties.

Two datasets are generated from time-varying Gaussian and GEV distributions, in which data are spatially dependent. These datasets are respectively modeled with a single-parameter Gaussian and GEV model. Although both of the estimation results are qualitatively in agreement, the estimation with the GEV model is less regular than the one with the Gaussian model.

2 Spatial dependence for extremes: Copulas vs. maximum stable processes

Copulas and maximum stable processes are two different approaches for modeling the spatial dependence of block maxima. In the general framework developed in this thesis, we opted for the use of copulas, because they are not specific to extremes and are hence more adaptable to a wide range of situations. Moreover, they can be easily incorporated into a Bayesian inference framework, which is not yet the case for max-stable processes.

However, a multivariate distribution derived from an elliptical copula does not belong to the family of multivariate extreme value distributions. This raises questions on the suitability of elliptical copulas when the modeling framework is applied to extreme data.

The objective of this study is to illustrate the differences between max-stable processes and elliptical copulas in the description of data dependence. More precisely, we are interested in two distinct questions:

1. Is using a copula to model data generated from a max-stable process an acceptable approximation as long as one is only interested in estimating marginal parameters (e.g. location, scale and shape parameters of local GEV distributions)?

2. What are the differences in terms of estimating joint or conditional probabilities of exceedance of high values for several sites?

2.1 Basics of maximum stable processes

Maximum stable processes have been proposed for modeling the spatial dependency of extreme block maximum data. They are defined as follows:

Suppose $\left(Y_i(s)\right)_{i=1,n}, s \in \mathbb{R}^d$ are n independent identically distributed random fields in a multi-dimensional space $\mathbb{R}^d$ (typically, d=2 for spatial data). If there exist suitable sequences $a_n(s)$ and $b_n(s) > 0$, such that the limit

$$Z(s) = \lim_{n \to +\infty} \frac{\max_{i=1,\ldots,n} Y_i(s) - a_n(s)}{b_n(s)} \tag{5.9}$$

exists for all $s \in \mathbb{R}^d$, then $Z(s)$ is a maximum stable process.

Without loss generality on the characterization of maximum stable processes, the margins can be transformed to one particular extreme distribution. For the convenience, a unit Fréchet distribution is assumed:

$$\Pr\left(Y(s) \le z\right) = \exp\left(-\frac{1}{z}\right) \tag{5.10}$$

Thus, setting $a_n(s) = 0$ and $b_n(s) = n$, we obtain:

$$\Pr\left(Z(s) \le z\right) = \exp\left(-\frac{1}{z}\right) \tag{5.11}$$

There are two most widely used characterizations of maximum stable processes known as the Smith model and the Schlather model. Details of these two models will be discussed in the following sections.

2.1.1 The Smith Model

Let us start by considering a "rainfall-storms" construction. In an area of $\mathbb{R}^d$, the intensity of the i^{th} storm episode is ξ_i. We denote $\xi_i f(s_i, s)$ the amount of rainfall at position s from a storm with shape function f centered at s_i and intensity ξ_i. Then

$$Z(s) = \max_i \left\{ \xi_i f(s_i, s) \right\}, s \in \mathbb{R}^d \tag{5.12}$$

is the observation of maximum rainfall over independent storms in area of $\mathbb{R}^d$.

The Smith model [*Smith*, 1990] is often referred to as a "rainfall-storms" models illustrated above, which is defined as follows:

Let $\{(\xi_i, s_i), i \geq 1\}$ denote the points of a Poisson process on $\mathbb{R}_*^+ \times \mathbb{R}^d$ with intensity measure $\xi^{-2} d\xi \times \upsilon(ds)$ where $\upsilon(ds)$ is a positive measure on $\mathbb{R}^d$. Then Eq (5.12) is one characterization of a max-stable process, where $\left\{ f(t,s); s, t \in \mathbb{R}^d \right\}$ is a non-negative function such that:

$$\int_{\mathbb{R}^d} f(t,s)\upsilon(dt) = 1, \text{ for all } s \in \mathbb{R}^d \tag{5.13}$$

The Smith model defined in Eq(5.12) is a very a general form, whose margins are unit Fréchet:

$$\begin{aligned} &\Pr\left(Z(s) \leq z\right) = \Pr\left(\#\left\{(\xi, t) \in \mathbb{R}_*^+ \times \mathbb{R}^d : \xi f(t,s) > z\right\} = 0\right) \\ &= \exp\left[-\int_{\mathbb{R}^d} \left(\int_{z/f(t,s)}^{+\infty} \xi^{-2} d\xi \right) \upsilon(dt) \right] = \exp\left[-\int_{\mathbb{R}^d} z^{-1} f(t,s) \upsilon(dt) \right] = \exp\left(-\frac{1}{z} \right) \end{aligned} \tag{5.14}$$

where # means the number of element of the set.

The joint distribution at two sites is given by setting υ as the Lebesgue measure, and setting $f(t,s) = f_0(t-s)$ as a multivariate normal pdf centered at 0 and covariance matrix $\boldsymbol{\Sigma}$. Thus, the joint cdf at two sites is given by the following formula:

$$\Pr\left(Z(s_1) \leq z_1, Z(s_2) \leq z_2\right) = \exp\left[-\frac{1}{z_1} \Phi\left(\frac{a}{2} + \frac{1}{a} \log \frac{z_2}{z_1} \right) - \frac{1}{z_2} \Phi\left(\frac{a}{2} + \frac{1}{a} \log \frac{z_1}{z_2} \right) \right] \tag{5.15}$$

where Φ is the standard Normal cdf, and

$$a^2 = \Delta s^T \boldsymbol{\Sigma}^{-1} \Delta s \tag{5.16}$$

where $\Delta s = s_1 - s_2$ is vector of s_1, s_2 coordinates difference.

In particular, if $z_1 = z_2 = z$, then the equation (5.15) becomes:

$$\Pr\left(Z(s_1) \le z, Z(s_2) \le z\right) = \exp\left[-\frac{2}{z}\Phi\left(\frac{a}{2}\right)\right] \tag{5.17}$$

Note that independence corresponds to $\Phi\left(\frac{a}{2}\right)$=1, which happens when a tends to infinity. This is the case when the distance between two sites tends to infinity. Therefore, the Smith model generates virtually independent values for very distant sites, which seems realistic for rainfall applications.

The multivariate normal pdf f in $\mathbb{R}^d$ could also be replaced by some other pdf satisfying Eq(5.13), such as the Student pdf. In this study, we limit ourselves to using the Gaussian Smith model.

Figure 5.7 illustrates two simulations of the Smith model in $\mathbb{R}^2$ with different covariance matrices. The figure on the left assumes non-correlation between *x*-axis and *y*-axis, and the figure on the right assumes a positive correlation between *x*-axis and *y*-axis. The fields generated by the Smith model are very smooth, which is a consequence of each individual storm having a deterministic Gaussian shape. This may seem at odds with real rainfall fields shown by e.g. radar images, which are much more irregular. The Schlather model presented in the next section addresses this issue.

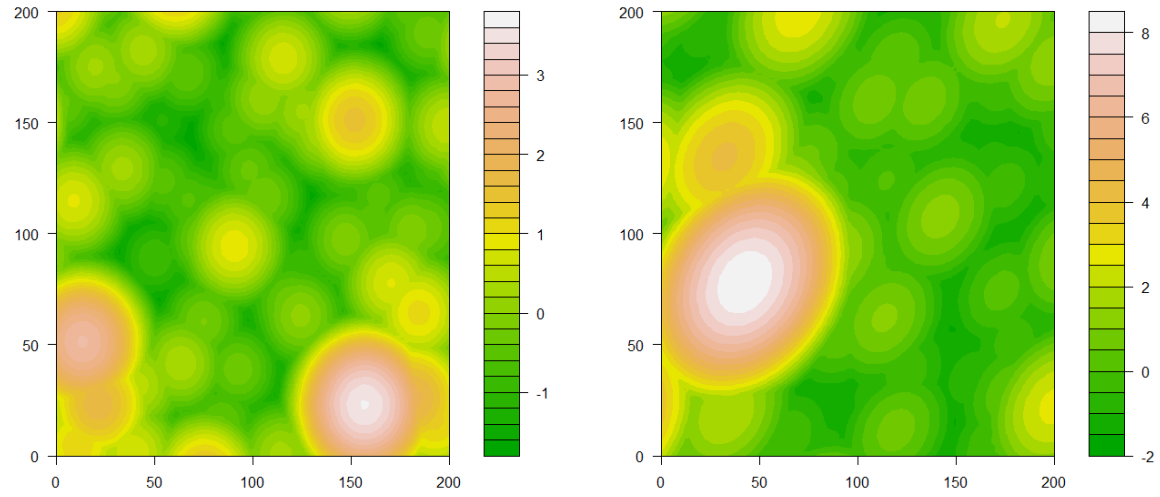


Figure 5.7-Two simulations of the Smith model with different covariance matrices. Left: Σ_{11}= Σ_{22}=200, Σ_{12}= Σ_{21}=0. Right: Σ_{11}= Σ_{22}=200, Σ_{12}= Σ_{21}=50. The data are transformed to unit Gumbel margins for viewing purposes.

2.1.2 The Schlather Model

Schlather [2002] described another characterization of max-stable processes, which is defined as follows:

Let $\{\xi_i, i \ge 1\}$ denote the points of a Poisson process on $\mathbb{R}_*^+$ with intensity measure $\xi^{-2}d\xi$, and $Y(.)$ be a stationary process on $\mathbb{R}^d$ such that $E\left[\max(0, Y(s))\right] = 1$. Then

$$Z(s) = \max_i \{\xi_i \max(0, Y_i(s))\} \tag{5.18}$$

is one characterization of a max-stable process, where $Y_i(.)$ are iid copies of $Y(.)$.

The main difference between the Smith model and the Schlather model is that the Smith model has a deterministic shape for the "storms" such as the multi-dimensional Normal pdf described in the previous section. However, the Schlather models do not impose a deterministic shape for each individual storm, but rather describe them as random processes $Y_i(.)$. Figure 5.8 presents an illustration of a field generated from the Schlather model, whose spatial structure seems more realistic for rainfall applications.

The margins of the Schlather model defined in Eq(5.18) are unit Fréchet. The proof is similar to Eq(5.14) by changing the shape function f to the random process $\max(0, Y_i(s))$.

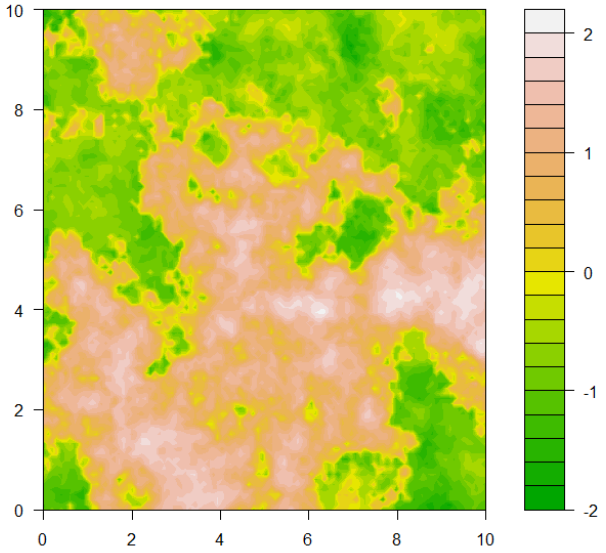


Figure 5.8-A simulation of Schlather model with Powered Exponential covariance, in which nugget, range and smoothness parameters are equal to 0, 3 and 1.

The joint distribution at two sites is given by setting $Y_i(.)$ as a stationary standard Gaussian process with correlation function $\rho(h)$ such that $E\left[\max(0, Y_i(s))\right] = 1$. Thus, the joint cdf at two sites is given by the following formula:

$$\Pr\left(Z(s_1) \le z_1, Z(s_2) \le z_2\right) = \exp\left[-\frac{1}{2}\left(\frac{1}{z_1} + \frac{1}{z_2}\right)\left(1 + \sqrt{1 - 2(\rho(h)+1)\frac{z_1 z_2}{(z_1+z_2)^2}}\right)\right] \tag{5.19}$$

where $h > 0$ is the Euclidean distance between sites s_1 and s_2. Table 5.1 lists several commonly used correlation models for $\rho(h)$.

In particular, if $z_1 = z_2 = z$, the equation (5.19) becomes:

$$\Pr\left(Z(s_1) \le z, Z(s_2) \le z\right) = \exp\left[-\frac{1}{z}\left(1+\sqrt{\frac{1-\rho(h)}{2}}\right)\right] \tag{5.20}$$

As $\rho(h)$ is positive, there is an upper limit for Eq(5.20), which is $\exp\left[-\frac{1}{z}\left(1+\sqrt{\frac{1}{2}}\right)\right]$. Since independence would correspond to $\exp\left[-\frac{2}{z}\right]$, this implies that the Schlather model will never generate independent values, even for infinitely distant sites. This may be problematic for rainfall applications, since it is difficult to imagine how annual maxima at very distant sites could be dependent, given the physics of rainfall generation but also physical frontiers like orography.

Table 5.1-Correlation functions for Schlather model

Model	Formula	Condition
Whittle-Matérn	$\rho(h) = \frac{2^{1-\upsilon}}{\Gamma(\upsilon)}\left(\frac{h}{c_2}\right)^{\upsilon} K_{\upsilon}\left(\frac{h}{c_2}\right)$	$c_2 > 0, \upsilon > 0$
Cauchy	$\rho(h) = \left[1+\left(\frac{h}{c_2}\right)^2\right]^{-\upsilon}$	$c_2 > 0, \upsilon > 0$
Powered Exponential	$\rho(h) = \exp\left[-\left(\frac{h}{c_2}\right)^{\upsilon}\right]$	$c_2 > 0, 0 < \upsilon \le 2$
Bessel	$\rho(h) = \left(\frac{2c_2}{h}\right)^{\upsilon} \Gamma(\upsilon+1) J_{\upsilon}\left(\frac{h}{c_2}\right)$	$c_2 > 0, \upsilon \ge \frac{d-2}{2}$

where c_2 and υ are the range and smooth parameters, Γ is the gamma function, and J_υ and K_υ are the Bessel and modified Bessel functions.

2.2 Gaussian copula inference with various spatial data

In this section, we aim to investigate whether the Gaussian copula model provides correct parameter inferences for the data generated from maximum stable process.

2.2.1 Moderate dependence data

We suppose that there are 20 observation sites in a region. Each site contains 50 years of data. The coordinates (in kilometer) of these 20 sites are generated from a uniform distribution Unif[0,200].

At each replication, data are generated with a GEV distribution, in which a temporal trend exists on the location parameter. The GEV parameters are specified as follows:

$$\forall s, \mu(s,t) = 30 + 0.01t \tag{5.21}$$

$$\forall s,t, \sigma(s,t) = 10 \tag{5.22}$$

$$\forall s,t, \xi(s,t) = -0.1 \tag{5.23}$$

In the moderate dependence case, spatial dependence is modeled with a Gaussian copula and a Smith model, respectively. The Schlather model is not considered in the moderate dependence case, because a limit exists for the spatial dependence of Schlather models, thus low dependence data cannot be simulated. Figure 5.9 and Figure 5.10 present one simulated storm with the Gaussian copula model and the Smith model. The corresponding dependence function for the Copula is $\Sigma_{ij} = \exp\left(-0.03 * \| s_i, s_j \|\right)$, and the covariance matrix for Smith model is $\Sigma_{11} = \Sigma_{22} = 100$ and $\Sigma_{12} = \Sigma_{21} = 0$. We use the R package "SpatialExtremes" developed by Dr. M. Ribatet to generate data. Note that the parameters used for the Gaussian copula and the Smith model correspond to the typical dependence observed in real rainfall data (see e.g. *Renard et al.* [2013]).

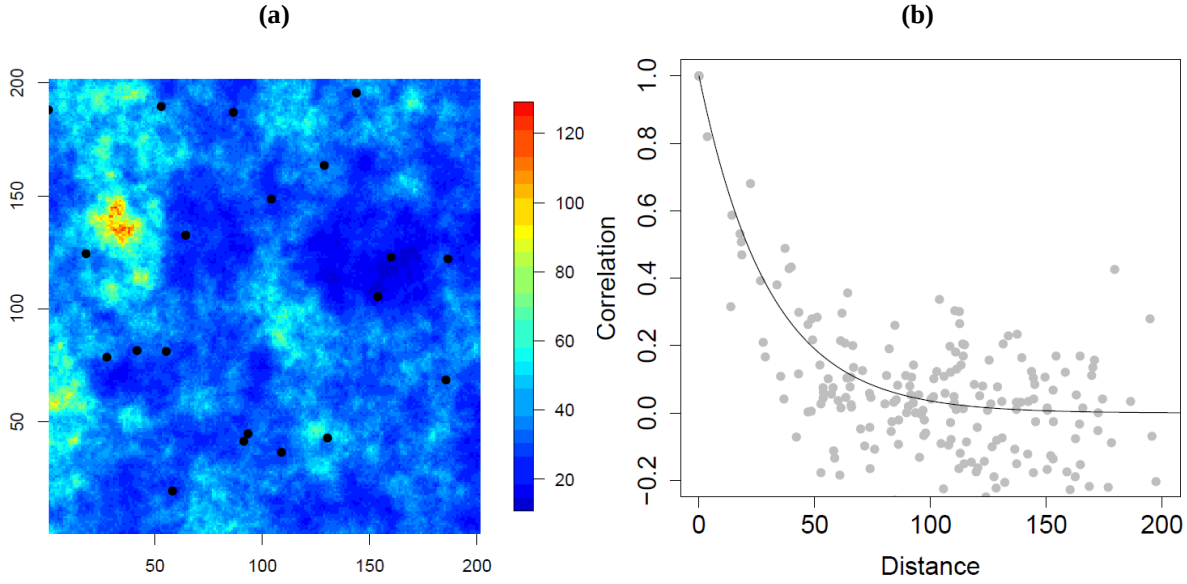


Figure 5.9-(a) Simulated field with a Gaussian Copula for the moderate dependence case. Black dots are the observation sites. (b) Corresponding dependence-distance function. Gray dots are the correlation between pairs of sites, estimated from 50 replicated fields. The black curve is an exponential correlogram, fitted to the pairwise correlations using least squares.

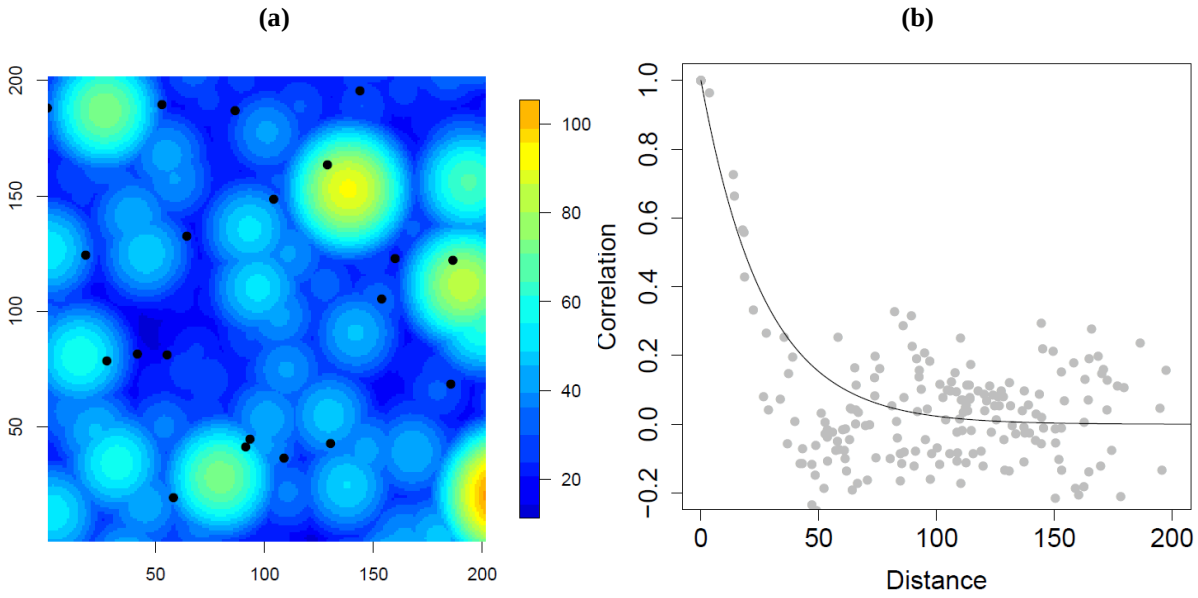


Figure 5.10-(a) Simulated field with a Smith model for the moderate dependence case. Black dots are the observation sites. (b) Corresponding dependence-distance function. Gray dots are the correlation between pairs of sites, estimated from 50 replicated fields. The black curve is an exponential correlogram, fitted to the pairwise correlations using least squares.

The simulated data are modeled with the following GEV distribution, in which all the parameters are regional:

$$Y(s,t) \sim GEV\left(\mu_0 + \mu_1 t, \sigma, \xi\right) \tag{5.24}$$

and the dependence-distance function of the Gaussian copula is:

$$\forall i \neq j, \Sigma(i,j) = \exp(-\theta_1 * \| s_i, s_j \|) \tag{5.25}$$

where $\| s_i, s_j \|$ is the distance between site s_i and s_j.

The GEV parameters are estimated both considering and ignoring copula. Figure 5.11 shows the estimation results based on 100 replications of (20 sites * 50 years) datasets. For readability, the 100 replications have been reordered by increasing values of parameters estimated with the Gaussian copula. Note however that, for one given replication in this figure, the estimations with and without copula are performed on exactly the same simulated dataset.

For the data generated with the Gaussian copula model, the 90% credibility intervals cover the true value with a rate very close to 90% for all parameters. However, ignoring spatial dependence leads to the under-estimation of uncertainty: a lower coverage rate of the true value is detected, which is consistent with the result of Section 1.2. More precisely, this under-estimation of uncertainty is particularly noticeable for the location and the trend parameters, with coverage rates of 65% and 69%, respectively.

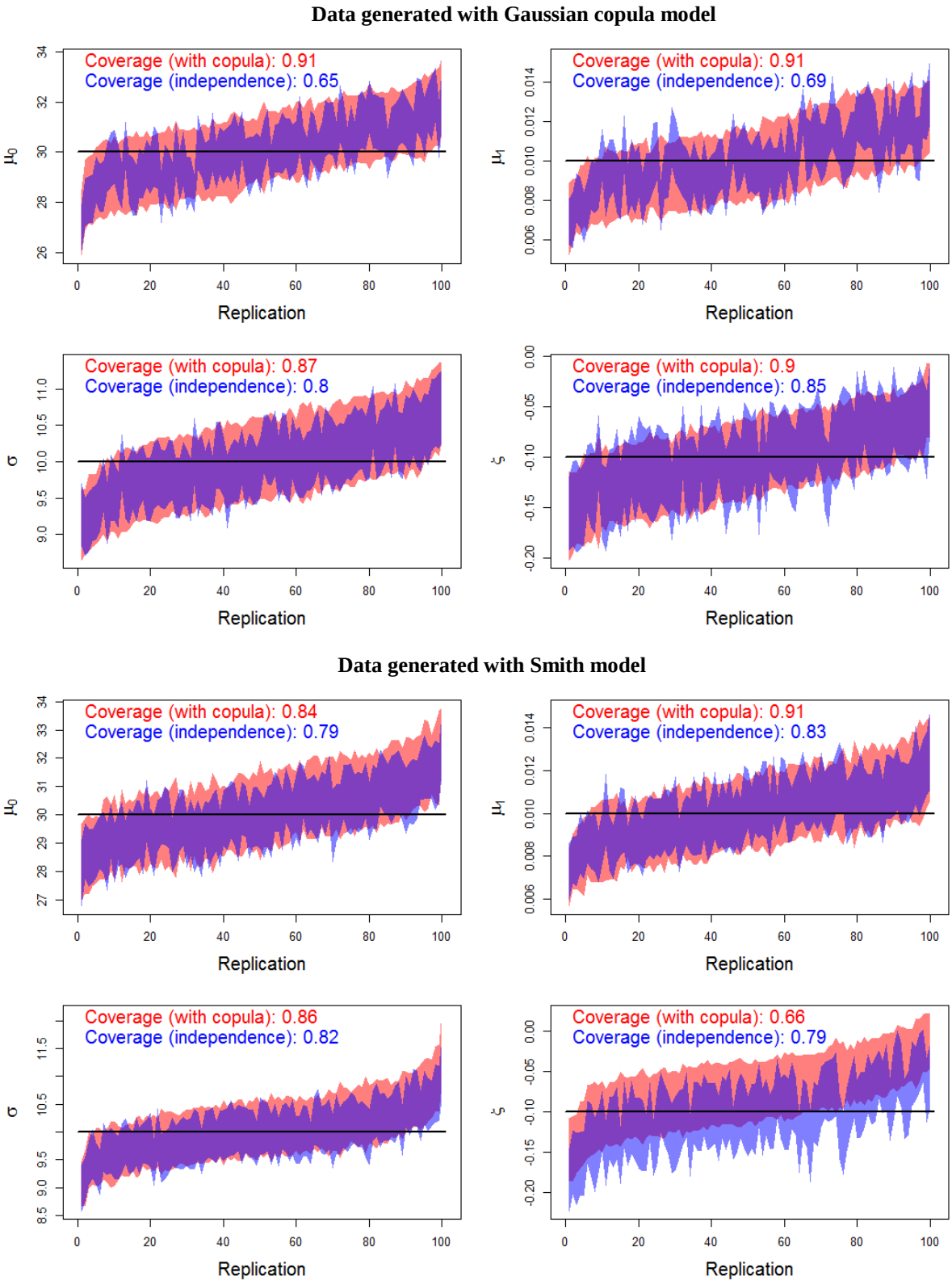


Figure 5.11-90% credibility intervals of the regression parameters for the moderate dependence case. For each parameter, replications are sorted according to the posterior mean of the copula-based estimation.

For the data generated with the Smith model, the estimation results with the Gaussian copula are good for the location, trend and scale parameters. Ignoring spatial dependence still leads to an under-estimation of uncertainty for these three parameters. However, for the shape parameter ξ, the coverage rate is too low for both estimations. Moreover, the true value tends to be slightly over-estimated when considering spatial dependence. This result suggests that the use of a Gaussian copula, while improving the coverage rate for the location, trend and scale parameters, induces a slight bias for the shape parameter.

2.2.1 High dependence data

In the high dependence case, data are generated with the same GEV distribution with parameters shown by Equations (5.21) (5.22) and (5.23). The spatial dependence is modeled with a Gaussian copula, the Smith model and the Schlather model (with Whittle-Matérn correlation). The dependence function for the Copula is $\Sigma_{ij} = \exp\left(-0.06 * \| s_i, s_j \|\right)$. The covariance matrix for the Smith model is $\Sigma_{11} = \Sigma_{22} = 300$ and $\Sigma_{12} = \Sigma_{21} = 0$. For the Schlather model, nugget, sill and range of the Whittle-Matérn correlation function are 0, 1 and 30, respectively. Figure 5.12, Figure 5.13 and Figure 5.14 present one simulated field simulated by these three models. The simulated data are fitted with the same model as in the moderate dependence case (Eq(5.24)(5.25)).

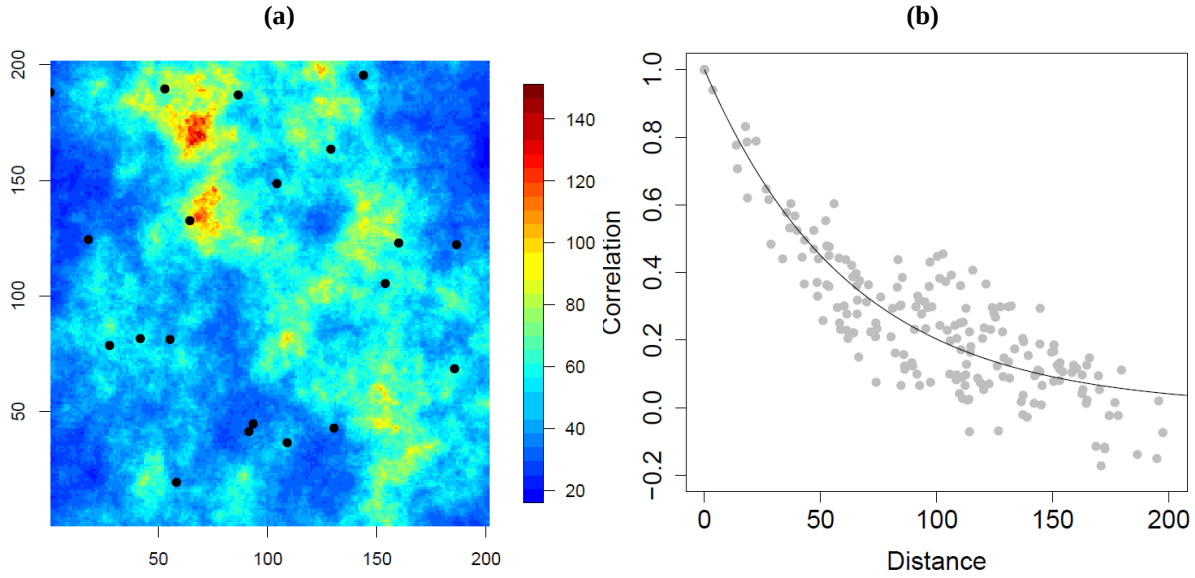


Figure 5.12-(a) Simulated field with a Gaussian Copula for the high dependence case. Black dots are the observation sites. (b) Corresponding dependence-distance function. Gray dots are the correlation between pairs of sites, estimated from 50 replicated fields. The black curve is an exponential correlogram, fitted to the pairwise correlations using least squares.

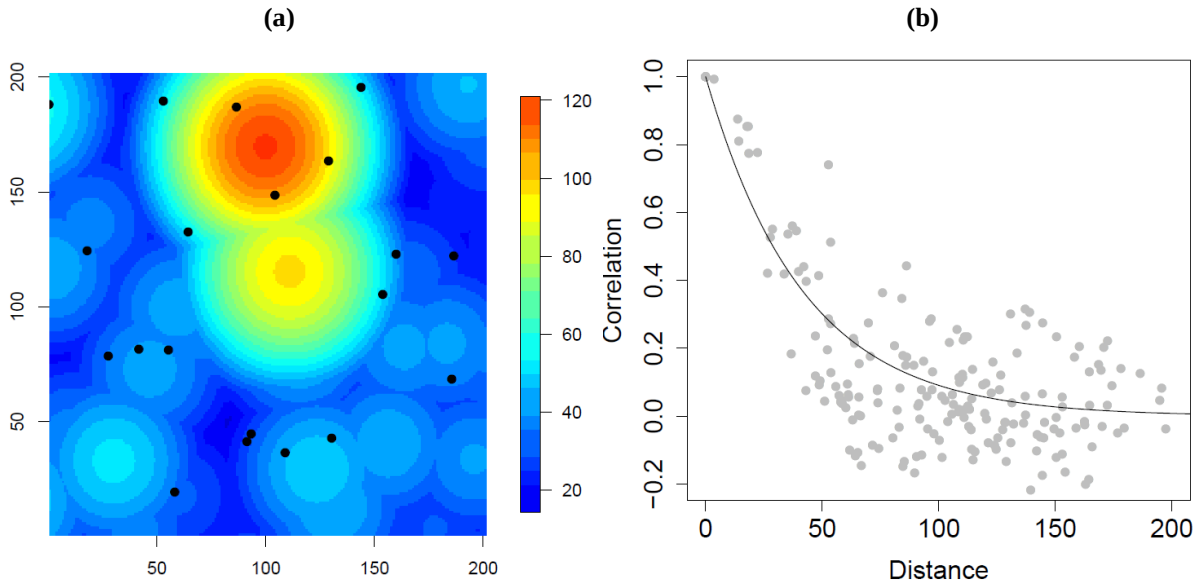


Figure 5.13- -(a) Simulated field with a Smith model for the high dependence case. Black dots are the observation sites. (b) Corresponding dependence-distance function. Gray dots are the correlation between pairs of sites, estimated from 50 replicated fields. The black curve is an exponential correlogram, fitted to the pairwise correlations using least squares.

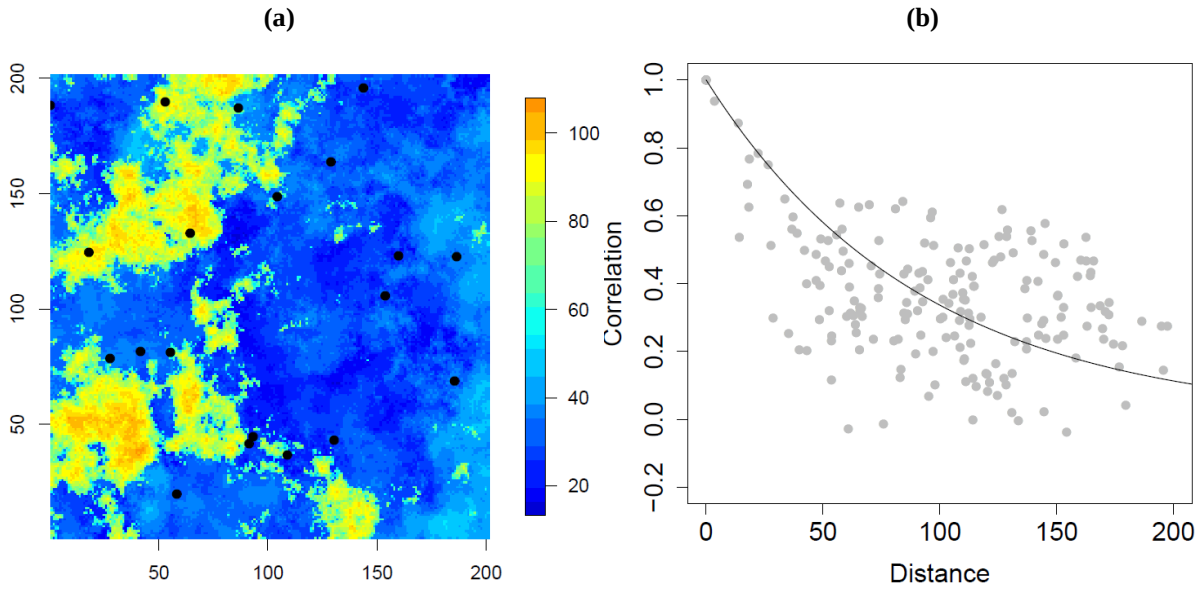


Figure 5.14- (a) Simulated field with a Schlather model for the high dependence case. Black dots are the observation sites. (b) Corresponding dependence-distance function. Gray dots are the correlation between pairs of sites, estimated from 50 replicated fields. The black curve is an exponential correlogram, fitted to the pairwise correlations using least squares.

The GEV parameters are estimated both considering and ignoring dependence. Figure 5.15 show the estimation results of 100 replications with data dependence simulated by a Gaussian copula. More evident than in Section 2.2.1, when ignoring the spatial dependence,

the coverage rate of the true value significantly drops for the location, trend and scale parameters, while it remains very close to 90% with the copula model: this highlights the advantage of the copula model by considering the spatial dependence. However, results are not as convincing when data are generated from max-stable models (Figure 5.16). Using a Gaussian copula significantly improves the coverage rate for the location and trend parameters. However, the Gaussian copula tends to over-estimate the shape parameter ξ, yielding a quite low coverage rate for this parameter.

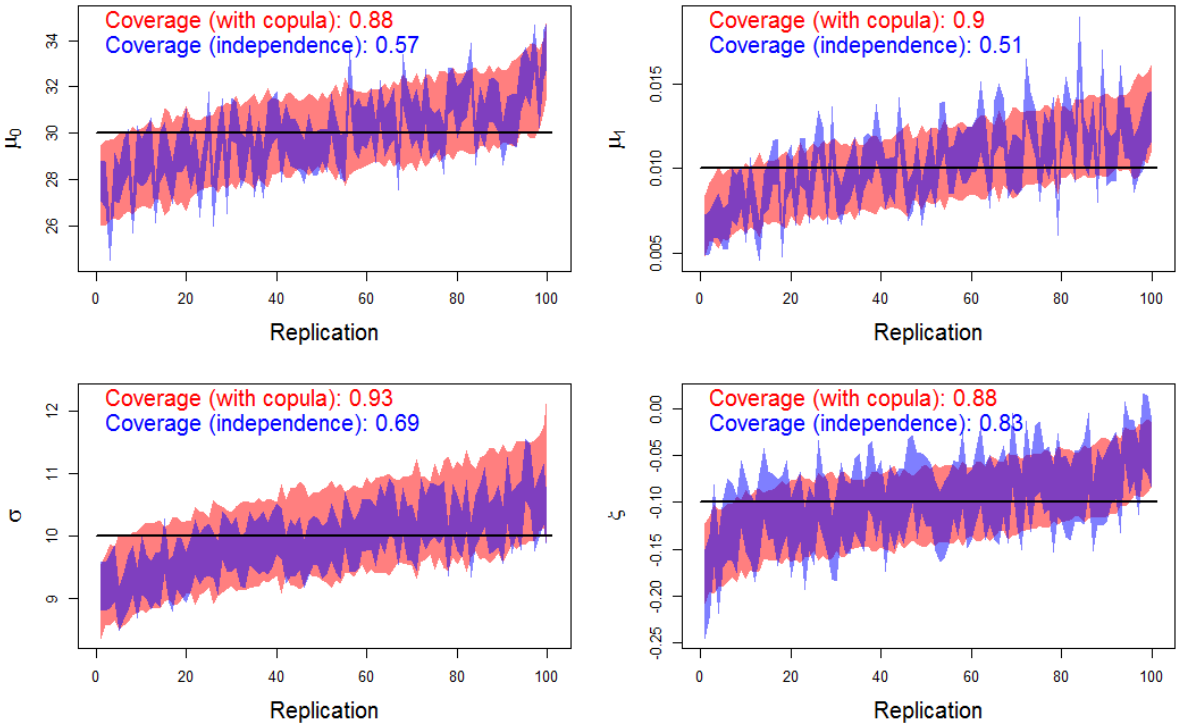


Figure 5.15-90% credibility intervals of the regression parameters for the high dependence case. Data are simulated with a Gaussian copula model. For each parameter, replications are sorted according to the posterior mean of the copula-based estimation.

2.3 Comparison with different spatial dependence models

In this section, we aim to investigate the differences in terms of estimating joint or conditional probabilities of exceedance of high values for several sites with different spatial dependence models. Spatial datasets are generated with a moderate dependence Smith model. These data are modeled with different spatial dependence models, including max-stable processes and copulas. Joint and conditional probabilities of exceedance are computed with both the estimated and the true marginal parameters.

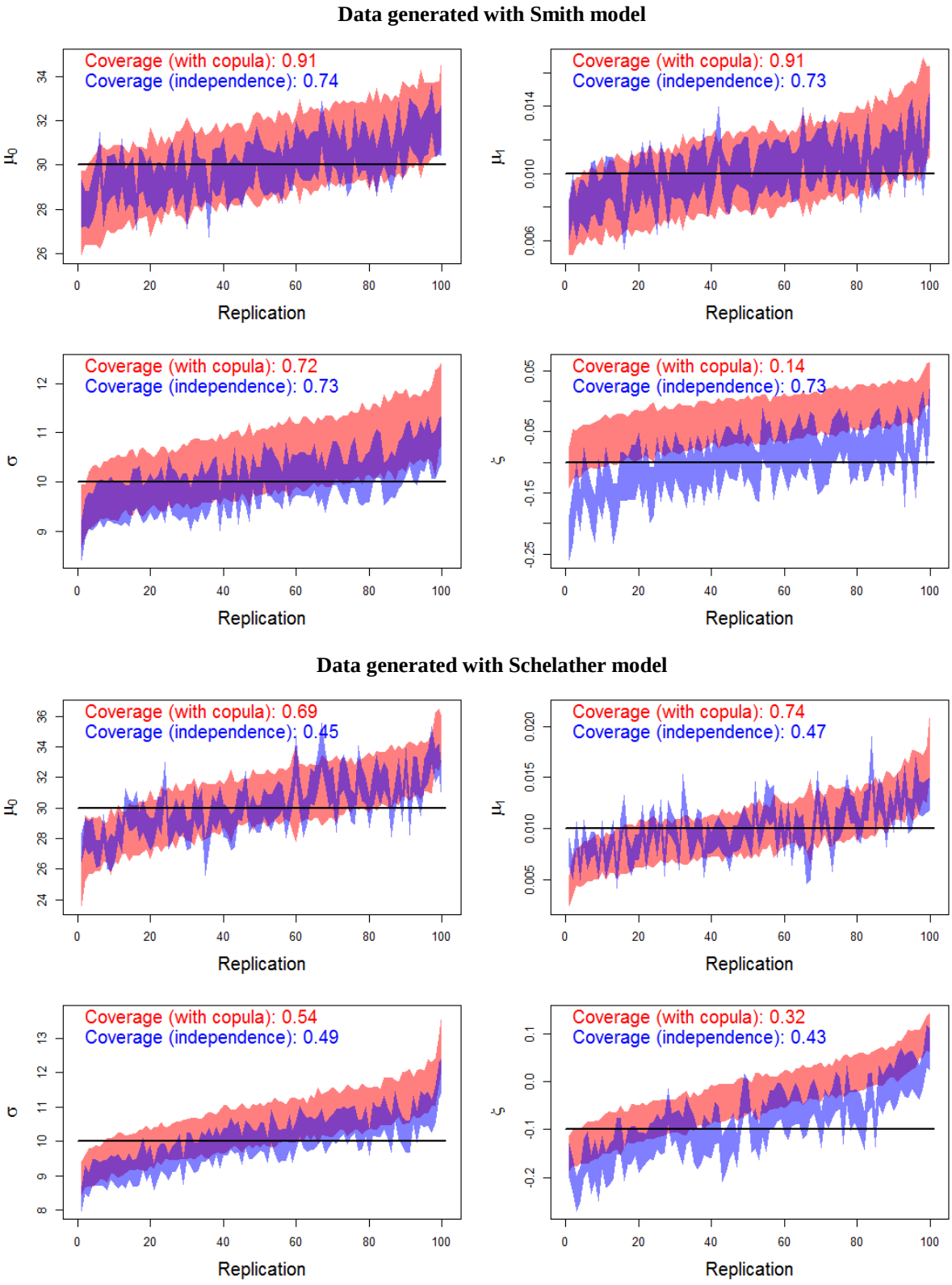


Figure 5.16-90% credibility intervals for the regression parameters for the high dependence case. Data are simulated with Smith and Schlather models. For each parameter, replications are sorted according to the posterior mean of the copula-based estimation.

2.3.1 Data simulation

Similar to Section 2.2, we suppose that there are 20 observation sites in a region. Each site contains 50 years of data. The coordinates (in kilometer) of these 20 sites are generated from a uniform distribution Unif[0,200]. Data are generated with a Gaussian Smith model. To facilitate the analysis of the spatial dependence, we assume that the GEV parameters do not vary with time. Thus data are generated from a GEV(30,10,-0.1), and the covariance matrix for the Smith model is $\Sigma_{11} = \Sigma_{22} = 200$ and $\Sigma_{12} = \Sigma_{21} = 0$. Figure 5.17 presents the simulated data with the Smith model for one time step. Blue dots denote the location of observation stations.

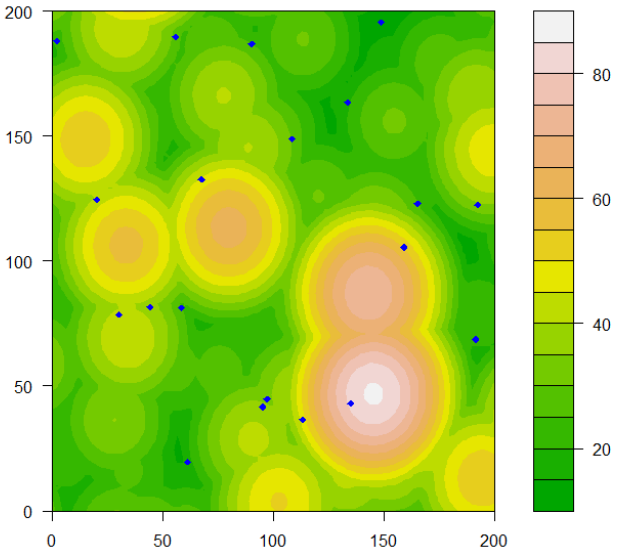


Figure 5.17-Simulated data with the Smith model at one time step. Blue dots are the observation stations. Color represents the intensity.

2.3.2 Inference on posterior distributions

The simulated data are modeled with a GEV distribution under different spatial dependence model assumptions listed in Table 5.2. Estimation with the Gaussian copula uses the general framework developed in Chapter 4. The dependence-distance function of the Gaussian copula is given by:

$$\forall i \neq j, \Sigma(i,j) = \exp(-\theta_1 * \| s_i, s_j \|) \tag{5.26}$$

Estimation with maximum stable processes uses the R package "SpatialExtremes".

Data are modeled with a stationary $GEV(\mu, \sigma, \xi)$ model. In total, $\mu, \sigma, \xi, \theta_1, \theta_2$ are the five R-parameters that need to be estimated. A uniform Unif[0,1] prior is used for θ_2 and flat priors are used for the other four parameters. Figure 5.18 presents the posterior distributions of the five parameters estimated within the Gaussian copula. The true parameter values are included in the "high density" area of the posterior distribution.

Table 5.2-Spatial dependence models

Model	Characterization
Copula	Gauss
Max Stable process: Smith	Gauss
Max Stable process: Schlather	Whittle-Matérn Cauchy Powered exponential Bessel

The estimations with the maximum stable processes are listed in Table 5.3. For all five maximum stable process models, the location and scale parameters are close to the true values, and are also close to the values estimated with a Gaussian copula. However, the shape parameter with Schlather models is further away from the true shape parameter value. The estimation results for the four variants of the Schlather models are very similar. The purpose of this analysis is to compare the difference between maximum stable processes and the copula model in terms of joint/conditional probability estimation, thus we are not going to provide further investigation on the results for marginal parameter estimation.

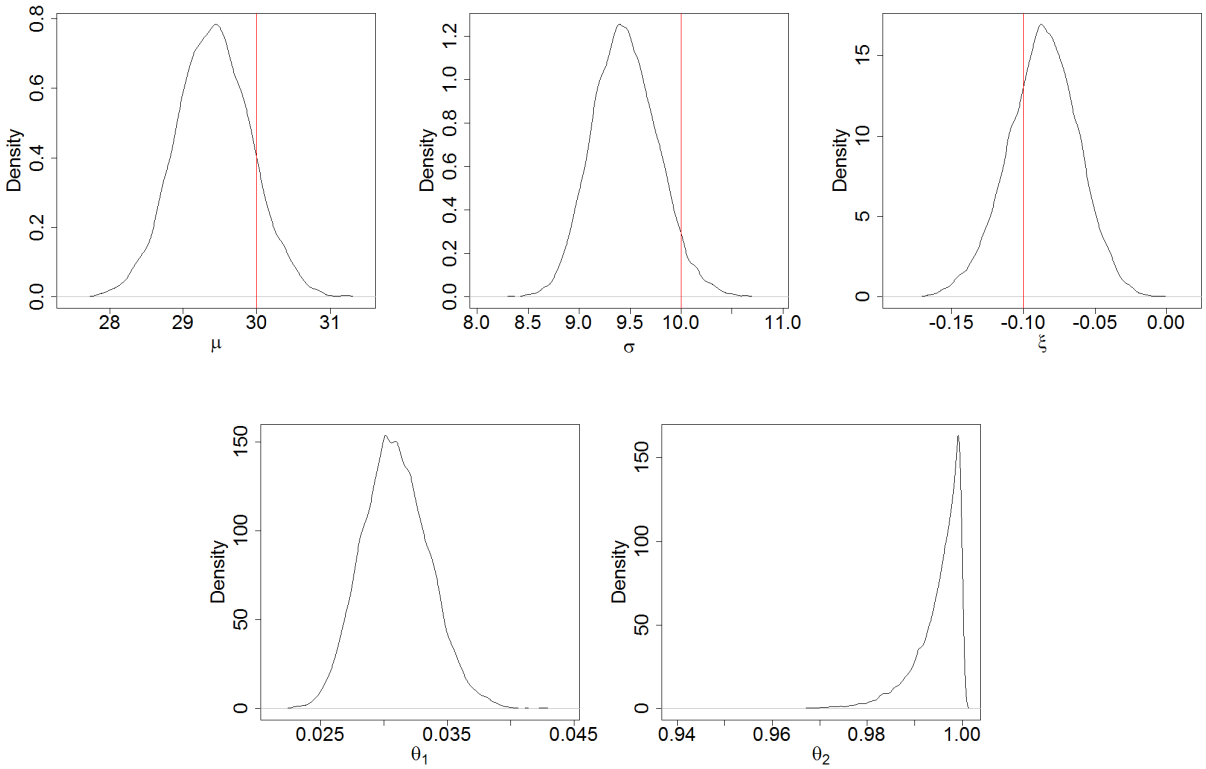


Figure 5.18-posterior distributions of the five parameters estimated with a copula-based model. Red lines are the true values of GEV parameters.

Table 5.3-Estimation results for different spatial dependence models

	Location	Scale	Shape
True value	30	10	-0.1
Smith	29.06	9.37	-0.14
Whittle-Matérn	29.22	9.87	-0.19
Cauchy	29.22	9.86	-0.19
Powered exponential	29.21	9.85	-0.19
Bessel	29.17	9.81	-0.19

2.3.3 Calculation of joint and conditional probabilities

An objective of this section is to compare joint and conditional probabilities of exceedance estimated using a copula and a maximum stable process. In this study, the marginal distribution at any site is the same. We restrict this comparison to two sites. More precisely, we are interested in the following two probabilities:

1. The probability of the annual maximum at both sites being larger than a certain value *z*: $\Pr\left(Z(s_1)>z, Z(s_2)>z\right)$, which is calculated as follow:

$$\begin{aligned}&\Pr\left(Z(s_1)>z, Z(s_2)>z\right)=1-\Pr\left(\{Z(s_1)\le z\}\cup\{Z(s_2)\le z\}\right)\\&=1-\left(\Pr\left(Z(s_1)\le z\right)+\Pr\left(Z(s_2)\le z\right)-\Pr\left(Z(s_1)\le z, Z(s_2)\le z\right)\right)\\&=1-2\Pr\left(Z(s_1)\le z\right)+\Pr\left(Z(s_1)\le z, Z(s_2)\le z\right)\end{aligned} \tag{5.27}$$

2. The probability of the annual maximum at one site being larger than a certain value *z*, conditional on the other site having also recorded a value larger than *z*: $\Pr\left(Z(s_1)>z \mid Z(s_2)>z\right)$, which is calculated as follow:

$$\begin{aligned}&\Pr\left(Z(s_1)>z \mid Z(s_2)>z\right)=\Pr\left(Z(s_1)>z, Z(s_2)>z\right)/\Pr\left(Z(s_2)>z\right)\\&=\Pr\left(Z(s_1)>z, Z(s_2)>z\right)/\left(1-\Pr\left(Z(s_2)\le z\right)\right)\\&=\left(1-2\Pr\left(Z(s_1)\le z\right)+\Pr\left(Z(s_1)\le z, Z(s_2)\le z\right)\right)/\left(1-\Pr\left(Z(s_1)\le z\right)\right)\end{aligned} \tag{5.28}$$

The joint probability $\Pr\left(Z(s_1)\le z, Z(s_2)\le z\right)$ is given by Eq(5.17) and Eq(5.20) for maximum stable processes and Eq(4.10) for copulas. As the joint probability depends on the distance *h* between two sites, in the following, it is denoted by $\Pr\left(Z(s_1)\le z, Z(s_2)\le z \mid h\right)$.

2.3.4 Difference in joint probabilities of exceedance

In this section, we analyze the relationship between the joint exceedance probability, the inter-site distance and the exceeded rainfall level *z*:

$$g_1(h,z) = \Pr\left(Z(s_1) > z, Z(s_2) > z \mid h\right) \tag{5.29}$$

In particular, we are interested in the following two questions:

1. With a fixed distance *h*, how does the joint probability vary with threshold *z*?
2. With a fixed threshold *z*, how does the joint probability vary with distance *h*?

Figure 5.19 presents the joint probability with respect to the threshold value for fixed distances 0, 20, 50 and 100 km. The case $h = 0$ corresponds to the marginal distribution. In this case, results from different maximum stable process models are identical. In order to understand whether discrepancies between models are due to the modeling of spatial dependence or the estimation of marginal parameters, we also report probabilities computed with the true marginal GEV parameter values.

In general, the joint probability calculated with Schlather models is larger than that calculated with Smith model and Gaussian copula. For small distance ($h \leq 20$), the joint probability with copula model is the smallest. For large distance ($h \geq 50$), the results with Gaussian copula and Smith model become similar. However, results with Schlather models remain markedly above those two models. This is a consequence of the dependence not vanishing to zero at infinite distances with the Schlather model, thus yielding larger joint exceedance probabilities than the Gaussian copula and the Smith models.

Figure 5.20 presents the joint probability with respect to the distance for fixed threshold value of 50, 80 and 100. It is logical that the joint probability goes down when the inter-site distances increase. However, a limit exists for the joint probability when the distance grows. In fact, according to Eq(5.27), with a fixed *z*, the joint probability only varies with $\Pr\left(Z(s_1) \leq z, Z(s_2) \leq z \mid h\right)$. With a Gaussian copula and the dependence-distance relationship assumed in Eq(5.26), this term converges to the square of the marginal probability $\Pr\left(Z(s) \leq z\right)$ since the dependence vanishes to zero at large distances. A similar behavior is observed with the Smith model, since the dependence also vanishes to zero at large distances (see Eq(5.17)). However, Schlather models yield a much higher joint probability at large distance. This is because the dependence does not vanishes to zero at larges distances (see Eq(5.20)), and the joint probability therefore does not converge to the square of the marginal probability. Note however that at short distances, the Smith and Schlather models yield similar joint probabilities that are consistently higher than the one observed with a Gaussian copula model.

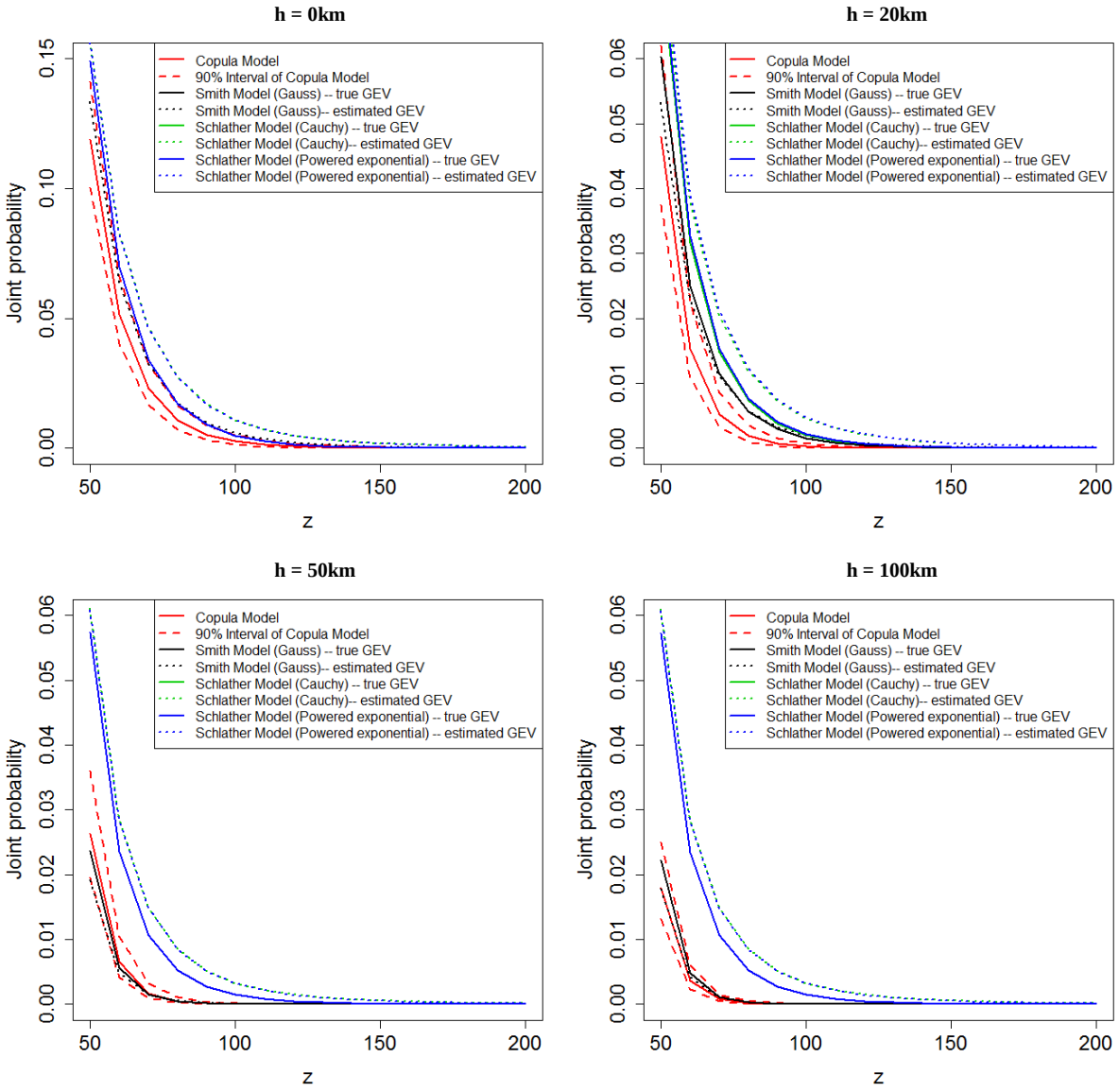


Figure 5.19-Joint exceedance probability with fixed distance

The peculiar behavior of Schlather models at large distances seems at odds with the expected behavior of rainfall fields, whose correlation should decrease and tend to zero when the inter-site distance increases. In the following, we will therefore only consider the Smith model and the Gaussian copula model.

2.3.5 Difference in the conditional probability

In this section, we are going to analyze how the conditional probability of exceedance varies according to the dependence model (Gaussian copula or Smith model). The same notation as in Section 2.3.4 is used.

$$g_2(h,z) = \Pr\left(Z(s_1) > z \mid Z(s_2) > z, h\right) \tag{5.30}$$

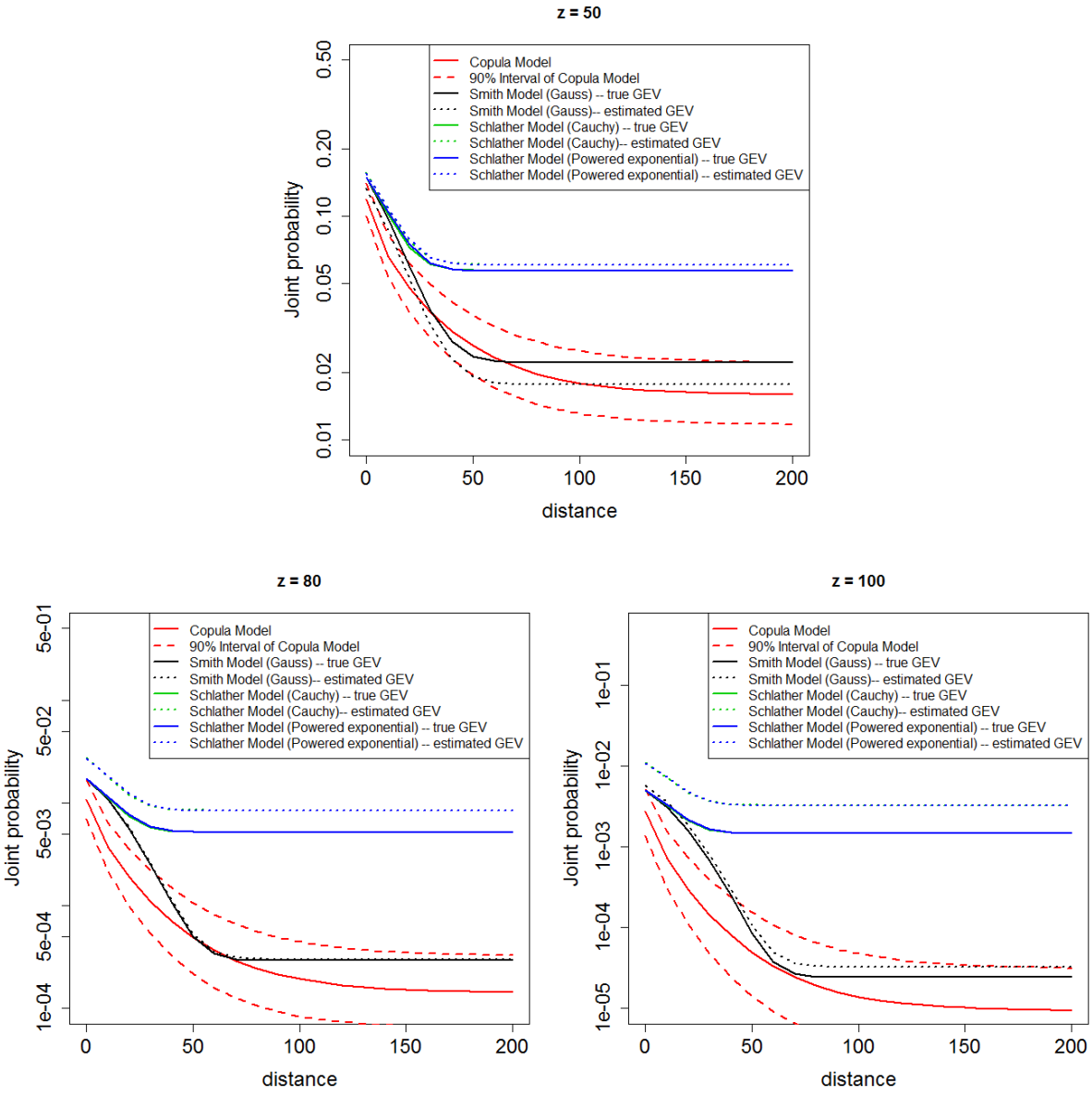


Figure 5.20-Joint probability with fixed threshold value

In particular, we are intested in two specific questions:

1. With a fixed distance *h*, how does the conditional probability vary with threshold *z*?
2. With a fixed threshold *z*, how does the conditional probability vary with distance *h*?

Figure 5.21 presents the conditional probability $g_2(h,z)$ with respect to the threshold value for fixed distances of 0, 20, 50 and 100km. For *h*=0, two sites are at the same location, then the conditional probability should be 1 as found by the Smith model, while a nugget exists for the copula based model. For distances h = 20 or 50, the conditional probability calculated with the Smith model tends to a non-zero value, while that calculated with the Gaussian copula tends to zero with increasing values of *z*. This is because the Smith model is asymptotically dependent and the copula model is asymptotically independent. Note that with

the log-scale in Figure 5.21, the difference of conditional probability reaches one order of magnitude for z=200. In practical terms, this implies that conditional exceedance probabilities may be markedly underestimated with a Gaussian copula model if data are actually asymptotically dependent. However, the asymptotic value of the Smith model tends to zero at a large distance h=100, and both the copula and the Smith model therefore yield similar conditional probabilities, that are close to the marginal probabilities since independence is virtually reached at such a large distance.

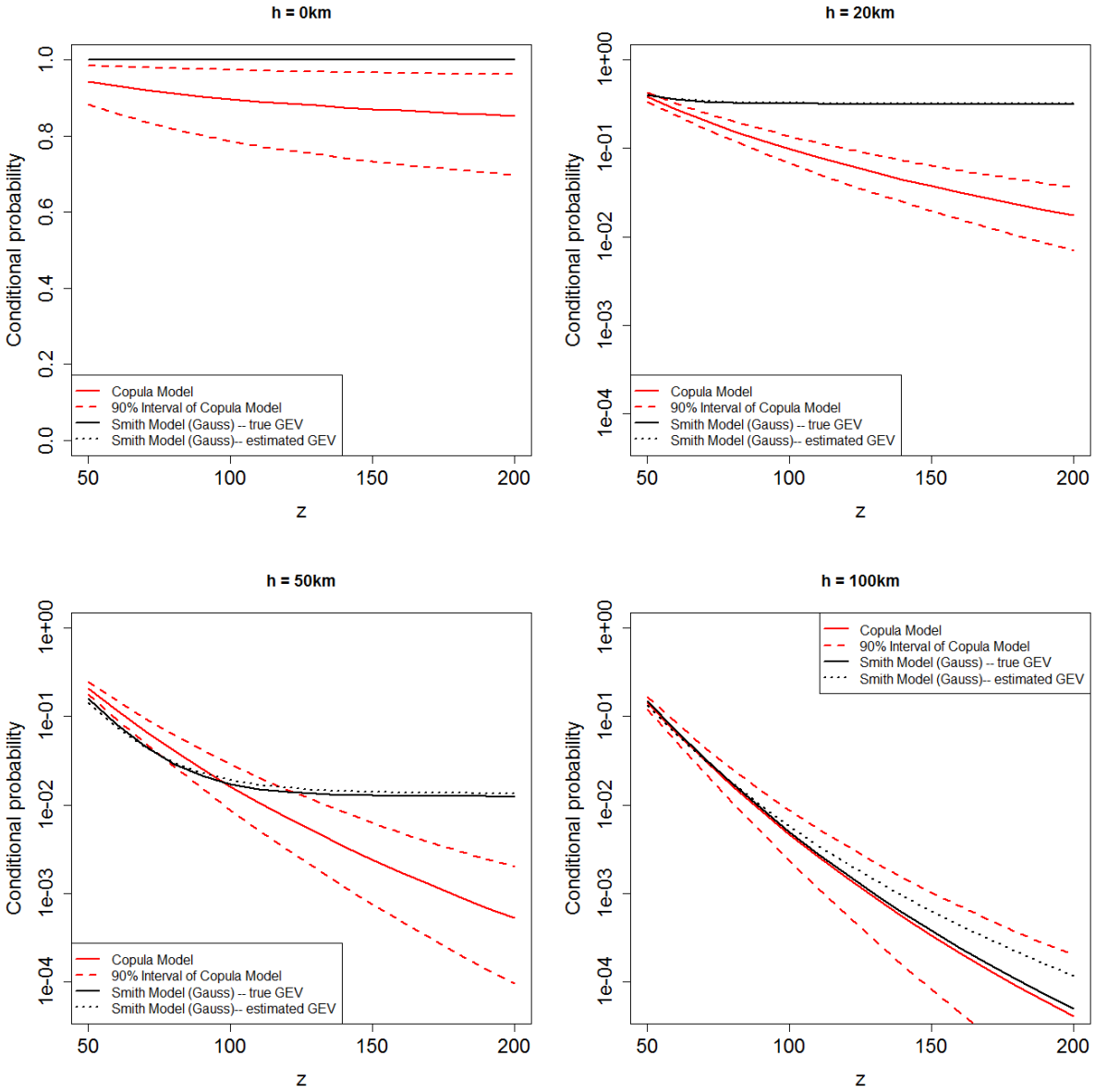


Figure 5.21-Conditional probability with fixed distance. For the Smith model, the estimated GEV is overlapped with the true GEV for h=0 km.

Figure 5.22 presents the conditional probability with respect to the distance for fixed threshold values of 50, 80 and 100. This figure has exactly the same shape as Figure 5.20, because for a fixed z, the conditional probability is the joint probability divided by the

marginal probability according to Eq(5.27)(5.28). Similar conclusions can therefore be drawn: at large distances, the conditional probabilities calculated with both the Smith and copula models converge to the marginal probability, but the Smith model yields larger conditional probabilities at short distances.

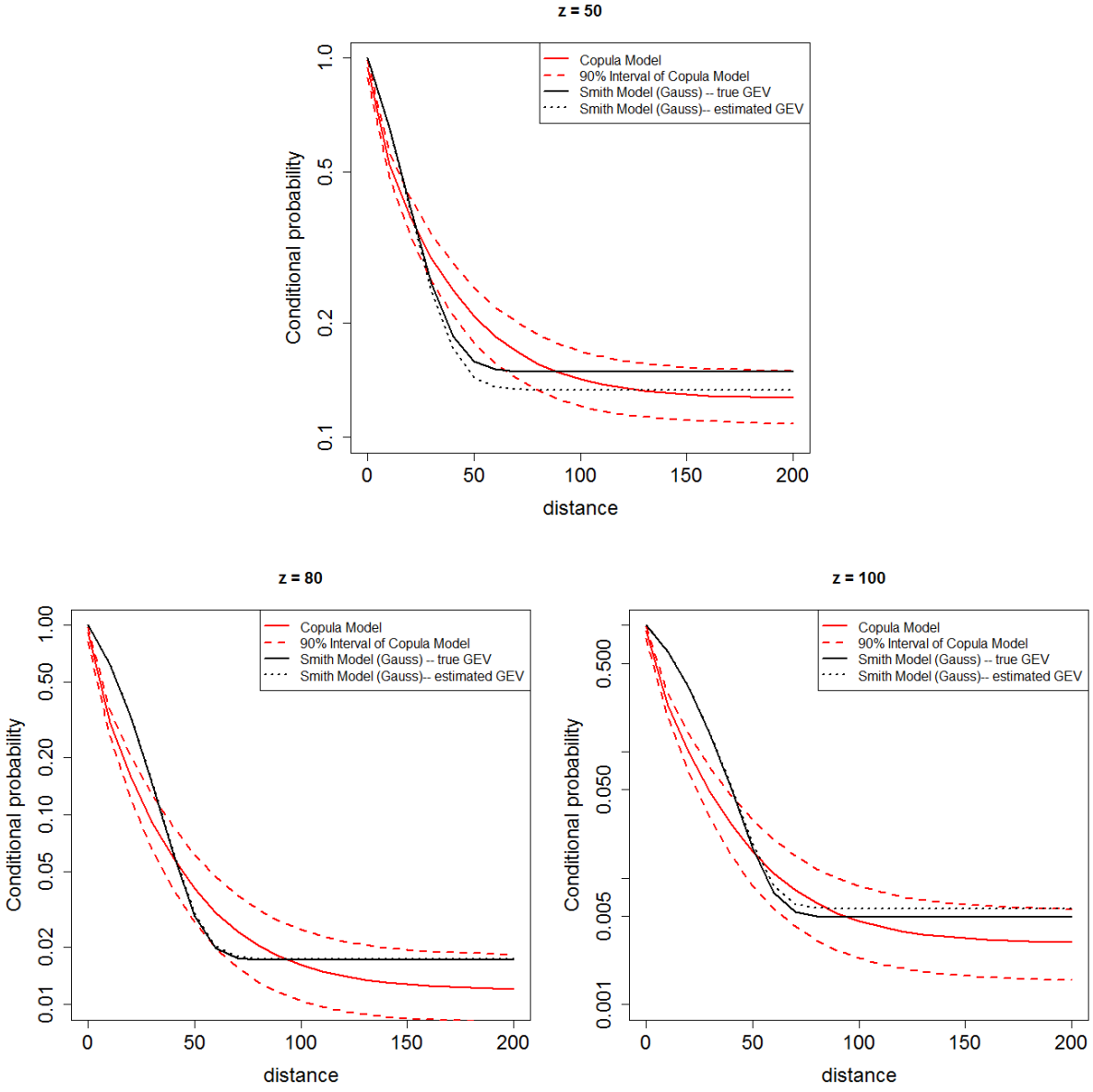


Figure 5.22-Conditional probability with fixed threshold value

2.4 Conclusion

In this section, we first investigated the reliability of using a copula to model data generated from a maximum stable process. In a moderate dependence case, regression parameters of a GEV distribution can be well-estimated with the copula. In particular, the estimation quality is better than that ignoring the spatial dependence: using a Gaussian copula yields a more realistic quantification of uncertainties. This is especially the case for the trend

parameter, which is important in the context of this thesis, since a major objective of the modeling framework is to improve the identification of trends or climate variability effects.

Results are not as convincing in a high dependence case. Estimation with a Gaussian copula provides also a good estimation for the location and trend regression parameters, with a much better quantification of uncertainty than that the estimation ignoring spatial dependence. The estimation of the scale parameter is similar when considering or ignoring spatial dependence. However, the shape parameter tends to be over-estimated when using the copula, while this behavior is not observed if spatial dependence is simply ignored.

These results indicate that the use of a Gaussian copula is reliable for moderate dependence cases. However, for datasets with a high degree of spatial dependence, using a Gaussian copula may yield a marked overestimation of the shape parameter. It is worth noting that these results are conditional to the two particular representations of max-stable processes that were used here (Smith and Schlather models). However, none of these models is fully realistic as a representation of rainfall fields: the Smith model yields much too smooth fields, while spatial dependence do not vanishes to zero at infinite distances with the Schlather model. This calls for additional studies with alternative representations of max-stable fields.

In a second step, Gaussian copula, Smith model and Schlather model are compared in terms of estimating joint or conditional probabilities of exceedance of high values for several sites. In general, Schlather models yield higher conditional and joint probabilities, because a minimum dependence between sites exists even at infinite distances according to the construction of Schlather models. The joint probabilities calculated with Smith model and Gaussian copula are generally similar at large distances, but the Smith model yields higher joint probabilities at shorter distances. The conditional probabilities calculated with the Gaussian copula and the Smith model can be markedly different, which is a consequence of the Gaussian copula being asymptotically independent, while the Smith model is asymptotically dependent.

Overall, these results suggest that even if a Gaussian copula approximation may yield acceptable results in terms of estimating marginal parameters, the computation of joint or conditional exceedance probabilities is much more sensitive to the representation of spatial dependence.

Part III General Applications:

ENSO impact on precipitation

General Introduction about the impact of ENSO on precipitation

Every year worldwide, extreme rainfall results in flooding that leads to loss of life and infrastructure and damages economies. Not only are the direct effects of floods felt for many years, but the indirect effects, including disease, trauma and social dislocation, reduced agricultural production or loss of manufacturing capacity, can be felt for decades. The El-Niño/La Niña Southern Oscillation (ENSO) is the single most influential climate phenomenon producing global extremes of precipitation [*Dai et al.*, 1997], and has been researched extensively since a major El Niño event in 1982–83.

The quality of ENSO forecasts has consequently dramatically improved in the last two decades, and scientists can predict ENSO events with more than 70% accuracy one year before their occurrence [*Weiher*, 1999]. Nevertheless, if the variation of extreme precipitation conditional on ENSO could be better understood, planners and engineers could improve operating rules and plan better emergency responses before predicted floods occur. The current studies provide important new insights into the effects of ENSO on extreme precipitation, including methods for quantifying the impact.

A large number of studies have discussed the relationship between ENSO and average regional precipitation. In boreal winter, positive anomalies are found in southwestern North America during El Niño episodes and northwestern North America during La Niña episodes [*Castello and Shelton*, 2004; *Meehl et al.*, 2007]. Positive anomalies are detected in southeastern South America during austral winter, spring and summer, and northeastern South America during autumn [*Fernandez and Fernandez*, 2002; *Grimm*, 2011; *Kayano and Andreoli*, 2006]. La Niña enhances the rainfall in South Africa in summer [*Kruger*, 1999; *Vanheerden et al.*, 1988], and significant impact is also found in Asia [*Kane*, 1999; *Kripalani and Kulkarni*, 1997; *Li and Ma*, 2012; *Wu et al.*, 2003] and Australia [*Cai and Cowan*, 2009]. A global pattern of the impact of ENSO on precipitation was described by *Dai et al.* [1997] and *Dai and Wigley* [2000]. *Ropelewski and Halpert* [1987] provided a global pattern of magnitude, phase and duration of ENSO-related precipitation.

Compared with average regional precipitation, extreme precipitation events are considerably more uncertain in terms of timing and intensity. The study of the impact of ENSO on extremes is therefore complicated. Several studies have been carried out for specific regions of the world. Significant influence on the frequency of extreme precipitation in North America [*Cayan et al.*, 1999; *Gershunov and Barnett*, 1998; *Higgins et al.*, 2011; *Jones and Carvalho*, 2012; *Schubert et al.*, 2008] and South America [*Grimm and Tedeschi*, 2009; *Pscheidt and Grimm*, 2009] has been reported. *Wan et al.* [2013] analyzed the relationship between ENSO and extreme monthly precipitation in China, while *Min et al.* [2013] studied the impact of ENSO on extreme rainfall in Australia using gridded rainfall data.

At the global scale, *Curtis et al.* [2007] investigated the correlation between ENSO and the frequency of estimated precipitation extremes for different seasons, while *Lyon and Barnston* [2005] discussed the spatial extent of tropical land precipitation extremes related to ENSO extremes. *Kenyon and Hegerl* [2010] reported on the response of global extreme precipitation to ENSO. Their study was based on the at-site (local) analysis, thus the significance of the impact of ENSO in a great number of sites was masked. *Alexander et al.*

[2009] studied the global response of precipitation extremes to global SST variability. While all of these studies added to the knowledge of ENSO-precipitation teleconnection, few of them established a quantified relationship between ENSO intensity and the severity of the impact on extreme precipitation.

Several studies have demonstrated that the teleconnection between climate variables (e.g., pressure and precipitation) and ENSO is not symmetric [*Hannachi*, 2001; *Hoerling et al.*, 1997; *Hoerling et al.*, 2001; *OrtizBevia et al.*, 2010; *Sardeshmukh et al.*, 2000]. Asymmetry has also been reported for the teleconnection with regional rainfall, and *Cai et al.* [2010]; [2012], *King et al.* [2013]. Asymmetry has also been described for the winter precipitation response in North America *Wu et al.* [2005], and the impact of ENSO on the spring rainfall in south China [*Feng and Li*, 2011] and Taiwan [*Chen et al.*, 2008] is also considered asymmetric. No study, however, provides a global pattern of the asymmetry of the impact of ENSO on precipitation in general, and on extreme precipitation in particular.

As discussed in Chapter 1 (Section 3.2), the peculiarity of extreme precipitation (as opposed to seasonal or annual averages) requires the development of particular statistical models to describe the impact of ENSO or, more generally, of any other type of climate variability. Several time-varying and climate-informed models have been developed for analyzing and predicting the impact of climate variability on the intensity of extreme precipitation. At the local scale, *Renard et al.* [2006a] developed a time-varying frequency analysis (FA) model for quantifying temporal trends in extreme hydrological events, while *Tramblay et al.* [2012] used a similar model to study the impact of the North Atlantic Oscillation (NAO) and the Mediterranean Oscillation (MO) on extreme precipitation in Morocco.

Ouarda and El-Adlouni [2011] discussed the impact of ENSO on annual maximum precipitation at the Tehachapi station in California using a time-varying Bayesian FA model, after *Kwon et al.* [2008] used a hierarchical Bayesian logistic regression model incorporating different climate components to analyze extreme summer rainfall in South Korea. Again at the regional scale, *Renard et al.* [2008] and *Hanel et al.* [2009a] described the use of time-varying regional models for extreme streamflow and precipitation. *Aryal et al.* [2009] and *Lima and Lall* [2010] used time-varying hierarchical approaches for climate teleconnections on extreme precipitation in the Swan-Avon River basin of southwestern Western Australia and rainfall occurrence in northeastern Brazil, respectively. *Shang et al.* [2011] used a maximum stable process for analyzing the impact of ENSO on extreme rainfall over California, USA.

Limitations of previous studies on the impact of ENSO on extreme precipitation can be summarized as follows:

1. Few of the studies focus on at-site observed extremes, which are much more uncertain than temporally or spatially averaged precipitations, and for which the impact of ENSO is more difficult to quantify.
2. The current analyses using extreme-specific models with data extremes observed on-site mostly provide individual studies from site to site. Such at-site (local) analysis could mask the impact of ENSO due to huge uncertainties. On the other hand, most analyses using regional models do not consider the spatial dependency of the data.

3. Most analyses at the global scale are based on gridded datasets, as opposed to station observations. Gridded data are not suitable for at-site extremes, because they involve some form of spatial averaging.
4. Most of the studies do not quantify the increase or decrease in extreme quantiles with respect to ENSO intensity at the regional or global scale.

In this part, two case studies are presented, in which the limitations listed above are essentially overcome by using the time-varying RFA framework developed in this thesis. To avoid ambiguity, we clarify that both studies describe the link between ENSO and precipitation variability. However, being a purely statistical analysis, it does not formally establish causality. Chapter 6 describes the impact of ENSO in Southeast Queensland (SEQ) Australia, and Chapter 7 gives a global analysis of the asymmetric impact of ENSO on extreme precipitation.

CHAPTER 6 Quantifying the impact of ENSO on summer rainfall in Southeast Queensland, Australia

There is increasing evidence that the distribution of hydrologic variables such as average or extreme rainfall/runoff is modulated by modes of climate variability in many regions of the world. This chapter presents an application of the general spatio-temporal RFA framework for quantifying the effect of climate variability on the distribution of hydrologic variables.

This modeling framework is applied to two case studies assessing the effect of El Niño Southern Oscillation (ENSO) on summer rainfall in Southeast Queensland. The first case study focuses on summer rainfall totals while the second analysis focuses on extremes using summer daily rainfall maxima. The reason for choosing data from this area is that *Cai et al.* [2010] found an asymmetric impact of ENSO on the summer rainfall in Southeast Queensland (SEQ): La Niña episodes correspond to marked positive rainfall anomalies in SEQ, with the anomalies being linearly related to the strength of the La Niña, while El Niño episodes do not appear to have any noticeable effects on rainfall. *Cai et al.* [2010] focused on spatially averaged seasonal rainfall over a large region. In this chapter, we will investigate two questions: (i) Is the asymmetric impact still evident on the summer rainfall totals of individual sites? (ii) Does a similar asymmetric impact also hold for extremes? The Southern Oscillation Index (SOI), a measure of ENSO, is considered as a time-varying covariate. In order to account for different effects during La Niña and El Niño episodes, an asymmetric piecewise-linear regression is used to analyze the rainfall data using both local and regional models.

1 Quantifying the impact of ENSO on summer rainfall totals using local models

This study uses a local model to verify and quantify the asymmetric impact of ENSO on summer rainfall totals over SEQ.

1.1 Data and covariates

Rainfall data are provided by the Australian Bureau of Meteorology (BOM). High quality summer (Dec, Jan, Feb) totals [*Lavery et al.*, 1997] are available over 16 observation sites until 2011, with the record starting year among these sites ranging from 1870 to 1913, with most having a record longer than one hundred years. Figure 6.1 shows the location of rain gauges.

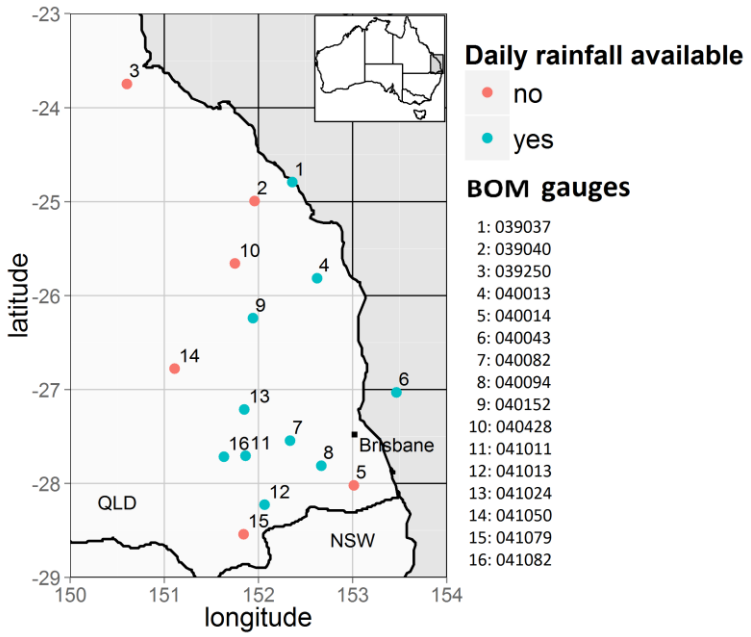


Figure 6.1-Locations of the rain gauges. Summer rainfall totals are available in all 16 gauges. The blue dots are the gauges in which daily rainfall data are available, which will be used to compute the summer daily maxima.

The Southern Oscillation Index (SOI) and Niño 3.4 are two main indices that characterize the strength of ENSO phenomenon. The SOI index is calculated from the mean sea level pressure difference between Tahiti and Darwin. Sustained negative SOI values below about –8 indicate an El Niño event while sustained positive values above +8 indicate a La Niña event. El Niño events are associated with a warming of the central and eastern tropical Pacific, while La Niña events are associated with a sustained cooling of these same regions. These temperature gradients across the Pacific are the most important driver of ENSO. To

simplify this usage, in this study, we consider negative (resp. positive) SOI periods as El Niño (resp. La Niña) phases. The SOI data (1877-2011) used in this study was obtained from the Australian Bureau of Meteorology (BOM) (http://www.bom.gov.au/climate/ current/soi2.shtml).

Niño 3.4 is computed through the sea surface temperature (SST) of the area 170E to 120W and 5N to 5S in the central Pacific. Due to the difficulty of measurement, the historical data is available only since 1950. Opposite to SOI, positive (resp. negative) Niño 3.4 describes El Niño (resp. La Niña) phase. In preliminary analyses for the period 1950-2011 (Section 1.3.1), the two indices are compared as potential covariates: SOI (1877-2011) and Niño 3.4 (1950-2011).

1.2 Local model for the summer rainfall totals

The previous study by *Cai et al.* [2010] suggests separating the effect of La Niña (positive SOI or negative Niño 3.4) and El Niño (negative SOI or positive Niño 3.4) episodes on the summer rainfall in SEQ. Thus the specific implementation of the parent distribution is to use a lognormal model for the summer total rainfall, as follow:

$$Y(t) \sim \log N(\mu(t), \sigma(t)) \tag{6.1}$$

where the mean $\mu(t)$ is asymmetric with respect to the positive and negative values of the temporal covariate, while the standard deviation is assumed to be constant:

$$\mu(t) = \begin{cases} \mu_0 + \mu_1^- * Cov(t); Cov(t) < 0 \\ \mu_0 + \mu_1^+ * Cov(t); Cov(t) > 0 \end{cases} \tag{6.2}$$

$$\sigma(t) = \sigma_0 \tag{6.3}$$

where μ_0, μ_1^-, μ_1^+ and σ_0 are the regression parameters and $Cov(t)$ is the temporal covariate (summer averaged SOI or Niño 3.4). In this study, flat priors are used for the regression parameters.

1.3 Results

1.3.1 Preliminary analyses with the standardized SOI and Niño 3.4 indices over the period 1950-2011

In this section, we first evaluate the similarity of these two indices for the model by using the data from the period 1950-2011.

Figure 6.2 illustrates the standardized summer averaged Niño 3.4 and SOI indices during 1950-2011. To facilitate the comparison, Niño 3.4 is inversed. Evidently, these two indices are highly correlated and show a similar temporal pattern, suggesting that they should provide similar covariate information for the model in Eq(6.2).

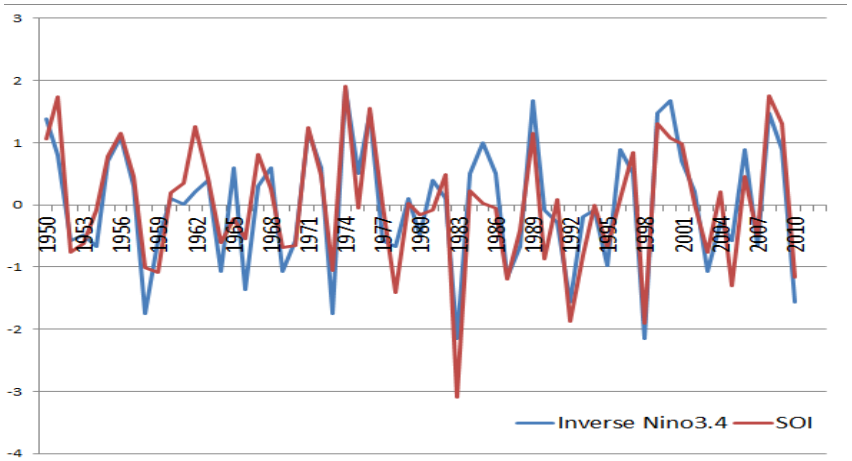


Figure 6.2-Standardized Niño 3.4 and SOI indices

This is further investigated by comparing the posterior distributions of the parameters quantifying the ENSO effect. During El Niño phases (Figure 6.3), the distributions are similar for all stations and do not suggest any significant effect of ENSO. During La Niña phases (Figure 6.4), the distributions are still similar for all stations, but indicate a positive effect of ENSO.

Given the similarity of the results obtained with covariates Niño 3.4 and SOI, we will restrict to the latter in the remainder of this chapter due to its longer period of availability.

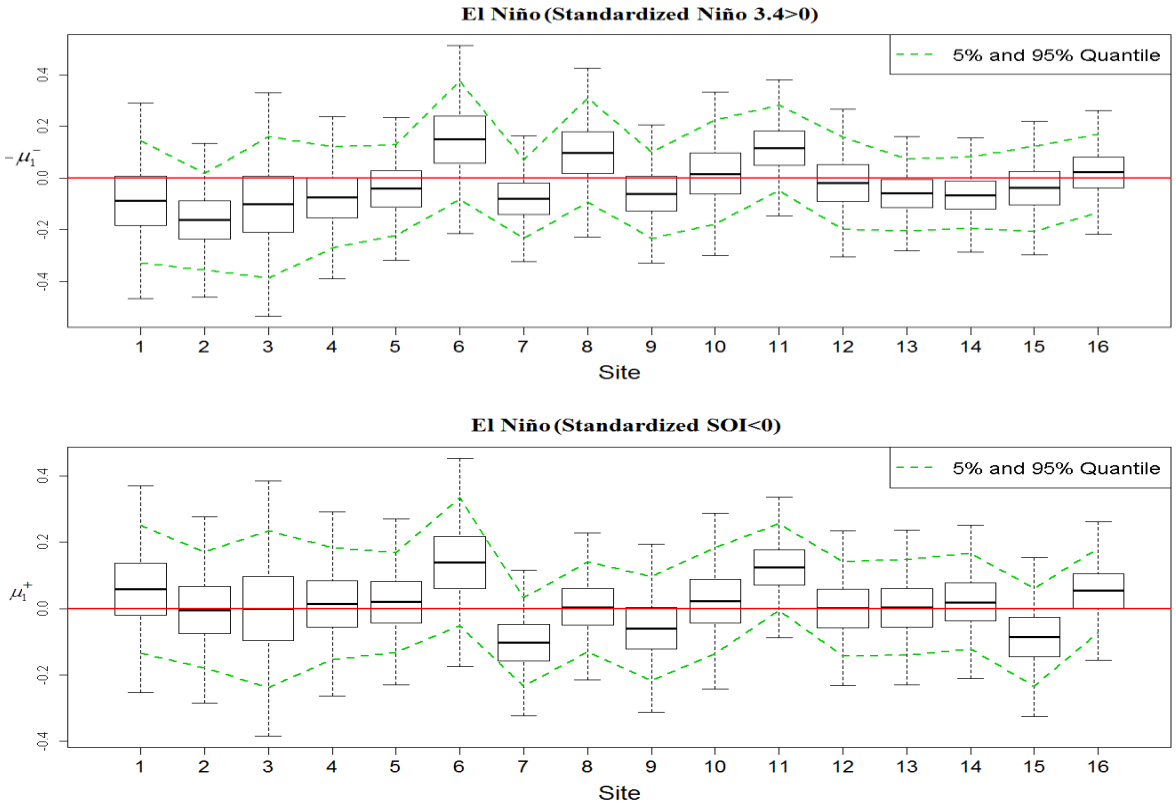


Figure 6.3- Boxplot of the posterior distribution of $-\mu_1^-$ (Niño 3.4) and μ_1^+ (SOI) during El Niño phase.

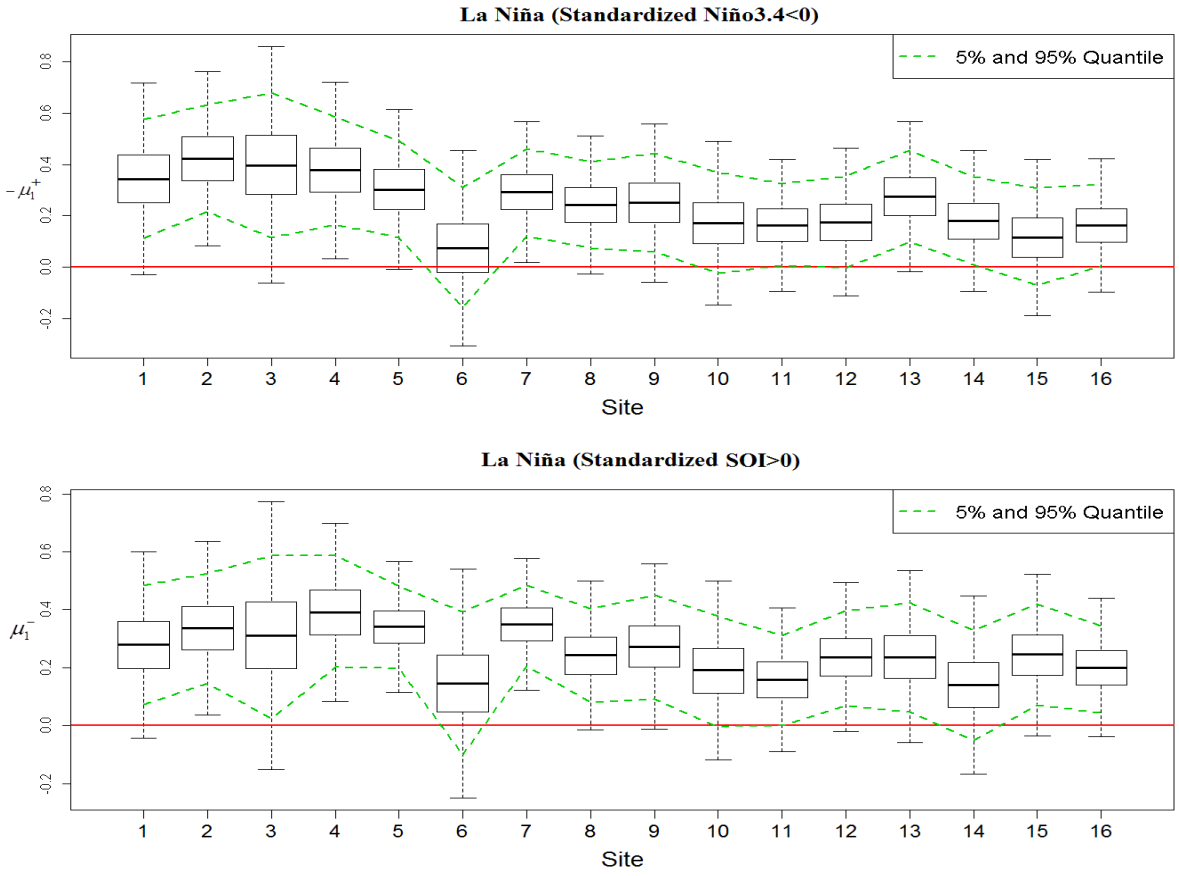


Figure 6.4- Boxplot of the posterior distribution of $-\mu_1^+$ (Niño 3.4) and μ_1^- (SOI) during La Niña phase.

1.3.2 Results with SOI used as covariate over the period 1877-2011

The goodness-of-fit is evaluated graphically by using a PP plot. In this study, the PP plot of each site is close to the diagonal (not shown), which indicates that lognormal distribution provides a good fit to the observations.

The impact of El Niño (negative SOI) and La Niña (positive SOI) on the summer rainfall is characterized by μ_1^- and μ_1^+ respectively. If such impact is significant, the posterior distributions of μ_1^-, μ_1^+ should be significantly different from zero. Figure 6.5 indicates that most sites are significantly influenced by La Niña, whereas El Niño influence is not detected.

To further illustrate the effect of La Niña and El Niño, the p-value of 0 is calculated for the regression parameters μ_1^- and μ_1^+. This p-value is equal to $\Pr[\mu \le 0 \,|\, \boldsymbol{Y}]$, which refers to the probability of the posterior distribution of μ_1^- or μ_1^+ being smaller than 0. Figure 6.6 shows the p-value for all 16 sites on a map. During El Niño episodes, the majority of sites show little effect. However, during the La Niña episodes, the significance is quite clear.

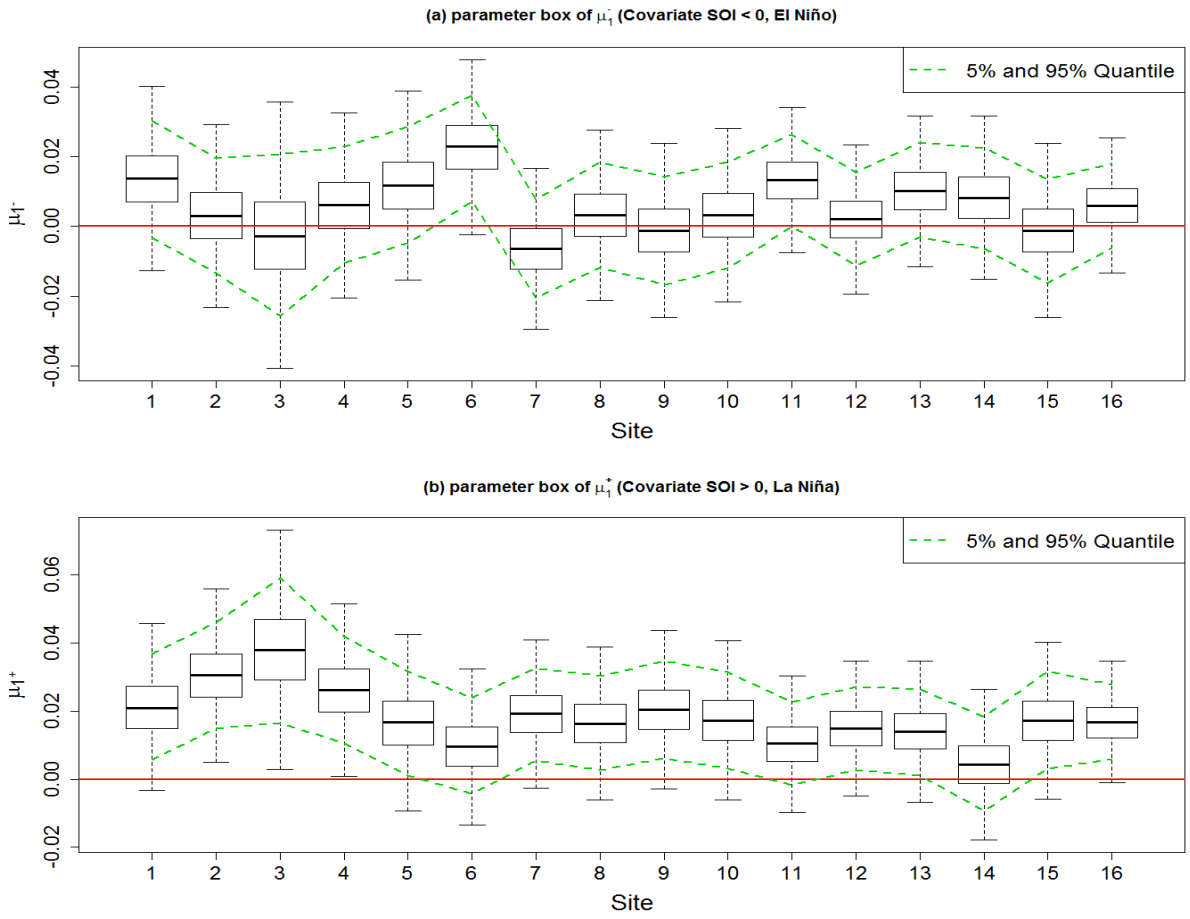


Figure 6.5-Boxplot of the posterior distribution of (a) μ_1^- (El Niño) and (b) μ_1^+ (La Niña) for each site for the summer rainfall totals

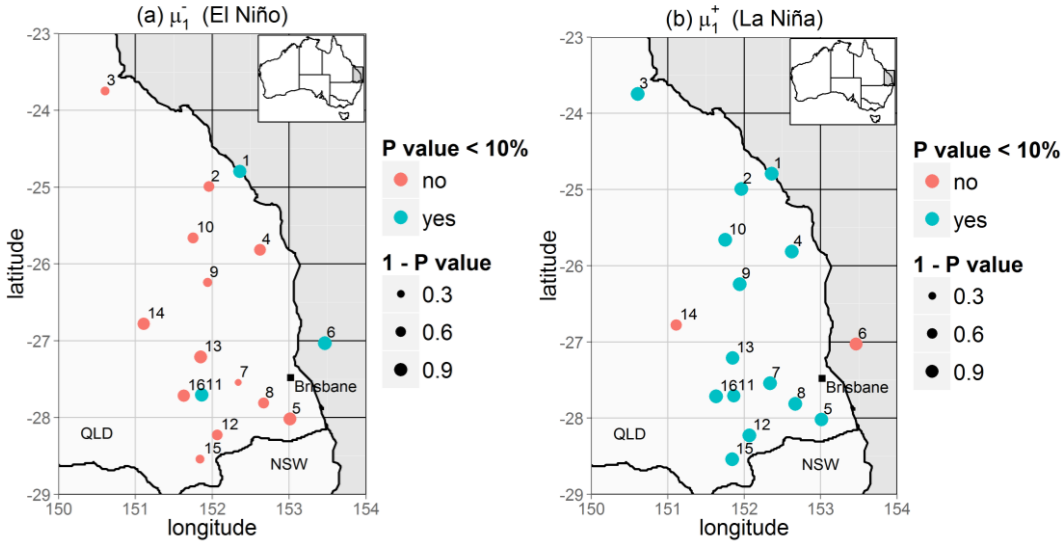


Figure 6.6-P value of zero of (a) μ_1^- (El Niño) and (b) μ_1^+ (La Niña) for each site for the summer rainfall totals. A p-value smaller than 10% (blue dots) indicates that the parameter is significantly larger than 0.

The general time-varying framework allows computing rainfall quantiles that vary with SOI values. Figure 6.7 therefore shows the evolution of several quantiles as a function of SOI. It indicates that during La Niña episodes, each unit of SOI value increases the summer rainfall by almost 5mm for the 0.5-quantile and by 10mm for the 0.99-quantile (1 in 100 year rainfall). However, during the El Niño episode, no clear trend is found.

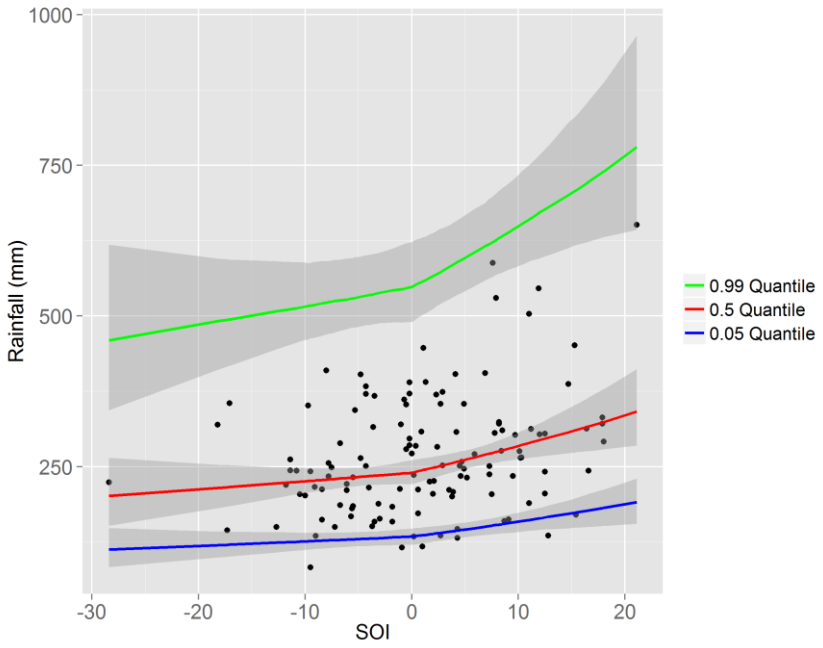


Figure 6.7-Quantiles of summer total rainfall with respect to SOI value for site 16. The blue, red and green lines are respectively the 0.05, 0.5 and 0.99 quantiles with 90% credibility intervals (grey shaded areas). Black dots are the observations with respect to the SOI value of each year.

1.4 Summary

The analysis of summer rainfall totals shows a clear impact of La Niña but no strong impact of El Niño, thereby confirming the results of *Cai et al.* [2010]. This impact can be detected even using a local model. In the remainder of this case study, we assess whether a similar relationship can be detected for extreme summer daily rainfall.

2 Quantifying the impact of ENSO on summer maximum daily rainfalls using local and regional models

We will now focus on the summer maximum daily rainfall over SEQ. *King et al.* [2013] performed a linear correlation analysis between the 5-day maximum of gridded rainfall data and the SOI, and found that the asymmetry in the ENSO-rainfall teleconnection over SEQ exists also for extreme rainfall. In this study, we focus on investigating whether the asymmetric impact of ENSO (that was evident for the summer rainfall totals) is also found for

the observed summer maximum daily rainfalls. Moreover, our objective is also to quantify the intensity of the impact in terms of high quantiles. For this study, the analysis is extended to include both local and regional models. In the case of extreme rainfall, there is considerably more uncertainty in the parameter estimates (compared with summer rainfall totals) – this uncertainty may mask the impact of ENSO. The use of a regional model to reduce parameter uncertainty and better identify the impact of ENSO impact is expected. Furthermore, comparison of different models is undertaken to answer questions such as: "Is the impact of ENSO on summer maximum daily rainfall symmetric or asymmetric?"

2.1 Data and covariates

Among the 16 high quality sites (Figure 6.1), daily rainfall data is available in 10 sites. The record starting years among these sites are ranging from 1880 to 1906. Summer maximum daily rainfall is extracted from the daily data of these 10 rain gauges. The same covariate (summer averaged SOI) as in Section 1.1 is used in this section.

2.2 Models for summer rainfall maximum

2.2.1 Local model with temporal covariates

In this study, the specific implementation of the parent distribution is a GEV model for the summer maximum daily rainfall [*Coles et al.*, 2003; *Katz et al.*, 2002]:

$$Y(t) \sim GEV(\mu(t), \sigma(t), \xi(t)) \tag{6.4}$$

To consider the ENSO impact on the location and scale of the GEV distribution, these parameters are assumed to be dependent on SOI, while the shape parameter is assumed to be constant. This is because the shape parameter ξ is difficult to estimate at a local scale even in the stationary context, hence it is more robust to assume the shape parameter to be constant.

To determine whether the asymmetric impact of ENSO found in the summer rainfall totals is also observed in the summer maximum daily rainfall, two different regression models are considered. The first one is a symmetric linear model and the other one is an asymmetric piecewise-linear model. To distinguish these two models, the asymmetric model uses the same symbols as in equations (6.2) and (6.3), and the symmetric model parameters are denoted with a tilde.

Model 1 (Symmetric linear model)

$$\tilde{\mu}(t) = \tilde{\mu}_0 + \tilde{\mu}_1 * SOI(t) \tag{6.5}$$

$$\tilde{\sigma}(t) = \tilde{\sigma}_0 + \tilde{\sigma}_1 * SOI(t) \tag{6.6}$$

$$\tilde{\xi}(t) = \tilde{\xi}_0 \tag{6.7}$$

Model 2 (Asymmetric piecewise-linear model)

$$\mu(t) = \begin{cases} \mu_0 + \mu_1^- * SOI(t); SOI(t) < 0 \\ \mu_0 + \mu_1^+ * SOI(t); SOI(t) > 0 \end{cases} \tag{6.8}$$

$$\sigma(t) = \begin{cases} \sigma_0 + \sigma_1^- * SOI(t); SOI(t) < 0 \\ \sigma_0 + \sigma_1^+ * SOI(t); SOI(t) > 0 \end{cases} \tag{6.9}$$

$$\xi(t) = \xi_0 \tag{6.10}$$

where $\boldsymbol{\theta}_{M_1} = \{\tilde{\mu}_0, \tilde{\mu}_1, \tilde{\sigma}_0, \tilde{\sigma}_1, \tilde{\xi}_0\}, \boldsymbol{\theta}_{M_2} = \{\mu_0, \mu_1^-, \mu_1^+, \sigma_0, \sigma_1^-, \sigma_1^+, \xi_0\}$ are the regression parameters of Model 1 and Model 2. Flat priors are used in this study as well.

2.2.2 Regional models

In order to better identify the parameters which quantify the impact of ENSO, regional models are applied in this case study. Following the two-step construction introduced in Chapter 4, in the first step, the time-varying structure at each site is prescribed using the same regression functions as in the equations (6.5)-(6.7) and (6.8)-(6.10). In the second step, two sets of parameters are spatialized. The first set comprises the ENSO effect parameters (e.g. for the asymmetric model $\mu_1^-, \mu_1^+, \sigma_1^-$ and σ_1^+). Indeed, climate indices, like ENSO, are expected to have similar effects on all the observation sites within a region. We also assume a regional shape parameter ξ_0. Thus we assume these parameters are the same over the region. Conversely, all other parameters are assumed purely local. The regionalized equations for the asymmetric model thus become:

$$\mu(s,t) = \begin{cases} \mu_{loc_0}^{(s)} + \mu_{reg_1}^- * SOI(t); SOI(t) < 0 \\ \mu_{loc_0}^{(s)} + \mu_{reg_1}^+ * SOI(t); SOI(t) > 0 \end{cases} \tag{6.11}$$

$$\sigma(s,t) = \begin{cases} \sigma_{loc_0}^{(s)} + \sigma_{reg_1}^- * SOI(t); SOI(t) < 0 \\ \sigma_{loc_0}^{(s)} + \sigma_{reg_1}^+ * SOI(t); SOI(t) > 0 \end{cases} \tag{6.12}$$

$$\xi(s,t) = \xi_{reg_0} \tag{6.13}$$

where $\boldsymbol{\theta}_{loc}^{(s)} = (\boldsymbol{\mu}_{loc}^{(s)}, \boldsymbol{\sigma}_{loc}^{(s)}) = ((\mu_{loc_0}^{(s_1)}, \mu_{loc_0}^{(s_2)}, ..., \mu_{loc_0}^{(s_p)}), (\sigma_{loc_0}^{(s_1)}, \sigma_{loc_0}^{(s_2)}, ..., \sigma_{loc_0}^{(s_p)}))$ are local regression parameters and $\boldsymbol{\theta}_{reg} = (\mu_{reg_1}^-, \mu_{reg_1}^+, \sigma_{reg_1}^-, \sigma_{reg_1}^+, \xi_{reg_0})$ are the regional regression parameters.

For all the models considered therein, a Gaussian copula is utilized to describe the spatial dependence. The distance-dependence relationship is characterized by the following function:

$$\Sigma(s_i, s_j) = \eta_1 * \exp(-\eta_2 * \| s_i, s_j \|) \tag{6.14}$$

where η_1 and η_2 are the dependence parameters.

2.3 Assessing competing hypotheses of ENSO impact on summer maximum daily rainfalls

In this case study, we consider three competing hypotheses, which lead to different regression models as follows:

1. There is no ENSO influence on the maximum rainfall, leading to a time-invariant model.
2. The second hypothesis is that ENSO influence is linear with respect to the GEV location and scale parameters, thus a symmetric linear model is trialed [Equations (6.5) and (6.6)].
3. The third hypothesis, motivated by the results found for the summer rainfall totals (Section 1), is that ENSO has an asymmetric impact during two different phases (El Niño and La Niña). Therefore, an asymmetric model is also used [Equations (6.11) and (6.12)].

The combination of these three regression models on the location and scale parameters of a GEV distribution gives several possible candidate models for this case study.

Furthermore, an important research question we are interested in is whether the multi-site information from regional analysis provides improved identification of the impact of ENSO. Hence, we are interested in comparing local and regional versions of the same models.

In a GEV distribution, the location parameter μ and the scale parameter σ characterize the intensity and the variability of the maximum rainfall. Thus the climate impact could be expressed through the location and scale parameters. Based on the three hypotheses above, we list the suitable GEV models by setting different regression models (Table 6.1). In Section 2.4.5, the performance of these models will be compared.

In Table 6.1, the first three are local models, and the next six are regional models. In the local models, all parameters are local. In the regional models, the parameters quantifying the effect of SOI and the shape parameter are regional. To simplify the notation, we use “L” for local and “R” for regional. The name of the models is denoted by their regression functions on the location and scale parameter (<location regression function>_<scale regression function>). *Stat* is for the identical (stationary) function. *Sym* is for the symmetric function. *Asy1* is for the asymmetric function as in equation (6.11)(6.12). *Asy2* is another asymmetric function in which the slope during negative *SOI* episode is fixed to 0 since the El Niño impact is not significant for summer total rainfall as shown in Section 1.4.

Table 6.1-Possible candidate models

Models	Regression functions for $\mu(s,t)$	Regression functions for $\sigma(s,t)$	Regression functions for $\xi(s,t)$
	Local models		
L_Stat_Stat	$\hat{\mu}_{loc}^{(s)}$	$\hat{\sigma}_{loc}^{(s)}$	$\hat{\xi}_{loc_0}^{(s)}$
LSym_LSym	$\tilde{\mu}_{loc_0}^{(s)} + \tilde{\mu}_{loc_1}^{(s)} * SOI(t)$	$\tilde{\sigma}_{loc_0}^{(s)} + \tilde{\sigma}_{loc_1}^{(s)} * SOI(t)$	$\tilde{\xi}_{loc_0}^{(s)}$
LAsy1_LAsy1	$\begin{cases} \mu_{loc_0}^{(s)} + \mu_{loc_1}^{+\ (s)} * SOI(t); SOI(t) < 0 \\ \mu_{loc_0}^{(s)} + \mu_{loc_1}^{-\ (s)} * SOI(t); SOI(t) > 0 \end{cases}$	$\begin{cases} \sigma_{loc_0}^{(s)} + \sigma_{loc_1}^{-\ (s)} * SOI(t); SOI(t) < 0 \\ \sigma_{loc_0}^{(s)} + \sigma_{loc_1}^{+\ (s)} * SOI(t); SOI(t) > 0 \end{cases}$	$\xi_{loc_0}^{(s)}$
	Regional models		
R_Stat_Stat	$\hat{\mu}_{loc}^{(s)}$	$\hat{\sigma}_{loc}^{(s)}$	$\hat{\xi}_{reg}$
RSym_RSym	$\tilde{\mu}_{loc}^{(s)} + \tilde{\mu}_{reg} * SOI(t)$	$\tilde{\sigma}_{loc}^{(s)} + \tilde{\sigma}_{reg} * SOI(t)$	$\tilde{\xi}_{reg}$
RAsy1_RAsy1	$\begin{cases} \mu_{loc_0}^{(s)} + \mu_{reg_1}^{-} * SOI(t); SOI(t) < 0 \\ \mu_{loc_0}^{(s)} + \mu_{reg_1}^{+} * SOI(t); SOI(t) > 0 \end{cases}$	$\begin{cases} \sigma_{loc_0}^{(s)} + \sigma_{reg_1}^{-} * SOI(t); SOI(t) < 0 \\ \sigma_{loc_0}^{(s)} + \sigma_{reg_1}^{+} * SOI(t); SOI(t) > 0 \end{cases}$	ξ_{reg}
RAsy2_RAsy2	$\begin{cases} \mu_{loc_0}^{(s)}; SOI(t) < 0 \\ \mu_{loc_0}^{(s)} + \mu_{reg_1}^{+} * SOI(t); SOI(t) > 0 \end{cases}$	$\begin{cases} \sigma_{loc_0}^{(s)}; SOI(t) < 0 \\ \sigma_{loc_0}^{(s)} + \sigma_{reg_1}^{+} * SOI(t); SOI(t) > 0 \end{cases}$	ξ_{reg}
RAsy1_Stat	$\begin{cases} \mu_{loc_0}^{(s)} + \mu_{reg_1}^{-} * SOI(t); SOI(t) < 0 \\ \mu_{loc_0}^{(s)} + \mu_{reg_1}^{+} * SOI(t); SOI(t) > 0 \end{cases}$	$\sigma_{loc_0}^{(s)}$	ξ_{reg}
RAsy2_Stat	$\begin{cases} \mu_{loc_0}^{(s)}; SOI(t) < 0 \\ \mu_{loc_0}^{(s)} + \mu_{reg_1}^{+} * SOI(t); SOI(t) > 0 \end{cases}$	$\sigma_{loc_0}^{(s)}$	ξ_{reg}

2.4 Results

2.4.1 Goodness-of -fit

Figure 6.8 illustrates empirical probability vs. model probability for the local (*LAsy1_LAsy1*) and regional (*RAsy1_RAsy1*) asymmetric models for all ten sites. The lines are all close to the diagonal, which indicates that both GEV local and regional asymmetric models have a good fit with the observation data.

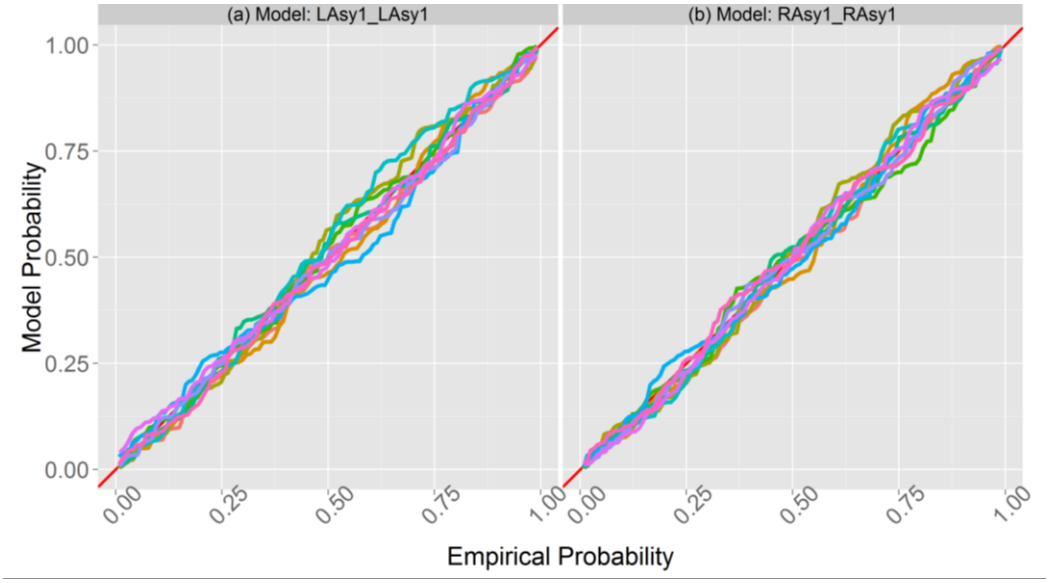


Figure 6.8-Probability-Probability plot of summer maximum daily rainfall with (a) local model LAsy1_LAsy1 and (b) regional model RAsy1_RAsy1. Each color represents one site.

2.4.2 Identifying the impact of ENSO on summer maximum daily rainfall: none, symmetric or asymmetric? Local analysis

The symmetric model (*LSym_LSym*) doesn't separate El Niño and La Niña episodes. The p-value shown in the Figure 6.9 (a) and (b) indicates that 6 out of 10 sites detect a significant ENSO impact on location or scale parameter or both. The asymmetric model (*LAsy1_LAsy1*) separates the impact of El Niño and La Niña episodes. Similar to the result of the summer total rainfall, the El Niño impact is found neither on the location nor on the scale parameter (not shown) for almost all sites. However, the La Niña impact is detected on either location or scale parameter or both (Figure 6.9 (c) (d)). The significance on the scale parameter indicates that La Niña also increases the variability of the summer maximum rainfall over the majority of sites. With both models, summer maximum rainfall is found to be affected by ENSO effect, at least during the La Niña episode.

Compared with the asymmetric model, the symmetric model has two main differences. One is the value of the slope and the other is the significance of the trend. An overview of all ten sites (Figure 6.10) indicates that 8 out of 10 sites have significant positive slope for the 1 in 100 year rainfall for the asymmetric model during the La Niña episode, and values of the slope are ranging from 4 to 10 mm/unit SOI. In comparison for the symmetric linear model, only half of the sites show a significant trend, and values of the slope are much lower ranging from 1 to 5 mm/unit SOI. From the asymmetric model, a significant trend is found during the La Niña episodes, but not during El Niño episodes, which explains why the trend analysis based on the symmetric model (which forces the same effect during El Niño and La Niña) leads to less significant results.

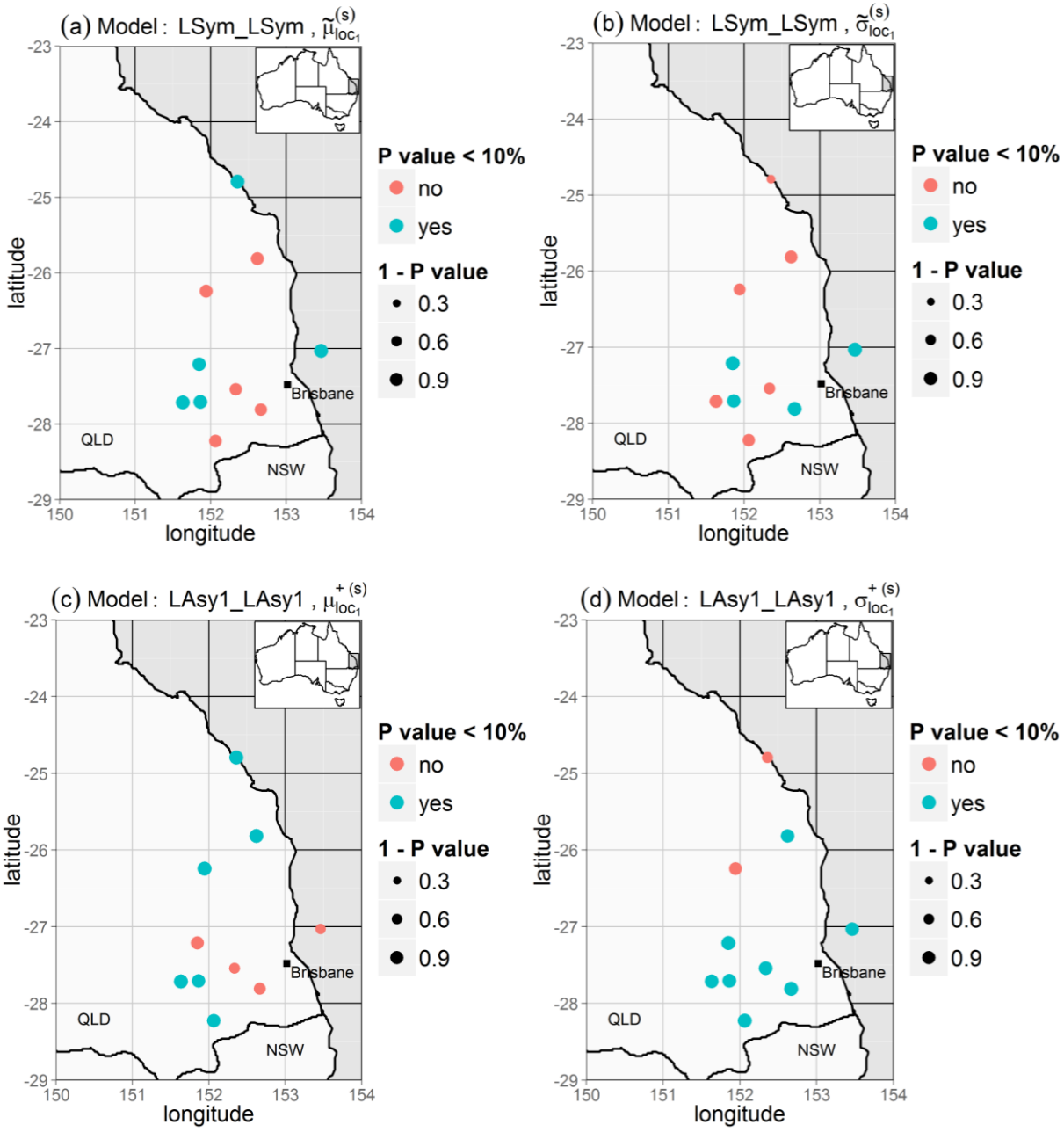


Figure 6.9-Summer maximum daily rainfall. P-value of zero of (a) $\tilde{\mu}_{loc_1}^{(s)}$ and (b) $\tilde{\sigma}_{loc_1}^{(s)}$ of each site for the symmetric model LSym_LSym, and p-value of zero of (c) $\mu_{loc_1}^{+(s)}$ and (d) $\sigma_{loc_1}^{+(s)}$ of each site (during La Niña episode) for the asymmetric model LAsy1_LAsy1. A p-value smaller than 10% (blue dots) indicates that the parameter is significantly larger than 0.

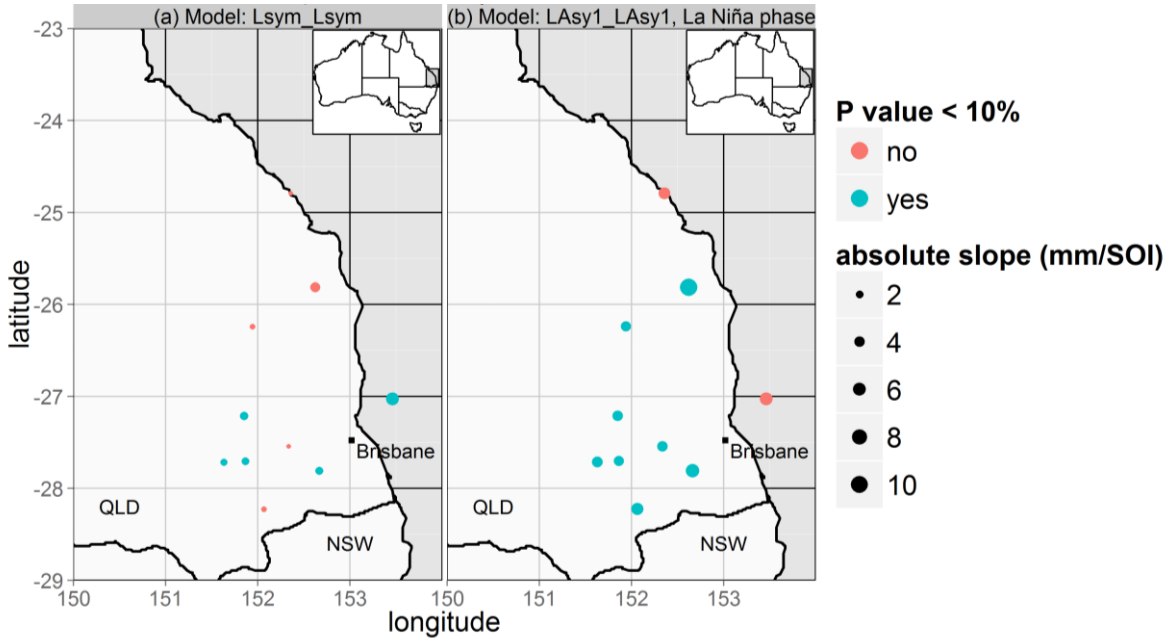


Figure 6.10-P value of zero for the slope of 1 in 100 year summer maximum daily rainfall with (a) the symmetric model LSym_LSym and (b) the asymmetric model LAsy1_LAsy1 during the La Niña episode.

2.4.3 ENSO-conditional predictions for summer maximum extreme rainfall: local analysis

Figure 6.11 illustrates the relationship between the 1 in 100 year rainfall (0.99-quantile) and the SOI index for site 16. The large slope of the asymmetric model (red) indicates that, for the positive SOI, each incremental unit increase in the SOI value will increase the 1 in 100 year rainfall by nearly 5mm, whereas the negative SOI doesn't have a statistically significant trend. Figure 6.11 also illustrates that these estimations are affected by very large uncertainties. During a strong La Niña (e.g. SOI = 20), the asymmetric model estimates that the posterior median of the 1 in 100 year rainfall is almost 25% higher than with the symmetric model and 45% higher than with a stationary model (Figure 6.11). Over all sites (not shown here), these two values can be up to 33 % and 50% respectively.

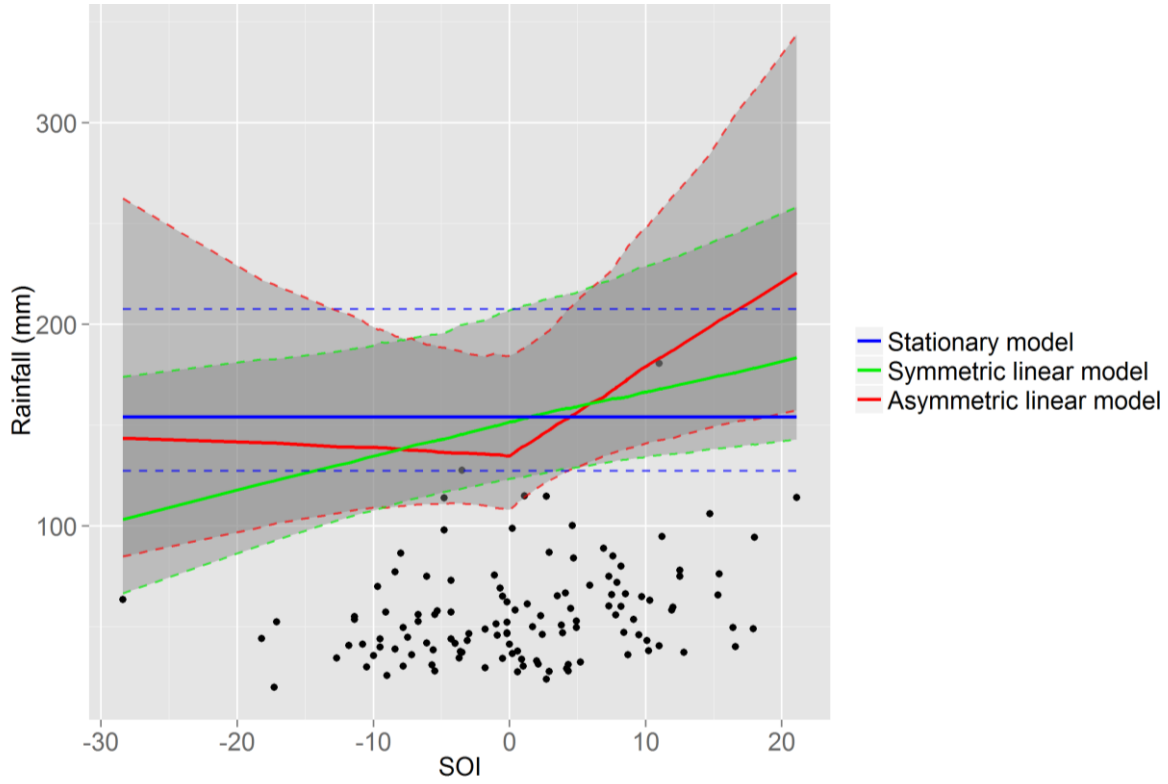


Figure 6.11-1 in 100 year summer maximum daily rainfall at site 16. The blue line is based on the stationary model (L_Stat_Stat). The green and red lines are respectively based on the symmetric (LSym_LSym) and asymmetric (LAsy1_LAsy1) models. The solid lines are median and areas inside the dashed line are 90% credibility intervals of each model. Black dots are the observations with respect to the SOI value of each year.

Although the asymmetric model detects a significant ENSO effect during La Niña, the ENSO-conditional predictions are affected by large uncertainties. This is due to the difficulty of precisely identifying the parameters with a local analysis. The regional analysis aims to reduce parameter uncertainties, hence better quantifying the impact of El Niño and La Niña.

2.4.4 Does regional analysis improve the identification of the impact of ENSO on summer maximum daily rainfall?

Figure 6.12 gives the distributions of the La Niña effect parameters on the GEV location parameter in local (*LAsy1_LAsy1*) and regional (*RAsy1_RAsy1*) models. There is a significant reduction of the distribution width for the regional model. Figure 6.13 illustrates that, for the asymmetric model *RAsy1_RAsy1*, $\mu^{+}_{reg_1}$ and $\sigma^{+}_{reg_1}$ (associated to La Niña) are found significantly larger than 0, whereas $\mu^{-}_{reg_1}$ and $\sigma^{-}_{reg_1}$ (associated to El Niño) are not. This regional analysis gives a more robust conclusion that La Niña has a significant influence on the summer maximum daily rainfall, whereas El Niño has not. Furthermore, the reduction of uncertainty on the La Niña effect parameter ($\mu^{+}_{reg_1}$, $\sigma^{+}_{reg_1}$) and the shape parameter (ξ_{reg_0})

provides an important improvement to decrease the uncertainty on high quantiles. In Figure 6.14, the 1 in 100 year rainfall of the local and regional asymmetric models are compared. During a strong El Niño (e.g. SOI = -20), the uncertainty of the regional model (measured by the interval width) is reduced by 50% compared with the local model, and during strong La Niña, this reduction is up to 60%. This clearly shows the benefit of a regional analysis in better identifying the impact of ENSO on extreme rainfall.

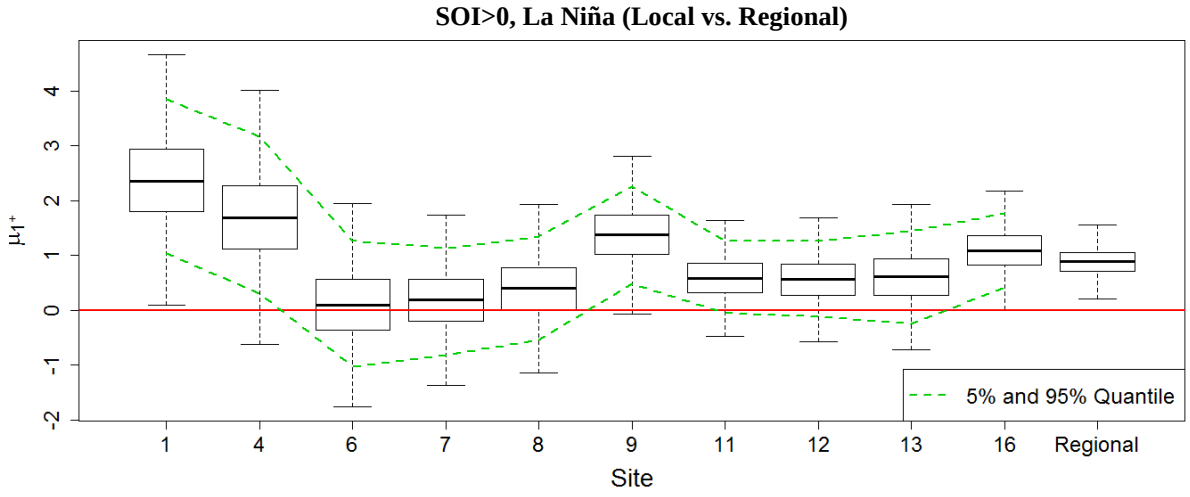


Figure 6.12-Boxplot of the posterior distribution of location parameter μ_1^+ ***(*** $\mu_{loc_1}^+$ ***in local model LAsy1_LAsy1 of each site and*** $\mu_{reg_1}^+$ ***in regional model RAsy1_RAsy1).***

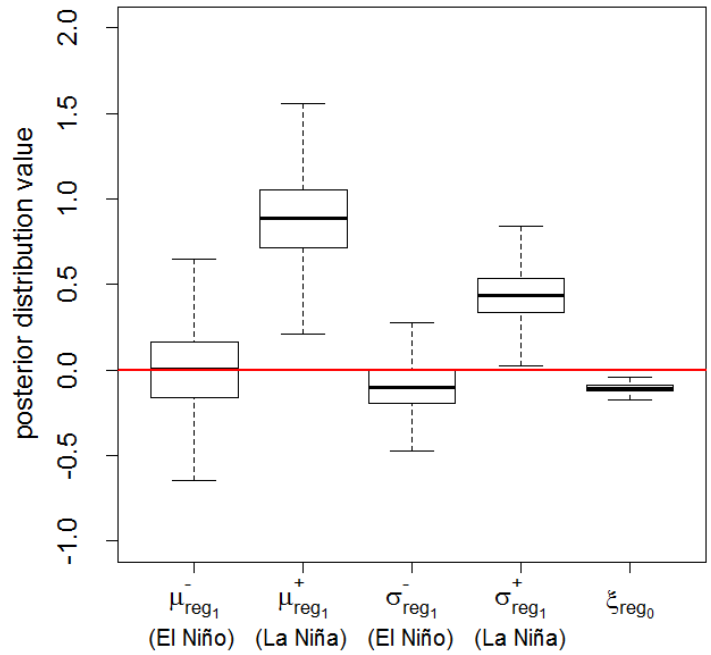


Figure 6.13-Boxplot of the posterior distribution of the regional parameters of model RAsy1_RAsy1 for the summer maximum daily rainfall

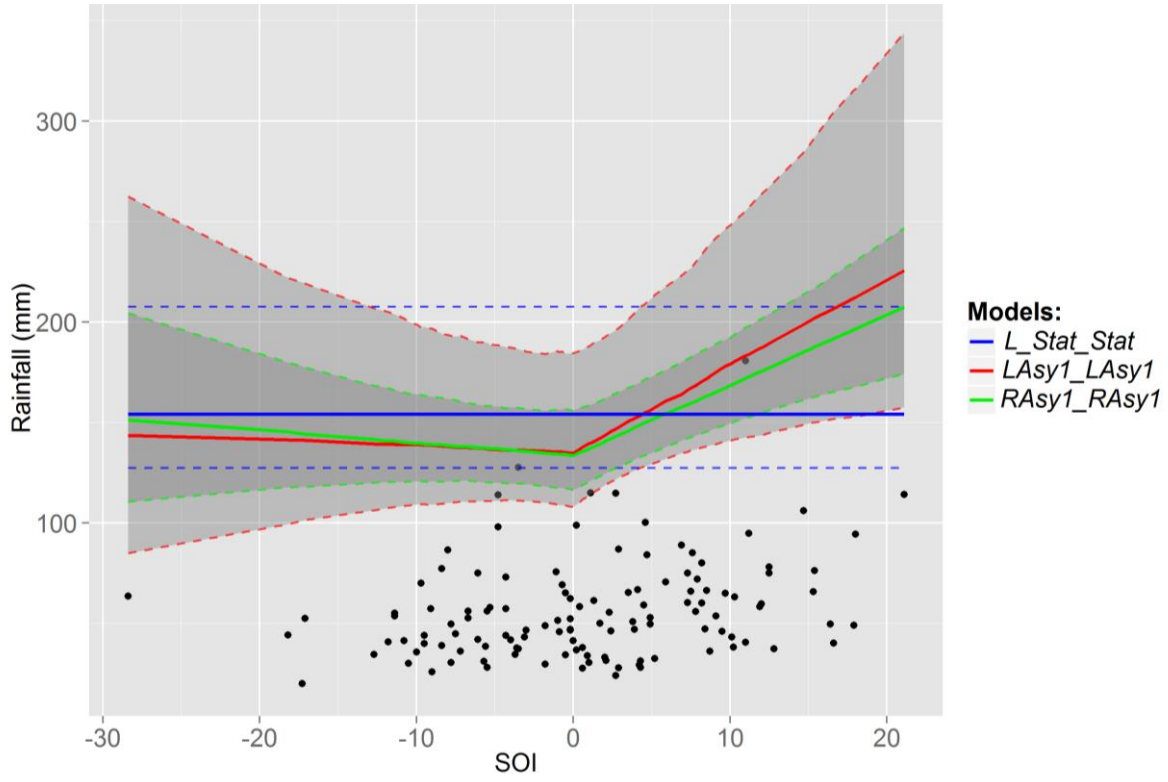


Figure 6.14-1 in 100 year summer maximum daily rainfall with local (L_Stat_Stat & LAsy1_LAsy1) and regional (RAsy1_RAsy1) models at site 16. The blue line is based on the stationary model (L_Stat_Stat). The red and green lines are respectively based on the local (LAsy1_LAsy1) and regional (RAsy1_RAsy1) models. The solid lines are median and areas inside the dashed line are 90% credibility intervals of each model. Black dots are the observations with respect to the SOI value of each year.

2.4.5 Model comparison for summer rainfall maxima

In this section, we use the DIC criterion to compare the following three pairs of models. A better model is denoted by a smaller DIC.

1. Local vs. regional modeling

Figure 6.15 illustrates the DIC values for the models in Table 6.1. The DIC values of at-site models (*L_Stat_Stat*, *Lsym_Lsym*, *LAsy1_LAsy1*) are much larger than the regional models (*R_Stat_Stat*, *RSym_RSym*, *RAsy1_RAsy1*, *RAsy2_RAsy2*, *RAsy1_Stat*, *RAsy2_Stat*). Compared with the regional models, the local models have many more parameters, which lead to a large penalty on the model complexity. Thus regional models are preferred (according to the DIC criterion) to at-site models.

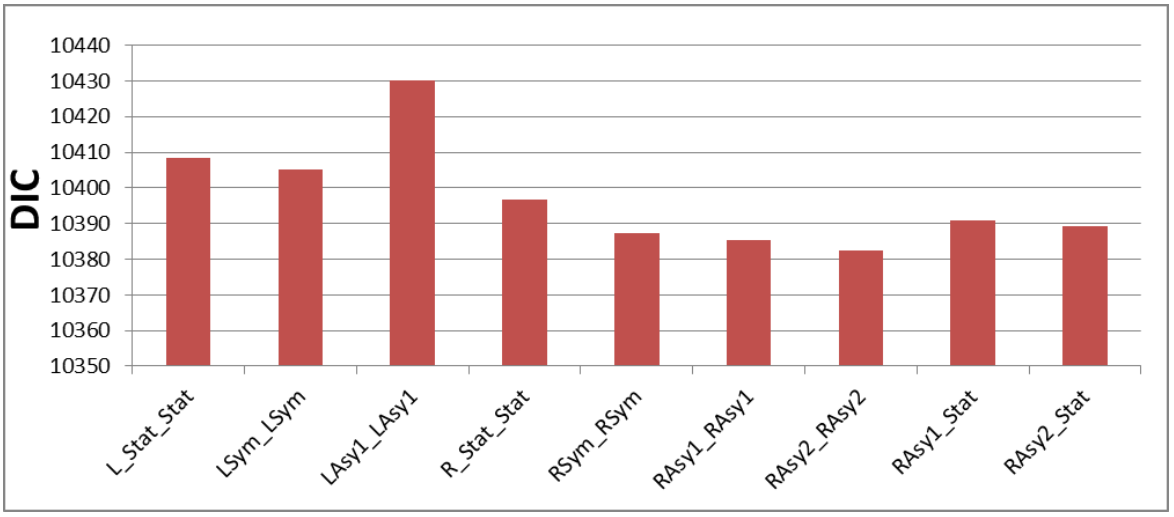


Figure 6.15-DIC value for the models in Table 6.1 for the summer maximum daily rainfall. L_Stat_Stat, LSym_LSym and LAsy1_LAsy1 are local models. R_Stat_Stat, RSym_RSym, RAsy1_RAsy1, RAsy2_RAsy2, RAsy1_Stat and RAsy2_Stat are regional models.

2. Stationary model vs. climate-informed model

According to the first point, we make this comparison with regional models only. Among the six regional models (*R_Stat_Stat*, *RSym_RSym*, *RAsy1_RAsy1*, *RAsy2_RAsy2*, *RAsy1_Stat*, *RAsy2_Stat*), the DIC value of the stationary model (*R_Stat_Stat*) is the largest (Figure 6.15). Thus, this result shows once again that ENSO influences the summer maximum rainfall over SEQ and suggests that a climate-informed model is better.

3. Symmetric vs. asymmetric effect of ENSO

The comparison between the symmetric and asymmetric models is established with the regional models listed in Table 6.1. Table 6.2 summarizes the DIC difference between the regional models in the list and the preferred model (*RAsy2_RAsy2*) with the smallest DIC. This preferred model has asymmetric regressions on both location and scale parameters. The difference between *RAsy1_RAsy1* and *RAsy2_RAsy2* is small, indicating that not inferring the slope of ENSO has little impact on DIC values. The difference between the remaining regional models and *RAsy2_RAsy2* is larger, which suggests that the models with asymmetric ENSO impact are preferred. In particular, model *R_Stat_Stat* (no ENSO effect) is strongly discredited according to the DIC. Lastly, these results also suggest that modeling a trend on the scale parameter is preferable, since models *RAsy1_Stat* and *RAsy2_Stat* have a lesser performance than the reference model *RAsy1_RAsy1* and *RAsy2_RAsy2*.

Table 6.2-DIC difference between the regional models listed on the table and RAsy2_RAsy2 model

R_Stat_Stat	RSym_RSym	RAsy1_RAsy1	RAsy2_RAsy2	RAsy1_Stat	RAsy2_Stat
14.0	4.6	2.9	0	8.4	6.7

2.5 Summary

We use both at-site and regional models to analyze ENSO effects on the summer rainfall maximum over SEQ. The link between ENSO and summer maximum daily rainfall is strong during La Niña phase and weak during El Niño phase. We demonstrate that using a regional model helps to reduce the uncertainty and provides more robust results. With the DIC criterion, competing models are compared. It is found that the asymmetric regression on both location and scale parameters is the preferred representation of ENSO effect on summer maximum daily rainfall.

3 Discussion

This chapter discusses key assumptions and current limitations of the modeling framework, and their consequences on the SEQ case study. It also proposes avenues for future improvements.

3.1 Assumption of homogeneous regions

An assumption of the regional model is that all data should be subject to similar climate impacts. This raises the question of defining such climatically homogenous regions. *Ouarda et al.* [2001] described some approaches to determine homogeneous hydrologic regions. Some Southeast Australian basins have also been classified into homogeneous regions by *Bates et al.* [1998]. The SEQ is a relatively small area, thus SEQ is assumed to be inside a same climatic homogenous region. However, when studying larger areas, the classification of different homogeneous regions will play an important role.

3.2 Spatial dependence modeling

The reason for using simple copulas, like Gaussian and Student copulas, is that they are applicable to any marginal distribution, which is convenient in the context of the general framework proposed in this thesis. Moreover the parameterization by a dependence matrix enables using geostatistical-like models (dependence is a function of distance). However, different copulas have different asymptotic behavior: asymptotically dependent (e.g. Student copula) and asymptotically independent (e.g. Gaussian copula). The extrapolation of copula is risky because the asymptotic dependence properties exert a strong leverage on joint probability of exceedance, but the limited sample size is not enough to identify such asymptotic properties. Therefore, to have a good decision between asymptotic dependent and independent copulas, more physical knowledge on spatial extent of rainfall or meteorological events is required. Further work could be therefore interested in the difference of using different copulas.

3.3 Spatial regression modeling

Inside a homogenous region, the distribution of the rainfall may depend on the spatial information at each site. For example the ENSO effect could vary with elevation or distance to sea. However, in the case study, we simply assume the same ENSO effect and shape parameter for all sites. Spatial effects could be investigated in the case study in several aspects. First, some parameters are purely local, which prevents transferring quantile estimates to ungauged sites. This could be improved by spatializing these parameters using a spatial regression. Moreover, a more flexible model could be considered by allowing spatial variations in purely regional parameters (ENSO effects and shape parameter). This was not attempted in this case study because identifying such spatial effects is difficult with only ten sites: we therefore favored the identification of ENSO effects. However, future case studies based on a spatially denser dataset will investigate in more depth the construction of such spatial models.

3.4 Practical Implications: utilizing predictions of extreme rainfall distributions from the climate-informed framework

One of the advantages of using a fully probabilistic model for extremes (as opposed to a simple linear regression between SOI and rainfall, as undertaken in *Cai et al.* [2010] and *King et al.* [2013]) is that it enables the prediction of frequency of extreme rainfall conditioned on climate variability indices. Figure 6.7, Figure 6.11 and Figure 6.14 all provide prediction of the 1 in 100 year rainfall conditional on values of *SOI*. For the preferred regional model (*RAsy1_RAsy1*), during strong La Niña phases (with high SOI), the 1 in 100 summer maximum daily rainfall is to 33% higher than the corresponding estimate obtained with the stationary model (e.g. Figure 6.14, with SOI = 20).

From an operational perspective, the knowledge that summer maximum daily rainfall is 33% higher during a strong La Niña could provide useful information for planners, engineers, water resource managers, emergency response organizations, in order to design operational/response strategies to mitigate the potential impact due to the increased risk of extreme rainfall. Consider the recent example of the summer of 2010-2011, when there was a strong La Niña (SOI = 27.1, in December) and a series of floods hit Queensland, which impacted on more than 70 towns and 200,000 people. The damage bill was over 5 billion $AUD (page 4, [*Operation Queenslander: the State Community, Economic and Environmental Recovery and Reconstruction Plan, 2011-2013*]). One of the major impact was a major flood in the city of Brisbane (a major Australian city with a population of 2.15 million), caused by the release of water from the major Wivenhoe dam upstream of Brisbane (see Chapter 16, [*Queensland Floods Commission of Inquiry*]). Armed with this knowledge of the impact of ENSO on extreme rainfall, planners/engineers/water resource managers, would be able to undertake better planning of emergency response, and potentially improve reservoir operating rules to better control floods, and reduce the impact of extreme rainfall during

strong La Niña's. On the other side of the hydrologic spectrum, climate-informed frameworks have the same importance for predicting extreme droughts [*Henley et al.*, 2011].

From a design perspective, the unconditional marginal distribution of extreme rainfall would be needed (e.g. for designing a dam or other hydraulic structures). Evaluating the marginal probabilities involves integrating out the *SOI*. This requires determining the distribution of *SOI*. Historical information could be used to inform this distribution, or alternatively predictions of the future variations in SOI from climate changes models could also be used. This climate-informed framework which provides a quantitative link between climate variability and rainfall provides far more useful information than that derived from a stationary model. Comparison of the extreme rainfall risk from a stationary model to the ones obtained by integrating out SOI in a climate-informed model is an important question that will be investigated in future work.

4 Conclusions

In this chapter, we describe the usage of the general spatio-temporal regional frequency analysis framework developed in Chapter 2 and Chapter 4, geared towards detecting and quantifying the effect of climate variability on hydrological variables. This is undertaken by using temporal regression models where the parameters of the probability distribution of hydrological events are a function of climate drivers (here, ENSO). A flexible framework is adopted, which allows testing different temporal regression functions to describe the impact of climate variability.

The first case study with the dataset of summer rainfall totals in Southeast Queensland shows that La Niña exerts a significant influence in the region for summer rainfall totals, while the impact of El Niño is not significant.

In the second case study of summer daily rainfall maxima, the flexible framework enables comparing numerous models to incorporate the impact of ENSO on extreme rainfall over SEQ. Stationary, symmetric and asymmetric models in both local and regional setups are compared using a model selection criterion (in this case, the DIC). Overall, the use of regional models yielded better identification of the impact of ENSO on extreme rainfall over SEQ compared with using only local models, for which there was too much uncertainty to enable clear identification. A variety of regional models, with different representations of the impact of ENSO (linear symmetric versus asymmetric) were also compared. Asymmetric models are found to be the best among them. More precisely, it is found that an asymmetric model, distinguishing between ENSO effect on location and scale parameters during the positive and negative phases of the SOI, is the most suitable in this case. These results corroborate the findings of other recent studies (*Cai et al.* [2010] and [*King et al.*, 2013]).

From a practical perspective, it was found that during a strong La Niña the 1 in 100 year rainfall for different sites can be 20% to 50% higher than estimates using a stationary model which ignores the influence of ENSO. This information has the potential to be used by engineers/planners to provide better informed response strategies.

CHAPTER 7 A global analysis of the asymmetric impact of ENSO on extreme precipitation

The El-Niño Southern Oscillation (ENSO) exerts a significant influence on average and extreme precipitation all over the world. In this study, a new database of monthly maxima of daily precipitation from 11,588 high quality observation sites was used to analyze the global impact of ENSO on extreme precipitation. Data were retrieved from selected regions identified after marking a 5^{o} latitude by 5^{o} longitude grid on a world map. For each season and region, a regional climate-informed statistical model was applied, in which the Southern Oscillation Index (SOI) was used as a covariate. The results of the study (i) quantify and describe the spatial pattern of the impact of ENSO on extreme precipitation; (ii) reveal the extent to which ENSO exhibits asymmetric impacts between El Niño and La Niña phases; and (iii) describe the seasons in which ENSO exerts the strongest impact.

1 Data and method

1.1 Data

The Hadley Center Global Climate Extremes Index 2 (HadEX2) dataset [*Donat et al.*, 2013] records the monthly maxima of daily precipitation from 11,588 high quality observation sites (Figure 7.1). Between 40 and 135 years of data recorded from approximately 7000 of these sites were used for this study. The median amount of data was about 60 years, and it covered most of the land on the globe. For each site, we calculated the seasonal maximum of daily precipitation for December-January-February (DJF), March-April-May (MAM), June-July-August (JJA) and September-October-November (SON). The precipitation data recorded by HadEX2 provides very good information about rainfall in Europe, the United States and South Africa, and correct information for India, China and some specific places. However, the coverage for central South America (Amazon), most of Africa, central Asia, tropical areas of Asia and central Australia is relatively low.

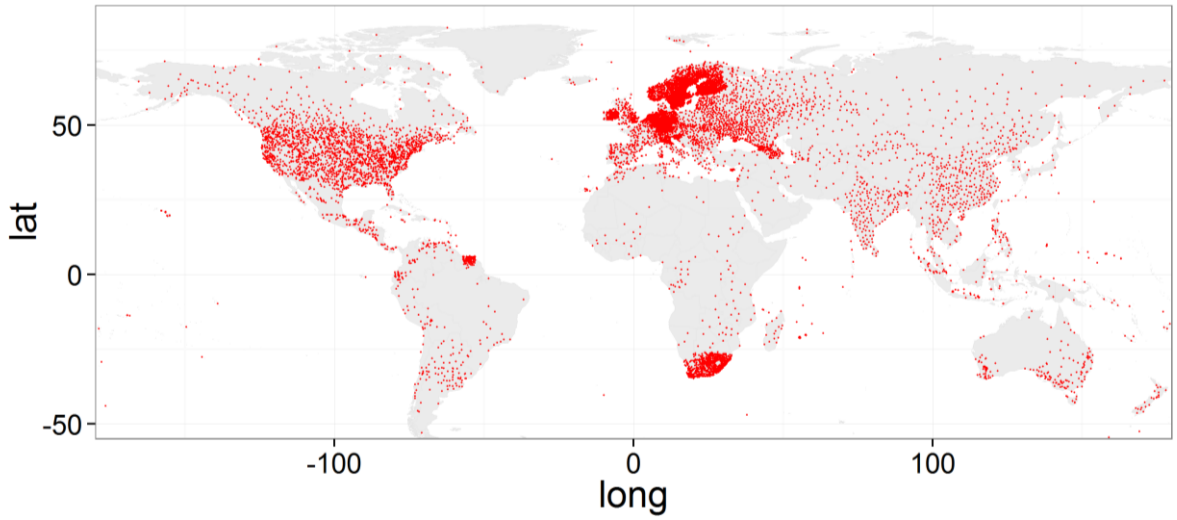


Figure 7.1-Location of high quality observation sites, with data available for 40 years or more.

As in Chapter 6, we used the Southern Oscillation Index (SOI) as a measure of ENSO. Seasonally-averaged SOI values were used in this study as covariates to explain the temporal variability of extreme precipitation.

1.2 A regional extreme value model

1.2.1 Probabilistic regional model

In a given region, $Y(s,t)$ denotes the observed seasonal maximum at site *s* and time *t*. A *GEV* distribution is assumed for all sites [*Coles et al.*, 2003; *Katz et al.*, 2002] with D-parameters varying in both space and time:

$$Y(s,t) \sim GEV(\mu(s,t), \sigma(s,t), \xi(s,t)) \tag{7.1}$$

A piecewise linear regression function for the location and scale parameters is used in order to separately evaluate the impact of ENSO during the El Niño and La Niña phases. In Chapter 6, we showed that using this function for both location and scale parameters could provide a better assessment of the impact of ENSO on extreme precipitation.

Reasoning that ENSO, as a global mode of variability, would have a fairly uniform impact across a homogeneous region, we assumed that some parameters of these regressions would also be identical for all sites within a region, including the parameters describing the effect of the SOI. In addition, we also assumed that the shape parameter was time-invariant and regional. The advantage of using regional parameters in terms of uncertainty reduction has been demonstrated in Chapter 6 (Section 2.4.4), as well as in previous studies (e.g., *Renard et al.* [2008]; *Hanel et al.* [2009a]; *Sun et al.* [2013]).

The regression function for each D-parameter is therefore given by:

$$\mu(s,t) = \begin{cases} \mu_{loc_0}^{(s)} + \mu_{reg_1}^{-} * SOI(t); SOI(t) < 0 \\ \mu_{loc_0}^{(s)} + \mu_{reg_1}^{+} * SOI(t); SOI(t) > 0 \end{cases} \tag{7.2}$$

$$\sigma(s,t) = \begin{cases} \sigma_{loc_0}^{(s)} + \sigma_{reg_1}^{-} * SOI(t); SOI(t) < 0 \\ \sigma_{loc_0}^{(s)} + \sigma_{reg_1}^{+} * SOI(t); SOI(t) > 0 \end{cases} \tag{7.3}$$

$$\xi(s,t) = \xi_{reg} \tag{7.4}$$

where $\boldsymbol{\theta}_{loc}^{(s)} = (\mu_{loc_0}^{(s)}, \sigma_{loc_0}^{(s)})$ and $\boldsymbol{\theta}_{reg} = (\mu_{reg_1}^{-}, \mu_{reg_1}^{+}, \sigma_{reg_1}^{-}, \sigma_{reg_1}^{+}, \xi_{reg})$ are regression parameters that need to be estimated. $\boldsymbol{\theta}_{loc}^{(s)}$ are site-specific (local) parameters, while $\boldsymbol{\theta}_{reg}$ are regional parameters which are common for all sites within the region.

In this model, the spatial dependence between data inside a region is described with a Gaussian copula. More detailed discussions on the use of the Gaussian copula can be found in Chapter 5, as well as in previous studies [*Favre et al.*, 2004; *Renard and Lang*, 2007; *Renard et al.*, 2013; *Sun et al.*, 2013]. The dependence matrix $\boldsymbol{\Sigma}$ for the Gaussian copula is parameterized as follows:

$$\Sigma(s_i, s_j) = \eta_1 \exp(-\eta_2 * \| s_i, s_j \|) \tag{7.5}$$

where $\| s_i, s_j \|$ is the distance between sites s_i and s_j, and $\boldsymbol{\eta} = (\eta_1, \eta_2)$ are two parameters that need to be estimated.

All regression parameters were estimated using MCMC methods (Algorithm 5, Chapter 2) under a Bayesian framework. Flat priors were used in this study. To ensure the convergence of the MCMC sampling, we ran two chains with two different starting points for each region.

1.2.2 Defining regions

Regional analyses produce more robust results than at-site (local) analyses. However, in order to obtain the spatial pattern of ENSO impact, regions should be neither too large nor too small, because regions with small size may not contain enough data and regions with large size will not concur with the homogeneity assumption. We adopted a grid size of 5˚ by 5˚ (about 309,000 km^2 in area at the equator and 155,000 km^2 at 60˚N), leading to 2592 regions, and applied the regional model described in section 1.2.1 repeatedly to each region.

In order to keep computation time reasonable, we used only some of the available observation sites in a region (the computational bottleneck being the inversion of the dependence matrix $\boldsymbol{\Sigma}$). Selection of the sites was achieved by subdividing each region into 16 sub-regions, from which if there are available gauges, the gauge with the longest record was selected (Figure 2). Therefore, in each region, there were at most 16 sites used for regional analysis. We considered a region to possess enough data to apply the regional model if there were at least three sub-regions containing available gauges in a region. The reliability of this method of data selection is discussed in Section 3.1.

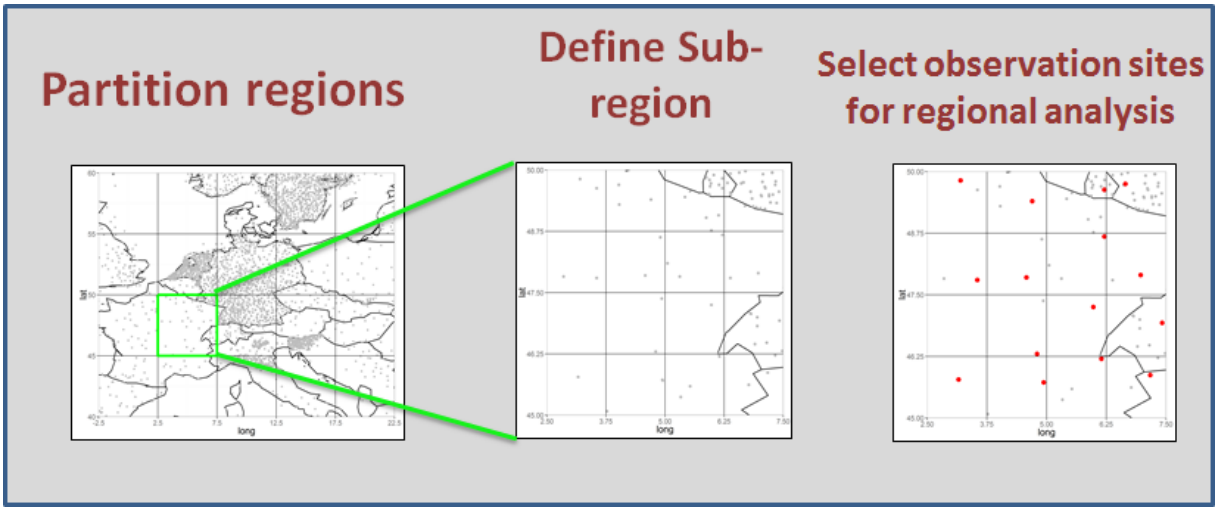


Figure 7.2-Schematic of choosing observation sites for each grid cell

1.2.3 The impact of ENSO on precipitation quantiles

The impact of ENSO on precipitation quantiles is presented through a slope value associated to SOI. More precisely, for a fixed exceedance probability 1-α, the associated quantile y_α is computed through the inverse cdf of a *GEV* distribution:

$$y_\alpha = \frac{\sigma}{\xi} K_\alpha + \mu \tag{7.6}$$

where $K_\alpha = 1 - \left(-\log(\alpha)\right)^{\xi}$.

By applying the regression functions (Equations (7.2)(7.3)(7.4)) to each *D*-parameter, conditional on a specific SOI value, the quantile y_α of site *s* becomes:

$$y_\alpha(s) = \mu_{loc_0}^{(s)} + \frac{\sigma_{loc_0}^{(s)}}{\xi_{reg}} K_\alpha + slp_\alpha * SOI \tag{7.7}$$

where $slp_\alpha = \begin{cases} \mu_{reg_1}^{-} + \frac{\sigma_{reg_1}^{-}}{\xi_{reg}} K_\alpha ; SOI < 0 \\ \mu_{reg_1}^{+} + \frac{\sigma_{reg_1}^{+}}{\xi_{reg}} K_\alpha ; SOI > 0 \end{cases}$ is the slope of the quantile with respect to SOI. Note that, as the slope slp_α only depends on the regional parameters, it is itself regional and hence does not depend on site *s*.

2 Results

In this section, we first describe the inference for the parameters of the model (Section 2.1). Then we discuss further estimates derived from the model. Section 2.2 describes the impact of ENSO on precipitation quantiles. Section 2.3 presents the asymmetry of the impact of ENSO and Section 2.4 discusses the seasonal impact of ENSO.

2.1 Regional parameter estimates

In the asymmetric model proposed in equations (7.2)(7.3) and (7.4), the impact of El Niño/La Niña is characterized by slope regression parameters $\mu_{reg_1}^{-}$ and $\sigma_{reg_1}^{-}$ for El Niño and $\mu_{reg_1}^{+}$ and $\sigma_{reg_1}^{+}$ for La Niña. If the impact is significant, the posterior distribution of $\mu_{reg_1}^{-}$ and/or $\sigma_{reg_1}^{-}$ (conversely, $\mu_{reg_1}^{+}$ and/or $\sigma_{reg_1}^{+}$) should be significantly smaller or larger than zero.

Figure 7.3 illustrates significance and intensity for $\mu_{reg_1}^{-}$ (El Niño) and $\mu_{reg_1}^{+}$ (La Niña) for each grid cell during DJF. Red (resp. Blue) contours denote significantly positive (resp. negative) values for $\mu_{reg_1}^{-}$ and $\mu_{reg_1}^{+}$ with respect to the SOI. Thus, during an El Niño phase (SOI<0), blue (resp. red) contours mean a stronger El Niño corresponding to a larger (resp. smaller) location parameter, while the relation between colour and location parameter is the opposite for the La Niña phase (SOI>0). Some cells in southern India and Africa are not convergent due to frequent zero precipitation during DJF.

During an El Niño phase (SOI <0), the location parameter is increased in southwest North America, southern South America, southeast coast of China and northern Europe.

Conversely, the location parameter is decreased in northwest North America, the Asian tropical islands and weakly in South Africa.

During a La Niña phase (SOI > 0), the location parameter is increased in northern North America, South Africa, Australia and northern Europe. Conversely, the location parameter is decreased in southern North America and northeast India.

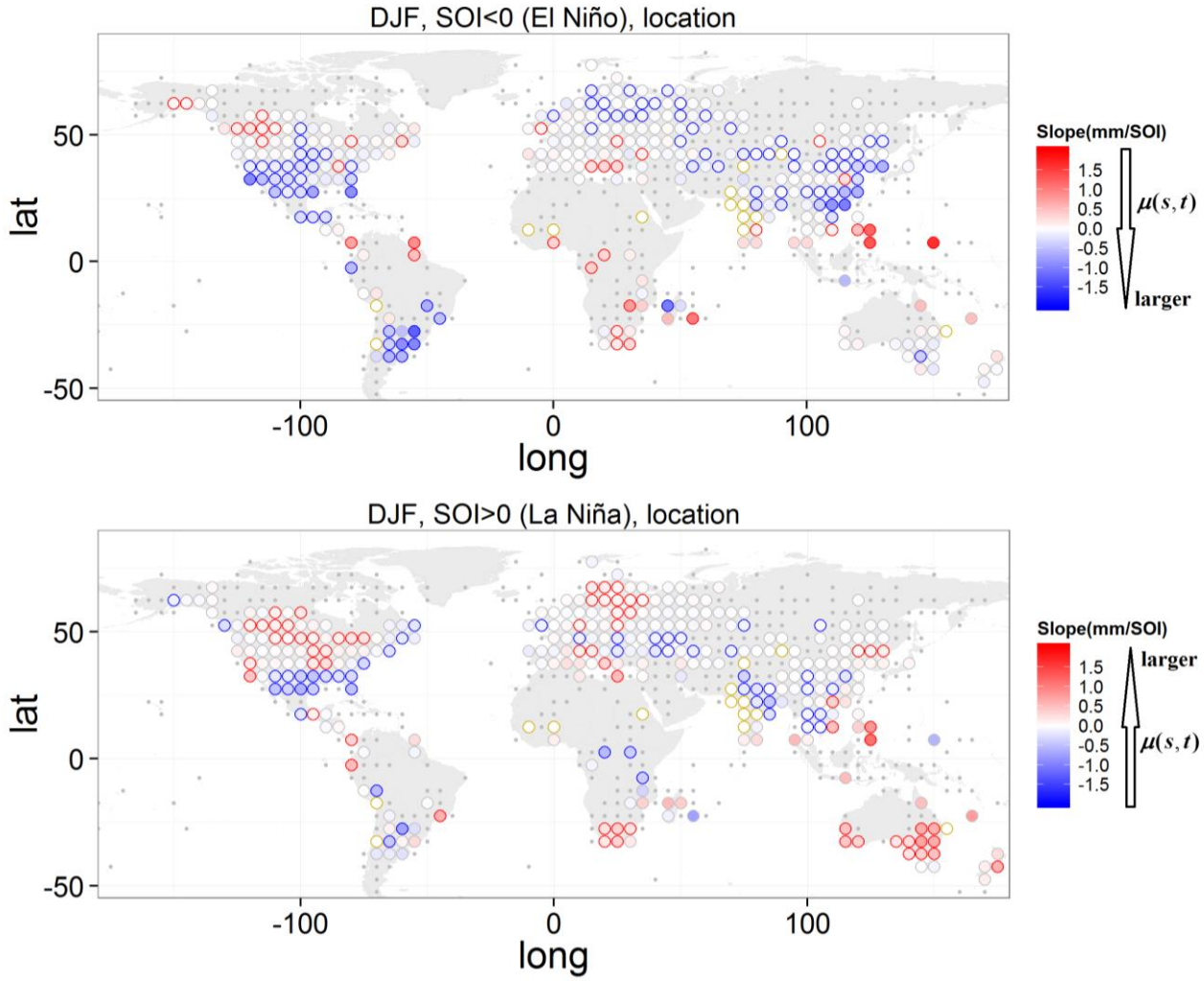


Figure 7.3-Slope of the location parameter with respect to SOI during El Niño ($\mu^{-}_{reg_1}$) and La Niña ($\mu^{+}_{reg_1}$) phases. Grey dots denote cells with too few data stations to perform a regional analysis. Dots with red (resp. blue) contours denote significantly positive (resp. negative) slopes, while dots with grey contours denote non-significant slopes. Dots with yellow contours denote cells where the MCMC algorithm did not converge, and correspond to specific locations in mountainous areas or with frequent zero precipitation during DJF.

This global pattern for both phases is consistent with the study of *Kenyon and Hegerl* [2010] for the season from November to April. However, since their study was based on at-site (local) analysis, the significance in a great number of sites was masked. The result is also

consistent with the study of ENSO-precipitation teleconnection in particular regions: North America [*Castello and Shelton*, 2004; *Cayan et al.*, 1999]; South America [*Grimm and Tedeschi*, 2009]; Eastern China [*Wu et al.*, 2003]; and Australia [*Cai et al.*, 2010; *King et al.*, 2013].

For the other seasons (see Section 5.1), the impact of ENSO is weak in MAM and JJA, and moderate in SON. In MAM, during an El Niño phase, the location parameter is higher in western North America and decreased in Central America. La Niña enhances the location parameter in the Philippines. In JJA, a significant decrease of the location parameter is found in Australia during El Niño, and a significant increase is found in south India during La Niña. In SON, El Niño strengthens the location parameter in western North America and western Mediterranean regions, while reducing the location parameter in eastern Australia and the Philippines. La Niña leads to a higher location parameter in northwestern North America, the Philippines and Australia, and reduces it in southern South America.

The impact on the scale parameter (not shown) shows a similar pattern. However, the strength is lower and less significant than for the location parameter.

To test the reliability of the method of data selection described in section 1.2.2, we ran our model on one or two regions on each continent using all available data from those regions (Figure 7.4). The results of these trials were highly consistent with the results we obtained using only a maximum of 16 selected sites from each region.

2.2 The impact of ENSO on precipitation quantiles

A more intuitive picture of the impact of ENSO on extreme precipitation can be obtained by expressing it in terms of quantiles rather than in terms of location and scale parameters. In this section, we illustrate the percentage change in 1 in 10 year precipitation during a strong El Niño or La Niña phase and a neutral phase. The 1 in 10 year precipitation (0.9 quantile) is computed from equation (7.7) with α equal to 0.1, which requires the usage of a local regression parameter. Thus the illustration of a global pattern is expressed on single sites instead of regional effects.

Figure 7.5 illustrates the percentage change for the 1 in 10 year precipitation during DJF between an extreme ENSO event (|SOI| = 20) and a neutral phase (SOI = 0). Red (blue) indicates that extreme El Niño/La Niña increases (decreases) the 1 in 10 year precipitation. During a strong El Niño event, it can be increased by more than 50% in Central America, 40% on the south coast of China, and nearly 20% in central North America and southeast South America. A decrease of about 15% can be observed in northwest America and a more than 20% decrease in the Philippines.

During a strong La Niña episode, the intensity of a 1 in 10 year precipitation increases by about 15% in northern North America, north Europe and the Mediterranean region, 10% to 40% (from east to west) in South Africa, about 20% in northeast China, more than 40% in eastern Australia and 60% in western Australia. However, the intensity of a 1 in 10 year precipitation is decreased by more than 50% in Mexico and nearly 25% in northeast India. As Figure 7.5 is a derivation of the model, the illustration of the global pattern of the impact of ENSO on 1 in 10 year precipitation events broadly coincides with that of the slope parameter

shown in Figure 7.3. The results of other seasons (see Section 5.2) correspond to the location slope parameter of that season as well.

2.3 Asymmetry of the impact of ENSO on extreme precipitations

In each phase, ENSO can have positive, negative or no impact on precipitation. Thus there are nine possible combinations of the impact of ENSO during two phases. Figure 7.6 illustrates these nine combinations between the quantiles and SOI. A very negative value of SOI corresponds to a strong El Niño, while a positive value indicates a La Niña.

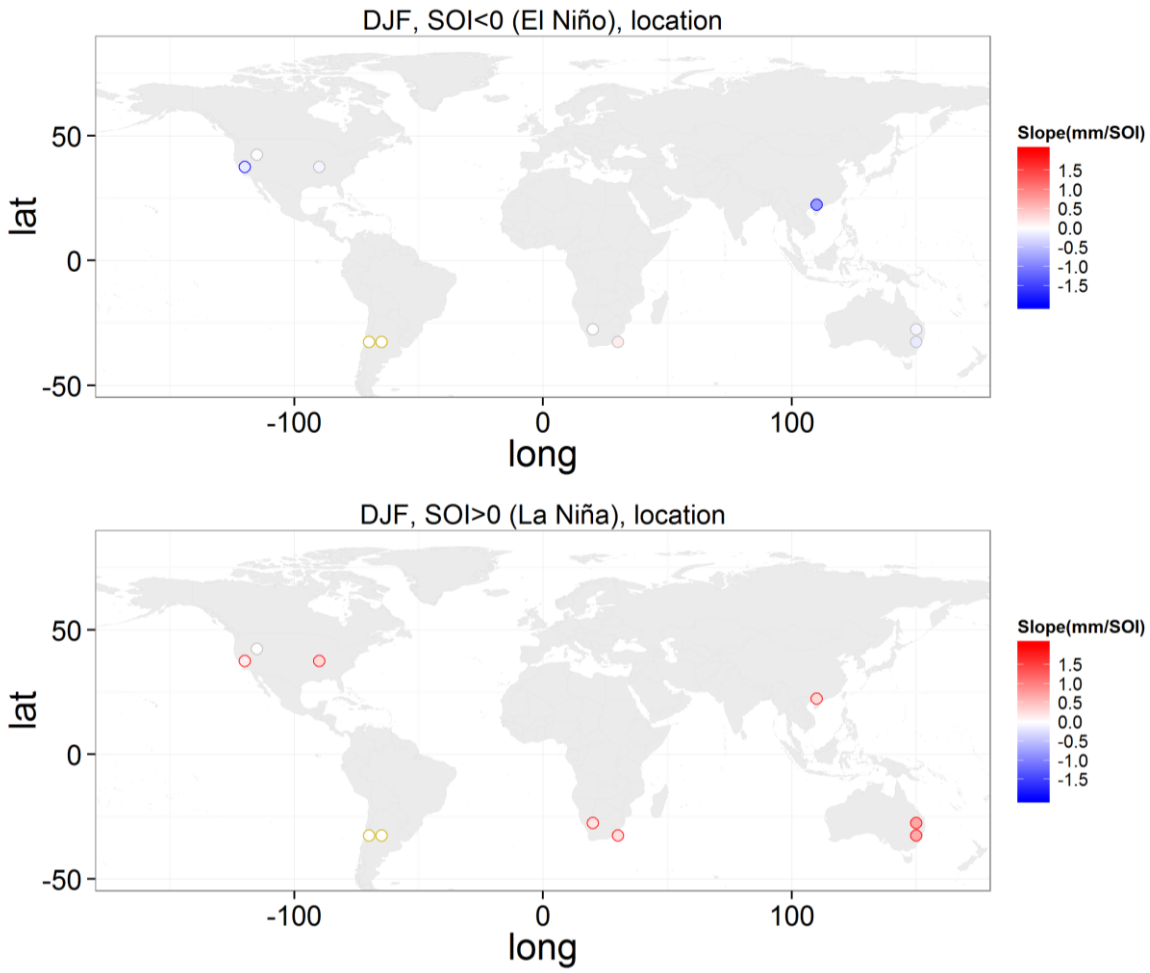


Figure 7.4-Same as Figure 7.3, but the model is run on randomly selected regions on each continent using all available data from those regions

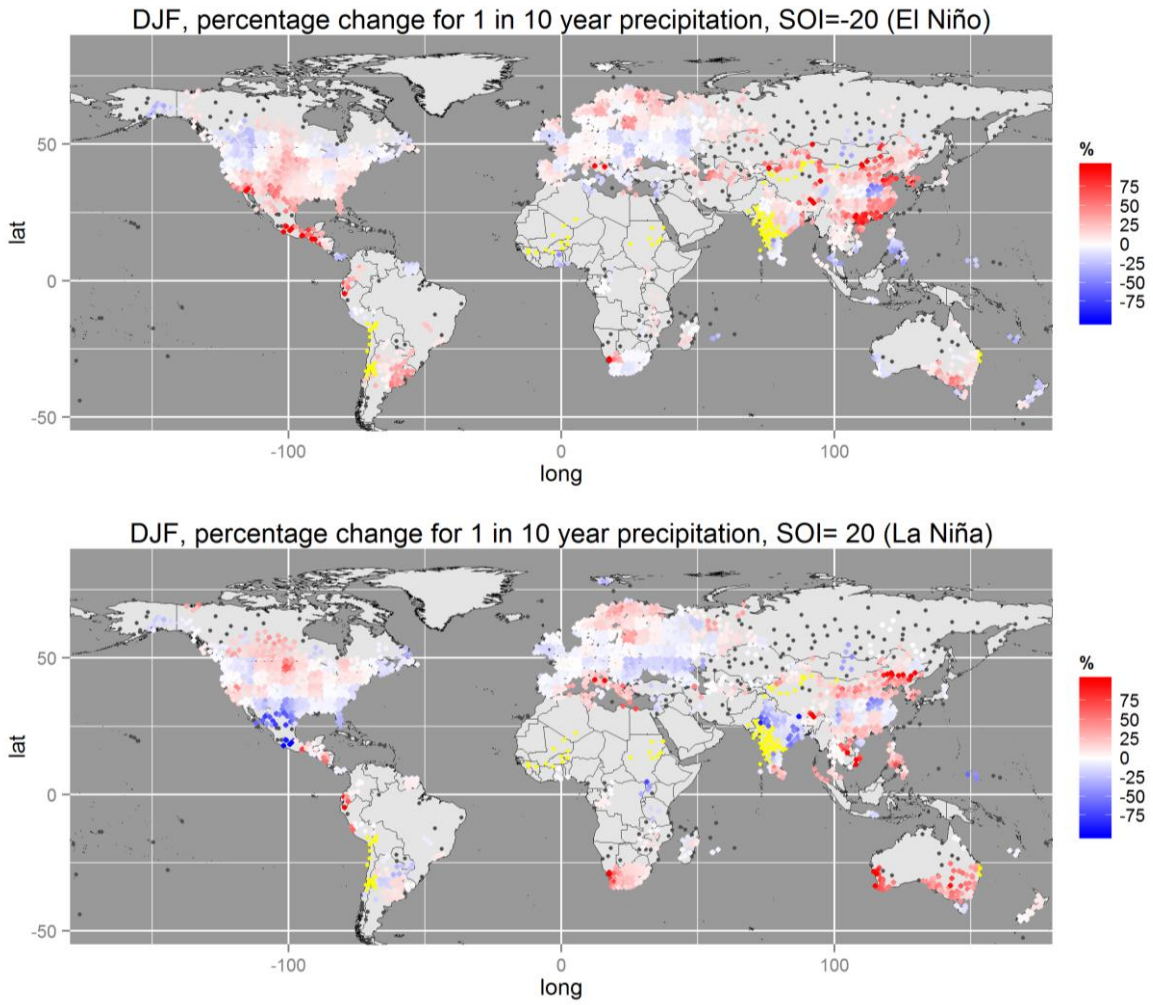


Figure 7.5-Percentage change for the intensity of 1 in 10 year precipitation relative to SOI=0. Grey dots denote cells with too little station data to perform a regional analysis. Red (resp. blue) dots denote an increase (resp. decrease) in the intensity of a 1 in 10 year precipitation for strong El Niño/La Niña phases compared with a neutral phase. Yellow dots denote cells where the MCMC algorithm did not converge. They correspond to specific locations in mountainous areas or to frequent zero precipitations during DJF.

The asymmetric behaviors could be summarized in two types: (1) no effect for one phase and one significant effect for another phase (Figure 7.6 (a), (b), (d), (e)); (2) precipitation increases or decreases during both El Niño and La Niña phases (Figure 7.6 (c), (f)). Figure 7.6 (g), (h) illustrate the symmetric behaviors and (i) presents no ENSO effect. The simplest way to detect the asymmetry is to compute the difference of slope during La Niña and El Niño phases. A positive value will be obtained for the relation of type (a), (b), (c), and a negative value for the relation of type (d), (e), (f) in Figure 7.6. Conversely, the slope differences for (g), (h) and (i) should be close to zero. Figure 7.7 illustrates the difference in the regional slope for the 1 in 10 year precipitation in DJF between the La Niña and El Niño phases. Red contours denote a significantly positive difference, while blue contours indicate the difference in slope is negative. The asymmetric type for a particular region could be obtained easily by

comparing the results of two phases in Figure 7.3. The asymmetric impact is mainly found in a part of North America, southeast South America, eastern China, Australia, and to a lesser extent, northern Europe and central Asia.

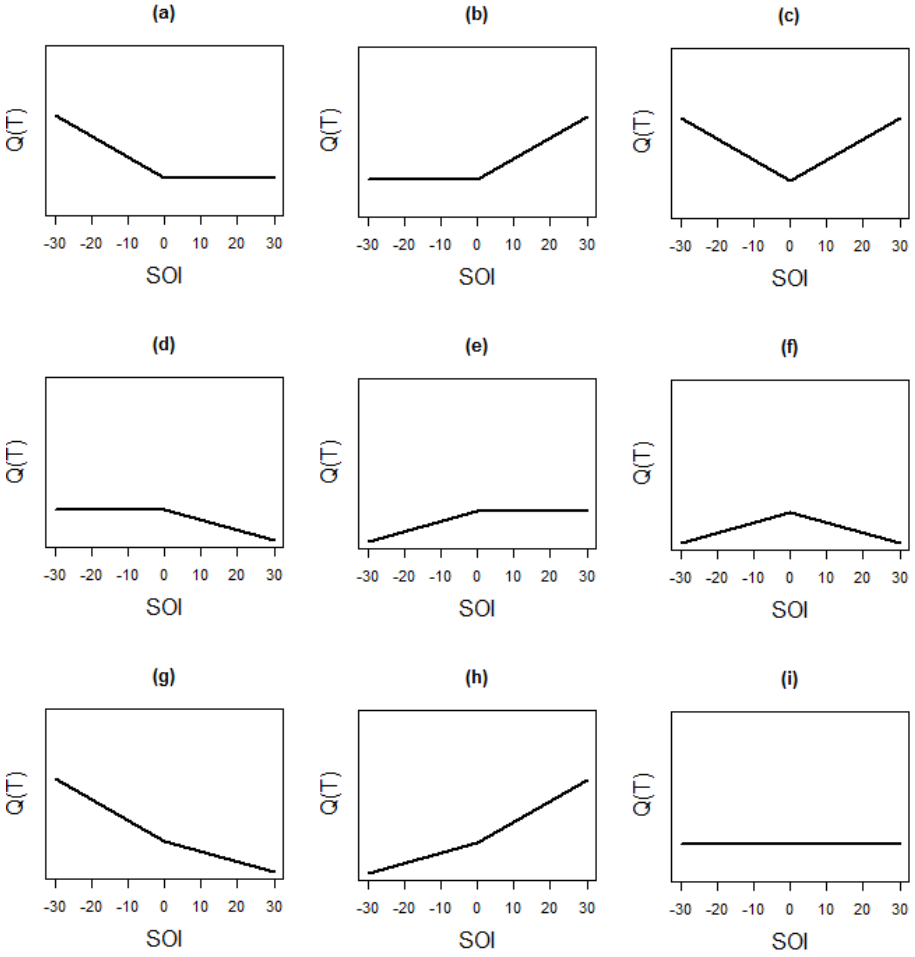


Figure 7.6-Nine possible combinations of the relation between quantile and SOI. An upward trend denotes a significantly positive slope for the SOI effect on the quantile; while a downward trend denotes a significantly negative slope. A flat line means the impact of ENSO is not significant on the quantile. For instance, (a) illustrates the case where the slope is negative during El Niño (SOI<0) and not significant during La Niña (SOI>0).

This asymmetric ENSO pattern coincides with the difference between the observed precipitation intensity during El Niño and La Niña described by *Meehl et al.* [2007, fig 3a] for North America. *Grimm and Tedeschi* [2009] pointed out that the El Niño impact on the frequency of extreme precipitation is not quite consistent with the impact of La Niña in South America. However, there is not enough data from Brazil in our analysis; thus we cannot make further verification in this area. The asymmetry found in south China during MAM (see Section 5.3) is consistent with the study of *Feng and Li* [2011]. However, we find asymmetric

impact is stronger in DJF than in MAM. The asymmetry found in Australia is also consistent with the studies of *Cai et al.* [2010] and *King et al.* [2013].

The results for other seasons (see Section 5.3) reveal that the asymmetry in eastern China and Australia is significant during MAM, and the asymmetry in western North America is more significant in SON than in DJF. There is no clear evidence that asymmetry exists in JJA, probably because JJA is the season with the lowest ENSO impact.

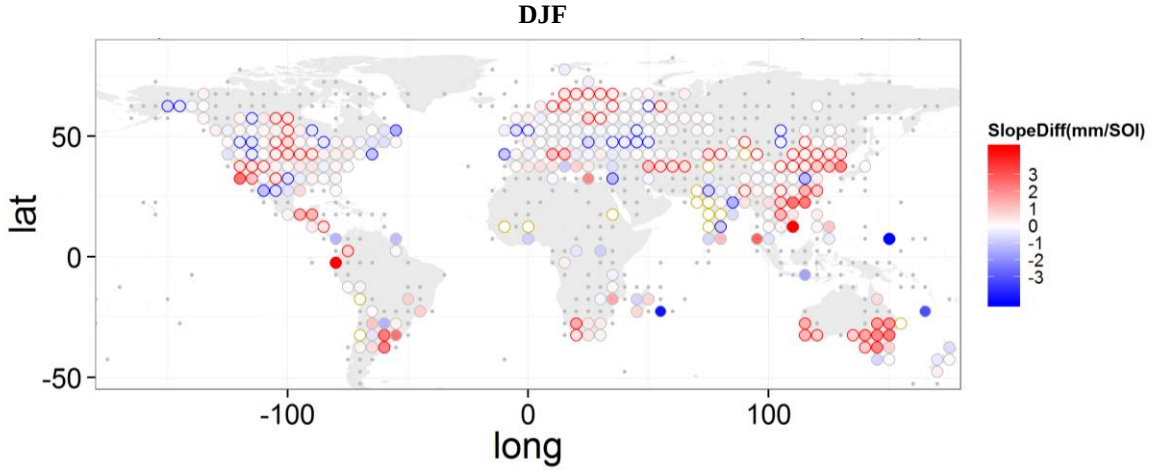


Figure 7.7-Difference between the slope of SOI during La Niña and El Niño phases (La Niña - El Niño) for 1 in 10 year precipitation. Grey dots denote cells with too little station data to perform a regional analysis. Dots with red contours denote a significant difference between the impact of the La Niña and El Niño phases, while blue contours denote a significantly negative difference. Dots with grey contours denote non-significant slope differences and dots with yellow contours denote cells where the MCMC algorithm did not converge, and correspond to specific locations in mountainous areas or with frequent zero precipitation during DJF season.

2.4 Seasonality of the impact of ENSO on extreme precipitations

Figure 7.8 illustrates the season during which ENSO exerts the greatest influence on the 1 in 10 year precipitation. Table 7.1 summarizes regions and seasons when the impact of ENSO is strongest. The data indicated that in most regions SON experiences the most marked effects of ENSO, although MAM can be equally influenced in some areas. In general, JJA is the season least affected by ENSO.

In some regions, the El Niño and La Niña phases do not exert their influence in the same season, which reinforces the asymmetric teleconnection between ENSO and extreme precipitation. In Figure 7.8, we only observe the season in which the impact is the largest. The other seasons may have also significant ENSO impacts, but they are masked in this figure.

In general, these findings are consistent with those from empirical orthogonal function (EOF) analyses for land precipitation: the ENSO signal is strongest in SON and weakest in JJA [*Dai et al.*, 1997]. However, there are some differences in particular regions. *Dai et al.* [1997] found that the ENSO signal is strongest during DJF in North America and SON in the

Australian-Indonesian region. The results of the current study in terms of the season with the strongest impact of ENSO in North America and Australia clearly differ from those of *Dai et al.* [1997]. It is likely that this difference results from different hypotheses in the two studies (symmetric vs. asymmetric ENSO impact) and different target variables (percentage of variance in the total precipitation explained by ENSO vs. intensity of the ENSO effect on extreme precipitation).

Table 7.1-Region and seasons experiencing the strongest impact of ENSO on 1 in 10 year precipitation

Regions	Strongest impact seasons	increase (+)/ decrease(-) precipitation
	El Niño episode	
Eastern part of North America	SON	-
Western part of North America	SON	+
Southern part of South America	SON, DJF	+
Southeast China	DJF	+
Eastern Australia	JJA, SON	-
	La Niña episode	
Middle north of North America	SON	+
Southern part of North America	DJF, MAM	-
Southern part of South America	SON, DJF	-
North India	SON	+
South Africa	DJF	+
Eastern part of China	MAM, JJA	+
Australia	DJF	+

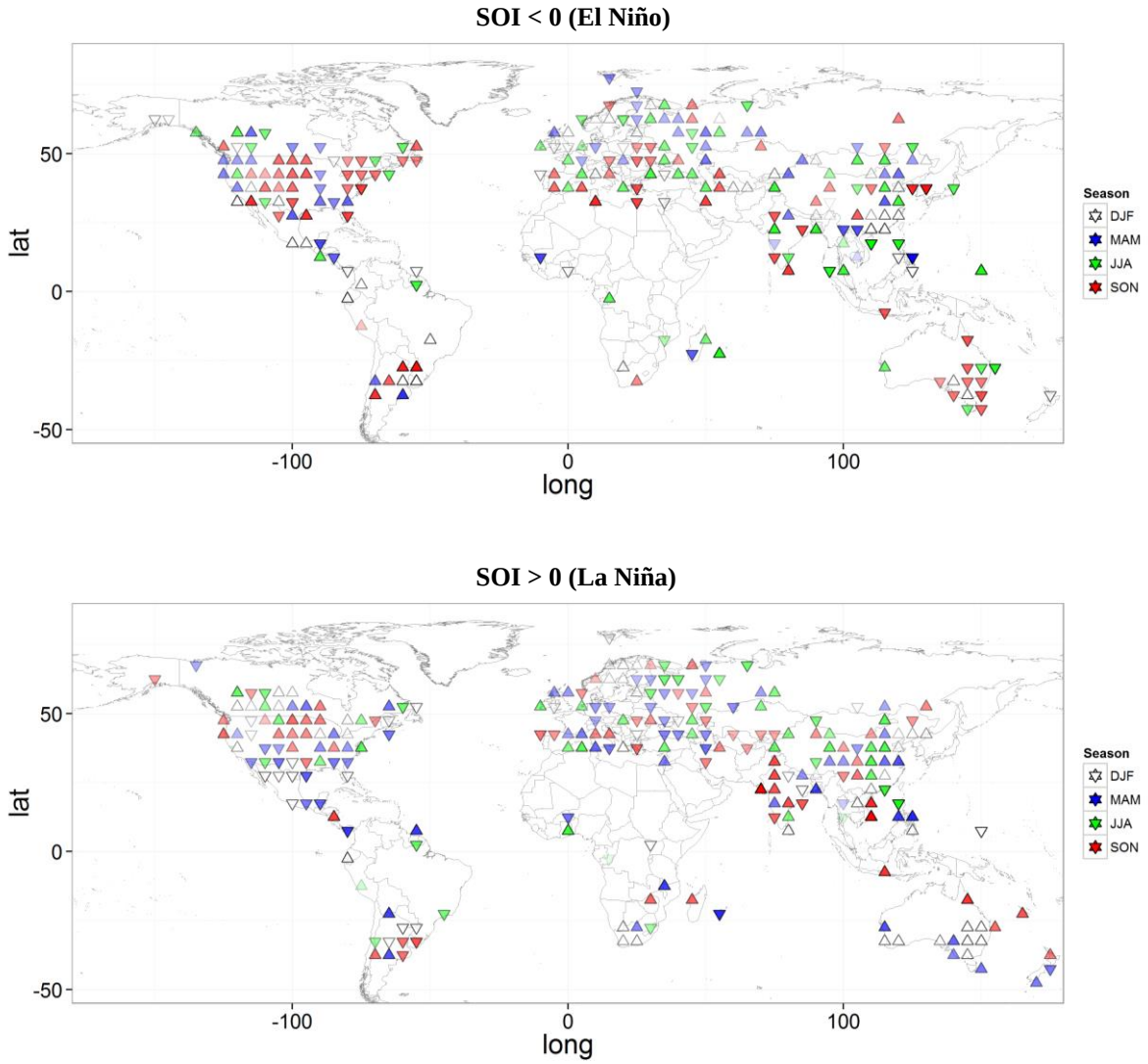


Figure 7.8-Map of the season with the largest ENSO impact. The color illustrates the season in which the ENSO effect is the strongest for the 1 in 10 year precipitation for each grid cell. Upward pointing triangles denote increases; downward pointing triangles denote decreases in extreme precipitation intensity. The color intensity is proportional to the intensity of the ENSO impact.

3 Discussion

3.1 Limitation of the model and reliability of the definition of a region

The use of regional models requires defining a homogenous region. In this study, we defined the region (grid) according to its latitude and longitude. It is important to note that for the purposes of the current study only the shape parameter and the ENSO effect parameters are regional. Other parameters are site-specific, which provides the flexibility to account for between-site differences within the region.

Experience during the research indicated that the definition of homogeneous regions might be unduly simplistic for some geographical areas. In mountainous areas, in particular, where humid air sometimes cannot get across the mountains, precipitation patterns differ on either side of the ranges. For example, because of the Andes Mountains, stations in the grid cells defined across Chile and Argentina reveal clearly different patterns. In the austral summer, precipitation is very low on the Chilean side, and much greater on the Argentinean side.

Because of this geographical influence, the MCMC sampling is not convergent for such grid cells. A possible solution is to redefine regions by subdividing the grid cells into smaller grids. However, the number of available observation sites might be too low for a meaningful regional analysis. A better outcome might be achieved if the homogeneous regions were based on climate or physiographic variables, as opposed to the simple regular grid adopted here. This will be explored in future work.

3.2 Changes in ENSO teleconnections

An important question in terms of ENSO is how the evolution of such an influential global process will affect extreme precipitation in a future climate. Several challenges need to be met to address this issue. Firstly, GCM simulations suggest the possible occurrence of "super-ENSO" events [*Latif et al.*, 2013] in the future, which may require extrapolating from the regional model used in this study well beyond the range of observed SOI values. Whether or not the "double-slope" relationship used in this model can be extrapolated to very high SOI values is an open question. Secondly, the ability of GCMs to describe the physical mechanisms governing the development of ENSO events remains limited, as discussed, for example, by *Bellenger et al.* [2013]. In the present state of GCMs, it is therefore unclear whether projections of future ENSO events can be reliably used to deduce the evolution of extreme precipitation in a future climate.

3.3 Impact of other large scale modes of climate variability

Besides ENSO, other large scale modes of climate variability could also affect regional extreme precipitation. For instance, the Indian Ocean Dipole (IOD) affects the precipitation in Southeast Asia and Australia, and the North Atlantic Oscillation (NAO) influences precipitation in North America and Europe. Research also indicates that in many regions precipitation is influenced by the combined effect of distinct large scale modes (e.g., [*Keim and Verdon-Kidd*, 2009]). For example, IOD and ENSO both influence precipitation in Australia. However, the ways in which the two modes combine to influence precipitation in Australia have not been established, meaning that there is still research to be undertaken in this area.

The current study was solely a purely statistical analysis of the impact of ENSO on extreme precipitation, and does not provide any answer to this question. If such a physical mechanism could be understood, however, a more appropriate regression model could be

integrated instead of the piecewise linear regression model to provide a better evaluation of ENSO impact.

4 Conclusions

In the current study, we quantified and described the global pattern of the impact of ENSO on extreme precipitation by applying a climate-informed regional frequency analysis framework.

It was found that for boreal winter (DJF), El Niño enhances extreme precipitation in southwest North America, southern South America, the southeast coast of China and northern Europe. Conversely, extreme precipitation is decreased in northwest North America and (more weakly) in South Africa. La Niña enhances extreme precipitation in northern North America, South Africa, Australia and northern Europe. Conversely, extreme precipitation is decreased in southern North America and northeast India. We also demonstrated how the result can also be used for predicting the impact of ENSO on the high return period precipitation.

In addition, the possible asymmetry of the impact of ENSO during its El Niño and La Niña phases was assessed. The asymmetry is highlighted in many regions. In particular, during the boreal winter, data from western North America, southeast South America, eastern China, Australia, and to a lesser extent, northern Europe and central Asia reveal an asymmetric ENSO impact.

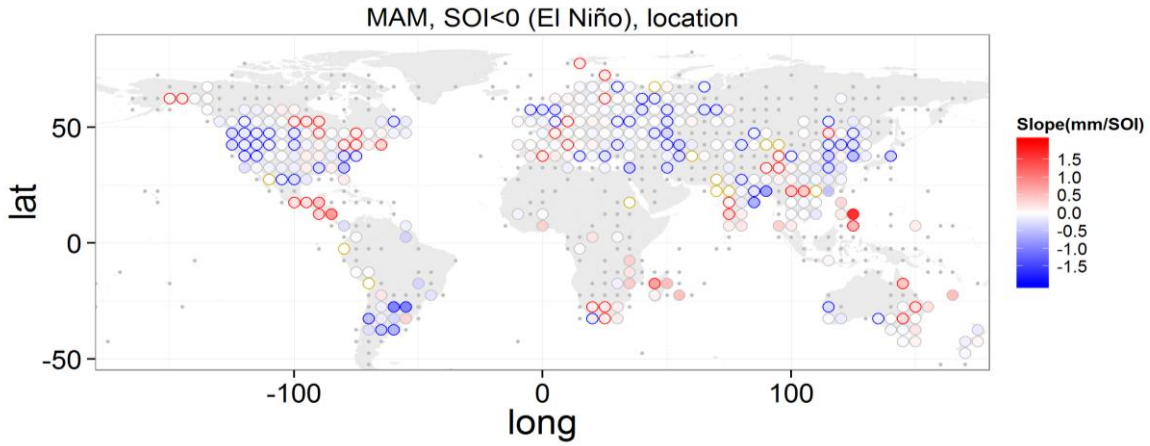
We also determined the season most affected by ENSO. It was found that during the El Niño phases, DJF is the season during which the southern part of South America and southeast China are the most affected by ENSO processes. SON is the season during which ENSO most strongly affects North America (increasing precipitation in the eastern part and decreasing it in the west), the southern part of South America and eastern Australia (decrease). During the La Niña phase, the greatest impact is during DJF in the southern part of North America (decrease), the southern part of South America (decrease), South Africa and Australia, while MAM is the season most strongly affected in the southern part of North America (decrease) and Southeast China. In the middle north of North America, the southern part of South America (decrease) and northern India, the greatest effects are felt during SON. In general, the impact of ENSO processes is weak in JJA.

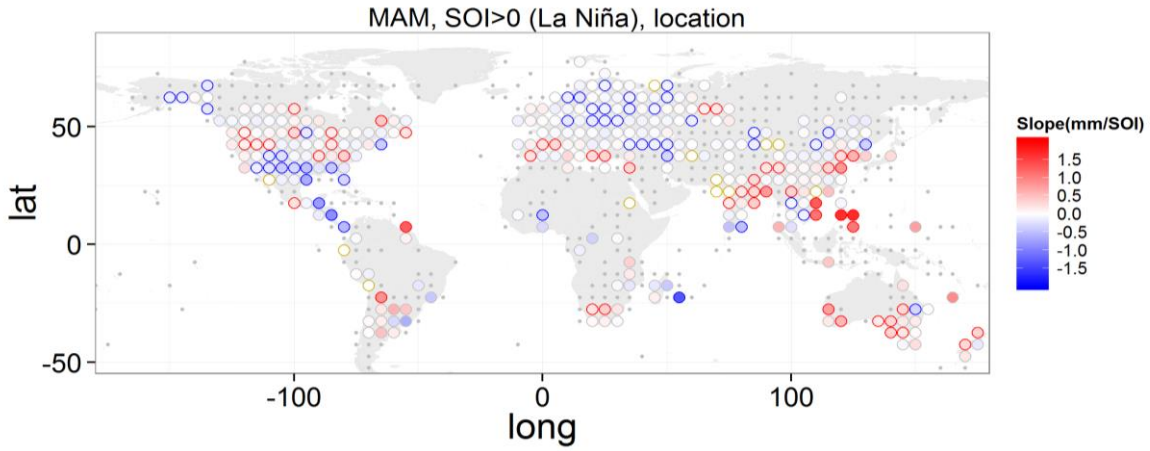
From an engineering perspective, the information provided in this study could help to forecast the seasonal floods and droughts caused by ENSO, and planners could decide in advance how much water needs to be held in dams.

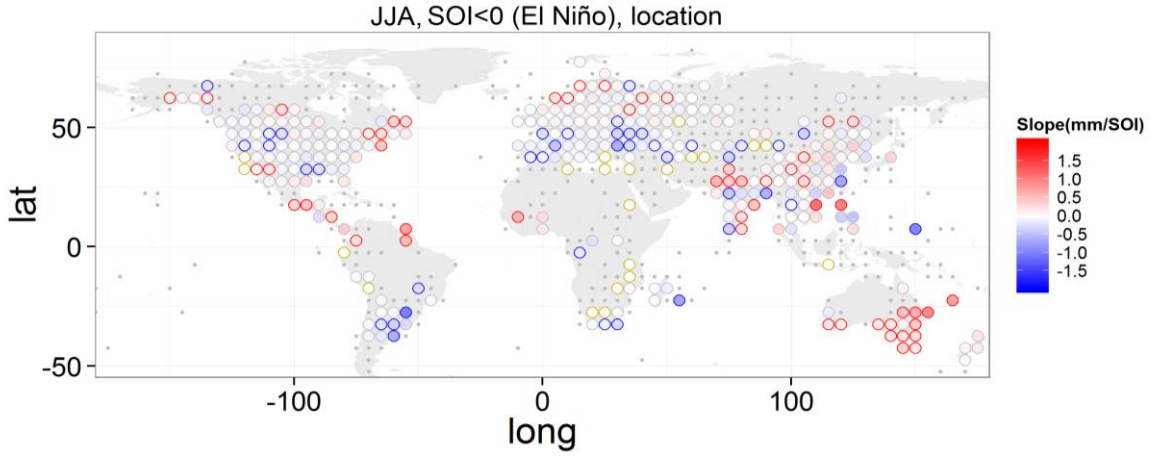
5 Figures for the other seasons

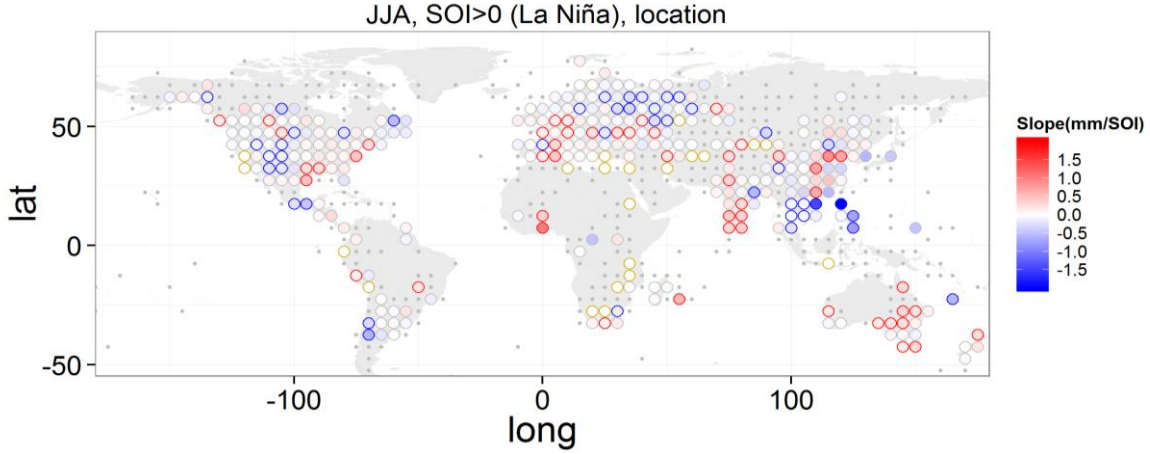
5.1 Slope of the location parameter with respect to SOI

The following figures illustrate the slope of the location parameter with respect to SOI for the other three seasons (MAM, JJA and SON). Captions are the same as Figure 7.3.


MAM, SOI<0 (El Niño), location
lat
50
0
-50
-100
0
100
long
Slope(mm/SOI)
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5


MAM, SOI>0 (La Niña), location
lat
50
0
-50
-100
0
100
long
Slope(mm/SOI)
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5


JJA, SOI<0 (El Niño), location
lat
50
0
-50
-100
0
100
long
Slope(mm/SOI)
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5


JJA, SOI>0 (La Niña), location
lat
50
0
-50
-100
0
100
long
Slope(mm/SOI)
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5

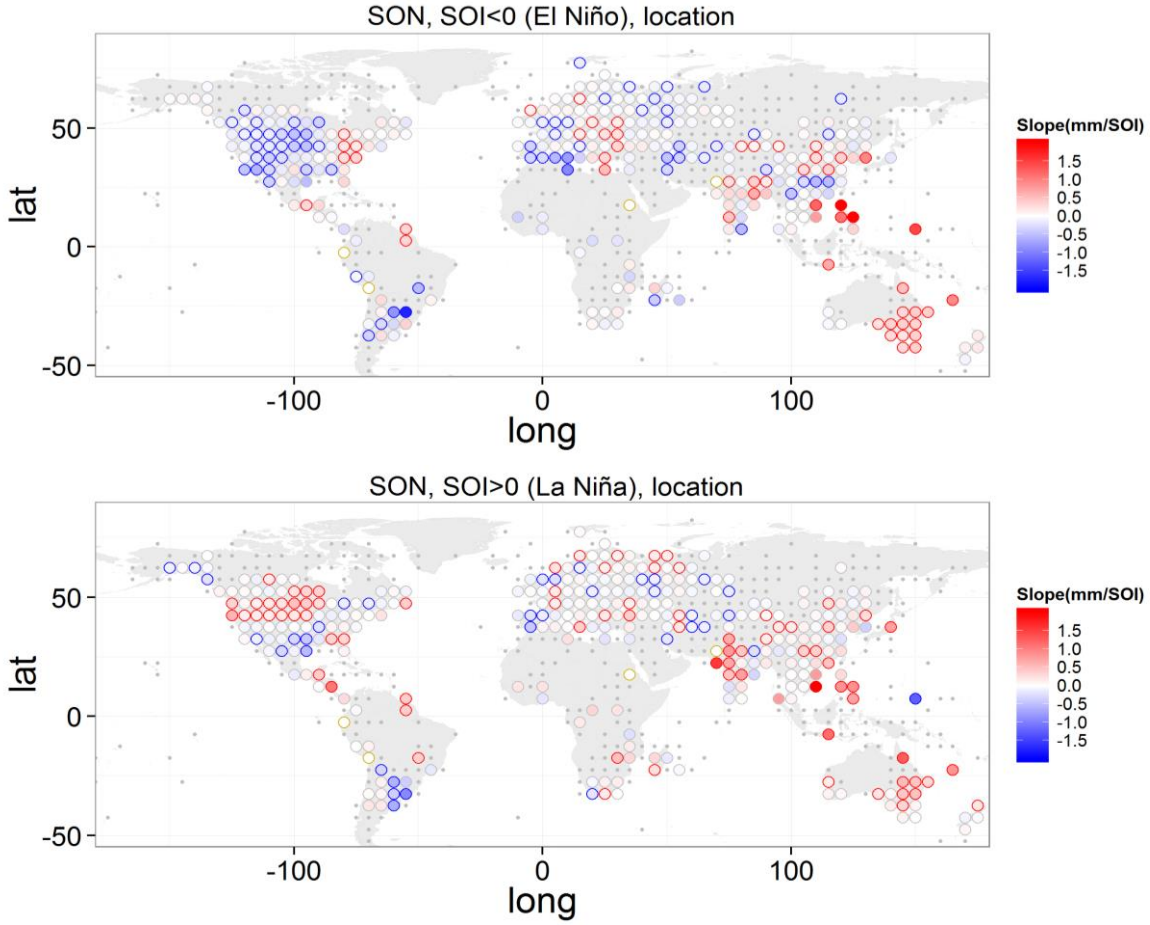
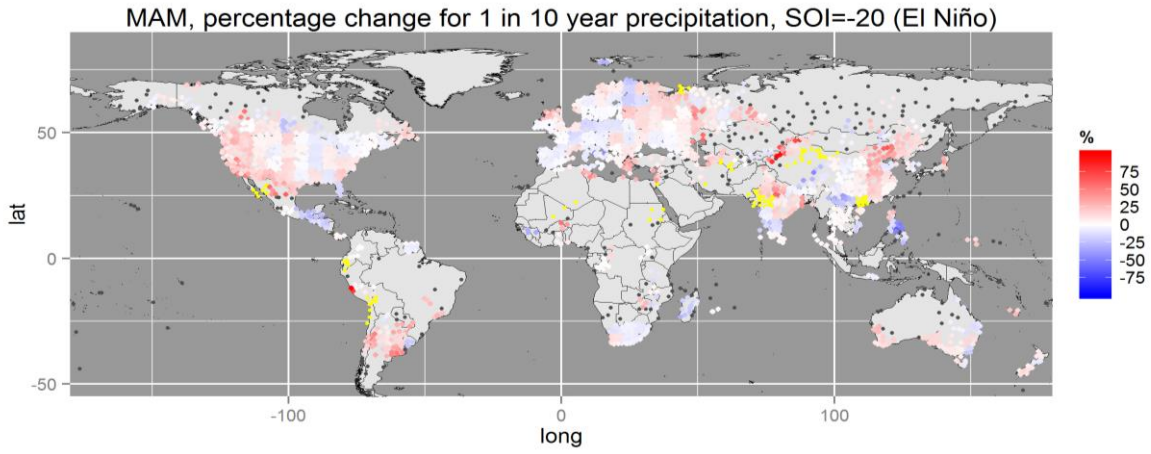
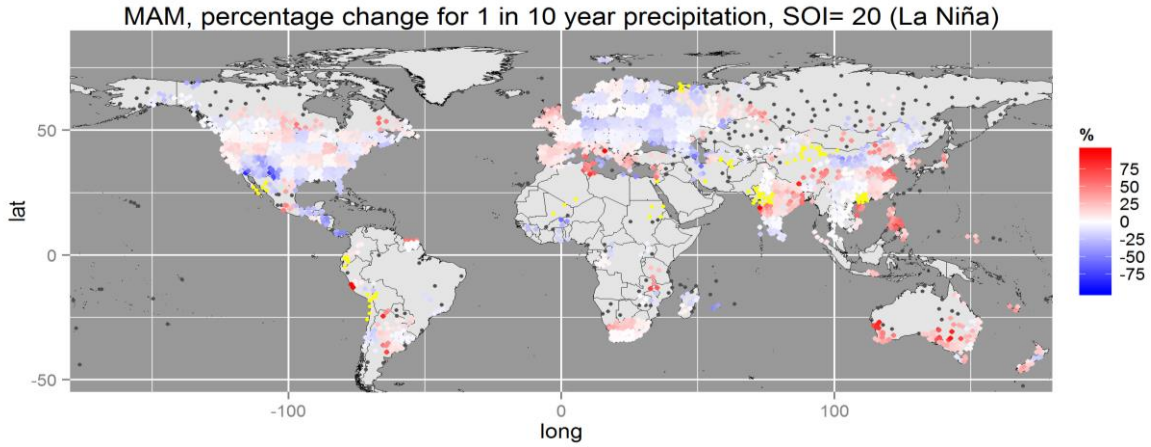


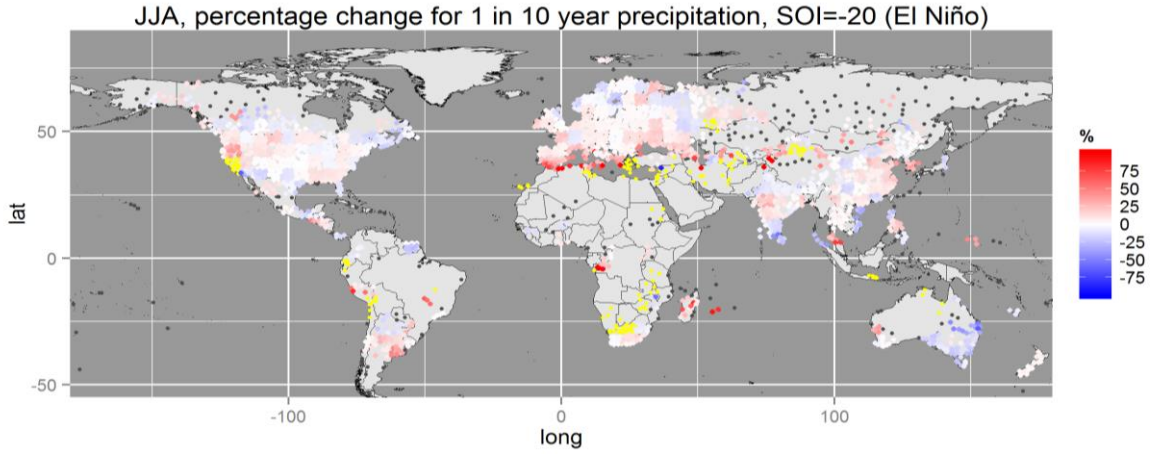
Figure 7.9- Slope of the location parameter with respect to SOI during El Niño ($\mu^{-}_{reg_1}$) and La Niña ($\mu^{+}_{reg_1}$) phases for the other three seasons (MAM, JJA and SON).

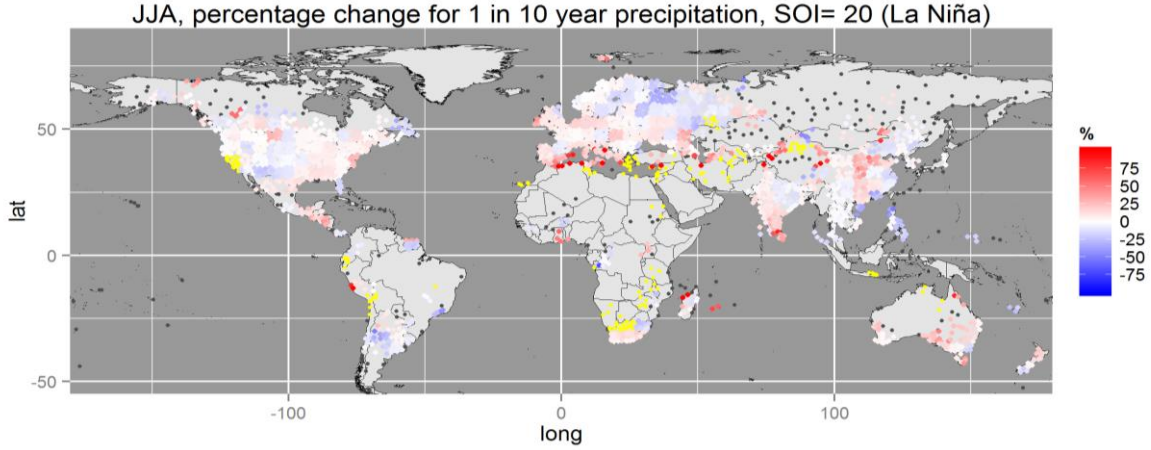
5.2 Percentage change for the intensity of 1 in 10 year precipitation relative to SOI=0

The following figures illustrate the percentage change for the intensity of 1 in 10 year precipitation relative to SOI=0 for the other three seasons (MAM, JJA and SON). Captions are the same as Figure 7.5.


MAM, percentage change for 1 in 10 year precipitation, SOI=-20 (El Niño)
lat
50
0
-50
-100
0
100
long
%
75
50
25
0
-25
-50
-75


MAM, percentage change for 1 in 10 year precipitation, SOI= 20 (La Niña)
lat
50
0
-50
-100
0
100
long
%
75
50
25
0
-25
-50
-75


JJA, percentage change for 1 in 10 year precipitation, SOI=-20 (El Niño)
lat
50
0
-50
-100
0
100
long
%
75
50
25
0
-25
-50
-75


JJA, percentage change for 1 in 10 year precipitation, SOI= 20 (La Niña)
lat
50
0
-50
-100
0
100
long
%
75
50
25
0
-25
-50
-75

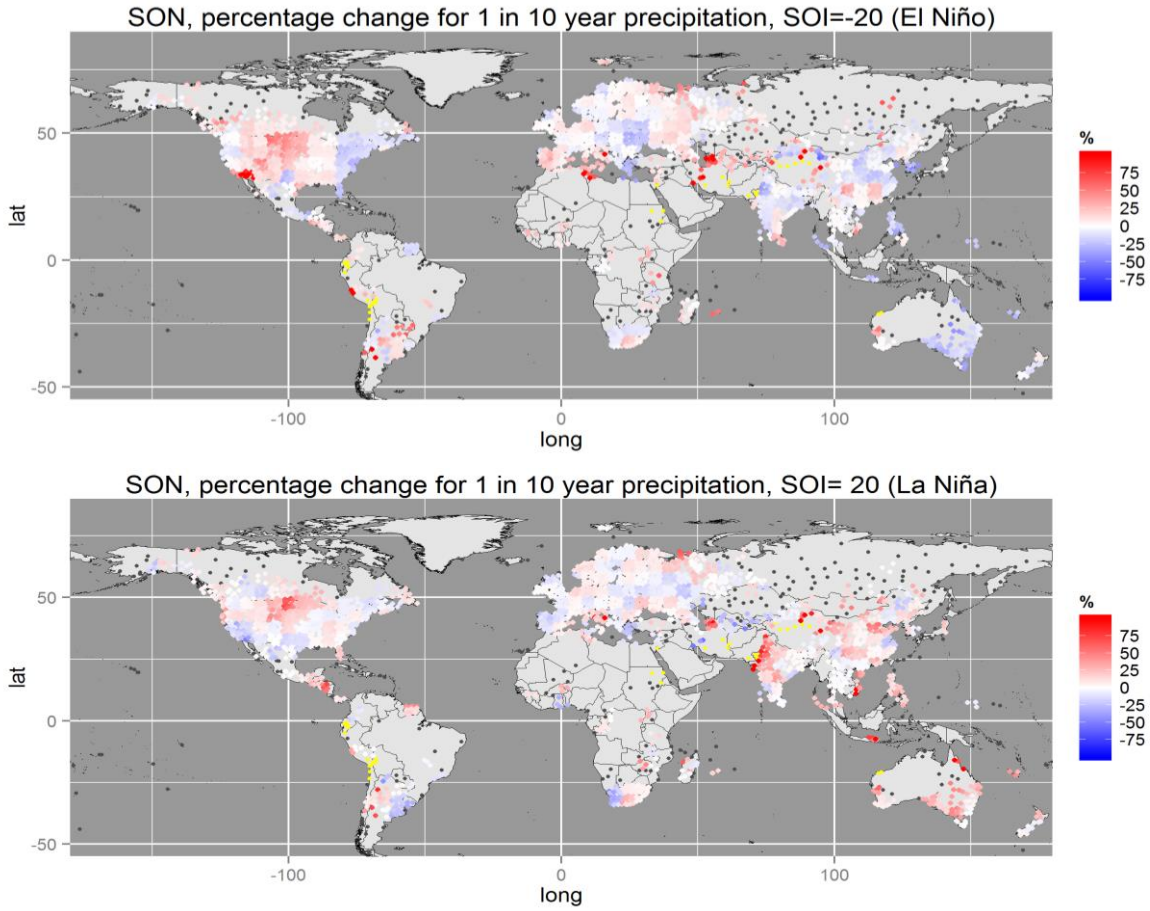


Figure 7.10-Percentage change for the intensity of 1 in 10 year precipitation relative to SOI=0 for the other three seasons (MAM, JJA and SON).

5.3 Difference between the slope of SOI during La Niña and El Niño phases

The following figures illustrate the difference between the slope of SOI during La Niña and El Niño phases for the other three seasons (MAM, JJA and SON). Captions are the same as Figure 7.7.

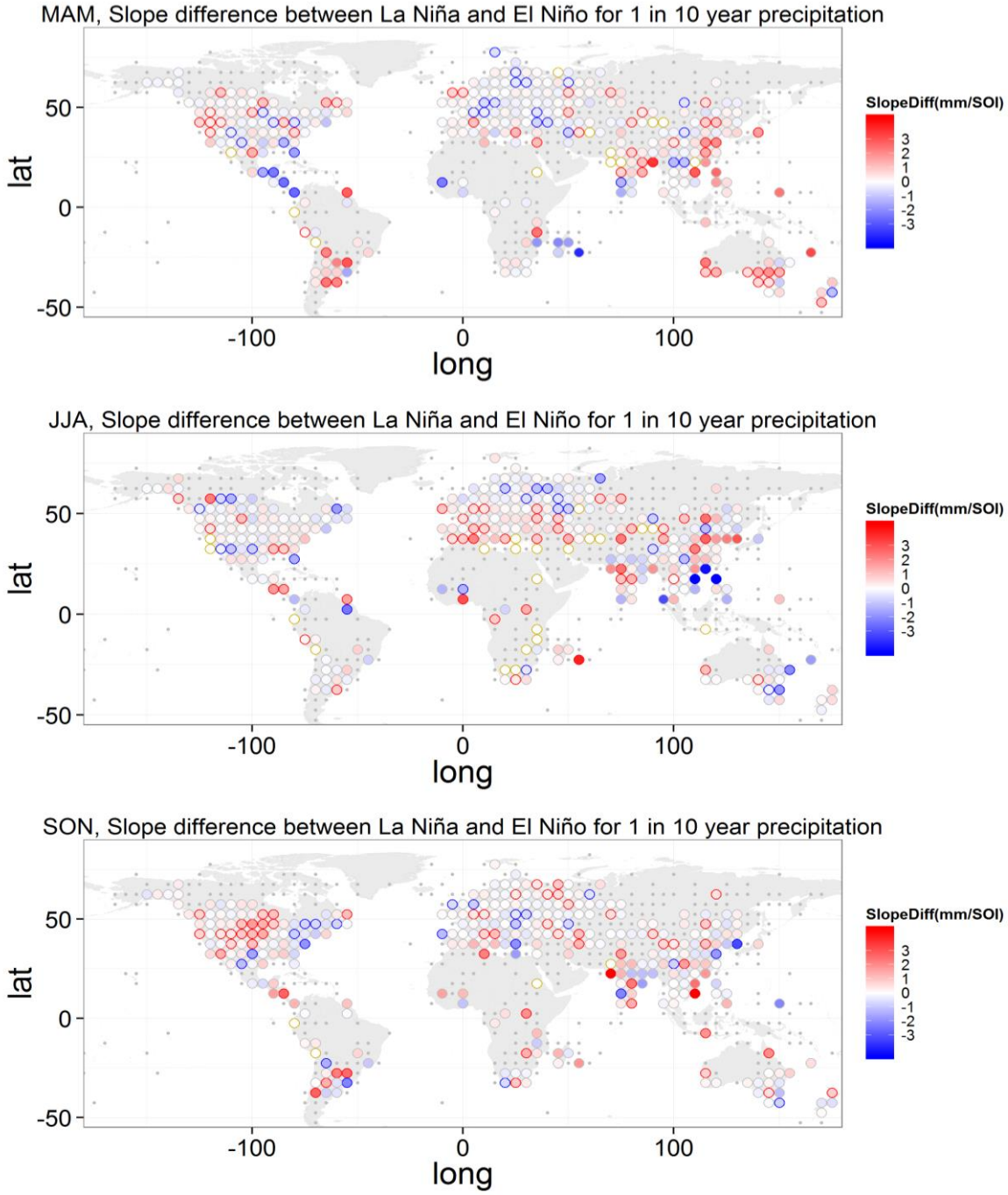


Figure 7.11- Difference between the slope of SOI during La Niña and El Niño phases (La Niña - El Niño) for 1 in 10 year precipitation for the other three seasons (MAM, JJA and SON).

Conclusion of the thesis

In this thesis, a general spatio-temporal regional frequency analysis framework was developed, geared towards analyzing and predicting hydrological hazards incorporating temporal trends and climate change/variability effects.

Development of the general spatio-temporal regional frequency analysis framework

The development of the model was undertaken through two steps: (i) local time-varying model construction, in which only temporal variation is considered; and (ii) regional spatio-temporal model construction, in which both temporal and spatial variations are involved. The former can be considered as a particular case of the latter. This spatio-temporal framework provides a very flexible platform for analyzing hydrological variables by incorporating three types of covariates: temporal, spatial and spatio-temporal covariates. In particular, it provides a free choice of the parent distribution for the observation, including discrete or continuous distributions. The relationship between the temporal (or spatio-temporal) covariates and data is modeled with a temporal regression where parameters of the parent distribution are a function of temporal covariates. The selection of temporal covariates is also flexible, since both deterministic variables (e.g. time) and stochastic variables (e.g. ENSO) can be used. In addition to the flexibility on the treatment of temporal effects, spatial effects are also involved in the framework. In order to introduce spatial effects in the parameters, a spatial regression function is applied to describe the relationship between the spatial covariates and the parameter.

A main advantage of the modeling framework is that spatial dependence is incorporated. In the thesis, we described the usage of elliptical copulas for modeling the spatial dependence of precipitation. Compared with models ignoring the spatial dependence, this model provides an important improvement on the estimation of parameter uncertainties, which enables a more realistic analysis in terms of detecting trends and climate affects.

In addition to the model construction, this spatio-temporal framework is integrated with other tools for facilitating the treatment of missing values in the data, the estimation of parameters and model diagnosis and comparison. In particular, thanks to carefully implemented algorithms for calculating the likelihood with non-missing data at each time step, no observation data are wasted in the inference. Furthermore, the parameter estimation uses a Bayesian approach, in which prior information of the parameters can be included. The Bayesian techniques also enable easily and naturally quantifying the uncertainties. In this thesis, a newly developed MCMC algorithm, which combines the adaptive block Metropolis method, an adaptive Metropolis-Hastings method and the classical Metropolis-Hastings method, is used to provide fast and efficient parameter estimation under the Bayesian framework. In a time-varying context, the probability-probability plot is used to evaluate the goodness-of-fit graphically. Model comparison tools are incorporated within the Bayesian framework, through the Deviance Information Criterion. These tools can be used to compare

specific models under different assumptions or with different regression functions. Completed with these surrounding tools, the general spatio-temporal framework developed in this thesis provides a flexible and convenient way for better identifying temporal trends and impacts of climate variability on hydrological events, as well as the induced hydrological hazard.

Assessment of the framework

Several studies are undertaken for evaluating the general framework. With the GCM-projected precipitation data for the 21st century in the Durance catchment, we demonstrate the flexibility in terms of choosing parent distributions and setting regression functions when analyzing variables without strong distribution guidance. Among the tested variables, a temporal trend is detected on the variable "first snowy day". This variable is also analyzed by using the same distribution with a time-invariant model for data from two sub-periods (current and future). There is a clear shift on the mean value and quantiles between the two periods. This result is consistent with the one found with our continuous time-varying model, which provides the estimation with much smaller uncertainties. Failure probabilities are also evaluated with both models, and results show that the time-varying model is more adapted for a risk assessment over a long duration.

The flexibility of the model is further discussed with the case study of French Mediterranean precipitation. Temporal trends and NAO effects on the annual maximum daily precipitation are analyzed at the local scale. With a GEV parent distribution, six regression models under different assumptions, including stationarity, temporal trend only and both temporal trend and NAO effect, are compared. Globally there is no strong indication of the existence of temporal trends and NAO effects, except for a few isolated sites.

The second assessment is to verify the advantage of considering spatial dependence through an elliptical copula. We simulate two datasets from time-varying Gaussian and GEV distributions. The parameter estimation is performed both considering and ignoring spatial dependence. As expected, in both cases, ignoring spatial dependence leads to an under-estimation of the parameter uncertainty. A further analysis is based on datasets simulated from maximum stable processes. Estimations are also performed both ignoring and considering spatial dependence with a Gaussian copula. With the dataset simulated with a moderate dependence, a significant improvement can be found in the estimation considering spatial dependence. However, in a high dependence case, results are not as convincing: using a Gaussian copula yields a realistic quantification of uncertainty for the location and trend parameters, but it leads to an over-estimation of the shape parameter.

The third assessment is to compare the Gaussian copula and maximum stable processes in modeling the spatial dependence of extreme data. The joint and conditional probabilities of an event exceeding a high threshold at two sites are compared. Results show that Schlather models systematically yield higher joint and conditional probabilities, due to its peculiar handling of dependence at large distances. On the other hand, the Smith model and the Gaussian copula yield markedly different estimates of conditional probabilities, which can be explained by the fact that the Smith model is asymptotically dependent, while the Gaussian copula is asymptotically independent. Overall, these results suggest that even if a Gaussian copula approximation may yield acceptable results in terms of estimating marginal parameters,

the computation of joint or conditional exceedance probabilities is much more sensitive to the representation of spatial dependence, and that unduly using a Gaussian copula may lead to underestimate conditional probabilities of exceedance.

The impact of ENSO on the global precipitation

One of the main contributions of the thesis is the quantification of the impact of ENSO on precipitation. Two case studies are performed. The first one analyzes the impact of ENSO on the summer total and summer maximum rainfall in Southeast Queensland, Australia. The second one provides a global analysis on the impact of ENSO on extreme precipitation. The general framework developed in this thesis provides a pragmatic way for these analyses. The temporal covariate used in the model is SOI and the impact of ENSO was quantified through the parameters characterizing SOI.

Summer precipitation over SEQ

In the case of Southeast Queensland, we first evaluate the asymmetric impact of ENSO on the summer total rainfall. An asymmetric linear regression function that separates different phases of ENSO is used on the mean of a lognormal distribution. Results show that La Niña exerts a significant influence in the region, while the impact of El Niño is not significant. The phenomenon is consistent with the literature review. Then the analysis turns to the extreme data, which is considered to be more uncertain than the total rainfall. Thanks to the flexibility of the general framework, various hypotheses are compared in this study, which includes the local and regional parameter settings with the assumption of stationarity, symmetric and asymmetric impact of ENSO. The DIC of these models indicates that the use of regional models yields a better identification of the impact of ENSO on extreme rainfall over SEQ. Among the assumptions, an asymmetric impact of ENSO is the most adapted, where La Niña has a strong effect on extreme summer rainfall, while El Niño has no effect. The results corroborate the findings with the dataset of summer total rainfall.

From an engineering perspective, it is found that the 1 in 100 year rainfall during a strong La Niña can be 20% to 50% higher than the estimates under the stationary assumption. This provides useful information for planners to organize in advance operational management and emergency responses during strong La Niña years.

The impact of ENSO on global extreme precipitation

Seeing the asymmetric impact of ENSO on the summer maximum daily rainfall in SEQ, we are further interested in the impact of ENSO on global extreme precipitation. This analysis is based on a new global high quality observation dataset, which includes about 7000 stations worldwide whose record length is longer than 40 years. The global map is gridded with 5˚ latitude by 5˚ longitude cells, which are considered as homogeneous regions. Although it varies with latitude, the surface of each cell is reasonable to perform regional analyses. Seeing the advantage of an asymmetric regional model in the case study of SEQ, we apply the same model to the global dataset.

It was found that during boreal winter (DJF), during El Niño phase, extreme precipitation is increased in southwest North America, southern South America, the southeast coast of China and northern Europe, while extreme precipitation is decreased in northwest North America and (more weakly) in South Africa. During La Niña phase, extreme precipitation is increased in northern North America, South Africa, Australia and northern Europe, while extreme precipitation is decreased in southern North America and northeast India. The 1 in 10 year precipitation calculated under a strong El Niño (SOI=-20) or La Niña (SOI=20) phase compared with that calculated during a neutral phase (SOI=0) showed a significant difference in some regions. For example, during a strong El Niño event, the intensity of a 1 in 10 year precipitation can be increased by more than 50% in Central America, 40% on the south coast of China, and nearly 20% in central North America and southeast South America. A decrease of about 15% can be observed in northwest America and a more than 20% decrease in the Philippines. During a strong La Niña episode, the intensity of a 1 in 10 year precipitation increases by about 15% the in northern North America, north Europe and the Mediterranean region, 10% to 40% (from east to west) in South Africa, about 20% in northeast China, more than 40% in eastern Australia and 60% in western Australia. A decrease by more than 50% in Mexico and nearly 25% in northeast India is also observed.

The impact of ENSO is found to be asymmetric in many regions, such as (during DJF) western North America, southeast South America, eastern China, Australia, and to a lesser extent in northern Europe and central Asia.

Lastly, the effect of ENSO is found to vary significantly according to the season. For the El Niño phase, DJF is the season during which the southern part of South America and southeast China are the most affected by ENSO processes. SON is the season during which ENSO most strongly affects North America (increasing precipitation in the eastern part and decreasing it in the west), the southern part of South America and eastern Australia (decrease). During the La Niña phase, the greatest impact is during DJF in the southern part of North America (decrease), the southern part of South America (decrease), South Africa and Australia, while MAM is the season most strongly affected in the southern part of North America (decrease) and Southeast China. In the middle north of North America, the southern part of South America (decrease) and northern India, the greatest effects are felt during SON. In general, the impact of ENSO processes is weak in JJA.

Perspectives

While the usefulness and the flexibility of the modeling framework developed in this thesis was demonstrated, several avenues for improvement exist as discussed in the following sections.

Using spatial and spatio-temporal covariates

While implemented in the modeling framework, spatial regressions have been scarcely tested in this work. Indeed in all case studies, we assumed identical regional parameters, which correspond to using identical spatial regressions. The usefulness of such structure needs further examination by case studies. In the ENSO case studies of Chapter 6 and Chapter 7, the ENSO effect could for instance vary with elevation or distance to sea. Alternatively, spatializing purely local parameters using a spatial regression would enable predicting quantiles at ungauged sites. This was not attempted in this case study because identifying such spatial effects is difficult with a limited number of sites (10 to 16): we therefore favored the identification of ENSO effects. However, future case studies based on a spatially denser dataset will investigate in more depth the construction of such spatial models.

Similarly, spatio-temporal covariates have not been used, although this possibility is implemented. An example of interesting use of a spatio-temporal covariate would be to include weather type information: at each site and each time step, the weather type observed on the day of the annual maximum could be used as covariate. *Garavaglia et al.* [2010]; [2011] demonstrated the interest of weather type information for predicting extreme precipitation.

Construction of hierarchical models

In the regional model, we proposed two kinds of parameters: local and regional parameters. Local parameters are different for each site, which offers a good flexibility. Regional parameters are common for all sites, yielding reduced uncertainties (e.g. Figure 6.13 and Figure 6.14). However, this distinction may be too "rigid". Some parameters may be different at each site, but still have some spatial consistency. A possible improvement is to use hierarchical models to enable constrained variations of parameters in space. *Wikle et al.* [1998] described a general hierarchical Bayesian framework in a non-stationary context. *Lima and Lall* [2009] [2010] used hierarchical models to describe the daily rainfall occurrence and extreme runoff. *Renard* [2011] and *Renard et al.* [2013] proposed a general hierarchical approach to regional frequency analysis. Future work could therefore be to generalize the model proposed in this thesis to a hierarchical setup. In particular, this may yield a more realistic quantification of uncertainties at ungauged sites, and hence have interesting applications for hazard mapping.

Treatment of spatial dependence

One of the advantages of the framework is to consider the spatial dependence of data. However, Chapter 5 showed that the treatment of spatial dependence using elliptical copulas

is perfectible. Additional work is needed to include alternative dependence structures, especially for extremes. Maximum stable processes provide a good alternative. However, as discussed in Chapter 5, the Smith and Schlather representations are also perfectible. Both representations have properties that do not seem very realistic when compared with real rainfall patterns: the Smith representation generates much too smooth rainfall fields, while intersite dependence does not vanishes to zero at infinite distances with the Schlather representation. Alternative representations that are more consistent with the properties of rainfall fields could be explored in the future. More generally, incorporating maximum stable models within a Bayesian framework is challenging as well, because of the difficulty in writing the joint distribution of a large number of sites. Thus statistical developments are still required to improve the practical handling of max-stable spatial dependence models.

On the other hand, the dependence-distance models integrated in the framework are convenient for spatially continuous variables such as precipitation. However, such dependence model cannot directly be used for streamflow variables, because streamflow is not a punctual variable but is instead related to a catchment, and his dependence is therefore structured by the hydrologic network, rather than the sole inter-site distance. The framework could be completed in the future by developing new dependence models adapted to streamflow applications. This would certainly involve the definition of non-Euclidean distances that can account for the structure of the hydrologic network. *Blanchet and Davison* [2011] developed some practical solutions to model the spatial dependence of snow depth using non-Euclidean distances, which may be extended to the streamflow variables.

Models for peak-over-threshold (POT)

In this thesis, we favored the use of a block maxima (GEV) representation to describe extreme precipitations. An alternative description would be to use a POT (GPD) representation. This has not been attempted because it is much more difficult to model spatial dependence for non-concomitant peaks over threshold. Block maxima are easier because they can be related to the same year, making the derivation of a multivariate dataset straightforward. While models adapted to multivariate threshold exceedance have been proposed in moderate dimension (e.g. *Heffernan and Tawn* [2004], *Boldi and Davison* [2007] and *Sabourin and Naveau* [2013]), the extension of such approaches to the highly-dimensional spatial case is challenging. An important advantage of using peaks over threshold data is that more data could be used in the regressions. Moreover, extremes could be described both in terms of severity and frequency. A further extension of the framework could be the development of models adapted to the POT representation, following progresses made in the statistical community.

Further analysis on the impact of ENSO

In the global analysis of the impact of ENSO, regional models were applied on grid cells. As discussed in Chapter 6 (Section 3.1), using cells as regions may lead to an incorrect representation of different climate regions. In future work, improvements can be achieved if well-defined homogeneous regions are developed. Furthermore, once homogenous regions are

well-defined, spatial effects can be included using spatial regression models with spatial covariates (e.g. elevation) to improve the analysis.

Besides the study of the ENSO impact on precipitation, similar analyses could be performed on flood variables. This would rely on the development of specialized spatial dependence models for floods as discussed previously. With the study on floods, planner/water resource managers could benefit from more direct information to better organize their operational rules. At the other side of the hydrologic spectrum, studying the impact of ENSO on meteorological and hydrological droughts could also have significant operational applications.

Lastly, besides ENSO, other large scale modes of climate variability could also affect regional extreme precipitation. As discussed in Chapter 7 (Section 3.3), it will be also interesting to use similar models for analyzing the impact of climate modes such as NAO or IOD on the regional and global precipitation. An interesting open question to investigate is whether the effects of distinct modes of variability simply add up to each other, or if significant interactions can be detected.

Frequency analysis based on the output of GCM simulations

In the context of projecting climate change impacts on hydrological variables, most analyses are still performed using time-invariant models with sub-periods data representing current and future climates. With the framework developed in this thesis, we can already establish a time-varying model with an entire transient GCM run, rather than restricting to sub-periods, which unnecessarily wastes GCM-simulated data [*Hanel et al.*, 2009b]. However, a limitation of the analysis performed in this thesis is that it relied on a single GCM, and moreover, it did not evaluate the ability of the GCM to reproduce the observed climate. Further developments could therefore focus on how to use such time-varying FA framework with the combined observations (historical data) and multi-GCM simulations (future data) to provide more reliable projections for the future.

For the engineering perspective, an important question is still how to design future hydraulic constructions in a changing climate. In many countries, the current approach is to fix an annual exceedance probability (generally prescribed by regulation) and to design the structure in order to withstand the corresponding quantile. However, in a context where distributions change with time, this approach is ambiguous: should the quantiles be calculated based on the distribution at the time of construction, at the end of the structure's expected lifetime, or at some other fixed future date? Isn't possible to base the design on the failure probability computed over the expected lifetime of the structure? Given the uncertainties affecting future climate and hydrologic projections, is a calculation based on the non-stationary assumption systematically justified? How to account for these uncertainties in designing future hydraulic structures? All these questions remain open for future works.

Bibliography

AghaKouchak, A., A. Bardossy, and E. Habib (2010), Copula-based uncertainty modelling: application to multisensor precipitation estimates, *Hydrological Processes*, *24*(15), 2111-2124.

Akaike, H. (1974), New look at statistical-model identification, *Ieee Transactions on Automatic Control*, *AC19*(6), 716-723.

Alexander, L. V., P. Uotila, and N. Nicholls (2009), Influence of sea surface temperature variability on global temperature and precipitation extremes, *Journal of Geophysical Research-Atmospheres*, *114*.

Arnaud, P., and J. Lavabre (1999), Using a stochastic model for generating hourly hyetographs to study extreme rainfalls, *Hydrological Sciences Journal-Journal des Sciences Hydrologiques*, *44*(3), 433-446.

Aryal, S. K., B. C. Bates, E. P. Campbell, Y. Li, M. J. Palmer, and N. R. Viney (2009), Characterizing and modeling temporal and spatial trends in rainfall extremes, *Journal of Hydrometeorology*, *10*(1), 241-253.

Balkema, A. A., and L. De Haan (1974), Residual life time at great age, *The Annals of Probability*, 792-804.

Bardossy, A., and J. Li (2008), Geostatistical interpolation using copulas, *Water Resources Research*, *44*(7), W07412.

Bardossy, A., and G. G. S. Pegram (2009), Copula based multisite model for daily precipitation simulation, *Hydrology and Earth System Sciences*, *13*(12), 2299-2314.

Barnston, A. G., and R. E. Livezey (1987), Classification, seasonality and persistence of low-frequency atmospheric circulation patterns *Mon Weather Rev*, *115*(6), 1083-1126.

Bates, B. C., A. Rahman, R. G. Mein, and P. W. Weinmann (1998), Climatic and physical factors that influence the homogeneity of regional floods in southeastern Australia, *Water Resources Research*, *34*(12), 3369-3381.

Bellenger, H., E. Guilyardi, J. Leloup, M. Lengaigne, and J. Vialard (2013), ENSO representation in climate models: from CMIP3 to CMIP5, *Climate Dynamics*, 1-20.

Berger, J. O. (1985), *Statistical decision theory and Bayesian analysis*, Springer-Verlag, New York.

Blanchet, J., and A. C. Davison (2011), Spatial Modeling of Extreme Snow Depth, *Ann Appl Stat*, *5*(3), 1699-1725.

Boe, J., L. Terray, F. Habets, and E. Martin (2006), A simple statistical-dynamical downscaling scheme based on weather types and conditional resampling, *Journal of Geophysical Research-Atmospheres*, *111*(D23), D23106.

Boldi, M. O., and A. C. Davison (2007), A mixture model for multivariate extremes, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(2), 217-229.

Boughton, W., and O. Droop (2003), Continuous simulation for design flood estimation - a review, *Environ. Modell. Softw.*, *18*(4), 309-318.

Brigode, P. (2013), Changement climatique et risque hydrologique: évaluation de la méthode SCHADEX en contexte non-stationnaire, PhD thesis, Université Pierre Marie Curie.

Brockwell, P. J., and R. A. Davis (2006), *Time series: theory and methods*, Springer.

Brunetti, M., M. Maugeri, and T. Nanni (2001), Changes in total precipitation, rainy days and extreme events in northeastern Italy, *International Journal of Climatology*, *21*(7), 861-871.

Burnham, K. P., and D. R. Anderson (2002), *Model selection and multi-model inference: a practical information-theoretic approach*, Springer.

Cai, W., and T. Cowan (2009), La Nina Modoki impacts Australia autumn rainfall variability, *Geophysical Research Letters*, *36*.

Cai, W., P. van Rensch, T. Cowan, and A. Sullivan (2010), Asymmetry in ENSO teleconnection with regional rainfall, its multidecadal variability, and impact, *Journal of Climate*, *23*(18), 4944-4955.

Cai, W., P. van Rensch, T. Cowan, and H. H. Hendon (2012), An asymmetry in the IOD and ENSO teleconnection pathway and its impact on Australian climate, *Journal of Climate*, *25*(18), 6318-6329.

Castello, A. F., and M. L. Shelton (2004), Winter precipitation on the US Pacific Coast and El Nino Southern oscillation events, *International Journal of Climatology*, *24*(4), 481-497.

Cayan, D. R., K. T. Redmond, and L. G. Riddle (1999), ENSO and hydrologic extremes in the western United States, *Journal of Climate*, *12*(9), 2881-2893.

Chen, J. M., T. Li, and C. F. Shih (2008), Asymmetry of the El Nino-spring rainfall relationship in Taiwan, *J. Meteorol. Soc. Jpn.*, *86*(2), 297-312.

Chowdhury, J. U., J. R. Stedinger, and L. H. Lu (1991), Goodness-of-fit tests for regional generalized extreme value flood distributions, *Water Resources Research*, *27*(7), 1765-1776.

Cohn, T. A., W. L. Lane, and J. R. Stedinger (2001), Confidence intervals for Expected Moments Algorithm flood quantile estimates, *Water Resources Research*, *37*(6), 1695-1706.

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

Coles, S., L. R. Pericchi, and S. Sisson (2003), A fully probabilistic approach to extreme rainfall modeling, *Journal of Hydrology*, *273*(1-4), 35 - 50.

Cooley, D. (2013), Return Periods and Return Levels Under Climate Change, in *Extremes in a Changing Climate*, edited, pp. 97-114, Springer.

Cooley, D., D. Nychka, and P. Naveau (2007), Bayesian spatial modeling of extreme precipitation return levels, *Journal of The American Statistical Association*, *102*(479), 824-840.

Cunderlik, J. M., and D. H. Burn (2003), Non-stationary pooled flood frequency analysis, *Journal of Hydrology*, *276*(1-4), 210-223.

Curtis, S., A. Salahuddin, R. F. Adler, G. J. Huffman, G. J. Gu, and Y. Hong (2007), Precipitation extremes estimated by GPCP and TRMM: ENSO relationships, *Journal of Hydrometeorology*, *8*(4), 678-689.

Dai, A., and T. M. L. Wigley (2000), Global patterns of ENSO-induced precipitation, *Geophys Res Lett*, *27*(9), 1283-1286.

Dai, A., I. Y. Fung, and A. D. DelGenio (1997), Surface observed global land precipitation variations during 1900-88, *Journal of Climate*, *10*(11), 2943-2962.

Donat, M. G., et al. (2013), Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century: The HadEX2 dataset, *Journal of Geophysical Research: Atmospheres*, *118*(5), 2098-2118.

Durrans, S. R., and S. Tomic (2001), Comparison of parametric tail estimators for low-flow frequency analysis, *J Am Water Resour As*, *37*(5), 1203-1214.

Durrans, S. R., and J. T. Kirby (2004), Regionalization of extreme precipitation estimates for the Alabama rainfall atlas, *Journal of Hydrology*, *295*(1-4), 101-107.

Favre, A. C., S. El Adlouni, L. Perreault, N. Thiemonge, and B. Bobee (2004), Multivariate hydrological frequency analysis using copulas, *Water Resources Research*, *40*(1), W01101.

Feng, J., and J. P. Li (2011), Influence of El Nino Modoki on spring rainfall over south China, *Journal of Geophysical Research-Atmospheres*, *116*.

Fernandez, H. W., and B. Fernandez (2002), The influence of ENSO in the precipitation regime in southern South America, *Ingenieria Hidraulica En Mexico*, *17*(3), 5-16.

Fisher, R. A., and L. H. C. Tippett (1928), Limiting forms of the frequency distribution of the largest or smallest member of a sample, *Mathematical Proceedings of the Cambridge Philosophical Society*, *24*(02), 180-190.

Frost, A. J. (2004), Spatio-temporal hidden markov models for incorporating interannual variability in rainfall, PhD thesis, University of Newcastle.

Garavaglia, F., J. Gailhard, E. Paquet, M. Lang, R. Garcon, and P. Bernardara (2010), Introducing a rainfall compound distribution model based on weather patterns sub-sampling, *Hydrology and Earth System Sciences*, *14*(6), 951-964.

Garavaglia, F., M. Lang, E. Paquet, J. Gailhard, R. Garcon, and B. Renard (2011), Reliability and robustness of rainfall compound distribution model based on weather pattern sub-sampling, *Hydrology and Earth System Sciences*, *15*(2), 519-532.

Gelman, A., and D. B. Rubin (1992), Inference from iterative simulation using multiple sequences, *Statistical science*, 457-472.

Genest, C., A. C. Favre, J. Beliveau, and C. Jacques (2007), Metaelliptical copulas and their use in frequency analysis of multivariate hydrological data, *Water Resources Research*, *43*(9), W09401.

Gershunov, A., and T. P. Barnett (1998), ENSO influence on intraseasonal extreme rainfall and temperature frequencies in the contiguous United States: Observations and model results, *Journal of Climate*, *11*(7), 1575-1586.

Ghosh, S., and B. K. Mallick (2011), A hierarchical Bayesian spatio-temporal model for extreme precipitation events, *Environmetrics*, *22*(2), 192-204.

Gibelin, A. L., and M. Déqué (2003), Anthropogenic climate change over the Mediterranean region simulated by a global variable resolution model, *Climate Dynamics*, *20*(4), 327-339.

Goswami, B. N., V. Venugopal, D. Sengupta, M. S. Madhusoodanan, and P. K. Xavier (2006), Increasing trend of extreme rain events over India in a warming environment, *Science*, *314*(5804), 1442-1445.

Gregersen, I. B., H. Madsen, D. Rosbjerg, and K. Arnbjerg-Nielsen (2013), A spatial and non-stationary model for the frequency of extreme rainfall events, *Water Resources Research*, doi:10.1029/2012WR012570, in press.

Grimm, A. M. (2011), Interannual climate variability in South America: impacts on seasonal precipitation, extreme events, and possible effects of climate change, *Stochastic Environmental Research and Risk Assessment*, *25*(4), 537-554.

Grimm, A. M., and R. G. Tedeschi (2009), ENSO and Extreme Rainfall Events in South America, *Journal of Climate*, *22*(7), 1589-1609.

Hanel, M., T. A. Buishand, and C. A. T. Ferro (2009a), A nonstationary index flood model for precipitation extremes in transient regional climate model simulations, *Journal of Geophysical Research-Atmospheres*, *114*.

Hanel, M., T. A. Buishand, and C. A. Ferro (2009b), A nonstationary index flood model for precipitation extremes in transient regional climate model simulations, *Journal of Geophysical Research: Atmospheres (1984–2012)*, *114*(D15).

Hannachi, A. (2001), Toward a nonlinear identification of the atmospheric response to ENSO, *Journal of Climate*, *14*(9), 2138-2149.

Hastie, T., and R. Tibshirani (1986), Generalized additive models, *Statistical science*, 297-310.

He, J., and C. Valeo (2009), Comparative Study of ANNs versus Parametric Methods in Rainfall Frequency Analysis, *Journal of Hydrologic Engineering*, *14*(2), 172-184.

Heffernan, J. E., and J. A. Tawn (2004), A conditional approach for multivariate extreme values (with discussion), *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *66*(3), 497-546.

Henley, B. J., M. A. Thyer, G. Kuczera, and S. W. Franks (2011), Climate-informed stochastic hydrological modeling: Incorporating decadal-scale variability using paleo data, *Water Resources Research*, *47*(11), W11509.

Higgins, R. W., V. E. Kousky, and P. Xie (2011), Extreme Precipitation Events in the South-Central United States during May and June 2010: Historical Perspective, Role of ENSO, and Trends, *Journal of Hydrometeorology*, *12*(5), 1056-1070.

Hoerling, M. P., A. Kumar, and M. Zhong (1997), El Nino, La Nina, and the nonlinearity of their teleconnections, *Journal of Climate*, *10*(8), 1769-1786.

Hoerling, M. P., A. Kumar, and T. Y. Xu (2001), Robustness of the nonlinear climate response to ENSO's extreme phases, *Journal of Climate*, *14*(6), 1277-1293.

Hosking, J. R. M. (1990), L-Moment - Analysis and Estimation of Distributions Using Linear-Combinations of Order-Statistics, *J Roy Stat Soc B Met*, *52*(1), 105-124.

Hosking, J. R. M., J. R. Wallis, and E. F. Wood (1985), An appraisal of the regional flood frequency procedure in the UK flood studies report, *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*, *30*(1), 85-102.

Hundecha, Y., and B. Merz (2012), Exploring the relationship between changes in climate and floods using a model-based analysis, *Water Resources Research*, *48*(4).

Hurrell, J. W. (1995), Decadal trends in the North-Atlantic Oscillation - regional temperatures and precipitation, *Science*, *269*(5224), 676-679.

Hurrell, J. W., and H. VanLoon (1997), Decadal variations in climate associated with the north Atlantic oscillation, *Climatic Change*, *36*(3-4), 301-326.

Hurvich, C. M., and C.-L. Tsai (1989), Regression and time series model selection in small samples, *Biometrika*, *76*(2), 297-307.

IPCC (2007a), *Climate change 2007-the physical science basis: Working group I contribution to the fourth assessment report of the IPCC*, Cambridge University Press.

IPCC (2007b), Climate change 2007. Synthesis report. Contribution of Working Groups I, II and III to the fourth assessment report*Rep.*, Intergovernmental Panel on Climate Change.

IPCC (2012), *Managing the risks of extreme events and disasters to advance climatechange adaptation. A special report of working groups I and II of the Intergovernmental Panel on Climate Change [Field, C.B., V. Barros, T.F. Stocker,D. Qin, D.J. Dokken, K.L. Ebi, M.D. Mastrandrea, K.J. Mach, G.-K. Plattner, S.K. Allen,M. Tignor, and P.M. Midgley (eds.)]*. Cambridge University Press, Cambridge CB2 8RU England.

Jones, C., and L. M. V. Carvalho (2012), Spatial-Intensity Variations in Extreme Precipitation in the Contiguous United States and the Madden-Julian Oscillation, *Journal of Climate*, *25*(14), 4898-4913.

Jones, P. D., T. Jonsson, and D. Wheeler (1997), Extension to the North Atlantic Oscillation using early instrumental pressure observations from Gibraltar and south-west Iceland, *International Journal of Climatology*, *17*(13), 1433-1450.

Jónsdóttir, J., C. Uvo, and A. Snorrason (2004), Multivariate analysis of Icelandic river flow and its relation to variability in atmospheric circulation, in *XIII Nordic Hydrological Conference*, edited, Tallinn, Estonia.

Kaczmarek, Z. (2003), The impact of climate variability on flood risk in Poland, *Risk Analysis*, *23*(3), 559-566.

Kane, R. P. (1999), El Nino timings and rainfall extremes in India, Southeast Asia and China, *International Journal of Climatology*, *19*(6), 653-672.

Kass, R. E., and A. E. Raftery (1995), Bayes Factors, *Journal of the American Statistical Association*, *90*(430), 773-795.

Katz, R. W., M. B. Parlange, and P. Naveau (2002), Statistics of extremes in hydrology, *Advances in Water Resources*, *25*(8-12), 1287 - 1304.

Kayano, M. T., and R. V. Andreoli (2006), Relationships between rainfall anomalies over northeastern Brazil and the El Nino-Southern Oscillation, *Journal of Geophysical Research-Atmospheres*, *111*(D13).

Keim, A. S., and D. C. Verdon-Kidd (2009), Climatic drivers of Victorian streamflow: Is ENSO the dominant influence?, *Australian Journal of Water Resources*, *13*(1), 17.

Kenyon, J., and G. C. Hegerl (2010), Influence of Modes of Climate Variability on Global Precipitation Extremes, *Journal of Climate*, *23*(23), 6248-6262.

Khaliq, M. N., T. B. M. J. Ouarda, J. C. Ondo, P. Gachon, and B. Bobee (2006), Frequency analysis of a sequence of dependent and/or non-stationary hydro-meteorological observations: A review, *Journal of Hydrology*, *329*(3-4), 534-552.

Kiely, G. (1999), Climate change in Ireland from precipitation and streamflow observations, *Adv. Water Resour.*, *23*(2), 141-151.

King, A. D., L. V. Alexander, and M. G. Donat (2013), Asymmetry in the response of eastern Australia extreme rainfall to low-frequency Pacific variability, *Geophys Res Lett*, 1-6.

Kingston, D. G., D. M. Hannah, D. M. Lawler, and G. R. McGregor (2006), Interactions between large-scale climate and river flow across the northern North Atlantic margin, *Iahs-Aish P*, *308*, 350-355.

Kousky, V. E., M. T. Kagano, and I. F. A. Cavalcanti (1984), A review of the Southern Oscillation - oceanic-atmospheric circulation changes and related rainfall anomalies, *Tellus Series a-Dynamic Meteorology and Oceanography*, *36*(5), 490-504.

Kripalani, R. H., and A. Kulkarni (1997), Rainfall variability over South-east Asia - Connections with Indian monsoon and enso extremes: New perspectives, *International Journal of Climatology*, *17*(11), 1155-1168.

Kroll, C. N., and J. R. Stedinger (1996), Estimation of moments and quantiles using censored data, *Water Resources Research*, *32*(4), 1005-1012.

Kruger, A. C. (1999), The influence of the decadal-scale variability of summer rainfall on the impact of El Nino and La Nina events in South Africa, *International Journal of Climatology*, *19*(1), 59-68.

Kunkel, K. E., D. R. Easterling, D. A. R. Kristovich, B. Gleason, L. Stoecker, and R. Smith (2010), Recent increases in U.S. heavy precipitation associated with tropical cyclones, *Geophys Res Lett*, *37*.

Kwon, H.-H., A. F. Khalil, and T. Siegfried (2008), Analysis of extreme summer rainfall using climate teleconnections and typhoon characteristics in South Korea, *J Am Water Resour As*, *44*(2), 436-448.

Kysely, J. (2008), A cautionary note on the use of nonparametric bootstrap for estimating uncertainties in extreme-value models, *Journal of Applied Meteorology and Climatology*, *47*(12), 3236-3251.

Kysely, J. (2009), Trends in heavy precipitation in the Czech Republic over 1961-2005, *International Journal of Climatology*, *29*(12), 1745-1758.

Kysely, J., J. Picek, and R. Beranova (2010), Estimating extremes in climate change simulations using the peaks-over-threshold method with a non-stationary threshold, *Global and Planetary Change*, *72*(1-2), 55-68.

Kysely, J., L. Gaal, and J. Picek (2011), Comparison of regional and at-site approaches to modelling probabilities of heavy precipitation, *International Journal of Climatology*, *31*(10), 1457-1472.

Lang, M., T. B. M. J. Ouarda, and B. Bobee (1999), Towards operational guidelines for over-threshold modeling, *Journal of Hydrology*, *225*(3-4), 103-117.

Latif, M., V. Semenov, and P. Wonsun (2013), Super El Niños in a Warming World, paper presented at EGU General Assembly Conference Abstracts.

Lavery, B., G. Joung, and N. Nicholls (1997), An extended high-quality historical rainfall dataset for Australia, *Aust Meteorol Mag*, *46*(1), 27-38.

Lecam, L. (1990), Maximum-Likelihood - an Introduction, *Int Stat Rev*, *58*(2), 153-171.

Leclerc, M., and T. B. M. J. Ouarda (2007), Non-stationary regional flood frequency analysis at ungauged sites, *Journal of Hydrology*, *343*(3-4), 254-265.

Li, C., and H. Ma (2012), Relationship between ENSO and winter rainfall over Southeast China and its decadal variability, *Adv Atmos Sci*, *29*(6), 1129-1141.

Lima, C. H. R., and U. Lall (2009), Hierarchical Bayesian modeling of multisite daily rainfall occurrence: Rainy season onset, peak, and end, *Water Resources Research*, *45*, W07422.

Lima, C. H. R., and U. Lall (2010), Spatial scaling in a changing climate: A hierarchical bayesian model for non-stationary multi-site annual maximum and monthly streamflow, *Journal of Hydrology*, *383*(3-4), 307-318.

Limanówka, D., Z. Nieckarz, and J. Pociask-Karteczka (2002), The North Atlantic Oscillation impact on hydrological regime in Polish Carpathians, in *ERB and Northern European FRIEND Project 5 Conference*, edited, Demanovska dolina, Slovakia.

Lins, H. F., and T. A. Cohn (2011), Stationarity: Wanted dead or alive?, *JAWRA Journal of the American Water Resources Association*, *47*(3), 475-480.

Lyon, B., and A. G. Barnston (2005), ENSO and the spatial extent of interannual precipitation extremes in tropical land areas, *Journal of Climate*, *18*(23), 5095-5109.

Madsen, H., and D. Rosbjerg (1997a), The partial duration series method in regional index-flood modeling, *Water Resources Research*, *33*(4), 737-746.

Madsen, H., and D. Rosbjerg (1997b), Generalized least squares and empirical bayes estimation in regional partial duration series index-flood modeling, *Water Resources Research*, *33*(4), 771-781.

Madsen, H., P. F. Rasmussen, and D. Rosbjerg (1997a), Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events .1. At-site modeling, *Water Resources Research*, *33*(4), 747-757.

Madsen, H., C. P. Pearson, and D. Rosbjerg (1997b), Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events .2. Regional modeling, *Water Resources Research*, *33*(4), 759-769.

Madsen, H., K. Arnbjerg-Nielsen, and P. S. Mikkelsen (2009), Update of regional intensity-duration-frequency curves in Denmark: Tendency towards increased storm intensities, *Atmos. Res.*, *92*(3), 343-349.

Madsen, H., P. S. Mikkelsen, D. Rosbjerg, and P. Harremoës (2002), Regional estimation of rainfall intensity-duration-frequency curves using generalized least squares regression of partial duration series statistics, *Water Resources Research*, *38*(11), 1239.

Maraun, D., H. W. Rust, and T. J. Osborn (2010), Synoptic airflow and UK daily precipitation extremes Development and validation of a vector generalised linear model, *Extremes*, *13*(2), 133-153.

Maraun, D., T. J. Osborn, and H. W. Rust (2011), The influence of synoptic airflow on UK daily precipitation extremes. Part I: Observed spatio-temporal relationships, *Climate Dynamics*, *36*(1-2), 261-275.

Marshall, L., D. Nott, and A. Sharma (2004), A comparative study of Markov chain Monte Carlo methods for conceptual rainfall-runoff modeling, *Water Resources Research*, *40*(2), W02501.

Meehl, G. A., C. Tebaldi, H. Teng, and T. C. Peterson (2007), Current and future US weather extremes and El Nino, *Geophys Res Lett*, *34*(20).

Meshgi, A., and D. Khalili (2009), Comprehensive evaluation of regional flood frequency analysis by L- and LH-moments. II. Development of LH-moments parameters for the generalized Pareto and generalized logistic distributions, *Stochastic Environmental Research and Risk Assessment*, *23*(1), 137-152.

Metropolis, N., and S. Ulam (1949), The Monte Carlo Method, *Journal of the American Statistical Association*, *44*(247), pp. 335-341.

Micevski, T., S. W. Franks, and G. Kuczera (2006), Muttidecadal variability in coastal eastern Australian flood data, *Journal of Hydrology*, *327*(1-2), 219-225.

Milly, P. C. D., J. Betancourt, M. Falkenmark, R. M. Hirsch, Z. W. Kundzewicz, D. P. Lettenmaier, and R. J. Stouffer (2008), Climate change - Stationarity is dead: Whither water management?, *Science*, *319*(5863), 573-574.

Min, S. K., W. J. Cai, and P. Whetton (2013), Influence of climate variability on seasonal extremes over Australia, *Journal of Geophysical Research-Atmospheres*, *118*(2), 643-654.

Nakicenovic, N., J. Alcamo, G. Davis, B. de Vries, J. Fenhann, S. Gaffin, K. Gregory, A. Grubler, T. Y. Jung, and T. Kram (2000), Special report on emissions scenarios: a special report of Working Group III of the Intergovernmental Panel on Climate Change*Rep.*, Pacific Northwest National Laboratory, Richland, WA (US), Environmental Molecular Sciences Laboratory (US).

Naulet, R., M. Lang, T. Ouarda, D. Coeur, B. Bobee, A. Recking, and D. Moussay (2005), Flood frequency analysis on the Ardeche river using French documentary sources from the last two centuries, *Journal of Hydrology*, *313*(1-2), 58-78.

Nelder, J. A., and R. W. Wedderburn (1972), Generalized linear models, *Journal of the Royal Statistical Society. Series A (General)*, 370-384.

Neppel, L., B. Renard, M. Lang, P. A. Ayral, D. Coeur, E. Gaume, N. Jacob, O. Payrastre, K. Pobanz, and F. Vinet (2010), Flood frequency analysis using historical data: accounting for random and systematic errors, *Hydrological Sciences Journal-Journal des Sciences Hydrologiques*, *55*(2), 192-208.

Nogaj, M., P. Yiou, S. Parey, F. Malek, and P. Naveau (2006), Amplitude and frequency of temperature extremes over the North Atlantic region, *Geophys Res Lett*, *33*(10).

O'Connell, D. R. H., D. A. Ostenaa, D. R. Levish, and R. E. Klinger (2002), Bayesian flood frequency analysis with paleohydrologic bound data, *Water Resources Research*, *38*(5).

OrtizBevia, M. J., I. Perez-Gonzalez, F. J. Alvarez-Garcia, and A. Gershunov (2010), Nonlinear estimation of El Nino impact on the North Atlantic winter, *Journal of Geophysical Research-Atmospheres*, *115*.

Ouarda, T. B. M. J., and S. El-Adlouni (2011), Bayesian Nonstationary Frequency Analysis of Hydrological Variables, *J Am Water Resour As*, *47*(3), 496-505.

Ouarda, T. B. M. J., C. Girard, G. S. Cavadias, and B. Bobée (2001), Regional flood frequency estimation with canonical correlation analysis, *Journal of Hydrology*, *254*(1), 157-173.

Overeem, A., A. Buishand, and I. Holleman (2008), Rainfall depth-duration-frequency curves and their uncertainties, *Journal of Hydrology*, *348*(1-2), 124-134.

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

Pagé, C., L. Terray, and J. Boé (2008), Projections climatiques à échelle fine sur la France pour le 21ème siècle : les scénarii SCRATCH08, *Technical note, Climate Modelling and Global Change Rep. TR/CMGC/08/64*, CERFACS, Toulouse, France.

Paquet, E., F. Garavaglia, R. Garçon, and J. Gailhard (2013), The SCHADEX method: a semi-continuous rainfall-runoff simulation for extreme flood estimation, *Journal of Hydrology*.

Payrastre, O., E. Gaume, and H. Andrieu (2011), Usefulness of historical information for flood frequency analyses: Developments based on a case study, *Water Resources Research*, *47*.

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

Pociask-Karteczka, J., Z. Nieckarz, and D. Limanowka (2003), Prediction of hydrological extremes by air circulation indices, in *Water Resources Systems - Water Availability and Global Change*, edited by S. Franks, G. Bloschl, M. Kumagai, K. Musiake and D. Rosbjerg, pp. 134-141.

Pscheidt, I., and A. M. Grimm (2009), Frequency of extreme rainfall events in Southern Brazil modulated by interannual and interdecadal variability, *International Journal of Climatology*, *29*(13), 1988-2011.

Pujol, N., L. Neppel, and R. Sabatier (2007), Regional tests for trend detection in maximum precipitation series in the French Mediterranean region, *Hydrological Sciences Journal-Journal des Sciences Hydrologiques*, *52*(5), 956-973.

Queensland Floods Commission of Inquiry 2012 report (http://www.floodcommission.qld.gov.au/publications/final-report)*Rep*.

Queensland Government Operation Queenslander: The State Community, economic and environmental recovery and reconstruction plan 2011–2013 (http://www.qldreconstruction.org.au/publications-guides/reconstruction-plans/state-plan)*Rep*.

Reis, D. S., and J. R. Stedinger (2005), Bayesian MCMC flood frequency analysis with historical information, *Journal of Hydrology*, *313*(1-2), 97-116.

Renard, B. (2006), Détection et prise en compte d'éventuels impacts du changement climatique sur les extrêmes hydrologiques en France, PhD thesis, Université de Grenoble.

Renard, B. (2011), A Bayesian hierarchical approach to regional frequency analysis RID G-1524-2011, *Water Resources Research*, *47*, W11513.

Renard, B., and M. Lang (2007), Use of a Gaussian copula for multivariate extreme value analysis: Some case studies in hydrology, *Advances in Water Resources*, *30*(4), 897-912.

Renard, B., M. Lang, and P. Bois (2006a), Statistical analysis of extreme events in a non-stationary context via a Bayesian framework: case study with peak-over-threshold data, *Stochastic Environmental Research and Risk Assessment*, *21*(2), 97-112.

Renard, B., V. Garreta, and M. Lang (2006b), An application of Bayesian analysis and Markov chain Monte Carlo methods to the estimation of a regional trend in annual maxima, *Water Resources Research*, *42*(12), W12422.

Renard, B., X. Sun, and M. Lang (2013), Bayesian methods for non-stationary extreme value analysis, in *Extremes in a changing climate: Detection, analysis and uncertainty*, edited by A. AghaKouchak, D. Easterling, K. Hsu, S. Schubert and S. Sorooshian, pp. 39-95, Springer Netherlands.

Renard, B., et al. (2008), Regional methods for trend detection: Assessing field significance and regional consistency, *Water Resources Research*, *44*(8).

Ribatet, M., E. Sauquet, J. M. Gresillon, and T. B. M. J. Ouarda (2007), Usefulness of the reversible jump Markov chain Monte Carlo model in regional flood frequency analysis, *Water Resources Research*, *43*(8), W08403.

Rigby, R., and D. Stasinopoulos (2005), Generalized additive models for location, scale and shape, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(3), 507-554.

Ropelewski, C. F., and M. S. Halpert (1987), Global and Regional Scale Precipitation Patterns Associated with the El-Nino Southern Oscillation, *Mon Weather Rev*, *115*(8), 1606-1626.

Rust, H. W., D. Maraun, and T. J. Osborn (2009), Modelling seasonality in extreme precipitation, *European Physical Journal-Special Topics*, *174*, 99-111.

Sabourin, A., and P. Naveau (2013), Bayesian Dirichlet mixture model for multivariate extremes: A re-parametrization, *Computational Statistics & Data Analysis*.

Saji, N. H., B. N. Goswami, P. N. Vinayachandran, and T. Yamagata (1999), A dipole mode in the tropical Indian Ocean, *Nature*, *401*(6751), 360-363.

Salas, J. D., and J. Obeysekera (2013), Revisiting the concepts of return period and risk for nonstationary hydrologic extreme events, *Journal of Hydrologic Engineering*.

Sang, H. Y., and A. E. Gelfand (2009), Hierarchical modeling for extreme values observed over space and time, *Environmental and Ecological Statistics*, *16*(3), 407-426.

Sankarasubramanian, A., and K. Srinivasan (1999), Investigation and comparison of sampling properties of L-moments and conventional moments, *Journal of Hydrology*, *218*(1-2), 13-34.

Sardeshmukh, P. D., G. P. Compo, and C. Penland (2000), Changes of probability associated with El Nino, *Journal of Climate*, *13*(24), 4268-4286.

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

Schubert, S. D., Y. Chang, M. J. Suarez, and P. J. Pegion (2008), ENSO and wintertime extreme precipitation events over the contiguous united states, *Journal of Climate*, *21*(1), 22-39.

Schwarz, G. (1978), Estimating the dimension of a model, *The Annals of Statistics*, *6*(2), pp. 461-464.

Shang, H., J. Yan, and X. Zhang (2011), El Nino-Southern Oscillation influence on winter maximum daily precipitation in California in a spatial model, *Water Resources Research*, *47*.

Sklar, A. (1959), Fonctions de répartition à n dimensions et leurs marges, *Publ. Inst. Stat. Univ. Paris*, *8*, 229-231.

Smith, R. L. (1990), Max-stable processes and spatial extremes, *Unpublished manuscript*.

Somot, S., F. Sevault, M. Deque, and M. Crepon (2008), 21st century climate change scenario for the Mediterranean using a coupled atmosphere-ocean regional climate model, *Global and Planetary Change*, *63*(2-3), 112-126.

Spiegelhalter, D. J., N. G. Best, B. R. Carlin, and A. van der Linde (2002), Bayesian measures of model complexity and fit, *Journal of The Royal Statistical Society Series B-Statistical Methodology*, *64*, 583-616.

Stahl, K., S. Demuth, H. Hisdal, M. J. Santos, R. Verıssimo, and R. Rodrigues (2001), The North Atlantic Oscillation (NAO) and the drought, In: Assessment of the Regional Impact of Droughts in Europe. Final Report, ARIDE*Rep.*, Institute of Hydrology, Freiburg.

Stedinger, J. R. (1983), Confidence-intervals for design-events, *Journal of Hydraulic Engineering-Asce*, *109*(1), 13-27.

Stedinger, J. R., and G. D. Tasker (1985), Regional hydrologic analysis 1. ordinary, weighted, and generalized least-squares compared, *Water Resources Research*, *21*(9), 1421-1432.

Stedinger, J. R., and T. A. Cohn (1986), Flood frequency-analysis with historical and paleoflood information, *Water Resources Research*, *22*(5), 785-793.

Stedinger, J. R., and V. W. Griffis (2011), Getting from here to where? Flood frequency analysis and climate, *JAWRA Journal of the American Water Resources Association*, *47*(3), 506-513.

Stedinger, J. R., R. M. Vogel, S. U. Lee, and R. Batchelder (2008), Appraisal of the generalized likelihood uncertainty estimation (GLUE) method, *Water Resources Research*, *44*.

Sun, X., M. Thyer, B. Renard, and M. Lang (2013), A general regional frequency analysis framework for quantifying the impact of ENSO on summer rainfall in Southeast Queensland (submitted), *Journal of Hydrology*.

Suppiah, R., and K. J. Hennessy (1998), Trends in total rainfall, heavy rain events and number of dry days in Australia, 1910-1990, *International Journal of Climatology*, *18*(10), 1141-1164.

Thyer, M., A. J. Frost, and G. Kuczera (2006), Parameter estimation and model identification for stochastic models of annual hydrological data: Is the observed record long enough?, *Journal of Hydrology*, *330*(1-2), 313-328.

Tramblay, Y., L. Neppel, and J. Carreau (2011), Brief communication "Climatic covariates for the frequency analysis of heavy rainfall in the Mediterranean region", *Natural Hazards and Earth System Science*, *11*(9), 2463-2468.

Tramblay, Y., W. Badi, F. Driouech, S. El Adlouni, L. Neppel, and E. Servat (2012), Climate change impacts on extreme precipitation in Morocco, *Global and Planetary Change*, *82-83*, 104-114.

Trigo, R. M., D. Pozo-Vazquez, T. J. Osborn, Y. Castro-Diez, S. Gamiz-Fortis, and M. J. Esteban-Parra (2004), North Atlantic oscillation influence on precipitation, river flow and water resources in the Iberian peninsula, *International Journal of Climatology*, *24*(8), 925-944.

van Loon, H., and J. C. Rogers (1978), The Seesaw in Winter Temperatures between Greenland and Northern Europe. Part I: General Description., *Mon Weather Rev*, *106(3)*, 296-310.

Vanheerden, J., D. E. Terblanche, and G. C. Schulze (1988), The Southern Oscillation and South-African Summer Rainfall, *J Climatol*, *8*(6), 577-597.

Vicente-Serrano, S. M., and J. M. Cuadrat (2007), North Atlantic oscillation control of droughts in north-east Spain: evaluation since 1600 A. D, *Climatic Change*, *85*(3-4), 357-379.

Vidal, J.-P., E. Martin, L. Franchistéguy, M. Baillon, and J.-M. Soubeyroux (2010), A 50-year high-resolution atmospheric reanalysis over France with the Safran system, *International Journal of Climatology*, *30*(11), 1627-1644.

Wagener, T., and J. Kollat (2007), Numerical and visual evaluation of hydrological and environmental models using the Monte Carlo analysis toolbox, *Environ. Modell. Softw.*, *22*(7), 1021-1033.

Wan, S. Q., Y. L. Hu, Z. Y. You, J. P. Kang, and J. G. Zhu (2013), Extreme monthly precipitation pattern in China and its dependence on Southern Oscillation, *International Journal of Climatology*, *33*(4), 806-814.

Weiher, R. F. (1999), *Improving El Niño forecasting: the potential economic benefits*, National Oceanic and Atmospheric Administration, Office of Policy and Strategic Planning.

Westra, S., L. V. Alexander, and F. W. Zwiers (2012), Global increasing trends in annual maximum daily precipitation, *Journal of Climate*(2012).

Wikle, C. K., L. M. Berliner, and N. Cressie (1998), Hierarchical Bayesian space-time models, *Environmental and Ecological Statistics*, *5*, 117-154.

Wilby, R. L., G. O'Hare, and N. Barnsley (1997), The North Atlantic Oscillation and British Isles climate variability, 1865–1996, *Weather*, *52*, 266-276.

Wilk, M. B., and R. Gnanadesikan (1968), Probability plotting methods for the analysis for the analysis of data, *Biometrika*, *55*(1), 1-17.

Wu, A. M., W. W. Hsieh, and A. Shabbar (2005), The nonlinear patterns of North American winter temperature and precipitation associated with ENSO, *Journal of Climate*, *18*(11), 1736-1752.

Wu, R. G., Z. Z. Hu, and B. P. Kirtman (2003), Evolution of ENSO-related rainfall anomalies in East Asia, *Journal of Climate*, *16*(22), 3742-3758.

Yu, P. S., T. C. Yang, and C. S. Lin (2004), Regional rainfall intensity formulas based on scaling property of rainfall, *Journal of Hydrology*, *295*(1-4), 108-123.

Zhai, P. M., X. B. Zhang, H. Wan, and X. H. Pan (2005), Trends in total precipitation and frequency of daily precipitation extremes over China, *Journal of Climate*, *18*(7), 1096-1108.

Regional frequency analysis of precipitation accounting for climate variability and change

Extreme precipitations and their consequences (floods) are one of the most threatening natural disasters for human beings. In engineering design, Frequency Analysis (FA) techniques are an integral part of risk assessment and mitigation. FA uses statistical models to estimate the probability of extreme hydrological events which provides information for designing hydraulic structures. However, standard FA methods commonly rely on the assumption that the distribution of observations is identically distributed. However, there is now a substantial body of evidence that large-scale modes of climate variability (e.g. El-Niño Southern Oscillation, ENSO; Indian Ocean Dipole, IOD; etc.) exert a significant influence on precipitation in various regions worldwide. Furthermore, climate change is likely to have an influence on hydrology, thus further challenging the "identically distributed" assumption. Therefore, FA techniques need to move beyond this assumption. In order to provide a more accurate risk assessment, it is important to understand and predict the impact of climate variability/change on the severity and frequency of hydrological events (especially extremes).

This thesis provides an important step towards this goal, by developing a rigorous general climate-informed spatio-temporal regional frequency analysis (RFA) framework for incorporating the effects of climate variability on hydrological events. This framework brings together several components (in particular spatio-temporal regression models, copula-based modeling of spatial dependence, Bayesian inference, model comparison tools) to derive a general and flexible modeling platform. In this framework, data are assumed to follow a distribution, whose parameters are linked to temporal or/and spatial covariates using regression models. Parameters are estimated with a Monte Carlo Markov Chain method under the Bayesian framework. Spatial dependency of data is considered with copulas. Model comparison tools are integrated. The development of this general modeling framework is complemented with various Monte-Carlo experiments aimed at assessing its reliability, along with real data case studies.

Two case studies are performed to confirm the generality, flexibility and usefulness of the framework for understanding and predicting the impact of climate variability on hydrological events. These case studies are carried out at two distinct spatial scales:

- Regional scale: Summer rainfall in Southeast Queensland (Australia): this case study analyzes the impact of ENSO on the summer rainfall totals and summer rainfall maxima. A regional model allows highlighting the asymmetric impact of ENSO: while La Niña episodes induce a significant increase in both the summer rainfall totals and maxima, the impact of El Niño episodes is found to be not significant.
- Global scale: a new global dataset of extreme precipitation including 11588 rainfall stations worldwide is used to describe the impact of ENSO on extreme precipitations in the world. This is achieved by applying the regional modeling framework to 5x5 degrees cells covering all continental areas. This analysis allows describing the pattern of ENSO impact at the global scale and quantifying its impact on extreme quantiles estimates. Moreover, the asymmetry of ENSO impact and its seasonal pattern are also evaluated.